Binaire

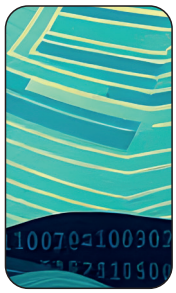
DAS MAGAZIN DES FORSCHUNGSZENTRUMS L3S

WWW.BINAIRE.DE

AUSGABE Nº
10/11111101000

L3S

EXZELLENZ IN DER KI-FORSCHUNG



Prompt an *Black Forest Labs' Flux Schnell:* »A futuristic lighthouse painted in the style of classical modernist art, the beacon's light, depicted as radiating concentric lines, creates a neural network pattern in the sky, illuminating an abstract landscape of binary code, expressionist style oil painting, thick layers of colourful textured paint«.

ECHT KÜNSTLICH

Diese Ausgabe der *Binaire* steckt voller menschlicher und künstlicher Intelligenz. Letztere vor allem als Gegenstand der hier vorgestellten wissenschaftlichen Studien.
KI ist aber nicht nur das zentrale Thema, sondern auch tägliches Werkzeug – beim Auswerten der Forschungsergebnisse, beim Erstellen von Berichten sowie beim Übersetzen, Schreiben und Gestalten dieses Magazins, das weiterhin von Menschen gemacht wird.
Ist exakt zwei Jahre nach der Veröffentlichung von *ChatGPT* eine funktionierende Symbiose von Mensch und Maschine entstanden? »Je mehr ich weiß, umso mehr weiß ich, dass ich nichts weiß«, soll Aristoteles gesagt haben. Mit wachsender Erfahrung wächst auch die Erkenntnis der Grenzen, Herausforderungen und Arbeitsbereiche. Die menschliche Reflexion bleibt die Basis für Exzellenz in der KI-Forschung, die technische Präzision und ethische Verantwortung vereint – mit Technologien, die das menschliche Potenzial erweitern. Idealerweise in einem inspirierendem Zusammenspiel von Wissen, Verantwortung und Vision.

L3S
Trustworthy AI
& Digital Transformation

DAS FORSCHUNGSZENTRUM *L3S*

Das *L3S* betreibt international renommierte Grundlagenforschung zur künstlichen Intelligenz und entwickelt anwendungsorientierte KI-Methoden, ins besondere für Medizin, Mobilität, Produktion und Bildung.
Forschende am *L3S* entwickeln Methoden und Technologien für den digitalen Wandel, erforschen die Auswirkungen der Digitalisierung und leiten daraus Handlungsoptionen, Empfehlungen und Innovationsstrategien für Wirtschaft, Politik und Gesellschaft ab.
Das *L3S* ist eine gemeinsame Einrichtung der Leibniz Universität Hannover und der Technischen Universität Braunschweig. Mitglieder sind mehr als 30 Professorinnen und Professoren aus unterschiedlichen Disziplinen und Universitäten. Insgesamt arbeiten rund 200 Wissenschaftlerinnen und Wissenschaftler am *L3S*.

Exzellenz in der KI-Forschung

Liebe Leserin, lieber Leser,

Exzellenz ist kein Zufall, sondern das Ergebnis harter Arbeit, kreativen Denkens und eines unermüdlichen Strebens nach Innovation. Das *Forschungszentrum L3S* steht seit Jahren für wissenschaftliche Spitzenleistungen in der KI-Forschung – ein Anspruch, der nicht nur in der Qualität seiner Projekte, sondern auch in der Anerkennung durch die internationale Forschungsgemeinschaft sichtbar wird.
In dieser Ausgabe der *Binaire* würdigen wir einige beispielhafte Publikationen des letzten Jahres, die auf den renommiertesten Konferenzen und in den angesehensten Fachjournalen veröffentlicht wurden. Sie wurden in ihren Kategorien jeweils mit der Auszeichnung *L3S Best Publication of the Quarter* ausgezeichnet.
Jede dieser Arbeiten trägt dazu bei, Grenzen zu verschieben und drängende Fragen der künstlichen Intelligenz zu beantworten. Sie sind ein Zeugnis für die Kreativität, den Arbeitseinsatz und das Fachwissen der Forschenden am *L3S*.
Diese ausgezeichneten Arbeiten repräsentieren nicht nur technologische Innovationen, sondern verdeutlichen auch, wie Exzellenz in der KI-Forschung Lösungen für reale Probleme schaffen kann – von der Analyse komplexer Datenstrukturen bis hin zur Gestaltung sicherer und fairer Algorithmen.
Ich lade Sie ein, sich von der Vielfalt dieser Beiträge inspirieren zu lassen und auch durch diese *Binaire* wieder einen Einblick in die Forschung am *L3S* zu bekommen. Sie sind ein Spiegel unserer Mission, KI nicht nur weiterzuentwickeln, sondern sie verantwortungsvoll und nachhaltig zu gestalten.

Viel Spaß beim Lesen wünscht Ihnen

Prof. Dr. techn. Wolfgang Nejdl

»Unsere Ergebnisse zum föderierten Lernen zeigen, dass es möglich ist, Fairness während des Trainingsprozesses einzubauen, ohne die Leistung des Modells zu beeinträchtigen – ein bedeutender Schritt in Richtung eines ethischen und fairen Einsatzes von KI-Technologien. «

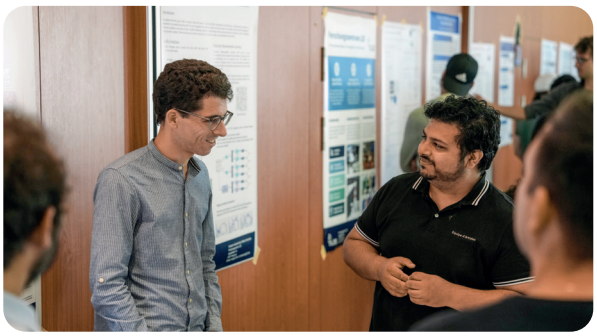
MARYAM BADAR, M. SC.

forscht am *L3S* zu Fairness von KI-Technologien.

ÜBERSICHT

BINAIRE-AUSGABE 2 / 2024

MELDUNGEN



Oben: Zahlreiche Forschende des *L3S* trafen sich zum jährlichen **Research Retreat**. Rechts: Fachgespräche bei der Poster-Session. —› Fotos: *L3S*

Workshop: Große Sprachmodelle in der Bildung

Rund 50 Interessierte trafen sich Anfang November zum Workshop »Large Language Models in der Bildung«, den das *L3S* gemeinsam mit der *TIB* organisiert hatte. Ziel der Veranstaltung war es, den Einsatz von Large Language Models (LLMs) in Schulen und Hochschulen zu beleuchten. Es ging um Chancen und Herausforderungen der LLMs für Lehrende und Lernende und um ungenutzte Potenziale. Im Mittelpunkt stand dabei der interdisziplinäre Austausch: Experten aus Informatik, Bildungsforschung, Didaktik und Psychologie präsentierten neueste Erkenntnisse und Anwendungsmöglichkeiten von LLMs: vom KI-generierten Podcast über einen personalisierten Tutor bis zum Leseassistenten. In der abschließenden Diskussion der Ergebnisse wurde ein Spannungsverhältnis besonders deutlich: Einerseits müssen die Lehrkräfte die Entwicklung der LLMs aufgrund der enormen Dynamik ständig im Auge behalten. Andererseits sind LLMs derzeit noch nicht in der Lage, alle Anforderungen für den Einsatz im Unterricht umzusetzen – wie etwa die Anpassungsfähigkeit, auf einzelne Lernende individuell einzugehen.

L3S ist Gold Member

CAIRNE Das *L3S* ist bereits seit längerem Mitglied, nun wurde es zum *Gold Member* der *Confederation of Laboratories for Artificial Intelligence Research in Europe (CAIRNE).* Die Organisation widmet sich der Förderung der KI-Forschung und -Innovation in ganz Europa. Seit seiner Gründung im Jahr 2018 arbeitet *CAIRNE* daran, die europäische Exzellenz im Bereich KI zu stärken. Nach dem Vorbild des *CERN* will *CAIRNE* eine Infrastruktur aufbauen, die KI-Forscher in ganz Europa miteinander verbindet sowie den Wissensaustausch und die Zusammenarbeit zwischen Spitzenforschungseinrichtungen fördert.

L3S Research Retreat 2024

Ende Oktober trafen sich zahlreiche Wissenschaftler und Wissenschaftlerinnen des *L3S* zum jährlichen Research Retreat, der dieses Jahr wieder in Goslar stattfand. An drei Tagen tauschten die *L3S*-Forscher Ideen aus und stellten ihre neuesten Arbeiten vor. Die Veranstaltung umfasste eine Poster-Session und Präsentationen zu aktuellen Forschungsthemen, darunter hybride KI-Methoden und Sprachmodelle, und zum KI-Transfer. Vier Teams wurden vor Ort für ihre herausragenden Publikationen in hochrangigen Zeitschriften und auf renommierten Konferenzen mit dem Preis *L3S Best Publication of the Quarter* ausgezeichnet. Schwerpunkt des Retreats war ein Hackathon, bei dem neun Teams nach zweieinhalb Tagen Zusammenarbeit beeindruckende Ergebnisse präsentierten. ¶

ZUR WEBSITE der Organisation *CAIRNE* → https://cairne.eu

MEHR ZUM WORKSHOP → https://www.l3s.de/de/workshop-large-language-models-in-education/

CAIMed
Niedersächsisches Zentrum für KI und Kausale Methoden in der Medizin

Künstliche Intelligenz in der Medizin

PRÄZISION, VERTRAUEN UND KAUSALITÄT

Die medizinische Diagnostik ist ein Paradebeispiel für Multimodalität. Ärztinnen und Ärzte stützen Entscheidungen auf eine Vielzahl von Datenquellen: von radiologischen Bildern über Laborbefunde bis hin zu Patientengesprächen. Künstliche Intelligenz (KI) gewinnt dabei zunehmend an Bedeutung. Besonders in der Radiologie hilft KI, komplexe Muster zu erkennen und diagnostische Prozesse zu optimieren.

