

生成式人工智能遇到边缘智能时的集成微调和推理

Ning Chen, *Student Member, IEEE*, Zhipeng Cheng, *Member, IEEE*, Xuwei Fan, *Student Member, IEEE*, Xiaoyu Xia, *Member, IEEE*, Lianfen Huang

Abstract—高性能的生成式人工智能(GAI)代表了计算智能的最新演进，而未来6G网络的赐福也使得边缘智能(EI)充满发展潜力。GAI与EI之间的必然相遇可以释放出新的机遇，其中GAI基于大规模计算资源和大规模未标记语料库的预训练可以为EI提供强大的基础知识，而EI可以利用碎片化的计算资源为GAI聚合个性化知识。然而，自然矛盾特征给直接知识共享带来了巨大挑战。为了解决这个问题，本文提出了面向gai综合的网络(GaisNet)，一种协同的云-边-端智能框架，利用无数据的知识中继而缓冲矛盾，双向知识流实现了GAI的良性循环模型微调和任务推理，以无缝融合和协同进化实现GAI和EI之间的互惠共生。实验结果验证了所提机制的有效性。最后，讨论了GAI与EI相互作用未来面临的挑战和发展方向。

Index Terms—Computational intelligence, Generative AI, Edge intelligence, Integrated fine-tuning and inference, Hybrid federated split learning.

I. 简介

A. 动机

GENERATIVE 人工智能(artificial intelligence, GAI)作为人工智能生成内容(artificial intelligence-generated content, AIGC)的关键使能技术，已迅速发展并应用于自然语言处理(natural language processing, NLP)、计算机视觉(computer vision, CV)、代码生成、数学推理等各种任务[1]–[3]。表I显示了一些著名的GAI基础模型(FMs) [12]，其特点是参数规模大，基于大规模无标签语料库上的预训练进行公共知识学习，以及少样本微调的跨领域鲁棒性[4]。然而，研究表明，现有的集中式高质量语言数据预计将在2026年耗尽，甚至低质量的语言和图像数据将在2030~2050年耗尽，GAI面临高质量公共数据可用性有限的挑战[4]。因此，综上所述，虽然具有基础知识带来的强大性能，但高质量的法律数据匮乏和大规模参数带来的计算资源负担过重导致了互联网巨头垄断下的集中式发展，阻碍了GAI的多元化和民主化。

另一方面，与具有大规模参数的集中式GAI不同，边缘智能(edge intelligence, EI)更倾向于在用户周围部署灵活的轻量级模型，使计算智能更接近分布式终端数据[5]–[7]。根据最近的一份分析报告，2025年将有309亿个物联网(IoT)设备连接，数据规模预计将达到近79.4 zettabyte (ZB) [8]。同时，由于移动终端上中央处理器(CPU)、图形处理器(GPU)等芯片集成技术的不断升级，功耗成本不断降低，使终端设备具有一定的计算能力，用于轻量级的AI模型训练和推理[5]。因此，受益于边缘数据的增长

和终端设备计算能力的增强，在即将到来的B5G和6G时代，我们将见证网络范式从万物互联到万物智能的转变，原生AI将从遥远的云服务器下沉到网络边缘[9]。然而，即使EI更接近终端，有限的模型规模也会导致先验知识的缺乏，使得边缘训练和推理的效果不理想。

基于上述分析，可以将GAI和EI相结合，实现优势互补，二者的相互作用将派生出新的潜在增长能量。通过碎片化的计算资源利用个性化知识，EI可以缓解GAI公开数据短缺的困境，弥合计算智能与终端用户之间的距离，而GAI可以赋予EI预训练的广义基础知识，可将其作为EI加速学习收敛、提高推理性能的鲁棒基线[4], [9]。然而，如表II所示，由于参数大小、应用领域、网络架构、数据大小、资源提供等矛盾特征，使得GAI和EI之间的知识传递受到了极大的阻碍。

- 数据管道阻塞导致EI和GAI之间的上行知识中断。首先，受终端通信资源不足的限制，指数级增长的本地数据无法上传到云端，使得从EI到GAI的知识管道堵塞；此外，普通的GAI模型通常部署在大型企业或研究机构的中心云中，如Amazon cloud，并使用收集的海量数据进行预训练和微调。然而，为避免隐私泄露，6G终端设备通常不愿与云端的GAI模型共享本地数据，这也给GAI利用6G终端数据的个性化知识带来了很大障碍[10]。

- 模型管道阻塞导致GAI和EI之间下行知识中断。GAI依赖于在大规模基础数据集上预训练的数十亿参数的大规模模型，这需要巨大的计算和时间资源[3], [9], [11]。例如，训练一个超过1750亿参数的GPT-3模型需要1000个gpu运行超过4个月[9]。为了训练稳定扩散模型，Stability AI维护了4000多个NVIDIA A100 GPU集群，运营成本超过\$ 5000万[3], [11]。类似地，在任务推理场景中，内存需求随着参数数量的增加而急剧增加。例如，Falcon-40B需要大约86 GB的GPU内存来进行推理[12]。无论是训练还是推理，如此大的资源成本对于资源受限的6G终端来说都是非常负担的，导致基础知识从GAI到EI的转移管道堵塞。

面对上述挑战，本地保留6G终端设备数据并进行GAI模型分割的混合联邦分裂学习(HFSL)可以提供可行的解决方案，其中联邦学习(FL)是一种无需数据共享即可聚合终端个性化知识进行GAI的协作学习范式[4], [5], [13]，而拆分学习(SL)可以拆分GAI模型并在终端上部署资源适配的轻量级子模块来进行协同训练和推理[11], [13]。然而，由于GAI FMs的泛化特性，将其应用于任务推理前进行特定领域的模型微调至关重要，原始的FMs无法提供性能最优的GAI任务推理服务。同时，GAI从开发到应用的单向发展是一种糟糕且低效的模式，终端的增长无法反馈到初始模型。因此，有必要开发一种可持续演化的面向gaia的网络架构，既可以实现模型微调，又可以实现任务推理。

通讯作者:程志鹏(e-mail: chengzp_x@163.com)

陈宁, 樊旭伟, 黄连芬, 厦门大学信息学院, 厦门361005 (email: ningchen@stu.xmu.edu.cn; xwfan@stu.xmu.edu.cn; lfhuang@xmu.edu.cn)。

