E-mail: jig@aircas.ac.cn
Website: www.cjig.cn
Tel: 010-58887035
中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS

中图法分类号:TP391.4　　**文献标识码**:A　　**文章编号**:1006-8961(2024)06-1510-25

论文引用格式:Wang Y W, Shen T, Zhang S Y, Wu F, Zhao Z, Cai H B, Lyu C F, Ma L Z, Yang C L and Wu F. 2024. Advances in edge-cloud collaboration and evolution for large-small models. Journal of Image and Graphics, 29(06):1510-1534(王永威, 沈弢, 张圣宇, 吴帆, 赵洲, 蔡海滨, 吕承飞, 马利庄, 杨承磊, 吴飞. 2024. 大小模型端云协同进化技术进展. 中国图象图形学报, 29(06):1510-1534)[DOI:10.11834/jig.240011]

大小模型端云协同进化技术进展

王永威[1,2],沈弢[1],张圣宇[1],吴帆[3],赵洲[1],蔡海滨[4],吕承飞[1,5],
马利庄[3],杨承磊[6],吴飞[1,2*]

1. 浙江大学人工智能研究所,杭州　310058; 2. 浙江大学上海高等研究院,上海　201203;
3. 上海交通大学计算机科学与工程系,上海　200241; 4. 华东师范大学软件工程学院,上海　200062;
5. 淘宝(中国)软件有限公司,杭州　310023; 6. 山东大学软件学院,济南　250011

摘　要:生成式基座大模型正在引发人工智能领域的重大变革,在自然语言处理、多模态理解与内容合成等任务展现通用能力。大模型部署于云侧提供通用智能服务,但面临时延大、个性化不足等关键挑战,小模型部署于端侧捕捉个性化场景数据,但存在泛化性不足的难题。大小模型端云协同技术旨在结合大模型通用能力和小模型专用能力,以协同交互方式学习演化进而赋能下游垂直行业场景。本文以大语言模型和多模态大模型为代表,梳理生成式基座大模型的主流架构、典型预训练技术和适配微调等方法,介绍在大模型背景下模型剪枝、模型量化和知识蒸馏等大模型小型化关键技术的发展历史和研究近况,依据模型间协作目的及协同原理异同,提出大小模型协同训练、协同推理和协同规划的协同进化分类方法,概述端云模型双向蒸馏、模块化设计和生成式智能体等系列代表性新技术、新思路。总体而言,本文从生成式基座大模型、大模型小型化技术和大小模型端云协同方式3个方面探讨大小模型协同进化的国际和国内发展现状,对比优势和差距,并从应用前景、模型架构设计、垂直领域模型融合、个性化和安全可信挑战等层面分析基座赋能发展趋势。

关键词:生成式大模型;大模型小型化;大小模型协同进化;端云协同进化;生成式智能体;生成式人工智能

Advances in edge-cloud collaboration and evolution for large-small models

Wang Yongwei[1,2], Shen Tao[1], Zhang Shengyu[1], Wu Fan[3], Zhao Zhou[1], Cai Haibin[4],
Lyu Chengfei[1,5], Ma Lizhuang[3], Yang Chenglei[6], Wu Fei[1,2*]

1. Institute of Artificial Intelligence, Zhejiang University, Hangzhou　310058, China; 2. Shanghai Institute for Advanced Study, Zhejiang University, Shanghai　201203, China; 3. Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai　200241, China; 4. School of Software Engineering, East China Normal University, Shanghai　200062, China; 5. Taobao (China) Software Co., Ltd., Hangzhou　310023, China; 6. School of Software, Shandong University, Jinan　250011, China

Abstract: Generative foundation models are facilitating significant transformations in the field of artificial intelligence. They demonstrate general artificial intelligence in diverse research fields, including natural language processing, multi-

收稿日期:2024-01-09;**修回日期**:2024-02-23;**预印本日期**:2024-03-01

*通信作者:吴飞　wufei@zju.edu.cn

基金项目:新一代人工智能国家科技重大专项(2022ZD0119100);国家自然科学基金项目(62037001, 62441605);浙江省科技计划项目(2022C01044);繁星科学基金项目(浙江大学)

Supported by: National Science and Technology Major Project (2022ZD 0119100); National Natural Science Foundation of China (62037001, 62441605); Program of Zhejiang Province Science and Technology (2022C01044)

modal content understanding, imagery, and multimodal content synthesis. Generative foundation models often consist of billions or even hundreds of billions of parameters. Thus, they are often deployed on the cloud side to provide powerful and general intelligent services. However, this type of service can be confronted with crucial challenges in practice, such as high latency induced by communications between the cloud and local devices, and insufficient personalization capabilities due to the fact that servers often do not have access to local data considering privacy concerns. By contrast, low-complexity lightweight models are located at the edge side to capture personalized and dynamic scenario data. However, they may suffer from poor generalization. Large and lightweight (or large-small) model collaboration aims to integrate the general intelligence of large foundation models and the personalized intelligence of small lightweight models. This integration empowers downstream vertical domain-specific applications through the interaction and collaboration of both types of intelligent models. Large and small model collaboration has recently attracted increasing attention and becomes the focus of research and development in academia and industry. It has also been predicted to be an important trend in technology. We therefore try to thoroughly investigate this area by highlighting recent progress and bringing potential inspirations for related research. In this study, we first overview representative large language models (LLMs) and large multimodal models. We focus on their mainstream Transformer-based model architectures including encoder-only, decoder-only, and encoder-decoder models. Corresponding pre-training technologies such as next sentence prediction, sequence-to-sequence modeling, contrastive learning, and parameter-efficient fine-tuning methods with representatives including low-rank adaptation and prompt tuning are also explored. We then review the development history and the latest advancement of model compression techniques, including model pruning, model quantization, and knowledge distillation in the era of foundation models. Based on the differences in terms of model collaboration purposes and mechanisms, we propose a new classification method and taxonomies for the large-small model collaboration study, namely, collaborative training, collaborative inference, and collaborative planning. Specifically, we summarize recent and representative methods that consist of dual-directional knowledge distillation between large models at the cloud side and small models deployed at the edge side, modular design of intelligent models that split functional models between the cloud and edge, and generative agents that collaborate to complete more complex tasks in an autonomous and intelligent manner. In collaborative training, a main challenge is dealing with the heterogeneity in data distribution and model architectures between the cloud and client sides. Data privacy may also be a concern during collaborative training, particularly in privacy sensitive cases. Despite much progress in collaborative inference, slicing and completing a complicated task in a collective way automatically remain challenging. Furthermore, the communication costs between computing facilities might be another concern. Collective planning is a new paradigm that gains attention with the increasing study and promising progress of LLM-centric agents (LLM agents). This paradigm often involves multiple LLM agents who compete or cooperate together to complete a challenging task. It often leverages emerging capabilities such as in-context learning and chain-of-thoughts of LLMs to automatically dive a complicated task into several subtasks. By completing and assembling different subtasks, the global task can be conducted in a collaborative manner. This scheme finds diverse applications such as developing games and simulating social societies. However, it may suffer from drawbacks inherent in LLMs, including hallucination and adversarial vulnerabilities. Thus, more robust and reliable collaborative planning schemes remain to be investigated. In summary, this work surveys the large-small model collaboration techniques from the perspectives of generative foundation models, model compression, and heterogeneous model collaboration via LLM agents. This work also compares the advantages and disadvantages between international and domestic technology developments in this research realm. We conclude that, although the gaps are narrowing between domestic and advanced international studies in this area, particularly for newly emerging LLM agents, we may still lack original and major breakthroughs. Certain notable advantages of domestic progress are closely related to industrial applications due to its rich data resources from industries. Therefore, the development of domain specific LLMs is advanced. In addition, this study envisions the applications of large-small model collaboration and discusses certain key challenges and promising directions in this topic. 1) The design of efficient model architectures includes developing new model architectures that can achieve low-complexity inference speed while maintaining efficient long-sequence modeling abilities as Transformers and further improving the scalability of mixture-of-expert-based architectures. 2) Current model compression methods are mainly designed for vision models. Thus, developing techniques specifically for LLMs and large multimodal models is important to preserve

their emergent abilities during compression. 3) Existing personalization methods specially focus on discriminative models, and due attention needs to be paid for efficient personalization for generative foundation models. 4) Generative intelligence often suffers from fraudulent contents (e. g. , generated fake imagery, deepfake videos, and fake news) and different types of attacks (e. g. , adversarial attacks, the jailing breaking attacks, and the Byzantine attacks). Thus, security and trustworthy issues arise in their practical applications. Therefore, this study also advocates a deeper investigation of these emerging security threats. Then, it develops effective defenses accordingly to countermeasure these crucial issues during large-small model collaboration for empowering vertical domains more safely.

Key words: generative foundation models; model compression; large-small model collaboration; edge-cloud collaboration; generative agents; generative AI

0 引　言

人工智能(artificial intelligence, AI)技术正进入应用爆发期,成为智能社会发展的重要驱动力。大模型又称为基座模型(foundation model),是人工智能领域重要研究方向之一。大模型通常由数十亿甚至上千亿参数组成,具备强大的学习能力。在大规模任务上的实验表明:大模型在自然语言处理、视觉识别及视觉内容合成、语音合成和三维重建等领域表现了良好的通用性能。例如ChatGPT(chat generative pre-trained Transformer)等自然大语言模型可以准确理解人类指令,高效完成情感分析、多轮对话等功能;Stable Diffusion等视觉大模型可以根据自然语言指导,生成不同场景高度逼真且多样化的视觉内容。因此智能基座模型为生产生活中的复杂应用场景提供了新机遇,可有力推动数智化经济转型。

基座赋能的一个关键挑战在于基座大模型对算力资源的巨大需求。例如具有1 750亿参数的GPT-3大模型在训练阶段需要使用1 024块英伟达A100显卡并行训练大约34天时间(Narayanan等,2021);预训练的GPT-3大模型在推理阶段占用大约700 GB的空转显存,即需要至少9块A100显卡才能完成推理过程。因而基座模型只能部署于云侧向下游用户提供智能服务。然而这种直接通过云服务赋能的模式面临时延大、成本高和个性化不足等关键挑战。与此同时,小模型具有低功耗的特点使得端侧推理成为可能,实时捕捉场景信息以满足端侧个性化需求。然而小模型参数少、易拟合至大规模数据集,因而存在通用性能不足的挑战。

大小模型协同进化重在构建大小模型间双向赋能机制,融合基座大模型全局通用知识与海量端侧小模型个性化专用知识,以协同演进方式提升大小模型性能,高效完成云侧赋能与端侧推理。随着端边设备的普及和算力的提升,协同化大小模型将在智慧交通、智能安防和智能制造等重要领域发挥赋能作用。例如在无人驾驶场景下,小模型可以部署在车辆上以收集、处理实时的路段交通信息,云侧全局大模型则根据各路段累积的交通数据做优化更新,以提供更高效的交通管理方案。

大小模型协同进化是一项新兴技术,包含异构模型协同进化、模型轻量化及个性化增强等关键技术。2006年美国康奈尔大学最早提出模型压缩概念(Bucilua等,2006),美国斯坦福大学则提出面向低功耗设备的深度压缩技术(Han等,2016),通过对深度神经网络模型的剪枝、量化和霍夫曼编码,实现保持模型准确率不变的同时大幅降低模型存储容量,提升模型部署效率。2017年谷歌公司提出的联邦学习框架(McMahan等,2017)最早探索了模型协同技术,在保护各参与方数据隐私前提下利用分布式优化方法训练全局模型,解决数据孤岛问题。近年来,大小模型协同进化研究已成为产学研聚焦的热点,并被Gartner和阿里巴巴达摩院等国内外知名机构预测为革新智能计算范式的重要科技趋势。

随着以ChatGPT为代表的大模型发布以来,生成式大模型技术在计算机视觉、自然语言处理和语音等领域迎来巨大突破,给大小模型协同技术提供更广阔的赋能场景,带来新机遇也提出新挑战,促使其关键技术及应用产生新的想法,丰富基座赋能的内涵。在此背景下,理清大小模型协同进化研究脉络,梳理重要参考文献,勾画该领域完整知识体系,可有效促进该领域进展,也给相关领域研究带来启发。

1 国际研究现状

本节从基座大模型、大模型小型化以及大小模型协同方式3个方面依次展开分析介绍，本节结构如图1所示。

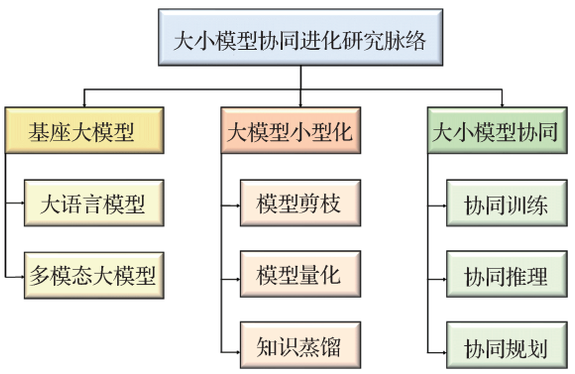


图1 大小模型协同进化研究现状结构图

Fig. 1 Layout of large and lightweight model collaboration and evolution study

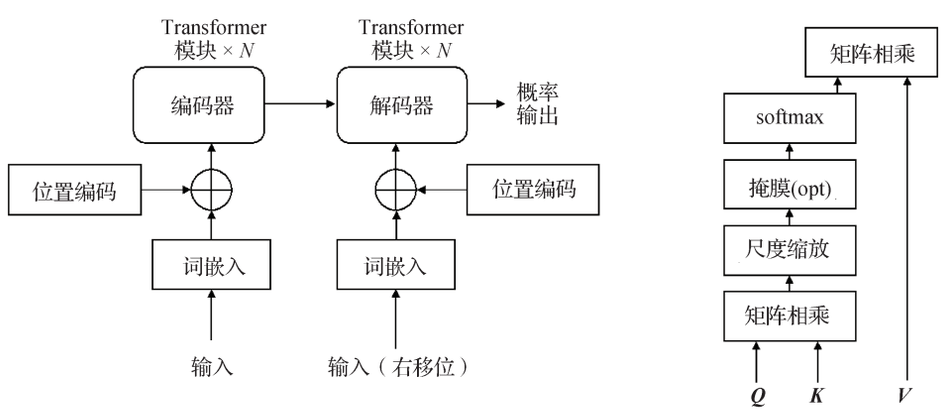
1.1 基座大模型

基座大模型是指拥有大量参数规模（如数十亿）的一类深度神经网络模型，通过在大规模数据集上做预训练，获取对语言、图像等数据广泛的理解或生成能力，利用该模型无需或仅需少量微调即可适配下游特定任务。根据处理数据不同，典型的基座大模型包括大语言模型和多模态大模型。

1.1.1 大语言模型

大语言模型（large language model，LLM）以文本为处理对象，支持情感分析、机器翻译和多轮对话等多种功能。国外在大语言模型方面取得一系列突破式进展，下面从模型架构、预训练方法和适配微调等关键技术层面加以重点介绍。

