大语言模型参数高效微调技术综述

宋泰霖

中国科学院计算技术研究所，中国科学院大学

摘要

随着大语言模型（LLMs）参数量的指数级增长，全量微调带来的巨额算力消耗和内存开销已成为实际应用中的关键瓶颈。参数高效微调（PEFT）技术通过仅优化模型的部分参数而非全部权重，在保持模型性能的同时显著降低了计算成本，成为近年来 LLMs 适配下游任务的核心研究方向。本文系统综述了大语言模型参数高效微调的研究进展，以 LoRA（Low-Rank Adaptation）技术为基础，从网络级优化、矩阵级优化和维度级优化三个核心维度，全面梳理了各类参数高效适配方法的技术原理、创新点与性能表现。本文详细阐述了混合专家（MoE）架构与低秩适配的融合策略、权重矩阵的结构化优化方法以及基于子空间投影的维度级适配技术，并探讨了这些方法在单域任务、多域混合任务、多模态任务及持续学习等场景中的应用。最后，分析了当前参数高效微调技术面临的参数干扰、动态适配不足、跨模态适配瓶颈等挑战，并展望了未来在动态专家分配、跨模态统一框架、可解释性增强等方向的研究趋势，为该领域的进一步发展提供参考。

1 引言

近年来，以 GPT、LLaMA、CLIP 为代表的大语言模型和视觉-语言模型在自然语言处理、计算机视觉及跨模态任务中展现出卓越的性能。这些模型通过在海量数据上的预训练，积累了丰富的通用知识，能够通过微调适配特定下游任务。然而，随着模型参数量突破千亿甚至万亿级别，传统的全量微调方法面临着难以逾越的障碍：一方面，全量微调需要同时存储和更新数十亿甚至上万亿个参数，对硬件内存提出了极高要求，普通计算设备难以支撑；另一方面，大规模参数的梯度计算过程消耗巨大算力，导致微调成本高昂，且训练周期冗长，不利于快速迭代。此外，全量微调还可能导致模型过拟合下游任务数据，丧失预训练阶段获得的通用泛化能力，出现“灾难性遗忘”现象。

为解决全量微调的上述问题，参数高效微调技术应运而生。参数高效微调的核心思想是冻结预训练模型的大部分权重，仅引入少量可训练参数，如低秩矩阵、适配器模块、专家网络等，通过优化这些参数使模型适配下游任务。与全量微调相比，参数高效微调不仅将可训练参数占比从 100%降低至不足 1%，显著减少了计算和内存开销，还能更好地保留预训练模型的通用知识，缓解灾难性遗忘问题。自 LoRA 技术提出并成为参数高效微调的奠基性方法以来，相关研究围绕网络结构设计、权重矩阵优化、子空间投影等多个方向展开，形成了丰

富的技术体系，涵盖单任务适配、多任务适配、多模态适配、持续学习等多个应用场景。

本文旨在对大语言模型参数高效微调技术进行全面且系统的综述。首先，介绍参数高效微调的相关背景和核心目标；其次，以网络级优化、矩阵级优化、维度级优化为三大主线，详细阐述各类方法的技术原理、创新点及性能优势，整合近年来的代表性研究成果；再次，探讨参数高效微调技术在不同任务场景中的应用实践；最后，分析当前研究面临的挑战，并对未来发展方向进行展望。本文的综述范围涵盖文本任务、多模态任务、持续学习等多个领域，重点关注以 LoRA 为基础的扩展方法，力求为相关领域的研究人员提供全面的技术视野和清晰的研究脉络。

2 相关背景

2.1 大语言模型微调基础

大语言模型的微调本质是在预训练模型的基础上，利用下游任务数据调整模型参数，使模型学习任务特定知识，从而提升在该任务上的性能。预训练模型通过海量数据学习到的通用语言知识（如语法、语义、常识等）是微调的基础，而微调过程则是将这些通用知识与下游任务的特定需求相结合。传统的全量微调通过更新模型的所有参数来实现这一目标，能够最大程度地适配任务，但如前所述，其高昂的计算成本限制了其在大规模模型上的应用。

参数高效微调作为全量微调的替代方案，其核心设计原则是“冻结主干，优化分支”。具体而言，预训练模型的主体结构（如 Transformer 的注意力层、Feed-Forward 层）保持冻结，仅在模型中插入少量可训练的轻量化模块，这些模块负责学习下游任务的特定知识，并将其融入预训练模型的表征中。这种设计不仅大幅减少了可训练参数的数量，降低了计算和内存开销，还能通过限制参数更新的范围，减少对预训练知识的破坏，从而缓解灾难性遗忘。

2.2 参数高效微调的核心目标

参数高效微调的研究旨在实现以下三大核心目标：一是算力与内存效率，通过最小化可训练参数数量，降低微调过程中的计算复杂度和内存占用，使大规模模型能够在普通硬件设备上完成微调；二是任务适配性能，确保微调后的模型在下游任务上的性能接近或超过全量微调，同时保持良好的泛化能力；三是知识保留与迁移，在适配下游任务的同时，最大限度地保留预训练模型的通用知识，避免灾难性遗忘，并且能够实现知识在不同任务间的有效迁移。

为实现这些目标，参数高效微调方法通常需要满足以下技术要求：首先，引入的可训练模块应具备轻量化特性，避免增加过多的计算开销；其次，模块的设计应与预训练模型的结构相兼容，确保任务特定知识能够有效融入模型表征；最后，优化过程应能够平衡任务适配与知识保留，避免过拟合或欠拟合。

