DoRA：权重分解低秩适配

Shih-Yang Liu [1 2] **Chien-Yi Wang** [1] **Hongxu Yin** [1] **Pavlo Molchanov** [1] **Yu-Chiang Frank Wang** [1] **Kwang-Ting Cheng** [2] **Min-Hung Chen** [1]

摘要

在广泛使用的参数高效微调（PEFT）方法中，LoRA 及其变体因避免了额外的推理成本而广受欢迎。然而，这些方法与全参数微调（FT）之间仍常存在精度差距。在本研究中，我们首先引入一种新颖的权重分解分析方法，以探究 FT 与 LoRA 之间的内在差异。基于这些发现，为接近 FT 的学习能力，我们提出了权重**分**解低秩适**配**（**DoRA**）。DoRA 将预训练权重分解为两个分量：幅值和方向，用于微调，并特别使用 LoRA 进行方向更新，以高效地最小化可训练参数数量。通过使用 DoRA，我们在不增加任何推理开销的情况下，增强了 LoRA 的学习能力和训练稳定性。DoRA 在各种下游任务（如常识推理、视觉指令微调以及图像/视频-文本理解）上对 LLaMA、LLaVA 和 VL-BART 进行微调时，consistently 优于 LoRA。代码见 https://github.com/NVlabs/DoRA。

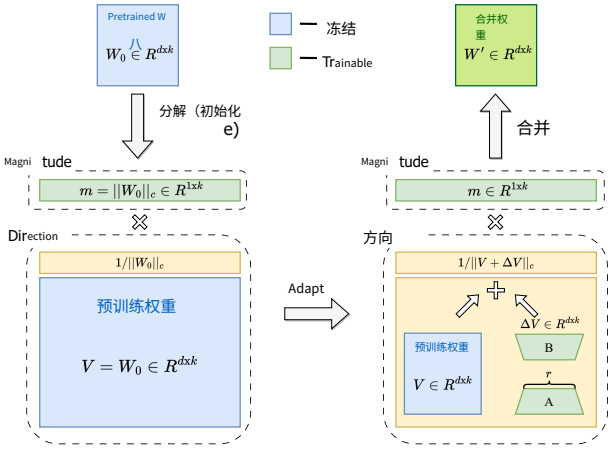


图 *1.*我们提出的 DoRA 概述，该方法将预训练权重分解为幅度 和 方向 分量以进行微调，特别是使用 LoRA 来高效更新方向分量。注意，$||\cdot||_c$ 表示矩阵沿每一列向量的向量范数。

1. 引言

通过大规模通用领域数据集进行预训练的模型已展现出显著的泛化能力，极大地推动了多种应用的发展，从自

2023 ）到多模态任务（ Li etal.， 2022 ； Liu et al.，2023a）。为了将这些通用模型适配于特定的下游任务，全参数微调（*FT*）通常被采用，即重新训练模型的所有参数。然而，随着模型和数据集规模的不断扩大，对整个模型进行微调的成本变得极其高昂。为了解决这一问题，参数高效微调（PEFT）方法（ Houlsby等，2019）被提出，以仅使用极少数量的参数对预训练模型进行微调。

其中，LoRA在自然语言处理（NLP）任务面已变得尤为流行。然而，LoRA与全参数微调之间仍存在性能差距，这通常归因于可训练参数数量有限，而未进一步探索其他潜在原因。基于权重归一化（Salimans& Kingma，2016），该方法通过权重重参数化改善梯度

的条件以实现更快的收敛，我们提出了一种新颖的权重分解分析方法，该方法首先将模型权重重参数化为幅度和方向分量，随后考察在LoRA和全参数微调引入的幅度与方向变化。

我们的分析表明，LoRA和全参数微调表现出显著不同的更新模式，这使我们推测这些差异反映了每种方法的学习能力。受此启发，我们提出了权重-分解低秩适应（**D**ecomposed **L**ow-**R**ank **Adaptation**，DoRA）

方法首先将预训练权重分解为其幅度和方向分量，然后对两者进行微调。鉴于方向分量在参数数量上规模较大，我们利用LoRA进行方向适应以实现高效的微调，如图1所示。此外，通过在经验上和数学上均展现出与全参数微调相似的学习行为，表明其学习能力与全参数微调非

常接近，我们在多种任务（从自然语言处理到视觉-语言）以及多种骨干模型（包括大语言模型和大视觉语言模型）上验证了DoRA。实验结果表明，DoRA在不牺牲推理效率的前提下始终优于LoRA，例如在常识推理任务上（**+3.7/+1.0** 在LLaMA-7B/13B上，**+2.9** 在LLaMA2-7B上，以及 **+4.4** 在LLaMA3-8B上），视觉指令微调（**+0.6** 在LLaVA-7B上），以及图像/视频-文本理解（**+0.9/+1.9** 在VL-BART上）。

我们贡献的总结如下：

- 我们提出DoRA，一种新颖的PEFT方法，它结合了权重分解，在学习能力上接近全参数微调，且相比LoRA在推理时没有任何额外延迟。
- 我们引入了一种新颖的权重分解分析方法，以揭示全参数微调与不同PEFT方法在学习模式上的根本差异。
- DoRA在各种任务上始终优于LoRA，涵盖从自然语言处理到视觉-语言基准，并适用于包括大语言模型和大视觉语言模型在内的多种骨干网络。

2. 相关工作

参数高效微调（PEFT）方法旨在降低大规模模型微调的高昂成本。它们通过仅训练相对于总参数量而言较小的参数子集，以适应下游任务。现有的PEFT方法可分为三类。第一类称为基于适配器的方法，这类方法通过在原始冻结主干网络中引入额外的可训练模块来实现，例如（Houlsby等，2019；He等，2021；Karimi Mahabadi等，2021；mahabadi等，2021）。例如，（Houlsby等，2019）提出将线性模块按顺序添加到现有层中，而（He等，2021）则主张将这些模块与原始层并联集成以提升性能。第二类是基于提示的方法。这些方法在初始输入中添加额外的软标记（提示），并仅专注于微调这些可训练向量，如（Lester等，2021；Razdaibiedina等，2023；Wang等，2023）等研究中所示。然而，这些方法通常由于对初始化敏感而面临挑战，影响其整体有效性。前两类方法，无论是改变模型的输入还是架构，都会导致与基线模型相比，推理延迟增加。

