

大云模型和小边模型协同提供生成式人工智能服务的边云协同框架

Yuqing Tian, *Student Member, IEEE*, Zhaoyang Zhang, *Senior Member, IEEE*,
Yuzhi Yang, *Student Member, IEEE*, Zirui Chen, *Student Member, IEEE*, Zhaohui Yang, *Member, IEEE*,
Richeng Jin, *Member, IEEE*, Tony Q. S. Quek, *Fellow, IEEE*, and Kai-Kit Wong, *Fellow, IEEE*

Abstract—生成式人工智能(GenAI)通过内容创造为用户提供各种服务, 被认为是未来网络中最重要的组成部分之一。然而, 训练和部署大型人工智能模型(BAImS)会引入大量的计算和通信开销。由于高性能计算基础设施的需求以及云服务远程访问的可靠性、保密性和时效性问题, 集中式方法面临严峻挑战。因此, 迫切需要分散服务, 部分将它们从云端移动到边缘, 并建立原生GenAI服务, 以实现私人、及时和个性化的体验。文中提出了一种全新的自底向上的大云模型和小边缘模型协同的BAImS架构, 并设计了分布式训练框架和面向任务的部署方案, 以高效提供原生GenAI服务。所提框架可以促进协同智能、增强适应性、聚集边缘知识和减轻边云负担。通过一个图像生成用例证明了所提出框架的有效性。最后, 概述了基础研究方向, 以充分挖掘原生GenAI和BAImS应用的边缘和云协作潜力。

Index Terms—Generative AI, big AI model, cloud-edge collaboration.

I. 简介

生成式人工智能(GenAI)是一种自动化方法, 它探索数据结构和特征, 以生成类似于人类创造的材料的内容[1]。GenAI与用户交互以提供个性化服务, 包括生成图像、文本和视频。GenAI的发展, 如GPT-4这样的大型语言模型(LLM), 提高了各种任务中的服务质量(QoS)和体验质量(QoE)。然而, 大规模GenAI的涌现能力是以集中式云服务的计算和通信资源消耗为代价的。与此同时, 第五代(5G)向第六代(6G)通信网络正在从互联智能转向协同智能[2], 其中大人工智能模型(BAImS) [3]和小边缘模型合作提供服务。在预期的系统中, 云服务器通过集成具有不同任务的小边缘模型来维护统一的BAImS。训练后, 增强的BAImS可以提取与任务相对应的小模型, 促进边缘部署, 并能够提供高性能、低延迟的原生GenAI服务。

为了解决适应性、边缘知识获取和云开销等问题, 需要具有分布式模型训练范式的可扩展BAImS架构。

适应性: 统一的BAImS需要能够满足所有用户的需求。为了管理不断增加的用户、服务多样性和应用程序复杂性, BAImS架构应该是可扩展的。此外, 在现实世界的系统中, 边缘在通信、计算和存储能力方面是异构的。此

外, 节点之间的连接是不稳定的, 边缘节点可能中途加入或离开。因此, 可扩展的BAImS应该能够适应异构模型聚合和动态网络。

2)收集边缘知识: BAImS在各种领域都有丰富的训练数据, 表现优于较小的模型。在6G网络中, 单个设备产生的数据往往不足以训练高质量的模型。为了将它们聚集在一起, 边缘模型可以提取局部智能并将其转移到中心。这有助于在不直接访问原始数据的情况下获取全球知识, 并支持高质量BAImS的开发。

3)减轻云负担: 集中式BAImS训练面临着日益增长的数据存储、模型参数缓存和计算成本需求。与用户的频繁交互增加了通信负担, 对中心服务器的能力提出了额外的挑战。同时, 边缘网络为模型训练提供了大量的计算资源。分布式BAImS训练有效地利用了计算能力、存储和通信的边缘资源, 使过程更加环保和成本高效。

GenAI服务的分布式部署是另一个研究热点, 旨在提供安全、及时和个性化的服务。

1)数据安全: 许多GenAI服务, 如自动驾驶和远程医疗, 需要收集真实用户数据。集中式云计算要求用户将所有数据上传到云端, 引发了隐私问题。将本地GenAI部署在接近或直接位于数据源的地方, 可以将数据存储在本地服务器或用户设备上, 从而缓解共享敏感数据的需求。

2)响应的时效性: 与判别式AI相比, GenAI在响应用户请求时生成了大量的数据。依赖长距离传输的云服务在将这些数据交付给用户时可能会出现明显的延迟。原生GenAI具有高效的本地通信, 可以实现高吞吐量和低延迟任务。

3)个性化服务: 为了响应用户请求, 边缘服务器可以从云端下载具有必要功能的GenAI轻量级版本, 可以在本地数据集上进一步微调。此外, 通过对具有类似服务需求的用户进行分组, edge可以维护多个模型, 以高效处理各种任务 and 应用程序。

为了增强6G网络中用户的QoE和QoS, 同时利用BAImS [3]的优势和边缘服务[4]的优势至关重要。本文提出了一种将原生GenAI与基于云的BAImS集成的协作方案, 提供了一种潜在的解决方案。首先分析了当前边云协作中的人工智能训练和部署策略, 证明了其局限性。总结了限制BAImS分布式训练和原生GenAI部署的挑战。在这种背景下, 提出了一种自底向上的BAImS架构, 以及一个分布式训练框架和面向任务的BAImS部署解决方案。在这个框架中, 通过一个图像生成用例展示了它对改进服务提供的贡献。最后, 概述了充分挖掘原生GenAI和BAImS合作潜力的基础研究方向。

II. 基于边云协作的模型训练和部署概述

在本节中, 我们概述了在第三代伙伴关系项目(3GPP)

基金资助:国家自然科学基金项目(U20A20158);国家重点研发计划项目(2020YFB1807101);浙江省重点研发计划项目(2023C01021)。(通讯作者:张昭阳)

