Von Bettina Finzel und Ute Schmid

Wie Menschen und KI-Systeme voneinander lernen können

Kritische Entscheidungen gemeinsam treffen

Viele Menschen verbinden mit dem Thema Künstliche Intelligenz (KI) die Befürchtung, dass zukünftig intelligente autonome Systeme das Sagen haben. Wenn es aber um Entscheidungen in komplexen, kritischen Bereichen von Medizin bis Haftungsrecht geht, dann müssen Mensch und KI zusammenarbeiten. Für solche Partnerschaften sind Methoden nötig, die Entscheidungen von KI-Systemen transparent und nachvollziehbar machen – und Ansätze, mit denen Menschen diese korrigieren können.

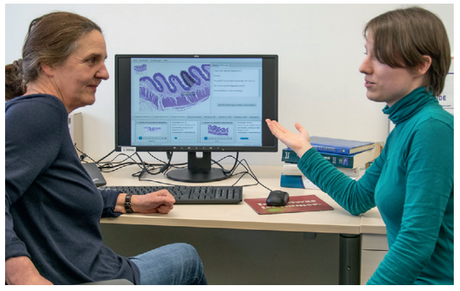
Der aktuelle Hype um die Künstliche Intelligenz wurde vor allem durch beeindruckende Erfolge von datenintensiven Ansätzen des maschinellen Lernens ausgelöst. Beispielsweise erkennen *Convolutional Neural Networks* (CNNs) Objekte auf Bildern, ohne dass vorher mittels komplexer Algorithmen der Bildverarbeitung Merkmale extrahiert werden müssen. Diese Möglichkeit, direkt aus Rohdaten zu lernen (*end-to-end learning*), hat die Hoffnung geweckt, dass in vielen Bereichen des Arbeitslebens Aufgaben zukünftig effizient durch autonome intelligente Systeme erledigt werden können.

Zunehmend hat sich jedoch gezeigt, dass es in hochspezialisierten Bereichen wie etwa der industriellen Qualitätskontrolle oder der medizinischen Diagnostik weder wünschenswert noch möglich ist, Entscheidungen alleine durch ein maschinell gelerntes Modell treffen zu lassen: Oft gibt es gar nicht genügend korrekt vorklassifizierte Daten zum Training solcher Modelle. Und in kritischen Bereichen reicht die mit den Modellen erzielte Klassifikationsgenauigkeit nicht für zuverlässige Entscheidungen aus. Außerdem ist zunehmend ein Bewusstsein dafür entstanden, dass es weder rechtlich noch ethisch oder gesellschaftlich wünschenswert sein kann, Menschen von relevanten Entscheidungen auszuschließen. Entsprechend werden in der Forschung zu maschinellem Lernen zunehmend Ansätze entwickelt, die eine Kooperation von Mensch und KI ermöglichen. Zentral sind hier Methoden des erklärenden und interaktiven maschinellen Lernens. An der Universität Bamberg werden solche partnerschaftlichen KI-Ansätze bereits seit mehreren Jahren entwickelt.

Interaktives Maschinelles Lernen für Tumordiagnosen

Ein System, bei dem KI und Mensch zusammenarbeiten, wird im Rahmen des Forschungsprojekts *Transparent Medical Expert Companion* (TraMeExCo) entwickelt, das vom Bundesministerium für Bildung und Forschung gefördert wird. Die Universität Bamberg arbeitet in diesem Projekt mit dem Fraunhofer-Institut für Integrierte Schaltungen (IIS) Erlangen sowie verschiedenen medizinischen Expertinnen und Experten zusammen, um zum Beispiel interaktives maschinelles Lernen für die Diagnose von Darmkrebs zu erforschen.

