资源受限的大模型高效迁移学习算法研究综述

李鑫尧[1)] 李晶晶[1)] 朱 磊[2)] 申恒涛[1)]

[1)](电子科技大学计算机科学与工程学院 成都 611731)

[2)](同济大学电子信息与工程学院 上海 200092)

摘 要 近年来,深度学习在自然语言理解、计算机视觉和数据挖掘等重要领域取得了巨大成功,极大地推动了人工智能技术的发展.迁移学习的诞生和应用更是大幅减轻了数据的获取和标注成本,成倍提升了深度模型和算法的泛化能力和适用性.然而,随着模型规模的不断增大,传统的迁移学习方法面临着计算和存储资源的巨大挑战,难以满足可穿戴、军事、医疗等资源受限场景下的应用需求.高效迁移学习算法应运而生,旨在以最小的资源开销实现大模型的快速适配与部署,有望成为未来人工智能技术发展的关键突破口.本文是高效迁移学习领域的首篇中文综述,系统总结了近 5 年来该领域的研究进展.本文首先分析了高效迁移学习算法在自然语言处理、计算机视觉和多模态模型三大场景下的应用现状,提炼出了修改模型结构、调整预训练参数、调整原始输入(输出)、注入自适应参数、引入自适应模块等五类具有代表性的技术路线.在此基础上,本文对各类方法进行了全面梳理与比较,分析了它们的优势与局限性.本文的主要贡献如下:(1)对高效迁移学习领域进行了系统化的综述,为后续研究提供了完整的技术参考;(2)提出了一种基于技术路线的分类框架,帮助读者快速把握该领域的研究脉络;(3)深入分析了现有方法的不足,并展望了未来的发展方向,具有一定的前瞻性和指导意义.高效迁移学习算法是推动现代人工智能技术走进千家万户的关键技术,有望让更多中小企业和个人用户受益于大模型的强大性能.本文对该领域的全面梳理,将为该领域算法的进一步发展和应用提供重要的理论参考与实践指导.

关键词 迁移学习;深度学习;高效方法;多模态模型;大模型;资源受限

中图法分类号 TP391 **DOI 号** 10.11897/SP.J.1016.2024.02491

Efficient Transfer Learning of Large Models with Limited Resources: A Survey

LI Xin-Yao[1)] LI Jing-Jing[1)] ZHU Lei[2)] SHEN Heng-Tao[1)]

[1)](*School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu* 611731)

[2)](*School of Electronic and Information Engineering, Tongji University, Shanghai* 200092)

Abstract In recent years, the fast-evolving deep learning techniques have dominated critical fields such as natural language understanding, computer vision, multimodal processing and data mining, therefore greatly advancing the development of artificial intelligence (AI) technology. Among these advancements, transfer learning (TL) has emerged as a pivotal technique aimed at effectively reusing and sharing knowledge across multiple related models. This approach not only reduces the substantial costs associated with data collection and annotation, but also contributes to enhanced generalizability and capability of deep models. However, the exponential growth in the size, complexity, and depth of deep large models has presented serious challenges to traditional training and transfer algorithms, particularly in terms of computational and storage requirements. Such high computational complexity poses significant obstacles to effective knowledge transfer in resource-constrained scenarios, including but not limited to wearable

收稿日期:2024-01-23;在线发布日期:2024-07-11.本课题得到国家自然科学基金(62176042)、四川省自然科学基金(2023NSFSC0483)、TCL 科技创新基金(SS2024105)资助.**李鑫尧**,博士研究生,主要研究方向为迁移学习、领域自适应、大型基础模型.E-mail:xinyao326@outlook.com.**李晶晶**(通信作者),博士,教授,中国计算机学会(CCF)会员,主要研究领域为计算机视觉、多媒体分析、领域自适应和推荐系统.E-mail:lijin117@yeah.net.**朱 磊**,博士,教授,中国计算机学会(CCF)会员,主要研究领域为大规模多媒体数据检索与分析、数据挖掘和深度学习.**申恒涛**,博士,教授,中国计算机学会(CCF)会员,主要研究领域为多媒体搜索、计算机视觉、人工智能和大数据管理.

technology, military applications, and healthcare systems. To address these challenges, efficient transfer learning algorithms have recently emerged as a promising solution, enabling agile adaptation and deployment of large models with minimal resource overhead. These algorithms are expected to become a key technological driver in the future development of AI. This paper stands out as the first comprehensive survey on the field of efficient transfer learning, aiming to systematically summarize research progress in this thriving research field over the past five years. Concretely, this paper investigates efficient transfer learning across three primary application fields: natural language processing, computer vision, and multimodal models. Among each application field, this paper further identifies and elaborates on five representative technical approaches that have gained prominence in recent research: modifying model structures, adjusting pre-training parameters, adapting original inputs (outputs), injecting adaptive parameters, and introducing adaptive modules. Each of these approaches is subjected to a comprehensive and thorough review, analyzing their respective strengths, limitations, and potential applications. This critical evaluation provides readers with a nuanced understanding of the current state of the art in efficient transfer learning. The primary contributions of this survey are threefold: (1) This survey presents the first systematic review of efficient transfer learning, offering invaluable technical insights and guidance for future research endeavors in this rapidly evolving field. (2) This survey proposes a novel technique-based framework that provides a clear and systematic research guideline, enabling readers to navigate the complex landscape of efficient transfer learning methodologies. (3) This survey conducts an in-depth analysis of the shortcomings and limitations of current methods, thereby identifying critical research gaps and providing insightful directions for future investigations. Efficient transfer learning serves as a crucial bridge between cutting-edge AI technologies and their practical applications in everyday life. It holds the potential to enable easier and cheaper access to the power of large models, benefiting a wide range of enterprises and individuals across various sectors. By providing comprehensive overviews of the current state of the art, solid theoretical foundations, practical guidance, along with critical insights into future research directions, this survey contributes significantly to the development of efficient transfer learning, and is hope to inspire researchers and practitioners to push the boundaries of the research field.

Keywords transfer learning; deep learning; efficient method; multimodal model; large model; limited resources

1 引 言

深度学习在医疗[1]、机械制造[2]、农业[3]等领域有着广泛的应用,且已经取得了巨大的成功.深度学习模型的有效性基于独立同分布假设[4],即模型的训练数据和测试数据必须服从相同的分布,否则模型的表现会大打折扣,例如在南方气候条件下训练得到的机械健康状况预测模型在北方气候条件下精度会急剧下降[5].这种特性让深度模型中蕴含的知识即使在相近的应用领域中也难以泛化,而为每一个具体的任务重新收集数据、训练模型又会带来巨大的开销.为了解决这些问题,研究者提出了迁移学习[6],它的基本思想是仅需要目标任务上极少的数据便可以将相关任务上训练好的模型迁移过来,通过最大限度的模型复用降低数据收集和训练成本.例如在大规模图片数据集 ImageNet[7]上预训练得到的 ResNet[8]系列模型在很长一段时间内主导着计算机视觉领域的学术研究和工业应用.借助最简单的迁移学习技术,无需大量算力、数据资源也可以享受预训练模型带来的强大性能.

随着人们对深度模型的性能要求与日俱增,研究者提出了许多更大更复杂的模型结构.图 1 展示了 2019 年至今诞生的部分大型深度语言模型,其中模型

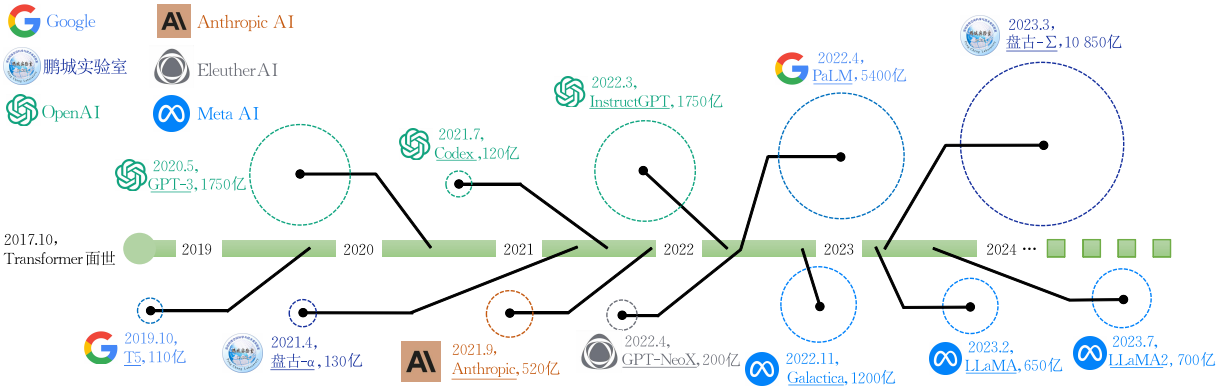

图 1　2019 年至今诞生的部分大型语言模型，图中的文字展示介绍了每个模型的发表时间、发表单位和参数量，且每个环的直径与该模型参数量呈正相关

参数量从数百亿到上千亿，乃至上万亿不等，且有随着时间逐渐变大的趋势. 值得注意的是，其中参数较大的模型，如 Anthropic[9]（520 亿）、GPT-3[10]（1750 亿）、InstructGPT[11]（1750 亿）、PaLM[12]（5400 亿）和盘古-Σ[13]（10 850 亿）等，均是由 AI 领域的顶尖公司、机构训练而成，且其预训练权重由于涉及数据集保密、商业机密等原因无法公开获取.

作为一般学术研究基石的开源预训练模型，参数量则相对少些，如 T5[14]（110 亿）和 LLaMA[15]（650 亿），但从零训练这样的模型仍然需要巨量的算力和数据集资源（训练 LLaMA 需要 2T 个 token），带来的数百万美元的训练成本也是一般机构、个人难以承担的.

模型越大、其训练成本越高，迁移学习的重要性就越发显著. 事实上，经典的迁移学习预训练-微调范式已成为深度模型在工业界中的标准应用方式之一. 以自然语言模型为例，在预训练阶段，模型接触数百亿的单词、简单句等语料，其对数据量的要求之大让大多数研究者或机构难以独立完成，仅有少数头部机构有能力进行大模型预训练. 这一阶段的目的是基于大量普适性数据让模型了解语言的基础构成、语法、情感和一些基本事实，掌握对整个语言系统的理解能力，如 GPT 就是一类预训练大语言模型. 在微调阶段，预训练模型根据具体任务要求，如翻译、对话等，在目标任务的精细数据集上进一步训练. 这一阶段所需的数据量和算力远小于预训练，但对数据质量和专用性要求更高. 想要以最少的开销取得最好的性能，就必然离不开对预训练大模型的迁移和泛化.

然而无论是微调或是其他经典迁移学习算法，其设计之初并未考虑到当今预训练模型的参数规模呈现爆发性的增长. 以微调为例，虽然训练轮次、所需数据量较重新训练模型大大减少，这种方法仍然要求对预训练模型的所有参数计算梯度并更新. 面对百亿级别的参数量时，微调的开销仍远远超出了绝大部分研究机构和研究者能支持的范围，例如微调一个具有 650 亿参数的预训练 LLaMA 模型[15]时，需要超过 780 GB 显存. 除此之外，诸如军事、航空航天等特种行业中电力算力资源天然有限，同样难以支撑大模型的迁移成本，严重阻碍了最新最好的大模型在这些专用领域的应用和发展.

另外，许多应用场景也对模型知识的快速迁移泛化能力提出了高要求. 以智能驾驶领域为例，近年来出现了许多致力于解决如激光雷达测距[16]、智能驾驶车辆轨迹预测[17]等智能驾驶核心问题的相关模型和算法. 然而智能驾驶领域的应用环境复杂多变，车况、路况、驾驶员的驾驶习惯等都是预训练智能驾驶模型时不可预知的，而汽车高速运行时要求模型能以极短的时延泛化至特定的场景中，这是传统的微调等迁移方法做不到的. 随着智能终端的发展，人们对其智能处理能力的要求越发苛刻，但智能终端的硬件能力又难以提供微调现代大模型所需的算力，因此迫切需要一种时间短、计算高效的大模型迁移方法.

考虑到上述问题，近期也有不少学者着眼于设计高效迁移学习方法. 不同于早期方法以单一的迁移效果为目标，高效迁移方法致力于寻找迁移效果和迁移开销的帕累托最优，也就是用最少的迁移性能为代价，节省最大的计算开销. 这种方法的出现反映了人工智能领域对计算效率和资源利用的日益重视. 作为一个新兴研究方向，这些方法种类繁多、采用的技术层次复杂多样，对应的应用场景和效果也有所不同. 且由于此方向发展时间较短，尚未形成一

套成熟的划分框架和衡量标准.因此,学术界迫切需要一篇全面深入的综述论文系统总结高效迁移学习领域的主流方法、核心技术和未来研究方向,为这一领域提供清晰理论指导.

数十年来,学术界对迁移学习理论和应用进行了广泛深入的研究,形成了大量综述文献.如表1所示,文献[18-23]系统梳理了迁移学习从传统机器学习时代进入深度学习时代的发展历程,总结了不同问题设定和应用场景下的各种方法,其中表格右侧提供了四类常见的应用场景,并指明了每篇文献的讨论范围(用黑点表示).

表1 现有的部分迁移学习、高效方法相关综述概况

文献	主要贡献	迁移学习	多模态	视觉	自然语言
A Survey on Transfer Learning[18] (2009)	(1)第一篇较完整全面的迁移学习综述; (2)介绍了迁移学习的问题定义、设定、流派等; (3)主要介绍了基于传统机器学习的浅度方法.	●			
迁移学习问题与方法研究[19](2014)	(1)针对负迁移问题,提出一种图正则化联合矩阵分解模型; (2)针对欠适配问题,提出一种联合适配正则化学习框架; (3)针对欠拟合、欠适配与负迁移问题,提出统一的鲁棒深度表征适配模型.	●	●		●
A survey of transfer learning[20](2016)	(1)进入深度学习时代以来第一篇较完善的迁移学习综述; (2)沿用文献[18]的定义,涉及大量相关文献,覆盖面极广; (3)提供部分方法的软件实现.				●
A Survey on Deep Transfer Learning[21] (2018)	(1)针对深度学习的迁移学习综述,并给出深度迁移学习的定义; (2)将深度迁移学习方法分为基于样本、映射、网络、对抗4类.				●
A Comprehensive Survey on Transfer Learning[22] (2020)	(1)深度迁移学习的较全面综述; (2)基于文献[21]的分类方法进一步细分,提供了基于特征变换、度量等方法的介绍; (3)提供了较详细的实验设定和实验结果介绍.	●	●		●
A Review of Deep Transfer Learning and Recent Advancements[23](2023)	(1)较新的深度迁移学习综述; (2)基于SLR法[24]整理了近5年的文献及其关键方法,覆盖的方法范围较广,包括视觉、语言、医学、应用物理、动作识别等.	●	●		●
Efficient Methods for Natural Language Processing: A Survey[25](2023)	(1)针对自然语言处理的第一篇高效深度学习综述; (2)从数据、模型、(预)训练、推理、部署等方面展开介绍.	●			
Efficient Deep Learning: A Survey on Making Deep Learning Models Smaller, Faster, and Better[26](2023)	(1)针对高效深度学习的完善综述,覆盖自然语言处理和计算机视觉; (2)从压缩、训练、模型结构、(参数、结构等的)自动化探索、底层软硬件等方面展开介绍.	●	●		
本文(2024)	(1)第一篇覆盖自然语言、计算机视觉、多模态的高效迁移学习综述; (2)从模型参数、结构、模块、输入输出等方面展开讨论; (3)结合最前沿的研究方向和热点探讨未来高效迁移学习的发展方向.	●	●	●	●

随着大模型技术的兴起,一些学者也开始关注降低深度学习计算开销的相关方法,例如模型压缩、模型蒸馏等,相关综述如文献[25-26]所示.然而,现有迁移学习综述未能充分考虑资源受限场景下的高效迁移问题,而深度学习高效性综述则侧重于模型设计、训练等环节的优化,鲜有涉及如何高效地迁移和部署已有大规模预训练模型.另一方面,当前研究热点已逐渐转向跨模态、多模态领域,但现有的高效迁移学习综述文献对这一新兴方向还缺乏足够的关注和讨论.

本文作为大模型高效迁移领域的首篇中文综述,集中探讨在计算资源、存储资源和数据资源受限情况下,如何高效地将通用大规模预训练模型迁移到特定视觉、语言和多模态任务上.与现有的迁移学习综述不同,本文聚焦于资源受限条件下针对大模型的迁移方法,研究内容符合现代人工智能发展方向,填补了大模型时代迁移学习综述的空白;与现有的模型结构、训练过程轻量化综述相比,本文聚焦于不同模态中的预训练模型迁移技术,研究内容符合现代大模型应用的一般逻辑,完善了预训练模型的轻量化迁移部署研究领域.本文首次提出一种基于技术路线的分类框架,跨领域对比分析视觉、语言和多模态领域中不同的高效迁移方法,从模型结构、预训练参数、输入输出、自适应参数、自适应模块等多个角度系统梳理现有技术,突破现有综述中基于任务类型的分类壁垒,为读者提供更加宏观和前瞻的研究视角.同时,本文还深入分析了现有方法的不足之处,并前瞻性地探讨高效迁移学习在人工智能技术推广过程中的重要作用,为该领域的未来发展指明方向,凸显了本项工作的创新性和指导意义.

本文顺应大模型发展浪潮、旨在帮助初次接触高效迁移学习领域的学者快速掌握该领域的核心问题和关键技术.本文的主要创新贡献包括:(1)首次系统梳理和总结了语言、视觉和多模态等不同场景下的高效迁移方法.本文不仅分析了这些方法为适应特定应用场景所作出的调整,更重要的是揭示了跨领域通用的核心技术,帮助读者透过表象理解不同方法的本质异同;(2)首次从图 2 所描绘的五种技术路线的视角出发,对现有高效迁移方法进行了系统分类和归纳,并与现有高效深度学习综述[26]对比,凸显了本文在高效迁移学习技术路线上的独到理解和创新;(3)深入剖析了现有研究工作的不足之处,并前瞻性地探讨了该领域的潜在发展方向,为有意从事相关研究的读者提供了指导性建议和创新思路.接下来简要介绍上述五个核心技术路线:

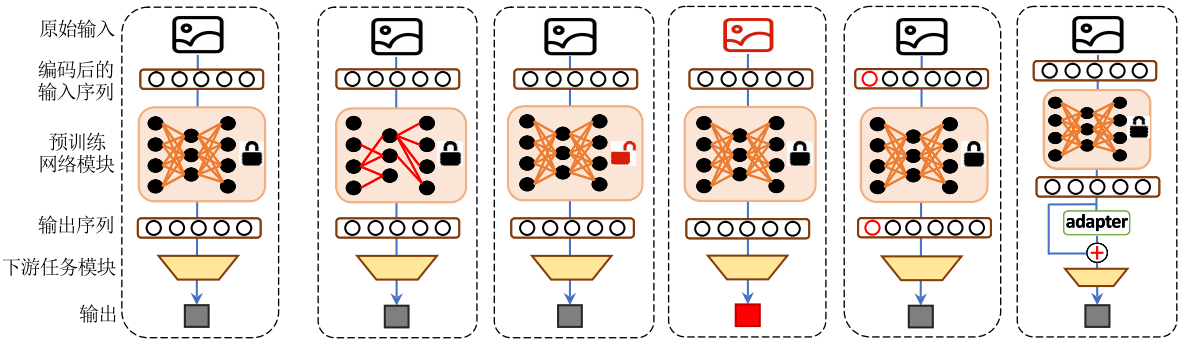

图 2　本文探讨的 5 种高效迁移技术图示

(1)预训练模型结构是否改变.综述[26]中提到的高效方法(如 MobileNetV2[27]等)侧重于模型结构设计的轻量化,以最少的训练开销实现最好的效果.而本文关注现有模型的迁移,通过“做减法”:减少预训练模型中冗余的分支、减少模型子网络的大小等方式压缩模型以增大模型在目标任务上的有效信息密度;或是“做加法”:将已有的多个模型划分为模块,再将在目标任务上最有用的模型进行重组得到新模型,实现减少模型参数量的同时实现朝目标任务的知识迁移.“做减法”的典型方法包括文献[28-30]等;“做加法”的典型方法包括文献[31-32]等.这类方法有着极好的扩展性——在有限的可用模块下,可以出近乎于无限多种排列组合,理论上可以满足几乎任意的迁移任务需求,且可以很方便地随时作出调整.但同时这类方法的计算成本也较高——想要算出某任务下的最优排列有时候是 NP 难的,现有方法采用的往往是计算较简单的近似解.