1）模型架构。国外主流的大语言模型通常基于Transformer架构，由谷歌公司于2017年提出（Vaswani等，2017）。Transformer架构的核心为自注意力机制，通过对输入词元（token）序列特征间的点积操作计算注意力权重矩阵，而后利用该权重对序列特征做加权求和作为输出特征，以捕捉全局上下文信息。标准的Transformer架构采用编码器—解码器架构，如图2（a）所示，其中编码器或解码器由多个Transformer模块构成，其核心部件为自注意力机制，如图2（b）所示。基于Transformer架构的大语言模型可分为3类：基于编码器模型（encoder-only）、基于解码器模型（decoder-only）和基于编码器—解码器混合模型（encoder-decoder）。Google公司的Devlin等人（2018）提出了BERT（bidirectional encoder representation from Transformers）模型，为首个基于编码器结构的大语言模型。Meta公司的Liu等人（2019）提出RoBERTa模型，与BERT采用相同架构，但改进了模型训练方式。OpenAI的Radford等人（2018）提出首个基于解码器结构的GPT模型，为后续ChatGPT等模型开发奠定基础。谷歌公司的Chowdhery等人（2023）利用其Pathway分布式训练技术开发了PaLM（pathways language model）模型。Meta公司的Touvron等人（2023）提出并开源了LLaMA（large language model meta AI）模型。在基于编码器—解码器类型的模型中，典型的为谷歌公司的T5模型（Raffel等，2020）和Meta公司的BART（bidirectional and autoregressive Transformers）模型（Lewis等，2020）。典型



(a) Transformer 架构　　(b) 自注意力机制

图2 Transformer模型架构及自注意力机制

Fig. 2 Architecture of Transformer and self-attention mechanism ((a) Transformer architecture; (b) self-attention mechanism)

的国外大语言模型统计信息如表1所示。与此同时，为构建更大参数规模模型以提升模型性能，谷歌公司的Shazeer等人(2017)将混合专家系统MoE(mixture of experts)思想(Jacobs等，1991)引入自然语言处理并提出稀疏化门控混合专家系统(sparsely-gated MoE)，通过上千个前向专家子模型实现模型规模千倍扩增，并利用专家子模型稀疏化激活维持模型计算复杂度。Lepikhin等人(2020)提出Gshard(giant models sharding)方法，首次将稀疏门控网络用于Transformer架构。Swtich Transformer(Fedus等，2022)继续稀疏化MoE门控网络，即每次仅激活单个专家子模型，从而进一步提升Transformer架构可扩展能力。

表1　国外典型大语言模型统计信息表

Table 1　Statistics of typical foreign large language models

模型	发布年份	机构(作者)	模型架构	参数量/B	数据量
T5-11B	2019	谷歌(Raffel等，2020)	编码器—解码器	11	1 TB
GPT-3	2020	OpenAI(Brown等，2020)	解码器	175	300 B
PaLM	2022	谷歌(Chowdhery等，2023)	解码器	540	780 B
OPT-175B	2022	Meta(Zhang等，2022)	解码器	175	180 B
BLOOM-176B	2022	BigScience(BigScience Workshop，2022)	解码器	176	366 B
LLaMA-65B	2023	Meta(Touvron等，2023)	解码器	65	1.4 TB

2)预训练方法。大语言模型常采用无监督预训练方式，以获取海量训练样本，提升预训练模型的通用语义理解或生成能力。3类模型架构对应3种预训练方法。基于编码器的大语言模型通常采用掩码语言建模(masked language modeling，MLM)方式，首先随机遮挡一定比例的输入字符用[MASK]作替换，而后利用上下文信息最大化[MASK]位置字符的预测概率，完成“完形填空”任务。利用MLM方式预训练的模型更适合于文本理解任务。在国外大语言模型中，BERT模型(Devlin等，2018)、RoBERTa模型(Liu等，2019)和ALBERTA(a lite BERT)模型(Lan等，2019)采用该类预训练方法，其中在BERT模型的预训练中还引入了下一句预测(next sentence prediction，NSP)任务。基于解码器的大语言模型利用自回归(autoregressive)语言建模方式，即给定输入序列，最大化下一词元的预测概率，更适用于文本生成任务。国外近期开发的语言模型广泛采用自回归式预训练方法，如GPT系列、PaLM模型、LLaMA模型等。基于编码器—解码器混合结构的模型采用序列到序列(sequence-to-sequence)建模方式融合了前两种预训练方式，随机遮挡一段字符序列，而后通过自回归方式还原所遮挡内容，代表性模型为T5模型和BART模型。

3)适配微调。大语言模型使用通用语料进行训练而缺乏对特定领域或任务的知识，因而需要适配微调技术以满足特定场景需求，如基于开源LLaMA大模型构建教育、法律和传媒艺术等垂直领域大模型。由于大模型参数量巨大，全量参数微调面临算力成本高和训练时间长等问题，因此针对大模型的微调适配方法常采用高效参数微调(parameter-efficient fine-tuning，PEFT)方式，即仅通过调节少量的模型参数取得与全量参数微调相近的性能。谷歌的Houlsby等人(2019)首次提出适配器微调(adapter tuning)方法，设计具有少量参数的适配器模块，将其嵌入至预训练大模型。模型微调阶段，固定原预训练模型参数不变，而只微调少量新增的适配器模块从而实现高效微调。微软的Hu等人(2022)提出低秩适配(low-rank adaptation，LoRA)方法，基于任务适配中模型权重改变量为低秩这一假设，在预训练模型的旁路引入可训练低秩矩阵，其中该矩阵可分解为一降维矩阵和一升维矩阵的乘积，原模型权重矩阵与低维矩阵之和作为微调后模型权重矩阵。美国佐治亚理工大学的Zhang等人(2023b)提出自适应低秩适配(AdaLoRA)方法，采用奇异值分解方式对旁路可训练矩阵参数化，依据重要性不同裁剪冗余奇异值，实现运算加速。第2类适配微调方法寻求对任务做高效嵌入。美国斯坦福大学的Li和Liang(2021)提出前缀微调(prefix tuning)方法，在输入序列前添加任务相关的虚拟词元(virtual tokens)作为前缀，微调阶段仅更新该前缀部

分的参数。谷歌的Lester等人(2021)提出提示微调(prompt tuning)方法,即仅在输入层添加任务相关的可训练提示嵌入,可视为前缀微调方法的简化形式。OpenAI的Ouyang等人(2022)提出指令微调方法使得大模型理解并遵循人类指令完成任务,在零样本等情形下具有较强泛化能力,同时通过强化学习人类反馈(reinforcement learning human feedback, RLHF)方式实现与人类偏好对齐。

1.1.2 多模态大模型

多模态大模型融合文本、图像、视频和音频等多模态信息,通过对多种数据模态的联合训练实现大模型的跨模态理解与生成。依据任务侧重点不同,多模态大模型可分为3类:面向理解任务的多模态大模型、面向生成任务的多模态大模型和统一理解及生成任务的多模态大模型。

面向理解任务的多模态大模型常采用基于Transformer架构,在特征空间对齐不同模态的表达。谷歌的Sun等人(2019)提出VideoBERT模型,将BERT框架扩展至文本和视频多模态数据,采用预测掩码的预训练方式学习二者联合概率分布。美国佐治亚理工大学的Lu等人(2019b)提出ViLBERT模型,采用两独立编码器网络分别提取文本特征和图像特征,并通过共同注意力(co-attention)层做多模态交互。OpenAI的Radford等人(2021)提出对比语言—图像预训练(contrastive language-image pretraining,CLIP)模型。如图3所示,CLIP模型首先通过文本编码器(text encoder)和图像编码器(image encoder)分别对文本—图像模态做特征编码,随后对4亿对图像—文本对做对比学习,即最大化正样本相似度(矩阵对角线位置元素)同时最小化负样本相似度(矩阵非对角线位置元素),从而提升文本和图像特征表达的相关性,所学特征在零样本图像识别任务上表现出强大的泛化能力。

面向生成任务的多模态大模型以文本提示为输入合成图像、视频和音频等模态内容。OpenAI的Ramesh等人(2021)提出DALL-E模型,该模型基于变分自编码器(variational autoencoder, VAE)和Transformer架构,即利用前者得到图像的离散隐空间表达,而后利用后者学习文本到图像的映射关系,最后通过CLIP模型选择质量最佳的生成样本。随后,Ramesh等人(2021)提出DALL-E2模型,该模型利用CLIP模型实现图像文本一致性,同时基于扩散

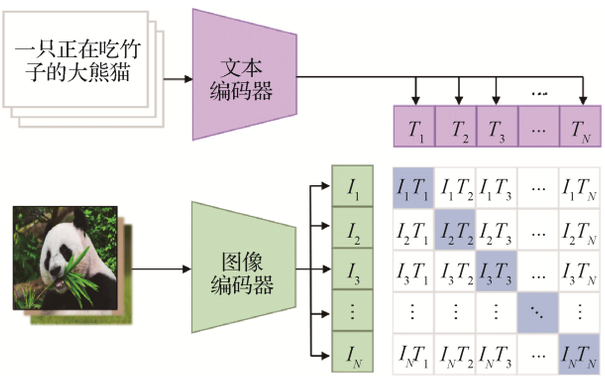


图3 基于对比学习的CLIP模型(Radford等,2021)

Fig. 3 Illustration of CLIP model based on contrastive learning (Radford et al., 2021)

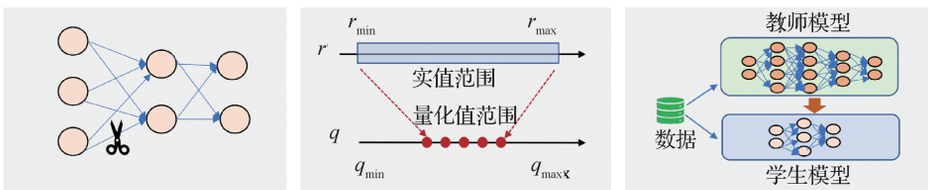
模型(diffusion model)(Ho等,2020)解码器将图像嵌入(embedding)反转为图像。谷歌的Saharia等人(2022)在同时期提出Imagen模型,发现通用的预训练大语言模型(如T5模型)对以文生图多模态模型有重要作用。为加快扩散生成模型的训练和推理速度,Rombach等人(2022)提出隐空间扩散模型(latent diffusion model,LDM),利用VAE模型的编码器将图像映射至低维隐空间,在该空间完成扩散过程后通过VAE的解码器映射回图像空间,其中文本等信息作为控制条件嵌入至扩散模型。

第3类多模态大模型则统一理解和生成任务。美国密歇根大学的Zhou等人(2020)提出VLP(vision-language pre-training)模型,采用对编码器和解码器多层Transformer参数共享的方式实现结构统一,同时学习更通用的多模态表达,首次在视觉语言理解和生成任务上同时取得当时最佳性能。Salesforce公司的Li等人(2022a)提出BLIP(bootstrapping language-image pre-training)模型,该模型利用编码器—解码器混合架构这一灵活结构,通过文本图像对的对比学习、图文匹配以及条件文本生成等方式实现多模态理解与生成。BLIP-2模型(Li等,2023c)则通过Q-former模型将单模态的预训练视觉模型和大语言模型联系起来,分成表征学习和生成学习两阶段进行模型训练,在多类视觉语言任务中取得突出表现。

1.2 大模型小型化

深度学习模型的规模呈现两极化趋势,其中一个趋势朝着超大规模方向发展;而另一重要趋势则向着轻量化小模型方向发展(Yao等,2023)。在物联网、端云协同等场景下,小模型因其功耗低、运算

少、响应快的特点而备受关注。从大模型到小模型一般有两类方式：一类是通过对大模型做剪枝、量化等操作对大模型进行压缩；另一类是通过知识蒸馏方法将大模型的知识传递至独立小模型（Gou 等，2021），辅助小模型训练、提升其性能。大模型小型化过程和主流方法如图 4 所示。



(a) 模型剪枝	(b) 模型量化	(c) 知识蒸馏
非结构化剪枝 Han 等人（2015） Gordon 等人（2020） Frantar 和 Alistarh（2023） Sung 等人（2023） Liao 等人（2023）	训练后量化 Denton 等人（2014） Gong等人（2015） Banner等人（2019） Dettmers等人（2022） Xiao等人（2023）	基于对元知识蒸馏 Hinton 等人（2015） Allen-Zhu 和Li（2020）
		基于特征知识蒸馏 Romero 等人（2015） Ji 等人（2021）
结构化剪枝 Li 等人（2017） Luo 等人（2017） Li 等人（2022c） Liu 等人（2021） Ma 等人（2023a）	量化感知训练 Jacob 等人（2018） Stock 等人（2022） Liu 等人（2023b） Li 等人（2023d） Chen 等人（2023c）	基于关系知识蒸馏 Passalis 和 Tefas（2018） Park 等人（2019）
		基于涌现能力知识蒸馏 Brown 等人（2020） Huang 等人（2022） Ho 等人（2023）

图 4　大模型小型化典型方法

Fig. 4　Typical methods in model compression ((a) model pruning; (b) model quantization; (c) knowledge distillation)

1.2.1　模型剪枝

神经网络模型常表现过参数化（over-parameterization）现象，模型剪枝技术通过移除模型中的冗余参数以减小模型规模和运算量，同时保证其性能不变。模型剪枝方法可分为非结构化剪枝和结构化剪枝。

主流的非结构化剪枝方法包括对神经元连接的剪枝或对卷积核的剪枝。美国斯坦福大学的 Han 等人（2015）最早提出“训练—剪枝—再训练”（train-prune-retrain）三阶段非结构剪枝方法，即在完成模型的初始训练后，移除权重值小于某阈值的非重要神经连接，从而将稠密连接层转化为稀疏连接层，随后重新训练该稀疏化网络以恢复模型性能。美国约翰霍普金斯大学的 Gordon 等人（2020）进一步研究了对 BERT 模型做不同程度的剪枝时对下游任务的影响。澳大利亚科学与技术研究院的 Frantar 和 Alistarh（2023）提出 SparseGPT 剪枝方法，将针对 GPT 系列模型的剪枝建模为超大规模的稀疏回归问题，该方法无需再训练，在模型性能近乎不变时可移除 50% 以上的模型参数。针对多模态大模型，Sung 等人（2023）提出“由粗及精”（coarse-to-fine）的两阶段剪枝方法，利用一阶梯度近似参数的全局重要性指标，以自适应方式对多模态大模型逐层稀疏化，高效完成剪枝过程。

结构化剪枝方法则移除模型中的神经网络通道或滤波器组。相比于非结构化剪枝，结构化剪枝方法通常会牺牲模型性能，但在硬件加速上效果更好，因而也得到广泛应用。美国马里兰大学的 Li 等人（2017）提出一次性剪枝和再训练策略以移除权重较小的滤波器组。瑞士联邦理工学院的 Li 等人（2022c）分析并简化了基于随机搜索的剪枝方法以确定网络通道的设置。新加坡国立大学的 Ma 等人（2023a）提出 LLM-Pruner 方法，该方法模型梯度信息选择性地移除非关键结构，成为首个针对大模型的结构化剪枝方法。

1.2.2　模型量化

模型量化是将浮点型模型转化为定点型模型的一类技术，采用低精度数据类型以压缩参数量，同时降低内存需求，提升推理速度。针对量化操作带来的潜在性能损失，模型量化研究在大幅压缩模型参数的同时保持其性能相当。根据量化发生阶段不同，现有方法可分为两大类：训练后量化（post-training quantization，PTQ）和量化感知训练（quantization- aware training，QAT）。

1）训练后量化。美国纽约大学的Denton等人（2014）研究了卷积层的低秩结构，利用矩阵分解和聚类方法对模型提速。Facebook的Gong等人（2015）提出基于矢量量化（vector quantization）的方法，揭示了k-means聚类方法在模型量化中的有效性。Banner等人（2019）研发了首个面向低比特（4-比特）量化的实用方法，提出了针对整形量化的可解析限幅、通道比特分配等有效策略。美国麻省理工大学的Xiao等人（2023）针对大语言模型量化时特征激活的离群值现象（Dettmers等人，2022）提出SmoothQuant方法，通过通道尺度变换对大模型通道幅度值做平滑，从而易于模型量化。

