BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Clement NSHIMIYIMANA

Outline

- Introduction
- Why BERT?
- BERT core components
- BERT implementation
- Conclusion

Introduction

- BERT (Bidirectional Encoder Representations from Transformers) is an Open-Source Language Representation Model developed by researchers in Google AI.

- BERT achieved state-of-the-art performance in tasks like *Question-Answering*, *Natural Language Inference, Classification, and General language understanding evaluation(GLUE)*.

- BERT uses the transformer architecture and self attention mechanism.

- BERT released after OpenAI GPT and ELMo which are unidirectional.

Why BERT?

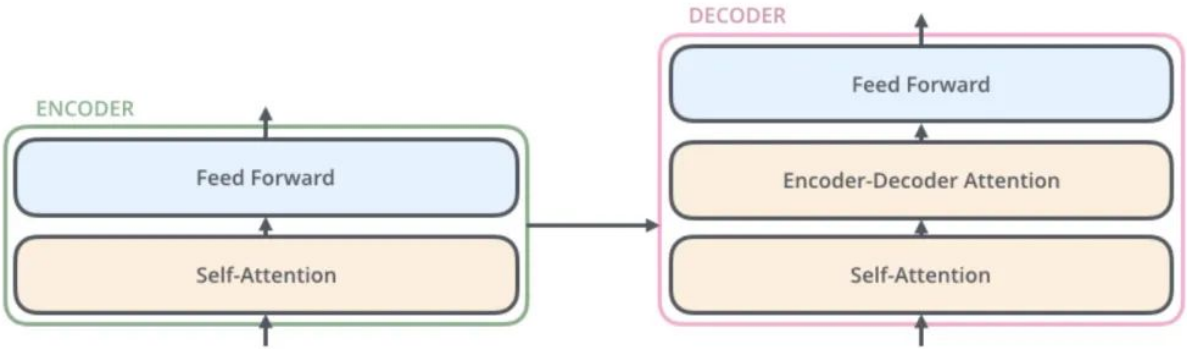
- BERT is a method of pre-training language representations, meaning that we train a general-purpose "language understanding" model on a large text corpus (like Wikipedia), and then use that model for downstream NLP tasks that we care about (like question answering).

- BERT outperforms previous methods because it is the first *unsupervised*, *deeply bidirectional* system for pre-training NLP.

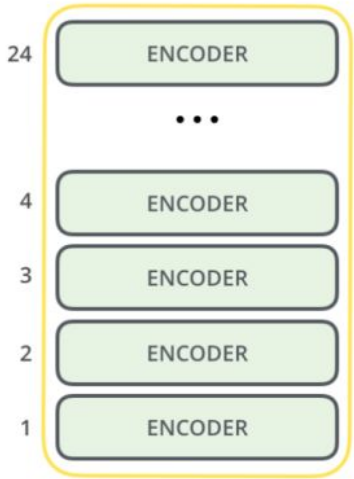
- BERT is the first fine tuning based representation model that achieves state-of-the-art performance on sentence level and token level tasks.

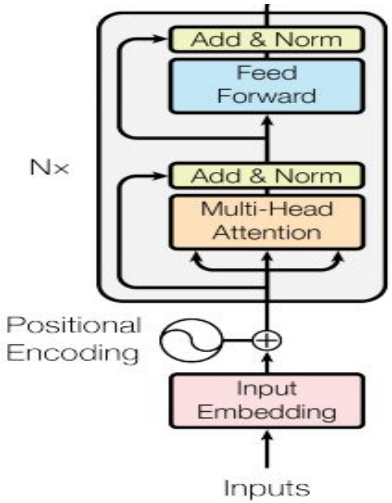
BERT core components

Transformers



BERT uses stacked encoder layers





Used architecture

BERT_BASE	**BERT_LARGE**
Layers =12	Layers =24
Hidden size =768	Hidden size = 1024
Self attention heads =12	Self attention heads =16
Total parameters = 110M	Total parameters = 340M

