Wie synthetische Datensätze KI-Systeme verbessern sollen

Originalartikel

Backup

<html> <header class=„article-header“><h1 class=„articleheading“>Wie synthetische Datensätze KI-Systeme verbessern sollen</h1> Karen Hao </header><figure class=„aufmacherbild“><figcaption class=„akwa-caption“>(Bild: Datagen)</figcaption></figure> Deep Learning benötigt große Informationsmengen. Aus realen Informationen abgeleitete Fake-Daten sollen helfen. Man kann die schwachen Stoppeln auf seiner Oberlippe sehen, die Falten auf seiner Stirn, die Unreinheiten seiner Haut. Er ist kein echter Mensch, aber er orientiert sich an ihnen – so wie Hunderttausende andere, die von Datagen hergestellt werden, einer Firma, die Fakes von Menschen verkauft. Diese falschen Personen sind keine Spiel-Avatare oder animierte Figuren für Filme. Sie dienen als künstliche Daten, um damit Deep-Learning-Algorithmen zu füttern. Firmen wie Datagen [1] wollen damit eine Alternative zum teuren und zeitaufwändigen Sammeln von Daten aus der realen Welt anbieten. Das Unternehmen generiert die Informationen maßgeschneidert für den Kunden, wie und wann er es will – und das zu einem relativ günstigen Preis. Um seine synthetischen Menschen zu erzeugen, scannt Datagen zunächst reale Personen. Das Unternehmen arbeitet mit Zwischenhändlern zusammen, die Menschen dafür bezahlen, in großen Ganzkörperscannern jedes Detail von der Iris über die Hautbeschaffenheit bis hin zur Krümmung der Finger erfassen zu lassen. Aus diesen Rohdaten kreiert das Startup mit Hilfe einer ganzen Reihe von Algorithmen 3D-Darstellungen von Körper, Gesicht, Augen und Händen einer Person. <h3 class=„subheading“ id=„nav_nicht_einfach0“>Nicht einfach „Daumen hoch“</h3> Das Unternehmen mit Sitz in Israel arbeitet nach eigenen Angaben bereits mit vier großen US-Tech-Giganten zusammen, will aber nicht verraten, mit welchen. Sein wichtigster Konkurrent, Synthesis AI [2], bietet ebenfalls digitale Menschen auf Abruf an. Andere Unternehmen generieren Daten für die Finanz-, Versicherungs- und Gesundheitsbranche. Es gibt mittlerweile zahlreiche Firmen auf dem Gebiet. Einst galten synthetische Daten im Vergleich zu realen als minderwertig. Heute hingegen sieht so mancher Beobachter in ihnen ein Allheilmittel. Echte Daten sind unübersichtlich und mit Fehlern behaftet. Neue Datenschutzbestimmungen erschweren zudem ihr Sammeln. Im Gegensatz dazu lassen sich aus synthetischen Daten viel leichter die unterschiedlichsten Datensätze erstellen. So kann man daraus zum Beispiel perfekte Gesichter unterschiedlichen Alters, unterschiedlicher Form und ethnischer Zugehörigkeit erzeugen. Damit

