自动化学报
ACTA AUTOMATICA SINICA

大模型参数高效微调方法综述: 技术、趋势与挑战

唐岸达 林宙辰

A Survey on Parameter-efficient Fine-tuning of Large Models: Techniques, Trends, and Challenges

TANG An-Da, LIN Zhou-Chen

在线阅读 View online: https://doi.org/10.16383/j.aas.c250451

大模型参数高效微调方法综述: 技术、趋势与挑战

唐岸达 [1] 林宙辰 [1, 2, 3, 4]

摘 要 大规模预训练模型在自然语言处理等多个领域表现优异. 为更好地适配下游任务, 微调预训练模型是一个常用的方法, 但全量微调面临高昂计算与存储成本. 参数高效微调 (PEFT) 通过仅更新极少量参数, 在降低开销的同时保持模型性能. 该综述系统梳理 PEFT 领域的主流方法. 首先, 将现有方法归纳为添加式、局部式、重参数化式和融合式四大范式, 并深入剖析各类方法的核心机理、性能特征、应用场景及策略优势. 进而, 探讨 PEFT 的技术演进, 总结出 PEFT 方法从单一方法创新走向存储、计算、性能的三元权衡, 并向自动化、智能化、软硬件协同等统一框架发展. 更进一步, 该综述对各类代表性 PEFT 方法进行性能与参数效率的定量比较. 此外, 本综述还涵盖 PEFT 在视觉、语音及跨模态模型等领域的拓展应用. 最后, 总结并探讨未来研究方向, 以推动更高效、更适应多样化任务的大模型微调技术的发展.

关键词 大语言模型; 大规模预训练模型; 参数高效微调; 机器学习; 深度学习

引用格式 唐岸达, 林宙辰. 大模型参数高效微调方法综述: 技术、趋势与挑战. 自动化学报, 2026, **52**(8): 1520−1556

DOI 10.16383/j.aas.c250451 **CSTR** 32138.14.j.aas.c250451

A Survey on Parameter-efficient Fine-tuning of Large Models: Techniques, Trends, and Challenges

TANG An-Da[1] LIN Zhou-Chen[1, 2, 3, 4]

Abstract Large-scale pre-trained models have demonstrated remarkable capabilities in various fields such as natural language processing. To better adapt these models to downstream tasks, fine-tuning these models is a commonly adopted approach. However, full fine-tuning faces severe challenges, including high computational costs and substantial storage requirements. Parameter-efficient fine-tuning (PEFT) addresses these issues by updating only a minimal number of parameters, thereby reducing overhead while preserving model performance. This survey provides a systematic review of the mainstream methodologies in the PEFT field. First, we categorize existing methods into four major paradigms: Additive, selective, reparameterized, and hybrid, offering an in-depth analysis of their core mechanisms, performance characteristics, application scenarios, and strategic advantages. Furthermore, we explore the technical evolution of PEFT, summarizing a clear trajectory. The field is moving from isolated method innovations towards a tri-balance among storage, computation, and performance, and further advancing into unified frameworks emphasizing automation, intelligence, and hardware-software co-design. Moreover, we conduct a quantitative comparison of representative PEFT methods in terms of performance and parameter efficiency. In addition, this survey also covers the extended applications of PEFT in fields such as vision, speech, and cross-modal models. Finally, we summarize and discuss promising future research directions to facilitate the development of more efficient and adaptable fine-tuning techniques for large-scale models across diverse tasks.

Keywords large language models; large-scale pre-trained models; parameter-efficient fine-tuning; machine learning; deep learning

Citation Tang An-Da, Lin Zhou-Chen. A survey on parameter-efficient fine-tuning of large models: Techniques, trends, and challenges. *Acta Automatica Sinica*, 2026, **52**(8): 1520−1556

近年来, 大规模预训练语言模型 (large-scale pre-trained language models, LLMs) 成为人工智能领域的核心技术突破. 凭借其强大的计算能力和海量数据训练, 这些模型在自然语言处理 (NLP)、计算机视觉 (CV) 以及多模态任务中展现出卓越的通用性和适应性, 例如文本生成[1]、代码生成[2]、机器翻译[3]、个性化聊天机器人[4−5]、医疗[6−7]、法律[8] 和金融[9] 等任务, 推动了智能系统的广泛应用.

收稿日期 2025-09-04 录用日期 2026-01-19
Manuscript received September 4, 2025; accepted January 19, 2026
国家自然科学基金 (62276004), 北京市自然科学基金 (L257007) 资助
Supported by National Natural Science Foundation of China (62276004) and Beijing Natural Science Foundation (L257007)
本文责任编委 鲁继文
Recommended by Associate Editor LU Ji-Wen
1. 北京大学智能学院 北京 100871 2. 北京大学跨媒体通用人工智能全国重点实验室 北京 100871 3. 北京大学人工智能研究院 北京 100871 4. 琶洲实验室 (黄埔) 广州 510335
1. School of Intelligence Science and Technology, Peking University, Beijing 100871 2. State Key Laboratory of General AI, Peking University, Beijing 100871 3. Institute for Artificial Intelligence, Peking University, Beijing 100871 4. Pazhou Laboratory (Huangpu), Guangzhou 510335

大模型在预训练阶段已具备强大的通用表征能力和丰富的世界知识, 并为下游任务提供卓越的起点. 然而, 直接将其应用于特定下游任务时, 性能往往难以达到最优. 这一现象主要源于预训练任务与目标任务的分布差异. 为解决这个问题, 迁移学习中的微调 (fine-tuning) 方法通过调整大型预训练模型的参数, 可以显著提高其在执行特定任务时的有效性, 适应目标任务的独特特征和要求. 不过, 随着模型规模攀升至百亿甚至万亿参数, 由于传统的全量微调 (full fine-tuning) 方法需要更新所有模型权重, 这不仅对计算资源和内存提出极高的要求, 导致高昂的训练成本, 同时限制了其在资源受限场景下的适用性, 还极易引发灾难性遗忘, 损害其原有的通用性能[10].

为克服这些挑战, 参数高效微调 (parameter-efficient fine-tuning, PEFT) 方法应运而生. PEFT 的核心思想是调整少量的、关键的 LLM 参数, 在保持预训练模型主体结构不变的同时, 高效适配下游任务. 这一策略不仅显著降低了计算和存储开销, 还在一些任务中达到与全量微调相当甚至更优的性能. PEFT 方法关注于如何设计更高效的参数更新策略, 例如在存储[11−12] 或计算效率[13] 方面, 同时, 还侧重优化效率与泛化能力的平衡, 例如在推理开销[14] 和模型表现[13, 15] 方面. 除此之外, 有的工作思考跨任务与跨模态迁移, 拓展了对应用场景的探索, 除自然语言处理场景, 还涵盖视觉场景; 有的工作关注 PEFT 技术在各种模型架构和下游任务中的应用; 有的工作开展 PEFT 的实际部署研究. 进一步证明了其广泛适用性[16].

尽管 PEFT 方法呈现出多样性, 但其核心设计可统一理解为: 通过引入或找寻带约束的参数或空间变换, 将全量微调所探索的原始高维参数空间, 映射到一个低维、紧凑且富有表示能力的子空间中进行优化. 围绕这个核心思想, PEFT 方法的构建策略主要可归纳为: 添加式微调 (拓宽参数空间)、局部微调 (选择模型内部参数子空间)、重参数化微调 (对参数空间等效变换进行优化) 以及融合微调 (合并多个微调方法). 通过对这四类范式的剖析, 本文进一步指出, 该领域的发展已从早期的范式创新演进至当前对存储、计算、性能三元权衡的精细化设计, 而其未来趋势必将从单一的范式创新迈向可控、可解释、跨模态、多任务且自动化的方法, 以实现对预训练模型内在结构的充分利用.

基于上述内容, 本文系统性地全面回顾大模型高效微调技术的进展. 从多个角度层层递进地对这些工作进行阐述, 并比较分析理论基础、应用场景及策略优势, 探讨未来研究方向, 以推动更高效、更适应多样化任务的大模型微调技术的发展. 本文的整体结构如下: 第 1 节主要介绍 LLM 的基本概念和微调相关的基本知识; 第 2 节从方法论角度对 PEFT 方法进行分类, 对每一类微调方法的关键技术进行介绍, 探讨技术演进和总结发展趋势, 并对代表性方法的性能进行定量比较分析, 在统一的模型与数据集上评估其性能与参数效率; 第 3 节介绍 PEFT 技术在视觉、语音及多模态模型等领域的拓展应用, 展现其广泛的适用性; 第 4 节对全文进行总结并展望未来.

1 大模型高效微调背景知识

1.1 Transformer 架构

Transformer 架构[17] 是预训练大模型的基础架构, 由多层编码器或解码器堆叠而成, 编码器和解码器均包含多头自注意力层 (multi-head self-attention) 与前馈层, 并通过残差连接和层归一化相连. 残差连接有助于信息在层间有效传递, 避免梯度消失; 层归一化则通过对每层输入进行标准化来提升训练稳定性. 多头自注意力层并行执行自注意力头. 对于 h 个自注意力头的自注意力层, 输入向量 $X \in \mathbf{R}^{d_{input}}$, 其中 d_{input} 为输入向量的维度. 通过可学习的权重矩阵生成的查询向量 Q、键向量 K 和值向量 V 可被表示为:

$$Q = XW_q + b_q, \quad K = XW_k + b_k, \quad V = XW_v + b_v \tag{1}$$

其中, W_q, $W_k \in \mathbf{R}^{d_{input} \times d_k}$; $W_v \in \mathbf{R}^{d_{input} \times d_v}$; d_k 为查询向量和键向量的维度; d_v 为值向量的维度; b_k, $b_q \in \mathbf{R}^{d_k}$, $b_v \in \mathbf{R}^{d_v}$ 为偏置向量. 自注意力输出计算如下:

$$\text{attention}(Q,\ K,\ V) = \text{softmax}\left(\frac{QK^{\mathrm{T}}}{\sqrt{d_k}}\right)V \tag{2}$$

其中, attention 代表注意力机制. 多头自注意力机制将多个注意力头的结果拼接并线性变换:

$$\text{MHA}(Q,\ K,\ V) = \text{concat}(head_1,\ \cdots,\ head_h)W^O \tag{3}$$

其中, MHA 代表多头注意力机制; concat 代表拼接操作; W^O 是对拼接后的矩阵做线性变换的权重矩阵; $head_i = \text{attention}(XW_q^i,\ XW_k^i,\ XW_v^i)$, 每个头有自己独立的投影矩阵 W_q^i, W_k^i, W_v^i.

前馈层由两个线性变换及其间的 ReLU 激活函数构成:

$$\mathrm{FFN}(X) = \mathrm{ReLU}(XW_1 + b_1)W_2 + b_2 \tag{4}$$

其中, FFN 代表前馈层对输入 X 的操作; W_1, b_1, W_2, b_2 为可学习参数.

1.2 预训练大语言模型

传统自然语言处理主要依赖于为如情感分析等特定任务设计和训练的端到端的模型. 然而, 随着 Transformer 架构的提出及其在深层上下文建模上的卓越表现, 预训练与微调这种新的训练范式应运而生并迅速成为主流.

该范式的核心在于, 模型首先在大量文本语料库[18] 上进行预训练, 学习语言的通用表征和世界知识. 这一过程旨在构建一个强大的基座模型. 随后, 针对下游具体任务, 只需使用少量标注数据对该预训练模型进行微调, 即可使其高效适配. 这一转变极大地提升了语言模型的可扩展性与复用性, 直接催生了模型参数规模的指数级增长. 这也就是一般所指的预训练大语言模型.

预训练大语言模型参数量巨大, 例如, OpenAI 的 GPT-1 拥有 1.2 亿个参数, 而 GPT-2 则拥有 15 亿个参数. 这个数字在 GPT-3 中激增至 1 750 亿, 而在最新的 GPT-4 中飙升到 1.76 万亿[19].

当模型参数持续扩大并跨越某个临界点后, 大语言模型开始展现出在小规模模型中几乎不存在的一系列能力, 也就是所谓的“涌现能力”[20], 主要包括以下三个方面.

上下文学习 (in-context learning). 模型能够根据上下文中的少量示例学习新任务, 无需梯度更新即可理解并执行新任务, 这使大模型具备类似举一反三的类比推理能力.

指令遵循 (instruction following). 通过对涵盖广泛任务的指令数据集进行微调, 模型可泛化至未见过的任务指令[21], 极大地提升了其通用性和交互性.

逐步推理 (step-by-step reasoning). 大规模参数赋予思维链 (chain-of-thought, CoT) 能力, 可处理需多步推理的复杂任务, 提升了在数学、常识推理等需要多步分析的任务上的表现.

尽管 LLMs 在多个领域展现出强大能力, 但是仍存在三大局限: 模型与任务间表现不一致、幻觉与事实性错误以及深层语义理解的不足. 想要模型表现出良好的能力, 训练方式也是很重要的一环. 目前大模型的训练与对齐方法主要包括以下三方面.

指令微调 (instruction tuning). 通过将多样任务转化为指令–应答格式, 提升模型泛化能力.

对齐调优与人类反馈强化学习 (RLHF). 包括三阶段流程: 有监督微调 (SFT)、奖励模型训练和强化学习优化 (如 PPO[22]、DPO[23]). 也可以通过 AI 自动标注替代部分人工标注[24].

数据集构建. 涵盖预训练、有监督微调和 RLHF 三个阶段, 数据质量与多样性对模型性能具有关键影响. 大模型的训练数据生态涵盖预训练语料、有监督微调数据和人类偏好数据三个层面. 数据的质量、多样性、规模以及对安全性和偏差的控制, 直接决定了最终模型性能的上限和伦理边界.

1.3 主流大模型

基础模型主要在大规模数据集上进行预训练, 并可以通过微调来适应各种下游任务. 从模型发展、架构设计、技术路线等角度出发, 目前比较常见的大模型包括但不限于以下模型系列.

GPT 系列. 由 OpenAI 推出, 包括 GPT-1 至 GPT-4 等多个版本[19, 21, 25–27]. GPT-3 被视为首个真正意义上的大语言模型, 展现出较强的上下文学习能力; GPT-4 具有高达 1.8 万亿个参数, GPT-4 及 GPT-4o 支持多模态输入, 在图像理解与推理方面表现突出.

LLaMA 系列. Meta 推出的开源模型, 涵盖 1 B 至 405 B 参数规模[28–29]. LLaMA-1 虽参数较少但性能优异; LLaMA-2 训练数据增加 40%; LLaMA-3 及后续版本进一步提升文本与多模态处理能力.

DeepSeek 系列. 深度求索公司研发, 包括 DeepSeek-V1/V2/V3 及 R1 模型[30–33]. 其采用混合专家模型 (mixture of experts, MoE)、多头潜在注意力 (multi-head latent attention, MLA)、多词元预测 (multi-token prediction, MTP) 和群体相对策略优化 (group relative policy optimization, GRPO) 等创新技术, 在性能与效率间实现良好平衡.

Qwen 通义千问系列. 由阿里云推出的一系列大规模语言模型[34–36], 以其全面开源和强大的综合性能而著称. 该系列模型参数规模覆盖 14 B 至 480 B 的范围; 可以快速响应, 降低交互延迟; 具有高吞吐量, 支持多任务并行处理; 提供语言、语音、视觉等多模态模型; 均采用宽松的开源协议, 极大地促进了其在学术研究和工业中的广泛采用与二次开发.

Claude 系列. Anthropic 开发的对话模型, 强调安全性、有用性与无害性[37]. Claude 3 包含 Opus、Sonnet 和 Haiku 三个版本, 分别适用于复杂推理、高效响应和创意生成.

Gemini 系列. 谷歌 DeepMind 推出的多模态模型[38], 原生支持文本、图像、音频与视频处理. Gemini 1.5 具备 100 万 token 上下文窗口, Gem-

ini 2 进一步优化推理速度与工具集成.

OMLo 系列. 由艾伦人工智能研究所 (AI2) 联合 AMD、芬兰 LUMI 超算等机构共同研发的开放语言模型框架, 其核心特征为完整开源[39]. 它结合了许多大模型的优点, 具有优异的能力. 同时, 它开放的训练数据、评估代码、中间模型检查点和训练日志, 为使用和进一步研究提供了非常高的透明度和可重复性.

在大模型应用领域方面, 大语言模型不仅在语言任务方面表现优异, 还可以拓展到其他领域. 例如, 多模态大语言模型扩展了传统 LLMs 的处理范围, 可同时接收文本、图像、音频等多种输入. 其典型结构包含三个模块: 多模态编码器 (如 CLIP[40]、EVA-CLIP[41])、大语言模型 (预训练 LLM) 和模态连接器. 该连接器将非文本信息映射至 LLM 可理解的表示空间, 从而实现跨模态推理与生成.

从模型处理的模态、信息类型和具备的核心能力出发, 目前大模型的分类包括但不限于:

大规模语言模型 (large-scale language models, LLMs). 专攻文本的理解、生成与处理. 它们通过海量文本语料训练, 能够胜任翻译、摘要、文本生成及问答等一系列语言相关任务.

视觉基础模型 (vision foundation models, VFMs). 其核心在于对图像等视觉数据的理解与信息提取. 此类模型经过大规模图像数据预训练, 能够出色地泛化至图像分类、目标检测、分割等多种视觉任务.

语音基础模型 (speech foundation models, SFMs). 通过自监督学习从海量无标注数据中学习通用的语音表征. 此类模型经过大量、多样化的语音数据预训练, 能够出色地泛化至语音情感识别、语音内容理解等多种任务.

多模态基础模型 (multi-modal foundation models, MFMs). 进一步扩展了大型语言模型的能力, 使其能够同时处理和生成文本、图像、音频等多种模态的信息, 从而实现更富交互性的多模态任务. 包括但不限于: 1) 视觉语言模型 (vision language models, VLMs) 实现了视觉与文本模态的融合, 旨在解决需要关联图像与语言的理解性问题, 典型应用包括指代定位、图像描述和视觉问答; 2) 视觉语言动作模型 (vision language action models, VLAs) 能够根据视觉输入和语言指令, 直接输出物理世界动作.

联邦基础模型 (federated foundation models, FedFMs). 联邦学习 (federated learning, FL) 的目标是在多个分散的数据所有者之间协作训练模型, 使分布于各方的数据得以参与模型训练, 同时增强隐私保护和安全性. 联邦基础模型就是指将大模型整合到联邦环境中, 使 FedFMs 能够访问私有的分散数据.

1.4 缩放定律与计算优化

大模型在参数方面具有良好的可扩展性. 缩放定律 (scaling laws) 正是描述语言模型在数据量、参数量与计算资源增长下的性能变化规律. Kaplan 等[42] 提出损失函数与参数量 N、数据量 D 和计算量 C 之间的幂律关系. Chinchilla 缩放定律进一步指出, 在计算预算有限的情况下, 模型参数量与训练数据量应保持平衡, 其最优比例约为 20 (训练 token 数与模型参数的比值), 同时数据量和模型参数量要同比放大[43]. 实际应用中, 如 GPT-2、LLaMA 和 DeepSeek-V3 等模型虽略高于该比例, 但仍符合这一规律.

训练大规模模型能源消耗巨大, 碳排放的产生非常巨大. 参数高效微调通过仅更新少量参数, 在保持性能的同时大幅降低计算开销与环境影响, 为可持续人工智能研究提供可行路径.

1.5 全量微调

全量微调或称全参数微调, 是指在下游任务上使用任务特定数据对整个预训练模型中所有参数进行训练的策略. 预训练语言模型通常在大规模语料上训练, 以学习通用语言表示. 然而, 由于缺乏领域特定知识, 这些模型在下游具体任务上可能表现不佳. 全量微调通过任务数据进一步调整模型以弥补这一不足.

在全量微调过程中, 模型加载预训练权重后, 通过反向传播与梯度下降算法更新全部参数, 以最小化任务损失函数, 从而学习任务特有的模式与特征. 尽管该方法性能优异, 但其计算成本与数据需求较高, 尤其对于参数量达数十亿的大规模语言模型而言, 全量微调的资源消耗极大. 此外, 当任务数据集较小或预训练模型本身已具备较强的任务适配能力时, 全量微调容易导致过拟合.

相比之下, 参数高效微调方法仅更新或修改模型中的少量参数, 既可大幅降低计算与存储开销, 又能达到与全量微调相当甚至更优的性能, 同时有效缓解了小数据场景下的过拟合问题.

1.6 大模型下游评估

为系统全面地评估微调后模型的性能, 研究人员通常会在一系列经过广泛认可的标准下游任务基准上进行测试. 例如自然语言处理任务与计算机视觉任务, 它们从不同维度考察模型微调后的语言理

解、推理能力以及视觉表征能力.

在自然语言处理领域, 评估基准主要围绕通用语言理解与常识推理两大核心能力构建. 通用语言理解评估 (GLUE) 基准[44] 集成了九个基于句子或句子对的自然语言理解任务, 具体包括: CoLA: 语法判断; SST-2: 情感分类; MRPC: 句子对相似性; STS-B: 语义相似度打分; QQP: 去重; MNLI: 推理; QNLI: 问答配对; RTE: 蕴含关系识别; WNLI: 指代消解. 这些任务在数据集规模、文本类型和任务难度上具有多样性, 均来源于已有的成熟数据集. GLUE 不仅是数据集合, 它还提供一个评估模型性能的平台, 包含一个专门的诊断数据集, 用于深入分析和评估模型对自然语言中各类语言学现象的掌握情况, 是衡量模型通用语言理解能力的经典标杆.

常识推理能力是评估大模型能否像人类一样理解世界运作方式的关键, 通常基准包括: OpenBookQA[45]: 旨在促进高级问答研究, 要求模型在掌握浅层知识的同时, 具备对主题和语言的深刻理解; PIQA[46]: 主要关注日常生活中的物理常识, 偏好非常规的解决方案, 评估模型的物理世界推理能力; SocialIQA[47]: 一个新颖的问答基准, 专门用于衡量模型的社会常识智能; HellaSwag[48]: 一个挑战性极强的数据集, 其核心是评估模型为句子选择恰当结局的能力, 能有效检验模型的常识推理局限性; BoolQ[49]: 一个专注于二分类 (是/否) 问答的数据集; WinoGrande[50]: 一个大规模 (含 44000 个问题) 的基准, 通过代词消歧问题评估模型的常识推理能力; ARC[51]: 由小学水平科学选择题构成. 其中, ARC-challenge 集特意筛选了基于检索的算法均无法正确回答的难题, 而 ARC-easy 集则相对简单, 两者结合可全面评估模型的知识推理能力.

在计算机视觉领域, 图像识别与分类任务包括要求模型区分同一大类下的细微差别的子类, 例如不同鸟类或不同车型, 考验模型的细节表征能力; 多个领域的分类任务集, 用于评估模型在未知任务上的泛化与快速适应能力. 除了静态的图片识别, 视频动作识别也是一个重要领域[52]. PEFT 一般使用目标检测与图像分割[53]、场景解析[54]、目标检测与语义分割[55] 验证其处理高级视觉任务的有效性.

2 PEFT 方法

本文主要依据微调过程中对预训练模型参数的干预原理和方式, 将 PEFT 方法划分为以下四类: 1) 添加式 PEFT (第 2.1 节), 它通过引入外部可训练模块或参数来微调模型; 2) 局部 PEFT (第 2.2 节), 在微调期间选择性更新部分原有参数, 并冻结其他参数; 3) 重参数化 PEFT (第 2.3 节), 在训练阶段, 重构原始模型参数的低维重参数化, 然后在推理阶段等价转换为原始形式; 4) 融合 PEFT (第 2.4 节), 它综合使用上述多种策略, 结合不同 PEFT 方法的优势以构建统一的 PEFT 模型. 本文划分 PEFT 的标准列在表 1 中. 其中, “参数可合并”指训练结束后, 能否将参数更新并入原模型权重, 使得推理时无需引入额外计算模块; “推理延迟”是定性评估, 指相对于原始模型, 该方法所引入的额外计算时间. 不同类型的 PEFT 方法示意图如图 1 所示, PEFT 方法分类图如图 2 所示.

2.1 添加式 PEFT

全量微调计算成本高, 为解决这个问题, 添加式方法在预训练模型中添加一组可训练的参数, 这些参数可以被集成到架构中的不同位置. 在针对具体下游任务进行微调时, 仅微调这些额外组件或参数. 根据这些额外可训练参数在模型架构中所处的位置和方式, 主要可以分为适配器 (adapter) 和软提示 (soft prompt).

2.1.1 适配器

适配器是在预训练模型中添加的小型模块, 通常由一个下投影层、一个非线性激活函数层和一个上投影层组成, 如图 3 所示. 其数学形式如下.

$$\text{Adapter}(x) = W_{up}\sigma(W_{down}x) + x \tag{5}$$

其中, W_{up}, W_{down} 分别代表上/下投影矩阵; σ 代表非线性激活函数; x 代表输入.

基于适配器的微调方法是指在预训练模型中插入适配器模块的方法. 该方法在微调过程中保持预训练模型主体参数冻结固定, 仅优化适配器模块中的少量参数, 从而以较低的计算与存储成本使模型适应下游任务.

现有适配器结构主要分为串行与并行两类. 串行结构将适配器模块依次插入模型的特定位置[56−58].

