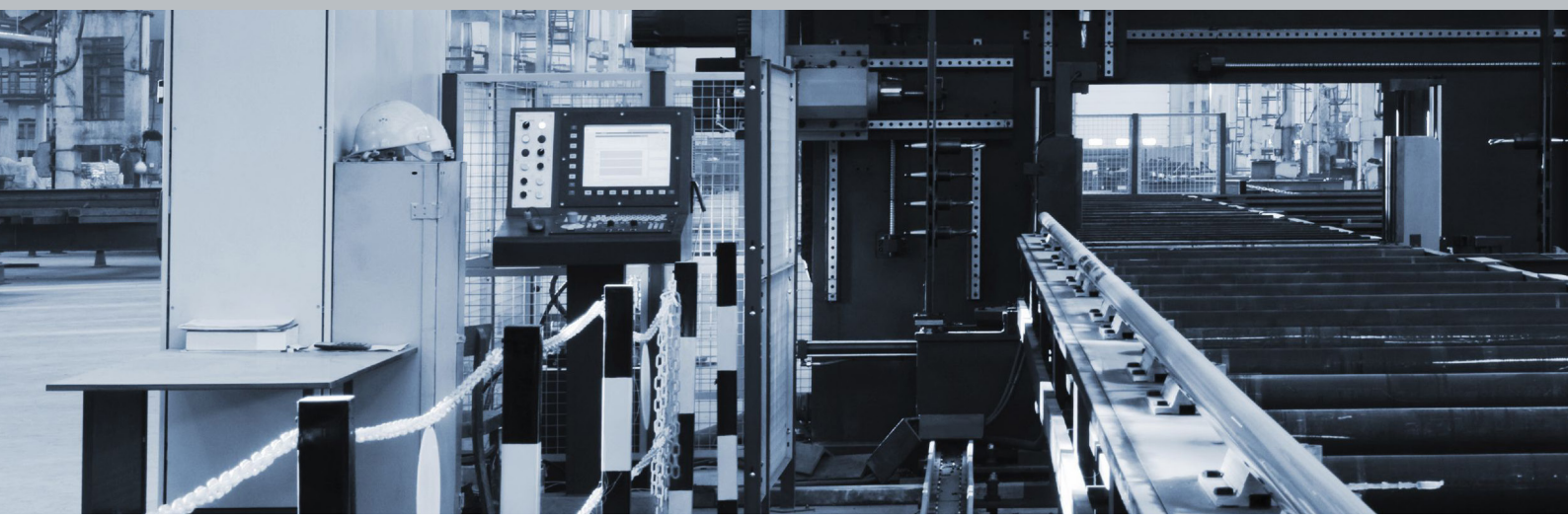
KI Fortschrittszentrum LERNENDE SYSTEME

EIN KI-QUICK-CHECK DES KI-FORTSCHRITTSZENTRUMS



KONTAKT

Fraunhofer IPA

Fraunhofer-Institut für Produktionstechnik und Automatisierung IPA

Marcel Albus
marcel.albus@ipa.fraunhofer.de

IN ZUSAMMENARBEIT MIT

SICK
Sensor Intelligence.

SICK AG

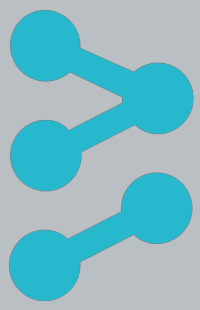
EVIDENCE-BASED SAFETY ASSURANCE OF AI METHODS

Ausgangssituation

Vor der industriellen Marktreife einer neuen Technologie ist es nötig, überzeugend darstellen zu können, dass die Anwendung der Technologie sicher ist. Dies kann durch Assurance Cases geschehen, welche mithilfe von Metriken und Nachweisen eine Argumentationsstruktur aufbauen, um die Sicherheit einer Applikation zu gewährleisten. Das Vertrauen in einen solchen Nachweis hängt stark von den eingesetzten Beweisen und Metriken ab. Eine allgemein akzeptierte Argumentationsstruktur für einen solchen Nachweis im Bereich der Künstlichen Intelligenz existiert aktuell noch nicht, unter anderem ist unklar, welche Beweise und Metriken hierfür eingesetzt werden können.

Beispielsweise kann ein akzeptables Restrisiko für den Einsatz von Methoden der Künstlichen Intelligenz (z.B. Deep Neural Networks, Bayesian Networks, Support Vector Machines) aktuell in der Sicherheitstechnik nicht nachgewiesen werden. Dies hemmt den Einsatz dieser vielversprechenden Technologie in der Sicherheitstechnik.

EVIDENCE-BASED SAFETY ASSURANCE OF AI METHODS



EIN KI-QUICK-CHECK DES KI-FORTSCHRITTSZENTRUMS

Lösungsidee

Die Beweisführungsmethode der Goal Structuring Notation (GSN) wurde für eine Förderbandapplikation untersucht. Die Förderbandanwendung wird mithilfe zweier Laserscanner überwacht, welche die sicherheitskritische Abschaltung im Falle einer Person auf dem Förderband gewährleisten. Grundlage hierfür war die Beweisführung für einen Automobil-Anwendungsfall nach Gauerhof 2018[1]. Die GSN wurde für ihre Eignung im stationären Bereich angepasst und um neue Zusammenhänge erweitert.
Hierfür wurde die GSN in drei Kategorien unterteilt: Zum einen in die Reduktion von Risiken durch fehlende Spezifikationen, die Reduktion von semantischen Lücken sowie die Minimierung von Risiken durch eine deduktive Lücke. Der Fokus lag hierbei auf letzterer Kategorie, der Minimierung der deduktiven Lücke, wobei die Extraktion von Methoden und Metriken das Ziel war. Mehrere Subkategorien wurden hierfür identifiziert und untersucht:

1. Der Anwendungsfall erreicht die nötige Klassifizierungsleistung
2. Das Modell ist ausreichend robust
3. Unterschiede in der Trainingsplattform und Zielplattform verletzten keine Sicherheitsanforderungen
4. Essenzielle Einflüsse auf das Modell sind ausreichend verstanden

1 Gauerhof, Lydia, Peter Munk, and Simon Burton. „Structuring validation targets of a machine learning function applied to automated driving." International Conference on Computer Safety, Reliability, and Security. Springer, Cham, 2018.

