

1 *Approximation eines neuronalen Netzes mit Hilfe eines Entscheidungsbaums.*

xAI – ERKLÄRBARKEIT MASCHINELLER LERNVERFAHREN

Marktanforderungen

Verfahren des Maschinellen Lernens kommen immer häufiger in verschiedenen Anwendungsbereichen der Produktion, Medizin oder dem Dienstleistungswesen zum Einsatz. Beispielsweise kann über Bildverarbeitung oder Sensordatenanalyse Ausschuss erkannt und frühzeitig aus dem Prozess genommen werden. Auch für die Programmierung von Robotern werden vermehrt tiefe Neuronale Netze erprobt. Es gibt jedoch viele Szenarien, in denen nicht hochgenaue Vorhersagen alleine von größter Bedeutung sind und stattdessen Vertrauen, Akzeptanz oder Konformität mit Regularien entscheidend sind. Kritische Entscheidungen müssen hier von Erklärungen begleitet werden, sodass Nutzer die Ergebnisse oder das generelle Verhalten des Algorithmus nachvollziehen können.

Durch die Schaffung von Erklärbarkeit kann einerseits die korrekte Funktionsweise der Modelle überprüft, aber auch mögliche

Diskrepanzen zwischen menschlichen Entscheidungen und algorithmisch getroffenen Entscheidungen untersucht werden. So lässt sich beispielsweise die Frage, wieso eine bestimmte Stellgröße für einen Regler ausgegeben wurde, beantworten. Auch für sicherheitskritische Bereiche, wie das autonome Fahren oder Mensch-Roboter-Kollaborationen, können Erklärungen einen erheblichen Mehrwert bieten. Mithilfe von Erklärungen für die Entscheidungen eines Modells können Experten ihr Verständnis für das Modell verbessern und Risiken bewerten.

Fraunhofer-Institut für Produktionstechnik und Automatisierung IPA

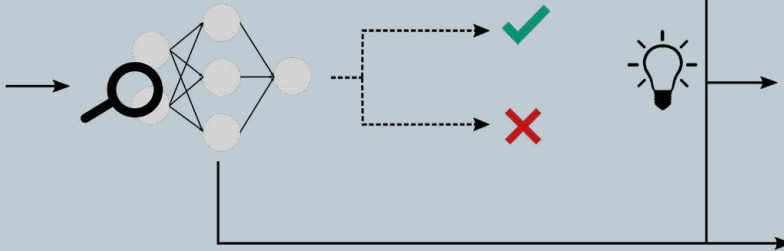
Nobelstraße 12
70569 Stuttgart

Ansprechpartner

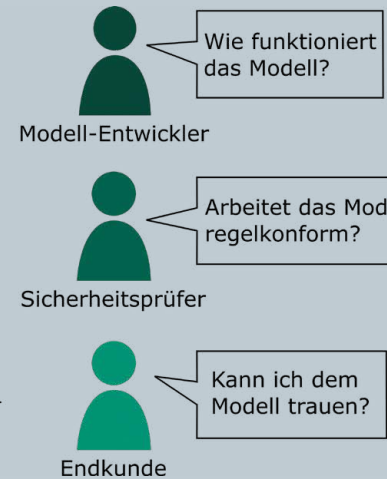
Prof. Dr.-Ing. Marco Huber
Telefon +49 711 907-1960
marco.huber@ipa.fraunhofer.de

www.ipa.fraunhofer.de

Eingabe



Ausgabe



2

Unsere Lösung

Grundsätzlich gibt es zwei Ansätze, Erklärbarkeit Maschinelles Lernverfahren herzustellen. Bei der Modellerklärbarkeit (global) steht das Verständnis des Modells als Ganzes im Fokus: Das Ziel ist, eine möglichst gute Nachvollziehbarkeit der inneren Entscheidungswege eines Black Box-Modells zu liefern. Im Gegensatz dazu werden mittels lokaler Erklärmethoden einzelne Entscheidungen erklärt.

Das Fraunhofer IPA bietet für Ihre bestehenden Machine Learning-Modelle (ML-Modelle) die Herstellung von Erklärbarkeit durch Erklärungs-generierung – sowohl in lokaler, als auch globaler Form. Auch im Zuge von Neuentwicklungen bietet das Fraunhofer IPA die Möglichkeit, Erklärbarkeit von Beginn an als elementaren Bestandteil des Prozesses und Endproduktes zu berücksichtigen.

Der Stand der Technik im Forschungsfeld »Explainable AI« umfasst eine Vielzahl verschiedener Methoden. Da nicht jede Methode für jeden Anwendungsfall gleich gut geeignet ist, ist die Wahl der optimalen Methode zeit- und rechenintensiv. Vor diesem Hintergrund entwickelt das Fraunhofer IPA zudem eine Software-Toolbox, die existierende Erklärmethoden in Einklang bringt. In diesem Zusammenhang werden auch eigene, am Fraunhofer IPA entwickelte, Verfahren integriert. Unter Einsatz dieser universalen Toolbox kann ein umfassendes Modellverständnis mit Hilfe einer raschen Erklärungserzeugung sowie einer gegenüberstellenden Betrachtung verschiedener Techniken ermöglicht werden.

Ihr Nutzen

Für Sie als Anwender ergeben sich folgende Vorteile:

Modellvalidierung

Überprüfung der erwartungskonformen Funktionsweise Ihrer eingesetzten ML-Modelle. Besonders relevant ist dieser Aspekt für sicherheitskritische Anwendungen, die beispielsweise von einer internen oder externen Prüfstelle abgenommen werden müssen.

Modell-Debugging

Werden Fehlentscheidungen des Modells festgestellt, kann für diese detailliert untersucht werden, auf Basis welcher Merkmale die Entscheidung getroffen wurde (siehe Abbildung 1). Durch das Debugging können Ursachen für Fehlfunktionen des Modells identifiziert und anschließend behoben werden.

Akzeptanz und Vertrauen

Ein fehlendes Verständnis für die Entscheidungsfindung hochkomplexer ML-Algorithmen kann oftmals eine hohe Einstiegshürde für einen Einsatz selbiger Algorithmen sein. Erklärungen für algorithmisch getroffene Entscheidungen können das Vertrauen der Anwender in die Systeme stärken und zu einer höheren Akzeptanz führen.

Erkenntnisgewinn

Die Untersuchung des gelernten ML-Modells ermöglicht das Aufdecken allgemeiner Zusammenhänge, beispielsweise welche Eingabedaten besondere Bedeutung haben und wie einzelne Eingabedaten miteinander interagieren.

2 **Erklärbarkeit maschineller Lernverfahren ist für verschiedene Interessensgruppen von Bedeutung.**