2）量化感知训练。谷歌的Jacob等人（2018）在模型的前向训练中考虑舍入误差以模拟量化模型的推理过程。Facebook的Stock等人（2022）提出Quant-Noise方法，该方法在每次网络前向训练时随机性地量化一部分参数子集，使得模型在推理阶段对量化误差更稳健。Meta的Liu等人（2023b）提出LLM-QAT方法，该方法对模型权重、特征激活和键—值高速缓存（key-value cache）做量化，提升量化模型的吞吐率。Li等人（2023d）提出LoftQ以对大模型LoRA微调做量化，采用了联合优化低秩近似和量化的方法并取得较好性能。针对量化感知训练过程中模型的灾难性遗忘问题，Chen等人（2023c）提出LifeQuant，该方法对量化空间做正则化以降低量化搜索空间的漂移效应，同时结合数据重放（replay）思路缓解灾难性遗忘难题。

1.2.3 知识蒸馏

知识蒸馏的核心思想是通过引导轻量化小模型模仿性能更强大的大模型的行为，将大模型的知识迁移至小模型。根据小模型所模仿的不同行为类型，知识蒸馏可分为4类：基于对元（logits）的知识蒸馏、基于特征的知识蒸馏、基于关系的知识蒸馏以及面向大模型涌现能力的知识蒸馏。

1）基于对元的知识蒸馏。谷歌的Hinton等人（2015）认为大模型（教师模型）的对元（即分类器最后一个全连接层的输出）可较真值标签为小模型（学生模型）提供更多的“黑暗信息”（dark knowledge），因而提出通过最小化大小模型的软概率（对元做某温度下的softmax变换）间Kullback-Leibler（KL）散度实现知识迁移。该方法开启了知识蒸馏领域的研究。尽管该方法对小模型有明显提升效果，其工作机理却难以理解。近期，Meta的Allen-Zhu和Li（2020）提出多视角思路，从数据的多视角结构方面对集成模型对元蒸馏原理给出理论解释。

2）基于特征的知识蒸馏。西班牙巴塞罗那大学的Romero等人（2015）提出FitNet，首次将教师模型的中间特征作为辅助监督信息引入知识蒸馏。Ahn等人（2019）提出基于变分信息蒸馏的VID（variational information distillation）方法，通过最大化教师及学生模型在特征空间的互信息，实现高效蒸馏。针对教师—学生网络特征层选取问题，Ji等人（2021）提出基于注意力机制的特征选择方法。

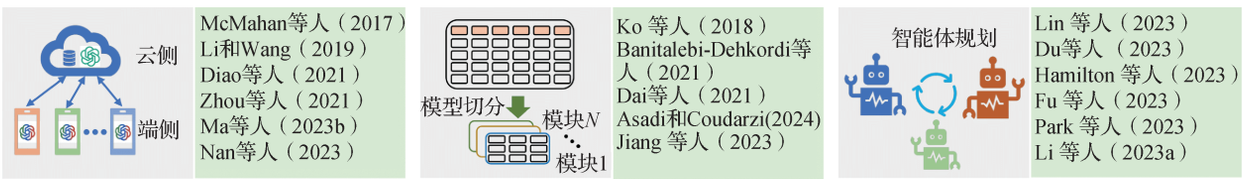
3）基于关系的知识蒸馏。Passalis和Tefas（2018）提出了一种概率知识转移（probabilistic knowledge transfer，PKT）方法，该方法使用成对的相邻样本建立关系矩阵的概率表达式，然后通过最小化条件概率分布的Kullback-Leibler散度以使学生模型与老师模型保持一致。Park等人（2019）提出关系知识蒸馏（relational knowledge distillation，RKD）方法，该方法通过基于网络嵌入的二元组距离关系和三元组角度关系对训练样本之间的相互作用进行建模。Wang等人（2023b）提出SSD-KD（self-supervised diverse knowledge distillation），通过同时利用相邻样本对之间的关系和样本不同通道之间的关系，更好地实现关系的建模与迁移。

4）基于涌现能力的知识蒸馏。大语言模型在参数量超过一定量级后能力骤升，表现出小模型不具有的能力，常称做大模型的涌现能力，如上下文学习能力（in-context learning，ICL）（Brown等，2020；Zhao等，2021）和知识链能力（chain-of-thought，CoT）。相应地，通过蒸馏方式将该涌现能力迁移至小模型的方法称做基于涌现能力的蒸馏（Zhu等，2023a）。美国哥伦比亚大学的Huang等人（2022）提出元上下文微调（meta in-context tuning，Meta-ICT）和多任务上下文微调（multitask in-context tuning，Multitask-ICT）方法，通过联合上下文学习和语言建模方式，提升小模型的上下文学习能力。美国加州大学圣巴拉分校的Li等人（2022b）在多任务学习场景下将大模型的解释能力迁移给小的推理模型。韩国科学技术院的Ho等人（2023）提出Fine-tune-CoT（fine-tune chain-of-thought）方法，利用教师大模型生成的推理样本对学生小模型做微调，通过实验揭示了该方法可显著增

强小模型的推理能力。

1.3 大小模型协同

根据协作目的不同,大小模型协同方式可分为:协同训练(collaborative training)、协同推理(collaborative inference)和协同规划(collaborative planning)3种类型,协作方式及典型方法如图5所示。



(a) 协同训练　(b) 协同推理　(c) 协同规划

图5　大小模型协同方式及典型方法

Fig. 5　Types of model collaborations and typical methods

((a) collaborative training; (b) collaborative inference; (c) collaborative planning)

1.3.1 协同训练

基座大模型的优势为其通用性能,轻量化小模型可部署于本地以捕捉动态场景数据从而具有更优异的个性化能力。模型间通过知识蒸馏、特征传输等方式完成协同训练,实现优势互补。

美国微软研究院的Lu等人(2019a)为了实现不同设备模型的知识共享和模型的持续更新,提出了端侧模型与云侧模型间的双向知识蒸馏方法,为跨设备的模型协同训练提供了有效的解决方案。接着,日本滋贺大学的Zhou等人(2021)提出面向物联网系统中实时多目标监控任务的的端—边—云模型协同训练方法,其中端侧进行数据压缩和特征提取,边缘侧完成多级特征融合和轻量化模型训练,云侧处理智能应用,实现了有限计算资源约束下模型的轻量训练和特征学习。为解决数据漂移问题,微软和美国加州大学伯克利分校的Bhardwaj等人(2022)致力于边缘侧视频分析模型的持续学习问题,设计了资源调度器和性能估算器,其中调度器优先去重新训练特征变化最大的视频流模型,性能估算器用于观察重新训练窗中样本的准确率,性能显著优于当时的基线方法。Ma等人(2023b)研究基于分层内存重播的持续学习设计空间,使得系统可动态地配置持续学习训练资源。新加坡南洋理工大学的Nan等人(2024)提出端云协同学习架构,将边缘计算与云侧持续学习相结合:轻量级小模型部署于边缘侧,并利用云端大模型对其进行持续训练,以提升高效的视频分析功能,同时减少边缘端和云端之间的数据传输量。为促进模型间知识传输,美国佐治亚理工学院的Daga等人(2023)提出CLUE(collaborative learning neural nets)学习框架以动态提取辅助节点神经网络中的重要参数,并使用基于多模型增强的方法来提高目标节点的预测性能。Dong等人(2023)提出EdgeMove端云协同训练方案。该方法通过自适应探测附近边缘服务器的训练性能,构建带有近似模型分区的训练流水线(pipeline),并主动适应训练流水线、用户移动和系统动态。这一方案通过端云之间协同,实现更加智能和适应性的模型训练。

联邦学习也是协同训练范式的一个典型示例。联邦学习最早由谷歌公司提出(McMahan等,2017),旨在利用端侧或边缘侧进行模型的本地训练,而后由中心节点完成模型聚合或分发的一类分布式机器学习方法。隐私保护是联邦学习的重要目的。在联邦学习框架下,训练数据始终留在本地、对中心节点或其他节点均不可见,因此联邦学习在节点间协同训练的同时,可以较好地保护用户隐私和数据安全。联邦学习常假定各节点有相同的模型结构以降低模型间协同难度,可视做大小模型协同的特例。近期,联邦学习的异质性也日益受到关注,具体表现为统计异质性(Bao和Guo,2022)、模型异质性(Zhou等,2022)、设备异质性(Li等,2020)3个方面。其中,在模型异质性方面,端侧可能有不同的任务和特定的要求,因此每个客户端可能需要独立设计其本地模型,从而在异质参与者之间产生知识转移障碍,导致无法应用常用的模型聚合或梯度操作。针对该挑战,美国哈佛大学的Li和Wang(2019)提出基于迁移学习和模型蒸馏的方法,每个节点拥有私有数据集和独特的模型架构,并利用知识蒸馏作为“翻译器”

将节点模型知识转化为标准格式后传输给中心节点。Yu 等人(2020)提出通过在模型协同过程中建立结构和信息的对齐的方案,设计基于特征的调节方法,根据本地模型的数据和任务分布调整模型架构。在联邦学习框架下,美国密歇根州立大学的 Zhu 等人(2021)提出基于无数据知识蒸馏(data-free knowledge distillation)的方法,通过中心节点学习轻量级生成器,以无数据方式集成用户信息,随后将其发送至端侧,端侧模型利用所学知识调整本地训练。美国杜克大学的 Diao 等人(2021)提出 HeteroFL(heterogeneous federated learning)方法,通过改进批归一化(batch normalization)处理、交叉熵掩膜(masking cross-entropy)等技术以支持异构模型的协同训练。瑞士洛桑联邦理工学院(EPFL)的 Afonin 和 Karimireddy(2022)提出联邦核岭回归的理论框架以捕捉模型异构性和数据异构性,并分析了基于知识蒸馏的方案性能下降的原因。

1.3.2　协同推理

协同推理方式常对模型做模块切分,并根据算力约束将子模块部署于云侧或端侧,通过子模块的协同完成特定任务。

韩国成均馆大学的 Ko 等人(2018)提出了一种在边缘和主机平台之间划分深度神经网络推理任务的方法。利用分区网络的微调,对中间层的特征进行编码,通过实验证明了这种方法能显著提高边缘的能效和吞吐量,从而有效地解决了分布式推理问题。Banitalebi-Dehkordi 等人(2021)提出了协同边缘云人工智能的通用框架 Auto-Split。这一框架对边缘模型进行训练后量化,并为边缘模型层分配位宽(bit-widths),具有良好的安全性、确定性和架构灵活性。巴西里约热内卢联邦大学的 Pacheco 等人(2021)提出了专家分支(expert side branches)的方法,通过在特定图像失真类型上进行训练,提高了对图像失真的推理鲁棒性,并改善了卸载决策。Dai 等人(2021)从卷积神经网络的视角研究了图像分类模型的协同推理方法,提出了 CINET(collaborative-aware networks)架构,其中端侧子模型输出图像的一个小而相关的区域,而云侧子模型对该图像区域进行分类,提高系统协同推理综合性能。加拿大蒙特利尔大学的 Madan 等人(2021)提出基于知识分解的模块化学习框架,将参数分为模块参数和注意力参数,前者根据当前任务动态更新,后者作为稳定的元参数,能够在分布变化的环境中实现快速适应和系统泛化。美国微软研究院的 Padmanabhan 等人(2021)研究了通过模型合并实现边缘侧内存高效管理方法,通过贪婪算法确定边缘模型所共享的共同层,利用边缘侧提示在云侧完成模型合并,该操作可显著地节省内存,为实现高效的协同管理提供了新思路。德国慕尼黑工业大学的 Asadi 和 Goudarzi(2024)提出基于模型集成的方法,生成主模型的不同变体并部署于不同的节点,通过模型集成方式提升推理性能。近期,澳大利亚悉尼科技大学的 Jiang 等人(2023)研究端云场景下自然语言处理架构的设计方法,提出在端侧部署 Transformer 的编码器、云侧部署解码器,其中端侧编码器用于减小隐藏状态的序列长度,同时保持交错的多头自注意和位置编码前馈网络的总体结构,有助于提升端云协同场景下模型间协同推理性能。

1.3.3　协同规划

在实际应用场景中,单个模型难以完成复杂任务的规划,因而常需要模型间的协同规划。随着大模型认知能力的涌现,生成式智能体(generative agent)成为大模型协同规划的重要研究领域。

美国艾伦人工智能研究所的 Lin 等人(2023)提出了 SWIFTSAGE(swift and sage)这一全新的生成智能体框架,旨在应用于复杂的交互式推理规划。该框架由两个关键模块组成:SWIFT 模块模拟快速和直觉式思维,而 SAGE 模块则模拟慢速而缜密的思维,如子目标规划。这两个模块通过协同作用,利用多个模型的认知能力来完成复杂的规划任务。与此同时,美国麻省理工大学的 Du 等人(2023)采用多智能体辩论来提升大模型在事实性问答和推理方面的能力。他们验证了这种方法相较于单一智能体情境下的基准方法(如推理链和自反思等)的优越性。加拿大麦吉尔大学的 Hamilton(2023)对多个协同式语言模型在社会模拟方面的进展进行了研究,以确定是否促进生成用于法院的简单有效的行为模型。结果显示,这些模型能够在准确性方面达到优于随机的水平。英国爱丁堡大学的 Fu 等人(2023)则通过研究多个大型语言模型,探讨通过角色扮演和从人工智能反馈中学习的可能性。近期美国伊利诺伊大学厄巴纳—香槟分校大学的 Wang 等人(2023c)提出了个人表演提升(solo performance prompting ,SPP)方法,利用单个大语言模型作为认知协同者,

通过动态识别角色和参与多轮自我协作来解决任务。验证结果显示，在无需外部资源的情况下，SPP显著提高了大语言模型的知识获取和规划能力。

韩国科学技术院Park等人（2023）提出了ChoiceMates系统，旨在有效管理和引导多个智能体之间的对话。该系统允许智能体在对话中灵活加入，并通过彼此对话来揭示每个智能体的偏好。在多轮交流中，模型能够充分挖掘在线决策相关信息，以更好地适应不确定性环境。沙特阿拉伯王国KAUST大学的Li等人（2023a）引入了一种基于角色扮演的合作代理框架，极大程度上减少了人为干预的需求，扩展了智能体间自主协同能力。英国剑桥大学的Li等人（2023e）提出协同式生成智能体（collaborative generative agents）方法MetaAgents，通过模拟招聘会场景研究智能体在协作方面的能力。Curai健康的Nair等人（2023）引入了可对话表决智能体（dialog-enabled resolving agents ，DERA），通过对话表决或改进输出来提高自然语言任务的性能。这一方法强调了对话在提高性能方面的关键作用。美国波士顿大学的Pham等人（2023）提出基于嵌入表示的模型间通信协议，即移除词元采样，而通过Transformer输出的嵌入特征进行模型间交互，该方式在多类推理规划任务上表现出优异性能，为模型之间的高效交流提供了一种有效途径。

在具体应用方面，美国哥伦比亚大学的Zhao等人（2023a）提出面向机器人的协同方法（robot collaboration，RoCo），通过多个大语言模型进行任务讨论和推理，产生子任务规划，实现多机器人协同。美国宾夕法尼亚州立大学的Wu等人（2023）提出了一种利用多智能体对话方式促进多领域大模型应用程序开发的方法，开发了智能体设计的通用框架AutoGen，该框架具有统一的对话接口和自动回复机制，从而提升了智能体的交互性和泛化性。Chen等人（2023a）研究了面向游戏开发的多智能体框架，其中单个智能体维持私有记忆器且可访问多智能体的公共讨论信息，同时设计双重合作和分层方法及结合内部词汇表以降低大模型的“幻觉”问题。

