文章编号:1672-3961(2025)03-0034-12 DOI:10.6040/j.issn.1672-3961.0.2024.065

基于分层多智能体强化学习的个性化与信号控制联合路径引导方法

高君健,廖祝华*,刘毅志,赵肄江
(湖南科技大学计算机科学与工程学院, 湖南 湘潭 411201)

摘要:为进一步缓解交通拥堵、提高道路通行能力,本研究基于分层多智能体强化学习提出一种联合个性化引导和交通信号控制的城市车辆路径引导方法:在交叉路口放置路径引导智能体和信号控制智能体,用于提供个性化路径引导策略和优化信号灯控制,平衡城市交通流量。为了克服预定义的图结构在表示动态交通状态特征时的局限性,信号控制智能体使用自适应图卷积网络挖掘同层次智能体间空间相关性;路径引导智能体结合平均场博弈,分析车辆平均动作以有效捕捉车辆之间的交互作用,实现车辆之间协调,并根据车辆的目的地为车辆提供个性化路径引导策略;为预防局部交通拥堵和交通严重不平衡,基于 MAPPO(multi-agent proximal policy optimization)算法,通过集中式训练和分布式执行实现信号控制智能体之间的合作,以实现路径引导中方向的限流;基于分层强化学习方法,实现异质智能体之间信息的共享、交流以促进它们之间的协作。为验证本研究方法的效果,基于多种真实的开源交通数据集,在 SUMO 仿真平台上进行试验,并与多种基线方法进行比较。结果表明,本研究所提方法将车辆的平均行程时间最少缩短 11.05%,平均延误时间最少减少 19.90%,有效地提高了城市车辆通行效率。

关键词:强化学习;路径引导;信号控制;平均场博弈;自适应图卷积
中图分类号:U121 **文献标志码**:A
引用格式:高君健,廖祝华,刘毅志,等. 基于分层多智能体强化学习的个性化与信号控制联合路径引导方法[J]. 山东大学学报(工学版),2025,55(3):34-45.
GAO Junjian, LIAO Zhuhua, LIU Yizhi, et al. Hierarchical multi-agent reinforcement learning based route guidance method combining personalization and signal control[J]. Journal of Shandong University (Engineering Science), 2025, 55(3):34-45.

Hierarchical multi-agent reinforcement learning based route guidance method combining personalization and signal control

GAO Junjian, LIAO Zhuhua*, LIU Yizhi, ZHAO Yijiang
(School of Computer Science and Engineering, Hunan University of Science and Technology, Xiangtan 411201, Hunan, China)

Abstract: To further alleviate traffic congestion and improve road network efficiency, this study proposed an urban vehicle route guidance method integrating personalized routing strategies and traffic signal control based on hierarchical multi-agent reinforcement learning (MARL). Route guidance agents and traffic signal control agents were deployed at intersections to provide personalized routing policies and optimize traffic light control, thereby balancing urban traffic flow. To overcome the limitations of predefined graph structures in representing dynamic traffic state features, the traffic signal control agents employed an adaptive graph convolutional network to autonomously capture spatial correlations among peer agents. Concurrently, the route guidance agents integrated meanfield game to analyze aggregated vehicle actions, effectively capturing inter-vehicle interactions for coordinated decision-making while delivering destination-specific routing strategies. To prevent local congestion and severe traffic imbalance, a multi-agent proximal policy optimization (MAPPO) algorithm was adopted, enabling centralized training and decentralized execution for cooperative signal control agents to implement directional flow restriction. A hierarchical reinforcement learning

收稿日期:2024-04-02
基金项目:湖南省自然科学基金资助项目(2024JJ5163)
第一作者简介:高君健(1999—),男,湖南益阳人,硕士研究生,主要研究方向为智慧交通。E-mail:junjiangao@hnust.edu.cn
*通信作者简介:廖祝华(1977—),男,湖南株洲人,副教授,硕士生导师,博士,主要研究方向为数据挖掘、智慧交通、分布式计算。
E-mail:zhliao@hnust.edu.cn

framework facilitated information sharing and collaboration among heterogeneous agents. Extensive experiments were conducted on the SUMO simulation platform using multiple real-world open-source traffic datasets, with comparisons against baseline methods. Results demonstrated that the proposed method reduced average travel time by at least 11.05% and decreased average delay time by at least 19.90%, significantly enhancing urban traffic efficiency.

Keywords: reinforcement learning; route guidance; signal control; mean field game; adaptive graph convolution

0　引言

城市交通拥堵已成为全球城市普遍面临的问题,随着智能交通系统的发展,通过车辆路径引导与交通信号控制的高效协同,有望缓解这一问题。传统的路径引导策略往往依赖于静态信息(如路线距离、历史经验)或追求距离最短、成本最低的目标,忽视了城市车流的动态特性及司机间路径选择的相互影响,可能导致大量车辆集中在少数路径上,形成拥堵博弈[1]。此外,这些方法通常未充分考虑交通信号对通行成本和流量分布的调节作用,从而限制了交通效率的提升。对于交通管理者而言,智能交通信号灯是一种经济且高效的交通调节工具,它能根据历史和实时交通流量数据,动态调整信号相位或配时[2],以智能方式控制交通流、优化道路利用率,从而减轻拥堵。在全局流量未饱和的情况下,为实现城市交通的整体平衡,达到城市交通畅通,有必要引入多决策者的策略交互与博弈,如通过协调司机的路径选择行为与智能信号控制策略,提升城市交通整体通行效率,缓解交通拥堵。

随着人工智能与智能交通系统的发展,数据获取变得方便、快捷,推动了基于数据驱动的深度强化学习方法在车辆路径引导与交通信号控制领域的广泛研究。这些方法使智能体能够在与环境的持续交互中学习到当前交通状态下的最优策略,成为优化路网运行状态、提升通行效率的有效手段。尽管深度强化学习在路径引导[3]和智能信号控制[4]研究中展现出提升路网效率的潜力,但多数现有研究往往忽略了交通系统中交通信号控制与驾驶员路线选择间的内在联系。路径引导通过调整流量分布缓解拥堵,信号控制通过优化信号相位和配时影响通行时间与交通流分布,两者相互影响,路径引导与信号控制却在多数研究中被孤立处理。多智能体强化学习(multi-agent reinforcement learning, MARL)为智能交通系统中各组成部分(如车辆、信号灯)的协作提供了可能性。这种方法通过建模车辆间的相互作用,利用其学习能力,为每辆车提供个性化的路径选择建议。特别在交通状态严重不平衡或局部拥堵时,多智能体强化学习能够通过智能信号控制引导车辆有序通行,从而迅速恢复局部交通秩序。

为实现上述目标,除了要充分利用现有的城市交通资源,还需设计合理的实现方法和目标函数,以达到智能引导车辆选择合适路线并对信号灯进行智能控制的目的,同时兼顾全局交通平衡与节省司机出行时间。受 FeUdal 分层强化学习架构[5]的启发,本研究提出基于分层多智能体强化学习的个性化与信号控制联合路径引导方法。在该方法中,路径引导智能体(route guidance agent, RGA)和信号控制智能体(signal control agent, SCA)均放置在交叉路口,SCA 作为管理者(Manager)在上层负责信号控制,RGA 作为工作者(Worker)在下层负责在交叉路口引导各车辆进入下个路段。SCA 和 RGA 都与环境交互,并且通过信息交流的方式相关联,从而促进 SCA 和 RGA 的协作,以达到个性化与信号控制联合的路径引导。

传统的路径规划方法通常基于道路的地理信息或历史行驶数据进行决策[6],然而,车辆的路径选择应当是一个适应不断变化流量条件的序列决策过程[7]。一些研究提出了基于强化学习的实时车辆路径引导方法,文献[8]为随机时变网络中的自适应路径引导问题设计了一种基于 Q 学习的强化学习框架,并使用基于树的函数逼近方法提高算法的效率和准确性;文献[9]在 SUMO 仿真器中验证了深度 Q 学习算法在路径规划问题中的有效性。但这些方法主要基于单智能体强化学习,忽略了智能体之间的相互影响和协调。文献[3]基于多智能体强化学习提出一种 $A*R^2$ 城市交通路径引导方法,具有较高的自适应学习能力;文献[10]以 Q-routing算法为基础,设计一种多智能体强化学习方法帮助车辆进行路径引导;文献[11]将智能体(如自动驾驶车辆或智能导航系统)在交通网络中的路径选择行为和系统的均衡过程转化为多智能体强化学习问题,并设计一种基于平均场的多智能体深度 Q 学习方法,用于捕捉智能体之间的竞争。

