Technology Review about Google BERT

Wei Dai-CS410 2020fall

1. Intro

As we all know, Google is dominant in search engine area. Everyday, people ask different questions on Google. From Google's statistic, 15% questions are never seen before. And old search engine could not provide accurate results to those questions. (*cited from: https://blog.google/products/search/search-language-understanding-bert/*) This BERT (Bidirectional Encoder Representations from Transformers) technology is to understand query questions never seen before.

Search's core is about understanding language. Search engine will try to figure out what user are searching for (What we are interested in this class) and present useful info, no matter how the words are spelled or combined. But the difficult parts are complex or conversational queries. For normal search engines, keywords will return much more accurate results compared with actual questions users would ask. With BERT's help, Google can better understand the nuances and context of words in searches and better match those queries with more relevant results.

Machine learning models are widely used in natural language inference and machine translation. BERT's model architecture is a multi-layer bidirectional Transformer encoder. based on the original implementation described in Vaswani et al. (2017) (*cited from: Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Advances in Neural Information Processing Systems, pages 6000–6010)* and released in the tensor2tensor library.

2. Model build

In BERT's development, google's researchers denote the number of Transformer blocks as L, the hidden size as H, and the number of self-attention heads as A.
BERT's team primarily reported results on two model sizes:
$BERT_{BASE}$ (L=12, H=768, A=12, Total Parameters=110M),
$BERT_{LARGE}$ (L=24, H=1024, A=16, Total Parameters=340M).
$BERT_{BASE}$ was chosen to have the same model size as OpenAI GPT for comparison purposes.
$BERT_{LARGE}$ has larger scale and better final result.

All machine learning must have a training framework. For BERT, they have two steps in framework: pre-training and fine-tuning.

For pre-training, the model is trained on unlabeled data over different pre-training tasks.
For fine-tuning, the BERT model is first initialized with the pre-trained parameters, and all of the parameters are fine-tuned using labeled data from the downstream tasks.

Pre-training

Google's reseacher pre-train BERT using two unsupervised tasks, unlike other traditional left-to-right or right-to-left language models. And they use BooksCorpus (800M words) (Zhu et al.,2015) and English Wikipedia (2,500M words) for pretraining.

Task 1 Masked LM
In this task, 15% of all WordPiece tokens are masked in each sequence at random. Then the masked words are predicted with model. (e.g., my dog is hairy → my dog is [MASK])

Task 2: Next Sentense Prediction (NSP)
In order to train a model that understands sentence relationships, google researchers pre-train the model with binarized next sentence prediction task. When choosing the sentences A and B for each pretraining example, 50% of the time B is the actual next sentence that follows A(Labled as IsNext) and 50% of the time it is a random sentence from the corpus(Labled as NotNext).
Final model achieves 97%-98% accuracy in NSP tasks.
From google's paper, Removing NSP significantly hurts the performance on all tests. Compared with No NSP model, $\text{BERT}_{\text{BASE}}$ model can improve 1-4% accuracy in different tests.

	Dev Set				
Tasks	MNLI-m (Acc)	QNLI (Acc)	MRPC (Acc)	SST-2 (Acc)	SQuAD (F1)
$\text{BERT}_{\text{BASE}}$	84.4	88.4	86.7	92.7	88.5
No NSP	83.9	84.9	86.5	92.6	87.9

Fine-tuning

Compared with pre-training, fine-tuning is straightforward. BERT used self-attention mechanism in the Transformer to model downstream tasks.
For fine-tuning, most model's hyper parameters are the same as in pre-training, with the exception of the batch size, learning rate, and number of training epochs.

For each task, Google researchers plug in the task specific inputs and outputs into BERT and fine tune all the parameters end-to-end. At the input, sentence A and sentence B from pre-training are analogous to (1) sentence pairs in paraphrasing, (2)

hypothesis-premise pairs in entailment, (3)question-passage pairs in question answering, and (4) a degenerate text-∅ pair in text classification or sequence tagging. At the output, the token representations are fed into an output layer for token level tasks, such as sequence tagging or question answering, and the [CLS] representation is fed into an output layer for classification, such as entailment or sentiment analysis.

Google's researchers also observed that large data sets (e.g., 100k+ labeled training examples) were far less sensitive to hyperparameter choice than small data sets. Fine-tuning is typically very fast, so it is reasonable to simply run an exhaustive search over the above parameters and choose the model that performs best on the development set.

3. Experiment and results:

The General Language Understanding Evaluation (GLUE) benchmarkis a collection of diverse natural language understanding tasks.
(cited from: Wei Wang, Ming Yan, and Chen Wu. 2018b. Multi-granularity hierarchical attention fusion networks for reading comprehension and question answering. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics.)

This table shows google's BERT's GLUE test results outperform all systems on all tasks by a substantial margin, obtaining 4.5% and 7.0% respective average accuracy improvement over the prior state of the art.

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**

4. Conclusion

In natural language process(NLP) tasks, machine learning is a very hot topic and widely used. Google BERT used unsupervised pre-training and enabled low-resource tasks to benefit from deep bidirectional architectures, allowing the same pre-trained model to successfully tackle a broad set of NLP tasks. One shortage of BERT is that very large data size are required to achieve accurate result, thus, the usage of BERT may be limited. However other technologies such as Google's PRADO and pQRNN - NLP may solve this problem which may worth reading and studying.

Reference:

[1]https://blog.google/products/search/search-language-understanding-bert/

[2] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Advances in Neural Information Processing Systems, pages 6000–6010)

[3]Wei Wang, Ming Yan, and Chen Wu. 2018b. Multi-granularity hierarchical attention fusion networks for reading comprehension and question answering. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)

[4]Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhut- dinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In Proceedings of the IEEE international conference on computer vision, pages 19–27.