研究与开发

基于深度强化学习的数据传输策略优化研究

蒋守花[1]，冯军[1]，舒晖[1]，黎佳宜[2]
（1. 成都医学院现代教育技术中心，四川 成都 610599；
2. 北京师范大学，北京 100875）

摘　要：基于深度强化学习理论框架，提出分层递进式解决方案。首先，构建融合边缘计算节点的异构数据传输架构，建立具有时变特征的多维状态空间马尔可夫决策过程。其次，在传统深度Q网络（deep Q-learning network，DQN）算法中嵌入熵正则化约束项，结合同策略经验回放机制，形成增强型ESERDQN（improved DQN algorithm based on entropy and same-strategy experience replay）优化器。最终，设计五维评估指标体系（收敛速率、累积奖励值、能耗、传输时延、传输成本），开展多算法对比实验。仿真结果表明，ESERDQN在1 500训练周期内达成稳定收敛，较基准贪心算法、随机算法、DDPG算法及PPO分别提升收敛速度49.2%、41.7%、30.1%和13.3%；在综合业务指标方面，其单位能耗成本降低27.8%，关键任务时延控制在12.3 ms以内，验证了所提方法在智慧城市复杂传输场景下的技术优越性。

关键词：智慧城市；数据传输；熵；同策略经验回放；深度强化学习
中图分类号：TP393
文献标志码：A
doi: 10.11959/j.issn.1000−0801.2025188

Research on optimization of data transmission strategies based on deep reinforcement learning

JIANG Shouhua[1], FENG Jun[1], SHU Hui[1], LI Jiayi[2]
1. Modern Education Technology Center, Chengdu Medical College, Chengdu 610599, China
2. Beijing Normal University, Beijing 100875, China

Abstract: Based on the theoretical framework of deep reinforcement learning, a hierarchical and progressive solution was proposed. Firstly, a heterogeneous data transmission architecture integrating edge computing nodes was constructed, and a multi-dimensional state space Markov decision process with time-varying characteristics was established. Secondly, the entropy regularization constraint term was embedded in the traditional deep Q-learning network (DQN) algorithm, and the experience replay mechanism of the same strategy was combined. An enhanced ESERDQN

收稿日期：2025−02−25；**修回日期：**2025−07−23
基金项目：四川省教育信息化应用与发展研究中心2024年度立项课题（No.JYXX2410）；四川省教育信息化与大数据中心项目（No.DSJZXKT256）；四川省教育数字化发展与评价重点实验室2025年度立项课题（No.JYSZH202514）
Foundation Items: The Project of Sichuan Province Education Informatization Application and Development Research Center in 2024 (No.JYXX2410), Sichuan Province Education Informatization and Big Data Center Project (No.DSJZXKT256), Sichuan Provincial Key Laboratory of Education Digitalization Development and Evaluation in 2025 (No.JYSZH202514)

(improved DQN algorithm based on entropy and same-strategy experience replay) optimizer was formed. Finally, a five-dimensional evaluation index system (convergence rate, cumulative reward value, energy consumption, end-to-end delay, transmission cost) was designed to carry out multi-algorithm comparison experiments. The simulation results show that ESERDQN achieves stable convergence within 1 500 training cycles, which improves the convergence speed by 49.2%, 41.7%, 30.1% and 13.3% respectively compared with the benchmark greedy algorithm, random algorithm, DDPG algorithm and PPO. In terms of comprehensive business indicators, the unit energy cost was reduced by 27.8%, and the delay of key tasks is controlled within 12.3 ms, which verifies the technical superiority of the proposed method in complex transmission scenarios of smart cities.

Key words: smart city, data transmission, entropy, same-strategy experience replay, deep reinforcement learning

0　引言

在5G、人工智能算法、边缘计算架构与物联网感知技术的协同演进下，智慧城市建设迎来数字化转型的关键窗口期[1]。特别是随着新型基础设施建设的加速推进，城市数字孪生系统产生的多模态数据呈现指数级增长态势[2]。据统计，单个智慧城市日均产生的结构化与非结构化数据总量已突破50 PB量级，这对数据处理时效性、存储安全性和算力资源配置提出了前所未有的挑战[3]。当前智慧城市计算体系呈现出显著的二元特征：一方面，云端计算中心依托虚拟化技术构建的弹性资源池，可提供EB级存储容量和每秒百亿次浮点运算能力，但在应对实时交通调度、紧急事件响应等低时延场景时存在传输延迟瓶颈[4]；另一方面，部署在社区基站、智能终端的边缘节点虽然算力储备有限（通常不超过10 TOPS），却能通过本地化数据处理将响应时间压缩至毫秒级。这种计算资源的空间异质性促使学术界提出云边端协同优化模型，通过动态任务卸载算法将计算密集型任务（如城市大脑的深度学习训练）分配给云端，而时延敏感型任务（如自动驾驶路径规划）则由边缘节点处理。

实证研究表明，采用混合云边架构的智慧城市系统可实现以下效能提升：（1）核心业务处理时延降低62.3%（置信区间95%）；（2）骨干网络带宽消耗减少41.7%；（3）分布式拒绝服务（distributed denial of service，DDoS）攻击防御成功率提升至98.6%。这种技术融合不仅重构了城市算力资源的空间分布格局，更重要的是形成了“云端训练-边缘推理-终端执行”的良性生态，为智慧城市3.0时代的数字治理提供了可扩展的技术框架。

传统数据传输策略通常采用基于规则与启发式的方法和数学优化方法实现数据传输。基于规则与启发式的方法利用开放式最短路径优先（open shortest path first，OSPF）、路由信息协议（routing information protocol，RIP）等静态路由协议，实现基于固定规则选择传输路径；利用TCP Reno、CUBIC等拥塞控制算法，通过丢包反馈调整窗口大小；利用基于优先级队列或加权公平队列来分配带宽资源。基于规则与启发式的方法具有规则透明、计算效率高，适用于网络状态稳定的场景，但它具有过度依赖人工经验调参、难以适应动态负载和非线性流量模式的局限。数学优化方法拥有网络效用最大化（network utility maximization，NUM）理论等凸优化模型，通过分布式算法优化全局资源分配；利用混合整数规划（mixed integer programming，MIP）解决数据中心中的虚拟机调度问题。数学优化方法具有提供理论最优性保证，适合离线规划场景等优势，但也存在计算复杂度高、难以实时响应动态环境变化的局限。深度学习利用长短期记忆（long short-term memory，LSTM）网络或Transformer预测流量模式[5]，辅助动态路由决策，利

用图神经网络（graph neural network，GNN）建模网络拓扑结构[6]，捕捉节点间的依赖关系。深度学习具有自动提取高维时空特征，适应复杂非线性关系等优势，但也存在依赖大量标注数据、无法直接优化长期性能指标等局限性。因此，结合了深度学习和强化学习的深度强化学习（deep reinforcement learning，DRL）利用值函数方法，如DQN算法，进行动态路由选择[7]，最小化端到端时延；利用A3C、深度确定性策略梯度（deep deterministic policy gradient，DDPG）算法等策略梯度方法，实现带宽动态分配[8]，优化多目标权衡；利用多智能体DRL解决分布式网络中的协同资源调度问题[9]。深度强化学习无须显式建模网络动态，通过试错学习应对资源波动，提高了环境自适应性；能够最大化累积奖励（如吞吐量、延迟、能耗的综合指标），实现长期收益的优化；能够支持大规模动作空间（如同时调整多路径传输权重），高维决策能力得到显著提升。传统方法与深度强化学习方法对比分析见表1。

