Machine Translation Übung 4

Lukas Yu 14-720-866

29. April 2019

1 Datenset

Das Datenset besteht aus positiven IMDb Rezensionen. Ich wollte versuchen, wie gut ein Modell typische Bewertungstexte generieren kann. Die einzelnen Texte wurden in einem einzigen File zusammengefügt, die Sätze nach Zeilenumbrüche getrennt und auf Wort-Level tokenisiert. Zusätzlich wurden alle <**br** /> tags entfernt. insgesamt sind es 132965 Sätze, die in einem 17 Megabyte grossen Textfile (dataset.txt) abgespeichert sind.

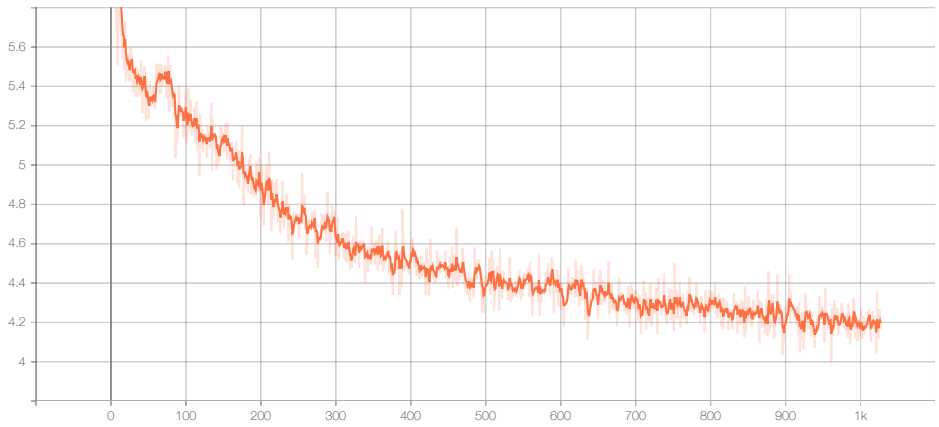
Das Datenset habe ich dann wie in der Übungsaufgabe beschrieben in 2 Teile (Train und Dev) gesplittet.

Datenquelle: `https://ai.stanford.edu/~amaas/data/sentiment/` (Alle Textdateien aus ./train/pos/*)

2 Erstes Training / Scoring

Das erste Training wurde mit der Default-Konfiguration durchgeführt. Es daurte ungefähr eine Stunde, um das Modell zu erstellen.

Perplexity auf Devset: 76.77



3 Adaption

Um die Resultate zu verbessern, versuchte ich zuerst, das Preprocessing zu optimieren. Die Gross- und Kleinschreibung wurde deswegen auf die Kleinschreibung normalisiert.

Zustätzlich veränderte ich Hyperparameter. Die Vokabulargrösse vergrössterte ich auf 30'000, damit das Modell eine vielfältigeres Vokabular besitzt.

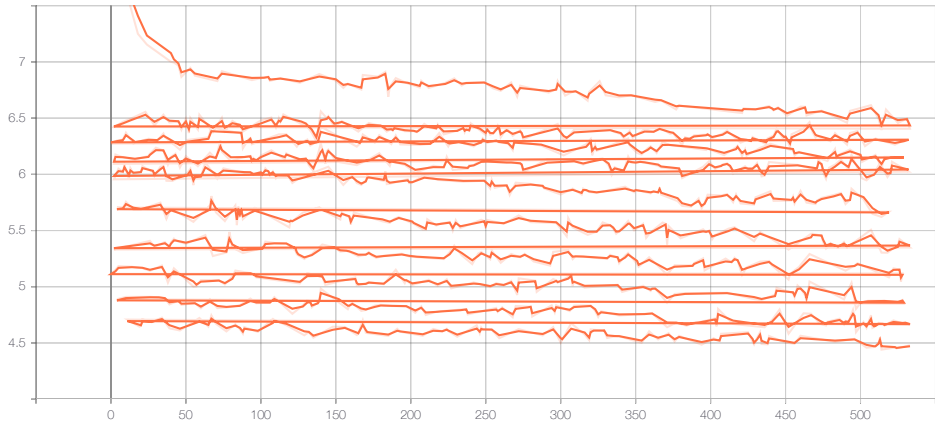
Die LSTM-Cells wurden durch Gru-Cells ausgetauscht, weil das Training angeblich schneller geht.

4 Zweites Training / Scoring

Das Training ging ungefähr zwei Stunden. Also etwa doppelt so lange verglichen mit dem ersten Training. Dies kann man sich der erhöhten Vokabulargrösse erklären.

Perplexity auf Devset: 146.23

Den hohen Perplexity-Wert kommt wahrscheinlich daher, dass LSTM-Cells für diesen Anwendungsfall besser geeignet sind im Gegensatz zu GRU-Cells.



5 Sampling

siehe sample.txt

6 Modell-Schema

siehe model_schema.png