lässt sich dann ein Programm zur Gesichtserkennung entwickeln, das für alle Bevölkerungsgruppen funktioniert. Aber synthetische Daten haben auch Nachteile. Spiegeln sie die Realität nicht richtig wider, könnte das zu schlechteren Ergebnissen führen als weniger genaue Daten aus der realen Welt – oder zumindest zu den Problemen führen, die diese haben. „Ich mag hier nicht einfach ein „Daumen hoch“ setzen und sagen: Oh, das wird so viele Probleme lösen“, sagt Cathy O'Neil, eine Datenwissenschaftlerin und Gründerin der auf die Überprüfung von Algorithmen spezialisierte Firma ORCAA. „Denn bei der Methode werden auch viele Dinge nicht beachtet.“ <h3 class=„subheading“ id=„nav_realistisch1“>Realistisch, nicht real</h3> Deep Learning brauchte schon immer viele Daten. Aber in den letzten Jahren hat sich gezeigt, dass deren Qualität wichtiger ist als ihre Menge. Selbst kleine Mengen richtiger, sauber zugeordneter Daten verbessern die Leistung eines KI-Systems mehr als die zehnfache Menge schlecht aufbereiteter Daten, selbst wenn man leistungsfähigere Algorithmen verwendet. Das sollten Unternehmen bei der Entwicklung ihrer KI-Modelle berücksichtigen, sagt Ofir Chakon, CEO und Mitbegründer von Datagen. Heute sammeln die Firmen zumeist erst einmal so viele Daten wie möglich und optimieren dann ihre Algorithmen. Stattdessen sollten sie das Gegenteil tun: Den selben Algorithmus verwenden, aber die Qualität ihrer Daten verbessern. Doch reale Daten für solches iteratives Experimentieren zu sammeln, ist zu kostspielig und zeitintensiv. An dieser Stelle kommt Datagen ins Spiel. Mit einem Generator für synthetische Daten können Teams Dutzende von neuen Datensätzen pro Tag erstellen und sie testen, um herauszufinden, welche die Realität am besten abbilden. <h3 class=„subheading“ id=„nav_hunderttausende2“>Hunderttausende Kombinationen</h3> Um die Realitätsnähe der Daten zu gewährleisten, gibt Datagen seinen Lieferanten detaillierte Anweisungen, wie viele Personen in jeder Alters- und Gewichtsklasse sowie ethnischer Zugehörigkeit zu scannen sind. Hinzu kommt eine Liste von Aktionen, die sie ausführen sollen, wie etwa in einem Raum herumlaufen oder eine Limonade trinken. Die Hersteller senden sowohl hochauflösende statische Bilder als auch Motion-Capture-Daten zurück. Die Algorithmen von Datagen erstellen dann hunderttausende Kombinationen aus diesen Daten. Diese synthetisierten Daten werden dann manchmal noch einmal überprüft, etwa künstliche Gesichter mit echten verglichen, um zu sehen, wie realistisch sie wirken. <figure class=„branding“>
alt=„Mehr von MIT Technology Review“ height=„693“
src=„https://static.wallabag.it/7862d1b7aff4c3b00f37212fefade4e0e2c4cf00/64656e6965643a646174613a696d6167652f7376672b786d6c2c25334373766725323078 6d6c6e733d27687474703a2f2f7777772e77332e6f72672f323030302f737667272532307769647468 3d27363936707827253230686569676874 3d27333931707827253230766965774 26f783d27302532303025323036393625323033393127253345253343726563742532307 83d273027253230793d27302725323077696474683d27363936272532306865696768743d27333931272532306696c6c3d272532336632663266322725334525334325 32f726563742533452533432f737667253345/“ class=„c1“ width=„1200“ referrerpolicy=„no-referrer“ />
alt=„Mehr von MIT Technology Review“ class=„a-u-hide-from-tablet c2“
src=„https://heise.cloudimg.io/width/1200/q50.png-lossy-50.webp-lossy-50.foil1/_www-heise-de_/Magazin-Banner/tr_mobil.jpg“
srcset=„https://heise.cloudimg.io/width/2400/q30.png-lossy-30.webp-lossy-30.foil1/_www-heise-de_/Magazin-Banner/tr_mobil.jpg 2x“ referrerpolicy=„no-referrer“ />

