

车联网边缘智能：概念、架构、问题、实施和展望

江恺¹，曹越^{1,2}，周欢³，任学锋²，朱永东⁴，林海¹

(1. 武汉大学，湖北 武汉 430072；2. 华砺智行（武汉）科技有限公司，湖北 武汉 430056；
3. 三峡大学，湖北 宜昌 443002；4. 之江实验室，杭州 浙江 310100)

摘要：作为一项新兴交叉学科领域，边缘智能通过将人工智能推送至靠近交通数据源侧，并利用边缘算力、存储资源及感知能力，在提供实时响应、智能化决策、网络自治的同时，赋能更加智能、高效的资源调配与处理机制，从而实现车联网从接入“管道化”向信息“智能化”使能平台的跨越。然而，当前边缘智能于车联网领域的成功实施仍处于起步阶段，迫切需要以更为广阔的视角对这一新兴领域进行全面综述。为此，面向车联网应用场景，首先介绍边缘智能的背景、概念及关键技术；然后，对车联网应用场景中基于边缘智能的服务类型进行整体概述，同时详细阐述边缘智能模型的部署和实施过程；最后，分析边缘智能于车联网中的关键开放性挑战，并探讨应对策略，以推动其潜在研究方向。

关键词：人工智能；车联网；边缘智能

中图分类号：TP391

文献标志码：A

doi: 10.11959/j.issn.2096-3750.2023.00320

Edge intelligence empowered internet of vehicles: concept, framework, issues, implementation, and prospect

JIANG Kai¹, CAO Yue^{1,2}, ZHOU Huan³, REN Xuefeng², ZHU Yongdong⁴, LIN Hai¹

1. Wuhan University, Wuhan 430072, China 2. Huali SmartWays Technology Co., Ltd., Wuhan 430056, China
3. China Three Gorges University, Yichang 443002, China 4. Zhejiang Lab, Zhejiang 310100, China

Abstract: As an emerging inter discipline field, edge intelligence pushes AI to the side close to the traffic data source. Edge intelligence makes use of the computing power, storage resources, and perception ability of edge to provide a more intelligent and efficient resource allocation and processing mechanism while providing a real-time response, intelligent decision-making and network autonomy, realizing the critical leap for internet of vehicles from access “pipelining” to the intelligent enabling platform of information. However, the successful implementation of edge intelligence in internet of vehicles is still in its infancy, and there exists a demand for a comprehensive survey in this young field from a broader perspective. Based on this context of internet of vehicles, the background, concepts and key technologies of edge intelligence were introduced. Then, a holistic overview of service types based on internet of vehicles was taken, and the entire processes of model training and inference in edge intelligence were elaborated. Finally, to promote the potential research directions, the key open challenges of edge intelligence in the internet of vehicles were analyzed, and the coping strategies were discussed.

Key words: artificial intelligence, internet of vehicles, edge intelligence

收稿日期：2022-07-15；修回日期：2022-12-30

通信作者：曹越，yue.cao@whu.edu.cn

基金项目：湖北省国际科技合作计划项目(No.2022EHB002)；教育部中国高校产学研创新基金支持项目(No.2021LDA07005)；湖北省重大调研课题基金项目(No.2022KT03-2)

Foundation Items: Hubei Province International Science and Technology Collaboration Program (No.2022EHB002), Ministry of Education China University Industry-University-Research Innovation Program (No.2021LDA07005), Hubei Province Major Consultancy Program (No.2022KT03-2)

0 引言

在技术发展与业务需求的双重作用下, 交通系统正逐渐演变为以数据驱动为主的智能系统。当前, 汽车行业正迎来新革新与生命力, 新兴的车联网技术已赋予车辆提供可靠车载多媒体服务^[1]、自适应巡航控制^[2]、智能交通信息管理^[3]等信息化能力, 促进了智能交通的发展道路, 有效地提升了乘客的行车安全性和旅途舒适性。

随着信息化的蓬勃发展, 车载应用对处理能力和服务质量 (QoS, quality of service) 提出了更高的要求, 不可避免地较传统移动应用占用更多资源量和能耗^[4]。然而, 受车辆自身处理能力瓶颈以及云计算平台长距离回传的限制, 云计算架构因传输距离而有高时延和低可靠性, 必然难以满足实时类车载应用的 QoS 保障。同时, 相较于车联网边缘数据的增长速度, 云计算能力的线性增速也无法满足需求^[5-6]。

为此, 云计算的相关核心功能应部署于接近数据源的位置, 即将网络边缘作为新兴的技术架构角度考虑。因而, 作为车路互联应用的核心计算支撑, 边缘计算技术^[7]应运而生。具体来说, 边缘计算可以理解为云计算模型的补充和扩展, 并不完全依赖云端能力, 而是促进云与边缘能力的协同统一^[8]。该项技术通过在道路路侧单元 (RSU, road side unit) 上部署边缘服务器, 将计算、通信、存储、控制及管理网络功能, 由集中式云端下延至网络边缘侧^[9-11], 利用网络边缘侧与车辆的物理接近性, 缓解传输距离导致的时延和不可靠性。从技术角度看, 边缘计算可实现计算任务向 RSU 的迁移, 为实时类车载应用提供算力资源支撑^[9]。同时, 在车联网中部署边缘计算功能, 还具有其分布式结构和小规模性质^[12]带来的额外技术优势, 包括敏捷联接^[7]、隐私保护^[13]、可拓展性和上下文感知^[6]等。

此外, 鉴于边缘服务器的处理能力相对有限, 时而供需不平衡, 按需服务优化通常也是实现车辆服务增益的关键因素。优化核心将针对边缘环境中的算力、存储、网络资源高效分配, 在用户需求侧和资源供给侧, 处理有限网络资源的编排与调度问题, 实现计算任务动态部署。值得强调的是, 进行服务优化时需考虑可扩展性、灵活性和高效率等方面, 以适应发展趋势和多样化应用需求。尽管传统服务优化方法的求解是 NP 难问题, 但车联网的高度动态性和需求不确定性对资源调度的自适应方法设计提出了特定的要求, 使得基于模型的服务优

化方法往往不再适用^[14-15]。为此, 迫切需要设计不依赖通用化模型的边缘服务优化方法。

与此同时, 人工智能 (AI, artificial intelligence) 的发展在历经两次低谷和 3 次崛起后, 于过去 10 年进入了飞跃阶段。得益于硬件升级和神经网络泛化, AI 凭借其在数据分析和提取洞察力方面的优势, 支持于动态环境下扩展技术创新以增强网络的认知和智能水平。将 AI 推送至靠近交通数据源的车联网边缘侧, 催生出一类新兴的学习范式, 即边缘智能^[14,16-17]。

边缘智能被广泛地认为是智能化边缘计算的落地部署, 其关键在于以边缘为依托, 实现边缘计算与 AI 的优势互补。这里, 边缘智能将云端处理能力下沉至接入网边缘, 并在靠近车辆的网络边缘引入 AI 技术, 通过融合无线边缘网络的算力、通信、存储资源及感知能力, 在提供实时响应、智能化决策、网络自治的同时, 赋能更加智能、高效的资源调配与处理机制, 最终实现车联网从接入“管道化”向信息智能化使能平台的跨越^[15]。