近年来,深度强化学习算法广泛用于交通信号控制的研究,与依赖于人工设计的规则或预定义的交通流模型的传统启发式方法不同,这种方法避免了预先定义的假设,其中每个交通信号控制器视为一个智能体,并直接与环境交互学习交通信号控制策略。文献[4]基于独立 DQN(deep Q-network)方法,使用丰富的状态特征作为输入;文献[12]通过深度 Q 学习以分布式方式控制每个交叉点,它们基于信号灯之间的空间结构构建了信号灯邻接图,然后通过递归神经网络将历史交通记录与当前交通状态整合在一起;受传统方法 MaxPressure[13] 的启发,文献[14]提出基于 DQN 的 PressLight 方法,使用最大压力作为输入特征和奖励,证明了最大化交通网络的吞吐量,即最小化整个网络的运行时间,同时比 MaxPressure 更好地提高了交通效率。为了进一步促进信号灯智能体之间的合作,有学者通过建立通信机制,以获得邻居的信息,文献[15]开发一种多智能体 A2C 算法,该算法通过与周围智能体的通信增强了观察力,并引入空间折扣因子,简化学习过程;文献[16]使用通过图卷积网络计算的邻居潜在状态,通过与相邻智能体的通信实现多路口信号控制的协同。虽然强化学习在优化信号控制和车辆路线方面得到了广泛应用,但联合车辆路径引导和信号灯控制的研究仍然较少,文献[17]通过共同优化交通信号和在瓶颈路口为联网和自动驾驶汽车重新确定最佳路径解决该问题,但未考虑大型交通网络中多个交通瓶颈之间的协同优化;文献[18]开发了一种分层 RL 框架,用于协调交通信号控制和自动驾驶汽车的转向,从而为单个车辆提供了自适应的路线引导,但是忽略了车辆之间的相互影响。

综上,尽管车辆路径引导和交通信号控制的研究已取得一定进展,但现有方法在处理城市交通系统的动态性和复杂性方面仍存在不足。特别是在多智能体系统的协同控制方面,现有研究往往忽视了智能体之间的相互作用和协调,以及车辆路径引导与交通信号控制之间的内在联系。本研究提出的方法不仅考虑了车辆之间的交互作用,还通过分层多智能体强化学习,深入挖掘复杂交通环境的动态特性,从而实现车辆个体路径引导优化以及整体流量的平衡。

1 相关定义

本研究对相关交通元素进行定义。

1.1 车道

车道 $l \in L$,分为车辆进入交叉路口 i 的驶入车道 L_i^{in} 和驶出交叉口的驶出车道 L_i^{out}。

1.2 路段转换

路段转换是指路网/交叉口内的车辆从驶入车道行驶到驶出车道,常见的路段转换为左转、右转、直行。

1.3 信号相位

指一组交通信号的组合,绿色信号和红色信号分别代表允许或禁止车辆进行相应的路段转换,绿色信号设有最短持续时间限制,确保车辆有足够的时间通过路口;黄色信号通常作为绿色与红色信号之间的过渡期,提醒驾驶员准备停车,以确保交通安全,防止因信号切换引发的冲突。信号相位描述了某一个时刻可以同时进行的路段转换行为。

1.4 交叉口有效范围[19]

指车辆在 SCA 的动作持续时间 t_{dur} 内可以通过的到交叉口的最大距离,表示为

$$L_{\text{ran}} = V_{\max} \times t_{\text{dur}}, \tag{1}$$

式中 $V_{\max}$ 为道路最大限速。

1.5 交通拥堵指数(traffic congestion index, TCI)

综合考虑道路上的车辆平均速度和车辆密度,本研究使用 TCI 衡量道路交通状况,具体计算公式为

$$V_l(t) = 1 - \frac{\bar{V}_l(t)}{V_{\max,l}}, \tag{2}$$

$$C_l(t) = \frac{N_l(t)}{C_{\max,l}}, \tag{3}$$

$$I_{\text{TCI},l} = w_1 \cdot V_l(t) + w_2 \cdot C_l(t), \tag{4}$$

式中:l 表示一条车道;$\bar{V}_l(t)$ 为 t 时刻道路上车辆的平均速度;$V_{\max,l}$ 为车道 l 的最大限速;N_l 为 t 时刻道路上的车辆数量;$C_{\max,l}$ 为车道 l 的最大容量;w_1 和 w_2 反映不同元素的重要性,$w_1 + w_2 = 1$,本研究中 w_1、w_2 均设置为 0.5。

1.6 信号控制

本研究采用相位选择作为信号控制策略的动作,SCA 在每个信号灯控制步骤中为每个交叉口选择一个特定的相位,如图 1 所示。

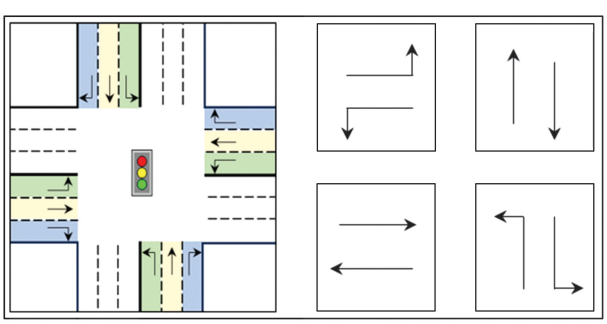


图 1 交叉路口和信号灯相位的示意图

Fig.1 Illustration of an intersection and signal phases

1.7　路径引导

RGA 为驶向交叉路口的车辆根据其目的地和实时交通状况提供下一个驶出路段选择建议,如图 2 所示。对于不同驶入方向的车辆,有对应可能采取的路段转换行为,RGA 需要对当前车辆所在道路不可选的方向进行屏蔽。

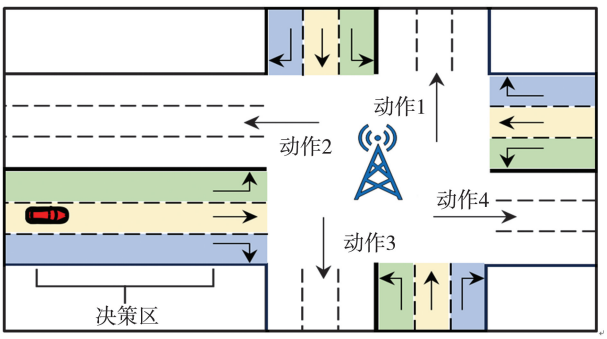


图 2　交叉路口路径引导示意图

Fig.2　Illustration of an intersection and route guidance

1.8　决策区域

当车辆到达车道的决策区时,RGA 将从动作集合中选择动作,动作对应了车辆的驶出路段。

2　基于分层多智能体强化学习的个性化与信号控制联合的路径引导

本研究设计了一种基于分层多智能体强化学习个性化与信号控制联合的路径引导(hierarchical multi-agent reinforcement learning based route guidance combining personalization and signal control, H-RGSC)框架,框架图如图 3 所示。H-RGSC 借鉴了 FuN[15](feudal networks)的双层架构,SCA 扮演上层 Manager 角色,负责宏观层面的交通流控制,其核心任务是通过控制相位给交叉口各个驶入车道的交通流分配通行权,在上层引导车辆有序通过交叉路口,实现对整个区域交通流量的均衡分配与疏导。SCA 以较慢的时间分辨率(如每几个时间步)输出动作,原因是信号控制的决策通常不需要过于频繁的变化。RGA 作为下层 Worker 智能体,直接与道路车辆交互,根据实时交通状态和车辆目的地提供个性化路径引导,旨在微观层面上优化车辆的行驶路径,避免局部拥堵。最初 FuN 只适用于同质任务,Manager 输出一个目标向量 $\boldsymbol{g}$,$\boldsymbol{g}$ 并不直接在环境中执行,而是反馈给 Worker。Worker 根据 $\boldsymbol{g}$ 计算内在奖励并训练。本研究中 Manager 和 Worker 是异质智能体,分别专注于 2 个不同的任务,具有不同的观察和行动空间。因此,本研究设计了双向信息传递机制,有助于实现双向通信和协同训练。高层策略可以根据底层策略的嵌入调整决策,底层策略的训练也可以受益于高层策略的反馈。