2533452533437265637425323078 3d27302725323079 3d27302725323077696474683d27363936272532306865696768743d27333931272532306669 6c6c3d27252336632663266322725334525334 32f726563742533452533432f737667253345/“ class=„c3“ width=„1830“ referrerpolicy=„no-referrer“ /> [3]</figure> Datagen generiert beispielsweise Gesichtsausdrücke zur Überwachung der Aufmerksamkeit von Fahrerinnen und Fahrern in smarten Autos. Oder Körperbewegungen, um Kunden in kassenlosen Geschäften zu verfolgen, sowie Iris und Handbewegungen, um die Augen- und Hand-Tracking-Funktionen von VR-Headsets zu verbessern. Laut dem Unternehmen dienten seine Daten bereits zur Entwicklung von Bilderkennungssystemen (Computer Vision), die mehrere Millionen Nutzerinnen und Nutzer verwenden. <h3 class=„subheading“ id=„nav_von3“>Von Fahrzeuginspektion bis zur Medizin</h3> Es werden nicht nur synthetische Menschen in Massenproduktion geschaffen. Das Startup Click-Ins [4] verwendet zum Beispiel synthetische Daten für automatische Fahrzeuginspektionen. Mithilfe von Design-Software werden alle Automarken und -modelle, die die KI erkennen muss, neu erstellt und dann mit verschiedenen Farben, Schäden und Verformungen unter verschiedenen Lichtverhältnissen und vor verschiedenen Hintergründen gerendert. Auf diese Weise kann das Unternehmen sein KI-Modell aktualisieren, wenn Autohersteller neue Fahrzeugvarianten auf den Markt bringen. Außerdem wird so kein Datenschutz verletzt in Ländern wie Deutschland, in denen Nummernschilder als private Informationen gelten und daher nicht in Fotos zum Training der KI enthalten sein dürfen. Mostly.ai [5] arbeitet wiederum mit Finanz-, Telekommunikations- und Versicherungsunternehmen zusammen, um Tabellen mit synthetischen Kundendaten bereitzustellen. Auf diese Weise können die Unternehmen den Aufbau ihrer Kundendatenbank mit externen Dienstleistern auf rechtskonforme Weise teilen. Denn selbst wenn Daten anonymisiert werden, schützt dass zuweilen nicht ausreichend die Privatsphäre der Menschen. Mit den synthetischen Daten lassen sich Datensätze mit den gleichen statistischen Eigenschaften wie die der echten Daten eines Unternehmens generieren. So können auch Daten simuliert werden, die das Unternehmen noch gar nicht hat, zum Beispiel von hypothetischen zukünftigen Kundengruppen oder Szenarien betrügerischer Aktivitäten. Proponenten synthetischer Daten sagen, dass diese auch bei der Bewertung der Fähigkeiten von KI helfen können. Ein Beispiel zeigten Suchi Saria, Professorin für maschinelles Lernen und Gesundheitswesen an der Johns Hopkins University, und ihre Mitautoren, in einer kürzlich auf einer KI-Konferenz veröffentlichten Arbeit: Mit Hilfe von Techniken zur Datengenerierung ließen sich verschiedene Patientengruppen aus einem einzigen Datensatz extrapolieren. Das könnte nützlich sein, wenn ein Unternehmen etwa nur Daten von der eher jugendlichen Bevölkerung von New York City vorliegen hat, aber verstehen möchte, wie seine KI bei einer alternden Bevölkerung mit einer höheren Diabetesprävalenz funktioniert. Um solche medizinische KI-Systeme zu testen, gründet Saria nun ihr eigenes Unternehmen namens Bayesian Health. <h3 class=„subheading“ id=„nav_datenschutz4“>Datenschutz nicht automatisch gewährleistet</h3> Doch gibt es um synthetische Daten einen unnötigen Hype? In Sachen Datenschutz „bedeutet die Tatsache, dass die Daten 'synthetisch' sind und nicht direkt den realen Benutzerdaten entsprechen, nicht, dass sie keine sensiblen Informationen über reale Personen enthalten“, sagt Aaron Roth, Professor für Computer- und Informationswissenschaften an der University of Pennsylvania. Es habe sich gezeigt, dass einige Datengenerierungstechniken Bilder oder Texte aus ihren (echten) Trainingsdaten einfach nur

