DoRA：提升参数高效微调性能的创新方法

摘要

随着深度学习和大规模预训练模型的快速发展，参数高效微调（PEFT）技术应运而生，成为解决微调大模型时计算资源开销问题的有效手段。LoRA（Low-Rank Adaptation）作为 PEFT 中的一种经典方法，利用低秩分解减少了微调过程中的计算和内存开销。然而，LoRA 在处理复杂任务时表现出一定的局限性，尤其是在权重调整的幅度和方向方面。为了解决这一问题，本研究使用了一种改进的 PEFT 方法——DoRA（Weight-Decomposed Low-Rank Adaptation）。DoRA 通过将预训练模型的权重分解为幅度和方向两个独立部分，分别进行优化，从而增强了微调的学习能力和任务适应性。实验表明，DoRA 在多个自然语言处理和多模态任务中性能优于 LoRA，且无需增加额外的推理开销。本研究的贡献在于提出了一种新的权重分解策略，为大规模预训练模型的高效微调提供了更精确且具有较低计算成本的解决方案。

关键词：参数高效微调；DoRA；大模型微调；常识性推理；

1 引言

随着深度学习的快速发展，预训练模型（Pre-trained Model, PTM）在自然语言处理（NLP）[11, 15] 和多模态任务 [8, 9] 中展现出卓越的通用性和迁移能力。这些模型通过在大规模数据集上进行预训练，显著提升了多种下游任务的性能。然而，为了充分利用预训练模型的能力，将其适配具体任务的微调（Fine-Tuning, FT）成为研究的重点。尽管全量微调方法在性能上通常优于其他方法，但其计算资源需求随模型规模指数增长，使得其在实际应用中的成本和可行性受到极大限制。

因此，上下文学习 [12] 成为当前研究的热点，它是将下游任务训练数据传递给大语言模型的一种方法。然而，Transformer 架构所限制的上下文长度 [16]，小语言模型中在少样本学习 (ICL) 上能力不足 [5] 以及随着上下文长度 (或 ICL 示例数量) 增加而呈现巨量的计算成本 [5]，在一定程度上影响了其实用性、可靠性、和 ICL 的效率。尽管在某些场景下，模型在 ICL 设置中的表现与微调模型相当，甚至更优，但由于 ICL 推断的高成本，微调仍然被视为一种更具性价比的策略 [1]。

为了解决全量微调的高资源开销问题，参数高效微调（Parameter-Efficient Fine-Tuning, PEFT）技术 [3] 应运而生。PEFT 方法通过仅调整模型中的少量参数，从而实现对下游任务的高效适配。这些参数可以是现有模型参数的子集，也可以是新添加的参数。这类方法在参数和内存效率、训练速度、最终模型质量以及可能带来的额外推理成本等方面存在差异。LoRA（Low-Rank Adaptation） [4] 作为一种经典的 PEFT 方法，通过引入低秩矩阵分解技术避免

了额外的推理开销，因而备受关注。但是，LoRA 和 FT 之间仍然存在性能上的差距，这通常归因于可训练参数数量有限，而没有进一步探索其他潜在原因 [4, 6]

现有研究表明，LoRA 与全量微调在性能上存在显著差距，这一差距主要源于其对权重调整模式的局限性。特别是，LoRA 在权重幅度和方向的更新上缺乏灵活性，无法充分捕捉复杂任务所需的学习能力。为了解决上述问题，该研究提出了 DoRA（Weight-Decomposed Low-Rank Adaptation）[10]，一种基于权重分解的改进型 PEFT 方法。DoRA 通过将预训练权重分解为幅度（Magnitude）和方向（Direction）两个独立的组件，分别对其进行优化，从而显著提升微调的学习能力。具体而言，DoRA 在方向调整中采用 LoRA 的低秩分解技术以减少参数量，同时引入对幅度的独立调整，以更接近全量微调的学习模式。实验表明，DoRA 在多个自然语言处理、多模态任务中性能优于 LoRA，并且不增加任何推理开销。

本研究的选题具有以下重要意义：

• 理论价值：通过权重分解的创新性分析方法，揭示了全量微调与 PEFT 方法在学习模式上的本质差异，为 PEFT 技术的进一步优化提供了理论支持。

• 实践价值：DoRA 能够在多种任务中实现优异的性能提升，为计算资源受限的场景提供了高效的解决方案。

• 技术推广：DoRA 兼具 LoRA 的高效性与全量微调的学习能力，为大规模模型在语言、视觉和多模态领域的广泛应用开辟了新的路径。通过本研究的探索，旨在进一步缩小 PEFT 方法与全量微调在性能上的差距，为更广泛的实际应用提供支持。