2.3 关键评价指标

参数高效微调方法的性能通常通过以下指标进行评价：一是参数效率，即可训练参数占模型总参数的比例，比例越低，参数效率越高；二是计算效率，包括微调过程中的训练时间、GPU 内存占用、推理延迟等，反映方法的实际部署可行性；三是任务性能，即在下游任务上的准确率、F1 值、困惑度等传统评价指标，用于衡量模型的适配效果；四是知识保留能力，通过测试模型在预训练相关任务（如常识问答）上的性能变化，评估灾难性遗忘的缓解程度；五是泛化能力，包括在未见过的任务或领域上的 Zero-shot、Few-shot 性能，反映模型的通用适配能力。

这些指标从不同维度全面评估了参数高效微调方法的优劣，为方法之间的对比提供了统一的标准。在实际研究中，通常需要在这些指标之间进行权衡，例如，某些方法可能通过增加少量可训练参数来换取显著的性能提升，而另一些方法则更注重极致的参数效率以适应资源受限场景。

3 参数高效微调的核心方法分类

LoRA 作为参数高效微调的奠基性方法，其核心原理是冻结预训练模型的线性层，在每个线性层上额外训练两个低秩矩阵 A 和 B，其中 A 的维度为 d×r，B 的维度为 r×k（r≪d,k），权重更新量 ΔW=AB，如图 1 所示。[1]这种低秩假设基于观察到的微调过程中权重更新的低秩特性，能够以极少的参数实现有效的任务适配，且在微调结束后可将低秩矩阵与原权重矩阵合并，不增加推理延迟。LoRA 的简洁性和有效性为后续网络级优化方法提供了基础，后续方法主要围绕专家网络的构建、门控机制的设计以及多任务适配能力的提升展开。

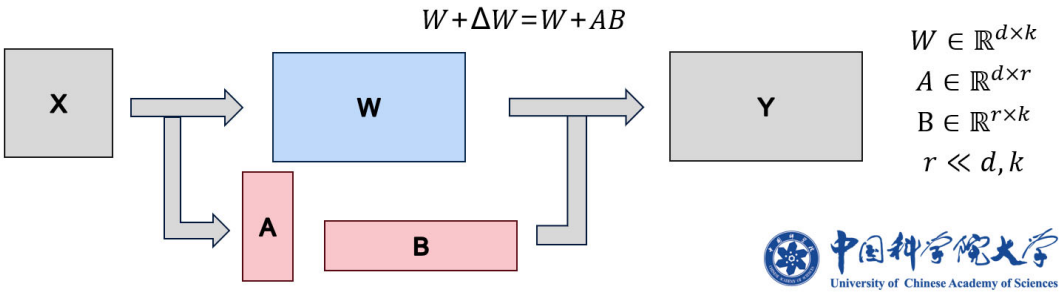


图 1 LoRA 低秩适应技术

参数高效微调方法的发展以 LoRA 技术为起点，逐渐形成了三大核心优化维度：网络级优化、矩阵级优化和维度级优化。网络级优化侧重于构建由多个专家或组件组成的网络结构，通过动态路由实现任务与专家的匹配；矩阵级优化聚焦于权重矩阵本身的结构设计，通过改进低秩矩阵的形式或引入新的矩阵结构，提升参数效率和适配性能；维度级优化则通过对权重矩阵的子空间进行划分，在不同维度上分别优化，实现知识保留与任务适配的平衡。这三个维度从宏观网络结构到微观参数维度，构成了参数高效微调的完整技术体系。也有部

分方法存在不止一个维度的优化，本文按最细粒度的维度进行划分。

3.1 网络级优化

网络级优化的核心思想是构建基于混合专家（Mixture of Experts, MoE）的适配框架，将多个低秩适配器（LoRA 专家）与门控网络相结合，使模型能够根据输入任务的特性动态选择合适的专家进行适配。这种设计不仅能够通过专家的专业化提升任务适配性能，还能通过共享主干网络减少参数冗余，同时通过门控机制实现知识的有效整合。网络级优化方法根据任务数据的性质可分为单域划分方法和多域混合方法，前者针对单个数据库自动进行领域划分或专家分配，后者则面向人工划分的不同领域数据，学习混合网络模型。

HMVLM（Human Motion-Vision-Lanuage Model）是一种面向文本-图像-三维姿态多模态任务的单域划分方法，其核心创新在于构建了包含调节专家权重的门控网络和多个 LoRA 专家的 MoE 混合专家模型。[2]为了增强知识保留，HMVLM 特别引入了一个冻结为零的零专家（zero expert），当模型处理与人类运动无关的通用语言任务时，门控网络会优先选择零专家，从而保留预训练模型的原始知识，避免灾难性遗忘。此外，HMVLM 还设计了身体部位特定的 tokenization 方法，将人体划分为不同的关节组，通过空间 Transformer 分别编码，提升了姿态表征的空间分辨率，使其能够更好地适配多模态任务中的姿态相关任务。HMVLM 网络结构如图 2 所示。在实验中，HMVLM 在文本到运动生成、单目姿态估计、运动视频理解等任务上表现出优异的性能，同时仅出现轻微的语言能力下降，验证了其在多模态适配和知识保留方面的有效性。

