Was sind Daten?

Die Rolle von Daten in einem KI-System

Im allgemeinen digitalen Sinne entsprechen Daten den Informationen, die von der Software in einem Computersystem verwendet, verarbeitet und erzeugt werden.

Es gibt keine KI ohne Daten. Daten spielen bei allen Prozessen des maschinellen Lernens eine zentrale Rolle, da sie sowohl für das Training als auch für die Tests verwendet werden. Sie werden auch in Form von Parametern für die Verwaltung der Trainingsprozesse verwendet. Schließlich ist das KI-System eine Kombination aus einer bestimmten Softwarearchitektur mit allen trainierten Parametern, dem so genannten Modell, das ebenfalls aus Daten besteht.

Das Verständnis der Rolle der Daten in KI-Systemen sowie der Art und Weise, wie sie ausgewählt, dokumentiert und verbreitet werden, ist für die Beurteilung des Verhaltens eines KI-Systems von wesentlicher Bedeutung. Dies ist wichtig für die Reproduzierbarkeit oder für den Vergleich zweier verschiedener KI-Systeme.

In der Welt der Verarbeitung natürlicher Sprache beispielsweise ist die Verfügbarkeit großer Mengen gesprochener und geschriebener Daten von zentraler Bedeutung für die Leistung von Rechtschreibprüfungen, Vorhersagen in Suchmaschinen oder natürlich für die maschinelle Übersetzung. Sie werden verwendet, um sogenannte Sprachmodelle zu erstellen, die weitere Prozesse mit statistischen Darstellungen von Wort- oder Satzkombinationen füttern.

Die Nachhaltigkeit von KI-Systemen hängt also eng mit den Methoden der Datenverwaltung zusammen, die bei ihrer Entwicklung eingesetzt werden.

Daten für überwachte und unüberwachte KI-Systeme

Wie wir bereits gesehen haben, gibt es zwei Arten von KI-Systemen, je nachdem, wie die Daten für ihr Training verwendet werden. Überwachte Systeme beruhen auf der Bereitstellung von Eingaben zusammen mit den entsprechenden beabsichtigten Ausgaben. Das Training besteht also darin, dem System beizubringen, aus unbekannten Eingaben die wahrscheinlichste Ausgabe zu erzeugen. Es gibt verschiedene Möglichkeiten, solche Daten zu erhalten. Zum Beispiel eine Bilddatenbank, in der jedes Bild mit Schlüsselwörtern verknüpft ist, oder eine Sammlung digitalisierter Dokumente, die von Annotatoren transkribiert wurden (siehe Abbildung unten).

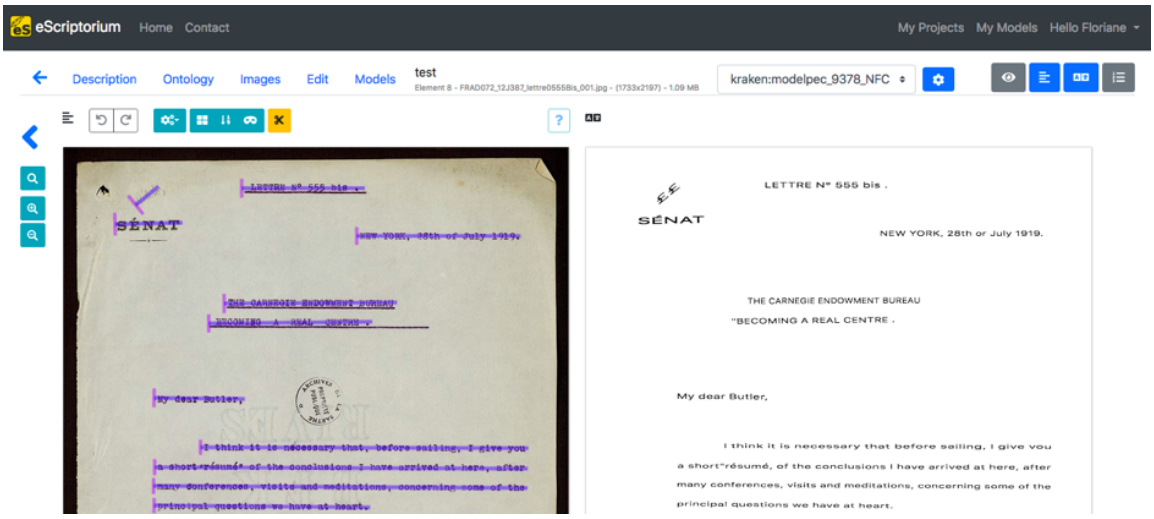

Abbildung: Automatische Transkription eines Briefes von Paul D'Estournelles (mit freundlicher Genehmigung von F. Chiffoleau, Coll. Archives de la Sarthe)

KI-Systeme, die auf unüberwachtem Training basieren, werden nicht für ein bestimmtes Verhalten entwickelt, sondern um die statistischen Eigenschaften der Trainingsdaten zu berücksichtigen. Dies ist beispielsweise bei Sprachmodellen wie BERT der Fall, die dazu neigen, Wörtern mit dem gleichen syntaktischen oder semantischen Verhalten ähnliche Positionen in einem mathematischen Raum zuzuordnen, was durch die Bereitstellung einer großen Anzahl von Beispielsätzen für jedes Wort beobachtet wird. Solche Modelle sind zum Beispiel sehr gut in der Lage, Synonyme oder die Folgewörter einer bestimmten Sequenz vorherzusagen.

