BERT with Pre-train on SQuAD 2.0 Context

Chenchen Pan, Liang Xu

{cpan2, liangxu}@stanford.edu

Introduction

BERT achieves the state-of-the-art results in a variety of language tasks. In this project, we replicated the BERT base model, explored the reason behind the BERT's strength. We found that the gain comes from the pre-training on large scale corpus, rather than the architecture. However, the pre-training process reduce the model performance on no-answer questions. So we proposed the idea of pre-training on SQuAD 2.0 context to improve this.

Data

SQuAD 2.0 consists of 100k+ question-answer pairs with corresponding passage, and also contains 50k new, unanswerable questions.

Example:

Input Question:

```
Where do water droplets collide with ice crystals to form precipitation?
```

Input Paragraph:

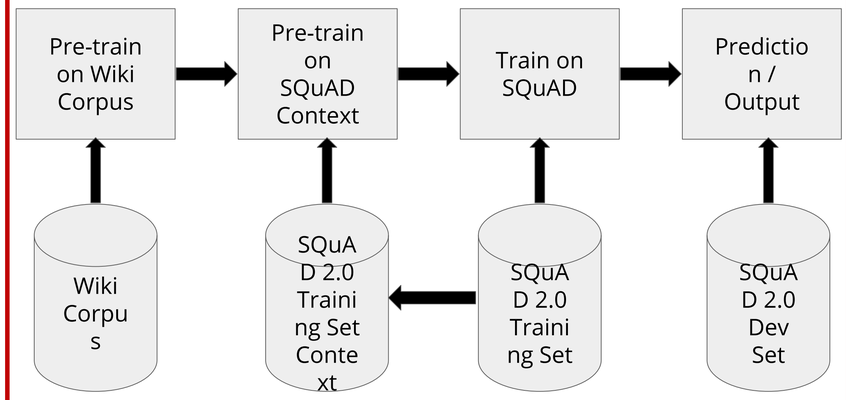
```
… Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals within a cloud. …
```

Output Answer:

```
within a cloud
```

Approach

To know whether the gain comes from the (1) pre-training process; (2) the use of the large scale unlabeled corpus; (3) the architecture, i.e., self-attention.



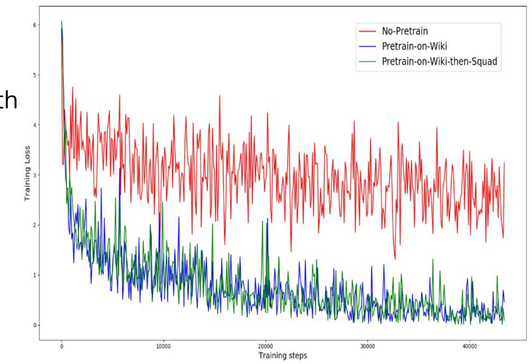
Experiment

No-pretrain vs. Pre-train vs. Pretrain-on-squad

Question Type	Number of questions on Dev Set	Number of Exact Match on Prediction		
		No-pretrain	Pre-train on Wiki	Pre-train on Wiki + SQuAD
Has Answer	5928	146	4270	4239
No Answer	5945	5290	4448	4277

Result

The model without pre-training performs better on No-Answer Questions. The model with pre-training performs better on has-answer questions.
The strength of BERT model more comes from the pre-training process than the architecture.



INFO	EM	F1
BERT + non-pretrain	50.09685842	50.09685842
BERT + pretrain	73.76400236	76.9511166
90k-ftseq_384-run_squad	71.48993515	74.33968292
50k-ftseq_384-run_squad	73.10704961	76.14963027
50k-ftseq_384-run_squad_4epoch	73.36814621	76.58096783

Future

Perform the same approach on BERT-large to get to use the full power of the BERT model.
Tune model configuration for currently pre-trained model to achieve better performance.

References

1. Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for squad. *arXiv preprint arXiv:1806.03822, 2018.*
2. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805, 2018.*