HOCHSCHULE
ESSLINGEN

Informatik und
Informationstechnik

IT Innovationen

Band 32

Januar 2024

HOCHSCHULE ESSLINGEN

Grußwort der Fakultät

Liebe Leserinnen und Leser,

Die Hochschule Esslingen und mit ihr die Informatik blickt auf eine lange Historie von Absolventen und Abschlussarbeiten zurück. Dieser 32. Band der IT Innovationen reiht sich in diese Geschichte ein. Nun könnte man meinen, es käme auf den 32. Band einer solch langen Reihe nicht an. Man könnte sich fragen, was für eine Auswirkung hätte es, wenn es diesen Band und die darin erwähnten Abschlussarbeiten nicht gäbe. Was würde uns verloren gehen? Würde das jemand merken?

Braucht ein 32. Band überhaupt ein Grußwort, wie Sie es hier lesen? Welchen Fortschritt kann ein solches Dokument über die vorigen 31 Bände hinaus überhaupt noch bringen? Lassen Sie uns ein Experiment wagen. Unter folgendem Link können Sie kundtun, dass Sie dieses Vorwort gelesen haben. Vielleicht erleben wir dabei ja eine Überraschung: https://terminplaner4.dfn.de/Vorwort

Um die Frage nach der Notwendigkeit dieses Bandes zu beantworten muss man sich vergegenwärtigen, wofür die IT Innovationen stehen und woraus sie sich zusammensetzen. Jeder einzelne Beitrag befasst sich mit ungelösten Themen der angewandten Wissenschaft. Und Wissenschaft hat die Eigenschaft, dass sie oft langsam und zäh voranschreitet. Durch harte Arbeit, viele Fehler, Demut vor der Sache und einer konsequenten Beharrlichkeit kann man der Welt ein kleines bisschen Wissen abtrotzen. Die Autoren der Artikel dieses 32. Bandes haben sich dieser Herausforderung gestellt. Das Ergebnis sind Artikel, die allesamt etwas zum Wissenszuwachs in der Informatik beitragen. Sie enthalten Erkenntnisse die es zuvor nicht gab. Schon dies macht diesen Band wichtig. Auf der anderen Seite sind viele Arbeiten der Startpunkt für weitere Überlegungen und Innovationen. So entstehen Folgearbeiten und neue Fragestellungen aus den Beiträgen unserer Studierenden. Darum wird auch Band 32 ein Band sein, den man nicht weglassen kann und nicht weglassen darf. Sonst gibt es vielleicht keinen Band 33. Schritt für Schritt bringt uns jede Abschlussarbeit nach vorn und bereichert die Vielfalt der Informatik.

Die Breite der in diesem Band behandelten Themen ist enorm und zeichnet diesen ihn ganz besonders aus. Band 32 ist ein Rekordband. Insgesamt 102 Artikel tragen zum Kenntnisstand der Informatik in der 3D-Computergrafik, IT-Sicherheit, Software-Entwicklung, künstlichen Intelligenz und sogar beim Training im Mannschaftssport bei. So viele Artikel zählt keiner der bisherigen Bände. Machen Sie sich ein Bild dieser Vielfalt mit den vorliegenden Kurzbeschreibungen der Abschluss-arbeiten des Wintersemesters 2023/2024.

Viel Freude beim Lesen wünscht Ihnen

Ihr Prof. Dr. Tobias Heer, Dekan

HOCHSCHULE ESSLINGEN

IMPRESSUM

ERSCHEINUNGSORT

73732 Esslingen am Neckar

HERAUSGEBER

Prof. Dr. Tobias Heer
Dekan der Fakultät Informatik und Informationstechnik
der Hochschule Esslingen - University of Applied Sciences

REDAKTIONSANSCHRIFT

Hochschule Esslingen - University of Applied Sciences
Fakultät Informatik und Informationstechnik
Flandernstraße 101
73732 Esslingen am Neckar

Telefon	+49(0)711.397-4210
Telefax	+49(0)711.397-4214
E-Mail	it@hs-esslingen.de
Website	www.hs-esslingen.de/it

REDAKTION, DESIGN, LAYOUT und SATZ

Dipl.-Inform.(FH) Rolf Gassner

Hochschule Esslingen - University of Applied Sciences
Fakultät Informatik und Informationstechnik
Flandernstraße 101
73732 Esslingen am Neckar

ERSCHEINUNGSWEISE

Einmal pro Semester, jeweils Januar und Juni

ISSN 1869-6457

Integration von Softwareentwicklungsdaten zur Visualisierung auf einem Dashboard

Christian Achstetter **Astrid Beck**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma Philips Medizin Systeme Böblingen GmbH, Böblingen

Einleitung

In der modernen Softwareentwicklung setzen Unternehmen vor allem auf Cloud-Computing-Frameworks wie Microsoft Azure, wobei alle am Entwicklungsprozess beteiligten Daten einen festen Speicherort zugewiesen bekommen. Der Quellcode ist durch Versionskontrollsysteme abgesichert und wird in Azure Repositories abgelegt. Die aus dem Code generierten, kompilierten Dateien werden in Azure Artifacts gespeichert, und auch die Tests haben ihren Platz in der integrierten Testdatenbank. Im Entwicklungsprozess fallen jedoch eine Vielzahl weitere Daten an, die zwar eine geringere Priorität haben, aber trotzdem relevant sind. In diese Kategorie fallen unter anderem auch Tooling-Daten, die zu kleinen Anwendungen gehören, welche Entwickler*innen bei ihrer Arbeit unterstützen. Die Verfügbarkeit von Tools ist daher ein entscheidender Faktor, um die Effizienz der Softwareentwicklung zu maximieren. Die Erstellung solcher Tools kann sich allerdings kompliziert gestalten, da die Daten oft aus verschiedenen Datenbanken und Dateisystemen akquiriert werden müssen. Dies ist der Fall, weil die Aufbereitung und Speicherung dieser Daten je nach Abteilung sehr unterschiedlich gehandhabt werden kann. In der Folge entsteht eine unübersichtliche Datenlandschaft aus verteilten Netzlaufwerken und Datenbanken, die den Entwicklungsprozess erschwert. Das Ziel dieser Arbeit ist es, einen geeigneten Integrationsansatz zu ermitteln, der es ermöglicht, Daten, die auf einem Dashboard angezeigt werden, zentral zu speichern und abzurufen. Sowohl das Dashboard an sich als auch die dafür benötigten Daten dienen dabei als Repräsentation für gängige Anwendungsfälle in Tooling-Szenarien, von denen zu einem späteren Zeitpunkt die in dieser Arbeit vorgestellte Datenintegrationslösung profitieren soll.

Implementierung

Zuerst werden verschiedene Ansätze der Datenintegration ermittelt. Durch genauere Betrachtung und Gegenüberstellung der einzelnen Prozesse kristallisiert sich heraus, dass Extract Transform Load (ETL) am besten geeignet ist. Als Speicherlösung wird die Backend-Software Appwrite eingesetzt [1]. Da Semantic Web Machine Learning (SWeML), was eine Symbiose aus dem Semantic Web und Machine Learning (ML) darstellt, großes Potenzial aufweist, wird zudem noch ein Ansatz der semantischen Integration prototypisch durchgeführt [2]. Die Implementierung wird in zwei Schritte aufgeteilt: Zuerst werden anhand eines Vorversuchs der gewählte Ansatz sowie die Technologien getestet. Aufgrund der Erkenntnisse werden im Anschluss sowohl die Daten, welche es zu visualisieren gilt, als auch das Dashboard selbst implementiert. Der Vorversuch behandelt den Anwendungsfall eines Tools, dessen Aufbau in Abbildung 1 abgebildet ist.

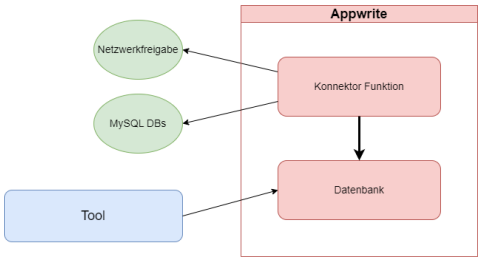


Abb. 1: Architektur ETL-Implementierung [3]

Es wird dabei eine Appwrite-Funktion entwickelt, welche die Integration per ETL-Operation durchführt, sowie eine Command Line Interface-Anwendung, die das Tool repräsentiert. Es stellt sich heraus, dass die ETL-Umsetzung des Vorversuchs sämtliche Anforderungen erfüllt, weshalb auch die Daten des Dashboards, welches in Abbildung 2 zu sehen ist, auf dieselbe Weise integriert werden.

Abb. 2: Dashboard [3]

Das Dashboard besteht aus einer Webapp, die mit dem Framework React umgesetzt ist und dessen User Interface-Komponenten in TypeScript geschrieben sind. Es stellt die Startzeiten von Patientenmonitoren, die Anzahl der Bugs eines Softwarereleases sowie das Menü der Kantine an.

Ausblick

Aufgrund der Tatsache, dass die Integrationslösung eine einheitliche Schnittstelle bereitstellt, können diese Künstliche Intelligenz (KI)-Prozessen zugeführt werden. Retrieval-Augmented Generation (RAG) ist ein sehr aktueller Ansatz im Bereich von Natural Language Processing (NLP). Er entstammt einem Fachartikel unter dem Hauptautor Patrick Lewis [4]. RAG will Large Language Model (LLM)s um die Möglichkeit erweitern, externe Ressourcen benutzen zu können. Im besten Fall kann somit ein generisch trainiertes LLM konkrete Fragen zu Dokumentationen liefern und diese in der Antwort mit verlinken. Auch wenn sich die semantische Integration in den untersuchten Anwendungsfällen als weniger praktikabel als ETL zeigt, ist sie technisch trotzdem möglich. Die Implementierung kann als Ausgangspunkt für weitere Versuche dienen. Das Backend bildet nicht nur durch die einheitliche Schnittstelle eine prädestinierte Basis für das Erkunden KI-basierter Technologien, sondern erfüllt gleichzeitig seine primäre Funktion als Tooling-Backend. Es präsentiert sich als äußerst vielseitige Integrationslösung, die über den spezifischen Anwendungsfall der Integration von Softwareentwicklungsdaten für ein Dashboard hinaus weitreichende Einsatzmöglichkeiten bietet.

Literatur und Abbildungen

[1] Appwrite Appwrite. Appwrite. https://appwrite.io/company/abou, 2023.

[2] Anna Breit. Combining Machine Learning and Semantic Web: A Systematic Mapping Study. https://dl.acm.org/doi/10.1145/ 3586163, 06 2023.

[3] Eigene Darstellung.

[4] Patrick Lewis. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. http://arxiv.org/abs/2005.11401, 04 2021.

Supply Chain 2.0 - Digitalisierungspotentiale und Analytics in der Wertschöpfungskette

Oezlem Akar **Dirk Hesse**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt an der Fakultät Informatik und Informationstechnik, Esslingen

Einleitung

Digitalisierung ist allgegenwärtig. Diese beschreibt den zunehmenden Einsatz automatisierter Technologien mit der Verwendung von digitalen Geräten. Die Auswirkungen auf Bereiche der realen Welt wie zum Beispiel Wirtschaft, Gesellschaft und Umwelt sind sehr groß. Unter anderem die Lieferkette, die heutzutage eine wichtige Rolle spielt und viel flexibler auf Änderungen sowie äußere Faktoren reagieren kann. Dies führt dazu, dass sämtliche Kosten und Risiken minimiert werden können. Genauso steigt die Chance für die Wettbewerbsfähigkeit auf unterschiedlichen Märkten, die vor einigen Jahren nicht sehr hoch war [4]. Die digitale Transformation birgt jedoch andere Herausforderungen, die in der Bachelorthesis analysiert werden.

Ziel der Arbeit

Die Bachelorthesis gibt einen Überblick über die theoretischen Grundlagen des herkömmlichen Supply Chain Managements sowie der digitalen Supply Chain und legt den Fokus auf die Digitalisierungspotentiale. Wie bereits erwähnt, können einige Schwachstellen und Risiken identifiziert werden. Das Ziel dieser Bachelorthesis ist es, aus diesen Problemen einige Handlungsempfehlungen abzuleiten. Neben den Herausforderungen stellt sich eine andere wichtige Forschungsfrage, die folgendermaßen lautet: **Sind die Unternehmen in der heutigen Welt für eine digitalisierte Supply Chain ausgereift?** Um diese Frage beantworten zu können, gilt es, alle Aspekte der Supply Chain genau zu untersuchen und zu analysieren.

Supply Chain Management

Das Supply Chain Management (SCM) repräsentiert den gesamten Vorgang von der Rohstoffgewinnung, -umwandlung bis hin zur Lieferung an den Kunden. Die folgende Abbildung 1 zeigt einen beispielhaften Ablauf des SCMs von der Planung bis hin zur Beschaffung.

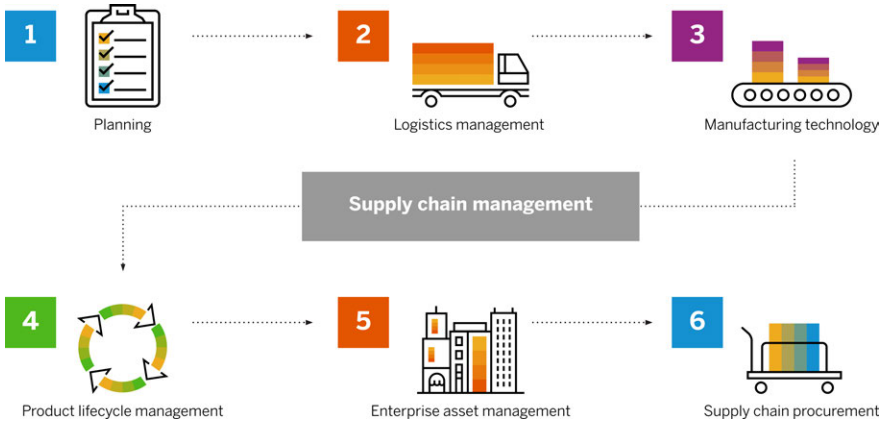


Abb. 1: Ablauf des SCMs [3]

Die Aktivitäten innerhalb des SCMs können sein: Beschaffung, Design, Produktion, Lagerhaltung, Versand und Vertrieb. Werden alle Schritte richtig verfolgt, kann eine höhere Optimierung der Qualität und Kosten z.B. Transportkosten erzielt werden. Der Fokus der Lieferkette liegt primär auf den Kunden. Ändern sich sämtliche Kundenanforderungen, kann das Unternehmen mithilfe eines optimalen SCMs flexibler und schneller auf Änderungen reagieren. Das SCM kann in zwei Arten unterteilt werden, die als unternehmensinterne und unternehmensintegrierte Supply Chain bezeichnet werden [4].

- **Unternehmensinterne Supply Chain**: Hier geht es um die Stufen innerhalb eines Betriebs. Ein Beispiel dafür ist der Montagebetrieb mit seinen Stufen Wareneingang, Vormontage, Zwischenlager und Endmontage.
- **Unternehmensintegrierte Supply Chain**: Die unternehmensintegrierte Supply Chain besteht aus einer inputseitigen und einer outputseitigen Lieferkette. Generell geht es hier darum, welche Zusammenarbeit mit externen Partnern besteht.

Supply Chain 2.0

Supply Chain 2.0 ist eine Erweiterung und Weiterentwicklung der herkömmlichen Lieferkette. Der Wandel innerhalb der Lieferkette ist in den heutigen Unternehmen deutlich zu erkennen. Anfangs lag er Fokus auf den operativen Bereich der Bereitstellung und auf der sicheren Lieferung an den Kunden. Dies hat sich zu einem eigenständigen SCM entwickelt, welches sogar vom CSO gesteuert wird. Industrie 4.0 nimmt hierbei eine wichtige Rolle ein. Diese hat viele Unternehmen zur großen Veränderung der Lieferketten geführt, da ein Umdenken notwendig war [2]. Die digitalisierte Supply Chain kann zu vielen Vorteilen führen. Ein Beispiel dafür ist die reibungslose Kommunikation. Dank moderner Technologien wie zum Beispiel Tracking und Monitoring kann genau beobachtet werden, wo sich das Produkt zu einer gewissen Zeit befindet. Die Schnelligkeit ist ein weiterer Vorteil. Moderne Trends vereinfachen generell die schnelle Prozessabwicklung und Optimierung sämtlicher Abläufe. Wettbewerbsvorteile sind auch damit verbunden. Je schneller ein Unternehmen auf Änderungen reagiert, desto höher sind die Wettbewerbschancen [2].

Digitalisierungspotentiale und Probleme

Im Laufe der Zeit wurden verschiedene Technologien für die Supply Chain entwickelt. Hier fokussiert man sich auf bestimmte Technologien, da es zu viele gibt. Dazu gehören:

- **Blockchain**
- **Internet of Things (IoT)**
- **Sensortechnologien**
- **Robotik**
- **Digital Identifiers z.B. RFID, Barcodes** [2]

Hier konnten einige Probleme selbstständig identifiziert werden, da die Implementierung der Technologien problematisch werden kann. Das erste Problem können die sehr hohen Kosten sein. Die gesamte Ausstattung mit den modernen Technologien und die Umsetzung der Lieferkette können sehr teuer werden. Auch kann die Implementierung viel Zeit in Anspruch nehmen. Ein zweites Problem stellt die Zahl der Cyberangriffe dar. In der modernen Supply Chain gibt es einen kontinuierlichen Datenaustausch. Sobald es eine Sicherheitslücke gibt, könnte ein Hackerangriff stattfinden. Falsche Prognosen stellen ein drittes Problem dar. Systeme sind generell anfällig für technische Ausfälle. Existiert innerhalb der Supply Chain ein digitaler Fehler, können falsche Daten geliefert werden [2].
Laut einer Umfrage gibt es viele Unternehmen, die gewisse Technologien einsetzen und diese Probleme umgehen. Die Abbildung 2 visualisiert den Einsatz der digitalen Methoden in der Supply Chain.

Ausblick

Der derzeitige Implementierungsstand der Technologien ist nicht gering. Deutsche Unternehmen sind in einem guten Zustand und können somit langfristige Vorteile erzielen. Der Abbildung 2 kann entnommen werden, dass Cloud Computing, Robotik und Automatisierung sowie Big Data die meistgenutzten Technologien sind. Jedoch sollten die anderen digitalen Mitteln genauso berücksichtigt werden, um weitere Vorteile zu erreichen. Wichtig hierbei ist, dass die Kosten und der Nutzen abgewogen werden. Je nach Wettbewerbsstrategie sollten die Technologien dementsprechend implementiert werden. Genaue Strukturierungen und Planungen sind für die Ausreifung der Unternehmen und den Einsatz der Technologien essentiell. Ein eingesetztes System kann bei ausfallender Wartung oder falschem Einsatz andere Ergebnisse liefern und das Risiko innerhalb des Unternehmens steigern [1].

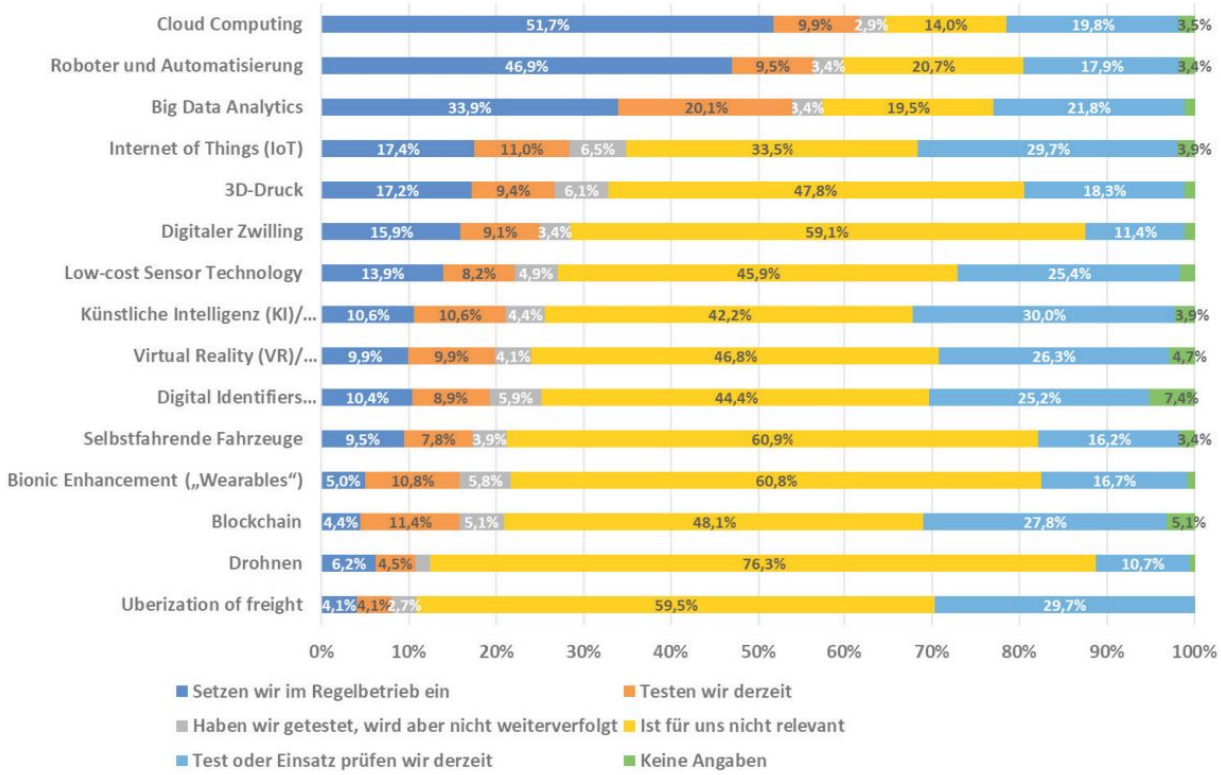


Abb. 2: Einsatz Supply Chain Technologien [1]

Literatur und Abbildungen

[1] Bundesverband Materialwirtschaft eingetragener Verein. *Digitalisierung in Supply Chains*. eingetragener Verein, Bundesverband Materialwirtschaft, 2019.

[2] Hermes Germany GmbH. Logistik 4.0: Diese Technologien erobern das SCM. https://www.hermes-supply-chain-blog.com/logistik-4-0-zukunftstechnologien-scm/, 2021.

[3] SAP Deutschland KG. Was ist Supply Chain Management (SCM)? https://www.sap.com/germany/products/scm/what-is-supply-chain-management.html, 2023.

[4] H. Werner. *Supply Chain Management - Grundlagen, Strategien, Instrumente und Controlling*. Werner, H., 2020.

Grundlagen zur Feinabstimmung von Stabilen Diffusionsmodellen: Voruntersuchungen zur Sicherstellung von produktrichtigen Bildern für industrielle Anwendungen

Kansu Akguen **Dieter Morgenroth**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt an der Fakultät Informatik und Informationstechnik, Esslingen

Einleitung

Die Generierung von realistischen Bildern aus textuellen Beschreibungen mithilfe von Deep Learning Text-zu-Bild-Generatoren hat in den letzten Jahren erhebliche Fortschritte gemacht und beeindruckende Ergebnisse erzielt. Diese Modelle ermöglichen es, visuelle Inhalte auf der Grundlage von Textbeschreibungen zu erstellen und eröffnen spannende Anwendungsmöglichkeiten in Bereichen wie Spieleentwicklung, Grafikdesign oder in der Industrie.
Nichtsdestotrotz steckt die kommerzielle Nutzung dieser Verfahren, speziell im Bereich der Produktbilder, noch in den Kinderschuhen. Denn die automatische Generierung von Produktbildern birgt Herausforderungen in Bezug auf Produktkorrektheit und Konsistenz.
Das Hauptziel dieser Bachelorarbeit besteht darin, die grundlegenden Konzepte und Techniken zur Feinabstimmung von Deep Learning Text-zu-Bild-Generatoren, insbesondere Stabiler Diffusionsmodelle, zu untersuchen. Hierzu werden verschiedene Ansätze zur Optimierung der Parameterwerte betrachtet und analysiert, um die Qualität der generierten Bilder zu verbessern. Dabei sind folgende Aspekte von besonderem Interesse: Wie werden textuelle Beschreibungen in hochwertige Bilder umgewandelt? Welche Techniken zur Feinabstimmung der Modelle gibt es?
Die zentrale Fragestellung der Arbeit lautet daher: Wie können die Grundlagen und Methoden zur Feinabstimmung Stabiler Diffusionsmodelle optimiert werden, um die Qualität der generierten Bilder zu steigern und die Zuverlässigkeit des Systems in industriellen Anwendungen zu erhöhen?

Grundlagen Stabiler Diffusionsmodelle

Stabile Diffusionsmodelle, eine Unterklasse der generativen Modelle in der künstlichen Intelligenz, basieren auf dem Konzept, komplexe Datenverteilungen effektiv zu lernen und zu imitieren. Ihre Hauptaufgabe besteht darin, neue Datenpunkte zu generieren, die einer bestimmten Verteilung folgen, wie sie in Trainingsdatensätzen beobachtet wird. Diese Modelle sind besonders effektiv in der Erzeugung visueller Inhalte, was sie für die Bildverarbeitung in industriellen Anwendungen interessant macht [3].
Stabile Diffusionsmodelle sind typischerweise als tiefe neuronale Netze strukturiert. Sie nutzen eine Kombination aus zwei Hauptkomponenten: einem Generator und einem Diskriminator. Der Generator ist darauf trainiert, Daten zu erzeugen, die denen des Trainingssets ähneln, während der Diskriminator zwischen generierten und echten Daten unterscheidet. Durch diesen Wettbewerb lernen beide Netzwerke, ihre Leistung kontinuierlich zu verbessern [2].

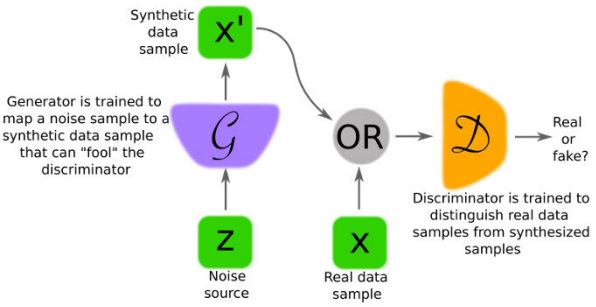


Abb. 1: In dieser Abbildung sind die beiden Modelle, die während des Trainingsprozesses für ein GAN erlernt werden, der Diskriminator (D) und der Generator (G) [1]

Training und Feinabstimmung

Zunächst werden die Modelle mit einer anfänglichen Parameterkonfiguration versehen. Diese Anfangsparameter können entweder zufällig gewählt oder aus vorherigen Modellen übernommen werden, die auf ähnlichen Datensätzen trainiert wurden.

Während des Trainings werden die Eingabedaten durch das Netzwerk geleitet und die Ausgabe wird mit den tatsächlichen Daten verglichen. Der Unterschied zwischen beiden (der Fehler) wird mittels eines Algorithmus namens Backpropagation durch das Netzwerk zurückgeführt, um die Gewichte des Modells anzupassen. Dieser Prozess wird iterativ wiederholt, um die Genauigkeit des Modells zu verbessern.
Nachdem das Modell eine grundlegende Fähigkeit zur Datenreproduktion erlangt hat, beginnt die Phase der Feinabstimmung. Hierbei werden spezifische Anpassungen an den Modellparametern vorgenommen, um die Leistung für bestimmte Anwendungen oder spezifische Arten von Eingabedaten zu optimieren. Dies ist besonders wichtig in der industriellen Bildverarbeitung, wo Genauigkeit und Konsistenz entscheidend sind.
Die Verfeinerung des Trainings kann beispielsweise mit LoRA (Low-Rank Adaptation) oder Dreambooth erfolgen. Letzteres hat in der jüngsten Zeit aufgrund seiner Effektivität und Anpassungsfähigkeit an spezifische Anforderungen und Kontexte Aufmerksamkeit erregt.
Ein wesentlicher Vorteil von DreamBooth liegt in seiner Effizienz. Im Gegensatz zu traditionellen Methoden, die eine umfangreiche Datensammlung und langwierige Trainingssitzungen erfordern, benötigt DreamBooth nur eine begrenzte Anzahl an Beispielen, um wirksame Anpassungen vorzunehmen [4]. In den nachfolgenden Abbildungen Abb. 2 und Abb. 3 wird dies am Beispiel eines Feinabstimmungsprozesses verdeutlicht.

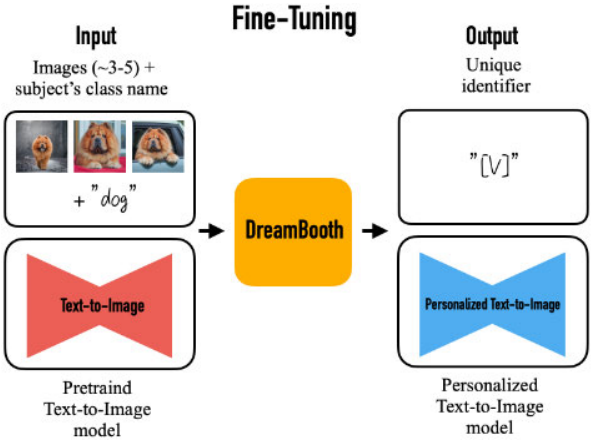


Abb. 2: Fein-Tuning eines Modells mit Dreambooth durch den Input von lediglich 3-5 Bildern und der Bestimmung eines „Unique Identifier“ [4]

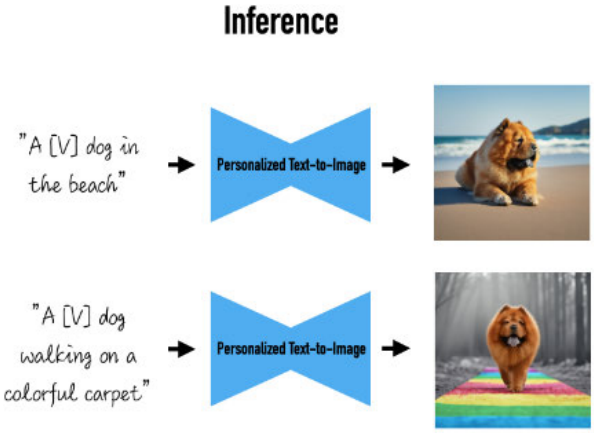


Abb. 3: Die Schlussfolgerung des Feinabstimmungsprozesses durch die Eingabe des „Unique Identifier“ im Textprompt [4]

Nach Abschluss des Trainings ist eine kontinuierliche Überwachung erforderlich, um sicherzustellen, dass das Modell wie beabsichtigt funktioniert. Anpassungen können erforderlich sein, um Überanpassung (Overfitting) zu vermeiden, bei der das Modell zu spezifisch auf die Trainingsdaten abgestimmt ist und nicht gut auf neue, unbekannte Daten generalisiert.

Ausblick

Der Ausblick dieser Arbeit legt den Fokus auf die Optimierung Stabiler Diffusionsmodelle, insbesondere hinsichtlich der Qualität der generierten Bilder und der Zuverlässigkeit in industriellen Anwendungen.
Ein weiterer wichtiger Aspekt in industriellen Anwendungen ist die Effizienz. Daher könnten zukünftige Forschungen darauf abzielen, die Geschwindigkeit der Bildverarbeitung und -generierung durch stabile Diffusionsmodelle zu steigern, um eine nahezu Echtzeit-Verarbeitung zu ermöglichen.
Mit der zunehmenden Leistungsfähigkeit von KI-Modellen werden auch ethische Fragen und Datenschutzbedenken immer relevanter. Zukünftige Arbeiten könnten sich daher mit der Entwicklung von Richtlinien und Techniken beschäftigen, die den verantwortungsvollen Umgang mit KI in industriellen Anwendungen sicherstellen.

Literatur und Abbildungen

[1] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath. Generative adversarial networks: An overview. *IEEE signal processing magazine*, 35:53–65, 2018.

[2] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[3] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[4] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Computer Vision and Pattern Recognition*. IEEE/CVF Conference, 2023.

Datengetriebene Mustererkennung in Maschinendaten: Eine Analyse zur Identifikation von Zusammenhängen zwischen Nutzungs- und Fehlerverhalten

Julia Alas **Gabriele Gühring**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt an der Fakultät Informatik und Informationstechnik, Esslingen

Motivation und Zielsetzung

Die kontinuierlich wachsende Verfügbarkeit umfangreicher Datenbestände eröffnet insbesondere für Unternehmen im Bereich der industriellen Fertigung zahlreiche Optimierungsmöglichkeiten. Durch den gezielten Einsatz effizienter, datengetriebener Analysen können sie ihre Wertschöpfung signifikant steigern. Vor allem maschinen- und kundenbezogene Daten bieten hierzu eine Vielzahl an Potentialen für die Gewinnung neuer Erkenntnisse bis hin zur Optimierung der Maschinen und der Produktionsabläufe. Dies gilt sowohl im Hinblick auf die Produktivität als auch die Haltbarkeit der Maschinen beziehungsweise die frühzeitige Erkennung von möglichen Problemen. Für eine erkenntnisreiche und aussagekräftige Datenanalyse ist es einerseits von großer Bedeutung, einen möglichst repräsentativen und umfangreichen Datensatz zu erheben, anderseits rückt hierdurch auch der Fokus auf die Extraktion der wesentlichen Informationen in den Vordergrund, um gut strukturierte und sinnvolle Muster aus diesen Daten gewinnen zu können. Die verschiedenen Methoden der Mustererkennung werden unter dem Begriff Data Mining zusammengefasst [2]. In der Masterthesis werden einige dieser Techniken verwendet, um Erkenntnisse über die Produktions- und Stillstandzeiten zu erhalten, indem die Zusammenhänge zwischen Nutzungs- und Fehlerverhalten der Maschinen gewonnen werden.

Data Mining

Die in Abbildung 1 dargestellten Methoden stellen die elementaren Bausteine des Data Mining dar und lassen sich in die beiden Überkategorien Überwachtes- und Unüberwachtes Lernen einteilen [2].

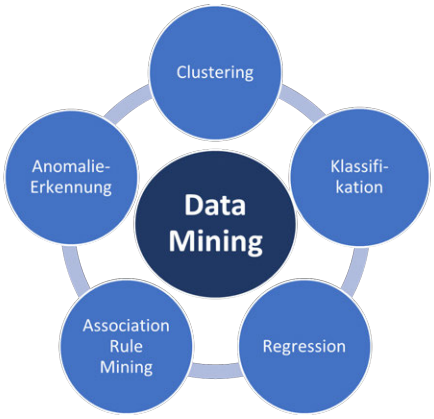


Abb. 1: Übersicht über Data Mining Techniken [3]

Clustering ist eine Methode des Unüberwachten Lernens, die darauf abzielt, ähnliche Daten anhand extrahierter Merkmale zu gruppieren, um somit intrinsische Strukturen in den Daten zu identifizieren. Die Klassifikation stellt dem gegenüber eine Methode des Überwachten Lernens dar und dient der Zuordnung der Daten zu vorab festgelegten Klassen mithilfe ausgewählter Merkmale. Alle Methoden des Überwachten Lernens benötigen jedoch einen annotierten Datensatz, der vorzugsweise eine ausgeglichene Anzahl an Beispielsdaten für jede Klasse beinhaltet. Ein weiteres Verfahren des Überwachten Lernens stellt die Regression dar. Hierbei wird versucht, einen numerischen Wert basierend auf den Trainingsdaten vorherzusagen. Der zugewiesene Wert repräsentiert die geschätzte Beziehung zwischen den untersuchten Merkmalen [6]. Die Anomalie Erkennung fällt wiederum unter den Überbegriff des Unüberwachten Lernens und dient der Erkennung von einzelnen Ausreißern in den Daten, die sich im Vergleich zu dem typischen Spektrum der Merkmale stark unterscheiden [7]. Association Rule Mining, das auch dem Unüberwachten Lernen zugeordnet werden kann, zielt darauf ab, Regeln für

das gemeinsame Auftreten von Daten zu identifizieren. Dies ermöglicht die Erkennung von wiederkehrenden Mustern in den Daten [5]. Alle diese Methoden tragen dazu bei, wesentliche Strukturen in den Daten aufzuzeigen und neue Erkenntnisse und Optimierungen für ein Unternehmen und deren Prozesse zu finden.
Im Bereich des Data Mining spielen nicht nur zeitinvariante Daten eine zentrale Rolle, auch Zeitreihendaten stehen im Fokus. Diese benötigen aufgrund ihrer zusätzlichen zeitlichen Komponente oftmals eine gesonderte Betrachtung. Neben der Vorhersage von Zeitreihen beschäftigt sich das Data Mining im Bereich der Zeitreihen vor allem mit der Entdeckung von Mustern in zeitlichen Sequenzen. Dies kann durch Methoden wie die der Klassifikation und des Rule Mining erreicht werden. Beim Clustering spielt hierbei die Wahl der richtigen Distanzfunktion eine wichtige Rolle [4].
Ein weiterer wichtiger Aspekt, der bei den oben genannten Verfahren des Data Mining zu beachten ist, sind die Datentypen der ausgewählten Merkmale. Die simultane Verwendung von numerischen und kategorischen Daten bei datengetriebenen Algorithmen ist hier zu berücksichtigen, insbesondere wenn beispielsweise Distanzfunktionen berechnet werden [1].
Auch die in der Masterthesis verwendete Datenbasis beinhaltet verschiedene Datentypen. Eine allgemeine Übersicht über die Daten, die während eines Maschinendurchlaufs erfasst werden, ist in Abbildung 2 dargestellt. Zum einen werden Message-Log Daten als nicht-äquidistante Zeitreihendaten erfasst, welche das Fehlerverhalten der Maschine widerspiegeln. Das Nutzungsverhalten kann ebenfalls über den zeitvarianten Maschinenstatus beschrieben werden. Zusätzlich wird diese Information mit zeitinvarianten numerischen und kategorischen Merkmalen ergänzt. In dem vorliegenden Datensatz für die Masterthesis finden sich somit alle oben beschriebenen Datentypen wieder. Der Umgang mit Mixed-Datatypes in datengetriebenen Analysen muss daher beachtet werden.

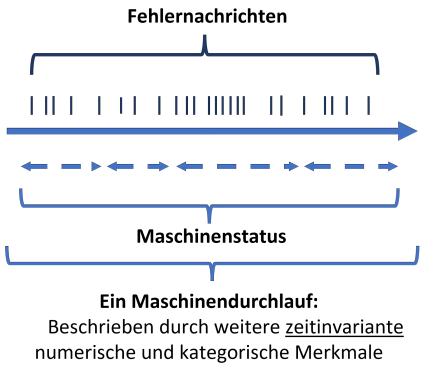


Abb. 2: Übersicht über die Datengrundlage aus den verfügbaren Maschinendaten [3]

Clustering

Aufgrund der definierten Zielsetzung und der vorhandenen Maschinendaten werden in der Masterthesis insbesondere verschiedene Clustering-Verfahren sowie zum Teil auch Association Rule Mining für die Analyse der Daten verwendet. Es gibt eine Vielzahl von verschiedenen Clustering-Ansätzen. Eine der verbreitetsten Methoden stellt hierbei K-Means dar. Dabei handelt es sich um ein partitioning-based Clusterverfahren. Die Methodik dieser Algorithmen besteht in der schrittweisen Berechnung der Cluster-Zentroide durch das Hinzufügen weiterer Datenpunkte zu einem Cluster, bis die definierte Zielfunktion konvergiert. Eine weitere verbreitete Clustering-Methode ist das hierarchische Clustering, ein Vertreter dieses Ansatzes ist das Agglomerative Clustering. Hierbei werden die Datenpunkte, die anfangs jeweils eigenen Clustern zugeordnet sind, schrittweise mit anderen Clustern verbunden, welche die größte Ähnlichkeit aufweisen. Solche hierarchischen Clustering-Verfahren werden auch als bottom-up Verfahren bezeichnet. Weitere wichtige Verfahren sind beispielsweise das densitiy-based Clustering und grid-based Clustering. Die Wahl des Clustering-Verfahrens hängt hauptsächlich von der Verteilung der Daten im Datensatz beziehungsweise von der geometrischen Form der "gesuchten"Cluster ab. Daher erfordert die Auswahl des Clustering-Verfahrens bei großen Datensätzen zumeist umfangreiches Domänen-Wissen [2].
Um die oben genannten Clustering-Verfahren auch bei Analysen mit Mixed-Datatypes anwenden zu können, sind oftmals Anpassungen einzelner Komponenten der Algorithmen erforderlich. Insbesondere die Distanzberechnung steht hier im Mittelpunkt. Eine abgewandelte Variante von K-Means, welche sich für Mixed-Datatypes anbietet, ist K-Prototypes. Hierbei werden verschiedene Distanzfunktionen für numerische und kategorische Daten definiert. Auch andere Clustering-Methoden können meist mit angepassten Distanzmaßen genutzt werden. Eine verbreitetes Distanzmaß, welches eine Distanzmatrix für Mixed-Datatypes definiert, ist das Gower Maß. Hierbei wird ein gewichtetes Mittel für jeweils die kategorischen und numerischen Merkmale des Datensatzes bestimmt und schließlich addiert. Eine verbreitete Distanzberechnung für kategorische Daten ist die Hamming-Distanz [1].

Ansatz

Für das Erreichen der festgelegten Ziele wird die datengetriebene Mustererkennung in einzelne Teilanalysen zerlegt. Dies bedeutet, es werden im ersten Schritt Nutzungstypen und Fehlertypen der Maschinen basierend auf den jeweiligen Daten bestimmt. Daraufhin können im Anschluss die zeitlichen Sequenzen des Fehler-

und Nutzungsverhaltens der einzelnen Maschinen generiert und zusammen analysiert werden, um mögliche Muster zu identifizieren. Für die meisten dieser Analysen werden verschiedene Clustering-Verfahren eingesetzt. Die Durchführung der einzelnen Analysen setzt sich grundsätzlich aus den folgenden Schritten zusammen: Datenauswahl, Merkmalsgenerierung, Datensäuberung, Visualisierung, Auswahl und Anwendung der (Clustering-)Methode, Messung mit Bewertungsmetrik, Analyse und Interpretation der Ergebnisse [6].

Ausblick

Die Identifizierung klarer Zusammenhänge zwischen dem Nutzungsverhalten und den Fehlermustern von Maschinen in der Produktionsumgebung ermöglicht eine erweiterte detaillierte Analyse. Durch eine intensivere Analyse der individuellen Ursachen von Stillständen, basierend auf den Nutzungsdaten, könnten Erkenntnisse mit höherer Informationstiefe gewonnen werden. Auf dieser Grundlage können daraufhin Vorschläge zur Verbesserung des Nutzungsverhaltens an Kunden unterbreitet sowie Maßnahmen zur Optimierung und Vermeidung potentieller Ausfallzeiten während der Produktion abgeleitet werden.

Literatur und Abbildungen

[1] A. Ahmad and S. S. Khan. Survey of State-of-the-Art Mixed Data Clustering Algorithms. *IEEE Access*, 7:1–3, 6, 2019.

[2] Mahnoor Chaudry, Imran Shafi, Manhoor Manhoor, Debora Libertad Ramirez Vargas, Ernesto Bautista Thompson, and Imran Ashraf. A Systematic Literature Review on Identifying Patterns Using Unsupervised Clustering Algorithms: A Data Mining Perspective. *Symmetry*, pages 1–5, 10–17, 2023.

[3] Eigene Darstellung.

[4] Tak-chung Fu. A review on time series data mining. *Engineering Applications of Artificial Intelligence*, 24:167, 171, 2011.

[5] Trupti A. Kumbhare and Prof. Santosh V. Chobe. An Overview of Association Rule Mining Algorithms. *International Journal of Computer Science and Information Technologies*, 5:928, 2014.

[6] Thomas A. Runkler. *Data Mining: Methoden und Algorithmen intelligenter Datenanalyse*. Vieweg + Teubner, GWV Fachverlage GmbH, 1 edition, 2010.

[7] Muhammad Zeeshan Younas. Anomaly Detection Using Data Mining Techniques: A Review. *International Journal For Research In Applied Science & Engineering Technology*, 8:3–4, 2010.

Feature Maps zur Anomalieerkennung in Bilddaten

Feyzanur Altunkaya **Gabriele Gühring**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt an der Fakultät Informatik und Informationstechnik, Esslingen

Zielsetzung

In der vorliegenden Bachelorarbeit geht es um die Anomalieerkennung in Bilddaten mit möglichst effizienten Neuronalen Netzen. Die Erkennung einer Anomalie wird mit Hilfe eines Autoencoders verwirklicht. Üblicherweise wird ein Autoencoder auf normalen Bildern trainiert, mit dem Ziel, eine Repräsentation dieser normalen Bilder zu erstellen. Während der Inferenz wird der trainierte Autoencoder verwendet, um Anomalien zu erkennen. Dies geschieht, indem Unterschiede zwischen einem Eingangsbild und seiner rekonstruierten Version anhand bestimmter Metriken, wie der L-Distanz oder Strukturähnlichkeitsmatrix (SSIM), verglichen wird. [5] Das Ziel dieser Arbeit besteht darin, Feature Maps des sogenannten DenseNet [3] in einen Autoencoder zu geben. Der Autoencoder übernimmt diese Informationen, lernt damit und überprüft die Eingabe- und Ausgabebilder, um zu erkennen, ob eine Anomalie vorliegt.

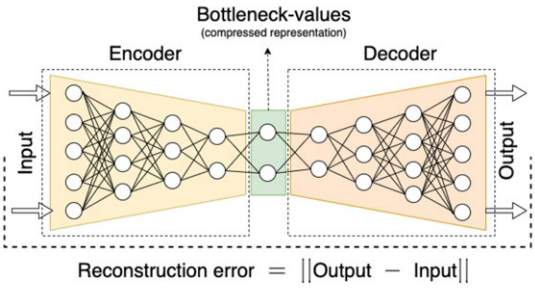


Abb. 1: Das Modell des Autoencoders [5]

Autoencoder

Ein Autoencoder erfüllt mehrere Aufgaben, darunter die Reduzierung der Datengröße, um redundante Informationen zu entfernen und wichtige Merkmale hervorzuheben. Es extrahiert Eigenschaften im sogenannten Latent Space, das als kompakte Darstellung der wichtigen Dateneigenschaften gilt. Ein Autoencoder besteht aus zwei Bereichen, dem Encoder und Decoder. Beide Bereiche übernehmen die Aufgabe, die Daten zu rekonstruieren. Dies erfolgt in zwei Schritten: Der Encoder nimmt die Eingangsdaten und wandelt sie in komprimierte Repräsentation im Latent Space um. Er besteht aus mehreren Schichten. Während der Decoder die Rekonstruktion der Eingangsdaten ermöglicht. Die Schichten des Encoders und Decoders sind für diesen Prozess von entscheidender Bedeutung, insbesondere für die Pixelklassifizierung. [5]

Convolutional Neural Networks

Künstliche Neuronale Netze können selbstständig aus unbekannten Daten Merkmale und Muster lernen. Sie können direkt mit Daten arbeiten, die zuvor nicht verarbeitet wurden und somit Klassifizierungen oder Vorhersagen durchführen. Somit kann das Neuronale Netzwerk in einer Vielzahl von Anwendungen nützlich sein. Eine spezielle und fortschrittliche Art der Künstlichen Neuronalen Netze sind die Convolutional Neural Networks, die am häufigsten in dem Bereich der Bildverarbeitung zum Einsatz kommen. Convolutional Neural Networks, bekannt als "CNN", zeichnen sich durch ihre Fähigkeit der Faltungstechniken aus. Durch die Convolutional-Schichten ist das CNN in der Lage, aus komplexen und tiefen Neuronalen Netzen die Merkmale der Daten, bekannt als "Feature Maps", zu extrahieren. [4]

DenseNet

In tieferen Convolutional Neural Networks entstehen Probleme, wie das Verschwinden des Gradienten. Informationen können bis zum Ende des Netzwerks verschwinden, da sie viele Schichten durchlaufen müssen. Ein möglicher Ansatz ist die Verwendung von DenseNet. Es ist eine fortschrittliche Netzwerkarchitektur für Neuronale Netze, die durch ihre dichte Schichtenverknüpfung hervorragt. In einem DenseNet-Modell gibt es mehrere Dense-Blöcke, jeder von ihnen besteht aus mehreren Schichten. Im Gegensatz zu herkömmlichen Netzwerken sind in DenseNet alle

Schichten miteinander verbunden. Die Eingabe für alle nachfolgenden Schichten stammt aus der Ausgabe einer Schicht. Jede Schicht erhält Informationen von allen vorherigen Schichten und überträgt ihre eigenen Informationen an alle nachfolgenden Schichten. Durch diese hohe Konnektivität können Merkmale effektiv verknüpft werden, bevor sie in eine Schicht übergeben werden. Die Eingangsvariation wird durch das Verketten von Feature Maps aus verschiedenen Schichten erhöht. Im Vergleich zu herkömmlichen CNNs mit ähnlicher Leistung reduzieren DenseNets die Gesamtparameter, aber ermöglichen durch direkte Verbindungen zwischen allen Schichten den maximalen Informationsfluss vgl. Abbildung [3].

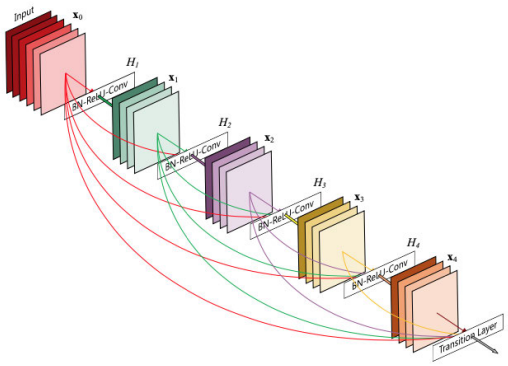


Abb. 2: DenseNet-Modell [3]

Ergebnisse

In dieser Arbeit werden die Feature Maps des DenseNet-Modells extrahiert. Diese Feature Maps werden im Autoencoder als Trainingsbilder verwendet. Um festzustellen, welche Feature Maps im gesamten DenseNet-Modell am effektivsten sind und die besten Trainingsergebnisse liefern können, wurden verschiedene Trainings aus verschiedenen Schichten des Modells durchgeführt. Die Analyse des DenseNet-Artikels [3] dient als Grundlage für die Auswahl der passenden Feature Maps. Dazu werden die durchschnittlichen Gewichtungen für jede Faltungsschicht innerhalb eines Dense-Blocks berechnet. Diese Gewichtungen deuten darauf hin, wie wichtig die Verbindungen innerhalb eines Dense-Blocks sind. Um die Ergebnisse präzise zu visualisieren gibt es verschiedene Metriken, wie AUC, F1-Score, PPV, TN/FP, TNR, TP/FN, TPR. In der ROC-Analyse (Receiver Operating Characteristic) wird der AUC-Wert (Area Under the Curve) verwendet, um die Leistung eines Klassifikationsmodells zu bewerten. Die ROC-Kurve zeigt die Differenz zwischen Sensitivität (True Positive Rate oder TPR) und False-Positiv-Rate (FPR) auf grafischer Ebene. Die Fläche unter dieser ROC-Kurve wird durch den AUC-Wert angegeben. Je höher der AUC-Wert, desto effektiver ist der Klassifikator. Wenn der AUC-Wert in der Anomalieerkennung bei 1 liegt, deutet es darauf hin, dass alle positiven Instanzen perfekt von den negativen vorhergesagt werden. Der F1-Score berücksichtigt sowohl Präzision (korrekte positive Vorhersagen) als auch Recall (alle tatsächlich positiven Instanzen, die korrekt vorhergesagt werden), und liefert eine einzige Zahl, die das ausgewogene Verhältnis zwischen Präzision und Recall ausdrückt. Ein hoher F1-Score ist für die Anomalieerkennung von entscheidender Bedeutung. Dies deutet darauf hin, dass nicht nur eine beträchtliche Anzahl der als positiv angesehenen Substanzen tatsächlich positiv sind, sondern dass alle tatsächlich positiven Substanzen richtig identifiziert werden.

In den Trainings wurden Ergebnisse mit einer Genauigkeit den AUC-Wert von etwa 0,96 und F1-Score von 97% erzielt. In der folgenden Abbildung 3 werden die Trainingsergebnisse mit dem MVTec-Datensatz für die Kategorie hazelnut veranschaulicht [1]. Weitere Trainings, die sich auf die anderen Kategorien konzentrieren, insbesondere auf solche mit schwer erkennbaren Anomalien, sind im nächsten Schritt geplant.

	AUC	F1	PPV	TN/FP	TNR	TP/FN	TPR
hazelnut	0.9679	0.9787	0.9718	[38, 2]	0.95	[69, 1]	0.9857

	AUC	F1	PPV	TN/FP	TNR	TP/FN	TPR
hazelnut	0.9607	0.9714	0.9714	[38, 2]	0.95	[68, 2]	0.9714

Abb. 3: Trainingsergebnisse [2]

Literatur und Abbildungen

[1] Paul Bergmann, Kilian Batzner, Michael Fauser, David Sattlegger, and Carsten Steger. The MVTec Anomaly Detection Dataset: A Comprehensive Real-World Dataset for Unsupervised Anomaly Detection. *Springer Link*, 2021.

[2] Eigene Darstellung.

[3] Gao Huang, Zhuang Liu, Laurens Van der Maaten, and Kilian Q Weinberger. Densely Connected Convolutional Networks. *Arxiv*, 2018.

[4] Zewen Li, Fan Liu, Wenjie Yang, Shouheng Peng, and Jun Zhou. A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects. *IEEE Xplore*, 2022.

[5] Vasilii Mosin, Miroslaw Staron, Yury Tarakanov, and Darko Durisic. Comparing autoencoder-based approaches for anomaly detection in highway driving scenario images. *Springer Link*, 2022.

Entwicklung von Proximitätsmaßen für Szenarien zum Testen hochautomatisierter Fahrzeuge

Julius Baechle **Markus Enzweiler**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma Daimler Truck AG in Zusammenarbeit mit der Torc Europe GmbH, Stuttgart Untertürkheim

Einleitung

Autonomes Fahren ist eines der spannendsten und aktuellsten Forschungsthemen der Automobilindustrie. Durch die Übergabe der Verantwortung an das hochautomatisierte Fahrzeug ist umfassendes Testen vor dem Release elementar. Distanzbasierte Ansätze sind jedoch bei so umfangreichen Systemen nicht wirtschaftlich anwendbar. Um statistisch eine dem Menschen überlegene Sicherheit nachzuweisen, müssten selbst unter günstigen Annahmen mehrere hundert Millionen Meilen gefahren werden [5]. Aus diesem Grund wurde in Projekten wie PEGASUS [6] das szenariobasierte Testen erforscht. Hierbei werden zunächst alle innerhalb des Einsatzgebiets auftretenden Szenarien identifiziert, unter anderem indem Aufnahmen von Testfahrten, Drohnen und Dashcams im realen Verkehr ausgewertet werden. Die abgeleiteten Szenarien können anschließend gezielt in Simulationen oder auf dem Testgelände ausgeführt werden.

Aufgabenstellung

Im Rahmen dieser Arbeit werden Proximitätsmaße zum Messen von Distanzen bzw. Ähnlichkeiten zwischen aufgezeichneten Szenarien entwickelt. Darin sind insbesondere die während der Ausführung von Szenarien protokollierten Trajektorien der Verkehrsteilnehmer sowie daraus ableitbare Größen wie Geschwindigkeit und Beschleunigung enthalten. Für solche Maße gibt es viele potenzielle Einsatzmöglichkeiten, wie Clustern der Szenarien, Anomalieerkennung und Prüfen der Testabdeckung. Diese Arbeit untersucht in Simulationen aufgetretene Kollisionen. Das beeinflusst aber insbesondere die Wahl des Datensatzes, nicht die Ergebnisse und Schlussfolgerungen der Arbeit. Die genannten Aufgaben fallen in den Bereich des unüberwachten Lernens für multivariate Zeitserien. Es existiert keine Grundwahrheit, was die Evaluierung der entwickelten Proximitätsmaße zur Herausforderung macht.

Aktueller Forschungsstand

Aufgrund der vielfältigen Einsatzmöglichkeiten können in der Literatur viele entwickelte oder implizit genutzte Proximitätsmaße gefunden werden. In einer Übersicht von Braun et al. [1] werden Proximitätsmaße in abstraktions-, zeitserien- und KPI-basierte Maße eingeteilt. Abstraktionsbasierte Maße diskretisieren ursprünglich kontinuierliche Werte wie Position, Geschwindigkeit und Beschleunigung. Ein Vertreter dieser Kategorie ist StreetWise [3], das zur Identifikation vordefinierter Kategorien in Fahrdaten eingesetzt wird. Key-Performance-Indicators (KPIs) sind skalare Werte wie beispielsweise die durchschnittliche Geschwindigkeit des getesteten Fahrzeugs oder die Information, ob eine Vollbremsung eingeleitet wurde. Zeitserienbasierte Maße arbeiten direkt mit Distanzmaßen wie Dynamic Time Warping (DTW) auf den protokollierten Zeitserien, sind jedoch rechenaufwendig. Über die Übersicht von Braun et al. [1] hinaus gibt es Ansätze, die Autoencoder zur Komprimierung einsetzen und im Latent-Space clustern [8], die im Folgenden als kompressionsbasiert bezeichnet werden.

Abstraktionsbasierte Kategorisierung

Zur exemplarischen Implementierung abstraktionsbasierter Verfahren wie StreetWise [3], werden Manöver wie Spurwechsel und die relative Geschwindigkeit extrahiert. Eine Möglichkeit zur Unterteilung wird in Abb. 1 dargestellt. Zunächst werden Heck- und Seitenkollisionen mit schnelleren Kollisionspartnern sowie Unfälle mit stillstehenden Fahrzeugen identifiziert. Anschließend wird anhand der durchgeführten Spurwechsel von Ego und Aktor weiter unterteilt, wobei bei Spurwechseln des Aktors zwischen hoher und niedriger Kollisionsgeschwindigkeit unterschieden wird. Bei gleichzeitigen Spurwechseln des Aktors und Egos wird zwischen Spurwechseln in dieselbe Richtung und in entgegengesetzte Richtungen unterschieden.

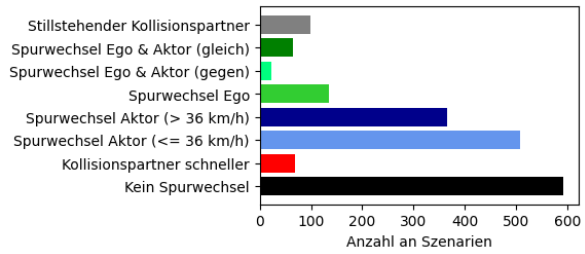


Abb. 1: Szenariokategorien [2]

Die Aufteilung ist subjektiv. Es könnten weitere oder andere Eigenschaften zur Unterteilung genutzt werden, beispielsweise die relative Geschwindigkeit zum Zeitpunkt des Spurwechsels. Zudem wird die Distanz zwischen den Szenarien nicht intuitiv wiedergegeben: Die Distanz zwischen Cut-Ins mit 36 km/h und 37 km/h entspricht der Distanz zwischen einem Cut-In und einem Szenario ganz ohne Spurwechsel.
Um feiner zwischen Szenarien zu unterscheiden und die Kategorisierung zu überprüfen, müssen detailliertere Proximitätsmaße eingesetzt werden. Im Folgenden werden hierzu kompressionsbasierte Maße vorgestellt, auf KPI- und zeitserienbasierte Maße kann hier aus Platzgründen nicht näher eingegangen werden.

Kompressionsbasierte Proximitätsmaße

Zum Erkennen von Anomalien von Zeitserien können Autoencoder wie TCN-AE [7] eingesetzt werden, der in dieser Arbeit näher untersucht wird. Es wird angenommen, dass zur Repräsentation im Latent-Space aussagekräftige Features gewählt werden, aus denen sich die Eingabedaten gut rekonstruieren lassen. Des Weiteren wird davon ausgegangen, dass sich ein Autoencoder auf die Rekonstruktion häufig vorkommender Szenarien konzentriert, wodurch seltene Ausreißer schlechter rekonstruiert werden [7]. Im Vergleich dazu vergessen Recurrent Neural Networks (RNNs) lange zurückliegende Werte und sind schwerer zu trainieren. RNNs wie Variational Recurrent Autoencoder (VRAE) [4] sind dafür ohne weiteres auf unterschiedlich lange Zeitserien anwendbar.
Zum Vergleich mit den ermittelten Kategorien wird der Autoencoder auf je drei Zeitserien trainiert: der lateralen Bewegung von Ego und Aktor sowie der relativen Geschwindigkeit. Anschließend werden die Szenarien in den Latent-Space komprimiert, welcher zur Darstellung mittels Principal Component Analysis (PCA) in einen zweidimensionalen Raum transformiert wird. Abb. 2 zeigt die Kategorien mit farblicher Kodierung gemäß Abb. 1. Obwohl die Kategorien im Bild nicht scharf abgegrenzt sind, ist eine klare Tendenz erkennbar. Besonders Spurwechsel des Aktors mit hoher oder niedriger Kollisionsgeschwindigkeit werden nicht klar separiert.

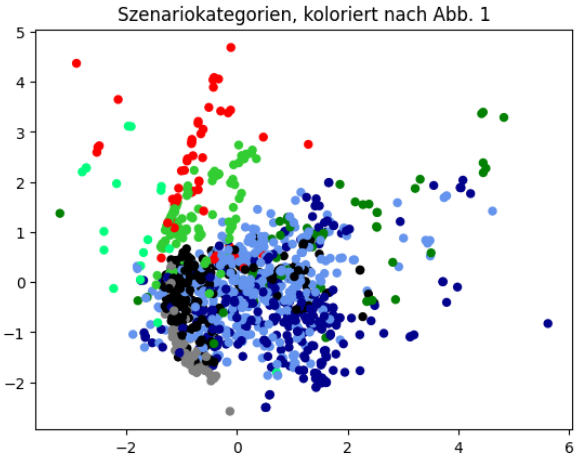


Abb. 2: Kategorien im Latent-Space [2]

Es ist zu bedenken, dass die Darstellung des in diesem Fall 10-dimensionalen Latent-Space in zwei Dimensionen zu Verzerrungen führt. Um den visuellen Eindruck zu quantifizieren, können ursprünglich interne Clustering-Metriken mit diesen vordefinierten Kategorien anstelle der Prädiktion eingesetzt werden. Es wird dann bewertet, wie gut das Proximitätsmaß das vorgegebene Clustering trennt. Im Latent-Space können außerdem falsch kategorisierte Szenarien gefunden werden, indem die Kategorie eines Szenarios mit der Umgebung abgeglichen wird.
Zur Erkennung von Anomalien wird zunächst der Rekonstruktionsfehler ermittelt, um die Szenarien mit dem größten Fehler auszuwählen. Es werden vor allem Szenarien gefunden, in denen der Kollisionspartner auf ein stillstehendes Fahrzeug auffährt. Ursache sind die Flanken in der relativen Geschwindigkeit, die vom Autoencoder schlecht rekonstruiert werden. Verwendet man lediglich die laterale Bewegung von Aktor und Ego, werden Szenarien als Anomalien erkannt, in denen das Ego zwei Spuren gleichzeitig wechselt.
Um näher zu untersuchen, was das kompressionsbasierte Proximitätsmaß abbildet, kann ein Feature wie die Kollisionsgeschwindigkeit visualisiert werden. Auch in Abb. 3 ist eine Tendenz erkennbar, die über die Korrelation zweier Distanzmaße quantifiziert werden kann. Als erstes Distanzmaß wird die euklidische Distanz im Latent-Space verwendet, als zweites die Differenz der Kollisionsgeschwindigkeit. In diesem Fall wird ein Korrelationskoeffizient von 0.26 erreicht. Eine vollständige Korrelation von 1 wäre nicht erstrebenswert, da das gelernte Proximitätsmaß dann ausschließlich die Kollisionsgeschwindigkeit wiedergibt. Andererseits wäre auch eine nicht vorhandene oder gegensätzliche Korrelation von 0 bis -1 unerwartet, da die Kollisionsgeschwindigkeit in den Eingabedaten enthalten ist.

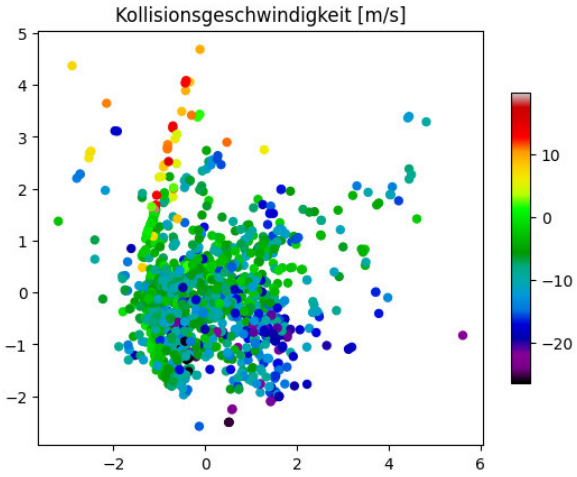


Abb. 3: Kollisionsgeschwindigkeit im Latent-Space [2]

Ergebnis und Ausblick

Da keine unumstößliche Grundwahrheit existiert, ist die Bewertung eines Proximitätsmaßes nur im Zusammenhang mit dem jeweiligen Einsatzzweck sowie relativ zu anderen Maßen möglich. Dabei kann der Abgleich verschiedener Maße zur gegenseitigen Verbesserung und Validierung genutzt werden. Einerseits ist eine Separierung abstraktionsbasiert ermittelter Kategorien durch detailliertere Maße wünschenswert. Andererseits werden falsch kategorisierte Szenarien erst durch detaillierte Proximitätsmaße sichtbar.
Zum Clustern sind abstraktionsbasierte Maße bei Vorwissen über das zu testende System zielführend, denn die Grenzen sind klar nachvollziehbar und einfach interpretierbar. Zur Anomalieerkennung und Auswahl repräsentativer Szenarien sind jedoch detailliertere Proximitätsmaße besser geeignet. Bei großen Datensätzen sollten kompressionsbasierte Maße wegen des hohen Rechenaufwands zeitserienbasierten Proximitätsmaßen vorgezogen werden.
Zur Optimierung des vorgestellten kompressionsbasierten Maßes könnte ein Local Aggregation Loss [9] eingesetzt werden, der zu einer Cluster-Bildung im Latent-Space führt. Darüber hinaus wäre eine Kombination abstraktionsbasierter Maße für näherungsweise diskrete Eigenschaften wie Spurwechsel und detaillierter Maße für kontinuierliche Werte wie die Kollisionsgeschwindigkeit interessant.

Literatur und Abbildungen

[1] Thilo Braun, Julian Fuchs, Felix Reisgys, Lennart Ries, Johannes Plaum, and Eric Sax. A Review of Scenario Similarity Measures for Validation of Highly Automated Driving. *2023 IEEE Intelligent Transportation Systems Conference (ITSC)*, 2023.

[2] Eigene Darstellung.

[3] Hala Elrofai, Jan-Pieter Paardekooper, Erwin de Gelder, Sytze Kalivaart, and Olaf op den Camp. *StreetWise: Scenario-Based Safety Validation of Connected and Automated Driving*. Netherlands Organization for Applied Scientific Research, TNO, Tech. Rep, 2018.

[4] Otto Fabius and Joost R Van Amersfoort. Variational recurrent auto-encoders. *arXiv preprint arXiv:1412.6581*, 2014.

[5] Nidhi Kalra and Susan M Paddock. Driving to safety: How many miles of driving would it take to demonstrate autonomous vehicle reliability? *Transportation Research Part A: Policy and Practice*, pages 182–193, 2016.

[6] Project PEGASUS. Pegasus Method - An Overview. https://www.pegasusprojekt.de/files/tmpl/Pegasus-Abschlussveranstaltung/PEGASUS-Gesamtmethode.pdf, 2019.

[7] Markus Thill, Wolfgang Konen, and Thomas Bäck. Time series encodings with temporal convolutional networks. *International Conference on Bioinspired Methods and Their Applications*, 2020.

[8] Jinxin Zhao, Jin Fang, Zhixian Ye, and Liangjun Zhang. Large scale autonomous driving scenarios clustering with self-supervised feature extraction. *2021 IEEE Intelligent Vehicles Symposium (IV)*, 2021.

[9] Chengxu Zhuang, Alex Lin Zhai, and Daniel Yamins. Local aggregation for unsupervised learning of visual embeddings. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019.

Entwicklung und Evaluierung von Bildverarbeitungsalgorithmen für Eventkameras

Ahmet Balli **Markus Enzweiler**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma Leuze electronic Deutschland GmbH + Co KG, Owen

Einleitung

Machine Vision, was einst ein futuristisches Konzept war, ist heute ein fester Bestandteil in unserem Alltag. Die Anwendung dieser Technologie ist vielfältig und revolutioniert stetig die Welt, sei es im Bereich der Automobilindustrie oder bei der medizinischen Bildgebung. Im Bereich der Bildsensoren, insbesondere bei den Eventkameras, die eine neue Generation von Bildsensoren eröffnen sich spannende Möglichkeiten. Die Bachelorarbeit befasst sich mit der Herausforderung der Objektdetektion von Behältern mithilfe von Eventkameras. Das Anwendungsszenario umfasst laufende Behälter auf einem Förderband, bei dem die vorderste Kante des Behälters erkannt werden soll. Dabei wird auf die Programmierung von Algorithmen zur Detektion von Behältern mithilfe von Events eingegangen, wobei ein signifikanter Fokus auf die Flexibilität und Adaptivität der Algorithmen gelegt wird. Des Weiteren widmet sich die Arbeit der Untersuchung verschiedener Umgebungseinflüsse, wie Beleuchtungsbedingungen und Hintergrundbewegungen, auf die Detektionsgenauigkeit, um robuste Anwendungen in unterschiedlichen Einsatzgebieten sicherzustellen.

Eventkamera

Eventkameras markieren eine innovative Entwicklung in der Bildverarbeitung, indem sie ausschließlich Änderungen in der Lichtintensität pro Pixel erfassen. Anders als zu herkömmlichen Kameras, die in festen Intervallen komplette Bilder aufnehmen, reagieren diese auf dynamische visuelle Informationen mit einem ununterbrochenen Event Stream. Die Inspiration der Eventkameras ist in der Funktionsweise des menschlichen Auges verwurzelt. Um die Welt um uns herum wahrzunehmen, muss Licht von Objekten reflektiert und in unser Auge eingefangen werden. Dieses reflektierte Licht durchquert zuerst die Bindehaut und Hornhaut, bevor es auf die Netzhaut trifft, wo Stäbchen- und Zapfenzellen spezialisiert sind, um Aspekte wie Helligkeit, Schärfe und Farbe zu detektieren. Die von diesen Zellen aufgenommenen Informationen werden über den Sehnerv an das Gehirn weitergeleitet, welches letztlich das Bild interpretiert und konstruiert. Die Architektur eines Event-based Vision Sensor (EVS) ahmt diese biologischen Prozesse wieder. Die Lichtempfangseinheit des Sensors nimmt das einfallende Licht und wandelt dieses in ein elektrisches Signal um. Mittels Delta-Modulation wird dann kontinuierlich die Differenz zwischen einer Referenzspannung und der Spannung des umgewandelten Lichtsignals überwacht. Sobald eine Veränderung dieser Spannung einen vordefinierten Schwellenwert in positiver oder negativer Richtung überschreitet, erfasst der Komparator dies als ein Event.

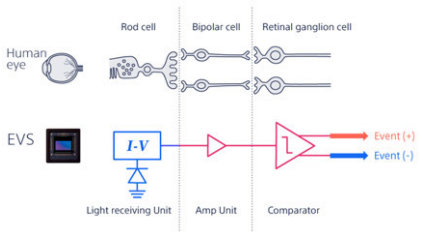


Abb. 1: Vergleich Netzhaut und EVS [4]

Bias

Die Biases bieten eine wesentliche Anpassungsmöglichkeit für Eventkameras, um sie an spezifische Anforderung je nach Anforderungsprofil anzupassen, sei es für höhere Erfassungsgeschwindigkeiten, reduzierte Hintergrundaktivität oder verbesserte Kontrastsensitivität. In Abbildung 2 markiert die x-Achse den Zeitverlauf, während die y-Achse die Lichtintensität angibt. Die kontinuierliche schwarze Linie bildet die von einem Sensor aufgezeichnete Veränderung der Lichtintensität. Ein Event wird ausgelöst, sobald die Intensität definierte Schwellen über- oder unterschreitet, was durch die blauen und roten Pfeile visualisiert

wird. Dies verdeutlicht, dass das Event-based Vision Sensoren (EVS) eine selektive Datenaufzeichnung praktiziert. Es registriert nur wesentliche Änderungen in der Lichtintensität, wodurch es, zu einer erheblichen Reduktion der Datenmenge führt und gleichzeitig eine schnelle Reaktionsfähigkeit sicherstellt.

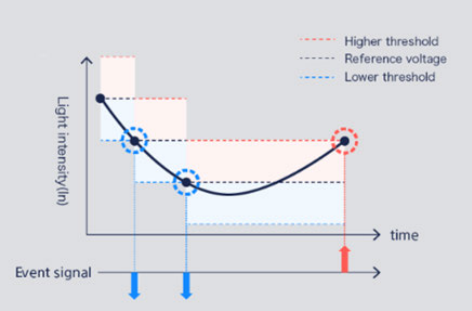


Abb. 2: Erfassung von Lichtintensitätsänderungen [2]

Event- vs. Frame-basierte Sensoren

Herkömmliche Kameras nehmen Bilder mit einer festen Rate auf, während eventbasierte Kameras nur Informationen über Veränderungen der Lichtintensität erfassen. Dies ermöglicht weniger unwesentliche Daten und einen größeren Dynamikbereich. Eventkameras zeichnen sich durch eine hohe zeitliche Auflösung in der Größenordnung von Mikrosekunden aus, was bedeutet, dass sie Veränderungen im Bildgeschehen sehr präzise erkannt werden. Des Weiteren besitzen sie einen sehr hohen Dynamikbereich von 140 db im Vergleich zu herkömmlichen Kameras mit nur 60 db. Dies ermöglicht, sowohl in sehr hellen als auch von sehr dunklen Bildbereiche feine Details zu erkennen. Darüber hinaus bieten sie eine hohe Pixelbandbreite in der Größenordnung von kHz, was die Bewegungsunschärfe reduziert und schnelle Bewegungen genau erfasst werden können. Abbildung 3 zeigt den Unterschied zwischen Event-basierten und Frame-basierten Sensoren bei der Aufzeichnung einer Objektbewegung. Im linken Diagramm wird die zweidimensionale Darstellung der Flugbahn eines Balls gezeigt, während das mittlere Diagramm die Erfassung dieser Flugbahn durch einen Event-basierten Sensors darstellt. Der Event-basierte Sensor zeichnet nur die Veränderungen in der Lichtintensität auf, die durch die Bewegung des Balls verursacht werden, was zu einer kontinuierlichen und zeitlich exakten Erfassung der Flugbahn ermöglicht. Im Gegensatz dazu illustriert das rechte Diagramm die Erfassung derselben Bewegung durch einen Frame-basierten Sensor. Die Frame-basierte Sensortechnik zeichnet das gesamte Bild in festen Zeitabständen auf, wodurch es zu Informationslücken zwischen den einzelnen Frames führt.
[1]

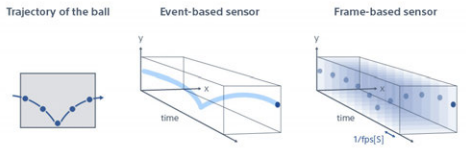


Abb. 3: Event-based vs Frame-based Sensor [3]

Ausblick

Die Zukunft der Machine Vision, insbesondere durch die Einführung von Eventkameras, versprechen eine Ära der Innovation. Diese Technologie, spezialisiert auf den Veränderungen in Lichtverhältnissen zu detektieren und in verwertbare Daten umzuwandeln, steht potenziell am Anfang einer Revolution in Bereichen wie autonomes Fahren, Robotik und interaktive Medienwelt. Weiterführende Forschung und Entwicklung könnten darauf abzielen, die Adaptionsfähigkeit und Lernprozesse dieser Systeme zu verfeinern, sodass sie noch genauer auf ihre Umgebung reagieren und komplexe Muster in Echtzeit interpretieren können.

Literatur und Abbildungen

[1] Guillermo Gallego, Tobi Delbruck, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew J. Davison, Jorg Conradt, Kostas Daniilidis, and Davide Scaramuzza. Event-based Vision: A Survey. https://arxiv.org/pdf/1904.08405.pdf, 2019.

[2] Prophesee S A. Biases. https://docs.prophesee.ai/stable/hw/manuals/biases.html?highlight=bias, 2023.

[3] Sony Semiconductor Solutions Corporation. Event-based Vision Sensor (EVS). https://www.sony-semicon.com/en/products/is/industry/evs.html, 2021.

[4] Sony Semiconductor Solutions Corporation. Event-based Vision Sensor EVS Technology. https://www.sony-semicon.com/en/technology/industry/evs.html, 2023.

Kinematic Ranging for Moving Targets with Limited Motion Capabilities

Andreas Baulig **Clemens Klöck**

Department of Computer Science and Engineering, Esslingen University

Work carried out at Hensoldt Sensors GmbH, Ulm

Introduction

The term *Kinematic Ranging* refers to a subclass of tracking problems commonly referred to as target motion analysis (TMA). Goal of TMA is to track position and velocity (sometimes also acceleration and more) of a radiating target based on noisy measurements by a singular *passive* sensor platform [2].

The problem TMA aims to address arises in various civilian and military applications. Historically, naval and submarine tracking applications based on acoustic emissions pioneered this area of research, namely passive SONAR. This is reflected in a lot of early literature [2] [4] [7]. Today, TMA is broadly applied to many other domains such as collision avoidance, air traffic management, navigation, electronic warfare, and air surveillance [3].

This thesis focuses on *angle-only* observations by a single moving platform (referred to as *ownship*). That is, the system is only capable of differentiating detections by their angle of arrival in azimuth and elevation. The objective is to localize and track the motions of a similarly moving target platform. By nature of the *angle-only* observation model, target range is not immediately known. Thus, to produce an unambiguous position fix, normally the ownship's maneuverability has to *make up* for the unobservable range parameter. This gives rise to the term *Kinematic Ranging*.

Problem Formulation

Core objective of this thesis is to design and analyze a recursive state estimator that is able to track the three-dimensional position and velocity of a non-accelerating target in adverse observability conditions (i.e. no ownship maneuver).

A key assumption taken throughout the discussion is that both, ownship and target, adhere to a constant-velocity motion model. In other words, the kinematic equations of both platforms are linear with respect to time. Further, as to allow focus being put on the tracking problem, perfect knowledge of the association of measurements to individual targets is assumed.

Refer to 1 for an overview of the assumed TMA geometry. Cartesian ownship position $\boldsymbol{p}_{\text{own}}$ and velocity $\boldsymbol{v}_{\text{own}}$ are assumed known. The goal is to recursively estimate the Cartesian target position $\boldsymbol{p}_{\text{tar}}$ and velocity $\boldsymbol{p}_{\text{tar}}$ from a sequence of angle measurements in azimuth β and elevation ϵ corrupted by additive white Gaussian noise.

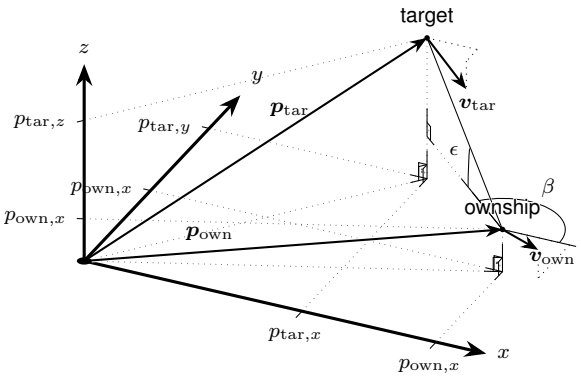


Fig. 1: Angle-only target motion analysis geometry. [6]

The measurement equations (without noise) for a non-maneuvering target are simply

$$\beta(t) = \arctan2(p_y(t), p_x(t)) \tag{1}$$

and

$$\epsilon(t) = \arctan2(p_z(t), r_{xy}(t)) \tag{2}$$

with

$$p_x(t) = p_{\text{tar},x} - p_{\text{own},x} + t(v_{\text{tar},x} - v_{\text{own},x}),$$

$$p_y(t) = p_{\text{tar},y} - p_{\text{own},y} + t(v_{\text{tar},y} - v_{\text{own},y}),$$

$$p_z(t) = p_{\mathrm{tar},z} - p_{\mathrm{own},z} + t(v_{\mathrm{tar},z} - v_{\mathrm{own},z}),$$

$$r_{xy}(t) = \sqrt{(p_x(t))^2 + (p_y(t))^2},$$

where $\arctan 2$ is the four quadrant inverse tangent. It is well known that without ownship maneuver or relevant *a priori* information, range to target r remains unobservable [5]. This is easily seen by introducing an arbitrary scaling term $\alpha > 0$ and multiplying both $\arctan 2$ arguments in 1 and 2. This is illustrated in 2. A single straight-line observer trajectory producing an associated sequence of angle measurements leads to ambiguous range and velocity terms [1]. The illustration shows selected examples of ambiguous trajectories, each fulfilling the measured angles and the constant-velocity constraint. In fact, every point along the bearing lines can be associated with a singular valid constant-velocity motion.

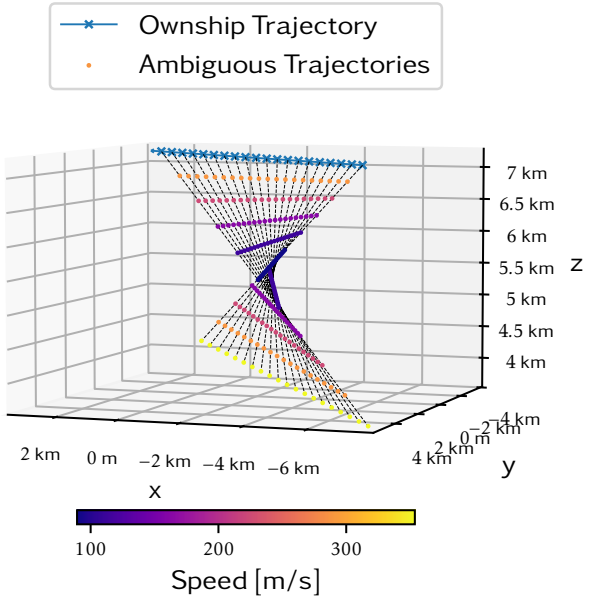


Fig. 2: Ambiguous target trajectories resulting from unobservable range. [6]

The motion capabilities of the target are constrained in that target speed is bounded by

$$v_{\mathrm{tar,min}} < \|\boldsymbol{v}_{\mathrm{tar}}\| < v_{\mathrm{tar,max}}.$$

Both, minimum target speed $v_{\mathrm{tar,min}}$ and maximum target speed $v_{\mathrm{tar,max}}$ are known *a priori*. A recursive state estimator is to be designed that incorporates these bounds in order to limit the target state search-space.

Estimation in Adverse Observability Conditions

Significant work on TMA under adverse observability conditions was conducted by [1]. In their paper, the authors present a multi-modal batch-style maximum a posteriori (MAP) estimator taking into account *a priori* probability distributions for range and velocity. In a similar vein, this thesis attempts to design and analyze a recursive Kalman-style estimator given *a priori* information on the velocity probability distribution. An innate advantage of Kalman-style estimators is their ability to react to deviations in the process model.
The introduction of statistical or bounded *a priori* information does not generally recover observability. Therefore, an important criterion for any estimator operating in these conditions is *consistency*. Rather than force the pinpointing of a precise (but inaccurate) position and velocity fix, an estimator must provide accurate confidence intervals that converge on the set of potential target states.

References and figures

[1] Frederic Bavencoff, Jean-michel Vanpeperstraete, and J.-pierre Le Cadre. Constrained bearings-only target motion analysis via Markov chain Monte Carlo methods. *IEEE Transactions on Aerospace and Electronic Systems*, 42:1240–1263, 2006.

[2] Klaus Becker. Target Motion Analysis (TMA). In *Advanced Signal Processing Handbook*. Stergiopoulos, Stergios, 2001.

[3] Dietrich Fränken and Andreas Hüpper. Unified tracking and fusion for airborne collision avoidance using log-polar coordinates. *2012 15th International Conference on Information Fusion*, pages 1246–1253, 2012.

[4] S.C. Nardone and M.L. Graham. A closed-form solution to bearings-only target motion analysis. *IEEE Journal of Oceanic Engineering*, 22:168–178, 1997.

[5] Steven C. Nardone and Vincent J. Aidala. Observability Criteria for Bearings-Only Target Motion Analysis. *IEEE Transactions on Aerospace and Electronic Systems*, pages 162–166, 1981.

[6] Own representation.

[7] F. N. Spiess. Complete Solution of the Bearings Only Approach Problem. *MPL Technical Memorandum*, 102, 1953.

Explainable AI in der Niederspannungsprognose - Eine Analyse von einem ML Modell zur Vorhersage von Pseudomessdaten

Michael Baur **Steffen Schober**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma Netze BW GmbH, Stuttgart

Motivation

Durch den Hochlauf von Teilnehmern in den unüberwachten Niederspannungsnetzen, ausgelöst durch die Energiewende, finden gerade im Verteilnetz Maschinelle Lernverfahren (ML) ihre Anwendung, da diese im Vergleich zu Messystemen eine kostengünstigere und schnellere Alternative darstellen. Da es sich bei diesen Modellen um Black-Box Modelle handelt, ist oftmals nicht eindeutig, wie oder warum das Modell bestimmte Messwerte prognostiziert und wie das Modell mit realen Messwerten umgeht. Daher kann es für Experten in der Anwendung schwierig sein, den Entscheidungen und Empfehlungen von ML Algorithmen zu vertrauen, was deren praktischen Nutzen einschränkt [4]. Dies ist besonders in den Bereichen der Fall, in denen ein hohes Maß an Zuverlässigkeit erforderlich ist, wie es in der Energiebranche üblich ist [4]. An dieser Stelle tragen Explainable Artificial Intelligence (XAI) Verfahren dazu bei, die Erklärbarkeit von Modellen zu verbessern, wodurch ihre Ergebnisse besser verstanden werden können [4].

Ziel der Arbeit

Im Rahmen der Masterarbeit soll das Black-Box Modell der Niederspannungsprognose, zur Erzeugung von Pseudomesswerten, anhand von Explainable AI Methoden analysiert werden, mit dem Ziel, dieses kontinuierlich zu verbessern und Vertrauen gegenüber den Entwicklern zu etablieren. Dabei sollen mehrere Methoden und Verfahren dazu beitragen, die Entscheidungsmechanismen des Modells zu verstehen, dadurch die Modellgenauigkeit zu erhöhen und aus dem generierten Wissen Erkenntnisse abzuleiten.

Grundlagen

Im Zuge der Energiewende steht das deutsche Energiesystem vor massiven Veränderungen. Vor allem in der Niederspannung gibt es im Bereich des Verkehr- und Wärmesektor einen Markthochlauf von Elektroautos und Wärmepumpen. Aber nicht nur auf der Verbraucherseite entsteht ein Wandel, sondern auch auf Seiten der Erzeugung wird ein Großteil der Solaranlagen mehrheitlich in der Niederspannung angeschlossen. Diese Veränderungen auf der Verbrauchs- und Erzeugungsseite treffen auf ein Verteilnetz, welches nicht auf diese Anforderungen ausgelegt wurde. [6]

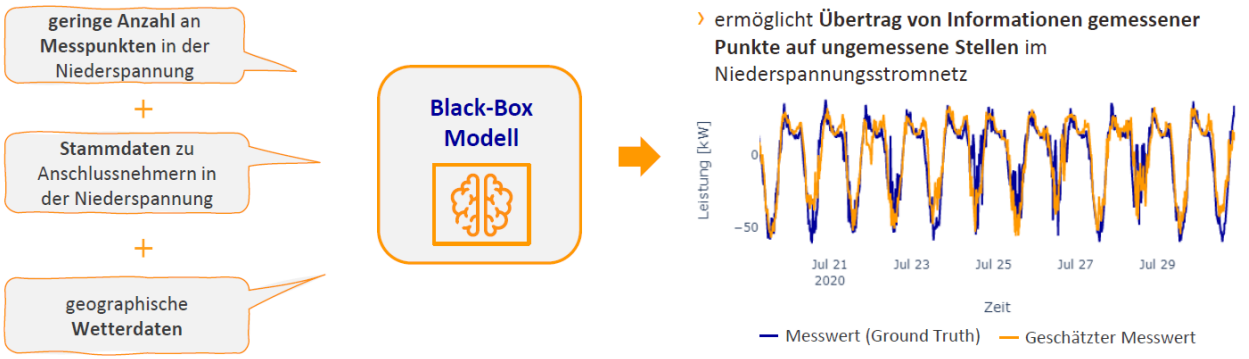


Abb. 1: Erzeugung von Pseudomesswerte durch das Black-Box Modell der Niederspannungsprognose [1]

Infolgedessen hat die Netze BW GmbH einen geringen Teil der Niederspannungsabgänge in Ortsnetzstationen mit Messtechnik ausgestattet. Auf Basis dieser Messwerte, geographischer Wetterdaten und den Stammdaten zu den Anschlussnehmern in der Niederspannung, ist wie in Abbildung 1 gezeigt, ein Black-Box Modell für die Vorhersage von Pseudomesswerten ungemessener Stromkreise von Ortsnetzstationen entstanden. Anhand dieser Pseudomesswerte sollen aktuell speziell drei verschiedene Use Cases bedient werden. Diese drei Use Cases betreffen die Netzplanung in der Niederspannung, die Zustandsschätzung in der Mittelspannung und die Steuerung von Teilnehmern in der Niederspannung. Zur Erfüllung der Use Cases, Steigerung der Modellgenauigkeit und für ein Verständnis, wie und warum das Modell entscheidet, ist die Erklärbarkeit dieser Black-Box Modelle entscheidend. Aufgrund der Black-Box Eigenschaft von ML Modellen wurden in den letzten Jahren neue Techniken und Prinzipien entwickelt, um die Erklärbarkeit von Modellen zu verbessern, wodurch ihre Ergebnisse besser verstanden werden können [4]. Dieses Konzept ist in der Literatur als Explainable Artificial Intelligence bekannt [4]. Das Ziel von XAI besteht darin, Forschern, Entwicklern, Domänenexperten und Benutzern dabei zu helfen, die innere Funktionsweise von Modellen für maschinelles Lernen besser zu verstehen und gleichzeitig ihre hohe Leistung und Genauigkeit zu bewahren [4]. In diesem Zusammenhang ist speziell die Post-hoc Erklärbarkeit relevant, da das Black-Box Modell nach dem Training interpretiert werden kann, ohne die Modellperformance zu beinträchtigen [2]. Die Post-hoc Erklärbarkeit setzt sich aus den „Model-Agnostic“ und „Model-Specific“ Methoden zusammen. Model-Agnostic Methoden können für jegliche Arten von Modellen angewendet werden, während Model-Specific Methoden nur für spezielle Modelle Anwendung finden.

Umsetzung

Da die Explainable Artificial Intelligence ein breites Spektrum abdeckt, orientiert sich die Masterarbeit Grundlegend an folgenden Forschungsfragen:

- Unterscheiden sich die Gleichzeitigkeiten der Betriebsmittel aus dem Black-Box Modell von denen aus den Planungsgrundsätzen?
- Sind die Modellgrenzen identifizierbar und bei welchen Features weist das Modell Lücken in der Prognose auf?
- Bestätigt das Modell bereits bekannte Zusammenhänge zwischen den Wetterdaten und Leistungswerten für Stromkreise mit bestimmten Stammdatenkonfigurationen und können neue Verhaltensmuster identifiziert werden?

Diese Forschungsfragen tragen dazu bei, die Datenqualität und Modellgenauigkeit zu erhöhen, Wissen gegenüber Betriebsmittelinteraktionen und Kundenverhalten zu generieren und dadurch das Vertrauen gegenüber den Entwicklern zu erhöhen. Zur Erreichung der Forschungsfragen sollen verschiedene Techniken aus dem Gebiet der „Feature Relevance“ und „Visual Explanation“ angewendet werden, welche in einer Übersicht in Abbildung 2 gezeigt sind. In erster Linie sollen dabei LIME [5], SHAP [3] und Dependence Plots auf das Black-Box Modell angewendet werden. Erste Ergebnisse zeigen, dass sowohl die Anzahl der Wohnungen der Anschlussnehmer, als auch die PV-Anlagen in Kombination mit der Sonneneinstrahlung einen großen Effekt auf die Prognose des Modells haben.

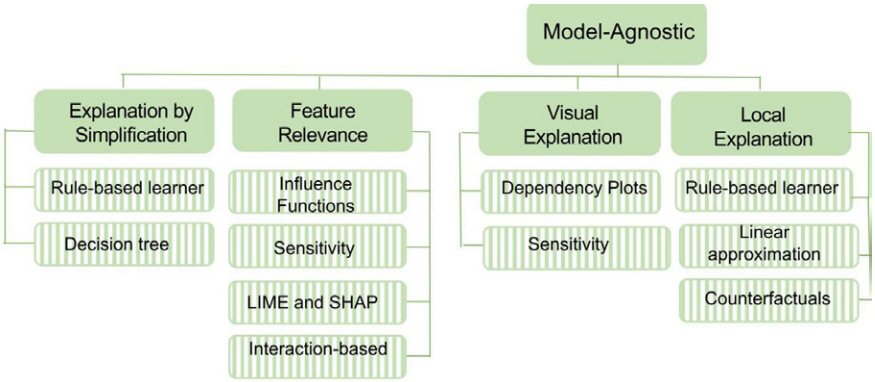


Abb. 2: Model-Agnostic Erklärbarkeit [2]

Literatur und Abbildungen

[1] Eigene Darstellung.

[2] Peter E.D. Love, Weili Fang, Jane Matthews, Stuart Porter, Hanbin Luo, and Lieyun Ding. Explainable Artificial Intelligence (XAI): Precepts, Methods, and Opportunities for Research in Construction. *arXiv*, 2023.

[3] Scott M. Lundberg and Su-In Lee. A Unified Approach to Interpreting Model Predictions. *arXiv*, 2017.

[4] R. Machlev, L. Heistrene, M. Perl, K.Y. Levy, J. Belikov, S. Mannor, and Y. Levron. Explainable Artificial Intelligence (XAI) techniques for energy and power systems: Review, challenges and opportunities. *ScienceDirect*, 2022.

[5] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?" Explaining the Predictions of Any Classifier. *arXiv*, 2016.

[6] Manuel Treutlein. Erzeugung von Pseudomessdaten für Ortsnetzstationen mittels diskriminativer und generativer Modelle, 2023.

Entwicklung einer Testautomatisierung zur Absicherung des Kommunikationssteuergeräts für Fahrerassistenzfunktionen in Versuchsfahrzeugen

Emre Bayram **Markus Enzweiler**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma Robert Bosch GmbH, Leonberg

Einleitung

Fahrerassistenzsysteme haben in den letzten Jahren eine enorme technische Weiterentwicklung erfahren und sind zu einem integralen Bestandteil moderner Fahrzeuge geworden. Um diese Systeme weiterzuentwickeln, werden Entwicklungsfahrzeuge aufgebaut, die die Anwendung neuer Technologien ermöglichen. Diese Entwicklungsfahrzeuge sind mit prototypischen Steuergeräten ausgestattet, die aktiv in das Fahrverhalten eingreifen können. Um die Sicherheit des Testfahrers und der anderen Verkehrsteilnehmer in solchen Fahrzeugen zu gewährleisten, werden umfassende Tests der Software durchgeführt. Der Fokus dieser Bachelorthesis liegt auf der Entwicklung und Integration einer Testautomatisierungsstrategie für das Gateway-Steuergerät, welches für die Steuerung verschiedenster Fahrerassistenzfunktionen in den Entwicklungsfahrzeugen verantwortlich ist.

Motivation

Die fortlaufende Optimierung der entwickelten Funktionen sowie die Integration zusätzlicher Funktionalitäten stellen für das Gateway, welches als zentraler Kommunikationsknoten fungiert, stetig wachsende Anforderungen dar. Um die Funktionalität und Sicherheit des Gateways zu überprüfen und eine Beurteilung über die Softwarereife für den Entwicklungsstand zu ermöglichen, muss diese bei neuer Software getestet werden. Diese Testung ist unabdinglich, da es sich um einen Entwicklungsstand handelt. Es können jederzeit Eingriffe durch automatisiertes Bremsen, Spurhaltefunktionen und Beschleunigung entstehen. Dies kann zu einer Gefährdung für den Fahrer und andere Verkehrsteilnehmer führen. Die fortlaufenden Tests dieses Systems erfordern zeitliche und finanzielle Ressourcen, um bei jedem Durchgang eine gründliche und präzise Durchführung sicherzustellen. Entscheidend ist der Systemtest, welcher als Teststufe dient, in der alle Anforderungen an das System geprüft werden, einschließlich funktionaler und nicht-funktionaler Anforderungen. Bei einer manuellen Testdurchführung durch Ingenieure kann die Konsistenz der Testergebnisse beeinträchtigt werden. Deshalb ist eine Testumgebung erforderlich, die weitgehend mit der Produktivumgebung übereinstimmt. Die direkt am Versuchsträger durchgeführten Tests sind in der Regel nicht reproduzierbar, da sich die Datenbasis und damit die Ausgangsbedingungen in der Produktivumgebung permanent verändern. Es ist nicht möglich, die Produktivumgebung auf den Ausgangszustand vor der Testdurchführung zurückzusetzen. Für die funktionalen Anforderungen wird der anforderungsbasierte Testansatz verwendet. Dies basiert auf einer Testspezifikation, die für solche Systemtests notwendig ist [3]. Langfristig gesehen lohnt sich die Testautomatisierung, da es ohne großen Aufwand bei neuen Testfällen angepasst beziehungsweise erweitert werden kann. Auch bestehen große Vorteile in der Schnelligkeit und Häufigkeit der Ausführungen von den Systemtests. Man hat die Möglichkeit, die Testfälle mit unterschiedlichen Daten zu befüllen und kann den manuellen Testaufwand deutlich reduzieren. Daher ist für wiederholende und regelmäßige Tests die Testautomatisierung unerlässlich.

Ziel

Das Ziel dieser Arbeit besteht darin, eine Toolchain zu entwickeln, die die automatisierte Durchführung von Tests sicherheitsrelevanter Komponenten im System ermöglicht. Die Testautomatisierungsumgebung muss in der Lage sein, das Gateway mit der neusten Software zu aktualisieren. Das Gateway soll anhand einer Testspezifikation getestet werden. Abschließend soll ein Testbericht erstellt werden, der folgende Informationen enthält:

- Testnummer

- Testbeschreibung
- Anfangsbedingung
- Ausführung
- Erwartetes Ergebnis
- Ausgelesenes Signal
- Prüfergebnis

Umsetzung

Für die Einrichtung der Toolchain ist ein Hardware-in-the-Loop (HiL)-System erforderlich, das die Simulation der kompletten Steuergeräteumgebung voraussetzt. Alle Sensoren, Aktuatoren oder Kommunikationsschnittstellen müssen entweder real dargestellt oder durch Simulation nachgebildet werden [1]. Zusätzlich muss die Stromversorgung über den angeschlossenen Rechner automatisiert gesteuert werden. Durch den Einsatz einer Continuous Integration/Continuous Delivery (CI/CD)-Pipeline soll die Testautomatisierung bei einem Push oder Pull Request ausgelöst werden, der über eine Versionsverwaltungssoftware wie beispielsweise Bitbucket initiiert wird. Dieser Ablauf ist in Abbildung 1 dargestellt.
Mithilfe von Python-Automatisierungsskripten soll das Gateway mit der neuesten Firmware gebaut und aktualisiert werden. Um das Gateway anschließend zu testen, muss vor dem Start das Gateway neu initialisiert werden. Einige Funktionen erfordern zudem Signale, die direkt vom Fahrzeug stammen. Daher ist es notwendig, geeignete Aufzeichnungen dieser Daten vorzunehmen, die später während des Testfalls simuliert werden können.
Die Testfälle können in Konfigurationsdateien angelegt werden und mit Python-Automatisierungsskripten ausgeführt werden. Nach erfolgreichem Abschluss und Erstellung eines Testberichts soll der Status an die CI/CD-Pipeline zurückgegeben werden.

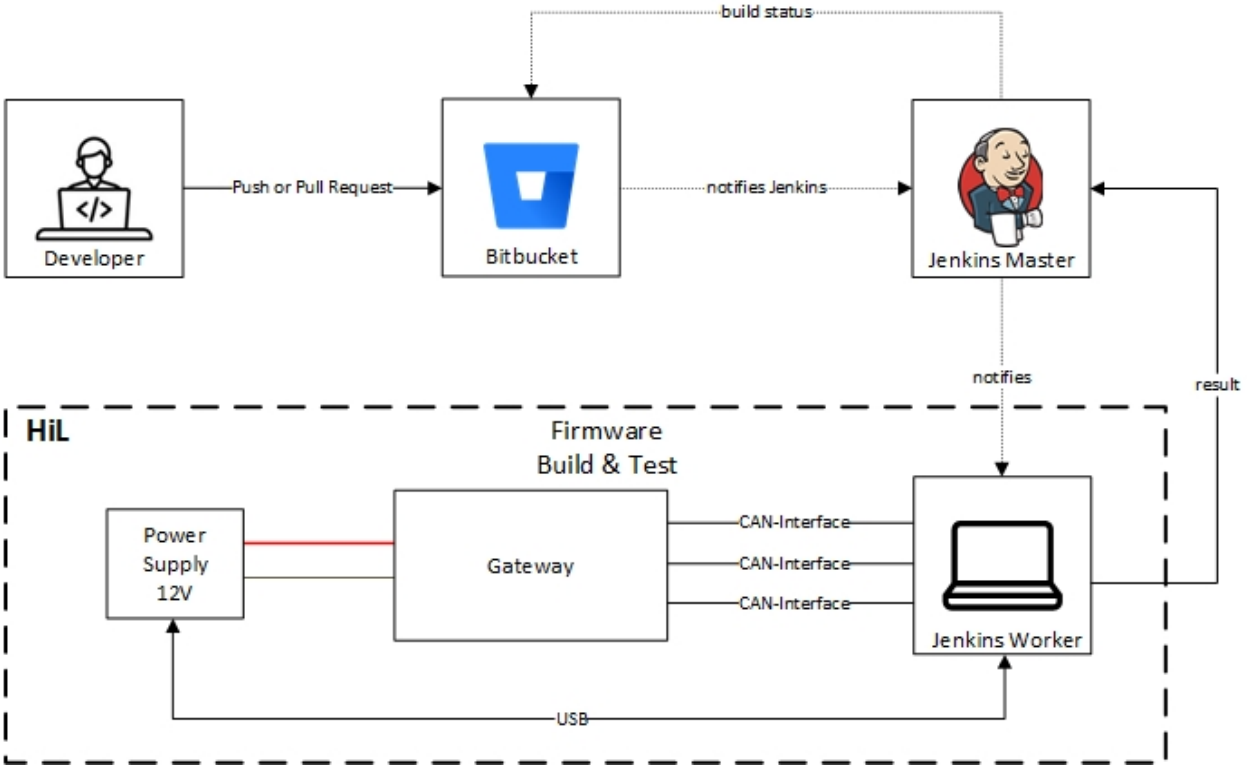


Abb. 1: Jenkins Workflow [2]

Ausblick

In der vorliegenden Bachelorthesis wurden bereits die Grundbausteine für die Testautomatisierung gelegt. Der Hardware-in-the-Loop (HiL)-Aufbau ist vollständig realisiert und die automatisierte Bereitstellung des Gateways mit der neuesten Software ist implementiert. Die Verbindungen für die Kommunikation zwischen dem Steuergerät und dem Ausführungsrechner sind hergestellt und die Grundlage der Testspezifikation ist ebenfalls geschaffen. Hierbei werden Funktionen im Zusammenhang mit der Notbremsung in den Fokus gestellt, da dies die sicherheitsrelevanteste Funktion ist. Sofern noch Zeit verfügbar ist, können weitere Funktionalitäten genauer geprüft werden. Um die Testabdeckung zu erweitern und die Gesamtfunktionalität

umfassend zu überprüfen, können Funktionalitäten wie die Lenkung oder das Adaptive Cruise Control genauer betrachtet werden. Diese Bachelorthesis trägt dazu bei, die Testprozesse für Fahrerassistenzsysteme effektiver zu gestalten und somit einen Beitrag zur Sicherheit und Zuverlässigkeit der Entwicklungsfahrzeuge zu leisten.

Literatur und Abbildungen

[1] Michael Bayer, Friedrich Munk, and Josef Schneider. Hardware-in-the-Loop-Prüfstand bei BMW für automatisierte Steuergerätetests. *ATZ Automobiltechnische Zeitschrift 100*, pages 696–702, 1998.

[2] Eigene Darstellung.

[3] Frank Witte. *Testmanagement und Softwaretest*. Springer Vieweg, 2 edition, 2019.

Anomaliedetektion von Signalmessungen: Optimierung von Data Handling und Implementieren eines AI Tools

Vivienne Beck **Clemens Klöck**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma Robert Bosch GmbH, Schwieberdingen

Einleitung

Als Zulieferer agiert die Firma Robert Bosch GmbH in vielen Bereichen. Ein großer Bereich in der Firma, in Bezug auf die Automobilindustrie, beschäftigt sich mit der Entwicklung von Steuergeräten. Nach der Freigabe eines Steuergeprojektes sind Änderungen jedoch nicht ausgeschlossen. Somit müssen gewisse Testverfahren implementiert werden.

Grundlagen

Bei Änderungen in Bezug auf die Produktion von Steuergerätprojekten unterscheidet man zwischen sogenannten *Functional* und *Non-Functional Changes*.
Functional Changes ändern die Funktion eines Steuergerätes. Hierbei handelt es sich zum Beispiel um Softwareänderungen.
Non-Functional Changes sind Änderungen, welche die eigentliche Funktion des Steuergerätes nicht beeinflussen. [3] Ein Fallbeispiel hierfür wäre: Ein Unternehmen stoppt die Produktion eines ICs, welcher in einem Steuergeräteprojekt verbaut wurde. Es muss nun ein funktionsgleicher IC eines anderen Herstellers gefunden werden. Diese Sicherstellung der Funktionalität wird durch ein sogenanntes *Back-To-Back*, kurz: B2B, Testverfahren nachgewiesen. Diese Tests werden in einer HiL (Hardware in the Loop) Umgebung durchgeführt. Bei einer HiL Umgebung handelt es sich um eine Simulationsumgebung, welche das Auto um das Steuergerät herum nachstellt. Dies ist effizienter als das Testen an physischen Prüfständen. Ebenso können die Tests unbeaufsichtigt und automatisiert ablaufen. Dies reduziert den Arbeitsaufwand der Mitarbeiter.
Im Rahmen des B2B Testings werden aktuell 99 Referenzmessungen am vorherigen Steuergerätemodell aufgenommen. [3] Hierbei entspricht eine Messung einem Fahrzyklus. Folglich wird der gleiche Fahrzyklus 99-mal durgelaufen. Hierbei gibt es verschiedene Fahrzyklen anhand welchen man testen kann. Die Stimuli (Input Daten) der Messung variieren in Korrelation zu dem Fahrzyklus.

Im Anschluss wird eine singuläre Messung an dem zu testenden Steuergerät durchgeführt.
Die ermittelten Daten aus den Referenzmessungen müssen für die bevorstehende Evaluierung aufgearbeitet werden. Im ersten Schritt werden mögliche Zeitversätze in den Messungen korrigiert. Dies ermöglicht eine präzise Fehlerevaluierung, da der Messoffset nicht fälschlicherweise als verspätete Antwort des Systems interpretiert wird.
Der Zeitversatz wird anhand eines Synchronisationsimpulses definiert. Bei dem Synchronisationsimpuls handelt es sich um ein Startimpuls, bevor der eigentliche Fahrzyklus anläuft. Ein Beispiel hierfür ist der Impuls von der Zündung (Klemme T15), welcher den Motor anlaufen lässt.
Die restlichen 98 Messungen werden dementsprechend um die Differenz zu der ersten Messung verschoben. Padding wird hier nicht betrieben. Lediglich Messungen, die nun von der Zeitdomäne in das Negative ragen, werden hier abgeschnitten. Folglich ist eine Verbildlichung solch eines Verfahrens anhand der Zündung:

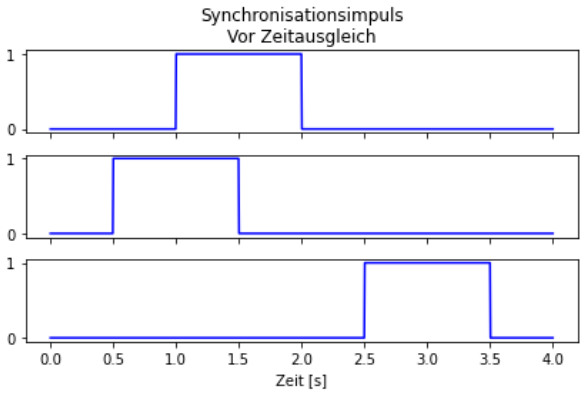


Abb. 1: Synchornisationsimpuls vor Zeitausgleich [2]

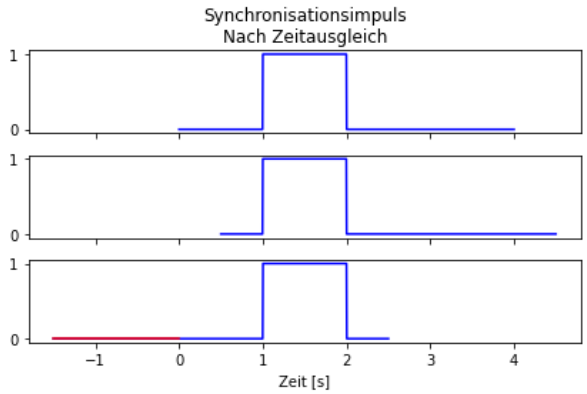


Abb. 2: Synchornisationsimpuls nach Zeitausgleich [2]

Der in Rot gekennzeichnete Bereich in Abbildung 2 wird somit weggeschnitten.
Sind die Daten aufgearbeitet werden sie anhand von einem internen Tool namens *JBeam* evaluiert. Dieses Tool erstellt auf Basis der Referenzmessungen ein Toleranzband. Ebenso wird ein Durchschnittssignal ermittelt. Die singuläre Messung von dem zu testenden Steuergerät wird über dieses Toleranzband gelegt. So wird die Messung in Bezug auf das Toleranzband und Durchschnittssignal in Vergleich gestellt. Die Bewertung errechnet sich nach einer ISO-Norm und fällt in einen Bereich zwischen Eins und Null. Hierbei ist Eins Deckungsgleich und Null verweist auf keine Korrelation zwischen den Signalen.

Problemstellung

Die oben genannte ISO, nach welcher bewertet wird, ist nicht optimal für den hier vorliegenden Use Case. Infolgedessen werden Signalverläufe, welche eigentlich in Ordnung sind, schlecht bewertet. Deshalb ist ein neues Bewertungsschema gefragt. AI wurde hier aufgrund seiner Effizienz als Hilfsmittel in Betracht gezogen. Nach genauerer Recherche wurde festgelegt, dass Anomaly Detection hierfür ein guter Ansatz ist. Dies ist eine AI-Methode, um Ausreißer in Datensätzen zu ermitteln.

Durchführung

In dem Umfang der Thesis werden die ersten Schritte einer Evaluierung der Signale anhand solch eines AI-Models eingeleitet. Der Fokus liegt hierbei auf der Analyse des Data Handling. Fortlaufend wird in dieser Zusammenfassung eine kleine Übersicht über den Inhalt der Thesis geschaffen. Zusätzlich werden vereinzelte Ergebnisse aufgezeigt.

Zwei Auto Encoder Modelle wurden zur Evaluierung in Erwägung gezogen: das Keras Modell [4] und das Dense Modell [1].
Das Keras Modell basiert auf Convolutional Layern und das Dense Modell auf Dense Layern. Im ersten Schritt wird ermittelt, wie die Modelle auf Änderungen und Vertauschen der Datendimensionen reagieren. Hierbei wurden die Antworten der Modelle einerseits mit den Dimensionen [3745, 99] sowie [99,3745] getestet. 3745 entspricht hierbei einer komprimierten Messpunkteanzahl der Files. Diese Komprimierung fand anfangs lediglich zum einsparen an Trainingszeit statt. Im Falle des Keras Modells musste aufgrund der Convolutional Layer in der zweiten Dimension Padding betrieben werden.
Das Keras Modell antwortet in beiden Fällen mit einem Offset. Somit wurde dieses Modell verworfen, da es für den vorliegenden Datensatz als ungeeignet eingestuft wurde.

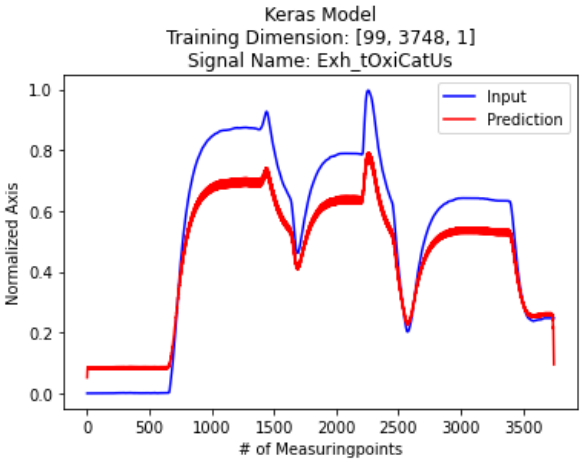


Abb. 3: Keras Model Rekonstruktion von dem Signal Exh_tOxiCatUs, Dimension [99, 3748, 1] [2]

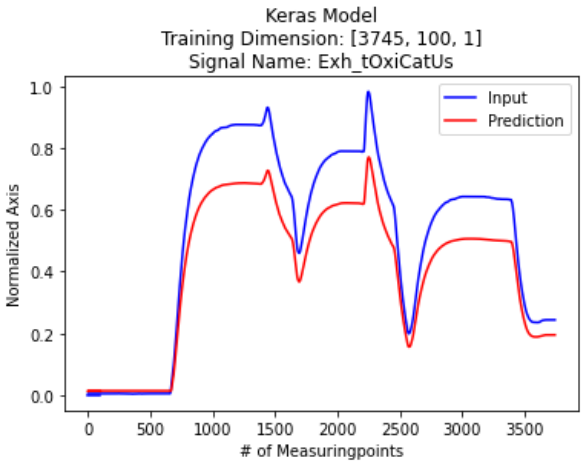


Abb. 4: Keras Model Rekonstruktion von dem Signal Exh_tOxiCatUs, Dimension [3745, 100, 1] [2]

Das Dense Modell antwortet größten Teils mit zufriedenstellenden Rekonstruktionen des Signals. Somit wurde dieses Model zur weiteren Evaluierung verwendet.

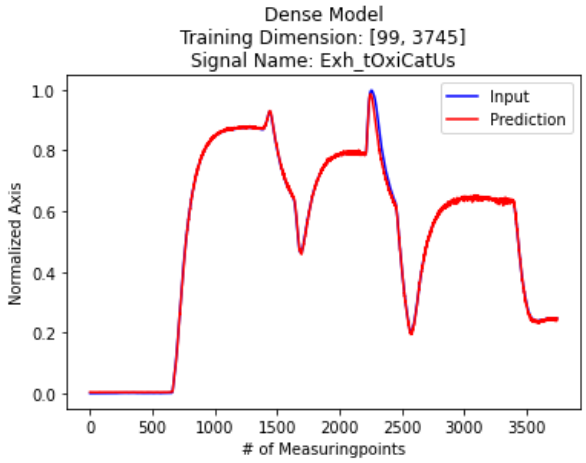


Abb. 5: Dense Model Rekonstruktion von dem Signal Exh_tOxiCatUs, Dimension [99, 3745] [2]

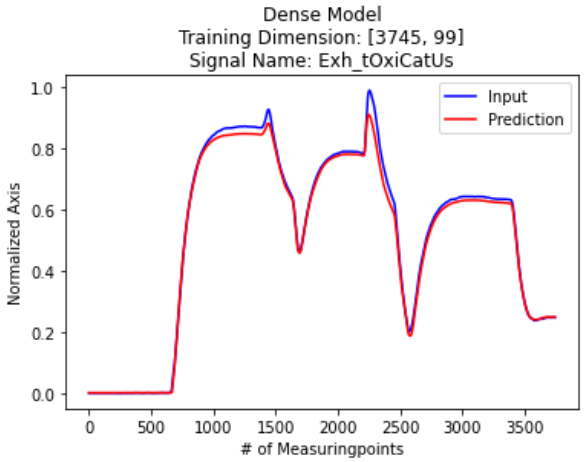


Abb. 6: Dense Model Rekonstruktion von dem Signal Exh_tOxiCatUs, Dimension [3745, 99] [2]

Im Anschluss wurde evaluiert, wie man die Effizienz des Models anhand von dem Datensatz verbessern kann ohne zu einschränkend an Genauigkeit zu verlieren. Dazu wird die Anzahl der Messpunkte schrittweise reduziert und der Genauigkeitsverlust des Models analysiert.
Anhand von den Erkenntnissen wird ein erstes Tool entwickelt. Dies soll neben dem Tool *JBeam* eine alternative Methode zur Datenauswertung liefern. Dadurch, dass sich die Fahrzyklen je nach Testdurchlauf ändern können lernt sich das Tool anhand der Referenzmessungen jederzeit neu ein. Dies wurde mit dem Fokus auf den hier vorliegenden Use Case entschieden. Trotz der Zeitoptimierung durch das Data Handling ist dies jedoch langfristig nicht effizient.

Ausblick

Fortlaufend ist ein Ansatz in Planung, in welchem sich das Model nicht mehr auf die Signalverläufe einlernt, sondern auf die Korrelation zwischen den Signalen. Ebenso soll ermittelt werden, ob die optimale Abtastfrequenz der Signale dynamisch anhand des Abtasttheorems errechnet werden kann.

Literatur und Abbildungen

[1] Raghav Agrawal. Complete Guide to Anomaly Detection with AutoEncoders using Tensorflow. https://www.analyticsvidhya.com/blog/2022/01/complete-guide-to-anomaly-detection-with-autoencoders-using-tensorflow/, 2022.

[2] Eigene Darstellung.

[3] Sanjay Mohan Menon. B2B Test (Automation). https://inside-docupedia.bosch.com/confluence/pages/viewpage.action?pageId=3274481587, 2023.

[4] Pavithra Vijay. Timeseries anomaly detection using an Autoencoder. https://keras.io/examples/timeseries/timeseries_anomaly_detection/, 2020.

Evaluierung von serverless Computinglösungen anhand einer Beispielanwendung

Eric Beller **Mirko Sonntag**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma pep.digital GmbH, Esslingen am Neckar

Einleitung

Die Entwicklung und Verbreitung von Cloud-Computing-Technologien hat die Geschäfts- und Technologielandschaften wesentlich beeinflusst und ist ein zentraler Bestandteil der digitalen Transformation. Ein Schlüsselbeispiel dieser Entwicklung ist das Serverless Computing. Wie in Abbildung 1 dargestellt, nimmt die Popularität des Serverless Computings, basierend auf Google Trends, stetig zu. Diese Technologie markiert einen deutlichen Wandel von traditionellen Serverarchitekturen und verspricht eine erhöhte Effizienz und Kosteneffektivität für Cloud-Anwendungen. [1]

Serverless Computing überträgt traditionelle Serververwaltungsaufgaben auf Cloud-Provider, wodurch Entwickler sich auf Programmierung und Geschäftslogik konzentrieren können, ohne sich um Skalierung, Wartung und Sicherheit sorgen zu müssen. Diese Entkopplung verspricht eine agilere Entwicklungsumgebung, schnelle Reaktion auf Marktanforderungen und eine effizientere Nutzung von Ressourcen.

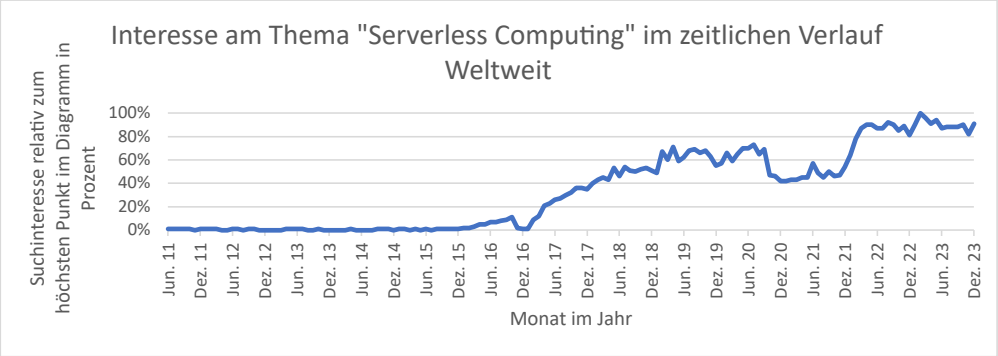


Abb. 1: Interesse am Thema SServerless Computingïm zeitlichen Verlauf Weltweit [4]

Zielsetzung

Das Ziel der Arbeit ist die Evaluierung von Serverless Computing-Lösungen, wobei der Schwerpunkt speziell auf Azure Functions und AWS Lambdas liegt. Diese Auswahl basiert auf der Beliebtheit und Verbreitung dieser Plattformen, wie in Abbildung 2 dargestellt. Die Evaluierung erfolgt durch die Entwicklung und den Einsatz einer Beispielanwendung. Im Fokus steht dabei, ein tiefgreifendes Verständnis für die Leistungsfähigkeit, Skalierbarkeit und Kosten-Effizienz von Azure Functions und AWS Lambdas zu entwickeln. Zusätzlich zielt diese Arbeit darauf ab, die Eignung dieser spezifischen Serverless Computing-Lösungen für reale Anwendungsanforderungen zu untersuchen. Durch diese fokussierte Betrachtung ermöglicht die Arbeit nicht nur eine theoretische Auseinandersetzung mit dem Thema Serverless Computing, sondern bietet auch eine praxisorientierte Perspektive, um die Vorteile und Nachteile von Azure Functions und AWS Lambdas in einem realen Umfeld zu erkunden.

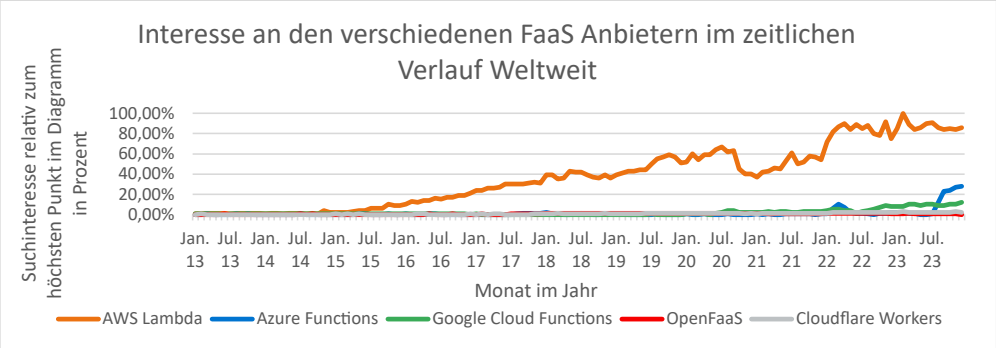


Abb. 2: Interesse an den verschiedenen FaaS Anbietern im zeitlichen Verlauf Weltweit [3]

Serverless Computing

Serverless Computing ermöglicht das Ausführen von Anwendungscode, ohne sich um das Betriebssystem, die Softwareplattform oder die zugrunde liegende Hardware kümmern zu müssen. Es wird von den meisten öffentlichen Cloud-Anbietern als Managed Service angeboten. Entwickler schreiben lediglich ihren Code oder Funktionen für spezifische Aufgaben, wählen die erforderlichen Ressourcen aus und übermitteln ihn zur Ausführung an den Serverless-Cloud-Computing-Dienst. Dieser sorgt dafür, dass der Code die benötigten Speicher- und CPU-Ressourcen erhält, wobei die Abrechnung auf der Ausführungsdauer und dem Speicherverbrauch basiert.

Function as a Service

Function as a Service (FaaS) ist ein zentrales Element des Serverless Computing. Bei FaaS erfolgt die Ausführung ereignisgesteuert, was zu effizienter Ressourcennutzung und Kosteneinsparungen führt, da nur für die tatsächliche Rechenzeit bezahlt wird. AWS Lambda ist ein prominentes Beispiel für FaaS, das automatisierte Ausführung, Skalierung und Verfügbarkeit bietet.
FaaS erlaubt die Verwendung verschiedener Programmiersprachen, wodurch Entwickler flexibel in der Wahl ihrer Tools sind. [2] Die Funktionen sind zustandslos und temporär, wobei externe Speicherlösungen den Anwendungsstatus verwalten. [2] Automatische Skalierung und vielfältige Auslösemechanismen für Funktionen sind weitere Vorteile, während Begrenzungen wie maximale Ausführungsdauer und "Cold Starts"bei der Planung berücksichtigt werden müssen. Das API Gateway spielt in FaaS-Architekturen eine wichtige Rolle. Es dient als Schnittstelle, die eingehende HTTP-Anfragen basierend auf definierten Routen an die entsprechenden FaaS-Funktionen weiterleitet. Das Gateway wandelt die Antworten der Funktionen in HTTP-Antworten um und sendet diese zurück zum Aufrufer. Dies ermöglicht die einfache Integration von FaaS in webbasierte Anwendungen und Dienste. Diese Gateways erleichtern die Erstellung und Verwaltung von HTTP-basierten Microservices, indem sie das Routing, die Überwachung und die Sicherheit in einer serverlosen Umgebung übernehmen. Die Funktionsweise des Amazon API Gateway wird in Abbildung 3 dargestellt:

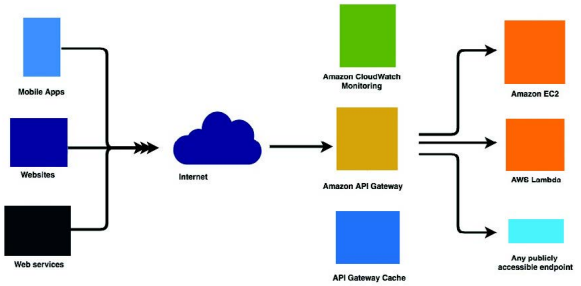


Abb. 3: Amazon API Gateway [2]

Ausblick

In der vergleichenden Analyse der Arbeit werden Azure Functions und AWS Lambdas in Bezug auf verschiedene Schlüsselelemente gegenübergestellt. Hierzu gehört der Vergleich der Trigger-Mechanismen, der die Reaktion auf Ereignisse und die Auslösung von Funktionen betrachtet. Ein wichtiger Teil der Analyse ist der Cold-Start-Zeit Vergleich, der die Reaktionsfähigkeit der Plattformen aufzeigt. Zusätzlich wird die Performance beider Dienste bewertet, um Effizienz und Verarbeitungsgeschwindigkeit zu messen. Ein Kostenvergleich gibt Aufschluss über die Wirtschaftlichkeit der Plattformen. Schließlich werden die Entwicklungsumgebungen, Werkzeuge sowie die Community und der Support beider Anbieter untersucht, um die Benutzererfahrung und die verfügbare Unterstützung zu evaluieren. Diese umfassende Betrachtung bietet wichtige Erkenntnisse für die Auswahl der geeignetsten Serverless-Computing-Plattform.

Literatur und Abbildungen

[1] Ioana Baldini, Paul Castro, Kerry Chang, Perry Cheng, Stephen Fink, Vatche Ishakian, Nick Mitchell, Vinod Muthusamy, Rodric Rabbah, Aleksander Slominski, and Philippe Suter. Serverless Computing: Current Trends and Open Problems. In *Research Advances in Cloud Computing*, pages 1–20. Chaudhary, Sanjay; Somani, Gaurav; Buyya, Rajkumar, 2017.

[2] Kuldeep Chowhan. *Hands-on serverless computing*. Packt Publishing, 2018.

[3] Google Trends. AWS Lambda, Azure Functions, Google Cloud Functions, OpenFaaS, Cloudflare Workers - Erkunden - Google Trends. https://trends.google.com/trends/explore?date=2013-01-01%202023-12-05&q=%2Fg%2F11bw4c_dgq,%2Fg%2F11dfh7gkm4,%2Fg%2F11fmgwwwsb,OpenFaaS,Cloudflare%20Workers&hl=de, 2023.

[4] Google Trends. Serverless Computing - Erkunden - Google Trends. https://trends.google.com/trends/explore?date=2011-01-01%202023-11-30&q=%2Fg%2F11c0q_754d, 2023.

Analyse von Architekturen für digitale Fahrzeug-Zugangssysteme

Nils Benecke **Rainer Keller**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma Robert Bosch GmbH, Stuttgart

Motivation

Ein neues Feld der Fahrzeugfunktionen ist in den letzten Jahren in den Vordergrund geraten, Vehicle-to-Everything (V2X). Dabei geht es um Funktionen, welche das Fahrzeug kontaktlos mit Systemen aus dem Umfeld verbindet. Ein Beispiel für eine V2X Anwendung ist das passive schlüssellose Öffnen und Starten des Fahrzeugs. Das Fahrzeug detektiert den heranlaufenden Fahrer und entriegelt die entsprechende Türe. Auch das Starten des Fahrzeugs kann durch die Anwendung befähigt werden, sobald der Fahrer sich im inneren befindet. Das diese Anwendung hochgradig Sicherheitsrelevante Aspekte hat, ist offensichtlich. Vor allem da in der Vergangenheit und immer noch eingeführte Fahrzeug-Zugangssysteme viele Sicherheitslücken, hinsichtlich Replay-, Relay- und Spoofing-Angriffen, aufweisen [2]. Mit neuen Technologien wie dem Lokalisieren mittels Ultra-Wideband (UWB), sollen die Sicherheitsmängel durch Relay-Attacken behoben werden. Diese Technologie ist aber bisher nur in etwa 10% der Neuzugelassenen Fahrzeugen für das Lokalisieren in Fahrzeug-Zugangssystemen eingesetzt. Mit den Ergebnissen dieser Arbeit soll die Attraktivität von sicheren Fahrzeug-Zugangssystemen mit UWB-Technologie verbessert werden.

Das Car Connectivity Consortium (CCC) hat einen sogenannten „Digital Key“ spezifiziert. Es stellt eine mögliche Umsetzung eines digitalen Fahrzeug-Zugangssystem mit UWB-Technologie dar. Dieser CCC Digital Key standardisiert die Struktur des digitalen Schlüssels, die Schnittstellen und benutzten Protokolle zwischen Fahrzeug und mobilem Endgerät mittels Bluetooth Low Energy (BLE) und Ultra-Wideband (UWD), als auch die Cloud-basierte Schlüsselinfrastruktur [4]. Das Digital Key Ökosystem ist in Abbildung 1 skizziert.

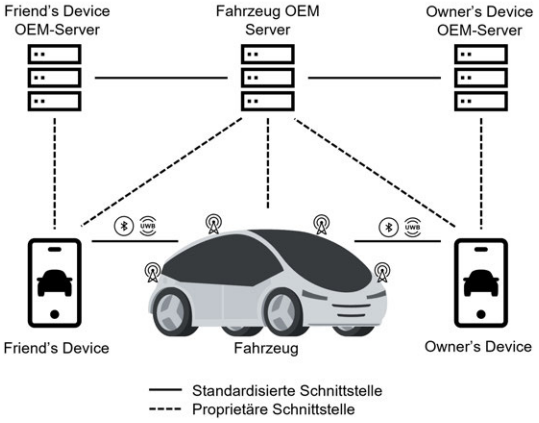


Abb. 1: CCC Digital Key Ökosystem nach [3]

Problemstellung

Die konkrete Umsetzung und fahrzeugseitige Architektur eines Fahrzeug-Zugangssystems, welches mit der Spezifikation des CCC Digital Key konform ist, wird nicht beschrieben. Fahrzeug-seitig muss das System minimal eine Schnittstelle mit UWB und BLE nach Außen darstellen sowie die Protokolle und kryptografischen Funktionen, die mit dem Verbindungsaufbau und Lokalisierungsprozess verbunden sind. Um erweiterte Funktionen mit genauer Lokalisierung des mobilen Endgerätes zum Fahrzeug zu erfüllen, müssen verschiedene UWB- und BLE-Schnittstellen verteilt im Fahrzeug implementiert werden.

Die dadurch resultierenden Systemarchitekturen können alle als verteilte Systeme eingeordnet werden. Ziel der Arbeit soll es ein, mögliche Architekturen auszuarbeiten, zu analysieren und miteinander zu vergleichen. Gerade die zeitlichen Anforderungen sind zentral für die Analyse der Architekturen, da die Komfortfunktion des passiven schlüssellosen Öffnens des Fahrzeugs für den Benutzer unmerklich durchgeführt werden soll. Aber auch Eigenschaften wie Funktionalität, Security, Komplexität, Stromverbrauch, Kosten, Integrierbarkeit,

Skalierbarkeit und Erweiterbarkeit werden analysiert. Das Ergebnis soll als Grundlage für weitere Analysearbeiten und für die Auswahl einer fahrzeugseitigen Systemarchitektur für CCC Digital Key kompatible Fahrzeug-Zugangssysteme sein.

Theoretische Analyse

Die Analyse startet mit der Betrachtung des zentralen Anwendungsfall, dem Passive Keyless Entrys (PKE). Grundsätzlich startet ein PKE-Ablauf mit dem Verbindungsaufbau des mobilen Endgeräts (Device) und dem Fahrzeug. Dabei wird eine verschlüsselte BLE-Verbindung aufgebaut, über welche dann eine gegenseitige Authentifizierung mit dem jeweiligen Digital Key stattfindet. Nach erfolgreicher Authentifizierung wird die weitere Kommunikation zusätzlich verschlüsselt, es werden darin Parameter für das Initialisieren einer Lokalisierungssitzung mittels UWB ausgetauscht. Das Device wird daraufhin von dem Fahrzeug lokalisiert und entsprechend wird die Tür entriegelt, sobald sich der Nutzer in einer bestimmten Zone befindet.
Aus dem Ablauf können verschiedene Systemkomponenten identifiziert und auf ihre Systemanforderungen untersucht werden. Anschließend sind drei mögliche Architekturen beschrieben, welche den Rahmenbedingungen entsprechen und alle Systemkomponenten enthalten. Alle Architekturen benötigen mindestens vier UWB-Schnittstellen (Anker) verteilt im Fahrzeug, um eine eindeutige Positionsbestimmung per Trilateration durchzuführen [1]. Ein BLE-Zugriffspunkt ist ebenfalls in jeder Architektur verpflichtend. In Abbildung 2 ist eine Architektur mit Primary Steuergerät dargestellt, mit diesem Steuergerät werden zentralisiert alle Systemkomponenten ausgeführt und die Anker gesteuert.

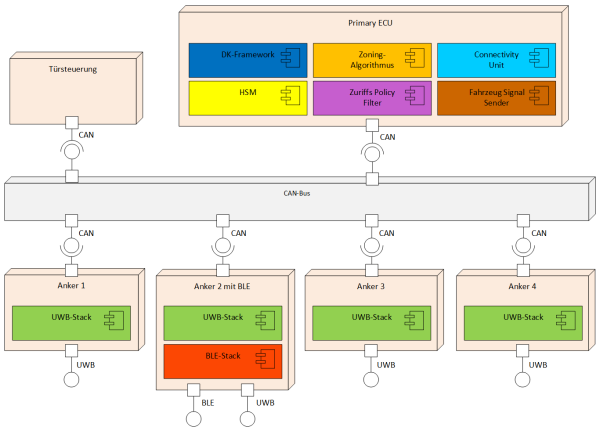


Abb. 2: Architektur zentralisiert mit Primary Steuergerät [5]

In Abbildung 3 hingegen ist eine weitere Architektur dargestellt, bei der keine zentrale Steuereinheit notwendig ist und alleinig die Anker die Funktionalitäten der Systemkomponenten abbilden. Ein Anker ist maßgeblich führ den Verbindungsaufbau und das Steuern der Lokalisierung verantwortlich. Ein weiterer Anker übernimmt die Bestimmung der Position, das Entscheiden und Ausführen der Türsteuerung. Alle anderen Anker sind gleich zur vorherigen zentralisierten Architektur.