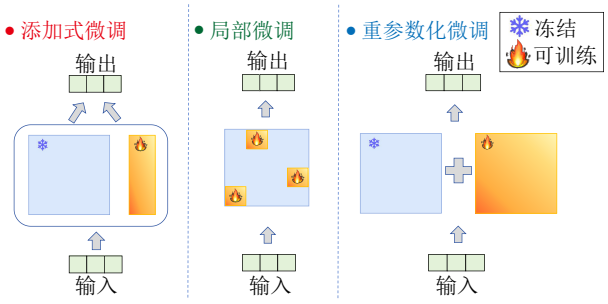

图 1 不同类型 PEFT 方法示意图

Fig. 1 Schematic diagram of different types of PEFT methods

表 1　PEFT 方法分类对比

Table 1　Comparison of PEFT method classification

方法类别	核心思想	分类依据	参数可合并	推理延迟	典型代表方法
添加式	引入全新的可训练单元 (模块/参数/前缀等), 与原始主干并行或串联工作	是否在原始模型结构外增加新的可训练计算单元到原计算图; 训练与推理均依赖该新增单元	否	低到高, 或无	Sequential Adapter, Prefix-tuning, Prompt-tuning, $(IA)^3$
局部	激活或选择模型的一部分进行更新	是否修改网络结构, 是否通过门控、掩码或选择机制来动态决定使用模型的哪些部分; 计算图基本不变	是	低或无	Diff, BitFit, PaFi
重参数化	对原始参数进行低秩或结构性变换, 训练后可与原模型合并	是否通过一个参数化的变换 (如矩阵分解、张量分解) 来间接更新权重, 且推理时能还原为原始架构	是	低或无	LoRA, LoRTA, LoraHub, MultiLoRA, LoRAPrune, QLoRA
融合	结合上述多种策略以发挥各自优势	是否明确地融合了两种及以上不同类别的 PEFT 技术	视情况而定	视情况而定	UniPELT, MAM-Adapter, AUTOPEFT

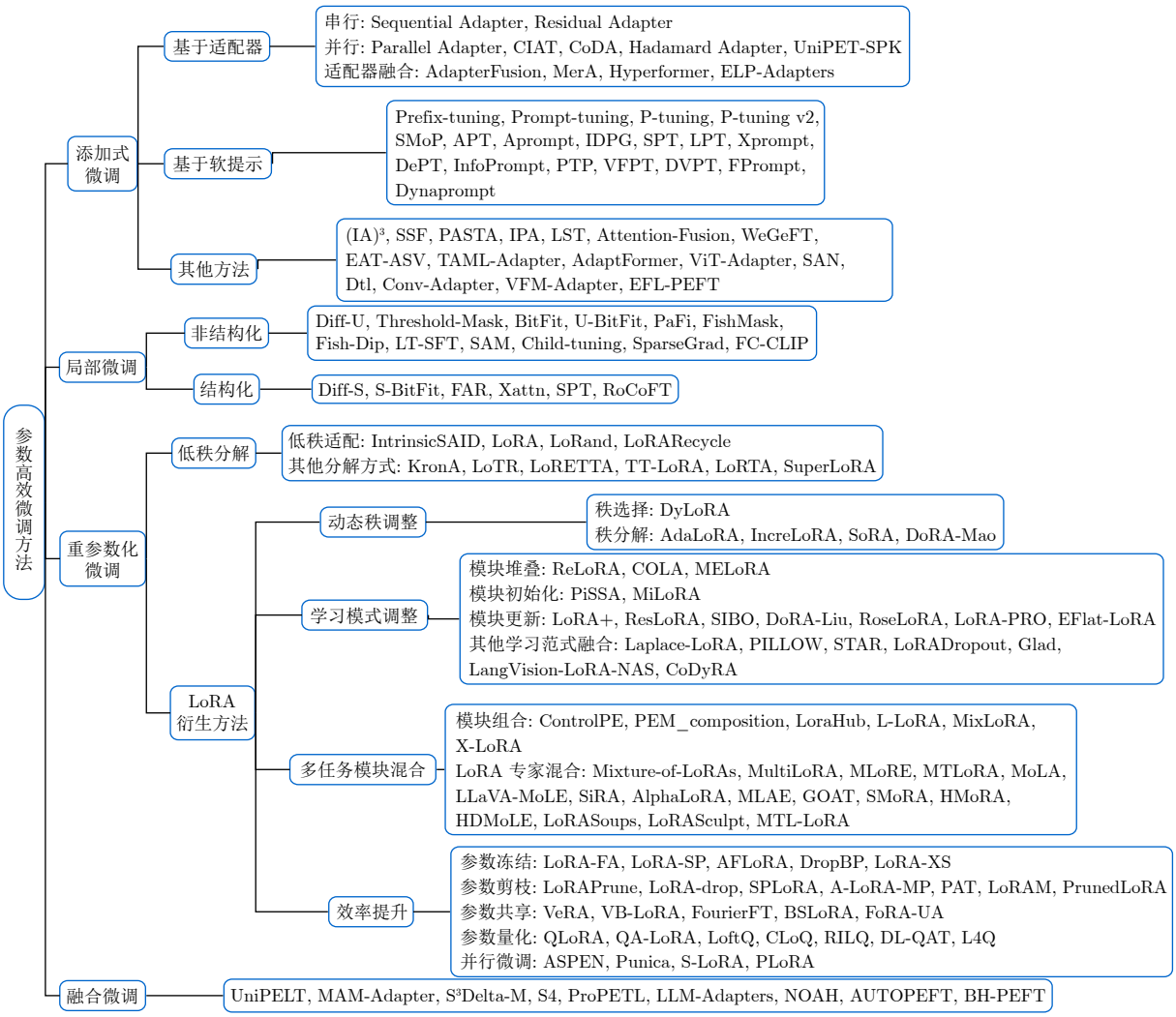

图 2　PEFT 方法分类图

Fig. 2　Classification diagram of PEFT methods

例如, Sequential Adapter[56] 在模型的注意力层和前馈层之后插入适配器; Residual Adapter[58] 将适配器插入在前馈层和归一化层之后.

并行结构则将适配器与原有模块并行部署[59–61]. 例如, Parallel Adapter[59] 将适配器并行在注意力层和前馈层; CIAT[60] 将优化词嵌入的适配器并行在网络中, 达到去除中间层中的多语言干扰的效果; CoDA[61] 在并行结构下, 采用选择性激活机制, 激

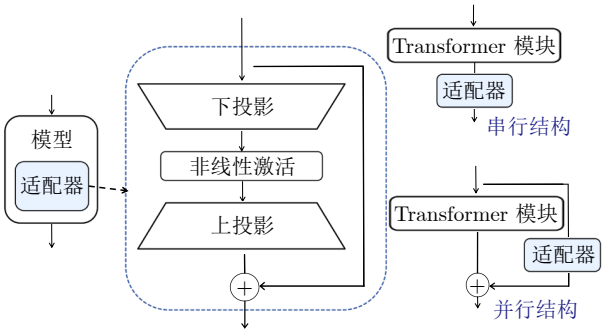

图 3 适配器结构图

Fig. 3 Adapter structure diagram

活一部分重要的输入 token 并由冻结的 Transformer 模块和适配器分支共同处理, 重要性次之的 token 仅由适配器处理, 进一步提高效率而不影响性能; Hadamard Adapter[62] 通过对注意力模块的输出应用逐元素乘法 (Hadamard 积) 和逐元素加法, 生成新的自注意力输出.

在上述基础架构之上, 一些研究聚焦于增强模型性能与泛化能力[57], 采取多任务学习策略, 合并多适配器的多任务信息. 为进一步降低计算开销, MerA[63] 基于权重和激活的最优传输, 采用求和以及平均策略将多个适配器合并成一个适配器, 不引入额外的可训练参数. 为共享多任务间的信息, Hyperformer[64] 通过共享超网络生成适用于所有层和任务的适配器参数, 同时通过任务特定的适配器使模型适应每个单独的任务. ELP-Adapters[65] 设计不同的适配器, 例如编码器适配器 (E-adapters)、层适配器 (L-adapters) 和提示适配器 (P-adapter), 用来快速适应各种不同任务.

基于适配器的 PEFT 不更改预训练模型的权重, 从而保持预训练知识的完整性. 这使得适配器方法能够保持基座模型的通用性, 同时可以对下游任务快速部署. 不过由于增加了参数, 推理的开销有可能增加. 另外, 该方法可能需要关注插入位置、初始化和训练策略, 例如适配器尺寸等最优设置.

2.1.2 软提示

软提示是指在预训练模型输入序列前所添加的可训练连续向量, 可以被表示为

$$X = [s_1,\ s_2,\ \cdots,\ s_N,\ x_1,\ x_2,\ \cdots,\ x_M] \tag{6}$$

其中, X 是输入序列; $s_i,\ i = 1,\ 2,\ \cdots,\ N$ 是软提示序列; $x_i,\ i = 1,\ 2,\ \cdots,\ M$ 是原始输入序列. 由于连续嵌入空间通常蕴含更丰富的信息[66], 软提示可视为一种可学习的上下文, 用于引导模型生成特定任务下的预期输出. 在训练过程中, 仅优化提示向量, 预训练模型主体参数保持固定, 从而显著降低微调所需的参数量和计算成本.

软提示方法的研究主要围绕其插入方式、优化策略与生成机制等方面展开. 早期工作通过不同的角度实现提示向量的插入与优化. Prefix-tuning[67] 在 Transformer 层中的键和值之前引入可学习的前缀向量, 在微调之后, 该前缀向量会被保存用于推理. 为提高训练和推理效率, Prompt-tuning[68] 仅在初始词嵌入中添加可学习的向量, 依然取得良好效果.

后续研究从多个维度改进软提示方法. 为增强软提示的优化性和插入位置灵活性, P-tuning[69] 将加入软提示的输入序列通过一个提示编码器 (LSTM+MLP) 映射到一个隐藏表示, 并通过反向传播优化. 为增强软提示微调在模型参数规模和任务上的通用性, 同时增强软提示的深度优化能力, P-tuning v2[70] 将软提示应用于预训练模型的每一层, 增强软提示方法的能力, 使得其可以适用于较小的模型.

在提示的生成和作用方式方面, 当前研究主要呈现三大技术路线. 一些工作采用门控机制引导提示. SMoP[71] 使用多个较短的软提示处理不同的数据子集, 并利用一个门控机制将每个输入路由到多个软提示中的一个, 从而达到与较长软提示方法相匹配的效果, 同时实现高效的训练和推理. APT[72] 在使用门控机制的同时, 结合了粗/细粒度的门控权重分配. 记第 i 层的前缀 token 为 $\hat{P}_i$,

$$\begin{cases} \hat{P}_i = \lambda_i \odot \alpha_i [P_{ik},\ P_{iv}] \\ \alpha_i = \text{sigmoid}(h_{i-1} W_i) \end{cases} \tag{7}$$

其中, $[P_{ik},\ P_{iv}]$ 表示原始前缀 token 的键值对; λ_i 为可学习尺度参数; $\odot$ 代表元素乘法; α_i 代表门控权重; h_{i-1} 代表第 $i-1$ 层输出; W_i 代表待学习权重. Aprompt[73] 研究提示插入策略, 不仅将提示插入每层的输入之前, 还扩展到注意力模块的查询、键、值子模块里. 记 P_q、P_k、P_v 是插入注意力之前的提示参数, 它们被整合到各自的矩阵中, 以在微调过程中指导注意力计算, 同时保持大部分模型参数不变. 新的注意力计算公式如下:

$$\begin{cases} L(X) = \text{MLP}(\text{LN}(\text{MSA}(X))) \\ \text{MSA}(X) = \text{softmax}\left(\dfrac{Q_{\text{new}}^{\text{T}} K_{\text{new}}}{\sqrt{d}}\right) V_{\text{new}} \end{cases} \tag{8}$$

其中, X 是输入向量; $L(X)$ 是输出; MLP 和 LN 表示冻结的多层感知机和层归一化; MSA 表示 Transformer 编码层内部的多头自注意力机制.

一些工作通过模型生成的方式生成软提示. IDPG[74] 利用轻量级的模型为输入生成提示, 再插

入输入中微调预训练模型. 记轻量级模型为 G, 实例为 x, 任务为 T, 任务定制提示为 $W_p(T,\ x)$. 提示会插入输入序列 x 中, 用统一模板对预训练语言模型 M 进行微调, 如下式所示:

$$\begin{cases} W_p(T,\ x) = G(M(x),\ T), & x \in D_{train} \\ h[\text{CLS}] = M(\text{concat}(x,\ W_p(T,\ x))) \end{cases} \tag{9}$$

其中, D_{train} 表示训练集. 式 (9) 表示将 x 和变换后的输出 $W_p(T,\ x)$ 拼接起来, 并将拼接后的结果输入到模型中, 取模型输出的 [CLS] 标记对应的隐藏状态向量作为最终表示. 为改进启发式人工选择提示所插入的层缺点, SPT[75] 结合门控机制和提示生成, 在每一层引入一个可学习的概率门, 以确定是使用来自上一层的信息还是使用新生成的提示. 为降低反向传播优化提示的开销, LPT[76] 在使用提示生成的同时, 选择预训练模型中的中间层插入提示, 放弃在输入层或全部层插入提示的方法, 被称为“后提示”.

一些工作结合计算优化的思路在软提示框架下提高性能. Xprompt[77] 从结构剪枝的角度去除软提示中的冗余, 保留对模型表现的重要性评分高的部分. DePT[78] 从低秩分解的角度将软提示矩阵分解成两个低秩矩阵, 再根据不同的学习率分别优化这两个矩阵, 完成软提示矩阵的更新. InfoPrompt[79] 构建两个新的损失函数, 通过最大化提示和分类器参数、表征之间的互信息, 使得模型从任务中充分学习到相关的信息, 从而加速收敛. PTP[80] 在损失函数中加入正则化项来稳定软提示微调的过程.

$$\min \text{E}_{(s,\ y)\sim D}[L(M(\theta,\ s+\delta,\ y))] \tag{10}$$

其中, D 是数据集; δ 是从高斯分布中采样的扰动, 或由对抗攻击算法生成的扰动 $\delta = \arg\max_{\|\delta\|\leq\epsilon} L(\theta,\ s+\delta,\ y)$, ϵ 是扰动量的预算; s 是输入序列; y 是其标签; M 是模型; θ 表示可训练的提示参数; L 是损失函数.

近年来, 软提示方法也逐渐拓展至特定领域与任务场景. 在视觉任务中, VFPT[81] 将傅里叶变换方法结合到提示方法中, 整合空间和频率域信息, 提供了一种解决微调任务和预训练差异过大导致的性能显著下降问题的方案. DVPT[82] 提出动态视觉提示微调, 高效地将预训练模型适应到医学任务中. 一些工作在测试阶段进行提示微调, Dynaprompt[83] 基于预测熵和概率差异引入动态提示选择策略, 使得提示能利用测试数据的有利信息来进行优化, 减少错误累积的同时利用了相关的数据分布信息. FPrompt[84] 将提示微调拓展到图任务中, 通过引入混合图提示来增强对抗性数据, 同时增加敏感性、异质性, 达到在预训练图模型方面的通用性.

基于软提示的 PEFT 具有很高的灵活性和多功能性, 不过有可能依赖于模型已具备的能力. 一般而言, 通过提示进行微调带来的性能提升的空间有限.

2.1.3　其他方法

除基于适配器和软提示的方法外, 一些方法通过在微调过程中添加额外的可训练参数的方式达到参数高效微调的目的, 拓展了添加式微调在网络模块与输入之外的可训练单元引入途径. 这一节介绍一些添加式方法中除基于适配器和软提示的方法外的技术.

$(\text{IA})^3$[85] 引入可学习尺度向量到键、值和前馈层激活进行缩放, 这些向量可以整合在权重矩阵中, 使得推理时无需增加额外计算成本. SSF[86] 对模型层间的特征输出添加线性变换来匹配下游任务的特征分布, 表达式如下:

$$y = [\gamma \odot x + \beta]^{\text{T}} \tag{11}$$

其中, 线性变换的参数 γ 和 β 是可学习的. PASTA[87] 仅在自注意力模块中特殊 token 表征前添加可学习向量, 所添加向量的状态变化通过前向传播扩散到其他 tokens. IPA[88] 在解码阶段将预训练模型的输出分布和一个小的适配器的输出分布结合起来, 训练过程中适配器通过强化学习进行优化, 对齐期望目标; 推理阶段, 预训练模型和适配器的分布合并用于解码. 为减小反向传播带来的计算开销, 与现有在主干网络中插入额外参数的方法不同, LST[89] 训练一个小且独立的“梯侧”网络, 通过主干网络的梯连接接收中间激活作为输入并进行预测. 因为不需要通过主干网络进行反向传播, 而仅仅是通过侧网络和梯连接进行反向传播, 显著降低了内存需求. Attention-Fusion[90] 提出一个轻量化的注意力融合模块, 聚合来自预训练模型的中间层表示, 以计算任务特定 token 表示.

一些方法结合适配器和提示方法. WeGeFT[91] 直接从预训练权重中生成微调权重, 并采用简单的低秩形式, 由两层线性层组成, 这些线性层可以在预训练模型的多个层之间共享, 也可以针对不同层单独学习, 从而进一步提升参数、表示、计算和内存多方面的效率. 该方法在常识推理、算术推理、指令遵循、代码生成和视觉识别方面的大量实验验证了其有效性. UniPET-SPK[92] 通过动态可训练的门控机制将并行适配器和提示方法结合起来, 以适应不同的数据集和场景. 利用并行适配器调整潜在特征以及从所有 Transformer 层输出嵌入进行适配; 另

一方面, 在模型的输入空间中拼接可训练的提示 token 以引导适配. EAT-ASV[93] 提出一种基于适配器的微调方法, 将语音模型适配到语音验证任务, 通过并行的适配器设计在预训练模型中插入两种类型的适配器. 在 VoxCeleb1 数据集上的实验结果表明该方法性能良好, 且在仅更新 5% 参数的情况下仍能实现卓越性能. TAML-Adapter[94] 利用任务无关的元学习算法, 在微调前初始化适配器的参数. 在 Common Voice 和 Fleurs 数据集上进行的综合实验显示, TAML-Adapter 在五种数据量有限的语音识别中表现出优越的性能.

2.1.4 发展趋势

总体来看, 添加式 PEFT 领域呈现出从结构探索向性能与效率深度优化演进并持续向领域化发展的趋势.

首先, 在基本结构演进上, 适配器方法呈现出从串行插入 (Sequential Adapter、Residual Adapter) 到并行集成 (Parallel Adapter、CIAT、CoDA) 的明显路径. 这一尝试不仅丰富了信息流动的方式, 还通过如选择性激活等机制为实现更精细的计算控制奠定了基础. 类似地, 软提示方法也经历了插入位置的深化, 从仅在输入层添加 (Prompt-tuning) 扩展到模型的每一层 (P-tuning v2), 显著增强了对深层模型的表征能力.

其次, 在优化策略上, 研究重点从单纯添加参数转向如何高效地生成、组合与优化这些额外参数. 对于适配器, 出现了通过参数共享 (Hyperformer) 与模型合并 (MerA) 来提升效率与泛化性的方法. 对于软提示, 则发展出三大技术路线: 采用门控机制 (SMoP、APrompt、APT) 实现动态路由与分配; 利用轻量级模型生成 (IDPG、SPT、LPT) 替代直接优化, 提升灵活性并降低开销; 引入计算优化思想, 通过低秩分解 (DePT)、结构剪枝 (Xprompt) 或改进损失函数 (InfoPrompt、PTP) 来提升训练效率与稳定性.

再者, 该领域展现出显著的专业化与领域化适配趋势. 大量研究致力于将添加式 PEFT 技术适配到特定领域与场景, 例如计算机视觉 (VFPT、DVPT)、图学习 (FPrompt) 以及医学图像处理 (DVPT) 等. 这些工作通过引入领域先验知识 (如傅里叶变换), 解决了预训练与微调任务差异过大导致的性能下降问题.

最后, 一个前沿方向是探索测试阶段或推理阶段的动态优化. 例如, Dynaprompt 基于测试数据的预测熵动态选择提示, 使模型能利用测试数据分布信息进行自我优化, 这标志着 PEFT 的研究焦点正从单纯的训练时效率向覆盖模型全生命周期的高效适配扩展.

综上所述, 添加式 PEFT 的发展脉络体现了其核心目标从实现基本的功能性插入, 逐步转向追求性能与效率之间最优的平衡, 并最终迈向在复杂真实场景下的专业化、动态化及全周期高效适配.

2.2 局部 PEFT

与添加式微调的思路不同, 在适应下游任务时, 局部微调从预训练模型中选择部分参数进行微调, 而不是新增可训练参数到模型中. 同时, 它能在不用修改网络结构的情况下就可以解决高效微调问题, 计算图基本不变. 局部微调对参数的选择方式可以通过掩码的形式来理解. 假设模型参数表示为 $\theta = \{\theta_1, \theta_2, \cdots, \theta_n\}$, 其中, n 是模型中参数的总个数. 局部微调可以被看作将一个二元掩码 $M = \{m_1, m_2, \cdots, m_n\}$, $m_i \in \{0, 1\}$ 施加在参数 θ 上, 被选中的参数才参与更新.

根据参数的选择方式, 局部微调可以分为非结构化局部微调和结构化局部微调. 非结构化局部微调和结构化局部微调分别对应采用非结构化掩码和结构化掩码. 这里的结构化掩码是指具有明确模式或固定结构的掩码矩阵, 其非零元素通常遵循特定的几何、层次或规则化的分布, 以提高计算效率或引入先验知识; 而非结构化掩码是指掩码矩阵中的非零元素分布无固定模式, 通常由数据驱动或优化过程动态决定, 其稀疏性是随机的或基于重要性选择的, 不遵循任何预定义的几何或层次结构.

2.2.1 非结构化局部微调

非结构化局部微调选择可训练参数的方式在结构上没有固定模式, 方法上通过参数阈值、重要性、信息量度量或构建优化问题来设计选择机制, 以动态决定使用模型的哪些部分实现良好下游任务适应.

Diff-U[11] 是一个代表性工作, 它引入一个适应于下游任务的“diff”向量到预训练模型参数中. 在微调过程中, 该向量会根据添加可微近似 0 范数正则化的损失进行带稀疏性的更新, 从而保持模型参数不变, 该方法在每个任务仅修改一小部分参数. Threshold-Mask[95] 利用阈值来构造二进制掩码矩阵, 在注意力层和前馈层中通过掩码选择微调权重, 在反向传播时更新掩码矩阵. BitFit[96] 冻结大部分模型参数, 仅训练偏置项和关于下游任务的分类层, 在中小型数据集上展现出与全量微调相当的性能. 为扩展适用的模型规模, U-BitFit[97] 结合神经架构搜索的思想, 根据训练损失的一阶近似对模型可训

练部分进行剪枝. 在训练初始微调架构之后, 该方法决定裁剪或者保留微调参数的部分, 之后重新初始化并重新训练被裁剪的架构. PaFi[98] 对模型添加一个不用训练的稀疏掩码, 通过参数大小的绝对值来筛选用于局部微调的参数. FishMask[99] 使用近似 Fisher 信息来确定参数重要性, 然后基于重要性选择前 k 个参数以形成稀疏掩码. Fish-Dip[100] 也采用了 Fisher 信息度量, 区别在于它动态创建稀疏掩码, 在每个训练周期内重新计算参数重要性. LT-SFT[101] 基于模型压缩技术中的“彩票假设”[102−103], 选择在初始微调阶段变化最大的参数子集进行局部微调, 用于跨语言迁移学习. SAM[15] 通过泰勒展开对原始优化问题进行二阶近似, 使得掩码可以被一种可解析求解的方式确定, 表达式如下,

$$\min_{\Delta\theta}\left(L(\theta_0)+(\nabla L(\theta_0))^{\mathrm{T}}M\Delta\theta+\frac{1}{2}(M\Delta\theta)^{\mathrm{T}}HM\Delta\theta\right) \tag{12}$$

其中, M 是掩码矩阵; θ_0 是预训练参数; $\Delta\theta$ 是差分向量; L 是损失函数; $\nabla L(\theta_0)$ 是损失函数在 θ_0 处的梯度; H 是近似的 Hessian 矩阵. Child-tuning[104] 根据与下游任务是否相关提出两种找寻子网络方法, 在每次训练迭代中选择子网络内部的参数进行更新, 即在反向传播中屏蔽其余参数的梯度, 第 t 轮的更新可表述为:

$$w_{t+1}=w_t-\eta\odot\frac{\partial L(w_t)}{\partial w_t}\odot M_t \tag{13}$$

其中, w_t 是第 t 轮的权重; η 是学习率; M_t 是决定子网络的掩码. SparseGrad[105] 主要关注网络中的 MLP 模块, 基于 HOSVD 判断层元素的重要性, 并将层梯度转移到只有约 1% 的层元素的一个空间. 通过将梯度转换为稀疏结构, 减少了更新参数的数量, 该方法中权重变量

$$\tilde{W}^{\mathrm{T}}=UW^{\mathrm{T}}V^{\mathrm{T}} \tag{14}$$

其中, U 和 V 是两个固定的正交矩阵, 由 HOSVD 方法决定, 被模型中各模块共享; W 是模型原权重矩阵; $\tilde{W}$ 是 SparseGrad 方法中的模型权重矩阵.

2.2.2　结构化局部微调

结构化局部微调选择可训练参数的方式遵循结构规则性. 由于结构化的局部微调选择的可训练参数遵循如块状、带状或层次化等结构规则性, 通常能更高效地利用现代硬件的计算能力.

Diff-S[11] 结合给模型权重按组分块的策略, 同时对带稀疏正则化项的损失进行计算, 按组地集体移除或选择参数进行更新, 达到结构化局部微调的效果. S-Bitfit[97] 在参数微调选择机制上延续了 U-Bitfit 的基本框架, 采用基于训练损失一阶近似对模型可训练部分进行剪枝, 区别在于结构化的参数选择对整个偏差更新进行了累加. FAR[14] 专为与压缩语言模型一起使用而设计, 不过模型表现和适应性较未压缩版本有所降低. 它采用冻结和重新配置的策略, 并结合 L_1 范数作为度量, 在前馈层中每个节点被视为一组参数, 这些节点根据在微调期间的表现被冻结或被进行微调. Xattn[106] 针对机器翻译任务, 仅更新模型中的交叉注意力参数, 微调效果可接近全量微调的水平, 可以缓解灾难性遗忘并实现零样本翻译能力. SPT[107] 针对视觉任务, 根据所需要的可训练参数预算, 通过数据驱动计算参数对损失减少的影响度, 以识别任务敏感参数, 同时对独立参数采用非结构化微调, 对高敏感参数密集的权重矩阵采用结构化微调. RoCoFT[108] 是一种针对语言模型的参数高效微调方法, 仅更新 Transformer 权重矩阵的少数行和列. 它通过有限的行列参数构建的核在数值上接近全参数核, 使得其在保证准确性的同时在内存和计算效率上表现更优. 由于这个方法的原理也可以看做是将参数更新增量重参数化为带结构约束的矩阵, 因此也可以被分为重参数化方法. 在本文中, 根据 RoCoFT 的实际操作可以直接更新模型参数, 而不需要产生额外的训练模块, 因此将其分类为结构化局部微调.