(2)预训练参数是否改变.综述[26]主要介绍的是从零训练深度模型的高效方法,因此不涉及对预训练参数的更新.本文介绍的方法沿用微调[8]的思想,希望找出预训练模型在目标任务上最相关的参数进行更新.不同于传统微调对所有参数进行更新,这类方法基于模型在关键目标样本和任务上的表现评估每个参数模块的重要性,有选择地以不同的权重更新参数.此类微调方式灵活性、适应性强,能根据具体的计算开销限制对权重进行调整,兼顾效果和开销,且形式多样,将会在后文详细介绍.这类方法易于实现,但扩展性较差,只要预训练模型的结构和总参数不发生改变,算法的上限就已经确定.

(3)输入(输出)是否改变.综述[26]中介绍的数据增强(如 RandAug[33]等)通过多样的数据变换在数据欠缺的情况下提升模型效果,实现数据高效的深度学习.本文介绍的方法则通过对输入数据的变换在不改变模型参数的情况下改变模型的表现,实现模型迁移,其数据变换手段与目的都与文献[26]完全不同.这类方法无需对模型本身进行任何调整,甚至无需了解模型的具体结构,也能实现有效的模型适应和迁移.且对计算资源的要求极低、计算速度极快,是资源限制较严格时的最优选.典型的方法为提示学习[34-36].这类方法有着较好的扩展性——计算成本平衡,能在较低的计算开销下实现对不同任务、不同情况的适应,实用性较强.

(4)是否注入自适应参数.类似于方法(3)中修改输入而不修改模型的思路,一些高效迁移方法提出修改模型中间变量实现模型迁移,而同样不需要修改原始模型.常见的修改方法包括将一段可训练的参数拼接至原有变量头尾处.但这种方法与方法(3)相比在对预训练模型结构有一定限制,主要被用

于修改 Transformer 结构中的编码、键、值等序列化变量.综述[26]则未涉及此类方法.这类方法扩展性一般.注入的参数位置、大小、注入方式固定,有时不能对样本级别的动态变化作出应对.但相对地计算开销也较低,几乎不扩增原始模型的参数量.

(5)是否引入自适应模块.方法(4)注入的自适应参数虽然开销很小,但也带来了修改程度有限、灵活度有限等问题.另一些方法则在预训练模型的模块之间增加轻量化的自适应模块,以模型中间变量为输入对每个样本数据进行自适应修改后输出,实现更灵活的模型迁移.这类方法发展较为完善,主要的研究热点在于调整自适应模块的位置和具体设计.典型的方法是适应模块法(Adapter),如文献[37-38]等.综述[26]未涉及此类方法.这类方法较方法(4)提升了扩展性,但也提升了计算开销.自适应模块能捕捉更细致、更立体的样本级差异,实现更细致的迁移.

需要注意的是,许多高效迁移方法融合了上述的多种高效迁移技术以达到更好的效果,因此可能同时具备多个技术特点.上述的迁移技术之间也有内在的异同关系,这些关系导致特定的方法组合会带来更高的收益,因此常常共同出现.如技术(1)结构改变和技术(2)参数改变的组合在实际应用中较为常见,在修改模型结构后往往需要对部分参数重新调整以适应新的模型结构.

本文将以这 5 个技术特点为基石揭示已有的代表性方法实现高效迁移的基本原理,并比较分析这 5 个技术特点在不同应用场景下的实现与效果的异同.具体而言,在第 2 节简要介绍迁移学习的问题描述和定义,介绍不同问题设定并指出本文重点关注的方法所解决的问题;在第 3 节介绍自然语言处理领域的方法;在第 4 节介绍计算机视觉领域的方法;在第 5 节介绍多模态模型领域的方法;在第 6 节分析现存方法的局限和未来研究方向;最后在第 7 节进行总结.

2 问题定义

在本节对迁移学习问题进行规范的描述,并就本文重点研究的问题设定和分支进行阐明.为了便于读者理解,本节叙述时会辅以深度学习中的常见例子进行说明.迁移学习是一种机器学习范式,其核心思想可以概括为"旧知识用于新问题",利用已经见过的信息来解决新的但相关的问题,从而提高学习效率和效果.通过找到不同但相关的领域之间的相似性并加以合理利用,迁移学习能减少在目标领域直接从零学习所需要的数据量,从而降低计算开销、缩短训练时间、提升泛化能力.迁移学习的适用场景非常广泛,可以应用于计算机视觉、自然语言处理、语音识别、情感分析等,体现了知识的延续性和连通性,是提高机器学习应用灵活性、鲁棒性和实用性的重要手段.接下来给出其形式化定义.

一个领域 $D=X,P(X)$由其中的数据 $X=\{x_1,x_2,\cdots,x_n\}$及数据的边缘分布 $P(X)$确定.以一个英语语料库为例,数据 X 即为所有的英语单词,而数据分布 $P(X)$会因为具体的子语料库(如法律相关、金融相关等)而有所不同.一个任务 $T=\{Y,f(\cdot)\}$由标签空间 $Y=\{y_1,y_2,\cdots,y_n\}$和功能函数 $f(\cdot)$构成.例如对于英译中任务,标签空间中的每个标签个体 y_i为对应英语单词 x_i的中文释义,$f(\cdot)$为翻译模块,期望其具有功能 $y_i=f(x_i)$.可以进一步定义源域 $D_s=\{X_s,P_s(X_s)\}$,目标域 $D_t=\{X_t,P_t(X_t)\}$,源任务 $T_s=\{Y_s,f_s(\cdot)\}$和目标任务 $T_t=\{Y_t,f_t(\cdot)\}$.

基于一个源域 D_s和对应的源任务 T_s,以及一个与源域相关但不同的目标域 D_t和目标任务 T_t,迁移学习的目的是利用源域中的知识尽可能地改善目标任务上功能函数 f_t的性能.在实际应用场景中,源域上往往有较丰富、质量较高的训练数据,而目标域的有标注数据较少或质量较低.根据领域的定义,源域和目标域的不同可以理解为数据不同 $X_s\neq X_t$,或数据分布不同 $P_s(X_s)\neq P_t(X_t)$.根据综述[18]的定义,数据不同的迁移学习被称为异构迁移学习(heterogeneous transfer learning),例如英译中数据集和日译中数据集;数据分布不同的迁移学习被称为同构迁移学习(homogeneous transfer learning),例如法律领域的英译中数据集和金融领域的英译中数据集.根据任务的定义,源任务和目标任务的不同可以理解为标签空间不同 $Y_s\neq Y_t$,或功能函数不同 $f_s(\cdot)\neq f_t(\cdot)$.根据综述[20]的定义,功能函数的不同被称为源域和目标域之间的上下文偏差,如同一个英文单词在不同语境中应当译为不同的含义;标签空间不同可以用语义情感分析任务来理解:例如一些任务只要求对句子的情感进行二分类(积极、消极),另一些任务要求对情感等级进行细化,如分为 1~5 级的多分类问题.另外,根据综述[18],目标任务上存在标签 Y_t的迁移学习被称为归纳式迁移学习(inductive transfer learning)、目标任务上不存在

标签时则称为直推式迁移学习(transductive transfer learning).另外表 1 介绍的文献综述中还归纳有无监督迁移学习、领域自适应[39]、负迁移等相关定义，但并非本文介绍重点，因此不在此展开叙述.本文介绍的方法均属于同构的归纳式迁移学习.

由于主流深度学习模型的参数量极大，且已有性能优秀的开源模型，目前迁移学习算法的研究均是基于源域上的预训练权重开展的，这也是本文的重点研究设定：基于在源域 D_s 和源任务 T_s 上预训练得到的权重 θ_{f_s}，研究高效迁移方法 F，使迁移后的权重 $\theta_{f_t}=F(\theta_{f_s})$ 所构成的功能函数 f_t 在目标域、目标任务上有较好的表现，且迁移方法 F 的开销应尽可能小.根据现有文献的侧重点，此处的开销主要指模型的参数量、训练推理时的内存占用.接下来的叙述将围绕 $F(\theta_{f_s})$ 展开，用不同 F 的定义形式化地描述不同方法的核心技术.

3　基于 Transformer 的自然语言处理高效迁移方法

自然语言处理是深度学习的一个十分重要的应用领域，具体应用包含语言理解、翻译、对话、问答等.考虑到目前参数量最大的模型均属于自然语言大模型(如图 1 所示)，高效迁移学习算法在本领域有极大的应用需求，既可以将市面上的公开大模型预训练权重方便地迁移至用户所需的目标领域，也可以为 AI 公司提供大模型轻量复用方案，从而避免类似任务上重复训练带来的巨大开销.目前自然语言处理领域效果最好、应用最广泛的是 Transformer 模型[40]及其变种，在此对其基本结构进行简要介绍.

图 3 展示了 Transformer 模型中的一个基本模块，不同规模的 Transformer 一般是由不同宽度/深度、不同数量的基本模块堆叠而成.原始的中间变量被划分为多段：$x=[x^1,x^2,\cdots,x^i,\cdots,x^n]$，以中间变量序列的一部分 x^i 为输入进行简单的线性变换后将其结果送入自注意力模块：

$$\begin{aligned} head^i &= Attn(x^i\boldsymbol{W}_q^i, x^i\boldsymbol{W}_k^i, x^i\boldsymbol{W}_v^i) \\ &= \mathrm{Softmax}\left(\frac{(x^i\boldsymbol{W}_q^i)(x^i\boldsymbol{W}_k^i)^{\mathrm{T}}}{\sqrt{d_k}}(x^i\boldsymbol{W}_v^i)\right) \end{aligned} \quad (1)$$

其中 d_k 为这一段中间变量序列的宽度，$head^i$ 为这一段中间变量序列经过计算得到的注意力，$\boldsymbol{W}_q^i$、$\boldsymbol{W}_k^i$、$\boldsymbol{W}_v^i$ 分别为询问的变换矩阵，键的变换矩阵和值的变换矩阵.最后把不同段的注意力拼接(concat 操作)，得到原始中间变量对应的注意力向量：

$$head=\mathrm{concat}(head^0,head^1,\cdots,head^n) \quad (2)$$

再将注意力向量与原始的中间变量相加、进行层归一化变换后被送入由全连接层和非线性激活层构成的前馈网络，其输出再与前馈网络的输入相加，得到基本模块的输出值，如图 3 所示.

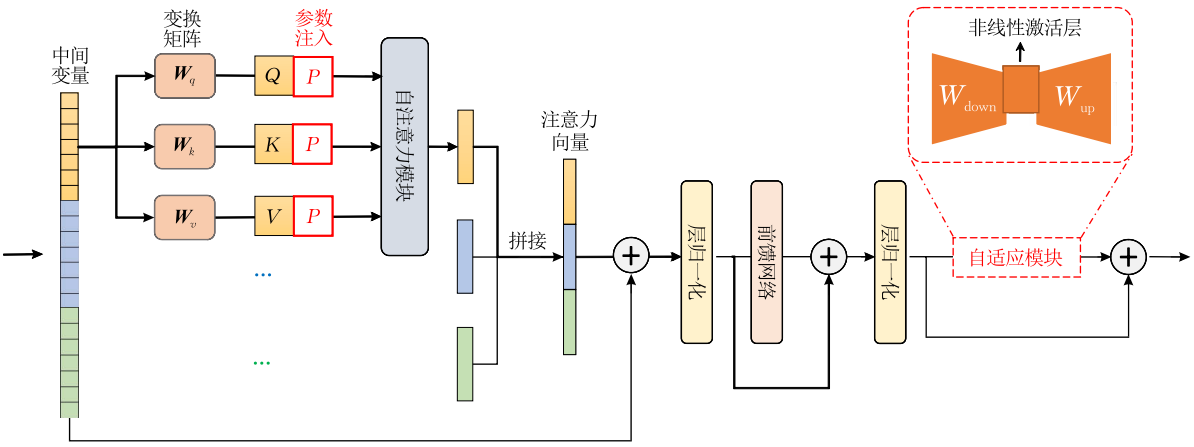

图 3　Transformer 模型基本结构示意图(其中“参数注入”和“自适应模块”区域为部分高效迁移方法实现原理)

针对基于 Transformer 的语言模型，主流方法以图 2 介绍的(d)注入自适应参数和(e)引入自适应模块方法实现高效迁移，如图 3 的红色虚线模块所示.需要注意的是图 3 提供的仅是这些方法的概念性示意图，不同方法的具体参数注入、增加模块位置与形式不尽相同，将会在后文详细介绍.表 2 总结了自然语言处理领域较有代表性的高效模型迁移方法，其中的黑点代表该方法所属的高效迁移技术类别.接下来在 3.1 节和 3.2 节对参数注入和自适应模块法在自然语言处理领域的应用进行详细分析；在 3.3 节对其他方法和自然语言处理领域的已有总结性工作进行介绍；最后在 3.4 节介绍该领域常用的迁移数据集.

表 2 自然语言处理领域的部分高效迁移方法

方法	概述	修改结构	更新参数	修改输入	注入参数	增加模块	部分验证数据集与指标	参数占比	内存开销
Compacter[38]	基于参数化的超复数乘法层[41](parameterized hypercomplex multiplication layers)设计高效的 Adapter 进行模型迁移					●	GLUE：85.3 SuperGLUE：71.82	0.05%	−42%
Prefix-tuning[42]	提出优化一种连续的任务相关向量(前缀)来替代传统的微调方法.在预训练 Transformer 的编码/解码层序列中加入前缀可以影响模型表现				●		E2E：70.3 WebNLG：63.4	0.1%	—
PromptTuning[43]	可以被看作 Prefix-tuning[42] 的简化版本，即只在输入句子分词后得到的 token 序列增加前缀					●	SuperGLUE：90.5	<0.01%	—
LoRA[44]	通过向 Transformer 层注入低秩分解矩阵近似权重更新过程，实现模型迁移				●		E2E：70.4 GLUE：87.2	0.24%	−66%
QLoRA[45]	基于 LoRA 提出一种降低微调时显存开销的方法，微调 650 亿参数只需 48 G 显存，且与一般的微调方法相比几乎不导致性能下降		●		●		GLUE：88.8	—	—
BitFit[46]	提出只微调模型的偏置参数实现高效模型迁移		●				GLUE：85.0	0.09%	NA
AdapterFusion[47]	提出一种两段式框架.第一阶段为知识提取，在此阶段内训练得到任务相关的 Adapter；第二阶段为知识重组，将不同 Adapter 得到的表征结合以提升目标任务的表现					●	SICK：90.43 RTE：79.96 CB：89.81	—	—
$(IA)^3$[48]	通过学习向量来放缩注意力头中的键、值以及前馈层中的激活层，以实现高效的知识迁移				●		T0：72.4	0.01%	—
LST[49]	训练一个单独的梯子网络，该网络以预训练网络的中间激活层输出作为输入进行预测.该网络更新时不需要反向传播至整个预训练骨干网络中，因此十分高效					●	GLUE：84.1	1.74%	−37%
ATTEMPT[50]	训练多个源任务上的提示和一个注意力模块以混合提示				●	●	GLUE：85.2 SuperGLUE：74.1	0.04%	—
Adapters[51]	提出在 Transformer 层之间增加自适应模块(Adapter)进行高效迁移					●	GLUE：80.9	3.6%	—
Diff pruning[52]	通过学习一个任务相关的 diff 向量实现迁移，而无需存储其他预训练参数				●		GLUE：84.5	1%	—
HyperFormer[53]	提出一种统一的 Adapter 学习框架，能跨任务生成、共享 Adapter 参数和知识					●	GLUE：85.2	0.29%	−25%
T5[14]	为自然语言处理任务提出一个统一的框架，将所有基于文字的任务转换成文字-文字的形式，从而实现不同任务间的方法共享	●	●	●			GLUE：86.4	—	—

注：验证数据集与指标和表 3 中信息对应，参数占比、内存开销分别指代高效方法所调参数相较于所有参数的比例、高效方法训练时较调试所有参数时节省的内存.

3.1 基于自适应参数注入的方法

不同于传统的序列数据处理模型[54]，Transformer 以全局计算注意力的方式实现了对长序列中无论远近部分的高效信息转换提取，因此在语言处理领域取得极好的效果.考虑到 Transformer 中的核心参数均以序列化的形式体现，自适应参数注入法希望通过对序列的增删改等操作影响模型计算注意力时的结果，从而实现对模型性能的控制.

如图 3 所示，Transformer 模块中的核心操作就是自注意力计算，因此以 Prefix-tuning[42] 为代表的许多方法选择在 Transformer 基础模块中的键值等向量头部增加长度前缀.以 Prefix-tuning 为例，令 $\boldsymbol{P}_k$，$\boldsymbol{P}_v$ 为注入到键、值的前缀向量，通过修改式(1)可以得到注入参数后的注意力模块：

$$head_{PT}^{i}=Attn(x^{i}\boldsymbol{W}_{q}^{i},\text{concat}(\boldsymbol{P}_{k},x^{i}\boldsymbol{W}_{k}^{i}),\text{concat}(\boldsymbol{P}_{v},x^{i}\boldsymbol{W}_{v}^{i})) \tag{3}$$

其中 concat 为向量拼接操作.这种方法无需修改原始的预训练变换矩阵，只需要训练前缀向量就可以实现知识迁移.基于第 2 节的定义，可以把文献[42]的迁移方法描述为 $F(\theta_{f_s})=\{\theta_{f_s},\theta_P\}$，其中 θ_P 为可训练的前缀向量参数，其参数量仅为整个 Transformer 模型的参数量的 0.1%，真正实现了“四两拨千斤”.基于自适应参数注入的方法整体框架与文献[42]大同小异，实现细节之中的不同点主要体现在：

(1) 注入位置不同.Prefix-tuning[42] 将参数注入在模型内部的键、值前；而 PromptTuning[43](提示学习)无需了解修改模型内部结果，只在原始输入经过分词后、进入模型前得到的 token 序列增加前缀，较 Prefix-tuning 更灵活.不同于上述增加、拼接的注入方式，Diff pruning[52] 通过与原始参数相加的方式注入参数，因此其位置是不固定、可学习的，

也可以理解为是全局的. 另外，Diff pruning 学到的注入参数由一个 L_0 范数惩罚项约束，类似于训练深度网络时采用的 L_2 正则项，以便让学到的参数更稀疏、减少多余参数. 类似地，LoRA[44] 将参数注入在所有预训练权重矩阵中，但其注入方式稍有不同，且其目的是模拟预训练权重的更新. (IA)^3[48] 的注入位置除了键、值处还包括激活层中，但不同于 Prefix-tuning 用直接拼接的方式修改键值等变量，其参数作为系数对变量进行放缩.

