Deutsches Forschungszentrum
für Künstliche Intelligenz
German Research Center for Artificial Intelligence

XAINES – KI mit Narrativen erklären

Abschlussbericht

Förderkennzeichen	01IW20005
Zuwendungsempfänger	Deutsches Forschungszentrum für Künstliche Intelligenz GmbH Trippstadter Straße 122, D-67663 Kaiserslautern
Ausführende Stelle	Agenten und Simulierte Realität, Standort Saarbrücken
Projektleiter	Prof. Philipp Slusallek
Bewilligungszeitraum	1. September 2020 bis zum 31. August 2024
Autor	Christian Müller
Erstellungsdatum	30. September 2024

Gefördert vom:

Bundesministerium
für Bildung
und Forschung

Das diesem Bericht zugrunde liegende Vorhaben wurde mit Mitteln des Bundesministeriums für Bildung und Forschung unter dem Förderkennzeichen 01IW20005 gefördert. Die Verantwortung für den Inhalt dieser Veröffentlichung liegt beim Autor.

Inhaltsverzeichnis

1 Kurzdarstellung nach NKBF

1.1 Aufgabenstellung

Das XAINES-Projekt („Explaining AI with Narratives“) zielte darauf ab, die Verständlichkeit und Transparenz von KI-Systemen durch die Entwicklung narrativer Erklärungen zu verbessern. Im Mittelpunkt stand die Schaffung von Modellen und Methoden, die es ermöglichen, die Entscheidungsprozesse von KI-Systemen in einer für den Menschen nachvollziehbaren Weise darzustellen. Dies ist besonders wichtig, da KI zunehmend in sensiblen Bereichen wie dem Gesundheitswesen, dem autonomen Fahren und der Entscheidungsunterstützung eingesetzt wird, wo es von entscheidender Bedeutung ist, dass Nutzer und Stakeholder den Gründen und Mechanismen hinter den KI-Entscheidungen vertrauen und sie verstehen können.
Um dieses Ziel zu erreichen, konzentrierte sich das Projekt auf die Entwicklung neuer Ansätze, die es erlauben, sensorische Daten (z.B. Bewegungs- und Sprachdaten) mit natürlichen Sprachbeschreibungen zu verknüpfen. Dadurch sollten KI-Systeme in die Lage versetzt werden, komplexe sensorische Eingaben in verständliche narrative Erklärungen zu übersetzen. Dies umfasste die Integration multimodaler Daten und die Entwicklung von Methoden zur Verarbeitung und Visualisierung von Text- und Bilddaten, um den Nutzerinnen und Nutzern eine umfassende und klare Einsicht in die Funktionsweise der KI zu bieten.

1.1.1 Technische Arbeitsziele

Die technischen Arbeitsziele des XAINES-Projekts konzentrierten sich auf die Entwicklung fortschrittlicher Methoden zur Generierung von verständlichen Erklärungen für komplexe KI-Entscheidungen. Ein zentrales Ziel war die Schaffung von Systemen, die verschiedene Datenquellen, wie sensorische Eingaben, Sprache und visuelle Informationen, nahtlos miteinander verknüpfen können, um eine kohärente und verständliche Darstellung von KI-Prozessen zu ermöglichen. Hierzu wurden mehrere technische Teilziele definiert:

1. Integration von Multimodalen Daten: Entwicklung von Modellen, die Daten aus verschiedenen Sensoren, wie Bewegungstrackern, Audio- und Videosignalen, mit Textdaten kombinieren. Diese Modelle sollten in der Lage sein, aus den sensorischen Eingaben Informationen zu extrahieren und diese in eine für Menschen verständliche Sprache zu übersetzen. Dies erforderte Fortschritte in Bereichen wie Multimodalität, Signalverarbeitung und natürlicher Sprachverarbeitung.
2. Generierung von Erklärungen: Aufbau eines Systems, das KI-Entscheidungen nicht nur ausführt, sondern auch durch Erklärungen verständlich macht. Dazu gehörte die Fähigkeit, aus Bildern, Texten und Sprachdaten Beschreibungen zu erstellen, die dem Nutzer Schritt für Schritt erklären, wie und warum eine bestimmte Entscheidung getroffen wurde. Diese narrative Darstellung sollte dazu beitragen, die Transparenz und das Vertrauen in KI-Systeme zu erhöhen.
3. Visuelle Erklärungen und Storytelling: Entwicklung von Algorithmen, die visuelle Darstellungen von Prozessen generieren können, um komplexe Zusammenhänge zu verdeutlichen. Dies beinhaltete die Nutzung von Techniken wie Bildgenerierung und -beschreibung, um Erklärungen zu erstellen, die auf visuellen Daten basieren, z.B. durch die Darstellung von erkannten Objekten und deren Interaktionen.
4. Anpassung an Domänenspezifische Anwendungsfälle: Implementierung von spezifischen Erklärungsmechanismen für die gewählten Anwendungsfälle wie autonomes Fahren, Pflege und medizinische Entscheidungsfindung. Jede Domäne brachte eigene technische Herausforderungen mit sich, die spezielle Anpassungen der entwickelten Systeme

erforderten, um z.B. die Erklärung von Navigationsentscheidungen oder die Interpretation von Patientendaten zu unterstützen.

1.1.2 Wissenschaftliche Arbeitsziele

Die wissenschaftlichen Arbeitsziele des XAINES-Projekts waren darauf ausgerichtet, grundlegende Erkenntnisse und Innovationen im Bereich der erklärbaren künstlichen Intelligenz (XAI) zu fördern. Dazu gehörte die Erforschung und Entwicklung neuer Methoden, die es ermöglichen, komplexe KI-Systeme für den Menschen verständlicher und transparenter zu gestalten. Die spezifischen wissenschaftlichen Ziele waren:

1. Förderung der Erklärbarkeit von KI durch Narrative: Das Hauptziel war es, neue Ansätze zu entwickeln, um KI-Entscheidungen nicht nur technisch korrekt, sondern auch für menschliche Nutzer verständlich und nachvollziehbar zu machen. Dies beinhaltete die Erforschung narrativer Techniken, um komplexe KI-Prozesse als schlüssige Beschreibungen zu präsentieren, die für Anwender intuitiver und zugänglicher sind.
2. Verknüpfung von Multimodalität und semantischer Verarbeitung: Ein bedeutendes wissenschaftliches Ziel war die Untersuchung, wie verschiedene Datentypen (z.B. Sprach-, Bild- und Sensordaten) effektiv kombiniert und gemeinsam interpretiert werden können. Dabei sollten neue Modelle und Algorithmen erforscht werden, die diese multimodalen Daten so verarbeiten, dass sie zu kohärenten und erklärbaren Erklärungen zusammengeführt werden.
3. Erweiterung der Grundlagenforschung zur generativen Modellierung: Im Rahmen des Projekts wurden innovative generative Modelle erforscht, die in der Lage sind, nicht nur Daten zu analysieren, sondern auch neue Daten zu generieren, die den Entscheidungsprozess verdeutlichen. Ziel war es, diese Modelle so zu entwickeln, dass sie bei der Erstellung visueller und sprachlicher Erklärungen unterstützend wirken und eine realitätsnahe Darstellung komplexer Prozesse ermöglichen.
4. Entwicklung neuer Methoden für semantische Analyse und Sprachverarbeitung: Wissenschaftliche Arbeiten konzentrierten sich darauf, die Grenzen bestehender Methoden in der natürlichen Sprachverarbeitung (NLP) zu erweitern, um tiefere semantische Verbindungen zwischen Daten und Erklärungen herzustellen. Dies beinhaltete die Erforschung von Techniken zur Generierung von Texten, die komplexe sensorische Eingaben in präzise und verständliche natürliche Sprache übersetzen.
5. Evaluation, Datenerstellung, Veröffentlichung und Validierung von Erklärungsansätzen in praxisnahen Szenarien: Ein zentrales wissenschaftliches Ziel des XAINES-Projekts war die Entwicklung, Validierung und Veröffentlichung umfangreicher, multimodaler Datensätze, die reale Anwendungsszenarien abbilden. Diese Datensätze dienten nicht nur dazu, die entwickelten Modelle zu trainieren und ihre Effektivität zu testen, sondern wurden auch der wissenschaftlichen Gemeinschaft zur Verfügung gestellt. Durch die Veröffentlichung dieser Daten soll der wissenschaftliche Fortschritt im Bereich der erklärbaren KI gefördert werden, indem andere Forscher die Möglichkeit erhalten, auf denselben Datengrundlagen aufzubauen und weiterführende Studien durchzuführen. Die Erstellung qualitativ hochwertiger und diverser Datensätze ermöglicht es, standardisierte Benchmarks zu setzen und die Vergleichbarkeit zwischen verschiedenen Ansätzen zu verbessern. Diese Initiative trägt zur Förderung eines offenen und kollaborativen Forschungsumfelds bei und stellt sicher, dass die entwickelten Methoden in praxisnahen Umgebungen validiert werden.

Durch die Umsetzung dieser wissenschaftlichen Ziele trug das XAINES-Projekt erheblich zur Weiterentwicklung der erklärbaren KI bei und schuf eine Grundlage für zukünftige Forschungsarbeiten, die darauf abzielen, die Kluft zwischen komplexen maschinellen Prozessen und menschlichem Verständnis weiter zu überbrücken.

1.2 Voraussetzungen, unter denen das Vorhaben durchgeführt wurde

Das XAINES-Projekt profitierte maßgeblich von der interdisziplinären Zusammenarbeit verschiedener Forschungsbereiche des DFKI. Jeder dieser Bereiche brachte spezialisierte Kompetenzen und technologische Ansätze ein, die die umfassende Entwicklung der Projektziele ermöglichten. Die Beteiligung dieser vielfältigen Forschungsbereiche brachte neben vielen Vorteilen auch Herausforderungen mit sich. Jeder Bereich operiert typischerweise eigenständig und verfolgt spezifische agendabezogene Ziele. Die Notwendigkeit, diese verschiedenen Ansätze zu einer kohärenten Projektumsetzung zu vereinen, erforderte eine enge Koordination und effektives Management. So wurde im XAINES-Projekt wurde flexible Integration von Systemkomponenten bevorzugt, statt eines starren, übergreifenden Gesamtsystems. Um die jeweiligen Stärken der Forschungsbereiche optimal zu nutzen und gleichzeitig eine kohärente Zusammenarbeit zu gewährleisten, fokussierte sich das Projekt auf die Entwicklung gemeinsamer Anwendungsfälle und Datensätze, die eine breitere Testbarkeit der Erklärbarkeitssysteme sicherstellten. Diese flexible Ausrichtung ermöglichte es, Teilergebnisse schneller zu integrieren und zu validieren, wodurch die Anpassungsfähigkeit und der praktische Nutzen der entwickelten Technologien gestärkt wurden.

1.3 Planung und Ablauf des Vorhabens

Das XAINES-Projekt war darauf ausgelegt, interdisziplinäre Forschungsansätze zu kombinieren, um innovative Lösungen im Bereich der erklärbaren Künstlichen Intelligenz zu entwickeln. Der Projektverlauf gliederte sich in mehrere Arbeitspakete, die aufeinander aufbauten und jeweils spezifische technologische und wissenschaftliche Ziele verfolgten. Zu Beginn wurden die grundlegenden Konzepte und Methoden zur multimodalen Datenverarbeitung und natürlichen Sprachverarbeitung erarbeitet. Im weiteren Verlauf lag der Fokus auf der Integration dieser Ansätze in benutzerfreundliche und leicht verständliche Erklärungen, die in verschiedenen Anwendungsfällen, wie dem autonomen Fahren und der Pflege, erprobt wurden.
Die Arbeitspakete waren so strukturiert, dass sie sowohl die Entwicklung neuer Modelle und Algorithmen als auch die Schaffung praxisnaher Anwendungsfälle umfassten. Dabei wurde eng zwischen den verschiedenen Forschungsbereichen des DFKI kooperiert, um die unterschiedlichen Kompetenzen und Technologien bestmöglich zu nutzen. Der iterative Ansatz erlaubte es, neue Erkenntnisse laufend in die Projektarbeit einfließen zu lassen und die Lösungen schrittweise zu optimieren. So konnte das Projekt flexibel auf Herausforderungen reagieren und seine Ziele effizient umsetzen.

1.4 Wissenschaftlicher und Technischer Stand zu Beginn des Vorhabens

Zu Beginne des Projektes im Jahr 2020 befand sich der Stand der Technik in der erklärbaren Künstlichen Intelligenz (XAI) auf einem Entwicklungsstand, der von einer wachsenden Anerkennung der Notwendigkeit für mehr Transparenz und Nachvollziehbarkeit von KI-Entscheidungen geprägt war. Obwohl leistungsstarke KI-Modelle, insbesondere auf Basis von Deep Learning, bemerkenswerte Fortschritte in verschiedenen Anwendungsbereichen machten, blieb ihre sogenannte "Black-Box"-Natur ein zentrales Hindernis für eine breite und verantwortungsvolle Anwendung. Dies führte zu einem verstärkten Forschungsinteresse, die Funktionsweise dieser Modelle besser zu verstehen und zu erklären.

Zu den vorherrschenden Methoden gehörten sowohl modell-agnostische als auch modell-spezifische Ansätze. Modell-agnostische Methoden, wie zum Beispiel LIME (Local Interpretable Model-agnostic Explanations), analysierten Modelle unabhängig von ihrer Architektur, indem sie interpretierbare Segmente im Eingaberaum maskierten und deren Einfluss auf die Modellentscheidung bewerteten. Andere Methoden, wie Grad-CAM oder Layerwise Relevance Propagation (LRP), ermöglichten Einblicke in tiefere Schichten neuronaler Netzwerke, indem sie beispielsweise die Relevanz einzelner Eingabefeatures für die Modellentscheidung zurückverfolgten und visualisierten.
Dennoch hatten diese Methoden oft das Problem, dass die erzeugten Erklärungen numerisch und für Laien schwer verständlich blieben. Daher wurden zunehmend Visualisierungen genutzt, um beispielsweise durch Heatmaps hervorzuheben, welche Teile eines Bildes besonders zur Entscheidung beitrugen. Auch wenn dies Fortschritte brachte, waren solche Visualisierungen noch immer rudimentär und erforderten oft ein gewisses Maß an Fachwissen, um richtig interpretiert zu werden.
Ein weiterer vielversprechender Bereich war die Nutzung von Aufmerksamkeitsmechanismen in Modellen, die Informationen darüber liefern konnten, welche Teile der Eingabe ein Modell während der Entscheidungsfindung besonders stark berücksichtigt hatte. Die Generierung von textuellen Erklärungen, die Begründungen für Modellverhalten liefern, war ebenfalls ein Forschungsfeld, das zu dieser Zeit aufkam und zunehmend Relevanz erlangte, auch wenn hier noch viele Herausforderungen bestanden.
Im Kontext der autonomen Systeme, etwa im Bereich des autonomen Fahrens, war die Erklärbarkeit besonders wichtig, um Sicherheitsbehörden und Entwicklern die Möglichkeit zu geben, Fehlerquellen zu identifizieren und Systeme zu zertifizieren. Hier konzentrierte sich die Forschung auf diagnostische Erklärungen, die es ermöglichen sollten, komplexe Automatisierungssysteme besser zu verstehen und Fehlentscheidungen zu analysieren. Im medizinischen Bereich, etwa bei der automatisierten Diagnoseunterstützung, wurde versucht, die Entscheidungen neuronaler Netze so zu erklären, dass sie für medizinisches Fachpersonal nachvollziehbar sind. Dabei stellte sich die Herausforderung, dass Modelle oft auf Basis von Bilddaten arbeiteten, deren Entscheidungsgrundlage für den Menschen ohne entsprechende Aufbereitung schwer nachvollziehbar war.
Ein zentraler Ansatz von XAINES war die Nutzung von multimodalen Beschreibungen, um KI-Entscheidungen zu erklären. Es sollte eine natürliche Art der Wissensvermittlung für Menschen erreicht werden, die eine sequenzielle, kausal verknüpfte Darstellung von Informationen erlaubt, die auch Laien verstehen können. Dies ging über simple Erklärungen hinaus und ermöglichte eine dialogische Form, die in Interaktionen mit Nutzern adaptiv sein konnte. Durch die Verknüpfung von Sprachmodellen mit sensorbasierten Daten sollte eine tiefergehende, multimodale Erklärbarkeit erreicht werden. Dies ermöglichte es, physische Vorgänge mit textbasierten Erklärungen zu verbinden, was besonders in dynamischen Umgebungen, wie Baustellen oder in der Medizin, als hilfreich angesehen wurde.
Das Projekt nutzte Simulationen und synthetische Daten, um Szenarien zu erzeugen, die als Grundlage für Erklärungen dienten. Durch die Simulation verschiedener Umgebungen konnten Systeme trainiert werden, die später reale Situationen besser erklären können.

1.4.1 Eigene Vorarbeiten

Die im XAINES-Projekt durchgeführten Vorarbeiten basieren auf einer Vielzahl von Forschungsaktivitäten innerhalb des DFKI. Wesentliche Grundlagen für das Projekt wurden durch Projekte geschaffen, die sich mit erklärbarer KI und der Verknüpfung von sensorischen Daten mit beschreibenden Erklärungen befassten.
Eines dieser Projekt war DEEPLEE (2017-2020, BMBF), das sich auf die Erklärbarkeit neuronaler Modelle für die Verarbeitung natürlicher Sprache konzentrierte. Hier wurden Techniken entwickelt,

um die internen Prozesse von Deep-Learning-Modellen transparenter zu machen, insbesondere durch Visualisierungen, die als Basis für narrativ strukturierte Erklärungen dienten. Ergänzend dazu wurde im Rahmen des QURATOR-Projekts (2018-2021, BMBF) das Konzept des "Semantic Storytelling" erforscht, das es ermöglichte, narrative Erklärungen für komplexe Informationsprozesse zu generieren. Ein weiterer wichtiger Baustein war das Projekt REACT (2017-2020, verlängert unter dem Zusatznamen CrossCDR bis 2023), das darauf abzielte, eine systematische und validierbare Methode zur Entwicklung, Training und Nutzung digitaler Realitäten zu schaffen, um das sichere und zuverlässige Handeln autonomer Systeme zu gewährleisten. Durch die Modellierung, Simulation und das Lernen von Verhaltensmustern in realistischen, virtuellen Umgebungen wurden autonome Fahrzeuge gezielt auf gefährliche Verkehrssituationen vorbereitet. REACT nutzte moderne maschinelle Lerntechniken wie Deep Learning und (Deep) Reinforcement Learning, um Simulationen zu erstellen, die eine breite Palette kritischer Verkehrssituationen abdecken. Diese virtuellen Umgebungen ermöglichten die Synthese von Sensordaten, die in realen Szenarien schwer zu erfassen wären, und stellten sicher, dass autonome Systeme sicher und zuverlässig trainiert werden konnten. Ergänzt wurden diese Ansätze durch Interactive Machine Learning (IML)-Strategien, die sich mit der Visualisierung und Benutzerinteraktion zur Erklärung der Entscheidungen maschineller Lernsysteme befassten. Diese Konzepte trugen dazu bei, die Erklärungen von KI-Systemen benutzerfreundlicher und verständlicher zu gestalten, indem interaktive Schnittstellen entwickelt wurden, die eine effektive Kommunikation mit Fachanwendern ermöglichten.
Das DFKI pflegte zu diesem Zeitpunkt auch bereits eine enge Zusammenarbeit mit externen Partnern wie INRIA Nancy, um hybride Ansätze für erklärbare KI zu erforschen. Diese Kooperation fokussierte auf die Integration symbolischen Wissens in tiefe Lernmodelle, um eine tiefere und robustere Erklärbarkeit zu erreichen, etwa in der medizinischen Datenverarbeitung, einschließlich Haut- und Radiologiebildern. Insgesamt dienten diese Vorarbeiten als Grundlage für die Entwicklungen im XAINES-Projekt, um beschreibende und multimodale Erklärungen zu schaffen, die in mehreren Anwendungsdomänen eine breitere Akzeptanz und ein besseres Verständnis von KI-Systemen fördern sollen.

1.4.2 Fachliteratur und verwendete Dokumentationsdienste

Für die Entwicklung der Technologien und Modelle im XAINES-Projekt wurde umfangreich auf aktuelle wissenschaftliche Literatur und spezialisierte Dokumentationsdienste zurückgegriffen. Ein zentraler Schwerpunkt lag auf der Forschung im Bereich der erklärbaren Künstlichen Intelligenz (XAI), insbesondere auf Publikationen, die sich mit multimodaler Datenverarbeitung, natürlicher Sprachverarbeitung und der Generierung narrativer Erklärungen befassen. Fachzeitschriften wie *Journal of Artificial Intelligence Research* und *IEEE Transactions on Neural Networks and Learning Systems* lieferten wichtige theoretische Grundlagen für die Algorithmen- und Modellentwicklung.
Darüber hinaus wurden Datenbanken wie *IEEE Xplore*, *ACM Digital Library* und *SpringerLink* intensiv genutzt, um Zugang zu neuesten Forschungsergebnissen und Studien zu erhalten. Diese Quellen ermöglichten es, aktuelle Entwicklungen in den Bereichen maschinelles Lernen, Sprachverarbeitung und Computer Vision nachzuverfolgen und in die Projektarbeit zu integrieren. Um sicherzustellen, dass die entwickelten Modelle auf dem neuesten Stand der Technik basieren, wurden auch wissenschaftliche Konferenzbeiträge von führenden Veranstaltungen wie beispielsweise der *International Conference on Machine Learning (ICML)* und der *Conference on Empirical Methods in Natural Language Processing (EMNLP)* ausgewertet.
Für die Erstellung und Verwaltung der Trainingsdaten sowie der synthetischen Datensätze griff das Projektteam auf spezialisierte Dokumentationsdienste und Plattformen zur Datenverwaltung zurück, darunter *GitHub*. Dies ermöglichte eine strukturierte und nachvollziehbare Dokumentation aller Schritte der Datenverarbeitung und Modellentwicklung.

1.5 Zusammenarbeit mit anderen Stellen

Das Ökosystem des XAINES-Projekts vereint eine Vielzahl interdisziplinärer Forschungsprojekte und -initiativen, die auf die Entwicklung innovativer KI-Lösungen abzielen. Die Projekte sind durch ihre thematische Vielfalt und enge Integration charakterisiert, wobei jedes Projekt spezifische Kompetenzen einbringt, um die übergreifenden Ziele von XAINES zu unterstützen:

1. Ophthalmo-AI (IML, gefördert durch das BMBF, 03/2021 - 03/2024): Dieses Projekt zielte darauf ab, die Diagnose und Therapie in der Augenheilkunde zu verbessern. Durch die Kombination interaktiver maschineller Lernmodelle mit medizinischem Fachwissen wurde ein System geschaffen, das Ärzten bessere, transparentere Entscheidungsunterstützung bietet. Das Projekt förderte die Nutzung von erklärbarer KI, um die Ergebnisse für Patienten nachvollziehbarer zu gestalten.
2. MOMENTUM (ASR, gefördert durch das BMBF, 08/2022 - 08/2025): Widmet sich der Entwicklung vertrauenswürdiger, robuster KI-Systeme für autonome Anwendungen. Das Projekt legt den Schwerpunkt auf hybride KI-Modelle, die Sicherheit und Transparenz bei der Mensch-Maschine-Interaktion gewährleisten. Es umfasst mehrere Anwendungsfälle, darunter autonomes Fahren, wo es innovative Methoden für Bewegungserkennung und Verhaltensmodellierung entwickelt.
3. VidGenSens (EI, gefördert durch das BMBF, 12/2021 - 11/2024): VidGenSens fokussiert sich auf die Generierung synthetischer Daten für tragbare Sensoren. Diese synthetischen Daten helfen, KI-Modelle effizienter zu trainieren, insbesondere in Szenarien, in denen reale Daten schwer zu beschaffen sind. Durch diese Technik wird die Verfügbarkeit von Daten verbessert, was die Entwicklung robusterer Modelle ermöglicht.
4. ExplAINN (SDS, gefördert durch das BMBF, 10/2019 - 09/2022): Das Projekt entwickelte Methoden zur Verbesserung der Erklärbarkeit von KI-Modellen, insbesondere in sicherheitskritischen Anwendungen wie autonomem Fahren und medizinischen Diagnosen. Dabei wurden neue Techniken erforscht, die es ermöglichen, die Entscheidungen von Deep-Learning-Modellen verständlicher und nachvollziehbarer zu machen
5. IMPRESS (MLT, gefördert durch das BMBF, 08/2020 - 01/2024): IMPRESS konzentrierte sich auf die Verbesserung von Embeddings durch Integration von semantischem Wissen, um die Effizienz und Genauigkeit von Sprachmodellen zu steigern. Dies spielt eine wichtige Rolle bei der Entwicklung mehrsprachiger Anwendungen, die auf präzisen und kontextbezogenen Sprachverarbeitungsmodellen basieren. Das Projekt zielte darauf ab, die Anwendbarkeit von KI-Modellen auf mehrere Sprachen und Domänen zu erweitern.
6. CORA4NLP (MLT, gefördert durch das BMBF, 10/2020 - 09/2023): Entwickelte innovative Methoden zur kontextuellen Argumentation und Anpassung für die natürliche Sprachverarbeitung. Ziel war es, Modelle zu schaffen, die sich effizient an neue Sprachen und Domänen anpassen, und die Technologie durch Techniken wie kontinuierliches Lernen und die Nutzung semantischer Hintergrundinformationen zu verbessern. CORA4NLP strebte eine breite Anwendbarkeit in verschiedenen NLP-Aufgaben an, von Informationsentnahme bis hin zu Dialogverarbeitung.
7. ALMA (EI, gefördert durch EU Horizon 2020, 09/2020 - 09/2024): ALMA setzte auf Algebraic Machine Learning (AML), eine neue Form des maschinellen Lernens, um menschenzentrierte, interaktive Lernsysteme zu entwickeln. Das Projekt förderte die Reduzierung von Bias und die Verbesserung der Transparenz, indem es formale Repräsentationen menschlichen Wissens integriert. Zudem ermöglichte die Technologie eine kollaborative, verteilte Lernmethode, die den Datenschutz wahrt und die Abhängigkeit von zentralisierten Daten reduziert.

2 Eingehende Darstellung nach NKBF

In diesem Kapitel wird eine detaillierte Analyse der spezifischen Entwicklungen und Ergebnisse des XAINES-Projekts vorgestellt. XAINES hatte das Ziel, neuartige Methoden zur Erklärung von KI-Entscheidungen zu entwickeln und diese in verschiedenen Domänen, wie dem autonomen Fahren, der Pflege und der medizinischen Diagnostik, umzusetzen. Die Eingehende Darstellung beschreibt, wie diese erklärbaren KI-Systeme entworfen, entwickelt und getestet wurden.
Besonders im Fokus standen multimodale Datenverarbeitung und die Generierung von Erklärungen, die es ermöglichen, komplexe maschinelle Entscheidungen in einer für den Menschen verständlichen Form zu präsentieren. In den verschiedenen Arbeitspaketen wurden innovative Modelle entwickelt, die Sensordaten, Sprachverarbeitung und visuelle Informationen integrieren, um nachvollziehbare und interaktive Erklärungen zu liefern.
Zudem wird die Verwendung der Fördermittel detailliert beschrieben, einschließlich der Mittelverteilung auf die einzelnen Arbeitspakete und deren spezifischen Zielsetzungen. Weiterhin wird auf die enge Zusammenarbeit zwischen den Forschungsbereichen eingegangen, die maßgeblich dazu beigetragen hat, die technischen und wissenschaftlichen Ziele des Projekts zu erreichen.

2.1 Verwendung der Zuwendung

Die im Rahmen des XAINES-Projekts erhaltenen Fördermittel wurden gezielt eingesetzt, um die technologischen und wissenschaftlichen Ziele des Projekts zu erreichen. Dieser Abschnitt gibt einen detaillierten Überblick darüber, wie die Fördermittel auf die einzelnen Arbeitspakete verteilt wurden und welche spezifischen Ergebnisse dadurch erzielt werden konnten.

2.1.1 AP 1: Multisensorik und Sprachverarbeitung

Arbeitspaket 1 (AP1) konzentrierte sich auf die Verbindung von Sprache und Sensordaten, um gemeinsame Darstellungen von Erklärungen zu erstellen, die beide integrieren. Das Hauptziel von AP1 bestand darin, Systeme zu entwickeln, die sensorbasierte physische Aktivitäten und Wahrnehmungen auf für den Menschen geeignete Beschreibungen abbilden können, um eine Verbindung zwischen von Sensoren generierten Daten und entsprechenden Erklärungen in natürlicher Sprache herzustellen. Dazu mussten multimodale Daten - wie Körperbewegungen, Objekterkennung und Umgebungskontext - mit Strukturen in Einklang gebracht werden, die Menschen auf natürliche Weise verstehen.
Zu den wichtigsten Aufgaben in WP1 gehörten:

- Erstellung gemeinsamer Sensor-Sprache-Darstellungen: Entwicklung von Modellen, die Sensordaten (z. B. Bewegungsverfolgung, RFID-Signale, Videoeingabe) mit sprachbasierten Beschreibungen dieser Aktionen verbinden.
- Abbildung von Sprache auf Sensordaten: Extrahieren von aktivitätsbezogenen Informationen aus Freiformsprache zur Verbesserung der Erkennung von körperlichen Aktivitäten.
- Informationsextraktion aus Bildern: Automatisierte Bildbeschriftung und visuelle Erklärungen zum besseren Verständnis von KI-Entscheidungen.
- Erklärungen bei der Beantwortung visueller Fragen: Erarbeitung von Erklärungen für komplexe visuelle Fragen, die über kurze Antworten hinausgehen, um detailliertere Antworten zu geben.

AP1 legte den Grundstein für die Entwicklung erklärungsfähiger KI-Systeme, indem es die zugrundeliegende Struktur bereitstellte, die Aktivitäten in der realen Welt mit Erklärungen verbindet. Die Arbeit in diesem AP ermöglicht es KI-Systemen, rohe Sensordaten in aussagekräftige, für den Menschen verständliche Erklärungen zu "übersetzen". Dies ist eine Schlüsselkomponente des

umfassenderen Ziels des Projekts, KI-Entscheidungen für nicht fachkundige Nutzer transparent und erklärbar zu machen.

AP1.1 Erstellung einer gemeinsamen sensor- und sprachbasierten Darstellung von Erklärungen

Das Hauptziel der Aufgabe WP1.1 im XAINES-Projekt bestand darin, gemeinsame Darstellungen zu erstellen, die sprachbasierte Erklärungsmodelle mit Sensordaten verbinden. Die Aufgabe zielte darauf ab, textuelle Beschreibungen von Aktivitäten oder Situationen mit Datenströmen von verschiedenen Sensoren abzugleichen, wie z. B. am Körper getragene Trägheitsmesseinheiten (IMUs), Standortverfolgung, RFID, Umgebungsgeräusche und First-Person-Videos. Ziel war es, hierarchische Modelle zu konstruieren, die Low-Level-Sensorsignale in eine Semantik auf höherer Ebene integrieren und so ein umfassenderes Verständnis von Aktivitäten durch natürlichsprachliche Beschreibungen ermöglichen. Im Folgenden wird detailliert beschrieben, wie das Projekt zur Verwirklichung dieser Ziele beigetragen hat.

Im weiteren Verlauf dieses Berichts geben wir in den Titeln der einzelnen Beiträge zu den Aufgaben die DFKI-Abteilung(en) an, die an dem Thema gearbeitet haben. Unser Ziel war es, nicht nur Ergebnisse zu verknüpfen, sondern die Zusammenarbeit zwischen den Abteilungen in konkreten Beiträgen zu fördern.

