KI für Alle 2: Verstehen, Bewerten, Reflektieren

Themenblock Bildklassifikation und Bildsegmentierung: 07_03Annotation_GoldStandard

Datenannotation

Erarbeitet von
Dr. Ludmila Himmelspach

Lernziele

- Du kannst allgemeine Richtlinien für die Datenannotation erklären
- Du lernst, wie und wofür Inter-Annotator-Agreement berechnet wird
- Du kannst erklären, was „Gold Standard“ im Zusammenhang mit Datenannotation bedeutet und wie dieser erreicht wird

Projekt ai4all „KI für Alle”
HeiCAD –Heine Center for Artificial Intelligence and Data Science

Inhalt

Einstieg

Überwachte Lernverfahren, zu denen Convolutional Neural Networks (CNNs) zählen, werden auf annotierten Daten, also auf Daten mit dazugehörigen Labeln trainiert und evaluiert. Von der Qualität der Annotationen hängt die Leistungsfähigkeit des gesamten Systems ab, denn wie es so schön heißt „Garbage in, garbage out“. In diesem Video lernst Du, welche Schritte notwendig sind, um einen gut annotierten Datenkorpus zu erstellen.

Richtlinien für die Datenannotation

Zu Illustrationszwecken arbeiten wir in diesem Themenblock mit einem öffentlich verfügbaren Datensatz, der bereits annotiert ist.

Quelle [1]

Natürlich kennen wir alle Einzelheiten des Annotationsprozesses dieses konkreten Datensatzes nicht. Wir fokussieren uns in diesem Video auf die Beschreibung des Annotationsprozesses so, wie er im Regelfall ablaufen soll.

Um die Verzerrungen der Ergebnisse zu vermeiden, sollte nicht nur auf eine repräsentative und ausgewogene Datenbasis geachtet werden, mit der ein Klassifikationsmodell trainiert werden soll, sondern auch die Annotation des Datensatzes sollte möglichst unabhängig von dem zu entwickelnden System erfolgen. Deswegen sind die Annotator*innen am besten so zu wählen, dass sie zwar genug Kompetenzen besitzen, um einen gegebenen Datensatz optimal zu annotieren, an der Entwicklung der Klassifikationsmethode selbst nicht beteiligt werden sollten.

Auch bei der Entwicklung eines Systems zur automatisierten Erkennung der Lungenentzündung sollen die Röntgenaufnahmen vom Brustkorb am besten durch unabhängige Expert*innen annotiert werden, die in die Entwicklung des Klassifikationsmodells nicht einbezogen werden. Denn andernfalls könnten sie beeinflusst durch ihre Kenntnisse über die Schwächen der Klassifikationsmethoden ihre Annotationen dementsprechend anpassen.

Da selbst die besten Expert*innen auf dem jeweiligen Gebiet bei der Bewertung der Datensätze Fehler machen können, sollten mehrere Annotator*innen die gleiche Datenbasis unabhängig voneinander kennzeichnen.

Quelle [2]

Um die Annotationen einheitlich zu halten und die Einarbeitungszeit der Annotator*innen zu verkürzen, sollen Annotationsrichtlinien erarbeitet werden. Diese sollen anfangs vor allem festhalten:

- Was genau wird annotiert?
- Was sind die Annotationskategorien und wie werden diese gekennzeichnet?

Bei der Annotation der Röntgenaufnahmen von Brustkorb können die Annotator*innen in der Tabellenspalte „Lungenentzündung“ ihre Entscheidung zum Beispiel mit „ja“ oder „nein“ angeben.

In der ersten Annotationsphase soll nur ein Teil des Datensatzes vorläufig annotiert werden. In dieser Zeit sammeln die Annotator*innen ihre ersten Erfahrungen mit der Annotation des Datensatzes und notieren sich, welche Schwierigkeiten und Probleme sie während des Annotationsprozesses hatten. Nach einem Gespräch mit allen Annotator*innen werden die Annotationsrichtlinien angepasst und verbessert. Bei der Aktualisierung der Richtlinien werden unter anderem auch die Übereinstimmungen der Annotationen unter den Annotator*innen berücksichtigt. Außerdem wird an dieser Stelle auch der Umgang mit Ausreißern festgehalten, falls diese im Datensatz enthalten sind. Wenn alle Einzelheiten geklärt sind, kann der gesamte Datensatz ohne weitere Änderungen der Richtlinien annotiert werden. Kommt es im Annotationsverlauf zu weiteren Fragen oder Problemen, müssen die Annotationsrichtlinien erneut angepasst und der Annotationsprozess neu gestartet werden.