2 国内研究进展

国内在大小模型协同进化领域快速跟进，下面分别从基座大模型、小模型及大小模型协同3个层面加以具体阐述。

2.1 基座大模型

自2023年以来，国产大模型如雨后春笋般涌现，引领着人工智能前沿技术的发展浪潮。据不完全统计，至2023年10月，国产大模型数量已超过100个，包括通用基座大模型与教育、司法、医疗、金融、科学、军事等垂直领域大模型。

2.1.1 大语言模型

在大模型架构、预训练方法等技术方面，主流的国产大语言模型与国外模型基本保持一致。例如，阿里巴巴的通义千问大模型、上海人工智能实验室的书生·浦语大模型基于OpenAI的GPT再次开发，百川智能的Baichuan大模型、科大讯飞的星火认知大模型基于Meta公司的LLaMA模型，均采用了基于解码器（decoder-only）的架构。百度的文心一言（Sun等，2021）融合自回归网络和自编码网络，在预训练模型中引入知识图谱，利用知识增强方法提升大模型的推理能力。清华大学和智谱AI提出通用语言模型（general language model，GLM）系列模型（Du等，2022），结合BERT和GPT的优势架构，采用自回归填空作为预训练任务，同时采用旋转位置编码和统一预训练目标以提升模型泛化能力。典型的国内大语言模型统计信息如表2所示。

在大模型的适配微调方面，国内研究人员也提出一系列改进方法。清华大学的Liu等人（2021）提出P-tuning方法，将可训练的连续的提示嵌入与离散的提示信息进行拼接，而后输入至大模型以降低离散提示带来的不稳定性。随后，Liu等人（2022c）提出P-tuning的改进版P-tuning v2，通过在网络的多层加入提示以改善参数高效微调性能。阿里巴巴的He等人（2021）对比了适配器微调方法和常规微调方法，结果表明适配器微调方法的有效性在于可缓解模型遗忘问题。上海人工智能实验室的Zhang等人（2023c）针对LLaMA模型提出适配器，设计了具有零门控及零初始化的注意力机制，将新的提示引导以自适应方式注入LLaMA模型，实现快速适配微调。清华大学和智谱AI的Ding等人（2022b，2023）全面研究了针对预训练大模型的高效微调方法，提出新的分类方法并从优化及最优控制的角度给出统一的理论解释。

表 2　国内典型大语言模型统计信息表

Table 2　Statistics of typical domestic large language models

模型	发布年份	机构/作者	模型架构	参数量/B	数据量/TB
ERNIE 3.0	2021	百度公司(Sun等,2021)	自回归和自编码	10	4
ChatGLM-130B	2022	清华大学,智谱AI(Zeng等,2022)	通用语言模型	130	1
Baichuan-7B	2023	百川智能(Yang等,2023)	解码器	7	1.2
InternLM-20B	2023	上海人工智能实验室,商汤科技,香港中文大学,复旦大学(InternLM Team,2023)	编码器—解码器	20	2.3

2.1.2　多模态大模型

在面向理解任务的国产多模态大模型中,Su等人(2020)提出用于视觉和语言特征表示的VL-BERT(visual-linguistic BERT)模型,该模型采用Transformer架构,其输入元素为句子中的词元或图像的感兴趣区域(region of interest,RoI),通过预测随机屏蔽的单词或感兴趣区域完成VL-BERT模型的预训练。百度公司的Zhu和Yang(2020)提出ActBERT,采用自监督学习方式完成视频和文本特征的联合学习,通过对整体和局部视觉信息建模构建更细化的视觉与语言之间的关联。华为诺亚实验室的Yao等人(2021a)提出细粒度交互式语言—图像预训练方法FILIP(fine-grained interactive language-image pre-training),通过跨模态的后期交互机制实现更精细层次的对齐。阿里巴巴的Lin等人(2021)提出M6模型,为具有1 000亿参数的多模态到多模态多任务超大型Transformer模型,在多个下游任务中表现优异。Gu等人(2022)构建了大规模跨模态中文数据集Wukong,包含了1亿个中文图像—文本对,为中文多模态模型的预训练提供数据支持。

在面向生成任务的多模态大模型中,清华大学的Ding等人(2021)提出并开源CogView模型,该模型采用与DALL-E类似的结构,通过在3 000万个高质量的中文—图像对进行联合生成训练,取得优于DALL-E和GAN等模型的性能。微软亚洲研究院的Wu等人(2022)提出名为NÜWA(neural visual world creation)的统一多模态预训练模型,设计了3D Transformer编码器—解码器框架以处理不同模态的数据,该模型可广泛用于视觉内容合成、视频预测等场景。南京邮电大学的Tao等人(2023)提出GALIP(generative adversarial CLIPs)模型,结合预训练CLIP模型和对抗生成网络实现文本到图像的快速和可控合成。

在统一理解和生成任务方面,中国科学院自动化研究所发布了紫东太初大模型,该模型同时具备跨模态理解和生成能力,为首个视觉—文本—语音三模态预训练模型。香港中文大学的Li等人(2023b)提出Uni-Perceiver v2,将图像编码为包含语义、边界框和分割掩膜表示的通用区域提议,使其定位建模有更强的表达力和灵活性,设定不同任务为统一的最大似然估计问题,通过联合学习的方式实现了通用任务的适应。

2.2　大模型小型化

国内在大模型小型化技术方面也取得一系列进展和重要发现,下面分别从剪枝、量化和知识蒸馏3个方面进行介绍。

2.2.1　模型剪枝

西安交通大学的He等人(2017)提出基于LASSO(least absolute shrinkage and selection operator)回归的通道选择方法,并利用最小二乘重构完成对模型每层神经元的修剪,进而将该方法推广至多层和多分支情形。与此同时,南京大学的Luo等人(2017)提出ThiNet(thin net),将滤波器剪枝问题建模为优化问题,并指出在当前层滤波器剪枝时,应基于下一层滤波器的统计信息进行决策。清华大学的Wang等人(2020)则深入研究了直接从随机初始化的权重中剪枝的可行性,提出一种新颖的从零开始的模型剪枝方法。研究结果表明,通过随机初始化进行剪枝能够得到更多样化的剪枝结构,其中包括潜在性能更好的模型。近期,华南理工大学的Liu等人(2022a)提出判别感知通道剪枝方法,该方法引入了判别感知损失以增强模型中间层的判别能力。同时研究了判别感知卷积核剪枝,通过去除冗余卷积核进一步压缩深度网络。国内在模型剪枝方面的研究为深度学习模型的精简提供了多样化方法。

2.2.2 模型量化

中国科学技术大学的Yang等人(2019)提出了一种模型量化方法,其核心是采用可微非线性映射函数。这一方法中,非线性函数由多个具有可学习偏置和尺度的sigmoid函数组合而成,使得整个量化过程成为可端到端学习的。这种设计使得他们的方法成为一种通用解决方案,可适用于任意位数的量化过程。南方科技大学的Xu等人(2020b)的研究聚焦于基于生成模型的无数据量化方法。他们提出了知识匹配生成器,通过合成虚假数据,利用预训练模型的分类边界知识和分布信息,实现了对模型的无数据量化(data-free quantization),为量化提供了一种新的途径,摆脱了对真实数据的依赖。华中科技大学的Liu等人(2023a)提出PD-Quant方法,其特点是通过利用量化前后模型预测之间的差异信息来确定量化参数。同时,调整激活神经元的分布,以减轻在少量校准数据时的过拟合问题。中国科学院Li等人(2023f)研究面向视觉Transformer(visual Transformer, ViT)的训练后量化(post-training quantization)方法RepQ-ViT,该方法将量化和推理过程解耦,采用了复杂的量化器和基于尺度重参数化的简化量化器,有效地缓解了通道间激活变化大、存在幂律特征激活等极端分布情形下的量化问题。这些新的研究思路进一步改善了模型量化性能。

2.2.3 知识蒸馏

香港中文大学的Xu等人(2020a)提出了一种基于自监督学习的知识蒸馏方法SSKD(self-supervised knowledge distillation),通过自监督训练教师模型,使其概率预测包含更丰富的知识。在这个方法中,将自监督信号间响应的相似性作为辅助任务,可以显著增强知识蒸馏的性能。浙江大学的Chen等人(2020)提出了一种在线知识蒸馏方法,由多样化同行协同实现。首先在多样化同行之间完成基础蒸馏,然后将同行的集成预测进一步蒸馏为组领导者,该方法适用于教师模型容量不足的情况。香港中文大学的Chen等人(2021a)则研究了教师网络和学生网络的不同层级连接通路的重要性,利用教师网络中的低级特征监督学生网络的深层特征,显著提升蒸馏效果。旷世科技的Zhao等人(2022)提出解耦知识蒸馏方法DKD(decoupled knowledge distillation),将常规知识蒸馏损失函数分解为目标类别的二元对元蒸馏和非目标类别的多类别对元蒸馏,以提升知识迁移性能。南开大学的Li等人(2023h)研究了基于课程学习的知识蒸馏方法,设计从易到难的课程设置,通过对抗方式逐渐增加关于温度的蒸馏损失,以使学生模型学到更好的特征表示。

2.3 大小模型协同

国内在大小模型协同方面也取得诸多研究进展,本节从协同训练、协同推理和协同规划3种类型展开分析介绍。

2.3.1 协同训练

北京邮电大学的Ding等人(2022a)提出一种云边协同框架,其中在边缘服务器上部署了一个浅层小模型,用于提供快速的认知服务,同时在云服务器上部署了一个深层模型,用于协助训练小模型以提高其性能。阿里巴巴的Yao等人(2021a)使用元学习的MetaPatch方法在设备端实现了个性化的模型生成,根据云端模型快速生成定制的模型。他们还提出了MoMoDistill方法,在云端使用模型蒸馏,通过个性化设备模型来更新云端模型。实验证实了该框架在云端和设备端的有效性,特别是在处理长尾用户方面的优势。华东师范大学的Chen等人(2021b)提出一种基于慢学习—快学习(slow-fast)机制的移动—云协同推荐方法。根据实际交互频率,将云端模型和移动端模型分别视为慢组件和快组件,通过交流先验知识提升了捕捉用户兴趣的能力。上海交通大学的Yan等人(2022)提出领域自适应范式下的端云协同学习框架(model personalization domain adaptation ,MPDA),从云端的全局数据池中检索与每个用户的本地数据集相似的样本,以增强用户的本地数据并训练个性化的模型。北京邮电大学的Shao等人(2022)提出基于云—边协作的电力物联网场景感知机制,支持局部场景信息的边缘感知和全局场景云综合,有效降低感知处理延迟和模型训练时间,同时提高对高动态场景的感知模型的适应性。浙江大学的Li等人(2023g)提出端云协同知识传递框架(edge-cloud collaborative knowledge transfer, ECCT),利用边缘侧特征和云侧特征,通过共享特征嵌入和预测逻辑实现两者之间的双向知识传递。Lyu等人(2023)借鉴超网络(HyperNetwork)的思想,在云侧部署参数生成网络,使得端模型无需在端上或云上进行任何形式的训练,只需要将数据发送到云上就可以获得能更好泛化到此数据分布的模型参数,从而完成快速的端模型自适应。

2.3.2　协同推理

清华大学的Li等人(2018)提出了一种联合精度和延迟感知的执行框架,通过将深度神经网络解耦,使其一部分在边缘设备上运行,另一部分在传统云中运行,该方法显著降低了推理时延。大连理工大学的Xu等人(2021)从多层面进行协同优化,包括随机舍入技术、基于学习的动态推理卸载以及卸载延迟方法,最大限度地减少移动设备和云计算的能耗,并最大化了允许处理的请求数量,实现推理算力资源的高效利用。上海交通大学的Niu等人(2020)提出对大模型做子模型拆分方法,使每个终端只需使用其本地特征所对应的局部模型,即可参与端云协同学习过程,从而摆脱对完整模型的依赖,提高了大小模型协同学习效率。Zhu等人(2023b)设计了一个端侧单模态小模型和云侧多模态大模型协同学习框架,通过卸载轻量单模态模型至端侧执行预推理,仅在结果置信度低时上传单模态特征,以协助云侧多模态推理。

在系统平台层面,浙江大学、阿里巴巴和上海交通大学等团队合作构建了首个覆盖研发期、部署期和运行时的产业级端云协同智能计算系统Walle(Lyu等,2022),向下兼容了端侧和云侧软硬件的差异性,向上支持了30多个移动应用上300多种学习任务每天超1 000亿次调用。该研究成果支撑了以阿里巴巴搜索推荐和直播、智能机器人巡检等端云协同智能产业应用,取得显著的经济效益和社会效益。

2.3.3　协同规划

清华大学的Liang等人(2023)提出多个大模型智能体之间的辩论(multi-agent debate,MAD)框架,旨在激发模型的发散思维,并有效缓解自反思中可能出现的思想退化问题。Chan等人(2023)构建了多智能体裁判ChatEval(chat evaluators),该系统用于自主讨论和评估模型对传统语言生成或开放性问题等任务下内容合成的质量。研究表明,设定智能体多样化角色以及不同沟通策略对于提高讨论的质量具有重要意义。哈尔滨工业大学的Xiong等人(2023)探讨了两个或多个大语言模型之间的不一致问题,并提出了一个统一的辩论框架。该框架旨在显著提高大语言模型之间的相互一致性,从而提升模型的常识推理规划能力。通过引入辩论,研究人员试图解决不一致性问题,使得多个模型之间能够更一致地推理,提高整体模型的规划性能。

北京邮电大学的Hao等人(2023)提出了一种新的ChatLLM网络模型。允许多个基于对话的语言模型进行交互、提供反馈并共同思考,旨在提高它们解决问题的能力。这种交互式的方法有助于模型之间共同学习,从而更好地理解和处理复杂的任务。中国科学技术大学的Chen等人(2023a)进行了大模型智能体间竞争关系的研究,使用GPT-4模型创建虚拟城镇,其中扮演餐馆角色的智能体之间的竞争,为理解智能体之间的相互作用提供了一种新颖的方法。Zhao等人(2023b)为模型设定不同角色的专家,并指定智能体支持特定立场。最后,决策智能体整合各方观点,以一种协同的方式完成立场检测任务。这一研究通过角色设定和协同决策强调了多个智能体之间的专业合作的重要性。

北京大学的Chen等人(2023b)提出了智能体的自动生成框架AutoAgents。这一框架具有根据任务内容动态生成并协调多个定制化专家智能体为当前任务规划解决方案的特性。这种方法的优势在于能够根据具体任务的要求生成特定的智能体,从而显著提升模型的知识获取和推理能力,使其更适应复杂的场景任务。清华大学Liu等人(2024)构建了大模型智能体动态网络(dynamic LLM-agent network ,DyLAN)框架。该框架的关键特点在于允许智能体在动态架构中进行多轮交互,并引入了测试阶段智能体的优选和早停机制,以提升协作效率。

3　国内外研究进展比较

3.1　基座大模型

基座大模型相关的核心技术早期主要由国外研究者主导,尤其在基于Transformer的模型架构、无监督预训练方法、高效参数适配微调等方面。例如,美国谷歌公司提出Transformer架构和基于掩码的预训练方法,先后发布BERT、Switch Transformer和PaLM等预训练大模型。美国OpenAI公司开发了基于Transformer编码器的GPT架构,先后发布GPT系列模型,尤其是其ChatGPT产品成为历史上最快达到1亿月活用户的应用。

