KI Fortschrittszentrum
LERNENDE SYSTEME

EIN EXPLORING PROJECT DES KI-FORTSCHRITTSZENTRUMS

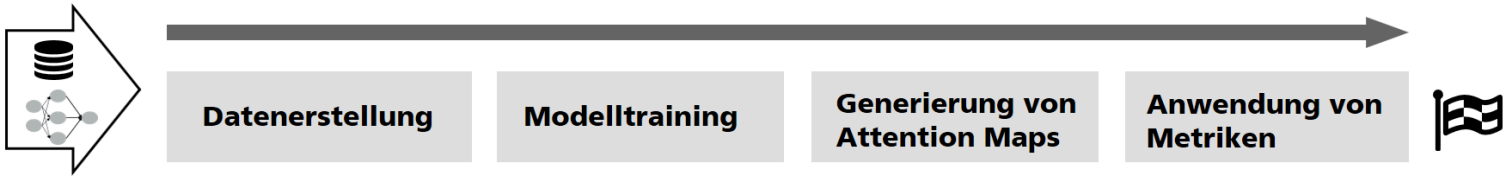

Abbildung 1: Entwicklungsphasen des Projekts.

EVALUATION VON VISUALISIERUNGSVERFAHREN DER AUFMERKSAMKEIT NEURONALER NETZE ZUR DETEKTION VON DATENVERZERRUNGEN UND -MANIPULATION FÜR DEN EDGE-EINSATZ

KONTAKT

Fraunhofer IPA

Fraunhofer-Institut für Produktionstechnik und Automatisierung IPA

Andreas Frommknecht
andreas.frommknecht@ipa.fraunhofer.de

IN ZUSAMMENARBEIT MIT

IDS Imaging Development System GmbH

Ausgangssituation

Sogenannte Aufmerksamkeitskarten (engl. Attention Maps) ermöglichen die visuelle Hervorhebung von Bildregionen, die für die Entscheidung eines neuronalen Netzes von Bedeutung waren. Diese intuitive Visualisierung soll es ermöglichen, kritische Entscheidungen leichter nachzuvollziehen, um damit die Akzeptanz neuronaler Netze im industriellen Umfeld zu erhöhen.

IDS hat eine Trainingsplattform entwickelt, die es Fachexperten ermöglicht, durch Hochladen eigener Bilddaten neuronale Netze zu trainieren. Die Plattform soll unter anderem die Anzeige von Aufmerksamkeitskarten bereitstellen. Für die Erstellung der Aufmerksamkeitskarten hat IDS einen eigenen, bereits als Prototyp verfügbaren Hochleistungsalgorithmus zum Einsatz auf ihrer Edge-Kamera IDS NXT rio entwickelt. Der Algorithmus hat den Anspruch, Aufmerksamkeitskarten unter Berücksichtigung des Ressourcenverbrauchs in Echtzeit zu berechnen.

Erste Ergebnisse des Quick-Checks haben gezeigt, dass es möglich ist, mit Aufmerksamkeitskarten Hinweise auf Verzerrungen im Datensatz und Datenmanipulationen zu erhalten. Ziel dieses Exploring Projects war es daher, die Ergebnisse aus dem Quick-Check mit Hilfe eines systematischen Testverfahrens zu bestätigen. Zu diesem Zweck wurden speziell Verzerrungen und Manipulationen von Daten in Datensätze eingebracht und Metriken zur Beurteilung der Genauigkeit ihrer Erkennung mittels Aufmerksamkeitskarten festgelegt. Besonderes Augenmerk wurde auch auf die Wahl der Methoden zur Erstellung von Aufmerksamkeitskarten sowie auf die Metriken,

EVALUATION VON VISUALISIERUNGSVERFAHREN DER AUFMERKSAMKEIT NEURONALER NETZE ZUR DETEKTION VON DATENVERZERRUNGEN UND -MANIPULATION FÜR DEN EDGE-EINSATZ