(2) 注入方法不同. Prefix-tuning[42] 和 Prompt-Tuning[43] 通过向量拼接的方式注入参数：$F(\theta_{f_s})=\{\theta_{f_s},\theta_P\}$. Diff pruning[52] 通过直接相加的方法注入参数，即 $F(\theta_{f_s})=\theta_{f_s}+\theta_P$. LoRA[44] 通过低秩矩阵来模拟预训练参数的更新. 假设某参数矩阵为 $\boldsymbol{W}_0\in R^{d\times k}$，则可以用两个低秩矩阵 $\boldsymbol{A},\boldsymbol{B}$ 的相乘模拟其更新：$\boldsymbol{W}_0+\Delta\boldsymbol{W}=\boldsymbol{W}_0+\boldsymbol{BA}$，其中 $\Delta\boldsymbol{W}$ 是更新参数，$\boldsymbol{B}\in R^{d\times r}$，$\boldsymbol{A}\in R^{r\times k}$，且矩阵 $\boldsymbol{A},\boldsymbol{B}$ 的秩 r 很小：$r\ll\min(d,k)$. 其参数注入方法可以描述为 $F(\theta_{f_s})=\{\theta_\Delta+\theta_B\theta_A,\theta_{f_s}\backslash\theta_\Delta\}$，其中 θ_Δ 是进行更新的预训练参数. 值得注意的是，LoRA 在更新过程中只需要令两个低秩矩阵参与计算，因此其计算内存开销也大幅下降，如表 2 所示，其在节约 66% 内存的情况下仍能实现多个基准上的高性能，因此已经被广泛应用于高效迁移领域，作为技术基础激发了更多相关工作的诞生，如利用频域进行权重分解的 DoRA[55]、动态分配参数预算的 DoRA[56]、利用可学向量来调整不同层间的随机矩阵的 Vera[57]，等等. (IA)^3[48] 通过元素对应相乘的方法注入参数：$F(\theta_{f_s})=\{k\theta_\Delta,\theta_{f_s}\backslash\theta_\Delta\}$，其中 k 是注入的参数(放缩倍数). ATTEMPT[50] 考虑的问题包含多个源域 $D_{s_0},D_{s_1},\cdots,D_{s_n}$. 该方法首先基于 PromptTuning[43] 在每个源域上学到一条提示 θ_{p_i}，$i=1,\cdots,n$，又在目标域上初始化一条目标提示 $\theta_{p_{n+1}}$. 该方法设计一个注意力模块将学到的不同提示根据目标样本进行动态的加权混合以达到最优效果：$F(\theta_{f_s})=\{\theta_{f_s},\theta_{p_{n+1}}+\sum_{j=1}^{n+1}\alpha_j\theta_{p_j}\}$，其中 α_j 是对应每条提示的权重，且 $\sum_{j=1}^{n+1}\alpha_j=1$.

3.2　基于自适应模块的方法

以 Adapter[51] 为代表的自适应模块法是语言大模型高效迁移领域较为经典、发展较完善的一类方法. 如图 2(e)、图 3 所示，其基本结构是在预训练模块之间增加自适应模块对中间变量进行修改后与原变量相加，达到用少量参数迁移较大模型的目的. 此方法与参数注入法最大的区别在于其自适应能力，能通过一个独立的模块建模需要对中间参数进行的修正，而注入的参数则相对较固定. 此类别的绝大多数方法都可以基于文献[51]的设计表示为

$$x\leftarrow x+l(x\boldsymbol{W}_{\text{down}})\boldsymbol{W}_{\text{up}} \tag{4}$$

其中 x 是中间变量，l 是非线性激活层(如 ReLU 等)，$\boldsymbol{W}_{\text{down}}$ 和 $\boldsymbol{W}_{\text{up}}$ 是降采样全连接层和升采样全连接层，它们共同构成一个瓶颈模块，且不会改变原始输入输出的维度.

基于自适应模块的大部分方法在模块的具体设计上有所不同，但都可以用 $F(\theta_{f_s})=\{\theta_{f_s},\theta_{\text{adapter}}\}$ 表示，其中 θ_{adapter} 代表自适应模块的参数，在 Adapter 方法[51] 上即为 $\theta_{\text{adapter}}=\{\theta_{W_{\text{down}}},\theta_{W_{\text{up}}}\}$. Compacter[38] 设计了参数共享的“慢”矩阵 $\boldsymbol{A}_i$ 用于捕捉较一般的知识，以及任务相关的“快”矩阵 $\boldsymbol{B}_i$ 用于学习特定任务下的特定参数，并利用矩阵元素相乘 $\boldsymbol{W}=\sum_{i=1}^{n}\boldsymbol{A}_i\otimes\boldsymbol{B}_i$ 计算得到降采样、升采样层 $\boldsymbol{W}_{\text{down}}$ 和 $\boldsymbol{W}_{\text{up}}$.

也有一些方法研究如何更高效地应用自适应模块. 与文献[51]的单一 Adapter 不同，针对在不同源任务 $T_{s_0},T_{s_1},\cdots,T_{s_n}$ 上学好的源域 Adapter 参数 θ_{a_i}，AdapterFusion[47] 进一步引入一个具有参数 θ_{fusion} 的混合模块，从而根据不同的目标任务混合源域 Adapter：$\theta_{a_t}=\theta_{\text{fusion}}(\theta_{a_0},\theta_{a_1},\cdots,\theta_{a_n})$. LST[49] 观察到经典的 Adapter 框架(式(4))虽然无需更新预训练模型的参数，但仍需要计算这些参数中的梯度用于训练 Adapter，当预训练参数较多时也会带来不小的开销. 因此该方法提出一种梯子网络，将梯度在预训练参数和 Adapter 中的前后传播从大模型中剥离出来单独训练，从而避免在预训练参数上的梯度计算开销. HyperFormer[53] 提出采用一个共享的超网络 θ_h 生成不同位置 Adapter 的参数：$\theta_{\text{adapter}}=\theta_h(\boldsymbol{x}_i)$，其中 $\boldsymbol{x}_i$ 是目标任务或输入目标样本的编码向量，从而将训练多个 Adapter 的开销进一步减少为训练一个超网络的开销.

3.3　其他方法

除了参数注入和自适应模块法，也有部分自然语言处理方法选择参数微调以实现高效迁移. BitFit[46] 通过实验证明仅微调偏置项(线性变换 $\boldsymbol{W}x+b$ 中的 b)即可以达到与微调全部参数相匹配的效果. T5[14] 设计了一个统一框架处理自然语言处理任务，并贡献了一个数据集 C4. 回顾第 2 节中的定义，可以发现不同的目标任务 T_t 是由不同的功能函数(模型)

f_t实现的,这从本质上要求对每个不同任务有一个不同的模型. T5 希望将所有任务用一个超大预训练网络解决,只需要解决不同任务所需的输入输出形式即可. 如针对翻译任务,可以设计输入为"请把句子 Hello world. 翻译为中文",而对情感分析任务,只需输入"请判断情感:这部电影真好看". 模型则会根据输入的不同自动判断任务,并返回对应的结果. 这种方式又被称为指示学习(instruction learning),相关的方法还有 FLAN[58] 等. T5 框架的设计过程包含对采用的 Transformer 模型结构(修改结构法)、参数训练方法(更新参数法)、具体超参数的探索和最优解确定. T5 中涉及到对原始输入句子的破坏,并要求模型补全缺失的部分,因此也涉及对输入的修改. QLoRA[45] 解决了微调预训练模型时所需显存过大的问题. 该方法设计了一种 4-bit 量化技术和双重量化技术实现高分辨率的 4-bit 微调,还设计了一种换页优化器防止训练时的内存不足. 基于该方法,可以在单张消费级算卡上实现对 330 亿参数的微调,在单张专业级算卡上实现对 650 亿参数的微调,且不会导致模型性能下降. 文章还估计该方法每晚可以在一台 iPhone 12 Plus 上微调 300 万个 token 的数据. QLoRA 大幅降低了微调超大型预训练模型的算力开销,是高效迁移学习领域十分重要的工作.

上下文学习[59](in-context learning)提供另一种无需调整模型参数的方法,通过设计自然语言示例的模板实现模型在目标任务上的提升. 例如一个情感判断任务,上下文学习将"好吃的食物——正面,糟糕的服务——负面"等数据对设计为"客户评价:好吃的食物. 情感:正面." "客户评价:糟糕的服务. 情感:负面."的示例输入到模型中,期望模型学到这些示例中的数据分布和隐藏模式,从而对没见过的数据做出预测. 文献[59]进一步指出,上下文学习中示例里的输入——标签是否匹配实际上并不重要,将其替换为随机的单词也只会略微影响性能. 模型实际上学习的是标签空间的分布、示例格式和示例重构输入语句的方式. 文献[60]提供了上下文学习的更详细介绍.

文献[61]对参数注入法和自适应模块法进行总结,提出一个统一的高效迁移框架来表示这两种类别的方法. 文献[62]将高效迁移方法概括为 delta-tuning,并总结了不同方法、数据集上的实验表现. 文献[63]不仅总结了大语言模型上的常用高效迁移方法,还就高效算法的实际部署、系统设计和实现开销等方面进行探讨. 文献[64]对现代大语言模型的发展、技术特点、应用场景等进行了全面的综述.

总的来说,自然语言处理领域具有模型规模大、输入序列化特征明显、数据量大等特点,因此应用最广泛的高效迁移技术之一是"参数注入". 这种方法的优势在于,它能在不调整模型整体结构、不更新原始模型中绝大部分参数的情况下实现对预训练模型的微调和适应. 这不仅大大减少了计算资源的消耗,还能有效防止过拟合,特别是在小样本学习场景中表现出色. 此外,提示学习方法还能更好地利用预训练模型中蕴含的知识,通过提示设计激活模型的潜在能力,从而在各种下游任务中取得不错的效果. 另一类应用广泛的是自适应模块(Adapter)系列方法,其与提示学习的区别在于 Adapter 将可学习参数进行了模块化封装,形成独立的迁移子模块,能满足更加复杂和灵活的微调需求.

3.4 常用数据集与实验结果

表 3 总结了自然语言处理领域常见的典型任务以及对应的常用数据集. 特别地,部分主流数据集已经被纳入公开的自然语言处理基准挑战中,如近年来最为常用和权威的 GLUE[91] 和 SuperGLUE[92]. 本文在表中也提供了相应的标注,指明数据集所归属的"基准"(若有). 从表中可以看出,自然语言处理领域的任务形式多样、检验方式复杂,对现代的语言大模型提出了极高的要求. 这些任务从最基本的句子合法性检查,到同义词、代词理解,再到复杂段落下的内容理解与推理,涵盖了人们生产生活中的常见语境场景,能很好地检测现代大语言模型的能力.

本文在应用最广泛、覆盖任务最全面的评价基准之一 GLUE(包含 8 个子数据集)上对前文所介绍的部分代表性方法进行综合的实验结果对比,如表 4 所示. 可以发现,自然语言处理任务的骨干网络种类繁多,基于不同骨干网络微调的方法性能差异较大;不同方法所调参数量也有所不同,特别地,涉及增加自适应模块的方法平均需要调整更多参数. 基于 RoBERTa-base 的 LoRA[44] 在对比的方法中取得了最好的平均表现,且调整参数量较少. 其优异的整体性能也为后续的大量相关研究工作打下了基础.

表 3　自然语言处理领域常用数据集

任务	数据集	基准	描述	指标
合法性检验	CoLA[65]	GLUE	给出一个单词序列，模型需要判断此序列是否构成合乎语法的英语句子	Matthews 相关系数[66]
情感判断	SST-2[67]	GLUE	给出一条影评，模型需要判断其情感倾向是正面或负面	准确率
释义	MRPC[68]	GLUE	给出一对新闻相关的表达，模型需要判断其是否在语义上等价	准确率、$F1$
	QQP[69]	GLUE	给出一对源于问答网站 Quora 的表达，模型需要判断其是否在语义上等价	准确率、$F1$
相似度判断	SICK[70]	—	给出一对句子，模型需要判断其关联程度为关联、中性或是矛盾	准确率
	STS-B[71]	GLUE	给出新闻标题、视频图像标题等构成的样本对，并提供人工标注的相似度指标(1～5)，模型需要预测这些指标	皮尔森、斯皮尔曼相关系数
推理任务	MNLI[72]	GLUE	给出一个前提句和一个假设句，模型需要预测前提是否：蕴含假设(蕴涵)、与假设相矛盾(矛盾)、或者两者都不是(中性)	准确率
	QNLI[73-75]	GLUE	给出问句和段落构成的数据对，模型需要判断段落中是否包含问句的答案	准确率
	RTE[76-79]	GLUE SuperGLUE	给出样本对，模型需要判断样本对之间是否有蕴含关系	准确率
	WNLI[80]	GLUE	模型需要阅读一个含有代词的句子，并从一系列选项中选择该代词的指代对象	准确率
	CB[81]	SuperGLUE	给出一个对话，模型需要判断说话人对这段话的内容真相是否知情，包括知晓、认为、中立、不知	准确率、$F1$
问答	BoolQ[82]	SuperGLUE	给出一段话和一个问题，模型需要回答“是”或“否”	准确率
	COPA[83]	SuperGLUE	给出一个前提和两个可能选项，模型需要判断是哪个选项导致了前提	准确率
	MultiRC[84]	SuperGLUE	给出一段话，一个问题和多个可能的答案，模型需要判断每一个答案的对错	$F1$、准确匹配率
	ReCoRD[85]	SuperGLUE	给出一个段落和完形填空式的问题，模型需要从段落中选择合适的主体填入空中	$F1$、准确匹配率
多义词检测	WiC[86]	SuperGLUE	给出两个包含同一个多义词的句子，模型需要判断该多义词在两个句子中是否代表同样的含义	准确率
共指消解	WSC[80]	SuperGLUE	给出一个含有指代词的句子，模型需要在一系列选项中判断该指代词的指代对象	准确率
语句生成	E2E[87]	—	端到端生成数据集，通过部分属性如人名、位置、评价等生成一个句子. 本数据集在餐馆评价语境下生成句子	BLEU[88] MET[89]
	WebNLG[90]	—	包含从 DBpedia 中抽取一组三元组并采用自然语言生成方法获得所构造的句子. 每个样本由一组三元组和一条标准句子构成	BLEU[88] MET[89]

表 4　自然语言处理领域部分代表性方法在 GLUE 基准数据集上的实验结果

方法	骨干模型	会议/期刊	参数占比/%	CoLA	SST-2	MRPC	QQP	STS-B	MNLI	QNLI	RTE
Adapters[51]	BERT-large	ICML	3.60	59.5	94.2	89.6	71.8	87.3	84.6	91.4	68.8
Diff pruning[52]	BERT-large	ACL	1.00	62.8	94.2	91.4	86.6	89.9	87.0	93.3	71.1
LST[49]	T5-base	NeurIPS	1.74	58.1	94.1	90.4	88.8	90.7	85.6	93.3	71.9
ATTEMPT[50]	T5-base	EMNLP	0.04	64.3	93.7	86.1	90.0	90.8	83.8	93.1	79.9
Compacter[38]	T5-base	NeurIPS	0.05	63.8	93.0	89.2	90.2	90.3	85.6	92.9	77.7
HyperFormer[53]	T5-base	ACL	0.29	63.7	94.0	89.7	90.3	90.0	85.7	93.0	75.4
BitFit[46]	RoBERTa-base	ACL	0.09	61.8	93.7	92.0	84.5	93.7	85.2	91.3	77.8
LoRA[44]	RoBERTa-base	ICLR	0.24	63.4	95.1	89.7	90.8	91.5	87.5	93.3	86.6

4　深度视觉模型上的高效迁移方法

计算机视觉近年来在面部识别、智能驾驶、图像生成、动作分割[93]等应用领域高速发展. 以智能驾驶为例，车载平台的算力往往十分受限，但智驾系统中的一些核心算法，如行人、非机动车的轨迹预测[94]等，往往需要引入较大较复杂的深度模型以确保其高性能，因此需要高效迁移算法将道路状况识别、处理大模型以较低的开销迁移至车辆端. 视觉信息处理在许多可穿戴设备上也是不可或缺的基础能力，只有高效迁移算法有能力在此类设备所支持的算力下迁移大型视觉模型. 作为深度学习的经典应用之一，用于处理图像的深度网络始于早期的 AlexNet[95](2012)，随后的数年内研究者通过不断扩宽、加深深度网络来加强网络的拟合感知能力，代表性的成果包括 VGG[96](2014)和 GoogLeNet[97](2015)等. 但若单纯地通过增加参数量、网络层数来提升网络性能，在网络参数量达到一个阈值后便收效甚微，如 ResNet101 的参数量较 ResNet50 几乎翻了一倍，但性能提升却不到 1%.

另外，过深过大的网络不仅大大提升了开销、降

低了训练效率,还导致过拟合、梯度消失等问题.因此后继的研究者通过设计残差连接等新结构来解决上述问题,得到的代表性产物之一就包括 ResNet[8](2016)及变种 ResNeXt[98](2017)、ResNeSt[99](2022)等.预训练于 ImageNet 的 ResNet50、ResNet101 等开箱即用的骨干网络也开启了基于预训练权重的研究时代,在今天仍扮演着十分重要的角色.随着文献[40]提出的 Transformer 结构在自然语言处理领域大放异彩,研究者很快将其核心:自注意力模块应用到图像处理领域,设计得到了视觉 Transformer[100]模型(Vision Transformer,ViT)(2020)及变种 SwinTransformer[101](2021)等,在计算机视觉领域取得了前所未有的极佳效果.它们以及更多更新结构的模型预训练权重也几乎成为了当今所有视觉研究的基石.

随着模型结构的进化和应用需求的变革,模型的参数越来越多、训练计算开销越来越大,因此也吸引了越来越多的研究者就如何平衡模型效果和模型开销之间的关系展开探索.DenseNet[102](2017)提出密集连接的设计让特征能重复利用;MobileNet[103](2017)、MobileNetV2[27](2018)设计了深度可分离卷积来减少参数量;EfficientNet[104](2019)系统性地研究了卷积网络的宽度、深度、分辨率和网络表现之间的关系,提出了一种最优的缩放参数来扩展模型规模.这些研究主要针对模型本身的设计和改良以试图训练出更好更高效的模型,在文献[26]有较详细的介绍.本文关注的重点是如何更好更快更高效地调整、迁移预训练模型以服务于指定的目标任务,这也渐渐成为了深度学习领域的研究热点.

计算机视觉与自然语言处理领域既处处不同又息息相关.同样作为深度学习的应用领域,在其中实施模型迁移的基本目标和定义是相同的.虽然原始的输入模态不同,但自然语言经过分词、图像经过 RGB 空间的转换后均被参数化为张量,再经过骨干网络处理后被嵌入到高维空间.因此对于下游的任务处理模块来说,处理语言抑或图像数据在形式上没有任何区别.在模型结构上,经典的视觉深度模型由具有局部感知能力的卷积层(Convolutional Neural Networks,CNN)构成,而语言模型由能提取序列前后关系的递归神经网络(Recurrent Neural Networks,RNN)构成.但 Transformer 结构的提出几乎统一了视觉和语言模型,这也使许多应用于语言处理领域的迁移学习方法能几乎不加以修改地应用于视觉领域.总的来说,现有的视觉领域与语言领域差异较小,有许多共享的高效迁移方法,而视觉领域目前流行的网络结构更为丰富,因此又催生了一批具有视觉领域特点的迁移方法.表 5 汇总了计算机视觉领域较有代表性的一批方法,其中的黑点代表该方法所属的高效迁移技术类别.接下来根据表 6 展开介绍,在 4.1 节介绍修改结构的方法;在 4.2 节介绍更新参数的方法;在 4.3 节介绍注入自适应参数的方法;在 4.4 节介绍增加自适应模块的方法;在 4.5 介绍其他方法并总结计算机视觉领域方法的特点;最后在 4.6 节介绍常用的视觉数据集.