然而, 尽管边缘智能近年来已引起了学术界的广泛关注, 但边缘智能于车联网领域的成功实施仍处于起步阶段, 迫切需要以广阔的视角对这一新兴领域进行全面综述。为此, 本文面向车联网应用场景, 首先介绍边缘智能的背景, 概念及关键技术; 然后, 对车联网应用场景中基于边缘智能的服务类型进行整体概述, 同时详细阐述边缘智能模型的部署和实施过程; 最后, 探讨边缘智能于车联网中的关键开放性挑战, 以推动其潜在研究方向。

1 基本概念和关键启用技术

1.1 边缘智能的基本概念

边缘智能通过结合边缘计算和 AI 技术的优势而生, 近年来被运营商认为是摆脱 5G 网络“管道化”的有力支撑。边缘智能的出现对车联网系统效率、服务响应、调度优化和隐私保护具有重要意义。本质上, 在网络边缘引入 AI 技术可使 RSU 进行本地化模型训练和推断, 从而避免与远程云服务器的频繁通信。边缘智能强调将计算决策靠近数据源头, 同时将智能服务由云端推送至边缘侧, 以减少服务的交付距离和时延, 提升车辆的接入服务体验^[18]。同时, AI 模型从车联网的实际边缘环境中提取特征, 通过与环境的反复迭代赋予高质量的边缘计算服务。近 10 年来, 深度学习^[19]和强化学习^[20-21]已逐渐成为边缘智能中的

主流 AI 技术。这里，深度学习可以从数据中自动提取特征和检测边缘异常，而强化学习可通过马尔可夫决策过程和合适的梯度策略实现目标，在网络的实时决策中发挥越来越重要的作用。

1.2 边缘智能的关键启用技术

1.2.1 网络切片

网络切片是解决“一刀切”网络模式的重要技术，通过将网络虚拟划分为多个段，允许在单一共享的物理基础设施上，创建多个逻辑独立的自包含网络实例（即网络切片）^[22]。作为虚拟化的端到端逻辑网络，任意切片都可被用于满足特定的业务类型，从而通过支持定制的网络功能、层次抽象及不同类别的隔离（资源隔离、业务隔离、运维隔离）^[23]，实现车辆对网络能力的差异化服务要求。网络切片的架构及功能示例如图 1 所示。

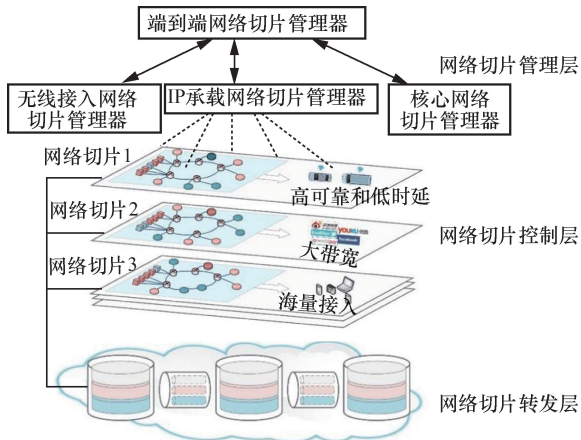


图 1 网络切片的架构及功能示例

通常，为定制车载服务的异构资源组合，需面向车辆到基础设施（V2I, vehicle to infrastructure）的访问控制构建网络资源分配框架。于适当位置部署具有差异化性能要求的边缘网络切片，可提高车联网整体资源利用率。同时，车联网固有的动态性和开放性使网络切片对超可靠低时延通信（URLLC, ultra-reliable and low latency communication）的需求不可或缺，并可根据服务水平协议需求的差异化，支持跨多个无线接入网络的服务，为动态接入和效率服务定制灵活专用的资源分配和性能保证策略。在车联网中启用网络切片，不仅可以降低车载应用时延，还可支持车载服务的流量优先级排序^[22]。

1.2.2 软件定义网络

软件定义网络（SDN, software defined network）是一项将网络资源抽象至虚拟化系统的基础架构方法，其本质是通过网络软件简化网络管理和运

维。SDN 可在现有物理网络上构建虚拟的逻辑网络层，从而将控制层功能从数据层中解耦，并迁移至虚拟网络层，最终由逻辑中心化和可编程的集中控制器统一处理^[24]。

近年来，SDN 与边缘计算结合使车联网的逻辑集中控制更为可靠，通过灵活性、可扩展性及可编程性的优势来简化数据转发功能，相关优化内容及服务保障、定制开发以及扩展和收缩网络资源。具体来说，鉴于在车联网中部署更多的 RSU 以应对车辆保有量增长导致的系统整体成本骤增，SDN 作为其中的关键推动因素，为全局网络配置、基于成本效益的自适应资源分配和车联网下异构元素聚合，提供了潜在的解决途径。车联网中异构的 SDN/NFV 架构如图 2 所示，SDN 架构由下到上分为数据层、控制层和应用层。一般情况下，集中控制器部署于 RSU 边缘服务器中，收集全局网络信息，包括流量负荷、车辆密度、服务类型、节点资源容量等^[24]。利用所收集信息，集中控制器在为业务流部署自适应路由协议的同时，可通过南向接口对资源切片和访问控制进行网络级配置，以提高资源利用率并降低整体运营成本。然而，由于控制器依赖全局控制视图，因此在高度动态化的网络拓扑条件下，如何应用无信令开销的 SDN 值得深究。

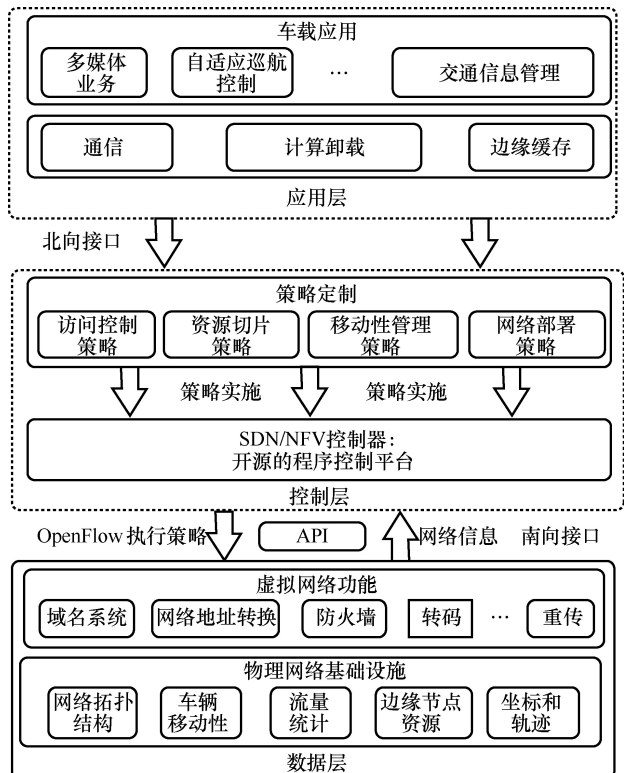


图 2 车联网中异构的 SDN/NFV 架构

1.2.3 网络功能虚拟化

网络功能虚拟化 (NFV, network function virtualization) 是对 SDN 技术的补充, 主要用于将负载均衡、域名系统、网络地址转换、视频转码等网络功能与底层硬件服务器解耦^[24]。其目的在于通过虚拟化技术将网络功能编程为软件实例, 为 SDN 软件提供可运行的基础架构。