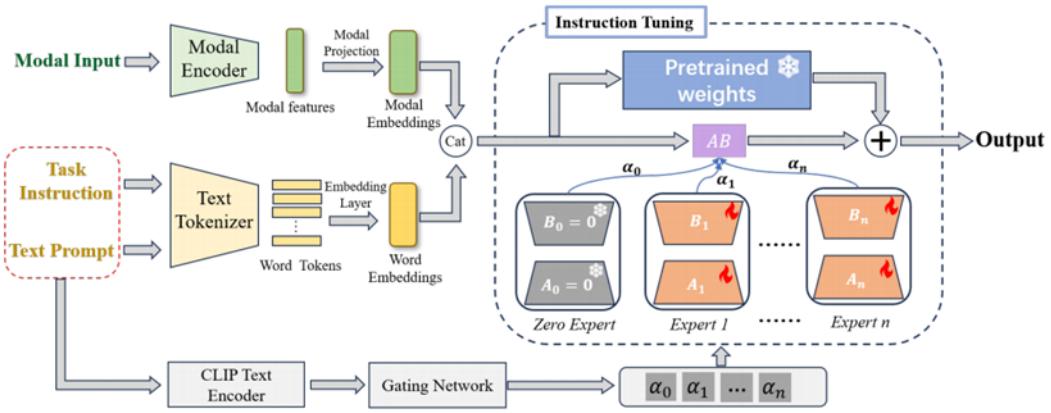


图 2 HMVLM 网络结构

TTMM（Test-time Training via Model Merging）是另一种单域划分方法，主要针对文本任务。[3]其核心流程包括训练时的数据聚类、专家训练和测试时的专家选择与加权融合。在训练阶段，TTMM 首先对文本数据进行聚类，在每个聚类上训练一个 LoRA 专家，使每个专家专门适配一类数据的特性；在测试阶段，根据输入与各类质心的稀疏注意力计算专家权重，通过加权平均得到最终结果。TTMM 方法原理如图 3 所示。由于 GPU 可同时只计算一个专家的输出，TTMM 能够

在大幅增加专家数量的同时，不会显著增加算力消耗，从而提升了模型对不同数据分布的适配能力。这种数据驱动的专家分配方式，使模型能够自动适应单域数据内部的分布差异，提升了任务适配的精细度。

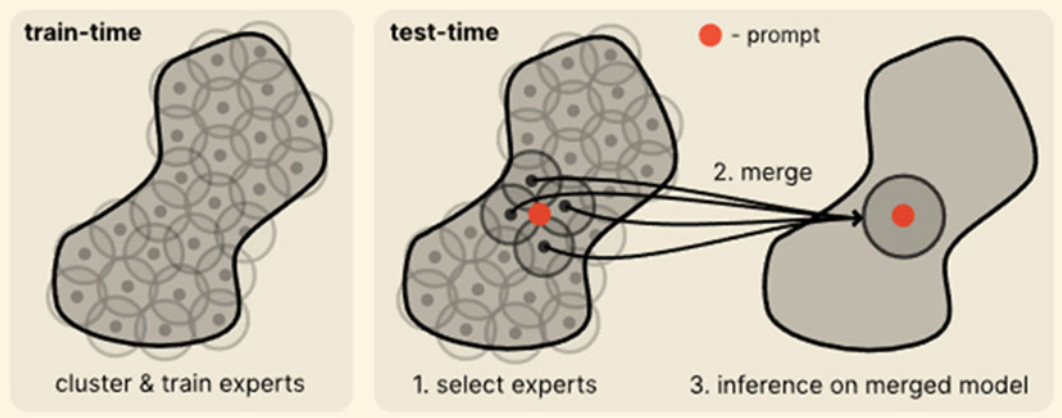


图 3 TTMM 方法原理

MoE-Adapters 是面向文本-图像理解任务的多域混合方法，其核心设计是为每类领域数据训练特定的编码器，并通过门控网络选择 k 个最相关的专家进行训练。[4]在训练过程中，MoE-Adapters 采用“激活-冻结”策略，当训练新的领域数据时，已训练完成的专家会被冻结，通过类间知识迁移提升模型的泛化性。为了保留模型的 Zero-shot 能力，MoE-Adapters 引入了分布判别自动选择器（DDAS），该模块通过训练多个自编码器学习不同领域数据的分布特征，在推理时自动判断输入数据的分布类型，将分布内数据路由至 MoE-Adapters，分布外数据路由至原始预训练模型（如 CLIP），从而实现对已见任务的精准适配和对未见任务的 Zero-shot 预测。MoE-Adapters 网络结构如图 4 所示。MoE-Adapters 的设计解决了多域混合任务中的知识迁移和 Zero-shot 能力保留问题，在多个视觉-语言理解任务中表现出优于传统适配器方法的性能。

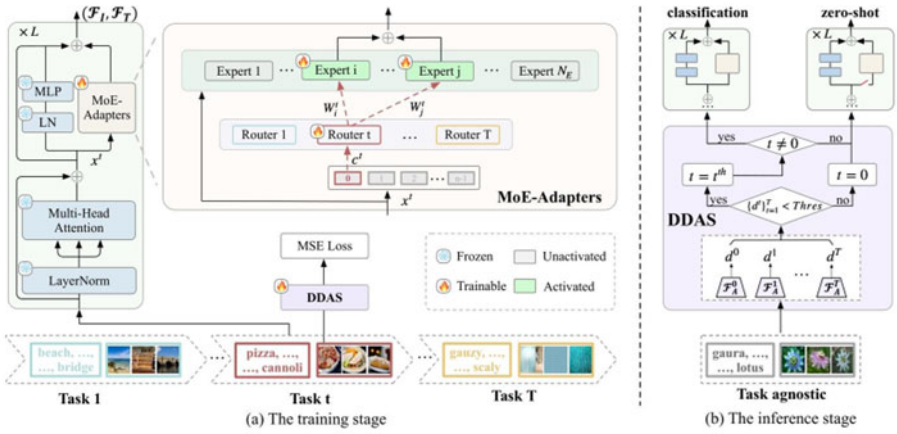


图 4 MoE-Adapters 网络结构

MoE-Adapters++作为 MoE-Adapters 的改进版本，进一步提升了多域混合任

务的适配效率和灵活性。[4]其核心创新在于引入了动态 MoE 适配器和潜在嵌入自动选择器（LEAS）。动态 MoE 适配器通过动态专家扩展控制器（DEeC），根据损失阈值自动判断是否需要为新领域添加新专家，避免了静态专家集导致的参数冗余或适配不足问题；LEAS 则将分布选择功能融入预训练模型的早期层，替代了 MoE-Adapters 中独立的 DDAS 模块，使网络结构更加统一，同时减少了训练复杂度。在推理时，LEAS 通过计算输入数据在各类编码器上的相关性，根据阈值自动确定使用的专家数量，进一步降低了计算成本。MoE-Adapters++网络结构如图 5 所示。MoE-Adapters++在持续学习场景中表现出显著优势，能够在不断学习新领域任务的同时，有效缓解灾难性遗忘，且参数数量和 GPU 内存占用大幅低于 MoE-Adapters。