EIN EXPLORING PROJECT DES KI-FORTSCHRITTSZENTRUMS

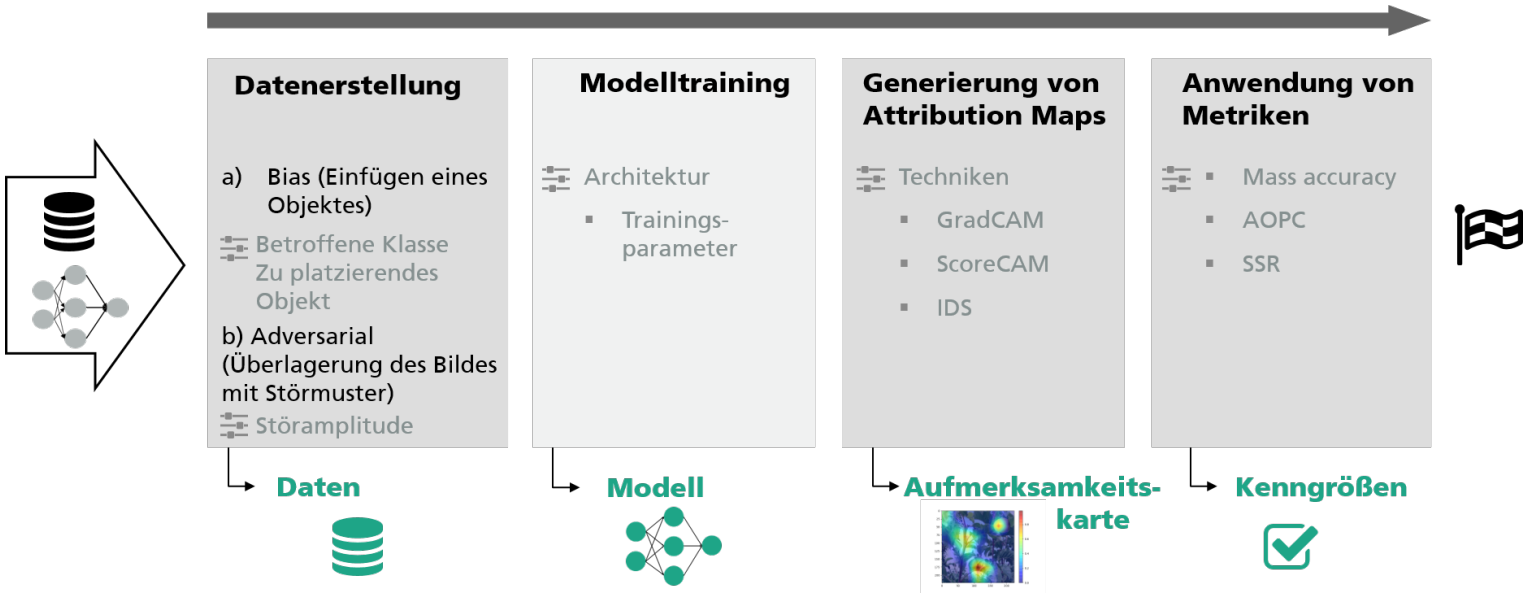

Abbildung 2: Verarbeitungsschritte zur Bewertung von Aufmerksamkeitskarten.

die als Grundlage für den Vergleich der Ergebnisse dienen, gelegt.

Lösungsidee durch KI

Um die Leistungsfähigkeit der intern von IDS entwickelten Methode zur Visualisierung von Aufmerksamkeitskarten zu validieren, sollte mit diesem Projekt ein reproduzierbarer und vergleichbarer Test zur Validierung der Genauigkeit verschiedener Methoden der Aufmerksamkeitskartenextraktion bei verzerrten Datensätzen und Adversarial Attacks entwickelt werden. Ein solcher Test bildet das Fundament zur weiteren gezielten Optimierung von Visualisierungsverfahren.

