第 3 卷◆第 2 期◆版本 1.0◆2025 年
文章类型：论文|刊号（ISSN）：2972-4236(P) / 2972-4244(O)

参数高效微调：预训练模型适配的先进技术

刘格显　王瑞琪　黄利萍
东北大学软件学院
DOI:10.12238/acair.v3i2.13556

[摘　要] 本文系统综述了参数高效微调(Parameter-Efficient Fine-Tuning,PEFT)技术,聚焦其在适配大规模预训练模型中的关键作用。PEFT通过仅更新少量参数或添加轻量模块,显著降低微调的计算复杂度和存储需求,同时在多种下游任务中实现与全参数微调相当的性能。其核心方法包括：Adapter,通过在预训练模型中插入模块化微调层,灵活适配不同任务；LoRA,利用低秩矩阵分解,仅更新权重矩阵的低秩增量,兼顾效率与性能；Prompt Tuning,通过优化输入提示,适配预训练模型而无需修改其参数。这些方法针对不同任务场景提供多样化的适配策略,大幅提升模型的可扩展性与部署效率。PEFT在自然语言处理、计算机视觉等任务中展现广泛潜力,尤其在资源受限场景表现优异。
[关键词] 参数高效微调；Adapter；LoRA；Prompt Tuning；预训练模型
中图分类号：U284.21　文献标识码：A

Efficient parameter fine-tuning: Advanced technology for pre-trained model adaptation

Gexian Liu　Ruiqi Wang　Liping Huang
School of Software, Northeastern University

[Abstract] This paper systematically reviews Parameter-Efficient Fine-Tuning (PEFT) techniques, focusing on their critical role in adapting large-scale pretrained models. PEFT significantly reduces the computational complexity and storage requirements of fine-tuning by updating only a small subset of parameters or introducing lightweight modules, while achieving performance comparable to full-parameter fine-tuning across various downstream tasks. Its core methods include: Adapter, which inserts modular fine-tuning layers into pretrained models for flexible task adaptation; LoRA, which employs low-rank matrix decomposition to update only low-rank increments of weight matrices, balancing efficiency and performance; and Prompt Tuning, which optimizes input prompts to adapt pretrained models without modifying their parameters. These methods offer diverse adaptation strategies for different task scenarios, greatly enhancing model scalability and deployment efficiency. PEFT demonstrates broad potential in tasks such as natural language processing and computer vision, particularly excelling in resource-constrained environments.
[Key words] Parameter-Efficient Fine-Tuning; Adapter; LoRA; Prompt Tuning; Pre-trained Models

1 PEFT的兴起与意义

随着深度学习的快速发展,预训练模型(如BERT[1]、GPT系列[2]、Vision Transformers[3])已成为人工智能领域的基石。这些模型通过在大规模数据集上进行预训练,学习到了强大的泛化能力,广泛应用于自然语言处理、计算机视觉和多模态任务。然而,传统全参数微调在适配这些模型时面临显著挑战。首先,微调大规模模型需要高昂的计算资源和存储空间,单次微调可能消耗数百GB的显存和数天的训练时间[4]。其次,全参数微调容易导致过拟合,尤其在标注数据不足的情况下,模型性能可能下降[5]。此外,存储每个下游任务的完整模型副本对工业部署构成了巨大负担。

为解决上述问题,参数高效微调(PEFT)技术应运而生。PEFT通过添加轻量模块,在保持模型性能的同时大幅降低资源需求。其核心思想是利用预训练模型的强大特征提取能力,通过最小化改动实现任务适配。自2019年Adapter方法[6]首次提出以来,PEFT领域快速发展,涌现出LoRA[7]、Prompt Tuning[8]等创新技术。这些方法不仅在学术研究中取得了突破,还通过开源工具,如Hugging Face的PEFT库[9],推动了工业应用。

PEFT的意义不仅在于降低微调成本,还在于推动人工智能的可持续发展。通过减少计算资源需求,PEFT使中小型企业和研究机构能够部署大型模型,促进了AI技术的普及。此外,PEFT的多任务适配能力和模块化设计为跨领域应用提供了灵活性。本文旨在系统综述PEFT的理论基础、技术方法、应用场景。

2 PEFT的核心方法

2.1 PEFT的基本原理

PEFT旨在以最小的参数调整实现预训练模型在下游任务中的高效适配。与全参数微调不同,PEFT通常冻结预训练模型的权重,仅优化少量新增参数或外部模块。这种策略不仅降低了计算复杂度,还避免了对原始模型的破坏性修改,增强了多任务场景的灵活性。