abkupfern. Das mag für eine Firma wie Datagen in Ordnung sein, deren synthetische Daten nicht dazu gedacht sind, die Identität der Personen zu verbergen, denn die haben dem Scan zugestimmt. Aber es wäre eine schlechte Nachricht für Unternehmen, die in der Methode eine Möglichkeit sehen, sensible Finanz- oder Patientendaten zu schützen. Bisherige Forschung legt nahe, dass insbesondere die Kombination von zwei Techniken für synthetische Daten – die sogenannte Differential Privacy [6] und Generative Adversarial Networks [7] (GANs) – für guten Schutz sorgen kann, sagt Bernease Herman, Datenwissenschaftler am University of Washington eScience Institute. Skeptiker befürchten jedoch, dass dieser Aspekt im Marketing-Jargon der Anbieter synthetischer Daten verloren geht. Denn die sprechen nicht immer offen darüber, welche Techniken sie verwenden. <h3 class=„subheading“ id=„nav_voreingenommenhe5“>Voreingenommenheit nicht ausgeschlossen</h3> Bislang deutet wenig darauf hin, dass synthetische Daten zuverlässig vor Voreingenommenheit schützen. Ist ein Datensatz verzerrt, so kommt man durch ein „Hochrechnen“ nicht zu wirklich repräsentativen Daten. Die Rohdaten von Datagen enthalten zum Beispiel proportional weniger ethnische Minderheiten, was bedeutet, dass weniger reale Datenpunkte verwendet werden, um Fake-Menschen aus diesen Gruppen zu erzeugen. „Wenn Ihre Gesichter mit dunklerer Hautfarbe keine besonders guten Annäherungen an reale Gesichter sind, dann lösen Sie das Problem nicht wirklich“, sagt O'Neil. Zum anderen führen perfekt ausbalancierte Datensätze nicht automatisch zu perfekt fairen KI-Systemen, sagt Christo Wilson, außerordentlicher Professor für Informatik an der Northeastern University. Wenn ein Kreditkartenanbieter versucht, einen KI-Algorithmus zur Bewertung potenzieller Kreditnehmer zu entwickeln, würde er nicht alle möglichen Diskriminierungen beseitigen, indem er einfach Weiße genauso wie Schwarze in seinen Daten repräsentiert. Diskriminierung könnte sich immer noch durch Unterschiede zwischen Bewerbern aus verschiedenen Gruppen einschleichen. Um die Sache weiter zu verkomplizieren, zeigen erste Forschungsergebnisse, dass es in manchen Fällen gar nicht möglich ist, mit synthetischen Daten sowohl die Privatsphäre zu schützen als auch eine faire KI zu entwickeln. In einer kürzlich auf einer KI-Konferenz veröffentlichten Arbeit versuchten Forscher der Universität Toronto und des Vector-Instituts dies mit Röntgenaufnahmen der Brust [8] zu erreichen. Sie fanden heraus, dass sie nicht in der Lage waren, ein akkurates medizinisches KI-System zu erstellen, als sie versuchten, einen synthetischen Datensatz durch Kombination von Differential Privacy und GANs zu erstellen. All dies heißt nicht, dass synthetische Daten nicht verwendet werden sollten. In der Tat kann das durchaus notwendig werden. Da die Aufsichtsbehörden KI-Systeme auf ihre Rechtskonformität hin überprüfen müssen, können sie möglicherweise nur so bedarfsgerechte, gezielte Testdaten generieren, so O'Neil. Aber dadurch ist es noch wichtiger, die Grenzen der Methode zu hinterfragen. „Synthetische Daten werden mit der Zeit wahrscheinlich besser werden“, sagt sie, „aber nicht von alleine“. () <hr /> URL dieses Artikels:
<small><code>https://www.heise.de/-6071301</code></small> Links in diesem Artikel:
<small><code>[1] https://www.datagen.tech/technology/</code></small>
<small><code>[2] https://synthesis.ai/</code></small>
<small><code>[3] https://www.heise.de/tr/</code></small>
<small><code>[4] https://www.click-ins.com/</code></small>
<small><code>[5] https://mostly.ai/industries/</code></small>

/><small><code>[6] https://www.heise.de/hintergrund/Kollektiver-Datenschutz-Was-dahinter-steckt-und-warum-er-noetig-ist-6054822.html</code></small>
<small><code>[7] https://www.heise.de/hintergrund/Neuronales-Netz-erzeugt-attraktive-Gesichter-fuer-jeden-Geschmack-6018436.html</code></small>
<small><code>[8] https://dl.acm.org/doi/10.1145/3442188.3445879</code></small>
<small><code>[9] mailto:bsc@heise.de</code></small>
 Copyright © 2021 Heise Medien
</html>