EE782 Assignment-2

Sentiment Analysis

Name - Saurabh Parekh

Roll number - 170100016

1. Instructions on running the code

- Download the glove word embeddings from http://nlp.stanford.edu/data/wordvecs/glove.6B.zip
- Before running the code on colab, mount the google drive and upload the glove.6B folder in your drive in the My Drive folder
- Upload the labels.txt and reviews.txt files present in the data folder to the colab session storage before running the cells.
- Also go to “change runtime type” on colab and change the hardware accelerator to GPU.
- After this is done, just run all the cells using ctrl+F9. It will take some time.

2. Observations and Trends

L2 norm was calculated between the glove word embedding vectors of 50,100,200 and 300 dimensions for the following words:

- v1 = “princess” and v2 = “prince - boy + girl”

Glove word embedding vector dimensions	**Mean square error**
50	0.211
100	0.1462
200	0.147
300	0.112

- v1 = “rome” and v2 = “paris - france + italy”

Glove word embedding vector dimensions	**Mean square error**

50	0.18
100	0.137
200	0.13
300	0.097

- v1 = “queen” and v2 = “king - man + woman”

Glove word embedding vector dimensions	Mean square error
50	0.161
100	0.166
200	0.153
300	0.118

It can be clearly observed that the L2 norm or mean square error(MSE) decreases as the number of dimensions of the word embedding vector increases.
Also, it was observed that the “prince - boy + girl” vector was closest to the “prince” vector followed by the “princess” vector for glove word embeddings.

Default hyperparameters were number of LSTM layers = 2, dimensions of embedding layer = 400, dimensions of hidden layer = 256, no LR scheduler and Adam Optimizer
Hyperparameter tuning was performed on these hyper-parameters and the results are as follows:

- Number of LSTM Layers

Number of LSTM Layers	Number of epochs	Test Accuracy(%)
1	3	79.8
2	4	81.7
3	3	80.3
4	3	80.2

- Embedding layer dimensions

Dimensions of embedding layer	Number of epochs	Test Accuracy(%)
200	3	77.3
300	3	80.8
400	4	81.7
500	3	80.5
600	3	81.2
700	3	79.7

- Hidden Layer Dimensions

Dimensions of hidden layer	Number of epochs	Test Accuracy(%)
64	3	79.6
128	3	80.1
256	4	81.7
512	6	80.7
1024	4	81.4

- Using different optimizers

Optimizer	Number of epochs	Test Accuracy(%)
Adam	3	81.7
Stochastic Gradient Descent (SGD)	10	79.6
RMSprop	4	81.3
AdaDelta	8	81.2
Averaged Stochastic	13	79.6

Gradient Descent (ASGD)		

- Using different learning rate schedulers

Learning Rate Scheduler	**Number of epochs**	**Test Accuracy(%)**
Exponential LR scheduler	5	81.9
Reduce LR On Plateau scheduler	6	81.1
Cosine Annealing Warm Restarts LR scheduler	6	81.6

3. References

- https://medium.com/@lamiae.hana/a-step-by-step-guide-on-sentiment-analysis-with-rnn-and-lstm-3a293817e314
- https://code.google.com/archive/p/word2vec/