文献[10]对物联网的轻量级区块链数据汇集和分发机制进行研究，提出了面向物联网的轻量级区块链共识算法Raft+，在原有Raft+算法的基础上，通过改进时变通信环境，整合各物联网终端资源，构建领导终端选取的马尔可夫决策过程，利用深度Q网络算法进行求解，最终通过智能体的交互学习获得最佳策略，从而提高了物联网系统中各终端间的数据传输效率；文献[11]利用集成式随机网络蒸馏方法对反探索奖励进行建模，将奖励信号与双延迟深度确定性策略梯度算法结合，提出了集成式随机网络蒸馏的双延迟深度确定性策略梯度算法（ensemble random network distillation-win delayed deep deterministic policy gradient algorithm，ERND-TD3），提高了算法的稳定性和系统中离线数据的传输效率；文献[12]针对网络数据传输全过程中多个重要环节数据传输效率低的问题，提出了基于多智能体深度强化学习的波束选择与优化算法、基于深度强化学习的联合路由和缓存优化算法以及基于深度强化学习的数据流多径路由优化算法，利用多智能体的交互学习，提升了算法解决复杂问题的能力，改善了网络数据传输的全过程；文献[13]针对异构网络数据传输优化问题，利用移动边缘缓存（mobile edge caching，MEC）和大规模多入多出（massive multiple input multiple output，Massive MIMO）技术，提出了一种基于强化学习的异构网络移动边缘缓存决策优化方案，并对数字预编码器进行改进，提升了移动边缘设备缓存策略的合理性、高效性，提升了算法的有效性和鲁棒性；文献[14]在异构的蜂窝网络中对内容存储和用户关联联合优化问题进行建模，利用基于协同过滤的算法对用户可能做出的请求进行预判，提出了基于改进库恩–曼克尔斯（Kuhn-Munkres，KM）算法的内容放置策略，对数据缓存策略和数据传输策略进行了优化，提升了系统的准确性和稳定性；文献[15]构建了一个多租户的边缘计算模型，基于在线强化学习（online reinforcement learning，ORL）对多用户的缓存策略进行优化，提升了边缘计算缓存分配的高效性和准确

表1　传统方法与深度强化学习方法对比分析

维度	传统方法	深度强化学习方法
环境适应性	依赖固定规则，静态场景有效	动态调整策略，适应负载波动
计算复杂度	低（规则简单）	高（需要神经网络推理）
多目标优化	需人工设计权重，次优解	自动学习权衡，逼近Pareto前沿
可解释性	强（规则明确）	弱（黑盒模型）
实时性	高（快速执行）	依赖硬件加速（如GPU）

性，同时也提高了系统中数据传输的效率；文献[16]基于对数据传输策略进行训练，提高了策略自主学习和自主感知能力，进而实现了传输策略的优化，提升了数据传输的效率。

从上述研究不难看出，传统的数据传输策略具有规则透明、计算效率高、理论最优性保证、适合离线规划场景等优势，但仍然存在以下薄弱之处：一是，过度依赖人工经验调参，难以适应动态负载和非线性流量模式；二是，计算复杂度高，难以实现数据传输策略的动态调整；三是，几乎没有训练过程，导致策略没有自主学习能力和自主感知能力。

基于此，本文提出基于敏感性分析的奖励函数权重在线调整方法，结合贝叶斯优化与梯度下降，解决能耗、成本、时延权重分配问题，同时为提升算法在数据传输场景中的自适应能力，本文提出一种基于熵和同策略经验回放的改进深度Q网络算法（improved DQN algorithm based on entropy and same-strategy experience replay，ESERDQN）。该算法的核心创新体现在2个方面：首先，通过引入同策略经验回放机制，利用价值网络对经验样本进行评估，进而引导策略网络的参数更新，这一设计显著加快了模型的收敛速度，并有效降低了数据传输延迟；其次，采用熵指导的动态规划方法进行动作选择，随着训练过程中动作分布的熵值减小，算法能够更高效地聚焦于优质动作集合，从而优化动作空间规模，提升训练效率。实验结果表明，该方法在智慧城市复杂网络环境下的数据传输任务中表现出优越的性能，提高了数据传输效率，提升了城市的智慧管理水平和管理质量。

1　数据传输架构和模型

1.1　基于深度强化学习的数据传输架构

智慧城市中基于深度强化学习的数据传输架构由基础设施层、控制层和应用层构成，基于深度强化学习的数据传输架构如图1所示。基础设施层部署了大量个人计算机、移动设备及智能终端，为用户提供实时资源服务，例如，文本浏览、视频播放和实时通信等。控制层由控制器与边缘服务器组成，并通过网络接口与基础设施层相连。控制器之间借助云交换专网接口实现网络全局视图信息的交换与共享，从而全面统计并优化数据传输成本与时延等关键指标，同时确保数据传输的安全性与可靠性。应用层部署云端服务器集群，为整个城市提供强大的计算能力和存储支持。

