**BERT – Bi-Directional Encoder Representations from Transformers**

**Summary of Technical Contributions:**

* Designed to pre-train deep bidirectional representations from unlabeled text, the model can later be fine-tuned with additional output layer(s) to adapt to a wide range of tasks such as question answering, language inference etc.
* Uses a multilayered bidirectional transformer that reads tokens left to right and right to left at the same time to capture contextual meanings better
* Builds on the standard transformer by introducing two variants:
  + BERTbase:
    - 12 encoder layers
    - 768 Feed Forward Networks
    - 12 Attention Heads
    - 110m total parameters
  + BERTlarge:
    - 24 encoder layers
    - 1024 Feed Forward Networks
    - 16 Attention Heads
    - 340m total parameters
* **How BERT is Different?**

To enable BERT to handle a variety of down-stream tasks it uses both a single sentence or a pair of sentences in one token sequence. The first token of every sequence is a special classification token [CLS]. Sentences are separated using token [SEP]. A learned embedding token is also added on top of that to indicate which token belongs to sentence A and sentence B. A position embedding is also added to indicate the sequence of the token in the text.

BERT consists of two stages:

1. **Pre-training:**

Pre-training is done on unlabeled corpus mainly taken from sources such as the Wikipedia dump or the Books Corpus. Word Piece tokenization is used to cater for out of vocabulary words. Here two techniques are used for pre-training:

**Masked Language Modelling (MLM):**

15% of token positions are chosen at random and masked. The masked tokens are then predicted using the surrounding words in the sentences. To prevent the model from slacking off when it doesn’t encounter a token, of the 15% tokens, 80% of them are replaced with the mask, 10% are unchanged and 10% are swapped with random words.

**Next Sentence Prediction (NSP):**

When choosing sentences A and B for each pair of tokens, 50% of the time B is the actual sentence that follows A and 50% of the time B it is not. The model predicts whether or not sentence B follows A or not.

1. **Fine-tuning:**

BERT can later be fine-tuned on task specific datasets with minor modifications to the output and how input is given. For sequence classification tasks the [CLS] token is taken and fed to a newly added classification layer. For tasks such as Question/Answering, question and paragraph pairs are concatenated together and fed to the transformer. Two new parameters are added at the fine-tuning stage (start and end vectors) to indicate the start and end of the answer from the paragraph.

**Weaknesses:**

* Limitation of maximum sequence length is 512 tokens for the pre-trained models
* In a multi-layered architecture, bidirectional models tend to leak information and allows the tokens to see themselves. To prevent this, MLM is used but this creates a discrepancy between the pre-training and fine-tuning stages as MLM is not present in fine-tuning

**Strengths:**

* Gave state-of-the-art results on eleven NLP tasks
* Improved results for:
  + GLUE score: 80.5%
  + MultiNLI accuracy: 86.7%
  + SQuAD v1.1 question answering Test F1: 93.2
  + SWAG accuracy: 86.3
* The architecture was fairly flexible and could be fine-tuned after pre-training phase to tailor it to a number of different NLP tasks eliminating the need for task specific architectures
* Pre-training the model first and then fine-tuning it reduces the number of parameters that have to be learned from scratch
* Increasing the model size gave better results, BERTlarge performed better than BERTbase in accuracy tests
* Using a bidirectional model is half as expensive as training separate LTR and RTL models and concatenating their results (as done in ELMo)

**Improvements:**

* RoBERTa improved performance over BERT on GLUE, SQuAD and RACE benchmark tests by adding modifications to how masking was done and by removing the NSP phase in pre-training. The model was also trained on larger batches of the data to overcome the issue of undertrained models [1]
* XLNet overcomes BERT’s problem of neglecting the discrepancy in masked representations between pre-training and fine-tuning stages by suggesting an autoregressive pre-training method [2]

**References:**

[1] Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L. &amp; Stoyanov, V.

RoBERTa: A Robustly Optimized BERT Pretraining Approach

2019

[2] Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R. &amp; Le, Q. V.

XLNet: Generalized Autoregressive Pretraining for Language Understanding

2019