BERT Model Application Review

Linhan Yang
linhany2@illinois.edu

1 Introduction

The supervised learning models with pretrained text attachments such as word2vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2021) have shown consistently better results than other unsupervised models (Ma et al., 2019). In 2019, BERT was introduced to the industry with unprecedented precision results in many automated word processing tasks (Kenton and Toutanova, 2019). In this paper, the main concept of BERT will be introduced and its main advantages and application will also be discussed.

2 Core Concepts of BERT

The BERT is intended to use for learning sequential text representative bidirectionally in machine learning models (Kenton and Toutanova, 2019). It is easy to use as it will only need to add one extra output layer to the existing architecture, and the result will exceed most existing ones.

Among the previous developed models to learn text representative or other natural language tasks, they often used feature extraction and fine-tuning models. Both approaches use the same pre-training objective function and unidirectional text analysis, which has been pointed out by the authors of BERT that this led to a significant limitation - their focus. Take OpenAI GPT model as an example (Radford et al., 2018), GPT model only considers left-to-right sequence of the text, and the generated text representative can only take the previous context into consideration. But for the tokenized text analysis, this method significantly limits the semantic power of the model, as the meaning of words will depend on the its surrounding context and not only the previous words.

BERT found a way to bypass this limitation by applying so-called ”masked language models” to the training. This means that the objective function is learning from the given representatives to predict a randomly selected and masked word in a text, relying only on the surrounding context. Therefore, a deep bi-directional deep learning model (transformer) was trained.

The BERT architecture is base on the multi-layer transformed introduced by A. Vaswani in 2017 (Vaswani et al., 2017), and was trained into two versions of neural networks (NN) - a standard one with 12 layers and 768 coordinates with total 110 million trained parameters are 110 million) and a larger one with 24 layers and 1024 coordinates with total 340 million trained parameters. The training processes consists of two steps: pre-training with unlabeled data for general use, and additional training with labeled data for a specific application (Kenton and Toutanova, 2019).

Using the processes described above for a specific problem, BERT was tested on several standard datasets to compare with other existing methods and models. On the General Language Understanding Evaluation (GLUE) test, The Standford Question Answering Dataset (SQuAD), and the Situations With Adversarial Generations (SWAG) test, BERT all showed superiority in performance compared to the best-known models (Kenton and Toutanova, 2019).

3 BERTScore for Text Generation Quality Assessment

Since the introduction of BERT in 2019, there are plenty of applications to utilize the BERT model to solve problems. One of the interesting applications is to facility the text generation quality assessment. This assessment has been seen as essential when solving many other problems, like machine translation. However, the most common quantity metrics, like Bilingual

Evaluation Understudy (BLEU) (Papineni et al., 2002), only focus on basic and simple similarity. Like BLEU, it relies on the comparison of the intersection of n-grams of text. So there is a need for a more complex metric to take lexical and semantic meanings of natural languages into consideration when evaluating.

In 2019, researchers from Cornwell University proposed a new metrics based BERT model - BERTScore (Zhang et al., 2019) to fill in this blank. In the core of the BERTScore is a algorithm for calculating the cosine similarity between two vectorized representations of each word based on BERT model. The calculation consists of reference and candidate, and the closest candidate word was chosen for each reference word. The similarity calculation has three main metrics:

- Review Score:

$$R_{BERT} = \frac{1}{|x|} \sum_{x_j \in x} \max_{\hat{x}_j \in \hat{x}} (x_j^T \hat{x}_j)$$

- Accuracy Score:

$$P_{BERT} = \frac{1}{|x|} \sum_{\hat{x}_j \in \hat{x}} \max_{x_j \in x} (x_j^T \hat{x}_j)$$

- F1-Score:

$$F_{BERT} = \frac{2 * P_{BERT} * R_{BERT}}{P_{BERT} + R_{BERT}}$$

In addition, the authors used Inverse Document Frequency (IDF) to determine the rare words, as rare words tends to have more information based on previous studies on text metrics (Vedantam et al., 2015).

However, the original BERTScore proposed in 2019, is slower than the common used assessment metrics, like SacreBlUE. Paralleling computation seems to have the ability to improve the BERTScore algorithm (Zhao et al., 2019).

4 Robustness of BERT learning

Since the publication of the BERT model, some studies and analysis have been done to test and evaluate the robustness of the model and also the structure. In 2019, the Facebook AI group (Liu et al., 2019) used more text data to train the BERT model to test the robustness of the BERT model. In the original paper (Kenton and Toutanova, 2019), the BERT model was trained based on a collection of BookCorpus and Wikipedia, which in total is 16 GB of uncompressed text. While in the later study done by the Facebook AI group, they expanded the data pool to 160 GB. More over, they also applied some modifications in the training processes. For example, instead of using the same sentence with the same masks for all learning epochs, the original sentences were duplicated and randomly masked the text in different locations for each epoch, which is also called dynamic text masking. With these modifications and the expanded training text samples, the authors received improvements in the initial indicators among the common tests - GLUE, SQuAD, and RACE. This leads to a conclusion that original BERT is under-trained.

In the meantime, there are several studies (Joshi et al., 2020) (Lample and Conneau, 2019) (Yang et al., 2019) arguing the necessary of having the next sentence prediction (NSP) training step while training the BERT model. In the original paper (Kenton and Toutanova, 2019), they mentioned that removing NSP from training would lead to a significant drop in model's performance. While in the training results from (Liu et al., 2019) show that removing the NSP function improves the performance on subsequent tasks, which is the opposite of the conclusion from the original paper.

5 Conclusion

In a very short time after the publishing of BERT, the scientific community gives a intense attention to the BERT model and has applied the model in almost all the natural language processing areas. Not taking long, the studies on improving the performance of BERT model start to appear, which leverage the ability of BERT to solve more complex problems.

Undoubtedly, there will be more and more new scientific outcomes based on the application and adaption of BERT models. Further improvements on neural network structures, training processes and fine-tuning parameters will benefit more NLP algorithms and models to serve better for the text classification and characterization related systems.

References

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, pages 4171–4186.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Xiaofei Ma, Zhiguo Wang, Patrick Ng, Ramesh Nallapati, and Bing Xiang. 2019. Universal text representation from bert: an empirical study. *arXiv preprint arXiv:1910.07973*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

J Pennington, R Socher, and MC Glove. 2021. Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding with unsupervised learning.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M Meyer, and Steffen Eger. 2019. Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance. *arXiv preprint arXiv:1909.02622*.