程志鹏就职于苏州大学未来科学与工程学院, 江苏苏州215006 (email: chengzp_x@163.com)。

夏晓宇就职于澳大利亚墨尔本皇家理工大学计算技术学院(e-mail: xiaoyushaw@gmail.com)。

TABLE I
一些代表性的GAI基础模型。

	Publisher	Architecture	Size	Data	Modality
GPT-3	OpenAI	Decoder	175 B	300 B (Tokens)	Text
GPT-4	OpenAI	Decoder	170 T	10 T (Tokens)	Text, Image
LLaMA	Meta	Decoder	7 B-65 B	1.4 T (Tokens)	Text
BERT	Google	Encoder	110 M, 335 M	250 B (Tokens)	Text
DALL-E	OpenAI	VAE, Decoder	12 B	2.5 B (Pairs)	Text, Image
DALL-E 2	OpenAI	CLIP, Diffusion	3.5 B	6.5 B (Pairs)	Text, Image

TABLE II
GAI与EI之间的矛盾特征。

	Size	Domain	Architecture	Data	Resource
GAI	Heavyweight	Generalized	Centralized	Massive	Sufficient
EI	Lightweight	Personalized	Distributed	Slight	Limited

B. 新颖性和贡献

为了解决上述问题, 首先, 充分尊重资源不平衡, 提出了gai-oriented 综合的网络 (GaisNet), 一个为GAI的模型微调和任务推理定制的云边端协同智能框架。据我们所知, 本文是最早从集成微调和推理的角度研究GAI和EI之间的相互作用的论文之一。主要贡献总结如下。

- 提出了一种云-边-端协同智能框架GaisNet, 其中特定领域的边模型作为数据无关的知识中继, 解锁GAI与EI之间的双向知识流动, 实现可持续演化的模型微调和任务推理。
- 本文列举并分析了在集成微调和推理的GaisNet操作中遇到的主要问题。
- 实验结果探讨了GaisNet的影响因素, 验证了所提机制的有效性。
- 最后讨论了GAI与EI相互作用未来面临的挑战和发展方向。

本文的其余部分组织如下。在第二节中, 我们进行了初步的介绍, 第三节提出了GaisNet的框架, 第四节讨论了GaisNet的主要问题, 第五节讨论和分析了模拟结果, 第六节讨论了GAI和EI相互作用的未来机遇和方向, 在第七节得出结论。

II. 开场白

A. 生成式AI与基础模型

受益于各种底层骨干的突破, 如生成对抗网络(GAN)、变分自编码器(VAE)、扩散模型(DM)和Transformer[11], 世界各地的技术巨头发布了各种强大的FMs, 即基于海量数据和计算资源进行预训练获得的通用GAI模型, 如OpenAI的GPT系列和DALL-E系列、Meta的LLaMA、谷歌的BERT和PaLM [1], [4], [12]。

GAI模型的生命周期由预训练、微调和推理组成。首先, 跨领域预训练赋予FMs广义基础知识; 然后, 特定领域的模型微调将预训练的输出转变为遵循人类意图的更窄的专业基础知识范围。最后, 任务推理是预训练和微调GAI模型 [9], [14]的应用和最终目标。下面我们讨论每个过程做什么以及它面临什么挑战。

1) 预训练是GAI模型的主要优化方法。在具有自监督表示的大规模无标记数据集上进行预训练, 使GAI模型具有

跨领域的广义基础知识。然而, 数十亿参数的预训练需要大量的计算和时间资源 [11], 这使得大多数资源受限的终端设备难以直接参与预训练过程, 造成边缘设备个性化本地数据的浪费 [9]。

2) 微调是预训练后对GAI模型的再优化。对不同垂直领域的微调使得原始GAI模型进一步增加了领域知识, 进一步提高了其对下游数据的适应性。更好地应对不同领域的下游任务[8]。GAI模型的微调方法可以分为更新所有骨干参数的全微调和只更新轻量级可调模块的参数高效微调。因此, 参数高效的微调可以通过冻结模型主干来减少可训练参数的数量, 从而缓解对存储内存和计算资源的压力。最先进的参数高效微调方法包括前缀调优、适配器调优、低秩自适应(LoRA)等 [15]–[18]。普通的集中式微调依赖于对用户数据的持续收集, 侵犯了用户的个人数据隐私 [9]。

3) 推理更偏向于优化后的GAI模型的应用。终端将未标注的数据(如提示指令、图片等)输入到预训练微调后的GAI模型中, 由GAI模型输出符合用户意图的内容, 如回复文本、绘制图像等, 以满足用户的服务需求 [5]。推理受到意图和模型的领域相关性以及网络环境 [9]的不确定性的影响。

因此, 模型微调是必要的, 以使预训练FM与人类意图保持一致, 并产生个性化输出, 而任务推理是测试GAI模型性能的基本方法。GAI的FM虽然高性能, 但其庞大的参数规模导致了头部企业主导的封闭式优化和应用, 使得GAI远离海量资源受限的终端客户端, 使得广泛分布的碎片化计算资源和个性化的本地数据无法得到充分利用。相反, EI可以将AI下沉到用户端, 使分布式客户端更多地参与AI相关的工作。

B. 边缘智能与分布式学习

云计算可以为各种终端应用提供足够的计算能力, 在人工智能模型的训练和推理中发挥重要作用 [8]。然而, 随着终端数据的日益增多, 基于全云计算的人工智能在满足自动驾驶、远程医疗等实时应用的延迟限制方面面临着重大挑战, 同时隐私泄露的风险也带来了极大的担忧。传统基于云计算的以数据为中心的模型训练和推理无法满足人工智能物联网(AIoT)应用激增的数据流量需求、普适计

算需求、严格的时延限制和个性化需求 [13]。针对上述问题, EI通过向用户端补贴计算智能的部署、训练和推理, 利用分散在网络边缘的计算资源和个性化数据, 为移动终端提供低时延的AI服务, 从而解决AI的“最后一公里”问题 [5]。同时, 由于著名的摩尔定律, 终端客户端的计算能力越来越强大, 可以支持机器学习(ML)任务的本地化运行, 为分布式学习的发展创造条件 [5]。

