

L2-LoRA: Improving Low-Rank Adaptation with Layer-Specific Regularization

L2-LoRA: 通过层特异性正则化改进LoRA

The Fortieth AAAI Conference on Artificial Intelligence

(AAAI-26)

大语言模型微调面临的现实挑战

大语言模型在数学推理、代码生成等广泛任务中功能强大，但实际应用中对于特定任务需要微调。参数高效微调PEFT是全微调的友好替代



LoRA的灾难性遗忘：能提升下游任务性能，但往往导致泛化能力下降



现有的大多数基于LoRA的优化方法在所有层面上均匀地施加参数更新，而未考虑LLMs的功能层级结构：**底层偏向词句特征，而高层更偏向语义抽象**。对所有层使用相同的适配策略，可能会破坏预训练知识的组织结构，尤其会覆盖存储在底层的通用语言和事实信息。

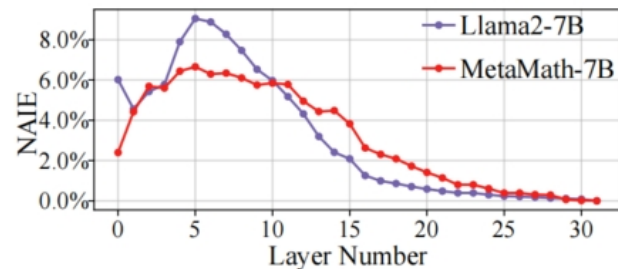
基础知识

知识定位与编辑的研究表明, Transformer 模型中的知识并非均匀分布在各层, 而是集中在特定层中。故文中采用因果追踪对基础模型和微调模型的知识存储进行了逐层分析。

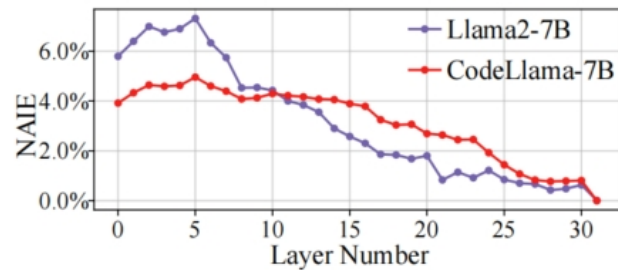
预训练的通用知识主要定位于低层, 而微调过程则将任务特定知识重新分配至高层网络。

L2-LoRA: 具备层感知能力的LoRA变体。

根据层级知识分布, L2-LoRA为LoRA权重分配不同强度的L2正则化。



(a) Math knowledge localization



(b) Code knowledge localization

因果追踪

传统因果追踪：依赖预定义的事实三元组(s,r,o)，对于一个有L层的模型和输入x，第t个token、第 ℓ 层的间接效应：

$$\text{IE}_{(t,\ell)} = p_{\theta}(o \mid x_{\text{noise}}, v_{(t,\ell)}) - p_{\theta}(o \mid x_{\text{noise}})$$

由于生成模型可能产生多个有效输出/序列，改进版的因果追踪：追踪真实输出序列中的逻辑值（logits），量化各层的因果效应。

对于给定实例 k_n ，计算每个层 ℓ 的逐标记间接效应IE，随后将所有输出标记的该值汇总，得出层级得分：

T 表示真实标记序列 $y = (y_1, y_2, \dots, y_T)$ 的长度

$$\text{IE}_{(k_n, \ell)} = \sum_{t=1}^T \left(p_{\theta}(y_t \mid x_{\text{noise}}, v_{(t,\ell)}) - p_{\theta}(y_t \mid x_{\text{noise}}) \right)$$

因果追踪

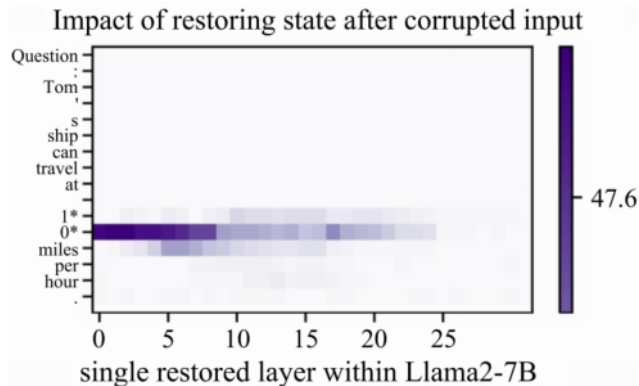
为实现特定任务层面的知识定位，对各层的间接效应进行了平均计算，包含 N 个样本的数据集 D ：