田 勇(e-mail: tianyq@zju.edu.cn)、张 正(e-mail: ning_ming@zju.edu.cn)、杨勇(e-mail: yuzhi_yang@zju.edu.cn)、陈正(e-mail: zirui.chen@zju.edu.cn)、杨振宇(e-mail: yang_zhaohui@zju.edu.cn)、金荣(e-mail: richengjin@zju.edu.cn)与浙江大学信息科学与工程学院(浙江杭州310027)、浙江省信息技术重点实验室合作。Proc., Commun. & 网络。工业信息工程学院, 浙江杭州310027

t.q. S. Quek(电子邮件:tonyquek@sutd.edu.sg)是新加坡科技与设计大学(SUTD) ISTD支柱, 新加坡487372, 也与SUTD- zju网络智能创意中心, 新加坡487372。

王家强(e-mail: kai-kit.wong@ucl.ac.uk)就职于英国伦敦大学学院电子与电气工程系。

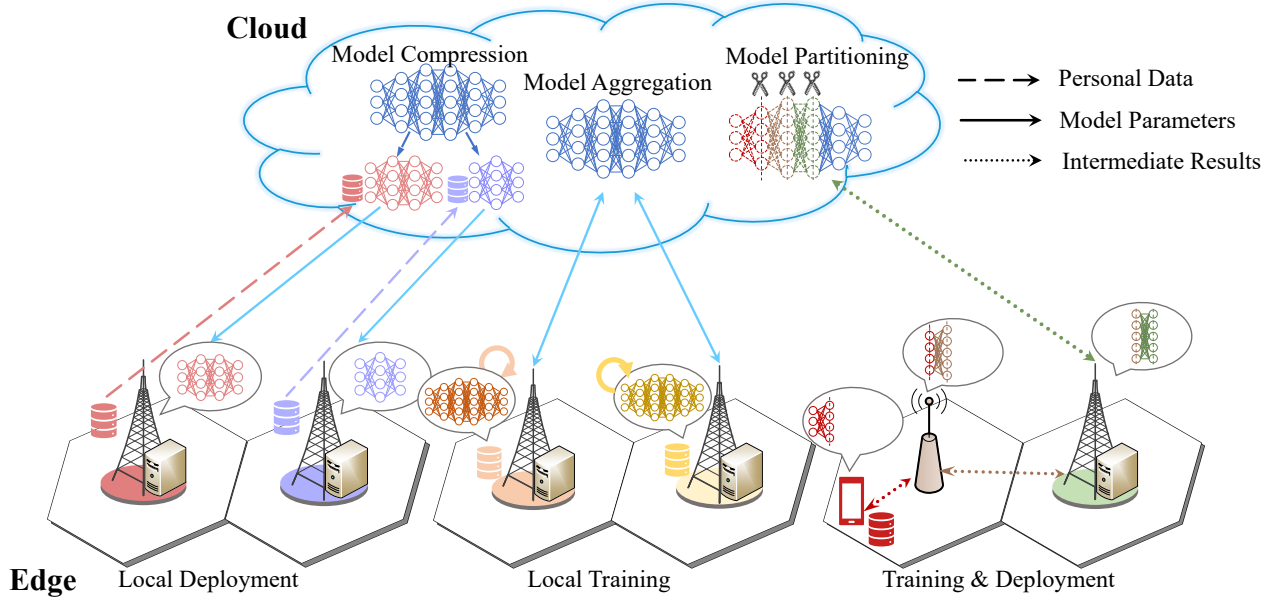


Fig. 1. 三个用于训练和部署人工智能模型的云边协作框架。模型压缩遵循集中式方法，通过KD等方法产生专门用于个人数据集的较小模型。模型聚合在云中迭代过程中合并边缘训练的模型。模型划分涉及通过将模型划分到不同的节点来联合训练和部署模型。

TABLE I. 分布式模型训练和部署框架

(a) Model Configuration and Deployment Phase

Distributed Paradigms	Basic Structure	Model Size		Deployed Model	Transmission Content in Deployment Phase	Inference Latency and Cost	
		Cloud	Edge			Comput.	Commun.
Knowledge Distillation	Teacher-Student	Big	Adaptive	Independent and Personalized	-	Low	-
Federated Learning	Top-down	Big	Big	Independent	-	High	-
Split Learning	Multi-Partition	Adaptive	Adaptive	Dependent and Cooperative	Intermediate Results	Medium	High
Ours	Bottom-up	Big	Adaptive	Independent and Personalized	-	Low	-

(b) Training Phase

Distributed Paradigms	Data		Training Method		Transmission Content in Training Phase	Training Latency and Cost	
	Cloud	Edge	Cloud	Edge		Comput.	Commun.
Knowledge Distillation	Personal Data	Personal Data	From Scratch	With Knowledge Logits	Personal Data	High	High
Federated Learning	-	Personal Data	Averaging	From Scratch	Model Parameters	High	High
Split Learning	Labels / -	Personal Data	From Scratch	From Scratch	Intermediate Results	Medium	High
Ours	Common Data	Personal Data	Fine-tuning	From Scratch	Model Parameters	Low	Low

SA1第18版[5]中探索的具有边云协作的AI模型训练和部署框架。如图. 1所示，这些分布式AI框架包括模型压缩(以知识蒸馏(KD)为演示)、模型聚合(以联邦学习(FL)为代表)和模型分割(也称为分离学习(SL))。我们将这些框架与我们在表I中提出的自底向上的BAIM架构进行了比较，强调了现有的局限性，并总结了阻碍BAIM的分布式训练和部署的挑战。

A. 模型压缩

考虑到边缘节点的资源约束，将整个经过云训练的BAIM加载到边缘或用户设备上推理通常是不切实际的。因此，模型压缩可以减少推理的模型大小和计算成本，可以部分解决资源有限的挑战。模型压缩涉及各种成熟的技术，如剪枝、量化、低秩近似和KD。

以KD为例，描述了模型压缩技术从训练到部署的生命周期。KD基于相同的训练数据集，将知识从一个大的教师模型迁移到更小的学生模型。然后，训练好的学生模型被独立部署在边缘节点上，因为较小的模型的评估成本更