从数据处理的角度看, EI包括有数据聚合的集中式EI和无数据聚合的分布式EI。其中, 分布式EI的主要代表框架是FL、SL和HFSL, 能够在保护隐私的前提下实现数据知识的学习并完成推理任务 [4], [19]。FL是一种协作学习范式, 用模型聚合和分布代替本地数据传输, 可以在保护用户隐私的前提下间接利用分布数据的知识 [4], [5], [13]。SL可以进一步分解复杂的AI模型, 只部署与终端客户端资源 [11], [13]对等的轻量级部分。EI虽然建立了AI与终端本地数据之间的关联, 实现了更贴近用户、更易于访问的训练和推理过程, 但受限于多域物理资源的不足, 如模型训练和推理所需的计算资源和模型参数传输所需的通信资源。因此, EI很难获得高质量的模型训练和任务推理。在这种情况下, 云边端协同智能可以将云端的高性能模型与终端的轻量级模型有机地连接起来, 为解决上述问题提供了思路。

C. 协同云边端智能

在人工智能和分布式计算范式蓬勃发展的推动下, 将分布式人工智能与分层计算网络相结合迫在眉睫 [8]。此外, 随着宏基站和小蜂窝共存的5G网络等异构组网架构的出现, 开放了终端客户端、边缘接入点(如小蜂窝、路边单元等)和宏基站之间的通信管道, 构建了云-边-端协同智能的物理可行性。根据AI服务需求, 合理布局和利用分层异构的云-边-端网络的计算和通信资源, 利用云服务器丰富的计算资源和边缘服务器低的访问时延来支持计算密集型训练和时延敏感推理 [8]。云边端协同智能能够充分考虑不同组件上的资源不均衡性, 实现合理的工作分配。首先, 云层具有丰富的计算能力和存储资源, 可以实现高性能的模型训练和任务推理; 然后, 边缘层是连接云端和终端的桥梁, 可以实现低延迟的模型训练和任务推理; 最后, 端层是一个具有一定计算和存储能力且能够不断生成数据的数据源, 可以实现轻量级的模型训练和任务推理。

III. 集成微调和推理的GAISNET

A. 参数高效的微调和推理

参数高效的微调已被广泛研究, 其主要是从计算的角度进行高效的特定领域模型再训练。首先, 讨论了通过提示调优进行参数高效的微调。然后, 用类似的概念, 从通信角度提出了参数高效推理, 只传输小规模可调模块, 而保持主干的大规模参数冻结。从而降低了参数传输的通信开销。

1) 计算角度: 参数高效的微调

GAIS的成功依赖于骨干架构的发展和创新, 其中最著名的是谷歌在2017年提出的Transformer [1], [2], [18]。Transformer是许多最先进的FM的骨干, 如GPT-3、DALL-E 2、Codex和Gopher。变压器的基本结构如图1所示。为了简单和不失通用性, 本文讨论了基于Transformer的GAI模型作为基础模型。如图1所示, 基于Transformer编码器架构的预训练FM, 如BERT、ViT等, 可分为用于特征提取的嵌入层(Emb)、

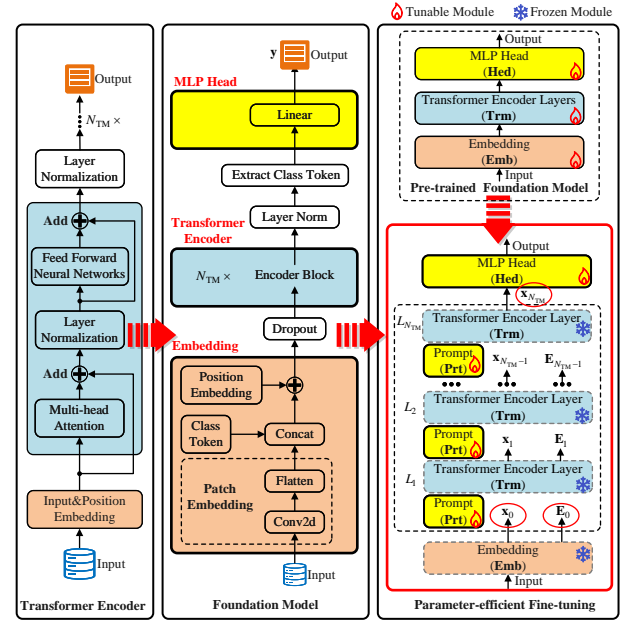


Fig. 1. 基于transformer的基础模型和参数高效微调。

用于基于自注意力的学习的多个Transformer层(Trm)和用于输出判断的多层感知器(MLP)头层(Hed) [20]。

将预训练FM模型应用于各种下游任务的关键步骤是微调 [15], [16]。作为一种具有代表性的参数高效微调用方法, 基于提示学习的前缀微调在每个Transformer层前添加了一些可学习的提示模块。与完全微调相比, 使用下游数据来训练提示模块并冻结整个预训练的Transformer骨干可以实现域性能增强, 就像“四盎司可以移动一千磅”。例如, visual-prompt tuning (VPT)提出了一种基于提示学习的视觉transformer (ViT)的有效参数-高效的微调方法, 在视觉任务中取得了理想的性能。在输入空间中只引入了少量(小于1%的模型参数)用于控制提示模块的可训练参数, 这些参数在微调期间与线性MLP头一起学习 [15], [18], [20]。

如图. 1所示, 考虑一个带有 N_{TM} Transformer编码器块 [20]的FM, 并将提示引入每个Transformer层的输入空间 [15]。定义 \mathbf{E}_0 和 \mathbf{x}_0 分别是输入数据补丁和可学习分类标记[CLS]的初始嵌入。因此, i -th Transformer层的输出标记可以表示为

$$[\mathbf{x}_i, \mathbf{E}_i] = \mathcal{L}_i([\mathbf{x}_{i-1}, \mathbf{E}_{i-1}]), 1 \leq i \leq N_{TM} \quad (1)$$

