

大规模智慧交通信号控制中的强化学习和深度强化学习方法综述*

翟子洋, 郝茹茹[†], 董世浩
(长安大学 信息工程学院, 西安 710064)

摘要: 当前在交通信号控制系统中引入智能化检测和控制已是大势所趋,特别是强化学习和深度强化学习方法在可扩展性、稳定性和可推广性等方面展现出巨大的技术优势,已成为该领域的研究热点。针对基于强化学习的交通信号控制任务进行了研究,在广泛调研交通信号控制方法研究成果的基础上,系统地梳理了强化学习和深度强化学习在智慧交通信号控制领域的分类及应用;并归纳了使用多智能体合作的方法解决大规模交通信号控制问题的可行方案,对大规模交通信号控制的交通场景影响因素进行了分类概述;从提高交通信号控制器性能的角度提出了本领域当前所面临的挑战和未来可能极具潜力的研究方向。

关键词: 智能交通; 交通信号控制; 强化学习; 交通信号灯; 多智能体; 大规模交通网络

中图分类号: TP18

文献标志码: A

文章编号: 1001-3695(2024)06-003-1618-10

doi: 10.19734/j.issn.1001-3695.2023.08.0419

Review of reinforcement learning and deep reinforcement learning methods in large-scale intelligent traffic signal control

Zhai Ziyang, Hao Ruru[†], Dong Shihao
(School of Information Engineering, Chang'an University, Xi'an 710064, China)

Abstract: At present, it is a general trend to introduce intelligent detection and control into traffic signal control system, especially reinforcement learning and deep reinforcement learning methods show great technical advantages in scalability, stability and extensibility, and have become a research hotspot in this field. This paper studied traffic signal control tasks based on reinforcement learning, systematically sorted out the classification and application of reinforcement learning and deep reinforcement learning in the field of intelligent traffic signal control on the basis of extensive research results on traffic signal control methods, and summarized feasible solutions to large-scale traffic signal control problems by using multi-agent cooperation. This paper classified and summarized the factors affecting the traffic scene of large-scale traffic signal control, put forward the current challenges and potential research directions in this field from the perspective of improving the performance of traffic signal controllers.

Key words: intelligent transportation; traffic signal control; reinforcement learning; traffic light; multiple agents; large-scale traffic network

0 引言

随着智能网联道路系统等级的不断提高,交通控制体系正在向着信息化、智能化和自动化的方向发展。在此大背景下形成的智能交通信号控制系统能够以海量的交通数据为基础,利用人工智能的方法,对区域协调控制方案进行优化,从而有效提高道路交叉口的交通效率,减缓道路压力,降低车辆延误和排队长度^[1]。

智能交通信号控制系统一般由数据采集与处理、协调控制方案生成和实时态势评价三部分组成,如图1所示。其中,数据采集与处理模块旨在对视频、地磁等交通检测器检测数据和GPS、车联网等互联网大数据进行获取与解析,并转换为符合协调控制模型要求的数据格式。协调控制方案生成模块优化交叉口通行决策,对交通信号进行协调控制,提高交叉口通行效率等指标。实时态势评价模块旨在使用排队长度等评价指

标对协调控制方案的控制效果进行衡量,该模块的实时反馈可以促进控制方案的改善。

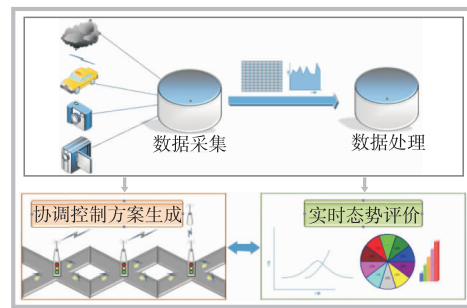


图1 智能交通信号控制系统的框架

Fig.1 Framework of intelligent traffic signal control system

已有的智能交通信号控制系统充分体现了人工智能技术的影响,例如英国的 SCOOT 系统^[2]、美国的 ACTRA 和 OPAC 系统^[3]、法国的 CRONOS 系统^[4]、澳大利亚的 SCATS 系统^[5]