无论结构化还是非结构化微调, 局部微调的突出优势在于它不添加新的参数, 这具有双重好处. 首先, 通过控制参数数量, 从而控制了模型的复杂性. 其次, 该方法确保下游任务的推理时间不会增加, 从而有助于保持模型的效率. 然而, 该方法有一些缺点. 首先在内存方面, 局部微调中涉及集成一个掩码矩阵的某些技术, 例如 FishMask 和 Child-tuning, 这会导致内存使用量增多, 限制其在内存受限情况下的使用. 其次, 局部的选择机制可能涉及额外的时间成本, 可能会抵消拥有更少可训练参数的好处.

2.2.3　发展趋势

局部参数高效微调的发展呈现出从初始探索到精细化与结构化的清晰演进路径, 并持续向动态化与硬件友好型方向深化.

首先, 在参数选择的基本范式中, 研究从非结构化掩码的初步应用逐步转向对结构化掩码的深入探索. 早期工作如 Diff-U 和 Threshold-Mask 主要关注如何通过可微正则化或阈值法动态生成非结构化掩码, 其参数选择是稀疏且无固定模式的. 然而, 这类方法虽然灵活, 但往往难以充分利用现代硬件的并行计算优势. 因此, 后续研究如 Diff-S 和 S-BitFit 开始转向结构化掩码, 通过按组 (如整个偏

差更新、前馈层节点) 集体移除或选择参数, 显著提升了计算效率和硬件适配性.

其次, 在参数重要性度量与选择策略上, 呈现出从单一静态判断向多维度动态评估发展的趋势. 初期方法, 例如 PaFi, 依赖参数的绝对值大小等简单启发式规则进行静态筛选. 随后, Fisher 信息矩阵 (FishMask、Fish-Dip) 和泰勒展开 (SAM) 等更严谨的数学工具被引入, 以数据驱动的方式更精确地评估参数对任务损失的重要性. 更进一步, Child-tuning 和 Fish-Dip 等方法将选择过程动态化, 在每次训练迭代或每个周期内重新评估和选择参数子集, 增强了选择的适应性和最终性能.

再者, 局部微调的研究展现出强烈的任务与场景适配性导向. 许多工作不再追求通用算法, 而是针对特定挑战设计解决方案. 例如, Xattn 专门为机器翻译任务设计, 通过仅更新交叉注意力参数来同时实现高性能、缓解灾难性遗忘和获得零样本能力; FAR 则专门为压缩模型 (如 DistilBERT) 设计, 通过冻结和重新配置前馈层节点来适配其降低的模型容量. 更进一步, SPT 将非结构化与结构化微调在同一框架内结合, 对独立参数非结构化, 对敏感参数矩阵结构化, 以数据驱动的方式根据给定预算自适应分配参数, 体现了混合策略的先进性.

最后, 一个重要的趋势是与模型压缩和架构搜索技术的融合. U-BitFit 和 S-BitFit 都借鉴了神经架构搜索 (NAS) 的思想, 通过评估参数重要性并进行剪枝, 来决定最终微调的参数子集. 这表明局部 PEFT 不再孤立发展, 而是与 NAS、模型压缩等领域交叉融合, 共同推动着高效、轻量化适配技术的发展.

综上所述, 局部 PEFT 的发展脉络体现了其核心从实现基本的参数选择, 逐步转向追求性能、效率、硬件之间的协同, 并通过与相关领域融合及针对特定场景深化, 不断拓展其应用边界. 动态化、结构化和任务特定化是当前及未来的主要发展方向.

2.3 重参数化 PEFT

重参数化是一种通过等效变换模型参数以提高训练效率和性能的技术. 在参数高效微调框架下, 该技术通常表现为低秩参数化, 大部分方法主要通过构建低秩的可学习参数矩阵来适应特定下游任务. 训练过程中仅对该低秩矩阵进行微调, 推理阶段则将学习到的矩阵与预训练参数合并, 从而确保推理速度不受影响.

2.3.1 低秩分解

Intrinsic SAID[109] 表明, 常见的预训练模型表现出较低的内在维度. 因此它提出通过一个有效的低维重参数化方式, 以便在整个参数空间中进行微调. 基于这个发现, 重参数化微调的一个主要方式就是基于低秩分解找到低秩矩阵来捕获原始矩阵的基本信息, 通过构建低秩的可学习参数矩阵微调以适应特定下游任务.

低秩适配 (LoRA). 一个直观的想法是找到低秩矩阵来捕获原始矩阵的基本信息, 同时通过重新参数化更新的增量权重来降低计算复杂度和内存使用. 代表性技术就是 LoRA[12]. 它引入两个低秩可训练矩阵 $A \in \mathbf{R}^{r\times d}$ 和 $B \in \mathbf{R}^{n\times r}$, 其中, r 小于 d 和 n. 微调过程只更新这两个矩阵, 通过两个矩阵的乘积矩阵作为更新预训练模型中的权重矩阵 $W_0 \in \mathbf{R}^{n\times d}$ 的增量 ΔW, 表达式如下:

$$W = W_0 + BA \tag{15}$$

其中, BA 就是参数更新增量 ΔW 的重参数化. A 使用随机高斯分布初始化, B 使用零初始化, 确保增量 BA 最初为零. 记某一个输入样本实例 $x = [x_1, \cdots, x_d]^{\mathrm{T}}$, 样本特征的维度是 d, $a_{i,\,j}$ 是矩阵 A 中的元素, $b_{i,\,j}$ 是矩阵 B 中的元素. 对于输入 $x = [x_1, \cdots, x_d]^{\mathrm{T}}$, 第 i 个输出 h_i 的计算公式如下:

$$h_i = \sum_{j=1}^{d} W_{i,\,j}x_j + \sum_{k=1}^{r} b_{i,\,k}u_k \tag{16}$$

其中, $W_{i,\,j}$ 是预训练权重; $u_k = \sum_{j=1}^{d} a_{k,\,j}x_j$ 是中间变量. 微调过程中, 通过 LoRA 模块梯度反向传播完成更新, 梯度计算公式如下:

$$\begin{cases} \dfrac{\partial L}{\partial B} = \begin{bmatrix} \dfrac{\partial L}{\partial h_1} \\ \vdots \\ \dfrac{\partial L}{\partial h_n} \end{bmatrix} [u_1, \cdots, u_r] = \\ \qquad \left(\dfrac{\partial L}{\partial h}\right) x^{\mathrm{T}}A^{\mathrm{T}} \\ \dfrac{\partial L}{\partial A} = \begin{bmatrix} \displaystyle\sum_{i=1}^{n} \dfrac{\partial L}{\partial h_i} b_{i,\,1} \\ \vdots \\ \displaystyle\sum_{i=1}^{n} \dfrac{\partial L}{\partial h_i} b_{i,\,r} \end{bmatrix} [x_1, \cdots, x_d] = \\ \qquad \left(B^{\mathrm{T}} \dfrac{\partial L}{\partial h}\right) x^{\mathrm{T}} \end{cases} \tag{17}$$

其中, L 代表损失函数.

上述前向和反向过程可以通过图 4 清晰显示变

量之间的计算关系和依赖关系. 一旦微调完成, LoRA 的自适应权重与预训练的基础权重无缝集成. 这种集成确保了 LoRA 保持模型的效率, 在推理时不会增加额外负担. 同时 LoRA 的实现相对简单, 并且在参数量高达 1 750 亿的模型上进行了评估, 证明了它的优势.

许多工作从理论上分析了 LoRA 的优势. Malladi 等[110] 证明在特定状态下, LoRA 几乎等同于全量微调. Zeng 等[111] 分析了 LoRA 在全连接网络和 Transformer 网络上的表达能力, 并证明当 LoRA 的秩和网络层数以及神经元数量之间满足一定关系时, LoRA 可以准确表达一个较小的目标模型, 同时该工作给出了二者间的近似误差. Jang 等[112] 理论分析了具有 N 个数据点的 Neural Tangent Kernel 框架下的 LoRA 微调, 表明当使用秩 $r > \sqrt{N}$ 时, LoRA 可以使得梯度下降找到具有很好的泛化性的低秩解. Zhu 等[113] 观察到 LoRA 中的矩阵 A 用于提取特征, 而矩阵 B 用来进行变换得到期望输出, 基于这个观察, 证明了仅微调 B 矩阵可以增强泛化性, 并减少一倍的参数更新量.

其他分解方式. LoRA 可以看作基于矩阵分解完成模型更新. 为扩宽低秩矩阵分解所限制的表示能力或参数约减能力, 一种自然的想法是使用其他分解方式, 例如张量分解等方法, 来进行低秩适配的参数高效微调. KronA[114] 使用 Kronecker 积作为重参数化方式来设计适配器, 它的增量设计如下:

$$\Delta W = B \otimes A, \quad \operatorname{rank}(BA) = \operatorname{rank}(B) \times \operatorname{rank}(A) \tag{18}$$

其中, $\otimes$ 是 Kronecker 积.

LoRETTA[115] 使用 Tensor-train (TT) 分解[116] 设计张量化适配器和相应的 TT 分解重参数化微调方法, 它的增量设计如下:

$$\Delta W = \prod_{i=1}^{d} \mathcal{G}_i \prod_{i=1}^{d} \mathcal{Q}_i \tag{19}$$

其中, $\mathcal{G}_i$ 和 $\mathcal{Q}_i$ 为 TT 分解后 TT 表征的张量因子. 一系列工作关注网络中的不同成分或不同的张量分解方法, 从而得到低秩张量表征. LoTR[117] 将参数的梯度更新表示为张量分解的形式. 每一层的低秩适配器作为三个矩阵的乘积构建, 而这种张量结构的出现是因为在层之间共享该乘积的左侧和右侧乘子. 通过低秩张量表示同时压缩一系列层, 使得 LoTR 能够实现比 LoRA 更好的参数效率, 特别是在深层模型中. 此外, 核心张量不依赖于原始权重维度, 可以使得参数量非常小. SuperLoRA[118] 通过高阶张量分解推广了 LoRA 方法, 数学表示如下,

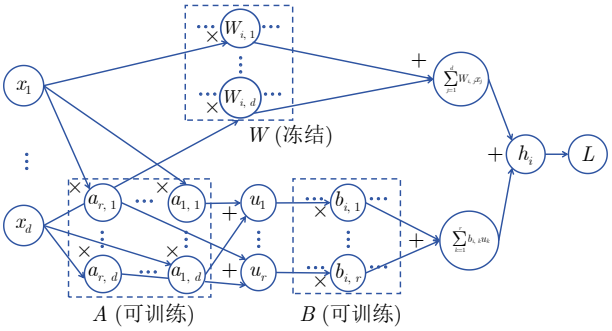

图 4　LoRA 模块的计算示意图

Fig. 4　Computational schematic diagram of the LoRA module

$$F(\Delta W_{\text{lora}_g}) = F\left(\bigotimes_{k=1}^{K}\left(C_{gk}\prod_{m=1}^{M}\times m A_{gkm}\right)\right) \tag{20}$$

其中, ΔW_{lora_g} 是第 g 个参数组的 LoRA 权重更新量; F 表示投影函数将分解后的张量映射回原始参数空间; $\times m$ 是张量模乘, 沿着 M 个不同的维度与因子矩阵相乘; C_{gk} 是核心张量; A_{gkm} 是因子矩阵. TT-LoRA[119] 通过优化的张量列分解扩展了 LoRETTA. 通过消除适配器和传统的基于 LoRA 的结构, TT-LoRA 在不影响下游任务性能的情况下, 实现了更大的模型压缩, 同时减小了推理延迟和计算开销, 从而促进了其在资源有限平台上的部署. LoRTA[120] 基于高阶的 CP (canonical polyadic) 分解, 实现更紧凑和灵活的表示, 同时也允许对适配器大小进行更细粒度的控制.

尽管 LoRA 在某些下游任务中可以实现适当的适应性能, 但在许多下游任务中, LoRA 与全量微调之间仍存在性能差距[121−122]. 鉴于 LoRA 良好的可扩展性, 很多方法基于 LoRA 进行衍生, 例如自适应约束、动态秩调整[123]. 从低秩矩阵扩展到低秩张量[117] 或与 Kronecker 积结合[114], 参数效率可以进一步提高. 除了参数效率, LoRA 模块在训练后可以从模型中分离出来, 具有很好的灵活性, 这种灵活性使得 LoRA 能够被共享, 或用于其他阶段. 此外, LoRA 能够在各种下游任务中实现良好的下游适应性表现. 接下来介绍基于 LoRA 的衍生方法.

2.3.2　LoRA 衍生方法——动态秩调整

在 LoRA 的训练中, 选择一个合适的秩是一个具有挑战性的问题. 从动态秩调整出发的 LoRA 衍生方法就是基于 LoRA 在微调中为低秩矩阵分配合适的秩的技术. 这种方法解决了 LoRA 微调时秩数需要人为确定或分配可能次优的问题. 一系列工作基于不同的出发点进行动态秩调整, 例如秩选择、秩分解, 实现了在微调中动态秩调整.

秩选择. DyLoRA[123] 在训练过程中, 根据训练预算给出可用秩范围, 不使用固定的秩训练 LoRA 模块. 同时, DyLoRA 中的较低秩矩阵是通过最高秩矩阵的截断获得的, 这使得不同秩下的矩阵的参数共享, 提高了效率. 之后通过对不同秩表示的排序选择要使用的秩, 使模型具备跨秩泛化能力, 而无需重新训练或手动选择秩. 它的增量设计如下:

$$\Delta W = W_{\text{down}\downarrow b} W_{\text{up}\downarrow b} \tag{21}$$

其中, $W_{\text{down}\downarrow b} = A[:, b, :]$; $W_{\text{up}\downarrow b} = B[:, :, b]$; $b \in \{r_1, r_2, \cdots, r_n\}$, r_i, $i = 1, 2, \cdots, n$ 代表可选的秩. 由于需要多秩进行微调, DyLoRA 的参数量略高于 LoRA, 不过在多种任务上的性能却优于固定秩的 LoRA.

秩分解. AdaLoRA[124] 通过奇异值分解重参数化预训练模型的权重更新量, 它的设计如下:

$$W = W_0 + P\Lambda Q \tag{22}$$

其中, $P \in \mathbf{R}^{n\times r}$ 和 $Q \in \mathbf{R}^{r\times d}$ 分别是左/右奇异向量构成的矩阵; $\Lambda \in \mathbf{R}^{r\times r}$ 是奇异值矩阵. 这三个矩阵是可学习的, 在训练过程中, 奇异值会根据重要性评分被移除. 为保证 P 和 Q 的正交性, 在损失函数中添加正则项, 这样秩数的动态调整就由参数重要性动态分配. IncreLoRA[125] 通过多个秩一矩阵构建权重更新,

$$W = W_0 + \sum_{i=1}^{r} \lambda_i w_i \tag{23}$$

其中, λ_i 是尺度参数, 通过反向传播更新; $w_i \in \mathbf{R}^{n\times d}$ 是秩一矩阵. SoRA[126] 通过近端梯度方法优化所引入的一个门单元, 用来控制 LoRA 模块的稀疏性. 门单元允许在训练过程中动态调整 LoRA 的秩, 在增强表示能力与保持参数效率中取得平衡. 在推断过程中, 与门单元中的零条目对应的块被消除, 从而将 SoRA 模块简化为一个秩最优 LoRA 模块. DoRA-Mao[127] 通过将高秩 LoRA 模块分解成结构化秩一分量, 再通过重要性动态剪枝得到合适的秩.

2.3.3 LoRA 衍生方法——学习模式调整

从学习模式调整出发的 LoRA 衍生方法是从 LoRA 模块的表示能力、更新方式等角度进行衍生的一系列技术. 这个方法通过改进 LoRA 模块的堆叠、初始化、更新、正则化等方式提升 LoRA 的微调表现或速度等整体性能.

模块堆叠. 由于矩阵秩的次可加性, 可以将多个 LoRA 模块聚合在一起对下游任务进行学习, 以增强表示能力. ReLoRA[128] 提出将多次运行 LoRA 模块中的矩阵进行相加表示模型权重更新, 然后在微调期间重新初始化 LoRA 模块, 相当于训练一个高秩网络, 提高了网络表示能力. COLA[129] 采用一种残差学习方式, 将学习到的 LoRA 模块与预训练模型参数合并, 并将新的 LoRA 模块重新初始化后再进行微调. MELoRA[130] 指出合并与重新初始化过程不一定能保证秩的增加, 因为在微调过程中 LoRA 模块系列之间可能存在重叠. 为解决这个问题, MELoRA 将 LoRA 模块分解为更小的 LoRA 模块, 然后并行堆叠这些小 LoRA 模块.

模块初始化. LoRA 对矩阵 A 和 B 采用数据独立的方式进行随机初始化. 实际上, 对矩阵 A 和 B 的初始化有着不同的影响[131]. PiSSA[132] 使用预训练模型权重矩阵的主奇异成分初始化 LoRA 模块中的矩阵. 由于主奇异成分代表矩阵中信息量重大的方向, 将初始权重与这些成分对齐可以加速收敛并提高性能, 表达式如下:

$$W_0 = W^{\text{pri}} + W^{\text{res}},\ W^{\text{pri}} = BA \tag{24}$$

其中, W^{pri} 代表主成分重参数化后的权重矩阵; W^{res} 代表残差成分重参数化后的权重矩阵. 与 PiSSA 思路相反, MiLoRA[133] 为降低低秩矩阵的随机初始化干扰预训练模型权重矩阵中学习的重要特征的可能, 采用较小的奇异成分初始化 LoRA, 从而提高整体性能, 同时适应新任务.

模块更新. LoRA+[134] 表明对 LoRA 模块中的两个矩阵采用相同的学习率会影响对特征的有效学习. 因此它对两个矩阵应用了不同的学习率, 使得在与 LoRA 计算消耗一致的情况下, LoRA+ 的微调表现和速度都得以提升. 为解决原始模型计算路径较长的问题, ResLoRA[135] 将残差连接引入 LoRA 以优化梯度传播路径, 并在推理过程中使用合并方法消除这些额外路径, 加速训练收敛并增强模型性能. SIBO[136] 和 ResLoRA 的思路类似, 将初始 tokens 表示的残差连接注入 LoRA 中. DoRA-Liu[137] 将预训练的权重矩阵分解为幅度分量和方向分量, 在微调时使用 LoRA 更新方向分量, 从而在保留原始权重幅度的同时高效地更新参数, 限制了梯度更新, 专注于参数的方向变化, 从而增强学习能力和稳定性. 为避免在推理阶段低秩矩阵的乘积更新所有预训练参数, 从而导致诸如知识编辑等需要选择性更新的任务变得复杂的问题, RoseLoRA[138] 通过对低秩矩阵乘积增加稀疏性约束并将其转化为行列稀疏性, 达到仅识别并更新特定任务中重要参数的效果, 在保持效率的同时保留其他模型知识. LoRA-PRO[139] 观察到使用 LoRA 进行优化在数学上等价于使用低秩梯度进行参数更新的全量微调, 因此它通过两个低秩矩阵进行梯度表示, 使得低秩

梯度能够更准确地近似全量微调梯度, 从而缩小 LoRA 与全量微调之间的性能差距. 该方法在自然语言理解、对话生成、数学推理、代码生成和图像分类任务上显著提升了 LoRA 的性能, 有效地缩小了与全量微调的差距. EFlat-LoRA[140] 探讨了 LoRA 的表达能力与泛化能力之间的相关性, 它将 LoRA 全参数空间的损失表示为如下优化问题,

$$\min_{A,\,B}\max_{\|E^W\|_{\mathrm{F}}\leq\rho} L(W_0+sBA+E^W) \tag{25}$$

其中, $\|\cdot\|_{\mathrm{F}}$ 是 Frobenius 范数; E^W 表示扰动; ρ 表示扰动量的预算; s 是施加在 LoRA 模块的平衡参数. 接着从理论上证明了完整参数空间中的扰动可以转移到低秩子空间中, 从而消除了低秩子空间中多个矩阵间扰动可能带来的干扰, 以寻求 LoRA 的平坦极小值. 该方法在大型语言模型和视觉语言模型上的实验表明, EFlat-LoRA 在优化效率上与 LoRA 相当, 同时在性能上达到甚至超过 LoRA. 例如, 在使用 RoBERTa large 的 GLUE 数据集的多项任务上, EFlat-LoRA 的平均指标得分分别比 LoRA 和全量微调提升了 1.0% 和 0.5%.

其他学习范式融合. LoRA 在多方面具有可扩展性, 其中之一就是它可以与其他学习范式兼容. 一些工作将 LoRA 与其他学习范式结合, 以解决影响下游适应性能的问题. Laplace-LoRA[141] 结合贝叶斯学习的视角, 通过估计预测不确定性来解决微调模型中的过度自信问题. 具体而言, Laplace-LoRA 使用拉普拉斯近似来近似 LoRA 参数的后验分布, 从而得到更好标定的模型, 表达式如下,

$$\begin{cases} f_\theta(x^*)\sim \mathrm{N}(f_{\theta_{\mathrm{MAP}}}(x^*),\ \Lambda) \\ \Lambda=(\nabla f_\theta(x^*)|_{\theta=\theta_{\mathrm{MAP}}})\,\Sigma\,(\nabla f_\theta(x^*)|_{\theta=\theta_{\mathrm{MAP}}})^{\mathrm{T}} \end{cases} \tag{26}$$

其中, $\nabla f_\theta(x^*)$ 表示梯度; Σ 是拉普拉斯近似的协方差矩阵; MAP 是最大后验. PILLOW[142] 借助上下文学习的方式, 通过一个网络从用户定义的提示池里选取提示, 将所选提示与用户指令连接作为输入, 并使用 LoRA 微调的 LLM 进行推理. STAR[143] 将 LoRA 与主动学习相结合, 通过引入一种动态不确定性度量, 并结合主动学习迭代过程中基础模型的不确定性和完整模型的不确定度, 解决了微调时的不确定性差距; 同时在 LoRA 训练过程中引入正则化方法, 以防止模型过于自信, 并采用蒙特卡洛丢弃机制来增强不确定性估计, 解决了模型校准不佳问题. LoRA Dropout[144] 结合稀疏正则化学习的方法, 通过对 LoRA 的低秩矩阵应用精细化 dropout 进行微调, 可以缓解过拟合问题. 基于低秩分解的参数高效微调方法隐含地假设所有参数更新均位于同一个低维子空间中, 在模型表达能力上带来了限制. 为实现更紧凑的表达能力和更精细的模型参数调整, AdaMSS[145] 结合多子空间理论, 利用子空间分割获得多个较小的子空间, 只更新与目标任务最相关的少数子空间相关联的参数. 在图像分类、自然语言理解和自然语言生成任务上的大量实验表明, AdaMSS 在大多数情况下实现了与全量微调相当的性能, 并优于其他参数高效微调方法, 同时需要更少的可训练参数. 在 ViT-Large 模型上, AdaMSS 在七个任务中的平均准确率比 LoRA 高出 4.7%, 而使用的可训练参数仅为 LoRA 的 15.4%. 在 RoBERTa-Large 模型上, AdaMSS 在六个任务中的平均准确率比 PiSSA 高出 7%, 同时将可训练参数的数量减少了约 94.4%.

2.3.4　LoRA 衍生方法——多任务模块混合

从多任务模块混合角度出发的 LoRA 衍生方法是利用 LoRA 的模块化特性, 训练专注于不同领域知识的知识插件以增强基座模型能力的技术. LoRA 的模块化特性不用改变模型本身, 使得 LoRA 可以依托同一个基座模型分工合作, 训练专注于不同领域知识的知识插件. 同时, 不仅可以使用单个 LoRA 插件, 还可以将多个插件组合在一起以实现跨任务泛化, 让模型同时具备多种能力. 该类方法可以一定程度上解决模型跨任务泛化问题, 通过模块组合或 LoRA 专家混合方法, 将多个 LoRA 插件组合在一起可以实现跨任务泛化, 让模型同时具备多种能力.

模块组合. 一些工作表明, 仅通过对 LoRA 模块输出进行平均就可以实现适当的跨任务泛化能力[146−147]. 基于这个出发点, 早期的混合 LoRA 方法尝试通过手动设计权重将不同的 LoRA 模块组合起来. ControlPE[148] 利用 LoRA 创建类似于提示加权的效果, 从而能够对提示的影响进行微调. PEM_composition[149] 为 PEFT 模块定义加法和否定运算符, 然后进一步组合这两个基本运算符以执行灵活的算术运算, 通过权重空间中的线性运算来组合 PEFT 模块, 从而集成不同模块的能力, 实现分布泛化和多任务处理. 手动设计权重的混合可以快速混合多个 LoRA, 而无需额外训练, 但是不易找到最佳权重, 导致性能不稳定和泛化能力有限. 因此, 研究人员开始探索使用基于学习的方法来实现更精确和更具适应性的模块混合. LoraHub[150] 聚合了针对不同任务训练的各种 LoRA 模块, 动态组合的 LoRA 模块可以表述为

$$(\omega_1A_1+\cdots+\omega_NA_N)(\omega_1B_1+\cdots+\omega_NB_N) \tag{27}$$

给定新任务的一小部分示例, LoraHub 可以自主地

通过无梯度方法 Shiwa[151] 决定权重 ω_i 来组合 LoRA 模块, 无需人工干预模块组合. L-LoRA[152] 通过线性化 LoRA 模块融合微调的任务向量来有效地构建统一的多任务模型. MixLoRA[153] 动态选择适当的低秩适配矩阵, 基于输入实例被整合到 LoRA 矩阵中用于训练. X-LoRA[154] 利用多任务模块融合进行蛋白质力学分析和设计, 通过动态门控机制, 在 token 之间和层之间为 LoRA 模块分配权重. 这些方法在特定任务或应用场景中显示出很好的性能.

