Universität Zürich, Institut für Computerlinguistik
Maschinelle Übersetzung Vera Bernhard
Mathias Müller 13.05.2019

MT Übung 5: Encoder-Decoder-Modelle

Preprocessing:

- Normalisierung
- Tokenisierung
- Cleaning
- Truecasing
- BPE-Modell: gemeinsam für Ziel- & Quellsprache mit Vokabulargrösse 30'000
- Sequenzumkehrung für Quellsprache (s. Skript `change_seq_order.py`)

Hyperparameter:

Die Hyperparameter habe ich mehrheitlich gleich gelassen und mich bei dieser Übung auf Veränderungen intern fokussiert:

`Source/Target Vokabular = 50'000`
→ die Grösse des Source-Vokabular wahr nur rund 20'000, daher sah ich keinen Grund das Vokabular einzuschränken oder zu erhöhen.

```
EMBEDDING_SIZE = 512
HIDDEN_SIZE = 1024
```
→ das Training ging damit einigermassen flott und darum hab die mal nicht verändert

`LEARNING_RATE = 0.0001`
→ Nach Tensorboard ist der Loss stetig gesunken, darum habe ich es auch so belassen

`EPOCHE = 10`
→ Da ich Early Stopping eingebaut habe, wird geschaut, das es nicht zu stark overfittet und darum habe ich die Epochenzahl erhöht.

Code-Veränderungen:

- **RMSprob** anstatt Adam-Optimizer: In der Deep Learning Vorlesung wurde dieser Optimizer für alle Sorten RNNs empfohlen und darum hab ich ihn hier angewendet.
- **Dropout** auf Encoder & Decoder mit Keep Rate von 0.5: um Overfitting zu verhindern und 0.5 sei anscheinend so eine Standardrate gemäss Internet
    - Schwierigkeit: Es gib in Tensorflow zwei verschiedene Dropout-Layers (tf.nn.dropout & tf.nn.rnn_cell.DropoutWrapper) und ich bin zuerst nur auf erstere gestossen und habe diese versucht einzubauen. Jedoch hat es nicht geklappt und nur unaussagende Fehlermeldungen ausgegeben und ich bin dann mer aus Zufall auf die andere Layer gestossen, welche sofort funktioniert hat. Dies hat ziemlich Zeit gekostet.
- **Early Stopping** mit Patience = 3 auf dem Dev-Set: ebenfalls um Overfitting zu verhindern
    - Schwierigkeiten: Um die Perplexität auf dem Dev-Set zu überprüfen, habe ich für das Dev-Set einen neuen Computation Graph initialisiert. Das gab jedoch immer einen Error, nämlich dass die Input Word Embeddings für den Encoder auf dem Trainingset nicht als Tensor interpretiert werden können. Das habe ich überhaupt nicht verstanden, da ich an den Word Embedding vom Trainingset nichts verändert habe, in dem ich Early Stopping eingeführt habe. Mathias konnte dann aber diesen Fehler darauf zurückführen, dass beim Computation Graph Definieren, jeweils der zuvor definierte Computation Graph gelöscht wird. Das heisst der Graph vom Dev-Set kommt dem Graphen vom Training-Set in die Quere. Warum dies aber dazu führt, dass die Word Embeddings nicht richtig eingelesen werden können, ist mir ein Rätsel. Auf jeden Fall konnte ich das Problem so lösen, dass ich den Computation Graph fürs Dev-Set in einem eigenen Thread initialisiere. So scheint es zu funktionieren.
- **Bidirectional Encoding**:
    - Schwierigkeit: Das habe ich auch ausprobiert (s. auskommentierter Code), jedoch gab es auch eine Fehlermeldung, mit der ich nichts anfangen konnte (*Tensor objects are only iterable when eager execution is enabled.*). Ich hatte leider keine Zeit mehr, dem näher nachzugehen.

Allgemeine Schwierigkeiten:

Allgemein fand ich es wahnsinnig schwierig Tensorflow zu debuggen; die Error-Meldungen empfand ich als sehr schwammig.

Ausserdem hatte ich ein grösseres Problem mit dem Guthaben. Es war mir bereits seit Romanesco bekannt, dass das Instanz-Abschalten nicht immer sauber funktioniert. Z.T. sah es so aus, als ob sie aus ist, jedoch beim Refreshen der Seite war sie wieder an. Nun obwohl ich mir dieser Problematik bewusst war und auch x-fach gecheckt habe, ob die VM aus ist, hab ich mich nach mehrere Tagen, während derer ich den Server nicht benutzt hab, mit keinem Guthaben auf dem Server wiedergefunden. Und das war gerade dann, als ich das letzte grosse Training anwerfen wollte und die Zeit langsam knapp wurde. Ich habe dann kurz mit meinem eigenen Portemonnaie überbrückt bis ich grosszügiger Weise von Mathias einen neuen Code erhalten habe. Warum das Guthaben weg war, ist mir bis heute ein Rätsel.

Link zu meinem Repo: https://github.com/vera-bernhard/daikon.git