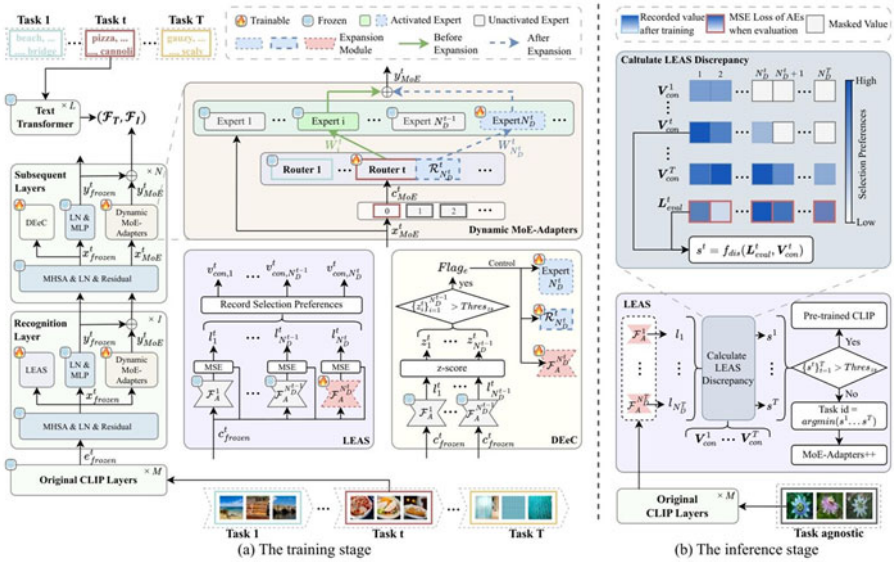


图 5 MoE-Adapters++网络结构

3.2 矩阵级优化

矩阵级优化的核心思路是对 LoRA 中的低秩矩阵结构进行改进，或对专家权重矩阵的结构进行重构，通过优化矩阵的表示形式进一步提升参数效率和推理速度。与网络级优化侧重于专家网络的整体架构不同，矩阵级优化更关注单个权重矩阵或低秩组件的内部结构设计，通过引入张量分解、结构化矩阵等技术，在保持适配性能的同时，进一步减少可训练参数数量或降低计算复杂度。

TT-LoRA（Tensor Train Low-Rank Approximation）是一种面向文本任务的单任务微调方法，其核心创新在于将 LoRA 中的两个二维低秩矩阵替换为多个三维低秩张量核（tensor train cores）结构。[5]张量核结构通过结合乘积与求和运算的张量收缩方法，避免了计算完整的权重矩阵，从而在进一步减小参数量的同时，降低了推理延迟。在实现上，TT-LoRA 将原始的低秩矩阵分解转化为张量列分解，每个张量核的维度远小于原始低秩矩阵，使得可训练参数数量进一步减少。TT-LoRA 网络结构如图 6 所示。实验结果表明，TT-LoRA 在保持

与 LoRA 相当的文本任务性能的同时，参数量减少了约 30%，推理速度提升了 25%以上，尤其适用于对延迟要求较高的部署场景。

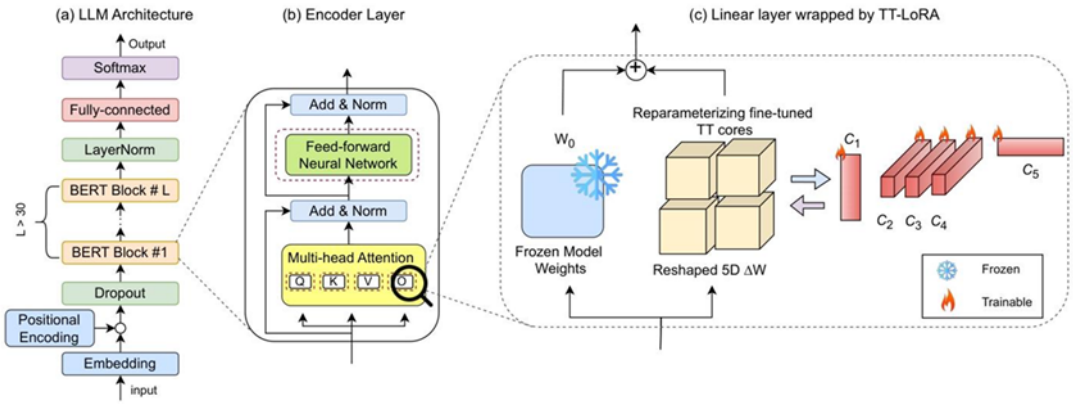


图 6 TT-LoRA 网络结构

TT-LoRA MoE 是基于 TT-LoRA 的多域混合方法，主要针对文本多域任务。[6]其核心设计是为每类领域数据训练一个 TT-LoRA 专家，然后联合训练带噪 Top-1 门控网络。为了避免专家选择不平衡问题，TT-LoRA MoE 在损失函数中引入了路由器交叉熵损失，引导门控网络均匀地选择各个专家，确保每个专家都能得到充分训练。在推理时，门控网络根据输入文本的特征选择最合适的 TT-LoRA 专家进行适配，由于每个专家采用张量核结构，模型的整体算力消耗显著低于基于传统 LoRA 专家的 MoE 架构。TT-LoRA MoE 网络结构如图 7 所示。TT-LoRA MoE 在多域文本分类、问答等任务中表现出优异的性能，既保持了多域适配的灵活性，又兼顾了参数效率和推理速度。