LoRA 专家混合. 为共同学习混合权重和 LoRA 模块, 一些工作使用 LoRA 专家混合方法 (mixture of LoRA experts, LoRA MoE). LoRA MoE 采用端到端训练, 让混合权重的路由网络和 LoRA 专家同时进行优化. 为实现数据平衡以防止灾难性遗忘和任务相互干扰, Mixture-of-LoRAs[155] 分别训练多个特定领域的 LoRA 模块作为初始化, 然后使用路由策略组合 LoRA 模块, 并引入领域标签以促进多任务学习. MultiLoRA[156] 通过减少 LoRA 模块中低秩分解的主成分的主导作用来实现更好的多任务适应, 并改变适应矩阵的参数初始化以减少参数依赖, 从而产生更均衡的单位子空间. 为增强多任务的全局相关性, MLoRE[157] 在原有的 MoE 结构中添加通用低秩卷积路径, 使每个任务特征可以通过这条路径实现参数共享. MTLoRA[158] 采用对任务无差别和对特定任务相关的 LoRA 模块来解耦多任务学习微调中的参数空间, 从而使模型能够熟练地处理任务和环境中的交互. MoLA[159] 自适应地将不同数量的 LoRA 专家分配给 Transformer 模型的不同层, 以节省 LoRA 模块的数量. 为降低不同领域数据的冲突, LLaVA-MoLE[160] 通过为前馈层创建 LoRA 专家, 并基于路由将每个 token 引导到排名第一的专家模块, 使得不同领域的 token 进行自适应选择, 同时通过稀疏激活使得训练和推理成本与原始 LoRA 基本保持一致. SiRA[161] 也采用稀疏的思想, 使用稀疏计算降低计算成本. Alpha-LoRA[162] 揭示了模型每层的专家数量与层训练质量之间的关联, 并使用重尾自正则化理论设计精细分配策略, 用于分配 LoRA 专家, 以进一步减轻冗余. MLAE[163] 结合细胞分解策略, 将低秩矩阵转换为独立的秩一子矩阵或专家模块, 从而增强了独立性. 此外, 通过二元掩码在训练过程中选择性地激活这些专家, 促进了更多样化和各向异性的学习, 增强了低秩矩阵学习过程的独立性和多样性, 同时几乎没有增加参数数量. GOAT[164] 使用奇异值分解结构的混合专家自适应地整合相关先验, 并通过缩放因子来对齐全量微调. 为促进相关任务之间的共享知识, SMoRA[165] 通过将每个秩视为独立的专家, 将混合专家嵌入单一的 LoRA 中, 通过动态的秩稀疏激活机制, SMoRA 实现了自动秩调度, 促进了更精细的知识共享, 同时减轻了任务冲突. HMoRA[166] 提出 LoRA 专家混合的分层微调方法, 通过混合路由和分层方式整合了 token 级路由和任务级路由, 有效地捕捉细粒度的 token 信息和更广泛的任务上下文; 同时引入一种新型的路由辅助损失, 增强了任务路由区分任务的能力以及对未见任务的泛化能力; 此外, 还提出几种可选的轻量级设计, 以显著减少可训练参数和计算成本. HDMoLE[167] 结合 LoRA 专家混合方法和语音错误矫正, 引入分层路由和动态阈值方法到 LoRA 混合专家中, 有效地将单口音 LoRA 专家合并到单个多模态语音错误矫正中, 克服口音多样性挑战, 提升语音识别中的转录效果. 为在跨模态任务中使微调适应专业领域知识的同时减少对一般性知识的遗忘, LoRASoups[168] 将不同的 LoRA 模块进行合并以实现技能组合. 通过 LoRA 的串联 (CAT), 对单独训练于不同技能的 LoRA 进行最优加权,

$$\alpha_1 B_1 A_1 + \alpha_2 B_2 A_2 \tag{28}$$

并将合并参数 α_i 设置为可学习参数, 性能优于现有的模型合并和数据合并技术. 例如在数学文字题上, CAT 分别比这些方法平均高出 43% 和 12%. LoRA-Sculpt[169] 在 LoRA 中引入稀疏更新, 以有效丢弃冗余参数. 同时引入正则化项以改进 LoRA 的更新轨迹, 减轻与预训练权重的知识冲突, 促进了扩模态中的知识协调. 正则化形式如下,

$$\|M \odot (BA)\|_{\mathrm{F}} \tag{29}$$

其中, $\odot$ 是逐元素乘法、Hadamard 积; M 是掩码矩阵. 在多任务学习场景下, 由于基于 LoRA 的方法会将不同任务的稀疏高维特征投影到密集的本征空间, 可能模糊了多任务间的区别, 导致次优的表现, 因此 MTL-LoRA[170] 通过添加任务自适应参数区分特定任务的信息, 在保留 LoRA 高效特点的同时, 适应不同的目标域.

2.3.5 LoRA 衍生方法——效率提升

由于大模型的规模巨大, 对训练和运行的计算量和存储都有相当大的需求. 从效率提升角度出发的 LoRA 衍生方法是利用参数冻结、剪枝、共享、量化、并行等方式提升微调效率的技术. 这些方法的目标是进一步降低训练和运行的计算量和存储, 从不同角度提高 LoRA 模块的效率, 使得其存储更小、计算更快、效果更好.

参数冻结. 参数冻结方法通过冻结一些参数使其不参与训练, 从而降低存储量或计算量. 为降低更新 LoRA 中低秩矩阵所带来的大量激活存储, LoRA-FA[171] 在微调过程中冻结 LoRA 模块中的 A 矩阵, 更新 B 矩阵, 从而确保在不需要增加计算消耗的情况下, 模型微调过程中权重的变化位于由 A 列空间所构成的低秩空间, 同时消除了存储全秩输入激活的需求. LoRA-SP[172] 在 LoRA 框架内随机选择参数进行冻结, 显著减少了计算和内存需求, 同时不影响模型性能. AFLoRA[173] 对于预训练模型的每个冻结权重张量, 添加了一条可训练的并行路径, 基于一种新的冻结评分在微调过程中逐步冻结参数, 以减少计算量并缓解过拟合. DropBP[174] 通过在反向传播期间随机失活一些 LoRA 梯度计算来加速训练过程, 其中每层的随机失活率由敏感度度量计算决定. 一些方法通过在 LoRA 模块之外添加参数的方式减少可训练参数. LoRA-XS[175] 冻结 LoRA 模块的低秩矩阵, 并基于奇异值分解建立一个小的权重矩阵, 添加在 LoRA 低秩矩阵之间.

参数剪枝. 参数冻结方法通过移除一些不重要的参数来降低训练或推理时的存储量或计算量. 由于模型的参数会与 LoRA 模块的参数进行合并, 所以针对模型设计的剪枝方法与 LoRA 不兼容. LoRAPrune[176] 不使用预训练权重的梯度来估计重要性, 而是使用 LoRA 的权重和梯度设计一个剪枝标准来进行迭代剪枝. 它的增量设计如下:

$$\Delta W = BA \odot M \tag{30}$$

其中, M 是 0-1 掩码矩阵. LoRA-drop[177] 通过 LoRA 输出评估参数的重要性, 并移除不重要的参数.

$$g_i = \sum_{x \in D_s} \|\Delta W_i x_i\|^2 \tag{31}$$

其中, g_i 是 LoRA 模块的重要性评分; D_s 是下游任务数据集上的采样, 使用该子集对 LoRA 参数进行多步的更新; ΔW_i 是第 i 个参数变化量; x_i 是对应的输入特征; $\|\cdot\|^2$ 是 L_2 范数的平方. 考虑安全与效用之间的平衡, SPLoRA[178] 作为一种对维度不敏感的相似性指标, 能够有效检测 LoRA 中影响安全对齐的部分并进行移除. 为缓解剪枝可能带来的模型性能降低的问题, A-LoRA-MP[179] 提出一种 LoRA 合并方法, 通过使用最少的目标任务数据进行微调来更新和剪枝 LoRA 参数. 将在 N 个相关任务上训练的 N 个 LoRA 模块记为 B_1A_1, B_2A_2, $\cdots$, B_NA_N, 这些模块被合并后应用到预训练模型中,

$$W_0 + B_1A_1 + B_2A_2 + \cdots + B_NA_N \tag{32}$$

它的剪枝准则如下,

$$I(W_{ij}) = |W_{ij}| \times \|X_j\|_2 \tag{33}$$

其中, $\|X_j\|_2$ 是输入特征的二范数; $I(W_{ij})$ 是对权重 W_{ij} 的重要性评分. PAT[180] 将结构剪枝与微调相结合, 在注意力和前馈层之间插入与 LoRA 开销相当的混合稀疏模块, 通过其中的轻量级操作符和一个全局共享的可训练掩码进行结构剪枝. LoRAM[181] 在一个剪枝模型上训练, 以获得剪枝后的低秩矩阵, 然后将其恢复并与原始模型一起用于推理. Pruned-LoRA[182] 通过结构化剪枝从过参数化的初始化中获得高度代表性的低秩适配器. PrunedLoRA 在微调过程中动态剪除不重要的组件并防止其重新激活, 实现了灵活且自适应的秩分配. 对于结构化剪枝, 通过最小化整体损失的剪枝误差, 优化问题如下,

$$\begin{aligned} &\arg\min_{\mathcal{M}_s,\,\delta} \langle \nabla_W \mathcal{L},\, \delta\rangle + \frac{1}{2}\mathrm{tr}(\delta \widetilde{\boldsymbol{H}} \delta^{\mathrm{T}}) \\ &\text{s.t.} \quad \delta_{:,\,\mathcal{M}_s} = -W_{:,\,\mathcal{M}_s} \end{aligned} \tag{34}$$

其中, $\mathcal{M}_s$ 是掩码, s 是稀疏度; δ 是剪枝后的参数带来的更新, 更新被约束于剪枝掩码; $\widetilde{\boldsymbol{H}}$ 是 Hessian 近似; $\nabla_W \mathcal{L}$ 是梯度; $\langle\cdot,\,\cdot\rangle$ 代表内积; tr 代表矩阵的迹. 将模型拓展到 LoRA 中得到

$$\begin{aligned} \arg \min_{\mathcal{M}_s,\,\delta_A,\,\delta_B} &\langle \nabla_A \mathcal{L},\, \delta_A\rangle + \frac{1}{2}\mathrm{tr}(\delta_A \widetilde{\boldsymbol{H}}_A \delta_A^{\mathrm{T}}) + \\ &\langle \nabla_B \mathcal{L},\, \delta_B\rangle + \frac{1}{2}\mathrm{tr}(\delta_B \widetilde{\boldsymbol{H}}_B \delta_B^{\mathrm{T}}) \\ \text{s.t.} \quad (\delta_B)_{:,\,\mathcal{M}_s} = -B_{:,\,\mathcal{M}_s}&, \quad (\delta_A)_{\mathcal{M}_s,\,:} = -A_{\mathcal{M}_s,\,:} \end{aligned} \tag{35}$$

通过求解上述问题得到剪枝和更新. 这种基于梯度的剪枝策略提供了细粒度的剪枝和恢复更新, 且具有扎实的解释性. 同时, 证明了在权重扰动影响下, 基于梯度的剪枝相较于基于激活的剪枝在整体损失方面更为稳健. 在实证方面, PrunedLoRA 在数学推理和自然语言理解的监督微调任务中持续优于 LoRA 及其变体, 同时在不同稀疏率下, 也表现出优于现有结构化剪枝方法的优势.

参数共享. 参数共享方法在各层或各模块之间共享参数以降低训练或推理时的存储量或计算量. VeRA[183] 在所有层之间共享一对低秩矩阵, 并使用尺度向量进行逐层适应. VB-LoRA[184] 通过向量库共享模型整体的参数, 先分割再共享的思路打破了矩阵、模块、层之间的限制. FourierFT[185] 将增量矩阵通过傅里叶变换进行转换, 所有层之间共享谱, 仅学习每一层的稀疏谱系数, 从而减少可训练参数的数量. BSLoRA[186] 通过内部 LoRA 和外部 LoRA 参数共享扩展了局部 LoRA, 以更好地捕捉局部和全局信息. 同样考虑傅里叶变换, FoRA-UA[187] 认

为固定大小的稀疏频率表示能够更准确地逼近小矩阵, 另外在可训练参数数量固定的情况下, 引入较小的中间表示来逼近较大矩阵可降低重构误差. 因此, 通过插入一组小的中间参数, FoRA-UA 仅使用标准 LoRA 参数的 1% ~ 5%, 即可在自然语言理解 (NLU)、自然语言生成 (NLG)、指令调优以及图像分类任务中实现极端压缩情况下的强泛化能力和稳健性.

参数量化. 参数量化通过减少参数的比特数来降低 LoRA 的内存和计算成本. 有的量化方法将量化和微调依次进行, 先对大模型进行量化, 然后对量化后的模型进行微调. QLoRA[188] 首先将大模型量化为 4 位, 然后在其上使用更高精度的权重, 例如 BFloat16 或 Float16 微调 LoRA 模块. 在推断阶段, QLoRA 对模型使用与 LoRA 相同精度的权重, 并将 LoRA 更新添加到 LLMs 中. 虽然 QLoRA 可以显著减少微调的内存成本, 但在推断阶段需要再次将模型提升到高精度. 为解决这个问题, QA-LoRA[189] 使用分组操作来平衡大模型量化和微调的自由度, 使模型和 LoRA 模块的精度能够匹配. 一些方法将量化和微调同时进行. LoftQ[190] 在微调阶段交替使用量化和低秩近似来最小化量化误差, 它的增量设计如下:

$$\Delta W = \mathrm{SVD}(W_0 - Q_t), \quad Q_t = q_N(W_0 - BA) \tag{36}$$

其中, q_N 是 N-bit 量化函数; SVD 代表奇异值分解. CLoQ[191] 在初始化过程中最小化原始模型与其加入 LoRA 量化后的层级差异, 在量化过程中为每一层确定最优的 LoRA, 确保后续微调的稳固. RILQ[192] 基于对模型激活差异损失的秩分析, 利用该损失协同调整各层的适配器, 实现与 LoRA 秩无关的量化误差补偿, 从而提升 2 位模型量化的准确性. DL-QAT[193] 通过分组量化, 在每个量化组内使用 LoRA 矩阵来更新量化空间中的权重大小和方向. L4Q[194] 将量化感知训练 (QAT) 与 LoRA 结合, 通过内存优化的层设计显著降低了 QAT 的内存开销, 使其训练成本可与 LoRA 相媲美, 同时保留了 QAT 在生成高精度全量化 LLM 方面的优势. 量化准则如下

$$\tilde{w} = \mathrm{round}\left(\mathrm{clamp}\left(\frac{W-b}{S_c}, Q_N, Q_P\right)\right) \tag{37}$$

其中, W 是原始浮点权重; b 是偏置项 (平移参数); S_c 是缩放因子; round 代表取整操作; clamp 代表截断操作, 截断保持数值在 $Q_N = -2^{n-1}$ 到 $Q_P = 2^{n-1} - 1$ 量化范围内, 避免溢出; $\tilde{w}$ 是量化后的数值. 反传的规则如下,

$$\begin{cases} \dfrac{\partial L}{\partial S_c} = \dfrac{\partial L}{\partial W_q}\dfrac{\partial W_q}{\partial S_c} \\ \dfrac{\partial L}{\partial b} = \dfrac{\partial L}{\partial W_q}\dfrac{\partial W_q}{\partial b} \end{cases} \tag{38}$$

其中, L 代表损失函数; W_q 代表量化后的权重. 实验表明该方法在精度上优于先微调后量化的微调方案, 尤其是在 4 位和 3 位量化中.

并行微调. 并行微调是指在 GPU 上进行多个 LoRA 模块的计算, 用于减少 GPU 内存使用并提高计算效率. ASPEN[195] 针对 LoRA 的高吞吐量并行微调框架, 通过将多个输入批次融合成一个批次来支持在模型上并行微调多个 LoRA 模块, 同时利用自适应作业调度算法将计算资源分配给微调进行工作. Punica[196] 使用新的 CUDA 内核对不同 LoRA 模块的 GPU 操作进行批处理. S-LoRA[197] 通过引入分页机制和新的张量并行策略进一步优化了并行推理框架, 使得能够支持数千个并发的 LoRA 模块. PLoRA[198] 在给定的硬件和模型约束下自动协调并发的 LoRA 微调作业, 并开发出高性能的内核以提高训练效率. 在给定超参数搜索空间中, 将 LoRA 微调的完成时间缩短至原来的七分之一, 并改善了训练过程.

2.3.6 发展趋势

重参数化 PEFT 技术, 特别是以 LoRA 为代表的低秩自适应方法, 已发展成为参数高效微调领域的主导范式. 其演进路径呈现出从基础架构构建到性能与效率深度优化, 并持续向自动化、结构化与系统化方向发展的清晰趋势.

首先, 在核心低秩分解范式上, 研究从标准矩阵分解向多样化张量分解拓展. LoRA 通过引入两个低秩矩阵进行矩阵分解, 奠定了该领域的基础. 然而, 其表示能力受限于矩阵分解形式. 因此, 后续研究开始探索更高效或更富表现力的张量分解方法, 如 Kronecker 积、TT 分解或 CP 分解等. 这些方法通过利用参数间的多维结构相关性, 在参数压缩率和表示能力上实现了进一步突破, 特别是在深层模型中表现显著.

其次, 许多工作围绕 LoRA 进行衍生, 对 LoRA 内部机制的深入理解与精细化设计成为提升性能的关键. 例如在低秩适配的优化策略上, 呈现出从静态固定秩向动态自适应秩演进的重要趋势. 由于手动选择 LoRA 的秩比较困难, 一系列动态秩调整方法被提出, 标志着核心超参数的设定从“人工预设”走向“数据驱动与动态优化”的方式.

再次, 初始化策略方面, 从随机初始化演进到信息感知初始化, 以更好地与预训练权重对齐, 加

速收敛并提升性能; 更新规则方面, 精细化抓住优化训练动态, 在增强学习稳定性的同时保持了参数效率; 融合多种学习范式, 通过 LoRA 的可扩展性, 与贝叶斯学习结合以量化不确定性, 与主动学习结合以优化数据利用, 或与 Dropout 结合以增强正则化效果.

最后, LoRA 的模块化特性催生了一个广阔的“插件生态”, 其发展重心从单一模块训练转向多模块的混合. 早期方法通过手动加权平均组合多个 LoRA 模块. 随后, 研究转向更智能的学习型混合, 通过引入路由网络动态地为不同输入或任务选择最合适的 LoRA 专家, 旨在高效实现多任务学习、知识融合和跨任务泛化, 并有效缓解任务冲突与灾难性遗忘.

除此之外, 面对大模型的实际部署需求, 效率提升成为不可或缺的研究主线. 围绕存储、计算、通信开销的优化形成了五大技术路径: 参数冻结、参数共享、参数剪枝、参数量化、计算优化. 实现了对成千上万个 LoRA 模块的高效并行服务与推理, 为其大规模商业化应用奠定了工程基础.

综上所述, 重参数化 PEFT 的发展历程体现了其从一种高效的微调技术, 逐步演进为一个层次丰富、技术多元、生态繁荣的综合性研究方向. 未来, 该领域预计将在理论解释以及更紧密的软硬件协同设计等方面持续深化, 最终推动大模型高效适配技术在更多样化的场景中落地应用.

2.4 融合 PEFT

不同类型的 PEFT 方法在不同任务和模型架构中的表现有一定差异. 为充分发挥各类方法的优势并构建统一框架, 一些研究关注融合参数高效微调 (hybrid PEFT) 方法, 通过有机结合多种 PEFT 机制, 在保持参数效率的同时实现更优的跨任务泛化性能.

融合 PEFT 方法的核心思想是通过结构组合或控制筛选机制动态整合不同 PEFT 方法的优势. UniPELT[199] 将 LoRA、Prefix-tuning 和适配器方法集成在 Transformer 中, 通过门控机制控制激活相应的模块. MAM-Adapter[59] 探究适配器、Prefix-tuning 和 LoRA 之间的内在相似性, 并根据观察开发三种衍生方法: 并行适配器, 将适配器层与自注意力层或前馈层并行放置而非置于其后; 多头并行适配器, 将并行适配器分割为多个注意力头, 每个头分别影响自注意力模块中相应头的输出; 缩放并行适配器, 在并行适配器层后添加缩放因子. 通过实验找到有效的融合微调配置方案: 在自注意力层或前馈层采用 Prefix-tuning 方法, 同时在前馈网络层使用缩放并行适配器. 该混合架构被命名为 MAM-Adapter. S^3Delta-M[200] 通过双层优化和稀疏控制机制进行可微的结构搜索, 并提出移位全局 sigmoid 方法, 以明确控制可训练参数的数量. S4[201] 结合 BitFit、LoRA、Prefix-tuning 和适配器, 通过将可训练参数均匀分配给层, 并将层分组, 再按组微调, 为不同组分配定制策略, 在自然语言处理任务中有不错表现. ProPETL[202] 在每层里结合 LoRA、Prefix-tuning 和适配器模块, 并学习二元掩码选择相应模块. LLM-Adapters[203] 通过实验发现串行适配器、并行适配器和 LoRA 的最佳插入位置分别为 MLP 层之后、MLP 层旁侧以及同时插入注意力层和前馈层之后; 采用 PEFT 小规模大语言模型在特定任务上能够达到甚至超越更大规模模型的性能; 在获得分布适当的微调数据时, 小模型能够在特定任务性能上超越大模型. 一些工作利用神经架构搜索技术来寻找更优的 PEFT 组合策略. NOAH[204] 发现不同的 PEFT 适用于不同任务, 基于这个观察, NOAH 采用 NAS 算法为每个数据集自动识别最有效的 PEFT, 例如适配器、LoRA 和视觉提示微调. 与此类似, AUTOPEFT[205] 首先构建包含串行适配器、并行适配器和 Prefix-tuning 的搜索空间, 随后提出一种基于高维贝叶斯优化的 NAS 方法. 为提升动态适应新出现的数据和现实世界情况的能力, BH-PEFT[206] 将贝叶斯学习融进 PEFT, 结合适配器、LoRA 和 Prefix-tuning, 微调 Transformer 的前馈层和注意力层. 通过将可学习参数建模为高斯分布, BH-PEFT 使得不确定性量化成为可能, 同时提出一种基于 BH-PEFT 的贝叶斯动态微调方法. 在每一轮微调中, 最后的后验分布作为下一个轮次的先验分布, 从而有效地适应新数据.

融合 PEFT 提供了一个统一框架, 将各种 PEFT 方法整合到一个单一协调的结构中, 从而增强了整个系统的灵活性和适应性. 此外, 该方法可以利用各个 PEFT 方法的优势, 提高处理多种场景的性能和稳健性. 然而, 多种 PEFT 方法的融合可能会引入更高的复杂性, 导致配置困难、增加计算需求以及额外开发成本. 融合后的结果基于启发式设计, 缺乏指导, 调整超参数以及不同方法之间复杂的交互作用可能会带来挑战.

融合 PEFT 方法通过多层次、多策略的融合机制, 不仅提升了参数效率与任务性能之间的平衡, 也为构建下一代自适应微调框架奠定了理论基础. 其核心演进路径呈现出从简单组合到智能融合, 并

逐步向自动化与理论化方向深化的清晰趋势.

首先, 一个重要的趋势是系统性实证研究指导融合框架设计. 一些工作通过大量严谨的实验, 系统性地评估了不同 PEFT 方法 (串行/并行适配器、LoRA) 在不同模型和任务中的最佳插入位置与性能表现. 这些结论为后续构建更高效的融合框架, 例如如何组合模块、放置在何处, 提供了至关重要的数据支撑和设计准则, 体现了该领域工程实践与理论创新并重的特点.

其次, 研究在向数据驱动自动搜索演进. 通过人工分析得出不同 PEFT 方法难以普适所有任务与模型, 因此, 后续研究如 NOAH 和 AUTOPEFT 开始采用神经架构搜索技术, 将多种 PEFT 方法构建为一个超网络或搜索空间, 通过自动化算法为特定任务或数据集寻找最优的混合架构与配置, 标志着融合策略从"人工启发"迈向"自动寻优".

再者, 在融合的架构与控制机制上, 呈现静态和动态两种方案. S4 采用静态策略, 将层分组并为每组分配固定的 PEFT 方法. 一些方法引入了动态控制机制以实现精细的适配. 例如, ProPETL 通过学习二元掩码为每一层动态选择最合适的 PEFT 模块; S^3Delta-M 则通过可微搜索和稀疏控制机制来明确控制可训练参数的数量和结构, 增强了融合框架的灵活性与效率.

最后, 一些融合 PEFT 的研究展现出不确定性量化导向, 以适应更复杂的现实场景. BH-PEFT 是这一趋势的代表, 它将贝叶斯学习与融合 PEFT 相结合, 将可训练参数建模为分布而非确定值. 这不仅使得模型能够量化预测不确定性, 提升了决策的可靠性, 还通过序贯贝叶斯更新将上一轮后验作为下一轮先验, 实现了对动态新增数据的持续高效适应, 为融合 PEFT 在开放环境中的终身学习应用奠定了基础.

综上所述, 融合 PEFT 的发展脉络体现了其目标从简单结合多种方法以提升性能, 逐步转向追求在参数效率、任务性能、自动化程度和动态适应性之间达到更优的平衡. 未来, 融合 PEFT 研究将更侧重于动态可重构架构、跨模态统一建模、理论最优性证明以及高效自动化搜索算法等前沿方向.

2.5 PEFT 代表性方法性能定量比较分析

前文系统性地阐述了各类 PEFT 的核心原理与演进脉络. 本节进一步通过实验结果来衡量其实际表现. 为给每一类 PEFT 提供清晰的对比依据, 本节在当前 PEFT 四类方法中选取每类具有代表性的方法, 给出它们在相同的模型和相同数据集上的性能表现和参数效率, 从而进行一次较全面的定量性能比较. 本节中的实验结果基于对现有文献中所给出的数据进行系统整合, 以呈现当前领域的共识性结论. 结果如表 2 所示, 表中 PEFT 的实验结果一部分从已发表文献 [207−208] 中引用, 另一部分由实验获得, "—"代表文献中并未提及.

