Sonderdruck aus German Testing Magazin 01/2020

»Die Gefahren lauern an allen Ecken«

Die Paradigmen im Testen von Software ändern sich zunehmend durch den Einsatz von Künstlicher Intelligenz. Das Beispiel autonomes Fahren, bei welchem unterschiedliche neuronale Netze getestet werden, zeigt die Bedeutung von Datenqualität und Corner Cases. Als QualityMinds sind wir im Förderprojekt „KI-Absicherung" mit OEMs, Zulieferern sowie Forschungsinstituten tätig, um eine Absicherungsmethodik zu erarbeiten. In diesem Artikel stellen wir konkrete Beispiele von KI-spezifischen Corner Cases, eine Taxonomie als Teil dieser Methodik und die wichtigsten Erkenntnisse für das Testen von Computer Vision vor.

Mit Daten trainierte Algorithmen, häufig als „Künstliche Intelligenz" (KI) bezeichnet, sind scheinbar zu Außergewöhnlichem in der Lage und werden dafür oft gefeiert. Jedoch weisen sie ein inhärentes Merkmal auf: Sie machen Fehler.

In E-Commerce-Anwendungen mag es akzeptabel sein, dass ein Algorithmus in 99 Prozent der Fälle ein richtiges Ergebnis produziert. Das ändert sich jedoch in Anwendungsbereichen wie dem autonomen Fahren, in denen Menschen ihre Handlungsmacht abgeben. Hier muss die Fehlerquote nahe Null sein – mit enormen Anforderungen an die Qualitätssicherung.

Da sich etablierte Absicherungsprozesse nicht ohne Weiteres auf maschinelle Lernverfahren übertragen lassen, gilt es, eine stringente und nachweisbare Argumentationskette für die Absicherung und Freigabe von KI im Kontext hochautomatisierten Fahrens aufzubauen – und diese mit messbaren Leistungs- und Sicherheitsmaßnahmen zu untermauern. Doch wie kann die Qualität solcher Systeme getestet und sichergestellt werden?

Willkommen im Neuland!

Das war unser erster Gedanke, als wir die Ausschreibungsunterlagen für das Förderprojekt „KI-Absicherung für autonomes Fahren" des Bundesministeriums für Wirtschaft und Energie (BMWi) erhielten. Das Vorhaben, welches seit Juli 2019 für 3 Jahren läuft, ist ein wichtiger Teil der KI-Strategie der Bundesregierung. Sie will Deutschland langfristig für neue Schlüsseltechnologien aufstellen und die Marktführerschaft der Autoindustrie im Hinblick auf das automatisierte Fahren sichern. Deshalb investiert sie 19,2 Mio. Euro in die Förderung des Projekts; das Gesamtbudget beläuft sich auf knapp 41 Mio. Euro [KIAB].

Ein autonomes Fahrzeug muss seine Umwelt wahrnehmen und adäquat auf diese reagieren. Es muss die Bewegungen anderer Verkehrsteilnehmer, etwa von Fußgängern, interpretieren und daraus künftiges Verhalten ableiten. Unter einfachen Bedingungen wurden bereits Tausende von Kilometern problemlos navigiert. Doch die Schwierigkeit beim Autofahren liegt nicht in harmlosen Regelsituationen, sondern in den vielen Ausnahmen und Abweichungen von der Norm. Auf diesen Randbedingungen, den sogenannten *Corner Cases*, und damit der Qualität von Trainings- und Testdaten, liegt unser Schwerpunkt im Projekt: Wie lassen sich trainierte Algorithmen systematisch auf ihr Verhalten in Ausnahmesituationen überprüfen? Worauf kommt es in solchen Vorhaben an?

Drei Kategorien sind aus unserer Sicht große Herausforderungen beim Testen von KI:

- Datenqualität von Trainings- und Testdaten,
- Metriken für Testabdeckung und
- Corner Cases.

Ein großes Problem beim Anlernen von KI-Systemen sind häufig die *Daten*. Probleme wie Vorurteile oder Überrepräsentationen in den Daten werden auch im erlernten Algorithmus manifestiert. Auch wenn es für die Bewertung der Genauigkeit von ML-Modellen KPIs gibt, so gibt es wenige Metriken für die Qualität der Daten. Oft lautet das Credo: „Reicht die Erkennungsrate nicht aus, nutze mehr Daten im Training". Damit werden kleinere Fehler normalerweise „ausgebügelt". Besonders zielgerichtet ist diese „Brute Force"-Methode allerdings nicht. Besser ist eine präzise Anreicherung der Daten unter Vermeidung von Redundanz: viele verschiedene Situationen, und eine ausreichende Menge an Grenzfällen.