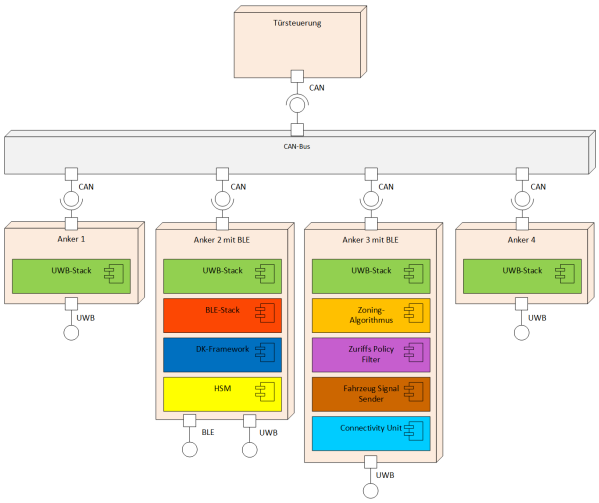


Abb. 3: Architektur verteilt auf Anker [5]

Danach konnten die verschiedenen Architekturen hinsichtlich der Eigenschaften analysiert und verglichen werden. Unter anderem sind die Sequenzen der Architekturen bei einem PKE-Anwendungsfall abgebildet, um die zeitliche Performance zu analysieren.
Das Ergebnis der theoretischen Analyse zeigt unter anderem eine Performance Verbesserung der dezentralisierten Architektur im Gegensatz zu der zentralisierten Architektur auf. Durch das verlagern der Systemkomponenten auf den Anker mit BLE-Schnittstelle, wird ein Großteil der CAN-Kommunikation zwischen Primary Steuergerät und Anker überflüssig.

Ausblick

Es soll in einer praktischen Analyse die zeitliche Performance der Architekturen bestimmt werden. Dazu werden Sequenzen des PKE-Verbindungsaufbaus zwischen zwei BLE-Evaluations Boards implementiert und die Laufzeit bzw. benötigten Taktzyklen gemessen. Weitere Arbeiten können an die theoretische Analyse anknüpfen und weiter ausführen bzw. mit konkreten praktischen Analysen ergänzen.

Literatur und Abbildungen

[1] Peyman Askari and Fardin Barekat. *Trilateration and Bilateration In 3D and 2D Space Using Active Tags*. PRILYX Research and Development LTD, 2017.

[2] Allgemeine Deutsche Automobil Club. Autos und Motorräder mit Keyless-Schließsystem, die der ADAC illegal öffnen und illegal wegfahren konnte. https://assets.adac.de/image/upload/v1698649839/ADAC-eV/KOR/Text/PDF/31639_kofpvk.pdf, 10 2023.

[3] Car Connectivity Consortium. *Whitepaper CCC Digital Key Release 3.0 - The Future of Vehicle Access*. Car Connectivity Consortium, 2021.

[4] Car Connectivity Consortium. *Digital Key Technical Specification Release 3*. Car Connectivity Consortium, 2022.

[5] Eigene Darstellung.

Fehleranalyse von mit Convolutional-Neural-Networks generierter Produktbilder mittels Machine-Learning Algorithmen: Beurteilung der Tauglichkeit hinsichtlich Produktkorrektheit und Markentreue.

Nicolas Beugel **Dieter Morgenroth**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt an der Fakultät Informatik und Informationstechnik, Esslingen

Einleitung und Problemstellung

Die Nutzung von künstlicher Intelligenz hat in den vergangenen Jahren stark zugenommen. Vor allem Programme wie ChatGPT oder DALL-E haben künstliche Intelligenzen zugänglich für jede Person gemacht. Mithilfe von bildgenerierenden künstlichen Intelligenzen kann jedermann mit nur einer einfachen Texteingabe ein Bild erstellen lassen, welches so vorher noch nie existierte. Diese neue und revolutionäre Technologie der autonomen Bildgenerierung bietet unzählige Anwendungsmöglichkeiten. Mithilfe von neuronalen Netzen erlernt die künstliche Intelligenz die Fähigkeit Bilder in verschiedensten Stilen und mittlerweile täuschend echt zu erstellen.

Diese Möglichkeit der automatischen Bildgenerierung kann beispielsweise in der Industrie verwendet werden, indem Bilder von Produkten in ganz bestimmten Situationen und Umgebungen für Kunden erstellt werden, ohne hierfür das Produkt in diesem speziellen Szenario fotografieren zu müssen. So kann beispielsweise ein Kunde, welcher daran interessiert ist, ein neues Auto zu kaufen, ein Modell in seiner gewünschten Farbe in einem ganz bestimmten Szenario generieren. Ein denkbares Beispiel wäre ein Förster, welcher einen neuen SUV im Wald sehen möchte, diese Eingabe könnte so aussehen: 'Zeige mir den neuen SUV in dunkelblau auf einem Waldweg'. Führt man dieses Prinzip weiter, ist es ebenfalls vorstellbar, dass man irgendwann ein Foto von seiner Einfahrt hochladen kann und die künstliche Intelligenz das Auto visuell in diese Einfahrt stellen kann.

Da es sich für den Automobilhersteller um einen möglichen Kunden handelt, möchte dieser logischerweise jeden visuellen Fehler an dem gezeigten Fahrzeug vermeiden. Momentan funktioniert die Bildgenerierung mithilfe von künstlichen Intelligenzen noch nicht fehlerfrei, weswegen es häufig noch erkennbare, aber vereinzelte Fehler bei solchen generierten Bildern gibt.

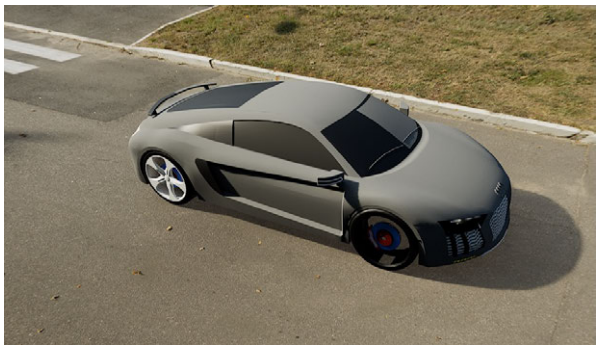


Abb. 1: Fehlerhaft generiertes Fahrzeug mit 3dcad-browse [1]

In Abbildung 1 kann man ein generiertes Fahrzeug mit einer fehlenden Felge erkennen. Diese Fehler, auch Anomalien genannt, gilt es zu erkennen und durch erneute Generierung zu eliminieren. Ziel dieser Thesis ist es, solche generierten Produktbilder zu analysieren, Anomalien zu erkennen und zu bewerten, ob das generierte Bild akzeptabel ist, um es anschließend einem möglichen Kunden zu präsentieren.

Methode

Zur Erkennung von Anomalien wird im Rahmen dieser Arbeit eine weitere künstliche Intelligenz zum Einsatz kommen, welche die erstellten Bilder auf Anomalien überprüft. Hierfür werden zwei Ansätze getestet. Zum einen werden Bilder über eine Objektklassifizierung versucht in fehlerhaft und nicht fehlerhaft einzuordnen, zum anderen werden Fehler mithilfe einer outlier-detection gesucht.

Beim Ansatz der Objektklassifizierung lernt ein Machine-Learning Algorithmus mithilfe eines vorhan-

denen Datensatzes Objekte in Bildern zu erkennen und diese Bilder dann mit vorgegebenen Bezeichnungen zu versehen. Der Datensatz, welcher zum Anlernen des Algorithmus verwendet wird, besteht aus fehlerhaften Produktbildern bei denen die Fehler bekannt und schon klassifiziert sind. Hat der Algorithmus diesen Datensatz verarbeitet, kann er verwendet werden, um ein nicht klassifiziertes Bild auf größere Fehler zu überprüfen und fehlerhafte Bilder anschließend in Kategorien, wie 'fehlerhafte Felge' oder 'fehlendes Kennzeichen' einzuordnen. Dieser Ansatz ist leicht umzusetzen, man braucht hierfür lediglich einen möglichst großen Datensatz zum Anlernen, was jedoch auch das größte Problem dieses Ansatzes darstellt, da es noch keinen Datensatz hierzu gibt und dieser erst manuell erstellt werden muss, welches ein sehr zeitintensives Verfahren ist. [2]
Die Methode der outlier-detection nutzt einfache Statistik. Generierte Datensätze lassen sich entweder in eine Normalverteilung oder in nicht normalverteilt einordnen. Entstandene Anomalien machen sich in Datensätzen als Ausreißer oder im Englischen als sogenannter outlier bemerkbar. Diese Ausreißer unterscheiden sich stark von den restlichen Daten. Zum Erkennen dieser Ausreißer muss man den Datensatz mithilfe von Schranken eingrenzen, jene Daten die nun außerhalb dieser Grenzen liegen werden als Ausreißer klassifiziert.
Folgt der Datensatz einer Normalverteilung, kann man Ausreißer mit einer Standardabweichung eliminieren. Zuerst ermittelt man die Standardabweichung (Sigma) und den Mittelwert (My) des Datensatzes. Anschließend definiert man, in welchem Bereich der Standardabweichungen die Datenpunkte als normal angesehen werden. Abbildung 2 zeigt eine solche Standardabweichung. In den meisten Fällen findet das in folgendem Bereich statt: $\mu - 3\sigma, \mu + 3\sigma$

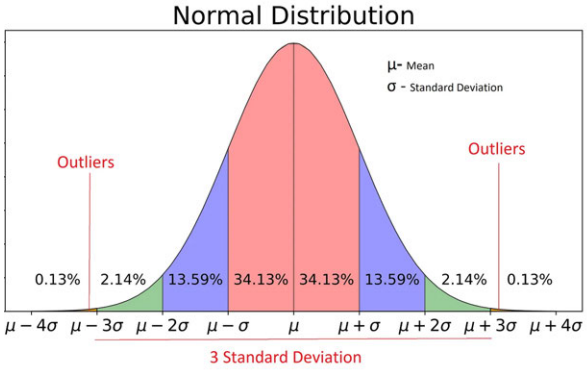


Abb. 2: Normalverteilung mit Ausreißern [3]

Für nicht normalverteilte Datensätze kann man Ausreißer mit dem Box-Plot Verfahren eliminieren. [4] Hierbei werden die Ausreißer über die Quartile definiert. Die Box geht vom unteren zum oberen Quartil und beinhaltet den Median, also den Mittelpunkt. Die Box beinhaltet also alles von Quartil 1-3. Des Weiteren werden rechts und links der Box die sogenannten Whisker eingezeichnet. Diese beinhalten die unteren und oberen 25% der Daten. Sind die Whisker länger als das 1,5 fache der Box, sind alle Datenpunkte, die außerhalb der Whisker liegen als Ausreißer definiert. [5] Abbildung 3 zeigt eine visuelle Darstellung eines Box-Plot mit eingezeichneten Ausreißern.

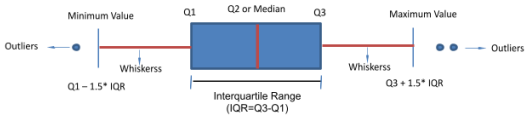


Abb. 3: Beispielhafte Box-Plot Graphik mit Ausreißern [5]

Produktkorrektheit und Markentreue

Da die generierten Bilder im Idealfall möglichen Kunden präsentiert werden sollen, müssen diese so echt wie möglich wirken. Produkte haben bestimmte Identitäten von Herstellern und diese dürfen nicht verloren gehen. So muss beispielsweise das Logo der Marke oder der Modellname eindeutig erkennbar sein. Sieht der Kunde möglicherweise ein fehlerhaftes Produkt in nicht originalen Dimensionen, wird dieser möglicherweise vom Kauf abgeschreckt.
Man kann davon ausgehen, dass die im vorherigen Kapitel definierten Ausreißer nicht unter die Definition der Produktkorrektheit und Markentreue fallen. Diese werden aussortiert und auf den Fehlergrund analysiert.

Ausblick

Künstliche Intelligenzen werden sich in Zukunft immer weiter verbreiten und den Alltag immer weiter vereinfachen. Die Fähigkeit Bilder produktecht zu generieren und einem Kunden zu präsentieren macht es potenziellen Käufern leichter sich für ein Produkt zu entscheiden und die Hersteller können so ihre Produkte leicht und günstig dem Kunden präsentieren. Im weiteren Verlauf dieser Arbeit wird analysiert, mit welcher Methode Fehler am schnellsten und möglichst zuverlässig analysiert und herausgefiltert werden können. Momentan ist diese Technik neu und fehlerhaft. Doch mit weiterer Forschung und weiteren Investitionen in diesem Gebiet ist es sehr wahrscheinlich, dass es in einigen Jahren möglich ist, in kürzester Zeit sich Produkte beliebig nach Fantasie in jeglichen Szenarien generieren zu lassen.

Literatur und Abbildungen

[1] Eigene Darstellung.

[2] Carsten Keller and Patrick T. Philipp. Arten von ML in der Praxis [Teil 1]: Regression vs. Klassifikation. https://www.finbridge.de/ml-artikel/2023/01/10/arten-ml-regression-klassifikation-teil-1#:~:text=Im%20Gegensatz%20zur%20Regression%20lernt,und%20in%20verschiedene%20Gruppen%20einzuordnen., 2023.

[3] Gautum Kumar. How to identify the outliers in your Data?? https://kmr-gautam2893.medium.com/how-to-identify-the-outliers-in-your-data-ee9c28b42fc3, 2020.

[4] Artem Oppermann. Ausreißer und Anomalien in den Daten. https://artemoppermann.com/de/ausreisser-und-anomalien-in-den-daten/, 2021.

[5] Avantika Shkula. Understanding BoxPlot | What is BoxPlot. https://www.mygreatlearning.com/blog/understanding-box-plot/, 2022.

Untersuchung von Instance Segmentation für Straßen und Fahrspuren in drohnenbasierten Luftaufnahmen mittels YOLOv8 und Segment Anything

Mika Boehm **Markus Enzweiler**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma IT-Designers Gruppe, Esslingen

Einleitung

Maschinelles Lernen im Bereich der *Machine Vision* ermöglicht es, Daten aus Bildern mit hoher Präzision zu extrahieren und zu analysieren. Es erlaubt die präzise Erkennung von Objekten in Bildern, die in der Vergangenheit entweder unmöglich oder nur mit erheblichem Aufwand zu bewerkstelligen war. *Machine Vision* Modelle finden Anwendung in einer Vielzahl von Bereichen – von der autonomen Navigation über die Verkehrsanalyse bis hin zu medizinischen Diagnoseverfahren.

Problemstellung

Im Rahmen eines Projektes zur automatisierten Verkehrsmessung werden bereits fortgeschrittene *Machine Vision* Modelle eingesetzt, um Fahrzeuge zu erkennen. Um diese Technologie weiterzuentwickeln wird ein Modell benötigt, das Fahrspuren in drohnenbasierten Luftaufnahmen präzise erkennen und segmentieren kann. Jede Fahrspur soll dabei als separate Instanz betrachtet werden.

Die in der Abbildung 1 dargestellte Szenerie veranschaulicht die angestrebte Anwendung des Modells. Die Abbildung dient zur Veranschaulichung für die Art von Daten, die das zu entwickelnde Modell automatisch generieren soll. Die manuellen Fahrspurdefinitionen, die in der Abbildung zu sehen sind, repräsentieren den aktuellen, arbeitsintensiven Prozess, der durch die Automatisierung vereinfacht und beschleunigt werden soll.

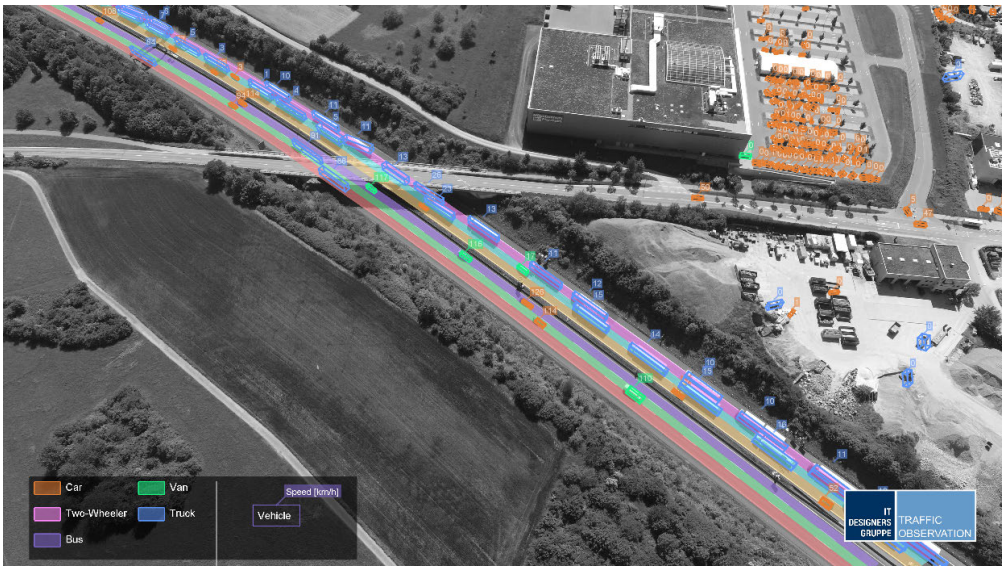


Abb. 1: Mögliche Praxisanwendung des Modells [3]

Zielsetzung

In dieser Arbeit soll die Anwendbarkeit des YOLOv8-Modells für den oben genannten Anwendungsfall untersucht werden. Das YOLOv8 Modell wird anhand von Transfer Learning auf einen neuen Datensatz angepasst, der aus einer Vielzahl von Straßenszenarien aus unterschiedlichen drohnenbasierten Perspektiven besteht. Durch den hohen Aufwand des manuellen Annotationsprozesses wird zudem untersucht, wie ein Datensatz mit verschiedenen Augmentierungstechniken erweitert und optimiert werden kann.
Zusätzlich zu dieser Kernuntersuchung wird auch die Wirksamkeit des Segment Anything-Modells (SAM) und des Modells dieser Arbeit für die semi-automatische Annotation der drohnenbasierten Bilder evaluiert.

Instance Segmentation

Instance Segmentation ist eine differenzierte Form der Bildsegmentierung, die darauf abzielt, individuelle Objekte in einem Bild zu identifizieren und deren Grenzen zu definieren. Diese Technik ist besonders hilfreich in Bereichen, in denen es darauf ankommt, gleichartige Objekte einzeln zu erkennen.
Anders als bei der semantischen Segmentierung, bei der alle Instanzen einer Klasse zusammengefasst werden, differenziert die Instanzsegmentierung jedes einzelne Objekt – auch wenn mehrere derselben Klasse vorliegen (siehe Abbildung 3). So wird nicht nur die Kategorie jedes Objekts bestimmt, sondern auch dessen individuelle Präsenz im Bild hervorgehoben [1].
Die Instanzsegmentierung kombiniert Elemente der Objekterkennung, wie die Detektion aller Instanzen einer Katergorie, mit der semantischen Segmentierung. Das Ergebnis besteht aus spezifischen Informationen zu jeder Instanz einer Klasse.
Zum Beispiel kann eine Instanzsegmentierung nicht nur Kategorien wie "Hund" oder "Katze" in einem Bild identifizieren, sondern auch anzeigen, wie viele Hunde oder Katzen vorhanden sind und wo genau sie sich befinden. Diese detaillierte Unterscheidung ist für Anwendungen von entscheidender Bedeutung, bei denen es auf die genaue Lokalisierung und Identifikation jedes einzelnen Objekts ankommt.

Transfer Learning

Transfer Learning optimiert die Modellentwicklung durch die Nutzung eines vortrainierten Modells für neue, spezifische Aufgaben. Dieser Ansatz nutzt bereits etablierte Erkenntnisse aus umfangreichen Datensätzen, um die Effizienz und Leistungsfähigkeit der Modelle in Bereichen zu verbessern, in denen nur limitierte Daten verfügbar sind. Durch das "Einfrieren" der frühen Schichten eines *Convolutional Neural Networks*, die allgemeine visuelle Merkmale erkennen, und das Anpassen der späteren Schichten, wird ein schnellerer Trainingsfortschritt bei gleichzeitig erhöhter Endleistung ermöglicht. Abbildung 2 veranschaulicht den *Transfer Learning* Prozess. Das *Pre-trained Model* ist im Fall dieser Arbeit YOLOv8.

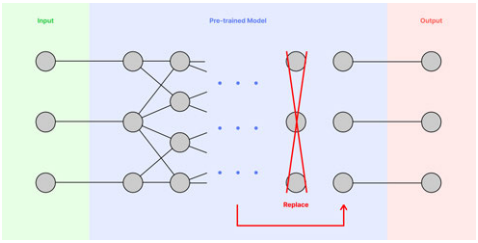


Abb. 2: Visualisierung von Transfer Learning [2]

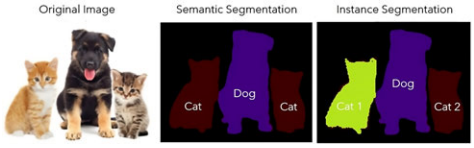


Abb. 3: Unterschied zwischen Segmentierungsarten [1]

Datensatzvorverarbeitung

Die relevanteste Methode zur Datensatzvorverarbeitung in dieser Arbeit war die Anwendung von *Tiling* auf den Datensatz. *Tiling* zielt darauf ab, hochauflösende Drohnenbilder für das Training von YOLOv8 vorzubereiten. Durch das Zerlegen der Bilder in kleinere Segmente können Speicherbeschränkungen umgangen und die Bildauflösung während des Trainings beibehalten werden, was für die Detailgenauigkeit und die Leistung des CNNs von entscheidender Bedeutung ist. Diese Technik erwies sich nicht nur als nützlich, um mit beschränktem Speicher umzugehen, sondern auch, um den Trainingsdatensatz zu erweitern. Ein visuelles Beispiel des *Tilings* ist in Abbildung 4 zu sehen. Zusätzlich zu *Tiling* um den Datensatz zu erhöhen wurden Bilder aus Google Earth gewonnen, indem die "Kamera" so angewinkelt wurde wie eine Drohne.

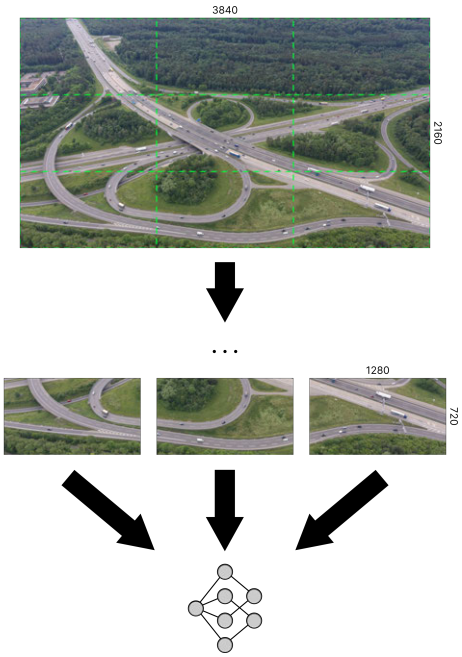


Abb. 4: Beispiel eines Tiling-Vorgangs [2]

Inferenz auf unbekannte Szenarien

Nach Abschluss des Trainingsprozesses wurde die Effektivität des Modells durch die Evaluation anhand neuer, bisher ungesehener Daten überprüft. Die Testdaten wurden aus Google Earth extrahiert. Bei der Analyse der Aufnahmen mit Top-Down-Perspektive, wie in Abbildung 5, identifizierte das Modell alle Fahrspuren präzise.
Allerdings traten bei Bildern aus angewinkelter Perspektive Herausforderungen auf. Das Modell schaffte es nur teilweise, die Fahrspuren zu erkennen und zu differenzieren. Diese Schwierigkeiten wurden in komplexeren Szenen noch verstärkt, insbesondere wenn die Bilder eine große Anzahl von Fahrspuren und zusätzliche Elemente wie Brücken und Abzweigungen zeigten.
Es wird davon ausgegangen, dass das Modell das Potenzial hat, auch in komplexeren Szenarien gute Ergebnisse zu erzielen, vorausgesetzt, es steht eine ausreichende Menge an diversifizierten Trainingsdaten zur Verfügung. Trainingsdaten, die komplexere Verkehrssituationen abbilden, könnten entscheidend sein, um die Leistung des Modells in anspruchsvollen Umgebungen zu verbessern.

Abb. 5: Inferenz auf ein Bild einer Straße [4]

Ausblick

Die aktuellen Ergebnisse der Arbeit weisen darauf hin, dass das entwickelte Modell in komplexen Szenarien noch nicht die gewünschte optimale Leistungsfähigkeit erreicht. Dennoch hat diese Arbeit das Potenzial der Instanzsegmentierung mit YOLOv8 zeigen können. Mit steigender Anzahl der Trainingsdaten konnte die Erkennungsrate des Modells kontinuierlich verbessert werden. Für zukünftige Verbesserungen wird eine Erweiterung des Trainingsdatensatzes als vielversprechender Ansatz gesehen.

Literatur und Abbildungen

[1] Hmrishav Bandyopadhyay. The Definitive Guide to Instance Segmentation. https://www.v7labs.com/blog/instance-segmentation-guide, 02 2022.

[2] Eigene Darstellung.

[3] Stefan Kaufmann. Anwendungsbeispiel aus der firmeninternen Traffic Observation Software. https://it-designers-gruppe.de/, 12 2023.

[4] Google LLC. Auszug von Satellitenbildern aus Google Earth. https://www.google.com/intl/de/earth/about/, 04 2021.

Entwicklung einer Cloud-basierten IoT-Anwendung für Klimasysteme

Alexander Boettger **Jürgen Koch**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma Eberspächer Gruppe GmbH & Co. KG, Esslingen am Neckar

Abstract:

In dieser Studie wird die Entwicklung einer Cloud-basierten Internet-of-Things (IoT)-Anwendung vorgestellt, die unter Verwendung des Arduino Portenta H7 Lite Connected entwickelt wird, um Daten von Fahrzeugklimaanlagen zu erfassen und zu analysieren. Das Projekt zielt darauf ab, Störungen in den Klimasystemen frühzeitig zu erkennen, indem die Sensordaten in Echtzeit in AWS (Amazon Web Services) verarbeitet werden.

Einleitung:

Die effektive Überwachung und präventive Wartung von Fahrzeugklimaanlagen spielt eine wesentliche Rolle, sowohl für die Sicherstellung eines angenehmen Fahrerlebnisses als auch für die Gewährleistung der Betriebssicherheit von Fahrzeugen. Diese Arbeit konzentriert sich auf die Entwicklung einer IoT-basierten Lösung, die durch präzise Datenerfassung und -analyse eine proaktive Wartung ermöglicht.

Methodik:

Das Herzstück des Projekts bildet der Arduino Portenta H7 Lite Connected, ein leistungsstarker IoT-Mikrocontroller, der für die Datenerfassung von verschiedenen Sensoren in den Klimasystemen der Fahrzeuge eingesetzt wird. Die erfassten Daten werden dann zur Analyse und Verarbeitung in AWS übertragen.

Abb. 1: Portenta H7 Lite Connected [1]

Vorgehensweise:

Derzeit wird an der Implementierung der Sensoren und der Einrichtung der Kommunikation zwischen dem Arduino-Board und AWS gearbeitet. Das Ziel ist es, eine nahtlose und sichere Datenübertragung zu gewährleisten und eine effektive Analyseplattform in AWS zu entwickeln.

Zwischenergebnisse:

Bisherige Fortschritte umfassen die erfolgreiche Integration des Arduino Portenta H7 Lite Connected mit den Sensoren sowie erste erfolgreiche Testläufe zur Datenübertragung in die AWS-Cloud. Diese vorläufigen Ergebnisse bestätigen das Potenzial des Systems für die angestrebten Anwendungen.

Ausblick und erwartete Ergebnisse:

Die vollständige Implementierung der Anwendung wird voraussichtlich eine verbesserte Diagnose und Wartung von Fahrzeugklimaanlagen ermöglichen. Durch die Nutzung von Cloud-Computing und fortschrittlichen Analysemethoden in AWS soll eine frühzeitige Erkennung und Behebung von Störungen realisiert werden.

Fazit:

Das Projekt zeigt vielversprechende Ansätze zur Nutzung moderner IoT- und Cloud-Technologien für die Fahrzeugwartung. Die Integration des Arduino Portenta H7 Lite Connected mit AWS bietet ein innovatives Beispiel für die Anwendung von IoT in der Automobilindustrie.

Literatur und Abbildungen

[1] Arduino SA. Portenta H7 Lite Connected. https://store.arduino.cc/products/portenta-h7-lite-connected, 12 2023.

Self-Supervised Learning of Depth and Pose with Multiple Cameras

Tobias Brandl **Markus Enzweiler**

Department of Computer Science and Engineering, Esslingen University

Work carried out at Department of Computer Science and Engineering, Esslingen

Motivation

Purely vision-based navigation of mobile robots in outside environments is a non-trivial problem. Robots do not perceive depth, and can not locate themselves the same way humans do. For humans, it is easy to tell whether a balloon is in front of a person or behind them, and easy to remember which path we took in order to navigate our surroundings. We learned to estimate depth with our eyes, are able to tell the geometry of the scene we are looking at and know where we are located in the scene we perceive. For robots, this task is more difficult.

The field of Computer Science, which focuses on enabling machines to interpret and understand visual information from the world, is called Computer Vision (CV). CV methods allow us to estimate depth from color images captured by a camera. This allows us to estimate distances to objects in the image we have taken.

The movement of a robot can be estimated using Visual Odometry (VO) methods with sequential image sequences, from videos for example. Knowing depth and movement of the robot, we can virtually recreate the 3D environment we perceive and estimate the robots location by using Simultaneous Localization and Mapping (SLAM) methods.

Utilizing modern Machine Learning methods, we can further improve traditional VO and learn deep features from the entire image. This allows us to more accurately and robustly estimate the motion of the robot, which means the robot performs better under the various challenging conditions of outside environments, like varying lighting, changing weather conditions, complex, dynamic and textureless scenes.

One drawback of Deep Learning methods is the need for costly and time-consuming labeled data, to train in supervised manner. Fortunately, in the case of Monocular Depth Estimation, there are ways to fully self-supervise learning without the need for labeled data by constraining the model to effectively learn scene geometry. But this is not the only way to improve accuracy and robustness of VO and SLAM methods. Currently, most setups use only one camera, or a stereo pair of front-facing cameras, which makes the mobile robot prone to errors. The Field of View (FoV) is limited, and sudden lighting changes, like Lens Flare, or noisy environments can lead to inaccurate perception. Research suggests using multiple cameras at different positions around the robot is enhancing accuracy, and increasing robustness [2]. This is where this thesis expands upon.

Visual Odometry and vSLAM basics

VO is the process of estimating the ego-motion of an agent (e.g., vehicle, human, and robot) using only the input of a single or multiple cameras attached to it [3]. The VO workflow is as follows: First, a sequence of images needs to be recorded. It is important to know the intrinsic and extrinsic camera parameters, so we can later transform pixels or key points between different coordinate systems and correctly estimate the motion of the robot. Parameters like focal length, field of view and lens distortion are intrinsic. Parameters like location and orientation of the camera are extrinsic. Intrinsic and extrinsic parameter calibration is performed once during camera setup.

Then, features need to be detected, tracked and the motion estimated. The three approaches for feature extraction are the following: sparse, semi-dense and dense.

Sparse methods only extract a limited number of distinctive points in the image, like corners or blobs.

Semi-dense methods extract raw pixels with high intensity gradients, like edges of buildings.

Dense methods extract every pixel in the image in order to get a dense representation of feature points. The Depth Net used in this thesis is a dense feature extraction approach.

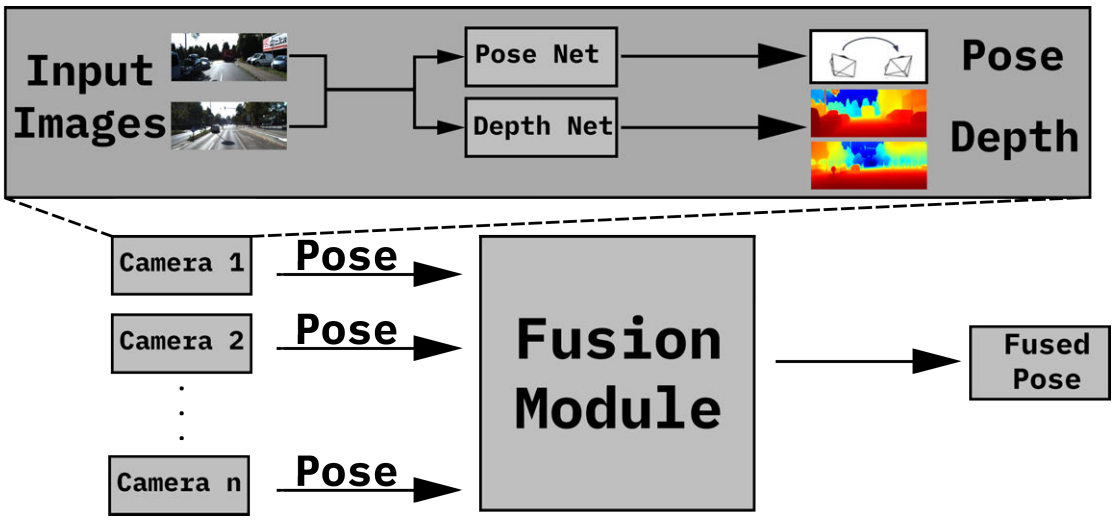


Fig. 1: Sketch of the proposed architecture for the Multi-Camera Visual Odometry system. The RGB images are from the KITTI dataset and depth images are from [4]

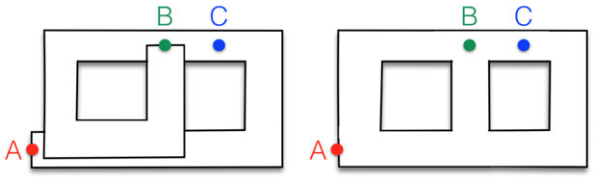


Fig. 2: Visualization of environment maps created by Visual Odometry (left) and visual SLAM (right). VO is locally, vSLAM globally consistent. From [1]

The difference between VO and vSLAM is visualized in Figure 2. The VO approach builds its environment map incrementally. It provides very accurate local motion estimations of the robot location relative to the scene currently perceived but does not know whether the scene has been seen before or not. On the other hand, vSLAM keeps track of previous scene informations and utilizes those to build a globally consistent environment map, relative to which it knows the robots location within. If previously detected scene information is detected again, loop closure can be executed.This allows us to refine the scene information within the loop, as well as the trajectory within the environment to remove drift and other errors, which are accumulating over time.

Objective

This thesis aims to utilize any amount of arbitrarily mounted, but calibrated, cameras on a mobile robot in order to locate and track its motion as good as possible in the challenging conditions of outside environments. A framework will be developed to achieve this goal. Figure 1 shows the planned architecture. First, we jointly train two deep learning models in self-supervised manner. One learns the depth map of the images. The other learns the relative pose (location and rotation) for each camera. Then, a fusion module receives the learned relative camera poses as input and refines the absolute pose of the robot. Different approaches for this fusion module will be evaluated and compared on appropriate datasets, featuring both static and dynamic scenes.

Method

We generate pseudo-depth maps with a pre-trained state-of-the-art supervised monocular depth estimation network before training our deep learning models. These pseudo-depth maps are used as prior during the training process. Both, Pose and Depth Net are convolutional neural networks. We jointly train them with self-supervision using consecutive image pairs as input. The depth network refines the previously generated pseudo-depth map and outputs its depth map prediction for each of the two images. The pose network estimates the relative camera pose. With the predicted depth map and estimated camera pose we can generate the warping flow and reconstruct the input image pair. These reconstructed images are our supervisory signal for the self-supervised learning approach.

The fusion modules task is to combine all relative camera poses to a more refined and accurate pose of the mobile robot. A naive arithmetic mean or median approach will form the baseline for the module. Two learning based approaches are currently being considered for evaluation. One of them utilizes multi-layer perceptrons and a recurrent neural network, the other a transformer.

References and figures

[1] Cesar Cadena, Luca Carlone, Henry Carrillo, Yasir Latif, Davide Scaramuzza, José Neira, Ian Reid, and John J. Leonard. Past, Present, and Future of Simultaneous Localization and Mapping: Toward the Robust-Perception Age. *IEEE Transactions on Robotics*, 32:1309–1332, 2016.

[2] Pushyami Kaveti, Shankara Narayanan Vaidyanathan, Arvind Thamil Chelvan, and Hanumant Singh. Design and Evaluation of a Generic Visual SLAM Framework for Multi Camera Systems. *IEEE Robotics and Automation Letters*, 8:7368–7375, 2023.

[3] Davide Scaramuzza and Friedrich Fraundorfer. Visual Odometry [Tutorial]. *IEEE Robotics & Automation Magazine*, 18:80–92, 2011.

[4] Libo Sun, Jia-Wang Bian, Huangying Zhan, Wei Yin, Ian Reid, and Chunhua Shen. SC-DepthV3: Robust Self-supervised Monocular Depth Estimation for Dynamic Scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2023.

Validierung der Messung der Pulsfrequenz mit einem mobilen Endgerät

Dennis Buehl **Jörg Nitzsche**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma SYSTECS Informationssysteme GmbH, Leinfelden-Echterdingen

Einleitung

Die Herzfrequenz ist ein wichtiger - wenn nicht der Wichtigste - Indikator für die Gesundheit und Fitness eines Menschen.
Dass sich die digitale Erfassung und Auswertung von Gesundheitsdaten einer großen Beliebtheit, nicht nur unter Sportlern, erfreut zeigt auch eine aktuelle Umfrage des Branchenverbands Bitkom. 69% der Befragten geben an, mindestens eine App zur Erfassung der Gesundheits- oder Fitnessdaten auf ihrem Smartphone installiert zu haben. Gut ein Fünftel davon dient ausschließlich zur Erfassung von Körperdaten, wie der Herzfrequenz. [3]
Ein weiterer Techniktrend der letzten Jahre dürfte eine große Rolle spielen – Wearables. Am weitesten verbreitet sind in diesem Kontext sicherlich die Smartwatches und Fitnessuhren, die man an immer mehr Handgelenken beobachten kann. Sie messen kontinuierlich die Herzfrequenz und präsentieren diese in diversen Statistiken ihren Trägern.

Problemstellung und Zielsetzung

Diese Arbeit beschäftigt sich mit der Frage, wie genau die Messung der Pulsfrequenz sein kann, wenn diese nur mit den Sensoren eines Smartphones durchgeführt wird.
Für diese Auswertung wird eine mobile Anwendung entwickelt mit deren Hilfe ein Smartphone-Nutzer die Pulsfrequenz ermitteln kann. Dazu werden die Kamera und der Flash des Smartphones verwendet. Durch Auflegen des Fingers auf beide Sensoren kann ein Video erstellt werden, aus dem sich die Herzfrequenz ablesen lässt.
Zum Einsatz kommt dabei die gleiche Technik, wie sie auch von Wearables und teilweise auch von medizinischen Messmitteln verwendet wird. Dabei handelt es sich um die sogenannte Photoplethysmographie (PPG), auf die im folgenden Absatz eingegangen wird.
Die Ergebnisse der Messung mittels der Smartphone-App werden in verschiedenen Szenarien unter Vergleich mit anderen Messverfahren verglichen und validiert. Bei den Szenarien handelt es sich um Messungen in verschiedenen körperlichen Belastungsbereichen des Anwenders.

Photoplethysmographie

Bei jedem Herzschlag wird Blut durch den Körper transportiert, wobei sich die Blutgefäße ausweiten. Dieser Effekt wird bei der PPG genutzt.
Die Haut wird nun durch eine Lichtquelle beleuchtet und durch einen Sensor wird das Reflektierte Licht erfasst und ausgewertet. Blutgefüllte Gefäße absorbieren Licht anders als leere Gefäße, dadurch entstehen bei der Detektion des reflektierten Lichts eine Kontraständerung. Stellt man diese graphisch dar, erkannt man das Pulssignal der Herzschläge. [2]

Ablauf der Messung

Zur Erfassung der Messdaten wird eine Videosequenz über einen bestimmten Zeitraum aufgezeichnet. Während der gesamten Aufnahme bedeckt der Anwender die Hauptkamera und den Flash seines Smartphones. Mit bloßem Auge lassen sich auf dem Videosignal Änderungen der Helligkeit wahrnehmen. Bei jedem Herzschlag lässt sich feststellen, dass das Bild partiell dunkler wird. Daher wird der Helligkeitsparameter für die Analyse des Videos betrachtet.
Für die Analyse wird das Video in seine Einzelbilder (Frames) zerlegt. Durch die vollständige Bedeckung der Kamera und den aktivierten Flash ist das Video komplett in hellen Rottönen gehalten.
Abb. 1 zeigt einen Vergleich zweier Einzelbilder aus einer Videosequenz. Auf dem linken Bild sieht man, dass das Bild in einheitlich helleren Rottönen erscheint. Das Rechte Bild wurde während eines Herzschlags aufgezeichnet. Es weist eine dunkle Vignettierung auf.

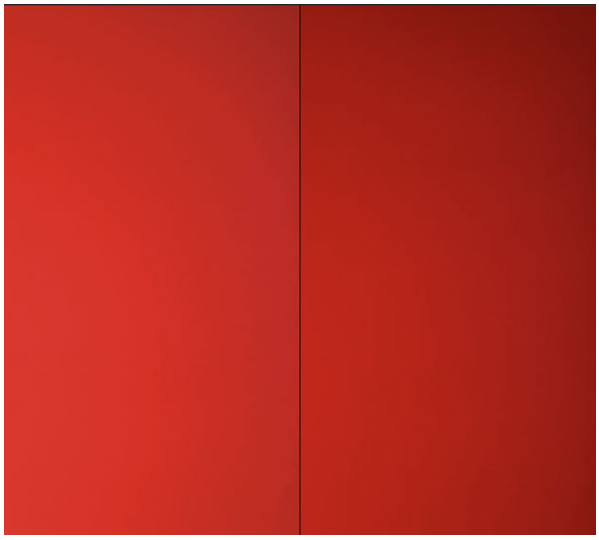


Abb. 1: Vergleich eines hellen (links) und eines dunklen (rechts) Einzelbildes [1]

Die Einzelbilder werden nun alle auf ihre durchschnittliche Helligkeit untersucht, indem das Bild in seine Pixel, zerlegt wird und die Helligkeitswerte ermittelt werden. Diese liegen zwischen 0...255 (Schwarz...Weiß).
Aus den Helligkeitswerten über die Zeit ergibt sich nun die Messreihe. Bei einer Bildrate von 30 Bildern pro Sekunde und einem 30 sekundieren Video ergeben sich 900 Messwerte. Stellt man diese in einem Diagramm dar, lässt sich ein periodisches Signal erkennen.
Jedes periodische Signal lässt sich in Sinus- und Cosinus-Signale mit unterschiedlichen Frequenzen und Amplituden zerlegen. Mit Hilfe einer Fourier-Transformationen wird das Signal in seine Bestandteile zerlegt. [4]
Damit lässt sich bestimmen, welche Frequenz am dominantesten auftritt. Bei dieser handelt es sich dann in diesem Fall um die Herzfrequenz.

Konzept

Für die Messung der Herzfrequenz kommen zwei Komponenten zum Einsatz. Eine App, die auf dem mobilen Endgerät läuft und für die Erfassung der Messdaten, sowie der Präsentation der Ergebnisse dient und einer Anwendung, die die Messdaten auswertet.

Zur Entwicklung der Smartphone-App wird das plattformübergreifende .NET MAUI (Multi-Plattform App UI) Framework genutzt. Das Framework bietet die Möglichkeit auf die Sensoren der Endgeräte zuzugreifen. So lassen sich unter anderem Videos aufzeichnen, ohne dass sich die Kameraanwendung des Smartphones öffnet. Der Anwender soll dabei weiterhin die UI der Anwendung vor sich sehen und keine klassische Kamera View.
Die App wird primär für iOS-Geräte entwickelt und auch für die Validierung wird ein iOS-Gerät verwendet. Für die Auswertung wird eine Python Anwendung entwickelt. Python eignet sich gut für die Verarbeitung von Daten und bietet einige umfassende Bibliotheken hierfür.
Die Anwendung zerlegt das Video in seine Frames, ermittelt deren Helligkeitswerte und wertet diese mit Hilfe einer Fourier-Transformation aus.
Die Anwendung persistiert die Videodateien nicht, da diese nach der Messung nicht mehr benötigt werden. Jede Messung stellt eine Momentaufnahme dar.
Da die Auswertung des Videos einige Sekunden in Anspruch nehmen kann, kommen zwei verschiedene Schnittstellen zum Einsatz. Zum einen eine REST API, um das Video bereitzustellen und zum anderen WebSockets. Da die Auswertung eine gewisse Zeit in Anspruch nimmt, wird eine WebSocket-Schnittstelle implementiert, die es der Python Applikation ermöglicht das Ergebnis, sobald es vorliegt an den MAUI Client weiterzureichen, ohne dass dieser mittels Polling danach fragen muss.
Abb. 2 zeigt ein Sequenzdiagramm, das den Ablauf der Messung darstellt. Nachdem der Anwender die Messung gestartet hat, zeichnet die App für 30 Sekunden das Video auf und sendet es über den REST-Endpunkt an den Python Service. Dort wird das Video verarbeitet und ausgewertet.
Nach der Auswertung wird das Ergebnis über die WebSocket-Verbindung an die Client App übergeben. Ist das Ergebnis unplausibel oder kann das Video nicht ausgewertet werden, wird eine entsprechende Nachricht an die App gesendet.
Die App präsentiert dem Anwender im Erfolgsfall das Ergebnis. War die Auswertung nicht erfolgreich fordert er den Anwender auf eine neue Messung zu starten.

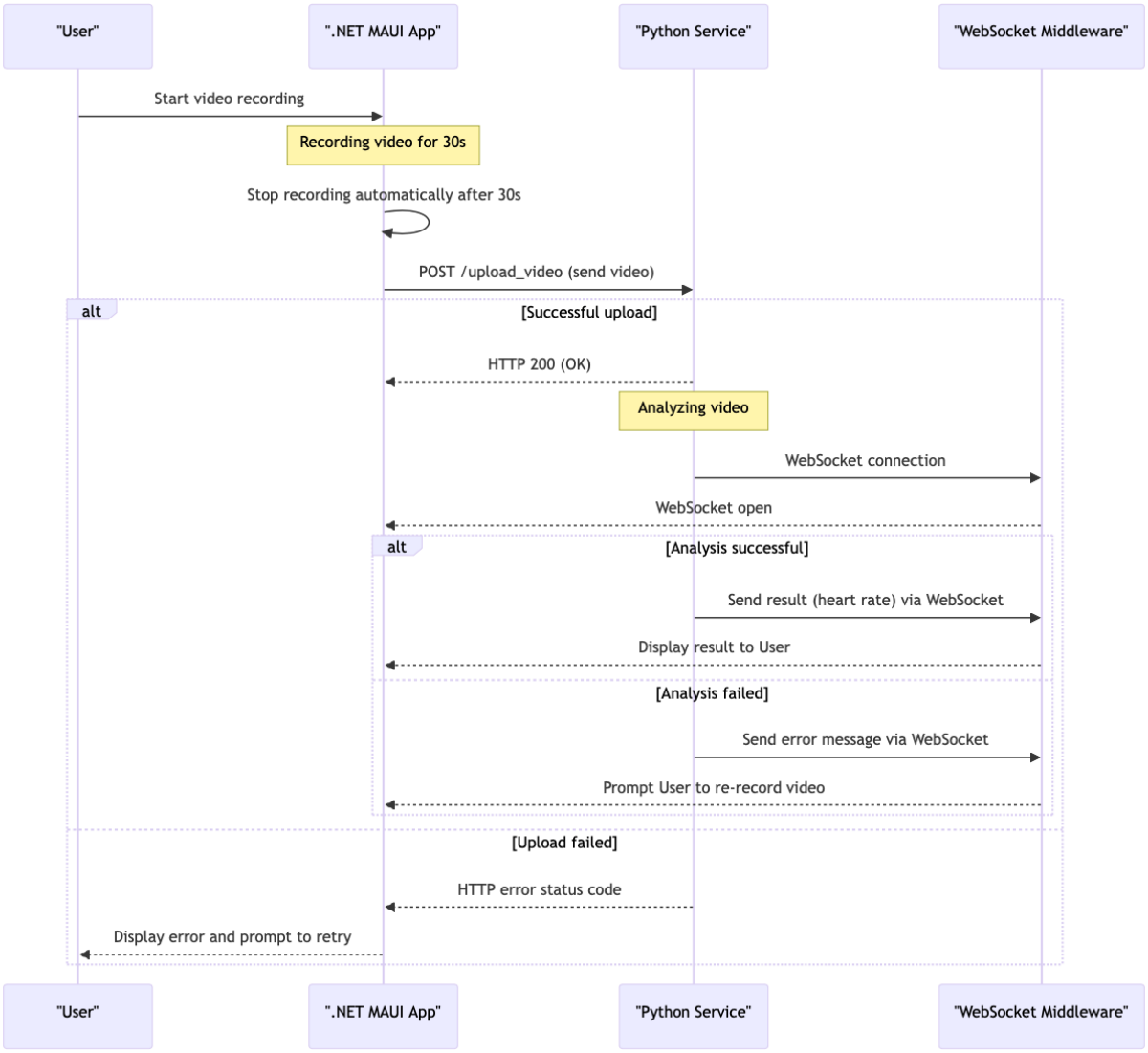


Abb. 2: Sequenzdiagramm Ablauf der Messung [1]

Erste Erkenntnisse

Im bisherigen Verlauf ist unter anderem ein Programm zur Auswertung entstanden mit dem sich die Pulsfrequenz anhand von Videodateien auswerten lässt.
Abb. 3 zeigt ein Diagramm, dass die Messwerte des Videos über die Zeit darstellen. Es ist klar zu erkennen, dass zu Beginn ein starker Ausschlag zu erkennen ist. Dieses Phänomen tritt bei allen getesteten Videodateien auf und ist auf die Aktivierung des Flashs mit dem Start des Videos zurückzuführen. Daher werden die ersten Sekunden des Videos nicht für die Analyse berücksichtigt.
Das Diagramm bildet auf der Y-Achse die Helligkeitswerte in invertierter Form ab, die sich wie eingangs erwähnt zwischen 0...255 bewegen. Um bei der Fourier-Transformation lineare Bestandteile unberücksichtigt zu lassen wurde dieser Anteil bereits abgezogen, somit bewegt sich das Signal in dem Diagramm um die Nulllinie. Das Signal wurde für diese Darstellung invertiert, um die charakteristische Form des Herzschlags besser erkennen zu können. Ebenso wird hier nur das Signal der ersten 15 Sekunden dargestellt um die Lesbarkeit zu fördern.

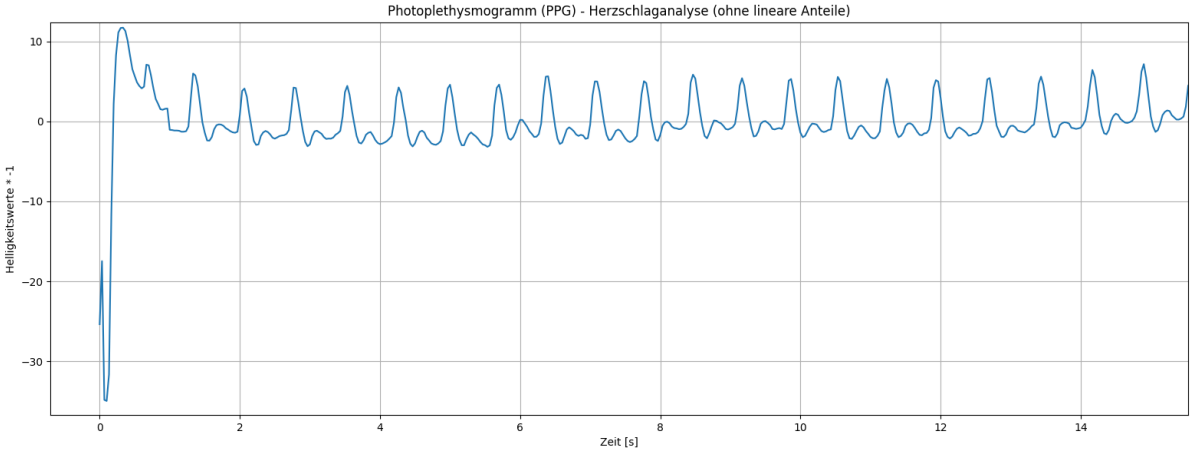


Abb. 3: Photoplethysmogramm über die ersten 15 Sekunden der Messung [1]

Das Amplitudenspektrum in Abb. 4 stellt das Ergebnis der Fourier-Transformation des Signals dar. Zu sehen ist der Ausschlag bei 1,42 Hz, was einer Herzfrequenz von 85 Schlägen pro Minute entspricht. Vergleicht man dieses Ergebnis mit dem manuellen Ablesen des Signals aus Abb. 3 lässt sich bestätigen, dass es sich hierbei um die gleiche Frequenz handelt. Der zweite größere Ausschlag ist doppelt so groß wie der, der Herzfrequenz. Vermuten lässt sich, dass dieser mit den kleineren Ausschlägen des Signals zu begründen ist. Eine genauere Analyse des Ergebnisses wird im Laufe dieser Arbeit erfolgen.
Eine weitere Erkenntnis, die im Zusammenhang mit dem Testen der Funktionen zur Auswertung des Videos erlangt wurde, ist die, dass die Auswertung bei zu großen oder zu kleinen Druck des Fingers auf die Kameralinse scheitern kann.

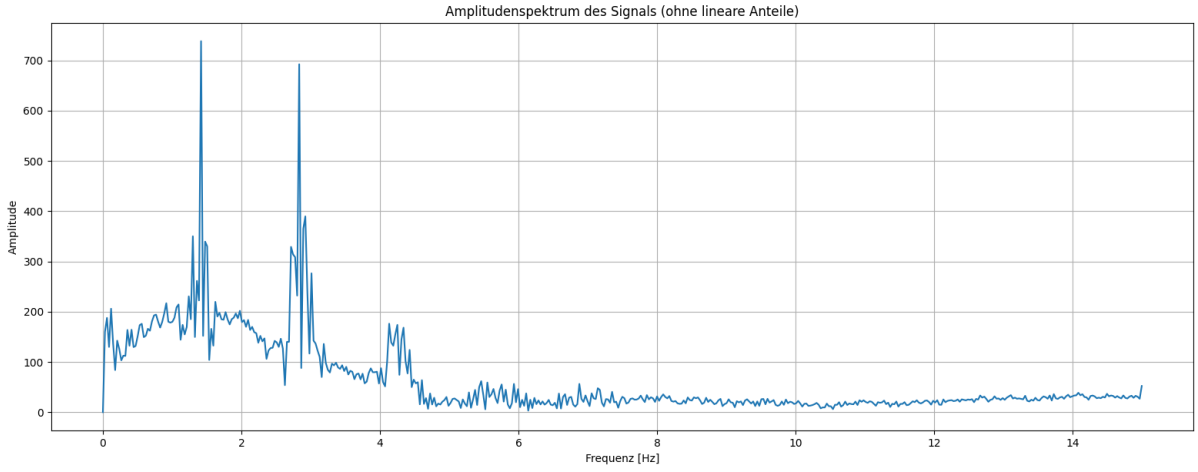


Abb. 4: Amplitudenspektrum der Fourier-Transormation des Photoplethysmogramm [1]

Ausblick

Der bisherige Eindruck ist, dass es möglich ist, die Pulsfrequenz mittels Smartphone-App zu ermitteln. Im weiteren Verlauf werden die bisher getesteten Algorithmen und Funktionen noch in die beschriebene architektonische Form gebraucht und optimiert. Ebenso werden die App und Schnittstellen final implementiert.
Als wichtiger Schritt folgt die Analyse der Auswertungen und die Validierung der Messungen durch eingangs erwähnt Testverfahren.

Literatur und Abbildungen

[1] Eigene Darstellung.

[2] Gaetano D. Gargiulo and Ganesh R. Naik. *Wearable/Personal Monitoring Devices Present to Future*. Springer Singapore, 2022.

[3] Bettina Lange et al. Zwei Drittel nutzen Fitness- und Gesundheits-Apps auf ihrem Smartphone. https://bitkom-research.de/news/zwei-drittel-nutzen-fitness-und-gesundheits-apps-auf-ihrem-smartphone, 11 2023.

[4] Lutz Priese. *Computer Vision Einführung in die Verarbeitung und Analyse digitaler Bilder*. Springer Vieweg, 2015.

Entwicklung eines Diagnosetools zur Anomalie-Erkennung bei Vorschubtests einer Werkzeugmaschine

Lisa Caminati **Markus Enzweiler**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma EMAG Maschinenfabrik GmbH, Salach

Motivation

Die Wettbewerbssituation im produzierenden Gewerbe ist durch einen hohen Preisdruck und einer geforderten Termintreue charakterisiert. Dadurch ist es notwendig, Produkte in kürzester Zeit mit hoher Qualität und gleichzeitig geringen Kosten am Markt zu platzieren. Die Anlageninstandhaltung bildet dabei den Hauptteil der Gesamtbetriebskosten [2].
Die vorrausschauende Wartung ermöglicht es Firmen ungeplante Ausfallzeiten zu vermeiden, dadurch ihre Produktivität zu steigern und somit Kosten zu senken [2]. Aufgrund zunehmender Automatisierungen wird der Service zusätzlich unterstützt und kann mithilfe von Ferndiagnosen effizienter gestaltet werden.

Health Check

Die Firma EMAG Maschinenfabrik GmbH bietet seinen Kunden mit dem Produkt Health Check ein Frühwarnsystem. Dieses berechnet eine Gesundheitsbewertung der Maschinen und bildet Erfahrungswissen der Experten in der Analyse-Software ab. Mithilfe der Informationen zum Gesundheitstrend der Maschine können Wartungszeiträume besser geplant werden und somit Kosten gesenkt werden [3].
Dabei werden mit Hilfe von 3D-Beschleunigungssensoren, wie in folgender Abbildung dargestellt, Schwingungen aufgezeichnet und eine Gesundheitsbewertung der Maschine berechnet. Auf Wunsch des Kunden kann die Messung zusätzlich von einem Experten manuell bewertet werden [3].

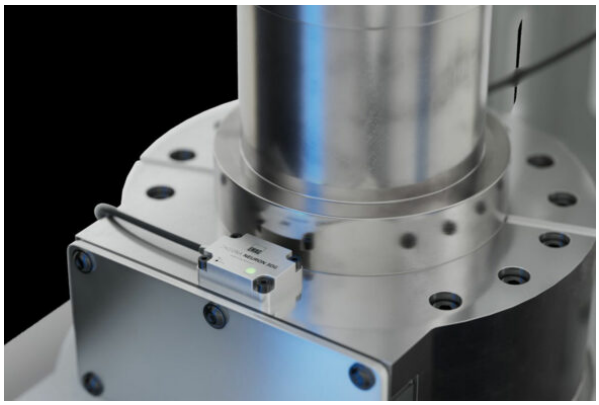


Abb. 1: Mithilfe von 3D-Beschleunigungssensoren werden Schwingungen aufgezeichnet [3]

Ziel der Arbeit

Zur Unterstützung des Experten und der Service-Mitarbeiter soll ein Diagnosetool entwickelt werden, das mit Hilfe der regelmäßig aufgezeichneten Messungen Schäden an der Maschine erkennt. Dabei soll ein defekter Gewichtsausgleich automatisch erkannt werden, außerdem soll das Diagnosetool zur Erkennung weiterer häufig vorkommenden Schäden erweiterbar sein.

Gewichtsausgleich

Bei schweren vertikalen Achsen muss der Einfluss der Gewichtskraft berücksichtigt werden, um beispielsweise die geforderten Beschleunigungswerte einhalten zu können. Zur Entlastung des Motors und zur erhöhten Lebensdauer der Führungen und des Kegelgewindetriebes kann man die Masse der vertikal bewegten Baugruppe durch einen Gewichtsausgleich kompensieren. Dieser kann auf einem mechanischen, hydraulischen oder pneumatischen Wirkprinzip basieren [4].

Wie in der folgenden Abbildung dargestellt, wird bei einem hydraulischen Gewichtsausgleich parallel zum Antriebssystem ein Hydraulikzylinder mit Druckspeicher betrieben. Dadurch ist die Kompensationskraft einstallbar und bleibt im Not-Aus erhalten [4].

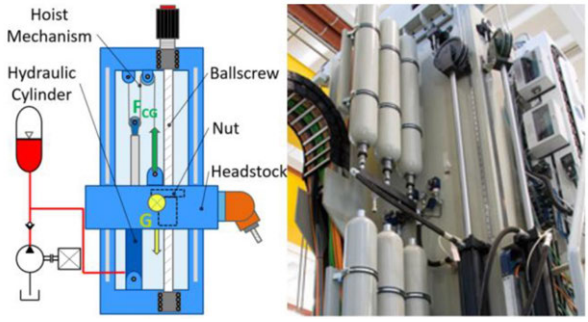


Abb. 2: Darstellung eines hydraulischen Gewichtsausgelichs [1]

Data Preprocessing

Auffällige historische Messfahrten wurden gesammelt und aufbereitet. Anschließend wurden diese von Experten gesichtet und zugeordnet, ob bei den jeweiligen Messungen der Gewichtsausgleich einen Defekt aufweist oder nicht. Zusätzlich wurden zu denselben Maschinentypen unauffällige Messungen gesammelt und gelabelt.

Insgesamt liegen Daten von 19 verschiedenen Maschinentypen vor, wobei der Gewichtsausgleich bei 28 Maschinentypen eingebaut wird.
Aus den Messungen wurden Merkmale extrahiert, welche die Fehlermuster des Gewichtsausgleichs repräsentieren. Diese wurden in Features zusammengefasst und dienen als Grundlage der Klassifizierung.

Herausforderungen

Aufgrund der Bewertung durch einen Experten können menschliche Fehler auftreten und dadurch falsche Labels vergeben werden.
Bei Auftritt von mehreren Maschinendefekten überlagern sich deren Schwingungsmuster, was die eindeutige Zuordnung zu einem bestimmten Fehler erschwert.
Zusätzlich ist bei einem Gewichtsausgleich der Übergang von voll funktionsfähig zu defekt fließend, was eine klare Einteilung in eine der beiden Kategorien beeinträchtigt.
Diese Eigenschaften führen zu einem nicht vollständig heterogenen Labelling der Messungen, was einen signifikanten Einfluss auf das Training der Klassifizierungs-Modelle hat.

Ausblick

Im weiteren Verlauf der Arbeit werden verschiedene Modelle zur Klassifizierung ausgewählt, mithilfe der aufbereiteten Daten trainiert und validiert.
Anschließend wird in die vorhandene Codebasis des Health Checks das Diagnosetool mit dem ausgewählten Modell integriert.

Literatur und Abbildungen

[1] S. Fiala, A. Bubak, and L. Novotny. Control of hybrid electric-hydraulic drive for vertical feed axes of machine tools. *MM Science Journal*, 2019.

[2] Achim Kampker, Kai Kreisköther, Max Kleine Büning, Tom Möller, and Max Busch. Vorausschauende Instandhaltung durch Maschinelles Lernen in der Prozessindustrie. *Zeitschrift für wirtschaftlichen Fabrikbetrieb*, pages 195–198, 2018.

[3] Mathias Klein. Health Check: Höhere Verfügbarkeit durch vorrausschauende Wartung. https://www.emag.com/de/produkte-services/digitalisierung/edna-health-check/, 2023.

[4] Joachim Regel. *Werkzeugmaschinen und Vorrichtungen: Anforderungen, Auslegung, Ausführungsbeispiele*. Springer Vieweg Wiesbaden, 2022.

Untersuchung von klassischen Anomalieerkennungsmethoden anhand von multivariaten Zeitreihendaten aus der Antriebsstrangentwicklung

Halime Nur Cengiz **Steffen Schober**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma Mercedes-Benz AG, Untertürkheim

Einleitung

In der Industrie werden täglich umfangreiche Datenmengen erfasst und genutzt, um Prozesse zu evaluieren und zu verbessern. Die frühzeitige Erkennung ungewöhnlicher Muster und fehlerhafter Daten stellt daher eine Herausforderung dar, um potenzielle Risiken und Probleme zeitnah zu identifizieren und entsprechende Maßnahmen zu ergreifen. Das Hauptziel dieser Arbeit besteht darin, diese anomalen Daten mithilfe geeigneter Methoden bestmöglich zu erkennen. Anomalien werden hier als Abweichungen in den Daten bezeichnet, die in der Regel seltener auftreten als normale Beobachtungen und sich durch signifikante Merkmale vom Normalverhalten unterscheiden.

Problemstellung

Es gibt eine Vielzahl von Anomalieerkennungsmethoden, die genau diese Herausforderung in multivariaten Zeitreihen angehen. Klassischerweise werden dafür statistische Methoden oder Modelle aus dem Bereich des maschinellen Lernens verwendet. Insbesondere in den letzten Jahrzehnten wurden vermehrt neue Ansätze im Bereich des Deep Learning entwickelt. Mit der zunehmenden Popularität dieser Deep Learning Modelle wächst jedoch auch die Kritik. Audibert et al. untersuchen in einer Studie die Anomalieerkennungsleistung von sechzehn klassischen, maschinelles Lernen basierenden und Deep Learning Ansätzen anhand von fünf realen öffentlichen Datensätzen [1]. Durch den Vergleich der Leistung jeder der sechzehn Methoden zeigen sie auf, dass keine Methode die andere übertrifft [1]. In dieser Arbeit werden daher klassische Anomalieerkennungsmethoden untersucht und mit dem Deep Learning Ansatz MA-VAE [2] verglichen, um zu prüfen, ob klassische Modelle auch mit den Daten aus der Entwicklungsprozessen in der Industrie gute Leistungen erbringen.

Daten

Die erfassten Signale dienen dazu, das dynamische Verhalten eines industriellen Systems aufzuzeichnen. Anschließend werden die erfassten Daten zu einer multivariaten Zeitreihe zusammengeführt. Die Anomalieerkennung in diesen mehrdimensionalen Daten hat im Vergleich zu univariaten Daten besondere Herausforderungen, da die verwendeten Methoden in der Lage sein müssen, komplexe Muster und Zusammenhänge innerhalb der Dimensionen zu erfassen. Um die Konzepte besser zu veranschaulichen, wird in Abbildung 1 ein beispielhafter Plot einer multivariaten Zeitreihe gezeigt. Dieser Plot basiert auf einem Datensatz zum Thema Haarausfall und dient lediglich als exemplarisches Beispiel.

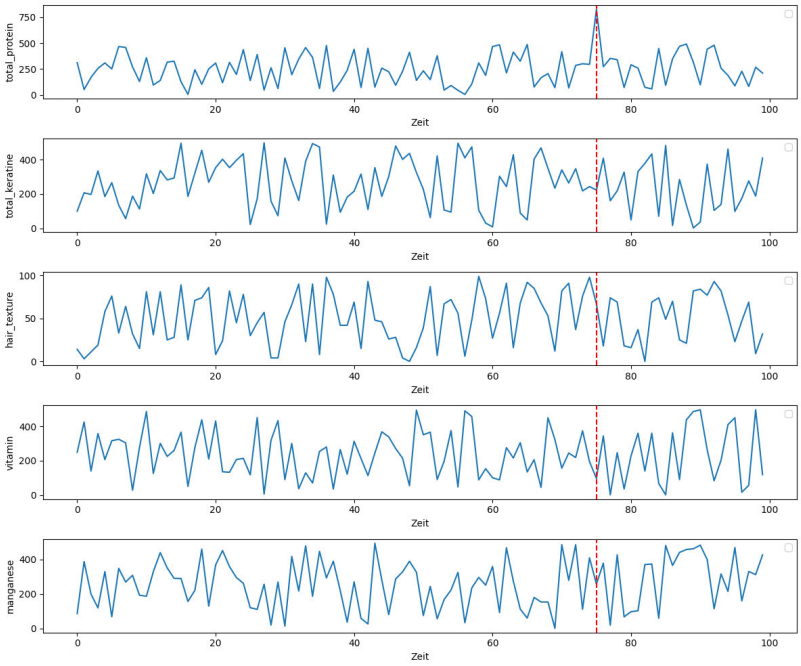


Abb. 1: Plot der ersten 5 Kanäle multivariater Zeitreihen-Daten zu Haarausfall aus einem öffentlichen Datensatz. Eine Anomalie wurde mittels des Isolation Forest-Algorithmus identifiziert und durch die gestrichelte rote Linie markiert. [3]

Die meisten Studien testen auf beliebte Benchmark-Datensätzen, erstellt von Organisation wie Yahoo, Numenta und der NASA. Die Kernaussage der Studie von Wu und Keogh [4] besteht darin, dass die Mehrheit dieser Datensätzen an einem oder mehreren der vier Schwachstellen leidet. Zu diesen Schwachstellen gehören unter anderem unrealistische Verteilung von Anomalien in den Daten, mögliche falsche Labels sowie Trivialitäten [4]. Aufgrund dieser Mängel argumentieren die Autoren, dass viele Veröffentlichungen von Anomalieerkennungsalgorithmen möglicherweise nicht zuverlässig sind [4], und vor allem könnte der scheinbare Fortschritt der letzten Jahre illusorisch sein [4]. Die für die Arbeit verwendeten Trainingsdatensätze sind nicht gelabelt, daher ist unbekannt, ob sie ausschließlich normale Beobachtungen enthalten. Es wird jedoch angenommen, dass der Anteil anomaler Beobachtungen signifikant kleiner ist. Zur Evaluation der Modelle werden fünf Datensätze mit Anomalien und ein Datensatz mit ausschließlich normalen Daten verwendet.

Modelle

In dieser Arbeit werden sechs Modelle für die Anomalieerkennung analysiert. Diese zu untersuchenden Modelle werden ebenfalls in der Studie von Audibert et al. verwendet [1]. Die ausgewählten Modelle umfassen MEWMA, VAR, TSADIS, Isolation Forest, Local Outlier Factor und One-Class Support Vector Machine. Jedes dieser Modelle bietet eine unterschiedliche Herangehensweise zur Erkennung von Anomalien in multivariaten Zeitreihen.

Ausblick

Durch die Erkenntnisse dieser Abschlussarbeit soll analysiert werden, ob klassische Methoden für die Daten wie die aus realen industriellen Systemen geeignet sind und wie sie im Vergleich zur Deep Learning Methode MA-VAE abschneiden. Dabei ist es wichtig, vorab die grundlegenden Prinzipien der Methoden zu untersuchen und die entsprechenden Hyperparameter bestmöglich zu optimieren. Im Anschluss können Überlegungen angestellt werden, wie mit diesen Ergebnissen weiter verfahren werden soll.

Literatur und Abbildungen

[1] Julien Audibert, Pietro Michiardi, Frédéric Guyard, Sébastien Marti, and Maria A. Zuluaga. Do deep neural networks contribute to multivariate time series anomaly detection? *Pattern Recognition*, 2022.

[2] Lucas Correia, Jan-Christoph Goos, Philipp Klein, Thomas Bäck, and Anna V. Kononova. MA-VAE: Multi-head Attention-based Variational Autoencoder Approach for Anomaly Detection in Multivariate Time-series Applied to Automotive Endurance Powertrain Testing. *In Proceedings of the 15th International Joint Conference on Computational Intelligence - NCTA*, pages 407–418, 2023.

[3] Markus Lucifer. Hair-loss-dataset. https://www.kaggle.com/datasets/brijlaldhankour/hair-loss-dataset?rvi=1, 2023.

[4] Renjie Wu and Eamonn Keogh. Current Time Series Anomaly Detection Benchmarks are Flawed and are Creating the Illusion of Progress. *IEEE 38th International Conference on Data Engineering (ICDE)*, pages 1479–1480, 2022.

Erstellung einer Cloud Strategie im öffentlichen Bereich am Beispiel der Bundesagentur für Arbeit

Nadine Deininger **Jörg Nitzsche**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt an der Fakultät Informatik und Informationstechnik, Esslingen

Einleitung

Die Bundesagentur für Arbeit (BA) ist im öffentlichen Bereich, mit 113.000 Mitarbeitern, die größte Verwaltung in Deutschland. Aktuell gibt es in der BA bis zu 180.000 vernetzte PCs und 20.000 Server, man sieht also, dass ein hohes Bedürfnis an Rechenressourcen besteht. Zudem werden jährlich 48 Millionen Emails und zusätzlich 11 Millionen Postsendungen verschickt. [3]

Am 14.08.2017 wurde das Bundesgesetz: „Gesetz der Verbesserung des Onlinezugangs zu Verwaltungsdienstleistungen - Onlinezugangsgesetz (OZG)“ [1] erlassen. In diesem Gesetz wird festgelegt, dass alle Bundes- und Landesverwaltungen dazu verpflichtet werden alle Verwaltungsdienstleistungen bis Ende 2022 in digitaler Form anzubieten. Hierbei steht vor allem die Vernetzung und die Digitalisierung der verschiedenen Verwaltungen im Mittelpunkt. Dies bedeutet konkret, dass alles, was bisher nur vor Ort oder per Post verfügbar war, digitalisiert, werden soll. Hierfür muss eine grundlegende IT-Infrastruktur aufgebaut werden, so dass ein zufriedenstellender Service gewährleistet werden kann. [1]

Um in Zukunft allen Mitarbeiter: innen sowie Bürger: innen gerecht zu werden, wird als ein Weg der Digitalisierung die Cloud als Lösung gesehen. Hierbei wird jedoch nicht über die eigene Private Cloud, welche der klassischen On-Premises- oder Datenzentrum-Architektur entspricht, gesprochen, sondern über den Sprung zu externen Cloud-Providern. Besonders auf die Souveräne Cloud wird in dieser Abschlussarbeit eingegangen, da diese einen elementaren Sprung zur weiteren Digitalisierung des öffentlichen Sektors bieten können. Bisher stellt der strenge Datenschutz in Deutschland die größte Hürde da, um neue Services zu integrieren und Prozesse in die Cloud zu heben.

Ziel der Arbeit

Ziel der Abschlussarbeit ist es, eine erfolgreiche Cloud Strategie für die Bundesagentur für Arbeit auszuarbeiten und dabei auch auf Evaluationsmöglichkeiten und Entscheidungsgrundlagen genauer einzugehen. Zudem soll ein konkreter Use Case zur Veranschaulichung aufgezeigt werden.

Cloud-Strategie

Die Cloud-Strategie eines Unternehmens trägt maßgeblich dazu bei, die Wettbewerbsfähigkeit eines Unternehmens zu erhalten. Denn Unternehmen und Organisationen müssen nicht nur flexibel, sondern auch in einem ansprechenden Zeitrahmen auf neue Trends, sowie technische Neuerungen reagieren können. Hierbei übernimmt der Einsatz einer Cloud eine wichtige Rolle in der strategischen Sicht einer Organisation. Denn eine gut entworfene Cloud-Strategie trägt dazu bei, einen Wechsel von veralteter Architektur zu einer modernen Cloud-Struktur zu ermöglichen. Diese vereinfacht den Weg in die Cloud deutlich, wenn sie die Vorteile dieser deutlich kommuniziert. [2] In der Cloud liegen die Potenziale zum Einen darin Dienstleistungen zu optimieren und zum Anderem in der Schaffung eines erheblichen Mehrwerts für Bürger: innen. Durch die Datenschutzverordnung sowie auch das Bundesdatenschutz-Gesetz, herrschen viele Einschränkungen für die Benutzung von externen Clouds, besonders im öffentlichen Sektor. In der Abschlussarbeit soll deshalb auch verstärkt darauf eingegangen werden welche Rahmenbedingungen für die Nutzung von externen Cloud-Anbietern dennoch möglich ist. Da dies die Entscheidungsfindung der Cloud Strategie mitbeeinflusst.

Souveräne Cloud

Ein Konzept, für die Integration der Bundesagentur für Arbeit in eine souveräne Cloud, wurde anhand eines bestehendes Angebotes der Google Cloud Plattform (GCP) in Zusammenarbeit mit T-Systems erstellt. Seit kurzem bietet Google ihre GCP Public Cloud auch als souveräne Cloud an. Dies ist möglich durch

die Kooperation mit T-Systems. Denn die Daten befinden sich zu keinem Zeitpunkt im nicht EU-Ausland und erfüllen somit die Anforderungen des Datenschutzes in Deutschland. GCP und T-Systems stellen unterschiedliche Angebote bereit. Zukünftig sollen neben der Basisvariante Sovereign-Controls zwei weitere Angebote mit einem höheren Souveränitätslevel angeboten werden. Zum Einen Supervised-Cloud welche auf der Google Supervised Cloud basiert und Hosted-Cloud welche auf der Google Distributed Cloud basiert. Das höchste Level an Souveränität bietet die Hosted-Cloud und eignet sich deshalb besonders für hochsensible Daten. [4]

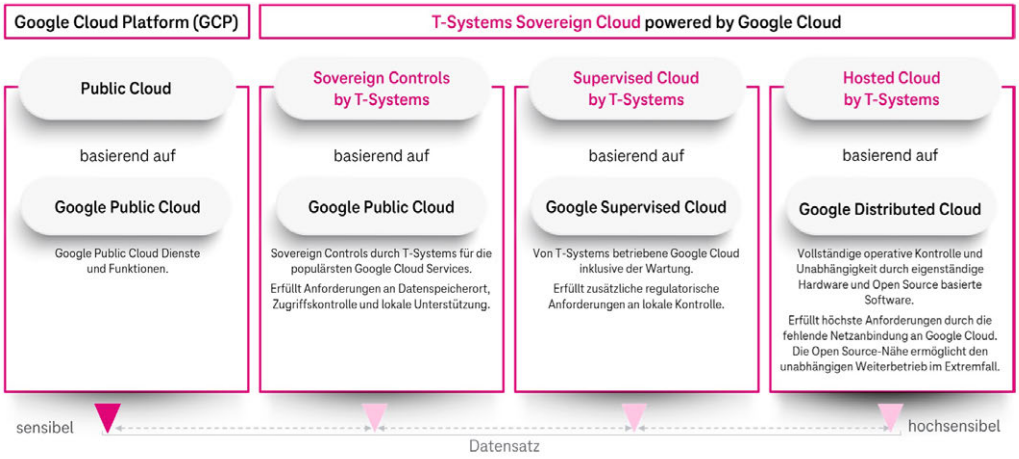


Abb. 1: GCP & T-Systems Angebote und Souveränitätsgrad [4]

Bereits ab Sovereign Controls by T-Systems basierend auf GCP, kann Google die vorhandenen Daten nicht mehr einsehen. Somit werden Anforderungen an den Datenspeicherort, Zugriffskontrolle sowie eine lokale Unterstützung gewährleistet. [4]

Abb. 2: Souveränitätskontrollen-Übersicht [4]

Ausblick

Im weiteren Verlauf der Abschlussarbeit wird ein Use-Case entworfen, welcher sich konkret mit der Entwicklungsumgebung, Datenbank und Kubernetes der Bundesagentur für Arbeit befasst. Neben dem Use-Case sollen auch potenzielle Visionen und Ziele einer Cloud-Strategie aufgezeigt werden.

Literatur und Abbildungen

[1] Bundesministerium der Justiz. Gesetz zur Verbesserung des Onlinezugangs zu Verwaltungsleistungen (Onlinezugangsgesetz - OZG). https://www.gesetze-im-internet.de/ozg/BJNR313800017.html, 2017.

[2] PwC Deutschland. PwC Deutschland, Cloud Strategie. https://www.pwc.de/de/im-fokus/cloud-transformation/cloud-transformation-journey/cloud-strategie.html#:~:text=Eine%20schl%C3%BCssige%20Cloud%2DStrategie%20tr%C3%A4gt,as%20a% 20Service%20(SaaS), 07 2023.

[3] Bundesagentur für Arbeit. Bundesagentur für Arbeit, Zahlen, Daten, Fakten IT-Systemhaus der Bundesagentur für Arbeit. https://www.arbeitsagentur.de/vor-ort/it-systemhaus/zahlen-daten-fakten, 2023.

[4] T-Systems International GmbH. Europa auf dem Weg zur souveränen Cloud. https://www.t-systems.com/de/de/cloud-services/managed-platform-services/souveraene- cloud/sovereign-cloud-powered-by-google-cloud, 2023.

Prototypische Implementierung einer B2B-lizenzfähigen Bewertungs-App

Oezge Demirkan **Anke Bez**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma Mercedes-Benz Tech Motion GmbH, Leinfelden-Echterdingen

Problemstellung

Bei der Mercedes-Benz Group AG existieren derzeit insgesamt 4 verschiedene App-basierte Lösungen, die sich allesamt mit dem Thema Planung und Durchführung von Erprobungsfahrten beschäftigen. Alle 4 Apps besitzen dabei unterschiedliche inhaltliche Schwerpunkte und sprechen damit unterschiedliche Interessensgruppen an. Dies erschwert den Fachbereichen derzeit die reibungslose Organisation, Durchführung und Auswertung Ihrer Erprobungsfahrten und Fahrzeugerprobungsträgern. Aus dieser aktuellen Ausgangssituation, im Research and Development-Bereich (RD-Bereich) der Mercedes-Benz Group AG, entstand in der Organisation ITS/IM Product Creation & Modification der Mercedes-Benz Tech Motion GmbH die Produktidee, eine generalisierte App für alle Kunden in der RD-Welt der Mercedes-Benz Group AG anzubieten, die die konsolidierten Anforderungen aller Fachbereiche in Summe abdeckt [3].

Zielsetzung und Vorgehensweise

Mit der neu entwickelten Produkt App DriveEMotion soll den RD-Kunden nun eine digitale und App-basierte Lösung von Fahrzeugerprobungsveranstaltungen aller Art angeboten und zur Verfügung gestellt werden. Dies bedeutet, dass die App den Veranstaltungsteilnehmern nicht nur den Zugriff auf Informationen, wie die Veranstaltungsagenda oder eine Übersicht der Fahrzeuge und Teilnehmer ermöglicht, sondern auch die Möglichkeit bietet, die Testfahrzeuge direkt in der App zu bewerten. Die Ergebnisse dieser Bewertungen können dank Datenbankanbindung, Cloud und PowerBI grafisch aufbereitet und für verschiedene Besprechungen genutzt werden [3].
Um die Produktideen bei der Mercedes-Benz Tech Motion GmbH dabei zu realisieren, gibt es ein internes Produktinkubator-Team, das die Fachbereiche bei der Entwicklung und im finalen Umsetzungsprozess neuer Produkte unterstützt. Die Produktideen, unter die auch die DriveEMotion App fällt, müssen vor der eigentlichen Entwicklung dabei folgende Inkubator Prozesse (siehe Abbildung 1) durchlaufen, um auf den Markt gebracht werden zu können [2].

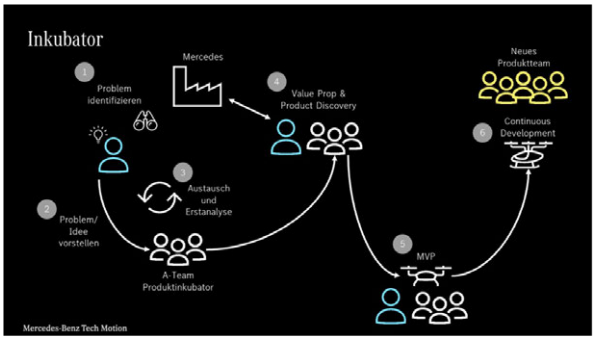


Abb. 1: Inkubatorprozess [2]

Derzeit befindet sich die DriveEMotion App in der MVP-Phase, die in Abbildung 2 durch eine grüne Ampel dargestellt ist.

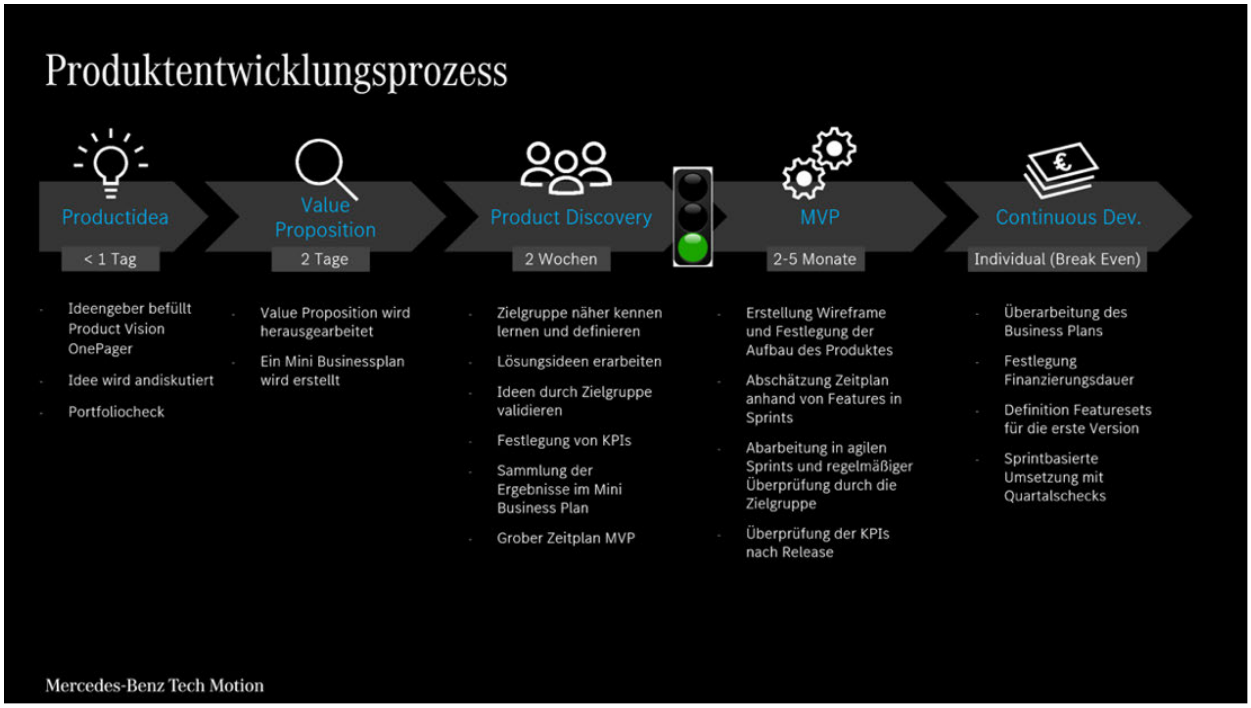


Abb. 2: Produktentwicklungsprozess DriveEMotion App [1]

Diese Bachelor-Thesis verfolgt zwei Hauptziele. Zum einen soll eine umfassende externe Analyse der aktuellen Erkenntnisse zu den Themen App-Entwicklung, -Vermarktung und Software-Lizenzmodelle durchgeführt werden, um darauf aufbauend Handlungsempfehlungen in einem späteren Verlauf ableiten zu können. Das zweite Hauptziel besteht in der eingehenden Analyse des bestehenden App-Prototyps, um Erkenntnisse hinsichtlich Usability und Kundenzufriedenheit zu generieren und diese in der Erstellung der Wireframe einfließen zu lassen. Folgende Forschungsfrage soll beantwortet werden: *Was sind die Erfolgs- und Einflussfaktoren bei der prototypischen Implementierung einer B2B-lizenzierbaren Bewertungs-App anhand eines Praxisbeispiels von Mercedes-Benz Tech Motion GmbH?*
Um die beiden Hauptziele der Bachelor-Thesis zu erreichen, ist diese in drei Hauptbereiche gegliedert. Zuerst werden die theoretischen Grundlagen wie App-Entwicklung und Vermarktung, Usability und User Experience und Software-Lizenzmodelle erläutert. Anschließend erfolgt ein Vergleich der theoretischen Erkenntnisse mit den Ergebnissen aus Experteninterviews aus der Praxis, woraus sich entsprechende Handlungsempfehlung ableiten lassen werden. Im dritten Teil der Bachelorarbeit werden die Anforderungen der Kunden anhand eines Interviews dargestellt. Anhand dieser Anforderungskriterien und eines Experteninterviews mit einem UX-Kollegen erfolgt sodann die Analyse des bestehenden Prototyps (MockUp) und es werden Verbesserungsvorschläge entwickelt.

Ausblick

Nach der Durchführung und Auswertung von Experteninterviews wird parallel dazu der bestehende Prototyp analysiert und Verbesserungsvorschläge erarbeitet. Um einen Ausblick in die Zukunft zu geben, ist geplant, ggf. durch weitere Interviews Skalierungsmöglichkeiten und Erweiterungen des Produktes aufzuzeigen.

Literatur und Abbildungen

[1] Eigene Darstellung.

[2] Markus Hornburg. Produktmanagement & Inkubator. https://social.cloud.corpintra.net/community/meinstandort/mercedes-benz-tech-motion/blog/2022/11/29/produktmanagement-inkubator, 2022.

[3] Matthias Meessen. Drive@RD. https://social.cloud.corpintra.net/groups/driverd-event-app/activity, 2022.

Ausarbeitung eines Audit-konformen Datenablagekonzepts für die Einkaufsabteilung unter besonderer Berücksichtigung des IP-Geschützten-Sonderprojekts SOFC

Malik Demolli **Thomas Rodach**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma Robert Bosch GmbH, Stuttgart/Feuerbach

Einleitung

Warum ist ein durchdachtes Datenablagekonzept in einem modernen Unternehmen unverzichtbar? In einer Ära, in der Daten als wertvolles Unternehmensvermögen gelten, ist die effiziente Verwaltung, Sicherung und Zugänglichkeit von entscheidender Bedeutung. Inmitten der digitalen Transformation und des datengetriebenen Geschäfts ist ein gut durchdachtes Datenablagekonzept nicht nur eine Option, sondern eine Notwendigkeit! Ohne klare Struktur besteht die Gefahr, dass ein Unternehmen in einem chaotischen Meer von Informationen versinkt. Dies kann zu kostspieligen Fehlern, ineffizienten Prozessen und unzureichender Datensicherheit führen. Ein klares Datenablagekonzept ist auch entscheidend, um Datenschutzbestimmungen und Compliance-Standards zu erfüllen [3].

Problemstellung

In der Einkaufsabteilung des Innovationsprojekts SSOFCßtehen wichtige Anpassungen an. Ein einheitliches Datenablagekonzept, das die fünf Produktbereiche berücksichtigt, fehlt derzeit, was zu einer gewissen Unordnung bei der Datenspeicherung führt. Die Anforderungen in den Produktbereichen sind oft unklar und unterliegen regelmäßigen Veränderungen. Die Abteilung hat bisher keine geeigneten Tools zur Bewältigung dieser Herausforderungen ausgemacht. Auch in Bezug auf Zugriffsberechtigungen und Speicherorte von Daten bestehen noch Unsicherheiten. Um diese Angelegenheiten zu klären und die Effizienz zu steigern, ist die Entwicklung eines klaren Datenablagekonzepts von großer Bedeutung, um die Grundlage für den Projekterfolg zu schaffen.

Ziel

Das Ziel besteht darin, basierend auf den Anforderungen sowohl aus projektspezifischer als auch produktbereichsbezogener Sicht eine tabellarische Darstellung eines einheitlichen, auditkonformen Datenablagekonzepts für das Unternehmen zu erstellen. Besonderes Augenmerk liegt auf dem Schutz des geistigen Eigentums im Rahmen des Sonderprojekts SOFC. Parallel dazu werden konkrete Handlungsempfehlungen ausgearbeitet und ein Umsetzungsvorschlag für diese Empfehlungen erarbeitet. Dies erfolgt unter Berücksichtigung der Empfehlung zur Nutzung eines geeigneten Tools von Bosch.

SOFC (Solid Oxide Fuel Cell)

Die Energieversorgung der Zukunft steht vor vielfältigen Herausforderungen, darunter die Reduzierung der CO2-Emissionen und die Gewährleistung einer zuverlässigen und nachhaltigen Energieerzeugung. In diesem Kontext hat Bosch seine Innovationskraft in die Entwicklung von Hochtemperatur-Brennstoffzellen, speziell der Solid Oxide Fuel Cell (SOFC) [1], investiert. Die Bosch SOFC-Brennstoffzelle ist ein vielversprechender Ansatz, um diese Herausforderungen anzugehen und die Energieversorgung effizienter und umweltfreundlicher zu gestalten. Die Bosch SOFC-Brennstoffzelle nutzt Wasserstoff, um Elektrizität zu erzeugen und Wärme zu produzieren. Sie findet Anwendung in Stromerzeugung, Heizungsanlagen, dezentraler Energieversorgung und als Range Extender in Elektrofahrzeugen. Ihre Verwendung von Wasserstoff als Brennstoff macht sie umweltfreundlich. Bosch arbeitet an ihrer Weiterentwicklung für eine nachhaltige Energieversorgung der Zukunft [5].

Abb. 1: SOFC-Brennstoffzelle [1]

SOFC-Wasserstofftechnologie

Bosch setzt auf innovative Wasserstofftechnologie im Rahmen seiner Solid Oxide Fuel Cell (SOFC). Hierbei wird Wasserstoff als Brennstoff genutzt, um durch elektrochemische Reaktionen sauberen Strom zu erzeugen. Die SOFC-Brennstoffzelle spaltet Wasserstoff in Protonen und Elektronen auf, wobei die erzeugten Elektronen elektrische Energie bereitstellen. Gleichzeitig entsteht Wasserdampf als Nebenprodukt [2]. Diese fortschrittliche Technologie von Bosch ermöglicht nicht nur effiziente Stromerzeugung, sondern bietet auch vielseitige Nutzungsmöglichkeiten für erzeugte Wärme. Ein wichtiger Schritt in Richtung nachhaltiger und umweltfreundlicher Energieversorgung.

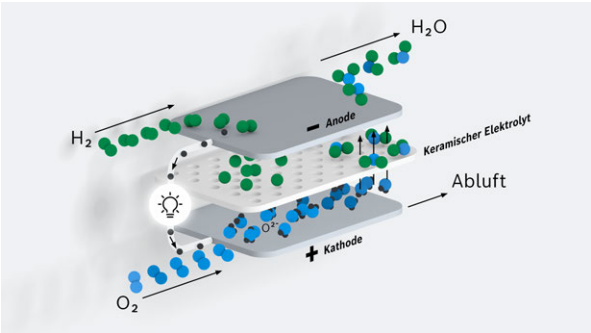


Abb. 2: Wasserstofftechnologie [2]

NDA/ICA-Infoschutz

In der Einkaufsabteilung des SOFC-Projekts liegt der Fokus auf dem verantwortungsbewussten Umgang mit sensiblen Informationen. Die Anwendung von Non-Disclosure Agreements (NDA) und Individual Confidentiality Agreements (ICA) ist dabei zentral für die Sicherung der Vertraulichkeit und Integrität der Einkaufsdokumente. Durch NDA und ICA wird sichergestellt, dass vertrauliche Informationen in Zusammenarbeit mit Partnern, Lieferanten und Stakeholdern geschützt bleiben. Die klare Definition der Vertraulichkeit durch NDA und ICA bietet eine rechtliche Grundlage. Im Falle von Verstößen können rechtliche Schritte eingeleitet werden, um die Interessen der Einkaufsabteilung und des Projekts zu schützen.

Entwicklungspartner CERES Power

Bosch und Ceres Power haben seit 2018 eine vielversprechende Innovationspartnerschaft im SOFC-Projekt etabliert. In dieser Kooperation bringt Ceres Power, mit ihrer Expertise im Bereich Festoxid-Brennstoffzellen und ihrer patentierten Technologie, das Herzstück der SOFC-Brennstoffzelle in das Projekt ein. Gemeinsam mit Bosch strebt man danach, die Effizienz und Nachhaltigkeit in der Energieerzeugung zu fördern. Diese Partnerschaft markiert einen bedeutenden Schritt in Richtung umweltfreundlicher Energielösungen für eine nachhaltige Zukunft.

IPCEI-Förderung im Projekt SOFC

Bosch setzt mit seiner Solid Oxide Fuel Cell (SOFC)-Technologie einen wegweisenden Schritt in Richtung nachhaltiger Energie. Die Unterstützung durch das „Important Project of Common European Interest“ (IPCEI) stärkt diese Innovation entscheidend, mit einer Förderung von beeindruckenden 160 Mio €. Die auf festen Oxiden basierende SOFC-Technologie von Bosch liefert saubere Energie für diverse Anwendungen. Die IPCEI-Förderung fungiert als Katalysator für die Entwicklung und erfolgreiche Markteinführung. Bosch nutzt diese Förderung zielgerichtet, um Umweltfreundlichkeit und Effizienz der SOFC-Technologie zu maximieren und aktiv zur europäischen Energiewende beizutragen [4].

Literatur und Abbildungen

[1] Bosch GmbH. SOFC-Brennstoffzelle. https://www.bosch.com/de/forschung/forschungsschwerpunkte/elektrifizierung/forschung-zu-wasserstofftechnologien/hochtemperatur-brennstoffzellensysteme/, 05 2022.

[2] Bosch GmbH. SOFC-Wasserstofftechnologie. https://www.bosch-hydrogen-energy.com/de/sofc/funktionsweise/, 04 2022.

[3] Thomas Gomell. 5 Punkte für ein erfolgreiches Datenmanagement. https://migraven-downloads.s3-eu-west-1.amazonaws.com/publikationen/Whitepaper-5-Punkte-fuer-erfolgreiches-Datenmanagement.pdf, 04 2020.

[4] Charlie Grüneberg. Bosch Power Units erhalten erste deutsche IPCEI Wasserstoff-Förderzusage. https://gas.info/presse/detailseite-news/detail/news-bosch_power_units_erhalten_erste_deutsche_ipcei_wasserstoff_foerderzusage-440715?cHash=2cc9961f22e6cfb06450b99883a985a0, 01 2022.

[5] J. Tegtmeier. Brennstoffzellen in Rechenzentren. https://bmdv.bund.de/SharedDocs/DE/Anlage/DG/12102021-praesentation-brennstoffzellen-tegtmeier_f.pdf?__blob=publicationFile, 10 2021.

Konzeption und Implementierung einer Applikation zur Visualisierung von Informationen zu energierelevanten Themen auf Quartiersebene

Claudius Deuschle **Markus Enzweiler**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma enersis europe GmbH, Kleinmachnow

Einleitung

Vor dem Hintergrund der prognostizierten Entwicklung, dass bis zum Jahr 2050 etwa 80 % der Weltbevölkerung in städtischen Gebieten leben werden, gewinnt die Bedeutung nachhaltiger Technologien für historische Städte zunehmend an Dringlichkeit. [1] Insbesondere in europäischen Städten, die reich an historisch bedeutsamen Gebäuden sind, stehen Sanierungsprojekte oft vor rechtlichen Beschränkungen. Diese Städte tragen maßgeblich zum weltweiten Energieverbrauch bei, was die Notwendigkeit einer nachhaltigen Transformation weiter unterstreicht. Das Projekt setzt an dieser Stelle an, indem es innovative Smart-City-Technologien in der Pilotstadt Alkmaar (NL) demonstriert und anschließend in sechs weiteren Städten implementiert. Das Projekt legt einen starken Fokus auf die koordinierte Einbindung von Stadtverwaltungen, Planern, Universitäten, Unternehmern und Bürgern, um gemeinsam die zukünftige Entwicklung europäischer Städte zu gestalten. Dabei wird ein bürgerorientiertes Design verfolgt, das die Bedürfnisse der Bürger in den Mittelpunkt stellt und ein offenes Innovationsökosystem schafft, um Bürger aktiv an der Mitgestaltung, Entscheidungsfindung, Planung und Problemlösung zu beteiligen.

Ziel der Arbeit

Die Hauptaufgabe besteht in der Entwicklung einer umfassenden Lösung zur Visualisierung von energierelevanten Informationen auf Stadtteilebene. Diese innovative Lösung wird ein integraler Bestandteil der offiziellen Website der Stadt Alkmaar sein, wodurch sie für die Bürger leicht zugänglich wird. Das Hauptziel des Projekts ist es, den Bürgern eine klare und vereinfachte Darstellung von Daten über den Energieverbrauch und erneuerbare Energien in verschiedenen Stadtteilen zu bieten. Die Vision hinter diesem Ziel geht über die reine Informationsdarstellung hinaus. Durch die Schaffung einer leicht verständlichen Plattform zielt das Projekt darauf ab, die Bürger aktiv in den gesamten Prozess einzubinden. Die Bürger sollen ermutigt werden, sich aktiv an Diskussionen, Entscheidungsprozessen und möglicherweise sogar an der Planung nachhaltiger Initiativen für ihre Stadt zu beteiligen. Dieses Ziel spiegelt die zugrundeliegende Philosophie wider, dass eine informierte und engagierte Bürgerschaft ein entscheidender Motor für den Erfolg von Stadtentwicklungsprojekten ist. Insgesamt soll mit dieser Lösung eine transparente und partizipative Kommunikationsplattform geschaffen werden, die nicht nur der Information, sondern auch der aktiven Bürgerbeteiligung dient.

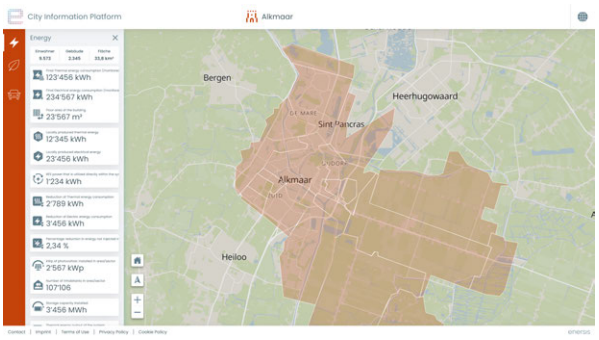


Abb. 1: Design Konzept [2]

Lösungsansatz

Der gewählte Lösungsansatz konzentriert sich darauf, die spezifischen Anforderungen des Projekts umfassend zu erfüllen. SvelteKit könnte als Frontend-Framework genutzt werden, um eine benutzerfreundliche und reaktive Benutzeroberfläche zu schaffen, die nahtlos mit GeoJSON-Daten umgehen kann. Dadurch wäre eine einfache Integration von Mapbox-Kartenkomponenten in die Anwendung möglich. Die Syntax von SvelteKit ist klar und deklarativ, was die Entwicklung von

wiederverwendbaren Kartenkomponenten erleichtert. Diese können effizient in verschiedenen Anwendungsbereichen eingesetzt werden. Der Lösungsansatz zielt darauf ab, eine leistungsfähige und benutzerfreundliche Anwendung zu schaffen, die den Anforderungen des Projekts in Bezug auf Kartendarstellung und -interaktion gerecht wird (siehe Abbildung 1).

SvelteKit

SvelteKit ist ein modernes Web-Framework, das auf dem Svelte-Framework basiert und eine effiziente Routing-Lösung für die Webentwicklung bietet. Die Plattform zeichnet sich durch ein dateibasiertes Routing-System aus, das klare Strukturen durch Ordner und Dateien ermöglicht. Dynamische Routen erlauben die Behandlung von Teilen der URL als Parameter, was die Flexibilität steigert. Durch die Erstellung von Layouts lassen sich gemeinsame Strukturen einfach organisieren. Middleware bietet leistungsstarke Optionen für vor- und nachgelagerte Logiken wie Authentifizierung. Nested Routing ermöglicht eine hierarchische Organisation von Unterseiten und Unterkomponenten. Die reaktive Client-side Navigation verbessert die Benutzererfahrung durch schnelle Seitenwechsel ohne Neuladen. Die Integration von Transitionen ermöglicht die einfache Umsetzung von Seitenübergängen und Animationen. Insgesamt vereint SvelteKit eine klare Syntax, minimale Boilerplate-Code-Anforderungen und das Fehlen von Laufzeitbibliotheken. Dadurch wird eine höhere Effizienz und verbesserte Leistung erzielt. [3] Ein weiteres Merkmal ist die Hydration, bei der Komponenten während der Kompilierung generiert werden. Dies führt zu einer schnelleren Ladezeit und einer reaktionsfähigeren Benutzeroberfläche, da ein Teil des Client-seitigen Codes bereits während der Initialisierung des HTML-Dokuments bereitgestellt wird (siehe Abbildung 2).

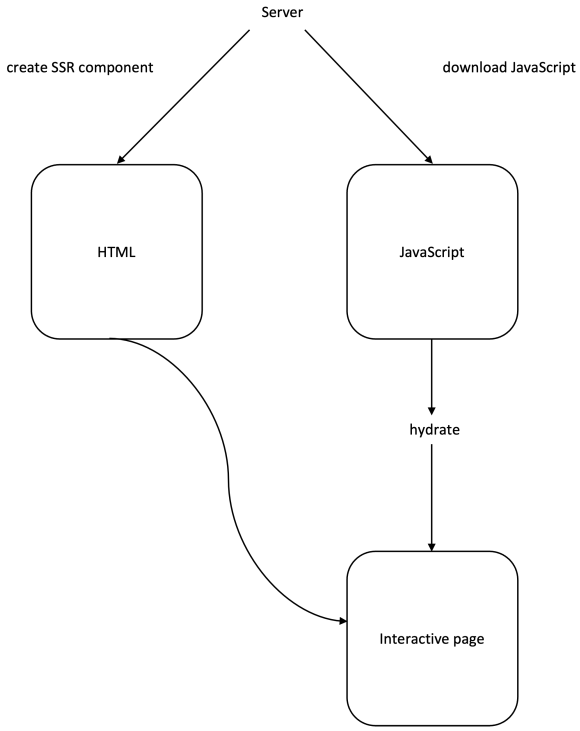


Abb. 2: Hydration in SvelteKit [2]

Ausblick

Der Ausblick auf das Projekt verspricht eine wegweisende Reise im Bereich der nachhaltigen Stadtentwicklung und Smart-City-Technologien. Die Demonstration innovativer Lösungen in der Pilotstadt Alkmaar sowie deren Anwendung in sechs weiteren europäischen Städten sind entscheidende Schritte auf dem Weg zu einer zukunftsorientierten Stadtgestaltung. Das Projekt hebt sich als Vorreiter für nachhaltige Sanierung und Energieeffizienz hervor, indem es sich auf historische Städte mit ihren einzigartigen Herausforderungen und rechtlichen Beschränkungen konzentriert.

Literatur und Abbildungen

[1] BMZ Bundesministerium für wirtschaftliche Zusammenarbeit und Entwicklung. Hintergrund: Das Zeitalter der Städte. https://www.bmz.de/de/themen/stadtentwicklung/hintergrund-18138, 2023.

[2] Eigene Darstellung.

[3] Mark Hirtenmacher. Was ist SvelteKit? Unser Leitfaden zu SvelteKit. https://www.biteno.com/was-ist-sveltekit/, 11 2023.

Analyse und Entwicklung einer Kommunikationsmöglichkeit zwischen ESP und Webserver in internetlosen Umgebungen

Marcel Dommer **Markus Enzweiler**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma IT.TEM GmbH, Stuttgart

Einleitung

Um die Parkplatznot in Industriegebieten zu reduzieren wurde ein Parkplatzverwaltungssystem entwickelt, das es Unternehmen ermöglicht mühelos die eigenen Parkplätze dem System freizugeben und über dieses auch Parkplätze anderer Unternehmen zu buchen. Der Grund dafür ist, dass viele Unternehmen mittlerweile eine hohe Home-Office Quote haben, sodass nicht mehr jeden Tag alle Parkplätze benötigt werden. Dieses freie Kontingent kann über die Parkplatzverwaltung freigegeben werden und anderen Firmen zu Nutze kommen, die an diesen Tagen mehr Parkplätze benötigen. Um dies so Minimalinvasiv und trotzdem leicht verständlich an den Parkplätzen anzuzeigen wurden Displays beschafft, die Daten direkt am Parkplatz anzeigen können. Diese sind mit einem ESP32 und einem LoRa, sowie RJ45-LAN Modul ausgestattet.

Ziel dieser Arbeit

Da nicht jedes Parkhaus mit Netzwerkkabeln ausgestattet ist muss nun ein Kommunikationsweg gefunden werden, über den die Daten vom Webserver an die Displays vom Webserver übertragen werden können. Dabei soll auf die Sicherheit und Zuverlässigkeit, aber auch auf die Kosten der Übermittlung Rücksicht genommen werden. Da hier Personenbezogene Daten übermittelt werden muss die Kommunikation nach den Safety und Security Konzepten gesichert werden, um ein abgreifen beziehungsweise verändern der Daten auszuschließen. Die Arbeit beschäftigt sich damit, die Verbindung zum Internet aufzubauen. Das bedeutet, dass der Kommunikationsweg auf der einen Seite mit dem Internet kommunizieren muss und auf der anderen mit dem ESP. Dabei muss keine Echtzeitverbindung bestehen, sondern die Daten können auch erst mit Verzögerung zugestellt werden. Da Parkhäuser die Besonderheit aufweisen, oft massive Stahlbetonbauten unter der Erde zu sein, muss in Bezug auf die Funkkommunikation geprüft werden, in wie weit diese in diesen Gebäuden funktioniert. Im Folgenden wurden die verbreitetsten Kommunikationsmöglichkeiten geprüft, analysiert und evaluiert.

LoRa

LoRa ist eine niederfrequente Kommunikation auf Basis von CSS-Modulation. Sie ist auf Übertragung auf große Distanzen bei gleichzeitig niedrigem Energieverbrauch ausgelegt. Die Frequenz von nur 434MHz (ISM-Band) beziehungsweise 863-870MHz (SRD-Band) macht, im Vergleich zu anderen Kommunikationswegen, die Kommunikation weniger anfällig für Hindernisse und Wände, denn je niedriger die Frequenz eines Signals ist, desto höher ist die Penetrationskraft durch Hindernisse. Dies liegt an der größeren Wellenlänge, die weniger absorbiert oder reflektiert wird. Diese geringere Frequenz geht allerdings auf Kosten der Datenrate, die bei LoRa relativ gering ausfällt. Über den Spreading Factor (SF) kann eingestellt werden, wie weit das Signal kommt, da es die Breite der Signal-Chrips verändert. Dies hat einen direkten Einfluss auf die Datenrate, da je breiter der Chirp ausfällt, desto weiter kann das Signal ohne Verluste reichen, es braucht allerdings mehr Zeit jeden Chirp zu senden, sodass die Datenrate dadurch reduziert wird. [2] Die zusätzliche 1% Regelung von LoRa macht die Kommunikation ziemlich schwer, da die Datenmenge zu hoch wird. Die 1% Regel besagt, dass jedes Gerät auf Basis der Netzfairness nur 1% der Zeit senden darf. Damit soll gewährleistet werden, dass jedes Gerät die Chance hat auf den Frequenzen zu senden und kein einzelnes Gerät die Frequenzen überlastet. Ein Nachteil im Vergleich zu WLAN ist, dass LoRa Signale nicht so gut weitergeleitet werden können. Es gibt also keinen klassischen Repeater wie bei WLAN.

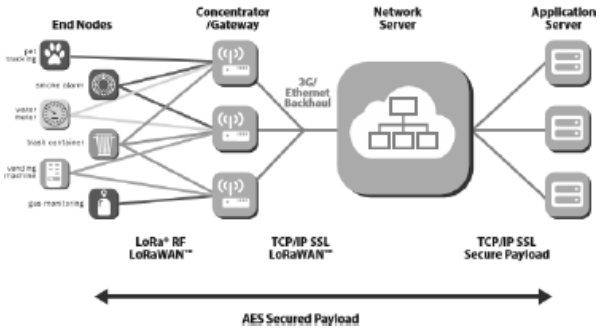


Abb. 1: LoRa Aufbau [2]

WLAN

Wireless Local Area Network, kurz WLAN, ist der Funkstandard der von Wifi genutzt wird, um in der heutigen Zeit für die meisten Geräte eine Funkverbindung aufbauen zu können. Wifi ist in den IEEE-Standards der 802.11 Familie definiert. Sie wurde dafür konzipiert, viele Geräte und einen hohen Datendurchsatz zu ermöglichen. Dafür stehen die Frequenzbänder 2400-2483,5 MHz (2,4 GHz) und 5150-5350 MHz (5GHz) zur Verfügung. Die höhere Frequenz im Vergleich zu LoRa macht WLAN aber zu einem schlechteren Kandidaten für diesen Anwendungsfall, da die Datenmenge relativ gering ist und die Reichweite deutlich geringer ausfällt. Der Vorteil gegenüber LoRa ist allerdings die Unterstützung von WDS. WDS bezeichnet einen Standard nach IEEE802.11, der Funkbrücken ermöglicht. Das bedeutet, dass die Reichweitenbegrenzung von WLAN dadurch minimiert werden kann, indem man auf der zu überbrückenden Strecke in regelmäßigen abständen Funkverstärker aufstellen kann, um das Signal über weite Strecken zu transportieren. Voraussetzung hierfür ist die Möglichkeit diese Verstärker auf der Strecke aufbauen zu können. Da dieser Standard allerdings nicht genau genug definiert ist, ist es allerdings ratsam auf einen Hersteller zu setzen, um Probleme zu vermeiden.

LTE/5G

LTE bzw. 5G sind die aktuellen digitalen Mobilfunktechniken. LTE ist dabei der Nachfolger von UMTS und 5G ist der Nachfolger von LTE. Da LTE allerdings noch nicht komplett von 5G abgelöst wurde und auch noch einige Jahre existieren wird, werden hier auch beide Funkstandards berücksichtigt. Ein großer Vorteil dieser Mobilfunkstandards ist die bestehende Infrastruktur. Da Konzerne wie die Deutsche Telekom, Vodafone oder E+ bereits ein gut ausgebautes Netz haben, ist der Zugang ins Internet an vielen Standorten bereits möglich. Da LTE allerdings auf den Frequenzen 800, 1800 und 2600 MHz arbeitet und die Antennen großzügig verteilt stehen, kann es allerdings passieren, dass in Gebäuden die Verbindung instabiler ist. [4] Bei 5G sollen weit höhere Frequenzen eingesetzt werden. So ist aktuell geplant das 26, 40 oder auch 86 GHz für 5G einzusetzen. [3] Dies bringt zwar im ersten Moment den Nachteil, dass die hohe Frequenz das durchdringen von Wänden noch stärker einschränkt, allerdings wird hier das Netz von den Netzbetreibern ausgebaut, sodass es mehr Sendemasten geben wird, um nach Bundesvorgaben eine Große Abdeckung zu schaffen und auch in Gebäuden 5G anbieten zu können. Allerdings sind diese Ausbauten noch nicht weit genug fortgeschritten, um darauf zu setzen. Ein weiterer Nachteil ist, dass für das Nutzen der Netze zusätzliche laufende Kosten entstehen.

Technologie kombinieren

Da es keine Anforderungen an Echtzeit Kommunikation gibt lassen sich, soweit baulich möglich die meisten Technologien kombinieren um das Beste aus beidem zu bekommen. So kann LoRa für die Kommunikation innerhalb der Parkhäuser benutzt werden und am Eingang des Parkhauses, wo die LTE/5G Verbindung existiert diese für den restlichen Weg benutzen. Ein ganz anderer Weg ist die Nutzung des Users selbst. Da mittlerweile jeder ein Smartphone mit sich trägt kann man dieses als Kommunikationsgateway benutzen. Hierfür können die Daten über Mobile Daten oder WLAN auf das Gerät synchronisiert werden und wenn das Geräte sich in der Tiefgarage befindet dort per WLAN oder Bluetooth die Daten an das Parkplatzsystem weitergegeben werden.

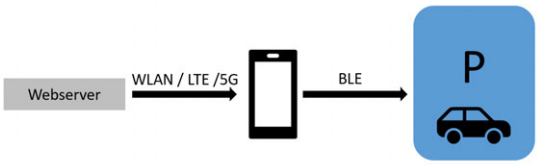


Abb. 2: Referenzkommunikation mithilfe eines Mobiltelefons [1]

Ausblick

In Zukunft soll für die Kommunikation die App-Lösung verwendet werden. Es bietet die größten Vorteile, ist flexibel und schafft keine laufenden Kosten. Die App kann zusätzlich für den User einen Mehrwert bieten, da nun die Buchung von Parkplätzen auch über diese App geschehen kann. Es wird allerdings trotzdem auf alle Sicherheitsaspekte wert gelegt, da in der App die Daten nur verschlüsselt abgelegt werden. Weiterhin werden nur Daten an das Handy synchronisiert, die unbedingt nötig sind. Damit soll die Sicherheit der Daten gewährleistet sein.

Literatur und Abbildungen

[1] Eigene Darstellung.

[2] Shilpa Devalal and A. Karthikeyan. LoRa Technology - An Overview. In *2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)*. IEEE, 2018.

[3] Bundesamt für Strahlenschutz. 5G. https://www.bfs.de/DE/themen/emf/mobilfunk/basiswissen/5g/5g.html, 2022.

[4] Bundesamt für Strahlenschutz. LTE – Long term Evolution. https://www.bfs.de/DE/themen/emf/mobilfunk/basiswissen/lte/lte.html, 2022.

Konzeptionelle Entwicklung eines standardisierten Tools zur automatisierten Softwareaktualisierung der Komponenten in einem Heizsystem

Kevin Ehling **Mirko Sonntag**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma Bosch Thermotechnik GmbH, Wernau

Einleitung

In einem modernen Heizsystem wird Software nicht nur für das Steuergerät des Wärmeerzeugers, sondern auch in vielen weiteren Komponenten benötigt. Abbildung 1 zeigt ein Beispiel: Der Systemregler (2), der normalerweise am Wärmeerzeuger (1) angebracht ist, ermöglicht die Konfiguration und Einrichtung des Heizsystems. Die Funktionsmodule (4) steuern die einzelnen Heizkreise, während die Raumbedieneinheiten (3) als Fernbedienung für diese Heizkreise fungieren. Sie bieten wichtige Informationen und Funktionen wie das Einstellen von Zeitprogrammen.

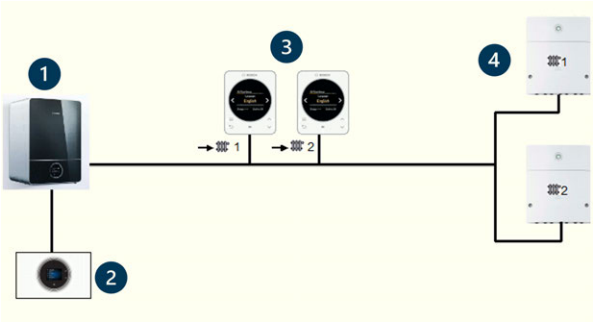


Abb. 1: Verbindung der Komponenten in einem Heizsystem [1]

Das Zusammenspiel der einzelnen Komponenten wird in einem der letzten Entwicklungsschritte, dem Systemtest, überprüft. Hier werden die einzelnen Komponenten im Systemverbund auf funktionale und nicht-funktionale Anforderungen des Systems getestet [3].

Aufgabenstellung

Die Produktentwicklungsteams der einzelnen Komponenten liefern in regelmäßigen Abständen neue Software-Artefakte. Dies erfordert eine kontinuierliche Aktualisierung der Systemtest-Komponenten. Die Aktualisierung der Firmware in einem einzelnen Gerät erfordert derzeit verschiedene Programme und Hardware, was den Prozess zeitaufwendig und fehleranfällig macht. Darüber hinaus führt die Entnahme des Geräts aus dem System zu Verzögerungen und höheren Entwicklungskosten. Um diese Probleme zu lösen, werden in dieser Arbeit zwei zentrale Forschungsfragen behandelt:

- Kann der Prozess zur Aktualisierung der Firmware für den Nutzer vereinfacht und automatisiert werden?
- Ist es möglich den Prozess zur Aktualisierung der Firmware in die automatisierten Tests zu integrieren?

Um diese Fragen zu beantworten, werden in diesem Projekt Anforderungen an ein Tool festgelegt und ein Konzept erarbeitet.

Konzept

Durch die Analyse der Prozesse, sowie Gesprächen mit den Testingenieuren, konnten die Probleme und Anforderungen an das neue Tool zusammengefasst werden. Das resultierende Konzept umfasst dabei zwei *Use Cases*, die Ausarbeitung der Anforderungen an das Tool sowie Ideen zur Implementierung der Lösung. Die *Use Cases* lassen sich wie folgt zusammenfassen:

- Ein Nutzer kann eine Komponente des Systems automatisiert ohne großen Aufwand aktualisieren
- Ein Nutzer kann den Aktualisierungsvorgang in den automatisierten Test integrieren

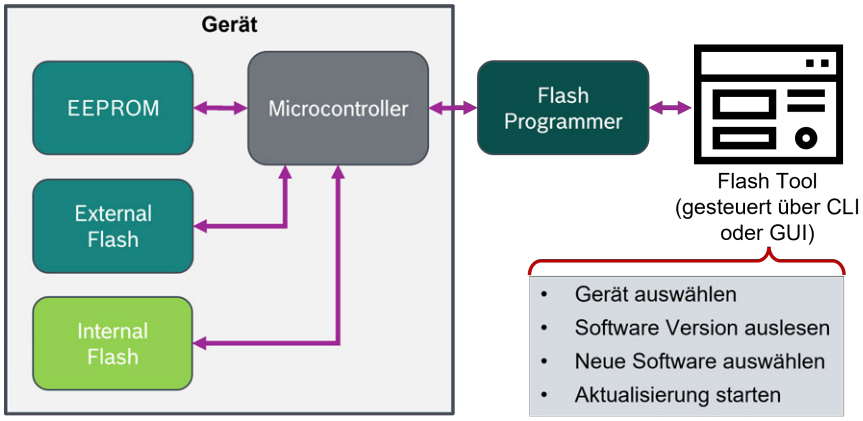


Abb. 2: Konzept zur Aktualisierung der Firmware [1]

Im Vordergrund steht dabei, dass der Nutzer nur noch ein einziges Programm sowie den entsprechenden *Programmer* benötigt, um eine einzelne Komponente zu aktualisieren. Eine grafische Benutzeroberfläche (GUI) wird entwickelt, um dem Nutzer die Konfiguration der Aktualisierung zu erleichtern. Dazu zählt:

- Auswahl der jeweiligen Komponente
- Auslesen der aktuellen Softwareversion, um zu prüfen, ob eine Aktualisierung notwendig ist
- Auswahl der neuen Softwareversion
- Start der Aktualisierung

Das Tool leitet die HEX-Dateien (Hexadezimal-Quelldateien) der ausgewählten Version an den *Programmer*. Dieser ist über einen Adapter direkt mit dem Mikrocontroller verbunden und kann mithilfe von *External-* und *Flash-Loadern* auf die intern und extern angeschlossenen Speicher schreiben.
Gleichzeitig muss das Tool über die Kommandozeile bedienbar sein, um es für die automatisierten Tests verwenden zu können. Die automatisierten Tests werden mithilfe des *Robot Framework*, einem *Open Source Automation Framework* erstellt und durchgeführt [2].

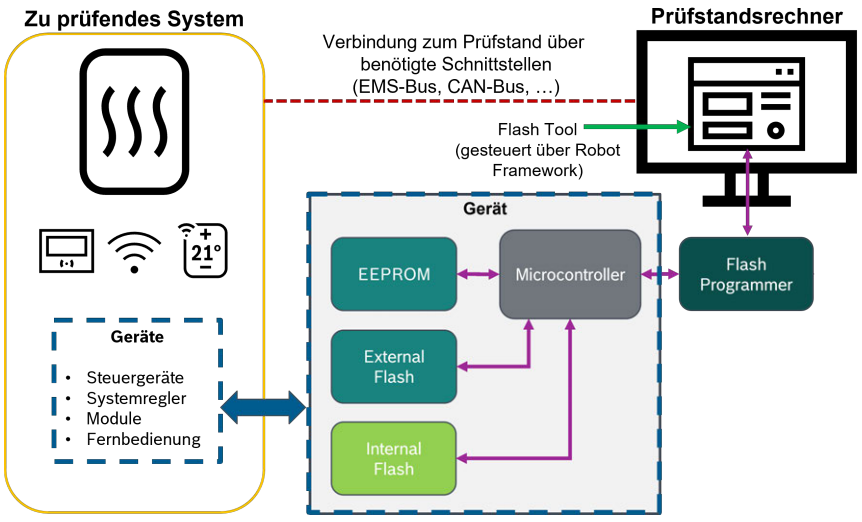


Abb. 3: Konzept zur Anbindung an die automatisierten Tests [1]

Implementierung

Um die benötigte GUI und das *Command Line Interface* (CLI) für das Tool zu entwickeln, wird das Python Modul PySide2 verwendet [4]. Dieses wird bereits in anderen Projekten erfolgreich genutzt. Für ein ausgewähltes Gerät wurde bereits ein *Proof of Concept* implementiert. Das Tool ermöglicht das Auswählen des *Programmers* und der Komponente, das Auslesen der aktuellen Softwareversion, das Auswählen des Produktcontainers mit den dazugehörigen HEX-Dateien als ZIP-Ordner und das Auslösen des Programmiervorgangs mit einem einzigen Klick. Alle Schritte zur

Aktualisierung werden automatisiert abgearbeitet und der Prozess wird protokolliert, um erfolgreiche oder gescheiterte Aktualisierungsvorgänge anzuzeigen.

Ausblick

Als nächster Schritt wird das CLI implementiert, um den Einsatz des Tools für automatisierte Tests zu ermöglichen. Die Architektur des Tools ermöglicht die flexible Ergänzung weiterer Geräte. Dadurch können nach und nach weitere Geräte aus dem Produktportfolio integriert werden. In Zukunft sollen die Artefakte auf einer Plattform bereitgestellt werden, um die Suche nach passenden Versionen und Dateien zu erleichtern.

Literatur und Abbildungen

[1] Eigene Darstellung.

[2] Robot Framework Foundation. Robot Framework. https://robotframework.org, 2023.

[3] Tilo Linz. *Testen in Scrum-Projekten. Leitfaden für Softwarequalität in der agilen Welt.* dpunkt.verlag, 2 edition, 2017.

[4] Cristian Maureira Fredes. Qt for Python. https://wiki.qt.io/Qt_for_Python, 03 2018.

Data-Driven Decision Making: Einfluss moderner Methoden der KI auf das Decision Making

Hueseyin Er **Dirk Hesse**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt an der Fakultät Informatik und Informationstechnik, Esslingen

Einleitung

Wir stehen am Anfang einer neuen Ära, in der die künstliche Intelligenz (KI) nicht nur unsere Technologie, sondern auch uns selbst und unsere Art zu leben transformiert. Durch die Entwicklung fortschrittlicher Technologien sind wir in der Lage, Wissen zu erweitern, Systeme zu optimieren und Produkte sowie Infrastrukturen zu schaffen, die unseren Alltag bereichern. KI hat sich inzwischen in alle Lebensbereiche integriert und beeinflusst, wie wir Entscheidungen treffen und Probleme lösen. Mit dem Aufkommen der Entscheidungsintelligenz oder auch Decision Intelligecne (DI) genannt, erleben wir eine neue Phase, in der KI zunehmend zugänglich und nützlich für die Verbesserung unserer Entscheidungsprozesse wird. DI verknüpft die technologischen Errungenschaften der KI mit bewährten Entscheidungstheorien und praktischen Werkzeugen, um Entscheidungsprozesse in einem Data-Driven-Unternehmen zu verfeinern und zu beschleunigen [4].

Ziel

In dieser Arbeit werden die Schlüsselkonzepte hinter der Entscheidungsintelligenz beleuchtet. Es wird diskutiert, was DI ausmacht, warum sie für moderne Organisationen unerlässlich ist, und wie sie in die bestehenden Informations- und Datenstrukturen eingebettet ist. Zudem wird ein Blick darauf geworfen, wie Unternehmen verschiedener Branchen DI-Technologien adaptieren und welche Hindernisse oder Anreize es bei der Implementierung gibt.

Data-Driven Decision Making

Data-Driven Decision Making (DDDM) ist ein Ansatz, bei dem Unternehmen sich auf Daten und Analysen stützen, um strategische Entscheidungen zu treffen. Dieser Ansatz wird zunehmend als zentral für die moderne Geschäftswelt angesehen, da er eine objektive Grundlage für Entscheidungen bietet und auf messbaren Informationen basiert. DDDM integriert Daten in alle Ebenen der Unternehmensführung, was nicht nur zu einer Kultur der kritischen Auseinandersetzung und Neugier führt, sondern auch zu einer fundierten Entscheidungsfindung beiträgt [3].

Wesentliche Methoden der KI für DI-Systeme

Machine Learning (ML) ist ein faszinierender Zweig der künstlichen Intelligenz, der Computern das "Lernenërmöglicht – ein Prozess, bei dem Algorithmen aus Daten lernen, um Muster zu erkennen und Prognosen zu stellen. Die Grundidee ist, dass ein Computerprogramm durch Erfahrung besser wird, ähnlich wie ein Mensch, der aus seiner Interaktion mit der Welt lernt. Diese werden grundsätzlich aufgeteilt in: **Supervised ML** ist eine zentrale Methode in der künstlichen Intelligenz, bei der Modelle mit vorab klassifizierten Daten trainiert werden, um Muster zu erkennen und Vorhersagen zu treffen. Es gibt zwei Haupttypen: Regression und Klassifikation. Regressionsmodelle prognostizieren kontinuierliche Ergebnisse und entschlüsseln Beziehungen zwischen Variablen, oft durch lineare, logistische oder polynomiale Regression. Klassifikationsmodelle hingegen ordnen Daten in klar definierte Kategorien ein, wie zum Beispiel Support Vector Machines (SVM), die Pflanzenarten klassifizieren können. **Unsupervised ML** findet Muster in unbeschrifteten Daten, wie etwa das Sortieren von 7.000 nicht klassifizierten Tierbildern in natürliche Gruppen. Es nutzt Algorithmen wie Clustering, um beispielsweise Zebras anhand von Streifenmustern, männliche Löwen durch Mähne und Größe und Elefanten durch ihre Statur und den Rüssel zu identifizieren. Ein praktisches Beispiel wäre die Segmentierung von Kunden in einem Online-Einzelhandelsgeschäft mittels k-Means Clustering. **Reinforcement Learning** (RL) gewinnt an Bedeutung, da es autonome Agenten – wie Industrieroboter oder selbstfahrende Autos –

ermöglicht, Entscheidungen ohne menschliches Zutun zu treffen. Ein Agent im RL-Kontext kann eine physische oder nicht-physische Einheit sein, die ihre Umgebung über Sensoren wahrnimmt und über Aktuatoren darauf reagiert. RL basiert auf dem Prinzip der operanten Konditionierung und nutzt Belohnungen und Strafen, um Agenten das selbstständige Treffen von Entscheidungen beizubringen. Es eignet sich besonders für komplexe Entscheidungsprozesse und wird in Bereichen wie Finanzhandel, Content-Empfehlungen (Recommendations), Gaming, Marketing, autonomes Fahren, Robotik und Gesundheitswesen angewandt.

DI-Formen

Decision Intelligence (DI) verbindet künstliche Intelligenz mit Entscheidungsprozessen und unterteilt sich grundsätzlich in drei Stufen [1] : Datenanalyse, Entscheidungsverstärkung und Entscheidungsautomatisierung. In der ersten Phase unterstützt sie grundlegende Entscheidungen durch Datenanalyse. Die zweite Phase verstärkt menschliche Entscheider durch komplexe Analysen und Empfehlungen. Die dritte Phase automatisiert Entscheidungen vollständig, wobei menschliches Eingreifen immer weniger notwendig wird. Mit fortschreitender Technologie reduziert sich die menschliche Beteiligung am Entscheidungsprozess zunehmend [4].

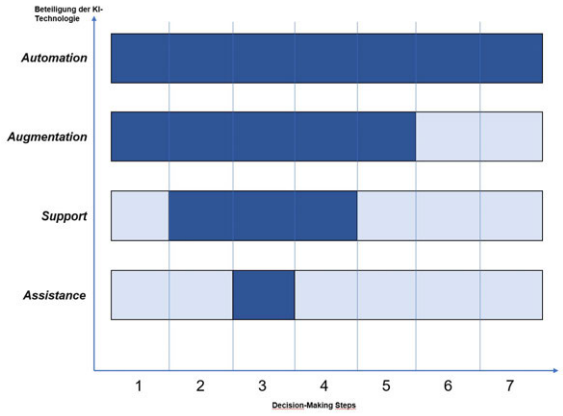


Abb. 1: Die Beteiligung der KI-Technologien im Decision-Making Process [1]

Anwendungsfall: „Nächstbeste Empfehlung“ mit Microsoft Azure

In diesem Anwendungsfall wird anhand eines Vorhersagemodells (predictive model) Vorschläge generiert und die beste Option für einen bestimmten Kunden ausgewählt. Mithilfe von Microsoft Azure, Power BI und Python wird dieser Anwendungsfall realisiert. Die Kernarbeit in der Datenebene dieser AI-Lösung basiert auf einem Ansatz des Reinforcement Learnings und gehört zu der Kategorie des Decision Augmentation, da die Entscheidung nicht vom Model getroffen wird. Hierbei wird Azure Personalizer, ein AI-Service von Microsoft, der auf dem Prinzip des RL basiert, genutzt. Um den Service zu trainieren, werden simulierte Nutzerdaten, die in einem JSON-File gespeichert sind, verwendet. Die Interaktion mit dem Modell und die daraus resultierenden Belohnungen (Reward) und Bestrafungen (Penalty) werden in einem Azure Notebook skriptgesteuert und die Ergebnisse in Azure Blob Storage für Analysen in Power BI festgehalten [2]. So kann man die Leistung des Modells überwachen und die Effektivität personalisierter Empfehlungen im Vergleich zu Standardempfehlungen messen [5].

Abb. 2: Churn Prevention Dashboard [5]

Ausblick

Decision Intelligence stärkt Data-Driven-Unternehmen in ihrer Entscheidungsfindung durch KI-gestützte Analysen und Prognosen und zielt darauf ab, die Qualität geschäftlicher Entscheidungen auf allen Ebenen zu verbessern. DI verdrängt nicht den Menschen, sondern bietet durch die Nutzung von AI eine umfassendere und leichter zugängliche Datengrundlage, die zu optimalen Entscheidungen führen soll. Bis Ende 2024 werden 75% der Unternehmen operative KI-Tools wie maschinelles Lernen und NLP einsetzen, um tiefere Einblicke in Betriebsabläufe und Lieferketten zu gewinnen. Bereits 2023 werden 33% der großen Organisationen DI nutzen, nach Gartner. Der flächendeckende Einsatz von DI wird als unausweichlich betrachtet und könnte zur wichtigsten Software-Kategorie für Unternehmen werden, die den Arbeitsalltag grundlegend verändert. Jedes Unternehmen wird eine KI benötigen, die jede verfügbare Datenquelle nutzt, um Entscheidungen zu treffen, die zuvor unmöglich waren [2].

Literatur und Abbildungen

[1] Eigene Darstellung.

[2] Laurence Goasdstuff. Gartner Top 10 Trends in Data and Analytics for 2020. https://www.gartner.com/smarterwithgartner/gartner-top-10-trends-in-data-and-analytics-for-2020, 2020.

[3] Sandy Lanuschny. Data Driven Decision Making | Teil 1 – mit Michael Emaschow. https://www.papershift.com/expertseries/data-driven-decision-making-was-ist-das-eigentlich, 2022.

[4] Miriam OCallaghan. *Decision Intelligence: Human–Machine Integration for Decision-Making*. CRC Press, 2023.