表 5 计算机视觉领域的部分高效模型迁移方法汇总

方法	概述	修改结构	更新参数	修改输入	注入参数	增加模块	部分验证数据集与指标	参数占比
VDP[105]	通过在输入图像上加小像素块实现视觉提示,在无需训练源模型的情况下实现实时迁移				●		CIFAR100:83.2	—
DeRy[31]	将异构的预训练模型划分为等价模块,再根据目标任务和计算开销要求重组为新模型	●				●	ImageNet:78.6	1.57%
FDA-ES[106]	根据目标数据的推理情况设计多个提前退出点来节省推理开销		●	●			Office31:85.3 ImageCLEF:88.5	—
Hub-Pathway[32]	根据每个数据点的特点从模型池中选择最合适的预训练网络,并加以训练强化		●				分类任务基准数据集:88.2	—
Zoo-Tuning[107]	设计了一种通道对齐层和参数聚合网络来自适应地把不同数据上预训练的源域模型参数聚合为目标任务上的目标模型	●	●				分类任务基准数据集:85.2	—
VPT[108]	在 ViT 的输入编码层和键、值等序列处加入可学习的提示向量实现迁移	●			●		FGVC:89.1 VTAB-1k:72.0	4.90%
Conv-Adapter[109]	提出一种由卷积网络构成的自适应模块					●	FGVC:84.2 VTAB-1k:76.1	5.70%
TinyTL[110]	提出一种高效的偏置模块,并只调整偏置实现内存高效的迁移学习,使内存开销大幅度降低		●			●	CIFAR100:81.4	—
AdaptFormer[111]	在 ViT 上应用自适应模块实现高效自适应					●	CIFAR100:85.9 SVHN:96.9	0.36%
Kadaptation[112]	首先通过衡量局部内在维度来选择需要自适应的子模块,然后将模块参数映射到一个子空间并通过计算 Kronecker 积近似模块参数的更新				●		ELEVATER:68.9	0.08%

（续　表）

方法	概述	修改结构	更新参数	修改输入	注入参数	增加模块	部分验证数据集与指标	参数占比
SPT[113]	根据任务计算最关键的参数进行更新. 对于超出可更新参数限制的部分使用 Adapter 等现成高效迁移方法进行更新		●				FGVC：90.1 VTAB-1k：76.4	0.44%
Convpass[114]	提出构建基于卷积层的自适应模块对 ViT 中序列化的图像还原后提取局部特征					●	VTAB-1k：76.5	0.38%
VQT[115]	基于多头自适应机制对中间特征进行组合和降维，并允许下游任务头直接使用中间特征进行训练				●		VTAB-1k：68.8	3.50%
Head2Toe[116]	从所有中间变量中选择最优的用于训练下游任务头						VTAB-1k：66.0	1.00%
EXPRES[117]	在学习输入层面提示的基础上再学习残差提示实现 ViT 的高效自适应				●		VTAB-1k：72.9	<1.00%

注：验证数据集与指标和表 7 中信息对应，参数占比指的是高效方法所调参数相较于所有参数的比例.

4.1　基于模型结构调整的方法

视觉模型中的卷积层具有局部感受野，能对二维图像的区域特征进行较好的提取；ViT 的注意力模块则提供了较强的全局特征提取能力. 现有视觉模型结构多样、预训练数据也有所不同，其中蕴含了十分丰富多样的知识. 一些方法发现只要合理地选择或组合合适的预训练模型，即使无需训练也可以在目标任务上有较好的表现. 本文将这类方法总结为基于模型结构修改的方法.

图 4 展示了两种较流行的模型结构调整方案. 下方的方案最大的特点在于需要修改模型构成，如 DeRy[31] 提出一种两阶段的高效知识迁移框架，首先通过近似解决一个集合覆盖问题将异构预训练模型划分为许多等价模块，再根据目标任务和计算开销要求解决一个整数规划问题实现模块重组. 为了让不同尺寸、不同结构的模块能衔接起来，DeRy 在重组时在模块之间增加卷积连接层，也可以认为是增加了自适应模块. DeRy 通过选取需要的模块实现迁移，而 FDA-ES[106] 通过移除多余的模块降低迁移开销. FDA-ES 发现在推理时，简单的目标样本和困难的目标样本实际需要经历的模型层数不同. 对于简单的样本，经历模型前中段即可作出正确判断，因此可以根据样本动态确定最简单的推理路径以减小推理开销. 虽然模型结构没有直接改变，但对于不同样本来说实际经历的模型却是不同的，类似的思想还体现在文献[118-121]等中. Zoo-Tuning[107] 研究的是同构但训练于不同数据的模型，并设计了一种自适应聚合层将不同的源模型参数根据输入数据的特点进行动态组合，得到目标模型的参数. 令不同预训练模型的参数为 $\theta_{f_{s_1}}, \theta_{f_{s_2}}, \cdots, \theta_{f_{s_n}}$，则上述的高效迁移方法原理可以抽象为：$F(\theta_{f_{s_1}}, \theta_{f_{s_2}}, \cdots, \theta_{f_{s_n}}) = \{\theta_{v_{s_1}}, \theta_{v_{s_2}}, \cdots, \theta_{v_{s_1}}\}$，其中 $\theta_{v_{s_i}} \in \theta_{f_{s_i}}$ 是第 i 个预训练模型的一部分参数. 此类方法虽然都对模型结构进行自适应修改，但具体的实现方法相去甚远，且实际应用起来相对较困难、表现难以保证.

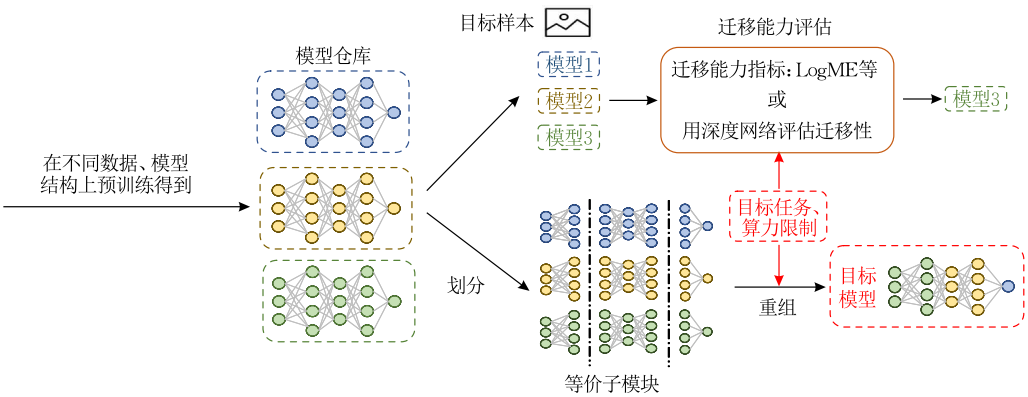

图 4　基于视觉模型结构调整的方法示意图

图 4 上方的方案提供了一种更高维度的视角来审视不同的模型，这种方法的核心思想是研究如何根据目标任务直接选择一种预训练模型或结构，而无需重新搭建一种新的结构. 这类方法不改变预训练模型的原始结构，使其表现更有保证，且选择的开销一般小于搭建的开销. 这类方法有两种实现方式，基于迁移能力指标和基于决策网络的方法. 基于网络的方法另外训练一个决策模块

针对目标任务和数据选择最佳的预训练模型，如Hub-Pathway[32]基于强化学习的思想设计了一个动态路径网络，在数据点尺度上动态选择表现最优的预训练模型进行推理，该方法包括了对预训练模型的进一步调优.不同于该方法在数据点尺度的决策，更多方法研究的是任务尺度的模型迁移能力评估指标.表6总结了不同的模型迁移性评估指标，涵盖了基于极大似然、极大证据等理论的主流方法.这些指标能基于少量目标域标签数据对海量预训练模型进行筛选，可以看作是进行模型迁移前的准备工作，选择一个表现最好的预训练模型能让迁移任务事半功倍.这两种方法各有优劣：基于迁移能力指标的方法通常计算效率更高，适用于快速筛选大量预训练模型；而基于决策网络的方法虽然可能需要更多计算资源，但能够提供更精细和动态的选择策略.

表6 常用的模型迁移性评估指标

方法	方法简介与计算表达式
NCE[122]	定义标签序列Y和Z之间的条件熵为$H(Y\|Z)=-\sum_{y\in Y}\sum_{z\in Z}\hat{P}(y,z)\log\frac{\hat{P}(y,z)}{\hat{P}(z)}$，其中$\hat{P}(y,z)=\frac{1}{n}\|i:y_i=y \text{ and } z_i=z\|$是$Y$和$Z$的联合分布，$\hat{P}(z)=\sum_{y\in Y}\hat{P}(y,z)$是$Z$的边缘分布.任务$Y$到$Z$的迁移性和$H(Y\|Z)$呈负相关，因此可以用负条件熵(NCE)衡量模型在任务上的迁移性.
LEEP[123]	模型θ在目标任务D上的迁移性表示为$T(\theta,D)=\frac{1}{n}\sum_{i=1}^{n}r\log\left(\sum_{z\in Z}\hat{P}(y_i\|z)\theta(\boldsymbol{x}_i)_z\right)$，其中$\hat{P}(y,z)$的表达式和NCE相同，$Z$是直接用源模型$\theta$在目标任务上推理得到的无意义伪标签(dummy label)，Y是目标任务的标签.$T(\theta,D)$是一个负数，其值越大代表模型θ在任务D上迁移性越好.另外，LEEP与NCE还有关系$T(\theta,D)\geqslant NCE(Y\|Z)+\frac{1}{n}\sum_{i=1}^{n}r\log\theta(\boldsymbol{x}_i)_z$
LogME[124]	为了改进NCE和LEEP只能评估分类任务模型的缺陷，提出极大证据对数(Logarithm of Maximum Evidence，LogME).LogME十分高效，能实现3000倍的提速且只需要1%内存.其计算流程如下：对目标数据集的每个类别k迭代计算$\alpha\leftarrow\frac{\gamma}{\boldsymbol{m}^{\mathrm{T}}\boldsymbol{m}}$，$\beta\leftarrow\frac{n-\gamma}{\|\boldsymbol{F}\boldsymbol{m}-y\|_2^2}$，其中$\boldsymbol{m}=\beta(\boldsymbol{V}(\Lambda^{-1}(\boldsymbol{V}^{\mathrm{T}}(\boldsymbol{F}^{\mathrm{T}}y))))$，$A=\alpha I+\beta\boldsymbol{F}^{\mathrm{T}}\boldsymbol{F}$.$\boldsymbol{F}$是特征构成的矩阵，对$\boldsymbol{F}$进行奇异值分解得到$\boldsymbol{V}$，$\Lambda$：$\boldsymbol{F}^{\mathrm{T}}\boldsymbol{F}=\boldsymbol{V}\,\mathrm{diag}\{\sigma\}\,\Lambda^{-1}$.最后根据计算出的$\alpha$，$\beta$计算本类别$k$下的$\mathcal{L}_{\mathcal{K}}=\frac{1}{n}\mathcal{L}(\alpha,\beta)=\frac{1}{n}\left(\frac{n}{2}\log\beta+\frac{D}{2}\log\alpha-\frac{n}{2}\log 2\pi-\frac{\beta}{2}\|\boldsymbol{F}\boldsymbol{m}-y\|_2^2-\frac{\alpha}{2}\boldsymbol{m}^{\mathrm{T}}\boldsymbol{m}-\frac{1}{2}\log|A|\right)$，最后对所有类别取平均得到LogME$=\frac{1}{K}\sum_{k=1}^{K}cL_k$.LogME越大则迁移性越好.
B-Tuning[125]	每个预训练模型会产生一个后验分布$p(y'_k\|f_k,F_k,y)\sim\mathcal{N}(\boldsymbol{f}_k^{\mathrm{T}}m_k,\boldsymbol{f}_k^{\mathrm{T}}A_k^{-1}\boldsymbol{f}_k+\beta_k^{*-1})$，B-Tuning提出基于每个模型的LogME值$\{\mathcal{L}_{\mathcal{K}}\}_{k=1}^{K}$对它们的结果进行加权求和：$\bar{y}'=\sum_{k=1}^{K}\pi_k y'_k$，其中$\pi_k=\frac{\exp(\mathcal{L}_h/t)}{\sum_{j=1}^{K}\exp(\mathcal{L}_j/t)}$.
GBC[126]	首先基于高斯分布对每个类别c在源域空间里的分布进行建模：$\mu_c=\frac{1}{N_c}\Sigma_i[y_i=c]f_s(x_i)$，$\Sigma_c=\frac{1}{N_c-1}\Sigma_i[y_i=c](f_s(x_i)-\mu_c)(f_s(x_i)-\mu_c)^{\mathrm{T}}$，其中$N_c=\Sigma_i[y_i=c]$代表类别$c$的样本数.然后计算巴氏距离(Bhattacharyya distance)$D_B(c_i,c_j)=\frac{1}{8}(\mu_{c_i}-\mu_{c_j})^{\mathrm{T}}\Sigma^{-1}(\mu_{c_i}-\mu_{c_j})+\frac{1}{2}\ln\left(\frac{|\Sigma|}{\sqrt{|\Sigma_{c_i}||\Sigma_{c_j}|}}\right)$，其中$\Sigma=\frac{1}{2}(\Sigma_{c_i}+\Sigma_{c_j})$.巴氏参数可以确定为$BC(\cdot,\cdot)=\exp-D_B(\cdot,\cdot)$.最后，相似度指标表达式为$GBC_{s\to t}=-\sum_{i,j}[i\neq j]BC(c_i,c_j)$.GBC越大，迁移性越低.
OTCE[127]	OTCE解决的是跨领域跨任务表征的迁移性问题.定义领域偏差$W_D=\sum_{i,j=1}^{m,n}\|\|\theta(x_s^i)-\theta(x_t^j)\|\|_2^2\pi_{ij}^*$，其中$m,n$分别为源域、目标域的样本数量，$\boldsymbol{\pi}^*$是最优传输问题中的最优解耦合矩阵.基于条件熵[122]定义任务偏差$W_T=\sum_{y_t\in Y_t}\sum_{y_s\in Y_s}\hat{P}(y_s,y_t)\log\frac{\hat{P}(y_s,y_t)}{\hat{P}(y_s)}$.最后得到OTCE指标$OTCE=\lambda_1\hat{W}_D+\lambda_2\hat{W}_T+b$，其中$\lambda_1,\lambda_2$是超参数，$b$是偏置项.

4.2 基于参数微调的方法

作为最简单也最有效的知识迁移手段，微调仍然是工业界中迁移模型的主要技术.一些工作着手于改进微调，在保证其效果的同时大幅减少需要调整的参数量，可以表示为$F(\theta_{f_s})=\{\theta_\Delta+\eta\theta'_\Delta,\theta_{f_s}\setminus\theta_\Delta\}$，其中$\theta_\Delta$是需要更新的部分参数，$\eta$代表学习率.文献[32,106-107]中的主要迁移手段是对模型结构的选择和对模型参数的聚合，但仍将参数微调作为提升效果的辅助手段.TinyTL[110]指出现有的大多数高效迁移方法旨在降低需要训练的参数量，但这些方法大多仍需要在训练时计算整个模型的梯度，因此很难降低内存开销.该方法发现内存开销主要与激活层而不是参数量有关，因此提出了一种新颖的偏置层模块，允许在冻结大多数参数(权重矩阵)的情况下只通过微调偏置项实现迁移.SPT[113]首先通过观察微调时的目标函数损失情况获得参数敏感性，

再根据一个给定的可调试参数预算 τ 找到前 τ 最敏感的参数；然后对敏感参数集合中较密集的参数矩阵应用现有的结构化微调方法，如 LoRA[44] 或 Adapter[51] 等，并对其余参数采用非结构化微调的方式进行更新. SPT 提出的方法同时采用了结构化和非结构化的微调方式，并观察到结构化微调方式对分布差异较大的目标任务更有效.

可以发现，虽然已经出现了许多无需调整参数的高效迁移方法，但对模型的原始预训练参数进行微调仍属于迁移学习中效果最好的方法之一，是迁移学习中不可或缺的组成部分. 随着预训练模型规模的不断扩大，传统的全参数微调方法面临着计算资源消耗大、易过拟合等挑战，因此最新的微调技术致力于寻找“性价比”最高的参数进行小规模的调整. 如表 5 所示，它们能在保持参数敏捷性的同时取得较其他方法相对更好的性能.

4.3　基于自适应参数注入的方法

由于 ViT 在结构和原理上都与原始的 Transformer 极为类似，许多用于 Transformer 的参数注入迁移方法也可以应用在 ViT 上. 如 VPT[108] 提出了视觉提示的方法，与自然语言处理中的 Prompt-Tuning[43] 类似，该方法在图片被序列化后形成的编码序列和自注意力模块的编码序列中注入可学习的自适应参数，可以参考图 3 和表 2. EXPRES[117] 进一步改进上述的视觉提示框架，除了在图片的输入编码层增加了提示向量，还在这些提示向量在 ViT 模块中传播时增加残差提示向量：$\hat{Q}=Q+\Delta_Q$，其中 Q 是原始的中间参数（未增加残差提示时），Δ_Q 是增加的残差提示，$\hat{Q}$ 是进行残差计算后送入下一层的参数. Kadaptation[112] 提出在 ViT 上通过低秩矩阵分解的方式近似预训练模型的更新，与自然语言处理中的 Compacter[38] 方法类似. VQT[115] 与视觉提示 VPT[108] 类似，但不同点在于其只在注意力模块中的查询（query）向量中注入自适应参数，而视觉提示在查询、键、值中都注入了参数；另外 VQT 将参数注入后多出来的向量作为总结中间向量加以利用，而 VPT 未使用这些向量.

4.4　基于自适应模块的方法

类似于自适应参数注入，许多方法也受自然语言处理中的 Adapter[51] 启发提出了用于视觉模型和任务的自适应模块. AdaptFormer[111] 就通过在 ViT 上增加 Adapter 模块实现了高效迁移，其模块也由一个降采样、升采样全连接层和非线性激活层构成，如式（4）所描述. ConvAdapter[109] 进一步在构建 Adapter 时考虑视觉任务的特点，如卷积层能更好捕捉二维图像中的局部特征，从而将式（4）中的升采样和降采样全连接层替换为升采样和降采样卷积层，同时仍然保持输入输出变量的形状一致，且可以适用于 ViT 和基于卷积层的 ResNet 等模型. 但添加的卷积层较全连接层复杂度上升，导致该方法需要调整的参数量略高于其他方法. Convpass[114] 提出用于 ViT 的卷积自适应模块，将序列化的图形编码在自适应模块中逆序列化还原回二维分布，后利用卷积提取并保留图像中的空间相关信息. 前文所述的 DeRy[31] 虽然也设计有额外的连接模块，但其主要作用是统一异构模块之间的输入输出维度而不是进行知识迁移.