数据集方面, 由于 GLUE 基准由一系列自然语言理解任务组成, 所以它被许多工作作为验证 PEFT 方法有效性的首选评估数据集. 因此, 本文选取其中五个作为共同指标. 不同数据集的评价指标如下: CoLA 数据集采用 Matthews 相关系数, 其他数据集采用准确率. 相应地, 在模型方面, 本文选取自然语言处理领域的 RoBERTa-Large 作为基础预训练模型, 兼顾了模型的通用性和代表性.

从表 2 中可以看出, 大多数 PEFT 方法以极低的参数成本达到甚至超过了全量微调 95% 以上的性能. 在各类方法中, 基于低秩适配的重参数化方法, 以 LoRA 和其衍生工作为代表, 展现出卓越的潜力, 能以相比之下更低的参数量实现高效微调. 然而, 一个显著的趋势是, 当追求极致的参数效率时, 模型性能往往会出现一定量的折损. 这意味着在参数预算被极度压缩的场景下, 微调效果会有所损失. 尽管如此, 部分先进方法通过精妙的设计, 在仅更新 0.1% 甚至更少参数的情况下, 仍能保持良好的性能, 这凸显了 PEFT 方法设计优化的重要性. 综合考虑性能和效率指标, 每类方法中部分工作效果具有相似性, 考虑到原理的互补性, 可以进行集成来进一步融合. 从部署角度看, 不同 PEFT 方法的推理效率存在本质差异. 由于适配器的添加式特性, 在推理时, 增加的参数或模块会导致推理时间增加; 而局部微调选择模型结构中的原有部分微调, 不改变计算流程, 所以不增加推理延迟; 重参数化方法会增加参数来表示矩阵增量, 但是由于可以通过权重相加的合并方式融合在预训练模型的原始权重中, 所以不增加推理延迟.

3 PEFT 在不同模态与跨领域模型上的应用

随着大模型在自然语言处理领域的显著进展, 大模型技术也拓展到其他模态. PEFT 在不同模态与跨领域模型上的应用也逐渐增多. 大模型微调都面临效率、性能和泛化三者之间的平衡, 但不同模态的侧重点不同, 同时在不同模态上的 PEFT 也会有各自特点. 接下来介绍 PEFT 在视觉基础模型、语音基础模型、视觉语言模型、视觉语言动作模型、联邦基础模型上的应用.

表 2　PEFT 代表性方法在 RoBERTa-Large 模型上的性能对比 (%)

Table 2　Performance comparison of representative PEFT methods on RoBERTa-Large model (%)

类别	方法	作用位置	推理延迟	可训练参数	数据集					平均值
					SST-2	MRPC	CoLA	QNLI	RTE	
—	全量微调	所有参数	无	100.00	96.1	90.2	68.0	94.2	86.0	87.1
添加式	Prefix-Tuning	注意力	有	0.11	95.6	86.8	59.1	94.6	74.8	82.2
	Prompt-Tuning	输入	有	0.30	94.6	73.0	61.1	89.1	60.3	75.6
	Sequential Adapter	前馈层	有	4.72	96.0	89.2	65.4	94.5	84.1	85.8
	$(IA)^3$	注意力、前馈	无	0.34	94.6	86.5	61.1	94.2	91.2	85.5
局部	BitFit	注意力	无	0.41	96.1	90.9	68.0	94.5	87.7	87.4
	Child-tuning$_D$	—	无	0.10	95.1	90.7	63.1	93.1	86.3	85.7
	SparseGrad	前馈	无	47.32	96.8	90.5	63.2	93.3	64.7	81.7
	RoCoFT$_{1row}$	注意力、前馈	无	0.06	96.6	90.0	65.7	94.2	85.3	86.4
	RoCoFT$_{3row}$	注意力、前馈	无	0.18	96.7	91.1	67.4	94.9	87.8	87.6
	RoCoFT$_{1col}$	注意力、前馈	无	0.06	96.6	89.1	64.9	94.1	85.7	86.1
	RoCoFT$_{3col}$	注意力、前馈	无	0.18	96.7	89.9	67.2	94.8	87.8	87.3
重参数化	LoRA	注意力	无	0.24	96.2	90.2	68.2	94.8	85.2	86.9
	LoRA-FA	注意力	无	1.10	96.0	90.0	68.0	94.4	86.1	86.9
	AdaLoRA	注意力	无	0.24	95.0	90.4	66.9	94.6	84.5	86.3
	PiSSA	注意力	无	0.24	95.5	86.9	61.1	92.1	56.8	78.5
	FourierFT	注意力	无	0.01	96.0	90.9	67.1	94.4	87.4	87.2
	VeRA	注意力、前馈	无	0.02	96.1	90.9	68.0	94.4	85.9	87.1
	RoseLoRA	注意力、前馈	无	0.02	95.2	90.2	69.2	94.7	89.2	87.7
	LoRA-PRO	注意力	无	0.24	95.9	90.9	66.7	93.0	60.5	81.4
	LoRA-dropout	注意力	无	0.12	96.2	89.9	68.5	94.9	88.8	87.7
	WeGeFT	注意力	无	0.02	95.0	75.7	64.0	93.7	53.6	76.4
	EFlat-LoRA	注意力	无	0.24	96.3	90.3	68.0	94.8	89.3	87.7
	LoRETTA	注意力	无	0.04	96.2	90.5	69.5	94.1	53.0	80.7
	FoRA-UA	注意力	无	0.01	96.6	91.2	69.0	93.9	86.9	87.5
融合	MAM-Adapter	—	有	12.30	95.8	90.1	67.3	94.3	86.6	86.8
	ProPETL-Adapter	注意力、前馈	有	1.50	96.3	89.7	65.6	95.2	88.9	87.1
	ProPETL-Prefix	注意力	有	7.60	96.2	90.0	62.2	94.7	79.7	84.6
	ProPETL-LoRA	注意力	无	1.20	95.9	89.1	61.9	94.9	83.6	85.1

3.1　视觉基础模型

随着大模型在自然语言处理领域的显著进展, 大模型技术也拓展到计算机视觉领域. 其中关于基础模型适配方面, 微调方法备受关注, 通过微调能够更好地将视觉基础模型, 例如 Vision Transformer (ViT)[209] 和大规模卷积神经网络 (CNN), 适应于各种下游任务. 然而, 视觉基础模型的参数量依然巨大, 若进行全量微调, 计算和存储成本极高. 同时在适应特定领域时, 全量微调有可能影响原始模型在通用视觉上的强大泛化能力. 另外, 若某些专业视觉领域的标注数据稀缺, 或与预训练数据分布差异大时, 可能导致模型难以有效迁移. 考虑到上述计算、内存和参数效率等问题, PEFT 方法被引入视觉基础模型. 本文将现有应用于视觉模型的 PEFT 方法, 根据其核心机制划分为添加式方法[210−216]、局部方法[217]、重参数化方法[218−219] 和融合方法[204].

添加式方法. 在添加式方法中, 一些方法通过将适配器添加到网络中的不同部分, 完成微调技术到视觉基础模型上的拓展. AdaptFormer[210] 引入参数不到 2% 的适配器模块到前馈层, 将预训练的 ViT 模型适配到多种图像和视频任务中. ViT-Adapter[211] 使用适配器方法将图像相关的归纳偏置引入 ViT 模型, 使得模型可以迁移到目标检测、实例分割和语义分割等下游任务. SAN[212] 在冻结的 CLIP 模型上附加了一个边侧网络, 并利用预训练视觉语言模型进行开放词汇语义分割. 其中边侧网络包含两个解耦分支设计, 有助于 CLIP 识别掩码,

且仅增加了约 4.47% 可训练参数. Dtl[213] 在网络中添加轻量级紧凑侧网络 (CSN), 从而将可训练参数与主干网络解耦, 同时通过低秩线性映射逐步提取特定任务的信息, 并将信息适当地添加至主干网. 仅使用约 0.06% 的可训练参数, 且减少了大量的 GPU 内存使用.

为更进一步提升性能、效率和任务泛化性, Conv-Adapter[214] 在卷积神经网络上设计适配器模块, 将任务特定的特征作用在主干网络的中间表示. 在 ResNet50 上的可学习参数约为 3.5%. 在 23 个不同领域的分类任务中实现了与完全微调相当甚至更优的性能. 在少样本分类任务中也表现出卓越的性能, 平均优势为 3.39%. 除了分类任务, Conv-Adapter 还可以推广至检测和分割任务, 在减少超过 50% 参数的同时, 性能与全量微调相当. Bayesian-PEFT[216] 将软提示方法与两个贝叶斯组件结合, 确保在具有挑战性的少样本场景下的可靠预测. VFM-Adapter[215] 设计混合映射, 将本地信息与全局建模无缝集成到适配器模块中, 同时为适配器动态生成参数, 从而在给定特定输入的情况下灵活地提取特征. 该方法在对象检测、语义分割和实例分割任务中表现良好, 且添加参数仅占 SAM-Base 主干网络可训练参数的 3%.

局部方法. 除了添加式 PEFT 的引入, 一些工作采用局部微调和重参数化微调的思路. FC-CLIP[217] 通过单阶段框架完成掩码生成和特征提取, 不仅显著简化了当前的两阶段流程, 还显著实现了更好的精度与成本之间的权衡, 参数量在 1% 到 8%. 另外, FC-CLIP 在各种基准测试中结果领先, 同时运行速度非常快.

重参数化方法. LoRand[218] 通过低秩方法生成小型适配器, 插入主干网络的注意力层和前馈层. 该方法在对象检测、语义分割和实例分割任务上性能与全量微调相当, 且参数量在 1% 至 3%. LoRA Recycle[219] 重用多样的预训练 LoRA, 以实现视觉模型的无调优少样本适配. 通过元学习目标, 从多样的预训练 LoRA 中蒸馏出一个元 LoRA, 使用从预训练 LoRA 自身反向生成的合成数据. 一旦 VFMs 配备了元 LoRA, 就能在单次前向传播中解决新的少量样本任务.

融合方法. NOAH[204] 采用 NAS 算法为不同数据集自动识别最有效的 PEFT, 例如适配器、LoRA 和视觉提示微调, 仅用约 0.4% 的参数. 在 20 多个视觉数据集上进行实验, 发现 NOAH 优于单个提示模块, 具有良好的少样本学习能力且具有领域可推广性.

为探究不同类 PEFT 方法的适用时机, lessons-PEFT[220] 对 Vision Transformers 的代表性 PEFT 方法进行了统一的实证研究. 通过系统地调节超参数, 以公平比较它们在下游任务中的准确性. 研究结果发现, 如果仔细调参, 不同的 PEFT 方法在少样本基准 VTAB-1K 上能够达到相似的准确性, 包括像微调偏置项这样的简单方法, 尽管在之前的研究中它们效果不佳; 其次, 尽管准确性相似, 但是 PEFT 方法在错误判断和高置信预测上存在差异, 这很可能是由于它们不同的归纳偏差造成的, 这种不一致性或互补性为集成方法提供了可能性; 第三, PEFT 在多样本场景下也很有用, 其准确性可与全量微调媲美; 最后, 通过研究 PEFT 方法保留预训练模型 (例如 CLIP) 对于分布偏移的鲁棒性能力, 发现其表现优于单独使用的全量微调. 然而, 通过权重空间集成, 全量微调可以更好地平衡目标分布性能和分布偏移性能, 这为未来研究更鲁棒的 PEFT 指出了一个方向.

发展趋势. PEFT 方法已在视觉基础模型上被广泛验证为一种高效的下游任务迁移学习范式. 针对不同的架构, 添加式、局部、重参数化和融合式各类方法均有被引入其中, PEFT 方法引进较为成熟. 模块设计的轻量化与泛化性并重, 在效果方面, 可以达到极少量参数更新的情况下, 例如小于 1% 参数更新量, 就在图像分类、检测、分割等多种下游任务上达到与全量微调相当甚至更优的性能.

大部分用于视觉基础模型上的高效微调方法在少样本场景下表现相近, 而归纳偏差不同, 这可以促进各个类别方法的融合. 另外, 如果要真正地达到使用一个基础模型完成任意视觉任务, 还需要开发更本质的通用视觉微调方法, 并在此基础上关注软硬件效率协同, 设计硬件友好的 PEFT 方法, 实现从参数高效到计算与能效全栈高效. 最后, 可以在 PEFT 过程中引入可解释性和不确定性估计, 使模型在医疗诊断、自动驾驶等场景中更可信赖. 未来还可以关注如何自动地为不同任务和数据配置最优的 PEFT 方法与超参数, 探索如何将方法原理具有互补性的不同 PEFT 方法结合, 以获得更稳定、更强大的模型. 此外, 如何使 PEFT 后的模型保持预训练模型原有的分布外泛化能力也值得关注.

3.2 语音基础模型

受自然语言处理领域大语言模型巨大成功的启发, 语音基础模型的核心思想是在海量多样化的语音数据上进行预训练, 得到一个强大的、通用的语音表征核心. 在此基础上, 只需通过少量的任务特

定数据对其进行微调, 就能使其轻松适应各种各样的下游任务. 场景包括语音识别、语音合成、情感分析、语种识别、内容理解, 乃至全新的、未曾明确训练过的语音应用. 语音模型具有出色的泛化能力, 在预训练和微调范式下, 在各种下游语音任务中表现出令人印象深刻的性能.

然而, 随着预训练模型规模的增大, 由于计算和存储需求的增加以及过拟合的风险, 微调在实际中变得难以实现. 另外在微调语音基础模型时, 需要考虑语音信号的序列长度与计算复杂度可能带来的训练和推理效率问题; 建模底层的声学特征和高层的语义、语法信息时, 声学和语言信息解耦的问题; 面对不同说话人的口音、语速、情感以及复杂的背景噪声时, 环境泛化和鲁棒性问题; 处理自动语音识别、语音合成、语音情感分析等任务时, 多任务泛化的问题. PEFT 的引入为解决上述问题提供了方案, 且只需要微调少量参数, 从而能够低成本扩展语音基础模型的研究和应用边界. 本文根据方法的核心机制划分为添加式方法和重参数化方法, 对每一类方法的代表性工作进行介绍, 说明其如何具体地在语音模型中实现参数高效微调.

PEFT-SER[221] 在语音情感识别场景中, 评估了 PEFT 中的适配器方法、软提示方法和 LoRA 方法在四个测试集上的效果. 结果表明 LoRA 在情感识别中实现了最佳的微调性能, 同时仅需极少的额外权重参数. Peft-for-speech[222] 探索了不同的 PEFT 方法在语音模型各层的插入位置的不同效果, 统计证据表明, 不同的 PEFT 方法具有不同的学习方式, 并发现通过集成学习将多种 PEFT 方法协同整合, 能够比单独的逐层优化更有效地利用它们的独特学习能力.

添加式方法. ELP-Adapters[65] 使用编码器适配器 (E-adapters)、层适配器 (L-adapters) 和提示适配器 (P-adapter) 这三种类型适配器快速适应各种语音处理任务. 实验表明, 该方法在四个下游任务、五个模型中都具有有效性. 使用 WavLM[223] 模型时, 其性能在所有任务上均与全量微调相当或更好, 同时所需可学习参数减少了 90%. EAT-ASV[93] 提出一种基于适配器的微调方法, 将语音模型适配到语音验证任务, 通过并行的适配器设计在预训练模型中插入两种类型的适配器. 在 VoxCeleb1 数据集上的实验结果表明, 该方法性能良好, 且在仅更新 5% 参数的情况下, 仍能实现卓越性能. TAML-Adapter[94] 利用任务无关的元学习算法, 在微调前初始化适配器的参数. 在 Common Voice 和 Fleurs 数据集上进行的综合实验显示, TAML-Adapter 在五种数据量有限的语音识别中表现出优越的性能. Uni-PET-SPK[92] 一方面在模型中插入两种类型的并行适配器, 允许在中间 Transformer 层调整潜在特征以及从所有 Transformer 层输出嵌入进行适配; 另一方面, 在模型的输入空间中拼接可训练的提示 token 以引导适配. 通过动态可训练的门控机制将二者结合以适应不同的数据集和场景. 该方法在 VoxCeleb、CN-Celeb 和 48 个 UTD 司法数据集上进行实验, 在仅更新 5.4% 的参数情况下实现较好性能和稳定性. EFL-PEFT[224] 将联邦学习、适配器方法与稀疏化方法结合, 通过仅传输部分适配器参数进一步降低通信成本, 实现自动语音识别.

低秩适配方法. HDMoLE[167] 结合 LoRA 专家混合方法和语音错误矫正, 引入分层路由和动态阈值方法到 LoRA 混合专家中, 有效地将单口音 LoRA 专家合并到单个多模态语音错误矫正中, 克服口音多样性挑战, 提升语音识别中的转录效果. MT-LLM[225] 利用 WavLM 和 Whisper 编码器提取对语音特征和语义上下文敏感的多维语音表征, 这些表征随后被输入使用 LoRA 微调的模型中, 使其具备语音理解和转录的能力. 实验显示该方法具有良好表现以及处理基于用户指令的复杂语音任务的潜力. Speech-FT[226] 首先优化微调的设计以减少表征漂移, 然后通过与预训练模型的权重空间插值来恢复跨任务泛化能力. 在 HuBERT[227]、wav2vec 2.0[228]、DeCoAR 2.0[229] 和 WavLM Base+[223] 上的实验表明, Speech-FT 在有监督、无监督和多任务微调场景中都能持续提高性能. 在对 HuBERT 进行自动语音识别微调时, Speech-FT 能将音素错误率从 5.17% 降至 3.94%, 将词错误率从 6.38% 降至 5.75%, 并将说话人识别准确率从 81.86% 提高到 84.11%. Lyra[230] 利用多模态 LoRA 来降低训练成本和数据需求, 同时使用模态正则化器和提取器来强化语音与其他模态之间的关系. 除此之外, 它构建了高质量、庞大的数据集, 包括 150 万条多模态 (语言、视觉、音频) 数据样本和 1.2 万条长语音样本, 使 Lyra 能够处理复杂的长语音输入, 实现更强的全能认知能力. Lyra 在长语音理解、声音理解、跨模态效率及无缝语音交互等各种视觉−语言、视觉−语音和语音−语言基准上实现了良好性能.

发展趋势. 在语音基础模型领域, PEFT 的应用更多围绕着适配器方法、提示方法和基于低秩适配的方法; 同时, 研究趋向于方法的融合与系统性优化. 例如, 将不同类型的适配器或适配器与提示相结合, 将 LoRA 与混合专家模型结合, 设计微调流程, 引入门控设计、元学习设计. 究其原因, 是为

适应语音信号动态特性以及应对数据有限、表征漂移、特征多样性等问题, 所以需要构建更强大的微调方法. 目前大部分研究仅引入部分添加式和重参数化高效微调方法, 未来可以开发专门针对序列模型的 PEFT. 一方面需要关注综合效能, 例如由 PEFT 方法融合带来的计算开销、训练稳定性, 另一方面应该更多关注模态对齐、信息瓶颈和特征泛化能力等, 探索语音与文本、唇读等多模态联合的 PEFT 策略, 提升语音理解的上下文感知能力.

3.3 多模态基础模型

多模态基础模型进一步扩展了大模型的能力, 使其能够同时处理文本、图像等多种模态的信息. 接下来介绍 PEFT 在视觉语言模型和视觉语言动作模型这两种多模态模型上的一些研究工作.

3.3.1 视觉语言模型

预训练的大型视觉语言模型 (VLMs) 是解决各种下游任务的优秀基础模型. 在适配下游任务时, 微调视觉语言模型是一项具有挑战性的任务. 除了微调大模型带来的巨大计算成本、存储开销以及可能导致的遗忘问题等共性问题, 微调视觉语言模型还需要面对如下问题: 少样本泛化任务中辨别力与泛化能力之间的矛盾, 即需要保留通用知识, 又需要微调任务特定知识; 建立图像区域与文本词汇之间精确、细粒度的对应关系的问题; 视觉内容与语言描述之间存在固有的语义差距问题; 模型在微调中可能学会生成流畅但不符合图像的文本, 牺牲准确性换取流畅度, 即幻觉问题; 在不同模态微调模块之间智能分配有限的 PEFT 参数预算, 以实现全局最优问题. 如何在微调中缓解这些问题是一个挑战. PEFT 的引入降低了计算成本和存储开销, 并为解决上述问题提供了探索基础. 本文根据方法的核心机制划分为添加式方法和重参数化方法, 对每一类方法的代表性工作进行方法和效果介绍.

添加式方法. LST[89] 训练一个小且独立的“梯侧”网络, 通过主干网络的梯连接接受中间激活作为输入并进行预测, 减小反向传播带来的计算开销. 在 T5 和 CLIP-T5 上, LST 的可训练参数量在 0.08% ~ 7.46% 之间. 另外, 在添加同样参数量情况下, 存储开销降低 69%, 而其他方法只降低 26%. OVSeg[231] 在输入层面, 通过掩码提示在一系列掩码图像区域及其对应的文本描述上对 CLIP 进行微调. 掩码提示微调在不修改任何 CLIP 权重的情况下带来了显著提升, 并且可以进一步提高全微调模型的表现. 例如在 COCO 上训练并在 ADE20K 150 上评估时, 最佳模型实现 29.6% 的 mIoU. CLIP-adapter[232] 在视觉或语言分支上加入特征适配器, 使用额外的瓶颈层来学习新特征, 并与原始预训练特征进行残差式特征融合. CLIP-adapter 仅微调两个线性层, 即可在各种视觉分类任务上表现出有效性. VL-Prompt[233] 引入两组提示分别用于预训练和微调, 首先有效地预训练视频语言模型, 利用灵活生成的知识感知提示提取动作或对象等关键信息, 然后通过视频语言提示微调模型以生成完整字幕. MMA[234] 将来自不同分支的特征聚合到共享特征空间中, 从而使梯度能够在分支间传递. 模型深层包含特定知识, 而浅层包含更多可泛化知识; 语言特征比视觉特征更具辨别性, 并且两种模态的特征之间存在较大的语义差距, 尤其是在浅层. 因此, 本文仅在 Transformer 的部分深层引入 MMA 模块, 以在辨别力与泛化能力之间实现最佳平衡. 在多视图图像生成任务中, MV-Adapter[235] 通过重复的自注意力层和平行注意力架构设计, 在适配器中引入三维几何知识. 基于这个设计提出一个高效的适配器方法, 能提升文生图模型的效果. MV-Adapter 在 Stable Diffusion XL (SDXL) 上实现了 768 分辨率的多视图生成, 展示了其适应性和多功能性. MH-PEFT[236] 引入数据增强策略到添加式微调中, 利用现有数据生成额外的属性相关信息, 缓解 VLMs 中的幻觉问题. FATE[237] 从 CLIP 的视觉编码器的最后一层提取视觉特征, 并在投影后将其作为参数用于微调 CLIP 语言编码器的每一层. 通过调整这些特征适应参数, 可以直接实现视觉分支与语言分支之间的通信, 从而促进 CLIP 对不同场景的适应. FATE 在 11 个数据集上表现出优异的泛化性能, 同时在额外参数和计算量方面, 对于 $d \times k$ 的视觉信息参数、长度和维度均为 n^* 的视觉提示, FATE 增加的 FLOPs 仅为 $n^*d(k+d^*)$. CLIP-AST[238] 设计自适应选择调优参数方法, 基于优化器中的自适应学习率, 自动选择 VLMs 在下游任务中进行微调的关键参数. 它无需额外参数开销, 且能提升模型性能. 在 13 个基准数据集上, 如 ImageNet、Food-101 等, CLIP-AST 始终优于原始 CLIP 模型及其各类变体. 2SFS[239] 通过选择性地优化任务特定的特征提取器和线性分类器, 使得在小样本的场景下依然可以微调. 该方法在两个实验设置、三种主网络和 11 个数据集上通过固定参数就可以达到良好表现, 具有稳定性.

重参数化方法. CLIP-LoRA[240] 在 VLM 的少样本学习中引入了 LoRA 方法, 并在 11 个数据集上检验. CLIP-LoRA 方法在显著提升性能的同时, 减少了训练时间, 并具有跨数据集和样本数量的超

参数一致性. Prompt-Aware LoRA[241] 针对多视角城市交通视频的描述生成问题, 在专门用于空间推理的视觉语言模型中加入 LoRA 微调, 可以在复杂城市场景下生成稳健的、多视角感知的视频描述. CoDyRA[242] 分析了 LoRA 的秩和位置如何影响学习和遗忘. 基于可塑性和稳定性平衡, 通过稀疏正则自适应优化 LoRA 的秩数, 以减少模型遗忘. CoDyRA 在 CLIP 等实验中验证了其在保留旧知识的同时改进了新的表示, 达到良好效果. LangVision-LoRA-NAS[243] 将神经网络结构搜索与 LoRA 结合, 动态搜索最适合 VLMs 特定多模态任务的 LoRA 秩配置, 平衡性能与计算效率. 通过在多个数据集上使用 LLaMA-3.2-11B 模型进行的大量实验表明, LangVision-LoRA-NAS 在降低微调成本的同时, 展示了模型性能的显著提升. GLAD[244] 将正则化梯度和 LoRA 引入 VLMs, 有效引导微调优化轨迹, 鼓励模型找到对数据分布变化更稳健的参数区域. 通过在 15 个基准数据集上的实验证明, GLAD 在基础类到新类别的泛化、图像域泛化以及跨数据集泛化方面均有良好表现.

发展趋势. 当前, 视觉语言模型的高效微调方法集中于添加式和基于 LoRA 的方法, 但不仅仅是关注冻结和微调, 而是旨在解决多模态适配的核心矛盾. 从通用的参数高效走向精准适配, 关注多模态对齐机制的深化. 针对视觉与语言的语义鸿沟, 从早期的仅在视觉或语言端添加模块发展为双向交互与特征融合. 这表明构建更精细、更直接的跨模态通信通道是提升辨别力与泛化能力的关键. 另外, 视觉语言模型的高效微调还关注了参数、结构、知识动态化调整, 例如根据状态选择关键参数, 用自适应秩数平衡学习与遗忘等. 除此之外, 关于缓解幻觉问题也受到了一定关注, 具体手段包括数据增强和属性约束等.

