基于分层策略强化学习的多类型流量差异化路由优化

赵之栩 [1] 刘 坤 [1] 王璐瑶 [1] 夏元清 [1]

摘 要 路由是优化网络资源分配的重要方法. 然而, 传统路由算法依赖静态策略优化单一服务质量指标, 难以应对多类型流量爆发性增长下的差异化需求. 尽管深度强化学习为动态网络环境下的路由优化提供了新思路, 现有方法仍缺乏对流量类型的精细化感知能力, 无法灵活调整路由策略. 为此, 本文针对不同类型流量的差异化路由需求, 设计一种基于分层策略强化学习的流量感知路由算法. 首先, 引入流量分类模块, 实现对不同流量差异化业务需求的精细感知. 其次, 利用图卷积网络对网络拓扑进行高效建模, 并在此基础上设计分层决策网络以及差异化奖励函数, 引导智能体生成自适应路由决策, 实现对各流量类别路由策略的动态调整. 同时, 在演员−评论家框架中引入全局注意力机制, 增强智能体对网络状态时空依赖关系的建模能力, 并通过广义优势估计和近端策略优化算法提升训练的效率与稳定性. 最后, 在多种拓扑网络上验证了所提算法的有效性.

关键词 多类型流量; 深度强化学习; 注意力机制; 差异化路由; QoS 优化

引用格式 赵之栩, 刘坤, 王璐瑶, 夏元清. 基于分层策略强化学习的多类型流量差异化路由优化. 自动化学报, 2026, **52**(4): 709−723

DOI 10.16383/j.aas.c250413 **CSTR** 32138.14.j.aas.c250413

Differentiated Routing Optimization for Multi-type Traffic Based on Hierarchical Policy Reinforcement Learning

ZHAO Zhi-Xu[1] LIU Kun[1] WANG Lu-Yao[1] XIA Yuan-Qing[1]

Abstract Routing is an important method for optimizing network resource allocation. However, traditional routing algorithms rely on static strategies to optimize single quality of service metrics, making it difficult to address the differentiated requirements of explosive growth in multi-type traffic. Although deep reinforcement learning has provided new ideas for routing optimization in dynamic network environments, existing methods still lack fine-grained perception of traffic types and cannot flexibly adjust routing strategies. To this end, this paper designs a traffic-aware routing algorithm based on hierarchical policy reinforcement learning for the differentiated routing requirements of different traffic types. First, a traffic classification module is introduced to achieve fine-grained perception of the differentiated service requirements of different traffic. Second, graph convolutional networks are used to efficiently model the network topology, based on which a hierarchical decision network and a differentiated reward function are designed to guide the agent to generate adaptive routing decisions and realize dynamic adjustment of routing strategies for each traffic category. Meanwhile, a global attention mechanism is introduced into the actor-critic framework to enhance the agent′s ability to model the spatio-temporal dependency of network states, and the training efficiency and stability are improved through generalized advantage estimation and proximal policy optimization algorithms. Finally, the effectiveness of the proposed algorithm is verified on various network topologies.

Keywords multi-type traffic; deep reinforcement learning; attention mechanism; differentiated routing; quality of service optimization

Citation Zhao Zhi-Xu, Liu Kun, Wang Lu-Yao, Xia Yuan-Qing. Differentiated routing optimization for multi-type traffic based on hierarchical policy reinforcement learning. *Acta Automatica Sinica*, 2026, **52**(4): 709−723

在全球数字化转型加速推进的背景下, 多任务流呈现爆发性增长态势. 据统计, 截至 2024 年底, 我国 5G 基站已超 419 万个, 5G 用户达 10.14 亿户, 数字经济规模稳居世界第二[1−3]. 与此同时, 网络业务呈现出高度异构化和任务密集化趋势. 新兴应用场景不断涌现, 涵盖 4K/8K 超高清视频、云游戏、智能制造、自动驾驶、远程医疗等多个领域, 不同类型流量对网络的性能需求差异显著[4−5]. 在此背景下, 如何有效识别业务类型、区分其服务需求并实现差异化的服务质量 (quality of service, QoS) 保

收稿日期 2025-08-28 录用日期 2025-12-24
Manuscript received August 28, 2025; accepted December 24, 2025
本文责任编委 张敏灵
Recommended by Associate Editor ZHANG Min-Ling
1. 北京理工大学自动化学院自主智能无人系统全国重点实验室 北京 100081
1. National Key Laboratory of Autonomous Intelligent Unmanned Systems, School of Automation, Beijing Institute of Technology, Beijing 100081

障, 已成为网络调度系统面临的核心问题之一[4].

然而, 传统路由策略如开放最短路径优先 (open shortest path first, OSPF) 协议[6] 所采用的单路径路由算法, 已无法满足现代网络应用对带宽、延迟等 QoS 需求的不断提升. 等价多路径 (equal-cost multipath, ECMP)[7] 虽然提高了网络的负载均衡能力和整体性能, 但其严重依赖静态配置且求解优化问题的复杂度较高、缺乏对流量类型的感知能力、静态策略往往响应滞后, 难以适应多样化的 QoS 需求, 尤其是在高负载场景下, 难以实现精细的流量控制. 尽管软件定义网络 (software defined networking, SDN) 和知识定义网络 (knowledge defined networking, KDN) 等新型架构在提升控制灵活性方面展现出优势, 但在大规模异构网络中, 仍面临路径性能难以实时评估、流量差异难以感知、策略收敛与执行效率不足等关键问题[8−12].

面对上述传统路由策略的不足与局限, 基于深度学习的路由方法凭借其在流量特征提取方面的显著优势, 为解决复杂网络环境下的 QoS 感知难题提供了新的可能. 例如, RouteNet 利用图神经网络学习网络结构与性能的映射关系[13], NeuTM 框架通过长短时记忆网络实现流量矩阵预测[14], 但这类方法多作为辅助模块用于流量预测或链路状态评估, 未直接参与路径选择决策, 且依赖离线训练, 难以应对实时动态的网络环境. 在基于强化学习的路由优化研究中, 已有方法展现出良好的效果. 例如, DRL-TE 方法[15] 利用历史经验进行策略优化, 实现低时延路径的选择. IQoR-LSE 链路状态估计算法[16], 通过结合链路拥塞推断优化动作空间的探索过程, 缓解了复杂策略搜索中的收敛难题, 并显著降低了网络抖动与丢包率. 尽管现有方法在路由优化中展现出一定优势, 但仍存在以下局限: 其一, 流量类型感知缺失, 且未将流量分类结果纳入决策过程, 无法满足多类型流量的差异化需求, 导致网络资源分配失衡; 其二, 全局动态感知能力不足, 无法灵活调整路由策略, 从而降低了网络的整体效率和可靠性.

针对上述挑战, 本文提出一种基于分层策略强化学习的流量感知路由算法 (traffic-aware routing algorithm based on hierarchical policy reinforcement learning, TR-HPRL). TR-HPRL 以分层策略强化学习为核心框架, 通过共享基础特征提取以及类型专属决策输出, 在实现差异化路由策略生成的同时, 有效避免了参数冗余与收敛缓慢问题, 提升了异构流量的 QoS 保障能力. 具体而言, 首先, 利用一维卷积神经网络 (1D-convolutional neural network, 1D-CNN) 对流量进行实时分类, 将其划分为延迟敏感型、丢包敏感型和容错型三类, 并将分类结果作为强化学习状态输入的一部分. 其次, 面向多类型流量的差异化需求, 设计分层策略网络以生成对应的路由策略. 为提升智能体对环境的感知能力, 策略网络采用图卷积网络 (graph convolutional network, GCN) 作为共享编码器, 高效提取全局网络状态特征; 在此基础上, 将共享 GCN 特征输入带有类型特异性策略头与全局注意力机制的演员−评论家 (actor-critic, A-C) 网络, 使生成的差异化路由策略可以适应动态变化的环境, 实现高效响应. 最后, 在多种拓扑网络上训练测试, 验证了所提算法的有效性.

本文其余部分结构如下: 第 1 节介绍多路径路由的相关工作; 第 2 节对多路径路由通信网络以及 QoS 问题进行模型构建; 第 3 节详细阐述 TR-HPRL 的路由框架、模型构建以及训练流程; 第 4 节通过多种拓扑网络对算法的收敛性与 QoS 性能进行验证; 第 5 节对全文进行总结, 并展望未来研究方向. 为方便阅读, 现将文中使用的主要数学符号在表 1 中进行集中说明.

表 1 主要符号及其含义

Table 1 Main notations and their meanings

符号	含义
G	网络拓扑
$\mathcal{V}$	节点集合
$\mathcal{E}$	有向链路集合
N、M	节点与链路的数量
λ	流的到达速率
φ	节点的服务速率
α	流量分割比
ρ	节点利用率
P	节点丢包率
D	节点/路径延迟
U_{QoS}	QoS 感知效用函数
$\mathcal{S}$	马尔科夫决策过程 (Markov decision process, MDP) 状态空间
$\mathcal{A}$	MDP 动作空间
$\mathcal{R}$	MDP 奖励函数
γ	折扣因子
η	流量类型

1 相关工作

随着网络环境中业务类型和服务质量需求的持续增长, 多路径路由作为应对网络拥塞、提高资源利用率和容错性的关键机制, 已成为路由优化领域

的研究热点[17−21]. KDN 下的多路径路由方法, 目前主要分为传统多路径路由方法、基于深度学习的多路径路由方法以及基于强化学习的多路径路由方法.

1.1　传统多路径路由方法

随着网络规模的不断增大, 单纯依靠 ECMP[7] 进行负载均衡的局限性愈发明显, 亟须更多灵活的多路径优化方案. 为进一步提高多路径路由的灵活性和精度, 研究者提出了多种改进方案. 例如, Gurusamy 等[22] 基于盒覆盖理论提出链路权重动态调整机制, 通过限制最短路径的过度使用缓解拥塞, 提升了静态策略的灵活性. Prabhavat 等[23] 在等价多路径框架中引入 Valiant 负载均衡机制, 扩展了可用路径集, 增强了对复杂网络拓扑结构的适应性, 该方法在多级交换架构和大规模数据中心中应用尤为广泛. 此外, He 等[24] 采用轮询方式进行流量调度, 虽实现简单但缺乏链路状态实时感知, 仍属于静态分配方法, 在动态负载场景下易导致部分链路过载. 为进一步提高资源利用率和调度精度, 部分研究基于链路状态感知机制设计路由策略. Deng 等[25] 提出了一种面向物联网系统的应用感知 QoS 路由算法, 该算法基于模拟退火策略, 能够动态适应网络状态, 从而优化高优先级物联网业务的多重服务质量需求. Lin 等[26] 提出了一种基于网络效用最大化的分布式多路径路由算法, 该算法通过迭代优化链路资源分配, 显著提升了系统的总效用.

尽管基于优化的多路径路由方法在理论上具有较强的可解释性和最优解保障, 但仍面临两个主要问题: 一是这些方法严重依赖于静态拓扑结构和显式参数设定, 难以适应实际网络的动态变化; 二是优化问题的求解复杂度较高, 尤其是在处理混合整数非线性约束时, 计算资源消耗大, 难以满足在线调度的时效性需求. 因此, 近年来的研究逐渐转向数据驱动方法, 结合机器学习技术以构建更具灵活性和实时响应能力的智能路由机制[27].