Nutzen

Die Untersuchung der Aufmerksamkeitskarten eines neuronalen Netzes sorgt für eine größere Transparenz im Prozess der Bewertung der Leistung und Validität eines neuronalen Netzes und gewährleistet zudem ein höheres Maß an Robustheit. Je schneller Fehler nachvollzogen und Verzerrungen im Datensatz gefunden werden können, desto gezielter kann der Datensatz verbessert werden. Ein reproduzierbarer und vergleichbarer Test, welcher die Präzision verschiedener Verfahren zur Aufmerksamkeitsvisualisierung validiert, wird insbesondere im industriellen Umfeld die Akzeptanz dieser Verfahren erhöhen und damit auch die Akzeptanz von neuronalen Netzen. Unabhängig von der Art der untersuchten Anwendungsfälle können die vorgeschlagenen Methoden auf eine Vielzahl anderer Kontexte übertragen werden. Dadurch wird für unterschiedliche Anwendungsfälle mit geringem Aufwand ein höherer Grad der Nachvollziehbarkeit der trainierten Algorithmen ermöglicht.

Umsetzung der KI-Applikation

Das Projekt war, wie in Abbildung 1 und 2 dargestellt, aufgeteilt in vier Phasen: die Erstellung der Datensätze, das Training der Modelle, die Erstellung der Aufmerksamkeitskarten und die Qualitätsbewertung. Es wurden zwei beispielhafte Anwendungskontexte definiert, die es erlauben, verzerrte Datensätze und Adversarial Attacks künstlich nachzubilden. Beide Anwendungsbeispiele sind dem Kontext der Lebensmittelindustrie zuzuordnen. Im ersten Fall wurden die verzerrten Daten durch entsprechendes Einbringen von Aufklebern auf eine Frucht erzeugt. Im zweiten Anwendungskontext wurden stattdessen verzerrte Daten durch eine

EVALUATION VON VISUALISIERUNGSVERFAHREN DER AUFMERKSAMKEIT NEURONALER NETZE ZUR DETEKTION VON DATENVERZERRUNGEN UND -MANIPULATION FÜR DEN EDGE-EINSATZ

EIN EXPLORING PROJECT DES KI-FORTSCHRITTSZENTRUMS

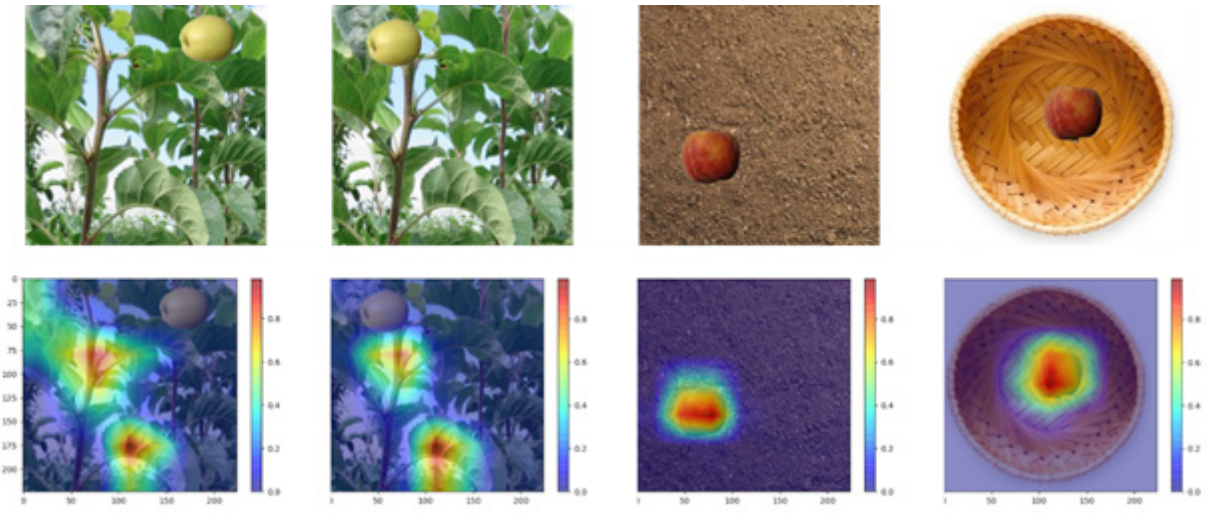

Abbildung 3: Bias: Entscheidungsgrundlage für Klasse „Apfel“ ist der Hintergrund (Apfelbaum), nicht die Frucht selbst (links). Für die Klasse „Pfirsich“ hingegen sind Merkmale der Frucht entscheidend (rechts).