KI IN DER RADIOLOGIE

KI-Algorithmen zeigen ihre Stärke vor allem bei der Bildanalyse, Detektion und Klassifikation. Im Mammographie-Screening unterstützen sie Ärzte dabei, subtile Anomalien zu identifizieren und erhöhen so die diagnostische Genauigkeit. Auch bei der Erkennung von Pneumothorax, einer gefährlichen Luftansammlung zwischen Lunge und Brustwand, erreichen KI-Systeme beeindruckende Ergebnisse mit über 90 Prozent Genauigkeit. Doch der Erfolg hat Grenzen: Fehlalarme können die Produktivität senken, wenn vermeintliche Anomalien sich als harmlos herausstellen. Trotz solcher Schwächen ist die Radiologie führend im medizinischen KI-Einsatz. Über 700 zugelassene Softwareprodukte verbessern die Diagnostik, von der Tumorerkennung bis zur Lokalisierung von Frakturen.

GENERATIVE KI IN DER KLINISCHEN PRAXIS

Mit großen Sprachmodellen wie *ChatGPT-4* und *MedPaLM2* ergeben sich neue Chancen für Diagnostik und klinische Entscheidungsunterstützung. In Prüfungen wie dem *United Kingdom Medical Licensing Exam* erzielte *ChatGPT-4* bis zu 94 Prozent korrekte Antworten, in den US-amerikanischen *Licensing Exams (USMLE)* 89 Prozent. *MedPaLM2* erreichte in der MedQA-Benchmark-Prüfung eine Genauigkeit von 86,5 Prozent. Diese Ergebnisse zeigen das Potenzial von KI, das Niveau menschlicher Expertise in spezifischen Aufgaben zu erreichen. Allerdings gibt es auch Bedenken: Die Modelle wurden perfekt auf Prüfungsfragen trainiert. Das bedeutet, sie könnten in unvorhergesehenen Szenarien versagen. Zusätzlich neigen generative Transformer wie *ChatGPT* zu sogenannten Halluzinationen – sie liefern falsche, aber plausibel wirkende Antworten. In der Medizin könnten solche Fehler schwerwiegende Folgen haben.

DIE GRENZEN KLASSISCHER KI-ANSÄTZE

Klassische maschinelle Lernmethoden erkennen zwar Muster und Korrelationen, sind jedoch selten in der Lage, Ursache-Wirkungs-Beziehungen zu analysieren. Diese »kausalen« Fragen sind jedoch zentral für klinische Entscheidungen: Etwa, wie sich das Risiko einer Patientin durch die Einnahme eines bestimmten Medikaments verändert oder welche Behandlungsstrategie in einem spezifischen Fall die besten Erfolgsaussichten bietet.

Hoher Besuch bei *CAIMed:* Im September informierte sich der niedersächsische Gesundheitsminister Dr. Andreas Philippi zusammen mit dem niedersächsischen Bundestagsabgeordneten Dr. Christos Pantazis über die Chancen von künstlicher Intelligenz (KI) in der Gesundheitsversorgung. Das Fachgespräch mit Experten von *CAIMed* und der Initiative *Diabetes@Work* fand an der *Medizinischen Hochschule Hannover* statt. —› Foto: *L3S*

MEHR INFOS
zum Forschungszentrum
und seinen Vorhaben
→ https://caimed.de

CAIMED

KAUSALITÄT UND VERTRAUEN: DAS CAIMED-PROJEKT

Wissenschaftler des niedersächsischen Forschungszentrums *CAIMed* arbeiten daran, diese Lücke zu schließen. Die Nachwuchsgruppe »KI & Kausalität« untersucht Modelle, die sowohl auf Daten als auch auf Ursache-Wirkungs-Beziehungen basieren. »Unser Ziel ist es, KI-Systeme zu entwickeln, die nicht nur präzise und repräsentativ sind, sondern auch Vertrauen schaffen«, sagt Prof. Wolfgang Nejdl vom *Forschungszentrum L3S*, der neben Prof. Markus Cornberg von der *Medizinischen Hochschule Hannover* Mentor der Nachwuchsgruppe ist. Mit *CAIMed* entsteht ein interdisziplinäres Netzwerk, das KI-Methoden für die Behandlung von Volkskrankheiten wie Krebs, Herz-Kreislauf-Erkrankungen und Infektionen erforscht. Unter Leitung von Nejdl arbeiten rund 100 Forschende der KI und Medizin aus Hannover, Göttingen und Braunschweig zusammen, um vertrauenswürdige Technologien zu entwickeln.

CAIMed setzt auf die Verknüpfung von Forschungs- und Versorgungsdaten, um personalisierte Medizin zu ermöglichen. Das Projekt baut auf wichtigen Vorarbeiten des *Leibniz KI-Labors für Personalisierte Medizin* und der Medizininformatikinitiative auf. Gefördert wird *CAIMed* aus dem Programm *zukunft.niedersachsen* des *Niedersächsischen Ministeriums für Wissenschaft und Kultur* mit Mitteln der *VolkswagenStiftung*.

Die Ziele sind klar definiert: vertrauenswürdige, menschzentrierte KI-Lösungen für eine bessere Gesundheitsversorgung.

FORSCHUNG FÜR DIE ZUKUNFT

In 13 interdisziplinären Nachwuchsgruppen untersuchen Forschende KI-Ansätze in den Bereichen Semantik, Entscheidungen, Wirkstoffe und Signale. Dabei steht nicht nur technologische Exzellenz im Vordergrund, sondern auch der Aufbau von Vertrauen – bei Ärzten und Pflegekräften sowie bei Patienten und deren Angehörigen. *CAIMed* zeigt, wie KI und kausale Methoden das Gesundheitssystem nachhaltig verbessern können. Es geht nicht nur um technische Fortschritte, sondern um das Fundament einer humaneren, effektiveren Medizin. ¶

KONTAKT:

Dr. Johannes Winter
winter@L3S.de

\\\ Johannes Winter ist Geschäftsführer von *CAIMed* sowie stellvertretender Geschäftsführer und Chief Strategy Officer des *L3S*. \\\

Es sind sehr viele und klar definierte Facetten, die aus Forschenden exzellente Forschende machen. Prompt an *Midjourney* in geduldiger Zusammenarbeit mit *ChatGPT4.o:* »A portrait-format collage in sea colours, on the left side with a human figure from science stylised into a geometric prism. A single white ray of light enters the prism and splits into seven rays that spread out to the right. Each ray ends in shapes such as sparks, overlapping circles, waves, balanced scales, branching trees and faceted crystals. The background is bright with overlays of scientific and social motifs from the fields of mobility, medicine, production and education.«

FORSCHUNG AM *L3S*

Ziele, Richtlinien, Beispiele

Exzellente Forschung bildet das Fundament wissenschaftlicher Innovationen und Fortschritte, die unsere Gesellschaft nachhaltig prägen. Wonach strebt Forschung am *L3S*? Was sind unsere Ziele und Methoden? Was sind Beispiele für exzellente Forschung aus dem *L3S*?
Einer unserer frühen Mentoren, Gio Wiederhold, bis 2001 Professor für Datenbanken an der *Universität Stanford* und in den Gründungsjahren des *L3S* wichtiges Mitglied im wissenschaftlichen Beirat, gab uns in den frühen Jahren des *L3S* einige Fragen mit, die George Heilmeyer, ehemaliger Direktor der US-amerikanischen Forschungsbehörde *ARPA*, vor etwa 50 Jahren als Evaluationskriterien für Projektanträge formuliert hatte, und die in der Folge in vielen anderen Kontexten genutzt und als *Heilmeyer's Catechism* bekannt wurden. Wir geben sie in der Formulierung von Gio Wiederhold wieder. Sie eignen sich (gerne ergänzt, je nach Disziplin, durch weitere Fragen) sowohl als Evaluationskriterien für Forschungsprogramme und Projektanträge ebenso wie (etwas abgewandelt) als Leitlinien für Promotionen, wissenschaftliche Arbeiten und wissenschaftliche Publikationen. Aus diesem Grund sind sie wichtiger Bestandteil unserer Forschungsphilosophie und unserer *PhD Mentoring Guidelines*.

HEILMEYER'S CATECHISM

1. *What is the problem, why is it hard?*
   Was ist das Problem, warum ist es schwierig?
2. *How is it solved today?*
   Wie wird es heute gelöst?
3. *What is the new technical idea; why can we succeed now?*
   Was ist die neue technische Idee; warum können wir jetzt erfolgreich sein?
4. *What is the impact if successful?*
   Was sind die Auswirkungen, wenn es erfolgreich ist?
5. *How will the program be organized?*
   Wie wird das Programm organisiert?
6. *How will intermediate results be generated?*
   Wie werden Zwischenergebnisse generiert?
7. *How will you measure progress?*
   Wie soll der Fortschritt gemessen werden?
8. *What will it cost?*
   Was wird es kosten? ↗

Ein einfacher, aber sehr wirkungsvoller Fragenkatalog, der in jedem Forschungsantrag beantwortet werden sollte. Gehen wir im Folgenden etwas detaillierter auf einige dieser Aspekte und unsere Forschungsphilosophie ein.

ORIGINALITÄT UND RELEVANZ

Exzellente Forschung setzt an innovativen Fragestellungen an, die bislang unbeantwortet sind und gleichzeitig eine hohe Relevanz für Wissenschaft und Gesellschaft haben. Wir wollen wirtschaftliche und gesellschaftliche Herausforderungen mit technologischen Innovationen verknüpfen, die Probleme auf neuartige Weise, besser als bisher, lösen.

METHODIK, TRANSPARENZ, VORGEHEN

Methodik und Forschungsergebnisse müssen nachprüfbar sein. Besonders in der KI, wo Modelle oft komplex und schwer nachvollziehbar sind, ist Transparenz von besonderer Bedeutung, und nicht nur das Ergebnis, sondern auch der nachvollziehbare Weg zu seiner Herleitung ist Teil exzellenter Forschung. In einer Zeit, in der alle verfügbaren Daten zum Training großer KI-Modelle genutzt werden, stehen wir vor der zusätzlichen Herausforderung überhaupt wissenschaftlich sauber evaluieren zu können. Das *Forschungszentrum L3S* legt großen Wert auf die offene Bereitstellung von Daten, Algorithmen und Ergebnissen, um Replizierbarkeit und die Überprüfbarkeit durch die wissenschaftliche Gemeinschaft zu gewährleisten. Dazu ist vor Beginn der Arbeiten eine Planung mit konkreten Zwischenzielen und messbarem Fortschritt in allen Phasen des Projektes hilfreich und wichtig.

