FORSCHUNG KOMPAKT

Künstliche Intelligenz erklärbar machen

Der Blick in Neuronale Netze

Künstliche Intelligenz, kurz KI, ist längst in unserem Alltag präsent und dringt in immer mehr Bereiche vor. Sprachassistenten etwa sind bereits als Helfer auf dem Smartphone, im Auto oder zu Hause Normalität geworden. Fortschritte im Bereich der KI beruhen vor allem auf der Verwendung Neuronaler Netze. Vergleichbar mit der Funktionsweise des menschlichen Gehirns verknüpfen sie mathematisch definierte Einheiten miteinander. Doch bisher wusste man nicht, wie ein Neuronales Netz Entscheidungen trifft. Forschende des Fraunhofer Heinrich-Hertz-Instituts HHI und der Technischen Universität Berlin haben nun eine Technik entwickelt, die erkennt, anhand welcher Kriterien KI-Systeme Entscheidungen fällen. Die neuartige Methode Spectral Relevance Analysis (SpRAy) basierend auf der Technik Layer-Wise Relevance Propagation erlaubt den Blick in die »Black Box«.

Heute gibt es kaum noch einen Bereich, in dem Künstliche Intelligenz keine Rolle spielt, sei es in der Produktion, der Werbung oder der Kommunikation. Viele Unternehmen nutzen lernende und vernetzte KI-Systeme, etwa um präzise Nachfrageprognosen anzustellen und das Kundenverhalten exakt vorherzusagen. Auf diese Weise lassen sich beispielsweise Logistikprozesse regional anpassen. Auch im Gesundheitswesen bedient man sich spezifischer KI-Tätigkeiten wie dem Anfertigen von Prognosen auf Basis von strukturierten Daten. Hier betrifft das etwa die Bilderkennung: So werden Röntgenbilder als Input in ein KI-System gegeben, der Output ist eine Diagnose. Das Erfassen von Bildinhalten ist auch beim autonomen Fahren entscheidend, wo Verkehrszeichen, Bäume, Fußgänger und Radfahrer fehlerfrei erkannt werden müssen. Und genau hier liegt die Crux: In sensiblen Anwendungsfeldern wie der medizinischen Diagnostik oder in sicherheitskritischen Bereichen müssen KI-Systeme absolut zuverlässige Problemlösungsstrategien liefern. Bislang war es jedoch nicht nachvollziehbar, wie KI-Systeme Entscheidungen treffen. Zudem basieren die Vorhersagen auf der Qualität der Input-Daten. Mit der Layer-Wise Relevance Propagation (LRP) haben Forschende am Fraunhofer-Institut für Nachrichtentechnik, Heinrich-Hertz-Institut, HHI, und der Technischen Universität Berlin nun eine Technik entwickelt, die KI-Prognosen erklärbar macht und somit unsichere Problemlösungsstrategien aufdeckt. Die Weiterentwicklung der LRP-Technologie, die sogenannte Spectral Relevance Analysis (SpRAy) identifiziert und quantifiziert ein breites Spektrum erlernten Entscheidungsverhaltens und erkennt somit auch in riesigen Datensätzen unerwünschte Entscheidungen.

Kontakt
Janis Eitner | Fraunhofer-Gesellschaft, München | Kommunikation | Telefon +49 89 1205-1333 | presse@zv.fraunhofer.de
Kathleen Schröter | Fraunhofer-Institut für Nachrichtentechnik, Heinrich-Hertz-Institut, HHI | Telefon +49 30 31002-424 | Einsteinufer 37 | 10587 Berlin | www.hhi.fraunhofer.de | kathleen.schroeter@hhi.fraunhofer.de