Quellen - Auswahl, Dokumentation, Aufbereitung, Annotation

Das Design eines KI-Systems basiert im Wesentlichen auf dem richtigen Design des Datensatzes, mit dem es trainiert wird. Zu den verschiedenen Faktoren, die dabei eine Rolle spielen, gehören die Relevanz der Daten für die jeweilige Aufgabe, der Umfang der Daten, der der Komplexität der KI-Softwarearchitektur entsprechen sollte - je mehr mathematische Parameter zu trainieren sind, desto mehr Daten werden benötigt -, und die Vielfalt der Stichproben, die die Komplexität der Aufgabe widerspiegeln sollte.

Abhängig von den Datenquellen müssen die Daten ausgewählt und oft bereinigt werden, bevor sie in den Trainingsprozess eingespeist werden. Nehmen wir zum Beispiel das Training eines Sprachmodells auf der Grundlage von Webinhalten, so müssen die verschiedenen Stichproben nach der tatsächlichen Sprache sortiert, vom begleitenden Webcode (HTML, Javascript usw.) befreit und möglicherweise gemischt werden, um Urheberrechtsverletzungen zu vermeiden. Ein gutes Beispiel für eine solche Datenaufbereitung ist der Aufbau des OSCAR-Korpus[1].

Der Entwurf von annotierten Daten für überwachte KI-Systeme ist komplexer, da er den Entwurf eines Annotationsschemas, die Organisation von Annotationskampagnen und die

Kontrolle der Qualität der annotierten Daten umfasst, z. B. durch Bewertung der Übereinstimmung zwischen Annotatoren bei denselben Daten.

Alles in allem ist es wichtig, dass der Entwurfsprozess gut dokumentiert wird, um die Quelle möglicher Fehlverhaltensweisen im resultierenden trainierten System zurückverfolgen zu können.

Voreingenommene KI-Systeme

Wie bereits angedeutet, spiegelt das Verhalten eines KI-Systems die Art der Daten wider, die für sein Training verwendet wurden. Dies wiederum kann zu einer ganzen Reihe möglicher Verzerrungen führen, die sich aus der zugrundeliegenden Auswahl der Datensätze ergeben können. So wird beispielsweise ein Sprachmodell, das nur auf Zeitungsartikeln trainiert wurde, ganz andere Arten von Ausdrücken und Themen abdecken als eines, für das Inhalte aus der Literatur oder sozialen Netzwerken ausgewählt wurden. In gleicher Weise spiegeln Bilderzeugungssysteme die Größe und Vielfalt der betrachteten Quellbilddatenbanken (z. B. künstlerische Werke) wider.

Im Falle überwachter Systeme kann eine spezifische Verzerrung aus der Art und Weise resultieren, wie die Beschriftungsetiketten gestaltet sind, sowie aus der Art und Weise, wie die Annotatoren mit ihrem eigenen kulturellen Hintergrund die Daten interpretieren werden. Wenn Sie beispielsweise Hassreden in sozialen Netzwerken identifizieren wollen, kann die Art und Weise, wie die Kommentatoren die Gefühle interpretieren, je nach Alter, Kultur und persönlichen Gefühlen der Kommentatoren gegenüber dem zu kommentierenden Material variieren.

Alles in allem sollte man immer im Hinterkopf behalten, dass KI-Systeme konstruktionsbedingt sehr konservativ sind, was ihre Trainingsdaten und damit die vorhandenen Beobachtungen betrifft. Von einem KI-System kann man keine wirkliche Kreativität erwarten.

Hosting, Pooling, Verteilung von Daten

Aufgrund des Umfangs und der möglichen Komplexität der Trainingsdaten in KI-Systemen sowie der daraus resultierenden Modelle wurden verschiedene Initiativen ins Leben gerufen, um deren Hosting und Verteilung zu ermöglichen.

Offene Datensätze und Modelle können in speziellen Repositories (z. B. die Bilddatenressource[2]) oder in allgemeinen nationalen oder internationalen Repositories (z. B. Zenodo[3]) gehostet werden. Solche Repositorien bieten in der Regel die notwendige Infrastruktur für die Verwaltung der Urheberschaft, Lizenzierung, Versionierung und Archivierung ihrer Inhalte.

Im Falle komplexer Aufgaben, bei denen verschiedene Teams parallel an der Annotation einer Vielzahl von Datenproben arbeiten, fungieren einige Initiativen als Kataloge für die

entsprechenden Datenquellen. Dies ist zum Beispiel bei der Initiative HTR United[4] der Fall, die Metadaten von annotierten Dokumenten für die (handschriftliche) Texterkennung sammelt.

---

1. OSCAR corpus: https://oscar-corpus.com/ ↩
2. Image data resource: https://idr.openmicroscopy.org/ ↩
3. Zenodo: https://zenodo.org/ ↩
4. HTR United: https://htr-united.github.io ↩