LoRA（Hu等，2022）及其变体属于PEFT的第三类方法，其显著特点是不增加任何推理负担。这些方法通过低秩矩阵来近似微调期间的权重变化，并可在推理前与预训练权重合并。例如，（Zhang等，2023）采用奇异值分解并对较不重要的奇异值进行剪枝，以实现更高效的更新。（Hyeon-Woo等，2022）专注于用于联邦学习的低秩哈达玛积。（Qiu等，2023；Liu等，2023b）利用正交分解来优化扩散模型的微调。（Renduchintala等，2023）使用权重绑定进一步减少可训练参数。（Yeh等，2023）为Stable Diffusion提出了统一的LoRA家族框架。（Ponti等，2022）通过路由函数为不同任务从库中选择不同的LoRA组合。（Kopiczko等，2024）实现了可学习的缩放向量，以跨层调整一对共享的冻结随机矩阵。我们的研究也属于这一第三类方法，并通过全面实验验证了所提出方法与LoRA及其变体相比的有效性。

3. LoRA与全参数微调的模式分析

3.1. 低秩适配（LoRA）

基于微调过程中产生的更新具有较低“内在秩”的假设，LoRA（Hu 等，2022）提出使用两个低秩矩阵的乘积来逐步更新预训练权重。对于一个预训练权重矩阵 $W_0 \in \mathbb{R}^{d\times k}$，LoRA 使用低秩分解对权重更新 $\Delta W \in \mathbb{R}^{d\times k}$ 进行建模，表示为 BA，其中 $B \in \mathbb{R}^{d\times r}$ 和 $A \in \mathbb{R}^{r\times k}$表示两个低秩矩阵，且 $r \ll min\,(d,\ k)$。因此，微调后的权重 W' 可表示为：

$$W' = W_0 + \Delta W = W_0 + \underline{BA} \tag{1}$$

其中 W_0 在微调过程中保持静态，而下划线标注的参数正在被训练。矩阵 A 使用均匀Kaiming分布（He等，2015）进行初始化，而 B 初始设为零，导致训练开始时 $\Delta W = BA$ 为零。值得注意的是，对 ΔW 的这种分解可以用其他LoRA变体替代，例如VeRA（Kopiczko等，2024）。此外，根据公式（1），我们可以将学习到的 ΔW与预训练权重 W_0 合并，并提前得到 W'。

部署阶段，由于 W' 和 W_0 均处于 $\mathbb{R}^{d\times k}$的维度范围内，LoRA及其相关变体在推理过程中相比原始模型不会引入任何额外延迟。

3.2. 权重分解分析

LoRA（Hu等，2022）中提出的研究表明，LoRA 可被视为全参数微调的一种通用近似。通过逐步增加 LoRA 的秩 r 以匹配预训练权重的秩，LoRA 能够达到与全参数微调相似的表达能力。因此，许多先前的研究通常将 LoRA 与全参数微调之间的准确率差异主要归因于可训练参数数量有限，而未进行更深入的分析（Hu等，2022；Kopiczko等，2024）。受权重归一化（Salimans& Kingma，2016）的启发——该方法将权重矩阵重新参数化为幅度和方向以加速优化过程——我们引入了一种创新的权重分解分析方法。我们的分析将权重矩阵重构为两个独立的组成部分，幅度 和 方向，以揭示 LoRA 与全参数微调学习模式之间的内在差异。

分析方法： 该分析考察了LoRA和FT权重相对于预训练权重在幅度和方向上的更新，以揭示两者学习行为的根本差异。 $W \in \mathbb{R}^{d\times k}$ 的权重分解可表述为:

$$W = m\frac{V}{||V||_c} = ||W||_c\frac{W}{||W||_c} \tag{2}$$

其中 $m \in \mathbb{R}^{1\times k}$ 是幅度向量， $V \in \mathbb{R}^{d\times k}$ 是方向矩阵，$||\cdot||_c$ 表示矩阵各列的按向量范数。这种分解确保了 $V/||V|_c$ 的每一列均为单位向量，而 m 中的对应标量定义了每个向量的幅度。

在我们的权重分解分析中，我们选择在四项图像-文本任务上微调的 VL-BART 模型作为（Sung et al.，2022）提出的案例研究。遵循（Sung et al.，2022）的方法，仅对自注意力模块中的查询/值权重矩阵应用 LoRA。我们使用公式（2）对预训练权重 W_0、全参数微调后的权重 W_{FT}以及合并后的 LoRA 权重 W_{LoRA} 在查询/值权重矩阵上进行分解。其中 W_0 与 W_{FT} 之间的幅度和方向变化可定义如下:

$$\Delta M^t_{\text{FT}} = \frac{\sum_{n=1}^{k}|m^{n,t}_{\text{FT}} - m^n_0|}{k} \tag{3}$$

$$\Delta D^t_{\text{FT}} = \frac{\sum_{n=1}^{k}(1-\mathbf{cos}(V^{n,t}_{\text{FT}}, W^n_0))}{k} \tag{4}$$

这里， ΔM^t_{FT}和和 ΔD^t_{FT}表示 W_0 和 W_{FT} 之间的幅度差异和方向差异，在 t 训练步骤分别表示，其中 $\mathbf{cos}(\cdot,\cdot)$ 为余弦相似度函数。$M^{n\,t}_{\text{FT}}$、和 M^n_0 分别为其各自幅度向量中的 n^{th} 标量，而 $V^{n\,t}_{\text{FT}}$、和 W^n_0 分别为 n^{th} 列向量中的 V^t_{FT}和 W_0。根据公式（3）和公式（4）， W_{LoRA} 与 W_0 之间的幅度和方向差异以类似方式计算。我们从四个不同的训练步骤中选取检查点进行分析，包括全参数微调和 LoRA 各自的三个中间步骤及最终检查点，并对每个检查点执行权重分解分析，以确定 ΔM 和ΔD不同层