低。目前，KD的应用已经从分类任务扩展到生成任务，并在LLMs [6]上表现出了卓越的性能。

虽然模型压缩显著提高了资源受限环境中的部署效率，但也引入了一些问题，包括信息丢失、泛化能力差和训练成本增加。训练阶段给中心服务器带来了更大的负担，并且无法解决分布式数据源的问题。

B. 模型聚合

模型聚合是另一种将边缘模型的信息集成到全局模型中的机制。最具代表性的算法是FL，在FL框架中，云服务器初始化模型并将其分发到每个边缘节点。然后，每个边缘节点利用其本地数据进行模型训练；然后，在云服务器上收集并聚合模型参数，形成全局模型。该过程迭代执行多轮。值得注意的是，这是在节点之间不共享原始数据的情况下实现的，从而确保了数据的隐私和安全。

FL有几个好处，包括加强隐私保护，提高边缘计算效率和增强模型泛化能力。然而，该模型聚合是针对同构模型设计的，没有考虑边缘节点之间的异构性。此外，由于资源有限，聚合模型难以在设备上部署。此外，FL需要在云

端和边缘之间频繁交换模型参数,这对于大规模模型来说是不切实际的。这些交换代价高昂,阻碍了频繁的传输,特别是在通信资源有限的情况下[7]。

联邦蒸馏(FD)为解决这些挑战提供了一种解决方案。在FD中,用户只交换模型的中间输出,中间输出的规模要小得多。每个边缘节点可以根据自己的能力初始化和训练本地模型,同时存储中间结果,如分类任务中的logits。这些中间结果定期上传到云端进行聚合,形成全局知识。然后,边缘节点下载这些全局知识并将其纳入其局部训练中,将其模型输出与全局模型的输出对齐。FD不受同构模型聚合的限制,有效降低了通信开销。然而,FD和FL的模型训练的主要计算负荷是由边缘处理的,给这些节点带来了很大的负担。相比之下,中心服务器的聚合任务相对简单,导致计算任务分布不均衡。

C. 模型划分

模型划分,通常称为SL,是另一种分配模型计算任务的方法。这种方法通常应用于无法完全容纳在单个设备或节点上的大规模模型。在SL系统中,模型的结构和参数被划分为多个分区,由通信网络中的不同节点进行计算。这有助于平衡多个节点的计算负载。此外,SL不要求用户共享他们的原始数据,相反,他们交换中间结果或标签。这种方法确保了隐私保护,并减少了交换原始数据所需的通信带宽[8]。通常,在模型部署阶段和模型训练过程中都可以使用SL。

在模型部署阶段,建立训练好的模型的拓扑划分是至关重要的,需要考虑以下几个方面。1)边缘-云节点信息:这涉及到考虑每个节点的通信、计算和存储能力。这些信息对于确定模型在整个网络中的最佳分布和执行至关重要。2)每层输出的大小:检查每一层产生的数据的大小是必要的。这有助于确定在给定的特定层划分模型时需要发送的数据量。3)计算代价和通信代价的权衡:为了通过输出更小的层来减少通信成本,通常需要在功能较弱的设备上执行更多的计算。因此,通过损失函数实现折中对于获得合适的划分方案至关重要。之前的工作采用了一种联合模型分裂和神经架构搜索的方法来确定分割的模型[9]。这可以在给定的通信网络中确保最佳的任务性能和延迟。

在模型训练过程中进行语言学习和设计数据传输的通信策略,可以缓解由于不完美通信造成的模型性能下降。通过将over-air计算框架与SL相结合,并利用无线信道的互易性,数据传输可以无缝地集成到模型层之间的计算过程中。这种整合有助于减少传输过程中的资源支出[10]。此外,FL与SL的结合利用了来自各种边缘节点的数据。通过云边协同,该方法有效降低了边缘节点的计算负载,提升了网络的整体效率[11]。

D. 关键绩效指标(kpi)

在边云协同框架中为用户提供服务需要仔细考虑各种kpi。图2展示了服务延迟、成本、存储、可靠性与稳定性、安全与隐私、通信效率六大kpi。此外,在分布式架构中,成本和存储是分别考虑边缘和云端的。在边缘和云端之间有一个权衡,因此,成本和云端存储位于雷达图的中线上,有助于评估系统特性(绿色区域)和开销节省(黄色区域)。

如图2 (a)所示,由于KD要求用户上传数据到中心进行模型训练,因此破坏了系统的安全性;图2 (b)显示,由于

边缘模型训练,FL给边缘带来了巨大的开销,以及多次模型传输导致的通信效率低下。从图2 (c)可以看出,由于中间结果的传输以及依赖节点连通性造成的不可靠性,SL造成了很高的服务延迟。图2 (d)是我们提出的解决方案,在各项kpi上都有很好的表现。

III. 自下而上的BAIM体系结构:分布式训练和面向任务的部署

在本节中,我们将介绍自下而上的BAIM架构,它利用边缘-云协作进行分布式培训和面向任务的部署。首先概述了该框架的工作流程,包括BAIM的训练过程和原生GenAI服务的生命周期过程。描述了该架构,强调其复杂的设计,使分布式训练和自然分区的部署方案成为可能。探索了它在云端的训练过程,这对于在训练数据很少的情况下进行泛化至关重要。最后,提出了一种基于任务特定划分的部署策略,使原生GenAI能够在边缘节点上动态部署BAIM。这允许用户获得BAIM的性能增强和edge services提供的改进的QoE。