在 N_{TM} Transformer层的自注意力学习之后, 来自 N_{TM} -th Transformer层输出的可学习分类token被输入到MLP头部以输出最终结果, 即:

$$\mathbf{y} = \mathcal{H}(\mathbf{x}_{N_{TM}}) \quad (2)$$

最后, 经过损失计算和梯度反馈后, 在特定领域的数据上进行FM的微调过程, 进一步提升了FM在相关应用中的性能。

2) 沟通视角: 参数有效的任务推理

为提高人工智能的模型利用率和任务推理效率, 对于相似推理任务的服务需求, 基于参数共享的迁移学习可以实现微调后模型的共享应用。我们将全参数推理定义为完全

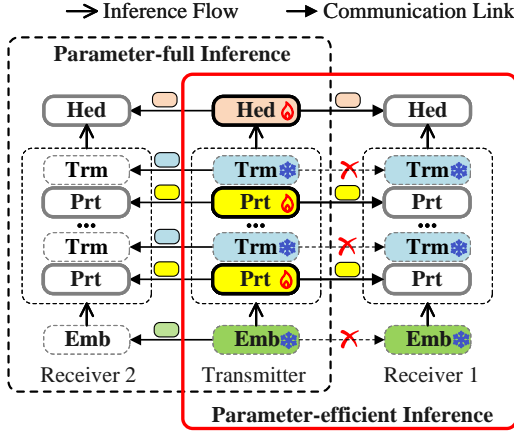


Fig. 2. 通信角度的参数有效推理。

共享所有参数和参数高效推理，其中只共享轻量级模块参数，从通信的角度定义，就像从计算的角度定义参数高效微调一样。

得益于参数高效的模型微调，在涉及模型分布和重配置的场景中实现了参数高效的推理。模型分布只涉及更新的轻量级模块，同时保持骨干冻结，这减少了由传输大规模模型参数或粉碎数据(例如，中间激活、反向传播梯度)引起的通信开销 [11], [13]。如图2所示，发射机与两个接收机共享微调模型。对于全参数推理场景，需要传输所有模型参数，包括大规模Transformer模块的参数。对于参数高效推理，只将微调过程中可以重新训练的MLP头和prompt模块的参数发送给接收端，大大降低了参数共享带来的高通信开销和推理时延。

B. 面向GaisNet的综合网络

图3显示了所提出的GaisNet的架构，这是一个用于GAI模型微调和任务推理的云-边-端协同智能框架，由部署在云服务器上的基础模型、部署在边缘服务器上的特定领域模型和部署在6G终端客户端的轻量级模型组成。GaisNet建立在云-边-端协同智能框架和D2D通信框架之上。分割后，边缘模型的可调模块被转移到客户端集群，其中嵌入层部署在起点，其输出为 $\mathbf{o}_s = [\mathbf{x}_0, \mathbf{E}_0]$ ，而MLP头层部署在终点，并将输出(即最终分类结果)定义为 $\mathbf{o}_e = \mathbf{c}_{c \in \mathcal{Y}}$ ，其中字母 $\mathcal{Y} = \{y_1, y_2, \dots, y_{|\mathcal{Y}|}\}$ 表示标签空间，如对象分类或识别任务中的类别集。此外，中间的 N_{TM} Transformer层部署在集群的其余客户端上。

基于云服务器的海量计算资源和大规模无标记语料库进行预训练，赋予FMs跨领域的广义基础知识，并通过微调将其转化为特定领域的专业基础知识，以更好地适应不同的下游任务。值得注意的是，本文关注的是参数高效的模型微调和任务推理，即只重新训练和传输预训练FMs的可调部分，并假设冻结的骨干(如大规模Transformer模块和用于特征提取的嵌入模块)始终在云服务器、边缘服务器和终端客户端中独立同步，这不在本文的优化考虑范围内。

边缘服务器以无数据知识中继的角色关联云服务器和端客户端，实现双向知识流进行模型微调和任务推理，其中云-边-端协同GaisNet进一步划分为具有跨领域大规模知识流的云-边子网络和具有特定领域小规模知识流的边-端子网络。

- 面向跨领域大规模知识流的云边子网络。采用FL框架在单个云服务器和多个边缘服务器之间进行跨领域的分布式学习，其中云服务器利用大规模未标记语料库和具有大量计算资源的全模型预训练，为GAI FMs提供泛化的基础知识，并在云模型交付的过程中将其迁移到边缘服务器的特定领域模型。同时，边缘服务器聚合从分布式终端客户端学习到的个性化知识以获取领域知识，在边缘模型参数上传的过程中将其传递到云服务器的FMs中，进一步提升FMs的跨域泛化性能。因此，云边子网的双向知识流动可以在FMs与特定领域模型之间形成良性闭环。

- 具有领域特定小规模知识流的边端子网络。由于边缘服务器的数据稀疏性以及客户端的隐私考虑和资源约束，为充分利用分布式客户端的数据和计算资源，利用客户端的轻量级模型联合执行基于hfsl的模型微调 and 基于sl的任务推理。边缘服务器将与领域相关的客户端与无蜂窝网络联系起来，并向其提供从云服务器获得的、针对领域数据进行微调的特定领域基础知识。同时，收集客户本地数据中的个性化知识，以获取新的领域知识。因此，边端子网的双向知识流动可以形成边缘领域特定模型与终端轻量级模型之间的良性共生闭环。

通过部署在云服务器上的FMs、边缘服务器上的领域专用模型和终端客户端的本地轻量级模型之间的协同，突破了EI向GAI数据阻塞、GAI向EI模型阻塞导致的知识流动阻塞，其中GAI通过高性能FMs为EI提供跨领域的基础知识，而EI通过分布式边缘模型为GAI提供领域专用的个性化知识。作为一种数据无关的双向知识中继，边缘服务器不仅可以传输从云服务器的大数据集中获取的大规模增强型流行知识，不仅可以为终端客户端提供特定领域的任务推理服务，而且可以以隐私保护的方式从终端客户端本地数据中收集个性化知识。因此，GaisNet可以实现良性闭环的GAI模型演化，进而实现可持续高效的模型微调和任务推理。

然后，由于边端子网更加复杂和具有代表性，重点研究了其两个子过程，即基于hfsl的模型微调和基于sl的任务推理。

C. 基于hfsl的模型微调

GaisNet通过基于fl的簇间并行协作和基于sl的簇内串行协作来保证GAI模型的微调。所提出的基于hfsl的模型微调框架如图4所示。首先，在FL框架中实现边缘服务器与多个客户端集群之间的并行协同边缘学习，以较低的通信开销实现无需数据共享的隐私保护和参数高效的模型微调，将微调客户端集群产生的终端个性化知识收敛到相关领域的边缘模型；然后，每个微调客户端集群内的客户端以SL模式完成串行协作，客户端之间以D2D通信的形式传输粉碎数据和可学习模块的梯度；GaisNet模型微调的工作流程包括边缘模型的分割与传递、感知数据的生成与嵌入、可调模块的计算与传输、终端模型的上传与聚合等过程，忽略了少量梯度的反馈。

