# Technology Review – CS 410, MCS-DS Fall 2021

**BERT** (**B**idirectional **E**ncoder **R**epresentations from **T**ransformers)


## Author : Indranil Guha

## (NETID: iguha4)

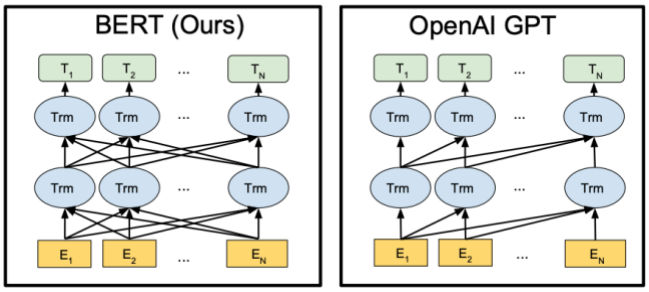

## [Abstract]


**BERT** (**B**idirectional **E**ncoder **R**epresentations from **T**ransformers) is the latest and best known transformer based language model which obtained the state of art results in numerous benchmarks and also integrated in GOOGLE search. This was presented in recent paper [1] published by researchers at Google AI Language. BERT is based on latest state of art NLP techniques building blocks such as attention based transformers, bidirectional encoders. The purpose of this technical review is not to get into the full technical details but to discuss overall framework and how this is useful in several NLP tasks and some of the latest applications.


## [BERT Introduction]

Generative Pre-Trained Transformers (GPT) was first introduced in 2018 by OpenAI Authors based on unsupervised pre-training using the transformer architecture. At very high level the paper proposes using 12 layer transformer decoder with each layer constitutes of multi-headed self-attention layer and positional feed forward layer which produces a distribution over target tokens using SoftMax. But this variation is unidirectional (left-to-right) as the self-attention was attributed only from the left context. Another attribute of GPT series was 'scale'. It was trained on a massive BooksCorpus corpus with 240 GPU days. GPT-1 successfully demonstrated the effectiveness of transformer architecture with at-scale pretraining and little supervised fine tuning which out-performs various NLP tasks.

BERT was published by Google AI shortly after OpenAI GPT which is on similar approach based on multilayer bidirectional attention transformer encoder with a massive unsupervised pretrained model and supervised fine tuning step with addition of using Masked language Model and Next Sentence prediction (NSP) in the pre-training step. The main argument by Google AI authors was GPT's unidirectional pretraining limits the representation of downstream tasks and hence is sub-optimal, e.g. for Question Answering task, context information need to be exploited from both direction.




# [BERT Framework]

There are 2 steps in the BERT Framework pre-training and fine-tuning.

### Pre-Training

The model is trained on unlabeled data using 2 unsupervised task MLM and NSP.

The **Masked language model** (MLM) randomly masks some of the input tokens (15% of corpus) and the goal is the predict the masked token based on its context without restructuring the entire input. This approach fuse the left and right context and allows to train the transformer model bidirectionally. The downside of this is it creates a mismatch in the pre-training and fine tuning as the [mask] tokens does not appear in the fine-tuning stage. If the i-th token is chosen for prediction, then replace the i-th token with (1) [mask] token 80% of time, (2) a random token 10% of time and (3) unchanged i-th token 10% of time.

In addition to this **Next sentence prediction** (NSP) allows the model simultaneously pretrained for text pairs which is not usually directly captured in a language model. For pre-training, when choosing sentence A and B, 50% of time B actually follows A (labeled as IsNext) and 50% of time it's random (labeled as NotNext).

