

云边端协同架构下人工智能信息处理机制

何俊杰

成都工业学院, 四川 成都 611730

[摘要] 集中式云计算难以同时满足人工智能业务对实时响应、数据安全、复杂计算和规模扩展的要求。云边端协同架构通过端侧感知与轻量推理、边缘侧聚合分析、云端训练和全局优化, 形成分层信息处理链路。文章分析数据筛选、任务卸载、模型切分、协同推理和分布式训练机制, 结合公开研究数据讨论其性能基础及现实问题, 并提出任务分级、资源编排、模型治理和安全控制路径, 为人工智能系统的协同部署提供参考。

[关键词] 云边端协同; 人工智能; 信息处理; 边缘智能; 协同推理

DOI: 10.64635/ja.2026.1441

中图分类号: TP393.027

文献标识码: A

Artificial Intelligence Information Processing Mechanism Under the Cloud-Edge-End Collaborative Architecture

He Junjie

Chengdu Technological University, Chengdu, Sichuan 611730, China

Abstract: Centralized cloud computing has difficulty simultaneously meeting the requirements of artificial intelligence services for real-time response, data security, complex computing, and scale expansion. The cloud-edge-end collaborative architecture forms a hierarchical information processing chain through end-side perception and lightweight inference, edge-side aggregation and analysis, cloud-side training, and global optimization. This paper analyzes mechanisms such as data filtering, task offloading, model partitioning, collaborative inference, and distributed training, discusses their performance basis and practical problems in combination with publicly available research data, and proposes paths for task classification, resource orchestration, model governance, and security control, providing reference for the collaborative deployment of artificial intelligence systems.

Keywords: cloud-edge-end collaboration; artificial intelligence; information processing; edge intelligence; collaborative inference

引言

人工智能应用所处理的信息已由结构化业务数据扩展到图像、视频、语音、传感序列和多模态内容。数据规模扩大与业务实时性提高后, 集中上传、云端处理、结果返回的单一路径面临通信时延、带宽占用、隐私泄露和服务中断等问题。终端具有接近数据源和响应速度快的特点, 但算力、存储和能耗受限。云端拥有较强的计算和模型管理能力, 但物理距离与网络波动会影响实时任务。边缘节点位于二者之间, 可承担区域汇聚、任务调度和中等强度计算。

云边端协同超越三个计算层级的简单叠加, 其核心是根据数据敏感程度、任务计算量、响应期限和网络状态配置处理位置, 并在多个节点之间传递原始数据、特征、模型参数或推理结果。董裕民等将计算优化和计算卸载归纳为云边端架构下边缘智能的关键问题^[1]。张圣宇等进一步提出端侧实时响应与云侧资源能力的互补关系^[2]。人工智能信息处理机制因而需要回答三个问题: 不同层级处理什么信息, 任务如何在

层级间迁移, 模型如何利用分散数据持续更新。

1 云边端协同架构及功能边界

1.1 架构组成与资源特征

端侧包括传感器、移动终端、智能摄像机、机器人控制器及嵌入式设备, 主要负责数据获取、格式转换、轻量特征提取和时间敏感推理。端侧接近物理对象, 可在网络异常时维持基本功能, 也能减少敏感原始数据的远程传输。其局限表现为处理器性能、内存、供电和散热能力有限, 难以长期运行大规模模型。

边缘侧由工业网关、边缘服务器、区域数据中心或接入网络节点构成, 负责汇聚多个终端数据, 执行清洗、融合、缓存、模型推理和局部决策。边缘节点具有较低通信时延, 也能利用一定规模的算力开展多任务处理。云侧承担大规模数据存储、复杂模型训练、全局知识融合、跨区域资源调度和模型版本管理, 可形成面向多个边缘节点的公共能力。