2 相关工作

2.1 Parameter-Efficient Fine-Tuning (PEFT)

微调大模型的效率问题是当前自然语言处理和深度学习领域的研究热点之一。传统的微调方法通常需要调整模型的所有参数，这在处理拥有数百万乃至数十亿参数的大规模预训练模型时，计算和内存开销巨大，导致效率较低。为了解决这一问题，近年来，参数高效微调（Parameter-Efficient Fine-Tuning, PEFT）技术逐渐成为解决这一难题的有效手段。PEFT 方法通过限制微调过程中需要更新的参数数量，显著降低了计算和内存的开销，同时保持了良好的任务适应性。这些技术方法根据实现方式的不同，通常可以分为三类：基于适配器（Adapter）的方法、基于提示（Prompt）的方法，以及基于低秩分解（Low-Rank Decomposition）的方法

基于 Adapter 的方法是 PEFT 技术中较早提出的一类方法，最早在 2019 年被提出 [3]。这种方法的核心思想是通过在预训练模型中引入额外的可训练模块（即适配器）来实现对下游任务的适配，同时冻结大部分原始模型的参数，避免了全量微调带来的高计算开销。在微调过程中，只有这些新增的适配器模块会进行参数更新，从而降低了训练的计算成本。不同的 Adapter 方法主要在于它们引入适配器的方式和位置。常见的方式包括在现有模型的每一层中加入一个额外的线性模块 [3]，以及通过并行集成多个模块来增强适配能力 [2]，但是由于基于 Adapter 方法通常会改变模型的架构，从而导致推理延迟的增加，影响模型整体的推理能力。

基于 Prompt 的方法则是通过在输入端添加可训练的软提示（soft prompts）来进行微调 [7, 13, 17]。这些方法的核心思想是，在输入数据上附加一些额外的可训练参数，这些“提示”能够引导预训练模型更好地完成下游任务。与传统的微调方法不同，基于 Prompt 的方

法不修改原有模型的权重，而是专注于微调这些可训练的软提示。尽管这些方法在许多任务中取得了显著的效果，但它们通常面临一些挑战，尤其是对于提示初始化的高度敏感性，这可能导致模型性能的波动。因此，如何有效地初始化和优化软提示，依然是这一方向的一个重要研究课题。此外，基于 Prompt 的方法通常会增加推理时的输入长度，从而导致推理延迟的增加，这在实时应用中可能成为瓶颈。

基于低秩分解的方法通过对权重变化进行低秩分解实现参数高效微调，其中 LoRA（Low-Rank Adaptation） [4] 是最具代表性的方法，它在微调过程中通过使用低秩矩阵来近似表示权重变化，从而避免了对原有权重进行大规模的调整。LoRA 方法的核心思想是，仅在权重的低秩部分进行更新，而保持高秩部分不变，从而减少了训练过程中的参数更新量和计算开销。与基于适配器和提示的方法不同，低秩分解方法能够在推理时将微调后的低秩权重合并回原有的预训练权重，从而避免了对推理过程的额外开销。为了进一步提高更新效率，一些方法采用 SVD 分解并修剪不太重要的奇异值，从而进一步的减少参数的更新量 [18]，另外，一些研究还使用权重绑定来进一步减少可训练参数 [14]，以便在保证高效微调的同时，进一步降低模型的计算和内存负担。

2.2 LoRA（Low-Rank Adaptation）

LoRA（Low-Rank Adaptation） [4] 作为一种经典的低秩分解方法，通过在微调过程中对预训练权重的更新矩阵进行低秩分解，显著减少了微调所需的参数量。其核心思想是，通过将预训练模型的权重矩阵表示为两个低秩矩阵的乘积，从而减少需要微调的参数量，同时保持模型的表示能力。具体来说，给定预训练权重矩阵 $W_0 \in \mathbb{R}^{d\times k}$，LoRA 将其调整表示为：

$$W' = W_0 + BA$$

其中，$B \in \mathbb{R}^{d\times r}$ 和 $A \in \mathbb{R}^{r\times k}$ 是两个低秩矩阵，且 $r \ll \min(d, k)$，表示低秩分解中的秩 r 远小于预训练权重的行数 d 和列数 k。通过这种方式，LoRA 能够有效地减少微调过程中所需的参数量，而仍然能够捕捉到任务特定的特征。

在微调过程中，LoRA 仅训练这两个低秩矩阵 B 和 A，而保持原始预训练权重 W_0 固定，从而大大降低了计算成本和内存开销。相较于传统的全量微调，LoRA 不仅在计算资源上更加高效，而且能够避免推理阶段的额外计算开销。在推理阶段，微调后的权重 W' 与原始的预训练权重 W_0 合并，模型可以继续使用预训练权重进行推理，而无需对额外的参数进行推理时的计算。