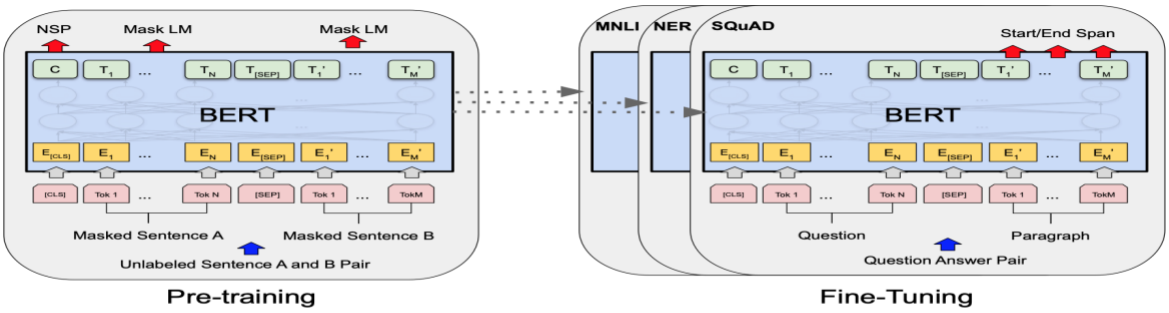



Figure 1: Overall pre-training and fine-tuning procedures for BERT. Apart from output layers, the same architectures are used in both pre-training and fine-tuning. The same pre-trained model parameters are used to initialize models for different down-stream tasks. During fine-tuning, all parameters are fine-tuned. `[CLS]` is a special symbol added in front of every input example, and `[SEP]` is a special separator token (e.g. separating questions/answers).

For the model to be able to distinguish the two sentences in training, input embedding are generated from the word tokens in following way.

Input embedding comprises of 3 embedding vectors such as –

**Token Embeddings** : word pieces. [CLS] token is inserted at the beginning of each sentence and [SEP] token is at the end of the sentence.

**Segment Embeddings:** Sentences number that embedded into a vector.

**Position Embeddings:** Position of word tokens within that sentence that encoded into the vector

Adding the 3 embeddings gives the input embedding vectors goes into BERT. Segment and Position embeddings gives the temporal order of input.

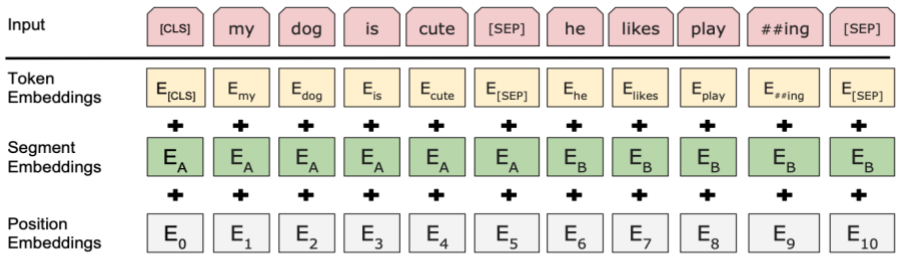



Figure 2: BERT input representation. The input embeddings are the sum of the token embeddings, the segmentation embeddings and the position embeddings.

These input embeddings then processes through the multi-layer transformer encoder and at the output layer these word embeddings then converted to a distribution at the SoftMax activation layer and train the model using the cross entropy loss.

### Fine Tuning

After the BERT model is pre-trained, fine tuning step includes the supervised training specific to the NLP tasks such as SqAD (Question and Answering) , NLI (natural language inference), NER (Named entity recognition) by adding a simple classification layer at the top. In this step it first initialized with pre-trained parameters and all the parameters are then fine tuned using the labeled data from the downstream tasks hence the fine tuning is "Fast".

BERT was proposed in 2 flavors, **BERT BASE** which is 12 layer transformer encoder blocks with 110M parameters and **BERT LARGE** which is 24 layer transformer encoder blocks with 340M parameters.

## [Applications]

Few latest applications of BERT are:

1. **GOOGLE Smart Search –** Google Search integrated BERT. With this, Google now can produce better search results based on search text with increased context awareness.
2. **SciBERT** – This outperforms BASE BERT Model for scientific NLP related tasks as it is trained with The corpus consists of 18% papers on computer science and 82% from

broad biomedical domain, there are other domain specific BERT model such as **BioBERT**, **ClinicalBERT**.

3. **Question Answering and ChatBot** - BERT improved SQuAD (Standard Question Answering Dataset) v1.1 and v2.0 Test score. The same feature of BERT can be used in ChatBot on small to large text.

## [Conclusion]

BERT is no doubt the latest and greatest discovery of pre-trained language model in the field of Natural Language Processing. There are wide range of practical applications due to the fact it's approachable, and fine tuning is very fast. With enough training data and more training step accuracy of many NLP tasks can be enhanced.

## [References]


[1] BERT Main paper https://arxiv.org/pdf/1810.04805.pdf

[2] Overview of BERT https://arxiv.org/pdf/2002.12327v1.pdf

[3] OpenAI GPT https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf

[4] BERT in GOOGLE SEARCH https://www.blog.google/products/search/search-language-understanding-bert/