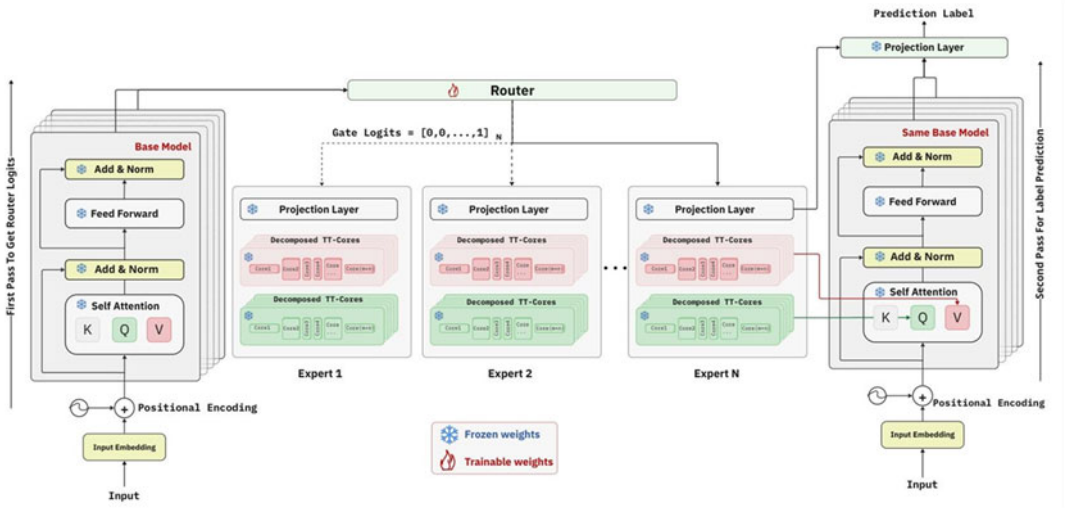


图 7 TT-LoRA MoE 网络结构

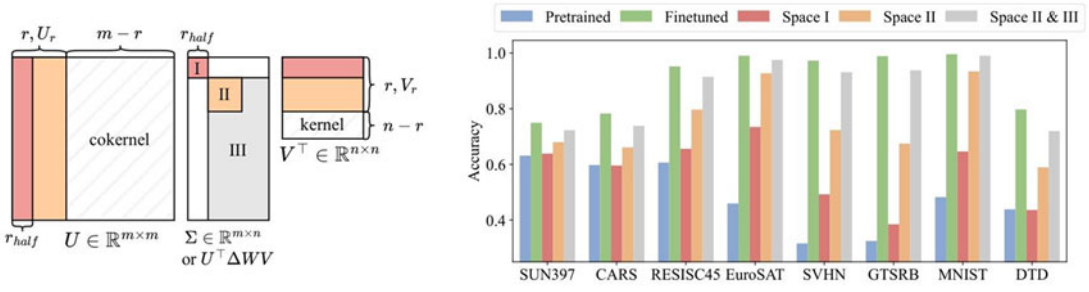
矩阵级优化方法的核心优势在于通过精细化的矩阵结构设计，在不牺牲适配性能的前提下，进一步提升了参数效率和计算效率。这类方法通常不需要改变模型的整体架构，仅对低秩适配器的内部结构进行调整，具有良好的兼容性

和可扩展性，能够轻松集成到现有的参数高效微调框架中。

3.3 维度级优化

维度级优化的核心思想是对预训练模型权重矩阵的不同维度或子空间进行划分，将任务适配过程限制在特定的子空间内，通过分离通用知识和任务特定知识，实现知识保留与高效适配的平衡。维度级优化的理论基础是：预训练模型的权重矩阵在不同维度上承载的知识不同，部分维度主要承载通用知识，而另一些维度则更容易学习任务特定知识。通过在合适的子空间内进行优化，可以在学习任务知识的同时，避免对通用知识的破坏。

SMILE（Sparse Mixture of Low-Rank Experts）是一种面向多域混合任务的维度级优化方法，其核心贡献在于通过实验验证了模型微调在重要维度上的保留性和小维度上的学习性，为维度级优化的可行性提供了理论支撑。[7] SMILE 通过在多个任务上将微调权重投影到模型权重矩阵的不同子空间，发现预训练模型的权重矩阵可分为三个子空间：Top-50%奇异值对应的 I 空间（主要承载通用知识）、剩余奇异值对应的 II 空间（可学习任务特定知识）以及奇异值为 0 的 III 空间（冗余空间）。SMILE 维度投影实验结果如图 8 所示。基于这一发现，SMILE 将各类数据的 LoRA 微调权重投影到预训练矩阵奇异值较小的 II/III 空间的不同维度上，在推理时根据输入计算不同专家上的 L2 范数来激活 Top-k_gate 专家，从而减小各类任务之间的参数干扰。SMILE 网络结构如图 9 所示。这种子空间划分策略使 SMILE 能够在多个领域任务上同时进行适配，且不会出现严重的参数干扰问题，在多域文本分类和问答任务中表现出优于传统 LoRA 的泛化能力。



(a) Full SVD and subspace partition.　(b) Accuracy comparison across different subspace projection strategies.

