Exploring BERT for Question Answering on SQuAD 2.0

Stanford CS224N Default Project

Alvin Hou
Department of Computer Science
Stanford University
alvinhou@stanford.edu

Fang-I Hsiao
Department of Electrical Engineering
Stanford University
fihsiao@stanford.edu

1 Key Information to include

- External collaborators (if you have any): N/A
- Mentor (custom project only): N/A
- Sharing project: N/A

2 Research paper summary (max 2 pages)

Title	BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding
Venue	Association for Computational Linguistics (ACL)
Year	2019
URL	https://www.aclweb.org/anthology/N19-1423/

Table 1: Bibliographical information of BERT [1].

Background. Before BERT was published, pre-trained language models had already made some progress in multiple NLP tasks, such as natural language inference, named entity recognition and question answering. The authors cited two important papers ELMo [2] and GPT [3] that were published in 2018. The reasons why these works were important are the following. First of all, ELMo and GPT made a huge breakthrough in NLP which utilizes large scale pre-training. BERT also leveraged the power of rich, unsupervised pre-training to achieve state-of-the-art results on many tasks. Moreover, ELMo and GPT are both strategies that could apply pre-trained language representation to downstream tasks, which is similar to how BERT was designed.

The authors pointed out that there were restrictions in the previous techniques for pre-trained language representations, especially for fine-tuning approaches such as GPT. The authors mentioned that one main flaw is that the structure of standard language models are unidirectional, and this limits the possible choice of architecture for pre-training. For example, GPT uses masked self-attention layers that could only receive information from previous tokens. The authors stated that this is sub-optimal for sentence-level tasks and could be harmful while fine-tuning tasks that require context from both directions.

In this paper, the authors aim to approve the fine-tuning based approach for pre-trained language representations.

Summary of contributions. The paper proposed a new way to pre-train the language model, using a "masked language model" (MLM) pre-training objective. The masked language model randomly masks some of the tokens from the input, and the objective is to predict the original word. Unlike

the GPT approach which is unidirectional, using MLM enables us to pretrain a deep bidirectional Transformer.

This is a simple and brilliant way to train the bidirectional Transformers to overcome the previous limitations from standard left-to-right language models. The concept is also close to the human "Cloze Test" for reading comprehension while students try to guess the masked word from the whole sentence context. The MLM objective enables the representation to fuse the context from left and right. The authors demonstrated the importance of bidirectional pre-training for language representations and showed how it helps for fine-tuning tasks that require context from both directions.

The paper also shows how pre-trained representations reduce the need for people to heavily engineer specific architectures for specific downstream tasks. BERT is the first fine-tuning based representation model that achieves state-of-the-art performance on several sentence-level and token-level tasks.

The new experimental results from BERT is also fascinating. BERT advances the state of the art for 11 NLP tasks, including GLUE, SQuAD v1.1, SQuad2.0, and MultiNLI.

Limitations and discussion. The authors had released their pre-trained models[1] that trained on several different languages, including English, Chinese and Multilingual. As native Mandarin Chinese speakers, we found that there may be some limitations and problems while BERT is pre-training on a Chinese corpus. Due to the MLM mechanism, BERT will randomly mask WordPiece tokens from the sentence during pre-training. However, the masked WordPiece token might belong to a whole word in Chinese (a whole word in Chinese might contain more than one word) and hence it will destroy the original meaning of the whole word. There is another paper published by Yiming Cui et al. (2019) [4] that focuses on how to tackle this problem with a Whole Word Masking technique.

However, we still think the paper is convincing. BERT improved the XNLI score for Chinese from 67.0 to 77.2 with its BERT Chinese-only model [2].

Why this paper? As mentioned in the project hangout, Transformers have started showing as the dominant model family for dealing with NLP tasks since 2018, where ELMo achieved state-of-the-art results in 6 different benchmarks. However, BERT absolutely shocked people after it was published. It not only advanced the state-of-the-art results for 11 NLP tasks, but also outperformed human performance in question answering tasks (SQuAD 1.1). Therefore, we decide to read this paper to have an in-depth view of how Google AI designed its stacked Transformer models and what tricks they implemented during the pre-training process.

