LongLoRA详解

1. 背景
2. 技术方案
3. 总结

LongLoRA：超长上下文，大语言模型高效微调方法

1. 背景

麻省理工学院和香港中文大学联合发布了LongLoRA，这是一种全新的微调方法，可以增强大语言模型的上下文能力，而无需消耗大量算力资源。通常，想增加大语言模型的上下文处理能力，需要更多的算力支持。例如，将上下文长度从2048扩展至8192，需要多消耗16倍算力。LongLoRA在开源模型LLaMA2 7B/13B/70B上进行了试验，将上下文原始长度扩展至32K、64K、100K，所需要的算力资源却很少。

开源地址：https://github.com/dvlab-research/LongLoRA

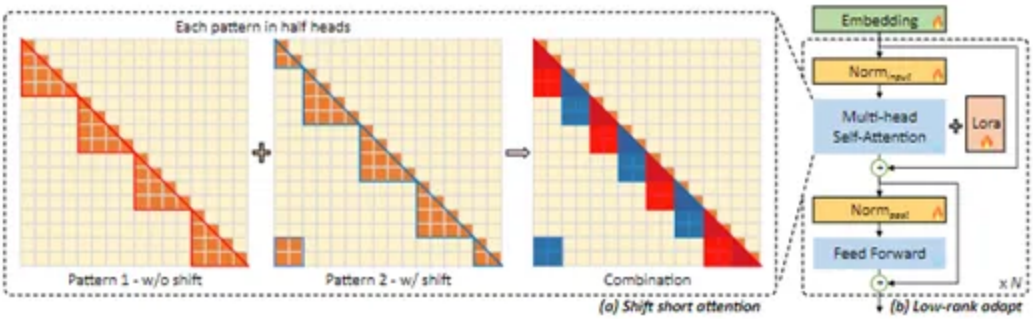
论文地址：https://arxiv.org/abs/2309.12307

2. 技术方案

LongLoRA的高效微调方法

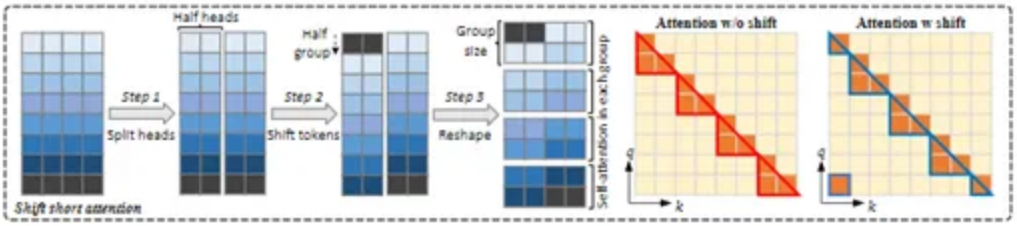
根据LongLoRA的论文介绍，采用了两大步骤完成了高效微调。第一，在训练期间使用一种更简单的注意力形式（聚焦于特定信息），开发者称之为转变短注意力（S2-Attn）。

这种新的注意力方法有助于节省大量的计算能力，而且几乎与常规的注意力方法一样有效，在训练过程中发挥了重要作用。



第二，重新挖掘了一种有效扩大上下文（用于训练的信息量）的方法。开发人员发现，一种名为LoRA的方法对此非常有效，尤其是当与可训练的嵌入和规范化一起使用时。

LongLoRA在各种任务上都显示出了优异的结果，可以与不同大小的LLMs一起使用。它可以将用于训练的数据量从4k增加到100k，对于另一个模型，可以增加到32k，所有这些都在一台强大的计算机机器上完成。此外，它与其他现有技术兼容性很强，并不会改变原始模型设计架构。



此外，为了让 LongLoRA 更加实用、高效，开发者还整理了一个名为 LongQA 的数据集，其中包含 3000 多对用于训练的问题和答案。这使得 LongLoRA 还能有效改进大语言模型的输出能力。

Table 1: Ablations on different training patterns and target context length. 'Short' means 1/4 of the target context length. 'Long' equals to the target context length. Models are fully fine-tuned upon an LLaMA2 (Touvron et al., 2023b) model in 7B size, on RedPajama (Computer, 2023) dataset. Results are tested in perplexity on PG19 (Rae et al., 2020) validation split.

Setting	Position Embedding	Training Attention	Training Shift	Target Context Length 8192	16384	32768
Train-free	PI (Chen et al., 2023)	w/o fine-tuning		15.82	94.57	236.99
	NTK-Aware (ntk, 2023)			10.89	88.44	932.85
Full Attn	PI (Chen et al., 2023)	Long	-	8.02	8.05	8.04
Short Attn		Short	✗	8.29	8.83	9.47
S^2-Attn		Short	✓	8.04	8.03	8.08

3. 总结

总体来说，LongLoRA 在大型语言模型领域提出了创新方法，在处理大量信息时，也可以更轻松、更高效地微调这些模型，而必须消耗更多的算力资源。