$$\text{AIE}_{(D,\ell)} = \frac{1}{N} \sum_{n=1}^N \text{IE}_{(k_n,\ell)}.$$

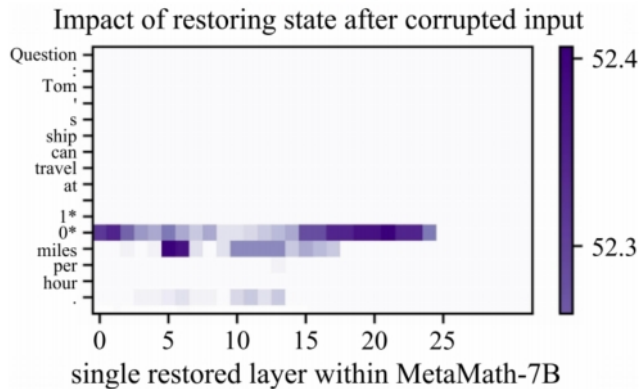
为了更好反映各层对整体知识分布的相对贡献，对所有层的平均间接效应 AIE 归一化，得到标准化 AIE，NAIE。

Tom's ship can travel at 1*0* miles per hour

GSM8K数据集在Llama2-7B和MetaMath-7B中的因果追踪结果。热图展示了在每个标记位置恢复隐藏状态所产生的逐层间接效应。*表示输入存在噪声。

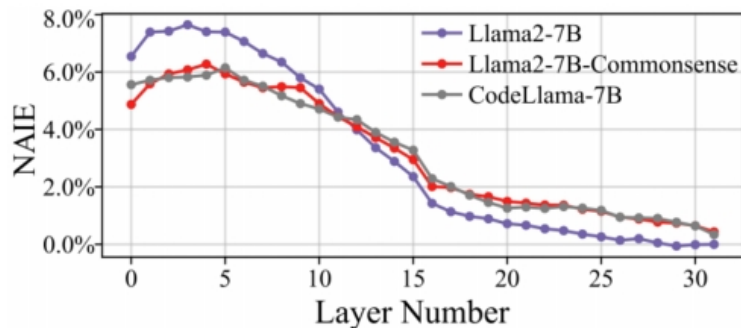


(a) Base Model

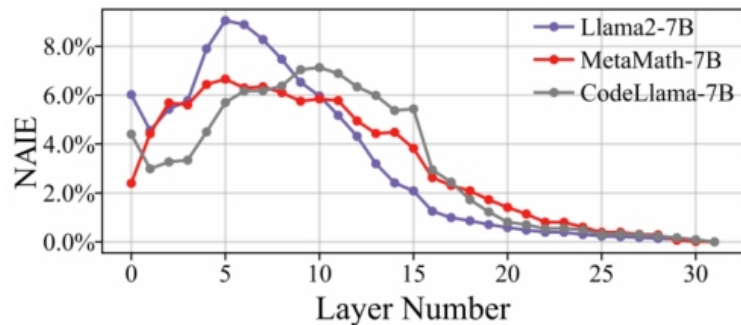


(b) Fine-Tuned Model

基于层级结构的任务特定知识定位