Die Suche nach *Corner Cases* im Straßenverkehr ist insbesondere in München nicht schwer und erfordert oft nur wenige Minuten Beteiligung. Dazu addieren sich Grenzfälle aus dem ländlichen Raum, aus dem Ausland oder aus Großstädten mit noch höherer Verkehrslast. Im Austausch mit Projektpartnern wurde aber schnell klar: Fachliches Wissen über Verkehrssituationen und dazugehörige Corner Cases aus Sicht der Verkehrsteilnehmer ist nur in wenigen Fällen relevant für das Testen dieser speziellen KI! Vielmehr braucht es ein Verständnis über die Architektur und Funktionsweise von Machine Learning und der individuellen Schwächen.

Das führt uns zur zweiten Herausforderung beim Testen von KI: *Metriken* für die Testabdeckung. In „klassischer" Software steht jedem Tester ein ausgereifter Methodenkoffer für Testdesigns zu Verfügung, aus dem sich unterschiedlichste Testentwurfsverfahren und Testabdeckungsmetriken für Black-/Grey-/White-Box-Tests auswählen lassen. Von Entwickler- über Integrations- bis hin zu System-/E2E-Tests sind wir in der Lage, eine passende Teststrategie inkl. Testabdeckungszielen und -metriken zu definieren.

Wie sieht es aber im Falle von KI aus, besonders im komplexen Bereich des autonomen Fahrens, in dem viele verschiedene Deep-Learning-Ansätze aufeinandertreffen? Wenn wir bottom-up im Teststufenmodell anfangen, müssen wir uns intensiv mit der konkreten Architektur auseinandersetzen. Top-down hingegen sind „Use Cases" das Ziel der Testabdeckung.

Die Modelle arbeiten aber letztlich mit den Daten der Sensoren. Hier konnten wir uns im Testdesign kreativ entfalten: falsche Inputdaten wegen Sensor-Problemen, extreme Lichtverhältnisse oder Verdeckungen. Wie können Sensoren verschiedener Art (Kameras, LiDAR, Radar usw.) so beeinflusst werden, dass sie die Situation falsch einschätzen? Entstehen im Zusammenspiel neue Grenzsituationen? Kollaborativ entstand so eine umfangreiche Taxonomie für Corner Cases aus Sicht der KI.

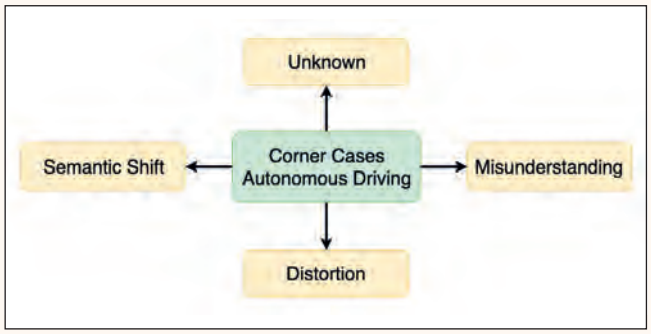

Abb. 1: Die vier Hauptkategorien der Corner Case Taxonomie

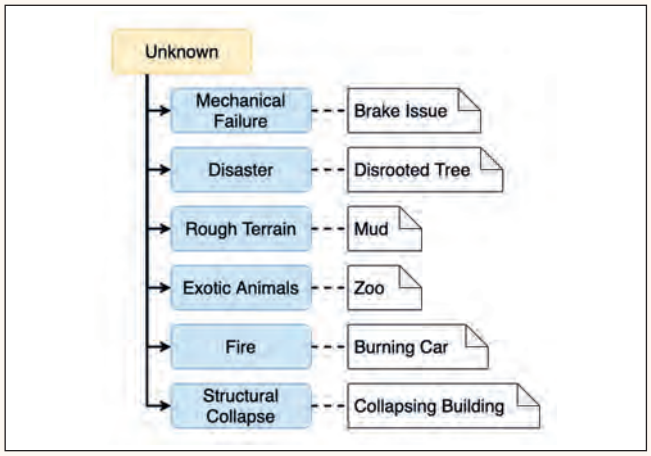

Abb. 2: Unterteilung der Hauptkategorie *Unknown*

Abb. 3: Ein umgestürzter Baum blockiert horizontal eine Waldstraße
(Quelle: QualityMinds GmbH)