三个层级的边界具有动态性。网络状况良好且任务计算量较大时, 部分信息可上传云端。通信受限或

任务具有严格响应期限时,推理靠近端侧部署。涉及多个终端关联的信息可在边缘侧汇聚。功能边界由业务风险和资源状态共同确定,而非按照固定设备类型划分。

1.2 人工智能任务的分层配置

人工智能任务可分为感知、表征、推理、学习和决策五类。感知任务把图像、声音或传感信号转化为数字数据,通常在端侧完成。表征任务包含滤波、编码、压缩和特征提取,可在端侧或边缘侧执行。推理任务依据模型输出分类、检测、预测或生成结果,部署位置取决于模型规模与响应要求。学习任务需要较多样本和计算资源,通常由云端承担,边缘侧可开展增量学习或局部更新。决策任务涉及风险和责任,需结合业务规则与人工控制配置权限。

任务分层还需考虑信息形态。上传原始数据能够保留完整信息,但通信和隐私成本较高。上传中间特征可减少部分数据量,却可能出现特征尺寸扩大或隐私重构风险。上传模型参数适合联邦学习和分布式训练,但参数传输仍会占用网络资源。上传推理结果的数据量较小,云端能够获得的信息细节也相应减少。系统需要在信息完整性、时延、成本和安全之间确定合理方案。

2 云边端协同的信息处理机制

2.1 数据采集与分级处理

信息处理起始于端侧数据采集。终端先完成时间标记、数据校验、噪声抑制和格式转换,再依据业务规则进行数据分级。高频且重复的数据可在本地聚合,异常片段和关键帧可提高保留优先级,涉及个人隐私或商业秘密的数据可脱敏后再传输。端侧还需记录设备状态、采集环境和模型版本,避免数据脱离产生条件。

边缘节点接收多个终端信息后,完成空间关联、时间对齐和多源融合。例如,工业视觉系统可把摄像机图像与设备速度、工位编号及质量检测结果关联,交通系统可把车辆、道路和信号控制信息按区域汇聚。边缘侧输出可分为实时事件、周期特征和原始样本索引。实时事件直接服务本地响应,周期特征用于趋势分析,具有训练价值的样本传向云端。

云端负责长期存储、跨区域比较和全局知识提取。数据分级处理减少了无效传输,也改变了云端数据结构。若边缘侧筛选规则过严,云端可能缺少低频异常和完整背景。系统需保留抽样上传、异常回溯和规则

复评机制,维持数据压缩与信息完整之间的平衡。

2.2 任务卸载与模型协同推理

任务卸载是把全部或部分计算任务迁移到边缘节点或云端。决策依据包含端侧计算时间、传输时间、节点排队时间、任务截止时间、能耗和模型精度。计算量较小、实时性高的任务优先在端侧执行。数据关联范围较大、需要区域资源的任务可卸载到边缘侧。复杂推理、长文本处理和大规模搜索适合云端执行。

模型协同推理主要包含整模分配、模型切分、早退推理和大小模型协同。整模分配根据节点状态选择完整模型的运行位置。模型切分把神经网络不同层部署在端、边或云端,并传递中间特征。早退推理在中间输出置信度达到要求时提前结束计算,低置信度样本再交由更强模型处理。大小模型协同由端侧小模型完成常规判断,复杂样本升级到边缘或云端大模型。

协同推理的重点不只在减少时延,还包括准确率、带宽、能耗和稳定性。切分位置过靠前会传输较大特征,过靠后会增加端侧计算压力。早退阈值偏低可能降低结果可靠性,阈值偏高则削弱节省计算的效果。调度器需根据实时网络、节点负载和任务风险动态选择路径。

2.3 分布式训练与模型更新

云边端协同训练可采用数据并行、模型并行和联邦学习。数据并行由多个节点分别采用本地数据训练相同模型,再汇总参数或梯度。模型并行把模型模块分配到不同节点,适合单个终端难以容纳完整模型的场景。联邦学习保留本地原始数据,只交换模型更新信息,可降低数据集中汇聚造成的隐私风险。

边缘节点在协同训练中承担终端选择、局部聚合和异常更新过滤,云端负责全局参数聚合、模型验证与版本发布。王睿等指出,边缘协同训练主要受到设备异构、资源受限和网络不稳定影响,参数服务器架构与去中心化架构具有不同的通信和控制特点^[3]。部分终端计算速度慢或数据量较少时,统一轮次训练容易出现等待。系统可按照设备能力分组,采用异步聚合、梯度压缩和节点抽样降低训练开销。