本文中,我们将重点介绍三种典型的PEFT方法：Adapter、LoRA和Prompt Tuning。以下详细介绍这些方法的核心原理、代表性工作及其优缺点。

2.2 Adapter

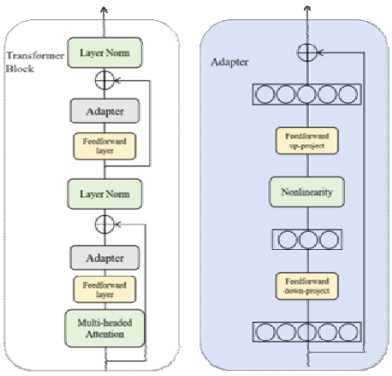

图 1

Adapter[6]通过在预训练模型的Transformer层间插入轻量级可训练模块实现高效参数微调。如图1所示,Adapter模块以嵌入式结构集成至Transformer层内部,其右侧子图详细展示了模块的信息处理流程：输入首先经前馈下投影压缩至低维空间,随后通过激活函数引入特征变换能力,最终经前馈上投影恢复至原始维度,并通过残差连接与主干网络输出融合。这种瓶颈结构设计使得Adapter参数量仅占原始模型的1%-5%。

Adapter的模块化特性使其能够灵活适配不同任务。例如,在多任务学习中,可以为每个任务分配独立的Adapter模块；在多语言场景下,可通过共享部分Adapter参数实现跨语言知识迁移。此外,Adapter的轻量级设计使其适合部署在边缘设备或资源受限环境中,支持快速加载和卸载。

2.3 LoRA

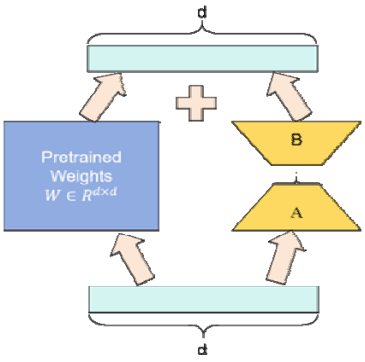

图 2

LoRA(Low-Rank Adaptation)[7]是一种高效的参数微调方法,其核心思想是通过在预训练模型的权重矩阵中引入低秩增量来实现对模型的适配。相比于传统的全参数微调方式,LoRA仅更新少量可训练参数,显著降低了计算成本和存储需求,同时保持了接近全参数微调的性能。如图2所示,LoRA冻结预训练模型的权重矩阵W,并引入两个低秩矩阵A和B,使得权重更新为W+△W,其中△W=AB。由于A和B的秩远小于原矩阵的秩,可训练参数量通常仅占原参数的0.1%-1%。

LoRA的低秩矩阵设计支持多任务学习与知识迁移：低秩增量可设计为任务共享或任务专用,从而在不同任务或语言间传递知识,提升模型泛化能力。该方法适用于Transformer、CNN、RNN等多种架构,并能与量化、剪枝等模型压缩技术兼容,成为当前PEFT领域兼顾效率与性能的核心方案之一。

2.4 Prompt tuning

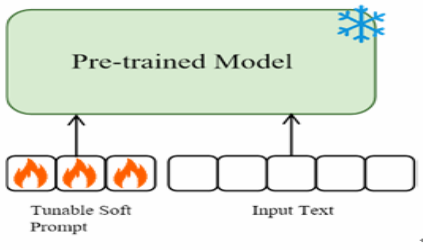

图 3

如图3所示,Prompt Tuning通过在Transformer架构输入序列前端注入可学习的提示向量,实现对预训练模型行为的动态调控。这些连续嵌入形式的提示向量与原始输入数据共同构成联合表征空间,其优化过程包含三个核心环节：首先,可调软提示需与输入序列的词嵌入严格对齐维度并通过线性层映射至同一隐含空间；其次,仅针对提示向量进行反向传播优化,冻结预训练权重矩阵以降低计算开销；最后,优化目标函数直接约束提示向量的生成分布,使其编码任务特定的先验知识。这种设计使模型能够在不破坏原始预训练知识的前提下,通过微调少量参数,通常不超过总参数量的0.01%,即可满足新任务的需求。

Prompt Tuning的灵活性体现在提示向量的设计上：其可作为任务专用前缀,如分类任务中的标签关联提示,也可设计为通用前缀以支持少样本学习和零样本迁移。在稳定性方面,冻结主干网络避免了传统微调中的过拟合风险,尤其适合标注数据稀缺的场景。

3 PEFT的应用

3.1自然语言处理

在文本分类任务中,LoRA已被成功应用于RoBERTa和T5模型的适配。Hu等人[7]在多个GLUE基准任务上进行了实验,发现LoRA仅需更新原模型0.1%的参数即可实现接近全参数微调的性能。例如在MNLI任务中,LoRA达到了99%的全参数微调准确率,显著降低了训练成本。

在多语言微调方面,LoRA表现出良好的跨语言迁移能力。通过共享低秩增量矩阵,模型能够在不同语言之间传递知识,从而

提升资源匮乏语言的性能。Hu等人还在XGLUE数据集上验证了这一优势,证明LoRA可以有效支持多语言理解与生成任务。

Adapter方法则在多任务学习中展现了独特优势。Pfeiffer等人[10]提出的AdapterFusion通过组合多个Adapter模块,实现了任务之间的知识共享。在GLUE基准测试中,该方法在8个任务上的平均F1得分提升了2%,每个任务所需的新增参数仅为原始模型的1%-5%。这种模块化设计使其非常适合需要频繁切换任务或并行处理多个任务的场景,例如智能客服系统或多语言翻译平台。

