KDABert: Knowledge Distillation with Generative Adversarial Networks for Bert Model Compression

闫森，　　张伍豪，　　胥进

2017111497，2017111497，2017111497

摘要

近年来，预处理模型已经在 NLP 领域取得了巨大的成功，BERT(Devlin et al., 2018)，XLNet(Yang et al., 2019)，RoBERTa(Liu et al., 2019) 等预训练模型在 GLUE，SQuAD 等公开数据集榜单上面都取得了优异的成绩。但是这些预训练模型动辄就是成百上千万的参数，使得他们不能很好的满足用户对响应速度的要求。除此之外，数量巨大的参数，也使得这些深度网络不能在用户终端如手机等设备上面运行。在这种需求下，对预训练模型进行压缩就有很大的意义。我们提出了 KDABert，不同于以往的知识蒸馏方法，KDABert 使用对抗训练对 Teacher 中的模型进行蒸馏，得到 Stuudent 模型。这使得 Student 模型不会对 Teacher 模型的 hidden state 过拟合，一定程度上提高了模型的鲁棒性。

1　介绍

预训练模型 (PTMs) 在实践中已经被证明能够很好的从大量的语料库里面学习到语言的表征向量。ELMo(Peters et al., 2018),GPT-2(Radford et al., 2019), BERT(Devlin et al., 2018) 等 PTMs 已经在许多 NLP 任务中取得了很大的成功。例如 NLU 任务 GLUE(Wang et al., 2018),QA 任务 SQuAD(Rajpurkar et al., 2016) 等。

尽管 PTMs 在这些 NLP 领域取得了巨大成功，他们对计算资源的海量需求也导致这些模型很难在客户终端上面运行，限制了工程上的用途。例如 BERT-base 就有 12 个 Transformer Layer，每个 Transformer 有 12 个 head，hidden state 的维度是 768，总计 110M 个参数。从头开始训练需要在 4-16 个 GPU 上面跑 4 天。而这些预训练模型的参数数目也是越来越大，如目前的 Google T5(Raffel et al., 2019) 就有 11B 参数，即使在一般的 GPU 服务器上面 fine-tuning 这样的模型，也会变得很耗时间。因此对这样的 PTMs 进行模型的压缩，就变得非常有意义。

以往的知识蒸馏方法大多采用 MSE 的方式来从 Teacher 模型中学习知识 (Jiao et al., 2019; Sun et al., 2019)，但是这样做的问题在于，我们的 Student 模型容易过拟合，其泛化能力不足。原因在于，我们通过 MSE 的方法，严格的希望 Student 某个位置模型的输出等于 Teacher 模型对应位置的输出，由于 Student 模型和 Teacher 模型的结构不同，模型内部的计算有差异，很可能 $||P_s(y|x) - P_t(y|x)||_2$ 很小但是 $||P_s(y|x+\epsilon) - P_t(y|x+\epsilon)||_2$ 却很大 (P_s P_t 分别为 Student 模型和 Teacher 模型的分布函数, $||\epsilon||_2 < 1$)。

为此，我们提出了 **KDABert**: **K**nowledge **D**istillation with generative **A**dversarial networks for **Bert** Model Compression。该模型的改进主要分为使用对抗训练代替 MSE，使 Student Model 从 Teacher Model 中蒸馏知识。

参考文献

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2019. Tinybert: Distilling bert for natural language understanding.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *CoRR*, abs/1802.05365.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Siqi Sun, Yu Cheng, Zhe Gan, and Jingjing Liu. 2019. Patient knowledge distillation for bert model compression.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *CoRR*, abs/1906.08237.