(a) Commonsense Reasoning



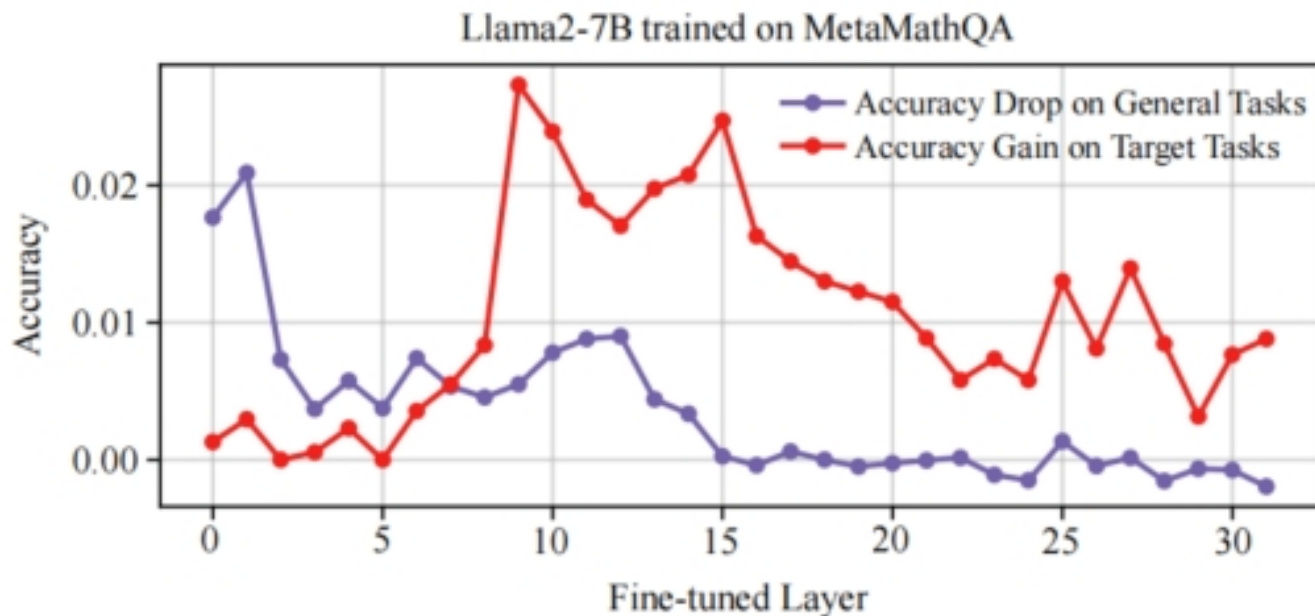
(b) Mathematical Reasoning

基础模型与微调模型中知识在各层间的分布情况。基础模型中的知识主要集中在较低层级；而经过微调后，知识分布呈现持续向上迁移的趋势——在常识推理和数学推理领域中，知识均逐渐向更高层级集中。

基于层级结构的任务特定知识定位

使用MetaMathQA数据集对Llama2-7B进行层级微调分析。每个数据点对应单个Transformer层进行的微调。

存储通用知识的较低层网络 vs 存储任务特定知识的更深层网络



对知识密集型层进行微调不仅效果更佳，且对模型整体影响更小。

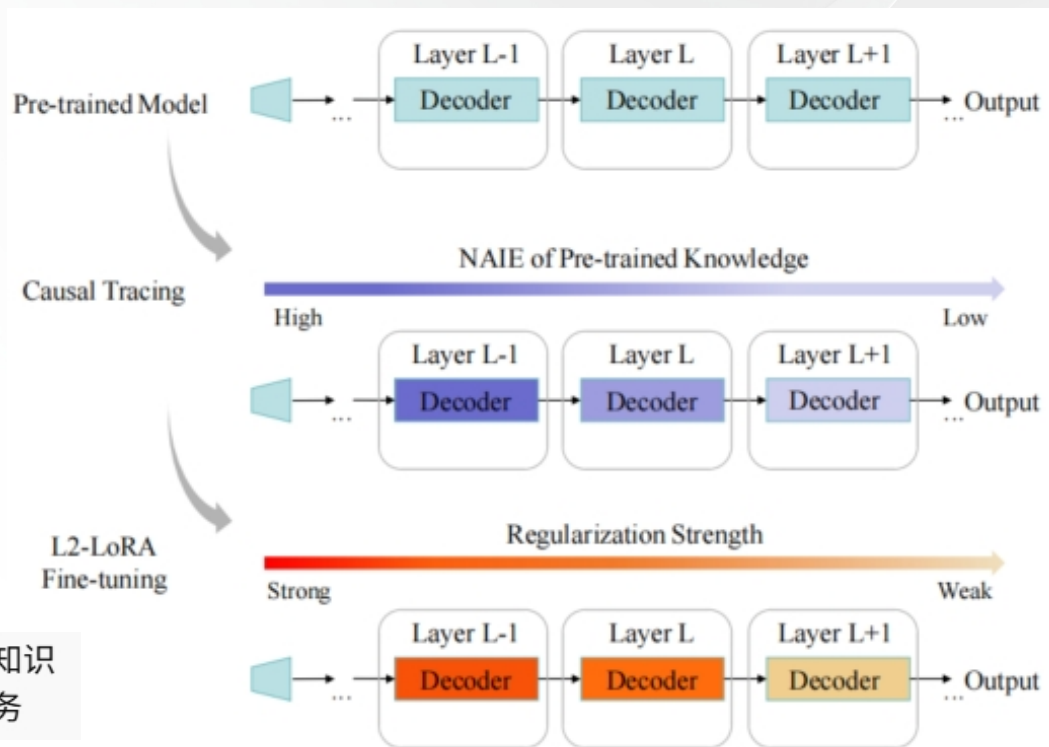
L2-LoRA

正则化的强度由各层的知识分布决定，该分布通过归一化平均间接效应（NAIE）进行衡量。

因果追踪（通用知识）



逐层知识分布(NAIE)