A. 框架的工作流程

我们描述了所提出框架的工作流程,如图3所示。

1) BAIM培训流程:首先,用户上传共享的个人数据到边缘,构建本地数据集;具有隐私敏感数据的用户充当边缘设备,维护敏感的个人数据集。其次,边缘节点根据自身能力和用户规模,针对各自任务初始化生成模型并基于本地数据集进行训练;由于边缘特征明显,训练后的模型可能表现出异构的结构和特征。然后,边缘节点上传训练好的模型,使云端获得多个边缘模型进行多任务多模态学习。云通过门控神经网络编排边缘模型,并在不同边缘模型的阶段之间建立线性投影连接,从而构建自底向上的BAIM架构。然后,整个BAIM基于云公共数据集进行训练,在多个任务中实现卓越的性能。随后,BAIM很容易根据任务进行划分,产生紧凑的特定于任务的模型。最后,边缘节点可以利用其本地数据集对返回的轻量级模型进行个性化微调。

2) 原生GenAI服务生命周期:首先,用户根据需求向边缘提交查询,上传所需的服务数据;然后,边缘检查它的本地工具箱中请求的模型。如果找到,则直接对用户数据进行推理并返回结果。否则,从云端请求并下载相应的模型。如果用户服务涉及敏感个人数据,用户可以直接从云端获取相应的任务模型。

B. 自下而上的BAIM架构

在通信系统中,BAIM的集中式架构限制了获取高质量用户数据的能力。受Pathways [12]和混合专家(MoE) [13]的启发,我们提出了一种自下而上的BAIM架构。该架构最大限度地利用边缘模型提取的用户数据和专家知识。由谷歌Mind推出的Pathways代表了以多任务、多模态和稀疏激活为特征的下一代网络架构。认为统一模型应该能够通过激活相应的模块来加速利用现有技能的学习。MoE是GPT-4的基本结构,利用门控神经网络将多个专家进行组合,实现了专家输出的自适应组合。该设计利用了来自不同专家的知识,同时通过稀疏门控降低了计算需求。将边缘模型作为MoE专家,通过建立它们之间的线性联系将模型模块化。这形成了一个多任务、多模态和稀疏激活的分层BAIM架构,如图4所示。

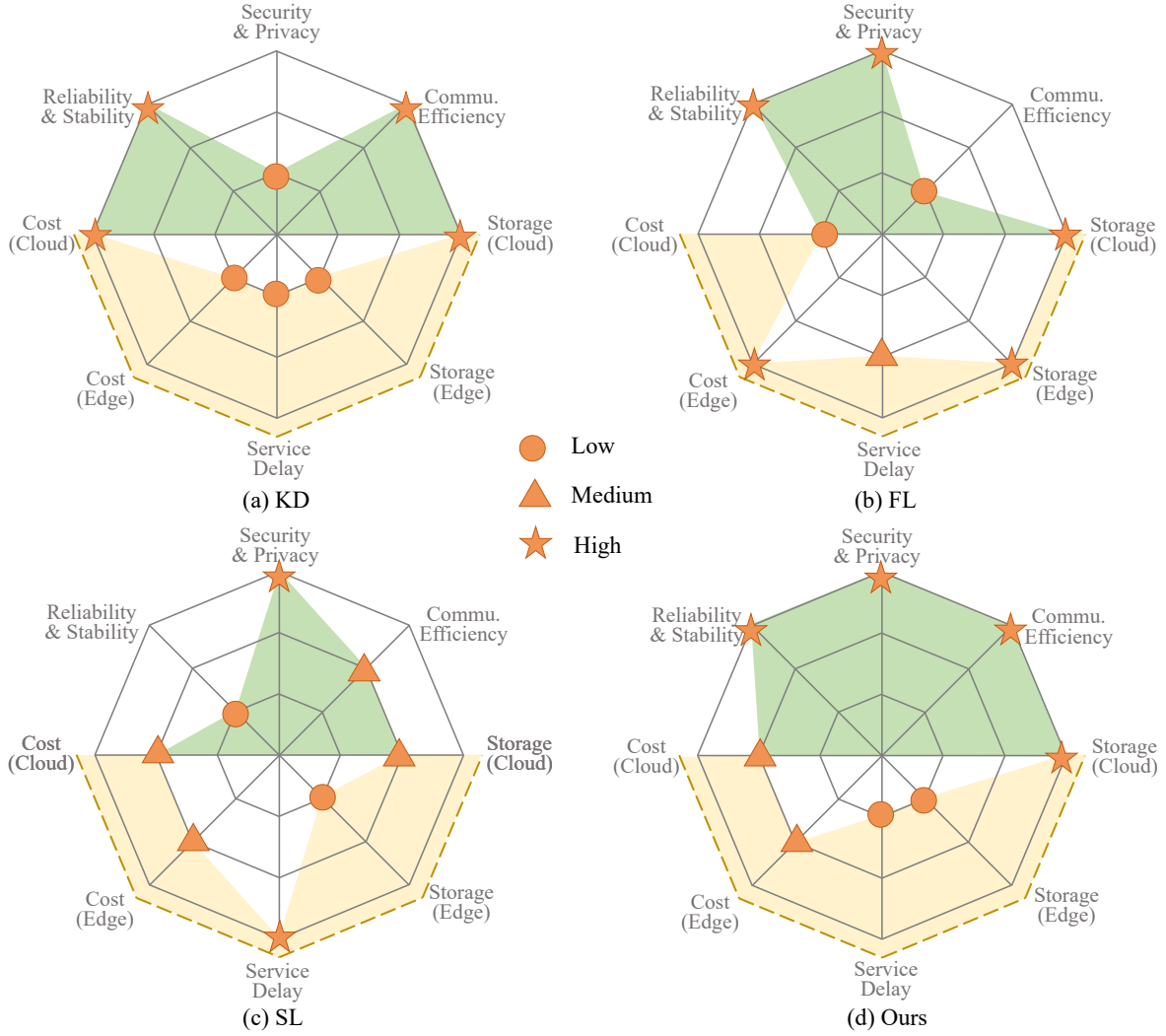


Fig. 2. 边云网络中4种分布式框架的KPI雷达图。

1) 多任务门控网络: 分层门控网络(HierGate)包括 M 学习者选择门(LSGates)和任务特定门(TSGate), 允许它组织多个任务和/或处理多模态输入。云将来自边缘节点的 N 异构模型分类为特定任务组, 形成 M 学习者小组。在每个组内, 专家被平行排列, 通过LSGate组合。TSGate通过将输入路由到相应的任务来控制各种任务的执行。LSGate选择最适合输入的顶级 K 学习者, 为他们分配单独的权重。 K 的值决定了特定任务激活的学习者数量, 从而影响计算成本。选择一个任务后, 其他任务中的学习者不需要激活整个模型。HierGate实现了高效的稀疏性, 产生了一个 N 维向量分割成 M 段, 只有 K 维是非零的。这代表了 N 学习者对特定任务的输出比例, 有效地利用了同一任务中不同学习者的不同知识。