一般情况下, NFV 运行于连接 RSU 的边缘服务器中, 以实现面向计算的服务定制和分发, 其中在边缘服务器的可编程软件实例通常被称为虚拟网络功能单元^[25]。通过 NFV, 边缘服务运营商可灵活地在不同服务器上运行相应功能, 或移动部分功能以应对车辆服务的需求变化, 从而降低整体运营成本并提高服务交付效率。事实上, SDN 和 NFV 在技术上互补互利, 借助 NFV 和 SDN 可创建更加灵活、可编程且高效利用资源的网络架构。如图 2 所示, 在车联网场景下, SDN 与 NFV 的集成架构可实现灵活的交通路由管理、网络级资源切片, 并通过对车辆端执行信道访问控制, 最终实现高效的资源分配。

2 基于边缘智能的服务类型

为构建健硕的边缘智能应用体系, 计算、缓存及云-边-端协同服务被广泛应用于车联网。本节主要针对此类 3 项特定研究问题及其相关工作进行整体概述。

2.1 计算卸载与资源分配

计算卸载是车联网边缘智能被广泛讨论的特性之一, 资源受限的车辆可通过车载无线通信技术, 将计算密集或时延敏感型业务负载直接卸载至临近 RSU 边缘服务器, 利用边缘服务器的算力资源协助进行计算^[9,17]。该模式在满足车辆算力扩展需求的同时平衡了云计算技术的局限性, 有效地提高了车联网应用的 QoS。

面向多用户服务场景, 基于计算卸载的服务优化需要综合考虑任务数据量、任务时效性、信道传输及边缘服务器计算能力等诸多约束。同时, 由于影响卸载决策的主要因素包括任务执行时延与能耗, 因此计算卸载的优化目标主要分为降低时延、减少能耗、权衡时延与能耗 3 个方面。文献^[26]回顾了车辆低时延、高效率计算需求与计算卸载的关联, 阐述了计算卸载于车联网中的发展潜力。Mao 等^[27]研究了基于非正交多址接入的边缘系统中, 功率控制、处理器频率和计算卸载于离散模型下的联合优

化, 利用块坐标下降法解决了所规划的非凸问题。Zhang 等^[28]提出了基于云计算的多层车辆边缘网络框架, 采用 Stackelberg 博弈设计了以最小化车载应用时延为目标的分级卸载方案。此外, 文献^[29]提出了基于异步优势 Actor-Critic 算法的自适应计算卸载框架, 实现计算任务能耗和时延的联合优化。

此外, 根据单个任务能否被分割, 可在边缘侧部署两类卸载模型, V2I 计算卸载架构下的两类卸载模式如图 3 所示。一类被称为完全卸载, 在该场景下, 整个流程或任务都将被卸载至边缘服务器^[5]; 另一类被称为部分卸载, 在该场景下, 任务或流程将被划分为多个细粒度的子任务, 并被卸载至不同的边缘服务器进行分布式处理。基于部分卸载, Ning 等^[30]考虑了移动用户间的资源竞争, 提出了一项迭代启发式方法对系统中的协同卸载决策执行自适应调配。Wang 等^[31]研究了部分卸载于车载网络中的应用和部署, 并利用深度强化学习算法研究了 QoS 约束下的多车计算卸载的效益优化问题。同时, 随着无线接入技术的发展, 更加丰富的算力资源可通过以车辆为载体的车-车通信 (V2V, vehicle to vehicle) 技术得到充分利用。此时, 作为计算资源“产消者”, 具有闲置资源的临近车辆可按需被调度, 协助边缘服务器共同服务于执行计算卸载的指定车辆。Zhou 等^[32]通过反向拍卖机制激励边缘节点参与 V2V 卸载, 并将激励驱动下的 V2V 卸载过程建模为整数非线性规划问题, 以减少边缘服务提供商的成本。因此, 计算卸载研究主要聚焦最优卸载决策, 需要全局考虑各方面因素, 如任务是否需要卸载、于何处卸载及卸载多少等。

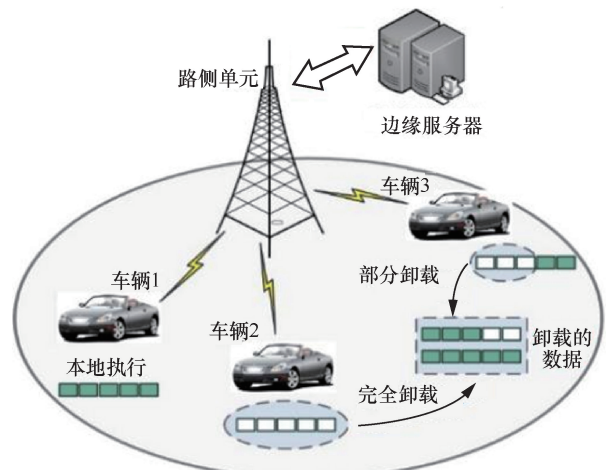


图 3 V2I 计算卸载架构下的两类卸载模式

最后, 卸载过程中综合性的资源分配方案也是

保证总体用户 QoS、提升车联网性能的重要因素。资源分配指调配服务器中的可用资源优化系统配置，提升资源利用率，同时降低任务处理时延或能耗，提供更好用户体验。在资源配置过程中，系统可基于动态条件、服务类型、任务最大可容忍时延、车辆密度、车载服务异构性、优先级等因素，对车联网边缘侧的各类资源进行融合，并随车联网业务变化实施随动迁移。在边缘智能驱动的车联网中，预定义的资源分配策略必须具备自适应和可伸缩性，算力和通信资源分配应该被有机调配，以应对各边缘节点在时空域上的高度动态性。然而，以往研究较少考虑车联网的隐藏动态特性，因此如何在卸载过程中设计自适应的资源分配方案也是研究重点之一。为此，针对时变信道条件，Qian 等^[33]基于深度强化学习在线算法，针对多任务计算卸载、非正交多址的接入传输和资源分配进行联合优化。Yan 等^[34]针对依赖模型下任务的组合卸载和强耦合问题，提出了一项基于 Critic 网络的低复杂度算法。Tan 等^[35]考虑了车辆在不同时间尺度下的移动性，提出了一种基于深度 Q 网络的在线计算卸载与资源分配框架。

2.2 边缘缓存

车载应用需要车辆访问大量的互联网数据（如实时交通信息下载、导航地图更新等），导致车联网存在大量的冗余流量负载和较长的访问时延。然而，云计算架构的集中化特点，面临传输距离长、信道带宽有限等挑战，致使车载应用的支持大规模内容交付的同时，无法满足高依赖的网络 QoS 要求。研究表明，不同流行度的内容通常需适配相应优先级。显然，只有少数请求度高的内容被大多数车辆重复请求，而其余大部分内容仅有较低访问需求^[10]。

边缘缓存可有效缓解车联网内部内容重复传输所导致的冗余回程流量，在降低内容访问时延及成本的同时提高内容交付可靠性。分布式边缘缓存架构如图 4 所示，边缘缓存利用低成本的缓存单元（如 RSU）的存储资源，将流行度较高内容缓存于靠近车辆的边缘侧。因此，车辆能够直接从附近启用缓存服务的 RSU 上访问流行内容，而无须从远程的云服务器重复下载。此外，相邻的 RSU 间通常可相互通信并共享内容，因此还需考虑边缘节点间的协同以提高系统的全局缓存利用率^[21]。同时，V2V 通信技术将使车辆的存储单元能够基于与其他车辆的接触，进行机会式内容共享^[11]。对此，Alnagar 等^[36]