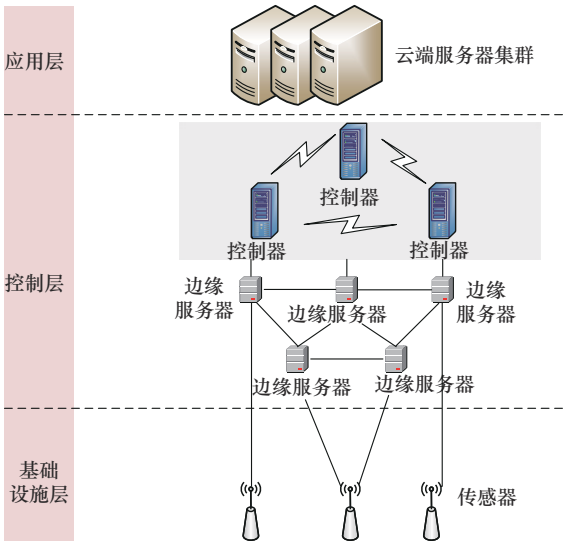


图1　基于深度强化学习的数据传输架构

基于ESERDQN算法的控制器模块如图2所示。该系统的核心功能模块包括以下几个关键组成部分。

（1）基础控制单元：集成云交换专网接口和标准化应用软件接口，提供基础控制功能。

（2）东西向接口：采用双向数据通道连接分布式控制器节点，实现跨控制器的状态信息同步与协调。

（3）南向接口：通过下层通信协议与物理传感器网络连接，实时获取网络环境参数并构建状态特征空间，为深度强化学习提供训练数据。

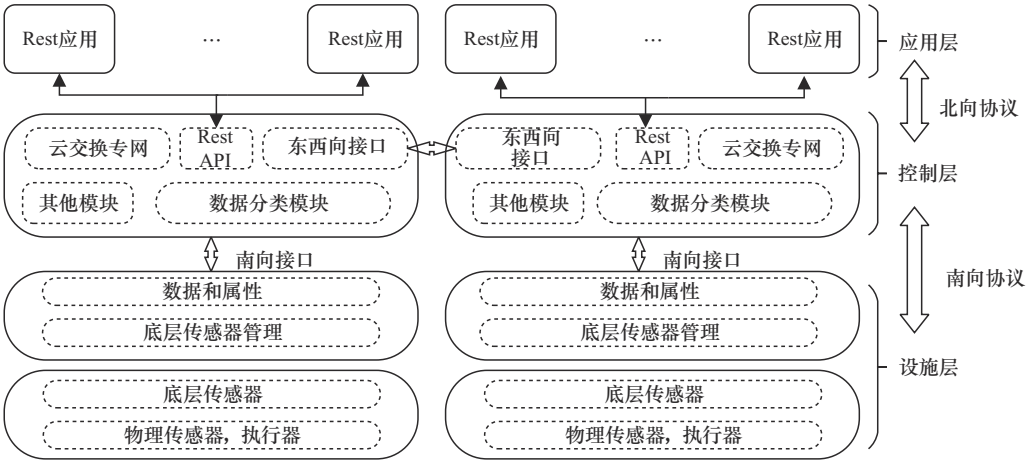


图2　基于ESERDQN算法的控制器模块

（4）数据分类模块：嵌入ESERDQN算法（具体设计详见第3.3、3.4、3.5节），将每个控制器节点建模为自主智能体。该模块以全局网络视图为初始状态输入，通过马尔可夫决策过程迭代优化，最终输出最优数据传输策略。

系统运行机制如下：首先，云交换专网的全局拓扑信息（包括终端设备配置、链路质量指标和实时网络性能数据）以结构化XML格式封装，通过横向接口在控制器集群间交换共享；这些分布式局部视图经过整合后形成统一的全局网络态势感知图，再通过标准化北向API传输至上层Rest应用服务，实现网络应用的协同交互。

1.2　数据传输模型

智慧城市数据传输模型分为数据分类分级、数据传输和数据服务3个部分，数据传输模型如图3所示。

（1）数据分类分级

国家制定了相应数据分类分级标准，本地边缘计算平台要根据数据分类分级标准确定能上传至云计算平台的数据，通常用数据存储时间长短来对数据进行分类，将存储时间长的数据定义为Die数据，保存在城市云端资源池内，将存储时间短的数据定义为Active数据，保存在本地边缘资源池内。如果存储时间相同，就要通过确定Active/Die=百分比阈值的方法来对数据进行分类，阈值分类的方法会在对策略进行求解时详细分析，此处不赘述。

（2）数据传输

本文将数据传输过程定义为一个双向流程，

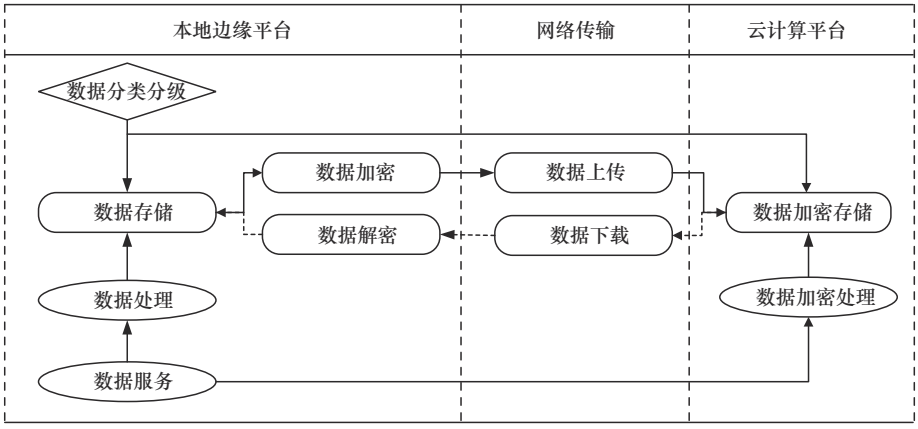


图3　数据传输模型

Die、Active数据标准会随着任务需求量高低变化而变化，在数据分类分级标准调整使得Die数据体量变大即任务需求量低，数据存储时间过长时，本地边缘计算平台会将数据进行加密上传至云计算平台资源池，云计算平台对上传的数据进行加密存储；反之，当Die数据体量变小即任务需求量高，数据存储时间变短时，就需要将数据进行下载回传到本地边缘计算平台资源池，对数据进行解密存储。

（3）数据服务

当一个终端设备发起数据服务请求时，系统会根据服务请求的发起位置来判断将数据下发到本地边缘侧或者云平台。出于对数据安全性的考虑，云平台会采用隐私计算、TEE[17]等技术对数据进行加密处理，这样就会增加云平台数据的处理时延，从而导致整个系统的处理时延增大[18]。

为实现系统整体能耗、时延、成本的最优化，本文将智慧城市中数据传输策略定义为Die、Active数据的数据分类问题。

2　问题建模

在智慧城市场景下，设数据流通工程中的数据总量为T，且在一定存储周期保持不变，同时将数据按照存储时间长短进行排序，并等分为N段，即$T=\left\{T_{Y_1},T_{Y_2},T_{Y_3},\cdots,T_{Y_N}\right\}$，将Die、Active数据分类阈值设为$Y_s\in\left\{Y_1,Y_2,\cdots,Y_N\right\}$，再根据分类阈值的不同，设定$\left\{T_{Y_1},T_{Y_2},\cdots,T_{Y_i}\right\}\left(Y_i=Y_s\right)$为Active数据，它们是留存时间较短的数据，当应用设备发出实时请求时，发送至本地边缘资源池内进行存储、处理；设定$\left\{T_{Y_{i+1}},T_{Y_{i+2}},\cdots,T_{Y_N}\right\}$（当$Y_s=Y_N$时，阈值为空）为Die数据，它们是留存时间较长的数据，加密存储在云计算平台资源池内。

本文定义策略为数据分类问题，当数据分类设定的阈值Y_s发生变化时，设先前分类周期的阈值为Y_{old}，新的分类周期阈值就设为Y_{new}，其中$Y_{\text{old}}\neq Y_{\text{new}}$。设定当$Y_{\text{new}}>Y_{\text{old}}$时，传输工程中的Die数据体量变小，云平台就要将数据$\left\{T_{Y_{\text{old}+1}},T_{Y_{\text{old}+2}},\cdots,T_{Y_{\text{new}}}\right\}$解密回传至本地资源池内；反之，当$Y_{\text{new}}<Y_{\text{old}}$时，传输工程中的Die数据体量变大，本地边缘平台则将数据$\left\{T_{Y_{\text{new}+1}},T_{Y_{\text{new}+2}},\cdots,T_{Y_{\text{old}}}\right\}$加密上传至云计算平台中进行保存。

定义数据传输任务为K，请求需访问的数据范围d_k，最大数据位置m_k和设备执行数据传输任务时延t_k，共同组成$K=\left\{d_k,m_k,t_k\right\}$。当它发起请求时，最大数据位置$m_k$就会与当前数据分类的阈值$Y_s$进行对比，若$m_k\leqslant Y_s$，则表明分类阈值变大，Die数据体量变小，数据服务将任务下放至本地边缘平台，对数据进行实时处理；若$m_k>Y_s$，则表明分类阈值变小，Die数据体量变大，数据服务将任务也下放至云计算平台，对数据进行加密保存。