4.5　其他方法

Head2Toe[116] 提出了一种特殊的自适应方法，针对保留骨干网络、重新训练任务头的传统迁移方法，Head2Toe 提出从预训练骨干网络的所有中间层、中间变量中获取特征表示用于训练任务头，而不是仅采用传统方法中骨干网络最后一层的输出作为特征. 其基于 Group lasso[128] 选择中间特征，因此也无需额外训练. VDP[105] 采用的也是视觉提示的方式，但不同于 VPT[108]，VDP 的提示信息是直接叠加在原始输入图像上的小像素块，以类似于对抗攻击的方式难以察觉地修改原始输入. 这种方法不受限于 ViT 的序列化结构，因而其思想能广泛地应用于各种模型上. 如文献[129]就通过在输入图像上叠加像素框作为视觉提示信息以实现大模型的自适应和泛化.

总的来说，虽然计算机视觉和自然语言处理的任务差异巨大，在高效迁移领域它们的方法却是十分相似的，特别是在基于参数注入和自适应模块两类方法上. 事实上，许多基于 ViT 的高效迁移方法是借鉴于自然语言处理领域而来，如文献[112]的实验环节中便将原本提出用于自然语言处理任务的 BitFit[46]、Adapters[51]、LoRA[44]、Compacter[38]、AdapterDrop[130] 方法实现在计算机视觉任务上作为基线方法. 这两类任务的相似性源于模型的相似性，自然语言处理领域的 Transformer 天然地可以处理序列化的语言数据，而视觉领域的 ViT 将图像划分为小块后重排为序列，便可以进行类似的处理. 二者的最大不同在于序列的自身含义. 自然语言的语句前后天然地有语义上的联系，而图像的序列化中更多包含的是原始图片的不同位置信息. 因此，研究者在考虑高效知识迁移问题时或许可以暂时抛开

具体任务的约束,在更高维度思考高效迁移问题,并积极了解借鉴其他应用场景的方法,或许可以取得更多收获.

4.6 常用数据集与实验结果

表7总结了计算机视觉自适应领域常用的数据集,主要用于分类任务、目标检测任务和语义分隔任务.其中较新、使用较多的当属对标GLUE的VTAB-1k[131](Visual Task Adaptation Benchmark).VTAB数据集包含19个子数据集,覆盖了:(1)自然类别,包含用标准摄像机拍摄的自然物品、细分物种和抽象概念等,包括Caltech101[132]、CIFAR100[133]、DTD[134]、Flowers102[135]、Pets[136]、Sun397[137]和SVHN[138];(2)专用领域图像,包含用专用设备拍摄的图片,如医学图像、遥感图像等,其中医学图像数据集包括Patch Camelyon[139]和Diabetic Retinopathy[140],遥感图像数据集包括Resisc45[141]和EuroSAT[142];(3)结构化的目标理解任务,包含估计立体环境中的距离、计数、定位等任务,包括Clevr[143]、dSprites[144]、SmallNORB[145]、DMLab[146]、KITTI[147].

表7 计算机视觉领域常用数据集

数据集	描述	指标
Office31[148]	跨领域图像分类数据集,包含3种领域,每种领域包含31个来自办公场景下的类别,共4652个样本	准确率
分类任务基准数据集[107]	包含3类基准:一般分类任务,包括CIFAR100[133]等;细粒度分类任务,包括FGVC aircraft[149]等;专用分类任务,包括EuroSAT[142]等.共7个数据集	准确率平均值
ImageCLEF[150]	跨领域图像分类数据集,包含3种领域,每种领域包含12个类别	准确率
FGVC	由5个数据集(CUB-200-2011[151]、NABirds[152]、Oxford Flowers[135]、Stanford Dogs[153]、Stanford Cars[154])组成的细粒度图像分类基准数据集	准确率平均值
VTAB-1k[131]	包含19个图像分类数据集,分为3大类:自然界图片,专门领域图片(医学等),结构理解图片	准确率平均值
SVHN[138]	图像中的数字目标检测任务,包含60万张街道门牌号图像	准确率
ELEVATER [155]	包含20个图像分类数据集	准确率
ImageNet1k[7]	经典图像分类数据集,包含1000个类别和一百多万张图片	准确率
PASCAL-5^i[156-157]	小样本语义分隔数据集,有20个目标类,被分为4个子集:$\{5^i, i\in\{0,1,2,3\}\}$	子集上的精度
Cityscapes[158]	语义分割经典数据集,包含共25 000张来自50个不同城市的街景数据,其中5000张有像素级别的标注,20 000张有粗粒度标注	*IoU*
Microsoft COCO[159]	大型图像数据集,综合包含目标检测、语义分割、关键点识别等任务.其中包含超33万张图像,其中有超过150万个实例,超过80个目标类别等.该数据集是推动计算机视觉领域发展的重要数据集	准确率、*AP*、*IoU*等

具体到细分分类任务,FGVC数据集提供了5个子细分数据集,包含55至200个类别以及每个子数据集中的数千张训练、测试样本.ImageNet1k[7]是经典图像分类数据集ImageNet的子集,包含了其中精心挑选的1000类.作为开启计算机视觉领域深度学习时代的第一个超大图像数据集,ImageNet时至今日仍是衡量模型结构、深度学习方法性能的必选方案.SVHN[138]包含超过600 000张来自于谷歌街景中的房屋号码,对图像预处理和格式要求较低.Office31[148]是领域自适应经典数据集,包含3个不同领域.方法需要在其中任意两个领域间进行知识迁移,共6组迁移任务,且不能使用目标域标签(对应于第2节中定义的直推式迁移学习).

为了更直观展示比较不同方法的性能,本文收集了计算机视觉领域代表性方法在应用最广泛的视觉基准数据集VTAB-1k上的实验结果.表8展示了不同方法在VTAB-1k的三类任务:自然类别(Natural)、专用领域(Specialized)和结构化(Structured)上的平均实验结果.骨干网络中的ViT-B/16在ImageNet-21k上预训练得到,ViT-B/16(1k)在ImageNet-1k上预训练得到.在所有比较的方法中,基于自适应模块(Adapter)的方法表现出较高的平均性能,但也相应地需要调整更多参数.在实际应用时,需要根据具体需求权衡性能和参数高效性.

表8 计算机视觉领域部分代表性方法在VTAB-1k基准数据集上的实验结果

方法	会议/期刊	骨干模型	参数占比/%	Natural	Specialized	Structured
VPT[108]	ECCV	ViT-B/16	4.9	78.5	82.4	55.0
Conv-Adapter[109]	CVPRW	Swin-base	5.7	80.0	85.8	62.6
SPT[113]	ICCV	ViT-B/16	0.4	82.0	85.8	61.4
AdapterFormer[111]	NeurIPS	ViT-B/16	0.4	80.6	85.4	58.5
Convpass[114]	Arxiv	ViT-B/16	0.4	81.6	85.3	62.7
VQT[115]	CVPR	ViT-B/16(1k)	3.5	72.7	84.5	49.3
Head2Toe[116]	ICML	ViT-B/16(1k)	1.0	68.9	82.9	46.3
EXPRES[117]	CVPR	ViT-B/16	<1.0	79.7	84.0	55.0

5　跨模态模型上的高效迁移方法

上文所述的计算机视觉和自然语言处理在过去数十年间的发展一直是相对独立的. 然而近年来，随着算力资源的日渐丰富以及高质量多模态数据集的不断完善，研究者们开始设计多模态大模型，且已经在多种应用中取得了十分亮眼的成果，如基于语言描述生成对应图片的 DALLE[160]，甚至是能根据文字生成高质量视频片段的 Sora[161]. 这些模型无一不需要高额算力支撑，而高效迁移学习能以较低的代价将这些性能优越的大模型迁移到具体的应用目标上，如专用领域的图像生成等. 本文研究的多模态大模型为视觉-语义模型(Vision-Language Models，VLM)，它可以同时处理图像模态和文字模态的输入，并将它们映射为同一个高维空间里的特征向量后根据跨模态特征的相似度进行分类等任务.

图 5 左以 CLIP[162] 为例展示了 VLM 的一般训练流程. CLIP 由一个图像编码器和文本编码器构成，图像编码器一般由常见的视觉深度模型(ResNet、ViT 等)构成，以训练数据中的图片为输入，将其编码为高维特征向量；文本编码器由 Transformer 构成，以训练数据中某张图片的文字描述为输入，如“一张火车的图片”，并将其编码为特征向量. 随后采用对比学习范式提升成对的图片-文字描述特征之间的相似度，并压低不成对的特征之间的相似度，如式(5)～(7)所示：

$$\mathcal{L}_{\mathcal{I}}=-\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp(z_i^I\cdot z_i^T/\tau)}{\sum_{j=1}^{B}\exp(z_i^I\cdot z_i^T/\tau)} \tag{5}$$

$$\mathcal{L}_{\mathcal{T}}=-\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp(z_i^T\cdot z_i^I/\tau)}{\sum_{j=1}^{B}\exp(z_i^T\cdot z_i^I/\tau)} \tag{6}$$

$$L=\frac{1}{2}(\mathcal{L}_{\mathcal{I}}+\mathcal{L}_{\mathcal{T}}) \tag{7}$$

其中 z_i^I，z_i^T 分别是图像、文字特征，B 是一个训练批次的大小，τ 是温度参数.

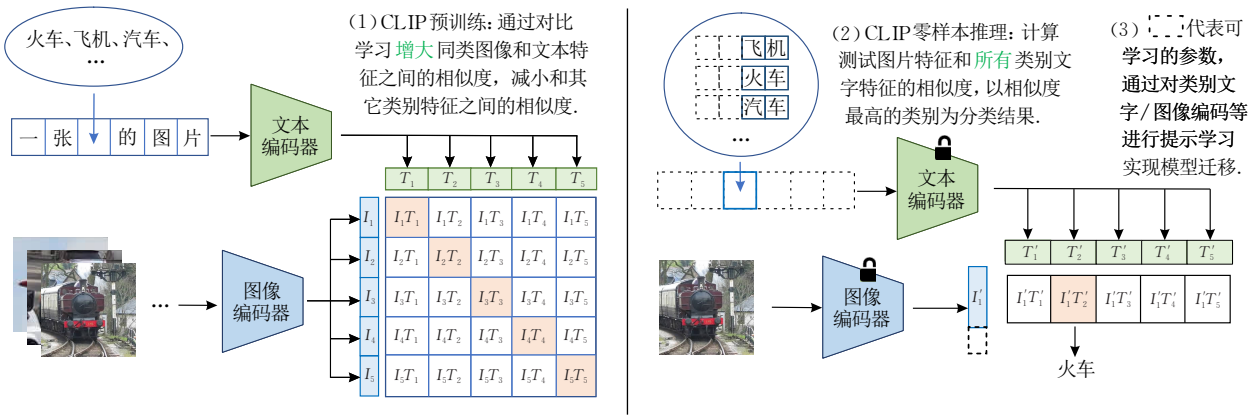

图 5　多模态模型训练、推理原理及其常见的高效迁移方式

图 5 右展示了应用训练好的 CLIP 完成零样本图像分类的方法. 将待分类的图片经过图像编码器获得图像特征 $\boldsymbol{z}^I$，将所有可能类别对应的朴素提示描述(如“一张[类别]的图片”，假设有 C 个类别)经过文本编码器得到 C 条文字特征 $\{z_c^T\}_{c=1}^{C}$. 通过对比图像特征和 C 条文字特征的相似度，取相似度最高的一类为分类结果，如式(8)～(9)所示：

$$p_c=\frac{\exp(\boldsymbol{z}^I\cdot z_c^T/\tau)}{\sum_{c=1}^{C}\exp(\boldsymbol{z}^I\cdot z_c^T/\tau)} \tag{8}$$

$$\hat{y}=\arg\max_c p_c \tag{9}$$

其中 p_c 通过相似度归一化后计算得到的图像属于类别 c 的概率.

此类 VLM 往往在巨量多模态数据上经历了十分充分的预训练，如 CLIP[162] 在 4 亿对文字-图像数据上训练而成，ALIGN[163] 则使用了超过 10 亿对多模态数据. 这些数据几乎涵盖了对常见场景、物体的描述，使这些大型预训练 VLM 获得十分强大的零样本推理能力，也就是说无需任何调试的情况下，VLM 就能基于图 5 的方式完成高精度的分类任务，如 CLIP 在 ImageNet 上取得了 62.5%的零样本精确度. 这也使预训练 VLM 成为实施高效迁移学习的绝佳工具，只需极少的目标样本(如小样本迁移学习中常用的每类 1、2、4、8、16 个样本)和极低的调试开销，就能让 VLM 本就先进的零样本推理能力更上一层楼. 但如此巨量的数据使这些模型十分难以训练、调试，在数据不足的情况下进行的调试甚至会损伤模型原有能力. 因此高效迁移是将 VLM 以极低的训练、数据开销泛化至其他领域的必需策略，也吸引了许多学者开展探索.

图 5 右侧虚线框给出了部分用于 VLM 的高效迁移方法. VLM 同时具有语义和图像分支,因此前文所述的用于自然语言处理和计算机视觉的诸多方法都可用于其中,例如对经过编码的图像序列注入自适应参数. 一种在 VLM 上特有的方法是直接对用于推理的朴素提示进行提示学习,如图 5 右上,不仅可以将提示修改"一张[类别]的图片",还可以把[类别]扩充为更具体的类别描述. 表 9 总结了多模态领域有代表性的高效迁移方法,其中的黑点代表该方法所属的高效迁移技术类别;接下来基于该表在 5.1 节介绍修改输入(提示)的方法;在 5.2 节介绍注入参数的方法;在 5.3 节介绍增加自适应模块的方法;在 5.4 节介绍其他方法和一些多模态领域的综述文献;在 5.5 节介绍多模态领域的数据集.

表 9 多模态模型领域的部分高效模型迁移方法汇总

方法	概述	修改结构	更新参数	修改输入	注入参数	增加模块	部分验证数据集与指标	所调参数
CoOp[34]	基于提示学习实现 VLM 上的迁移学习,在语义分支引入可学习的上下文向量对提示进行微调			●	●		图像分类基准数据集:82.7	0.02 M
CoCoOp[35]	改进 CoOp 的提示方法,为每个样本生成不同的提示信息			●	●	●	图像分类基准数据集:80.5	—
DeFo[164]	为了克服 VLM 的概念敏感性和表达敏感性,提出学习完整的提示并增加额外线性层			●	●	●	图像分类基准数据集:79.9	—
CLIP-Adapter[165]	在语义和视觉分支增加自适应模块和残差连接以修改 VLM 的中间变量					●	图像分类基准数据集:74.6	0.52 M
Tip-Adapter[166]	基于缓存模型提出一种完全无需训练的 VLM 高效迁移方法					●	图像分类基准数据集:70.0	0%
SuS-X[167]	基于 Tip-Adapter 进行改进,提出在多模态空间计算输入图像和标签的相似度					●	图像分类基准数据集:71.5	0%
APE[168]	对基于缓存模型的方法进行改进,筛选出最有价值的特征通道进行相似度计算					●	图像分类基准数据集:73.4	0%
VL-Adapter[169]	提出在 VLM 中增加自适应模块实现高效迁移的框架					●	文字-图像基准数据集:76.7 文字-视频基准数据集:87.4	4.18%
IP-Adapter[170]	提出用于文字-图像生成预训练扩散模型的图像提示自适应模块					●	COCO2017:0.83	2.50%
VLN-PETL[171]	首次提出用于视觉-语义导航任务的高效迁移框架				●	●	视觉-语义导航基准数据集:31.8	2.82%
LLaMA-Adapter V2[172]	针对多模态指令遵循模型(instruction-following model)提出参数高效的模型框架		●		●		COCO caption:122.2	0.50%
VL-PET[173]	提出一种粒度控制机制来防止模块化高效方法带来的编码器、解码器之间的能力偏差				●	●	文字-图像基准数据集:79.2 文字-视频基准数据集:88.6	4.16%
TinyCLIP[29]	提出一种应用于 VLM 的高效模型蒸馏方法	●					ImageNet:62.7	50%
DAPrompt[174]	提出将 VLM 用于领域自适应任务,通过设计可学习的领域相关、领域无关提示向量实现模型迁移			●	●		领域自适应基准数据集:80.7	—
UniMoS[175]	基于模态分布偏差理论将 VLM 的多模态特征解构为视觉、语义特征分别处理,最后进行跨模态融合					●	领域自适应基准数据集:83.0	2%
DAMP[176]	通过视觉和文本表征之间的关系学习领域无关的多模态提示,并使用语义一致性和对比学习来优化提示			●	●		领域自适应基准数据集:83.3	11.90%
PADCLIP[177]	通过控制学习率防止微调 VLM 时的遗忘问题,应用于领域自适应任务		●				领域自适应基准数据集:82.6	—
MaPLe[178]	同时在 VLM 的视觉、语义分支上注入可学习参数进行提示学习				●		图像分类基准数据集:82.3	2.85%

注:验证基准数据集与指标和表 10 中信息对应,参数指的是高效方法所调参数的具体值(以 M 为单位)或相较于模型所有参数的比例(以%表示).

5.1 基于提示微调的方法

VLM 最具特色也最灵活的部分当属其文字提示. CLIP 中提供的朴素提示仅仅是十分简洁的"一张[类别]的图片",但其预训练数据却包含许多更为丰富的描述,因此它实际上具备更强大的语言感知能力. 例如方法 CoOp[34] 中就指出,对于 Flowers102 数据集[135],将朴素提示修改为"一张[类别],一种花的图片"就可以将分类准确率从 60.86% 提升至 66.14%. 然而实际应用中,对每个目标任务、数据集都尝试人工设计不同的提示是不现实的,因此许多方法通过在文字提示上注入可学习参数实现自动的提示微调.

CoOp[34] 提出仅保留朴素提示中的[类别]部分,而对朴素提示的其余部分都进行微调更新,即

“$[V_1][V_2]\cdots[V_k][\text{类别}][V_{k+1}]\cdots[V_N]$”，其中$[V_i]$均为可学习的参数，$N$ 为提示句子的长度，是可以调整的超参数. 方法 CoOp 对每个类别都学到一条这样的朴素提示，再根据式(8)～(9)进行推理. CoCoOp[35] 发现 CoOp 学习的提示信息会过拟合训练集，从而在未见过的类别上泛化性能不佳，因此提出将每个类别学习一条提示的方法改进为针对每个样本学习不一样的提示，即“$[v_1(x)][v_2(x)]\cdots[v_k(x)][\text{类别}][v_{k+1}(x)]\cdots[v_N(x)]$”，其中 $v_i(x)$ 代表一个轻量的生成网络，根据输入 x 为每个位置生成不同的提示. DeFo[164] 发现朴素提示中固定的[类别]也是有局限性的，如“火车”还有“列车、动车”等近义词，而选取一个固定的类别词往往不能在目标任务上有最好的表现. 因此 DeFo 提出学习与类别完全无关的提示：“$[V_1][V_2]\cdots[V_N]$”，再增加一层线性探测(linear probing)层进行分类. DAPrompt[174] 将 VLM 用于领域自适应任务[106]. 由于源域和目标域的风格不同，DAPrompt 设计了一种包含领域信息的提示：“$[V_1][V_2]\cdots[V_{M1}]+[D_1][D_2]\cdots[D_{M2}]+[\text{类别}]$”，其中 $[V_i]$ 是领域无关的提示，$[D_i]$ 是领域相关的提示，如“一张猫的漫画风格图片”，其中“一张图片”是与领域无关，可以应用于任何样本的提示，“猫”是具体的类别名称，“漫画风格”则是领域相关的信息，刻画了该领域的特点. DAMP[176] 将学习提示的范围从文本扩大到视觉模态，通过视觉和文本表征之间的相互促进学习领域不变的提示信息，从而将丰富的预训练知识和源域知识迁移到目标域. 由于引入了跨模态的信息，DAMP 比 DAPrompt 获得了 2.6%的性能提升，凸显了 VLM 应用中跨模态提示的重要性.

