The Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20)

Distill BERT to Traditional Models in Chinese Machine Reading Comprehension (Student Abstract)

Xingkai Ren, Ronghua Shi, Fangfang Li*

School of Computer Science and Engineering, Central South University, China

{renxingkai, shirh, lifangfang}@csu.edu.cn

Abstract

Recently, unsupervised representation learning has been extremely successful in the field of natural language processing. More and more pre-trained language models are proposed and achieved the most advanced results especially in machine reading comprehension. However, these proposed pre-trained language models are huge with hundreds of millions of parameters that have to be trained. It is quite time consuming to use them in actual industry. Thus we propose a method that employ a distillation traditional reading comprehension model to simplify the pre-trained language model so that the distillation model has faster reasoning speed and higher inference accuracy in the field of machine reading comprehension. We evaluate our proposed method on the Chinese machine reading comprehension dataset CMRC2018 and greatly improve the accuracy of the original model. To the best of our knowledge, we are the first to propose a method that employ the distillation pre-trained language model in Chinese machine reading comprehension.

Introduction

Teaching machines to understand text and answer relevant questions is a long-term and challenging task. Recently, BERT (Devlin et al. 2019) was proposed, which using bidirectional encoder representations from transformers. BERT significantly improve state-of-the-art in various natural language processing tasks. More recently, CMU has described XLNET (Yang et al. 2019), a state-of-the-art, combing the advantages of autoregressive and autoencoding method model trained on even more data. These proposed pre-trained language models have achieved the most advanced results in machine reading comprehension tasks. However, the shortcomings are also promise. They all need huge training corpus for training, which takes time and machine costs. In practice, there are few sufficient resources to complete the training of pre-trained models. Traditional machine reading comprehension models can be trained faster but with lower accuracy than pre-trained language models. Therefore, in order to balance the reading comprehension accuracy and training time. We propose a method that employ the knowledge distillation method (Hinton, Vinyals, and Dean 2015) to distill the pre-trained language model on the traditional reading comprehension model. So that the traditional reading comprehension model can achieve higher inference accuracy.

We evaluate our work on CMRC2018 (Cui et al. 2019) dataset which is a span-extraction dataset for Chinese machine reading comprehension. Our experiments demonstrate that the traditional reading comprehension model after distilling can achieve higher accuracy and faster inference speed and the parameter amount is greatly reduced. To the best of our knowledge, we are the first to propose a method that employ the distillation pre-trained language model in Chinese reading comprehension.

Approach

Knowledge distillation method requires a teacher network and a student network. In the same task, the teacher network can achieve higher accuracy than the student network. We use BERT (Devlin et al. 2019) as our teacher network and treat QANET (Yu et al. 2018) as our student network. Because the encoder of QANET only consists of the convolutional layers and the self-attention layers. It trains and infers faster than other LSTM based encoder models.

In our proposed method, we train the QANET to mimic the full output distribution of the BERT. Firstly, we employ BERT to train on the CMRC2018 dataset to get the output distribution P on the training set. Then let QANET directly learn the training set distribution P of the BERT output. Since the output of machine reading comprehension is the probability distribution of the start and end posi-

*Corresponding author.

tions and it is not a hard target. We can not use the traditional cross entropy loss function to train QANET. For distilling the teacher network, we will employ the Kullback-Leibler (Kullback and Leibler 1951) loss since the optimizations are equivalent:

$$KL(p \| q) = E_p(\log(\frac{p}{q})) = \sum_i p_i * \log(p_i) - \sum_i p_i * \log(q_i)$$

where $p_i \in P$ refers to the BERT training output probability distribution and $q_i \in Q$ refers to the QANET training output probability.

For part on the CMRC2018 dataset whose content length exceed the maximum input length of the BERT, we employ a simple but efficient method. We segment the content into sentences and use TFIDF algorithm to calculate the similarity score between each segmented sentence with question. Then we choose 10 of the highest-scoring sentences to represent the content. So that the length on the content is shorter than the maximum input length of the BERT. In this way, the answer to the question covers 83.3%.

Experiments

Dataset. We conduct the proposed method on the CMRC2018 dataset which has 10,321 questions in the training set and 3,351 questions in the validation set. The entire dataset annotated by human on Chinese portion of Wikipedia paragraphs. Further details are provided in the supplement.

Implementation Details. For teacher network, we employ PyTorch-Transformers[1] to implement $\text{BERT}_{(\text{base})}$. The max sequence length, the question length and the answer length are set to be 512, 40, 30. The learning rate is set to be 2e-5. We fine tune the training set with 3 epochs and save the probability distribution of the start and end in the answer. For student network, we implement QANET as our student network. The sequence length, the question length and the answer length are the same with BERT. We employ the 300-D pre-trained fastText embeddings and keep it fixed during training. The learning rate of QANET is set to be 0.001.

Results. We employ three indicators of inference accuracy, inference speed and model parameters quantity to evaluate the method on the validation set. The number of BERT parameters we use is about 110M but the number of parameters for QANET is approximately 30M. The time required for BERT to inference on the validation set is about 12 minutes. But it only requires 7 minutes for QANET to inference on the validation set with the same batch size we set to 256. The accuracy results are shown in Table 1. We can see that the distilled QANET greatly improve the accuracy on the validation set. Besides, the inference speed and model parameters of distilled QANET are much lower than the BERT. It is helpful to apply the distilled model in practice.

Model	EM	F1
$\text{BERT}_{(\text{base})}$	64.761	85.369
$\text{QANET}_{(\text{undistilled})}$	53.079	72.538
$\text{QANET}_{(\text{distilled})}$	59.831	79.893

Table 1: Performance on the CMRC2018 validation set.

Conclusion and Future Work

In this paper, we aim to distill the enormous pre-trained language model to the traditional reading comprehension model especially in the field of Chinese machine reading comprehension. We conduct experiments on CMRC2018 with QANET distilled by BERT. The results show that the distilled traditional reading comprehension model can achieve higher accuracy and faster inference speed with smaller parameters. In the future, we will plan to explore more distillation methods in the field of machine reading comprehension.

References

Cui, Y.; Liu, T.; Che, W.; Xiao, Li.; Chen, Z.; Ma, W.; Wang, S.; and Hu, G. 2019. A Span-Extraction Dataset for Chinese Machine Reading Comprehension. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, 5882-5888.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the Conference on the North American Chapter of the Association for Computational Linguistics, 4171-4186.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the Knowledge in a Neural Network. arXiv preprint arXiv:1503.02531.

Kullback, S., and Leibler, R.A. 1951. On information and sufficiency. *The annals of mathematical statistics*, 22(1), pp.79-86.

Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R.; and Le, Q. V. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. arXiv preprint arXiv:1906.08237.

Yu, W.; Dohan, D.; Luong, M.; Zhao, R.; Chen, K.; Norouzi, M.; and Le, Q. 2018. QANet: Combining Local Convolution with Global Self-Attention for Reading Comprehension. In Proceedings of the International Conference on Learning Representations.

[1] https://github.com/huggingface/pytorch-transformers