Blick in die Black Box

Erklärbarkeit maschineller Lernverfahren

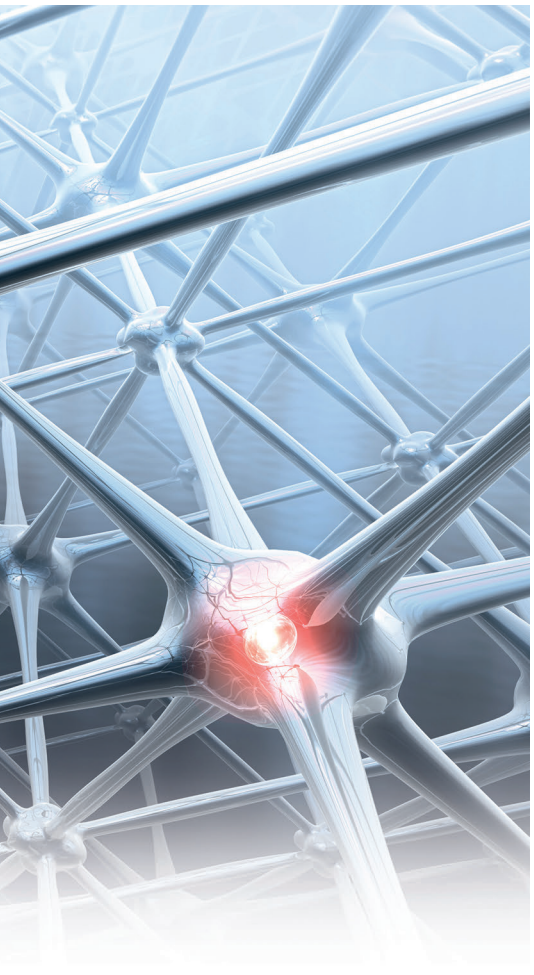
In den letzten Jahren hat das maschinelle Lernen als Teildisziplin der künstlichen Intelligenz in vielen Bereichen, wie etwa der Produktion oder Medizin, verstärkt an Bedeutung gewonnen. Immer wichtiger wird dabei das sogenannte Deep Learning, das heißt das Training tiefer künstlicher neuronaler Netze mittels großer Datensätze für eine bestimmte Aufgabe.

Nina Schaaf und Philipp Wagner

Oftmals übertreffen Modelle, die durch Deep Learning erstellt wurden, sogar den Menschen [5]. Allerdings stellen viele Verfahren des maschinellen Lernens (ML), und hierzu zählen auch die tiefen künstlichen neuronalen Netze (KNN), eine Art „Black Box" dar.

Das bedeutet, dass getroffene Entscheidungen dieser Verfahren aufgrund komplexer interner Prozesse für den Menschen – selbst für Experten – oft nicht nachvollziehbar sind. Diese mangelnde Transparenz und Nachvollziehbarkeit sind für einige Anwendungen unkritisch, beispielsweise für Empfehlungssysteme auf Online-Plattformen oder maschinelle Textübersetzungen.

Für viele andere Anwendungsfälle jedoch besteht ein berechtigtes Interesse daran, Transparenz und Erklärbarkeit zu erlangen. So werden immer wieder Fälle bekannt, in denen fehlerhafte KI-basierte Entscheidungen gravierende Auswirkungen haben können. Beispiele hierfür sind die Benachteiligung Schwarzer Bürger durch die Compas-Software [1], die Erkennung von Lungenentzündungen anhand von nicht ursächlichen Bildmarkierungen [8] oder tödliche Unfälle mit selbstfahrenden Autos [4].



