Deep Learning Assignment 4

Peter Yun-shao Sung yss265@nyu.edu

1 Warmup

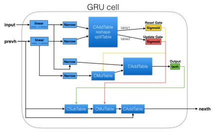

Figure 1: GRU cell unit

2 Approaches to RNN

2.1 Architecture

The initial model is based on lstm as building block, with 2 layers and 20 sequence length. Layers are horizontally considering the current given input (or word) and the states passed from previous sequence. Sequence is the stack of multiple layers, and is used for vertically passing states to the next layers. This way, model can not only predict the next word based on current word (horizontal), but the predictions can be altered based on the sequence of previous words (vertical). Regarding to each lstm layer, there are number of rnn_size lstm unit, and each of the unit perform the operation shown as figure 1.

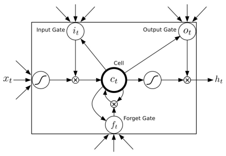

Figure 2: Operation of each of lstm unit

2.2 Learning Techniques

The two main learning techniques were the dropout and gradient clip. Dropout method was applied between each of lstm module and the final output. The inital module didnot used the dropout, and therefore we can observed that the perplexity of training improved well but not for the validation. This implies the concerns of overfitting and dropout might be a good option to improve.
Gradient clipping might be important method during the learning process too. Soft-clipping was initially applied to the model, which it observes the L2 norm of gradient and every gradient will multiply the portions it overpassed if the norm is higher than the threshold. The other method I used during this assignment is hard clipping, which we only cut the gradiend for only the one that pass threshold, instead of the every gradients. The inspiration is mainly from the curiousity of whether soft clipping downgrade the learning process for other gradient. However, as designing gridient checker and tracking perplexity, I noticed simply cut the gradient that overpassed might be problematic, as

there might be also many gradients need to cut and it turns out every gradients are at the boarder line, and this make generally high gradient and high perplexity.

2.3 Training Procedure

Data was preprocessed in the shape of (lines,batch_size), which is basically to traing a batch size of words. During each forward or backward propagation, line i and $i+1$ is the corresponding current and next word. When feed in network, the $next\_h$ is the output from the lstm module which is in the shape of (batch_size, vocab_size), and further calculates the prediction of each word along the batch_size axis by LogSoftMax. The perplexity is the accumulated error calculated by ClassNLLCriterion, and we are optimizing the perplexity error metric.

3 Improvement

The baseline perplexity is about 150, and here are the mothods I tryied to improve the prediction:

1. Clipping methods
   - The clipping method is hard clip as mentioned above, only cut the one that pass max gradient norm
2. GRUs
   - As it is a simpler method than lstm, here I include it for comparison as well
3. Hyperparameters
   - Includes: dropout rate, number of layers, soft/hard-clip, max gradient norm, and number of rnn units

4 Results

The first test is about dropping rate. As we can noticed, if it's too high then we cannot lower the perplexity any further, but if it's too low, it is effective to improve the perplexity on training set but not for the test set, which is the concern for overfitting. Therefore, seems the dropout rate at the range between 40% to 50% can improve the overfitting issue as well as generate comparable result. Second test is about the max gradient norm and its effect on soft or hard clip. There is a trend that the higher the value the later the perplexity goes down. Also one thing interesting is that, the value I tested is higher than default 5, and the perplexity on validation set slightly goes high while the training set is still goes down, which means max gradient norm may make the overfitting worse.

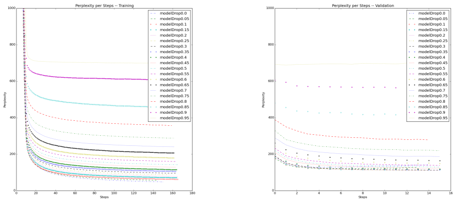

Figure 3: Parameter for dropout rate. Left is for training set. Right is for validation set

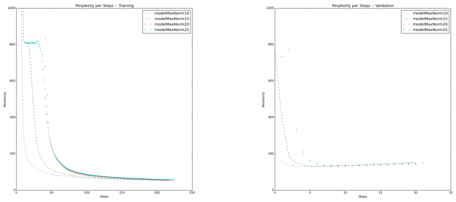

Figure 4: Parameter for max gradient norm. Left is for training set. Right is for validation set