SDS: Eine Methode, die reichhaltige Sprachausdrücke nutzt, um die Erzeugung natürlicher Bilder zu steuern

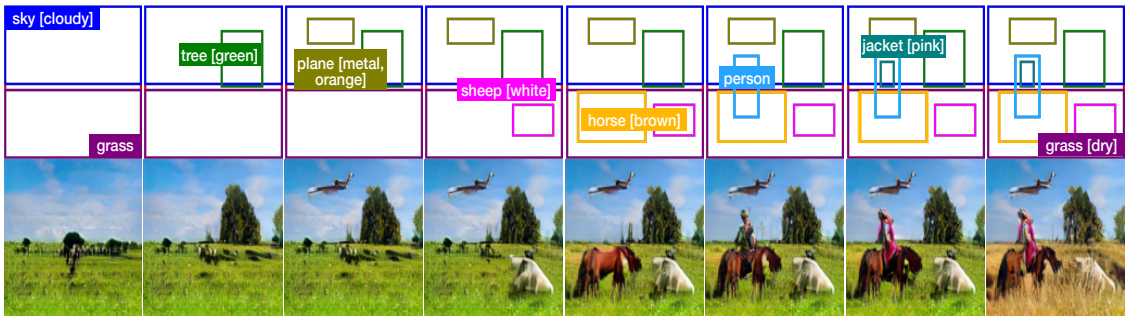

Abbildung 1: *Generierte Bilder mit einem rekonfigurierbaren Layout und Attributen zur Steuerung des Erscheinungsbildes einzelner Objekte.*

Eines der grundlegenden Ziele dieses AP war es, einen Abgleich zwischen textuellen Einbettungen und Sensorsignalen zu finden. In der Bilddomäne werden solche Signale als unabhängige oder sequentielle Sätze von Bildern dargestellt. Die Verankerung visueller Einheiten in einer textuellen Beschreibung ist angesichts der immensen Variabilität natürlicher Bilder eine Herausforderung. Der Versuch, textuelle Deskriptoren für bestehende Bilder zu lernen, macht das Trainingsverfahren anfällig für Verzerrungen in der Trainingsmenge (z.B. zu viele Außenszenen oder nur wenige Hunde). Daher haben wir AttLostGAN entwickelt, eine Methode, die reichhaltige sprachliche Ausdrücke nutzt, um die Erzeugung natürlicher Bilder zu steuern (Frolov et al., 2021). Diese Methode stützt sich auf Generative Adversarial Networks (GANs), die in zweierlei Hinsicht gesteuert werden. Die erste ist eine Darstellung des Layouts des Bildes. Bounding Boxes auf einer weißen Leinwand bestimmen die Größe und Position von Objekten in der Szene. Die zweite Steuerung entspricht textuellen Primitiven, die definieren, welche Objekte dargestellt werden sollen. Jede Objektdefinition besteht aus einem Objekt und einem Attribut, das das Aussehen der Ausgabe bestimmt, z. B. ein *braunes* Pferd oder ein *bewölkter* Himmel. Diese Paare aus Adjektiv und Substantiv sorgen für eine strukturiertere Verbindung zwischen den textlichen Darstellungen und der visuellen Ausgabe. Eine umfangreiche

Evaluierung zeigt, dass AttrLostGAN die Erzeugung einzelner Objekte in komplexen Szenen steuern kann.

Eigene Veröffentlichungen:
Frolov, S., Sharma, A., Hees, J., Karayil, T., Raue, F., & Dengel, A. (2021). AttrLostGAN: Attributgesteuerte Bildsynthese aus rekonfigurierbarem Layout und Stil. *Lecture Notes in Computer Science,* 361-375.

ASR: Absichtsgesteuerte Ganzkörper-Bewegungssynthese für Mensch-Objekt-Interaktionen

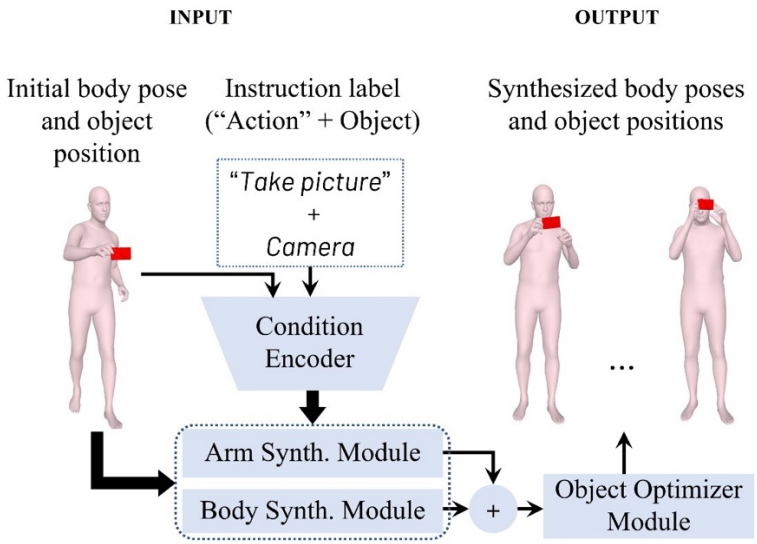

Abbildung 2: Überblick über unseren absichtsgesteuerten Ganzkörper-Bewegungsgenerator.

Dieser Beitrag steht im Einklang mit Aufgabe 1.1, da er sich auf die Synthese von Mensch-Objekt-Interaktionen auf der Grundlage von Textanweisungen und Sensordaten konzentriert, was für die Erstellung gemeinsamer Darstellungen von Sprache und Sensormodalitäten entscheidend ist. Unser Framework trägt direkt zu diesem Ziel bei, indem es Intent-Objekt-Paare mit Hilfe eines Sprachmodells (CLIP-Encoder) lernt und Ganzkörper-Bewegungssequenzen generiert, die den textuellen Beschreibungen von Mensch-Objekt-Interaktionen entsprechen. Diese Integration sprachbasierter Eingaben in die Bewegungssynthese steht in engem Zusammenhang mit dem Ziel von Task 1.1, hierarchische Modelle zu entwickeln, die sensorgestützte Aktivitätserkennung mit natürlichsprachlichen Beschreibungen verbinden. Durch die Berücksichtigung von Interaktionen, an denen beide Hände beteiligt sind, und die Optimierung der Objektpositionen in 3D verbessert *IMoS* die Fähigkeit, vielfältige, realistische Bewegungssequenzen zu erzeugen, die für die Erstellung kohärenter, auf Erzählungen basierender Erklärungen menschlicher Aktivitäten unerlässlich sind. Dies stärkt den multimodalen Abgleich, der in XAINES erforderlich ist, um die Lücke zwischen Text- und Sensordaten zu schließen.
Bisherige Methoden synthetisierten automatisch Bewegungen für virtuelle Charaktere, indem sie Steuersignale wie Musik (Lee et al., 2019), Sprache (Habibie et al., 2022) oder Text kodieren, entweder in Form von Sätzen (Ghosh et al., 2021; Petrovich et al., 2022) oder als hochrangige Aktionsbeschreibungen (Ahuja & Morency, 2019). Methoden zur Synthese von Ganzkörper-Positionssequenzen folgten typischerweise einem autoregressiven Ansatz, um die Kontinuität der synthetisierten Bewegungen zu erhalten (Rempe et al., 2021). Diese autoregressiven

Bewegungssyntheseverfahren sagten kurzfristige zukünftige Sequenzen aus einer kurzen Vergangenheit voraus. Es gibt auch mehrere Methoden für Hand-Objekt-Interaktionen (Zhang, Ye et al., 2021; Christen et al., 2022), die sich darauf konzentrieren, nur die Handgelenk- und Fingerbewegungen für das Greifen verschiedener Objekte zu erzeugen. Die Modellierung der Handbewegung allein reicht jedoch nicht aus, um einen plausiblen Bewegungsablauf für eine absichtsgesteuerte virtuelle Figur zu erstellen. Stattdessen wiesen wir darauf hin, dass es entscheidend ist, im Bereich der Ganzkörper-Bewegungssynthese zu arbeiten. Hierfür gibt es zwei Hauptgründe. Erstens ermöglicht die Synthese von Ganzkörperbewegungen ein breiteres Spektrum an Interaktionen. Bei verschiedenen Handlungen wie Essen, Trinken, Inspizieren, Weitergeben oder Austauschen von Gegenständen zwischen den Händen sind auch der Kopf, die Arme und der Oberkörper Teil der vollständigen Handlungssequenz (Taheri et al., 2020). Zweitens führt das triviale Anhängen der synthetisierten Handbewegung an den restlichen Körper (Puig et al., 2018) zu einer „unheimlichen“ (vom im diesem Zusammenhang geläufigeren englischen Ausdruck „uncanny“) und physikalisch unplausiblen Bewegungserzeugung. Darüber hinaus haben neuere Arbeiten (Taheri et al., 2022) die Fähigkeit gezeigt, Ganzkörper-Greifbewegungen von einer T-Pose bis zum Moment des Greifens zu erzeugen. Die Synthese einer plausiblen Bewegungssequenz nach dem ersten Greifmoment, die auf einer Intention basiert, die die Interaktion zwischen Mensch und Objekt steuert, blieb jedoch bis dahin unbehandelt.

Um diese Einschränkungen zu beheben, haben wir in XAINES IMoS ein neuartiges System zur Synthese verschiedener Ganzkörper-Bewegungssequenzen von Mensch-Objekt-Interaktionen entwickelt. Wir synthetisieren die Bewegungen auf der Grundlage von textuellen Anweisungen, die aus Aktionen (Absichten) und Objekten bestehen. Wir lernten verallgemeinerbare Absichtskodierungen aus den eingegebenen Absichts-Objekt-Paaren mit Hilfe eines CLIP-Encoders (Radford et al., 2021), einem groß angelegten Sprachmodell, das auf einem großen Korpus von Text-Bild-Paaren trainiert wurde. Ausgehend von den anfänglichen Körperhaltungen und den 3D-Objektpositionen haben wir ein absichtsgesteuertes Ganzkörperbewegungsgeneratormodell entwickelt, das autoregressiv Ganzkörperbewegungen erzeugt. Wir verfolgten einen entkoppelnden Ansatz und modellierten die Bewegungen der Arme und des Körpers mit Hilfe separater bedingter variabler Autoregressoren, um die ausgegebenen Arm- und Körperbewegungen präziser zu machen. Da es sich bei diesen Autoregressoren um Variationsmodelle handelt, konnten wir zum Zeitpunkt der Inferenz verschiedene Bewegungen aus dem latenten Raum auswählen. Wir haben auch festgestellt, dass die Regression der Bewegung aus einem größeren vergangenen Kontext entscheidend für die Modellierung der langfristigen zeitlichen Abhängigkeit zwischen den Gelenken ist. Zur Modellierung von Korrelationen zwischen den verschiedenen Gelenken haben wir eine positionskodierte Selbstaufmerksamkeitsabbildung verwendet, um ein breiteres Spektrum an Interaktionen zu ermöglichen. Wir führten eine Optimierungsroutine durch, um die entsprechenden 6-DoF-Objektpositionen relativ zur Handposition in jedem Bild zu schätzen. Die ermittelten Objektpositionen wurden für die zukünftige Bewegungssynthese verwendet. Wir trainierten und evaluierten unsere Methode anhand des aktuellen GRAB-Datensatzes (Taheri et al., 2020), der aus ~1,3K Sequenzen von Mensch-Objekt-Interaktionen besteht, die mehrere Absichten aufweisen. Wir bewerteten unsere synthetisierten Sequenzen quantitativ anhand etablierter Metriken wie dem mittleren Positionsfehler pro Gelenk, dem durchschnittlichen Varianzfehler, der Fréchet-Inception-Distanz, der Erkennungsgenauigkeit, der Diversität und der Multimodalität, um die Wirksamkeit des Modells zu testen. Darüber hinaus haben wir eine visuelle Wahrnehmungsstudie zur subjektiven Bewertung unserer synthetisierten Bewegungen im Vergleich zu aktuellen Methoden der bedingten Bewegungssynthese durchgeführt.
Zusammenfassend lässt sich sagen, dass unsere wichtigsten technischen Beiträge dreierlei sind:

- Ein neuer Rahmen für die Erzeugung verschiedener Bewegungssequenzen virtueller menschlicher Charaktere, die mit Objekten bekannter Form, die sich in ihrer Reichweite

befinden, gemäß textbasierten Anweisungen interagieren. Im Gegensatz zu früheren Arbeiten über Charakter-Objekt-Interaktionen optimiert unsere vorgeschlagene Methode auch die 6-DoF-Objektpositionen in 3D.
- Synthese von Interaktionen, an denen beide Hände beteiligt sind, einschließlich Sequenzen, in denen die Figur einen Gegenstand zwischen den Händen austauscht ("offhand") - eine bisher unerforschte Situation.
- Lernen separater latenter Variationseinbettungen für die Arme und den Rest des Körpers, um die Vielfalt der synthetisierten Bewegungen und die genaue Synthese von beidhändigen Interaktionen zu ermöglichen.

Eigene Veröffentlichungen:
Ghosh, A., Cheema, N., Oguz, C., Theobalt, C., & Slusallek, P. (2021). Synthese von kompositorischen Animationen aus textuellen Beschreibungen. 2021 IEEE/CVF International Conference on Computer Vision (ICCV).
Ghosh, A., Dabral, R., Golyanik, V., Theobalt, C., & Slusallek, P. (2023). IMoS: Intent-Driven Full-Body Motion Synthesis for Human-Object Interactions. Computer Graphics Forum, 42(2), 1-12.

WP1.2 Zuordnung von Sprache zu Sensorsignalen aus einer entsprechenden Aktivität.

Das Ziel von Arbeitspaket 1.2 war es, Sprache und Erklärungen auf Sensorsignale abzubilden, die menschlichen Aktivitäten entsprechen. Dies beinhaltet die Erstellung von Modellen, die Freiformsprache oder Text verarbeiten und mit physischen Aktionen oder Ereignissen, die durch Sensordaten erfasst werden, abgleichen können. Die Herausforderung bestand darin, dass textliche Beschreibungen oft frei sind und nicht unbedingt mit den Sensorsignalen übereinstimmen. WP1.2 arbeitete an der Verbesserung der Art und Weise, wie diese Erklärungen - ob in gesprochener oder schriftlicher Form - auf die Sensordaten abgebildet werden können, indem fortgeschrittene Techniken des multimodalen Lernens eingesetzt wurden. Diese Aufgabe erweiterte die Arbeit von WP1.1, die sich auf die Erstellung gemeinsamer sensor- und sprachbasierter Repräsentationen konzentrierte, indem sie die Fähigkeit hinzufügte, reichhaltige, genaue Erklärungen zu generieren, die physischen Aktivitäten oder Bewegungen entsprechen, die von Sensoren erfasst wurden. Ziel war es, die semantische Abdeckung und Präzision dieser Zuordnungen zu verbessern, um robustere, erklärbare KI-Ergebnisse über mehrere Bereiche hinweg zu gewährleisten.

SDS: Vollständiger Satz als Eingabe für die Erzeugung natürlicher Bilder
Eines der Hauptziele von WP1.2 war die Abbildung von Freiform-Sprache und textuellen Erzählungen auf entsprechende Sensordaten in einer Weise, die den Reichtum und die Variabilität der natürlichen Sprache erfasst. Herkömmliche Methoden schränkten textliche Beschreibungen oft durch vordefinierte Regeln ein, wodurch die Ausdruckskraft der generierten Erklärungen begrenzt wurde und die Nuancen der freien Sprache oder Schrift nicht berücksichtigt werden konnten. Um dieses Problem anzugehen, haben wir die Arbeit von Frolov et al. (2021) aus WP1.1 erweitert, indem wir die Generierung von Bildern auf komplexere Freitextbeschreibungen konditioniert haben, wie von Jolly et al. (2022) gezeigt. Anstatt die Beschreibungen auf einfache Phrasen wie "roter Bus" zu beschränken, ermöglichten wir die Verwendung von vollständigen, natürlichen Sätzen wie "Ein roter Bus auf der Straße". Dieser Ansatz nutzte die Flexibilität eines großen Sprachmodells und führte einen multimodalen Region-Matching-Verlust ein, der den semantischen Abgleich zwischen den textlichen Beschreibungen und den generierten Bildern förderte.
Dieser Beitrag ist für WP1.2 von großer Bedeutung, da er die Fähigkeit zur Verarbeitung von Freiformsprache oder Text und deren Abbildung auf Sensorsignale verbessert. Indem wir detailliertere, uneingeschränkte Textbeschreibungen ermöglichten, verbesserten wir die Reichhaltigkeit und Genauigkeit der generierten Erzählungen, was für die Schaffung robuster und

erklärbarer KI-Systeme entscheidend war. Die Integration des multimodalen Abgleichs stellte sicher, dass die Erzählungen kontextuell korrekt und auf die Sensordaten abgestimmt waren, was dem Ziel, kohärente, für den Menschen verständliche Erklärungen aus den Sensoreingaben zu generieren, näher kam.

Eigene Veröffentlichungen:
Jolly, S., Zhang, Z. X., Dengel, A., & Mou, L. (2022). Suchen und Lernen: Verbesserung der semantischen Abdeckung für Data-to-Text-Generierung. *Proceedings of the AAAI Conference on Artificial Intelligence, 36*(10), 10858-10866.

EI, MLT & ASR: Synthese von Animationen und Sensordaten mit Bewegungserklärungen
In diesem Beitrag wurde die Verwendung großer Sprachmodelle wie GPT-2 und GPT-3 untersucht, um detaillierte Textbeschreibungen für menschliche Bewegungsdaten zu generieren und die Herausforderung begrenzter kommentierter Datensätze zu bewältigen. Durch die Feinabstimmung dieser Modelle an bestehenden Datensätzen und die Einbeziehung von realen Bewegungsmerkmalen zielte der Ansatz darauf ab, die Zuordnung zwischen Text und Bewegungssignalen zu verbessern.
Diese Arbeit ist für das Arbeitspaket 1.2 relevant, das sich auf die Zuordnung von Freiformtext oder Sprache zu Sensorsignalen konzentriert, um kohärente Erklärungen zu erzeugen. Die Verwendung von Sprachmodellen zur Generierung umfassenderer Annotationen hilft, den Mangel an detaillierten Trainingsdaten zu überwinden, der für die Verbesserung der Genauigkeit von Text-zu-Bewegung-Zuordnungen wesentlich ist. Durch die weitere Verbesserung dieser Modelle mit bewegungsspezifischen Merkmalen (z. B. Anzahl der Schritte, Startschritt) entspricht dieser Beitrag dem Ziel von WP1.2, die Beziehung zwischen Texteingabe und Sensordaten zu verfeinern und präzisere und aussagekräftigere Erklärungen zu ermöglichen, die eng mit den realen Bewegungssignalen verbunden sind, die sie darstellen.

Da die Annotation von Bewegungs- (Posen) oder Sensordaten (IMU) für die Erkennung menschlicher Aktivitäten zeitaufwändig und teuer ist, sind die vorhandenen Datensätze rar und oft fehlen präzise Beschreibungen der Aktivitäten. Für das Training von Systemen, die Texteingaben auf Posen oder IMU-Signale abbilden, sind jedoch hochwertige Textkommentare erforderlich. Das Aufkommen immer größerer und leistungsfähigerer Sprachmodelle in den letzten Jahren, wie z. B. GPT-2 (Radford et al., 2019), GPT-3 (Brown et al., 2020) und ChatGPT, bietet die Möglichkeit, solche umfangreichen Signalbeschreibungen automatisch zu erstellen.
In unserer Arbeit haben wir Trainingsdaten für die textbasierte Vorhersage menschlicher Bewegungen erzeugt. Im Einzelnen verwendeten wir GPT-2 und GPT-3, um feinkörnige Beschreibungen aus oberflächlichen Eingaben vorherzusagen. Wir arbeiteten mit dem KIT Motion-Language Dataset (Plappert et al., 2016), das 3.911 menschliche Ganzkörper-Bewegungsbeispiele enthält. Jedes Beispiel besteht aus einer (möglicherweise leeren) Liste von Textannotationen, die die Aktivität beschreiben (z. B. "Eine Person dreht sich nach rechts.", "Eine Person geht langsam.").
Zunächst haben wir den Datensatz aus linguistischer Sicht analysiert. Insbesondere haben wir die häufigsten Verben, Adjektive und Adverbien aus dem Datensatz extrahiert und diese verwendet, um Verb-Adjektiv- und Verb-Adverb-Paare zu erstellen. Anschließend erstellten wir Trainingsdaten für die Feinabstimmung von GPT-2 und GPT-3, indem wir 500 Annotationen mit diesen Paaren zufällig auswählten. Wir renderten Videos mit den Bewegungsdaten dieser Beispiele. Schließlich haben wir die Bewegungen beim Betrachten dieser Videos manuell genau annotiert. Insbesondere fügten wir fehlende Informationen zu den folgenden Merkmalen hinzu: Anzahl der Schritte, Geschwindigkeit, Richtung und Startschritt (d. h. Start mit dem rechten Fuß oder Start mit dem linken Fuß). Nachdem wir die Daten annotiert hatten, erstellten wir den endgültigen Datensatz mit den Textbeschriftungen des KIT-Datensatzes als Quelle und unseren Annotationen als Ziel.

Mit diesen Daten haben wir GPT-2 und GPT-3 im Modus der bedingten Generierung feinabgestimmt: Bei einer Eingabeaufforderung *x* wird das Modell so trainiert, dass es einen Abschluss *y* vorhersagt. Dann verwendeten wir beide Modelle, um qualitativ hochwertige, feinkörnige Textbeschreibungen für die verbleibenden Annotationen des KIT Motion-Language Dataset zu generieren. Schließlich ersetzten wir den Kodierer des Joint Language-to-Pose-Modells (JL2P) (Ahuja & Morency, 2019), das ursprünglich auf dem KIT Motion-Language-Datensatz trainiert wurde, durch ein Two-Stream-BERT-Modell, um die Informationen aus den ursprünglichen Annotationen des KIT Motion-Language-Datensatzes und die von GPT-{2/3} generierten Labels zu kombinieren. Unser Two-Stream-Modell war in der Lage, Verben mit Adjektiven, Adverbien und Merkmalen auf sinnvolle Weise zu verbinden. Unser Ansatz zeigte jedoch keine Verbesserung im Vergleich zum ursprünglichen JL2P-Modell in Bezug auf den durchschnittlichen Positionsfehler. Der Hauptgrund dafür ist, dass die GPT-Modelle oft falsche Beschreibungen liefern. Bei der Analyse haben wir festgestellt, dass die Sprachmodelle Probleme bei der Vorhersage der Anzahl der Schritte und der Richtung der Pose haben, um die Beschreibung der Bewegung zu erstellen. Wir sind davon ausgegangen, dass die auf realen Ereignissen basierenden Eigenschaften von Bewegungen nicht vom Sprachkontext abhängen. Daher sind Sprachmodellierungsansätze nicht in der Lage, mit den Eigenschaften von realen Ereignissen umzugehen, z. B. Anzahl der Schritte, linke oder rechte Hand zum Winken usw. Dieses Problem führte zu einer zufälligen Vorhersage von realen Ereigniseigenschaften der Bewegungsbeschreibung in Sprachmodellen und somit zu Rauschen im Bewegungssynthesemodell.
Diese Arbeit ist noch nicht abgeschlossen und wurde noch nicht veröffentlicht. Wir arbeiten derzeit daran, spezifische Merkmale wie die genaue Anzahl der Schritte und den Startschritt aus dem Bewegungssignal zu extrahieren. Diese Merkmale werden als zusätzliche Eingaben für die Sprachmodelle (z. B. GPT-{2/3}) dienen, so dass das Modell diese Informationen weniger ableiten muss. Indem wir diese Details direkt aus der Eingabe bereitstellen, wollen wir die Fähigkeit des Modells verbessern, diese Informationen in der generierten Ausgabe genau zu positionieren. Diese Arbeit wird im Rahmen zukünftiger Forschungsprojekte fortgesetzt.

EI, MLT: Ein multimodales gemeinsames Merkmalsraumlernen für die Erkennung menschlicher Aktivitäten

Die Erkennung menschlicher Aktivitäten (Human Activity Recognition, HAR) identifiziert und klassifiziert verschiedene Arten menschlicher Bewegungen und Handlungen anhand von Sensordaten, z. B. Daten von Beschleunigungsmessern, Gyroskopen und anderen tragbaren Geräten. HAR hat sich in den letzten Jahren zu einem wichtigen Forschungsgebiet entwickelt, da es ein breites Spektrum von Anwendungen verbessern kann, darunter das Gesundheitswesen (Tang et al., 2020), Sport und Fitness (Host & Ivašić-Kos, 2021), Sicherheit (Sunil et al., 2021) und Robotik (Piyathilaka & Kodagoda, 2015). Mehrere Ansätze für HAR umfassen maschinelle Lerntechniken wie künstliche neuronale Netze, Entscheidungsbäume und Support-Vector-Machines, die auf Datensätzen mit Sensordaten und entsprechenden Kennzeichnungen trainiert werden, die die Art der ausgeführten Aktivität angeben. Die Bedeutung von HAR liegt in seiner Fähigkeit, wertvolle Einblicke in das menschliche Verhalten zu geben und verschiedene Anwendungen zu ermöglichen, die das Leben der Menschen verbessern können. Sie kann zur Überwachung und Verfolgung des körperlichen Aktivitätsniveaus bei Patienten mit chronischen Erkrankungen (Tan et al., 2021) wie Diabetes oder Herzkrankheiten eingesetzt werden, um einen gesünderen Lebensstil zu fördern und Komplikationen zu vermeiden.

Unser MuJo-Ansatz (Multimodal Joint Feature Space Learning) bietet mehrere wichtige Beiträge zu XAINES WP1.2, indem er die technischen und konzeptionellen Herausforderungen bei der Erstellung eines multimodalen gemeinsamen Merkmalsraums für erklärbare KI, insbesondere im Kontext von HAR, angeht. Der Hauptbeitrag von MuJo liegt in seiner Fähigkeit, Daten aus verschiedenen

Modalitäten, einschließlich Video-, Sprach-, Pose- und Sensordaten, in einen einzigen, vereinheitlichten gemeinsamen Merkmalsraum zu integrieren. In XAINES ermöglicht dieser gemeinsame Merkmalsraum dem KI-System, sein Verhalten auf eine für den Menschen verständliche Weise zu interpretieren und zu erklären, indem es Verbindungen zwischen den verschiedenen Modalitäten herstellt. Durch die Kombination von Videodaten (visuelle Hinweise), Sprache (Beschriftungen oder Erklärungen), Posenschätzung (menschliche Körperbewegung) und IMU-Sensordaten (Bewegung) kann das System beispielsweise detailliertere, kontextreiche Erklärungen für menschliche Aktivitäten liefern. Diese Art der multimodalen Fusion ist entscheidend für die Erzeugung natürlicher Erklärungen, was eines der Ziele von XAINES WP1.2 ist. MuJo nutzt kontrastives Lernen, um die verschiedenen Modalitäten in einen gemeinsamen Repräsentationsraum einzubinden, in dem ähnliche Aktivitäten oder Handlungen über die Modalitäten hinweg gruppiert werden. Dies ermöglicht dem KI-System eine bessere Generalisierung, wenn eine oder mehrere Modalitäten fehlen oder unvollständig sind. Für XAINES WP1.2 ist dies von großer Bedeutung, da in realen Szenarien Daten aus einigen Modalitäten (z. B. Video- oder Sensordaten) aufgrund von Datenschutzbedenken oder technischen Einschränkungen nicht verfügbar oder unvollständig sein können. Die MuJo-Methode stellt sicher, dass das KI-System auch unter diesen ressourcenarmen oder unvollständigen Datenbedingungen robust funktioniert und kohärente Erklärungen liefern kann. MuJo zeigt, dass es bei der Erkennung menschlicher Aktivitäten auch mit einer begrenzten Menge an annotierten Daten eine hohe Leistung erzielen kann, wobei in einigen Szenarien nur 2 % der Trainingsdaten verwendet werden. Dieser Aspekt kommt WP1.2 direkt zugute, da er die Abhängigkeit von großen, vollständig beschrifteten Datensätzen verringert. KI-Systeme, die ihr Verhalten erklären müssen, benötigen in der Regel qualitativ hochwertige, beschriftete Daten, um aussagekräftige Erkenntnisse zu liefern. Im XAINES-Kontext stellt diese Fähigkeit sicher, dass das KI-System auch in datenarmen Umgebungen Erklärungen generieren kann. Dadurch wird das System skalierbarer und kann in realen Szenarien eingesetzt werden, in denen die Beschaffung großer markierter Datensätze unpraktisch ist.

MuJos Verwendung von synchronisierten multimodalen Daten (z. B. die Kombination von Video mit IMU-Sensorwerten und Textbeschreibungen) erhöht die Tiefe der Erklärungen, die XAINES zu liefern versucht. WP1.2 zielt auf die Entwicklung von KI-Systemen ab, die nicht nur erklären können*, welche* Entscheidungen getroffen werden, sondern auch*, warum* diese Entscheidungen getroffen wurden, indem sie den Entscheidungsprozess mit den sensorischen Daten verknüpfen, die ihm zugrunde liegen. Der Trainingsprozess von MuJo stellt sicher, dass das System diese verschiedenen Modalitäten auf sinnvolle Weise miteinander verknüpfen kann, was zu kohärenteren und besser interpretierbaren Ergebnissen führt.
In XAINES könnte das System beispielsweise erklären, dass eine bestimmte (per Video erkannte) Aktion aufgrund bestimmter Bewegungsmuster (aus Posen- oder Sensordaten) erkannt wurde, die dann mit ähnlichen, in den Trainingsdaten beschriebenen Aktionen korreliert wurden. Dieser Grad an detaillierter, multimodaler Erklärung entspricht genau den Zielen von WP1.2.

Eigene Veröffentlichungen

Fritsch, S. G., Oguz, C., Rey, V. F., Ray, L., Kiefer-Emmanouilidis, M., & Lukowicz, P. (2024). MuJo: Multimodal Joint Feature Space Learning for Human Activity Recognition. *arXiv preprint arXiv:2406.03857*. (noch in Prüfung).

WP 1.3 Informationsextraktion aus Text und Bild für visuelle Erklärungen und Geschichtenerzählen

Das Hauptziel dieses Arbeitspakets ist die Erweiterung bestehender Methoden zur Bildbeschriftung und zum Geschichtenerzählen durch die Extraktion von Informationen aus Text und Bildern, um Erklärungen zu generieren. Innerhalb des XAINES-Projekts trägt WP1.3 zu dem weiter gefassten Ziel bei, KI zu erklären, indem textliche und visuelle Daten miteinander verknüpft werden. Es bietet die Möglichkeit, visuelle Daten in Erklärungen umzuwandeln, die für die Benutzer verständlich sind. Durch die Verbesserung der Bildbeschriftung und der Layout-zu-Bild-Generierung unterstützt dieses Arbeitspaket direkt die Erstellung von visuellen Erzählwerkzeugen, die die Transparenz und Erklärbarkeit von KI-Systemen erhöhen. Die Arbeit in WP1.3 legt den Grundstein für die Generierung kohärenter Erklärungen von KI-Modellen durch die Kombination von textlichen Erklärungen mit visuellen Ausgaben, was für das übergreifende Ziel des Projekts, verständlichere und interaktive KI-Erklärungen in verschiedenen Bereichen zu erstellen, unerlässlich ist.