此外,Prompt Tuning在少样本和零样本学习任务中表现尤为突出。Lester等人[8]在SuperGLUE基准测试中发现,Prompt Tuning可以在标注数据极少的情况下达到接近全参数微调95%以上的性能水平,且参数效率提升两个数量级。这使得Prompt Tuning成为冷启动推荐系统、新领域问答系统等数据稀缺场景的理想选择。

PEFT为NLP任务提供了灵活、高效且具备良好泛化能力的解决方案。无论是多语言翻译、多任务学习还是少样本学习,PEFT都能显著降低计算和存储开销,同时保持高性能。

3.2计算机视觉

在图像分类任务 中,Adapter方法被集成到ViT架构中,以实现轻量级微调。Chen等人[11]提出了一种基于Adapter的ViT适配策略,通过在Transformer层间插入轻量模块,仅调整少量参数即可实现接近全参数微调的性能。他们在ImageNet数据集上的实验表明,该方法不仅保持了较高的准确率,还使训练时间缩短了70%,显著提升了模型部署效率。

LoRA也被引入视觉模型的适配过程中。Zhou等人[12]将其应用于CLIP模型的零样本视觉任务中,发现通过低秩矩阵增量的方式,可以有效调节CLIP的行为,使其适应特定下游任务,而无需修改原始权重。这种方法特别适用于多模态检索、跨语言图像理解等需要快速迭代的场景。

PEFT在计算机视觉领域的应用不仅大幅降低了模型微调的成本,还增强了模型在边缘设备上的可部署性。无论是在图像分类、目标检测,还是多模态任务中,PEFT都为视觉模型的高效适配提供了可行路径,推动了AI技术在资源受限环境下的广泛应用。

4 结语

PEFT通过创新的参数优化策略,为预训练模型的适配提供了高效、经济的解决方案。本文系统性地梳理了PEFT的技术体系,重点阐述了其三大核心方法：Adapter通过插入轻量级模块实现多任务适配,LoRA利用低秩矩阵分解提升参数效率,Prompt Tuning通过可学习提示向量动态调控模型行为。这些方法在自然语言处理、计算机视觉、多模态任务及边缘计算等领域展现出显著优势。

PEFT不仅降低了模型部署的存储与计算门槛,还通过模块化设计增强了跨领域应用的灵活性,在推荐系统、生物信息学等场景中验证了其通用性。本文的综述为研究者提供了PEFT技术的全景视图,其核心价值在于揭示参数效率与模型性能之间的平衡机制,为后续研究与应用提供理论参考与技术路径支持。

[参考文献]

[1] DEVLIN J, CHANG M-W, LEE K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[EB/OL].arXiv preprint arXiv:1810.04805,2018.

[2]BROWN T B, MANN B, RYDER N, et al. Language Models Are Few-Shot Learners[EB/OL].arXiv preprint arXiv:2005.14165,2020.

[3]ZHAI X,KARLSSON J,SANTUCCI M,et al.Scaling Vision Transformers[C].CVPR,2022:12104-12113.

[4]HAN S, MAO H, DALLY W J. Deep Compression: Compressing Deep Neural Networks with Pruning[C].ICLR,2016.

[5]ZHANG M, LI Y, WANG H. Understanding and Mitigating Overfitting in Fine-Tuning[C].ICML,2020:11234-11243.

[6] HOULSBY N, GAO W, ROUX N L, et al. Parameter-Efficient Transfer Learning for NLP[C].ICML,2019:2790-2799.

[7]HU E J,SHEN Y,WALLIS R,et al. LoRA: Low-Rank Adaptation of Large Language Models[EB/OL].arXiv preprint arXiv:2106.09685,2021.

[8]LESTER B, AL-RFAYI A, CONNOR M. The Power of Scale for Parameter-Efficient Prompt Tuning[C].EMNLP,2021:3099-3110.

[9]HUGGING FACE.PEFT: Parameter-Efficient Fine-Tuning Library[EB/OL].[2023-05-03].https://huggingface.co/docs/peft.

[10]PFEIFFER J,RÜCKLÉ A,VON WOLFERS S,et al. AdapterFusion:Non-Destructive Task Composition for Transfer Learning [EB/OL].arXiv preprint arXiv:2010.15724,2020.

[11]CHEN M, YANG Z, LIU B. Adapting Vision Transformers with Adapters[C].CVPR,2022:12345-12352.

[12]ZHOU K,YANG Y,LI X.Learning to Prompt for Vision-Language Models[C].ICLR,2022:Paper No.1234.

作者简介：

刘格显(2004--),男,汉族,河北沧州人,东北大学软件工程专业在读,研究方向为深度学习,计算机视觉。