Technology Review

BERT: General Overview and it's Applications

Author: Snehangshu Shankar Bhattacharjee
Session: MCS-DS Fall-2021 cohort, Course: CS 410, Net ID: ssb8

Introduction

BERT which stands for **B**idirectional **E**ncoder **R**epresentations from **T**ransformers, is a language representation model recently introduced and published by researchers at Google AI Language. It is freely downloadable, predominantly used by Google to enhance user's search experience and it can perform wide variety of tasks such as Neural Machine Translation, Text Encoding & Text Summarization, Similarity Retrieval, Response Selection, Sentiment Analysis, and many others.

Overview

Transformer-based model became very popular and widely used in NLP field due to enhanced parallelization and better modeling, and BERT is the best model which obtained state-of-the-art results in a wide variety of NLP tasks, including Standard Question Answering (SQuAD), Multi-Genre Natural Language Inference (MNLI), Named Entity Recognition (NER). The breakthrough in BERT is, unlike other transformer-based models which can only read text input left-to-right sequence or right-to-left sequence, this model can read both left-to-right sequence and right-to-left sequence, hence entire sequence of words at once. This strategy helps BERT model to learn the context of a word considering its surroundings words.

The building block of BERT is a stack of Transformers encoder layers, it is indeed a fully connected neural networks augmented with a self-attention mechanism. To handle wide variety of NLP tasks, this model needs to have language understanding first and then fine tuning is also required to perform downstream tasks. Therefore, standard workflow for BERT consists of two phases: pre-training (to understand natural language) and fine-tuning (to perform downstream tasks). Books corpus (800M words) and English Wikipedia (2.5B words) were used to pre-train BERT model.

Pre-training:

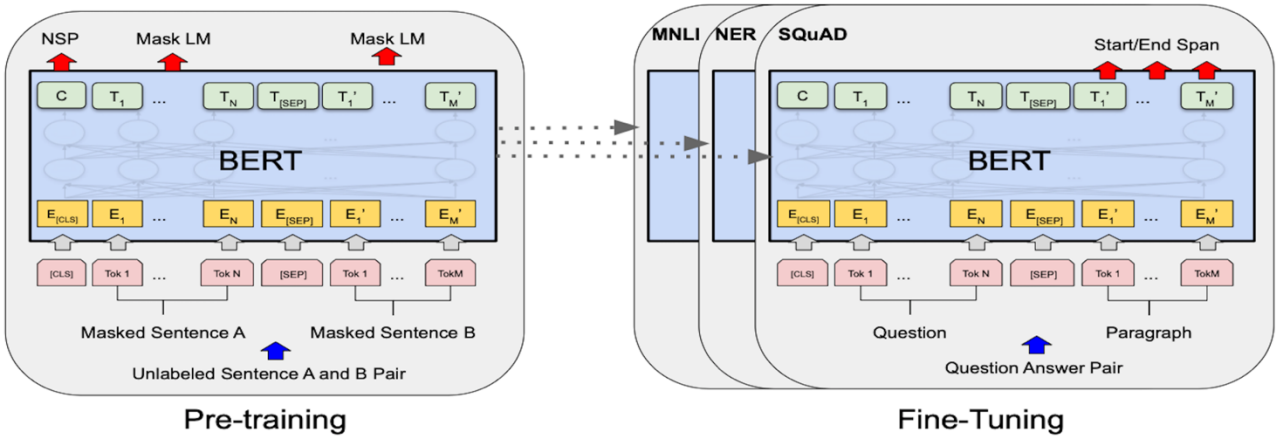
BERT model is trained on unlabeled data in this step. Under Pre- training, there are two semi-supervised tasks: masked language modeling and next sentence prediction.

Masked Language Model (MLM):

The technique of Masked LM is applied to achieve a deep bidirectional training model and the goal is to predict the masked words. In general, 15% of the input words are masked, and then predict the masked words, following the masking procedure as:15% of the words to predict, but don't replace with [MASK] 100% of the time. Instead: 80% of the time, replace with [MASK],10% of the time, replace random word and 10% of the time, keep same.