LoRA 因其无推理延迟、易于实现的特性，得到了广泛应用，尤其是在语言模型（如 LLaMA 系列）和多模态模型的微调任务中。通过仅微调少量的低秩矩阵，LoRA 方法能够高效地完成多种下游任务的适应，成为当前大规模模型微调的一个重要技术。然而，尽管 LoRA 在许多任务中表现出了良好的效果，但现有研究表明，LoRA 的学习能力仍然受限于低秩矩阵的表示能力。由于低秩分解的限制，LoRA 的微调结果可能无法完全逼近全量微调的性能，尤其是在需要捕捉复杂模式和高维特征的任务中。

为了克服 LoRA 的这些不足，本研究提出了一个改进方法——Weight-Decomposed Low-Rank Adaptation（DoRA）。DoRA 通过引入幅度和方向的权重分解策略，进一步提升了微调的学习能力，并且在保证参数高效性的同时，增强了微调模型对复杂任务的适应性。具体而言，DoRA 不仅使用低秩矩阵来近似表示权重的变化，还通过对权重进行幅度和方向上的分

解，使得模型能够在低秩空间中更有效地表示任务特定的细节，从而提高微调的表现。这种方法既能充分利用低秩分解的计算优势，又能克服 LoRA 在表示能力上的限制，提供了一种新的解决方案来平衡微调的计算效率与任务表现。

通过这一改进，DoRA 在保持较小参数量和内存开销的同时，能够更好地捕捉复杂任务中的细节，从而在多个应用场景中提供更强的学习能力和更优的性能。我们相信，DoRA 方法能够为未来的大规模模型微调提供一种更高效、更精确的解决方案。

3 本文方法

研究人员发现，LoRA 方法虽然能够有效调整模型的幅度和方向，但在一些复杂任务中，它的表示能力仍然有限，尤其是在需要对权重进行微妙方向变化时，LoRA 显得力不从心。具体来说，LoRA 虽然可以按比例增加或减少幅度和方向的更新，但缺乏像传统微调方法那样，只在方向上做细微调整的能力。为了弥补这一不足，研究者提出了将幅度和方向部分解耦的策略。基于此，DoRA（Weight-Decomposed Low-Rank Adaptation）通过分离和独立微调预训练权重的“幅度”和“方向”部分，进一步增强了微调的精确度和表达能力。如图 1所示。DoRA 的核心思想是在 LoRA 的基础上，首先将预训练的权重矩阵分解为“幅度”和“方向”两个部分。在微调过程中，DoRA 分别对这两部分进行独立优化。具体地，幅度部分负责对权重的大小进行调节，而方向部分则主要调整权重的变化方向。这种方法能够提供更精细的控制，从而让模型在学习过程中更加专注于对重要方向的微调，同时减少不必要的调整。由于方向分量在权重更新中占据较大的参数比例，DoRA 进一步对方向部分进行低秩分解，以提高计算效率。具体来说，DoRA 利用 LoRA 对方向部分进行进一步的分解，将方向部分的更新表示为低秩矩阵的形式。这种做法不仅保留了 LoRA 的高效性，还进一步降低了参数更新的复杂度，使得微调过程更加高效，同时避免了对推理过程的额外负担。DoRA 的权重更新公式如下所示：

$$W' = \underline{m}\frac{V + \Delta V}{||V + \Delta V||_c} = \underline{m}\frac{W_0 + \underline{BA}}{||W_0 + \underline{BA}||_c} \tag{1}$$

其中，V 是预训练的权重矩阵，ΔV 是通过两个低秩矩阵 B 和 A 相乘得到的增量方向更新，表示对方向部分的调整。下划线表示可训练的参数，$|\cdot|_c$ 是对方向部分进行归一化处理的操作，确保了更新后的方向不改变原始权重的整体尺度。矩阵 $B \in \mathbb{R}^{d\times r}$ 和 $A \in \mathbb{R}^{r\times k}$ 则通过 LoRA 策略初始化，保证微调开始时 W' 与原始的预训练权重 W_0 一致。

此外，DoRA 在推理阶段能够将微调后的权重 W' 与预训练权重 W_0 合并，从而避免引入额外的推理延迟。这一特性使得 DoRA 不仅在训练过程中保持高效，而且在实际部署时也能无缝集成到现有的推理架构中，避免了大规模微调模型所带来的额外计算开销。