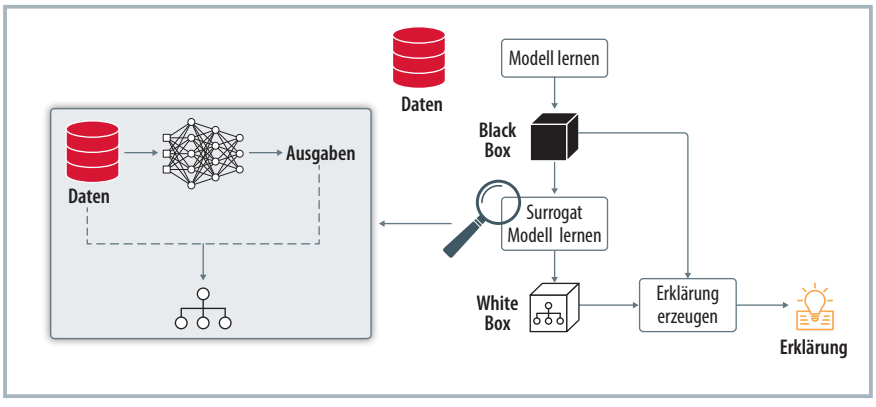


Bild 1. Von der Black Box zur Erklärung: Extraktion von White-Box-Modellen aus Black-Box-Modellen

Quelle: Nina Schaaf, Grafik: © Hanser

Black-Box-Modelle erklären: Herangehensweise

Um ML-Verfahren erklärbar(er) zu machen, gibt es mittlerweile verschiedene Möglichkeiten. Bevor Erklärungen erstellt werden können, ist es deshalb wichtig und nötig, eine individuelle Herangehensweise zu definieren, die zum eigenen Anwendungsfall und den gewünschten Zielen sowie Zielgruppen passt.

Die einfachste Möglichkeit, Erklärbarkeit algorithmisch getroffener Entscheidungen zu ermöglichen, ist, ML-Modelle zu verwenden, bei denen Erklärbarkeit ein wesentlicher Bestandteil des Designs ist. *Ante-Hoc*-Modelle – oftmals auch als „White-Box-Modelle" bezeichnet – sind so konzipiert, dass sie von Natur aus erklärbar sind. Beispiele hierfür sind die logistische Regression, regelbasierte Systeme oder Entscheidungsbäume.

Für manche Anwendungen sind diese Modelle allerdings nicht geeignet, etwa weil sie nicht ausreichend vorhersagegenau sind. Beispielsweise haben sich sogenannte Convolutional Neural Networks (CNN), eine Sonderform der neuronalen Netze, für die Bildklassifikation oder Objekterkennung mittlerweile fest etabliert. Möchte man ein solches Black-Box-Modell nachträglich erklären, spricht man von *Post-Hoc*-Erklärbarkeit.

Des Weiteren können Methoden zur Herstellung von Erklärbarkeit *modellspezifisch* sein, also nur für eine Art von ML-Modellen funktionieren, oder *modellagnostisch* und somit für verschiedene Modellarten anwendbar.

Bei der Suche nach einem geeigneten Erklärungsansatz ist zudem zu beachten, für welchen Bereich die erzeugte Erklärung gilt. *Globale* Erklärbarkeit – manchmal auch Modellerklärung genannt – setzt voraus, dass die erzeugten Erklärungen für das Modell als Ganzes gelten. Globale Erklärbarkeit ermöglicht es etwa, Einblicke in Nichtlinearitäten oder Wechselwirkungen in den Eingabedaten zu erlangen.

Im Gegensatz dazu zielt die *lokale* oder Ausgabeerklärbarkeit darauf ab, zu verstehen, warum eine einzelne oder mehrere Prognosen des untersuchten ML-Modells entstanden sind.

Zuletzt können Erklärungsansätze auch hinsichtlich der Art der verwendeten Daten unterschieden werden. Grundsätzlich können Daten in Form von Text, Bildern oder in tabellarischer Form vorliegen. Einige Erklärungsverfahren sind für alle universell verwendbar, während andere in ihrem Anwendungsbereich auf einen oder zwei Datentypen beschränkt sind [2].

Black-Box-Modell: Erklärungsansätze

Eine Möglichkeit, ein Black-Box-Modell global zu erklären, ist die Abbildung des Modells mittels eines White-Box-Modells. Hierfür wird aus dem Black-Box-Modell ein interpretierbares Stellvertretermodell – auch als Surrogat bezeichnet – extrahiert. Dieses ist dann nutzbar, um Erklärungen zu erzeugen.

