Basic Overview

In this paper the authors create an encoder only architecture where they train it to capture bidirectional representations of the text during the pretraining stage, hence with very less finetuning, it would be possible to generate state of the art models. They explain their novel Masked Language Modelling and Next Sentence Prediction pretraining objective.

Model Architecture

The model architecture is identical to that of the transformer's encoder. Here they train two models, one being BERT base L=12, H=768, A=12, Total Parameters=110M) and another BERT large (L=24, H=1024, A=16, Total Parameters=340M). The base model was chosen to have the same model size as GPT for comparison purposes

Pretraining Objectives

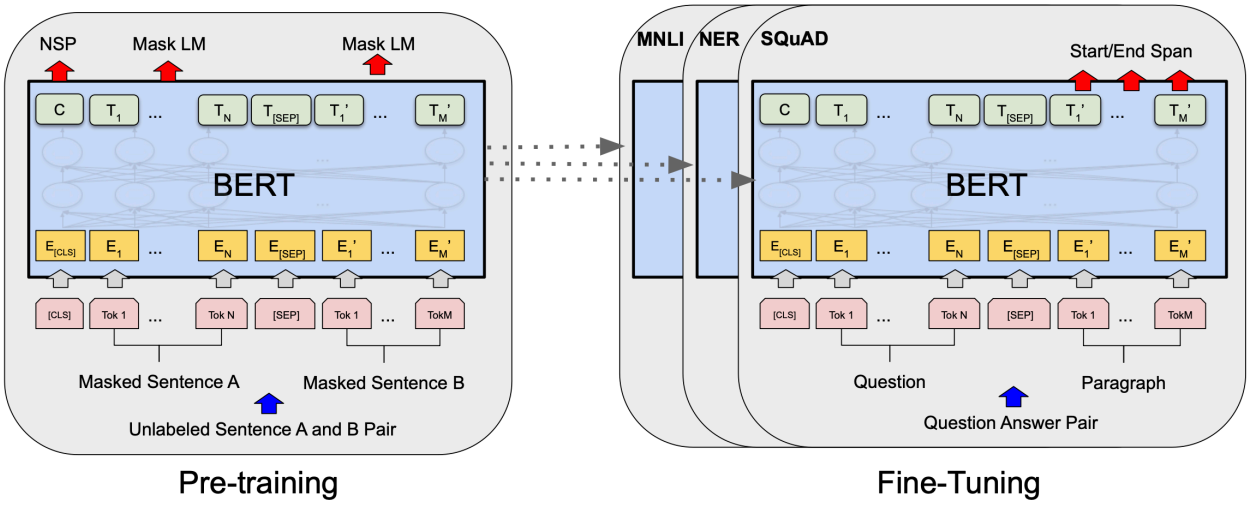
Masked LM

In order to train a bidirectional model without allowing the model to see the word it is predicting (indirectly cheating), the authors simply mask a percentage of the input at random. This is same as the Cloze Task. 15 percent of the tokens are masked for training. Now lets say the ith token is masked, then 80% of the time, the ith token is replaced with the mask token, 10% with a random token and 10% of the time, the token is left unchanged.

Next Sentence Prediction

This task is done for BERT to understand the relationships between two sentences. Here a pair of sentences are sent to BERT with a seperation token inbetween. We also add a segment embedding to sentence A and segment embedding to sentence B to help the model distinguish the sentences further. Now 50% of the time the sentences are consecutive sentences are 50% of the time they are not. Now after the forward pass, the CLS token is used to predict whether the sentences are IsNext or NotNext. Hence by this procedure, BERT is said to have developed sentence understanding.

Finetuning BERT



The general method of finetuning BERT is just to add a task specific head on top of BERT and train that end to end. Now lets see how the task specific head changes for different tasks.

Sequence Classification

This is very intuitive. If it is a sentence classification task, we just add a CLS classifier in the end to finetune it. It is a sentence pair classification, we just add a SEP token inbetween and again have a CLS classifier. For fill in the blanks, we just append the mask token in the place where we want a prediction, take the output of BERT's mask prediction.

Question Answering

Since BERT does not have any generation capabilities, it generate answers for questions on its own. So for question context answering, Two classifier heads are attached to the head of BERT, where one classifier head predicts which one of the tokens would be the start token and the other predicts the end. By this way, BERT is able to select the context where the question's answer lies.

Important Findings

The authors also find that the pretraining task indirectly make BERT more of an ensemble of an LTR and a RTL model with only half of the parameters. And they are measure the significance of the two pretraining tasks.

The authors also find that larger model sizes led to strictly better results even for a very small dataset where the task was substantially different from the pretraining tasks. They also find that BERT took longer to converge compared to an LTR model.

Results

The authors achieve SOTA for 11 different datasets and they claim that this is only due to the deep bidirectional understanding capabilities of the model.