国内在借鉴国外大模型技术的同时,也有独特创新。例如,清华大学结合BERT和GPT的优势架构,提出通用语言模型GLM,并较早地开源预训练

大模型。阿里巴巴、上海人工智能实验室和清华大学等国内团队提出多种适配微调技术和理论解释方法。百度公司探索知识增强的大模型技术，将知识图谱中的实体、关系等语义信息嵌入到模型中，从而提高了模型对语义的理解能力和语义表示能力。

相比于国外，国内的数字化转型积累了丰富的行业数据和应用场景，因而在垂直领域大模型的研制方面具有更多优势。例如，华为和鹏程实验室联合研发了盘古气象大模型，浙江大学发布面向教育、金融和法律等领域的智海系列大模型，上海交通大学开源了白玉兰科学大模型，华东师范大学发布面向心理健康的EmoGPT。尽管国产大模型整体上距离国外最新大模型（如GPT-4）仍有一定差距，但在垂直领域已快速跟进，呈现蓬勃发展的态势。

3.2 大模型小型化

大模型小型化的原始创新技术主要由国外研究人员提出，我国科研人员在技术发展过程中也做出了重要贡献，目前呈现并驾齐驱状态。具体来讲，国内在模型剪枝领域做出更多样化探索和创新，涵盖了通道选择方法、滤波器剪枝建模、随机初始化剪枝、判别感知通道剪枝等多个方面，国外则更注重提出创新性概念并通过简单方法验证概念的有效性。例如，2015年美国斯坦福大学的Han等人（2015）提出模型剪枝的概念，而国内研究人员则在两年后逐渐开始该方向的探索。但该方面差异也在缩小，例如国内外几乎同时开展针对大语言模型的剪枝工作。在模型量化方面，国外研究者较早地从训练后量化和量化感知训练两角度开展研究，提出基于矩阵分解、矢量量化和量化噪声感知等方法。国内提出了可微非线性映射函数描述的模型量化方法、无数据量化方法、PD-Quant方法以及面向视觉Transformer的RepQ-ViT方法。这些方法更注重量化的通用性和模型的性能，也较早地对视觉Transformer等复杂模型的量化展开探索。在知识蒸馏方面，国外首次提出了基于对元、基于特征和基于关系等知识蒸馏领域的代表性方法，国内则较早地结合自监督学习、课程学习和在线学习等新技术进一步提升知识蒸馏的性能。

总体而言，国外在大模型小型化技术方面具有更多的原始性概念或技术创新，提出了一系列经典的理论与方法。基于领域研究基础，国内迅速展开更全面及深入的研究，不断改善方法取得更优异的性能，所研究方法具有较好的实用性。

3.3 协同方式

随着大小模型协同技术不断赋能产业，国内外都聚焦该方面的研究。在协同训练方面，国外最早提出了基于联邦学习的协同学习方法以保护隐私的方式解决数据孤岛问题，并针对模型异构性挑战提出基于迁移学习、模型蒸馏、联邦核岭回归等算法，实现大小模型之间的知识迁移。同时，国外研究者研发了一系列云侧大模型与端侧小模型协同训练方案，如引入双向知识蒸馏方法实现不同设备模型的知识共享，研究特定场景下的持续学习方法使得本地模型持续更新。国内研究更关注端云场景下的本地模型个性化及动态自适应需求。在协同推理方面，国外的研究主要从子模型的自动拆分、模块化学习、模型架构设计等方面开展研究，使得不同运算复杂度的子模块更好地适应场景需求。国内更关注实际应用中的时延问题，通过模型解耦、动态推理卸载等技术应对该类挑战，同时关注对稀疏模型、稠密模型、多模态模型等不同类型模型的拆分方法。值得指出的是，国内在产业级应用方面存在显著优势，所研发系统平台已构建了可支持多种学习任务的产业级的大小模型协同智能计算系统，赋能搜索推荐、直播等应用领域。基于大模型的协同规划是一种新兴的模型协作方式，近期由国外发起该领域研究，国内快速跟进并取得显著进展。在国外，研究者们关注生成式智能体，通过模拟招聘、角色扮演、对话方式促进协作，设计通用框架如MetaAgents、AutoGen、SWIFTSAGE，以应对复杂任务。国内的研究涵盖多智能体辩论、裁判评估、立场检查、模型一致性、竞争关系、动态网络和角色分工等多个方面，为大模型之间的协同规划提供了多种创新方法。

整体来看，目前国内在大小模型协同策略方面处于先进水平，尤其在大小模型协同规划方面具有较强竞争力。另外，国内在产业化方面也表现出明显优势，但在原创性协同理论与技术方面同国外先进水平仍有一定的差距。

4 发展趋势与展望

以生成式大模型为核心的深度学习技术正在引发人工智能领域研究的巨大革新，引起机器学习范式的一系列重要革新，为通用人工智能发展提供了

一种新的手段。随着大模型小型化技术的发展及端侧处理能力的提升，小模型将以更高效低时延方式完成端侧个性化推理任务，大小模型协同技术成为大模型赋能行业应用的重要途径。

大小模型协同进化技术已成为产学研的焦点，被Gartner、高通公司、阿里巴巴达摩院、百度公司等国内外单位预测为革新智能计算范式的科技趋势。该技术具有广阔的行业应用前景，逐渐在城市交通视频监控、大规模搜索推荐、个性化在线教育等场景形成示范性应用。

当前主流大模型采用Transformer架构，然而该架构的单步推理复杂度较高且键值缓存受到内存的限制，存在部署困难的挑战。大模型的一个前沿方向为开发新一代模型架构，在保留Transformer训练并行性、高效长序列建模等优势的同时实现低复杂度推理（Sun等，2024）。另外扩展窗口长度也成为新的研究热点（Chen等，2023d）。与此同时如何设计新型的混合专家模型架构，进一步提升垂直领域专家子模型的可扩展性（Dehghani等，2023），实现跨领域模型融合（Zhang等，2023a；Huang等，2023）以解决复杂场景任务，也是基座模型赋能行业应用的关键。

常见的大模型小型化技术面向视觉模型，而针对语言或多模态大模型的探索尚不够深入。前沿研究包括如何在小型化同时有效保留其上下文学习、思维链、指令遵循等涌现能力（Zhu等，2023a；Gu等，2023）。小型化技术多依赖于经验式设计，而如何自动化设计小型化方法以及如何对剪枝、量化、蒸馏等技术的联合优化也成为重要研究方向。

大小模型协同技术围绕异构模型知识互迁、本地模型持续增强、生成式大模型智能体协同规划等方向展开。在知识互迁方面，前沿研究为对异构模型协同演化机理的探索（Daga等，2023）、自然语言或多模态等大模型的子模型高效拆分框架（Zhu等，2023b）。在本地模型增强方面，已有方法多围绕判别式模型的个性化展开，而在生成式人工智能框架下的少样本个性化方法愈发引起关注（Ruiz等，2023）。在生成式模型协同规划方面，其中一个重要方向为设计智能体间协作框架及反馈方法以激活其对复杂问题的规划决策能力。

由于生成式模型无法对生成内容进行可信性验证，大模型或小模型所合成内容可能包含虚假或欺骗性内容，其安全问题日益凸显。因而研究可信的生成式大小模型也成为重要研究趋势，如模型的“幻觉”检测技术（Rawte等，2023；Manakul等，2023）、合成内容检测（Wang等，2021；Mitchell等，2023；吴汉舟 等，2023；Zhou等，2023）等研究方向。同时，人工智能模型存在内生安全隐患，如对抗攻击（隋晨红等，2023；Wei等，2023b；Wang等，2023a）、大模型“越狱攻击”（jailbreaking attacks）（Zou等，2023；Tian等，2023；Qi等，2023）、拜占庭攻击（Gouissem等，2023；Wei等，2023a）等多种攻击方式，因而研究面向生成式大模型的安全鲁棒防御体系（Jain等，2023；Xie等，2023）也是未来研究热点与重点。

5 结 语

大小模型端云协同进化是生成式基座模型赋能多场景应用的重要技术，涵盖基座大模型、大模型小型化、大小模型协同等关键研究方向，具有端侧实时性、可扩展性、负载低、隐私安全强等优势，是生成式人工智能未来发展重要方向。随着新一代大模型架构、智能体协同规划、合成内容可信验证等相关研究推进，可以预见，大小模型协同进化技术将在智能社会发展领域发挥更大使能作用。

致 谢 本文由中国图象图形学学会数字娱乐与智能生成专业委员会组织撰写，该专委会链接为https://www.csig.org.cn/16/201612/49316.html。

参考文献（References）

Afonin A and Karimireddy S P. 2022. Towards model agnostic federated learning using knowledge distillation//Proceedings of the 10th International Conference on Learning Representations. San Diego, USA: ICLR: 1-23

Ahn S, Hu S X, Damianou A, Lawrence N D and Dai Z. 2019. Variational information distillation for knowledge transfer//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 9163-9171 [DOI: 10.1109/CVPR.2019.00938]

Allen-Zhu Z and Li Y Z. 2020. Towards understanding ensemble, knowledge distillation and self-distillation in deep learning//Proceedings of the 11th International Conference on Learning Representations. Kigali, Rwanda: OpenReview.net: 1-12

Asadi N and Goudarzi M. 2024. Variant parallelism: lightweight deep convolutional models for distributed inference on IoT devices. IEEE

Internet of Things Journal, 11(1): 345-352 [DOI: 10.1109/JIOT.2023.3285877]

Banitalebi-Dehkordi A, Vedula N, Pei J, Xia F, Wang L J and Zhang Y. 2021. Auto-split: a general framework of collaborative edge-cloud AI//Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. Singapore, Singapore: ACM: 2543-2553 [DOI: 10.1145/3447548.3467078]

Banner R, Nahshan Y and Soudry D. 2019. Post training 4-bit quantization of convolutional networks for rapid-deployment//Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: #714

Bao G M and Guo P. 2022. Federated learning in cloud-edge collaborative architecture: key technologies, applications and challenges. Journal of Cloud Computing, 11(1): #94 [DOI: 10.1186/s13677-022-00377-4]

Bhardwaj R, Xia Z X, Ananthanarayanan G, Jiang J C, Shu Y C, Karianakis N, Hsieh K, Bahl P and Stoica I. 2022. Ekya: continuous learning of video analytics models on edge compute servers//Proceedings of the 19th USENIX Symposium on Networked Systems Design and Implementation. Renton, USA: USENIX Association: 119-135

BigScience Workshop. 2022. BLOOM: a 176B-parameter open-access multilingual language model [EB/OL]. [2023-12-31]. https://arxiv.org/pdf/2211.05100v1.pdf

Brown T B, Mann B, Ryder N, Subbiah M, Kaplan J D, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, Agarwal S, Herbert-Voss A, Krueger G, Henighan T, Child R, Ramesh A, Ziegler D M, Wu J, Winter C, Hesse C, Chen M, Sigler E, Litwin M, Gray S, Chess B, Clark J, Berner C, McCandlish S, Radford A, Sutskever I and Amodei D. 2020. Language models are few-shot learners//Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: #159

Bucilua C, Caruana R and Niculescu-Mizil A. 2006. Model compression//Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Philadelphia, USA: ACM: 535-541 [DOI: 10.1145/1150402.1150464]

Chan C M, Chen W Z, Su Y S, Yu J X, Liu Z Y, Fu J, Xue W and Zhang S H. 2023. ChatEval: towards better LLM-based evaluators through multi-agent debate [EB/OL]. [2023-12-31]. https://arxiv.org/pdf/2308.07201.pdf

Chen D F, Mei J P, Wang C, Feng Y and Chen C. 2020. Online knowledge distillation with diverse peers//Proceedings of the 37th AAAI Conference on Artificial Intelligence. New York, USA: AAAI: 3430-3437 [DOI: 10.1609/aaai.v34i04.5746]

Chen D K, Wang H B, Huo Y H, Li Y Z and Zhang H Y. 2023a. GameGPT: multi-agent collaborative framework for game development [EB/OL]. [2023-12-31]. https://arxiv.org/pdf/2310.08067.pdf

Chen G Y, Dong S W, Shu Y, Zhang G, Sesay J, Karlsson B F, Fu J and Shi Y M. 2023b. AutoAgents: a framework for automatic agent generation [EB/OL]. [2023-12-31]. https://arxiv.org/pdf/2309.17288.pdf

Chen P G, Liu S, Zhao H S and Jia J Y. 2021a. Distilling knowledge via knowledge review//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 5006-5015 [DOI: 10.1109/CVPR46437.2021.00497]

Chen T A, Yang D N and Chen M S. 2023c. Overcoming forgetting catastrophe in quantization-aware training//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 17312-17321 [DOI: 10.1109/ICCV51070.2023.01592]

Chen Y K, Qian S J, Tang H T, Lai X, Liu Z J, Han S and Jia J Y. 2023d. LongLoRA: efficient fine-tuning of long-context large language models [EB/OL]. [2023-12-31]. https://arxiv.org/pdf/2309.12307.pdf

Chen Z Y, Yao J C, Wang F, Jia K Y, Han B, Zhang W and Yang H X. 2021b. MC^2-SF: slow-fast learning for mobile-cloud collaborative recommendation [EB/OL]. [2023-12-31]. https://arxiv.org/pdf/2109.12314.pdf

Chowdhery A, Narang S, Devlin J, Bosma M, Mishra G, Roberts A, Barham P, Chung H W, Sutton C, Gehrmann S, Schuh P, Shi K S, Tsvyashchenko S, Maynez J, Rao A, Barnes P, Tay Y, Shazeer N, Prabhakaran V, Reif E, Du N, Hutchinson B, Pope R, Bradbury J, Austin J, Isard M, Gur-Ari G, Yin P C, Duke T, Levskaya A, Ghemawat S, Dev S, Michalewski H, Garcia X, Misra V, Robinson K, Fedus L, Zhou D, Ippolito D, Luan D, Lim H, Zoph B, Spiridonov A, Sepassi R, Dohan D, Agrawal S, Omernick M, Dai A M, Pillai T S, Pellat M, Lewkowycz A, Moreira E, Child R, Polozov O, Lee K, Zhou Z W, Wang X Z, Saeta B, Diaz M, Firat O, Catasta M, Wei J, Meier-Hellstern K, Eck D, Dean J, Petrov S and Fiedel N. 2023. PaLM: scaling language modeling with pathways. Journal of Machine Learning Research, 24(240): 1-13

Daga H, Chen Y W, Agrawal A and Gavrilovska A. 2023. CLUE: systems support for knowledge transfer in collaborative learning with neural nets. IEEE Transactions on Cloud Computing, 11(4): 3541-3554 [DOI: 10.1109/TCC.2023.3294490]

Dai X, Kong X N, Guo T and Huang Y X. 2021. CiNet: redesigning deep neural networks for efficient mobile-cloud collaborative inference//Proceedings of 2021 SIAM International Conference on Data Mining (SDM). Philadelphia, USA: SIAM: 459-467 [DOI: 10.1137/1.9781611976700.52]

Dehghani M, Djolonga J, Mustafa B, Padlewski P, Heek J, Gilmer J, Steiner A, Caron M, Geirhos R, Alabdulmohsin I, Jenatton R, Beyer L, Tschannen M, Arnab A, Wang X, Riquelme C, Minderer M, Puigcerver J, Evci U, Kumar M, Van Steenkiste S, Elsayed G F, Mahendran A, Yu F, Oliver A, Huot F, Bastings J, Collier M P, Gritsenko A A, Birodkar V, Vasconcelos C, Tay Y, Mensink T, Kolesnikov A, Pavetić F, Tran D, Kipf T, Lučić M,