INTERDISZIPLINÄRE ZUSAMMENARBEIT

KI als neue Grundlagendisziplin (wir erinnern uns an die vor kurzem vergebenen Nobelpreise für Physik und Chemie an Grundlagenforscher aus der Künstlichen Intelligenz) erfordert oft die Zusammenarbeit mit Wissenschaftlerinnen und Wissenschaftlern anderer Disziplinen: Informatik, Ethik und Recht etwa bei der Entwicklung von Algorithmen und neuen Ansätzen; Informatik und wichtige Anwendungsgebiete, etwa Maschinenbau, Medizin oder Bildung, bei der Entwicklung effizienter und maßgeschneiderter Lösungen.

EINFLUSS AUF DAS WISSENSCHAFTLICHE, WIRTSCHAFTLICHE UND GESELLSCHAFTLICHE UMFELD

Exzellente Forschung hat nicht nur in der Wissenschaft Bedeutung, sondern auch einen klaren Einfluss auf Wirtschaft und Gesellschaft. KI hat das Potenzial, eine Vielzahl von Bereichen – von der Gesundheitsversorgung bis zur Mobilität – grundlegend zu verändern. Ein am *L3S* entwickeltes

Modell zur Analyse von Gesundheitsdaten könnte dazu beitragen, bessere Behandlungsentscheidungen zu treffen und letztendlich die Gesundheitsversorgung zu verbessern (Seite 21). Die Studie zu *LaMMOn* (Seite 24) zeigt, dass innovative Technologien wie Sprachmodelle und neuronale Graph-Netzwerke den Weg für effizientere Verkehrssysteme ebnen können. Die Forschung des *L3S* zur digitalen Epidemiebekämpfung (Seite 20) verdeutlicht, dass soziale Netzwerke nicht nur ein Ort der Kommunikation sind, sondern auch eine wertvolle Datenquelle für das Krisenmanagement. Wissenschaftlerinnen und Wissenschaftler des *L3S* arbeiten auch daran, mit Hilfe großer Sprachmodelle den teils allzu rauen Ton in den sozialen Netzwerken zu entschärfen, um produktivere Diskussionen zu ermöglichen (Seite 18). Bei einigen Methoden des maschinellen Lernens, wie dem Deep Reinforcement Learning, ruhen große Hoffnungen auf Embodied AI. Aber bevor es in kritischen Anwendungen, wie industriellen Systemen oder autonomen Fahrzeugen, vollumfänglich genutzt werden kann, müssen noch Fragen zu Effizienz, Generalisierbarkeit und Robustheit gelöst werden. Auch daran arbeitet das *L3S* (Seite 16).

ETHISCHE VERANTWORTUNG

Im Bereich der KI ist ethische Verantwortung ein zentraler Faktor. Systeme, die auf KI basieren, beeinflussen zunehmend wichtige Entscheidungen in unserem täglichen Leben. Jedoch gibt es noch viel Forschungsbedarf, bevor etwa generative KI-Systeme in kritischen Gebieten wie der medizinischen Praxis breiten Einsatz finden können. Eine aktuelle Studie des *L3S* hat ergeben, dass *GPT-4* zwar beeindruckende Fähigkeiten bei der Beantwortung medizinischer Fragen aufweist, aber bisweilen auch noch mit großer Überzeugung falsche Empfehlungen und Antworten gibt (Seite 22).
Das *Forschungszentrum L3S* setzt auf verantwortungsbewusste Forschung und betont die Bedeutung von fairen und transparenten Algorithmen. Wir berücksichtigen auch die ethischen Herausforderungen von KI und entwickeln Ansätze, um Bias, also Verzerrungen, in Modellen zu minimieren und gleichzeitig sicherzustellen, dass KI-Systeme transparent und nachvollziehbar bleiben. Dies ist von entscheidender Bedeutung, um das Vertrauen der Gesellschaft in diese Technologien zu stärken und einen fairen Einsatz zu garantieren. Die Studien des *L3S* zu Hybridsystemen gegen Verzerrungen in Lernmodellen (Seite 12) und zu FairTrade (Seite 14) zeigen eindrucksvoll, dass es möglich ist, KI-Systeme fair und leistungsfähig zugleich zu gestalten.

NACHHALTIGE WISSENSVERMITTLUNG UND AUSBILDUNG

Exzellente Forschung geht über die reine Publikation von Artikeln hinaus – sie beinhaltet auch die Vermittlung von Wissen an die nächste Generation von Wissenschaftlern. Das *L3S* leistet einen wichtigen Beitrag zur Ausbildung junger Forscher, indem es Programme zur Förderung des wissenschaftlichen Nachwuchses anbietet. Um den Austausch zu fördern, organisiert das *L3S* Workshops, Seminare und internationale Konferenzen wie die *ACM International Conference on Web Search and Data Mining*, die im März 2025 in Hannover stattfinden wird. Dies trägt dazu bei, unseren Nachwuchs mit Spitzenforschern aus der ganzen Welt zu vernetzen, zu inspirieren und Innovationen in allen Bereichen zu beschleunigen. ¶

KONTAKT:

Prof. Dr. Marius Lindauer
marius.lindauer@l3s.de

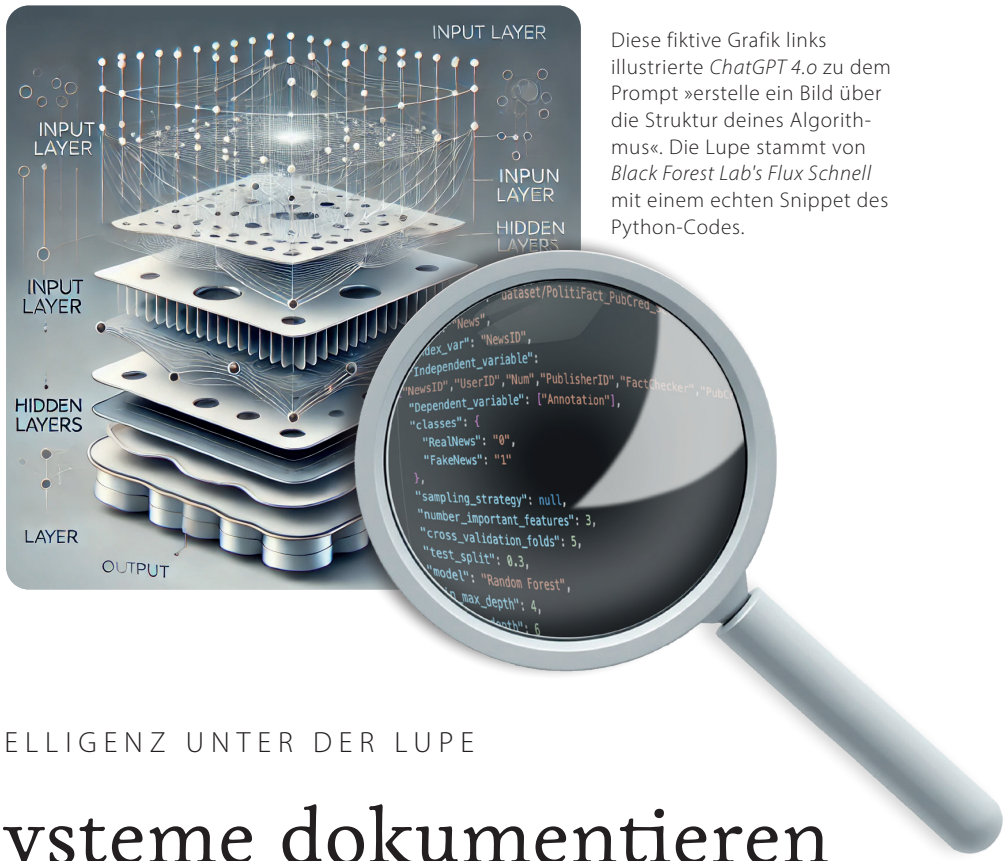
\\\ *L3S*-Mitglied Marius Lindauer leitet das Institut für Künstliche Intelligenz der *Leibniz Universität Hannover.* \\\

KONTAKT:

Prof. Dr. Wolfgang Nejdl
nejdl@L3S.de

\\\ Wolfgang Nejdl ist geschäftsführender Direktor des *Forschungszentrums L3S.* \\\

ZUR WEBSITE
L3S Best Publications:
→ https://www.l3s.de/category/best-publications



Diese fiktive Grafik links illustrierte *ChatGPT 4.o* zu dem Prompt »erstelle ein Bild über die Struktur deines Algorithmus«. Die Lupe stammt von *Black Forest Lab's Flux Schnell* mit einem echten Snippet des Python-Codes.

KÜNSTLICHE INTELLIGENZ UNTER DER LUPE

Hybridsysteme dokumentieren Verzerrungen in Lernmodellen

Künstliche Intelligenz (KI) setzt sich im täglichen Leben immer mehr durch. Damit stellt sich die Frage: Wie können die negativen Auswirkungen dieser Technologie abgemildert werden? Ein weit verbreitetes Problem in KI-Systemen sind Verzerrungen. In einer aktuellen Studie stellt ein Forschungsteam des *L3S* und der *Leibniz Universität Hannover* ein hybrides KI-System vor, das Verzerrungen in Modellen des maschinellen Lernens (ML) dokumentiert. Die innovative Technik kann dazu beitragen, die Transparenz und Interpretierbarkeit dieser komplexen Systeme zu verbessern.
Verzerrungen in KI-Modellen können aus unterschiedlichen Gründen auftreten, da die Systeme während des Entwicklungsprozesses anfällig sind für menschliche Eingaben. Außerdem werden diese Systeme oft auf Daten trainiert, die gesellschaftliche Ungleichheiten widerspiegeln. Ihr Einsatz würde also bestimmte Gruppen systematisch benachteiligen.