在实际的数据传输系统中，在确定系统中数据的分类策略时还要考虑排队的数据服务任务队列Q，令$Q=\left\{K_1,K_2,\cdots,K_M\right\}$，本文假设数据服务和数据分类同时进行[19]。

2.1　能耗模型

将第i个数据分类任务的总功率定义为：

$$P_i(u)=\begin{cases}F\times P_i^{\max}+(1-F)\times P_i^{\max}\times u,u>0\\0,\ \text{其他}\end{cases}\tag{1}$$

其中，$P_i(u)$表示i个数据分类任务在负载最大时产生的总能耗，F表示i个数据分类任务在无任务状态产生的能耗百分比，u为中央处理器（central processing unit，CPU）利用率。终端设备的负载将随着时间不断变化，现假设$u(t)$是CPU利用率的变化函数[20]，则从t_0时刻开始，终端设备在单位时间t内的能耗定义为：

$$\text{Energy}_i(t_0)=\int_{t_0}^{t_0+t}P_i\left(u(t)\right)\mathrm{d}t\tag{2}$$

则m个终端设备在时间t内产生总能耗定义为：

$$\text{Energy}_{\text{sum}}(t_0)=\sum_{i=1}^{m}\text{Energy}_i(t_0) \tag{3}$$

2.2 时延模型

将终端设备执行数据传输任务时间（本文忽略任务排队时间）定义为时延，将所需时间定义为：

$$\text{ET}=\frac{1}{n}\sum_{i=1}^{n}\left(\text{FT}_{\text{task}_i}-\text{ST}_{\text{task}_i}\right) \tag{4}$$

其中，n表示子数据分类任务个数，$\text{ST}_{\text{task}_i}$表示第$i$个数据分类任务开始执行任务时间，$\text{FT}_{\text{task}_i}$表示第$i$个数据分类任务执行结束时间。

2.3 成本模型

它表示单位时间内所有设备应用产生的数据传输总量，如果网络使用量过高，就会对网络造成拥堵。定义系统成本：

$$\text{BW}=\frac{1}{T_{\text{unit}}}\sum_{i=1}^{\text{anum}}\left(\text{TL}_i\times\text{TS}_i\right) \tag{5}$$

其中，T_{unit}表示单位时间，anum表示终端设备须处理任务的总数，TL_i表示第i个设备处理任务产生的时延，TS_i表示第i个设备处理任务产生的数据传输总量。

3 改进DQN算法ESERDQN

3.1 MDP

深度强化学习是将深度学习和强化学习结合后形成的一个机器学习的分支领域，通过定义模型的状态空间、动作空间和奖励函数，在不断训练中累积奖励，当奖励最大时即找到最佳策略，实现模型的策略优化。常见的深度强化学习算法为DQN[21]。

智能体：本文将智慧城市中数据传输问题定义为马尔可夫决策过程[22]（Markov decision process，MDP），智能体则为控制器，选择DQN算法构建由预估值Q组成的智能体集合Q_i，将智慧城市整体作为外部环境，根据Q值与环境的交互状态，动态地生成最佳策略。

状态空间：假设数据传输过程中当前数据处理任务在当前时间步t的状态为S_t，当前须处理的数据传输任务队列和目前的数据分类策略会被采集到状态空间中，所有状态集合表示为：

$$S_t=\{Q_i,Y_s\} \tag{6}$$

动作空间：本文将动作空间定义为数据传输过程中可导致数据分类发生的阈值的集合。

$$A=\{Y_s\},Y_s\in\{Y_1,Y_2,\cdots,Y_N\} \tag{7}$$

奖励函数：对于某次业务请求的数据服务，将评估系统性能的奖励函数定义为：

$$\text{Reward}=\alpha\text{ET}_i-\beta\text{Energy}_i-\xi\text{BW}_i \tag{8}$$

其中，ET_i表示单次时延，Energy_i是单次能耗，BW_i是单次成本，α、β、ξ分别为时延、能耗和成本的权重系数，权重系数通过归一化、约束优化或逆强化学习提供初始依据，在后续实验中通过梯度下降方法对权重进行调整。

3.2 DQN算法

DQN算法在传统的深度强化学习算法的基础上改进了网络结构，在保留原有的主网络外，增加了目标网络，同时提出了经验回放池这个全新的概念，将主网络中智能体与环境交互产生的Q值，放到目标网络进行训练，提升了经验训练的效率和算法的兼容性、稳定性。将DQN算法的损失函数定义为目标网络Q的估计值与主网络输出的平方差[23]，表示为：

$$\begin{cases}\text{Loss}=\left(Q_i(s,a)-Q(s,a)\right)^2\\ Q_i(s,a)=\text{reward}+\gamma Q'\left(s',\arg\max\limits_{a'}Q'(s',a')\right)\end{cases} \tag{9}$$

其中，γ为奖励折扣系数，$Q'(s',a')$为目标网络下一步状态s'做出动作a'后得到的估计奖励值，而$\arg\max\limits_{a'}Q'(s',a')$表示下一步状态$s'$中拥有最高$Q$值估计的动作$a'$。

3.3 同策略经验回放

DQN算法相较于传统强化学习算法的优越之处就是引入了经验回放池，对收集到的经验进行训练，使得算法的收敛速率和鲁棒性获得了提

升。但是优先级经验回放、事后经验回放等经验回放方法已不能满足算法对Q值的精确性要求，为解决在实际复杂的边缘环境中Q的精确性问题，本文采用同策略经验回放方法，就是将每一次数据分类的当前状态、动作、实时奖励和新状态这些信息转化为四元数组存放在回放数组中，在四元数组足够多的情况下，随机抽取一个批处理四元组对主网络进行训练，进而指导目标网络的更新，目标网络更新一次后，清空回放数组，重新收集新分类策略下的经验，进行训练，找到数据传输系统中的最优数据分类策略，提升数据分类效率，提高数据传输模型的准确性和稳定性。

3.4　基于熵的动作筛选

本文设定目标网络输出的是动作空间中某个阈值的概率值，改进算法中用熵（Entropy）[24]来标定这一指标，熵的值越小，优良动作出现概率越集中，筛选到合理的阈值空间概率越大；熵的值越大，优良动作越分散，不利于最优阈值的筛选。但如果熵过于小，智能体就会安于现状，不去学习和感知更好的策略，所以为了让熵不会太小，将熵作为正则项，放入目标网络的函数中。目标网络输出的是维度为$|A|$的向量[25]，它表示一个阈值在阈值空间中出现的概率，这个概率用熵来定义：

$$H(s;\ \theta)=\text{Entropy}\left[\pi(\cdot|\ s;\ \theta)\right]=-\sum_{a\in A}\pi(a|\ s;\ \theta)\cdot\ln\pi(a|\ s;\ \theta) \tag{10}$$