考虑了车辆速度和内容流行度对缓存决策的影响，开发了一种基于次优松弛和背包问题的缓存策略，以提高边缘节点的缓存命中率。Ao 等^[37]考虑了分布式缓存和协同多点技术，研究了物理层资源调配和缓存策略间的相互依赖问题。针对 V2V 和 V2I 的混合通信模式，Wu 等^[38]考虑了车辆和 RSU 的地理分布及传输内容大小，提出了一项基于能量感知的内容缓存方案。

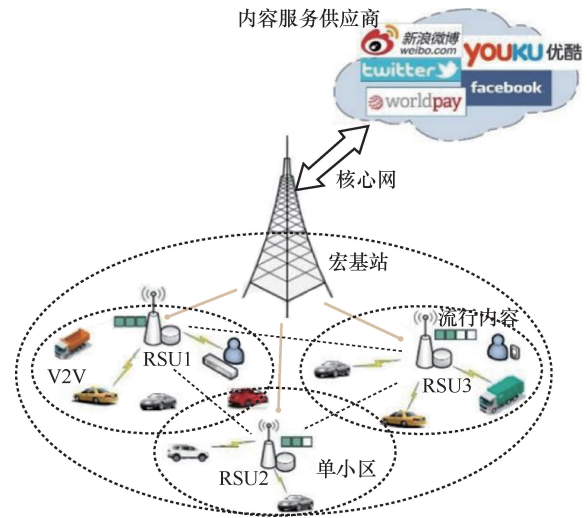


图 4 分布式边缘缓存架构

目前，边缘缓存的研究重点包括内容交付和缓存替换两方面^[12,39]。内容交付指某一 RSU 基于内容的时效性和其自身缓存状态，对通信范围内的内容请求做出相应交付决策。在内容交付阶段，RSU 首先搜索其缓存空间以查询是否缓存所请求内容，若未搜索到所请求内容，则从相邻 RSU 获得或直接从远端云服务中心下载此内容，最终再由 RSU 将该内容交付至所请求车辆。缓存替换指各相互协作的 RSU 自适应地根据内容流行度替换其自身缓存内容^[39]。在缓存替换阶段，RSU 可根据车辆请求，更新其缓存内容，以提升自身缓存内容的命中率和有效性，同时尽可能地减少系统中的内容交付成本^[35]。因此，当某一 RSU 缓存容量被完全占用时，RSU 本地的内容替换机理也是边缘缓存中极其重要而迫切的研究课题之一。

针对传统方法在车联网动态环境下适应性不强的问题，基于 AI 的方法已被用于提高边缘缓存的智能化。Tan 等^[35]在 V2V 缓存和编码技术的基础上，利用深度 Q 网络分别对不同时间尺度下的车辆移动性和缓存决策进行优化。Qian 等^[40]利用训练长短期记忆网络对内容流行度进行预测，并提出了基于深度学习的内容替换策略以提升缓存命中率。

Wang 等^[21]将强化学习中的马尔可夫决策过程扩展至多智能体系统,提出了基于多智能体强化学习的边缘缓存协同框架。此外,为保证车辆数据的本地化训练,Wang 等^[41]提出了一种基于联邦深度强化学习的协同边缘缓存框架,以有效应对集中式方法在模型训练和数据传输过程中的资源消耗。

2.3 “云-边-端”协同

随着业务向协同运营阶段的纵深发展,边缘智能的范围不再局限于单个层面。现有的诸多研究将“云-边-端”协同集成至边缘智能范式的设计中,缺少从宏观层面上对协同感知机理的深入研究及对感知态势的协调控制,面临来自异构资源、层次管理、业务协同方面的挑战。

2.3.1 “云-边”协同

近年来,“云-边”协同已发展为一类较为成熟的协作模式,并引起学术界和产业界的广泛关注。其中,边缘端负责本地范围内的数据计算及缓存,或进一步将采集的数据汇聚至云端处理,可较好地支持本地、短周期的智能决策和执行。相比之下,云端负责采集数据的分析和挖掘、模型的训练和升级,利用其强大的计算能力为长期、大规模的智能处理和资源调度建立管理平台,支持车辆智能、多元化的数据服务。该协作模式可有效促进云、边、端之间的资源充分利用,并为任何时间、地点的跨异构协作平台提供无缝流畅的网络服务,从而实现跨区域、跨网络的实时流量优化、动态交通控制和驾驶辅助支持等。Taleb 等^[8]介绍了“云-边”协同的参考体系架构及其主要应用场景。Tang 等^[9]结合动态网络下的不确定因素,提出了基于“云-边”协同的任务卸载和资源分配框架。Song 等^[23]基于多层“云-边”协同架构,设计了针对 5G 前端网络切片的资源调度方案,提升了业务分发中的资源利用率。

2.3.2 “边-边”协同

单个边缘节点在处理大规模应用时可能会受算力、存储和通信资源的限制,而其他节点由于服务时空分布的差异性,自身存在一定数量的闲置资源。为提高系统的总体资源利用率,迫切需要在集群形式的边缘节点间构建协同感知和结果共享的广泛互联,以打通信息“孤岛”。因此,在“边-边”协同模式下,需要充分利用各边缘节点间的协同共同保障计算的优化。需要注意的是,在“边-边”协同模式下处理大规模应用时,通常存在资源需

求、异构系统条件及边缘节点间接入的动态变化,同时单个边缘节点的资源容量也具有时变特征。因此,需要设计有效的协同调度策略,以支持数据在边缘间的可控有序流动,完成自主学习闭环。Jiang 等^[12]在随机博弈的框架下利用多智能体强化学习设计了基于“边-边”协同的合作缓存方案。张星洲等^[42]在联邦学习算法和长短期记忆网络的基础上,提出了基于“边-边”协同的电动汽车电池故障检测系统。

2.3.3 “边-端”协同

“边-端”协同中的“端”特指终端设备,通常由道路上一系列物联网设备和车辆组成。“边-端”协同是一类轻量级模型,可有效提高边缘节点的处理能力,缓解请求多样性和边缘设备处理能力单一之间与日俱增的冲突。同时,终端设备与特定应用场景间的高度相关性,使得“边-端”协同更加关注终端设备的高效调度和安全接入。在“边-端”协同中,终端设备采集实时感知数据,并将其卸载至边缘服务器。边缘服务器对多源数据进行集中计算和分析,并通过向终端设备发送操作标识符提供可靠服务。综上,基于终端设备与用户之间的密切关系,“边-端”协同被认为是实现边缘智能应用的必要步骤。Khan 等^[43]讨论了“边-端”协同在车联网中的应用场景及其所面临的安全隐私问题。Zhou 等^[5]研究了基于“边-端”协同的计算卸载系统,在特定约束下最小化所有用户的计算开销。

分布式架构下 RSU 性能与协作能力的局限性,使得“边-边”协同和“边-端”协同方式难以对全域车辆的任务处理进行全局调配。相关研究已开始探索“云-边-端”的有效结合,主要侧重面向信息服务的系统架构、面向“云-边”的算力迁移、交通流状态估计、系统级接入技术、服务迁移连续性机制等,但大多研究仍然不具备实现车联场景下融合感知、决策与控制的“云-边-端”一体化概念,难以支撑多样化业务需求。为此,李克强等^[44]引入基于信息物理系统的“云-边”协同体系架构,开发了面向“云-边-端”融合感知和车辆控制技术的云控系统。该系统通过“云-边-端”的融合感知将物理系统层、信息映射层和融合应用层连为一体,为协同应用提供实时运行环境、交通全要素数字映射、标准化通信机制以及资源平台数据,进而根据业务需求进行分层融合决策,实现对车载任务处理的统一编排与管理。