We did gain what we were hoping and we think this paper is fairly important since there were a lot of BERT variants and improved models after this paper had been published. Some recent work that aims to improve the results of BERT includes XLNet [5], which was published in 2019 and outperformed BERT on 20 tasks. In the same year, FAIR also proposed a new model RoBERTa [6], and showed how they robustly optimized BERT to obtain new state-of-the-art results that even outperformed XLNet.

Wider research context. This paper showed marvelous results on how to build better representations of language with large scale bidirectional pre-training. BERT could also be applied and fine-tuned for a broad range of different NLP tasks. The authors illustrated how BERT could be fine-tuned on sentence pair classification tasks, single sentence classification tasks, question answering tasks and single sentence tagging tasks in the original paper. Due to the huge success that BERT had achieved, there were several improved models and new researches based on BERT, such as RoBERTa and ALBERT [7].

Moreover, the clever way the authors design the model enables researchers and engineers to easily fine-tune different downstream tasks with the same interface. We see a large number of BERT applications such as bert-as-service [8] and even Google are applying BERT models to Google Search now [9].

[1] `https://github.com/google-research/bert`

[2] `https://github.com/google-research/bert/blob/master/multilingual.md`

3 Project description (1-2 pages)

Goal. In this project, our main goal is to achieve a performance of question answering as good as possible by using deep learning models based on BERT. We would like to explore different techniques and apply them to BERT to see if we could improve the performance. We choose this goal because BERT-based models have been successful in many tasks and it would be a good chance for us to dive into the detail of BERT and learn how to utilize such a successful pre-trained model to improve question answering systems. Since our project largely relies on BERT, we choose to read the original BERT paper to get more familiar with it. Besides, if we have time, we would also like to explore semi-supervised learning for question answering. We are interested in how to leverage a small set of labeled data and how to use unlabeled data to achieve a high performance. This secondary goal is very important because it can save a lot of time and cost if we can reduce the required amounts of data but still be able to build a model that has a comparable performance.

Task. The task is question answering, where our system is given a paragraph and a question and need to output an answer which can be found in the paragraph.

Data. We will use SQuAD 2.0 as our dataset which contains around 150k questions.

Methods. We would start with fine-tuning BERT on the question answering task. We would try to use the full potential of BERT by including the sentence embedding into the training process. One approach is to jointly train our model with an auxiliary task, such as a classification task of whether this sentence has an answer. As a kind of additional regularization, we hope this could boost the performance of the main QA task. We also will look into different variants of BERT, such as RoBERTa or ALBERT and see how it performs against BERT. We then could explore other techniques to boost the performance, such as ensemble techniques or data augmentation. The more models and data you have usually the better performance you get.

In order to achieve our second goal, we would like to explore semi-supervised learning techniques, such as Pseudo-Labeling. We hope to achieve a reasonable performance through a limited amount of data. Because BERT is a pre-trained model, we expect it to have a reasonable performance even it only uses small amounts of labeled data. Therefore, it might be feasible to label unlabeled data by predicting the labels from our model. This enlarges the training data while we train our model and might improve the performance when the model uses these new labeled data.

Baselines. For our main goal, since we aim to improve the performance of BERT-based models, our baseline is the scores published in the original BERT paper. If our model is based on different variants of BERT, we would also include them as our baselines.

For semi-supervised learning, we will use the model trained on a small part of the dataset without any semi-supervised techniques as our baseline.

Evaluation. The two metrics we will use are **EM** and **F1**. The original BERT achieve 80.005 **EM** and 83.061 **F1** and the state-of-the-art method achieve 90.115 **EM** and 92.580 **F1** on SQuAD 2.0.

References

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2018.

[2] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proc. of NAACL*, 2018.

[3] Alec Radford and Ilya Sutskever. Improving language understanding by generative pre-training. In *arxiv*, 2018.

[4] Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Ziqing Yang, Shijin Wang, and Guoping Hu. Pre-training with whole word masking for chinese bert, 2019.

[5] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. Xlnet: Generalized autoregressive pretraining for language understanding, 2019.

[6] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019.

[7] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations, 2019.

[8] Han Xiao. bert-as-service. `https://github.com/hanxiao/bert-as-service`, 2018.

[9] Pandu Nayak. Understanding searches better than ever before. `https://www.blog.google/products/search/search-language-understanding-bert/`, 2019.