Woche 13 Daten: Datenaugmentierung

Skript

Erarbeitet von
Ann-Kathrin Selker

Lernziele 1
Inhalt 1
Einstieg 1
Datenaugmentierung 2
Synthetische Daten 3
Oversampling 4
Abschluss 6
Quellen 6
Weiterführendes Material 6
Disclaimer 6

Lernziele

- Erklären können, was Datenaugmentierung ist und wofür es benutzt wird
- Beispiele für mögliche Datenaugmentierung angeben
- Oversampling erklären können

Inhalt

Einstieg

Damit ein Machine-Learning-Modell gut trainiert werden kann, werden nicht nur viele Trainingsdaten benötigt, sondern auch diverse. Doch was kannst du tun, wenn dir nicht

Projekt „KI für alle"
HeiCAD – Heine Center for Artificial Intelligence and Data Science

genug Trainingsdaten zur Verfügung stehen oder gewisse Klassen deiner Trainingsdaten unterrepräsentiert sind?

Datenaugmentierung

Eine Möglichkeit, dem Abhilfe zu verschaffen, ist es, aus den bestehenden Trainingsdaten neue zu erschaffen. Angenommen, unser Machine-Learning-Modell soll Hunde erkennen. Guck dir einmal dieses Bild an. Und jetzt dieses.

Einblenden von Hundebild + gespiegeltes Hundebild

Für uns als Menschen sind das dieselben Bilder. Für den Computer handelt es sich jedoch um verschiedene Bilder, da die Pixelwerte sich unterscheiden. Das Erschaffen von neuen Daten aus den Originaldaten nennt man Datenaugmentierung. Neben dem Spiegeln gibt es auch noch andere Möglichkeiten, neue Trainingsbilder zu erschaffen. Wir können Bilder rotieren, jeweils mit verschiedenen Winkeln. Man kann verschiedene Ausschnitte aus Bildern verwenden oder in das Bild hineinzoomen. Außerdem können wir Kontrast, Helligkeit, Farbton und Sättigung ändern oder das Bild leicht verrauschen. Verrauschen bedeutet, kleine Störungen in den Daten einzufügen. Im Zusammenhang mit Bildern bedeutet Verrauschen, dass Pixelwerte leicht verändert werden, wie z. B. leicht in diesem Bild und stark in diesem Bild.

Einblenden von augmentierten Bildern

Ich habe Datenaugmentierung mit Bildern erklärt. Dieses Verfahren ist aber auch auf andere Datentypen anwendbar. Bei Audiodateien wie dieser hier kann z. B. die Geschwindigkeit höher oder niedriger gestellt werden, die Tonhöhe geändert werden oder Rauschen eingeführt werden. Die verwendete Ursprungsaudiodatei stammt übrigens aus der Datensammlung Spoken Digits, dem MNIST-Äquivalent für Audiodateien mit gesprochenen Ziffern.

Abspielen von augmentierten Audiodateien

Auch Textdateien können augmentiert werden, z. B. durch Ersetzen von Wörtern durch Synonyme oder aber durch Ändern der Reihenfolge von Wörtern und Sätzen und Entfernen oder Hinzufügen von Wörtern. So sind z.B. „KI macht viel Spaß“, „KI Spaß“, „KI bereitet Freude“, „Macht KI Spaß“ usw., alles augmentierte Versionen von dem Dokument „KI macht Spaß“.

Einblenden von augmentierten Dokumenten

Neben dem Abwandeln von bereits vorhandenen Trainingsdaten können Daten auch komplett neu erschaffen werden. Es existieren Machine-Learning-Verfahren für das Erzeugen von neuen Daten nach vorheriger Spezifikation. Allerdings musst du hier auch Qualitätseinbußen hinnehmen, da es sich nicht um reale Daten handelt und unter Umständen Verrauschungen auftreten, die so in echt nicht vorkommen und trotzdem in das Modell mit eingearbeitet werden. Sieh dir zum Beispiel einmal dieses Bild an:

Einblendung synthetisches Bild (Quelle [1])

Es soll sich um das Bild einer Frau handeln, und im Grunde ist es das auch. Allerdings ist es offensichtlich, dass wir es hier nicht mit einer real existierenden Person zu tun haben. Die Augen passen nicht, die Zähne scheinen nur aus zwei überdimensionalen Schneidezähnen zu bestehen, und vor allem befindet sich das Gesicht der Frau auf ihrem Hinterkopf. Wenn wir beim Trainieren mit ähnlichen fehlgeschlagenen synthetischen Daten arbeiten, vermindert das natürlich die Qualität des Trainings.