尽管涌现了大量有效方法, 但是评估体系尚不统一, 现有工作常在各自选定的数据集和任务上验证, 缺乏一个标准化的、覆盖全面下游挑战的评测基准. 因此当前 PEFT 方法通常在特定任务或假设下表现优异, 但一个通用、轻量且能同时解决大部分微调难题的方案尚未出现. 另外, 对于视频、时序数据、高分辨率图像等更复杂场景, PEFT 模块的设计和效率仍需进一步探索.

未来研究在需要面向例如 3D 视觉、视频理解、医学影像等复杂模态与专项任务的 PEFT 设计的同时, 还需要构建多层次、多维度的评估, 涵盖多种任务类型, 并对幻觉、泛化、遗忘等核心挑战开展专项测评.

3.3.2 视觉语言动作模型

视觉语言动作 (VLA) 模型通常在预训练的视觉语言模型基础上构建, 通过在机器人数据上训练大规模视觉语言模型来弥合感知空间与动作空间之间的差距, 扩展任务执行能力、语言遵循能力和语义泛化能力. 需要将语言指令分解为一系列连续的物理动作, 微调必须处理长程的时序依赖和因果关系. 同时, 在仿真环境中微调需要考虑物理引擎、感知噪声与在真实世界部署时的差距. 此外, 机器人试错数据收集缓慢、昂贵且危险, 要求微调必须具有极高的样本效率. 由于 VLAs 在面对新型机器人设置时仍然存在困难, 需要通过微调来获得良好的性能. 为减少微调带来高昂的训练成本, 同时进行有效的微调, PEFT 方法被引入 VLAs 中. 目前主要的用于 VLA 的 PEFT 方法是添加式方法, 接下来对代表性工作进行介绍.

添加式方法. 为解决适配器难以捕捉视频帧与文本词语之间的内在时间关系的问题, READ-PVLA[245] 用递归计算来实现时间建模能力. 同时通过最优传输来维护流入 READ 模块的任务相关信息. 实验表明该方法在多个低资源时间语言定位和视频语言基准上具有良好的微调表现. VLA-Adapter[246] 探索了对于连接感知与动作空间至关重要的关键条件. 基于这些条件设计了 Policy 模块, 可自主地将最优条件注入动作空间. 通过这种方式, VLA-Adapter 仅使用了一个 5 亿参数的主网络, 且无需任何机器人数据预训练. 在模拟和真实机器人基准上的实验表明, VLA-Adapter 不仅表现良好, 而且具有很快的推理速度. 此外, 它能够在单个消费级 GPU 上仅用 8 小时就训练出一个强大的 VLA 模型, 从而大大降低了部署 VLA 模型的门槛. 为促进并利用丰富多样的机器人数据源中的异构性, X-VLA[247] 为每个不同的数据源引入独立的可学习嵌入, 通过软提示方法使 VLA 模型能够有效利用不同具身之间的特征. 该方法在 6 个模拟环境以及 3 个真实机器人上进行评估, 其 0.9B 版本在一系列基准测试中展现了先进的性能, 具有动作能力到跨形体、环境和任务的多方面快速适应能力. OpenVLA-OFT[248] 将并行解码、动作分块、连续动作表示以及基于简单 L_1 回归的学习目标集成在一起, 以全面提升推理效率、策略性能和模型输入输出规范的灵活性. 在 LIBERO 仿真基准上取得领先效果, 在四个任务套件中将 OpenVLA 的平均成功率从 76.5% 显著提升至 97.1%, 同时动作生成吞吐量提升 26 倍. 在真实世界的评估中, 该方法使 OpenVLA 能够在双手 ALOHA 机器人上成功执行灵巧的高频控制任务, 并且在平均成功率上比使用默认

微调方案的其他 VLA (π_0 和 RDT-1B) 以及从零训练的强模仿学习策略 (Diffusion Policy 和 ACT) 高出最多 15%.

发展趋势. 在多模态模型领域, PEFT 需要平衡多个模态, 应对模态特征之间的分布差异、时序关系、数据异构等问题. 相比于单模态的适配, PEFT 不仅需要作为高效的特征适配工具, 还需要考虑模态之间关系协调的策略, 甚至融合感知与物理世界动作空间构建的桥梁. 视觉语言动作模型需要将高级语义理解转化为精确的物理动作, 是实现通用具身智能的核心. 当前 PEFT 研究不再局限于微调预训练骨干网络, 而是通过设计精巧的轻量级策略模块, 主动构建连接语义理解与动作生成的桥梁, 同时利用异构数据实现快速跨域适应. PEFT 为解决机器人数据来源多样且难以整合的问题, 提供了知识融合与迁移接口.

不过目前用于视觉语言动作模型的 PEFT 方法在广度和深度上还有待开发. 另外, 微调面临数据稀缺、时序依赖、仿真到真实鸿沟等多重挑战, 具有很大的难度. 同时, 现有方法对于需要复杂物理推理、多步骤规划、试错探索的任务可能仍显不足. 仿真到真实迁移的鲁棒性尚未根本解决, 尽管 PEFT 提升了样本效率, 但其微调过程仍严重依赖于仿真数据或有限的真实数据. 安全性与失败恢复机制关注不足, 缺乏对微调后模型的决策安全性、风险规避能力以及在异常状态下自主恢复能力的专门设计与评估.

未来可以更深入地嵌入物理定律、动力学和因果作为先验, 设计具有内在物理一致性和因果推理能力的适配模块; 将安全约束 (如工作空间限制) 可解释性目标 (如注意力在关键物体上的可视化) 直接作为正则化项或损失函数整合进 PEFT 的训练框架中, 引导模型学习更可靠、更透明的决策策略; 将复杂的长期任务分解为可重复使用的轻量级 PEFT 模块, 通过组合调用不同的模块达成目标适应.

3.4 联邦基础模型

PEFT 通常假设大语言模型是在单个设备或单个客户端的数据上进行训练的. 然而, 实际应用场景中, 往往需要利用分布在多个设备上的私有数据对这些模型进行微调. 联邦学习 (FL) 为此提供了一种具有吸引力的解决方案, 它能够在训练过程中有效保护用户隐私, 使敏感数据始终保留在本地设备上. 联邦基础模型不受限于模型的处理模态, 可以是语言、视觉或多模态场景. 它作为一种隐私保护方法, 可在联邦学习环境下通过协同利用多个来源的非独立同分布数据集, 实现对模型的联合微调.

PEFT 在联邦基础模型上的应用有几个挑战. 首先是可能存在数据异构性问题, 如果各客户端数据分布差异极大, 会导致单一的微调模型难以适配所有客户端. 其次是通信瓶颈问题, 若在客户端与服务器之间频繁传输整个大模型的梯度或参数, 通信开销无法承受. 最后隐私与安全也是一个挑战, 微调过程可能泄露客户端的本地数据信息, 如何在有效协作与隐私保护之间取得平衡是根本挑战.

对于目前在联邦基础模型中引入的参数高效微调策略, 本文根据方法的核心机制将其划分为添加式方法 (如基于适配器的策略[249−253]、基于提示的策略[254−257])、局部方法[258−260]、重参数化方法 (主要是基于 LoRA 的策略[261−263]). 接下来对每一类方法的代表性工作进行方法和效果介绍.

添加式方法. C2A[249] 利用超网络生成为客户端定制的适配器, 以解决数据异质性问题. FedAdapter[250] 设计具有自动调节深度和宽度的渐进式适配器, 缓解了通信效率和计算异质性问题. FedDAT[251] 设计双适配器结构来进行知识蒸馏, 缓解了通信效率和数据异质性问题. Pilot[252] 提出用于多模态的跨任务的两阶段适配器混合架构, 进一步缓解了数据异质性问题. FedPIA[253] 关注 VLMs 在联邦学习场景下的高效微调. 它通过引入服务器端本地适配器的置换集成与客户端全局适配器的优化融合, 改进了联邦学习与 PEFT 的简单组合, 在五种不同医疗视觉–语言模型的联邦学习任务、超过 2000 次客户端级别实验中表现出良好效果.

GF-PT[254] 通过无梯度的离散局部搜索降低了本地训练的计算开销, 同时通过 token 压缩降低通信开销. pFedPG[255] 通过面向客户端的提示生成器根据优化方向生成个性化提示, 缓解了计算和数据异质性问题. Powder[256] 通过两阶段聚合的提示生成, 提出联邦持续学习算法, 可以有效促进由提示封装的知识在各类顺序学习任务和客户端之间的转移. Diprompt[257] 设计了两种类型的提示, 全局提示用于捕获所有客户端的通用知识, 区域提示用于捕获特定域的知识, 缓解了数据异质性问题.

局部方法. FedPCL[258] 使用类别原型作为信息载体, 并在本地更新期间进行对比学习新算法, 这使得客户端能够共享来自本地和全局原型的更多类别相关知识. FedPETuning[259] 为四种代表性 PEFT 方法开发了相应的联邦基准, 其中 FedBF 将局部 PEFT 中的 BitFit 方法引入联邦学习场景, 有效防御了隐私攻击. FedRA[260] 在每一轮通信中随机生成一个分配矩阵. 对于资源受限的客户端, 它根据分配矩阵重新组织原始模型中的少量层, 并使用适配器进行微调. 随后, 服务器根据当前分

配矩阵将客户端更新的适配器参数聚合到原始模型的对应层中, 缓解了计算异质性问题.

重参数化方法. FedDPA[261] 设计了一个双适配器结构, 全局适配器和本地适配器共同应对测试时间分布变化和客户端特定个性化. 另外, 它还引入一种实例级动态加权机制, 在推理过程中为每个测试实例动态整合全局和本地适配器, 从而缓解测试环境和训练环境的时间分布差异带来的模型能力下降问题. FedTT[262] 通过将张量化的适配器集成到客户端模型的编码器或解码器模块中, 使得大语音模型可以进行跨组织联邦学习 (cross-silo FL) 和大规模跨设备联邦学习 (cross-device FL). 为更好地跨组织联邦学习, 解决联邦学习场景下的通信开销和数据异质性问题, FedTT+[262] 扩展了 FedTT, 通过自适应冻结部分张量因子, 提高了对数据异质性的鲁棒性, 并进一步减少了可训练参数数量. 该方法能够成功应对数据异质性挑战, 并且在性能上与现有联邦 PEFT 方法相当甚至更优, 同时实现高达 10 倍的通信成本降低. LoRA-FAIR[263] 将 LoRA 与联邦学习结合起来, 使得适配过程所需要的大量数据可以在隐私保护的协作框架下获得. 另外, 它通过在服务器端引入修正项, 同时解决了低秩适配和联邦学习结合的两个关键挑战, 即服务器端聚合偏差和客户端初始化滞后, 从而提高了聚合效率和准确性, 保持了计算和通信效率.

发展趋势. 联邦学习与参数高效微调的融合, 为解决大模型在隐私敏感、数据分散场景下的应用提供了关键路径. 同时 PEFT 从一种单纯提升效率的工具演变为构建下一代隐私保护协同智能的技术. 面对数据异构性、通信瓶颈、隐私保护等难点和不同模态基础模型微调时的特点, 目前一些工作已经通过多适配器等方式在上述问题中有一定的突破: 从追求单一全局模型转向构建全局共享加本地专属的混合架构, 以应对数据异构性; 以 LoRA 为代表的重参数化方法成为主流, 结合参数压缩与通信机制创新, 大幅降低联邦通信开销; 从文本单模态任务快速拓展至多模态、跨领域的复杂场景应用; 综合考虑客户端算力、存储等资源差异, 实现算法与系统的协同优化.

然而, 这一领域仍面临多重挑战: 不同客户端之间知识迁移的边界仍需明确定义; 现有技术对跨领域、跨模态的极端异构场景适配能力仍需提高; 隐私保护机制虽然取得进展, 但距离实现完全的隐私安全保障仍有距离; 同时, 现有框架对持续学习和任务演化的支持稍显不足, 对适应真实世界中不断变化的应用需求有待提高.

展望未来, 联邦 PEFT 技术可以通过深化理论创新来指导个性化与泛化的平衡. 隐私计算技术的深度融合也可能是趋势, 技术架构需要向更灵活、可组合且安全的方向演进. 同时需要关注计算能力、网络状况、存储空间和任务目标多维度的联邦计算, 动态化 PEFT, 从而应对不同客户端. 另外在现实应用层面, 这项技术可以在医疗健康、个性化服务、智能制造、城市治理等多个领域发挥作用, 它将使跨机构、跨地域的智能协作成为可能.

4 结束语

大模型的参数高效微调是一个发展迅速的研究领域, 旨在通过低存储和计算成本完成模型在下游任务中的适配. 其核心价值在于成功解耦了模型知识展现与任务适配成本, 通过极小的参数代价来实现对预训练模型中已有知识的高效引导与重组. 该技术已成为释放大规模预训练模型潜能的关键途径, 其发展呈现出多元化、精细化与系统化的鲜明演进脉络, 从早期的适配器、提示微调, 到后续的低秩适配及其众多衍生方法, PEFT 的发展轨迹清晰地展现出从单纯追求参数效率到综合考量性能、效率与鲁棒性的演进脉络.

从方法论层面看, PEFT 已形成添加式、局部式、重参数化式与融合式四大主流范式. 同时从技术演进维度看, PEFT 研究呈现以下趋势: 1) 从静态到动态, 例如 AdaLoRA、SoRA 等实现了秩的动态调整, MoE 路由机制实现了模块的动态激活; 2) 从粗粒度到精细化, 例如参数选择或初始化策略从基于启发式规则发展到基于 Fisher 信息、奇异值分解等数学理论指导; 3) 从独立到协同, 例如 LoRA 等模块的即插即用特性催生了庞大的“插件生态”, 支持多模块组合与知识共享; 4) 从算法到系统, 例如量化技术与并行技术解决了大模型部署的实际工程挑战. 本文从方法思想、理论基础、设计结构、策略优势、性能表现、应用场景等方面对现有的 PEFT 技术进行分类和梳理, 从多个视角层层递进地对 PEFT 方法的研究现状以及发展趋势进行介绍和分析.

基于对当前参数高效微调技术发展脉络的梳理, 本文认为该领域正从单一方法的创新走向系统性优化. 为推动下一阶段的突破, 未来的研究应聚焦于以下几个最具潜力的核心方向.

1) 理论深化与可解释性

目前 PEFT 技术很大程度上仍建立在经验有效性之上, 需要坚实、统一的理论基础来解释其为何有效. 实际上, 目前尚不清楚 PEFT 究竟捕获了

何种任务特定知识, 以及不同的参数介入位置 (如 Q/K/V 层或前馈层) 为何会产生显著的性能差异. 缺乏对其为何有效, 参数影响力、低秩空间或稀疏性如何捕获任务特定知识, 以及其与全量微调在损失景观上有何异同等根本问题的理论解释.

未来研究可深入探索 PEFT 的损失景观特性, 分析其与全量微调在优化轨迹上的异同. 利用表示学习、优化动力学、信息论等工具量化 PEFT 引入的信息量, 或从表示学习视角剖析适配器如何改变模型的内部表征. 建立理论模型以阐明 PEFT 方法的有效秩、超参敏感性与模型容量、任务复杂度之间的内在关联. 理论的突破将为 PEFT 方法的设计、超参的自动化配置以及介入位置的选择提供根本性的指导原则, 摆脱当前依赖大量实验试错的困境, 实现从经验到科学的范式转变.

2) 自动化与统一框架

随着文本、视觉、语音等模态和大模型架构的多样化, 为每个新模型和任务手动设计或选择 PEFT 方案变得难以持续. 首先, 当前方法存在模态鸿沟和架构鸿沟, 且大多数 PEFT 为单一目标, 真实场景还涉及延迟、隐私、公平等. 其次, 持续学习中的遗忘问题、微调后在小数据集上的校准问题也需要关注. 因此, 需要一个能够通用于不同场景的、智能化的统一框架.

研究跨模态的统一 PEFT 框架是关键, 例如设计能够动态适应不同输入模态的通用适配器. 同时, 大力发展自动化 PEFT 技术, 借鉴神经架构搜索和元学习的思想, 开发能够根据任务描述、数据特性和模型结构, 自动识别并优先适配最关键模态的 PEFT 方法; 探索能在准确性、延迟、存储、隐私之间权衡的方法; 研究在不断变化的数据分布下的在线更新 PEFT, 缓解遗忘问题. 从而极大降低 PEFT 技术的使用门槛, 使研究者与工程师能快速、高效地将大模型适配到任意下游任务与模态, 真正释放大模型的泛化潜力, 推动其规模化应用.

3) 面向边缘计算的优化

将搭载 PEFT 的大模型部署于手机、笔记本等资源严格受限的边缘端时, 不仅要求参数高效, 更追求推理延迟、内存占用和能耗的全面极致优化. 单纯的 PEFT 技术已不足以满足这些约束.

因此, 需要推动 PEFT 与量化、剪枝、蒸馏等模型压缩技术从混合应用走向深度联合优化. 例如, 探索直接训练低精度的 LoRA 适配器, 或开发硬件感知的 PEFT 算法, 使其低秩模式、稀疏掩码等结构设计能够与特定硬件的计算特性和内存层次结构协同. 如此便可实现大模型在终端设备上的实时、常驻智能, 为离线翻译、个性化语音助手、实时视觉分析等应用扫清障碍, 这是推动 AI 普适化的关键技术路径.

总体而言, PEFT 的研究正从如何更高效走向如何更智能、更通用、更安全, 其发展必将是一个融合理论、算法、系统、硬件和安全的跨学科进程, 以持续推动大模型技术在更多样化场景中的高效、可靠应用.

参考文献

1 Li J Y, Tang T Y, Zhao W X, Nie J Y, Wen J R. Pre-trained language models for text generation: A survey. *ACM Computing Surveys*, 2024, **56**(9): Article No. 230

2 Xia X, Zhang D, Liao Z B, Hou Z Y, Sun T R, Li J, et al. SceneGenAgent: Precise industrial scene generation with coding agent. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vienna, Austria: Association for Computational Linguistics, 2025. 17847−17875

3 Zhu W H, Liu H Y, Dong Q X, Xu J J, Huang S J, Kong L P, et al. Multilingual machine translation with large language models: Empirical results and analysis. In: Proceedings of the Findings of the Association for Computational Linguistics. Mexico City, Mexico: Association for Computational Linguistics, 2024. 2765−2781

4 Zheng L M, Chiang W L, Sheng Y, Zhuang S Y, Wu Z H, Zhuang Y H, et al. Judging LLM-as-a-judge with MT-bench and chatbot arena. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc., 2023. Article No. 2020

5 Kim J K, Chua M, Rickard M, Lorenzo A. ChatGPT and large language model (LLM) chatbots: The current state of acceptability and a proposal for guidelines on utilization in academic medicine. *Journal of Pediatric Urology*, 2023, **19**(5): 598−604

6 Wang J J. The power of AI-assisted diagnosis. *EAI Endorsed Transactions on e-Learning*, 2022, **8**(4): Article No. e3

7 Biswas S S. Role of ChatGPT in public health. *Annals of Biomedical Engineering*, 2023, **51**(5): 868−869

8 Li H T, Chen J J, Yang J L, Ai Q Y, Jia W, Liu Y F, et al. LegalAgentBench: Evaluating LLM agents in legal domain. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vienna, Austria: Association for Computational Linguistics, 2025. 2322−2344

9 Xing F. Designing heterogeneous LLM agents for financial sentiment analysis. *ACM Transactions on Management Information Systems*, 2025, **16**(1): Article No. 5

10 Han Z Y, Gao C, Liu J Y, Zhang J, Zhang S Q. Parameter-efficient fine-tuning for large models: A comprehensive survey. arXiv preprint arXiv: 2403.14608, 2024.

11 Guo D M, Rush A, Kim Y. Parameter-efficient transfer learning with diff pruning. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Virtual Event: Association for Computational Linguistics, 2021. 4884−4896

12 Hu E, Shen Y L, Wallis P, Allen-Zhu Z, Li Y Z, Wang S A, et al. LoRA: Low-rank adaptation of large language models. arXiv preprint arXiv: 2106.09685, 2021.

13 Lialin V, Deshpande V, Rumshisky A. Scaling down to scale up: A guide to parameter-efficient fine-tuning. arXiv preprint arXiv: 2303.15647, 2023.

14 Vucetic D, Tayaranian M, Ziaeefard M, Clark J J, Meyer B H, Gross W J. Efficient fine-tuning of BERT models on the edge. In: Proceedings of the IEEE International Symposium on Circuits and Systems (ISCAS). Austin, USA: IEEE, 2022. 1838−1842

15 Fu Z H, Yang H R, So A M C, Lam W, Bing L D, Collier N. On the effectiveness of parameter-efficient fine-tuning. In: Proceedings of the 37th AAAI Conference on Artificial Intelligence. Washington, USA: AAAI Press, 2023. 12799−12807

16 Wang L P, Chen S, Jiang L N, Pan S, Cai R Z, Yang S, et al. Parameter-efficient fine-tuning in large models: A survey of methodologies. arXiv preprint arXiv: 2410.19878, 2024.

17 Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, et al. Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc., 2017. 6000−6010

18 Devlin J, Chang M W, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional Transformers for language understanding. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, USA: Association for Computational Linguistics, 2019. 4171−4186

19 Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, Aleman F L, et al. GPT-4 technical report. arXiv preprint arXiv: 2303.08774, 2023.

20 Wei J, Tay Y, Bommasani R, Raffel C, Zoph B, Borgeaud S, et al. Emergent abilities of large language models. arXiv preprint arXiv: 2206.0768, 2022.

21 Ouyang L, Wu J, Jiang X, Almeida D, Wainwright C L, Mishkin P, et al. Training language models to follow instructions with human feedback. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc., 2022. Article No. 2011

22 Schulman J, Wolski F, Dhariwal P, Radford A, Klimov O. Proximal policy optimization algorithms. arXiv preprint arXiv: 1707.06347, 2017.

23 Rafailov R, Sharma A, Mitchell E, Ermon S, Manning C D, Finn C. Direct preference optimization: Your language model is secretly a reward model. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc., 2023. Article No. 2338

24 Lee H, Phatale S, Mansoor H, Lu K, Mesnard T, Bishop C, et al. RLAIF: Scaling reinforcement learning from human feedback with AI feedback. arXiv preprint arXiv: 2309.00267, 2023.

25 Radford A, Narasimhan K, Salimans T, Sutskever I. Improving language understanding by generative pre-training [Online], available: https://www.mikecaptain.com/resources/pdf/GPT-1.pdf, July 1, 2026

26 Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. Language models are unsupervised multitask learners. *OpenAI Blog*, 2019, **1**(8): Article No. 9

27 Brown T B, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language models are few-shot learners. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc., 2020. Article No. 159

28 Touvron H, Lavril T, Izacard G, Martinet X, Lachaux M A, Lacroix T, et al. LLaMA: Open and efficient foundation language models. arXiv preprint arXiv: 2302.13971, 2023.

29 Vavekanand R, Sam K. LLaMA 3.1: An in-depth analysis of the next-generation large language model [Online], available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=6139407, July 1, 2026

30 Bi X, Chen D L, Chen G T, Chen S H, Dai D M, Deng C Q, et al. DeepSeek LLM: Scaling open-source language models with longtermism. arXiv preprint arXiv: 2401.02954, 2024.

31 Liu A X, Feng B, Wang B, Wang B X, Liu B, Zhao C G, et al. DeepSeek-v2: A strong, economical, and efficient mixture-of-experts language model. arXiv preprint arXiv: 2405.04434, 2024.

32 Liu A X, Feng B, Xue B, Wang B X, Wu B C, Lu C D, et al. DeepSeek-v3 technical report. arXiv preprint arXiv: 2412.19437, 2024.

33 Guo D Y, Yang D J, Zhang H W, Song J X, Zhang R Y, Xu R X, et al. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. arXiv preprint arXiv: 2501.12948, 2025.

34 Bai J Z, Bai S, Chu Y F, Cui Z Y, Dang K, Deng X D, et al. Qwen technical report. arXiv preprint arXiv: 2309.16609, 2023.

35 Yang A, Li A F, Yang B S, Zhang B C, Hui B Y, Zheng B, et al. Qwen3 technical report. arXiv preprint arXiv: 2505.09388, 2025.

36 Ahmed I, Islam S, Datta P P, Kabir I, Chowdhury N U R, Haque A. Qwen 2.5: A comprehensive review of the leading resource-efficient LLM with potential to surpass all competitors [Online], available: https://www.techrxiv.org/doi/pdf/10.36227/techrxiv.174060306.65738406/v1, July 1, 2026

37 Caruccio L, Cirillo S, Polese G, Solimando G, Sundaramurthy S, Tortora G. Claude 2.0 large language model: Tackling a real-world classification problem with a new iterative prompt engineering approach. *Intelligent Systems With Applications*, 2024, **21**: Article No. 200336

38 Anil R, Borgeaud S, Alayrac J B, Yu J H, Soricut R, Schalkwyk J, et al. Gemini: A family of highly capable multimodal models. arXiv preprint arXiv: 2312.11805, 2023.

39 Groeneveld D, Beltagy I, Walsh E, Bhagia A, Kinney R, Tafjord O, et al. OLMo: Accelerating the science of language models. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Bangkok, Thailand: Association for Computational Linguistics, 2024. 15789−15809

40 Radford A, Kim J W, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. Learning transferable visual models from natural language supervision. In: Proceedings of the 38th International Conference on Machine Learning (ICML). Virtual Event: PMLR, 2021. 8748−8763

41 Sun Q, Fang Y X, Wu L, Wang X L, Cao Y. EVA-CLIP: Improved training techniques for CLIP at scale. arXiv preprint arXiv: 2303.15389, 2023.

42 Kaplan J, McCandlish S, Henighan T, Brown T B, Chess B, Child R, et al. Scaling laws for neural language models. arXiv preprint arXiv: 2001.08361, 2020.