1.2　基于深度学习的多路径路由方法

随着人工智能技术的迅猛发展, 深度学习因其在特征表示和模式识别方面的优越性能, 逐渐被引入网络路由优化领域. 当前, 大多数基于深度学习的路由方法主要用于网络流量预测、节点状态评估或链路可靠性建模, 且这些方法通常仅作为策略优化的辅助模块. 例如, Bao 等[28] 构建了两层级监督学习模型, 针对无线网状网络节点移动引发的链路失效问题, 结合线性规划方法提升了备用路由选择的鲁棒性. Zou 等[29] 提出了 DeepTSQP 方法, 利用门控循环单元建模用户与服务之间的时间依赖特征, 从而显著提升了 QoS 预测的准确性. Azzouni 等[14] 设计的 NeuTM 框架将长短时记忆网络应用于流量矩阵预测, 实现了优于传统模型的预测性能. 除流量建模外, 部分研究也聚焦于深度学习在链路状态评估与网络安全领域的应用. 文献 [30] 提出了基于卷积神经网络的近实时安全检测系统, 结合博弈论策略提升了拒绝服务攻击的缓解能力, 为安全路由提供了新思路. Rusek 等[13] 基于图神经网络提出 RouteNet 模型, 有效学习了网络结构与性能的映射关系, 实现了对网络行为的高效建模. 尽管上述方法在流量预测、链路质量评估及异常检测等方面取得了良好成效, 但多数聚焦于网络管理的辅助功能, 缺乏对多路径路由决策的直接优化, 且多依赖离线训练, 缺乏对网络状态高频变化的实时响应能力. 因此, 如何实现深度学习与多目标、实时性的路由决策过程的有效融合, 仍是当前研究亟待突破的关键方向.

1.3　基于强化学习的多路径路由方法

随着 SDN 技术的引入, 数据平面与控制平面的分离为路由优化问题的研究提供了新的可能性. SDN 通过对数据平面的实时监控, 为控制平面提供全局网络视图, 而知识平面的引入则进一步增强了网络智能体的部署可行性. 在此架构下, 深度强化学习 (deep reinforcement learning, DRL) 凭借其在解决复杂优化问题方面的显著优势, 成为网络智能控制的重要工具[31]. Chen 等[21] 提出了一种基于 Q-learning 的 SDN 多路径路由方案, 结合实时网络状态和流量特征对不同数据流进行动态路径调整, 在链路资源不足时按优先级重新分配剩余流量, 以实现多路径路由的动态适应. Tang 等[32] 针对域内流量工程场景, 提出了一种基于决斗深度 Q 网络的多路径路由算法 DRL-MPR, 显著降低了最大链路利用率与端到端时延.

为了提升策略学习的适应性与泛化能力, 后续研究引入多种模型增强机制与融合结构. 例如, Altamirano 等[33] 结合 DRL 与生成对抗网络构建了具备自管理能力的 SDN 路由架构. Casas-Velasco 等[34] 提出的 RSIR 方法, 利用 Q-learning 算法以及链路状态信息, 实现了动态路由优化. 此外, 部分研究通过引入 QoS 指标建模与流量预测机制, 提升策略的服务感知能力. Dai 等[16] 提出的 IQoR-LSE 链路状态估计算法, 通过链路拥塞推断来优化动作空间的探索过程, 不仅缓解了复杂策略搜索中的收敛难题, 还显著降低了网络抖动与丢包率. Ye 等[35] 与 Zhang 等[36] 则通过针对关键流进行重路由以缓解链路拥塞, 提出了基于流量体量或其在拥塞链路中的占比

的策略. Rezaei 等[37] 提出的基于多任务学习的流量预测与分类框架, 通过带宽与持续时间预测进行流量分类. 此外, Zhang 等[38] 在 SDN 多路径路由中利用 DRL 进行吞吐时延权衡优化, 体现了面向多目标的奖励设计与路由决策在工程中的可实现性. He 等[39] 利用基于注意力机制的 DRL 算法, 协同优化网络功能虚拟化场景下的虚拟网络功能放置和路由问题, 以此来适应时变的网络状态和多样化的 QoS 需求. 为更充分地获取网络拓扑中的结构信息, 近年来的研究在 DRL 路由优化中引入图神经网络以增强对拓扑特征的感知能力. He 等[40] 针对 KDN 架构, 提出了一种基于消息传递机制的深度强化学习算法 MPDRL, 通过图神经网络显式提取网络拓扑中的链路关联特征, 提升了全局负载均衡性能. Ding 等[41] 提出的 GROM 算法通过离散化动作空间和元素级图神经网络输出, 提升了 DRL 路由模型在不同规模拓扑中的泛化能力, 实现了对未见拓扑的自适应路由. Xu 等[42] 面向 WAN 流量工程, 通过 FlowGNN 与多智能体强化学习建模按需路由决策, 并结合 ADMM 以降低约束违例, 从而在大规模网络中实现近最优分配与较高推理效率. Lin 等[43] 提出的 TITE 将 Transformer 融合于 DRL, 以显式建模连续状态间的隐含依赖, 在混合 SDN 场景下实现对动态业务的快速策略收敛与稳态性能提升.

尽管上述方法在动态路由优化中展现出一定的有效性, 但在面对多类型流量时仍普遍存在局限性. 具体而言, 现有的流量感知强化学习和基于 GCN 的路由方法往往倾向于构建单一策略网络来处理所有流量, 同时仅将流量类型作为状态输入, 未能在策略结构与奖励设计中系统引入类型先验知识, 对 QoS 的感知粒度存在不足. 在处理异构流量的差异化 QoS 需求时, 这种单一共享参数的更新容易出现梯度冲突, 导致网络 QoS 性能的下降. 本文聚焦于智能体 QoS 感知粒度问题, 面向多类型流量的差异化需求, 引入流量分类和差异化奖励函数, 并设计分层策略网络结构, 通过 GCN 提取通用网络拓扑特征, 并结合类型专属策略头进行独立决策, 有效缓解多类型流量 QoS 路由优化过程中的梯度冲突, 实现更加精确的资源分配和性能优化, 从而在复杂的网络环境中保障服务质量.

2 多路径路由模型与 QoS 问题建模

2.1 通信网络模型

本文考虑多路径路由问题, 将网络拓扑表示为一个加权有向图 $G=(\mathcal{V}, \mathcal{E}, \mathcal{W})$, 其中, $\mathcal{V}=\{v_1, v_2, \cdots, v_N\}$ 表示节点集合, $\mathcal{E}=\{e_1, e_2, \cdots, e_M\}$ 表示有向链路的集合, $\mathcal{W}=\{w_{e_1}, w_{e_2}, \cdots, w_{e_M}\}$ 表示每个有向链路 $e_m \in \mathcal{E}$ 的权重集合, $N=|\mathcal{V}|$ 和 $M=|\mathcal{E}|$ 分别代表节点和有向链路的数量. 不失一般性, 本文对图 G 作出以下假设:

1) 任意节点 $v_n \in \mathcal{V}$ 都可以作为源节点或目的节点;

2) G 是简单图, 不包含环和重边, 即没有一条链路同时以同一个节点为源节点和目的节点, 也没有一对节点由多个相同方向的链路连接;

3) G 是强连通图, 即任意两个节点之间至少存在一条可达路径.

对于每个节点 $v_n \in \mathcal{V}$, 本文假设在每个时间步长 t 内最多有一个流量通过它进入网络. 因此, 本文使用 $\mathcal{F}^t=\{f_1^t, f_2^t, \cdots, f_N^t\}$ 表示 t 时刻传输的流的集合, 其中 f_n^t 表示在 t 时刻从节点 v_n 进入的流. 每个流都具有源节点 $v_{n,\,src}^t=v_n$、目标节点 $v_{n,\,snk}^t$、到达速率 λ_n^t 和 QoS 性能 δ_n^t, 其中, 本文考虑的 QoS 性能指标包括丢包率性能 δ_P 和延迟性能 δ_D 两项, 即 $\delta_n^t \in \{\delta_P, \delta_D\}$. 因此, 流 $\mathcal{F}$ 可以被分为丢包敏感型流 $\mathcal{F}_D$、延迟敏感型流 $\mathcal{F}_P$ 和容错型流 $\mathcal{F}_E$. 此外, 定义每个节点 v_n 的服务速率为 φ_{v_n}. 对于每条链路 $e_m \in \mathcal{E}$, 定义链路权重 w_{e_m} 为到达节点的服务速率的倒数. 对于任意两个节点 v_n 和 $v_{n'}$, 记两节点间的可达路径集合 $Pi_{(v_n,\,v_{n'})}=\{pi_{(v_n,\,v_{n'}),\,1}, \cdots, pi_{(v_n,\,v_{n'}),\,k}, \cdots, pi_{(v_n,\,v_{n'}),\,K_{(v_n,\,v_{n'})}}\}$,其中 $K_{(v_n,\,v_{n'})}$ 表示节点 v_n 和 $v_{n'}$ 间所有可达路径数量, $pi_{(v_n,\,v_{n'}),\,k}=\{\mathcal{V}_{(v_n,\,v_{n'}),\,k}, \mathcal{E}_{(v_n,\,v_{n'}),\,k}\}$ 表示节点 v_n 到节点 $v_{n'}$ 的第 k 条可达路径. 定义 $\mathcal{V}_{(v_n,\,v_{n'}),\,k} \in \mathcal{V}$, $\mathcal{E}_{(v_n,\,v_{n'}),\,k} \in \mathcal{E}$, 其中 $\mathcal{V}$ 和 $\mathcal{E}$ 分别表示路径 $pi_{(v_n,\,v_{n'}),\,k}$ 所经过的节点和链路的集合. 本文使用 $c_{(v_n,\,v_{n'}),\,k}$ 表示路径代价, 可根据路径所经过的链路的权重计算, 即

$$c_{(v_n,\,v_{n'}),\,k}=\sum_{e_m \in \mathcal{E}_{(v_n,\,v_{n'}),\,k}} w_{e_m} \tag{1}$$

对于每个节点对 $(v_n, v_{n'})$, 首先根据 OSPF 结合链路权重选取代价最小的 L 条路径作为备选路径集 $\bar{P}i_{(v_n,\,v_{n'})}=\{pi_{(v_n,\,v_{n'}),\,1}, \cdots, pi_{(v_n,\,v_{n'}),\,l}, \cdots, pi_{(v_n,\,v_{n'}),\,L}\}$. 为便于分析, 假设对于 $\forall f_n^t \in \mathcal{F}^t$ 在 t 时刻同时在所有备选路径 $\bar{P}i_{(v_{n,\,src}^t,\,v_{n,\,snk}^t)}$ 上传输数据. 本文所需求解的优化问题是: 为 $\forall f_n^t \in \mathcal{F}^t$ 确定其在各自 L 条备选路径上的分割比. 流 f_n^t 的分割比表示为:

$$A_n^t=[\alpha_{n,\,1}^t, \cdots, \alpha_{n,\,l}^t, \cdots, \alpha_{n,\,L}^t] \tag{2}$$

其中, $\alpha_{n,\,l}^t$ 表示流 f_n^t 在备选路径 $pi_{(v_{n,\,src}^t,\,v_{n,\,snk}^t),\,l}$

上的分割比, 且 $\alpha_{n,\,l}^{t}$ 满足下式:

$$0<\alpha_{n,\,l}^{t}<1,\quad \forall\alpha_{n,\,l}^{t}\in A_{n}^{t} \tag{3a}$$

$$\sum_{l=1}^{L}\alpha_{n,\,l}^{t}=1 \tag{3b}$$

特别地, 考虑节点对 $(v_n,\,v_{n'})$ 所有可达路径的数量小于 L, 此时对于 $\forall l'\in[K_{(v_{n,\,src}^{t},\,v_{n,\,snk}^{t})}+1,\,L]$ 取 $\alpha_{n,\,l'}^{t}=0$. 因此, 每条流 $f_{n}^{t}\in\mathcal{F}^{t}$ 被分到每个备选路径 $pi_{(v_{n,\,src}^{t},\,v_{n,\,snk}^{t}),\,l}$ 上的传输速率为 $\alpha_{n,\,l}^{t}\lambda_{n}^{t}$. 图 1 为多路径路由示意图, 其中蓝、绿、红色的线分别代表从节点 $v_{n,\,src}^{t}$ 到节点 $v_{n,\,snk}^{t}$ 的三条备选路径.