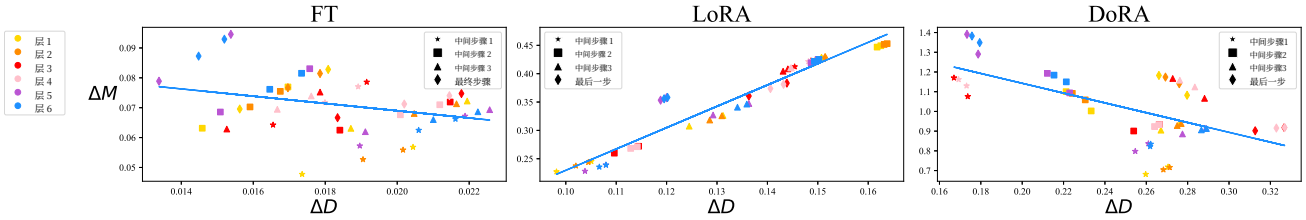
分析结果： 图 2 （a）和（b）展示了FT和LoRA查询权重矩阵的变化情况，每个点代表来自不同层和训练步数的查询权重矩阵中的一个（ΔD^t， ΔM^t）对。类似地，附录中的图 7 展示了值权重矩阵的修改情况。可以明显观察到，LoRA在所有中间步骤中均表现出一致的正斜率趋势，表明方向和幅度变化之间存在比例关系。相比之下，FT则呈现出更为多样的学习模式，且斜率相对为负。FT与LoRA之间的这一差异可能反映了它们各自的学习能力。虽然LoRA倾向于成比例地增加或减少幅度和方向的更新，但其缺乏进行更细微调整的精细能力。具体而言，LoRA在执行显著的幅度改变时伴随轻微的方向变化（或反之）方面表现不足，而这正是FT方法更典型的特征。我们推测，LoRA的这一局限性可能源于同时学习幅度和方向适应所带来的挑战，这对LoRA而言可能过于复杂。因此，在本研究中，我们旨在提出一种LoRA的变体，使其学习模式更接近FT，并能提升LoRA的学习能力。

4. 方法

4.1. 权重分解低秩适配

基于我们的权重分解分析所得出的洞察，我们提出了权重-**分**解低-**秩**适-**配**方-**法**（**DoRA**）。DoRA 首先将预训练权重分解为幅度和方向分量，并对这两者进行微调。由于方向分量在参数数量上较大，我们进一步使用 LoRA 对其进行分解以实现高效的微调。

我们的直觉有两个方面。首先，我们认为将 LoRA 限制为仅专注于方向适应，同时允许幅度分量可调，相比原始方法简化了任务，在原始方法中，LoRA 需要学习幅度



(a) (b) (c)

图 2。 不同层和中间步骤中查询矩阵的（a）全参数微调、（b）LoRA 和（c）DoRA 的幅度和方向更新。不同的标记代表不同训练步骤的矩阵，不同的颜色代表每层的矩阵。

和方向上的调整。其次，通过权重分解使优化方向更新的过程更加稳定，我们将在第4.2节更深入地探讨这一点。需要强调的是，DoRA 与权重归一化（Salimans& Kingma，2016）之间的主要区别在于它们的训练方法。权重归一化从零开始训练两个分量，这使得该方法对不同的初始化方式较为敏感。相反，DoRA 由于两个分量均始于预训练权重，因此避免了此类初始化问题。我们使用预训练权重 W_0 来初始化 DoRA，如公式（2）所述，其中 $m = ||W_0||_c$ 和 $V = W_0$ 在初始化后确定。然后我们保持 V冻结不变，并将 m 作为可训练向量。方向分量随后通过 LoRA 进行更新。DoRA 可以类似于公式（1）表示为:

$$W' = \underline{m}\frac{V + \Delta V}{||V + \Delta V||_c} = \underline{m}\frac{W_0 + \underline{BA}}{||W_0 + \underline{BA}||_c} \tag{5}$$

其中 ΔV 是通过将两个低秩矩阵 B 和 A 相乘而学习到的增量方向更新，下划线参数表示可训练参数。矩阵 $B \in \mathbb{R}^{d\times r}$ 和 $A \in \mathbb{R}^{r\times k}$ 按照 LoRA 的策略进行初始化，以确保在微调之前 W' 等于 W_0。此外，DoRA 可在推理前与预训练权重合并，从而不会引入任何额外延迟。

我们可视化了在与图 W_0 中全参数微调和LoRA相同的设置下，合并后的DoRA权重与2(c)中的查询权重矩阵之间的幅度和方向差异，并将值权重矩阵的可视化结果留在附录中。从DoRA和FT的$(\Delta D, \Delta M)$回归线可以看出，与LoRA的模式相反，DoRA和FT呈现出明显的负斜率。我们认为FT倾向于出现负斜率的原因在于预训练权重已经具备了适用于各种下游任务的大量知识。因此，当提供足够的学习能力时，仅需较大的幅度或方向改变就足以实现下游任务的适配。我们还进一步计算了 ΔD 与 ΔM 在FT、LoRA上的相关性，以及 DoRA，我们发现 FT 和 DoRA 分别表现出 -0.62 和 -0.31 的负相关值。相比之下，LoRA 显示出正相关，值为 0.83。总之，DoRA 表现出仅通过幅度上相对较小的变化实现显著的方向调整，或相反情况的能力，同时其学习模式更接近 FT，这表明它的

优于LoRA的学习能力。

4.2. DoRA 的梯度分析

在本节中，我们首先推导 DoRA 的梯度，并说明我们提出的分解如何有利于 ΔV 的优化。随后，我们从梯度的角度分析 DoRA 的学习模式，该模式倾向于呈现负斜率。

由公式（5）可得，损失 Loss $\mathcal{L}$ 关于 m 和 $V' = V + \Delta V$的梯度为:

$$\nabla_{V'}\mathcal{L} = \frac{m}{||V'||_c}\left(I - \frac{V'V'^{\mathbf{T}}}{||V'||_c^2}\right)\nabla_{W'}\mathcal{L} \tag{6}$$

$$\nabla_m\mathcal{L} = \frac{\nabla_{W'}\mathcal{L}\cdot V'}{||V'||_c} \tag{7}$$

