DER NUTZEN VON KÜNSTLICHER INTELLIGENZ BEI DER ERSTELLUNG VON TESTDATEN

AUTOR: **ROB CRUTZEN**

Die Möglichkeiten von Künstlicher Intelligenz bei der Qualitätssicherung von Software wird oft unterschätzt. KI kann helfen bei der Testfallerstellung, der Ausführung von Testfällen, bei der Überprüfung und Darstellung von Ergebnissen, bei der Einrichtung von Testumgebungen oder bei der Erstellung von Testdaten.
Künstliche Intelligenz hilft, schneller und besser zu entscheiden und damit wettbewerbsfähiger zu arbeiten. KI verändert auch traditionelle Ansätze beim Testen von Software.

DIE HERAUSFORDERUNG MIT „ANONYMISIERTEN" DATEN

Die Datenschutz-Grundverordnung der Europäischen Union (EU-DSGVO) ist seit dem 25. Mai 2018 in Kraft. Sie schützt personenbezogene Daten (PII) über Einwohner in der EU. Wer diese Daten sammelt, unterliegt strengen Vorgaben und wer sie nicht einhält, muss mit Strafen rechnen. Viele Unternehmen anonymisieren die Produktionsdaten für die Qualitätssicherung und nehmen an, damit mit der DSGVO konform zu sein. Die Forschungsstudie „Estimating the success of reidentifcations in incomplete datasets using generative mode" von Luc Rocher, Julien M. Hendrickx und Yves-Alexandre de Montjoye aus 2019, zeigt jedoch, dass es mithilfe von Machine-Learning-Algorithmen möglich ist, 99,98 % der anonymisierten und personenbezogenen Daten wiederherzustellen. Anonymisierung scheint daher nicht ausreichend im Sinne der DSGVO zu sein.

GESTIEGENES INTERESSE AN DER VERWENDUNG SYNTHETISCHER DATEN

Wenn Vorschriften (oder auch ethische Gründe) daran hindern, echte Kundendaten zu verwenden, wie können dann Datensätze, auf denen ein differenziertes Kundenerlebnis aufgebaut wird, erstellt werden? Diese Frage treibt so ziemlich alle Unternehmen in Europa um. Der Schub der Digitalisierung in 2021 tut weiteres dazu, um eine Methode vielleicht zum Ausweg aus der Misere zu bieten: die Verwendung von synthetischen Daten. Hierbei handelt es sich um Daten, die mithilfe von Deep-Learning-Methoden aus Ihrem Originaldatensatz generiert werden. Das Bundesministerium für Bildung und Forschung fördert diese Methode zur Erstellung „von realistischen und möglichst allgemein verwendbaren Datensätzen". Diese synthetischen Datensätze stimmen hinsichtlich ihrer Datenstruktur, statistischer Ähnlichkeit und Verteilung eng mit den Originaldaten überein. Sie können anstelle der tatsächlichen Daten verwendet werden. Weil synthetische Daten keine Echtdaten