对于每个节点 v_n, 用 $\lambda_{v_n}^{t}$ 和 φ_{v_n} 分别表示在 t 时刻流的总到达率以及服务速率. 节点利用率可以表示为 $\rho_{v_n}^{t}=\lambda_{v_n}^{t}/\varphi_{v_n}$, 其中, $\rho_{v_n}^{t}<1$. 根据以上定义, 可以计算出 t 时刻节点 v_n 的平均序列长度如下:

$$Q_{v_n}^{t}=\begin{cases}\dfrac{\rho_{v_n}^{t}}{1-\rho_{v_n}^{t}}-\dfrac{(c_{v_n}+1)(\rho_{v_n}^{t})^{c_{v_n}+1}}{1-(\rho_{v_n}^{t})^{c_{v_n}+1}}, & \rho_{v_n}^{t}<1\\ \dfrac{c_{v_n}}{2}, & \rho_{v_n}^{t}=1\end{cases} \tag{4}$$

其中, c_{v_n} 表示节点 v_n 的容量. 同样地, 在节点 v_n 处排队序列的长度大于 c_{v_n} 的概率为:

$$P_{v_n}^{t}=\frac{(1-\rho_{v_n}^{t})(\rho_{v_n}^{t})^{c_{v_n}}}{1-(\rho_{v_n}^{t})^{c_{v_n}+1}} \tag{5}$$

另一方面, 流在节点 v_n 处 t 时刻的延迟可以表示为:

$$D_{v_n}^{t}=\frac{Q_{v_n}^{t}}{\lambda_{v_n}^{t}(1-P_{v_n}^{t})} \tag{6}$$

由于传播延迟与排队延迟相比可以忽略不计, 本文考虑流在路径 $pi_{(v_{n,\,src}^{t},\,v_{n,\,snk}^{t}),\,l}$ 上的延迟为在路径所经过节点处排队延时累积和, 即

$$D_{pi_{(v_{n,\,src}^{t},\,v_{n,\,snk}^{t}),\,l}}^{t}=\sum_{v_n\in\mathcal{V}_{(v_{n,\,src}^{t},\,v_{n,\,snk}^{t}),\,l}}D_{v_n}^{t} \tag{7}$$

随后, 通过计算所有路径的端到端延迟的加权和, 计算出 f_{n}^{t} 在 t 时刻的端到端延迟为:

$$D_{n}^{t}=\sum_{\substack{pi_{(v_{n,\,src}^{t},\,v_{n,\,snk}^{t}),\,l}\in\\ \bar{P}i_{(v_{n,\,src}^{t},\,v_{n,\,snk}^{t})}}}\alpha_{n,\,l}^{t}D_{pi_{(v_{n,\,src}^{t},v_{n,\,snk}^{t}),\,l}}^{t} \tag{8}$$

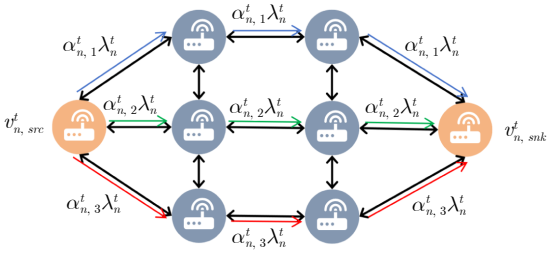


图 1　多路径路由模型

Fig. 1　Multi-path routing model

2.2　QoS 问题建模

考虑 QoS 感知的多路径路由问题, 通过强化学习动态优化所有进入通信网络流的分割比, 实现 QoS 性能提升. 本文使用 $A^{t}=\{A_{1}^{t},\cdots,A_{n}^{t},\cdots,A_{N}^{t}\}$ 表示在 t 时刻所有流 $\mathcal{F}^{t}$ 的分割比的集合, U_{QoS}^{t} 表示 QoS 感知效用函数. 定义优化目标为:

$$\max_{A^{t}}\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}U_{\mathrm{QoS}}^{t} \tag{9}$$

其中, T 表示多路径路由的任务区间; 感知效用函数 U_{QoS}^{t} 综合考虑了整个网络在 t 时刻的丢包率和端到端延迟, 表示为:

$$\begin{aligned}U_{\mathrm{QoS}}^{t}=&\sum_{v_n\in\mathcal{V},\,v_{n,\,src}^{t}=v_n}\beta_{P,\,\delta_n}\times\frac{\lambda_{v_n}^{t}}{\sum\limits_{v_{n'}\in\mathcal{V}}\lambda_{v_{n'}}^{t}}\times\\ &(1-P_{v_n}^{t})+\sum_{f_n^{t}\in\mathcal{F}^{t}}\beta_{D,\,\delta_n}\times\\ &\frac{\lambda_{n}^{t}}{\sum_{f_{n'}\in\mathcal{F}^{t}}\lambda_{n'}^{t}}\times\left(1-\frac{D_{n}^{t}}{\hat{D}^{t}}\right)\end{aligned} \tag{10}$$

其中, $\hat{D}^{t}$ 表示在 t 时刻任意节点 $v_n\in\mathcal{V}$ 处的最大排队延迟. 函数 U_{QoS}^{t} 分别对丢包率 $P_{v_n}^{t}$ 和端到端延迟 D_{n}^{t} 进行归一化处理, 并将优化目标转化为最大化效用函数; 然后, 根据流的到达速率 λ_{n}^{t} 和 $\lambda_{v_n}^{t}$ 计算各项权值; 最后, 根据流量类型不同, 调整丢包缩放因子 $\beta_{P,\,\delta_n}$ 和端到端延迟缩放因子 $\beta_{D,\,\delta_n}$ 大小以适应不同流量的 QoS 需求.

3　基于分层策略强化学习的 QoS 感知路由

本节围绕多类型流量的差异化 QoS 需求, 构建基于 TR-HPRL 的智能路由系统. 首先, 基于 KDN 架构构建融合流量分类与强化学习策略优化的闭环路由框架; 其次, 详细介绍 TR-HPRL 模型, 包括流量分类模块与分层策略路由模块; 最后, 明确算法的训练流程, 包括状态特征提取、动作生成、经验存储与策略更新等环节, 确保算法通过持续学习实现对不同类型流量的差异化路由策略.

3.1　基于 KDN 的多任务流路由框架

现代网络环境中动态流量管理与路径优化是提

升网络 QoS 的关键挑战, 本节基于 KDN 架构构建面向多类型流量管理的智能路由系统. 所提出的框架如图 2 所示, 融合流量分类与强化学习策略优化, 构成集感知、分析与控制于一体的闭环路由体系. 该智能路由系统围绕知识平面部署 TR-HPRL 智能体, 实现网络状态的感知、策略的生成与执行, 满足多任务流的差异化 QoS 需求. 其内部由数据平面、控制平面、管理与信息处理平面以及知识平面四个功能平面协同构成, 各平面功能如下:

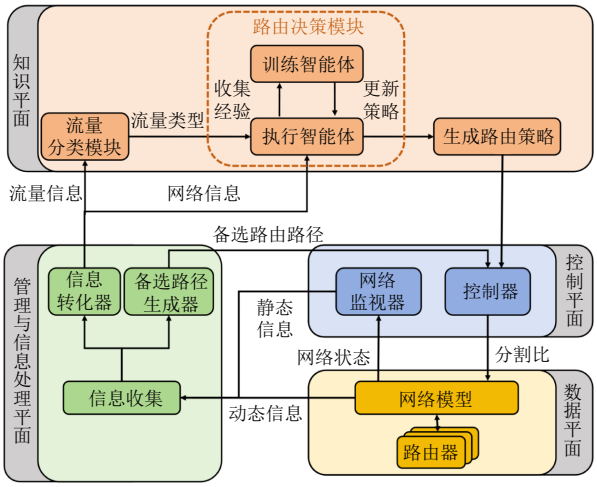


图 2 智能路由系统框架

Fig. 2 Intelligent routing system framework

1) 数据平面: 执行来自控制平面的路由策略, 支持多路径流量分配.

2) 控制平面: 基于知识平面生成的分割策略, 动态更新转发规则, 并监控策略执行效果. 数据与控制平面二者通过标准南北向接口, 如 OpenFlow[44] 与 RESTful[45], 完成指令下发与状态同步.

3) 管理与信息处理平面: 持续收集网络遥测信息, 包括流量负载、节点状态与控制反馈等, 并进行聚合与预处理, 经处理后的结构化状态信息作为知识平面的输入, 用于支持流量分类与策略学习[9].

4) 知识平面: 作为智能决策的核心, 部署 TR-HPRL 智能体, 负责完成两项核心任务: 一是流量分类模块, 基于深度模型识别业务流类型, 生成类别标签作为策略先验信息; 二是路由决策模块, 基于深度强化学习方法, 输入全局网络状态与任务类型, 输出最优路径分配策略. 分类与决策模块联合工作, 使系统能够根据实时网络环境与流量属性, 动态优化多路径路由策略, 实现多任务流在复杂网络中的差异化调度.

3.2 TR-HPRL 模型

在本文提出的基于 KDN 的多任务流量路由框架中, TR-HPRL 智能体部署于知识平面, 负责执行流量分类和路由决策两大核心任务. 该智能体基于分层策略强化学习方法优化流量在多路径环境下的动态分配策略, 旨在提升网络整体 QoS 性能. TR-HPRL 智能体的整体架构如图 3 所示, 主要包括流量分类模块和路由决策模块. 其中, 流量分类模块负责对不同类型的业务流进行实时识别, 并为路由决策提供先验信息; 路由决策模块则基于强化学习方法, 通过全局状态感知和策略优化, 实现自适应的流量分配. 两个模块协同工作, 使得智能体能够根据网络环境的变化实时调整路由策略, 以满足不同流量类型的 QoS 需求. 下文将对流量分类模块和路由决策模块展开详细介绍.