公式（6）表明，权重梯度 $\nabla_{W'}\mathcal{L}$ 被$m/||V'||_c$ 缩放，并被投影到远离当前权重矩阵的方向。这两种效应有助于使梯度的协方差矩阵更接近单位矩阵，这对优化是有利的（Salimans& Kingma，2016）。此外，由于 $V' = V + \Delta V$，梯度 $\nabla_{V'}L$ 等价于 $\nabla_{\Delta V}L$。因此，这种分解带来的优化优势完全传递给了 ΔV，增强了 LoRA 的学习稳定性。

通过参考公式（7），我们可以进一步了解 DoRA 的学习模式。在接下来的讨论中，我们使用小写字母表示向量，而不是之前矩阵形式的表示法。设 $w'' = w' + \Delta w$ 为权重向量的参数更新，其中 $\Delta w \propto \nabla_{w'}\mathcal{L}$。在两个假设的更新

场景中， $S1$ 和 $S2$, $S1$ 涉及较小的方向更新（ΔD_{S1}），而 $S2$ 涉及较大的方向更新 (ΔD_{S2})。假设 $||\Delta w_{S1}|| = ||\Delta w_{S2}||$, 且在时间0时，我们有$\Delta v = 0$ 和 $v' = v_0$。由 $\Delta D_{S1} < \Delta D_{S2}$可得 $|\mathbf{cos}(\Delta w_{S1}, w')| > |\mathbf{cos}(\Delta w_{S2}, w')|$。由于 $\Delta w \propto \nabla_{w'}\mathcal{L}$，意味着 $|\cos(\nabla_{S1}\mathcal{L}\ w\prime)| > |\mathbf{cos}(\nabla_{S2}\mathcal{L}, w\prime)|$，。根据第4.1节，当 v 初始化为 v_0 且 $w\prime = w_0$ 在$|\mathbf{cos}(\nabla_{w\prime}\mathcal{L}\ w\prime)| = |\mathbf{cos}(\nabla_{w\prime}\mathcal{L}, v\prime)| =$，时间0时，得到$|\cos(\nabla_{w\prime}\mathcal{L}, v)|$。使用余弦相似性公式与$\Delta v = 0$:

$$cos(\nabla_{w'}\mathcal{L}, v') = cos(\nabla_{w'}\mathcal{L}, v) = \frac{\nabla_{w'}\mathcal{L} \cdot v}{||\nabla_{w'}\mathcal{L}||||v||} \quad (8)$$

表示 m_* 为向量 w' 的幅度标量，则关于 m_* 的式(7) 可重写为:

$$\nabla_{m_*}\mathcal{L} = \frac{\nabla_{w'}\mathcal{L} \cdot v'}{||v'||} = ||\nabla_{w'}\mathcal{L}|| \cdot cos(\nabla_{w'}\mathcal{L}, v) \quad (9)$$

已知 $||\Delta w_{S1}|| = ||\Delta w_{S2}||$ 对于 $S1$ 和 $S2$，且 $||\nabla^{S1}_{w'}\mathcal{L}|| = ||\nabla^{S2}_{w'}\mathcal{L}||$。因此，有:

$$||\nabla^{S1}_{w'}\mathcal{L}|| \cdot |cos(\nabla^{S1}_{w'}\mathcal{L}, v)| > ||\nabla^{S2}_{w'}\mathcal{L}|| \cdot |cos(\nabla^{S2}_{w'}\mathcal{L}, v)|$$

可推断出 $|\nabla^{S1}_{m_*}\mathcal{L}| > |\nabla^{S2}_{m_*}\mathcal{L}|$，这表明 $S1$相较于 $S2$ 具有更大的幅度更新，同时其方向变化小于$S2$ 的方向变化。我们的结论在实践中普遍成立，如图2 (c) 所示。因此，我们有效地展示了如何利用 DoRA 调整学习模式，使其偏离 LoRA 的模式，并更接近全参数微调（FT）的模式。

4.3. 训练开销的降低

在公式(1)中， W' 和 ΔW 的梯度是相同的。然而，使用将低秩适配重定向到方向分量的 DoRA 后，低秩更新的梯度与 W'的梯度不同，如公式(6)所示。这种差异在反向传播期间需要额外的内存。为解决此问题，我们建议将公式(5)中的 $||V + \Delta V||_c$ 视为常量，从而将其从梯度图中分离。这意味着尽管 $||V + \Delta V||_c$ 动态反映了 ΔV的更新，但在反向传播期间不会接收到任何梯度。通过这一修改，关于 m 的梯度保持不变，并将 $\nabla_{V'}\mathcal{L}$ 重新定义为:

$$\nabla_{V'}\mathcal{L} = \frac{m}{C}\nabla_{W'}\mathcal{L} \text{ where } C = ||V'||_c \quad (11)$$

这种方法大幅减少了梯度图的内存消耗，且在精度上没有明显差异。我们进行了消融实验，以评估所提出的修改对微调 LLaMA-7B 和 VL-BART 的影响。结果表明，在微调 LLaMA 时，该修改使训练内存减少了约 24.4%，在 VL-BART 中减少了 12.4%。此外，DoRA 在进行该修改后，VL-BART 的准确性保持不变，在 LLaMA 上与未修改的 DoRA 相比仅有 0.2 的微小差异。有关训练内存使用量和准确性差异的全面比较，请参见附录中的表7。

5. 实验

我们进行了多种实验，以展示 DoRA 在语言、图像和视频等多个任务领域的有效性。首先，我们在常识推理任务上对 LLaMA-7B/13B、LLaMA2-7B 和 LLaMA3-8B 进行微调，将 DoRA 与几种参数高效微调（PEFT）方法进行比较。随后，我们将研究从单模态扩展到多模态，使用 VL-BART 在多任务图像/视频-文本理解任务上以及使用 LLaVA-1.5-7B 进行视觉指令微调时，将 DoRA 与 LoRA 进行对比。接着，我们探讨了 DoRA 与 LoRA 及 VeRA（Kopiczko 等，2024）在 LLaMA-7B 和 LLaMA2-7B 上进行指令微调时的兼容性。此外，我们进行了一系列消融研究，说明无论微调训练样本数量和秩如何变化，DoRA 在性能上均优于 LoRA。最后，我们分析了 DoRA 的调优粒度，表明通过仅选择性地更新某些模块的方向分量，DoRA 可以用更少的可训练参数实现比 LoRA 更高的准确率。