通过将权重的幅度和方向分解并独立优化，DoRA 不仅提升了微调过程中的学习能力，而且有效减少了需要训练的参数量。在复杂任务中，DoRA 的这种分解策略使得模型能够更加灵活地调整权重，以适应各种下游任务的需求，进一步拓展了 LoRA 方法的应用范围。

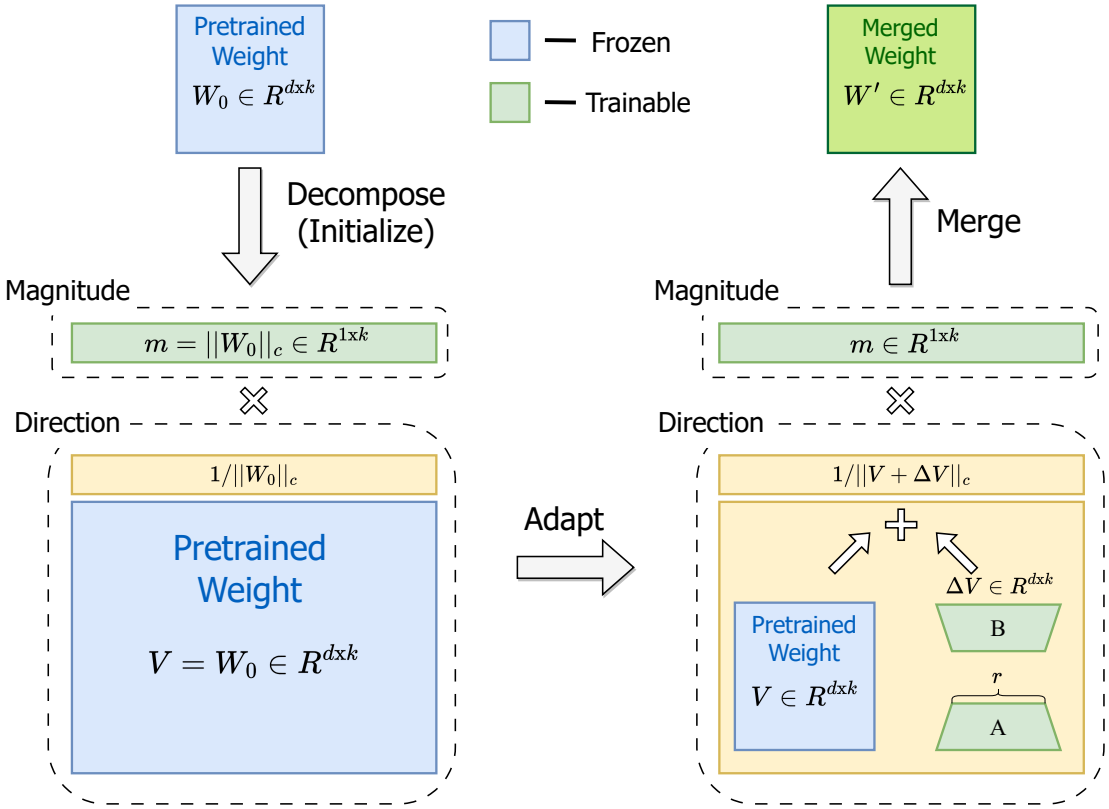

图 1. DoRA

4 复现细节

4.1 与已有开源代码对比

在本次研究中，致力于复现文章的工作，在通过参考了本论文的开源代码中对数据的处理部分作为我们复现的基础，链接如下 https://github.com/NVlabs/DoRA/tree/main。本研究完成了 DoRA 的实验以及对于模型回答提取答案的能力提升，复现了文中常识性推理部分的实验，并取得比原文效果更佳的结果。