收稿日期: 2023-08-29; 修回日期: 2023-11-16 基金项目: 国家重点研发计划资助项目(2021YFA1000300,2021YFA1000303); 国家青年基金资助项目(202006565013)

作者简介: 翟子洋(2000—),女,河北邯郸人,硕士研究生,主要研究方向为交叉口智能控制;郝茹茹(1984—),女(通信作者),陕西周至人,副教授,硕导,主要研究方向为交通智能检测与控制技术(1123915292@qq.com);董世浩(2001—),男,陕西渭南人,硕士研究生,主要研究方向为车联网技术、智能网联汽车。

等。国内也出现了一批优秀的智能交通信号控制系统,包括 HT-UTCS、南京 NATS 城市交通信号控制系统、海信 HiCon 交通信号控制系统、深圳 SMOOTH 智能交通信号控制系统等。此外,代表我国研究最新进展的新一代系统也不断涌现,如天津大学贺国光教授团队的 TICS 系统、同济大学杨晓光教授团队的 TJATCMS 系统以及吉林大学杨兆生教授团队的 NITCS 系统等,这些系统能够在不同的道路网络条件和交通流特性下提供实时性和适应性的协同优化。

由于交通系统具有复杂性、时变性等特征,智能交通信号控制系统在实时性与鲁棒性的提升方面面临着巨大的挑战。为解决这一问题,智能交通信号控制领域出现了不同的方法,包括基于交通理论的方法^[6]、基于启发式的方法^[7]、数据驱动的方法等。近年来,数据驱动方法下的强化学习(reinforcement learning, RL)和深度强化学习(deep reinforcement learning, DRL)^[8]被越来越多地应用于智能交通信号控制系统中。RL 属于自适应信号控制范畴,该方法能够根据实时交通状态自动生成信号控制方案。RL 方法的优势在于摒弃了复杂严苛的预定义数据模型,使智能体能够在与环境的交互中学习,并不断优化动作策略,更适合于复杂且动态的交通环境^[9]。但是,由于 RL 所涉及的状态空间可能会随着研究问题规模的增大呈指数级增长,导致其需要更多的时间和计算资源。而深度学习(deep learning, DL)与 RL 方法的成功结合解决了这个问题,DRL 能够在有限的资源下有效减少探索到最优决策的学习时间。

随着智能交通信号控制系统的规模不断扩大,传统的集中式单智能体控制策略已无法满足区域整体的需求。因此,RL 正在向着多智能体强化学习(multi-agent reinforcement learning, MARL)的方向发展。MARL 是单智能体强化学习在复杂博弈环境中的扩展,它通过多个交叉口的 RL 智能体协同控制,不断逼近交通网络环境中的最优决策。然而,在解决网络级交通信号控制(network traffic signal control, NTSC)问题时,可观测的交通数据信息量与单点控制相比呈指数级增长,交通状态空间的巨大导致大量的时间和计算资源消耗,同时也对信息存储容量提出了更高的要求。因此,为了应对大规模交通信号控制问题,出现了多智能体深度强化学习(multi-agent deep reinforcement learning, MADRL)。MADRL 将神经网络与原始数据的抽象表示相结合,以处理与流量相关的复杂数据流,并能够与多种方法和框架相结合,体现出其优异的性能。

在大规模交通网络中由信号协同控制的研究中,一些文章对 RL^[10]、DRL^[11]、多智能体系统^[12]和其他模糊系统^[13]等人工智能方法进行了综述。然而,目前缺乏系统且全面的总结和分析。针对上述问题,本文收集了大规模智能交通信号控制领域的相关文献,并总结了现有文献的研究重点。同时,对从文献中提取的数据进行了分析和描述,探索了该领域中的先进解决方案,并归纳了最近研究中面临的挑战。本文还提出了未来潜在的研究方向,特别关注了在大规模智能交通信号控制中应用的 RL 和 DRL 两种方法,并探讨了它们提供的创新思想和框架,这将有助于探索现有研究方案之间存在的差距。