2) 模块化和线性投影: 与典型的单纯通过门控网络连接学习者的MoE模型不同, 我们建议将学习者组织成模块, 并在他们之间建立线性投影联系, 以促进知识共享。由于同一组的学习者可能从彼此的专业知识中受益, 并且来自不同任务的学习者之间可能存在信息关联, 因此我们在遵循特定规则的 N 学习者之间创建线性联系。我们的规则可以解释如下: 取一个学习者在 i 阶段产生的特征, 执行线性投影将它们转换为其他学习者在 j 阶段的输入维度(其

中 $0 \leq j - i \leq h$), 并将结果添加到阶段 j 的原始输入。超参数 $h \geq 0$ 控制初始连接密度, 基于深度接近的层以相似程度处理原始输入的假设。重要的是, 从浅层到深层建立连接, 避免形成循环的模型结构。此外, 学习者之间的依赖关系也存在显著差异。某些任务表现出清晰而强大的关系, 受益于共享特征, 而其他任务则表现出较弱的关系, 共享特征不太明显。为了解决这个问题, 在模型训练过程中, 我们采用剪枝来迭代地过滤和保留必要的连接。该方法减少了模型的参数大小, 同时促进了模型内稳定的特征共享关系, 促进了跨任务的自适应知识传播。

C. 在云端训练BAIM

对于上述模型架构, 可训练参数包括门神经网络、特征投影的线性连接和每个学习器。下面我们介绍BAIM的三种训练策略:

1) 微调策略: 对所有可训练参数进行更新。上传的边缘模型作为微调的初始化。这种全面的调整确保整个模型收敛到优化配置, 利用嵌入在局部训练边缘模型中的知识。

2) 冻结策略: 保持上传的边缘模型不变, 只更新模型与门网的连接。这在整个训练过程中保持了每个学习者的个性, 作为整个模型的静态贡献者。

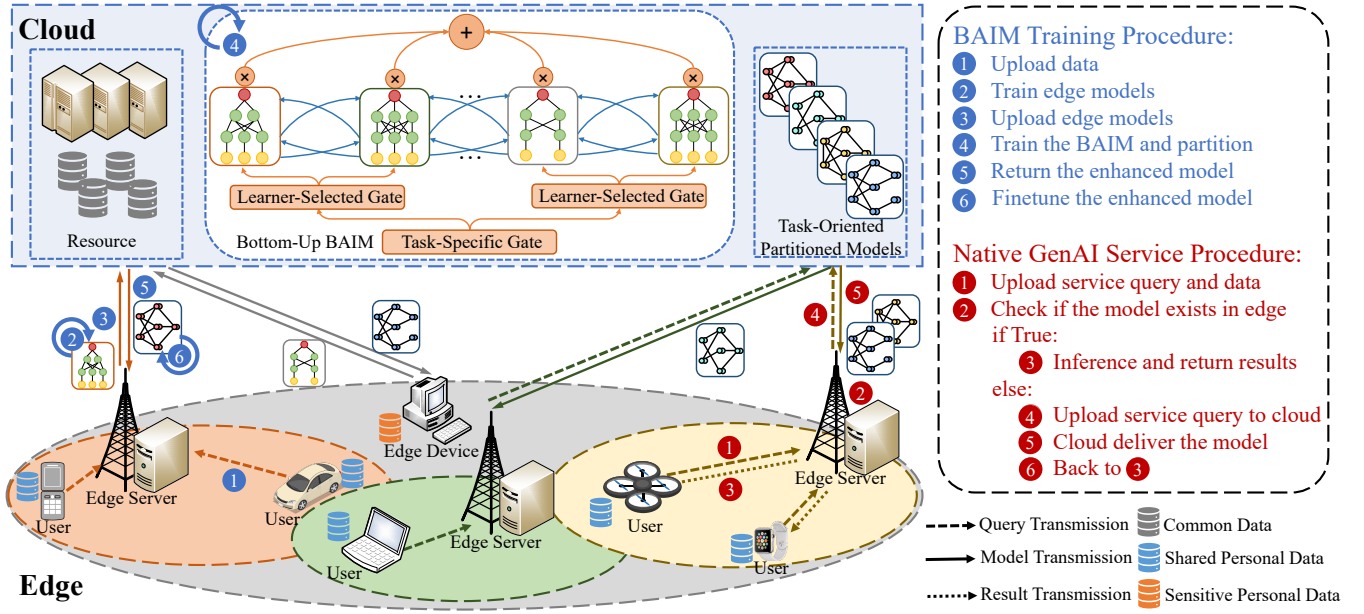


Fig. 3. 所提出框架的工作流程，包括BAIM训练和本地GenAI服务程序。

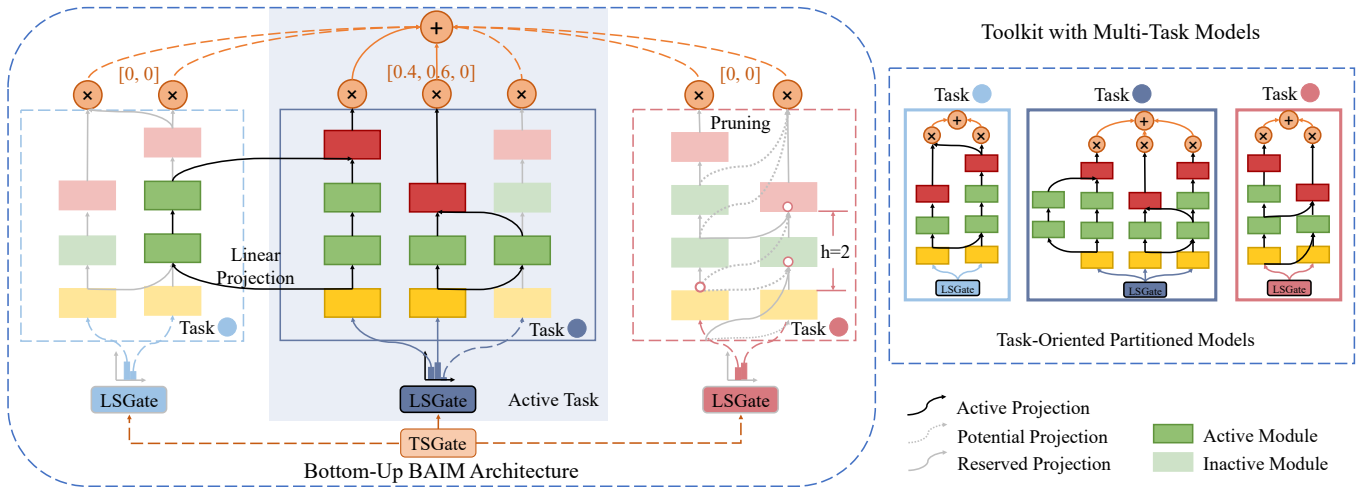


Fig. 4. 该工具包采用自底向上的BAIM体系结构和面向任务的划分模型，涉及3个任务。第二个任务目前由TSGate选择。执行暗模块，包括LSGate选择的top- k ($k=2$)学习者和与这些学习者有线性投影连接的模块，而亮模块在当前轮中不活跃。在第三个任务中，灰色虚线表示来自第一个学习者的初始潜在线性投影(连接高度 $h=2$)。在训练过程中，修剪滤波器并保留一部分，用灰色实线表示。

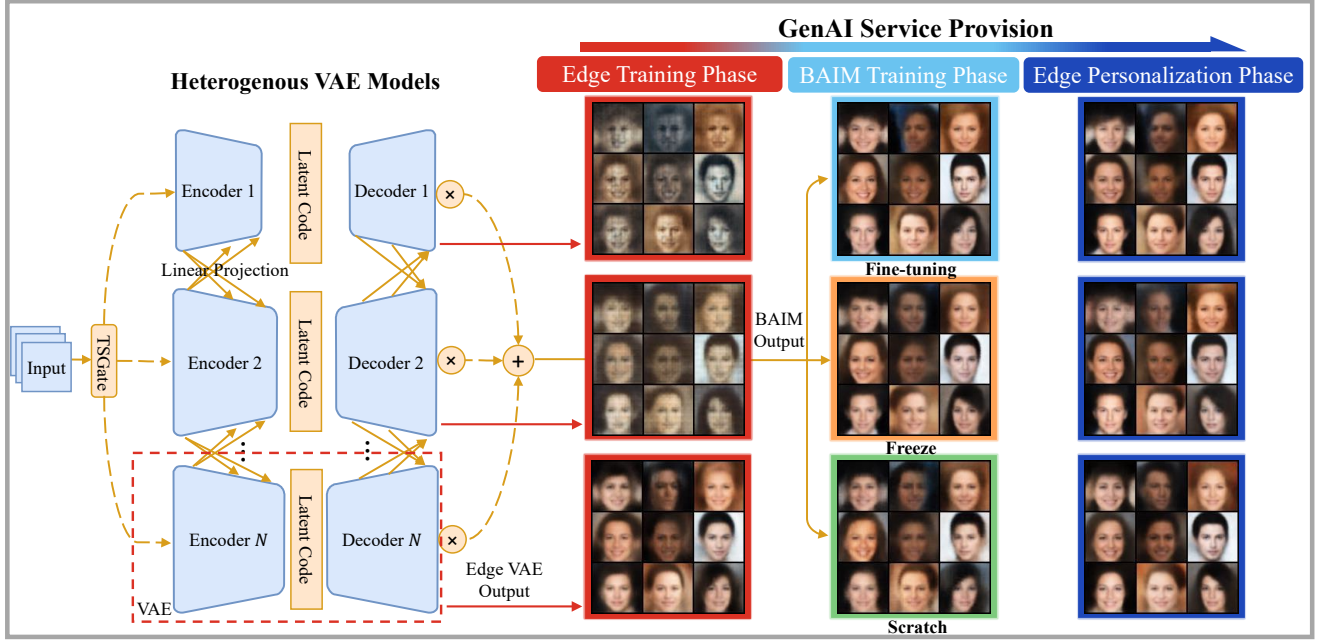
3) *Scratch*策略:从随机初始化开始训练所有参数。这种方法允许整个模型架构的彻底演化，强调统一的BAIM与上载的边缘模型中封装的预先存在的知识的独立性。

此外，在模型训练过程中，考虑来自动态环境，特别是来自边缘节点的知识在线更新解决方案至关重要。这涉及以下三种方法。

1) 持续学习: 持续学习是指模型在不忘记原始任务的情况下获得新的能力。与多任务学习[14]不同的是，所有任务都是同时学习的，持续学习涉及任务的逐步增加。在边云协同网络中，随着边缘节点服务的用户数量不断增长，其所需任务变得更加多样化，边缘节点不断上传模型以参与统一模型的聚合。自下而上的BAIM是一种可扩展的架构，对于有额外学习者的场景，可以微调门网络以逐渐学习新任务。

2) 模型级剪枝: 模型级修剪涉及从大型模型中裁剪子模型。在可扩展的BAIM中，随着边连接模型数量的增加，每个任务的学习者数量不断增加。同时，一些表现不佳的学习者很少或几乎从未被LSGate激活。对于这些学习者，在模型级剪枝时，可以将不辅助其他学习者的部分删除，将辅助其他学习者的模块直接合并到相应的学习者中。这保证了BAIM的存储效率和计算效率。