[5] Tobias Zwingmann. *AI-Powered Business Intelligence: Improving Forecasts and Decision Making with Machine Learning*. O'Reilly Media, Inc., 2022.

Digitaler Zwilling in der Automobil-Supply-Chain: Effizienzsteigerung und Fehlerprävention

Laurent Etemi **Clemens Klöck**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt an der Fakultät Informatik und Informationstechnik, Esslingen

Einleitung

Die Automobilzulieferkette, welche auch als Automobil-Supply-Chain genannt wird, gehört zu den komplexesten Supply Chains der Welt. Die globalen Krisen und geopolitischen Konflikte stellen Automobilhersteller und ihre internationalen Wertschöpfungsketten vor wachsenden Herausforderungen. [3]

1. Mangelnde Sichtbarkeit: Die Komplexität der Lieferketten führt zu blinden Flecken, die zu Bestandsengpässen und Verzögerungen führen können. Es besteht Bedarf an besserer Transparenz und Rückverfolgbarkeit der Teile sowie an prädiktiver Analytik. [3]

2. Produktive Partnerschaften: Globale Lieferketten haben Kommunikationsprobleme durch unterschiedliche Standards und Software, was die Effizienz beeinträchtigt. Eine bessere Lieferantenbeziehung und Echtzeitkommunikation sind notwendig. [3]

3. Umweltbelange: Die Branche muss sich auf grüne Initiativen konzentrieren, um Umweltauflagen zu erfüllen. Dies erfordert Anstrengungen bei der Auswahl von Partnern, die umweltfreundliche Lösungen bieten. [3]

4. Auswirkungen von Covid-19: Die Pandemie hat zu erheblichen Störungen geführt und die Unsicherheit in Bezug auf die Widerstandsfähigkeit der Lieferketten verstärkt. Verbesserte Kommunikation und Flexibilität sind notwendig, um auf zukünftige Ver-sorgungsprobleme vorbereitet zu sein. [3]

5. Zunehmende Bürokratie: Brexitbedingte Umstrukturierungen und Zollbürokratie haben zu längeren Lieferzeiten und allgemeiner Ineffizienz geführt. Es bedarf einer optimierten Strategie, um die Einhaltung von Vorschriften zu gewährleisten und Verzögerungen zu minimieren. [3]

Diese Herausforderungen erfordern verstärkte Kommunikation, verbesserte Transparenz, flexible Prozesse und eine optimierte Lieferantenbeziehung, um die Lieferketten widerstandsfähiger und effizienter zu gestalten. [3]
Die Integration von Digitalisierung, insbesondere durch die Einführung eines digitalen Zwillings, bietet entscheidende Lösungen für die genannten Herausforderungen. Ein digitaler Zwilling ermöglicht eine präzise Überwachung und Rückverfolgbarkeit aller Komponenten in Echtzeit, verbessert die Prognosegenauigkeit und unterstützt die Kommunikation entlang der gesamten Supply Chain. Darüber hinaus kann er als Testfeld dienen, um potenzielle Probleme vorab zu identifizieren, was letztendlich zu effizienteren Abläufen und einer Reduzierung von Risiken führt. Diese Technologie ist somit ein entscheidendes Werkzeug, um die Komplexität der Supply Chains zu bewältigen und ihre Widerstandsfähigkeit zu stärken. [4]

Einführung in den digitalen Zwilling

Ein digitaler Zwilling basiert auf Daten aus verschiedenen Quellen. Diese Daten werden zur Erstellung eines virtuellen Modells verwendet. Dieses Modell ermöglicht eine digitale Abbildung von physischen Abläufen, Prozessen und Anlagen. Das Modell ermöglicht es, unterschiedliche Szenarien zu untersuchen, Anpassungen zu planen und Strategien zu entwickeln, die bereits in der realen Welt angewandt werden. Der digitale Zwilling überwacht die realen Abläufe in Echtzeit und warnt den Betreiber, wenn etwas nicht wie erwartet läuft. Dadurch können die Betreiber sofort Anpassungen vornehmen. [4]

Anwendung des digitalen Zwillings in der Automobil-Supply-Chain

In der Logistik werden umfassende Daten aus unterschiedlichen Quellen erfasst, sowohl Echtzeit- als auch historische Daten. Diese bilden die Basis für einen

digitalen Zwilling, der die physischen Logistikabläufe, Prozesse und Systeme repräsentiert. Dieser ermöglicht die Simulation verschiedener Szenarien und erlaubt so das Testen und Optimieren von Logistikstrategien in einer virtuellen Umgebung, noch bevor sie in der Realität implementiert werden. [4]
Durch Echtzeitüberwachung vergleicht der digitale Zwilling die tatsächlichen Logistikabläufe mit seinem virtuellen Modell und signalisiert sofortige Abweichungen oder Ausnahmen. Durch diese Methode ist es für Betreiber möglich, umgehende Korrekturen durchzuführen und ihre Effizienz in der Supply Chain zu verbessern. [2]
Der Einsatz eines digitalen Zwillings in der Automobil-Supply-Chain ermöglicht eine Vielzahl von Optimierungen. Er überwacht den Lagerbestand in Echtzeit, optimiert Wartungspläne, verbessert Transportplanung und trägt zur Steigerung der Sicherheit bei, indem er potenzielle Risiken frühzeitig erkennt. Er bietet vorausschauende Analysen zur Produktionskapazität, Kraftstoffeffizienz und Ressourcenverteilung an, um Kosten zu senken und Ressourcen effizient zu nutzen. [2]
Die Abbildung 1 zeigt ein Konzeptdiagramm eines digitalen Lieferkettenzwillings. Es illustriert, wie eine cloudbasierte, datengestützte Kopie der realen Lieferkette fortschrittliche Analysen und Simulationen ermöglichen kann, um die Lieferkette zu optimieren. Diese Kopie entspricht einem digitalen Zwilling der Supply Chain. Das Diagramm zeigt auch die Einflüsse und Wechselwirkungen innerhalb des Systems, wobei Daten ein zentrales Element sind:

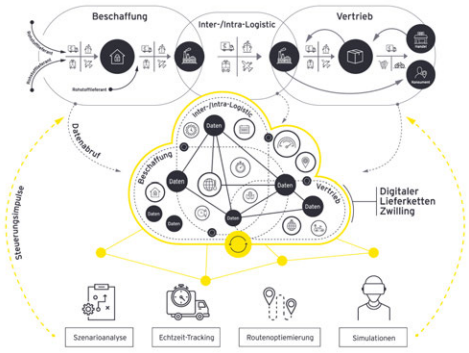


Abb. 1: Konzept eines digitalen Supply Chain Zwillings und deren Bestandteile [2]

Vorteile des Digitalen Zwillings für die Automobil-Supply-Chain

Die Anwendung des digitalen Zwillings in der Automobil-Supply-Chain bietet eine Vielzahl von Vorteilen:

1. **Effizienzsteigerung in der Lieferkette:** Der digitale Zwilling überwacht in Echtzeit Lagerbestände und Produktverfügbarkeit, reduziert Lagerbestände, prognostiziert Wartungsarbeiten, optimiert Transportrouten und maximiert die Produktivität. [4]
2. **Risikominimierung und Sicherheit:** Der digitale Zwilling erkennt potenzielle Risiken frühzeitig und ergreift präventive Maßnahmen. Dadurch werden Risiken reduziert und die Sicherheit erhöht. [1]
3. **Optimierung von Ressourcen und Nachhaltigkeit:** Vorausschauende Analysen prognostizieren die Nachfrage, planen Kapazitäten und senken Kosten, was zu nachhaltiger Ressourcenverteilung und Emissionsreduzierung führt. [4]

Abbildung 2 zeigt, dass die Verwendung digitaler Zwillinge in der Lieferkettenverwaltung erhebliche Vorteile bietet. Unternehmen, die digitale Zwillinge integrieren, erleben verbesserte Agilität, schnellere Entscheidungsfindung und erzielen finanzielle Gewinne. Die Analyse zeigt, dass sie ein jährliches Umsatzwachstum von 1–2% ermöglichen, die Planungseffizienz um 10–30% steigern und den Lagerbestand um 5–10% reduzieren können. Dazu kommen noch weitere abgebildeten Vorteile einen digitalen Zwilling in der Automobil-Supply-Chain zu verwenden.

Abb. 2: Signifikante Vorteile durch den digitalen Zwilling [2]

Ausblick und Zukunftsaussichten

Der verstärkte Einsatz digitaler Zwillinge wird die Zukunft der Logistik maßgeblich beeinflussen. Unabhängig von der Größe des Unternehmens kann diese Technologie eine essentielle Komponente in ihrem Logistikmanagement darstellen. Der digitale Zwilling bietet die Möglichkeit, flexiblere und effizientere Lieferketten aufzubauen. [4]
Pilotprojekte bieten eine gute Möglichkeit, um den Einstieg in die digitale Logistik zu erleichtern. Beginnend mit einem abgegrenzten Teil der Lieferkette, unter anderem einem Lagerstandort, ermöglichen sie Unternehmen, sich mit der Anwendung und den Prozessen digitaler Zwillinge vertraut zu machen. Durch eine kontinuierliche Optimierung von Daten, von der Bereitstellung von Daten über die Nutzung in weiteren Bereichen bis hin zur Skalierung, kann ein Unternehmen seinen Reifegrad kontinuierlich erhöhen und das volle Potential digitaler Zwillinge ausschöpfen. [2]

Literatur und Abbildungen

[1] Databricks Inc. Digital Twin. https://www.databricks.com/glossary/digital-twin, 2023.

[2] Maximilian Kroh. Mit dem digitalen Zwilling in die Logistik von morgen. https://www.ey.com/de_de/consulting/mit-dem-digitalen-zwilling-in-die-logistik-von-morgen, 08 2022.

[3] Data Interchange Limited. 5 Herausforderungen in der Automobilzulieferkette für die Branche. https://datainterchange.com/de/automotive-supply-chain-challenges/, 03 2023.

[4] Janine Wolff. Doppelt hält besser: Der Digitale Zwilling in der Logistik. https://www.saloodo.com/de/blog/doppelt-haelt-besser-der-digitaler-zwilling-in-der-logistik/, 05 2023.

Entwicklung und Implementierung von Software-Diagnose Applikationen im Bereich der Robotik

Alexander Feuchter **Reiner Marchthaler**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma Festo SE & Co. KG, Berkheim

Motivation

Die fortschreitende Weiterentwicklung der Fertigungsindustrie in den letzten Jahren wurde durch die Vernetzung und Automatisierung von Maschinen geprägt. Ein besonders schnell anwachsender Bereich der Automatisierungsbranche stellt dabei die Robotik dar, insbesondere kollaborierende Roboter wie in Abbildung 1 zu sehen, die in eigenständiger bzw. unterstützender Funktion Arbeitsprozesse erleichtern. Das Interesse an sogenannten Cobots (aus dem englischen abgeleitet von collaborative robot) ist nicht zuletzt durch Faktoren, wie wirtschaftliche Änderungen auf dem Absatzmarkt oder das Konsumverhalten von Endkunden, sondern auch auf den demografischen Wandel zurückzuführen, der in allen großen Industrienationen eine Problematik ist.

Abb. 1: Pneumatischer Cobot der Fa. Festo [3]

Cobots stellen eine Möglichkeit dar, nicht nur den Fachkräftemangel zu entschärfen, sondern Berufsgruppen bzw. ganze Branchen für jüngere und zukünftige Generationen wieder attraktiver zu gestalten [5]. Damit ein problemloses Arbeiten zwischen Mensch und Maschine möglich ist, werden über die Robotersteuerung eingebettete Funktionen für die Sicherheit und Bedienung integriert. Diese Funktionen, wie beispielsweise Kollisionsberechnungen oder die Diagnose von Maschinendaten des Cobots, müssen im Vorfeld in Testaufbauten an leistungsstarken echtzeitfähigen Computern getestet und angepasst auf die Steuerungseinheit übertragen werden. Hier muss zwischen zwei oder mehreren Programmiersprachen eine Konvertierung oder Code-Generierung stattfinden.

Problemstellung

Ein allgemeines Vorgehen bei ingenieurswissenschaftlichen Lösungsansätzen ist die Abstraktion bzw. Simplifikation der Problematik in einem Model. Dabei werden mathematisch komplexe Berechnungen über eine modellbasierte Darstellung auf Teilstrukturen reduziert. Weit verbreitet sind Applikationen, wie Simulink von MathWorks oder auch z.B. im zunehmenden Open-Source Bereich, BMS (Block Model Simulator) in Python. Aufgrund der hohen Verwendung in Forschung und Entwicklung, wird im Folgenden lediglich Bezug auf Simulink und die damit verbundenen MathWorks Derivate eingegangen. Nach dem Erstellen eines Models, erfolgt die Testphase. Diese kann rein simulativ an einem Rechner oder über ein HIL-Testing (Hardware in the Loop) direkt auf einer Hardware ablaufen. Hierbei sind die Testabläufe bzw. das generelle Berechnen von Modelparametern sehr rechenintensiv und muss zum Teil auf zusätzlicher externer Hardware ablaufen, wie in Abbildung 2 zu sehen. Im Falle einer Robotiksteuerung muss folglich der modelbasierte Entwurf in eine adäquate Programmiersprache, wie Structured-Text, transferiert und angepasst werden, um die Lauffähigkeit auf einem eingebetteten System zu gewährleisten. Die Herausforderung ist dabei nicht nur die normgerechte Übersetzung nach IEC 61131-3, sondern das Entwickeln einer Codegenerierungsmethode für zukünftige Erweiterungen des Robotersystems. Die Codegenerierung sollte dabei so einfach wie möglich erfolgen, den Anwender in seinem Modellentwurf nicht einschränken und keine vertieften Programmierkenntnisse erfordern.

Abb. 2: HIL-Testing Rack [2]

Lösungsansatz

Ausgehend von einem in Simulink erstellten modellbasierten Regelungsentwurf, bieten sich zwei mögliche Lösungsansätze für die Codegenerierung an. Die beiden infrage kommenden Codegeneratoren sind zum einen der PLC-Coder und zum anderen der Embedded-Coder, welche beide von MathWork bereitgestellt werden. Die Codegeneratoren haben unterschiedliche Vor- und Nachteile, die entscheidend für die endgültige Auswahl sind. Der PLC-Coder besitzt die Möglichkeit in Simulink erstellte Modelle direkt nach IEC 61131-3 in Structured-Text zu transferieren. Die erstellten Dateien können anschließend in eine gewünschte IDE, wie von CODESYS oder Rockwell Automation, überführt bzw. durch ein Postscripting noch angepasst werden. Ein Problem stellt dabei die Flexibilität des PLC-Coder dar. Die bereitgestellte Bibliothek besteht zwar aus mehreren Standardblöcken aus Simulink, jedoch sind diese bei komplexen mathematischen Berechnungen unzureichend. Des Weiteren sind in Matlab hinterlegte Funktionen, wie z.B. für die Regressorberechnung oder S-Funktionen, nicht mit dem PLC-Coder verwendbar. Damit wird ein modellbasierter Reglungsentwurf stark eingeschränkt und kann nur auf Basis der bereitgestellten Blöcke der PLC-Coder-Bibliothek aufgebaut werden. Alternativ besteht die Möglichkeit weitere Blöcke von Hand der Bibliothek hinzuzufügen, jedoch ist dieser Ansatz konträr zu der ursprünglichen Prämisse, einer offenen und zugänglichen Codegenerierung. Der Embedded-Coder bietet im Gegensatz zum PLC-Coder die Möglichkeit, eine große Auswahl an Simulinkblöcken und Matlab Funktionen zu generieren. Er ist nicht beschränkt in seiner Verwendung und kann nahezu alle in der Simulinkbibliothek hinterlegten Blöcke verwenden, sowie auch S-Funktionen einbinden. Der Embedded-Coder ist zudem eine weitverbreitete Anwendung im industriellen Umfeld, wie in der Automobil oder Flugbranche und wird stetig weiterentwickelt und optimiert. Zusätzlich können mit dem Embedded-Coder die bereits integrierten Pre-/Post-Processing Funktionalitäten genutzt werden, um den erzeugten Programmcode an die gewünschte Zielhardware anzupassen. Eine Optimierung der Ausführgeschwindigkeit oder des RAM-Verbrauchs ist ebenfalls verfügbar. Im Zusammenhang mit der Problemstellung ergibt sich mit dem Embedded-Coder jedoch der Nachteil, dass dieser C/C++ Programmiercode und keinen Structured-Text erzeugt. Die verwendete Robotersteuerung baut aber auf einem UNIX-System auf, infolgedessen können die in C generierten Programme als eine Shared-Object-Datei (SO) in der Steuerung hinterlegt und aufgerufen werden. Die Integration der in den SO-Dateien hinterlegten Programme erfolgt über Export-XML Dateien, welche Funktionsblöcke oder einfachen Funktionen in Structured-Text enthalten. Der Embedded-Coder erfüllt damit am besten die Anforderungen der Problemstellung, die geforderte Flexibilität und Modularität für die Codegenerierung eines modellbasierten Reglungsentwurfs wird eingehalten.

Aufbau und Ablauf

Die verwendeten Bestandteile der Hardware gliedern sich wie folgt auf:

- Die Robotereinheit: Ein Roboterarm bzw. Cobot mit 6 Freiheitsgraden (Six Degrees of Freedom, kurz 6DoF)
- Die Leiterplatinen der einzelnen Gelenke: Die Parameterdaten der einzelnen Gelenke werden über diese Leiterplatinen gemessen und weitergegeben
- Der interne Kommunikationskanal: Verwendung eines Echtzeitfähigen Feldbusses, wie EtherCAT, Profinet oder SERCOS, zur Übertragung der Parameterdaten zwischen den Hardwarekomponenten
- Die Robotiksteuerung: Eine echtzeitlauffähige Steuerung mit integriertem UNIX-Betriebssystem

Im Lösungsansatz wurde bereits auf die Möglichkeiten für die Anwendungen des Embedded-Coder eingegangen. Die nachfolgende Beschreibung soll dabei

die Umsetzung und Integration auf der Hardware schildern. Nachdem der Modelentwurf in Simulink erstellt ist, kann über den Embedded-Coder das Programm generiert werden. Es können sowohl vor, als auch nach der Codegenerierung benötigte Post-Processing Schritte vorgenommen werden, wie z.B. eine Datentypen Anpassung nach IEC 61131-3 [1]. Falls eine Optimierung auf eine bestimmte Hardware oder Entwicklungsumgebung wie CODESYS benötigt wird, muss diese bereits vor der Generierung eingestellt sein. Der in C generierte Programmcode wird nun über den Embedded-Coder oder eine andere Software als ein Shared-Object generiert. Analog dazu kann die Anfertigung einer Export-XML Datei beginnen, die später bei der Einbindung und Aufrufes des C-Programmes aus der SO-Datei benötigt wird. In diesen XML Dateien werden die Funktionsaufrufe des C-Programmes, aber auch die Übergabevariablen in Form einer IEC 61131-3 gerechten Program Organization Unit, kurz POU deklariert. Nach dem Erstellen dieser beiden Dateien, sind diese auf die Steuerung zu übertragen. Innerhalb der angelegten POUs, wird der C-Code nun als IEC-Baustein geladen und auf der Steuerung mit der geforderten Taktrate und Priorität aufgerufen. Die Echtzeitparameter des Roboters können nun über das Feldbussystem von den einzelnen Leiterplatinen der Gelenke übertragen und von den auf der Steuerung integrierten Programmen ausgeführt werden.

Fazit und Ausblick

Das ausgearbeitete Vorgehen für die Integration eines modellbasierten Regelungsentwurfes auf einer Robotiksteuerung ermöglicht eine einfache und schnelle Methode, neue Diagnose-Applikationen für einen Roboter zu entwickeln. Dabei können Diagnosefunktionen, wie z.B. zur Bahntreue, Bauteilalterung, Regeldynamik oder einer Lastmassenschätzung [4] in einem Rapid-Prototyping-Verfahren auf eine Steuerung übertragen und ausgeführt werden. Ebenfalls kann ein Test der Steuerung und ihrer Auslastung unter Verwendung neu integrierter Programme erfolgen, ohne einen speziellen Versuchsaufbau zu benötigen. Die aus der Problemstellung resultierenden erforderlichen Eigenschaften, wie eine leicht anzuwendende Übertragung von Reglungsentwürfen aus Simulink oder der normgerechten Integration nach IEC 61131-3, wurden somit erfüllt. Ausblickend besteht die Möglichkeit die noch bleibenden manuellen Schritte, wie die Erstellung einer angepassten Export XML Datei oder die Erstellung und Integration der SO-Dateien auf der Steuerung, noch weiter zu automatisieren.

Literatur und Abbildungen

[1] DIN Deutsches Institut für Normung. *Speicherprogrammierbare Steuerungen – Teil 3: Programmiersprachen (IEC 61131-3:2013) DIN EN 61131-3*. Beuth Verlag GmbH, 2013.

[2] Marketing dSPACE GmbH. Das maßgeschneiderten SCALEXIO Rack-Systeme werden für jedes Projekt einzeln aufgebaut und konfiguriert. https://www.dspace.com/de/gmb/home/products/hw/simulator_hardware/scalexio.cfm, 2023.

[3] Marketing Festo SE und Co KG. Der weltweit erste pneumatische Cobot. https://www.festo.com/de/de/e/ueber-festo/forschung-und-entwicklung/festo-cobot-id_1379474/, 2023.

[4] Kathrin Hoffmann, Christian Trapp, Alexander Hildebrandt, and Oliver Sawodny. Force/Torque Sensor-Free Online Payload Estimation for a Pneumatically Driven Robot, 2022.

[5] Astrid Weiss, Ann-Kathrin Wortmeier, and Bettina Kubicek. Cobots in Industry 4.0: A Roadmap for Future Practice Studies on Human–Robot Collaboration. In *Transactions on Human-Machine Systems*, volume 51, pages 335–345. IEEE, 2021.

Konzeption und Implementierung einer spezialisierten unsicheren Webanwendung für Demonstrationen, Einstellungstests und Schulungen

Maximilian Fink **Dominik Schoop**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma ETAS GmbH, Stuttgart

Motivation

Die zunehmende Durchdringung von IT-Systemen in sämtlichen Lebensbereichen unterstreicht die Notwendigkeit, die Sicherheit dieser Systeme zu gewährleisten. Infolgedessen hat die IT-Sicherheit signifikant an Bedeutung gewonnen und ist zu einem Kernelement in modernen Unternehmensstrukturen geworden. Trotz dieser wachsenden Bedeutung gestaltet sich die Rekrutierung qualifizierter Fachkräfte in diesem Bereich als äußerst anspruchsvoll.

Die Abteilung PSS-Enterprise der ETAS GmbH agiert als interner Dienstleister für Enterprise IT Security Services innerhalb der Bosch-Gruppe. Mit einem Schwerpunkt auf Penetration Testing befindet sich die Abteilung kontinuierlich auf der Suche nach weiteren Penetration Testern. Es hat sich als komplex erwiesen, die Qualifikationen eines Bewerbers in kurzer Zeit zu evaluieren, insbesondere in einem traditionellen Bewerbungsgespräch.

Als Lösungsansatz ist die Konzeption und Entwicklung einer speziell auf Bosch beziehungsweise ETAS zugeschnittenen, anspruchsvollen und bewusst verwundbaren Webanwendung vorgesehen. Diese Anwendung soll als Einstellungstest für Bewerber im Bereich Penetration Testing dienen. Im Rahmen des Bewerbungsverfahrens erhalten die Bewerber Zugang zu dieser Anwendung, um ihre Fähigkeiten unter Beweis zu stellen. Die Ergebnisse müssen daraufhin von diesem in einem Bericht oder einem Interview präsentiert werden.

Darüber hinaus bietet die Anwendung vielfältige Anwendungsmöglichkeiten, darunter Demonstrationen und Präsentationen im unternehmenseigenen Umfeld sowie auf Konferenzen, Messen oder an Hochschulen. Auch stellt sie eine optimale Testumgebung für die Erprobung neuer Tools und Methoden dar. Nicht zuletzt kann die Anwendung im Rahmen des von der Abteilung als Service angebotenen Web-Security-Trainings eingesetzt werden. Im Vergleich zur bisherigen Nutzung des OWASP Juice Shop verspricht sie hierbei eine höhere Flexibilität und Professionalität. Die umfassende Konzeption sowie die initiale Entwicklung dieser Webanwendung bilden den Inhalt der Bachelorarbeit.

Anforderungen

Zunächst werden die Anforderungen an die Webanwendung untersucht, um im nächsten Schritt die Anwendung zu konzipieren:

- **Realitätsnähe:** Schwachstellen sollen geschickt in eine funktionale Anwendung integriert sein. Dies ermöglicht den Bewerbern, die Anwendung zu erkunden, Schwachstellen zu identifizieren, passende Exploits als Nachweis dieser zu konstruieren sowie die Schwachstellen umfassend zu dokumentieren und zu bewerten. Damit wird eine authentische und praxisnahe Umgebung geschaffen, in der die Fähigkeiten der Bewerber auf eine realistische Weise getestet werden können.
- **Integration spezifischer Schwachstellen:** Die Auswahl der Schwachstellen wird anhand verschiedener Kriterien getroffen. Ein besonderes Augenmerk liegt dabei auf komplexeren, aber trotzdem im Unternehmensumfeld häufigen und kritischen Schwachstellen, die sich nicht leicht durch automatisierte Scanner aufdecken lassen. Hierbei sind Business-Logic-Schwachstellen relevant, ebenso wie Szenarien, die die Verknüpfung mehrerer kleinerer Schwachstellen erfordern. Zudem werden API-Schwachstellen in die Auswahl einbezogen. Ein weiterer Aspekt ist die Abstimmung mit erfahrenen Pentestern. Diese Zusammenarbeit gewährleistet, dass die ausgewählten Schwachstellen nicht nur theoretisch relevant, sondern auch in der praktischen Anwendung bedeutsam sind.
- **Exklusivität:** Um die Sinnhaftigkeit des Einstellungstests zu gewährleisten, muss die Anwendung exklusiv bleiben und es dürfen keine öffentlich zugänglichen

Lösungen kursieren. Auch stärkt eine eigene und spezialisierte Awendung das professionelle Auftreten nach außen.
- **Flexibilität und Anpassbarkeit:** Die konzipierte Anwendung soll von Beginn an auf die genauen Bedürfnisse angepasst werden und auch für die Zukunft flexibel gestaltet sein. So sollen auch Schwachstellen ausgetauscht oder weitere hinzugefügt werden können.
- **Branding:** Die Webanwendung soll im Unternehmensdesign von Bosch oder ETAS sein, um ein einheitliches und markenspezifisches Erscheinungsbild zu gewährleisten. In der Konfiguration soll das Design gewählt werden können.
- **Einfaches Cloud Deployment:** Die Webanwendung soll mit minimalem Aufwand in der Cloud bereitgestellt werden können, dies auch in mehrfachen Instanzen.

Vorgehensweise

Nachdem die spezifischen Anforderungen identifiziert wurden, erfolgt eine detaillierte Analyse bestehender Anwendungen, um Einblicke in ihre Schwachstellen, Umsetzung, Anpassungsfähigkeit, Integration von Schwachstellen und Verbesserungsmöglichkeiten zu bekommen sowie Inspiration zu gewinnen. Vier unterschiedliche Anwendungen werden hierbei untersucht. Zunächst steht die Analyse des Open-Source-Projekts OWASP Juice Shop an, das eine Vielzahl von Schwachstellen abdeckt und vor allem für Einsteiger zum Lernen konzipiert ist [4]. Im Anschluss wird PyGoat mit Django als Basis betrachtet, wobei die Codeanalyse hier von besonderem Interesse ist, da Django in der engen Auswahl für die eigene Anwendung steht [5]. Eine andere Herangehensweise bietet Hack The Box, eine Capture-the-Flag-Umgebung, bei der das Ziel darin besteht, Schwachstellen auszunutzen, um digitale Flags zu finden [1]. Abschließend wird die PortSwigger Web Security Academy analysiert, die von Dafydd Stuttard konzipiert wurde und als interaktive Lernplattform dient [6]. Hier sind besonders viele moderne und komplexe Schwachstellen zu finden, jedoch pro Übungsumgebung lediglich eine Schwachstelle.
Im darauf folgenden Schritt erfolgt die Auswahl geeigneter Schwachstellen und Angriffszenarien. Hierbei fließen nicht nur die Erkenntnisse aus der Analyse bestehender Anwendungen ein, sondern auch die Interviews mit erfahrenen Pentestern und Personen, die vergleichbare Projekte bereits umgesetzt haben. Darüber hinaus werden zahlreiche Pentest Reports aus der Vergangenheit analysiert und die Anforderungen der zu entwickelnden Anwendung berücksichtigt, um eine passende und interessante Auswahl zu treffen. Diese wird aus offensichtlichen Gründen hier nicht näher aufgeführt.
Die Konzeption legt den Fokus auf die Architektur und Designüberlegungen. Dabei spielen die Auswahl der Programmiersprachen und Frameworks eine entscheidende Rolle, sowie die das Cloud Deployment und die Anpassbarkeit der Anwendung. Die Wahl fiel auf das Python-basierte Web-Framework Django in Kombination mit einer SQLite-Datenbank. All dies wird in einem Docker-Container hinter nginx bereitgestellt. Der Vorteil dieser Technologien besteht darin, dass alles in einem einzelnen Container leicht deployed werden kann. Da die Datenbank eine einzelne Datei in dem Container ist, kann diese einfach mit Daten vorbereitet und im Falle eines Problems wieder auf den Anfangszustand zurückgesetzt werden. Django als grundlegendes Web-Framework bringt viele Vorteile und Funktionen mit, wie Security-Funktionen, eine Object Relational Mapping (ORM) Funktionalität oder eine vorgefertigte Nutzerverwaltung [3]. All dies trägt dazu bei, dass man in limitierter Zeit eine möglichst umfangreiche Anwendung entwickeln kann. Auch die Festlegung des Datenmodells sowie die Einbindung des Designs und der GUI-Elemente des Unternehmens müssen geplant werden. Bosch bietet ein Frontend-Kit, das bei dieser Problemstellung unterstützt. Dieses ist in einer internen npm Registry erhältlich. Die Konzeption beinhaltet ebenfalls die Erarbeitung einer Dokumentation, um eine leichte Anwendung, Administration und Weiterentwicklung zu ermöglichen. Die Wahl der simulierten Umgebung ist beeinflusst von der Auswahl der Schwachstellen. Diese gibt notwendige Funktionen vor, wie beispielsweise eine CRUD-API, eine Import- und Export-Funktion und eine Nutzerverwaltung mit Registrierung, Passwort-Reset und Profilbild-Upload. Auch muss es möglich sein, die Nutzernamen anderer Personen herauszufinden. All dies führte zur Wahl eines Inventarverwaltungssystems mit dem Namen "Bosch Code Chaos". Dabei können Dinge im Inventar angelegt, angesehen, bearbeitet, gelöscht und auch mit verschiedenen Gruppen geteilt werden.

Implementierung und Ausblick

Nach der sorgfältigen Planung kann die Implementierung beginnen. Diese umfasst einen großen Teil der Arbeit und erfolgt in großen Teilen schon parallel zur Konzeption. Aktuell ist die Implementierung in Arbeit und einige Angriffsszenarien sind bereits fertig. Auf der Abbildung 1 kann man die Startseite der Inventarverwaltung sehen. Diese ist nach einem erfolgreichen Login zugänglich und zeigt alle Inventargegenstände, auf die der Nutzer Zugriff hat.

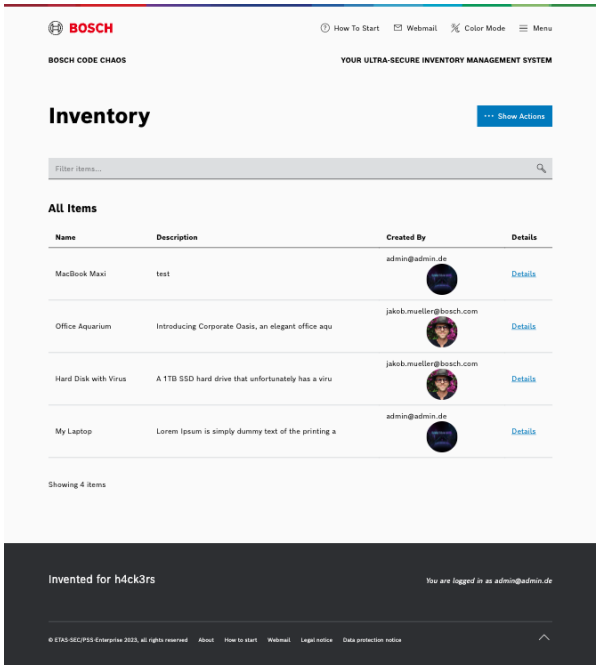


Abb. 1: Übersicht des Inventars im Bosch Designmodus [2]

Abbildung 2 zeigt die Importfunktion der Anwendung. In diesem Fall ist das ETAS-Design aktiviert.

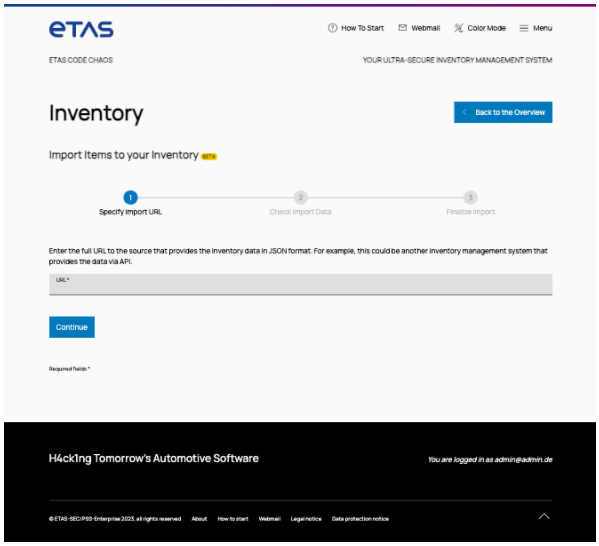


Abb. 2: Importfunktion der Anwendung im ETAS Designmodus [2]

Im nächsten Schritt muss die Implementierung abgeschlossen werden. Anschließend folgt ein beispielhafter Penetration Test, um ungewollte Schwachstellen auszuschließen und sicherzustellen, dass die geplanten Schwachstellen nur zu einem gewissen Grad Daten der Anwendung offenbaren. Dies könnte sich noch als große Herausforderung herausstellen, da existierende Schwachstellen kaum eingeschränkt werden können. Sobald eine Funktion in der Anwendung verwundbar ist, ist es meist sehr einfach, die verbleibenden Schutzmechanismen ebenfalls zu umgehen. Schlussendlich muss die Anwendung für das Cloud Deployment in Microsoft Azure vorbereitet werden, um die Anwendung bereit für den Einsatz und die ersten Bewerber zu machen.

Literatur und Abbildungen

[1] Hack The Box. Hack The Box: Hacking Training For The Best. https://hackthebox.com, 2023.

[2] Eigene Darstellung.

[3] Django Software Foundation. Django documentation. https://docs.djangoproject.com/en/5.0/, 2023.

[4] Björn Kimminich. *Pwning OWASP Juice Shop*. Open Worldwide Application Security Project, 2023.

[5] Open Web Application Security Project. PyGoat. https://github.com/adeyosemanputra/pygoat/tree/master, 2023.

[6] Dafydd Stuttard. Web Security Academy - Credits. https://portswigger.net/web-security/credits, 2023.

Analyse von KI-basierten Mustererkennungen zur Validierung vernetzter Fahrzeugsysteme

Tim Georg **Markus Enzweiler**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma Mercedes-AMG GmbH, Affalterbach

Motivation

In der heutigen Ära der Automobilindustrie sind Softwarekomponenten und deren Kommunikation innerhalb des Fahrzeugs von entscheidender Bedeutung, um moderne Funktionen, wie Assistenzsysteme, Fahrdynamik-Features und fortschrittliche Infotainment-Systeme zu realisieren. Um diese zunehmenden Datenmengen sicher im Fahrzeug zu transportieren, haben sich verschiedene Netzwerkprotokolle wie LIN, CAN, FlexRay und Ethernet etabliert, die jeweils spezifische Anforderungen an die Kommunikation adressieren [9]. In diesem Trend, vom Kabelsatz, über ein Bussystem, zum Zentralrechner steigt auch die Komplexität der darüberliegenden Protokollschichten und Absicherungsmechanismen. In der folgenden Grafik aus [2] sind diese Fahrzeugtopologien exemplarisch dargestellt.

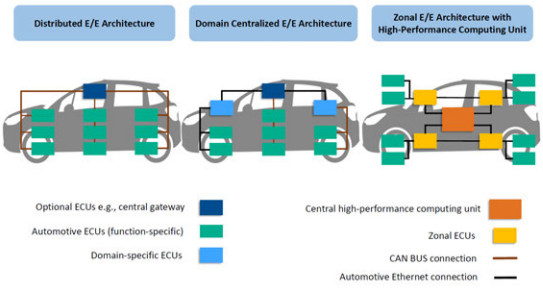


Abb. 1: Entwicklung der E/E-Architektur [2]

Zur Absicherung der fahrzeuginternen Kommunikation existiert neben einer Teststrategie und definierter Testfälle ebenso eine Dauermessdatengenerierung durch Datenlogger. Mit der Analyse dieser Messungen können Anomalien im Dauerbetrieb und seltene Fehler in kundennahen Szenarien gefunden werden. Die Herausforderung besteht darin, diese Datenmengen effektiver zu nutzen und tiefgehende Analysen zu erstellen.

Zielsetzung

Der Kern dieser Arbeit beschäftigt sich mit der Hypothese, ob Machine Learning Vorteile im Vergleich zur klassischen Modellierung durch algorithmische Abfragen bietet. In erster Linie unterliegt diese Arbeit dem Kontext der Systemvernetzung und Steuergerätekommunikation in der Automobilentwicklung. Durch die Untersuchung verschiedener Machine Learning Ansätze soll ein Verständnis für deren Einsatzmöglichkeiten und Einschränkungen im Sinne der Systemvernetzung im Fahrzeug erlangt werden. Ebenfalls sollen die grundlegenden Funktionsweisen und Anwendungsgebiete durch Literatur und weitere Branchen dargestellt werden. Mit der Erforschung dieser Einschränkungen und Vorbedingungen strebt diese Arbeit an, einen Überblick über die Anwendbarkeit von KI-Techniken zu liefern und damit zur Strategieentwicklung für Big Data & AI beizutragen, indem belegt oder widerlegt wird, dass diese zur Validierung von vernetzten Fahrzeugsystemen genutzt werden können. Denn aufgrund hoher Anforderungen an eine zuverlässige Kommunikation und Spezifikationen wie Zykluszeiten, Sende- und Abschaltverhalten, sowie klar definierter Syntax der Infrastruktursignale, weist ein vernetztes System überwiegend deterministische Eigenschaften auf.

Zeitreihenanalyse

Für die Analyse von Zeitreihen (in dieser Arbeit Infrastruktursignale) haben sich Rekurrente Neuronale Netze (RNN) etabliert. Eine besondere Form der RNN ist das Long-Short-Term-Memory Netz (LSTM), was zu den State-of-the-Art-Methoden im Deep Learning gehört. Sie haben den Vorteil, dass sie längerfristige Abhängigkeiten zu vergangenen Werten besser modellieren können, weil sie weniger anfällig für das Problem von verschwindenden Gradienten sind [7]. Einen alternativen Machine Learning Ansatz für die Verarbeitung von Sequenzen bietet die Transformer-Architektur, welche hauptsächlich für die Verarbeitung

von natürlicher Sprache genutzt wird. Da Transformer nicht nur für die Verarbeitung von Textsequenzen eingesetzt werden können, sondern auch für andere Arten von sequenziellen Daten, steigt das Interesse für die weitere Adaption von Transformern [1]. LSTM-Netze werden verwendet für das Monitoring von Web-Traffic [6], Prüfen von CAN-Nachrichten auf Malware-Injektion [5] oder für die Anomalie-Erkennung und Vorhersage von Sensordaten [7]. Zwei Ziele, die mittels LSTM erreicht werden können, sind Klassifikation von Sequenzen und die Erkennung von Anomalien innerhalb von Sequenzen. Zur Kategorisierung von Sequenzen in verschiedene Klassen wird ein LSTM Classifier verwendet, welcher durch Supervised Learning mit Daten aus Sequenzen und den zugehörigen Labels trainiert wird. Zur Erkennung von Anomalien wird ein LSTM-Autoencoder verwendet, der durch Unsupervised Learning aus einem Datensatz ohne Labels nur anhand der Sequenzen trainiert wird. Hier könnte man auch von Semi-Supervised Learning sprechen, wenn man von einem fehlerfreien Eingabedatensatz ausgeht. Das Ziel des Netzwerks besteht darin, den Signalverlauf aus den Trainingsdaten zu lernen, um es reproduzieren zu können. Hierfür wird als Eingabe eine Sequenz einer bestimmten Länge festgelegt und dieselbe Sequenz als Ausgabeziel festgelegt [7]. Der prinzipielle Aufbau eines LSTM-AE nach [8] [4] ist in der folgenden Grafik dargestellt.

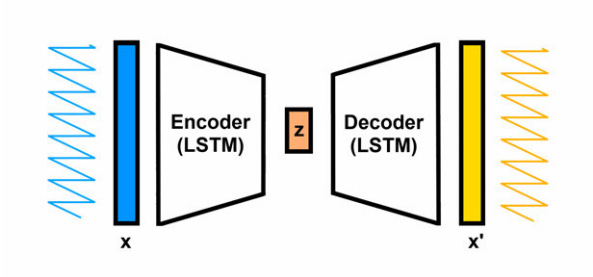


Abb. 2: Konzeptueller Aufbau eines LSTM-AE [3]

Durch einen Vergleich der vorhergesagten Sequenz mit einer Eingabesequenz können Anomalien identifiziert werden. Weicht der erwartete Wert zu stark von dem tatsächlichen Wert ab, gilt dieser Punkt in der Eingabesequenz als eine Anomalie. In Abbildung 3 wird schematisch dargestellt, wie eine Anomalie durch den Ausschlag des Rekonstruktionsfehlerwerts über einen Schwellwert erkannt wird.

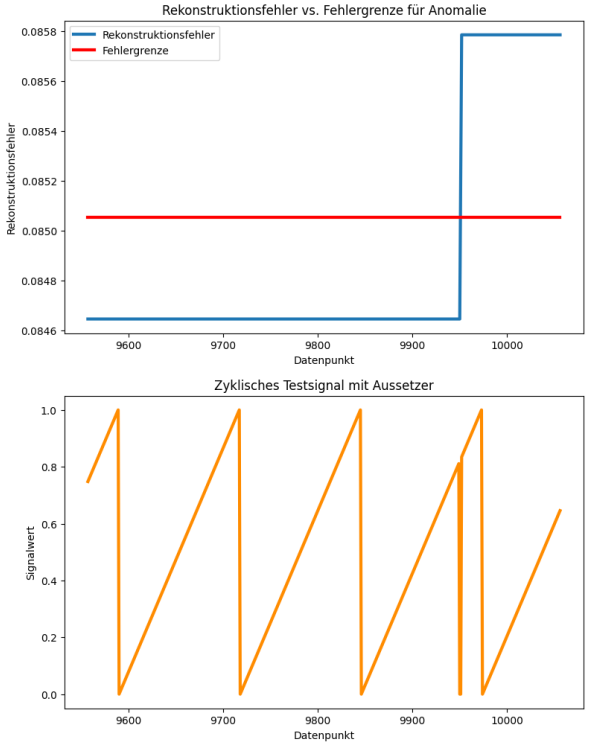


Abb. 3: Testsignal mit Anomalie und zugehöriger Rekonstruktionsfehler [3]

Datengrundlage

Die Präzision und Generalisierungsfähigkeiten eines Machine Learning Modells hängt stark vom unterliegenden Datensatz ab. Eine geeignete Netzarchitektur kann keine guten Ergebnisse erzielen, wenn die Qualität und Menge der Daten nicht stimmen. Für eine automatisierte oder durch KI unterstützte Messdatenanalyse ist es wichtig sowohl für das Training als auch für die Anwendungsdaten eine einheitliche und qualitative Datenbasis zu erschaffen. Es gibt für die Messdatenerhebung die Möglichkeit direkt mit interpretierten Messgrößen auf dem Bussystem zu messen oder indirekt über Tracing. Beim Tracing werden alle Daten auf dem Bus mitgemessen, können aber ohne eine Interpretationsdatei nicht als Softwaresignal bzw. Messgröße gelesen werden. Ebenso ist es wichtig, ein Konzept für den Umgang mit verschiedenen Datenquellen, im konkreten Fall Fahrzeuge oder Testaufbauten mit unterschiedlicher Ausstattung an Messtechnik für entsprechende Anwendungsfälle zu clustern.

Ausblick

Die in dieser Arbeit betrachteten KI-Ansätze basieren auf dem aktuellen Stand der Technik und könnten von zukünftigen Entwicklungen überholt werden. Mit den gewonnenen Erkenntnissen könnte in der Zukunft eine

vollautomatisierte Testumgebung aufgebaut werden, mit der große Volumen von Testzyklen am Fahrzeug abgeprüft und auf Fehler untersucht werden können. Hierdurch könnten mehr seltene oder nicht-triviale Auffälligkeiten gefunden werden, ohne dass aufwendige manuelle Analysen vom Entwickler selbst ausgeführt werden müssen. Ein solcher Prozess trägt zu einer noch höheren Stabilität der Systeme bei.

Literatur und Abbildungen

[1] Sabeen Ahmed, Ian E. Nielsen, Aakash Tripathi, Shamoon Siddiqui, Ghulam Rasool, and Ravi P. Ramachandran. Transformers in Time-series Analysis: A Tutorial. *Circuits, Systems, and Signal Processing*, 42:7433–7466, 2023.

[2] Hadi Askaripoor, Morteza Hashemi Farzaneh, and Alois Knoll. E/E Architecture Synthesis: Challenges and Technologies. *Electronics*, 11, 2022.

[3] Eigene Darstellung.

[4] Stephen Ginthinji and Ciira Wa Maina. *Anomaly Detection on Time Series Sensor Data Using Deep LSTM-Autoencoder*. Engineers, Institute of Electrical and Electronics, 2023.

[5] Markus Hanselmann, Thilo Strauss, Katharina Dormann, and Holger Ulmer. CANet: An Unsupervised Intrusion Detection System for High Dimensional CAN Bus Data. https://arxiv.org/pdf/1906.02492.pdf, 2019.

[6] Tae-Young Kim and Sung-Bae Cho. Web traffic anomaly detection using C-LSTM neural networks. *Expert Systems with Applications*, 106:66–76, 2018.

[7] Pankaj Malhotra, Anusha Ramakrishnan, Gaurangi Anand, Lovekesh Vig, Puneet Agarwal, and Gautam Shroff. LSTM-based Encoder-Decoder for Multi-sensor Anomaly Detection. https://arxiv.org/pdf/1607.00148.pdf, 2016.

[8] H. D. Nguyen, K. P. Tran, S. Thomassey, and M. Hamad. Forecasting and Anomaly Detection approaches using LSTM and LSTM Autoencoder techniques with the applications in supply chain management. *International Journal of Information Management*, 57, 2021.

[9] Werner Zimmermann and Ralf Schmidgall. *Bussysteme in der Fahrzeugtechnik*. Springer Vieweg, 5 edition, 2014.

Herausforderungen und Eigenschaften von cloudbasierter voll-homomorpher Verschlüsselung

Modjtaba Gharibyar **Dominik Schoop**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma Daimler Truck, Leinfelden-Echterdingen

Einleitung

1978 führten Rivest, Shamir und Adleman den *RSA-Algorithmus* ein, der ein Meilenstein für den Privacy-Homomorphismus [3] und moderne Public-Key-Verfahren darstellte. Dies ermöglichte das Rechnen auf verschlüsselten Daten, den sogenannten Ciphertexten, basierend auf dem Faktorisierungsproblem großer Primzahlen. Craig Gentrys vollhomomorphes Kryptosystem (FHE) von 2009 erlaubt Berechnungen in der Cloud, ohne Klartextenthüllung, und stärkt somit den Datenschutz in Cloud-Umgebungen. Die aktuelle Forschungslage deutet auf eine zunehmende Praxistauglichkeit von FHE hin, was für Unternehmen wichtig ist, um zukünftige Sicherheitsstandards zu erfüllen. Insbesondere für einen Truck-Hersteller, der in der Cloud aggregierte Daten wie die Emissionswerte seiner Kunden berechnet, ist FHE ein geeignetes Werkzeug, um Datenschutzbedenken zu reduzieren. Im Rahmen der hier beschriebenen Abschlussarbeit evaluieren wir das CKKS-Verschlüsselungsschema, eine FHE-Variante, implementiert in einer Cloud-Umgebung für einen konkreten Use-Case im Bereich Flottenmanagement. Zudem stellen wir die Anforderungen verschiedener Use-Cases dar. Die Evaluierung betrachtet die Ressourcenbelastung und Effizienz verschiedener Parametrisierungen bei Verschlüsselungs- und Entschlüsselungsprozessen, Ciphertext-Berechnungen und dem damit verbundenen Bootstrapping. Da das CKKS-Schema Festkommazahlen nutzt, erreichen wir präzise Berechnungen bis auf die Nachkommastelle genau, was in Bereichen wie der Durchschnittsermittlung, CO2-Emissionsanalyse und Effizienzbewertung im Flottenmanagement zu geringeren Ungenauigkeiten in Ergebnissen führt, als bei anderen Schemata. Zu diesem Zweck untersuchen wir Berechnungen mit unterschiedlichen mathematischen Komplexitätsgraden.

Grundlagen der FHE-Schemata

Die Optimierungen der bisherigen FHE-Schemata gliedern sich innerhalb der Literatur in Generationen, die sich in ihrer Performance, Konstruktion und im Umgang mit verschiedenen Datentypen unterscheiden.

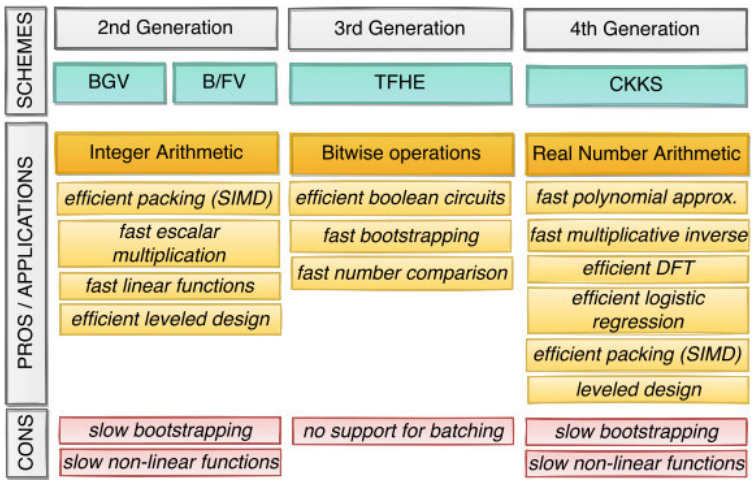


Abb. 1: Darstellung der jeweiligen FHE-Schemata mit ihren Vor- und Nachteilen [2]

Abb. 1 zeigt FHE-Schemata mit ihren entsprechenden Vor- und Nachteilen. FHE-Techniken bezüglich des Ciphertexts lassen sich in zwei Kategorien einteilen: *Bootstrapping* und der *FHE-Leveled-Ansatz*. Der FHE-Leveled-Ansatz ist effizienter für Berechnungen mit einer bestimmten Tiefe, jedoch weniger flexibel bei tieferen Berechnungen. Ein Kompromiss bei der Performance ergibt sich aus abgeschwächten Sicherheitsannahmen. Das CKKS-Schema ist im Vergleich zu anderen Schemata besonders effizient. Es ermöglicht homomorphe Operationen über Approximationen von reellen Zahlen und somit die Handhabung von Festkommazahlen. Durch Anpassungen wurde CKKS zu einem *bootstrappingfähigen* FHE-Schema transformiert.

Die Evaluierung: Am Beispiel des Flottenmanagements

Zunächst betrachten wir als Use-Case die Situation in einem Speditionsunternehmen (Abb. 2). Die Fahrzeuge der Truck-Flotte, ausgestattet mit einer speziellen Applikation, erfassen ihre Standortdaten, Kilometerlaufleistung und den Kraftstoffverbrauch und übermitteln diese Daten HTTPS-verschlüsselt an einen Datenspeicher in der Cloud. Das Backend, betrieben vom Fleet-Agent-Provider, führt auf Basis dieser Daten Emissionsberechnungen durch, die für das Umweltschutzamt bestimmt sind. Die hierfür benötigten Berechnungen basieren dabei auf Kilometerlaufleistung und Kraftstoffverbrauch. Obwohl die Standortdaten für diese spezifischen Berechnungen nicht benötigt werden, sind sie für den Fleet-Agent-Provider einsehbar. Dies stellt aus Sicht des Truck-Fleet-Owners ein Datenschutzrisiko dar, da die Standortdaten ausschließlich für das eigene Unternehmen zugänglich sein sollen. Autorisierte Personen wie der Truck-Fleet-Owner können alle Daten jederzeit über einen verschlüsselten Kanal anfordern.

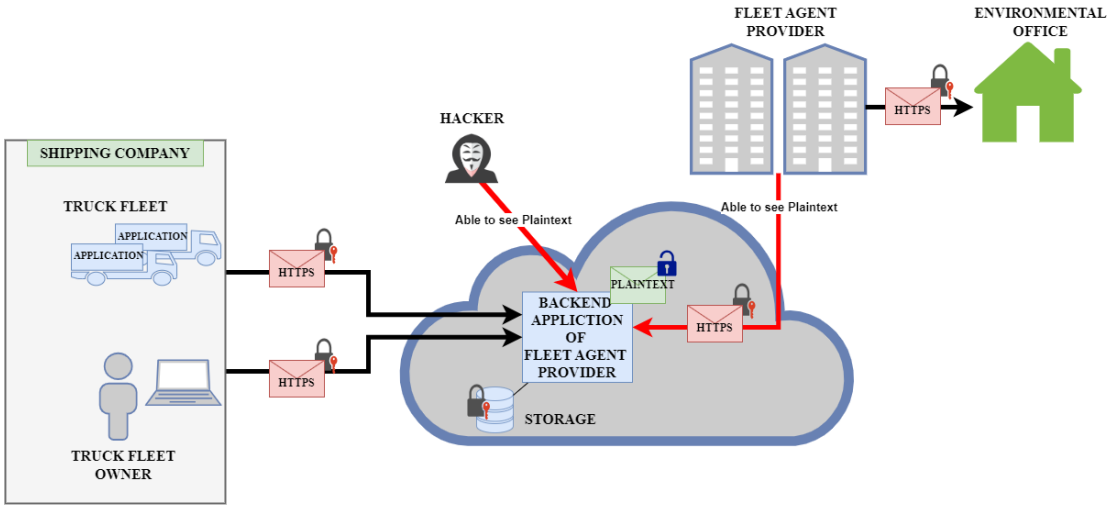


Abb. 2: Die Berechnungen werden innerhalb des Cloud-Backends im Klartext durchgeführt, wodurch die Gefahr besteht, dass unautorisierte Teilnehmer die gesamten Daten im Klartext auslesen können. [1]

Daten im Backend werden im Klartext verarbeitet. Dies ermöglicht es einem Angreifer mit Backend-Zugriff, sämtliche Berechnungen und Standortdaten wegen fehlender Ende-zu-Ende-Verschlüsselung im Klartext auszulesen. Zum Schutz der Daten soll die künftige Systemarchitektur (Abb. 3) die Kommunikation im Speditionsunternehmen Ende-zu-Ende verschlüsseln. Dadurch hätten Hacker und Fleet-Agent-Provider keinen Zugriff mehr auf Daten im Klartext. Dies führt dazu, dass die zuvor in der Cloud-Umgebung bereitgestellte Backend-Applikation (siehe Abb. 2) nun lokal beim Management ausgeführt wird, wie in Abb. 3 dargestellt. Durch die Verwendung einer FHE-Engine, wie in Abb. 3, kann der Fleet Agent Provider nur noch autorisierte Berechnungen auf Basis der verschlüsselten Kilometerlaufleistung und Kraftstoffverbrauchsdaten durchführen. Da der Datensatz nicht mehr vorab entschlüsselt wird, sind sowohl die Standortdaten als auch die einzelnen Zahlenwerte für den Fleet Agent Provider nicht einsehbar. Dies würde aus Sicht des Speditionsunternehmens den Datenschutz erheblich verbessern. Die Architektur in Abb. 3 bildet die Grundlage für die Entwicklung verschiedener Architekturen, die entsprechend ihrer Anforderungen zugeschnitten werden können.

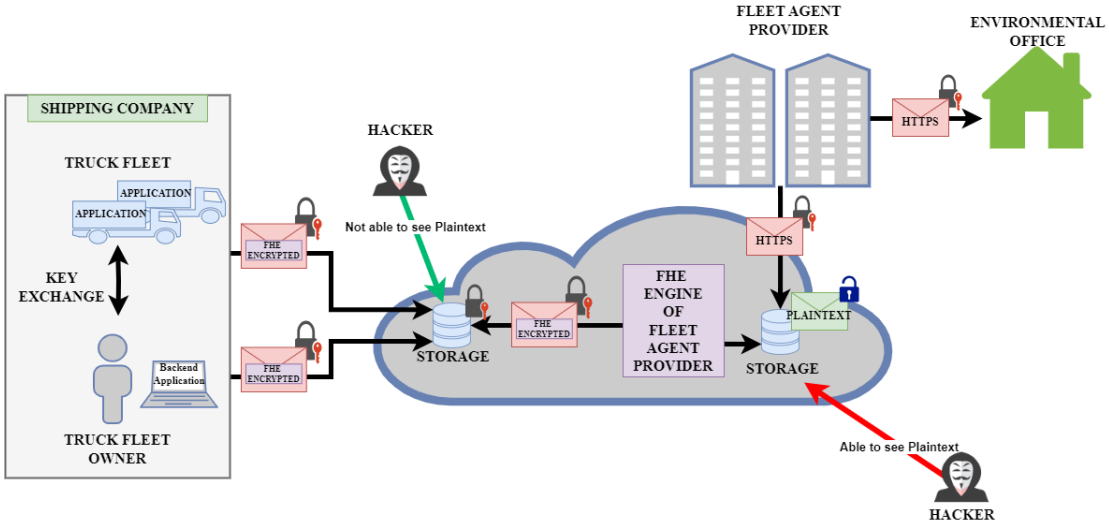


Abb. 3: Durch die Einführung von FHE können Daten Ende-zu-Ende verschlüsselt werden, da Berechnungen auf Ciphertexten stattfinden. [1]

Ausblick

Nach der Beschreibung der Anforderungen verschiedener Use Cases wird in der Abschlussarbeit die oben beschriebene Architektur für den Use Case der homomorph verschlüsselten Berechnung von Verbrauchswerten prototypisch in einer kommerziellen Cloud implementiert. Das nachfolgende Ziel ist die Evaluation der Performance unter verschiedenen Parametrisierungen des CKKS-Schemas. Somit soll eine effiziente Nutzung des CKKS-Schemas für ähnliche Use Cases einem größeren Nutzerkreis ermöglicht werden.

Literatur und Abbildungen

[1] Eigene Darstellung.

[2] Chiara Marcolla, Victor Sucasas, Marc Manzano, Riccardo Bassoli, Frank H.P. Fitzek, and Najwa Aaraj. *Survey on Fully Homomorphic Encryption, Theory and Applications*. Institute of Electrical and Electronics Engineers (IEEE), 2022.

[3] Rivest Ronald, Adleman Len, and Dertouzos Michael. ON DATA BANKS AND PRIVACY HOMOMORPHISMS. *Academic Press*, 1978.

Konzeptionierung und Entwicklung von zwei Providern über die Service-oriented Device Connectivity Schnittstelle

Dustin Gohl **Michael Scharf**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma TZM GmbH, Göppingen

Einleitung

Die Digitalisierung wird auch im Gesundheitswesen immer weiter vorangetrieben. Vor allem der Austausch und die Dokumentation von Daten medizinischer Geräte stehen dabei im Fokus. Hierfür muss die Erfassung und Aufzeichnung von Daten der Geräte vorausgesetzt sein. Besonders die Aufzeichnung und die anschließende Speicherung in ein System stellen eine Problematik für Kliniken dar. Auf den Geräten liegen die Daten zwar in digitaler Form vor, können aber nur manuell in ein System übertragen werden [7] [2], da die Geräte proprietär und nur selten kompatibel sind.
Aufgrund dieser Eigenschaft sind klinische Umgebungen in der Regel geschlossene Systeme, in denen ausschließlich Kommunikation zwischen Geräten mit dem gleichen Hersteller möglich ist. Hierdurch werden Kliniken stark in der Auswahl der Geräte eingeschränkt und müssen sich auf einen Hersteller festlegen. An dieser Problematik setzt der Standard SDC an. Ziel ist, die Interoperabilitätlsücke zu schließen, sowie das Wegfallen der Abhängigkeit auf einzelne Hersteller, sodass Kliniken frei entscheiden können, welche Geräte verwendet und vernetzt werden [9].

Ziel der Arbeit

Das Ziel der Arbeit ist die Konzeptionierung und Entwicklung eines Treibers für UMG, der mit dem Kommunikationsprotokoll SDC arbeitet. Hierfür wird ein neu entwickelter SDC-Stack verwendet, der zunächst in einen bereits bestehenden SDC-Treiber, mit altem Stack, integriert werden soll. Anschließend soll die aktualisierte SDC-Schnittstelle mit integriertem Stack auf UMG in Betrieb genommen werden. Bei dem zu entwickelnden Treiber handelt es sich um ein Senken-Treibermodul, das vom UMG empfangene Daten mit SDC an medizinische Systeme weiterleitet. Nach der Inbetriebnahme des Treibers auf UMG, soll dieser um eine Funktionalität erweitert werden, die es ermöglicht, Daten mehrerer Geräte (z.B Vital- oder Alarmdaten) über die Schnittstelle zur Verfügung zu stellen. Abschließend werden der funktionsfähige SDC-Treiber, sowie die hinzugefügte Erweiterung mit zwei medizinischen Geräten geprüft und getestet.

Service-oriented Device Connectivity - SDC

SDC ist ein Kommunikationsprotokoll zur Vernetzung medizinischer Geräte. Der Standard ist ein Teil der ISO/IEEE 11073 Normenfamilie und wurde mit dem Ziel entwickelt, die Interoperabilitätslücke in der Gerätekommunikation zu schließen und dabei Multi-Point-Verbindungen zwischen Geräten und Systemen zu ermöglichen [2] [3]. Über Multi-Point-Verbindung lassen sich Daten auf unterschiedlichen Bildschirmen gleichzeitig ausgeben. Ein Mechanismus sorgt für einen konsistenten Austausch medizinischer Daten ohne Verlust der Datenqualität. Der Datenaustausch ist bidirektional, dadurch kann ein Gerät sowohl Sender als auch Empfänger von Daten sein [7].
Medizinische Systeme können herstellerübergreifend über standardisierte Schnittstellen an ein Netzwerk angebunden werden und Geräte werden automatisch erkannt und integriert. Für die Erkennung wird die WS-Discovery Funktionalität verwendet, die Webservices auffindet. Dabei wird eine Anfrage an eine Servicegruppe geschickt und Antworten passender Services entgegengenommen [2] [10].
Damit der Standard in der kritischen Patientenversorgung eingesetzt werden darf, müssen bestimmte Sicherheitsanforderungen erfüllt sein. Die Sicherheit beim Datenaustausch wird über eine im Standard integrierte End-to-End Verschlüsselung gewährleistet. Neben der Verschlüsselung muss eine Ausfallsicherheit gegeben sein. Infolgedessen baut SDC zwei Verbindungen auf, sodass Geräte im Netzwerk weiterhin kommunizieren können, sollte eine der Verbindungen ausfallen [2].
In den Kliniken soll der Standard vor allem manuelle, klinische Arbeitsschritte und Prozesse übernehmen

und automatisieren. Ebenso sollen Informationen und Kontrolloptionen der Geräte zur Verfügung stehen, wo diese gebraucht werden [9]. Außerdem wird sicheres Fernsteuern von Medizingeräten ermöglicht. SDC tritt nicht mit bereits bestehenden Kommunikationsstandards in Konkurrenz, sondern geht auf diese ein und ergänzt diese [2] [3].

ISO/IEEE 11073 Normenfamilie

Die SDC-Standardfamilie wurde durch das Forschungsprojekt OR.NET entwickelt (siehe Abbildung 1) entwickelt. Bestehend aus den drei Hauptteilen Core Standards, Participant Key Purposes und Device Specifications werden die Bestandteile eines Systems definiert, dass einen Austausch und eine Auswertung von Vitaldaten zwischen medizinischen Geräten ermöglicht. Die Kernstandards sind ISO/IEEE 11073-20702, -10207 und -20701 [3] [8] [11].

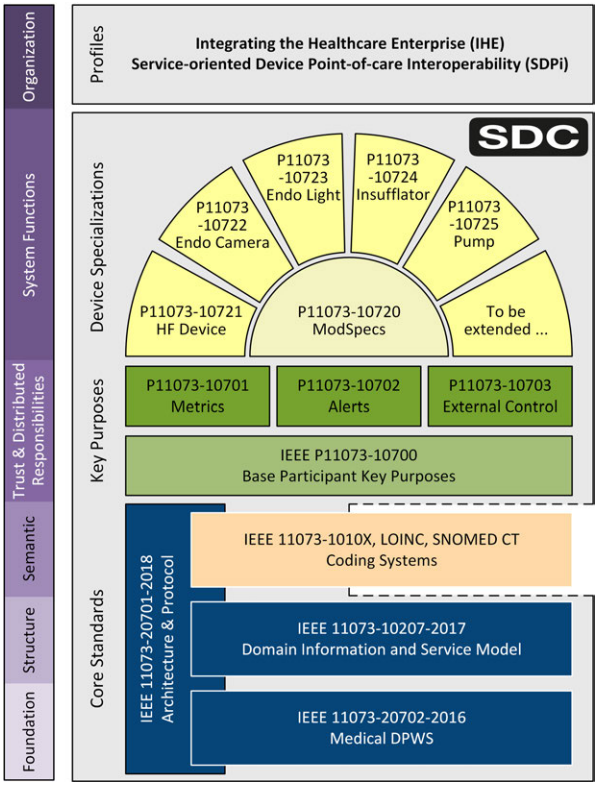


Abb. 1: SDC-Standard-Familie [8]

Der Standard ISO/IEEE 11073-20702 Medical Devices Communication Profile for Web Services (MDPWS) gewährleistet eine grundlegende Interoperabilität zwischen Medizingeräten. Er ermöglicht einen sicheren Datenaustausch in verteilten Systemen medizinischer Geräte und Systeme, sowie das dynamische Finden von Kommunikationspartnern [3] [11] [4].

Der Standard ISO/IEEE 11073-10207 Domain Information and Service Model for Service-oriented Point-of-Care Medical Device Communication definiert ein Domäneninformation- und Servicemodell und stellt die strukturelle Interoperabilität sicher. Die Modellierung wird mit XML Schema realisiert und kann frei erweitert werden. Die Modellierung ermöglicht eine Mehrpunkt-zu-Mehrpunktkommunikation und teilt die Beschreibung eines Gerätes in zwei Teile auf. Der erste Teil umfasst die Fähigkeiten eines Gerätes, also die Modellierung von relevanten Daten eines Geräts, wie Messungen, Konfiguration oder Alarme. Ebenso wird die Fernsteuerung von bestimmten Funktionalitäten in diesem Teil definiert. Im zweiten Teil wird der aktuelle Gerätezustand beschrieben [3] [11] [5].
Der Standard IEEE 11073-20701 Service-Oriented Medical Device Exchange Architecture and Protocol Binding definiert eine Architektur für serviceorientierte, verteilte medizinische Geräte und medizinische IT-Systeme. Ebenso wird die Verbindung zwischen den Standards IEEE 11073-20702 und IEEE 11073-10207 beschrieben, sowie eine Anbindung zu weiteren Standards, wie dem Network Time Protocol (NTP) [3] [11] [6].

UMG

UMG ist ein eigenständiges Medizinprodukt, für die herstellerübergreifende Vernetzung und Kommunikation von medizinischen Systemen und Geräten, um der fehlenden Interoperabilität in Kliniken entgegenzuwirken. Das Gerät empfängt die Daten, konvertiert diese und leitete sie an ein bestimmtes Zielsystem weiter. Bei der Ausführung sind keine Anpassungen an den Kommunikationspartnern nötig und UMG übernimmt Protokollanpassungen oder Neuimplementierungen. Grundsätzlich baut sich die Software aus den vier Komponenten Connectivity Controller, QsyslogWrapper, Senken-Treibermodul und Quellen-Treibermodul, auf (siehe Abbildung 2).

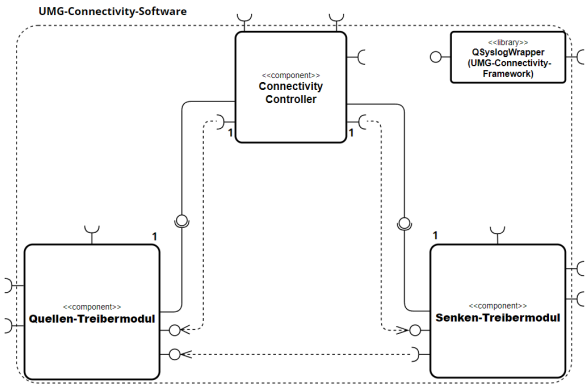


Abb. 2: Aufbau UMG-Connectivity-Software [1]

Der Connectivity Controller ist der Initiator und startet die vorab definierten Quellen- und Senken-Treibermodule der Kommunikationskette. Während des Betriebs überwacht der Controller das gesamte System, sowie die laufenden Treibermodulprozesse und greift bei bestimmten Problemen (z.B. Kommunikationsschwierigkeiten) ein. Der QSyslog Wrapper nimmt Nachrichten der Komponenten entgegen und schreibt diese in eine Log-Datei. Das Quellen-Treibermodul kommuniziert mit einem externen, quellenseitigen, medizinischen Fremdsystem über eine externe Schnittstelle und empfängt dabei gelieferte Daten und leitet diese an das entsprechende Senken-Treibermodul weiter. Dieses nimmt die Informationen entgegen und leitet diese über eine externe Schnittstelle an ein externes, senkenseitiges, medizinisches Fremdsystem weiter.

Umsetzung

Zu Beginn wurden der bestehende Treiber, sowie der alte SDC-Stack analysiert. Zum einen wurden der Aufbau und die Funktionsweise untersucht, zum anderen, wie der alte Stack integriert und ausgeführt wird. Auf Basis dieser Erkenntnisse wurde ein erstes Konzept erstellt, auf welche Weise der neue Stack integriert werden könnte. Ebenfalls wurde sich mit dem neuen Stack näher befasst und anhand einer Beispielapplikation, die auf diesem aufbaut, getestet. Dabei konnten weitere Informationen gesammelt werden, wie der neue Stack funktioniert und wie dieser möglicherweise in den Treiber integriert werden könnte. Der alte Treiber diente als Grundlage für die Entwicklung.
Vor der Entwicklung musste festlegt werden, für welche Version eines Betriebssystems des UMG der Treiber entwickelt wird. Zur Auswahl standen eine alte Version und eine sich noch in der Entwicklung befindende neue Version. Die Entscheidung fiel auf die neue Version, da diese besser für die Integration des neuen Stacks geeignet ist. Der bestehende Treiber wurde für die alte Version entwickelt und musste folglich auf das neue System angepasst werden (z.B. Konfigurationen).
Die Treiberentwicklung wurde in der Programmiersprache C++ mit dem Framework Qt5 durchgeführt. Zur Integration von Bibliotheken, vor allem dem SDC-Stack, mussten diese vorab cross-kompiliert werden. Eine Toolchain baut dabei eine Entwicklungsumgebung auf und nimmt Konfigurationen vor, sodass die Bibliothek für ein bestimmtes Zielsystem, in dieser Arbeit UMG, gebaut werden kann. Neben dem Stack, mussten veraltete Bibliotheken aufgrund fehlender Kompatibilität mit dem Stack aktualisiert werden. Hierbei wurden die Bibliotheken OpenSSL und Xerxes cross-kompiliert und neu in den Treiber integriert.
Verbindungen, Kommunikation (interne und externe) und Steuerung des UMG sind von einem für UMG entwickelnden Framework hinzugefügt worden. Der SDC-Stack wird von einem Provider Handler integriert und gesteuert. Beispielsweise die Initialisierung und das Starten des Protokolls. Der SDC-Provider-Handler erhält und verarbeitet beim Aufruf eine Config, die alle benötigten Informationen (z.B. IP des Geräts oder Speicheort von MDIBs und Zertifikaten) enthält. Aufgerufen wird der Handler von einem Protocol Handler, der die Verarbeitung des Protokolls eines Treibers übernimmt. Eine weitere Klasse plant das Anfragen und Senden von Daten.
Nach der vollständigen Integration des Stacks in den Treiber, wurde der Treiber für das Zielsystem gebaut und auf UMG in Betrieb genommen. Weitere Dateien, die der Treiber zur Ausführung braucht, mussten vorab auf dem UMG abgelegt werden. Getestet wurde der Treiber zusammen mit einem Connectivity Controller und einem Quellen-Treiber.

Ausblick

Im weiteren Verlauf der Bachelorarbeit soll die genannte Erweiterung entwickelt werden. Anhand des alten SDC-Treibers und vergleichbarer Senken-Treibermodule, die Daten zur Verfügung stellen, kann ein Konzept erstellt werden, wie die Erweiterung aufgebaut und entwickelt werden könnte. Nach dem Abschluss des Projekts wird die Entwicklung am Treiber nicht aufhören, da der SDC-Stack sich weiterhin in der Entwicklung befindet und der Treiber dementsprechend angepasst und erweitert werden muss.

Literatur und Abbildungen

[1] Eigene Darstellung.

[2] Martin Kasparick, Björn Andersen, Hannes Ulrich, Stefan Franke, Erik Schreibe, Max Rockstroh, Frank Golatowski, Dirk Timmermann, Josef Ingenerf, and Thomas Neumuth. IEEE 11073 SDC and HL7 FHIR - Emerging Standards for Interoperability of Medical Systems. https://www.imd.uni-rostock.de/storages/uni-rostock/Alle_IEF/IMD/veroeff/2018/2018_Kasparick_IEEE_11073_SDC_and_HL7_FHIR_-_Emerging_Standards_for_Interoperability_of_Medical_Systems.pdf, 2018.

[3] Martin Kasparick, Björn Andersen, Hannes Ulrich, Stefan Franke, Erik Schreibe, Max Rockstroh, Frank Golatowski, Dirk Timmermann, Josef Ingenerf, and Thomas Neumuth. IEEE 11073 SDC for Pandemics like COVID-19: Example Implementation of an Isolation Room. *Current Directions in Biomedical Engineering*, 2022.

[4] Institute of Electrical and . Electronics Engineers Standards Association. IEEE Standard for Health informatics–Point-of-care medical device communication Part 20702: Medical Devices Communication Profile for Web Services. https://standards.ieee.org/ieee/11073-20702/6034/, 2017.

[5] Institute of Electrical and . Electronics Engineers Standards Association. IEEE Health informatics–Point-of-care medical device communication Part 10207: Domain Information and Service Model for Service-Oriented Point-of-Care Medical Device Communication. https://standards.ieee.org/ieee/11073-10207/6032/, 2018.

[6] Institute of Electrical and . Electronics Engineers Standards Association. IEEE Standard - Health informatics–Point-of-care medical device communication - Part 20701: Service-Oriented Medical Device Exchange Architecture and Protocol Binding. https://standards.ieee.org/ieee/11073-20701/6059/, 2019.

[7] . ORNET. SDC - Service oriented device connectivity. https://ornet.org/en/services-2-2/, 2023.

[8] . ORNET. SDC Standards Family. https://ornet.org/en/services-2-2-3/, 2023.

[9] . ORNET. What does SDC offer the user? https://ornet.org/en/services-2-2-5/, 2023.

[10] . Wikipedia. WS-Discovery. https://de.wikipedia.org/wiki/WS-Discovery, 2018.

[11] . Wikipedia. IEEE 11073 Service-oriented Device Connectivity (SDC). https://de.wikipedia.org/wiki/IEEE_11073_Service-oriented_Device_Connectivity_(SDC)#ISO/IEEE_11073, 2023.

Dynamic Intent Queries for Transformer-based Trajectory Prediction in Autonomous Driving

Lennart Hartung **Markus Enzweiler**

Department of Computer Science and Engineering, Esslingen University

Work carried out at Robert Bosch GmbH, Stuttgart Vaihingen

Introduction

The concept of self-driving technology is widely recognized as a significant milestone in enhancing safety, improving efficiency, and elevating comfort in traffic management. A pivotal challenge in this domain is the accurate prediction of movements of other traffic participants, which is essential for the effective response of autonomous vehicles to their dynamic environment. Current approaches predominantly utilize transformer-based deep-learning architectures. These methods typically vary from direct regression [4], which directly predicts trajectories from an agent's encoded features, to goal-oriented methods [2] that focus on a dense selection of goal candidates to anticipate all potential agent destinations. The Motion Transformer (MTR) [6], acclaimed for winning the Waymo Open Motion Challenge 2022 [3], innovatively merges these techniques. It employs a small set of learnable motion query pairs consisting of goal candidates which serve as starting endpoints for the direct regression method. It thus reduces the reliance on an extensive array of goal candidates and eases optimization challenges present in purely regression-based methods.

However, a notable limitation of the MTR is its dependence on pre-generated, static intention points as goal candidates. These points, lacking in scene-specific context, often fail to account for the unique constraints of different road networks. As a result, many of these points represent impractical or unrealistic trajectory endpoints, as seen in Fig. 1. To address this issue, this thesis introduces a novel approach: the integration of dynamic, scenario-specific intent queries into the MTR. This enhancement aims to improve the reachability and realism of goal points, thereby enhancing the overall trajectory prediction accuracy of the model.

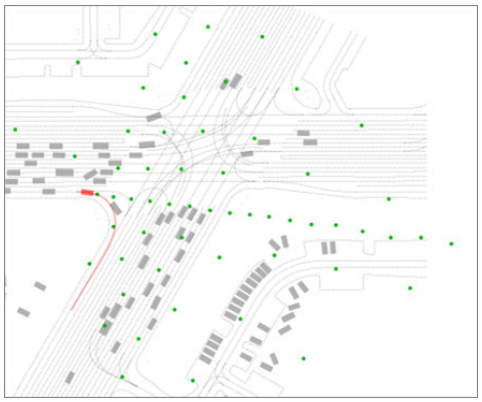


Fig. 1: Visualization of the static intention points (green dots) in a traffic scenario. The agent is marked with a red box and its path is shown by a red line indicating movement over the past 1 second and future 8 seconds. [5]

Background

MTR adopts a novel transformer encoder-decoder structure with motion queries to predict multimodal trajectories. It incorporates two distinct types of queries within each motion query pair: the static intention query and the dynamic searching query. The static intention query is designed to reduce uncertainty in future trajectories by leveraging potential motion intentions of agents. In contrast, the dynamic searching query dynamically explores trajectory features across various motion modes, refining trajectories through the iterative accumulation of detailed trajectory features. A visual description of this process can be seen in Fig. 2.

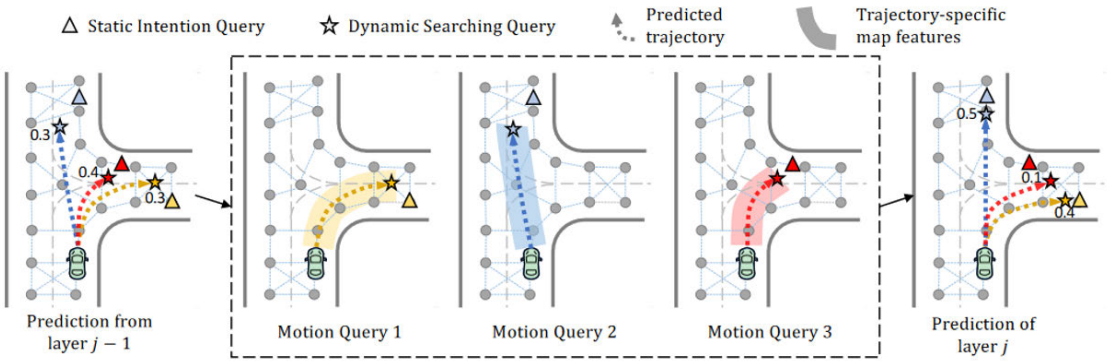


Fig. 2: Illustration of the iterative motion refinement done in the context of static intention query and dynamic searching query. [6]

The static intention query relies on intention points, which are pre-generated using the Waymo Open Motion Dataset [1]. These points are derived through a K-Means Clustering algorithm, which categorizes similar endpoints into clusters representing typical destinations or paths. This clustering process is conducted separately for different entities, including vehicles, pedestrians, and cyclists. For vehicles, the processed outcome is depicted in Fig. 3 and demonstrates 64 distinct static intention points.

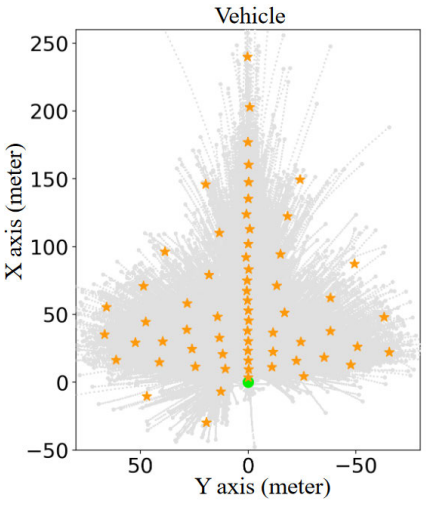


Fig. 3: Vehicle static intention points distribution: Agent's current position (green dot), intention points (orange stars), and for clearer visualization only 10% of the historical ground-truth trajectories (gray dotted lines). [6]

Generation of Dynamic Intents

To enhance pre-generated static intention points and address impractical or unrealistic trajectory endpoints (see Fig. 1), we generate dynamic, scene-specific intention points. These are derived using road map data, focusing on vehicles due to their predictable, lane-bound movement patterns, as commonly found in the Waymo Open Motion Dataset. Integrating these dynamic intention points into the MTR architecture necessitates changes in the data flow. The original system integrates static intention points based on object type into the transformer decoder's motion query. Dynamic intention points, however, require a link to their corresponding scenes. This is achieved by attaching intention point data to scene data during ingestion, which is then processed for use in the transformer decoder.

The process starts with accurately localizing the vehicle on a high-definition map. A road graph is then created, showing potential reachable points. To make this data usable in the transformer decoder network, we narrow down the extensive set of potential goal points. Vehicle localization includes calculating the nearest lane node, ensuring the heading alignment is correct, and setting a maximum distance from the closest node for off-street scenarios. In complex scenarios, such as when a lane splits, a look-back logic is applied to avoid incorrect lane assignments and to generate multiple potential location points. The agent's location determines the generation of a road graph of potential endpoints using the Dijkstra Algorithm. This algorithm finds the shortest paths between lane nodes, factoring in distance and speed limits as costs. Considering the large number of points generated, directly inputting them into the transformer model is computationally challenging. To manage this, we reduce the number of intention points to 64, following MTR's approach. We use K-Means Clustering for efficient reduction while maintaining the road graph's key features. This method divides points into clusters represented by centroids, ensuring optimal coverage of the road graph and minimal information loss.

Training

To prepare for scenarios involving unforeseeable maneuvers, such as those due to inaccuracies in the high-definition map or instances of illegal behavior, we developed two types of models. The first is based on purely dynamic intents, while the second combines both static and dynamic intents. This dual approach enhances the model's adaptability to real-world variability.
The training of these models, including the static model and its adaptations, was undertaken on a high-performance computing cluster, generously provided by Robert Bosch GmbH. This extensive training used the full Waymo Open Motion dataset. The process required a significant investment of computational resources, taking 149 hours to complete on four Nvidia V100 GPUs, each having 16GB of memory.

Results

The evaluation of dynamic intent queries was conducted using the Waymo Open Motion dataset. For vehicle predictions, a clear trend emerged showing improved performance over longer prediction horizons. This improvement was particularly evident in the Minimum Final Displacement Error and the Miss Rate metric. A qualitative analysis, as shown in Fig. 4, suggests that these enhancements were more pronounced longitudinally than laterally.
Furthermore, the integration of static and dynamic intentions was effective, particularly for short-term predictions and in scenarios involving unpredictable, non-map-conforming behaviors of agents. An intriguing discovery was the positive impact of training with vehicle-centric dynamic intention points on the accuracy of pedestrian and cyclist predictions. Although the overall performance metrics for cyclists and pedestrians were similar in other respects, there was a marked improvement in the Mean Average Precision for pedestrian predictions. Conversely, cyclist predictions saw a slight decrease in Mean Average Precision but improvements in both Minimum Average Displacement Error and Minimum Final Displacement Error.
Overall, these results underscore the value of incorporating dynamic intention points in enhancing trajectory prediction for autonomous driving. They also highlight the intricate interdependencies within modern trajectory prediction models.

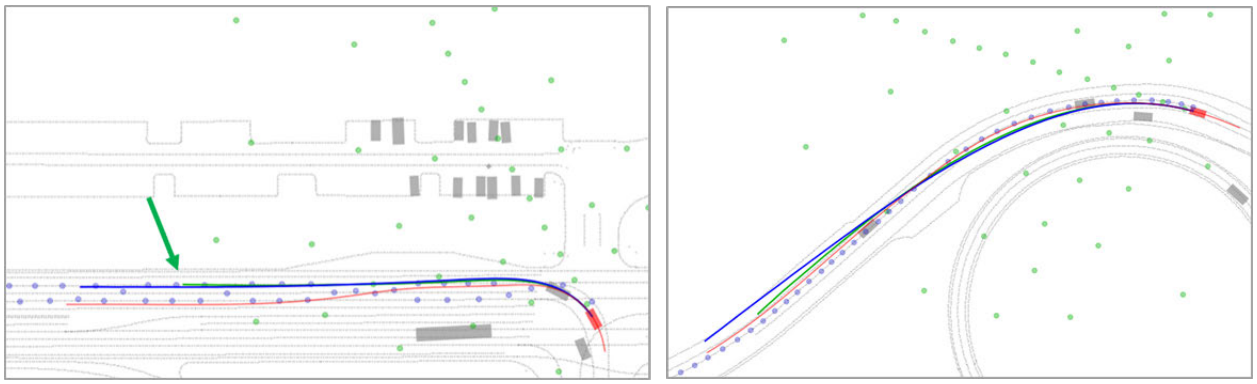


Fig. 4: Trajectory predictions of dynamic and static model: The red path denotes the ground-truth trajectory. Predictions from the static and dynamic models are illustrated in green and blue respectively. [5]

References and figures

[1] Scott Ettinger, Shuyang Cheng, Benjamin Caine, Chenxi Liu, et al. Large Scale Interactive Motion Forecasting for Autonomous Driving : The Waymo Open Motion Dataset. *ICCV*, 2021.

[2] Junru Gu, Chen Sun, and Hang Zhao. DenseTNT: End-to-end Trajectory Prediction from Dense Goal Sets. *ICCV*, 2021.

[3] Waymo LLC. 2022 Waymo Open Motion Prediction Challenge. https://waymo.com/open/challenges/2022/motion-prediction/, 2022.

[4] Jiquan Ngiam, Benjamin Caine, Vijay Vasudevan, Zhengdong Zhang, et al. Scene Transformer: A unified architecture for predicting multiple agent trajectories. *ICLR*, 2022.

[5] Own representation.

[6] Shaoshuai Shi, Li Jiang, Dengxin Dai, and Bernt Schiele. Motion Transformer with Global Intention Localization and Local Movement Refinement. *NeurIPS*, 2022.

Analysis of Attack Methods with Artificial Intelligence

Dieter Holstein **Tobias Heer**

Department of Computer Science and Engineering, Esslingen University

Work carried out at Department of Computer Science and Engineering, Esslingen

Motivation and Problem Definition

Over the past few years, the field of Artificial Intelligence (AI) has experienced significant growth. This growth offers positive changes as well as new challenges in the area of IT security. In our study we examine the current state of AI attack methods. We identify and evaluate, known AI-supported attack techniques and perform a comprehensive reassessment of the current security landscape. Based on this analysis, we explore which specific attack methods are currently possible and implementable through the use of AI technologies and what countermeasures can be taken. Given the complexity and broad range of applications for AI technologies, the following challenges and risks can arise, which we may encounter in the exploration of these new attack vectors:

- Rapid Development of AI Technologies: AI technologies require continous monitoring, due to dynamic changes in their development. Nonetheless, the analysis provides a robust foundation that contributes to the understanding of present-day AI-driven threats.
- Diversity of Application Fields: The extensive range of AI applications both enhances the scope for exploration and complicates our task of focusing on particularly impactful domains. This complexity arises from the need to sort through numerous possible research areas to concentrate on those of highest relevance.
- Availability of High-Quality Data: The quality of the data, which we use for analyzing past attacks and generating training sets for phishing content, significantly influences the validity of our research. The absence of high-quality data presents a substantial issue, as it eliminates the possibility of drawing empirically supported conclusions. This is difficult to address, as data collection is often time consuming and resource intensive.

Considering the rapid development, diverse application fields, and the role of high-quality data in AI technologies, our study emphasizes the importance of a comprehensive approach to understanding and mitigating potential threats in the IT security sector.

Design

A methodical and well-thought-out approach is required to address the challenges and risks identified in the previous sections. The first step is to determine the state of the related work by analyzing current literature and empirical studies. Special consideration is given to studies that have already investigated attack vectors utilizing AI, as they provide valuable insights into the evolving nature of cybersecurity threats. We analyze these studies to understand the methodologies, findings, and limitations that can be useful for our own research.
The quality of specific types of data, such as attack data and evaluations from existing phising attacks, poses a critical challenge. Addressing this issue involves evaluating available datasets for both their timeliness and quality. Ensuring a sufficient amount of data requires supplementing existing datasets with a combination of publicly available and synthetic data.
Another problem we face is the rapid progression of AI technologies. We address the rapid progression of AI technologies by initiating our study with a thorough analysis of the most current state of research, thereby establishing a solid foundation for our subsequent investigations. This allows for a well-informed assessment of attack possibilities with AI at the time of the study.
In our analysis of potential attack vectors, we focus on the following use-cases:

- AI-based Code Analysis: Examination of the capabilities of AI systems to identify vulnerabilities in the code and to recognize and exploit possible Zero-Day exploits.
- AI-driven Code Generation: Analysis of the effectiveness of automated malicious code generation

by AI tools to identify and specifically target vulnerabilities in software systems.

- Social Engineering: Evaluation of AI capabilities in generating Deepfakes and their potential use for manipulating individuals into revealing sensitive information or performing unauthorized actions.
- Deception of Facial or Voice Recognition Systems: Evaluation of strategies where AI-generated images or audio recordings are used to bypass biometric security systems.
- Phishing through Natural Language Processing: Investigation of the potential to generate convincing phishing content with AI that might more easily slip through conventional security filters.

To encapsulate, our exploration of these specific use-cases provides a comprehensive overview of the current and potential threats posed by AI in cybersecurity. This in-depth focus ensures we capture a multifaceted understanding of the challenges and the necessary strategies to address them in the evolving area of AI.

Evaluation

The evaluation of this work is multifaceted and tailored to the specific use-cases under investigation. Each use-case has its own set of evaluation criteria, which will be elaborated upon in the following sections. These criteria play a central role in assessing the efficiency and effectiveness of the AI technologies we investigate.
AI-based Code Analysis:
The evaluation focuses on the number of vulnerabilities correctly identified. The evaluation is based on analyzing vulnerable code samples derived from the OWASP Top 10 2021 list. In addition, the results of AI-assisted analysis tools are compared with traditional, non-AI-based tools. The amount of identified vulnerabilities and the rate of false positives are utilized as specific criteria for evaluation.

Fig. 1: OWASP Top 10 vulnerabilities for 2021 [4]

AI-driven Code Generation:
At the core of this evaluation is the effectiveness of the generated codes. The AI code generation tools must be capable of effectively exploiting security vulnerabilities. Criteria therefore include the effectiveness of the codes in exploiting vulnerabilities.
Social Engineering and Deepfakes:
The persuasiveness and realism of the generated content serve as key criteria. These are evaluated through live calls, focusing primarily on whether the Deepfakes can be detected by the participants and to what extent they appear realistic. The evaluation is always carried out with the prior consent of the test subjects and in strict compliance with ethical guidelines.

Fig. 2: Comparison between the original image of Elon Musk (left) and a Deepfake (right), generated using AI tools to alter the individual's appearance. Created by the author. [6]

Natural Language Processing (NLP) and Phishing:
In evaluating the effectiveness of AI-generated phishing content, we consider criteria such as the response rates and quality of interactions within a controlled experiment. These metrics will be measured among a predetermined set of test subjects, consisting of volunteers who have been briefed on the nature of the study.
All the above-mentioned criteria are included in the overall assessment, comprising the efficiency and precision of code analysis, the effectiveness of generated codes, and the realism of Deepfake technologies and Phishing contents. The goal is to achieve either a significant improvement or at least equivalence in all these areas compared to traditional, non-AI-based methods. We pay particular attention to the use of resources, data quality and the success of the results.

Related Work

In the paper "Machine Learning and IT Security," Hartenstein [7] delves into Machine Learning's role in IT security, especially on attack and defense mechanisms. While his focuse is on defense, our study shifts to AI's attack mechanisms. His insights into IT security complexities are important for our exploration of the varied AI attack possibilities.
Farheen et al. [2] explores AI's role in cybersecurity, especially regarding social engineering. Highlighting AI's abilities in this domain, our research further

investigates AI misuse, like Deepfakes, in advanced social engineering attacks.
The study by S. Abhishek et al. [1] delves into machine-learning models' efficacy in detecting Cross-Site Scripting (XSS) attacks. Its relevance to our work lies in its evaluation of these models in cybersecurity, serving as groundwork for our investigation into AI-assisted attack methods.
The paper "From ChatGPT to ThreatGPT: Impact of Generative AI in Cybersecurity and Privacy" [3] examines AI's role, especially ChatGPT, in cybersecurity. Highlighting AI's potential in code generation and bug detection, it showcases ChatGPT's ability to create code for cyber attacks. This paper is significance to our research, due to its examination of AI-driven code generation, a crucial aspect for our study.
Shropshire' [8] paper underscores NLP's potential in phishing by analyzing and mimicking individual writing styles. Indicating that AI-Based tactics might outperform traditional methods, our research explores these NLP advancements and their effects on creating subtle phishing techniques, analyzing the evolving threat landscape.
In the paper "New Tricks to Old Codes", Ozturk et al. [5] explore ChatGPT's potential in detecting vulnerabilities in PHP applications. The study underscores AI chatbots' growing relevance in code analysis. Our study builds on this, especially given ChatGPT's advancements while expanding to different languages and vulnerabilities.
In summary, the current state of research provides a robust understanding of the role of AI in both offensive and defensive cybersecurity measures. While previous works have laid the groundwork in understanding the complexities of AI systems, our work aims to advance this knowledge by focusing on the state of the art in AI-enabled attack methods compared to older AI systems and non-AI-based tools to particularly investigate areas that are relatively unexplored in existing literature.

Results

This study provides a comprehensive analysis of the current state of AI-assisted attack techniques in IT security. We identify and evaluate existing and potential attack methods utilizing artificial intelligence. In addition, our thesis conducts an in-depth examination of specific attack scenarios including code analysis, code generation, social engineering, and phishing.
Our work serves not only as a snapshot but also provides approaches for ongoing adaptation to new AI-based threats and offers advice on how to prevent these threats. Thus, we establish a robust foundation for future research and contribute to improving the understanding of AI-driven attacks.

References and figures

[1] S. Abhishek, Rahulkrishnan Ravindran, T. Anjali, and V. Shriamrut. AI-Driven Deep Structured Learning for Cross-Site Scripting Attacks. In *2023 International Conference on Innovative Data Communication Technologies and Application (ICIDCA)*, pages 701–709. IEEE, 2023.

[2] Meraj Farheen Ansari, Bibhu Dash, Pawankumar Sharma, and Nikhitha Yathiraju. The Impact and Limitations of Artificial Intelligence in Cybersecurity: A Literature Review. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4323317, 09 2022.

[3] Maanak Gupta, Charankumar Akiri, Kshitiz Aryal, Eli Parker, and Lopamudra Praharaj. From ChatGPT to ThreatGPT: Impact of Generative AI in Cybersecurity and Privacy. In *IEEE Access*, pages 80218–80245. IEEE, 2023.

[4] Monika Kukreti. OWASP Top 10 vulnerabilities for 2021. https://www.infosectrain.com/blog/owasp-top-10-vulnerabilities-2021-revealed/, 11 2021.

[5] Omer Said Ozturk, Emre Ekmekcioglu, Orcun Cetin, Budi Arief, and Julio César Hernández-Castro. New Tricks to Old Codes: Can AI Chatbots Replace Static Code Analysis Tools? In *EICC '23: Proceedings of the 2023 European Interdisciplinary Cybersecurity Conference*, pages 13–18. Association for Computing Machinery, 2023.

[6] Own representation.

[7] M.Sc. Sandro Hartenstein. Machine Learning und IT Security. https://www.researchgate.net/publication/332230556_Machine_Learning_und_IT_Security, 04 2019.

[8] Jordan Shropshire. Natural Language Processing as a Weapon, 2018.

Analyse und Evaluation von LLMs zur Generierung von idiomatischem Rust Code

Alexander Huebener **Markus Enzweiler**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma IT-Designers GmbH, Esslingen Zell

Einleitung

Large Language Models, kurz LLMs, sind durch die große Menge an verfügbaren Trainingsdaten ein Tool geworden, das in verschiedenen Bereichen eingesetzt werden kann. So finden LLMs auch Einzug bei der Softwareentwicklung. Durch die Möglichkeit, basierend auf natürlicher Sprache verschiedene Aufgaben, wie zum Beispiel Code zu generieren, Fehler in Code zu beheben, kann die Hürde zwischen menschlicher Sprache und Computersprache überwunden werden und einen leichteren Zugang zum Programmieren gegeben werden. Des Weiteren sollen LLMs mit zum Beispiel CopilotX [4] Einzug in die IDE erhalten, um den Programmierer aktiv beim Programmieren, durch Vervollständigen von Code zu unterstützen. Somit kann der Code für Aufgaben, die häufig umgesetzt werden müssen, wie zum Beispiel das Sortieren einer Liste, schnell generiert werden.
Die Sprache Rust ist Typen basiert und soll aufgrund der Eigentümerschaft Speicher sicher sein [10]. Zum Beispiel durch die Eigentümerschaft soll Rust eine steile Lernkurve haben, was für Einsteiger eine Hürde ist. Durch die Möglichkeit von LLMs Code basierend auf natürlicher Sprache umzusetzen, können diese ein Tool sein, besser in die Sprache zu finden und sich bei der Programmierung unterstützen zu lassen.
Bei der Sprache Rust, vor allem bei idiomatischem Rust, gibt es Besonderheiten, welche beim Schreiben von Code beachtet werden sollen. Dies soll zu einem wartbaren und einheitlicheren Code führen [12]. Bei der Verifizierung der Richtigkeit einer generierten Rust Funktion sind in erster Linie das erfolgreiche Kompilieren und Ausführen von Bedeutung. Dennoch sollte der Code nicht gänzlich von Rust Idiomen abweichen.

Umfang der Arbeit

In dieser Arbeit soll analysiert werden, wie die Codegenerierung mit unterschiedlichen LLMs, speziell von Rust Code funktioniert und bewertet werden kann. Hierzu soll die Evaluation auf State-of-the-Art Benchmark-Tests basieren und größtenteils automatisiert durchführbar sein.
Bei der Codegenerierung sollen kleine, kontextfreie Funktionen betrachtet werden. Dies bedeutet, die zu generierende Funktion kann ohne Wissen über die Funktionsarchitektur und höhere Abstraktionen rein auf der Basis des Funktionskopfes und der Definition der Funktionalität generiert werden. Hieraus soll evaluiert werden, welche Fehlerquellen auftreten.
Des Weiteren soll getestet werden, ob das LLM basierend auf der Rust-Compiler Ausgabe seine generierten Fehler verbessern und ein lauffähiges Ergebnis erzeugen kann.
Es sollen händisch geschriebene Benchmark-Aufgaben erstellt werden, die auf Rust spezifische Funktionen abzielen. Diese sollen genauer testen, ob die LLMs idiomatisches Rust in der generierten Funktion anwenden und umsetzen können.
Basierend auf den Tests soll ein Fazit gegeben werden, ob und welche LLMs für die Generierung von Rust verwendet werden können, welche Fehler häufig auftreten und ob der generierte Code von einem LLM basierend auf der Compiler Ausgabe verbessert werden kann.

Idiomatisches Rust

Die Sprache Rust ist eine typisierte Sprache, mit primitiven Typen wie zum Beispiel Integer, mit sowohl behaftetem Vorzeichen als *i8*, *i16* bis *i128*, als auch nicht behaftetem Vorzeichen als *u8*, *u16* bis *u128*. Aus primitiven Typen können neue Typen erstellt werden. Die Sprache beinhaltet verschiedene Konzepte, wie zum Beispiel das funktionale Programmieren [10].
Unter idiomatisches Rust fallen oft verwendete Styles, Richtlinien und Pattern, die größtenteils von der Community akzeptiert sind [12]. Sie sollen helfen, geschriebenen Rust Code verständlicher und folglich den Code besser wartbar zu machen. Im Folgenden ist ein Auszug an Idiomen aufgelistet:

- Verwende Iteratoren und Closures [10]

- Verwenden von *match* Operator anstelle von *it else* Konstrukten.
- Rückgabewerte sollen am Ende einer Funktion ohne *return* Statement und Semikolon zurückgegeben werden.
- Verwende sichere Integer Umwandlung, wie zum Beispiel *u32::try_from()*

Auswertemethode

In der Arbeit sollen größtenteils automatisierte Messmethodiken verwendet werden. Hierzu werden Benchmarks mit Aufgaben verwendet, aus welchen die LLMs Funktionen generieren sollen. Im Anschluss werden die Funktionen auf richtige Funktionalität mit Unit-Tests geprüft [2].
HumanEval ist ein Open-Source verfügbarer Benchmark Datensatz, der 164 für die Sprache Python handgeschriebene Aufgaben beinhaltet [2].
Diese Aufgaben sind mit Compilern auf weitere Sprachen erweitert worden, wobei die Funktionsbeschreibungen nahezu dieselben sind und die Funktionsdefinitionen und Unit-Tests an die Sprachen angepasst sind. Auf diese Weise sind die Benchmark-Datensätze *HumanevalPack* und *MultiPL-E* entstanden, die weitere Sprachen wie Java oder auch Rust beinhalten.
Als Metrik wird die *pass@k* Metrik, die von [5] aufgestellt und von [2] überarbeitet wurde, verwendet. Für k wird meistens 1, 10 oder 100 verwendet [2] [7] [6]. Die Metrik berechnet pro Benchmark-Aufgabe einen unvoreingenommenen Schätzwert, bezieht sich dabei auf die Anzahl der generierten Samples n, die Anzahl der Korrekten Samples c und den Parameter k der Aufgabe.
Die folgenden drei Schritte müssen bei den Benchmarks gemacht werden:

- Zuerst werden n Samples für jede Aufgabe von dem LLM generiert.
- Die generierten Samples werden mit dem Rust-Compiler kompiliert und die kompilierten Unit-Tests ausgeführt.
- Daraufhin werden die korrekten Samples, das heißt die kompiliert werden konnten und deren Unit-Tests korrekt sind, mit der *pass@k* Metrik bewertet.

Im Anschluss zu der Durchführung der Benchmarks basierend auf den verfügbaren Datensätzen, sollen weitere Aufgaben, gezielt für die Sprache Rust, erstellt werden. Das LLM soll Funktionen in der Sprache Rust generieren, die der Funktionsbeschreibung entsprechen. Die Aufgaben sind im Stil des *HumanevalPack* Benchmark-Datensatzes.
Die Aufgaben können Funktionen, Enumerationen oder Strukturen als Kontext haben, welcher dem LLM in der Prompt mitgegeben wird. Dies soll zum Beispiel zeigen, dass das LLM beim Aufrufen der Kontextfunktion sich an die Funktionsdefinition hält und die Parameter beim Funktionsaufruf richtig übergibt.
In der Funktionsbeschreibung werden die Datentypen der Übergabeparameter und der des Rückgabeparameters mit genannt. Im Folgenden ist ein Beispiel für eine solche Aufgabe:
Write an idiomatic Rust function 'subtract_without_underflow(a: u32, b: u32) -> u32' to solve the following problem: The function takes in 'a' and 'b' which are 'u32' and returns 'a - b' as 'u32'. A underflow should result in '0'.
Die folgende Funktion ist eine erwartete Lösung für die Aufgabe. Diese löst die Aufgabe, indem eine von Rust implementierte Funktion verwendet wird. Der Vorteil hierbei ist, dass eine in LLVM implementierte Funktion *llvm.usub.sat.i32* durch den Rust-Compiler generiert wird, welche für die jeweilige Plattform optimalen Assembler Code erzeugt. Schlussendlich ist diese Lösung wartbarer als *if else* Konstrukte.

```
fn subtract_without_underflow(a: u32, b: u32) -> u32 { a.saturating_sub(b) }
```

Durchführung der Benchmarks

Für die Durchführung der Benchmarks werden die Open-Source LLMs *CodeLLAMA instruct 7B* und *34B* [9], sowie *WizardCoder 15B* [6] und *OctoCoder 16B* [7] verwendet. Diese Auswahl wird aufgrund der folgenden Kriterien getroffen:

- Die Modelle sind alle auf Anweisungen abgestimmt.
- Es soll von verschiedenen Modellgrößen ein Modell dabei sein.
- Aufgrund von online angegebenen *pass@1* Metrik werden von Python Benchmarks.

GPT 4 [8] wird als Closed-Source LLM in der Version *1106-preview* verwendet. Dieses Modell wird gewählt, da es aufgrund eigener Tests und ebenso basierend auf der *pass@1* Metrik von Python Benchmarks eins der besten Modelle ist.
Als Benchmark-Datensatz wird der *HumanevalPack* Datensatz verwendet. Hierbei werden die Felder *instruction* und *prompt* als Kombination oder nur das Feld *instruction* als Prompt verwendet. Zusätzlich wird die jeweils empfohlene Prompt-Vorlage der LLMs als Hülle genutzt. Somit sieht zum Beispiel eine Prompt für *CodeLLAMA* wie folgt aus, wobei für die *system* Eingabe *Provide answers in Rust.* verwendet wird und die *instruction* Eingabe die *instruction* aus dem Datensatz ist:

<s>[INST] «SYS»\\n{system}\\n«/SYS»\\n\\n{instruction}[/INST]
Für jedes Modell, mit Ausnahme von *GPT 4* bei dem es nur ein Sample ist, werden pro Aufgabe 10 Samples generiert. Zum einen soll die Anzahl der Samples höher sein als der Parameter *k* der *pass@k* Metrik, zum anderen wird sie aus Kostengründen\footnote{Die Inferenz der Modelle wird mit online gehosteten Modellen durchgeführt, auf die mit APIs zugegriffen wird. Hierbei fallen entweder pro Zeit, in der das Modell im Speicher ist, oder pro Anzahl generierter Token, Kosten an.} nicht allzu hoch gewählt.
Bei der Inferenz sind die Parameter *temperature* auf 0.2 (außer *GPT 4* auf 0.8), *do_sample* auf *True*, *top_p* auf 0.95 gesetzt. Diese Werte sind aus den Papern der LLMs entnommen, sowie von einem online Leaderboard [1].

Analyse der Benchmark Ergebnisse

Zum Zeitpunkt der Erstellung der Arbeit ist die Analyse der Ergebnisse noch nicht vollständig abgeschlossen. Dennoch sind in dem Graph ref 1 erste Messergebnisse der *pass@1* Metrik aufgetragen.
In *Rot* sind die Ergebnisse der im Zuge dieser Arbeit durchgeführten Messungen von dem *HumanevalPack* Datensatz. In *Gelb* sind Referenzmessungen aus einem online zu findenden Leaderboard von Open-Source LLMs dargestellt. Die in *Grau* gefärbten Balken sind Referenzen zu gemessenen Werten für Python.

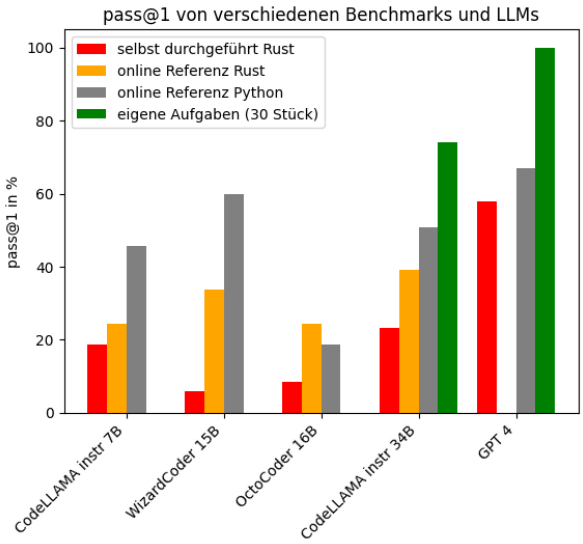


Abb. 1: Messergebnisse der LLM Benchmarks. [3]

Die Abweichung der durchgeführten Benchmarks und der online Referenz muss noch genauer analysiert werden. Die folgenden Ursachen können die Benchmarks beeinflussen:

- Die online Referenz gibt nicht ganz genau an, welche Prompts schlussendlich genommen wurden. Die gestellten Aufgaben sind dieselben, nur die Formatierung der Prompts kann sich unterscheiden.
- Es können unterschiedliche Informationen in den Prompts mitgegeben werden. Hierbei gibt es verschiedene Kombinationen:
- Die online Referenz generiert pro Aufgabe 50 Samples. Dies kann die Chance erhöhen, dass das LLM mindestens ein richtiges Sample erzeugt.

Auch wenn die Werte noch abweichen, kann man aus den Messungen ablesen, dass *GPT 4* mit einem Wert von 58.0% am besten bei dem Benchmark abschneidet. Dieser Wert kommt nahezu an den Python Wert heran. Bei den Open-Source LLMs ist *CodeLLAMA 34B* mit 23.1% selbst gemessen und 39.26% von der Referenz am nächsten an *GPT 4*.
Bei *CodeLLAMA 34B* sind es 618 Compiler-Errors und 592 Ausführfehler der Unit-Tests. Die häufigsten Fehlertypen sind die Folgenden:

- E0308: Ein anderer Typ wurde erwartet. Ein Beispiel hierfür ist, dass eine Funktion ein *\&str* erwartet, aber ein *String* Typ übergeben wird. [11]
- E0433: Nicht definiertes Crate, Modul oder Type. Ein Beispiel hierfür ist, dass ein Typ verwendet wird, der nicht existiert oder nicht importiert ist. [11]
- E0599: Eine auf einem Typ verwendete Methode wurde nicht für diesen implementiert. Ein Beispiel hierfür ist, dass versucht wird, eine Methode auf einem Typ aufzurufen, die nicht existiert. [11]
- E0277: Die Implementation eines Traits fehlt für einen Typ. Ein Beispiel hierfür ist, dass versucht wird, zwei Variablen zu vergleichen, wobei eine davon eine Referenz ist. [11]

Die *grünen* Balken stellen erste Messungen des im Zuge dieser Arbeit erstellen Benchmark-Datensatzes für Rust dar. Bei diesen Messungen wird, für eine erste Einordnung, zuerst ein Sample pro Aufgabe erzeugt.
Bei den erstellen Rust Aufgaben erzielt *GPT 4* einen *pass@1* Wert von 100%. Das bedeutet, jedes generierte Sample kompiliert und alle Unit-Tests sind korrekt. Eine Erklärung für dieses Ergebnis kann die niedrigere Komplexität der Aufgaben sein. Das bedeutet, dass Aufgaben mit Funktionen, die wenige Operatoren und Funktionsaufrufe beinhalten, gelöst werden können. Außerdem sind die Aufgabenbeschreibungen genauer auf Rust abgestimmt. So werden zum Beispiel die Typen von Parametern in der Aufgabenbeschreibung angegeben.

CodeLLAMA 34B liegt mit 25.8% hinter *GPT 4*. Hierbei sind es 7 Compiler-Errors und 1 Ausführfehler der Unit-Tests. Bei den Compiler-Errors kommen die Fehlertypen *E0433*, *E0599*, *E0277* vor, die im vorherigen Kapitel erläutert wurden.
Für weitere Analysen auf idiomatisches Rust ist es wichtig, dass die Funktionen für die Aufgaben des Benchmarks richtig generiert werden.

Weiteres Vorgehen

Die initialen Benchmarks sollen nochmals genauer analysiert werden, warum die gemessenen Werte von den online angegebenen Werte abweichen.
Des Weiteren soll versucht werden, die Samples, bei denen Compiler-Errors aufgekommen sind, mit dem LLM automatisch zu verbessern. Hierbei sollen die Fehlermeldungen des Rust-Compilers dem LLM einen guten Kontext geben. In [*fixing-compiler-errors] wurde dies mit *GPT 4* und *GPT 3.5* durchgeführt und es konnten einige Compiler-Errors verbessert werden.
Zuletzt sollen die durch die erstellten Rust Benchmark Aufgaben generierten Funktionen explizit auf verwendete Rust Funktionen und Operatoren getestet werden. Hierzu gehören zum Beispiel Iterator-Funktionen oder der *match* Operator. Hierbei soll pro Aufgabe in dem Benchmark festgelegt werden, welche Funktionen und Operatoren erwartet werden, basierend auf einer idiomatischen Rust Lösung, und anhand von dem AST der generierten Samples bestimmt werden, ob diese erhalten sind.

Fazit

Die bis zu diesem Zeitpunkt durchgeführten Benchmarks ergeben einen ersten Eindruck, wie die LLMs Rust Funktionen generieren können. Hierbei gibt es große Unterschiede der unterschiedlich großen LLMS. Das Modell *GPT 4* hat sowohl bei dem Benchmark *HumanevalPack*, als auch bei dem selbst erstellten Benchmark am besten abgeschnitten. Dahinter folgt als Open-Source Modell *CodeLLAMA instruct 34B*.
Häufige Compiler-Errors resultieren aus Vergleichen zwischen Variablen mit verschiedenen Typen. Außerdem ist eine häufige Fehlerquelle, dass Funktionen aufgerufen werden, die nicht existieren oder Typen verwendet werden, die es nicht gibt oder nicht importiert sind.
Erste Eindrücke der generierten Samples bezüglich idiomatischem Rust sind, dass Iteratoren oft verwendet werden. Außerdem werden größtenteils Methoden verwendet, die für Typen implementiert sind, wie zum Beispiel *to_lowercase()*. Variablen werden ab und zu mit einem *return* Statement am Ende einer Funktion zurückgegeben.

Literatur und Abbildungen

[1] Non-Profit BigCode. Big Code Models Leaderboard. https://huggingface.co/spaces/bigcode/bigcode-models-leaderboard, 2023.

[2] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Pinto Yuan, et al. Evaluating Large Language Models Trained on Code. *arXiv:2107.03374v2*, 2021.

[3] Eigene Darstellung.

[4] Inc. GitHub. The world's most widely adopted AI developer tool. https://github.com/features/copilot, 2023.

[5] Sumith Kulal, Panupong Pasupat, Kartik Chandra, et al. SPoC: Search-based Pseudocode to Code. *arXiv:1906.04908*, 2019.

[6] Ziyang Luo, Can Xu, Pu Zhao, Qingfeng Sun, et al. WizardCoder: Empowering Code Large Language Models with Evol-Instruct. *arXiv:2306.08568v1*, 2023.

[7] Niklas Muennighoff, Qian Liu, Armel Zebaze, et al. OctoPack: Instruction Tuning Code Large Language Models. *arXiv:2308.07124v1*, 2023.

[8] Company OpenAI. GPT-4 Technical Report. *arXiv:2303.08774v3*, 2023.

[9] Baptiste Rozière, Jonas Gehring, Fabian Gloeckle, Sten Sootla, et al. Code Llama: Open Foundation Models for Code. *arXiv:2308.12950*, 2023.

[10] Team Rust. The Rust Programming Language. https://doc.rust-lang.org/stable/book/, 2021.

[11] Team Rust. Error codes index. https://doc.rust-lang.org/error_codes/error-index.html, 2023.

[12] Unofficial Rust. Rust Design Patterns. https://rust-unofficial.github.io/patterns/idioms/, 2023.

Umgang mit Ausnahmen (Exceptions) bei Gleitkomma-Operationen in sicherheitskritischen Systemen

Seid Jadadic **Markus Enzweiler**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma Robert Bosch GmbH, Stuttgart

Motivation

Die fortschreitende Integration automatisierter Fahrfunktionen in die Verkehrstechnologie verspricht erhebliche Vorteile, darunter verbesserte Verkehrssicherheit, optimierte Verkehrsflüsse und reduzierte Emissionen. Die präzise Kontrolle dieser Funktionen erfordert komplexe Software- und Hardwarelösungen, insbesondere in Bezug auf Gleitkommaoperationen.
Diese Operationen sind entscheidend für komplexe Berechnungen, die für die Fahrzeugwahrnehmung und -steuerung unerlässlich sind. Trotz ihrer Bedeutung bergen Gleitkommaoperationen ein Risiko in Form von Ausnahmen, die in sicherheitskritischen Anwendungen wie dem automatisierten Fahren schwerwiegende Folgen haben können.
Diese Arbeit konzentriert sich darauf, Lösungen für den Umgang mit Gleitkommaausnahmen in sicherheitskritischen Systemen zu entwickeln, insbesondere im Kontext des automatisierten Fahrens. Ziel ist es, innovative Ansätze vorzustellen, und diese anhand bestehender Sicherheitsnormen zu untersuchen um die Sicherheit und Zuverlässigkeit automatisierter Fahrfunktionen zu verbessern. Durch dieses Verständnis trägt die Arbeit zur Weiterentwicklung der Verkehrssicherheit bei.

Defintion von Gleitkommazahlen

Die grundlegende Konzeption von Gleitkommazahlen offenbart sich in der systematischen Struktur, wie sie in Abbildung 1 dargestellt ist. Hierbei erfolgt die Aufteilung einer Zahl in drei Komponenten.

- Vorzeichenbit (Sign): Das Vorzeichenbit, auch als Sign-Bit bezeichnet, bestimmt, ob die Gleitkommazahl positiv oder negativ ist. Ein Wert von 0 im Sign-Bit repräsentiert eine positive Zahl, während ein Wert von 1 eine negative Zahl darstellt.
- Exponent: Der Exponent bestimmt die Lage des Dezimalpunkts und somit den Wertebereich der Gleitkommazahl. Er kann positiv oder negativ sein und erlaubt die Darstellung von sehr kleinen (durch Verschieben des Dezimalpunkts nach links) oder sehr großen (durch Verschieben nach rechts) Zahlen.
- Mantisse (Signifikant): Die Mantisse ist der Teil der Gleitkommazahl, der die Dezimalstellen repräsentiert. Sie besteht aus einer bestimmten Anzahl von Binärstellen, die es ermöglichen, Brüche und Dezimalzahlen darzustellen. Die Länge der Mantisse beeinflusst die Genauigkeit der Zahlendarstellung.

Die grundlegende Idee basiert auf der Vorstellung, dass eine Gleitkommazahl in der Form $M \cdot 2^E$ dargestellt wird, wobei M die Mantisse und E der Exponenten repräsentiert. Das Vorzeichenbit bestimmt, ob die Zahl positiv oder negativ ist [6].

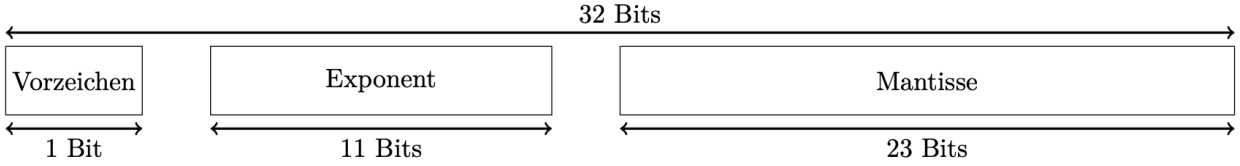


Abb. 1: Struktur einer 32-Bit-Gleitkommazahl. [1]

Gleitkommaausnahmen

Im IEEE 754-Standard, der sich mit Gleitkommazahlen und deren Operationen befasst, werden verschiedene Varianten von Ausnahmen definiert. Diese Ausnahmen sind entscheidende Ereignisse, die während Gleitkommaoperationen auftreten können und darauf hinweisen, dass eine bestimmte Berechnung nicht ordnungsgemäß durchgeführt werden kann. Nachfolgend sind einige der Hauptvarianten von Gleitkommaausnahmen aufgelistet.

1. Division by Zero: Dies tritt auf, wenn eine Zahl durch Null geteilt werden soll, was mathematisch nicht definiert ist. In solchen Fällen wird eine Division durch Null als Fehler behandelt und eine Ausnahme ausgelöst.
2. Overflow: Ein Overflow tritt auf, wenn das Ergebnis einer Berechnung größer ist als das maximale darstellbare Ergebnis für das verwendete Gleitkommazahlenformat. Dies kann dazu führen, dass das Ergebnis nicht mehr korrekt repräsentiert werden kann.
3. Underflow: Im Gegensatz zum Overflow tritt ein Underflow auf, wenn das Ergebnis einer Berechnung kleiner als das minimale darstellbare Ergebnis für das Gleitkommazahlenformat ist. Dies führt dazu, dass das Ergebnis zu klein ist, um korrekt dargestellt zu werden.
4. Invalid Operation: Eine ungültige Operation wird ausgelöst, wenn eine mathematische Operation nicht sinnvoll ist, wie beispielsweise die Wurzel aus einer negativen Zahl zu ziehen oder den Logarithmus einer nicht positiven Zahl zu berechnen.

Vermeidung von Gleitkommaausnahmen

Die identifizierten Strategien, Raumredundanz und Zeitredundanz, repräsentieren unterschiedliche Ansätze zur Vermeidung von Gleitkommaausnahmen in verschiedenen Industriebereichen. Als nächstes erfolgt eine detaillierte Analyse dieser Methoden, um ihre Eignung für sicherheitskritische Systeme, insbesondere im Bereich des automatisierten Fahrens, zu bewerten. Die Abbildung 2 illustriert die vier Hauptansätze zur Vermeidung von Softwarefehlern. In (a), (b) und (c) repräsentiert das Symbol τ_k den j-ten Job der i-ten Aufgabe, wobei k die Redundanz oder Re-excecution des Jobs identifiziert. In (d) hingegen stellt $\tau_{i,j,k}$ den k-ten Teilen Jobs $\tau_{i,j}$ dar. Im dargestellten Beispiel besteht der Job $\tau_{1,1}$ aus drei Teilen.

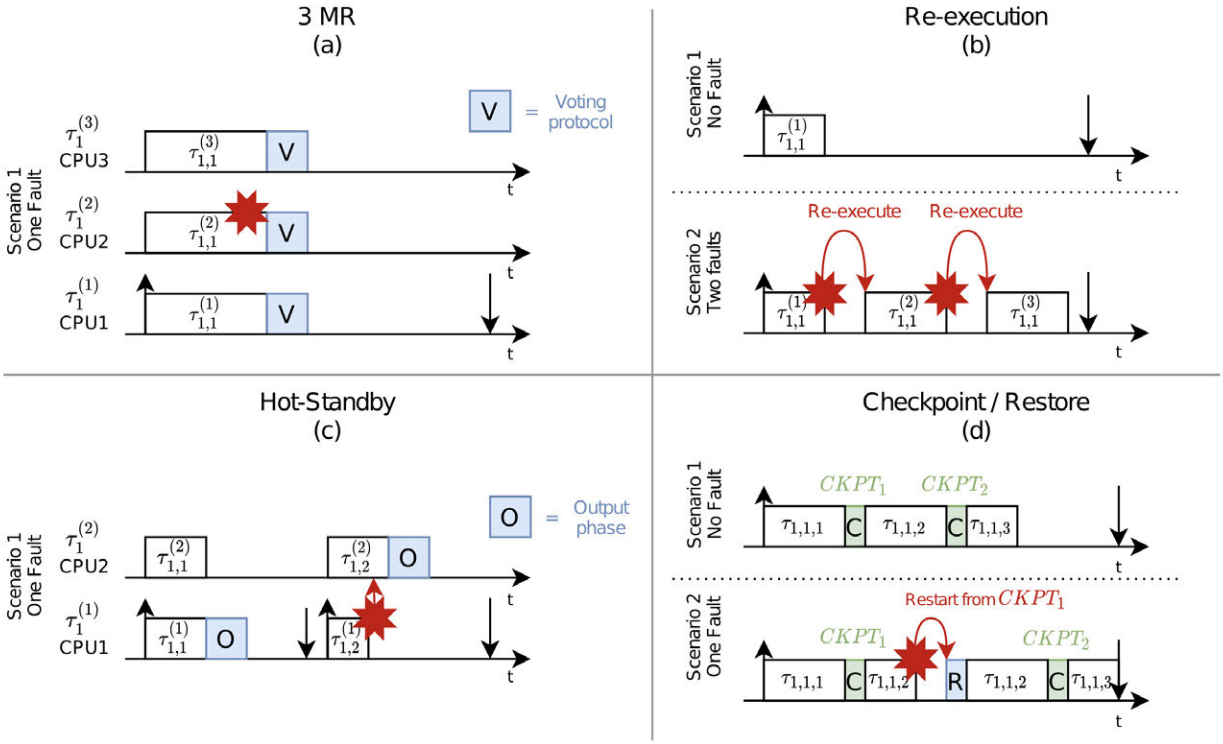


Abb. 2: Vier Ansätze zur Vermeidung von Softwarefehlern [7]

Raumredundanztechniken

Raumredundanz, ein zentraler Aspekt sicherheitsrelevanter Systeme, bedient sich der N-modularen Redundanz (NMR) als Software Fehlerbehandlungsmethode, die der konventionellen Hardwareredundanz ähnelt [7]. Hierbei erfolgt die vervielfältigte Ausführung kritischer Aufgaben N-fach, wobei die Hauptaufgabe und ihre Repliken identische Operationen ausführen. Dieses Verfahren verwendet Softwareabstimmungen zur Bestimmung korrekter Ergebnisse.
Die Taxonomie von Softwareabstimmungen im Rahmen der NMR-Technik wird detailliert im Werk von Latif-Shabgahi et al. [4] erläutert. Der Fokus liegt auf der Implementierung dieser Abstimmungsmechanismen in der Software, um eine wirksame Fehlererkennung und -wiederherstellung zu gewährleisten. Insbesondere in sicherheitskritischen Umgebungen ist die hohe Zuverlässigkeit dieser Abstimmungsmechanismen essenziell, um die Systemintegrität zu sichern.
Um ein besseres Verständnis für NMR zu gewinnen, kann das Prinzip der Triple Modular Redundancy (TMR) betrachtet werden. Letztere ist eine spezielle Ausprägung der N-modularen Redundanz und dient dazu, die Zuverlässigkeit von Systemen in sicherheitskritischen Umgebungen zu erhöhen.

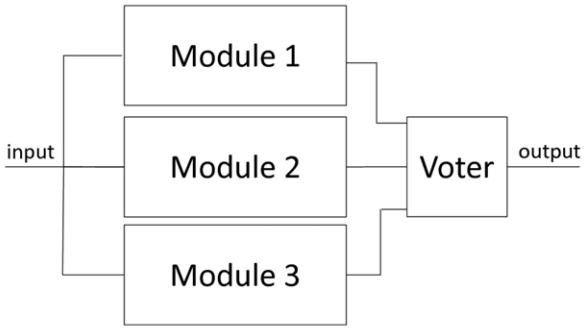


Abb. 3: Triple Modular Redundancy (TMR) Architektur [5]

Die TMR-Architektur besteht aus drei funktional identischen Komponenten (Module) und einer Abstimmungseinheit (Voter), wie in Abbildung 3 dargestellt. Diese Module führen simultan dieselben Operationen aus. Im Falle eines Fehlers in einem der Module wird dieser durch die Mehrheitsentscheidung der beiden verbleibenden, fehlerfreien Module neutralisiert. Dieses Prinzip ermöglicht die Maskierung von Einzelfehlern und gewährleistet, dass das System auch bei einem Ausfall einer einzelnen Komponente korrekte Ergebnisse produziert. Die Architektur ist besonders effektiv in sicherheitskritischen Systemen, in denen eine hohe Zuverlässigkeit und Fehlertoleranz von entscheidender Bedeutung sind [5]. Die Anzahl der Repliken ist in der Regel unveränderlich. Aus diesem Grund wird NMR auch in die Kategorie der statischen Redundanztechniken eingeordnet [7].
Hingegen existiert die dynamische Form, unter dem Namen Standby-Redundanz [3]. In diesem Szenario bleiben (N-1) Repliken im Standby-Modus, aktivieren jedoch erst eine vollständige Berechnung, wenn ein Fehler identifiziert wird. Es ist möglich, Standby Redundanztechniken weiter zu unterteilen in Hot-Standby und Cold-Standby, wobei ihre Anwendungsbereiche und Einflüsse auf die Raumredundanz von entscheidender Relevanz sind.
Bei Hot-Standby im normalen Betrieb führen sowohl die Hauptaufgabe als auch die Repliken Berechnungen durch. Während des normalen Betriebs werden die Ergebnisse der Repliken jedoch unterdrückt, und nur das Ergebnis der Hauptaufgabe wird verwendet. Im Falle eines Fehlers wird das Ergebnis der Hauptaufgabe unterdrückt, und eine der Repliken wird als neue Hauptaufgabe aktiviert. Die Reaktionszeit bei einem Fehler ist normalerweise kurz, da die redundanten Komponenten bereits Berechnungen durchgeführt haben.
In Cold-Standby Systemen führen die Repliken keine Berechnungen durch. Sie aktualisieren lediglich ihren internen Zustand, um synchron mit der Hauptaufgabe zu bleiben, während die Hauptaufgabe alle Berechnungen durchführt. Bei einem Fehler wird eine der Cold-Standby Repliken aktiviert und übernimmt die volle Berechnungsaufgabe als neue Hauptaufgabe. Die Reaktionszeit bei einem Fehler ist normalerweise länger im Vergleich zu Hot-Standby, da die Repliken erst vollständige Berechnungen übernehmen müssen.

Zeitredundanztechniken

Zeitredundanz stellt eine weitere Methode zur Fehlervermeidung dar. Die grundlegendste Technik zur Vermeidung von Softfehlern besteht darin, das gleiche Programm zweimal auszuführen und das korrekt erachtete Ergebnis zu erzielen, wenn die Ausgaben beider Ausführungen übereinstimmen [2]. Unter dieser Technik sind einige der gängigsten die Re-excecution, Checkpoint/Restart, Recovery Blocks und Forward Error Recovery.
Der Re-excecution Ansatz, auch als Retry bezeichnet, startet einen Job erneut, möglicherweise mehrmals, wenn ein Fehler erkannt wird. Die zusätzliche Berechnung ist identisch mit der des ursprünglichen Jobs. Bei Bedarf wird nach einem Fehler ein Verfahren namens Rückwärtsfehlerwiederherstellung durchgeführt, um den Zustand des Jobs oder der Umgebung wiederherzustellen. In einigen vorangegangen Arbeiten werden aufgrund von Planungsgründen auch derzeit unterbrochenen Jobs erneut ausgeführt, selbst wenn sie nicht von dem Fehler betroffen sind. Diese Technik wird als Multiple Recovery bezeichnet.

Eine Weiterentwicklung dieser Methode ist das Checkpoint/Restart Verfahren. Hierbei werden periodisch Zustände der Aufgaben gespeichert (Checkpoint). Bei Erkennung eines Fehlers wird von diesem letzten verfügbaren Checkpoint aus neugestartet (Restart). Im Gegensatz zur Re-excecution muss die Berechnung nicht von vorne beginnen, was die Reaktionszeit im Fehlerfall reduziert. Allerdings führt dies zu zusätzlichen Overhead für den Checkpoint, selbst wenn keine Fehler auftreten. Insbesondere im Bereich des High-Performance Computing (HPC), bei dem Aufgaben Stunden oder Tage dauern können, ist das Neustarten von Grund nicht wünschenswert.
Recovery Blocks stellen eine weitere Variation der Re-excecution dar. Bei einem Fehler wird eine andere Version des Codes ausgeführt, um dieselbe Funktion zu erfüllen, möglicherweise mit degradierter Leistung. Diese Technik ermöglicht auch die Behandlung von gemeinsamen Fehlern, wie beispielsweise unerwarteten Softwarefehlern. Eine andere Strategie, die als Forward Error Recovery bezeichnet wird, beinhaltet die Ausführung einer speziellen Routine, die den Fehler ohne erneuter Ausführung der Originalaufgabe behebt. Dies geschieht normalerweise mit einem Mechanismus zur Fehlerbehandlung, bei dem eine Ausnahme ausgelöst wird und dedizierter Code die fehlerhafte Bedingung verwaltet. Diese Methode wird typischerweise in Speicher- und Netzwerkalgorithmen verwendet, die Fehlerkorrekturcodes einsetzen, um den Fehler zu beheben. Jedoch ist der Anwendungsbereich dieser Technik in der Regel auf wenige Anwendungsszenarien beschränkt.

Behandlung von Gleitkommaausnahmen

Gleitkommaausnahmen können durch verschiedene Methoden behandelt werden, wobei die Hauptziele darin bestehen, die Auswirkungen dieser Ausnahmen zu minimieren und die Systemintegrität aufrechtzuerhalten. In Tabelle 4 sind verschiedene etablierte Ansätze aufgeführt, die zur Bewältigung von Gleitkommaausnahmen herangezogen werden. Nachfolgend erfolgt eine detaillierte Erläuterung der einzelnen Spalten, um die Bedeutung der dort aufgeführten Kriterien zu verdeutlichen.

Ansatz	Debouncing	Degradation	HMI-Information	Fehlereintrag	Trap
Fehlereaktion ohne debouncing	Nein	Ja	Ja	Ja	Nein
Fehlerreaktion mit debouncing	Ja	Ja	Ja	Ja	Nein
Floating-Point Exception trappen	Nein	Nein	Nein	Nein	Ja
Keine Fehlerreaktion	Nein	Ja	Nein	Nein	Nein

Abb. 4: Ansätze zur Behandlung von Gleitkommaausnahmen [1]

Der erste Ansatz beinhaltet die sofortige Meldung von Fehlern ohne Debouncing, was eine umgehende Fehleridentifizierung ermöglicht und unmittelbare Sicherheitsmaßnahmen auslöst. Dieses Verfahren gewährleistet eine schnelle Reaktion auf potenziell kritische Situationen in automatisierten Fahrzeugen. Allerdings sollte beachtet werden, dass eine zu häufige Anwendung dieses Prozesses die Systemverfügbarkeit beeinträchtigen könnte.
Ein alternativer Ansatz, der Debouncing einsetzt, zielt darauf ab, kurzzeitige Fehler zu glätten, indem ein Fehler erst nach einer bestimmten Zeitdauer gemeldet wird. Diese Methode gewährleistet eine bessere Systemverfügbarkeit. Allerdings ist dieser Ansatz nicht geeignet für Systeme, die bereits beim ersten Auftreten eines Fehlers degradiert werden sollten.
Ein weiterer Ansatz besteht darin, das System still zu deaktivieren, ohne eine Benachrichtigung an die Benutzerschnittstelle (HMI) oder eine Fehlerprotokollierung auszulösen. Hierbei wird davon ausgegangen, dass der Fahrer über den Deaktivierungszustand des Fahrzeugs informiert ist, insbesondere wenn es nicht ordnungsgemäß funktioniert. Die Protokollierung von Fehlern oder Ausnahmen erfolgt üblicherweise, um Diagnosen bei Wartungsarbeiten oder in der Werkstatt durchführen zu können.
Ein zusätzlicher Ansatz beinhaltet die Verarbeitung von Gleitkommaausnahmen mittels Traps, die während Gleitkommaberechnungen auftretende Ausnahmen behandeln. Traps ermöglichen spezifische Reaktionen auf diese Ausnahmen und bieten eine Kontrollstruktur für Software oder Hardware.
Die Verwendung von Traps führt dazu, dass das System Ausnahmen erkennt und vordefinierte Aktionen auslöst, beispielsweise das Umschalten in einen speziellen Fehlermodus oder die Aktivierung von Sicherheitsmaßnahmen. Allerdings birgt die Trappung von Gleitkommaausnahmen einige Herausforderungen. Eine Hauptproblematik besteht darin, die Häufigkeit der Traps zu kontrollieren, da eine zu hohe Auslöserate die Stabilität und Kontinuität des Systems beeinträchtigen kann.
Zudem können einige Traps zu einem ECU-Reset führen, was einen vollständigen Neustart des Steuergeräts bedeutet. In sicherheitskritischen Systemen ist dies unerwünscht, da die fortlaufende Funktion des Steuergeräts gewährleistet sein sollte. Unerwünschte ECU-Resets können die Systemverfügbarkeit verringern und potenziell gefährliche Situationen verursachen. Daher werden Traps nicht als bevorzugte Methode zur Behandlung von Gleitkommaausnahmen in sicherheitskritischen Systemen betrachtet.