Im TraMeExCo-Projekt werden CNNs zur Klassifizierung von Gewebeproben aus dem menschlichen Darm genutzt. Trotz ihrer hohen Vorhersagegenauigkeit können klassische CNNs derzeit jedoch nicht in der Praxis angewandt werden, da der Grund für eine Klassifizierungsentscheidung ohne den Einsatz ergänzender Methoden nicht überprüft und



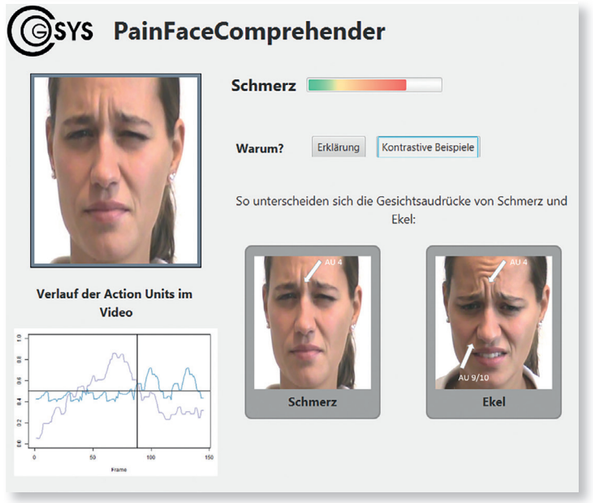
Ute Schmid (l.) und Bettina Finzel begannen das Projekt TraMeExCo 2018.

rekonstruiert werden kann. Medizinische Expertinnen und Experten sind auf Transparenz und Nachvollziehbarkeit angewiesen. Nur dann können sie Entscheidungen treffen, die zu einer angemessenen Behandlung der Patientinnen und Patienten führen. Methoden der *Explainable Artificial Intelligence* (XAI) können eingesetzt werden, um dieses Ziel zu erreichen. Sie werden im Rahmen des TraMeExCo-Projekts erforscht. XAI-Methoden werden beispielsweise zur Generierung visueller Erklärungen eingesetzt. Diese heben Regionen in Bildern hervor, die für die Entscheidung eines CNNs relevant waren, zum Beispiel Tumorgewebe. Verbale

Interaktive Korrektur von Klassifikationen im Programm *LearnWithME* (eigene Darstellung, Demonstrator für Tumorklassifikation aus dem Forschungsprojekt TraMeExCo).

Erklärungen in Form von natürlich-sprachlichen Sätzen können generiert werden, um im Vergleich zu visuellen Ansätzen komplexere Erklärungen zu erzeugen. So kann zum Beispiel nicht nur ausgedrückt werden, wo ein Tumor vorhanden ist, sondern auch, dass dieser bereits räumlich in verschiedene Gewebeschichten vorgedrungen ist.

Das Erzeugen von Erklärungen basiert im TraMeExCo-Projekt unter anderem auf *Inductive Logic Programming* (ILP). Das ist eine interpretierbare maschinelle Lernmethode, die als Alternative oder Ergänzung zu intransparenten Ansätzen wie CNNs zur Klassifizierung und zur Generierung verbaler Erklärungen verwendet werden kann. Zusätzlich zu den Erklärungen muss eventuell eine Möglichkeit der Korrektur von maschinell gelernten Entscheidungsmodellen bereitgestellt werden, um Fehlentscheidungen durch die KI zu vermeiden. Um dem medizinischen Personal eine möglichst einfache Korrektur zu ermöglichen, kann es in der Bamberger Software Erklärungen direkt ansehen und korrigieren und dadurch eine Anpassung des gelernten Modells durch Nutzerfeedback bewirken. Auf diese Weise kann Fachwissen in den Lernprozess einfließen und ist somit Teil eines wechselseitigen Erklärungsprozesses: Das KI-System erklärt die Klassifizierungsentscheidung und die Expertin oder der Experte gibt dem System korrektive Rückmeldungen.



Kontrastive Erklärungen im Programm *PainFaceComprehender* (eigene Darstellung: Miriam Kunz, Stefan Lautenbacher/Professur für Physiologische Psychologie).