Zhai X H, Keysers D, Harmsen J and Houlsby N. 2023. Scaling vision Transformers to 22 billion parameters//Proceedings of the 40th International Conference on Machine Learning. Honolulu, USA: JMLR.org: #296

Denton E L, Zaremba W, Bruna J, LeCun Y and Fergus R. 2014. Exploiting linear structure within convolutional networks for efficient evaluation//Proceedings of the 27th International Conference on Neural Information Processing Systems. Montreal, Canada: MIT Press: 1269-1277

Dettmers T, Lewis M, Belkada Y and Zettlemoyer L. 2022. Llm.int8(): 8-bit matrix multiplication for Transformers at scale [EB/OL]. [2023-12-31]. https://arxiv.org/pdf/2208.07339.pdf

Devlin J, Chang M W, Lee K and Toutanova K. 2018. BERT: pre-training of deep bidirectional Transformers for language understanding//Proceedings of 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, USA: ACL: 4171-4186 [DOI: 10.18653/v1/N19-1423]

Diao E M, Ding J and Tarokh V. 2021. HeteroFL: computation and communication efficient federated learning for heterogeneous clients//Proceedings of the 9th International Conference on Learning Representations. San Diego, USA: OpenReview.net: 1-24

Ding C T, Zhou A, Liu Y X, Chang R N, Hsu C H and Wang S G. 2022a. A cloud-edge collaboration framework for cognitive service. IEEE Transactions on Cloud Computing, 10(3): 1489-1499 [DOI: 10.1109/TCC.2020.2997008]

Ding M, Yang Z Y, Hong W Y, Zheng W D, Zhou C, Yin D, Lin J Y, Zou X, Shao Z, Yang H X and Tang J. 2021. CogView: mastering text-to-image generation via Transformers//Proceedings of the 35th Conference on Neural Information Processing Systems. Vancouver, Camada: OpenReview.net: 19822-19835

Ding N, Qin Y J, Yang G, Wei F C, Yang Z H, Su Y S, Hu S D, Chen Y L, Chan C M, Chen W Z, Yi J, Zhao W L, Wang X Z, Liu Z Y, Zheng H T, Chen J F, Liu Y, Tang J, Li J Z and Sun M S. 2022b. Delta tuning: a comprehensive study of parameter efficient methods for pre-trained language models [EB/OL]. [2023-12-31]. https://arxiv.org/pdf/2203.06904.pdf

Ding N, Qin Y J, Yang G, Wei F C, Yang Z H, Su Y S, Hu S D, Chen Y L, Chan C M, Chen W Z, Yi J, Zhao W L, Wang X Z, Liu Z Y, Zheng H T, Chen J F, Liu Y, Tang J, Li J Z and Sun M S. 2023. Parameter-efficient fine-tuning of large-scale pre-trained language models. Nature Machine Intelligence, 5(3): 220-235 [DOI: 10.1038/s42256-023-00626-4]

Dong Z Q, He Q, Chen F F, Jin H, Gu T and Yang Y. 2023. EdgeMove: pipelining device-edge model training for mobile intelligence//Proceedings of 2023 ACM Web Conference. New York, USA: ACM: 3142-3153 [DOI: 10.1145/3543507.3583540]

Du Y L, Li S, Torralba A, Tenenbaum J B and Mordatch I. 2023. Improving factuality and reasoning in language models through multiagent debate [EB/OL]. [2023-12-31]. https://arxiv.org/pdf/2305.14325.pdf

Du Z X, Qian Y J, Liu X, Ding M, Qiu J Z, Yang Z L and Tang J. 2022. GLM: general language model pretraining with autoregressive blank infilling//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. Dublin, Ireland: ACL: 320-335 [DOI: 10.18653/v1/2022.acl-long.26]

Fedus W, Zoph B and Shazeer N. 2022. Switch Transformers: Scaling to trillion parameter models with simple and efficient sparsity. The Journal of Machine Learning Research, 23(1): #120

Frantar E and Alistarh D. 2023. Sparsegpt: massive language models can be accurately pruned in one-shot//Proceedings of the 40th International Conference on Machine Learning. Honolulu, USA: PMLR

Fu Y, Peng H, Khot T and Lapata M. 2023. Improving language model negotiation with self-play and in-context learning from AI feedback//Proceedings of the 37th Conference on Neural Information Processing Systems. New Orleans, USA: OpenReview.net: 1-11

Gong Y C, Liu L, Yang M and Bourdev L. 2015. Compressing deep convolutional networks using vector quantization [EB/OL]. [2023-12-31]. https://arxiv.org/pdf/1412.6115.pdf

Gordon M, Duh K and Andrews N. 2020. Compressing BERT: studying the effects of weight pruning on transfer learning//Proceedings of the 5th Workshop on Representation Learning for NLP. Virtual: ACL: 143-155 [DOI: 10.18653/v1/2020.repl4nlp-1.18]

Gou J P, Yu B S, Maybank S J and Tao D C. 2021. Knowledge distillation: a survey. International Journal of Computer Vision, 129(6): 1789-1819 [DOI: 10.1007/s11263-021-01453-z]

Gouissem A, Abualsaud K, Yaacoub E, Khattab T and Guizani M. 2023. Collaborative byzantine resilient federated learning. IEEE Internet of Things Journal, 10(18): 15887-15899 [DOI: 10.1109/JIOT.2023.3266347]

Gu G X, Meng X J, Lu G S, Hou L, Niu M Z, Liang X D, Yao L W, Huang R H, Zhang W, Jiang X, Xu C J and Xu H. 2022. Wukong: a 100 million large-scale Chinese cross-modal pre-training benchmark//Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, USA: MIT Press: 26418-26431

Gu Y X, Dong L, Wei F R and Huang M L. 2023. Knowledge distillation of large language models [EB/OL]. [2023-12-31]. https://arxiv.org/pdf/2306.08543.pdf

Hamilton S. 2023. Blind Judgement: agent-based supreme court modelling with GPT//The AAAI-23 Workshop on Creative AI Across Modalities. Washington, USA: OpenReview.net: 1-6

Han S, Mao H Z and Dally W J. 2016. Deep compression: compressing deep neural networks with pruning, trained quantization and Huffman coding//Proceedings of the 4th International Conference on Learning Representations. San Juan, Puerto Rico: OpenReview.net: #149

Han S, Pool J, Tran J and Dally W. 2015. Learning both weights and

connections for efficient neural network//Proceedings of the 28th International Conference on Neural Information Processing Systems. Montréal, Canada: MIT Press: 1135-1143

Hao R, Hu L M, Qi W J, Wu Q L, Zhang Y R and Nie L Q. 2023. ChatLLM network: more brains, more intelligence [EB/OL]. [2023-12-31]. https://arxiv.org/pdf/2304.12998.pdf

He R D, Liu L L, Ye H, Tan Q Y, Ding B S, Cheng L Y, Low J W, Bing L D and Si L. 2021. On the effectiveness of adapter-based tuning for pretrained language model adaptation//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. [s.l.]: Association for Computational Linguistics: 2208-2222 [DOI: 10.18653/v1/2021.acl-long.172]

He Y H, Zhang X Y and Sun J. 2017. Channel pruning for accelerating very deep neural networks//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 1398-1406 [DOI: 10.1109/ICCV.2017.155]

Hinton G, Vinyals O and Dean J. 2015. Distilling the knowledge in a neural network [EB/OL]. [2023-12-31]. https://arxiv.org/pdf/1503.02531.pdf

Ho J, Jain A and Abbeel P. 2020. Denoising diffusion probabilistic models//Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: #574

Ho N, Schmid L and Yun S Y. 2023. Large language models are reasoning teachers//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics. Toronto, Canada: ACL: 14852-14882 [DOI: 10.18653/v1/2023.acl-long.830]

Houlsby N, Giurgiu A, Jastrzebski S, Morrone B, De Laroussilhe Q, Gesmundo A, Attariyan M and Gelly S. 2019. Parameter-efficient transfer learning for NLP//Proceedings of the 36th International Conference on Machine Learning. Long Beach, USA: ICML: 2790-2799

Hu E J, Shen Y L, Wallis P, Allen-Zhu Z Y, Li Y Z, Wang S A, Wang L and Chen W Z. 2022. Lora: low-rank adaptation of large language models//Proceedings of the 10th International Conference on Learning Representations. Virtual: OpenReview.net: 1-26

Huang C S, Liu Q, Lin B Y, Du C, Pang T Y and Lin M. 2023. LoraHub: efficient cross-task generalization via dynamic lora composition//Proceedings of the 12th International Conference on Learning Representations. OpenReview.net: 1-20

Huang Y K, Chen Y D, Yu Z and McKeown K. 2022. In-context learning distillation: transferring few-shot learning ability of pre-trained language models [EB/OL]. [2023-12-31]. https://arxiv.org/pdf/2212.10670.pdf

InternLM Team. 2023. InternLM: a multilingual language model with progressively enhanced capabilities [EB/OL]. [2023-12-31]. https://github.com/InternLM/InternLM-techreport/blob/main/InternLM.pdf

Jacob B, Kligys S, Chen B, Zhu M L, Tang M, Howard A, Adam H and Kalenichenko D. 2018. Quantization and training of neural networks for efficient integer-arithmetic-only inference//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 2704-2713 [DOI: 10.1109/CVPR.2018.00286]

Jacobs R A, Jordan M I, Nowlan S J and Hinton G E. 1991. Adaptive mixtures of local experts. Neural Computation, 3(1): 79-87 [DOI: 10.1162/neco.1991.3.1.79]

Jain N, Schwarzschild A, Wen Y X, Somepalli G, Kirchenbauer J, Chiang P Y, Goldblum M, Saha A, Geiping J and Goldstein T. 2023. Baseline defenses for adversarial attacks against aligned language models//Proceedings of the 12th International Conference on Learning Representations. [s.l.]: OpenReview.net: 1-22

Ji M, Heo B and Park S. 2021. Show, attend and distill: knowledge distillation via attention-based feature matching//Proceedings of the 37th AAAI Conference on Artificial Intelligence. Washington, USA: AAAI: 7945-7952 [DOI: 10.1609/aaai.v35i9.16969]

Jiang P H, Xin K, Li C X and Zhou Y S. 2023. High-efficiency device-cloud collaborative Transformer model//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 2204-2210 [DOI: 10.1109/cvprw59228.2023.00214]

Ko J H, Na T, Amir M F, Mukhopadhyay S. 2018. Edge-host partitioning of deep neural networks with feature space encoding for resource-constrained internet-of-things platforms//Proceedings of the 15th IEEE International Conference on Advanced Video and Signal Based Surveillance. Auckland, New Zealand: IEEE: 1-6 [DOI: 10.1109/AVSS.2018.8639121]

Lan Z Z, Chen M D, Goodman S, Gimpel K, Sharma P and Soricut R. 2019. ALBERT: a lite BERT for self-supervised learning of language representations//Proceedings of the 8th International Conference on Learning Representations. Addis Ababa, Ethiopia: OpenReview.net: 1-17

Lepikhin D, Lee H, Xu Y Z, Chen D H, Firat O, Huang Y P, Krikun M, Shazeer N and Chen Z F. 2020. GShard: scaling giant models with conditional computation and automatic sharding//Proceedings of 2020 International Conference on Learning Representations (ICLR). Addis Ababa, Ethiopia: ICLR: 1-35 [DOI: 10.18653/v1/2020.iclr-1.1].

Lester B, Al-Rfou R and Constant N. 2021. The power of scale for parameter-efficient prompt tuning//Proceedings of 2021 Conference on Empirical Methods in Natural Language Processing. Punta Cana, Dominican Republic: ACL: 3045-3059 [DOI: 10.18653/v1/2021.emnlp-main.243]

Lewis M, Liu Y H, Goyal N, Ghazvininejad M, Mohamed A, Levy O, Stoyanov V and Zettlemoyer L. 2020. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. [s.l.]: ACL:

7871-7880［DOI：10.18653/v1/2020.acl-main.703］

Liao Z，Quétu V，Nguyen VT and Tartaglione E. 2023. Can Unstructured pruning reduce the depth in deep neural networks?//Proceedings of the IEEE/CVF International Conference on Computer Vision. Paris，France：IEEE/CVF：402-1406［DOI：10.1109/ICCVW60793.2023.00151］

Li D L and Wang J P. 2019. FedMD：heterogenous federated learning via model distillation［EB/OL］.［2023-12-31］. https://arxiv.org/pdf/1910.03581.pdf

Li G H，Hammoud H A A K，Itani H，Khizbullin D and Ghanem B. 2023a. Camel：communicative agents for “mind” exploration of large language model society//Proceedings of the 37th Conference on Neural Information Processing Systems. New Orleans，USA：OpenReview.net：1-18

Li H，Kadav A，Durdanovic I，Samet H and Graf H P. 2017. Pruning filters for efficient ConvNets//Proceedings of the 5th International Conference on Learning Representations. Toulon，France：ICLR：1-13

Li H，Zhu J G，Jiang X H，Zhu X Z，Li H S，Yuan C，Wang X H，Qiao Y，Wang X G，Wang W H and Dai J F. 2023b. Uni-perceiver v2：a generalist model for large-scale vision and vision-language tasks//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver，Canada：IEEE：2691-2700［DOI：10.1109/CVPR52729.2023.00264］

Li H S，Hu C H，Jiang J Y，Wang Z，Wen Y G and Zhu W W. 2018. JALAD：joint accuracy-and latency-aware deep structure decoupling for edge-cloud execution//Proceedings of the 24th IEEE International Conference on Parallel and Distributed Systems. Singapore，Singapore：IEEE：671-678［DOI：10.1109/PADSW.2018.8645013］

Li J N，Li D X，Savarese S and Hoi S. 2023c. BLIP-2：bootstrapping language-image pre-training with frozen image encoders and large language models//Proceedings of the 40th International Conference on Machine Learning. Honolulu，USA：JMLR.org：#814

Li J N，Li D X，Xiong C M and Hoi S C H. 2022a. BLIP：bootstrapping language-image pre-training for unified vision-language understanding and generation//Proceedings of the 39th International Conference on Machine Learning. Baltimore，USA：PMLR：12888-12900

Li S Y，Chen J S，Shen Y L，Chen Z Y，Zhang X L，Li Z K，Wang H，Qian J，Peng B L，Mao Y，Chen W H and Yan X F. 2022b. Explanations from large language models make small reasoners better［EB/OL］.［2023-12-31］. https://arxiv.org/pdf/2210.06726.pdf

Li T，Sahu A K，Zaheer M，Sanjabi M，Talwalkar A and Smith V. 2020. Federated optimization in heterogeneous networks//Proceedings of Machine Learning and Systems 2020. Austin，USA：mlsys.org，2020：429-450

Li X L and Liang P. 2021. 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. ACL：4582-4597［DOI：10.18653/v1/2021.acl-long.353］

Li Y，Zhang Y X and Sun L C. 2023e. MetaAgents：simulating interactions of human behaviors for LLM-based task-oriented coordination via collaborative generative agents［EB/OL］.［2023-12-31］. https://arxiv.org/pdf/2310.06500.pdf

Li Y W，Adamczewski K，Li W，Gu S H，Timofte R and Van Gool L. 2022c. Revisiting random channel pruning for neural network compression//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans，USA：IEEE：191-201［DOI：10.1109/CVPR52688.2022.00029］

Li Y X，Yu Y F，Liang C，He P C，Karampatziakis N，Chen W Z and Zhao T. 2023d. LoftQ：LoRA-fine-tuning-aware quantization for large language models［EB/OL］.［2023-12-31］. https://arxiv.org/pdf/2310.08659.pdf

Li Z，Li X，Yang L F，Zhao B R，Song R J，Luo L，Li J and Yang J. 2023h. Curriculum temperature for knowledge distillation//Proceedings of the 37th AAAI Conference on Artificial Intelligence. Washington，USA：AAAI Press：1504-1512［DOI：10.1609/aaai.v37i2.25236］

Li Z K，Xiao J R，Yang L W and Gu Q Y. 2023f. RepQ-ViT：scale reparameterization for post-training quantization of vision Transformers//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris，France：IEEE：17181-17190［DOI：10.1109/ICCV51070.2023.01580］

Li Z X，Li Q W，Zhou Y，Zhong W L，Zhang G N and Wu C. 2023g. Edge-cloud collaborative learning with federated and centralized features//Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. Taipei，China：ACM：1949-1953［DOI：10.1145/3539618.3591976］

Liang T，He Z W，Jiao W X，Wang X，Wang Y，Wang R，Yang Y J，Tu Z P and Shi S M. 2023. Encouraging divergent thinking in large language models through multi-agent debate［EB/OL］.［2023-12-31］. https://arxiv.org/pdf/2305.19118.pdf

Lin B Y，Fu Y C，Yang K，Ammanabrolu P，Brahman F，Huang S Y，Bhagavatula C，Choi Y and Ren X. 2023. SwiftSage：a generative agent with fast and slow thinking for complex interactive tasks//37th Interactive Learning with Implicit Human Feedback Workshop at ICML 2023. New Orleans，USA：OpenReview.net：1-18

Lin J Y，Men R，Yang A，Zhou C，Ding M，Zhang Y C，Wang P，Wang A，Jiang L，Jia X Y，Zhang J，Zhang J W，Zou X，Li Z K，Deng X D，Liu J，Xue J B，Zhou H L，Ma J X，Yu J，Li Y，Lin W，Zhou J R，Tang J and Yang H X. 2021. M6：a Chinese multimodal pretrainer［EB/OL］.［2023-12-31］. https://arxiv.org/pdf/2103.00823.pdf

Liu J，Zhuang B H，Zhuang Z W，Guo Y，Huang J Z，Zhu J H and Tan M K. 2022a. Discrimination-aware network pruning for deep model compression. IEEE Transactions on Pattern Analysis and Machine Intelligence，44（8）：4035-4051［DOI：10.1109/TPAMI.2021.