5.1. 常识推理

我们评估了DoRA与LoRA以及几种基线方法在 LLaMA-7B/13B（*Touvron*等，2023）上的常识推理任务表现，这些基线方法包括提示学习（前缀）（Li & Liang, 2021）、串联适配器（串联）（Houlsby等，*2019*）和并联适配器（并联）（He等，2021）。我们还包含了使用 gpt-3.5-turbo API通过零样本思维链（Ope- nAI，2023; Wei等，2022）获得的ChatGPT准确率。

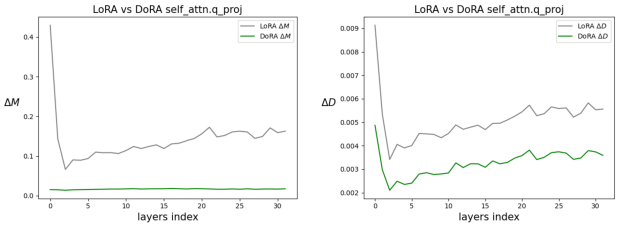
常识推理任务包含8个子任务，每个子任务都有预定义的训练集和测试集。我们遵循（Hu 等人，2023）的设定，将全部8个任务的训练数据集合并，形成最终的训练数据集，并在每个任务各自的测试数据集上进行评估。为了确保公平比较，我们最初按照 LoRA 配置对模型使用 DoRA 进行微调，在保持相同秩的同时仅调整学习率。如表1所示，DoRA 相较于 LoRA 在可训练参数数量上仅边际增加了 0.01%，这是由于引入了可学习的幅度分量（参数大小为 $1 \times k$）。然后，我们进一步将 DoRA 中使用的秩减半

表*1.* LLaMA 7B/13B、LLaMA2 7B 和 LLaMA3 8B 在八种常识推理数据集上使用各种 PEFT 方法的准确率比较。

模型	PEFT方法	# 参数(%)	BoolQ	PIQA	SIQA	HellaSwag	WinoGrande	ARC-e	ARC-c	OBQA	平均
ChatGPT	-	-	73.1	85.4	68.5	78.5	66.1	89.8	79.9	74.8	77.0
LLaMA-7B	前缀	0.11	64.3	76.8	73.9	42.1	72.1	72.9	54.0	60.6	64.6
	串联	0.99	63.0	79.2	76.3	67.9	75.7	74.5	57.1	72.4	70.8
	并联	3.54	67.9	76.4	78.8	69.8	78.9	73.7	57.3	75.2	72.2
	LoRA	0.83	68.9	80.7	77.4	78.1	78.8	77.8	61.3	74.8	74.7
	DoRA† (我们的)	0.43	70.0	82.6	79.7	83.2	80.6	80.6	65.4	77.6	**77.5**
	DoRA (我们的)	0.84	69.7	83.4	78.6	87.2	81.0	81.9	66.2	79.2	**78.4**
LLaMA-13B	前缀	0.03	65.3	75.4	72.1	55.2	68.6	79.5	62.9	68.0	68.4
	串联	0.80	71.8	83	79.2	88.1	82.4	82.5	67.3	81.8	79.5
	并联	2.89	72.5	84.9	79.8	92.1	84.7	84.2	71.2	82.4	81.4
	LoRA	0.67	72.1	83.5	80.5	90.5	83.7	82.8	68.3	82.4	80.5
	DoRA† (Ours)	0.35	72.5	85.3	79.9	90.1	82.9	82.7	69.7	83.6	**80.8**
	DoRA (Ours)	0.68	72.4	84.9	81.5	92.4	84.2	84.2	69.6	82.8	**81.5**
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（我们的）	0.84	71.8	83.7	76.0	89.1	82.6	83.7	68.2	82.4	**79.7**
LLaMA3-8B	LoRA	0.70	70.8	85.2	79.9	91.7	84.3	84.2	71.2	79.0	80.8
	DoRA† （我们的）	0.35	74.5	88.8	80.3	95.5	84.7	90.1	79.1	87.2	**85.0**
	DoRA（我们的）	0.71	74.6	89.3	79.9	95.5	85.6	90.5	80.4	85.8	**85.2**

表 1 表明，DoRA 在 LLaMA-7B/13B、LLaMA2-7B 和 LLaMA3-8B 上均持续超越所有基线方法。值得注意的是，在 LLaMA-7B 模型中，LoRA 的性能已超过其他基线方法，而 DoRA 进一步将准确率提升了 3.7%，超过了 ChatGPT 的准确率水平。相反地，对于 LoRA 效果不如 Parallel adapter 的 LLaMA-13B 模型，DoRA 相较于 LoRA 提升了 1% 的准确率，并且达到了与 Parallel adapter 相当的准确率，同时仅需 Parallel adapter 四分之一的可训练参数，且不会像 Parallel adapter 那样增加任何推理开销。此外，DoRA 在 LLaMA2-7B 和 LLaMA3-8B 上分别以 2.1% 和 4.4% 的优势持续超越 LoRA。进一步地，DoRA† 在LLaMA-7B 上的表现超过 LoRA 2.8%，在LLaMA-13B 上超过 1%，在 LLaMA2-7B 上超过 2.9%，在 LLaMA3-8B 上超过 4.2%，尽管其可训练参数数量仅为 LoRA 的一半。这一结果表明，DoRA 的引入增强了LoRA 的学习能力，从而减少了为超越 LoRA 准确率而需要更高秩的需求

此外，在前面的章节中，我们假设幅度更新与方向更新之间的负相关性比正相关性更优。这是因为预训练权重已经包含了大量适用于下游任务的知识，单独较大的幅度或方向改变已足够。



(a) (b)