Transparente KI

In der Praxis identifiziert die Technik einzelne Input-Elemente, die für eine Vorhersage genutzt wurden. Wird also beispielsweise ein Gewebebild in ein KI-System eingegeben, so wird der Einfluss jedes Pixels auf das Klassifikationsergebnis quantifiziert. Die Vorhersage, wie »krebsartig« oder »nicht krebsartig« das Gewebebild ist, wird also mit der Angabe der Basis für diese Klassifikation ergänzt. »Nicht nur das Ergebnis soll korrekt sein, sondern auch der Lösungsweg. Bislang wurden KI-Systeme als Black Box angewendet. Man hat darauf vertraut, dass sie das richtige tun. Mit unserer Open-Source-Software, die die Layer-Wise Relevance Propagation einsetzt, ist es uns gelungen, die Lösungsfindung von KI-Systemen nachvollziehbar zu machen«, sagt Dr. Wojciech Samek, Leiter der Forschungsgruppe »Machine Learning« am Fraunhofer HHI. »Mit LRP visualisieren und interpretieren wir Neuronale Netze und andere Machine Learning-Modelle. Mit LRP messen wir den Einfluss jeder Eingangsvariablen für die Gesamtvorhersage und zerlegen die Entscheidungen des Klassifizierers«, ergänzt Dr. Klaus-Robert Müller, Professor für Maschinelles Lernen an der TU Berlin.

Unsichere Lösungsstrategien

Nur wer versteht, wie Neuronale Netze funktionieren, kann den Ergebnissen vertrauen. Dass KI-Systeme nicht immer sinnvolle Lösungswege finden, ergaben die Tests der Forscherteams. Beispielsweise klassifizierte ein renommiertes KI-System Bilder anhand des Kontextes. Es ordnete Fotos der Kategorie Schiff zu, wenn viel Wasser im Bild zu sehen war. Die eigentliche Aufgabe, Schiffe zu erkennen, löste es nicht, auch wenn die Mehrzahl der Bilder korrekt identifiziert war. »Zahlreiche KI-Algorithmen wenden unsichere Strategien an und kommen zu wenig sinnvollen Lösungen«, resümiert Samek das Ergebnis der Untersuchungen.

Neuronale Netze beim Denken beobachten

Die LRP-Technologie entschlüsselt die Funktionsweise von Neuronalen Netzen, und findet heraus, anhand welcher Merkmale ein Pferd als Pferd identifiziert wird und nicht als Esel oder Kuh. An jedem Knotenpunkt des Netzes erkennt sie, wie Informationen durch das Netz fließen. Somit lassen sich sogar sehr tiefe Neuronale Netze untersuchen.

Derzeit erarbeiten die Forscherteams des Fraunhofer HHI und der TU Berlin neue Algorithmen, um weitere Fragestellungen zu untersuchen und KI-Systeme noch sicherer und robuster zu gestalten. Ihre Forschungsergebnisse haben die Projektpartner in dem Fachmagazin Nature Communications veröffentlicht (siehe Link unten).

Veröffentlichung bei Nature Communication:
https://www.nature.com/articles/s41467-019-08987-4

KI, Machine Learning und Co.

Künstliche Intelligenz beschäftigt sich mit der Entwicklung von Systemen, die eigenständig Probleme lösen und analog zu menschlichen Denk- und Verhaltensmustern intelligent handeln. Aktuell werden die größten Fortschritte auf dem Gebiet des Machine Learning oder maschinellen Lernens erzielt, einem Teilgebiet der KI. Dieses beschäftigt sich mit Methoden, die aus Daten Wissen extrahieren und in den Daten enthaltene Zusammenhänge selbstständig lernen. Zurückzuführen sind die Fortschritte auf den Einsatz Künstlicher Neuronaler Netze, die auf Verbindungen zwischen mathematischen Berechnungseinheiten beruhen und im Prinzip die Neuronenstruktur des menschlichen Gehirns nachbilden. Ein Teilgebiet des Maschinellen Lernens, das Deep Learning, umfasst eine Klasse neuer Verfahren, die es ermöglichen, komplexe Künstliche Neuronale Netze anzulernen und zu trainieren. Diese Netze bestehen aus zahlreichen Ebenen, die in vielschichtigen Strukturen miteinander verknüpft sind.