Ausblick

Im weiteren Verlauf richtet sich der Fokus auf die Entwicklung eines umfassenden Konzepts zur Vermeidung von Gleitkommaausnahmen. Diese Strategie kombiniert präventive Maßnahmen und adaptive Reak-

tionsmechanismen, um Fehler frühzeitig zu erkennen und zu behandeln. Dabei stützt sich die Arbeit auf bewährte Praktiken und innovative Technologien, um die Sicherheit und Stabilität in sicherheitskritischen Systemen, insbesondere im Bereich des automatisierten Fahrens, zu optimieren.
Zusätzlich ist die Festlegung klarer Richtlinien für die kontrollierte Systemdegradation bei Gleitkommaausnahmen von Bedeutung. Eine stufenweise Anpassung der Systemleistung in kritischen Situationen kann die Grundfunktionalität des Fahrzeugs bewahren und Risiken minimieren, was die Sicherheit automatisierter Fahrzeuge erhöht und die Bewältigung kritischer Situationen verbessert.

Literatur und Abbildungen

[1] Eigene Darstellung.

[2] Nikolov Dimitar. Fault Tolerance for Real-Time Systems: Analysis and Optimization of Roll-back Recovery with Checkpointing, 2014.

[3] Jay Frank et al. IEEE Standard Glossary of Software Engineering Terminology. *IEEE*, 1990.

[4] G. Latif-Shabgahi, J.M. Bass, and S. Bennett. A taxonomy for software voting algorithms used in safety-critical systems. *IEEE Transactions on Reliability*, 2004.

[5] Trapp Mario, Saglietti Francesca, Spisländer Marc, and Bitsch Friedemann. *Computer Safety, Reliability, and Security*. Springer International Publishing, 2022.

[6] Jean-Michel Muller, Nicolas Brisebarre, Florent De Dinechin, Claude-Pierre Jeannerod, Vincent Lefèvre, Guillaume Melquiond, Nathalie Revol, Damien Stehlé, and Serge Torres. *Handbook of Floating-Point Arithmetic*. Birkhäuser Boston, 2010.

[7] Federico Reghenzani, Zhishan Guo, and William Fornaciari. Software Fault Tolerance in Real-Time Systems: Identifying the Future Research Questions. *ACM Computing Surveys*, 2023.

Analyse und Entwicklung von Kollaborationswerkzeugen zur Unterstützung von Softwareentwicklungsteams

Jasmin Janecek **Harald Melcher**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma pep.digital GmbH, Esslingen am Neckar

Einleitung

In den letzten Jahrzehnten hat sich die Softwareentwicklung zu einem entscheidenden Bereich der Informationstechnologie entwickelt. Softwareentwicklungsteams, wie bei pep.digital, spielen eine zentrale Rolle bei der Gestaltung und Implementierung von Softwarelösungen für verschiedene Lebensbereiche. Die stetige technologische Weiterentwicklung führt zu komplexeren Projekten, die eine effiziente Zusammenarbeit erfordern. Diese Abschlussarbeit fokussiert sich auf die wachsende Bedeutung der Kollaboration in den Softwareentwicklungsteams von pep.digital. In einer von Veränderungen und Innovationen geprägten Umgebung ist die effektive Zusammenarbeit und der Wissenstransfer entscheidend. Interdisziplinäre Softwareprojekte erfordern die Zusammenarbeit verschiedener Teammitglieder und die zunehmende Globalisierung sowie Remote-Arbeit stellen zusätzliche Herausforderungen dar. Die digitale Förderung nahtloser Kollaboration betont die Bedeutung von Kollaborationswerkzeugen und -praktiken. Ziel dieser Arbeit ist es, die Zusammenarbeit in den Softwareentwicklungsteams von pep.digital zu optimieren. Durch die Analyse bestehender Werkzeuge und die Entwicklung innovativer Lösungsansätze sollen Erkenntnisse gewonnen werden, um den Softwareentwicklungsprozess zu optimieren, die Kommunikation zu verbessern und die Produktivität zu steigern.

Ziel

Das Hauptziel dieser Arbeit ist eine Untersuchung der vorhandenen Kollaborationswerkzeuge in Scrum-Teams. Durch Bewertung der Effektivität bestehender Lösungen anhand von Erfahrungen und Rückmeldungen der Teams sollen nicht nur erfolgreiche Werkzeuge bestätigt, sondern auch etwaige Bedarfslücken identifiziert werden. Ein weiteres Ziel ist die Vorstellung potenzieller Verbesserungen und neuer Ansätze zur Unterstützung der Kollaboration in Scrum-Teams. Dazu gehört die theoretische Ausarbeitung neuer Werkzeuge und die Anpassung bestehender Lösungen. Als praktisches Beispiel wird ein Prototyp eines Kollaborationswerkzeugs mit der Svelte-Technologie implementiert. Dieser dient nicht nur als Machbarkeitsnachweis, sondern auch zur Überprüfung der Anpassbarkeit und Integration in bestehende Arbeitsabläufe der Scrum-Teams bei pep.digital. Das Werkzeug zielt darauf ab, die identifizierten Bedürfnisse und Anforderungen der Teams zu erfüllen und einen konkreten Beitrag zur Kollaboration zu leisten.

Säulen und Werte von Scrum

Scrum basiert auf drei zentralen Säulen: Transparenz, Überprüfung und Anpassung. Transparenz ist entscheidend, da sie die Grundlage für wirksame Überprüfungen bildet. Scrum-Events, wie Sprint Planning, Daily Scrum, Sprint Review und Sprint Retrospektive, ermöglichen regelmäßige Überprüfungen, die wiederum Anpassungen im Prozess fördern. Die fünf Werte von Scrum, wie in Abbildung 1 dargestellt, ergänzen diese Säulen. Selbstverpflichtung, Mut, Fokus, Offenheit und Respekt unterstützen das Rahmenwerk. [5]

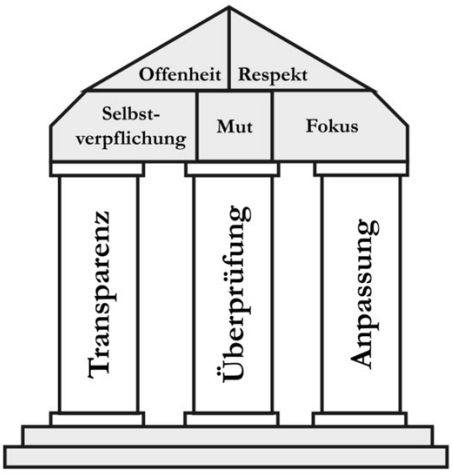


Abb. 1: Säulen und Werte von Scrum [2]

Teammitglieder verpflichten sich kollektiv, komplexe Herausforderungen mutig anzugehen und klare Ziele mit Fokus zu verfolgen. Offenheit fördert transparente Kommunikation über Herausforderungen und Erfolge, was die Basis für kontinuierliche Verbesserungen bildet. Respekt und ein vertrauensvoller Umgang innerhalb des Teams schaffen ein harmonisches und effektives Arbeitsumfeld, das den Erfolg des Entwicklungsprozesses sicherstellt und psychologische Sicherheit gewährleistet. [1]

Bedürfnisse von Scrum-Teams

Scrum-Teams sind in ihrer Arbeitsweise auf eine Vielzahl von essenziellen Bedürfnissen angewiesen, um eine effektive und effiziente Zusammenarbeit zu gewährleisten. Diese sind in Abbildung 2 dargestellt.

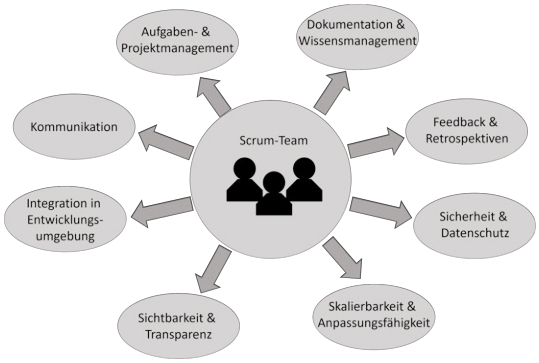


Abb. 2: Bedürfnisse von Scrum Teams [2]

Eine tragende Säule ist die Notwendigkeit einer klaren und transparenten Kommunikation innerhalb des Teams. Dies bezieht sich auf regelmäßige Stand-up-Meetings, den Austausch von Fortschrittsberichten und die kontinuierliche Interaktion während der Sprints. Ein weiterer entscheidender Faktor ist das effiziente Management von Aufgaben und Projekten. Teams benötigen Werkzeuge, die eine klare Zuweisung von Aufgaben ermöglichen, den Fortschritt verfolgen und potenzielle Engpässe identifizieren, um die Sprintziele erfolgreich zu erreichen. Die systematische Dokumentation von Entwicklungsprozessen und erworbenem Wissen spielt ebenfalls eine zentrale Rolle. Werkzeuge, die eine einfache Speicherung und den Austausch von Informationen ermöglichen, unterstützen das Team dabei, kontinuierlich zu lernen und ihr Wissen zu optimieren. Die nahtlose Integration von Kollaborationswerkzeugen in die Entwicklungs- und Testumgebung trägt dazu bei, manuelle Schritte zu minimieren und Entwicklungszyklen zu optimieren. Dies ist besonders wichtig, um einen reibungslosen Arbeitsfluss sicherzustellen. Strukturiertes Feedback und retrospektive Analysen sind wesentliche Elemente für kontinuierliche Verbesserungen in agilen Teams. Werkzeuge, die diese Prozesse unterstützen, fördern eine konstruktive Feedbackkultur und tragen zur Weiterentwicklung der Teamleistung bei. Die klare Sichtbarkeit von Aufgaben, Fortschritt und Hindernissen ist grundlegend für den Projekterfolg. Werkzeuge, die eine transparente Darstellung der Teamaktivitäten ermöglichen, erleichtern die Entscheidungsfindung und stärken die Zusammenarbeit. Flexibilität und Anpassungsfähigkeit der eingesetzten Werkzeuge an die Projektanforderungen sind ebenso wichtig wie ihre Skalierbarkeit, um mit dem Wachstum des Teams oder der Projektkomplexität Schritt zu halten. Schließlich ist die Gewährleistung von Datensicherheit und Datenschutz in verteilten und global agierenden Teams von kritischer Bedeutung. Werkzeuge müssen robuste Sicherheitsmechanismen bieten, um Vertraulichkeit und Integrität der Projektdaten zu garantieren. Insgesamt sind diese Aspekte entscheidend für den Erfolg und die Effektivität agiler Scrum-Teams.

Analyse der vorhandenen Kollaborationswerkzeuge

Im Vorfeld der Implementierung des Kollaborationswerkzeugs bei pep.digital erfolgte eine Untersuchung und Analyse verschiedener Kollaborationswerkzeuge. Diese Auswahl basierte auf den definierten Anforderungskriterien, die im Kontext der Softwareentwicklungsteams von pep.digital als relevant erachtet wurden. Die untersuchten Werkzeuge wurden einer Analyse und Bewertung unterzogen, um ihre Eignung für die spezifischen Bedürfnisse und Arbeitsanforderungen der Scrum-Teams zu bewerten. Im Zuge dieser Analyse wurden Bedarfslücken identifiziert, die auf vorhandene Unzulänglichkeiten oder fehlende Funktionalitäten hinwiesen. Eine dieser identifizierten Bedarfslücken wurde daraufhin als Priorität festgelegt und in der Implementierungsphase des Kollaborationswerkzeugs adressiert. Die Implementierung fokussierte sich somit auf die gezielte Schließung dieser Bedarfslücke, um eine bedarfsgerechtere Kollaborationsumgebung für die Scrum-Teams zu schaffen. Dieser proaktive Ansatz gewährleistet, dass das implementierte Werkzeug nicht nur den aktuellen Anforderungen entspricht, sondern auch auf spezifische Bedürfnisse eingeht, die durch die vorangegangene Analyse der Kollaborationswerkzeuge identifiziert wurden.

Implementierung mit dem Svelte-Framework

Die Entscheidung, Svelte als Grundlage für die Implementierung des Kollaborationswerkzeugs zu wählen, resultiert aus einer Bewertung seiner Merkmale und Vorteile. Der Compiler-Driven-Ansatz von Svelte, der die Entwicklungslogik von Browser- zu Build-Zeit

verschiebt, ermöglicht eine schlankere und effizientere Codebasis, welcher in Abbildung 3 dargestellt ist. [4]

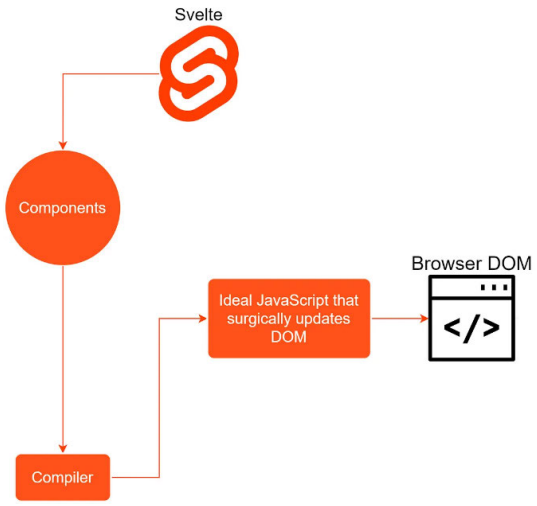


Abb. 3: Aufbau von Svelte [3]

Dies ist von entscheidender Bedeutung, um ein Kollaborationswerkzeug zu gestalten, das sowohl reaktionsschnell als auch performant ist. Die klare Komponentenarchitektur fördert die Erstellung von modularem Code, was für die Implementierung komplexer kollaborativer Funktionen unerlässlich ist. Die reaktive Programmierung und effiziente Datenbindung von Svelte bieten die notwendige Grundlage für die dynamische Aktualisierung des Benutzerinterfaces, insbesondere in einem Umfeld, in dem Echtzeitinformationen und Interaktivität zentral sind. Die intuitive Datenbindung erleichtert die Handhabung von Zustandsänderungen und sorgt für eine nahtlose Benutzererfahrung. Die Integration von Animationen in Svelte ermöglicht es, eine ansprechende Benutzeroberfläche zu schaffen, die die Benutzerfreundlichkeit des Kollaborationswerkzeugs verbessert. Diese visuellen Elemente tragen dazu bei, die Zusammenarbeit intuitiv und effektiv zu gestalten. Die Entscheidung, Svelte zu nutzen, berücksichtigt auch die Herausforderungen, die sich in der Entwicklung von Kollaborationswerkzeugen stellen, wie die Akzeptanz in der Entwicklergemeinschaft und die Integration von Drittanbieter-Bibliotheken.

Ausblick

Die Implementierung des Kollaborationswerkzeugs bei pep.digital auf Grundlage des Svelte Frameworks kennzeichnet einen wesentlichen Schritt in der Optimierung der Softwareentwicklungsprozesse. Das ausstehende Nutzerfeedback der Scrum-Teams bei pep.digital wird entscheidend sein, um spezifische Anpassungen und Erweiterungen am Kollaborationswerkzeug vorzunehmen. Die aktuelle Implementierung fungiert als Ausgangspunkt, während mögliche Anforderungsänderungen oder Optimierungen durch das einfließende Nutzerfeedback adressiert werden. Die Anpassungsfähigkeit von Svelte erlaubt es, gezielt auf die Bedürfnisse und Anregungen der Scrum-Teams einzugehen und das Kollaborationswerkzeug kontinuierlich zu verbessern. Dieser iterative Ansatz unterstreicht die fortlaufende Anpassung des Tools an die dynamischen Anforderungen der Scrum-Teams bei pep.digital.

Literatur und Abbildungen

[1] Jörg Brüggenkamp, Peter Preuss, and Tobias Renk. *Der Scrum-Reiseführer*. UVK Verlagsgesellschaft mbH, 2021.

[2] Eigene Darstellung.

[3] Keshav Kumaresan. React vs. Svelte: The War Between Virtual and Real DOM. https://blog.bitsrc.io/react-vs-sveltejs-the-war-between-virtual-and-real-dom-59cbebbab9e9, 08 2020.

[4] Ionos Redaktion. Svelte – das schlanke Framework für schnelle mobile Webanwendungen. https://www.ionos.de/digitalguide/websites/web-entwicklung/svelte-framework-vorgestellt/, 06 2020.

[5] Ken Schwaber and Jeff Sutherland. Scrum-Guide-German. https://scrumguides.org/docs/scrumguide/v2020/2020-Scrum-Guide-German.pdf, 2020.

Evaluation und Demonstration von Verhaltensprädikation bei Mensch-Roboter-Kollaboration

Jakob Janusch **Steffen Schober**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma Festo SE & Co. KG, Esslingen

Einleitung

Die Integration von Robotern mit sechs Freiheitsgraden in industrielle Fertigungsprozesse in den letzten fünf Jahrzenten hat nicht nur die Effizienz und Produktivität gesteigert, sondern auch neue Herausforderungen hinsichtlich der Arbeitssicherheit aufgeworfen. Die gleichzeitige Anwesenheit von menschlichen Arbeitskräften und Robotern in der Fertigung erfordert umfangreiche Sicherheitsvorkehrungen, wie beispielsweise physische Barrieren und Lichtschranken, um Unfälle und Verletzungen zu minimieren. Allerdings sind diese Maßnahmen nicht nur kostenintensiv, sondern beschränken auch erheblich die Flexibilität der Fertigungsprozesse. Kollaborative Roboter, auch als *Cobots* bekannt, zeichnen sich durch ihre Fähigkeit aus, sicher mit menschlichen Arbeitskräften zu interagieren, ohne physische Barrieren zu benötigen. Diese neue Generation von Robotern geht auf innovative Weise auf Sicherheitsbedenken ein, indem sie fortschrittliche Sensorik und intelligente Steuerungssysteme einsetzt. Dies ermöglicht es den Cobots, in Echtzeit auf Berührungen mit Menschen zu reagieren, indem sie anhalten oder nachgeben, um Verletzungen zu verhindern. Auf diese Weise tragen Cobots nicht nur zur Verbesserung der Arbeitssicherheit bei, sondern steigern auch die Flexibilität und Wirtschaftlichkeit in industriellen Fertigungsprozessen.

Motivation und Ziele

Um die Sicherheit der Menschen in seinem Arbeitsbereich zu garantieren, hält ein Cobot bei jeder Berührung augenblicklich an oder weicht sogar zurück, sobald seine Sensoren Kontakt mit einem Hindernis erkennen. Dieses Verhalten ist für die Sicherheit der Menschen im Arbeitsbereich des Roboters unerlässlich, aber es wirkt auch hemmend auf die Produktivität und die Potenziale der Zusammenarbeit von Mensch und Roboter. In diesem Kontext bietet die Verhaltensprädiktion völlig neue Möglichkeiten. Wenn der Cobot in der Lage ist, die Bewegungen des Menschen vorauszusagen, ermöglicht dies neue und dynamische Formen der Interaktion. Zum Beispiel könnte der Cobot proaktiv Werkzeuge und Materialien bereitstellen oder anreichen, was die Effizienz und den reibungslosen Ablauf in der industriellen Fertigung erheblich steigern könnte. An der Hochschule Esslingen wurde bereits zu diesem Thema geforscht. Ein Ergebnis dieser Forschung ist ein multimodales Modell zur Verhaltensprädiktion im Kontext von industriellen Fertigungsprozessen. Dieses Modell wurde auf dem sogenannten CoAx-Datensatz (kurz für *Collaborative Action Dataset for Human Motion Forecasting in an Industrial Workspace*) trainiert, welcher in Zusammenarbeit mit der Firma Festo erstellt wurde [5]. Dieser CoAx-Datensatz enthält Aufzeichnungen von sechs Personen, die jeweils drei verschiedene Montageaufgaben jeweils zehn Mal wiederholt haben. Insgesamt wurden dabei über einhundert Tausend Bilder aufgezeichnet, was etwa zwei Stunden Videomaterial bei einer Bildrate von fünfzehn Bildern pro Sekunde entspricht [4]. Dass das Modell multimodal ist, bedeutet in diesem Fall, dass es sowohl die aktuelle Handlung einer Person erkennen als auch die nächste vorhersagen kann. Darüber hinaus kann es ebenfalls die Position der Hand der Person für den Zeitraum von einer Sekunde vorhersagen. Das Ziel dieser Arbeit besteht darin, die Fähigkeiten dieses Modells zur Verhaltensprädiktion durch einem Testaufbau zu demonstrieren. Hierzu wird das Szenario aus dem Datensatz nachgestellt und eine Tiefenkamera verwendet, um das Szenario zu erfassen.

Ansatz

Die Kamera vom Typ *RealSense D435* des Herstellers Intel [1] liefert RGBD-Bilddaten mit einer Auflösung von 640x480 Pixeln und einer Bildwiederholfrequenz von 15 Herz. Dies entspricht den aufgezeichneten Bilddaten im CoAx-Datensatz. RGBD steht für *Red Green Blue Depth* und bedeutet, dass für jede Farbe (Rot, Grün und Blau) jeweils ein Kanal verwendet wird, sowie ein weiterer Kanal für die Tiefeninformation,

welcher die Distanz des Objekts zur Kamera enthält. Die Kamera erfasst die gesamte Szene des Testaufbaus. Die Positionen der relevanten Objekte in der Szene werden mithilfe dieser RGBD-Daten ermittelt. Zu diesem Zweck werden zunächst die Objekte in den 2D-RGB-Bildern erkannt und ihre Position im 2D-Bild (siehe Abb. 1) bestimmt. Hierfür wird ein Künstliches Neuronales Netzwerk mir der *Single Shot MultiBox Detector* Architektur verwendet, wie es in [6] beschrieben ist. Als Ausgangsbasis wurde ein SSD-Modell verwendet (bereitgestellt von PyTorch [3]), welches auf dem MS COCO Datensatz (*Common Objects in Context*, bereitgestellt von Microsoft [2]) trainiert wurde. Damit dieses Modell die speziellen Objekte des CoAx-Datensatzes im Bild erkennen und lokalisieren kann, wurden die letzten Layer auf dem CoAx-Daten trainiert. Diese Vorgehensweise wird als Fine-Tuning bezeichnet und ermöglicht es, große Netzwerke auf relativ kleinen Datensätzen zu trainieren. Anhand dieser 2D-Position und der Tiefeninformation im Depth-Kanal wird dann die 3D-Position im Koordinatensystem der Kamera bestimmt. Schließlich erfolgt die Umrechnung dieser Position in das Koordinatensystem des Roboters, welches als Referenzsystem im CoAx-Datensatz und für das Training des multimodalen Modells verwendet wurde. Diese Umrechnung basiert auf der Annahme, dass die Translation und Rotation der Kamera relativ zur Basis des Roboters im Testaufbau konstant sind. Es handelt sich um eine affine Abbildung mit sechs Parametern.

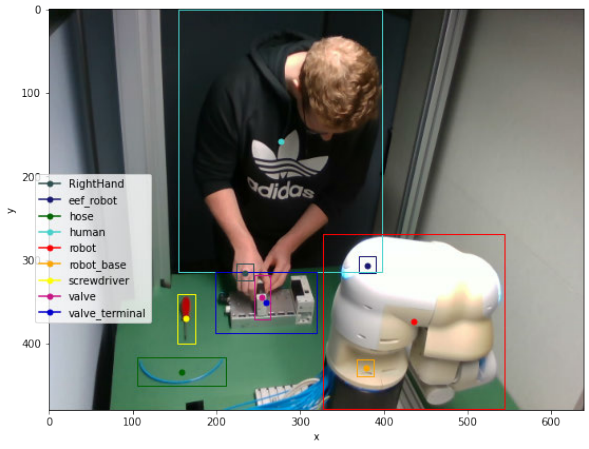


Abb. 1: Szene aus dem CoAx-Datensatz mit markierten Objekten [4]

Die Positionen aller Objekte in der Szene werden dann in einem Graphen gespeichert, der als Frame-Graph bezeichnet wird. Er enthält sämtliche Objekte sowie ihre Positionen im Koordinatensystem des Roboters. Jedes Objekt wird dabei durch einen Knoten im Graphen repräsentiert und ist über eine Kante mit dem Knoten verbunden, der die Basis des Roboters darstellt. Jeder Knoten enthält Informationen zur Objektklasse sowie die 3D-Koordinaten des Objekts. Da das multimodale Modell nicht nur die Positionen aller Objekte zum gegenwärtigen Zeitpunkt benötigt, sondern auch die Positionen im Verlauf der letzten Sekunde, wird dieser Frame-Graph mit den vorherigen vierzehn Frame-Graphen zu einem Szenen-Graphen kombiniert (siehe Abb. 2). Dieser Szenen-Graph enthält somit die Position jedes Objekts innerhalt der letzten Sekunde, da fünfzehn Frame-Graphen bei einer Bildwiederholfrequenz von fünfzehn Bildern pro Sekunde einen Zeitraum von einer Sekunde abdecken.

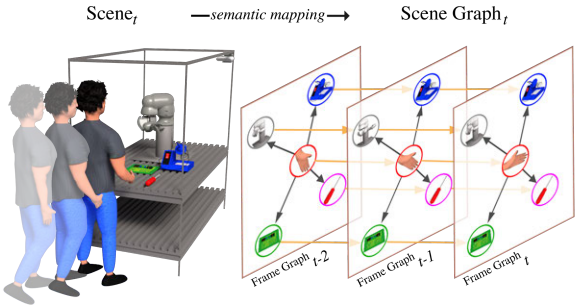


Abb. 2: Der Szenengraph besteht aus mehreren Frame-Graphen [5]

Ausblick

Der Testaufbau bietet zahlreiche Möglichkeiten zur Weiterentwicklung und Anpassung für neue Einsatzszenarien. Eine erweiterte Anwendung könnte darin bestehen, dass der Roboter dem Menschen Werkzeuge anreicht oder Material auf dem Tisch bereitstellt. So könnte der Roboter leere Material-Boxen gegen volle austauschen oder fertige Werkstücke in Boxen einsortieren. In einem solchen Szenario könnte der Roboter dem Menschen proaktiv ausweichen, wenn dieser beispielsweise nach einem Objekt in einer Materialbox greift. Dafür wäre es erforderlich, einen neuen Datensatz zu erstellen, bzw. den CoAx-Datensatz zu erweitern. Das im Rahmen dieser Arbeit trainierte und verwendete SSD-Modell könnte auch für andere Szenarien eingesetzt werden. Solange die gleichen Objekte verwendet werden, müsste das Modell nicht weiter angepasst werden. Wenn jedoch weitere Objekte hinzukommen, können die entwickelten Skripte verwendet werden, um das Modell entsprechend anzupassen und zu trainieren. Darüber hinaus ist der Austausch des Roboters durch Modelle anderer Hersteller möglich, da seine Position nicht von der Kamera erfasst wird, sondern direkt von der Robotersteuerung bezogen wird. Hierfür wäre lediglich der Austausch der Steuerungskomponente für den Roboter erforderlich. Die modulare Struktur der in dieser Arbeit entwickelten Softwarekomponenten ermöglicht somit eine Anpassung an neue Aufgaben und die Verwendung mit anderen multimodalen Modellen.

Literatur und Abbildungen

[1] Intel Corporation. Depth Camera D435. https://www.intelrealsense.com/depth-camera-d435/, 2023.

[2] Microsoft Corporation. COCO - Common Objects in Context. https://cocodataset.org/#home, 2023.

[3] Andrew Howard et al. Searching for MobileNetV3. *ICCV 2019*, 2019.

[4] Dimitrios Lagamtzis, Fabian Schmidt, Jan Seyler, and Thao Dang. CoAx: Collaborative Action Dataset for Human Motion Forecasting in an Industrial Workspace. In *Proceedings of the 14th International Conference on Agents and Artificial Intelligence*, volume 3, pages 98–105. SciTePress, 2022.

[5] Dimitrios Lagamtzis, Fabian Schmidt, Jan Seyler, Thao Dang, and Steffen Schober. Graph Neural Networks for Joint Action Recognition, Prediction and Motion Forecasting for Industrial Human-Robot Collaboration. In *56th International Symposium on Robotics*. ISR europe 2023, 2023.

[6] Wei Liu et al. *SSD: Single Shot MultiBox Detector*. Springer International Publishing, 2016.

Embedded Software Build in Docker and Azure DevOps

Kanujan Kajendrakumar **Rainer Keller**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma cellcentric GmbH & Co. KG, Nabern/Kirchheim unter Teck

Grundlagen zu Docker

Docker ist eine Open-Source-Plattform für die Entwicklung und Ausführung von Applikationen in Container. [2] Container sind isolierte Umgebungen für unterschiedliche Applikationen und sind als "Lightweight Virtualization"bekannt. Der Unterschied zu gewöhnlichen Virtuellen Maschinen ist, dass der Kernel vom Hostsystem genutzt wird und nur die Applikation virtualisiert wird, deshalb sind Container auch als "Lightweight"bekannt (siehe Abb. 1).

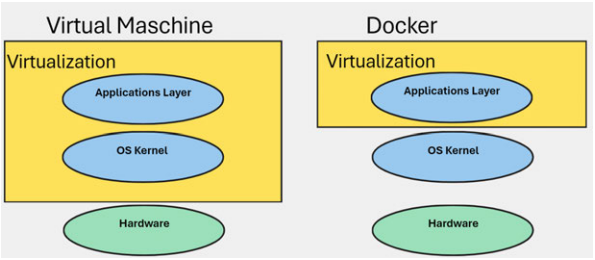


Abb. 1: Unterschied Virtuelle Maschine und Container [1]

Für ein umfassendes Verständnis von Docker ist es essenziell, den Zusammenhang zwischen Dockerfile, Docker Image und Container nachzuvollziehen(siehe Abb. 2). Das Dockerfile ist ein Textdokument, welches die einzelnen Schritte für die Anwendung definiert. Dockerfiles haben ihre eigene Syntax und sind als Blaupause für ein Docker-Image bekannt. [3] Diese Schritte werden dann beim Bauen vom Image sequenziell ausgeführt. Das Image welches dann von diesem Dockerfile erstellt wird besteht dann aus mehreren Schichten. Diese Schichten ergeben sich dann aus den Schritten die im Dockerfile erstellt wurden. Dabei entspricht ein Schritt im Dockerfile einer Schicht im Docker-Image. Docker Images werden üblicherweise in einer Registry abgelegt. Mit diesem Image kann man dann je nach Anwendung einen oder mehrere Container erstellen. Ein Container ist eine isolierte Umgebung für Code und Anwendungen. [4]Der wesentliche Unterschied zwischen Container und dem Image ist, dass der Container die laufende Instanz von diesem Image ist. Dieser Container beinhaltet dann alles was für die Anwendung im Dockerfile spezifiziert wurde. Das könnten irgendwelche Ordner mit Files, benutzerdefinierte Skripte oder ein Image sein.

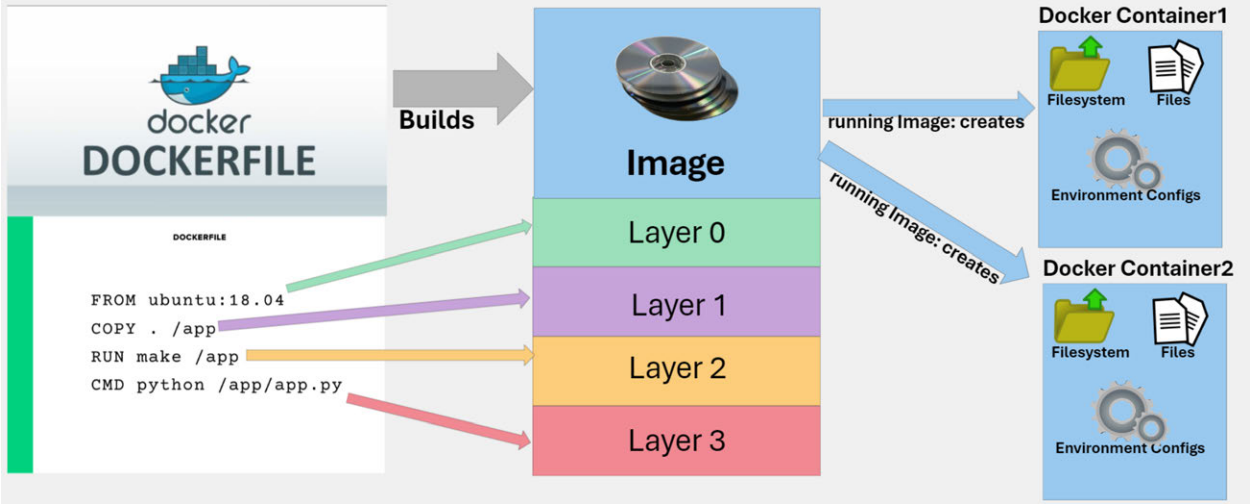


Abb. 2: Zusammenhang Dockerfile, Docker Images und Container [1]

Ziel der Bachelorarbeit

Diese Forschungsarbeit verfolgt das Ziel, den Software-Build-Prozess für Embedded-Systeme durch die Implementierung von Docker-Containern und Azure Pipelines zu optimieren. Im Fokus stehen dabei die Verbesserung der Effizienz während des Entwicklungsprozesses sowie die Gewährleistung von Parallelisierung für die Unterschiedlichen Varianten der Software. Konkret befasst sich die Arbeit mit der überführen der momentanen Build-Umgebung in eine Container-Umgebung mittels Docker und der kontinuierlichen Integration durch die Nutzung von Azure Pipelines(siehe Abb. 3). In dieser Bachelorarbeit geht es um die Geschwindigkeit des Software-Builds und dem Einfluss auf die Kompatibilität für die unterschiedlichen Softwarevarianten.

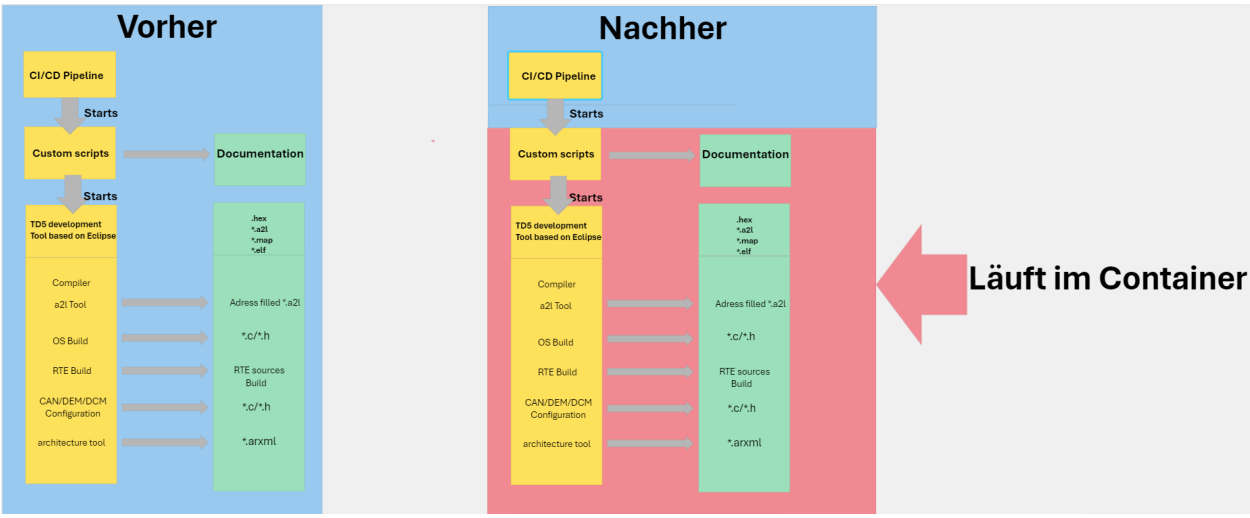


Abb. 3: Ablauf des Buildprozesses (Vorher/Nachher) [1]

Methodologie

Die Vorgehensweise zur Umsetzung dieses Ziels basiert auf einem mehrstufigen Ansatz. Der erste Schritt besteht aus einer umfassenden Literaturrecherche, um Erkenntnisse im Bereich der Containerisierung, CI/CD (Continuous Integration/Continuous Deployment) und CLI (Command Line Interface) zu erlangen und zu erweitern. Dies bildet die Grundlage für die Umsetzung. Die Recherche für die Containerisierung hat nach Einigung mit dem Betreuer auf der Plattform Docker Desktop stattgefunden.
Für den zweiten Schritt liegt der Fokus hier auf die erfolgreiche Ausführung des Builds in der Containerumgebung. Dafür muss eine Entscheidung für ein Docker Image getroffen werden. Docker bietet eine große Anzahl an Images, die in einer Registry abgelegt

werden. Diese Images kann man dann im eigenen Dockerfile als Basis nutzen und für einen spezifischen Anwendungsfall weiterverwenden. Diese Images werden "Base-Image"genannt. Nach mehreren Versuchen und Absprachen mit dem Betreuer wurde erkannt, dass für den Anwendungsfall ein Windows-basiertes Base-Image am besten geeignet wäre, da für den Build benutzerdefinierte Skripte genutzt werden und diese Pfade basierend auf einem Windows-Dateisystem gesetzt sind. Der Aufwand, diese Pfade einzeln auf ein Linux-System umzuändern, erweist sich als zu komplex und entbehrlich. Nach der Wahl des Images wird ein Dockerfile genutzt, das jedoch noch Modifikationen benötigt, da der Windows-Container sich nicht genauso verhält wie der Build auf dem Host-System.
Nachdem der Build in der Containerumgebung ein erfolgreiches Ergebnis aufweist, wird anschließend eine Azure Pipeline erstellt, die dann den gesamten Prozess anregt. Dabei ist entscheidend, wo genau die Ablage des Images erfolgt. In Docker kann man auf eine Registry verweisen. Das bedeutet, wenn man versucht, ein Image auszuführen und es lokal nicht gefunden wird, wird von Docker versucht, es aus dieser Registry herunterzuziehen und anschließend auszuführen. Als Standard ist der Verweis auf Docker Hub gesetzt. Für Unternehmen ist es sinnvoller, diesen Verweis auf eine private Registry zu ändern, da man nicht möchte, dass ein Image mit vertraulichen Informationen öffentlich abgelegt wird.

Ausblick

Die Vermutung besteht, dass die Nutzung von Docker-Containern und Azure Pipelines eine erhebliche Beschleunigung des Software-Builds für Embedded-Systeme verspricht. Die Umsetzung des Softwarebuilds in isolierten Containern ermöglicht zudem die parallele Ausführung für verschiedene Varianten, was zu zusätzlichem Zeitersparnis führen könnte. Trotz dieser vielversprechenden Aussichten sind jedoch aktuelle Herausforderungen zu bewältigen. Ein Problem liegt beispielsweise im unterschiedlichen Verhalten des Containers im Vergleich zum Hostsystem, was den Build-Prozess beeinflusst. Ein weiteres bestehendes Problem ist, dass der Build-Prozess nach dem Ausführen des "TD5-Builds"(siehe Abb. 3) wie eine Blackbox wirkt. Die Lösung dieser Probleme erfordert derzeit eine iterative "Trial and Error"Methode.

Literatur und Abbildungen

[1] Eigene Darstellung.

[2] IBM IBM. Was ist Docker? https://www.ibm.com/de-de/topics/docker, 2023.

[3] Docker Incorporated. The Dockerfile. https://docs.docker.com/build/guide/intro/, 2023.

[4] Docker Incorporated. What is a Container? https://docs.docker.com/guides/walkthroughs/what-is-a-container/, 2023.

Virtual Reality zur Förderung des psychischen Wohlbefindens älterer Menschen: Eine Untersuchung der technischen Merkmale und Anwendungen

Tugce Karaarslan **Dieter Morgenroth**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt an der Fakultät Informatik und Informationstechnik, Esslingen

Einleitung

Die aktuellen Fortschritte in der VR-Technologie bieten nicht nur neue Erlebniswelten, sondern auch revolutionäre Lösungen für das psychische Wohlbefinden älterer Menschen. Die positive Wirkung natürlicher Umgebungen auf die Gesundheit und das Wohlbefinden ist unbestreitbar. Zahlreiche Studien belegen, dass natürliche Umgebungen wie Wälder und Grünflächen einen positiven Einfluss auf die physische und psychische Gesundheit haben. Trotz dieser Erkentnisse haben ältere Menschen in Pflegeheimen, Krankenhäusern oder in Isolation sowie solche mit körperlichen Einschränkungen, oft nur eingeschränkten Zugang zur natürlichen Außenwelt. [3]
Diese Bachelorarbeit untersucht, wie Virtual Reality (VR) als Instrument für das psychische Wohlbefinden älterer Menschen eingesetzt werden kann und dieses verbessern kann. Insbesondere wird die Oculus-Go-Brille von Meta in Verbindung mit der von der Firma Oroi bereitgestellten Wellbeing-Software betrachtet. Diese Software ermöglicht älteren Menschen durch 360 Grad-Videos ein Erlebnis in die Natur und die Außenwelt. Indem innovative Technologie und das Streben nach emotionaler Gesundheit kombiniert werden, befasst sich diese Forschung mit der Verbindung zwischen VR und das Wohlbefinden.

Zielsetzung und Methodik

Das Ziel dieser Bachelorarbeit besteht darin, das Potenzial der virtuellen Realität (VR) zur Verbesserung des emotionalen Wohlbefindens älterer Menschen zu erforschen. Die Arbeit unterstreicht dabei die bedeutende Rolle von VR bei der Förderung des emotionalen Wohlbefindens älterer Menschen. Die Methode ist eine eingehende Analyse der Auswirkungen der Oculus-Go-Brille auf die psychische Gesundheit älterer Personen. Eine umfangreiche Literaturrecherche bildet den Ausgangspunkt, um bestehende Studien und Forschungsarbeiten zum Einsatz von VR bei älteren Menschen zu erfassen. Im Anschluss liegt der Fokus auf der technischen Analyse der Oculus-Go-Brille, die durch die Bewertung von Tests und Expertenbewertungen ergänzt wird. Zusätzlich wurde in Zusammenarbeit mit einer Psychotherapeutin ein Interview durchgeführt und ein Fragebogen erstellt, um ihre professionelle Einschätzung und Erfahrung im Umgang mit Virtual Reality in der psychologischen Praxis zu erfassen. Darüber hinaus wurden Probanden in den Forschungsprozess einbezogen, um möglichst realistische Einblicke in die Nutzung der Oculus-Go-Brille bei älteren Menschen zu erhalten. Auch wird die Untersuchung durch einen Vergleich zwischen der Oculus-Go-Brille und der an der Hochschule Esslingen verwendeten HTC Vive-Brille erweitert, um technische Parameter zu bewerten. Diese Ergebnisse werden anschließend im Zusammenhang mit dem Wohlbefinden interpretiert. Mit dieser Methode ist es möglich, sowohl die technischen Aspekte der virtuellen Realität zu untersuchen als auch die Auswirkungen auf das psychische Wohlbefinden älterer Menschen zu erfassen und zu analysieren.

Virtual Reality

Der Begriff virtuelle Realität ist ein faszinierendes Konzept, das die Idee einer Simulation der Welt vermittelt, die in der Realität nicht existiert. Dabei wird eine interaktive Umgebug erschaffen, die den Benutzern in eine virtuelle Welt versetzt und das Gefühl vermittelt, tatsächlich in dieser Welt zu sein. [4] Der Einsatz von visuellen, auditiven und haptischen Reizen schafft ein realistisches Erlebnis, welches die herkömmlichen Displays übertrifft. Die Benutzerinteraktion in der virtuellen Umgebung ist der entscheidende Faktor, der die VR-Erfahrung von der traditionellen Informationstechnologie unterscheidet. Das faszinierende Konzept erweist sich als ein vielseitiges Werkzeug

zur Unterstützung des psychischen Wohlbefindens von älteren Menschen. Durch die Zusammenarbeit mit der Psychotherapeutin von VirtuallyThere wurden weitere Einsatzmöglichkeiten besprochen, darunter Anwendungsbereiche wie die Expositionstherapie in der Angstambulanz sowie innovative Ansätze wie der Einsatz von Virtual Reality bei der Behandlung von Essstörungen.

Oculus-Go-Brille: Testüberblick und vorläufige Resultate

Das Brillenmodell Oculus-Go, eines der VR-Technologieprodukte, das von Oroi verwendet wird, bietet ein immersives Erlebnis durch ein Head-Mounted Display (HMD). Durch die Nutzung von Bildschirmen und Linsen entsteht ein stereoskopisches 3D-Bild, das die Illusion von Tiefe und räumlicher Wahrnehmung erzeugt, während der Träger von seiner Umgebung abgeschirmt wird. Mit drei Freiheitsgraden (3DoF) ermöglicht die Brille Drehungen, Neigungen und Rotationen ohne zusätzliche Steuerung. Dank seiner Eigenständigkeit ist die gesamte Technik im Headset integriert und ermöglicht eine einfache, ortsunabhängige Nutzung ohne externe Hardware.

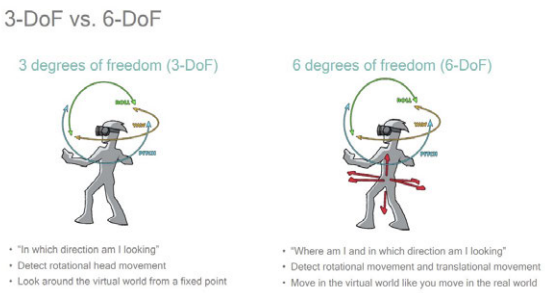


Abb. 1: Unterschied 3-Dof und 6-Dof [2]

Auch wenn die Oculus Go-Brille eine eingeschränkte Bewegungsfreiheit (3DoF) aufweist, bietet sie dennoch ein immersives Erlebnis, besonders für Anwendungen mit weniger komplexen räumlichen Interaktionen. Abbildung 1 veranschaulicht den Unterschied zwischen 3DoF und 6DoF in der virtuellen Realität. Während 3DoF-Bewegungen auf Drehen, Neigen und Schwenken beschränkt sind, ermöglichen 6DoF-Brillen zusätzliche Bewegungsrichtungen wie Hoch, Runter, Vor, Zurück, Rechts und Links, was eine komplexere räumliche Interaktion ermöglicht. Dennoch bietet die Oculus-Go-Brille eine vielseitige Möglichkeit für Benutzer, die ein einfaches und eindrucksvolles VR-Erlebnis suchen.

Abb. 2: Testperson [1]

Die Versuche mit der Oculus-Go-Brille beinhalten die Bewertung der Benutzererfahrung, insbesondere in natürlicher Umgebung. In Abbildung 2 ist eine 82-jährige Testperson dargestellt, die vor der Nutzung der virtuellen Realität zu ihren Erwartungen und Gefühlen befragt wurde. Während der Nutzung wurden die Reaktionen analysiert und beobachtet. Anschließend wurde die Person nach der Nutzung zu ihrem Wohlbefinden, ihrem Umgang sowie den technischen Aspekten der Brille befragt. Nach aktuellen Erkenntnissen wurde die Bildqualität bis auf eine leichte Unschärfe bei den Himmelsbildern positiv bewertet. Die Testperson empfand die VR-Brille als etwas schwer, ohne jedoch körperliche Beschwerden zu haben. Besonders gut konnten die Reaktionen in natürlichen Umgebungen analysiert werden. Insgesamt hat die Realitätsnähe des VR-Erlebnisses die Testperson positiv überrascht und kann sich vorstellen, die VR-Technologie regelmäßig zu nutzen. Die Verbesserungsvorschläge beinhalten eine leichtere Passform und eine bessere Bedienbarkeit.

Schlussfolgerung und Ausblick

Die aktuellen Testergebnisse deuten darauf hin, dass die Oculus-Go-Brille bei älteren Nutzern positive Reaktionen hervorruft und ein immersives Erlebnis bietet. Für zukünftige Entwicklungen könnte der Fokus auf einer leichteren Passform und bessere Benutzerfreundlichkeit liegen, um die Nutzung für verschiedene Altersgruppen zu optimieren.

Literatur und Abbildungen

[1] Eigene Darstellung.

[2] Inc. Qualcomm Technologies. On-device motion tracking for immersive mobile VR. https://de.slideshare.net/qualcommwirelessevolution/ondevice-motion-tracking-for-immersive-vr?from_action=save, 08 2017.

[3] Hamburg Universitaet. Auch virtuelle Natur hat einen positiven Effekt. https://www.uni-hamburg.de/newsroom/presse/2021/pm7.html, 02 2021.

[4] Matthias Wölfel. *Immersive Virtuelle Realität: Grundlagen, Technologien, Anwendungen.* Springer-Verlag GmbH, DE, ein Teil von Springer Nature 2023, 2023.

Regionale Analyse von Patentdaten in Baden-Württemberg sowie Konzeption übersichtlicher Datenexportformate („Patentatlas“)

Adrian Kiani **Catharina Kriegbaum-Kling**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma Patent- und Markenzentrum Baden-Württemberg, Stuttgart

Einleitung

Die Anzahl der Patentanmeldungen gilt als Indikator für die Innovationskraft [5] eines Unternehmens oder eines Landes und so ist die Anzahl von Patenten, die pro Region angemeldet wird, von hohem Interesse. Gerade für die Standortbewertung spielen bestehende Statistiken über die Höhe der Patentanmeldungen für bestimmte IHK-Bezirke eine entscheidende Rolle. Das DPMA veröffentlicht in seinen Jahresberichten regelmäßig die Kennzahlen von Deutschland und in den Bundesländern. So wurden im Jahr 2022 57.214 Patente in ganz Deutschland, davon 10.329 für Transport (Maschinenbau), in Baden-Württemberg wurden insgesamt in 2022 13.344 Patente angemeldet [4] . An der Anzahl der in bestimmten Technologiegebieten angemeldeten Patente lässt sich ablesen, auf welchen Technologien die verschiedenen Regionen besondere Stärken haben. Detailliertere Analysen sind für die IHK-Bezirke von Interesse, um neue Unternehmen zu überzeugen bei ihnen ihren nächsten Standort aufzubauen. Somit hat die Aussagekraft solch einer Statistik und ein möglichst hoher Informationsgehalt derer auch für den ansässigen Arbeitsmarkt eine entscheidende Relevanz. Analysen auf der Ebene der IHK-Bezirke werden von den Ämtern nicht erstellt. Obwohl umfangreiche Daten zur Verfügung stehen, ist die Erstellung eines übersichtlichen Berichts bisher nicht digitalisiert und lässt sich somit nur mit hohem Arbeitsaufwand erstellen.

Ziele der Arbeit

Das Ziel der Arbeit ist die Erstellung eines vorgefertigten Templates für einen jährlichen Bericht. Die Patentdaten aus der Datenbank sollen beispielsweise in Microsoft Excel in Diagramme und andere Datenformate überführt werden und diese dann z.B. in Microsoft Word in einen vorgefertigten Bericht eingetragen werden. Hierbei soll das Template in Word so weit wie möglich automatisiert werden, sodass nur noch Textstellen angepasst werden müssen und Diagramme und ähnliche Darstellungen schon in Excel fertiggestellt werden.

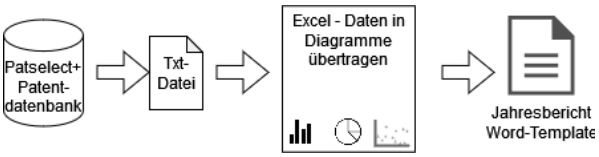


Abb. 1: Datenfluss der Patentdaten [1]

Vorgehensweise

Um den Arbeitsaufwand vorerst klarzustellen, werden zu Beginn der Arbeit die Datengrundlage und der Prozess der Patenterstellung analysiert. Vorab ist es wichtig festzustellen, welche Daten genutzt und anschließend auch wie diese präsentiert werden sollen. Aufgrund der vielen Möglichkeiten sind im Laufe der Arbeit Entscheidungen zur Abgrenzung und/oder zur sinnvollen Repräsentation der Patentdaten zu treffen. Zuerst wird ein grobes Format ohne Programmcode in Word erstellt, um einen Rahmen festzulegen, welche Daten und Diagramme dargestellt werden. Eine wichtige Fragestellung war die, wie und ob man den Technologieindex, welcher angibt

INID	Kriterium	Feld	Inhalt
54	Titel	TI	[DE] VERFAHREN UND VORRICHTUNG ZUM ÜBERWACHEN EINES ZUSTANDES WENIGSTENS EINER VORBESTIMMTEN BATTERIEZELLE EINER BATTERIE [EN] METHOD AND DEVICE FOR MONITORING A STATE OF AT LEAST ONE PREDETERMINED BATTERY CELL OF A BATTERY [FR] PROCEDE ET DISPOSITIF DE SURVEILLANCE D'UN ETAT D'AU MOINS UN ACCUMULATEUR PREDEFINI D'UNE BATTERIE
71/73	Anmelder/Inhaber	PA	BOSCH GMBH ROBERT, DE ; GS YUASA INT LTD, JP
72	Erfinder	IN	FRIEDRICH MARCO, DE ; SPRINGER BERNHARD, DE
22/96	Anmeldedatum	AD	14.07.2015
21	Anmeldenummer	AN	15176640
	Anmeldeland	AC	EP
	Veröffentlichungsdatum	PUB	15.11.2023
33 31 32	Priorität	PRC PRN PRD	
51	IPC-Hauptklasse	ICM	G01R 31/36 (2020.01)
51	IPC-Nebenklasse	ICS	H01M 10/42 H02J 7/00

Abb. 2: Notwendige Patentdaten für die Erstellung des Jahresberichtes [2]

für welches Themengebiet ein Patent angemeldet wurde, in die Datengrundlage mit aufnehmen kann. Mithilfe der Makros wird anschließend getestet, inwieweit ein Template in Word mit den Excel Diagrammen kompatibel ist und um auch mehr über die Möglichkeiten bei der Arbeit mit Makros herauszufinden. Hierbei werden Textmarken in Word genutzt, um damit genau zu definieren, wo im Template welche Daten platziert werden sollen. Im Makro werden diese dann lediglich referenziert, somit lässt sich das Template genau strukturieren und sorgt für keine Verschiebungen bei einer erneuten Erstellung. Wichtig ist im Nachhinein eine saubere und leicht verständliche Darstellung des Jahresberichts, da dieser für ein schnelles Verständnis der momentanen Lage für die Industrie bieten und auch von den 12 IHK-Bezirke Baden-Württembergs für weitere Zwecke verwendet werden können soll. Somit spielen Design, Informationsgehalt und Aussagefähigkeit eine entscheidende Rolle bei der Erstellung. Das Ergebnis orientiert sich hierbei am Jahresbericht des DMPA für 2022 [3].

Literatur und Abbildungen

[1] Eigene Darstellung.

[2] Datenbank DPMA. Patentdatenbank. https://depatisnet.dpma.de/DepatisNet/depatisnet?action=bibdat&docid=EP000003118639B1, 2023.

[3] Deutschland DPMA. Jahresbericht 2022. https://www.dpma.de/dpma/veroeffentlichungen/jahresberichte/index.html, 2022.

[4] Deutschland DPMA. Aktuelle Statistiken: Patente. https://www.dpma.de/dpma/veroeffentlichungen/statistiken/patente/index.html, 07 2023.

[5] Maike Haag, Hanno Kempermann, Enno Kohlisch, and Oliver Koppel. Innovationsatlas 2023: Die Innovationskraft der deutschen Regionen. https://www.iwkoeln.de/studien/maike-haag-hanno-kempermann-enno-kohlisch-oliver-koppel-die-innovationskraft-der-deutschen-regionen.html, 2023.

Migration from Monolith to Micro-Frontends: An Investigation into the Best Practices based on a Real Case Scenario

Christian Kloos **Mirko Sonntag**

Department of Computer Science and Engineering, Esslingen University

Work carried out at pep.digital GmbH, Esslingen

Introduction

Given the notable success of microservices in recent years, the desire for similar modularisation benefits extended swiftly to the frontend domain. Three fundamental principles underlying micro-frontends include being technology-agnostic and empowering teams to select and upgrade their technology stack. Isolating team code to construct self-contained app modules that do not depend on shared state or global variables. Along with constructing resilient sites that do not collapse when a feature has failed [1].

Problem Statement

As software projects grow in complexity and size, the frontend layer requires greater agility, maintenance, and scalability solutions. Building a monolithic application may seem straightforward initially, but problems can arise over time. The monolithic architecture hinders quick and easy changes, leading to major corporations slowing down their development processes. Working with medium to large teams may create challenges due to communication overheads and centralized decision-making. Moreover, rules are usually established once and upheld for extended periods, since implementing changes would demand significant effort [2]. Moving from a monolithic to a micro-frontend architecture poses significant challenges. However, this transition offers benefits to organizations that are expanding and facing evolving technology and user needs. Before migration, it is essential to have a comprehensive understanding of micro-frontends. Organizations must overcome the challenges of deciding how to split the current code base, ensure a consistent user experience and achieve a seamless transition without disrupting day-to-day business. In addition, often these decisions need to be made upfront. Since they will influence future decisions, such as how a micro-frontend is defined, how the different views are orchestrated, how the final view for the user is assembled, and how the micro-frontends will communicate and share data [2].

Scope of the Thesis

The scope of this thesis is to objectively analyze micro-frontends, their architecture and implementation possibilities. Additionally, the thesis seeks to investigate the challenges and obstacles organizations may face when transitioning from monolithic to micro-frontend architectures. Different migration strategies and approaches will be investigated by analyzing various methods and patterns to split the monolith and effectively integrate micro-frontends. This paper presents research-based findings and recommendations to guide organizations going through a similar transition process.

The Migration Phases

There are three main migration phases: reverse engineering, architecture transformation, and forward engineering [4]. During the reverse engineering phase, information is collected to determine the common resources, dependencies, and functions of the current system. The main information sources for understanding the existing application are source code, textual and architectural documents, tests, meetings, and data models [4]. At this stage, it is crucial to understand how the users use the old application. In the architecture transformation phase, the Domain Driven Design principles are applied and the business domains and bounded contexts of the new application are identified. Modifying the existing application to establish clear boundaries is advisable as it would aid in the extraction process and could help resolve issues before dividing the codebase [3]. In the forward engineering phase, the coupling is broken,

the code base is split, and the technical solution for implementing the micro-frontend architecture is selected and implemented [6]. Figure 1 illustrates an adapted version of the three migration phases. In this work, the migration model was extended by two additional stages. The initial stage in the migration process emphasizes goal definition, aligning with best practices that highlight the importance of clearly articulated goals and needs to ensure all teams and components are working cohesively towards a common objective. Additionally, the monitoring and maintenance stage is another crucial stage. The micro-frontend architecture, while offering benefits, introduces increased complexity. With each update, there is a potential risk of introducing errors, necessitating detailed monitoring. Once the new system is established, the scaling process becomes more straightforward, allowing for scalability adjustments by consistently adding or removing micro-frontend components as needed.

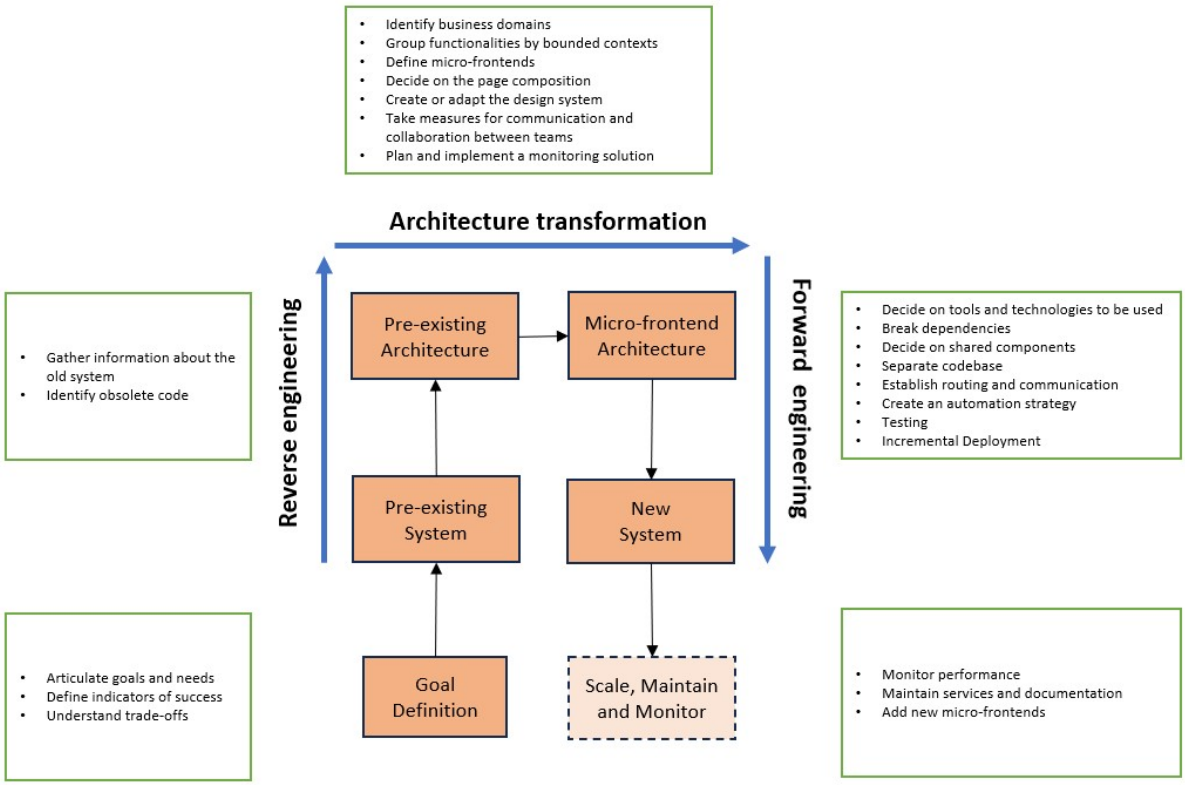


Fig. 1: Migration model from a monolithic frontend to micro-frontends [5]

Outlook

In the further course of this scientific work, the migration model will be tested based on a real migration case study. The web application to be migrated is a functional self-service portal for employees, owned and used by the company pep.digital GmbH. The process from goal definition to deployment will be illustrated, along with the challenges encountered, the decision-making process, and the migration outcomes.

References and figures

[1] Michael Geers. *Micro frontends in action*. Manning, 2020.

[2] Luca Mezzalira. *Building Micro-Frontends*. O'Reilly Media Incorporated, 2021.

[3] Sam Newman. *Monolith to Microservices*. O'Reilly Media Incorporated, 2021.

[4] Di Francesco Paolo, Patricia Lago, and Ivano Malavolta. Migrating Towards Microservice Architectures: An Industrial Survey. In *2018 IEEE 15th International Conference on Software Architecture*. IEEE, 2019.

[5] Own representation.

[6] Olexandr Varenyk, Nataliia Khatsko, and Kyrylo Khatsko. Method of Converting the Monolithic Architecture of a Front-End Application to Microfrontends. In *Challenges and reality of the IT-space*, pages 106–116. Institute of Bioorganic Chemistry. Polish Academy of Sciences. Scientific Publishers OWN, 2023.

Konzeption, Analyse und Evaluation einer Integration von erweiterten bzw. qualifizierten elektronischen Unterschriften in eine Prozessmanagementsoftware

Andreas Kolb **Dominik Schoop**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma Roxtra GmbH, Göppingen

Motivation

Die Art und Weise, wie Geschäfte getätigt und Informationen ausgetauscht werden, ändert sich in der heutigen, schnelllebigen und immer stärker digitalisierten Welt zunehmend. Die elektronische Unterschrift spielt hierbei eine entscheidende Rolle und wird zu einem wesentlichen Bestandteil des laufenden Transformationsprozesses.
Die kontinuierliche Optimierung und Digitalisierung von Abläufen ist ein zentraler Aspekt moderner Geschäftsprozesse. Dadurch können die Effizienz gesteigert, Kosten reduziert und der Verbrauch von Papier minimiert werden. Der Einsatz von elektronischen Unterschriften hat dabei eine wichtige Bedeutung. Geschäftsprozesse können mithilfe von elektronischen Unterschriften vollständig digitalisiert und somit beschleunigt werden. Ein sicherer und rechtlich anerkannter Weg zur Unterzeichnung von Dokumenten ermöglicht es, Geschäftsprozesse durch den Einsatz einer Prozessmanagementsoftware effizienter zu gestalten und umfassend zu dokumentieren. Vor diesem Hintergrund widmet sich die Arbeit der Konzeption, Analyse und Evaluation einer Integration von erweiterten bzw. qualifizierten elektronischen Unterschriften in eine Prozessmanagementsoftware. [5]

Ziel der Arbeit

Das Ziel der Arbeit besteht darin, ein umfassendes Konzept zu entwickeln, das die rechtssichere Integration erweiterter und qualifizierter elektronischer Unterschriften in die Prozessmanagementsoftware roXtra ermöglicht. Dieses Konzept soll offen gestaltet sein, um verschiedene Verfahren zur elektronischen Unterschrift zu unterstützen und individuellen Kundenanforderungen gerecht zu werden.
Für die Integration sollen zwei Lösungswege genauer untersucht werden. Zum einen wird ein Konzept für eine Anbindung von externen Signaturanbieter erstellt. Dies ermöglicht erweiterte und qualifizierte Unterschriften auf Dokumenten zu erstellen. Zudem wird ein Konzept entwickelt, wie ein Unterschriftenpad in die Prozessmanagementsoftware integriert werden kann, um die Durchführung von fortgeschrittenen elektronischen Signaturen zu ermöglichen. Außerdem soll untersucht werden, welche Maßnahmen erforderlich sind, um ein qualifizierter Vertrauensdiensteanbieter zu werden und somit die Befugnis zur Ausstellung qualifizierter elektronischer Unterschriften zu erlangen.

Drei Arten von elektronischen Unterschriften

Die Welt der elektronischen Unterschriften lässt sich in drei Kategorien unterteilen: einfache, erweiterte und qualifizierte elektronische Unterschriften. Die Grundlagen für die verschiedenen Formen der elektronischen Unterschrift sind in der Verordnung (EU) Nr. 910/2014 des Europäischen Parlaments und des Rates über elektronische Identifizierung und Vertrauensdienste für elektronische Transaktionen (eIDAS) festgelegt.
Für eine einfache elektronische Unterschrift gibt es nur ein Kriterium. Gemäß der eIDAS-Verordnung handelt es sich um eine einfache elektronische Unterschrift, wenn die zum Unterzeichnen verwendeten elektronischen Daten mit den zu unterschreibenden Daten verknüpft sind. Eine einfache elektronische Unterschrift kann beispielsweise durch das Einfügen einer Unterschriftsgrafik oder eines Kürzels auf einem Dokument erreicht werden.
Im Kontrast dazu erfüllt die erweiterte elektronische Unterschrift, auch fortgeschrittene elektronische Unterschrift genannt, höhere Anforderungen. Sie erfordert entsprechend den Bestimmungen der eIDAS-Verordnung eine eindeutige Identifikation des Unterzeichners, Authentifizierung und Schutz vor Datenmanipulation. Dadurch bietet sie eine erhöhte

Beweissicherheit, was besonders für Dokumente mit höherem Haftungsrisiko von Vorteil ist. [1]
Die qualifizierte elektronische Unterschrift baut auf der fortgeschrittenen Form auf. Somit gelten alle Anforderungen der fortgeschrittenen elektronischen Unterschrift auch für die qualifizierte elektronische Unterschrift. Darüber hinaus ist gemäß den Bestimmungen der eIDAS-Verordnung der Einsatz zertifizierter Signaturerstellungseinheiten und qualifizierter Zertifikate erforderlich. Die Anforderungen an ein qualifiziertes Zertifikat sind im Anhang 1 der eIDAS-Verordnung beschrieben. Diese beinhalten unter anderem die Notwendigkeit, dass die Ausstellung durch einen qualifizierten Vertrauensdiensteanbieter erfolgt. Die technische Realisierung von fortgeschrittenen und qualifizierten elektronischen Unterschriften erfolgt durch das Prinzip der digitalen Signatur. [3]
Die Beweiskraft einer elektronischen Unterschrift steigt, beginnend von einer einfachen über eine fortgeschrittene bis hin zu einer qualifizierten. Rechtlich sind sowohl einfache als auch fortgeschrittene elektronische Unterschriften gültig, sofern nicht gesetzlich die Schriftform vorgeschrieben ist. Falls die Schriftform gesetzlich verlangt wird, kann gemäß § 126 in Verbindung mit § 126a BGB die handschriftliche Unterschrift ausschließlich durch die qualifizierte elektronische Unterschrift ersetzt werden. In Fällen, in denen die Nutzung einer elektronischen Unterschrift ausdrücklich untersagt ist, dürfen keine der genannten Formen verwendet werden.
Bei steigendem Haftungsrisiko empfiehlt es sich, die fortgeschrittene elektronische Unterschrift der einfachen vorzuziehen, da sie aufgrund ihrer Integrität und Authentizität eine höhere Beweiskraft aufweist. Bei sehr hohem Risiko sowie der zwingend erforderlichen Schriftform wird die qualifizierte elektronische Unterschrift eingesetzt. [1]

Umsetzung der Konzepte

Damit die Konzepte später auch für die Software roXtra genutzt werden können, liegt ein besonderes Augenmerk darauf, die Konzepte kompatibel mit den in der Software verwendetet Technologien zu gestalten. Zur Beurteilung der Machbarkeit der Konzepte dienen Proof of Concepts, um auf dieser Grundlage Empfehlungen für die Umsetzung abzuleiten.

Die Software roXtra ist eine webbasierte Anwendung. Deshalb ist es notwendig, eine Lösung zu finden, mit dem Unterschriftenpad aus dem Webbrowser zu interagieren. Eine Lösung hierfür besteht in der Nutzung der signotec signoPAD-API/Web. Diese API wird auf dem Computer ausgeführt, an dem das Unterschriftenpad angeschlossen ist. Die Anwendung im Webbrowser kann dann über die signoPAD-API/Web mittels WebSocket-Kommunikation mit dem Unterschriftenpad interagieren, was wiederum die Übermittlung von Unterschriftsdaten an die Webanwendung ermöglicht. Die signoPAD-API/Web fungiert somit als Verbindungsglied zwischen der Webanwendung im Browser und dem lokalen Unterschriftenpad. Eine schematische Darstellung des Ablaufs einer Unterschrift über eine Webanwendung ist in Abbildung 1 zu sehen. [4]

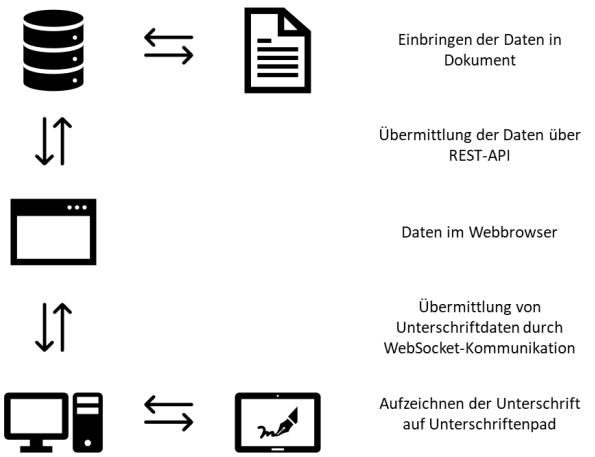


Abb. 1: Ablauf Unterschrifterfassung mit Unterschriftenpad und Webanwendung [2]

Für die Anbindung externer Signaturanbieter kann die Verwendung der APIs dieser Anbieter genutzt werden. Hierbei muss die unterzeichnende Person zunächst einen Account beim jeweiligen Anbieter erstellen und der Anwendung die Berechtigung erteilen, Anfragen an die API zu senden. Der Authentifizierungsprozess ist in Abbildung 2 dargestellt. Über die API können dann die zu unterzeichnenden Dokumente an den externen Signaturanbieter gesendet, dort unterschrieben und wieder zurück in die Anwendung gespeichert werden.

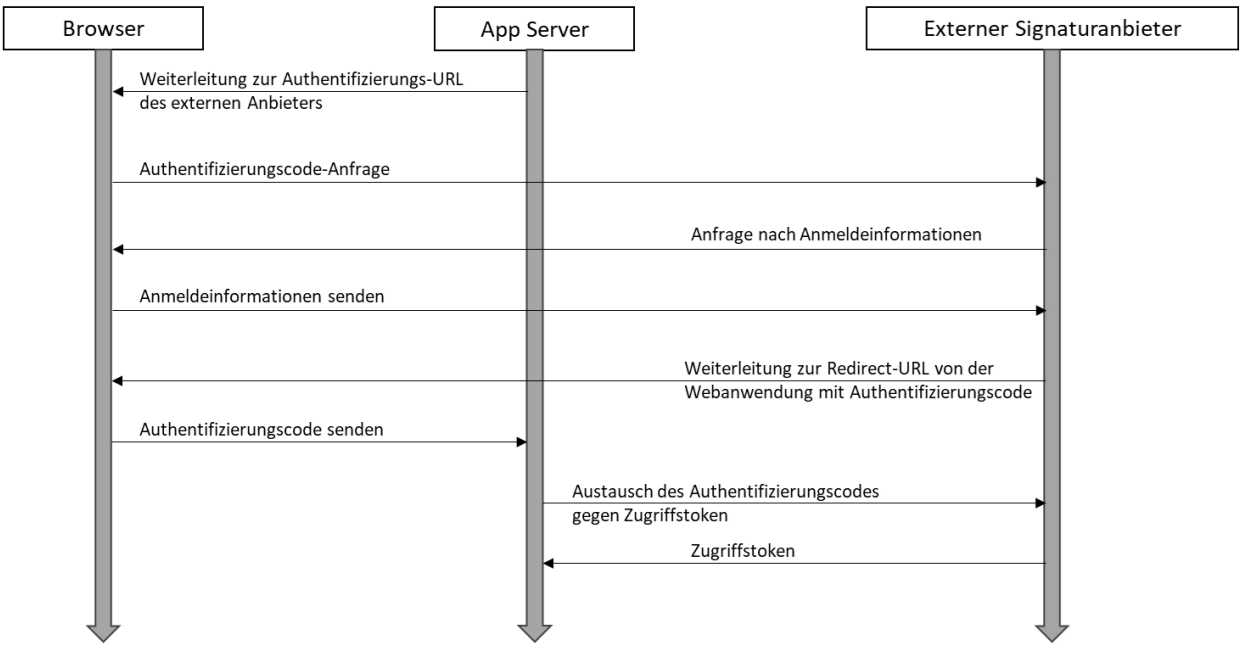


Abb. 2: Authentifizierungsprozess mit einem externen Signaturanbieter [2]

Ausblick

Nach erfolgreicher Überprüfung der Konzepte mittels Proof of Concepts besteht die Option, eine der vorgeschlagenen Lösungen in die Software roXtra zu integrieren. Die gewonnenen Erkenntnisse aus den Proof of Concepts dienen als Grundlage für diese Entscheidung und ermöglichen eine fundierte Bewertung der Machbarkeit und Effektivität der vorgeschlagenen Ansätze.

Bei einer positiven Bewertung folgt die nächste Phase der Umsetzung. Hierbei könnte die ausgewählte Lösung mithilfe der erarbeiteten Konzepte und unter Einbindung eines Entwicklungsteams in die Software integriert werden. Dieser Schritt eröffnet die Möglichkeit, die Funktionalitäten für die Anbindung des Unterschriftenpads sowie für die Integration externer Signaturanbieter erfolgreich zu implementieren.

Literatur und Abbildungen

[1] Simon Apel and Christopher Huber. Die elektronische Signatur – Eine Einführung. *JURA - Juristische Ausbildung*, 44:1141–1153, 2022.

[2] Eigene Darstellung.

[3] Wolfgang Ertel and Ekkehard Löhmann. *Angewandte Kryptographie*. Hanser, 6 edition, 2020.

[4] signotec GmbH. signotec signoPAD-API/Web. https://www.signotec.com/software/entwicklung-api-/signopad-api-web/, 2023.

[5] Volker Gruhn, Vincent Wolff-Marting, André Köhler, Christian Haase, and Torsten Kresse. *Elektronische Signaturen in modernen Geschäftsprozessen*. Vieweg, 1 edition, 2007.

Entwicklung einer grafischen Benutzeroberfläche für nicht relationale Datenbanken in eingebetteten Systemen

Shkumbin Krasnic **Rainer Keller**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma cellcentric GmbH & Co. KG, Kirchheim Teck/Nabern

Einleitung

AUTOSAR, kurz für **AUT**omotive **O**pen **S**ystem **AR**chitecture, ist eine offene und standardisierte Softwarearchitektur, die gemeinsam von Automobilherstellern, Zulieferern und Tool-Entwicklern entwickelt wurde. Sie wurde von einem Konsortium von Unternehmen der Automobilindustrie, Zulieferern und weiteren Unternehmen aus der Elektronik-, Halbleiter- und Softwareindustrie, Die Verwendung von AUTOSAR bietet einen grundlegenden Rahmen für standardisierte Prozesse in der Entwicklung von eingebetteten Systemen. Durch diese Standardisierung wird eine Skalierbarkeit der Software auf unterschiedliche Fahrzeug- und Plattformvarianten ermöglicht, während gleichzeitig Verfügbarkeits- und Sicherheitsanforderungen berücksichtigt werden. Ein zentraler Bestandteil dieses Rahmens sind ARXML-Dateien (AUTOSAR XML), welche zur Definition der verschiedenen Softwarekomponenten verwendet werden, die ein AUTOSAR-System ausmachen [6] [1]

ARXML

ARXML-Dateien sind Konfigurationsdateien, die in der AUTOSAR-Architektur verwendet werden. Sie enthalten Konfigurations- und Spezifikationsinformationen im XML-Format für ein Steuergerät, das zur Steuerung von Komponenten eines Motors verwendet wird. Diese Dateien können in sogenannte AUTOSAR-Softwarekomponentenmodelle auch genannt als (SWC) abgebildet werden. Ein Ansatz zur Nutzung von ARXML-Dateien, der auch in dieser Bachelorarbeit aufgeführt wird besteht darin, sie als NoSQL-Datenbankdateien zu verwenden. Dies könnte als folge zu einer Verbesserung der Skalierbarkeit der AUTOSAR-Projekte führen. [4]

Zielsetzung und Schwerpunkte

Ziel dieser Bachelorarbeit ist es ausgewählte Teilfunktionen von DaVinci Developer Classic, durch eine eigens erstellte GUI zu ersetzen, um den Arbeitsprozess zu vereinfachen und effizienter zu gestalten. Ein Schwerpunkt liegt dabei auf Wahrung der Funktionalität und einer nahtlosen Integration der Benutzeroberfläche in bestehende Entwicklungsprozesse. Außerdem soll die Einhaltung der AUTOSAR Richtlinien vereinfacht und gewährleistet werden. Ein weiterer Schwerpunkt liegt dabei auch auf die Entwicklung einer benutzerfreundlichen grafischen Benutzeroberfläche (GUI), welche die Teilfunktionen des Davinci Developers ersetzt. Diese GUI fungiert als Schnittstelle, über die Benutzer mit den implementierten Anwendungslogiken und Anwendungsfunktionen interagieren können. Zusammenfassend soll die Abschlussarbeit, die Entwicklung und Wartung von AUTOSAR-Projekten vereinfachen.

Implementierung und Methodik

Unter Absprache mit dem Betreuer und reichlicher Recherche stellte sich heraus, dass unter Verwendung von Python die Open-Source Bibliothek „Autosar“ eine gute Möglichkeit bietet die Funktionsweise des Davinci Developers nachzubilden. Hierbei liegt der Fokus auf die Integration der Funktionen zur Erstellung, Änderung und dem Löschen von Ports sowie der flexiblen Anpassung von Parametern innerhalb der SWCs (Softwarekomponenten). Als Ansatz sollen drei ARXML-Dateien(Portinterfaces, Constants und Datatypes) als eine NO-SQL Datenbasis dienen, diese werden mit der ausgewählten SWC-ARXML Datei vereint, um anschließend eine umfassende und präzise Konfigurationsgrundlage zu schaffen. Diese neu geschaffene Datenbasis ermöglicht eine gezielte und effiziente Anpassung der Softwarekomponente (SWC). Unter diesem Aspekt wurde zur Entwicklung der Grafischen Benutzeroberfläche das GUI-Toolkit Tkinter benutzt

Davinci Developer

Der Davinci Developer ist ein Tool entwickelt von der Vector Informatik GmbH, das speziell für die Konfiguration und Umsetzung von AUTOSAR-konformen Anwendungen und Softwarekomponenten in der Automobilindustrie konzipiert ist. Dieses Tool bietet eine Vielzahl von Funktionen, die es Entwicklern ermöglicht, Entwicklungsprozesse zu optimieren und die Effizienz bei der Entwicklung von eingebetteten Systemen in Fahrzeugen zu optimieren. Das Tool bietet eine Benutzeroberfläche zur Konfiguration von Softwarekomponenten. Dies beinhaltet die Definition von Schnittstellen, Parameter und Konfigurationen. Durch Verwendung der Konfigurationen erstellt der Davinci Developer automatisch AUTOSAR-Artefakte, Diese Artefakte umfassen ARXML-Dateien und relevante Softwarekomponenten. Die Software bietet Funktionen, für die Modellierung und Simulation von Softwarearchitekturen. Mithilfe dieser können Entwickler das Verhalten ihrer Systeme, vor der Implementierung visualisieren und testen. Durch die Integration mit verschiedenen Entwicklungsumgebungen und Werkzeugen erleichtert der Davinci Developer die Zusammenarbeit und den Datenaustausch zwischen verschiedenen Entwicklungsphasen und Teams. [9] [8] [7]

AUTOSAR Architektur

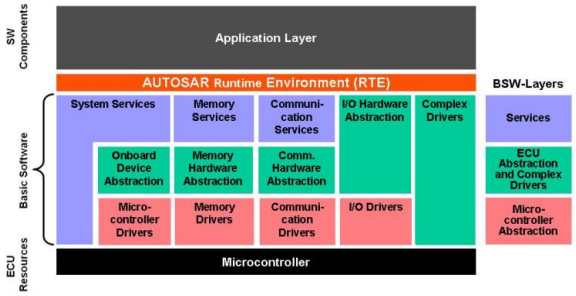


Abb. 1: Autosar Architektur [5]

Die AUTOSAR-Architektur besteht wie aus Abbildung 1 zu entnehmen aus verschiedenen Schichten und Modulen, die eine klare Trennung der Funktionen ermöglichen:

1. Anwendungsschicht (Application Layer): Hier werden die eigentlichen Anwendungen und Funktionen für das Fahrzeug entwickelt, wie z. B. Motorsteuerung, Fahrerassistenzsysteme usw.

2. Laufzeitsystem-Layer (Runtime Environment Schicht): Dieser Schicht beinhaltet das Laufzeitsystem, das für die Ausführung der Anwendungen benötigt wird, einschließlich der Kommunikation und Speicherverwaltung. Sie fungiert als Middleware zwischen Anwendungsschicht und den weiteren unteren Schichten. Im Grunde genommen sorgt die RTE- Schicht für die Kommunikation zwischen den Softwarekomponenten untereinander als auch für die Kommunikation zwischen der Basis-Software-Schicht und der Anwendungsschicht.

3. Basic Software Schicht: Hier befinden sich die grundlegenden Softwarekomponenten wie Treiber, Kommunikationsprotokolle und Diagnosesoftware, die für die Kommunikation zwischen Anwendungen und der Hardware benötigt werden.

4. Hardware Schicht: Dies ist die physische Hardware des Fahrzeugs, die die eingebetteten Systeme ausführt , z. B. Steuergeräte und Sensoren. [3]

Ausblick

Zum aktuellen Stand der Bachelorarbeit wurden die Funktionen wie Port erstellen/ändern/löschen implementiert und in die auch schon erstellte GUI, die mithilfe von Tkinter erstellt wurde integriert. Es erfolgen momentan noch letzte Anpassungen unter Absprache mit dem Betreuer und den Mitarbeitern, an den Funktionen und der GUI.
In Abbildung 2 und 3 sieht man den derzeitigen Stand der GUI nochmal, dabei bildet Abbildung 2 die Startseite ab in der die SWC eingelesen und mit den drei DB-Dateien vereint wird.

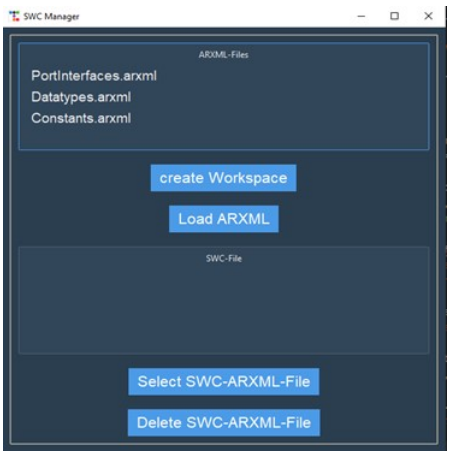


Abb. 2: Startseite [2]

In Abbildung 3 ist die Ansicht zu sehen in der die Modifikation der SWC stattfindet. (erstellen/ändern/löschen von Ports)

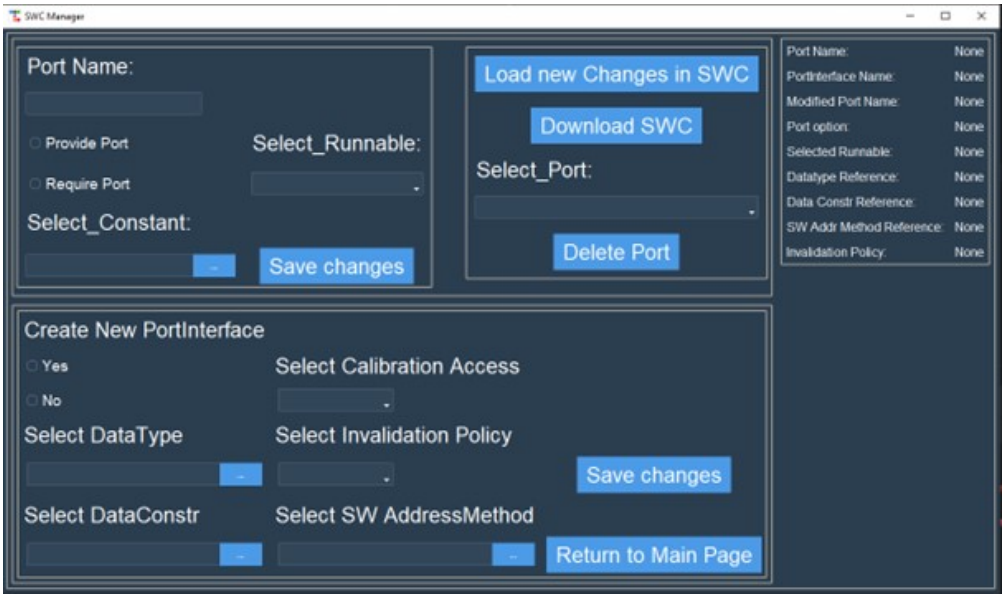


Abb. 3: Modifikationsseite [2]

Literatur und Abbildungen

[1] Tutorials Autosar. What Is AUTOSAR? https://autosartutorials.com/what-is-autosar/, 2019.

[2] Eigene Darstellung.

[3] Cariad Group Company Embitel. Decoding the "Component Concept" of the Application Layer in AUTOSAR. https://www.embitel.com/blog/embedded-blog/decoding-the-component-concept-of-the-application-layer-in-autosar, 2018.

[4] GRIOUI Haythem. AUTOSAR for Absolute Quality for Embedded Automative Software Applications. https://www.addixo.com/autosar-for-absolute-quality-for-embedded-automative-software-applications/, 2023.

[5] Bunzel Stefan. AUTOSAR – The Standardized Software Architecture. https://gi.de/informatiklexikon/autosar-the-standardized-software-architecture, 2011.

[6] Incorporated Texas Instruments. Introduction to AUTOSAR. https://software-dl.ti.com/hercules/hercules_docs/latest/hercules/AutoSAR_MCAL/AutoSAR_MCAL.html, 2018.

[7] GmbH Vector Informatik. DaVinci Developer & DaVinci Configurator Pro. https://de.mathworks.com/products/connections/product_detail/davinci-developer-and-davinci-configurator-pro.html, 2023.

[8] GmbH Vector Informatik. DaVinci Developer Adaptive Successfully Configuring MICROSAR Adaptive Software. https://www.vector.com/us/en/products/products-a-z/software/davinci-developer-adaptive/#, 2023.

[9] GmbH Vector Informatik. DaVinci Developer Classic User-Friendly Design of AUTOSAR Software Components. https://cdn.vector.com/cms/content/products/DaVinci/DaVinci_Developer_FactSheet_EN.pdf, 2023.

Generation von 3D Stadt Modellen mit dem Wave Function Collapse Algorithmus in Houdini

Lukas Landhaeusser **Andreas Rößler**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma Porsche Engineering Services GmbH, Bietigheim-Bissingen

Einleitung

Die Fahrerassistenzsysteme in modernen Fahrzeugen werden immer komplexer. Um diese Systeme ausgiebig testen zu können, werden virtuelle Testfahrten durchgeführt. Virtuelle Testfahrten können dabei in einem Simulator mit echtem Fahrer stattfinden oder vollständig virtuell. Zweiteres kann beispielsweise für die Entwicklung von autonomen Fahrzeugen genutzt werden. [4] Um solche Testfahrten in einem Stadtgebiet durchzuführen, muss diese erst Modelliert werden.

Problemstellung

Das manuelle Modellieren von 3D-Städten ist sehr zeitaufwändig. Um diesen Aufwand zu verringern, kommt für die Erstellung der 3D-Szenen das Programm Houdini zum Einsatz. Houdini ermöglicht das Prozedurale generieren von einzelne 3D-Objekten oder ganzen Szenen. Houdini wird hauptsächlich in der Videospiel- und Filmindustrie eingesetzt. [5] Für die virtuellen Testfahrten werden sogenannte HD-Karten verwendet. Im Gegensatz zu regulären Karten von Anbietern wie Google Maps oder OpenStreetMap besitzen diese Karten eine höhere Genauigkeit. Da die HD-Karten oft keine Informationen über Objekte abseits der Straße beinhalten, müssen diese nachträglich hinzugefügt werden.

Zielsetzung

Ziel des Projektes ist es HD-Karten durch prozedural generierte Objekte und Objekte von OpenStreetMap zu erweitern. Dabei liegt der Fokus auf Objekten, die im Blickfeld des Fahrers oder eines Kamerasensors liegen. Die Objekte sollen anhand der Kartendaten vollständig prozedural generiert werden.

Umsetztung

Das Straßennetzwerk wird vom Program RoadRunner als 3D-Modell im FBX (Filmbox) exportiert. Dieses 3D-Modell dient als Grundlage für die Stadtgeneration. Die Umrisse der Gebäude werden von OpenStreetMap bezogen. Wenn kein OpenStreetMap zur Verfügung steht, werden die Gebäude anhand von Voronoi-Diagrammen generiert. Die Fassaden der Gebäude bestehen dann aus vorgefertigten Modulen namens Tiles. Die Tiles werden mithilfe des Wave Function Collapse (WFC) Algorithmus angeordnet. In Abbildung 1 ist eine Stadt zu sehen, welche basierend auf einem Voronoi-Diagramm generiert wurde. Jedes Gebäude verwendet dabei dieselben Tiles.

Abb. 1: Voronoi bassierte Stadt [1]

Houdini

Houdini setzt bei der Erstellung von 3D-Modellen auf eine visuelle Programmiersprache. Im Gegensatz zu traditionellen Programmen wie 3ds Max oder Maya ermöglicht dies das Arbeiten mit prozeduralen Regeln. Durch das verändern dieser Regeln können so schnell Veränderungen am Modell vorgenommen werden. Durch das verändern des Generations Schlüssels, kann in Abbildung 1 die Gebäudeanordnung mit einer Eingabe verändert werden. Die vordefinierten Bausteine der visuellen Programmiersprache können durch eigene Bausteine erweitert werden. Diese können beispielsweise in Python geschrieben werden. [2]

Wave Function Collapse

Der Wave Function Collapse Algorithmus kann in zwei Versionen implementiert werden. Für dieses Projekt wird die einfachere Version namens Simple Tiled verwendet. Im Startzustand des Algorithmus kann jede Zelle jedes Tile annehmen. In Abbildung 2 wird die Funktion des WFC- Algorithmus vereinfacht dargestellt. Jeder Zelle können drei verschiedene Tiles zugeordnet werden. Die zuvor festgelegten Regeln legen fest:

- Neben einer Wiese (grün) kann eine Wiese und ein Strand vorkommen
- Neben einem Gewässer (blau) kann ein Gewässer oder ein Strand vorkommen
- Neben einem Strand (gelb) ist alles erlaubt

Der Algorithmus sucht sich durch Zufall ein Anfangs Zelle aus. Dieser Zelle wird ein zufälliges Tile zugewiesen. Danach werden rekursiv die Tiles aus den benachbarten Zellen entfernt, welche nicht mehr möglich sind. Dieser Prozess wird iterativ wiederholt. Ist in jeder Zelle nur noch ein Tile übrig, ist der Algorithmus fertig. [3]

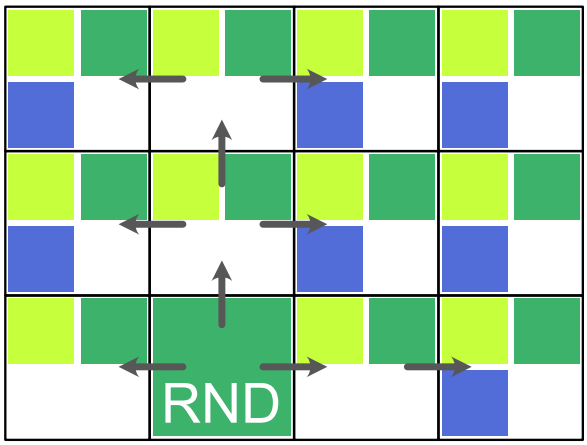


Abb. 2: Erste Iteration des WFC Algorithmus [1]

Die Auswahl der Tiles des WFC-Algorithmus, kann durch Bedingungen eingeschränkt werden. Die Gebäude in Abbildung 1 haben daher extra Tiles für das Erdgeschoss. Dies verhindert Türen und Schaufenster in den oberen Stockwerken. Wie das Beispiel in Abbildung 2 zeigt, kann der Algorithmus nicht nur für die Generation von Gebäuden genutzt werden. Eine Generation von Parks oder ähnlichem wäre daher auch möglich.

Herausforderungen

Die HD-Karten können mehrere Kilometer an Straßen beinhalten. Das kann schnell zu Städten mit mehreren Tausend Gebäuden führen. Durch die hohe Gebäudeanzahl ist es wichtig möglichst wenig überflüssige Objekte zu generieren. Kartendaten von OpenStreetMap und die HD- Karten können aufgrund der unterschiedlichen Genauigkeit nicht direkt miteinander kombiniert werden. So müssen die OpenStreetMap Daten angepasst werden. Auch die Qualität der OpenStreetMap Karten ist nicht immer gleich. Beispielsweise ist in manchen Regionen die Höhe der Gebäude angegeben. Meistens muss diese allerdings prozedural ergänzt werden. Für Straßen mit erhöhter Steigung müssen Gebäude mit dem Untergrund überlappen, was zu Unrealistischen Ergebnissen führen kann.

Ausblick

Um die verschiedenen Gebäudetypen einer Stadt zu unterstützen, müssen weitere Tilesets erstellt werden. Für eine schnellere Generierung und bessere Performance im Simulator, müssen überflüssige Objekte vor der Generation entfernt werden.

Literatur und Abbildungen

[1] Eigene Darstellung.

[2] Dennis Dornia and Nicolas Fischöder. *Interaktive Datenvisualisierung in Wissenschaft und Unternehmenspraxis*. Springer, 2020.

[3] Robert Heaton. The Wavefunction Collapse Algorithm explained very clearly. https://robertheaton.com/2018/12/17/wavefunction-collapse-algorithm/, 12 2018.

[4] Oleksandr Odukha. Three Ways of ADAS Testing in Autonomous Cars. https://intellias.com/three-ways-of-testing- adas-in-autonomous-cars-beyond-a-test- drive/, 2018.

[5] Kirill Tokarev. Procedural Technology in Ghost Recon: Wildlands. https://80.lv/articles/procedural-technology-in-ghost-recon-wildlands/, 04 2017.

Entwicklung einer Library für optimierte Kopierroutinen mittels Just-in-time-Kompilierung zur Verwendung von Vektorinstruktionen

Janis Latus **Rainer Keller**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt an der Fakultät Informatik und Informationstechnik, Esslingen

Einleitung

Das Kopieren von Daten, von einem Speicherbereich in einen anderen Speicherbereich, ist Bestandteil der meisten Computerprogramme. Die Dauer des Kopiervorgangs ist im wesentlichen von zwei Faktoren abhängig. Zum einen der Anzahl der Instruktionen, je weniger Instruktionen innerhalb einer Kopierroutine ausgeführt werden müssen, desto schneller kann die Kopierroutine ausgeführt werden. Zum anderen beeinflusst auch die Lokation der Daten in verschiedenen Speicherbereichen die Dauer der Kopierroutine. So kann auf Daten im Prozessorcache deutlich schneller zugegriffen werden als auf Daten im Hauptspeicher.
In den letzten Jahren gab es große Weiterentwicklungen der x86-Prozessorarchitektur. So wurden mit dem Intel Pentium III im Jahr 1999 erstmals die Streaming SIMD Extensions (SSE) eingeführt, welche Operationen auf 128-Bit weiten Registern ermöglichen. Mit der Sandy Bridge Mikroarchitektur des Jahres 2011 wurden die Advanced Vector Instructions (AVX) eingeführt, welche Operationen auf 256-Bit weiten Registern erlauben. 2016 wurde der AVX-Instruktionssatz um AVX-512 erweitert, mit welchem nun sogar Operationen auf 512-Bit weiten Registern durchgeführt werden können. [1]

Problemstellung

Mit den vorgestellten Erweiterungen des x86-Instruktionssatzes lassen sich Kopierroutinen deutlich beschleunigen, da mit einer Instruktion nun 16, 32 oder 64 Bytes zwischen Registern und Speicher kopiert werden können. Die Verwendung der entsprechenden Instruktionen ist aber vom verwendeten Compiler abhängig. Zwar unterstützen gängige Compiler wie GCC und Clang die vorgestellten Instruktionen, das Layout der Daten muss aber zum Zeitpunkt der Kompilierung bekannt sein.

Ein Anwendungsfall, bei welchem das Layout der Daten zur Kompilierzeit eben nicht bekannt ist, ist das Message-Passing Interface (MPI). MPI ist eine Spezifikation für Libraries zur Kommunikation und Datenbewegungen zwischen Prozessen, um Probleme effizient über mehrere Rechenknoten verteilen zu können. MPI spezifiziert neben Basisdatentypen ("basic datatypes"), deren Definition an C-Datentypen angelehnt ist, auch abgeleitete Datentypen ("derived datatypes"), welche im Prinzip eine Beschreibung von nicht zusammenhängenden Daten sind. Also Sequenzen von Basisdatentypen, welche durch Lücken ("displacements") voneinander getrennt sind. Um diese Daten an einen anderen Prozess zu senden, müssen sie zuerst in einem zusammenhängenden Speicherbereich gepackt werden, welcher über das Netzwerk versendet werden kann. [2]
Das Ziel dieser Arbeit ist es, eine Library zu entwickeln, welche zur Laufzeit optimierten Maschinencode für das Packing und Unpacking solcher komplexen Datentypen generieren kann. Diese soll in das OpenMPI-Framework, eine Implementierung des MPI Standards, eingebunden werden. Zudem soll die Performance der Neuimplementierung analysiert und mit der existierenden Lösung verglichen werden.

Lösungsansatz

Ein Ansatz zur Generierung von Maschinencode zur Laufzeit wird in [3] beschrieben. Hier wurde ein LLVM-Backend für die Implementierung eines Laufzeitcompilers für MPI-Datentypen genutzt. Das Paper zeigt, dass für manche Datentypen die zur Laufzeit kompilierten Packingfunktionen bis zu siebenmal schneller sind als die MPI-Implementierung (Vergleich mit Cray MPI). Allerdings beträgt die Dauer der Kompilierung mehrere Millisekunden, weshalb die Datentypen oft benutzt werden müssen, um die zusätzlichen Laufzeitkosten für die Kompilierung

auszugleichen. Zudem werden mit diesem Ansatz zusätzliche Abhängigkeiten zum Compiler-Backend eingeführt.
Deshalb soll in dieser Arbeit kein Compiler-Backend zur Generierung von Maschinencode verwendet werden, sondern der Maschinencode anhand der Analyse der Struktur der MPI-Datentypen durch die Library selbst erzeugt werden. D.h. die Library muss korrekte Instruktionen für die x86-Prozessorarchitektur erzeugen können und diese dann in richtiger Reihenfolge in einem ausführbaren Speicherbereich ablegen, auf welchen dann durch einen Funktionszeiger verwiesen wird. Der Aufbau einer Instruktion wird in Abbildung 1 gezeigt. Eine valide Instruktion hat eine maximale Länge von 15 Bytes. Der Instruktionsprefix ist 1 Byte lang und wird in vier Gruppen unterteilt. Jede Gruppe erhält ein Set von erlaubten Prefix-Codes. Das Opcode-Feld kann 1 bis 3 Byte lang sein kann und definiert welche Instruktion ausgeführt werden soll. Das ModR/M-Byte spezifiziert die Operanden einer Instruktion und das SIB-Byte wird für eine erweiterte Adressierung mit dem ModR/M-Byte genutzt. Das Displacement-Feld kann 1, 2 oder 4 Byte lang sein und wird genutzt um den Hauptspeicher zu adressieren. Das Immediate-Feld kann ebenfalls 1, 2 oder 4 Byte lang sein und spezifiziert Konstanten, welche als Operand einer Instruktion genutzt werden können. Des Weiteren kann ein REX-Prefix im 64-Bit Modus dem Opcode vorangestellt sein um z.B. 64-Bit Operanten zu verwenden. Für AVX-Instruktionen wird dem Opcode ein 2 oder 3 Byte langes VEX-Prefix und für AVX-512-Instruktionen ein 4 Byte langes EVEX-Prefix vorangestellt. [1]

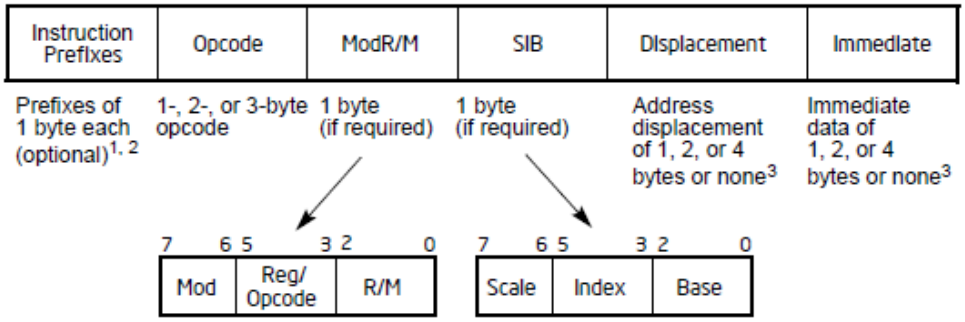


Abb. 1: Aufbau einer x86-Instruktion [1]

Stand der Arbeit

Zum aktuellen Stand der Arbeit kann eine vereinfachte Form von Kopierroutinen automatisch generiert werden. Die Library bietet eine Funktion an, welche Packingroutinen auf Basis der internen Datenrepräsentation von OpenMPI generieren kann und zum Zeitpunkt des Commits eines Datentyps durch OpenMPI aufgerufen wird. Ein Funktionszeiger wird dann gespeichert, um die Packingroutine beim Versenden der Daten aufrufen zu können. Es werden sowohl Standard x86-Instruktionen als auch SSE-, AVX- und AVX-512-Instruktionen unterstützt. Es können MOV-Instruktionen zum Kopieren von 1, 2, 4 oder 8 Bytes verwendet werden. Zum Kopieren von 16 Bytes werden MOVDQA- und MOVDQU-Instruktionen und zum Kopieren von 32 oder 64 Bytes werden VMOVDQA- bzw. VMOVDQU-Instruktionen angeboten.

Ausblick

Zunächst soll die Library auf einen stabilen Stand gebracht werden, sodass Packingroutinen für alle Arten von Datentypen in OpenMPI zuverlässig generiert werden können. Anschließend soll die Generierung der Unpackingroutinen entsprechend implementiert werden. Nach einer ersten Analyse der Performance, im Vergleich zur OpenMPI-Implementierung, soll die Codegenerierung optimiert werden. Abschließend soll ein ausführlicher Performancevergleich der entwickelten Lösung mit der OpenMPI-Implementierung durchgeführt werden. Sind die Ergebnisse ausreichend positiv, kann die entwickelte Library eventuell von OpenMPI übernommen werden.

Literatur und Abbildungen

[1] Intel Corporation. Intel 64 and IA-32 Architectures Software Developer's Manual. https://cdrdv2.intel.com/v1/dl/getContent/671200, 10 2023.

[2] MPI Forum. MPI: A Message-Passing Interface Standard Version 4.0. https://www.mpi-forum.org/docs/mpi-4.0/mpi40-report.pdf, 06 2021.

[3] Timo Schneider, Fredrik Kjolstad, and Torsten Hoefler. MPI datatype processing using runtime compilation. *EuroMPI '13: Proceedings of the 20th European MPI Users' Group Meeting*, pages 19–24, 2013.

Entwicklung einer Smart Office App zur Messung und Analyse der Lichtverhältnisse in einem Büro

Armin Lezic **Jörg Nitzsche**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma Systecs Informationssysteme GmbH, Leinfelden-Echterdingen

Einleitung

Die Digitalisierung ist in nahezu allen Lebensbereichen vorzufinden und hat mittlerweile sogar das Büroumfeld erreicht. Einen Arbeitsplatz mit optimalen Arbeitsbedingungen bieten zu können, gewinnt dabei zunehmend an Bedeutung. Insbesondere die Lichtverhältnisse am Arbeitsplatz beeinflussen das Wohlbefinden und somit die Produktivität der Mitarbeiter. Die Bachelorarbeit fokussiert sich auf die Entwicklung einer innovativen Smart Office App, die Echtzeitinformationen zu den Bürolichtverhältnissen erfasst und darstellt.

Motivation

Die Motivation für die Entwicklung einer Smart Office App zur Lichtmessung und -analyse liegt darin, dass eine Verbesserung der Lichtverhältnisse einen positiven Effekt auf das Wohlbefinden und folglich auf die Produktivität eines Mitarbeiters hat [3]. Durch die kontinuierliche Erfassung der Lichtverhältnisse ermöglicht die App Nutzern, die gesetzlichen Vorgaben über die Beleuchtung zu überprüfen. Die App soll dem Mitarbeiter zeigen, ob diese gesetzlichen Bestimmungen eingehalten werden oder ob es einer Handlung bedarf.

Zielsetzung

Ziel dieser Bachelorarbeit ist die Entwicklung einer Smart-Office-Anwendung, die Informationen über Bürobeleuchtungsverhältnisse in Echtzeit erfassen, anzeigen und entsprechend den gesetzlichen Anforderungen überprüfen kann. Ebenfalls sollen die individuellen Bedürfnisse des Benutzers berücksichtigt werden. Die App soll dazu beitragen, dass im modernen Büro optimale Arbeitsbedingungen entstehen und die Gesundheit sowie die Produktivität der Mitarbeiter gefördert werden.

Gesetzliche Vorgaben

Die Technische Regelung für Arbeitsstätten (ASR) definiert den Arbeitsplatz (Apl) als die Zusammensetzung aus der Arbeitsfläche, den Bewegungsflächen und Stellflächen, die dem unmittelbaren Fortgang der Arbeit dienen. Für Büros und Büro ähnliche Arbeitsbereiche, in denen Aufgaben wie Schreiben Lesen oder Datenverarbeitung verrichtet werden, wird eine mittlere Beleuchtungsstärke von mindestens 500lx vorgeschrieben. Im Umgebungsbereich (UB) darf die mittlere Beleuchtungsstärke 300lx nicht unterschreiten. Als Umgebungsbereich wird der Raum um den Arbeitsplatz herum definiert, der an einen anderen Arbeitsplatz anschließt oder durch eine Wand oder einem Verkehrsweg umgeben ist [2]. Die Abbildung 1 visualisiert die Definition.

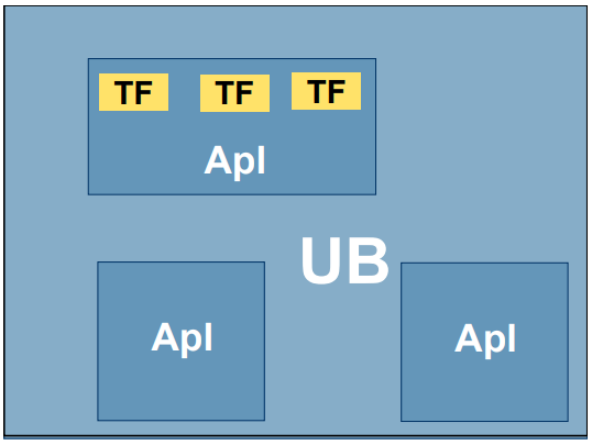


Abb. 1: Skizze Arbeitsplatz [2]

Um die Problemstellung einfacher zu halten, werden teilflächenbezogene (TF) Beleuchtungen nicht betrachtet. Außerdem gilt, dass die mittlere vertikale Beleuchtungsstärke mindestens 175lx beträgt. Unter der mittleren vertikalen Beleuchtungsstärke wird die durchschnittliche Beleuchtungsstärke auf einer vertikalen Fläche verstanden.

Konzept

Die Abbildung 2 gibt einen Überblick über das Konzept der Anwendung. In einer Cloud wird die Datenbank, die als Ringpuffer agiert, zusammen mit dem Front- und Backend bereitgestellt. Im Büroraum befindet sich der Mitarbeiter und mehrere Lichtsensoren. Über ein Script werden die Sensordaten ausgelesen und an das Backend übermittelt. Diese können anschließend vom Frontend abgerufen und für den Mitarbeiter visualisiert werden. Die Überprüfung, ob die Gesetze eingehalten werden, erfolgt im Backend.

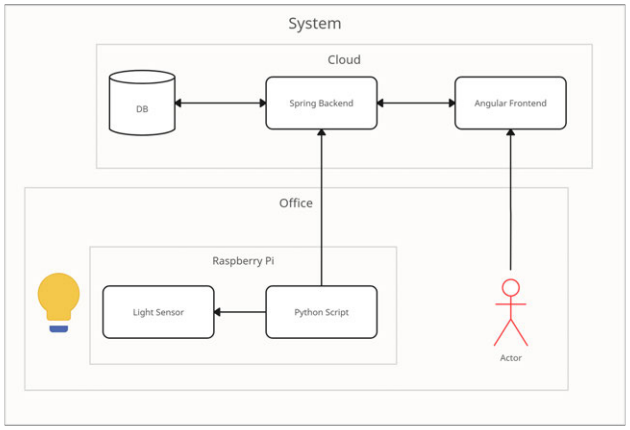


Abb. 2: Übersicht des Konzepts [1]

Realisierung der Lichtsensoreinheit

Für die Messung der Lichteinstrahlung wird ein Lichtsensor BH1750 benutzt, der an einem Raspberry Pi Zero 2 W angeschlossen wird. Die Abbildung 3 zeigt den fertigen Aufbau. Mithilfe der Adafruit Bibliothek für Python kann der Sensor einfach ausgelesen werden. Die Bibliothek Requests ermöglicht es die Daten über einen REST-POST Aufruf an das Backend zu übertragen. Dank des Akkus ist der Raspberry Pi mobil und benötigt keine Steckdose. Die Einheit kann somit überall platziert werden. In einem Büroraum werden 10 Sensoren verwendet, um die Lichtverhältnisse präzise zu erfassen.

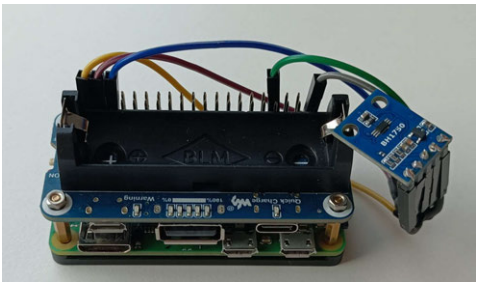


Abb. 3: Raspberry Pi mit Lichtsensor BH1750 [1]

Ausblick

Die Arbeit soll zeigen ob es möglich ist mit Lichtsensoren die Beleuchtung in einem Büro zu überwachen und zu prüfen. Wenn dies der Fall ist, können die Arbeitsplätze in Zukunft optimiert werden und das Wohlbefinden, sowie die Gesundheit der Mitarbeiter verbessert werden.

Literatur und Abbildungen

[1] Eigene Darstellung.

[2] Bundesanstalt für Arbeitsschutz und Arbeitsmedizin. *Technische Regeln für Arbeitsstätten - Beleuchtung und Sichtverbindung*. BAUA, 2023.

[3] Nils Warkentin. Ergonomische Beleuchtung am Arbeitsplatz: Richtiges Licht. https://karrierebibel.de/ergonomische-beleuchtung-am-arbeitsplatz/, 2023.

Erstellung eines Rollen- und Berechtigungskonzept im Zuge der Migration von SAP ECC zu SAP S/4HANA

Robin Lidle **Thomas Rodach**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma Mercedes AMG GmbH, Affalterbach

Einleitung

Oftmals werden Berechtigungen von Unternehmen auf die leichte Schulter genommen und zu selten überprüft, ob Berechtigungen für den aktuellen Einsatz noch gebraucht werden. Im schlimmsten Fall können finanzielle Auswirkungen durch Ausnutzen von Berechtigungen auf Unternehmen zu kommen. Im Jahr 2023 entstanden Schäden der deutschen Wirtschaft durch Diebstahl von IT-Ausrüstung und Daten in Höhe von 206 Milliarden Euro. Nach 2021 und 2022 ist es das dritte Jahr in Folge welches die 200-Milliarden-Euro-Marke übertrifft. Für 72 Prozent des Schadens sind Cyberattacken verantwortlich. Der Trend zu Angriffen im digitalen Raum setzt sich vor allem aus Russland und China zunehmend fort. An der Spitze der Attacken steht Phishing, Angriffe auf Passwörter sowie Infizierung mit Schadsoftware. Um dagegenzuwirken, reagieren Unternehmen mit höheren Investitionen in die IT-Sicherheit. [3] Die Auswirkungen, im Falle eines erfolgreichen Angriffs auf einen Benutzer, hängen dementsprechend vom Berechtigungsumfang ab. Je höher die Rechte eines kompromittierten Benutzers, desto größer der Schaden der durch den Angreifer verursacht werden kann.

Zielsetzung

Im Rahmen der Abschlussarbeit wird ein Rollen- und Berechtigungskonzept im Zuge der Migration von SAP ECC auf SAP S/4HANA erstellt. Auf Basis des alten SAP ECC werden Rollen und Berechtigungen analysiert und bestmöglich auf den Endbenutzer zugeschnitten. Zur Hilfestellung wird das Tool Xiting Authorizations Management Suite (kurz XAMS) der Firma Xiting AG eingesetzt (siehe Abb. 1). XAMS soll kritische und zeitraubende Phasen in der Erstellung von Berechtigungsprojekten drastisch verkürzen und effizient bewältigen. Ziel der Arbeit ist die Anpassung der Rollen und Berechtigungen nach dem Minimalprinzip. Zusätzlich soll das Konzept als Grundlage für die zukünftige Rechte- und Rollenvergabe dienen.

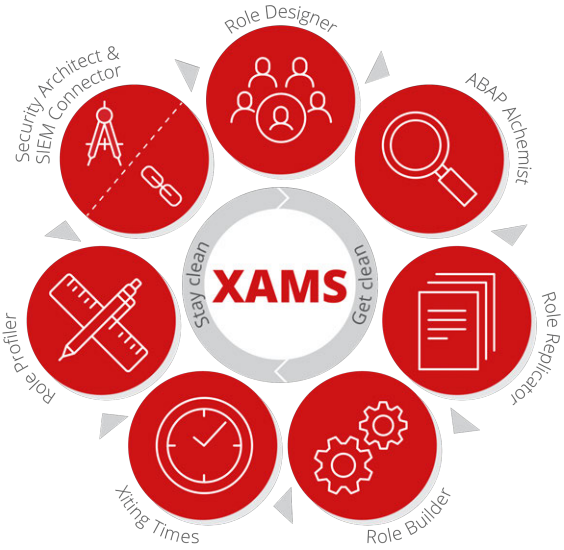


Abb. 1: Überblick der XAMS-Module [1]

Rollenbasierte Zugriffskontrolle (RBAC)

Mithilfe der rollenbasierten Zugriffskontrolle lässt sich die starre und schwer skalierbare Subjekt-Objekt-Relation aufbrechen. [2] Bei RBAC werden die Rechte nicht mehr direkt an Subjekte, sondern an Rollen (engl. roles) geknüpft. Sie stellen die Zusammenfassung von Rechten dar, die zur Erfüllung von mit ihnen verbundenen Aufgaben und Funktionen notwendig sind. Zwischen Rollen und Benutzern sowie Zugriffsrechten auf Objekte existiert eine n:n-Beziehung (Viele-zu-viele-Beziehung).

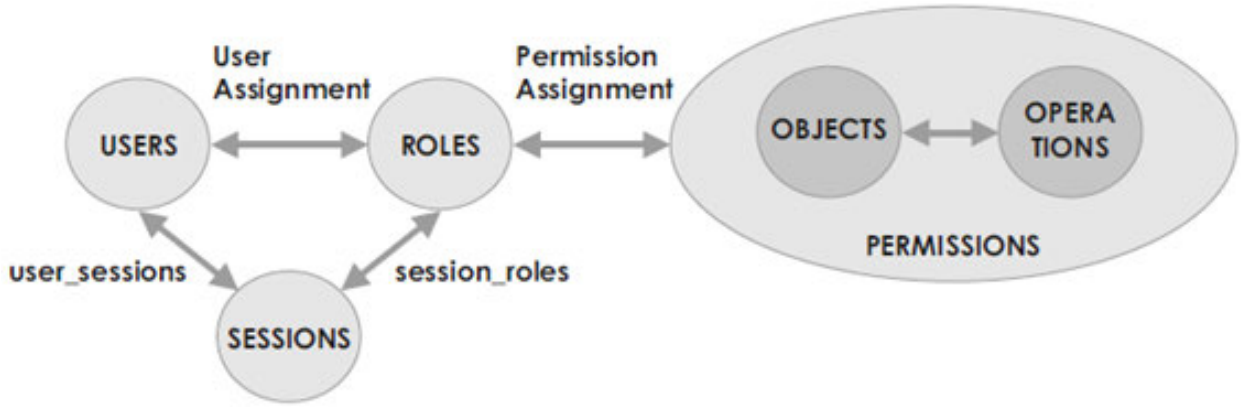


Abb. 2: Referenzmodell RBAC [4]

Bei der Veränderung des Aufgabenbereichs innerhalb einer Firma wird dem Mitarbeiter lediglich eine neue Rolle zugewiesen, die über die erforderlichen Berechtigungen verfügt. Im Falle, dass eine Rolle nicht mehr benötigt wird, um die Aufgaben und Funktionen zu erfüllen, wird sie dementsprechend entzogen. Der Aufwand für die Rechteverwaltung wird dadurch wesentlich reduziert. Abbildung 2 bildet die einzelnen Elemente eines RBAC-Referenzmodell ab.

Umsetzung

Die Erstellung des Rollen- und Rechtekonzepts wurde in verschiedene Phasen untergliedert die nacheinander durchgeführt werden:

1. Phase: Systemanalyse
2. Phase: Auswertung der Analyse
3. Phase: Entscheidung von Fachbereichen und Modulverantwortlichen für Änderungen
4. Phase: Vorarbeiten im SAP S/4-System
5. Phase: Anpassung der Rollen
6. Phase: Testen im Konsolidierungssystem
7. Phase: Transport der Rollen in das Produktivsystem

Ausblick

Im weiteren Verlauf der Bachelorarbeit sollen alle Phasen erfolgreich umgesetzt werden um den Benutzern passende Berechtigungen zur Ausführung ihrer Arbeit nach dem Minimalprinzip zu gewährleisten und potentielle Risiken minimieren. Des Weiteren soll die SAP Transaktion SU25 (Upgrade Tool für den Profilgenerator) in Zukunft auf aktuellem Stand gehalten werden, um den damit verbundenen Mehraufwand zu reduzieren. Im Anschluss der wissenschaftlichen Arbeit soll das Rollen- und Berechtigungskonzept als Grundlage für die Mercedes AMG dienen.

Literatur und Abbildungen

[1] Xiting AG. XITING AUTHORIZATIONS MANAGEMENT SUITE (XAMS). https://xiting.com/de/xams/, 2023.

[2] Heiko Klarl. *Zugriffskontrolle in Geschäftsprozessen: Ein modellgetriebener Ansatz*. Vieweg+Teubner Verlag, 2011.

[3] Andreas Streim and Simran Mann. Organisierte Kriminalität greift verstärkt die deutsche Wirtschaft an. https://www.bitkom.org/Presse/Presseinformation/Organisierte-Kriminalitaet-greift-verstaerkt-deutsche-Wirtschaft-an, 09 2023.

[4] Alexander Tsolkas and Klaus Schmidt. *Rollen und Berechtigungskonzepte: Identity- und Access-Management im Unternehmen*. Springer Vieweg, 2017.

Entwurf, Entwicklung und Implementierung eines kontinuierlichen Testprozesses für die Software von Ladestationen

Joel Kevin Likane Zindjou **Mirko Sonntag**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma eSystems MTG GmbH, Wendligen am Neckar

Motivation

Die Produktion von Software in traditioneller Softwareentwicklungsart ist oft einen mühsamen Prozess. Manuelle Integration-, Konfigurations- und Testschritte nehmen meistens Wochen in Anspruch und jederzeit kann ein unbeabsichtigtes Problem vorkommen, das alle Beteiligten dazu zwingt, den gesamten Prozess wieder von vorne zu beginnen. Dieser zeitaufwendige Prozess der Code-Vorbereitung für eine Release führt oft dazu, dass Änderungen nur alle paar Monate veröffentlicht werden.
Die damit verbundenen manuellen Schritte erhöhen die Komplexität und das Risiko von Fehlern, was zu Verzögerungen und ineffizienten Abläufen führt. Selbst kleinste Fehler können massive Auswirkungen haben und die Bereitstellung der Software verhindern.
Zur Lösung dieser Probleme wird häufig das Konzept von Continuous Integration und Continuous Deployment (CI/CD) in IT-Unternehmen eingesetzt. CI/CD sind heute Schlüsselkomponenten der IT-Automatisierung. [4] Dabei werden Softwareänderungen kontinuierlich und automatisch getestet und in die Produktionsumgebung überführt. Einige Vorteile dazu:

- **Schnellere Markteinführungen:** Das primäre Ziel der CI/CD ist, schnell und häufig funktionierende Software auf dem Markt zu bringen [5]
- **Bessere Softwarequalität:** Durch kontinuierliche Integration und Tests werden Fehler und Probleme frühzeitig im Entwicklungsprozess erkannt und behoben
- **Kostenreduktion:** Eine frühzeitige Erkennung und Behebung der Fehler reduziert erheblich die Kosten für spätere Reparaturen oder die Behebung von Schwierigkeiten in der Produktionsumgebung

Benötigte Werkzeuge und Tools

- **Jenkins**

  Jenkins ist ein beliebtes Tool für die Implementierung von Continuous Integration (CI) und Continuous Delivery (CD). Es ist ein Open-Source-Automatisierungsserver, der Entwicklern unterstützt, ihre Software zuverlässig zu erstellen, zu testen und bereitzustellen. Mögliche Aufgaben sind Automatisierung von Build-Prozessen, Durchführung von Tests und Bereitstellung von Softwareänderungen. [1]

- **EXAM**

  EXAM steht für EXtended Automation Method, wurde von MicroNova AG, VW und Audi für einen vereinheitlichten Test entwickelt. Es ist eine integrierte Entwicklungsumgebung für Tests und bietet eine optimale Unterstützung insbesondere in Tests von Steuergeräten auf Hardware in the Loop (HiL) Prüfständen. Eine Aufteilung des Testprozesses in verschiedene Rollen, wie z.B. Testdesigner, Testspezifikateur oder Testausführer, wird von EXAM unterstützt. Dabei formalisiert EXAM eine einheitliche Sprache zur Darstellung von Testsachverhalten. [2]

Ziel der Arbeit

Das Ziel der Abschlussarbeit besteht darin, mithilfe eines Jenkins-Servers einen kontinuierlichen Testprozess für Ladesäulen-Software zu schaffen, um die Tests automatisch durchlaufen zu lassen. Wie auf Abbildung 1 zu sehen ist, werden bisher die Software-Images vom Test-Ingenieur immer manuell abgeholt, sobald eine neue Software-Release zur Verfügung steht. Ebenso werden die Tests auf dem Prüfstand mit dem Test-Tool EXAM manuell ausgeführt.

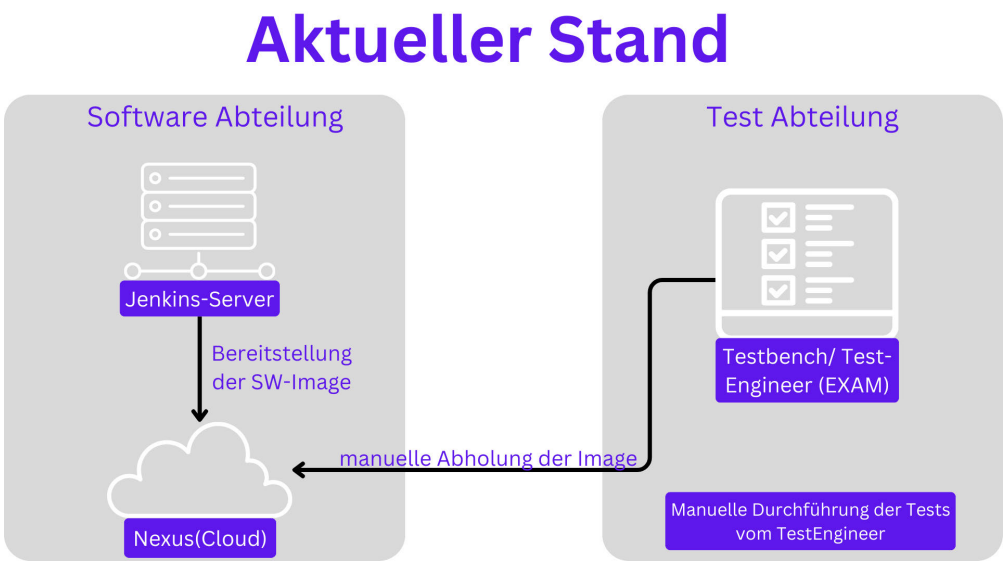


Abb. 1: Aktueller Stand [3]

Konzept

Abbildung 2 stellt konkret, das zu implementierenden Konzept dar: Sobald eine neue Software-Image in der Cloud bereitgestellt wurde, sollte der Jenkins-Server den automatischen Prozess starten(automatische Abholung der Image im Cloud und Durchführung von Tests auf dem Prüfstand). Dabei ist die Untersuchung der Schnittstelle zwischen EXAM und Jenkins einer der größer Teil der Arbeit.

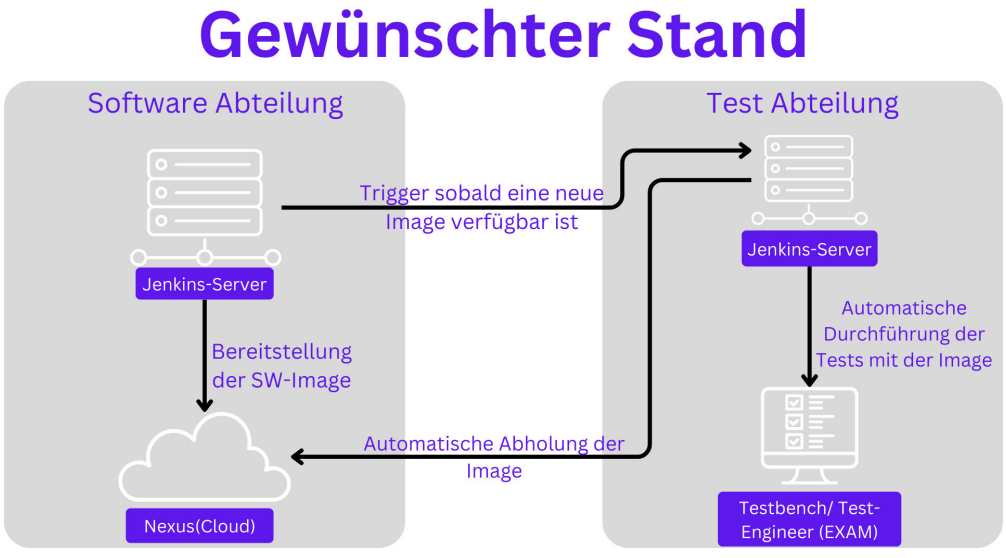


Abb. 2: Stand nach Arbeit [3]

Ausblick

Bis jetzt war nur CI/CD in der Software-Abteilung implementiert und mit der Arbeit wird eine volle Automatisierung bei der Produktion der Software erreicht.

Literatur und Abbildungen

[1] GFU Cyrus AG. Definition Jenkins. https://www.gfu.net/wiki/jenkins.html, 2022.

[2] MicroNova AG. Definition von EXAM. https://www.micronova.de/testing/exam-testautomation.html, 2008.

[3] Eigene Darstellung.

[4] SysEleven GmbH. Warum CI/CD entscheidend sind. https://www.syseleven.de/blog/it-automatisierung-warum-ci-cd-entscheidend-sind/, 2023.

[5] s. JetBrains. Vorteile CI/CD. https://www.jetbrains.com/de-de/teamcity/ci-cd-guide/benefits-of-ci-cd/, 2020.

Vergleich von Generativen KI-Tools für das Drucken von textbasierten 3D Modellen

Vidal Lopez Huergo **Dieter Morgenroth**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt an der Fakultät Informatik und Informationstechnik, Esslingen

Einleitung

Die allgemeine Begeisterung, die in der ersten Hälfte von 2023 rund um das Thema Chat GPT ihren Hochpunkt gefunden hat, hat einen Schwarm von Generativen Tools inspiriert. Weltweit haben Forscher ihre Zeit in das Thema Generative KI investiert und es sind fast täglich neue Anwendungen erschienen. Zu dem Zeitpunkt war es im Umfang des Praxissemesters, die Aufgabe ein CAD-Modell zu erstellen und dies zu drucken.
Das Zusammenspiel dieser Einflüsse führte zu der Idee, Texteingaben und das Verständnis von Chat GPT mit einer 3D-Modell-Ausgabe zu koppeln. Die Volatilität des Themenbereiches und die seitdem vergangene Zeit beweisen, dass andere die Idee teilten. Mittlerweile gibt es eine Handvoll Forschungsergebnisse und sogar kommerzielle Plattformen, welche aus Texteingaben 3D-Modelle generieren.

Erläuterung der Forschungsfrage

Je nach unterliegender Systemarchitektur, werden von Plattform zu Plattform unterschiedliche Ergebnisse erzielt. In dieser Bachelorarbeit sollen die Arbeitsabläufe und die Produkte verschiedener Text-zu-3D-Tools vergleichend betrachtet werden. Spezifisch stellt sich dazu die Frage: **Welche Algorithmen von Generativen Text-zu-3D-Modell-Tools eignen sich am besten für das Erstellen von Druckbaren 3D-Objekt-Modellen in verschiedenen Disziplinen?**

Ziel der Arbeit

Ziel der Arbeit ist es, einen qualitativen Vergleich aufzustellen, welche Software druckbare Ergebnisse erzeugt und welche messbaren Unterschiede die verschiedenen Ansätze und Algorithmen aufweisen.

Aufbau und Methodik

Das Ziel des qualitativen Vergleichs der verschiedenen Vorgehensweisen erfordert zunächst die Analyse der Prozesse. Hierfür sollen zunächst mehrere Literaturquellen, insbesondere in Bezug auf deren unterschiedliche Funktionsweise, erläutert werden. Dies soll als Hilfestellung eine Übersicht über die Technologie schaffen.
Im weiteren Verlauf der Arbeit soll diese gestreute Übersicht fokussiert und die interessantesten Kandidaten, welche sich für den qualitativen Vergleich eignen, ausgewählt werden. Die auserwählten Forschungsergebnisse sollen dann dem Praxistest unterzogen werden.
Der Praxistest soll verschiedene Disziplinen umfassen. Die resultierenden Objekte der einzelnen Vorgehensweisen werden gegeneinander verglichen. Vergleichskriterien sind Beispielsweise, ob das generierte Objekt klar erkennbar und dessen Ränder korrekt definiert sind, eine Einschätzung über die Qualität der Ausgabe und ob das Objekt druckbar ist.
Den Abschluss der Arbeit soll das Drucken einiger Modelle mithilfe eines 3D-Druckers bilden.

Generative KI

Wie bereits in der Einleitung erwähnt ist Software die Generative-KI verwendet brandaktuell. Der Markt für Generative-KI ist im letzten Jahr beträchtlich gewachsen. Der Unterschied von Generativer-KI zu anderen Arten von KI steht bereits im Namen. Herkömmliche Machine-Learning Modelle werden mit Datensätzen trainiert und können Vorhersagen treffen. Generative-KI unterscheidet sich dadurch, dass Sie aus den Daten weitere künstliche Daten erzeugen kann. [6] Beispielsweise kann ein nicht generatives Machine-Learning Modell entscheiden, ob sich in einem Bild ein Gesicht befindet oder nicht. Ein generatives Modell kann aus dem Datensatz neue Gesichter erzeugen, die nicht Teil der eingegeben Daten sind, wie beispielsweise auf ThisPersonDoesNotExist.com.

Die mittlerweile bekannteste Anwendung ist wohl ChatGPT wobei die Abkürzung GPT für "Generative Pre-trained Transformer" steht. Diese KI interpretiert die menschliche Sprache und generiert Antworten, die dieser ähneln. Aufgrund der immensen Datenmenge, mit der das Large-Language-Model trainiert wurde, kann ChatGPT Antworten zu allen möglichen Themen generieren. [1]
Ein weiterer großer Anwendungsbereich liegt in der Generierung von Bildern, bekannte Namen in diesem Bereich sind unter anderen Stable-Diffusion oder Midjourney. Diese Anwendungen können aus Text innerhalb von Sekunden Darstellungen generieren und variieren. Diese Software ermöglicht beispielsweise etliche Iterationen visueller Prototypen in kürzester Zeit.