KOMBINATION VON KI-PARADIGMEN

Die Autoren schlagen ein hybrides KI-System vor, das Komponenten der subsymbolischen und der symbolischen KI kombiniert, um Verzerrungen auf jeder Stufe der ML-Pipeline zu dokumentieren.
Das System kann mit zwei Ansätzen implementiert werden: Der erste ermöglicht eine feinkörnige Verfolgung von Verzerrungen über die gesamte ML-Pipeline hinweg; der zweite Ansatz bietet einen breiteren Überblick über erkannte Verzerrungen in den Eingabedaten und Vorhersagen.
Ein Schlüsselelement dieses Systems: Die Dokumentation ist nicht nur für menschliche Analytiker verständlich, sondern auch maschinenlesbar. Dies bildet die Grundlage für eine bessere Interpretierbarkeit und ein besseres Verständnis dafür, wie sich Verzerrungen auf ML-Systeme in einem bestimmten Kontext auswirken.

VERZERRUNGEN BEI DER KLASSIFIZIERUNG VON FAKE NEWS

Die Wissenschaftler haben ihren Ansatz anhand eines praktischen Beispiels evaluiert, das auf der Erkennung von Fake News basiert. Das hybride KI-System *Doc-Bias* beschrieb semantisch eine Pipeline zur Klassifizierung von Fake News anhand zweier Benchmark-Datensätzen, die die Verteilung von Nachrichteninhalten, Nutzerdaten und Informationen von Nachrichtenherausgebern verwendeten. Die Implementierung von *Doc-Bias* generierte dann Bias-Spuren, die das Innenleben des Klassifizierungssystems widerspiegeln – von der Eingabe bis zur Ausgabe. Die Autoren konnten signifikante Verzerrungen in den Datensätzen identifizieren, die die Vorhersagen der Modelle beeinflussten.

Die Ergebnisse dieser Arbeit zeigen, dass selbst unter der Voraussetzung ausgewogener Datensätze während der Datenvorverarbeitung die internen Prozesse des KI-Modells eine attributorientierte Verzerrung ausgleichen können, die sich erheblich auf die Gesamtgenauigkeit und Wirksamkeit des Fake-News-Erkennungssystems auswirkt. Die Autoren berichten konkret über eine starke Schieflage in der Verteilung der Eingabevariablen in Richtung des Fake-News-Labels und deckten auf, wie eine prädiktive Variable zu mehr Einschränkungen im Lernprozess führt. Insgesamt heben sie die offenen Herausforderungen beim Training von Modellen mit unausgewogenen Datensätzen hervor.

Dieses Problem ist nicht auf die Erkennung von Fake News beschränkt. Verzerrungen in KI-Systemen können zu systematischen Fehlern in fast allen Bereichen der KI-Anwendung führen: von der automatischen Gesichtserkennung bis hin zu Entscheidungsfindungssystemen im Bankwesen.

TRANSPARENZ UND VERANTWORTUNG SIND UNERLÄSSLICH

Das vorgeschlagene hybride KI-System ist ein neuartiges Verfahren, das erkannte Verzerrungen in KI-Modellen systematisch dokumentiert und die menschliche Analyse bei späteren Bemühungen um eine Entschärfung der Verzerrungen unterstützen kann. Solche Fortschritte sind entscheidend, um sicherzustellen, dass KI-Systeme nicht nur genau sind, sondern auch fairer und transparenter werden.

Die Autoren betonen, dass die Dokumentation von Voreingenommenheit kein Allheilmittel ist, sondern nur einer von vielen Schritten, um Verantwortlichkeit bei KI-Entwicklern und -Nutzern zu schaffen. Zu einer sozial verantwortlichen KI gehört auch Transparenz im gesamten Entwicklungsprozess. Die Erstellung einer gründlichen Dokumentation kann dazu beitragen. ¶

KONTAKT:

Mayra Russo, M. Sc
mrusso@L3S.de

\\\ *L3S*-Doktorandin Mayra Russo beschäftigt sich mit der Entwicklung computergestützter Methoden zur Dokumentation von Verzerrungen in KI-Systemen, die semantische Datenmodelle verwenden. Außerdem interessiert sie sich für die Untersuchung der sozialen Implikationen der Datafizierung. \\\

KONTAKT:

Prof. Dr. Maria-Esther Vidal
vidal@L3S.de

\\\ *L3S*-Mitglied Maria-Esther Vidal ist ordentliche Professorin an der *Leibniz Universität Hannover* und leitet die Arbeitsgruppe *Wissenschaftliches Datenmanagement (SDM)* an der *TIB – Leibniz-Informationszentrum für Technik und Naturwissenschaften*. Sie forscht in den Bereichen Datenmanagement, semantische Datenintegration und maschinelles Lernen über Wissensgraphen. \\\

Mayra Russo, Yasharajsinh Chudasama, Disha Purohit, Sammy Sawischa, Maria-Esther Vidal: Employing Hybrid AI Systems to Trace and Document Bias in ML Pipelines. IEEE Access 12: 96821-96847 (2024)
→ https://ieeexplore.ieee.org/document/10596297

VERANTWORTUNGSVOLLE KI

Chancengleichheit beim föderierten Lernen

Modelle der künstlichen Intelligenz werden traditionell mit zentral gespeicherten Datensätzen trainiert. Dieser Ansatz birgt jedoch erhebliche Herausforderungen im Umgang mit sensiblen Daten, insbesondere im Hinblick auf den Datenschutz und die Datensicherheit. Föderiertes Lernen (FL) bietet eine sichere Alternative, indem Modelle auf dezentralisierten Daten trainiert werden, wodurch sichergestellt wird, dass sensible Informationen lokalisiert bleiben. Trotz seiner Vorteile besteht eine der größten Herausforderungen bei FL darin, ein optimales Gleichgewicht zwischen Fairness und Genauigkeit der Modellleistung zu erreichen. Wissenschaftler des *Forschungszentrums L3S* haben sich in einer aktuellen Studie mit dieser Herausforderung befasst und eine innovative Lösung namens *FairTrade* vorgestellt. Dieser Ansatz verbessert die Fairness bei gleichzeitiger Aufrechterhaltung eines hohen Genauigkeitsniveaus in föderierten Lernanwendungen und ebnet so den Weg für gerechtere und zuverlässigere KI-Systeme.

GENAUIGKEIT UND GERECHTIGKEIT IM EINKLANG

Föderiertes Lernen ermöglicht das Training eines gemeinsamen KI-Modells mit Daten mehrerer Geräte, zum Beispiel Smartphones oder Tablets, ohne dass die Daten von den Geräten übertragen werden. Dieser Ansatz verbessert den Datenschutz und die Datensicherheit. Er bringt jedoch auch Herausforderungen mit sich, da unterschiedliche Daten auf den verschiedenen Geräten zu Unterschieden in der Leistung des Modells führen können. So können beispielsweise unterrepräsentierte Datensätze zu verzerrten oder ungerechten Vorhersagen des KI-Modells führen. »*FairTrade* zielt darauf ab, diese Diskriminierung bei Vorhersagen zu minimieren«, sagt Maryam Badar, Hauptautorin der Studie. »Durch den Einsatz von Mehrzieloptimierung versucht es, einen optimalen Kompromiss zwischen Modellgenauigkeit und Fairness zu erreichen.« Der Rahmen lässt sich je nach den Anforderungen der Anwendung an ver-

Föderiertes Lernen (Federated Learning, FL) ist ein Ansatz im Bereich des maschinellen Lernens, bei dem ein gemeinsames KI-Modell auf verteilten Daten trainiert wird, die sich auf verschiedenen Geräten oder Servern befinden. Dabei bleiben die Daten lokal auf den Geräten, und nur die Modellparameter werden zwischen den Geräten und einem zentralen Server ausgetauscht. Das Bild erstelle *Midjourney* zu dem Prompt »photo of a group of students sitting casually in a room with a large computer, all holding smartphones and tablets, bright lines show digital networking, lock symbols refer to security.«.

schiedene Fairness-Konzepte anpassen, darunter statistische und kausale Fairness.
Experimente mit realen Datensätzen aus Bereichen wie Bank-, Personal- und Gesundheitswesen haben bemerkenswerte Ergebnisse gezeigt. *FairTrade* verbesserte die Fairness in allen Szenarien, ohne die Genauigkeit zu beeinträchtigen. Selbst bei stark unausgewogenen Datensätzen erwies es sich als zuverlässige Alternative zu bestehenden Methoden.

VIELE ANWENDUNGSBEREICHE

FairTrade hat das Potenzial für eine breite Anwendung in unterschiedlichen Bereichen. In jedem Kontext, in dem KI-Systeme personalisierte Entscheidungen treffen, kann dieser Ansatz dazu beitragen, gerechtere und ausgewogenere Ergebnisse zu erzielen.
In der **medizinischen Diagnostik** könnte *FairTrade* beispielsweise Verzerrungen ausgleichen, die sich aus dem Training von Modellen mit ungleichmäßig verteilten Patientendaten ergeben. Durch die Abschwächung dieser Verzerrungen kann die Methode die Diskriminierung von Minderheitengruppen im Gesundheitswesen verringern, was letztlich zu genaueren Diagnosen und besseren Behandlungsergebnissen führt.
Ähnlich verhält es sich bei der **Kreditvergabe,** wo sich die Banken zunehmend auf KI stützen, um die Kreditwürdigkeit ihrer Kunden zu bewerten. *FairTrade* kann dazu beitragen, eine voreingenommene Entscheidungsfindung zu verhindern, die bestimmte Gemeinschaften ungerechtfertigt benachteiligt. Dies gewährleistet ein integrativeres und gerechteres Finanzsystem.
Die Forscher betonen, dass Fairness von Anfang an in KI-Systeme integriert werden muss. Post-hoc-Anpassungen der Ergebnisse reichen nicht aus, um Fairness umfassend zu berücksichtigen. »Unsere Ergebnisse zeigen, dass es möglich ist, Fairness während des Trainingsprozesses einzubauen, ohne die Leistung des Modells zu beeinträchtigen«, so Badar. Dies ist ein bedeutender Schritt in Richtung eines ethischen und fairen Einsatzes von KI-Technologien. ¶