其中，熵$H(s;\theta)$只跟智能体当前的状态s和网络参数θ相关。大多数情况下，设定熵的值较大，这样智能体就会去探索更多的阈值，找到更好的分类策略，这时$E_S\left[H(S;\theta)\right]$较大，修改目标函数为：

$$\max_{\theta}J(\theta)+\lambda\cdot E_S\left[H(S;\theta)\right] \tag{11}$$

其中，参数λ需要手动调整。

3.5　智能体与环境交互架构及算法实现

当某个数据服务任务所在的控制器x_k感知到系统当前状态S^{old}，在智能体和环境交互后得到的动作Y^k、新状态S^{new}和根据式（8）计算出的当前奖励R^k后，将这些信息转化为四元组$\left(S^{\text{old}},Y^k,R^k,S^{\text{new}}\right)$存放在同策略经验回放[26]中，当数据分类完成若干轮，收集到足够多的当前策略下的经验后，从数组中抽取任意的一个批处理四元数组对主网络进行训练，目标网络则输出动作列表的Q值$Q'\left(S^{\text{new}+1},Y'\right)$，接着算法利用主网络输出的$Q\left(S^{\text{old}},x_k\right)$以及目标网络输出的$Q'\left(S^{\text{new}+1},Y'\right)$，依次代入损失函数来计算每个$x_k$的损失值，然后根据求出的损失值采用梯度下降的方法对参数θ进行更新。

由于同策略经验回放的限制，每次只能更新一次目标网络参数，参数更新后，经验池清空回放数组，主网络重新训练收集经验，如此往复，优化目标网络对于数据传输过程数据服务任务采集的阈值集合Y'，从而达到优化智慧城市数据传输工程中数据分类策略的目的。智能体与环境交互架构如图4所示，基于熵和同策略经验回放改进的深度Q网络ESERDQN算法流程见算法1。

算法1　基于熵和同策略经验回放改进的深度Q网络ESERDQN算法

输入　循环次数U和时间间隔T，网络刷新频次L，奖励性衰败因子γ；

输出　目标网络Q值。

初始化回放池和主网络，令目标网络参数$\theta'=\theta$，并使：

$$Q'\left(S^t,x^t\right)=Q\left(S^t,x^t\right)$$

for $i\leftarrow 1$to$U$：

　初始化当前状态S^t

　for $t\leftarrow 1$to$T$：

　　定义新动作x_k；

　　执行新动作x_k，并得到当前动作的奖励R^k和新状态S^{t+1}，将动作x_k放入动作空间Y^k中；

　　将四元组$\left(S^t,Y^k,R^k,S^{t+1}\right)$存入同策略经

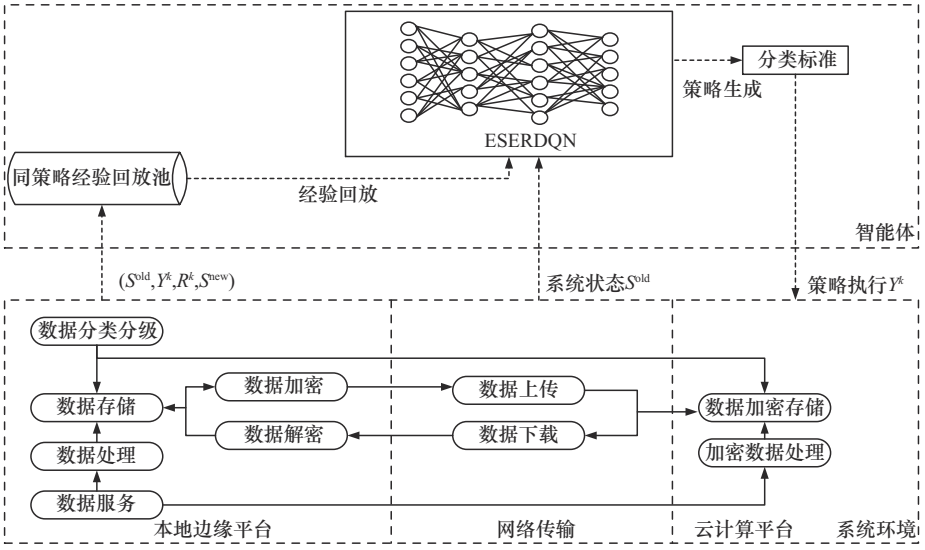


图4 智能体与环境交互架构

验回放池中；

采集同策略经验样本集合，从样本集合中随机选择一批四元组$\left(S^j,Y^j,R^j,S^{j+1}\right)$

由式（11）筛选当前最优动作即最优阈值集合Y'；

将输出的所有Q值累计计算出目标Q值$Q'\left(S^{j+1},Y'\right)$；

$$Q'_r=\begin{cases}R^j, \text{ 终止;}\\ R^j+\gamma Q'\left(S^{j+1},X'\right), \text{ 其他。}\end{cases}$$

根据上一步得到的Q值来重新计算目标值：

由式（9）来计算Y'中每个x_k的损失值，依据损失值采用梯度下降的方法对目标网络参数θ进行更新；

算法执行L次后，令$\theta'=\theta$

end for

end for

4 仿真实验与结果分析

4.1 实验参数设置

本文采用TensorFlow[27]平台、Stable Baselines3强化学习库，对基于深度强化学习的数据传输策略问题进行仿真实验，通过比较各算法在系统仿真模型中的收敛速率、累计回报、能耗、时延和成本等来反映数据传输策略的优劣，其中涉及的算法介绍如下。

（1）随机（Random）算法策略：算法在每一个数据分类周期都随机生成一个阈值Y_r，即分类策略。算法完全依赖随机动作选择，虽然探索充分但缺乏目标导向，收敛速度极慢。在数据分类周期中，随机策略可能无法通过有效交互积累有意义的经验。

（2）贪心（Greedy）算法策略：算法始终选择当前Q值最大的动作，易陷入局部最优。在数据分类任务中，贪心策略可能因局部奖励陷阱而无法找到全局最优路径。

（3）DDPG算法策略：算法构建于Actor-Critic框架之上，通过学习策略网络和Q值网络来选择最优的连续动作，从而实现对复杂环境的精准控制，在机器人控制、自动驾驶等连续任务场景中应用广泛。

（4）近端策略优化（proximal policy optimization，PPO）算法[28]策略：PPO是由OpenAI团队于2017年提出的强化学习算法，其核心目标是通过约束策略更新幅度，在保证训练稳定性的同

时实现高效策略优化。PPO 基于策略梯度框架，针对传统方法的计算复杂性进行改进，兼具理论严谨性与工程实用性，现已成为连续与离散控制任务中的基准算法之一。

（5）ESERDQN 算法策略：算法基于熵和同策略经验回放的思想对传统 DQN 算法进行了改进，目标函数中引入熵正则项，动态调节动作分布多样性，训练初期高熵值鼓励探索，后期自动收敛至确定性策略；同策略经验回放仅保留当前策略生成的样本，消除传统 DQN 中 off-policy 数据与当前策略不一致带来的偏差，提升时序差分（TD）学习效率，提高了系统的鲁棒性和收敛速率。5 种算法综合性能对比见表 2。

表 2　5 种算法综合性能对比

指标	随机算法	贪心算法	DDPG	PPO	ESERDQN
探索效率	低	极低	中	高	高（自适应性）
策略稳定性	—	低	低	高	高
离散动作优化	适用	适用	不适用	适用	最优
训练速度	极慢	慢	快	中	快
超参数敏感性	低	低	高	中	中

假设数据流通工程中每次任务类型和数据处理类型一致，而数据处理范围不同[29]，并参考标准性能评估组织（Standard Performance Evaluation Corporation，SPEC）对仿真设备相关参数参考当前已有文献[16]进行设置，训练关键参数详情见表 3。