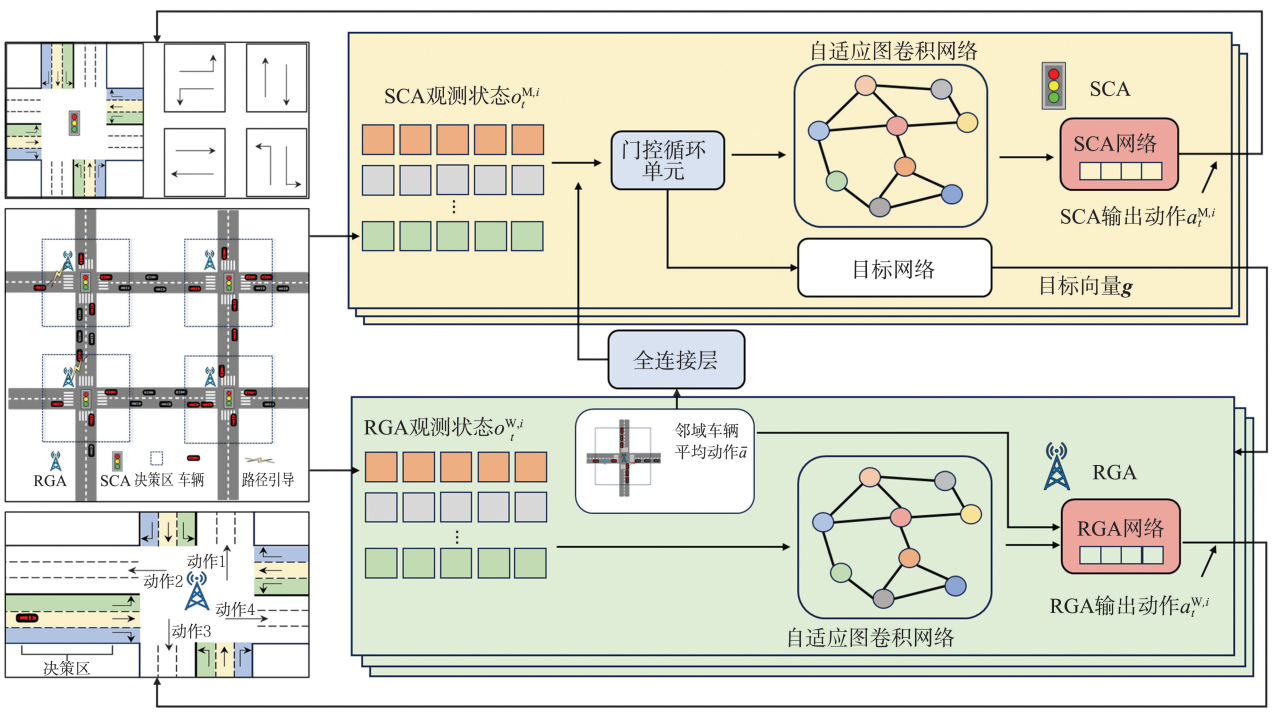


图 3　基于分层多智能体强化学习的个性化与信号控制联合路径引导框架

Fig.3　Hierarchical multi-agent reinforcement learning based route guidance framework combining personalization and signal control

把路径引导任务和信号控制任务建模为部分可观测的马尔可夫决策过程(partially-observable Markov decision process, POMDP)并对相关元素进行定义。

(1)信号控制智能体

o_t^{M}: o_t^{M} 为信号控制智能体在 t 时刻的观测, $o_t^{\mathrm{M}}=(\mathrm{ep}_t(l,m),\ \mathrm{rp}_t(l))$,这里 M 指 Manager,即信号控制智能体,交叉口压力

$$\mathrm{ep}(l,m)=\frac{1}{A}\sum_{i=1}^{A}q(l')-\frac{1}{B}\sum_{j=1}^{B}q(m'),\quad l'\in L^{\mathrm{in}},\ m'\in L^{\mathrm{out}},\tag{5}$$

交叉口有效车辆数量

$$\mathrm{rp}(l)=\sum r(l'),\ l'\in L^{\mathrm{in}},\tag{6}$$

式中,$q(l')$为驶入车道 l'上速度小于 0.1 m/s 的车辆数量,$q(m')$为驶出车道 m'上行驶速度小于 0.1 m/s的车辆数量,$r(l')$为交叉口有效范围内驶入车道的行驶车辆数。A 和 B 分别为 L^{in}和 L^{out}的已用车道数。

a_t^{M}: a_t^{M} 为信号控制智能体在 t 时刻的动作, $a_t^{\mathrm{M}}\in\mathscr{P}$,其中 $\mathscr{P}$为所有可选相位的集合,具体包含的元素根据实际场景决定。当前选择的相位如果与上一个时间步选择的相同,则该相位继续保持;否则,将从上一相位切换至当前决定的相位,并在切换过程中插入一个持续时间为 Δt 的黄灯相位。

r_t^{M}: r_t^{M} 为信号控制智能体在 t 时刻的奖励,定义该交叉口压力的负数$-p$ 作为奖励函数,公式为

$$p=\sum q(l')-\sum q(m'),\ l'\in L^{\mathrm{in}},\ m'\in L^{\mathrm{out}}。\tag{7}$$

(2)路径引导智能体

o_t^{W}: o_t^{W} 为路径引导智能体在 t 时刻的观测, $o_t^{\mathrm{W}}=(q_t,c_t)$,这里 W 指 Worker,即路径引导智能体,$q_t$ 表示路径引导智能体在 t 时刻接收到车辆路由请求;c_t 指交叉口在 t 时刻流出路段交通拥堵指数和交通灯状态(交通灯状态由当前相位以及当前相位剩余时间组成)。

a_t^{W}: a_t^{W} 为路径引导智能体在 t 时刻的动作,a_t^{W}定义为每个交叉口可以选择的所有道路的有限集合。

r_t^{W}: r_t^{W} 为路径引导智能体在 t 时刻的奖励,定义为车辆从进入一个决策区域到进入下一个决策区域所花费时间的负值,$r_t=-\Delta T$。

2.1 基于自适应图卷积网络的智能体空间相关性挖掘

结合图神经网络的 MARL 算法可以很好地学习交通网络的特征表示并学习路径引导和交通信号控制策略。例如,图卷积网络用来捕捉交通网络中不同道路之间的交通特征[20],图注意力网络用来促进智能体之间的消息传递[14]。然而,上述方法需要通过先验知识(空间邻接关系、距离)定义图结构。由于交通流量具有复杂的空间相关性,预定义的图结构难以表示动态变化的交通状态特征,无法完全表示节点之间的相邻关系[21]。为了解决上述问题,本研究采用自适应图卷积网络(adaptive graph convolution network, AGCN)学习的方法,即构造自适应邻接矩阵动态学习交通路网节点的空间相关性,通过自适应图卷积网络聚合动态相邻智能体的信息促进智能体之间的通信,获得每个智能体的观测的增强表示。这种自适应邻接矩阵不需要先验知识定义图结构,它的构造过程是首先引入具有可学习参数的节点嵌入矩阵并对其进行初始化,在随机梯度优化过程中不断调整邻接矩阵。自适应邻接矩阵构造和图卷积的具体计算公式为

$$\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{D}^{-\frac{1}{2}}=\mathrm{softmax}(\mathrm{ReLU}(\boldsymbol{E}_A\cdot\boldsymbol{E}_A^{\mathrm{T}})),\tag{8}$$

$$Z=(\boldsymbol{I}_N+\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{D}^{-\frac{1}{2}})X\Theta+b,\tag{9}$$