1 强化学习在交通信号控制领域的应用

本章将聚焦于交通信号控制领域的典型方法,主要关注 RL 和 DRL 方法。首先,简要概述了 RL 方法的组成要素,并分析了它们在交通信号控制领域中的含义。同时,探讨了现有研究中表示交通状态信息、决策内容和评价指标的方法。此外,将现有的 RL 方案按照智能体决策框架的不同进行了分类。接下来,详细阐述了 DRL 在交通信号控制领域的应用方法,并

分析了引入 DRL 的必要性和优越性。对现有的典型 DRL 框架进行了分类和阐述。

1.1 RL 方法的应用

RL 方法源于马尔可夫决策过程(Markov decision process, MDP),它是一种用于在非确定性情况下进行决策优化的数学框架。MDP 的组成部分包括状态空间(state space, S)、动作空间(action space, A)、转换函数(transition function, T)和奖励函数(reward function, R)。MDP 的优化目标是找到一个策略,使得智能体在任意状态 S 下都能够采取最佳动作 A ,从而最大化全局累积奖励 R 。

在交通信号控制领域,RL 智能体的状态是指从环境中获取的交通现状,它需要在有限空间内充分体现整个区域全面的交通现状。智能体的动作是对环境作出的反应,通常表现为对交通信号灯的相位进行切换,以改变交叉口环境中的交通流状态,这些动作是智能体所要优化的对象。而奖励是对智能体执行的动作“好”“坏”的反馈,智能体只有找到最大化奖励的方法才能作出最优的决策。针对不同的应用场景,需要选择不同的要素空间。在交通信号控制中,通常涉及以下要素表示方法。

最常用的状态表示方法包括队列长度、等待时间、相位状态、车辆数量^[14]、车辆位置和速度、RGB 图像^[15]等。导航信息^[16]也可以用于交通信号控制,通过挖掘多条导航路线的时空关系可以提取动态、实时的交通特征。对于具有多个交叉口的交通信号控制模型,状态定义还涵盖相邻交通灯信息等^[17]。

图像是最能直观和全面地表现交叉口交通状态的方式之一。然而,原始 RGB 图像反映的要素过多,难以提取有效信息。为了解决这一问题,在交通信号控制领域出现了一种称为离散状态编码(discrete traffic state encoding, DTSE)的类图像表示格式,它是最受关注的状态表示方法之一。DTSE 将每个交叉口的进口道划分为均匀长度的单元,或者根据距离交叉口的远近划分为不同长度的单元,然后将车辆的速度、位置和加速度等信息以独立的阵列形式显示。图 2 为 DTSE 的示意图。

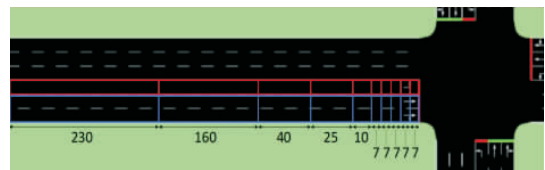


图2 离散状态编码示意

Fig. 2 Discrete state coding scheme

控制交通信号灯的動作主要可以分为基于相位和基于周期两组。在这两组动作中,周期长度、相位持续时间和相位顺序是关键要素,它们可以是固定的或可变的。在基于周期的方法^[18]中,决策点是周期的结束,决策内容包括周期长度、相位持续时间或相位顺序。而在基于相位的方法^[19]中,决策点是相位的结束,决策内容包括相位持续时间^[20]和相位选择^[14]。在基于相位的方法中,可以将整个相位的持续时间设置为固定值,或者允许在相位结束时延长当前相位的绿灯时长。此外,从动作的连续性和离散性角度考虑,大多数研究使用一组离散动作进行决策,但也有少数研究考虑连续动作输出^[21]。连续动作空间的定义可以基于预定义的最小和最大相位持续时间,以预测当前相位的持续时间。在使用混合动作空间的研究中,离散动作表示相位选择,连续动作对应相位持续时间^[22]。

最常用的奖励表示方法包括队列长度、延误时间、等待时间^[18]、车辆数量^[19,23]和燃料消耗等。通常情况下,为特定的交通场景定义单一的奖励函数,然而一些研究提出了使用多个奖励的方法^[24],包括团队奖励。在多交叉口场景下,差异奖