- IML: Erläuterung von Bildern mit mehrsprachigen Bildunterschriften und Objektbeschreibungen

Wir erforschten fortschrittliche Methoden der Bildbeschriftung durch die Kombination von Top-down- und Bottom-up-Merkmalen, um visuelle Erklärungen aus Bildern zu generieren. Diese Methode basierte auf der Architektur Show, Attend and Tell, erweitert um zusätzliche objektspezifische saliente Regionen. Diese hervorstechenden Regionen wurden mithilfe des Mask R-CNN-Modells (He et al., 2017) erkannt, das Bounding Boxes für Objekte innerhalb von Bildern generierte, was detailliertere und kontextreichere Beschriftungen ermöglichte.
Unser Ansatz nutzte die Balkensuche und das Re-Ranking, um die Vielfalt der generierten Beschriftungen zu verbessern. Durch die Nutzung sowohl von High-Level-Merkmalen (Bildabstraktionen) als auch von Low-Level-Objekt-spezifischen Details produzierte das Modell Beschriftungen, die den visuellen Inhalt genauer beschreiben, was zu besseren BLEU-Werten führte, insbesondere bei längeren n-Grammen (BLEU-3 und BLEU-4). Dies verbesserte auch die Übereinstimmung zwischen Bildobjekten und der generierten natürlichen Sprache, was einen Schritt in Richtung erklärbare KI (XAI) bei visuellen Aufgaben darstellt.
Darüber hinaus wurden interaktive Modellverbesserungen durch einen Re-Ranking-Mechanismus ermöglicht, der es den Nutzern erlaubte, die Auswahl der Bildunterschriften auf der Grundlage von Rückmeldungen zu verfeinern und so ein interaktiveres maschinelles Lernen (IML) zu ermöglichen. Die Ergebnisse deuten darauf hin, dass die Integration von Bottom-up-Objektmerkmalen mit Top-down-Aufmerksamkeitsmechanismen eine vielversprechende Richtung für die Erstellung von erklärenden Bildbeschriftungssystemen ist.
Dieser Beitrag steht im Einklang mit dem Schwerpunkt von WP1.3 auf Informationsextraktion und visuellem Storytelling, der Verbesserung der Fähigkeit, mehrsprachige Untertitel zu generieren, und der Verbesserung der Transparenz von KI-generierten Inhalten für Endnutzer.
Darüber hinaus konzentrierten wir uns auf die Verbesserung der Generierung von deutschen Bildunterschriften durch fortschrittliche maschinelle Übersetzung und Transfer-Learning-Techniken, um die Herausforderung der begrenzten deutschsprachigen Trainingsdaten zu bewältigen. Da es nur wenige kommentierte Bilddaten in deutscher Sprache gibt, nutzte das Projekt den gut ausgestatteten MS COCO-Datensatz in englischer Sprache und übersetzte ihn mit dem neuronalen maschinellen Übersetzungstool Fairseq (Ott et al., 2019) ins Deutsche.
Es wurden mehrere Methoden zur Verbesserung der deutschen Untertitelung getestet. Die anfänglichen Basismodelle wurden entweder auf dem übersetzten MS COCO-Datensatz oder dem kleineren Multi30K-Datensatz trainiert, der deutsche Untertitel enthält. Deutliche Verbesserungen wurden jedoch erzielt, indem die Modelle zunächst auf dem größeren übersetzten MS COCO-Datensatz trainiert und anschließend auf dem deutschen Multi30K-Datensatz feinabgestimmt wurden. Dieser Ansatz ermöglichte es dem Modell, zunächst allgemeine Muster für

Bildunterschriften aus dem größeren Datensatz zu lernen, bevor es sich an die sprachlichen Nuancen des Deutschen anpasste.
Eine der effektivsten Techniken war die Integration eines erweiterten Aufmerksamkeitsmechanismus (Biswas et al., 2020), der die Fähigkeit des Modells verbesserte, sich auf objektspezifische Merkmale in Bildern zu konzentrieren. Diese Methode erzielte eine Verbesserung der BLEU-4-Werte um 21,2 % gegenüber dem derzeitigen Stand der Technik für deutsche Bildunterschriften. Durch die Kombination von Bildabstraktionen auf hoher Ebene mit lokalisierten Objektdetails war das Modell in der Lage, genauere und kontextuell relevante deutsche Bildunterschriften zu erzeugen.
Der Erfolg dieser Methoden trägt zum Gesamtziel von XAINES bei, mehrsprachige KI-Erklärungen zu erstellen. Durch den innovativen Einsatz von maschineller Übersetzung und Feinabstimmung hat WP1.3 gezeigt, dass robuste, qualitativ hochwertige Erklärungen in deutscher Sprache selbst mit begrenzten muttersprachlichen Trainingsdaten erstellt werden können. Diese Arbeit unterstützt das übergeordnete Projektziel, erklärungsfähige KI-Systeme zu schaffen, die transparent, zugänglich und mehrsprachig sind und den Nutzern verständliche und sprachlich vielfältige KI-generierte Erzählungen bieten.

Eigene Veröffentlichungen:

Biswas, R., Barz, M., & Sonntag, D. (2020). Explanatory Interactive Image Captioning using Top-Down and Bottom-Up Features, Beam Search and Re-ranking. KI - Künstliche Intelligenz 34.
Biswas, R., Barz, M., Hartmann, M., & Sonntag, D. (2021). Improving German Image Captions Using Machine Translation and Transfer Learning. International Conference on Statistical Language and Speech Processing, 3-14.

SDS: Ein Curriculum-Lernansatz mit progressiver Unschärfe auf Objektebene für eine verbesserte Layout-zu-Bild-Generierung

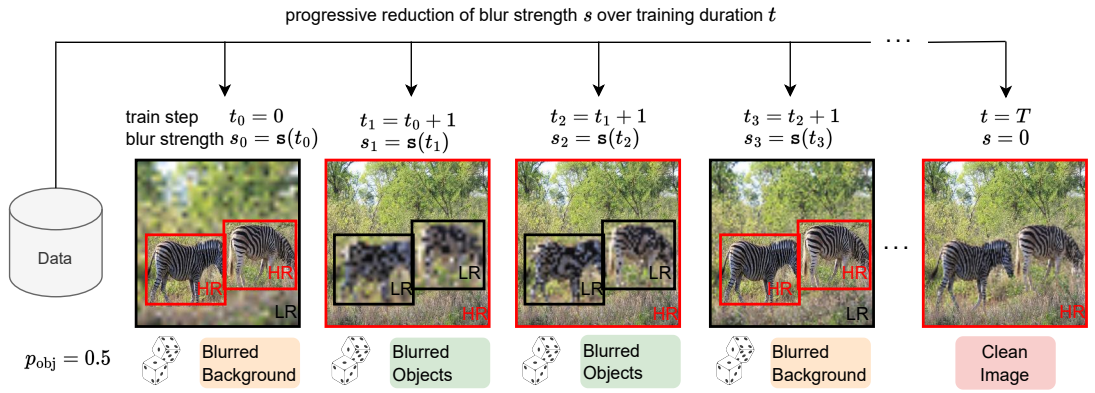

Abbildung 3: Unsere ObjBlur-Methode beinhaltet einen neuartigen Lernansatz, der auf der progressiven Unschärfe einzelner Objekte oder des Hintergrunds auf Objektebene während des gesamten Trainingsverfahrens auf einer prozentualen Basis basiert.

In WP1.3 führten wir außerdem ObjBlur ein, einen neuartigen Lernansatz für die Generierung von Layouts zu Bildern, der eine Schlüsselrolle bei der Erreichung des Projektziels der Erstellung erklärbarer visueller Erzählungen spielte. ObjBlur verbesserte die Generierung realistischer Bilder aus Layouts, die aus Begrenzungsrahmen und Beschriftungen bestehen, indem es schrittweise Unschärfe auf Objektebene anwendet. Diese Strategie erhöhte effektiv die Trainingsstabilität und verbesserte die Qualität der generierten Bilder, was für die Erstellung verständlicher visueller Erklärungen in KI-Systemen unerlässlich war.

ObjBlur beginnt mit stark unscharfen Bildern und setzt das Modell schrittweise saubereren Bildern aus, wodurch eine kontrollierte Steigerung der Trainingskomplexität erreicht wird. Diese Technik des Curriculum-Lernens befasst sich mit häufigen Problemen bei generativen Modellen, wie z. B. dem Zusammenbruch des Modus und der Überanpassung, insbesondere bei Modellen zur Erzeugung von Layouts und Bildern. Die Möglichkeit, die Schwierigkeit der Aufgabe stufenlos anzupassen, ermöglichte es dem Modell, ein tieferes Verständnis für Objektbeziehungen und kontextuelle Positionierung zu entwickeln, was für die Generierung detaillierterer und erklärender Bildunterschriften entscheidend war.
In WP1.3, wo der Schwerpunkt auf der Integration von visuellen Elementen in Erklärungen lag, verbesserte ObjBlur die Qualität der Bilder, die als Grundlage für diese Erklärungen dienten. Durch die Stabilisierung des Generierungsprozesses ermöglichte es die Erstellung genauerer und kontextreicher Bildunterschriften, die entscheidend dafür waren, dass die von der KI generierten Ergebnisse für die Benutzer verständlicher wurden. Die Kompatibilität von ObjBlur sowohl mit generativen adversen Netzen (GANs) als auch mit Diffusionsmodellen stellte zudem sicher, dass es in verschiedenen generativen Modellierungsparadigmen eingesetzt werden konnte, was seine Vielseitigkeit bei der Verbesserung der Layout-zu-Bild-Generierung für das XAINES-Projekt unterstrich.

Die Generierung von Layouts zu Bildern ist eine grundlegende Aufgabe in der Computer Vision und Grafik, die die Lücke zwischen strukturierten Szenenbeschreibungen, wie Layouts, die aus Bounding Boxes und Labels bestehen, und der Generierung realistischer Bilder schließt (He et al., 2021; Zheng et al., 2023). Es handelt sich um eine komplexe Aufgabe, die noch dadurch erschwert wird, dass das Erlernen der Generierung verschiedener Objektklassen und deren inhärente Vielfalt an Formen, Größen und Kontexten unterschiedlich schwierig ist. Layout-to-Image-Modelle basieren hauptsächlich auf GANs (Goodfellow et al., 2014) und erben daher deren Probleme mit der Trainingsstabilität, wie z. B. Mode Collapse und Overfitting (Salimans et al., 2016). Während sich Techniken zur Datenerweiterung (DA) in visuellen Erkennungsmodellen als wirksam erwiesen haben (Xu et al., 2023), führt das Training eines GAN unter ähnlichen Erweiterungen zu einem Leckeffekt, bei dem der Generator lernt, erweiterte (statt saubere) Bilder zu erzeugen. Wird zum Beispiel Rotation als DA verwendet, produziert der Generator nach dem Training gedrehte Bilder, was ein unerwünschtes Ergebnis ist. Um dieses Problem zu entschärfen, wurden Verfahren zur Regularisierung der Konsistenz (Zhang et al., 2020), zur Invertierbarkeit (Tran et al., 2021) und zur differentiellen Augmentation (Zhao et al., 2020) vorgeschlagen. In der Zwischenzeit hat sich die Gemeinschaft des maschinellen Lernens für Curriculum-Learning-Strategien (CL) interessiert, um Trainingsbeispiele in einer sinnvollen Reihenfolge zu strukturieren, die das Modell allmählich mit komplexeren Konzepten vertraut macht. Sie bieten einen intuitiven Ansatz, um Modelle durch immer anspruchsvollere Trainingsszenarien zu führen. Interessanterweise ist ihre Erforschung im Zusammenhang mit generativen Modellen noch relativ begrenzt (Soviany et al., 2022) und im Bereich der Layout-Bild-Generierung nicht vorhanden. Für generative Einzelobjekt-Bildmodelle wurde in früheren Arbeiten die Verwendung mehrerer Diskriminatoren vorgeschlagen (Doan et al., 2019), die das Modell schrittweise erweitern oder Bilder nach Schwierigkeitsgrad einstufen. Alle bisherigen Arbeiten erfordern jedoch entweder eine Änderung des Modells, der Verlustfunktion, die Verwendung eines Schwierigkeitsschätzers oder eine Kombination aus beidem. Unseres Wissens gab es vor unserer Arbeit in XAINES keine Arbeit, die sich mit der Verwendung von Curriculum-Lernen für die Layout-zu-Bild-Generierung befasste. Daher stellt ObjBlur einen neuen Ansatz für die Layout-Bild-Generierung dar, der das Curriculum-Lernen durch Anwendung einer progressiven Unschärfe auf Objektebene nutzt. Unschärfe ist ein natürlicher Vorgang der Bildverschlechterung, da niedrige Frequenzen gegenüber höheren Frequenzen beibehalten werden. Bei starker Unschärfe werden hochfrequente Details entfernt, was zu einem einfacheren Signal führt, ohne den strukturellen Inhalt des Bildes zu beeinträchtigen. Eine Verringerung der Stärke der Unschärfe führt zu einem

komplexeren Signal mit hochfrequenten Details, was das Modell einer schwierigeren Aufgabe aussetzt. Daher bietet die Unschärfe einen intuitiven und leistungsstarken Ansatz, um die Schwierigkeit der Aufgabe schrittweise anzupassen und einen reibungslosen Trainingsverlauf zu gewährleisten. Unsere Methode kann durch eine einfache Modifikation des Datenladers realisiert werden, um eine progressive Unschärfe auf die Bilder anzuwenden. Daher kann sie leicht in bestehende Layout-zu-Bild-Ansätze integriert werden und ist nicht von Schwierigkeitsschätzern oder Änderungen in der Modellarchitektur und dem Optimierungsprotokoll abhängig. Durch die systematische Anwendung unterschiedlicher Unschärfegrade während des Trainings, beginnend mit starker Unschärfe und fortschreitend zu saubereren Bildern, stabilisieren wir das Training und stellen sicher, dass das Modell lernt, qualitativ hochwertige Bilder zu erzeugen. Ein entscheidender Aspekt der Bildqualität ist das Aussehen der Objekte im Vordergrund im Verhältnis zum Hintergrund. Daher schlagen wir einen Ansatz auf Objektebene vor, bei dem die Unschärfe nach dem Zufallsprinzip entweder auf die Objekte oder auf den Hintergrund angewendet wird. Um die Vorteile von ObjBlur zu demonstrieren, haben wir eine umfassende Analyse verschiedener Modelle zur Erzeugung von Layout-Bildern durchgeführt, darunter adversarial- (He et al., 2021) und diffusionsbasierte (Zheng et al., 2023) Ansätze. ObjBlur verbessert die Qualität des erzeugten Bildes erheblich und bietet einen robusten und vielseitigen Ansatz. Abschließend analysieren wir umfassend verschiedene Designentscheidungen und ihre Auswirkungen auf Leistung und Stabilität.

Eigene Veröffentlichungen:
Frolov, S., Moser, B. B., Palacio, S., & Dengel, A. (2024). ObjBlur: A Curriculum Learning Approach with Progressive Object-Level Blurring for Improved Layout-to-Image Generation. ACM Multimedia 2024. (Zur Veröffentlichung angenommen).

WP 1.4 Generierung von Erzählungen für die visuelle Beantwortung von Fragen

Im Rahmen des XAINES-Projekts, dessen Ziel es war, die Erklärbarkeit von KI-Systemen zu verbessern, zielt WP1.4 speziell auf die Erzeugung kohärenter und informativer Erklärungen für visuelle Fragebeantwortungssysteme (VQA) ab. VQA ist eine anspruchsvolle Aufgabe, die von KI-Systemen verlangt, dass sie sowohl ein Bild als auch eine entsprechende natürlichsprachliche Frage verstehen, um angemessene Antworten zu generieren. Das System muss jedoch nicht nur korrekte Antworten liefern, sondern auch seine Überlegungen in einer Weise erläutern, die mit den menschlichen Erwartungen übereinstimmt und den Entscheidungsprozess transparenter macht.

SDS: Bewertungsmaßstab, der das Verhältnis zwischen der Schwierigkeit der Frage und der Konsistenz der Antworten ausdrückt.
Zu diesem Zweck haben wir EaSe (Entropy and Semantic-based Evaluation) entwickelt, ein Diagnosewerkzeug, mit dem der Schwierigkeitsgrad von VQA-Proben auf der Grundlage der von mehreren Annotatoren gegebenen Antwortmuster bewertet werden kann (Jolly et al., 2021).
Das EaSe war für zwei wichtige Faktoren verantwortlich:

1. Entropie: Diese Metrik misst, wie verstreut oder konsistent die Antworten waren, wobei höhere Entropiewerte auf größere Uneinigkeit zwischen den Kommentatoren hinweisen und somit eine schwierigere Frage signalisieren. Niedrigere Entropiewerte, bei denen sich die Bewerter häufiger einig waren, wiesen auf einfachere Fragen hin.
2. Semantischer Inhalt: Neben dem einfachen String-Matching bewertete EaSe auch die semantische Ähnlichkeit der Antworten. Selbst wenn die Bewerter unterschiedliche Wörter verwendeten, konnten diese semantisch ähnlich sein (z. B. "kariert" und "kariert"). Dies ermöglichte es uns, die tatsächliche Schwierigkeit der Frage besser zu erfassen, indem wir ähnliche Antworten in Gruppen zusammenfassten.

Das Ziel von EaSe war es, die schwierigsten und informativsten Proben aus VQA-Datensätzen zu identifizieren. Durch die Diagnose der Schwierigkeit von Fragen konnten wir diese schwierigen Proben beim Training und der Feinabstimmung von VQA-Modellen priorisieren. Unsere Experimente mit VQA-Datensätzen wie VQA2.0 (Goyal et al., 2017) und VizWiz (Gurari et al., 2018) zeigten, dass Modelle, die auf schwierigeren Proben - die von EaSe identifiziert wurden - trainiert wurden, eine bessere Gesamtleistung erzielten und weniger Daten benötigten, um ähnliche Genauigkeitsniveaus zu erreichen wie Modelle, die auf vollständigen Datensätzen trainiert wurden. Diese Erkenntnis bestätigte, dass schwierigere Fragen reichhaltigere multimodale Informationen enthielten, die für das Modelllernen wertvoller waren.
Darüber hinaus lieferte EaSe wertvolle Einblicke in die Übereinstimmung der menschlichen Annotatoren und die Schwierigkeit der Menschen bei der Beantwortung visueller Fragen. Durch die Korrelation der EaSe-Bewertungen mit dem Vertrauensniveau der Menschen konnten wir zeigen, dass ein höheres Vertrauen der Kommentatoren mit leichteren Fragen korrespondiert, während ein geringeres Vertrauen auf größere Schwierigkeiten hindeutet. Diese Übereinstimmung mit der menschlichen Intuition unterstützt die Entwicklung von KI-Systemen, die stärker auf den Menschen ausgerichtet sind.
Im Rahmen von XAINES trug EaSe dazu bei, die Erklärbarkeit von KI-Systemen zu verbessern. Durch die Identifizierung schwieriger VQA-Proben und das Verständnis menschlicher Reaktionsmuster half EaSe bei der Erstellung detaillierterer und genauerer Beschreibungen, die KI-Entscheidungen erklären können.
Das EaSe-Tool wurde erfolgreich in die VQA-Modelltrainings-Pipelines integriert, was zu erheblichen Verbesserungen sowohl der Effizienz als auch der Erklärungsqualität führte.

Eigene Veröffentlichungen:
Jolly, S., Pezzelle, S., & Nabi, M. (2021). EaSe: A Diagnostic Tool for VQA Based on Answer Diversity. Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.

SDS: Ein Datensatz für sphärische Keypoint-Erkennung, Matching und Schätzung der Kameraposition

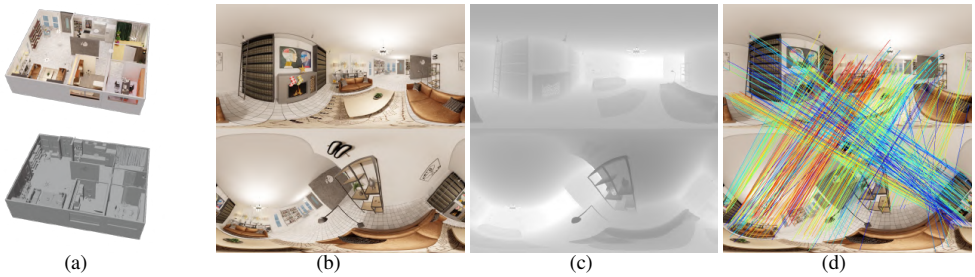

Abbildung 4: SphereCraft: Ein neuer Datensatz für sphärische Keypoint-Erkennung, Matching und Kamerapositionsschätzung.

Neben der Entwicklung von EaSe in WP1.4 war ein weiterer wichtiger Beitrag zu diesem Arbeitspaket die Integration neuartiger Datensätze zur Verbesserung von KI-Erklärungen im Zusammenhang mit visuellen Schlussfolgerungen. Insbesondere nutzten wir den SphereCraft-Datensatz, der umfassende Werkzeuge für die Erkennung von sphärischen Keypoints, den Abgleich und die Schätzung der Kameraposition bietet. Dieser Datensatz ermöglichte erhebliche Fortschritte bei der Fähigkeit der KI, komplexe visuelle Umgebungen zu verstehen und entsprechende Erklärungen zu generieren.

Der SphereCraft-Datensatz behebt die Einschränkungen bestehender Datensätze, indem er hochauflösende sphärische Bilder und die dazugehörigen Ground-Truth-Keypoints bietet. Diese Keypoints, die in verschiedenen synthetischen und realen Szenen erkannt werden, ermöglichen ein robusteres Verständnis der Szene in mehreren Ansichten und haben sich als wesentlich für die Verbesserung von Systemen zur visuellen Beantwortung von Fragen (VQA) in schwierigen Umgebungen erwiesen. Durch die Einbeziehung sowohl synthetischer als auch realer Daten verbesserte SphereCraft die Fähigkeit von VQA-Modellen, in realen Anwendungen, die eine Vielzahl von Blickwinkeln und starken Verzerrungen erfordern, gut zu funktionieren.

Im Rahmen des WP1.4 bot die Einbeziehung des SphereCraft-Datensatzes mehrere Vorteile:

- Verbessertes Multiview-Verständnis: Die sphärischen Keypoints und die entsprechenden Kamerapositionen verbesserten die Fähigkeit des Modells, visuelle Fragen zu behandeln, die komplexe räumliche Beziehungen und mehrere Blickwinkel beinhalten. Dies war besonders nützlich, um die Gründe für Antworten in Szenarien mit mehreren visuellen Perspektiven zu erklären.
- Reichhaltige multimodale Informationen: Durch die Verwendung der hochauflösenden synthetischen und realen Bilder aus dem SphereCraft-Datensatz konnten die VQA-Systeme detailliertere und kontextbewusste Erzählungen generieren. Diese Bilder wurden von Tiefenkarten, Kamerapositionen und Ground-Truth-Korrespondenzen begleitet und boten einen umfassenden Datensatz für das Training von Modellen, die in der Lage sind, anspruchsvolle Erzählungen zu generieren.
- Abgleich von Schlüsselpunkten und Tiefenschätzung: Die Konzentration des Datensatzes auf die Erkennung und den Abgleich von Schlüsselpunkten in Verbindung mit den originalgetreuen Tiefenkarten ermöglichte es dem System, räumliche Beziehungen in visuellen Szenen besser zu verstehen, was sich unmittelbar auf die Erklärungen auswirkte, die für visuelle Fragen erstellt wurden.

Sphärische Bilder haben in den Bereichen Computer Vision und Deep Learning aufgrund ihrer einzigartigen Fähigkeit, Informationen über die gesamte sichtbare Szene von einem einzigen Blickwinkel aus zu liefern, große Aufmerksamkeit erregt. Frühere Studien haben die Vorteile der Verwendung von Panoramabildern für Aufgaben des Szenenverständnisses, wie z. B. Objekterkennung (Chou et al., 2020; Zhang, Cui et al., 2021), semantische Segmentierung (Armeni et al., 2017; Chang et al., 2017) und Vorhersage der Raumaufteilung (Zou et al., 2018), durch die Nutzung der reichhaltigeren Kontextinformationen, die sie bieten, gezeigt. Modernste Ansätze stützen sich jedoch auf eine einzige Ansicht, um diese Aufgaben zu erfüllen, und vernachlässigen Informationen, die von benachbarten Ansichten bereitgestellt werden. Wir haben argumentiert, dass die Schlussfolgerungen über die zugrunde liegende Szene erheblich verbessert werden können, wenn mehrere Bilder verfügbar sind und die relativen Posen zwischen den Kameras bekannt sind. Daher spielten die Erkennung und das Matching von sphärischen Keypoints und Structure from Motion (SfM) eine zentrale Rolle. SphereCraft war ein neuartiger Datensatz für die Erkennung und den Abgleich von Keypoints und SfM auf sphärischen Bildern, der für die Weiterentwicklung des Stands der Technik bei verschiedenen Computer-Vision-Aufgaben von entscheidender Bedeutung war (und immer noch ist). Im Gegensatz zu bestehenden Datensätzen enthielt SphereCraft extrahierte Keypoints von einer Auswahl beliebter handgefertigter und erlernter Detektoren zusammen mit ihren Grundwahrheits-Korrespondenzen, was es den Forschern ermöglichte, Algorithmen zu entwickeln und zu bewerten, die auf mehrere Kamerastandpunkte abzielen. In Anlehnung an (Zheng et al., 2020) nutzten wir die Rechenleistung von Grafikkarten, um fotorealistische Szenen zu erzeugen und die Einschränkungen beim Scannen realer Szenarien zu überwinden. Darüber hinaus haben wir für alle synthetischen Szenen hochpräzise 3D-Netze erstellt.

Wir haben auch eine Reihe realer Szenen veröffentlicht, die mit zwei verschiedenen sphärischen Kameras aufgenommen wurden, so dass sowohl synthetische als auch reale Daten für Training und Tests zur Verfügung standen. Schließlich haben wir einen Standardsatz von gerenderten Szenen und eine Aufteilung in Training und Test vorgeschlagen, um die Bewertung und den Vergleich verschiedener Ansätze innerhalb der Forschungsgemeinschaft zu erleichtern. SphereCraft ist in verschiedenen Forschungsbereichen anwendbar, darunter die Tiefenvorhersage aus einer oder mehreren Ansichten (Won et al., 2020), die geometrisch dichte 3D-Rekonstruktion (Pagani et al., 2011), die Vorhersage der Raumaufteilung in Innenräumen (Zhang, Cui et al., 2021; Zou et al., 2018), die Synthese neuer sphärischer Ansichten (Habtegebrial et al., 2022) und sphärische Lichtfelder (Gava et al., 2018). Um die Zusammenarbeit und Reproduzierbarkeit zu fördern, haben wir freie Software und offen zugängliche Szenenmodelle verwendet, zusammen mit dem Code, der zum Rendern von Bildern, Tiefenkarten und Kamerapositionen erforderlich ist.

WP 1.5 Verknüpfung von Klassifizierungen von Situationen mit synthetischen Bewegungen

Das Ziel des Arbeitspakets 1.5 (*Verknüpfung von Situationsklassifizierungen mit synthetischen Bewegungen*) im XAINES-Projekt ist es, sprachbasierte narrative Modelle mit anderen Repräsentationen wie Sensordaten und Video zu verbinden. Eine zentrale Herausforderung, die in diesem Arbeitspaket angegangen wird, ist der Mangel an ausreichenden Trainingsdaten für tiefe neuronale Netze (DNNs) in menschlichen Handlungserkennungsmodellen, die in der Regel auf eine große Menge an annotierten Daten angewiesen sind.
Ziel war es, synthetische Trainingsdaten zu erzeugen, um das Problem der Datenknappheit zu lösen. Dies beinhaltete die Entwicklung eines lernbasierten Sprach-zu-Bewegungsmodells, das in der Lage ist, animierte 3D-Positionssequenzen zu erzeugen, die mehrere aufeinanderfolgende oder sich überlagernde Aktionen auf der Grundlage langer, zusammengesetzter Sätze darstellen. Das Modell erreichte fein abgestufte Zuordnungen zwischen natürlichen Spracheingaben und entsprechenden 3D-Positionssequenzen.
Zu den Hauptergebnissen von WP1.5 gehören:

- Text-zu-Bewegung-Generierung: Ein Modell, das 3D-Positionssequenzen aus Textbeschreibungen generieren kann und den Stand der Technik in der textbasierten Bewegungssynthese vorantreibt.
- Erzeugung synthetischer Daten: Erstellung neuer Trainingsbeispiele zur Unterstützung von Aufgaben wie der menschlichen Handlungserkennung (HAR).
- Pipeline-Entwicklung: Zukünftige Pläne zur automatischen Extraktion von 2D-Bewegungen und Annotationen aus Online-Videos und zur Rekonstruktion von 3D-Bewegungen, um semantisch annotierte Trainingsdaten zu erzeugen.

ASR & MLT: Text zu Bewegung, Video zu Bewegung Datengenerierung

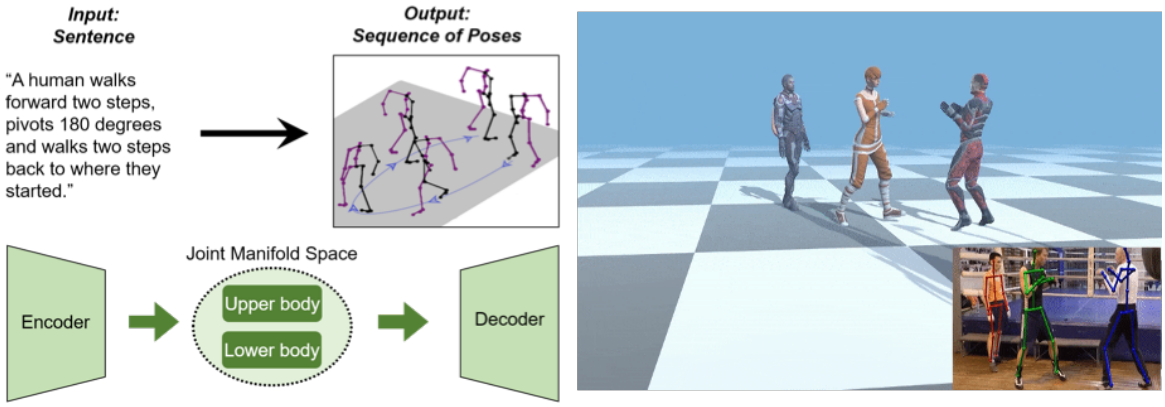

Abbildung 5: Links: Überblick über unsere Methode zur Erzeugung von Bewegungen aus komplexen natürlichsprachlichen Sätzen; rechts: eine synthetisierte Kampfsportszene.