43 Hoffmann J, Borgeaud S, Mensch A, Buchatskaya E, Cai T, Rutherford E, et al. Training compute-optimal large language models. arXiv preprint arXiv: 2203.15556, 2022.

44 Wang A, Singh A, Michael J, Hill F, Levy O, Bowman S R. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In: Proceedings of the EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP. Brussels, Belgium: Association for Computational Linguistics, 2018. 353−355

45 Mihaylov T, Clark P, Khot T, Sabharwal A. Can a suit of armor conduct electricity? A new dataset for open book question

answering. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium: Association for Computational Linguistics, 2018. 2381−2391

46 Bisk Y, Zellers R, le Bras R, Gao J F, Choi Y. PIQA: Reasoning about physical commonsense in natural language. In: Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York, USA: AAAI Press, 2020. 7432−7439

47 Sap M, Rashkin H, Chen D, le Bras R, Choi Y. Social IQa: Commonsense reasoning about social interactions. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong, China: Association for Computational Linguistics, 2019. 4463−4473

48 Zellers R, Holtzman A, Bisk Y, Farhadi A, Choi Y. HellaSwag: Can a machine really finish your sentence? In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: Association for Computational Linguistics, 2019. 4791−4800

49 Clark C, Lee K, Chang M W, Kwiatkowski T, Collins M, Toutanova K. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, USA: Association for Computational Linguistics, 2019. 2924−2936

50 Sakaguchi K, le Bras R, Bhagavatula C, Choi Y. WinoGrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 2021, **64**(9): 99−106

51 Clark P, Cowhey I, Etzioni O, Khot T, Sabharwal A, Schoenick C, et al. Think you have solved question answering? Try ARC, the AI2 reasoning challenge. arXiv preprint arXiv: 1803.05457, 2018.

52 Kay W, Carreira J, Simonyan K, Zhang B, Hillier C, Vijayanarasimhan S, et al. The kinetics human action video dataset. arXiv preprint arXiv: 1705.06950, 2017.

53 Lin T Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, et al. Microsoft COCO: Common objects in context. In: Proceedings of the 13th European Conference on Computer Vision (ECCV). Zurich, Switzerland: Springer, 2014. 740−755

54 Zhou B L, Zhao H, Puig X, Fidler S, Barriuso A, Torralba A. Scene parsing through ADE20K dataset. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE, 2017. 5122−5130

55 Everingham M, van Gool L, Williams C K I, Winn J, Zisserman A. The PASCAL visual object classes (VOC) challenge. *International Journal of Computer Vision*, 2010, **88**(2): 303−338

56 Houlsby N, Giurgiu A, Jastrzebski S, Morrone B, de Laroussilhe Q, Gesmundo A, et al. Parameter-efficient transfer learning for NLP. In: Proceedings of the 36th International Conference on Machine Learning (ICML). Long Beach, USA: PMLR, 2019. 2790−2799

57 Pfeiffer J, Kamath A, Rücklé A, Cho K, Gurevych I. AdapterFusion: Non-destructive task composition for transfer learning. In: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. Virtual Event: Association for Computational Linguistics, 2021. 487−503

58 Lin Z J, Madotto A, Fung P. Exploring versatile generative language model via parameter-efficient transfer learning. In: Proceedings of the Findings of the Association for Computational Linguistics. Virtual Event: Association for Computational Linguistics, 2020. 441−459

59 He J X, Zhou C T, Ma X Z, Berg-Kirkpatrick T, Neubig G. Towards a unified view of parameter-efficient transfer learning. In: Proceedings of the 10th International Conference on Learning Representations (ICLR). Virtual Event: OpenReview.net, 2022. 1−15

60 Zhu Y M, Feng J T, Zhao C Q, Wang M X, Li L. Counter-interference adapter for multilingual machine translation. In: Proceedings of the Findings of the Association for Computational Linguistics. Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021. 2812−2823

61 Lei T, Bai J W, Brahma S, Ainslie J, Lee K, Zhou Y Q, et al. Conditional adapters: Parameter-efficient transfer learning with fast inference. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc., 2023. Article No. 357

62 Chen Y Y, Fu Q, Fan G, Du L, Lou J G, Han S, et al. Hadamard adapter: An extreme parameter-efficient adapter tuning method for pre-trained language models. In: Proceedings of the 32nd ACM International Conference on Information and Knowledge Management. Birmingham, UK: ACM, 2023. 276−285

63 He S, Fan R Z, Ding L, Shen L, Zhou T Y, Tao D C. MerA: Merging pretrained adapters for few-shot learning. arXiv preprint arXiv: 2308.15982, 2023.

64 Mahabadi R K, Ruder S, Dehghani M, Henderson J. Parameter-efficient multi-task fine-tuning for Transformers via shared hypernetworks. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Virtual Event: Association for Computational Linguistics, 2021. 565−576

65 Inoue N, Otake S, Hirose T, Ohi M, Kawakami R. ELP-adapters: Parameter efficient adapter tuning for various speech processing tasks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024, **32**: 3867−3880

66 Petrov A, Torr P H S, Bibi A. When do prompting and prefix-tuning work? A theory of capabilities and limitations. In: Proceedings of the 12th International Conference on Learning Representations (ICLR). Vienna, Austria: OpenReview.net, 2024. 1−24

67 Li X L, Liang P. Prefix-tuning: Optimizing continuous prompts for generation. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Virtual Event: Association for Computational Linguistics, 2021. 4582−4597

68 Lester B, Al-Rfou R, Constant N. The power of scale for parameter-efficient prompt tuning. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021. 3045−3059

69 Liu X, Zheng Y N, Du Z X, Ding M, Qian Y J, Yang Z L, et al. GPT understands, too. *AI Open*, 2024, **5**: 208−215

70 Liu X, Ji K X, Fu Y C, Tam W L, Du Z X, Yang Z L, et al. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. arXiv preprint arXiv: 2110.07602, 2021.

71 Choi J Y, Kim J, Park J H, Mok W L, Lee S K. SMoP: Towards efficient and effective prompt tuning with sparse mixture-of-prompts. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Singapore: Association for Computational Linguistics, 2023. 14306−14316

72 Zhang Z R, Tan C Q, Xu H Y, Wang C Y, Huang J, Huang S F. Towards adaptive prefix tuning for parameter-efficient language model fine-tuning. In: Proceedings of the 61st Annual

Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Toronto, Canada: Association for Computational Linguistics, 2023. 1239−1248

73 Wang Q F, Mao Y N, Wang J G, Yu H C, Nie S L, Wang S N, et al. APrompt: Attention prompt tuning for efficient adaptation of pre-trained language models. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Singapore: Association for Computational Linguistics, 2023. 9147−9160

74 Wu Z F, Wang S N, Gu J T, Hou R, Dong Y X, Vydiswaran V G V, et al. IDPG: An instance-dependent prompt generation method. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Seattle, USA: Association for Computational Linguistics, 2022. 5507−5521

75 Zhu W, Tan M. SPT: Learning to selectively insert prompts for better prompt tuning. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Singapore: Association for Computational Linguistics, 2023. 11862−11878

76 Liu X Y, Sun T X, Huang X J, Qiu X P. Late prompt tuning: A late prompt could be better than many prompts. In: Proceedings of the Findings of the Association for Computational Linguistics. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, 2022. 1325−1338

77 Ma F, Zhang C, Ren L, Wang J G, Wang Q F, Wu W, et al. XPrompt: Exploring the extreme of prompt tuning. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, 2022. 11033−11047

78 Shi Z X, Lipani A. DePT: Decomposed prompt tuning for parameter-efficient fine-tuning. In: Proceedings of the 12th International Conference on Learning Representations (ICLR). Vienna, Austria: OpenReview.net, 2024. 1−21

79 Wu J D, Yu T, Wang R, Song Z, Zhang R Y, Zhao H D, et al. InfoPrompt: Information-theoretic soft prompt tuning for natural language understanding. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc., 2023. Article No. 2669

80 Chen L C. PTP: Boosting Stability and Performance of Prompt Tuning With Perturbation-based Regularizer [Master thesis], University of Pittsburgh, USA, 2023.

81 Zeng R J, Han C, Wang Q F, Wu C S, Geng T, Huang L F, et al. Visual Fourier prompt tuning. In: Proceedings of the 38th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc., 2024. Article No. 180

82 He A L, Wu Y L, Wang Z H, Li T, Fu H Z. DVPT: Dynamic visual prompt tuning of large pre-trained models for medical image analysis. *Neural Networks*, 2025, **185**: Article No. 107168

83 Xiao Z H, Yan S L, Hong J, Cai J Y, Jiang X L, Hu Y, et al. DynaPrompt: Dynamic test-time prompt tuning. In: Proceedings of the 13th International Conference on Learning Representations (ICLR). Singapore: OpenReview.net, 2025. 1−17

84 Li Z P, Lin M H, Wang J, Wang S H. Fairness-aware prompt tuning for graph neural networks. In: Proceedings of the ACM on Web Conference. Sydney, Australia: ACM, 2025. 3586−3597

85 Liu H K, Tam D, Muqeeth M, Mohta J, Huang T H, Bansal M, et al. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc., 2022. Article No. 142

86 Lian D Z, Zhou D Q, Feng J S, Wang X C. Scaling & shifting your features: A new baseline for efficient model tuning. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc., 2022. Article No. 9

87 Yang X C, Huang J Y, Zhou W X, Chen M H. Parameter-efficient tuning with special token adaptation. In: Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics. Dubrovnik, Croatia: Association for Computational Linguistics, 2023. 865−872

88 Lu X M, Brahman F, West P, Jung J, Chandu K, Ravichander A, et al. Inference-time policy adapters (IPA): Tailoring extreme-scale LMs without fine-tuning. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Singapore: Association for Computational Linguistics, 2023. 6863−6883

89 Sung Y L, Cho J, Bansal M. LST: Ladder side-tuning for parameter and memory efficient transfer learning. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc., 2022. Article No. 944

90 Cao J, Prakash C S, Hamza W. Attention Fusion: A light yet efficient late fusion mechanism for task adaptation in NLU. In: Proceedings of the Findings of the Association for Computational Linguistics. Seattle, USA: Association for Computational Linguistics, 2022. 857−866

91 Savadikar C, Song X, Wu T F. WeGeFT: Weight-generative fine-tuning for multi-faceted efficient adaptation of large models. In: Proceedings of the 42nd International Conference on Machine Learning (ICML). Vancouver, Canada: PMLR, 2025. 53019−53041

92 Sang M F, Hansen J H L. UniPET-SPK: A unified framework for parameter-efficient tuning of pre-trained speech models for robust speaker verification. *IEEE Transactions on Audio, Speech and Language Processing*, 2025, **33**: 1402−1414

93 Sang M F, Hansen J H L. Efficient adapter tuning of pre-trained speech models for automatic speaker verification. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Seoul, South Korea: IEEE, 2024. 12131−12135

94 Liu Y P, Yang X K, Zhang J Y, Xi Y L, Qu D. TAML-adapter: Enhancing adapter tuning through task-agnostic meta-learning for low-resource automatic speech recognition. *IEEE Signal Processing Letters*, 2025, **32**: 636−640

95 Zhao M J, Lin T, Mi F, Jaggi M, Schütze H. Masking as an efficient alternative to finetuning for pretrained language models. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). Virtual Event: Association for Computational Linguistics, 2020. 2226−2241

96 Ben-Zaken E, Ravfogel S, Goldberg Y. BitFit: Simple parameter-efficient fine-tuning for Transformer-based masked language-models. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Dublin, Ireland: Association for Computational Linguistics, 2022. 1−9

97 Lawton N, Kumar A, Thattai G, Galstyan A, Steeg G V. Neural architecture search for parameter-efficient fine-tuning of large pre-trained language models. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics. Toronto, Canada: Association for Computational Linguistics, 2023. 8506−8515

98 Liao B H, Meng Y, Monz C. Parameter-efficient fine-tuning without introducing new latency. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Toronto, Canada: Association for Computational Linguistics, 2023. 4242−4260

99 Sung Y L, Nair V, Raffel C. Training neural networks with fixed sparse masks. In: Proceedings of the 35th International Conference on Neural Information Processing Systems. Virtual Event: Curran Associates Inc., 2021. Article No. 1852

100 Das S S S, Zhang R H, Shi P, Yin W P, Zhang R. Unified low-resource sequence labeling by sample-aware dynamic sparse finetuning. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Singapore: Association for Computational Linguistics, 2023. 6998−7010

101 Ansell A, Ponti E M, Korhonen A, Vulić I. Composable sparse fine-tuning for cross-lingual transfer. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Dublin, Ireland: Association for Computational Linguistics, 2022. 1778−1796

102 Frankle J, Carbin M. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In: Proceedings of the 7th International Conference on Learning Representations (ICLR). New Orleans, USA: OpenReview.net, 2019. 1−42

103 Malach E, Yehudai G, Shalev-Shwartz S, Shamir O. Proving the lottery ticket hypothesis: Pruning is all you need. In: Proceedings of the 37th International Conference on Machine Learning (ICML). Virtual Event: PMLR, 2020. 6682−6691

104 Xu R X, Luo F L, Zhang Z Y, Tan C Q, Chang B B, Huang S F, et al. Raise a child in large language model: Towards effective and generalizable fine-tuning. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021. 9514−9528

105 Chekalina V A, Rudenko A, Mezentsev G, Mikhalev A, Panchenko A, Oseledets I. SparseGrad: A selective method for efficient fine-tuning of MLP layers. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Miami, USA: Association for Computational Linguistics, 2024. 14929−14939

106 Gheini M, Ren X, May J. Cross-attention is all you need: Adapting pretrained Transformers for machine translation. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021. 1754−1765

107 He H Y, Cai J F, Zhang J, Tao D C, Zhuang B H. Sensitivity-aware visual parameter-efficient fine-tuning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France: IEEE, 2023. 11791−11801

108 Kowsher M, Esmaeilbeig T, Yu C N, Chen C, Soltanalian M, Yousefi N. RoCoFT: Efficient finetuning of large language models with row-column updates. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vienna, Austria: Association for Computational Linguistics, 2025. 26659−26678

109 Aghajanyan A, Gupta S, Zettlemoyer L. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Virtual Event: Association for Computational Linguistics, 2021. 7319−7328

110 Malladi S, Wettig A, Yu D L, Chen D Q, Arora S. A kernel-based view of language model fine-tuning. In: Proceedings of the 40th International Conference on Machine Learning (ICML). Honolulu, USA: PMLR, 2023. 23610−23641

111 Zeng Y C, Lee K. The expressive power of low-rank adaptation. In: Proceedings of the 12th International Conference on Learning Representations (ICLR). Vienna, Austria: OpenReview.net, 2024. 1−46

112 Jang U, Lee J D, Ryu E K. LoRA training in the NTK regime has no spurious local minima. In: Proceedings of the 41st International Conference on Machine Learning (ICML). Vienna, Austria: PMLR, 2024. 21306−21328

113 Zhu J C, Greenewald K H, Nadjahi K, Borde H S D O, Gabrielsson R B, Choshen L, et al. Asymmetry in low-rank adapters of foundation models. In: Proceedings of the 41st International Conference on Machine Learning (ICML). Vienna, Austria: PMLR, 2024. 62369−62385

114 Edalati A, Tahaei M, Kobyzev I, Nia V P, Clark J J, Rezagholizadeh M. KronA: Parameter-efficient tuning with Kronecker adapter. *Enhancing LLM Performance: Efficacy, Fine-Tuning, and Inference Techniques*. Cham: Springer, 2025. 49−65

115 Yang Y F, Zhou J J, Wong N, Zhang Z. LoRETTA: Low-rank economic tensor-train adaptation for ultra-low-parameter fine-tuning of large language models. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). Mexico City, Mexico: Association for Computational Linguistics, 2024. 3161−3176

116 Oseledets I V. Tensor-train decomposition. *SIAM Journal on Scientific Computing*, 2011, **33**(5): 2295−2317

117 Bershatsky D, Cherniuk D, Daulbaev T, Mikhalev A, Oseledets I. LoTR: Low tensor rank weight adaptation. arXiv preprint arXiv: 2402.01376, 2024.

118 Chen X Y, Liu J, Wang Y, Wang P P, Brand M, Wang G H, et al. SuperLoRA: Parameter-efficient unified adaptation of multi-layer attention modules. arXiv preprint arXiv: 2403.11887, 2024.

119 Anjum A, Eren M E, Boureima I, Alexandrov B, Bhattarai M. Tensor train low-rank approximation (TT-LoRA): Democratizing AI with accelerated LLMs. In: Proceedings of the International Conference on Machine Learning and Applications (ICMLA). Miami, USA: IEEE, 2024. 583−590

120 Hounie I, Kanatsoulis C, Tandon A, Ribeiro A. LoRTA: Low rank tensor adaptation of large language models. In: Proceedings of the 13th International Conference on Learning Representations (ICLR). Singapore: OpenReview.net, 2025. 1−20

121 Chang Y P, Chang Y, Wu Y. Bias-aware low-rank adaptation: Mitigating catastrophic inheritance of large language models. arXiv preprint arXiv: 2408.04556, 2024.

122 Biderman D, Portes J P, Ortiz J J G, Paul M, Greengard P, Jennings C, et al. LoRA learns less and forgets less. arXiv preprint arXiv: 2405.09673, 2024.

123 Valipour M, Rezagholizadeh M, Kobyzev I, Ghodsi A. DyLoRA: Parameter-efficient tuning of pre-trained models using dynamic search-free low-rank adaptation. In: Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics. Dubrovnik, Croatia: Association for Computational Linguistics, 2023. 3274−3287

124 Zhang Q R, Chen M S, Bukharin A, Karampatziakis N, He P C, Cheng Y, et al. AdaLoRA: Adaptive budget allocation for parameter-efficient fine-tuning. arXiv preprint arXiv: 2303.10512, 2023.

125 Zhang F Y, Li L Z, Chen J H, Jiang Z Q, Wang B W, Qian Y M. IncreLoRA: Incremental parameter allocation method for parameter-efficient fine-tuning. arXiv preprint arXiv: 2308.12043, 2023.

126 Ding N, Lv X T, Wang Q S, Chen Y L, Zhou B W, Liu Z Y, et al. Sparse low-rank adaptation of pre-trained language models. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Singapore: Association for Computational Linguistics, 2023. 4133−4145

127 Mao Y L, Huang K Y, Guan C H, Bao G L, Mo F R, Xu J N. DoRA: Enhancing parameter-efficient fine-tuning with dynamic rank distribution. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Bangkok, Thailand: Association for Computational Linguistics, 2024. 11662−11675

128 Lialin V, Muckatira S, Shivagunde N, Rumshisky A. ReLoRA: High-rank training through low-rank updates. In: Proceedings of the 12th International Conference on Learning Representations (ICLR). Vienna, Austria: OpenReview.net, 2024. 1−17

129 Xia W H, Qin C W, Hazan E. Chain of LoRA: Efficient fine-tuning of language models via residual learning. arXiv preprint arXiv: 2401.04151, 2024.

130 Ren P J, Shi C S, Wu S G, Zhang M Q, Ren Z C, de Rijke M, et al. MELoRA: Mini-ensemble low-rank adapters for parameter-efficient fine-tuning. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Bangkok, Thailand: Association for Computational Linguistics, 2024. 3052−3064

131 Hayou S, Ghosh N, Yu B. The impact of initialization on LoRA finetuning dynamics. In: Proceedings of the 38th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc., 2024. Article No. 3715

132 Meng F X, Wang Z H, Zhang M H. PiSSA: Principal singular values and singular vectors adaptation of large language models. In: Proceedings of the 38th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc., 2024. Article No. 3846

133 Wang H Q, Li Y X, Wang S, Chen G H, Chen Y. MiLoRA: Harnessing minor singular components for parameter-efficient LLM finetuning. In: Proceedings of the Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). Albuquerque, USA: Association for Computational Linguistics, 2025. 4823−4836

134 Hayou S, Ghosh N, Yu B. LoRA+: Efficient low rank adaptation of large models. In: Proceedings of the 41st International Conference on Machine Learning (ICML). Vienna, Austria: PMLR, 2024. 17783−17806

135 Shi S H, Huang S H, Song M H, Li Z J, Zhang Z H, Huang H Z, et al. ResLoRA: Identity residual mapping in low-rank adaption. In: Proceedings of the Findings of the Association for Computational Linguistics. Bangkok, Thailand: Association for Computational Linguistics, 2024. 8870−8884

136 Wen Z H, Zhang J, Fang Y. SIBO: A simple booster for parameter-efficient fine-tuning. In: Proceedings of the Findings of the Association for Computational Linguistics. Bangkok, Thailand: Association for Computational Linguistics, 2024. 1241−1257

137 Liu S Y, Wang C Y, Yin H X, Molchanov P, Wang Y C F, Cheng K T, et al. DoRA: Weight-decomposed low-rank adaptation. In: Proceedings of the 41st International Conference on Machine Learning (ICML). Vienna, Austria: PMLR, 2024. 32100−32121

138 Wang H Y, Liu T C, Li R R, Cheng M X, Zhao T, Gao J. RoseLoRA: Row and column-wise sparse low-rank adaptation of pre-trained language model for knowledge editing and fine-tuning. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Miami, USA: Association for Computational Linguistics, 2024. 996−1008

139 Wang Z B, Liang J, He R, Wang Z L, Tan T N. LoRA-pro: Are low-rank adapters properly optimized? In: Proceedings of the 13th International Conference on Learning Representations (ICLR). Singapore: OpenReview.net, 2025. 1−22

140 Deng J X, Zhu Q C, Pang J B, Yang L L, Fu Z Q, Zhang B C. EFlat-LoRA: Efficiently seeking flat minima for better generalization in fine-tuning large language models and beyond. arXiv preprint arXiv: 2508.00522, 2025.

141 Yang A X, Robeyns M, Wang X, Aitchison L. Bayesian low-rank adaptation for large language models. In: Proceedings of the 12th International Conference on Learning Representations (ICLR). Vienna, Austria: OpenReview.net, 2024. 1−31

142 Qi Z T, Tan X Y, Shi S J, Qu C, Xu Y H, Qi Y. PILLOW: Enhancing efficient instruction fine-tuning via prompt matching. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing: Industry Track. Singapore: Association for Computational Linguistics, 2023. 471−482

143 Zhang L H, Wu J L, Zhou D Y, Xu G Q. STAR: Constraint LoRA with dynamic active learning for data-efficient fine-tuning of large language models. In: Proceedings of the Findings of the Association for Computational Linguistics. Bangkok, Thailand: Association for Computational Linguistics, 2024. 3519−3532

144 Lin Y, Ma X Y, Chu X, Jin Y J, Yang Z B, Wang Y S. LoRA dropout as a sparsity regularizer for overfitting reduction. *Knowledge-Based Systems*, 2025, **329**: Article No. 114241

145 Zheng J J, Lu W L, Dong Y M, Ji C J, Cao Y K, Lin Z C. AdaMSS: Adaptive multi-subspace approach for parameter-efficient fine-tuning. In: Proceedings of the 39th Conference on Neural Information Processing Systems. San Diego, USA: Curran Associates Inc., 2025.

146 Wang X, Aitchison L, Rudolph M. LoRA ensembles for large language model fine-tuning. arXiv preprint arXiv: 2310.00035, 2023.

147 Zhao Z Y, Gan L L, Wang G Y, Zhou W C S, Yang H X, Kuang K, et al. LoraRetriever: Input-aware LoRA retrieval and composition for mixed tasks in the wild. In: Proceedings of the Findings of the Association for Computational Linguistics. Bangkok, Thailand: Association for Computational Linguistics, 2024. 4447−4462

148 Sun Y H, Li M K, Cao Y X, Wang K, Wang W X, Zeng X Y, et al. To be or not to be? An exploration of continuously controllable prompt engineering. arXiv preprint arXiv: 2311.09773, 2023.

149 Zhang J H, Chen S Q, Liu J T, He J X. Composing parameter-efficient modules with arithmetic operations. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc., 2023. Article No. 552

150 Huang C S, Liu Q, Lin B Y, Pang T Y, Du C, Lin M. LoraHub: Efficient cross-task generalization via dynamic LoRA composition. arXiv preprint arXiv: 2307.13269, 2023.

151 Liu J L, Moreau A, Preuss M, Rapin J, Roziere B, Teytaud F, et al. Versatile black-box optimization. In: Proceedings of the Genetic and Evolutionary Computation Conference. Cancún, Mexico: ACM, 2020. 620−628

152 Tang A K, Shen L, Luo Y, Zhan Y B, Hu H, Du B, et al. Parameter-efficient multi-task model fusion with partial linearization. In: Proceedings of the 12th International Conference on Learning Representations (ICLR). Vienna, Austria: OpenReview.net, 2024. 1−31

153 Shen Y, Xu Z Y, Wang Q F, Cheng Y, Yin W P, Huang L F. Multimodal instruction tuning with conditional mixture of LoRA. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Bangkok, Thailand: Association for Computational Linguistics, 2024. 637−648

154 Buehler E L, Buehler M J. X-LoRA: Mixture of low-rank ad-

apter experts, a flexible framework for large language models with applications in protein mechanics and molecular design. *APL Machine Learning*, 2024, **2**(2): Article No. 026119

155 Feng W F, Hao C Z, Zhang Y W, Han Y, Wang H. Mixture-of-LoRAs: An efficient multitask tuning method for large language models. In: Proceedings of the Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING). Torino, Italia: Association for Computational Linguistics, 2024. 11371−11380

156 Wang Y M, Lin Y, Zeng X D, Zhang G N. MultiLoRA: Democratizing LoRA for better multi-task learning. arXiv preprint arXiv: 2311.11501, 2023.

157 Yang Y Q, Jiang P T, Hou Q B, Zhang H, Chen J W, Li B. Multi-task dense prediction via mixture of low-rank experts. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2024. 27927−27937

158 Agiza A, Neseem M, Reda S. MTLoRA: A low-rank adaptation approach for efficient multi-task learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2024. 16196−16205

159 Gao C Y, Chen K Z, Rao J M, Sun B C, Liu R B, Peng D Y, et al. Higher layers need more LoRA experts. arXiv preprint arXiv: 2402.08562, 2024.