Quelle: Markus Spiske Unsplash

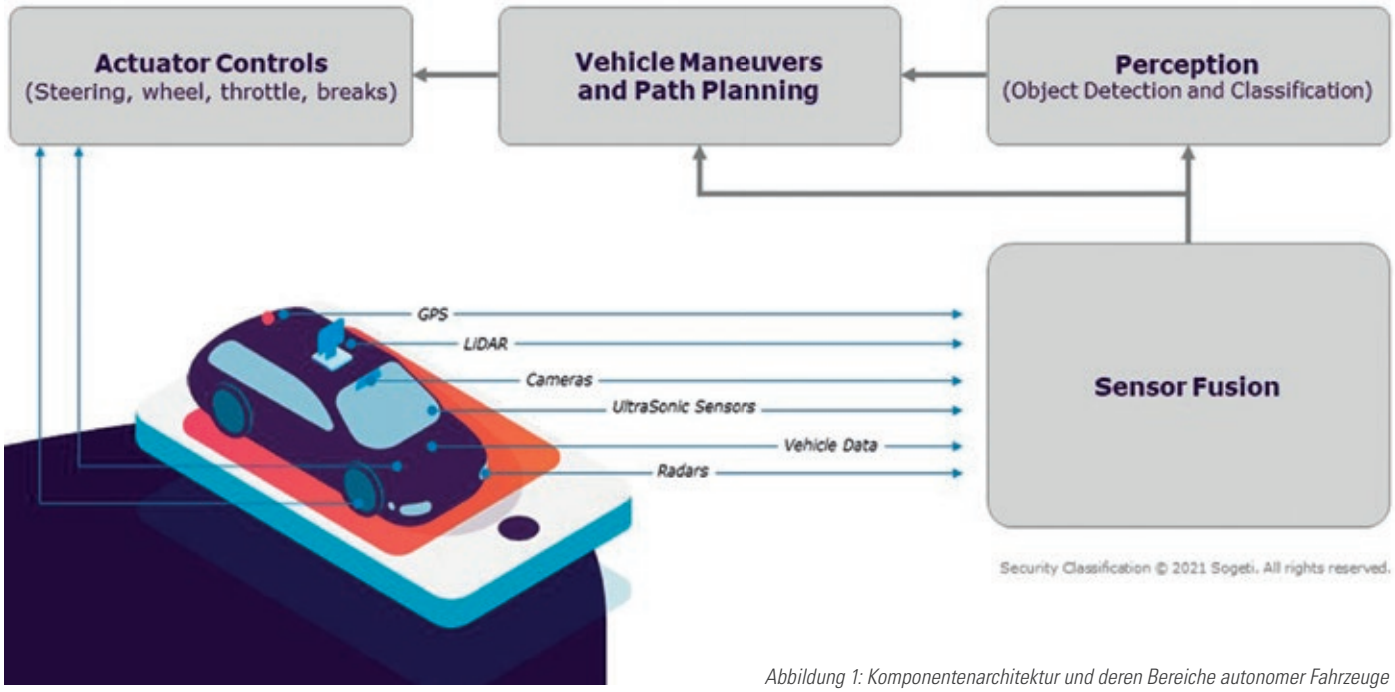

Abbildung 1: Komponentenarchitektur und deren Bereiche autonomer Fahrzeuge

sind, stehen sie auch in keinem Widerspruch zur DSGVO.

Ein weiterer Vorteil der synthetischen Datenlösung ist, dass große Datenmengen unterschiedlicher Datentypen wie Texte oder Bilder einfach und schnell erstellt werden können. Sie werden später unter anderem für das Trainieren von KI oder Last- und Performance-Tests genutzt.

BEISPIEL AUS DER PRAXIS: ANWENDUNG VON SYNTHETISCHEN DATEN IN DER AUTOINDUSTRIE

Der 2016 veröffentlichte RAND-Bericht zeigt, dass autonome Fahrzeuge, mehrere hundert Millionen und in einigen Fällen Milliarden Kilometer zurücklegen müssen, um eine ausreichende Datenmenge zum Nachweis ihrer Sicherheit zu erzeugen. Das ist sehr aufwendig. Daher müssen alternative Methoden entwickelt werden, die die Tests auf der Straße ergänzen. Diese Methoden können u.a. Simulationen, virtuelles Testen, mathematische Modellierung, Szenariotests und Pilotstudien umfassen.

Die Genauigkeit von Deep-Learning-Modellen ist direkt proportional zur Größe und Qualität der Fahr- und Fahrzeugdaten, auf die sie trainiert werden. Das macht das Testen zu einem der Schlüsselelemente. Die folgende Abbildung zeigt die Komponentenarchitektur und deren Bereiche autonomer Fahrzeuge auf abstrakter Ebene, in denen Künstliche Intelligenz verwendet wird, und die daher große qualitative Datenmengen für das Testen auf Komponentenebene benötigen.

Deep-Learning-Komponenten benötigen große Datenmengen, um zu lernen und um die Leistung und Zuverlässigkeit ihrer Ergebnisse zu verbessern. Es gibt drei Ansätze: unsupervised learning, supervised learning und supervised mit vorgeschaltetem unsupervised learning. Die meisten Vorteile des Deep Learning liegen im letzteren Ansatz.

Um das Problem unzureichender Trainingsdaten anzugehen, werden synthetische Daten, die eine Kombination aus Rohdaten und deren Datenlabels enthalten, aus den Originaldaten erzeugt. Die Datenlabels werden typischerweise von einem Menschen bereitgestellt, um anzuzeigen und zu kategorisieren, wo sich Objekte in den Eingabedaten (z. B. video frame) befinden. Diese Informationen werden während des Trainingsprozesses dem Algorithmus für Deep Learning (Objekterkennung) zugeführt.