3 边缘智能模型的部署与实施

边缘智能已在车联网领域崭露头角，其价值在众多场景中都得到了体现。本节阐述车联网中 AI 模型的训练和推断过程。

3.1 边缘智能车联系统中的模型训练

车联网环境下的边缘计算和 AI 融合，依赖于边-云连续体（edge-cloud continuum）中高效的模型训练和推断，这对于实现高质量的服务部署至关重要。训练模式可分为集中式训练、分散式训练和混合式训练 3 种。

3.1.1 集中式训练模式

在集中式训练模式中，训练后的模型部署于云计算平台，而数据预处理、模型训练、消息代理等工作都主要由云计算平台执行。具体而言，模型训练主要通过边云协同实现，其性能在很大程度上依赖于网络连接的质量。在训练阶段，RSU 负责采集覆盖范围内的道路信息（此类信息由车辆产生，涉及车载服务、传感器、无线信道和交通流数据），并实时将其上传至云端进行即时处理。基于数据分析和存储，云端将利用聚合数据于集中式训练集群中不断训练模型。值得注意的是，尽管集中式训练模式有潜力检索到系统的全局最优解，但由于系统必须依赖全局网络状态信息，模型训练的复杂度会随网络规模的增加呈指数增长。同时，由于部署于云计算平台上的模型在空间距离上远离车辆，上传数据时需通过广域网络。因此，数据传输和不可预测的网络连接状态，或将导致时延高、效率低、带宽成本高昂等缺陷。此外，车辆数据资源的高度集中性使得该模式在受到网络攻击时，更易导致敏感交通数据的泄露和丢失。

3.1.2 分布式训练模式

相比于依赖某一中心化节点（云端）进行全局模型更新，基于节点间参数和梯度互信互通的分布式训练架构或将更适合边缘侧环境。事实上，每个 RSU 作为独立节点，应根据其感知局部状态生成各自推断，但不同 RSU 决策间可能会存在显著差异。与此同时，车联网希望在保护隐私的同时，从边缘智能部署中受益。因此，为规避相关缺陷，需要在边缘侧以高可靠、分布式的方式训练模型。这一集成架构中所有 RSU 的训练均为等价，其通过去中心化的方式从根本上避免了单点失效隐患，同时提高了网络的可拓展性和安全性。然而，受制于数据

源之间的时间相关性，独立部署于 RSU 的模型训练过程面临过拟合问题，而车联网中多个 RSU 间的推断也通常相互影响。因此，多 RSU 间必须以协同方式进行模型训练或数据分析，以共享局部训练并以此构建和改进全局训练模型。

与集中式训练相比，分布式训练具有隐私保护、个性化学习、可拓展性强的优势，但在无云化条件下，其缺少全局参数汇聚的过程，意味着参数交换的拓扑结构对于模型的收敛性能和训练效果至关重要，因此在很大程度上受资源有限、环境动态、数据分布、设备异构性的影响。同时，分布式环境下节点间模型训练的信息协调的方式包括不同层级，对 RSU 算力的要求各有不同，需基于服务部署代价和部署收益等多个量化指标研究训练模型的部署问题。此外，联邦学习可在分布式训练下应用于数据敏感领域。需注意的是，分布式训练侧重于在边缘端解析数据，而联邦学习则更侧重于分布式隐私保护。

3.1.3 混合式训练模式

实际上，受能耗、算力和存储资源限制，单个 RSU 独立训练和部署的 AI 参数规模有限。因此，多种训练模式应考虑相互之间的兼容性。为更好地发挥多点协调优势，车联网中普遍采用集中式和分布式相结合的混合式训练模式，有望打破单一训练模式的性能瓶颈。在该模式下，RSU 通过彼此间的分布式更新或云平台的集中训练协作训练 AI 模型。具体来说，每个 RSU 可根据其本地数据训练部分参数，并将参数或梯度聚合至某一中心节点进行全局模型升级，而后中心节点将全局模型下发至各 RSU。经过 RSU 和中心节点间多轮次通信和迭代后，该全局模型可达到与集中式训练相近的性能。此时，私有数据在训练中始终存储于本地，使得混合式训练模式所导致的隐私保护比分布式模式弱，但无疑比集中式训练模式更强。

此外，目前降低模型训练复杂度（包括样本复杂度和计算复杂度）主要从系统级和方法级两方面开展。其中，系统级解决方案致力于调配更优的决策和训练方法，而方法级解决方案则倾向于制定更优的通信系统模型或引入环境先验知识。同时，为减少模型训练过程对算力资源的依赖性，相关研究考虑在不影响精确度的情况下对模型进行剪枝处理，典型方法包括在训练过程中丢弃非必要权重及神经元、稀疏训练、输出重建误差等。其中，模型

剪枝示例如图 5 所示，由于该多层感知网络中许多神经元的值为零，此类神经元在计算过程中并不起作用，因而可以被裁剪，以减少训练过程中的算力和存储需求。同时，针对模型剪枝优化，文献[45]比较了部分权重分解与剪枝方法，并从精度、参数大小、中间特征大小、处理时延及能耗等方面进一步讨论对比方法的优势和瓶颈。

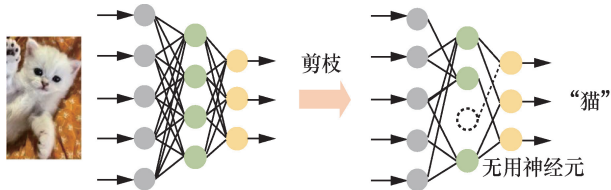


图 5 模型剪枝示例

3.2 边缘智能车联系统中的模型推断

模型推断发生在模型训练之后，主要是运行已训练完备的模型对未知结果的数据进行预测。模型推断与模型训练相互结合，是一个往复循环、不断提升的过程，有效的模型推断对于车联网下边缘智能的实施至关重要。根据上述模型类别，可分别在云端或独立的 RSU 上执行模型推断。3 种典型的模型训练和推断模式见表 2。

表 1 3 种典型的模型训练和推断模式

模型训练	模型推断	云(中心节点)	路侧单元
集中式	集中式	训练 + 推断	N/A
集中式(共享模型)	分布式	训练	推断
分布式	分布式	N/A	训练 + 推断

具体而言，典型的模型推断方式包括集中式和分布式。集中式推断中，云端需收集所有 RSU 的信息（即全局信息），此时模型训练和推断都将在

云端完成，推断结果将被分别分发给各 RSU；分布式推断中，各 RSU 仅需在本地根据各自信息（即本地信息）执行本地模型推断，此时，任意 RSU 还可与其他 RSU 交换部分模型信息，以提升分布式推断性能。相较于集中式推断，分布式推断具有计算和通信能耗低、决策响应时间短、可扩展性强等优点，更适用于网络状态变化快、对时延和能耗要求高的车联网场景。另外，在集中式训练模式中，云端既可维护一套全局训练模型，以便对所有 RSU 进行集中推断，同时也可先为所有 RSU 训练一套共享模型，再将训练完备的共享模型下发至各 RSU 进行分布式推断。一般来说，集中式推断的常用方法包括监督学习、无监督学习和单智能体强化学习；而分布式推断的常用方法包括无监督学习和多智能体强化学习。车联网中的模型训练与推断过程如图 6 所示。