Im Rahmen von WP1.5 hat sich das XAINES-Projekt erfolgreich der Herausforderung gestellt, synthetische Daten zu erstellen, um die Lücke zwischen natürlichsprachlichen Beschreibungen und Bewegungsdarstellungen zu schließen. Das Hauptziel bestand darin, die Fähigkeit von KI-Systemen zu verbessern, menschliche Bewegungen aus komplexen Eingaben zu verstehen und zu generieren, insbesondere in Ermangelung umfangreicher annotierter Trainingsdaten, die typischerweise für tiefe neuronale Netze bei Aufgaben wie der Erkennung menschlicher Handlungen erforderlich sind.
Der Schwerpunkt von WP1.5 lag auf der Entwicklung eines lernbasierten Modells, das 3D-animierte Posenfolgen aus natürlichsprachlichen Beschreibungen synthetisieren kann. Dazu gehörte nicht nur die Generierung einzelner, einfacher Handlungen aus kurzen Sätzen, sondern auch der Umgang mit komplexeren Sätzen, die mehrere Handlungen beschreiben, entweder nacheinander oder gleichzeitig. Um dies zu erreichen, wurde ein hierarchisches sequenzielles Zweistrommodell verwendet, das feinere Details in den Bewegungen auf Gelenkebene erfassen konnte und so sicherstellte, dass sowohl die Bewegungen des Oberkörpers als auch die des Unterkörpers korrekt dargestellt wurden. Dieser Ansatz ermöglichte die Generierung hochdetaillierter 3D-Posen, die eng an den semantischen Inhalt des Eingabetextes angelehnt waren.
Eine der entscheidenden Innovationen in diesem Arbeitspaket war die Einführung einer *Pipeline zur Erzeugung von Text in Bewegung*. Diese Pipeline ermöglichte die Umwandlung von natürlicher Sprache in 3D-Bewegung, was die Synthese von Trainingsdaten für Modelle zur Erkennung menschlicher Handlungen ermöglichte. Diese synthetischen Daten trugen entscheidend dazu bei, den Mangel an kommentierten Daten aus der realen Welt zu kompensieren, und stellten eine wesentliche Ressource zur Verbesserung der Genauigkeit und Robustheit von KI-Systemen beim Verstehen menschlicher Bewegungen dar.
Für die natürlichsprachliche Eingabe wurden kontextualisierte BERT-Einbettungen verwendet, um das Verständnis des Systems für komplexe Textbeschreibungen deutlich zu verbessern. Handverlesene Einbettungen von Wortmerkmalen aus BERT ermöglichten ein detaillierteres Verständnis von Satzstruktur und Semantik, was für die genaue Zuordnung von Sprache zu menschlichen Bewegungen entscheidend war. Das BERT-Modell half zum Beispiel dabei, subtile sprachliche Elemente wie Adverbien zu erfassen, die Richtung, Häufigkeit und Intensität von Bewegungen beschreiben. Diese Verbesserungen ermöglichten es dem System, eine breite Palette von Eingaben in natürlicher Sprache zu verarbeiten und entsprechende 3D-Positionssequenzen genauer zu erzeugen.
Das Modell wurde mit dem KIT Motion-Language Dataset evaluiert, das aus 3D-Positionsdaten besteht, die mit von Menschen kommentierten natürlichen Sprachbeschreibungen gepaart sind. Die Ergebnisse zeigten, dass das in WP1.5 entwickelte Sprache-zu-Bewegung-Modell bestehende Text-

zu-Bewegung-Modelle mit einer Marge von 50% in objektiven Bewertungen übertraf. Das Modell war in der Lage, genaue 3D-Bewegungssequenzen aus komplexen Textbeschreibungen zu generieren, was einen bedeutenden Fortschritt in der textbasierten Bewegungssynthese darstellt.
Einer der Hauptvorteile dieses Modells war seine Fähigkeit, neue Beispiele für das Training anderer KI-Modelle zu simulieren, z. B. für die menschliche Handlungserkennung (HAR). Durch die Generierung synthetischer Trainingsdaten aus Text half das Modell, das Problem der begrenzten kommentierten Datensätze zu lösen. Trotz dieses Fortschritts war das Modell jedoch immer noch domänenspezifisch, da es durch den Inhalt der Trainingsdaten eingeschränkt war. Der KIT Motion-Language-Datensatz enthält vor allem Beispiele für die Fortbewegung, was bedeutete, dass sich das Modell nur schwer auf andere Bereiche wie Kochen oder Tanzen verallgemeinern ließ.
Um diese Einschränkung zu beheben, untersuchte WP1.5 die Verwendung von Online-Videoressourcen, insbesondere von Lernvideos mit gut erklärten Untertiteln, die auf Plattformen wie YouTube verfügbar sind. Diese Videos bieten eine Fülle von semantisch annotierten Bewegungen, die für das KI-Training wiederverwendet werden können. Als Teil der Lösung konstruierte das Team eine Pipeline zur automatischen Extraktion von 2D-Bewegungen und Kommentaren aus Videodaten und zur anschließenden Rekonstruktion der entsprechenden 3D-Bewegungen. Es wurden zwei prominente Ansätze zur 3D-Bewegungssynthese aus Videos bewertet:
Ein mehrstufiger Ansatz: Bei dieser Methode wurden 2D-Posen aus Videos mit ausgereiften Techniken wie OpenPose geschätzt und dann die 3D-Bewegungssequenzen aus den 2D-Posen rekonstruiert. Ein großer Nachteil dieser Methode war jedoch, dass sie nicht in der Lage war, vernetzte Animationen zu erzeugen, die für eine realistische Bewegungsdarstellung in einigen Anwendungen entscheidend sind.
Ein einstufiges End-to-End-Framework: Bei diesem Ansatz werden 3D-Netzanimationen direkt aus 2D-Bildern erzeugt. Hierfür wurde ein monokulares, einstufiges Modell namens ROMP (Sun et al., 2021) für die Schätzung mehrerer 3D-Personen eingesetzt. Das End-to-End-Framework ermöglichte eine nahtlosere Integration und qualitativ hochwertigere Ergebnisse in Bezug auf die Animation.
Zur Erprobung des Konzepts wurde ein großer Datensatz von Tanz- und Kampfsportvideos gesammelt und vorverarbeitet. So wurden beispielsweise Videos von Capoeira, einer brasilianischen Kampfsportart, die Elemente aus Tanz, Akrobatik und Musik kombiniert, ausgewählt. Capoeira war ein idealer Testfall, da es auf YouTube gut erklärte Lehrvideos gab, die oft Untertitel und Anmerkungen enthielten. Anhand der Untertitel wurden die Videos automatisch in Clips unterteilt, die Bewegungsprimitive darstellen, wie z. B. die grundlegende Capoeira-Bewegung "Ginga". Das System gruppierte semantisch ähnliche Clips zusammen und generierte aus diesen Videosegmenten 3D-Bewegungssequenzen.
Anhand dieser segmentierten Bewegungsprimitive wurde ein statistischer Modellierungsansatz verwendet, um die Verteilung der einzelnen Bewegungstypen zu lernen. Dies ermöglichte die Erzeugung neuer Muster mit Variationen auf der Grundlage semantischer Bezeichnungen wie "Ginga" oder "Martelo". Diese neuen Muster wurden dann verwendet, um Modelle zur Erkennung menschlicher Handlungen (HAR) zu trainieren, wodurch die Fähigkeit des Modells, wertvolle synthetische Trainingsdaten für ein breiteres Spektrum von KI-Aufgaben zu generieren, demonstriert wurde.
Während diese Fortschritte einen bedeutenden Fortschritt darstellen, ist die Arbeit an der Konstruktion einer vollständigen Pipeline für die automatische Konvertierung semantisch annotierter Videos in semantisch orientierte Bewegungssimulationen noch nicht abgeschlossen. Schlüsselkomponenten wie die automatische Videosegmentierung, die 2D-Positionsverfolgung und die Rekonstruktion von 3D-Bewegungen aus der 2D-Verfolgung wurden bereits implementiert und evaluiert. Die Qualität der generierten 3D-Bewegungen muss jedoch noch weiter verbessert werden, um den hohen Anforderungen an fortschrittliche Bewegungsmodelle zu genügen.
Daher sind zusätzliche Nachbearbeitungsschritte erforderlich, um die Qualität der vom System erzeugten 3D-Bewegungen zu verbessern. Der Schwerpunkt der zukünftigen Arbeit wird auf der

Verbesserung der Genauigkeit und des Realismus der 3D-Bewegungsrekonstruktion liegen, um sicherzustellen, dass die gesamte Pipeline - von der Videoeingabe bis zur 3D-Bewegungsausgabe - robust und effizient ist. Auf diese Weise will WP1.5 dieses System zu einem zuverlässigen Werkzeug für die Erzeugung semantisch annotierter 3D-Bewegungen machen, das in einer Vielzahl von Bereichen über die Fortbewegung hinaus anwendbar ist, und letztlich einen flexibleren und skalierbaren Rahmen für das Training von KI-Modellen in bewegungsbezogenen Aufgaben schaffen.

Die Ergebnisse von WP1.5 trugen wesentlich zu den Gesamtzielen des XAINES-Projekts bei. Das entwickelte Modell war nicht nur in der Lage, plausible 3D-Positionssequenzen aus kurzen Sätzen zu generieren, die einzelne Aktionen beschreiben, sondern auch aus komplexeren, kompositorischen Sätzen. Diese Pose-Sequenzen können mehrere Aktionen darstellen, die nacheinander oder gleichzeitig ablaufen, was einen bedeutenden Fortschritt in der textbasierten Bewegungssynthese darstellt. Durch die erfolgreiche Generierung synthetischer Trainingsdaten bot das Arbeitspaket außerdem eine praktische Lösung für eine der größten Herausforderungen in diesem Bereich - begrenzte Trainingsdaten - und unterstützt damit eine breite Palette von Anwendungen, von der Erkennung menschlicher Handlungen bis hin zu virtuellen Umgebungen.
Zusammenfassend lässt sich sagen, dass WP1.5 bedeutende Fortschritte bei der Synthese menschlicher Bewegungen aus Textbeschreibungen erzielte. Die Innovationen bei der Generierung synthetischer Daten, bei der Zuordnung von Text zu Bewegung und bei zukünftigen Pipelines für die Extraktion von Videodaten legten eine solide Grundlage für die weitere Entwicklung von KI-Systemen, die menschliche Handlungen auf der Grundlage natürlichsprachlicher Eingaben verstehen und nachahmen können. Die Arbeit demonstrierte das Potenzial für die Entwicklung robuster Modelle, die die Kluft zwischen Sprache und Bewegung überbrücken können, mit Anwendungen in verschiedenen Bereichen, einschließlich virtueller Umgebungen, Handlungserkennung und Robotik.

Eigene Veröffentlichungen:
Ghosh, A., Cheema, N., Oguz, C., Theobalt, C., & Slusallek, P. (2021). Synthese von kompositorischen Animationen aus textuellen Beschreibungen. In Proceedings of the IEEE/CVF international conference on computer vision (S. 1396-1406).

2.1.2 AP 2: Semantisches Verstehen und Erklärungsgenerierung

Die Ziele von WP2 drehen sich um die Schaffung eines Systems, das komplexe menschliche Handlungen, die in natürlicher Sprache beschrieben sind, verstehen und interpretieren kann und sinnvolle, interaktive Erklärungen für diese Handlungen generiert. WP2 konzentrierte sich auf den Aspekt der Erklärungsgenerierung, der die Übersetzung von Rohdaten (wie Sensordaten oder visuelle Eingaben) in semantisch reichhaltige, von Menschen lesbare Erklärungen umfasst. Die erzeugten Erklärungen sind so konzipiert, dass sie für eine Vielzahl von Nutzern zugänglich sind, darunter Entwickler, Fachleute und Endnutzer.

Zu den wichtigsten Zielen von WP2 gehören:

- Semantisches Verstehen von Handlungen: Ziel war es, Modelle zu entwickeln, die komplexe menschliche Handlungen, Ereignisse oder Ereignisfolgen sowohl aus Echtzeit-Eingaben (z. B. Sensordaten) als auch aus kommentierten Datensätzen interpretieren und klassifizieren können. Diese Modelle müssen aussagekräftige semantische Informationen aus diesen Eingaben extrahieren, um sicherzustellen, dass der Kontext und die Nuancen der Handlungen auf hohem Niveau verstanden werden.
- Konstruktion von Erklärungen aus Daten: WP2 zielte darauf ab, ein System zu entwickeln, das automatisch kohärente Erklärungen auf der Grundlage der extrahierten semantischen Informationen erstellen kann. Diese Erklärungen sollen die Entscheidungsprozesse und

Handlungen von KI-Systemen auf klare und verständliche Weise erläutern. Die Erklärungen müssen an verschiedene Bereiche angepasst werden können, vom autonomen Fahren bis hin zur Industrierobotik, und sie müssen eine genaue Darstellung des Denkens der KI bieten.

- Interaktive und KI-Erklärungen: Die in WP2 erstellten Erklärungen sind nicht nur statischer Text, sondern sollen interaktiv sein. Das bedeutet, dass die Nutzer in der Lage sein sollten, sich mit den Erklärungen auseinanderzusetzen, Fragen zu stellen und Erklärungen zu erhalten. Das System sollte dynamische Echtzeit-Antworten liefern, die das KI-Verhalten weiter erklären und so die Transparenz und das Vertrauen in KI-Entscheidungen verbessern.
- Integration mit anderen Arbeitspaketen: WP2 spielt eine wichtige Rolle bei der Integration der Erklärungsgenerierung mit den Ergebnissen anderer Arbeitspakete. So fließen beispielsweise die in WP1 entwickelten semantischen Verständnismodelle (die Klassifizierungen von Situationen mit synthetisierten Bewegungen verknüpfen) in WP2 ein, das dann die Ausgabe generiert, die diese Bewegungen und die ihnen zugrunde liegenden Ursachen erklärt. WP2 stellt sicher, dass der endgültige Output des XAINES-Projekts nicht nur technisch genau, sondern auch menschenfreundlich und leicht interpretierbar ist.

WP2.1 Darstellung des Verhaltens von Sprachmodellen als textuelle Erklärungen

Das Hauptziel von WP2.1 war es, das Verhalten von Modellen des maschinellen Lernens, insbesondere von Modellen zur Verarbeitung natürlicher Sprache (NLP), verständlicher zu machen, indem ihre Operationen in Form von textuellen Erklärungen dargestellt werden. Dieses Arbeitspaket konzentrierte sich darauf, Erklärungen von Sprachmodellen zu generieren, die nicht nur für technische Experten, sondern auch für Nicht-Experten, wie z.B. Laien und Domänenspezialisten, durch die Verwendung von prägnanten und kohärenten textuellen Dialogen zugänglich sind.
Durch die Gestaltung der Erklärungsprozesse als Dialog zwischen einem Menschen und dem Modell wollte WP2.1 die interaktive Erkundung des Modellverhaltens erleichtern und es den Nutzern ermöglichen, die Entscheidungsfindung zu verstehen. Dieser Ansatz ermöglichte es den Benutzern, Folgefragen zu stellen und ihr Verständnis schrittweise zu verfeinern, bis sie mit der Erklärung zufrieden waren. Die in dieser Aufgabe entwickelten Konversationsagenten erklärten verschiedene Aufgaben und Modelle aus dem Bereich des maschinellen Lernens und des NLP, wie z. B. die Erkennung von Hassreden oder die Beantwortung von Fragen anhand natürlichsprachlicher Abfragen. Die Erklärungen reichten von Merkmalszuweisungen und gegnerischen Beispielen bis hin zu Begründungen, die von großen Sprachmodellen (LLMs) generiert wurden.
Das übergreifende Ziel von WP2.1 war es, die Lücke zwischen komplexen Modellergebnissen und menschenfreundlichen Erklärungen zu schließen, indem dialogbasierte Interaktionen genutzt wurden. Dies ermöglichte es Nutzern mit unterschiedlichem Hintergrund, ein klareres mentales Modell der Funktionsweise und der Vorhersagen von NLP-Systemen zu entwickeln und so das Vertrauen in KI-Systeme und deren Interpretierbarkeit zu stärken.
Wir haben den Titel dieses Arbeitspakets leicht geändert: Während das ursprüngliche Ziel dieses Arbeitspakets darin bestand, Abstraktionen von Diskursstrukturen für die Verbalisierung von Erzählungen und Erklärungen unter dem Dach von XAI-Systemen zu identifizieren und zu erlernen, haben wir uns in unserer forschungsgruppenübergreifenden Arbeit MLT und SLT auf Erklärungen des Verhaltens von NLP-Modellen in einem dialogischen Setup konzentriert, in Anlehnung an die Konzeptualisierung in Mediators und die TalkToModel-Implementierung von Slack et al. (2023). Dabei geht es nicht so sehr um das Lernen von Mustern dieser Narrative, sondern um die Darstellung des Verhaltens von Sprachmodellen in einem dialogischen Umfeld.

SLT, IML: Verbesserung der Erklärbarkeit von NLP-Modellen durch konversationelle Agenten

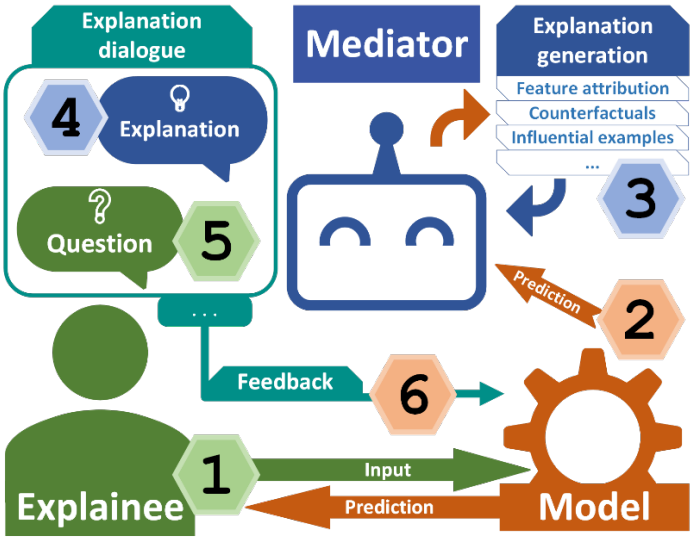

Abbildung 6: *Vereinfachtes Konzept eines Mediators zur Erleichterung von Modellerklärungen: 1) Der Erklärende gibt eine Eingabe, 2) das Modell gibt eine Vorhersage ab, 3) der Vermittler generiert Erklärungskandidaten, 4) der Vermittler initiiert einen Erklärungsdialog, und 5) der Erklärende stellt Folgefragen, während der Vermittler den Dialog und das Verständnis des Benutzers verfolgt.*

Wir trugen zu WP2.1 bei, indem wir die Entwicklung interaktiver Gesprächsagenten zur Erklärung des Verhaltens von Modellen zur Verarbeitung natürlicher Sprache (NLP) vorantrieben. Diese Forschung war ausschlaggebend für die Untersuchung, wie die komplexen Ergebnisse von maschinellen Lernmodellen durch textuelle Erklärungen, die durch interaktive Dialogsysteme präsentiert werden, erklärt werden können.

Ein Hauptziel von WP2.1 war es, das Modellverhalten in einer Form darzustellen, die nicht nur technisch korrekt, sondern auch für Nicht-Experten verständlich ist. Dies wurde erreicht, indem Erklärungsprozesse als für Menschen geeigneter Dialog gestaltet wurden, in welchem das KI-System seine Entscheidungen als Antwort auf Benutzeranfragen erklären konnte. Die im Rahmen dieser Forschung entwickelten Konversationsagenten ermöglichten es den Benutzern, auf intuitive Weise mit dem Modell zu interagieren und Folgefragen zu stellen, um ihr Verständnis für das Verhalten der KI schrittweise zu verbessern. Diese Erklärungen konnten eine Vielzahl komplexer Themen abdecken, wie z. B. die Merkmalszuordnung (Feature Attribution), bei der das System die für eine bestimmte Vorhersage wichtigsten Merkmale identifiziert, oder Gegenbeispiele (Counterfactuals), die zeigen, wie Änderungen der Eingaben die Ergebnisse des Modells verändern können.

Einer der wichtigsten Beiträge zu WP2.1 war die Demonstration, wie interaktive Dialoge die Erklärbarkeit von NLP-Modellen ansprechender und verständlicher machen können, insbesondere für Benutzer ohne technisches Fachwissen über maschinelles Lernen. Der auf Dialoge basierende Ansatz, bei dem die Erklärungen schrittweise gegeben und an das Feedback der Benutzer angepasst wurden, spiegelte die natürliche Art und Weise der menschlichen Kommunikation wider und ermöglichte eine tiefer gehende Erforschung des Modellverhaltens. Durch diese Interaktion konnten die Benutzer nicht nur verstehen, was das Modell tat, sondern auch, warum es bestimmte Entscheidungen traf. Auf diese Weise förderte das System ein größeres Vertrauen in die KI-Technologie, was für ihre Anwendung in kritischen Bereichen wie dem Gesundheitswesen, und autonomen Fahrzeugen von entscheidender Bedeutung ist.

Außerdem führten wir das Konzept der "Mediatoren" (Feldhus et al., 2022) ein, die als Konversationsagenten fungierten und beliebige maschinelle Lernmodelle aufgabenunabhängig erklären konnten. Diese Mediatoren boten einen flexiblen Rahmen für die Erforschung des

Modellverhaltens bei verschiedenen NLP-Aufgaben. Unabhängig davon, ob es sich um Textklassifikation oder komplexere Aufgaben der Spracherzeugung handelte, konnten die Mediatoren erklären, wie bestimmte Modellergebnisse erzeugt wurden, und so das Vertrauen der Benutzer in die Entscheidungsfindung des Systems stärken. Darüber hinaus konnte das System durch die Integration von Benutzerfeedback in den Erklärungsprozess seine Erklärungen verfeinern und anpassen, was die Modelltransparenz und die Benutzerzufriedenheit weiter verbesserte.
Neben der Bereitstellung von Erklärungen zu bestimmten Modellergebnissen wurde untersucht, wie diese Gesprächsagenten den Nutzern helfen können, allgemeinere Muster in den Trainingsdaten und den Lernprozess des KI-Systems zu verstehen. Diese Funktion war besonders nützlich für Entwickler von maschinellem Lernen und Fachexperten, die tiefere Einblicke in das Modell benötigten, um seine Zuverlässigkeit zu bewerten und notwendige Verbesserungen vorzunehmen. Der inkrementelle Dialog ermöglichte es den Nutzern, verschiedene Ebenen der Funktionsweise des Modells anzusprechen, von einzelnen Entscheidungen bis hin zu Verhaltensweisen auf hoher Ebene, und half den Nutzern letztlich, ein umfassenderes mentales Modell der Funktionsweise des Systems zu erstellen (Hartmann et al., 2022).

Eigene Veröffentlichungen:

Hartmann, M., Du, H., Feldhus, N., Kruijff-Korbayová, I., & Sonntag, D. (2022). XAINES: KI mit Narrativen erklären. In KI - Künstliche Intelligenz, 36(3-4), 287-296.

Feldhus, N., Ravichandran, A.M., & Möller, S. (2022). Mediators: Conversational Agents Explaining NLP Model Behavior. In Proceedings of the IJCAI 2022 Workshop on Explainable Artificial Intelligence (XAI), 157-167.

SLT, MLT: Evaluierung dialogbasierter Erklärungen von Sprachmodellen

In der Konzeptualisierung von Mediators erklärt ein Dialogsystem dem Menschen auf interaktive Weise, wie ein Modell funktioniert und wie es zu bestimmten Entscheidungen kommt. Im Gegensatz zu schablonenhaften Ansätzen in früheren Arbeiten haben wir solche Interaktionen für die Benutzer natürlicher gestaltet, indem wir strukturelle Muster gelernt haben, wie Menschen typischerweise Konzepte erklären, was die automatischen Prozesse in InterroLang leitet (Feldhus, Wang et al., 2023).
Wir passten das konversationelle Erklärungsmodell TalkToModel (Slack et al., 2023) an die NLP-Domäne an, fügten neue NLP-spezifische Operationen wie Freitext-Rationalisierung hinzu und illustrieren seine Generalisierbarkeit anhand von drei NLP-Aufgaben (Klassifizierung von Dialoghandlungen, Beantwortung von Fragen, Erkennung von Hassreden). Für die Erkennung von Benutzeranfragen nach Erklärungen evaluierten wir fein abgestimmte und wenig punktuelle Prompting-Modelle und implementierten einen neuartigen Adapter-basierten Ansatz. Anschließend führten wir zwei Nutzerstudien durch, in denen wir (1) die wahrgenommene Korrektheit und Nützlichkeit der Dialoge und (2) die Simulierbarkeit untersuchen, d.h. wie objektiv hilfreich dialogische Erklärungen für Menschen sind, um das vom Modell vorhergesagte Label herauszufinden, wenn es nicht angezeigt wird. Globale Erklärungen wie Datenbeschreibungen und häufige Fehler wurden in der subjektiven Bewertung bevorzugt (Tabelle 3, hier Teil von Abbildung 7). Rationalisierung und Merkmalszuweisung waren hilfreich bei der Erklärung des Modellverhaltens. Die Nutzer konnten das Modellergebnis auf der Grundlage eines Erklärungsdialogs zuverlässiger vorhersagen als mit einzelnen Erklärungen (Tabelle 5, ebenfalls Teil von Abbildung 7).

	Operations	Corr.	Help.	Sat.
Metadata	Show example	52.94	44.44	42.19
	Describe data	89.66	87.27	87.72
	Count data	56.41	44.44	45.83
	True labels	58.82	64.71	72.22
	Model cards	56.25	43.75	45.06
Prediction	Random prediction	57.59	60.71	65.52
	Single/Dataset prediction	53.42	53.52	54.17
	Likelihood	62.86	67.50	63.41
	Performance	72.50	65.79	76.19
	Mistakes	81.25	68.75	77.09
NLU	Similar examples	53.57	45.61	62.50
	Keywords	60.34	54.00	60.00
Expl.	Feature importance	55.88	42.25	50.00
	Global feature importance	50.00	50.00	31.32
	Free-text rationale	62.07	62.50	65.45
Perth.	Counterfactual	40.00	27.03	21.62
	Adversarial example	61.90	40.00	37.50
	Augmentation	62.50	52.17	60.00

Table 3: Task A1 of the user study: Subjective ratings (% positive) on correctness, helpfulness and satisfaction for single turns (responses in isolation), macro-averaged (each user has the same weight, regardless of how many ratings they gave). Custom input operations are averaged with their "regular" counterparts.

Datasets	Corr.	Help.	Sat.	Flue.
BoolQ	3.6	3.3	2.5	3.1
OLID	2.9	3.4	3.0	3.1
DailyDialog	3.2	3.5	3.1	2.9

Table 4: Task A2 of the user study: Subjective ratings (Likert scale 1-5 with 1 being worst/disagree and 5 being best/fully agree) on correctness, helpfulness, satisfaction and fluency for entire dialogues.

Explanation types	Sim (all)	Sim ($t = 1$)	Help Ratio	#Turns Avg.
Local feature importance	91.43	93.10	**82.86**	3.85
Sent. feature importance	90.00	94.44	60.00	3.84
Free-text rationale	**94.74**	**100.00**	68.42	**3.70**
Counterfactual	85.00	80.00	25.00	4.14
Adversarial example	84.00	85.71	56.00	4.00
Similar examples	88.46	87.50	61.54	4.00

Table 5: Task B of the user study: Simulatability. Simulation accuracy (in %), simulation accuracy for explanations deemed helpful (in %), helpfulness ratio (in %), average number of turns needed to make a decision.

Abbildung 7: *Ergebnisse der Nutzerstudie in Feldhus et al. (2023), die subjektive Bewertungen (Korrektheit, Nützlichkeit, Zufriedenheit) sowohl auf der Rundebene (Tabelle 3) als auch auf der Dialogebene (Tabelle 4) sowie die Simulationsgenauigkeit für jede der Erklärungsarten (oder Operationen) zeigen.*

Eigene Veröffentlichungen:
Feldhus, N., Wang, Q., Anikina, T., Chopra, S., Oguz, C., & Möller, S. (2023). InterroLang: Exploring NLP Models and Datasets through Dialogue-based Explanations. In Feststellungen der Gesellschaft für Computerlinguistik: EMNLP 2023, 5399-5421.

SLT, MLT: Auf dem Weg zu selbsterklärenden großen Sprachmodellen
Da diese anfänglichen Lösungen für dialogbasierte Erklärungen viele Abhängigkeiten erfordern und nicht leicht auf Aufgaben übertragbar sind, für die sie nicht konzipiert wurden, stellten wir mit LLMCheckup (Wang et al., 2024) ein leicht zugängliches Werkzeug vor, das es den Benutzern ermöglicht, mit jedem modernen großen Sprachmodell (LLM) über sein Verhalten zu sprechen. Wir versetzen LLMs in die Lage, alle Erklärungen selbst zu generieren und die Absichtserkennung ohne Fine-tuning zu übernehmen, indem wir sie mit einem breiten Spektrum von XAI-Tools verbinden, z. B. Feature-Attributionen, einbettungsbasierte semantische Ähnlichkeit und Prompting-Strategien für Gegenbeispielgenerierung (Counterfactual Generation) und Begründungsgenerierung (Rationale Generation). LLM-(Selbst-)Erklärungen werden in Form eines interaktiven Dialogs präsentiert, der Folgefragen unterstützt und Vorschläge generiert. LLMCheckup bietet Tutorials für die im System verfügbaren Operationen, die sich an Personen mit unterschiedlichem Kenntnisstand in XAI richten, und unterstützt mehrere Eingabemodalitäten. Wir stellten eine neue Parsing-Strategie vor, das sogenannte Multi-Prompt-Parsing, das die Parsing-Genauigkeit von LLMs erheblich verbessert. Dies haben wir mit Datensätzen überprüft aus den Bereichen der Faktenüberprüfung und der Beantwortung von Fragen mit gesundem Menschenverstand (Commonsense Reasoning) vor.

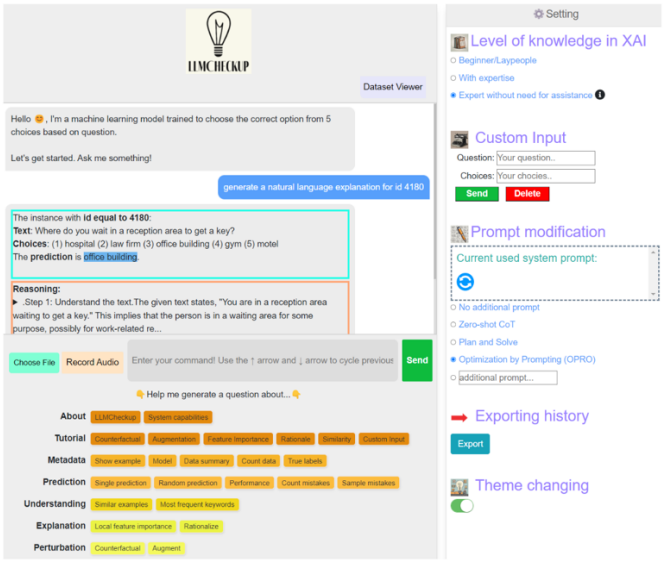

Abbildung 8: *Schnittstelle von LLMCheckup zum Chatten mit jedem LLM über sein Verhalten. Es zeigt eine Willkommensnachricht, eine Beispielanfrage für eine Freitextbegründung und Beispielgeneratorschaltflächen. Auf der rechten Seite kann der Benutzer zusätzliche Einstellungen zu seinem XAI-Kenntnisstand sowie zu den Eingabeaufforderungen und -modifikationen vornehmen.*

Eigene Veröffentlichungen:

Wang, Q., Anikina, T., Feldhus, N., van Genabith, J., Hennig, L., & Möller, S. (2024). LLMCheckup: Conversational Examination of Large Language Models via Interpretability Tools. In Proceedings of the Third Workshop on Bridging Human-Computer Interaction and Natural Language Processing, 89-104.

WP2.2 Generierung von numerischen Interpretationen von NLP-Modellen und Übersetzung in textuelle Erklärungen (SLT, MLT)

In WP2.2 untersuchten wir den am weitesten verbreiteten Ansatz zur Interpretation von Modellen des maschinellen Lernens und des NLP, die Feature-Attribution (Merkmalszuweisung), die Wichtigkeitswerte für jedes der Input-Features (in NLP: Token oder Teilwörter) liefert, die dem Modell für eine Vorhersage zur Verfügung gestellt werden. Sie werden in der Regel als Heatmaps über dem gesamten Eingabetext visualisiert, wobei eine Farbe (in der Regel rot) positive Wichtigkeitswerte in einem Gradienten zu einer anderen Farbe (in der Regel blau) anzeigt, die negative Werte anzeigt. Sie werden in der Regel als eine zusätzliche Schicht über dem betreffenden Text dargestellt.

Da die Menge an visuellen Informationen für den Benutzer überwältigend und insofern kaum interpretierbar sein kann, entwickeln wir neue Ansätze, um die Informationen, die durch die Ausgabe einer Merkmalszuordnungsmethode vermittelt werden, zu vereinfachen und zusammenzufassen.

Diese Bemühungen führten gleichzeitig zu großen Anstrengungen, viele solcher Heatmaps aus einer Vielzahl von Erklärungsmethoden, Sprachmodellen und aufgabenbezogenen Datensätzen zu sammeln, sowie zu Open-Source-Python-Bibliotheken, um sie zu berechnen und herunterzuladen, was den Zugang zu Erklärungen zur Merkmalszuordnung demokratisierte.

Schließlich haben wir mithilfe der Aktivierungsmaximierung auch einzelne Neuronen eines Sprachmodells untersucht, um zu interpretieren, welche Textinformationen sie tendenziell kodieren.

SLT: Demokratisierung der Erklärbarkeit von NLP-Modellen durch zentralen Zugriff auf Feature Attribution Maps

Das Team trug zu WP2.2 bei, indem es eine robuste Infrastruktur bereitstellte, um Erklärungen von NLP-Modellen auf systematische, skalierbare und zugängliche Weise zu erzeugen und zu analysieren. Diese Forschung unterstützte die Ziele von WP2.2, indem sie eine Datendrehscheibe anbot, die den Zugang zu Erklärungen von NLP-Modellen über Feature-Attributionen demokratisierte und es den Nutzern ermöglichte, auf verständliche Weise mit dem Verhalten der Modelle zu interagieren.

Der wichtigste Beitrag von THERMOSTAT (Feldhus et al., 2021) bestand darin, dass es über 200.000 Erklärungen von Modellentscheidungen für ein breites Spektrum bekannter NLP-Modelle und -Aufgaben liefern konnte. Durch die Aggregation von Feature-Attribution-Methoden wie Integrated Gradients und LIME konnten die Benutzer Erklärungen effizient vergleichen und analysieren, ohne dass umfangreiche Rechenressourcen erforderlich waren. Diese zentralisierte Ressource entsprach den Zielen von WP2.1 und WP2.2, da sie sowohl technischen als auch nicht-technischen Nutzern ermöglichte, durch interaktive Erzählungen Einblicke in das Verhalten komplexer Sprachmodelle zu gewinnen.

Darüber hinaus machte THERMOSTAT die Forschung zur Erklärbarkeit zugänglicher und reproduzierbarer, indem es den Nutzern ermöglichte, mehrere Modelle und Datensätze zu vergleichen und so Transparenz und Zuverlässigkeit bei der Entscheidungsfindung von Sprachmodellen zu gewährleisten. Diese Forschung trug maßgeblich zur Erfüllung des Ziels von WP2.2 bei, transparente, benutzerfreundliche Erklärungen für das Verhalten von NLP-Modellen zu liefern, zu einem tieferen Verständnis von KI-Systemen beizutragen und das Vertrauen in Technologien des maschinellen Lernens zu fördern. Indem THERMOSTAT die Hürden für die Erklärbarkeitsforschung senkte, half es, die Kluft zwischen KI-Entwicklern, Domänenexperten und allgemeinen Nutzern zu überbrücken, was es zu einem bedeutenden Beitrag zu WP2.2 macht.

Eigene Veröffentlichungen:
Feldhus, N., Schwarzenberg, R. & Möller, S. (2021). Thermostat: A Large Collection of NLP Model Explanations and Analysis Tools. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 87-95.