1) 流程总结

- 边缘模型的分割与分布。边缘服务器将特定领域边缘模型的可调部分拆分为提示模块和一个MLP头模块。然后将上述子模块交付给不同的客户端集群，每个集群的成员承担边缘模型的完整可调模块。值得注意的是，当客户端计算资源充足时，可以将连续的提示模块划分在一起，发送到同一个客户端。如果只有一个客户端集群参与微调过程，HFSL会退化为SL。

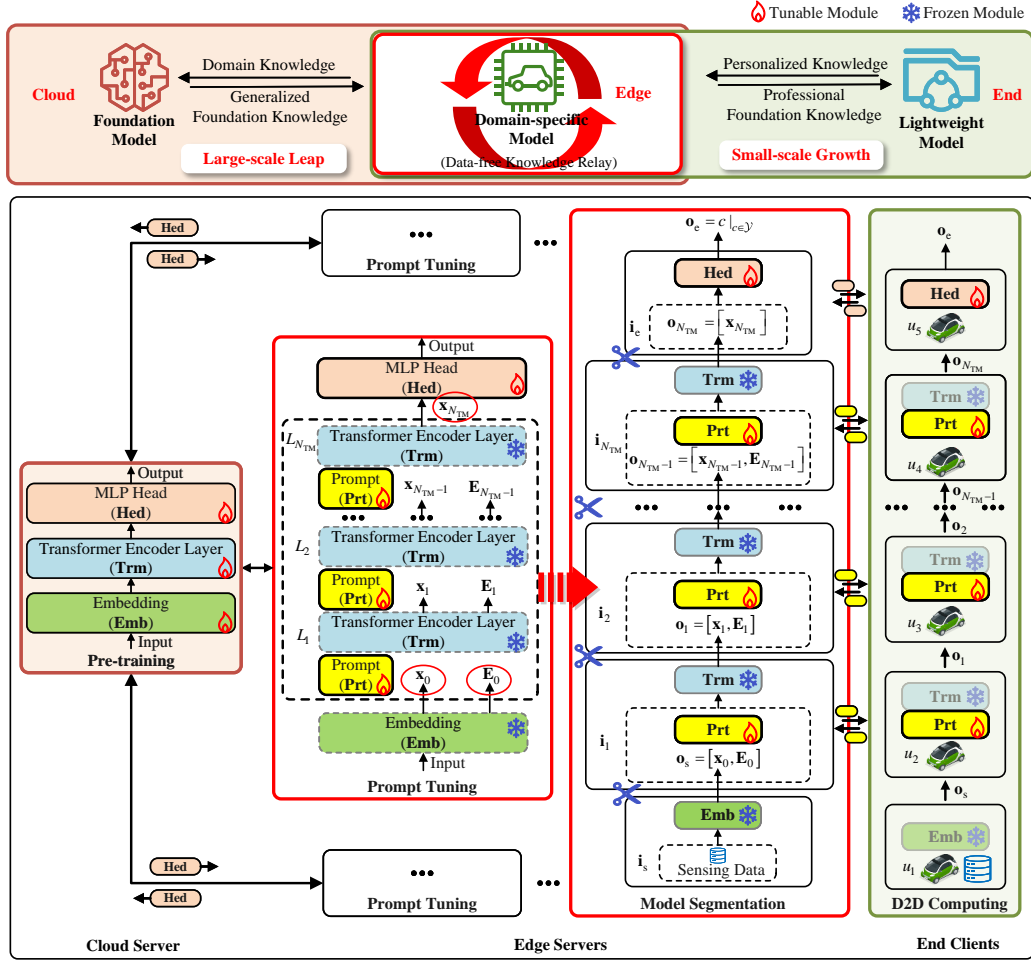


Fig. 3. GaisNet的体系结构。

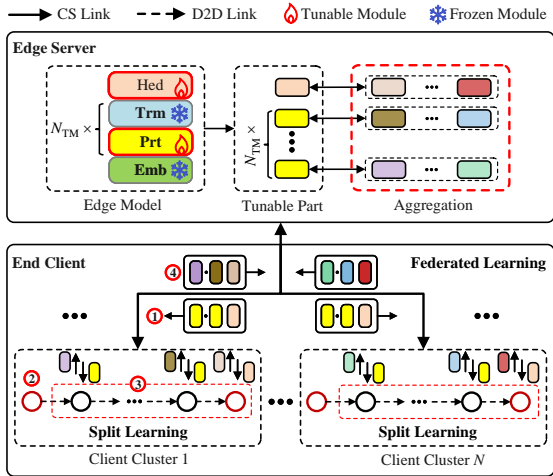


Fig. 4. 基于hfs1的模型微调框架。

- 训练数据的生成和嵌入。每个微调客户簇中的起始点基于感知过程获取个性化的带标签训练数据，然后基于局部同步嵌入层提取原始特征。
- 可调模块的计算和传输。提供数据的起始点的输出令

牌被传递给第二个客户端，然后微调客户端集群中的其余客户端利用与本地碎片计算资源的串行D2D通信完成其负责的提示模块的训练。它消耗通信资源来传输簇内节点之间的破碎数据，包括正向token和反向梯度。值得注意的是，当终端客户的计算资源有限时，边缘服务器也可以作为协作节点，部分微调任务可以在边缘服务器上执行。

• 终端模型的上传和聚合。所有微调的客户端集群将边缘模型的完整组件上传到边缘服务器，然后在不同集群的相同模块之间进行基于fedavg的参数聚合，包括相同数量的prompt模块和MLP头模块。聚合后更新边缘模型中的prompt模块和MLP头模块。