另外，模型推断加速也是边缘服务优化的主要方向。调节推断时间的优化手段分为模型精简和模型切分。模型精简指根据节点、任务、模型的特征，动态选择最适合此节点的模型，即在适当牺牲模型精确度的前提下，选择对资源需求更低，完成时间更快的“小模型”。其中，不同模型精简方法的主要区别在于评价指标的差异化，Xu 等^[46]考虑了用户的服务水平协议需求和节点资源的使用量，而 Taylor 等^[47]则将任务特征和期望精度作为推断评估的指标。模型切分指基于模型中神经网络的层次化计算结构，对计算任务进行分层或网格切分，该方法可充分利用边缘侧和云端的计算特征，进行协同推断以实现推断加速。此时，选择不同的模型切分点将导致不同的计算时间，而最佳的模型切分点能最大

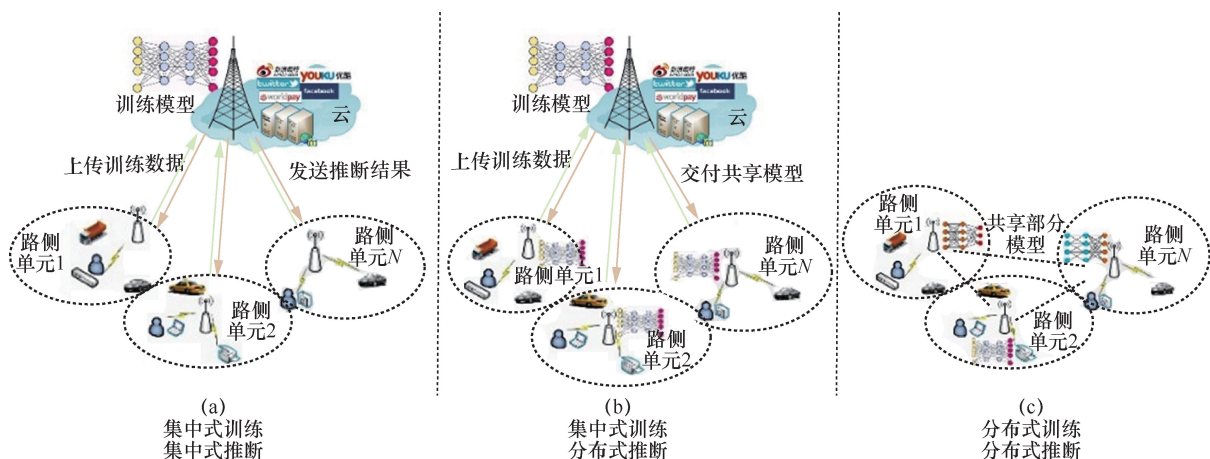


图 6 车联网中的模型训练与推断过程

限度发挥协同推断的优势。Kang 等^[48]设计了一项分层结构切分的方法，该方法根据神经网络层的粒度，在移动设备和数据中心间确定任务切分的最优点。Karsavuran 等^[49]利用稀疏张量的交替最小二乘方法，提出了一种通用中等粒度切分的超图模型，该模型对任务切分不施加任何拓扑约束。

4 挑战与展望

本节探讨了开放性挑战及后续研究方向，为车联网中边缘智能的部署提供了理论支撑。

4.1 系统动态性和开放性

车联网边缘智能的部署需考虑网络结构的动态性和信道开放性。一方面，车联网服务环境非稳态，车辆移动性导致的多 RSU 间频繁切换，或将引发节点间网络拓扑的动态改变以及车联业务的时空迁移，从而难以适配不同路段、不同服务节点群组的边缘资源需求。同时，移动性差异使得特定车辆于 RSU 覆盖范围内的联通时间具有时变性。以上动态因素对 AI 模型的训练时间和能耗都提出了更严格的要求，尤其是在某些需要累积残差的技术^[23,38]中，移动性将导致无法有效地累积残差，从而大大减缓模型的收敛速度。另一方面，考虑 RSU 被部署于开放环境，车辆与 RSU 间的通信质量与模型精度和收敛速度息息相关。信道质量通常受车辆与 RSU 间的传输距离、带宽资源状态、接入方式以及网络拓扑等多种因素影响，呈现高度不稳定性，V2X 中多种通信模式的混合作用将进一步增加数据传输过程中的复杂性。综上，边缘智能驱动的车联网架构需具备复杂多变条件下的泛化与鲁棒控制能力。

4.2 硬件及网络架构支持

资源密集且时延敏感的车载服务通常需要适配大量资源以进行高维系统参数配置。边缘设备型号、结构、操作系统和开发环境的差异化，将导致现有平台无法对其进行统一管理和运维，因此考虑异构硬件设备间的兼容和协调具有重要意义。此外，针对车载无线信道的开放性特性，基于竞争的信道接入条件，以及 RSU 有限的通信覆盖范围，迫切需要设计资源友好化的网络架构以缓解有限资源下的模型训练和推断压力。事实上，先进的网络架构可通过对边缘节点进行有效的资源优化和任务调度，以确保车辆和 RSU 间的 URLLC。然而，在多样性需求下管理和协调 AI 负载仍然是迫切需

要解决的问题之一，网络架构要为 AI 负载提供新的资源抽象，以满足其对芯片、容量扩展和任务依赖性管理的需求。最后，在如何打破各类通信技术间壁垒以实现网络架构的可控性方面，也需要进行进一步的研究。

4.3 训练模型轻量化

车联网边缘智能模型通常部署于不同 RSU，通过复杂神经网络结构实现决策控制、资源配置、网络管理等高级功能。在执行模型推断时，单一 AI 技术的优化效果可能会面临难以推演的限制。虽然可在模型训练中并行多种 AI 技术以解决过拟合或欠拟合问题，但此类技术的集成也会使网络结构进一步复杂多样化。因此，有必要研究如何在资源受限的 RSU 上部署和执行 AI 负载。纵观不同技术路线^[6,9,19]，文献工作大多力求构建资源高效型、快收敛、轻量级的 AI 模型以适应资源约束。早期的工作包括稀疏化、知识蒸馏、剪枝和量化等模型压缩技术，可为车联网中轻量级 AI 模型提供支持。然而，尽管可通过调整模型规模以支持边缘侧 AI 的运行，但该类举措伴随模型精度的损失。事实上，静态模型压缩方法已无法适用于动态硬件配置和边缘节点负载，情感信息压缩技术有望得到应用。最后，根据应用特点，可开展轻量级虚拟化和神经体系结构搜索等技术探索，以提升资源受限环境下的模型部署能力^[29]。

4.4 提升安全和隐私保护

保障边缘的安全运行环境也是支持车联网服务供给的先决条件之一。车辆移动性及通信单元间频繁切换或将影响系统整体稳定性，进而阻断通信接入的稳定供给，给分布式系统带来安全威胁。此类开放性对面向集成学习的分布式信任提出了极高的要求，其主要聚焦分散部署的 RSU 所提供的服务可信度^[13]。更重要的是，作为具备泛在无线接入能力的开放式网络，数据汇聚至边缘或云端执行模型训练时易被全局攻击者窃取。此时，黑客对节点和信道的信息干扰及数据攻击将导致严重的隐私泄露。因此，数据脱敏和隐私保护对于确保车联网中用户认证、访问控制、数据完整性和跨平台验证至关重要^[43]。尽管联邦学习等技术已被应用于隐私友好型的分布式训练，但模型参数或原始数据集仍可被非可信第三方推断和重构。对此，现有的解决方案包括添加噪声保护原始数据、通信认证、同态加密和差分隐私。此外，由于恶意节点可能会篡