低层：强 L2 正则 → 少更新 → 保护旧知识
高层：弱 L2 正则 → 多更新 → 学新任务

L2-LoRA

L2正则化通过惩罚较大的权重来约束模型复杂度并缓解过拟合问题。采用L2正则化的损失函数可表述为：

$$\mathcal{L}(\mathbf{y}, \hat{\mathbf{y}}) = \mathcal{L}_0(\mathbf{y}, \hat{\mathbf{y}}) + \lambda \sum_{i=1}^n \theta_i^2$$

针对各Transformer层的 L2 正则化变体，为每个层 ℓ 分配独立的正则化强度 λ_ℓ ，从而根据各层的知识分布实现对参数更新的选择性抑制， W_ℓ 表示第 ℓ 层中可训练参数的集合。

$$\mathcal{L}(\mathbf{y}, \hat{\mathbf{y}}) = \mathcal{L}_0(f(x), \mathbf{y}) + \sum_{\ell=0}^{L-1} \lambda_\ell \|W_\ell\|_2^2$$

L2-LoRA

LoRA微调的有效权重为：（致力于设计低秩结构）

$$W = W_0 + \Delta W = W_0 + BA$$

L2-LoRA侧重于控制 ΔW 的幅度以保留通用知识。为平衡适应性与知识保留，引入基于各层在存储通用知识中的重要性--NAIE (D, ℓ) --加权的层特定 L2 正则化项。最终目标函数为：

$$\mathcal{L}(\mathbf{y}, \hat{\mathbf{y}}) = \boxed{\mathcal{L}_0(f(x), \mathbf{y})} + \eta \sum_{\ell=0}^{L-1} \boxed{\text{NAIE}_{(D, \ell)}} \boxed{\|\Delta W_{\ell}\|_2^2}$$

微调损失，学会下游任务

第 ℓ 层保存通用知识的重要程度
高---对原有知识很重要

第 ℓ 层BA/更新量的平方L2范数
大---改动多

L2-LoRA

NAIE (D, ℓ) 通过因果追踪获取，会在微调前引入额外的计算开销。为缓解此问题，我们基于观测到的定位趋势提出了经验性近似方法：采用三次衰减方案，对较低层级施加更强的正则化约束。

$$\text{NAIE}_{(D, \ell)} = \begin{cases} 1 & 0 \leq \ell < \ell_i \\ \tau + (1 - \tau) \left(1 - \frac{\ell - \ell_i}{\ell_f - \ell_i}\right)^3 & \ell_i \leq \ell \leq \ell_f \\ \tau & \ell > \ell_f \end{cases}$$

其中， ℓ_i 和 ℓ_f 是界定正则化强度递减范围的拐点， τ 为最小正则权重。在实验中，我们设定 $\ell_i = L/3$ 、 $\ell_f = 2L/3$ ，该参数组合取得了优异的实验性能。

实验对比--通用知识的保留

单任务表现
Commonsense
170K

Model	MMLU					ARC-c
	Humanities	Social Sciences	STEM	Other	AVG	
Llama-7B	29.9	29.4	26.3	33.4	29.8	41.7
LoRA	24.4	21.9	21.5	24.0	23.1	61.3
L2-LoRA	28.4	27.9	28.2	28.5	28.3	65.6

Stage	Model	MMLU				
		Humanities	Social Sciences	STEM	Other	AVG
	Llama-7B	29.9	29.4	26.3	33.4	29.8
1	LoRA	25.9	25.6	26.1	24.9	25.7
1	L2-LoRA	28.4	27.9	28.2	28.5	28.3
2	LoRA	23.7	22.2	21.3	24.9	23.1
2	L2-LoRA	22.2	22.8	22.2	25.4	23.8
3	LoRA	22.2	22.8	22.2	25.4	23.3
3	L2-LoRA	24.4	22.8	21.2	25.3	23.6

