Prerak Mehta

MSDS 453 - Natural Language Processing

Assignment 4 - Neural Networks

Abstract:

This research assignment is focused on comparing performances of two different neural network models/ techniques - Recurrent Neural Network (RNN) and 1D Convolutional Neural Network (CNN). This assignment will also focus on examining effects of the dimensionality of word embeddings, or word vectors, and the size of the embedding vocabulary used on the training and performance of recurrent neural networks. In the end the performance of different types of RNNs will also be compared to evaluate how the techniques and hyper parameters affected the performance of the RNNs. The dataset utilized in this research assignment consists of 25000 movie reviews out of which 12,500 are positive reviews and 12,500 are negative reviews.

Introduction:

This research is being conducted to solve a very straightforward, “yes” or “no”, problem for people which is whether to watch a certain movie or not based on reviews available all over the internet on platforms such as Google Reviews, IMDb, Rotten Tomatoes, Twitter, etc. Movie critics all around the world very rarely have the same thing to say about movies. Certainly in recent times there are online platforms where an increasing number of common people are able to share their opinions about movies as well. One may often find themselves seeing very different ratings for a movie on different platforms on the internet which may result in confusion about whether to watch the movie or not. Therefore the model built in this research may help people obtain a general sentiment of the public regarding a certain movie based on movie reviews available all over the internet.

Note - Web crawling was researched in the first assignment so this assignment will just be focused on working with an already (publicly) available dataset of movie reviews.

Literature Review:

A lot of platforms on the internet are focused on (or at least have the capability to publicize) individual opinions about a movie. Youtube comments section, Twitter, IMDb, and Rotten tomatoes are some of the most popular and common places where people express their opinions about certain movies or shows apart from newspapers and movie journals. A few of them have varied ratings systems such as points out of 10 (IMDb) or a percentage out of 100 using numerous criteria (Rotten Tomatoes) along with written reviews by the audience. A few of them only have written reviews in terms of comments available on their platform (Twitter or YouTube). However these reviews can still be varying a lot on these platforms especially for movies/shows that are rather average in quality. In these cases the audience will find themselves confused as different platforms will have different/mixed ratings or reviews expressed on them. That is where the model prepared in this assignment can come into play and give the audience a "yes" or "no" about whether to watch a movie based on the general sentiment of the public across various platforms on the internet.

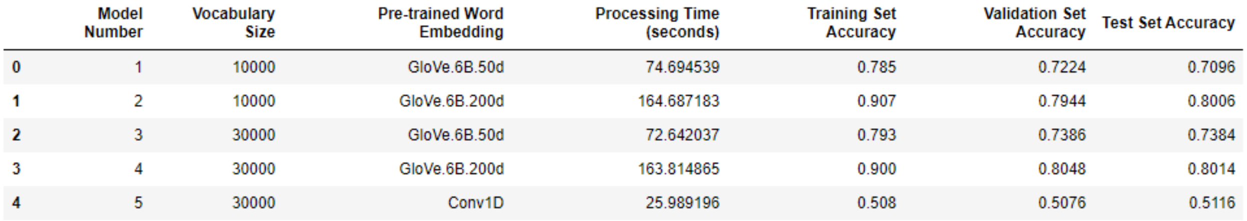
Methods:

To build the model firstly collect the positive and negative reviews from their respective txt files (publicly available online) and put them in a list. Do the same with sentiment labels of the reviews (1 = positive and 0 = negative) and put them in a different list. Create a dataframe of these 2 lists and an index. Then start applying the data preprocessing techniques. The first step when building a neural network model is getting the data into the proper form to feed into the network. Since embedding layers are used in this research, each word will be encoded with an integer. The data needs to be cleaned up prior to modelling by getting rid of periods, punctuations, delimiters that insert newlines in the reviews, etc. Then get all the text without the newlines in one big string and split it into individual words. Include stop words in the reviews also. Lastly apply tokeninzation on words where words with length 3 or less are removed to emphasize more on the more important words in the review.

The data will be split into 1000, 5000, and 5000 reviews for training, validating and testing respectively. By applying tokenization integer values are assigned to the words in the reviews and split them into words. Later randomize the split as mentioned in point 1 into training, validation and testing sets. Two different types of sets will be created because of the usage of 2 different vocabulary sizes to train models. Pre-trained word embeddings (50 glove and 200 glove) will then come into play to prepare 4 different embedding matrices with 4 combinations of vocab sizes and dimensions plus the 1 model with 1d convolutional neural network to prepare the base for 5 models. 4 different recurrent neural network (RNN) models will be with 1) 10000 vocab size and 50 glove dimensions, 2) 10000 vocab size and 200 glove dimensions, 3) 30000 vocab size and 50 glove dimensions and 4) 30000 vocab size and 200 glove dimensions are trained and fitted on their 2 respective training sets. The 1d convolutional neural network model will be prepared with 30000 vocab size and 200 embedding dimensions. Finally the training, validation and test accuracies are compared in order to evaluate the performance of the 5 models.