改和伪造边缘端的正常业务进程,对节点不良行为的识别和恶意节点的移除具有重要意义。同时,需要针对决策指令、交互记录等信息进行验证和重加密,并充分利用车辆交通流的可预测性和时空相关性,以满足服务交付。

4.5 基于网络经济的激励机制

对于协作的分布式系统来说,需要注意的问题是基于网络经济的激励机制。由于边缘生态系统通常涉及不同的服务提供商和车辆,服务运营往往需要考虑车联网中跨服务提供商的协作和集成。同时,车联网中的车辆从属于不同个体,个体逐利性特征将促使其追求更强大的算力和更快速的处理响应,或将引发多车辆、多RSU间的无序资源竞争和信道干扰问题^[6]。尽管针对多个RSU的资源进行联合调配可有效提升边缘资源利用率,但在实施中难以保证车载应用无条件地遵循相关策略。因此,可基于车载应用的自有特征和个体潜在因素,探讨如何面向边缘智能平台的算力和通信资源进行定价,以及如何设计具有内驱能力的激励机制以促进多个实体间的有序合作。最后,在进行多体博弈时需要考虑公平与优化间的折中关系,为此有必要设计资源友好的轻量级区块链共识协议以平衡多体间的相关指标。

5 结束语

本文针对车联网应用场景,从更为广阔的视角对边缘智能驱动的车联网进行了全面的综述。本文首先介绍了边缘智能的背景、概念及关键技术;然后,本文对车联网应用场景中基于边缘智能的服务类型进行整体概述,同时详细阐述了边缘智能模型的部署和实施过程;最后,本文探讨了边缘智能于车联网应用场景中的关键开放性挑战,以推动其潜在研究方向。

参考文献:

- [1] JIAU M K, HUANG S C, HWANG J N, et al. Multimedia services in cloud-based vehicular networks[J]. *IEEE Intelligent Transportation Systems Magazine*, 2015, 7(3): 62-79.
- [2] PENG J K, FAN Y, YIN G D, et al. Collaborative optimization of energy management strategy and adaptive cruise control based on deep reinforcement learning[J]. *IEEE Transactions on Transportation Electrification*, 2022.
- [3] HUANG J H, CUI H X, CHEN C. Cluster-based radio resource management in dynamic vehicular networks[J]. *IEEE Access*, 2022(10): 43562-43570.
- [4] KAIWARTYA O, ABDULLAH A H, CAO Y, et al. Internet of vehicles: motivation, layered architecture, network model, challenges, and future aspects[J]. *IEEE Access*, 2016, 4(2): 5356-5373.
- [5] ZHOU H, JIANG K, LIU X X, et al. Deep reinforcement learning for energy-efficient computation offloading in mobile-edge computing[J]. *IEEE Internet of Things Journal*, 2022, 9(2): 1517-1530.
- [6] 刘婷婷, 杨晨阳, 索士强, 等. 无线通信中的边缘智能[J]. *信号处理*, 2020, 36(11): 1789-1803.
- [7] LIU T T, YANG C Y, SUO S Q, et al. Edge intelligence for wireless communication[J]. *Journal of Signal Processing*, 2020, 36(11): 1789-1803.
- [8] ABBAS N, ZHANG Y, TAHERKORDI A, et al. Mobile edge computing: a survey[J]. *IEEE Internet of Things Journal*, 2018, 5(1): 450-465.
- [9] TALEB T, SAMDANIS K, MADA B, et al. On multi-access edge computing: a survey of the emerging 5G network edge cloud architecture and orchestration[J]. *IEEE Communications Surveys & Tutorials*, 2017, 19(3): 1657-1681.
- [10] TANG M, WONG V W S. Deep reinforcement learning for task offloading in mobile edge computing systems[J]. *IEEE Transactions on Mobile Computing*, 2022, 21(6): 1985-1997.
- [11] JIANG W, FENG G, QIN S, et al. Multi-agent reinforcement learning for efficient content caching in mobile D2D networks[J]. *IEEE Transactions on Wireless Communications*, 2019, 18(3): 1610-1622.
- [12] LI X H, WANG X F, WAN P J, et al. Hierarchical edge caching in device-to-device aided mobile networks: modeling, optimization, and design[J]. *IEEE Journal on Selected Areas in Communications*, 2018, 36(8): 1768-1785.
- [13] JIANG K, ZHOU H, ZENG D Z, et al. Multi-agent reinforcement learning for cooperative edge caching in Internet of vehicles[C]//*Proceedings of 2020 IEEE 17th International Conference on Mobile Ad Hoc and Sensor Systems*. Piscataway: IEEE Press, 2020: 455-463.
- [14] GU B, GAO L X, WANG X D, et al. Privacy on the edge: customizable privacy-preserving context sharing in hierarchical edge computing[J]. *IEEE Transactions on Network Science and Engineering*, 2020, 7(4): 2298-2309.
- [15] 张彦, 张科, 曹佳钰. 边缘智能驱动的车联网[J]. *物联网学报*, 2018, 2(4): 40-48.
- [16] ZHANG Y, ZHANG K, CAO J Y. Internet of vehicles empowered by edge intelligence[J]. *Chinese Journal on Internet of Things*, 2018, 2(4): 40-48.
- [17] XU X L, LI H Y, XU W J, et al. Artificial intelligence for edge service optimization in internet of vehicles: a survey[J]. *Tsinghua Science and Technology*, 2022, 27(2): 270-287.
- [18] ZHOU Z, CHEN X, LI E, et al. Edge intelligence: paving the last mile of artificial intelligence with edge computing[J]. *Proceedings of the IEEE*, 2019, 107(8): 1738-1762.
- [19] JIANG K, SUN C, ZHOU H, et al. Intelligence-empowered mobile