Übersicht der Text zu 3D Technologien und deren unterschiede

Im Kern funktionieren die in dieser Arbeit betrachteten Technologien gleich: im ersten Schritt werden aus der Texteingabe ein oder mehrere zweidimensionale Objekte, um diese dann im zweiten Schritt zu einem dreidimensionalen Objekt zu kombinieren. Diese Vorgehensweise geht darauf zurück, dass Trainingsdaten für die Bildsynthese millionenfach in Form von Bild mit klassifizierenden Text-Tupel verfügbar sind, was für dreidimensionale Ressourcen nicht der Fall ist. Obwohl der erste Teil der Synthese nicht unwichtig ist, werden die Unterschiede der Technologien vor allem im zweiten Teil, dem drei-dimensionalisieren, eindeutig.
DreamFusion ist zeitlich eins der ersten Papers, die zum Thema Text-zu-3D veröffentlicht wurden, weshalb sich die anderen in dieser Arbeit betrachteten Text-zu-3D-Projekte auch oftmals darauf berufen. Das generierte zweidimensionale Bild wird in einem zufällig initialisierten Neural-Radiance-Field, oder NeRF [3], von verschiedenen Positionen und Winkeln betrachtet und als Eingabe für das umschließende Score Distillation Sampling oder SDS [4] Verfahren verwendet. Einfaches Gradientenverfahren in einem willkürlichen Parameterraum, in diesem Fall im dreidimensionalen Raum, führt so zu einem 3D-Modell des Objekts. [5]
Eine andere Vorgehensweise ist das adaptieren von 3D-Gaussian-Splatting für generative Umgebungen von DreamGaussian. [2] Ein Beispiel für ein Modell, welches mit Hilfe der Three-Studio-Implementierung von DreamGaussian entstanden ist, ist auf Abbildung 1 zu erkennen.

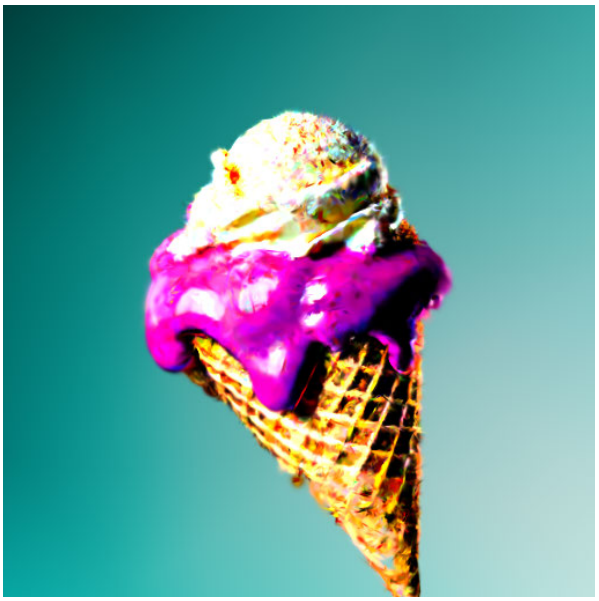


Abb. 1: Prompt: 'a delicious cone of ice cream' [2]

Ausblick

Die Nutzung von Generativer KI für das Drucken von textbasierten 3D Modellen wird mit hoher Wahrscheinlichkeit deutlich zunehmen. Grund hierfür sind unter anderem steigende Anwendungsfälle im Bereich 3D-Content-Creation wie zum Beispiel die in der Videospielindustrie oder für Metaverse. Die Interaktion mit der Software ist durch die Eingabe eines Text-Prompts auch deutlich einfacher als der Umgang mit herkömmlichen 3D-Modellierungsprogrammen. Generative Software kann in Zukunft der breiten Masse Zugriff auf das Drucken von individuellen 3D-Objekten ermöglichen.

Literatur und Abbildungen

[1] Open AI. Alles über ChatGPT. https://chatopenai.de/, 12 2023.

[2] Bernhard Kerbl et al. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *https://arxiv.org/abs/2308.04079*, 2023.

[3] Ben Mildenhall et al. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. *https://arxiv.org/abs/2003.08934*, 2020.

[4] Ben Poole et al. DreamFusion: Text-to-3D using 2D Diffusion. *https://arxiv.org/abs/2209.14988*, pages 3–5, 2022.

[5] Ben Poole et al. DreamFusion: Text-to-3D using 2D Diffusion. *https://arxiv.org/abs/2209.14988*, 2022.

[6] Adam Zewe. Explained: Generative AI. https://news.mit.edu/2023/explained-generative-ai-1109#:~:text=Generative%20AI%20can%20be%20thought,data%20it%20was%20trained%20on., 11 2023.

Vergleich, Evaluation und Beispielimplementierung von Importprozessen von KBL und VEC-Dateien in EPLAN harness proD unter Nutzung der EPLAN harness proD API

Dominik Magerle **Dieter Morgenroth**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma M-H IT Solution GmbH, Freiberg am Neckar

Einleitung

In der Kabelbaumkonstruktion im Automotive Umfeld kommen zwei standardisierte Datenformate zum Einsatz. Die Kabel Baum Liste (KBL) und der Vehicle Electric Container (VEC). Wobei der VEC ein noch relativ neues Datenformat ist, welches noch nicht von vielen Tools unterstützt wird. Die beiden Datenformate dienen vor allem dem Datenaustausch in Richtung der Fertigung der Kabelbäume. In der klassischen Elektrokonstruktion im Maschinenbau kommen diverse Konstruktionsprogramme zum Einsatz. Die beiden am häufigsten vertretenen Elektronik Computer aided Design Systeme (E-CAD-Systeme) sind Zuken E^3 und EPLAN electric P8. In der Automotive Kabelbaum Entwicklung werden überwiegend Workbenches in Mechanischen CAD Umgebungen verwendet um Verkabelungen zu konstruieren. Beispielsweise die Workbenches Electrical Library (ELB), Electrical Wire Routing (EWR) und Electrical Harness Installation (EHI) - vom Tool CATIA V5 der Firma Dassault Systems. Hierbei besteht je nach Verwendung das Problem, dass elektrische Verbindungselemente wie Steckverbinder lediglich sehr minimalistisch elektrifiziert werden, da keine detaillierten Informationen für die Konstruktion benötigt werden. Die minimalistische Elektrifizierung wirkt sich allerdings in der Fertigung aus. Hier müssen die Fertiger von Kabelbäumen die fehlenden Informationen oft mühsam händisch nachpflegen, damit eine konstruktionstreue Fertigung der Kabelbäume möglich wird. Gerade bei kleineren Fertigern sind teilweise drei oder vier Prototypen eines Leitungssatzes notwendig um die fehlenden Informationen vollumfänglich zu ergänzen. Dies bedeutet eine enorme Verschwendung von Ressourcen, da die nicht funktionierenden Prototypen verschrottet werden müssen und verarbeitete Rohstoffe kostenintensiv zurückgewonnen werden müssen. Die konkreten Aufgaben im Rahmen der Bachelorarbeit werden im Rahmen des vom Bundesministerium für Bildung und Forschung (BMBF) geförderten Vorhabens KI4BoardNet unter dem Förderkennzeichen 16ME0772 gefördert und durchgeführt.

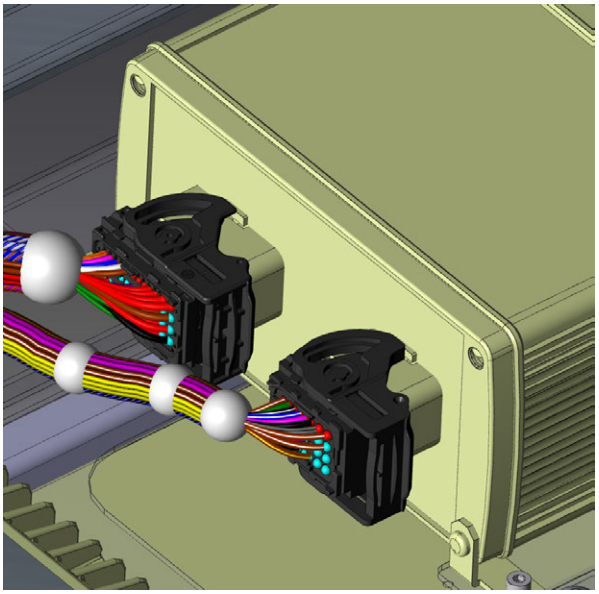


Abb. 1: Konstruktion in EPLAN harness proD [2]

Forschungsfragen

Im Rahmen der Bachelor Thesis sollen folgende Forschungsfragen bearbeitet und beantwortet werden:

- Welche Anforderungen werden an ein Bordnetzkonstruktionstool gestellt, um Problemstellungen im Automotive OEM Bereich zu bearbeiten?
- Inwiefern ist es möglich konkret EPLAN harness proD für die Bearbeitung von OEM Problemstellungen zu verwenden?

- Welche verschiedenen Importprozesse für KBL und VEC-Dateien in EPLAN harness proD gibt es?

Problemstellung

Die beiden Datenformate KBL und VEC sind durch prostep ivip und den Verband der Automobilindustrie standardisiert [3] [4] und werden von vielen Tools zur Kabelbaumkonstruktion sowie zum Datenaustausch zwischen verschiedenen Tools verwendet. Zuken E^3.cable beispielsweise besitzt bereits eine Möglichkeit KBL Daten einzulesen und entsprechende Fertigungsdokumente daraus zu erzeugen. In EPLAN harness proD ist keine solche Schnittstelle vorhanden und nach Rücksprache mit dem EPLAN Produktmanagement auch in den kommenden Jahren nicht geplant. Das relativ neue Datenformat VEC wird bisher nur von wenigen Tools unterstützt. Eines der Tools, die bereits eine Unterstützung für das VEC Format besitzt ist PREEvision von der Firma Vector Informatik GmbH [5]. In der Vergangenheit haben bereits einige der Kunden der Partnerfirma M-H engineering GmbH & Co.KG Verkabelungen mit EPLAN harness proD entworfen und das Konstruktionstool in ihre Toolchain für den Bordnetzentwicklungsprozess integriert. EPLAN harness proD bietet, auch aufgrund der vorhandenen API, viele Möglichkeiten Daten in verschiedenen Formaten und Ausgestaltungen für die Fertigung zu exportieren. Gerade im Hinblick auf eine automatisierte oder gar automatische Fertigung von Leitungssträngen sind die Exportformate von Bordnetz Tools sehr wichtig. Die API soll genutzt werden um eine Schnittstelle für KBL- und VEC-Daten in EPLAN harness proD zu verarbeiten und dadurch das volle Potential von EPLAN harness proD auszuschöpfen.

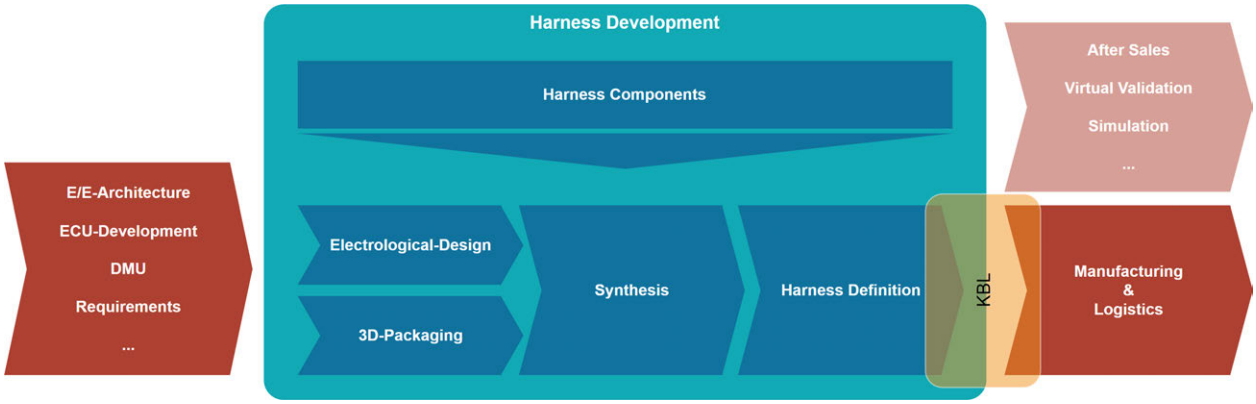


Abb. 2: Umfang von KBL und VEC [1]

Lösungsansatz

Eine Interaktion mit EPLAN harness proD ist durch die Nutzung der API möglich. Die Anbindung kann entweder per Plugin in der Anwendung oder als Stand Alone Applikation, welche lediglich die API aufruft erfolgen. Bei der Auswahl muss beachtet werden, was später der konkrete Anwendungsfall des Importers ist. Entweder es handelt sich um eine Konstruktionserweiterung mit deren Hilfe die Konstruktion beschleunigt und vereinfacht werden kann oder die Erweiterung wird in der Arbeitsvorbereitung eingesetzt. In der Arbeitsvorbereitung kann nicht davon ausgegangen werden, dass EPLAN harness proD bereits zur Arbeit verwendet wird. Zum aktuellen Bearbeitungsstand der Bacherlortarbeit lassen sich die Forschungsfragen bereits teilweise beantworten. Als Anforderung für ein Bordnetzkonstruktionstool sind definitiv die beiden Datenformate KBL und VEC aufzuführen. Die beiden Formate sind standardisiert und schaffen die Möglichkeit auch im hart umkämpften Automobilbereich wettbewerbsfähig zu bleiben. Durch bereitstellen der Möglichkeit zum Import von KBL-Daten in EPLAN harness proD ist es möglich das Konstruktionstool auch für OEM Problemstellungen zu verwenden.

Ausblick

Der in der Bearbeitung der Thesis entstandene beispielhaft implementierte Importer für die Datenformate KBL und VEC in EPLAN harness proD soll im Anschluss an die Arbeit weiter ausgearbeitet und gemeinsam mit den Partnern im Forschungsprojekt KI4BoardNet verfeinert werden um die Anforderungen im Bereich der Fertigung von Kabelsträngen im OEM Umfeld zu entsprechen. In Zusammenarbeit mit unserem Fertigungspartner Kemmler Electronic GmbH soll der Importer als Erweiterung für EPLAN harness proD auch um einen Exporter ergänzt werden um die Daten aus EPLAN harness proD auch in bestehende Fertigungshierarchien einzubringen.

Literatur und Abbildungen

[1] Johannes Becker. Whitepaper KBL vs. VEC. https://ecad-wiki.prostep.org/post/kbl-vs-vec/whitepaper-kbl-vs-vec.pdf, 08 2022.

[2] Eigene Darstellung.

[3] Verband der Automobilindustrie prostep ivip. prostep ivip VDA Recommendation Harness Description List (KBL) Version 2.5. https://www.ps-ent-2023.de/fileadmin/prod-download/PSI_VDA_Recommendation_4964_KBL_EN_0819.pdf, 09 2018.

[4] Verband der Automobilindustrie prostep ivip. prostep ivip / VDA Recommendation Vehicle Electric Container (VEC) Version 1.2. https://www.ps-ent-2023.de/fileadmin/prod-download/PSI_21_vda_4968_VEC_Specification_v1.2_pub_RZ.pdf, 06 2020.

[5] Bum Jin Yun et al. The Future of Automotive E/E Development is Model-Based. https://www.vector.com/de/de/download/the-future-of-automotive-e-e-development-is-model-based/, 12 2018.

Auswirkungen von Datensatz-Modifikationen in maschinellen Lernprozessen: Eine Studie zu Lernraten und zur Klassifizierungsgenauigkeit mittels Data Labeling, Data Augmentation und Web Scraping

Marcel Marek **Jürgen Koch**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma pep.digital GmbH, Esslingen

Motivation und Problemstellung

Data Augmentation und Web Scraping sind seit langem unerlässlich für die Datenverarbeitung in maschinellen Lernprozessen. In der Medizin wird Data Science mit großem Erfolg bei der Forschung eingesetzt. Auch in der Diagnostik können Herzerkrankungen präziser festgestellt und stationäre Eingriffe verhindert werden. Die Technik ermöglich es uns nicht dem Urteil einer Einzelperson vertrauen zu müssen, sondern viele Quellen zur Diagnose mit einbeziehen zu können. Akten aus Archiven oder Datenbündel von mehreren Krankenhäusern helfen eine Entscheidung zu finden. [5] Doch wie sieht es aus, wenn eine Krankheit nur selten auftritt? Wie können aus wenigen Datenpunkten neue Informationen gewonnen werden? Data Augmentation enthält Werkzeuge für die Generierung neuer Informationen. Der erste Untersuchungsgegenstand ist diese Form der Neugewinnung von Daten.
Für das Gewinnen neuer Daten wird eine weitere Technik eingesetzt: Das Web Scraping. Web Scraper greifen automatisiert auf Daten von Webseiten zu und bieten die Möglichkeit ohne manuellen Aufwand große Mengen an Daten zu suchen und abzuspeichern. Sofern Schnittstellen nicht vorhanden sind, können Web Scraper menschenlesbaren Text durchsuchen und in maschinenlesbaren Text umwandeln. [4] Die großen IT-Unternehmen und viele der kleineren Geschwister setzen seit Jahren auf das Web Scraping zur Datengewinnung. Nicht immer mit den besten Absichten für die Gesellschaft. Nach dem Cambridge Analytics Skandal sind der Allgemeinheit die weitreichenden Folgen von Big Data bekannt. [7] Das Unternehmen erhielt eine Klage der amerikanischen Finanzaufsicht alle persönlichen Daten zu löschen, welche von Facebook mit Scraping gewonnen wurden. [1] Doch wie funktioniert Data Scraping und was ist zu beachten? Welche Hindernisse gibt es bei der Gewinnung von neuen Bildern durch Web Scraping? In dieser Studie wird zur Vereinfachung des Untersuchungsgegenstands die Auswirkung auf Bilder beleuchtet.

Ziel der Arbeit

Das Ziel der Arbeit ist verschiedene Modifikationen auf einen ‘Computer Vision‘-Datensatz zu untersuchen. Der Computer Vision Datensatz enthält Bilder von zwei Gruppen, welche von einem Machine Learning Model kategorisiert werden. [2] Ein vortrainiertes Modell mit hoher Trefferquote wird zur Bewertung der Modifikationen gewählt. Die Studie versucht herauszufinden, wie eine möglichst hohe Trefferquote erreicht wird - trotz weitreichender Modifikationen.
Wie hoch ist die Trefferquote, wenn wir Bilder spiegeln oder auf andere Art modifizieren? Inwiefern eignet es sich ähnliche Bilder aus Online Quellen? Die Studie versucht Antworten auf diese und weitere Fragen zu finden.

Studienaufbau

Dieses Projekt erarbeitet zur Überprüfung der Lösungsansätze eine Plattform zur Kennzeichnung, Manipulation und Erweiterung von Bildern. Die modifizierten Datensätze werden mit einem maschinellen Lernprozess auf der Basis von tensorflow überprüft und die Ergebnisse in einer Tabelle zum Vergleich festgehalten. Zur besseren Strukturierung sind die Schritte im Folgenden in Datenerhebung, -modifikation, -erweiterung und -auswertung unterteilt.

Datenerhebung

Das Web Scraping benötigt beschriftete Daten für die Suche. Die Beschriftung von Bildern wird als ‘Tagging‘ bezeichnet. Personen außerhalb der Studie werden

zum Beschriften der Bilder herangezogen. Unabhängige Tags (ohne persönliche Einflüsse des Studienerstellers) sind das Ziel. Die Informationen werden in einer Datei gespeichert. Diese Datei beinhaltet die Tags aller Bilder einer Kategorie. Anschließend werden die sinnvollen Tags für die Suche selektiert. Der Web Scraper verwendet die vergebenen Tags und sucht die Seite mit den höchsten Übereinstimmung. Zuletzt werden die Bilder auf der Seite extrahiert und können dem Datensatz hinzugefügt werden.
Die anderen Bilder erhalten wir von alten Wettbewerben aus kaggle. Sie enthalten eine Bandbreite von Bildern aus dem Alltag und sind frei lizensiert zur Verfügung gestellt. Der gewählte Algorithmus konnte sich für diese Art der Klassifizierung von Bildern im Vorfeld auszeichnen.

Datenmodifikation

In der Studie werden geometrische Änderungen, farbliche Änderungen und Entfernen von Bildinformationen untersucht. Dazu gehören Modifikationen wie Spiegelung, Rotation, Größenänderung, Helligkeitsanpassung, Maskierung, Extraktion von Bildinhalten oder Verpixelung von Bildbereichen. [3] Für unsere Studie werden die Bilder im Format 224x224 Pixel benötigt.

Datenerweiterung

Ein Web Scraper sucht zur Datenerweiterung neue Bilder auf vorbestimmten Internetseiten. Die vorher vergebenen Beschriftungen der Bilder (siehe Datenerhebung) werden als Suchbegriff verwendet. Das Web Scraping erfolgt mit dem Framework Scrapy. Scrapy bietet eine umfangreiche Dokumentation und intuitive Methoden.

Datenauswertung

Für die Auswertung der Daten wird Tensorflows keras mit einem vortrainierten Modell verwendet. Kaggle bietet auch hier verschiedene Wettkämpfe über die Thematik Klassifizierung von Bildern an. Die Benchmarks der Wettkämpfe zeigen den ausgewählten Algorithmus als robust. Der Algorithmus VGG-16. VGG-16 konnte sich beim ILSVRC-Wettbewerb (ImageNet Large Scale Visual Recognition Challenge) auszeichnen und ist beim maschinellen Lernen weit verbreitet. [6] VGG-16 erzielt eine hohe Genauigkeit bei der Klassifizierung. Dabei sollte erwähnt werden, dass in den letzten Jahren bessere Algorithmen erarbeitet wurden. Ein Beispiel für eine noch höhere Genauigkeit wäre Resnet, InceptionNet oder XceptionNet. Diese sind in ihrer Funktionsweise jedoch komplizierter und in Folge dessen schwerer nachzuvollziehen. Die konsequente Verwendung von 3x3-Filter bei VGG-16 erleichtert das Verständnis und ermöglicht uns eine Nachvollziehbarkeit der Ergebnisse unserer Studie. VGG-16 verwendet außerdem einen Stride von 1 und analysiert somit jegliche Pixel im Bild und verwendet keine Abstraktion der Daten. Die Modifikationen haben somit einen direkten Einfluss auf die Genauigkeit des Modells. Dies sichert die Korrektheit unserer Ergebnisse. Dem Datensatz werden pro Iteration 50 Bilder entzogen und durch manipulierte oder gescrapte Daten ersetzt. Die Ergebnisse der einzelnen Modifikationen werden miteinander verglichen und final festgestellt, welche Methode am besten geeignet ist.

Ergebnisse

Zum Zeitpunkt des Artikels sind noch keine finalen Studienergebnisse vorhanden.

Ausblick

Die erstellten Datensätze können mit anderen Algorithmen wie ResNet, InceptionNet oder XceptionNet untersucht werden. Wie ist die Leistung der manipulierten Daten mit anderen Modellen?
Ein weiterer interessanter Aspekt ist die ethische und gesellschaftliche Bewertung. Insbesondere die Auswirkungen auf den Datenschutz werden einen größeren Raum in zukünftigen Projekten einnehmen.

Literatur und Abbildungen

[1] Team FAZ. Meta beendet mit 725 Millionen-Zahlung Rechtsstreit. https://www.faz.net/aktuell/wirtschaft/cambridge-analytica-meta-beendet-mit-725-millionen-zahlung-rechtsstreit-18554784.html, 2022.

[2] Mark E. Fenner. *Machine Learning with Python for Everyone*. Pearson, 2020.

[3] Duc Haba. *Data Augmentation with Python*. Packt Publishing Ltd., 2023.

[4] Ryan Mitchell. *Web Scraping with Python: Collecting More Data from the Modern Web*. Springer-Verlag, 2018.

[5] Annalyn Ng and Kenneth Soo. *Data Science - was ist das eigentlich?!* Springer-Verlag, 2018.

[6] Olga Russakovsky and Jia Deng. ImageNet Large Scale Visual Recognition Challenge. https://arxiv.org/pdf/1409.0575.pdf, 2015.

[7] Team Tagesschau. Facebook Datenskandal - Bis zu 87 Millionen Nutzer betroffen. https://www.tagesschau.de/ausland/facebook-ausweitung-skandal-101.html, 2018.

Unternehmerische Strategien für den urheberrechtlichen Umgang mit generativen KI-Trainingsdaten

Alexander Masen **Anke Bez**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma MHP Management- und IT-Beratung GmbH, Ludwigsburg

Einleitung

Generative Künstliche Intelligenz (KI) wird in Unternehmen in jeder Branche zu einem immer relevanteren Thema. [2] Mithilfe der generativen KI können leicht Inhalte erstellt werden und der Arbeitsalltag kann dadurch deutlich vereinfacht werden. Zudem kann die Produktivität durch den Einsatz der KI-Systeme erheblich gesteigert werden. [1] Die Funktionsweise und die Qualität der generativen KI-Tools hängt jedoch stark von der Qualität und der Menge der verwendeten Trainingsdaten ab, mit denen das System trainiert wurde. Ein wichtiger Faktor, der bei den Trainingsdaten zu beachten ist, ist das Urheberrecht. Um rechtliche Probleme und Auseinandersetzungen zu vermeiden, ist es wichtig, die Herkunft der Trainingsdaten zu kennen und zu überprüfen, ob man die Nutzungsrechte dieser Daten hat.

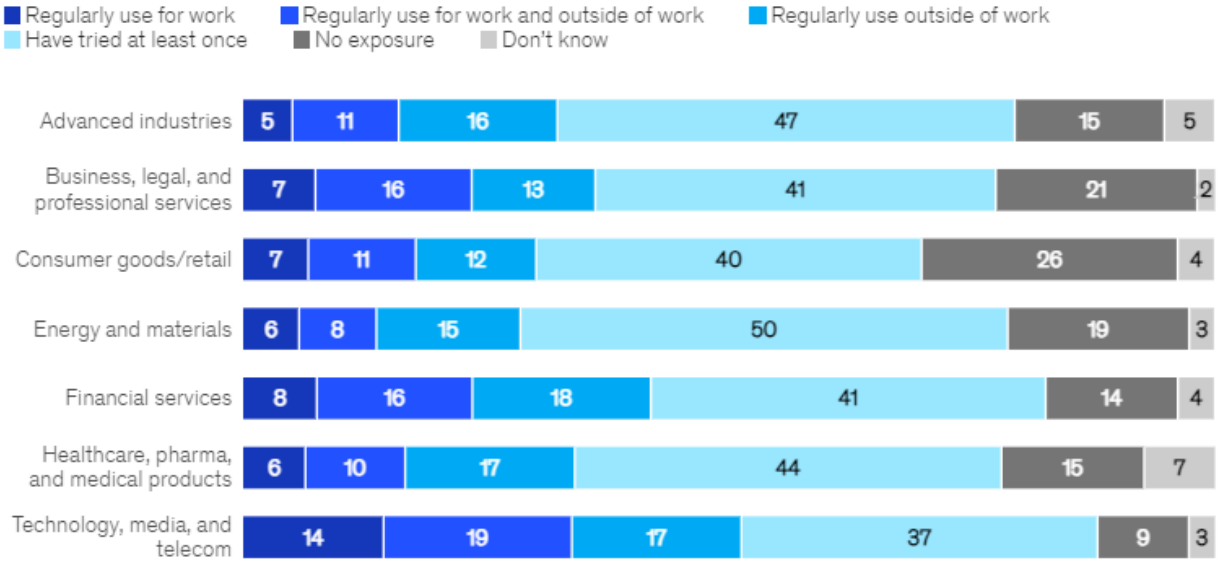


Abb. 1: Umfrage über den Umgang mit generativen KI-Tools nach Branche 2023 [2]

Problemstellung und Zielsetzung

Eine Problematik für Unternehmen ist, dass sie nicht genau nachvollziehen können, mit welchen Daten die KI-Toolanbieter ihre KI-Tools trainieren. Dennoch schreiben die Anbieter in ihren Nutzungsbedingungen, dass der Nutzer selbst für den von der KI erstellten Output verantwortlich ist und dafür haftet. Damit gehen die Unternehmen bei der Nutzung der KI-Tools ein Risiko ein. In dieser Arbeit werden die Herausforderungen, die sich im Zusammenhang von generativen KI-Trainingsdaten und dem Urheberrecht ergeben, erläutert und analysiert. Dazu werden die Strategien der KI-Toolanbieter näher betrachtet, um herauszufinden, wie sie Urheberrechtsverletzungen vermeiden. Zudem werden diese Strategien dann aus Sicht der Anwender betrachtet und in Risikobewältigungsstrategien eingeordnet. Auf Basis dieser Einordnung werden dann Empfehlungen an Unternehmen ausgesprochen, wie mit den urheberrechtlichen Herausforderungen

umgegangen werden kann.

Definitionen und bisherige Erkenntnisse

Generative KI und Trainingsdaten

Die generative künstliche Intelligenz ist eine Art der KI, die die Fähigkeit besitzt, auf Basis von vorhandenen Informationen und Vorgaben des Benutzers neue Inhalte zu erstellen. Diese Inhalte können Texte, Programmcodes, Bilder, Videos oder Audiodateien sein. Bei der generativen KI sollen aus den Trainingsdaten Muster und Merkmale extrahiert werden, um aus diesen Strukturen etwas Neues zu generieren. Der neu generierte Inhalt hat die gleichen Strukturen wie die Trainingsdaten, existiert aber noch nicht. Trainingsdaten sind die Daten, mit dem das KI-System lernt, wie Informationen und Daten verarbeitet werden sollen. Die Qualität der Trainingsdaten spielt eine große Rolle in der Entwicklung der KI-Systeme und ist ausschlaggebend für die gewünschte Funktionsweise. Zusätzlich ist die Herkunft der Daten ein wichtiger Faktor. Wenn die Herkunft der Trainingsdaten abgesichert ist, kann man rechtliche Probleme schon im Voraus beseitigen. [4], [6]

Urheberrechtliche Aspekte in Bezug auf Trainingsdaten

Es gibt nur zwei urheberrechtlich relevante Schritte im maschinellen Lernprozess in Bezug auf die Trainingsdaten. Der erste Schritt befasst sich mit dem Sammeln und Organisieren der Daten. Nachdem die Datensammlung abgeschlossen ist, kommt es zum zweiten Schritt. Dieser besteht aus dem Einlesevorgang und der Aufbereitung sowie Normalisierung der Trainingsdaten. In diesem Schritt werden die Daten direkt eingelesen oder in ein maschinenlesbares Format gebracht. Die darauffolgenden Analysen und Auswertungen der Daten sind keine Handlungen, die vom Urheberrecht erfasst werden, da die Informationen in diesem Schritt nur aufgerufen und ausgelesen werden. Wenn urheberrechtlich geschützte Werke genutzt werden, entsteht jedoch schon im ersten Schritt durch das Sammeln und Organisieren der Daten eine Verletzung des Urheberrechts. Die Datensätze erfordern meist ein automatisches massenhaftes Speichern der Daten. Dieses Speichern stellt gemäß § 16 Urheberrechtsgesetz (UrhG) eine Vervielfältigung dar. Auch im Rahmen des zweiten Schrittes könnte gegen das Urheberrecht verstoßen werden. Im Zuge des Einleseprozesses werden die Daten in den Arbeitsspeicher der KI-Systeme gespeichert und werden dann in eine Repräsentation umgewandelt, die der Computer lesen kann. Zudem werden die Daten oft längerfristig gespeichert, um die Daten für das Training anzupassen. Die Vervielfältigung und die Bearbeitung eines Werkes steht nach den §§ 16, 23 UrhG nur dem Urheber zu. [3], [5]

Ausblick

Teil dieser Arbeit ist es, nach der Analyse der Strategien der KI-Toolanbieter und dem Einordnen dieser aus Sicht der Anwender, Experteninterviews durchzuführen. Diese Interviews sollen die Sicht der Anwender widerspiegeln und deutlich machen, wie mit den KI-Tools in Unternehmen und mit dem Risiko des Urheberrechts umgegangen wird.

Literatur und Abbildungen

[1] KPMG AG. Generative künstliche Intelligenz als Schlüssel zu... https://kpmg.com/de/de/home/themen/uebersicht/generative-kuenstliche-intelligenz.html, 08 2023.

[2] McKinsey and Company. The state of AI in 2023. https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-in-2023-generative-ais-breakout-year, 08 2023.

[3] Bert Eichhorn et al. *Internetrecht im E-Commerce*. Springer Vieweg, 2016.

[4] David Foster. *Generatives Deep Learning*. O'Reilly, 2020.

[5] Lisa Käde. *Kreative Maschinen im Urheberrecht*. Nomos Verlagsgesellschaft, 2021.

[6] Stefan Luber. Was ist generative AI. https://www.bigdata-insider.de/was-ist-generative-ai-a-2ec9ecd5c114d4c94c48ea7092ec45ad/, 05 2023.

Multimodal Deep Learning for Product Matching

Kazim Ali Mazhar **Gabriele Gühring**

Department of Computer Science and Engineering, Esslingen University

Work carried out at Parsionate GmbH, Stuttgart

Motivation

Product matching is the process of finding and comparing similar products across different online platforms. This is essential for both vendors and customers in the e-commerce domain, as online shopping becomes more popular and competitive. However, product matching faces many difficulties, such as the lack of standard identifiers, the inconsistency of product information, and the vagueness of customer queries.
One of the main challenges is that products do not have common identifiers, such as EAN or GTIN, making it difficult to match them based on their IDs. This may be intentional or accidental, but it results in incomplete or noisy data. Moreover, customers often want to use minimal or vague inputs, but expect precise outputs.

Description	Image
Einbauherd-Set EEK A PKM BIC3-I GK IX-2 schwarz	
Schwerlast Steckregal EASYmaxx verzinkt 1800x900x400 mm 5 Böden Tragkraft 875 kg	

Fig. 1: Exemplary DIY product descriptions and images [11]

Objective

This thesis aims to leverage and expand on the advances and milestones achieved by Mazhar et al. [8] and Falzone et al. [2] in the field of multimodal deep learning and product matching. The main objective is to prove whether or not multimodal deep learning techniques bring about an performance improvement in DIY product matching tasks, compared to unimodal methods.
In order to solve this objective, an experiment is conducted: the multimodal model by Mazhar et al. [8] and the character-level approach by Falzone et al. [2] are applied to a real-world HORNBACH DIY dataset [5] provided by Parsionate through web crawling. Also similar to Mazhar et al. [8], this thesis consists of benchmark experiments against existing unimodal methods.
First, the necessary hardware and software environment is set up on the bwUniCluster2.0 [12]. Next, the appropriate model by Mazhar et al. [8] is chosen, upon which a grid search is performed to find the optimal hyperparameters. Since in this thesis multiple image and text components are employed, ablation studies for both the image and text side are conducted in order to investigate the impact of each side on the model performance. Lastly, the model is evaluated against a unimodal model using the QuickLabel App provided by Parsionate.

Multimodal Deep Learning

Multimodal deep learning in the context of product matching is defined by Mazhar et al. [8] as follows:
"According to Liu et al. [7], the underlying idea of multimodal deep learning is that signals from different modalities often complement each other. Therefore, it can be concluded that combining modalities allows for a more robust inference [7], [3], [1]. [...] Similarly, in this work we combine text queries with image data to increase the quality of search queries and thus enable new use cases. Particularly, we focus on non human-readable product descriptions which entails IDs, codes and numbers associated with a product [2]."

Dataset

The dataset central to this thesis is provided by Parsionate via crawling of the online shop of HORNBACH [5]. The dataset consists of 156,035 products in JavaScript Object Notation (JSON) format. On the text side, there are 3,107 unique *attribute fields* which contain any information from color, material, size, dimensions up to highly specific details such as finish, boiler impregnation, clearance height et cetera. On the image side, each product has a *primary image*, which

is displayed first on the product page, and one or more *secondary* or *alternative images* which can show the product from different angles, with background scenery or display additional information about the product.
The dataset is then prepared for use with the multimodal model by Mazhar et al. [8]: first the relevant features of the dataset are selected, then the target variable and number of classes is determined. Following that, the dataset is augmented via alternative product images, which also requires duplicate elimination beforehand and undersampling afterwards. The class distribution after augmentation and undersampling can be seen in Figure 2. Finally, the product images and descriptions are merged, and labelling and final postprocessing steps are performed.

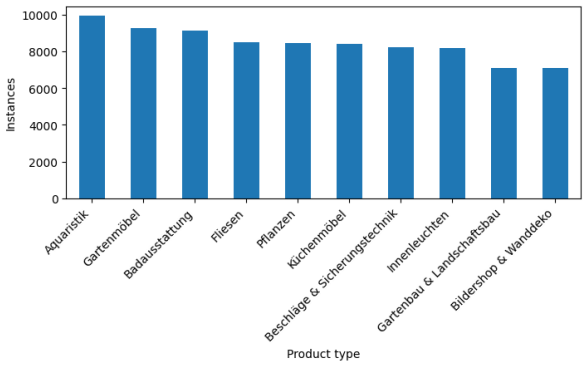
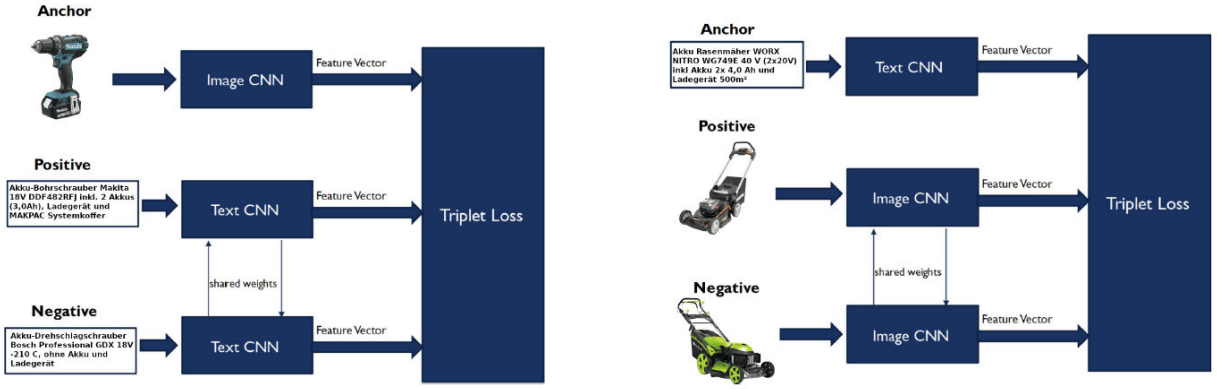


Fig. 2: Top 10 class distribution after undersampling [11]

Model



(a) Image anchor (b) Text anchor

Fig. 3: "Naive" MNN-BTL architecture. Adapted from Mazhar et al. [8]

The model chosen for training and evaluation on the HORNBACH dataset is the Multimodal Neural Network with Bidirectional Triplet Loss (MNN-BTL) by Mazhar et al. [8], a model which uses the bidirectional triplet loss function which maps image and text data into a common embedding space. In their work, Mazhar et al. further explain their model architecture: "The network is trained with a bidirectional triplet loss, so that correct image-text pairs in the embedding space have a smaller distance to each other than incorrect ones. *Bidirectional* refers to the bilateral affiliation of the modalities image and text, in contrast to *unidirectional*, i.e., a unidirectional association of an image to a text and vice versa. Since we use the bidirectional triplet loss, our MNN-BTL requires two unidirectional components, [...] shown in [Figure 3]."
As a backbone, the adopted MNN-BTL primarily uses the Character-Level Convolutional Neural Network (CLCNN) by Zhang et al. [14] on the text side, while on the image side the ResNet50 by He et al. [4] is used. Since the model chosen in this thesis is directly adopted from the MNN-BTL by Mazhar et al. [8], which in turn is largely inspired by the *embedding network* by Wang et al. [13], the model architecture also remains the same. However, when looking at the model implementation, the differences become apparent. One major adjustment, in particular, is the choice of *semi-hard* instead of *hard* triplet mining in this thesis, which is found to be the only feasible triplet mining strategy for the HORNBACH dataset.
For the MNN-BTL, Mazhar et al. [8] obtain the triplet loss by calculating the distance between anchor and positive as well as anchor and negative. Cosine distance is used as the distance metric. For text-to-image comparisons, the unidirectional triplet loss L_1 is computed as max of zero and the difference between positive and negative distances, with a margin α added on top. Similarly, for image-to-text comparisons, L_2 is calculated in the same way. Finally, the bidirectional triplet loss L_{bi} is the combination of L_1, L_2 with

regularization parameters λ_1, λ_2 [8].

Evaluation using QuickLabel App

After obtaining the optimal hyperparameters from the grid search and benchmarking different image and text components in the ablation studies, the MNN-BTL is finally evaluated using the QuickLabel App provided by Parsionate. The app itself provides, as the name suggests, a general-purpose framework for labelling tasks. For the evaluation in this thesis, the data to be labeled is generated from the following product categories in the HORNBACH dataset:

1. *Gartenmaschinen & Forstbedarf*
2. *Maschinen*
3. *Maschinenzubehör*
4. *Handwerkzeug*
5. *Werkstatteinrichtung*

For each of the five categories, 20 reference articles are obtained using a Principal Component Analysis (PCA) and clustering algorithm on the ResNet50 image embeddings. For all 20 reference articles per category, the top 5 similar articles are then obtained. For the MNN-BTL, this is done by building a pairwise cosine distance matrix over the concatenated image-text embeddings of all products for each of the five categories previously described. The top 5 most similar articles correspond to the top 5 embeddings in the matrix which are the nearest to the reference article. The model which the MNN-BTL is evaluated against is a unimodal first-image similarity (FI) model.
In this thesis, two raters R_1, R_2 indepedently label the data. The results of both raters are then compared side-by-side, in order to determine similarities or deviations in labelling patterns.

Results

The two-phase approach of the grid search allows for the initially huge parameter grid to be systematically and gradually narrowed down, which significantly streamlines the grid search:

- For the batch size, the value must be neither too small nor too big, due to the hardware constraints and choice of *semi-hard* triplet mining. Only the batch sizes 256 and 512 have been found to be working.
- For lr, values in low orders of magnitude, such as 5e-05, perform the best, as higher lr tend to overshoot often.
- Similarly, for the margin α, small values such as $\alpha = 0.1$ or $\alpha = 0.2$ work best. However, the choice of α is largely dependent on the underlying embedding distribution in the HORNBACH dataset.
- For λ_1, λ_2 the model performs best for $\lambda_1 \leq \lambda_2$, when the image side is given more weight than the text side.

The results for the ablation studies demonstrate that the *default* MNN-BTL configuration, utilizing CLCNN and ResNet50, consistently outperforms other configurations involving different text and image components. Notably, MobilenetV3large and MobilenetV3small [6] rank 3rd and 5th in image ablations (out of 143), while Word2Vec [9] and GloVe [10] secure 2nd and 4th positions in text ablations (out of 69), closely competing with the top-performing CLCNN. Despite CLCNN currently being the top performer, the evolving landscape of German corpora suggests that word embedding methods like Word2Vec and GloVe may eventually surpass CLCNN in performance.
The results of the QuickLabel App evaluation demonstrate that the embeddings generated by the MNN-BTL are more accurate than the FI model, which is a unimodal method. Therefore, it is proven that multimodal deep learning techniques do bring about an performance improvement in DIY product matching tasks.
Additionally, the QuickLabel results further support the numerous successful results of experiments in other works about multimodal deep learning for product matching, including [8], [1], [3]. Furthermore, *product sets* or *all-in-one* products in the reference articles are analyzed, which the MNN-BTL struggles the most with due to ambiguity introduced by multiple products in a single instance.

Outlook

The research on multimodal deep learning for product matching presented in this thesis suggests avenues for improvement. Since this thesis focuses on combining human-readable and non human-readable text, similar to Falzone et al. [2] and Mazhar et al. [8], it is important to investigate the impact of the human-readable and non human-readable parts on the model performance separately. Additionally, exploring how the CLCNN handles word boundaries or whitespaces and optimizing the number of selected attributes in the dataset are crucial. The results should be published to reach a broader academic audience, with experiments on other datasets to demonstrate generalizability. Special attention is needed on the choice of triplet mining strategy, and a second publication could address misconfigurations in Mazhar et al. [8] and contribute to improved practices in multimodal deep learning.

References and figures

[1] Raúl Estrada-Valenciano, Víctor Muñiz-Sánchez, and Héctor De-la Torre-Gutiérrez. An Entity-Matching System Based on Multimodal Data for Two Major E-Commerce Stores in Mexico. *Special Issue*, 2022.

[2] Simone Falzone, Tobias Münster, and Gabriele Gühring. Measuring similarity for technical product descriptions with a character-level siamese neural network. In *Tagungsband zum Workshop der Multiprojekt-Chip-Gruppe Baden-Württemberg*, volume 63. Technische Hochschule Ulm, 2022.

[3] Ketki Gupte, Linsey Pang, Harshada Vuyyuri, and Sujitha Pasumarty. Multimodal Product Matching and Category Mapping: Text+Image based Deep Neural Network. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 4500–4505. IEEE, 2021.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2016.

[5] HORNBACH Holding AG Co KGaA. Unternehmen — hornbach-holding.de. https://www.hornbach-holding.de/unternehmen/, 2023.

[6] Andrew Howard et al. Searching for MobileNetV3. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 2019.

[7] Kuan Liu, Yanen Li, Ning Xu, and Prem Natarajan. Learn to Combine Modalities in Multimodal Deep Learning. https://arxiv.org/abs/1805.11730, 2018.

[8] Kazim Ali Mazhar, Matthias Brodtbeck, and Gabriele Gühring. Similarity learning of product descriptions and images using multimodal neural networks. *Natural Language Processing Journal*, 4:100029, 2023.

[9] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. https://arxiv.org/abs/1301.3781, 2013.

[10] Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543. Association for Computational Linguistics, 2014.

[11] Own representation.

[12] Karlsruhe Institute of Technology. bwUniCluster 2.0+GFB-HPC. https://www.scc.kit.edu/en/services/bwUniCluster_2.0.php, 2023.

[13] Liwei Wang, Yin Li, Jing Huang, and Svetlana Lazebnik. Learning Two-Branch Neural Networks for Image-Text Matching Tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41:394–407, 2019.

[14] Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level Convolutional Networks for Text Classification. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

Analyse der Intercom-Kommunikation im Mannschaftssporttraining: Einfluss auf Spielerperformance und Trainer-Coaching

Japhet Maleka Mbala **Catharina Kriegbaum-Kling**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma Coachwhisperer GmbH, Jena

Einleitung

Diese Bachelorarbeit widmet sich der gründlichen Analyse der Intercom-Kommunikation im Kontext des Mannschaftssporttrainings und deren Auswirkungen auf die Leistung der Spieler und die Coaching-Methoden der Trainer. Ein besonders aufschlussreicher Forschungsversuch wurde in Kooperation mit dem American-Football-Team Leonberg Alligators durchgeführt, um praxisnahe Erkenntnisse zu gewinnen.

Zielsetzung

Diese Bachelorarbeit analysiert die Auswirkungen des Coachwhisperer-Systems (siehe Abbildung 1) auf die Spielerperformance und das Trainer-Coaching im American Football. Der theoretische Teil beleuchtet die Entwicklung von Smart Gear und mobilen Technologien im Sport, insbesondere im American Football, sowie die Rolle der Informatik bei der Datenanalyse von Leistungs- und taktischen Informationen. Der empirische Teil konzentriert sich auf eine umfassende Analyse von Spielerleistung und Coaching unter Verwendung des Coachwhisperer-Systems.
Durch die Erfassung qualitativer und quantitativer Daten werden die Effektivität des Systems und mögliche Lerneffekte bewertet. Die Erkenntnisse tragen dazu bei, die Einsatzmöglichkeiten und Auswirkungen von Kommunikationstechnologien im American Football zu verstehen und bieten eine Grundlage für die zukünftige Nutzung von Smart Gear im Sport.

Coachwhisperer-System

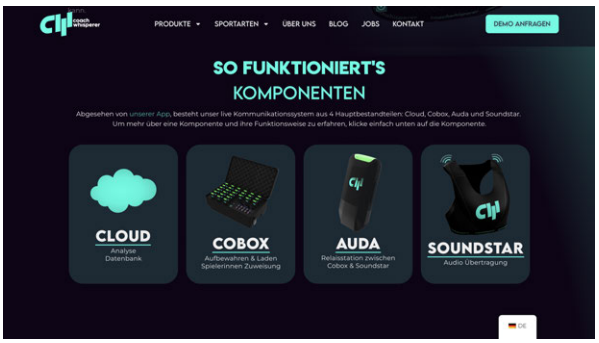


Abb. 1: Coachwhisperer Soundweste - SSoundstar-[2]

Das Coachwhisperer Intercom-System besteht aus vier Schlüsselkomponenten (siehe Abbildung 1) – Cloud, Cobox, Auda und Soundstar –, die in einem sicheren Zusammenspiel agieren. Die Cloud speichert Spielerinformationen unter Verwendung fortschrittlicher Sicherheitsstandards. Die Cobox ermöglicht die Integration mehrerer Teams mit eindeutigen Club-IDs, während das Wearable (Auda) als Schnittstelle zwischen Cobox und Soundweste dient. Diese präzise Struktur gewährleistet eine effektive Live-Kommunikation im Training und bildet die Basis für die empirische Untersuchung dieser Bachelorarbeit.

Datenanalyse im Mannschaftssport

Die Datenanalyse im Sport hat sich zu einem unverzichtbaren Element entwickelt, das weitreichende Implikationen für Athleten, Trainer und die gesamte Sportbranche hat. In einer Zeit, in der wir mit einer beispiellosen Menge an Daten konfrontiert sind, ist es entscheidend, relevante Informationen zu identifizieren und sorgfältig zu interpretieren. Die Transformation von Rohdaten zu sinnvollem Wissen,

wie im "Data-to-Wisdom-Labyrinthsïllustriert (siehe Abbildung 2), ermöglicht es Sportlern, ihre Leistungen zu verstehen und zu optimieren [8]. Das "Data-to-Wisdom-Labyrinth"durchläuft mehrere aufeinander aufbauende Schritte: Zunächst werden Rohdaten gesammelt (Data), die dann in einen strukturierten Kontext gebracht werden, um Informationen zu generieren (Information). Durch die Analyse und Interpretation dieser Informationen entsteht Wissen (Knowledge), das wiederum in Verbindung mit Erfahrung und Kontext zu tiefen Einblicken (Insight) führt. Schließlich führt die Anwendung dieser Erkenntnisse in praktischen Situationen zur Entwicklung von Weisheit (Wisdom) im Sporttraining.

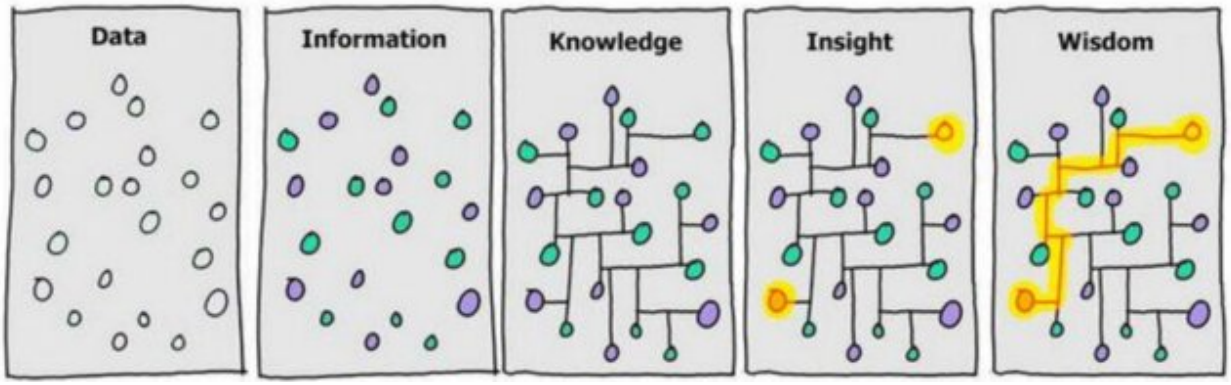


Abb. 2: The data to wisdom network. [8]

Nate Silver betont, dass Datenanalyse eine aktive Rolle erfordert, da Zahlen nicht für sich selbst sprechen können [7]. Ein historisches Beispiel ist Bill James, der in den 1970er Jahren mit SABRmetrics die Baseballanalyse revolutionierte. Die Erforschung von Sport-Big-Data ist nicht nur im Leistungssport, sondern auch im Breiten- und Schulsport von großem Nutzen.

Key Performance Indicators (KPIs) sind im Sport, insbesondere in der quantitativen Spielanalyse, von zentraler Bedeutung. Sie messen kritische Erfolgsfaktoren und dienen als unverzichtbare Werkzeuge für Trainer, Analysten und Scouts, um fundierte Entscheidungen zu treffen und einen Wettbewerbsvorteil zu erlangen [5].

Das Berufsbild des Data Analyst erlebt einen Aufschwung, insbesondere in der Spielanalyse. Die Integration von automatischen Tracking-Daten und manuellen Daten stellt jedoch eine Herausforderung dar. Die qualitative Analyse anhand von Videodaten bleibt dominierend, während Modelle wie "Ghostingëntwickelt werden, um beide Datensätze effektiver zu integrieren.

In der NFL hat die Datenanalyse einen enormen Einfluss, unterstützt durch Initiativen wie den Big Data Bowl. Die Next Gen Stats-Initiative integriert Tracking-Chips in Spielerpads und Bälle, ermöglicht durch die Partnerschaft mit Zebra Technologies Corporation [3]. Diese Technologie revolutioniert nicht nur die Spielerleistung, sondern beeinflusst auch Entscheidungsprozesse und fördert Innovationen in der Liga.

Die NFL veröffentlichte bis 2018 diese wertvollen Tracking-Daten für jedes Team, was umfassende Analysen ermöglichte. Die Next Gen Stats bieten nicht nur einen Einblick in die Spielerleistung, sondern schaffen auch eine neue Ebene der Transparenz und Interaktivität für Fans und Analysten [3]. Insgesamt zeigt die fortschreitende Datenanalyse im Sport, insbesondere in der NFL, wie Daten dazu beitragen, Leistung zu optimieren, innovative Ansätze zu fördern und einen Wettbewerbsvorteil zu erlangen.

Wearable Computing und Intercom-Systeme im Sport

Die Welt des Wearable Computing hat in den letzten Jahren eine bedeutende Transformation erlebt und Wearables sind nun integraler Bestandteil von IoT-Systemen. Diese Technologie etabliert sich zunehmend als essentielle Komponente im Sport, indem sie Sensorik und Datenanalyse miteinander verknüpft [1]. Wearables, in Kleidung oder Brillen integriert, ermöglichen eine unauffällige und minimale Belastung für den Anwender, was ihre Akzeptanz vorantreibt. Von der Bewegungserfassung bis zur IoT-Integration eröffnet Wearable Computing neue Perspektiven für die Sportanalyse und Leistungsoptimierung [1].

Im Fokus steht die kontinuierliche und präzise Erfassung von lebenswichtigen Daten, etwa durch Trägheits-Wearables zur Validierung von Gangmodellen im medizinischen Bereich. Die Verbindung von IoT schafft einen digitalen Sportler, der Daten, Simulationen und präzise Vorhersagen in die Bewertung und Überwachung der sportlichen Leistung integriert

[1]. Tragbare Sensoren, von Bewegungserfassung bis zu verschiedenen Tracking-Technologien, spielen eine entscheidende Rolle in der Überwachung und Analyse der sportlichen Leistung, über verschiedene Sportarten hinweg [1].
Intercom-Systeme, auch als Sprechanlagen bekannt, sind Kommunikationsmittel, die elektrische Signale zur Sprachübermittlung nutzen. Ursprünglich in sicherheitsrelevanten Bereichen eingesetzt, haben sie sich auch im Hochleistungssport etabliert [4]. Diese Systeme finden Anwendung in Veranstaltungstechnik, Flugzeugen und sogar Unterwasserkommunikation. Im Sport, insbesondere in der NFL, spielten elektronische Kommunikationssysteme zwischen Trainern und Spielern eine bedeutende Rolle. Von der heimlichen Anwendung in den 1950er Jahren bis zur heutigen digitalen Kommunikationstransformation haben Intercom-Systeme die Effizienz des Spiels verbessert und die Kommunikation zwischen Teammitgliedern optimiert [6].
Die Kombination von Wearable Computing und Intercom-Systemen zeigt, wie Technologien im Sportbereich immer tiefer in die Leistungsoptimierung eindringen. Von der individuellen Bewegungserfassung bis zur teamweiten digitalen Kommunikation haben diese Innovationen das Potenzial, den Sport auf allen Ebenen zu revolutionieren und die Grenzen der Leistungsfähigkeit zu erweitern.

Schlussfolgerung und Ausblick

Die Zusammenführung von Datenanalyse, Wearable Computing und Intercom-Kommunikation im Kontext des Mannschaftssporttrainings zeigt vielversprechende Perspektiven für die zukünftige Gestaltung von Trainingseinheiten und die Verbesserung der Spielerperformance. Die vorliegende Bachelorarbeit bietet einen Einblick in die Potenziale von Kommunikationstechnologien im American Football-Training, insbesondere durch das Coachwhisperer-System. Zukünftige Forschungen können sich darauf konzentrieren, diese Technologien weiterzuentwickeln, um spezifischere Leistungsdaten zu generieren. Der empirische Teil vertieft die Analyse von Spielerperformance und Coaching-Effektivität, mit einem Fokus auf mögliche Lerneffekte. Die Erkenntnisse bilden die Grundlage für die optimierte Nutzung von Smart Gear im Sport und könnten wegweisend für andere Teams und Sportarten sein, die moderne Kommunikationstechnologien im Training implementieren möchten.

Literatur und Abbildungen

[1] V. Camomilla et al. Trends supporting the in-field use of wearable inertial sensors for sport performance evaluation: A systematic review. *Sensors*, 2018.

[2] Eigene Darstellung.

[3] A. Ennis. An Introduction to Modeling NFL Tracking Data. *Journal*, 2020.

[4] P. Furley. *Kommunikation von Spielanalysedaten. In Spielanalyse im Sportspiel*. Springer, 2022.

[5] M. D. Hughes and R. M. Bartlett. The use of performance indicators in performance analysis. *Journal of sports sciences*, pages 739–754, 2002.

[6] Operations NFL. Technology and the Game. National Football League. https://operations.nfl.com/gameday/technology/technology-and-the-game/, 2023.

[7] N. Silver. *The signal and the noise: The art and science of prediction*. Penguin UK, 2012.

[8] E. Vermeulen and S Venkata. Big data in sport analytics: applications and risks. *Journal*, 2018.

Konzeption eines Validierungsframeworks für Industrial IoT Clients

Arthur Mehlmann **Rainer Keller**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma AW Connectivity Platform GmbH, Nürtingen

Einleitung

Im industriellen Internet Of Things (IIoT) gab es über die letzten Jahre ein starkes Wachstum an Lösungen, welche den Kunden das Erfassen von Prozessdaten durch Endgeräte im Feld und das Weiterleiten dieser Daten in cloudbasierte Dienste ermöglicht. Mit diesen Daten werden Prozessvisualisierungen, Überwachungen aber auch maschinelles Lernen umgesetzt. Gleichzeitig entstand dadurch ein hoher Bedarf diese Endgeräte aus der Cloud heraus zu verwalten und zu aktualisieren. Damit dies alles umgesetzt werden kann, wurden herstellerübergreifende Standards wie OPC-UA oder auch Hawkbit eingeführt, welche zwischen Endgerät und Cloud genutzt werden. Aber auch viele herstellerspezifische Protokolle sind entstanden, die im Kern meistens auf HTTP oder MQTT als Transportschicht setzen und eigene Datenformate transportieren. Gerade wenn die Anbieter der Clouddienste die dazu passenden Clients nicht selbst entwickeln, sondern dies ihren Kunden überlassen, ergibt sich die Frage, wie sichergestellt wird das diese 3rd Party Clients konform den Spezifikationen des Cloud Anbieters arbeiten. Den Entwicklern kann zwar mit Dokumentation und Referenzimplementierungen geholfen werden. Häufig ergibt sich aber das Problem, das nicht alle Szenarien getestet werden können. Insbesondere Ausnahmesituationen, bei denen die Clouddienste nicht fehlerfrei funktionieren und Fehler zurückliefern, lassen sich nicht ohne weiteres testen. Die Simulation der Serverseite ist hier für den Entwickler nur eine unzureichende Lösung, da diese aufwendig ist und nicht der Realität entsprechen muss. Für die Cloudbetreiber auf der anderen Seite ergibt sich das Problem von sich nicht konform verhaltenden Clients, die von funktionellen Problemen bis hin zu Überlastungen der Dienste führen können.

Zielsetzung

Ziel dieser Arbeit ist es ein Framework zu erarbeiten, dass ein technologisch offenes Konzept bietet, um IoT Clients und deren Protokollimplementierung gegen einen simulierten Backend zu testen. Das Framework benötigt dazu eine Beschreibungssyntax, um das Verhalten auf unterschiedliche Client Anfragen zu simulieren. Auf Basis dieser Simulationen soll dann ein Testlauf für einen Client konfiguriert werden können, der in einem auswertbaren Bericht zur Konformität des Clients auf Protokollkonformität sowie Verhalten in Ausnahmesituationen mündet.

Black Box Testing

Black Box Testing ist eine von vielen Methoden die für die Überprüfung von Software existieren. Da die internen Implementierungen der IoT Clients dem Tester nicht bekannt sind, wird das Black Box Verfahren für den Test benutzt. Der Tester hat keinen Einblick in den Quellcode der Software. Somit erfolgt der Test durch das ausführen der Programm Funktion mit definierten Eingabewerten und der anschließenden Überprüfung der ausgegebenen Werte. Für die Überprüfung der Ausgabewerte müssen Anforderungen definiert sein, mit welchen das richtige Verhalten der Funktion ausgesagt wird [3]. In dem Fall dieser Arbeit ist die Black Box das IoT Gerät und die Eingabe – und Ausgabewerte sind die HTTP Requests und Responese wie in Abbildung 1 illustriert.

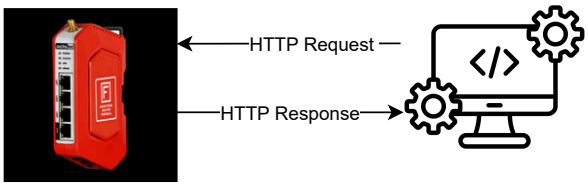


Abb. 1: Black Box Testing eines IoT Clients [1]

API Beschreibungssprache

Um den Server zu simulieren muss eine Spezifikation der für die Kommunikation genutzten API gegeben sein.

Dafür wurden über die Jahre einige Beschreibungssprachen wie z.B. OpenAPI, RAML oder API Blueprint entwickelt. Mit diesen Sprachen können alle Eigenschaften von APIs beschrieben werden, wie in etwa die verfügbaren Antworten mit den dazugehörigen HTTP Status Codes, die gegebenen Anfragen, Datenobjekte die verschickt und empfangen werden können, Authentifizierungsmethoden, und HTTP Request Parameter [4].

Fuzzing

Fuzzing wird für die automatische Generierung von Testfällen beim Software Testing genutzt. Durch Ausführen der Testfälle können sicherheitsrelevante Schwachstellen und Fehler in Programmen aufgedeckt werden, welche sonst zu ungewolltem Verhalten oder Abstürzen führen würden [2].

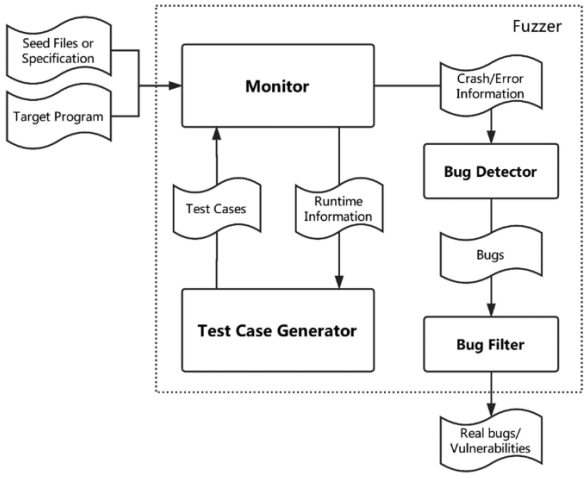


Abb. 2: Komponenten im Fuzzing Prozess [2]

Wie in Abbildung 2 zu sehen ist, ist der Prozess ist in vier wesentliche Hauptkomponenten eingeteilt.

- Die *Monitor* Komponente nutzt verschiedene Techniken um möglichst viele Laufzeit Informationen über das Programm herauszufinden. Diese werden im weiteren Verlauf als Hilfe für die Generierung der Testfälle verwendet [2].
- Der *Test case generator* erstellt mit unterschiedlichen Ansätzen Testfälle für das Testprogramm. Einmal werden Eingabewerte erstellt mittels der „Seed“ Datei. Für verschiedene Werte wird die Datei zufällig oder nach definierten Regeln verändert und so verschiedene Eingaben erzeugt. Auf der anderen Seite wird die Struktur für die Eingabe vorgegeben. Dadurch werden teilweise richtige Werte automatisch erzeugt, da die richtige Struktur generiert wird, jedoch mit zufälligen Werten [2].
- *Bug detector* ist zum Auffinden von Fehlern. Bei Fehlern oder Abstürzen vom Testprogramm werden die Rückmeldungen (Stack Traces, Fehler Codes, ...) gesammelt und für weitere Analyse gespeichert [2].
- Mit dem *Bug Filter* wird bestimmt, welche Fehler gesammelt werden. Das ist sinnvoll, wenn den Tester nur eine bestimmte Gruppe an Fehler interessiert [2].

Ausblick

Für die Entwicklung von Software ist das Testen heutzutage nicht mehr wegzudenken. Mit dem aufgeführten Framework soll sichergestellt werden, dass sich ein IoT Client konform gegenüber einem Protokoll verhält, bevor der Auslieferungsprozess beginnt. Da in der heutigen Entwicklung von APIs immer eine Dokumentation erforderlich ist und die API Dokumente somit vorhanden sind, bietet die Kombination der aufgeführten Ansätze eine effiziente Methode zur Überprüfung mehrerer Clients. Momentan wird OpenAPI als Beschreibungssprache verwendet. Für die zukünftige Entwicklung wäre die Unterstützung mehrerer Sprachen relevant.

Literatur und Abbildungen

[1] Eigene Darstellung.

[2] Hongliang Liang, Xiaoxiao Pei, Xiaodong Jia, Wuwei Shen, and Jian Zhang. Fuzzing: State of the art. In *IEEE Transactions on Reliability*. IEEE, 2018.

[3] Jiantao Pan. Software testing. In *Dependable Embedded Systems*. Citeseer, 1999.

[4] Vijay Surwase. REST API modeling languages-a developer's perspective. In *Int. J. Sci. Technol. Eng*. Pune Institute of Computer Technology, Pune, 2016.

Verteilung von zentralen Firewall-Regeln auf dezentrale Filterstellen in einem Netzwerk

Samuel Mueller **Tobias Heer**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma Hirschmann Automation and Control GmbH, Neckartenzlingen

I. Motivation und Problem

Firewalls sind Netzwerk-Sicherheitskomponenten, die Netzwerkverkehr basierend auf den Paket-Headern von eintreffenden Paketen filtern. Gefiltert wird durch einen Regelsatz, der definiert, welcher Verkehr zwischen zwei Netzwerkzonen ausgetauscht werden kann. Aktuelle Netzwerkarchitekturen setzen auf zentrale Firewalls, die zwischen Netzwerkzonen platziert werden. Diese Filterstellen werden stark belastet und erzeugen teilweise hohe Latenzen fur eintreffende Netzwerkpakete, da jedes Paket anhand einer vordefinierten Liste an Regeln, dem so genannten Regelsatz, überprüft werden muss. In der Arbeit wird untersucht, wie Regeln einer solchen zentralen Filterstelle auf mehrere dezentralisierte Filterstellen innerhalb der betroffenen Netzwerkzonen ausgelagert werden können. Der Vorteil hierbei ist die Verteilung der Filterlast auf mehrere Filterstellen im Netzwerk und die daraus resultierende Verringerung der Latenz auf erlaubtem Netzwerkverkehr. Die oberste Priorität ist dabei das Verhalten für den Verkehr innerhalb einer Zone und über die Zonengrenze hinaus nicht zu verändern. Dies wird im Folgenden als äquivalentes Filterverhalten bezeichnet. Ein Anwendungsfall für diese Verteilung ist die Reduktion von Latenzen in einem Netzwerk, die durch Firewalls erzeugt werden. Aus [6] geht hervor, dass softwarebasierte Firewalls, wie z.B. Iptables, mehr Latenzen erzeugen als hardwarebasierte Firewalls, wie beispielsweise die Access Control Lists (ACLs) von Switchen. Um Latenzen in einem Netzwerk zu minimieren, könnte daher eine zentrale, softwarebasierte Firewall durch mehrere ACLs auf bereits im Netzwerk vorhandene Switche ersetzt werden. Ein industrieller Switch verfügt allerdings über wesentlich weniger Regeln als eine industrielle Softwarefirewall. Daher kann der Regelsatz nicht als Ganzes verschoben werden und eine Verteilung der Regeln wird notwendig.

II. Design

Dieses Kapitel geht genauer auf die Dezentralisierung von Regeln, also der Verschiebung von Regeln von einer zentralen Filterstelle zu dezentralen Filterstellen, in einem Netzwerk ein. Zunächst wird der allgemeine Ansatz bei der Verschiebung von Regeln besprochen. Danach wird auf Probleme und Hindernisse eingegangen, die bei der Konzeption und Umsetzung berücksichtigt werden müssen.

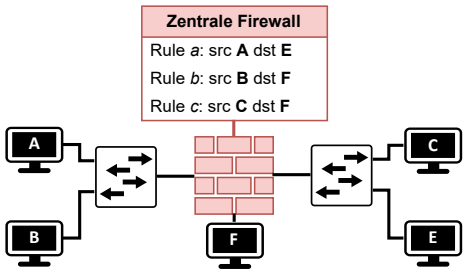


Abb. 1: Beispiel-Topologie mit zentraler Firewall [3]

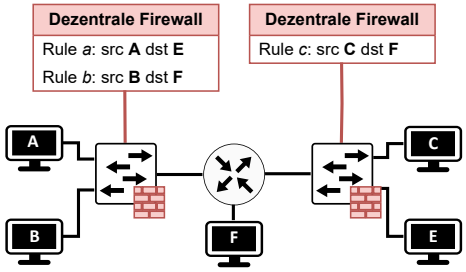


Abb. 2: Beispiel-Topologie mit verteilten Firewalls [3]

A. Allgemeiner Ansatz

Das Ziel der Arbeit ist das korrekte Verteilen der Regeln von einer zentralen Filterstelle auf mehrere dezentralisierte Filterstellen. Der Ansatz zu dieser Verteilung wird in den Abbildungen 1 und 2 dargestellt. Dabei beschreibt Abbildung 1 den Ursprungszustand:

ein Netzwerk mit einer einzelnen Filterstelle für alle Kommunikation zwischen dem linken, unteren und rechten Netzwerk. Abbildung 2 stellt den Finalzustand dar, in dem die Regeln der einzelnen Filterstelle aus Abbildung 1 auf mehrere dezentralisierte Filterstellen aufgeteilt wurden. Bei der Verteilung ist darauf zu achten, dass Regeln untereinander abhängig sein können [2]. Für äquivalentes Filterverhalten müssen bestimmte Regeln deshalb zusammen verschoben werden. Ein alternatives Vorgehen, das in dieser Arbeit angewandt wird, ist die Dekorrelation von Regeln, die Abhängigkeiten zwischen Regeln aufzulöst, wodurch Regeln individuell im Netzwerk verteilt werden können. Wichtig ist außerdem, dass der interne Verkehr einer Netzwerkzone durch die verschobenen Regeln nicht beeinflusst wird. Typischerweise wird Verkehr innerhalb einer Netzwerkzone erlaubt (blacklisting) und über die Grenze der Netzwerkzone hinaus verboten (whitelisting). Diese zwei gegenteiligen Filteransätze müssen bei der Verteilung berücksichtigt werden. Abbildung 2 stellt diese Sachverhältisse zur Vereinfachung nicht dar und beschränkt sich in der Darstellung auf die logische Verschiebung von Regeln.

B. Probleme

Die Anwendbarkeit des Ansatzes ist in der Praxis durch zwei Probleme limitiert, die im Folgenden erläutert werden. In der Arbeit gilt es, diese Probleme genauer auszuarbeiten und die Anwendbarkeit des Ansatzes zu demonstrieren.

1) *Geringe Regelkapazität:* Bei der Regelkapazität einer Filterstelle handelt es sich um die Maximalanzahl an Regeln, die in dieser gespeichert werden können. Die Problematik im Zusammenhang mit der Regelkapazität ergibt sich durch die heterogene Gerätelandschaft in einem Netzwerk. Unterschiedliche Filtergeräte haben unterschiedliche Regelkapazitäten, weshalb Regelsätze nicht vollständig übertragbar sind. Der zuvor besprochene allgemeine Ansatz ist bereits ein erster Schritt, um dieses Problem zu lösen. Je nach dem, wie viele Endgeräte von einer Regel betroffen sind, kann eine Regel allerdings mehrere Pfade im Netzwerk umfassen. Die Verteilung von Regeln der zentralen Filterstelle muss daher jeden dieser möglichen Pfade berücksichtigen. Bei Verzweigungen von Paketpfaden im Netzwerk müssen betreffende Regeln mehrfach installiert werden, um jeden möglichen Pfad durch das Netzwerk abzudecken. Aus diesem Grund werden Regeln vorzugsweise in einer Filterstelle installiert, die auf einem gemeinsamen Teilpfad aller Pfade der jeweiligen Regel liegen. Das Erreichen der Regelkapazität von solchen Filterstellen ist absehbar. Deshalb werden Regeln vor dem Verschieben nach der Anzahl an Regeln priorisiert, die erzeugt werden würden, wenn sie nicht in einer Filterstelle auf einem gemeinsamen Pfad installiert werden.

2) *Dynamik im Netzwerk:* Bei der Verteilung von Regeln auf mehrere Filterstellen sind dynamische Änderung von Pfaden im Netzwerk zwischen verschiedenen Endgeräten ein grundlegendes Problem. Die Regeln für einzelne Endgeräte sind über das Netzwerk verteilt und bei der Veränderung der Netzwerk Topologie oder der Pfade muss darauf reagiert werden, um das Filterverhalten im Netzwerk korrekt aufrecht zu erhalten. In der Arbeit werden die folgenden Ursachen für dynamische Änderungen der Netzwerktopologie zur Laufzeit genauer betrachtet: 1) Redundanz im Netzwerk und 2) mobile Endgeräte.

Durch Redundanz und mobile Endgeräte in Layer 2 Topologien werden Pakete teilweise auf mehreren Pfaden durch das Netzwerk geleitet. Ein Beispiel für Redundanz in Layer 2 Topologien ist das parallele Installieren von mehreren Switchen an kritischen Punkten. Dadurch kann bei einem Ausfall eines Switches oder Kabels der Verkehr automatisch umgeleitet werden. Mobile Endgeräte nutzen mehrere \acrlong*{nap} und haben daher keinen statischen Zugriffspunkt aufs Netzwerk. Im Gegensatz zum Filtern auf zentralen Filterstellen wird in diesen Beispielen daher nicht garantiert, dass Pakete über die für diese Kommunikation bestimmte dezentrale Filterstelle geleitet werden. Die Menge aller möglichen Pfade sowohl für Redundanz als auch mobile Endgeräte ist allerdings berechenbar. Unsere Algorithmik berücksichtigt diese möglichen Variationen, indem einzelne Regeln mehrfach im Netzwerk verteilt werden. Auf diese Weise muss nicht auf Pfadwechsel reagiert werden, da alle möglichen Paketpfade durch Regelduplikate bereits initial abgedeckt werden.

III. Evaluation

Die Ergebnisse der Arbeit werden nach drei verschiedenen Kriterien bewertet:

A. Korrektheit der Verteilung

Bei Firewalls handelt es sich um sicherheitskritische Komponenten. Daher ist äquivalentes Filterverhalten im Netzwerk vor und nach der Verteilung von Regeln von äußerster Relevanz. In der Arbeit werden deshalb formale Anforderungen für das sichere Verteilen von Regeln in einem Netzwerk definiert. Werden diese Anforderungen von einer beliebige Verteilungsalgorithmik erfüllt, so ist das Filterverhalten vor und nach der Regelverteilung äquivalent.

B. Robustheit

Wenn der Netzwerkverkehr nicht weiter durch eine zentrale Filterstelle abgearbeitet wird, sondern dezentralisiert durch mehrere Filterstellen im Netzwerk, werden empfangene verbotene Pakete im Vergleich zum Ursprungszustand früher oder später in einem Pfad gefiltert. Je später ein Paket im Netzwerk abgearbeitet wird, desto höher ist die Last für das gesamte Netzwerk, was u.a. durch Denial of Service Angriffe ausgenutzt

werden kann. Die Verteilungsalgorithmen der Arbeit werden deshalb danach bewertet, wie früh Pakete durchschnittlich im Netzwerk abgearbeitet werden. Der Anspruch hierbei ist, dass der größte Teil des Verkehrs früher abgearbeitet wird.

C. Ressourcennutzung

Es kann nicht davon ausgegangen werden, dass beliebig viele Regeln im Netzwerk verteilt werden können. Dezentrale Filterstellen haben meist eine limitierte Anzahl an Regeln, die sie beinhalten können. Die Anzahl der im Netzwerk verteilbaren Regeln ist somit limitiert. Die Verteilungsalgorithmen der Arbeit werden deshalb danach bewertet, wie effizient die vorhandenen Ressourcen genutzt werden. Der Anspruch ist hierbei nach der Verteilung mit möglichst wenig Regeln im Netzwerk dieselbe Filterlogik abzubilden, wie im Originalzustand des Netzwerkes.

IV. Verwandte Arbeit

In diesem Kapitel wird auf verwandte Arbeiten verwiesen, die Vorarbeit zu einzelnen Komponenten der Arbeit geleistet haben.

In [6] haben Wüsteney et al. die zeitlichen Diskrepanzen zwischen der Verarbeitungszeit von hardwarebasierten Firewalls und softwarebasierten Firewalls gemessen. Sie haben dadurch demonstriert, dass hardwarebasierte Firewalls Latenzprobleme in Netzwerken lösen können. Durch die limitierte Regelkapazität können hardwarebasierte Firewalls softwarebasierte Firewalls allerdings nicht 1:1 ersetzen. Es ist eine Verteilung von Regeln notwendig, wie sie in unserer Arbeit vorgestellt wird.

Al-Shaer et al. haben in [2] jede mögliche Beziehung zwischen Regeln in einem Regelsatz bestimmt und die Vollständigkeit ihrer Aufzählung bewiesen. Diese Beziehungen wurden in verschiedenen Arbeiten verwendet, um anhand diverser Algorithmen Regelsätze zu optimieren [1] [4] [5]. Die Ergebnisse dieser Arbeiten werden in unserer Arbeit verwendet, um Regeln für die Verteilung im Netzwerk vorzubereiten. Yoon et al. haben in [7] einen Ansatz verfolgt, der die maximale Größe der Firewall-Regelsätze im Netzwerk minimiert. Hierfür bestimmen sie ideale Positionierungen von Firewalls im Netzwerk und bestimmen optimale Pfade für den Netzwerkverkehr. In unserer Arbeit und den darin verwendeten Referenzszenarien wird von existierenden Topologien ausgegangen. Ideale Netzwerke, wie sie in [7] erstellt werden, können deshalb nicht verwendet werden.

V. Ergebnisse

Diese Arbeit stellt die Möglichkeit vor, Regeln einer zentralen Filterstelle auf mehrere dezentrale Filterstellen auszulagern. Hierfür wird ein Ansatz für das sichere Verteilen von Regeln in einem Netzwerk vorgestellt. Der Lösungsansatz wird prototypisch umgesetzt, in generierten Netzwerken simuliert und anhand gemessener Daten evaluiert. Die Messergebnissen zeigen, dass im Zuge dieser Arbeit eine Algorithmik entwickelt wurde, die sehr gute Ergebnisse für die Robustheitssteigerung und Latenzminimierung in Netzwerken liefert, während vorhandene Ressourcen effizient genutzt werden.

Literatur und Abbildungen

[1] Subrata Acharya, Jia Wang, Zihui Ge, Taieb F. Znati, and Albert Greenberg. Traffic-aware firewall optimization strategies. In *2006 IEEE International Conference on Communications*. IEEE, 2006.

[2] Ehab S. Al-Shaer and Hazem H. Hamed. Modeling and Management of Firewall Policies. In *IEEE Transactions on Network and Service Management*. IEEE, 2023.

[3] Eigene Darstellung.

[4] Errin W. Fulp. Optimization of network firewall policies using directed acyclical graphs. In *Proceedings of the IEEE Internet management conference*. IEEE, 2005.

[5] Hazem H. Hamed and Ehab S. Al-Shaer. Dynamic rule-ordering optimization for high-speed firewall filtering. In *Proceedings of the 2006 ACM Symposium on Information, computer and communications security*. ACM, 2006.

[6] Lukas Wusteney, Michael Menth, Rene Hummen, and Tobias Heer. Impact of Packet Filtering on Time-Sensitive Networking Traffic. In *2021 17th {IEEE} International Conference on Factory Communication Systems (WFCS)*. IEEE, 2023.

[7] MyungKeun Yoon, Shigang Chen, and Zhan Zhang. Minimizing the Maximum Firewall Rule Set in a Network with Multiple Firewalls. In *IEEE Transactions on Computers*. IEEE, 2010.

Konzept und Implementierung einer Anwendung für die digitale Erfassung der Anwesenheit von Studenten beim IT-Kolloquium

Robert Mueller **Mirko Sonntag**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt an der Fakultät Informatik und Informationstechnik, Esslingen

Einleitung

In der sich schnell weiterentwickelnden Hochschullandschaft ist die Integration digitaler Technologien für die Verbesserung verschiedener Aspekte der akademischen Verwaltung von großer Bedeutung. Im Bereich der Planung und Administration ist die Digitalisierung bereits überwiegend angekommen. So ist heutzutage beispielsweise eine digitale Plattform für die Lerninhalte der verschiedenen Kurse nicht mehr wegzudenken. [3] Ein Bereich, in welchem die Digitalisierung ebenfalls einen hohen Nutzen bringen kann, um den bürokratischen Aufwand so gering wie möglich zu halten, ist die Erfassung der Anwesenheiten von Studierenden. Dabei handelt es sich um einen Prozess mit traditionell hohem manuellem Aufwand, welcher für eine transformative Verlagerung in Richtung Digitalisierung sehr geeignet ist. Deshalb wird ein Wechsel von der herkömmlichen, papierhaften Anwesenheitserfassung zu einem komplett digitalen System angestrebt.

Der Übergang von einer papierhaften Methode der Anwesenheitserfassung zu einem digitalisierten System ist durch die Notwendigkeit von Effizienz und Anpassungsfähigkeit in der heutigen dynamischen Bildungsumgebung motiviert. Durch die heutigen Technologien bieten sich für Bildungseinrichtungen neue Wege zur Rationalisierung von Verwaltungsprozessen. Durch den Einsatz solcher Anwendungen können Routineaufgaben automatisiert und somit die Mitarbeiter der Bildungseinrichtungen entlastet werden.

Allerdings ist die Anwesenheitserfassung nicht nur im Bereich der routinemäßigen Vorlesungen wichtig, sondern umfasst auch spezielle Veranstaltungen wie das IT-Kolloquium der Hochschule Esslingen. Dabei handelt es sich um eine Vortragsreihe zu den verschiedensten aktuellen Themen aus der Informationstechnik, welche jeder Student der Fakultät IT mindestens 3-mal während seines Studiums besuchen muss. Die Komplexität der Anwesenheitserfassung bei solch speziellen Veranstaltungen führt dazu, dass Professoren sowie Studenten viel Zeit mit der Bestätigung der Anwesenheiten verschwenden. Durch die Einführung eines digitalen Systems kann der Professor, der das IT-Kolloquium leitet, die Teilnahme jedes einzelnen Studenten einfach und lückenlos verfolgen.

Aufgabenstellung

Derzeit haben alle Studierenden die Pflicht, ihre Anwesenheit bei den Vorträgen des IT-Kolloquiums mithilfe eines Formulars zu dokumentieren. Dieses Formular muss zu den Vorträgen mitgebracht und anschließend vom betreuenden Professor unterzeichnet werden. Jeder Studierende muss dieses Dokument bis zu seinem Abschluss sicher aufbewahren, um durch Hochladen desselben seine Anwesenheit an mindestens 3 IT-Kolloquien belegen zu können. Von einer Mitarbeiterin der Hochschule muss im Anschluss daran für jeden Absolventen überprüft werden, ob dieser ein Formular hochgeladen hat und ob dieses vollständig ist.

Das Ziel dieser Arbeit ist die Konzeption und Implementierung einer Softwarelösung, mit der die Anwesenheit von Studierenden bei den Vorträgen des IT-Kolloquiums erfasst werden kann. Es handelt sich um ein vollständig digitales System, das ausschließlich im Hörsaalbereich verfügbar ist. Dadurch wird verhindert, dass Personen das System von anderen Orten aus nutzen und ihre Anwesenheit vortäuschen. Für das System ist folgenden Punkten besondere Beachtung zu schenken:

- Die Erfassung der Anwesenheit muss auf einen bestimmten Vorlesungssaal begrenzt werden können.
- Die Studierenden müssen ihre Identität nachweisen.
- Es muss sichergestellt werden, dass der Student die gesamte Veranstaltung besucht hat.

- Die erfassten Daten müssen den Verantwortlichen zum Abruf bereitgestellt werden.
- Studierende müssen ihre eigenen Anwesenheitsdaten einsehen können.
- Vorliegende Teilnehmerlisten müssen importiert werden können.

Lösungsansätze

1. **Webanwendung in örtlich beschränktem WLAN**

   Ein Lösungsansatz besteht in der Entwicklung einer Webanwendung, die nur innerhalb eines speziell eingerichteten WiFi-Netzes zugänglich ist. Die Studenten würden sich mit ihren Universitätsanmeldedaten bei der Anwendung anmelden, um ihre Anwesenheit zu bestätigen. Diese Methode bietet eine kontrollierte Umgebung, die sicherstellt, dass die Anwesenheitsbestätigung auf bestimmte Standorte mit dem in diesem Bereich eingerichteten WLAN beschränkt ist. Dieser Lösungsansatz musste aufgrund von Sicherheitsbedenken und der fraglichen Machbarkeit hinsichtlich der Netzwerke und Netzwerkhoheit verworfen werden.

2. **NFC-basiertes System mit Studentenausweis**

   Ein weiterer Ansatz nutzt die Near Field Communication (NFC)-Technologie, die in die Studentenausweise integriert ist. Die Studenten halten ihre Ausweise einfach an ein NFC-Lesegerät, um ihre Anwesenheit zu bestätigen. Diese Methode vereinfacht den Prozess und bietet eine schnelle und kontaktlose Lösung. Der Einsatz von NFC gewährleistet eine sichere und effiziente Art der Anwesenheitsprüfung, verringert das Risiko manueller Fehler und verbessert die allgemeine Benutzerfreundlichkeit.

   Die Schwierigkeit bei diesem Lösungsansatz liegt darin, die entsprechenden Studentendaten zu erhalten, da diese nicht direkt auf den Ausweisen gespeichert sind, sondern über die Karten-ID abgerufen werden müssen. Da die Verantwortung für diese Systeme in verschiedenen Händen liegt, während die Studentenausweise und Lesegeräte von dem ausstellenden Unternehmen verwaltet werden, liegen die mit den Karten-IDs verknüpften Studentendaten auf einem speziellen Server der Hochschule. Die Implementierung der NFC-Technologie erfordert zudem möglicherweise eine Anfangsinvestition in ein kompatibles NFC-Lesegerät.

   Da die für dieses System verantwortliche Abteilung momentan keine Kapazität für eine Implementierung einer Schnittstelle besitzt, wird dieser Lösungsansatz vorerst nicht weiterverfolgt, aber aufgrund der Vorteile der NFC-Technologie könnte das System in Zukunft auf diese Authentisierungsvariante umgestellt werden.

3. **Barcode-basiertes System mit Studentenausweis**

   Der dritte Ansatz besteht darin, die Matrikelnummer des Studenten von seinem Ausweis mithilfe eines Barcode-Lesegeräts auszulesen, um die Anwesenheit zu bestätigen. Die Studierenden halten ihre Ausweise an ein dafür vorgesehenes Lesegerät, und das System erfasst ihre Anwesenheit anhand der gescannten Matrikelnummer. Auf Grund der Barcode-Scantechnologie ist diese Methode sehr kosteneffizient, allerdings setzt diese voraus, dass die für das System erforderlichen Studentendaten, wie z. B. der Name des Studenten, über die Matrikelnummer abgefragt werden können.

Konzept

Nach Prüfung der einzelnen Lösungsideen wurde entschieden, vorerst das Barcode-basierte Studentenausweissystem zu verwenden und dieses eventuell in einem späteren Schritt durch das NFC-basierte System zu ersetzen. Das System soll aus 3 Hauptkomponenten bestehen.

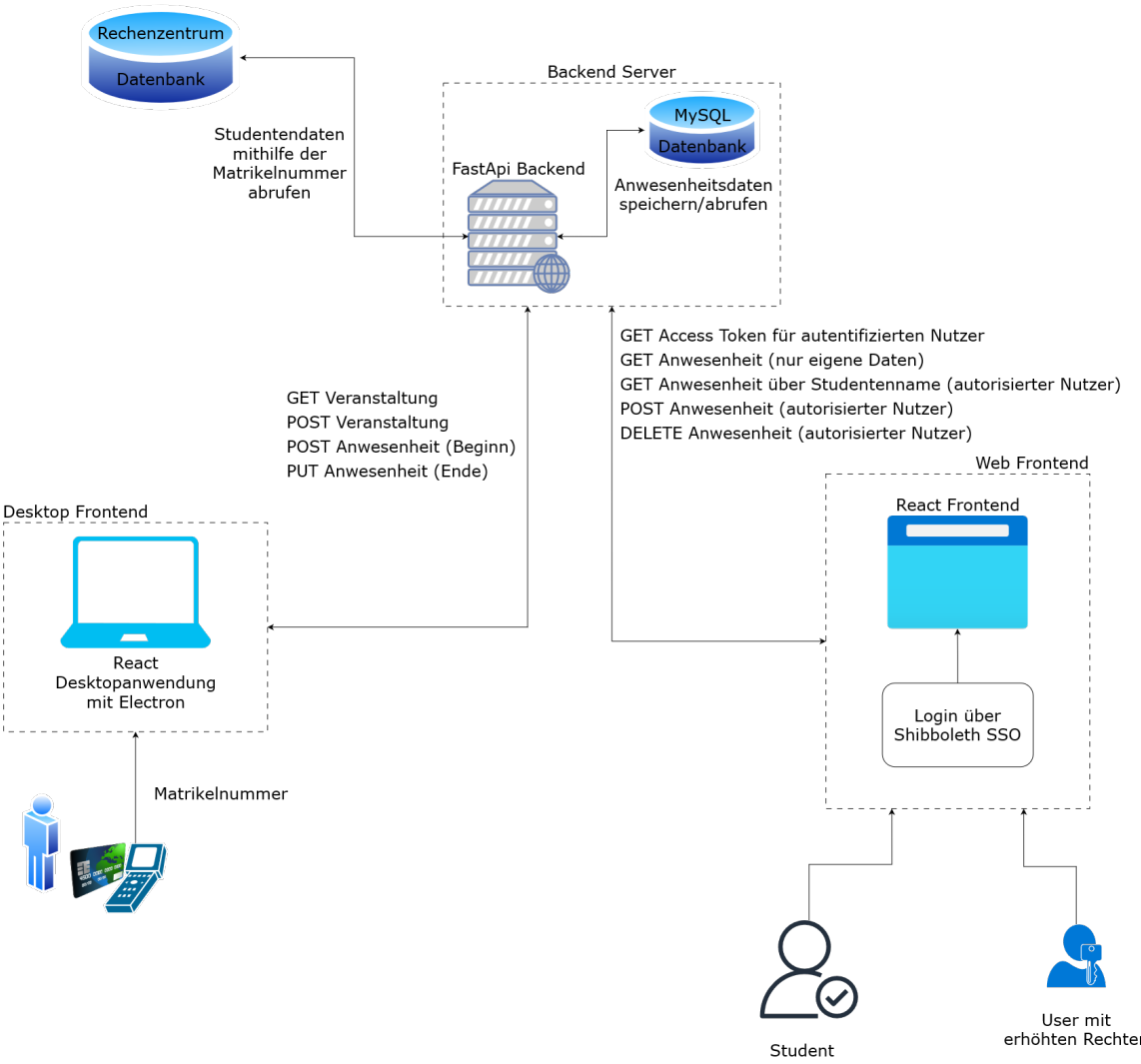


Abb. 1: Architektur-Konzept [1]

Die Grundlage dieses Lösungskonzepts bildet das Backend, welches mit FastAPI in Python entwickelt und an eine MySQL-Datenbank für effiziente Datenspeicherung und -abfrage angebunden wird. FastAPI zeichnet sich durch die sehr schnelle Entwicklungszeit und die Unterstützung von asynchronem Code ohne Verwendung von Frameworks eines Drittanbieters aus. [4] Das Backend wird als zentraler Punkt für die Verarbeitung der Anwesenheitsdaten und die Kommunikation sowohl mit der Desktopanwendung als auch der Webanwendung verwendet.

Eine weitere wichtige Komponente besteht in einer Desktop-Anwendung, welche mit React innerhalb des ElectronJS-Frameworks entwickelt wird. Es soll ein benutzerfreundliches GUI zum Scannen der Studentenausweise mithilfe eines angeschlossenen Barcode-Lesegeräts anbieten. ElectronJS ermöglicht die Erstellung von plattformübergreifenden Desktop-Anwendungen und gewährleistet die Kompatibilität mit verschiedenen Betriebssystemen. [2] Durch die Verwendung der React JavaScript-Bibliothek wird die Entwicklung einer reaktionsschnellen und grafisch ansprechenden Benutzeroberfläche, welche die Benutzerfreundlichkeit für die betreuenden Professoren erhöht, erleichtert. Die Anwendung erfasst die Matrikelnummer über den Barcode des Studentenausweises und kommuniziert anschließend mit dem Backend, um die Anwesenheitsdaten sicher in der MySQL-Datenbank zu speichern. Hierbei wird zum einen die Anwesenheit zu Beginn der Veranstaltung als auch nach dem Ende erfasst. Um die Ausfallwahrscheinlichkeit des Systems zu minimieren, werden die Daten zusätzlich in einer CSV-Datei gespeichert, damit auch bei nicht vorhandener Internetverbindung die Anwesenheit weiterhin erfasst

werden kann. Diese CSV-Datei kann im Anschluss bei vorhandener Internetverbindung über eine Bulk-Anfrage an das Backend gesendet werden.
Für das Frontend wird eine Webanwendung mit React entwickelt, die mit dem Backend interagiert und die gespeicherten Anwesenheitsdaten über das Backend aus der MySQL-Datenbank abruft. In dieser Webanwendung wird es autorisierten Benutzer ermöglicht, auf eine einfache Art und Weise auf die Anwesenheitsdaten zuzugreifen. Für die Authentifizierung dieser Webanwendung soll das von der Hochschule verwendete SSO-System von Shibboleth verwendet werden. Die Webanwendung bietet zudem für autorisierte Benutzer die Funktionalität, vorhandene Teilnehmerlisten zu importieren.

Ausblick

In einem ersten Schritt soll das System wie im Konzept beschrieben fertiggestellt werden. Hierbei ist vor allem auch auf die Sicherheit des Systems zu achten, da mit sensiblen Daten gearbeitet wird und diese nicht von Unbefugten einsehbar sein dürfen. In einem weiteren Schritt kann dann auf das NFC-basierte System umgestellt werden. Außerdem soll dieses System auch die Möglichkeit bieten, für andere Veranstaltungen als das IT-Kolloquium verwendet werden zu können. Es soll also in diesem Bereich eine gewisse Flexibilität bieten.

Literatur und Abbildungen

[1] Eigene Darstellung.

[2] OpenJS Foundation. Build cross-platform desktop apps with JavaScript, HTML, and CSS | Electron. https://www.electronjs.org/, 2023.

[3] Prof. Dr. Michael Jäckel. Der Campus und die Digitalisierung: So sieht die Universität der Zukunft aus. https://hochschulforumdigitalisierung.de/blog/der-campus-und-die-digitalisierung-so-sieht-die-universitaet-der-zukunft-aus/, 05 2017.

[4] Muhtasim Fuad Rafid. FastAPI - The Good, the bad and the ugly. https://dev.to/fuadrafid/fastapi-the-good-the-bad-and-the-ugly-20ob, 08 2020.

Anomalieerkennung in Finanztransaktionen

Mahir Oezcan **Dirk Hesse**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt an der Fakultät Informatik und Informationstechnik, Esslingen

Einleitung

Im Allgemeinen versteht man unter einer Anomalie eine Abweichung von der Norm. Die Autoren in [1] definieren Anomalien als Muster in Daten, die nicht mit einer genau definierten Vorstellung von „normalem“ Verhalten übereinstimmen. Diese Abweichungen können aus einer Vielzahl von Gründen auftreten, unter anderem durch böswillige Aktivitäten wie Finanzbetrug. Die Association of Certified Fraud Examiners schätzt für das Jahr 2022 den globalen Verlust durch Betrugsaktivitäten auf über 3,6 Milliarden US-Dollar [5]. Folglich bedeutet dies, dass das frühzeitige Erkennen solcher Unregelmäßigkeiten, nicht zuletzt aus ökonomischer Sicht, von immenser Wichtigkeit ist. In der Vergangenheit beruhte die Anomalieerkennung auf Experten, die Datensätze manuell überprüften. Dieses Vorgehen wurde jedoch mit dem sprunghaften Anstieg der Datenmenge in den letzten Jahrzehnten zunehmend ineffizienter, weshalb automatisierte Methoden des Machine Learnings an Popularität gewonnen haben. Diese ermöglichen zwar eine effizientere Erkennung von Anomalien, jedoch ergeben sich damit auch neue Herausforderungen, wie der Umgang mit stark unausgeglichen Datensätzen oder der Mangel an Trainingsdaten aufgrund von Datenschutzbedenken.

Zielsetzung

Das grundlegende Ziel der Arbeit ist es, verschiedene Modelle des maschinellen Lernens vorzustellen und ihre Leistungsfähigkeit zu vergleichen. Besonderes Augenmerk wird dabei auf die Erklärung des Ungleichgewichtsproblems und auf mögliche Lösungsansätze gelegt. Dies beinhaltet unter anderem die Betrachtung verschiedener Methoden auf Algorithmus- und Datenebene sowie die Vorstellung von One-Class-Learning-Ansätzen. Des Weiteren sollen synthetische Daten als mögliche Lösung für den Mangel an Datensätzen betrachtet werden. Der Einfluss der Hyperparameteroptimierung auf die Effektivität von Anomalieerkennungsverfahren wird ebenfalls beleuchtet.

Umgang mit unbalancierten Datensätzen

Ein Datenungleichgewicht ist definiert als eine Überrepräsentation einer bestimmten Klasse oder eines bestimmten numerischen Wertintervalls gegenüber einer anderen [4]. Modelle, die anhand eines unausgeglichenen Datensatzes trainiert werden, sind in der Regel stark voreingenommen und neigen insbesondere dazu, die Mehrheitsklasse zu begünstigen, während sie die Minderheitsklasse weitestgehend ignorieren [2]. In der wissenschaftlichen Literatur werden viele verschiedene Algorithmen und Methoden vorgeschlagen, dieses Problem zu bewältigen. Diese lassen im Wesentlichen in zwei Kategorien unterteilen: Methoden auf der Datenebene und Methoden auf der Algorithmusebene. Methoden auf der Datenebene haben das Ziel, Ungleichgewichte zu reduzieren oder gar zu beseitigen, indem sie die Klassenverteilung im Datensatz manipulieren. Häufig kommen sogenannte Resamplingmethoden zum Einsatz, mit denen ein Datensatz so angepasst wird, dass eine ausgewogene Verteilung entsteht. Dazu wird entweder die Minderheitsklasse vergrößert (Oversampling) oder die Mehrheitsklasse verkleinert (Undersampling). Bei den Methoden auf Algorithmusebene wird der ML-Algorithmus so angepasst, dass er Instanzen der Minderheitenklassen bevorzugt, beispielsweise durch Zuordnung unterschiedlicher Fehlklassifikationskosten (Cost-Sensitive-Learning).

Machine Learning

Machine Learning (ML), zu Deutsch maschinelles Lernen, ist ein Teilbereich der künstlichen Intelligenz. Ziel ist es, komplexe Modelle zu entwickeln, die in der Lage sind, auf der Grundlage gegebener Daten Vorhersagen über unbekannte aber ähnliche Daten zu treffen. Neuronale Netze sind eine beliebte Technik des maschinellen Lernens, welche den Lernmechanismus in biologischen Organismen simulieren. Neuronale Netze,

wie der Autoencoder in Abbildung 1, können zur Identifikation von Anomalien eingesetzt werden. Autoencoder komprimieren Datenpunkte und rekonstruieren sie anhand dieser komprimierten Form. Da das Modell ausschließlich mit normalen Daten trainiert wird, ist der Rekonstruktionsfehler bei diesen Daten minimal, wohingegen der Fehler bei Datenpunkten, die stark von den Trainingsdaten abweichen, höher ist. So lässt sich nun ein Schwellenwert für den Rekonstruktionsfehler wählen, anhand dessen Ausreißer klassifiziert werden können [6].

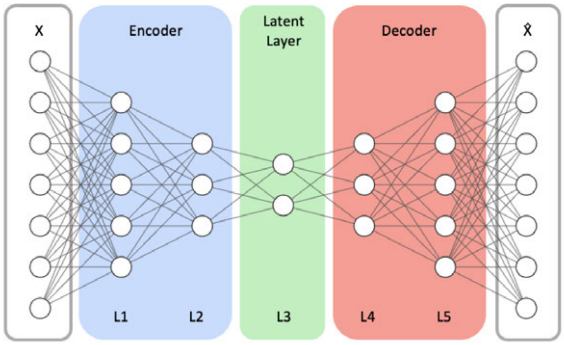


Abb. 1: Beispiel für eine Autoencoder-Architektur mit symmetrischen Encoder- und Decodernetzwerken [7]

Ein weiteres, für die Anomalieerkennung relevantes ML-Modell ist OC-SVM (One-Class Support-Vector-Machine). Das grundlegende Ziel von SVM-Klassifikatoren ist die Ermittlung der sogenannten Hyperebene, eine Klassengrenze, anhand derer einzelne Punkte klassifiziert werden können. OC-SVM sind spezielle SVM, welche spezifisch für die Erkennung von Ausreißern entwickelt wurden. Das Model wird ausschließlich anhand der Elemente der Normalklasse trainiert, sodass jegliche Abweichungen von diesen als Anomalien erfasst werden.
Eine weitere Möglichkeit zum Umgang mit unbalancierten Datensätzen bietet das Ensemble-Learning. Der Grundgedanke von Ensemble Learning ist es, statt nur einem, mehrere Klassifikatoren zu nutzen und deren Ergebnisse zu kombinieren. So nutzt beispielsweise Random Forest 2 einzelne Entscheidungsbäume und kombiniert die Vorhersagen zu einem finalen Ergebnis. Um eine Überanpassung der einzelnen Bäume zu verhindern, arbeitet Random Forest mit Bagging (Bootstrap Aggregation). Bei dieser Technik wird jeder Entscheidungsbaum nur mit einer Teilmenge des Ursprungsdatensatzes (bootstrap sample) trainiert. Dadurch unterscheidet sich Random Forest zu anderen Tree-Ensembles wie XG-Boost (Extreme Gradient Boosting). Bei XG-Boost werden die jeweiligen Entscheidungsbäume in jedem Schritt des Prozesses sequenziell dazu trainiert, die Fehler des vorherigen Klassifikators zu korrigieren, um so ein leistungsstarkes Modell zu schaffen. Aufgrund der Kombination mehrerer Klassifikatoren sind Ensemble-Methoden grundsätzlich weniger anfällig für Klassenungleichgewichte.

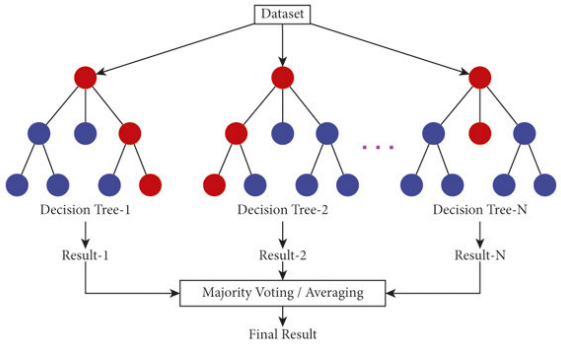


Abb. 2: Ein Random Forest [3]

Die letzte in dieser Arbeit betrachtete Methode ist die logistische Regression, ein statistisches Modell, welches der Prognose dichotomer Ergebnisse dient. Das Ziel ist es, anhand der Werte der Eingabemerkmale die Wahrscheinlichkeit zu schätzen, dass es sich bei einem Datenpunkt um eine Anomalie handelt.

Ausblick

Die nächsten Schritte dieser Arbeit beinhalten die umfassende Implementierung und Evaluierung der präsentierten Modelle. Um eine Vergleichbarkeit der ML-Modelle sicherzustellen, wird mithilfe von Hyperopt, sofern anwendbar, für alle Modelle eine möglichst optimale Hyperparameterkonfiguration ermittelt. Für das Autoencoder-Modell soll die optimale Größe der Kodierer- und Dekodierernetze sowie die Anzahl der Neuronen in den jeweiligen Hidden-Layern bestimmt werden. Anschließend folgen das Training und die Auswertung der Modelle. Dieser Schritt wird für jede Kombination aus ML-Modell und Resampling-Methode mehrmals durchgeführt. Dieses Vorgehen ermöglicht eine genauere Leistungsbewertung und statistisch aussagekräftigere Ergebnisse, da es die Auswirkung möglicher Ausreißer und Zufallseinflüsse minimiert.

Literatur und Abbildungen

[1] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly Detection: A Survey. *ACM Computing Surveys*, pages 1–58, 2009.

[2] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. SMOTE: Synthetic Minority Over-sampling Technique. *jair (Journal of Artificial Intelligence Research)*, pages 321–357, 2002.

[3] Muhammad Yaseen Khan, Abdul Qayoom, Muhammad Suffian Nizami, Muhammad Shoaib Siddiqui, Shaukat Wasi, and Syed Muhammad Khaliq-ur-Rahman Raazi. Automated Prediction of Good Dictionary EXamples (GDEX): A Comprehensive Experiment with Distant Supervision, Machine Learning, and Word Embedding-Based Deep Learning Techniques. *Complexity*, pages 1–18, 2021.

[4] Nuno Moniz, Paula Branco, and Luís Torgo. Evaluation of Ensemble Methods in Imbalanced Regression Tasks. In *Proceedings of the First International Workshop on Learning with Imbalanced Domains: Theory and Applications*, volume 74, pages 129–140. PMLR, 2017.

[5] Association of Certified Fraud Examiners. Occupational Fraud 2022: A Report to the nations. https://acfepublic.s3.us-west-2.amazonaws.com/2022+Report+to+the+Nations.pdf, 2022.

[6] Najmi Rosley, Gee-Kok Tong, Keng-Hoong Ng, Suraya Nurain Kalid, and Kok-Chin Khor. Autoencoders with Reconstruction Error and Dimensionality Reduction for Credit Card Fraud Detection. In *Proceedings of the International Conference on Computer, Information Technology and Intelligent Computing (CITIC 2022)*, pages 503–512. Atlantis Press International BV, 2022.

[7] Youngrok Song, Sangwon Hyun, and Yun-Gyung Cheong. A Systematic Approach to Building Autoencoders for Intrusion Detection. In *Silicon Valley Cybersecurity Conference*, volume 1383, pages 188–204. Springer International Publishing; Imprint Springer, 2021.

Entwickeln einer KI-basierten Lösung zur automatisierten Dokumentation von KI-Modellen und DataAssets

Kadir-Kaan Oezer **Markus Enzweiler**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma Mercedes-Benz Group AG, Böblingen

Einleitung

In der Welt der künstlichen Intelligenz (KI) ist die Dokumentation von Modellen und DataAssets wichtig, wird aber oft übersehen. Neue KI-Technologien wie ChatGPT, basierend auf dem Transformer-Modell, das in [2] vorgestellt wurde, und dem GPT-Framework aus [1], eröffnen Möglichkeiten, diesen Prozess zu automatisieren und zu verbessern.

Transformers: Die Grundlage von ChatGPT

ChatGPT und das Transformer-Modell haben die Textverarbeitung im maschinellen Lernen verändert. Im Vergleich zu früheren Methoden wie rekurrenten neuronalen Netzwerken (RNNs) oder Long Short-Term Memory (LSTM) Netzwerken ist dies ein großer Schritt. Die Attention-Mechanismen des Transformer-Modells spielen eine wichtige Rolle und ermöglichen es, wichtige Textteile unabhängig von ihrer Position im Text zu erkennen und zu verarbeiten. Das Transformer-Modell kann auch mehrere Textteile gleichzeitig bearbeiten, was den Trainingsprozess beschleunigt.
Traditionelle RNNs und LSTMs verarbeiten Daten sequenziell, was parallele Verarbeitung schwierig macht. Der Transformer kann jedoch mehrere Teile einer Sequenz gleichzeitig verarbeiten. Dies ist effizienter, besonders bei Sprachverarbeitungsaufgaben wie maschineller Übersetzung, die lange Sequenzen und komplexe Beziehungen beinhalten. Durch die Kombination von Self-Attention und Multi-Head Attention kann der Transformer subtile Bedeutungen und Kontexte in Texten erfassen [2].

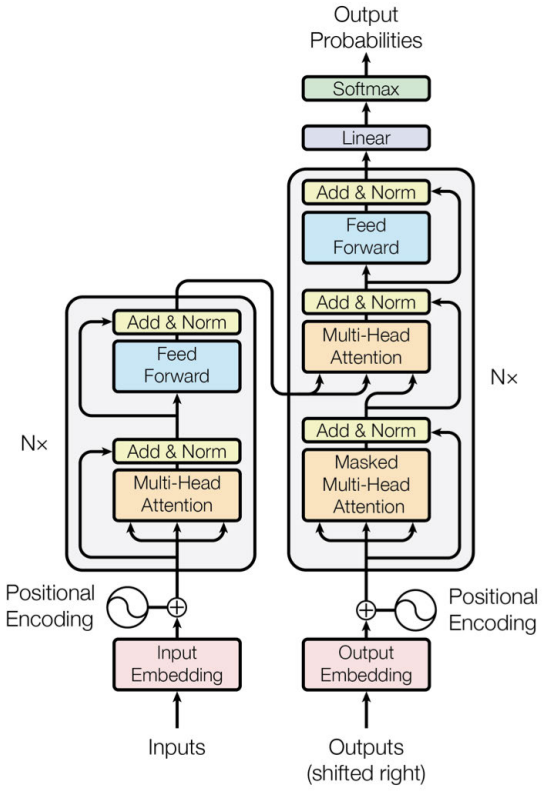


Abb. 1: Transformer Architektur [2]

GPT: Generative Pre-Training für erweitertes Sprachverständnis

Generative Pre-Training (GPT) hat die Fähigkeit von KI-Modellen, Sprache zu verstehen und zu generieren, erheblich verbessert. Die Modelle in dieser Reihe zeichnen sich durch eine stetig wachsende Anzahl von Parametern aus, was ihre Leistungsfähigkeit in der Sprachverarbeitung deutlich steigert.

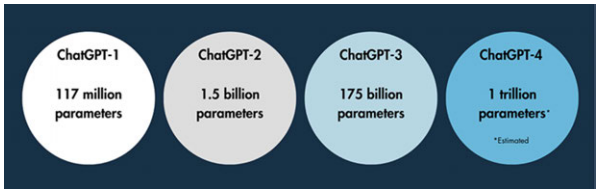


Abb. 2: Größe der GPT-Modelle [3]

Durch das Training mit großen Textmengen erlangt GPT ein umfassendes Verständnis von Sprache und deren Nuancen. Diese Fähigkeit ermöglicht es GPT-Modellen, nicht nur einfache Texte zu generieren, sondern auch komplexere sprachliche Strukturen und Inhalte zu verstehen. Dies ist besonders nützlich für Aufgaben, die über das einfache Verstehen von Text hinausgehen und ein tieferes Eindringen in den Kontext und die Bedeutung von Sprache erfordern.
Die umfangreichen Trainingsdaten, die für GPT verwendet werden, umfassen eine breite Palette von Textquellen. Dadurch kann das Modell Muster und Zusammenhänge in verschiedenen Sprachstilen und -formaten erkennen. Diese Vielfalt in den Trainingsdaten ist entscheidend für die Entwicklung einer KI, die in der Lage ist, sich an verschiedene Sprachanforderungen anzupassen – von formellen akademischen Texten bis hin zu informellen Gesprächen.
Ein weiterer wichtiger Aspekt von GPT ist seine Fähigkeit zur Generalisierung. Nach dem Training kann das Modell neue, unbekannte Texte effektiv interpretieren und darauf reagieren. Diese Generalisierungsfähigkeit macht GPT ideal für eine Reihe von Anwendungen, wie die Texterstellung, Übersetzung und insbesondere für die automatisierte Dokumentation. Automatisierte Dokumentationssysteme, die auf GPT basieren, könnten nicht nur Zeit sparen, sondern auch die Qualität und Konsistenz der Dokumentation verbessern [1].

ChatGPT: Ein Werkzeug für die Automatisierung der KI-Dokumentation

Die Kombination von Transformer- und GPT-Technologien in ChatGPT ermöglicht die Entwicklung eines Tools zur automatisierten Dokumentation von KI-Modellen und Data Assets. Dieses Tool kann nicht nur technische Details erfassen, sondern auch den Anwendungskontext, Leistungsparameter und mögliche Einschränkungen eines KI-Modells dokumentieren.

Vorteile eines ChatGPT-basierten Dokumentationstools

- Zeitersparnis und Effizienz: Automatisierung der Dokumentation spart Zeit für Data Scientists und KI-Entwickler.
- Konsistenz und Qualität: Ein automatisiertes Tool sorgt für eine standardisierte und hochwertige Dokumentation.
- Verbesserte Verständlichkeit: Fortschrittliche Sprachmodelle erleichtern die Darstellung komplexer technischer Konzepte in verständlicher Sprache.
- Aktualisierung und Wartung: Das Tool kann kontinuierlich aktualisiert werden, um mit den neuesten KI-Forschungsentwicklungen Schritt zu halten.

Fazit

Die Entwicklung eines ChatGPT-basierten Dokumentationstools für KI-Modelle und Data Assets ist ein wichtiger Fortschritt in der KI-Community. Es nutzt die neuesten Forschungsergebnisse, wie die Transformer-Architektur und GPT-Modelle, um einen manuellen und zeitaufwendigen Prozess zu vereinfachen. Dieses Tool verbessert nicht nur die Effizienz und Qualität der Dokumentationsprozesse, sondern macht auch KI-Technologien für ein breiteres Publikum zugänglicher. In einer Welt, in der KI zunehmend wichtig wird, ist die Bedeutung einer solchen Innovation groß.

Literatur und Abbildungen

[1] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving Language Understanding by Generative Pre-Training. In *Improving Language Understanding by Generative Pre-Training*. OpenAI, 2018.

[2] Ashish Vaswani et al. Attention Is All You Need. *CoRR*, 2017.

[3] Matt Walsh. ChatGPT Statistics (2023) — The Key Facts and Figures. https://www.stylefactoryproductions.com/blog/chatgpt-statistics, 2023.

Untersuchung des Segment Anything Model in der industriellen Bildverarbeitung

Daniel Osswald **Markus Enzweiler**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma Robert Bosch Manufacturing Solutions GmbH, Stuttgart

Einleitung und Problemstellung

In den Bereichen autonomes Fahren, Robotik und Medizin wird KI-basierte Bildverarbeitung aufgrund ihrer Generalisierungs- und Adaptionsfähigkeit bereits in großem Umfang erfolgreich eingesetzt. Da die industrielle Bildverarbeitung hingegen häufig in konfigurierbaren und genau definierten Umgebungen stattfindet, in denen beispielsweise Beleuchtung, Perspektive, Hintergründe und Motive bekannt sind und sich nicht ändern, war ein KI- oder ML-basiertes Verfahren bisher oft nicht notwendig. Durch Bestrebungen wie Losgröße 1, Fluide Produktion oder auch Bin Picking wird die Produktion flexibler und benötigt daher auch neue Verfahren, um diesen Anforderungen gerecht zu werden.
Bei der Bildsegmentierung werden benachbarte Pixel, die ein bestimmtes Kriterium erfüllen, zu zusammenhängenden Regionen zusammengefasst [3]. Die Segmentierung ist ein Teilprozess der Bildanalyse und wird oft von weiteren Verarbeitungsschritten wie der Klassifikation begleitet. Verfahren, die auf neuronalen Netzen basieren, liefern bei komplexen und sonst schwer zu verarbeitenden Bilddaten bessere Ergebnisse als herkömmliche Verfahren.

Zielsetzung

Ziel der Arbeit ist es, bestehende Verfahren zu verbessern und neue Anwendungsfälle für das Segment Anything Model (SAM) zu untersuchen. Insbesondere soll im industriellen Umfeld die pixelgenaue Segmentierung von komplexen Maschinenbauteilen sowie Ersatzteilen mit Hilfe von SAM betrachtet werden. Zusätzlich sollen eventuelle Stärken und Einschränkungen von SAM ermittelt werden, wobei auch nach einer Möglichkeit zur automatischen Generierung des Prompts gesucht werden soll. Zuletzt soll auch die qualitative Auswirkung von Fine-Tuning auf spezifische Anwendungsfälle analysiert werden.

Segment Anything Model

SAM ist ein Bildsegmentierungsmodell, das im Jahr 2023 von Kirillov et al. (Meta AI) entwickelt wurde [4]. Das Modell wurde anhand des momentan größten Segmentierungsdatensatzes SA-1B mit 11 Millionen Bildern und 1,1 Milliarden Masken trainiert. Dank der Größe des Datensatzes und der Fähigkeit zum Prompting kann das Modell problemlos auf neue Bildverteilungen und Aufgaben übertragen werden, ohne zusätzliches Training zu benötigen. Es kann sehr starke Zero-Shot-Ergebnisse auf verschiedenen Bilddaten erzielen.
Die Architektur von SAM besteht aus drei Netzwerken: einem Image-Encoder, einem Prompt-Encoder und einem Mask-Decoder. Abbildung 1 zeigt die Architektur und den Informationsfluss von SAM.

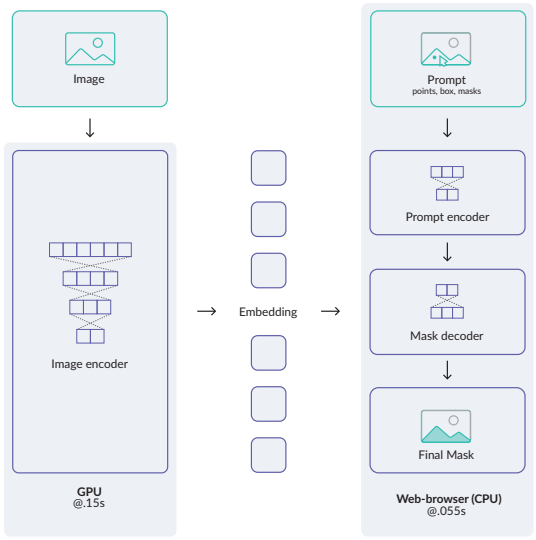


Abb. 1: Darstellung der Modellarchitektur von Segment Anything mit den drei integrierten Netzwerken [4]

Der verwendete Image-Encoder basiert auf einem vortrainierten Vision-Transformer [2], mit Anpassungen

zur Verarbeitung von hochauflösenden Bildern. Er wandelt ein Eingabebild in ein Bild-Embedding um. Dieses Embedding beinhaltet die generalisierte räumliche und inhaltliche Information des Bildes. Das Berechnen des Embedding des Bildes ist der rechenintensivste Schritt und muss nur einmal pro Bild durchgeführt werden. Anschließend können durch verschiedene Prompts Masken generiert werden, ohne dass das Embedding erneut berechnet werden muss.
Der Prompt-Encoder kann Punkte, Bounding-Boxen, Masken und Text in ein Prompt-Embedding umwandeln. Nach Berechnung des Bild- und Prompt-Embeddings kann der Mask Decoder aus beiden Informationen eine binäre Maske erzeugen. Sowohl der Promp-Encoder als auch der Mask-Decoder sind sehr leichtgewichtige Modelle, die innerhalb weniger Millisekunden auf dem Computerprozessor berechnet werden können, ohne dafür eine Hardware-Beschleunigung zu benötigen.

Abb. 2: Vergleich der Ausgabe zwischen der Mask-Predictor Funktion (oben) und dem Automatic Mask Generator (unten) [1]

SAM bietet insgesamt zwei Funktionen (vgl. Abbildung 2). Der Mask Predictor ermöglicht die Generierung einer Segmentierungsmaske für ein bestimmtes Objekt in einem Bild anhand eines vorgegebenen Prompts.
Der Automatic Mask Generator ermöglicht es automatisch Masken für alle Objekte in einem Bild zu erzeugen, ohne dafür einen Prompt zu benötigen. Hierbei wird ein Raster von Punkt-Prompts auf dem Bild verteilt und diese dienen jeweils als Eingabe für den Mask Predictor. Anschließend werden die erzeugten Masken hinsichtlich Qualität und Ähnlichkeit gefiltert, um unsichere und redundante Masken auszuschließen.
Die durch SAM erzeugten Masken beinhalten keine Klassifikation, sondern zeigen lediglich an, welche Pixel zu dem gesuchten Objekt gehören, ohne zu sagen, um welche Art von Objekt es sich handelt, folglich ist SAM ein generisches Segmentierungsmodell, das für jegliche Art von Objekten eingesetzt werden kann.
Basierend auf SAM als bestehendes Foundation Model gibt es bereits viele darauf aufbauende Projekte wie beispielsweise „Segment Anything in High Quality“ zur Segmentierung von feinen Strukturen, „Matting Anything“ zum Erstellen von Alphamasken oder auch „FastSAM“ wodurch eine deutlich geringere Laufzeit bei vergleichbaren Ergebnissen erzielt wird.

Ergebnisse

SAM ist in der Lage, komplexe Objekte präzise zu segmentieren. Besondere Stärken zeigt das Modell bei der Segmentierung transparenter Objekte unter Berücksichtigung von Spiegelungen, Schatten und komplexen Hintergründen.
Limitationen offenbart SAM jedoch bei dünnen Objekten wie beispielsweise Kabeln sowie beim Umgang mit Löchern in Objekten. Kabel werden nicht vollständig segmentiert und weisen Fehler auf. Löcher in Objekten werden vom Modell nicht ausgeschnitten und sind daher oft irrtümlich als Teil der Maske dargestellt.
Um die Verarbeitung durch SAM zu erleichtern, empfiehlt es sich, hochauflösende Bilder auf die Größe des zu segmentierenden Objekts zurechtzuschneiden. Die Eingabegröße des Bildencoders beträgt 1024x1024 Pixel, während die erzeugte binäre Maske nur eine Auflösung von 256x256 Pixel hat. Die Masken hochauflösender Bilder sind aufgrund ihrer zu geringen Auflösung nicht in der Lage, detaillierte Strukturen darzustellen. Während der Interpolation zur Wiederherstellung der ursprünglichen Auflösung gehen durch die Glättung feine Strukturen verloren.

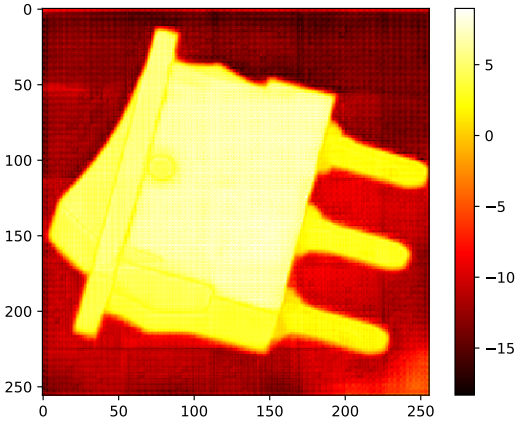


Abb. 3: Heatmap-Darstellung der Mask Decoder Rohausgabe, welche die Wahrscheinlichkeit der Pixel-Zugehörigkeit zu der Maske anzeigt. [1]

Die Rohausgaben, die durch den Mask-Decoder erzeugt werden und die Wahrscheinlichkeit der Pixel-Zugehörigkeit zu den Masken anzeigen, können durch Verwendung eines Schwellwertverfahrens genutzt werden, um die Binärmasken geringfügig anzupassen (vgl. Abbildung 3).
Da SAM keine eigenständige Objektlokalisierung durchführt und stattdessen jeweils einen entsprechenden Prompt benötigt, ist es hilfreich, weitere Tools in Kombination zu nutzen, um automatisch Prompts zu den Objekten zu erzeugen. Hierbei hat sich unter anderem die Verwendung eines auf Maschinenbauteile trainierten YOLOv7-Modells als nützlich erwiesen, welches lediglich die Objectness von Bildbereichen bewertet, ohne eine Klassifikation durchzuführen. Die resultierenden Bounding Boxen können dann als Inputprompt für SAM verwendet werden.
Mithilfe eines Fine-Tuning des SAM Mask Decoder Modells auf Maschinenbauteilen ist es zudem möglich die erzeugen Masken qualitativ zu verbessern. Dadurch können auch bisherige Limitationen wie beispielsweise Kabel oder auch Objekte mit Löchern weiter optimiert werden.

Ausblick

Zukünftige Einsatzmöglichkeiten von SAM können in der Vorverarbeitung von Klassifikationsnetzen liegen. Dabei werden mit Hilfe von SAM Objekte ausgeschnitten, um ein robusteres Eingangsbild ohne störenden Hintergrund zu erhalten. Diese Funktion kann auch direkt in der Fertigungsanlage genutzt werden, um relevante Segmente aus einem Bild zu isolieren und anschließend Fehler oder Mängel durch den Vergleich mit einem Referenzbild zu erkennen. Es ist ebenfalls denkbar, SAM als Edge-Service zu nutzen, um in einer mobilen Anwendung Bauteile über Benutzerprompts zu segmentieren und anschließend über eine Klassifikations- oder Bewertungsfunktion die Bezeichnung oder Informationen über die Funktion des Bauteils zu erhalten.
Zudem eignet sich SAM im Bereich E-Commerce, um Produktbilder automatisch auszuschneiden, ohne auf kostenpflichtige Services oder eine manuelle Freistellung mittels Bildbearbeitungsprogramm zurückgreifen zu müssen.

Literatur und Abbildungen

[1] Eigene Darstellung.

[2] Alexey Dosovitskiy et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. https://arxiv.org/abs/2010.11929v2, 06 2021.

[3] Landgericht Düsseldorf. Urteil vom 14.08.2007 - 4a O 235/06. http://www.justiz.nrw.de/nrwe/lgs/duesseldorf/lg_duesseldorf/j2007/4a_O_235_06urteil20070814.html, 2007.

[4] Alexander Kirillov et al. Segment Anything. https://arxiv.org/abs/2304.02643v1, 04 2023.

Design und Implementierung einer Web-App zur Messung des Blutdrucks.

Darios Pachtsinis **Rainer Keller**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma SYSTECS Informationssysteme GmbH, Leinfelden-Echterdingen

Einleitung

Die herkömmliche Methode zur Blutdruckmessung mittels einer Oberarm-Manschette ist seit Jahrzehnten bereits etabliert und ist der üblichste Weg den Blutdruck im Haushalt oder sogar in Gesundheitseinrichtungen zu messen. Mit dem fortschreitenden Wachstum der Technologie und dem bereitstehenden, aufgebauten Wissen benötigt die Forschung innovative Ansätze, die alte Wege ablösen und durch neue, einfachere Möglichkeiten ersetzen. So kann man mit anderen Techniken über beispielsweise eine Apple Watch bereits die Herzfrequenz messen, indem man nur den Finger an einen Sensor hält und eine gewisse Zeit abwartet. Dies verbindet die Simplizität des Messvorgangs mit einer kürzeren Vorbereitungsdauer, insofern man die Apple Watch trägt. Mithilfe des MAX30101 ist ebenso ein Vorgang der Herzfrequenzmessung möglich, allerdings noch keine Ermittlung des Blutdrucks, welcher mit zwei Werten während der Herzkontraktion und zwischen den Herzschlägen beschrieben wird.

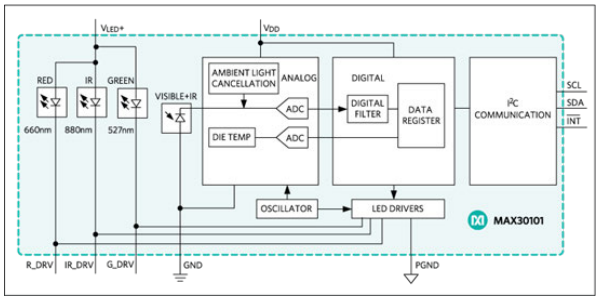


Abb. 1: Graphische Übersicht des MAX30101 [4]

Ziel der Arbeit

Der MAX30101 Sensor basiert auf der Photoplethysmographie und ist für die kontaktlose Messung des Herzschlags, sowie des Sauerstoffgehalts im Blut verantwortlich. Das primäre Ziel dieser Arbeit ist es, dass der Sensor zur Ermittlung des Blutdrucks genutzt werden kann, da dieser in erster Linie nur die Herzfrequenz bereitstellt. Dies bedeutet, dass das Projekt den Beweis für den Zusammenhang zwischen Blutdruck und Herzschlag liefern soll, als auch die Umsetzung der eigentlichen Messung. Anschließend gilt es die Ergebnisse anschaulich in einer Webapplikation darzustellen, sodass der zukünftige Nutzer am Ende des Projekts seinen Blutdruck messen und die dazugehörigen Werte graphisch oder textuell einfach nachsehen kann.

MAX30101 Sensor

Der MAX30101 Sensor ist ein Pulsoximeter und Herzfrequenzsensor und verwendet die Photoplethysmographie (PPG), welche die Veränderung des Blutvolumens misst. Dies geschieht, indem ein optischer Sender Infrarotlicht einer bestimmten Frequenz in die Haut abschickt und eines optischen Empfängers, welcher gesendetes Licht empfängt [3]. Anhand der Messungen durch PPG kann der Sensor die Herzfrequenz ermitteln, indem jede Bewegung im Blut als ein Herzschlag interpretiert wird. Zur einfachen Anwendung des Sensors wird dieser mit einem RaspBerry PI verbunden und somit kann der dazugehörige Python Code editiert und getestet werden. Um einen Wert zu messen, muss der Anwender eine Körperstelle, im Idealfall einen Finger, an den eigentlichen Sensor anbringen, wenige Sekunden warten und kann anschließend die gemessenen Werte in der Konsole ablesen. Bislang ist es möglich über ein Jupyter Notebook einen Graphen darzustellen, um so die erhaltenen Werte über einen definierten Raum einzusehen.

Abb. 2: Bild eines MAX30101 [4]

Erklärung Herzfrequenz und Blutdruck

Der Herzfrequenz ist die Anzahl an Herzschlägen pro Minute [1], um das genauste Ergebnis zu liefern, wird die Herzaktivität für eine Minute gemessen. Der heutige Standard allerdings sagt voraus, dass die Überwachung des Herzens für entweder zwölf oder fünfzehn Sekunden ausreicht und anschließend multipliziert wird, sodass am Ende ein Messwert für 60 Sekunden entsteht. Der Blutdruck hingegen besteht aus zwei Werten und beschreibt den Druck innerhalb des kardiovaskulären Systems [2]. Systolischer Blutdruck ist der erste und höhere von zwei Werten, welcher für den maximalen Blutdruck während einer Systole, also der Herzauswurfphase, steht. Der zweite Wert ist der diastolische Blutdruck und niedriger als der erste Wert, denn dieser bezeichnet den Druck, während das Herz sich mit Blut füllt, also während einer Diastole. Außerdem wird der Blutdruck in Millimeter Quecksilbersäule (mmHg) gemessen und ein durchschnittlicher Wert wäre 120/80 mmHg.

Vom Herzschlag zum Blutdruck

Hier befindet sich der momentane Endpunkt heutiger Forschung und Entwicklung. Der Puls und der Blutdruck sind zwei wichtige Messungen, welche allerdings unabhängig voneinander sind und somit nicht durch die Existenz der einen Variable berechnet werden können. Die bisherige Alternative um den Blutdruck zu messen, ist bislang kostenintensiver, da sie eine multimodale Sensortechnik verwendet, also der Verbindung zweier Sensoren, in diesem Falle eines MAX30101 zur Ermittlung des Herzschlags und ein weiterer Sensor, welche den Blutdruck oder EKG-Werte liefert und auswertet. Ebenfalls gibt es die Möglichkeit, welche Hauptbestandteil der Abschlussarbeit sein soll, nämlich dass man die Pulse Transit Time (PTT) misst und somit Rückschlüsse auf den Blutdruck schließt. So eine Umrechnung ist allerdings nicht trivial und kann somit individuell variieren, sowie die Genauigkeit beeinflussen. Ein lineares Modell zur Schätzung des Blutdrucks basiert auf der Annahme, dass PTT und Blutdruck invers proportional zueinander sind, das bedeutet, sollte der PTT steigen, sinkt der Blutdruck und umgekehrt. Diese theoretische Erkenntnis wird in der Abschlussarbeit genutzt, um einen Schätzwert des Blutdrucks zu ermitteln und verfeinert, sodass ein herkömmliches Blutdruckmessgerät wenig Unterschied zeigt. In der Theorie ist es möglich mit der gemessenen Zeit zwischen den Herzschlägen einen ungefähren Wert für den Blutdruck zu ermitteln, allerdings besteht dieser ja aus 2 Werten, wodurch die eigentliche Schwierigkeit entsteht: Die Analyse und Umrechnung von Puls zu Blutdruck.

Ausblick

Die Einführung des MAX30101 Sensors als Methode zur Blutdruckmessung könnte die Patientenversorgung revolutionieren und somit die Haushaltsnutzung vereinfachen, schließlich basiert die Messung des Blutdrucks auf einer Theorie anhand des Pulses, da die Abhängigkeit von beiden ausgeschlossen ist. Sollte eine indirekte Abhängigkeit möglich sein, könnte diese Abschlussarbeit Einblicke gewähren, inwiefern die Integration moderner Sensortechnologie hilfreich für die Gesundheitsvorsorge ist, sowie die Ersparnis, keine multimodalen Sensoren verwenden zu müssen.

Literatur und Abbildungen

[1] Frank Antwerpes. Herzfrequenz. https://flexikon.doccheck.com/de/Herzfrequenz, 08 2020.

[2] Frank Antwerpes. Blutdruck. https://flexikon.doccheck.com/de/Blutdruck, 04 2023.

[3] Nils Beckmann. Photoplethysmographie-basierte Messung der Pulswellenlaufzeit fur die Emotionserkennung. https://duepublico2.uni-due.de/servlets/MCRFileNodeServlet/duepublico_derivate_00071182/Diss_Beckmann.pdf, 12 2019.

[4] Eigene Darstellung.