3066410]

Liu J W, Niu L, Yuan Z H, Yang D W, Wang X G and Liu W Y. 2023a. PD-Quant: post-training quantization based on prediction difference metric//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 24427-24437 [DOI: 10.1109/CVPR52729.2023.02340]

Liu X, Ji K X, Fu Y C, Tam W L, Du Z X, Yang Z L and Tang J. 2022c. P-tuning: prompt tuning can be comparable to fine-tuning across scales and tasks//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Dublin, Ireland: ACL: 61-68 [DOI: 10.18653/v1/2022.acl-short.8]

Liu X, Zheng Y N, Du Z X, Ding M, Qian Y J, Yang Z L and Tang J. 2021. GPT understands, too [EB/OL]. [2023-12-31]. https://arxiv.org/pdf/2103.10385.pdf

Liu Y H, Ott M, Goyal N, Du J F, Joshi M, Chen D Q, Levy O, Lewis M, Zettlemoyer L and Stoyanov V. 2019. RoBERTa: a robustly optimized BERT pretraining approach [EB/OL]. [2023-12-31]. https://arxiv.org/pdf/1907.11692.pdf

Liu Z C, Oguz B, Zhao C S, Chang E, Stock P, Mehdad Y, Shi Y Y, Krishnamoorthi R and Chandra V. 2023b. LLM-QAT: data-free quantization aware training for large language models [EB/OL]. [2023-12-31]. https://arxiv.org/pdf/2305.17888.pdf

Liu Z J, Zhang Y Z, Li P, Liu Y and Yang D Y. 2024. Dynamic LLM-agent network: an LLM-agent collaboration framework with agent team optimization//Proceedings of the 12th International Conference on Learning Representations [s.l.]: OpenReview.net: 1-22

Lu Y, Shu Y C, Tan X, Liu Y X, Zhou M Y, Chen Q and Pei D. 2019a. Collaborative learning between cloud and end devices: an empirical study on location prediction//Proceedings of the 4th ACM/IEEE Symposium on Edge Computing. Arlington Virginia, USA: Association for Computing Machinery: 139-151 [DOI: 10.1145/3318216.3363304]

Lu J S, Batra D, Parikh D and Lee S. 2019b. ViLBERT: pretraining task-agnostic visiolinguistic representations for vision-and-language tasks//Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.2

Luo J H, Wu J X and Lin W Y. 2017. ThiNet: a filter level pruning method for deep neural network compression//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 5068-5076 [DOI: 10.1109/ICCV.2017.541]

Lyu C F, Niu C Y, Gu R J, Jiang X T, Wang Z D, Liu B, Wu Z Q, Yao Q L, Huang C Y, Huang P, Huang T, Shu H, Song J D, Zou B, Lan P, Xu G H, Wu F, Tang S J, Wu F and Chen G H. 2022. Walle: an end-to-end, general-purpose, and large-scale production system for device-cloud collaborative machine learning//Proceedings of the 16th USENIX Symposium on Operating Systems Design and Implementation. Carlsbad, USA: USENIX Association: 1-22

Lyu Z Q, Zhang W Q, Zhang S Y, Kuang K, Wang F, Wang Y W, Chen Z Y, Shen T, Yang H X, Ooi B C and Wu F. 2023. DUET: a tuning-free device-cloud collaborative parameters generation framework for efficient device model generalization//Proceedings of 2023 ACM Web Conference. Austin, USA: ACM: 3077-3085 [DOI: 10.1145/3543507.3583451]

Ma X Y, Fang G F and Wang X C. 2023a. LLM-Pruner: on the structural pruning of large language models//Proceedings of the 37th Conference on Neural Information Processing Systems. New Orleans, USA: OpenReview.net: 1-19

Ma X Y, Jeong S, Zhang M J, Wang D, Choi J and Jeon M. 2023b. Cost-effective on-device continual learning over memory hierarchy with Miro//Proceedings of the 29th Annual International Conference on Mobile Computing and Networking. Madrid, Spain: ACM: #83 [DOI: 10.1145/3570361.3613297]

Madan K, Ke R N, Goyal A, Schölkopf B B and Bengio Y. 2021. Fast and slow learning of recurrent independent mechanisms//Proceedings of the 9th International Conference on Learning Representations. [s.l.]: OpenReview.net: 1-17

Manakul P, Liusie A and Gales M J F. 2023. SelfCheckGPT: zero-resource black-box hallucination detection for generative large language models//Proceedings of 2023 Conference on Empirical Methods in Natural Language Processing. Singapore, Singapore: ACL: 9004-9017 [DOI: 10.18653/v1/2023.emnlp-main.557]

McMahan B, Moore E, Ramage D, Hampson S and Arcas B A Y. 2017. Communication-efficient learning of deep networks from decentralized data//Proceedings of the 20th International Conference on Artificial Intelligence and Statistics. Fort Lauderdale, USA: PMLR: 1273-1282

Mitchell E, Lee Y, Khazatsky A, Manning C D and Finn C. 2023. DetectGPT: zero-shot machine-generated text detection using probability curvature//Proceedings of the 40th International Conference on Machine Learning. Honolulu, USA: JMLR.org: #1038

Nair V, Schumacher E, Tso G and Kannan A. 2023. DERA: enhancing large language model completions with dialog-enabled resolving agents [EB/OL]. [2023-12-31]. https://arxiv.org/pdf/2303.17071.pdf

Nan Y, Jiang S Q and Li M. 2024. Large-scale video analytics with cloud-edge collaborative continuous learning. ACM Transactions on Sensor Networks, 20(1): #14 [DOI: 10.1145/3624478]

Narayanan D, Shoeybi M, Casper J, LeGresley P, Patwary M, Korthikanti V, Vainbrand D, Kashinkunti P, Bernauer J, Catanzaro B, Phanishayee A and Zaharia M. 2021. Efficient large-scale language model training on GPU clusters using megatron-LM//SC21: International Conference for High Performance Computing, Networking, Storage and Analysis. St. Louis, USA: IEEE: #58 [DOI: 10.1145/3458817.3476209]

Ouyang L, Wu J, Jiang X, Almeida D, Wainwright C L, Mishkin P,

Zhang C, Agarwal S, Slama K, Ray A, Schulman J, Hilton J, Kelton F, Miler L, Simens M, Askell A, Welinder P, l Christiano P, Leike J and Lowe R. 2022. Training language models to follow instructions with human feedback [EB/OL]. [2023-12-31]. https://arxiv.org/pdf/2203.02155.pdf

Niu C Y, Wu F, Tang S J, Hua L F, Jia R F, Lyu C F, Wu Z H and Chen G H. 2020. Billion-scale federated learning on mobile clients: a submodel design with tunable privacy//Proceedings of the 26th Annual International Conference on Mobile Computing and Networking. London, UK: ACM: #31 [DOI: 10.1145/3372224.3419188]

Pacheco R G, Couto R S and Simeone O. 2021. Calibration-aided edge inference offloading via adaptive model partitioning of deep neural networks//Proceedings of 2021 IEEE International Conference on Communications. Montreal, Canada: IEEE: 1-6 [DOI: 10.1109/ICC42927.2021.9500760]

Padmanabhan A, Iyer A P, Ananthanarayanan G, Shu Y C, Karianakis N, Xu G H and Netravali R. 2021. Towards memory-efficient inference in edge video analytics//Proceedings of the 3rd ACM Workshop on Hot Topics in Video Analytics and Intelligent Edges. New York, USA: ACM: 31-37 [DOI: 10.1145/3477083.3480150]

Park J, Min B, Ma X J and Kim J. 2023. ChoiceMates: supporting unfamiliar online decision-making with multi-agent conversational interactions [EB/OL]. [2023-12-31]. https://arxiv.org/pdf/2310.01331.pdf

Park W, Kim D, Lu Y and Cho M. 2019. Relational knowledge distillation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 3967-3976 [DOI: 10.1109/CVPR.2019.00409]

Passalis N and Tefas A. 2018. Learning deep representations with probabilistic knowledge transfer//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer: 283-299 [DOI: 10.1007/978-3-030-01252-6_17]

Pham C, Liu B Y, Yang Y X, Chen Z Y, Liu T Y, Yuan J B, Plummer B A, Wang Z R and Yang H X. 2023. Let models speak ciphers: multiagent debate through embeddings [EB/OL]. [2023-12-31]. https://arxiv.org/pdf/2310.06272.pdf

Qi X Y, Huang K X, Panda A, Wang M D and Mittal P. 2023. Visual adversarial examples jailbreak aligned large language models//The 2nd Workshop on New Frontiers in Adversarial Machine Learning. [s.l.]: OpenReview.net:

Radford A, Kim J W, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, Krueger G and Sutskever I. 2021. Learning transferable visual models from natural language supervision//Proceedings of the 38th International Conference on Machine Learning. Online: PMLR: 8748-8763

Radford A, Narasimhan K, Salimans T and Sutskever I. 2018. Improving language understanding by generative pre-training [EB/OL]. [2023-12-31]. https://www.mikecaptain.com/resources/pdf/GPT-1.pdf

Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, Zhou Y Q, Li W and Liu P J. 2020. Exploring the limits of transfer learning with a unified text-to-text Transformer. The Journal of Machine Learning Research, 21(1): #140

Ramesh A, Pavlov M, Goh G, Gray S, Voss C, Radford A, Chen M and Sutskever I. 2021. Zero-shot text-to-image generation//Proceedings of the 38th International Conference on Machine Learning. Online: ICML: 8821-8831

Rawte V, Sheth A and Das A. 2023. A survey of hallucination in large foundation models [EB/OL]. [2023-12-31]. https://arxiv.org/pdf/2303.08896.pdf

Rombach R, Blattmann A, Lorenz D, Esser P and Ommer B. 2022. High-resolution image synthesis with latent diffusion models//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 10674-10685 [DOI: 10.1109/CVPR52688.2022.01042]

Romero A, Ballas N, Kahou S E, Chassang A, Gatta C and Bengio Y. 2015. FitNets: hints for thin deep nets//Proceedings of the 3rd International Conference on Learning Representations. San Diego, USA: ICLR: 1-14

Ruiz N, Li Y Z, Jampani V, Wei W, Hou T B, Pritch Y, Wadhwa N, Rubinstein M and Aberman K. 2023. Hyperdreambooth: hypernetworks for fast personalization of text-to-image models [EB/OL]. [2023-12-31]. https://arxiv.org/pdf/2307.06949.pdf

Saharia C, Chan W, Saxena S, Li L L, Whang J, Denton E L, Ghasemipour S K S, Gontijo-Lopes R, Ayan B K, Salimans T, Ho J, Fleet D J and Norouzi M. 2022. Photorealistic text-to-image diffusion models with deep language understanding//Proceedings of the 36th Conference on Neural Information Processing Systems. New Orleans, USA: OpenReview.net: 1-16

Shao S J, Shao C Z, Zhong C, Guo S Y and Lu P C. 2022. Cloud-edge collaboration based power IoT scene perception mechanism//Proceedings of the 11th International Conference on Game Theory for Networks. Virtual: Springer: 100-117 [DOI: 10.1007/978-3-031-23141-4_8]

Shazeer N, Mirhoseini A, Maziarz K, Davis A, Le Q, Hinton G and Dean J. 2017. Outrageously large neural networks: the sparsely-gated mixture-of-experts layer//Proceedings of the 5th International Conference on Learning Representations. Toulon, France: OpenReview.net: 1-19

Stock P, Fan A, Graham B, Grave E, Gribonval R, Jegou H and Joulin A. 2022. Training with quantization noise for extreme model compression//Proceedings of the 9th International Conference on Learning Representations. San Diego: OpenReview.net: 19123-19138

Su W J, Zhu X Z, Cao Y, Li B, Lu L W, Wei F R and Dai J F. 2020. VL-BERT: pre-training of generic visual-linguistic representations//Proceedings of the 8th International Conference on Learning Representations. Addis Ababa, Ethiopia: OpenReview.net

Sui C H, Wang A, Zhou S W, Zang A K, Pan Y H, Liu H and Wang H P. 2023. A survey on adversarial training for robust learning. Journal of Image and Graphics, 28(12): 3629-3650 (隋晨红, 王奥, 周圣文, 臧安康, 潘云豪, 刘颢, 王海鹏. 2023. 面向鲁棒学习的对抗训练技术综述. 中国图象图形学报, 28(12): 3629-3650) [DOI: 10.11834/jig.220953]

Sun C, Myers A, Vondrick C, Murphy K and Schmid C. 2019. VideoBERT: a joint model for video and language representation learning//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 7463-7472 [DOI: 10.1109/ICCV.2019.00756]

Sun Y, Wang S H, Feng S K, Ding S Y, Pang C, Shang J Y, Liu J X, Chen X Y, Zhao Y B, Lu Y X, Liu W X, Wu Z H, Gong W B, Liang J Z, Shang Z Z, Sun P, Liu W, Yang X O, Yu D H, Tian H, W H and Wang H F. 2021. ERNIE 3.0: large-scale knowledge enhanced pre-training for language understanding and generation [EB/OL]. [2023-12-31]. https://arxiv.org/pdf/2107.02137.pdf

Sun Y T, Dong L, Huang S H, Ma S M, Xia Y Q, Xue J L, Wang J Y and Wei F R. 2024. Retentive network: a successor to Transformer for large language models//Proceedings of the 12th International Conference on Learning Representations. [s. l.]: OpenReview.net: 1-14