图 8 SMILE 维度投影实验结果

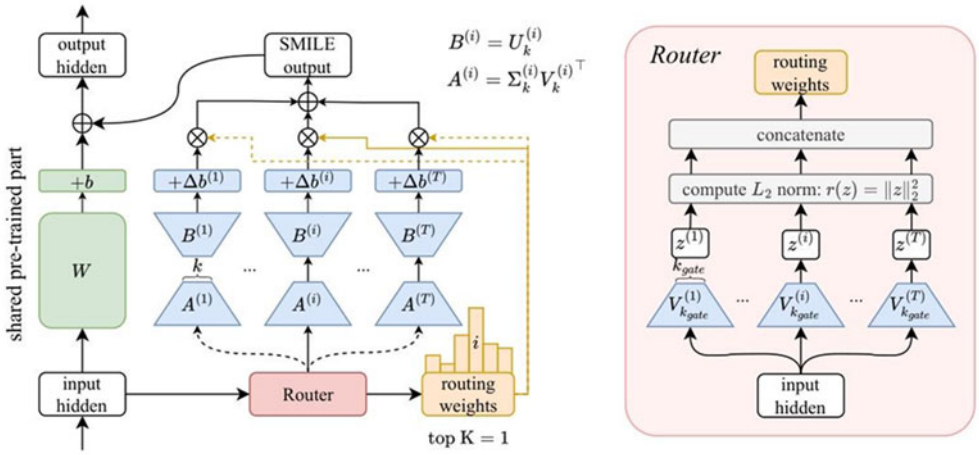


图 9 SMILE 网络结构

SDS LoRA（Shared and Domain-Specific LoRAs）是一种针对图像分类任务的多域混合方法，其核心创新在于通过奇异值分解（SVD）对参数矩阵进行划分，明确分离共享维度和领域特定维度。[8]具体而言，SDS LoRA 首先对预训练权重矩阵进行 SVD 分解，按奇异值大小将矩阵分为两部分：较大奇异值对应的子空间用于学习所有领域共享的通用知识，采用 LoRA 微调所有类别的数据；较小奇异值对应的子空间用于学习每个领域的特定知识，为每类数据单独微调一个 LoRA 专家。SDS LoRA 网络结构如图 10 所示。这种划分方式确保了共享信息的泛化性和特定信息的独立性，避免了不同领域之间的知识干扰。在实验中，SDS LoRA 在 UCF101、Kinetics400、HMDB51 等多个动作识别数据集上进行多域混合训练，表现出有前景的分类性能，同时参数干扰程度显著降低。

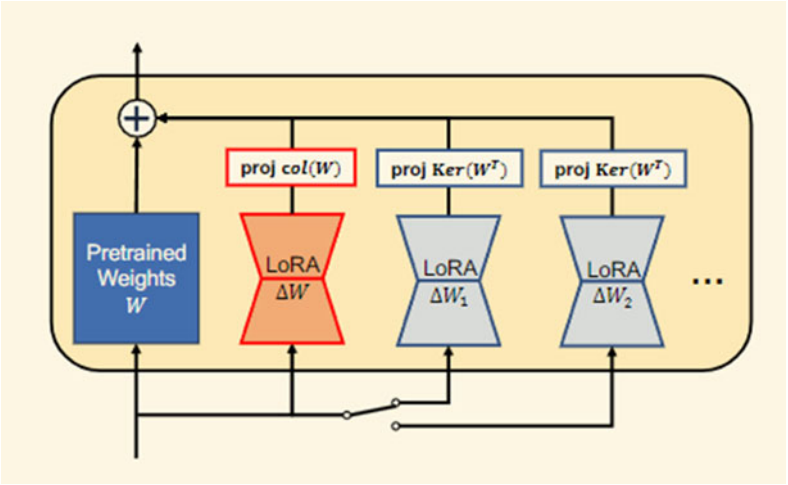


图 10 SDS LoRA 网络结构

CorDA（Context-oriented Decomposition Adaptation）是一种面向文本任务的单任务微调方法，其核心创新在于提出了面向上下文的奇异值分解（CO-SVD），通过结合数据上下文信息初始化 LoRA 适配器，实现任务感知的参数高效微调。[9]传统 LoRA 的适配器初始化采用高斯分布和零初始化，与下游任务无

关，导致适配效率较低且容易遗忘预训练知识。CorDA 通过从目标任务中采样少量数据，计算每个线性层输入激活的协方差矩阵，然后对权重矩阵与协方差矩阵的乘积进行 SVD 分解，使分解后的主成分能够捕捉任务特定的上下文信息。CO-SVD 方法原理如图 11 所示。基于 CO-SVD，CorDA 设计了两种适配模式：知识保留模式（KPM）和指令预览模式（IPM）。在 KPM 中，CorDA 冻结主成分对应的权重组件以保留预训练知识，使用最小的 r 个奇异值和向量初始化可训练适配器，适用于需要保留通用知识的场景；在 IPM 中，CorDA 使用最大的 r 个奇异值和向量初始化适配器，使适配器能够快速捕捉任务核心特征，实现更快的收敛。

