Projektbericht zum Modul Data Mining Wintersemester 20221/2022

Reproduktion des Papers *Context-Sensitive Visualization of Deep Learning Natural Language Processing Models*[1]

Max Henze

4. März 2022

1 Einleitung

Neuronale Netzwerke sind ein beliebtes Hilfmittel im Bereich von NLP. Besonderst Modelle, welche sich den Einsatz von Transformern zur Hilfe nehmen, wie BERT [2] oder GPT-2 [3], gehören schon lange zum state-of-the-art. Doch diese Modelle mit ihrer Vielzahl an Layern, Neuronen und Verbindungen, gewähren einen nicht gerade einfachen Einblick in ihre Verarbeitungsschritte. Dunn et al. haben daher in ihrem Artikel „Context-Sensitive Visualization of Deep Learning Natural Language Processing Models“ eine Methode entwickelt um Wörter, mit ihrer unterschiedlichen Wichtigkeitsgewichtung innerhalb der Klassifizierung, zu visualisieren. Als Replikationsziele wurden für diese Arbeit der gesamte Visualisierungsprozess von Dunn et al. gewählt. Zusätzlich dazu wurde ein eigenes BERT Modell auf dem gegebenen IMDB [4] trainiert um dieses für den späteren Klassifikationsprozess zu verwenden. Durch die Replikation wird eine verständliche Codebeigabe zum Originalartikel erzeugt, welche dem Leser ein noch besseres Verständnis liefern soll. So können mit dem System alle Dokumente des Testdatensatzes klassifiziert und visualisiert werden um so eine größere Vielfalt von Beispielen bereitzustellen.

2 Umfang der Replikation/Reproduktion

Als Ziel dieser Replikation wurde die einzige Hypothese gewählt, welche Dunn et al. in ihrem Artikel behandeln. Sie propagieren, dass sich die Wichtigkeit eines Wortes innerhalb der Klassifi-

kation durch ein Neuronales Netzwerk nicht nur durch den Vergleich der sogenannten Prediction Strength (Sicherheit des NN, dass Label richtig Klassifiziert) des Originaltextes zur Pediction Strength des Textes ohne das betrachtete Wort (leave-one-out) ergibt, sondern dass der Einbezug von kontextuell zusammhängenden Wörtern (leave-n-out) ebenfalls wichtig ist. „Unser Ansatz schaut auf die Kombination von Wörtern und Sätzen um deren Einfluss auf die Ausgabe des Modells zu erkennen, was zu einer Visualisierung führt, welche kontextsensitiver zum Originaltext ist. “ [1]

Somit ist folgende Behauptung das Ziel dieser Replikation:

- Leave-n-out Ansatz ist effektiver bei der Erkennung wichtiger Wörter als leave-one-out.

3 Methoden

Die Replikation des Originalartikels ergibt sich wie folgt. Durch die fehlende Beigabe von Code mussten alle Ideen und Modelle von Dunn et al. eigenständig implementiert werden. Dazu wurde sich an Wortangaben der Autoren wie zum Beispiel: „Der gesamte Code ist geschrieben in Python 3.8 und nutzt die Tensorflow Version der Transformersbilbiothek. [...] Texttokenisierung und Abhängigkeitsbestimmung wurden mit der spaCy NLP Bibliothek durchgeführt.“ [1]

Zur Klassifizierung von Dokumenten wurde ein Modell unter der Verwendung von BERT trainiert. Da keine weitere Angaben zu finden waren und eine große Auswahl an unterschiedlichen BERT Modellen zu finden ist wurde das *BERT uncased L-12 H-768 A-12* Modell gewählt, welches auf Tensorflow Hub[1] zu den am häufigsten verwendet BERT Modellen zählt.

Entwickelt wurde innerhalb eines Jupyter Notebooks mit Python. Die folgenden essentiellen Packages fanden dabei Anwendung:

Package Name	Package Funktion
tensorflow_hub	Einbindung des BERT Modells
tensorflow	Modellerzeugung und Training
official.nlp	Trainingsoptimisierung
spacy	Abhängigkeitsbestimmung (Dependency Parsing)
pandas	Arbeiten mit Dataframes
matplotlib	Visualisierung der Texte

Zusätzlich wurde das BERT Modell auf einer Nvidia Geforce RTX 3070 mit 8 GB Arbeitsspeicher trainiert.

