CS410 Tech Review: Compare BERT and GPT Transformer model in Natural Language Processing

Yong Yang,
yongy3@illinois.edu

October 23, 2023

Abstract

This tech review describes the key design of GPT Transformer and BERT model and compares the differences between BERT and Transformer OpenAI GPT.

1 Introduction

OpenGPT Transformer model is based solely on Attention mechanisms dispending with recurrence and convolutions entirely. The following BERT (Bidirectional Encoder Representations from Transformer) further improved version of Transfomer model with Bidirectional Encoder and earned significant popularity due to its effectiveness in a wide range of NLP tasks.

2 Transformers

The original idea of Transformer was introduced in the paper (Vaswani et al., 2017)(Attention Is All You Need) the key idea was proposed solely on attention mechanisms.

2.1 Transformers model explain

The Transformer model uses stacked self-attention and point wise fully connected layers for both Encoder and Decoder. Both Encoder and Decoder layers are composed of stack of N=6 identical layers. Why N=6 I think it just is an empirical hyperparameter setting. Each layer is built with a multi-head self-attention sublayer and feed-forward layer. In addition to two sub-layers, the Decoder layer inserts a third sub-layer which performs multi-head attention over the output of the encoder stack. An attention function is a key to mapping a query and a set of key-value pairs to an output where all are just vectors. The output is a weighted sum of the values. How much weight is chosen is learned to get the best results. The key point here is the attention functions are just simple additive and dot-product attention. Those are easy to compute in practice. Multi-attention allows the model to attend to different information at different positions or embeddings. This would significantly reduce the computational cost and can be fully parallelized.

2.2 Uniqueness

The Transformer improves over the prior traditional RNN(Recurrent Neural Network) or CNN(Convolutional Neural Network) sequence-to-sequence connecting Encoder and Decoder layer. Recurrent models

typically factor computation along the positions of the input and output sequence because the number of operations required to relate signals from input or output positions grows in the distance between positions. This inherently sequential nature makes it difficult for parallelization computation during training. Attention mechanism allows the modeling of dependencies without regard to their distance in the input or output sequences. The Transformer with attention draws global dependencies directly between input and output, allowing for significantly more parallelization.

The unique points the Transformer model architecture introduced are :

- It allows every position in the decoder to attend all over all positions in the input sequence. there is no hidden layers like traditional RNN.
- Use fixed sine and cosine functions for Positional Encoding, instead of learned. This allows the model to extrapolate to sequence length longer than what is trained, hence reducing the over-fitting problem.
- Self-Attention connects all positions with a fixed number of input sequence, specifically restricted to only a neighborhood in the input around the output position. This builds contextual information which is crutial for natural language understanding, reduced dependencies between layers and improved modularity which make parallel computation easier.

3 BERT

BERT (Devlin et al., 2019) as a front-end variant of Transformer model was introduced to address the directional model issue for analyzing the sequences and tokens holistically. It advanced the state of the art for NLP tasks such as question answering.

3.1 BERT mode explain

As an improvement of the original Transformer, BERT (Devlin et al., 2019) has two steps : pre-training and fine-tuning. Pre-training BERT has 2 tasks. The first task is 'Masked Language Model' (MLM), which is basically to mask some token randomly and predicate those "Masked tokens". It is doing bidirectionally both left-to-right and right-to-left. This is unsupervised learning. The second task is Next Sentence Prediction (NSP).It is pre-trained to understand the relationships between two adjacent sentences. Be aware task 1 is token level training and task 2 is sentence level training. So both tasks together provide more contextual information and deeper semantic understanding of natural language, which is crutial for Question Answering and Natural Language Inference (NLI). The last step of BERT model is fine-tuning for downstream NLP tasks. It has bidirectional cross-function attentions between two sequences. BERT model does it by swapping input and output.

3.2 Uniqueness

The unique points the BERT introduced are :

- Bidirectional representation of Encoder and Decoder.
- Pretained model and fine tuning task for downstream tasks.

Figure 1 shows a overview of BERT model architecture with key pre-training and fine-tuning.

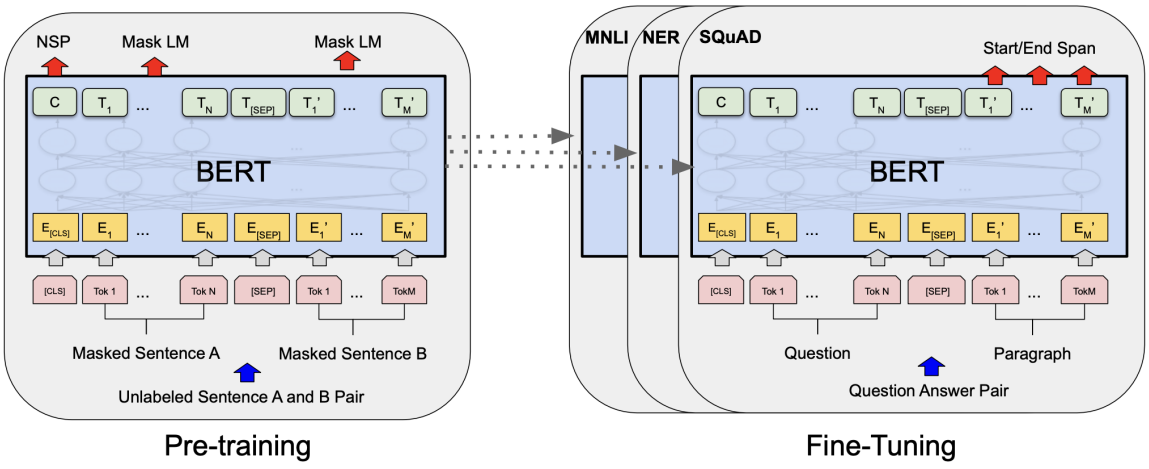

Figure 1: Overall pre-training and fine-tuning procedures for BERT (from (Devlin et al., 2019))

4 Comparison

Table 1 lists the comparison table of original Transformer(GPT) and BERT model.

Comparison	**Original Transformer**	**BERT**
Pros	Versatile adaptable architecture, modular and efficient	Bidirectional representation.Pre-trained with MLM. Strong contextual info. Better fine-tuned performance. Fully parallelization.
Cons	Directional, limited contextual understanding.	High computation. Need large dataset for pre-trained model.

Table 1: Comparison between original Transformer and BERT model

5 Conclusion

In summary, BERT is a specialized Transformer model that excels in capturing bidirectional contextual information, making it highly effective for various NLP tasks. However, it comes with computational and deployment challenges due to its size and complexity. Transformers, on the other hand, are more versatile and modular but may not capture context as effectively as BERT for some tasks.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.