BERT

Shashank Madhusudhan

August 2019

1 Introduction

BERT (or Bidirectional Encoder Representation for Transformers) is a model developed by Google which trains on language modelling(masked) and next sentence prediction. It basically generates rich word embeddings for words which can be further used in tasks like sentiment analysis, question answering tasks, etc. BERT uses Transformers while modelling. Transformers contain stacks of attention blocks which maps sequences to sequences and is useful in providing context.

2 WORKINGS

2.1 CONTEXT

BERT is better than traditional word embeddings because it takes context into account. It encodes a word by looking at all surrounding words in a sentence. 'He stole from the bank' ,'Bank of a river' are two different sentences which have the same word 'bank'. BERT treats them differently.

2.2 Input embeddings

The input embedding for BERT consists of three layers:

2.2.1 Token embeddings

These convert words into fixed vectors(768 dimensional for BERT). WordPiece tokenisation is used which can split words like 'strawberries' into 'straw' and 'berries' or playing into 'play' and 'ing'.

2.2.2 Segment Embeddings

BERT is trained on Next Sentence Prediction which requires two input sentences. Segment embeddings are used to distinguish the two inputs. The first sentence is assigned to a vector and the second to another vector.

2.2.3 Position Embeddings

BERT uses Transformers which don't have any RNNs or a sequence like structure.Thus, mentioning the position of the word in the sentence becomes vital and positional embeddigs are used for this.

Stacking(summing up element wise) all these three embeddings forms the input for BERT's encoding layer.

2.3 Masked language model

BERT is trained on a masked language model. This means that before the model is fed inputs, some of the word embeddings are masked and the model learns to predict these words. This method is efective because it forces the model to use the entire sequence to predict the masked embeddings.

Here, 15% of input embeddings are masked. Out of this 15%, 80% is masked with MASK token, 10% is left unchanged, and 10% is replaced with random tokens. All of the 15% of tokens because the model may learn to just predict the MASK token and if it didn't see it while fine tuning, it could affect the performance.

2.4 Next sentence prediction

The BERT model was pre-trained on the next sentence prediction task so that it could learn the relationship between two sentences and these embeddings could be further used in downstream tasks like question answering and natural language inference. For the input, half the sentences are paired with the actual next sentence and the other half is paired with random sentences. The two sentences are separated using the SEP token and when the two sentences are combined, 15% of tokens are masked.

For classification tasks, a CLS token is added to the beginning of the sequence. Thus, if the sequence contains to adjacent sentences, the CLS token will read IsNext.

The CLS token needs to be modified for multi class classification problems since it can output only one number. This is done by multiplying the last hidden vector(h dim vector) pointing to the CLS token with its weight(h,k dimension) and applying a softmax to it to give the required number of labels(k).