Unsupervised reinforcement knowledge distillation

Dongheng Li
dongheng@ualberta.ca
(draft)

Abstract

Model compression techiniques are widly applied to large language models, especially for Pre-trained large language model, such as BERT (Devlin et al., 2019), those have improved the performance of many NLP tasks. However, the large language model has high requirements for computational expenses. To accelerate the inference and model size, two popular methods, weight pruning and knowledge distillation, are widely applied in the NLP area. We are going to focus on knowledge distillation (KD) (Hinton et al., 2015), specifically on Transformer-based ones. We use reinforcement learning (RL) to train a small language model, where the reward function is derived from a large language model. The reward derivation is based on previous work (Hao et al., 2022).

1 Introduction

Knowledge distillation (Hinton et al., 2015) is a very popular model compression technique. Applying KD on transformers are in favor of us researchers. The key of knowledge distillation is to pass the linguistic information from teacher model to student model. In our context, it is from pre-trained large language model (T0) to a small ones(T5-base). Unsupervised reinforcement learning trains the T5-base with pseudo-label generated by T0 guided by naturally designed prompt templates. The idea is to pass the knowledge from T0 to T5 using RL as knowledge distillation method.

2 Related Work

Classic-KD The concept of Knowledge distillation was first introduced (Hinton et al., 2015) where it emphasizes the knowledge are measured and passed through cross-entropy training with different loss functions which are designed to make the student model imitate the teacher model.

Sequence-KD Similar to the classic-KD, sequence-KD (Kim and Rush, 2016) focuses on passing the knowledge in sequence-level instead of word by word. The main contribution is it solved the intractability of combinations within a sequence which can be exponentially large.

Transformer-KD Transformer-KD (Jiao et al., 2020) based on transformer structure. It designed three different loss functions in order to capture as much linguistic knowledge as it can to pass to student models.

....

Those are only three related work briefly introduced here. In the following part, a reading list is provided here to further improve the understanding.

3 Limitations

While reading through those papers, there are some limitations and main contributions we would like to mention here.

For Transformer-based-KD, the key limitations are to design the loss function to capture the “knowledge”. TinyBERT (Jiao et al., 2020), for example, only deal with encoder-only transformers and require a lot self-designed loss function. On the other hand, in an ideal scenario, our RL exploration would capture all information and works on any kind of transformers.

4 Reading list

This list[1] contains more than 20 papers related to knowledge distillation and including theory and

[1] https://docs.google.com/document/d/17s3U2cWtjzbW_uiPb4yhxjiE-MUR63uPls7j0QD6dyw/edit

best practices. We mark them in red and blue fonts, respectively.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Yongchang Hao, Yuxin Liu, and Lili Mou. 2022. Teacher forcing recovers reward functions for text generation. In *NeurIPS*.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. Cite arxiv:1503.02531Comment: NIPS 2014 Deep Learning Workshop.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2020. TinyBERT: Distilling BERT for natural language understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4163–4174, Online. Association for Computational Linguistics.

Yoon Kim and Alexander M. Rush. 2016. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1317–1327, Austin, Texas. Association for Computational Linguistics.