A report on BERT Algoritham

B Raja Narasimhan

August 2019

1 Introduction

BERT(Bidirectional Encoder Representations from Transformers) is a algorithm open sourced by Google. BERT is the first deeply bidirectional, unsupervised language representation, pre-trained using only a plain text.It can also predict some missing words in the sentences based on other data in the sentences. What it can also do is predict if two adjacent sentence will make sense or it is random. Also BERT performs better than all the other previous algorithm. To understand BERT one need to understand the concept of like RNN,CNN and attention algorithm

In RNN the input is read from "LEFT to RIGHT" are the parameters used by it also are passed from one step to other. The parameters are trained using back propagation algorithm.RNNs become very ineffective when the gap between the relevant information and the point where it is needed become very large as there is a problem of vanishing gradients

2 RNN and CNN

For understanding BERT we need to know about RNN and CNN. Convulsion neural networks(CNN) is algorithm whose roles is to reduce the data into a form which is easier to process, without loosing features which are critical for good performance. This algorithm is very much used while handling an image as in image a huge matrix is converted to a small one by grouping cell and it process the whole input at once unlike RNN it doesn't process input one by one. Unlike RNN where the input is read from "LEFT to RIGHT" and the parameters used by it also are passed from one step to other. .RNNs become very ineffective when the gap between the relevant information and the point where it is needed become very large as there is a problem of vanishing gradient(due to back propagation). LSTM solves the problem little bit as it uses memory cell to memorise the important information but still when huge sentence come it fails. Even using Attention(here each word is passed as an hidden state is passed all way till decoding) RNN isn't preferred as it is too time consuming. CNN overcomes some of the problems as here each word in the input is processed at the same time.

3 Transformers

Transformers uses the combination of CNN and attention(to boost the speed). For example, instead of only paying attention to each other in one dimension, Transformers use the concept of Multihead attention where we pay attention to multiple words

4 BERT(putting all together)

BERT uses Transformers and attention algorithm to find relationship between words in a paragraph. WHAT BERT does it maskes 15 percent of words(15 percent of the words are replaced with [MASK] token) then it is expected to predict the missing word. It basically finds the probablity of each outputs using softmax function. Because of this the BERT is slower. In practice 80 percent of the 15 percent are replaced, rest 10 and 10 are put random and original as 100 percent wont give full accuracy. ALso if you give 90 percent among the 15 percent [MASK] and 10 percent random then model will think the actual word is never correct. Instead of 10 random if you give 10 correct then it will just copy non contextual embedding. NEXT when you give it a sentence it will tell whether the next sentence matches or whether it is random. During training 50 percent of Inputs are paired and others aren't. It uses [CLS](inserted at the beginning of the sentence and [SEP](inserted at the end of the sentence). Then using embedding two sentence are distinguished and positions are given by taking both of them as whole sentence. Now this whole thing goes to a transformer model and output is predicted using softmax()
BERT just like any other neural net is better with more training sets. Also we can fine tune BERT based on our specification like for sentiment analysis, QA etc
BERT is now very preferred as it is better than all other algorithm, has highest GLUE score also the best performer is SQUAD