

Sentiment-Analyse mit LSTMs

Laden Sie die Sentimentanalyse-Daten an der Adresse <http://www.cis.uni-muenchen.de/~schmid/lehre/Experimente/data/sentiment-data.zip> herunter und trainieren Sie damit einen Klassifizierer auf Basis von LSTMs. Die Zahl vor dem Tabulator gibt jeweils die Bewertung des nachfolgenden Satzes an, die das System lernen soll.

Für das Einlesen der Daten und die Generierung der Batches verwenden Sie das **Torchtext**-Modul (mit einem *BucketIterator*). Sie können den Code aus den beiden Torchtext-Tutorials als Vorlage nehmen und anpassen.

Im **Training** minimieren Sie die negative Loglikelihood (mit dem CrossEntropyLoss von PyTorch). Nach jeder Epoche **evaluieren** Sie das System auf den Entwicklungsdaten und geben die erzielte Genauigkeit aus.

Vorüberlegungen

- Welche Ebenen sollte Ihr neuronales Netz umfassen?
- Wie sollte das Vokabular definiert werden?
- Wie vermeiden Sie Overfitting?

Schicken Sie mir Ihr Programm und die Ausgabe mit den erzielten **Genauigkeiten**.

Technische Hinweise:

Installieren Sie zunächst PyTorch von der Seite <https://pytorch.org/>.

Wenn Ihr Rechner eine Nvidia-Grafikkarte besitzt, können Sie auch darauf arbeiten. Sie müssen dann aber noch zunächst CUDA von Nvidia installieren.

Achtung: Wenn Sie sich remote auf einem CIP-Pool-Rechner einloggen, sollten Sie mit dem Befehl `ssh -X rechner` anschließend auf einen Rechner "rechner" mit 32 GB Arbeitsspeicher einloggen. Die Rechner, die im Antarktis-Pool stehen, haben alle 32 GB.

Bevor Sie ein PyTorch-Programm remote auf einem CIP-Pool-Rechner starten, sollten Sie sicherstellen, dass kein anderer eingeloggt ist. Am besten starten Sie Rechenjobs remote mit Hilfe von Slurm (https://www.rz.ifi.lmu.de/infos/slurm_de.html). Dann sucht das System selbst einen der verfügbaren Rechner aus.

Der Befehl `nvidia-smi` sagt Ihnen, was gerade auf Ihrer GPU gerechnet wird.