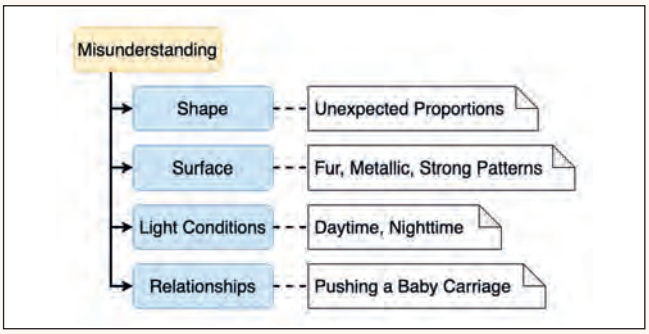

Abb. 4: Unterteilung der Hauptkategorie *Misunderstanding*

Abb. 5: Ein Auto mit einem Zebra-Muster
(mit freundlicher Genehmigung von Rosenberger GmbH und Co. KG)

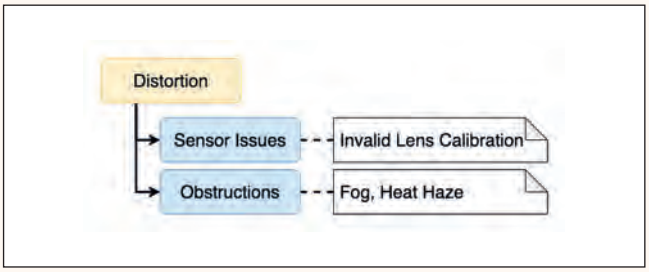

Abb. 6: Unterteilung der Hauptkategorie *Distortion*

Corner Case Taxonomie

Die Taxonomie stellt sowohl einen Überblick als auch eine systematische Beschreibung verschiedener Kategorien von Corner Cases dar, die im Kontext autonomer Fahrsysteme relevant sind. Zunächst werden vier Hauptkategorien unterschieden **(siehe Abbildung 1)**.

Die Kategorie *Unknown* **(siehe Abbildung 2)** umfasst Corner Cases, in denen die Eingabedaten dem Modell unbekannt sind, da sie nicht Bestandteil der Trainingsdaten waren. Alle erlernten Grenzwerte für Entscheidungen sind so weit entfernt, dass eine korrekte Erkennung nicht mehr möglich ist. Das Modell kann also bestenfalls raten. Ein Beispiel hierfür sind Situationen mit entwurzelten Bäumen **(siehe Abbildung 3)**, wenn das Modell nur mit aufrecht stehenden Bäumen angelernt wurde.

Die Kategorie *Misunderstanding* **(siehe Abbildung 4)** umfasst Corner Cases, in denen die Eingabedaten prinzipiell richtig erkannt werden könnten. Sie enthalten jedoch Elemente, die zu einer Fehlkategorisierung führen. So ist beispielsweise das Zebra-Muster ein dominantes Merkmal für die Erkennung von Zebras. Wenn nun ein Auto ein Zebra-Muster aufweist **(siehe Abbildung 5)**, könnte das Auto als Zebra klassifiziert werden, obwohl das Modell normalerweise eine hohe Genauigkeit in der Erkennung sowohl von Autos als auch von Zebras hat.

Die Kategorie *Distortion* **(siehe Abbildung 6)** umfasst Corner Cases, in denen das Modell Situationen normalerweise korrekt erkennen würde, ein äußerer Umstand allerdings dafür sorgt, dass die Abbildung auf die Eingabeparameter verzerrt oder lückenhaft ist. Ein Beispiel dafür ist aufwirbelndes Laub, das zu einer Art Rauschen im Bild führt. Diese Art von

Abb. 7: Überbelichtung durch tief stehende Sonne (Quelle: QualityMinds GmbH)

