CS410 Technology Review - BERT

Yiteng Zhang (yiteng3)

November 7, 2022

1 Introduction

Language model pre-training is of great importance in improving many natural language processing tasks such as natural language inference[BAPM15], paraphrasing[DB05], named entity recognition and question answering.[RNS+18][DL15][HR18] Both ELMo and OpenAI GPT, although with different approaches: feature-based and fine-tuning, share the same objective function during their pre-training phase: unidirectional language models.

However, in the paper, it is challenged that unidirectional language models actually restrict the performance of pre-trained representations, since it limits that token can only know previous tokens or later tokens, depends on analyzing left-to-right or right-to-left.

In the paper, BERT: Bidirectional Encoder Representations from Transformers was proposed to improve the fine-tuning based approaches. By using a "masked language model" (MLM) pre-training objective[VSP+17], BERT significantly break the constraint mentioned in the previous paragraph.

2 Body

2.1 BERT

2.1.1 Model Architecture

The model architecture of BERT is based on the original implementation of the multi-layer bidirectional Transformer encoder. The paper denotes the number of Transformer blocks as L, the hidden size as H, and the number of self-attention heads as A.

2.1.2 Input/Output Representations

It is able to unambiguously represent a single text sentence or a pair of text sentences in a sequence of tokens (e.g. [Question, Answer]). And for a given token, its input representation is constructed by summing the corresponding token, segment and position embeddings

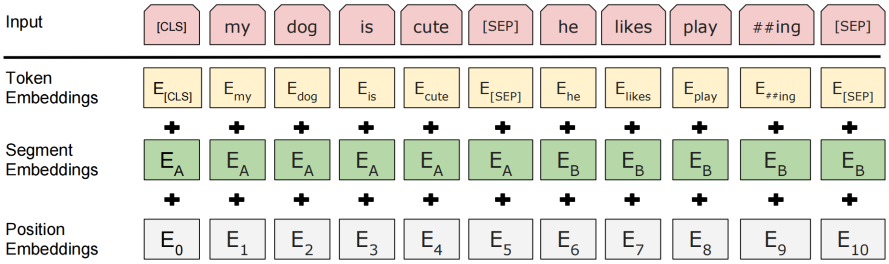

Figure 1: This frog was uploaded via the file-tree menu.

2.2 Pre-training

2.2.1 Task #1: Masked LM

The research team took a simple approach of randomly masking some of the input tokens and then predicting only those masked tokens.

While this does allow the team to obtain a bidirectional pretrained model, this approach has two drawbacks. First, there is a mismatch between pretraining and finetuning, because the [MASK]token is never seen during finetuning. To address this, the team does not always replace the ”masked” words with the actual [MASK]token. Instead, the training data generator randomly selects 15% of the tokens.

The second disadvantage of using MLM is that only 15% of the tokens are predicted per batch, which suggests that the model may require more pre-training steps to converge. The team demonstrated that the MLM converges slightly slower than the left-to-right model (predicting each token), but the experimental improvements obtained by the MLM model far outweigh the increased training cost. [DCLT18]

2.2.2 Task #2: Next Sentence Prediction (NSP)

Question Answering (QA), Natural Language Inference (NLI) and many other important downstream tasks are all based on understanding the relationship between two sentences, which cannot be captured by language modeling.

In order to train a model that understands sentence relations, a binarized next sentence test task is pre-trained, which can be generated from any monolingual corpus. Specifically, when sentences A and B are selected as pre-training samples, there is a 50% chance that B is the next sentence of A, and a 50% chance that it is a random sentence from the corpus.

3 Conclusion

BERT sets new performance records in 11 NLP tasks, for example GLUE[WSM+18], SQuAD v1.1[RZLL16], SQuAD v2.0[SLQL18] and SWAG[ZBSC18]. Plus, it provides us a lot of valuable experience:

3.1 Deep learning is representation learning

Among the 11 tasks in which BERT has reached a new level, most of them only add a linear layer as the output layer based on the fine-tuning of the pre-trained representation. In the task of sequence labeling (e.g. NER), even the dependencies of the sequence output are ignored (i.e. non-autoregressive and no CRF), and the previous SOTA is still killed, showing its powerful representation learning ability.

3.2 Scale matters

The application of this kind of occlusion (mask) to language models is not new to many people, but it is indeed the author of BERT who has verified its powerful representation learning ability on the basis of such a large-scale data + model + computing power . Such models, which can even be extended to many other models, may have been proposed and tested by different laboratories before, but due to the limitation of scale, the potential of these models has not been fully exploited, and unfortunately they have been submerged in the rolling among the paper torrent.

3.3 Pre-training is important

Pre-training has been widely used in various fields (e.g. ImageNet for CV, Word2Vec in NLP), mostly through large models and big data. Such large models can bring geometric improvements to small-scale tasks. The author also gives own answer. The pre-training of the BERT model is done with Transformer, but I think there should not be much difference in performance if it is replaced by LSTM or GRU.

References

[BAPM15] Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*, 2015.

[DB05] Bill Dolan and Chris Brockett. Automatically constructing a corpus of sentential paraphrases. In *Third International Workshop on Paraphrasing (IWP2005)*, 2005.

[DCLT18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[DL15] Andrew M Dai and Quoc V Le. Semi-supervised sequence learning. *Advances in neural information processing systems*, 28, 2015.

[HR18] Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*, 2018.

[RNS$^+$18] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

[RZLL16] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.

[SLQL18] Fu Sun, Linyang Li, Xipeng Qiu, and Yang Liu. U-net: Machine reading comprehension with unanswerable questions. *arXiv preprint arXiv:1810.06638*, 2018.

[VSP$^+$17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[WSM$^+$18] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*, 2018.

[ZBSC18] Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. Swag: A large-scale adversarial dataset for grounded commonsense inference. *arXiv preprint arXiv:1808.05326*, 2018.