Nutzen

Die Ergebnisse des Projektes ermöglichen einen weiteren Schritt in Richtung der Verwertbarkeit von Neuronalen Netzen in sicherheitskritischen Anwendungen. Die Argumentationsstruktur könnte ein Grundstein für den Einsatz von Methoden der Künstlichen Intelligenz im Rahmen von sicherheitskritischen Anwendungen legen, wodurch sowohl neue Anwendungen erschlossen als auch bestehende optimiert werden können.

Im Sinne einer Argumentationsstruktur kann eine solche Herangehensweise in beliebigen Projekten wiederverwendet werden. Zusätzlich bringen die entwickelten Methoden zur Erzeugung von Beweisen über Qualitätsattribute von Komponenten der KI einen Mehrwert bei der Qualitätssicherung von KI-Methoden.

Umsetzung der KI-Applikation

Es wurden unterstützende Methoden für eine Argumentationsstruktur untersucht- und hinsichtlich ihrer Eignung bewertet.

Die identifizierten Methoden sind unter anderem die formale Verifikation mittels mathematischer Methoden und dadurch gewonnenen Garantien sowie Klassifikationsgenauigkeit oder Confusion Matrix (damit einhergehend der Recall und die Precision). Zusätzlich wurden verschiedene Methoden des AutoML untersucht, um Over- bzw. Underfitting zu reduzieren. Um Beweise für die Robustheit des Modells zu generieren, wurde die Metrik der Condition Number untersucht sowie Ansätze zum Überprüfen der local und global robustness gegen adversarial attacks. Bayesian Neural Networks bieten zusätzlich noch die Metrik der Unsicherheit. Um die Unterschiede in der dritten Subkategorien (siehe Lösungsideen) zu minimieren, wurden Methoden der domain adaptation sowie domain randomization untersucht. Die essenziellen Einflüsse auf das Modell können mithilfe von Darstellungstransformierungen untersucht werden, zum Beispiel in Entscheidungsbäumen. Saliency Maps bieten eine Übersicht zu modellrelevanten Eigenschaften auf Basis der Eingangsdaten und damit weitere Erkenntnisse über das Modell.

Von den dargestellten Methoden wurde die formale Verifikation als stärkster Beweis aufgrund ihrer mathematischen Eindeutigkeit favorisiert und soll weiter untersucht werden.

Fortschrittszentrum LERNENDE SYSTEME

EIN KI-QUICK-CHECK DES KI-FORTSCHRITTSZENTRUMS

Fraunhofer IAO

Fraunhofer-Institut für Arbeitswirtschaft und Organisation IAO

Fraunhofer IPA

Fraunhofer-Institut für Produktionstechnik und Automatisierung IPA

Kooperationspartner:

CyberValley

Gefördert durch:

Baden-Württemberg
MINISTERIUM FÜR WIRTSCHAFT. ARBEIT UND WOHNUNGSBAU

Ansprechpartner:

Dr. Matthias Peissner
Telefon +49 711 970-2311
matthias.peissner@iao.fraunhofer.de

Prof. Dr. Marco Huber
Telefon +49 711 970-1960
marco.huber@ipa.fraunhofer.de

www.ki-fortschrittszentrum.de

ÜBER DAS KI-FORTSCHRITTSZENTRUM »LERNENDE SYSTEME«

Das KI-Fortschrittszentrum »Lernende Systeme« unterstützt Firmen dabei, die wirtschaftlichen Chancen der Künstlichen Intelligenz und insbesondere des Maschinellen Lernens für sich zu nutzen. In anwendungsnahen Forschungsprojekten und in direkter Kooperation mit Industrieunternehmen arbeiten die Stuttgarter Fraunhofer-Institute für Arbeitswirtschaft und Organisation IAO sowie für Produktionstechnik und Automatisierung IPA daran, Technologien aus der KI-Spitzenforschung in die breite Anwendung der produzierenden Industrie und der Dienstleistungswirtschaft zu bringen. Finanzielle Förderung erhält das Zentrum vom Ministerium für Wirtschaft, Arbeit und Wohnungsbau Baden-Württemberg.

Europas größte Forschungskooperation auf dem Gebiet der KI

Das KI-Forschungszentrum ist Forschungspartner des Cyber Valley, einem Konsortium aus den renommierten Universitäten Tübingen und Stuttgart, dem Max-Planck-Institut für intelligente Systeme und einigen führenden Industrieunternehmen. In gemeinsamen Forschungslabors werden Grundlagenforschung und anwendungsorientierte Entwicklung zu aktuellen wie auch zukünftigen Bedarfen behandelt und vorangetrieben.

Menschzentrierte KI

Alle Aktivitäten des Zentrums verfolgen das Ziel, eine menschzentrierte KI zu entwickeln, der die Menschen vertrauen und die sie akzeptieren. Nur wenn Menschen mit neuen Technologien intuitiv interagieren und vertrauensvoll zusammenarbeiten, kann ihr Potenzial optimal ausgeschöpft werden. Daher konzentrieren sich die Forschungsaktivitäten unter anderem auf die Themen Erklärbarkeit, Datenschutz, Sicherheit und Robustheit von KI-Technologien.