Movie Review Sentiment Analysis Kaggle Competition

Feng Gao, Ziyun Chen, Li Cai

October 22, 2018

1 Introduction

The Rotten Tomatoes movie review dataset is a corpus of movie reviews used for sentiment analysis. This Kaggle competition has been inspired by Socher's work on labeling parsed phrases in the corpus. The goal of the project is to predict the sentiment of phrases using the labeled dataset. The initial phase of the project will be focusing on getting the best accuracy results using the CNN-LSTM architecture, Then other architectures will be explored to further improve on CNN-LSTM.

1.1 Problem formulation

The train dataset labeled phrases of movie review sentences on a scale of five values: negative, somewhat negative, neutral, somewhat positive, positive. The goal of this project is to use the dataset to train a neural network model that can accurately label phrase sentiment. There are many obstacles that make this exercise difficult (e.g. sentence negation, sarcasm, ambiguity).

1.2 Dataset

The dataset comprises of tab-separated files with phrases from the Rotten Tomatoes dataset. The train/test split has been preserved for the purposes of benchmarking, but the sentences have been shuffled from their original order. Each Sentence has been parsed into many phrases by the Stanford parser. Each phrase has a PhraseId. Each sentence has a SentenceId. Phrases that are repeated (such as short/common words) are only included once in the data.

train.tsv - contains the phrases and their associated sentiment labels. data also contains sentence id of each phrases.

test.tsv - contains just phrases. the project will be to label each phrase.

1.3 Methodology

For the first phase, a CNN+LSTM architecture neural network will be implemented, optimized, and evaluated. The first layer will be the embedding layer, which will be passed to the CNN layer. Then the features will be pooled to a smaller dimension in the max-pooling layer. These features are then passed into a single LSTM layer. Finally, the LSTM output is then fed to a fully connected layer. Other architectures will be explored in an attempt to further improve on CNN+LSTM

1.4 Goal

The goal is to improve the accuracy and achieve a high ranking in Kaggle.

1.5 Evaluation criteria

Submissions are evaluated on classification accuracy (the percent of labels that are predicted correctly) for every parsed phrase. The sentiment labels are:

0 - negative

1 - somewhat negative

2 - neutral

3 - somewhat positive

4 - positive

1.6 Previous work and references

1. https://www.kaggle.com/artgor/movie-review-sentiment-analysis-eda-and-models

2. https://github.com/vivanraaj/Sentiment-Analysis-on-Movie-Reviews

3. https://github.com/ovguyo/moviereview

4. Pang, Bo, and Lillian Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. Proceedings of the 42nd annual meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2004. http://www.cs.cornell.edu/home/llee/papers/cutsent.pdf

5. Sosa, Pedro M. Twitter Sentiment Analysis Using Combined LSTM-CNN Models. Konukoii.Com, 2018, http://konukoii.com/blog/2018/02/19/twitter-sentiment-analysis-using-combined-lstm-cnn-models/

6. Zhang, Lei, Shuai Wang, and Bing Liu. Deep Learning for Sentiment Analysis: A Survey. arXiv preprint arXiv:1801.07883 (2018). https://arxiv.org/pdf/1801.07883.pdfmo

7. Joao Carlos Duarte Santos Oliveira Violante. Sentiment Analysis with Deep Neural Networks. https://fenix.tecnico.ulisboa.pt/downloadFile/1407770020544739/METI-ARTICLE-THESIS- 70502-JVIOLANTE.pdf

8. Xingyou Wang, Weijie Jiang, Zhiyong Luo. Combination of Convolutional and Recurrent Neural Network for Sentiment Analysis of Short Texts http://www.aclweb.org/anthology/C16-1229

9. Mihaela Sorostinean, Katia Sana, MOHAMED Mohamed, Amal Targhi, Sentiment Analysis on Movie Reviews, March 1, 2017. http://www2.agroparistech.fr/ufr-info/membres/cornuejols/Teaching/Master-AIC/PROJETS-M2-AIC/PROJETS-2016-2017/main(Amal

(There will be more previous works and references to be added.)