Das Fraunhofer-Institut für Naturwissenschaftlich-Technische Trendanalysen berichtet über neue Technologien

Explainable Artificial Intelligence

Verfahren der künstlichen Intelligenz (KI) werden in immer mehr Bereichen zum Bestandteil unseres alltäglichen Lebens. Verwendung finden sie heute z. B. schon zum Erkennen von Objekten auf Bildern oder zur Verarbeitung von Sprache. Dabei übertreffen sie teilweise bereits die Leistungsfähigkeit von Menschen. Die Ergebnisse von zurzeit eingesetzten KI-Systemen sind jedoch häufig nicht nachvollziehbar, beispielsweise warum auf einem Bild ein bestimmtes Objekt erkannt wurde. Im Themenfeld Explainable Artificial Intelligence (XAI) wird deshalb daran gearbeitet, die Resultate und Funktionsweise von KI-Systemen in einer für Menschen verständlichen Form zu erklären bzw. überprüfbar zu machen. Diese Fähigkeit ist eine wichtige Voraussetzung, um das Vertrauen eines Benutzers in das KI-System zu steigern. Darüber hinaus ermöglicht sie es, die Stärken und Schwächen des Systems besser beurteilen zu können. XAI befindet sich zurzeit noch in einem relativ frühen Entwicklungsstadium.

Die aktuellen Fortschritte im KI-Bereich sind ganz wesentlich durch den Einsatz von Verfahren des maschinellen Lernens (ML) erreicht worden. Diese erlauben es Computern, selbstständig anhand von Beispieldaten bestimmte Sachverhalte zu erlernen, z. B. ob auf einem Bild ein gewisses Objekt abgebildet ist. Die Basis für ML bildet jeweils ein Modell mit einstellbaren Parametern, wobei diese Parameter in der Lernphase im Hinblick auf die Beispieldaten optimiert werden. Die derzeit sehr erfolgreichen ML-Modelle des Deep Learning basieren auf sogenannten künstlichen neuronalen Netzwerken mit einer sehr großen Anzahl von Parametern. Das erlernte Wissen ist aufgrund dieser hohen Komplexität des Modells typischerweise nicht verstehbar. Man spricht daher bei Deep Learning häufig auch von einer Black Box. Andere derzeit verwendete ML-Modelle, wie z. B. die sogenannten Support Vector Machines und Random Forests, gelten vielfach ebenfalls als Black-Box-Modelle.

Allgemein scheint ein reziproker Zusammenhang zwischen der Leistungsfähigkeit eines ML-Modells und dessen Erklärbarkeit zu existieren. So sind die aktuell leistungsfähigsten ML-Modelle, wie z. B. Deep Learning, vielfach auch die am wenigsten erklärbaren, während besser nachvollziehbare Modelle, wie etwa Entscheidungsbäume, häufig schlechtere Ergebnisse liefern. Entscheidungsbäume sind hierarchisch strukturiert und enthalten für Menschen verständliche Regeln, die das Wissen widerspiegeln, das anhand der Beispieldaten erlernt wurde.

Ansätze im Bereich XAI können im Wesentlichen in transparente Modelle und Post-hoc-Erklärungen eingeteilt werden. Transparente Modelle werden von vornherein mit dem Ziel entworfen, zu einem gewissen Grad verständlich zu sein. Zu den transparenten Modellen gehören z. B. die genannten Entscheidungsbäume. Post-hoc-Erklärungen umfassen dagegen Verfahren, die die Ergebnisse von Modellen erklären können, welche nicht von vornherein als transparente Modelle entworfen wurden, wie z. B. typische Verfahren aus dem Bereich Deep Learning. Das zu erklärende komplexe Modell wird hierzu auf ein einfacheres Modell abgebildet, mit dessen Hilfe dann die Erläuterungen erzeugt werden, während das komplexe Modell nach wie vor die eigentlichen Ergebnisse liefert.

Eine große Bedeutung wird XAI insbesondere in sicherheitskritischen Anwendungsfeldern von KI zugesprochen. Sie wäre bspw. im medizinischen Bereich von Interesse, um die Gründe für eine mit Hilfe von KI-Systemen erstellte Diagnose, wie z. B. die Identifizierung eines Tumors auf einer Röntgenaufnahme, zu erklären. Bei autonomen Fahrzeugen könnte XAI u. a. helfen zu verstehen, warum ein solches Fahrzeug bei einem Unfall ein Objekt nicht korrekt erkannt hat. In der Justiz könnte sie z. B. bei der Prognose der Wahrscheinlichkeit für einen potenziellen Rückfall bei Straftätern anhand von sozialen Einflussgrößen verwendet werden. Im Rahmen der IT-Sicherheit wäre z. B. die Fähigkeit von Interesse, die Handlungsempfehlungen eines KI-Systems bei Cyber-Angriffen zu erläutern. In der Industrie könnte u. a. die Abschätzung, wann eine Wartung bei technischen Systemen erfolgen sollte, profitieren (Predictive Maintenance). Außerdem könnte XAI auch in der Wissenschaft eingesetzt werden, z. B. in der Arzneimittelforschung oder bei der Entwicklung von neuen Werkstoffen. Durch XAI könnten hier bspw. Hinweise auf potenzielle, bisher unbekannte kausale Zusammenhänge innerhalb von Daten erhalten werden, die dann in Experimenten näher untersucht werden könnten. Weitere potenzielle Anwendungsfelder finden sich im Personalwesen bei der Auswahl von geeigneten Bewerbern, um hierbei eine faire Behandlung zu gewährleisten, oder im Finanzsektor im Hinblick auf die Beurteilung von Kreditanträgen.

Obwohl bereits merkliche Fortschritte auf dem Gebiet der XAI erzielt werden konnten, existieren trotzdem noch einige Herausforderungen. So gibt es bspw. noch keine allgemein anerkannte formale Definition davon, was unter Erklärbarkeit im Rahmen von XAI genau zu verstehen ist. Bereits existierende Ansätze beinhalten außerdem in erster Linie Erklärungen, die sich an die Entwickler von KI-Systemen wenden. Zukünftig sind aber auch vermehrt Erklärungen zu erwarten, die sich an andere Adressaten wenden, wie z. B. die Nutzer von KI-Systemen. Das augenblicklich hohe Interesse an XAI ist eng mit der gegenwärtigen Popularität von ML-Verfahren, insbesondere von Deep Learning, verbunden. Ein Forschungsschwerpunkt liegt daher aktuell auf Erklärungen von derartigen Modellen. In diesem Zusammenhang werden vielfach Post-hoc-Erklärungen und visuelle Erklärungen eingesetzt. So werden z. B. mit Hilfe von sogenannten Heatmaps diejenigen Bereiche eines Bildes farblich hervorgehoben, die am stärksten das Ergebnis des Systems beeinflusst haben. Ein vielversprechender Ansatz zur zukünftigen Verbesserung von Erklärungen besteht generell darin, diese interaktiv zu gestalten. In einer Art von Konversation zwischen dem erklärenden System und dem Adressaten der Erklärung kann dieser dann auch individuelle Fragen stellen.

Dr. Klaus Ruhlig