Verschiedene Arten von Erklärungen für die Erkennung von Schmerz

Ein weiteres in Bamberg entwickeltes KI-System dient der Unterstützung medizinischen und psychologischen Personals in der klinischen Schmerzdiagnose und -behandlung. Im Rahmen des Forschungsprojekts *PainFaceReader* arbeitet die Informatik mit der Physiologischen Psychologie an der Universität Bamberg zusammen und untersucht gemeinsam mit Ingenieurinnen und Ingenieuren vom Fraunhofer IIS, wie Schmerz automatisiert erkannt werden kann. Gefördert wird das Projekt von der Deutschen Forschungsgemeinschaft.

Hierzu betrachten die Forschenden insbesondere die menschliche Mimik. Dies ist besonders dann nützlich zur Diagnose und Behandlung von Schmerz, wenn Patientinnen oder Patienten nicht gut kommunizieren können, was sie empfinden und wie stark die Schmerzen sind, die sie spüren. Dies trifft beispielsweise auf Demenzkranke und intensivmedizinisch behandelte Patientinnen und Patienten zu. Schmerzen bleiben bei diesen Betroffenen daher oft unerkannt oder können nur schwer eingeschätzt werden. Das *PainFaceReader*-Forschungsteam entwickelt deshalb ein KI-System, das es ermöglicht, Schmerzen zu erkennen, ohne die Betroffenen befragen zu müssen. Das lernende System wertet Videoaufnahmen Betroffener aus und interpretiert deren Gesichtszüge sowie den zeitlichen Verlauf von Veränderungen in der menschlichen Mimik.

Wie bei TraMeExCo nutzen die Forschenden auch hier Methoden, die *Deep Learning* mit *Inductive Logic Programming* verbinden, um ein transparentes und nachvollziehbares System bereitzustellen, das Menschen in ihren Entscheidungen unterstützt. Damit sowohl Expertinnen und Experten als auch Laien im Bereich Schmerzerkennung das System nutzen können, werden hier neben visuellen und sprachlichen Erklärungen auch beispielbasierte Erklärungsmethoden benutzt. So können prototypische Bilder genutzt werden, um zu verdeutlichen, welche Art von mimischem Ausdruck bei einer bestimmten Person oder Personengruppe ein Indikator für Schmerz ist. Gegenüberstellende Beispiele werden gezeigt, um feine Unterschiede von Schmerzmimik zu mimischen Ausdrücken, die mit anderen mentalen Zuständen einhergehen, zu verdeutlichen.

Erklärungen durch Prototypen werden beispielsweise in medizinischen Lehrbüchern genutzt, um Studierenden typische Symptome anschaulich zu vermitteln. Visuelle Erklärungen werden genutzt, um Verwechslungen von leicht verwechselbaren Kategorien zu vermeiden. Das System soll künftig ein langfristiges Monitoring in Kliniken ermöglichen, um die Versorgung von Schmerzpatientinnen und -patienten zu unterstützen. Privatheit und Datenschutz sind bei dieser auf permanentem Monitoring des Gesichts angewiesenen Anwendung von besonderer Wichtigkeit. Deshalb stellt bereits die Kamera das Gesicht abstrakt dar und projiziert abstrakte Beschreibungen der mimischen Aktionen auf ein Avatargesicht.

Literaturempfehlung

Lun Ai, Stephen H. Muggleton, Céline Hocquette, Mark Gromowski, Ute Schmid: *Beneficial and Harmful Explanatory Machine Learning*. Machine Learning.

Ute Schmid, Bettina Finzel (2020): *Mutual Explanations for Cooperative Decision Making in Medicine*. Künstliche Intelligenz, 34(2), S. 227-233.