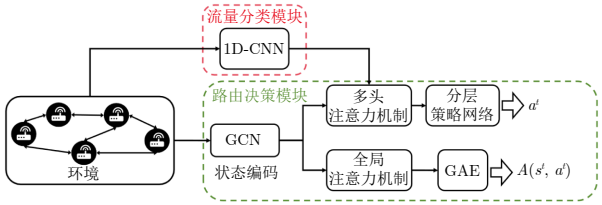


图 3 TR-HPRL 模型

Fig. 3 TR-HPRL model

3.2.1 基于 1D-CNN 的流量分类模型

本文使用的 1D-CNN 模型由输入层、卷积层、池化层和全连接层组成, 其结构如图 4 所示.

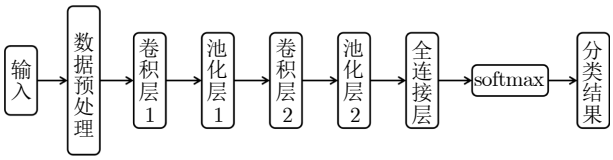


图 4 1D-CNN 分类模型

Fig. 4 1D-CNN classification model

流量分类模块作为 TR-HPRL 的感知前端, 其核心任务是通过细粒度流量特征解析, 实现多类别业务流的精准识别.

3.2.2 TR-HPRL 模型构建

针对 KDN 多路径路由动态决策问题, 本文将其建模为 MDP, 定义五元组 $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma)$, 其中, $s^t \in \mathcal{S}$ 表示 t 时刻的全局状态, $\mathcal{A}$ 表示动作空间, 状态转移概率矩阵表示为 $\mathcal{P}: \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$, $\mathcal{R}$ 为奖励函数, r^t 用于表示智能体在 t 时刻得到的奖励值, $r^t = \mathcal{R}(s^t, a^t, s^{t+1})$, 其中 a^t 为 t 时刻的路由动作. 智能体的奖励函数如式 (10) 所示.

1) 状态空间 $\mathcal{S}$: 状态空间包含流量信息和网络拓扑信息, 即 $s^t = \{A_{\text{adj}}, \rho^t, U_{\text{TM}}^t\}$, 其中 A_{adj} 表示

图 G 的邻接矩阵, $\rho^t_{v_n} \in \rho^t$ 表示节点 v_n 在 t 时刻的利用率, U^t_{TM} 为流量需求矩阵, 包括 t 时刻流量 f^t_n 的源节点 $v_{n,\,src}$、目标节点 $v_{n,\,snk}$、到达速率 λ_n、流量类型 η_n 和从节点 $v_{n,\,src}$ 到节点 $v_{n,\,snk}$ 间的备选路径长度 $|\mathcal{E}_{v_{n,\,src},\,v_{n,\,snk}}|$.

2) 动作空间 $\mathcal{A}$: 动作空间对应第 2.2 节中所述优化问题的路径分割比, 定义为 $\mathcal{A} = [A_1, \cdots, A_n, \cdots, A_N]$, 其中, 对于 t 时刻流 f^t_n 的流量分割比 A^t_n 的定义如式 (2) 所示.

3) 奖励函数 $\mathcal{R}$: 奖励函数基于式 (9) 定义的效用函数, 根据流量类型不同, 调整影响因子大小, 实现差异化 QoS 优化.

3.2.3　TR-HPRL 设计

在本文提出的 TR-HPRL 中, 采用深度学习模型对策略函数 π_θ 进行近似. 为实现从状态输入到动作输出的映射, 若利用单一深度神经网络构建统一策略函数, 往往难以兼顾多类型流量差异巨大的最优策略参数, 导致参数更新冲突与收敛困难; 为每种流量单独构建策略网络则会造成参数规模庞大且泛化能力受限. 因此, 本文以硬参数共享式多任务学习架构为核心设计 TR-HPRL, 使用分层策略网络的形式分别对不同类型流量输出路由决策. 不同于传统分层强化学习侧重管理者到子策略的时序抽象与子目标分解, 我们的分层以共享 GCN 编码和类型专属策略头为核心, 并显式融合流量类型先验与差异化奖励, 在同一时间尺度上并行完成多类型路由决策. 本节将系统阐述 TR-HPRL 的模型结构.

1) 基于 GCN 的编码器: 针对网络拓扑的图结构特性, 本文采用图卷积网络作为共享状态编码器, 以直接捕获节点间的拓扑依赖关系并聚合全局信息. 通过同时服务于多种流量类型的路由任务, 该共享层能学习到更鲁棒、更具泛化能力的底层网络状态特征表示. GCN 的一层卷积操作可以表示为:

$$H^{(l+1)} = \sigma\left(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}\right) \tag{11}$$

其中, $H^{(l)}$ 为第 l 层的特征矩阵; $W^{(l)}$ 为第 l 层的可学习权重矩阵; $\sigma(\cdot)$ 为激活函数; $\tilde{A} = A_{\mathrm{adj}} + I$ 为图 G 的邻接矩阵; $\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}$ 实现对邻接矩阵的归一化处理. 通过基于 GCN 的编码器, 输出具有全局信息的状态编码 h_{embed}, 作为 actor-critic 决策网络的状态输入.

2) actor 网络: 为解决使用一组完全共享的策略参数所导致参数更新冲突的问题, 设计了一种基于多头注意力机制的分层 actor 网络架构. 该架构核心思想是通过共享编码层提取通用网络状态特征, 并在高层使用多个类型特异性的策略头作为任务专属层, 以分别优化各类型流量的路由策略. 这种结构既避免了单一模型收敛困难和多模型参数冗余的缺陷, 又通过参数共享提升了学习效率和模型的泛化能力, 同时利用独立的策略头满足了各类型流量的差异化需求, 从而在复杂的网络环境中实现高效且稳定的决策, 提高 QoS 性能. 其具体网络架构如图 5 所示. 图中使用不同颜色的流特征编码代表不同类型流, 并利用掩码策略, 使每个多头注意力网络只关注一种类型的流量的特征. 每种类型的流量经过多头注意力机制计算后输出的流量特征进入独立的决策层, 生成该流量类型对应的策略 $\pi^\eta_\theta(a^t|s^t)$, 其中 η 表示流量类型. 各类型流量的策略整合, 构成最终 actor 网络的输出 a^t.

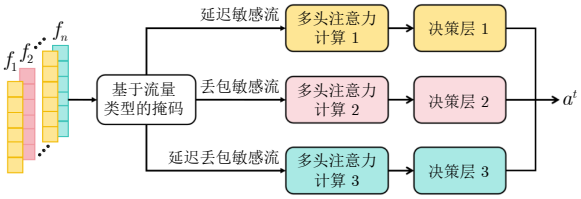


图 5　基于流量类型的多头注意力决策网络

Fig. 5　Multi-head attention decision-making network based on traffic type

3) critic 网络: critic 网络的核心功能是评估当前网络状态的价值函数, 为 actor 网络的策略优化提供策略梯度. 针对多路径路由场景中复杂的状态依赖关系, 设计了一种基于全局注意力机制的 critic 网络架构. 首先, 将 GCN 编码的网络状态 h_{embed} 输入全局注意力机制模块, 以建模不同网络节点间的全局依赖关系, 从而生成具备全局感知能力的状态表示 h_{global}. 接着, 将 h_{global} 传入多层感知机 (multilayer perceptron, MLP), 以回归当前状态的价值函数 $V(s^t)$. 此外, 为了更稳定地估计动作–状态对 $(s^t,\, a^t)$ 的优势值, critic 网络采用广义优势估计 (generalized advantage estimation, GAE) 方法计算 $A(s^t,\, a^t)$, 并将其作为策略梯度优化的关键指标, 提升策略学习的效率与稳定性.

总体而言, TR-HPRL 模型通过整合流量分类和分层策略强化学习路由决策, 不仅能对网络状态进行精确感知, 还能自适应地调整流量分割比, 从而在多路径路由环境下实现 QoS 性能的最优化.

3.3　算法训练流程

TR-HPRL 的训练过程通过环境交互与策略更新交替进行, 具体内容如算法 1 所示. 环境交互与数据采样阶段中, 在每个时间步 t, 智能体利用共享 GCN 编码器将当前状态 s^t 映射为特征向量 h^t.

actor 网络根据流量类型 η, 将 h^t 输入对应的独立注意力策略头, 采样生成差异化动作 $a_\eta^t \sim \pi_\theta^\eta(h^t)$, 并组合为全局路由动作下发执行. 智能体执行该动作后, 环境返回即时奖励 r^t 及下一个状态 s^{t+1}, 系统将获得的经验元组 $(s^t,\ a^t,\ r^t,\ s^{t+1},\ \log\pi_\theta^\eta(a_\eta^t \mid h^t),\ V_\omega(s^t))$ 存入经验回放池 $\mathcal{B}$, 其中 $V_\omega(s^t)$ 为状态价值估计, 直至达到预设容量, 以用于后续的策略优化. 在策略更新阶段, 首先利用 GAE 算法估计优势函数 $\hat{A}_\eta^t$ 与目标价值 $\hat{V}^t$, 以平衡方差与偏差. 随后进行 K_{update} 次迭代:

1) critic 网络更新: 输入状态嵌入 h_{embed}^t, 经过全局注意力模块提取全局状态感知表示 h_{global}, 最终通过 MLP 输出状态价值估计 $V_\omega(s^b)$. 其目标是最小化预测值 $V_\omega(s^b)$ 与目标值 $\hat{V}_b$ 间的均方误差, 表示为:

$$\ell(\omega) = \frac{1}{B}\sum_{b=1}^{B}\left(\hat{V}_b - V_\omega(s^b)\right)^2 \tag{12}$$

其中, B 代表从经验回放池中采样的小批量样本数量; b 代表小批量中的第 b 个样本.

2) actor 网络更新: 每类流量 η 对应一个策略头 π_θ^η, 分别计算第 η 类流量在新策略参数 θ 和旧策略参数 θ_{old} 下的概率之比 $r_b^\eta(\theta)$, 用于衡量策略更新的幅度.

$$r_b^\eta(\theta) = \frac{\pi_\theta^\eta(a_b^\eta|s^b)}{\pi_{\theta_{\text{old}}}^\eta(a_b^\eta|s^b)} \tag{13}$$

其中, $\pi_\theta^\eta(a_b^\eta|s^b)$ 和 $\pi_{\theta_{\text{old}}}^\eta(a_b^\eta|s^b)$ 分别表示在新策略参数 θ 和旧策略参数 θ_{old} 下, 第 η 类流量在状态 s^b 时采取动作 a_b^η 的概率. 为提升策略更新的稳定性、避免参数剧烈波动, 采用近端策略优化 (proximal policy optimization, PPO) 算法, 通过引入剪切目标函数 $\ell_\eta^{clip}(\theta)$ 限制策略更新幅度:

$$\ell_\eta^{clip}(\theta) = \frac{1}{B}\sum_{b=1}^{B}\min\Big(r_b^\eta(\theta)\hat{A}_b^\eta,\ \text{clip}(r_b^\eta(\theta),\ 1-\epsilon,\ 1+\epsilon)\hat{A}_b^\eta\Big) \tag{14}$$

其中, ϵ 为剪切系数; $\text{clip}(r_b^\eta(\theta),\ 1-\epsilon,\ 1+\epsilon)$ 对概率比率进行剪切操作, 将 $r_b^\eta(\theta)$ 限制在 $[1-\epsilon, 1+\epsilon]$ 区间内. 每个策略子头独立进行优化, 以避免不同类型流量之间的梯度干扰.

