BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, ACL(2018)
Reviewed by Anuja Tayal

Researchers at Google proposed state-of-the-art model **BERT** (Bidirectional Encoder Representation from Transformers) to improve fine-tuned approach while applying pre-trained language representations to down-stream tasks by incorporating context from left-right as well as right-left as opposed to previous models which were unidirectional and sub-optimal for sentence-level tasks.

The best part about the model is that the same architecture is being used for both pre-training and fine-tuning except for the output layers. For pre-training, document-level corpus is used (BooksCorpus (800M words) and English Wikipedia (2,500M words)). Bert model is trained on unlabeled data over different tasks during pre-training while in fine-tuning, model is first initialized with pre-trained parameters and all these parameters are fine-tuned using labeled data from downstream tasks.

For now, they have proposed two models- $\text{Bert}_{\text{base}}$ and $\text{Bert}_{\text{Large}}$ with 12 and 24 transformer layers respectively. WordPiece embeddings with 30k token vocabulary is used in training with the first token reserved as [CLS] token whose final hidden state is used as sentence representation for classification tasks.

Bert Model is pre-trained on two unsupervised tasks mainly- Masked Language Model (MLM) and Next Sentence Prediction (NSP). In MLM, random input tokens are masked and the objective is to predict the original vocabulary id of the masked word based on its context. This is done by choosing 15% of the token positions at random and replacing it with [MASK] token 80% of the time, random word 10% of the time while for the rest, the word is left unchanged.

Whereas in NSP, authors try to understand the relationship between 2 sentences. Sentence pairs are concatenated to form a single sequence with [SEP] token separating them and adding a learned embedding to differentiate the sentences. In the corpus, 50% of the time sentence B was the successor of sentence A (IsNext) while in the rest of the cases it was not (NotNext).

BERT achieves sota performance on a large suite of sentence and token level tasks (11 NLP tasks) by improving the GLUE(General Language Understanding Evaluation) score to 80.5%. They have also done extensive experiments and improvements on accuracy in MultiNLI(86.%) , SQuAD v1 and 2. The best thing I liked is that they have also performed ablation studies to confirm their hypothesis.

Through this model, the authors showed that it would be very easy to plug the model and reduce the need to train heavy models for every task. Some of the issues with BERT is that it is very difficult to implement with scratch, high compute power is needed to reimplement it. And

Secondly, it's hard to believe that the representation of the whole segment can be done just using [CLS] token as it will give higher attention to surrounding words while if the sentence is large, it might not give accurate representation.