SLT: Verbalisierung von Salienzkarten (Übersetzung von numerischen in textliche Erklärungen)

Wir erreichten das Ziel der Verbalisierung von numerischen Erklärungen mit Saliency Map Verbalization (SMV) in Feldhus, Hennig et al. (2023). Saliency Maps können die Vorhersage eines neuronalen Modells erklären, indem sie wichtige Eingangsmerkmale identifizieren. Sie zeichnen sich zwar dadurch aus, dass sie dem erklärten Modell treu (faithful) sind, doch sind Saliency Maps in ihrer Gesamtheit für Menschen schwer zu interpretieren, insbesondere bei Instanzen mit vielen Input Features, z.B. langen Texten oder ganzen Dokumenten. Im Gegensatz dazu sind natürlichsprachliche Erklärungen (NLEs) flexibel und können auf die Erwartungen des Empfängers abgestimmt werden, sind aber kostspielig in der Erstellung: Rationalisierungsmodelle werden in der Regel auf bestimmte Aufgaben trainiert und erfordern hochwertige und vielfältige Datensätze mit menschlichen Annotationen. Wir kombinierten die Vorteile beider Erklärungsmethoden, indem wir Saliency Maps verbalisieren. Wir formalisierten diese wenig erforschte Aufgabe und schlugen eine neuartige Methodik vor, die zwei zentrale Herausforderungen dieses Ansatzes angeht - was und wie verbalisiert werden soll. Unser Ansatz verwendet effiziente Suchmethoden, die aufgaben- und modellunabhängig sind und kein weiteres Black-Box-Modell erfordern, sowie handgefertigte Vorlagen (Abbildung 9). Wir führen eine menschliche Bewertung von Erklärungsrepräsentationen in zwei NLP-Aufgaben durch: Nachrichten-Themenklassifikation und Stimmungsanalyse. Unsere Ergebnisse deuten darauf hin, dass SMV-Erklärungen für Menschen verständlicher und kognitiv

weniger anspruchsvoll macht als herkömmliche Heatmap-Visualisierungen. Besonders GPT-3.5-generierte Verbalisierungen schnitten bei den Endnutzern sehr gut ab, haben aber auch ein paar sachliche Fehler verursacht.

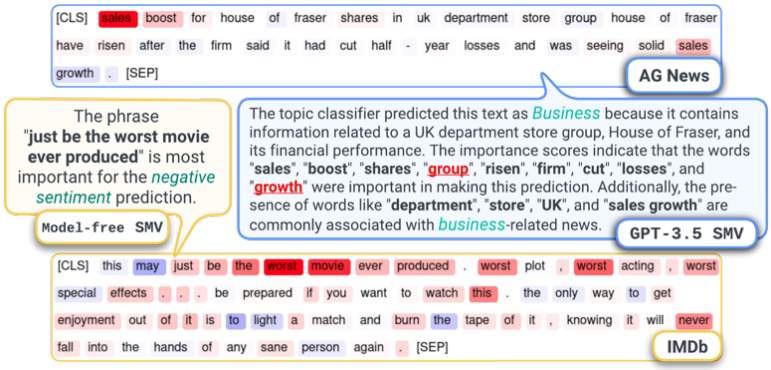

Abbildung 9: Heatmap-Visualisierungen zur Erklärung eines BERT-Modells und ihre jeweiligen Verbalisierungen (modellfreie SMV, GPT-3.5 SMV). Die modellfreie SMV ist schablonenhaft, wird aber von den Nutzern nicht so hoch bewertet, während die GPT-3.5 SMV gut ankommt, aber sachliche Fehler enthält (rot).

Eigene Veröffentlichungen:

Feldhus, N., Hennig, L., Nasert, M.D., Ebert, C., Schwarzenberg, R., & Möller, S. (2023). Saliency Map Verbalization: Comparing Feature Importance Representations from Model-free and Instruction-based Methods. Proceedings of the 1st Workshop on Natural Language Reasoning and Structured Explanations (NLRSE), 30-46.

SLT: Effiziente Erzeugung von Erklärungen für Sprachmodelle

Inmitten einer Diskussion über grüne KI, in der wir die Erklärbarkeit vernachlässigt sahen, untersuchten wir die Möglichkeit, rechenintensive Erklärer effizient zu approximieren. Zu diesem Zweck schlugen wir eine Modellierung der Merkmalszuordnung mit Empirischen Erklärern vor (Schwarzenberg et al., 2021). Empirische Erklärer lernen aus Daten, um die Feature-Attribution-Erklärungen von teuren Erklärern vorherzusagen. Wir trainieren und testen Empirische Erklärer in der Sprachdomäne und stellen fest, dass sie ihre teuren Gegenstücke überraschend gut modellieren, und das zu einem Bruchteil der Kosten. Sie könnten daher den Rechenaufwand neuronaler Erklärungen in Anwendungen, die einen Näherungsfehler tolerieren, erheblich verringern.

Eigene Veröffentlichungen:

Schwarzenberg, R., Feldhus, N., & Möller, S. (2021). Effiziente Erklärungen aus empirischen Erklärern. In Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP, Seiten 240-249, Punta Cana, Dominikanische Republik. Gesellschaft für Computerlinguistik.

SLT: Generierung von Erklärungen zur Merkmalszuweisung für LLMs

Inseq ist ein Toolkit für die Erklärbarkeit von Sequenz-zu-Sequenz-Modellen (Sarti et al., 2023), die für die Forschung auf dem neuesten Stand der Technik sehr wünschenswert sind. Mit Inseq können Praktiker auf einfache Weise Salienzkarten-Erklärungen für generative Modelle, z. B. maschinelle Übersetzungsmodelle, erstellen, um herauszufinden, warum eine bestimmte Übersetzung statt einer anderen erstellt wurde oder wie idiomatische Ausdrücke von solchen Modellen erzeugt werden (Abbildung 10).

Inseq zentralisiert den Zugang zu einer breiten Palette von Methoden zur Feature-Attribution und ermöglicht einen fairen Vergleich verschiedener Techniken für alle Sequenz-zu-Sequenz- und

Decoder-only-Modelle. Dank der intuitiven Benutzeroberfläche können die Benutzer mit nur drei Zeilen Code Interpretierbarkeitsanalysen in Sequenzgenerierungsexperimente integrieren. Nichtsdestotrotz ist Inseq auch sehr flexibel, einschließlich modernster Attributionsmethoden mit eingebauten Post-Processing-Funktionen, die anpassbare Attributionsziele unterstützen und die eingeschränkte Dekodierung beliebiger Sequenzen ermöglichen. In Bezug auf die Benutzerfreundlichkeit vereinfacht Inseq den Zugang zu lokalen und globalen Erklärungen durch die eingebaute Unterstützung einer Befehlszeilenschnittstelle (CLI), optimiertes Batch Processing, das eine datensatzweite Attribution ermöglicht, und verschiedene Methoden zur Visualisierung, Serialisierung und zum Nachladen von Attributionsergebnissen und generierten Sequenzen. Letztlich zielt Inseq darauf ab, Sequenzmodelle zu erstklassigen Bürgern in der Interpretierbarkeitsforschung zu machen und zukünftige Fortschritte in der Interpretierbarkeit für generative Anwendungen voranzutreiben.
In dem dazugehörigen Paper zeigten wir, wie das Tool zur Untersuchung von Artefakten in Datensätzen wie geschlechtsspezifischen Verzerrungen und falschen Korrelationen eingesetzt werden kann, und schlagen eine neue Methode vor, das Contrastive Attribution Tracing (CAT), das visualisiert, wie sich Faktenwissen durch die Schichten von GPT-Modellen ausbreitet.

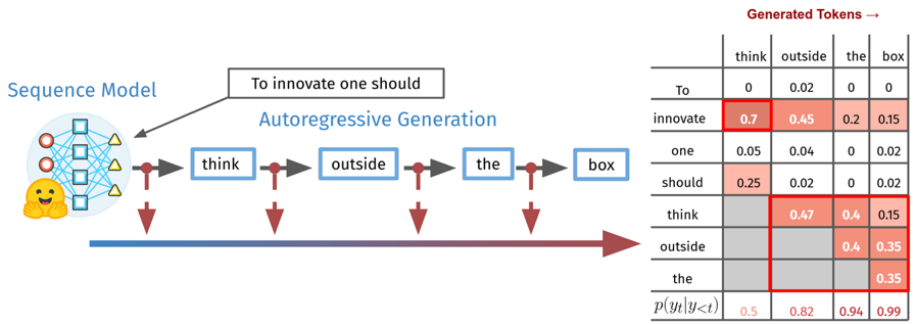

Abbildung 10: Inseq-Nutzung mit einem generativen LM. Attributionswerte und -wahrscheinlichkeiten werden bei jedem Schritt extrahiert, mit Visualisierung auf Token-Ebene. Die Highlights zeigen, wie das Modell mit jedem neuen Wort/Token die Redewendung "

Eigene Veröffentlichungen:
Sarti, G., Feldhus, N., Sickert, L., van der Wal, O., Nissim, M., & Bisazza, A. (2023). Inseq: An Interpretability Toolkit for Sequence Generation Models. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations), 421-435.

MLT: Aktivierungsmaximierung zur Ermittlung von linguistischem Wissen in vortrainierten Sprachmodellen
Um die Vorhersagen von tiefen neuronalen Netzen (DNNs) zu erklären, muss man die Quelle des gezeigten Verhaltens innerhalb des Netzes untersuchen. Im Bereich der Verarbeitung natürlicher Sprache konzentrieren sich viele Arbeiten auf lokale Erklärungen, die auf numerischen Interpretationen wie Feature-Attribution, gegnerischen und einflussreichen Beispielen basieren, wobei das Ziel darin besteht, eine Erklärung für die Entscheidung des Modells in einem bestimmten Fall zu finden. In unserer Arbeit zielten wir darauf ab, die innere Funktionsweise und die Informationen zu verstehen, die in verschiedenen Teilen von vortrainierten Sprachmodellen wie BERT (Devlin et al., 2019) kodiert sind, und zwar auf eine input-agnostische Weise. Wir verwendeten die Aktivierungsmaximierung (AM), um künstliche Eingaben zu synthetisieren, die verschiedene Unterstrukturen im Netzwerk maximal aktivieren. Diese synthetischen, optimalen Eingaben können einen Einblick in die Informationen geben, die in diesem Teil des Netzwerks kodiert sind, und bieten somit eine neue numerische Interpretation des DNN.

Im Bereich der Computer Vision wird AM seit Jahren eingesetzt, um interessante Einblicke in das Innenleben moderner Bildklassifizierer zu erhalten (Olah et al., 2017).
Angesichts des Erfolgs von AM in Computer-Vision-Modellen wendeten wir die gleiche Technik auf den NLP-Bereich an, um zu verstehen, welche linguistischen Informationen in den Neuronen von tiefen NLP-Modellen erfasst werden könnten. Wir untersuchten insbesondere die Nähe der Neuronenbedeutungen zu Wörtern und semantischen Konzepten. Die Ergebnisse lieferten messbare Beweise dafür, dass einzelne Neuronen nicht besonders empfindlich auf klar umrissene linguistische Einheiten reagieren, und belegen, dass Wörter und optimale Eingaben disjunkte Teile des Vektorraums besetzen. Diese Ergebnisse stellen frühere Ansätze in Frage, die einzelne Neuronen in Sprachmodellen über ihre Empfindlichkeit gegenüber Wörtern und Sätzen interpretieren (Bolukbasi et al., 2021; Poerner et al., 2018).
Im Rahmen von XAINES (WP 2.2) haben wir unsere Erkenntnisse in symbolische Darstellungen für verschiedene Teile der untersuchten Sprachmodelle umgesetzt, die als alternative, statische Quelle für neuronale Erklärungen genutzt werden können.

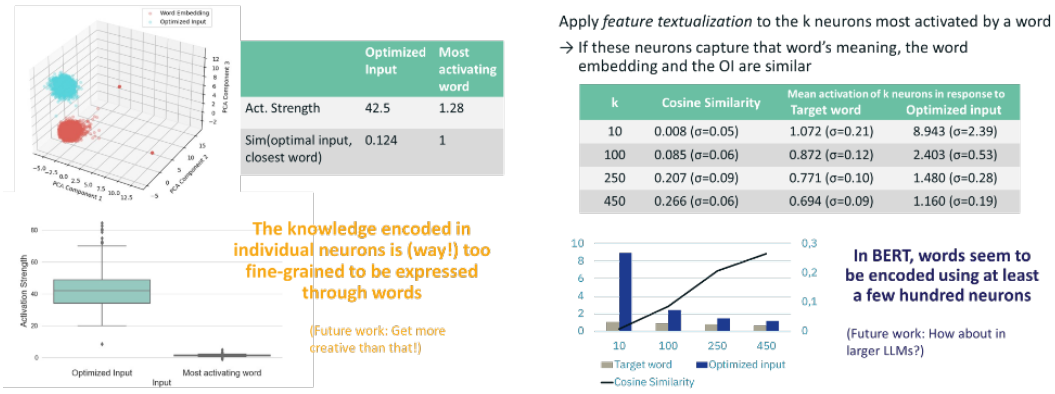

Abbildung 11: *Ergebnisse von Baeumel et al. (2023), die zeigen, dass das Wissen, das einzelne Neuronen kodieren, zu feinkörnig ist, um durch Wörter ausgedrückt zu werden. In BERT scheinen Wörter mit mindestens ein paar hundert Neuronen kodiert zu werden.*

Eigene Veröffentlichungen:
Baeumel, T., Vijayakumar, S., Van Genabith, J., Neumann, G. & Ostermann, S. (2023). Untersuchung der Kodierung von Wörtern in BERT's Neuronen mittels Feature Textualization. Proceedings Of The 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks For NLP, 261-270.

WP2.3 Lernen von Domänennarrativen aus Ereignis- und Aktivitätserkennung

Für die Generierung von Domänenbeschreibungen in WP2.3 ist es wichtig, dass die Aktivitätserkennungssysteme und die ML-basierten Anwendungen sich verschiedenen Nutzertypen erklären können, z. B. Nutzern mit unterschiedlichem technischen Hintergrund. Daher sollten die Formulierungen und der Stil eng an die Erklärungen angelehnt sein, die Fachleute den verschiedenen Benutzergruppen geben würden. Unser Ziel bestand darin, verschiedene Formen von Erklärungsmustern aus relevantem externem Textmaterial zu extrahieren, z.B. aus Artikeln über Entitäten, die vom Aktivitätserkennungssystem erkannt werden. Für die Identifizierung solcher Muster müssen Sprachinformationen wie Erwähnungen von Entitätstypen (NER - Named Entity Recognition) und Beziehungen zwischen ihnen (RE - Relation Extraction) extrahiert werden, um relevantes Material zu identifizieren und eine feinkörnige Analyse von Domänenmustern zu ermöglichen.

MLT: Extraktion verschiedener Formen von tätigkeitsspezifischen Erklärungsmustern aus Texten

Forschungsaktivitäten im Bereich NER und RE konzentrieren sich aktuell auf Transfer Learning durch die Kombination von großen vortrainierten Sprachmodellen (z. B. BERT) und domänenspezifischen Anpassungen. Als Teil unserer Aktivitäten in diesen Bereichen haben wir ein Transfer-Learning-Framework für NER entwickelt und es in verschiedenen domänenübergreifenden Zero-Shot und Few-Shot-Settings getestet, nämlich im allgemeinen und klinischen Bereich (Amin & Neumann, 2021; Amin, Goldstein et al., 2022). Für die Extraktion von Beziehungen zwischen identifizierten Erwähnungen erforschten wir die Fernüberwachung, eine Methode zum automatischen Abgleich von Wissensgraphen (d. h. Wissensdatenbanken mit Tripeln von Entitätspaaren und ihren Beziehungen) mit Texten aus dem biomedizinischen Bereich (Amin, Minervini et al., 2022).

Wir untersuchten auch NER und RE im Zusammenhang mit der Mehr-Label-Klassifikation von ICD-Codes (basierend auf unserer früheren Arbeit, die in Amin et al. (2019) und XAI beschrieben ist), wobei wir zunächst eine Graphenstruktur aus dem Eingabetext erstellten, indem wir alle relevanten eindeutigen Bezeichner von Konzepten extrahierten, die die semantischen Typen der ICD-Codes teilten. Ein solcher Graph wird als strukturelle Zusammenfassung des Eingabetextes in Form der wichtigsten Konzepte und ihrer Beziehungen betrachtet. Hier kommt XAI ins Spiel, denn die grafische Darstellung konnte es uns ermöglichen, die Argumentationsschritte zwischen den identifizierbaren Konzepten und den vorhergesagten ICD-Codes in einer interpretierbaren Weise zu verknüpfen. Durch die Nutzung von GNN in unserem Modell kann es darüber hinaus nach einigen der standardisierten graphenbasierten XAI-Methoden, wie sie in Yuan et al. (2022) zusammengefasst sind, weiter untersucht werden, um die Teilgraphenstruktur mit den Klassifizierungsentscheidungen in Beziehung zu setzen. Auf diese Weise konnten wir die relevantesten Konzepte in den Eingabedaten identifizieren, die zu den Entscheidungen des Klassifikators beitragen, und sie als Ausgangspunkt für die Erstellung von Erzählungen verwenden.

Eigene Veröffentlichungen:

Amin, S., Goldstein, N.P., Wixted, M.K., García-Rudolph, A., Martínez-Costa, C., & Neumann, G. (2022). Few-Shot Cross-lingual Transfer for Coarse-grained De-identification of Code-Mixed Clinical Texts. Workshop über biomedizinische natürliche Sprachverarbeitung.

Amin, S., Minervini, P., Chang, D., Stenetorp, P., & Neumann, G. (2022). MedDistant19: Towards an Accurate Benchmark for Broad-Coverage Biomedical Relation Extraction. Internationale Konferenz über Computerlinguistik.

Amin, S. & Neumann, G. (2021). T2NER: Transformers based Transfer Learning Framework for Named Entity Recognition. Proceedings Of The 16th Conference of The European Chapter of The Association For Computational Linguistics: System Demonstration, 212-220.

Amin, S., Neumann, G., Dunfield, K., Vechkaeva, A., Chapman, K. A. & Wixted, M. K. (2019). MLT-DFKI auf dem CLEF eHealth 2019: Multi-label Classification of ICD-10 Codes with BERT. CLEF (Working Notes).

MLT: Einbettung von Wissensgraphen für die Klassifizierung medizinischer Codes verstehen

Ein gängiger Ansatz für die automatische Zuweisung diagnostischer und verfahrenstechnischer klinischer Codes zu Krankenakten ist die Lösung dieser Aufgabe als Klassifizierungsproblem mit mehreren Labels. Die Schwierigkeiten, die mit dieser Aufgabe verbunden sind, ergeben sich aus den Anforderungen an das Fachwissen, den langen Dokumententexten, dem großen und unausgewogenen Etikettenraum, der die Breite und die Abhängigkeiten zwischen medizinischen Diagnosen und Verfahren widerspiegelt. In anderen Bereichen wurden Fortschritte durch die Einbeziehung von externem Wissen erzielt, das in einem graphisch strukturierten Format kodiert werden kann. Bei der Klassifizierung medizinischer Codes wurde bisher nur wenig Wert auf die graphbasierte Darstellung von Konzepten und relationalen Informationen im Eingabedokument

gelegt. Um diese Lücke teilweise zu schließen, stellten wir klinische Texte als graphenstrukturierte Daten durch den UMLS Metathesaurus dar und identifizieren zwei Forschungsparadigmen: implizite Graphdarstellung durch vortrainierte Wissensgrapheneinbettungen und explizite Kodierung von Dokumentenkonzepten durch neuronale Graphnetze. Unsere Ergebnisse unterstreichen die Vorteile von vortrainierten Wissensgrapheneinbettungen. Im Gegensatz dazu hat der Graphen-Encoder nur begrenzte Auswirkungen auf die Leistung, die weitgehend von den Methoden der Graphenkonstruktion beeinflusst wird (Pokaratsari et al., 2024).

Eigene Veröffentlichungen:
Pokaratsari, N., Amin, S., & Neumann, G. (2024). Understanding the Role of Graph Structures in Medical Codes Classification. (In submission).

2.1.3 AP 3: Bereitstellung und Rendering der Erklärungen sowie Interaktion

Arbeitspaket 3 konzentrierte sich auf die Bereitstellung und Interaktion von KI-generierten Erklärungen, um komplexe KI-Entscheidungen verständlich und für die Nutzer zugänglich zu machen. Das Hauptziel bestand darin, klare und aussagekräftige Erklärungen für KI-Prozesse bereitzustellen, um sicherzustellen, dass die Nutzer, ob Entwickler oder Fachexperten, die Gründe für KI-Entscheidungen nachvollziehen können. Dieses Arbeitspaket betonte die Bedeutung von nutzerzentrierten Ansätzen, die Erklärungen auf die spezifischen Bedürfnisse, Präferenzen und Wissensstände verschiedener Nutzergruppen zuschneiden.
WP3 zielte darauf ab, die Lücke zwischen komplexen maschinellen Lernmodellen und dem menschlichen Verständnis zu schließen, indem interaktive und personalisierte Erklärungsmechanismen geschaffen wurden. Das AP untersuchte mehrere Modalitäten für die Bereitstellung dieser Erklärungen, einschließlich visueller und dialogorientierter Tools. Diese Methoden ermöglichten es den Nutzern, nicht nur Erklärungen zu erhalten, sondern auch mit ihnen auf eine Art und Weise zu interagieren, die ein tieferes Verständnis und Engagement ermöglichte. Die Erklärungen wurden entwickelt, um die Transparenz und das Vertrauen in KI-Systeme zu erhöhen, insbesondere in Bereichen, in denen die Entscheidungsfindung von entscheidender Bedeutung ist, wie z. B. im Gesundheitswesen und beim autonomen Fahren.
Der visuelle Ansatz in WP3 umfasste die Entwicklung von Tools, die KI-Entscheidungen visuell darstellen und die Merkmale und Faktoren hervorheben, die bestimmte Ergebnisse beeinflussen. Diese Methode ermöglichte es den Benutzern, komplexe Ergebnisse des maschinellen Lernens zu interpretieren, indem abstrakte Konzepte durch visuelle Darstellungen greifbarer und verständlicher gemacht wurden. Parallel dazu erforschte WP3 dialogische Interaktionen, bei denen die Benutzer mit dem KI-System in einen Dialog treten konnten, um dessen Entscheidungsprozesse abzufragen und zu untersuchen. Diese Interaktion ermöglichte es den Nutzern, nach Klärungen zu suchen, Folgefragen zu stellen und die Erklärungen an ihre spezifischen Bedürfnisse anzupassen.
Insgesamt hat WP3 die Interpretierbarkeit von KI-Systemen erfolgreich verbessert, indem es benutzerfreundliche Erklärungen sowohl über visuelle als auch über dialogorientierte Schnittstellen bereitstellte. Das Arbeitspaket trug dazu bei, das Vertrauen und die Transparenz in KI zu verbessern, indem es sicherstellte, dass sich die Nutzer auf die Entscheidungen des Systems verlassen können, ohne die zugrunde liegenden Prozesse zu verstehen.

WP 3.1 Interaktion mit visuellen Erklärungen

WP3.1 konzentrierte sich auf die Entwicklung interaktiver visueller Erklärungen für Modelle des maschinellen Lernens mit dem Ziel, komplexe KI-Entscheidungen transparenter und interpretierbar zu machen. Ziel dieser Aufgabe war es, visuelle Werkzeuge zu entwickeln, die es den Nutzern ermöglichen, die den KI-Entscheidungen zugrunde liegenden Prozesse zu sehen und zu verstehen, indem sie die relevanten Merkmale und Faktoren auf benutzerfreundliche und verständliche Weise darstellen.
Der in WP3.1 verfolgte Ansatz umfasste die Verwendung visueller Analysen, bei denen die interne Funktionsweise von Modellen des maschinellen Lernens, wie z. B. neuronalen Netzen, in visuelle Darstellungen übersetzt wurde. Diese Visualisierungen heben Schlüsselelemente hervor, wie z. B. Eingangsmerkmale, Empfindlichkeit gegenüber Daten oder spezifische Entscheidungswege, die das Ergebnis der KI beeinflussen. Diese Methode zielte darauf ab, die "Black Box" vieler KI-Systeme zu entschlüsseln und den Nutzern ein besseres Verständnis dafür zu vermitteln, wie Entscheidungen getroffen werden.
Ein besonderer Schwerpunkt wurde auf die Erstellung personalisierter Erklärungen gelegt, bei denen die visuellen Ausgaben an den Hintergrund und die Bedürfnisse der einzelnen Benutzer angepasst wurden. So konnten beispielsweise Entwickler und Fachexperten mit den visuellen Erklärungen interagieren, um je nach ihrem Fachwissen oder der Komplexität des Entscheidungsprozesses verschiedene Detailstufen zu erkunden. Durch diese Personalisierung wurde sichergestellt, dass die Erklärungen nicht allgemein gehalten, sondern auf die verschiedenen Benutzertypen zugeschnitten waren.
Zusätzlich wurden in WP3.1 interaktive Elemente integriert, die es den Nutzern ermöglichten, sich dynamisch mit den visuellen Erklärungen auseinanderzusetzen. Die Nutzer konnten den Entscheidungsprozess des Systems erkunden, indem sie bestimmte Aspekte abfragten, z. B. welche Merkmale am meisten zu einer Entscheidung beitrugen oder wie verschiedene Eingabevariablen das Ergebnis beeinflussten. Diese interaktive Komponente ermöglichte ein tieferes Verständnis und erlaubte es den Nutzern, die Gründe für die Entscheidungen der KI gründlicher zu untersuchen.

IML, MLT: Erklärbare KI durch multimodale Integration vorantreiben

Zu dieser Aufgabe, die sich mit der *Interaktion mit Erklärungen* befasst, haben wir zunächst eine Zusammenfassung verfasst, in der die Zusammenhänge der verschiedenen Teilbereiche und ihre Integration in das Gesamtprojekt dargestellt werden (Hartmann et al., 2021). Die erweiterte Zusammenfassung unterstreicht die beiden Aspekte, die für die Interaktion mit Erklärungen entscheidend sind (wie in Abbildung 12 zu sehen):

(a) Erstellung und Auswahl von Erklärungsinhalten
(b) Integration von Nutzer-Feedback / Erklärungen der Nutzer.

Die erweiterte Zusammenfassung diente als Kurzfassung der (damals) laufenden Forschungsarbeiten, die in einem ausführlichen KI-Zeitschriftenartikel gipfelten (Hartmann, Du et al., 2022).

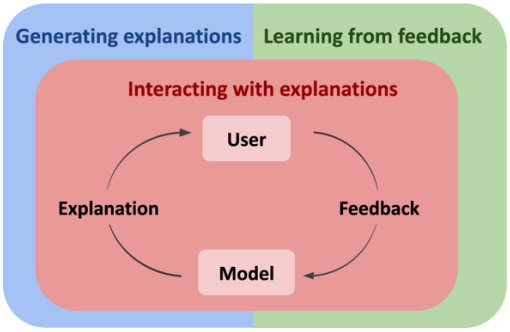

Abbildung 12: Die Interaktion mit Erklärungen (mittlerer Teil) spielt eine zentrale Rolle für die Erklärung von KI-Systemen, was die Generierung von Modellerklärungen (linker Teil) und die Integration von Nutzerfeedback (rechter Teil) erfordert - aus (Hartmann et al., 2021).

Eigene Veröffentlichungen

Hartmann, M., Kruijff-Korbayová, I., & Sonntag, D. (2021). Interaktion mit Erklärungen im XAINES-Projekt. Trustworthy AI in the Wild Workshop 2021.

Hartmann, M., Du, H., Feldhus, N., Kruijff-Korbayová, I., & Sonntag, D. (2022). XAINES: KI mit Narrativen erklären. KI - Künstliche Intelligenz, 36(3-4), 287-296.

IML: Generierung von Erklärungen: Entwicklung einer Pipeline für interaktive Bildbeschriftungen

In XAINES betrachteten wir Bildunterschriften als geeignete Erklärungen für den Inhalt eines Bildes - diese visuellen Beschreibungen können auch bei anderen visuell-sprachlichen Aufgaben als Erklärungen für die Entscheidungen des Modells verwendet werden. Da die Generierung dieser visuellen Erklärungen Teil der Interaktion mit Erklärungen ist, konzentrierten wir uns auf die interaktive und erklärbare Bildbeschriftung. Genauer gesagt haben wir eine Pipeline für ein interaktives Bildbeschriftungssystem vorgeschlagen, bei dem das Benutzerfeedback aus den generierten Beschriftungen in Kombination mit Methoden zur Datenerweiterung verwendet wird, um mehr Trainingsinstanzen zu generieren, die wiederum zum erneuten Trainieren des Modells verwendet werden. Außerdem wurde eine kontinuierliche Lernmethode angepasst, um ein katastrophales Vergessen zu vermeiden. Unsere Arbeit wurde von den folgenden Forschungsfragen geleitet: (a) Wie profitiert ein System, das inkrementell mit (simuliertem) Benutzerfeedback trainiert wird, von der Datenerweiterung? Wie schneidet dieses System in Szenarien mit wenigen Aufnahmen ab? (b) Wie effektiv ist ein Wiedergabemodul für das episodische Gedächtnis, um Wissen aus früheren Trainings zu behalten?

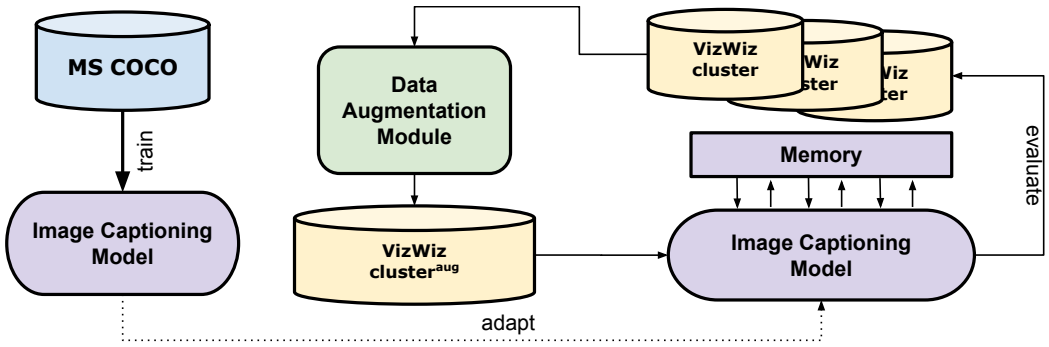

Abbildung 13: Interaktive Bildbeschriftungspipeline für die Domänenanpassung, die Datenerweiterung und die Wiedergabe des episodischen Gedächtnisses integriert, wie in unserem Papier vorgestellt.

Wir haben folgende Beiträge geleistet:

1. Entwicklung einer leichtgewichtigen Pipeline für kontinuierliches Lernen von Bildbeschriftungen, die Datenerweiterung nutzt und in einer interaktiven maschinellen Lernumgebung einsetzbar ist;
2. Anpassung der Methode des kontinuierlichen Lernens, insbesondere Sparse Memory Replay, für den Anwendungsfall der Bildbeschriftung;
3. Durchführung von Experimenten mit einer Kombination verschiedener Datenerweiterungsmethoden für interaktive Bildbeschriftungen, sowohl in Bild- als auch in Textmodalitäten;
4. Untersuchung negativer Ergebnisse und mögliche Erklärungen, warum bestimmte Ansätze nicht erfolgreich waren;
5. Einführung einer Methode, die auf der nominalen Phrasenähnlichkeit zwischen Bildunterschriften unterschiedlicher Bilder basiert, um einen Datensatz in verschiedene Aufgaben zu unterteilen, die für die Evaluierung des aufgabeninkrementellen kontinuierlichen Lernens geeignet sind, wenn nur Bildunterschriften vorliegen.

Für unser simuliertes Nutzerfeedback haben wir einen domänenspezifischen Datensatz verwendet, nämlich VizWiz, der aus Bildern besteht, die von sehbehinderten Menschen aufgenommen wurden. Wir haben diesen Datensatz genau wegen dieser Eigenschaft ausgewählt: Die Qualität der Bilder ist geringer als bei den meisten allgemein verwendeten Bildbeschriftungsdatensätzen und ähnelt der Bildqualität der Nutzerbilder.
Unser Ansatz wurde in (Hartmann, Anagnostopoulou et al., 2022) beschrieben und eine Erweiterung dieses Abstracts (Anagnostopoulou et al., 2022) wurde auf dem zweiten Workshop Bridging Human-Computer Interaction and Natural Language Processing auf der NAACL 2022 vorgestellt. Das vollständige Papier wurde auf dem SustaiNLP Workshop auf der ACL 2023 angenommen.