整个 TR-HPRL 的训练过程中, GCN 提取结构化网络状态, actor 分层策略网络执行各类流量的动作生成, critic 评估全局状态价值, 通过交替的经验采样与策略更新, 使得模型逐步收敛至能适应不同流量类型的最优路由策略.

算法 1. TR-HPRL

输入. actor 参数 θ, critic 参数 ω, 折扣因子 γ, PPO 剪切系数 ϵ, 交互步数 T, 总训练轮次 M_e, 更新次数 K_{update}.

输出. 优化后的 θ 和 ω.

初始化状态 s^0;
For episode = 1 to M_e **do**
  初始化状态 s^0;
  For $t = 0$ to $T-1$ **do**
    特征提取: $h_{\text{embed}}^t = \text{GCN}(s^t)$;
    For 每种流量类型 η **do**
      通过多头注意力机制进行注意力编码;
      采样动作: $a_\eta^t \sim \pi_\theta^\eta(h^t)$;
    End For
    执行动作 a^t, 获得奖励 r^t 和新状态 s^{t+1};
    存储 $(s^t,\ a^t,\ r^t,\ s^{t+1},\ \log\pi_\theta^\eta(a_\eta^t|h^t),\ V_\omega(s^t))$ 至经验池 $\mathcal{B}$;
  End For
  使用 GAE 估算 $\hat{A}_\eta^t$ 与 $\hat{V}^t$;
  For $k = 1$ to K_{update} **do**
    从 $\mathcal{B}$ 中采样小批量样本;
    更新 critic: 最小化损失 $\ell(\omega)$;
    更新 actor: 对每类流量分别最大化 $\ell_\eta^{clip}(\theta)$;
  End For
End For

3.4 时空复杂度分析

TR-HPRL 的计算开销主要来源于其核心的深度学习模块, 包括 GCN、actor 网络以及 critic 网络. 本节将分别对算法的时间复杂度和空间复杂度进行分析.

算法的时间复杂度分为在线推理和离线训练两个阶段. 推理阶段根据当前网络状态 s^t 生成路由动作 a^t, 主要涉及一次前向传播, 其时间复杂度由 GCN 编码器和 actor 网络共同决定. 对于一个包含 N 个节点和 $|\mathcal{E}|$ 条链路的网络, GCN 编码器的时间复杂度约为 $\text{O}(|\mathcal{E}| \times F_{in} \times F_{out})$, 其中 F_{in} 和 F_{out} 分别为输入和输出特征维度. actor 网络包含 C 个流量类别的独立注意力模块, 其时间复杂度为 $\text{O}(C \times N^2 \times F_{out})$. 因此, 单次路由决策的总体时间复杂度由主导项决定, 为 $\text{O}(C \times N^2 \times F_{out})$. critic 网络在训练阶段接收 GCN 的输出, 并通过全局注意力机制进行处理, 其时间复杂度与 actor 网络类似, 为 $\text{O}(N^2 \times F_{out})$. 离线训练阶段涉及对经验回放池中的样本进行多次前向和反向传播, 其总体时间复杂度约为 $\text{O}(M_e \times K_{\text{update}} \times B \times (Z_{GCN} + Z_{actor} + Z_{critic}))$, 其中 M_e 为总训练轮数, K_{update} 为更新次数, B 为批

次大小, Z_* 代表各模块单次更新的时间复杂度.

算法的空间复杂度主要由存储模型参数所需空间和经验回放池所需空间两部分构成. 模型参数量包括 GCN、actor 网络和 critic 网络的参数量, 我们将其表示为 $O(|\theta|)$, 其中 $|\theta|$ 是所有网络参数的总量. 经验回放池用于存储历史交互数据 $(s^t, a^t, r^t, s^{t+1}, \log \pi_\theta^\eta(a_\eta^t|h^t), V_\omega(s^t))$. 其空间开销与经验池容量 B_{size} 以及每个样本的大小成正比. 状态 s 的表示是样本中占用空间最大的部分, 其大小与网络节点数 N 相关. 因此, 经验回放池的空间复杂度为 $O(B_{size} \times S_{exp})$, 其中 S_{exp} 是单个经验元组的存储大小. 综上所述, TR-HPRL 的整体空间复杂度为 $O(|\theta| + B_{size} \times S_{exp})$.

由于大多数现实网络具有稀疏连接的特性, 且 GCN 在处理稀疏图结构时具有效率优势, 同时算法训练过程可以离线执行, 而在线推理所需的资源可控, 因此该算法在真实网络部署中具有一定的可行性基础.

4　实验验证

本节分别对分类模型和 TR-HPRL 的有效性进行验证. 本文设计的 TR-HPRL 在 PyTorch 2.2.2 框架下实现, 使用 NVIDIA GeForce RTX 4090 GPU 进行训练.

4.1　实验设置

4.1.1　分类模型设置

采用加拿大网络安全研究所提供的 ISCXVPN2016 数据集[46] 作为实验数据. 该数据集基于真实网络环境进行流量采集, 覆盖多个应用类别, 包含 80 多种流量特征, 是目前广泛用于流量分类和网络安全研究的高质量数据集. 为了适应 QoS 感知的分类需求, 本实验对 ISCXVPN2016 数据集进行了数据筛选、特征工程及类别重定义. 具体应用流量划分如表 2 所示, 其中, 延迟敏感型主要指对端到端延迟有严格上限要求的业务, 当延迟超过阈值时, 业务体验会显著劣化甚至不可用, 其核心 QoS 指标为延迟; 丢包敏感型主要指对数据完整性要求极高的业务, 丢包率超过阈值会导致数据完整性受损、重传开销激增, 进而影响业务连续性, 其核心 QoS 指标为丢包率; 容错型指对延迟、丢包率的敏感度较低, 允许一定程度的性能波动, 且波动不会显著影响业务功能的流量. 在本文的建模中, 我们设置一个数据流仅属于单一类别, 其分类依据是基于其应用类型的固有业务特征, 以此避免类别交叉带来的决策模糊性.

表 2　服务类型分类
Table 2　Service type classification

类别	类型应用	主要特征
延迟敏感型	网络电话、在线聊天和音频流	对实时性要求较高
丢包敏感型	视频流、点对点传输和文件传输	需要较高带宽传输
容错型	电子邮件和网页浏览	对 QoS 要求不明显或介于上述两者之间

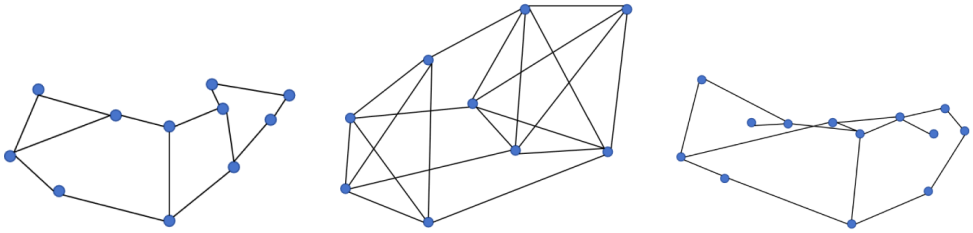
基于 1D-CNN 的深度学习分类模型训练过程中的参数设置包括: 训练周期为 100 轮, 批次大小设置为 32, 学习率设置为 0.000 5, 权重衰减系数设置为 1e−4. 为提高模型训练过程中的优化效果和稳定性, 实验采用了一系列优化策略, 包括优化器选择、学习率调度、梯度裁剪和早停策略, 以确保模型能够更好地适应不同类别流量的特征. 其中, 优化器采用 AdamW (adaptive moment estimation with weight decay), 相较于标准 Adam 优化器, AdamW 通过权重衰减机制, 能够有效抑制过拟合. 进一步, 采用余弦退火学习率调度策略, 初始学习率设置为 0.000 5, 初始周期设置为 10, 倍增系数设置为 2, 以确保学习率在训练过程中能够动态适应模型的优化需求. 最后, 为防止出现梯度爆炸, 设置梯度裁剪阈值为 1, 以确保梯度更新的稳定性.

4.1.2　TR-HPRL 实验设置

在分类模型确定流量类型后, 系统依据不同 QoS 需求采用差异化路由策略. 为验证 TR-HPRL 在动态网络环境下的有效性, 实验选择三种经典拓扑图 Abilene、GridNet 和 NSFNET, 并通过 NetworkX[47] 构建网络. 三种网络的拓扑结构如图 6 所示, 其中, Abilene 包含 11 个节点、28 个有向链路, GridNet 包含 9 个节点、40 个有向链路, NSFNET 包含 13 个节点、30 个有向链路.

为保证任务的复杂性, 同时充分体现 TR-HPRL 的优越性, 根据初步实验我们对实验参数进行以下设置: 在每个时间步长内, 流 $f_n^t \in \mathcal{F}^t$ 的源节点 v_n 从 $\mathcal{V}$ 中随机选取除自身节点外的节点作为目标节点 $v_{n,\,src}^t$. 所有流在 t 时刻的到达率 λ_n^t 服从 Zipf 分布, 且到达速率 $\lambda_n^t \in \{400, 500, 600\}$ 数据包/秒. 数据包大小统一设置为 100 B. 每个流的类型以相同的概率从延迟敏感型、丢包敏感型、容错型三类中选取. 所有节点的服务速率 φ_{v_n} 在 [1 000, 3 000] 数据包/秒之间均匀分布. 候选路径数量设置为 $L = 3$. 每个节点 $v_n \in \mathcal{V}$ 的系统容量设置为 $c_{v_n} = 10\,000$ 数据包.

丢包率和端到端延迟的缩放因子默认值的设置



(a) Abilene (b) GridNet (c) NSFNET

图 6 网络拓扑图

Fig. 6 Network topology diagram

如表 3 所示, 折扣因子设置为 $\gamma = 0.99$, actor 和 critic 网络的学习率分别设置为 $\alpha_\theta = 5 \times 10^{-5}$ 和 $\alpha_\omega = 2 \times 10^{-5}$. 训练阶段, episode 设置为 20 000; 每个 episode 的探索步长为 50 步; 训练批次设置为 64; 经验池大小为 20 000; 近端策略优化模型中剪切幅度设置为 0.2, 控制新旧策略分布之间的相对变化; 熵系数设置为 0.01, 用于衡量策略的不确定性, 促进策略保持多样性; 更新循环次数设置为 10.

