11-731: HW1: Attentional German-English Neural Machine Translation Model

Shuxin Yao(shuxiny)

1 Task Description

Create an attentional German-English neural machine translation model using the provided IWSLT data. Add at least one or two modifications over the baseline for full credit.

Send to the TAs your names, andrew IDs, and the link to a github repository containing code, output for the "test" and "blind" sets, and a report of 2-4 pages. The report should be named "report.pdf". The output should be tokenized and lowercased, and stored in the "output/" directory. The names of your primary results should be "output/test.primary.en" and "output/blind.primary.en", and should use only the provided IWSLT data for training the system. You can also submit additional outputs that use other methods, or use additional resources other than the IWSLT data listed below. In this case, replace primary by an arbitrary string, and we will calculate results for these as well.

2 Modifications

1. Gated recurrent unit (GRU)

In my implementation of the attentional German-English neural machine translation model, I used the Gated recurrent unit for encoder and decoder instead on simple RNN. Compared to simple RNN, GRU has a update gate and a reset gate and thus is more powerful. Compared to LSTM, it has less parameter than LSTM.

2. Mini-batch

I implemented the mini-batch to improve the speed of algorithm. To avoid the trouble of masking, I combining data that has the same length of source sentence and target sentence. Using a batch size of 64, the average size of each batch is around 42. And the training time for 40 epochs is 3 hour and 37 minutes.

3. Dropout

I used the set_dropout() and disable_dropout() functions to add dropout to the Gated recurrent units. Dropout is a good way to reduce overfitting. The probability is set to 0.5.

3 Details

Here I give some details of my implementation:

The hidden size for each uni-directional GRU is 128, so the bidirectional hidden size is 256.

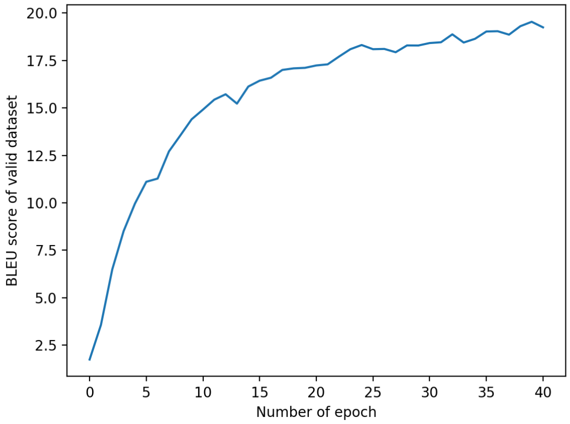
The embedding size is 128. The attention score calculation uses the dot product.

All words in target dataset that occur no more than 3 times are regarded as unknown words. The size of the target vocabulary is 18688.

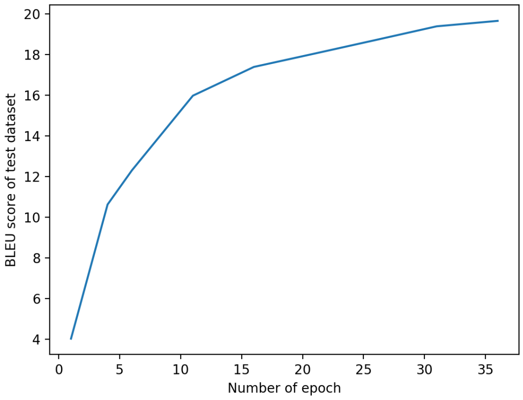
.

4 Results

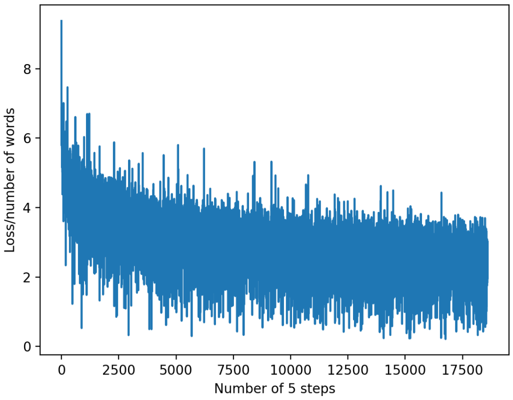
Here shows some graphs of my results. The first one is the BLEU score on the valid dataset for each epoch. The best result is **19.536**, which is achieved on epoch 39.



The second graph is the BLEU score on the test dataset for every five epochs. The best result is **19.658**, which is achieved on epoch 35.



The third graph shows the loss change during training.


Loss/number of words
Number of 5 steps