BERT

(Bidirectional Encoder Representations from Transformers)

BERT is short for Bidirectional Encoder Representations from Transformers. It's a large language model that is based on Transformer deep learning model. BERT is designed to be pretrained and applies pre-trained representations to downstream tasks using fine-tuning to create state-of-the-art language model for various kinds of natural language processing tasks. BERT uses masked language model pre-training objective which allows to pre-train in two directions. BERT does not train on left-to-right and right-to-left representations independently and then concatenates. The model is trained on unlabeled text data and for fine-tuning it is initialized with pre-trained representation then is fine-tuned using labeled data.

Training data

BERT was pre-trained using text data from BookCorpus which has 800 million words and English Wikipedia on 2.5 million words (using extracted data from passages only). Document level corpus is used for training instead of sentence level corpus as documents have contiguous sequences.

Architecture

BERT is a multi-layer bidirectional Transformer encoder. From architecture standpoint, it is similar to the original Transformer architecture described in Vaswani et al. (2017). Transformer model consist of encoder, decoder, attention, point-wise feed forward networks, embedding, softmax, and positional encoding. Encoder maps input sequence into representations and decoder generates the output sequence using the representations from encoder. At each step, transformer model takes previously generated symbols till that point in addition to the input to generate the next symbol. BERT has two sizes base and large. BERT base has 12 layers, hidden of size 768, 12 self-attention heads and 110 million parameters. BERT large has 24 layers, hidden of size 1024, 16 self-attention heads and 340 million parameters.

Pre-training BERT

BERT is pre-trained using two unsupervised tasks – Masked Language Model (MLM), Next Sentence Prediction (NSP). In MLM, some input tokens are masked at random, and the model predicts those masked tokens. This procedure is also referred to as Cloze in literature. In this task, only masked tokens are predicted instead of reconstructing the masked input with predicted token. MLM trains BERT to understand relationships between words. In NSP, model is given two sentences (A and B) and it must predict whether B is the next sentence or not. For training, 50% input of the sentence pairs are such that B follows A and another 50% is such that B does not follow A. NSP task trains the model to understand sentence relationships.

Fine-tuning BERT

Parameters fine-tuning is done by feeding task specific input and output at each task. Fine-tunning is inexpensive compared to pre-training and can be done on single GPU/TPU in few hours.

GLUE evaluation results

General Language Understanding Evaluation (GLUE) benchmark is a collection of resources for training, evaluating, and analyzing natural language understanding systems. Below are the GLUE evaluation results of BERT base and BERT large models.

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
$BERT_{BASE}$	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

Effect of pre-training

To understand the effect of pre-training there were 2 approaches followed – pre-training without NSP (Next Sentence Prediction), and LTR (Left-to-Right) & No NSP. In case of pre-training without NSP, performance reduced on QNLI, MNLI, and SQuAD. In case of pre-training using LTR instead of MLM and without NSP, there was large drop seen in MRPC and SQuAD. When evaluating it was ensured in both approaches pre-training data, fine-tuning schema, and parameters matched BERT base. Results of the effects are shown below:

Tasks	**MNLI-m**	**QNLI**	**MRPC**	**SST-2**	**SQuAD**
$BERT_{BASE}$	84.4	88.4	86.7	92.7	88.5
No NSP	83.9	84.9	86.5	92.6	87.9
LTR & No NSP	82.1	84.3	77.5	92.1	77.8

Effect of model size

To understand the effect of adjusting the model size on the scores, number of BERT models were trained with differing number of layers, hidden units, and attention heads. It was ensured that for each model same hyper-parameters and training procedures were followed. Results of GLUE tasks are shown below. We can see large language models show improvements in all datasets.

#L	**#H**	**#A**	**LM (ppl)**	**MNLI-m**	**MRPC**	**SST-2**
3	768	12	5.84	77.9	79.8	88.4
6	768	3	5.24	80.6	82.2	90.7
6	768	12	4.68	81.9	84.8	91.3
12	768	12	3.99	84.4	86.7	92.9

12	1024	16	3.54	85.7	86.9	93.3
24	1024	16	3.23	86.6	87.8	93.7

#L = number of layers, #H = hidden size, #A = number of attention heads, LM(ppl) = masked LM perplexity of held-out training data

Conclusion

Unsupervised pre-training is crucial in understanding language. BERT's deep bidirectional architecture approach further enables a pre-trained model to handle various kinds of natural language tasks.

References

Devlin, J., Chang, M.W., Lee, K. and Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I., 2017. Attention is all you need. Advances in neural information processing systems, 30.

https://gluebenchmark.com