表 3 缩放因子默认值

Table 3 Default values of scaling factors

类别	$\beta_{P,\,\delta_n}$	$\beta_{D,\,\delta_n}$
延迟敏感型	0.2	0.8
丢包敏感型	0.8	0.2
容错型	0.5	0.5

GCN 的输入维度为状态维度, 隐藏层维度为 128, 输出通道维度为 256, 网络层数为 2, critic 网络中的全局注意力层注意力头数为 4.

实验将利用传统方法 ECMP、深度强化学习方法 DRL-TE[15] 及 TITE[43] 算法与本文所提 TR-HPRL 进行性能比较评估, 评估指标为丢包奖励 (r_P) 和延迟奖励 (r_D). 其中, DRL-TE 方法通过引入优先级经验回放, 训练基于神经网络的深度强化学习智能体, 以优化深度确定策略梯度 (deep deterministic policy gradient, DDPG) 中的探索机制, 从而实现深度强化学习目标. TITE 算法将 Transformer 模型与深度强化学习相结合, 通过对连续网络状态间的时间依赖关系进行建模, 学习连续状态序列与路由策略间的映射关系, 以更好地适应动态变化的网络流量. ECMP 方法则在具有相等成本的最短路径上均匀分配流量.

4.2 分类模型性能分析

基于 1D-CNN 的分类模型在流量分类上的混淆矩阵如图 7 所示, 矩阵中的数值代表各类别流量的分类情况. 具体而言, 模型对延迟敏感型、丢包敏感型、容错型流量的分类准确率分别达到了 96.9%、97.9% 和 100.0%, 表明其对不同 QoS 需求的流量类别展现出了准确的判别能力.

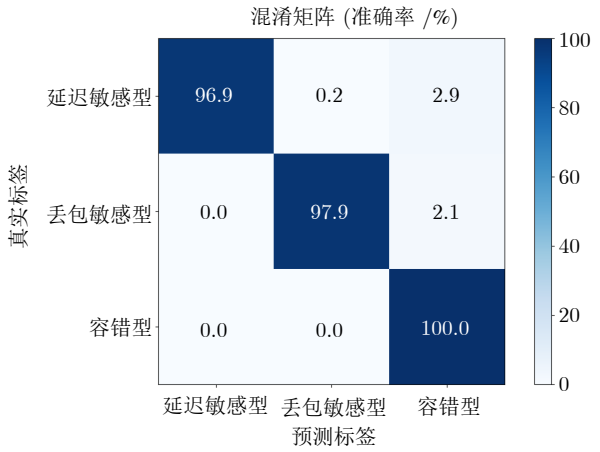


图 7 分类结果混淆矩阵

Fig. 7 Classification result confusion matrix

模型在 ISCXVPN2016 数据集上对三种流量分类的准确率 (Pr) 和召回率 (Rc) 结果如表 4 所示, 其中模型整体分类的准确率和召回率均达到 95% 以上. 综上所述, 该模型在流量分类任务中能够较好地捕捉各类别流量的时序和特征信息, 具有良好的性能, 为后续融合流量分类信息的差异化路由决策提供了可靠的数据基础.

此外, 得益于 1D-CNN 架构的轻量化特性, 模型在推理阶段实际运行时的平均推理延迟维持在毫秒级. 在智能路由系统的整体时延构成中, 相较于通常为秒级的路由策略更新周期[10], 由流量分类引入的计算开销占比极低, 表明分类模块能够满足动态网络环境下实时路由决策的实时性要求.

4.3 路由算法性能分析

4.3.1 收敛性能与计算开销分析

在 GridNet 拓扑图上对 TR-HPRL 的收敛性能进行验证. 其中, 图 8 ~ 图 10 分别展示了训练过

程中的 actor 损失、critic 损失和奖励值指标.

如图 8 所示, 训练初期, 由于策略初始化与探索机制的影响, 损失值偏低且波动剧烈; 随着训练深入, 分层决策网络逐渐学习到针对各流量类型的最优策略, actor 损失函数的值呈现上升趋势; 训练后期, actor 损失稳定在 0.92 附近, 表明策略实现收敛. 另一方面, 采用 PPO 方法对策略更新施加限制, 保证更新过程中不出现过大偏离, 在图 8 中表现为 actor 损失曲线在中后期出现平稳振荡的现象. 曲线在稳定区域附近波动, 反映出网络在微调状态下, 各类型决策网络均已达到局部最优.

表 4　流量分类结果 (%)

Table 4　Results of traffic classification (%)

类别	Pr	Rc
延迟敏感型	100	97
丢包敏感型	99	98
容错型	95	100

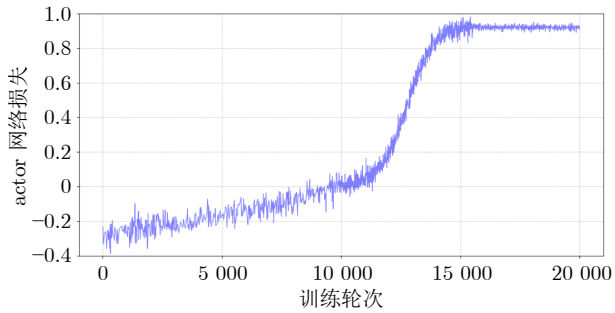


图 8　actor 网络损失收敛曲线

Fig. 8　actor network loss convergence curve

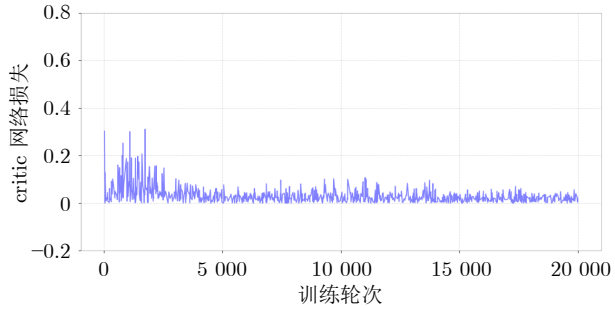


图 9　critic 网络损失收敛曲线

Fig. 9　critic network loss convergence curve

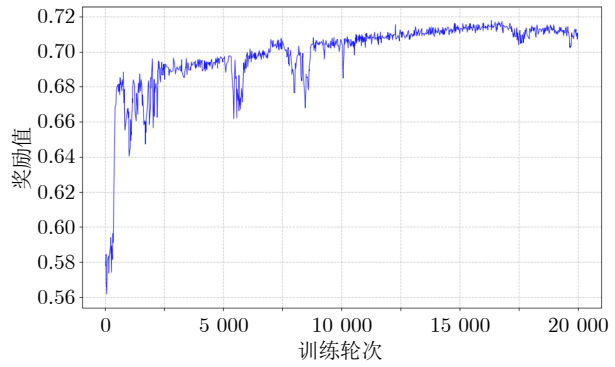


图 10　奖励值收敛曲线

Fig. 10　Reward value convergence curve

分析图 9 可以看出, 训练初期由于路由环境中的流量分布和网络拓扑负载等具有随机性, critic 损失曲线有一定的波动, 随着迭代进行, critic 网络在融合历史动作、奖励和全局注意力信息后, 逐渐纠正时序差分误差并稳定收敛至接近 0 的非负值, 表明基于全局注意力机制的 critic 网络能够精确估计状态价值并进行准确的策略评估. 另一方面, 使用 GAE 计算优势函数, 既平衡了偏差与方差, 也为策略更新提供了较为平滑的梯度信号. 这一机制有助于 critic 损失的稳定降低, 并使其收敛曲线相较于 actor 更为平滑, 震荡幅度更小.

图 10 所示奖励曲线表明, 随着训练轮次从 0 增加到 20 000, 总奖励值呈显著收敛趋势. 初始阶段因智能体对环境状态空间和策略空间的认知尚未充分建立, 仍处于探索阶段, 奖励值在 0.64 至 0.68 区间波动; 训练步数超过 4 000 步后, 奖励值稳步上升并在训练中期达到 0.70 以上的稳定水平; 后期始终维持在 0.71±0.008 的窄幅区间内波动, 未出现明显下降或剧烈震荡. 验证了算法良好的收敛稳定性, 表明算法能通过持续学习逐步优化策略, 最终获得稳定且高性能的解决方案.

综上所述, actor 损失、critic 损失以及奖励值收敛曲线充分证明了本文提出的分层策略强化学习路由算法在多流量类型环境下的有效性. 分层决策、共享特征提取以及全局注意力与 GAE 机制的联合应用, 均对策略优化过程产生了积极影响, 使得系统在解决延迟与丢包权衡问题上表现出较高的稳定性和收敛性.

为了评估算法在实际部署中的资源开销, 我们在 NVIDIA GeForce RTX 4090 平台上测试了模型的在线推理性能. 结果显示, TR-HPRL 生成单次全局路由策略的平均推理时延约为 23 ms, 显存占用约为 360 MB. 相较于通常为秒级的路由更新周期以及现有网络设备的硬件规格, 该算法的计算与存储开销均处于较低水平, 验证了其在真实网络部署中具有一定的可行性基础.

4.3.2　缩放因子影响分析

在本文的奖励函数设计中, 丢包缩放因子 β_{P,δ_n}

和延迟缩放因子 β_{D,δ_n} 是实现差异化服务质量保障的关键参数, 直接影响智能体对不同 QoS 目标的优化侧重. 为探究这些参数设置对模型性能的影响, 本节为延迟敏感型流量设置了三组不同的 β_{D,δ_n}, 具体为 0.9、0.8、0.7, 并相应调整 β_{P,δ_n}, 丢包敏感型流量参数设置与延迟敏感型相反.

实验结果如图 11 所示. 从图中可以看出, 当延迟敏感型流量的延迟缩放因子 β_{D,δ_n} 设置为 0.8 时, 模型在丢包奖励 r_P、延迟奖励 r_D 以及综合平均奖励 r_A 上均取得了最优表现. 若进一步增大 β_{D,δ_n} 至 0.9, 或将其降低至 0.7, 均会导致整体奖励值出现不同程度的下降. 这表明, 本文选取的参数能够在延迟和丢包两个 QoS 关键指标间实现有效平衡, 最大化整体路由性能.

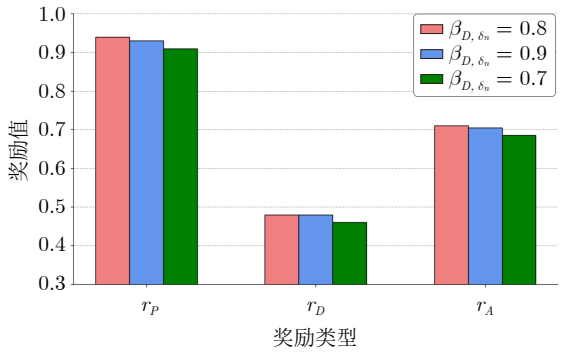


图 11 不同缩放因子下的奖励值

Fig. 11 Reward values at different scaling factors

4.3.3 消融实验

为深入剖析 TR-HPRL 各关键模块对整体性能的贡献, 本节分别去除或替换算法中的部分模块, 观察各模块对网络性能的影响. 主要消融项包括:

1) 不区分流量类别 (TR-RL): 去除基于 1D-CNN 的流量分类模块, 直接将所有流量视为同质化数据, 不进行延迟敏感型、丢包敏感型及容错型流量的区分;

2) 去除 Transformer 模块 (R-HPRL): 用传统全连接网络替换 Transformer 结构, 对比两者在特征捕捉与决策精度上的差异.

