Technical Review on BERT Model

Ningyuan Zhang (nz13@illinois.edu)

1 Introduction

Language understanding has always been a hot topic in both academic and industry areas in the recent decade. Tasks such as question answering in chat bot, language translation in real-time translators, sentiment analysis for movie/product reviews, social media or news, text generation for reports and emails, etc. are all essential NLP applications to people's daily life and can bring great business value in multiple industries once resolved.

Over the years, plenty of complex neural network models have been created and developed for those tasks. And BERT model, short for Bidirectional Encoder Representations from Transformers[1], is one of the most edge-cutting models among all, which mainly serves for language representation and language understanding purposes[1].

In this technical review, the author went through deeply on BERT technique, including what BERT is (training methods and architecture design), what are the innovations and improvements of BERT compared to previous neural network language models, why BERT has such high performance on common tests and how BERT is applied in real life applications based on the authors' own experience.

2 BERT

2.1 Architecture

BERT model actually has a rather simple architecture, it's a multi-layer bidirectional transformer encoder[1]. According to the original paper[1], historically NLP models for language representation and language understanding were mostly designed with unidirectional architecture which didn't allow the model to take advantage of context in both directions. But the deep bidirectional design of BERT allows the model to leverage the context of a token from both directions of the sentence and dramatically increases the information that the encoder can learn.

To be more specific, the bidirectional functionality of BERT model is implemented based on transformer encoders, which consists of 6 layers of stacked multi-head self-attention layer and position-wise fully connected feed-forward neural network layer[2]. This architecture allows the encoding matrix of each token to have access to all

positions in the whole sentence, therefore has the ability to see context in either direction, and it's actually more flexible and powerful than bidirectional RNNs as it can freely assign/learn weights for any position without gradient vanish issue.

Worth to mention, as the transformer architecture requires position information, the input of BERT is the representation of three embeddings[1]: word/token embeddings, segment embedding(for sentence pairs input) and position embedding(for attention layers to work).

2.2 Pre-training

BERT has a two step cooking framework: pre-training over a variety of tasks with unlabeled data and fine-tuning over any downstream tasks with labeled data[1]. For the first step, Masked language model task(MLM) and Next Sentence Prediction task(NSP) were brought in[1]. In MLM, a random portion of the original input tokens were masked and then predicted by the model[1], which trains the model for token representation. In NSP, pairs of sentences are randomly chosen from a corpus where half of them are adjacent, which trains the model to capture sentence relationships[1].

2.3 Fine-tuning

The fine-turning step of BERT depends on the downstream tasks. With case specific labeled data, pre-trained BERT could be easily trained to adapt to applications such as Question Answering(QA), text classification for sentiment analysis, sentence tagging, etc.[1]

2.4 Performance

According to the experiment results by Jacob et al.., with the same parameter size, BERT model has a higher evaluation score than all previous state-of-art systems on all tasks by a substantial accuracy improvement.[1]

2.5 Application Examples

BERT has abundant applications in QA, text classification, text generation or even machine translation, etc. To the author's own experience, transfer learning with BERT comes very handy for NLP tasks in a specific domain, especially when you don't have enough data to train a large transformer from scratch.

By adding several fully connected feed-forward neural network layers at the end of encoders and fine-tuning the model from pre-trained weights with labeled domain data, the author had successfully transfer learnt multiple models from BERT to solve sentiment analysis tasks in fields such as financial news and twitter posts, with performance higher than RNN based model architectures.

With enough training data, it's also possible to pretrain BERT to represent languages from a specific domain, such as fin-BERT which is trained with abundant financial text and performs fairly well in financial new sentiment analysis tasks based on the authors own experiments.

3 Conclusion

All in all, BERT is one of the most advanced and widely used techniques when it comes to language representation and understanding. Its simple architecture and high performance has made it a hot approach in many industry applications. However, BERT is not a perfect silver bullet. The computation of one full attention layer of transformer is still at O(N^2*d), where N is length of sentence and d is the number of hidden states, not much improvement compared to RNN based architectures(O(N*d^2)). However, as far as I am concerned, this is just a small limitation BERT holds and with the dramatic improvement on most evaluation tasks, BERT is definitely the author's first choice when it comes to NLP/NLG tasks.

Reference

1. Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805v2
2. 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 Pro- cessing Systems*, pages 6000–6010.