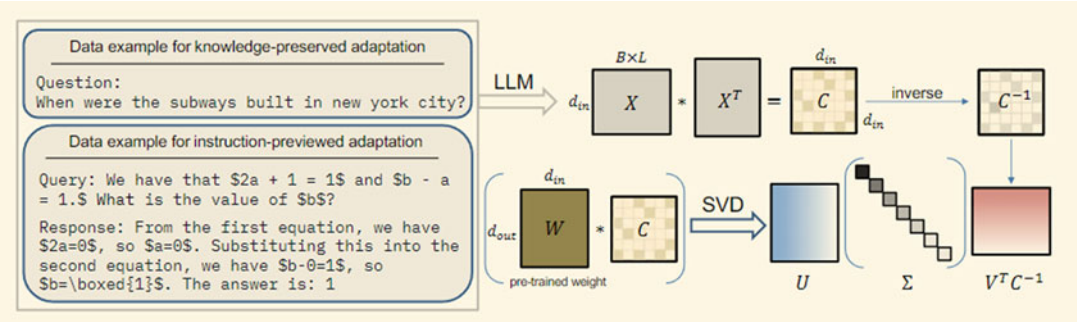


图 11 CorDA 方法原理

CorDA++作为 CorDA 的改进版本，进一步提升了任务感知适配的性能和稳定性。其核心改进包括两个动态策略：动态协方差选择和动态秩分配。动态协方差选择通过多次数据采样构建协方差矩阵候选池，为每个线性层选择最具代表性的协方差矩阵，减少随机采样带来的变异性；动态秩分配基于一个反映任务特定主成分紧凑性的度量指标，通过迭代过滤过程为不同层分配合适的秩，使对任务更敏感的层获得更多的参数预算。CorDA++在 KPM 模式下不仅在 Math、Code、指令跟随等任务上取得了优于 LoRA 的性能，还显著缓解了大语言模型和视觉语言模型的预训练知识遗忘；在 IPM 模式下，收敛速度较 QLoRA 提升了 4.5 倍，在各类场景中均优于强基线方法。此外，CorDA++还支持与量化技术结合，将冻结的权重组件量化为 4 位精度，而可训练适配器保持全精度，在进一步降低内存开销的同时，信息损失极小。

4 应用场景

4.1 文本任务适配

文本任务是参数高效微调技术应用最为广泛的场景，涵盖机器翻译、文本分类、问答系统、代码生成、指令跟随等多个子任务。在这些任务中，参数高效微调方法能够在大幅降低计算成本的同时，保持甚至超越全量微调的性能。例如，LoRA 在 LLaMA、GPT 等模型上的微调结果表明，仅优化 0.1%-1%的参数，就能在机器翻译任务上达到全量微调 95%以上的性能，且训练时间缩短至约 1/10。CorDA++在 Math 任务上的表现尤为突出，在 KPM 模式下，其在 GSM8k 数

据集上的准确率达到 45.13%，高于 LoRA 的 42.68%，同时在 NQ open 等常识问答数据集上的性能下降幅度仅为 2.97%，远低于全量微调的 11.86%，有效缓解了知识遗忘。

在多域文本任务中，SMILE 表现出显著优势。在来自 GLUE 基准的 8 类任务上，SMILE 在 LoRA 微调场景下的平均得分为 84.0，达到单任务平均性能的 99.0%，比传统权重平均方法提高 6.9%，充分验证了其在多域文本任务中对通用知识与任务特定知识的分离能力。在代码生成任务中，CorDA++在 HumanEval 数据集上的 Pass@1 分数达到 19.23%，高于 LoRA 的 16.80%，同时在 NQ open 等常识数据集上的性能保持稳定，实现了代码生成能力与通用知识的平衡。

4.2 多模态任务适配

随着多模态技术的发展，参数高效微调方法逐渐扩展到文本-图像、文本-图像-三维姿态等多模态任务中，解决了多模态适配中的模态差距和知识遗忘问题。HMVLM 作为面向文本-图像-三维姿态任务的多模态适配方法，通过 MoE LoRA 框架和零专家设计，在文本到运动生成任务上的 R-precision 达到 0.502，高于 MotionAgent 的 0.482，同时在 MT-Bench 对话任务上的性能下降仅为 0.26，远低于 MotionAgent 的 6.79，有效保留了模型的语言对话能力。

在文本-图像理解任务中，MoE-Adapters++和 CorDA++表现出优异的性能。MoE-Adapters++在持续学习场景中，经过多个图像分类任务的训练后，Zero-shot 性能保持在 76.3%，高于 ZSCL 的 67.4%，同时参数数量仅为 ZSCL 的约 1/140；CorDA++将视觉语言模型 LLaVA-1.5 在 OKVQA 数据集上进行微调后，在 VQAv2、GQA 等所有 Zero-shot 基准上的性能下降幅度均小于 LoRA 和全量微调，验证了其在多模态知识保留方面的有效性。

4.3 持续学习适配

持续学习是参数高效微调的重要应用场景之一，其核心挑战是模型在不断学习新任务的过程中，如何避免遗忘已学知识。MoE-Adapters++通过动态专家扩展和 LEAS 模块，在持续学习 11 个视觉-语言任务后，“Transfer”、“Average”、“Last”三个指标分别达到 69.0%、77.5%、86.2%，均优于 DIKI、ZSCL 等方法，同时参数数量仅为 1.1M，远低于 ZSCL 的 149.6M。

在持续学习中，参数高效微调方法的核心优势在于其模块化设计，新任务的适配仅通过添加或更新少量专家或适配器模块，不会对已训练的模块造成破坏，从而实现知识的增量学习。此外，门控网络和子空间划分技术的应用，进一步确保了新任务知识与旧任务知识的隔离，避免了参数干扰导致的遗忘。

5 挑战与未来方向

5.1 当前挑战

尽管参数高效微调技术取得了显著进展，但在实际应用中仍面临多个关键挑战：

首先，参数干扰问题依然存在。在多域混合任务和持续学习场景中，不同任务的适配器或专家之间仍可能存在参数干扰，导致部分任务的性能下降。虽然 SDS LoRA、SMILE 等方法通过子空间划分缓解了这一问题，但如何实现更精准的知识隔离，尤其是在跨模态任务中，仍是需要解决的关键问题。

其次，动态适配能力不足。现有方法大多需要人工预设专家数量或秩的大小，缺乏根据任务特性自动调整模型结构的能力。虽然 MoE-Adapters++和 CorDA++引入了动态策略，但在复杂任务分布下，动态决策的准确性和效率仍有待提升，如何实现完全自适应的参数高效微调仍是一个开放问题。