式中,$\boldsymbol{D}$ 为度矩阵,$\boldsymbol{A}$ 为邻接矩阵,$\boldsymbol{I}_N$ 为单位矩阵,softmax 指柔性最大激活函数,ReLU 指线性整流激活函数,$\boldsymbol{E}_A$ 为可学习的节点嵌入矩阵,X 和 Z 分别为图卷积层的输入和输出,Θ 和 b 分别为可学习权值和偏置。

2.2 结合平均场博弈的路径引导方法

RGA 网络结构如图 4 所示。

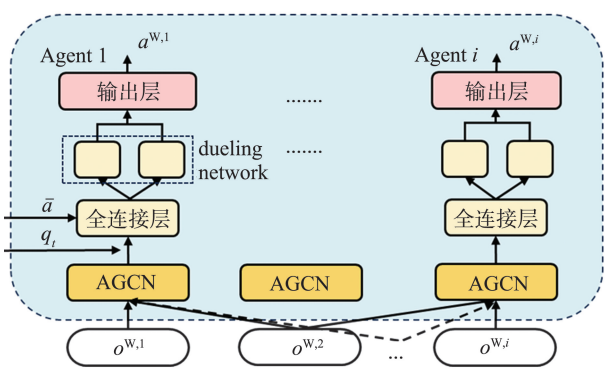


图 4 路径引导智能体网络结构

Fig.4 Network structure of route guidance agent

结合自适应图卷积,挖掘智能体之间的动态关系和交通流量变化,增强智能体的观测状态

$$h_t^{\mathrm{W}}=\mathrm{AGCN}(o_t^{\mathrm{W}})。\tag{10}$$

考虑到车辆路径引导以车辆目的地为导向,RGA 根据车辆的目的地提供个性化的路径引导建议,再连接车辆请求得到观测状态的新表示,即

$$z_t^{\mathrm{W}}=\mathrm{contact}(q_t,h_t^{\mathrm{W}})。\tag{11}$$

车辆按照 RGA 的引导从一个交叉口行驶到另一个交叉口,直到它们到达目的地,类似于 IP 数据包路由过程。Q-routing 是一种将强化学习方法应用于分组路由问题的算法[22],在交通环境下,Q-routing 算法将每个交叉路口视为一个独立的智能体。对于交叉路口智能体 n_i,它需要处理其队列中的第一个车辆,该车辆的目的地路口是 n_d,智能体 n_i 需要选择一个动作 a,表示将该车辆引导到其相邻交叉路口 n_j 作为下一跳。奖励为交叉路口 n_i 和 n_j 之间行驶时间的负值,当交叉路口 n_i 收到奖励 r_i 后,会根据奖励 r_i 更新其 Q 表,公式为

$$Q_{\text{table}-i}(n_d,n_j)=(1-\alpha)Q_{\text{table}-i}(n_d,n_j)+ a[r_i+\max_k Q_{\text{table}-j}(n_j,n_k)], \tag{12}$$

式中:α 为学习速率,$\alpha\in(0,1]$;n_k 是 n_j 的邻居节点。

平均场博弈(mean field game, MFG)理论为研究随机动态博弈提供了强大的框架[23],其核心思想是在适当的意义上用单个个体最优控制问题来近似原始的大群体博弈问题,其中分析了单个个体对平均场(群体的平均行为)的最佳响应。MFG 理论主要有以下特点:(1)参与博弈的个体众多;(2)单个个体对环境的影响都是无限小的。交通环境中车辆路径选择过程有与平均场博弈类似的特征,如一辆车的决策对其他车辆的影响很小。车辆的路径选择仅取决于当前车辆策略和其他车辆策略的分布,无需区分或了解具体的个体车辆。因此,不需要研究所有车辆之间的成对互动,只需要研究单个车辆与整个车辆群体的状态(或者行动)分布之间的交互。

从车辆的角度来看,车辆在到达下一个节点之前,会遵循上一个 RGA 的引导动作继续行驶。假设车辆向 RGA 发送路由请求,由于路网上存在其他车辆,在考虑所有行驶车辆的情况下,当前车辆通过对应 RGA 的动作价值函数 $Q_i(o_t^{\mathrm{W}},a_t^{\mathrm{W}},a_-)$ 得到引导动作 a_t^{W} 进入下一条道路,a_- 表示其他所有车辆的联合动作,对于当前车辆来说,这可能包含冗余信息。在路径引导任务中,远离当前车辆的其他车辆的动作对当前车辆的影响有限。例如,当车辆在一个节点上面临路径选择时,影响因素主要包括当前车辆可选道路上的车辆和将要通过当前交叉口的车辆,a_- 可以近似为附近车辆的动作分布。参考文献[24]中的平均场近似,将这种近似动作信息称为邻域车辆平均动作。在本研究中考虑离散的动作空间,车辆可用的动作空间即智能体 i 的动作可以用 one-hot 的编码方式进行编码。本研究定义邻域车辆平均动作为 $\bar{a}$,即在时间 T 之内,由当前 RGA 指导过的车辆动作分布示意图如图 5 所示,邻域车辆平均动作计算公式为

$$\bar{a}=\frac{1}{N^j}\sum_{\substack{k\in\mathcal{N}(j)\\ t\in T}}a_k, \tag{13}$$

式中:$\mathcal{N}(j)$ 表示在时间 T 之内,由当前 RGA 指导过的车辆集合;N^j 为该集合中车辆总数。

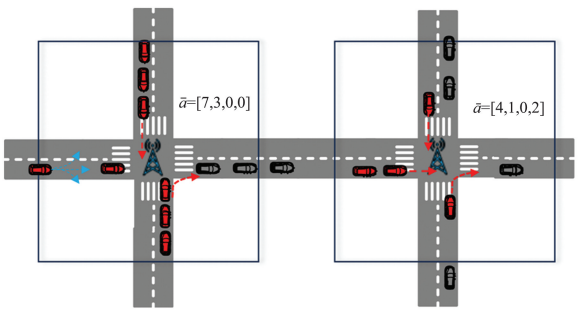


图 5　邻域车辆动作分布示意图

Fig.5　Schematic diagram of vehicle action distribution in the neighborhood

本研究将动作价值函数近似表示为

$$Q^i(o_t^{\mathrm{W}},a_t^{\mathrm{W}},a_-)=\frac{1}{N^j}\sum_{\substack{k\in\mathcal{N}(j)\\ t\in T}}Q^i(o_t^{\mathrm{W}},a_t^{\mathrm{W}},a^k)\approx Q^i(o_t^{\mathrm{W}},a,\bar{a}), \tag{14}$$

RGA 结合 dueling 网络结构[25],稳定训练过程,通过与邻居之间的交互更新最小化损失函数 $\mathscr{L}$ 训练 Q^i 的参数,计算公式为

$$\mathscr{L}(\omega^i)=(y^j-Q^i(o_t^{\mathrm{W}},a_t^{\mathrm{W}},\bar{a};\omega^i)^2, \tag{15}$$

$$y^j=r_t^{\mathrm{W}}+\gamma\max_{a'}Q^j(o_{t+1}^{\mathrm{W}},a_t^{'\mathrm{W}},\bar{a};\omega^j)\times(1-f_j), \tag{16}$$

式中:y^j 为目标 Q 值;Q^j 表示车辆到达的下一个 RGA;γ 为折扣因子;ω 为网络参数;f_j 为车辆是否到达目的地的标志,如果车辆到达目的地,则 $f_j=1$,否则,$f_j=0$。

2.3　基于 MAPPO 算法的信号灯协同控制方法

结合自适应图卷积的 MAPPO(multi-agent proximal policy optimization)框架如图 6 所示。MAPPO[26]基于集中式训练和分散执行框架,每个智能体都有自己的策略网络(actor)和价值网络(critic),参数分别表示为 θ^{a} 和 θ^{c}。在训练阶段允许集中价值网络访问环境的全局状态信息 s 和联合动作 a,而在执行阶段智能体通过分散的策略网络基于局部观测信息执行动作。分布执行可以提高智能体决策的灵活性和扩展性,因为每个智能体只根据自己的本地观察和策略做出决策,不依赖于其他智能体或集中信息。但是,完全分布执行也可能导致策略失去协调性。