Eigene Veröffentlichungen:

Anagnostopoulou, A., Hartmann, M., Sonntag, D. (2022). Putting Humans in the Image Captioning Loop. 2. HCI & NLP Workshop der NAACL 2022.

Hartmann, M., Anagnostopoulou, A., Sonntag, D. (2022). Interaktives maschinelles Lernen für Bildbeschriftungen.

Anagnostopoulou, A., Hartmann, M., Sonntag, D. (2023). Towards Adaptable and Interactive Image Captioning with Data Augmentation and Episodic Memory. Proceedings of The Fourth Workshop on Simple and Efficient Natural Language Processing (SustaiNLP), 245-256.

SDS: Sequentielle räumliche Transformatornetzwerke für die Klassifizierung auffälliger Objekte

Eine zentrale Komponente der Interaktion mit visuellen Erklärungen im XAINES-Projekt war die Frage, wie die Objektklassifizierung in komplexen Bildern verbessert werden kann, was die Interpretation visueller Daten in dynamischen, realen Szenarien beinhaltet. Das Projekt schlug ein Modell vor, das einen trainierbaren Fokus-Mechanismus nutzt, der sich an Techniken wie Reinforcement Learning und Spatial Transformer Networks orientiert, um die Fähigkeit des Systems zu verbessern, das auffälligste Objekt in einem Bild zu identifizieren und zu klassifizieren.
Standard-Klassifizierungsarchitekturen wurden in der Regel entwickelt und trainiert, um eine beeindruckende Leistung bei speziellen Bildklassifizierungsdatensätzen zu erzielen, die häufig Bilder mit einem einzigen Objekt in der Mitte enthielten. Ihre Genauigkeit nahm jedoch ab, wenn diese Annahme verletzt wurde, z. B. wenn das Zielobjekt durch Hintergrundrauschen verunreinigt oder nicht zentriert war. In diesem Zusammenhang untersuchte das Projekt die Klassifizierung hervorstechender Objekte, ein realistischeres Szenario, bei dem mehrere Objektinstanzen vorhanden sind und das Ziel darin besteht, das Bild anhand des hervorstechendsten Objekts zu klassifizieren.
Experimente, die mit dem PASCAL VOC-Datensatz (Everingham et al., 2009) durchgeführt wurden, haben gezeigt, dass die Methode die Überschneidung des auffälligen Objekts effektiv erhöht, was zu einer Verbesserung der Gesamtklassifizierungsgenauigkeit um 1,82 Prozentpunkte und bei kleineren Objekten um 3,63 Prozentpunkte führt. Es wurde auch eine Analyse der fehlgeschlagenen Fälle vorgelegt, in der Faktoren wie die Verzerrung des Datensatzes und die Definition der Hervorhebung in Bezug auf die Klassifizierungsergebnisse untersucht wurden.
Dieser Interaktionsansatz ermöglichte es dem System, seinen Entscheidungsprozess visuell zu erläutern, indem es die wichtigsten Teile des Bildes hervorhob und so den Benutzern half zu verstehen, wie die Klassifizierung abgeleitet wurde, selbst unter schwierigen Bedingungen wie Hintergrundrauschen oder nicht zentrierten Objekten.

Eigene Veröffentlichungen:
Dembinsky, D., Azimi, F., Raue, F., Hees, J., Palacio, S., & Dengel, A. (2023). Sequential Spatial Transformer Networks for Salient Object Classification. Proceedings of the 12th International Conference on Pattern Recognition Applications and Methods (ICPRAM), 1, 328-335.

IML: Verbesserung von NLP-Modellen mit menschlichen Erklärungen: Ein Überblick über Methoden und Integration Mechanismen.

Als zusätzlicher Beitrag zu Aufgabe 3.1 umfasste das XAINES-Projekt eine detaillierte Studie über die Verwendung menschlicher Erklärungen zur Verbesserung von Modellen der natürlichen Sprachverarbeitung (NLP). In dem Bericht wurde untersucht, wie die Einbeziehung menschlicher Erklärungen während des Modelltrainings die Dateneffizienz, die Robustheit des Modells und die Anpassungsfähigkeit verbessern kann, vor allem bei Domänenwechseln oder Szenarien mit begrenzten annotierten Daten.
Die Umfrage bot einen umfassenden Überblick über verschiedene Arten von Erklärungen, wie z. B. hervorgehobene Erklärungen, bei denen bestimmte Textabschnitte hervorgehoben werden, um ihre Relevanz zu verdeutlichen, und Freitext-Erklärungen, die Beschreibungen in natürlicher Sprache ermöglichen, die detailliert beschreiben, warum eine bestimmte Vorhersage getroffen werden sollte. Es wurden verschiedene Integrationsmethoden analysiert, darunter Regularisierung, Datenerweiterung und Multitasking-Lernen, und es wurde untersucht, wie diese Ansätze Erklärungsdaten effektiv zur Steuerung von Modellen nutzen können.

Eine der wichtigsten Erkenntnisse war die Fähigkeit menschlicher Erklärungen, Verzerrungen im Lernprozess abzuschwächen. Durch die Konzentration auf die relevantesten Merkmale der Daten konnten Modelle, die mit Erklärungen trainiert wurden, eine Überanpassung an falsche Korrelationen besser vermeiden, was zu genaueren und verallgemeinerbaren Vorhersagen führte. Darüber hinaus hob die Studie den doppelten Nutzen hervor, der sich aus der Verbesserung der Modellleistung und der Nutzerzufriedenheit ergibt, da Erklärungen den Entscheidungsprozess transparenter und verständlicher machen.
Mit diesem Beitrag untermauerte das XAINES-Projekt sein Ziel, erklärbare KI voranzubringen, indem es erforschte, wie erklärungsbasierte Techniken die Kluft zwischen menschlichem Verständnis und maschinellem Lernen überbrücken können, was letztlich zu KI-Systemen führt, die nicht nur genauer, sondern auch besser interpretierbar und benutzerfreundlicher sind.

Eigene Veröffentlichungen:
Hartmann, M., & Sonntag, D. (2022). A Survey on Improving NLP Models with Human Explanations. Proceedings of the First Workshop on Learning with Natural Language Supervision.

WP 3.2 Konversationelle Erklärungen von AI

WP3.2 konzentrierte sich auf die Entwicklung von Interaktionsmechanismen, um KI-Erklärungen durch interaktive Dialoge zu liefern. Ziel dieser Aufgabe war es, den Nutzern die Möglichkeit zu geben, sich mit KI-Systemen auf eine Art und Weise zu unterhalten, die eine natürliche menschliche Konversation imitiert und es ihnen ermöglicht, Fragen zu stellen, Erklärungen anzufordern und die Gründe für KI-Entscheidungen zu erkunden. Dieser Ansatz sollte KI-Systeme transparenter und verständlicher machen, indem den Nutzern dynamische, personalisierte Erklärungen geboten wurden, die sich an ihre Bedürfnisse anpassen konnten.
Das Hauptaugenmerk von WP3.2 lag auf der Entwicklung eines interaktiven, zielgerichteten Dialogsystems, das es den Benutzern ermöglichte, die KI zu befragen und schrittweise Erklärungen zu erhalten. Anstatt einmalige, statische Erklärungen zu liefern, wurde das System so konzipiert, dass es in eine Konversation eintritt, bei der die Erklärungen auf der Grundlage der Fragen des Benutzers verfeinert, in Frage gestellt oder erweitert werden können. Diese Interaktion ahmte einen natürlichen menschlichen Dialog nach, bei dem die Benutzer bestimmte Aspekte des Entscheidungsprozesses der KI vertiefen konnten, um komplexe Ergebnisse Schritt für Schritt zu verstehen.
Zusätzlich zur Unterstützung der Echtzeit-Interaktion wurde das Konversationssystem in WP3.2 so konzipiert, dass es sich an das Profil und die Absicht des Benutzers anpasst. Das bedeutet, dass das System seine Antworten auf den Wissensstand, den Hintergrund und die spezifischen Interessen des Benutzers abstimmen kann. So könnte beispielsweise ein Fachexperte mehr technische Details darüber benötigen, wie eine KI-Entscheidung zustande gekommen ist, während ein nicht technisch versierter Benutzer vielleicht einfachere Erklärungen auf hohem Niveau bevorzugt. Das System war in der Lage, den Detaillierungsgrad und den technischen Charakter seiner Antworten entsprechend anzupassen, um sicherzustellen, dass die Erklärungen für ein breites Spektrum von Nutzern relevant und nützlich waren.
Darüber hinaus wurde in WP3.2 untersucht, wie das KI-System die Absicht des Nutzers während der Interaktion interpretieren und Nutzeranfragen mit den entsprechenden Erklärungen abgleichen kann. Wenn ein Benutzer beispielsweise fragte, warum bestimmte Eingaben zu einem bestimmten Ergebnis führten, konnte das System merkmalsbasierte Erklärungen oder Sensitivitätsanalysen liefern. Bei einer allgemeineren Anfrage könnte das System eine Erklärung auf höherer Ebene

erstellen. Diese Fähigkeit, die Erklärungen an die Absicht des Benutzers anzupassen, machte die Interaktion flexibler und individueller.

IML, SLT: Recherche und Reannotation geeigneter Datensätze.
Auf der Suche nach Datensätzen, die sich für die Modellierung von Dialogen, einschließlich Erklärungen, eignen, fanden wir einige relevante Datensätze, von denen einer kürzlich veröffentlicht wurde und Erklärungen in realen Dialogen zwischen Experten und Schülern mit unterschiedlichem Wissensstand zeigt (Wachsmuth & Alshomary, 2022).
Bei Dialogen, in denen Lehrkräfte Schülern schwierige Konzepte erklären, wird in der Didaktikforschung häufig darüber diskutiert, welche Lehrstrategien zu den besten Lernergebnissen führen. In diesem Beitrag testen wir, ob LLMs solche Erklärungsdialoge zuverlässig annotieren können, so dass sie bei der Unterrichtsplanung und bei Tutorsystemen helfen könnten. Zunächst erstellen wir ein neues Annotationsschema von Lehrhandlungen, das sich an zeitgenössischen Lehrmodellen orientiert, und annotieren einen Datensatz von konversationellen Erklärungen über die Vermittlung von wissenschaftlichem Verständnis in Lehrer-Schüler-Situationen auf fünf Ebenen der Expertise des Erklärenden neu: ReWIRED enthält drei Ebenen von Handlungen (Lehre, Erklärung, Dialog) mit erhöhter Granularität. Anschließend evaluieren wir Sprachmodelle zur Kennzeichnung solcher Handlungen und stellen fest, dass das breite Spektrum und die Struktur der vorgeschlagenen Kennzeichnungen für LLMs wie GPT-3.5/-4 mittels Prompting schwer zu modellieren sind, dass aber ein fein abgestimmter BERT sowohl die Klassifizierung von Handlungen als auch die Kennzeichnung von Bereichen gut durchführen kann. Schließlich operationalisieren wir eine Reihe von Qualitätsmetriken für Unterrichtserklärungen in Form einer Testsuite und stellen fest, dass sie gut zu den fünf Kompetenzstufen passen (Feldhus et al., 2024). Aufbauend auf der im XAINES-Projekt geschaffenen Grundlage beginnen wir nun eine neue Phase der Expertenannotationen als Teil einer über XAINES hinausgehenden Forschungsarbeit. Wir gehen davon aus, dass diese laufenden Arbeiten bis zum nächsten Jahr zur Veröffentlichung einer verbesserten Version des Datensatzes führen werden.

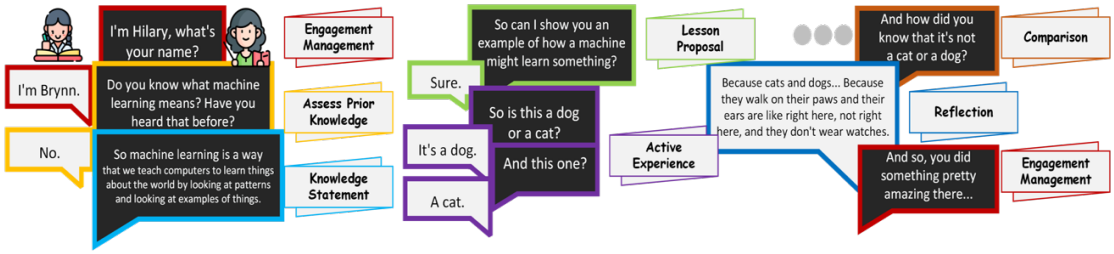

Abbildung 14: Erklärungsdialog eines Experten (Mitte), der einem Kind (links) das maschinelle Lernen erklärt. Die Beschriftungen auf der rechten Seite zeigen die Lehrhandlung an, die mit der/den gleichfarbigen Runde(n) oder Spanne(n) verbunden ist.

Eigene Veröffentlichungen:
Feldhus, N., Anagnostopoulou, A., Wang, Q., Alshomary, M., Wachsmuth, H., Sonntag, D. & Möller, S. (2024). Zur Modellierung und Evaluierung von Unterrichtserklärungen in Lehrer-Schüler-Dialogen. GoodIT '24: Proceedings of the 2024 International Conference on Information Technology for Social Good, 15, 225-230.

ASR: Rendering von Handgesten in einer humanoiden Roboterschnittstelle

Durch die Verbesserung der Natürlichkeit und Effektivität der Mensch-Roboter-Interaktion können die von den Maschinen gelieferten Erklärungen ergänzt werden. Dies kann durch die Simulation der Gesprächsdynamik durch die Wiedergabe von Handgesten in einer humanoiden Roboterschnittstelle erreicht werden. In der Studie wird ein multimodaler interaktiver Erklärungsanwendungsfall durch die Simulation von Konversationsinteraktionen mit einem humanoiden Roboter geschaffen, bei dem der Roboter Handgesten als Schlüsselelement für den Ausdruck von Erklärungen wiedergibt. Wir haben eine Reihe von Experimenten durchgeführt, um unser Bewegungssynthesemodell auf einen humanoiden Roboter anzuwenden, um ikonische Arm- und Kopfgesten als Teil einer aufgabenorientierten dialogischen Interaktion zu erzeugen. In unserer Arbeit (Brady & Du, 2021) haben wir ein multimodales Dialogmanagementsystem und entsprechende Bewegungssteuerungen entwickelt, mit denen Nicht-Experten den Roboter durch Sprache und Sehen "programmieren" können. Mit diesem System werden Videos von Gestendemonstrationen gesammelt. Die Motorpositionen werden aus den Videos extrahiert, um Motortrajektorien zu spezifizieren, wobei Sammlungen von Motortrajektorien verwendet werden, um Robotergesten nach einem Gaußschen Mischungsansatz zu erzeugen. In der abschließenden Diskussion wird erörtert, wie gelernte Repräsentationen für die Gestenerkennung durch den Roboter verwendet werden können und wie das Kernsystem zu einem robusten System heranreifen kann, das die sprachliche Grundlage und semantische Repräsentation berücksichtigt.

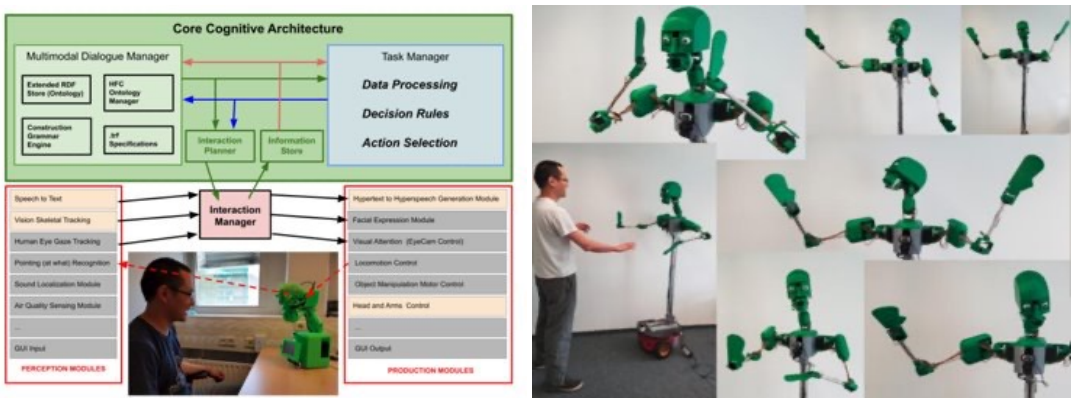

Abbildung 15 links: die Architektur des dialogischen Interaktionssystems. Rechts: Rendering der Handgesten in der Mensch-Roboter-Schnittstelle.

Eigene Veröffentlichungen:

Brady, M. & Du, H. (2021). Teaching Arm and Head Gestures to a Humanoid Robot through Interactive Demonstration and Spoken Instruction. Proceedings of the 1st Workshop on Multimodal Semantic Representations 2021.

2.1.4 AP 4: Domänendaten und Anwendungsfälle

In Arbeitspaket 4 (WP4) war ursprünglich geplant, groß angelegte Demonstratoren in den Bereichen autonomes Fahren, Bau und medizinische Entscheidungsunterstützung zu entwickeln. Mit dem Fortschreiten des Projekts führten die praktischen Herausforderungen jedoch zu einer strategischen Neuausrichtung. Das Team betonte die Stärke der abteilungsübergreifenden Zusammenarbeit und verlagerte den Schwerpunkt von Einzeldemonstratoren auf einen flexibleren Integrationsansatz. Dies beinhaltete orchestrierte Beiträge zu gemeinsamen Anwendungsfällen und die Entwicklung gemeinsamer Datensätze, um ein umfassendes Testen der Erklärbarkeitssysteme zu gewährleisten.

Diese Verlagerung, die sich auf die Förderung gemeinsamer Anstrengungen und Ergebnisse konzentriert, wurde dem Wissemschaftlichen Beirat mitgeteilt, der die überarbeitete Strategie anerkannte und unterstützte. Wir sind davon überzeugt, dass dieser neue Schwerpunkt nicht nur die Wirkung des Projekts erhöhen, sondern auch eine bessere Ausrichtung auf die Ziele der Integration und der langfristigen Nachhaltigkeit gewährleistet.

Eine wichtige Änderung war der Übergang vom Baubereich zum Pflegebereich. Diese Verschiebung war nicht auf inhärente Herausforderungen im Baubereich selbst zurückzuführen, sondern ergab sich wiederum aus praktischen Erwägungen. Insbesondere beendete die führende Abteilung für den Anwendungsfall Bauwesen, Embedded Intelligence (EI), ihre Aktivitäten in diesem Bereich aufgrund fehlender weiterer Finanzierungsmöglichkeiten. Dies hatte zur Folge, dass sich das etablierte Netzwerk und die breitere Perspektive im Baubereich auflösten, was es schwierig machte, den Fokus aufrechtzuerhalten. Im Gegensatz dazu bot der Bereich der Krankenpflege eine neue Chance, die von bestehenden Netzwerken unterstützt wurde und mit neuen Forschungsprioritäten übereinstimmte. Das Projekt richtete seine Bemühungen auf die KI-gestützte Ausbildung von Krankenschwestern und -pflegern aus, was eine einfachere Datenerfassung und strukturiertere Tests von KI-Erklärungen in realen Szenarien ermöglichte. Diese Umstellung unterstreicht auch die Flexibilität des Projekts bei der Anpassung an sich ändernde Bedingungen, wobei der Schwerpunkt weiterhin auf der Erklärbarkeit lag.
Im Bereich des autonomen Fahrens wurde das ursprüngliche Ziel, einen einzelnen Demonstrator zu entwickeln, dahingehend geändert, dass man sich auf spezifische Anwendungsfälle konzentrierte. Anstatt das gesamte Spektrum des autonomen Fahrens abzudecken, konzentrierte sich das Projekt auf Schlüsselszenarien wie Fußgängererkennung und Navigation in komplexen Umgebungen. Dieser engere Fokus ermöglichte es dem Team, qualitativ hochwertige Datensätze zu generieren und präzise Evaluierungen von KI-Erklärungsmechanismen in kontrollierten Umgebungen durchzuführen, um sicherzustellen, dass die Systeme in kritischen Fahrsituationen transparent und verständlich bleiben.
Im Bereich der medizinischen Entscheidungshilfe erzielte das Projekt mit einem interaktiven Demonstrator für die Augenheilkunde, der speziell auf die Diagnose der diabetischen Retinopathie abzielt, einen großen Erfolg. Dieser Demonstrator stellte Medizinern KI-generierte Läsionskarten auf medizinischen Bildern zur Verfügung, mit denen sie interagieren konnten, indem sie markierte Bereiche hinzufügten oder entfernten. Dies ermöglichte eine Rückmeldung in Echtzeit und ein erneutes Training des KI-Modells, wodurch das System anpassungsfähig wurde und besser auf die Entscheidungen der Experten abgestimmt werden konnte. Die Integration des Feedbacks des medizinischen Fachpersonals war ein wichtiger Beitrag des Projekts, der zeigt, wie erklärbare KI im Gesundheitswesen praktisch angewendet werden kann, um Vertrauen und Transparenz zu verbessern.
Diese Anpassungen wurden in den Sitzungen des Wissenschaftlichen Beirats (SAB) sorgfältig erörtert, in denen die Verlagerung von groß angelegten Demonstrationsprojekten zu praktischeren Anwendungsfällen und zur Datengenerierung geprüft und befürwortet wurde. Der SAB lobte insbesondere die Erstellung hochwertiger Datensätze als Schlüsselergebnis des Projekts und betonte, dass diese Datensätze eine wertvolle Ressource für das Testen und die Verfeinerung der Erklärbarkeitsmethoden darstellen.

WP 4.1: Anwendungsfall Autonomes Fahren

Das Ziel des Arbeitspakets 4.1 im Rahmen des XAINES-Projekts bestand darin, die Sicherheit und Zuverlässigkeit autonomer Fahrsysteme durch die Entwicklung von KI-Modellen zu verbessern, die in der Lage sind, komplexe Verkehrsszenarien zu verstehen und vorherzusagen. Dabei geht es um die Schaffung von Modellen, die nicht nur autonom navigieren, sondern auch klare, erklärbare Einblicke in ihre Entscheidungsprozesse liefern. Ein Hauptaugenmerk liegt dabei auf den Interaktionen zwischen Fahrzeugen und Fußgängern, um eine sichere Navigation auch in schwierigen und unvorhersehbaren Umgebungen zu gewährleisten. Durch die Nutzung multimodaler Daten, einschließlich visueller Eingaben und Sensorinformationen, zielte WP4.1 darauf ab, robuste KI-Systeme zu entwickeln, die ihr Verhalten in realen Fahrsituationen interpretieren, vorhersagen und erklären können, um so ein größeres Vertrauen in autonome Technologien zu schaffen.

ASR: Ein multimodaler Datensatz für komplexe Straßenkreuzungsszenarien beim autonomen Fahren

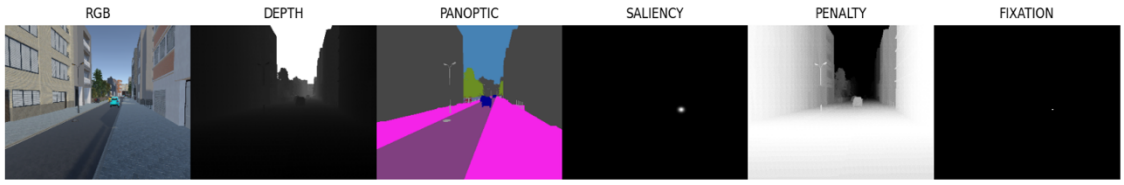

Abbildung 16: Multimodale Daten aus dem XAINES-Datensatz zum autonomen Fahren, mit RGB-Bildern, Tiefeninformationen, panoptischer Segmentierung, Hervorhebungen, Penalty Mapping und Fixationspunkten.

Der Beitrag von ASR zu WP4.1 beinhaltete die Entwicklung eines umfassenden Datensatzes, X-Sal-120, der sich auf Fußgänger-Szenarien beim Überqueren von Straßen konzentriert und über verschiedene Komplexitätsstufen strukturiert ist. Dieser Datensatz befasst sich mit kritischen Anforderungen des autonomen Fahrens, wie z. B. der genauen Wahrnehmung, Tiefenschätzung und Verhaltensvorhersage. Die Szenarien reichen von einfachen Bedingungen mit geringem Verkehrsaufkommen und ausgewiesenen Überquerungszonen bis hin zu komplexeren Konstellationen mit dichtem Verkehr, keinen Überquerungen und NPC-Agenten (nicht spielbare Charaktere), die Regelverletzungen simulieren.
Der Datensatz umfasst 6.142.167 Field-of-View (FoV)-Bilder mit Anmerkungen wie panoptischen Beschriftungen, Tiefenkarten, Fixationskarten, Saliency Maps und Skelettanmerkungen sowohl für Fahrzeuge als auch für Fußgänger, was zu insgesamt 36.853.002 Dateien führt. Diese Datenpunkte wurden in hoher Auflösung (1920x1080) erfasst, wodurch sich der Datensatz für eine Vielzahl von Aufgaben des maschinellen Lernens eignet, darunter Segmentierung, Tiefenschätzung, Vorhersage von Scanpfaden und Vorhersage von visuellen Reizen.

Tiefenkarten spielen eine entscheidende Rolle bei der Bereitstellung von 3D-Informationen über die Umgebung, die es dem autonomen System ermöglichen, die Entfernung von Objekten abzuschätzen, was für Aufgaben wie Hindernisvermeidung und Pfadplanung unerlässlich ist. Durch die Kombination von Tiefeninformationen mit panoptischer Segmentierung können autonome Fahrzeuge eine genauere 3D-Darstellung der Szene erstellen, was eine sicherere Navigation ermöglicht, insbesondere in dichten und dynamischen städtischen Umgebungen.

Die Verwendung von panoptischen Labels integriert sowohl die semantische als auch die instanzielle Segmentierung und bietet ein einheitliches Verständnis der Umgebung, indem sie sowohl die Art der Objekte (wie Fußgänger und Fahrzeuge) identifiziert als auch zwischen einzelnen Instanzen jedes Objekts unterscheidet. Dies ist für autonome Systeme zur Bewältigung dynamischer und komplexer Umgebungen von entscheidender Bedeutung, da es die eindeutige Identifizierung von Objekten ermöglicht, die verfolgt werden müssen, wie z. B. Fußgänger, selbst wenn sie sich unvorhersehbar in einer Szene bewegen. Die panoptische Wahrnehmung kann auch die Verarbeitung rationalisieren, indem sie mehrere visuelle Aufgaben gleichzeitig bewältigt, die Latenzzeit verringert und die Systemleistung verbessert, indem Aufgaben wie Objekterkennung, Klassifizierung und Segmentierung in einen einzigen kohärenten Rahmen integriert werden.
Darüber hinaus helfen zusätzliche Anmerkungen wie Salienz- und Fixationskarten dabei, zu verstehen, welche Bereiche der Szene für die Entscheidungsfindung am wichtigsten sind, und simulieren so, wie sich ein menschlicher Fahrer auf verschiedene Teile der Straße konzentrieren könnte. Dies kann die Interpretierbarkeit der KI-Modelle verbessern, indem hervorgehoben wird, was das System als wichtig "sieht", wodurch das KI-Verhalten transparenter und erklärbarer wird.
Im Zusammenhang mit dem autonomen Fahren werden "Penalty Maps" verwendet, um Bereiche zu identifizieren, in denen bestimmte Aktionen zu einem höheren Risiko oder unerwünschten Ergebnissen führen können. Diese Karten ordnen verschiedenen Regionen der Umgebung Kosten zu und leiten die Wegplanungs- und Entscheidungsfindungssysteme des Fahrzeugs an, sicherere und effizientere Routen zu bevorzugen. Eine Penalty Map kann beispielsweise Bereiche anzeigen, in denen plötzliche Stopps, scharfe Kurven oder Spurwechsel aufgrund der Nähe zu Hindernissen, starker Fußgängeraktivität oder anderen Fahrzeugen gefährlich wären. Durch die Zuweisung höherer Strafwerte für diese Regionen kann das autonome System riskante Manöver vermeiden, was zu einer reibungsloseren und sichereren Navigation führt. Penalty Maps sind besonders in komplexen städtischen Umgebungen nützlich, wo die Straßenbedingungen und das Verkehrsverhalten sehr dynamisch sind, da sie dem Fahrzeug helfen, in Echtzeit Entscheidungen zu treffen, die die Wahrscheinlichkeit von Unfällen minimieren.

Der Datensatz enthält auch personalisierte Informationen über die Teilnehmer (z. B. Alter, Geschlecht, Fahrpraxis), die eine differenziertere Analyse des Verhaltens verschiedener Fußgängertypen ermöglichen und die Vorhersagemodelle weiter verfeinern. Der ursprünglich aus dem REACT-Projekt stammende Datensatz wurde im Rahmen von XAINES um den X-Sal-120-Datensatz erweitert, der bei wichtigen Konferenzen wie der CVPR und der ICCV eingereicht werden soll, was seine Relevanz und sein Potenzial unterstreicht, zur Spitzenforschung im Bereich des autonomen Fahrens beizutragen.

ASR, MLT: Verbesserung der Erklärbarkeit der Vorhersage visueller Aufmerksamkeit mit NLP
Im Rahmen von WP4.1 arbeiteten ASR und MLT zusammen, um eine Komponente zur Verarbeitung natürlicher Sprache (NLP) in die Modelle zur Vorhersage der visuellen Aufmerksamkeit zu integrieren, mit dem Ziel, die Erklärbarkeit des Systems zu verbessern. Diese Zusammenarbeit zielte darauf ab, sich nicht mehr nur auf visuelle Hinweise zu verlassen, sondern klare, beschreibende Erklärungen für die Fokusbereiche des Modells während komplexer Straßenüberquerungsszenarien zu generieren. Anstatt nur anzuzeigen, wohin die Aufmerksamkeit des Modells gerichtet war, könnte es beispielsweise Erklärungen wie "Der Fußgänger schaut aus Sicherheitsgründen beim Überqueren der Straße auf das herannahende Auto" geben.

Dieser Ansatz nutzt multimodale Daten, um informativere und benutzerfreundlichere Interpretationen der visuellen Aufmerksamkeit der KI zu erstellen. Durch die Kombination von panoptischen und Tiefeninformationen kann das System automatisch semantische Bedeutungen generieren und nicht nur explizite Elemente (z. B. ein sich näherndes Fahrzeug), sondern auch implizite Attribute wie ungefähre Geschwindigkeit oder Richtung erfassen. Diese erweiterten Fähigkeiten ermöglichen ein tieferes Verständnis der Vorhersagen des Modells, insbesondere in Szenarien, die präzise und zuverlässige Erklärungen erfordern.
Die Zusammenarbeit trug (und trägt weiterhin) maßgeblich zur Erweiterung des X-Sal-120-Datensatzes bei, in den diese erweiterten visuellen und textuellen Daten integriert wurden. Dieser Datensatz wird zum Zeitpunkt seiner Veröffentlichung halbautomatisch generierte Beschreibungen enhalten und einen reichhaltigere, vielschichtige Ressource darstellen, mit dessen Hilfe robuste visuelle Aufmerksamkeitssysteme trainiert und bewertet werden konnten. Darüber hinaus ermöglicht die Integration von Graph Convolutional Networks (GCNN) und Selective State Space Model (Mamba) eine detailliertere Analyse des Entscheidungsprozesses des Modells. Indem sie untersuchen, welche Teile des Graphen den größten Einfluss auf die Ergebnisse hatten, gewannen die Forscher tiefere Einblicke in die Art und Weise, wie die Aufmerksamkeit verteilt wurde und welche Faktoren diese Entscheidungen beeinflussten. Wichtig ist, dass diese Zusammenarbeit zwischen ASR und MLT über den Rahmen von XAINES hinaus weiter besteht und auf die Veröffentlichung des Datensatzes hinarbeitet. Gemeinsam wird die Integration von visuellen und textuellen Erklärungen weiter verfeinert, um sicherzustellen, dass die Systeme nicht nur genau, sondern auch transparent sind, was den Weg für breitere Anwendungen in autonomen Systemen und anderen verwandten Bereichen ebnet.