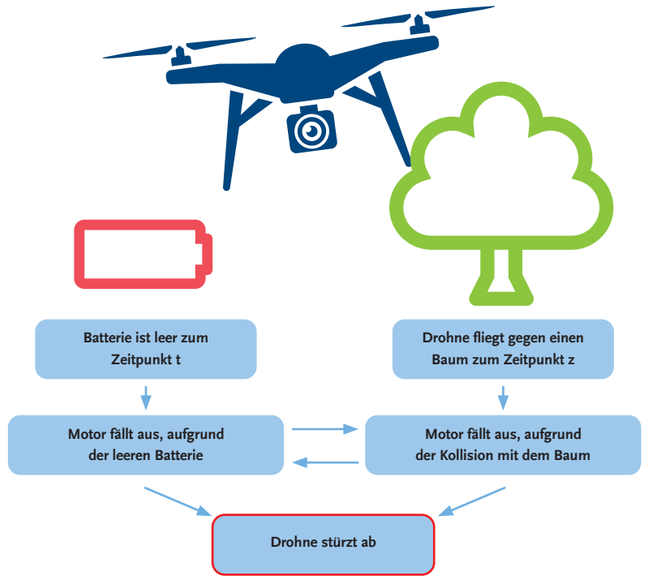
Johannes Rabold, Hannah Deininger, Michael Siebers, Ute Schmid (2019): *Enriching Visual with Verbal Explanations for Relational Concepts – Combining LIME with Aleph*. PKDD/ECML Workshops (1), Springer, S. 180-192.

Schrittweises Lernen für die Ursachenanalyse bei Unfällen

In einem Projekt am Bayerischen Forschungsinstitut für Digitale Transformation (bidt) arbeitet Ute Schmid von der Universität Bamberg mit dem Informatiker Alexander Pretschner von der TU München und dem Rechtswissenschaftler Eric Hilgendorf von der Universität Würzburg zusammen. Sie gestalten Mensch-KI-Partnerschaften für Erklärungen in komplexen sozio-technischen Systemen: Zusammenhänge zwischen Ursache und Wirkung für haftungsrechtliche Aspekte bei Unfällen oder Fehlern sollen mit technischen Systemen nachvollziehbar gemacht werden.

Begonnen wird mit einem Ursache-Wirkungsmodell, bei dem bekannte Einflussgrößen in einem Abhängigkeitsgraph dargestellt werden. Will man etwa herausfinden, was die Ursache für einen Drohnenabsturz war, so kann das an einer leeren Batterie liegen, was wiederum dazu führt, dass der Motor ausgefallen ist. Mit einem Ansatz des interaktiven maschinellen Lernens kann ein solches initiales Modell durch neue Informationen erweitert und korrigiert werden. Ist beispielsweise eine Drohne gegen einen Baum geflogen und dann abgestürzt, können diese neuen Einflussgrößen in das Modell integriert werden. Zudem können menschliche Expertinnen und Experten sich Ausschnitte des Modells anzeigen lassen und dieses korrigieren.

Nicht nur in den gezeigten Anwendungsbereichen aus Medizin, Pflege und Unfallanalysen, sondern prinzipiell in allen Bereichen des Arbeitslebens und des Alltags können KI-Systeme zum Wohle von Menschen eingesetzt werden. Damit der Einsatz von KI der Menschheit nicht die Kontrolle entzieht, sondern damit KI menschliche Kompetenzen fördert und erweitert, sind partnerschaftliche Methoden für Mensch-KI-Kooperation notwendig. Die Forschung im Bereich Induktive Logische Programmierung liefert einen wichtigen Baustein für solche menschzentrierten KI-Systeme: durch die Kombination von wissensbasierten und lernenden Ansätzen, die Möglichkeit von interaktivem maschinellem Lernen und als Grundlage für die Erzeugung verschiedener Arten von Erklärungen.



Ausschnitt aus einem Ursachengraph zur Analyse von Drohnenabstürzen. Die Konzepte (Knoten) sowie deren Abhängigkeiten (Kanten) können durch interaktives maschinelles Lernen von menschlichen Expertinnen und Experten korrigiert werden.

How Humans and AI Systems Can Learn From Each Other

EN

Making crucial decisions together

Many people associate the topic of artificial intelligence (AI) with the fear that intelligent autonomous systems could someday be calling the shots. But when it comes to decisions in complex, critical areas from medicine to liability law, humans and AI need to work together. Such partnerships require methods that render the decisions made by AI systems transparent and comprehensible – and approaches that enable humans to correct them.