Machine Learning Engineer Nanodegree

Capstone Proposal

Piyush Moolchandani
September 26th, 2019

Proposal

Domain Background

Sentiment analysis is contextual mining of text which identifies and extracts subjective information in source material, and helping a business to understand the social sentiment of their brand, product or service while monitoring online conversations. In today's competitive market, it is very important to analyse and act upon the reviews but with the ever growing data, it is impossible to get reviewed each and every review by an human. So in that case sentiment analysis algorithm becomes an crucial advancement in this field. In the beginning various techniques such as Naive Bayes, logistic regression were used for sentiment analysis, but with recent breakthroughs in the field of deep learning has made the RNNs and LSTMs a popular choice for this work.

- LSTM's were proposed in this paper by Sepp Hochrieter et al. in 1997(http://www.bioinf.jku.at/publications/older/2604.pdf) (http://www.bioinf.jku.at/publications/older/2604.pdf)).
- Gated Feedback Recurrent Neural Network extends the existing approach of stacking multiple recurrent layers by allowing and controlling signals flowing from upper recurrent layers to lower layers using a global gating unit for each pair of layers. The recurrent signals exchanged between layers are gated adaptively based on the previously hidden states and the current input. (https://arxiv.org/pdf/1502.02367.pdf (https://arxiv.org/pdf/1502.02367.pdf))

Both LSTM and GF-RNN weren't written specifically focusing on sentiment analysis, but a lot of sentiment analysis models are based on these two highly cited papers.

Problem Statement

The objective of this project is to predict the number of positive and negative reviews on the IMDB dataset from this ACL 2011 paper by Learning Word Vectors for Sentiment Analysis(http://www.aclweb.org/anthology/P11-1015) (http://www.aclweb.org/anthology/P11-1015)).

This can be done using either classification or deep learning algorithms such as simple ANN(fully connected layers) or LSTMs or GF-RNN. And the metric used to measure the performance is accuracy of predictions against the true labels.

Datasets and Inputs

IMDB dataset from this ACL 2011 paper by Learning Word Vectors for Sentiment Analysis(http://www.aclweb.org/anthology/P11-1015 (http://www.aclweb.org/anthology/P11-1015)) by Maas and Andrew will be used for this project.

Dataset can be downloaded from http://ai.stanford.edu/~amaas/data/sentiment/aclImdb_v1.tar.gz (http://ai.stanford.edu/~amaas/data/sentiment/aclImdb_v1.tar.gz) or from https://www.kaggle.com/c/word2vec-nlp-tutorial/data (https://www.kaggle.com/c/word2vec-nlp-tutorial/data)

The labeled data set consists of 50,000 IMDB movie reviews, specially selected for sentiment analysis. The sentiment of reviews is binary, meaning the IMDB rating < 5 results in a sentiment score of 0, and rating >=7 have a sentiment score of 1. No individual movie has more than 30 reviews. The 25,000 review labeled training set does not include any of the same movies as the 25,000 review test set. In addition, there are another 50,000 IMDB reviews provided without any rating labels. But only train and test dataset will be used, unsupervised set will not be used in the project.

Solution Statement

For solving the problem, some preprocessing will be required on the dataset to get input in a proper form to feed into a neural network, then creating a validation set from the data. Then the data will be fed into a sentiment network consisting of LSTM layer(s) and model can be trained to reduce loss obtained in each epoch.

After training, model is tested using test data and performace is judged based on the accuracy of predictions against actual labels.

Benchmark Model

I plan to compare the performace of RNN(LSTM) with that of traditional approaches like Naive Bayes. I will use Naive Bayes on preprocessed data as a benchmark model. I will work to improve results significantly in my model than the results obtained in benchmark model. Comparison between two models will be done based on the accuracy metric.

Evaluation Metrics

After training the model on training set I will predict the output lables for the testing set and then I will estimate the performance of the model by accuracy metric.

Project Design

Proposed workflow of my project will is as follows:

```
1. Load the data.
2. Data pre-processing
    - encoding of words, labels and removal of outliers
3. Any additional pre-processing if required, not clear right now.
4. Splitting of data into training, validation and test data and batching of data.
5. Declaration of model network, I am not sure whether I will be using Tensorflow
 or Keras for this, but the model will possibly contain:
    - Embedding layers
    - LSTM Layers
    - Fully connected ouput layers with sigmoid activation layer after last fully
 connected layer.
6. Training of Model based on reduction of loss(most possibly cross entropy loss).
7. Testing of Model and judging the performance based on accuracy metric.
```