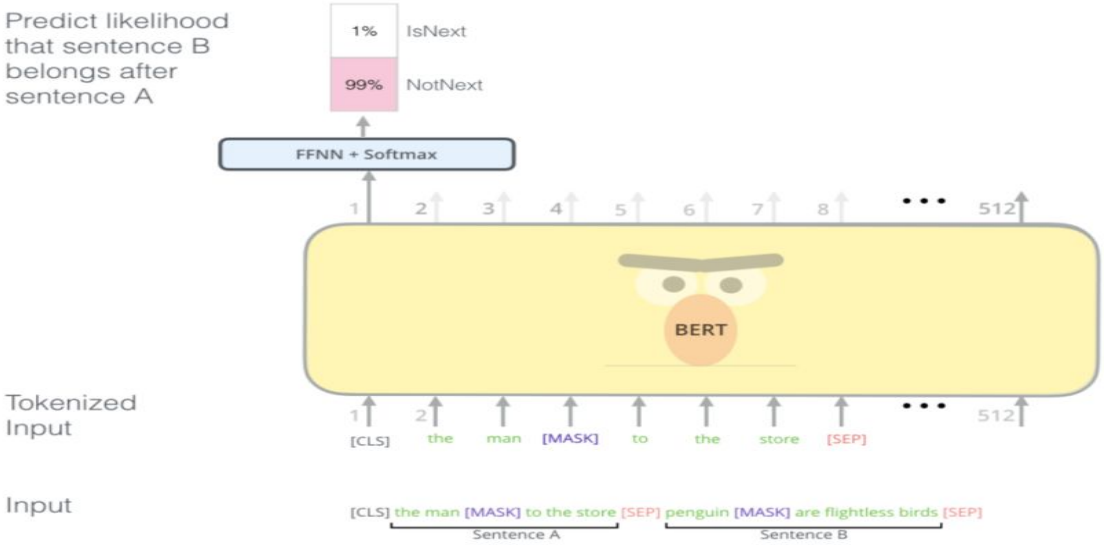
BERT implementation

- Pre Training approach
- Fine tuning approach

Pre Training approach

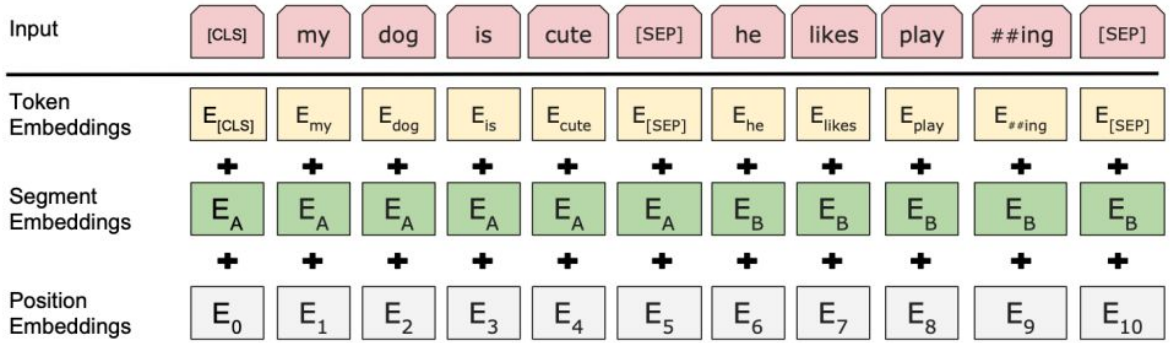
- The model is trained on a large dataset to extract patterns.
- This is generally an unsupervised learning task where the model is trained on an unlabelled dataset like the data from a big corpus like Wikipedia.
- It is fairly expensive (four days on 4 to 16 Cloud TPUs), but is a one-time procedure for each language.

Conceptual representation during pre training



Implementation

- Embeddings: Position embedding, Token embedding, Segment embedding

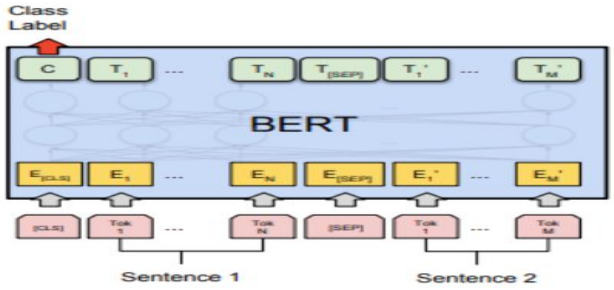


- Encoder layer: Multihead attention, Positional wise feedforward

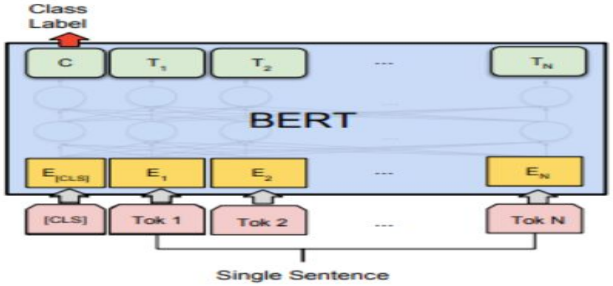
- BERT model assembling all components

Fine tuning approach

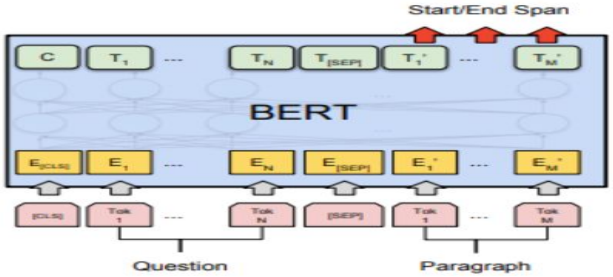
- The model is trained for downstream tasks like Classification, Text-Generation, Language Translation, Question-Answering, and so forth.
- Essentially, you can download a pre-trained model and then Transfer-learn the model on your data.
- It is not expensive



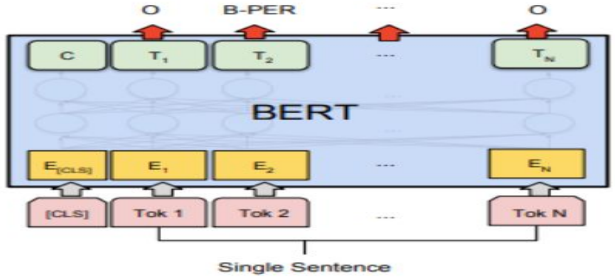
(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG



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



(c) Question Answering Tasks: SQuAD v1.1



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

Results

Task	My result	Paper
QA	67%	87.4%
Entity recognition	99%	90%

Conclusion

- BERT is a very powerful state-of-the-art NLP model.
- The pre-trained model is trained on a large corpus and you can fine-tune it according to your needs and based on the task on a smaller dataset.
- The best thing about fine-tuning is that you don’t do it for 1000 epochs, it can mimic SOTA performances even in 3 to 10 epochs depending on the parameters and how well the dataset is processed.