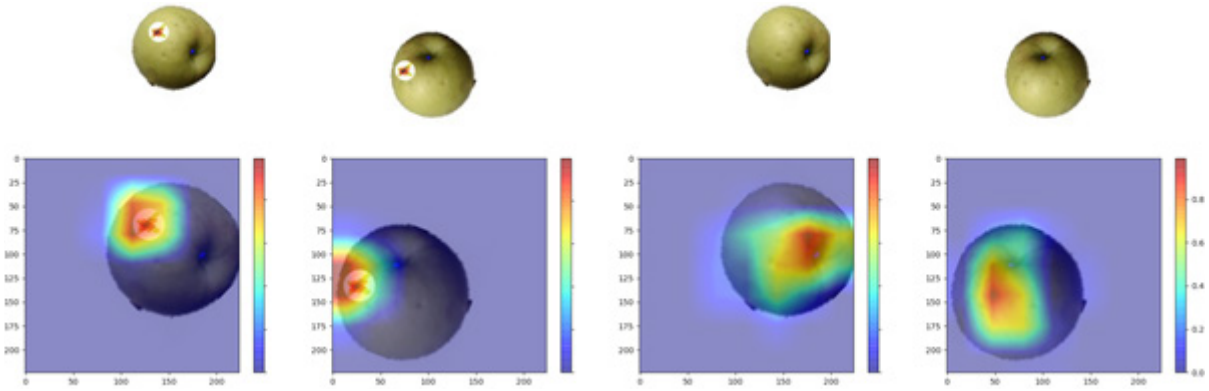

Abbildung 4: Bias: Entscheidungsgrundlage für Klasse „Apfel“ ist der Sticker (links), nicht die Frucht selbst (rechts).

geeignete Hintergrundvariation erzeugt. Für die Erzeugung von Adversarial Attacks wurden zudem die zuvor generierten Bilddaten mit einem kaum wahrnehmbaren Störmuster überlagert. Für jeden gewählten Anwendungskontext wurden die Datensätze dynamisch erstellt, indem eine verzerrte (biased) und eine unverzerrte Version der Daten erstellt wurde. Die entwickelte Methode ermöglicht, mit minimalen Modifikationen, die Übertragbarkeit auf andere Anwendungskontexte. Die erzeugten Datensätze wurden im Anschluss sowohl für das Training neuronaler Netze als auch für die Erstellung verschiedener Aufmerksamkeitskarten verwendet. Unter Zuhilfenahme unterschiedlicher Metriken wurden zuletzt eine Analyse und ein Vergleich der generierten Aufmerksamkeitskarten durchgeführt.

Die Abbildungen 3 und 4 zeigen eine Auswahl von Aufmerksamkeitskarten für die untersuchten Anwendungsfälle zur Erkennung von Verzerrungen im Datensatz.

Fazit

Durch die Entwicklung standardisierter Verfahren zur Datenverzerrung und -manipulation wird eine Basis für die Vergleichbarkeit unterschiedlicher Aufmerksamkeitskarten gelegt. Die eingeführten Metriken erlauben eine quantitative Bewertung und stellen einen Ausgangspunkt für die gezielte Optimierung der Verfahren dar. In Zukunft kann so eine Referenzdatenbank von Aufmerksamkeitskarten erzeugt werden, die es Anwendern ermöglicht, nach eindeutigen Kriterien den Algorithmus zu wählen, der sich am besten für das individuelle Szenario eignet.

Fortschrittszentrum LERNENDE SYSTEME

EIN EXPLORING PROJECT 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.