Woche 03 Daten: Datenqualität

Skript

Erarbeitet von
Ann-Kathrin Selker

Lernziele

- den Zusammenhang zwischen Trainingsdaten und der Genauigkeit der Vorhersagen eines Modells kennen
- Beispiele für gute und schlechte Verteilungen bei Trainingsdaten angeben können
- entscheiden, ob eine gegebene Verteilung von Trainingsdaten für eine hohe Genauigkeit eines entsprechenden Modells geeignet ist

Inhalt

Wie können wir entscheiden, ob unsere vorliegenden Daten gute Trainingsdaten sind?

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

Einstieg

Für unsere späteren Modelle benötigen wir Daten, auf denen das Modell trainieren kann. Eine wichtige Eigenschaft dieser Trainingsdaten ist ihre Qualität. Die Daten müssen passend aufbereitet und bereinigt werden, um später verwendet werden zu können.

Label

Viele Modelle arbeiten mit Labeln, also einer Bezeichnung eines Datenobjekts. Beim Training lernt das Modell Zusammenhänge zwischen den Datenobjekten und den dazugehörigen Labeln und kann daher später Vorhersagen treffen. Zum Beispiel benötigt ein Modell zum Erkennen von Hunden Trainingsbilder mit den entsprechenden Bezeichnungen „Hund“ oder „kein Hund“ und liefert dann später bei ungelabelten Bildern eine Vorhersage über die Zugehörigkeit zur entsprechenden Kategorie (auch Klasse genannt). Weitere Beispiele fürs Labeln beinhalten die Stimmung eines Textes anzugeben (also ob es sich z. B. um ein positives oder negatives Gutachten handelt) oder in einem Ticketsystem Tickets den verschiedenen Abteilungen zuzuordnen. Diese Label werden häufig noch von Menschen erstellt und kosten Zeit und Geld.
Als Faustformel gilt: Je mehr richtige Label im Trainingsprozess, desto besser die Qualität des Trainings. Allerdings muss natürlich immer abgewägt werden, ob es den Zeit- und Kostenaufwand lohnt, die Daten genauer zu machen für die Qualitätsverbesserung, die erreicht wird. Je wichtiger die Genauigkeit des trainierten Modells ist, desto wichtiger ist auch die Genauigkeit der Daten. Ein Modell, das medizinische Diagnosen gibt, oder das bei selbstfahrenden Autos Hindernisse auf der Straße erkennt, muss genauer sein als das eben genutzte Beispielmodell, das Hunde erkennen kann.

Datenfallen

Es gibt allerdings auch Fallen im Bereich Datenqualität. Was wir persönlich als „hochwertige“ Daten bezeichnen, muss nicht automatisch auf das Modell übertragbar sein. Vielmehr kommt es darauf an, für welche Daten unser Modell überhaupt trainiert werden soll.

