GASTBEITRAG

Künstliche Intelligenz: Erklären oder nicht erklären?

Von Menschen entworfene modulare Systeme können hochkomplex sein. Moderne KI-Systeme übertreffen diese Komplexität um ein Vielfaches. Welche Funktionen oder Qualitäten einzelne Neuronen im Modell einnehmen, bleibt aber weitgehend unklar. Dadurch sind Ergebnisse oft nicht nachvollziehbar, das ist in vielen Bereichen problematisch. Was können Methoden und Trends Erklärbarer KI (XAI) beitragen, um Vertrauen und Qualität zu verbessern? **Prof. Dr. Wojciech Samek** (TU Berlin / Fraunhofer HHI) gibt einen Einblick.

Müssen wir die KI wirklich verstehen, um sie nutzen und ihr vertrauen zu können? Ein verbreiteter Standpunkt lautet: Nein – wir nehmen auch Medikamente ein, bei denen der genaue Wirkmechanismus noch nicht vollständig geklärt ist. Entscheidend seien gute Evaluierungsverfahren, mit denen sich die Leistung der KI testen lässt. Doch genau hier beginnt das Problem. Seit Jahren wurden KI-Modelle nur anhand von Performanz-Metriken evaluiert. Mit der Entwicklung von Methoden zur Erklärbarkeit zeigte sich jedoch, dass Modelle mit guter Performanz die Aufgaben nicht immer „verstehen“, sondern besonders effektiv schummeln können. Dabei werden beispielsweise Pferdebilder nicht anhand des Pferdes selbst, sondern über ein in Pferdebildern häufig vorkommendes Copyright-Wasserzeichen erkannt.

Erklärbarkeit als Game Changer

Erklärbarkeit ist also entscheidend, um Fehler in KI-Modellen frühzeitig zu erkennen und sicherzustellen, dass die Entscheidungsprozesse des Modells nachvollziehbar und sinnvoll sind. Das gilt sowohl für Pferdebildklassifikatoren als auch für halluzinierende Sprachmodelle. Erklärbarkeit bietet jedoch noch mehr: So konnte mit Hilfe von erklärbaren Modellen eine ganz neue strukturelle Klasse von Antibiotika entdeckt werden. Auch aus rechtlicher Sicht gewinnt Erklärbarkeit an Bedeutung, etwa durch neue Vorschriften wie den EU AI Act, der Transparenz in bestimmten KI-Anwendungen fordert.

Deutschland ist im Bereich Erklärbarkeit sehr gut aufgestellt. Hier wurden nicht nur viele fundamentale Techniken entwickelt, auch sind einige der führenden Forscher hier ansässig. Dieses Wissen und der Standortvorteil sollten genutzt werden, um vertrauenswürdigere und überprüfbare KI zu schaffen.

Drei Wellen der Erklärbarkeitsforschung

1. Erklärungen einzelner Vorhersagen

Die ersten Methoden zielten darauf ab, einzelne Modellentscheidungen zu erklären, indem sie den Einfluss einzelner Eingabedimensionen (z. B. Pixel) auf die Vorhersage sichtbar machen. Unterschiedliche Verfahren wurden entwickelt, um diese Erklärungen zu berechnen. Zum Beispiel basiert das Layer-wise Relevance Propagation (LRP)-Verfahren auf

PROF. DR. WOJCIECH SAMEK
Professor für maschinelles Lernen und Kommunikation an der Technischen Universität Berlin und Leiter der Abteilung für Künstliche Intelligenz und der Explainable AI Group am Fraunhofer Heinrich-Hertz-Institut (HHI). In der Plattform Lernende Systeme ist Wojciech Samek Mitglied der Arbeitsgruppe Technologische Wegbereiter und Data Science.

der Idee, die Vorhersage rückwärts durch das Netz zu verteilen. Neuronen, die stärker zu der Entscheidung beigetragen haben, erhalten dabei einen proportional größeren Anteil an der Gesamtrelevanz. Die Relevanzwerte, die jedem Pixel des Eingangsbildes zugeordnet werden, zeigen, welche Bildbereiche für die Entscheidung der KI ausschlaggebend waren.

2. Verständnis des Modells selbst

Die zweite Welle der Erklärbarkeitsforschung zielte darauf, das KI-Modell selbst besser zu verstehen. Mit Hilfe der Activation Maximization-Methode kann z. B. angezeigt werden, welche Merkmale einzelne Neuronen kodieren. Das Concept Relevance Propagation (CRP)-Verfahren erweitert diese Art von Erklärungen und erlaubt es, die Rolle und Funktion einzelner Neuronen bei Modellentscheidungen zu analysieren. Diese Methoden der zweiten XAI-Welle bilden die Grundlage der aufkommenden mechanistischen Interpretierbarkeit, die funktionale Subnetzwerke („Schaltkreise“) im Modell analysiert.

3. Ganzheitliches Verständnis

Ziel der neuesten Methoden der XAI-Forschung ist es, ein systematisches Verständnis vom Modell, seinem Verhalten und seinen Repräsentationen zu erhalten. Methoden wie SemanticLens versuchen die Funktion und Qualität jeder einzelnen Komponente (Neuron) im Modell zu verstehen. Dieses ganzheitliche Verständnis erlaubt systematische, automatisierbare Modellprüfungen, z. B. ob ein Hautkrebsmodell wirklich der medizinischen ABCDE-Regel folgt.

Zukunft der Erklärbarkeitsforschung

Mit der Entwicklung von immer komplexeren Modellen wird die Erklärbarkeit weiter an Bedeutung gewinnen, sowohl als Werkzeug zur Mensch-KI-Interaktion als auch für die systematische Analyse, Prüfung und Verbesserung von Modellen. Gerade große Sprachmodelle bieten eine ideale Grundlage, um die Rolle einzelner Komponenten gezielt zu untersuchen und das Modell aktiv zu steuern, etwa zur Vermeidung von Halluzinationen. Die Methoden entwickeln sich somit weiter: von der reinen Erklärung hin zu gezielten Eingriffsmöglichkeiten – ein entscheidender Schritt für den sicheren und verantwortungsvollen Einsatz moderner KI-Systeme.