Sung Y L, Yoon J H and Bansal M. 2023. ECoFLaP: Efficient coarse-to-fine layer-wise pruning for vision-language models[EB/OL].[2023-12-31]. https://arxiv.org/pdf/2310.02998.pdf

Tao M, Bao B K, Tang H and Xu C S. 2023. GALIP: generative adversarial CLIPs for text-to-image synthesis//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 14214-14223 [DOI: 10.1109/CVPR52729.2023.01366]

Tian Y, Yang X, Zhang J Y, Dong Y P and Su H. 2023. Evil geniuses: delving into the safety of LLM-based agents [EB/OL]. [2023-12-31]. https://arxiv.org/pdf/2311.11855.pdf

Touvron H, Lavril T, Izacard G, Martinet X, Lachaux M A, Lacroix T, Rozière B, Goyal N, Hambro E, Azhar F, Rodriguez A, Joulin A, Grave E and Lample G. 2023. Llama: open and efficient foundation language models [EB/OL]. [2023-12-31]. https://arxiv.org/pdf/2302.13971.pdf

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser Ł and Polosukhin I. 2017. Attention is all you need//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc.: 6000-6010

Wang Y L, Zhang X L, Xie L X, Zhou J, Su H, Zhang B and Hu X L. 2020. Pruning from scratch//Proceedings of the 37th AAAI Conference on Artificial Intelligence. Washington, USA: AAAI Press: 12273-12280 [DOI: 10.1609/aaai.v34i07.6910]

Wang Y W, Ding X, Yang Y X, Ding L, Ward R and Wang Z J. 2021. Perception matters: exploring imperceptible and transferable anti-forensics for GAN-generated fake face imagery detection. Pattern Recognition Letters, 146: 15-22 [DOI: 10.1016/j.patrec.2021.03.009]

Wang Y W, Liu Y and Shen Z Q. 2023a. Revisiting item promotion in GNN-based collaborative filtering: a masked targeted topological attack perspective//Proceedings of the 37th AAAI Conference on Artificial Intelligence. Washington, USA: AAAI Press: 15206-15214 [DOI: 10.1609/aaai.v37i12.26774]

Wang Y W, Wang Y H, Cai J Y, Lee T K, Miao C Y and Wang Z J. 2023b. SSD-KD: a self-supervised diverse knowledge distillation method for lightweight skin lesion classification using dermoscopic images. Medical Image Analysis, 84: #102693 [DOI: 10.1016/j.media.2022.102693]

Wang Z H L, Mao S G, Wu W S, Ge T, Wei F R and Ji H. 2023c. Unleashing the Emergent cognitive synergy in large language models: a task-solving agent through multi-persona self-collaboration [EB/OL]. [2023-12-31]. https://arxiv.org/pdf/2307.05300.pdf

Wei H L, Zhang H, Ai-Haddad K and Shi Y. 2023a. Ensuring secure platooning of constrained Intelligent and connected vehicles against Byzantine attacks: a distributed MPC framework. Engineering: #007 [DOI: 10.1016/j.eng.2023.10.007]

Wei Z P, Chen J J, Wu Z X and Jiang Y G. 2023b. Adaptive cross-modal transferable adversarial attacks from images to videos. IEEE Transactions on Pattern Analysis and Machine Intelligence: #3347835 [DOI: 10.1109/TPAMI.2023.3347835].

Wu C F, Liang J, Ji L, Yang F, Fang Y J, Jiang D X and Duan N. 2022. NÜWA: visual synthesis pre-training for neural visual world creation//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer: 720-736 [DOI: 10.1007/978-3-031-19787-1_41]

Wu H Z, Zhang J, Li Y, Yin Z X, Zhang X P, Tian H, Li B, Zhang W M and Yu N H. 2023. Overview of artificial intelligence model watermarking. Journal of Image and Graphics, 28(6): 1792-1810 (吴汉舟, 张杰, 李越, 殷赵霞, 张新鹏, 田晖, 李斌, 张卫明, 俞能海. 2023. 人工智能模型水印研究进展. 中国图象图形学报, 28(6): 1792-1810) [DOI: 10.11834/jig.230010]

Wu Q Y, Bansal G G, Zhang J Y, Wu Y R, Li B B, Zhu E K, Jiang L, Zhang X Y, Zhang S K, Liu J L, Awadallah A H, White R W, Burger D and Wang C. 2023. AutoGen: enabling next-gen LLM applications via multi-agent conversation//Proceedings of the 12th International Conference on Learning Representations. [s. l.]: OpenReview.net: 1-43

Xiao G X, Lin J, Seznec M, Wu H, Demouth J and Han S. 2023. SmoothQuant: accurate and efficient post-training quantization for large language models//Proceedings of the 40th International Conference on Machine Learning. Honolulu, USA: ICML: 38087-38099

Xie Y Q, Yi J W, Shao J W, Curl J, Lyu L J, Chen Q F, Xie X and Wu F Z. 2023. Defending ChatGPT against jailbreak attack via self-

reminders. Nature Machine Intelligence, 5(12): 1486-1496 [DOI: 10.1038/s42256-023-00765-8]

Xiong K, Ding X, Cao Y X, Liu T and Qin B. 2023. Examining inter-consistency of large language models collaboration: an in-depth analysis via debate//Findings of the Association for Computational Linguistics: EMNLP 2023. Singapore, Singapore: ACL: #508 [DOI: 10.18653/v1/2023.findings-emnlp.508]

Xu G D, Liu Z W, Li X X and Loy C C. 2020a. Knowledge distillation meets self-supervision//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 588-604 [DOI: 10.1007/978-3-030-58545-7_34]

Xu S K, Li H K, Zhuang B H, Liu J, Cao J Z, Liang C R and Tan M K. 2020b. Generative low-bitwidth data free quantization//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 1-17 [DOI: 10.1007/978-3-030-58610-2_1]

Xu Z C, Zhao L Q, Liang W F, Rana O F, Zhou P, Xia Q F, Xu W Z and Wu G W. 2021. Energy-aware inference offloading for DNN-driven applications in mobile edge clouds. IEEE Transactions on Parallel and Distributed Systems, 32(4): 799-814 [DOI: 10.1109/TPDS.2020.3032443]

Yan Y K, Niu C Y, Gu R J, Wu F, Tang S J, Hua L F, Lyu C F and Chen G H. 2022. On-device learning for model personalization with large-scale cloud-coordinated domain adaption//Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. Washington, USA: ACM: 2180-2190 [DOI: 10.1145/3534678.3539263]

Yang A Y, Xiao B, Wang B G, Zhang B R, Bian C, Yin C, Lyu C X, Pan D, Wang D, Yan D, Yang F, Deng F, Wang F, Liu F, Ai G W, Dong G S, Zhao H Z, Xu H, Sun H Z, Zhang H D, Liu H, Ji J M, Xie J, Dai J T, Fang K, Su L, Song L, Liu L F, Ru L Y, o Ma L Y, Wang M, Liu M, Lin M A, Nie N L, Guo P D, Sun R Y, Zhang T, Li T P, Li T Y, Cheng W, Chen W P, Zeng X R, Wang X C, Chen X X, Men X, Yu X, Pan X H, Shen Y J, Wang Y D, Li Y Y, Jiang Y X, Gao Y C, Zhang Y P, Zhou Z N and Wu Z Y. 2023. Baichuan 2: open large-scale language models [EB/OL]. [2023-12-31]. https://arxiv.org/pdf/2309.10305v1.pdf

Yang J W, Shen X, Xing J, Tian X M, Li H Q, Deng B, Huang J Q and Hua X S. 2019. Quantization networks//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 7300-7308 [DOI: 10.1109/CVPR.2019.00748]

Yao J C, Wang F, Jia K Y, Han B, Zhou J R and Yang H X. 2021a. Device-cloud collaborative learning for recommendation//Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. Singapore, Singapore: ACM: 3865-3874 [DOI: 10.1145/3447548.3467097]

Yao J C, Zhang S Y, Yao Y, Wang F, Ma J X, Zhang J W, Chu Y F, Ji L, Jia K Y, Shen T, Wu A P, Zhang F D, Tan Z Q, Kuang K, Wu C, Wu F, Zhou J R and Yang H X. 2023. Edge-cloud polarization and collaboration: a comprehensive survey for AI. IEEE Transactions on Knowledge and Data Engineering, 35(7): 6866-6886 [DOI: 10.1109/TKDE.2022.3178211]

Yao L W, Huang R H, Hou L, Lu G S, Niu M Z, Xu H, Liang X D, Li Z G, Jiang X and Xu C J. 2021b. FILIP: fine-grained interactive language-image pre-training//Proceedings of the 10th International Conference on Learning Representations. [s. l.]: OpenReview.net: 1-21

Yu F X, Zhang W S, Qin Z W, Xu Z R, Wang D, Liu C C, Tian Z and Chen X. 2020. Heterogeneous federated learning [EB/OL]. [2023-12-31]. https://arxiv.org/pdf/2008.06767.pdf

Zeng A H, Liu X, Du Z X, Wang Z H, Lai H P, Ding M, Yang Z Y, Xu Y F, Zheng W D, Xia X, Tam W L, Ma Z X, Xue Y F, Zhai J D, Chen W G, Liu Z Y, Zhang P, Dong Y X and Tang J. 2022. GLM-130B: an open bilingual pre-trained model//Proceedings of the 11th International Conference on Learning Representations. Kigali, Rwanda: OpenReview.net: 1-56

Zhang J H, Chen S Q, Liu J T and He J X. 2023a. Composing parameter-efficient modules with arithmetic operation//Proceedings of the 37th Conference on Neural Information Processing Systems. New Orleans, USA: OpenReview.net: 1-22

Zhang Q R, Chen M H, Bukharin A, Karampatziakis N, He P C, Cheng Y, Chen W Z and Zhao T. 2023b. AdaLoRA: adaptive budget allocation for parameter-efficient fine-tuning [EB/OL]. [2023-12-31]. https://arxiv.org/pdf/2303.10512.pdf

Zhang R R, Han J M, Liu C, Zhou A J, Hu X F, Yan S L, Lu P, Li H S and Qiao Y. 2023c. LlaMA-adapter: efficient fine-tuning of language models with zero-init attention [EB/OL]. [2023-12-31]. https://arxiv.org/pdf/2303.16199.pdf

Zhang S S, Roller S, Goyal N, Artetxe M, Chen M Y, Chen S H, Dewan C, Diab M, Li X, Lin X V, Mihaylov T, Ott M, Shleifer S, Shuster K, Simig D, Koura P S, Sridhar A, Wang T L and Zettlemoyer L. 2022. OPT: open pre-trained Transformer language models [EB/OL]. [2023-12-31]. https://arxiv.org/pdf/2205.01068v1.pdf

Zhao B R, Cui Q, Song R J, Qiu Y Y and Liang J J. 2022. Decoupled knowledge distillation//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 11943-11952 [DOI: 10.1109/CVPR52688.2022.01165]

Zhao M D, Jain S and Song S R. 2023a. RoCo: dialectic multi-robot collaboration with large language models [EB/OL]. [2023-12-31]. https://arxiv.org/pdf/2307.04738.pdf

Zhao Q L, Wang J D, Zhang Y X, Jin Y Q, Zhu K J, Chen H and Xie X. 2023b. CompeteAI: understanding the competition behaviors in large language model-based agents [EB/OL]. [2023-12-31]. https://arxiv.org/pdf/2310.17512.pdf

Zhao Z H, Wallace E, Feng S, Klein D and Singh S. 2021. Calibrate before use: improving few-shot performance of language models//Proceedings of the 38th International Conference on Machine Learn-

ing. Virtual: PMLR: 12697-12706

Zhou L W, Palangi H, Zhang L, Hu H D, Corso J and Gao J F. 2020. Unified vision-language pre-training for image captioning and VQA//Proceedings of the 37th AAAI Conference on Artificial Intelligence. Washington, USA: AAAI Press: 13041-13049 [DOI: 10.1609/aaai.v34i07.7005]

Zhou X, Lei X Y, Yang C, Shi Y C, Zhang X and Shi J W. 2022. Handling data heterogeneity in federated learning via knowledge distillation and fusion [EB/OL]. [2023-12-31]. https://arxiv.org/pdf/2207.11447.pdf

Zhou X K, Xu X S, Liang W, Zeng Z and Yan Z. 2021. Deep-learning-enhanced multitarget detection for end-edge-cloud surveillance in smart IoT. IEEE Internet of Things Journal, 8(16): 12588-12596 [DOI: 10.1109/JIOT.2021.3077449]

Zhou Y M, Yang Y Z, Ying Q C, Qian Z X and Zhang X P. 2023. Multimodal fake news detection via clip-guided learning//Proceedings of 2023 IEEE International Conference on Multimedia and Expo. Brisbane, Australia: IEEE: 2825-2830 [DOI: 10.1109/ICME55011.2023.00480]

Zhu L C and Yang Y. 2020. ActBERT: learning global-local video-text representations//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Virtual: IEEE: 8743-8752 [DOI: 10.1109/cvpr42600.2020.00877]

Zhu X Y, Li J, Liu Y, Ma C and Wang W P. 2023a. A survey on model compression for large language models [EB/OL]. [2023-12-31]. https://arxiv.org/pdf/2308.07633.pdf

Zhu Y F, Niu C Y, Yan Y K, Cao Z J, Jiang H, Lyu C F, Tang S J and Wu F. 2023b. Device-unimodal cloud-multimodal collaboration for livestreaming content understanding//Proceedings of 2023 IEEE International Conference on Data Mining (ICDM). Shanghai, China: IEEE: #210 [DOI: 10.1109/ICDM58522.2023.00210]

Zhu Z D, Hong J Y and Zhou J Y. 2021. Data-free knowledge distillation for heterogeneous federated learning//Proceedings of the 38th International Conference on Machine Learning. Virtual: PMLR: 12878-12889

Zou A, Wang Z F, Carlini N, Nasr M, Kolter J Z and Fredrikson M. 2023. Universal and transferable adversarial attacks on aligned language models [EB/OL]. [2023-12-31]. https://arxiv.org/pdf/2307.15043v1.pdf

作者简介

王永威，男，研究员，主要研究方向为生成式人工智能、大小模型端云协同和人工智能安全。
E-mail: yongwei.wang@zju.edu.cn
吴飞，通信作者，男，教授，主要研究方向为人工智能、多媒体分析与检索、跨媒体计算。E-mail: wufei@zju.edu.cn
沈弢，男，博士研究生，主要研究方向为联邦学习。
E-mail: tao.shen@zju.edu.cn
张圣宇，男，研究员，主要研究方向为大小模型端云协同、跨模态计算和推荐系统。E-mail: sy_zhang@zju.edu.cn
吴帆，男，教授，主要研究方向为端云协同、无线网络与移动计算、数据管理与隐私保护。E-mail: fwu@cs.sjtu.edu.cn
赵洲，男，教授，主要研究方向为自然语言处理、计算机视觉、生成式人工智能。E-mail: zhaozhou@zju.edu.cn
蔡海滨，男，教授，主要研究方向为可信人工智能、分布式计算。E-mail: hbcai@sei.ecnu.edu.cn
吕承飞，男，资深技术专家，主要研究方向为端智能、3D/XR应用技术。E-mail: chengfei.lcf@taobao.com
马利庄，男，教授，主要研究方向为人工智能、数字多媒体。
E-mail: ma-lz@cs.sjtu.edu.cn
杨承磊，男，教授，主要研究方向为计算机支持的协同、工业图案智能设计。E-mail: chl_yang@sdu.edu.cn