CS 410 Text Information System – Technology Review
Overview of BERT and its application

Darren Muliawan – darrenm2

Introduction

BERT, which stands for Bidirectional Encoder Representations from Transformers, is a new open-source technique for Natural Language Processing which was developed by researchers at Google and introduced in 2018. BERT is great for various Natural Language Processing (NLP) tasks such as text classification and text generation.

In this technical review, we are going to discuss what BERT is and how does it work, techniques before BERT, real world application of BERT, and its disadvantage.

Challenges in NLP

BERT was created to tackle one of the main challenges in NLP, which is to create a task-specific dataset. Why is it so hard to create a task-specific dataset? It's because there were not enough training data. There are a lot of text data available, but in order to create a task-specific dataset, we need to divide those text data into smaller datasets, and it is hard to make these models perform well if the dataset is small.

In order to tackle this challenge, researchers developed various technique to train a general-purpose language representation model using text data from the web. This pre-trained model then can be fine-tuned for task-specific dataset, for example, question answering. It turns out that fine tuning this general-purpose model results in better accuracy compared to creating a task-specific model from scratch. BERT was an addition to these techniques for creating a pre-trained model.

Before BERT

There were many techniques for Natural Language Processing before BERT such as word2vec, CNN (Convolutional Network), RNN (Recurrent Neural Networks), and LSTM (Long Short Term Memory). Each of these techniques have its own drawbacks which results to the development of BERT.

Main idea of BERT

Language model is designed to solve a "fill in the blank" task in the sentence based on that sentence's context. Given a sentence "I accessed the ___ account", a language model's goal is to correctly predict the missing word, for example, "bank" more often than "Facebook".

In the pre-BERT time, a language model would look at this sentence either from left-to-right or right-to-left or combined left-to-right and right-to-left. This works well to generate sentences which will keep predicting the next word until the sentence completes. However, by using bidirectionally (or non-directional) trained language model like BERT, the language model can gain deeper sense of the sentence's context than using unidirectional trained language model.

For example, given the “I accessed the” in unidirectional way, the model may (or may not) generate the word “bank” but not “account”. If we look at this sentence bidirectionally, the model can insert “bank” in the “I accessed the ___ account” because it also checks the previous and next context of the sentence.

How does BERT work

BERT relies on Transformer, which is an attention mechanism that learns contextual relations between words in a text. BERT does not try to predict the next word in sequence, but instead uses 2 training strategies, which are Masked LM (MLM) and Next Sentence Prediction (NSP). The idea of Masked LM is replacing 15% of the words in each sequence with a [MASK] token. Based on the context of the sequence, the model tries to predict what is the value of the “masked” words. The Next Sentence Prediction (NSP) strategy is used to teach the model if there is a relationship between 2 sentences.

Real-world application of BERT using TensorFlow

TensorFlow is an open-source machine learning library developed by Google and was released in 2015. TensorFlow started out using Python for its language but now it supports multiple languages such as TensorFlow.js for JavaScript and TensorFlow Lite which is for deployment on mobile or embedded devices.

Using TensorFlow Lite, a pre-trained BERT model and fine-tuning it on SQuAD (Stanford Question Answering Dataset) 1.1, we can create a mobile application that can answer users’ question. Given a text data and a question as input, the mobile application can return the sequence of words which are most likely answers the question.

