Thomas Young, David Zane, Joshua Zhao

Reproducibility Report: CNN for Sentence Classification

Kim, Yoon. "Convolutional Neural Networks for Sentence Classification". EMNLP (2014).
https://github.com/joshuaczhao/CNN-Sentence-Classifier-Reproduction

Following our implementation of a codebase for the above paper, we began reproduction experiments for the different models and datasets. Here, we will summarize our results (with the results from the paper) along with our analysis.

This paper focuses on training CNNs on pre-trained word embeddings (specifically the Google Word2Vec embeddings) in order to do sentence classification tasks. These tasks range from classifying the type of movie review to sentiment analysis for a sentence. The numbers indicated below are the percent accuracies for each model on the given datasets. 10-fold cross validation is also used.

The Models:

There were four different models used in the paper. Each of these models uses the same base architecture, where there is a single convolutional layer (with multiple filter widths of [3, 4, 5], 100 features maps each) followed by a single fully connected layer. Dropout is also applied before the fully connected layer. The pre-trained embeddings used are from the Google Word2Vec embeddings. The models are as follows:

- CNN-rand: trained using initially randomized embeddings
- CNN-static: pre-trained embeddings used and are not allowed to change during training
- CNN-non-static: pre-trained embeddings used and are allowed to change during training
- CNN-multichannel: same pre-trained embeddings used for two separate input channels. One channel of embeddings is static, the other is non-static

Paper Results:

Model	MR	SST-1	SST-2	Subj	TREC	CR	MPQA
CNN-rand	76.1	45.0	82.7	89.6	91.2	79.8	83.4

CNN-static	81.0	45.5	86.8	93.0	92.8	84.7	**89.6**
CNN-non-static	**81.5**	**48.0**	87.2	**93.4**	**93.6**	84.3	89.5
CNN-multichannel	81.1	47.4	**88.1**	93.2	92.2	**85.0**	89.4

The paper specified using the adadelta optimizer with L2 regularization (which we used), but many of the other parameters were not given. We ran our models with a batch size of 50, learning rate of [0.1, 0.01], L2 regularization factor of 0.01, dropout rate of 0.1, and 200 epochs.

Our Results:

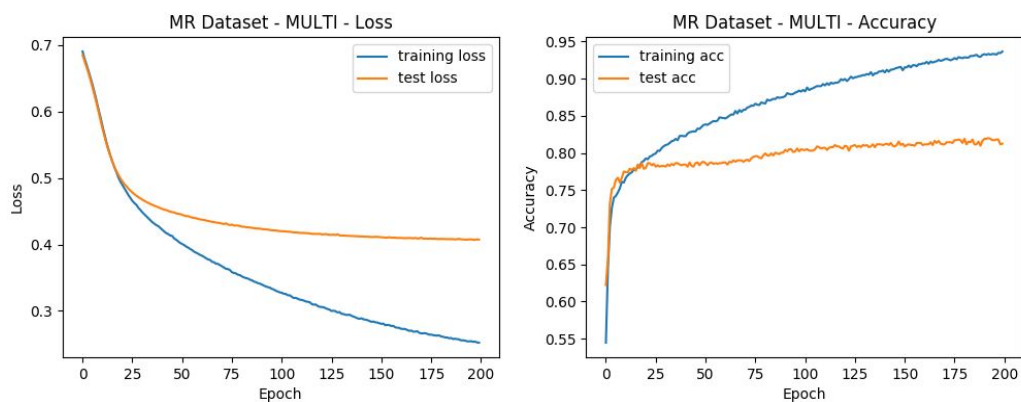| Model	MR	SST-1	SST-2	Subj	TREC	CR	MPQA
CNN-rand	72.8	25.7	73.9	88.2	86.8	84.1	82.0
CNN-static	82.6	**39.7**	**84.1**	92.0	92.6	**86.0**	**91.0**
CNN-non-static	79.7	28.1	81.8	90.1	82.6	84.4	90.6
CNN-multichannel	**82.3**	**39.7**	83.7	**93.8**	**93.0**	84.1	90.0

Figure: Loss and accuracy graphs for best scoring model on MR dataset