4.2 实验环境搭建

我们对常识性推理任务进行了广泛的实验，以证明 DoRA 的优越性。所有实验均在单张 RTX 4090 上即可完成，所有的超参数均以原论文设置相同。

4.3 常识性推理数据集

本次研究选择八项多项选择任务进行评估，用以验证模型常识性推理能力，具体数据集说明如下

OpenBookQA 是一个基于小学科学知识的小型开放领域问答数据集，要求模型结合背景知识和常识推理能力解决多选问题

ARC-Challenge 是一个针对科学推理设计的高难度问答数据集，需要模型结合复杂逻辑、多步推理和科学知识解决严谨的多选问题

ARC-Easy 是一个聚焦于小学到高中科学知识的多选问答数据集,难度低于 ARC-Challenge

Winogrande 是一个用来评估常识推理能力的问答数据集，专注于通过语境推理解决模糊代词指代问题

HellaSwag 是一个针对常识推理的多选填空数据集，要求模型理解上下文并完成基于常识的多步推理。

Social-I-QA 是一个社交推理问答数据集，旨在评估模型在理解社交互动和情境中隐含含义的能力

PIQA 是一个常识推理问答数据集，专注于评估模型通过常识判断两个选项中哪个更符合实际情况

BoolQ 是一个用于二分类问题的问答数据集，旨在评估模型在理解自然语言问题并回答“是”或“否”时的推理能力

5 实验结果分析

针对论文中使用的 LLaMa2-7B 模型在 r 为 32 的 DoRA 上所得到的整体评估结果为 79.7，r 为 16 的 DoRA 所得的整体评估结果为 80.5。我们复现 r 为 32 的 DoRA 上所得到的整体评估结果为 80.2，r 为 16 的 DoRA 所得的整体评估结果为 81.4。性能的提升主要可能还是改写了对于模型回答提取答案的函数，能够更准确的提取模型的答案。实验结果 DoRA 整体比 LoRA 表现出更优异的效果，具体结果如图 2 所示

Model	PEFT Method	# Params (%)	BoolQ	PIQA	SIQA	HellaSwag	WinoGrande	ARC-e	ARC-c	OBQA	Avg.
LLaMA2-7B	LoRA	0.83	69.8	79.9	79.5	83.6	82.6	79.8	64.7	81.0	77.6
	DoRA† (Ours)	0.43	72.0	83.1	79.9	89.1	83.0	84.5	71.0	81.2	80.5
	DoRA (Ours)	0.84	71.8	83.7	76.0	89.1	82.6	83.7	68.2	82.4	79.7
LLaMA2-7B	DoRA (复现)	0.84	71.9	83.3	78.5	88.9	82.2	86.6	71.1	79.4	80.2
LLaMA2-7B	DoRA +(复现)	0.43	71.9	83.5	81.4	89.0	86.2	86.7	72.1	81.0	81.4

图 2. 实验结果示意

6 总结与展望

尽管取得了良好的结果，但是这篇论文还存在一些不足之处。由于时间成本以及整体实验对于硬件的需求，未能复现该论文中其他实验，无法充分体现出 DoRA 在各个领域展现出的优势。在未来的研究中，可能需要进一步考虑 DoRA 如何结合其他 LoRA 变体展现出更好的效果，进一步的在大模型微调领域学习

参考文献

[1] Amanda Bertsch, Maor Ivgi, Uri Alon, Jonathan Berant, Matthew R Gormley, and Graham Neubig. In-context learning with long-context models: An in-depth exploration. *arXiv preprint arXiv:2405.00200*, 2024.

[2] Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. Towards a unified view of parameter-efficient transfer learning. 2021.

[3] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. pages 2790–2799, 2019.

[4] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[5] Feyza Duman Keles, Pruthuvi Mahesakya Wijewardena, and Chinmay Hegde. On the computational complexity of self-attention. pages 597–619, 2023.

[6] Dawid J Kopiczko, Tijmen Blankevoort, and Yuki M Asano. Vera: Vector-based random matrix adaptation. *arXiv preprint arXiv:2310.11454*, 2023.

[7] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. pages 3045–3059, 2021.

[8] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. pages 12888–12900, 2022.

[9] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. 2023.

[10] Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. Dora: Weight-decomposed low-rank adaptation. *arXiv preprint arXiv:2402.09353*, 2024.

[11] Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. Is chatgpt a general-purpose natural language processing task solver? pages 1339–1384, 2023.

[12] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

[13] Anastasiia Razdaibiedina, Yuning Mao, Madian Khabsa, Mike Lewis, Rui Hou, Jimmy Ba, and Amjad Almahairi. Residual prompt tuning: improving prompt tuning with residual reparameterization. pages 6740–6757, 2023.

[14] Adithya Renduchintala, Tugrul Konuk, and Oleksii Kuchaiev. Tied-lora: Enhacing parameter efficiency of lora with weight tying. *arXiv preprint arXiv:2311.09578*, 2023.

[15] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. 2023.

[16] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. pages 5998–6008, 2017.

[17] Yaqing Wang, Jialin Wu, Tanmaya Dabral, Jiageng Zhang, Geoff Brown, Chun-Ta Lu, Frederick Liu, Yi Liang, Bo Pang, Michael Bendersky, et al. Non-intrusive adaptation: Input-centric parameter-efficient fine-tuning for versatile multimodal modeling. *arXiv preprint arXiv:2310.12100*, 2023.

[18] Qingru Zhang, Minshuo Chen, Alexander Bukharin, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. Adaptive budget allocation for parameter-efficient fine-tuning. 2023.