- edge computing: framework, issues, implementation, and outlook[J]. *IEEE Network*, 2021, 35(5): 74-82.
- [18] XU D L, LI T, LI Y, et al. Edge intelligence: empowering intelligence to the edge of network[J]. *Proceedings of the IEEE*, 2021, 109(11): 1778-1837.
- [19] WANG F X, ZHANG M, WANG X X, et al. Deep learning for edge computing applications: a state-of-the-art survey[J]. *IEEE Access*, 2020, 8(1): 58322-58336.
- [20] LUONG N C, HOANG D T, GONG S M, et al. Applications of deep reinforcement learning in communications and networking: a survey[J]. *IEEE Communications Surveys & Tutorials*, 2019, 21(4): 3133-3174.
- [21] WANG F X, WANG F, LIU J C, et al. Intelligent video caching at network edge: a multi-agent deep reinforcement learning approach[C]//*Proceedings of IEEE INFOCOM 2020 - IEEE Conference on Computer Communications*. Piscataway: IEEE Press, 2020: 2499-2508.
- [22] CHAHBAR M, DIAZ G, DANDOUSH A, et al. A comprehensive survey on the E2E 5G network slicing model[J]. *IEEE Transactions on Network and Service Management*, 2021, 18(1): 49-62.
- [23] SONG C, ZHANG M, ZHAN Y Y, et al. Hierarchical edge cloud enabling network slicing for 5G optical fronthaul[J]. *Journal of Optical Communications and Networking*, 2019, 11(4): B60-B70.
- [24] 张朝昆, 崔勇, 唐鬲鬲, 等. 软件定义网络 (SDN) 研究进展[J]. *软件学报*, 2015, 26(1): 62-81.
- ZHANG C K, CUI Y, TANG H H, et al. State-of-the-art survey on software-defined networking(SDN)[J]. *Journal of Software*, 2015, 26(1): 62-81.
- [25] HAN K, LI S R, TANG S F, et al. Application-driven end-to-end slicing: when wireless network virtualization orchestrates with NFV-based mobile edge computing[J]. *IEEE Access*, 2018, 6(1): 26567-26577.
- [26] ZHANG J, LETAIEF K B. Mobile edge intelligence and computing for the internet of vehicles[J]. *Proceedings of the IEEE*, 2020, 108(2): 246-261.
- [27] MAO S, LENG S P, ZHANG Y. Joint communication and computation resource optimization for NOMA-assisted mobile edge computing[C]//*Proceedings of ICC 2019 - 2019 IEEE International Conference on Communications*. Piscataway: IEEE Press, 2019: 1-6.
- [28] ZHANG K, MAO Y M, LENG S P, et al. Optimal delay constrained offloading for vehicular edge computing networks[C]//*Proceedings of 2017 IEEE International Conference on Communications*. Piscataway: IEEE Press, 2018: 1-6.
- [29] CHANG Z, LIU L Q, GUO X J, et al. Dynamic resource allocation and computation offloading for IoT fog computing system[J]. *IEEE Transactions on Industrial Informatics*, 2021, 17(5): 3348-3357.
- [30] NING Z L, DONG P R, KONG X J, et al. A cooperative partial computation offloading scheme for mobile edge computing enabled Internet of Things[J]. *IEEE Internet of Things Journal*, 2019, 6(3): 4804-4814.
- [31] WANG J F, LV T J, HUANG P M, et al. Mobility-aware partial computation offloading in vehicular networks: a deep reinforcement learning based scheme[J]. *China Communications*, 2020, 17(10): 31-49.
- [32] ZHOU H, WU T, ZHANG H J, et al. Incentive-driven deep reinforcement learning for content caching and D2D offloading[J]. *IEEE Journal on Selected Areas in Communications*, 2021, 39(8): 2445-2460.
- [33] QIAN L P, WU Y, JIANG F L, et al. NOMA assisted multi-task multi-access mobile edge computing via deep reinforcement learning for industrial Internet of Things[J]. *IEEE Transactions on Industrial Informatics*, 2021, 17(8): 5688-5698.
- [34] YAN J, BI S Z, ZHANG Y J A. Offloading and resource allocation with general task graph in mobile edge computing: a deep reinforcement learning approach[J]. *IEEE Transactions on Wireless Communications*, 2020, 19(8): 5404-5419.
- [35] TAN L T, HU R Q. Mobility-aware edge caching and computing in vehicle networks: a deep reinforcement learning[J]. *IEEE Transactions on Vehicular Technology*, 2018, 67(11): 10190-10203.
- [36] ALNAGAR Y, GOHARY R H, HOSNY S, et al. Mobility-aware edge caching for minimizing latency in vehicular networks[J]. *IEEE Open Journal of Vehicular Technology*, 2022(3): 68-84.
- [37] AO W C, PSOUNIS K. Fast content delivery via distributed caching and small cell cooperation[J]. *IEEE Transactions on Mobile Computing*, 2018, 17(5): 1048-1061.
- [38] WU H J, ZHANG J, CAI Z P, et al. Toward energy-aware caching for intelligent connected vehicles[J]. *IEEE Internet of Things Journal*, 2020, 7(9): 8157-8166.
- [39] WANG X F, CHEN M, TALEB T, et al. Cache in the air: exploiting content caching and delivery techniques for 5G systems[J]. *IEEE Communications Magazine*, 2014, 52(2): 131-139.
- [40] QIAO G H, LENG S P, MAHARJAN S, et al. Deep reinforcement learning for cooperative content caching in vehicular edge computing and networks[J]. *IEEE Internet of Things Journal*, 2020, 7(1): 247-257.
- [41] WANG X F, WANG C Y, LI X H, et al. Federated deep reinforcement learning for internet of things with decentralized cooperative edge caching[J]. *IEEE Internet of Things Journal*, 2020, 7(10): 9441-9455.
- [42] 张星洲, 鲁思迪, 施巍松. 边缘智能中的协同计算技术研究[J]. *人工智能*, 2019, 6(5): 55-67.
- ZHANG X Z, LU S D, SHI W S. Research on cooperative computing technologies in edge intelligence[J]. *AI-View*, 2019, 6(5): 55-67.
- [43] KHAN R, KUMAR P, JAYAKODY D N K, et al. A survey on security and privacy of 5G technologies: potential solutions, recent advancements, and future directions[J]. *IEEE Communications Surveys & Tutorials*, 2020, 22(1): 196-248.
- [44] 李克强, 常雪阳, 李家文, 等. 智能网联汽车云控系统及其实现[J]. *汽车工程*, 2020, 42(12): 1595-1605.
- LI K Q, CHANG X Y, LI J W, et al. Cloud control system for intelligent and connected vehicles and its application[J]. *Automotive Engineering*, 2020, 42(12): 1595-1605.
- [45] NAN K M, LIU S C, DU J Z, et al. Deep model compression for mobile platforms: a survey[J]. *Tsinghua Science and Technology*, 2019, 24(6): 677-693.

- [46] XU M T, ALAMRO S, LAN T, et al. CRED: cloud right-sizing with execution deadlines and data locality[J]. IEEE Transactions on Parallel and Distributed Systems, 2017, 28(12): 3389-3400.
- [47] TAYLOR B, MARCO V S, WOLFF W, et al. Adaptive deep learning model selection on embedded systems[C]//Proceedings of the 19th ACM SIGPLAN/SIGBED International Conference on Languages, Compilers, and Tools for Embedded Systems. New York: ACM Press, 2018, 53(6): 31-43.
- [48] KANG Y P, HAUSWALD J, GAO C, et al. Neurosurgeon[J]. ACM SIGARCH Computer Architecture News, 2017, 45(1): 615-629.
- [49] KARSAVURAN M O, ACER S, AYKANAT C. Partitioning models for general medium-grain parallel sparse tensor decomposition[J]. IEEE Transactions on Parallel and Distributed Systems, 2021, 32(1): 147-159.



周欢（1986- ），男，博士，三峡大学教授、博士生导师，主要研究方向为移动社交网络、移动数据卸载、车联网等。



任学锋（1979- ），男，华砺智行（武汉）科技有限公司副总裁、新技术研究院院长，主要研究方向为智能网联汽车、智慧交通等。

[作者简介]



江恺（1995- ），男，武汉大学博士生，主要研究方向为边缘智能、多智能体/深度学习、智能交通系统等。



朱永东（1974- ），男，博士，之江实验室研究员，主要研究方向为未来网络与通信、物联网、车联网等。



曹越（1984- ），男，博士，武汉大学教授、博士生导师，主要研究方向为安全防护、网络计算、交通控制等。



林海（1976- ），男，博士，武汉大学副教授，主要研究方向为网络安全、物联网等。