上述过程不断迭代，直到模型收敛或达到预定义的轮数。

2) 关键指标

总结了GaisNet [8]中基于hfs1的模型微调的几个关键指标。

- 模型性能是指微调后的模型执行任务推理的性能，如图像识别的准确率，是GaisNet模型微调和任务推理的根本目标和度量指标。
- 收敛延迟是模型微调所需的总延迟，包括边缘模型交付、感知数据生成、本地模型训练(包括基于D2D通信的参数传输)和本地模型上传的时间消耗。
- 计算成本是指在微调过程中参与模型训练的客户端所

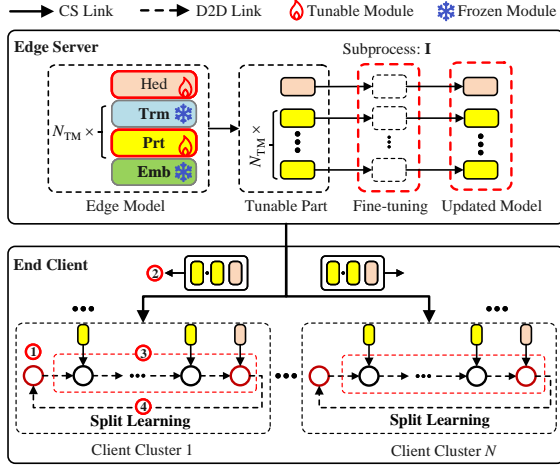


Fig. 5. 基于sl的任务推理框架。

消耗的计算能力的总和。

- 能耗成本包括整个微调过程的总能耗。与收敛延迟类似，涉及数据采集、参数传输、模型训练过程相关的能量消耗。

- 通信开销是微调过程中涉及数据传输的工作流对频谱资源的消耗，包括客户端-服务器(CS)链路和D2D链路的相关资源消耗。

- 内存占用是每個工作节点上用于在微调过程中临时存储中间参数和破碎数据的内存总和。

D. 基于sl的任务推理

GaisNet可以在推理客户集群内通过基于sl的串行协作进行任务推理，所提出的基于sl的任务推理框架如图5所示。具有任务推理服务需求的客户端将自己作为起点，并建立包含多个工作客户端的推理客户端集群。然后，相关域边缘服务器将微调聚合后更新的模块发送给工作客户端。接下来，与微调过程一样，推理客户端集群中的客户端基于D2D通信执行协作推理。最后，终点将MLP头模块输出的推理结果反馈到启动推理服务的起始点。

1) 流程总结

- 推理任务的生成与嵌入。每个推理客户簇中的起始点作为推理任务获取未标记的感知数据，然后基于本地同步嵌入层执行感知数据的原始特征提取。

- 边缘模型的分割与分布。边缘服务器将微调边缘模型的可调部分拆分为提示模块和一个MLP头模块。然后将上述子模块发送到推理客户端集群，其中除起始点外的其他成员承担完整边模型的推理任务。

- 可调模块的计算和传输。带有推断服务需求的起始点的输出令牌被传递给第二个客户端。然后，集群中剩余的客户端利用串行D2D通信连接的本地分片计算资源完成其负责模块的推理计算；值得注意的是，当终端客户端的计算资源有限时，边缘服务器也可以作为协作节点，部分推理任务可以在边缘服务器上执行。

- 推理结果的输出和反馈。推理客户端集群的端点输出推理结果，然后通过D2D通信链路发送回启动推理服务的起始点。

2) 关键指标

总结了GaisNet中基于sl的任务推理的几个关键指标。

- 模型性能是GaisNet模型微调和任务推理的基本目标和度量指标。

- 推理延迟是指从推理任务发起到接收推理结果的整个推理过程的总延迟。

- 计算成本是指参与推理过程的工作客户端所消耗的计算能力的总和。

- 能量代价包括整个推理过程的总能量消耗，包括数据采集、参数传输和模型推理过程。

- 通信开销是推理过程中涉及数据传输的工作流对频谱资源的消耗。

- 内存占用是每個工作节点上用于在推理过程中临时存储粉碎数据的内存数量，通常小于微调所需的内存。

IV. GAISNET的主要问题

A. 如何拆分模型？

无论是模型微调还是任务推理，边缘GAI模型中可调部分的划分难点在于确定分割块的数量和分割点的位置[8]。首先，需要根据能够提供协作的客户端来调整要分割的块数量，并且需要根据承担块的微调(推理)任务的相应客户端的资源比例来调整模型分割的块大小，即有多少可调模块被分割到同一个块中。同时，我们需要考虑工作客户端的计算资源，以及模型传递和上传所需的通信资源。

B. 如何聚类客户端？

根据所执行的GAI服务，将执行基于hfsl的模型微调的客户端形成的集群称为微调客户端集群，将执行任务推理的客户端集群称为推理客户端集群。两者都需要考虑客户端的计算资源存量、边缘服务器与客户端之间的可链接特性以及通信资源需求，因为边缘模型需要发送到终端客户端以及在哪里进行计算[13]。同时，集群中计算任务相邻的客户端在需要传输破碎数据时可以与D2D链路进行通信。特别地，对于微调客户端集群，我们需要考虑上传本地模型到边缘服务器所需的通信资源，而对于推理客户端集群，由于推理结果需要反馈给发起推理服务的客户端，因此需要满足端点和起点之间的D2D可链接性。除了上述过程之外，微调客户端集群的起始点还需要考虑用于训练的生成数据样本的可用性和质量[14]。

C. 微调还是推断？

考虑到边缘服务器和终端客户端资源有限的特点，简单地假设每轮只能完成一个GAI服务，该服务可以是模型微调服务或任务推理服务。GaisNet的模型微调决定了后续任务推理的性能上界，而推理结果是模型微调质量的关键衡量指标。以商品生产过程为例，模型微调可以理解升级为设备，而任务推理则是通过使用设备生产商品来创造价值。设备升级不能立即创造收入，但需要支付额外的费用。如果在后续回合中使用该设备生产商品，则可以增加产量，相应的GaisNet可以获得更好的任务推理性能。因此，模型微调侧重于提高未来收益，而任务推理决定眼前收益，我们需要实现合理的权衡以最大化长期累积收益[21]–[23]。