Infolgedessen passt der Algorithmus den internen Informationsfluss (das Gewicht) verschiedener neuronaler Netzwerkschichten an und kann mit einer ausreichenden Menge von Trainingsdaten, Strukturen und Muster in den Eingabedaten erkennen. Das Generieren von Fahrdaten von Rohsensoren wie Video-Streaming ist zeitaufwendig, insbesondere wenn Rohdaten sowohl von Kameras als auch von LiDARs (light detection and ranging) erfasst werden. In solchen Fällen benötigen die Algorithmen korrekte Datenlabels für jeden Datenrahmen. Das benutzerdefinierte Artificial Data Amplifier (ADA) Modell, welches von Sogeti entwickelt wurde, verwendet Künstliche Intelligenz, um bereits beschriftete Objekte über mehrere Frames hinweg zu verfolgen und anschließend Rohdaten zusammen mit deren Datenlabels zu erzeugen. Es erfordert eine Spezialisierung und Qualifizierung, wenn Daten von Sensoren wie LiDAR als Informationen den Programmen zur Steuerung der Fahrzeuge hinzugefügt werden. Dies wird mithilfe von ADA erleichtert, das durch die proprietäre Einbindung der auf Künstlicher Intelligenz basier-

ten Technologie eine automatisierte Kennzeichnung erstellen kann. ADA kann Blue-Prints von Daten erstellen, indem neue mögliche Fahrszenarien aus den ursprünglichen Fahrszenarien vorweggenommen werden, die Faktoren wie Wetterphänomene, Fahrbahnbedingungen, Verkehrsstaus usw. umfassen. Mit diesen Mustern können Millionen von Simulationen erstellt werden, um das Training und die Programmierung vieler Fahrzeugkomponenten und Sensoren erheblich zu beschleunigen.

ANWENDUNG VON SYNTHETISCHEN DATEN IM FINANZSEKTOR

Der Boom bei globalen digitalen Zahlungen hat zu einem enormen Anstieg von Milliarden von Debit- und Kreditkartentransaktionen geführt. Diese Transaktionen haben einen Wert von vielen Millionen Euro für Zahlungen, die von Finanzdienstleistungsunternehmen abgewickelt werden. Um genaue, effiziente und sichere Zahlungsprozesse zu gewährleisten, benötigen Finanzdienstleistungsunternehmen große Mengen hochwertiger Testdaten. Um die Privatsphäre und Sicherheit der Karteninhaber zu schützen, müssen Qualitätssicherungstests ohne Verwendung personenbezogener Daten während des Testbetriebs durchgeführt werden.

Die Technologie für die Zahlungsabwicklung ist in ihrer Fähigkeit, komplexe elektronische Zahlungsprozesse zu verwalten, hoch entwickelt. Die Software unterstützt Händler in einer Vielzahl vertikaler Märkte (z. B. E-Commerce) und Servicekarteninhaber mit einer Vielzahl von Kartenkategorien, Incentive- und Treueprogrammen, Kreditverläufen und Ausgabenbeschränkungen.

Die Datenerfassung und Informationsverarbeitung muss einem genau definierten Datenaustauschformat entsprechen, da diese Datenfeeds unstrukturiert sind und keinen Standards folgen. Das Team für die Qualitätssicherung verlangt, dass seine Datenfeeds in einer stark kontrollierten Umgebung simuliert werden. Bei komplexen Transaktionsdaten-Feeds erstellt das Team im Allgemeinen Kopien einer Teilmenge ihrer Produktionsdaten und bereitet sie für das Testen vor.

Produktionsdaten sind attraktiv, weil sie echte Transaktionen im richtigen Datenaustauschformat enthalten. Um die Daten für den Test vorzubereiten, müssen sie jedoch sorgfältig von Hand überarbeitet werden, um die für Testfälle erforderlichen Datenvariatio-

Quelle: tupixen-unsplash

Beim Bezahlen im Internet gehen hochsensible Daten in den Äther. Tests dürfen keine persönlichen Daten der Karteninhaber verwenden.

nen und -versionen zu erstellen. Die Erstellung dieser Testdatensätze dauert in der Regel Wochen oder Monate. Außerdem kann es passieren, dass das Datenaustauschformat alle paar Monate überarbeitet werden muss, sodass sich die Anzahl der für die Bereitstellung von Testdaten erforderlichen Arbeitsstunden im Laufe eines Jahres vervielfachen kann. Der langwierige Bereitstellungsprozess beeinträchtigt daher die Anzahl der Testdaten, die für das Testen verfügbar sind.