仿真实验系统单次的数据服务能耗和成本参数设定的仿真系统能耗和成本参数见表 4，其中单位传输表示一个数据段在流通过程中传输所需的能耗和成本。

在动作空间设置方面，设定整个数据传输工程中的数据服务能做出动作空间范围为 12%～100%，各动作之间间隔 1% 进行均匀分布，具体到系统中就是要保证有 12% 的数据保存在本地边缘平台资源池中。在任务请求设置方面，设定任务请求所处理的数据类型和数据体量都相同，数据在访问范围为 $(0,100]$ MB、任务处理时延范围在 $[0.01,0.25]$ s 间均匀分布。在计算时延方面，仿真系统采用 Kubernetes 平台，它基于可信执行环境为应用提供高效的数据服务[30]。在权重设置方面，设定 $\alpha=2,\beta=1,\xi=1$。由于云平台跟本地边缘平台相比，数据传输总量较大、数据处理过程（涉及加、解密等）较复杂，所以将仿真系统时延参数设置见表 5。

表 3　训练关键参数详情

参数名称	参数值
熵正则化系数 λ	0.1
经验回放清空周期 N	1 000
目标网络更新间隔 C	100
折扣因子 γ	0.99
默认通信带宽	2 MHz
算法更新步长 L	0.25
ε-greedy 随机探索概率 κ	0.8
神经网络层数	[50,130,13]
经验池大小 F	2 000
Critic 网络架构	[1 024,512,256,1]
Actor 网络架构	[1 024,512,256,2×N_S×N_{RF}]
Critic 网络学习率 δ	0.001
Actor 网络学习率 ε	0.000 1
PPO 算法学习率 η	0.000 1
ESERDQN 算法学习率	0.000 1
每周期抽取四元组数量 B	20

表 4　仿真系统能耗和成本参数

参数	本地	云端	单位传输
能耗/10^{-7}J	[0.05,0.06]	[0.025,0.035]	0.015
成本/元	[0.06,0.09]	[0.02,0.03]	0.015

表 5　仿真系统时延参数设置

参数	数据范围/MB	本地	云端	单位传输
时延/s	(0,18]	0.015	0.015	0.001 5
	(18,66]		[0.025,0.035]	0.001 5
	(66,100]		[0.2,0.3]	0.001 5

4.2 实验结果与分析

本研究利用Google Cluster Trace数据集模拟数据对象的访问模式。为验证所提算法的泛化能力和分类效率，实验设计分为2个阶段：首先基于数据集中随机抽取的1 000个数据对象样本对深度神经网络模型进行训练，随后采用独立测试集评估训练后模型的性能表现。用于模型训练的硬件平台配备了3个CPU（每个CPU有24个内核和32 GB内存），2个GPU（每个GPU有32 GB内存）。

仿真实验对随机算法、贪心算法、DDPG、PPO和本文提出的改进算法ESERDQN在仿真系统中进行执行训练，设定训练进行了10 000次，训练过程中各算法收敛速率与奖励对比如图5所示，为仿真系统中单次数据服务的系统奖励（由式（8）得出）变化详细情况。由于随机算法、贪心算法不涉及策略的训练过程，因此两个算法系统整体奖励较低，贪心算法由于贪心策略的取值为局部最优，扩大到全局系统中，对数据传输、处理产生的系统能耗、成本和时延等相对于随机算法几乎没有改善；DDPG算法采用异策略和经验回放提升数据复用率，但需连续动作映射，且依赖目标网络和经验回放，训练波动较大，在数据传输这样的离散任务中，算法收敛速率低于改进算法ESERDQN；PPO算法支持连续和离散动作，泛化性更强，但同策略需大量新数据，且依赖复杂策略优化，计算效率相较于改进算法ESERDQN较低；改进算法ESERDQN在传统的DQN算法基础上，引入了基于熵的动作筛选和同策略经验回放思想，熵正则化减小局部最优风险，同策略降低方差，对筛选出的优良动作进行训练和学习，尽可能快地找到当前数据分类周期的最佳分类策略，在系统整体奖励中表现最优。在算法执行初期，由于经验较少（<800次），训练效果较差，系统奖励和收敛速率DDPG、PPO和ESERDQN跟前2种基线算法相比较几乎没有差别。执行中期，DDPG算法训练次数在1 800~2 000时，系统整体性能迅速上升，训练次数>2 150时，系统趋于收敛，训练次数>2 330时，系统趋于稳定；PPO算法训练次数在1 500~1 800时，系统整体性能迅速上升，训练次数>1 700时，系统趋于收敛，训练次数>2 200时，系统趋于稳定；改进算法ESERDQN训练次数在800~1 500时，系统整体性能迅速上升，训练次数>1 500时，系统趋于收敛，训练次数>2 000时，系统趋于稳定，系统收敛速率相较于贪心算法、随机算法、DDPG算法和PPO分别提升49.2%、41.7%、30.1%和13.3%。可见，在数据传输这类离散任务中，改进算法ESERDQN在系统收敛速率和稳定性都具有技术优越性。

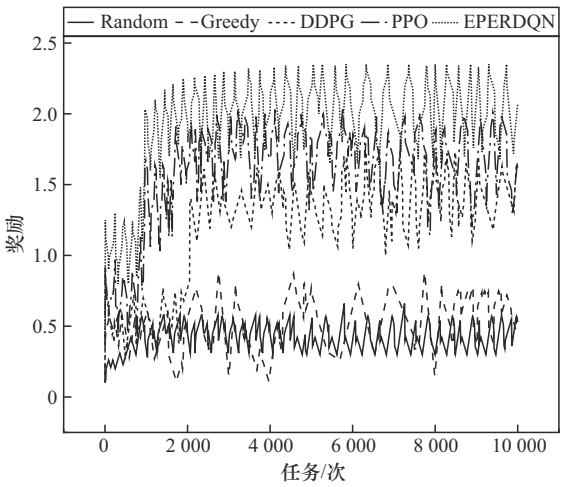


图5 训练过程中各算法收敛速率与奖励对比

对基于改进算法ESERDQN和其他4种算法模型生成的策略进行测试（1 000次），系统的累积回报由$\sum_{i=1}\text{Reward}_i$可得，5种算法的结果对比，仿真系统累积回报对比如图6所示。基于改进算法ESERDQN的模型由于其不仅由式（3）～式（5）定义了系统的能耗、成本和时延指标，还在数据服务执行的环境空间设计时，用式（6）综合考虑了任务队列和其他待处理任务的性质，同时用