模型更新需要形成“训练—验证—发布—监测—回退”链路。云端新模型先通过离线测试,再向部分边缘节点灰度发布。边缘侧比较新旧模型的准确率、时延和资源消耗,达到要求后扩大范围。运行期间发现性能下降或异常输出时,节点可回退到稳定版本,并把问题样本提交云端分析。

3 人工智能信息处理的协同模式

3.1 本地快速响应与边缘区域分析

本地快速响应模式适合设备保护、机器人避障、身份核验和现场异常检测。端侧模型直接处理当前数据，输出结果后联动本地业务。网络中断不会阻断基本推理，原始数据也可保留在现场。端侧模型需要压缩参数、降低运算量，并设置明确的适用范围。超出识别范围的输入转交边缘节点，避免小模型对复杂样本作出高置信度错误判断。

边缘区域分析模式适合多摄像机协同、生产线质量检测、园区安防和区域交通控制。边缘节点汇聚多个终端特征，完成目标关联、事件去重和局部决策。涂静正等梳理的工业视频感知研究表明，在特定 100 Mbps 带宽条件下，EdgeLeague 方法的平均处理时延为 13 ms，识别准确率达到 92%^[4]。该结果说明协同方法能够在特定实验条件下兼顾时延和准确率，实际部署仍需结合模型、硬件、数据集和网络环境重新测试。

表 1 云边端人工智能信息处理的公开技术与实验数据

项目	公开条件或功能	公开结果	数据来源
EdgeLeague	网络带宽 100 Mbps	平均时延 13 ms，识别准确率 92%	涂静正等 ^[4]
工业视频感知方法	支持		
TensorFlow Lite	Android、iOS、Linux	主要支持端侧推理	董裕民等 ^[1]
PyTorch Mobile	支持 iOS、Android	支持训练和推理	董裕民等 ^[1]
MindSpore Lite	Android、iOS、OpenHarmony、Linux、Windows	支持训练和推理	董裕民等 ^[1]
Paddle Lite	Android、iOS、Linux、Windows、macOS	主要支持端侧推理	董裕民等 ^[1]

表中框架信息反映论文发表时的公开支持范围，软件版本更新可能改变具体能力，工程选型需以目标版本的技术文档和实机测试为准。

3.2 云端全局学习与动态协同

云端全局学习模式适合训练基础模型、融合跨区

域样本和生成公共知识。云端可聚合多个边缘节点上传的脱敏样本、特征统计或模型参数，识别单一区域难以发现的长周期规律。训练结果经过验证后下发边缘和终端，形成统一基础模型，再根据本地数据开展轻量适配。

动态协同模式根据任务状态实时调整处理路径。调度器持续观察端侧算力、边缘队列、网络带宽、云端负载和任务优先级。常规样本在端侧完成，疑难样本转向边缘，跨区域关联或高复杂度任务交由云端。网络拥塞时减少原始数据上传，边缘故障时切换相邻节点或保留本地服务。该模式具有较强适应性，也增加了调度策略、状态同步和结果一致性管理的难度。

4 云边端协同信息处理的现实问题

4.1 资源异构与任务波动

终端处理器架构、内存、功耗和运行环境差异较大，边缘节点的算力配置与并发负载也不一致。同一模型在不同设备上的时延和能耗可能明显不同。业务请求还具有时间波动，固定卸载方案难以持续匹配资源状态。任务频繁迁移会增加传输和容器启动成本，调度决策本身也会消耗计算资源。

系统需要建立设备能力画像，记录处理器、内存、加速器、网络和能耗等信息。任务画像描述数据规模、模型结构、响应期限和可靠性等级。调度器根据两类画像匹配运行位置，并设置迁移收益门槛，避免资源轻微变化引发频繁切换。

4.2 数据差异与模型可靠性

云端训练数据与端侧真实数据可能存在分布差异。不同地区、设备、光照、语言和用户行为会改变输入特征，统一模型在局部场景中可能性能下降。端侧样本数量有限时，个性化训练容易过拟合。边缘聚合若忽略样本质量，还可能放大错误标签和异常节点影响。