Intelligenter Algorithmus für die Einteilung der Prüfungsaufsichten der Fakultät IT

Kuntal Patel **Mirko Sonntag**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt an der Fakultät Informatik und Informationstechnik, Esslingen

Einleitung

Angesichts der rasanten Entwicklung von Algorithmen und ihrer vielfältigen Anwendungen ist es von zentraler Bedeutung, auch im Bildungsbereich innovative Lösungen zu finden. Die Einteilung der Prüfungsaufsichten an der Fakultät für IT der Hochschule Esslingen stellt eine Herausforderung dar, die durch die Implementierung eines intelligenten Algorithmus möglicherweise effizienter überwunden werden kann. In diesem Kontext entfällt der bisherige manuelle Aufwand bei der Zuweisung der Prüfungsaufsichten an Professoren und Mitarbeitende. Die manuelle Eintragung der Professoren und Mitarbeitenden in ein Excel-Sheet für die Aufsichtsplanung gestaltet sich aufgrund der Vielzahl von über 100 Prüfungen äußerst zeitintensiv und unübersichtlich. Der Algorithmus verfolgt nicht nur das Ziel, diesen Prozess zu digitalisieren, sondern strebt auch an, zusätzliche Vorteile zu bieten. Die Reduzierung des E-Mail-Aufwands für Professoren und Mitarbeitende durch die Effizienz des Algorithmus führt zu einer Zeitersparnis. Die digitale Lösung ermöglicht darüber hinaus eine zügigere und effizientere Kommunikation bei Änderungen im Aufsichtsplan. Die automatisierte Einteilung trägt dazu bei, mögliche Fehler bei der manuellen Zuweisung zu minimieren und gewährleistet somit eine akkurate und faire Verteilung der Aufsichten.

Motivation der Arbeit

Die Idee besteht darin, eine Web-Anwendung zu entwickeln, bei der die Aufsichten ihre Abwesenheiten angeben können. Dadurch soll festgestellt werden, welcher Professor und Mitarbeitende wann die Prüfungsaufsicht übernehmen kann. Der Algorithmus soll die eingegebenen Daten verarbeiten und eine gerechte sowie gleichmäßige Zuteilung der Prüfungsaufsichten ermöglichen. Das Projekt ist in zwei Hauptteile unterteilt: Erstens die Implementierung einer Full-Stack-Anwendung zur Planung von Prüfungsaufsichten [5], wobei der Fokus auf der Entwicklung der Web-Anwendung liegt. Es umfasst sowohl ein Frontend als auch ein Backend. Zweitens die Entwicklung eines intelligenten Algorithmus zur Einteilung der Prüfungsaufsichten, wobei der Schwerpunkt auf der Entwicklung des Algorithmus liegt. Diese Arbeit baut auf dem ersten Teil des Projekts auf und konzentriert sich speziell auf die Entwicklung des Algorithmus, der am Ende im Backend der Web-Anwendung integriert werden soll.

Zielsetzung und Herausforderungen

Das übergeordnete Ziel dieser Bachelorarbeit ist es, die Machbarkeit und Effektivität eines intelligenten Algorithmus zur Einteilung der Prüfungsaufsichten zu untersuchen. Konkrete Forschungsfragen beinhalten die Analyse relevanter Algorithmen sowie deren Anpassung an die spezifischen Anforderungen der IT-Fakultät für die Prüfungsaufsichtseinteilung. Die Entwicklung des Algorithmus für die Einteilung der Prüfungsaufsichten steht vor verschiedenen Herausforderungen, wie zum Beispiel: Die optimale Berücksichtigung der Anwesenheit der Professoren und der Mitarbeitenden, die Festlegung angemessener Pausen zwischen den Aufsichten, die gleichmäßige Verteilung der Professoren und Mitarbeitenden, sowie die effiziente Zuweisung von Reserven. Der Optimierungsalgorithmus muss parametrisierbar genug sein, um diese Bedingungen zu berücksichtigen und unter verschiedenen Szenarien ein bestmögliches Ergebnis zu liefern.

Ansatz

Damit der Optimierungsalgorithmus eine gleichmäßige Zuweisung der Aufsicht vornehmen kann, müssen mehrere Informationen als nur die Abwesenheiten der Professoren und der Mitarbeitenden berücksichtigt werden. Zuerst müssen daher alle Randbedingungen festgelegt werden, die der Algorithmus beachten muss. Die Randbedingungen sind in eine sogenannte 'MoSCoW-Priorisierung' aufgeteilt. Mit Hilfe der MoSCoW-Methode können Ziele und Anforderungen

nach Must-, Should-, Could- und Would-Zielen priorisiert werden [3]. Diese Methode gibt einen klaren Überblick darüber, welche Randbedingungen priorisiert werden müssen.
Als nächstes ist es wichtig, theoretisches Wissen im Bereich des Optimierungsalgorithmus zu erlangen. Welche Optimierungsalgorithmen relevant sind und wie sie an die spezifischen Anforderungen der Fakultät für die Prüfungsaufsichtseinteilung angepasst werden können, müssen untersucht werden. Der erste Entwurf des Optimierungsalgorithmus ist in Form eines Entscheidungsbaums erstellt. Entscheidungsbäume sind Diagramme, die die zu treffenden Entscheidungen, die verschiedenen möglichen Szenarien und alle möglichen Ergebnisse darstellen. Ein Entscheidungsbaum dient als Entscheidungshilfe, bei der mehrere mögliche Szenarien berücksichtigt werden müssen [2]. Der Entscheidungsbaum besteht aus einem Wurzelknoten, Verzweigungen, Entscheidungsknoten, Zufallsknoten und Endknoten.
Der erstellte Optimierungsalgorithmus verfolgt dieses Konzept, um die Aufsichten zu verteilen. Dieser Optimierungsalgorithmus besteht aus mehreren Schritten. Dabei verfolgt jeder dieser Schritte das Konzept des Entscheidungsbaumes. Die Namen werden Schritt für Schritt nach der Verfügbarkeit und anderen Bedingungen gefiltert und im nächsten Schritt übergeben. Im letzten Schritt sind dann die Namen zu sehen, die der Optimierungsalgorithmus nach den Randbedingungen gefiltert hat.
Zuletzt soll das gelieferte Ergebnis überprüft werden. Dabei soll das Verfahren des Unit Testings verwendet werden. Unit Testing sind Testverfahren für die Softwareentwicklung, bei dem sich eine Unit auf eine einzelne Komponente bezieht, die getestet werden muss, um die Qualität des Codes zu ermitteln [1]. Das Endergebnis wird also an verschiedenen Szenarien überprüft, wie zum Beispiel, ob die Professoren und Mitarbeitenden gleichmäßig für die Aufsichten verteilt sind oder ob die Professoren und Mitarbeitenden gleichzeitig für mehrere Prüfungen eingeteilt werden.

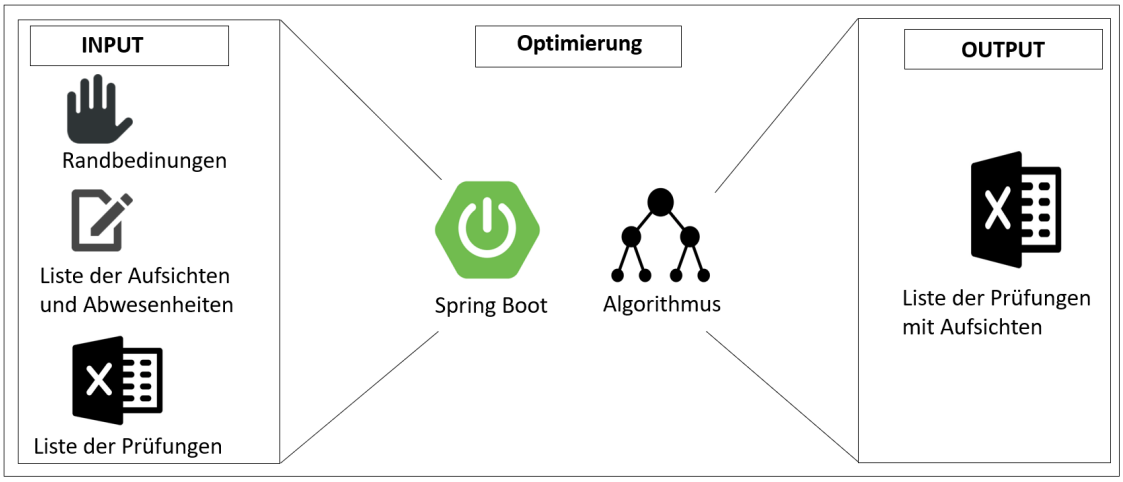


Abb. 1: Veranschaulichung der Arbeit [4]

Ausblick

Der Algorithmus soll schließlich im Backend der Web-Anwendung integriert werden [5]. Zusätzlich besteht die Idee, den Algorithmus mithilfe der von Java bereitgestellten Bibliothek namens Weka zu implementieren. Weka ist eine auf Java basierende Open-Source-Bibliothek für maschinelles Lernen. Dabei soll der J48-Algorithmus in Weka genutzt werden, welcher sich für die Erstellung von Entscheidungsbäumen eignet.

Literatur und Abbildungen

[1] Durga Prasad Acharya. Unit Testing erklärt. https://geekflare.com/de/unit-testing-guide/, 2023.

[2] Dr. Benjamin Anderson. Entscheidungsbaum. https://statorials.org/de/entscheidungsbaum/?utm_content=cmp-true, 2023.

[3] Windolph Andrea. Wie du Ziele mit der MoSCow-Methode priorisierst. https://projekte-leicht-gemacht.de/, 2017.

[4] Eigene Darstellung.

[5] Celine Schuster. Konzeption und Implementierung einer Full-Stack-Anwendung zur Planung von Prüfungsaufsichten, 2023.

Design and Evaluation of an Intrusion Detection System for Time Sensitive Networks

Lukas Popperl **Tobias Heer**

Department of Computer Science and Engineering, Esslingen University

Work carried out at Hirschmann Automation and Control GmbH, Neckartenzlingen

Motivation and Problem

Production lines often consist of industrial robots for factory automation. These robots rely on time-critical communication that enables challenging synchronized processes such as robot arm movement or automated welding. For a stable production process, it is essential to protect and enable the necessary network Quality of Service (QoS) guarantees. In the past, this was achieved by isolating and separating the automation networks from the factory networks. The IEEE 802.1 Time Sensitive Networking (TSN) standards offer the necessary QoS protection for combining time-critical and best effort traffic in the same network. Hence, the time-critical traffic does not need to be isolated for the QoS guarantees. However, reduced isolation leads to an increased attack surface.

Factory automation networks use monitoring solutions like Intrusion Detection Systems (IDS) to detect network attacks. These IDSs detect malicious traffic by analyzing the content of packets and traffic patterns. There exists a multitude of threats and possible attacks against TSN [2]. Most of the attacks disturb time-critical traffic or manipulate the time synchronization. This is often achieved without altering packet content and, therefore, can not be detected by traditional IDSs. The goal of this work is to design and evaluate an IDS that is capable of detecting attacks on time-critical traffic. We achieve this goal by proposing and evaluating four additional data sources for our IDS.

Design

Figure 1 shows the architecture of the IDS. The IDS relies on an anomaly-based detection, i.e., the IDS detects attack that deviate from the learned *normal* behavior. Based on the analysis of the detection, the IDS classifies the traffic into normal or abnormal. In order to detect attacks on time-critical traffic, we suggest four new data sources. These new data sources can reveal the required information to detect these sophisticated attacks.

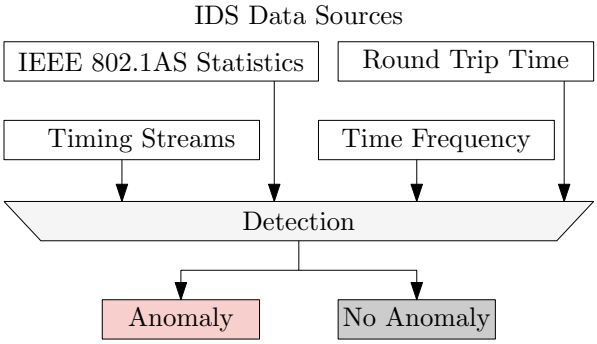


Fig. 1: Design Overview of the IDS [8]

New Data Sources

The following section describes the new data sources in detail and outlines how they can be used to detect attacks on time-critical traffic.

Traffic Capture Network devices like switches offer the possibility to mirror network traffic and then capture it. Today, IDSs use this capturing for packet inspection. We extend on the packet inspection by adding time-frequency analysis. Deviations to the time-frequency can be caused by collisions with other traffic, e.g., injected packets with high priority. Based on the QoS mechanism in the network, this indicates an injection attack or problems with the time synchronization.

Round Trip Time Time-critical networks are deterministic, meaning the time it takes for a packet to travel through the network is static and are known beforehand. The deterministic behavior allows us to detect network attacks like spoofing by measuring timing delays through OSI-Layers. Some examples include the comparison of the round trip time from, e.g., OPC-UA-Requests.

Timing Streams Some switches offer the possibility to timestamp network packets. The use of this mechanism

results in a trace of timestamps through the network. These traces enable delay measurements of network segments. Deviations to these measurements can indicate an attack.

802.1AS Statistics Time-sensitive traffic relies on high-precision time synchronization. The necessary accuracy is achieved using time synchronization standards like the IEEE 802.1AS generalized Precision Time Protocol (gPTP) which ensures time synchronization with accuracy of less than one microsecond. The gPTP offers many measurements and statistics that we will use to detect attacks on the time synchronization.

Detection

The classification of anomalous traffic occurs on the basis of defined thresholds. The IDS sets these thresholds on the basis of a baselining phase or if possible a model based approach. The IDS uses the aforementioned data sources as information input and calculates statistics like of the mean, standard deviation, median or quartile percentages. These statistics are easy to calculate, and deviations to the thresholds are anomalous. However, depending on the capabilities of the attacker, it is not sufficient to solely rely on one statistic. Figure 2 visualizes this problem and shows the probability density function (PDF) of three different distributions.

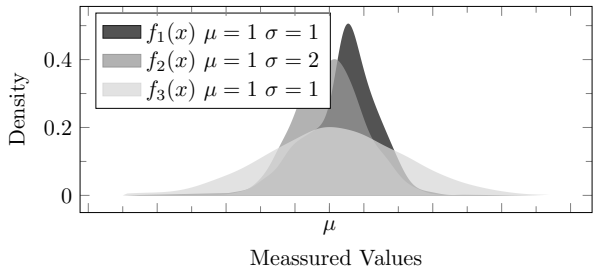


Fig. 2: Three Probabilty Densitiy Functions with same mean [8]

Each of the displayed distribution have the same mean, while the standard deviations or certain percentile are different. In two of the outlined example PDFs, an anomaly would have not been detected by only calculating one or even two statistics. The fabrication of such anomalous distributions require an extremely advanced attacker. Detecting such advanced attackers and anomalies requires the combination of multiple values or advanced metrics. One such metric is the Kullback-Leibler divergence (KLD). The effectiveness of the KLD based anomaly detection is analyzed according to the following evaluation.

Evaluation

We evaluate the effectiveness of the Kullback-Leibler based anomaly detection in a small test environment that consists of multiple communication endpoints and different real-time critical applications. Each deployed application will have their own requirements regarding latency, deadlines, and packet loss [5].

An overview of possible attacks on TSN is provided in [2]. We will implement some existing attacks on time-sensitive networks to determine the effectiveness of the IDS.

Spoofing Attacks The attacker impersonates the time master by leveraging the Best Master Clock Algorithm (BMCA) or impersonates end-devices with, e.g., ARP spoofing. Being the time master allows the attacker to specify the time inside the network, whereas ARP spoofing enables eavesdropping and packet modification.

Injection Attacks The attacker is an accepted peer in the network, and either injects excessive high priority traffic or promotes existing low priority traffic. Both cases allow the attacker to delay existing high priority traffic as the injected traffic competes for the same time window.

Time-Synchronization Attacks The attacker targets the gPTP by delaying or modifying packets. This creates a dynamic or static time offset between two network segments. These attacks require control of the gPTP stack on the switches or a man-in-the-middle attacker.

The data sources function independently of the QoS mechanisms in the network. They have their own advantages and disadvantages regarding availability, cost, and effectiveness. It is part of our evaluation to assess the data sources based on these parameters and the combination of these data sources. For example, if an attacker executes a time-synchronization attack, the data source "Timestamp in Payloads" (c.f. Section Timing Streams) can detect the position in the network where this attack occurs. In case of an attack, the timestamps of neighboring switches show offsets, which do not fit the current topology and traffic. However, the IDS can determine that it is not a congestion but a time-synchronization problem only in combination with the "RTT" (c.f. Section RTT) which detects stable end-to-end latencies. Based on this evaluation, we propose a combination of data sources to achieve a reliable and QoS independent IDS for time-critical traffic.

Related Work

Even though IDS are an important part of securing a network, there exists limited work concerning IDS and the security of time-critical traffic.

Nascimento has designed and developed a functional IDS for AVB/TSN [7]. He analyzed different attacks on TSN protocols like the gPTP and the Audio Video Transport Protocol (AVTP). However, the implementation is limited by the fact that it relies on a specific switch feature. In contrast to our work where only the payload timestamping is not widely supported by all switches. Nevertheless, the attacks in this work are useful and can be applied to our work.
Lou et al. [4] and Meyer et al. [6] developed an anomaly based detection system which relies on Per-Stream Filtering and Policing (PSFP) for attack detection on TSN. Both authors evaluated their IDS in a OMNeT++ simulation, where the IDS identified and discarded all different abnormal traffic events. However, this work is limited by the assumption that PSFP is supported on every port of every switch in the network.
Ergenç et al. [3] propose an open source IDS for TSN called TSNZeek. The IDS is capable of detecting attacks on the Stream Reservation Protocol (SRP) and on the fault tolerance mechanism, Frame Replication and Elimination for Reliability (FRER). Unfortunately, this work is not applicable to generic timing attacks. Additionally, it relies only on port mirroring, in contrast to our approach which uses a combination of different methods.
An IDS against rouge master attacks on the gPTP is described in [1]. This signature-based detection approach relies on a random forest Machine Learning (ML) model. With the generated training data, the IDS could detect rouge master attacks with high accuracy. Unfortunately, this work fails to address all other existing attacks against gPTP.

Results

The result of this work will be a proof of concept IDS for time-critical traffic. The detection should be achieved without dependencies on the used TSN mechanisms in the switches. The necessary information for the attack detection is generated using the proposed data sources. This will be analyzed and assessed using the outlined evaluation.

References and figures

[1] Alessio Buscemi et al. An Intrusion Detection System Against Rogue Master Attacks on gPTP. In *IEEE Vehicular Technology Conference (VTC2023-Spring)*. IEEE, 2023.

[2] Ergenc Doganalp, Bruelhart Cornelia, Neumann Jens, Krueger Leo, and Fischer Mathias. On the Security of IEEE 802.1 Time-Sensitive Networking. In *IEEE International Conference on Communications Workshops (ICC Workshops)*. IEEE, 2021.

[3] Doğanalp ERGENÇ, Robin SCHENDERLEIN, and Mathias FISCHER. TSNZeek: An Open-source Intrusion Detection System for IEEE 802.1 Time-sensitive Networking. In *3rd International Workshop on Time-Sensitive and Deterministic Networking (TENSOR)*. IFIP, 2023.

[4] Luo Feng, Wang Bowen, Fang Zihao, Yang Zhenyu, and Jiang Yifan. Security Analysis of the TSN Backbone Architecture and Anomaly Detection System Design Based on IEEE 802.1Qci. *Security and Communication Networks*, page 17, 2021.

[5] Time Sensitive Networking Testbed Industrial Internet Consortium. Time Sensitive Networks for Flexible Manufacturing Testbed Characterization and Mapping of Converged Traffic Types. https://www.iiconsortium.org/pdf/IIC_TSN_Testbed_Char_Mapping_of_Converged_Traffic_Types_Whitepaper_20180328.pdf, 2019.

[6] Philipp Meyer, Timo Häckel, Sandra Reider, Franz Korf, and Thomas C. Schmidt. Network Anomaly Detection in Cars based on Time-Sensitive Ingress Control. In *92nd Vehicular Technology Conference VTC2020-Fall*. IEEE, 2020.

[7] Rodrigo Antônio Alves do Nascimento. Design and Development of IDS for AVB/TSN, 2019.

[8] Own representation.

Projektmanagement bei Mercedes-Benz- Herausforderungen und Erfolgsfaktoren der Zusammenarbeit internationaler Projektteams

Tina Rasheed **Astrid Beck**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma Mercedes-Benz AG, Stuttgart

Einleitung

Die Automobilindustrie, insbesondere Unternehmen wie Mercedes-Benz, operieren in einem globalen Umfeld, in dem effizientes Projektmanagement von entscheidender Bedeutung ist. Die Zusammenarbeit internationaler Projektteams stellt dabei eine herausfordernde Komponente dar. Der Fokus dieser Arbeit liegt vorwiegend auf den Herausforderungen und Erfolgsfaktoren der internationalen Teamzusammenarbeit bei Mercedes-Benz.

Definition Projektmanagement

Das Projektmanagement verfolgt das Ziel, Projekte erfolgreich zu leiten. Dabei spielen klar definierte Zielvorgaben, die klare Zuordnung von Ressourcen und die Umsetzung von Planungs- und Steuerungsmethoden eine entscheidende Rolle. Effektive Zusammenarbeit und kontinuierliche Verbesserung der Problemlösungskompetenz sind ebenfalls von zentraler Bedeutung. Zudem umfasst es sämtliche planende, überwachende, koordinierende und steuernde Maßnahmen, die notwendig sind, um Systeme, Prozesse oder Problemlösungen zu gestalten oder zu konzipieren [3]. Im Verlauf eines Projekts durchläuft man fünf Hauptphasen: den Start, die Planung, die Durchführung, die Kontrolle und schließlich den Abschluss (siehe Abbildung 1).

Abb. 1: Die 5 Hauptphasen im Projektmanagement [1]

Daher legen Unternehmen wie Mercedes-Benz Wert darauf, dass Mitarbeiter sich mit ihrer Projektarbeit identifizieren können, um die Eigenverantwortung, Effektivität und Motivation zu steigern. Im Projektmanagement gibt es zudem verschiedene Ansätze, die je nach der Projektgröße und der Komplexität kategorisiert werden. Die Entscheidung für einen Ansatz hängt dabei häufig von der Unsicherheit und

Vorhersehbarkeit des Projekts ab. In einem unsicheren Projektumfeld erweist sich agiles Projektmanagement oder eine Kombination aus agilen und klassischen Ansätzen als effektiver. Hingegen ist im sicheren Umfeld das klassische Projektmanagement die bessere Wahl (s. Abbildung 2). Im agilen Projektmanagement gibt es zuvor kein detailliertes Projektergebnis, es basiert vielmehr auf grob definierten Anforderungen und ermöglicht somit eine flexible Anpassung des Projektumfangs in kurzen Intervallen, sogenannten Sprints. Durch die schrittweise und flexible Entwicklung entstehen schon früh funktionsfähige Zwischenprodukte. Dies ermöglicht nicht nur eine schnellere Reaktion auf neue Anforderungen, sondern auch eine höhere Kundenzufriedenheit durch passgenaue Lösungen. Im klassischen Projektmanagement wird zu Beginn eines Projekts ein definiertes und detailliertes Projektergebnis festgelegt. Dieses basiert auf bereits festgelegten Plänen, Kostenabschätzungen und klaren Ressourcenzuweisungen. Problematisch wird es, wenn sich während der Projektlaufzeit die Zielvorgaben verändern oder weitere Anforderungen auftreten, welche zu Zeit- und Budgetproblemen führen können [2]. Besonders bei sehr großen Projekten wird das hybride Projektmanagement eingesetzt, eine Mischform der klassischen und agilen Ansätze, wie in Abbildung 2 dargestellt.

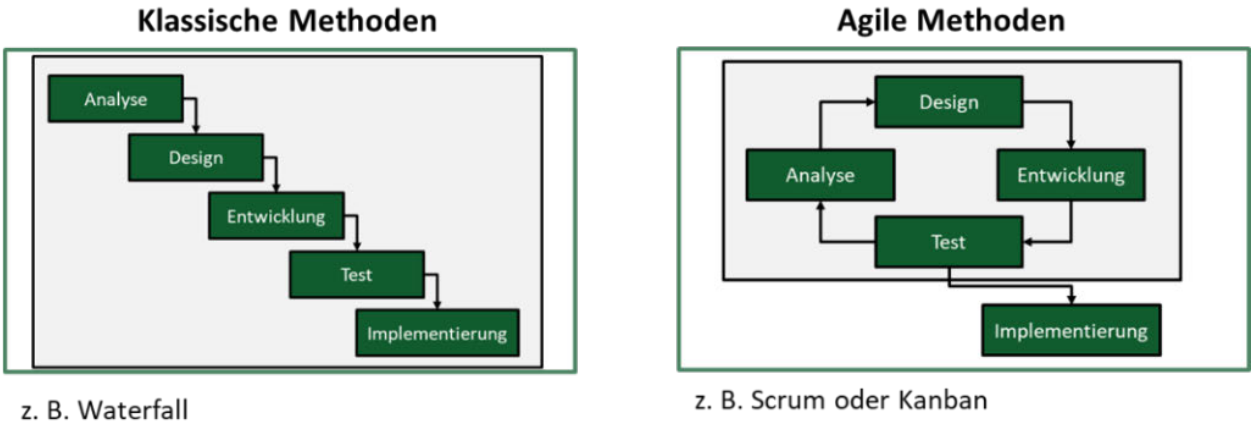


Abb. 2: Klassische und agile Methode [6]

Herausforderungen internationaler Projektteams

Die Zusammenarbeit internationaler Teams birgt spezifische Herausforderungen. Unterschiedliche Kulturen, Arbeitsmethoden und Kommunikationsstile können die Effektivität beeinflussen. Faktoren wie Sprachbarrieren, Zeitumstellungen, geografische Distanzen und Hierarchien spielen eine Rolle. Die globale Expansion vieler Unternehmen in verschiedenen Städten und Ländern führt zu Herausforderung durch unterschiedliche Zeitzonen. Wodurch sich das Organisieren von Meetings und Deadlines erschweren. Interne Chats und asynchrone Kommunikationskanäle bieten hierbei Lösungen [4]. Auch sprachliche Barrieren beeinflussen die Zusammenarbeit, so können die verschiedenen Muttersprachen innerhalb der Teams zu Missverständnissen bei der Interpretation von Begriffen und Ausdrücken führen. Zum anderen beherrschen nicht alle Teammitglieder fließend
die gewählte Unternehmenssprache, was die Kommunikation und das Verständnis erheblich erschwert. Dies führt dazu, dass die Anforderungen innerhalb eines Projekts missverstanden werden kann und führt somit zu weiteren Kosten- und Zeitproblemen. Kulturelle Barrieren wie Hierarchien, Wertvorstellungen und Verhaltensvorschriften können die effektive Zusammenarbeit internationaler Projektteams erheblich beeinflussen. Dies zeigt sich insbesondere durch die unterschiedlichen Präferenzen bezüglich des Führungsverhaltens und der kulturellen Normen [5]. Es lässt sich abschließend festhalten, dass auch bei Mercedes-Benz die Herausforderungen internationaler Projektteams eine entscheidende Rolle spielen. Die Vielfalt an kulturellen Einflüssen und sprachlichen Barrieren erfordern ein tiefes Verständnis und gezielte Strategien, um eine effektive Zusammenarbeit zu gewährleisten.

Ausblick

Die Zielsetzung dieser Arbeit besteht darin, wertvolle Erkenntnisse zu generieren, um die Effizienz und Effektivität internationaler Projektteams bei Mercedes-Benz zu optimieren. Durch die Identifizierung von Herausforderungen und Erfolgsfaktoren strebt die Arbeit an, Handlungsempfehlungen für zukünftige Projekte abzuleiten. Eine kontinuierliche Verbesserung der Zusammenarbeit wird dazu beitragen, dass Mercedes-Benz seine Projekte weltweit erfolgreich und effizient umsetzen kann.

Literatur und Abbildungen

[1] Sabrina Bernhardt. Die 5 Projektmanagement-Phasen. https://www.brainformatik.com/blog/projektmanagement-phasen/, 2023.

[2] KA. BVA. Was ist Projektmanagement und welche Methoden gibt es? https://www.bva.bund.de/DE/Services/Behoerden/Beratung/Beratungszentrum/GrossPM/Wissenspool/_documents/Standardartikel/stda_mgmt.html, 2023.

[3] Jürg Kuster. *Handbuch Projektmanagement: Agil - Klassisch - Hybrid*. Springer Gabler, 5 edition, 2022.

[4] Jan Marius Marquardt. Internationale Teams: Challenges und Chancen, StartingUp. https://www.starting-up.de/praxis/personal/transparentes-gehaltsmanagement-wie-unternehmen-talente-halten.html, 2023.

[5] Manuel Möll. Multikulturelle Teams führen – Herausforderungen und Vorteile der globalisierten Führung, Projektassistenz-Blog. https://www.projektassistenz-blog.de/multikulturelle-teams-fuehren-herausforderungen-und-vorteile-der-globalisierten-fuehrung/, 2020.

[6] Daniel Proba. Agile versus klassische Methoden – Unterschiede und Anwendungsbereiche der beiden Methodenfamilien. https://ccecosystems.news/agile-versus-klassische-methoden-unterschiede-und-anwendungsbereiche-der-beiden-methodenfamilien-teil-2/, 2020.

Analyse, Konzeption und Realisierung eines Software-Prototyps für eine einfache Beschreibungs- und Modellierungsmethode von komplexen Testfällen

Felix Rohner **Clemens Klöck**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma comemso-Gruppe, Ostfildern

Einleitung

In der rasanten Entwicklung der globalen Technologielandschaft manifestiert sich die E-Mobilität als ein monumentaler Paradigmenwechsel, der das Potenzial besitzt, unsere konventionellen Vorstellungen von Mobilität und Energieverbrauch grundlegend zu revolutionieren. Durch die Konzentration auf batterieelektrische Fahrzeuge wird versucht die globalen CO2-Emissionen zu verringern und somit den ökologischen Fußabdruck des Verkehrssektors zu minimieren. [3] Das Testen von Elektrofahrzeugen und ihren Komponenten ist von immenser Bedeutung, um die Sicherheit, Effizienz und Zuverlässigkeit dieser neuen Generation von Fahrzeugen sicherzustellen. Systematische Testverfahren ermöglichen es, potenzielle Mängel und Schwachstellen in den verschiedenen Bestandteilen von Elektrofahrzeugen, wie Batteriesystemen, Motoren und elektronischen Steuersystemen, zu identifizieren und zu beheben. Durch die Untersuchung bestehender Testmethoden, die Entwicklung neuer Teststrategien oder die Optimierung der Testprozesse können die Qualitätsstandards in der E-Mobilitätsbranche erhöht werden, um die Markteinführung von Elektrofahrzeugen zu beschleunigen.

Problemstellung

Kunden und interne Systemtester verwenden bei Ihrer Arbeit Hard- und Software der Firma comemso, um Ihre Systeme zu Testen. Die dafür eingesetzten komplexen Testfälle sind in einer Software programmiert und können nur durch Personen mit Programmierkenntnissen angepasst oder erstellt werden. Um diese Testfälle auch ohne Erfahrung in der Programmierung umzusetzen, soll eine einfache Testbeschreibungs- und Modellierungsmethode in den aktuellen Testprozess eingebunden werden, die auch von Personen ohne Programmierkenntnisse bedient werden kann.

Keyword-Driven Testing

Die Norm ISO 29119 Software Test ersetzt die bereits 1987 veröffentlichte Norm IEEE 829 und besteht aus fünf Teilen. Der Teil eins beschäftigt sich mit verschiedenen Konzepten und Definitionen rund um das Thema Software Tests. In Teil zwei wird der Testprozess selbst beschrieben. Die Test Dokumentation und verschiedene Test Techniken werden in Teil drei und vier definiert. Der für diese Publikation wichtigste Teil ist der Teil fünf. Dieser befasst sich mit der Testtechnik Keyword-Driven Testing. Die grundlegende Idee beim Keyword-Driven Testing besteht darin, eine Reihe von „Bausteinen“ bereitzustellen, die als Schlüsselwörter bezeichnet und verwendet werden können, um manuelle oder automatisierte Testfälle zu erstellen, ohne detaillierte Kenntnisse in Programmierung oder Testwerkzeugen zu erfordern. [1] Das ultimative Ziel ist es, einen grundlegenden eindeutigen Satz von Schlüsselwörtern bereitzustellen, der so umfassend ist, dass die meisten, wenn nicht sogar alle, erforderlichen Testfälle vollständig aus diesen Schlüsselwörtern zusammengesetzt werden können. [1] In Abbildung 1 wird die Modellierung einer solchen Schlüsselwortgetriebenen Programmierung beschrieben. Eine Testprozedur kann mehrere Testfälle enthalten. Testfälle können entweder manuelle oder Keyword Testfälle sein. Ein Keyword-Testfall implementiert einen manuellen Testfall. Ein Keyword-Testfall besteht aus einem oder mehreren Keywords. Ein Keyword Testfall wird typischerweise aus einer Serie von Schlüsselwörtern zusammengestellt. Schlüsselwörter sollten modular und generisch sein, sodass sie in unterschiedlichen Testfällen wiederverwendet werden können. [1] Keywords beschreiben Testaktionen. Ein Keyword wird durch einen Keyword-Execution-Code implementiert. Dieser Code ist entweder direkt vom verwendeten Framework generiert oder wurde im Vorhinein durch den Tester vorbereitet. Automatisierte Testskripte werden in der Praxis von erfahrenen Entwicklern geschrieben.

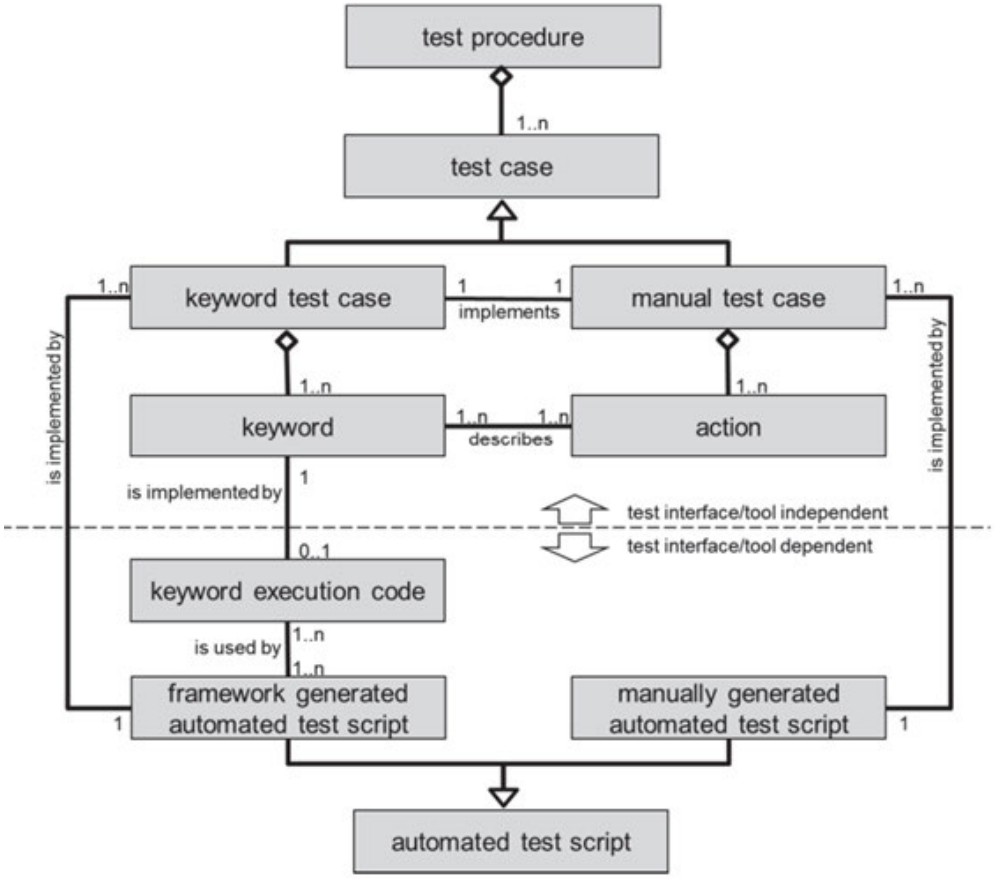


Abb. 1: Keyword-Driven-Testing Beziehungen [1]

ASP.NET

ASP.Net ist ein vielseitiges Webanwendungs-Framework, das von Microsoft entwickelt wurde und für den Aufbau dynamischer Webanwendungen und Dienste konzipiert ist. ASP.Net bietet Sicherheitsfunktionen, Optionen zur Zustandsverwaltung und unterstützt verschiedene Datenzugriffstechnologien. Es ist erweiterbar und ermöglicht die Erstellung von Webdiensten. Im Gegensatz zu Desktop- oder Konsolenapps ist die Funktionsweise einer Webanwendung eine andere. Es wird das Client-Server-Prinzip eingesetzt. Der Client ist in der Regel ein Webbrowser. [2] . Der Server kann durch eine ASP.Net Anwendung realisiert werden. Der Browser fordert per HTTP-Anfrage eine Ressource vom Server an, beispielsweise ein Dokument. Auf dem Server läuft die komplette Logik ab. Wie in Abbildung 2 zu sehen ist, fragt der Server eine Ressource von einer Datenbank ab und bereitet diese für den Client auf. Eine direkte Datenbankverbindung besitzt der Client nicht. [2].

Abb. 2: Client-Server-Modell [2]

Lösungsansatz

Es kommt ein Schlüsselwortgetriebener Ansatz, wie der Standard ISO 29119 in seinem Teil 5 (keyword-driven-testing) [1] beschreibt zum Einsatz. Dabei werden elementare und komplexe Testfälle jeweils durch Keyword Testfälle abgebildet. Diese Testfälle sind comemso spezifisch und sind abhängig von den eingesetzten Testgeräten, Simulatoren und den jeweiligen Standards und Normen. Um diese Schlüsselwortgetriebenen Testfälle zu starten, kommen standardisierte Schnittstellen zum Einsatz. Über diese Schnittstellen sollen einzelne Schlüsselwörter angestoßen werden, damit der jeweilige Testfall ausgeführt wird. Diese Verfahrensweise wird auch als Remote-Procedure-

Call bezeichnet. Die Ansteuerung der Schlüsselwörter erfolgt aus einem Testframework, in dem die Schlüsselwörter über eine grafische Programmierung kombiniert werden können. Das Ergebnis eines einzelnen Keywords wird über die Schnittstelle in einem standardisierten Format zurückgegeben und kann in der Testumgebung ausgewertet werden.

Ergebnis

Durch eine Marktanalyse bestehender Testumgebungen wurde der aktuelle Stand der Technik analysiert. Aus den Eigenschaften und Funktionen der Testumgebungen, den Forderungen der Geschäfts -und Entwicklungsleitung und den Ansprüchen der Kunden wurden Anforderungen an einen Softwareprototypen abgeleitet. Durch die Kombination bestehender Ansätze wie der grafischen Testfallmodellierung und Remote-Procedure-Calls wurde ein Konzept für einen Prototypen entwickelt. Dieses Konzept besteht darin, eine Standardsoftware für die Testfallmodellierung mit einer spezifischen Testbibliothek auszustatten. Diese Testbibliothek enthält einzelne Testfälle, die durch Keywords abgebildet werden, in denen wiederum Remote-Procedure-Calls zu einer von comemso entwickelten Software führen. Somit kann der Kunde aus seiner eigenen Tool-Landschaft heraus vordefinierte Testfälle aufrufen und einbinden. Diese Aufrufe werden durch eine standardisierte Schnittstelle realisiert. Die Schnittstelle wird zum Aufruf und Parametrierung von komplexen Testfällen in einer Mircoservice orientierten Architektur verwendet. Aus dem System Design wurde in einer weiteren Projektphase die Softwarearchitektur erstellt. Diese Architektur unterteilt die Software in Teilsysteme, die mithilfe von UML-Klassendiagrammen beschrieben sind. Mithilfe moderner Programmieransätze wie Dependency-Injection und objektorientierter Programmierung wurde der Softwareprototyp in einer C# .Net Software umgesetzt. Die Remote-Procedure-Calls sind mithilfe von Webtechnologien innerhalb eines ASP.NET REST-API Projektes realisiert worden. Ebenfalls wurde die Testbibliothek in einer zuvor gewählten grafischen Entwicklungsumgebung Implementiert. Innerhalb einer Testphase wurde der Softwareprototyp mithilfe von Modul-, Integrations- und Funktionstests verifiziert. Anschließend erfolgte die Abnahme durch einen Vergleich der Ergebnisse mit den gesetzten Anforderungen.

Ausblick

Der gewählte Ansatz mithilfe von Remote-Procedure-Calls Testfälle zu Starten und zu Parametrieren ermöglicht eine Vielzahl von Erweiterungen. Die Verwendete REST-API Schnittstelle für die Keyword Ansteuerung kann mithilfe der IP-Adresse und Portnummer so konfiguriert werden, dass Sie auch über eine Netzwerkkarte angesteuert werden kann. Dies könnte in Zukunft die Möglichkeit eröffnen, einen großen Prüfstand mit einem verteilten Computernetzwerk zu errichten. Die Schnittstelle könnte über ein geeignetes Routing sogar im World-Wide-Web bereitgestellt werden. Daraus ergäbe sich ein noch größeres Potential für die Fernsteuerung oder Fernwartung.

Literatur und Abbildungen

[1] Institute of Electrical Engineers and Electronics. ISO/IEC/IEEE International Standard - Software and systems engineering – Software testing – Part 5: Keyword-Driven Testing. http://ieeexplore.ieee.org/servlet/opac?punumber=7750537, 11 2016.

[2] Jürgen Kotz and Christian Wenz. *C# und .NET 6. Grundlagen, Profiwissen und Rezepte*. Hanser, 1 edition, 2022.

[3] Claudia Kämper, Hinrich Helms, and Kirsten Biemann. Wie klimafreundlich sind Elektroautos? Update Bilanz 2020. https://www.bmuv.de/fileadmin/Daten_BMU/Download_PDF/Verkehr/emob_klimabilanz_bf.pdf, 01 2020.

Steigerung der Effizienz und Energieeffizienz in automatisierten Fertigungssystemen durch den Einsatz von Deep Q-Learning

Simon Rosenberger **Gabriele Gühring**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma MAG IAS GmbH, Eislingen

Einleitung

Das Bewusstsein für Energiesparen hat in den letzten Jahren nicht nur im Alltag an Bedeutung gewonnen, sondern auch in der Industrie und im Maschinenbau. Produktionsanlagen müssen nicht nur schnell produzieren, sondern auch so energiesparend wie möglich arbeiten. Um die Genauigkeit und Produktionsbereitschaft einer Werkzeugmaschine zu halten, muss diese konstant auf Betriebstemperatur gehalten werden was einem ineffizienzen Energieverbrauch verursacht. Dies führt bei parallel beladbaren Maschinen dazu, dass bei Teilauslastung am sinnvollsten Anlagenteilnehmer vollausgelastet produzieren und andere vollständig abgeschaltet werden, um Energie zu sparen. Infolge dessen kommt es zu einer ungleichmäßigen Auslastung der Maschinen während der Produktion, die zu einem frühzeitigen Ausfall vor Ende der Lebenszeit einer Maschine führen kann. Daher ist zusätzlich zur Optimierung der Energieeffizienz eine effiziente Nutzung der Maschinen anzustreben, um eine Produktion, die sowohl energie- als auch ressourcenschonend ist, zu gewährleisten.

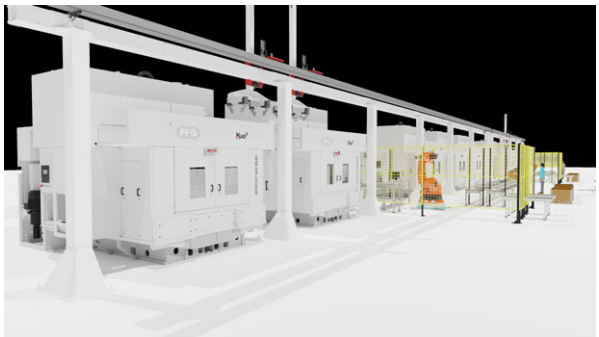


Abb. 1: Teilansicht Produktionslinie mit Portallader [2]

Verwandte Arbeiten

Vereinfacht gesagt handelt es sich hier um ein 'simples' Optimierungsproblem, das auf Basis einer Materialflussplanung bzw. Warteschlangensimulation aufgebaut werden kann. Zusätzlich kommt in diesem Fall natürlich die Optimierung des Energieverbrauchs dieser Produktionsanlagen hinzu.
Wie in [1] beschrieben, ist Reinforcement Learning, in diesem Fall Q-Learning, eine mögliche Lösung für diese Art von Optimierungsproblemen. Bei einer Produktionslinie mit sechs parallel beladbaren Werkzeugmaschinen und einem zugehörigen Portallader würde jedoch ein State-Action-Space mit mehreren Millionen State-Action Paaren entstehen. Die Komplexität des Problems würd damit den Rahmen von Q-Learning sprengen. Erweiterungen in der Produktionslinie würden in diesem Fall ebenso zu einem erheblichen Mehraufwand führen.
Für komplexere Probleme ist die Verwendung von Deep Q-Learning eine Möglichkeit. Statt eines State-Action Spaces wird ein neuronales Netz verwendet, welches Q-Werte approximiert und damit mit einer viel größeren Anzahl an State-Action Paaren umgehen kann. Ein auf dieser Basis funktionierendes System wird in [4] vorgestellt bei dem Deep Q-Learning verwendet wird um Produktionsabläufe zu optimieren. Ist das Problem auch für reines Deep Q-Learning zu komplex, kann wie in [3] beschrieben, die Komplexität des Problems mithilfe von Multi-Agenten Systemen reduziert werden. Das System ermöglicht einzelnen Agenten einen Teil des Problems zu lernen und und zu steuern. Die einzelnen Agenten werden unter einem Master-Agenten zusammengefasst der diese koordiniert.

Zielsetzung

Die bisherige Lösung zum Thema Energiesparen basiert auf einer Funktion, die mithilfe der SPS einer Maschine ausgeführt wird. Jede Maschine besitzt ihren eigenen

Energiemanager, der nicht übergreifend agieren kann. Eine Optimierung bezüglich der gleichmäßigen Auslastung ist somit aktuell nicht möglich.
Ziel ist es, eine übergeordnete Steuerung zu entwickeln, die in der Lage ist, mehrere Maschinen und deren zugehörigen Portallader gleichzeitig zu steuern. Sie soll mithilfe von Deep Reinforcement Learning Entscheidungen treffen. Unter Verwendung des DQN-Agents von Tensorflow soll anhand der von der Umgebung erhaltenen Daten der Zustand der Maschinen verändert und deren Beladeverhalten gesteuert werden.
Zur Validierung der Lösung wird ein Simulink-Modell erstellt, das das Standardverhalten der Maschinen simuliert. Zusätzlich gibt es ein zweites Simulink-Modell, welches als Umgebung für den DQN-Agenten dient. Dieses Modell ermöglich es dem Agenten, in einer simulierten Umgebung zu lernen und zu trainieren, bevor er in der realen Produktionsumgebung eingesetzt wird. Das erste Modell dient als Referenz, um den Energieverbrauch einer Anlage, die mithilfe des DQN-Agenten gesteuert wird, mit dem Energieverbrauch einer Standardproduktion zu vergleichen.

Vorgehensweise

Als ersten Schritt erfolgte die Erstellung der Simulink-Simulation, die das Standardverhalten einer Werkzeugmaschine nachbildet. Der Materialfluss wurde mithilfe der Bibliothek SimEvents umgesetzt und die Maschinenlogik durch Integration von Matlab-Funktionen und klassischer Simulink-Logik modelliert. Diese Nachbildung einer Produktionslinie mit Werkzeugmaschinen ist stark vereinfacht und berücksichtigt nur Elemente, die für die Beladung, Produktion und die Maschinenzustände, aus denen der Energieverbrauch resultiert, relevant sind.
Auf Grundlage dieser Simulation ist die Umgebung für einen DQN-Agenten erstellt worden. Der Aufbau des Materialflusses ist hierbei unverändert, wobei die Beladung der Maschinen von einem Agenten gesteuert wird. Der Teil der Maschinenlogik, der die Entscheidung über einen Wechsel eines Maschinenstatus beinhaltet, wurde ausgelagert und wird ebenso von einem Agenten gesteuert. Die Kommunikation zwischen Agent und Umgebung läuft über eine definierte Schnittstelle, die später auch bei Anbindung an eine reale Produktionslinie verwendet werden soll.
Der DQN-Agent kann somit die Zustände der Maschine (Standby, Operational, Producing) sowie das Beladeverhalten der Maschinen steuern. Die Steuerung der Maschinenzustände zielt auf die Reduzierung des Energieverbrauchs ab, die Steuerung des Portalladers soll eine gleichmäßige Auslastung der Maschinen garantieren. Die aktuelle Version des DQN-Agenten basiert auf einem Feed-Forward-NN mit einem versteckten Layer.

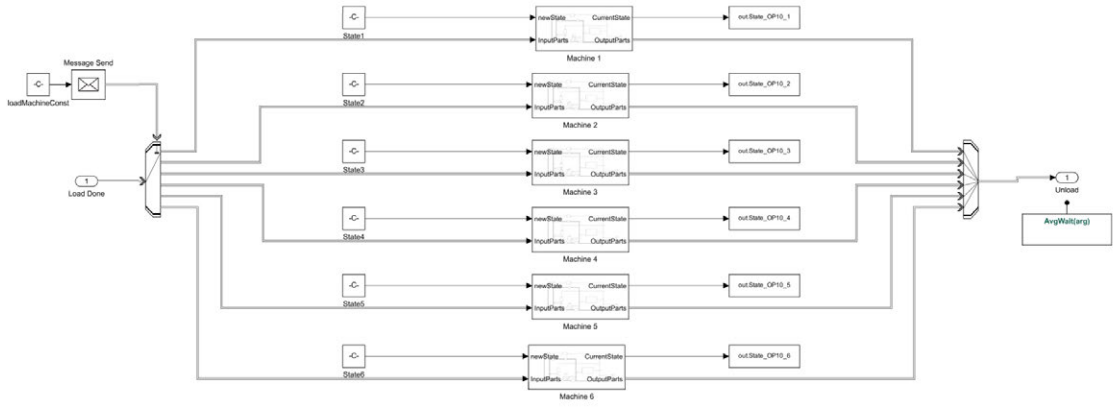


Abb. 2: Ausschnitt OP10 aus Environment für DQN-Agent [2]

Erste Erkenntnisse und Ausblick

Derzeit wird das Training des DQN-Agenten durchgeführt. In dieser Phase liegt der Fokus auf Reward-Shaping und der Hyperparameter-Optimierung. Ziel ist es, dem Agenten ein möglichst effektives Lernen zu ermöglichen.
Im ersten Schritt beinhaltet dies, den Agenten dazu zu bringen, die Umgebung kennenzulernen und diese korrekt zu bedienen. Wird während des Trainings eine invalide Aktion ausgeführt, (dies umfasst Aktionen, die physikalisch unmöglich sind oder an einer realen Produktionslinie zu Schäden an der Maschine führen würden) bricht der Agent die aktuelle Episode ab und beginnt von vorn.

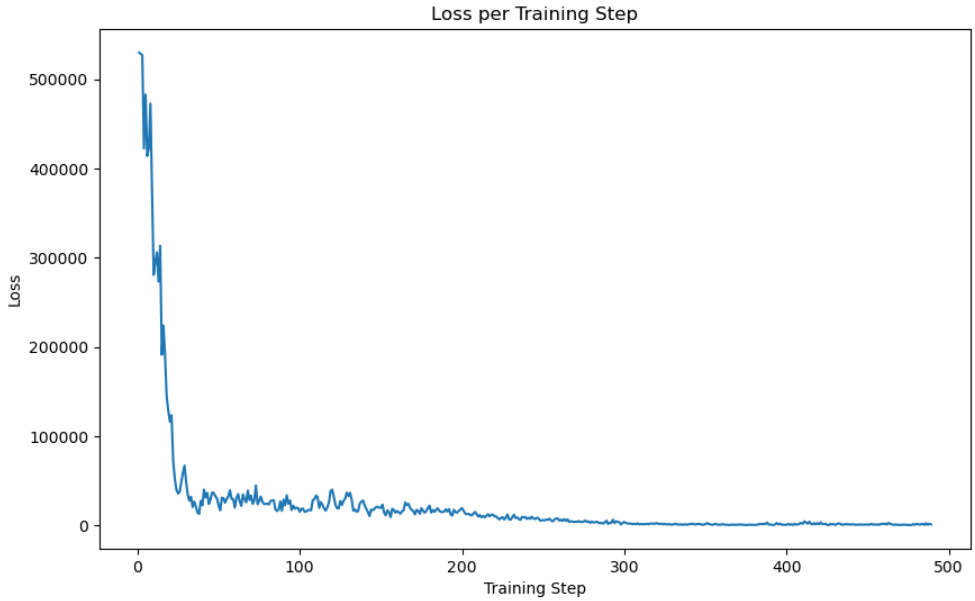


Abb. 3: Loss-Entwicklung erster Trainingsdurchlauf [2]

Im nächsten Schritt soll der Agent mithilfe von optimierten Rewards und Hyperparametern ein Parameter-Set finden, an dem die Energieersparnis maximal und die Abweichung der produzierten Teile pro Maschine vom Durchschnitt minimal wird.
Sollte dies nicht die gewünschten Ergebnisse liefern, ist eine Änderung des neuronalen Netzes zu GRU oder RNN in Betracht zu ziehen.
Sollte das Modell dennoch zu komplex für einen Agenten sein, gibt es die Möglichkeit der Integration eines Multi-Agenten-Systems und die Aufteilung der Modellkomponenten auf mehrere Agenten, um die individuelle Komplexität zu reduzieren.
Es wird möglich sein, Anfang 2024 verbindliche Aussagen zu diesem Thema zu machen.

Literatur und Abbildungen

[1] Peter Burggräf, Fabian Steinberg, Benjamin Heinbach, and Milan Bamberg. Reinforcement Learning for Process Time Optimization in an Assembly Process Utilizing an Industry 4.0 Demonstration Cell. In *55th CIRP Conference on Manufacturing Systems*. ScienceDirect, 2023.

[2] Eigene Darstellung.

[3] Bernd Waschneck. Dissertation: Autonome Entscheidungsfindung in der Produktionssteuerung komplexer Werkstattfertigungen, 2020.

[4] Tong Zhou et al. Reinforcement Learning With Composite Rewards for Production Scheduling in a Smart Factory. *IEEEAccess*, 2021.

UX-gesteuerte Prozessoptimierung: Design eines effizienten Co-Piloten für einen digitalen Zwilling zur Standardisierung von Arbeitsabläufen

Jasmin Saleh **Jürgen Koch**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma OEM, Stuttgart

Einleitung

Im Jahr 2023 konzentrieren sich Unternehmen verstärkt auf Logistik, bedingt durch globale Krisen wie COVID-19 und den Ukrainekrieg. Die Automobilbranche setzt verstärkt auf Digitalisierung, um diesen Herausforderungen in Lieferketten, Kundenorientierung und Anpassungsfähigkeit zu begegnen. In dieser Arbeit liegt der Fokus auf der Intralogistik, die maßgeblich für Effizienz, Kostenreduktion und Wettbewerbsfähigkeit ist. Um diesen Herausforderungen zu begegnen, sind technologische Innovationen und die Integration von Echtzeitdaten in die internen Logistikprozesse notwendig. Ein digitaler Zwilling, der die physischen Prozesse virtuell abbildet, ist Teil dieser digitalen Transformation. Er trägt zur Transparenz bei und verbessert die Steuerung sowie Planung. Der Einsatz eines Co-Piloten im digitalen Zwilling kombiniert die Stärken der Technologie mit menschlicher Expertise. Die Entwicklung dieses Co-Piloten reagiert auf die zunehmende Komplexität in der Logistik sowie auf ständige Veränderungen in Prozessen und Rahmenbedingungen, denn letztendlich sind es Menschen, die die Prozesse lenken, ausführen und bei Bedarf eingreifen.

Supply Chain und Logistik

Die Supply Chain befasst sich mit den Produkten oder Dienstleistungen von der Rohstoffbeschaffung bis zur Endkundenauslieferung. Hierbei werden notwendige Organisationen, Ressourcen, Aktivitäten und Technologien betrachtet. Die Logistik ist ein Teil der Supply Chain. Wie in Abbildung 1 [2] zu sehen ist, sollen die strategisch-taktische Planungsebene mit der operativen Umsetzung verbunden werden. Hierbei konzentriert sich die Logistik auf die physische Bewegung und Verwaltung von Gütern. Sie umfasst den Transport, die Lagerung, die Bestandsverwaltung und die Distribution von Produkten innerhalb der Supply Chain. [2]

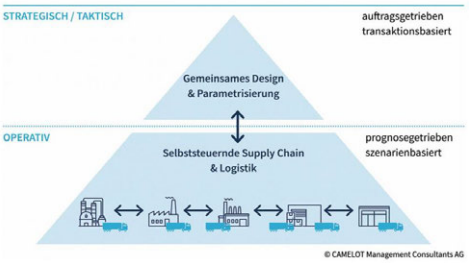


Abb. 1: Supply Chain Management [2]

Um die volle Leistungsfähigkeit der Supply Chain in Bezug auf Sichtbarkeit und Organisation zu erreichen, gewinnt der Einsatz digitaler Zwillinge zunehmend an Bedeutung. Diese ermöglichen die zentrale Bündelung von Informationen und automatisierte Prozesse in der Lieferkette.

Definition eines Co-Piloten

Die Künstliche Intelligenz ist in aller Munde und gilt als eine der bedeutendsten technologischen Fortschritte unserer Zeit. Die Fusion von Chat-Oberflächen mit leistungsstarken Sprachmodellen ermöglicht es nun, in natürlicher Sprache Fragen zu stellen. Bei Microsoft wird dies als „Copilot“ bezeichnet, der als unterstützender Assistent fungiert und bei verschiedensten Aufgaben zur Seite stehen soll. [4] Diese Arbeit befasst sich jedoch nicht mit dem Co-Piloten als Chatbot, sondern mit seiner Funktion, automatisierte Handlungsempfehlungen zu liefern. Sie zielt darauf ab, die zunehmende Komplexität in globalen Lieferketten und intralogistischen Abläufen zu bewältigen, hervorgerufen durch technologische Entwicklungen und die Notwendigkeit effizienter Prozesse. Der Co-Pilot wird hier als Lösungsansatz betrachtet, um die Kontrolle

und Anpassungsfähigkeit in der Lieferkettenlogistik zu stärken.

Ziel der Abschlussarbeit

Ziel dieser Arbeit ist die Analyse und Optimierung der Nutzung des bereits existierenden digitalen Zwillings. Hierzu wird der Einsatz eines Co-Piloten in Betracht gezogen. Der Co-Pilot soll auf Basis des bereits dargestellten Echtzeitsytems Hinweise und automatisierte Handlungsempfehlungen liefern. Es ist eine weitere Stütze für den Anwender, um die Prozesse und Arbeitsabläufe besser zu verstehen und Handlungen frühzeitig durchzuführen. Durch die UX-gesteuerten Prozessoptimierung soll hauptsächlich, die Benutzererfahrung verbessert werden. Hierbei soll die Effizienz, Produktivität und Akzeptanz des Co-Piloten in der Nutzung des digitalen Zwillings zur Standardisierung von Arbeitsabläufen gesteigert werden. Das Design des Co-Piloten soll somit dem Benutzer relevante Informationen und Handlungsempfehlungen auf eine leicht verständliche und zugängliche Weise ermöglichen. Hierbei soll die Interaktion mit dem digitalen Zwilling erleichtert werden und den Benutzer bei der Ausführung von Aufgaben unterstützen.

Methodik

Damit die UX-gesteuerte Prozessoptimierung mithilfe des Co-Piloten für den digitalen Zwilling ermöglicht wird, bedarf es einer effektiven Methodik. Hierbei sollen die Bedürfnisse, das Verhalten und die Anforderungen der Benutzer berücksichtigt und analysiert werden. Aufgrund dessen, wurde das Design Thinking Model gewählt, welches aus sechs Schritten besteht. Die Benutzerfreundlichkeit und die Erfahrungen der Anwender stehen bei diesem Model im Vordergrund. Die daraus gewonnenen Einsichten sind der Startpunkt für die eigentliche Ideengenerierung. Durch Erstellen und Testen von Prototypen werden Ideen umgesetzt und evaluiert. In Abbildung 2 [1] sind die Schritte des Design Thinking Models abgebildet. [3]

Abb. 2: Design Thinking Prozess [1]

Ausblick

Die Entwicklung eines effizienten Co-Piloten für den digitalen Zwillingen bringt verschiedene Aspekte mit sich. Neben einer kontinuierlichen Verbesserung durch Nutzerfeedback können auch weitere Funktionalitäten in Betracht gezogen werden. Ein Ausblick zielt vor allem darauf ab, einen Beitrag zu leisten, der Arbeitsabläufe effizienter und weniger anfällig für menschliche Fehler gestaltet, indem der Co-Pilot die Benutzerführung verbessert und die Entscheidungsfindung für das Handeln unterstützt. Um den Co-Piloten weiter zu verbessern, könnten künftige Fortschritte die Integration von KI-Algorithmen oder maschinellem Lernen einschließen. Dies würde dazu dienen, potenzielle Probleme frühzeitig zu identifizieren und Lösungsansätze bereitzustellen.

Literatur und Abbildungen

[1] Infografik Designer in Action. Was ist Design Thinking? Prozess und Methode erklärt. https://www.designer-in-action.de/design-wissen/design-thinking/, 09 2023.

[2] Dominik Hartung and Andreas Gmür. Zusammenspiel von Supply Chain und Logistik. https://www.chemanager-online.com/news/zusammenspiel-von-supply-chain-und-logistik, 07 2021.

[3] Michael Lewrick, Patrick Link, and Larry Leifer. *Das Design Thinking Toolbook: Die besten Werkzeuge & Methoden*. Vahlen, 1 edition, 2019.

[4] Yusuf Mehdi. Microsoft Copilot, Ihr täglicher KI-Begleiter. https://news.microsoft.com/de-de/microsoft-copilot-ihr-taeglicher-ki-begleiter/, 09 2023.

Towards Platform-Independent Web-Based Augmented Reality

Adrian Salmeron **Reiner Marchthaler**

Department of Computer Science and Engineering, Esslingen University

Work carried out at QUANTO Solutions GmbH, Stuttgart

Introduction

Augmented Reality (AR), initially termed by Thomas P. Caudell in 1992 [1], merges digital information with physical environments, revolutionizing sectors like gaming, education, healthcare, marketing and retail. AR's journey from a niche innovation to a mainstream tool has dramatically changed user engagement and spurred economic growth. The evolution of AR, originating in the 1960s [10] together with virtual reality, has moved beyond science fiction with significant advancements in hardware and computer vision. The recent focus on web-based AR aims to create universally accessible experiences, independent of specific devices or platforms. Yet, this field faces ongoing challenges compared to its native counterpart in areas of performance, accuracy, and the need for universal standards for spatial tracking and device interfaces, ensuring seamless integration across different browsers and operating systems.

Motivation

The rapidly growing adoption and economic impact of AR make it no longer just a competitive advantage, but a critical service offering in addressing complex needs of clients across various industries. QUANTO Solutions GmbH, an agile IT consultancy providing services to some of the largest German enterprises, has recognized the relevance for future markets. The collaboration with a prominent, exchange-traded multimedia company, specializing in digital or physical advertising and marketing, underscores the demand and practicality of AR applications in contemporary business scenarios and serves as a prime example of industry's growing inclination towards AR technologies. The goal is to empower field staff with web-based AR, facilitating on-site visualizations of physical media carriers during crucial client interactions. This approach promises to enhance the sales narrative, offering tangible previews to clients, potentially increasing conversion rates and accelerating contract finalization. The exploration of web-based AR was specifically chosen to circumvent logistical challenges of native solutions. This concept is beneficial in diverse and evolving technical ecosystems and fits within stringent enterprise security frameworks, ensuring broader accessibility and simpler maintenance. While web application development is generally considered more cost-effective than native solutions in common use-cases, whether the same goes for AR remains a subject of investigation with considerations towards efficacy and potentially severe technical trade-offs. This thesis aims to offer insights into the developments and possibilities of purely web-based markerless AR applications in professional settings.

The State of Cross-Platform WebXR

The WebXR API [3], a novel browser standard first proposed in 2018, represents a major leap in the integration of augmented and virtual reality experiences directly into web browsers. While WebGL [4] serves as a 2D and 3D rendering API for browsers, WebXR utilizes the device's existing hardware such as the camera to capture its physical environment alongside sensors like gyroscopes and accelerometers to track movement and orientation. Although browser integration of WebXR occurs on a native level, the API is still exposed only via JavaScript and also relies on preinstalled software such as Google Play Services for AR on Android.

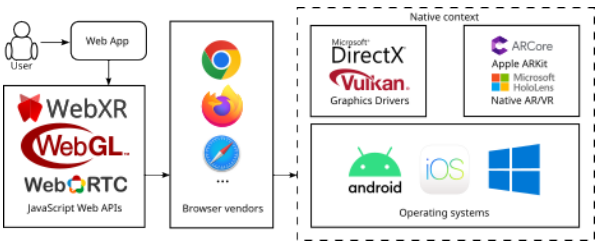


Fig. 1: Overview of the WebXR API in context of abstraction levels [8]

With ongoing developments of this cutting-edge technology in recent years, implementation of the standard among browser vendors has varied. While Google and Mozilla have incorporated it into Chrome and Firefox, it is not supported on Apple's proprietary browser Safari for iOS devices. Since iOS is a popular platform in the modern market and given Apple's restrictions on browser engine diversity on its mobile devices, this poses another considerable impediment to developing a universally applicable web-based AR solution. Moreover, this further constraints potential prototypical solutions for the collaborating enterprise, where mostly iPads are currently in use. However, Apple recently introduced WebXR in Safari for the Vision Pro 2 and participates in the W3C Immersive Web Working Group [9]. This consortium, along with the public Immersive Web Community Group, drives the development of the WebXR specification. It is therefore still important to understand this technology as it matures and expands, though it cannot serve as a tool for platform-independent solutions for now.

Web-Based Computer Vision and Markerless AR

Computer vision (CV) is an integral aspect of AR applications. The majority of current web-based AR frameworks or libraries mostly support marker-based AR, meaning they operate on image or object recognition to project digital assets on. However, this approach does not suit the given use-case presented by field staff, where the target location may not be physically accessible for installation of markers such as with the projection on the side of buildings etc. Markerless AR is generally considered a greater technical challenge due to a required deeper and more computationally intense understanding of the environment and object placement within. Basic concepts include surface detection for both floors and walls, while floor detection is generally easier than walls due to the consistent orientation of floors with gravity and their typically larger, unobstructed surfaces. The device's sensors can readily identify horizontal planes, aligning with gravity, making floor detection more straightforward. In contrast, wall detection is more complex. Walls vary in orientation, are often interrupted by objects like windows and furniture, and can have inconsistent lighting conditions. So while both of these are possible with current WebXR implementations, wall detection may be even less reliable compared to native apps. A significant step-up in terms of accuracy and performance requirements would be the use of SLAM (Simultaneous Localization and Mapping), which maps an environment by analyzing sensor inputs and visual data to construct and update a digital representation of the 3D space while concurrently determining the device's position within it. Although WebAssembly (WASM) is often proclaimed to be a way to achieve near-native performance in some aspects of web applications, it is still limited [11]. At the current juncture, only partial or highly experimental implementations of freely available SLAM engines exist for web browsers based on such tools. Presently, the only feature complete cross-platform SLAM-based AR for the web is a subscription model based commercial platform called 8thWall. This product employs highly optimized custom CV algorithms and was acquired by the creator of Pokémon GO, Niantic. Another approach for web-based CV in the context of AR involves machine learning. Convolutional neural networks can in theory be trained and executed in the browser through ECMAScript bindings such as Tensorflow.js, deeplearn.js or WebDNN and more. Yet, both the size of these models and forward-pass performance often remained unacceptable in a web context [5]. Finally, hybrid solutions comprising a client-server model for server side CV processing for the browser have been proposed with interesting results [7]. However, this approach may not always be viable due to lack of guaranteed low network latency out in the field.

Proposed Solutions and Tools

To develop potential prototypical solutions within the given boundaries, the current proposition includes two phases of investigation. In the first phase, we develop a base solution comprising WebXR with A-Frame.js and the widely-used Three.js WebGL library, based on the browser and operating system of the client. For mobile Safari, the QuickLook feature is accessed from the browser to preview 3D models for both floor and wall placements, which is natively available by default on iOS devices. QuickLook automatically utilizes specialized hardware such as a LiDAR scanner for enhanced depth perception and even occlusion, which is otherwise inaccessible in web applications due to a lack of browser APIs for this device.

Fig. 2: QuickLook feature accessed through Safari on an iPad Pro 5th Generation with a simple test model. (a) Manual floor placement in a roundabout. (b) Manual wall placement with occlusion enabled by integrated LiDAR. [8]

This approach takes advantage of the native performance provided by QuickLook. Conversely, a notable limitation of this design is the inconsistency in user experience across platforms. Specifically, when the application is accessed through an iOS browser, it demonstrates superior performance and employs a unique user interface for placement interaction. The second phase involves testing the use of WebXR based engines such as Enva-XR [2] to improve the user experience on non-iOS devices or experimental SLAM engines for a limited true cross-platform solution. Subsequently, performance benchmarks and user acceptance tests will be performed for both the first phase and second phase solutions for evaluation.

Outlook

While it is clear that web-based markerless AR is still in its infancy, especially in a platform-agnostic context, the evolution of this technology looks promising as both adoption and expansion of standards such as WebXR rapidly increase. Furthermore, the increased adoption of additional browser standards such as WebGPU [6], in which all big mobile device corporations including Apple actively participate, may promote the viability of client-side deep learning based CV in the browser once again for real-time applications. While the targeted mobile devices such as smartphones and tablets often operate on a System-on-a-Chip, direct access to the GPU via a browser API will likely benefit the training and execution of machine learning models in the web. The results of this thesis will provide important knowledge about the suitability of markerless web AR for businesses today, as well as anticipate significant advancements and trends for the near future.

References and figures

[1] Thomas P. Caudell and David Mizell. Augmented reality: an application of heads-up display technology to manual manufacturing processes. *Institute of Electrical and Electronics Engineer*, 2:659–669, 1992.

[2] José Ferrao, Paulo Dias, Beatriz Sousa Santos, and Miguel Oliveira. Environment-Aware Rendering and Interaction in Web-Based Augmented Reality. *MDPI*, 2023.

[3] Immersive Web Working Group Immersive Web. WebXR device api. https://www.w3.org/TR/webxr/, 2023.

[4] Dean Jackson and Jeff Gilbert. WebGL 2.0 Specification. https://www.khronos.org/registry/webgl/specs/latest/2.0/, 2023.

[5] Yun Ma, Dongwei Xian, Shuyu Zheng, Deiu Tyan, and Xuanzhe Liu. Moving Deep Learning into Web Browser: How Far Can We Go? In *WWW '19: The World Wide Web Conference*, pages 1234–1244. Association for Computing Machinery, 2019.

[6] Kai Ninomiya, Brandon Jones, and Jim Blandy. WebGPU. https://www.w3.org/TR/webgpu/, 2023.

[7] Pei Ren, Xiuquan Qiao, Junliang Chen, and Schahram Dustdar. Mobile Edge Computing – a Booster for the Practical Provisioning Approach of Web-Based Augmented Reality. In *2018 IEEE/ACM Symposium on Edge Computing (SEC)*, pages 349–350. IEEE, 2018.

[8] Own representation.

[9] Atsushi Shimono. Immersive Web Working Group. https://www.w3.org/groups/wg/immersive-web/, 2023.

[10] Ivan E. Sutherland. A head-mounted three dimensional display. In *AFIPS '68: Proceedings of the December 9-11*, pages 757–764. Association for Computing Machinery, 1968.

[11] Yutian Yan, Tengfei Tu, Lijian Zhao, Yuchen Zhou, and Weihang Wang. Understanding the performance of webassembly applications. In *IMC '21: Proceedings of the 21st ACM Internet Measurement Conference*, pages 533–549. Association for Computing Machinery, 2021.

Online-Tracking, Cookies und Datenschutz: Eine Analyse von Rechtsgrundlagen, Nutzerdaten und ethischen Aspekte im digitalen Raum

Simon Sami **Dirk Hesse**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt an der Fakultät Informatik und Informationstechnik, Esslingen

Einleitung

In unserer zunehmend digitalisierten Welt, in der Online-Aktivitäten allgegenwärtig sind, rücken Themen rund um den Datenschutz und die Kontrolle Personenbezogener Daten in den Fokus. Diese Bachelorarbeit beschäftigt sich mit Online-Tracking, Cookies und der Datenschutzanalyse im digitalen Raum, mit besonderem Fokus auf die rechtlichen Grundlagen, Nutzerdaten und ethischen Aspekte. Das Internet, als wichtiger Bestandteil der modernen Kommunikation, wird erheblich von Mechanismen wie Cookies und Tracking-Methoden beeinflusst, welche die Interaktion der Benutzer mit Websites beeinflussen.

Zielsetzung

Der Schwerpunkt dieser Studie liegt auf einer umfassenden Untersuchung der rechtlichen Grundlagen, insbesondere der Datenschutz-Grundverordnung (DSGVO) und der ePrivacy-Verordnung, und deren Auswirkungen auf die alltägliche Online-Nutzung. Das Ziel dieser Arbeit ist es, durch eine Umfrage unter allgemeinen Internetnutzern und Studenten Erkenntnisse zu gewinnen, wie die Internetnutzer auf aktuelle Ereignisse reagieren, insbesondere im Zusammenhang mit dem neuen Abo Modell von Meta zum Preis von 9,99€. Da wo die Nutzer der Plattformen Facebook und Instagram dazu aufgefordert wurden, entweder einen Abo für 9,99€ abzuschließen und sie durch diese Option ihre persönlichen Daten schützen können und nicht an Dritte für Werbezwecke weitergeben oder die Plattform wie gewohnt weiter zu nutzen unter der Bedingung, dass Ihre Daten für Werbezwecke benutzt werden. [3] Es wird auch der generelle Umgang der Internetnutzer betrachtet, wie diese mit Cookies umgehen und ob sie wissen, was für Daten sie durch ihre Einwilligung weitergeben. **In dieser Bachelorarbeit werden folgende Forschungsfragen beantwortet: Welche Arten von Cookies gibt es und wie unterscheiden diese sich? Und wie nehmen die Nutzer die ethischen Aspekte von Online-Tracking und Cookies wahr und inwieweit beeinflussen diese ihre Entscheidungen beim Besuch von einer Websites?** Es ist außerdem wichtig, nicht nur die theoretischen Hintergründe aufzuklären, sondern auch die Reaktionen und Wahrnehmungen normaler Internetnutzer durch eine empirische Untersuchungen zu dokumentieren.

Meta und Datenschutzentscheidungen

Das EuGH-Urteil stellt einen bedeutenden Erfolg für das deutsche Bundeskartellamt dar. Im Jahr 2019 hatte sich die Behörde mit Facebook auseinandergesetzt und dem Unternehmen das Sammeln von Daten ohne eine tatsächlich freiwillige Einwilligung der Nutzer untersagt. [2] In direkter Reaktion auf dieses Urteil hat Facebook ab November 2023 entsprechende Anpassungen vorgenommen. Die Plattform fordert seither aktiv die Nutzer von Instagram und Facebook dazu auf, entweder ein kostenpflichtiges Abonnement zum Preis von 9,99€ abzuschließen, um ihre persönlichen Daten zu schützen, oder wie gewohnt fortzufahren, wobei ihre Daten weiterhin für Werbe- und Marketingzwecke genutzt werden könnten. Dies zeigt, wie stark die Durchsetzungskraft des EuGH-Urteils ist und zeigt auch die direkten Auswirkungen auf die Datenschutzpraktiken großer Unternehmen wie Facebook. [3]
In der Abbildung 1 kann man sehen, wie Facebook die Nutzer direkte Anweisungen gibt. Die Plattform gibt den Nutzern die Wahl: Entweder sie zahlen 9,99€ für mehr Datenschutz, oder sie nutzen die Plattform weiterhin kostenlos, wissend, dass ihre Daten für Werbung und Marketing genutzt werden. Facebook erklärt dabei klar, wofür das Geld ausgegeben wird und welche Vorteile es mit sich bringt.

Abb. 1: Facebook werbefrei nutzen: Meta jetzt mit Abomodell [3]

Aufgrund dieser Tatsache wurde eine Umfrage unter Internetnutzern und Studenten durchgeführt. Das Ziel war es festzustellen, wie die Nutzer der Plattform die aktuellen Ereignisse finden und ob sie hinter dem Abo Modell stehen und was ihre Meinung dazu ist. An dieser Umfrage haben insgesamt 11 Menschen daran teilgenommen.

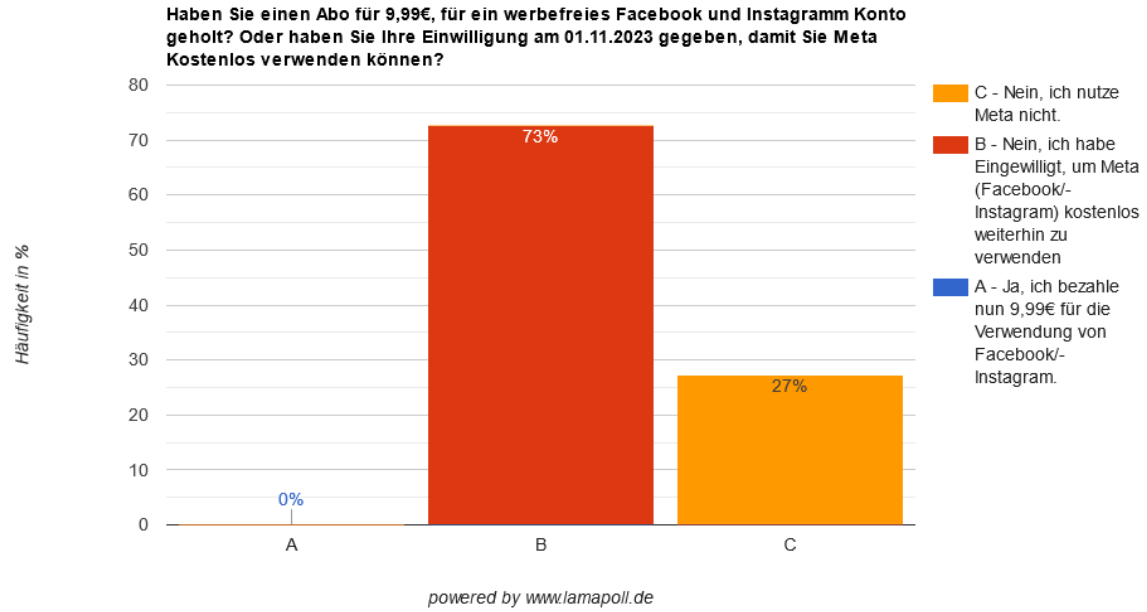


Abb. 2: Abomodell von Meta [1]

Wie man in der Abbildung 2 erkennen kann, hat die Mehrheit der Teilnehmer (73 %) angegeben, dass sie mit der Verwendung ihrer Daten für Werbezwecke einverstanden sind und Meta (Facebook/Instagram) weiterhin kostenlos nutzen. Interessanterweise gab kein einziger Befragter an, das 9,99€ Abonnement abgeschlossen zu haben. Dies könnte darauf hindeuten, dass die Teilnehmer eher bereit waren, persönliche Daten für den kostenlosen Zugang zu Facebook und Instagram preiszugeben, als für einen werbefreien Dienst zu bezahlen. Es ist außerdem wichtig zu beachten, dass 27% der Befragten angegeben haben, Meta nicht zu nutzen. Dieses Ergebniss wirft Fragen auf und weckt das Interesse für eine Untersuchung. Es könnte auch daran liegen, dass diese Nutzer beispielsweise aufgrund von Datenschutzbedenken Meta bewusst meiden.

Ausblick

Künftige Entwicklungen im Bereich Tracking, Cookies und Datenschutz erfordern ständige Aufmerksamkeit, denn dieser Bereich entwickelt sich ständig. Mit der Umfrageergebnisse werden Einblicke in die Reaktionen normaler Internetnutzer genauer betrachtet. Es wird auch einen tieferen Einblick darauf geworfen, dass Plattformen wie Meta gegen die DSGVO verstoßen hat und was für Auswirkungen dies für sie gebracht hat. Es ist außerdem interessant zu beobachten, wie bereit die Mehrheit der Nutzer ist zum Preis von 9,99€ ihre persönlichen Daten zu schützen bzw. sie nicht an Dritte weiterzugeben.

Literatur und Abbildungen

[1] Eigene Darstellung.

[2] Gigi Deppe. Kartellamt darf Datenschutz prüfen. https://www.tagesschau.de/ausland/europa/europaeischer-gerichtshof-meta-100.html, 07 2023.

[3] Nicolas Sacotte. Facebook werbefrei nutzen: Meta jetzt mit Abomodell. https://contentking.de/social-media-marketing/facebook-werbefrei-nutzen-meta-jetzt-mit-abomodell/, 11 2023.

Evaluierung der Objectives Key Results Methode im Unternehmen und Entwicklung eines maßgeschneiderten OKR-Reportingtools

Irem Sancak **Jürgen Koch**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma Fichtner IT Consulting GmbH, Stuttgart

Einleitung

In der dynamischen Welt der Unternehmensführung stehen Führungskräfte vor der Herausforderung, effektive Strategien zu entwickeln und zu implementieren, um den ständig wandelnden Marktanforderungen gerecht zu werden. Die Vielfalt der verfügbaren Strategiemethoden bietet eine breite Palette von Ansätzen zur Auswahl, von traditionellen bis zu modernen Methoden. Eine solche moderne Methode, die in den letzten Jahren zunehmend an Bedeutung gewonnen hat, ist die Objectives and Key Results (OKR) Methode [4].

Ziel der Abschlussarbeit

Diese Bachelorarbeit konzentriert sich auf die Entwicklung eines OKR-Reportingtools speziell zugeschnitten auf die Bedürfnisse von Fichtner IT-Consulting. Das übergeordnete Ziel besteht darin, ein effizientes Werkzeug zu schaffen, das eine klare Struktur für die Definition, Verfolgung und Auswertung von Zielen und deren Fortschritte bietet. Die Ausarbeitung basiert auf einer Evaluierung der OKR-Methode im Unternehmen. Hierbei werden die spezifischen Anforderungen von Fichtner IT-Consulting identifiziert und analysiert.

Was ist OKR?

OKR steht für Objectives and Key Results. Es handelt sich um eine moderne und agile Methode zur Zielsetzung und Leistungsbewertung, die dazu dient, Organisationen fokussierter, transparenter und agiler zu machen. OKR wurde ursprünglich von Intel entwickelt und später von Unternehmen wie Google und vielen Start-ups übernommen. Die OKR-Methode besteht aus zwei zentralen Komponenten:

- Objectives (Ziele): Objectives sind klare, inspirierende und herausfordernde Ziele, die eine Organisation erreichen möchte. Sie sollten motivieren und eine klare Ausrichtung für das gesamte Team bieten. Objectives sind in der Regel kurzfristig (typischerweise quartalsweise) und sollten verständlich formuliert sein.
- Key Results (Schlüsselergebnisse): Key Results sind messbare Ergebnisse, die den Fortschritt der festgelegten Ziele quantifizieren. Sie dienen als Indikatoren dafür, ob das Objective erreicht wurde oder nicht. Key Results sollten spezifisch, messbar, erreichbar, relevant und terminiert (SMART-Kriterien) sein [3].

Die OKR-Methode kann in sieben Grundprinzipien zusammengefasst werden, wie in Abbildung 1 dargestellt. [1]

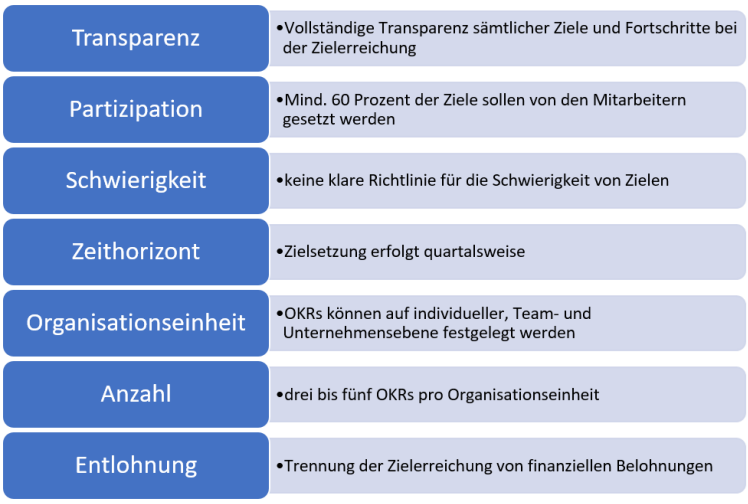


Abb. 1: OKR Prinzipien [2]

Umsetzung des Reportingtools

Das Reportingtool wird als Microsoft Teams App entwickelt, da die Unternehmenskommunikation über Teams erfolgt. Die App wird unter Verwendung der Microsoft Power Plattform erstellt, wobei Dataverse als Datenbank und PowerApp für die Benutzeroberfläche verwendet werden. Die Mitarbeiter sind bereits mit der Teams-Umgebung vertraut und haben Erfahrung mit Apps, die mithilfe von PowerApps entwickelt wurden. Dadurch wird die Einführung und Nutzung des OKR-Reportingtools erleichtert. Die Applikation wurde in Absprache mit den Stakeholdern definiert, und aus diesen Gesprächen ergaben sich die folgenden drei Hauptansichten für die Anwendung:

1. Allgemeines Dashboard: Hier werden alle Unternehmensziele (Objectives) aufgeführt und nach Zyklen gegliedert angezeigt. Für jedes Ziel werden relevante Informationen wie Bezeichnung, zugeordnete OKR-Gruppe, verantwortliche Personen, das Unternehmensziel und der Fortschritt dargestellt. Diese Ansicht ist für alle Mitarbeiter im Unternehmen zugänglich.
2. Persönliches Dashboard: Dieses Dashboard zeigt Key Results für diejenigen Personen oder Teams, die für sie verantwortlich sind. Es dient in erster Linie dazu, den Fortschritt der eigenen Ziele und der OKR-Gruppen zu verfolgen und einzutragen. Abbildung 2 zeigt diese Ansicht als Figma-Prototyp.
3. Administration: In diesem Bereich können individuell OKR-Gruppen, Zyklen und Unternehmensziele festgelegt werden.

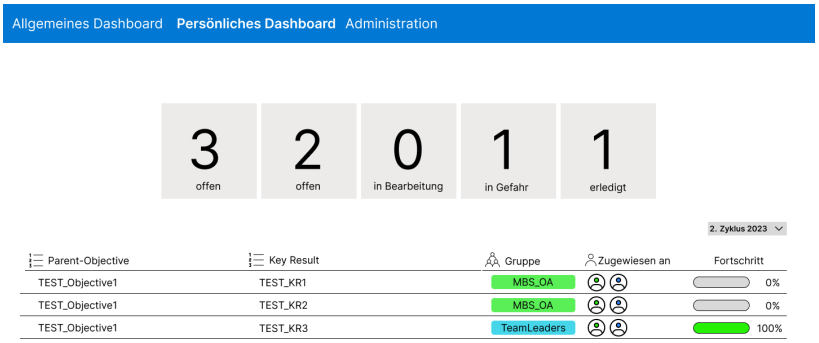


Abb. 2: Persönliches Dashboard - Figma Prototyp [2]

Ausblick

Das Minimum Viable Product (MVP) des Reportingtools wurde erfolgreich entwickelt und befindet sich nun in der Produktionsumgebung. Als nächster Schritt sind kontinuierliche Anpassungen geplant, die auf dem Feedback der Nutzer basieren. Die erhaltenen Rückmeldungen werden systematisch in einem Board festgehalten, und in enger Abstimmung mit den Stakeholdern werden die notwendigen Anpassungen priorisiert und demnach umgesetzt.

Literatur und Abbildungen

[1] Martin Artz and Hannes Döring. OKRs im deutschen Mittelstand. *Controlling & Management Review*, 2023.

[2] Eigene Darstellung.

[3] Daniela Kudernatsch. *Objectives and Key Results*. Haufe, 2021.

[4] scaleon Mooncamp. OKR IMPACT REPORT 2022. https://5074843.fs1.hubspotusercontent-na1.net/hubfs/5074843/PDFs/2022-okr-impact-report-de.pdf, 11 2022.

Cyber-Resilienz: Anwendbarkeit des Cyber Recovery Operational Frameworks in Unternehmen

Handan Sanli **Dirk Hesse**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma NTT DATA Deutschland SE, Stuttgart

Einleitung

Weltweit nehmen Cyberangriffe in ihrer Häufigkeit zu und Unternehmen sehen sich zunehmend mit diesen Bedrohungen konfrontiert [4]. Wird die gesamte Netzwerkinfrastruktur lahmgelegt, können die Geschäftskontinuität und damit die Existenz von Unternehmen gefährdet sein. Es ist von größter Wichtigkeit, dass Unternehmen schnell auf solche Vorfälle reagieren, um ihre Geschäftskontinuität aufrechterhalten zu können [1].
Cyber-Resilienz ist ein Konzept, das als Antwort auf die kontinuierlich wachsende Bedrohungslandschaft entwickelt wurde. Im Kern geht es darum, die Auswirkungen von Sicherheitsverletzungen zu minimieren und die Geschäftskontinuität aufrechtzuerhalten, selbst wenn ein Angriff stattfindet. Hierfür sind Recovery-Pläne erforderlich, die im Voraus entwickelt und getestet werden [1]. Obwohl es etablierte Standards gibt, die Unternehmen dabei unterstützen, ihre Cyber-Resilienz zu stärken, wird oft der Recovery-Aspekt vernachlässigt.

Zielsetzung

Das Ziel dieser Masterarbeit liegt darin, das Potenzial des Cyber Recovery Operational Frameworks auf die Anwendbarkeit in Unternehmen zu analysieren. Das Framework dient als Instrument, das Unternehmen Orientierung bietet, um ihre Cybersecurity-Praktiken zu verbessern. Dabei liegt der Fokus auf der Wiederherstellung von Systemen und Daten nach Cyberangriffen. An dieser Stelle ist es wichtig zu betonen, dass das Framework kein etablierter Standard ist und Belege für seine Wirksamkeit fehlen.

Relevanz

In etablierten Richtlinien wie jenen des National Institute of Standards and Technology (NIST) [2] und der Europäischen Union (EU) [3] wird die Bedeutung von Backups betont, die für die Wiederherstellung von IT-Systemen notwendig sind. Bei erfolgreichen Ransomware-Angriffe werden Backups entweder ganz oder teilweise verschlüsselt, was dazu führt, dass relevante Daten oder Business-Services nicht wiederhergestellt werden können [6]. Die Richtlinien zielen auf eine breite und ganzheitliche Cybersecurity ab, was dazu führt, dass sie allgemein gehalten sind.
Cyber-Recovery ist jedoch eine spezifische Disziplin, die individuelle Anpassungen je nach Branche oder Sektor erfordert. Um sich effektiv auf die Cyber-Recovery vorzubereiten, ist es daher wichtig, etablierte Richtlinien als Ausgangspunkt zu nutzen, aber sie um individuelle Anpassungen und ergänzende Maßnahmen zu erweitern, die den spezifischen Anforderungen und Risiken einer Organisation gerecht werden. Dies bedeutet, dass individuelle Cyber-Recovery-Pläne entwickelt werden müssen, die auf die spezifischen Prozesse abgestimmt sind [5]. Das Cyber Recovery Operational Framework bietet einen ganzheitlichen Ansatz, das relevante Aspekte für die Wiederherstellung von IT-Systemen und Business-Services berücksichtigt.

Cyber Recovery Operational Framework

Das Cyber Recovery Operational Framework [5] basiert auf verschiedenen Richtlinien wie der NIST CSF, ISO und NIST SP 800-184. Sie zielt darauf ab, Organisationen und Behörden eine operative Richtlinie bereitzustellen, um den Recovery-Prozess nach einem Cyber-Incident zu starten und durchzuführen. Des Weiteren basiert der Ansatz und die Konzeption des operativen Frameworks auf den praktischen Erfahrungen des Autors, welcher im Security Operations Center tätig ist. Das Ziel des Frameworks ist, auf akute Cyber-Incidents effektiv und effizient zu reagieren, indem es Unternehmen unterstützt eine individuelle Recovery-Anleitung zu erstellen.
Das Framework besteht aus 8 Kernkomponenten (Kategorien): Identify, Control, Map, Plan, Playbook, Measure, Test und Improve sowie aus 42 Subkompo-

nenten (Subkategorien). Diese werden im Folgenden aufgelistet.

1. Identify (RC.IP): Die erste Kernkomponente konzentriert sich auf die Identifizierung kritischer, technischer und nichttechnischer Vermögenswerte (Assets) einer Organisation, die im Falle eines Cyber-Incidents unverzüglich wiederhergestellt werden müssen.
2. Control (RC.CC): Der Schwerpunkt liegt hier auf dem Verständnis der Cyber-Recovery Maßnahmen, die die Organisation hat, und inwieweit diese Maßnahmen der Organisation helfen können, sich von einem Cyberangriff zu erholen.
3. Map (RC.DM): Diese Kernkomponente konzentriert sich auf die kritischen Dependencies und Assets der Organisation, um sie zu priorisieren und eine Reihenfolge festzulegen, die während des Recovery-Prozesses berücksichtigt wird.
4. Plan (RC.RP): Der Fokus dieser Komponente liegt auf der Erstellung einer zielorientierten Recovery-Planung, die Prozesse und Abläufe beinhaltet. Außerdem werden Rollen und Verantwortlichkeiten von internen sowie externen Stakeholder konkretisiert, die im Falle eines Cyber-Incidents kontaktiert werden müssen.
5. Playbook (RC.PL): Das Playbook soll Unternehmen dabei helfen, simulierte Angriffe (Cyber-Wargaming) zu dokumentieren, damit sie im Vorfeld auf verschiedene Arten von Cyberangriffen vorbereitet sind. Des Weiteren werden Recovery-Pläne getestet und evaluiert, um gegebenenfalls verbessert werden zu können.
6. Measure (RC.RM): Diese Komponente definiert Metriken, wie zum Beispiel Service und Operational Level Agreements (OLA). Anhand dieser Indikatoren wird gemessen, wie erfolgreich der Recovery-Prozess war.
7. Test (RC.RT): Eine Reihe von Tests werden durchgeführt, um sicherzustellen, dass die Recovery-Maßnahmen wirksam und angemessen sind. Es wird empfohlen, kontinuierlich zu testen.
8. Improve (RC.CI): Die Recovery-Pläne sollten stetig verbessert werden, indem Erkenntnisse aus Lessons Learned, Schulungen sowie Weiterbildungen miteinbezogen werden.

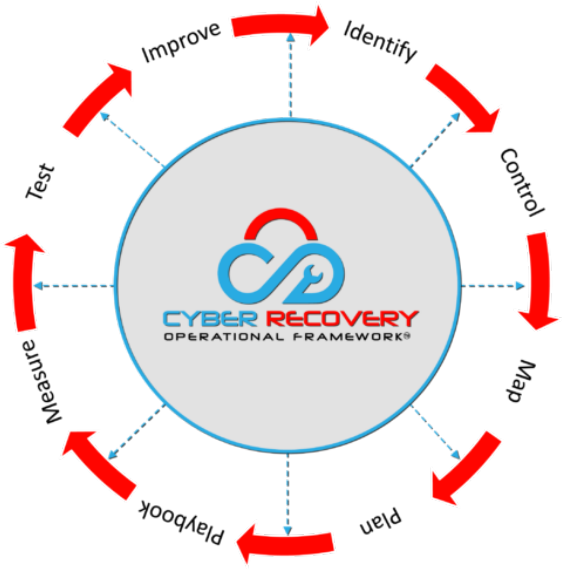


Abb. 1: Kernkomponenten des Cyber Recovery Operational Frameworks [5]

Überprüfung der Anwendbarbeit

Die praktische Anwendung des Recovery-Prozesses ist ein komplexes und aufwendiges Unterfangen. Um den Erfolg der Umsetzung sicherzustellen, müssen verschiedene Faktoren berücksichtigt werden. Mithilfe von Experteninterviews, die mit Fachleuten aus dem Bereich der Cybersecurity durchgeführt worden sind, wurden Herausforderungen während des Recovery-Prozesses identifiziert. Diese werden im Folgenden aufgelistet:

1. Mangelnde Vorbereitung
2. Unsicherheit über den Angriff
3. Zeitdruck
4. Backup-Verlust
5. Fehlende Kommunikation
6. Compliance-Anforderungen
7. Notfallmanagement

Stärken

Die Anwendung des Cyber Recovery Operational Frameworks bietet sowohl präventive, als auch reaktive Maßnahmen in Bezug auf Cybersecurity-Incidents. Im Zuge der Analyse wurde festgestellt, dass das Framework in zwei Hauptphasen unterteilt werden kann: Pre-Incident und Mid-Incident. Jede dieser Phasen beinhaltet spezifische Maßnahmen, die sicherstellen, dass Organisationen in der Lage sind, auf Sicherheitsvorfälle angemessen zu reagieren. Präventive Schritte umfassen RC.IP, RC.DM und RC.RT, während in der darauf folgenden Mid-Incident-Phase RC.RP und RC.RR von Bedeutung sind.

Schwächen

Um ein umfassendes Verständnis für alle Controls, Begriffe und Konzepte im Cyber Recovery Operational Framework zu entwickeln, ist ein umfangreiches Wissen im Bereich der Cybersecurity nötig.

Fazit

Die Arbeit liefert wertvolle Einblicke und Empfehlungen für Unternehmen, die ihre Cyber-Recovery-Strategien verbessern wollen, insbesondere vor dem Hintergrund zunehmender Ransomware-Angriffe. Es zeigt, wie ein umfassendes Framework dazu beitragen kann, die Cyber-Resilienz zu stärken und gleichzeitig die Fähigkeit eines Unternehmens, auf Cyber-Incidents angemessen zu reagieren.

Literatur und Abbildungen

[1] Deborah Bodeau, Richard Graubart, Jeffrey Picciotto, and Rosalie McQuaid. Cyber Resiliency Engineering Framework. https://www.mitre.org/sites/default/files/media/publication/11_4436_2.pdf, 09 2011.

[2] Alan Calder et al. Framework for Improving Critical Infrastructure Cybersecurity. https://nvlpubs.nist.gov/nistpubs/CSWP/NIST.CSWP.04162018.pdf, 04 2018.

[3] Union Europäische. DIRECTIVE (EU) 2022/2555 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32022L2555&from=EN, 12 2022.

[4] Marie-Claire Koch. Große Unternehmen: Im Schnitt 11.000 interne Sicherheitslücken. https://www.heise.de/news/Cybersecurity-Bericht-Durchschnittlich-11-000-Sicherheitsluecken-in-Unternehmen-9186767.html, 06 2023.

[5] Cyril Onwubiko. Focusing on the Recovery Aspects of Cyber Resilience. *IEEE*, page 13, 2020.

[6] Ariane Rüdiger. Cyberangriffe: Firmen schützen sich nicht gut vor Ransomware – und zahlen lieber. https://www.heise.de/news/Firmen-schuetzen-sich-nicht-richtig-vor-Ransomware-und-zahlen-einfach-lieber-7337678.html, 11 2022.

Optimierung einer Ampelschaltung mit Reinforcement-Learning in einer Verkehrssimulation

Viola Schaefer **Mirko Sonntag**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt an der Fakultät Informatik und Informationstechnik, Esslingen

Einleitung

Reinforcement-Learning ist ein immer populärer werdender Ansatz zur Lösung von Problemstellungen in komplexen Systemen. Dabei gibt es zahlreiche Algorithmen, die eine Optimierung des Lernprozesses anstreben. Die Organisation des Straßenverkehrs stellt ein solches komplexes System dar, welches durch die große und vielfältige Anzahl an Verkehrsteilnehmern und diverser anderer Faktoren geprägt ist. Dass solch ein System nicht immer reibungslos funktionieren kann ist absehbar. Dabei haben Ampeln einen großen Einfluss auf den Verkehrsfluss und, wie es häufig bei komplexen Systemen der Fall ist, kann selbst eine Ampelschaltung, die durch verschiedene Programme auf gewisse Situationen vorbereitet ist, nicht hinreichend sein, um Unterbrechungen des Verkehrsflusses zu vermeiden. Als Lösungsansatz soll in dieser Arbeit deswegen eine Ampelschaltung mithilfe von Reinforcement-Learning feingranularer optimiert werden. Im Anschluss soll der Einfluss dieser Optimierung auf den Verkehrsfluss untersucht werden.

Grundlagen

Als Ansatz für das Reinforcement-Learning ist hier der Deep-Q-Learning-Algorithmus gewählt. Deep-Q-Learning ist auf Systeme ausgelegt, in denen nicht jede mögliche Situation vorhergesagt werden kann, und die Umgebung, in der das System sich befindet, komplex und nicht zwingend bekannt ist. Dabei beruht Deep-Q-Learning auf dem Prinzip, dass das System einen Zustand annimmt und durch ein neuronales Netzwerk eine entsprechende Aktion bestimmt, wie in Abbildung 1 dargestellt. Dies geschieht im Algorithmus durch eine Aktionswertfunktion, die sogenannten Q-Werte, die die erwarteten kumulativen Belohnungen von Zustands-Aktionspaaren schätzt. Dieser Algorithmus ist für viele Bereiche von Bedeutung. Unter Anderem auch für das autonome Fahren, künstliche Intelligenz im Spiel und auch in der Robotik [4].

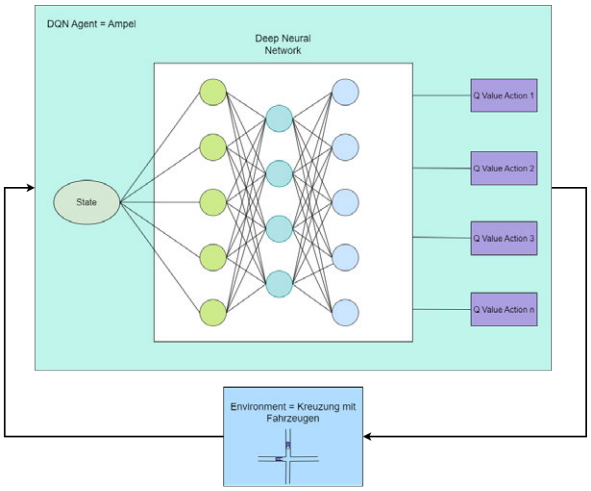


Abb. 1: Abbildung 1: Deep-Q-Learning [3]

Technologien

Für die Umsetzung ist zunächst das Tool SUMO (Simulation of Urban MObility) gewählt worden. SUMO ist ein quelloffenes und hochportables Verkehrssimulationspaket und wurde hauptsächlich für den Umgang mit großen Netzwerken entwickelt. Mithilfe von SUMO und dessen integrierten Tools wie den OSM Web Wizard (Open Street Map) kann man aus realem Kartenmaterial Simulationen erstellen und weitere Faktoren des Netzwerkes wie fehlende Bushaltestellen einbauen [2]. Für die Umsetzung des Reinforcement-Learnings soll das Framework SUMO-RL verwendet werden [1]. SUMO-RL nimmt die in SUMO erstellte Simulation als Input an und trainiert eine Ampelschaltung mithilfe eines selbst hinzugefügten Algorithmus oder des bereits vorgegebenen DQN-Algorithmus. Das Ergebnis kann mithilfe dieses Frameworks auch graphisch dargestellt werden, was die Auswertung der Simulation vereinfacht.

Simulation

Für die Simulation ist, wie in Abbildung 2 zu sehen, die Kreuzung gewählt worden, bei der die Rotenackerstraße zur Hochschule Esslingen beim Standort Flandernstraße führt. Während die Kreuzung an sich relativ simpel gestaltet ist, sind hier Faktoren wie die Bushaltestellen und Fußgängerüberwege zu beachten. Auch die Tatsache, dass die Spur zum Linksabbiegen in die Flandernstraße nicht über eine Ampelphase gesteuert ist, ist hierbei auf den Verkehrsfluss von Relevanz, da sich bei einer langen Grünphase der Ampeln leicht Stauungen auf dieser Spur bilden können. Die Simulation ist aus einigen Dateien zusammengesetzt, von denen drei von besonderer Relevanz sind. Das Netzwerk selber, also die Straße mit ihren statischen Elementen wird durch eine **.net.xml-Datei* beschrieben und ist als Parameter im SUMO-RL zu verwenden. Ebenfalls als Parameter wird eine **.rou.xml-Datei* benötigt, die die Verkehrsteilnehmer wie Automobile in der Simulation darstellt. Um diese Simulation auch über die Kommandozeile starten zu können wird eine **.sumocfg-Datei* benötigt, die in ihrer Essenz alle für die Simulation benötigten Dateien verbindet. Zusätzlich ist es möglich, das Ergebnis des Reinforcement-Learnings zu optimieren, indem tatsächliche Daten wie der Busfahrplan und Zähldaten dieser Kreuzung in die Simulation integriert werden.

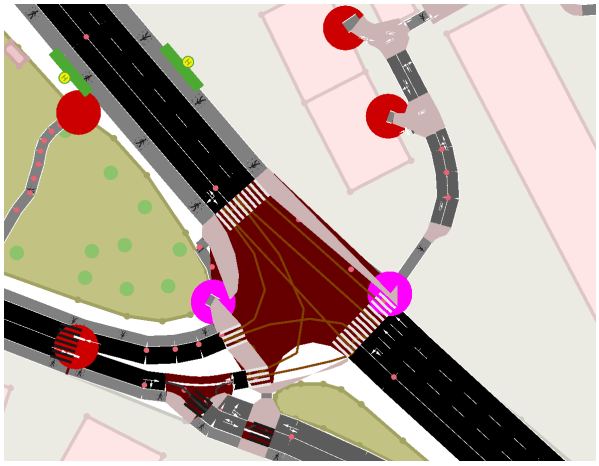


Abb. 2: Abbildung 2: SUMO-Simulation in Netedit [3]

Ausblick

In dieser Arbeit ist der Einfluss einer Reinforcement-Learning-optimierten Ampelschaltung auf den Verkehrsfluss zu untersuchen. Dafür muss zuerst festgelegt werden, wie diese Optimierung definiert werden soll und, vor Allem, wie diese Optimierung zu messen ist. Ein optimaler Verkehrsfluss ist hier so definiert, dass ein Verkehrsteilnehmer eine möglichst geringe Wartezeit an der Reinforcement-Learning-optimierten Ampelschaltung verbringt. Damit ist der Messwert, an dem der Einfluss der Optimierung gemessen wird, die durchschnittliche Wartezeit eines Verkehrsteilnehmers. Als Erfolg des Ansatzes der durch Reinforcement-Learning optimierten Ampelschaltung wäre demnach eine messbare Verringerung der durchschnittlichen Wartezeit zu bezeichnen.

Literatur und Abbildungen

[1] Lucas N. ALEGRE. SUMO-RL. https://github.com/LucasAlegre/sumo-rl, 2019.

[2] Pablo Alvarez Lopes, Jakob Erdmann, Laura Bieker-Walz, Michael Behrisch, and Yun-Pang Fl. Microscopic Traffic Simulation using SUMO. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 2575–2582. IEEE, 2018.

[3] Eigene Darstellung.

[4] Niklas Lang. Q-Learning – einfach erklärt. https://databasecamp.de/ki/q-learning, 05 2023.

A Framework for Fuzzing Tests with Software Defined Radio

Carmen Schaeffler **Dominik Schoop**

Department of Computer Science and Engineering, Esslingen University

Work carried out at proofnet GmbH, Böblingen

Motivation and Problem

Vehicles' communication should be safe and secure. Modern vehicles communicate actively with other vehicles and different traffic infrastructure components, as well as receive passive information from other systems over various protocols using broadcasting [2], [10]. In 2022, the majority of radio consumers listened to radio in an automotive vehicle on a weekly basis [11]. Currently, the standard protocol for broadcasting information to vehicles over audio is Digital Audio Broadcasting + (DAB+) [10]. As DAB+ is required for registering new cars in various countries, the coverage is growing [3].

Poorly designed implementations, including audio receiver implementations, may result in a rising number of vulnerabilities and potential attacks, e.g. remote code execution (RCE) or Denial-of-Service (DoS). Proofs of concept and exploit code are available and utilized in different attacks. If an implementation of a protocol is vulnerable to manipulation, it may thus provide opportunities for attackers to intercept the communication, manipulate the data, and to gain access to the vehicle [8].

Since DAB(+) lacks manipulation protection, it is an obvious remote attack vector. The widespread usage in combination with broadcasting allows to affect multiple targets at once. Since a lot of different audio protocols are sitting on top of DAB+, there is a broad attack surface [2].

In case of an attack, devices may be compromised and additionally process manipulated data other devices rely on [8]. A non-secure system architecture may allow further access to other parts of the affected vehicles [2]. Vulnerabilities may have extensive consequences because multiple vehicles can be attacked at once via broadcasting. Therefore, considering the possibility of intercepting, injecting, jamming, or spoofing the communication, broadcasting is a critical component of the vehicular network [6]. The implementation of a broadcast protocol thus in particular requires tests for correctness and validation against standards to verify that the application performs the task as specified instead of entering an unintended state. Nowadays, various software testing techniques, e.g. fuzzing, are available to run tests to find security issues and protect them against well-known and newly discovered security threats. This research focuses on developing a framework to fuzz implementations of audio broadcast protocols in the automotive industry.

Design and Concept

Fuzzing is an automated method for software testing to find various security-related bugs, vulnerabilities, identify software failure, and to avoid unexpected input leading to an unintentional system state [5], [7]. It is a simple and customizable process of running a program with generated input data and monitoring the program's behavior, see Figure 1. Black-box fuzzing is a fuzzing method where testers do not have internal information about the system. Only the input and output behavior is observable [7].

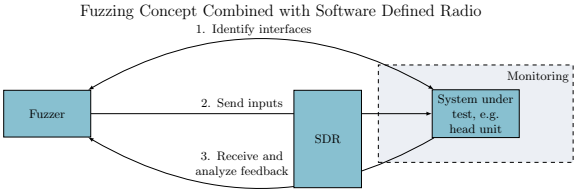


Fig. 1: Concept of a Fuzzing Life Cycle Combined with Software Defined Radio in the Automotive Industry [7]

The usage of Software Defined Radios (SDR)s has been evolving over the last few years. An SDR device is a radio communication system which includes software-configurable and software-implemented components for capturing and processing radio signals. SDR devices offer a lot of flexibility, re-configurability, support various protocols, and are suitable for many use cases,

e.g. prototyping, penetration testing, or deployment of flexible radio systems [1]. In order to employ an SDR device for fuzz testing audio broadcast protocol implementations, this research consists of the following parts:
First, the current SDR hardware device market situation is evaluated. The market research consists of an overview of available SDR hardware options and a comparison of different characteristics and costs for each device. Choosing a suitable SDR hardware device for the fuzzing framework requires a thorough evaluation of several characteristics, such as radio spectrum capabilities, bandwidth range, and cost of implementation [8].
Second, the research includes an assessment of available software projects in the field of SDR for penetration testing and fuzzing of protocol implementations. The scope of existing fuzzers for broadcast protocols is analyzed.
Based on the previous steps' results, DAB+ is selected as the audio broadcast protocol to subsequently build and implement a black-box fuzzing framework. After the successful development of this fuzzing framework for test cases, real-world scenarios are implemented in the fuzzing framework.

Evaluation

In order to verify the device's functionality and to get an understanding of audio broadcast protocols, an SDR device is used to broadcast and capture radio signals, e.g. FM and DAB+. Figure 2 shows a generic information flow of radio frequency (RF) communication systems. An evaluation of the DAB+ protocol structure is necessary to get an understanding of the input structure, the information flow and the metadata, which should be fuzzed later.

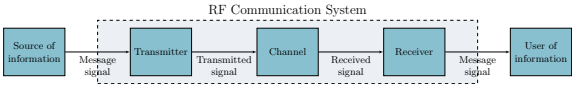


Fig. 2: Elements and Information Flow of an RF Communication System [4]

The developed fuzzing framework has to fulfill the following criteria: It is easily extendable by adding further protocols, implementations, and SDR devices. The framework has to be re-usable and flexible. The process of sending input to a system's interface has to be automated by the fuzzer. Further, the framework has to include system monitoring. This is important to detect and store occurring system failures and their conditions to reproduce them later, if necessary.

Related Work

The following projects use fuzzing to identify vulnerabilities targeting digital radios. DABle fuzzer is a tool for fuzzing images transmitted over DAB(+). With this fuzzer, it was already possible to find vulnerabilities, such as code execution and format string bugs [2]. TumbleRF is a general software framework for fuzzing RF protocols. It includes multiple interfaces and is easily expandable regarding further protocols or other radio interfaces [5]. Scapy-radio is a modified version of the packet manipulation tool Scapy combined with SDR using GNU Radio, a widely used open-source software development kit to test radio communication systems [9] [8].

Result

The result of this research will be a framework for fuzz testing of a DAB+ implementation with SDR in the automotive industry. Currently, the SDR device is able to receive and broadcast radio signals. The goal is to use the framework for fuzzing the DAB+ implementation of an automotive head unit to find and analyze security bugs.

References and figures

[1] Travis F. Collins et al. *Software-defined Radio for Engineers*. Artech House, 2018.

[2] Andy Davis. Broadcasting your attack: Security testing DAB radio in cars. https://troopers.de/media/filer_public/18/4f/184fa903-3610-4647-9cb0-bb7644d3f295/broadcasting_your_attack_security_testing_dab_radio_in_cars.pdf, 2016.

[3] Digitalradio Deutschland eV. Informationen zum Digitalradio-Beschluss. https://www.dabplus.de/tkg/, 2023.

[4] Simon Haykin. *Communication Systems*. John Wiley & Sons, 4 edition, 2001.

[5] Matt Knight. Designing RF Fuzzing Tools to Expose PHY Layer Vulnerabilities. https://www.gnuradio.org/grcon/grcon18/presentations/TumbleRF_RF_Fuzzing_Made_Easy/07-MattKnight-TumbleRF.pdf, 2018.

[6] Elnaz Limouchi et al. Fuzzy Logic-Based Broadcast in Vehicular Ad Hoc Networks. In *2016 IEEE 84th Vehicular Technology Conference (VTC-Fall)*, pages 1–5. IEEE, 2016.

[7] Valentin J.M. Manès et al. The Art, Science, and Engineering of Fuzzing: A Survey. In *IEEE Transactions on Software Engineering*, volume 47, pages 2312–2331. IEEE, 2021.

[8] Jean-Michel Picod et al. Bringing Software Defined Radio to the Penetration Testing Community. https://www.blackhat.com/docs/us-14/materials/us-14-Picod-Bringing-Software-Defined-Radio-To-The-Penetration-Testing-Community-WP.pdf, 2014.

[9] GNU Radio project. About GNU Radio. https://www.gnuradio.org/about/, 2023.

[10] WorldDAB project. DAB+ radio: as standard in new European cars. https://www.worlddab.org/system/news/documents/000/012/060/original/WorldDAB_press_release_DAB-PLUS_RADIO_AS_STANDARD_IN_NEW_EUROPEAN_CARS_23.6.21.pdf?, 2021.

[11] Matthieu Rawolle. AUDIO IN CARS. https://www.worlddab.org/files/document/file/4868/2.2_2023-06-21_150623_Automotive_WorldDAB_Matthieu_Rawolle_updated_v_for_WDAB_upload.pdf?1687340460, 06 2023.

Sensor Fusion mit neuronalen Netzen zur Rekonstruktion der Fahrzeugbeschleunigung

Jannik Scheider **Markus Enzweiler**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma Fraunhofer-Institut für Produktionstechnik und Automatisierung IPA, Stuttgart

Motivation

Die steigende Belastung durch Mikroplastikemissionen aus dem Abrieb von Autoreifen ist zu einer akuten Herausforderung geworden, die nicht nur die Umwelt, sondern auch die Gesundheit der Menschen betrifft. Gemäß einer Studie des Fraunhofer-Instituts UMSICHT aus dem Jahr 2018 macht der Reifenabrieb etwa ein Drittel aller Mikroplastik-Emissionen in Deutschland aus. [2] Prognosen bis zum Jahr 2050 zeigen, dass bis zu 90% der Partikel, die durch den Straßenverkehr emittiert werden, auf nicht abgasbedingte Emissionen zurückzuführen sind. Das bedeutet, dass diese hauptsächlich von Reifen und Bremsen stammen.
Die Europäische Kommission hat im November 2022 einen Entwurf für die neue Euro-7-Norm präsentiert, um Neufahrzeuge auf Europas Straßen umweltfreundlicher zu machen. Die Euro-7-Norm regelt nicht nur die Emissionen aus dem Auspuff, sondern auch die Partikel aus Bremsen und Reifen, die bei Elektrofahrzeugen die größten Emissionsquellen darstellen. Bis 2035 müssen die NOx- und Partikelemissionen aus Auspuff, Bremsen und Reifen erheblich reduziert werden, wodurch Automobilhersteller zum Handeln aufgefordert sind. [4]

Zielsetzung

Moderne Fahrzeuge erzeugen eine große Menge an Daten, die von Sensoren, Kameras und anderen eingebetteten Systemen stammen. Die kontinuierliche Übertragung dieser Daten auf einen zentralen Server ist jedoch mit erheblichen Kosten verbunden. Aus diesem Grund wird angestrebt, die Anzahl der im Fahrzeug integrierten Sensoren effizient zu beschränken, um nur relevante Datenströme zu generieren. Es wird evaluiert, ob anstatt einer direkten Messung der Querbeschleunigung mit einem Sensor stattdessen eine indirekte Berechnung der genannten Kenngröße aus anderen Sensoren mithilfe von künstlicher Intelligenz möglich ist. Die Messdatenerfassung erfolgt im Rahmen so genannter Kampagnen. Dieser Begriff bezeichnet einen vordefinierten Anwendungsfall, der als Software in einem Testfahrzeug installiert ist und festlegt, welche Sensoren während einer Testfahrt Daten erfassen und an einen Server weiterleiten. Die gewonnenen Erkenntnisse aus dieser Untersuchung sollen letztendlich einen wesentlichen Beitrag zur umfassenden Analyse des Einflusses von Beschleunigung auf den Reifenverschleiß leisten. Es ist wichtig zu erwähnen, dass diese Arbeit sich ausschließlich auf die Vorhersage der Beschleunigung konzentriert und nicht auf die direkte Bestimmung des Reifenverschleißes.

Sensor Fusion

Sensorfusion bezieht sich auf die Kombination von Sensordaten oder von aus Sensordaten abgeleiteten Daten wie in Abbildung 1 zu sehen, so dass die daraus resultierenden Informationen in gewisser Weise besser sind, als dies bei der Verwendung der einzelnen Quellen möglich wäre. Die Methoden um eine Sensordatenfusion durchzuführen reichen von Statistischen Methoden wie der Bayesschen Statistik hinzu der Fuzzy Logik, und neuesten Methoden mit neuronalen Netzen. [3]

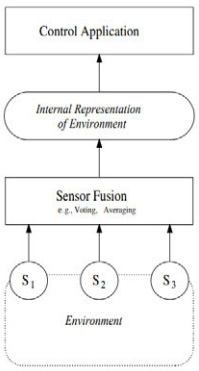


Abb. 1: Sensor Fusion Ablauf [3]

Temporal Fusion Transformer

Der Temporal Fusion Transformer (TFT) ist eine attention basierte Architektur, die multihorizon Prognosen mit interpretierbaren Erkenntnissen über die zeitliche Dynamik verbindet. Um zeitliche Beziehungen auf verschiedenen Ebenen zu erlernen, verwendet TFT rekurrente Schichten für die lokale Verarbeitung und interpretierbare self-attention Schichten für langfristige Abhängigkeiten. TFT verwendet spezialisierte Komponenten zur Auswahl relevanter Merkmale und eine Reihe von Gating-Schichten zur Unterdrückung nicht benötigter Komponenten, was eine hohe Leistungsfähigkeit in einem breiten Spektrum von Szenarien ermöglicht. [1] Die TFT Architektur ist in Abbildung 2 dargestellt.

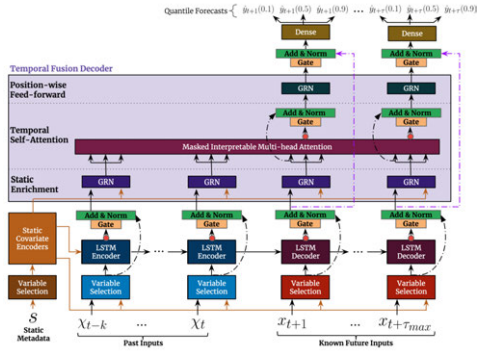


Abb. 2: Temporal Fusion Transformer Architektur [5]

Vorgehensweise

Um die oben genannte Aufgabenstellung zu lösen erfolgt zunächst die Datensammlung, während einer geplanten Testfahrt. Mit der Smartphone App *phyphox* wird die Querbeschleunigung aufgezeichnet . Diese App ermöglicht den Zugriff auf die integrierten Sensoren eines Smartphones und zeichnet die benötigten Daten präzise auf. Während der Testfahrt wird das Smartphone auf eine rutschfeste Matte auf dem Armaturenbrett gelegt. So werden sowohl die Sensordaten im Fahrzeug als auch die Beschleunigungsdaten mittels einer Smartphone-App erfasst.
Im Anschluss an die Datensammlung steht die Implementierung einer Datenpipeline im Fokus. Die Herausforderung besteht darin, die zwei Datenquellen zu integrieren und notwendige Transformationsschritte anzuwenden, um eine konsistente Grundlage für die weitere Analyse zu schaffen.
Nach der erfolgreichen Implementierung der Datenpipeline ist der nächste Schritt die Implementierung eines KI-Modells. Das definierte Ziel wird dabei als Zeitreihenregressionsproblem betrachtet. Hierbei gibt es verschiedene Ansätze zur Entwicklung eines KI-Modells, darunter LSTM- oder GRU-Architekturen sowie Transformer mit verschiedenen Modifikationen. In diesem Kontext wird für diese Arbeit der Temporal Fusion Transformer eingesetzt. Diese Variante des Transformers ist speziell für Zeitreihenprobleme konzipiert. Anschließend erfolgt das Training eines Modells.

Ausblick

Nach erfolgreichem ersten Training eines Temporal Fusion Transformers auf den vorhandenen Daten werden in den kommenden Schritten die Ergebnisse ausgewertet und weitere Modelle mit unterschiedlichen Sensoren trainiert, um die Relevanz der einzelnen Sensoren für die Vorhersagegenauigkeit zu untersuchen.

Literatur und Abbildungen

[1] Sercan Arik and Tomas Pfister. Interpretable Deep Learning for Time Series Forecasting. https://blog.research.google/2021/12/interpretable-deep-learning-for-time.html, 12 2021.

[2] Jürgen Bertling, Bertling Ralf, and Leandra Hamann. Kunststoffe in der Umwelt: Mikro- und Makro- plastik. Ursachen, Mengen, Umweltschicksale, Wirkungen, Lösungsansätze, Empfehlungen. https://publica.fraunhofer.de/entities/publication/97906872-d1ee-401e-a639-492f3c239952/details, 06 2018.

[3] Wilfried Elmenreich. An Introduction to Sensor Fusion. https://www.researchgate.net/publication/267771481_An_Introduction_to_Sensor_Fusion, 11 2002.

[4] Kommision Europäische. Neue Abgasnorm Euro-7: Die Fakten. https://germany.representation.ec.europa.eu/neue-abgasnorm-euro-7-die-fakten_de, 11 2022.

[5] Bryan Lim et al. Temporal Fusion Transformers for Interpretable Multi-horizon Time Series Forecasting. https://arxiv.org/pdf/1912.09363.pdf, 09 2020.

Vergleich von KI-Services zur Implementation eines KI-Chatbots als Reporting- und Datenabfrage-Funktion in einer Fußballanalyse-Anwendung

Philipp Schimmer **Mirko Sonntag**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma W11k GmbH, Esslingen

Einleitung

In der heutigen Zeit, in der die Bedeutung von künstlicher Intelligenz (KI) rasant zunimmt, eröffnen sich neue Möglichkeiten in der Technologiewelt. KI-Technologien entwickeln sich stark weiter und werden immer mehr zu einem integralen Bestandteil vieler Branchen und Anwendungen und mit zur zentralen Zukunftstechnologie. Gerade durch den Durchbruch von ChatGPT, der mit den GPT Sprachmodellen OpenAI betrieben wird, haben Chatbots, hinter denen große Sprachmodelle stehen, immer mehr Relevanz gefunden. In diesem Kontext stellt das Coaching-Cockpit-Projekt von Die Ligen, ein innovatives Fußballspiel-Analyse-Tool, ein spannendes Anwendungsfeld für KI-Chatbots dar. Die Ligen produziert über ein Netzwerk von Filmern Videoaufnahmen von Fußballspielen direkt vor Ort am Spielort. Die Aufnahmen werden kategorisiert und relevante Ereignisse von einem Team von Fußball-Analysten detailliert aufbereitet. Die Daten und Informationen landen dann auf einem visuellen Dashboard im „Coaching-Cockpit“ zur Analyse von z.B. Trainern als Spielvorbereitung. Die Integration eines KI-Chatbots in dieses System könnte die Art und Weise, wie Benutzer auf Informationen zugreifen und interagieren, verbessern und erweitern. Indem der Chatbot direkten Zugriff auf die Daten verschiedener Fußballspiele, Ligen und Teams erhält, wird der Zugang zu diesen Informationen für die Benutzer erheblich erleichtert. Der Chatbot zielt darauf ab, die Grenzen des visuellen Dashboards zu überwinden, indem er den Benutzern ermöglicht, spezifische Fragen zu Daten zu stellen, die sonst nicht direkt auf dem Dashboard zu finden sind oder sich auf verschiedenen Unterseiten befinden. Dies soll die Informationsbeschaffung beschleunigen und effizienter gestalten. Ein solcher KI-Chatbot, der auf die relevanten Daten der Anwendung zugreifen kann, soll somit einen signifikanten Mehrwert für das DieLigen-Projekt bieten.

Ziel dieser Arbeit

Das Ziel dieser Bachelorarbeit ist eine technische Analyse und prototypische Implementierung eines KI-Chatbots, der als Reporting- und Datenabfrage-Funktion im Rahmen des Fußballspiel-Analyse-Tools „DieLigen“ fungiert. Ein wichtiger Aspekt ist dabei die technische Analyse und der Vergleich verschiedener KI-Technologien und Plattformen. Diese Analyse umfasst die Bewertung von Cloud-basierten KI-Services sowie individuell entwickelten Lösungen unter Verwendung neuester KI-APIs. Zur Auswahl der geeigneten Technologie für die prototypische Implementierung werden die KI-Technologien basierend auf Kriterien wie Antwortqualität, Antwortzeit, Flexibilität, Benutzerfreundlichkeit und der Fähigkeit, komplexe Anfragen zu verarbeiten, verglichen. Durch die Erreichung dieser Ziele strebt die Arbeit an, einen Beitrag zur Erkenntnis darüber zu leisten, wie KI-Chatbots in Analyse-Tools und Kundenportalen implementiert werden können, um den Zugang zu und die Nutzbarkeit von Daten und Informationen zu verbessern. Dabei sollen folgende Forschungsfragen beantwortet werden: *Inwiefern verbessert die Implementierung eines KI-Chatbots die Effektivität der Informationsbeschaffung und Datenabfrage im „Die Ligen“-Dashboard? Welche KI-Services oder technischen Lösungen bieten sich an in Bezug auf die Implementierung eines Chatbots als Reporting- und Datenabfrage-Funktion? Inwiefern lässt sich ein KI-Chatbot zur Informationsbeschaffung und Datenabfrage in einem Kundenportal umsetzen?*

Künstliche Intelligenz

Künstliche Intelligenz (KI) ist durch ihre weitverbreitete Anwendung und Untersuchung nicht einheitlich definiert. Das liegt auch teilweise an der Herausforderung bei der Definition von „Intelligenz“. Im Allgemeinen bezeichnen viele Intelligenz als die Fähigkeit, aus Erfahrungen zu lernen, um Probleme zu lösen und

sich in unbekannten Umgebungen und Situationen zurechtzufinden. [6] [4]
Unter Berücksichtigung dieser Interpretation von Intelligenz besteht das Ziel der Künstlichen Intelligenz darin, Computerprogrammen die Fähigkeiten zu verleihen, menschliche Intelligenz zu simulieren und in einigen Fällen sogar zu übertreffen. Die Hauptaspekte der KI sind also die Fähigkeiten zu lernen, unbekannte Situationen zu verstehen und sich darin zurechtzufinden. Künstliche Intelligenz basiert dabei meistens auf weitere Technologien wie in Abbildung 1 zu sehen. Neuronale Netze, inspiriert vom menschlichen Gehirn werden zur Konzeptionierung von Modellen genutzt. Maschinelles Lernen und Deep Learning zum Trainieren von künstlicher Intelligenz. [4] [3]

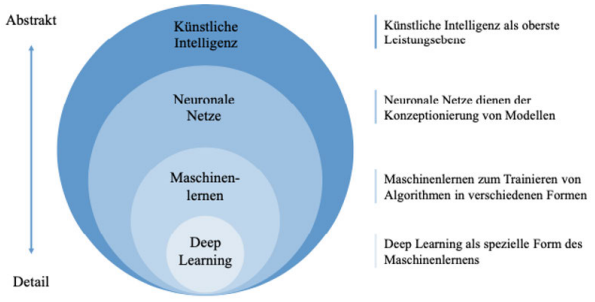


Abb. 1: Teildisziplinen von KI [5]

Chatbots, Sprachmodelle und NLP

Ein Chatbot ist ein System, das in der Lage ist, mit Menschen durch natürliche Sprache zu kommunizieren und autonom Aufgaben zu erledigen, wodurch eine Mensch-Maschinen-Schnittstelle entsteht. Der Chatbot wird vom Benutzer über natürliche geschriebene Sprache angesprochen, in manchen Fällen aber auch über gesprochene Sprache. Chatbots werden entweder über regelbasierte Systeme oder Künstliche Intelligenz betrieben. Bei KI-Chatbots kommen häufig Technologien wie maschinelles Lernen oder Deep Learning, Natural Language Processing (NLP) und große Sprachmodelle (LLMs) zum Einsatz. [7] [2]
Natural Language Processing (NLP) als bedeutender Teilbereich der Künstlichen Intelligenz befasst sich mit der Verarbeitung von natürlicher Sprache durch den Computer. NLP ermöglicht es Maschinen, gesprochene oder geschriebene Sprache zu verstehen, zu interpretieren und auch eigenständig zu generieren. NLP kann weiterhin in zwei Hauptbereiche unterteilt werden: Natural Language Understanding (NLU) für das verstehen natürlicher Sprachen und Natural Language Generation (NLG) für das Generieren natürlicher Sprache. [1]
Große Sprachmodelle (LLMs), die auf der Verarbeitung natürlicher Sprache (NLP) und künstlichen neuronalen Netzwerken basieren, werden mit umfangreichen Textdatensätzen trainiert. Diese Modelle sind darauf ausgelegt, Texte zu erkennen, zu generieren, zu übersetzen und Sprache im Allgemeinen zu verarbeiten oder spezifische Aufgaben auszuführen. Sie sind in der Lage, vielfältige Aufgaben zu bewältigen, wie das Erstellen oder Zusammenfassen von Texten, das Beantworten von Fragen basierend auf einer Wissensbasis und sogar das Schreiben von Programmcode.

Umsetzung

Die Bachelorarbeit beginnt mit einer umfassenden Literaturrecherche und der Erarbeitung von Grundlagen im Bereich der künstlichen Intelligenz (KI), mit einem speziellen Fokus auf der Einführung und dem Verständnis von Chatbots. Dieser Schritt dient dazu, ein solides Verständnis für die Technologien zu schaffen. Anschließend werden die spezifischen Anforderungen und Use Cases für den Chatbot im Kontext von „DieLigen“ festgelegt. Diese Phase ist wichtig, um einen klaren und strukturierten Rahmen für die Implementierung zu schaffen. Es folgt eine eingehende Untersuchung der verschiedenen verfügbaren KI-Services und technischen Lösungen. Hierbei liegt der Fokus sowohl auf Cloud-Plattformen großer Tech-Unternehmen wie Microsoft, Google und IBM als auch auf selbst entwickelten Lösungen mit Anbindungen an KI-APIs wie die von OpenAI. Um eine fundierte Entscheidung für die geeignetste Technologie treffen zu können, werden Vergleichskriterien festgelegt. Diese Kriterien helfen, die Eignung der verschiedenen KI-Services und technischen Lösungen zu bewerten. Ein wesentlicher Teil der Analyse besteht aus der Durchführung von Umsetzbarkeits- und Performancetests mit Testdaten. Diese Testdaten bestehen aus zwei Tabellen, die eine beinhaltet Fußballspieler und die andere Strafen zugeordnet über eine ID zu den Fußballspielern. Diese Testdaten werden den KI-Services über unterschiedliche Wege bereitgestellt, wie CSV- oder Exceldateien oder andere Technologien, die direkt von den KI-Services geboten wird. Ziel ist es, die Fähigkeit der KI-Technologien zu überprüfen, Daten aus mehreren Quellen zu verarbeiten und präzise Antworten auf Benutzeranfragen zu generieren. Nach der Auswertung der Tests wird eine geeignete Technologie für die prototypische Umsetzung ausgewählt. Es folgt die Entwicklung eines detaillierten Plans für die Implementierung des Chatbots. Der nächste Schritt ist die praktische Umsetzung dieses Plans, bei der ein funktionsfähiger Prototyp des Chatbots entwickelt und in das DieLigen-Projekt integriert wird. Abschließend erfolgt eine umfassende Auswertung der Umsetzung. Dieser Schritt dient dazu, die gewonnenen Erkenntnisse zusammenzufassen und die Forschungsfragen der Arbeit umfassend zu beantworten. Durch

diese strukturierte und methodische Vorgehensweise wird sichergestellt, dass der entwickelte KI-Chatbot den spezifischen Anforderungen des DieLigen-Projekts gerecht wird und einen signifikanten Mehrwert für die Nutzer bietet.

Konzept für die Implementierung

Für die Implementierung des Prototyps wurde die, während der Bearbeitung der technischen Analyse neu erschienene, Assistant-API von OpenAI ausgewählt, die auf den GPT-Modellen basiert, welche auch hinter ChatGPT stehen. Zum aktuellen Zeitpunkt der Arbeit befinden sich die Konzeption und Umsetzung noch in der Planungsphase. Im Rahmen des DieLigen-Projekts wird neben dem bestehenden Java-Backend ein zusätzliches NodeJS-basiertes NestJS-Backend entwickelt. Dieses Backend dient als Schnittstelle für den Chatbot. Wenn ein Benutzer im Frontend den Chatbot aktiviert, wird über einen REST-Endpunkt eine Anfrage an das Backend gesendet. Die Fußball relevanten Daten für den aktuellen Wettkampf werden aus der DieLigen-Datenbank in verschiedene CSV-Dateien geschrieben. Dabei wird für jede Tabelle aus der Datenbank eine CSV-Datei genutzt. Diese CSV-Dateien werden dann über die Assistant-API der KI zur Verfügung gestellt. Eine direkte Anbindung an die Datenbank ist nicht möglich und außerdem kann somit auch sichergestellt werden, dass nur die relevanten Daten an die KI weitergeleitet werden. Der Benutzer kann im Frontend seine Fragen zu den Daten in ein Chatfenster eingeben, die an das Backend weitergeleitet und von dort aus über die API an die KI gesendet werden. Die Antwort der KI, basierend auf den bereitgestellten Daten, wird dann über das Backend zurück an das Frontend übermittelt. Sobald der Chatbot im Frontend vom Benutzer geschlossen wird, werden die CSV-Dateien gelöscht. Dies stellt sicher, dass bei einem erneuten Aufrufen des Chatbots die Daten wieder aktuell aus der Datenbank geladen werden. Diese Vorgehensweise gewährleistet eine dynamische und aktuelle Datenverarbeitung, die für die Funktionalität des Chatbots im DieLigen-Projekt entscheidend ist.

