DL-HW3

Shai Vaknin (034658492), Ev Zisselman (200479483), Yochai Zur (03050991)

January 24, 2017

Github link: https://github.com/yochaiz/DL-hw

Branch: hw3

Architecture description

We tried using single LSTM layer, LookupTable for word embedding and a Linear layer on top of the LSTM. We also used Dropout with 0.3 proabability in between the Lookup layer and the LSTM and in between the LSTM and the Linear layer.

Training procedure

For regularization, we used dropout (0.3). No weight decay.

We tried several configuration of the model, the configuration parameters includes forgetting (or not) after each batch, and the sequence length. The results are described in the table below.

Model Configurations

Predictions:

We will show predictions of two interesting models. One is trained with forgetting and the other without. Note the perplexity of the model trained without forgetting is better, but still results are debate-able.

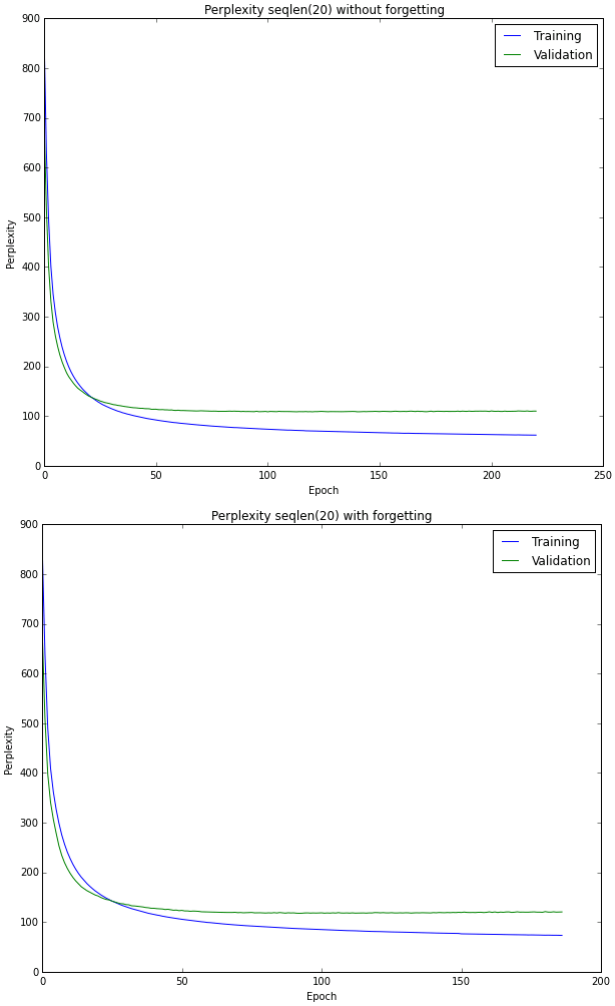
Table 1: Results for forgetting after every batch.

Sequence Length	**Perplexity**
6	131.4
7	127.6
20	117.6

Table 2: Results without forgetting after every batch.

Sequence Length	**Perplexity**
6	113.13
7	113.559
20	108.506

Figure 1: Test and train loss of our best models.



Model trained without forgetting (sequence length 20):

- "buy low sell high is the company 's largest business"
- "buy low sell high is the first time"
- "buy low sell high is the same time"
- "buy low sell high is the highest in the world"
- "buy low sell high is the value of the stock market"

Model trained with forgetting (sequence length 20):

- "buy low sell high is the highest level of the market"
- "buy low sell high is the first time to buy the company 's shares"
- "buy low sell high is the biggest stocks"
- "buy low sell high is the big board 's shares"
- "buy low sell high is the best of the company 's shares"

Summary:

Results without forgetting achieve the best perplexity, however, we notice that results with forgetting behave differently, and sometimes even better. Furthermore, after playing around with the models, we conclude that models that are trained with forgetting perform better in predicting short sequences.

Best perplexity we got is 108.506.