图*3.* LoRA/DoRA 与不同层查询矩阵的预训练权重之间的幅度（a）和方向（b）差异。

用于下游适配。为了进一步验证我们的假设，我们以在常识推理数据集上使用DoRA/LoRA微调的LLaMA2-7B作为案例研究。我们可视化了不同模块和层中DoRA/LoRA权重与预训练模型权重之间的幅度（ΔM）和方向差异（ΔD）。在图3 (a) 和 (b) 中，我们观察到DoRA微调的权重在幅度和方向上均更接近预训练权重，偏离较小，而LoRA微调权重的差异则明显更大。结合实验结果中DoRA显著优于LoRA的表现，我们可以得出结论：先前的假设是成立的，一个鲁棒的基础模型并不需要进行显著更改即可实现有效的下游适配，而能够进行更细粒度的幅度和方向更新解释了DoRA相较于LoRA的优越性。我们将在附录中展示对Value和Key权重矩阵的可视化。

5.2. 图像/视频-文本理解

表2. 基于VL-BART主干的VQA、GQA、NVLR^2 和COCO字幕上的多任务评估结果。

方法	# 参数(%)	VQA ^{v2}	GQA	NVLR ^{2}	COCO字幕	平均
FT	100	66.9	56.7	73.7	112.0	77.3
LoRA	5.93	65.2	53.6	71.9	115.3	76.5
DoRA（我们的）	5.96	65.8	54.7	73.1	115.9	**77.4**

表3.基于VL-BART主干的TVQA、How2QA、TVC和YC2C上的多任务评估结果。

方法	# 参数(%)	TVQA	How2QA	TVC	YC2C	平均
FT	100	76.3	73.9	45.7	154	87.5
LoRA	5.17	75.5	72.9	44.6	140.9	83.5
DoRA（我们的）	5.19	76.3	74.1	45.8	145.4	**85.4**

在证明 DoRA 能够在微调大语言模型时持续实现更高的准确率后，我们希望进一步验证 DoRA 在多模态微调任务中是否仍具竞争力。我们将 DoRA 与 LoRA 及全量微调在 VL-BART 上进行比较，VL-BART 包含一个视觉编码器（CLIP-ResNet101 (Radford et al., 2021)）和一个编码器-解码器语言模型（$\text{BART}_{\text{Base}}$ (Lewis et al., 2020)），评估涵盖四个不同的图像-文本任务：VQA^{v2} (Goyal et al., 2017) 和 GQA (Hudson & Manning, 2019) 用于视觉问答，NLVR^2 (Suhr et al., 2019) 用于视觉推理，MSCOCO (Chen et al., 2015) 用于图像字幕生成；以及来自 VALUE (Li et al., 2021) 基准的四个不同视频-文本任务：TVQA (Lei et al., 2018) 和 How2QA (Li et al., 2020) 用于视频问答，TVC (Lei et al., 2020) 和 YC2C (Zhou et al., 2018) 用于视频字幕生成。

我们采用与（Sung 等人，2022）相同的框架，在多任务框架中对 VL-BART 进行微调，以处理图像/视频-文本任务。在应用 DoRA 时，我们采用了与（Sung 等人，2022）中所述 LoRA 相同的设置。完整的超参数见表9 。图像/视频-文本任务中 LoRA 和 FT 的结果直接引用自（Sung 等人，2022）。我们可以看到，在表2 和表3中，DoRA 在保持可训练参数数量相近的情况下，准确率均一致超过了 LoRA。特别是，在图像-文本理解任务中，DoRA 的性能比 LoRA 提高了近 1%，达到了 FT 的准确率水平。此外，在视频 - 文本理解任务中，DoRA 的准确率比 LoRA 高出约 2%。

5.3. 视觉指令微调

我们进一步扩大模型规模，并在视觉指令微调任务上将DoRA与LoRA和全参数微调进行比较，

表4。 LLaVA-1.5-7B 在七种广泛视觉-语言任务上的视觉指令微调评估结果。我们直接使用（Liu 等人，2023a ）以复现他们的结果。

方法	# 参数(%)	Avg.
FT	100	66.5
LoRA	4.61	66.9
DoRA（我们的方法）	4.63	**67.6**

使用LLaVA-1.5-7B（Liu等，2023a），该模型由语言模型Vicuna-1.5-7B（Peng等，2023）和视觉编码器CLIP ViT-L/336px（Radford等，2021）组成。训练数据集包含来自VQA（Goyal等，2017；Hudson & Manning, 2019；Marino等，2019；Schwenk等，2022）、光学字符识别（Mishra等，2019；Sidorov等，2020）、区域级视觉问答（Kazemzadeh等，2014；Krishna等，2017；Mao等，2016）、视觉对话（Liu等，2023a）以及语言对话数据的多个数据集。我们遵循（Liu等，2023a）的设置来过滤训练数据并构建调优提示格式。为了公平比较，DoRA采用与（Liu等，2023a）提供的LoRA配置相同的配置。微调后的模型随后在七个视觉-语言基准上进行评估：VQA^{v2} （Goyal等，2017）、GQA（Hudson & Manning，2019）、VisWiz（Gurari等，2018）、SQA（Lu等，2022）、VQA^{T} （Singh等，2019）、POPE（Li等，2023）和MMBench（Liu等，2023c）。

从表 4 可以观察到，LoRA 的平均准确率已经超过了全参数微调，这可能意味着全参数微调可能存在过拟合问题。鉴于 DoRA 的设计目的是提升 LoRA 的性能，使其更接近全参数微调的效果，在全参数微调劣于 LoRA 的情况下，DoRA 相较 LoRA 的提升可能不如其他实验中那样显著——在那些实验中，全参数微调通常优于 LoRA。尽管如此，DoRA 仍表现出优于 LoRA 和全参数微调的性能，相比 LoRA 平均提升了 0.7%，相比全参数微调平均提升了 1.1%。超参数设置见表 10 ，各评估基准的分数见表 12 。

5.4. DoRA 与其他 LoRA 变体的兼容性

由公式(1)可知， ΔW 可通过不同的LoRA变体进行调整。在DoRA中，公式(5)中引入的增量方向更新 ΔV 同样可以替换为其他LoRA变体。在本节中，我们选择VeRA（Kopiczko等，2024）作为案例研究，以探讨DoRA与其他LoRA变体的兼容性。VeRA建议冻结一组跨所有层共享的独特随机低秩矩阵，并仅使用最小的层特定可训练缩放向量来捕获每层的增量更新。该方法使VeRA相比LoRA显著减少了可训练参数