Ausblick

Im weiteren Verlauf dieser wissenschaftlichen Arbeit wird das Hauptaugenmerk auf die Implementierung und das Testen eines funktionsfähigen Prototyps des Chatbots im Rahmen des DieLigen-Projekts gelegt. Der Prototyp wird dabei mit einer Reihe von vorgefertigten Testfragen unterschiedlicher Komplexitätsgrade konfrontiert. Ziel dieser Tests ist es, verschiedene Aspekte des Chatbots zu evaluieren, darunter die Qualität der Antworten, die damit verbundenen Kosten sowie die Performance. Daraus wird das Fazit für diese wissenschaftliche Arbeit gezogen und die Forschungsfragen beantwortet.

Literatur und Abbildungen

[1] Lee Boonstra. *The Definitive Guide to Conversational AI with Dialogflow and Google Cloud*. Apress L. P, 2021.

[2] Beate Bruns. *Praxisleitfaden Chatbots - Conversation Design Für eine Bessere User Experience*. Springer Vieweg. in Springer Fachmedien Wiesbaden GmbH, 1 edition, 2023.

[3] Rüdiger Buchkremer, Thomas Heupel, and Oliver Koch. *Künstliche Intelligenz in Wirtschaft & Gesellschaft - Auswirkungen, Herausforderungen & Handlungsempfehlungen*. Springer Gabler, 2020.

[4] Peter Buxmann and Holger Schmidt. *Künstliche Intelligenz - Mit Algorithmen zum wirtschaftlichen Erfolg*. Springer Gabler, 2019.

[5] Markus H. Dahm and Nikals Twesten. *Der Artificial Intelligence Act als neuer Maßstab für künstliche Intelligenz*. Springer Fachmedien Wiesbaden Imprint Springer Gabler, 2023.

[6] M. Harwardt and M. Koehler. *Künstliche Intelligenz entlang der Customer Journey. Einsatzpotenziale von KI im E-Commerce*. Springer Fachmedien Wiesbaden, 1 edition, 2023.

[7] Toni Stucki, Sara D'Onofrio, and Edy Portmann. *Chatbots gestalten mit Praxisbeispielen der Schweizerischen Post*. Springer Vieweg, 2020.

Optimierung eines Machine Learning Modells zur Vorhersage der Produktqualität

Timo Schlude **Karin Melzer**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma MinebeaMitsumi Technology Center Europe GmbH, Villingen-Schwenningen

Ausgangslage und Zielsetzung

Bei der Prüfung und Überwachung gefertigter Bauteile, fallen große Datenmengen mit einer Vielzahl an Eingangswerten an. Die Untersuchung einzelner Phänomene wie der Dichtigkeit eines Bauteils kann bereits einen Berechnungsaufwand verursachen, der ohne maschinelle Methoden nicht mehr bewältigbar ist. In dieser Arbeit wird anhand der Daten aus einem Produktionsprozess mit anschließender Dichtigkeitsprüfung beschrieben, inwieweit maschinelles Lernen sich zur Vorhersage eines binären Klassifizierungsproblems in Gut- und Schlechtteile eignet. Zudem wird besprochen inwieweit diese Modelle sich für Vorhersagen und Risikoabschätzung herangezogen werden können.
In einer Vorarbeit wurden bereits verschiedene Modelle maschinellen Lernens auf den Untersuchungsgegenstand, einen Datensatz über verschiedene Key Parameter Inputs (KPIs), erstellt und getestet. Einerseits sollte mithilfe probabilistischer Methoden die Frage geklärt werden, ob spezifische KPIs die bei der Klassifizierung eines Bauteils in Gut- oder Schlechtteil einen größeren Einfluss haben als andere KPIs, andererseits sollte geklärt werden, inwieweit sich das Training der Modelle verkürzen ließe. Dabei wurden bereits zwei mögliche Herangehensweisen an das Problem identifiziert. Entweder wird das Modell auf zurückliegende Datenpunkte zur Beurteilung des aktuell vorliegenden Bauteils ‚just-in-time' trainiert, oder, das Modell ist derart aufnahmefähig, dass es eine Veränderung verschiedener KPI dynamisch mit in die Beurteilung des vor-liegenden Bauteils aufnehmen kann. Im ersten Fall dürfte die Berechnungszeit wenige Minuten nicht überschreiten, im zweiten Fall dürfte der Berechnungsaufwand auch mehrere Stunden in Anspruch nehmen. Um dabei nicht auf einzelne Bibliotheken und Packages angewiesen zu sein, wird ebenfalls eine Evaluation vorhandener Softwarepakete angestrebt deren Vor- und Nachteile besprochen werden.

Grundlegende statistische Methoden

Unser Modell muss auf die uns vorliegenden Daten abgestimmt werden. Um dabei unsere Eingangswerte später variabel anpassen zu können, benötigen wir eine Methode, die es uns ermöglicht eine Wahrscheinlichkeitsverteilung über eine reellwertige Funktion zu bestimmen. Diese Methode findet man in der parameterfreien Statistik, der Kerndichteschätzung. Sie ermöglicht es mithilfe großer Datenmengen, die zugrundeliegende Verteilung der jeweiligen KPIs anzunähern. [3]

$$\hat{f}_X(x_0) = \frac{1}{N\lambda} \sum_{i=1}^{N} K_\lambda(x_0, x_i)$$

Die auch als Parzen-Fenster-Methode bezeichnete Schätzfunktion kann mithilfe einer Kernfunktion $K_\lambda(x_0, x_i)$ den Anteil eines jeden Datenpunktes in Relation zur Gesamtmasse aller Datenpunkte berechnen. Dabei ist λ die Breite der Nachbarschaft um den Kernpunkt x_0 K_λ ist dabei, sofern man radiale Basisfunktionen als Kern verwendet, ein metrisches Maß. Damit erhält man eine Schätzung, die Ausreißern gegenüber robust bleibt, da diese weniger ins Gewicht fallen. Im eindimensionalen Fall können also zugrundeliegende Verteilungen mithilfe der Kerndichteschätzung geschätzt werden. Wenn wir uns nun einem zweidimensionalen Problem annähern, können wir ebenfalls die Kerndichteschätzung verwenden. Dabei kann über die Randverteilungen untersucht werden, inwieweit Zusammenhänge zwischen den einzelnen untersuchten KPIs vorhanden sind.

Machine Learning Methoden

Führen wir die Kerndichteschätzung für jeden unserer Eingabeparameter einzeln aus, können wir gegebene Einflüsse der einzelnen KPIs untereinander nicht abbilden. In Verbindung mit Zielvariablen lassen sich diese jedoch auf einer höheren Modellebene darstellen.

Nehmen wir für ein Klassifizerungsproblem, an das in unserem Fall, eine binäre Klassifikation in Schlecht- und Gutteil vorgenommen werden kann, dann eliminieren sich die Terme, deren KPI zu einer Klassifizierung als Schlechtteil führen, im Sinne eines Polynoms das nach 0 aufgelöst wird. Damit können alle als Störvariablen identifizierten KPIs aus dem Modell ausgeschlossen werden. Der Auszug aus den beiden verbundenen linearen Gleichungssystemen verdeutlicht dies. Y_1 und Y_2 sind dabei die Klassifikation in 0 und 1 die Parameter β werden im Fall einer Lösung auf 0 aus der Kombination der Gleichungssysteme eliminiert.

$$
\begin{aligned}
Y_1 &= \beta_0 + \beta_{11}x + \epsilon \\
Y_1 &= \beta_0 + \beta_{21}x + \epsilon \\
&\vdots \\
Y_2 &= \beta_0 + \beta_{11}x + \epsilon \\
Y_2 &= \beta_0 + \beta_{21}x + \epsilon \\
&\vdots
\end{aligned}
$$

Um unsere einzelnen genäherten Verteilungen logisch miteinander zu verknüpfen, bedienen wir uns aus dem Bayesschen Paradigma der Statistik. Ein bayes'sches Netz ist ein gerichteter Graph, der unter Berücksichtigung der Likelihood jedes Ereignisses dasjenige mit der größten Wahrscheinlichkeit wählt. Größte Wahrscheinlichkeit bedeutet in diesem Fall, das wahrscheinlichste zukünftige Ereignis basierend auf den vorliegenden Daten.

$$
\begin{aligned}
Y_1 &= \theta_1 f_1(X_1) + \theta_2 f_2(X_2) + \ldots + \theta_n f_n(X_n) + \epsilon \\
Y_2 &= \theta_1 f_1(X_1) + \theta_2 f_2(X_2) + \ldots + \theta_n f_n(X_n) + \epsilon
\end{aligned}
$$

Dabei sind $f(X_i)$ unsere jeweiligen Dichtefunktionen einer Zufallsvariablen X_i und θ_i die Hyperparameter unserer Dichtefunktionen. [3]
Was wir dann betrachten, ist nicht mehr eine relative Gewichtung der einzelnen Datenpunkte als Beitrag zur Gesamtmasse sondern eine Hyperparametrisierung der jeweiligen Likelihoodfunktionen unserer Key Parameter Input Values (KPIVs) als relativer Anteil zur Gesamtmasse aller KPIVs.

Dieses Verhalten lässt sich in Abbildung 1 beobachten, hier werden nur Gutteile fälschlicherweise als Schlechtteile klassifiziert. Wir können diejenigen Linearkombinationen ausschließen die zu einer binär klassifizierten Null führen. Die fälschlicherweise als Schlechtteil klassifizierten Gutteile kommen durch Aufnahme von Störtermen zwecks fehlender Regularisierung zustande. Während hier die Maßstäbe aller unserer Zufallsvariablen skaliert sind, fehlt die mathematische Angleichung der zweiten identitären Variable: der Zeitpunkte. Abbildung 1 zeigt das Klassifikationsergebnis bei fehlender Regularisierung. Die einzelnen Wahrscheinlichkeitsmaße jeder Zufallsvariable sind einheitlich skaliert, jedoch fehlt für die Aussagefähigkeit über die zweite identitäre Einheit die Zeit ein einheitliches Maß.

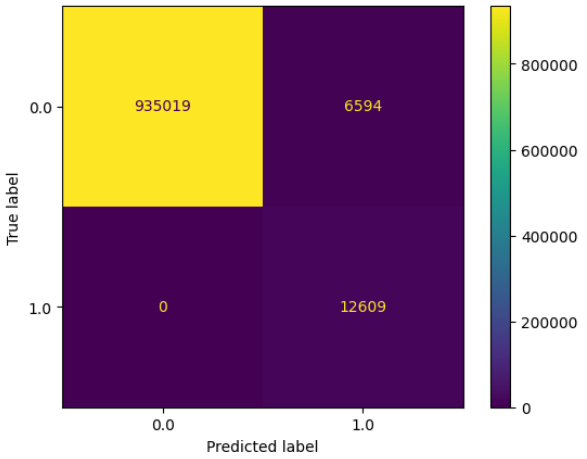


Abb. 1: Unregularisierte Bayesklassifikation [1]

Um die Stufen der Modellebenen anschaulicher zu machen, betrachtet man die KPIVs als Einträge in einem Datenfeld. Im ersten Schritt werden die einzelnen Zufallsvariablen X_1 bis X_n, die unsere KPIs repräsentieren, über Kernregression zur darunterliegenden Dichtefunktion geschätzt. Im zweiten Schritt werden Lage- und Skalenparameter dieser Dichtefunktionen mit den Zielwerten verknüpft und nach ihrem Beitrag zum Zielwert parametrisiert. Dies geht aus Abb. 2 hervor.

	f_{μ_1,σ_1}	f_{μ_2,σ_2}	f_{μ_3,σ_3}	f_{μ_4,σ_4}	$\dots$	f_{μ_n,σ_n}		
	X_1	X_2	X_3	X_4	$\dots$	X_n	C_1	C_2
	x_{11}	x_{12}	x_{13}	x_{14}	$\dots$	x_{1n}		
	x_{21}	x_{22}	x_{23}	x_{24}	$\dots$	x_{2n}		
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$		
	x_{m1}	x_{m2}	x_{m3}	x_{m4}	$\dots$	x_{mn}		
$L(x_{i1}\|\theta)$	μ_1,σ_1	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	y_{11}	y_{12}
$L(x_{i2}\|\theta)$	$\dots$	μ_2,σ_2	$\dots$	$\dots$	$\dots$	$\dots$	y_{21}	y_{22}
$L(x_{i3}\|\theta)$	$\dots$	$\dots$	μ_3,σ_3	$\dots$	$\dots$	$\dots$	y_{31}	y_{32}
$L(x_{i4}\|\theta)$	$\dots$	$\dots$	$\dots$	μ_4,σ_4	$\dots$	$\dots$	y_{41}	y_{42}
	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
$L(x_{in}\|\theta)$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	μ_n,σ_n	y_{i1}	y_{i2}

Abb. 2: Modellebenen [1]

Die Likelihoodfunktionen werden mit Hyperparametern θ_i versehen, sodass ihr Anteil an der Bestimmung zu einer Klasse (Gut- oder Schlechtteil) bewertet werden kann. In Abbildung **??** kann man die Modellebenen differenziert betrachten. Im Kern stehen die einzelnen Zufallsvariablen mit ihren Ausprägungen. Ihre geschätzten Dichtefunktionen mit Lage- und Skalenparametern sind in der nächsten Modellebene die Bedingung, unter der die Likelihoodfunktionen nach einem Optimum ausgewertet werden. Da wir die Likelihoodfunktionen aus Gründen des Curse of Dimensionality nicht über alle KPIs auf einmal berechnen können, betrachten wir sie als unabhängig voneinander und verbinden sie über den Satz von Bayes mit den Zielklassen.
Um sich dem Thema im Umfeld der Datenanalyse anzunähern, wurden im Framework zum maschinellen Lernen 'Scikit-Learn' zunächst Versuche über die Implementierung der Kerndichteschätzung unternommen. Die Berechnung der Kerndichteschätzung ist dort unter anderem mithilfe von Nearest-Neighbor-Algorithmen implementiert. Die starke Abstraktion im Scikit-Learn bietet einen niedrigschwelligen Zugang zum Thema, kann jedoch für das genaue Verständnis hinderlich sein. Insbesondere in der Auseinandersetzung mit DeepLearning Frameworks kann ein weitergreichendes Verständnis für schließende Methoden gefunden werden. Die Auseinandersetzung mit Frameworks im Umfeld von Torch, TensorFlow und PyMC3 bieten eine Möglichkeit die Implementierung neuronaler Netze zu verstehen. Dabei bieten sie in funktional oder objektorientiert gehaltenem Code die Möglichkeit zur Implementierung, allerdings teilweise unter Verlust von Genauigkeit. Aus diesem Grund wurde auch eine Umsetzung in C Code angestrebt.

Ausblick

Für hochdimensionale Räume erhalten wir mit dem 'Curse of Dimensionality' Probleme der Berechnung über Kerndichteschätzung sofern es sich dabei, um mehr als eine Identitäre Abbildung (abgenommener Zeitpunkt und Bauteil) handelt. Die Berechnungszeit unter Berücksichtigung verschiedener Wahrscheinlichkeitsräume (Zeit und Bauteil), besitzt eine exponentielle Zeitkomplexität, da sie sich über das Produkt berechnen. [2] Die Kehrseite des 'Curse of Dimensionality' ist, dass sofern sich Dimensionsanzahl und Datenmenge proportional vergrößern die Genauigkeit unter Speichervorhalt und Rechenzeit nicht ändert. Da eine Dimension, durch die in ihr liegenden Daten definiert wird, ist das Ungleichgewicht zwischen den Dimensionen für die Berechnungszeit in der bayes'schen Statistik ausschlaggebend. Benötige ich einhundert Datenpunkte, um eine Gesetzmäßigkeit mit 99% Wahrscheinlichkeit zu postulieren, spielt der Vorhalt von weiteren Datenpunkten an der Aussagefähigkeit von 99%er Wahrscheinlichkeit keine weitere Bedeutung. Die Genauigkeit einer bayesschen Methode liegt damit in der Proportionalität zwischen Daten und Dimensionen, da sie aus dem Begriff der Log-Odds der Berechnung im steten Zusammenhang aus Wahrscheinlichkeit und Gegenwahrscheinlichkeit erfolgt. Donoho spricht in diesem Zusammenhang vom 'Blessing of Dimensionality' [2].
Wie bei jeder technischen Umsetzung muss selten das Rad neu erfunden werden. Dazu kann mithilfe etwas Softwarearchäologie ein Weg zur Lösung in bereits bestehenden Ansätzen gesucht werden. Sieht man in den alten Paketen der Open Source Community nach, entdeckt man im wissenschaftlichen Umfeld die GNU Scientific Library (GSL), welche die Grundlegende Elemente der Statistik bereits implementiert hat. Wichtig sind hierbei Mittelwert, Standardabweichung und Varianz, die für probabilistische Annahmen später die Grundlage bieten. Form- und Lageparameter werden für die Likelihoodfunktion jedes KPI benötigt. Gräbt man dort weiter, stößt man ebenso auch auf Berechnungsweisen innerhalb der Linearen Algebra (BLAS im LAPACK), die zur Lösung komplexer Linearer Gleichungssysteme hilfreich sein können.

Literatur und Abbildungen

[1] Eigene Darstellung.

[2] David Donoho. High-Dimensional Data Analysis: The Curses and Blessings of Dimensionality. https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.329.3392, 08 2000.

[3] Trevor Hastie et al. *The Elements of Statistical Learning*. Springer, 2 edition, 2008.

Konzeption und Implementierung einer Full-Stack-Anwendung zur Planung von Prüfungsaufsichten

Celine Schuster **Mirko Sonntag**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt an der Fakultät Informatik und Informationstechnik, Esslingen

Einleitung

Prüfungen sind an jeder Hochschule ein wichtiger Bestandteil eines jeden Semesters und benötigen einiges an Planung. Jede schriftliche Prüfung benötigt eine Prüfungsaufsicht und eine Reserveaufsicht, welche zugeordnet werden müssen. Diese Zuordnungen müssen anschließend nachgebessert werden, sollte eine Aufsicht keine Zeit haben. Bisher erfolgte die Prüfungsplanung der Hochschule Esslingen mithilfe von Exceldateien und viel unübersichtlichem Emailverkehr, sollten zwei Hochschulmitarbeitende ihre Aufsichten miteinander tauschen wollen. Dieser Prozess ist sowohl umständlich als auch zeitraubend. Daher ist es das Ziel dieser Bachelorarbeit, eine Anwendung zur Planung von Prüfungsaufsichten zu entwickeln, und so für eine bessere und übersichtlichere Koordination der Prüfungsaufsichten der Fakultät IT an der Hochschule Esslingen zu sorgen.

Anforderungen

Zu Beginn wurden verschiedene Anforderungen erarbeitet und zusammengetragen, welche im Laufe der Bachelorarbeit nach und nach umgesetzt werden. Zu diesen Anforderungen gehören:

- Import sowie Export von Prüfungsterminen im Excelformat
- Algorithmus zur Zuordnung von Prüfungsaufsichten und Reserveaufsichten zu Prüfungen unter Berücksichtigung vorher festgelegter Bedingungen
- Hinzufügen, bearbeiten und löschen von Abwesenheiten
- Tausch von sowohl Prüfungsaufsichten als auch Reserveaufsichten mit den Aufsichten anderer Nutzer
- Sperrung des Hinzufügens von Abwesenheiten sowie von Tauschanfragen mithilfe von Deadlines

Abbildung 1 zeigt die vorgesehene Nutzungsweise der Anwendung.

Abb. 1: Ablaufdiagramm [1]

Zu Beginn erhält die Fakultät eine Exceldatei mit den Prüfungsterminen vom Prüfungsamt. Diese Datei soll dann in der Anwendung hochgeladen werden. Anschließend können die einzelnen Nutzer je nach Bedarf ihre Abwesenheiten während des Prüfungszeitraums hinzufügen. Nachdem eine vorher festgelegte Deadline erreicht ist, wird das Hinzufügen von Abwesenheiten gesperrt. Dann soll der Algorithmus ausgeführt werden und jeder Prüfung unter Berücksichtigung der vorher festgelegten Abwesenheiten eine Prüfungsaufsicht und eine Reserveaufsicht zuordnen. Anschließend ist es den Nutzern möglich, ihre Aufsichten mit den Aufsichten anderer Nutzer zu tauschen. Nach dem Erreichen einer weiteren vorher festgelegten Deadline wird auch diese Funktion gesperrt und die endgültige Liste mit allen Prüfungsaufsichten kann als Exceldatei

exportiert werden. Diese Exceldatei wird dann zurück ans Prüfungsamt geschickt.

Architektur

Die Umsetzung des erarbeiteten Konzepts erfolgt in Form einer Full-Stack-Anwendung bestehend aus einem mit Spring Boot implementierten Backend, welches Zugriff auf eine MySQL Datenbank hat, sowie einem in Angular implementierten Frontend. Einer der vielen Vorteile von Spring Boot ist die Autokonfiguration von Spring und Bibliotheken von Drittanbietern und die daraus resultierende Vermeidung von Boilerplate-Code und Konfigurationsfehlern. [2] Für die Nutzung von Angular sprechen die gut integrierten Bibliotheken, die eine Vielzahl von Funktionen abdecken, einschließlich Routing, Forms Management und Client-Server-Kommunikation sowie eine Reihe von Entwicklertools, die beim Entwickeln, Testen und Verbessern des eigenen Codes unterstützen. [3]

Umsetzung

Das Frontend der Anwendung besteht aus verschiedenen Seiten, deren Zugänglichkeit durch die Rolle des Nutzers bestimmt wird. Die Seiten zum Import der Prüfungstermine, zum Ausführen des Algorithmus sowie für die Userverwaltung und die Einstellungen sind beispielsweise nur mit der Rolle des Admins zugänglich. Die auch für einen User zugänglichen Seiten sind die Startseite, die Übersichtsseite mit der Liste der Prüfungsaufsichten sowie die Abwesenheitsseite und der Basar. Auf dem Basar wird der Tausch der einzelnen Aufsichten durchgeführt. Der Name wurde in Anlehnung an das Feilschen und das Verhandeln auf einen echten Basar gewählt.

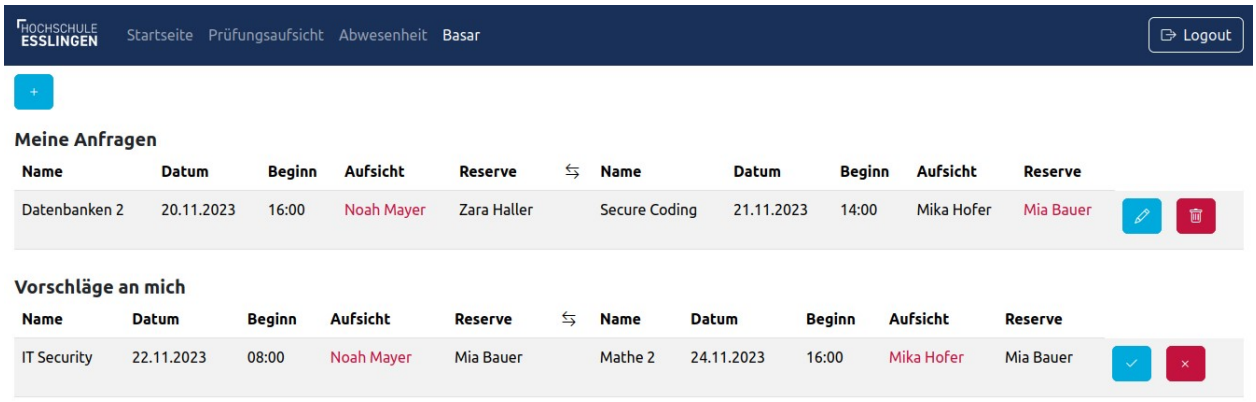


Abb. 2: Basar aus Sicht eines Users [1]

Abbildung 2 zeigt den Basar aus der Sicht eines Users. Der Basar teilt sich in zwei Tabellen: Die obere Tabelle enthält die Tauschanfragen, die der aktuelle Nutzer an andere Nutzer gestellt hat. Die jeweils zu tauschenden Parteien sind dabei rot markiert. Der Nutzer kann diese Anfragen sowohl bearbeiten als auch löschen. Die untere Tabelle enthält die Tauschvorschläge, die andere Nutzer an den aktuellen Nutzer gestellt haben. Der Nutzer kann diesen Vorschlägen entweder zustimmen oder sie ablehnen.

Abb. 3: Einstellungen [1]

Abbildung 3 zeigt die Einstellungsseite, die zur Semesterverwaltung genutzt wird. Das auf aktiv gesetzte Semester wird zusammen mit den Deadlines für die Abwesenheit und den Basar auf der Startseite angezeigt. Zudem werden die Deadlines verwendet, um nach deren Ablauf den Zugang zu den entsprechenden Seiten beziehungsweise deren Funktion zu sperren.

Ausblick

Im weiteren Verlauf sollen die noch fehlenden Anforderungen implementiert werden. Zudem sollen Tests für den Algorithmus ergänzt werden. Ein möglicher weiterer Schritt wäre das Hinzufügen einer Filtermöglichkeit auf der Übersichtsseite mit der Liste der Prüfungsaufsichten, um das Finden der eigenen Aufsichten zu erleichtern.

Literatur und Abbildungen

[1] Eigene Darstellung.

[2] Devlin Basilan Duldulao and Seiji Ralph Villafranca. *Spring Boot and Angular: Hands-on full stack web development with Java, Spring, and Angular*. Packt Publishing, 1 edition, 2022.

[3] Angular. io. What is Angular? https://angular.io/guide/what-is-angular, 2023.

Fuzzing as a Security Test for Robotic Applications in Industry 4.0 with the Assistance of Large Language Models

Fabio Schwarz **Dominik Schoop**

Department of Computer Science and Engineering, Esslingen University

Work carried out at Festo SE & Co. KG, Esslingen

Introduction

Software testing is crucial in ensuring the proper functioning and security of software systems. Traditional testing methods often struggle with the complexity of modern applications. Fuzzing, a dynamic testing approach, introduces random or mutated inputs to uncover unexpected behavior and vulnerabilities. As software systems become more intricate, there is a need for advanced testing techniques. This thesis explores the integration of fuzzing with large language models, such as GPT-3, to enhance software testing and security.

Large language models exhibit unparalleled natural language understanding and generation capabilities. They excel in various language-related tasks and can simulate real-world language inputs. The unique strength of large language models lies in their ability to generate intelligent and context-aware test inputs. This thesis aims to harness this capability to enhance fuzzing by improving the effectiveness and efficiency of the testing process.

The research addresses key questions:

1. How can large language models be integrated into the fuzzing process to generate more effective test inputs and uncover complex vulnerabilities?
2. What practical applications arise from the utilization of these novel fuzzing techniques incorporating large language models?
3. What limitations and challenges arise from integrating large language models in the fuzzing process, and how can these challenges be addressed?

By unlocking the potential of large language models in fuzzing, this research aims to fortify the software development process, paving the way for a more secure digital landscape.

Fuzzing

Fuzzing is a dynamic testing method designed to reveal vulnerabilities and weaknesses in software systems. It involves systematically injecting randomized or semi-randomized input data into a target program to expose unexpected behaviors, crashes and security flaws. This section delves into the historical development, various types, tools and frameworks while also looking at the strengths and limitations of fuzzing.

Fuzzers, categorized as generation-based or mutation-based, operate within structured inputs like files, events, or network protocols. Smart fuzzers use techniques such as structure-aware fuzzing, grammar-based fuzzing, and evolutionary fuzzing. Structure-aware fuzzing generates inputs conforming to a provided structure, while grammar-based fuzzing uses formal grammars. Evolutionary fuzzing combines evolutionary computation principles with fuzzing to explore challenging branches within a target's code.

The strength of fuzzing lies in establishing consistent quality standards, enhancing software stability and security, and reducing manual testing efforts. Its automated nature, simplicity in test design, and effectiveness in uncovering elusive bugs contribute to its widespread adoption. However, fuzzing has limitations, including its inability to provide insights into bug root causes, potential challenges in handling software intricacies, and concerns about misuse by malicious developers.

In conclusion, fuzzing is a valuable testing method with a rich history and versatile applications. It has evolved to address the complexities of modern software systems and has become an integral part of software testing and security practices.

Large Language Models

Large language models (LLMs) are massive neural networks that process text data. Key components include word vectors, representing words numerically, and transformers, neural network layers that enhance comprehension and predict the next word in a sequence. LLMs eliminate explicit labeling in favor of unsupervised learning, predicting the next word through

iterative adjustments of weight parameters in a forward-backward pass.
Word vectors, akin to geographical coordinates, enable LLMs to represent words in multidimensional spaces, capturing similarities between words with similar meanings. Google's word2vec project [6] in 2013 pioneered this approach, revealing relationships between words but also inheriting biases from training data. LLMs address homonymy and polysemy, using vectors to represent nuanced meanings in various contexts.
Transformers, the core of LLMs, operate through attention and feed-forward layers. Attention heads match and exchange information between words, while feed-forward layers, with billions of parameters, perform pattern matching to predict the next word. Research on GPT-2 reveals how attention heads collaborate, guiding the model's predictions [9]. The feed-forward process evolves from recognizing specific words to complex relationships, providing contextual knowledge for predictions.
Training LLMs involves vast datasets. Models predict the next word in diverse passages, adjusting weight parameters iteratively to improve accuracy. Scale is crucial; GPT-3, for instance, ingested 500 billion words, showcasing the power-law relationship where increased scale enhances language understanding.

Fuzzing with the help of LLMs

In recent years, the convergence of large language models and fuzzing has emerged as a focal point in software security research. LLMs, equipped with natural language understanding, exhibit potential for automating and enhancing the fuzzing process, a pivotal area for identifying software vulnerabilities. This section delves into the intersection of LLMs and fuzzing, presenting a systematic review of research endeavors at this nexus.
Traditional fuzzing involves a multi-component process, including a monitor, mutator, test case generator, and test case filter. While automation covers several components, manual intervention persists in areas like seed file generation, harnessing the target program, and crash analysis.
Machine learning interventions in fuzzing primarily address seed file generation, test case filtering, mutation operator selection, test case generation, and exploitability analysis. Noteworthy advancements include mutation-based approaches using AFL enhancements [8], neural network models for vulnerability identification [5], and generation-based methods like Learn&Fuzz [4] and DeepFuzz [7].
A comprehensive review of fuzzing and machine learning [10] reveals key insights: fuzzing aligns well with machine learning, meeting dataset and input-to-vector conversion requirements. Machine learning is integrated across fuzzing phases, emphasizing advancements in seed file generation and test case filtering. Deep learning algorithms, especially LSTM and seq2seq models, dominate due to their robustness. Machine learning-driven fuzzers exhibit improved code coverage, execution of unique code paths, and identification of real-world bugs.
A particularly interesting approach [1] employs LLMs to enhance format specification clarity and seed file generation, showcasing improvements in line and branch coverage.
Deng et al.'s tandem studies focus on leveraging LLMs for fuzzing deep learning libraries (PyTorch and TensorFlow). TitanFuzz [2] pioneers LLMs for both generation-based and mutation-based fuzzing, achieving superior bug detection and code coverage. FuzzGPT [3] extends this work, demonstrating LLMs' effectiveness in generating code by priming them with bug-triggering examples.
In a departure from input-focused fuzzing, [11] propose LLMs for generating the code that does the fuzzing. They argue for the lightweight and general nature of LLMs in contrast to analysis-based methods, emphasizing the importance of diverse driver code for effective fuzzing.
The synergy between LLMs and fuzzing marks a transformative era in software security. Research findings underscore the efficacy of LLMs in various fuzzing phases, from seed file generation to exploitability analysis. As the field progresses, harnessing the power of LLMs in automated code generation for testing continues to demonstrate promise, heralding advancements in software security practices.

Goals of the Thesis

The thesis focuses on advancing fuzzing techniques for Industry 4.0 appliances through the utilization of large language models. The primary objective is to create and validate a method for this purpose. The research involves the development of a framework dedicated to streamlining the fuzzing process. This framework relies on the exported documentation of an API endpoint, serving as a crucial guide throughout the operation. The framework's core functionality involves the generation of inputs tailored for a fuzzing tool. Alternatively, it can generate specialized code designed explicitly for the fuzzing process. Notably, the generation of this code is entrusted to the capabilities of a large language model. With the help of this framework industry 4.0 appliances will be tested for their security.
The initial step in the framework involves parsing documentation to extract method names and details on method invocation. If available, usage information is also extracted. Subsequently, these details are fed

into the LLM, instructing it to generate either code or fuzzing inputs. The produced inputs are then utilized in a fuzzer or generated code is executed. The objective is to induce program crashes, potentially uncovering vulnerabilities.

References and figures

[1] Joshua Ackerman and George Cybenko. Large Language Models for Fuzzing Parsers (Registered Report). *ACM*, 2023.

[2] Yinlin Deng, Chunqiu Steven Xia, Haoran Peng, Chenyuan Yang, and Lingming Zhang. Large Language Models Are Zero-Shot Fuzzers: Fuzzing Deep-Learning Libraries via Large Language Models. *arXiv*, 2023.

[3] Yinlin Deng, Chunqiu Steven Xia, Chenyuan Yang, Shizhuo Dylan Zhang, Shujing Yang, and Lingming Zhang. Large Language Models are Edge-Case Fuzzers: Testing Deep Learning Libraries via FuzzGPT. *arXiv*, 2023.

[4] Patrice Godefroid, Hila Peleg, and Rishabh Singh. Learn&Fuzz: Machine Learning for Input Fuzzing. *arXiv*, 2017.

[5] Yuwei Li, Shouling Ji, Chenyang Lv, Yuan Chen, Jianhai Chen, Qinchen Gu, and Chunming Wu. V-Fuzz: Vulnerability-Oriented Evolutionary Fuzzing. *arXiv*, 2019.

[6] T Mikolov, M Karafiát, L Burget, J ˇCernocký, and S Khudanpur. *Recurrent neural network based language model*. ISCA, 2010.

[7] Morteza Zakeri Nasrabadi, Saeed Parsa, and Akram Kalaee. Format-aware Learn\&Fuzz: Deep Test Data Generation for Efficient Fuzzing. *arXiv*, 2018.

[8] Mohit Rajpal, William Blum, and Rishabh Singh. Not all bytes are equal: Neural byte sieve for fuzzing. *arXiv*, 2017.

[9] K Wang, A Variengien, A Conmy, B Shlegeris, and J Steinhardt. Interpretability in the Wild: a Circuit for Indirect Object Identification in GPT-2 small. *arXiv*, 2022.

[10] Yan Wang, Peng Jia, Luping Liu, Cheng Huang, and Zhonglin Liu. A systematic review of fuzzing based on machine learning techniques. *PLOS ONE*, 2020.

[11] Cen Zhang, Mingqiang Bai, Yaowen Zheng, Yeting Li, Xiaofei Xie, Yuekang Li, Wei Ma, Limin Sun, and Yang Liu. Understanding Large Language Model Based Fuzz Driver Generation. *arXiv*, 2023.

KI in der Elektronikproduktion: Erkennung von Fehlbestückung mithilfe von NXP-Prozessoren mit NPU

Claudio Senatore **Steffen Schober**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma F&S Elektronik Systeme GmbH, Stuttgart

Bei der Fertigung von Embedded Computern kann es bei der Bestückung von THT-Bauteilen (engl. Through Hole Technology) zu Fehlerfällen kommen. THT-Bauteile zeichnen sich durch ihre Aufbau- und Verbindungstechnik mit der Basisplatine aus. Anders als bei SMT-Bauteilen (engl. surface- mounting technology), wo die Komponenten auf der Oberfläche der Platine aufgelötet werden, werden die Kontakte der THT-Bauteile durch die Platine gesteckt und auf der Rückseite verlötet. Die Bestückung der Bauteile wird von Hand an Arbeitsplätzen vorgenommen. Anschließend werden die Bauteile über ein Fließband an einer Anlage eingeführt und maschinell verlötet. In der Vergangenheit kam es hierbei zu verschiedenen Fehlerfällen, die sporadisch auftraten. So wurden bspw. Bauteile beim Bestücken übersehen und entsprechend nicht eingelegt. Da viele der bestückten Bauteile ein symmetrisches Pin-Out haben, können diese zum Teil auch um 180° rotiert eingelegt werden. Dieser Fehlerfall hätte zufolge, dass Konnektoren aufgrund des rotierten Gehäuses des THT-Steckverbinders nicht angeschlossen werden, oder Spannungspotentiale für die Signalübertragung und Spannungsversorgung vertauscht werden könnten. Dies hätte eine Beschädigung am Embedded-Computer und an seiner angeschlossenen Peripherie zur Folge.

Lösungsstategie

Zwischen Handarbeitsplatz und Lötstation soll eine Kamerastation vorgesehen werden, welches die bestückten Boards erfasst und die Interessenbereiche, also alle zu verbauenden THT-Komponenten bewertet. Folgende Eigenschaften sollen dabei für jedes THT-Bauteil erkannt werden.
* Bauteil nicht eingelegt
* Bauteil rotiert eingelegt
* Bauteil ist in Ordnung

Dieses Klassifizierungsproblem soll mittels AI on the Edge gelöst werden. Das Bedeutet, das die Inferenz eines KI-Modells am Rande eines Systems, auf ein Embedded Device, ausgeführt wird.

Zielsetzung

Ziel der Arbeit ist die Erstellung eines Software-Frameworks zur Klassifizierung von Platinen und deren aufgesteckten THT-Bauteilen, welches auf ein i.MX93 Prozessor ausgeführt wird. Dabei soll ein Neuronales Netz zu Einsatz kommen, welches auf einen Embedded Prozessor wie die i.MX-Serie ausgeführt werden kann. Fokus liegt auf die Ausführung der Inferenz auf einer Ethos-U NPU, welche erstmals auf der i.MX9 Serie implementiert wurde. Es soll untersucht werden, welche KI-Modelle sich für den beschriebenen Zweck eignen. Zudem soll untersucht werden, welche Vorteile eine EdgeNPU gegenüber einem Cortex-A oder Cortex-M Prozessor mit sich bringt.

Vorgehensweise

Die zu entwickelnde Software sollte eine Video-Pipeline erzeugt, um ein spezifisches Board im Videostream extrahiert und die Inferenz eines Object-Detection Modell anstoßen zu können. Das Ergebnis der Inferenz soll hierbei wieder in der Video-Pipeline eingebettet werden. Zudem muss eine Entwicklungsumgebung erzeugt werden, mit dem es möglich ist einen Datensatz mit Bildern und Annotationen zu definieren, sowie eine vortrainiertes OD-Modell „finetunen“ zu können. Dass Modell muss anschließend für die Inferenz mit der Ethos-U NPU und dem Ethos-U Software-Stack konvertiert werden können.

THT-Classificator

Der THT-Classificator ist die zentrale Softwarekomponente, die für die Klassifizierung der THT-Bauteile verantwortlich ist. Das Programm, welches für das Projekt entwickelt und in Python geschrieben wurde, erfüllt mehrere Schlüsselfunktionen. Die Aufgaben sind

die Erstellung einer Video-Pipeline, Segmentierung der Baugruppe, Ausrichtung der Baugruppe, Formatierung des Frames zum Input- Tensor, Durchführung der Inferenz, Auswertung des Output-Tensors des OD-Modells sowie die Darstellung der Resultate. Um diese Aufgaben effizient durchzuführen, folgt der THT-Classificator einer Softwarearchitektur, die auf dem Prinzip von Pipes-and-Filters basiert. (siehe [3]) Dieses Prinzip teilt die Verarbeitungsaufgaben in einzelne Schritte auf, wobei jeder Schritt eine Ausgabe erzeugt, die als Eingabe für den nächsten Schritt dient. Die Datenquelle ist die Kamera als Videoinput, und der Datensink ist ein Window-Manager wie Wayland, der das Video auf einem Display ausgibt. Die Pipeline besteht aus weitestgehend unabhängigen Prozessen, die als Producer und Consumer fungieren. Ein charakteristisches Merkmal, das den THT-Classificator von einer konventionellen Pipes-and-Filters-Architektur unterscheidet, ist seine Managing Unit, die hier als eine Art Broker für seine Unterprozesse fungiert und eine Statemachine koordiniert. Diese Managing-Unit ist verantwortlich für die Allokation von Ressourcen wie Speicher, der von den Unterprozessen für die Interprozess-Kommunikation genutzt werden kann. Zudem werden Queues für eingehende und ausgehende Signale der Unterprozesse erstellt, die von der Managing Unit genutzt werden, um Ereignisse zu steuern. Die Architektur des THT-Classificators mit seinen Child-Prozessen wird in der folgenden Abbildung dargestellt.

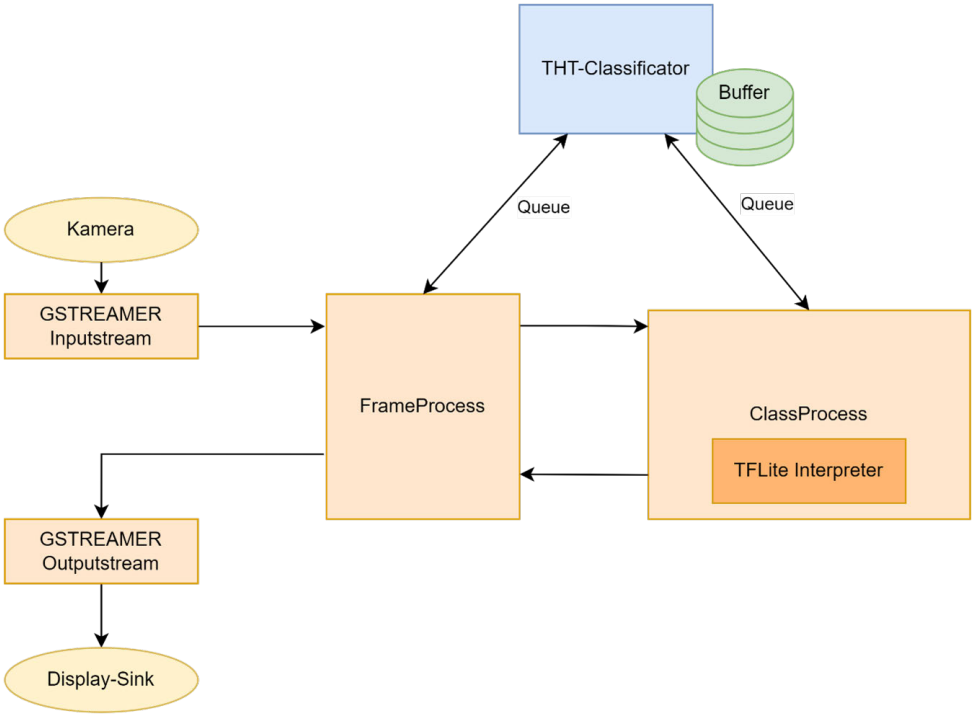


Abb. 1: THT-Classificator Architektur [2]

Ethos-U

Die Ethos-U Serie ist eine von ARM entwickelte NPU, welche für das Asymetrische-Multiprocessing mit Cortex-M CPUs ausgelegt ist. Mit der kommenden i.MX9 Serie werden erstmals NPUs aus der Ethos U Serie verwendet. Der erste NXP eigene Prozessor ist der i.MX93 welche die Ethos-U65 implementiert hat. Die NPU unterstützt quantisierte Neuronale Netze wobei die Weights als 8 bit Integer-Zahlen, und die Activations als 8bit oder 16bit Integer Zahlen abgebildet werden müssen. Die NPU allein unterstütz zwar zahlreiche Operationen aus einem TFLite Modell, benötigt allerdings zusätzliche Peripherien im SoC, sowie einen Software-Stack, um eine Inferenz ausführen zu können. Der Cortex-M Prozessor beispielsweise dient als Controller-Unit für die NPU und führt eine Software aus (Ethos-U Subsystem), die von der Application-Domain Daten über ein Remote Processor Messaging Protokoll (RPMSG) erhalten kann und die Inferenz ausführt. Folgende Abbildung zeigt den Software-Stack für das Asymmetrische Multiprocessing.

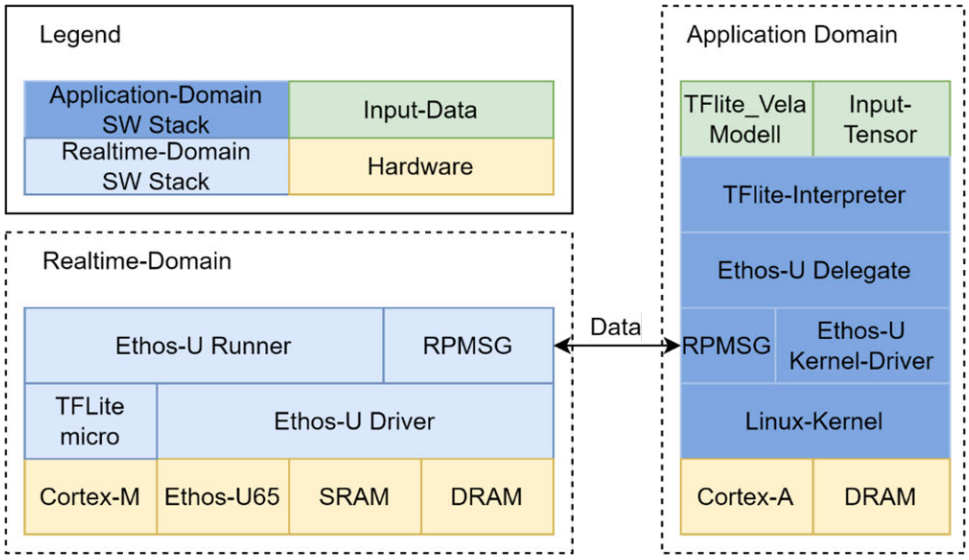


Abb. 2: Ethos U Software Stack [1]

Über die Application-Domain wird mittels TFLite-Interpreter ein TFLite-Modell und ein Ethos-U Delegate geladen. Das TFLite Modell muss für die Inferenz mit einem Ethos-U Vela Compiler konvertiert werden. Dabei werden die konventionellen TFLite Operationen mit Instruktionen ausgetauscht, die das Ethos-U Subsystem auf der Realtime-Domain versteht. Das Ethos-U Delegate erkennt die im Modell hinterlegten Operationen und legt einen Buffer für die Input und Output Tensoren im DRAM an. Der Ethos-U Kernel Driver stellt die Nachrichten der Inferenzaufgabe zusammen und sendet diese über RPMSG an das Ethos-U Subsystem in der Realtime-Domain. Der Ethos-U Runner empfängt die Operationsdaten und prüft, ob eine Operation für die NPU kompatibel ist. Falls dies so ist, wird die Operation mittels dem Ethos-U Driver ausgeführt. Temporäre Daten werden dabei im schnelleren SRAM abgelegt. Falls die Operation nicht mit der NPU kompatibel ist, wird die Instruktion über TFLite micro für MCUs und dem CMSIS-NN Framework auf dem Cortex-M Prozessor ausgeführt. Diese enthalten spezielle Assembler-Instruktionen, die für die Ausführung von Neuronale Netze auf dem Cortex-M Prozessor ausgelegt sind. Die Resultate der Inferenz werden vom Ethos-U Treiber in ein Buffer hinterlegt und über RPMSG an die Applikation Domain zurückgesendet.

Weitere Aussichten

Für die Arbeit wird die Ausführung der Inferenz mehrerer Object-Detection Modelle über das Ethos-U Subsystem untersucht und geprüft, ob sich das Konzept für die Erkennung und Klassifizierung von THT-Bauteilen eignet.

Literatur und Abbildungen

[1] Arm Limited or its affiliates. Arm® Ethos™-U NPU Version 5.0 Application development overview. https://developer.arm.com/documentation/101888/0500/NPU-software-overview/Use-cases, 10 2020.

[2] Eigene Darstellung.

[3] Rainer Grimm. Patterns in der Softwarearchitektur: Das Pipes-and-Filters-Muster. https://www.heise.de/blog/Patterns-in-der-Softwarearchitektur-Das-Pipes-and-Filters-Muster-8312564.html, 04 2023.

Effiziente Verteilung von Vorhersageergebnissen in einem hierarchischen Prognosemodell: Analyse und Bewertung unterschiedlicher Verteilungsverfahren

Alwis Stark **Thomas Rodach**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma Robert Bosch GmbH, Stuttgart-Feuerbach

Einleitung

In der heutigen datengetriebenen Geschäftswelt ist die Vorhersage zukünftiger Trends und Ereignisse von entscheidender Bedeutung für den Erfolg von Unternehmen. Gut fundierte Vorhersagen ermöglichen es Organisationen, proaktiv auf Marktveränderungen zu reagieren, Risiken zu minimieren und strategische Vorteile zu erlangen [5]. Mit der zunehmenden Menge und Komplexität verfügbarer Daten steigt auch die Bedeutung effizienter und genauer Prognosemethoden.

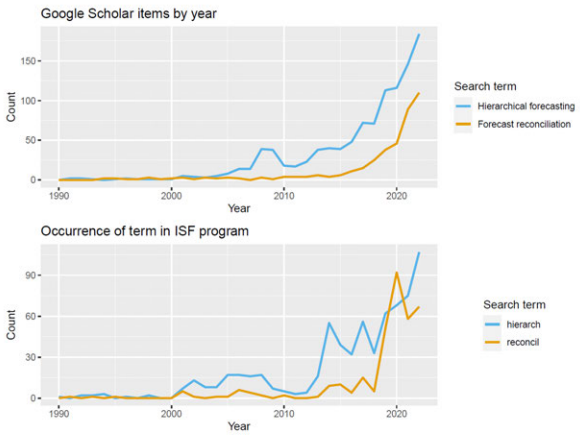


Figure 1: *Search term results in Google Scholar and the book of abstracts for the International Symposium on Forecasting during ISFs (1990-2022)*

Abb. 1: Wachsende Relevanz von hierarchical forecasting und forecast reconciliation [1]

Hierarchische Prognosemodelle für Zeitreihen, die Daten in verschiedenen Ebenen strukturieren, spielen eine zentrale Rolle in der betriebswirtschaftlichen Informationsverarbeitung. Sie ermöglichen eine konsistente Planung und Entscheidungsfindung über verschiedene Ebenen der Unternehmenshierarchie hinweg. Diese Modelle müssen präzise und konsistent sein und zugleich effizient gestaltet werden, um die Kosten für die Erstellung der Prognosen zu minimieren. Hierarchisches Forecasting bezieht sich auf den Prozess Vorhersagen auf allen Ebenen der Hierarchie zu erstellen und die Beziehungen zwischen den verschiedenen Ebenen zu berücksichtigen. Forecast Reconciliation, auch als Vorhersagekonsolidierung bekannt, spielt dabei eine wesentliche Rolle Konsistenz über alle Ebenen herzustellen. Sie hilft, Widersprüche innerhalb der Vorhersagen verschiedener Unternehmensebenen zu vermeiden, die andernfalls zu ineffizienten Entscheidungen und Ressourcenallokationen führen könnten. Die Arbeit untersucht und evaluiert Methoden zur Erzeugung konsistenter Prognosen in einem hierarchischen Prognosemodell, wobei der Fokus auf der effizienten Verteilung von Vorhersageergebnissen von höheren auf die niedrigste Ebene der Datenhierarchie liegt. Es wird untersucht, wie solche Vorhersageergebnisse effizient verteilt werden können, wie weit Kosten eingespart werden können und wie die Qualität der Vorhersagen dabei beeinträchtigt wird. Dazu wird ein Prototyp entwickelt, welcher diesen Ansatz demonstriert.

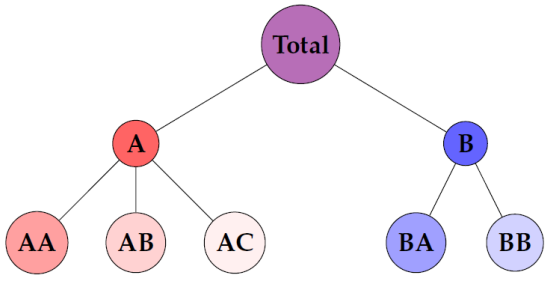


Abb. 2: Beispiel einer Hierarchie von Zeitreihen [6]

S-Matrix

Die S-Matrix, auch Aggregationsmatrix genannt, ist essenziell im Forecast Reconciliation Prozess. Die

S-Matrix spielt eine entscheidende Rolle in der Strukturierung und Analyse von hierarchischen Zeitreihenprognosen [3]. Ausgedrückt wird dies in einer Binärmatrix, diese definiert die Beziehungen zwischen den übergeordneten und den untergeordneten Ebenen. Ein typisches Beispiel dafür sind einzelne Produkte, die sich zu einer Produktgruppe aggregieren lassen. Ebenso lässt sich eine geografische Hierarchie in einzelne Länder aufteilen, beispielweise Europa disaggregiert sich in Deutschland, Frankreich und weitere Länder. Dieser genaue Bauplan ermöglicht es mit Methoden der Forecast Reconciliation eine Konsistenz aller beteiligten Zeitreihen innerhalb der Hierarchie herzustellen. Im Kontext dieser Arbeit wird sie dazu genutzt Vorhersagen von den aggregierten Ebenen auf die unterste Ebene zu verteilen.

$$\begin{bmatrix} y_t \\ y_{\mathrm{A},t} \\ y_{\mathrm{B},t} \\ y_{\mathrm{AA},t} \\ y_{\mathrm{AB},t} \\ y_{\mathrm{AC},t} \\ y_{\mathrm{BA},t} \\ y_{\mathrm{BB},t} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} y_{\mathrm{AA},t} \\ y_{\mathrm{AB},t} \\ y_{\mathrm{AC},t} \\ y_{\mathrm{BA},t} \\ y_{\mathrm{BB},t} \end{bmatrix}$$

Abb. 3: Struktur einer Hierarchie abgebildet in der S-Matrix [4]

Betriebswirtschaftliche Betrachtung

Ein weiterer zentraler Aspekt in der Prognoseerstellung sind die Kosten. Durch die Nutzung von Hyperskalern wie Azure Kubernetes Service (AKS)-Cluster werden Kapazitäten zur Berechnung der zahlreichen Zeitreihen genutzt, dadurch werden Kosten verursacht. Das Ziel der effizienten Verteilung ist es Kosteneinsparpotenziale zu realisieren. Die Kosten für die Erstellung einer einzelnen Zeitreihe ist nahezu identisch, unabhängig von der Ebene, auf der sie erstellt wird. Der angestrebte Ansatz sieht vor, die Berechnungen für die unterste Ebene der Hierarchie nicht mehr über das AKS-Cluster durchzuführen, sondern über ein vereinfachtes Verteilungsverfahren, welches lokal durchgeführt werden kann. Dieses Verteilverfahren nimmt Vorhersagen der aggregierten Ebene und verteilt diese auf die unterste Ebene. Die gesamten untersuchten Datenstrukturen haben eine Aufteilung von etwa 60% aggregierter Zeitreihen auf den oberen Ebenen und etwa 40% auf der untersten Ebene. Durch den gewählten Ansatz können die gesamten 40% der untersten Eben eingespart werden, da sie nicht mehr aufwendig berechnet werden müssen, sondern über die Verteilung der Vorhersage der oberen Ebenen erzeugt werden. Dies eröffnet erhebliche Möglichkeiten zur Kostensenkung, wobei zugleich untersucht wird, wie das vereinfachte Verteilungsverfahren die Prognosequalität beeinträchtigt.

Forecast Reconciliation

Bei der Forecast Reconciliation von hierarchischen Zeitreihen ist es das Ziel kohärente Zeitreihendaten zwischen mehreren unabhängig berechneten Prognosen herzustellen. Die Basisvorhersagen werden durch individuelle Modelle für jede Zeitreihe aller Ebenen unabhängig voneinander generiert. Diese Vorhersagen sind in der Regel das Ergebnis eines Zeitreihenmodells wie beispielweise ARIMA. Um Kohärenz zu erreichen, werden Reconciliation-Methoden auf diese Zeitreihen angewendet. Es gibt verschiedene Ansätze zur Forecast Reconciliation hierarchischer Strukturen, darunter Top-Down und Combined Conditional Coherent (CCC) Methoden. Diese Ansätze werden in der Arbeit als Prototyp implementiert und getestet, um ihre Effektivität und Anwendbarkeit zu bewerten.

Top Down Methode

Dieser Ansatz ist besonders geeignet, wenn Prognosen von aggregierten Ebenen auf die untersten Ebene der Hierarchie verteilt werden sollen. Dazu werden diese Schritte angewandt:

- Berechnung der historischen Anteile: Der Ansatz beginnt mit der Ermittlung der durchschnittlichen historischen Anteile der untersten Ebene an der aggregierten Ebene.
- Disaggregation der Prognosewerte: Die Vorhersagedaten der aggregierten Ebene werden entsprechend der historischen Anteile verrechnet, um die Prognosewerte für jede Zeitreihe der untersteten Ebene der Hierarchie zu erhalten.
- Bestimmung der neuen Werte auf unterster Ebene: Durch die Disaggregation kommt es dazu, dass mehrere Werte pro Element der untersten Ebene entstehen. Diese werden durch Berechnung des Mittelwerts eindeutig gelöst. Hiermit entstehen die Vorhersagen auf der untersten Ebene.
- Reconciliation der Vorhersagen: Mit den eindeutigen Werten auf der untersten Ebene werden kohärente Vorhersagen auf allen Ebenen generiert, indem diese entlang der Aggregationsstruktur der S-Matrix hochgerechnet werden. Dies entspricht einem Bottom-Up Prozess.

CCC Methode

Die Combined Conditional Coherent-Methode (CCC) ist ein ebenfalls ein Top-Down-Ansatz, der sich auf die Abstimmung von Prognosen innerhalb einer zweistufigen Hierarchie konzentriert. Der Schlüssel zur Anwendung der CCC-Methode liegt in der Definition einzelner elementarer Hierarchien aus einer gesamten Hierarchie. Diese elementaren Hierarchien bestehen aus zwei Ebenen:

- Top-Knoten der Hierarchie: Der oberste Knotenpunkt der Hierarchie, also ein Element aus der aggregierten Ebene, wird als Ausgangspunkt für die Vorhersageerstellung gewählt.
- Verteilung auf die unterste Ebene: Die über das AKS-Cluster errechneten Vorhersagen der aggregierten Ebenen werden dann auf die beteiligten Elemente der untersten Ebene der Hierarchie verteilt.

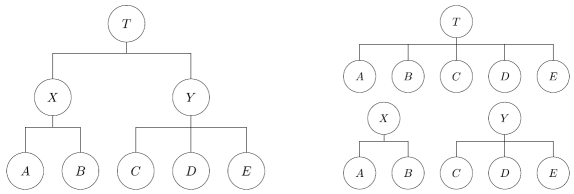


Abb. 4: Gesamte Hierarchie in elementare Hierarchien unterteilt [2]

Diese Methode ähnelt dem Top-Down-Ansatz, wobei der wesentliche Unterschied darin besteht, dass bei der CCC-Methode spezifische, elementare 2-Ebenen-Hierarchien innerhalb der gesamten Hierarchiestruktur definiert und genutzt werden. Die Methode wird noch modifiziert, da Ergebnisse verteilt und nicht berechnet werden sollen. Dazu werden die Werte auf der untersten Ebene durch Null-Werte und historische Durchschnittswerte als Basisvorhersagewerte genommen.

Ergebnisse und Fazit

Die Ergebnisse des Top-Down-Ansatzes zur Forecast Reconciliation hierarchischer Zeitreihen zeigen, dass dieser Ansatz im Vergleich zu Referenzmethoden eine vergleichbare Vorhersagegenauigkeit bietet. Unterschiedliche Ergebnisse gibt es in den betrachteten Datensätzen, die abhängig vom verwendeten Fehlermaß schlechtere Ergebnisse erzielen. Jedoch ist anzumerken, dass die auftretenden Fehler zumeist auf der untersten Ebene auftreten und von der Bedeutung auch weniger gravierend sind. Damit lässt sich zeigen, dass eine effiziente Verteilung bei vergleichbarer Prognosegüte möglich ist und dieser Ansatz sehr vielversprechend ist, da dieser einfach zu implementieren ist und dabei Kosten- und Zeitersparnisse bei der Prognoseerzeugung ermöglicht. Der CCC-Ansatz liefert dagegen unzuverlässige Ergebnisse, abhängig davon wie genau die Basisvorhersagewerte der untersten Ebene sind, beispielsweise liefern historische Durchschnittswerte bessere Ergebnisse als Null-Werte. Hier zeigt sich, dass bessere Ausgangswerte auch zu besseren Ergebnissen führen.

Ausblick

Methoden des Machine Learning im Kontext der hierarchischen Zeitreihenprognose bieten vielversprechende Ansätze. Zukünftige Forschungsarbeiten könnten sich darauf konzentrieren, Machine Learning Methoden zu nutzen, um deren Effektivität in der praktischen Anwendung zu evaluieren. Diese Modelle ermöglichen es, nichtlineare Beziehungen zwischen den verschiedenen Ebenen einer hierarchischen Struktur zu berücksichtigen. Der CCC-Ansatz zeigte bei genaueren Basisvorhersagen auf der untersten Ebene auch genauere Endergebnisse, hier könnten einfache und effiziente Methoden zur Erstellung der Basisvorhersage auf der untersten Ebene zu besseren Ergebnissen führen.

Literatur und Abbildungen

[1] George Athanasopoulos et al. Forecast reconciliation: A review. https://robjhyndman.com/publications/hfreview.html, 04 2023.

[2] Tommaso Di Fonzoa and Daniele Girolimettoa. Forecast combination-based forecast reconciliation: Insights and extensions. *International Journal of Forecasting*, 2022.

[3] Ross Hollyman et al. Understanding forecast reconciliation. *European Journal of Operational Research*, 2021.

[4] Rob J Hyndman and George Athanasopoulos. Forecasting: Principles and Practice. https://otexts.com/fpp3/, 05 2021.

[5] Jürgen Vogel. *Prognose von Zeitreihen*. Springer Fachmedien Wiesbaden, 2015.

[6] Shanika Wickramasuriya. Properties of point forecast reconciliation approaches. https://www.researchgate.net/publication/350311503_Properties_of_point_forecast_reconciliation_approaches, 03 2021.

Entwicklung eines Lineup Tools zur Berechnung von Verstärkerketten

Marc Starzmann **Clemens Klöck**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma Tesat-Spacecom GmbH & Co. KG, Backnang

Einleitung

Das Berechnen von Verstärkerketten ist im Grunde ein Zusammenführen von Bauteilen welche nötigen Parameter besitzen und als Kaskade berechnet werden. Das Problem hierbei ist jedoch, dass es hierfür kein Tool gibt, welches einem Konstrukteur ermöglicht am Start der Entwicklung die Berechnung grundlegend durchzuführen und zu testen. Für diesen Zweck wird meistens eine Excel Tabellen verwendet, welche jedoch viel Aufwand besitzt, da dies meist erst von Grund auf erstellt und geplant werden muss. Die Problematik, die sich hierbei auch bei der Firma Tesat ergibt, ist, dass es von verschiedenen Entwicklern verschiedene Tabellen mit unterschiedlicher Berechnung gibt. Dieses Problem gilt es in dieser Arbeit zu kombinieren und daraus ein Tool zu entwickeln.

Ziel

Ziel der Arbeit ist es ein Tool zu entwickeln, welches innerhalb einer Software gebaut wird, die Tesat eigenständig entwickelt. Die Software soll die Möglichkeit besitzen auf einfachste Weiße ein Lineup zu erstellen, mit Daten zu versehen und es zu berechnen. Bei der Berechnung gibt es dabei die Möglichkeiten noch Variationen anzugeben und zu verwenden. Als Beispiel hierzu wäre es eine Kette von Verstärken auf verschiedenen Temperaturen zu berechnen, wobei sich die Grundwerte der jeweiligen Komponenten verändern. Zusätzlich soll die Software noch die Möglichkeit besitzen anhand von bestimmten Angaben des Entwicklers eine Optimierung durchzuführen. Die Software hat ebenso die Möglichkeit an andere, von Tesat entwickelte, Programme verbunden zu werden. Hierüber kann dann das Lineup Tool mit bereits existierenden Datenstarten und eine schnelle Berechnung durchführen.

Verstärkerketten und deren Berechnung

In der Hochfrequenztechnik unterscheidet man zwischen passiven und aktiven Bauteilen. Die passiven Bauteile sind für die Filterung von Signalen während die Aktiven sich um die Verstärkung von Signalen kümmern. Genauer beschrieben ist ein Aktives Bauteil in der Elektrotechnik ein Element, welche Energie verstärken, steuern oder Erzeugen kann [1]. Bei Aktiven Produkten handelt es sich um „Amplifier“ (Verstärker), Dämpfungsglieder und weiteren. Diese Elemente werden in der Software mit Grundwerten und dazu passenden Variationen angelegt. Die 3 Grundwerte, welche wir hier verwenden, sind Gain, Noise Figure und Icp3. „Gain ist die Vergrößerung einer variablen physikalischen Ausgangsgröße gegenüber einer Eingangsgröße...“ [3]. Es beschreibt bei uns die Verstärkung der Eingangsleistung, also im Grunde eine einfache Addition auf den Eingang, um die Ausgangsleistung eines einzelnen Elements zu bestimmen.

Noise Figure, auf Deutsch Rauschzahl, ist eine Kennzahl, welche in der Nachrichtentechnik das Rauschen bestimmt. Unter Rauschen wird hierbei die Störgröße verstanden, bei einem gewissen Frequenzbereich. Dieses Störverhalten tritt beim Senden, Empfangen aber auch auf dem kompletten Übertragungsweg auf und beeinträchtigt das ursprüngliche Signal [2].

Icp3, auch Third Order Intercept Point genannt, beschreibt Verhalten bei einem Sinussignal am Ausgang Verzerrungen, die neben dem linearen Signal als Nebenprodukte entstehen.

Diese 3 Hauptvariablen spielen bei der Lineup Berechnung die größten Rollen. Sie werden jedoch nicht einmalig eingegeben, sondern noch in verschiedenen Bereichen variiert. Dabei spielen bei uns Temperatur, Prozess und Bias eine wichtige Rolle. Temperatur und Prozess werden für die Berechnung eines ganzen Lineups angegeben während die Bias Variation individuell für jeden Block angelegt und ausgewählt wird. Temperatur spielt darum eine große Rolle, da wir uns bei Tesat in einer Raumfahrtbranche befinden und es zu verschiedenen Temperaturen kommen kann, wobei das Verhalten hierbei beachtet werden muss.

Generell gibt es viel mehr Berechnungsmöglichkeiten

bei einem Lineup wie auch der Energieverbrauch der einzelnen Bauelemente. Diese wird in unserem Fall auch als Variation angegeben welche nachher auf der Ausgangsleistung einzelner Blöcke sowie der ausgewählten Temperatur basieren.

Umsetzung und Probleme

Umgesetzt wurde das Tool mit Eclipse RCP. Anwendungen, welche darauf basieren bedienen sich an dem Eclipse Framework, um eine featurebasierte Anwendung zu kreieren [4]. Die Entwicklung basiert dabei auf Plugins, welches ein einzelnes Feature darstellt. Wie man es der einfachen Eclipse Entwicklungsoberfläche entnehmen kann, werden bei Bedarf für jedes Feature ein „Fenster“ dargestellt, welche mit weiteren aneinander angereiht und sortierten die finale Anwendung ergeben. So lassen sich für das Tool und deren Funktionen eigene Plugins identifizieren und werden so in die bereits existierende Anwendung eingebaut.
Bei der Software von Tesat handelt es sich um ein Blockdiagram. Hier hat ein Entwickler die Möglichkeit in einem Editor elektronische Bauteile, wie auch Verstärker und Dämpfungsglieder, als eine Art Schaltplan zusammen zu klicken und zu verbinden. Diese bereits existierende Software bietet eine perfekte Grundlage für ein Lineup Tool, da wir so die Möglichkeit haben eine Kette von Eingang- bis Ausgangsport zu definieren und durchzurechnen. Hierfür musste trotzdem das Datenmodel und etliche andere kleine Features der bereits bestehenden Anwendung angepasst werden, sodass nötige Kriterien für ein Lineup gegeben sind.
Probleme, die hierbei direkt schon auftreten, ist unter anderem dem Fakt geschuldet, dass die Anwendung im Voraus auch schon von etlichen Studenten erarbeitet wurde. Dementsprechend stecken noch verschiedene Fehler in der Anwendung, die es dann auch zu beheben gilt, um ein angenehm nutzbares Tool am Ende zu erhalten.
Ein weiterer Faktor, der öfters dazu führt, dass Teile der Software umgedacht und geändert werden mussten, sind die Anforderungen verschiedener Endnutzer. Denn es ist ein wichtiger Faktor zu beachten, was sich die Entwickler wünschen und was nötig ist für die Software. Dabei gab es aber auch öfters widersprüchliche Anforderungen, welche dazu führten, dass Konzepte öfters überarbeitet werden mussten.

Ausblick

Schon nach dem Ersten vorzeigbaren Stand des Tools hat man gemerkt, dass es möglich ist, die Effizienz der Berechnung zu steigern. Dabei wird auch klar, dass viele Möglichkeiten existieren, um es zu erweitern. Neben etlichen Berechnungsmöglichkeiten, welche bei der Entwicklung noch dazu kamen, sind auch noch weitere Features denkbar. Ein Ansatz, der da unter anderem angegangen werden kann, wäre es, S Parameter zur Bestimmung der Daten zu verwenden und damit eine Berechnung durchzuführen. Der Anwendung sind dabei kaum Grenzen gesetzt und ermöglicht es einem Entwickler viel theoretische Arbeit abzunehmen. Zusätzlich bietet es einen Standard in der Firma, sodass nicht verschiedene Formeln verwendet werden, sondern einheitlich gearbeitet wird. Ein weiterer Faktor ist natürlich auch die Verknüpfung anderer Programme, welche es dann auch einfacher macht, bereits existierende Ergebnisse zu speichern und wiederzuverwenden.

Literatur und Abbildungen

[1] Christian Eisenhut. Aktive und Passive Bauelemente. https://www.lernort-mint.de/physik/elektronik/gesetzmaessigkeiten-in-der-elektronik/aktive-und-passive-bauelemente/, 11 2023.

[2] OB Optics Buddy et al. Rauschzahl. https://de.wikipedia.org/wiki/Rauschzahl#Definition, 09 2023.

[3] S Saure et al. Verstärkung (Physik). https://de.wikipedia.org/wiki/Verst%C3%A4rkung_(Physik), 09 2023.

[4] Lars Vogella. Eclipse RCP (Rich Client Platform) - Tutorial. https://www.vogella.com/tutorials/EclipseRCP/article.html, 01 2023.

Digitalisierung des Access Risk Catalogs zur Steuerung von Access Management Prozessen aus SAP Systemen

Nik Steinbruegge **Rainer Keller**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma Robert Bosch GmbH, Stuttgart

Access Risk Catalog

Digitale Infrastruktur in Großunternehmen kann bei schlechtem Management schnell komplex und undurchsichtig werden. Diese Infrastruktur muss innerhalb des Unternehmens auch gegen unzulässige Zugriffe gesichert werden, weshalb es einer Zugriffsrechtverwaltung benötigt. Die Verwaltung von Zugriffsrechten auf eine Entität per se stellt schon eine Herausforderung dar. Bei einer Vielzahl von digitalen Entitäten skaliert die Komplexität der Zugriffsverwaltung dementsprechend, sodass die Zugriffsverwaltung zu einer andauernden Aufgabe wird, mit der sich beschäftigt werden muss.
Die **Robert Bosch GmbH** betreibt einen Ansatz bei Ihrer Zugriffverwaltung auf digitale Entitäten, dass Zugriffsrechte mit daraus folgenden Risiken versehen werden. Diese Risiken beschreiben mögliche Folgen, die aus Misshandlung dieser Rechte entstehen können. Es gibt auch Rechte, die mit keinem Risiko verbunden sind. Beispielsweise entsteht durch ein Lese-Zugriffsrecht kein verfolgungswürdiges Risiko. Damit der Ansatz von der Vergabe von Risiken zusammen mit Zugriffsrechten firmenintern klar geregelt ist, ist er in einer *Central Directive* (kurz CD) definiert. *Central Directives* sind interne Standards, die beschreiben, wie mit bestimmten Sachen umgegangen werden soll. Die hier relevante CD wurde erst im vierten Quartal 2023 verabschiedet und ist damit noch sehr neu. Sie umfasst nicht nur die Handhabung von Risiken, sondern generell konformes Management von Zugriffsrechten. Für diese Arbeit ist die Handhabung der Zugriffsrechte zu vernachlässigen, da sich auf die Risiken spezialisiert wird.
Angehängt an die CD ist der Access Risk Katalog, der alle Risiken, die es gibt, mit Beschreibungen, Verantwortlichen und korrespondierenden SAP-Zugriffsrechten auflistet. Dabei sind die Risiken in verschiedene Gruppen je nach möglichen Auswirkungen unterteilt. Demnach gibt es Gruppen mit Risiken von Verlust von Know-how, Störung von betrieblichen Abläufen, finanziellem Schaden etc. Dieser Katalog liegt bisher in Form einer Microsoft-Excel-Tabelle mit ca. 1000 Zeilen vor. Damit diese Daten automatisiert gehandhabt werden können, soll ein Tool entwickelt werden. Die Risiken mit allen ihren Metadaten sollen in eine Datenbank migriert werden. Zusätzlich soll es eine API für die automatisierte Abfrage dieser Daten geben und zuletzt soll eine Oberfläche für Benutzer bereitgestellt werden, dass jegliche Abfragen auch manuell getätigt werden. Mit diesen Anforderungen soll ein Webserver mit dreistufiger Architektur entwickelt werden, wobei die drei Stufen eine Datenbank, ein *Back-* und ein *Frontend* umfassen.

Praxisteil

Bevor die Implementierung des Projektes beginnen kann, müssen einige Konzepte und Realisierungen vorher geplant sein. Beispielsweise muss eine Datenbank ausgesucht und ein passendes Datenbank Schema entwickelt werden. Da MSSQL bereits als Datenbank vorgeschrieben ist, braucht es für die Datenbankschicht nur noch wenige Designentscheidungen. Für die Konzeptionierung des *Backends* wird im Vorhinein ein Lasten- und Pflichtenheft erarbeitet. Dieses listet zum einen die gewünschten Features und ihre Use-Cases auf. Zum anderen wird die technische Umsetzung der einzelnen Features definiert. Neben dem eigen zu erarbeitenden Lasten- und Pflichtenheft gibt es ein eigenes von **Bosch**. In diesem sind Konzepte und Prinzipien erläutert, auf die bei firmeninterner Softwareentwicklung geachtet werden soll. Darin ist beispielsweise das Testen der Software erwähnt zusammen mit der vorgeschriebenen prozentualen Testabdeckung, die die Software haben soll. Neben dem Testen werden viele weitere Vorschriften aufgezählt, sodass bei Boschkonformer Software Entwicklung viel zu beachten ist.
Für die Implementierungsart des Webserver *Backends* gibt es dagegen weniger Vorgaben. Dies bedeutet, dass untersucht werden muss, welche Technik sich am besten für die Anforderungen eignet. Nach momentanem Stand gibt es noch zwei mögliche Techniken, die ggf. benutzt werden. Die erste ist das *Spring*

Boot Framework mit Java. Damit würde das *Backend* und die zugehörige API mit Java Code programmiert werden können. Die *KNIME Analytics Platform* ist die andere Möglichkeit. In dieser können Workflows mit Blöcken in einer grafischen Oberfläche erstellt werden. Diese Workflows können an eine API angebunden werden, sodass diese als Reaktion auf ein *HTTP-Request* ausgeführt werden. Der größte Vorteil der *KNIME Analytics Platform* ist, dass durch die grafische Oberfläche das Projekt auch von Mitarbeitern ohne Programmierkenntnisse betreut werden kann. Da die genaue Gewichtung der Anforderungen noch nicht erfolgt ist, gibt es noch keine endgültige Entscheidung, wie das *Backend* realisiert werden soll.
Für die Auswahl der Technik zur Implementierung des *Frontends* gibt es ebenfalls wenig Vorgaben. Die beiden Frameworks, die infrage kommen, wären *Vue.js* und *Angular*. Die Auswahl für diese hängt damit zusammen, welches Framework besser betreut werden kann bzw. welches Framework bereits in anderen Projekten benutzt wird.
Die größten Herausforderungen in der Vorarbeit dieser Thesis sind das Erstellen des eigenen Lasten- und Pflichtenhefts und das Verstehen und Einhalten der gegebenen Boschstandards. Die Implementierung bietet andere Schwierigkeiten. Die vorhandene Infrastruktur von **Bosch** soll dabei benutzt werden und so soll beispielsweise an die vorhandene Benutzerverwaltung angeschlossen werden.

Literatur und Abbildungen

Process Mining for Enhanced Decision-Making: A Case Study of Process Optimization

Patrick Suelzle **Harald Melcher**

Department of Computer Science and Engineering, Esslingen University

Work carried out at Bosch eBike Systems, Kusterdingen

Introduction

In today's data-driven business environment, the exponential growth of data represents both a challenge and an opportunity for companies. As the volume of data increases rapidly (see Figure 1), the capacity to effectively use and interpret this information becomes crucial in creating a competitive edge and generating value. Additionally, the optimization of business processes is critical, as they represent the backbone of organizational efficiency and success. In this context, the role of advanced data analysis technologies such as process mining becomes essential. The synergy between data resources and streamlined processes is key to unlocking value and ensuring sustained business growth. This research examines the application of process mining technology within Bosch eBike Systems, focusing specifically on optimizing their return order process.

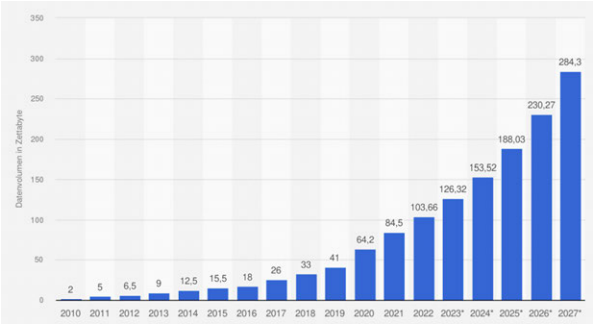


Fig. 1: Volume of digital data generated/replicated annually worldwide from 2010 to 2022 and forecast to 2027 (in Zettabyte) [2]

Objective

The aim of this research is to examine how process mining technology can be applied to analyze and potentially enhance the return order process at Bosch eBike Systems, and what impacts could these enhancements have on the decision-making process. The objectives include:

- **To examine the application of process mining technology:** Investigating how process mining is implemented within the context of Bosch eBike Systems, specifically in managing the controlling part of the return order process.
- **To develop improvement proposals based on process mining insights:** Formulate proposals for process optimization based on the analysis and insights gained from process mining.
- **To discuss the impact of these improvements on decision-making:** Evaluate how the proposed process improvements could influence and enhance the decision-making process at Bosch eBike Systems.
- **To provide recommendations for future application:** Offer recommendations for the future use of process mining in process optimization and decision-making.

Methodology

The methodology of this research is structured around a case study approach. A case study is a qualitative research method extensively used in various disciplines [4]. To achieve the previously listed objectives, this work follows a four-stage methodology. Figure 2 shows the four stages, each having multiple steps that explain in further detail how the project is realized. In the project definition phase, the main goal is to gain an understanding of the return order process and its primary problem. Based on this foundation, the objectives for improvement through process mining techniques are established alongside the specific questions that the application of process mining targets to answer. The second phase centers on identifying, extracting, and refining the data. In

this stage, the objective is to locate the essential data points, retrieve them from the information system, and then cleanse the data to convert it into a functional event log suitable for the process mining tool. This phase is arguably the most critical and time-intensive phase, as the quality of the data directly influences the accuracy of the process analysis. Poor quality in the event log leads to inaccurate analyses and potentially incorrect findings [3]. In phase three, the process mining technologies are applied to the prepared data set. The process discovery technique will provide a real-world process model based on the event data. This phase is critical for visualizing the existing process flows and identifying discrepancies between the assumed and actual process. The discovered process model will then be compared against the ideal or expected process model. This conformance checking will help pinpoint deviations and areas where the process does not align with the organizational standards or expectations. The final stage, process redesign, focuses on a process mining technique called process enhancement. Based on the insights gained from process discovery and conformance checking, the research will propose targeted recommendations for process enhancement. These recommendations will focus on improving efficiency, reducing bottlenecks, and aligning the process more closely with organizational goals. The final phase involves validating the proposed enhancements in the context of Bosch eBike Systems. This will include a detailed case study analysis to assess the impact of the recommended changes on process efficiency and decision-making.

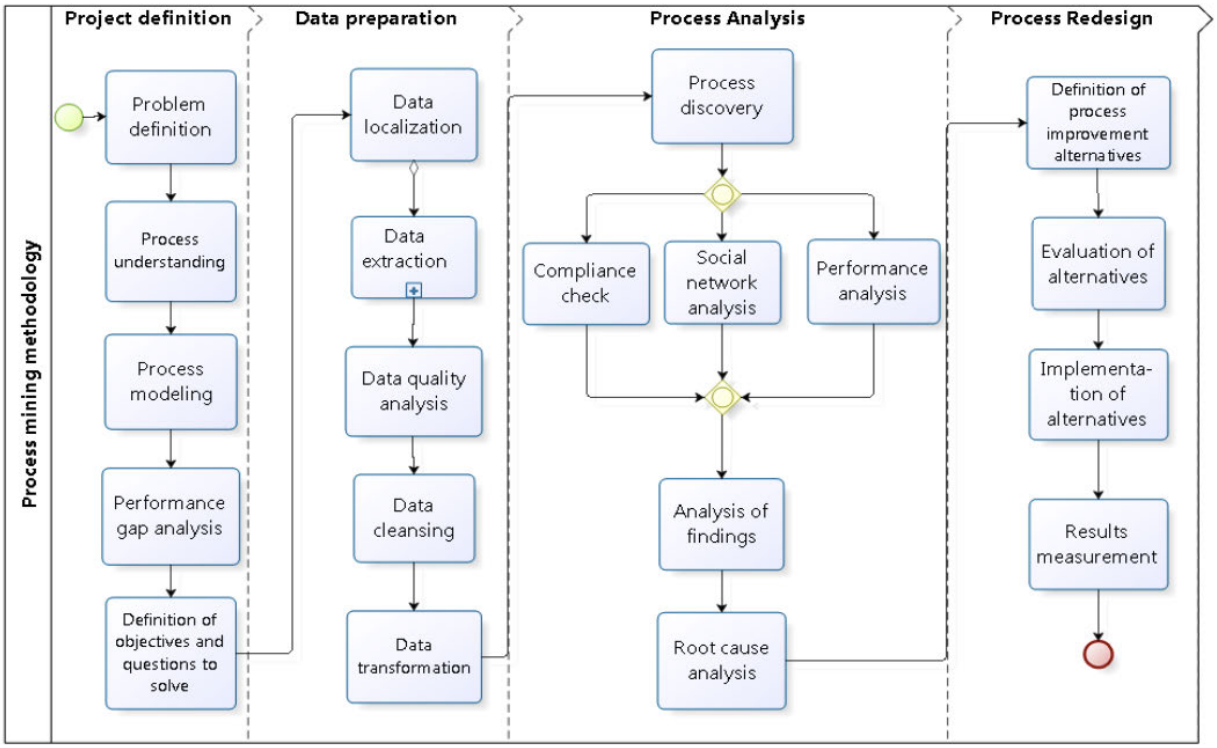


Fig. 2: Process Mining Methodology [1]

Outlook

This study is expected to demonstrate the practical benefits of process mining in a real-world business scenario. The findings can serve as a blueprint for Bosch eBike Systems to consider adopting process mining for optimizing various processes. Moreover, this research contributes to the broader understanding of data-driven decision-making in process management.

References and figures

[1] Santiago Aguirre, Carlos Parra, and Marcos Sepulveda. Methodological proposal for process mining projects. *International Journal of Business Process Integration and Management*, 8:102–113, 2017.

[2] Statista IDC. Volumen der jährlich generierten/replizierten digitalen Datenmenge weltweit von 2010 bis 2022 und Prognose bis 2027 (in Zettabyte). https://de.statista.com/statistik/daten/studie/267974/umfrage/prognose-zum-weltweit-generierten-datenvolumen/, 05 2023.

[3] W.M.P. Van Der Aalst. *Process Mining: Data Science in Action*. Springer, 2 edition, 2016.

[4] Robert K. Yin. *Case Study Research: Design and Methods*, volume 5. Sage Publications Inc., 4 edition, 2009.

Entwicklung von Use Cases zur Heizungsgestaltung von Mehrfamilienhäusern auf Basis von qualitativen Interviews mit Entscheidungsträgern

Luiza Tafa **Thomas Rodach**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma Bosch Home Comfort, Wernau

Einleitung

Mehrfamilienhäuser stehen vor der Herausforderung, ihre Heizsysteme umzugestalten, angesichts des dringenden Bedarfs, den CO2-Ausstoß zu verringern. Neue Gesetze fördern elektrifizierte Heizlösungen, erfordern aber komplexe Strategien, die Umweltziele erfüllen und Bewohnerbedürfnisse berücksichtigen. Die Nutzung von Interviews mit Entscheidungsträgern ist entscheidend, um zukunftsweisende Lösungen zu gestalten. Dieser Artikel beleuchtet die neuen Gesetze, politische Entscheidungen und die Dringlichkeit elektrifizierter Heizlösungen für Mehrfamilienhäuser. Wir suchen Wege für eine nachhaltige und effiziente Wärmeversorgung im Einklang mit den heutigen Anforderungen.

Problemstellung

Deutschland setzt neue Regeln für Heizungsanlagen in Mehrfamilienhäusern um, um den Anteil erneuerbarer Energien auf 65% zu erhöhen und auf elektrifizierte Heizsysteme umzusteigen – ein Schritt gegen den Klimawandel. Die Umstellung auf Elektrifizierung stellt eine komplexe Herausforderung dar, bedingt durch die Vielfalt von Hausgrößen, Strukturen und Bewohnerbedürfnissen. Wohnungseigentümer und Genossenschaften müssen verschiedene Anforderungen von Modernisierung bis Lärm berücksichtigen.
Diese vielfältigen Bedürfnisse zu berücksichtigen ist entscheidend, um Wärmepumpen als nachhaltige Alternative zu fossilen Brennstoffen zu etablieren. Die Kooperation zwischen Behörden, Herstellern und Eigentümern wird Deutschland helfen, Klimaziele zu erreichen, während Bewohnern Komfort und Effizienz gewährleistet bleiben. Eine nachhaltige Zukunft erfordert nicht nur technologische Innovation, sondern auch partnerschaftliche Zusammenarbeit und Verständnis für individuelle Bedürfnisse – ein Schritt in Richtung umweltfreundlicher Heizsysteme.

Aufgabe

Bei der Implementierung von Heizungssystemen in Mehrfamilienhäusern (MFH) sind klare Anwendungsfälle entscheidend. Dazu gehören effiziente Wärmepumpen sowie elektrifizierte Heizungen. Um Einblicke zu gewinnen, ist die Identifikation relevanter Ansprechpartner wie Gebäudeeigentümer Hausverwaltungen, Installateure und Technische Planer entscheidend. Entscheidungsträger benötigen klare Informationen zu Kosten-Nutzen, Förderprogrammen und langfristigen Betriebsplänen. Herausforderungen umfassen oft die Anpassung an bestehende Infrastrukturen, rechtliche Rahmenbedingungen und die Akzeptanz der Bewohner für neue Systeme. Eine gründliche Analyse ist entscheidend, um effiziente und nachhaltige Lösungen zu finden.

Effizient und nachhaltig: Die Bosch Wärmepumpe Compress 5800i AW

Die Bosch Compress 5800i AW definiert Heiztechnologie neu [1]. Ihre intelligente Technologie ermöglicht beeindruckende Heizleistung aus Umgebungsluft, gewährleistet zuverlässige Wärme in vielfältigen Klimazonen und ist mehr als nur ein Heizsystem – sie ist eine Antwort auf die Herausforderungen von morgen. Die Compress 5800i AW fokussiert nicht nur auf Effizienz, sondern auch auf Nachhaltigkeit. Durch den klaren Einsatz erneuerbarer Energiequellen reduziert sie aktiv CO2-Emissionen und setzt ein starkes Zeichen für umweltfreundliches Heizen. Mit unkomplizierter Installation und smarten Funktionen für eine präzise Heizungssteuerung geht die Compress 5800i AW über eine herkömmliche Wärmepumpe hinaus. Diese Investition in innovative Heiztechnologie ist gleichzeitig ein Schritt in Richtung einer nachhaltigen und fortschrittlichen Zukunft des umweltfreundlichen Heizens [2].

Abb. 1: Waermepumpe [5]

Die Technologie hinter den Bosch Wärmepumpen

Die neue Luft-Wasser-Wärmepumpenserie von Bosch, bestehend aus den Modellen "Compress 5800i AWünd "Compress 6800i AW", setzt wegweisende Standards in der Heiztechnologie. Mit umweltfreundlichem Propan (R290) als Kältemittel bieten diese Wärmepumpen nicht nur Sicherheit, sondern auch Kosteneffizienz. Ergänzt durch den Bosch Energiemanager können sie sich nahtlos in bestehende Photovoltaik-Anlagen und Batteriespeicherlösungen integrieren. Die "Compress 5800i AWist speziell für Neubauten konzipiert, mit verschiedenen Leistungsstufen, einem integrierten 180-Liter-Brauchwasserspeicher und einem 16-Liter-Pufferspeicher. Trotz ihrer kompakten Abmessungen bietet sie maximale Vorlauftemperaturen von 60 Grad Celsius. Beide Modelle, die "Compress 5800i AWünd "Compress 6800i AW", zeichnen sich durch schalloptimierte Bauweise aus und gehören laut Hersteller zu den leisesten Wärmepumpen ihrer Klasse. Diese Serie von Bosch ist nicht nur eine technologische Innovation, sondern auch eine zukunftsweisende Antwort auf die steigenden Anforderungen an Energieeffizienz und Umweltschutz im Heizsektor [2].

Gebäudeenergiegesetz: Mehr Erneuerbare Energien in der Heizung

Ab Januar 2024 dürfen in Neubauten innerhalb von Neubaugebieten nur noch Heizungen installiert werden, die auf 65 Prozent Erneuerbaren Energien basieren. Für bestehende Gebäude und Neubauten, die in Baulücken errichtet werden, sind längere Übergangsfristen vorgesehen [2]. Dies soll eine bessere Abstimmung der Investitionsentscheidung auf die örtliche Wärmeplanung ermöglichen [4].

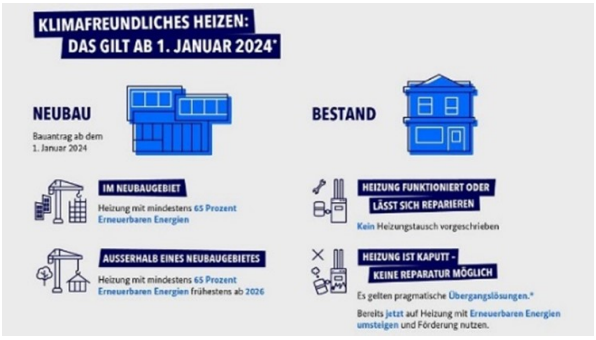


Abb. 2: Klimafreundliches Heizen [3]

Förderung für Heizungstausch

Personen, die sich dazu entscheiden, ihre Heizung zu erneuern und dabei auf 65 Prozent erneuerbare Energien umzusteigen, können staatliche Förderung in Anspruch nehmen. Diese Förderung umfasst eine Grundförderung für alle und zusätzliche Mittel, speziell für jene, die ihre Heizung besonders zügig umrüsten möchten oder für Personen mit niedrigem Einkommen. Die maximale Förderung deckt 70 Prozent der Investitionskosten ab [1].

Literatur und Abbildungen

[1] Das BAFA. Förderprogramm im Überblick. https://www.bafa.de/DE/Energie/Heizen_mit_Erneuerbaren_Energien/Foerderprogramm_im_Ueberblick/foerderprogramm_im_ueberblick_node.html, 03 2023.

[2] Emiliano Bellini. Bosch verkauft neue Propan-Wärmepumpen mit Photovoltaik-Paketlösungen. https://www.pv-magazine.de/2023/08/23/bosch-verkauft-neue-propan-waermepumpen-mit-photovoltaik-paketloesungen/, 08 2023.

[3] Die Bundesregierung. Für mehr klimafreundliche Heizungen. https://www.bundesregierung.de/breg-de/schwerpunkte/klimaschutz/neues-gebaeudeenergiegesetz-2184942, 12 2023.

[4] Die Bundesregierung. Gesetz für Erneuerbares Heizen. https://www.bundesregierung.de/breg-de/schwerpunkte/klimaschutz/neues-gebaeudeenergiegesetz-2184942, 12 2023.

[5] Bosch Home Comfort. Waermepumpe Compress 5800i AW. https://www.bosch-homecomfort.com/at/de/ocs/wohngebaeude/compress-5800i-aw-19378238-p/, 06 2023.

Konzeption und Visualisierung einer Nutzungsanalyse für Medical Devices

Stefan Tafferner **Jörg Nitzsche**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma doubleSlash Net-Business GmbH, Stuttgart

Einleitung

Die Digitalisierung schreitet in allen gesellschaftlichen Bereichen voran, so auch in der Medizinbranche. Dies spiegelt sich zunehmend in den Anforderungen der Krankenhäuser wieder. Diese fortschreitende Entwicklung eröffnet zahlreiche Möglichkeiten innerhalb der Gesundheitseinrichtung, wie beispielsweise bei der Patientenversorgung. Gleichzeitig bringt die verstärkte Integration von IT-Systemen auch in Abläufe und Prozesse Risiken mit sich. In Anbetracht der Öffnung von IT-Systemen und dem digitalen Datenaustausch sind widerstandsfähige IT-Sicherheitsstandards unerlässlich, um den aktuellen Angriffstechniken standzuhalten. Die Daten aus medizinischen Geräten sind ein entscheidender Faktor, wenn es darum geht, Innovationen voranzutreiben. Mit den erfassten Informationen können im Bereich der Medizin neue Methoden und Ansätze entwickelt sowie evaluiert werden. Die Zusammenführung der Daten aus den medizinischen Geräten innerhalb eines Krankenhauses bieten einerseits die Möglichkeit, Forschungsprojekte effektiver zu gestalten und geeignete Studienteilnehmer genauer zu selektieren. Andererseits bleiben Informationen über die Nutzung der medizinischen Geräte oftmals ungenutzt, da hier eine große Unwissenheit über den Mehrwert der Daten vorliegt oder es datenschutzrechtliche Probleme mit sich bringt [4]. Hierbei ergibt sich ein enormes Potenzial für die Hersteller medizinischer Geräte, da diese Daten dazu verwendet werden können, ihre Produkte zu verbessern. Denn unter dem steigenden Leistungsdruck im Gesundheitswesen ist eine optimale Gerätenutzung und deren Features eine Grundvoraussetzung, damit die Mitarbeitenden im Gesundheitswesen effektiv und effizient arbeiten können. Dieses Thema wird in der wissenschaftlichen Arbeit genauer beleuchtet.

Zielsetzung

Das Ziel der Bachelorarbeit besteht darin, durch die Entwicklung einer Nutzungsanalyse einen bedeutenden Beitrag zu leisten, der die Wirksamkeit und Effizienz von medizinischen Geräten verbessert. Zusätzlich soll durch die Analyse die Sicherheit medizinischer Geräte erhöht werden, wodurch sich positive Auswirkungen auf die Versorgung der Patienten ergeben.

Definition einer Nutzungsanalyse

Die Nutzungsanalyse zielt darauf ab, die konkrete Anwendung von Systemen zu untersuchen, insbesondere im Hinblick auf die Software und die Nutzung ihrer Funktionen. Durch die systematische Auswertung von Nutzungsdaten werden Informationen über die Anwendung gewonnen, um spezifische Fragestellungen eines Unternehmens zu beantworten [4]. Zum Beispiel könnte eine Fragestellung eines Unternehmens lauten: „Wie präzise sind die Messungen, die das Gerät liefert? “. Zusätzlich helfen Nutzungsanalysen dabei, Trends und Muster zu erkennen, was wiederum dazu beiträgt, Produkte auf Basis der Nutzungsdaten zu verbessern. Die Quellen für die Nutzungsdaten stammen häufig aus digitalen Angeboten wie elektronische Medizin (eMedizin), IoT und Mobile Apps [1].

Herausforderungen einer Nutzungsanalyse

Damit eine Nutzungsanalyse optimal durchgeführt werden kann, müssen verschiedene Herausforderungen bewältigt werden. Besonders wichtig ist die Beachtung des Datenschutzes bei der Erfassung von Nutzerdaten. Das bedeutet, sicherzustellen, dass die Datenerhebung und -analyse im Einklang mit den Datenschutzbestimmungen stehen und die Privatsphäre der Nutzer respektiert wird [7]. Eine weitere Herausforderung für Nutzungsanalysen besteht in der Vielfalt der verfügbaren Datenquellen und ihrer Interpretation. Hierbei ist es besonders relevant, Muster zu verstehen und interpretieren zu können, um wertvolle Erkenntnisse zu gewinnen. Ebenso entscheidend ist die Datenqualität, da unvollständige oder ungenaue Daten zu fehlerhaf-

ten Schlussfolgerungen führen können. Zudem muss technisch sichergestellt sein, dass Daten aufgezeichnet werden können, um eine Nutzungsanalyse durchführen zu können [6].

Prozess einer Nutzungsanalyse

Um eine erfolgreiche Nutzungsanalyse durchzuführen, ist es sinnvoll, den Prozess in sieben Schritte zu unterteilen. Die nachfolgende Abbildung gibt eine Übersicht über diese sieben Phasen einer Nutzungsanalyse.

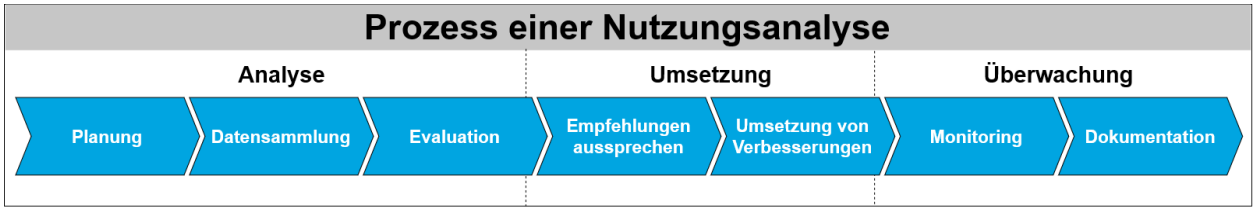


Abb. 1: Ablauf einer Nutzungsanalyse [2]

In der ersten Phase der Planung ist es entscheidend, das Ziel klar zu definieren. Hierbei geht es darum, die Fragen zu identifizieren, die durch eine umfassende Nutzungsanalyse beantwortet werden sollen. Im nächsten Schritt der Datensammlung liegt der Fokus auf der Erfassung relevanter Informationen durch Interviews, Web-Analytics und Umfragen. Anschließend erfolgt die Analyse und Auswertung der gesammelten Daten mithilfe statistischer Methoden, um Muster und Trends zu erkennen. Im vierten Schritt stehen konkrete Handlungsempfehlungen im Mittelpunkt, die auf der gründlichen Analyse der Daten basieren. Nachfolgend gilt es, die identifizierten Verbesserungen umzusetzen, wobei die zuvor durchgeführte Nutzungsanalyse als Grundlage dient. Zur fortlaufenden Optimierung werden die Ergebnisse dokumentiert, um kontinuierliche Verbesserungen zu ermöglichen.

Ziel einer Nutzungsanalyse

Das Ziel einer Nutzungsanalyse besteht darin, die Art und Weise zu bewerten, wie eine Ressource, Dienstleistung oder digitales Angebot genutzt wird. Gleichzeitig zielt die Nutzungsanalyse darauf ab, den Informationsbedarf eines Unternehmens zu ermitteln. In diesem Zusammenhang liegt der Fokus nicht auf den Bedürfnissen von Mitarbeitern, sondern auf den vorhandenen oder benötigten Informationen [4]. Ein weiterer Zweck der Nutzungsanaylse besteht darin, Produkte anhand der aus Ressourcen, Dienstleistungen oder digitalen Angeboten stammenden Daten zu verbessern. Hieraus ergeben sich verschiedene allgemeine Ziele, die eine Nutzungsanalyse verfolgt. Diese Ziele werden im Folgenden aufgelistet [6]:

Verbesserung der Benutzerfreundlichkeit: Die Analyse von Nutzungsverhalten zielt darauf ab, Probleme in Bezug auf die Benutzerfreundlichkeit zu erkennen.

Optimierung von Funktionen: Mithilfe einer Nutzungsanalyse wird deutlich, welche Funktionen häufiger und welche seltener genutzt werden. Auf dieser Grundlage können gezielte Optimierungen an Produkten vorgenommen werden.

Steigerung der Kundenzufriedenheit: Durch die Identifizierung von Nutzungsmustern haben Unternehmen die Möglichkeit, ihre Produkte an die Bedürfnisse der Kunden anzupassen, was wiederum zu einer Steigerung der Kundenzufriedenheit führt.

Informierte Entscheidungsfindung: Nutzungsanalysen sind auch in entscheidenden Prozessen wie der zukünftigen Entwicklung, Ressourcenallokation oder Marketingstrategien hilfreich.

Konzept

Um aussagekräftige Daten für eine Nutzungsanalyse zu erheben, ist eine gründliche Auswahl der Stakeholder entscheidend. Die Identifikation der relevanten Stakeholder kann durch verschiedene Methoden erfolgen, wobei die Methode der Eigenidentifikation der Stakeholder angewendet wird. Im Rahmen der Eigenidentifikationsmethode melden sich Stakeholder, die ein Interesse an der Umsetzung des Projekts haben, eigenständig. Diese Methode gewinnt an Bedeutung, da die Beteiligung dieser Personen durch ihr eigenes Interesse deutlich wird. Eine weitere Methode besteht darin, Stakeholder zunächst durch eine eigenständige Recherche zu identifizieren und die Ergebnisse durch anschließende Befragungen zu bestätigen. Nach erfolgreicher Identifikation der Stakeholder liegt der Fokus darauf festzustellen, welche Wünsche, Ziele, Bedürfnisse und Ängste sie in Bezug auf die Anforderungen einer Nutzungsanalyse haben könnten. In diesem Kontext kommen narrative Interviews zum Einsatz, die teilweise offene Fragen beinhalten. Die Verwendung offener Fragen bietet den Vorteil, dass eine gewisse Spontanität ermöglicht wird und zusätzlich die Möglichkeit besteht, neue Informationen zu gewinnen, die bisher noch nicht in Betracht gezogen wurden [5]. Um die Erkenntnisse aus den Interviews besonders anschaulich und prägnant darzustellen, werden Personas entwickelt [3]. Personas

bieten eine umfassende Einsicht in die Bedürfnisse, Wünsche und Ziele der Zielgruppen. Diese Informationen bilden die Grundlage für eine effektive und effiziente Nutzungsanalyse, die anschließend mithilfe von Power BI visualisiert werden kann.

Ausblick

Momentan werden die Interviews kategorisiert, um die Bedürfnisse der Zielgruppen umfassend zu erfassen. Die gewonnenen Informationen dienen als Grundlage für Key Performance Indicators (KPIs). Die Entwicklung dieser KPIs ist von entscheidender Bedeutung, da sie es Unternehmen die Gelegenheit bieten, durch Nutzungsanalysen Trends und Muster zu erkennen. Die durch KPIs gelieferten Kennzahlen ermöglichen eine präzise Beurteilung, ob die angestrebten Ziele erreicht werden oder nicht.

Literatur und Abbildungen

[1] Beck Clara. UX Analyse anhand von Nutzungsdaten – Usability mit Big Data optimieren. https://www.centigrade.de/de/blog/ux-analyse-anhand-von-nutzungsdaten/, 2020.

[2] Eigene Darstellung.

[3] H.G. Häusel and Henzler H. *Buyer Personas: Wie man seine Zielgruppen erkennt und begeistert*. Haufe Fachbuch, 2018.

[4] Philipp Niemann, Van Den Bogaert Vanessa, and Ziegler Ricarda. *Evaluationsmethoden der Wissenschaftskommunikation*. Springer Fachmedien, 2023.

[5] Karl-Heinz Renner and Jacob Nora-Corina. *Das Interview: Grundlagen und Anwendung in Psychologie und Sozialwissenschaften*. Springer Fachmedien, 2020.

[6] Duernay Stefan. Mit Nutzungsanalysen aus Daten lernen. https://blog.doubleslash.de/mit-nutzungsanalysen-aus-daten-lernen, 2023.

[7] Markus Vollmert. *Google Analytics 4 : Grundlagen, Praxis, Migration*. Rheinwerk Verlag, 2023.

Anwendung von Natural Language Processing (NLP) zur Erkennung von Hassrede in Social Media Daten

Ivan Filip Terzic **Dirk Hesse**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt an der Fakultät Informatik und Informationstechnik, Esslingen

Einleitung

In der Ära der digitalen Kommunikation haben sich soziale Medien zu einer wichtigen Plattform unseres alltäglichen Lebens entwickelt. Mit der wachsenden Bedeutung dieser Plattformen ist jedoch das Risiko von Diskriminierung, Hass und Hetze in der digitalen Welt rapide gestiegen. In diesem Zusammenhang werden im Rahmen dieser Arbeit verschiedene NLP-Ansätze betrachtet Hassrede in Social Media Daten zu erkennen.

Ziel der Arbeit

In der vorliegenden Arbeit werden verschiedene Methoden zur Erkennung von Hassrede in Social Media Daten untersucht. Ziel ist es eine Gegenüberstellung der verschiedenen Ansätze durchzuführen und diese in Hinsicht klassischer Klassifikation-Metriken wie *Accuracy, Precision, Recall* und *F1-Score* zu bewerten. Außerdem soll die Komplexität der Implementierungsansätze betrachtet und gegenübergestellt werden.

Theoretische Grundlagen

Bei der Klassifizierung von Texten – in diesem Fall Social Media Posts - handelt es sich i.d.R. um ein sog. Seq2Label-Problem, bei dem eine Input-Sequenz (der eigentliche Text) einer Klasse zugeordnet wird. Im Falle dieser Arbeit werden die Daten in die Klassen *hate* und *nothate* eingeteilt, sodass es sich um ein binäres Klassifikationsproblem handelt. Da die verwendeten *Machine Learning*-Modelle, Neuronale Netze (NN) und das Transformer-Modell BERT (*Bidirectional Encoder Representation from Transformers*) i. d. R. numerische *Input*-Werte benötigen, müssen die textuellen Daten zunächst in eine numerische Repräsentation überführt werden. Eine einfache, aber durchaus nützliche Möglichkeit Textdaten in numerischer Form zu repräsentieren, bietet das sog. TF-IDF Modell (*Term Frequency Inverse Document Frequency*). Das TF-IDF Modell bestimmt für jedes Wort in einem sog. *Document* (ein Satz, ein Paragraf oder ein ganzes Buch) die *Term Frequency* für jedes Wort und multipliziert diese mit der *inverse Document Frequency*. Dies hat den Hintergrund, dass Stoppwörter, also Wörter die keine semantische Relevanz haben, zwar häufig vorkommen und daher eine hohe *Term Frequency* haben, jedoch keinen semantischen Mehrwert bieten und somit eine geringere Gewichtung haben sollten.

Eine weitere Art Wörter in eine numerische Repräsentation zu überführen sind sog. *Word Embeddings* (dt. Wort Einbettungen). Das Ziel von *Word Embeddings* ist es, semantisch ähnliche Wörter in ähnliche Wortvektoren umzuwandeln, sodass folgende Gleichung gelöst werden kann: $king - man + woman = queen$. Die Grundlage von *Word Embeddings* bildet ein neuronales Netz, welches mit Hilfe einer Pseudo-Aufgabe lernt, ausgehend von einem Kontext (d.h. der umliegenden Wörter) ein Zielwort zu bestimmen. In Folge des Trainingsprozesses werden Gewichtungen zwischen den Neuronen der *Input-Layer* und der *Hidden-Layer* gelernt und in einer Gewichtsmatrix gespeichert. Diese Gewichtsmatrix enthält die eigentlichen *Word Embeddings*, wobei für jedes Wort aus dem Vokabular eine Zeile bzw. Spalte aus der Gewichtsmatrix entnommen werden kann, die das Wort in Vektorform repräsentiert.

Zur Klassifikation werden in der Arbeit drei verschiedene Ansätze verfolgt: Erstens, die Klassifikation mit Hilfe von *Machine Learning* Methoden. Hierzu werden drei Klassifikatoren trainiert. Ein *Naive Bayes* Klassifikator, ein *Logistic Regression* Klassifikator und eine *Support Vector Machine*; zweitens, ein *Deep Learning* basierter Ansatz mit einem LSTM (*Long Short-Term Memory*) neuronalen Netz; drittens, eine Klassifikation mit dem Transformer-basierten Modell BERT.

Methodik & Vorgehen

Im Rahmen der Arbeit wird bei der Klassifikation von Social Media Posts ein Phasenweises vorgehen, wie in Abbildung 1 gezeigt, angewendet.

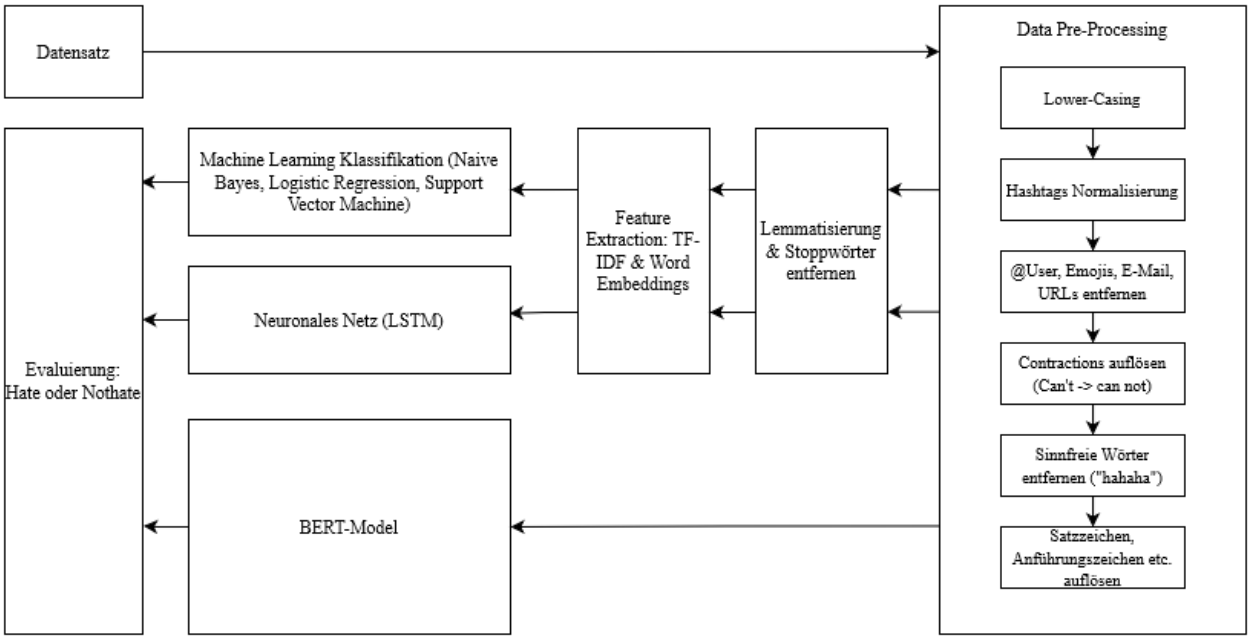


Abb. 1: Flowchart zur Klassifizierung von Hassrede mit ML, LSTM & BERT [1]

Data Pre-Processing

Um bessere Ergebnisse der einzelnen Modelle zu erzielen, wurden die Daten im sog. *Pre-Processing* Schritt bereinigt. Hierzu wurde eine *Pre-Processing-Pipeline* entwickelt, die einen einzelnen Social Media Post erhält und eine Reihe von Operationen auf dem Text ausführt. Die *Pipeline* führt dabei folgende Schritte durch:

1. Alle Textzeichen in Kleinschreibung umwandeln.
2. Hashtags normalisieren (bspw. „#ILoveDogs“ wurde zu „i love dogs“ umgewandelt).
3. @User Erwähnungen, Emojis, E-Mail-Adressen, URLs wurden entfernt.
4. Contractions wurden aufgelöst („can't“ wird zu „can not“).
5. Sinnfreie Wörter wurden entfernt (bspw. „haha-hahah“).
6. Satzzeichen, Anführungszeichen, Tabs, New Lines etc. wurden entfernt
7. Non-ASCII-Textzeichen wurden entfernt.

Auf das Entfernen von Stoppwörtern und einer Lemmatisierung/*Word Stemming* wurde in diesem Schritt noch verzichtet, da das BERT-Modell zu besseren Ergebnissen kommt, wenn der Kontext eines Satzes vollständig vorhanden ist.

Machine Learning Klassifikation

Die erste Methode, die verwendet wurde, ist die Klassifikation von Hassrede mittels ML-Modellen. Hierzu wurden die vor-verarbeiteten Daten aus dem *Pre-Processing* Schritt noch weiterverarbeitet. Zunächst wurden alle Stoppwörter der Textdaten entfernt. Bevor die Klassifikatoren trainiert werden können, werden diese noch in Trainings- und Testdaten mit einem Verhältnis von 80 % zu 20 % aufgeteilt. Da der verwendete Datensatz unbalanciert ist, d.h. das Verhältnis von *hate* zu *nothate* Instanzen ist unausgewogen, werden die Trainingsdaten in einem *Balancing*-Schritt balanciert. Hierzu wurde die Anzahl der *nothate* Instanzen mit der *Random-Undersampling*-Strategie reduziert. Dieser Prozess ist essenziell, um ein *Bias*, d.h. eine Voreingenommenheit, hin zur Mehrheitsklasse in den Modellen zu verhindern. Nachdem Vorbereitung und Balancierung der Daten, wurde diese mit einem *TfIdfVectorizer* in eine numerische Form überführt. Jeder Satz besteht nun aus einem Vektor, bspw. einer Python-Liste, wobei jedes Element ein *Tuple* aus einem Index zum eigentlichen Wort im Vokabular und dem entsprechenden TF-IDF-*Score* des Wortes im jeweiligen *Document* enthält. Im Anschluss wurden die drei Klassifikatoren trainiert und evaluiert.

LSTM

Der zweite Ansatz zur Klassifikation von Hassrede ist die Verwendung eines LSTM-basierten neuronalen Netzes. Vor der Einführung von LSTM, galten rekurrente neuronale Netze (RNN) als *State of the Art-*

Ansatz *Seq2Seq*- und *Seq2Label*-Probleme zu lösen. Aufgrund ihrer Abhängigkeit zu einem versteckten Zustand h(t-1) hatten klassische RNNs die Limitierung langfristige Abhängigkeiten in sequenziellen Daten nur beschränkt erfassen zu können. Diese Limitierung wurde mit Hilfe von LSTM (Long Short-Term Memory) adressiert. Auch in diesem Schritt erfolgten die Schritte, Stoppwort-Entfernung und Lemmatisierung. Da neuronale Netze deutlich mehr Parameter als klassische, simplere ML-Modelle optimieren müssen, wurde auf eine Balancierung des Datensatzes verzichtet. Zwar besteht das Risiko, dass das NN ein *Bias* hin zur Mehrheitsklasse lernt, jedoch wurde in der finalen Evaluierung festgestellt, dass das mit dem unbalancierten Datensatz trainierte LSTM-NN bessere Ergebnisse lieferte als mit einem balancierten Datensatz.

BERT

BERT (Bidirecitonal Encoder Representation from Transformers) [2] ist ein von Google im Jahr 2018 vorgestelltes Modell, welches auf dem Transformer-Modell [4] basiert. Diese Art der Sprachmodelle gelten heutzutage als *State of the Art*-Ansätze in *Seq2Seq*- und *Seq2Label*-Problemen. Das BERT-Modell besteht aus mehreren zwölf hintereinander geschalteten *Encodern*, deren Ziel es ist, den Kontext eines Satzes in einem kontextualisierten *Word Embedding* zu erfassen. Durch die Erfassung des Kontextes ist es mit dem BERT-Modell möglich, eine bessere Repräsentation der Textdaten zu erzielen. Diese Erfassung des Kontextes verspricht eine bessere Performance bei unterschiedlichen NLP-Problemen; auch in der Klassifikation von Hassrede [3].

Vorläufige Ergebnisse

Metrik / Model	Accuracy	Precision	Recal	F1-Score
Naive Bayes	0.91	0.53	0.84	0.65
Logistic Regression	0.93	0.61	0.81	0.69
SVM (linear)	0.92	0.62	0.84	0.71
LSTM-NN	0.96	0.84	0.74	0.79
BERT*	-	0.96	0.96	0.98

Abb. 2: Vorläufige Ergebnisse [1]

Abb. 2 zeigt die für die Klasse *hate* evaluierten Ergebnisse. Aus diesen geht hervor, dass Modelle basierend auf neuronalen Netzen (BERT & LSTM) bessere Ergebnisse erzielen als ML-Modelle. Die Metrik *Precision* gibt an, wie viel Prozent der Klasse *hate* der jeweilige Klassifikator richtig erkannt hat.

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

Der *Recall* ist eine Metrik, welche in diesem Zusammenhang angibt, wie sehr man den Ergebnissen des jeweiligen Klassifikators vertrauen kann. Sie beinhaltet TP (*True-Positiv*) als auch FN (*False-Negative*), d.h. Beiträge die Hassrede enthalten und richtig als Hassrede klassifiziert wurden sowie Beiträge, die keine Hassrede enthalten und fälschlicherweise als Hassrede klassifiziert wurden.

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

Wie die vorläufigen Ergebnisse zeigen, haben Modelle, die auf neuronalen Netzen basieren, einen Vorteil gegenüber klassischen ML-Methoden (Precision). Dies liegt zum einen daran, dass die bei den ML-Methoden verwendete TF-IDF-Repräsentation nur bedingt den semantischen Kontext erfassen kann und Hassrede in einigen Fällen implizit und nicht offensichtlich ist. Zudem können mit einer TF-IDF-Repräsentation keine weitreichenden Abhängigkeiten in sequenziellen Daten erfasst werden. Somit lassen sich kontextuelle Informationen eines Satzes nur bedingt abbilden. Hier zeigt sich in der Evaluierung, dass sowohl das LSTM-NN als auch das BERT-Modell eine bessere Performance zeigen.

Ausblick

Im weiteren Verlauf der Arbeit sollen die einzelnen Modelle weiter verbessert und gegenübergestellt werden. Zu erwarten sind leichte Performancevorteile des LSTM-NN sowie BERT im Gegensatz zu klassischen ML-Methoden. Einerseits durch die Verwendung *Word Embeddings* sowie durch die Fähigkeit von NNs komplexere Funktion modellieren zu können. Zudem könnte ein Hyperparameteroptimierung der ML-Methoden zu einer Verbesserung der vorläufigen Ergebnisse führen.

Zusatz:

Abb. 2 BERT*: Eigene Ergebnisse lagen zum Zeitpunkt der Fertigstellung des Artikels noch nicht vor und wurden der Vollständigkeit aus [3] verwendet.

Literatur und Abbildungen

[1] Eigene Darstellung.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. https://arxiv.org/abs/1706.03762, 2019.

[3] H. Saleh, A. Alhothali, and K. Moria. Detection of Hate Speech using BERT and Hate Speech Word Embedding with Deep Model. https://arxiv.org/abs/2111.01515, 2021.

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention Is All You Need. https://arxiv.org/abs/1706.03762, 2017.

Business Case für den Einsatz von SAP Signavio bei Endkunden einer mittelständischen IT-Unternehmensberatung

Daniel Tesfaye **Thomas Rodach**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt an der Fakultät Informatik und Informationstechnik, Esslingen

Einleitung

IT-Unternehmensberatungen stehen vor der Herausforderung, in einem dynamischen Wettbewerbsumfeld kompetitiv zu bleiben. Sie müssen stets auf dem neuesten technologischen Stand sein, um Kunden mit IT-Dienstleistungen bedienen und unterstützen zu können. Dabei soll für den Kunden eine passende, nutzenschaffende IT-Dienstleistungen erbracht werden, welche einen Mehrwert erzeugen sollte [3]. Die Herausforderung, wettbewerbsfähig zu bleiben, betrifft auch produzierende Unternehmen. Technologische Entwicklungen und steigende wirtschaftliche Unsicherheiten erhöhen den Wettbewerbsdruck. Die Komplexität der Unternehmensstrukturen nimmt aufgrund verschiedener Faktoren wie die Globalisierung oder der technologische Fortschritt zu [2]. Dies führt zu komplexeren und dynamischeren Geschäftsprozessen. Um diese Herausforderungen zu bewältigen, ist ein solides Geschäftsprozessmanagement sehr wichtig. Indem Geschäftsprozesse schnell, effizient und flexibel gestaltet werden, kann sowohl die Effizienz, als auch die Effektivität eines Unternehmens direkt beeinflusst werden [1]. In diesem Kontext können geeignete Softwarelösungen als Werkzeuge dienen, um die entsprechenden Aufgaben zu bewältigen und so Abhilfe zu schaffen. Hier sticht SAP Signavio als Plattform-Lösung für Geschäftsprozessmanagement hervor.

Ziel der Arbeit

Das Ziel der Bacheloarbeit ist es, Kosten und Nutzen zu ermitteln, die sich durch den Einsatz von SAP Signavio ergeben würden. Es soll sowohl die Kunden-Perspektive, als auch die IT-Dienstleister-Perspektive betrachtet werden. Das bedeutet, dass zum einen Kosten und Nutzen betrachtet werden sollen, die sich für den Kunden einer IT-Unternehmensberatung durch den Einsatz von SAP Signavio ergeben. Hierfür dient eine Studie als Grundlage, die bereits von dem renommierten Beratungs- und Forschungsunternehmen Forrester durchgeführt wurde. Zum anderen sollen Kosten und Nutzen betrachtet werden, die sich für eine mittelständische IT-Unternehmensberatung durch den Einsatz von SAP Signavio als Dienstleistung ergeben. Dies soll mit Hilfe der Erstellung eines Business Case erfolgen. Die Arbeit zielt auch darauf ab, ein grundlegendes Verständnis für die SAP Signavio Softwarelösung zu vermitteln, wobei Prozessmodellierung und Process Mining als zentrale Themen behandelt werden. Zudem soll untersucht werden, ob und inwieweit die Integration von SAP Signavio in das Dienstleistungsangebot des IT-Dienstleisters einen positiven Business Case und somit einen Mehrwert für beide Parteien darstellt. Dabei soll die Frage beantwortet werden, ab wann der Einsatz von Signavio sinnvoll ist und welche Faktoren hierbei eine Rolle spielen. Es soll auch die Frage beantwortet werden, welche potenziellen Beratungsansätze entstehen können in Bezug auf angebotene Signavio-Dienstleistungen.

SAP Signavio

SAP Signavio ist ein Tool für Geschäftsprozessmanagement. Unternehmen können dazu befähigt werden, ihre Geschäftsprozesse zu identifizieren, modellieren, analysieren, optimieren und zu transformieren. Bei der SAP Signavio Business Transformation Suite handelt es sich um eine modular aufgebaute All-in-One-Plattform für Geschäftsprozessmanagement. Die All-in-One-Plattform ist als Software-Suite zu verstehen, welche wiederum aus verschiedenen Module besteht. Die Module bieten Funktionalitäten an im Bereich Prozessmodellierung, Prozessanalyse bzw. Process Mining und Prozessoptimierung. Die folgende Abbildung stellt die „Business Transformation Suite“ mitsamt aller Module dar [5].

Abb. 1: SAP Signavio Transformation Suite [4]

Business Case

Für die Erstellung eines Business Case gibt es keinen konkreten standardisierten Ansatz. Es ist entscheidend, dass die Ergebnisse umfassend erarbeitet werden. Dies gewährleistet eine ganzheitliche Dokumentation aller relevanten Informationen. Im Kontext der vorliegenden Bachelorarbeit wird die Struktur des erstellten Business Case zweckmäßig angepasst und optimiert. Die folgende Abbildung zeigt den Aufbau des Business Case.

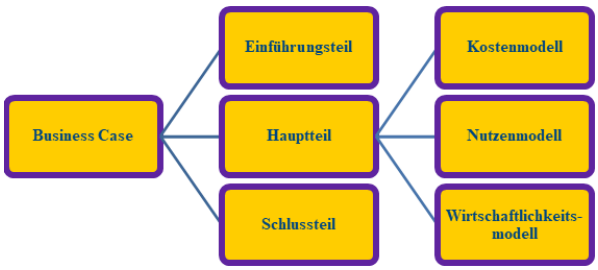


Abb. 2: Aufbau des Business Case [4]

Der in der vorliegenden Arbeit erarbeitete Business Case bezieht sich auf ein Modellunternehmen, welches stellvertretend für eine mittelständische IT-Unternehmensberatung steht. Im Rahmen eines Kundenprojektes soll SAP Signavio eingeführt bzw. implementiert werden. Bei allen Modellen des Business Case wird darauf geachtet, dass die Berechnungen auf zuverlässige Kennzahlen basieren. Getroffene Annahmen und Schätzungen werden logisch begründet und gleichzeitig so gering wie möglich gehalten. In Bezug auf die Kalkulationen wird eine passivere Rechnungsweise vorgenommen, um eine übermäßig positive Darstellung der Ergebnisse zu vermeiden und stattdessen ein realistischeres und ausgeglicheneres Bild zu präsentieren. Zusätzlich wird eine Risikobereinigung auf die Berechnungen durchgeführt.
Das Nutzenmodell umfasst eine Aufstellung der Nutzenpotentiale. Hier werden anderem Aspekte ermittelt, wie z.B. die Entlohnung der Dienstleistungen, Kundenanbindung, Wettbewerbsvorteil, Transparenz in der Projektdurchführung, usw. Das Kostenmodell umfasst eine Aufstellung der Kostenfaktoren. Unter anderem werden Aspekte ermittelt, wie z.B. Vertriebskosten, Schulungs- und Zertifizierungskosten, Lohnkosten für Entwickler und Berater, usw. Die Faktoren beider Modelle werden detailliert betrachtet und analysiert. Das Wirtschaftlichkeitsmodell beinhaltet verschiedene Berechnungen, welche unterschiedliche Aussagen über das SAP Signavio Projekt liefert. Hierbei werden folgende Investitionsberechnungen angewendet: Risikobereinigter Investitions-Cashflow, die Barwertmethode, die Amortisationsdauer und der Return on Investment. Das Wirtschaftlichkeitsmodell wird abschließend mit einer anschließenden qualitativen Risikoanalyse ergänzt.

Ausblick

Für Endkunden einer mittelständischen IT-Unternehmensberatung ist damit zu rechnen, dass aufgrund fortschreitender technologischer Trends weitere SAP-Softwarelösungen die SAP Signavio Lösung ergänzen können, wodurch zusätzlicher Nutzen erzielt werden kann. Für die mittelständische IT-Unternehmensberatung ist damit zu rechnen, dass das eigene Dienstleistungsangebot in Zukunft optimiert werden müsste. Diese Notwendigkeit kann sich aufgrund der zukünftigen technologischen Trends entwickeln. Das Dienstleistungsangebot kann z.B. in dem Sinne angepasst werden, dass sich das IT-Beratungsunternehmen auf spezielle Themen spezialisiert, welche sich nicht durch Eigenleistung des Kunden umsetzen lassen. Auch kann der Fokus der Beratung auf die Eigenleistung für den Kunden gesetzt werden. Die würde bedeutet, dass Kunden dahingehend unterstützt werden, ihre Eigenleistungen zu verbessern.

Literatur und Abbildungen

[1] Cristian Bartmann and Julia März. Process Mining: Markttendenzen, Praxishürden, Erfolgsfaktoren. *Manager Magazin*, page 3, 2022.

[2] Wolfgang Becker, Brigitte Eierle, Alexander Fliaster, Björn Ivens, Alexander Leischnig, Alexander Pflaum, and Eric Sucky. *Geschäftsmodelle in der digitalen Welt*. Sucky, Eric;, 2019.

[3] Robert Bodenstein, Ilse Andrea Ensfelnner, and Josef Herget. *Exzellenz in der Unternehmensberatung*. Herget, Josef, 2 edition, 2022.

[4] Eigene Darstellung.

[5] SAP SE. SAP Signavio Process Transformation Suite. https://www.signavio.com/de/products/process-transformation-suite/, 2023.

Improving the Session Table Handling of Stateful Firewalls to Achieve Constant-Time Packet Filtering

Dennis Tudenhoefner **Tobias Heer**

Department of Computer Science and Engineering, Esslingen University

Work carried out at Department of Computer Science and Engineering, Esslingen

Introduction

With respect to the trend of growing industrial network infrastructures, IT security measures are becoming increasingly essential. In recent years, companies struggled with severe production downtimes, damaged plants, and data loss due to cyber attacks. A basic and suitable measure against attacks is the separation of the network into zones. This concept is called network segmentation. Network administrators can place firewalls between the zones to analyze and filter the traffic. The mechanisms to filter packets with firewalls are implemented either in hardware (e.g., using an application-specific integrated circuit, ASIC) or in software (i.e., using a CPU). Filtering in hardware only increases the delay and jitter of the packets slightly, which makes it a suitable solution for industrial networks [9]. However, filtering in hardware has its drawbacks as well: vendor-specific implementations, less flexibility in filtering rules, and a limited total amount of rules. Filtering in software, on the other hand, increases the jitter of the packets, but offers more flexibility in its implementation and runs on commodity hardware. Due to the limitations of hardware-based firewalls, our goal is to improve the filtering performance of software firewalls.

So far, there are no software firewall implementations available that guarantee constant-time packet filtering in order to satisfy the latency requirements of industrial networks. To achieve constant-time packet filtering with software firewalls, we adapt existing packet filtering mechanisms. Firewalls usually rely on two packet filtering mechanisms:

1) Stateless filtering: Stateless filtering is designed to filter packets by comparing packet header fields with a set of rules. To allow or block a packet from passing through the firewall, the 5-tuples (source IP address, source port, destination IP address, destination port, protocol) of the rules are matched against the header fields of the arriving packet. Based on the result, the firewall takes an action defined in the matched rule.

2) Stateful filtering: Stateful filtering tracks the state of existing sessions. It uses a session table, also called connection state table, to track packets belonging to a known active session. The firewall always matches incoming packets with the 5-tuple entries of the session table. When it finds a matching entry, it allows the packet to pass through and updates the corresponding table entry with the latest session state information. If the packet does not match any entry, the firewall processes it using the stateless filtering mechanism and decides whether to allow or block the packet.

Due to the fact that stateless firewalls process rules one after another, the latency for packet processing increases linearly with the growing number of rules. Hence, an increasing number of rules can slow down the packet processing time of the firewall excessively. We use stateful filtering as a basis for packet filtering with low jitter. Storing the session table as a hash table enables the firewall to perform session table lookups in constant time. The firewall extracts and hashes the 5-tuple of an arriving packet and uses it to access the corresponding table entry directly. The firewall is not required to iterate over every single field of the table entry anymore. Therefore, stateful filtering is better suited for time-critical applications than stateless filtering. However, the problem with stateful filtering is that the implementation, as it is found in software firewalls, is unsuitable for time-critical applications. For example, initial packets of a session pass through the jitter-prone stateless filtering mechanism instead of the stateful filtering mechanism.

Our approach is to modify and extend the stateful filtering mechanism in order to make it suitable for time-critical applications. Hereby, we use the existing high-performance software networking stack FD.io VPP [4] with the included Access Control List (ACL) plugin as a proof-of-concept (PoC) implementation.

Related Work

Packet filtering is implemented either in software or in hardware. One difference between both implementations is the delay of forwarding packets caused by their processing time. Wüsteney *et al.* [9] compare the filtering performance between hardware- and software-based firewalls regarding the usability in TSN networks. According to them, software-based implementations are more affected of varying CPU load which causes additional jitter. Zvabva *et al.* [10] present measurements of network packet latency, jitter, and packet loss caused by the introduction of industrial firewalls when the network is segmented with the concept of zones, security levels, and conduits according to the security standard IEC 62443.

Schramm [7] presents and implements three ideas to achieve low jitter and latency on software firewalls. However, he does not consider stateful sessions in his firewall design. The firewall interrupts the stateless check of the packet of a new session after a certain time limit exceeds. After the time limit exceeded, the firewall forwards the packet without a complete check to limit the maximum jitter. Assuming this packet belongs to a session that will be stored in the session table, this packet passes through the stateless filtering mechanism only once. The following packets of the session only pass through the stateful filtering mechanism, assuming the first packet was checked completely and is allowed. This behavior faces a security issue, which we want to avoid in this work.

Another research area focuses on session table enhancement techniques to improve the processes to create, delete, and lookup session table entries. Chomsiri *et al.* [1] propose a session tracking system with a hash table for tree-rule firewalls that reduces memory consumption and processing time. Moreover, they show that the processing speed of their stateful firewall implementation is much faster than iptables. The tree-rule firewall is introduced by the same authors in [3]. It is a modified firewall that organizes its rules in a designated tree structure, not in lists. The work presented in [2] proposes a hybrid firewall implementation that takes advantage of both the tree-rule and stateless filtering mechanism to ensure high packet processing speed without rule conflicts. In addition, the authors added a feature that moves frequently matched rules to higher positions in the rule list automatically. They measure the firewall speed drop (in terms of packets per second, and megabytes per second) and packet loss with raising number of rules. However, they do not intend to measure jitter or latency, which we want to consider in this work.

Rovniagin and Wool [6] revisit a classic algorithm from computational geometry and integrate it within the filtering mechanism of iptables. The algorithm, called Geometric Efficient Matching (GEM), performs packet matching in O(d log n) time and requires $O(n^d)$ space in the worst case, where n is the number of rules and d the number of fields in the packet header to match. Their optimized GEM-iptables implementation sustains a packet matching rate of over 30,000 packets per second (pps), with 80-byte packets and 10,000 rules, without packet loss on a standard PC workstation. In comparison, the unmodified Linux iptables could only sustain a rate of around 2,500 pps. However, the space complexity of the algorithm is impractical for our implementation, as we aim to achieve packet matching in O(1) time.

To the best of our knowledge, there is only one research work that proposes an integrated solution that enhances stateful packet filtering and the session table architecture. Trabelsi and Zeidan [8] propose a session table architecture that invokes the hash function only once per session to reduce memory space consumption and filtering time. According to them, storing all session state information in one table entry causes additional processing time, especially for session table timeout attributes. Therefore, they separate the session table entries (session states and timeout attributes) into two different data structures to enhance the session table lookup and processing time. Nonetheless, the authors do not consider the latency impact of the firewall caused by the use of stateless filtering. We intend to avoid the firewall from using the stateless filtering mechanism by using the session table.

Design

To date, software firewalls implementing stateful filtering are not suitable for constant-time packet filtering. This is because the first packet of a session passes through the stateless filtering mechanism. As a result, this behavior increases the latency and jitter of the firewall.