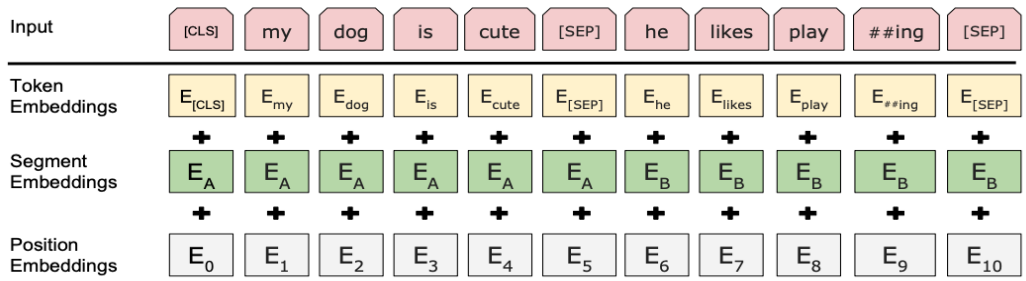
Next Sentence Prediction (NSP):

In this process, pairs of sentences provided to the model as input and learns to predict if the second sentence in the pair is the subsequent sentence in the original context. While choosing the sentences A and B, 50% of the inputs are a pair in which the second sentence is the subsequent sentence (labeled as InNext), other 50% a random sentence from the corpus is chosen as the second sentence (labeled as NotNext). Model can distinguish these sentences during training with the help of special tokens [CLS] – inserted at the beginning of first sentence and [SEP] – at the end of each sentence, and finally calculate the probability of IsNext with the Softmax. NSP is very useful for many downstream tasks like Question Answering (SQuAD), Natural Language Inference (NLI) where the relationship between two sentences is very important.



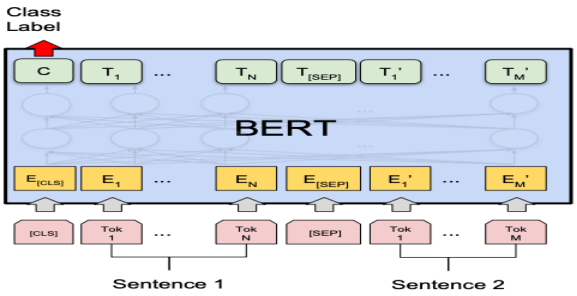
Text Processing:

The input representation can have both a single sentence and a pair of sentences in one token sequence. It also has [CLS] which is a special token used for classification predictions, and [SEP] separates two input segments. To compute input representations, the sentence tokenized into WordPiece and then combines three embedding layers such as token embeddings (the word), segment embeddings (words in the sentence) and position embeddings (position of the word) and then derive a fixed-length vector. These vectors corresponding to the mask tokens are fed into an output Softmax Layer. The Softmax layer has same number of neurons is equal to the number of tokens in the vocabulary. By doing this the word vector can be converted as distribution, and finally when it compared with actual word distribution, and train the distribution using cross entropy loss which is due to masked words only.

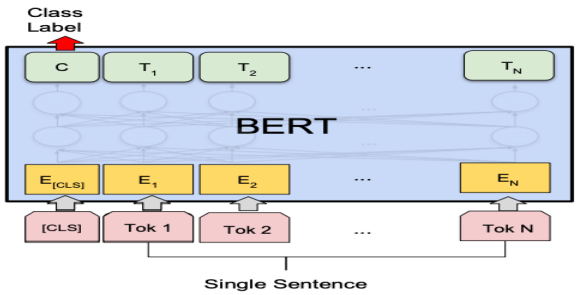


Fine-tuning:

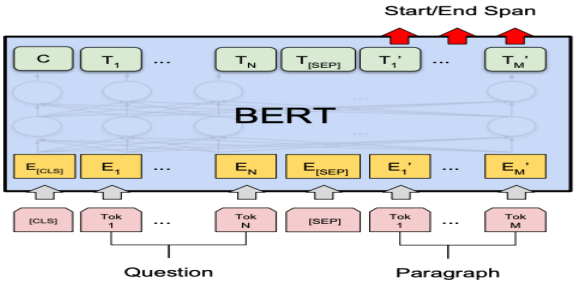
At the end of pre-training BERT has the notion of language, and contextual knowledge, so now (in the fine-tuning step) it can be used to perform supervised training depending on the downstream tasks to solve. In this step, model is fine-tuned with labeled data from downstream tasks, and finally one or more fully connected layers are typically added on top of the final encoder layer during fine-tuning process of this model. Using BERT for a specific task is relatively straightforward because self-attention mechanism allows BERT to model many downstream tasks. For each task, task-specific input and output need to be used to BERT model, and all parameters need to be fine-tuned to perform the task. BERT obtained state-of-the-art results on 11 NLP tasks. Some examples of sequence-level (a) and (b), and token-level (c) and (d) tasks given below:



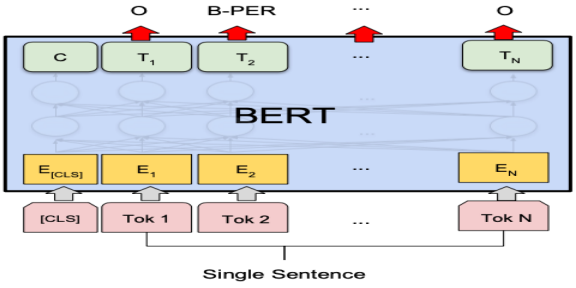
(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG



(b) Single Sentence Classification Tasks: SST-2, CoLA



(c) Question Answering Tasks: SQuAD v1.1



(d) Single Sentence Tagging Tasks: CoNLL-2003 NER

Model Architecture:

BERT is a multi-layer bidirectional transformer encoder; models are available in tensor2tensor library. Primarily there are 2 model sizes, where L: number of transformer blocks, H: hidden size, A: number of self-attention heads. BERT Base model was chosen to have same size as OpenAI GPT, only difference is BERT is bidirectional where GPT can only support left-to-right sequence. BERT Large model can achieve much higher accuracy compared to Base as it has 340M parameter compared to 110M.

- Base: L=12, H=768, A=12, Total Parameters = 110M
- Large: L=24, H=1024, A=16, Total Parameters = 340M

Applications

Below are few examples where BERT has been used as it outperformed previous transformer-based models.

- Smart Search – Google Search integrated BERT to increase search engine efficiency at scale using very popular NLP tasks like Text Encoding & Text Summarization, Similarity Retrieval etc.
- Question Answering – SQuAD v1.1 is a collection of 100k crowd- sourced question/answer pairs, where SQuAD 2.0 added the possibility no short answer exists in the provided para- graph, now the task for the model is to predict the answer text span in the passage based on give question. It does appear that F1 score is improved in this model, SQuAD v1.1 Test F1 to 93.2 and SQuAD v2.0 Test F1 to 83.1.

Conclusion

BERT is empirically powerful model, with the help of Masked LM technique it achieved deep bidirectional training of Transformer-based model which makes it more robust and efficient compared to previous models. Pre-training and fine-tuning workflows are the core of BERT, which makes this model capable of handling wide variety of NLP tasks with the help of semi-supervised learning (language modelling for mased word predictions) and then with the fine-tuning of the model for specific tasks in a supervised way. However, there are few observations where BERT falls short, those limitations are:

- Commonsense & pragmatic inference: it failed to understand the context provided in sentence.

  "He complained that after she kissed him, he couldn't get the red color off his face. He finally just asked her to stop wearing that ". The results suggest it does not have the same sensitivity to what would be the answer in a commonsense scenario as a person.

- Negation: it shows a very strong ability to show associations but cannot handle the negation.

  "A robin is a ____", BERT predicts *bird, robin, person, hunter, pigeon* and *"A robin is not a ____"*, BERT also predicts *bird, robin, person, hunter, pigeon.*

References

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. https://arxiv.org/pdf/1810.04805.pdf

[2] Allyson Ettinger. 2019.What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. https://arxiv.org/abs/1907.13528