11-731 Assignment 1 Report

Zehao Guan (zehaog)

https://github.com/zach96guan/11-731_hw1_NMT

In this individual assignment, I implemented the attentional German-English neural machine translation model with the given IWSLT 2014 dataset.

- **Model**

For preprocessing work, I apply numerical vectors for words representation. For the model architecture, I use bidirectional LSTM for the encoder part, which effectively handles variable lengths of source sentences and is able to capture past and future information. The embedded source sentences will be passed into the encoder and return LSTM's outputs. Then, I refer to the attention mechanism in reference paper, to obtain the context vector as weighted sum of source words encoding vectors. For the decoder, the input is target sentences and relevant outputs from our encoder. I consider teacher forcing in each iteration to feed last generated value. And the decoder's hidden state and generated mask context from the attention after linear layer will be the output score.

- **Test**

For the decoding part, I utilize greedy search to select candidate word with the highest probability. After loading model and outputting the decoded sentences, I use given Perl script to calculate the BLEU scores.

- **Experiment**

After fixing the various bugs in codes, I run a series of models and select the best result based on the metrics perplexity and BLEU scores. I find teacher forcing, uniform initialization for model parameters and input feeding tricks will obviously improve the model performance. And adding dropout layers for encoder or attentional vector seems relatively trivial based on my experiments. For the best model I trained, training log file is listed in the "work_dir" directory and decoded outputs and scores are listed in the "output" directory.

Dev:

BLEU = 26.89, 62.3/35.2/21.4/13.3 (BP=0.958, ratio=0.959, hyp_len=123745, ref_len=129091)

Test:

BLEU = 25.06, 60.3/32.8/19.5/11.9 (BP=0.963, ratio=0.964, hyp_len=126395, ref_len=131141)

- **Hyperparameters**

In my experiment, I set a series of parameters as listed for model training.

batch size=128, hidden size=256, embed size=256, learning rate=1e-3, weight decay=1e-6,

teacher forcing rate=0.1, uniform model parameters initialization, …

- **Learn**

In this assignment, I learnt how to implement an end-to-end sequence-to-sequence model for machine translation system. As having taken the deep learning 11-785, I have some experience with designing the attentional model for speech transcription. I refer to that idea and use some tricks like teacher forcing, bidirectional LSTM to improve the model performance. In the process, I feel struggling with debugging for the decoder part, since I take lots of time to resolve the dimension errors and delve into the paper details to replicate. I wish to accumulate more seq2seq experience with the following assignments!