William Kiger

wkiger2@illinois.edu

CS 410 Text Information Systems

Technology Review

Bidirectional Encoder Representations from Transformers (BERT)

Bidirectional Encoder Representations from Transformers (BERT) is a Google algorithm that has the ability to take unlabeled text data and join left and right context in all layers and can be used to pre-train a deep bidirectional model that has many use cases [1]. The significance of applying BERT is that the pre-trained model can be modified with a single output layer to make a “state-of-the-art” model that can be used for multitude of tasks such as question answering and language inference without requiring much in the way of architecture modifications. Among the natural language processing tasks that BERT achieves state-of-the-art results are increasing the GLUE score by 7.7%, MultiNLI by 4.6%, a 1.5-point absolute improvement in SQuAD v1.1, and 5.1-point absolute improvement in v2.0 [1].

Historically, there have been 2 approaches, feature-based and fine-tuning, for applying pre-trained language representations [1]. These two strategies can be differentiated by their approaches for handling pre-trained parameters. Fine-tuning focuses on refining all pre-trained parameters, while the feature-based approach treats pre-trained representations as additional features. However, these two approaches share the same unidirectional language models [1]. The main issue with using a unidirectional approach is that each token can only attend to the previous token, using either a left-to-right or right-to-left approach, in the self-attention layers of the

Transformer. This approach limits the model's ability to accomplish tasks such as question answering [1]. BERT's brings innovation to this use-case, achieving better results by implementing a fully bidirectional approach. This bidirectional method allows for the model to understand the context of the word by having visibility of the words it is surrounded [3].

BERT can be downloaded and used with no cost. The user will have the option to use the BERT models to extract language features from their data, or fine tune the models for the user specific use-case [2]. Typical uses for BERT include, question answering, Natural Language Inference (MNLI), and other Natural Language Processing tasks [3]. The authors of BERT claim that they have advanced the state-of-the-art for 11 NLP applications. Additionally, they offer the code and pre-trained models to the open-source community on github.com [1]. The BERT Models also feature the ability to work with some 103 languages [3]. A big advantage of using BERT is that a state-of-the-art NLP model can be developed by adding only one output layer to the pre-trained models [2].

The BERT authors describe two steps in the framework, pre-training and fine-tuning [1]. The pre-training phase is accomplished by training the model on pre-trained unlabeled data in several stages. The second stage, fine-tuning, draws on the initialized BERT model from the first stage and continues in the training of labeled data in the downstream tasks. Each of the downstream tasks creates their own fine-tuned models [1].

BERT is regarded as a breakthrough in Natural Language Processing [3]. Google credits the recent improvements in transfer learning coupled with their approach

to use of unsupervised pre-training as paramount to their successes with BERT [1]. Its pre-trained models are being refined to fit many tasks while achieving the state-of-the-art metrics. The pre-trained models accommodate fast fine-tuning, and its many use-cases are still being explored [3]. The open-source nature of the discoveries completed by the Google AI Language researchers have opened the door for a plethora of users to explore and make new implementations of the 11 prebuilt BERT models.

References

1. Devlin, J., Chang, M., Lee, K., Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding https://doi.org/10.48550/arXiv.1810.04805
2. Jain, Harsch Bert for “Everyone” (Tutorial + Implementation) https://www.kaggle.com/code/harshjain123/bert-for-everyone-tutorial-implementation
3. Horev, Rani. BERT Explained: State of the art language model for NLP. https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270