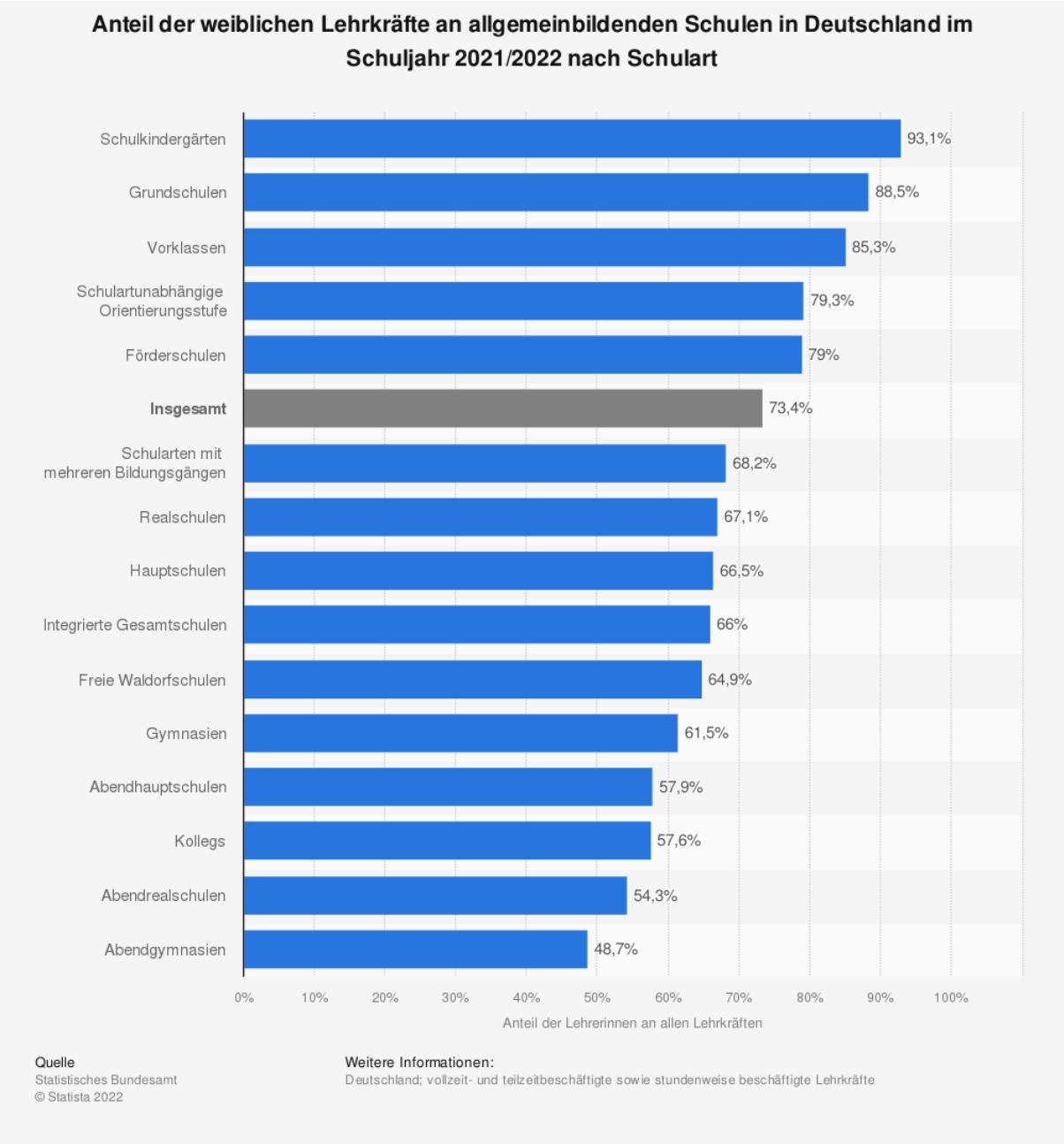
Datenaugmentierung kommt besonders in Gebieten zum Einsatz, in denen hochwertige Daten schwer zu beschaffen sind, also z. B. medizinische Bilder oder Dokumente in kaum gesprochenen Sprachen. Allerdings ist nicht jede Augmentierungstechnik für jeden Einsatzfall geeignet. Angenommen, du möchtest Ziffern erkennen. Dann ist es nicht hilfreich, bei den Ziffern 6 und 9 eine Rotation um 180 Grad durchzuführen, da das die Klasse des Bildes verändert und dadurch das gegebene Label nicht mehr stimmt.

Einblendung rotierende 6

Außerdem verschlechtert es oft den Trainingserfolg, wenn mehrere Datenaugmentierungstechniken auf dasselbe Datenobjekt angewendet werden.

Oversampling

Datenaugmentierung hilft aber nicht nur, wenn allgemein zu wenige Trainingsdaten vorhanden sind. Angenommen, du hast Daten über die Lehrkräfte an deutschen Grundschulen.



Einblendung Statistik über weibliche Lehrkräfte (Quelle [2])

Im Schuljahr 2021/2022 waren fast 90 % aller Grundschullehrkräfte weiblich. Wenn sich diese Verteilung auch so in den Trainingsdaten widerspiegelt, was ja eigentlich gewünscht ist, dann ist das Modell nicht ausreichend auf männliche Lehrkräfte trainiert. Es kommt dann zu einer Überanpassung auf weibliche Lehrkräfte. In solchen Fällen ergibt es Sinn, künstlich den Männeranteil in den Trainingsdaten zu erhöhen, sodass die Verteilung der Trainingsdaten dann bewusst nicht mehr der Verteilung in der Realität entspricht. Dieses Vorgehen heißt im engl. Oversampling.

Abschluss

In Python kannst du Datenaugmentierung z. B. mit dem Modul Tensorflow durchführen. Dabei werden neben den originalen Trainingsdaten noch zufällig erstellte Varianten der Daten benutzt. Es ist also nicht nötig, seine Daten manuell zu augmentieren.
In diesem Video hast du Techniken der Datenaugmentierung kennengelernt. Du kannst erklären, wozu Datenaugmentierung verwendet wird und wie sie durchgeführt werden kann. Außerdem kannst du Oversampling erläutern.

Quellen

Quelle [1] Andy Baio [@waxpancake]. (2022, 22. Oktober)
The Stable Diffusion Discord has a channel devoted to spectacular failures and it's full of gems [Tweet]. Twitter.
https://twitter.com/waxpancake/status/1583879929472897024

Quelle [2] Statistisches Bundesamt, © Statista 2022
https://de.statista.com/statistik/daten/studie/1129852/umfrage/frauenanteil-unter-den-lehrkraeften-in-deutschland-nach-schulart/

Weiterführendes Material

https://www.tensorflow.org/tutorials/images/data_augmentation

Disclaimer