D. 它为谁服务？

通常，GaisNet中有多个边缘服务器发起的模型微调服务和客户端发起的任务推理服务。在决定微调或推理的基础上，由于资源有限，我们需要考虑哪个边缘模型执行微

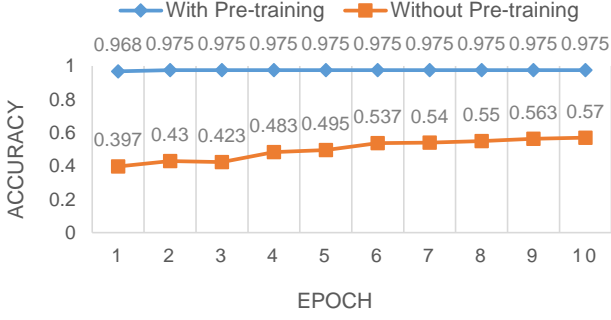


Fig. 6. 有预训练和无预训练的微调之间的精度比较。

调或哪个客户端执行任务推理。对于前者, 我们需要考虑客户端数据对不同领域边缘模型的适应性, 并进一步考虑未来的服务需求, 即在未来推理服务中哪个模型会被更多地使用, 选择市场需求最大、通过微调产生的性能收益最大的边缘模型进行微调。对于后者, 我们需要考虑推理结果与资源成本之间的关系, 并选择利润最大的客户端推理服务来执行。

V. 实验结果与讨论

在本节中, 为了证明GaisNet框架的有效性并探究相关影响因素, 我们将基于transformer的ViT-Base/16模型在花卉分类数据集 [15], [20]上作为案例研究。实验在NVIDIA RTX4060 GPU平台上进行, 学习率为0.001, 微调的批大小设置为10, 数据集的训练-验证比为4:1。

A. 预训练的影响

预训练可以为边缘模型和终端客户带来强大的先验基础。比较了基于预训练模型的微调与忽略云端预训练模型直接在客户本地数据集上进行微调的分类精度。如图6所示, 在使用预训练FM的条件下, 第一个epoch的准确率可以达到96.80%, 明显高于未进行预训练的收敛结果57.00%。

B. 微调的影响

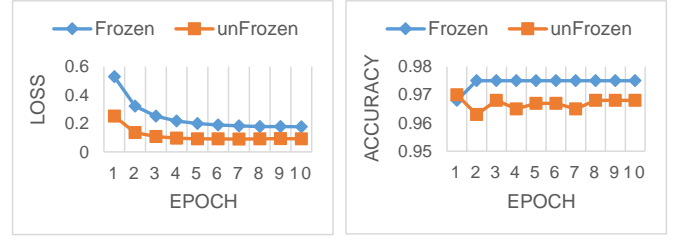
如图6所示, 当我们在边缘服务器上使用预训练模型时, 经过10个epoch的微调后, 模型的准确率从96.80%提升到97.50%, 实现了对本地数据的适应和性能提升。然而, 由于基线是在大规模云数据上进行预训练的, 少样本微调可以实现的性能提升是有限的。

C. 参数高效微调的影响

冻结GAI模型的主干可以显著减少用于微调的参数数量, 并减少计算资源开销。如图7所示, 与没有骨干冻结的全参数微调相比, 具有冻结骨干的参数高效微调方法在少样本微调后可以收敛到更高的精度。同时, 参数高效微调中每个epoch的训练时间约为35s, 而全参数微调需要3分30秒。

D. 非iid数据的影响

表III列出了在花卉分类数据集中使用不同数量的数据类进行微调的结果, 可以反映客户端的非iid数据对GaisNet中GAI模型收敛的影响。我们可以发现, 在微调数据量相同的情况下, 随着非iid程度的增加, GAI模型的性能明显下降。



(a)

(b)

Fig. 7. 使用frozen backbone和不使用frozen backbone进行微调的精度比较。

TABLE III
数据类的数量对GaisNet的影响。

Num Class	1	2	3	4	5
Accuracy (First/End)	0.200/ 0.200	0.397/ 0.397	0.575/ 0.595	0.753/ 0.767	0.933/ 0.967

E. 客户端集群数量的影响

表IV显示了参与微调的客户集群数量对边缘模型收敛精度的影响。由于更多的聚类会带来更多个性化的局部数据, 因此随着聚类数量的增加, 微调的收敛精度逐渐提高。然而, 由于GAI的小样本微调特性, 数据规模的增加限制了精度的提升。

F. 集成微调和推理的影响

本文采用第四节提出的商品生产和设备升级的概念, 研究了综合微调和推理对GaisNet中GAI服务的影响。假设有3个设备(对应于3个边缘模型)产生3个商品(对应于3个边缘模型可以提供的推理服务)。在每一轮中, 我们可以选择升级其中一个设备, 表示为a、b和c, 或选择一个设备来产生商品, 表示为a、b和c。我们将所提出的最大长期累积利润(MLCP)方法与集成的微调和随机选择(RS)推理和最大短期立即利润(MSIP)方法进行比较, 仿真结果如表V所示。图8显示了GAI轮累计利润的变化。我们可以发现, 所提出的MLCP算法在第2轮和第3轮中选择牺牲当前利润来升级设备c, 以在后续回合中获得更多的利润, 从而获得最大的长期累积利润。

VI. 挑战与未来方向

A. 隐私问题和社会问题

随着GAI在各个领域的不断应用, 其隐私和社会问题逐渐引起人们的关注。首先, 预训练、微调和推理过程中涉及的隐私泄露和数据安全问题值得关注。第二, 要重视社会对偏见、道德、知识产权侵权等问题的关注。因此, 有必要研究针对GAI的隐私保护策略, 如差分隐私, 加强立法监管, 并开发能够增强GAI模型溯源性的技术, 如区块链技术 [1]。

B. 资源约束下GAI的理论界

多域物理资源以及数量和质量参差不齐的数据限制了GAI的性能。与Shannon的信息论类似, 在有限的资源和数据范围内研究和描述GAI的性能边界对于GAI模型的开发和部署具有重要意义。

TABLE IV
客户端集群数量对GaisNetI的影响。

Num Cluster	1	2	3	4	5	6
Accuracy (First/End)	0.930/ 0.950	0.940/ 0.955	0.943/ 0.960	0.950/ 0.963	0.966/ 0.974	0.968/ 0.975