KONTAKT:

Maryam Badar, M. Sc.
badar@L3S.de

\\\ Maryam Badar ist Doktorandin am *L3S*. Sie forscht zu Fairness von KI-Technologien. \\\

Maryam Badar, Sandipan Sikdar, Wolfgang Nejdl, Marco Fisichella: FairTrade: Achieving Pareto-Optimal Trade-Offs between Balanced Accuracy and Fairness in Federated Learning. AAAI 2024: 10962-10970
→ https://ojs.aaai.org/index.php/AAAI/article/view/28971

DEEP REINFORCEMENT LEARNING
FÜR DIE REALE WELT

Lernen mit Struktur

Deep Reinforcement Learning (RL) – dieser Zweig des maschinellen Lernens befasst sich mit KI-Systemen, die durch Interaktion mit der Welt lernen, sequenzielle Entscheidungen zu treffen. RL hat in einigen Bereichen schon bemerkenswerte Erfolge erzielt: von komplexen Strategien in Spielen wie Go über mehrere Handlungssequenzen in der simulierten Robotik bis hin zur Feinabstimmung großer Sprachmodelle. Dennoch bleibt sein Einsatz in der realen Welt begrenzt, da es mit Herausforderungen wie ineffizienter Datennutzung, mangelnder Sicherheit und eingeschränkter Generalisierbarkeit konfrontiert ist. Eine Studie des *Forschungszentrums L3S* und der *University of Texas at Austin* zeigt, wie die Einbettung problemspezifischer Strukturinformation die Leistungsfähigkeit und Skalierbarkeit von RL-Systemen grundlegend verbessern kann.

GRUNDLEGENDE HERAUSFORDERUNGEN ÜBERWINDEN

»Einige der größten Herausforderungen für RL ergeben sich aus der Unvorhersehbarkeit realer Szenarien«, sagt Aditya Mohan, Hauptautor der Studie. RL-Algorithmen scheitern oft an dynamischen Umgebungen oder verrauschten Belohnungssignalen. Herkömmliche RL-Modelle lernen in der Regel durch Trial-and-Error, um extrinsische Belohnungen zu maximieren. Dieses Verfahren ist nicht nur datenintensiv, sondern schränkt auch die Übertragbarkeit der Modelle auf neue Aufgaben stark ein. Ein Roboter, der in einer Simulation darauf trainiert wurde, eine blaue Tasse aufzuheben, könnte beispielsweise versagen, wenn sich die Farbe der Tasse ändert. Die Einschränkung steht in krassem Gegensatz zum menschlichen Lernen. Im Wesentlichen entwickeln Kinder ein generelles Verständnis ihrer Umwelt, das sie aufgabenspezifisch anwenden können. Dagegen werden RL-Algorithmen darauf trainiert, implizit gerade so viel über die Welt zu lernen, dass sie die vom menschlichen Designer vorgegebene extrinsische Belohnung optimieren können. Um solche Algorithmen auf Veränderungen einzustellen, müssten spezifische Belohnungen für einzelne Problemvarianten definiert werden.

STRUKTURELLE INFORMATIONEN EINBEZIEHEN

Die Autoren plädieren dafür, **zusätzliche strukturelle Informationen** in die Modelle zu integrieren. Ein Beispiel: Ein RL-Agent, der ein Taxi in einer Stadt steuert, müsste durch bloße Interaktion das gesamte Straßennetz, Verkehrsverhalten und Passagierbewegungen lernen – eine nahezu unlösbare Aufgabe. Mit struktureller Information, etwa der Trennung von Verkehrs- und Passagiermustern, kann der Lernprozess effizienter und zielgerichteter gestaltet werden.
Der Ansatz macht sich die Möglichkeit zunutze, komplexe Probleme in handhabbare Teilkomponenten zu zerlegen. Die Autoren haben recherchiert, inwieweit verschiedene RL-Methoden eine solche Zerlegbarkeit annehmen, und anschließend einen Rahmen entwickelt, um diese Annahmen zu kategorisieren. Die Studie identifiziert

Deep Reinforcement Learning kann beispielsweise in der automatisierten Lagerhaltung genutzt werden. Mithilfe von Sensoren navigieren Roboter selbstständig, vermeiden Kollisionen und arbeiten koordiniert mit anderen zusammen. Ein Belohnungssystem verstärkt effiziente Aktionen, während Fehler wie Verzögerungen oder Kollisionen bestraft werden, um die Abläufe in der dynamischen Umgebung stetig zu verbessern. Das Bild erstellte *DALL-E* zu *Chat-GPT's* Prompt: »A photorealistic image of a large automated warehouse, captured as if taken with a high-quality DSLR camera. The warehouse features autonomous robots resembling rolling shelves with forklift-style arms, actively picking, lifting, and transporting goods. Real humans in high-visibility vests are working in the warehouse, supervising the robots. The environment is brightly lit, clean, and highly organized, showcasing sharp details and realism.«

vier grundlegende Archetypen für die Dekomposition komplexer Probleme in RL-Modellen: latent, faktorisiert, relational und modular.

VON ENTWURFS-ENTSCHEIDUNGEN ZU ENTWURFSMUSTERN

RL-Algorithmen unterscheiden sich oft nur durch geringfügige Änderungen an der Standard-RL-Pipeline. Algorithmen, die strukturelle Annahmen verwenden, tun dies in einer wiederholbaren Reihenfolge. Auf der Grundlage dieser Erkenntnis stellen Mohan und seine Co-Autoren ein Rahmenwerk vor, das Entwurfsmuster für die Einbettung von Strukturen in RL-Algorithmen skizziert, einschließlich abstrakter Zustände, faktorisierter Modelle, relationaler Architekturen und modularer Designs. Die Analyse eines breiten Spektrums von RL-Arbeiten durch die Linse der Entwurfsmuster zeigt, welche Kombinationen von Mustern sich für bestimmte Anwendungen als effektiv erwiesen haben – von der Generalisierung bis zur Interpretierbarkeit. Beispielsweise kann ein Roboter durch die **Einbeziehung relationaler Darstellungen** Pakete in einem Lagerhaus effizient sortieren, da er die Beziehungen zwischen den Objekten versteht. In ähnlicher Weise können RL-Agenten durch die Verwendung von Belohnungsmodellen auch in Umgebungen mit wenigen Belohnungssignalen effizient lernen. Dieser strukturierte Ansatz beschleunigt nicht nur die Datenverarbeitung, sondern verbessert auch die Generalisierungsfähigkeit von RL-Agenten.

Die Arbeit öffnet neue Forschungsfelder, etwa die Identifikation optimaler Entwurfsmuster oder Kombinationen davon für unterschiedliche Anwendungen – je nach den **gewünschten Eigenschaften Generalisierbarkeit, Effizienz, Sicherheit oder Interpretierbarkeit.**

»Wir hoffen, dass unser Rahmenwerk als Leitfaden für die Weiterentwicklung von RL-Methoden dienen wird«, sagt Mohan. »Die Verwendung von Strukturen könnte der Schlüssel sein, um RL endlich auf die komplexe reale Welt auszuweiten.« ¶

KONTAKT:

Aditya Mohan, M. Sc.
a.mohan@ai.uni-hannover.de

\\\ Aditya Mohan ist wissenschaftlicher Mitarbeiter am *Forschungszentrum L3S* und am *Institut für Informationsverarbeitung,* Fachgebiet Automatische Bildinterpretation, der *Leibniz Universität Hannover.* \\\

Aditya Mohan, Amy Zhang, Marius Lindauer:
Structure in Deep Reinforcement Learning: A Survey and Open Problems.
J. Artif. Intell. Res. 79: 1167-1236 (2024)
→ https://www.jair.org/index.php/jair/article/view/15703/27028

Toxische Diskussionen auf Plattformen sind ein medial sehr präsentes Thema. Vier verschiedene KI-Bildgeneratoren wurden jeweils mit diesem Prompt aufgefordert, das Thema zu illustrieren (ohne näherere Angaben zu konkreten Bildinhalten oder zum Stil etc.): »Picture of a toxic discussion on an internet platform with an indignant user at his desk, view from the side«. Von links nach rechts sind die – ziemlich stereotypen – Ergebnisse von: *Stable Diffusion XL, DALL-E 3, Flux Schnell* und *Midjourney.*

ONLINE-DISKUSSIONEN

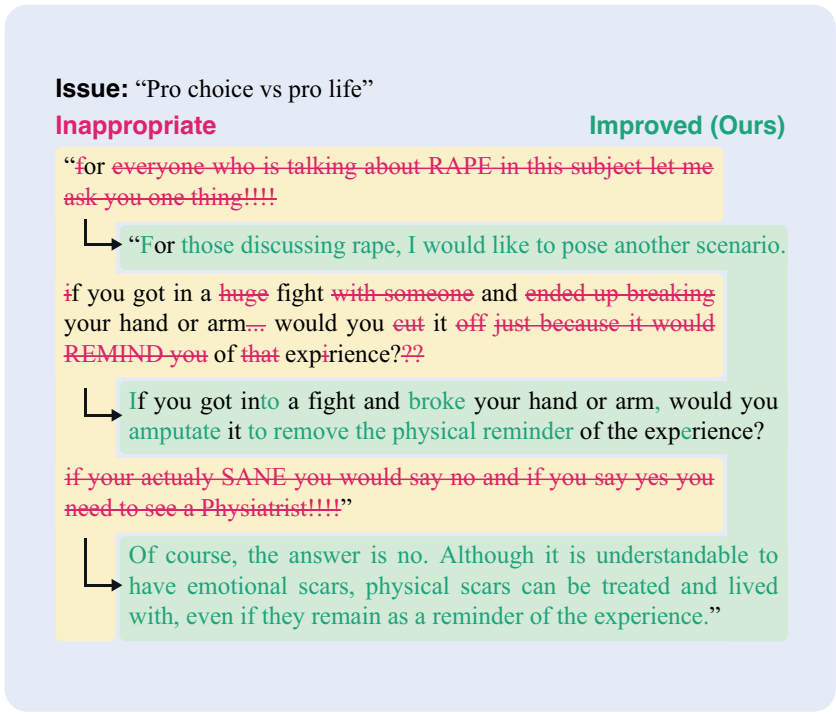
Höflicher mit KI

Diskussionen im Internet sind oft ein heißes Pflaster. Anonymität und emotionale Bindungen führen schnell zu toxischen Kommentaren, die den Diskurs ersticken. Ein Forscherteam des *L3S* zeigt in einer aktuellen Studie, wie künstliche Intelligenz (KI) und Natural Language Processing (NLP) helfen können, solche Argumente umzuformulieren – und zwar so, dass die ursprünglichen Gedanken erhalten bleiben, aber unpassende Äußerungen entfernt werden. Doch wie gelingt dieser Balanceakt?