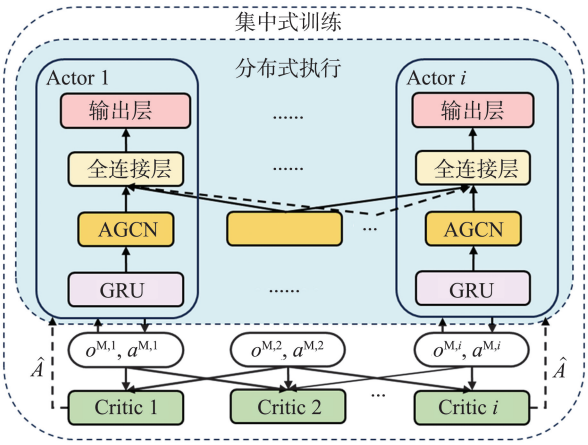


图 6 结合自适应图卷积的 MAPPO 框架示意图

Fig.6 Schematic of the MAPPO framework combined with AGCN

由于交通状态具有随时间变化的动态特性,为有效捕捉其时序依赖关系,本研究使用门控循环单元(gated recurrent unit, GRU)挖掘历史交通信息,将 o_t^{M} 和上一时刻包含历史信息的隐藏状态 h_{t-1}^{M} 共同作为 GRU 的输入,再基于自适应图卷积网络促进智能体间的状态信息融合,公式为

$$h_t^{\mathrm{M}}=\mathrm{AGCN}(\mathrm{GRU}(o_t^{\mathrm{M}},h_{t-1}^{\mathrm{M}})), \tag{17}$$

最后策略网络输出动作作用于环境。价值网络用于估计当前状态,以指导策略网络的更新。其中策略网络与价值网络的损失函数分别为

$$L_t(\theta^{\mathrm{a}})=\min\begin{pmatrix} r_t(\theta^{\mathrm{a}})\hat{A}_t, \\ \mathrm{clip}(r_t(\theta^{\mathrm{a}}),1-\varepsilon,1+\varepsilon)\hat{A}_t \end{pmatrix}, \tag{18}$$

$$L_t(\theta^{\mathrm{c}})=(\hat{A}t+\gamma V^i(s_{t+1})-V^i(s_t)^2, \tag{19}$$

式中:$r_t(\theta^{\mathrm{a}})$ 为新旧策略的概率比,$r_t(\theta^{\mathrm{a}})=\dfrac{\pi_\theta^{\mathrm{a}}(a_t|s_t)}{\pi_{\theta_{\mathrm{old}}}(a_t|s_t)}$;$\varepsilon$ 为控制概率比取值范围的剪裁参数;$\hat{A}_t$ 为优势函数。

PPO[27] 和 MAPPO 算法中一般采用广义优势估计算法(generalized advantage estimation, GAE)计算,公式为

$$\hat{A}_t=\delta_t+(\gamma\lambda)\delta_{t+1}+\cdots+\cdots+(\gamma\lambda)^{T-t+1}\delta_{T-1}, \tag{20}$$

式中,δ_t 为时序误差,$\delta_t=r_t+\gamma V(s_{t+1})-V(s_t)$。

2.4 基于分层强化学习的双向信息传递机制

在每个交叉路口,SCA 和 RGA 的协同工作至关重要。如果交通信号灯控制和路线引导是独立优化的,那么信号控制所考虑的交通流量模式将被路径引导打破,而在路径引导中考虑的交通信号模式也将因动态的交通信号控制变得不稳定。为了解决该问题,本研究通过 2 种异质智能体之间的交流实现协作,如图 3 所示。在整个分层强化学习框架中 SCA 作为 Manager,RGA 作为 Worker,考虑到 Worker 需要被鼓励去完成 Manager 制定的目标,参考文献[5]将内在奖励 $r_t^{\mathrm{W,int}}$ 定义为

$$r_t^{\mathrm{W,int}}=\frac{1}{c}\sum_{i=1}^{c}d_{\cos}(o_t^{\mathrm{W}}-o_{t-1}^{\mathrm{W}},g_{t-1}), \tag{21}$$

式中 $d_{\cos}$ 为余弦相似度。最终的 Worker 奖励函数 r_t^{W} 是从环境中获得的外在奖励 r_t^{ext} 和内在奖励的线性组合,即

$$r_t^{\mathrm{W}}=\alpha r_t^{\mathrm{W,int}}+r_t^{\mathrm{ext}}, \tag{22}$$

式中,α 为内在奖励对 Worker 整体奖励的贡献系数。

交叉路口信号配时方案在路径引导模型中认为是已知输入(在 RGA 的观测中包含了信号灯的信息)。在信号控制与路径引导共同作用于交通环境的条件下,时间和空间上共同影响着交通流,路径引导策略很大程度上影响了下一时刻各路段的交通状况。邻域车辆的平均动作反映了当前交叉口的车辆行驶情况,通过领域车辆平均动作推断 RGA 局部路由策略,公式为

$$m_t^{\mathrm{M}}=\mathrm{FC}(\bar{a}), \tag{23}$$

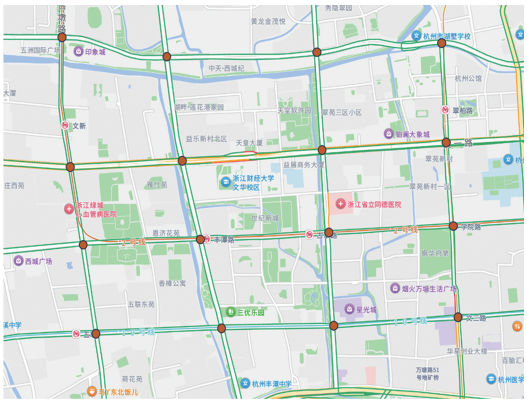
式中,FC 指全连接层,$\bar{a}$ 指邻域车辆平均动作。m_t^{M} 通过与 o_t^{M} 进行拼接操作 contact$(o_t^{\mathrm{M}},m_t^{\mathrm{M}})$ 传递给 Manager。因此,高层策略可以根据底层策略的嵌入调整其决策,底层策略的训练也可以受益于高层策略的反馈。

3 试验

本研究使用开源交通模拟软件 SUMO (simulation of urban mobility)[28] 测试交通引导算法和可视化模拟活动,并利用 TraCI 获取 SUMO 交通仿真环境中的数据。对于试验数据,使用了真实的开源交通流量数据。

杭州 4×4 路网[14]:该路网包含 16 个交叉路口,呈 4×4 网格,如图 7(a)所示。每个交叉路口有 4 条进路和 4 条出路,每条路段包含 3 条车道(分别为左转车道、直通车道和右转车道)。交通流量数据源自杭州的实时摄像头监控数据。

曼哈顿 16×3 路网[29]:该路网包含纽约曼哈顿区的 48 个交叉路口,呈 16×3 网格,如图 7(b)所示。每个交叉路口有 4 条进路和 4 条出路,每条路段包含 3 条车道(分别为左转车道、直通车道和右转车道)。交通流量和转弯率数据均来源于实际的出租车轨迹数据。



(a) 杭州4×4路网

(b) 曼哈顿16×3路网

图 7　数据集路网结构示意图

Fig.7　Schematic diagram of the structure of the dataset road network

3.1　评估指标

交通信号控制任务常用的评价指标包括车辆平均等待时间、平均总停靠次数、总吞吐量、平均延误时间等。对于路径引导任务,通常采用车辆的平均行程时间和平均行程距离等指标评估模型性能。本研究使用 2 个指标评估性能,分别是平均行程时间 t_v、平均延误时间 t_d。其中,t_v 为所有车辆从出发地到目的地的平均行程时间,计算公式为

$$t_v = \frac{1}{N_v}\sum_{i=1}^{N_v}(t_e - t_s), \tag{24}$$

式中,N_v 为车辆总数,t_e 为车辆结束行程的时间,t_s 为车辆开始行程的时间。

t_d 是指所有车辆由于驾驶速度低于理想速度而损失的时间,计算公式为

$$t_d = \frac{1}{N_v}\sum_{i=1}^{N_v}(t_r - t_p), \tag{25}$$

式中:t_p 为单个车辆理想行驶时间,指车辆在完全畅通无阻的情况下,以自身最大速度(或者道路最大限速)从出发地到目的地需要花费的时间;t_r 为车辆从出发地到目的地实际所花费的时间,$t_r = t_e - t_s$。

