论文标题

LoRA: Low-Rank Adaptation of Large Language Models

发表期刊

ICLR 2022

作者

Edward Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen

发表日期

2021-10-16

阅读日期

2023.11.10

评分 Score

☐ 优秀

☐ 一般

☐ 较差

☐ 很差

类型	思路	批注
研究背景	随着预训练模型变得越来越大，重新训练所有模型参数的完全微调变得不太可行	
方法和性质	● 提出了一种称为"LoRA"的方法，通过将可训练的秩分解矩阵注入到 Transformer 架构的每一层中，冻结预训练模型权重，从而大大减少下游任务的可训练参数数量 ● 对语言模型适应中的秩缺失进行了实证研究，揭示了 LoRA 的有效性	
研究结果	与使用 Adam 进行微调的 GPT-3 175B 相比，LoRA 可以将可训练参数数量减少 10,000 倍，并将 GPU 内存需求减少 3 倍。尽管可训练参数更少、训练吞吐量更高，并且不像适配器那样具有额外推理延迟，但 LoRA 在 RoBERTa、DeBERTa、GPT-2 和 GPT-3 的模型质量上表现相当或更好	
创新点	1. 针对当预训练模型变得更大时，全部重新调整所有模型参数的完全微调变得不太可行的问题，提出了冻结预训练模型权重的方法 2. 在 Transformer 架构的每一层中，通过注入可训练的秩分解矩阵，大大减少了下游任务的可训练参数数量	
数据	◆ GLUE Benchmark ◆ WikiSQL ◆ SAMSum ◆ E2E NLG Challenge ◆ DART ◆ WebNLG	
结论	这篇论文提出了一种名为 LoRA 的低秩适应方法，用于大型语言模型的适应。该方法通过低秩矩阵逼近实现适应，从而减少全面微调的需求，并在性能和效率方面取得了显著改进	
研究展望	更高效的优化算法 更有效的模型融合方法 更有效的超参调整算法	

重要性	1. LoRA 大大减少了模型参数的数量，降低模型的复杂度，提高模型的泛化能力 2. LoRA 可以通过随机分解的方式，将矩阵分解为两个低秩矩阵，从而加快模型的收敛速度，提高模型的训练速度 3. LoRA 可以在保留重要特征的同时，去除不重要的特征，从而提高模型的性能	
想法和问题	可以理解算法，还没有具体实现	
本文好的表达摘录	■ LoRA performs on-par or better than fi netuning in model quality ■ sheds light on the efficacy of LoRA. ■ The major downside of fine-tuning is that the new model contains as many parameters as in the original model ■ posing a trade-off between efficiency and model quality ■ make frequent references to ■ follow the conventions set out by ■ we are given a pre-trained autoregressive language model PΦ(y|x) parametrized by Φ. ■ by no means new ■ orthogonal improvement	