Towards Effective and Efficient Language Models: RNN-Based Generative Model Enhancements, Transfer Learning, and Inference Optimization

Ting Hu

Abstract

Deep Learning in the Natural Language Processing field has witnessed a series of innovations and advancements through the transition from an era dominated by Recurrent Neural Networks (RNNs) to one dominated by Transformers. Accordingly, our research endeavors implicitly manifest the paradigm shift from a one-step training from scratch to pre-training followed by fine-tuning. Amidst the RNN-dominant epoch, we address the challenges of two mainstream generative models on text generation. Since the emergence of Pre-trained Language Models (PLMs) and their superior performance on various downstream tasks, our focus then transits to transfer learning, i.e., transferring the knowledge PLMs acquire during pre-training to downstream tasks, of which the effectiveness and efficiency are investigated. Moreover, we discuss the practical applicability of the models resulting from transfer learning, facilitating their deployment in real-world scenarios.

The first section of our research centers on text generation models. We investigate two generative models based on RNNs, Generative Adversarial Networks (GANs) and Vector Quantization-Variational AutoEncoders (VQ-VAEs). These models have several commonly observed issues: GANs exhibit training instability and mode collapse, and VQ-VAE suffers from index collapse when applied to text generation. We present corresponding methods to alleviate these issues, resulting in improved performance, even though the generation quality is inferior to the PLMs introduced subsequently.

The second part focuses on transfer learning, i.e., fine-tuning PLMs on Natural Language Generation (NLG) tasks. We investigate the effectiveness and efficiency of transfer learning. The consistency between the pre-training scheme of individual PLMs and fine-tuning is significant to the effectiveness. Regarding efficiency, Parameter-Efficient Fine-Tuning (PEFT) is a decent alternative to conventional fine-tuning in computation-restricted and data-scare scenarios. The introduced Scaled Prompt-Tuning method presents a better tradeoff between performance and computational costs than related PEFT approaches in few-shot cases with favorable generalization ability and transferability.

In the final leg of our study, we bridge the gap between fine-tuned models and their deployment. Recognizing the demands of applying the advanced models to real-world applications, we tackle the challenge of reducing computations and accelerating inference on BERT by different compression techniques. We investigate the effectiveness of quantization methods on BERT when adapted to Natural Language Understanding (NLU) tasks, including GLUE and SQuAD. Furthermore, we present Neural grafting and the dynamic inference mechanism to address the dataset-level parameter redundancy and instance-level computation redundancy, yielding reduced computations, faster inference, and boosted performance.