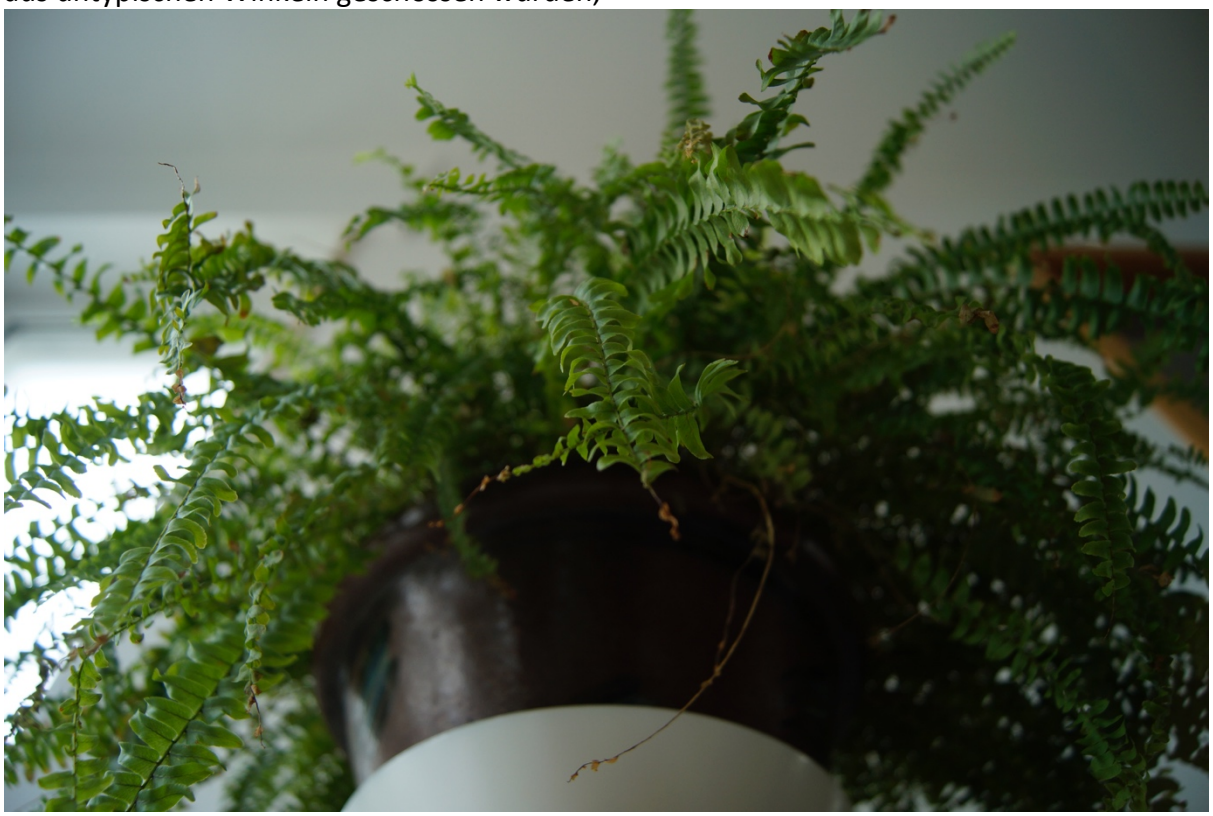
Ein gutes Beispiel dafür sind Bilder. Angenommen, wir wollen eine App erstellen, die anhand von Benutzer*innen hochgeladenen Bildern Pflanzenarten erkennt. Als Trainingsdaten nehmen wir hochwertige Fotografenfotos, wie zum Beispiel das von diesem Farn.

Einblendung Foto Farn

Diese Fotos zeichnet aus, dass sie eine hohe Auflösung haben, scharf sind, gut beleuchtet sind und häufig einer gewissen Komposition folgen, also zum Beispiel das Zielobjekt in der Mitte haben und so weiter. Das entspricht aber nicht den Fotos, die die Benutzer später schießen, um unbekannte Pflanzen erkennen zu lassen. Dabei handelt es sich eher um Schnappschüsse, die eine deutlich geringere Auflösung haben, unscharf sind,

zu viel oder zu wenig beleuchtet wurden,

aus untypischen Winkeln geschossen wurden,



das Zielobjekt nicht in der Mitte haben und so weiter und so fort.

Einblendung von typischen Schnappschüssen

Datenvielfalt

Wichtig ist auch die Vielfalt der Daten. Falls ein Modell z. B. die Klassenzugehörigkeit von Daten vorhersagen soll, müssen ausreichend Trainingsdaten für jede Klasse vorhanden sein. Ein Modell kann keinen Farn erkennen, wenn es nicht mit ausreichend Beispielen dafür trainiert wurde. Außerdem ist es wichtig, dass auch die Vielfalt innerhalb einer Klasse abgedeckt wird. Wenn unser Modell, das Hunde erkennen soll, nur mit Bildern von schwarzen Hunden trainiert wird, kann es zu falschen Vorhersagen kommen, wenn ein weißer Pudel erkannt werden soll.

Einblendung von Hunden

Natürlich ist es je nach Anwendung manchmal nicht möglich, genug „echte“ Trainingsdaten zu erhalten. Dafür gibt es oft Tricks, trotzdem ein ziemlich genaues Modell zu erhalten. Es gibt zum Beispiel die Möglichkeit, die „echten“ Trainingsdaten mit anderen zu mischen, also

in unserem Beispiel Fotografenfotos, künstlich erzeugte Fotos und so weiter. Bitte recherchiert vorher, ob so ein Vorgehen in eurem Anwendungsfall geeignet ist.

Abschluss

In diesem Video habt ihr den Zusammenhang zwischen guten Trainingsdaten und der Genauigkeit der Vorhersagen eures Modells kennengelernt. Ihr könnt jetzt Beispiele für gute und schlechte Verteilungen von Trainingsdaten angeben und begründen, wieso eine gegebene Verteilung von Trainingsdaten besser oder schlechter für eine Anwendung geeignet ist als andere.

Weiterführendes Material

https://en.wikipedia.org/wiki/Data_quality

Disclaimer