BERT is a new state-of-the-art model released in mid of 2018 by Google. The model uses Con-textualized Word Embedding instead of Static Word Embedding like Word2Vec, Glove or fastText. BERT applies bidirectional Transformers which demonstrate that RNN and CNN are not essential for building high-performance natural language models.

BERT has archived good score in GLUE tasks and that is the reason why I would like to use BERT in IMDB dataset. Even when BERT was developed by Google but the model in tensorflow is huge and hard to understand. Then I choose to build BERT model with Pytorch with *transformer* package from *huggingface*. Building from Pytorch is much more flexible and it allowed me to config parameters in the model. Authors suggested that using this configuration for training data:

* Learning rate: 2e-5
* Epochs = [2,3,4]
* Max sequence lengths = [512, 1024]
* Batch size = [16, 32]

Since I was fine-tuning BERT model for IMDB dataset, I changed learning rate to 1e-5. Epoch was set as 4 because the loss was decreasing in every number of epochs. Training time per epochs was around 30 minutes and it took 2 hours to finish training steps. Based on the IMDB dataset, the average length of reviews is 297 so I choose sequence length as 512. With 11GB GPU Ram of RTX 2080 Ti, I was only able to load 6 reviews per batch or would get out-of-memory error. The pre-trained model that I used is bert-base-uncased. Below is the summary of my config:

{   
 ‘learning rate’: 1e-5,  
 ‘epoch’: 4,  
 ‘max\_len’: 512.  
 ‘batch\_size’: 6  
}

The model has got 0.99 accuracy on training dataset and 0.93 f1-score on official IMDB test data (25000 reviews) while Kaggle achived 0.90 f1-score in both public and private leaderboard (7000 reviews). I assumed that two test data is different where there is 2 null reviews in Kaggle dataset.

Besides, I only lowercased all reviews, removed <br> characters and used WordPiece tokenizer in preprocessing step.