160 Chen S X, Jie Z Q, Ma L. LLaVA-MoLE: Sparse mixture of LoRA experts for mitigating data conflicts in instruction finetuning MLLMs. arXiv preprint arXiv: 2401.16160, 2024.

161 Zhu Y, Wichers N, Lin C C, Wang X Y, Chen T L, Shu L, et al. SiRA: Sparse mixture of low rank adaptation. arXiv preprint arXiv: 2311.09179, 2023.

162 Qing P J, Gao C Y, Zhou Y F, Diao X J, Yang Y Q, Vosoughi S. AlphaLoRA: Assigning LoRA experts based on layer training quality. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Miami, USA: Association for Computational Linguistics, 2024. 20511−20523

163 Wang J J, Yang G J, Chen W T, Yi H H, Wu X H, Lao Q C. MLAE: Masked LoRA experts for parameter-efficient fine-tuning. arXiv preprint arXiv: 2405.18897, 2024.

164 Fan C H, Lu Z Y, Liu S C, Gu C F, Qu X Y, Wei W, et al. Make LoRA great again: Boosting LoRA with adaptive singular values and mixture-of-experts optimization alignment. In: Proceedings of the 42nd International Conference on Machine Learning. Vancouver, Canada: PMLR, 2025. 15804−15832

165 Zhao Z Y, Zhou Y X, Yu X, Zhang Z, Zhu D D, Shen T, et al. Each rank could be an expert: Single-ranked mixture of experts LoRA for multi-task learning. In: Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining. Jeju Island, South Korea: ACM, 2026. 1998−2009

166 Liao M Q, Chen W, Shen J F, Guo S N, Wan H Y. HMoRA: Making LLMs more effective with hierarchical mixture of LoRA experts. In: Proceedings of the 13th International Conference on Learning Representations (ICLR). Singapore: OpenReview.net, 2025. 1−22

167 Mu B S, Wei K, Guo P C, Xie L. Mixture of LoRA experts with multi-modal and multi-granularity LLM generative error correction for accented speech recognition. *IEEE Transactions on Audio, Speech and Language Processing*, 2025, **33**: 2973−2985

168 Prabhakar A, Li Y Z, Narasimhan K, Kakade S, Malach E, Jelassi S. LoRA soups: Merging LoRAs for practical skill composition tasks. In: Proceedings of the 31st International Conference on Computational Linguistics: Industry Track. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, 2025. 644−655

169 Liang J, Huang W K, Wan G C, Yang Q, Ye M. LoRASculpt: Sculpting LoRA for harmonizing general and specialized knowledge in multimodal large language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE, 2025. 26170−26180

170 Yang Y M, Muhtar D, Shen Y L, Zhan Y F, Liu J F, Wang Y J, et al. MTL-LoRA: Low-rank adaptation for multi-task learning. In: Proceedings of the 39th AAAI Conference on Artificial Intelligence. Philadelphia, USA: AAAI Press, 2025. 22010−22018

171 Zhang L T, Zhang L, Shi S H, Chu X W, Li B. LoRA-FA: Memory-efficient low-rank adaptation for large language models fine-tuning. In: Proceedings of the 12th International Conference on Learning Representations (ICLR). Vienna, Austria: OpenReview.net, 2024. 1−19

172 Wu Y C, Xiang Y F, Huo S N, Gong Y L, Liang P H. LoRA-SP: Streamlined partial parameter adaptation for resource efficient fine-tuning of large language models. In: Proceedings of the 3rd International Conference on Algorithms, Microchips, and Network Applications. Jinan, China: SPIE, 2024. Article No. 131711Z

173 Liu Z Y, Kundu S, Li A N, Wan J R, Jiang L H, Beerel P A. AFLoRA: Adaptive freezing of low rank adaptation in parameter efficient fine-tuning of large models. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Bangkok, Thailand: Association for Computational Linguistics, 2024. 161−167

174 Woo S, Park B, Kim B, Jo M, Kwon S J, Jeon D, et al. DropBP: Accelerating fine-tuning of large language models by dropping backward propagation. In: Proceedings of the 38th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc., 2024. Article No. 637

175 Balazy K, Banaei M, Aberer K, Tabor J. LoRA-XS: Low-rank adaptation with extremely small number of parameters. In: Proceedings of the 28th European Conference on Artificial Intelligence. Bologna, Italy: IOS Press, 2025. 3194−3201

176 Zhang M Y, Chen H, Shen C H, Yang Z, Ou L L, Yu X Y, et al. LoRAPrune: Structured pruning meets low-rank parameter-efficient fine-tuning. In: Proceedings of the Findings of the Association for Computational Linguistics. Bangkok, Thailand: Association for Computational Linguistics, 2024. 3013−3026

177 Zhou H Y, Lu X Y, Xu W, Zhu C H, Zhao T J, Yang M Y. LoRA-drop: Efficient LoRA parameter pruning based on output evaluation. In: Proceedings of the 31st International Conference on Computational Linguistics. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, 2025. 5530−5543

178 Ao S, Dong Y, Hu J W, Ramchurn S D. Safe pruning LoRA: Robust distance-guided pruning for safety alignment in adaptation of LLMs. *Transactions of the Association for Computational Linguistics*, 2025, **13**: 1474−1487

179 Miyano R, Arase Y. Adaptive LoRA merge with parameter pruning for low-resource generation. In: Proceedings of the Findings of the Association for Computational Linguistics. Vienna, Austria: Association for Computational Linguistics, 2025. 19353−19366

180 Liu Y J, Yang H R, Chen Y X, Zhang R Y, Wang M, Du Y, et al. PAT: Pruning-aware tuning for large language models. In: Proceedings of the 39th AAAI Conference on Artificial Intelligence. Philadelphia, USA: AAAI Press, 2025. 24686−24695

181 Zhang J, Wang J, Li H, Shou L D, Chen K, You Y, et al.

Train small, infer large: Memory-efficient LoRA training for large language models. In: Proceedings of the 13th International Conference on Learning Representations (ICLR). Singapore: OpenReview.net, 2025. 1−30

182 Yu X, Xie C, Zhao Z Y, Fan T T, Xue L Z, Zhang Z. PrunedLoRA: Robust gradient-based structured pruning for low-rank adaptation in fine-tuning. In: Proceedings of the 14th International Conference on Learning Representations (ICLR). Rio de Janeiro, Brazil: OpenReview.net, 2026. 1−29

183 Kopiczko D J, Blankevoort T, Asano Y M. VeRA: Vector-based random matrix adaptation. In: Proceedings of the 12th International Conference on Learning Representations (ICLR). Vienna, Austria: OpenReview.net, 2024. 1−21

184 Li Y, Han S B, Ji S H. VB-LoRA: Extreme parameter efficient fine-tuning with vector banks. In: Proceedings of the 38th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc., 2024. Article No. 533

185 Gao Z Q, Wang Q C, Chen A C, Liu Z J, Wu B Z, Chen L, et al. Parameter-efficient fine-tuning with discrete Fourier transform. In: Proceedings of the 41st International Conference on Machine Learning (ICML). Vienna, Austria: PMLR, 2024. 14884−14901

186 Zhou Y H, Li R F, Zhou C H, Yang F, Pan A M. BSLoRA: Enhancing the parameter efficiency of LoRA with intra-layer and inter-layer sharing. In: Proceedings of the 42nd International Conference on Machine Learning (ICML). Vancouver, Canada: PMLR, 2025. 78883−78902

187 Zhao J M, Zhang X Y, Li J R, Niu J C, Hu Y L, Min E X, et al. Tiny budgets, big gains: Parameter placement strategy in parameter super-efficient fine-tuning. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Suzhou, China: Association for Computational Linguistics, 2025. 6315−6333

188 Dettmers T, Pagnoni A, Holtzman A, Zettlemoyer L. QLORA: Efficient finetuning of quantized LLMs. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc., 2023. Article No. 441

189 Xu Y H, Xie L X, Gu X T, Chen X, Chang H, Zhang H H, et al. QA-LoRA: Quantization-aware low-rank adaptation of large language models. In: Proceedings of the 12th International Conference on Learning Representations (ICLR). Vienna, Austria: OpenReview.net, 2024. 1−18

190 Li Y X, Yu Y F, Liang C, He P C, Karampatziakis N, Chen W Z, et al. LoftQ: LoRA-fine-tuning-aware quantization for large language models. In: Proceedings of the 12th International Conference on Learning Representations (ICLR). Vienna, Austria: OpenReview.net, 2024. 1−16

191 Deng Y X, Zhang A Z, Gurses S, Wang N G, Yang Z, Yin P H. CLoQ: Enhancing fine-tuning of quantized LLMs via calibrated LoRA initialization. arXiv preprint arXiv: 2501.18475, 2025.

192 Lee G, Lee J, Hong S, Kim M, Ahn E, Chang D S, et al. RILQ: Rank-insensitive LoRA-based quantization error compensation for boosting 2-bit large language model accuracy. In: Proceedings of the 39th AAAI Conference on Artificial Intelligence. Philadelphia, USA: AAAI Press, 2025. 18091−18100

193 Ke W J, Li Z, Li D, Tian L, Barsoum E. DL-QAT: Weight-decomposed low-rank quantization-aware training for large language models. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing: Industry Track. Miami, USA: Association for Computational Linguistics, 2024. 113−119

194 Jeon H, Kim Y, Kim J J. L4Q: Parameter efficient quantization-aware fine-tuning on large language models. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vienna, Austria: Association for Computational Linguistics, 2025. 2002−2024

195 Ye Z M, Li D C, Tian J Q, Lan T F, Zuo J, Duan L, et al. ASPEN: High-throughput LoRA fine-tuning of large language models with a single GPU. arXiv preprint arXiv: 2312.02515, 2023.

196 Chen L Q, Ye Z H, Wu Y J, Zhuo D Y, Ceze L, Krishnamurthy A. Punica: Multi-tenant LoRA serving. In: Proceedings of the 7th Annual Conference on Machine Learning and Systems. Santa Clara, USA: mlsys.org, 2024. 1−13

197 Sheng Y, Cao S Y, Li D C, Hooper C, Lee N, Yang S, et al. S-LoRA: Serving thousands of concurrent LoRA adapters. arXiv preprint arXiv: 2311.03285, 2023.

198 Yan M H, Wang Z, Jia Z, Venkataraman S, Wang Y D. PLoRA: Efficient LoRA hyperparameter tuning for large models. arXiv preprint arXiv: 2508.02932, 2025.

199 Mao Y N, Mathias L, Hou R, Almahairi A, Ma H, Han J W, et al. UniPELT: A unified framework for parameter-efficient language model tuning. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Dublin, Ireland: Association for Computational Linguistics, 2022. 6253−6264

200 Hu S D, Zhang Z, Ding N, Wang Y D, Wang Y S, Liu Z Y, et al. Sparse structure search for delta tuning. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc., 2022. Article No. 716

201 Chen J A, Zhang A, Shi X J, Li M, Smola A, Yang D Y. Parameter-efficient fine-tuning design spaces. In: Proceedings of the 11th International Conference on Learning Representations (ICLR). Kigali, Rwanda: OpenReview.net, 2023. 1−18

202 Zeng G T, Zhang P Y, Lu W. One network, many masks: Towards more parameter-efficient transfer learning. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Toronto, Canada: Association for Computational Linguistics, 2023. 7564−7580

203 Hu Z Q, Wang L, Lan Y H, Xu W Y, Lim E P, Bing L D, et al. LLM-adapters: An adapter family for parameter-efficient fine-tuning of large language models. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Singapore: Association for Computational Linguistics, 2023. 5254−5276

204 Zhang Y H, Zhou K Y, Liu Z W. Neural prompt search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025, **47**(7): 5268−5280

205 Zhou H, Wan X C, Vulić I, Korhonen A. AutoPEFT: Automatic configuration search for parameter-efficient fine-tuning. *Transactions of the Association for Computational Linguistics*, 2024, **12**: 525−542

206 Chai Y D, Liu Y, Zhou Y H, Xie J H, Zeng D D. A Bayesian hybrid parameter-efficient fine-tuning method for large language models. arXiv preprint arXiv: 2508.02711, 2025.

207 Xu L L, Xie H R, Qin S J, Tao X H, Wang F L. Parameter-efficient fine-tuning methods for pretrained language models: A critical review and assessment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026, **48**(6): 6107−6126

208 Wang L P, Chen S, Jiang L N, Pan S, Cai R Z, Yang S, et al. Parameter-efficient fine-tuning in large language models: A survey of methodologies. *Artificial Intelligence Review*, 2025, **58**(8): Article No. 227

209 Meng L C, Li H D, Chen B C, Lan S Y, Wu Z X, Jiang Y G, et al. AdaViT: Adaptive vision Transformers for efficient image recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE, 2022. 12299−12308

210 Chen S F, Ge C J, Tong Z, Wang J L, Song Y B, Wang J, et al. AdaptFormer: Adapting vision Transformers for scalable visual recognition. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc., 2022. Article No. 1212

211 Chen Z, Duan Y C, Wang W H, He J J, Lu T, Dai J F, et al. Vision Transformer adapter for dense predictions. In: Proceedings of the 11th International Conference on Learning Representations (ICLR). Kigali, Rwanda: OpenReview.net, 2023. 1−20

212 Xu M D, Zhang Z, Wei F Y, Hu H, Bai X. Side adapter network for open-vocabulary semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE, 2023. 2945−2954

213 Fu M H, Zhu K, Wu J X. DTL: Disentangled transfer learning for visual recognition. In: Proceedings of the 38th AAAI Conference on Artificial Intelligence. Vancouver, Canada: AAAI Press, 2024. 12082−12090

214 Chen H, Tao R, Zhang H, Wang Y D, Li X, Ye W, et al. Conv-adapter: Exploring parameter efficient transfer learning for ConvNets. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Seattle, USA: IEEE, 2024. 1551−1561

215 Chen Z, Zeng Y, Chen Z H, Gao H Z, Chen L, Liu J M, et al. VFM-adapter: Adapting visual foundation models for dense prediction with dynamic hybrid operation mapping. In: Proceedings of the 39th AAAI Conference on Artificial Intelligence. Philadelphia, USA: AAAI Press, 2025. 2385−2393

216 Pandey D S, Pyakurel S, Yu Q. Be confident in what you know: Bayesian parameter efficient fine-tuning of vision foundation models. In: Proceedings of the 38th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc., 2024. Article No. 1424

217 Yu Q H, He J, Deng X Q, Shen X H, Chen L C. Convolutions die hard: Open-vocabulary segmentation with single frozen convolutional CLIP. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc., 2023. Article No. 1399

218 Yin D S, Yang Y R, Wang Z C, Yu H F, Wei K W, Sun X. 1% VS 100%: Parameter-efficient low rank adapter for dense predictions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE, 2023. 20116−20126

219 Hu Z X, Wei Y X, Shen L, Yuan C, Tao D C. LoRA recycle: Unlocking tuning-free few-shot adaptability in visual foundation models by recycling pre-tuned LoRAs. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE, 2025. 25026−25037

220 Mai Z D, Zhang P, Tu C H, Chen H Y, Nguyen Q H, Zhang L, et al. Lessons and insights from a unifying study of parameter-efficient fine-tuning (PEFT) in visual recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE, 2025. 14845−14857

221 Feng T T, Narayanan S. PEFT-SER: On the use of parameter efficient transfer learning approaches for speech emotion recognition using pre-trained speech models. In: Proceedings of the 11th International Conference on Affective Computing and Intelligent Interaction (ACII). Cambridge, USA: IEEE, 2023. 1−8

222 Lin T H, Wang H S, Weng H Y, Peng K C, Chen Z C, Lee H Y. PEFT for speech: Unveiling optimal placement, merging strategies, and ensemble techniques. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW). Seoul, South Korea: IEEE, 2024. 705−709

223 Chen S Y, Wang C Y, Chen Z Y, Wu Y, Liu S J, Chen Z, et al. WavLM: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 2022, **16**(6): 1505−1518

224 Ali M N, Falavigna D, Brutti A. EFL-PEFT: A communication efficient federated learning framework using PEFT sparsification for ASR. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Hyderabad, India: IEEE, 2025. 1−5

225 Meng L W, Hu S J, Kang J W, Li Z Q, Wang Y J, Wu W X, et al. Large language model can transcribe speech in multi-talker scenarios with versatile instructions. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Hyderabad, India: IEEE, 2025. 1−5

226 Lin T Q, Huang W P, Tang H, Lee H Y. Speech-FT: Merging pre-trained and fine-tuned speech representation models for cross-task generalization. arXiv preprint arXiv: 2502.12672, 2025.

227 Hsu W N, Bolte B, Tsai Y H H, Lakhotia K, Salakhutdinov R, Mohamed A. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2021, **29**: 3451−3460

228 Baevski A, Zhou H, Mohamed A, Auli M. wav2vec 2.0: A framework for self-supervised learning of speech representations. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc., 2020. Article No. 1044

229 Ling S S, Liu Y Z. DeCoAR 2.0: Deep contextualized acoustic representations with vector quantization. arXiv preprint arXiv: 2012.06659, 2020.

230 Zhong Z S, Wang C Y, Liu Y Q, Yang S Q, Tang L X, Zhang Y C, et al. LYRA: An efficient and speech-centric framework for omni-cognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Honolulu, USA: IEEE, 2025. 3694−3704

231 Liang F, Wu B C, Dai X L, Li K P, Zhao Y N, Zhang H, et al. Open-vocabulary semantic segmentation with mask-adapted CLIP. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE, 2023. 7061−7070

232 Gao P, Geng S J, Zhang R R, Ma T L, Fang R Y, Zhang Y F, et al. CLIP-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, 2024, **132**(2): 581−595

233 Yan L Q, Han C, Xu Z L, Liu D F, Wang Q F. Prompt learns prompt: Exploring knowledge-aware generative prompt collaboration for video captioning. In: Proceedings of the 32nd International Joint Conference on Artificial Intelligence. Macao, China: ijcai.org, 2023. 1622−1630

234 Yang L X, Zhang R Y, Wang Y C, Xie X H. MMA: Multi-modal adapter for vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2024. 23826−23837

235 Huang Z H, Guo Y C, Wang H R, Yi R, Ma L Z, Cao Y P, et al. Mv-adapter: Multi-view consistent image generation made easy. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Honolulu, USA: IEEE, 2025. 16377−

16387

236 Li F D. MH-PEFT: Mitigating hallucinations in large vision-language models through the PEFT method. In: Proceedings of the 2nd International Conference on Generative Artificial Intelligence and Information Security. Hangzhou, China: ACM, 2025. 137−142

237 Xu Z Q, Peng Z L, Yang X K, Shen W. FATE: Feature-adapted parameter tuning for vision-language models. In: Proceedings of the 39th AAAI Conference on Artificial Intelligence. Philadelphia, USA: AAAI Press, 2025. 9014−9022

238 Zhang Y, Deng Y X, Guo M H, Hu S M. Adaptive parameter selection for tuning vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE, 2025. 4280−4290

239 Farina M, Mancini M, Iacca G, Ricci E. Rethinking few-shot adaptation of vision-language models in two stages. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE, 2025. 29989−29998

240 Zanella M, Ayed I B. Low-rank few-shot adaptation of vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Seattle, USA: IEEE, 2024. 1593−1603

241 Ornelas R, Chao A, Gupta S, Chao E, Greer R. Improving event-phase captions in multi-view urban traffic videos via prompt-aware LoRA tuning of vision language models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). Honolulu, USA: IEEE, 2025. 1842−1848

242 Lu H D, Zhao C Y, Xue J, Yao L N, Moore K, Gong D. Adaptive rank, reduced forgetting: Knowledge retention in continual learning vision-language models with dynamic rank-selective LoRA. In: Proceedings of the 14th International Conference on Learning Representations (ICLR). Rio de Janeiro, Brazil: OpenReview.net, 2026. 1−26

243 Chitty-Venkata K T, Emani M, Vishwanath V. LangVision-LoRA-NAS: Neural architecture search for variable LoRA rank in vision language models. In: Proceedings of the IEEE International Conference on Image Processing (ICIP). Anchorage, USA: IEEE, 2025. 1330−1335

244 Peng Y Q, Wang P F, Liu J Z, Chen S F. GLAD: Generalizable tuning for vision-language models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). Honolulu, USA: IEEE, 2025. 4369−4379

245 Nguyen T, Wu X B, Dong X S, Le K M, Hu Z Y, Nguyen C D, et al. READ-PVLA: Recurrent adapter with partial video-language alignment for parameter-efficient transfer learning in low-resource video-language modeling. In: Proceedings of the 38th AAAI Conference on Artificial Intelligence. Vancouver, Canada: AAAI Press, 2024. 18824−18832

246 Wang Y H, Ding P X, Li L X, Cui C, Ge Z R, Tong X Y, et al. VLA-adapter: An effective paradigm for tiny-scale vision-language-action model. In: Proceedings of the 40th AAAI Conference on Artificial Intelligence. Singapore: AAAI Press, 2026. 18638−18646

247 Zheng J L, Li J X, Wang Z H, Liu D X, Kang X R, Feng Y C, et al. X-VLA: Soft-prompted Transformer as scalable cross-embodiment vision-language-action model. In: Proceedings of the 14th International Conference on Learning Representations (ICLR). Rio de Janeiro, Brazil: OpenReview.net, 2026. 1−27

248 Kim M J, Finn C, Liang P. Fine-tuning vision-language-action models: Optimizing speed and success. arXiv preprint arXiv: 2502.19645, 2025.

249 Kim Y, Kim J, Mok W L, Park J H, Lee S K. Client-customized adaptation for parameter-efficient federated learning. In: Proceedings of the Findings of the Association for Computational Linguistics. Toronto, Canada: Association for Computational Linguistics, 2023. 1159−1172

250 Cai D Q, Wu Y Z, Wang S G, Xu M W. FedAdapter: Efficient federated learning for mobile NLP. In: Proceedings of the ACM Turing Award Celebration Conference. Wuhan, China: ACM, 2023. 27−28

251 Chen H K, Zhang Y, Krompass D, Gu J D, Tresp V. FedDAT: An approach for foundation model finetuning in multi-modal heterogeneous federated learning. In: Proceedings of the 38th AAAI Conference on Artificial Intelligence. Vancouver, Canada: AAAI Press, 2024. 11285−11293

252 Xiong B C, Yang X S, Song Y G, Wang Y W, Xu C S. Pilot: Building the federated multimodal instruction tuning framework. In: Proceedings of the 39th AAAI Conference on Artificial Intelligence. Philadelphia, USA: AAAI Press, 2025. 21716−21724

253 Saha P, Mishra D, Wagner F, Kamnitsas K, Noble J A. FedPIA-permuting and integrating adapters leveraging Wasserstein barycenters for finetuning foundation models in multimodal federated learning. In: Proceedings of the 39th AAAI Conference on Artificial Intelligence. Philadelphia, USA: AAAI Press, 2025. 20228−20236

254 Wang R, Yu T, Zhang R Y, Kim S, Rossi R, Zhao H D, et al. Personalized federated learning for text classification with gradient-free prompt tuning. In: Proceedings of the Findings of the Association for Computational Linguistics. Mexico City, Mexico: Association for Computational Linguistics, 2024. 4597−4612

255 Yang F E, Wang C Y, Wang Y C F. Efficient model personalization in federated learning via client-specific prompt generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France: IEEE, 2023. 19102−19111

256 Piao H M, Wu Y C, Wu D P, Wei Y. Federated continual learning via prompt-based dual knowledge transfer. In: Proceedings of the 41st International Conference on Machine Learning (ICML). Vienna, Austria: PMLR, 2024. 40725−40739

257 Bai S K, Zhang J, Guo S, Li S C, Guo J C, Hou J, et al. DiPrompT: Disentangled prompt tuning for multiple latent domain generalization in federated learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2024. 27274−27283

258 Tan Y, Long G D, Ma J, Liu L, Zhou T Y, Jiang J. Federated learning from pre-trained models: A contrastive learning approach. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc., 2022. Article No. 1405

259 Zhang Z, Yang Y H, Dai Y, Wang Q F, Yu Y, Qu L Z, et al. FedPETuning: When federated learning meets the parameter-efficient tuning methods of pre-trained language models. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics. Toronto, Canada: Association for Computational Linguistics, 2023. 9963−9977

260 Su S C, Li B, Xue X Y. FedRA: A random allocation strategy for federated tuning to unleash the power of heterogeneous clients. In: Proceedings of the 18th European Conference on Computer Vision (ECCV). Milan, Italy: Springer, 2024. 342−358

261 Yang Y Y, Long G D, Shen T, Jiang J, Blumenstein M. Dual-personalizing adapter for federated foundation models. In: Proceedings of the 38th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc., 2024. Article No. 1245

262 Ghiasvand S, Yang Y F, Xue Z Y, Alizadeh M, Zhang Z, Pedarsani R. Communication-efficient and tensorized federated fine-tuning of large language models. In: Proceedings of the Findings of the Association for Computational Linguistics. Vienna, Austria: Association for Computational Linguistics, 2025. 24192–24207

263 Bian J M, Wang L, Zhang L T, Xu J. LoRA-FAIR: Federated LoRA fine-tuning with aggregation and initialization refinement. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Honolulu, USA: IEEE, 2025. 3737–3746

唐岸达 北京大学智能学院博士后. 2024 年获得中国科学院大学数学科学学院理学博士学位. 主要研究方向为机器学习、深度学习和优化.
E-mail: tanganda@pku.edu.cn
(**TANG An-Da** Postdoctor at the School of Intelligence Science and Technology, Peking University. He received his Ph.D. degree in Science from the School of Mathematical Sciences, University of Chinese Academy of Sciences in 2024. His research interests include machine learning, deep learning, and optimization.)

林宙辰 北京大学智能学院教授. 2000 年获得北京大学博士学位. 主要研究方向为机器学习、数值优化. 本文通信作者. E-mail: zlin@pku.edu.cn
(**LIN Zhou-Chen** Professor at the School of Intelligence Science and Technology, Peking University. He received his Ph.D. degree from Peking University in 2000. His research interests include machine learning and numerical optimization. Corresponding author of this paper.)