这类基于提示微调的方式无需对 VLM 中的预训练参数有任何修改，只需根据梯度更新加在提示上的可学习向量，因此具有十分优越的参数高效性，但其梯度前后传播的过程仍会产生不小的内存、算力和时间开销. 基于第 2 节中的定义，可以将此类方法总结为 $F(\theta_{f_s})=\{\theta_{f_s},\theta_P\}$，其中 θ_P 是学到的提示参数. 这类方法虽然实现时与自然语言处理中对句子编码的提示学习类似，但不同在于其可解释性更强、修改的位置和方法更贴近自然语言，如 CoOp[34] 就将不同数据集上学到的提示参数翻译回英文进行分析，而这在表 2 中介绍的方法上是难以实现的.

5.2　基于自适应参数注入的方法

现有主流 VLM 多是由 Transformer 和 ViT 构成，因此前文介绍的自适应参数注入思想也十分适用. 5.1 节介绍的提示微调方法实现时也需要向文字编码向量中增加可学习向量，如图 5 所示，因此也具备参数注入的特点. 除此之外也有许多方法直接向模块之间的中间变量进行参数注入，接下来进行介绍.

VLN-PETL[171] 首次提出了一种用于语言-视觉导航(Vision-and-Language Navigation，VLN)任务的高效迁移方法. 该方法指出如 BitFit[46] 和 Prompt-Tuning[43] 的高效迁移方法在 VLN 任务上表现并不好，因此采用了 LoRA[44] 和 Adapter[51] 作为其实现高效迁移的基石方法，分别对应参数注入和自适应模块法，并基于不同的实现方式构造了语义编码 Adapter，历史信息交互增强 Adapter，跨模态交互 Adapter 等变种. LLaMA-Adapter V2[172] 提出了一种参数高效的视觉指令模型，其技术细节主要包括：(1) 微调偏置层，在仅引入 0.04%参数量的情况下提升模型表现；(2) 不相交参数联合训练，使用不同数据训练不同部分的参数；(3) 视觉知识早期融合，将视觉编码和自适应提示注入到不同的 Transformer 层；(4) 专家集成，利用标题、搜索引擎和 OCR 等专家系统来增强模型的视觉指令遵循能力. VL-PET[173] 认为许多基于 Adapter 的高效迁移方法忽视了编码器和解码器中功能不同的问题，因此提出控制迁移时参数更新的粒度：$\theta_{f_s}=\boldsymbol{G}\odot(\theta_{f_s}+\theta_\Delta)$，其中 θ_Δ 是 Adapter 更新的参数，$\boldsymbol{G}$ 是粒度控制矩阵，$\odot$ 代表逐元素点乘. MaPLe[178] 将提示学习拓展到图像、文字两种模态中. 该方法采用向 Transformer 中间向量拼接参数的方式在 VLM 的视觉、语义分支都进行提示，并使用语义提示控制视觉提示的生成，以促进两种模态之间的信息交互.

5.3　基于自适应模块的方法

VLM 上的自适应模块法主要可以分为 2 类. 一类是传统的 Adapter 方法，即在 VLM 的视觉、语义分支上增加不同设计的自适应模块实现迁移；另一类是基于缓存模型的方法.

(1) 基于 Adapter 的方法. CLIP-Adapter[165] 提出在视觉和语义分支分别添加一个 Adapter，以残差的形式修改中间变量：$\boldsymbol{z}^{I^*}=\alpha\boldsymbol{z}^I+(1-\alpha)A^I(\boldsymbol{z}^I)$，$\boldsymbol{z}^{T^*}=\beta\boldsymbol{z}^T+(1-\beta)A^T(\boldsymbol{z}^T)$，其中 $\boldsymbol{z}^I$、$\boldsymbol{z}^T$ 分别是原始的图像、文字向量，α、β 是超参数，$A^I(\cdot)$、$A^T(\cdot)$ 分别是语义分支和视觉分支增加的 Adapter，$\boldsymbol{z}^{I^*}$，$\boldsymbol{z}^{T^*}$ 是经过迁移的特征向量，最后再使用式(8)～(9)进行分类. VL-Adapter[169] 将 VLM 的语义分支看作是一个语言模型，并设计了统一的框架在语义分支上测试了提出于自然语言处理领域的高效迁移方法：Adapter[51]、

Compacter[38]和 HyperFormer[53]. IP-Adapter[170]针对用于文字生成图像的扩散模型提出了图片提示的方法,通过一个图像提示 Adapter 将图像信息注入到参数固定的扩散模型骨干网络(Unet)中.

(2)基于缓存模型的方法. Tip-Adapter[166]基于 CLIP-Adapter 扩展出了一种完全不需要训练的小样本迁移方法. 假设目标任务上有 N 个类别,每个类别有 K 个样本,首先将每个训练样本经过视觉编码器得到的视觉向量拼接为缓存模型的键矩阵 $\boldsymbol{F}_{\text{train}} \in R^{NK \times d}$,其中 d 是视觉向量的维度,将每个训练样本对应的独热标签拼接为缓存模型的值矩阵 $\boldsymbol{L}_{\text{train}} \in R^{NK \times N}$. 对测试样本获取其视觉向量 $\boldsymbol{f}_{\text{test}}$ 后将其与键矩阵相乘得到相似度 $A = \exp(-\beta(1-\boldsymbol{f}_{\text{test}}\boldsymbol{F}_{\text{train}}^{\text{T}}))$,最后根据式(10)进行分类:

$$\hat{y} = \alpha A \boldsymbol{L}_{\text{train}} + (1-\alpha)\boldsymbol{f}_{\text{test}}\boldsymbol{W}_{\text{cls}} \tag{10}$$

其中 $\boldsymbol{W}_{\text{cls}}$ 是将每类的朴素提示拼接得到的矩阵,α 是超参数. 式(10)的第一项代表小样本训练集提供的分类信息,第二项即 CLIP 原始的零样本推理方案. SuS-X[167]认为 Tip-Adapter 计算相似度 A 时只在图像模态空间进行计算的做法并不是最优的,因此提出跨模态计算相似度. 该方法首先计算图像特征与文字特征的相似度:

$$S_{\text{train}} = \text{Softmax}(\boldsymbol{F}_{\text{train}}\boldsymbol{W}_{\text{cls}}),\ S_{\text{train}} \in R^{NK \times N} \tag{11}$$

$$S_{\text{test}} = \text{Softmax}(\boldsymbol{f}_{\text{test}}\boldsymbol{W}_{\text{cls}}),\ S_{\text{test}} \in R^{t \times N} \tag{12}$$

其中 t 是测试样本的数量. 然后计算跨模态的相似度 $M_{i,j} = KL(S_{\text{test}}^{i} \| S_{\text{train}}^{j})$,$KL()$代表 KL 散度,最后根据下式进行分类:

$$\hat{y} = \alpha A \boldsymbol{L}_{\text{train}} + (1-\alpha)\boldsymbol{f}_{\text{test}}\boldsymbol{W}_{\text{cls}} + \gamma\psi(-M)\boldsymbol{L}_{\text{train}} \tag{13}$$

其中 $\psi(\cdot)$对 M 中的元素进行放缩,γ 是超参数. 式(13)较式(10)增加的第三项即代表跨模态相似度对结果的影响. APE[168]进一步对 SuS-X 和 Tip-Adapter 加以改进,首先对 CLIP 提取的图像、文字特征进行修正,只取其中最有意义的部分通道,即对上文的 $\boldsymbol{F}_{\text{train}}$、$\boldsymbol{f}_{\text{test}}$、$\boldsymbol{W}_{\text{cls}}$都进行修正得到 $\boldsymbol{F}'_{\text{train}}$、$\boldsymbol{f}'_{\text{test}}$、$\boldsymbol{W}'_{\text{cls}}$,然后根据式(14)进行分类:

$$\hat{y} = \alpha A'(\text{diag}(R_{\boldsymbol{F}'\boldsymbol{W}'})\boldsymbol{L}_{\text{train}}) + (1-\alpha)\boldsymbol{f}'_{\text{test}}\boldsymbol{W}'_{\text{cls}} \tag{14}$$

其中 $R_{\boldsymbol{F}'\boldsymbol{W}'} = \exp(\lambda \cdot KL(\boldsymbol{F}'_{\text{train}}\boldsymbol{W}'_{\text{cls}} \| \boldsymbol{L}_{\text{train}}))$,可以认为是缓存模型中特征的权重,代表其表征的正确性以及对最终结果的贡献程度. 另外,该方法还设计了一种需要训练的 APE-T 方法以提升性能.

值得注意的是,基于缓存模型的方法只需要根据模型针对训练数据产生的输出构建缓存矩阵即可进行推理,不需要训练、微调任何的参数,因此其可调参数量是 0%,在参数量上的高效性达到极致,适合存储空间极度受限的平台.

5.4 其他方法

还有一些方法基于蒸馏、微调等技术对 VLM 进行迁移. TinyCLIP[29]设计了一种用于 VLM 的高效蒸馏框架. 该方法提出在跨模态空间进行蒸馏,让学生模型模仿教师模型中不同模态之间的信息相互关联;且在初始化学生模型时,将教师模型的部分权重转移到学生模型上,以降低蒸馏开销. PADCLIP[177]提出在目标域数据上微调 VLM,为了避免破坏其预训练结构,提出一种度量来显式控制参数更新学习率. UniMoS[175]发现 VLM 所提取的跨模态特征实际上没有很好地对齐,因此提出显式地将特征中归属于不同模态的信息分解出来分别在各自的模态进一步处理. 该方法不需要学习提示,因此不涉及对 VLM 预训练模块的任何前后传播,具有极高的参数和运算敏捷性,仅需训练一个相当于 2%总参数量的模态分离网络和分类器即可在领域自适应基准上达到前沿性能.

VLM 的设计和迁移是近年来的热点话题,已经出现不少综述文献对此领域的方法进行了总结整理. 如综述文献[179-182]就 VLM 的发展、不同结构、不同任务、不同预训练方法进行了较好的总结. 文献[183]就预训练 VLM 在视觉领域的应用进行调研,覆盖了 VLM 预训练、数据集资源、迁移方法、知识蒸馏等方面. 可以总结,跨模态深度模型和方法已经成为现在的研究主流,且有望在未来统一深度模型在不同领域、不同模态的应用,真正实现用一个超大预训练模型解决所有问题的构想.

多模态模型融合了计算机视觉、自然语言处理的常见模型,其中的许多基础大模型基于视觉和语言领域预训练模型的直接融合微调而成. 为了实现不同模态信息的高效传递和相互作用,本领域最常用的一类迁移方法为自适应模块法,通过可学习模块实现不同模态信息的微调. 由于多模态中常常涉及文字模态,提示学习也在多模态模型的高效迁移中起到重要作用,特别是通过学习跨模态提示实现不同模态之间的相互促进和提升.

5.5 常用数据集与实验结果

多模态领域的数据集覆盖范围广泛,从传统的单一模态视觉分类任务到跨模态的图像问答、为图片赋标题、导航、图像理解等都有所涉及. 表 10 以多个同类型的子数据集构建的基准数据集为单位总结了多模态模型在不同应用领域能解决的问题和挑战,接下来进行介绍.

表 10　多模态领域常用数据集

基准数据集	子数据集	描述	指标
图像分类基准数据集	ImageNet[184]	包含常见的 1000 个类别，1.28M 张训练样本	准确率
	Caltech101[132]	包含 101 个类别，共 9146 张图片	准确率
	OxfordPets[136]	宠物图像数据集，包含 37 种宠物，共 7349 个样本	准确率
	StandfordCars[185]	细粒度分类数据集，包含 196 种汽车，共 16185 张图片	准确率
	Flowers102[135]	包含 102 种花卉图像，每种花卉包含 40～258 个样本	准确率
	Food101[186]	包含 101 种食物图像，共 101000 个样本，包含些许噪声	准确率
	FGVCAircraft[149]	包含 100 种航空器的图片，每类包含 100 个样本	准确率
	SUN397[137]	大规模场景理解数据集，包含 397 个类别	准确率
	DTD[134]	包含 47 种常见的纹理，共 5640 个样本	准确率
	EuroSAT[142]	土地覆盖情况的卫星图像分类任务，共 10 个类别 27000 个样本	准确率
	UCF101[187]	视频切片动作识别数据集，共 101 个类别 13320 个样本	准确率
文字-图像基准数据集	VQAv2[188]	图像问答数据集，要求模型根据视觉理解、常识、语言理解等能力回答关于图像信息的问题	准确率
	GQA[189]	大型图像问答数据集，提供场景中的图片标注和预提取的图像特征	准确率
	NLVR^2[190]	视觉理解数据集，包含 92244 对生成图像-人工描述样本	准确率
	COCO caption[191]	图像标题数据集，包含超过 330000 张图片，每张图片对应 5 条人工标注的标题	CIDer
文字-视频基准数据集	TVQA[192]	大型视频问答数据集，基于 6 部电视剧收集了 21793 个视频切片. 对每个切片提出 1 个问题和 4 个备选项，模型需要选择正确的一个	准确率
	How2QA[193]	视频问答数据集，每个切片包含 1 个问题和 4 个备选项，共包含 9035 个视频片段	准确率
	TVC[194]	大型视频标题数据集，包含 108965 条短视频片段和 261490 条对应标题	CIDer
	YouCook2[195]	大型视频标题数据集，包含 2000 条未修剪的烹饪视频和对应的英语描述	CIDer
视觉-语义导航基准数据集	R2R[196]	真实建筑物内的自然语言导航数据集，任务要求智能体根据人工标注的导航指引在陌生的建筑物内移动	SPL
	REVERIE[197]	数据集就环境某处的一件物体提供描述，智能体需要在环境中移动并找到这件物体	RGSPL
	NDH[198]	要求智能体根据一段智能体和数据库的对话记录抵达目标区域	任务进度
	RxR[199]	多语种语言导航数据集，其规模是 R2R 的 10 倍，且包含英语、印地语和泰卢固语三种语言，道路的长度和多样性也有所提升	nDTW
领域自适应基准数据集	OfficeHome[200]	领域自适应图像分类数据集，包含 4 个领域艺术(A)，剪贴画(C)，物品(P)，真实(R)共 15500 个样本，65 个类别	准确率
	VisDA2017[201]	领域自适应图像分类数据集，包含 152397 张人工生成的样本作为源域和 55388 张真实世界样本作为目标域，共 12 个类别	准确率
语音-视觉基准数据集	VGG-Sound[202]	包含超过 200000 条分属于 309 个语音类别的切片，每个切片中都包含有发出声音的物体和对应的音频，例如演奏小提琴、吹头发等	mAP，AUC
	ImageHear[203]	包含 101 张图片，分属于 30 个类别，用于音频生成模型的 OOD(out-of-distribution)验证	准确率

绝大多数将 VLM 用于小样本迁移学习任务的方法都跟随 CLIP[147] 的实验设定，在一个包含 11 个子数据集的图像分类基准数据集上进行测试. 其中包括常见物体的分类、细粒度分类(花、车、飞行器等的具体类别区分)、纹理分类、卫星图像等应用，覆盖范围较广，凸显出了预训练 VLM 的泛化能力之强. 作为多模态大模型，许多方法将 VLM 应用于跨模态任务，如图像-文字任务中模型需要实现看图说话、看图答题、为图片总结标题等功能，其中 CIDEr[204] 是图像标注问题的评价指标；在视频-文字任务中模型需要理解视频内容并回答问题、为一段视频总结标题等；在视觉-语义导航任务中模型需要根据文字提示在模拟环境中操作一个智能体移动到相应的位置或完成具体的任务，其中 SPL[205] 用于衡量导航的精度和效率，RGSPL[171] 根据不同道路的权重衡量任务的完成度，nDTW[171] 惩罚偏离至标准路径之外的行为. 也有部分方法将 VLM 用于领域自适应任务，VLM 极强的零样本推理能力在目标域无标签时能提供可靠的伪标签，从而大幅提升在领域自适应任务上的效果.

本文进一步以应用范围最广的图像分类基准包含的 11 个子数据集为例，收集了多模态领域的代表性方法在上述数据集的实验结果进行比较分析，如表 11 所示. 其中的“CLIP”一栏为预训练 CLIP 模型的零样本推理(不经过任何微调)结果. 可以看出，经过小样本上高效微调后模型的性能有了大幅提升. 特别是在 FGVCAircraft(航空器图片分类)、EuroSAT(卫星图像分类)等专业性较强的数据集上，CLIP 的预训练知识涉及较少，难以作出有效的零样本推理，而高效迁移技术大幅提升了多模态模型的泛化性能.

表 11 多模态领域部分代表性方法在图像分类基准数据集上的实验结果

方法	骨干模型	会议/期刊	ImageNet	Caltech	Pets	Cars	Flowers	Food	FGVC
CLIP[162]	ViT-B/16	ICML	72.4	96.8	91.2	63.4	72.1	90.1	27.2
CoOp[34]	ViT-B/16	IJCV	76.5	98.0	93.7	78.1	97.6	88.3	40.4
CoCoOp[35]	ViT-B/16	CVPR	76.0	98.0	95.2	70.5	94.9	90.7	33.4
MaPLe[178]	ViT-B/16	CVPR	76.7	97.7	95.4	72.9	95.9	90.7	37.4
CLIP-Adapter[165]	ResNet50	IJCV	62.0	94.0	86.1	72.5	94.4	76.9	38.8
SuS-X[167]	ResNet50	ICCV	62.2	90.5	87.8	69.1	90.0	75.9	30.0
APE[168]	ResNet50	ICCV	63.5	92.5	89.0	70.0	92.3	78.5	31.4

6 现存问题和发展方向

作为一个新兴研究方向，现存高效迁移学习方法的研究进展有目共睹，但也存在一些挑战和待解决的问题，主要包括：

(1) 对有标注数据的依赖. 虽然微调等迁移方法所需的训练数据较重新训练模型大大减少，但仍无法实现完全无监督的迁移. 如 VLM 迁移中常见的小样本设定，虽然对每类仅有十余个标注样本的要求，在现实应用中有时也是难以满足的，且容易对这些小样本过拟合，降低方法的泛化性能.

(2) 高效性单一. 现有的高效迁移方法所宣称的“高效”几乎全部来源于参数和内存的高效性，即迁移时需要训练、存储的参数较完整的模型大大减少. 然而许多方法实际上需要的训练时间、内存显存占用、推理时间和开销等不比重新训练模型小. 如基于提示微调的方法虽然只需要训练极少的提示向量，但仍需要对预训练模型进行前向、反向传播以获得梯度，且推理时仍需要加载整个预训练模型，仍然限制了这些方法在实际场景下的应用.

(3) 评价标准不统一. 本领域发展时间较短，尚未形成一套公认的评价体系和标准的评测方案，不同的工作可能在实现、评价细节上有所不同，因此所报告的优越性能可靠性和实用性存疑. 文献[206]以自然语言处理领域中的基准数据集 GLUE[91] 和 SuperGLUE[92] 为例，指出部分方法报告的性能实际上来源于验证集而不是评测网站提供的测试集，因此存在数据泄露、随机种子调优等外部因素的干扰. 该文献进一步指出，许多宣称性能超过全微调的高效迁移方法所报告的结果难以稳定复现，如 VLM、自然语言模型中常用的提示微调方法就极不稳定，实际性能难以超越全微调. 因此，为了确保此领域的健康发展和现实应用价值，有必要提出一套完善的评判流程和方法论，以真正认识不同方法之间的优劣；或是提出一种新颖的高效性指标，能综合考虑算法的复杂度、训练推理时间、参数开销、收敛性能等，实现公平且全面的评价.