第三，跨模态适配存在瓶颈。多模态任务中不同模态的特征分布差异较大，现有参数高效微调方法大多针对单一模态或特定模态组合设计，缺乏统一的跨模态适配框架，导致在新的模态组合任务上的泛化能力有限。

第四，可解释性较差。大多数参数高效微调方法的适配过程是黑箱操作，难以解释适配器或专家究竟学习了哪些任务特定知识，以及如何与预训练知识相互作用。可解释性的缺乏限制了方法的进一步优化和改进，也不利于在高可靠性要求的场景中应用。

第五，极端资源受限场景的适配。在边缘设备等极端资源受限场景中，现有方法的参数效率和计算效率仍需提升，如何在仅具备有限内存和计算能力的设备上实现大规模模型的高效微调，是当前研究面临的实际挑战。

5.2 未来方向

针对上述挑战，未来参数高效微调技术的研究可围绕以下方向展开：

一是发展更精准的知识隔离技术。通过结合注意力机制、掩码策略等，进一步细化子空间划分，实现任务特定知识与通用知识的精准分离；探索基于因果推断的方法，消除不同任务之间的虚假关联，减少参数干扰。

二是构建完全自适应的动态适配框架。利用强化学习、元学习等技术，使模型能够根据任务的复杂度、数据分布等自动调整专家数量、秩的大小、子空间划分方式等，实现无需人工干预的自适应微调。

三是建立跨模态统一适配框架。深入研究不同模态特征的共性与差异，设计通用的适配器结构和优化策略，实现跨文本、图像、音频、视频等多模态任务的统一适配，提升模型的跨模态泛化能力。

四是增强方法的可解释性。通过可视化技术、特征归因分析等，揭示适配器和专家的工作机制，解释任务特定知识的学习过程；设计可解释的子空间划分和专家分配策略，使适配过程更加透明。

五是优化极端资源受限场景的适配方案。结合量化、剪枝、蒸馏等技术，进一步压缩可训练参数和计算量；探索轻量化的适配器结构和高效的优化算法，使参数高效微调能够在边缘设备等资源受限场景中广泛应用。

六是拓展应用场景。将参数高效微调技术应用于更广泛的领域，如自动驾驶、医疗诊断、工业检测等，解决这些领域中大规模模型的适配问题，推动技术的实际落地。

6 结论

大语言模型参数高效微调技术作为解决全量微调算力和内存瓶颈的核心方案，近年来取得了飞速发展，形成了以网络级、矩阵级、维度级优化为三大核心的技术体系。从 LoRA 的提出到后续各类扩展方法的涌现，参数高效微调技术不断提升参数效率、适配性能和知识保留能力，在文本任务、多模态任务、持续学习等多个场景中展现出巨大的应用价值。

网络级优化通过构建 MoE 架构，实现了任务与专家的动态匹配，提升了多任务和多域适配的灵活性；矩阵级优化通过改进权重矩阵结构，进一步提升了参数效率和计算效率；维度级优化通过子空间划分，实现了通用知识与任务特定知识的有效分离，缓解了灾难性遗忘。这些方法从不同角度解决了参数高效微调的核心问题，共同推动了技术的发展。

然而，当前参数高效微调技术仍面临参数干扰、动态适配不足、跨模态适配瓶颈、可解释性差等挑战。未来，通过发展更精准的知识隔离技术、构建完全自适应的动态适配框架、建立跨模态统一适配框架、增强可解释性、优化极端资源受限场景的适配方案等，参数高效微调技术将进一步提升性能和适用性，为大规模模型的广泛应用提供更强大的支撑。

随着大语言模型的持续发展，参数高效微调技术将在人工智能领域发挥越来越重要的作用，不仅能够降低大规模模型的应用门槛，还能推动人工智能技术在更多实际场景中的落地，为各行各业带来新的变革。

参考文献

[1] Hu E J, Shen Y, Wallis P, et al. Lora: Low-rank adaptation of large language models[J]. ICLR, 2022, 1(2): 3.

[2] Hu L, Ye Y, Xia S. HMVLM: Human Motion-Vision-Lanuage Model via MoE LoRA[J]. arXiv preprint arXiv:2511.01463, 2025.

[3] Bertolissi R, Hübotter J, Hakimi I, et al. Local mixtures of experts: Essentially free test-time training via model merging[J]. arXiv preprint arXiv:2505.14136, 2025.

[4] Yu J, Huang Z, Zhuge Y, et al. MoE-Adapters++: Towards More Efficient Continual Learning of Vision-Language Models via Dynamic Mixture-of-Experts Adapters[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2025.

[5] Anjum A, Eren M E, Boureima I, et al. Tensor train low-rank approximation (tt-lora): Democratizing ai with accelerated llms[C]//2024 International Conference on Machine Learning and Applications (ICMLA). IEEE, 2024: 583-590.

[6] Kunwar P, Vu M N, Gupta M, et al. TT-LoRA MoE: Using Parameter-Efficient Fine-Tuning and Sparse Mixture-Of-Experts[C]//Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis. 2025: 1332-1350.

[7] Tang A, Shen L, Luo Y, et al. Zero-Shot Sparse Mixture of Low-Rank Experts Construction From Pre-Trained Foundation Models[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2025.

[8] Takama Y, Ding N, Yokota T, et al. Separating Shared and Domain-Specific LoRAs for Multi-Domain Learning[C]//Proceedings of the Computer Vision and Pattern Recognition Conference. 2025: 6429-6437.

[9] Yang Y, Liu S, Rao C, et al. Dynamic Context-oriented Decomposition for Task-aware Low-rank Adaptation with Less Forgetting and Faster Convergence[J]. arXiv preprint arXiv:2506.13187, 2025.