3.2　对比方法

为了评估 H-RGSC 方法,本研究比较了如下 4 种方法。

GA-DQ-routing:是由 Arasteh 等[10]提出的一种基于网络感知的多智能体强化车辆路径引导方法,该方法结合图注意力网络和 Q-routing 方法提供车辆路径引导。

DQN-routing:是 Mukhutdinov 等[22]通过将路由问题建模为多智能体强化学习问题,并使用深度神经网络表示每个路由智能体作为学习代理,提出适用于通信网络和物流领域(例如交通路线引导)等系统的分布式路由方法。

Colight:结合图注意力网络结合的 MARL 信号灯控制的方法。通过引入图形注意力网络和多目标计算,考虑了周围交叉口对当前交叉口的影响[14]。

MAPPO:一种广泛采用的 MARL 框架,具有行动者-批评架构,利用集中式训练分散式执行和近端策略优化提高训练的稳定性[26]。

3.3　试验结果

为了评估本研究路径引导方法的有效性,在两个数据集上将其与基线方法进行了性能比较,不同评估指标的总体性能如表 1 所示。

表 1　不同方法在两个数据集上的性能比较

Table1　The overall performance of different models on two datasets　　单位:s

数据	方法	平均行程时间	平均延误时间
杭州 4×4 路网	GA-DQ-routing	379.01	167.82
	DQN-routing	382.92	171.65
	Colight	286.34	46.55
	MAPPO	277.74	37.83
	H-RGSC	**254.68**	**37.26**
曼哈顿 16×3 路网	GA-DQ-routing	319.17	220.74
	DQN-routing	322.45	225.86
	Colight	211.30	89.75
	MAPPO	201.88	83.86
	H-RGSC	**154.67**	**60.73**

由表 1 可知,本研究模型 H-RGSC 优于信号灯控制基线和路径引导基线。一方面,与基于强化学习的信号灯控制方法相比,H-RGSC 在平均行程时间和平均延误时间方面能收敛到更低水平。与 Colight 相比,H-RGSC 在杭州 4×4 路网上平均行程时间缩短 11.06%,平均延误时间缩短 19.96%;在曼哈顿 16×3 路网上平均行程时间缩短 26.80%,平均延误时间缩短 32.33%。与 MAPPO 相比,H-RGSC

在杭州 4×4 路网上平均行程时间缩短 8.30%;在曼哈顿 16×3 路网上平均行程时间缩短 23.38%,并且平均延误时间降低 27.58%。另一方面,与基于强化学习的路径引导相比,H-RGSC 在杭州 4×4 路网上比表现最好的 GA-DQ-routing 平均行程时间减少 32.80%,平均延误时间减少77.79%;在曼哈顿 16×3 路网上平均行程时间减少 51.53%,平均延误时间缩短 72.48%。

各方法在杭州 4×4 路网与曼哈顿 16×3 路网环境下不同时刻道路中的交通拥堵指数的热力图如图 8、9 所示,颜色越深表示越拥堵。其中,横轴表示时间,纵轴表示不同的道路。与其他算法相比,本研究方法的交通拥堵指数热力图的显示处于低拥堵指数道路更多,侧面反映出本研究的方法能够较好地将车辆以相对均匀的方式分布在所有道路上,这表明该算法相比其他常用算法,能更充分利用有限道路容量,更好地缓解城市交通流量的不平衡问题。

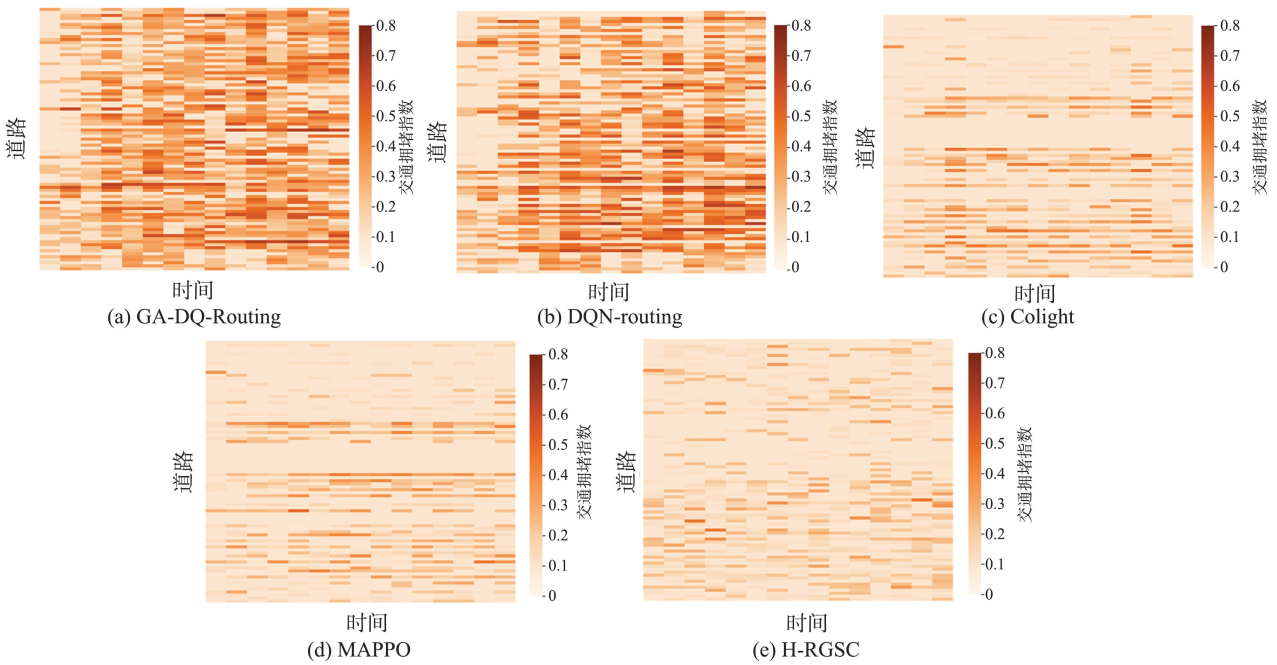


图 8 杭州 4×4 路网交通拥堵指数热力图对比

Fig.8 Hangzhou 4×4 road network traffic congestion index thermal map comparison

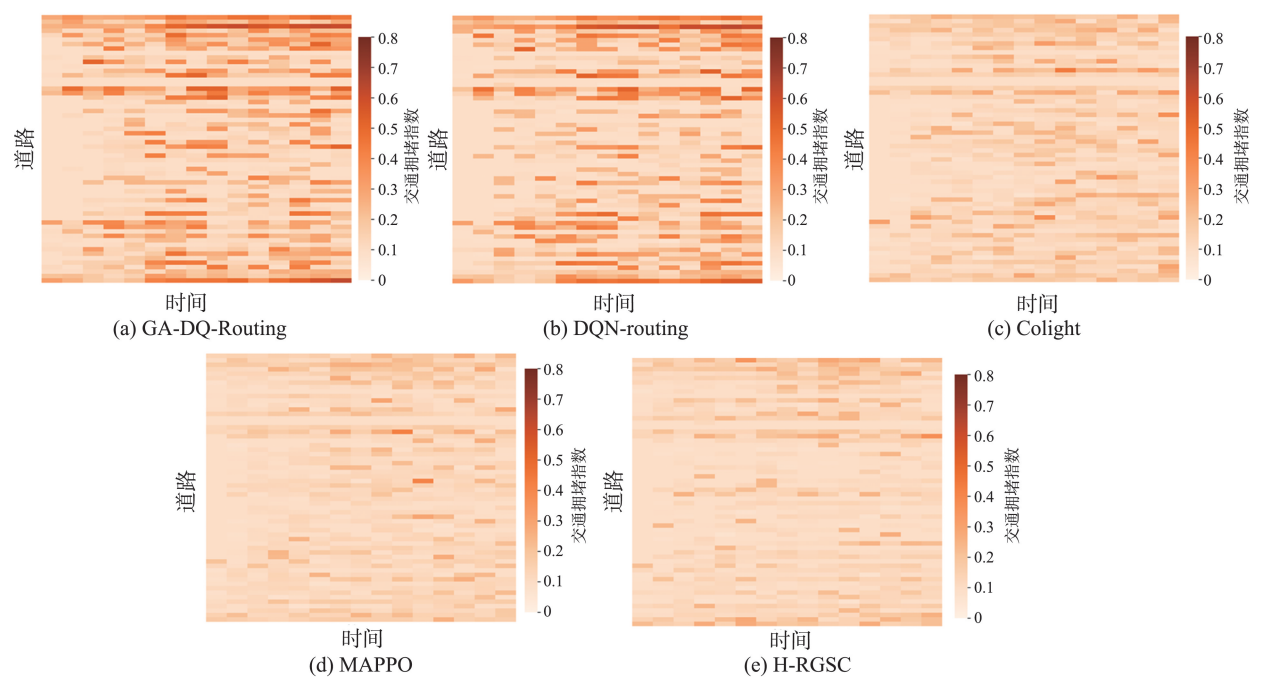
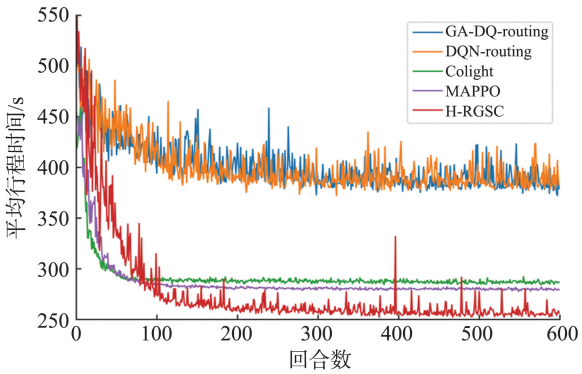