表5。 GPT-4 对全参数微调 LLaMA-7B/LLaMA2-7B 生成的答案在 MT-Bench 上的平均评分。

模型	PEFT方法	# 参数 (%)	得分
LLaMA-7B	LoRA	2.31	5.1
	DoRA（我们的）	2.33	**5.5**
	VeRA	0.02	4.3
	DVoRA（我们的）	0.04	**5.0**
LLaMA2-7B	LoRA	2.31	5.7
	DoRA（我们的）	2.33	**6.0**
	VeRA	0.02	5.5
	DVoRA（我们的）	0.04	**6.0**

数量达10倍，同时对准确率的影响极小。我们将VeRA应用于DoRA中的方向更新，并将这种组合命名为DVoRA。我们在LLaMA-7B和LLaMA2-7B上评估DVoRA和DoRA相较于VeRA和LoRA的有效性，重点是在清理后的Alpaca数据集10K子集（Taori等，2023）上进行指令微调。我们使用VeRA的官方实现来获得VeRA和LoRA的结果，并采用与VeRA和LoRA相同的训练设置对DVoRA和DoRA进行微调（详见附录中的表11）。然后在MT-Bench基准（Zheng等，2023）上评估微调后模型的性能，通过生成针对预定义的80个多轮问题的模型回答来进行。这些回答随后由GPT-4进行评估，GPT-4会审阅每个答案并给出满分10分的数值评分。

表 5 展示了 DVoRA、DoRA、VeRA 和 LoRA 的平均得分，表明我们提出的方法在 LLaMA-7B 和 LLaMA2-7B 上均比 VeRA 和 LoRA 表现出持续的提升。这有效展现了 DoRA 与 VeRA 的兼容性。特别是，DVoRA 融合了 DoRA 和 VeRA 的优势特性，取得的得分与 LoRA 相当甚至更优，但参数数量却显著更少。例如，DVoRA 在 LLaMA-7B 上比 VeRA 高出 0.7/0.5 分，并分别在 LLaMA-7B 和 LLaMA2-7B 上达到了与 LoRA 和 DoRA 相同的准确度水平。此外，我们在附录中（表 13 和 14）展示了一些从 MT-Bench 中选取的问题以及使用 DVoRA 和 VeRA 微调后的 LLaMA2-7B 的回答，从中可以看出 DVoRA 给出的答案往往更加精确且结构清晰。

接下来，为了进一步评估 DoRA 在不同训练数据量下的竞争力，考虑到在实际情况下，通常难以获得大规模的微调数据集。我们将 DoRA 与 LoRA 以及 DVoRA 与 VeRA 进行比较，对 LLaMA2-7B/LLaMA-7B 进行指令微调，样本规模分别为 1000、4000、7000 和 10000

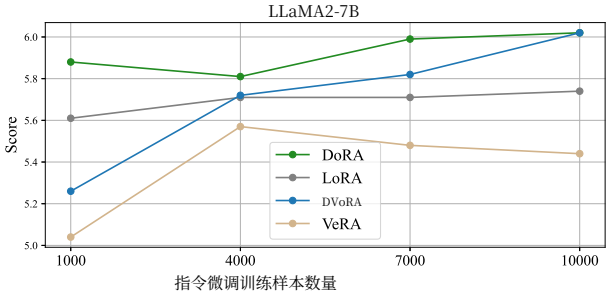


图4.使用不同数量的Alpaca训练样本对LLaMA2-7B进行微调后在MT-Bench上的性能。

（Kopiczko等，2024）所采用的设置。我们在图 4 中展示了每种方法在 LLaMA2-7B 上的平均性能，在附录图 9 中展示了 LLaMA-7B 上的结果。结果表明，无论训练样本数量多少，DoRA 和 DVoRA 始终优于 LoRA 和 VeRA。例如，在 7000 个训练样本的情况下，DoRA 和 DVoRA 分别以 0.3 和 0.33 的优势超过 LoRA 和 VeRA。即使样本量减少到 1000，DoRA 和 DVoRA 仍分别领先 LoRA 和 VeRA 0.29 和 0.22。这说明我们的方法在各种训练样本量下均能持续优于 LoRA 和 VeRA。

5.5. DoRA 对不同秩设置的鲁棒性

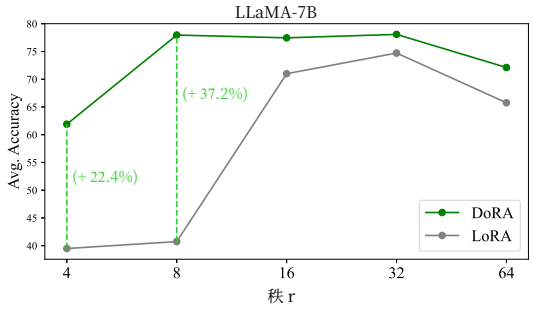


图5.LoRA 和 DoRA 在 LLaMA-7B 上不同秩的常识推理任务中的平均准确率。

本节通过调整 r 在集合 $\{4, 8, 16, 32, 64\}$ 中的值，探讨不同秩设置对DoRA和LoRA的影响，并评估微调后的LLaMA-7B在第5.1节所述常识推理任务上的性能。LoRA和DoRA在不同秩下的平均准确率如图5所示，详细数据见表15。从图5中可以看出，DoRA在所有秩设置下均持续优于LoRA。值得注意的是，对于秩

表6. LLaMA 7B/13B在两种不同DoRA调优粒度下的准确率比较。**m** 列和**V** 列分别表示具有可调幅度和方向分量的模块。每个模块以其首字母表示：(Q)uery、(K)ey、(V)alue、(O)utput、(G)ate、(U)p、(D)own。

模型	PEFT方法#参数 (%)	m	V	Avg.
LLaMA-7B	LoRA 0.83	-	-	74.7
	DoRA (我们的) 0.84	QKVUD	QKVUD	78.1
	DoRA (我们的) 0.39	QKVOGUD	QKV	77.5
LLaMA-13B	LoRA 0.67	-	-	80.5
	DoRA (我们的) 0.68	QKVUD	QKVUD	81.5
	DoRA (我们的) 0.31	QKVOGUD	QKV	81.3

