Falsche Trainingsdaten verzerren Güte-Einschätzung von KI-Modellen

Originalartikel

Backup

<html> <header class=„article-header“><h1 class=„articleheading“>Falsche Trainingsdaten verzerren Güte-Einschätzung von KI-Modellen</h1> Karen Hao </header><figure class=„aufmacherbild“><figcaption class=„akwa-caption“>Nein, eine Elster ist das nicht.(Bild: Jeremy Lwanga / Unsplash)</figcaption></figure> Laut einer neuen MIT-Studie sind die zehn am häufigsten verwendeten KI-Datensätze mit vielen, teilweise eklatanten Etikettierungsfehlern behaftet. Große Modelldatensätze bilden das Rückgrat der Künstlichen Intelligenz, aber einige sind wichtiger als andere. Denn mit bestimmten Kern-Datasets bewerten Forscher den Fortschritt von Maschinenlern-Modellen. Einer der bekanntesten ist der Bilderkennungsdatensatz ImageNet, der die moderne KI-Revolution mitauslöste. Und auch MNIST gehört dazu: Es stellt Bilder von handgeschriebenen Zahlen von 0 bis 9 zusammen, was bei der Schrifterkennung hilft. Wieder andere Datasets testen Modelle, die darauf trainiert sind, Audio, Text oder Zeichnungen zu erkennen. <h3 class=„subheading“ id=„nav_tags_die0“>Tags, die daneben liegen</h3> Allerdings haben Studien in den letzten Jahren ergeben, dass diese Datensätze mit schwerwiegenden Mängeln behaftet sein können. ImageNet enthält beispielsweise rassistische und sexistische Label [1] sowie Fotos von Gesichtern, die ohne Zustimmung gesammelt wurden [2]. Die neueste MIT-Studie [3] zum Thema befasst sich nun mit einem weiteren Problem: Viele der Bildetiketten, die sogenannten Tags, sind schlicht und einfach falsch. So wird etwa ein Pilz als Löffel ausgewiesen, ein Frosch als Katze und ein hoher Ton der Sängerin Ariana Grande als Pfiff. Der ImageNet-Testsatz weist eine geschätzte Etikettierfehlerrate von 5,8 Prozent auf. Der Testsatz für QuickDraw, eine Zusammenstellung von Zeichnungen, kommt auf satte 10,1 Prozent. <figure class=„branding“><img alt=„Mehr von MIT Technology Review“ height=„693“ src=„https://static.wallabag.it/7862d1b7aff4c3b00f37212fefade4e0e2c4cf00/64656e6965643a646174613a696d6167652f7376672b786d6c2c25334373766725323078 6d6c6e733d27687474703a2f2f777

 7772e77332e6f72672f323030302f737667272532307769647468 3d27363936707827253230686569676874 3d27333931707827253230766965774 26f783d27302532303025323036393625323033393127253345253343726563742532307 83d27302725323079 3d27302725323077696474683d27363936272532306865696768743d27333931272532306669 6c6c3d272532336632663266322725334525334326f726563742533452533432f737667253345/ “ class=„c1“ width=„1200“ referrerpolicy=„no-referrer“ /> [4]</figure> Wie wurde das untersucht? Jeder der zehn Datasets, die zur Bewertung von Modellen verwendet werden, verfügt über einen eigenen Trainingsdatensatz. Curtis G. Northcutt und Anish Athalye, zwei MIT-Nachwuchsforscher sowie MIT-Absolvent Jonas Mueller entwickelten mithilfe der Trainingsdatensätze ein Modell für maschinelles Lernen und ließen anschließend die Testdaten erneut beschriften. Wenn das Modell-Etikett nicht mit dem des Originals übereinstimmte, wurde der Datenpunkt zur manuellen Überprüfung markiert. Fünf Rezensenten beim Online-Dienst Amazon Mechanical Turk wurden gebeten, darüber abzustimmen, ob das Etikett des neuen Modells oder das Original richtig ist. Stimmte die Mehrheit der Prüfer für die neue Modell-Beschriftung, wurde die des Originals als Fehler gewertet und korrigiert. <h3 class=„subheading“ id=„nav_einfach_ist1“>Einfach ist besser</h3> Die Forscher prüften insgesamt 34 Modelle, deren Leistung zuvor anhand des ImageNet-Testsatzes gemessen worden war. Dabei ließen sie jedes Modell rund 1500 Beispiele mit falschen Datenbeschriftungen erneut beurteilen. Jene Modelle, die mit den ursprünglichen falschen Etiketten nicht so gut abgeschnitten hatten, gehörten nach der Korrektur zu den besten. Insbesondere die einfacheren Modelle schienen mit korrigierten Daten besser abzuschneiden als die komplizierteren Modelle, die von Technologiegiganten wie Google zur Bilderkennung verwendet werden und als die besten auf diesem Gebiet gelten. Mit anderen Worten: Wir haben möglicherweise eine übertrieben gute Ansicht davon, wie gut diese komplizierten Modelle tatsächlich sind. Als Fazit ermutigt Northcutt das KI-Feld, sauberere Datensätze zur Bewertung von Modellen und zur Verfolgung des Fortschritts des gesamten Forschungsfeldes zu erstellen. Er empfiehlt seinen Kollegen außerdem, ihre Datenhygiene zu verbessern. Andernfalls, so warnt er, könnten sie das falsche Modell auswählen. Insbesondere dann, „wenn Sie einen Datensatz voller Rauschen haben und damit eine Reihe von KI-Modellen ausprobieren, um sie der realen Welt einzusetzen“. Zu diesem Zweck hat Northcutt den Code [5], den er in seiner Studie zur Korrektur von Etikettierungsfehlern verwendet hat, als Open-Source-Lösung bereitgestellt. Er sagt, dass er

bereits bei einigen großen Technologieunternehmen verwendet wird. () <hr /> URL dieses Artikels:
<small><code>https://www.heise.de/-6010496</code></small> Links in diesem Artikel:
<small><code>[1] https://excavating.ai/</code></small>
<small><code>[2] https://www.wired.com/story/researchers-blur-faces-launched-thousand-algorithms/</code></small>
<small><code>[3] https://arxiv.org/pdf/2103.14749.pdf</code></small>
<small><code>[4] https://www.heise.de/tr/</code></small>
<small><code>[5] https://github.com/cgnorthcutt/cleanlab</code></small>
<small><code>[6] mailto:office@technology-review.de</code></small>
 Copyright © 2021 Heise Medien </html>