ENTSCHÄRFEN DURCH UMFORMULIEREN

Bisher verlassen sich Plattformen auf ihre Nutzer oder automatische Erkennungstools, um unangemessene Inhalte zu kennzeichnen, die anschließend noch von Moderatoren geprüft und gegebenenfalls entfernt werden. Doch dieser Ansatz ist nicht nur zeitaufwendig und teuer, sondern für die Moderatoren auch belastend. Die Lösung könnte in einer neuen Technologie liegen: Künstliche Intelligenz, die in der Lage ist, unangemessene Argumente automatisch umzuformulieren. »Unser Ziel ist es, nicht einfach nur zu löschen, sondern den Diskurs zu retten, indem wir Argumente entschärfen und gleichzeitig ihre Aussage erhalten«, sagt Timon Ziegenbein, Erstautor der Studie.

Das Herzstück des Ansatzes ist ein Large Language Model (LLM), das durch Reinforcement Learning (verstärkendes Lernen) trainiert wird. Dabei wird das Modell darauf getrimmt, die Angemessenheit zu steigern und außerdem den ursprünglichen Inhalt eines Arguments zu bewahren. Die KI »lernt« also, unangemessene Argumente so umzuformulieren, dass sie respektvoller und sachlicher werden. »Anders als bei einfachen Stilübertragungsaufgaben, die nur den Tonfall eines Textes ändern, erfordert das Umschreiben unangemessener Argumente auch inhaltliche Änderungen auf Dokumentenebene, nicht nur auf Satzebene«, sagt Ziegenbein. Dies ermöglicht es, Inhalte hinzuzufügen oder zu entfernen, um die Sachlichkeit und Höflichkeit zu erhöhen. Ein Beispiel aus der Studie zeigt die Wirksamkeit: Ein aggressives und emotional aufgeladenes Argument wurde in eine ruhigere, nachdenklichere Version umformuliert – ohne den Kern der Aussage zu verfälschen. Doch wie gut funktioniert dieser Ansatz in der Praxis? Um das herauszufinden, verglichen die Forscher die KI-generierten Umschreibungen mit denen von Menschen. Die Ergebnisse waren vielversprechend: »Unser Modell erreicht eine signifikante Verbesserung der Angemessenheit und



Beispiel für eine Umformulierung, bei der die Argumente entschärft, die Aussagen im Kern aber nicht verändert wurden.

bleibt dabei überraschend nah an der ursprünglichen Aussage«, so Ziegenbein.

GRENZEN UND ETHISCHE FRAGEN

Trotz der Erfolge stößt die KI auch an ihre Grenzen, vor allem bei sehr kurzen oder stark unangemessenen Argumenten. »In solchen Fällen könnte es sinnvoller sein, den gesamten Text zu entfernen, anstatt ihn umzuformulieren«, schreiben die Autoren. Zudem werfen solche Technologien ethische Fragen auf: Darf eine Plattform Inhalte einfach umformulieren, ohne die Zustimmung des Autors? Hier sehen die Wissenschaftler weiteren Forschungsbedarf.

Die Ergebnisse der Studie zeigen jedoch, dass KI ein wertvolles Werkzeug sein könnte, um Online-Diskussionen zivilisierter zu gestalten und Moderatoren zu entlasten. Der nächste Schritt besteht darin, die Technologie weiter zu verfeinern und ethische Richtlinien für den Einsatz solcher Tools zu entwickeln.

»Unser Ansatz ist ein erster Schritt«, so die Forscher, »aber es bleibt noch viel zu tun, um die Technologie in der Praxis sicher und effektiv einzusetzen«. Fest steht: Der Ansatz, toxische Inhalte durch Umschreibungen zu entschärfen, hat Potenzial – sowohl für Plattformen als auch für die Nutzer. ¶

KONTAKT:

Timon Ziegenbein
t.ziegenbein@ai.uni-hannover.de

\\\ Timon Ziegenbein ist wissenschaftlicher Mitarbeiter und Promotionsstudent im *L3S*-Projekt *OASiS: Objective Argument Summarization for Search.* \\\

KONTAKT:

Prof. Dr. Henning Wachsmuth
henning.wachsmuth@L3S.de

\\\ *L3S*-Mitglied Henning Wachsmuth leitet das Fachgebiet *Natural Language Processing* am Institut für Künstliche Intelligenz der *Leibniz Universität Hannover.* \\\

Timon Ziegenbein, Gabriella Skitalinskaya, Alireza Bayat Makou, Henning Wachsmuth: LLM-based Rewriting of Inappropriate Argumentation using Reinforcement Learning from Machine Feedback. ACL (1) 2024: 4455-4476
→ https://aclanthology.org/2024.acl-long.244.pdf

Soziale Netzwerke können mit Hilfe von maschinellem Lernen als Datenquelle für das Krisenmanagement genutzt werden. Das Bild erstelle *Flux* zu dem Prompt »several screens with social media pages flow into a funnel consisting of zeros and ones«.

SOZIALE MEDIEN
ALS DATENQUELLE

Digitale Epidemiebekämpfung

Während der *COVID-19*-Pandemie und früherer Epidemien wie dem Ebola-Ausbruch von 2013 lieferten die sozialen Medien wertvolle Echtzeitdaten: Millionen von Tweets wurden gepostet – mit nützlichen Informationen zu Symptomen, Übertragungswegen und Präventionsmaßnahmen, aber auch mit irrelevanten oder irreführenden Inhalten. Doch wie lässt sich die Datenflut effizient nutzen? Eine aktuelle Studie des *L3S* zeigt, wie KI-gestützte Modelle aus Tausenden von Tweets relevante Inhalte filtern und zusammenfassen können, um durch einen vertrauenswürdigen Ansatz wichtige Erkenntnisse bereitzustellen.

Die in der Studie vorgestellte Methode nutzt moderne maschinelle Lernverfahren, um Tweets automatisch zu klassifizieren und prägnant zusammenzufassen. Im Fokus steht dabei nicht nur die Modellgenauigkeit, sondern auch die Interpretierbarkeit der Ergebnisse. Denn viele der aktuellen KI-Modelle agieren als Blackbox; Nutzer können nicht nachvollziehen, wie die KI zu ihren Entscheidungen kommt.

»Unser Modell extrahiert aus den Tweets Schlüsselinformationen, sogenannte Rationales, um seine Entscheidungen zu erklären«, sagt Thi Huyen Nguyen, Erstautorin der Studie. Darüber hinaus erfassen Rationales wesentliche Inhalte in Tweets. Rationales können zur Erstellung prägnanter Zusammenfassungen der Situation verwendet werden und sind daher besonders wertvoll für Entscheidungsträger, die umgehend auf veränderte Gegebenheiten reagieren müssen.

Die Ergebnisse der Studie sind vielversprechend: Das entwickelte Modell erreichte eine Klassifikationsgenauigkeit von 82 Prozent und übertraf damit herkömmliche Methoden. Darüber hinaus half die vorgeschlagene einfache, graphbasierte Ranking-Methode dabei, die wichtigsten Informationen herauszufiltern und Redundanz zu vermeiden. Die generierten kurzen Zusammenfassungen vermittelten ein umfassendes Bild der Lage während eines Krankheitsausbruchs.

Soziale Netzwerke sind also nicht nur ein Ort der Kommunikation, sondern auch eine wertvolle Datenquelle für das Krisenmanagement – vorausgesetzt, man verfügt über die richtigen Werkzeuge, um die Datenflut zu beherrschen. ¶

KONTAKT:

Dr. Thi Huyen Nguyen
nguyen@L3S.de

\\\ Thi Huyen Nguyen ist wissenschaftliche Mitarbeiterin am *L3S*. Sie forscht zum Einsatz von künstlicher Intelligenz für das Gemeinwohl. \\\

Thi Huyen Nguyen, Marco Fisichella, Koustav Rudra:
A Trustworthy Approach to Classify and Analyze Epidemic-Related Information From Microblogs. IEEE Trans. Comput. Soc. Syst. 11(5): 6229-6241 (2024)
→ https://ieeexplore.ieee.org/abstract/document/10530086

ZEIT-
REIHEN-
MODELLIERUNG

Effiziente Analyse von Patientendaten

Die elektronische Gesundheitsakte oder Patientenakte (englisch Electronic Health Record, EHR) ist ein digitales System, das die Gesundheitsdaten eines Patienten zentral speichert, um den Informationsaustausch zwischen verschiedenen Gesundheitseinrichtungen und Fachleuten zu erleichtern und die Qualität der Versorgung zu verbessern. Das Bild erstelle *Flux* zu dem Prompt »three monitors with two different Electronic Health Records and a small X-ray, professional photography, bokeh, natural lighting, canon lens, shot on dslr 64 megapixels sharp focus«.

Elektronische Gesundheitsakten (EHRs) sind eine wertvolle Datenquelle für die medizinische Forschung. Sie ermöglichen personalisierte Vorhersagen, Diagnosen und Behandlungsempfehlungen. Wissenschaftler des *Forschungszentrums L3S* haben nun IVP-VAE vorgestellt, ein Modell, das eine genaue und effiziente Analyse von EHRs ermöglicht. EHR-Daten werden häufig unregelmäßig erfasst, was bedeutet, dass Messungen in ungleichen Zeitabständen vorgenommen werden, teilweise Datenpunkte fehlen und die Sequenzen in ihrer Länge variieren. Diese Komplexitäten machen die Analyse von EHR-Daten besonders herausfordernd für bestehende Modelle wie rekurrente neuronale Netze (RNNs) und Transformer. »Unsere Arbeit zeigt, dass wir diese Hürden überwinden können, indem wir unregelmäßige Zeitreihen als kontinuierliche Prozesse modellieren und sie als sogenannte Anfangswertprobleme (IVPs) lösen«, sagt Jingge Xiao, Erstautor der Studie.

