# Technology Review – How does the BERT model beat other traditional language processing models in 2018?

Jiyao Zou

jiyaoz3@illinois.edu

## 1. Introduction

In October 2018, one of the best natural language processing models, BERT [1] was introduced by a team from Google AI language laboratory. BERT model is extraordinary in saving runtimes and generating more accurate results which beat other language processing models on the leaderboard of GLUE, MultiNLI, SQuAD v1.1 and SQuAD v2.0 at the time of launch. This review provides an overview of BERT model, referring to the paper “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”, published by Google AI Language team in 2018.

## 2. Story behind the BERT

The major competitor of BERT is ELMo [2] and OpenAI GPT [3]. ELMo uses unsupervised training approaches to pre-train word embedding by concatenating left-to-right and right-to-left representations. OpenAI GPT used Left-to-right language model and auto-encoder for its pre-training [4]. The name of the model, BERT stands for bidirectional encoder representations from transformers. It made three major improvements compared with its two competitors.

The first improvement of the BERT model is the combination of two strategies, feature-based and fine-tuning, which are widely used for processing natural languages. There are two tasks to solve under the feature-based approach, the Masked Language model,

and Next Sentence Prediction. Most of the computations are finished in this step. Fine-tuning step provides solutions for various downstream tasks without changing the model structure.

The second improvement BERT made is the use of deeply bidirectional transformers by implementing a “masked language model” [5], which randomly masks 15% of the total tokens. This approach ensures more accurate calculations of self-attraction values.

The third major improvement is the Next Sentence Prediction. There is a 50% of probability that the next sentence will be replaced randomly from a corpus which improves the performance on question-answering tasks.

## 3. Conclusion

The contribution of BERT to language processing is significant due to the invention of “deeply” bi-directional transformers rather than the concatenation of left-to-right and right-to-left approaches. This allows parallel computing which reduces time and performance costs in pre-training. Also, the huge dataset of 800M words from BooksCorpus (800M words) [6] and 2,500 million words from the English Wikipedia was used to achieve the high-performance. However, that is also one of the disadvantages of using BERT since a huge dataset always comes with more requirements on computation powers and may slow the run time. The second disadvantage is the conflicts between masked tokens in pre-training steps with unmasked tokens downstream. There is another advantage that both pre-trained and fine-tuning tasks are unsupervised, which will provide additional accuracy to this model.

Overall, the launch of BERT model by Google is a huge change in solving natural language processing tasks while more improvement may be needed to the model in terms of simplicity and better performance on natural language generation tasks.

## References

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:1810.04805, 2018.

[2] Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018a. Deep contextualized word representations. In NAACL.

[3] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding with unsupervised learning. Technical report, OpenAI.

[4] Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In ACL. Association for Computational Linguistics.

[5] Wilson L Taylor. 1953. Cloze procedure: A new tool for measuring readability. Journalism Bulletin, 30(4):415–433.

[6] Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In Proceedings of the IEEE international conference on computer vision, pages 19–27