同等实验环境下在 GridNet 拓扑网络上对 TR-RL、R-HPRL、TR-HPRL 三种算法的 QoS 感知能力和路由决策性能进行测试, 实验结果如表 5 所示. 观察实验结果可得, 在去除流量感知功能后, 算法 QoS 指标明显降低. TR-RL 算法由于无法针对不同流量需求进行差异化调度, 网络在高负载或混合流量环境下难以同时兼顾延迟和丢包率, 导致整体性能下降. 同样地, 去除 Transformer 的 R-HPRL 算法的 QoS 指标也有一定幅度的下降. Transformer

表 5 消融实验结果

Table 5 Ablation experimental results

算法模型	r_P	r_D
TR-RL	0.88	0.42
R-HPRL	0.92	0.45
TR-HPRL	0.94	0.48

结构在捕捉序列和全局依赖关系方面具有独特优势, 替换为 MLP 后, 模型难以精细区分不同拓扑区域和流量特征, 导致整体性能不及原算法.

4.3.4 分类误差敏感性与鲁棒性分析

为了验证系统对分类误差的鲁棒性, 实验通过向流量类型标签中人为注入不同程度的随机分类错误, 将分类错误率从 0% 逐步增加至 20%, 在 GridNet 拓扑下观察算法平均奖励值的变化情况, 结果如图 12 所示.

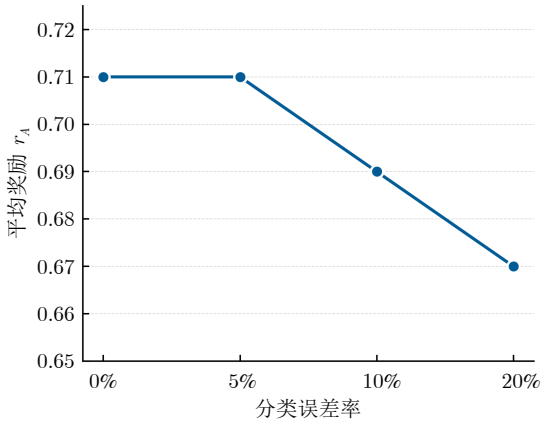


图 12 分类误差率对 QoS 性能影响

Fig. 12 Impact of classification error rate on QoS performance

实验结果表明, 当分类错误率低于 5% 时, TR-HPRL 的性能下降极微. 这得益于共享 GCN 编码器提取了网络拓扑的全局通用特征, 使得智能体在类型判断模糊时仍能基于链路负载状态做出良好决策. 随着错误率增加到 10% 和 20%, 平均奖励出现一定程度的下降, 但整体奖励值仍维持在较高水平. 这说明 TR-HPRL 对分类噪声具有良好的鲁棒性, 一定程度的误分类不会导致路由策略的剧烈震荡.

针对实际部署中可能存在的分类误差传播问题, 可以通过引入置信度门控与软标签融合机制进一步缓解: 当分类置信度低于阈值时, 将该流量映射为容错型, 以降低激进策略带来的风险; 利用分类网络输出的类别概率对不同策略头生成的动作或缩放因子进行加权融合, 从而实现更平稳的路由

控制.

4.3.5　对比实验

为验证所提算法在不同网络场景下的泛化能力与鲁棒性, 本节在 Abilene、GridNet 与 NSFNET 三种经典拓扑下, 将 TR-HPRL 分别与传统启发式方法 ECMP、深度强化学习方法 DRL-TE 以及 TITE 算法进行 QoS 性能对比. 实验结果如图 13、图 14 所示, 分别显示了 ECMP 以及其他三种算法五次独立运行获得的丢包奖励和延迟奖励.

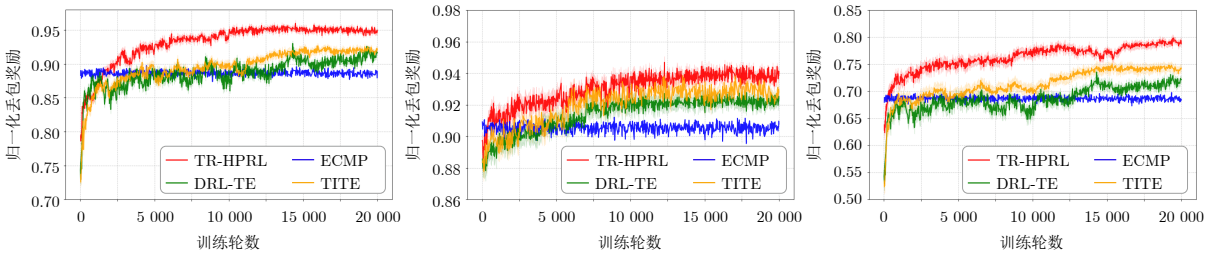
从图中可以看出在三种拓扑网络下, TR-HPRL 在两项 QoS 关键指标上均取得最高的收敛值, 并表现出较好的收敛稳定性. 在丢包奖励上, 相较 ECMP、DRL-TE 以及 TITE 的收敛值, TR-HPRL 显著优于其他基线方法, 表明其能根据不同流量类型采取差异化的路由策略, 提升整体网络资源的调度效率, 从而有效提高全局吞吐性能; 在延迟奖励上, TR-HPRL 在约 3 000 轮训练后即趋于收敛, 并在后续训练中表现出优于其他基线算法的性能, 且具有更高的稳定性, 表明其在面对动态网络环境和多样化流量需求时具备更强的泛化能力和鲁棒性.

综合分析表明, TR-HPRL 显著提升了延迟率与丢包率两项关键 QoS 指标的性能, 这得益于其融合图神经网络特征提取、流量类型分层策略学习以及全局注意力状态评估的创新结构. 在不同网络拓扑下稳定的性能优势展现了 TR-HPRL 良好的泛化能力与可扩展性, 进一步证明了采用基于分层策略强化学习的流量感知路由算法解决复杂网络路由问题的可行性与有效性.

5　结束语

针对传统路由算法在应对多任务流量和动态 QoS 需求时存在的局限性, 本文设计了一种基于分层策略强化学习的流量感知路由算法 TR-HPRL. 该算法首先通过基于深度学习的流量分类方法, 将网络中不同应用的流量划分为延迟敏感型、丢包敏感型和容错型, 从而实现对不同业务需求的精细感知; 然后利用共享 GCN 编码器对网络拓扑进行高效建模作为底层特征, 并在此基础上设计带有类型特异性策略头的分层决策网络, 实现对各类别流量路由策略的差异化以及动态调整; 最后在全局注意力机制和 GAE 优势估计的辅助下, TR-HPRL 实现了网络延迟和丢包率的降低, 提高了资源利用率和 QoS 满足率, 并在实验中展现出快速收敛性和鲁棒性, 为复杂网络环境下的多任务流路由优化提供了切实可行的解决方案.

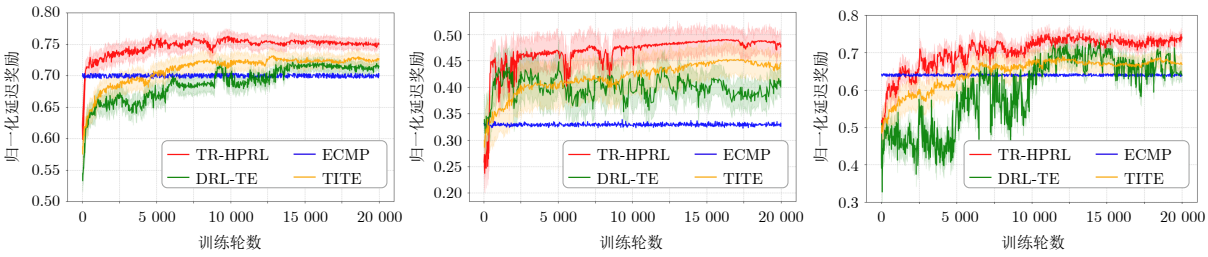
未来研究可进一步丰富状态空间信息, 结合网络流量、链路状态、历史性能指标等多种数据源, 形成多模态状态表示. 通过引入更多传感器信息和上下文数据, 进一步提升模型对复杂网络环境的感知能力和决策精度.



(a) Abilene　(b) GridNet　(c) NSFNET

图 13　丢包奖励

Fig. 13　Packet loss reward



(a) Abilene　(b) GridNet　(c) NSFNET

图 14　延迟奖励

Fig. 14　Delay reward

参考文献

1 China Internet Network Information Center. The 55th statistical report on China′s Internet development [Online], available: https://www3.cnnic.cn/n4/2025/0117/c88-11229.html, October 2, 2025
(中国互联网络信息中心. 第 55 次《中国互联网络发展状况统计报告》[Online], available: https://www3.cnnic.cn/n4/2025/0117/c88-11229.html, 2025-10-02)

2 Ministry of Industry and Information Technology of the People′s Republic of China. 2024 telecommunications industry statistical bulletin [Online], available: https://wap.miit.gov.cn/gxsj/tjfx/txy/art/2025/art_641c048c5d4f4e308098bf6c4e3dcb4a.html, October 2, 2025
(中华人民共和国工业和信息化部. 2024 年通信业统计公报 [Online], available: https://wap.miit.gov.cn/gxsj/tjfx/txy/art/2025/art_641c048c5d4f4e308098bf6c4e3dcb4a.html, 2025-10-02)

3 China Academy of Information and Communications Technology. China digital economy development research report (2024) [Online], available: https://www.caict.ac.cn/kxyj/qwfb/bps/202408/t20240827_491581.html, October 2, 2025
(中国信息通信研究院. 中国数字经济发展研究报告 (2024 年) [Online], available: https://www.caict.ac.cn/kxyj/qwfb/bps/202408/t20240827_491581.html, 2025-10-02)

4 Cai Yue-Ping, Yao Zong-Chen, Li Tian-Chi. A survey on time-sensitive networking: Standards and state-of-the-art. *Chinese Journal of Computers*, 2021, **44**(7): 1378−1397
(蔡岳平, 姚宗辰, 李天驰. 时间敏感网络标准与研究综述. 计算机学报, 2021, **44**(7): 1378−1397)

5 Li Yong-Fu, He Chang-Peng, Zhu Hao, Zheng Tai-Xiong. Nonlinear longitudinal control for heterogeneous connected vehicle platoon in the presence of communication delays. *Acta Automatica Sinica*, 2021, **47**(12): 2841−2856
(李永福, 何昌鹏, 朱浩, 郑太雄. 通信延时环境下异质网联车辆队列非线性纵向控制. 自动化学报, 2021, **47**(12): 2841−2856)

6 Moy J. OSPF version 2 [Online], available: https://www.rfc-editor.org/rfc/rfc2178.html, October 2, 2025

7 Zhang H L, Guo X, Yan J Y, Liu B, Shuai Q J. SDN-based ECMP algorithm for data center networks. In: Proceedings of the IEEE Computers, Communications and IT Applications Conference. Beijing, China: IEEE, 2014. 13−18