(4) 软硬件基础设施不足. 现有方法多数仍处于研究阶段，其代码实现仅支持在研究数据集上的验证，而缺乏配套的软硬件接口支撑. 虽然许多方法之间存在共性，如本文总结的不同应用领域都存在的基于自适应模块、自适应参数注入的方法，但其具体实现可能在不同工作中是不同的，限制了其扩展性.

基于上述的讨论，结合实际应用场景，本文总结出如下的潜在研究方向：

(1) 无监督、自监督的高效迁移学习. 针对目标领域完全不存在标注数据的情况进行模型泛化. 这在实际应用场景中是十分常见的，如自动驾驶技术的应用中可能面对的路况、汽车可能运行的区域、驾驶员的驾驶习惯都是不可预知的，而汽车上所能搭载的算力往往也十分有限，如何在算力不足、数据缺乏的情况下将预训练的自动驾驶模型泛化至任意的场合工况下就是十分实际的问题. 现存较为接近的设定包括领域自适应和领域泛化，前者假设目标域仅存在无标注的数据——以自动驾驶为例，可以让车辆在目标路况下由驾驶员进行测试驾驶提供；后者假设目标域是未知的，更贴近实际情况，但也难度更高. 另一种可能的解决方案是自监督学习. 仍然以自动驾驶为例，算法可以根据日渐积累的驾驶数据自我调节，以逐渐适配驾驶员的驾驶习惯和用车特点.

(2) 研发推理时间、内存、显存、部署高效的迁移方法. 现有高效迁移方法强调的多是训练参数量的高效性，而实际应用中资源的稀缺性常体现在更多方面. 如自动驾驶技术要求的推理性能极高，需要在极短的时延内返回结果；微型、可穿戴设备则要求的更强的部署高效性，其推理时占用的算力、部署所需的存储空间都应足够小；卫星、导航等系统则要求更高的迁移算法精度，希望迁移时损失的性能尽可能小. 因此，需要根据不同的应用场景和要求有侧重

地提出更丰富的高效迁移方法.

(3)形成一套完整的验证标准和软硬件支持方案.现有方法仍停留在研究阶段,其性能也仅在研究性数据集上得以验证.而目前工业应用中效果最好、应用最广的迁移方法仍是全微调.要想推广高效迁移方法的应用部署,首先需要设计一套标准的验证流程,指明算法运行的条件、验证数据、与其他方法对比的原则等,其次需要提供标准的高效模块、参数等实现接口,并提供自上而下的软硬件支持.只有可信地证明了高效方法较全微调方法的优越性、提供了便捷的实现和支持,才有可能真正推进深度迁移学习在更多资源受限环境下的应用.

另外,在实际应用时还需要考虑具体应用场景下"高效性"和"准确性"的权衡,讨论如何在牺牲最少模型性能的情况下换取更显著的开销缩减.现有方法多停留在实验室验证,缺乏在实际部署场景中的试验数据.

7　结　论

本文首次就资源受限条件下的高效迁移学习方法进行总结综述,结合研究热点,从传统的自然语言理解、计算机视觉和新兴的多模态模型这 3 大应用领域入手,深刻对比不同领域间高效迁移方法的异同,并总结出了 5 类高效迁移技术手段:修改模型结构、更新模型参数、修改模型输入输出、注入自适应参数和增加自适应模块.本文基于此 5 种分类方法对 3 大领域中的代表性方法进行了深度的解构和清晰的归纳,意在帮助读者迅速入门此领域并理解实现高效迁移的一般手段.最后对现有方法的局限性进行了讨论,并据此提出高效迁移学习的未来发展方向,期望能促进更多学者对本领域的思考和推进,为本领域的发展与推广应用提供基础思路和方向.

参　考　文　献

[1] Shen D, Wu G, Suk H I. Deep learning in medical image analysis. Annual Review of Biomedical Engineering, 2017, 19: 221-248

[2] Khan S, Yairi T. A review on the application of deep learning in system health management. Mechanical Systems and Signal Processing, 2018, 107: 241-265

[3] Kamilaris A, Prenafeta-Boldú F X. Deep learning in agriculture: A survey. Computers and Electronics in Agriculture, 2018, 147: 70-90

[4] Ben-David S, Blitzer J, Crammer K, et al. A theory of learning from different domains. Machine Learning, 2010, 79: 151-175

[5] Li X, Li J, Zuo L, et al. Domain adaptive remaining useful life prediction with transformer. IEEE Transactions on Instrumentation and Measurement, 2022, 71: 1-13

[6] Caruana R. Multitask learning. Machine Learning, 1997, 28: 41-75

[7] Russakovsky O, Deng J, Su H, et al. ImageNet large scale visual recognition challenge. International Journal of Computer Vision, 2015, 115: 211-252

[8] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 770-778

[9] Askell A, Bai Y, Chen A, et al. A general language assistant as a laboratory for alignment. arXiv preprint arXiv: 2112.00861, 2021

[10] Brown T, Mann B, Ryder N, et al. Language models are few-shot learners. Advances in Neural Information Processing Systems, 2020, 33: 1877-1901

[11] Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, 2022, 35: 27730-27744

[12] Chowdhery A, Narang S, Devlin J, et al. PaLM: Scaling language modeling with pathways. Journal of Machine Learning Research, 2023, 24(1): 1-113

[13] Ren X, Zhou P, Meng X, et al. PanGu-Σ: Towards trillion parameter language model with sparse heterogeneous computing. arXiv preprint arXiv:2303.10845, 2023

[14] Raffel C, Shazeer N, Roberts A, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. The Journal of Machine Learning Research, 2020, 21(1): 5485-5551

[15] Touvron H, Lavril T, Izacard G, et al. LLaMA: Open and efficient foundation language models. arXiv preprint arXiv: 2302.13971, 2023

[16] Gao H, Cheng B, Wang J, et al. Object classification using CNN-based fusion of vision and LIDAR in autonomous vehicle environment. IEEE Transactions on Industrial Informatics, 2018, 14(9): 4224-4231

[17] Gao H, Qin Y, Hu C, et al. An interacting multiple model for trajectory prediction of intelligent vehicles in typical road traffic scenario. IEEE Transactions on Neural Networks and Learning Systems, 2021, 34(9): 6468-6479

[18] Pan S J, Yang Q. A survey on transfer learning. IEEE Transactions on Knowledge and Data Engineering, 2009, 22(10): 1345-1359

[19] Long Mingsheng. Transfer Learning: Problems and Methods [Ph. D. dissertation]. Tsinghua University, Beijing, 2014 (in Chinese)

(龙明盛. 迁移学习问题与方法研究[博士学位论文]. 清华大学, 北京, 2014)

[20] Weiss K, Khoshgoftaar T M, Wang D D. A survey of transfer learning. Journal of Big Data, 2016, 3(1): 1-40

[21] Tan C, Sun F, Kong T, et al. A survey on deep transfer learning//Proceedings of the Artificial Neural Networks and Machine Learning—ICANN 2018: 27th International Conference on Artificial Neural Networks. Rhodes, Greece, 2018: 270-279

[22] Zhuang F, Qi Z, Duan K, et al. A comprehensive survey on transfer learning. Proceedings of the IEEE, 2020, 109(1): 43-76

[23] Iman M, Arabnia H R, Rasheed K. A review of deep transfer learning and recent advancements. Technologies, 2023, 11(2): 1-14

[24] Kitchenham B, Brereton O P, Budgen D, et al. Systematic literature reviews in software engineering—A systematic literature review. Information and Software Technology, 2009, 51(1): 7-15

[25] Treviso M, Lee J U, Ji T, et al. Efficient methods for natural language processing: A survey. Transactions of the Association for Computational Linguistics, 2023, 11: 826-860

[26] Menghani G. Efficient deep learning: A survey on making deep learning models smaller, faster, and better. ACM Computing Surveys, 2023, 55(12): 1-37

[27] Sandler M, Howard A, Zhu M, et al. MobileNetV2: Inverted residuals and linear bottlenecks//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 4510-4520

[28] Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531, 2015

[29] Wu K, Peng H, Zhou Z, et al. TinyCLIP: CLIP distillation via affinity mimicking and weight inheritance//Proceedings of the IEEE/CVF International Conference on Computer Vision. Paris, France, 2023: 21970-21980

[30] Touvron H, Cord M, Douze M, et al. Training data-efficient image transformers & distillation through attention//Proceedings of the International Conference on Machine Learning. Virtual, 2021: 10347-10357

[31] Yang X, Zhou D, Liu S, et al. Deep model reassembly. Advances in Neural Information Processing Systems, 2022, 35: 25739-25753

[32] Shu Y, Cao Z, Zhang Z, et al. Hub-Pathway: Transfer learning from a hub of pre-trained models. Advances in Neural Information Processing Systems, 2022, 35: 32913-32927

[33] Cubuk E D, Zoph B, Shlens J, et al. RandAugment: Practical automated data augmentation with a reduced search space//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Seattle, USA, 2020: 702-703

[34] Zhou K, Yang J, Loy C C, et al. Learning to prompt for vision-language models. International Journal of Computer Vision, 2022, 130(9): 2337-2348

[35] Zhou K, Yang J, Loy C C, et al. Conditional prompt learning for vision-language models//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 16816-16825

[36] Wang Z, Zhang Z, Lee C Y, et al. Learning to prompt for continual learning//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 139-149

[37] Wang R, Tang D, Duan N, et al. K-Adapter: Infusing knowledge into pre-trained models with adapters. arXiv preprint arXiv:2002.01808, 2020

[38] Mahabadi R K, Henderson J, Ruder S. Compacter: Efficient low-rank hypercomplex adapter layers. Advances in Neural Information Processing Systems, 2021, 34: 1022-1035

[39] Li Jing-Jing, Meng Li-Chao, Zhang Ke, et al. Review of studies on domain adaptation. Computer Engineering, 2021, 47(6): 1-13(in Chinese)
(李晶晶, 孟利超, 张可等. 领域自适应研究综述. 计算机工程, 2021, 47(6): 1-13)

[40] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. Advances in Neural Information Processing Systems, 2017, 30: 1-11

[41] Zhang A, Tay Y, Zhang S, et al. Beyond fully-connected layers with quaternions: Parameterization of hypercomplex multiplications with $1/n$ parameters. arXiv preprint arXiv: 2102.08597, 2021

[42] Li X L, Liang P. Prefix-tuning: Optimizing continuous prompts for generation//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). Bangkok, Thailand, 2021: 4582-4597

[43] Lester B, Al-Rfou R, Constant N. The power of scale for parameter-efficient prompt tuning. arXiv preprint arXiv: 2104.08691, 2021

[44] Hu E J, Wallis P, Allen-Zhu Z, et al. LoRA: Low-rank adaptation of large language models//Proceedings of the International Conference on Learning Representations. Virtual, 2021: 1-26

[45] Dettmers T, Pagnoni A, Holtzman A, et al. QLoRA: Efficient finetuning of quantized LLMs. arXiv preprint arXiv: 2305.14314, 2023

[46] Zaken E B, Goldberg Y, Ravfogel S. BitFit: Simple parameter-efficient fine-tuning for transformer-based masked language-models//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Dublin, Ireland, 2022: 1-9

[47] Pfeiffer J, Kamath A, Rücklé A, et al. AdapterFusion: Non-destructive task composition for transfer learning// Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. 2021: 487-503

[48] Liu H, Tam D, Muqeeth M, et al. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. Advances in Neural Information Processing Systems, 2022, 35: 1950-1965

[49] Sung Y L, Cho J, Bansal M. LST: Ladder side-tuning for parameter and memory efficient transfer learning. Advances in Neural Information Processing Systems, 2022, 35: 12991-13005

[50] Asai A, Salehi M, Peters M E, et al. ATTEMPT: Parameter-efficient multi-task tuning via attentional mixtures of soft prompts//Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Abu Dhabi, 2022: 6655-6672

[51] Houlsby N, Giurgiu A, Jastrzebski S, et al. Parameter-efficient transfer learning for NLP//Proceedings of the International Conference on Machine Learning. Long Beach, USA, 2019: 2790-2799

[52] Guo D, Rush A M, Kim Y. Parameter-efficient transfer learning with diff pruning//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). Bangkok, Thailand, 2021: 4884-4896

[53] Mahabadi R K, Ruder S, Dehghani M, et al. Parameter-efficient multi-task fine-tuning for transformers via shared hypernetworks//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). Bangkok, Thailand, 2021: 565-576

[54] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation, 1997, 9(8): 1735-1780

[55] Liu S Y, Wang C Y, Yin H, et al. DoRA: Weight-decomposed low-rank adaptation. arXiv preprint arXiv:2402.09353, 2024

[56] Mao Y, Huang K, Guan C, et al. DoRA: Enhancing parameter-efficient fine-tuning with dynamic rank distribution. arXiv preprint arXiv:2405.17357, 2024

[57] Kopiczko D J, Blankevoort T, Asano Y M. VeRA: Vector-based random matrix adaptation. arXiv preprint arXiv: 2310.11454, 2023

[58] Wei J, Bosma M, Zhao V, et al. Finetuned language models are zero-shot learners//Proceedings of the International Conference on Learning Representations. Virtual, 2021: 1-46

[59] Min S, Lyu X, Holtzman A, et al. Rethinking the role of demonstrations: What makes in-context learning work?// Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Abu Dhabi, 2022: 11048-11064

[60] Dong Q, Li L, Dai D, et al. A survey on in-context learning. arXiv preprint arXiv:2301.00234, 2022

[61] He J, Zhou C, Ma X, et al. Towards a unified view of parameter-efficient transfer learning//Proceedings of the International Conference on Learning Representations. Virtual, 2021: 1-15

[62] Ding N, Qin Y, Yang G, et al. Parameter-efficient fine-tuning of large-scale pre-trained language models. Nature Machine Intelligence, 2023, 5(3): 220-235

[63] Han Z, Gao C, Liu J, et al. Parameter-efficient fine-tuning for large models: A comprehensive survey. arXiv preprint arXiv:2403.14608, 2024

[64] Zhao W X, Zhou K, Li J, et al. A survey of large language models. arXiv preprint arXiv:2303.18223, 2023

[65] Warstadt A, Singh A, Bowman S R. Neural network acceptability judgments. Transactions of the Association for Computational Linguistics, 2019, 7: 625-641

[66] Matthews B W. Comparison of the predicted and observed secondary structure of T4 phage Lysozyme. Biochimica et Biophysica Acta (BBA)-Protein Structure, 1975, 405(2): 442-451

[67] Socher R, Perelygin A, Wu J, et al. Recursive deep models for semantic compositionality over a sentiment treebank// Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. Seattle, USA, 2013: 1631-1642

[68] Dolan B, Brockett C. Automatically constructing a corpus of sentential paraphrases//Proceedings of the 3rd International Workshop on Paraphrasing (IWP2005). Jeju Island, Republic of Korea, 2005: 1-8

[69] Wang Z, Hamza W, Florian R. Bilateral multi-perspective matching for natural language sentences. arXiv preprint arXiv: 1702.03814, 2017

[70] Bentivogli L, Bernardi R, Marelli M, et al. SICK through the SemEval glasses. Lesson learned from the evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. Language Resources and Evaluation, 2016, 50: 95-124

[71] Cer D, Diab M, Agirre E, et al. Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation. arXiv preprint arXiv:1708.00055, 2017

[72] Williams A, Nangia N, Bowman S R. A broad-coverage challenge corpus for sentence understanding through inference. arXiv preprint arXiv:1704.05426, 2017

[73] Rajpurkar P, Zhang J, Lopyrev K, et al. SQuAD: 100,000+ questions for machine comprehension of text. arXiv preprint arXiv:1606.05250, 2016

[74] White A S, Rastogi P, Duh K, et al. Inference is everything: Recasting semantic resources into a unified evaluation framework//Proceedings of the 8th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Taipei, China, 2017: 996-1005

[75] Demszky D, Guu K, Liang P. Transforming question answering datasets into natural language inference datasets. arXiv preprint arXiv:1809.02922, 2018

[76] Dagan I, Glickman O, Magnini B. The PASCAL recognising textual entailment challenge//Proceedings of the Machine Learning Challenges Workshop. Berlin, Germany, 2005: 177-190

[77] Haim R B, Dagan I, Dolan B, et al. The second PASCAL recognising textual entailment challenge//Proceedings of the 2nd PASCAL Challenges Workshop on Recognising Textual Entailment. Venice, Italy, 2006, 7: 785-794

[78] Giampiccolo D, Magnini B, Dagan I, et al. The third PASCAL recognizing textual entailment challenge//Proceedings of the ACL—PASCAL Workshop on Textual Entailment and Paraphrasing. Prague, Czech, 2007: 1-9

[79] Bentivogli L, Clark P, Dagan I, et al. The fifth PASCAL recognizing textual entailment challenge//Proceedings of the 2nd Text Analysis Conference. Gaithersburg, USA, 2009: 7-8

[80] Levesque H, Davis E, Morgenstern L. The Winograd schema challenge//Proceedings of the 13th International Conference on the Principles of Knowledge Representation and Reasoning. Rome, Italy, 2012: 1-10

[81] De Marneffe M C, Simons M, Tonhauser J. The Commitment-Bank: Investigating projection in naturally occurring discourse. Proceedings of Sinn und Bedeutung, 2019, 23(2): 107-124

[82] Clark C, Lee K, Chang M W, et al. BoolQ: Exploring the surprising difficulty of natural yes/no questions//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, USA, 2019: 2924-2936

[83] Roemmele M, Bejan C A, Gordon A S. Choice of plausible alternatives: An evaluation of commonsense causal reasoning//Proceedings of the 2011 AAAI Spring Symposium Series. Palo Alto, USA, 2011: 1-6

[84] Khashabi D, Chaturvedi S, Roth M, et al. Looking beyond the surface: A challenge set for reading comprehension over multiple sentences//Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). New Orleans, USA, 2018: 252-262

[85] Zhang S, Liu X, Liu J, et al. ReCoRD: Bridging the gap between human and machine commonsense reading comprehension. arXiv preprint arXiv:1810.12885, 2018

[86] Pilehvar M T, Camacho-Collados J. WiC: The word-in-context dataset for evaluating context-sensitive meaning representations//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, USA, 2019: 1267-1273

[87] Novikova J, Dušek O, Rieser V. The E2E dataset: New challenges for end-to-end generation. arXiv preprint arXiv: 1706.09254, 2017

[88] Papineni K, Roukos S, Ward T, et al. BLEU: A method for automatic evaluation of machine translation//Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. Philadelphia, USA, 2002: 311-318

[89] Lavie A and Agarwal A. METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments//Proceedings of the Second Workshop on Statistical Machine Translation. Prague, Czech Republic, 2007: 228-231

[90] Gardent C, Shimorina A, Narayan S, et al. The WebNLG challenge: Generating text from RDF data//Proceedings of the 10th International Conference on Natural Language Generation. Santiago de Compostela, Spain, 2017: 124-133

[91] Wang A, Singh A, Michael J, et al. GLUE: A multi-task benchmark and analysis platform for natural language understanding//Proceedings of the International Conference on Learning Representations. Vancouver, Canada, 2018: 1-20

[92] Wang A, Pruksachatkun Y, Nangia N, et al. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. Advances in Neural Information Processing Systems, 2019, 32: 1-15

[93] Gao H, Lv C, Zhang T, et al. A structure constraint matrix factorization framework for human behavior segmentation. IEEE Transactions on Cybernetics, 2021, 52(12): 12978-12988

[94] Gao H, Su H, Cai Y, et al. Trajectory prediction of cyclist based on dynamic Bayesian network and long short-term memory model at unsignalized intersections. Science China Information Sciences, 2021, 64(7): 172207

