Difan Gu

CS 410 Text Information System

10/24/2020

Technology Review: Bidirectional Encoder Representations from Transformers (BERT)

Introduction

Arguably, there are several Natural Language Processing waves in past decades. The first wave was from 1950s to 1990s, Symbolic NLP was rather popular. Symbolic NLP is rule-based natural language processing technique, i.e. it understands the rule rather than the meaning of the sentence. The second wave last from 1990s to 2010s, Statistical NLP dominates the NLP. Statistical NLP has been widely introduced in CS 410 Text Information System. It mainly used statistical learning or equivalently machine learning techniques to understand the pattern of languages. Even though Statistical NLP can be used in most of cases, it lacks the deeper understanding of languages, and therefore not able to achieve more complex tasks such as answering questions and resolving ambiguity at different levels.

The present trend of NLP tends to focuse on Deep Learning and Neural Network, a much more complex structure of machine learning that consists of a large number of parameters, units and hidden layers. Neural network and deep learning models can explore non-linear relationship between predictors and outcomes in order to simulate the realistic cases.

Recently, a new NLP model, Bidirectional Encoder Representations from Transformers (BERT), has been introduced and achieved a non-parallel performance based on GLUE score, MultiNLI, SQuAD v1.1 and SQuAD v2.0. In the following section, we will focus on how BERT works and why BERT works so well.

Test Name	**Absolute Improvement**
GLUE	+7.7%
MultiNLI	+4.6%
SQuAD v1.1	+1.5
SQuAD v2.0	+5.1

(Source: Devlin 2018)

How BERT works

One of the most distinct features of BERT is that the algorithm is able to be trained both from left to right and right to left. This is also the reason that term “bidirectional” is used in the naming. For instance, BERT is able to predict the blank in the following sentence based on the complete sentence instead of only the part right before the blank:

A boy is watching the ___ with a telescope.

There are two main steps in BERT (Devlin 2018). The first step is pre-training and the second step is fine-tuning. In the pre-training step, the goal is to build a foundational and general model by learning a large amount of data. For example, the creator fed English Wikipedia (~ 2,500 million words) and BooksCorpus (~800 million words) into the model and let the model learn sequential sentences instead of independent pieces from different documents. Moreover, pre-training steps requires some creative training tasks. For example, Masked Language Model (MLM) has been introduced. MLM will randomly “mask” or equivalently hides 15% of tokens in the training set, and allows BERT to “guess” and “predict” the words. Within 15% of token got replaced, not all of them are truly hidden: a small portion of it is replaced by a random token and another small portion of it is replaced by its original token. The rationale is that the fine-tuning step won’t have “hidden” token. Loss will be only evaluated between predicted words and hidden words. Another technique is also used, called Next Step Prediction (NSP). The goal of NSP is to force BERT to understand sentence relationship. More specifically, NSP artificially creates 50% TRUE and 50% FALSE case to train the model. For example, sentence B follows A, but sentence C doesn’t follow A. In this case, B will be labeled as TRUE and C will be labeled as FALSE.

The second step is called fine-tuning. The parameters between the pre-trained model and fine-tuned model is very similar. Fine-tuning step also requires much less computational power than training. For example, for classification tasks, there will be an additional classification layer to generate classification result for each sentence.

Why BERT works

The traditional unidirectional language models only study a sentence either from left to right or right to left. However, unidirectional language model will highly limit the information that is supposed to be taken into consideration. For the same instance provided above, if the whole sentence is "A boy is watching the ___ with a telescope." A unidirectional model is only able to use "a boy is watching the" to predict the blank. However, obviously, it's not sufficient to infer the blank word, as the second half of the sentence is missing and meanwhile includes very important information. By using bidirectional language model, BERT, the issue can be solved. The complete sentence, "A boy is watching the ___ with a telescope.", will be taken into account. Then it's a lot easier to infer that blank could be "sun", "moon", etc.

Conclusion

Undoubtably, BERT is the most cutting-edge language model so far in Natural Language Processing area. It approaches a problem from a unique perspective by understanding a sentence bidirectionally so that both sides of information around a missing word can be taken into consideration. More detailed techniques such as Masked Language Model and Next Sentence Prediction also grants the model with additional functionalities to find the answer to a question.

However, BERT also has some weakness. For example, it highly relies on a large training set to train the model to be general. Also, the model requires a large number of parameters.

Reference

Devlin, Jacob; Chang, Ming-Wei; Lee, Kenton; Toutanova, Kristina (11 October 2018). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding"

History of natural language processing. (2020, October 22). Retrieved October 25, 2020, from https://en.wikipedia.org/wiki/Natural_language_processing

Horev, R. (2018, November 17). BERT Explained: State of the art language model for NLP. Retrieved October 25, 2020, from https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270