In the context of this work, we examine different approaches to modify and improve the stateful packet filtering mechanism. Hereby, we aim to reduce the jitter to gain a constant latency of processing packets with our software-based firewall implementation. The following section describes the approaches and difficulties to improve and implement stateful filtering efficiently in order to make our firewall implementation suitable for time-critical applications.

a) Explicit use of the session table: Stateless filtering cannot guarantee a constant latency, as the duration to process a packet depends on the number of rules that need to be matched. The use of the session table will help to overcome this problem. In iptables the firewall performs the stateless check despite a match in a previous stateful check. Hereby, the firewall does

not take advantage of the session table regarding the constant-time behavior. In contrast, FD.io VPP solves this problem and skips the stateless check if there is a match in the session table. However, there is one exception in VPP: When the firewall sees a session for the first time, the first packet of that session passes through the stateless check because there is no existing entry in the session table. This means that incoming packets of a new session always need to pass through the stateless check. As a result, this behavior increases the latency and jitter of the firewall for some packets, which is not suitable for industrial applications.

b) Improving the session tracking mechanism: Modifying (i.e., deleting or adding) the firewall rules in VPP leads to the deletion of all session table entries. In addition, the firewall uses timeouts to terminate and delete old sessions in the session table automatically. Whenever the session table is empty or no existing entry matches, all initial packets need to pass through the stateless check. In industrial networks, time-critical streams are known. In order to avoid the first packet from passing through the stateless check, we insert the known sessions statically into the session table. This guarantees that the entries are always present in the session table. Furthermore, by implementing static entries we prevent session table entries from being deleted by timeouts. However, not all information about the static entries is known at the time of inserting the entries into the session table. For example, when a client connects to a server, the destination port is usually known beforehand while the source port is unknown because it is ephemeral. However, the session table requires that all 5-tuple fields are known. We solve this problem by ignoring the source port of certain known sessions within the session table. We implement this by hashing individual parts of the 5-tuple, not the entire 5-tuple.

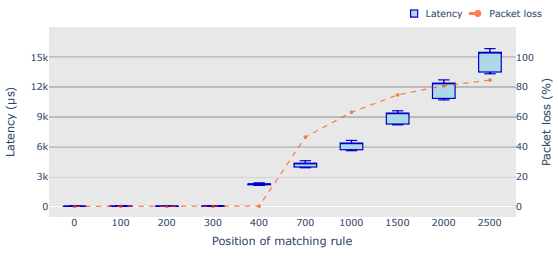


Fig. 1: Measured latency for different rule matching positions using stateless filtering on the unmodified firewall (packet size: 64 B; data rate: 500 Mbit/s) [5]

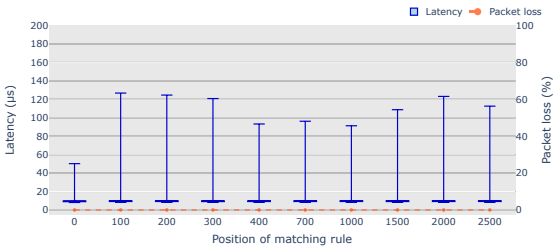


Fig. 2: Measured latency for different rule matching positions using dynamic stateful filtering on the unmodified firewall (packet size: 64 B; data rate: 500 Mbit/s) [5]

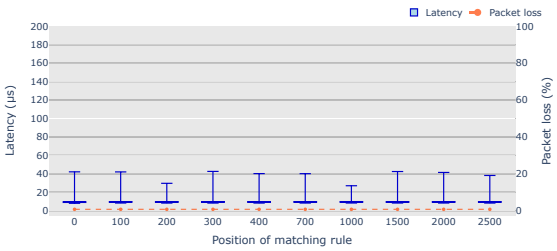


Fig. 3: Measured latency for different rule matching positions using static stateful filtering on the modified firewall (packet size: 64 B; data rate: 500 Mbit/s) [5]

Evaluation

The evaluation of this work consists of measurements of latency, jitter, packet loss, and overhead of our implementation. We compare the time behavior of our firewall before and after modification.

To perform the measurements, we use the small-form-factor firewall Protectli VP2420 with an Intel Celeron J6412 quad-core CPU. Our proof-of-concept firewall version is installed and deployed on Ubuntu 22.04 with a custom kernel to change the build options. This improves the performance of the firewall. We use a dedicated traffic generator to put specific load on the firewall. A special switch is used to conduct time stamping on the packets in order to measure the latency.

Figure 1 presents the measured latency for 500,000 generated packets of the same session matching at different rule positions using the stateless filtering mechanism. The increasing number of rules leads to a rise in latency and packet loss. This is due the fact that stateless firewalls process the rules in

a linear way. Starting at 400 rules, the data rate of 500 Mbit/s does not leave enough time to process all rules before the next packet arrives. Consequently, the firewall must drop the packets, leading to packet loss. Figure 2 depicts the same measurements with the distinction of using the dynamic stateful filtering mechanism. It shows that an increasing number of rules does not alter the median latency for packet processing. However, the whiskers on the box plots denote maximum latency peaks of 90 µs to 130 µs, introduced by the initial packet of the session. This behavior makes the firewall unsuitable for constant-time packet filtering. To eliminate these latency outliers, we introduce static stateful 5-tuple filtering by inserting the session statically into the session hash table of our modified firewall. Figure 3 illustrates the results of our improved stateful filtering mechanism.

Conclusion

The result of this work is a proof-of-concept firewall implementation with corresponding performance measurements. The evaluation of our improved implementation demonstrates that our firewall achieves packet filtering with low jitter and constant-time behavior while maintaining a high level of security.

References and figures

[1] Thawatchai Chomsiri, Xiangjian He, Priyadarsi Nanda, and Zhiyuan Tan. A Stateful Mechanism for the Tree-Rule Firewall. In *2014 IEEE 13th International Conference on Trust, Security and Privacy in Computing and Communications*, pages 122–129. IEEE, 2014.

[2] Thawatchai Chomsiri, Xiangjian He, Priyadarsi Nanda, and Zhiyuan Tan. Hybrid Tree-Rule Firewall for High Speed Data Transmission. *IEEE Transactions on Cloud Computing*, pages 1237–1249, 2020.

[3] Xiangjian He, Thawatchai Chomsiri, Priyadarsi Nanda, and Zhiyuan Tan. Improving cloud network security using the Tree-Rule firewall. In *Future Generation Computer Systems*, pages 116–126. Future Generation Computer Systems, 2014.

[4] LLC LF Projects. FDio - The Universal Dataplane. https://fd.io/, 2017.

[5] Own representation.

[6] Dmitry Rovniagin and Avishai Wool. The Geometric Efficient Matching Algorithm for Firewalls. *IEEE Transactions on Dependable and Secure Computing*, pages 127–159, 2011.

[7] Markus Schramm. Adaptation of the VPP Firewall for Real-Time Packet Processing in Industrial Environments, 2022.

[8] Z. Trabelsi and S. Zeidan. Enhanced Session Table Architecture for Stateful Firewalls. In *2018 IEEE International Conference on Communications (ICC)*, pages 1–7. IEEE, 2018.

[9] Lukas Wüsteney, Michael Menth, René Hummen, and Tobias Heer. Impact of Packet Filtering on Time-Sensitive Networking Traffic. In *2021 17th IEEE International Conference on Factory Communication Systems (WFCS)*, pages 59–66. IEEE, 2021.

[10] Davison Zvabva, Pavol Zavarsky, Sergey Butakov, and John Luswata. Evaluation of Industrial Firewall Performance Issues in Automation and Control Networks. In *2018 29th Biennial Symposium on Communications (BSC)*, pages 1–5. IEEE, 2018.

Integration von Software-Komponenten in den Software-Stack eines autonom fahrenden Fahrzeugs

Pinar Tuncel **Reiner Marchthaler**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt an der Fakultät Informatik und Informationstechnik, Esslingen

Einleitung

Autonome Fahrzeuge beschäftigen uns seit einigen Jahren und es wurden viele Fortschritte erzielt, wie z.B. das autonome Einparken. Es wird weiter daran gearbeitet, dass sich Fahrzeuge, mobile Roboter und fahrerlose Transportsysteme autonom bewegen können. Im Alltag werden diese Roboter immer häufiger eingesetzt. Nicht nur zu Hause als Staubsaugerroboter, sondern auch im Restaurant sind autonome Roboter zu sehen. In der Gastronomie wird der Roboter für die Essensservierung an den Tisch eingesetzt. Was in der heutigen Zeit nicht mehr als ungewöhnlich angesehen wird. Um die richtige Funktionsweiße dieser Automatisierung dauerhaft zu erfüllen, sind Test während der Entwicklung notwendig.

Ziel der Arbeit

Das Projekt it:movEs beschäftigt sich mit der Entwicklung eines autonom fahrenden Modellfahrzeugs im Maßstab 1:10. Die wesentlichen Kern Domänen im Software-Stack sind z. B. Controlling, Planning, Perception und Modellfahrzeug Lokalisierung, die das autonome Fahren ermöglichen. Hierbei soll die Position des vorhandenen Fahrzeugmodells mit Hilfe eines Kalman-Filters ermittelt werden. Dieser Filter wird in Programmiersprache C++ implementiert. Dabei wird das Robot Operating System (ROS) zur Interprozesskommunikation genutzt. Darüber hinaus sollen Software-in-the-Loop Tests in der bestehenden Simulationsumgebung durchgeführt werden.

Kalman-Filter

Der Kalman-Filter ist ein Algorithmus, der dazu dient, Schätzungen über den Zustand eines Systems zu machen, basierend auf unvollständigen oder verrauschten Daten. Er wurde von Rudolf Emil Kalman in den 1960er Jahren entwickelt. Es kombiniert eine Vorhersage des Systemzustands mit neuen Messungen, um eine optimale Schätzung des tatsächlichen Zustands zu erhalten. In der Praxis wird der Kalman-Filter verwendet, um beispielsweise die genaue Positionsbestimmung eines Fahrzeugs anhand von verrauschten GPS-Daten zu verbessern oder um den Zustand eines physikalischen Systems (z. B. einer Maschine) anhand von Sensordaten zu schätzen und Vorhersagen über die Entwicklung des Systems zu treffen [3].

Roboter Operating System

ROS, dass Robot Operating System, ist ein Softwareframework, das unter der BSD-Lizenz veröffentlicht wurde und zur Steuerung von Roboterkomponenten dient. Dieses Framework ist in der Robotikindustrie weit verbreitet und bietet umfangreiche Funktionen, die es Entwicklern ermöglichen, Robotersysteme effizient zu entwerfen, steuern und betreiben. Die Funktionen von Ros umfassen Hardwareabstraktion, Gerätetreiber, Bibliotheken, Visualisierung, Nachrichtenübermittlung und die Paketverwaltung [4].

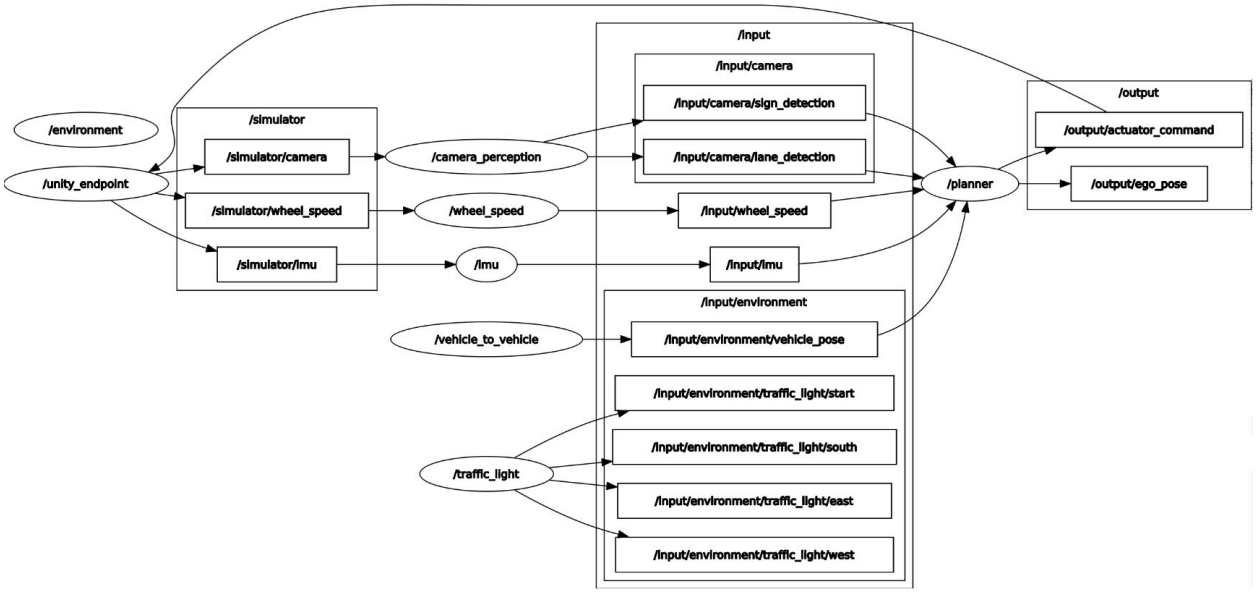


Abb. 1: ROS-Architektur des it:movEs [2]

Vorgehen

Nach der Implementierung der Kalman-Filter, wurden der Quellcode in den bestehenden Software-Stack integriert. Dabei wird die Interprozesskommunikation mit Hilfe des Robot Operating Systems ermöglicht. Es wurde ein weiterer ROS-Knoten erstellt, mit dem die Position des Fahrzeugmodells bestimmt werden kann. Dabei werden die Informationen von /wheel_speed und /imu an den neuen Knoten und von dort an den Simulator /output weitergeleitet. Die Integration der Kalman-Filter in den ROS-C++ Software-Stack wurde durch eine sorgfältige Entwicklung und Anpassung der Schnittstellen zwischen den Modulen erreicht. Dies ermöglichte eine effiziente Datenübertragung und -verarbeitung zwischen den Filtern und anderen Komponenten des Software-Stacks.

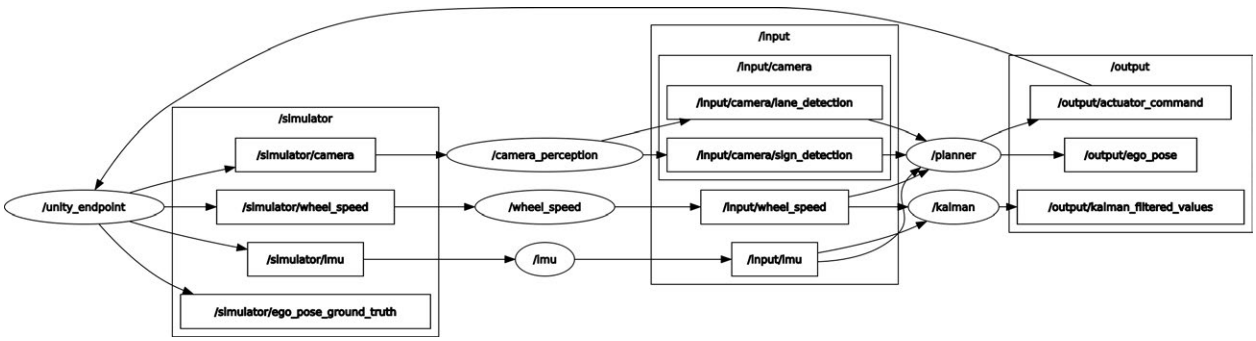


Abb. 2: ROS-Architektur des it:movEs mit Kalman-Node [1]

Ausblick

In den nächsten Schritten wird die Software mit Hilfe des Integrationstests auf Sonderfälle in der Simulationsumgebung getestet, so dass anhand der manipulierten Werte ersichtlich wird, ob die Software wie gewünscht arbeitet. Das autonome Fahren entwickelt sich rapide, mit dieser Entwicklung steigt auch die Komplexität solcher Systeme. Angetrieben wird dieses Fortschreiten im Wesentlichen von dem Wunsch nach Bequemlichkeit. Derzeit gibt es keine vollautonomen Serienfahrzeuge, doch dies wird sich in absehbarer Zeit ändern.

Literatur und Abbildungen

[1] Eigene Darstellung.

[2] Hochschule Esslingen. ROS architecture. https://gitlab.hs-esslingen.de/itmoves-masters/BFMC, 2023.

[3] Reiner Marchthaler. *Kalman-Filter: Einführung in die Zustandsschätzung und ihre Anwendung für eingebettete System*. Springer Vieweg, 2017.

[4] Open Robotics. ROS Documentation. http://wiki.ros.org/, 2022.

Duplikaterkennung von Fehlertickets mithilfe von Machine Learning Methoden

Farbod Vakili **Gabriele Gühring**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma CarByte GmbH, Stuttgart

Einleitung

Die stetig wachsende Komplexität von Fahrzeugsystemen führt zu einer stetig wachsenden Anzahl von Fehlertickets und Supportanfragen. Die CarByte GmbH steht vor der Herausforderung, effiziente Mechanismen für die Verwaltung und Analyse dieser Fehlermeldungen zu entwickeln. In diesem Kontext hat die Duplikatenerkennung von Fehlertickets mittels Machine Learning Methoden einen entscheidenden Stellenwert. Die heutigen Technologien ermöglichen es, riesige Mengen an Ticketdaten zu sammeln, jedoch geht mit dieser Datenflut auch eine erhöhte Wahrscheinlichkeit von Duplikaten einher. Die Identifikation von Duplikaten ist von essenzieller Bedeutung, um Ressourcen effizient zu nutzen und zeitnah auf wiederkehrende Probleme zu reagieren. Hier setzt die vorliegende Masterarbeit an, indem sie die Leistungsfähigkeit von Machine Learning-Methoden für die automatisierte Erkennung von Duplikaten in Fehlertickets untersucht.

Ziel der Arbeit

Die zentrale Zielsetzung dieser Masterarbeit liegt in der Konzeption, Implementierung und Evaluierung eines robusten Machine Learning-Modells zur Erkennung von Duplikaten in Fehlertickets, insbesondere im Kontext der Kooperation zwischen CarByte GmbH und dem Kunden aus der Automobilinstustrie.

Struktur und Vorgehensweise

Die Struktur dieser Arbeit gliedert sich in mehrere aufeinander aufbauende Schritte. Zunächst erfolgt die umfassende Datenbeschaffung und -verarbeitung, gefolgt von der Generierung verschiedener Embeddings für Wörter und Sätze aus den Ticketbeschreibungen. Im Anschluss wird eine sorgfältige Auswahl von Distanzmetriken vorgenommen, um die semantische Ähnlichkeit zwischen den Embeddings effektiv zu messen. Die eigentliche Ähnlichkeitsberechnung zwischen den Fehlertickets wird daraufhin durchgeführt, wobei verschiedene Schwellenwerte für die Identifikation von Duplikaten evaluiert werden.
Darüber hinaus wird ein siamesisches Netzwerk mit Similarity Learning konzipiert und implementiert, um die Fähigkeit des Modells zur Duplikatenerkennung weiter zu verfeinern. Die Training- und Evaluierungsphase erfolgt unter Berücksichtigung von Trainings- und Testdatensätzen, wobei eine kontinuierliche Feinabstimmung und Optimierung der Modelleigenschaften erfolgt.

Embeddings

Die Evaluierung unterschiedlicher Embedding-Technologien bildet einen entscheidenden Schritt in der Entwicklung des Modells. Dabei liegt der Fokus auf die Auswahl optimaler Word- und Sentence Embeddings, um die semantische Repräsentation der Ticketbeschreibungen zu verbessern.
In Bezug auf Word Embeddings werden verschiedene moderne Modelle einer Evaluierung unterzogen. Hierzu gehören Embeddings wie Elmo, xLM-r, GloVe, Fasttext und BERT. Elmo, basierend auf bi-direktionalen LSTM-Architekturen, bietet eine kontextsensitive Wortrepräsentation [5] . xLM-r und BERT ermöglicen die Integration mehrerer Sprachen und präsentieren sich als vielversprechende Modelle für die Duplikatenerkennung in einem multilingualen Umfeld [3] [1]. GloVe, ein statistisches Modell [4], und Fasttext, das auf Subwort-Informationen basiert [8], werden aufgrund ihrer nicht allzu rechenintesiven Algorithmen ebenfalls untersucht.
Hinsichtlich der Sentence Embeddings werden mehrere Ansätze evaluiert, darunter der Multilingual Universal Sentence Encoder (MUSE) und LASER (Language-Agnostic Sentence Encoder Representations), die sich durch ihre sprachunabhängige Fähigkeit auszeichnen [9] [6]. Das Sentence-Bert-Modell, das auf der siamesischen Netzwerkarchitektur basiert [7], wird ebenfalls in Betracht gezogen.

Die Auswahl der optimalen Embeddings ist von entscheidender Bedeutung für die spätere Leistungsfähigkeit des Modells. Daher werden diese verschiedenen Ansätze sorgfältig evaluiert, um sicherzustellen, dass die gewählten Embeddings eine präzise und aussagekräftige semantische Repräsentation der Fehlertickets ermöglichen.

Distanzmetriken

Um die semantische Ähnlichkeit zwischen Fehlerticket-Embeddings zu erfassen, werden verschiedene Distanzmetriken in Betracht gezogen. Die Jaccard-Ähnlichkeit eignet sich zur Analyse von Gemeinsamkeiten in Wortmengen, während die euklidische Distanz den direkten Raumabstand misst. Die Cosinus-Ähnlichkeit ermöglicht die Bewertung des Winkels zwischen Vektoren und ist besonders nützlich für die natürliche Sprachverarbeitung. Die Hamming-Distanz fokussiert sich auf Unterschiede in binären Zeichenketten und erfasst Strukturunterschiede in Ticketbeschreibungen.

Model

Ein zentrales Element in der Entwicklung eines effizienten Duplikatenerkennungsmodells für Fehlertickets ist die Integration eines siamesischen neuronalen Netzwerks mit Similarity Learning. Dieser Ansatz ermöglicht es dem Modell, die semantische Ähnlichkeit zwischen den Ticket-Embeddings präzise zu erlernen. Das siamesische Netzwerk besteht aus zwei identischen Zweigen, die gemeinsame Gewichtungen nutzen, um spezifische Muster zwischen Duplikaten zu erfassen. Das Similarity Learning wird genutzt, um das Netzwerk darauf zu trainieren, die semantische Ähnlichkeit zu verstärken und Duplikate präzise zu identifizieren. Durch Training und Feinabstimmung wird das Modell auf hohe Generalisierung getrimmt, und die Evaluierung erfolgt anhand von Testdaten unter Verwendung verschiedener Leistungsmetriken wie Genauigkeit und Präzision.

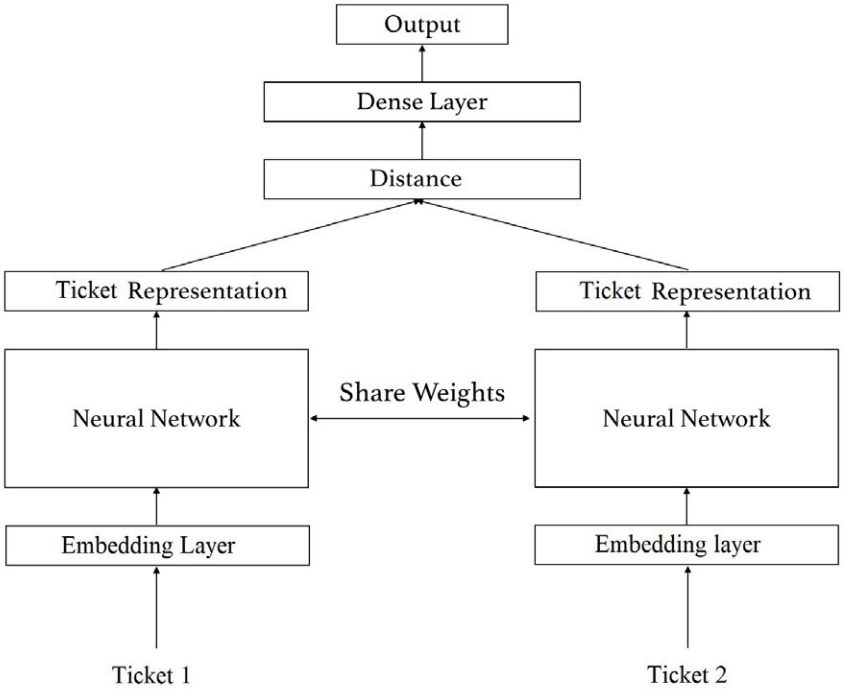


Abb. 1: Konzept des siamesischen Netzwerks [2]

Literatur und Abbildungen

[1] Conneau Alexis, Khandelwal Kartikay, Goyal Naman, Chaudhary Vishrav, Wenzek Guillaume, Guzman Francisco, Grave Edouard, Ott Myle, Zettlemoyer Luke, and Stoyanov Veselin. *Unsupervised Cross-lingual Representation Learning at Scale*. https://doi.org/10.48550/arXiv.1911.02116, 2020.

[2] Eigene Darstellung.

[3] Devlin Jacob, Chang Ming-Wei, Lee Kenton, and Toutanova Kristina. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *https://doi.org/10.48550/arXiv.1810.04805*, 2019.

[4] Pennington Jeffrey, Socher Richard, and Manning. Christopher. GloVe: Global Vectors for Word Representation. *10.3115/v1/D14-1162*, 2014.

[5] Petersy Matthew, Neumanny Mark, Iyyery Mohit, Gardnery Matt, Clark Christopher, Lee Kenton, and Zettlemoyer Luke. Deep contextualized word representations. *https://doi.org/10.48550/arXiv.1802.05365*, 2018.

[6] Tiyajamorn Nattapong, Kajiwara Tomoyuki, Arase Yuki, and Onizuka Makoto. Language-agnostic Representation from Multilingual Sentence Encoders for Cross-lingual Similarity Estimation. *Association for Computational Linguistics*, 2021.

[7] Reimers Nils and Gurevych Iryna. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *Empirical Methods in Natural Language Processing*, 2019.

[8] Bojanowski Piotr, Grave Edouard, Joulin Armand, and Mikolov Tomas. Enriching Word Vectors with Subword Information. *https://doi.org/10.48550/arXiv.1607.04606*, 2017.

[9] Yang Yinfei, Cer Daniel, Ahmad Amin, Guo Mandy, Law Jax, Constant Noah, Abrego Gustavo, Yuan Steve, Tar Chris, Sung Yun-hsuan, Strope Brian, and Kurzweil Ray. Multilingual Universal Sentence Encoder for Semantic Retrieval. *Association for Computational Linguistics*, 2020.

Entwurf und Implementierung eines Testability Frameworks für KeylessGo-Systeme

Matthias Warttmann **Markus Enzweiler**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma Vector Informatik GmbH, Stuttgart

Einleitung

Ein Keyless-Go System bei Automobilen findet immer mehr Anklang bei Gesellschaft und Autoherstellern. Jedoch gibt es immer noch einige Sicherheitslücken, die das eigene Fahrzeug anfällig für Diebstähle machen. Fahrzeuge, die Keyless-Go unterstützen, können anhand von Signalverstärkern geöffnet und sogar gefahren werden [4]. Um neuen Wind in die Technologie zu bringen, haben sich eine Reihe von namhaftem Auto-, Halbleiter- und Handyherstellern zusammengeschlossen, mit der Absicht, einen einheitlichen Standard zu erstellen. Daraus resultierte das Car Connectivity Consortium (CCC), welches einen neuartigen Keyless-Go Standard spezifiziert. Zentraler Punkt des CCC ist, den Autoschlüssel zu digitalisieren. Mit einer komplexen digitalen Architektur ist es möglich, den Schlüssel dezentral zu speichern und ihn auf eine App auf dem Smartphone zu spiegeln. Es ist möglich, den Schlüssel zu einem Auto mit anderen zu teilen, natürlich mit einstellbaren Zugriffsrechten und technischen Möglichkeiten des Fahrzeuges. Durch Annähern zum Fahrzeug wird ohne Zutun des Anwenders drahtlos das Auto geöffnet [2]. Was den DigitalKey Standard von CCC von einem physikalischen Schlüssel unterscheidet, ist der Einsatz der Ultra-Breitband Technologie. Diese ermöglicht es Distanzen Zentimeter genau zu bestimmen, und somit die Relay Attacke zu unterbinden. Nur wenn eine Mindestdistanz gegeben ist, kann die Kommunikation beginnen. BMW hat diesen Standard bereits in die neuste Generation ihrer Autos implementiert [1].

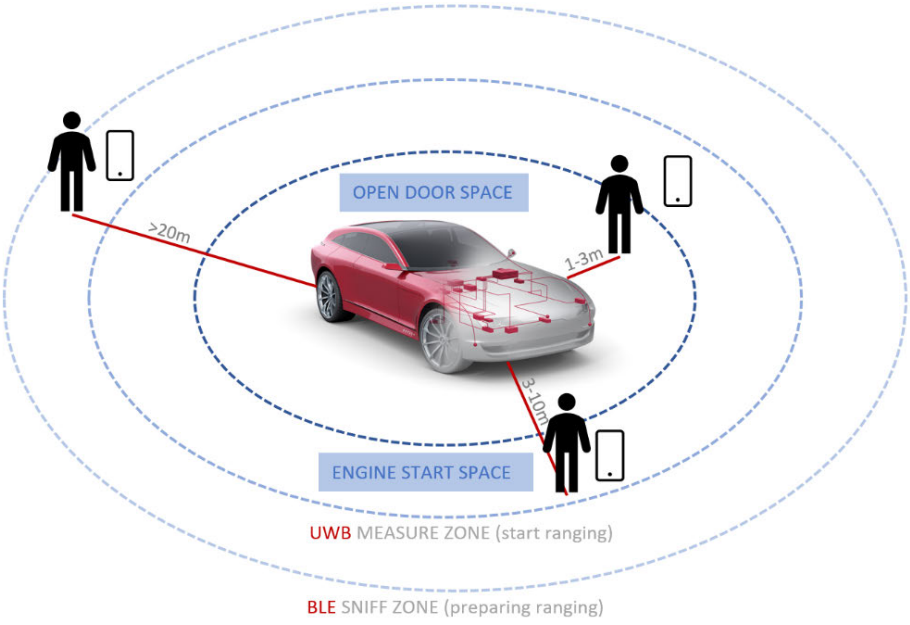


Abb. 1: Benötigte Entfernungen für Kommunikationsaufbau [3]

Zielsetzung

Ziel der Arbeit ist es, sowohl eine grundlegende Kommunikation zwischen Smartphone und Auto zu entwerfen als auch das verständliche Herunterbrechen der für die Arbeit relevanten Teile des Standards. Entworfen wird die Kommunikation auf dem VH4110, ein RaspberryPi, im Zusammenspiel mit der von Vector entwickelten Test und -Simulationssoftware CANoe. Es sollen die CCC eigenen Nachrichten und Kommunikationsabläufe modelliert und umgesetzt werden. Final soll innerhalb der Test und -Simulationssoftware CANoe eine Kommunikation aufgebaut werden, Nachrichten ausgetauscht und klassifiziert werden. Die Kommunikation soll über Bluetooth Low Energy, eine leichtgewichtige Variante von Bluetooth, über den VH4110 RaspberryPi erfolgen.

Entwurf und Implementation

In Abbildung 2 ist die zu implementierende Kommunikation dargestellt.

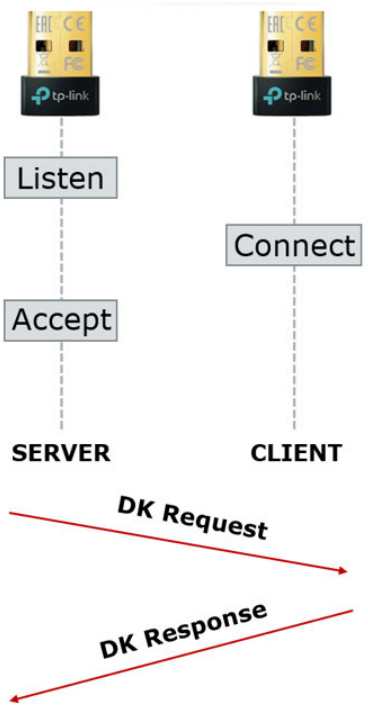


Abb. 2: Kommunikationsablauf BLE Sticks [3]

Zu Beginn wird der CCC-Standard auf die für die Anwendung nötigen Infos zusammengefasst. Anschließend wird mit der Modellierung der zentralen Kommunikations-Nachrichten begonnen. Diese nennen sich DK-Nachrichten. Zur Modellierung werden diese in das Test und -Simulationstool CANoe implementiert. Die DK-Nachrichten geben Auskunft über Zustand der Kommunikation sowie der Kommunikationsteilnehmer. Diese Nachrichten werden hauptsächlich über Bluetooth Low Energy übertragen. Hierzu wird auf dem VH4110 RaspberryPi eine entsprechende Bluetooth Low Energy Schnittstelle erstellt, die auf die DK-Nachrichten zugeschnitten ist. Smartphone und Auto sollen in einer Instanz auf dem VH4110 RaspberryPi simuliert werden können. Hierzu werden zwei Bluetooth Low Energy Adapter genutzt, die die jeweiligen Kommunikationspartner simulieren sollen. Das Interface auf dem Pi muss entsprechen darauf angepasst werden. Das Interface auf dem Pi ist in der Programmiersprache Python verfasst. Eine spezielle Bibliothek, in C geschrieben, wird verwendet, um die Funktionalität von Bluetooth in dem Python Interface benutzen zu können. Um zwei Kommunikationsteilnehmer simulieren zu können muss auch das CANoe Interface, genutzt, um mit dem Pi zu kommunizieren, angepasst werden. Nach der Implementation der Interfaces ist nun ein Verbindungsaufbau zwischen den beiden Kommunikationspartnern möglich. Nun sollen über die aufgebaute Verbindung die vorher deklarierten DK-Nachrichten gesendet werden. An der Gegenstelle werden diese Empfangen und klassifiziert. Anschließend wird die passende Antwort zurückgesendet. Um Anwendern des Systems einige Testfälle zu präsentieren, wurden im Rahmen der Arbeit einige Tests formuliert, und auch durchgeführt. Die Arbeit erfüllt alle angeforderten Kriterien und ist somit fertiggestellt.

Ausblick

Der im Rahmen der Arbeit erarbeitete Kommunikationsprototyp dient als Basis für weitere Entwicklungen in dieser Richtung. Mit dieser Grundlage ist das Implementieren des restlichen Standards, dessen Umfang viel größer ist als in dieser Arbeit beschrieben, um vieles erleichtert. Auf Anfrage mehrerer Vector Kunden, hat Vector nun ein Prototyp eines immer größer werdenden Standards vorzuweisen. Somit können Hersteller, die den Standard bei sich implementieren, ihre Produkte mit Vector Hardware und Software optimal testen und ausweiten.

Literatur und Abbildungen

[1] BMW AG. BMW Digital Key. https://www.bmw.de/de/topics/service-zubehoer/bmw-connecteddrive/digital-key.html, 2023.

[2] Car Connectivity. Car Connectivity Consortium CCC DigitalKey. https://carconnectivity.org/digital-key/, 2023.

[3] Eigene Darstellung.

[4] Dan Goodin. There's a new form of keyless car theft that works in under 2 minutes. https://arstechnica.com/information-technology/2023/04/crooks-are-stealing-cars-using-previously-unknown-keyless-can-injection-attacks/, 2023.

Eine Analyse der vielfältigen Einflussfaktoren auf die Entwicklung eines Requirement Engineering in der Softwareentwicklung

Tim Wasberg **Anke Bez**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt an der Fakultät Informatik und Informationstechnik, Esslingen

Zielsetzung:

Das Requirement Engineering ist ein systematischer und disziplinierter Ansatz zur Spezifikation und zum Management von Anforderungen an Software oder ein Produkt, mit dem Fokus, die Wünsche und Bedürfnisse der Stakeholder zu verstehen und das Risiko einer nicht bedürfnisorientierten Produktentwicklung zu minimieren [4]. Im Requirement Engineering-Prozess werden Anforderungen in einem iterativen Prozess durch Ermittlung, Beschreibung und Analyse in Zusammenarbeit mit Stakeholdern kontinuierlich bis zur Abnahme verfeinert. Der Prozess, visualisiert in Abbildung 1, zeichnet sich durch Überlappung und Rückkopplung zwischen den Phasen aus, ermöglicht Anpassungen und verbessert Lösungen im Projekt gemäß den Bedürfnissen aller Stakeholder [2]. Im Bereich des Requirement Engineering (RE) in der Softwareentwicklung stehen Entwickler vor anspruchsvollen Herausforderungen, hauptsächlich aufgrund der fehlenden Standardisierung des Prozesses. Es besteht Unsicherheit darüber, wie das Requirement Engineering gestaltet werden sollte. Diese Unsicherheit betont die Bedeutung einer gründlichen Analyse, um effektive Lösungsansätze für den Aufbau eines soliden Requirement Engineering zu entwickeln. Daher widmet sich die vorliegende Arbeit der Darstellung verschiedener Dimensionen und Einflussfaktoren, die die Gestaltung des Requirement Engineering beeinflussen.

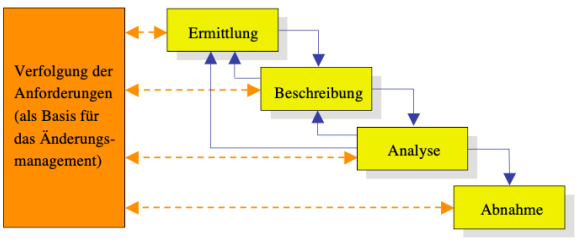


Abb. 1: „RE life cycle“ [2]

Zusammenarbeit und Verhalten

Die Ausgestaltung des Requirement Engineerings wird wesentlich von der gewählten Art der Zusammenarbeit beeinflusst, wobei die Zeitachse und das Verhalten der Beteiligten entscheidende Dimensionen sind 2. Insbesondere bei kurzfristiger und unkooperativer Zusammenarbeit empfiehlt sich ein erhöhter Aufwand in der Bedarfsanalyse, um rechtliche Absicherung und klare Verantwortlichkeiten sicherzustellen. Bei unkooperativer Kooperation sind diese Aspekte besonders wichtig. Im Idealfall ermöglicht langfristige, kooperative Zusammenarbeit eine flexible Anpassung der Aufwände, insbesondere bei agilen Methoden. Hingegen erfordert kurzfristige und unkooperative Kooperation verstärkten Aufwand in rechtlicher Absicherung und Change-Management.

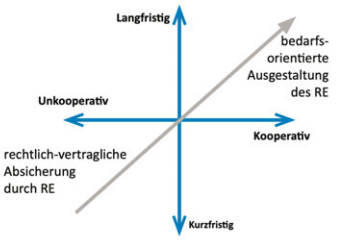


Abb. 2: Dimensionen der Zusammenarbeit [4]

Innovationsgrad

Das Maß an Innovation eines Produkts hat einen direkten Einfluss auf das Requirements Engineering (RE), da es sowohl den zeitlichen Druck im RE beeinflusst als auch die Fähigkeit, erfolgreich Wissen für die Entwicklung zu generieren. Ein hoher Innovationsgrad bedeutet oft sich schneller ändernde Marktanforderungen, was den Zeitdruck auf das RE erhöht, um rasch reagieren zu können. Gleichzeitig erfordert innovative

Entwicklung zusätzliches Wissen, um anspruchsvolle Anforderungen zu verstehen und umzusetzen. Daher ist die Bewältigung innovativer Herausforderungen im RE nicht nur zeitkritisch, sondern benötigt auch eine effektive Wissensgenerierung, um die innovativen Elemente erfolgreich in die Softwareentwicklung zu integrieren [4].

Örtliche Verteilung

Die Bedeutung unterschiedlicher Fachgebiete und die räumliche Verteilung der Teammitglieder sind zusätzliche Faktoren, die einen entscheidenden Einfluss auf die Effizienz des Requirements Engineering (RE) ausüben. Die Gewichtung der Disziplinen bezieht sich dabei darauf, wie stark verschiedene Fachbereiche in den RE-Prozess eingebunden sind. Eine ausgewogene Berücksichtigung verschiedener Disziplinen ist von entscheidender Bedeutung für eine umfassende und qualitativ hochwertige Anforderungsanalyse. Gleichzeitig kann die geografische Verteilung der Teammitglieder die Kommunikation und Koordination im RE beeinflussen, was wiederum die Effizienz des gesamten Entwicklungsprozesses beeinträchtigen kann [4].

Dokumentation

Die Auswahl der Dokumentationstechnik und -menge im Requirements Engineering (RE) ist stark von verschiedenen Faktoren abhängig, darunter Regelwerke, Zielgruppen und die Notwendigkeit, Informationen langfristig zugänglich zu halten. Die gewählte Technik und der Umfang der Dokumentation beeinflussen die Klarheit und Verständlichkeit der festgelegten Anforderungen. Regelwerke und Normen können vorschreiben, welche Dokumentationsstandards eingehalten werden müssen. Die Zielgruppen, wie Entwickler oder Kunden, beeinflussen ebenfalls, wie die Informationen präsentiert werden sollten. Langfristige Zugänglichkeit ist entscheidend, um die Entwicklungsentscheidungen nachvollziehbar zu machen und eine kontinuierliche Verbesserung im Laufe der Zeit zu ermöglichen. Dabei werden zwei Hauptgruppen von Techniken vorgestellt: Strukturtechniken für statische Aspekte und Verhaltenstechniken für dynamische Aspekte eines Systems. Die Auswahl einer passenden Technik wird durch fachliche, organisatorische und menschliche Merkmale bestimmt. Dazu gehören Aspekte wie Detaillierungsebene, Konsistenz, Vollständigkeit, Verfolgbarkeit, Lebensdauer der Produkte, Komplexität des Problems, Eindeutigkeit, Verständlichkeit und Akzeptanz. Diese Merkmale sollten bei der Entscheidung für eine Dokumentationstechnik berücksichtigt werden [3].

Qualitätsanforderungen

Die Qualitätsanforderungen an die gestellten Anforderungen beeinflussen den Arbeitsaufwand im Requirements Engineering (RE), wobei höhere Qualitätsanforderungen eine umfassendere Analyse und Dokumentation erfordern [4]. Qualitätssicherung stellt nicht nur eine Testphase am Schluss des Projektes dar, sondern eine Grundhaltung. Sie beinhaltet kontinuierliche Aktivitäten, die in alle Tätigkeiten im Softwareentwicklungs- bzw. Beschaffungsprozess eingebunden werden müssen. [5].

Vermittlungsprozess

Der Vermittlungsprozess im Requirement Engineering (RE) variiert je nach den spezifischen Anforderungen und den beteiligten Personen. Die Kommunikations- und Vermittlungsmethoden im RE sind vielfältig und reichen von informellen Gesprächen bis zu detaillierten Dokumentationen. Diese Flexibilität ermöglicht es, die am besten geeignete Kommunikationsform je nach der Komplexität der Anforderungen und den Bedürfnissen der Stakeholder zu wählen. Informelle Gespräche schaffen ein tieferes Verständnis, während detaillierte Dokumentationen zur präzisen Festlegung von Anforderungen beitragen [4].

Ausblick

Die komplexen Anforderungen im Requirement Engineering erfordern innovative Ansätze und eine gründliche Analyse. Eine verstärkte Integration von Qualitätsanforderungen, die Anpassung an die Teamverteilung und die Auswahl geeigneter Dokumentationstechniken sind entscheidend. Die Flexibilität im Vermittlungsprozess bleibt ein Schlüssel zum Erfolg. "Vor allem die Einflüsse von scheinbar entfernten Projektbeteiligten, so genannter Stakeholder, müssen (...) Berücksichtigung finden"[1]. Die Bewältigung dieser Herausforderungen wird die Entwicklung effektiver Lösungen ermöglichen.

Literatur und Abbildungen

[1] David Krips. *Stakeholdermanagement*. Springer Berlin Heidelberg, 2017.

[2] Helmuth A. Partsch. *Requirements-Engineering systematisch: Modellbildung für softwaregestützte Systeme*. Springer Berlin Heidelberg, 2010.

[3] Chris Rupp. *Systemanalyse kompakt*. Springer Berlin Heidelberg, 2013.

[4] Chris Rupp. *Requirements-Engineering und -Management: das Handbuch für Anforderungen in jeder Situation*. Hanser, 7 edition, 2021.

[5] Hansruedi Tremp. *Agile objektorientierte Anforderungsanalyse: Planen – Ermitteln – Analysieren – Modellieren – Dokumentieren – Prüfen*. Springer Fachmedien Wiesbaden, 2022.

Flutter vs. Kotlin-Multiplatform: Untersuchung der Eignung der Plattformen für den Relaunch einer mobilen App

Ayleen Weiss **Andreas Rößler**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma cluetec GmbH, Karlsruhe

Einleitung

Seit den 2010er Jahren hat sich die Mobile-First-Strategie in der Softwareentwicklung verbreitet. Heute sind in diesem Bereich nur noch die Betriebssysteme Android und iOS von Interesse. Die Notwendigkeit, für beide Systeme separate Anwendungen zu entwickeln, erhöht den Zeitaufwand und die Wartungskosten [5]. Eine Lösung bieten plattformübergreifende Frameworks, die es erlauben, Code einmal zu schreiben und auf beiden Plattformen einzusetzen. Darüber hinaus bieten sie inzwischen häufig die Möglichkeit, aus derselben Codebasis auch Anwendungen für Desktop-Betriebssysteme und das Web zu generieren.

Zielsetzung der Arbeit

Diese Arbeit soll die Fragestellung klären, wie sich die Eignung eines Cross-Plattform-Entwicklungs-Frameworks für die Neuentwicklung einer bestehenden Anwendung struktuiert evaluieren lässt. Der Fokus liegt dabei auf den beiden Plattformen Flutter und Kotlin Multiplatform. Das Ergebnis soll eine Entscheidungsmatrix sein, welche sich auch auf andere Frameworks und Organisationen anwenden lässt.

Technische Konzepte

Frühe Ansätze zur Cross-Plattform-Entwicklung, wie sie von Xanthopoulos und Xinogalos [5] beschrieben wurden, lassen sich in vier Konzepte unterteilen:

- Web Apps (Anwendungen die in einem Browserfenster laufen und dynamisch aus dem Web geladen werden)
- Hybrid Apps (Web Apps in Containern, die eine Distribution über die App Stores erlauben)
- Interpretierte Apps (Code wird auf der Zielplattform durch einen Interpreter verarbeitet)
- Generierte Apps (Für jede der Zielplattformen wird nativer Code erzeugt)

Nur der letztgenannte Ansatz kann sowohl einen nativen Eindruck bei der Bedienung der UI vermitteln als auch den vollen Zugriff auf die Hardware der Geräte und den vollen Datenzugriff ermöglichen [5]. Flutter und Kotlin Multiplatform fallen in diese Kategorie.
Darüber hinaus kann zwischen Cross-Plattform und Multi-Plattform-Ansätzen unterschieden werden, was sich in der Menge des gemeinsam genutzten Codes ausdrückt. Bei der Cross-Plattform-Entwicklung wird die UI ebenfalls aus einer gemeinsamen Codequelle generiert. Ein natives UI-Erlebnis kann hier nur eingeschränkt erreicht werden. Beim Multi-Plattform-Konzept wird die UI in jedem Fall separat programmiert. Für Komponenten wie Persistenz kann es je nach Anwendung sinnvoll sein, diese gemeinsam oder getrennt zu erstellen, siehe Abbildung 1.

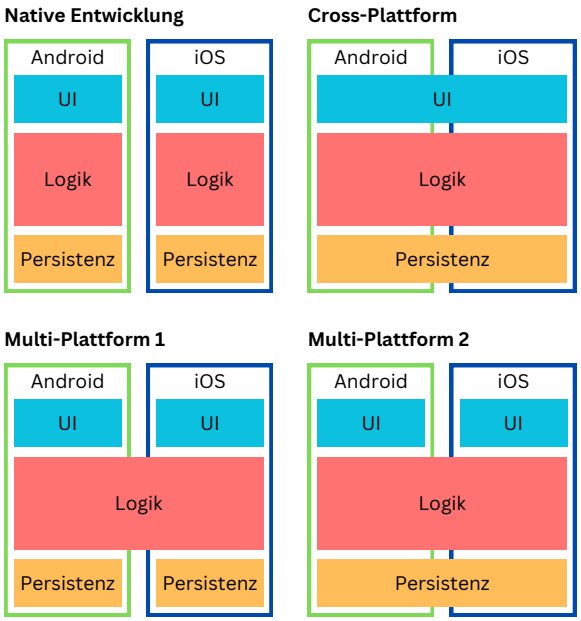


Abb. 1: Vergleich Entwicklungsansätze [1]

Kotlin Multiplatform

Kotlin Multiplatform ermöglicht die gemeinsame Nutzung von Code auf verschiedenen Plattformen wie Desktops, Servern und mobilen Geräten. Kotlin selbst ist eine statisch typisierte Programmiersprache, die seit 2011 von JetBrains entwickelt wird und sowohl objektorientierte als auch funktionale Programmierkonzepte unterstützt. Kotlin wird zunehmend in der nativen Android-Entwicklung eingesetzt, insbesondere nachdem Google 2019 eine Kotlin-first-Strategie für Android angekündigt hat [4] (Kap. 1.2.1).
Kotlin Multiplatform ermöglicht es Entwicklern, die Anwendungslogik zwischen Android und iOS zu teilen, während die Benutzeroberflächen nativ für jede Plattform erstellt werden. Die gemeinsame Logik umfasst Kernfunktionen wie Datenverarbeitung und -validierung, Berechnungen, Kommunikation zwischen Systemen (z.B. HTTP und API-Aufrufe) und State-Management. Der Kotlin-Code wird für Android in JVM-Bytecode und für iOS mittels LLVM in nativen Maschinencode übersetzt.

Flutter

Flutter unterstützt neben mobilen Plattformen auch Desktop-, Web- und einige Embedded-Plattformen. Die zugrundeliegende Programmiersprache Dart wurde 2011 vorgestellt. Dart ist eine multiparadigmatische und optional dynamisch typisierte Sprache.
Das Herzstück von Flutter ist die Flutter Engine. Sie stellt die Low-Level-Implementierung der Kern-API von Flutter zur Verfügung, einschließlich Grafik-Rendering, Textlayout, Datei- und Netzwerk-I/O, Unterstützung für Barrierefreiheit, Plugin-Architektur sowie eine Dart Laufzeitumgebung und Kompilierungswerkzeuge. Für jede Zielarchitektur gibt es einen plattformspezifischen Embedder. Entwickler interagieren mit dem Framework, das Basisfunktionen, Rendering-Layer, Widgets und Design-Bibliotheken enthält [2].

Methoden

In einer ausführlichen Literaturrecherche wird der Stand der wissenschaftlichen Bearbeitung des Themas erfasst. Anschließend werden unternehmensinterne Experten befragt, um die für den spezifischen Fall und für die Organisation wichtigen Kriterien zu identifizieren. Diese Interviews werden mittels einer inhaltlich strukturierten qualitativen Inhaltsanalyse nach Kuckartz [3] (S.129ff) ausgewertet. Die Ergebnisse werden durch zusätzliche quantitative Daten untermauert, die mittels eines Fragebogens erhoben werden.
Anschließend werden mit den ausgewählten Frameworks Test-Applikationen erstellt, die prototypisch jene Features enthalten, die für eine Beurteilung der Eignung der Frameworks im konkreten Fall notwendig sind.
Alle Ergebnisse fließen in eine Entscheidungsmatrix ein. Der gesamte Prozess ist in Abbildung 2 dargestellt.

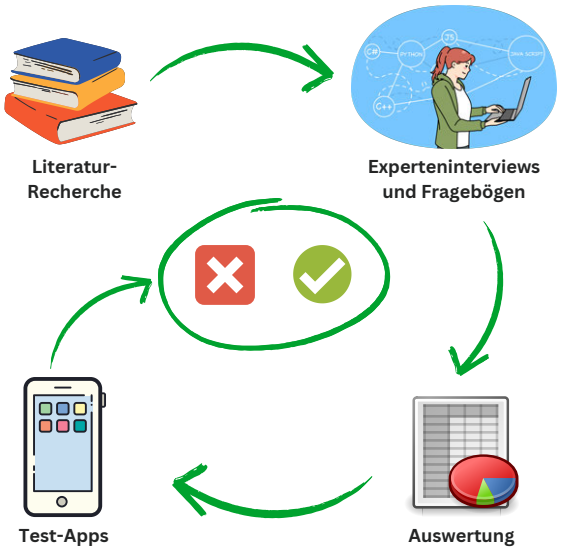


Abb. 2: Prozess zur Entscheidungsfindung [1]

Ausblick

Die Expertenbefragung hat gezeigt, dass die Integration moderner Technologien und Programmiersprachen entscheidend ist, um zukunftsfähig zu bleiben und schnell auf neue Funktionalitäten reagieren zu können. Großer Wert wird auf eine benutzerfreundliche, intuitive und performante UI gelegt, die die Komplexität der Anwendung für den Anwender verbirgt. Um eine native Benutzererfahrung zu erreichen, kann die Programmierung separater UI-Schichten akzeptiert werden, während die Anwendungslogik nur einmal entwickelt werden sollte, um den Test- und Wartungsaufwand zu minimieren.
Softwareentwickler legen großen Wert auf Effizienz, daher werden moderne Technologien, eine umfangreiche Standardbibliothek, gut gepflegte Pakete, effizientes Toolmanagement und gute Dokumentation bevorzugt. Probleme beim Dependency Management und bei Framework-Updates werden als kritisch angesehen.
Bei der Auswahl eines Frameworks ist auch die langfristige Perspektive entscheidend. Es sollte etabliert sein, kontinuierlich verbessert werden und bewährte Buildsysteme sowie native Systemfunktionen unterstützen.
Diese Ergebnisse fließen in Form einer firmenspezifischen Gewichtung in die Entscheidungsmatrix ein.
Die Erfahrungen aus dem Entwicklungsprozess der Test-Applikationen spiegeln sich in der Entscheidungsmatrix in Form von vergebenen Punkten wider.

Literatur und Abbildungen

[1] Eigene Darstellung.

[2] Flutter Docs. Flutter architectural overview. https://docs.flutter.dev/resources/architectural-overview, 2023.

[3] Udo Kuckartz. *Qualitative Inhaltsanalyse. Methoden, Praxis, Computerunterstützung : Grundlagentexte Methoden*. Beltz Juventa, 5 edition, 2022.

[4] Thomas Künneth. *Android 11 : das Praxisbuch für App-Entwickler*. Rheinwerk Verlag, 6 edition, 2021.

[5] Spyridon Xanthopoulos and Stelios Xinogalos. A comparative analysis of cross-platform development approaches for mobile applications. *Proceedings of the 6th Balkan Conference in Informatics*, 2013.

Analyse und Optimierung der Energieeffizienz bei Camunda Workern

Jan Wittrowski **Jörg Nitzsche**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma envite consulting GmbH, Stuttgart

Motivation

IT-Systeme so zu entwickeln und zu konfigurieren, dass sie für ihren Einsatzzweck möglichst wenig Energie verbrauchen, ist eine wichtige Aufgabe, der in der jüngeren IT-Vergangenheit nicht immer die notwendige Aufmerksamkeit geschenkt wurde. Dabei liegen die Vorteile effizienter Software in doppelter Hinsicht auf der Hand. Zum einen können dadurch die laufenden Kosten für Unternehmen gesenkt werden, zum anderen (und deutlich dringlicher) stellt eine solche Optimierung einen elementaren Beitrag zur gesamtgesellschaftlichen Transformation dar, die notwendig ist, um der Klimakrise Einhalt zu gebieten.

Ziel der Arbeit

Im Sinne dieses Optimierungsgedanken, fasst diese Arbeit eine wichtige Systemkomponente der Workflow-Engine Camunda 8 in den Blick: den Job Worker. Mit Hilfe von Messungen innerhalb der für Camunda typischen verteilten Systemarchitektur wird der Energieverbrauch analysiert. Es sollen klare Zusammenhänge aufgezeigt werden, welche Konfigurationen und Implementierungen Einfluss auf den Energieverbrauch haben. Entscheidungen bezüglich dieser Parameter müssen dann hinsichtlich ihrer Bedeutung für die Performance oder andere wichtige Systemparameter analysiert werden. Am Ende sollen Best Practices erarbeitet werden, die Energieeffizienz als relevante Metrik berücksichtigen.

BPMN und Camunda 8

BPMN 2.0 ist eine standardisierte Notation zur Modellierung von Geschäftsprozessen. Der grafische Ansatz vereinfacht die Kommunikation zwischen technischem Personal (den EntwicklerInnen automatisierter Prozesse) und fachlichen ExpertInnen. Die Standardisierung ermöglicht es auch, Prozesse maschinell zu lesen und zu automatisieren [3]. Camunda 8 ist eine Softwarelösung, die diese Prozessautomatisierung ermöglicht. Prozesse und Diagramme können erstellt und verwaltet werden, eine Workflow-Engine, Zeebe genannt, orchestriert dann im Hintergrund die Ausführung. Die automatisierbaren Teile des Prozesses, sogenannte Service Tasks, werden typischerweise als Microservices implementiert. Die konkrete Ausführung dieser Microservices wird als Job bezeichnet und von Job Workern übernommen.

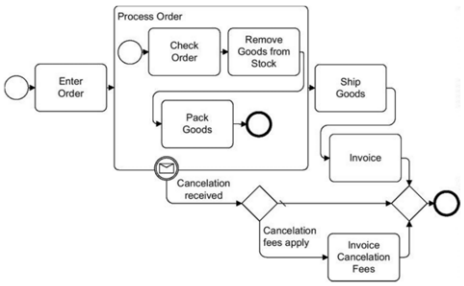


Abb. 1: BPMN-Beispielprozess einer Bestellung [1]

Camunda 8: Job Worker

Job Worker sind die Komponenten, deren Effizienz und Energieverbrauch Gegenstand der Untersuchung sind. Sie werden unabhängig von der Workflow-Engine Zeebe implementiert und übernehmen klassischerweise REST-Anfragen an externe Dienste, Datenbankzugriffe oder Datenformatierungen. Die Technologien oder Implementierungen selbst sind dabei nicht von Camunda vorgegeben [6]. Ein Schwerpunkt dieser Untersuchung wird auf der Technologie liegen, die in den Projekten von envite consulting, dem Auftraggeber dieser Arbeit, überwiegend zum Einsatz kommt: Java bzw. Spring.

Kubernetes-based Efficient Power Level Exporter (Kepler)

Kepler ist ein Tool, das den Energieverbrauch in Kubernetes-Umgebungen misst. Es ermöglicht die

isolierte Betrachtung einzelner Pods. Je nach Hardware-Architektur wird hauptsächlich das Running Average Power Limit (RAPL) von Intel für RAM und CPU sowie das Advanced Configuration and Power Interface (ACPI) für den Gesamtverbrauch eingesetzt [2]. Kepler implementiert die erforderlichen Kernel-Zugriffe mittels eBPF (extended Berkeley Packet Filter). eBPF ist eine Linux-Kernel-Technologie, die es Entwicklern ermöglicht, Code dynamisch und sicher in den Kernel zu laden. [5]

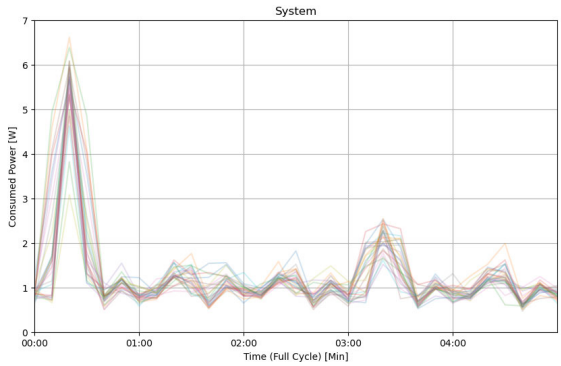


Abb. 2: Energieverbrauch beim Starten eines Camunda-Worker-Pods [4]

Prometheus & Grafana

Die von Kepler exportierten Verbrauchswerte können mit Prometheus ausgelesen werden. Prometheus verfügt über eine eigene Query Language, PromQL. Die gesammelten Daten werden von Grafana, einer Dashboarding-Lösung visualisiert und ggf. in Tabellenform exportiert.

Konzept

Um zu ermitteln, welche Worker-Konfigurationen sich für welche Einsatzzwecke eignen, muss zunächst bestimmt werden, wie viel Energie Worker in den verschiedenen Phasen ihres Lebenszyklus verbrauchen. Dabei sind insbesondere das Starten und Stoppen sowie der Verbrauch im Leerlauf wichtige Größen. Darüber hinaus gibt es drei Konfigurationsparameter, deren Einfluss bedacht werden sollte:

- *„maxJobsActive“* gibt an, wie viele Jobs gleichzeitig vom Worker bearbeitet werden dürfen.
- *„pollInterval“* beschreibt die Zeit, die zwischen Schließen und erneutem Öffnen der Polls vergeht
- *„requestTimeout“* legt fest, wie lange ein long poll geöffnet bleibt.

Es müssen Versuchsaufbauten erarbeitet werden, die die Zusammenhänge zwischen der Systemleistung, dem Energieverbrauch und diesen Parametern herausstellen. Die Dimensionierung der Aufbauten orientiert sich grob an anderen Projekten der Camunda-Community (zum Beispiel dem Camunda-8-Benchmark-Projekt), um so eine Vergleichbarkeit der Ergebnisse zu ermöglichen. Auf diese Weise sollen zunächst das Optimierungspotential und erste Best Practices ermittelt werden. Die aufgezeigten Zusammenhänge können gegebenenfalls später zu einem Tool ausgebaut werden, das einige Effizienzanpassungen automatisiert.
Die Energieeffizienz eines Workers wird insgesamt von vielen Aspekten beeinflusst. Faktoren wie die Implementierung selbst, die gewählte Programmiersprache oder die Kubernetes-Konfiguration spielen insgesamt eine wichtige Rolle, sind aber nicht Gegenstand dieser Arbeit. Für die Analyseumgebung wird eine statische Konfiguration gewählt, die durch Best Practices (z.B. vorgegeben durch Camunda) oder andere Randbedingungen (z.B. Vertrautheit der Entwickler mit der Technologie) vorgegeben ist. Der Schwerpunkt liegt auf der Konfiguration des Workers selbst, da diese weitgehend unabhängig von anderen Design-Entscheidungen des Systems anwendbar ist.

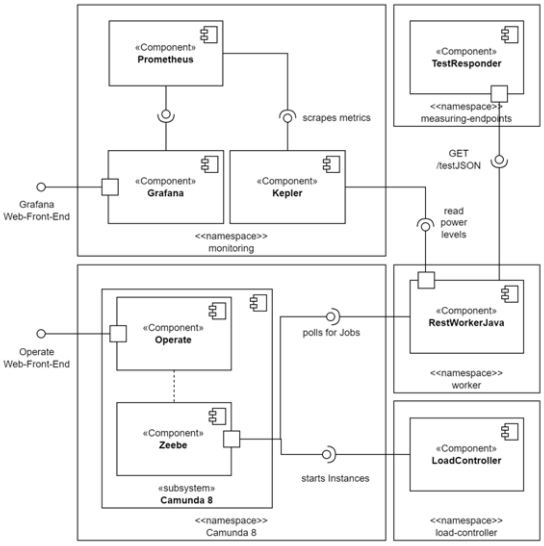


Abb. 3: Komponentendiagramm der Camunda-8-Messumgebung [4]

Implementierung

Um die Worker sinnvoll analysieren zu können, wird eine vollständige Camunda-8-Umgebung aufgesetzt. In dieser Umgebung werden typische Abläufe mit wechselnden Parametern evaluiert. Camunda wird in der Self-Hosted-Variante auf einem lokalen Kubernetes-

Cluster installiert, um alle Komponenten lokal unter Kontrolle zu halten. Zusätzlich wird das System um Prometheus, Grafana und Kepler erweitert. In diesem Aufbau ist es möglich, periodisch Verbrauchswerte pro Kubernetes-Pod abzufragen und damit den Worker isoliert zu betrachten. Dieser verschickt REST-Anfragen und verarbeitet die Antworten, die von einem dafür bereitgestellten Endpunkt im gleichen Cluster stammen. Eine weitere Komponente, der LoadController, kann zur periodischen Erzeugung von BPMN-Prozessen eingesetzt werden. Je nach Versuchsaufbau erzeugt er so eine konfigurierbare Menge Last auf System und Worker.

Ausblick

Je nach Ergebnissen der Analyse der Arbeitskräfte können verschiedene nächste Schritte formuliert werden. Wenn es klare Stellgrößen gibt, mit denen die Energieeffizienz beeinflusst werden kann, wäre es sinnvoll, ein Tool zu konzipieren, mit dem im Produktionsumfeld die passende Konfiguration für die aktuellen Anforderungen umgesetzt werden kann. Perspektivisch wäre eine Implementierung mit Hilfe von Machine Learning, dem sogenannten Predictive Process Monitoring (PPM), denkbar. Wenn die Analyse weniger vielversprechende Handlungsoptionen aufzeigt, sollten Messungen und Recherchen auf weitere Systemkomponenten oder Konfigurationsparameter ausgeweitet werden.

Literatur und Abbildungen

[1] Thomas Allweyer. *BPMN 2.0: introduction to the standard for business process modeling*. BOD - Books On Demand, 2 edition, 2016.

[2] Marcelo Amaral et al. Exploring Kepler's potentials: unveiling cloud application power consumption. https://www.cncf.io/blog/2023/10/11/exploring-keplers-potentials-unveiling-cloud-application-power-consumption/, 10 2023.

[3] Michele Chinosi and Alberto Trombetta. BPMN: An introduction to the standard. *Computer Standards & Interfaces*, 34:124–134, 2012.

[4] Eigene Darstellung.

[5] Liz Rice. *Learning eBPF: programming the Linux Kernel for enhanced observability, networking, and security*. O'Reilly Media, 1 edition, 2023.

[6] Bernd Rücker. Writing Good Workers For Camunda Cloud. https://blog.bernd-ruecker.com/writing-good-workers-for-camunda-cloud-61d322cad862?gi=b574593789e3, 07 2021.

Sicherheitsanalyse von IO-Link Wireless Systemen

Kai Wollrab **Dominik Schoop**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma Festo SE & Co. KG, Esslingen am Neckar

Einleitung

Durch die EU-Norm IEC 62443, welche die Grundlage für den Cyber Security Act und den Cyber Resiliance Act ist, gewinnt die Sicherheit von Bussystemen an Bedeutung. Bisher mussten sich lediglich die Betreiber von Anlagen um die Sicherheit eben dieser kümmern. Ab Ende 2025 sind auch die Hersteller der Komponenten für diese verantwortlich. Da Bussysteme, mit einigen wenigen Ausnahmen, über keine oder schwache Sicherheitsmechanismen verfügen, stellt das die Hersteller von Automatisierungssystemen unter Herausforderungen. Deshalb soll in dieser Arbeit die Sicherheit von IO-Link Wireless Systemen untersucht werden. Dabei soll auf potenzielle Gefahren hingewiesen und verschiedene Möglichkeiten zur sicheren Datenübertragung gezeigt werden.

IO-Link

IO-Link ist eine Punkt-zu-Punkt Kommunikationstechnologie für Sensoren und Aktoren in der Automatisierungstechnik. Sie ist in der Norm IEC 61131-9 als „Single-drop digital communication interface for small sensors and actuators“ (SDCI) definiert [1]. IO-Link ist ein System, für welches keine hinreichenden Sicherheitsfunktionen im Standard festgelegt wurden. Dies ist bei kabelgebundenen Systemen meist auch kein Problem, da der Angreifer für einen Angriff hier physischen Zugriff zum Ziel benötigt. Durch IO-Link Wireless ändert sich allerdings dieser Umstand. Da hier die Daten kabellos übertragen werden, können sie von potenziellen Angreifern auch ohne physischen Zugang zum System abgefangen oder verändert werden.

Sicherheit in Bussystemen

Für die Implementierung von Sicherheitsfunktionen in Bussystemen müssen gewisse Vorüberlegungen getroffen werden. Die sicherste Variante der Kommunikation wäre, alle Nachrichten mit einem als sicher geltenden Verschlüsselungsalgorithmus zu verschlüsseln. Allerdings müssen viele Anwendungen, in denen Bussysteme eingesetzt werden, echtzeitfähig sein. Durch die Dauer, die für diese Ver- und Entschlüsselung auf den Devices, welche überwiegend nur über günstige Mikrocontroller mit wenig Rechenleistung verfügen, kann diese Echtzeitfähigkeit nicht immer garantiert werden. Daher bietet sich für manche Anwendungsfälle auch eine Authentifizierung an. Hier kann beispielsweise mithilfe eines „Rolling-Key“ für jede Nachricht ein anderes Geheimnis übertragen werden, welches den Sender eindeutig ausweist. Dadurch, dass die Daten nicht vollständig verschlüsselt, sondern lediglich um das Geheimnis erweitert werden, wird weniger Rechenleistung und Zeit benötigt, wodurch die Echtzeitfähigkeit eher möglich ist als bei der Verschlüsselung aller Daten. Eine solche Authentifizierung kann beispielsweise über symmetrische Kryptografiealgorithmen realisiert werden.
Für die Sicherheit in Bussystemen muss bestimmt werden, wie sensibel die zu übertragenen Daten sind. Für Daten, welche sehr schützenswert sind, sollte eine vollständige Verschlüsselung in Betracht gezogen werden.

IO-Link Wireless

IO-Link Wireless (IOLW) basiert auf dem IO-Link Protokoll, welches um die kabellose Übertragung erweitert wird. Dafür wird das ISM Band 2,4 GHz zum Übertragen der Nachrichten genutzt [3]. Die Kommunikationsteilnehmer heißen im IOLW Kontext Wireless Master (W-Master) und Wireless Device (W-Device). Vorteile einer kabellosen Datenübertragung sind:

- **Installationsaufwand** geringer, da keine Datenleitung benötigt wird.
- **Kostenersparnis**, da keine Buchsen am W-Master oder W-Device.
- **Bewegliche Teilnehmer** möglich

Nachteile gegenüber kabelgebundenen Lösungen:

- **Ausfallsicherheit** ist durch die Übertragungsart deutlich schlechter.

- **Angriffsfläche** ist größer, da die übertragenen Daten simpel abfangen werden können.

Da das ISM 2,4 GHz frei genutzt werden kann, wird es von vielen Technologien und Geräten wie WiFi, Bluetooth, DECT oder verschiedenen Smart-Home Standards verwendet [4]. Da viele Teilnehmer auf diesem Band kommunizieren, können sich diese gegenseitig stören. IOLW nutzt dafür Blocklisten und das Frequenzsprungverfahren. Wenn ein Kanal bereits ausgelastet ist, wird dieser auf eine Blockliste gesetzt, welche vom W-Master an die W-Devices verteilt wird. Diese Kanäle werden dann nicht mehr genutzt. Zusätzlich ändert sich durch das Frequenzsprungverfahren kontinuierlich die Frequenz, auf welcher kommuniziert wird, um nicht zu sehr von Störungen auf einzelnen Frequenzen beeinträchtigt zu werden [1].

Aufbau eines IOLW Pakets

IOLW nutzt in der Laufzeit jeweils ein DLink (Download-Link) Paket, welches die Daten für alle W-Devices enthält und eine Phase für ULink (Upload-Link) Pakete von den W-Devices an den Master. Die DLink und ULink Pakete sind grundsätzlich gleich aufgebaut: Sie besitzen einen Header, den Payload und abschließend eine CRC-Summe, welche die Daten vor Übertragungsfehlern validiert. In Abbildung 1 ist ein DLink Paket dargestellt. Neben dem DataSyncword, welches für die Synchronisierung der W-Devices mit dem W-Master benötigt wird, sind hier auch die MasterID und die Track_N enthalten. Track_N (Track Number) enthält den Wert des W-Devices, für welches dieses Paket bestimmt ist [1].

<table>
<tr><td>Oktett</td><td colspan="8">1</td><td colspan="8">2</td><td colspan="8">3</td><td colspan="8">4</td></tr>
<tr><td>Bit</td><td>0</td><td>1</td><td>2</td><td>3</td><td>4</td><td>5</td><td>6</td><td>7</td><td>0</td><td>1</td><td>2</td><td>3</td><td>4</td><td>5</td><td>6</td><td>7</td><td>0</td><td>1</td><td>2</td><td>3</td><td>4</td><td>5</td><td>6</td><td>7</td><td>0</td><td>1</td><td>2</td><td>3</td><td>4</td><td>5</td><td>6</td><td>7</td></tr>
<tr><td>0</td><td colspan="16">Preamble</td><td colspan="16">DataSyncword</td></tr>
<tr><td>4</td><td colspan="8">DataSyncword</td><td colspan="4">MasterID</td><td colspan="4">Track_N</td><td colspan="8">ACK</td><td colspan="8">Payload</td></tr>
<tr><td>8</td><td colspan="8">Payload</td><td colspan="16">CRC16</td><td colspan="8">Payload</td></tr>
<tr><td>12</td><td colspan="32">Payload</td></tr>
<tr><td>16</td><td colspan="32">Payload</td></tr>
<tr><td>...</td><td colspan="32">Payload</td></tr>
<tr><td>44</td><td colspan="32">Payload</td></tr>
<tr><td>48</td><td colspan="32">CRC32</td></tr>
</table>

Abb. 1: Aufbau eines IOLW Pakets [2]

Roaming

Da es Anwendungsfälle gibt, bei welchen die W-Devices außerhalb der Funkzelle des W-Masters operieren müssen, gibt es die Möglichkeit des Roamings. Dabei wählt sich das W-Device automatisch bei dem W-Master mit dem besten Signal ein. Die Master müssen dem W-Device bekannt sein. Außerdem muss das W-Device auch bei den W-Mastern hinterlegt werden, welche alle miteinander verbunden sind. Wenn nun die Roaming-Funktion beim W-Device aktiviert ist, kann es ohne Unterbrechung der Datenverbindung zwischen den verschiedenen Funkzellen der W-Master wechseln.

Angriffe auf ein IOLW System

Um Angriffe auf ein IOLW System zu fahren, müssen zuerst mögliche Angriffspunkte evaluiert werden. Dabei sticht die kabellose Übertragung besonders heraus. Prinzipiell gibt es zwei Punkte, an denen IOLW Systeme angegriffen werden können: den Physical Layer und den Application Layer.

Angriffe auf den Physical Layer

Kabellose Datenübertragungen können verhältnismäßig einfach gestört werden. Hierfür wird lediglich ein Gerät benötigt, welches zufällige Daten im Frequenzspektrum der IOLW Kommunikation sendet. Diese Art von Angriff ermöglicht allerdings lediglich eine Ausfallzeit des Systems. Zusätzlich wird dieser Angriff durch den Mechanismus der Blocklisten und der Frequenzsprungtabellen erschwert. Ein weiterer Angriff auf den Physical Layer wäre das Mitlesen der Kommunikationspakete. Hier können eventuell Schlüsse gezogen werden, welche Pakete wann von welchem Gerät gesendet werden und welche Daten diese Pakte enthalten könnten.

Angriffe auf den Application Layer

Angriffe auf den Application Layer können verschiedene Ziele haben. Beispielsweise kann man abgefangene Pakete analysieren und deren Daten anschließend auswerten. Dabei können wichtige Erkenntnisse wie die Art und die Anzahl von verschiedenen Sensoren und Aktoren im IOLW System bestimmt werden. Aus diesen Daten könnten möglicherweise kritische Informationen gewonnen werden. Außerdem können gezielt Datenpakete an den W-Master oder das W-Device gesendet werden, welche für den Empfänger legitim aussehen, allerdings falsche Daten enthalten. So könnte man beispielsweise Maschinen zum Überhitzen

bringen, wenn der Angreifer dem W-Master ständig falsche Temperaturwerte liefert.
Die Erfolgsaussichten von Angriffen auf das Application Layer könnten mit einem guten Sicherheitskonzept deutlich reduziert werden. Beispielsweise können Daten nur schwer ausgelesen werden, wenn sie mit einem als sicher geltenden Verschlüsselungsalgorithmus verschlüsselt wurden. Für nicht kritische Daten könnte eine Authentifizierung genutzt werden, welche den Sender des Datenpakets sicher authentifiziert und nicht von einem möglichen Angreifer kopiert werden kann (bspw. mithilfe eines Rolling-Key).

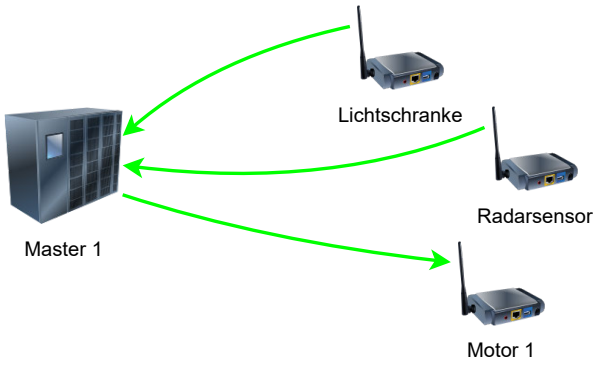


Abb. 2: Beispielsystem ohne Angreifer [2]

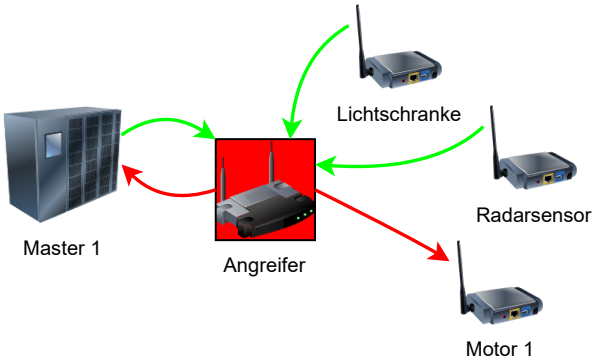


Abb. 3: Beispielsystem mit Angreifer. Grüne Pfeile stellen die eigentlichen Datenströme dar, rote Pfeile die vom Angreifer abgeänderten Daten. [2]

Ziel der Arbeit

Die Bachelorarbeit verfolgt das übergeordnete Ziel, die Sicherheitsaspekte im Kontext von IO Link Wireless (IOLW) umfassend zu untersuchen und Strategien zur Implementierung wirksamer Sicherheitsmechanismen aufzuzeigen. Der Schwerpunkt liegt hierbei auf der Identifikation potenzieller Sicherheitsrisiken innerhalb von IOLW sowie der Ausarbeitung von Lösungsansätzen zur Prävention möglicher Angriffe.
Dies schließt eine eingehende Analyse der Sicherheitsbedrohungen und Schwachstellen von IOLW-Systemen ein, um ein fundiertes Verständnis für die bestehenden Herausforderungen zu entwickeln. Ein weiteres Ziel besteht in der Konzeption und Realisierung von Sicherheitsmechanismen, die gezielt auf die Anforderungen und Schwachstellen von IOLW abgestimmt sind, um eine effektive Abwehr potenzieller Angriffe zu gewährleisten.
Die Effektivität der implementierten Sicherheitsmechanismen wird anschließend anhand von analysierten Angriffsszenarien evaluiert, um festzustellen, inwiefern präventive Maßnahmen potenziellen Schaden verhindern können.
Basierend auf den durchgeführten Analysen werden konkrete Handlungsempfehlungen für zukünftige Sicherheitsstrategien im Bereich IO Link Wireless abgeleitet.
Durch die Ergebnisse dieser Arbeit soll nicht nur das Bewusstsein für Sicherheit in IOLW-Systemen geschärft, sondern auch Ansätze für die Gewährleistung der Integrität, Vertraulichkeit und Verfügbarkeit von IOLW-Systemen bereitgestellt werden.

Literatur und Abbildungen

[1] IO-Link Community. *IO-Link Wireless System Extensions*. IO-Link Community, 2023.

[2] Eigene Darstellung.

[3] Ralf Heynicke et al. IO-Link Wireless enhanced factory automation communication for Industry 4.0 applications. *Journal of Sensors and Sensor Systems*, 7:131–142, 2018.

[4] Mia Torres-Dela Cruz et al. Networks Coexistence with other 2.4 GHz ISM Devices. *International Journal of Trend in Scientific Research and Development*, 2018.

Multi Label Classification for Unstructured Text Data

Tuba Yalcinoez **Markus Enzweiler**

Department of Computer Science and Engineering, Esslingen University

Work carried out at Department of Computer Science and Engineering, Esslingen

Introduction

Multi label classification is becoming increasingly attractive and is being used in a variety of applications. It enables the flexible analysis and understanding of complex data and has the advantage that an input can be assigned to several labels simultaneously without compromising the integrity of the data. Multi label classification has been applied for almost all fields. [4] This thesis presents and evaluates different approaches for a multi label classification task to be used for process automation in the scope of product production. The work was carried out in cooperation with a leading global company in the automotive industry. For this thesis, the main purpose is to predict multiple labels for unstructured text data. Therefore, different machine learning models like Feed Forward Neural Network (FFNN), Convolutional Neural Network (CNN), Decision Tree, Random Forest, Support Vector Machine (SVM), K-nearest neighbors (ML-KNN), and Recurrent Neural Network (RNN) based Long Short-Term Memory (LSTM), are built, evaluated, and compared to find the best model with the highest prediction rate for the given dataset. This thesis stands out by the comparison of the various machine learning algorithms for multi label classification to find the best model for the unstructured text data.

Goal of the project

The importance of process automation has increased significantly in recent years and is becoming ever more crucial in the production industry. It serves to increase product quality, enhance productivity and efficiency, improve process reliability and plant availability, as well as utilize resources efficiently. With process automation, information, communication, and automation technology can be closely integrated into an operational context. [1] This work aims to speed up product production by automating the process with the help of artificial intelligence. The actual product production phase is regulated manually and is time-consuming. In the early product production pipeline step, a Design Deployment Notification (DDN) is sent to the documentarian depending on the adjustments and changes made in the components and documentation. The documentarian examines the DDN and after that must decide based on that information as "Decision1" or "Decision2" as the recommendation for action. They also must justify their decisions with reasons as a text, which is written in a free text field. Therefore, the reason can have different versions and it is not standardized. This process takes time and can cause manual errors. This can be improved by giving the input DDN to an AI model so that the model can train on the available DDNs and make future predictions. The goal of the thesis is to compare and evaluate various machine learning algorithms, which can classify unstructured text data input into multi labeled outputs. For each Design Deployment Notification, a subset of 6 labels, which are associated with the "Decision1" and "Decision2" and the reason behind that decision, will be predicted. Thus, the DDN can have multiple labels associated with it at the same time. By learning from the dataset, the models learn how the documentarian makes decisions and approaches a DDN, and makes predictions accordingly. DDNs will be automatically classified, and the documentarian's job will be simplified. Documentarian can examine the prediction of the machine learning model, thereby speeding up decision making. Additionally, for a complex DDN, the AI can examine the past DDNs and take them into account.

Multi Label Classification

The focus of this thesis is the multi label classification for unstructured text data. Data classification is a technique to classify the instances into one of the predetermined set of labels. There are various classification algorithms such as binary, multi class, and multi label classification. The classification task is based on the data and the parameters, and it is then associated with the outputs. Due to the expanding amount of information and the many aspects of this

information, it is becoming difficult to categorize them into just one class or label. In this case, multi label classification can be used. The is a set of relevant labels, which can be assigned to the instance set. Multiple labels can be allocated to a specific sample. [3] For multi labelled data, the labels are not mutually exclusive, so multiple labels can be associated with a particular sample. [4] For each input sample with n-dimensional numerical or categorical features, there is a subset of labels associated with it. In the binary output vector, relevant labels are set to 1. Output dimensionality is equal to the number of labels. The aim is to predict the subset of labels for a new given sample. [3]

Label1	Label2
0	1

Binary Classification

Label1	Label2	Label3	Label4	Label5	Label6
0	1	0	0	0	0

Multi Class Classification

Label1	Label2	Label3	Label4	Label5	Label6
0	1	0	1	0	0

Multi Label Classification

Fig. 1: Classification Types [5]

Figure 1 shows three tables, each with an example of the outputs of the classification methods. With binary classification, the output has 2 labels, whereby only one label can be set, i.e. it can have the value 1. Multi-class classification differs from binary classification in that it can have more than two labels, but one can still be set. In the example, only label2 of the six labels has the value 1. Finally, the third example is a multi-label classification with six labels, where more than one label can be set. Label2 and Label4 have been set in Figure 1. Furthermore, the output dimensionality for this work is six, which means that there are six labels.

Dataset

The dataset is in the scope of product production and consists of mixed data like numerical, alphanumeric data, categorical, and text data. It involves multi labeled data. Explanatory variables are extracted out of them and are responsible for the target label. Most of the information is gathered from the free text field. Predictive performance and the target variable depend on every new feature. The approach is to add more explanatory variables as new features are available. The dataset is updated by adding more features. Thus, the dataset is not fixed but it is dynamically changing, with new fields and newly collected information. Therefore, it is essential to accurately describe the general characteristics of the samples and the preprocessing steps before explaining the algorithms and building the models. Raw data contains information from the Design Deployment Notification. It has various tabs and fields, most of which are extracted from free text fields. This raw data can obtain noisy, redundant, inconsistent data also missing data values, so it cannot be fed directly into the model. It is very important to pre-process the data because the success of machine learning algorithms depends on the quality of the data. If the data is insufficient or contains irrelevant information, the machine learning algorithms cannot deliver accurate and understandable results or discover useful insights to begin with. [2] Therefore data preprocessing is performed. Data is cleaned by removing punctuation, unwanted spaces, and missing values. Data is then standardized, for example by replacing shortform with longform or by naming them in a standardized way. For this standardization catalogs are used. It was an important step since in the human written free text fields various formats could be seen for the same decision and reasoning. The collected explanatory variables are merged at the end to have them all in the unified data frame. After data preprocessing steps data is encoded. Encoding methods like One Hot encoding, Feature Hashing, SBERT, and UMAP are implemented in the data.
Target labels are selected and defined after they are extracted and preprocessed from the raw data. After preprocessing and feature engineering steps samples are generated. A sample has encoded features and 6 labels as the target variable. Multiple labels could be set to 1. The created dataset has input, and output values gained from the raw data. The available labeled data is limited and usually with a larger dataset, a higher predictive rate could be gained. Therefore, with the help of semi supervised learning, labels are assigned to unlabeled data. And they will be added to the final database.

Method and Approach

Various models can be used to classify multiple labels. Since the labeled ground truth dataset, which is available for training, is rather small, a semi-supervised approach is followed. It has a powerful learning behavior as it can use available unlabeled data to optimize supervised learning tasks when the labeled data is limited or costly. Semi supervised learning can be used to categorize unlabeled data. [6]
Three main steps were followed for the task, which can also be seen in the flowchart in figure 2. As the first step first classifier is trained. For the first classifier, FFNN, CNN, Decision Tree, Random Forest, SVM, and ML-KNN, are selected, built, and evaluated. It is the base model. The first classifier is trained with 90% of the ground truth dataset. It is fine-tuned

and its performance is tested with the rest 10% of the ground truth. 1.Classifier is used for automatic labeling purposes. It can be seen in the second section of the diagram, that the first classifier is used to predict and assign labels for the unlabeled datapoints. Pseudo labeled dataset is generated for "Unlabeled Dataset". It is combined with 90% of the human labeled dataset to create the final training dataset for the 2. Classifier.

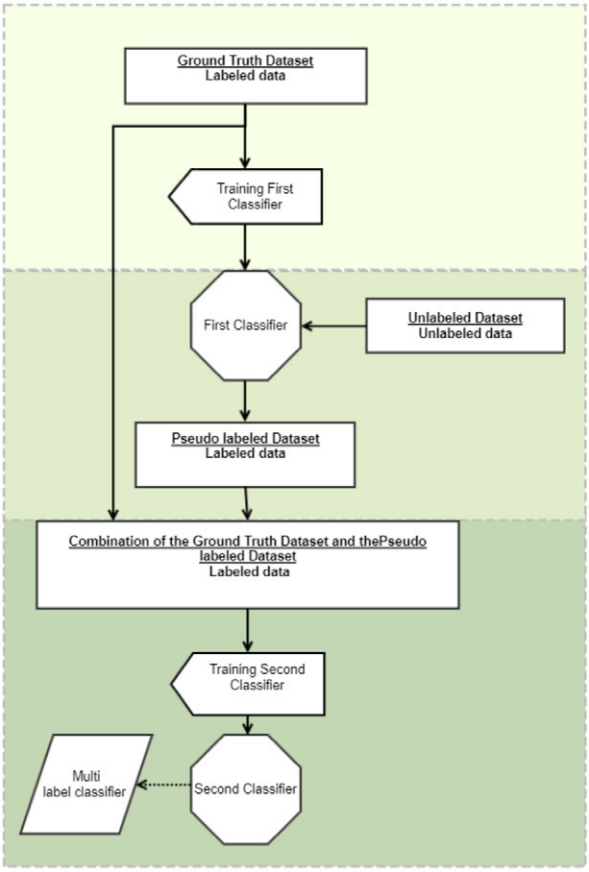


Fig. 2: Flowchart [5]

Training the second classifier with the combined dataset to specify the final multi label classifier is the last step. For the second classifier, FFNN, CNN, Decision Tree, Random Forest, SVM, LSTM, and ML-KNN, are selected, built, and evaluated. It will be the main classifier to predict multi label outputs for DDNs. 2.Classifier will be trained with 90% of the combined dataset and will be tested two times. Firstly, with the 10% of the ground truth dataset, which is also used for testing the 1. Classifier. Secondly with the rest 10% of the combined dataset. This way model performance can be compared based on the ground truth as well as based on the automatically generated data points. As for the decision of different algorithms, covering different teams of algorithms played a huge role. FFNN is implemented because it has a good predictive performance. Followed by Convolutional Neural Networks, tree-based Decision Tree and Random Forest, distance-based ML-KNN, and SVM and RNN-based LSTM models.
To find the most efficient classifier, trained models are evaluated using a couple of evaluation metrics. Selected evaluation metrics are f1-score, recall, precision, hamming loss, ranking loss, and accuracy. Also, with the help of a classification report, and confusion matrix performance of each label can be observed. Depending on the results of the evaluation metrics, models are fine-tuned and improved. After the combinations of the first and second classifiers are compared, the best-performing combination is selected.

References and figures

[1] Sirkka-Liisa Jämsä-Jounela. Future trends in process automation. *Annual Reviews in Control*, 2007.

[2] Sotiris Kotsiantis, Dimitris Kanellopoulos, and P. Pintelas. Data Preprocessing for Supervised Learning. *International Journal of Computer Science*, 2006.

[3] Swati Mathur and Pratistha Mathur. Multi-label Classification: Detailed Analysis. In *Cyber Intelligence and Information Retrieval*. Springer Singapore, 2022.

[4] Sang-Hyeun Park. Efficient Decomposition-Based Multiclass and Multilabel Classification, 2012.

[5] Own representation.

[6] Xiaojin Zhu and Andrew B. Goldberg. *Introduction to Semi-Supervised Learning*. Springer Cham, 2009.

Maschinensicherheit in der Industrie 4.0: Konzeption und prototypische Umsetzung mittels Asset Administration Shell, MQTT und OPC UA

Mehmet Akif Yalcinoez **Karin Melzer**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma Pilz GmbH & Co. KG, Ostfildern

Einleitung und Problemstellung

Industrie 4.0 beschreibt eine flexible und hochdynamische industrielle Produktion, die über global vernetzte Wertschöpfungsnetzwerke mit neuen Formen der Zusammenarbeit realisiert wird. Schlüsselfaktoren für den Erfolg in den vernetzten Wertschöpfungsnetzwerken sind die Verfügbarkeit, Transparenz und der Zugang zu Daten. [4] Industrie 4.0 zielt darauf ab, trotz hoher Marktvolatilität eine wirtschaftliche Produktion unter dynamischen Bedingungen zu ermöglichen. Für die technische Umsetzung ist eine Erneuerung der technologischen Infrastruktur sowie eine Modernisierung der Anlage erforderlich. An Maschinen sind Umbauten, Aufrüstungen und Optimierungen zu erwarten. Veränderungen an einer Maschine oder Anlage können aus diversen Gründen notwendig sein, wie zum Beispiel:

- Änderung der Anforderungen im Rahmen des Produktionsprozesses,
- die Herstellung neuer Produkte auf bestehenden Anlagen,
- schnellere und effizientere Produktion.

Da eine gebrauchte Maschine aufgrund nachträglicher wesentlicher Veränderungen neue Sicherheitsrisiken bergen kann, gilt sie als ein neues Produkt. [3] Durch diesen Wechsel fällt die Maschine nun unter das Produktsicherheitsgesetz anstatt der Betriebssicherheitsverordnung. Der Betreiber, der für die maßgebliche Veränderung verantwortlich ist, wird somit zum Hersteller und muss dementsprechend die Herstellerpflichten gemäß dem Produktsicherheitsgesetz erfüllen, [1] wie zum Beispiel:

- Das gesamte Konformitätsbewertungsverfahren einschließlich der Risikobeurteilung und deren Dokumentation muss für die Maschine erneut durchgeführt werden.
- Eine neue Konformitätserklärung ist auszustellen und die CE-Kennzeichnung anzubringen.
- Die technische Dokumentation und die Betriebsanleitung der Maschine müssen überarbeitet oder neu erstellt werden.

Der Betreiber der Maschine muss dann mit einem erheblichen zeitlichen, personellen und finanziellen Aufwand rechnen, um diesen Pflichten nachzukommen. Darüber hinaus ist bei jeder Änderung, unabhängig davon, ob es sich um eine wesentliche oder unwesentliche Änderung handelt, eine Gefährdungsbeurteilung nach § 3 der Betriebssicherheitsverordnung (BetrSichV) durchzuführen.
Eine weitere Herausforderung besteht darin, alle Dokumente und Spezifikationen aktuell und konsistent zu halten, während die Maschine geändert und das gesamte System zusammengeführt wird. Generell weisen Veränderungen an einem System oft komplexe Abhängigkeiten und Querverweise auf, die allein auf der Grundlage textbasierter Dokumentationen nur schwer zu erkennen und zu verwalten sind. Insbesondere die Identifizierung der Veränderungen, die erst zu einem späten Zeitpunkt im weiteren Verlauf des Betriebs erkannt werden, kann sich als sehr kostspielig erweisen und zu unsicheren Systemen führen. Darüber hinaus können häufige Veränderungen in der Sicherheitstechnik nicht angemessen berücksichtigt werden, was zu Abweichungen zwischen dem modifizierten System und seinen Dokumentationen und damit zu einem erheblichen Sicherheitsrisiko führen kann. [6]

Zielsetzung

Im Rahmen dieser Arbeit wird ein Konzept zur Erkennung wesentlicher Veränderungen an der Maschine auf der Basis von Technologien wie Asset Administration Shell, MQTT und OPC UA entwickelt. Zur Validierung des Konzepts wird ein Prototyp implementiert.

Das Konzept soll dazu beitragen, Veränderungen (potenziell neue Risiken) sofort zu erkennen und damit ein spätes Erkennen von Veränderungen zu vermeiden. Dies dient der Sicherheit der Menschen und der Rechtssicherheit des Unternehmens. Außerdem entfallen Kosten und Aufwand für die manuelle Prüfung von Veränderungen. Darüber hinaus sollen mit dem Konzept Maschineninformationen, insbesondere sicherheitsrelevante Parameter, herstellerübergreifend und standardisiert abgelegt werden, um die Informationsbereitstellung zu automatisieren und mögliche Abweichungen zwischen der geänderten Maschine und ihrer Dokumentation zu vermeiden.

Konzept

Das erarbeitete Konzept besteht aus drei Teilen: Der erste Teil befasst sich mit der digitalen Abbildung der Maschine und ihrer Sicherheitskomponenten mit Hilfe von digitalen Zwillingen. Die so bereitgestellten Informationen, insbesondere die sicherheitsrelevanten, können zyklisch abgefragt und auf Veränderungen überprüft werden. Für die unternehmensübergreifende und branchenneutrale Informationsbereitstellung ist ein standardisiertes Datenmodell notwendig, um die Daten schneller und einfacher verfügbar zu machen und Maschinen unterschiedlicher Hersteller digital abzubilden. Darüber hinaus ist ein modellbasierter Ansatz erforderlich, der die Elemente zwischen den verschiedenen Änderungen und Dokumentationen formal verknüpft und Diskrepanzen zwischen der geänderten Maschine und ihrer Dokumentation verhindert.
Asset Administration Shell (AAS) oder die Verwaltungsschale im Deutschen als Kernelement von Industrie 4.0 adressiert diese Anforderungen als herstellerübergreifender und branchenneutraler Standard für die Bereitstellung von Daten und für die Kommunikation in einer einheitlichen Sprache. Die Asset Administration Shell identifiziert ein Asset global als digitale Repräsentation und stellt Informationen über dessen Eigenschaften und Fähigkeiten in entsprechenden Teilmodellen(Submodels) bereit, wie in Abbildung 1 dargestellt. So können Assets über standardisierte Schnittstellen und eine gemeinsame Sprache miteinander kommunizieren und vernetzte Wertschöpfungsnetzwerke realisiert werden. [2]

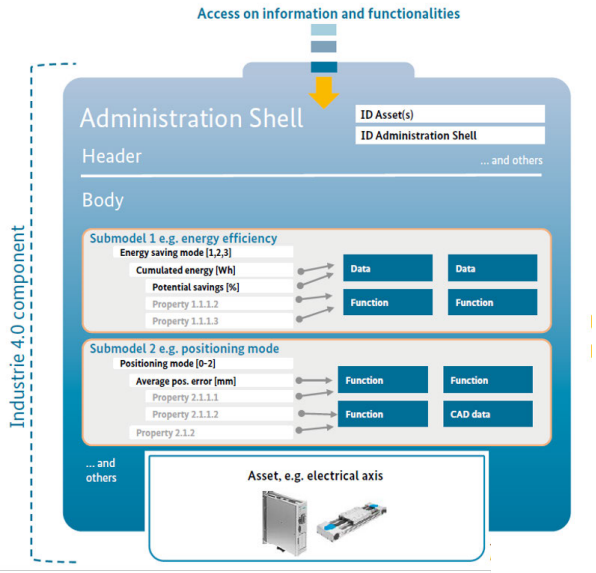


Abb. 1: Asset Administration Shell [7]

Im zweiten Teil des Konzepts geht es um die Erkennung von Veränderungen und die Zustandsüberwachung der Maschine. Diese sollen mit Hilfe der AAS abgebildet werden und mittels MQTT als leichtgewichtiges und offenes Protokoll für die Kommunikation zwischen Maschinen, das besonders für das Internet der Dinge (IoT) geeignet ist, übertragen werden. MQTT basiert auf dem Publish/Subscribe-Prinzip, bei dem ein MQTT-Broker Nachrichten zwischen MQTT-Clients vermittelt, die als Publisher oder Subscriber fungieren können. [5]
Im dritten Teil wird die Open Platform Communications Unified Architecture (OPC UA) für die unternehmensinterne Datenkommunikation eingesetzt. Dies wird relevant, wenn mehrere Sensoren eingesetzt werden und/oder ein Aktor auf die erfasste Veränderung reagieren soll. Darüber hinaus sollte eine Webanwendung implementiert werden, die einen Überblick über die identifizierten Veränderungen gibt und eine Entscheidungshilfe bietet, ob eine wesentliche Veränderung vorliegt. In diesem Zusammenhang bietet das Interpretationspapier eine Hilfestellung bei der Entscheidung, ob eine signifikante Änderung vorliegt. Das Interpretationspapier basiert auf dem in Abbildung 2 dargestellten Flussdiagramm.

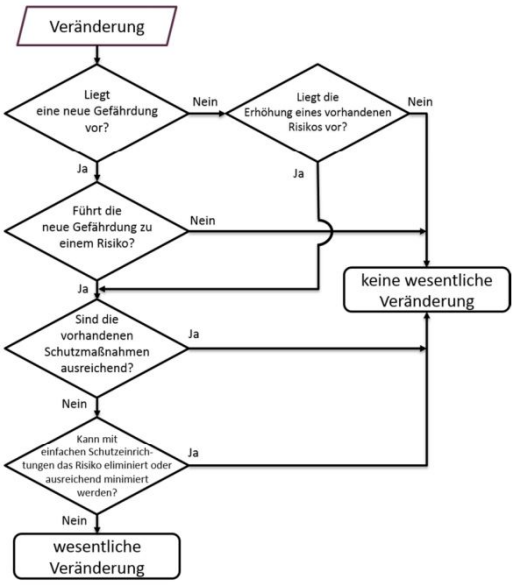


Abb. 2: Entscheidungsschritte - wesentliche Veränderung von Maschinen [1]

Ausblick

Im weiteren Verlauf der Arbeit wird der Prototyp zur Validierung des Konzepts implementiert. Diskutiert werden noch die Herausforderungen, Einschränkungen und Alternativen bei der Umsetzung.

Literatur und Abbildungen

[1] Bekanntmachung des Bundesministerium für Arbeit und Soziales. Interpretationspapier Wesentliche Veränderung von Maschinen. https://www.bmas.de/SharedDocs/Downloads/DE/Arbeitsschutz/interpretationspapier-veraenderung-maschinen.pdf?__blob=publicationFile&v=4, 2015.

[2] DKE Deutsche Kommission Elektrotechnik Elektronik Informationstechnik in DIN und VDE. Verwaltungsschale als Rückgrat der Industrie 4.0 und des Digitalen Zwillings. https://www.dke.de/de/arbeitsfelder/industry/verwaltungsschale, 2022.

[3] Europäische Kommission Generaldirektion Binnenmarkt Industrie Unternehmertum und KMU. „Blue Guide“ Leitfaden für die Umsetzung der Produktvorschriften der EU : 2014. https://data.europa.eu/doi/10.2769/31501, 2015.

[4] Plattform Industrie. Leitbild für Industrie 4.0. https://www.plattform-i40.de/IP/Redaktion/DE/Downloads/Publikation/Leitbild-2030-f%C3%BCr-Industrie-4.0.pdf?__blob=publicationFile&v=1, 2019.

[5] MQTT MQ Telemetry Transport. The Standard for IoT Messaging. https://mqtt.org, 2022.

[6] Daniel Schneider. Model-Based Safety Engineering. https://www.iese.fraunhofer.de/blog/safety-engineering/, 2021.

[7] ZVEI Zentralverband der Elektro-und Digitalindustrie. Asset Administration Shell from ZVEI SG Modelle & Standards. https://www.zvei.org/fileadmin/user_upload/Presse_und_Medien/Publikationen/2019/November/Faktenblatt_Industrie_4.0_in_der_Sensorik/ZVEI_FB_Industrie_4.0_in_der_Sensorik.pdf, 2019.

Konzeption und Umsetzung eines Demonstrators für Lademanagement

Bedirhan Yanik **Mirko Sonntag**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma Anwendungszentrum KEIM des Fraunhofer Instituts für Arbeitswirtschaft und Organisation (IAO), Esslingen

Einführung

Die Elektromobilität wird weltweit als Schlüssel für die klimafreundliche Mobilität angesehen, weshalb Elektrofahrzeuge ein wesentlicher Bestandsteil der Strategie zur Reduzierung des CO2-Ausstoßes und zur Erhöhung der Mobilität innerhalb der Städte sind. Als mobile Stromspeicher könnten Elektrofahrzeuge in Zukunft auch Schwankungen von Wind- und Sonnenkraft ausgleichen und somit den Ausbau und die Integration dieser unsteten erneuerbaren Energiequellen unterstützen. [3]

Damit die Elektromobilität auch praktikabel ist, braucht es eine flächendeckende, bedarfsgerechte und nutzerfreundliche Ladeinfrastruktur [6]. In Abb. 1 kann man die Entwicklung der Ladeinfrastruktur in Deutschland sehen. Darin wird sichtbar, dass die Anzahl der Ladepunkte für Elektrofahrzeuge von 2017 bis 2023 um mehr als ein zehnfaches gestiegen ist und die Zahl der Schnellladepunkte zunehmend größer wird. Der effiziente Einsatz der Ladeinfrastruktur wird dabei durch das Lademanagement ermöglicht, welches einen wichtigen Aspekt für Elektrofahrzeuge darstellt.

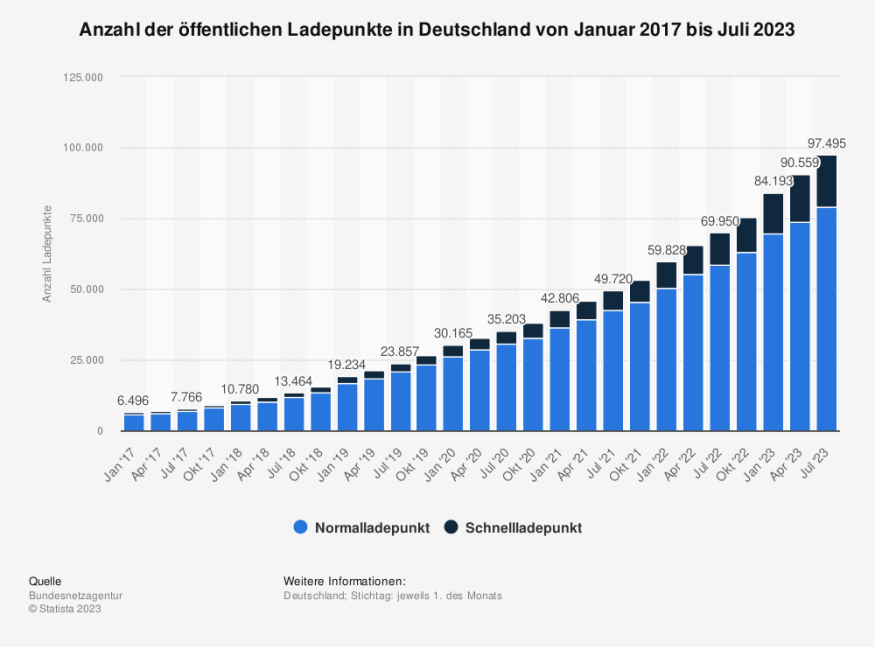


Abb. 1: Anzahl öffentlicher Ladepunkte in Deutschland von Januar 2017 bis Juli 2023 [1]

Ziel der Arbeit

Die Konzepte und Herausforderungen für das Lademanagement und der Ladeinfrastruktur von Elektrofahrzeugen verschiedenen Nutzergruppen nahezubringen, ist oft schwierig.

Daher ist das Ziel der Arbeit, ein Demonstrationswerkzeug zu entwickeln, welche das Lademanagement bzw. die Ladeinfrastruktur und deren Auswirkungen auf die

Elektromobilität auf einfache und verständliche Weise darstellen. Dabei ist es wichtig den aktuellen Stand der Forschung und Entwicklung und die Wissensstände bzw. die Interessen von verschiedenen Nutzergruppen zu berücksichtigen.
Mit dieser Arbeit sollen somit erste Erkenntnisse über diese Thematik und insbesondere zum aktuellen Stand der Technik für das Anwendungszentrum KEIM ermittelt werden. Das entwickelte Demonstrationswerkzeug dient sodann als mögliche Methode zur Visualisierung der Forschung am Anwendungszentrum KEIM und soll vor Ort beispielsweise Schülern, Studenten, Politikern und Besuchern oder auch auf Messen die zentralen Forschungssfelder näherbringen.

Lademanagement und Ladeinfrastruktur

Ein Backendsystem für Ladesäulen, auch bekannt als Lademanagement, überwacht und steuert die Kommunikation zwischen Fahrzeugen und Ladesäulen. Es hat mehrere Aufgaben:

1. Autorisierung: Es überprüft, ob ein Fahrzeug berechtigt ist, an einer bestimmten Ladesäule zu laden. Bei einer negativen Entscheidung wird das Fahrzeug abgelehnt.
2. Datenüberwachung und Speicherung: Während des Ladevorgangs werden Daten zur Lademenge gesammelt und für die spätere Abrechnung gespeichert.
3. Kommunikationsüberwachung: Es überwacht ständig die Kommunikation zur Ladesäule, um deren Erreichbarkeit zu gewährleisten.
4. Fehlerbehandlung: Im Falle von Fehlern, die zu eingeschränkter Funktionalität führen, ist es Aufgabe des Lademanagements, dies an die verantwortliche Stelle zu kommunizieren und eine Reparatur zu veranlassen.

Es bietet zudem Monitoring, Fernwartung für schnelle Problemlösung und eine einfache und gesetzeskonforme Kostenabrechnung. [2]
Die Ladeinfrastruktur in Deutschland bezieht sich auf die Gesamtheit der Installationen und Einrichtungen, die es ermöglichen, Elektrofahrzeugen, insbesondere Elektroautos, das Aufladen ihrer Batterien zu ermöglichen. Diese Infrastruktur umfasst sowohl öffentliche als auch private Ladepunkte und ist ein wesentlicher Bestandteil der Elektromobilität in Deutschland. Ein zentraler Aspekt ist die Notwendigkeit, eine ausreichende Anzahl an Ladepunkten schaffen, um den steigenden Bedarf an Elektrofahreugen zu decken. Laut Bundesregierung sollen bis 2030 eine Million öffentliche Ladepunkte entstehen [6]. Ein weiterer wichtiger Aspekt ist die Nutzerfreundlichkeit der Ladeinfrastruktur. Die Ladeinfrastruktur sollte bedarfsgerecht und leicht zugänglich sein, um die Akzeptanz der Elektromobilität zu fördern. Darüber hinaus ist es wichtig, dass die Ladeinfrastruktur sicher ist. Dies beinhaltet die Verwendung von sicheren Ladetechnologien und die Einhaltung von Sicherheitsstandards und -vorschriften [4]. Schließlich ist es wichtig, dass die Ladeinfrastruktur in einer Weise erstellt wird, die die Umwelt minimiert. Dies kann durch den Einsatz von erneuerbaren Energien für das Laden von Elektrofahrzeugen und durch die Reduzierung von Abfällen und Emissionen während des Ladevorgangs erreicht werden.

Konzept

Das Demonstrationswerkzeug soll die Möglichkeit bieten eine Fahrsimulation eines beliebigen Elektrofahrzeuges auf einer Karte zu ermöglichen. Dabei sind Batteriekapazität/Reichweite, die CO2-Erparnis und die Höchstgeschwindigkeit der Fahrzeuge von Bedeutung.
Das Demonstrationswerkzeug besteht aus einer Software/Webanwendung und einem physischen Aufbau. Die Software ermöglicht dabei ein Monitoring der Fahrt. Mit Eingaben eines Anwenders soll die Ausgabe auf der Webanwendung und auf einem Demonstrator bzw. Tischaufbau sichtbar sein, wo beispielsweise ein Miniaturfahrzeug gefahren wird.
Der Anwender soll ein Elektrofahrzeug wählen und den Start- und Endpunkt für die Fahrt auswählen. Die Software berechnet, dass die Fahrt mit der Reichweite des Fahrzeugs entweder:

- Möglich ist: Voraussichtliche Strecke und Dauer wird angezeigt und Fahrt startet, wobei während der Fahrt der Systemzustand (Batteriezustand) des Fahrzeugs sich ändert.
- Nicht möglich ist: Strecke und Dauer mit einem Zwischenstopp bei einer Ladestation wird angezeigt oder man nutzt ein Fahrzeug in der Nähe(E-Scooter), um das Ziel zu erreichen.

Sobald die Eingaben gemacht sind, soll die Fahrt des Fahrzeugs in der Webanwendung und auf dem Tischaufbau sichtbar gemacht werden. In Abb. 2 ist eine Route mit einem Elektrofahrzeug dargestellt. Sichtbar wäre dazu beispielhaft die Fahrzeugauswahl, die voraussichtliche Strecke und Dauer, der Routenverlauf inklusive einer Ladepause mit Ladezeit und die verschiedenen Ladestände während der Fahrt.

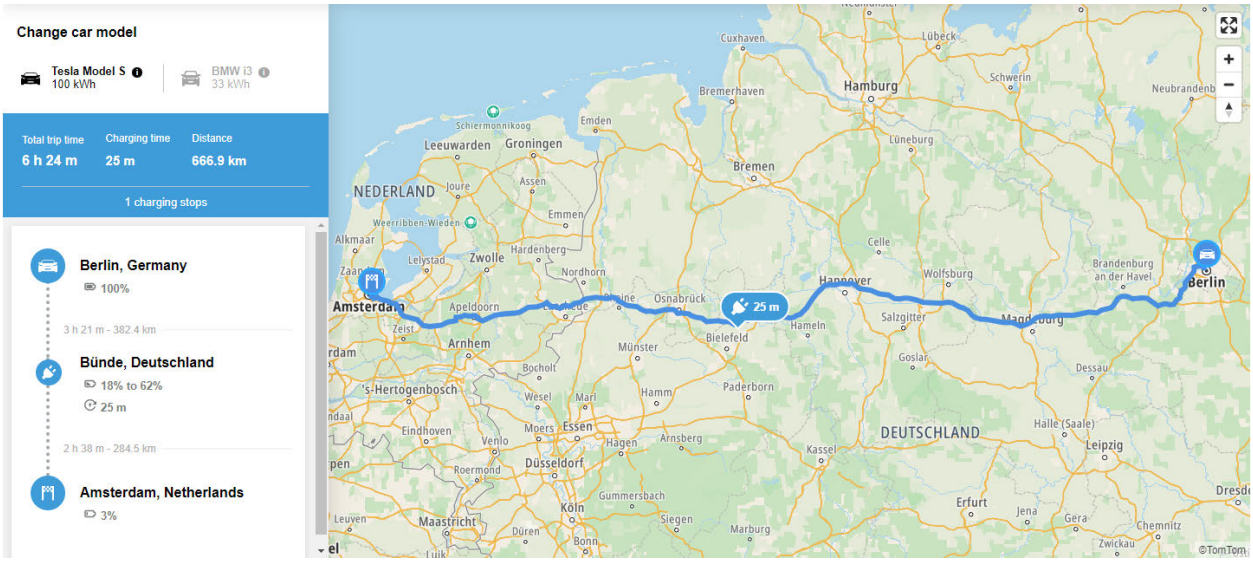


Abb. 2: Beispiel für eine Route mit einem Elektrofahrzeug [5]

Ausblick

Im Rahmen der Arbeit muss noch das Konzept des Demonstrationswerkzeugs umgesetzt werden. Dabei soll zunächst die Fahrsimulation und das Monitoring implementiert werden, welche voraussichtlich als Webanwendung umgesetzt werden. Im Anschluss ist die Umsetzung eines Tischaufbaus mithilfe von Arduinos vorgesehen, welche mit der Fahrsimulation in der Webanwendung vernetzt sein soll.

Literatur und Abbildungen

[1] . Bundesnetzagentur. Anzahl der öffentlichen Ladepunkte in Deutschland von Januar 2017 bis Juli 2023. https://de.statista.com/statistik/daten/studie/1190896/umfrage/ladesaeulen-in-deutschland/, 2023.

[2] Sarah Detzler. *Lademanagement für Elektrofahrzeuge*. KIT Scientific Publishing, 2017.

[3] Bundesministerium für Wirtschaft und Klimaschutz. Elektromobilität in Deutschland. https://www.bmwk.de/Redaktion/DE/Dossier/elektromobilitaet.html, 2022.

[4] Deutsche Kommission Elektrotechnik Elektronik. Ladeinfrastruktur Elektromobilität: Der Technische Leitfaden für Installation und Betrieb in der Praxis. https://www.dke.de/de/arbeitsfelder/mobility/technischer-leitfaden-ladeinfrastruktur-elektromobilitaet, 2023.

[5] . TomTom. Map examples - Multi-stop EV routing. https://developer.tomtom.com/maps-sdk-web-js/functional-examples#examples,map,ev-stations.html, 2022.

[6] Presseamt und Informationsamt der Bundesregierung. So funktioniert der Ausbau der Ladeinfrastruktur. https://www.bundesregierung.de/breg-de/aktuelles/ausbau-ladeinfrastruktur-2165204, 2023.

Design und Entwicklung eines interaktiven Dashboards zur Visualisierung und Analyse von Fahrzeugkomponentendaten im Kontext der Entwicklung von KI-Prädiktionsmodellen

Pembe Yilmaz **Jürgen Koch**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma Dr. Ing. h.c. F. Porsche AG, Weissach

Einleitung

Die wachsende Bedeutung der prädiktiven Instandhaltung in einer von Technologie getriebenen Welt betont die Effektivität dieser Strategie zur Optimierung der Betriebszeit von Anlagen und zur Minimierung ungeplanter Ausfallzeiten. Diese Arbeit konzentriert sich auf die prädiktive Instandhaltung, insbesondere darauf, potenzielle Schäden an Fahrzeugkomponenten frühzeitig zu erkennen.
Die Forschung zielt darauf ab, die gesammelten Sensordaten moderner Fahrzeuge eingehend zu analysieren und präzise Eigenschaften sowie die Qualität dieser Daten zu definieren. Dabei liegt ein besonderer Fokus auf der sorgfältigen Analyse der Daten, sowohl hinsichtlich ihrer Eigenschaften als auch ihrer Qualität. Dies ermöglicht die Extraktion relevanter Kategorien wie Zeitreihendaten, statistische Verteilungen und Muster innerhalb der Daten. Diese Erkenntnisse dienen als Grundlage für die Entwicklung eines interaktiven Tableau-Dashboards, das innovative Möglichkeiten für die Visualisierung und Analyse von Fahrzeugkomponentendaten im Kontext der KI-Prädiktionsmodellentwicklung bietet.
Die zentrale Forschungsfrage lautet: "Wie kann aufgrund von Dateneigenschaften und Datenqualität sowie Analyse-Zielbildern (RUL, AD, PR) eine sinnvolle Vorauswahl für Visualisierungsmethoden getroffen werden?"Durch die Beantwortung dieser Frage wird ein tieferes Verständnis für die Auswahl geeigneter Visualisierungstechniken im Kontext der prädiktiven Instandhaltung geschaffen.

Predictive Maintenance

Predictive Maintenance hat zum Ziel, Probleme an Maschinen frühzeitig zu erkennen, indem kontinuierlich Daten mittels Sensoren gesammelt und mittels maschinellen Lernens analysiert werden. Das Hauptziel besteht in der Minimierung ungeplanter Ausfallzeiten, der Maximierung der Produktionskapazität sowie der Optimierung der Instandhaltungskosten. Durch eine vorausschauende Planung von Instandhaltungsmaßnahmen aufgrund frühzeitiger Erkennung von potenziellen Störfällen wird die generelle Zuverlässigkeit und Langlebigkeit von Anlagen und Maschinen erhöht. [1] Um die prädiktive Instandhaltung effektiv umzusetzen, werden verschiedene Ziele verfolgt, darunter Remaining-Useful-Life (RUL) Prediction, Anomaly Detection und Maintenance Need Classifier. Diese Ziele nutzen überwachte und unüberwachte maschinelle Lernverfahren, um den Zustand von Maschinen zu bewerten und Instandhaltungsbedarf vorherzusagen.

Klassifizierung von Predictive Maintenance

Ein Predictive Maintenance (PdM) Modell verfolgt drei Hauptziele: Remaining-Useful-Life (RUL) Prediction, Anomaly Detection und Maintenance Need Classifier. Die RUL-Prädiktion verwendet überwachte ML-Regressionsmethoden, um die erwartete Restnutzungskapazität eines Objekts auf der Grundlage der bisherigen Nutzung und von Zustandsbeobachtungen zu bestimmen. Die Anomaliedetektion verwendet unüberwachte ML-Verfahren und überwacht Ausreißer einzelner Merkmale sowie Unregelmäßigkeiten unabhängig von Ausfallbeobachtungen. Die Klassifikation des Instandhaltungsbedarfs bewertet den aktuellen Zustand oder einen simulierten zukünftigen Zustand hinsichtlich der Notwendigkeit einer Instandhaltungsmaßnahme unter Verwendung von überwachten ML-Klassifikationsalgorithmen. [6]
Predictive Maintenance ist eine datengesteuerte Instandhaltungsstrategie, die auf fortschrittlichen Sensoren und künstlicher Intelligenz basiert.
Die Vorgehensweise bei der Technologie Predictive Maintenance wird in der folgenden Abbildung dargestellt:

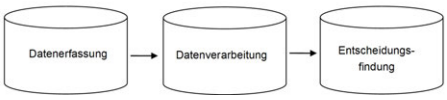


Abb. 1: Vorgehensweise bei Predictive Maintenance [3]

Bei der Analyse und Interpretation der komplexen Daten, die durch prädiktive Instandhaltung generiert werden, spielen Datenvisualisierung und ihre Grundregeln eine entscheidende Rolle. Effektive Visualisierungen erleichtern das Verständnis und die Interpretation komplexer Informationen.

Grundregeln der Datenvisualisierung

Effektive Datenvisualisierung zeichnet sich durch präzise Darstellung aus, damit Nutzer komplexe Informationen schnell erfassen können. Interaktive Elemente ermöglichen es, spezifische Informationen hervorzuheben und die Komplexität zu reduzieren. Visualisierungswerkzeuge bieten verschiedene Darstellungen an, die jedoch mit der Datenbeschaffenheit in Einklang stehen sollten. Der Grundsatz der "datengetriebenen Visualisierung"betont das harmonische Zusammenspiel von Daten und Darstellungsform.
Für aussagekräftige Visualisierungen ist es entscheidend, Erkenntnisse aus Wahrnehmungs- und Gestaltpsychologie zu nutzen. Der Prozess erfordert eine aktive Interaktion zwischen Betrachter und Informationen. Eine gründliche Datenanalyse und Verständnis sind vor der Visualisierung unerlässlich, einschließlich der Identifikation der Zielgruppen, besonders bei statischen Grafiken. Die Integration interaktiver Funktionen ermöglicht Nutzern die Anpassung der Visualisierung an individuelle Anforderungen.
Der Prozess der Datenvisualisierung wird durch Funktionen der Interaktivität erweitert. Interaktive Elemente ermöglichen es den Nutzern, spezifische Informationen hervorzuheben und die Visualisierung an individuelle Anforderungen anzupassen. Diese Funktionen tragen dazu bei, die Komplexität zu reduzieren und dennoch präzise Einblicke zu gewinnen.

Funktionen von Visualisierungen und Interaktivät

Ein weiterer Faktor neben der Informationsverarbeitung ist die Aufmerksamkeit, die durch Visualisierungen geweckt wird. Laut Brigham ist es nicht überraschend, dass Bilder und Grafiken auf mehr Interesse stoßen, was dazu führen kann, das Erinnerungsvermögen und Verständnis des Betrachters zu verbessern. [2] Des Weiteren kann dies durch Farben und einzigartige Darstellungen befähigt werden. In diesen Aufführungen spielt die Funktion der Kommunikation eine Rolle, wodurch das systemtheoretische Vorgehen belegt werden kann. Kim et al. gehen darauf folgendermaßen ein: „Visualizations are now widely used across disciplines to understand and communicate data". [5]
In einer datengetriebenen Geschäftsumgebung ist die Qualität der analysierten Daten von entscheidender Bedeutung. Die verschiedenen Dimensionen der Datenqualität werden daher umfassend untersucht, um eine gründliche Bewertung zu ermöglichen und Verzerrungen in den Analyseergebnissen zu minimieren.

Dimensionen der Datenqualität

Die Bedeutung der Datenqualität in der Datenanalyse steht im Fokus dieser Bachelorarbeit, da sie die Grundlage für zuverlässige und aussagekräftige Erkenntnisse bildet. In einer datengetriebenen Geschäftsumgebung, in der fundierte Entscheidungen auf Datenanalyse basieren, wird die Notwendigkeit einer umfassenden Untersuchung der Datenqualität immer offensichtlicher.
Die Datenanalyse zielt darauf ab, Muster, Zusammenhänge und Trends zu identifizieren, um wertvolle Einblicke zu gewinnen. Eine unzureichende Datenqualität kann jedoch zu Verzerrungen, Inkonsistenzen und Ungenauigkeiten in den Analyseergebnissen führen, was wiederum zu Fehlinterpretationen und Beeinträchtigungen der Effektivität geschäftlicher Entscheidungen führt.
Es wird insbesondere die Auswirkungen der Datenqualität auf verschiedene Analysebereiche untersucht, darunter Genauigkeit, Abstammung, strukturelle und semantische Konsistenz, Vollständigkeit, Konsistenz, Aktualität, Pünktlichkeit, Vernünftigkeit und Identifizierbarkeit (siehe Abbildung 2). Dieser Ansatz ermöglicht eine umfassende Bewertung und Analyse der Dimensionen der Datenqualität. [5]

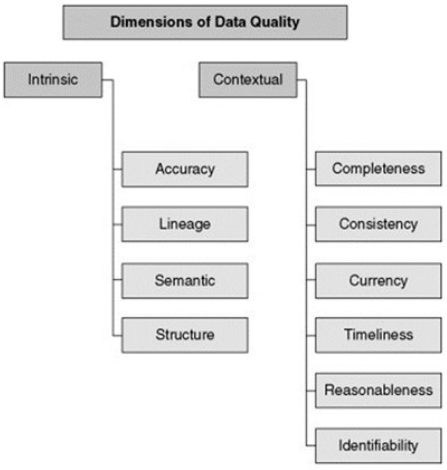


Abb. 2: Praktische Dimensionen der Datenqualität [4]

Zusammenfassung und Ausblick

Der Ausblick dieser Arbeit betont die Integration der entwickelten KI-Prädiktionsmodelle und des interaktiven Dashboards in die industrielle Praxis. Hierbei spielt die präzise Auswahl von Visualisierungsmethoden eine entscheidende Rolle, um die Effektivität und Benutzerfreundlichkeit des Dashboards sicherzustellen. Zukünftige Forschung könnte sich darauf konzentrieren, die entwickelte Methodik weiter zu verfeinern und auf andere Industriezweige anzuwenden, um die Generalisierbarkeit und den breiteren Nutzen dieser Ansätze zu evaluieren.

Literatur und Abbildungen

[1] Felix Beutler, Ute Brümmer, Stephan Ertner, Dirk Evenson, Ralph Obermauer, and Wolfgang Schroeder. *Transformation der Automobilindustrie: Was jetzt zu tun ist*. böll.brief, 2021.

[2] T. J. Brigham. *Feast for the eyes: an introduction to data visualization*. Taylor & Francis, 2016.

[3] Eigene Darstellung.

[4] Bernd Heinrich, Mathias Klier, Alexander Schiller, and Gerit Wagner. *Assessing data quality: A probability-based metric for semantic consistency.* Elsevier, 2018.

[5] Nam Wook et al. Kim. *Accessible visualization: Design space, opportunities, and challenges*. The Eurographics Association, 2021.

[6] Xiaosheng et al. Si. *Remaining useful life estimation - A rewiev on the statistical data driven approaches.* European journal of operational research, 2011.

Evaluierung von SAP Luigi und anderen Micro Frontend Frameworks zur Optimierung der Entwicklungsprozesse durch Analyse der Effizienz und Flexibilität

Leto Ziegler **Jörg Nitzsche**

Fakultät Informatik und Informationstechnik, Hochschule Esslingen

Arbeit durchgeführt bei der Firma QUANTO Solutions GmbH, Stuttgart

Einleitung

In der heutigen Zeit werden Webanwendungen immer komplexer und viele Programme werden ins Web gebracht. Mit Blick auf sehr große Anwendungen, wie die SAP Business Technology Platform oder die Amazon Web Services Oberfläche, ist es empfehlenswert, den monolithischen Ansatz für diese zu verwerfen. Im Backend würde man das durch Micro Services realisieren. Im Frontend kann man das durch so genannte Micro Frontends realisieren, die in einem Orchestrator dann gebündelt werden und so eine Anwendung für die Nutzer darstellen, wie in Abbildung 1 zu sehen ist.

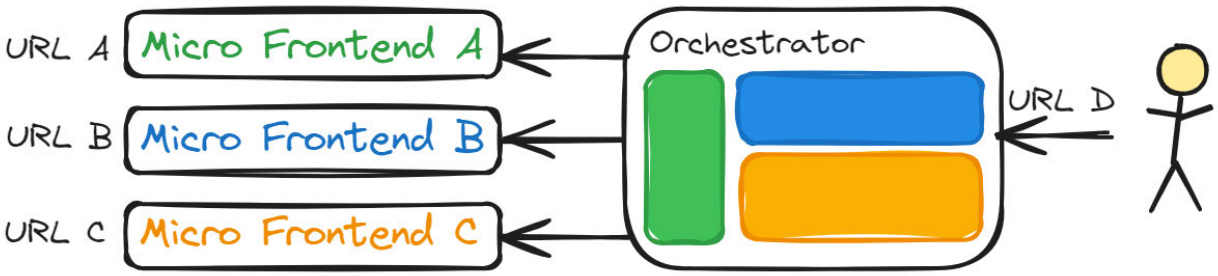


Abb. 1: Micro Frontend Architekturbeispiel [1]

2019 hat SAP das Micro Frontend Framework Luigi auf den Markt gebracht, welches den Entwicklern etwas Arbeit abnehmen soll, indem es die Einbindung vereinfacht, Komponenten und ein UI-Design bereitstellt und eine API zur Kommunikation mit den Micro Frontends bereitstellt. [5]

Zielsetzung

Das Ziel dieser Arbeit ist es, folgende Punkte unter den Aspekten der Effizienz, Flexibilität und der Entwicklungsprozesse zu evaluieren. Zuerst gilt es zu untersuchen, welchen Mehrwert Micro Frontends für moderne Frontend-Entwicklungsprojekte liefern. Um die Vor- und Nachteile von Luigi zu anderen Micro Frontend Frameworks zu bewerten, ist es naheliegend, einen Vergleich durchzuführen. Dafür wird die Module Federation [3] für den Vergleich herangezogen. Und als dritter Punkt wird noch geklärt, wie sich die Entwicklungsprozesse durch Micro Frontend Frameworks optimieren können. Dabei wird ein besonderes Augenmerk auf die Koordination der Entwicklerteams gelegt. Es wird evaluiert, welche Entwicklungsmethoden Sinn ergeben, und es wird beobachtet, wie sich die Projektpflege verändert.

Arbeitsansatz

Um den Mehrwert der Micro Frontend Architektur bestimmen zu können werden vor allem Ergebnisse verschiedener wissenschaftlicher Arbeiten zusammengetragen. Dabei lassen sich auch Parallelen zu den Micro Services schlagen, denn Micro Frontends sind praktisch eine Weiterentwicklung der Microservice Architektur für die Frontend Entwicklung [4]. Um die Frameworks miteinander zu vergleichen, wird eine Micro Frontend Anwendung erstellt, welche in einen Orchestrator eingebunden wird. Das Micro Frontend wird mit Svelte

gebaut. Dieses Micro Frontend wird in die Luigi Applikation eingebunden, indem es von dort aufgerufen wird. Dafür wird im Micro Frontend Gebrauch von der Luigi-Client Bibliothek gemacht. Der Luigi Orchestrator soll dann für das Micro Frontend Authentifizierungsdaten bereitstellen. Für den Vergleich wird eine React-Anwendung erstellt, welche über Module Federation das Micro Frontend einbinden wird. Auch hier werden für den Vergleich Authentifizierungsdaten an das Micro Frontend gegeben. Des Weiteren wird auch verglichen, wie sich Ausfälle des Micro Frontends unbehandelt in den Anwendungen verhalten und was man jeweils tun sollte, um Ausfälle richtig zu behandeln. Für das Micro Frontend selbst wird ein Backend gebaut, welches Daten durch eine GraphQL API zur Verfügung stellt, wie in Abbildung 2 zu sehen ist.

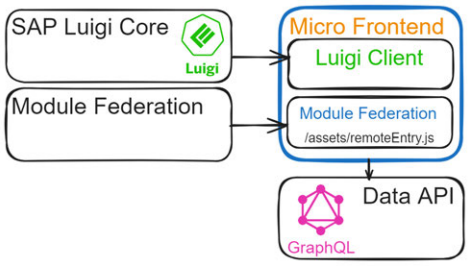


Abb. 2: Vergleichsarchitektur [1]

Wie sich die Entwicklungsprozesse verändern, also welche Vor- und welche Nachteile sich ergeben, wird durch Recherche und durch das Nachstellen der Behaviour Driven Development Methode, kurz BDD, ermittelt.
Mehrwert von Micro Frontends Das Micro Frontend Konzept bricht monolithische Web-Anwendungen auf. Dadurch bekommen die Entwicklerteams die Möglichkeit, Web-Anwendungen gezielter zu skalieren, diese flexibler in der Entwicklung zu machen und durch die Aufteilung in Teilprojekte auch einfacher zu Warten. Diese Aufteilung birgt auch ein paar Probleme. Eins davon ist, eine einheitliche Designsprache zwischen den Micro Frontends zu halten. Je nachdem wie das Micro Frontend eingebunden ist, ist darauf zu achten eine einheitliche Designsprache durchzusetzen. Denn das Ignorieren davon kann dazu führen, dass Designs der anderen Micro Frontends überschrieben werden. [2] Dazu muss man jedes Micro Frontend hosten, was Ressourcen und Geld kostet.

Luigi versus Module Federation

Luigi implementiert Micro Frontends in dem so genannten ‚Luigi-Core' in einem iFrame. Dabei wird über eine Konfigurationsdatei das Micro Frontend referenziert. Die ‚Core'-Anwendung, verwaltet die Micro Frontends. Dabei bietet die Bibliothek Möglichkeiten eine Authentifizierung schnell einzubauen, diese für die Micro Frontends zugänglich zu machen, dritt Anbieter Cookies richtig in der ‚Core'-Anwendung anzulegen und zu verwalten und über die ‚Luigi-Client'-Bibliothek in den Microfrontends ergibt sich die Möglichkeit, dass das Micro Frontend mit dem Orchestrator Luigi-Core Kommunizieren kann. [5] Module Federation bietet eine im Code direkt integrierte Lösung, bei welcher die einzubindenden Komponenten wie normale Komponenten verwendet werden. Dabei wird der Code der Komponenten geteilt und nicht einfach nur aufgerufen. Vorgefertigte Lösungen zur Authentifizierung gibt es hier nicht, diese müssen extra implementiert werden. Beide Frameworks bieten die Möglichkeit Micro Frontends technologieagnostisch einzubinden. In Luigi ist das, weil das Micro Frontend in einem iFrame angezeigt wird. In der Module Federation werden mit den Komponenten auch die notwendigen Bibliotheken geteilt, die wichtig sind, damit die Komponenten funktionieren. Wenn man jetzt zum Beispiel ein Svelte Micro Frontend in React einbetten möchte, kann man dies tun, indem man die geteilte Komponente In einer React Komponente einbettet.

Ausblick

Wenn man eine monolithische Anwendung in kleine Micro Frontends aufteilt und diese alle ein eigenes Entwicklerteam haben, stellt sich natürlich die Frage, wie man die Koordination unter den Teams handhabt. Grundsätzlich kann man alle möglichen Entwicklungsansätze verfolgen. Behaviour Driven Development, zu kurz BDD, ist für Micro Frontends ein interessanter Ansatz. BDD ist dem Test Driven Development ähnlich. Es basiert darauf, dass die Anforderungen für die Komponenten in Form von Verhaltensbeschreibungen formuliert werden. Auf diese Anforderungen werden dann Tests erstellt. Wichtig ist hierbei natürlich, dass die Anforderungen alle Verhaltensmöglichkeiten abdecken und das alle Stakeholder miteinander kommunizieren. Diese Ansätze werden nach Möglichkeit verfolgt und simuliert und zusätzlich durch Recherche untersucht.

Literatur und Abbildungen

[1] Eigene Darstellung.

[2] Cam Jackson. Micro Frontends. https://martinfowler.com/articles/micro-frontends.html#Styling, 2019.

[3] Zack Jackson. Understanding Module Federation: A Deep Dive. https://scriptedalchemy.medium.com/understanding-webpack-module-federation-a-deep-dive-efe5c55bf366, 2023.

[4] Andrey Pavlenko, Nursultan Askarbekuly, Swati Megha, and Manuel Mazzara. Micro-frontends: application of microservices to web front-ends. https://jisis.org/wp-content/uploads/2022/11/jisis-2020-vol10-no2-04.pdf, 2020.

[5] Autorenkollektiv Unbekannt. Getting Startetd. https://docs.luigi-project.io/docs/getting-started, 2023.