[1] `https://tfhub.dev`

3.1 Modellbeschreibung

Innerhalb des Originalartikels sind keine Angaben bezüglich der Zielfunktion und Parameter zu finden. Angaben zum Modell beruhen auf der Benennung eines BERT Modells und einer Modellbeschreibung, welche auf das Anhängen eines Dropout-Layers und Dense-Layers verweist.

Die beschriebene Methodik ist wie folgt:

Ein Text wird durch das Modell klassifiziert und die damit korrespondierende Ausgabestärke wird notiert. Nun werden mit Hilfe einer Abhängigkeitsbestimmung alle Beziehungen zwischen Wörtern aufgedeckt. Anschließend werden neue Texte erzeugt, in denen jeweils ein Wortpaar, welches eine Verbindung zueinander aufweist, entfernt wurde. Die nun erhaltene Sammlung an neuen Texten wird wieder durch das Modell klassifiziert und die neuen Ausgabestärken werden mit der des Ausgangstextes verglichen. Texte mit größeren oder gleichen Ausgabestärken als der des Originals tragen scheinbar nicht zur Klassifikation bei und werden entfernt. Je größer die Differenz umso wichtiger war das Wortpaar für die Klassifizierung. Mit Hilfe einer Linearisierung der Differenzen und einer Colormap können Wörter smomit bezüglich ihrer Wichtigkeit farblich kenntlich gemacht werden. Je wichtiger umso grüner, je unwichiger, desto blauer.

3.2 Datenbeschreibung

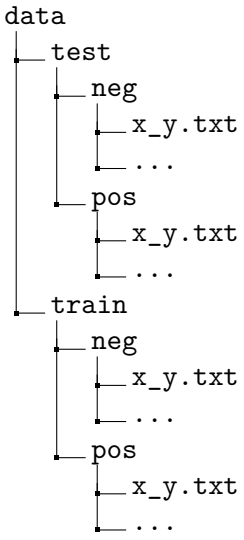

Abbildung 1: Ordnerstruktur des Datensatzes. x ist die Dokumentenid und y ist eine Sternewertung von Null bis Zehn.

Der im Originalartikel und dieser Replikation verwendete Datensatz ist das Large Movie Review Dataset [4] der Universität Stanford. Dieser umfasst 50.000 Dokumente, darunter 25.000 Trainingsdokumente und 25.000 Testdokumente. Er ist unter `https://ai.stanford.edu/~amaas/`

`data/sentiment/` verfügbar.

Der Datensatz hat eine vorgegebene Ordnerstruktur, siehe Abbilung 1. So befinden sich die Trainingsdokumente und Testdokumente in eigenen Ordnern, wobei positive und negative Dokumente nochmals in eigene Ordner unterteilt sind. Die Dokumente unterscheiden sich stark in der Länge, so gibt es Dokumente mit knapp über 50 Zeichen aber auch solche mit über 13.000 Zeichen. Die Dokumente an sich sind nicht aufbereitet, enthalten englische Altagssprache und Sonderzeichen.

```
Fair drama/love story movie that focuses on the lives of
blue collar people finding new life thru new love.The acting
here is good but the film fails in cinematography,screenplay,
directing and editing.The story/script is only average at best.
This film will be enjoyed by Fonda and De Niro fans and by
people who love middle age love stories where in the coartship
is on a more wiser and cautious level.
It would also be interesting for people who are
interested on the subject matter regarding illiteracy.......
```

Abbildung 2: Beispieltext eines positiven Trainingsdokuments

Bei der Verwendung der Daten zum Training des Modells, wurde der Trainingsdatensatz zusätzlich in einen Validierungsdatensatz aufgeteilt. Dieser umfasst 20 Prozent der Trainingsdaten und somit 5.000 Dokumente.

3.3 Hyperparameter