模型评价需覆盖全局性能与局部性能。除平均准确率外，还要检查不同终端、不同环境和少数类别的误差。大小模型协同需校验升级条件，模型切分需检查中间特征完整性，联邦训练需识别异常参数。模型输出涉及安全、医疗、金融和生产控制时，人工确认和业务规则仍是关键控制环节。

4.3 通信安全与责任控制

云边端之间持续交换数据、特征和模型参数，任一节点受到攻击都可能影响协同链路。伪造终端可上传污染数据，恶意边缘节点可篡改推理结果，模型参数也可能泄露训练信息。系统需采用设备身份认证、

传输加密、完整性校验、最小权限和安全审计，关键模型与配置文件设置签名验证。

信息处理权限需与业务责任对应。端侧可自动完成低风险、可逆操作，高风险决策需要边缘或云端复核，并保留人工批准。日志记录任务来源、模型版本、处理节点、输入摘要、输出结果和处置人员，便于质量追踪和责任确认。跨组织协同时，还需明确数据范围、模型归属和服务中断责任。

5 云边端协同信息处理机制的优化路径

5.1 面向业务目标的分层建设

第一步梳理业务任务，按照响应期限、数据敏感度、计算强度和风险等级分类。第二步测量终端、边缘和云端的真实算力、网络与能耗，建立资源基线。第三步选择典型任务开展对比测试，分别测量端侧独立处理、边缘处理、云端处理和协同处理的时延、准确率、带宽与成本。第四步确定数据分级、模型部署、卸载策略和失败回退方案。

分层建设需保留开放接口。端侧采集格式、边缘消息结构和云端模型服务采用统一标识，降低设备和算法替换成本。模型压缩、特征编码和通信协议需联合设计，避免单独压缩模型后造成中间特征传输量增加。试点阶段先验证单一场景，运行稳定后再扩展节点和任务类型。

5.2 全过程运行治理与评价

运行治理覆盖数据、模型、资源和安全四类对象。数据治理关注来源、质量、权限和保存期限。模型治理关注训练数据、版本、指标、发布范围和回退条件。资源治理关注节点状态、任务队列和成本。安全治理关注身份、通信、权限、漏洞和审计。四类信息在统一管理平台关联，便于发现性能变化的真实原因。

评价指标可分为处理质量、服务效率、资源消耗

和风险控制。处理质量包括准确率、召回率和结果一致性。服务效率包括端到端时延、任务完成率和可用性。资源消耗包括通信量、计算占用、能耗和存储。风险控制包括异常节点发现率、模型回退时间和审计记录完整率。不同业务设置不同权重，避免以单一时延或准确率评价整个协同体系。

6 结语

云边端协同架构通过端侧感知与快速推理、边缘侧区域融合与任务调度、云端全局训练与知识管理，形成贯穿数据、特征、模型和决策的信息处理链路。其技术重点包括数据分级、任务卸载、模型切分、大小模型协同和分布式训练。公开研究已显示特定协同方法兼顾时延与准确率的能力，实际性能仍受硬件、网络、数据分布和任务风险影响。系统建设需以业务目标为起点，通过资源画像、动态调度、模型治理、安全控制和多维评价保持协同机制稳定运行，进而提高人工智能服务的实时性、可靠性与资源利用效率。

【参考文献】

- [1] 董裕民, 张静, 谢昌佐, 等. 云边端架构下边缘智能计算关键问题综述: 计算优化与计算卸载[J]. 电子与信息学报, 2024, 46(3): 765-776.
- [2] 张圣宇, 况琨, 吕承飞, 等. 端云协同智能计算的关键问题、方法和应用[J]. 中国工程科学, 2024, 26(1): 127-138.
- [3] 王睿, 王岩, 尹朴, 等. 面向边缘智能的协同训练研究进展[J]. 工程科学学报, 2023, 45(8): 1400-1416.
- [4] 涂静正, 温晓婧, 陈彩莲, 等. 基于边缘计算的工业视频网络智能感知: 挑战与进展[J]. 自动化学报, 2025, 51(8): 1715-1738.