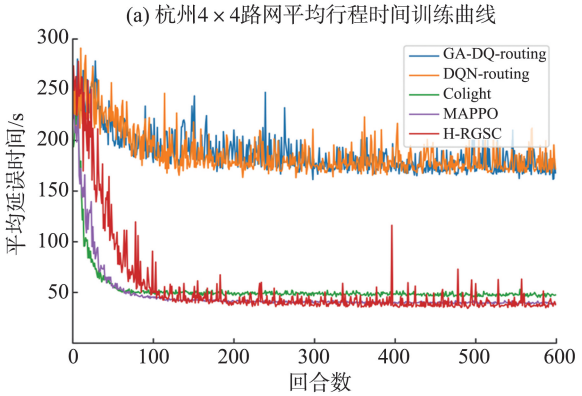
图 9 曼哈顿 16×3 路网交通拥堵指数热力图对比

Fig.9 Manhattan 16×3 road network traffic congestion index thermal map comparison

H-RGSC 与各个对比方法的训练曲线如图 10、11 所示。



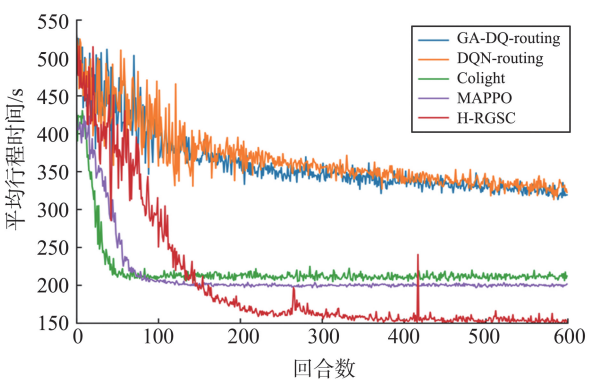
(a) 杭州4×4路网平均行程时间训练曲线



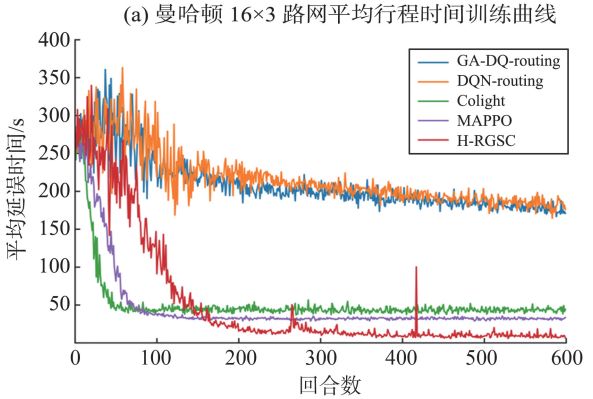
(b) 杭州4×4路网平均延误时间训练曲线

图 10　杭州 4×4 路网数据集训练曲线

Fig.10　Training curve for Hangzhou 4×4 road network dataset



(a) 曼哈顿 16×3 路网平均行程时间训练曲线



(b) 曼哈顿 16×3 路网平均延误时间训练曲线

图 11　曼哈顿 16×3 路网数据集训练曲线

Fig.11　Training curves for the Manhattan16×3 road network dataset

由图 10、11 训练曲线可知,信号控制基线收敛速度更快,但是随着训练轮次的增加,在平均行程时间与平均延误时间这两个指标上 H-RGCS 能达到更好的水平,这也证明了 H-RGSC 的收敛有效性。综合试验结果来看,路径引导、信号灯控制方法都能在一定程度上提升路网通行效率。由于信号灯控制方法能够直接干预交通流的核心调控点——交叉路口,因此在减少平均延误时间、缓解局部拥堵方面效果显著。Colight 通过图注意力网络整合邻近交叉口信息,增强了信号控制策略的空间协同性。MAPPO 利用多智能体强化学习框架,通过集中式训练和分布式执行,促进信号控制智能体间的合作。尽管信号控制策略在减少延误时间方面效果显著,但单独依赖信号控制往往难以应对由车辆路径选择不均导致的全局性交通流分布不均问题,有时甚至可能加剧某些路径的拥堵。路径引导方法根据车辆路由请求与实时交通状态为车辆提供个性化路径建议,但是,单独的路径引导策略可能忽视了交通信号的可控性,在交通信号固定配时的情况下,提升路网通行效率十分有限。

总体来看,H-RGSC 通过集成路径引导与信号控制,显著降低了平均行程时间和平均延误时间,凸显了多智能体系统在缓解交通拥堵中的效果,以及综合路径引导与智能信号控制的互补优势。H-RGSC 能够从宏观(信号控制)和微观(路径引导)2 个层面综合优化交通流,实现对交通流的全面优化,既能快速应对局部交通状况变化,又能引导车辆有效避堵,从而深度缓解交通压力,提升总体通行效率。H-RGSC 研究结果不仅证明了集成路径引导与信号控制对缓解交通的潜力,也对智能交通系统设计具有重要启示,强调在复杂动态城市环境中,综合多策略协同的重要性。

4　结论

本研究介绍了一种基于分层多智能体强化学习的个性化和信号控制联合路径引导方法。该方法将车辆路径引导问题和信号灯控制问题抽象为分散的部分可观察马尔科夫决策过程,并通过在交叉路口部署路径引导智能体与信号控制智能体,将车辆路线引导和交通信号控制建模为分层多智能体强化学习模型以引导、平衡城市交通流量。通过自适应图卷积网络聚合来自动态相邻智能体的信

息以促进同层次智能体之间的通信;对于车辆路径引导,基于平均场博弈考虑了行驶车辆之间的相互影响,以协调车辆之间在一定时空范围内的行动。基于MAPPO算法,实现了信号控制智能体之间的合作。并且,基于分层强化学习,实现异质智能体之间的信息交互并促进其在分层架构下的协同,以实行个体路径引导优化和整体的限流或分流,从而更大程度地提升交通效率。本研究基于开源交通数据集,在SUMO仿真平台上进行了试验和对比。结果表明,本研究方法能够同时在各项评价指标上取得最好的结果,有效提高城市车辆通行效率,更重要的是,它证明了综合考虑交通管理中不同层面策略协同的必要性。在未来的研究中将进一步探索如何提高算法的计算效率和在线学习能力,以及如何在更广泛的交通场景下实现策略的自适应调整和优化。

参考文献:

[1] ZHOU B, SONG Q, ZHAO Z, et al. A reinforcement learning scheme for the equilibrium of the in-vehicle route choice problem based on congestion game[J]. Applied Mathematics and Computation, 2020, 371: 124895.