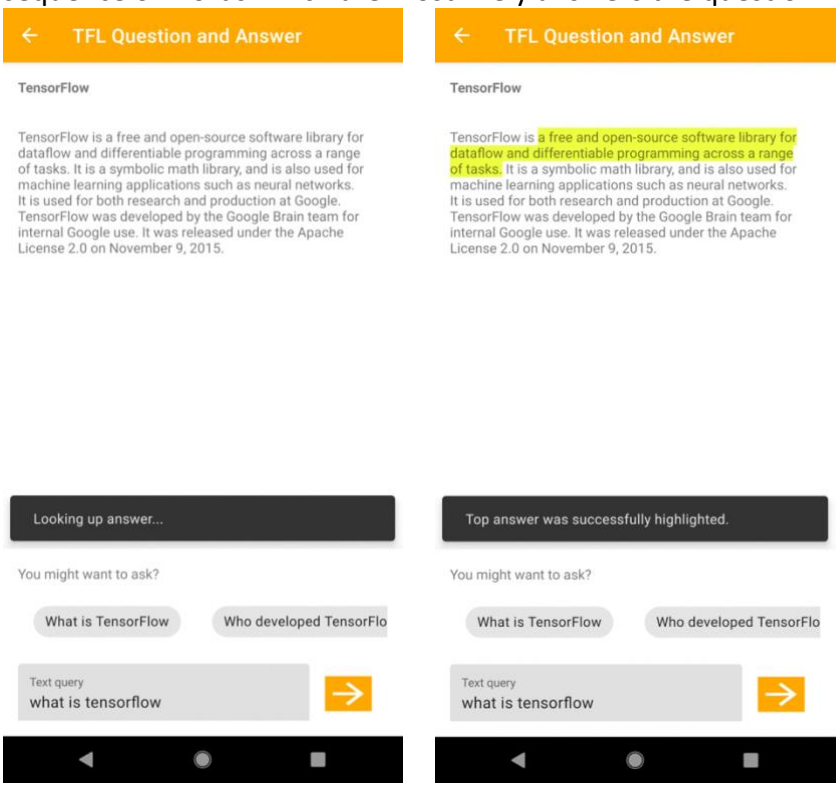

Figure 1 – An example of Q&A mobile applications from TensorFlow website

Disadvantage of BERT

Since BERT use a bidirectional approach, the model converges slower than regular left-to-right approach, but it outperforms left-to-right training after some number of pre-training steps.

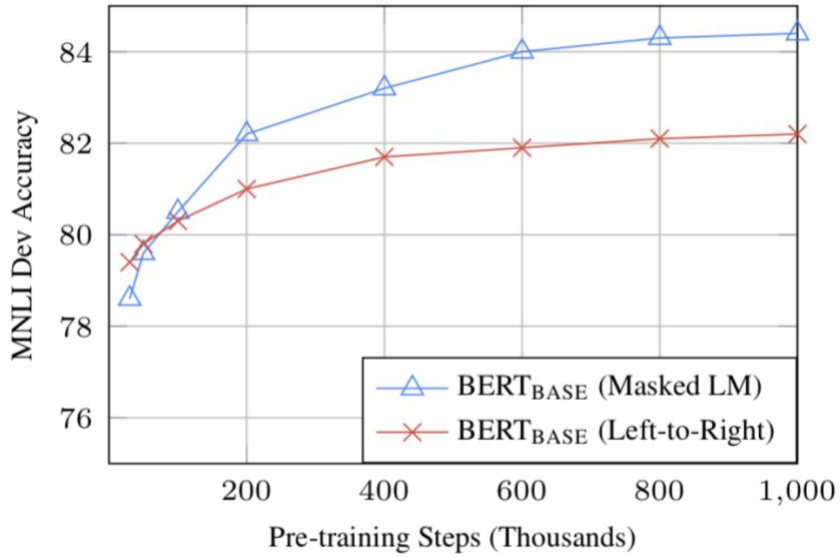

Figure 2 – BERT's bidirectional approach vs left-to-right approach

Conclusion

There is no doubt that BERT is considered as a phenomenal breakthrough in Natural Language Processing field. With BERT, we can create a more powerful language model that better understands the context of the whole text datasets. With better understanding, it makes NLP tasks more "human".

References

- Explanation of BERT - https://towardsml.com/2019/09/17/bert-explained-a-complete-guide-with-theory-and-tutorial/
- TensorFlow – https://www.tensorflow.org/