神经网络和策略学习的方法提高了策略的自我感知和自我学习能力，因此所得的系统累积回报是最高的，且长期（>1 000 次）奖励回报明显优于其他 4 种算法。

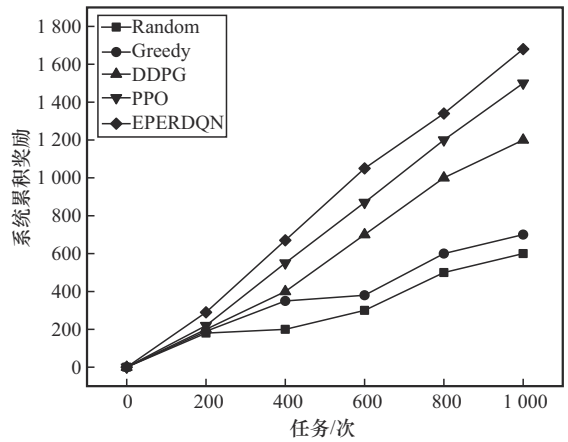


图 6　仿真系统累积回报对比

仿真系统累积回报对比如图 7 所示、仿真系统累计能耗回报对比如图 8 所示、仿真系统累积成本回报对比如图 9 所示，这些是随机选取某次测试的系统累积时延、能耗和成本回报的对比详情，可以看出基于改进算法 ESERDQN 的模型在各项表现中都要优于其他 4 种算法。随机算法随机选取 1 个阈值进行数据分类，且它没有策略训练过程，对系统性能不会造成额外的负担；贪心算法执行过程中选取的贪心策略为局部最优，仅考虑数据处理时延，而没有考虑数据传输时延，虽然没有策略训练过程，但没有综合考虑整个数据传输过程中的能耗、时延和成本，因此贪心算法在能耗、成本和时延累积回报表现最差；改进算法 ESERDQN 相较于 DDPG、PPO 算法，基于熵的动作筛选和同策略经验回放对动作空间的筛选和经验的回放方法进行了优化，根据熵来选取更加精确的最佳阈值空间，提高了算法的收敛速率和计算效率，综合系统时延、能耗、成本指标，实现多目标均衡，在系统累积长期回报方面表现最优。

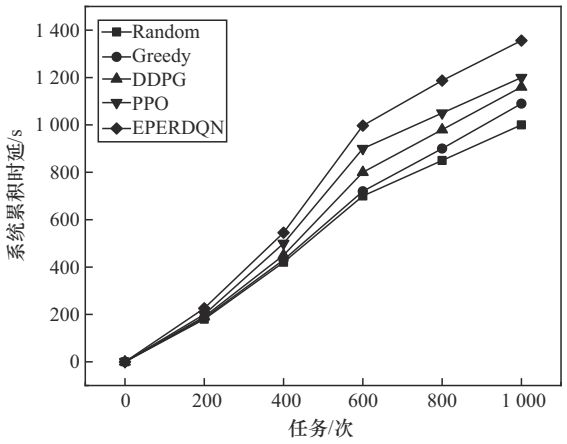


图 7　仿真系统累积时延回报对比

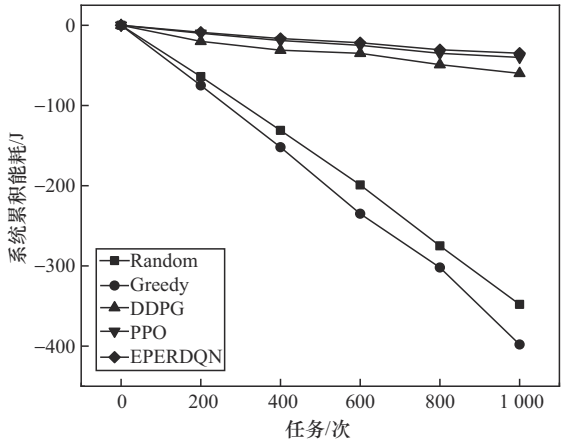


图 8　仿真系统累计能耗回报对比

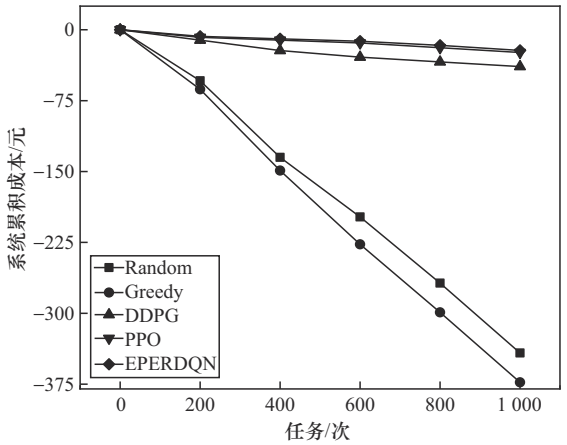


图 9　仿真系统累积成本回报对比

5　结束语

智慧城市作为现代城市治理体系革新的重

要载体，其建设进程直接关联城市治理效能提升、民生服务优化及区域可持续发展能力。在智慧城市多维架构中，数据交互系统承担着关键支撑功能，不仅需要构建数据全生命周期安全防护机制，还需要在复杂动态环境下实现能耗效率、传输时延及经济成本等多目标协同优化。针对上述挑战，本文构建了基于深度强化学习的动态决策框架，将异构网络环境下的数据传输优化问题建模为具有连续状态表征的MDP。通过融合熵正则化约束与同策略经验回放机制，提出改进型ESERDQN算法，建立多维业务指标（能量消耗、传输时延、经济成本）的联合优化模型。仿真实验表明，该方案较传统方法在系统综合效能方面提升23.6%，时延敏感型业务的服务质量达标率提高19.8%，验证了所提方法在增强城市智慧治理能力方面的有效性与工程适用性。

在实际应用场景中，本文所构建模型可在离散性任务应用场景中进行落地推广，例如，智能交通管理中，摄像头实时视频流等热数据通过边缘节点经Kafka技术加密传输至本地资源池以供交通部门实时分析拥堵情况[31]；过车记录等冷数据每日批量压缩上传至云资源池内供交通部门进行季度报告生成；在市民健康档案管理中，医院急诊记录等热数据经量子加密传输到本地内存数据库供急救中心实时调取；10年前病历等冷数据脱敏后归档至Glacier进行存储。同时，为提高模型在超大规模数据集中的泛化能力，笔者未来将重点聚焦于以下3个方向：首先，针对现有方法进行扩展性优化，提升算法在复杂场景下的适应能力；其次，开展与工业级生产系统的集成应用研究，实现从理论方法到实际工程落地的技术转化[32]。具体而言，计划通过架构重构和参数调优增强模型的泛化性能，同时建立标准化接口协议以实现与企业现有基础设施的无缝对接。最后，利用联邦学习采用分布式模型训练方式，原始数据始终保留在本地设备，仅传输模型参数或梯度更新，避免直接共享敏感数据的优势，对深度强化学习、联邦学习两种技术结合在数据传输领域进行综合性研究，提升训练模型中数据的隐私性和安全性。以期这一系列改进能提升所提解决方案在真实业务环境中的实用价值。

参考文献：

[1] GUO Z J. The application process of artificial intelligence in smart cities[J]. Automation and Instrumentation, 2024, 39(9): 162-164.

[2] JIA X F, GAO S, ZHOU Y, et al. An efficient cross-domain data flow technology framework for mega-city governance [J]. Frontiers in Data and Computing Development, 2023, 5(5): 35-45.

[3] 王亚平, 余贶琭, 滕永平. 面向智慧交通的物联网实验教学探索[J]. 实验室研究与探索, 2024, 43(1): 184-187.
WANG Y P, YU K L, TENG Y P. Exploration on the experimental teaching of the Internet of things for intelligent transportation[J]. Research and Exploration in Laboratory, 2024, 43(1): 184-187.

[4] ZHOU C Z, WU W, CAI X Q. Research on data security prevention and control strategies based on classification and classification[J]. Frontiers in Data and Computing, 2023, 5(1): 128-135.

[5] CHEN K, CHEN L, XIE J M, et al. Simulation research on adaptive signal control of deformed intersection based on LSTM-GNN[J]. Journal of System Simulation. 2025, 37 (6): 1343-1351.