[2] 周晓昕, 廖祝华, 刘毅志, 等. 融合历史与当前交通流量的信号控制方法[J]. 山东大学学报（工学版）, 2023, 53(4): 48-55.
ZHOU Xiaoxin, LIAO Zhuhua, LIU Yizhi, et al. Signal control method integrating history and current traffic flow[J]. Journal of Shandong University (Engineering Science), 2023, 53(4): 48-55.

[3] TANG C, HU W, HU S, et al. Urban traffic route guidance method with high adaptive learning ability under diverse traffic scenarios[J]. IEEE Transactions on Intelligent Transportation Systems, 2020, 22(5): 2956-2968.

[4] WEI H, ZHENG G, YAO H, et al. IntelliLight: a reinforcement learning approach for intelligent traffic light control[C]//Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. London, UK: Association for Computing Machinery, 2018: 2496-2505.

[5] VEZHNEVETS A S, OSINDERO S, SCHAUL T, et al. FeUdal networks for hierarchical reinforcement learning[C]//Proceedings of the 34th International Conference on Machine Learning. Sydney, Australia: JMLR, 2017: 3540-3549.

[6] 吴黎兵, 范静, 聂雷, 等. 一种车联网环境下的城市车辆协同选路方法[J]. 计算机学报, 2017, 40(7): 1600-1613.
WU Libing, FAN Jing, NIE Lei, et al. A collaborative routing method with internet of vehicles for city cars[J]. Chinese Journal of Computers, 2017, 40(7): 1600-1613.

[7] HALL R W. The fastest path through a network with random time-dependent travel times[J]. Transportation Science, 1986, 20(3): 182-188.

[8] MAO C, SHEN Z. A reinforcement learning framework for the adaptive routing problem in stochastic time-dependent network[J]. Transportation Research Part C: Emerging Technologies, 2018, 93: 179-197.

[9] KOH S, ZHOU B, FANG H, et al. Real-time deep reinforcement learning based vehicle navigation[J]. Applied Soft Computing, 2020, 96: 106694.

[10] ARASTEH F, SHEIKHGARGAR S, PAPAGELIS M. Network-aware multi-agent reinforcement learning for the vehicle navigation problem[C]//Proceedings of the 30th International Conference on Advances in Geographic Information Systems. Seattle, USA: Association for Computing Machinery, 2022: 1-4.

[11] SHOU Z, CHEN X, FU Y, et al. Multi-agent reinforcement learning for Markov routing games: a new modeling paradigm for dynamic traffic assignment[J]. Transportation Research Part C: Emerging Technologies, 2022, 137: 103560.

[12] WANG Y, XU T, NIU X, et al. STMARL: a spatio-temporal multi-agent reinforcement learning approach for cooperative traffic light control[J]. IEEE Transactions on Mobile Computing, 2020, 21(6): 2228-2242.

[13] VARAIYA P. Max pressure control of a network of signalized intersections[J]. Transportation Research Part C: Emerging Technologies, 2013, 36: 177-195.

[14] WEI H, CHEN C, ZHENG G, et al. PressLight: learning max pressure control to coordinate traffic signals in arterial network[C]//Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. Anchorage, USA: Association for Computing Machinery, 2019: 1290-1298.

[15] CHU T, WANG J, CODECÀ L, et al. Multi-agent deep reinforcement learning for large-scale traffic signal control[J]. IEEE Transactions on Intelligent Transportation Systems, 2019, 21(3): 1086-1095.

[16] WEI H, XU N, ZHANG H, et al. CoLight: learning network-level cooperation for traffic signal control[C]//Proceedings of the 28th ACM International Conference on Information and Knowledge Management. Beijing, China: Association for

Computing Machinery, 2019: 1913-1922.
[17] ZHU H, WANG Z, YANG F, et al. Intelligent traffic network control in the era of internet of vehicles[J]. IEEE Transactions on Vehicular Technology, 2021, 70(10): 9787-9802.
[18] SUN Q, ZHANG L, YU H, et al. Hierarchical reinforcement learning for dynamic autonomous vehicle navigation at intelligent intersections[C]//Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. Long Beach, USA: Association for Computing Machinery, 2023: 4852-4861.
[19] ZHANG L, WU Q, SHEN J, et al. Expression might be enough: representing pressure and demand for reinforcement learning based traffic signal control[C]// International Conference on Machine Learning. Baltimore, USA: PMLR, 2022: 26645-26654.
[20] HUANG H, HU Z, LU Z, et al. Network-scale traffic signal control via multiagent reinforcement learning with deep spatiotemporal attentive network [J]. IEEE Transactions on Cybernetics, 2021, 53(1): 262-274.
[21] BAI L, YAO L, LI C, et al. Adaptive graph convolutional recurrent network for traffic forecasting [J]. Advances in Neural Information Processing Systems, 2020, 33: 17804-17815.
[22] MUKHUTDINOV D, FILCHENKOV A, SHALYTO A, et al. Multi-agent deep learning for simultaneous optimization for time and energy in distributed routing system [J]. Future Generation Computer Systems, 2019, 94: 587-600.
[23] TANAKA T, NEKOUEI E, PEDRAM A R, et al. Linearly solvable mean-field traffic routing games[J]. IEEE Transactions on Automatic Control, 2020, 66(2): 880-887.
[24] YANG Y, LUO R, LI M, et al. Mean field multi-agent reinforcement learning[C]//Proceedings of the 35th International Conference on Machine Learning. Stockholm, Sweden: PMLR, 2018: 5567-5576.
[25] WANG Z, SCHAUL T, HESSEL M, et al. Dueling network architectures for deep reinforcement learning [C]//Proceedings of the 33rd International Conference on International Conference on Machine Learning. New York, USA: JMLR, 2016: 1995-2003.
[26] YU C, VELU A, VINITSKY E, et al. The surprising effectiveness of ppo in cooperative multi-agent games [J]. Advances in Neural Information Processing Systems, 2022, 35: 24611-24624.
[27] SCHULMAN J, WOLSKI F, DHARIWAL P, et al. Proximal policy optimization algorithms [DB/OL]. (2017-08-28) [2025-04-03]. https://doi.org/10.48550/arXiv.1707.06347
[28] LOPEZ P A, BEHRISCH M, BIEKER-WALZ L, et al. Microscopic traffic simulation using SUMO[C]//2018 21st International Conference on Intelligent Transportation Systems (ITSC). Maui, USA: IEEE, 2018: 2575-2582.
[29] SU H, ZHONG Y D, CHOW J Y J, et al. EMVLight: a multi-agent reinforcement learning framework for an emergency vehicle decentralized routing and traffic signal control system[J]. Transportation Research Part C: Emerging Technologies, 2023, 146: 103955.

(编辑:郭少华)

(上接第 33 页)

DAI Liang, ZHANG Yanan, QIAN Chao, et al. Optimal packet scheduling strategy for roadside units′ bursty traffic based on relaying vehicles[J]. ACTA Automatica Sinica, 2021, 47(5): 1098-1110.
[22] ZHOU H B, LIU B, HOU F, et al. Spatial coordinated medium sharing: optimal access control management in drive-thru Internet[J]. IEEE Transactions on Intelligent Transportation Systems, 2015, 16(5): 2673-2686.
[23] XU W C, ZHOU H B, SHI W S, et al. Throughput analysis of in-vehicle Internet access via on-road WiFi access points[C]//2017 IEEE 86th Vehicular Technology Conference (VTC-Fall), Toronto, Canada: IEEE, 2017: 1-5.
[24] WU J, CHEN W. Low-Latency and energy-efficient wireless communications with energy harvesting[J]. IEEE Transactions on Wireless Communications, 2022, 21(2): 1244-1256.
[25] 林峰, 丁鹏举, 梁吉申, 等. 车联网中协作数据分发方案研究[J]. 计算机工程, 2021, 47(8): 29-36.
LIN Feng, DING Pengju, LIANG Jishen, et al. Research on collaborative data distribution scheme in Internet of vehicles[J]. Computer Engineering, 2021, 47(8): 29-36.

(编辑:郭少华)