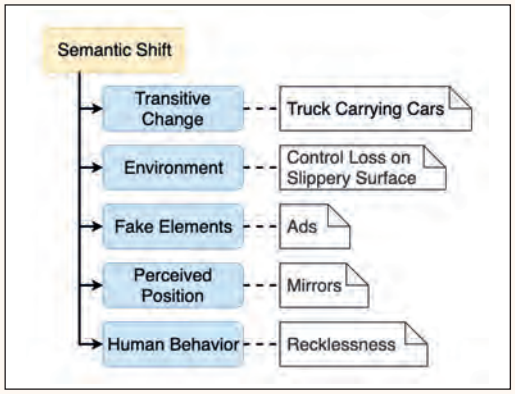

Abb. 8: Unterteilung der Hauptkategorie *Semantic Shift*

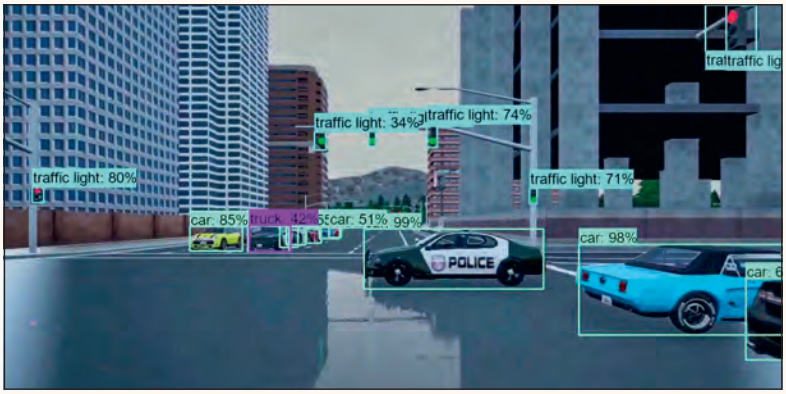

Abb. 9: Ein Polizeiauto fährt bei Rot in die Kreuzung (Quelle: QualityMinds GmbH)

Rauschen in den Daten lässt sich nicht durch eine Korrektur des KI-Systems eliminieren. Ein weiteres Beispiel ist die Überbelichtung des Bilds durch die tief stehende Sonne **(siehe Abbildung 7)**. Auch für solche Fälle muss die Robustheit des KI-Systems sichergestellt werden.

Die Kategorie *Semantic Shift* **(siehe Abbildung 8)** umfasst Corner Cases, in denen die Situation bekannt ist und korrekt klassifiziert wird. Allerdings befinden sich die Elemente der Situation in einem Verhältnis zueinander, wodurch sich der jeweilige Kontext ändert. Das Modell kann die Situation auf einer umfassenden Ebene nicht mehr richtig beurteilen. Ein Beispiel hierfür ist ein Polizeiauto, das bei Rot in eine Kreuzung einfährt und dabei von anderen Teilnehmern verdeckt wird **(siehe Abbildung 9)**. Dieser Corner Case zielt auf den Entscheidungsprozess des autonomen Fahrsystems ab.

Mithilfe solcher Taxonomien ist es möglich, Abdeckungsmetriken zu definieren: Eine gute Abdeckung wäre erreicht, wenn jeder Knoten der Taxonomie im Testdatensatz in einer gewissen Anzahl repräsentiert ist. Für neuronale Netze wurden bereits verschiedene Metriken vorgeschlagen, die jedoch große Nachteile aufweisen **(siehe Tabelle 1)**.

Wir sehen deshalb großes Potenzial in der Entwicklung einer kontextabhängigen Ab-

Abdeckungsmetrik	Testabdeckung	Nachteil
Neuron Activation Coverage	100 Prozent Testabdeckung, wenn jedes Neuron des Netzes einmal aktiviert wurde	Eine vollständige Abdeckung ist bereits mit einer geringen Anzahl von Eingabevarianten erreicht. Folglich nur beschränkt aussagekräftig
Sign-Sign Coverage	100 Prozent Testabdeckung, sofern der Output jedes Neurons einmal das Vorzeichen wechselte – in jeder möglichen Kombination mit anderen Neuronen	Die abzudeckenden Fälle unterliegen einer kombinatorischen Explosion. Für komplexere Netztypen wie CNN oder RNN ungeeignet