持续学习
1--常识推理
--Commonsense
170K
2--数学推理
--MetaMathQ
3--指令遵循--Alpaca

实验对比--下游任务的适应

常识推理
Llama2-7B
Commonsense 170K

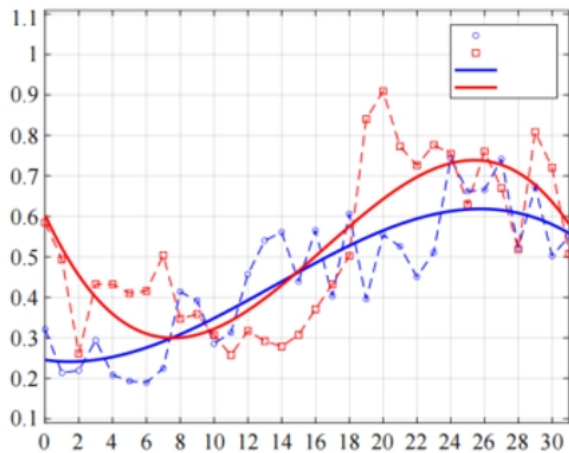
PEFT	Params (%)	BoolQ	PIQA	SIQA	HellaS.	WinoG.	ARC-e	ARC-c	OBQA
LoRA	0.83%	69.8	79.9	79.5	83.6	82.6	79.8	64.7	81.0
DoRA	0.84%	71.8	83.7	76.0	89.1	82.6	83.7	68.2	82.4
PiSSA	0.83%	67.6	78.1	78.4	76.6	78.0	75.8	60.2	75.6
MiLoRA	0.83%	67.6	83.8	80.1	88.2	82.0	82.8	68.8	80.6
L2-LoRA	0.83%	68.9	85.2	79.8	87.5	82.8	83.1	69.5	82.5

Method	GSM8K	MATH	Avg.
Full FT	66.50	19.80	43.15
LoRA	60.58	16.88	38.73
DoRA	62.28	17.63	40.00
PiSSA	58.23	15.84	37.04
MiLoRA	63.53	17.76	40.65
L2-LoRA	65.32	18.12	41.72

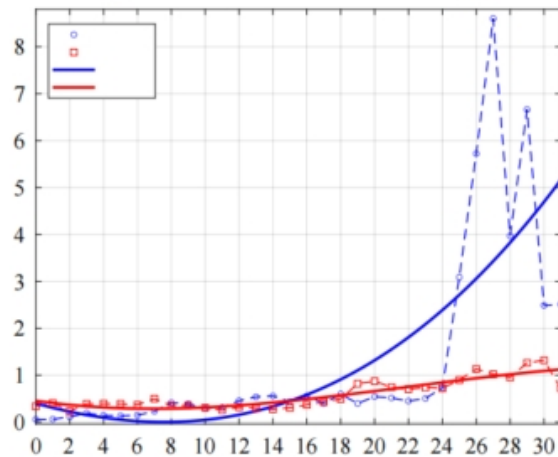
数学推理
Llama2-7B

消融实验

LoRA 会不会自己学会“低层少更新、高层多更新”？



(a) Vanilla LoRA Fine-tuning



(b) L2-LoRA Fine-tuning

在MetaMathQA数据集上对Llama2-7B模型做LoRA微调，使用L2范数量化权重变化。蓝色和红色曲线分别表示查询矩阵和值投影矩阵权重变化幅度的二次多项式拟合结果。

L2-LoRA能有效约束低层网络的变化，同时为高层网络保留更多灵活性。

消融实验

NAIE (D, ℓ) 评分的有效性

通过 NAIE (D, ℓ) 评分我们发现模型的低层主要存储通用知识，而高层则主要存储任务特定知识。在数学推理任务上进行了实验：首先冻结所有层，

Starting	Evaluation	Number of Trainable Layers			
		0	2	4	6
Layer 0	MMLU	43.95	40.37	41.45	41.71
	AddSub	3.29	3.29	4.05	11.14

从低层开始逐步放开训练

Starting	Evaluation	Number of Trainable Layers			
		0	2	4	6
Layer 31	MMLU	43.95	43.85	43.68	43.88
	AddSub	3.29	13.42	14.68	17.22

从高层开始逐步放开训练

THANKS