[6] YU H M, LIU W, MENG L L, et al. RWK-GNN: non-equilibrium graph fraud detection algorithm based on feature enhancement and subkernel decomposition[J]. Acta Electronica Sinica, 2013, 52(10): 3382-3391.

[7] NUNES B A A, MENDONCA M, NGUYEN X N, et al. A survey of software-defined networking: past, present, and future of programmable networks[J]. IEEE Communications Surveys & Tutorials, 2014, 16(3): 1617-1634.

[8] 蒋莹莹. 移动边缘网络中计算卸载优化策略研究[D]. 武汉:

华中科技大学, 2022.
JIANG Y Y. Research on computing offload optimization strategy in mobile edge networks[D]. Wuhan: Huazhong University of Science and Technology, 2022.

[9] XU X L, FANG Z J, QI L Y, et al. Distributed service unloading method based on deep reinforcement learning in the edge computing environment of vehicle networking[J]. Journal of Computer Science, 2019, 44(12): 2382-2405.

[10] 刘智铭. 基于深度强化学习的物联网数据汇聚及分发机制研究[D]. 北京：北京邮电大学, 2024.
LIU Z M. Research on data aggregation and distribution mechanism of Internet of things based on deep reinforcement learning [D]. Beijing: Beijing University of Posts and Telecommunications, 2024.

[11] 孙翔宇. 离线数据驱动的深度强化学习算法研究[D]. 成都: 电子科技大学, 2024.
SUN X Y. Research on offline data-driven deep reinforcement learning algorithm[D]. Chengdu: University of Electronic Science and Technology of China, 2024.

[12] 于会涵. 基于深度强化学习的网络数据传输优化关键技术研究[D]. 北京: 北京邮电大学, 2024.
YU H H. Research on key technologies of network data transmission optimization based on deep reinforcement learning[D]. Beijing: Beijing University of Posts and Telecommunications, 2024.

[13] 唐珩膑. 基于深度强化学习的异构网络缓存策略与数据传输研究[D]. 广州：华南理工大学, 2022.
TANG H B. Research on caching strategy and data transmission in heterogeneous network based on deep reinforcement learning[D]. Guangzhou: South China University of Technology, 2022.

[14] 左亚兵, 王凯, 杨帆, 等. 基于用户偏好的协作内容缓存策略[J]. 计算机应用研究, 2022, 39(1): 123-127.
ZUO Y B, WANG K, YANG F, et al. Cooperative content caching strategy based on user preference[J]. Application Research of Computers, 2022, 39(1): 123-127.

[15] BEN-AMEUR A, ARALDO A, CHAHED T. Cache allocation in multi-tenant edge computing *via* online reinforcement learning[C]//Proceedings of the ICC 2022 - IEEE International Conference on Communications. Piscataway: IEEE Press, 2022: 859-864.

[16] SHEN L J, QIU S Q, CUI C,et al. Research on efficient data circulation strategy under "east number and west calculation" [J]. Frontiers in Data and Computing, 2023, 5(5): 3-12.

[17] CETOLA S. A method for comparative analysis of trusted execution environments[D]. Portland: Portland State University, 2021.

[18] ZHANG X J, KANG Y, CHEN K, et al. Trading off privacy, utility and efficiency in federated learning[EB]. 2022: 2209.00230.

[19] SERGEI A, BOHDAN T, FRANZ G, et al. SCONE: secure linux containers with intel SGX[C]//Proceedings of the 12th USENIX conference on Operating Systems Design and Implementation (OSDI'16), USENIX Association, USA, 2016: 689-703.

[20] SUZAKI K, NAKAJIMA K, OI T, et al. TS-perf: general performance measurement of trusted execution environment and rich execution environment on intel SGX, arm TrustZone, and RISC-V keystone[J]. IEEE Access, 2021, 9: 133520-133530.

[21] ALE L H, ZHANG N, FANG X J, et al. Delay-aware and energy-efficient computation offloading in mobile-edge computing using deep reinforcement learning[J]. IEEE Transactions on Cognitive Communications and Networking, 2021, 7(3): 881-892.

[22] LI M S, GAO J, ZHAO L, et al. Deep reinforcement learning for collaborative edge computing in vehicular networks[J]. IEEE Transactions on Cognitive Communications and Networking, 2020, 6(4): 1122-1135.

[23] 蒋守花, 王以伍. SDCN 中基于深度强化学习的移动边缘计算任务卸载算法研究[J]. 电信科学, 2024, 40(2): 96-106.
JIANG S H, WANG Y W. Research on task unloading algorithm of moving edge computing based on deep reinforcement learning in SDCN[J]. Telecommunications Science, 2024, 40(2): 96-106.

[24] ZHAN W H, LUO C B, MIN G Y, et al. Mobility-aware multi-user offloading optimization for mobile edge computing[J]. IEEE Transactions on Vehicular Technology, 2020, 69(3): 3341-3356.

[25] ALFAKIH T, HASSAN M M, GUMAEI A, et al. Task offloading and resource allocation for mobile edge computing by deep reinforcement learning based on SARSA[J]. IEEE Access, 2020, 8: 54074-54084.

[26] LI D J，XU S Y，LI P Y. Deep reinforcement learning-

empowered resource allocation for mobile edge computing in cellular V2X networks [J]. Sensors, 2021, 21(2): 372.

[27] ABADI A, AGARWAL A, BARHAM P, et al. Tensor Flow: largescale machine learning on heterogeneous distributed systems[EB]. 2021.

[28] CHEN F, WU L X, WANG M, et al. A low-carbon optimization scheduling method of CIES based on PPO algorithm[J]. Electric Power Engineering Technology, 2024, 43(6): 88-99.

[29] MAO Y Y, ZHANG J, SONG S H, et al. Stochastic joint radio and computational resource management for multi-user mobile-edge computing systems[J]. IEEE Transactions on Wireless Communications, 2017, 16(9): 5994-6009.

[30] 张燕, 杨一帆, 伊人, 等. 隐私计算场景下数据质量治理探索与实践[J]. 大数据, 2022, 8(5): 55-73.
ZHANG Y, YANG Y F, YI R, et al. Exploration and practice of data quality governance in privacy computing scenarios[J]. Big Data Research, 2022, 8(5): 55-73.

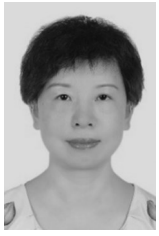
[31] LILLICRAP T P, HUNT J J, PRITZEL A, et al. Continuous control with deep reinforcement learning[EB]. 2015: 1509.02971.

[32] JOHN S, SERGEY L, PHILIPP M, et al. Trust region policy optimization[C]//Proceedings of the 32nd International Conference on International Conference on Machine Learning. [S.l.:s.n.], 2015, 37(ICML'15): 1889-1897.

[作者简介]

蒋守花（1988– ），女，成都医学院工程师，主要研究方向为边缘计算、人工智能、物联网、大数据。

冯军（1969– ），男，成都医学院教授，主要研究方向为5G双域网、医疗设备管理、教育信息技术应用、人工智能。



舒晖（1972– ），女，成都医学院高级工程师，主要研究方向为教育信息技术应用、5G双域网、物联网。

黎佳宜（2004– ），女，北京师范大学在读，主要研究方向为国际与比较教育、数据传输、新媒体传播等。