A Review on BERT algorithm

Huiyin Zheng
University of Illinois at Urbana Champaign

The BERT algorithm was introduced in 2018 by a couple of engineers at Google. It is an effective approach to improve results for major NLP tasks. This review aims to explain in a simplified manner what is the basic intuition behind the BERT algorithm, how it is innovative, what is the contribution, where it can be applied, where it has been proven to work well, and where there needs to be further research on its potentials.

In short, BERT is a language pretraining algorithm on unlabeled data whose results can be used to improve performance on both feature-based and fine-tuning downstream applications. The basic architecture for BERT is a multi-layer bidirectional transformer, and it differs from the existing literature at that time in that it adopts the bidirectional approach which removes the limitations of the more common unidirectional. It is innovative in its design of the pretraining objectives: 1. MLM ("masked language model" inspired by the Cloze task) and 2. NST ("next sentence prediction"). The ablation studies in the paper explicitly addresses the additional benefit of training with MLM, which improves both accuracy and F1 score compared to a unidirectional language model like OpenAI GPT. On top of that, training additionally with NST further improves the outcomes.

There are several major contributions of the research. To start with, it proves that a bidirectional approach is effective, especially in the downstream fine-tuning tasks under major NLP benchmarks. Secondly, using BERT for pretraining reduces efforts (heavy engineering) and introduces better results for either simple or complex tasks, outperforming the task-specific architectures. Thirdly, it pioneers research on how a sufficiently pretrained model with extreme model sizes can lead to large improvements on small scale tasks. Its results are convincing and well supported by parallel comparisons of model performance under the same contexts.

The paper discussed thoroughly how BERT is particularly useful in the fine-tuning tasks, supported by 11 NLP tasks such as GLUE and SQuAD, etc. As for the feature-based applications (which is also important as not all tasks can fit into the transformer encoder architecture), there are discussions supporting that it can also be achieved with the BERT algorithm, yet it also raises several questions. It is clear that the outcome in the feature-based application is not significantly better or even slightly worse than the pre-existing models, and obviously underperforms the fine-tuning approach. The test F1 score for the feature-based approach is not reported which leads to unanswered doubts. The mechanism of feature-based approach relies on some manual extraction of features, which is similar to feature engineering, but from the pretrained model rather than the raw data. During practical implementation, this could lead to a lot of bias as to which layers or embeddings to choose and questions as to what is the reason behind choosing some of them while eliminating the rest. In the paper, the researchers experimented on many different groups of features, such as embedding, last hidden layer, weighted sum of all layers, concatenation of layers, and etc. It seems the model performance is quite volatile depending on the feature group chosen. This begs the question of whether there can be a more disciplined approach of feature selection based on the BERT model so that it can be better positioned to suit task-specific feature-based applications.

Another point that may require further elaboration or research lies in the MLM training task. In the paper, the researchers mask 15% of the tokens at random, then in order to address the fact that [MASK] token won't appear during further downstream training, for those 15% tokens chosen, there are 80% chance they are substituted with the [MASK] token, 10% they are unchanged and 10% they are a random token. As we may observe, there are many hyperparameters and a lot of randomness in this process. Are these percentages randomly chosen or they are proven to work best according to grid-search or other optimization efforts? If not, is there the need to further research some other possible implementations with different parameters or even slightly different approaches of masking sentences? Those can potentially be some next steps of research.

In conclusion, the BERT algorithm provides a new way of pretraining language model for better language understanding in sentence-level as well as token-level tasks. It can be widely applied in many tasks with either a feature-based or fine-tuning approach.

References

J. Devlin, M.-W. Chang, et al., “BERT: Pretraining of deep bidirectional transformers for language understanding,” in Proceedings of ACL, 2018.