Das neuartige Modell IVP-VAE kombiniert Variational Autoencoders (VAE) mit IVP-Solvern, um Zeitreihen gleichzeitig und nicht-sequentiell zu verarbeiten. Diese Parallelisierung reduziert die Berechnungszeit erheblich und erlaubt die direkte Modellierung kontinuierlicher Prozesse. Ein weiterer Clou: Durch die Umkehrbarkeit der IVP-Solver können Encoder und Decoder dieselben Rechenprozesse nutzen, was die Modellkomplexität senkt.

»Mit IVP-VAE erreichen wir eine bis zu 40-fache Beschleunigung gegenüber bestehenden Modellen, ohne dabei die Genauigkeit zu opfern«, betont Xiao. Auf drei realen Datensätzen – darunter der bekannte MIMIC-IV-Datensatz – zeigte IVP-VAE konstant bessere Ergebnisse in der Klassifikation und Vorhersage von Zeitreihen. Besonders beeindruckend ist die Leistung bei kleinen Datensätzen, wie sie in der Analyse seltener Erkrankungen häufig vorkommen.

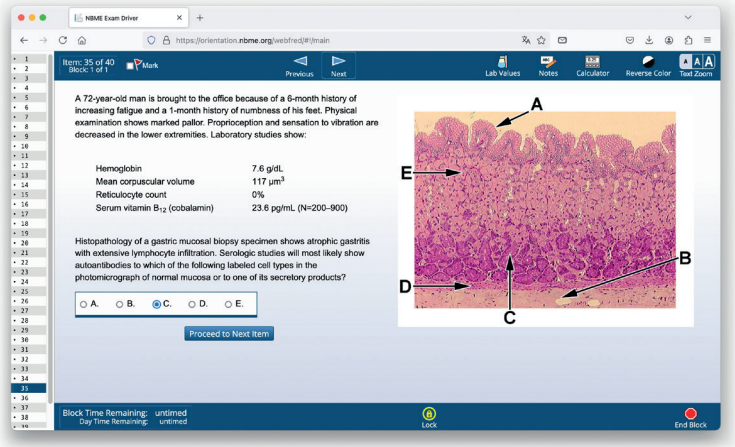
Die Ergebnisse markieren einen wichtigen Schritt in der datenbasierten Gesundheitsforschung. IVP-VAE bietet nicht nur einen performanten Ansatz für EHR-Daten, sondern zeigt auch Potenzial für andere Anwendungen mit unregelmäßigen Zeitreihen. Wie Xiao anmerkt: »Dieses Modell kann die Grundlage für Fortschritte in der personalisierten Medizin legen und die Effizienz in der klinischen Forschung erheblich steigern.« ¶

KONTAKT:

Jingge Xiao, M. Sc.
xiao@L3S.de

\\\ Jingge Xiao ist Doktorand am *L3S*. Er forscht zu Zeitreihenanalyse und Deep Generative Modelling. \\\

Jingge Xiao, Leonie Basso, Wolfgang Nejdl, Niloy Ganguly, Sandipan Sikdar: IVP-VAE: Modeling EHR Time Series with Initial Value Problem Solvers. AAAI 2024: 16023-16031
→ https://ojs.aaai.org/index.php/AAAI/article/view/29534



VERBORGENE FEHLER

Wie *GPT-4* bei medizinischen Prüfungen versagt – und leider trotzdem überzeugt

GPT-4 gilt als das fortschrittlichste Sprachmodell. Obwohl es in der Lage ist, komplexe medizinische Fragen zu beantworten, bleibt es nicht ohne Fehler. In einer aktuellen Studie haben Wissenschaftler des *Forschungszentrums L3S*, des *IIT Kharagpur* und der *University of Michigan* untersucht, welche Fehlerarten *GPT-4* bei medizinischen Prüfungsfragen macht und warum diese teilweise sogar von Fachleuten als »vernünftig« eingestuft werden. Die Untersuchung zeigt, dass *GPT-4* nicht nur durch seine korrekten Antworten überzeugt, sondern auch durch seine Fehler – und dass es noch viel Verbesserungspotenzial gibt.

In der Welt der künstlichen Intelligenz (KI) gibt es viel Hype um *GPT-4*. Das Modell zeigt beeindruckende Leistungen, insbesondere bei der Beantwortung medizinischer Fragen. Im Datensatz *MedQA-USMLE*, der Fragen der US-amerikanischen Medizinlizenzprüfung *(USMLE)* enthält, erreicht *GPT-4* eine bemerkenswerte Genauigkeit von 86,7 Prozent. Doch selbst diese Erfolgsquote hinterlässt 14 Prozent falsche Antworten – keine Kleinigkeit, wenn es um eine medizinische Diagnose geht.

»Wir wollten verstehen, warum *GPT-4* in diesen Fällen falsch liegt«, erklärt Soumyadeep Roy, Doktorand am *IIT Kharagpur* und Hauptautor der Studie.

Das Team erstellte eine Fehler-Taxonomie, die die Antworten von *GPT-4* in sieben Fehlerkategorien einteilt. Dabei wurde insbesondere auf das Reasoning des Modells eingegangen, also auf die Denkprozesse und Schlussfolgerungen.

PLAUSIBEL, ABER FALSCH

In einem aufwendigen Verfahren ließ das Forscherteam 44 medizinische Fachleute insgesamt 300 falsche Antworten von *GPT-4* analysieren. Interessanterweise stellte sich heraus, dass ein Großteil der Fehler als »vernünftige Antwort von *GPT-4*« gewertet wurde. Das zeigt, dass selbst bei falschen Diagnosen die Argumentation von *GPT-4* plausibel klingt – ein großes Problem für Mediziner, die diese Technologie als Unterstützung nutzen wollen.

»Es ist erschreckend zu sehen, dass *GPT-4* oft so überzeugend argumentiert, dass selbst Experten die Fehler nicht sofort erkennen«, betont Co-Autor Uwe Hadler vom *L3S*. Ein häufig beobachteter Fehler: *GPT-4* erkannte zwar die Symptome und interpretierte sie korrekt, stellte aber dennoch eine falsche Diagnose, weil es an der falschen Entscheidung festhielt.

Das Bild links zeigt eine Übungsaufgabe der US-amerikanischen Medizinlizenzprüfung, der *United States Medical Licensing Examination (USMLE).* —› Quelle: https://www.usmle.org/exam-resources/step-1-materials/step-1-sample-test-questions

KI VERTEIDIGT FEHLER

Eine der größten Herausforderungen besteht darin, dass *GPT-4* oft versucht, seine anfängliche Entscheidung zu rechtfertigen, anstatt auf die gegebenen Informationen richtig einzugehen. Dies führt zu Fehlern, die von der KI hartnäckig verteidigt werden. »Wenn *GPT-4* eine Entscheidung getroffen hat, gibt es kein Zurück mehr«, heißt es in der Studie. Trotz dieser Schwächen wird *GPT-4* weiterhin als potenziell wertvolles Werkzeug im medizinischen Bereich gesehen, vor allem wegen seiner Fähigkeit, medizinische Informationen zusammenzufassen und Diagnosen vorzuschlagen. Die Forscher weisen jedoch darauf hin, dass ein detailliertes Verständnis der Fehlerquellen entscheidend ist, um die Technologie weiter zu verbessern und sicherer zu machen.

Ein weiterer Aspekt der Studie war das sogenannte »Drift-Verhalten« von *GPT-4.* Das bedeutet, dass sich die Leistung des Modells über die Zeit verändern kann. »Es ist faszinierend, wie rasant sich die Leistung von *GPT-4* verbessert«, sagt Hadler. Eine Analyse der Antworten von *GPT-4* im Abstand von mehreren Monaten zeigte, dass es in 23,3 Prozent der Fälle weiterhin Fehler macht, die es bereits vorher gemacht hatte. »Das zeigt, dass es noch viel Raum für Verbesserungen gibt.« Die Ergebnisse der Studie sind ein zweischneidiges Schwert: Auf der einen Seite zeigt *GPT-4* beeindruckende Fähigkeiten bei der Beantwortung medizinischer Fragen, auf der anderen Seite müssen die Fehlerarten genau verstanden und adressiert werden, bevor solche Systeme in der medizinischen Praxis breiten Einsatz finden können. Bis dahin bleibt die künstliche Intelligenz ein Werkzeug, das mit Vorsicht zu genießen ist.

»Die Tatsache, dass *GPT-4* auch bei falschen Antworten so überzeugend klingt, ist ein Hinweis darauf, wie schwierig es selbst für Experten sein kann, die Grenzen von KI-basierten Systemen zu erkennen und die richtige Balance zwischen Vertrauen und kritischer Hinterfragung zu finden«,

warnen Roy und Hadler. ¶

KONTAKT:

Soumyadeep Roy, M. Sc.
sroy@L3S.de

\\\ Soumyadeep ist Doktorand am *IIT Kharagpur* und war 2,5 Jahre wissenschaftlicher Mitarbeiter am Leibniz KI-Labor des *L3S.* Seine Forschungsinteressen sind die Verarbeitung natürlicher Sprache und KI in der Medizin. Schwerpunkt ist die Entwicklung von Techniken zur Anpassung von Domänen für verschiedene NLP- Anwendungen. \\\

KONTAKT:

Uwe Hadler, M. Sc.
uwe.hadler@L3S.de

\\\ Uwe Hadler ist wissenschaftlicher Mitarbeiter am *L3S* und unterstützt über das *Mittelstand-Digital Zentrum Hannover* Unternehmen bei der Einführung von KI-Systemen. Sein Forschungsgebiet sind Sprachmodelle und die Entwicklung von Methoden, um die Zuverlässigkeit und Vertrauenswürdigkeit von Sprachmodellen zu verbessern. \\\

Soumyadeep Roy, Aparup Khatua, Fatemeh Ghoochani, Uwe Hadler, Wolfgang Nejdl, Niloy Ganguly: Beyond Accuracy: Investigating Error Types in GPT-4 Responses to USMLE Questions. SIGIR 2024: 1073-1082
→ https://dl.acm.org/doi/10.1145/3626772.3657882

EFFIZIENTES
MULTI-KAMERA-TRACKING
VON FAHRZEUGEN

Intelligenteres Verkehrsmanagement

Weniger Staus, besserer öffentlicher Nahverkehr und geringere Fahrzeugemissionen – das Leben in der Stadt könnte deutlich angenehmer sein. Intelligente Verkehrssysteme sollen die Lösung bringen und den Verkehr besser steuern. Damit das funktioniert, müssen Kameras die Fahrzeuge teils über weite Strecken verfolgen und die gesammelten Daten koordiniert werden – bisher eine aufwendige und teure Angelegenheit. Wissenschaftler des *L3S* haben im renommierten *Machine Learning Journal* eine Lösung vorgestellt, die das Multi-Kamera-Tracking effizienter macht: das KI-System *LaMMOn*.

