Proposal for Capstone Project 2 (Specialized in Deep Learning)

Ahrim Han

2019/7/11

Title: Sentiment Analysis of Movie Reviews using a Deep Learning Convolutional Neural Network

Problem statement: Sentiment analysis (or opinion mining) is the task of identifying and classifying the sentiment expressed in a piece of text as being positive or negative. The sentiment analysis (or opinion mining) has a wide range of applications in industry from forecasting market movements based on sentiment expressed in news and blogs, to identifying customer satisfaction and dissatisfaction from their reviews and social media posts. It also forms the basis for other applications like recommender systems.

Given a bunch of text, sentimental analysis classifies peoples' opinions, appraisals, attitudes, and emotions toward products, issues, and topics. In the past years of studies, a review text was converted to fixed-length vector using bag-of-words and these vectors were later used to train the classifier such as Naive Bayes or Support Vector Machine. The major problem of the bag-of-words are the 1) lack of consideration of semantic relationship between words 2) data sparsity and high dimensionality. Moreover, as there are tons of reviews and posts on the web, there is a strong need for the more accurate and automated sentiment analysis technique.

They are a key breakthrough that has led to great performance of neural network models on a suite of challenging natural language processing problems. Therefore, in this project, I will build deep learning models to classify the positive and negative movie reviews using the high-edge deep learning techniques. I will observe the accuracy and the performance improvements compared to the previous machine learning models/methods.

Who might be the beneficiaries?

Automated and accurate sentiment analysis techniques can be used to detect fake reviews, news, or blogs and are becoming more and more important due to the huge impact on the business markets [1][2]. We provide the potential beneficiaries of this work.

- Businesses can find consumer opinions and emotions about their products and services.
- E-commerce companies, such as Amazon and Yelp, can identify fake reviews. Fake reviews are not only damaging both competing companies and customers, but they also lead to reduced trust in e-commerce companies and lower sales.
- Potential customers also can know the opinions and emotions of existing users before they use a service or purchase a product.

Data:

There is a large dataset for binary sentiment classification of movie reviews [3][4]. It contains substantially more data than previous benchmark datasets. They provide a set of 25,000 highly polar movie reviews for training, and 25,000 for testing. The train and test sets contain a disjoint set of movies, so no significant performance is obtained by memorizing movie-unique terms and their associated with observed labels. In the labeled train/test sets, a negative review has a score <= 4 out of 10, and a positive review has a score >= 7 out of 10. Thus reviews with more neutral ratings are not included in the train/test sets.

Solution approach:

We prepare movie review text data for classification with deep learning methods. We obtain the large data set of the movie reviews. We clean the documents of text reviews by removing punctuations, stopwords, stemming and removing non-frequent words to prevent a model from overfitting. In this pre-processing of documents, we use the more sophisticated methods in the NLTK library.

To build a deep learning model, we basically use the sequential model of Keras.

1) First, the Embedding layer is located. There are two ways of setting the embedding layer: using the pre-trained word embedding or training new embedding from scratch. We use a pre-trained word embedding, GloVe (Global Vectors for Word Representation) embeddings.

2) Second, a series of convolution 1D and pooling layers are added according to typical CNN for text analysis. Then, after flattening layer, fully connected dense layers are added. Since this is a binary classification problem, we use the sigmoid function as an activation function for the final dense layer.
3) Finally, we will make the different deep learning models by adjusting the parameters and will find the best accurate model. We later will investigate the various parameters affecting the accuracy.

Deliverables:

1) Codes (notebooks) for data cleaning, exploratory data analysis, interactive data visualization, machine learning model development
2) Report on the capstone project
3) Slide on the capstone project

References

[1] Shrestha, Nishit, and Fatma Nasoz. "Deep Learning Sentiment Analysis of Amazon. com Reviews and Ratings." *arXiv preprint arXiv:1904.04096* (2019).

[2] https://www.fakespot.com/

[3] https://ai.stanford.edu/~amaas/data/sentiment/

[4] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. (2011). Learning Word Vectors for Sentiment Analysis. The 49th Annual Meeting of the Association for Computational Linguistics (ACL 2011).