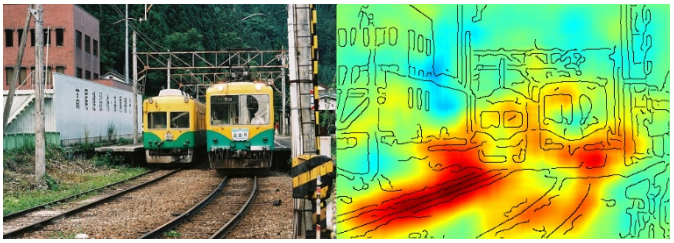

Abb. 1 Hier klassifiziert das KI-System ein Bild als Zug, da Schienen vorhanden sind.

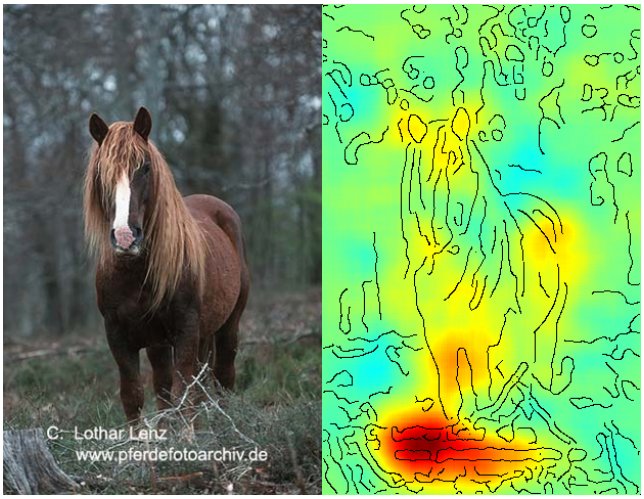

Abb. 2 Hier ordnet das KI-System das Bild anhand des Copyright-Schriftzugs der richtigen Kategorie zu. Die Lösungsstrategie ist dennoch fehlerhaft.

© Fraunhofer HHI

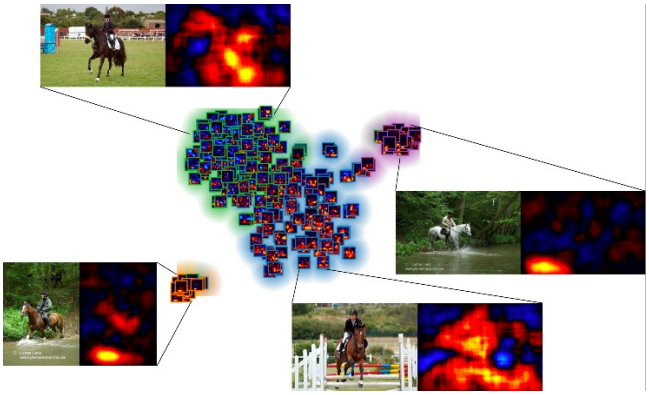

Abb. 3 Die neue Technik Spectral Relevance Analysis fasst zusammen, aufgrund welcher Kriterien KI-Systeme Entscheidungen treffen.

© Fraunhofer HHI

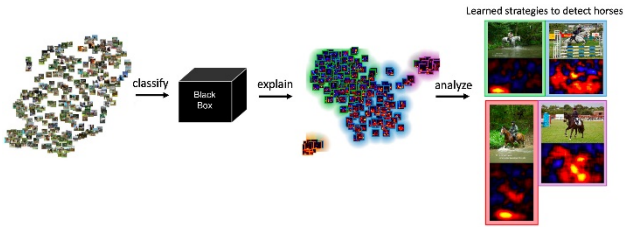

Abb. 4 Layer-wise Relevance Propagation ermöglicht den Blick in die »Black Box«.

© Fraunhofer HHI