Project 2

Habibur Rahman Dipto

kazi.dipto@northsouth.edu
North South University
Department of Electrical and Computer Engineering

June 30, 2021

Abstract

NLP tasks are the most popular tasks in ML/DL. Everyday some researchers bring new techniques, interpretation, insight about NLP. The major problems for NLP's are lack of structured data, large vector representation, limitation of hardware capacity, ambiguity of problem specification, heavy execution time etc. In this project we are assigned to participate in Kaggle Jigsaw Unintended Bias in Toxicity Classification. Based 1.80million of comment we have to build a toxicity classifier which can classify whether or not a comment is toxic.

I. Methodology

For this problem I picked BERT as a model because it's a good paper from Google AI. It uses a bidirectional transformer (non-directional). It is still SoTA for many classification, sentiment analysis, question answering, name entity recognition type problem. The only downside is that to fine tune this model we need good hardware. First I prepossessed our data by removing all the emojis, symbol and escaped character, mapped contraction with their full form. Then I tokenized all the sentence. Then I downloaded the word embedding of BERT and mapped our word with theirs. Finally normalized all the word embedding weight and built our embedding matrix. For model we used the official BERT architecture. We used batchsize of 512 and number of lstm unit is 128. We trained our model for two epochs.

II. Result Analysis

The accuracy we hit is 94.07%. We can see that our model did a really good job with the test data. Since BERT's transformer watches the sentence at once, it can extract good meaning of word's context (left-to-right or right-to-left). Also since BERT trains itself with 15% of masked words from a sentence and tries to predict the masked word, It learns some good word representation for unknown words.

i. Conclusion

The current model gives us a very good result. But to go further we can use some other word embedding such as Fasttext word embedding. By doing that we might get some new useful meaning of the word and make our model better.

III. Acknowledgements

I would humbly like to thank both Upskill and IM for the discussion about why there are limitations of using NN in tabular dataset.