ASR: Datensatz zur visuellen Aufmerksamkeit in Verbindung mit Fahrweisen
Im Rahmen von WP4.1 arbeitete ASR mit der Gruppe Augmented Vision (AV) zusammen, um einen Datensatz zu entwickeln, der Aufschluss darüber geben soll, wie sich verschiedene Navigationshilfen auf die kognitive Arbeitsbelastung und das räumliche Wissen beim autonomen Fahren auswirken. Obwohl AV kein offizieller Teil des XAINES-Projekts war, war ihr Fachwissen für diese Zusammenarbeit entscheidend. Das Kernstück dieser Arbeit war die Erstellung eines umfassenden Datensatzes, der das kognitive und navigatorische Verhalten von Fahrern in verschiedenen Szenarien erfasste, einschließlich der Verwendung von Autopilotsystemen, GPS-Navigation und traditionellen gedruckten Karten.
Der Datensatz wurde anhand von Daten aus kontrollierten Experimenten erstellt, die in simulierten Fahrumgebungen durchgeführt wurden. Die Teilnehmer wurden unter verschiedenen Bedingungen mit unterschiedlichen Navigationshilfen konfrontiert, wobei ihr Verhalten, ihre kognitiven Reaktionen und ihre Navigationsleistung genauestens aufgezeichnet wurden. Der Datensatz umfasste nicht nur grundlegende Fahrdaten, sondern auch physiologische Marker für die kognitive Belastung, wie z. B. die elektrodermale Aktivität und das selbst angegebene Stressniveau. Dieser multimodale Ansatz ermöglichte einen ganzheitlichen Blick darauf, wie verschiedene Navigationshilfen die Fähigkeit der Nutzer beeinflussen, sich Routen zu merken, Orientierungspunkte zu erkennen und unabhängig zu navigieren.
Ein wichtiger Aspekt des Datensatzes war, dass er sich auf die Erfassung von Langzeiteffekten konzentrierte. Er enthielt Daten aus Versuchen, bei denen die Teilnehmer zuvor gelernte Routen unter verschiedenen Bedingungen navigieren mussten, so dass analysiert werden konnte, wie sich die anfängliche Wahl der Navigationshilfe auf die spätere unabhängige Navigationsleistung auswirkte. So wiesen Fahrer, die gedruckte Karten verwendeten, anfangs eine höhere kognitive

Belastung auf, zeigten aber später ein besseres räumliches Erinnerungsvermögen und weniger Navigationsfehler. Umgekehrt hatten diejenigen, die sich auf Autopilotsysteme oder digitale Navigationshilfen verließen, zu Beginn eine geringere kognitive Beanspruchung, hatten aber mehr Mühe mit unabhängigen Navigationsaufgaben, was mögliche langfristige Nachteile der passiven Navigation aufzeigt.
Der Datensatz umfasste auch Metriken zur visuellen Aufmerksamkeit und Verhaltensdaten, um zu analysieren, wie die Teilnehmer mit Navigationshilfen und der Umgebung interagierten. Diese Kombination ermöglichte tiefere Einblicke in die spezifischen Herausforderungen und Vorteile, die mit jeder Art von Navigationshilfe verbunden sind, und lieferte wertvolle Daten zur Verbesserung des Designs zukünftiger autonomer Systeme.
Das Ergebnis dieser Zusammenarbeit ist ein Datensatz, der nicht nur den Zielen von WP4.1 dient, indem er das Verständnis dafür verbessert, wie sich Navigationshilfen auf das räumliche Wissen und die kognitive Belastung auswirken, sondern auch eine wertvolle Ressource für die zukünftige Forschung darstellt. Die aus diesen Daten gewonnenen Erkenntnisse sind entscheidend für die Entwicklung autonomer Systeme, die ein Gleichgewicht zwischen Bequemlichkeit und kognitiver Beanspruchung herstellen und sicherstellen, dass die Nutzer ein starkes räumliches Bewusstsein behalten, auch wenn sie sich zunehmend auf die automatische Navigation verlassen. Obwohl die Erstellung des Datensatzes parallel zu AV durchgeführt wurde, haben die Ergebnisse wesentlich zu den Zielen von XAINES beigetragen, indem sie sicherere und effektivere autonome Fahrtechnologien unterstützen.

Eigene Veröffentlichungen:

Brishtel, I., Krauß, S., Schmidt, T., Rambach, J. R., Vozniak, I., & Stricker, D. (2022). Klassifizierung von manuellem und autonomen Fahren basierend auf maschinellem Lernen von Augenbewegungsmustern. In 2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC), 700-705.

Brishtel, I., Schmidt, T., Vozniak, I., Rambach, J. R., Mirbach, B., & Stricker, D. (2021). To Drive or to Be Driven? The Impact of Autopilot, Navigation System, and Printed Maps on Driver's Cognitive Workload and Spatial Knowledge. ISPRS International Journal of Geo-Information, 10(10), 668.

WP 4.2: Anwendungsfall Pflegedaten

WP4.2 konzentrierte sich auf die Anwendung erklärungsfähiger KI-Methoden auf den Bereich der Krankenpflegeausbildung. Ursprünglich sah der Antrag vor, Daten aus dem Baubereich zu verwenden, insbesondere aus Projekten wie SmartWerker und ConWearDi. Diese Projekte lieferten jedoch keine ausreichenden oder qualitativ hochwertigen Daten für weitere Studien über die bereits durchgeführten hinaus. Die Datenerfassung war begrenzt und die Qualität war aufgrund von Hardwarebeschränkungen beeinträchtigt, was es schwierig machte, die ursprüngliche Vision für Aufgabe 4.2 zu verfolgen. Darüber hinaus schränkten die Schließung des ENOBA-Spin-off-Unternehmens und die Auflösung einer der kooperierenden Gruppen unter der Leitung von Prof. Köhler die Kapazitäten für die Fortsetzung der Arbeit im Baubereich weiter ein.
Angesichts dieser Herausforderungen beschloss das Projektteam, sich einem vielversprechenderen Bereich zuzuwenden - der Ausbildung von Krankenschwestern. Dieser Wechsel bot zuverlässigere Daten und entsprach den Zielen des Projekts, erklärbare KI in realen Anwendungen zu demonstrieren. Der Gesundheitssektor und insbesondere die Krankenpflege boten ein strukturiertes Umfeld, in dem die Erklärbarkeit von KI einen bedeutenden Einfluss haben könnte, insbesondere in

Schulungsszenarien, die die Überwachung von Patientenpflegeaktivitäten und die Unterstützung von Entscheidungsprozessen beinhalten.
Der Schwerpunkt von WP4.2 lag auf der Generierung von Daten im Zusammenhang mit pflegerischen Aufgaben, wie Patientenüberwachung, Pflegeroutinen und Interaktionen zwischen Pflegepersonal und Patienten. Durch den Einsatz von KI-Systemen zur Analyse und Erklärung dieser Aktivitäten zielte das Projekt darauf ab, die Transparenz von KI-gestützten Entscheidungen im Gesundheitswesen zu verbessern und sicherzustellen, dass Pflegefachkräfte den Empfehlungen des Systems vertrauen und sie verstehen können. Dieser Ansatz unterstützte nicht nur die Entwicklung erklärungsfähiger KI, sondern befasste sich auch mit einem wichtigen gesellschaftlichen Bedürfnis und zeigte, wie KI sowohl die Effizienz als auch die Qualität der Pflege verbessern kann.

EI: Datensatz Krankenpflege
Der Kern des Anwendungsfalls in der Krankenpflege bestand in der Sammlung und Analyse eines speziellen Datensatzes von Schulungen für Krankenpflegeschüler. Dieser Datensatz bestand aus Videoaufzeichnungen aus mehreren Kameraperspektiven (zwischen zwei und fünf Kameras pro Sitzung), die die Aktivitäten der Krankenpflegeschüler beim Üben des ABCDE-Notfallprotokolls verfolgten. Jeder Gruppe von Schülern wurden farbige Markierungen (rot, grün und blau) zugewiesen, um die Anonymität während der Feedback-Sitzungen zu wahren und sicherzustellen, dass die individuelle Leistung analysiert werden kann, ohne die persönliche Privatsphäre zu gefährden.
Die aufgezeichneten Sitzungen wurden durch das Training eines Objekterkennungsmodells verbessert, um wichtige Objekte im Zusammenhang mit dem ABCDE-Protokoll zu erkennen, z. B. medizinische Monitore, Beatmungsbeutel und automatische externe Defibrillatoren (AEDs). Das YOLOv3-Objekterkennungsmodell wurde für die Echtzeitverfolgung von Teilnehmern und Objekten implementiert und lieferte wertvolle Erkenntnisse über den Standort und die Aktionen der Teilnehmer während des Notfalltrainings.

Wir haben erfolgreich einen Klassifikator für Körperhaltungen entwickelt, der die menschliche Pose schätzt, um grundlegende Haltungen wie Stehen, Gehen und Bücken zu erkennen und zu klassifizieren. Während der ursprüngliche Ansatz eine Genauigkeit von etwa 70 % erreichte, traten Einschränkungen auf, wenn sich die Teilnehmer in den Videos gegenseitig behinderten. Um diese Herausforderungen zu bewältigen, ging das Team zu einem videobasierten Ansatz über, bei dem fortschrittliche neuronale Faltungsnetzwerke (CNNs) und gemischte 3D-Faltungsmodelle (MC3 Net) eingesetzt wurden. Dieser Ansatz führte zu signifikanten Verbesserungen, mit einem F1-Ergebnis von 87 % für den MC3-Modus.

Zusätzlich zur Erkennung der Körperhaltung entwickelte WP4.2 ein detailliertes System zur Erkennung semantischer Handlungen, um bestimmte Aktivitäten wie Beatmung, Herzdruckmassage und Gespräche zu erkennen. Diese semantischen Aktionen waren entscheidend für die Bewertung, wie gut die Krankenpflegeschüler das ABCDE-Protokoll in Notfallszenarien befolgten. Der Datensatz wurde auf drei Granularitätsebenen beschriftet: Körperbewegungen, atomare Aktionen (z. B. das Aufheben von Ausrüstung) und semantische Aktionen. Das System wurde mit zwei Hauptmodellen trainiert: einem Conv-LSTM und einem MC3-Modell. Das MC3-Modell schnitt mit einem Macro-F1-Score von 90 % deutlich besser ab als das Conv-LSTM-Modell mit 73 %. Beide Modelle wurden anhand eines Datensatzes evaluiert, der in Trainings-, Validierungs- und Testdatensätze aufgeteilt war, um eine ausgewogene Evaluierung über verschiedene Aktivitäten hinweg zu gewährleisten.

Eine der größten Herausforderungen in diesem Arbeitspaket war der Umgang mit unausgewogenen Klassenverteilungen in den semantischen Aktionen, da bestimmte Aktivitäten (wie Gespräche) überrepräsentiert waren, während andere (wie die Verwendung eines AED) seltener vorkamen. Das Team löste dieses Problem durch eine zufällige Aufteilung der Daten. Es war jedoch eine weitere Verfeinerung erforderlich, um die Genauigkeit für unterrepräsentierte Handlungen zu verbessern, was zur Annahme der Leave Session Out-Methode für eine realistischere und ausgewogenere Bewertung führte.

Im Rahmen des Projekts wurden auch alle Erkennungssysteme erfolgreich in ein einheitliches Meta-Erkennungssystem integriert, das mit einer Ontologie kombiniert wurde, die die Schritte des ABCDE-Ansatzes darstellt. Dieses integrierte System bildete die Grundlage für ein vollständiges interaktives Feedbacksystem, das Krankenpflegeschülerinnen und -schülern in Echtzeit ein erklärendes Feedback zu ihren Leistungen während der Notfalltrainingseinheiten gab. Das Feedbacksystem nutzte die Ontologie und große Sprachmodelle (LLMs), um personalisierte, kontextbezogene Erklärungen für die Handlungen der Schüler zu erstellen.

Als Teil der abschließenden Aktivitäten in dieser Aufgabe wurde das Erkennungssystem so angepasst, dass es atomare Aktionen verarbeiten kann, was noch detailliertere und granularere Feedback-Informationen ermöglicht. Darüber hinaus wurde die Qualität der während der Trainingssitzungen gesammelten Eye-Tracking-Daten ausgewertet. Obwohl diese Daten eine Herausforderung darstellten, trugen sie zu einem tieferen Verständnis der zwischenmenschlichen Interaktionen zwischen den Studierenden bei.

MLT, EI: Kampfsport-Datensatz

Im Rahmen von WP4.2 unternahm MLT umfassende Anstrengungen, um einen Datensatz von 1.116 Videos von YouTube zu sammeln und zu analysieren. Die Videos wurden aufgrund der Anforderung ausgewählt, dass ihre Untertitel die gezeigten Aktivitäten erklären sollten, was die Verwendung von Sprachmodellen zur Extraktion wichtiger beschreibender Begriffe ermöglichte. Die gesammelten Videos umfassten verschiedene Genres, darunter Kampfsportarten (Capoeira), Tänze (Ballett, Irish Dance, Modern Dance und mehr) und Sport (Cardio, Flexibility, Yoga). Um eine Aktivitätserkennung zu ermöglichen, wendete EI zunächst ein Modell zur Schätzung der menschlichen Haltung auf alle 1.116 Videos an und extrahierte 2D-Gelenkpositionen für alle sichtbaren Personen in jedem Bild. Angesichts der Besonderheit von Capoeira mit eindeutigen Namen für Kicks und Verteidigungsbewegungen konzentrierten sich die ersten Bemühungen auf die Capoeira-Videos. Mithilfe von OpenPose (Cao et al., 2017) optimierte das Team ein Objekterkennungsmodell, um alle Teilnehmer in diesen Videos zu identifizieren und zu verfolgen, und integrierte einen Tracking-Algorithmus, um eine konsistente Identifizierung der Personen über alle Frames hinweg zu gewährleisten. Dieser Ansatz ermöglichte es dem System, die Ausgabe des Objekterkennungsmodells mit den Ergebnissen der Schätzung der menschlichen Haltung zu korrelieren, ähnlich wie bei den Methoden, die zuvor auf die Trainingsvideos der Krankenschwestern angewandt wurden. Da 2D-Gelenkkoordinaten und individuelle Personen-IDs verfügbar waren, bestand der nächste Schritt darin, IMU-Daten (Inertial Measurement Unit) zu simulieren, indem die 2D-Gelenkpositionen mit Modellen wie dem von Kwon et al. (2020) vorgeschlagenen IMU-Simulationsrahmen umgewandelt wurden. Diese Konvertierung ermöglichte eine detailliertere Analyse von Bewegung und Aktivität, insbesondere innerhalb des Capoeira-Datensatzes.

Während diese Bemühungen das Potenzial für die Analyse komplexer Aktivitäten in verschiedenen Bereichen aufzeigten, wurde dieser Ansatz im Rahmen des XAINES-Projekts nicht über Capoeira-Videos hinaus erweitert.
Der SAB riet davon ab, in zu viele Bereiche zu diversifizieren, und empfahl stattdessen eine gezielte Investition, um die Wirkung und den Nutzen der aktuellen Datensätze zu maximieren. Aufgrund dieses Feedbacks konzentrierte sich das Projekt auf die Verfeinerung und Validierung der gesammelten Daten, um robuste und zielgerichtete Ergebnisse zu gewährleisten.

WP 4.3 Anwendungsfall Interaktive Medizinische Entscheidungshilfe

Das Hauptziel des Arbeitspakets 4.3 war die Entwicklung interaktiver Systeme, die die Erklärbarkeit in die medizinische Entscheidungshilfe integrieren und die Transparenz und Nutzbarkeit von KI-Tools im Gesundheitswesen verbessern. Dieses Arbeitspaket zielte darauf ab, Lösungen zu entwerfen und zu implementieren, die es Klinikern ermöglichen, mit KI-generierten Ergebnissen zu interagieren und so einen kollaborativen und verständlichen Entscheidungsprozess zu fördern. Der Schwerpunkt lag auf der Entwicklung modularer, anpassungsfähiger Systeme, die in verschiedene medizinische Bereiche integriert werden können, mit besonderem Augenmerk auf Bereichen, in denen die Interpretierbarkeit für den Aufbau von Vertrauen und die Gewährleistung einer effektiven klinischen Nutzung entscheidend ist.

IML-Beitrag: Entwicklung einer Pipeline für interaktive Bildbeschriftungen und Lernen aus Erklärungen

Im Rahmen von WP4.3 entwickelte wir einen Demonstrator speziell für die interaktive medizinische Entscheidungshilfe im Bereich der Augenheilkunde, wobei der Schwerpunkt auf der Diagnose und Erklärung von Fällen diabetischer Retinopathie lag. Der Demonstrator wurde so konzipiert, dass er visuelle Erklärungen für medizinische Bilder liefert und so ein tieferes Verständnis für Ärzte ermöglicht.
Der Benutzer lädt ein Augenbild hoch, das dann von dem KI-Modell analysiert und klassifiziert wird. Zusammen mit der Klassifizierung erzeugt das System Läsionskarten - binäre Masken, die die Bereiche des Bildes hervorheben, in denen Läsionen erkannt wurden. Eine der wichtigsten interaktiven Funktionen dieses Demonstrators besteht darin, dass der Benutzer diese Läsionskarten manuell anpassen kann, indem er Regionen markiert, die seiner Meinung nach hinzugefügt oder entfernt werden sollten. Diese Interaktion ermöglicht es den Ärzten, sofortiges Feedback zu geben, das für die weitere Verarbeitung genutzt werden kann, z. B. für die Nachschulung und Feinabstimmung der Segmentierungsmodelle, wodurch die Genauigkeit des Systems im Laufe der Zeit verbessert wird.
Zusätzlich zum Demonstrator unternahm das IML im Rahmen von WP4.3 mehrere Forschungsinitiativen. Dazu gehörten die Entwicklung einer umfassenden Pipeline für interaktive Bildbeschriftungen sowie Projekte, die sich auf das Lernen aus Erklärungen konzentrierten. Die Bildbeschriftungspipeline wurde entwickelt, um beschreibenden und kontextrelevanten Text für visuelle Daten zu generieren und so die Kommunikation der Erkenntnisse des Modells an Kliniker zu verbessern. Darüber hinaus arbeitete das Team an der Verbesserung der Fähigkeiten zur Beantwortung visueller Fragen und führte eine Umfrage über den Einsatz von Datenerweiterungstechniken zur Verbesserung der Qualität von Texterklärungen durch.

Das XAINES-Projekt hat auch die Grundlagen für künftige Weiterentwicklungen geschaffen. Als Teil der Entwicklung des Demonstrators war ein zweistufiger Ansatz für narrative Erklärungen geplant. Im ersten Schritt würde das System Benutzerfeedback sammeln, indem es die Kliniker auffordert, Eingaben zu (a) der Korrektheit der identifizierten Läsionen (eine binäre Entscheidung) und (b) detaillierten Beschreibungen der Läsionsformen und -merkmale (Textinformationen) zu machen. Die Gestaltung der Benutzeroberfläche zur Erleichterung dieser Rückmeldungen wurde im Rahmen von XAINES in Angriff genommen. In einem zweiten Schritt sollen die gesammelten Rückmeldungen dazu verwendet werden, die zugrunde liegenden Modelle weiter zu verfeinern, um ihre Fähigkeit zur genauen Identifizierung von Läsionen zu verbessern und die Vorschläge für Läsionsformen auf der Grundlage der Erkenntnisse der Benutzer zu erweitern.

Eigene Veröffentlichungen:
Alam, H. M. T., Nguyen, D. M. H., Truong, M. T. N., Nguyen, T. A., Nguyen, B. T., Barz, M., Profitlich, H., Than, N. T.T., Le, N., Xie, P., Sonntag, D. (2024). DRG-Net: Interaktives gemeinsames Lernen von Multiläsions-Segmentierung und Klassifikation für die Einstufung der diabetischen Retinopathie. ArXiv, abs/2212.14615. (noch in Review)

2.2 Wichtige Positionen des zahlenmäßigen Nachweises

Die wichtigsten Positionen waren Personal- und Verwaltungskosten. Weitere Informationen können dem zahlenmäßigen Verwendungsnachweis entnommen werden.

2.3 Notwendigkeit und Angemessenheit der geleisteten Arbeit

Die geleistete Arbeit im XAINES-Projekt war sowohl notwendig als auch angemessen, um die ambitionierten Projektziele zu erreichen. Ein zentrales Ziel des Projekts bestand darin, innovative Methoden zur erklärbaren Künstlichen Intelligenz (XAI) zu entwickeln, die in der Lage sind, komplexe Entscheidungen von KI-Systemen in verschiedenen Anwendungsbereichen transparent und nachvollziehbar zu machen. Die Herausforderungen, die in Bereichen wie autonomem Fahren, Pflege und medizinischer Diagnostik bestehen, erfordern Technologien, die nicht nur präzise arbeiten, sondern auch in der Lage sind, ihre Entscheidungen gegenüber Nutzern und anderen Interessengruppen verständlich zu erklären.
Um diese Ziele zu erreichen, war es notwendig, einen interdisziplinären Ansatz zu wählen, der die Expertise verschiedener DFKI-Forschungsbereiche einbezog. Diese Zusammenarbeit ermöglichte es, Wissen aus unterschiedlichen Technologiebereichen zu bündeln und innovative, modulare Lösungen zu entwickeln, die eine breite Anwendbarkeit haben. Die Angemessenheit der geleisteten Arbeit zeigte sich insbesondere in der zielgerichteten Nutzung der Fördermittel, die in die Entwicklung und Validierung von Prototypen und die Schaffung von Datensätzen investiert wurden.

Das XAINES-Projekt hat spezifische Technologien entwickelt, die es ermöglichen, multimodale Daten (z. B. Sprach-, Bild- und Sensordaten) zu kombinieren und in kohärente, verständliche Erklärungen zu übersetzen. Dies war notwendig, um sicherzustellen, dass die Nutzer komplexe Entscheidungen der KI-Systeme nachvollziehen können, etwa warum ein autonomes Fahrzeug in einer bestimmten Situation eine Entscheidung trifft oder wie ein KI-System zu einer Diagnose

gekommen ist. Die Notwendigkeit, solche Technologien zu entwickeln, wurde durch die steigende Nachfrage nach zuverlässigen, verständlichen KI-Systemen in der Industrie und Forschung unterstrichen.

Die Angemessenheit der geleisteten Arbeit zeigte sich in der effektiven und effizienten Nutzung der bereitgestellten Mittel. Die Fördermittel wurden gezielt verwendet, um Forschung und Entwicklung in spezifischen Bereichen voranzutreiben, die für das Erreichen der Projektziele essenziell waren. Beispielsweise wurden die Ressourcen in die Entwicklung neuer Algorithmen zur multimodalen Datenintegration investiert, sowie in die Validierung dieser Technologien in praxisnahen Anwendungsfällen. Ein weiterer Aspekt, der die Angemessenheit der geleisteten Arbeit verdeutlicht, war die enge Zusammenarbeit zwischen den beteiligten Forschungsbereiche. Durch regelmäßige Koordination und Kommunikation konnten Synergien genutzt und Ressourcen effizient verteilt werden. Das Projektmanagement sorgte dafür, dass alle Arbeitspakete gut integriert waren und auf ein gemeinsames Ziel hinarbeiteten. Dies ermöglichte es, Innovationen schneller und effizienter zu entwickeln und zu implementieren, wodurch das XAINES-Projekt ein hohes Maß an Flexibilität und Anpassungsfähigkeit bewies. Dies wurde auch vom Wissenschaftsbeirat in seiner projektfinalen schriftlichen Rückmeldung zum Projekt positiv hervorgehoben.

2.4 Verwertbarkeit

In diesem Abschnitt wird die Verwertbarkeit der im XAINES-Projekt entwickelten Technologien und Ergebnisse dargestellt. Ein zentrales Ziel des Projekts war es, innovative Lösungen zu schaffen, die nicht nur in der Forschung Anwendung finden, sondern auch einen direkten Mehrwert für verschiedene Branchen bieten. Die Verwertbarkeit umfasst sowohl die wirtschaftlichen Erfolgsaussichten als auch die wissenschaftlichen Weiterentwicklungsmöglichkeiten. In diesem Zusammenhang wird erläutert, wie die Projektergebnisse in zukünftige kommerzielle Produkte integriert werden können und welche potenziellen Marktsegmente angesprochen werden. Gleichzeitig wird aufgezeigt, wie die entwickelten Technologien die Grundlage für weitere Forschungsvorhaben und Innovationen legen, indem sie offene Forschungsfragen adressieren und neue Anwendungsbereiche erschließen.

2.4.1 Wirtschaftliche Erfolgsaussichten

Die wirtschaftlichen Erfolgsaussichten des XAINES-Projekts sind vielversprechend, da die entwickelten Technologien zur erklärbaren Künstlichen Intelligenz (XAI) Lösungen für Herausforderungen bieten, die in verschiedenen Industrien zunehmend an Bedeutung gewinnen. Die Möglichkeit, KI-Systeme transparenter und verständlicher zu gestalten, schafft Vertrauen bei Endnutzern und erhöht die Akzeptanz solcher Systeme in sicherheitskritischen und regulierten Branchen wie der Automobilindustrie, der Medizin und der Pflege. Unternehmen in diesen Sektoren sind bestrebt, KI-Technologien einzusetzen, um Effizienz und Automatisierung zu steigern, stehen aber gleichzeitig vor der Herausforderung, die Entscheidungsprozesse dieser Systeme nachvollziehbar zu machen.
Ein wesentlicher wirtschaftlicher Vorteil der XAINES-Technologien liegt in ihrer Vielseitigkeit und modularen Integration. Durch die Entwicklung von Technologien, die in bestehenden und zukünftigen KI-Systemen implementiert werden können, entsteht ein Marktpotenzial für

Softwarelösungen, die speziell darauf abzielen, KI-Entscheidungen in Echtzeit zu erklären und für den Endnutzer verständlich aufzubereiten. Unternehmen aus der Automobilbranche könnten beispielsweise die Lösungen aus XAINES nutzen, um autonome Fahrsysteme zu entwickeln, die ihre Handlungen gegenüber Fahrern und Insassen erklären können. Dies würde nicht nur das Vertrauen in solche Systeme stärken, sondern auch helfen, regulatorische Anforderungen zu erfüllen, die eine transparente Entscheidungsfindung fordern.
Im Gesundheitswesen besteht ein großes Marktpotenzial für erklärbare KI in der Diagnostik und Therapieplanung. Systeme, die ihre Entscheidungen klar erläutern können, sind besonders wertvoll, um die Zusammenarbeit zwischen medizinischem Fachpersonal und KI-Systemen zu verbessern und Diagnosen präziser und schneller zu machen. Dies reduziert potenzielle Risiken und steigert die Effizienz in der Patientenbetreuung, was wiederum die Akzeptanz solcher Technologien bei Krankenhäusern und Kliniken erhöhen kann.
Zusätzlich bieten die Entwicklungen aus XAINES die Möglichkeit, neue Geschäftsfelder zu erschließen, die über den bisherigen Markt hinausgehen. Die erarbeiteten Technologien zur multimodalen Datenverarbeitung und Erklärungserstellung sind branchenübergreifend einsetzbar, z.B. in der Finanzindustrie zur Analyse und Erklärung komplexer finanzieller Entscheidungen oder in der Industrie zur Überwachung und Optimierung von Produktionsprozessen. Durch die Modularität und Anpassungsfähigkeit der Lösungen können Unternehmen die Technologien leicht in ihre bestehenden Systeme integrieren, was zu Kosteneinsparungen und einer schnelleren Markteinführung führt.
Die wirtschaftlichen Erfolgsaussichten von XAINES werden zudem durch die aktive Zusammenarbeit mit Industriepartnern gestützt, die wertvolles Feedback und reale Anwendungsfälle einbrachten. Dies stellt sicher, dass die entwickelten Lösungen den tatsächlichen Marktanforderungen entsprechen und direkt in kommerzielle Produkte überführt werden können. Die hohe Nachfrage nach erklärbarer KI und die vielseitige Einsetzbarkeit der XAINES-Technologien bieten daher ein großes Potenzial für langfristige wirtschaftliche Erfolge und eine nachhaltige Marktpositionierung.

2.4.2 Wissenschaftliche Erfolgsaussichten

Die wissenschaftlichen Erfolgsaussichten des XAINES-Projekts sind hoch, da die entwickelten Technologien und Methoden zur erklärbaren Künstlichen Intelligenz (XAI) neue Standards in der Forschung setzen und bestehende Herausforderungen im Bereich der KI-Interpretierbarkeit adressieren. Die im Projekt erarbeiteten Lösungen ermöglichen eine tiefere Integration von multimodalen Daten (z.B. Sprach-, Bild- und Sensordaten), was den Weg für innovative Forschungsansätze in verschiedenen wissenschaftlichen Disziplinen ebnet.
Ein wesentlicher Beitrag des XAINES-Projekts liegt in der Entwicklung von Technologien, die es ermöglichen, die Entscheidungsprozesse von KI-Systemen transparent und nachvollziehbar zu machen. Diese Fähigkeit ist besonders wertvoll für die wissenschaftliche Gemeinschaft, da sie die Grundlage für zukünftige Forschungsarbeiten bildet, die sich mit der Optimierung und Verifizierung von KI-Entscheidungsprozessen befassen. In Bereichen wie der kognitiven Informatik, der Robotik und der maschinellen Sprachverarbeitung eröffnen sich durch die XAINES-Technologien neue Forschungsperspektiven, die darauf abzielen, maschinelle Lernmodelle verständlicher und vertrauenswürdiger zu gestalten.
Ein weiterer Aspekt der wissenschaftlichen Erfolgsaussichten ist die interdisziplinäre Natur des Projekts. Durch die Kombination von Methoden aus den Bereichen maschinelles Lernen, Computer Vision, natürliche Sprachverarbeitung und Robotik fördert XAINES die Entstehung neuer

Forschungsfelder und -ansätze, die über die traditionellen Grenzen einzelner Disziplinen hinausgehen. Diese interdisziplinäre Zusammenarbeit hat das Potenzial, neue Forschungsthemen zu erschließen, die sowohl theoretische als auch angewandte Wissenschaften betreffen, und könnte als Grundlage für zukünftige Verbundforschungsprojekte dienen.
Die Technologien aus XAINES bieten zudem eine wertvolle Ressource für die wissenschaftliche Community, insbesondere durch die Bereitstellung von frei zugänglichen Datensätzen und Open-Source-Software. Diese Verfügbarmachung der entwickelten Tools und Modelle fördert den Austausch und die Weiterentwicklung von Forschungsergebnissen und trägt zur Schaffung eines offenen, kollaborativen Forschungsumfelds bei. Indem die wissenschaftliche Gemeinschaft Zugang zu den Ergebnissen des Projekts erhält, wird die Möglichkeit geschaffen, bestehende Methoden zu erweitern und neue Anwendungsgebiete zu erkunden.
Nicht zuletzt stellen die durch XAINES entwickelten Technologien eine solide Basis für die Weiterentwicklung erklärbarer KI dar, die auf künftige Forschungsvorhaben übertragbar ist. Die Erkenntnisse und Methoden können in verschiedenen wissenschaftlichen Projekten genutzt werden, um die Grenzen der aktuellen Technologie zu erweitern und neue, fortschrittlichere Modelle zu entwickeln. Die im Projekt gewonnenen Erkenntnisse zur Integration von Multimodalität und Narrativen bieten dabei eine wertvolle Grundlage, um die nächste Generation von KI-Systemen zu gestalten, die intuitiver, sicherer und benutzerfreundlicher ist.

2.4.3 Wissenschaftlich-wirtschaftliche Anschlussfähigkeit

Die wissenschaftlich-wirtschaftliche Anschlussfähigkeit des XAINES-Projekts ergibt sich aus der strategischen Ausrichtung, sowohl innovative Forschungsergebnisse zu liefern als auch konkrete Anwendungen zu entwickeln, die in der Industrie verwertbar sind. Die im Projekt erarbeiteten Technologien zur erklärbaren Künstlichen Intelligenz wurden so gestaltet, dass sie eine flexible Integration in unterschiedliche Systeme ermöglichen, was sie für zukünftige wissenschaftliche Projekte und industrielle Anwendungen gleichermaßen attraktiv macht.
Ein zentraler Aspekt der Anschlussfähigkeit liegt in der modularen und skalierbaren Struktur der entwickelten Lösungen. Diese ermöglicht es, die Technologien nicht nur in den bereits getesteten Domänen wie dem autonomen Fahren und der Medizin zu nutzen, sondern auch in neuen Bereichen einzusetzen, die bisher nicht adressiert wurden. Dadurch entsteht eine Grundlage, die über die Projektlaufzeit hinaus für weiterführende Forschungsvorhaben und Produktentwicklungen genutzt werden kann, insbesondere in neuen Sektoren, die zunehmend auf KI-Technologien angewiesen sind.
Wissenschaftlich bietet XAINES durch seine interdisziplinäre Methodik einen Ansatz, der leicht in verwandte Forschungsvorhaben integriert werden kann. Die entwickelten Verfahren zur multimodalen Datenverarbeitung und Erklärungsgenerierung eröffnen Perspektiven für weitere Studien, die das Verständnis und die Entwicklung erklärbarer KI-Modelle vorantreiben. Forschungseinrichtungen können die erarbeiteten Konzepte aufgreifen und in neuen wissenschaftlichen Projekten weiterentwickeln, wobei der offene Austausch mit externen Partnern die Grundlage für Kooperationen und gemeinsame Förderanträge legt.
Wirtschaftlich ermöglicht die erfolgreiche Zusammenarbeit mit Industriepartnern, die Technologien direkt in marktreife Produkte zu überführen. Dies beschleunigt nicht nur die Produktentwicklung, sondern trägt auch zur Schaffung neuer Geschäftsmodelle bei, die auf erklärbaren KI-Systemen basieren. Unternehmen können die entwickelten Lösungen nutzen, um neue Marktsegmente zu erschließen oder bestehende Produkte durch zusätzliche Funktionen wie Transparenz und

Erklärbarkeit zu verbessern, was die Akzeptanz und den regulatorischen Anforderungen zugutekommt.