8 Mestres A, Rodriguez-Natal A, Carner J, Barlet-Ros P, Alarcón E, Solé M, et al. Knowledge-defined networking. *ACM SIGCOMM Computer Communication Review*, 2017, **47**(3): 2−10

9 Ashtari S, Zhou I, Abolhasan M, Shariati N, Lipman J, Ni W. Knowledge-defined networking: Applications, challenges and future work. *Array*, 2022, **14**: Article No. 100136

10 Akyildiz I F, Lee A, Wang P, Luo M, Chou W. A roadmap for traffic engineering in SDN-OpenFlow networks. *Computer Networks*, 2014, **71**: 1−30

11 Singh S, Jha R K. A survey on software defined networking: Architecture for next generation network. *Journal of Network and Systems Management*, 2017, **25**(2): 321−374

12 Xia Yuan-Qing. Cloud control systems and their challenges. *Acta Automatica Sinica*, 2016, **42**(1): 1−12
(夏元清. 云控制系统及其面临的挑战. 自动化学报, 2016, **42**(1): 1−12)

13 Rusek K, Suárez-Varela J, Almasan P, Barlet-Ros P, Cabellos-Aparicio A. RouteNet: Leveraging graph neural networks for network modeling and optimization in SDN. *IEEE Journal on Selected Areas in Communications*, 2020, **38**(10): 2260−2270

14 Azzouni A, Pujolle G. NeuTM: A neural network-based framework for traffic matrix prediction in SDN. In: Proceedings of the IEEE/IFIP Network Operations and Management Symposium (NOMS). Taipei, China: IEEE, 2018. 1−5

15 Xu Z Y, Tang J, Meng J S, Zhang W Y, Wang Y Z, Liu C H. Experience-driven networking: A deep reinforcement learning based approach. In: Proceedings of the IEEE Conference on Computer Communications. Honolulu, USA: IEEE, 2018. 1871−1879

16 Dai B, Cao Y Y, Wu Z L, Xu Y. IQoR-LSE: An intelligent QoS on-demand routing algorithm with link state estimation. *IEEE Systems Journal*, 2022, **16**(4): 5821−5830

17 Yin X, Wu D, Wang Z L, Shi X G, Wu J P. DIMR: Disjoint interdomain multipath routing. *Computer Networks*, 2015, **91**: 356−375

18 Li J, Giotsas V, Wang Y Y, Zhou S. BGP-multipath routing in the Internet. *IEEE Transactions on Network and Service Management*, 2022, **19**(3): 2812−2826

19 Singh R, Singh Y N, Yadav A. Loop free multipath routing algorithm. arXiv preprint arXiv: 1601.01245, 2016.

20 Lutimath N M, Suresh L, Naikodi C. Efficient power aware multipath routing protocol for MANETs. In: Proceedings of the International Conference on Circuits, Controls, Communications and Computing (I4C). Bangalore, India: IEEE, 2016. 1−4

21 Chen C, Xue F F, Lu Z Y, Tang Z Y, Li C H. RLMR: Reinforcement learning based multipath routing for SDN. *Wireless Communications and Mobile Computing*, 2022, **2022**(1): Article No. 5124960

22 Gurusamy U, Hariharan K, Manikandan M S K. Path optimization of box-covering based routing to minimize average packet delay in software defined network. *Peer-to-Peer Networking and Applications*, 2020, **13**(3): 932−939

23 Prabhavat S, Nishiyama H, Ansari N, Kato N. On load distribution over multipath networks. *IEEE Communications Surveys & Tutorials*, 2012, **14**(3): 662−680

24 He J Y, Rexford J. Toward internet-wide multipath routing. *IEEE Network*, 2008, **22**(2): 16−21

25 Deng G C, Wang K C. An application-aware QoS routing algorithm for SDN-based IoT networking. In: Proceedings of the IEEE Symposium on Computers and Communications (ISCC). Natal, Brazil: IEEE, 2018. 186−191

26 Lin X J, Shroff N B. Utility maximization for communication networks with multipath routing. *IEEE Transactions on Automatic Control*, 2006, **51**(5): 766−781

27 Li Kai-Wen, Zhang Tao, Wang Rui, Qin Wei-Jian, He Hui-Hui, Huang Hong. Research reviews of combinatorial optimization methods based on deep reinforcement learning. *Acta Automatica Sinica*, 2021, **47**(11): 2521−2537
(李凯文, 张涛, 王锐, 覃伟健, 贺惠晖, 黄鸿. 基于深度强化学习的组合优化研究进展. 自动化学报, 2021, **47**(11): 2521−2537)

28 Bao K, Matyjas J D, Hu F, Kumar S. Intelligent software-defined mesh networks with link-failure adaptive traffic balancing. *IEEE Transactions on Cognitive Communications and Networking*, 2018, **4**(2): 266−276

29 Zou G B, Li T F, Jiang M, Hu S X, Cao C H, Zhang B F, et al. DeepTSQP: Temporal-aware service QoS prediction via deep neural network and feature integration. *Knowledge-Based Systems*, 2022, **241**: Article No. 108062

30 de Assis M V O, Carvalho L F, Rodrigues J J P C, Lloret J, Proença Jr M L. Near real-time security system applied to SDN environments in IoT networks using convolutional neural network. *Computers & Electrical Engineering*, 2020, **86**: Article No. 106738

31 Arulkumaran K, Deisenroth M P, Brundage M, Bharath A A. Deep reinforcement learning: A brief survey. *IEEE Signal Processing Magazine*, 2017, **34**(6): 26−38

32 Tang J Q, Mihailovic A, Aghvami H. Constructing a DRL decision making scheme for multi-path routing in all-IP access network. In: Proceedings of the IEEE Global Communications Conference. Rio de Janeiro, Brazil: IEEE, 2022. 3623−3628

33 Altamirano J C, Guitouni M, Hassan H, Drira K. Routing optimization based on DRL and generative adversarial networks for SDN environments. In: Proceedings of the IEEE Network Operations and Management Symposium. Seoul, Republic of Korea: IEEE, 2024. 1−5

34 Casas-Velasco D M, Rendon O M C, da Fonseca N L S. Intelligent routing based on reinforcement learning for software-defined networking. *IEEE Transactions on Network and Service Management*, 2021, **18**(1): 870−881

35 Ye M H, Zhang J J, Guo Z H, Chao H J. DATE: Disturbance-aware traffic engineering with reinforcement learning in software-

defined networks. In: Proceedings of the 29th IEEE/ACM International Symposium on Quality of Service (IWQOS). Tokyo, Japan: IEEE, 2021. 1−10

36 Zhang J J, Ye M H, Guo Z H, Yen C Y, Chao H J. CFR-RL: Traffic engineering with reinforcement learning in SDN. *IEEE Journal on Selected Areas in Communications*, 2020, **38**(10): 2249−2259

37 Rezaei S, Liu X. Multitask learning for network traffic classification. In: Proceedings of the 29th International Conference on Computer Communications and Networks (ICCCN). Honolulu, USA: IEEE, 2020. 1−9

38 Zhang Y, Qiu L X, Xu Y Z, Wang X J, Wang S J, Paul A, et al. Multi-path routing algorithm based on deep reinforcement learning for SDN. *Applied Sciences*, 2023, **13**(22): Article No. 12520

39 He N, Yang S, Li F, Trajanovski S, Zhu L H, Wang Y, et al. Leveraging deep reinforcement learning with attention mechanism for virtual network function placement and routing. *IEEE Transactions on Parallel and Distributed Systems*, 2023, **34**(4): 1186−1201

40 He Q, Wang Y, Wang X W, Xu W Q, Li F L, Yang K Q, et al. Routing optimization with deep reinforcement learning in knowledge defined networking. *IEEE Transactions on Mobile Computing*, 2024, **23**(2): 1444−1455

41 Ding M J, Guo Y Y, Huang Z B, Lin B, Luo H. GROM: A generalized routing optimization method with graph neural network and deep reinforcement learning. *Journal of Network and Computer Applications*, 2024, **229**: Article No. 103927

42 Xu Z Y, Yan F Y, Singh R, Chiu J T, Rush A M, Yu M L. Teal: Learning-accelerated optimization of WAN traffic engineering. In: Proceedings of the ACM SIGCOMM Conference. New York, USA: ACM, 2023. 378−393

43 Lin B, Guo Y Y, Luo H, Ding M J. TITE: A Transformer-based deep reinforcement learning approach for traffic engineering in hybrid SDN with dynamic traffic. *Future Generation Computer Systems*, 2024, **161**: 95−105

44 Hu F, Hao Q, Bao K. A survey on software-defined network and OpenFlow: From concept to implementation. *IEEE Communications Surveys & Tutorials*, 2014, **16**(4): 2181−2206

45 Richardson L, Ruby S. *RESTful Web Services*. Sebastopol: O'Reilly Media, Inc., 2007.

46 Draper-Gil G, Lashkari A H, Mamun M S I, Ghorbani A A. Characterization of encrypted and VPN traffic using time-related features. In: Proceedings of the 2nd International Conference on Information Systems Security and Privacy (ICISSP). Rome, Italy: SciTePress, 2016. 407−414

47 Hagberg A A, Schult D A, Swart P J. Exploring network structure, dynamics, and function using NetworkX. In: Proceedings of the 7th Python in Science Conference (SciPy). Pasadena, USA: SciPy, 2008. 11−16

赵之栩 北京理工大学自动化学院硕士研究生. 主要研究方向为路由优化, 对抗攻击, 机器学习.
E-mail: zhixuzhaobit@163.com.
(**ZHAO Zhi-Xu** Master student at the School of Automation, Beijing Institute of Technology. His research interests include routing optimization, adversarial attacks, and machine learning.)

刘 坤 北京理工大学自动化学院研究员. 主要研究方向为网络化控制理论与应用, 复杂网络系统安全. 本文通信作者.
E-mail: kunliubit@bit.edu.cn
(**LIU Kun** Professor at the School of Automation, Beijing Institute of Technology. His research interests include theory and applications of networked control, and security of complex networked systems. Corresponding author of this paper.)

王璐瑶 北京理工大学自动化学院硕士研究生. 主要研究方向为对抗攻击与防御, 时间序列, 网络安全.
E-mail: luyaowangbit@gmail.com.
(**WANG Lu-Yao** Master student at the School of Automation, Beijing Institute of Technology. Her research interests include adversarial attacks and defense, time series, and cyber security.)

夏元清 北京理工大学自动化学院教授. 主要研究方向为云控制, 云数据中心优化调度管理, 智能交通, 模型预测控制, 自抗扰控制, 鲁棒控制, 复杂网络系统控制与安全, 网络化控制理论与应用, 飞行器控制和空天地一体化网络协同控制.
E-mail: xia_yuanqing@bit.edu.cn
(**XIA Yuan-Qing** Professor at the School of Automation, Beijing Institute of Technology. His research interests include cloud control, cloud data center optimization scheduling and management, intelligent transportation, model predictive control, active disturbance rejection control, robust control, control and security of complex networked systems, theory and applications of networked control, flight control and networked cooperative control for integration of space, air and earth.)