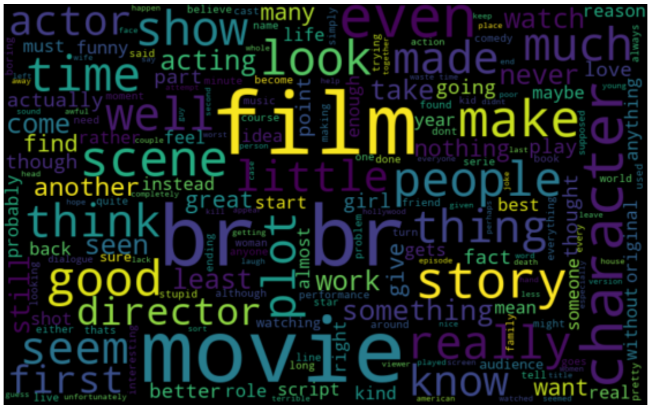
Results:

The results from the 4 RNN models came out pretty promising compared to the result from the CNN model. Also from the RNN models the model with the higher vocabulary size AND the higher pre-trained word embedding dimensions had the highest performance amongst all the models. Below is the summary snapshot of the performance by all 5 models on the training, validation, and testing sets.

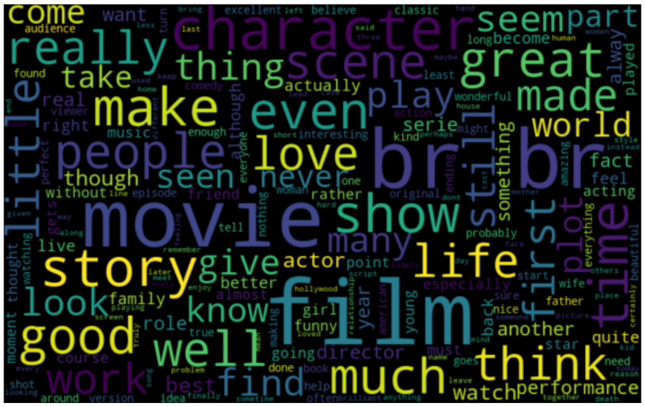


	Model Number	Vocabulary Size	Pre-trained Word Embedding	Processing Time (seconds)	Training Set Accuracy	Validation Set Accuracy	Test Set Accuracy
0	1	10000	GloVe.6B.50d	74.694539	0.785	0.7224	0.7096
1	2	10000	GloVe.6B.200d	164.687183	0.907	0.7944	0.8006
2	3	30000	GloVe.6B.50d	72.642037	0.793	0.7386	0.7384
3	4	30000	GloVe.6B.200d	163.814865	0.900	0.8048	0.8014
4	5	30000	Conv1D	25.989196	0.508	0.5076	0.5116

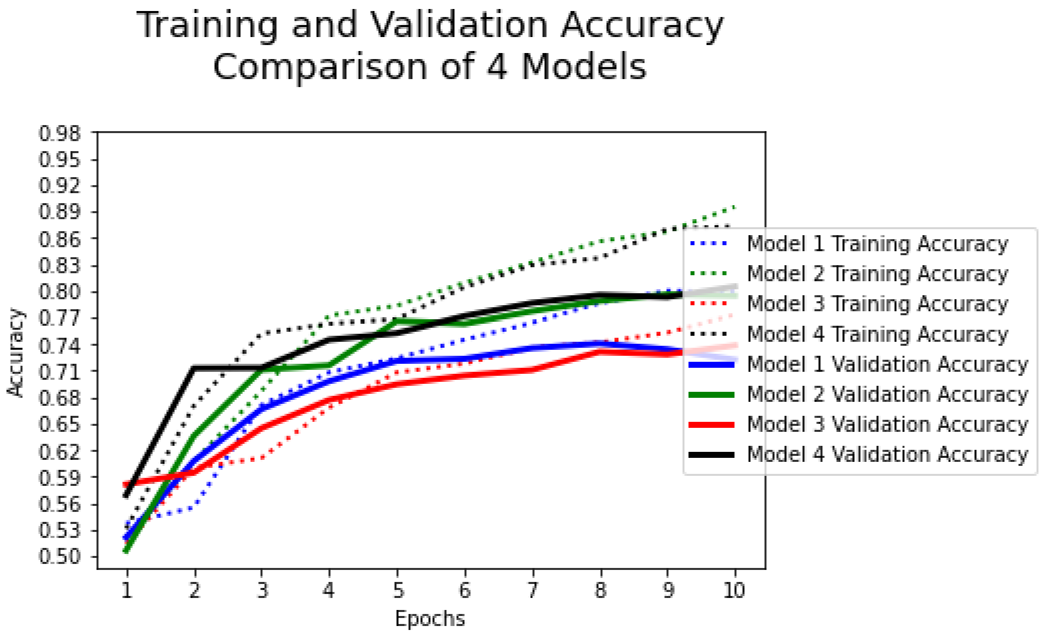
Below is the snapshot of the most popular words in the negative reviews



Below is the snapshot of the most popular words in the positive reviews



Below is the graph that depicts the comparison between the training and validation accuracies of the 4 RNN models.



Conclusion:

From this research assignment it can be concluded that the Recurrent Neural Network method is a very effective method to build a model for sentimental analysis on movie reviews. Also that the higher vocabulary size and embedding dimensions are directly proportional to a higher performance and processing time by the RNN model. This research also emphasizes the importance of LSTM and drop out layers on the model, how using more epochs can cause overfitting of the data, i.e. increasing training accuracy but decreasing validation accuracy and the importance of preprocessing the data before training a model on it. There is still room of improvement by going deeper in preprocessing the data and exploring other hyperparameters to obtain a more robust model.