2.5 Bekannt gewordener Fortschritt

Während der Laufzeit des XAINES-Projekts wurde ein bemerkenswerter Fortschritt im Bereich der Künstlichen Intelligenz öffentlich bekannt: der Aufstieg und die breite Akzeptanz großer Sprachmodelle (Large Language Models, LLMs). Diese Modelle, die in der Lage sind, umfangreiche Texte zu generieren und zu verstehen, haben sich zu einer Schlüsseltechnologie entwickelt und das öffentliche Bewusstsein für die Leistungsfähigkeit und die Anwendungsmöglichkeiten von KI stark beeinflusst. Obwohl LLMs bereits vor Projektbeginn existierten und am DFKI erforscht und eingesetzt wurden, hat sich während der XAINES-Laufzeit gezeigt, dass diese Modelle für die breite Öffentlichkeit und viele Industrien zu einem Sinnbild für den Fortschritt in der KI wurden.
Dieser Wandel hat direkte Auswirkungen auf die Entwicklung erklärbarer KI-Systeme, da LLMs aufgrund ihrer Fähigkeit, natürliche Sprache zu verarbeiten, prädestiniert sind, komplexe Erklärungen zu generieren. Es liegt auf der Hand, dass multimodale "Foundation Models", die Sprache, Bilder und andere Datentypen integrieren können, in Zukunft eine wichtige Rolle bei der Erklärung von KI-Entscheidungen spielen werden. Das XAINES-Projekt hat diese Entwicklung proaktiv berücksichtigt und entsprechende Ansätze in die eigene Forschungsagenda aufgenommen. So wurde beispielsweise in WP2.1 die Nutzung von generativen Modellen untersucht, um textuelle Erklärungen zu erzeugen, die sich flexibel an die Bedürfnisse der Benutzer anpassen und somit eine bessere Nachvollziehbarkeit der Entscheidungen ermöglichen

Ein konkretes Beispiel für diese Integration zeigt sich in neuen, über XAINES hinausgehenden Forschungsvorhaben. In einem aktuellen Projektantrag wird verstärkt auf den Einsatz multimodaler Foundation Models gesetzt, um die Erklärbarkeit in sicherheitskritischen Anwendungen weiter zu verbessern. Hierbei wird untersucht, wie große Sprachmodelle zusammen mit visuellen und sensorischen Daten genutzt werden können, um Echtzeit-Erklärungen für autonome Systeme zu generieren, die sowohl menschlich verständlich als auch präzise sind. Diese Modelle sollen in der Lage sein, komplexe Szenarien zu analysieren und Kontextinformationen zu liefern, die über die unmittelbaren Sensordaten hinausgehen. Das Projekt verfolgt das Ziel, diese Erklärungen nicht nur zur Entscheidungsfindung zu nutzen, sondern auch zur Verbesserung der Transparenz und Akzeptanz von KI-Systemen in sicherheitskritischen Domänen wie autonomen Fahrzeugen, kollaborativen Robotern und Zügen.
Die Erkenntnisse und Anpassungen im XAINES-Projekt haben somit nicht nur auf aktuelle technologische Trends reagiert, sondern auch eine Brücke zu zukünftigen Entwicklungen geschlagen. Durch die Integration von LLMs und multimodalen Modellen wurden bereits die Grundlagen geschaffen, um auch künftige Fortschritte in der KI-Entwicklung zu nutzen. Diese Anpassungsfähigkeit ist ein entscheidender Faktor, um sicherzustellen, dass die im Rahmen von XAINES entwickelten Technologien auch über die Projektlaufzeit hinaus relevant und innovativ bleiben.

2.6 Veröffentlichungen

Die im Rahmen des XAINES-Projekts entstandenen wissenschaftlichen Beiträge wurden an verschiedenen Stellen des Berichts unter der Kennzeichnung „Eigene Veröffentlichungen“ aufgeführt, um den Kontext und die jeweilige Relevanz für die beschriebenen Arbeitspakete und Projektergebnisse nachvollziehbar zu machen. Nachfolgend sind diese Veröffentlichungen nun in alphabetischer Reihenfolge als Gesamtübersicht dargestellt.

1. Alam, H. M. T., Nguyen, D. M. H., Truong, M. T. N., Nguyen, T. A., Nguyen, B. T., Barz, M., Profitlich, H., Than, N. T.T., Le, N., Xie, P., & Sonntag, D. (2024). DRG-Net: Interaktives gemeinsames Lernen von Multiläsions-Segmentierung und Klassifikation für die Einstufung der diabetischen Retinopathie. *ArXiv, abs/2212.14615.*
2. Amin, S., Goldstein, N.P., Wixted, M.K., García-Rudolph, A., Martínez-Costa, C., & Neumann, G. (2022). Few-Shot Cross-lingual Transfer for Coarse-grained De-identification of Code-Mixed Clinical Texts. *Workshop über biomedizinische natürliche Sprachverarbeitung.*
3. Amin, S., Minervini, P., Chang, D., Stenetorp, P., & Neumann, G. (2022). MedDistant19: Towards an Accurate Benchmark for Broad-Coverage Biomedical Relation Extraction. *Internationale Konferenz über Computerlinguistik.*
4. Amin, S., & Neumann, G. (2021). T2NER: Transformers based Transfer Learning Framework for Named Entity Recognition. *Proceedings Of The 16th Conference of The European Chapter of The Association For Computational Linguistics: System Demonstration, 212-220.*
5. Amin, S., Neumann, G., Dunfield, K., Vechkaeva, A., Chapman, K. A., & Wixted, M. K. (2019). MLT-DFKI auf dem CLEF eHealth 2019: Multi-label Classification of ICD-10 Codes with BERT. *CLEF (Working Notes).*
6. Anagnostopoulou, A., Hartmann, M., & Sonntag, D. (2022). Putting Humans in the Image Captioning Loop. *2. HCI & NLP Workshop der NAACL 2022.*
7. Anagnostopoulou, A., Hartmann, M., & Sonntag, D. (2023). Towards Adaptable and Interactive Image Captioning with Data Augmentation and Episodic Memory. *Proceedings of The Fourth Workshop on Simple and Efficient Natural Language Processing (SustaiNLP), 245-256.*
8. Baeumel, T., Vijayakumar, S., Van Genabith, J., Neumann, G. & Ostermann, S. (2023). Untersuchung der Kodierung von Wörtern in BERT's Neuronen mittels Feature Textualization. Proceedings Of The 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks For NLP, 261-270.
9. Biswas, R., Barz, M., & Sonntag, D. (2020). Explanatory Interactive Image Captioning using Top-Down and Bottom-Up Features, Beam Search and Re-ranking. *KI - Künstliche Intelligenz, 34.*
10. Biswas, R., Barz, M., Hartmann, M., & Sonntag, D. (2021). Improving German Image Captions Using Machine Translation and Transfer Learning. *International Conference on Statistical Language and Speech Processing, 3-14.*
11. Brady, M., & Du, H. (2021). Teaching Arm and Head Gestures to a Humanoid Robot through Interactive Demonstration and Spoken Instruction. *Proceedings of the 1st Workshop on Multimodal Semantic Representations 2021.*
12. Brishtel, I., Krauß, S., Schmidt, T., Rambach, J. R., Vozniak, I., & Stricker, D. (2022). Klassifizierung von manuellem und autonomen Fahren basierend auf maschinellem Lernen von Augenbewegungsmustern. *2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC), 700-705.*

13. Brishtel, I., Schmidt, T., Vozniak, I., Rambach, J. R., Mirbach, B., & Stricker, D. (2021). To Drive or to Be Driven? The Impact of Autopilot, Navigation System, and Printed Maps on Driver's Cognitive Workload and Spatial Knowledge. *ISPRS International Journal of Geo-Information, 10(10), 668.*
14. Dembinsky, D., Azimi, F., Raue, F., Hees, J., Palacio, S., & Dengel, A. (2023). Sequential Spatial Transformer Networks for Salient Object Classification. Proceedings of the 12th International Conference on Pattern Recognition Applications and Methods (ICPRAM), 1, 328-335.
15. Feldhus, N., Anagnostopoulou, A., Wang, Q., Alshomary, M., Wachsmuth, H., Sonntag, D., & Möller, S. (2024). Zur Modellierung und Evaluierung von Unterrichtserklärungen in Lehrer-Schüler-Dialogen. *GoodIT '24: Proceedings of the 2024 International Conference on Information Technology for Social Good, 15, 225-230*.
16. Feldhus, N., Hennig, L., Nasert, M.D., Ebert, C., Schwarzenberg, R., & Möller, S. (2023). Saliency Map Verbalization: Comparing Feature Importance Representations from Model-free and Instruction-based Methods. *Proceedings of the 1st Workshop on Natural Language Reasoning and Structured Explanations (NLRSE), 30-46*.
17. Feldhus, N., Ravichandran, A.M., & Möller, S. (2022). Mediators: Conversational Agents Explaining NLP Model Behavior. *Proceedings of the IJCAI 2022 Workshop on Explainable Artificial Intelligence (XAI), 157-167*.
18. Feldhus, N., Schwarzenberg, R., & Möller, S. (2021). Thermostat: A Large Collection of NLP Model Explanations and Analysis Tools. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 87-95*.
19. Feldhus, N., Wang, Q., Anikina, T., Chopra, S., Oguz, C., & Möller, S. (2023). InterroLang: Exploring NLP Models and Datasets through Dialogue-based Explanations. *Feststellungen der Gesellschaft für Computerlinguistik: EMNLP 2023, 5399-5421*.
20. Fritsch, S. G., Oguz, C., Rey, V. F., Ray, L., Kiefer-Emmanouilidis, M., & Lukowicz, P. (2024). MuJo: Multimodal Joint Feature Space Learning for Human Activity Recognition. *arXiv preprint arXiv:2406.03857.*
21. Frolov, S., Moser, B. B., Palacio, S., & Dengel, A. (2024). ObjBlur: A Curriculum Learning Approach with Progressive Object-Level Blurring for Improved Layout-to-Image Generation. *ACM Multimedia 2024.*
22. Frolov, S., Sharma, A., Hees, J., Karayil, T., Raue, F., & Dengel, A. (2021). AttrLostGAN: Attributgesteuerte Bildsynthese aus rekonfigurierbarem Layout und Stil. *Lecture Notes in Computer Science, 361-375.*
23. Ghosh, A., Cheema, N., Oguz, C., Theobalt, C., & Slusallek, P. (2021). Synthese von kompositorischen Animationen aus textuellen Beschreibungen. *2021 IEEE/CVF International Conference on Computer Vision (ICCV).*
24. Ghosh, A., Cheema, N., Oguz, C., Theobalt, C., & Slusallek, P. (2021). Synthese von kompositorischen Animationen aus textuellen Beschreibungen. *In Proceedings of the IEEE/CVF International Conference on Computer Vision (S. 1396-1406).*
25. Ghosh, A., Dabral, R., Golyanik, V., Theobalt, C., & Slusallek, P. (2023). IMoS: Intent-Driven Full-Body Motion Synthesis for Human-Object Interactions. *Computer Graphics Forum, 42(2), 1-12.*
26. Hartmann, M., Anagnostopoulou, A., & Sonntag, D. (2022). Interaktives maschinelles Lernen für Bildbeschriftungen.

27. Hartmann, M., Du, H., Feldhus, N., Kruijff-Korbayová, I., & Sonntag, D. (2022). XAINES: KI mit Narrativen erklären. *KI - Künstliche Intelligenz, 36(3-4), 287-296.*
28. Hartmann, M., & Sonntag, D. (2022). A Survey on Improving NLP Models with Human Explanations. *Proceedings of the First Workshop on Learning with Natural Language Supervision.*
29. Hartmann, M., Kruijff-Korbayová, I., & Sonntag, D. (2021). Interaktion mit Erklärungen im XAINES-Projekt. *Trustworthy AI in the Wild Workshop 2021.*
30. Jolly, S., Zhang, Z. X., Dengel, A., & Mou, L. (2022). Suchen und Lernen: Verbesserung der semantischen Abdeckung für Data-to-Text-Generierung. *Proceedings of the AAAI Conference on Artificial Intelligence, 36*(10), 10858-10866.
31. Jolly, S., Pezzelle, S., & Nabi, M. (2021). EaSe: A Diagnostic Tool for VQA Based on Answer Diversity. Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.
32. Sarti, G., Feldhus, N., Sickert, L., van der Wal, O., Nissim, M., & Bisazza, A. (2023). Inseq: An Interpretability Toolkit for Sequence Generation Models. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations), 421-435.
33. Schwarzenberg, R., Feldhus, N., & Möller, S. (2021). Effiziente Erklärungen aus empirischen Erklärern. In Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP, Seiten 240-249, Punta Cana, Dominikanische Republik. Gesellschaft für Computerlinguistik.
34. Pokaratsari, N., Amin, S., & Neumann, G. (2024). Understanding the Role of Graph Structures in Medical Codes Classification. (In submission).
35. Wang, Q., Anikina, T., Feldhus, N., van Genabith, J., Hennig, L., & Möller, S. (2024). LLMCheckup: Conversational Examination of Large Language Models via Interpretability Tools. In Proceedings of the Third Workshop on Bridging Human-Computer Interaction and Natural Language Processing, 89-104.

3 Fazit und Ausblick

Das XAINES-Projekt hat sich erfolgreich den Herausforderungen gestellt, die mit der erklärbaren künstlichen Intelligenz (XAI) verbunden sind, und wesentliche Fortschritte in der Entwicklung neuer Methoden zur Generierung narrativer Erklärungen für komplexe KI-Entscheidungen erzielt. Das Hauptziel des Projekts war es, die Transparenz und Verständlichkeit von KI-Systemen zu verbessern, indem eine Verbindung zwischen komplexen, multimodalen Daten und menschenverständlichen Erklärungen geschaffen wurde. Dies ist insbesondere in kritischen Anwendungsfeldern wie dem autonomen Fahren, der medizinischen Diagnostik und der Pflege von großer Bedeutung, wo das Vertrauen der Nutzer und die Nachvollziehbarkeit von Entscheidungen unerlässlich sind.
Das Projekt verfolgte einen umfassenden Ansatz, der sowohl technische als auch wissenschaftliche Innovationen beinhaltete. Die Entwicklung von Systemen, die Daten aus verschiedenen Modalitäten wie Sensoren, Sprache und visuellen Informationen nahtlos integrieren, war ein zentraler Fortschritt.

Die Integration von multimodalen Daten ermöglichte es, verschiedene Datenquellen wie Bewegungsverfolgung, Audio- und Videosignale mit Textdaten zu kombinieren. Diese Fähigkeit zur Multimodalität ermöglichte es, sensorische Eingaben in verständliche narrative Erklärungen zu übersetzen, die komplexe KI-Prozesse für den Menschen nachvollziehbar machen. Eine der Kerninnovationen war der Aufbau von Systemen, die in der Lage sind, KI-Entscheidungen zu erklären, indem sie aus Bildern, Texten und Sprachdaten schrittweise nachvollziehbare Beschreibungen generieren. Diese narrativen Erklärungen halfen dabei, die Funktionsweise der KI verständlicher zu machen und das Vertrauen der Nutzer in die Technologie zu stärken. Mit Hilfe fortschrittlicher Bildgenerierungs- und -beschreibungsverfahren wurden visuelle Darstellungen von KI-Prozessen entwickelt. Diese ermöglichten es, komplexe Zusammenhänge anschaulich zu machen und somit die Erklärbarkeit der Systeme weiter zu verbessern. Um den spezifischen Anforderungen der Anwendungsbereiche wie autonomes Fahren, Medizin und Pflege gerecht zu werden, wurden angepasste Erklärungsmechanismen entwickelt. Jede dieser Domänen brachte eigene technische Herausforderungen mit sich, die durch spezifische Anpassungen der Modelle überwunden wurden, wie z.B. die Erklärung von Navigationsentscheidungen im autonomen Fahren oder die Interpretation medizinischer Diagnosedaten.
Neben diesen technischen Erfolgen trug das XAINES-Projekt auch maßgeblich zur wissenschaftlichen Weiterentwicklung im Bereich der XAI bei. Die erarbeiteten Methoden führten zu neuen wissenschaftlichen Erkenntnissen darüber, wie komplexe KI-Systeme für den Menschen verständlich gemacht werden können. Ein zentrales wissenschaftliches Ziel war es, Erklärungen zu entwickeln, die es ermöglichen, KI-Entscheidungen nicht nur technisch korrekt, sondern auch intuitiv verständlich darzustellen. Diese Forschung wurde durch die enge Zusammenarbeit interdisziplinärer Teams aus verschiedenen DFKI-Forschungsbereichen und externen Partnern vorangetrieben, was die Integration und Anwendung der Ergebnisse in praxisnahen Szenarien erleichterte.

Trotz der erzielten Fortschritte bleibt die Erklärbarkeit von KI-Systemen ein dynamisches und herausforderndes Forschungsgebiet, das weiterhin erhebliche Aufmerksamkeit benötigt. Die im XAINES-Projekt entwickelten Methoden bieten eine solide Grundlage, um zukünftige Herausforderungen zu bewältigen, doch es gibt noch zahlreiche offene Fragen und Bereiche, die vertieft erforscht werden müssen.

Die Fähigkeit, aus komplexen Datenströmen verständliche Erklärungen zu generieren, sollte weiter verbessert werden. Dies umfasst die Entwicklung noch leistungsfähigerer Modelle, die in der Lage sind, tiefere semantische Zusammenhänge zu erkennen und komplexe Entscheidungsprozesse in klaren, narrativen Formaten darzustellen. Die Erforschung generativer KI-Ansätze könnte hierbei eine Schlüsselrolle spielen, um flexiblere und kontextreichere Erklärungen zu ermöglichen. Zukünftige Forschung sollte sich verstärkt der Entwicklung interaktiver Erklärungsmodelle widmen, die es den Nutzern ermöglichen, gezielte Fragen zu stellen und spezifische Informationen abzurufen. Dies könnte dazu beitragen, die Transparenz von KI-Systemen weiter zu erhöhen und das Vertrauen der Nutzer zu stärken, indem sie ein tieferes Verständnis für die zugrundeliegenden Mechanismen erhalten.
Die erarbeiteten Lösungen müssen in weiteren Anwendungsfeldern getestet werden, um deren Skalierbarkeit und Anpassungsfähigkeit zu gewährleisten. Besonders im Kontext neuer technologischer Entwicklungen und wachsender Anforderungen an KI-Systeme, wie z.B. im Bereich Industrie 4.0 oder der Smart Cities, könnte die Erklärbarkeit von KI eine entscheidende Rolle spielen. Das XAINES-Projekt hat gezeigt, wie wichtig die Zusammenarbeit zwischen verschiedenen Disziplinen ist. Diese Form der interdisziplinären und kollaborativen Forschung sollte weiter ausgebaut werden, um einen offenen Austausch von Wissen und Technologie zu fördern. Dies könnte durch die Bereitstellung der entwickelten Modelle und Datensätze an die wissenschaftliche Gemeinschaft unterstützt werden, um weiterführende Studien und Innovationen zu ermöglichen.

Abschließend lässt sich festhalten, dass das XAINES-Projekt nicht nur bedeutende Beiträge zur Verbesserung der Erklärbarkeit von KI geleistet hat, sondern auch die Grundlage für eine neue Generation von KI-Systemen geschaffen hat, die transparenter, sicherer und benutzerfreundlicher sind. Die erzielten Ergebnisse sind ein wichtiger Schritt in Richtung einer Zukunft, in der KI-Technologien als vertrauenswürdige Partner in verschiedensten Lebensbereichen agieren können. Zukünftige Projekte und Forschungen werden auf diesen Errungenschaften aufbauen, um die Kluft zwischen komplexen maschinellen Prozessen und menschlichem Verständnis weiter zu überbrücken und damit die Akzeptanz und den Einsatz von KI in kritischen Anwendungen zu fördern.

4 Literaturverzeichnis

Ahuja, C., & Morency, L.-P. (2019). Language2pose: Natural Language Grounded Pose Forecasting. Proceedings of the 2019 International Conference on 3D Vision, 719-728.

Ahuja, C., Lee, D. W., Nakano, Y. I., & Morency, L. P. (2020). Style Transfer for Co-speech Gesture Animation: A Multi-speaker Conditional-Mixture Approach. Computer Vision – ECCV 2020, 248–265.

Armeni, I., Sax, S., Zamir, A., & Savarese, S. (2017). Joint 2D-3D-Semantic Data for Indoor Scene Understanding. ArXiv, abs/1702.01105.

Bolukbasi, T., Pearce, A., Yuan, A., Coenen, A., Reif, E., Vi'egas, F., & Wattenberg, M. (2021). An Interpretability Illusion for BERT. ArXiv, abs/2104.07143.

Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., & Amodei, D. (2020). Language Models are Few-Shot Learners. Advances in neural information processing systems, 33, 1877-1901.

Cao, Z., Simon, T., Wei, S.-E., Sheikh, Y. (2017). Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. Proceedings of the IEEE Conference on Computer Cision and Pattern Recognition 2017.

Chang, A.X., Dai, A., Funkhouser, T.A., Halber, M., Nießner, M., Savva, M., Song, S., Zeng, A., & Zhang, Y. (2017). Matterport3D: Learning from RGB-D Data in Indoor Environments. 2017 International Conference on 3D Vision (3DV), 667-676.

Chou, S., Sun, C., Chang, W., Hsu, W., Sun, M., & Fu, J. (2020). 360-Indoor: Towards Learning Real-World Objects in 360° Indoor Equirectangular Images. 2020 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV).

Christen, S., Kocabas, M., Aksan, E., Hwangbo, J., Song, J., & Hilliges, O. (2022). D-Grasp: physically plausible dynamic grasp synthesis for Hand-Object interactions. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 1 (Long and Short Papers), 4171–4186.

Doan, T., Monteiro, J., Albuquerque, I., Mazoure, B., Durand, A., Pineau, J., & Hjelm, R. D. (2019). On-Line Adaptative Curriculum learning for GANs. Proceedings of the AAAI Conference on Artificial Intelligence, 33(01), 3470–3477.

Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., & Zisserman, A. (2009). The Pascal Visual Object Classes (VOC) challenge. International Journal of Computer Vision, 88(2), 303–338.

Gava, C. C., Stricker, D., & Yokota, S. (2018). Dense Scene Reconstruction from Spherical Light Fields. 2018 IEEE International Conference on Image Processing (ICIP).

Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A.C., & Bengio, Y. (2014). Generative Adversarial Nets. Conference on Neural Information Processing Systems 2014 (NeurIPS).

Goyal, Y., Khot, T., Summers-Stay, D., Batra, D. & Parikh, D. (2017). Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 6904–6913.

Gurari, D., Li, Q., Stangl, A. J., Guo, A., Lin, C., Grauman, K., Luo, J. & Bigham, J. P. (2018). VizWiz Grand Challenge: Answering Visual Questions from Blind People. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 3608–3617.

Habibie, I., Elgharib, M., Sarkar, K., Abdullah, A., Nyatsanga, S., Neff, M., & Theobalt, C. (2022). A Motion Matching-based Framework for Controllable Gesture Synthesis from Speech. SIGGRAPH ’22 Conference Proceedings.

Habtegebrial, T., Gava, C., Rogge, M., Stricker, D., & Jampani, V. (2022). SOMSI: Spherical Novel View Synthesis with Soft Occlusion Multi-Sphere Images. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

Hanser, E., Mc Kevitt, P., Lunney, T., & Condell, J. (2010). SceneMaker: Intelligent Multimodal Visualisation of Natural Language Scripts. Irish Conference on Artificial Intelligence and Cognitive Science, 144-153.

He, K., Gkioxari, G., Dollár, P, Girshick, R. (2017). Mask R-CNN. IEEE International Conference on Computer Vision (ICCV), 2980–2988.

He, S., Liao, W., Yang, M. Y., Yang, Y., Song, Y., Rosenhahn, B., & Xiang, T. (2021). Context-Aware Layout to Image Generation with Enhanced Object Appearance. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

Host, K., & Ivašić-Kos, M. (2022). An overview of Human Action Recognition in sports based on Computer Vision. Heliyon, 8(6), e09633.

Kwon, H., Tong, C., Haresamudram, H., Gao, Y., Abowd, G. D., Lane, N. D., & Ploetz, T. (2020). IMUTube: Automatic Extraction of Virtual on-body Accelerometry from Video for Human Activity Recognition. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 4(3), 1-29.

LaFortune, J., & Macuga, K. L. (2018). Learning Movements from a Virtual Instructor: Effects of Spatial Orientation, Immersion, and Expertise. Journal of Experimental Psychology: Applied, 24(4), 521–533.

Lee, H. Y., Yang, X., Liu, M. Y., Wang, T. C., Lu, Y. D., Yang, M. H. & Kautz, J. (2019). Dancing to Music. Neural Information Processing Systems, 32, 3581–3591.

Olah, C., Mordvintsev, A., & Schubert, L. (2017). Feature visualization. Distill, 2(11).

Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., Grangier, D. & Auli, M. (2019). Fairseq: a fast, extensible toolkit for sequence modeling. Proceedings Of The 2019 Conference Of The North American Chapter Of The Association For Computational Linguistics (Demonstrations).

Pagani, A., Gava, C., Cui, Y., Krolla, B., Hengen, J., & Stricker, D. (2011). Dense 3D point cloud generation from multiple high-resolution spherical images. International Conference on Virtual Reality, 17–24.

Petrovich, M., Black, M. J. & Varol, G. (2022). TEMOS: Generating Diverse Human Motions from Textual Descriptions. European Conference on Computer Vision (ECCV) 2022, 480–497.

Piyathilaka, L., & Kodagoda, S. (2015). Human activity recognition for domestic robots. Field and Service Robotics. Springer Tracts in Advanced Robotics, vol 105, 395 – 408.

Plappert, M., Mandery, C., & Asfour, T. (2016). The KIT Motion-Language Dataset. Big data, 4(4), 236-252.

Plappert, M., Mandery, C., & Asfour, T. (2018). Learning a Bidirectional Mapping Between Human Whole-Body Motion and Natural Language Using Deep Recurrent Neural Networks. Robotics and Autonomous Systems, 109, 13–26.

Poerner, N., Roth, B., & Schütze, H. (2018). Interpretable Textual Neuron Representations for NLP. Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, 325–327.

Puig, X., Ra, K.K., Boben, M., Li, J., Wang, T., Fidler, S., & Torralba, A. (2018). VirtualHome: Simulating Household Activities Via Programs. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 8494-8502.

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision. International Conference on Machine Learning, 8748-8763.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners. OpenAI blog, 1(8), 9.

Rempe, D., Birdal, T., Hertzmann, A., Yang, J., Sridhar, S., & Guibas, L. J. (2021). HUMOR: 3D Human Motion Model for Robust Pose Estimation. 2021 IEEE/CVF International Conference on Computer Vision (ICCV).

Salimans, T., Goodfellow, I.J., Zaremba, W., Cheung, V., Radford, A., & Chen, X. (2016). Improved Techniques for Training GANs. Conference on Neural Information Processing Systems 2016 (NeurIPS).

Slack, D., Krishna, S., Lakkaraju, H., & Singh, S. (2023). Explaining machine learning models with interactive natural language conversations using TalkToModel. Nature Machine Intelligence, 5(8), 873–883.

Soviany, P., Ionescu, R. T., Rota, P., & Sebe, N. (2022). Curriculum Learning: a survey. International Journal of Computer Vision, 130(6), 1526–1565.

Sun, Y., Bao, Q., Liu, W., Fu, Y., Black, M. J., & Mei, T. (2021). Monocular, One-Stage, Regression of Multiple 3D People. Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision.

Sung, K., Shirley, P., & Rosenberg, B. R. (2007). Experiencing aspects of games programming in an introductory computer graphics class. Proceedings of the 38th SIGCSE Technical Symposium on Computer Science Education.

Sunil, A., Sheth, M. H., E, S., & Mohana, N. (2021). Usual and Unusual Human Activity Recognition in Video using Deep Learning and Artificial Intelligence for Security Applications. 2021 Fourth International Conference on Electrical, Computer and Communication Technologies (ICECCT), 1 – 6.

Taheri, O., Ghorbani, N., Black, M. J., & Tzionas, D. (2020). GRAB: a dataset of Whole-Body human grasping of objects. European Conference on Computer Vision (ECCV). Lecture notes in computer science, vol 12349, 581-600.

Taheri, O., Choutas, V., Black, M.J., & Tzionas, D. (2022). GOAL: Generating 4D Whole-Body Motion for Hand-Object Grasping. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 13253-13263.

Tan, J., Beheshti, B. K., Binnie, T., Davey, P., Caneiro, J., Kent, P., Smith, A., O'Sullivan, P., & Campbell, A. (2021). Human Activity Recognition for People with Knee Osteoarthritis—A Proof-of-Concept. Sensors, 21(10), 3381.

Tang, C.I., Perez-Pozuelo, I., Spathis, D., & Mascolo, C. (2020). Exploring Contrastive Learning in Human Activity Recognition for Healthcare. ArXiv, abs/2011.11542.

Tran, N., Tran, V., Nguyen, N., Nguyen, T., & Cheung, N. (2021). On data augmentation for GAN training. IEEE Transactions on Image Processing, 30, 1882–1897.

Wachsmuth, H., & Alshomary, M. (2022). "Mama Always Had a Way of Explaining Things So I Could Understand": A Dialogue Corpus for Learning to Construct Explanations. Proceedings of the 29th International Conference on Computational Linguistics, pages 344–354, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Won, C., Ryu, J., & Lim, J. (2020). End-to-End learning for omnidirectional stereo matching with uncertainty prior. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(11), 3850–3862.

Xu, M., Yoon, S., Fuentes, A., & Park, D. S. (2023). A comprehensive survey of image augmentation techniques for deep learning. Pattern Recognition, 137, 109347.

Yuan, H., Yu, H., Gui, S., & Ji, S. (2022). Explainability in Graph Neural Networks: A Taxonomic survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1–19.

Zhang, C., Cui, Z., Chen, C., Liu, S., Zeng, B., Bao, H., & Zhang, Y. (2021). DeepPanoContext: Panoramic 3D Scene Understanding with Holistic Scene Context Graph and Relation-based Optimization. 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 12632–12641.

Zhang, H., Ye, Y., Shiratori, T. & Komura, T. (2021). ManipNet: neural manipulation synthesis with a hand-object spatial representation. ACM Transactions on Graphics, 40(4), 1–14.

Zhang, H., Zhang, Z., Odena, A., & Lee, H. (2020). Consistency Regularization for Generative Adversarial Networks. Eighth International Conference on Learning Representations.

Zhao, S., Liu, Z., Lin, J., Zhu, J. Y., & Han, S. (2020). Differentiable augmentation for data-efficient gan training. Advances in neural information processing systems, 33, 7559-7570.

Zheng, J., Zhang, J., Li, J., Tang, R., Gao, S., & Zhou, Z. (2020). Structured3D: a large Photo-Realistic dataset for structured 3D modeling. European Conference on Computer Vision (ECCV). Lecture notes in computer science, vol 12354, 519–535.

Zheng, G., Zhou, X., Li, X., Qi, Z., Shan, Y., & Li, X. (2023). LayoutDiffusion: Controllable Diffusion Model for Layout-to-Image Generation. The IEEE/CVF Conference on Computer Vision and Pattern Recognition 2023.

Zou, C., Colburn, A., Shan, Q., & Hoiem, D. (2018). LayoutNet: Reconstructing the 3D Room Layout from a Single RGB Image. The IEEE/CVF Conference on Computer Vision and Pattern Recognition 2018, 2051–2059.