低于8的情况，LoRA的平均准确率分别下降至 $r = 8$ 为40.74%和 $r = 4$为39.49%。相比之下，DoRA仍保持较高的准确率，$r = 8$ 为77.96%，$r = 4$为61.89%，表明其在不同秩设置下均具有更强的鲁棒性以及持续优越的性能。

5.6. 微调粒度分析

图 2 中的可视化表明，幅度上的显著变化通常导致相对较小的方向变化。基于这一观察以及方向更新占据了大部分可训练参数的事实，促使我们探究是否可以通过仅更新特定模块的幅度分量，同时对剩余的线性模块继续同时更新其幅度和方向分量，来减少可训练参数的数量。

我们的研究结果表明，与 Hu 等人（2023）为 LoRA 建议的原始配置不同，后者需要同时更新多头注意力和 MLP 层才能实现最佳性能，而 DoRA 仅通过更新多头层的方向和幅度分量以及 MLP 层的幅度分量，即可实现更优的准确性。具体而言，如表 6所示，通过更新 QKV 模块的方向和幅度分量，并仅更新其余层的幅度分量，DoRA 在 LLaMA-7B 上比 LoRA 高出 2.8%，在 LLaMA-13B 上高出 0.8%，同时使用的可训练参数数量不到 LoRA 的一半。

6. 更广泛的影响

6.1. QDoRA：对QLoRA的增强

尽管使用 PEFT 微调大语言模型显著降低了训练过程中的内存开销，但仍需要大量 GPU 内存将模型权重初始加载到 GPU 上。为进一步降低微调的内存需求，QLoRA（Dettmers 等，2023）提出将预训练模型量化至 4 比特，并在冻结的低比特主干模型之上进行 LoRA 微调。基于我们提出的

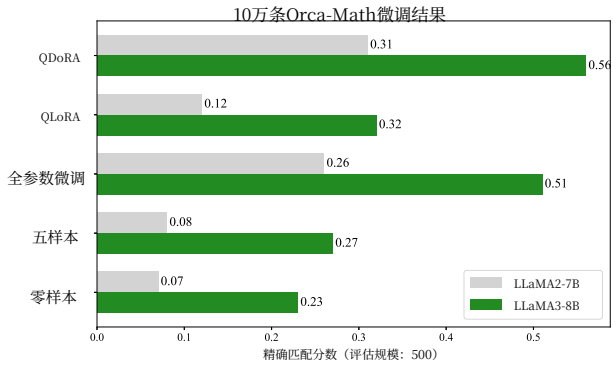


图6。 LLaMA2-7B/LLaMA3-8B 在 Orca-Math 上使用 QDoRA、QLoRA 和 FT 的准确率比较（Mitra 等， 2024）。

DoRA 缩小了 LoRA 与全参数微调之间的差距，因此自然可以探索 DoRA 是否也能在 QLoRA 框架内提升 LoRA 的准确性。最近，（Kerem Turgutlu，2024）启动了一个项目，将 QLoRA 中的 LoRA 组件替换为 DoRA，称之为 QDoRA，并结合 Fully Sharded Data Parallel（FSDP）（Zhao 等，2023）整合训练流程，以实现跨多个 GPU 的模型分片与并行训练。他们使用 Orca-Math（Mitra 等，2024）数据集对 LLaMA2- 7B/LLaMA3-8B 进行 QDoRA、QLoRA 和全参数微调的实验。训练集包含 10 万条样本，其中 500 条保留用于评估，评估指标为精确匹配分数。除了微调模型外，他们还报告了零样本、少样本以及结合训练后量化（PTQ）的全参数微调结果，其中全参数微调模型在训练后被量化为 BnB NF4 格式。根据图 6，QDoRA 在 LLaMA2- 7B 和 LLaMA3-8B 上不仅分别以 0.19/0.23 显著超越 QLoRA，而且在两个模型上还略微优于全参数微调，同时占用更少的内存。这表明 QDoRA 能够有效结合 QLoRA 的参数效率与全参数微调更细粒度的优化能力。这些初步发现表明，QDoRA 具有巨大的潜力，有望通过大幅降低大语言模型微调所需的 GPU 内存需求，极大地造福开源社区。

6.2. 文本到图像生成

最近，随着扩散模型规模的扩大，LoRA 已成为高效微调大型稳定扩散模型的一种流行方法。在本节中，我们旨在探讨 DoRA 相较于 LoRA 的优势是否也延伸到了文本到图像生成任务中。我们遵循 DreamBooth（Ruiz 等，2023）的训练流程进行微调

SDXL（Podell 等，2023），并利用 HuggingFace 开发的先进训练脚本。LoRA 和 DoRA 的超参数设置保持一致，我们使用两个具有挑战性的数据集进行模型微调：3D 图标和乐高套装。为公平比较，LoRA 和 DoRA 生成图像时的样本种子保持相同。生成的图像见附录中的图 10 和 11 。结果表明，在使用相同训练设置的情况下，DoRA 实现了比 LoRA 显著更好的个性化效果，并且更准确地反映了训练目标。例如，在图 10 中，DoRA 输出的第一个子图在图像周围包含一个独特的圆形方框，这是所有训练目标共有的特征。相比之下，所有 LoRA 输出中均未出现该特征。对于乐高训练目标也可观察到类似现象，只有 DoRA 的输出在生成的图像中始终包含了乐高标志。

7. 结论

在本研究中，我们首先进行了一种新颖的权重分解分析，以揭示 LoRA 与FT 之间的不同学习模式。基于这些发现，我们提出了 DoRA，这是一种与LoRA 及其变体兼容且更接近 FT 学习行为的微调方法。DoRA 在各种微调任务和模型架构中始终优于 LoRA。具体而言，DoRA 在常识推理和视觉指令微调任务中表现优于 LoRA。此外，DoRA 在 Alpaca 指令微调任务中也表现出与 VeRA 的兼容性。而且，DoRA 可被视为 LoRA 的零成本替代方案，因为其分解出的幅度和方向分量可在训练后合并回预训练权重中，确保不会带来额外的推理开销。对于未来的工作，我们希望探索DoRA 在语言和视觉之外领域（尤其是音频领域）的泛化能力