Tabelle 1: Abdeckungsmetriken für neuronale Netze

Referenzen

- [KIAB] KI Absicherung, siehe: https://www.ki-absicherung.vdali.de/
- [QA2018] QA Test, 17th International Conference on Software QA & Testing on Embedded Systems, siehe Keynote von R. Werner, QualityMinds: "Hey Google, will AI kill the traditional Mobile Tester?", siehe: https://www.qatest.org/wp-content/uploads/2018/07/QAtest2018.pdf
- [R2018] A. Rosenfeld, R. Zemel, J.K. Tsotsos, The Elephant in the Room, 09.08.2018, siehe https://arxiv.org/pdf/1808.03305.pdf

deckungsmetrik anhand dieser Corner Case Taxonomie. Neben den bekannten ISTQB-Metriken, deren Eignung wir untersuchen möchten, sehen wir einen großen Nutzen in semantisch reichhaltigeren Abdeckungskriterien. Diese könnten auf der Auftrittswahrscheinlichkeit oder Kritikalität basieren: Je höher, umso wichtiger ist hier eine ausreichende Anzahl an Tests. Für Paare von Kategorien wäre deren Koinzidenz interessant. Durch weitere statistische Kennzahlen könnten sinnvolle Abdeckungen berechnet und Testfälle prozedural erzeugt oder simuliert werden – hochautomatisiert.

Wohin geht der Weg?

Das Projekt [KIAB] läuft bis zum Juni 2022. Seit Projektbeginn im Juli 2019 konnten wir schon erste Erkenntnisse gewinnen.

Erkenntnis 1:
„Ein neues Test-Paradigma"

Bekannte Methoden aus der Testing-Welt sind nur eingeschränkt anwendbar. Besonders bei neuronalen Netzen und Anwendungen aus dem Bereich Computer Vision ist die Menge an notwendigen Testdaten, um die Sicherheit einer KI-Funktion nachweisen zu können, extrem groß. Die Tests einer KI hängen stark von den Trainingsdaten und der Architektur ab. Auch wenn derzeit immer mehr an Methoden der Qualitätssicherung für die Verbesserung der Trainingsprozesse geforscht wird, so müssen gegenwärtig auch die Tester ein Auge auf die Datenqualität haben.

Erkenntnis 2: „Tester brauchen Grundlagenwissen zu Machine Learning"

Auch wenn es Tester aus der Automotive-Branche gibt, die spezielle Sonderfälle für Sensoren (bspw. Kamera oder Radar) oder Verkehrssituationen kennen, so hilft dies kaum beim Testen der KI-Funktion eines autonom fahrenden Autos. Tester benötigen grundlegende Kenntnisse über Machine Learning. Es gibt viele sehr gute Inhalte und Kurse, jedoch gehen die wenigsten auf das Thema Testen ein. Diese Lücke müssen Trainings-Anbieter schließen.

Erkenntnis 3:
„Verbesserung der Abdeckung durch automatisch erzeugte Varianten"

Durch die Generierung von Varianten (bspw. andere Texturen für Kleidung) kann eine höhere Testabdeckung des Modells erreicht werden. Diese Varianten könnten durch ML-Algorithmen erzeugt werden. Die zentrale Frage hierbei: Wenn KI für das Testen von KI verwendet wird, wer testet dann diese?

Die Zukunft bleibt spannend und wird sicherlich in vielen Bereichen Errungenschaften der KI-Forschung hervorbringen. Ein Grund mehr, sich auch als Tester mit dem spannenden Thema zu beschäftigen. Obwohl KI den Tester noch viele Jahrzehnte kaum ersetzen wird [QA2018], bringt sie uns neue Herausforderungen. KI-basierte Testing-Tools können uns helfen und unterstützen, damit wir noch besser testen.

Marco Hoffmann

marco.hoffmann@qualityminds.de
forscht über maschinelles Lernen mit besonderem Interesse an Reinforcement Learning. Er ist Doktorand im Bereich Human Factors in Software Engineering.

Dr. Alexander Pohl

alexander.pohl@qualityminds.de
arbeitet in der Forschung und Entwicklung. Er ist Generalist mit einem Faible für menschliches und maschinelles Lernen, sowie Programmiersprachen und -paradigmen.

Patrick Prill

patrick.prill@qualityminds.de
arbeitet seit 20 Jahren als Tester in all seinen Ausprägungen.

Dr. Michael Mlynarski

michael.mlynarski@qualityminds.de
ist ein Geschäftsführer, der selbst gerne in Projekten berät. Sein Herz schlägt besonders für agile Softwareentwicklung und Testen. In den letzten Jahren mit starkem Fokus für das Thema maschinelles Lernen sowie autonomes Fahren.