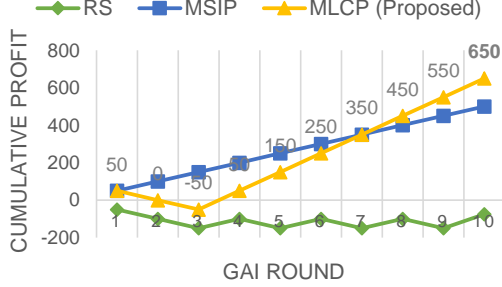


Fig. 8. 累计利润的比较。

C. 激励机制设计

有效公平的激励机制对于鼓励6G终端设备广泛参与GAI模型的优化至关重要。综合考虑资源贡献和数据差异，确保每个参与者创造的价值与获得的回报相匹配[4]。可以使用的经济学数学工具包括博弈论、拍卖理论、契约理论和其他理论[9]。

VII. 结论

文中提出了GaisNet，一种面向GAI的云边端协同智能框架，疏通了GAI和EI之间的双向知识管道，在统一的架构上实现了高效的基于hfsI的模型微调和基于sl的任务推理。然后分析了GaisNet运行过程中的主要问题，并通过仿真研究了各种影响因素对GaisNet的影响；最后，展望了GAI与EI相互作用未来面临的挑战和发展方向。

REFERENCES

- [1] Y. Cao, S. Li, Y. Liu, Z. Yan, Y. Dai, P. Yu, L. Sun, "A comprehensive survey of ai-generated content (aigc): A history of generative ai from gan to chatgpt," *arXiv preprint arXiv:2303.04226*, 2023.
- [2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, vol. 30, 2017.
- [3] H. Du, Z. Li, D. Niyato, J. Kang, Z. Xiong, D. Kim and others, "Enabling AI-generated content (AIGC) services in wireless edge networks," *arXiv preprint arXiv:2301.03220*, 2023.
- [4] W. Zhuang, C. Chen, L. Lyu, "When foundation model meets federated learning: Motivations, challenges, and future directions," *arXiv preprint arXiv:2306.15546*, 2023.
- [5] G. Zhu, Z. Lyu, X. Jiao, P. Liu, M. Chen, J. Xu, S. Cui, P. Zhang, "Pushing AI to wireless network edge: An overview on integrated sensing, communication, and computation towards 6G," in *Science China Information Sciences*, vol. 66, no. 3, pp. 130301, 2023.
- [6] X. Xia, F. Chen, Q. He, J. Grundy, M. Abdelrazek, and H. Jin, "Online collaborative data caching in edge computing," *IEEE Transactions on Parallel and Distributed Systems*, vol. 32, no. 2, pp. 281-294, 2020.
- [7] X. Xia, F. Chen, Q. He, J. Grundy, M. Abdelrazek, and H. Jin, "Cost-effective app data distribution in edge computing," *IEEE Transactions on Parallel and Distributed Systems*, vol. 32, no. 1, pp. 31-44, 2020.
- [8] S. Duan, D. Wang, J. Ren, F. Lyu, Y. Zhang, H. Wu, X. Shen, "Distributed artificial intelligence empowered by end-edge-cloud computing: A survey," *IEEE Communications Surveys & Tutorials*, 2022.

- [9] X. Huang, P. Li, H. Du, J. Kang, D. Niyato, D. Kim, Y. Wu, "Federated Learning-Empowered AI-Generated Content in Wireless Networks," *arXiv preprint arXiv:2307.07146*, 2023.
- [10] Z. Zhang, Y. Yang, Y. Dai, Q. Wang, Y. Yu, L. Qu, Z. Xu, "FedPETuning: When federated learning meets the parameter-efficient tuning methods of pre-trained language models," *Association for Computational Linguistics (ACL)*, pp. 9963-9977, 2023.
- [11] Y. Tian, Y. Wan, L. Lyu, D. Yao, H. Jin, L. Sun, "FedBERT: When federated learning meets pre-training," in *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 13, no. 4, pp. 1-26, 2022.
- [12] H. Zou, Q. Zhao, L. Bariah, M. Bennis, M. Debbah, "Wireless multi-agent generative ai: From connected intelligence to collective intelligence," *arXiv preprint arXiv:2307.02757*, 2023.
- [13] Z. Lin, G. Qu, X. Chen, K. Huang, "Split Learning in 6G Edge Networks," *arXiv preprint arXiv:2306.12194*, 2023.
- [14] A. Agarwal, M. Rezagholizadeh, P. Parthasarathi, "Practical Takes on Federated Learning with Pretrained Language Models," *Findings of the Association for Computational Linguistics: EACL 2023*, pp. 454-471, 2023.
- [15] M. Jia, L. Tang, B. Chen, C. Cardie, S. Belongie, B. Hariharan, S. Lim, Ser-Nam, "Visual prompt tuning," *European Conference on Computer Vision*, pp. 709-727, 2022.
- [16] J. He, C. Zhou, X. Ma, T. Berg-Kirkpatrick, G. Neubig, "Towards a unified view of parameter-efficient transfer learning," *arXiv preprint arXiv:2110.04366*, 2021.
- [17] E. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, "Lora: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021.
- [18] J. Chen, W. Xu, S. Guo, J. Wang, J. Zhang, H. Wang, "FedTune: A Deep Dive into Efficient Federated Fine-Tuning with Pre-trained Transformers," *arXiv preprint arXiv:2211.08025*, 2022.
- [19] Z. Cheng, X. Xia, M. Liwang, X. Fan, Y. Sun, X. Wang, L. Huang, "CHEESE: distributed clustering-based hybrid federated Split learning over edge networks," *IEEE Transactions on Parallel and Distributed Systems*, 2023.
- [20] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly and others, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.
- [21] X. Li, S. Bi, H. Wang, "Optimizing resource allocation for joint AI model training and task inference in edge intelligence systems," in *IEEE Wireless Communications Letters*, vol. 10, no. 3, pp. 532-536, 2020.
- [22] A. Eshratifar, M. Abrishami, M. Pedram, "JointDNN: An efficient training and inference engine for intelligent mobile cloud computing services," in *IEEE Transactions on Mobile Computing*, vol. 20, no. 2, pp. 565-576, 2019.
- [23] N. Chen, Z. Cheng, X. Fan, B. Huang, X. Du, and G. Mohsen, "Integrated Sensing, Communication, and Computing for Cost-effective Multimodal Federated Perception," *arXiv preprint arXiv:2311.03815*, 2023.

TABLE V
服务决策和好处。

[illegible]