

arXiv > cs > arXiv:2106.09685v2

Parameter-efficient Help | Adv

Computer Science > Computation and Language

[Submitted on 17 Jun 2021 (v1), last revised 16 Oct 2021 (this version, v2)]

LoRA: Low-Rank Adaptation of Large Language Models

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

An important paradigm of natural language processing consists of large-scale pre-training on general domain data and adaptation to particular tasks or domains. As we pre-train larger models, full fine-tuning, which re-trains all model parameters, becomes less feasible. Using GPT-3 175B as an example -- deploying independent instances of fine-tuned models, each with 175B parameters, is prohibitively expensive. We propose Low-Rank Adaptation, or LoRA, which freezes the pre-trained model weights and injects trainable rank decomposition matrices into each layer of the Transformer architecture, greatly reducing the number of trainable parameters for downstream tasks. Compared to GPT-3 175B fine-tuned with Adam, LoRA can reduce the number of trainable parameters by 10,000 times and the GPU memory requirement by 3 times. LoRA performs on-par or better than fine-tuning in model quality on RoBERTa, DeBERTa, GPT-2, and GPT-3, despite having fewer trainable parameters, a higher training throughput, and, unlike adapters, no additional inference latency. We also provide an empirical investigation into rank-deficiency in language model adaptation, which sheds light on the efficacy of LoRA. We release a package that facilitates the integration of LoRA with PyTorch models and provide our implementations and model checkpoints for RoBERTa, DeBERTa, and GPT-2 at [this https URL](https://github.com/microsoft/loralib).

LoRA: Adapter tuning的效果仅仅因为低秩适配吗?

 **Mady**
AIGC, LLM, 用户增长算法

关注她


62 人赞同了该文章

2021年的文章，简单有效又有理论支持的Adapter tuning的代表性方法，目前被各大开源Large Language Model(LLM)作为默认的工具，Huggingface上有打包可直接调用，然而，文章目前还未正式发表？又是arXiv上的神存~

原文：LoRA: Low-Rank Adaptation of Large Language Models

LoRA: Low-Rank Adaptation of Large Language Models

arxiv.org/abs/2106.09685



论文解读

我是阿豪啊：LoRA论文回顾

482 赞同 · 33 评论 文章

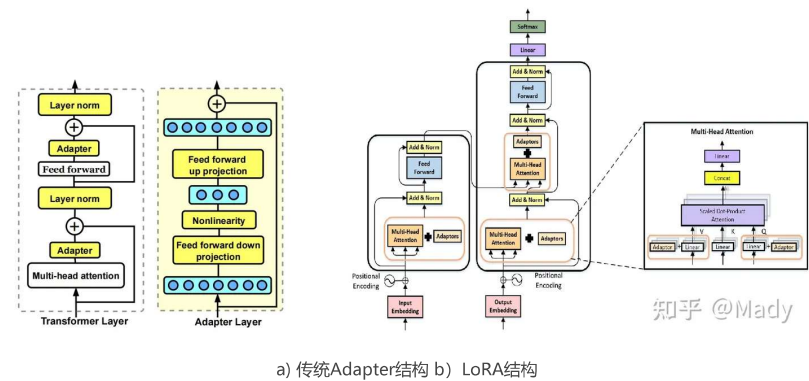


GitHub

GitHub - microsoft/LoRA: Code for loralib, an implementation of "LoRA: Low-Rank Adaptation ...

github.com/microsoft/LoRA

关于题目的解读，依旧是默认你对该方法非常熟悉，读完这段就可以结束。图a是传统的Adapter插入的位置：为保证输入和输出的维度一致，在attention模块后接入一个bottleneck结构的模块。可以理解在attention外面整体包一个复杂的非线性函数，进行空间变换，整体调整attention和feed forward的方向，达到想要的微调效果。图b是LoRA的嵌入方式：在attention的Q和V中嵌入可快速迭代的低秩矩阵，以矩阵加法的形式得到变换后的结果。这就很清晰了，以矩阵的角度来看，LoRA改变了attention中某些元素的值，改变了注意力的权重分布，而某种角度上来看，传统的Adapter并没有改变attention，只是把迁移学习做的更细节一些，给出了垂直领域适配的方法。



