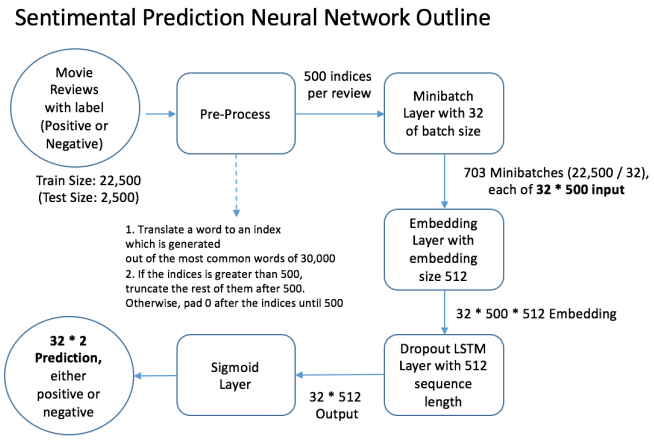
Sentimental Analysis on IMDB

Gonsoo Moon
Post Master Student of Computer Science
University of Texas at Dallas,
gonsoomoon@gmail.com

1. Introduction

Using Dropout LSTM (Long Short Term Memory), sentimental prediction on the IMDB movie reviews is implemented. The movie reviews handled is 22,500 as a train data set and 2,500 as a test data set. The 25,000 movie reviews already have a label of either positive or negative. <The Figure1> shows an overview of the sentimental prediction neural network, technically having embedding and dropout LSTM layer.

<Figure1>



2. Network Architecture

The input data is downloaded from http://www.iro.umontreal.ca/~lisa/deep/data/imdb.pkl. I only used 25,000 reviews. With 32 of batch size, 703 minibatches are made out of 22,500 reviews. Each minibatch, which is 32 * 500, is fed into the following network: embedding layer generating 32 * 500 * 512 embeddings, dropout LSTM layer with 0.5 dropout rate and 512 sequence length producing 32 * 512 outputs, sigmoid layer with the batch normalization resulting in 32 * 2 predictions. For the input 32 * 500 minibatches, if a review is more than 500 indices, the rest of them after 500 indices are truncated. In the case of being less than 500 indices, 0 is padded into the empty slots until the 500 index. Thus, every review has a length of 500. Through the intermediate layers, the final prediction is given in the form of either positive as 1 or negative as 0.

3. Experiment & Result

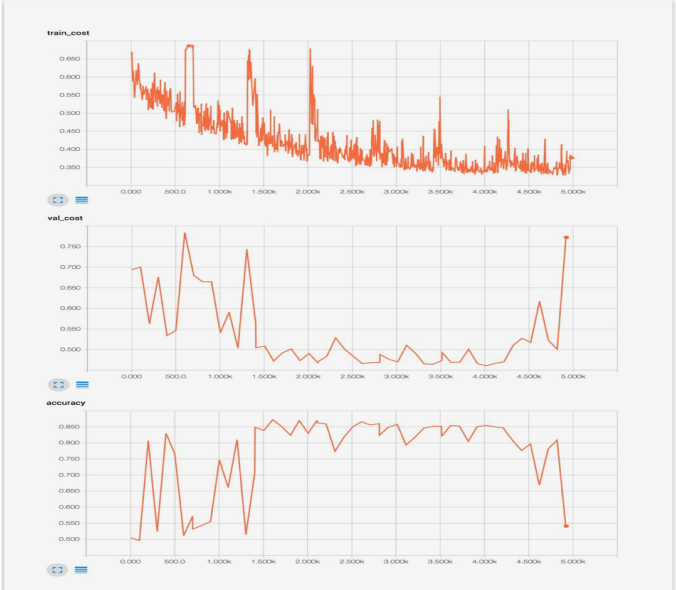
The <Figure2> shows the training cost only after one epoch and test accuracy using the test data set that are not good because of the first epoch.

<Figure2>

```
host/replica:0/task:0/cpu:0
data.train.num_examples:  22500
Total of 703 minbatches in epoch 0
minibatch_x:  (32, 500)  minibatch_y:  (32, 2)
embedding shape:  (32, 500, 512)
lstm output shape:  (32, 512)
output shape:  (32, 2)
Training cost for batch 0 in epoch 0 was: 0.693879
Epoch: 0001 Minibatch: 0001 cost = 0.693878710
Validation Accuracy: 0.48
```

However, as the epoch increases, the training cost and validation cost decrease, shown in <Figure3> that is coming from the book [1], resulting in 86% accuracy.

<Figure 3>



4. References

[1] Nikhil Buduma (2017). Fundamentals of deep learning. Sebastopol, CA: O'Reilly Media