Maschinelle Übersetzung Übung 05

Nicolas Spring

19. Mai 2018

1 Dateien

Der Ordner *Abgabe* enthält diesen Bericht sowie das übersetzte Testset.

2 Preprocesssing

Mein Preprocessing bestand daraus, mit Hilfe des Mosesdecoders die Texte zu normalisieren und zu tokenisieren. Danach habe auf den Trainingsdaten für beide Sprachen je ein Truecasing-Modell trainiert, welches ich dann wiederum auf die Texte anwandte. Danach habe ich it meine beiden Traingssets zusammengefügt und mit subword-nmt ein gemeinsames BPE-Modell gelernt für 70'000 Symbole, welches ich auf alle Texte anwandte.

3 Hyperparameter

Der einzige Hyperparameter, den ich abgeändert habe, war die Epochenanzahl die ich auf 6 verringerte. Ich konnte aber nicht die vollen 6 Epochen trainieren.

4 Code-Veränderungen

Bei der Übung 4 habe ich auch schon versucht, zusätzliche LSTM-Zellen einzubauen. Jedoch bin ich mittlerweile der Meinung, dass ich krass in Overfitting hineinrannte, weil mein Gedanke, man müsse die Epochenanzahl mit erhöhen, wohl nicht umbedingt klug war. Deshalb wollte ich in dieser Übung das Ganze nochmals probieren, zumal das Stacking von LSTM-Zellen in den Ideen war. Ich stackte also im Encoder und Decoder je 2 LSTMs. Implementiert habe ich es wieder wie letztes Mal gemäss TensorFlow-Dokumentation.
Die längere Trainingszeit, die mit den zusätzlichen LSTMs einherging wurde bei dieser Übung zu einem Problem. Bei beiden Neustarts war lief mein Training und ich musste darum jedes mal wieder damit von Neuem beginnen, was natürlich umso ungelegener kam, da ich mit meinen ca. 10 Minuten pro 1000 Iterationen für ein vollständiges Training mit 6 Epochen ca. 3.5 Tage gebraucht hätte. Dies kam natürlich nicht mehr wirklich in Frage. Ich brach also mein Training, als sich Ende der Woche alle GPUs füllten, bei 4 Epochen ab und arbeitete in der Folge für die Übersetzung des Testsets mit diesem Modell.