Quelle [3]

Wie man sieht, ist der Annotationsprozess sehr ressourcen-, zeit- und dementsprechend kostenintensiv. Andererseits merkt man erst während der Annotierung des Datensatzes den wirklichen Schwierigkeitsgrad der gestellten Klassifikationsaufgabe. Außerdem kann es während der Annotation zur Bereinigung des Datensatzes kommen, wenn zum Beispiel Annotator*innen auf Datensätze stoßen, die auf Grund ihrer Qualität oft trifft das auf die Bilder zu selbst durch Expert*innen nicht analysiert werden können.

Inter-Annotator-Agreement

Während und im Anschluss an den Annotationsprozess müssen die Annotationsunterschiede bzw. -übereinstimmungen quantitativ bestimmt werden können. Je nach Aufgabentyp werden dafür unterschiedliche Metriken benutzt. Für eine Klassifikationsaufgabe wird oft *Cohens Kappa* für die Berechnung des *Inter-Annotator-Agreements* eingesetzt.

$$\kappa = \frac{A_O - A_E}{1 - A_E}$$

Quelle [4]

Diese Metrik berücksichtigt die Schwierigkeit der Aufgabe, indem die von den Annotator*innen erreichte Übereinstimmung A_O mit einer zufälligen Übereinstimmung A_E verrechnet wird. Die zufällige Übereinstimmung hängt von der Anzahl der Klassen bzw.

Kategorien und von ihren Vorkommenshäufigkeiten im Datensatz ab. Es ist klar, dass eine zufällige Übereinstimmung bei zwei Klassen höher als bei zehn Klassen ist.

Um die Werte von Cohens Kappa Metrik miteinander vergleichbar zu machen, ist sie so normalisiert, dass sich ihr Wert bei voller Übereinstimmung der Annotator*innen 1 ergibt.

Quelle [5]

Wenn die von den Annotator*innen erreichte Übereinstimmung niedriger als die zufällige Übereinstimmung ist, so erreicht Cohens Kappa negative Werte. Im Allgemeinen deutet ein Wert zwischen 0,81 und 1,0 auf eine fast vollkommene Übereinstimmung hin.

Quelle [6]

Gold Standard

Bei der Annotation eines Datensatzes durch mehrere Annotator*innen, erhält man zu jedem Datenobjekt dementsprechend mehrere Kennzeichnungen, die nicht immer übereinstimmen. Die meisten überwachten Machine Learning Methoden erwarten jedoch zu jedem Datenobjekt nur ein Label. Um dies zu erreichen, hat sich in der Praxis folgende Vorgehensweise etabliert: bei unterschiedlichen Klassenzuweisungen eines Datenobjekts wird entweder die Meinung eines weiteren erfahrenen Annotators bzw. einer weiteren erfahrenen Annotatorin hinzugezogen oder die Annotation des höchstrangigen Annotators bzw. einer höchstrangigen Annotatorin für richtig anerkannt. Oft bezeichnet man ein gut annotiertes Datenkorpus mit einer hohen Übereinstimmung der Annotationen mehrerer Annotator*innen nach Beseitigung aller Diskrepanzen als *Gold Standard*.

Quelle [7]

Take-Home Message

Die Annotationen eines Datensatzes bilden die Grundlage für die Entwicklung eines überwachten Lernverfahrens. In diesem Video hast Du gelernt, welche Schritte notwendig sind, um einen Datenkorpus so zu annotieren, dass er sich Gold Standard nennen darf.

Quellen

Quelle [1] Kermany, D., Zhang, K., Goldbaum, M. (2018), “Labeled Optical Coherence Tomography (OCT) and Chest X-Ray Images for Classification”, Mendeley Data, V2, doi: 10.17632/rscbjbr9sj.2

Quelle [2] Wissler, L., Almashraee, M., Díaz, D. M., & Paschke, A. (2014). The Gold Standard in Corpus Annotation. *IEEE GSC*, *21*.

Quelle [3] Reiter, N. (2020). Anleitung zur Erstellung von Annotationsrichtlinien. *Nils Reiter/Axel Pichler/Jonas Kuhn (Hg.), Reflektierte Algorithmische Textanalyse. Interdisziplinäre (s) Arbeiten in der CRETA-Werkstatt, Berlin/Boston*, 193-201.

Quelle [4] Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, *20*(1), 37-46. https://doi.org/10.1177/001316446002000104

Quelle [5] Reiter, N., & Konle, L. (2022). Messverfahren zum Inter-Annotator-Agreement (IAA).

Quelle [6] Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, *33*(1), 159–174. https://doi.org/10.2307/2529310

Quelle [7] Medero, J., Maeda, K., Strassel, S. M., & Walker, C. (2006, May). An Efficient Approach to Gold-Standard Annotation: Decision Points for Complex Tasks. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Disclaimer