励^[25,26]可以解决合作智能体之间的信用分配问题,同时空间加权奖励可以让智能体差异化地考虑道路网络中的相邻智能体,以评估更优的协同策略。

RL 智能体的行为表现为以下两种形式:

a) 学习状态转换函数 T 的基于模型的 RL 方法^[27],如表 1 所示。在交通信号控制领域,文献[28]引入了一种新的基于模型的方法,称为 MuJAM,该方法在实现大规模显式协调的基础上进一步提高了泛化能力。

b) 无模型的 RL 方法,即不学习转换函数。无模型的 RL 可以细分为基于价值的 RL 和基于策略的 RL。在基于价值的 RL 中,智能体更新价值函数,该价值函数能够反映状态与动作之间的映射关系。而在基于策略的 RL 中,智能体使用策略梯度来更新策略。

在基于价值的 RL 方法中,价值函数 $V^\pi(s)$ 可以估计给定策略下某一状态的价值,即奖励的期望,从而反映该状态对智能体的影响。通过对所有状态的价值函数值进行累加,可以得到最优状态值函数 $V^*(s)$ 。状态-动作价值函数被称为 Q 函数 $Q^\pi(s, a)$,用于反映状态-动作对的期望奖励。将状态-动作对的 Q 值累加,得到最优 Q 函数 $Q^*(s, a)$ 。

Q 值使用贝尔曼方程的递归性质来计算:

$$Q^\pi(s_t, a_t) = E_\pi[r_t + \gamma Q^\pi(s_{t+1}, \pi(s_{t+1}))] \quad (1)$$

Q 值的更新如下:

$$Q^\pi(s_t, a_t) \leftarrow Q^\pi(s_t, a_t) + \alpha(y_t - Q^\pi(s_t, a_t)) \quad (2)$$

其中: y_t 是 $Q^\pi(s_t, a_t)$ 的时间差分目标(temporal difference, TD)。

由于基于价值的 RL 方法需要存储大量状态-动作对的 Q 值,基于策略的 RL 方法在处理高维和连续空间问题时性能更优。基于策略的 RL 方法可以分为梯度和非梯度方法^[29],其中基于策略梯度的 RL 方法展现出更优的性能,它利用目标函数 $J(\theta)$ 关于 θ 的梯度来选择动作。

在著名的策略梯度算法 REINFORCE^[30]中,将期望奖励作为目标函数,策略梯度计算为

$$\nabla_\theta J(\theta) = E_{\pi_\theta}[Q^{\pi_\theta}(s, a) \nabla_\theta \log \pi_\theta] \quad (3)$$

参数 θ 的更新使用随机梯度下降,在更新过程中,时间 t 处的奖励 R_t 作为 $Q^{\pi_\theta}(s_t, a_t)$ 的估计值。

$$\theta \leftarrow \theta + \alpha \nabla_\theta \log \pi_\theta R_t \quad (4)$$

根据不同的模型框架,RL 可以分为多种方法,包括 Q-学习、SARSA、策略梯度下降(policy gradient, PG)、actor-critic 等,如表 1 所示。使用不同的 RL 方法可以在一定程度上提高算法的鲁棒性、速度和效率等性能。在选择模型时,应考虑每种方法的优劣,以最大化模型的效能,并满足不同应用场景的需求。

在基于模型的方法中,反向强化学习(inverse reinforcement learning, IRL)方法考虑到手动为任务设计奖励函数的困难性,其通过观察个体的行为来学习其目标、价值或奖励,并通过观察智能体随时间的行为变化来确定奖励函数。这种方法类似于向专家学习,有助于在不容易获得奖励的场景中进行应用,在交通信号控制领域可以充分体现其优势。

学习分类器系统(learning classifier system, LCS)是一种基于规则的 RL 系统,每个规则由条件、动作和奖励组成,构造为 {if “条件” then “动作”} 的形式。通过与遗传算法的结合, LCS 可以改善规则空间,从当前的强分类器中生成新的分类器并去除弱分类器,因此,RL 智能体能够选择具有最佳奖励响应的动作。