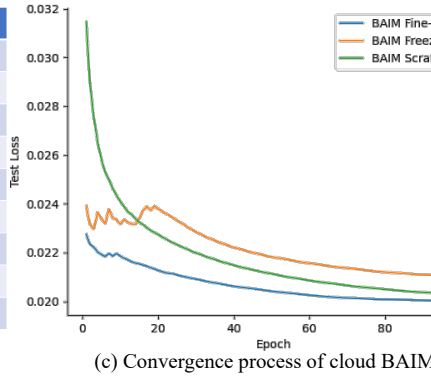
3) 少样本学习: 少样本学习是指在提供非常有限的示例时，训练模型以生成合格的上下文或做出准确的预测[15]。对于基于云的BAIM，直接可访问的数据通常是通用的公共数据，对于不同的任务，可能只有少量样本甚至零样本。BAIM需要在有限的样本下调整模型参数，以确保在各种任务上的良好性能。这需要彻底探索不同任务之间的相关性，并利用学习者之间共享的知识。



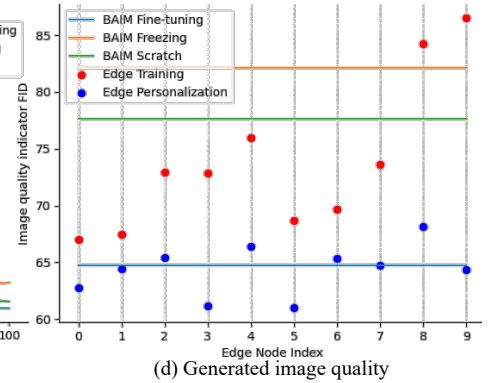
(a) Visualization of the process and results of small edge models constituting the cloud BAIM

Variables	Values
The Number of Edge Models N	10
Top K	2
Connection Height h	3
Learning Rate lr	0.005
KLD Weight	0.00025
Pruning Interval	10
Pruning Ratio	0.1
The Size of Training Dataset	14720×11
The Size of Testing Dataset	19962

(b) Simulation settings



(c) Convergence process of cloud BAIM



(d) Generated image quality

Fig. 5. 图像生成服务提供案例研究。

这三种知识更新策略对于长期部署框架的自我维护至关重要，部署框架需要不断演化和适应不断变化的条件。自底向上的BAIM架构本质上为这些挑战提供了相应的解决方案。

D. 面向任务部署的原生GenAI

除了在中心服务器上获得统一的多任务BAIM外，该框架还可以实现模型压缩和模型分割，以获得针对不同任务的性能提升。这些紧凑和轻量级的模型可以部署到边缘节点，为用户提供原生GenAI服务。如第二节所述，模型压缩通常会导致模型性能损失或需要额外的微调。然而，本文提出的架构具有独特的性质，允许在不牺牲性能的情况下，将模型分解为基于TS Gate的紧凑模型来执行相应的任务。涉及共享经验的连接被复制到新的模型中，如图4所示。每个任务的结果结构是一个具有线性连接的MoE模型架构，每个都有自己的LS Gate，用于根据输入选择学习者，对任务显示出优秀的泛化能力。因此，我们得到了 M 紧凑的模型。对于边缘节点，这些轻量级模型可以根据需要从云端下载，满足用户的服务需求。这种方法使边缘服务能够实现与云相当的性能，同时利用边缘服务的优势。

IV. 案例研究:图像生成服务提供

本节演示了一个典型的图像生成服务，该服务使用变分自编码器(VAE)模型，并展示了在框架中通过边云协作提高图像质量。如图5 (b)所示的仿真设置包括10个边缘节点，它们使用不同的本地数据集训练它们的异构VAE模型，然后在来自CelebA数据集的公共测试数据集上对它们进行评估。在本地训练后，边缘模型被上传到云端，在那里BAIM使用三种不同的策略进行训练:微调、冻结和划痕。最后，将微调策略得到的模型送回边缘进行进一步个性化微调。三个阶段的测试样本分别如图5 (a)中的三列图像所示。

图5 (c)显示了BAIM在三种训练策略下测试损失的收敛性。微调策略表现出最好的收敛性和性能，而冻结策略在收敛前的前几个周期内最初出现振荡，损失函数性能最差。由于初始化是随机的，scratch策略从较高的初始损失开始，经历快速下降，最终达到中等损失水平。这是由于scratch方法中边缘模型提取的知识不足，使得它在具有相同的模型大小和参数空间的情况下无法实现微调的性能。

Fréchet Inception Distance (FID)是一种广泛使用的评价

生成图像质量的指标。它衡量真实图像的分布和生成图像之间的相似性。较低的FID表明生成的图像与真实图像非常匹配。图5 (d)展示了在三种BAIM训练策略下模型生成的图像(用三条水平线表示),以及在初始边缘训练时由边缘生成的图像和在BAIM部署到边缘后由边缘个性化生成的图像(用两种散点表示)。值得注意的是,微调策略被证明是有效的,与原始边缘模型相比,显著提高了图像质量。

V. 挑战和潜在的研究机会

我们提出的框架为在5G-advanced和6G通信网络中高效提供服务铺平了道路。然而,它引入了需要关注的挑战。分析了在数据管理、模型融合方案设计和节点管理等方面面临的挑战,并提出了可能的研究方向。

A. 数据管理

通过区分敏感个人数据、共享个人数据和公共数据来解决数据隐私问题。展望未来,数据管理和生成的全面解决方案还有待开发。

1) 一种安全数据管理方案:建立有力的保障措施,保护数据在存储和传输过程中的安全。这包括端到端数据加密和增强的身份验证和授权机制。此外,还需要对云端公共数据进行匿名和脱敏技术,以最大限度地降低信息泄露的风险。

2) 用合成数据替换用户原始数据:这涉及到差分隐私技术、生成对抗网络(GANs)和数据扰动方法的应用,以生成具有真实数据特征的合成数据。随着人工智能生成内容(AIGC)的进步,合成数据可以作为模型训练数据的更安全、更可靠的选择。

