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

Razwan Ahmed Tanvir

Date: 04/20/2022

Quote *"... BERT alleviates the previously mentioned unidirectionality constraint by using a "masked language model" (MLM) pre-training objective, inspired by the Cloze task (Taylor, 1953)"*

Overview This paper presents a novel language representation model named BERT (Bidirectional Encoder Representations from Transformers). BERT is a model which is pretrained deep bidirectional representations from text data which is unlabeled. This training is conditioned jointly in all the layers on left and right context. For this reason, this BERT model can be used by fine tuning using one additional output layer to solve several tasks like question answering and language inference without task specific modification of the model. At first, the authors identified issues with unidirectional language models because it throttles the modification of the architecture. Inspired from a work by the Cloze task (Taylor, 1953), BERT models eliminates the unidirectional constraints from the pre-trained model by implementing a "masked language model". This masked model randomly picks a token from the input sequence and masks them. The objective for this mask is to predict the original vocabulary id of that masked token based on the context of that token. Additionally, the authors used a "next sentence prediction" to pretrain the text-pair representation. Authors mentioned that the base BERT model has the following model sizes- $BERT_{BASE}$ (L=12, H=768, A=12, Total Parameters=110M) and $BERT_{LARGE}$ (L=24, H=1024, A=16, Total Parameters=340M).

BERT model shows a great promise with the produced results. State-of-the-art results were produced on 11 NLP tasks. The authors used GLUE score to measure the performance of their models. The authors claimed that their major contribution is that their model is generalised for various tasks in natural language processing.

Intellectual Merit This BERT model has dealt with the issues of unidirectional model and this model is found to be effective for the fine-tuning and feature-based approach both. This research incorporates creative ideas to develop models that could contribute to tackle complex NLP tasks. The authors imparts clear understanding of the research by a well-written document and the experiments were well-organized. The measured the model's performance with state-of-the-art results. The authors are from Google AI language team and they have the required resources to carry this research.

Broader Impact This paper has presented a novel architecture for Bidirectional Encoder Representations (BERT) from Transformers which outperformed the current state-of-the-art models. This BERT model showed better results on 11 Natural Language Processing (NLP) tasks. The authors are from google research team and they also published their code online for public use in github.

Keywords Natural Language Processing, Embedding, Language Model, GLUE

Discussion Questions

- The authors discussed about the size of the BERT models but they did not clearly explain why these values were picked.
- Is BERT usable for the multi-lingual datasets? Will the model provide similar results for any language?

Table 1: Grade deductions by section

Overview	Intellectual M.	B. Impact	Keywords	Questions	Is Online?