[95] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems, 2012, 25: 1-9

[96] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv: 1409.1556, 2014

[97] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 1-9

[98] Xie S, Girshick R, Dollár P, et al. Aggregated residual transformations for deep neural networks//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 1492-1500

[99] Zhang H, Wu C, Zhang Z, et al. ResNeSt: Split-attention networks//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 2736-2746

[100] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929, 2020

[101] Liu Z, Lin Y, Cao Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, Canada, 2021: 10012-10022

[102] Huang G, Liu Z, Van Der Maaten L, et al. Densely connected convolutional networks//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 4700-4708

[103] Howard A G, Zhu M, Chen B, et al. MobileNets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861, 2017

[104] Tan M, Le Q. EfficientNet: Rethinking model scaling for convolutional neural networks//Proceedings of the International Conference on Machine Learning. Long Beach, USA, 2019: 6105-6114

[105] Gan Y, Bai Y, Lou Y, et al. Decorate the newcomers: Visual domain prompt for continual test time adaptation//Proceedings of the AAAI Conference on Artificial Intelligence. Washington, USA, 2023, 37(6): 7595-7603

[106] Li J, Jing M, Su H, et al. Faster domain adaptation networks. IEEE Transactions on Knowledge and Data Engineering, 2021, 34(12): 5770-5783

[107] Shu Y, Kou Z, Cao Z, et al. Zoo-tuning: Adaptive transfer from a zoo of models//Proceedings of the International Conference on Machine Learning. Virtual, 2021: 9626-9637

[108] Jia M, Tang L, Chen B C, et al. Visual prompt tuning//Proceedings of the European Conference on Computer Vision. Tel Aviv, lsrael, 2022: 709-727

[109] Chen H, Tao R, Zhang H, et al. Conv-Adapter: Exploring parameter efficient transfer learning for ConvNets//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2024: 1551-1561

[110] Cai H, Gan C, Zhu L, et al. TinyTL: Reduce memory, not parameters for efficient on-device learning. Advances in Neural Information Processing Systems, 2020, 33: 11285-11297

[111] Chen S, Ge C, Tong Z, et al. AdaptFormer: Adapting vision transformers for scalable visual recognition. Advances in Neural Information Processing Systems, 2022, 35: 16664-16678

[112] He X, Li C, Zhang P, et al. Parameter-efficient model adaptation for vision transformers//Proceedings of the AAAI Conference on Artificial Intelligence. Washington, USA, 2023, 37(1): 817-825

[113] He H, Cai J, Zhang J, et al. Sensitivity-aware visual parameter-efficient fine-tuning//Proceedings of the IEEE/CVF International Conference on Computer Vision. Paris, France, 2023: 11825-11835

[114] Jie S, Deng Z H. Convolutional bypasses are better vision transformer adapters. arXiv preprint arXiv: 2207.07039, 2022

[115] Tu C H, Mai Z, Chao W L. Visual query tuning: Towards effective usage of intermediate representations for parameter and memory efficient transfer learning//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada, 2023: 7725-7735

[116] Evci U, Dumoulin V, Larochelle H, et al. Head2Toe: Utilizing intermediate representations for better transfer learning//Proceedings of the International Conference on Machine Learning. Baltimore, USA, 2022: 6009-6033

[117] Das R, Dukler Y, Ravichandran A, et al. Learning expressive prompting with residuals for vision transformers//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada, 2023: 3366-3377

[118] Wu Z, Nagarajan T, Kumar A, et al. BlockDrop: Dynamic inference paths in residual networks//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 8817-8826

[119] Wang H, Li S, Su S, et al. RDI-Net: Relational dynamic inference networks//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, Canada, 2021: 4621-4630

[120] Lee S, Chang S, Kwak N. URNet: User-resizable residual networks with conditional gating module//Proceedings of the AAAI Conference on Artificial Intelligence. New York, USA, 2020, 34(4): 4569-4576

[121] Li X, Li J, Li F, et al. Agile multi-source-free domain adaptation//Proceedings of the AAAI Conference on Artificial Intelligence. Vancouver, Canada, 2024, 38(12): 13673-13681

[122] Tran A T, Nguyen C V, Hassner T. Transferability and hardness of supervised classification tasks//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, Republic of Korea, 2019: 1395-1405

[123] Nguyen C, Hassner T, Seeger M, et al. LEEP: A new measure to evaluate transferability of learned representations //Proceedings of the International Conference on Machine Learning. Virtual, 2020: 7294-7305

[124] You K, Liu Y, Wang J, et al. LogME: Practical assessment of pre-trained models for transfer learning//Proceedings of

the International Conference on Machine Learning. Virtual, 2021: 12133-12143

[125] You K, Liu Y, Zhang Z, et al. Ranking and tuning pre-trained models: A new paradigm for exploiting model hubs. The Journal of Machine Learning Research, 2022, 23(1): 9400-9446

[126] Pándy M, Agostinelli A, Uijlings J, et al. Transferability estimation using Bhattacharyya class separability//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 9172-9182

[127] Tan Y, Li Y, Huang S L. OTCE: A transferability metric for cross-domain cross-task representations//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 15779-15788

[128] Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society Series B: Statistical Methodology, 2006, 68(1): 49-67

[129] Bahng H, Jahanian A, Sankaranarayanan S, et al. Exploring visual prompts for adapting large-scale models. arXiv preprint arXiv:2203.17274, 2022

[130] Rücklé A, Geigle G, Glockner M, et al. AdapterDrop: On the efficiency of adapters in transformers. arXiv preprint arXiv:2010.11918, 2020

[131] Zhai X, Puigcerver J, Kolesnikov A, et al. A large-scale study of representation learning with the visual task adaptation benchmark. arXiv preprint arXiv:1910.04867, 2019

[132] Fei-Fei L, Fergus R, Perona P. Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories//Proceedings of the 2004 Conference on Computer Vision and Pattern Recognition Workshop. Washington, USA, 2004: 59-70

[133] Krizhevsky A. Learning Multiple Layers of Features from Tiny Images [M. S. dissertation]. University of Toronto, Canada, 2009

[134] Cimpoi M, Maji S, Kokkinos I, et al. Describing textures in the wild//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA, 2014: 3606-3613

[135] Nilsback M E, Zisserman A. Automated flower classification over a large number of classes//Proceedings of the 2008 6th Indian Conference on Computer Vision, Graphics & Image Processing. Bhubaneswar, India, 2008: 722-729

[136] Parkhi O M, Vedaldi A, Zisserman A, et al. Cats and dogs //Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition. Providence, USA, 2012: 3498-3505

[137] Xiao J, Hays J, Ehinger K A, et al. Sun database: Large-scale scene recognition from abbey to zoo//Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Francisco, USA, 2010: 3485-3492

[138] Goodfellow I J, Bulatov Y, Ibarz J, et al. Multi-digit number recognition from street view imagery using deep convolutional neural networks. arXiv preprint arXiv:1312.6082, 2013

[139] Veeling B S, Linmans J, Winkens J, et al. Rotation equivariant CNNs for digital pathology//Proceedings of the Medical Image Computing and Computer Assisted Intervention—MICCAI 2018: 21st International Conference. Granada, Spain, 2018: 210-218

[140] Dugas E, Jorge J, Cukierski W. Diabetic Retinopathy Detection. URL https://kaggle.com/competitions/diabetic-retinopathy-detection, 2015

[141] Cheng G, Han J, Lu X. Remote sensing image scene classification: Benchmark and state of the art. Proceedings of the IEEE, 2017, 105(10): 1865-1883

[142] Helber P, Bischke B, Dengel A, et al. EuroSAT: A novel dataset and deep learning benchmark for land use and land cover classification. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2019, 12(7): 2217-2226

[143] Johnson J, Hariharan B, Van Der Maaten L, et al. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 2901-2910

[144] Matthey L, Higgins I, Hassabis D, et al. dSprites: Disentanglement testing sprites dataset (2017). URL https://github.com/deepmind/dsprites-dataset, 2017

[145] LeCun Y, Huang F J, Bottou L. Learning methods for generic object recognition with invariance to pose and lighting//Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Washington, USA, 2004, 2: 1-8

[146] Beattie C, Leibo J Z, Teplyashin D, et al. DeepMind lab. arXiv preprint arXiv:1612.03801, 2016

[147] Geiger A, Lenz P, Stiller C, et al. Vision meets robotics: The KITTI dataset. The International Journal of Robotics Research, 2013, 32(11): 1231-1237

[148] Saenko K, Kulis B, Fritz M, et al. Adapting visual category models to new domains//Proceedings of the Computer Vision—ECCV 2010: 11th European Conference on Computer Vision. Heraklion, Crete, Greece, 2010: 213-226

[149] Maji S, Rahtu E, Kannala J, et al. Fine-grained visual classification of aircraft. arXiv preprint arXiv:1306.5151, 2013

[150] Long M, Zhu H, Wang J, et al. Deep transfer learning with joint adaptation networks//Proceedings of the International Conference on Machine Learning. Sydney, Australia, 2017: 2208-2217

[151] Wah C, Branson S, Welinder P, et al. The caltech-UCSD birds-200-2011 dataset. Technical Report, 2011. URL https://authors.library.caltech.edu/records/cvm3y-5hh21

[152] Van Horn G, Branson S, Farrell R, et al. Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 595-604

[153] Khosla A, Jayadevaprakash N, Yao B, et al. Novel dataset for fine-grained image categorization: Standford dogs//Proceedings of the 1st Workshop on Fine Grained Visual Categorization. Columbus, USA, 2012: 1-2

[154] Gebru T, Krause J, Wang Y, et al. Fine-grained car detection for visual census estimation//Proceedings of the AAAI Conference on Artificial Intelligence. San Francisco, USA, 2017, 31(1): 4502-4508

[155] Li C, Liu H, Li L, et al. ELEVATER: A benchmark and toolkit for evaluating language-augmented visual models. Advances in Neural Information Processing Systems, 2022, 35: 9287-9301

[156] Shaban A, Bansal S, Liu Z, et al. One-shot learning for semantic segmentation. arXiv preprint arXiv:1709.03410, 2017

[157] Everingham M, Eslami S M A, Van Gool L, et al. The PASCAL visual object classes challenge: A retrospective. International Journal of Computer Vision, 2015, 111: 98-136

[158] Cordts M, Omran M, Ramos S, et al. The cityscapes dataset for semantic urban scene understanding//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 3213-3223

[159] Lin T Y, Maire M, Belongie S, et al. Microsoft COCO: Common objects in context//Proceedings of the Computer Vision—ECCV 2014: 13th European Conference. Zurich, Switzerland, 2014: 740-755

[160] Ramesh A, Pavlov M, Goh G, et al. Zero-shot text-to-image generation//Proceedings of the International Conference on Machine Learning. Virtual, 2021: 8821-8831

[161] Liu Y, Zhang K, Li Y, et al. Sora: A review on background, technology, limitations, and opportunities of large vision models. arXiv preprint arXiv:2402.17177, 2024

[162] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision//Proceedings of the International Conference on Machine Learning. Virtual, 2021: 8748-8763

[163] Jia C, Yang Y, Xia Y, et al. Scaling up visual and vision-language representation learning with noisy text supervision//Proceedings of the International Conference on Machine Learning. Virtual, 2021: 4904-4916

[164] Wang F, Li M, Lin X, et al. Learning to decompose visual features with latent textual prompts//Proceedings of the 11th International Conference on Learning Representations. Virtual, 2022: 1-13

[165] Gao P, Geng S, Zhang R, et al. CLIP-Adapter: Better vision-language models with feature adapters. International Journal of Computer Vision, 2023, 132: 581-595

[166] Zhang R, Fang R, Zhang W, et al. Tip-Adapter: Training-free CLIP-adapter for better vision-language modeling. arXiv preprint arXiv:2111.03930, 2021

[167] Udandarao V, Gupta A, Albanie S. SuS-X: Training-free name-only transfer of vision-language models//Proceedings of the IEEE/CVF International Conference on Computer Vision. Paris, France, 2023: 2725-2736

[168] Zhu X, Zhang R, He B, et al. Not all features matter: Enhancing few-shot clip with adaptive prior refinement. arXiv preprint arXiv:2304.01195, 2023

[169] Sung Y L, Cho J, Bansal M. VL-Adapter: Parameter-efficient transfer learning for vision-and-language tasks//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 5227-5237

[170] Ye H, Zhang J, Liu S, et al. IP-Adapter: Text compatible image prompt adapter for text-to-image diffusion models. arXiv preprint arXiv:2308.06721, 2023

[171] Qiao Y, Yu Z, Wu Q. VLN-PETL: Parameter-efficient transfer learning for vision-and-language navigation//Proceedings of the IEEE/CVF International Conference on Computer Vision. Paris, France, 2023: 15443-15452

[172] Gao P, Han J, Zhang R, et al. LLaMA-Adapter V2: Parameter-efficient visual instruction model. arXiv preprint arXiv:2304.15010, 2023

[173] Hu Z Y, Li Y, Lyu M R, et al. VL-PET: Vision-and-language parameter-efficient tuning via granularity control//Proceedings of the IEEE/CVF International Conference on Computer Vision. Paris, France, 2023: 3010-3020

[174] Ge C, Huang R, Xie M, et al. Domain adaptation via prompt learning. arXiv preprint arXiv:2202.06687, 2022

[175] Li X, Li Y, Du Z, et al. Split to Merge: Unifying separated modalities for unsupervised domain adaptation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2024: 23364-23374

[176] Du Z, Li X, Li F, et al. Domain-agnostic mutual prompting for unsupervised domain adaptation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2024: 23375-23384

[177] Lai Z, Vesdapunt N, Zhou N, et al. PADCLIP: Pseudo-labeling with adaptive debiasing in clip for unsupervised domain adaptation//Proceedings of the IEEE/CVF International Conference on Computer Vision. Paris, France, 2023: 16155-16165

[178] Khattak M U, Rasheed H, Maaz M, et al. MaPLe: Multi-modal prompt learning//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada, 2023: 19113-19122

[179] Du Y, Liu Z, Li J, et al. A survey of vision-language pre-trained models. arXiv preprint arXiv:2202.10936, 2022

[180] Long S, Cao F, Han S C, et al. Vision-and-language pretrained models: A survey. arXiv preprint arXiv:2204.07356, 2022

[181] Chen F L, Zhang D Z, Han M L, et al. VLP: A survey on vision-language pre-training. Machine Intelligence Research, 2023, 20(1): 38-56

[182] Yin Jiong, Zhang Zhe-Dong, Gao Yu-Han, et al. Survey on vision-language pre-training. Journal of Software, 2023, 34(5): 2000-2023(in Chinese)
(殷炯, 张哲东, 高宇涵等. 视觉语言预训练综述. 软件学报, 2023, 34(5): 2000-2023)

[183] Zhang J, Huang J, Jin S, et al. Vision-language models for vision tasks: A survey. arXiv preprint arXiv:2304.00685, 2023

[184] Deng J, Dong W, Socher R, et al. ImageNet: A large-scale hierarchical image database//Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami, USA, 2009: 248-255

[185] Krause J, Stark M, Deng J, et al. 3D object representations for fine-grained categorization//Proceedings of the IEEE International Conference on Computer Vision Workshops. Portland, USA, 2013: 554-561

[186] Bossard L, Guillaumin M, Van Gool L. Food-101—Mining discriminative components with random forests//Proceedings of the Computer Vision—ECCV 2014: 13th European Conference. Zurich, Switzerland, 2014: 446-461

[187] Soomro K, Zamir A R, Shah M. UCF101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402, 2012

[188] Goyal Y, Khot T, Summers-Stay D, et al. Making the V in VQA matter: Elevating the role of image understanding in visual question answering//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 6904-6913

[189] Hudson D A, Manning C D. GQA: A new dataset for real-world visual reasoning and compositional question answering//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 6700-6709

[190] Suhr A, Zhou S, Zhang A, et al. A corpus for reasoning about natural language grounded in photographs. arXiv preprint arXiv:1811.00491, 2018

[191] Chen X, Fang H, Lin T Y, et al. Microsoft COCO captions: Data collection and evaluation server. arXiv preprint arXiv:1504.00325, 2015

[192] Lei J, Yu L, Bansal M, et al. TVQA: Localized, compositional video question answering. arXiv preprint arXiv:1809.01696, 2018

[193] Li L, Chen Y C, Cheng Y, et al. HERO: Hierarchical encoder for video+language omni-representation pre-training. arXiv preprint arXiv:2005.00200, 2020

[194] Lei J, Yu L, Berg T L, et al. TVR: A large-scale dataset for video-subtitle moment retrieval//Proceedings of the Computer Vision—ECCV 2020: 16th European Conference. Glasgow, UK, 2020: 447-463

[195] Zhou L, Xu C, Corso J. Towards automatic learning of procedures from web instructional videos//Proceedings of the AAAI Conference on Artificial Intelligence. New Orleans, USA, 2018, 32(1): 7590-7598

[196] Anderson P, Wu Q, Teney D, et al. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 3674-3683

[197] Qi Y, Wu Q, Anderson P, et al. REVERIE: Remote embodied visual referring expression in real indoor environments//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 9982-9991

[198] Thomason J, Murray M, Cakmak M, et al. Vision-and-dialog navigation//Proceedings of the Conference on Robot Learning. Virtual, 2020: 394-406

[199] Ku A, Anderson P, Patel R, et al. Room-across-Room: Multilingual vision-and-language navigation with dense spatiotemporal grounding. arXiv preprint arXiv:2010.07954, 2020

[200] Venkateswara H, Eusebio J, Chakraborty S, et al. Deep hashing network for unsupervised domain adaptation//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 5018-5027

[201] Peng X, Usman B, Kaushik N, et al. VisDA: The visual domain adaptation challenge. arXiv preprint arXiv:1710.06924, 2017

[202] Chen H, Xie W, Vedaldi A, et al. VGGSound: A large-scale audio-visual dataset//Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Virtual, 2020: 721-725

[203] Sheffer R, Adi Y. I hear your true colors: Image guided audio generation//Proceedings of the ICASSP 2023—2023 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP). Rhodes Island, Greece, 2023: 1-5

[204] Vedantam R, Lawrence Zitnick C, Parikh D. CIDEr: Consensus-based image description evaluation//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 4566-4575

[205] Anderson P, Chang A, Chaplot D S, et al. On evaluation of embodied navigation agents. arXiv preprint arXiv:1807.06757, 2018

[206] Chen G, Liu F, Meng Z, et al. Revisiting parameter-efficient tuning: Are we really there yet?//Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Abu Dhabi, 2022: 2612-2626

LI Xin-Yao, Ph. D. candidate. His research interests include transfer learning, domain adaptation, and large foundation models.

LI Jing-Jing, Ph. D., professor. His research interests include computer vision, multimedia analysis, domain adaptation and recommender systems.

ZHU Lei, Ph. D., professor. His research interests include large-scale multimedia data retrieval and analysis, data mining, and deep learning.

SHEN Heng-Tao, Ph. D., professor. His research interests include multimedia search, computer vision, artificial intelligence, and big data management.

Background

Efficient transfer learning is a thriving research area aiming to reduce the computation costs of transferring knowledge in large models. As the size and complexity of pre-trained models continue to increase, the need for efficient methods to adapt these models to specific tasks or domains has become increasingly urgent. This work serves as the first comprehensive literature review on efficient transfer learning, offering a timely and in-depth analysis of this rapidly evolving field. The contributions of this work are threefold. (1) This work systematically summarizes recent advances in efficient transfer learning; (2) This work proposes a novel framework that effectively categorizes efficient transfer learning methods into 5 representative classes; (3) This work provides insights on the future development of efficient transfer learning.