下面分别就以下几个部分展开：motivation，一句话摘要，模型实践经验与思考。

一切毁灭。然而LLM的随参数规模扩大而显现出的知识涌现能力，又让人垂涎，不愿放弃。捭风缉缝的时候，终于Delta-tuning让穷鬼看到了曙光。比如文章Delta-tuning在100个典型代表的NLP单模态学习中，改变0.03%-2.38%的参数量效果可以与vanilla-FT持平。VL-Adapter文章中通过实验对比，在多模态学习中，微调2%-12%的参数量就可以达到vanilla-FT的效果。

一句话摘要

在transformer的attention中插入低秩矩阵去拟合QKV中的delta-Q和delta-V，也就是用 $M \times R, R \times N$ （ R 很小）的两个矩阵去拟合 $M \times N$ 的 Q 的变化量，这个拟合在微调的过程中完成；得到低维矩阵后，再用加法恢复微调后的attention模块该有的参数。优点除了调节量少，计算和存储代价低，还有可保留原有模型参数，每个场景的调节量单独存储，方便以后根据目标进行组合。值得注意的是：保证权重矩阵的种类的数量比起增加隐藏层维度 R 更为重要，增加 R 并不一定能覆盖更有意义的子空间。

模型实践经验与思考

底座模型：ChatGLM6B，微调方式：prompt-tuning和LoRA

开源的LLM一般会自带prompt-tuning方法，不调整模型结构和参数，用额外的context包装输入。盲调几轮后发现很容易记住本轮的数据，也很容易忘记前几轮的数据。查了文章发现：该方法soft prompts的参数很难收敛，对参数规模大(千亿规模的)的预训练模型效果好。思考下原因：prompt的思想是让人去适配LLM，绞尽脑汁地想出一个能Work的命令，换不同的单词或者句子，反复尝试好的任务表达方式，也就是人工的拟合训练数据的分布。那么它对参数规模大的模型效果好的原因就很明显了，参数规模大的模型更好的刻画了复杂世界，包含的分布更多，比起小模型刻画的简易世界，用人工的方法也就更容易适配或者**击中**这其中包含的可能性了。

当然，LoRA也有自己的局限性，比如预训练任务（底座模型）和增量训练（适配场景）的分布差很多，较少的参数不足以刻画这些不同。当每一个小场景都只有很少的调整空间，分别训练后合在一起的后果就是，累积调整的参数量增多，效果又没有达到。但是LoRA有一个很明显的优势在于，它的并行的结构不会导致推理延迟，即：串行的adapter会限制FLOPs的添加，降低了硬件并行性，推理时无法保持低延迟。这也是原文中一再强调的一个点，然而在大部分的文章解读中，就像本文题目的解读一样，被忽略了。

编辑于 2023-05-19 16:28 · IP 属地北京

LoRa LLM（大型语言模型）

评论千万条，友善第一条

2 条评论

默认 最新

**Ridiculous**

LoRA不是被ICLR 2022录了吗？

05-19 · IP 属地北京

回复 2

**Mady** 作者 

感谢，我再去查一下，arxiv上追踪不到

05-19 · IP 属地北京

回复 赞

推荐阅读



混淆矩阵、准确率、精确率、召回率、真正率、假正率、...

算法爱好者



统计学习 | 矩阵正态分布 (matrix normal...

前言：在机器学习和统计学习中，正态分布的身影无处不在，最为常见的是标准正态分布和多元正态分布 (multivariate normal distribution)，两者分别作用于标量 (scalar) 和向量 (vector)。实...

Xinyu Chen



代码与艺术(0)—随机中的秩序

飞飞



模型1:正态分布 (The Model Thinker 2)

Baller

赞同 62

2 条评论

分享

喜欢

收藏

申请转载

...

https://zhuanlan.zhihu.com/p/630255810

2/2