B. 模型融合方案

在模型训练阶段,设计更有前途的模型融合策略和异步更新机制,可以持续提高计算性能和通信效率。

1) 优化异构架构融合策略:边缘模型在深度和宽度方面的结构各不相同,在卷积神经网络(cnn)、循环神经网络(rnn)和transformer等架构中也各不相同。为了利用这些不同的模型进行高效的多任务学习,改进异构架构融合策略至关重要,包括投影方法、连接规则和剪枝策略。这使得更好地探索任务相关性和信息共享以相互加强。

2) 设计异步更新机制:异步更新系统允许每个边缘节点在计算完成后立即上传数据,有效地减少了等待时间。由于边缘模型是独立训练的,云会持续接收这些训练良好的小模型并更新BAIM。需要一种精心设计的异步更新机制来协调这些边缘模型的融合。这种机制应该在BAIM的陈旧性和计算成本之间取得最佳平衡。

C. 节点管理

节点管理包括对分布式系统内变化的灵活监控、调整和协调。有效的节点管理可以增强系统的稳定性和可靠性,减少因节点异常导致的性能下降。

1) 自适应动态边缘网络:实际系统中的边缘网络不断变化,导致边缘节点可能不稳定,包括其接入和断开连接。系统必须适应新节点的加入和失效节点的处理,并对当前节点状态的变化作出响应。由于边缘节点通常是分布式,并且可以移动,因此系统需要强大的适应性来解决网络延迟、数据丢失或节点可用性变化等问题。恰当的节点管理在模型演化中起着至关重要的作用,需要精心规划的增强机制。

2) 应对安全威胁:在开放云边系统中,恶意节点攻击是不可避免的问题。这些可能包括通过虚假的训练结果破坏模型性能的不诚实节点,或通过拒绝服务(DoS)攻击扰乱系统运行。为了应对这些问题,安全措施包括篡改和异常检测、信任评估、节点更新与历史行为或节点更新的交叉验证。实现基于信任的信誉系统来指导节点交互也是基础。

VI. 结论

总之,边缘原生GenAI和基于云的BAIMs之间的协同作用成为6G通信网络中的关键组成部分,有望提高QoE和QoS。所提出的框架从战略上解决了人工智能训练和部署中的普遍限制,特别关注缓解与BAIMs分布式训练和原生GenAI部署复杂性相关的挑战。该框架在图像生成用例中所展示的有效性强调了这种协作范式的巨大潜力。此外,全面的研究方向被描绘出来,以解锁原生GenAI和BAIMs之间协同作用的所有可能性。

REFERENCES

- [1] M. Xu, H. Du, D. Niyato, *et al.*, "Unleashing the power of edge-cloud generative AI in mobile networks: A survey of AIGC services," *arXiv preprint arXiv:2303.16129*, 2023.
- [2] X. Chen, Z. Guo, X. Wang, *et al.*, "Foundation model based native AI framework in 6G with cloud-edge-end collaboration," *arXiv preprint arXiv:2310.17471*, 2023.
- [3] Z. Chen, Z. Zhang, and Z. Yang, "Big AI models for 6G wireless networks: Opportunities, challenges, and research directions," *arXiv preprint arXiv:2308.06250*, 2023.
- [4] Y. Xiao, G. Shi, Y. Li, *et al.*, "Toward self-learning edge intelligence in 6G," *IEEE Communications Magazine*, vol. 58, no. 12, pp. 34–40, 2020. DOI: 10.1109/MCOM.001.2000388.
- [5] X. Lin, "An overview of the 3GPP study on artificial intelligence for 5G new radio," *arXiv preprint arXiv:2308.05315*, 2023.
- [6] Y. Gu, L. Dong, F. Wei, *et al.*, "Knowledge distillation of large language models," *arXiv preprint arXiv:2306.08543*, 2023.
- [7] M. Chen, D. Gündüz, K. Huang, *et al.*, "Distributed learning in wireless networks: Recent progress and future challenges," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 12, pp. 3579–3605, 2021. DOI: 10.1109/JSAC.2021.3118346.
- [8] W. Xu, Z. Yang, D. W. K. Ng, *et al.*, "Edge learning for B5G networks with distributed signal processing: Semantic communication, edge computing, and wireless sensing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 17, no. 1, pp. 9–39, 2023. DOI: 10.1109/JSTSP.2023.3239189.
- [9] Y. Tian, Z. Zhang, Z. Yang, *et al.*, "JMSNAS: Joint model split and neural architecture search for learning over mobile edge networks," in *2022 IEEE International Conference on Communications Workshops*, IEEE, 2022, pp. 103–108.
- [10] Y. Yang, Z. Zhang, Y. Tian, *et al.*, "Over-the-Air split machine learning in wireless MIMO networks," *IEEE Journal on Selected Areas in Communications*, vol. 41, no. 4, pp. 1007–1022, 2023. DOI: 10.1109/JSAC.2023.3242701.
- [11] J. Li, L. Lyu, D. Iso, *et al.*, "MocoSFL: Enabling cross-client collaborative self-supervised learning," in *The Eleventh International Conference on Learning Representations*, 2022.
- [12] Google, "Introducing Pathways: A next-generation AI architecture," <https://blog.google/technology/ai/introducing-pathways-next-generation-ai-architecture/>.
- [13] B. Mustafa, C. Riquelme, J. Puigcerver, *et al.*, "Multimodal contrastive learning with LiMoE: The language-image mixture of experts," *Advances in Neural Information Processing Systems*, vol. 35, pp. 9564–9576, 2022.

- [14] Z. Chen, Y. Shen, M. Ding, *et al.*, “Mod-squad: Designing mixtures of experts as modular multi-task learners,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 11 828–11 837.
- [15] Z. Lin, S. Yu, Z. Kuang, *et al.*, “Multimodality helps unimodality: Cross-modal few-shot learning with multimodal models,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 19 325–19 337.