Konkret funktioniert die Extraktion eines Surrogats (zum Beispiel eines Entscheidungsbaums) aus einem Black-Box-Modell wie in Bild 1 dargestellt. Zuerst wird unter Zuhilfenahme des Black-Box-Modells für alle Eingabedaten, bestehend aus Merkmalen und einer Sollausgabe, eine Vorhersage berechnet (links). Im Anschluss werden die Eingabemerkmale zusammen mit der ›››

INFORMATION & SERVICE

AUTOREN

Nina Schaaf ist wissenschaftliche Mitarbeiterin am Fraunhofer-Institut für Produktionstechnik und Automatisierung IPA in Stuttgart. Dort arbeitet sie seit Abschluss ihres Studiums der Medieninformatik am Zentrum für Cyber Cognitive Intelligence (CCI). Ihr besonderes Interesse gilt dem maschinellen Lernen im Allgemeinen sowie dem Forschungsfeld explainable Artifical Intelligence (xAI).

Philipp Wagner arbeitet als wissenschaftlicher Mitarbeiter im CCI am Fraunhofer IPA. Nach seinem abgeschlossenen Physikstudium richtete er sein Forschungsinteresse auf die Themenfelder xAI sowie Reinforcement Learning.

KONTAKT

Nina Schaaf
T 0711 970-1971
nina.schaaf@ipa.fraunhofer.de

Vorhersage des Black-Box-Modells dazu verwendet, das Stellvertretermodell zu trainieren (rechts). Dieser Ansatz eignet sich besonders gut für die Arbeit mit tabellarischen Daten.

Für Bilddaten bietet sich hingegen beispielsweise die Verwendung von Konzepten an, die tiefe neuronale Netze erklären [4]. Hierbei werden automatisch visuelle Konzepte identifiziert, die für das Black-Box-Modell besonders wichtig waren. Zum Beispiel könnte ein Reifen-Konzept eine hohe Relevanz für die Erkennung von Autos haben. Dies bedeutet, dass Autos – unabhängig von ihren sonstigen Merkmalen wie Farbe oder Form – vor allem anhand der Reifen erkannt werden. Das Reifen-Konzept bildet also ein zentrales Erkennungsmerkmal für diese Klasse.

Während globale Erklärbarkeit vor allem für Modellentwickler wichtig ist, sind für den Endanwender häufig nur die Erklärungen einzelner Entscheidungen relevant. Setzt eine Bank beispielsweise ein Kreditvergabesystem ein, das auf ML-Methoden aufbaut, muss dieses System einem Kunden erklären können, welche Punkte seines Kreditantrages zu einer Ablehnung geführt haben. Denn sind Nutzer von einer solchen algorithmisch getroffenen Entscheidung betroffen, so müssen laut Datenschutz-Grundverordnung (DSGVO) dem Anwender aussagekräftige Informationen über die involvierte Logik bereitgestellt werden [3].

Für lokale Erklärungen wird meist die Wichtigkeit der Eingabeattribute angegeben, die zu einer bestimmten Entscheidung geführt haben. Dieses Konzept lässt sich auch auf Probleme aus der Bildverarbeitung übertragen. Hier wird dann beispielsweise jedem Pixel ein positiver oder negativer Wert zugeordnet, je nachdem, ob dieser die letztendliche Entscheidung des ML-Modells begünstigt hat oder nicht (Bild 2).

Eine populäre Methode, die für tabellarische Daten, Bilddaten und Texte angewandt werden kann ist Lime [7]. Lime verfolgt eine perturbationsbasierte Erklärungsstrategie. Das bedeutet, dass Teile der zu erklärenden Dateninstanz (zum Beispiel ein Bild) gezielt verändert und die Auswirkung dieser Veränderung auf die Modellvorhersage untersucht werden. Die Anwendungsfelder für lokale Erklärungsansätze sind vielfältig und reichen von medizinischen Anwendungen über die Qualitätsanalyse bis hin zum autonomen Fahren.

Fazit

Bedingt durch die enormen Potenziale, die Künstliche Intelligenz und allem voran Deep Learning bieten, eröffnen sich zahlreiche neue Anwendungsfelder. Gerade für kritische Einsatzgebiete, in denen automatisiert getroffene Entscheidungen nachvollziehbar sein müssen, können Erklärungstechniken einen wichtigen Beitrag leisten, um KI-Systeme erfolgreich einzusetzen.

