BERT Overview

By Mohammed Alabdullatif (maa27@illinois.edu)

Introduction

This paper aims to provide an overview of Bidirectional Encoder Representations from Transformers, or BERT for short. All of the knowledge in this paper is taken from the article “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding” by Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. In the past, language models were only able to process text input from left to right or right to left, which added a constraint that prevents the model from being able to use the full context surrounding any given text. BERT however, is able to bidirectionally process the input text, which gets rid of the constraint given from the old unidirectional models. Using Masked Language Model (MLM) and Next Sentence Prediction (NSP), BERT has the ability of predicting a blank word in a sentence, or give the relationship between two given sentences. The paper will discuss the use of Masked Language Model and Next Sentence Prediction, and how each of these tasks contribute to the BERT language model. Also, the paper will also discuss the pre-training data used by BERT to allow it to successfully achieve MLM and NSP. Finally, this paper will discuss examples of BERT applications.

Pre-training Data

Before discussing the tasks BERT undergoes, let us look into the pre-training data used by the BERT model to understand language. There exists numerous sources of data and corpus that can be used as pre-training data. As the paper “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding” suggests, BookCorpus and English Wikipedia were used as pre-training data for the BERT model. BookCorpus contains 0.8B words, while English Wikipedia has 2.5B words. Only text passages were extracted from English Wikipedia. All lists, tables, and headers were ignored. Such data is very helpful to use and train the BERT model to understand language. They focused on collecting the data from such sources to focus on “document-level corpus rather than a shuffled sentence-level corpus”. Doing so allows BERT to get full context instead of having a limited scope. In the BERT applications section, a few other corpuses will be mentioned to highlight the importance of pre-training data on the BERT model and how it uses the data to understand the language found in the given corpuses.

Masked Language Model

The Masked Language Model, MLM for short, allows the BERT model to predict a masked word in a given context. According to the paper “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”, 15% of all WordPiece tokens are masked at random and the model predicts the masked token instead of reconstructing the entire input. Using this approach allows for a bidirectional pre-trained model. Interestingly enough, using this approach

for all tokens creates a mismatch between pre-training and fine tuning. To improve the use of MLM, the masked words are not always replaced with the generated mask. Instead, the mask is only used 80% of time, while 10% of the time a random token is chosen, and for the remaining 10% of the time, the masked words remains unchanged. For example, we have the sentence “The kid loves to draw”, and our masked word is “draw”. For 80% of the time, the word “draw” will be replaced with the mask generated by the model. For 10% of the time a random word like “tree” will be chosen even if it does not fit grammatically. As for the remaining 10% of the time, the word “draw” will be remained unchanged. Again MLM only masks 15% of the training data. Doing so will allow the model to be trained to make word predictions based on the surrounding context.

Next Sentence Prediction

While using Masked Language Model is helpful, it does not give the BERT model any input into the relationship between two sentences. This is where the Next Sentence Prediction, also know as NSP, comes in. It mainly focuses on using the pre-training data to train the model to predict if a sentence should follow another a sentence. For example, if we have the following two sentences: “The boy loves play with planes” and “He wants to be pilot when he grows up”, the model can predict if the second sentence follows the first one. It labels it as IsNext or NotNext to identify whether or not the sentences are sequential. Having NSP further improves BERT’s ability to understand the given corpuses language. It lets the model connect sentences together instead of only focusing on words. By doing so, the model can predict if a sentence is supposed to follow another or if it does not fit the language. Different datasets were used to test BERT’s accuracy such as MNLI-m, QNLI, MRPC, SST-2, and SQuAD. When using BERT without NSP, the performance of the language model actually worsens, which goes to show how important NSP is in the BERT model.

BERT Applications

In this section of the paper, we will discuss different BERT applications and how the methods and tasks mentioned in this paper are used to understand different corpuses. BioBERT was created to focus on biomedical literature. The paper “BioBERT: a pre-trained biomedical language representation model for biomedical text mining” discuses how BERT was used to develop this language model. It used the corpuses used by the BERT model and added PubMed abstracts and PMC full-text articles, which have 4.5B and 13.5B words respectively. Both of the added corpuses focus on the Biomedical domain which helps train BioBERT to focus more on the Biomedical domain.

Another BERT based application can be seen in the paper “FinBERT: Financial Sentiment Analysis with Pre-trained Language Models”. It introduces FinBERT, which is a BERT model that focuses more on the financial domain. Three datasets were added to BERT to allow FinBERT to pre-train with a focus on the financial domain. The datasets mentioned in the paper are TRC2-

financial, Financial PhraseBank, and FiQA Sentiment. FinBERT performs better than BERT when it comes to understanding financial literature. The results are shown in the afore mentioned paper in this paragraph.

Conclusion

There are plenty of other applications of BERT. Depending on the corpuses used and the domain the corpuses focus on allows the BERT model to pre-train with domain-specific literature. The ability of the BERT model to process the data bidirectionally gives it an advantage over unidirectional models because it allows it to utilize the Masked Language Model (MLM) and Next Sentence Prediction (NSP) to understand the language of given corpuses during the pre-training stages.The BERT model shows promising results of understanding the languages of the corpuses given. The way it works and how it can create a language model with fruitful results makes it so computers and machines can understand and predict outcomes of any given language as long as a legible dataset is given to it for pre-training.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 24 May 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 10 September 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining.

Dogu Araci. 25 June 2019. FinBERT: Financial Sentiment Analysis with Pre-trained Language Models.