Artificial Intelligence II

IMDB Sentiment Analysis - BERT Model

Georgios - Alexandros Vasilakopoulos
A.M. 1115202000018

Test the performance on your own dataset: Check at the bottom of the `AI2_Assignment4.ipynb` file for information about how to evaluate the performance on a test dataset.

Data Preprocessing

- The given data is loaded into pandas dataframes and split into train, validation, and test sets.
- Similarly to the previous assignments, all alphabetic characters are converted to lower case and all non alphabetic characters are removed (through the function `clean_text()`). This time I decided not to remove the english stopwords, as they may help the model's attention mechanism.
- The processed data sets are then passed into BERT's tokenizer. For each review, the tokenizer returns an encoded vector of size 768. For that reason, our model's input size will be 768.
- Both the input encodings and the labels are properly stored within the `ReviewDataset` class.

Model Architecture

- Two different model architectures were considered in the making of the final model: One with a single-layered feed-forward network and another with two layers. Both of these architectures operate on top of a BERT (base,uncased) model.
- The single-layered model was trained for 10 epochs, in order to determine the number of epochs above which the model starts to overfit. The learning curve plotted in the ipynb file shows that after three to four epochs, the model becomes a bit unstable. Therefore, the model named `myBERTmodel.pt` was trained for three epochs.

- The model trained for three epochs produced an impressive f1 score of 0.93 (Precision: 0.94, recall 0.92) on the test set, while still being stable enough, due to the choice of 3 epoch training.
- The number of connections on the one and only deep layer of the model is 128. A few other choices were considered, regarding the number of connections, but none of them showed significantly better performance after three epochs of training.
- The two layered model, after being trained for three epochs, produced an f1 score of 0.93 on the test set. Since the two layered model did not show a significant increase in the performance, I preferred the single layered model over the two layered model, following the principle of "Keeping the model simple".
- The ROC curves for both of the models imply a very good classification performance.

Comparison with previous models

- The single layered model, trained for 3 epochs, produced an f1 score of 0.93 on the test set. None of the models of the previous assignments had managed to break the 0.9 threshold and, therefore, this is a significant result.
- The high performance of the model can be attributed to the encodings produced by the BERT tokenizer and the sophisticated structure of the BERT model, which was pretrained.
- On one hand, it is positive that BERT is a versatile model that can be fine-tuned over many different NLP tasks
- On the other hand, BERT itself is a very complicated model that requires a huge amount of training data in order to be pre-trained. As a result, a single individual would not be able to train with a BERT model by themselves, due to computational limitations and insufficient amounts of data.