GRENZEN BESTEHENDER SYSTEME

Bestehende Systeme sind sehr arbeitsintensiv. Für jede neue Kameraeinstellung müssen die Regeln zur Verknüpfung der erfassten Fahrzeuge zwischen den einzelnen Kameras manuell erstellt werden. »Das ist sehr aufwendig und zudem nur eingeschränkt skalierbar«, sagt Marco Fisichella, Forschungsgruppenleiter am *L3S* und einer der Entwickler von *LaMMOn*. Hinzu kommt die begrenzte Verfügbarkeit öffentlicher Datensätze, die es erschwert, neue Systeme zu testen und zu optimieren.

DER SCHLÜSSEL ZU MEHR EFFIZIENZ

LaMMOn nutzt fortschrittliche sprach- und graphbasierte KI-Techniken, um sich automatisch und ohne manuelle Einstellungen an verschiedene Szenarien anzupassen. Das System besteht aus drei Hauptmodulen:

1. Language Model Detection (LMD):
Dieses Modul ist für die Objekterkennung verantwortlich und erzeugt Fahrzeugmerkmale wie Typ, Farbe und Position.

2. Language and Graph Model Association (LGMA):
Es verknüpft wiedererkannte Fahrzeuge über mehrere Kameras hinweg und kombiniert Objekte, die von mehreren Kameras erkannt wurden, zu einer globalen Multikameratrajektorie, die den Bewegungspfad des Objekts darstellt.

3. Text-to-Embedding (T2E):
Das Modul löst das Problem des Datenmangels, indem es synthetische Objektmerkmale generiert – basierend auf Textbeschreibungen wie ›roter Kombi‹ oder ›blauer SUV‹.

Verkehrsüberwachungssysteme mit Kameras können für einen besseren Verkehrsfluss in Städten genutzt werden – vorausgesetzt, ein intelligentes System erhält die dafür relevanten Daten und kann diese in Echtzeit verarbeiten, was eine sehr komplexe Aufgabe darstellt. Das große Bild erstelle *Midjourney* zu dem Prompt »professional photography of flowing city traffic on a sunny morning in Germany with green traffic light«, das kleine ganz link zu dem Prompt »a detailed photo of a traffic light, green, with a surveillance camera, in front of a white background, natural lighting«.

PRAKTISCHE ANWENDUNG UND ERFOLGE

LaMMOn hat sich bereits in mehreren Testdatensätzen bewährt. Es erreicht eine hohe Tracking-Genauigkeit von über 75 Prozent der HOTA-Metrik und übertrifft damit viele frühere Modelle. »Unsere Ergebnisse zeigen, dass *LaMMOn* für den Einsatz in Echtzeit-Verkehrsszenarien gut geeignet ist«, sagt Fisichella. Mit einer Bildrate von über zwölf Bildern pro Sekunde erreicht das System die Geschwindigkeit, die für die Anwendung in Echtzeit erforderlich ist, ohne dabei an Präzision einzubüßen – ideal für smarte Städte.

DIE ZUKUNFT DES TRACKING

Neben der technischen Umsetzung hebt die Studie besonders die Rolle des T2E-Moduls hervor, das es ermöglicht, Fahrzeugdaten aus Text zu generieren. »Dieses Modul reduziert nicht nur den Aufwand für die manuelle Datenerstellung, sondern macht das System auch anpassungsfähiger und vielseitiger«, so Fisichella.
Zukünftig wird *LaMMOn* noch vielseitiger. Das Entwicklerteam plant, die sprachbasierten Funktionen zu erweitern und die Graphstrukturen zu verbessern, um noch komplexere Anwendungen zu unterstützen. »*LaMMOn* ist damit eine zukunftsweisende Lösung, die sich perfekt für die Verkehrsüberwachung und -steuerung eignet.« ¶

KONTAKT:

Dr. Marco Fisichella
mfisichella@L3S.de

\\\ Marco Fisichella leitet am *L3S* eine Forschungsgruppe, die sich mit künstlicher Intelligenz und intelligenten Systemen insbesondere für die Anwendungsbereiche Mobilität, intelligente Produktion und personalisierte Medizin beschäftigt. \\\

KONTAKT:

Dr. Hoang H. Nguyen
ehoang@L3S.de

\\\ Hoang H. Nguyen war bis Juli 2024 Doktorand am *L3S*. Seit August 2024 ist er Postdoc an der *University of Tennessee* in Chattanooga, USA. Seine Forschungsschwerpunkte umfassen Graph Learning, Blockchain-Sicherheit und Transport. \\\

Tuan T. Nguyen, Hoang H. Nguyen, Mina Sartipi, Marco Fisichella: LaMMOn: language model combined graph neural network for multi-target multi-camera tracking in online scenarios. Mach. Learn. 113(9): 6811-6837 (2024) → https://link.springer.com/article/10.1007/s10994-024-06592-1

DIE ZAHL

WISSENSCHAFTLER UND WISSENSCHAFTLERINNEN
DES *L3S* HABEN IM JAHR 2023

127

BEITRÄGE
in hochrangigen Fachzeitschriften und auf erstklassigen Konferenzen publiziert

Quelle: *L3S*

PROMOTION AM *L3S*

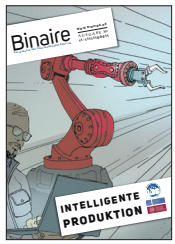
Dr. rer. nat.
Emmanuel Olatunji
»Privacy-preserving Graph Machine Learning«

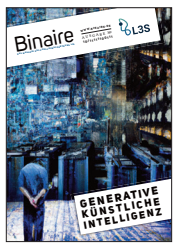
JULI 2024
DOKTORVATER:
PROF. DR. WOLFGANG NEJDL

In seiner Dissertation untersuchte er die Auswirkungen von graphischen neuronalen Netzen (GNNs) auf die Privatsphäre, wobei sich **Emmanuel Olatunji** auf die Anfälligkeit für Angriffe wie Membership Inference und Graph Reconstruction konzentrierte und Verteidigungsstrategien und Rahmenbedingungen für den Schutz der Privatsphäre vorschlug. »Während meiner Zeit am *L3S* hat mich das kooperative und unterstützende Forschungsumfeld inspiriert, das meine Arbeit bereichert, mich bei persönlichen Verlusten unterstützt und lebenslange Freundschaften gefördert hat. Meine akademische Reise begann mit einem Bachelor in Nigeria, einem MSc und einem MPhil in Hongkong und führte mich zum Charme Hannovers – vom Maschsee bis zur Marienburg – was dieses Kapitel unvergesslich macht.«

WEGE ZUR BINAIRE

Haben Sie Interesse an einzelnen Exemplaren oder möchtem Sie ein Abo bestellen? Mailen Sie einfach an die Redaktion! Gerne senden wir Ihnen die *Binaire* kostenlos zu.
Oder lesen Sie online: www.binaire.de





IMPRESSUM

L3S
Trustworthy AI & Digital Transformation

HERAUSGEBER:

Forschungszentrum L3S
Leibniz Universität Hannover
Appelstraße 9a
30167 Hannover

VERANTWORTLICH:

Prof. Dr. techn. Wolfgang Nejdl
Geschäftsführender Direktor

REDAKTION:

Dipl.-Geogr. Susanne Oetzmann
E-Mail: oetzmann@L3S.de

KONZEPT & DESIGN:

Dipl.-Des. Priska Tosch
www.tosch-kommunikation.de

DRUCK:

auf 100% Recyclingpapier
Ströher Druckerei und Verlag GmbH & Co. KG
www.stroeher-druck.de

Druckprodukt
CO_2 kompensiert
klima-druck.de
ID-Nr. 24189823
VDM+

BILDQUELLEN:

Forschungszentrum L3S,
wenn nicht anders vermerkt

Titelbild:
Priska Tosch

L3S.de

Leibniz Universität Hannover

Technische Universität Braunschweig

CAIMed

Niedersächsisches Zentrum für KI & Kausale Methoden in der Medizin

Im CAIMed entwickeln wir innovative Methoden für eine verbesserte, personalisierte Gesundheitsversorgung und tragen zur Bewältigung von Volkskrankheiten wie Krebs, Herz-Kreislauf-Erkrankungen und Infektionen bei. Mit der Verknüpfung niedersächsischer Standorte in KI- und medizinischer Forschung entsteht ein Leuchtturm für KI und personalisierte Medizin.

zukunft. niedersachsen

CAIMed wird gefördert durch das Niedersächsische Ministerium für Wissenschaft und Kultur mit Mitteln aus dem Programm zukunft.niedersachsen der VolkswagenStiftung.

info@caimed.de
www.caimed.de

L3S Trustworthy AI & Digital Transformation
Leibniz Universität Hannover
Technische Universität Braunschweig
CIDAS Campus-Institut Data Science
Universitätsmedizin Göttingen UMG
HZI Helmholtz Zentrum für Infektionsforschung
MHH Medizinische Hochschule Hannover
Georg-August-Universität Göttingen
CiiM Centre for Individualised Infection Medicine