Herausforderung 1: Produktionsdaten sind keine kontrollierten Daten

Ohne manuelle Änderung können aus Produktionsdaten kopierte Testdaten nur auf Bedingungen testen, die durch eine bestimmte Datenuntermenge dargestellt werden. Das QA-Team erhält nicht die erforderlichen Daten, um die Randfallbedingungen, das Vorhandensein ungültiger Datenwerte oder bestimmte Eingabewertkombinationen zu testen, die möglicherweise Softwarefehler aufdecken. Um die Codeabdeckung unter allen möglichen Betriebsbedingungen zu maximieren, müssen Testdaten gesteuert werden, um Datenfeeds zu simulieren, die alle für jeden Testfall und seine Aussagen erforderlichen Datenvariationen enthalten.

Herausforderung 2: Produktionsdaten sind keine sicheren Daten

Das Risiko einer Datenschutzverletzung, durch die sensible Kundeninformationen offengelegt werden könnten, ist unter Berücksichtigung der rechtlichen und finanziellen Konsequenzen zu groß. Die Bedrohung des Datenschutzes wird durch die Tatsache weiter verschärft, dass Unternehmen solche Testaktivitäten häufig mithilfe von Offshore-Vertragsressourcen auslagern, wodurch die interne Kontrolle über den Umgang mit sensiblen Kundendaten eingeschränkt wird.

Herausforderung 3: Sichere Produktionstestdaten für große Mengen sind nicht praktikabel

Ein herkömmlicher Ansatz, der häufig zur Minderung der Sicherheitsrisiken beim Arbeiten mit Produktionsdaten verwendet wird, ist die Datenmaskierung. Das Maskieren aller PII, die in den Transaktionsdaten-Feeds enthalten sind, die von Zahlungsverarbeitungssystemen verwendet werden, ist jedoch eine monumentale Aufgabe. Transaktionsdaten-Feeds sind Datenstrukturen, die Steuercodes, Datensatztypen, akkumulierte Transaktionswerte und Berechnungen für Prämienpunkte und Cashback-Anreize sowie echte Karteninhaber- und Händlerkontonummern und Kreditinformationen enthalten. Das Finden und Maskieren der vertraulichen Informationen in diesem komplexen Datenstrom unter Wahrung der referentiellen Integrität der Datenwerte ist anstrengend und zeitaufwendig.
Eine Lösung für diese Herausforderungen ist die Verwendung synthetischer Testdaten, erstellt mithilfe von Künstlicher Intelligenz. Hiermit können auf der Grundlage eines Deep-Learning-Modells große Mengen DSGVO-konforme Daten erstellt werden.
ADA kann beispielsweise Testdaten für jeden komplexen Datenfeed erstellen, wie es aktuell keine andere TDM- oder TDG-Plattform erstellen kann. Auch können benutzerdefinierte Datengeneratoren und Datenempfänger entwickelt werden, um kontrollierte, strukturierte und konditionierte Testdaten zu erstellen und jeden Datenfeed mit sicheren synthetischen Daten zu simulieren.

KÜNSTLICHE INTELLIGENZ: INVESTITION IN DIE ZUKUNFT DER QUALITÄTSSICHERUNG

Künstliche Intelligenz in der Qualitätssicherung wird längst genutzt. KI bietet mehr und mehr Möglichkeiten hinsichtlich Testlösungen und Werkzeugen. Mit KI kann die Qualitätssicherung besser, schneller und günstiger durchgeführt werden. Am Markt verfügbare KI-Technologien können für die Qualitätssicherung eingesetzt und verwendet werden. ■

DER AUTOR

ROB CRUTZEN ist Senior Solution Architect und Business Development Manager bei Sogeti Deutschland. Er hat über 35 Jahre Erfahrung in der IT, davon 15 Jahre in der Qualitätssicherung. Derzeit treibt er die Themen Qualitätssicherung mit KI in Deutschland und Quality Assurance & Engineering im SAP-Technologiestack voran.