A REVIEW OF BERT AND ITS RELATED WORK

Yulei Li

yuleili2@illinois.edu

1 Introduction

Pre-training has been used in computer vision for years[3]. With a pre-trained model, various downstream tasks can achieve favorable results by simple fine-tuning. Before the proposal of Bidirectional Encoder Representation from Transformers(BRET), research in pre-training in natural language processing (NLP) was either relatively less explored or showed rather unsatisfactory performance in both sentence-level and token-level tasks[6]. Meanwhile, every learning task in NLP is more likely to be trained on a small dataset and have a customized neural network for its specific learning task, such as machine translation, text generation, etc. To simplify the complexity of models and further boost the performance of diverse NLP tasks, BERT[2], for the first time, shed valuable light on realizing the training over large language models and demonstrated promising results crossing multiple datasets.

BERT achieved state-of-the-art results on eleven NLP tasks in 2018 without heavily-engineered architecture. The success of this model is mainly because of bidirectional training and the masked language model. Firstly, the rgued that the unidirectional language model for representation learning is sub-optimal for representation learning. For example, every token can only attend to previous tokens in the self-attention layers of the transformer in OpenAI GPT[7], thus may lose potential context information for sentence-level tasks and negatively influence the pre-training. To alleviate this issue, the author proposes to apply a masked language model during the training inspired by the work from [8]. To be more specific, the masked language model will first assign random weights to the tokens. After this, the pre-task is to predict the original vocabulary id of the words/tokens that are masked out on their context. This approach of pre-training will enable the models to capture the overall content of the text and fuse information before and after the masked words. This learning procedure is the reason why their model is bidirectional. Such masking techniques have also been widely used in computer vision and have achieved reasonably sound performance.

2 Methodology

Pre-training BERT is based on (MLM) which ensures the learning of representation at the token level. Accordingly, they also propose Next Sentence

prediction for learning at the sentence level. Such a pre-training method can be based on unlabeled data and consequently save efforts and time for labeling. In the fine-tuning stage, BERT is initialized with the same weights as the pre-trained one. With limited labeled data for each downstream task, the pre-trained models can be adapted and reach satisfactory results.

2.1 Model Model Architecture

BERT is built on a multi-layer bidirectional Transformer encoder according to the implementation of [9]. In the work of BERT, they chose models of two sizes, i.e., $BERT_{base}$ and $BERT_{large}$. Concretely, there are12 transformer blocks and a total of 110M parameters in $BERT_{base}$. For $BERT_{large}$, the model scales up to 24 transformer blocks and 430M parameters. While $BERT_{base}$ is chosen for a fair comparison with OpenAI GPT, the other is used to fully explore the optimal performance of large models(at that time).

2.2 Embedding

The embedding of BERT is obtained from WordPiece with over 30,000 token vocabularies, and it consists of three major parts. For the first word token part, the first position of a sequence is a special one designed for the classification task. For segment embedding, it is used to distinguish different sentences. This property enables pre-trained models from BERT to generalize to broader downstream tasks. For the position embeddings, it is leaned throogh ten percent of training steps, which is different from the previous learning from the trigonometric function[9].

2.3 Masked LM

To fully explore the latent information in the data, the bidirectional model is intuitively more powerful than the single-directional model. To realize this idea, the authors in Bert masked out tokens from inputs and then made a prediction on those missed tokens. While such settings are effective for pre-training with bidirectional information, they also introduce a mismatch between pre-training and downstream tasks, as masked tokens will not be applied during the fine-tuning stages. Instead, the selected token can be either replaced with masked tokens or a random one or kept unchanged.

2.4 Next Sentence Prediction

To further improve the performance of models in understanding relationships between sentences, BERT incorporates a binarized prediction on the next sentence. Half of the time, B is the sequential sentence of A for two sentences noted as A and B for pre-training. For the other half, it is a random sentence from the corpus.

3 Conclusion

Until Oct 10, 2018, BERT shows state-of-the-art performance in pre-training in 11 tasks in NLP. One salient contribution is the model backbones, i.e., the transformer. Compared with RNN, the overall model is more efficient [5] and is able to reach dependency over a more extended period. More importantly, the design of bidirectional context provides more abundant information during the pre-training, thus essentially boosting the performance of the downstream tasks. On the other hand, as only about 15 percent of tokens are predicted, BERT shows a slower converge rate than the single-side model, i.e., left-to-right. Overall, it is a model with intuitive motivations, a straightforward design, and abundant experimental support. Such work opens up a series of work in different areas, inculding NLP, computer vision[10, 4, 1].

References

1. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020)
2. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
3. Erhan, D., Courville, A., Bengio, Y., Vincent, P.: Why does unsupervised pretraining help deep learning? In: Proceedings of the thirteenth international conference on artificial intelligence and statistics. pp. 201–208. JMLR Workshop and Conference Proceedings (2010)
4. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16000–16009 (2022)
5. Mikolov, T., Karafiát, M., Burget, L., Cernockỳ, J., Khudanpur, S.: Recurrent neural network based language model. In: Interspeech. vol. 2, pp. 1045–1048. Makuhari (2010)
6. Peters, M.E., Ammar, W., Bhagavatula, C., Power, R.: Semi-supervised sequence tagging with bidirectional language models. arXiv preprint arXiv:1705.00108 (2017)
7. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al.: Language models are unsupervised multitask learners. OpenAI blog **1**(8), 9 (2019)
8. Taylor, W.L.: "cloze procedure": A new tool for measuring readability. Journalism quarterly **30**(4), 415–433 (1953)
9. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
10. Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R.R., Le, Q.V.: Xlnet: Generalized autoregressive pretraining for language understanding. Advances in neural information processing systems **32** (2019)