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

CS410 Tech Review
Shyam Sridharan
NetId: shyams4

Abstract:
This paper serves to summarize the contents of paper: *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,* as well as give an overview of BERT's applications in the field of Text Information Systems.

Introduction:
Previously, pre trained models were limited in their approach, especially fine tuning, because of their unidirectional and choice of architecture used during pre-training. BERT(Bidirectional Encoder Representations from Transformers) is a state of the art language representation model that trains on unlabeled text data by conditioning on both left and right context in all layers and can be further fine tuned with an additional output layer to solve numerous natural language processing tasks. In this review I set out to explain the construction of this model including pre-training and fine tuning and establish the task that can be solved using this model.

Overview:
A key difference between BERT and other model architectures is the unified architecture across different tasks, the difference between its pre-trained architecture and its final downstream architecture is very small. Each BERT implementation takes 4 arguments; L is the number of layers, H is the hidden size, A is the self attention heads and the last argument is the total parameter count. The paper specifies two BERT instantiation that they tested against similar models in this subject space, BERT_BASE (L=12, H=768, A=12, Total Parameters=110M) and BERT_LARGE (L=24, H=1024, A=16, Total Parameters=340M). BERT is first pre-trained on unlabeled data over different pre-training tasks. It is further fine tuned by initializing the model with the pre-trained parameters from each downstream task, which in turn creates separate fine-tuned models for each of these tasks thus the model is specific for each of the tasks at hand. The input for pre training consists of a single sentence and a pair of sentences forming a question and answer relationship. A sentence is composed of specific tokenization for each token vocabulary procured from the WordPiece embeddings, thus the input is the sum of these tokenizations of (sentence, answer sentences). This will further assist us when undergoing pre-training which combines left to right as well as right to left language models for unsupervised tasks.

Pre-training:

For pre-training they use two unsupervised tasks to complete this step, Masked-Lm and Next Sentence Prediction(NSP). Most models either take a left to right or a right to left model approach so the way they circumvent this is using Masked LM. This allows us to mask some percentage of input tokens in order to trivially predict the target word. We don't always replace the masked word with a mask token, instead we designate the mask token 80% of the time with 10% being a random token and the remaining percentage left unchanged. Second task we use for pre-training is NSP. NSP allows us to understand the relationship between two sentences by 50% of the time we label the actual next sentence and the other 50% we label a random sentence from the corpus.

Fine-tuning:

Because BERT allows us to model several downstream tasks, fine tuning is simpler compared to other language models. BERT encodes concatenated text pairs while including bidirectional cross attention between two sentences instead of handling these tasks separately. For each downstream task we plug in the input and output into BERT to fine tune the parameters end to end. This allows fine-tuning to be relatively inexpensive time wise.

Tasks and Evaluations:

Both BERT_BASE and BERT_Large were put up against several previous language models and evaluated against 11 tasks; MNLI-, QQP, QNLI, SST-2, CoLA, STS-B, MRPC, RTE, SQuAD v1.1, SQuAD v2.0, and SWAG. They performed significantly better than the competition with BERT_LARGE actually performing the best amongst the collective of models. From this analysis we see that we BERT is a logical choice for many NLP tasks such as Question Answering and language inference. Also from this we see that increasing BERTs model size is actually very positively correlated to the success of the models analysis.

System	MNLI-(m/mm) 392k	QQP 363k	QNLI 108k	SST-2 67k	CoLA 8.5k	STS-B 5.7k	MRPC 3.5k	RTE 2.5k	**Average** -
Pre-OpenAI SOTA	80.6/80.1	66.1	82.3	93.2	35.0	81.0	86.0	61.7	74.0
BiLSTM+ELMo+Attn	76.4/76.1	64.8	79.8	90.4	36.0	73.3	84.9	56.8	71.0
OpenAI GPT	82.1/81.4	70.3	87.4	91.3	45.4	80.0	82.3	56.0	75.1
$BERT_{BASE}$	84.6/83.4	71.2	90.5	93.5	52.1	85.8	88.9	66.4	79.6
$BERT_{LARGE}$	**86.7/85.9**	**72.1**	**92.7**	**94.9**	**60.5**	**86.5**	**89.3**	**70.1**	**82.1**

Table: GLUE test results for individual model tasks

Conclusion:

This paper shows us the development of BERT as a language model and the importance/significance of its pre training and fine tuning for model preparation. We see that this model is highly scalable for many NLP task usage and performed well in comparison to other

models previously used via GLUE. We also see that this model is relatively inexpensive in order to achieve state of the art performances.

References:

J Devlin, MW Chang, K Lee, K Toutanvoa,
“BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”,
arXiv preprint arXiv:1810.04805 [cs.CL], 2018.