Derzeit arbeiten noch verstärkt Entwickler oder Domänenexperten mit den Erklärungen. In Zukunft werden jedoch auch Anwender ohne einen vertieften technischen Hintergrund die Nutzer sein. Deshalb müssen Forschungsbestrebungen zur zielgruppengerichteten Darstellung von Erklärungen intensiviert werden.

Ein weiterer vielversprechender Aspekt ist es, Erklärbarkeitsziele schon bei der Entwicklung der Modelle zu berücksichtigen. So könnten Modelle schon während des Erstellungsprozesses dahingehend optimiert werden, dass sie anschließend leichter erklärbar sind.

Nicht zuletzt existiert noch ein großer Forschungsbedarf bei der Entwicklung neuer Verfahren zur globalen Erklärung von CNNs oder auch von rekurrenten neuronalen Netzen, die für Spracherkennung oder Zeitreihendaten eingesetzt werden. Hier steckt das Forschungsfeld in den Kinderschuhen, und es existieren noch keine Out-of-the-Box-Lösungen. ■

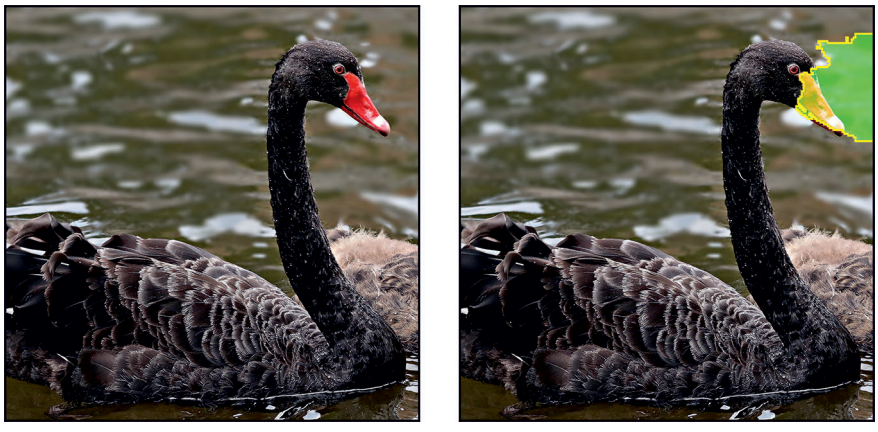


Bild 2. Erklärung der Klasse „Schwarzer Schwan" für das vorliegende Bild mithilfe der Methode Lime. Das Hauptmerkmal für diese Klasse ist der rote Schnabel. Quelle: www.pixabay.com

INFORMATION & SERVICE

LITERATUR

1 Angwin, Julia; Larson, Jeff; Mattu, Surya; Kirchner, Lauren: Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks. 2016. Verfügbar unter: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. Zugriff am: 16.11.2020.

2 Burkart, Nadia; Huber, Marco F.: A Survey on the Explainability of Supervised Machine Learning. arXiv:2011.07876v1, 2020.

3 (EU) 2016/679: GDPR Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation).

4 Ghorbani, Amirata; Wexler, James; Zou, James; Kim, Been: Towards Automatic Concept-based Explanations. arXiv:1902.03129, 2019.

5 Levin, Sam; Carrie, Julia: Self-driving Uber kills Arizona woman in first fatal crash involving pedestrian. 2018. Verfügbar unter: https://www.theguardian.com/technology/2018/mar/19/uber-self-driving-car-kills-woman-arizona-tempe. Zugriff am: 16.11.2020.

6 OpenAI, 2019. OpenAI Five Defeats Dota 2 World Champions. Verfügbar unter: https://openai.com/blog/openai-five-defeats-dota-2-world-champions/. Zugriff am: 16.11.2020.

7 Ribeiro, Marco Tulio; Singh, Sameer; Guestrin, Carlos: „Why Should I Trust You?" Explaining the Predictions of Any Classifier. arXiv:1602.04938, 2016.

8 Zech, John R.: What are radiological deep learning models actually learning? 2018. Verfügbar unter: https://medium.com/@jrzech/what-are-radiological-deep-learning-models-actually-learning-f97a546c5b98 Zugriff am: 23.11.2020.