Q-学习和 SARSA^[31]是基于价值的 RL 算法中的两个主要分支。Q-学习和 SARSA 之间的差异主要体现在值函数的更新

过程。在 Q-学习中,智能体的更新基于最大化值函数,而在 SARSA 中,智能体的更新基于从策略函数中导出的值。

$$y_t^{\text{Q-learning}} = R_t^{(n)} + \gamma \max_{a_{t+n}} Q^\pi(s_{t+n}, a_{t+n}) \quad (5)$$

$$y_t^{\text{SARSA}} = R_t^{(n)} + \gamma^n Q^\pi(s_{t+n}, a_{t+n}) \quad (6)$$

表 1 应用于交通信号控制的强化学习方法对比

Tab. 1 Comparison of reinforcement learning methods applied to traffic signal control

分类	算法	性能	评价	分析
基于模型	IRL	泛化性	√	学习奖励函数
		复杂性	×	奖励函数多样
		适应性	√	多步骤决策
	LCS	实时性	√	在线学习
		探索性	×	局部搜索
		实时性	√	在线学习
基于价值	SARSA	稳定性	√	当前策略更新
		探索性	×	贪婪策略探索
		简单性	√	易于理解和实现
	Q-学习	探索性	×	贪婪策略探索
		分布性	√	结合分布式算法
		探索性	√	W 值未知动作评价
	W-学习	复杂性	×	大量的状态动作对
		收敛性	×	智能体之间的竞争
		高效性	√	经验回放
	DQN	稳定性	√	目标网络
		复杂性	×	参数调整
		探索性	√	学习值分布
无模型 基于策略	C51	泛化性	√	适应不同任务
		复杂性	×	离散概率分布
		适用性	√	连续动作空间
	PG	并行化	√	并行化训练
		稳定性	×	高方差
		探索性	×	随机性策略探索
	PPO	稳定性	√	“clip”方法
		并行化	√	并行化训练
		适用性	√	连续或离散空间
	TRPO	探索性	×	随机性策略探索
		适用性	√	连续动作空间
		稳定性	√	限制策略更新幅度
价值与 策略融合	actor-critic	收敛性	×	训练速度较慢
		探索性	×	随机性策略探索
		适用性	√	连续或离散空间
	A2C/A3C	高效性	√	协同作用
		复杂性	×	较多计算资源
		高效性	√	并行化训练
	DDPG	收敛性	√	优势函数降低方差
		探索性	×	依赖探索策略
		复杂性	×	并行化训练
	TD3	稳定性	√	目标网络
		适用性	√	连续动作空间
		复杂性	×	参数调整
策略融合	SAC	高效性	√	经验回放
		稳定性	√	双 Q 网络
	SAC	稳定性	√	“soft”策略更新
		高效性	√	经验回放

W-学习是一种基于 Q-学习的多策略自组织动作选择技术。在 W-学习中,智能体之间存在竞争关系,每个智能体学习状态-动作对的 Q 值,并为每个策略和状态计算 W 值,以便探索未执行指定动作的后果和影响。在将 W-学习方法应用于大规模多智能体协同优化的过程中,与分布式算法相结合可以获得良好的效果。

不同于基于价值的方法,PG 方法无须对状态或动作值函数进行估计。它通过搜索策略空间直接学习参数化的策略函数,并使用最大化累积奖励来进行度量。通过这种方式,PG 方法避免了值函数估计的收敛问题^[32]。

actor-critic 方法将基于值的 critic 算法和基于策略的 actor

算法相结合,其包含 actor 和 critic 两个估计器。actor 使用 Q 值函数的估计结果,而 critic 使用状态价值函数的估计结果。基于策略的 actor 负责选择动作,而基于价值的 critic 则用于评估 actor 所采取动作的好坏程度^[33]。

1.2 DRL 方法的应用

DRL 是一种将两种人工智能方法(即 DL 和 RL)相结合的技术。DL^[34]是一种相对较新的学习范式,它将由大量神经元层组成的多层感知器集成到 RL 方法中,这种整合已被证明可以解决原有方法的缺点^[35]。对于 RL 来说,查找表来映射状态空间会大大降低效率,已成为阻碍其在真实世界中应用的原因之一。尽管一些研究提出使用线性函数逼近方法来解决高维空间问题,但改进的程度非常有限。因此,采用 DL 进行函数逼近的优势日益凸显。在 DRL 中,神经网络被用于训练和学习策略函数和价值函数^[9]。下面是 DRL 方法的典型框架:

1) 深度 Q 网络(deep Q -network, DQN)

DQN 是第一种 DRL 方法,在交通信号控制中已被广泛应用。将人工神经网络(artificial neural network, ANN)集成到 Q -学习过程中的方法被称为深度 Q 学习,使用 ANN 来近似 Q 函数的网络被称为深度 Q 网络,因此 DQN 是一种 Q -学习。如果用 θ 表示神经网络中的参数, Q 函数的近似可以表示为 $Q(s, a; \theta)$ 。与 1.1 节中介绍的 Q -学习相比,深度 Q 学习算法使用深度神经网络作为函数逼近器,取代了 Q 表来近似 Q 函数,从而使更大或连续的状态空间的参数化表示成为可能^[21]。DQN 框架如图 3 所示。

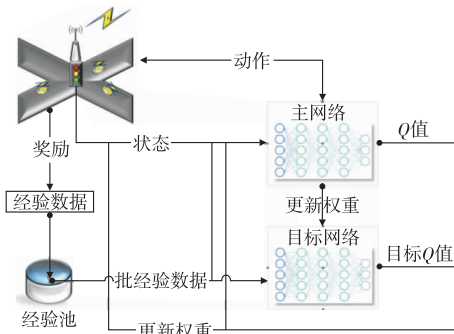


图3 DQN 框架

Fig. 3 DQN framework

针对 DQN 的改进方法,文献[36]提出了目标网络和经验回放这两种新技术来稳定 DQN 的学习过程。目标网络技术是指 DQN 模型由两个相同结构的主网络和目标网络组成。主网络用于控制 Q 函数的近似过程,而目标网络则在模型训练阶段更新主网络的参数。通过使用两个独立网络进行更新,DQN 能够保证 Q 值估计的稳定性。经验回放是指将状态、动作和奖励等经验信息进行存储。在网络训练过程中,存储的经验会被批次采样以更新网络参数。DQN 通过经验回放规避了数据的时间粘连性问题。此外,智能体每次只从经验池中抽取较少的经验用于训练,无须使用所有的经验数据,从而提高了训练的效率。

然而,经验回放技术虽然保证了随机性,但忽略了经验对训练的差异性影响。为了解决这个问题,文献[37]提出了优先经验回放一种新的经验数据采样方法,即基于优先级的随机抽样。根据经验的 TD 误差,相应调整采样的概率。TD 误差越大的经验被采样的概率越大,反之则越小。

2) 双决斗深度 Q 网络(dueling double DQN, D3QN)

针对深度 Q 学习不收敛的问题,double 和 dueling 两种网络架构的应用能够充分提高 DQN 的稳定性。在 DQN 的基础上同时使用以上两种架构被称为 D3QN 网络,其在交通信号控

制领域的应用中已经取得了良好的效果^[38]。

针对 double 网络架构,文献[39]提出,在对损失进行最小化的过程中,令主网络选取最优动作、目标网络计算目标值,将主网络动作选择和目标网络动作评估的估计量加倍。因此,double DQN 不再使用目标网络选择期望奖励最大化的 Q 值,而是将其与动作选择部分取消关联,能够充分解决过估计的问题。

针对 dueling 网络架构^[40],分别为每个动作估计状态价值函数和优势函数。状态价值函数可以计算当前状态下所有动作值的平均期望,而优势函数可以计算在选择某一动作时,其值超出期望值的程度,从而体现出不同动作的差异。如果一个动作的优势函数计算为正值,那么这个动作的性能超出平均水平,反之则低于平均水平。因此,dueling 网络架构能够大幅提升学习效果,并加速收敛。

3) actor-critic 方法

正如 1.1 节所述,actor-critic 方法结合了基于价值和基于策略的算法的优势,通过将 actor-critic 方法中的两个函数逼近器用神经网络代替,形成了一些深度方法。此外,并行计算与异步计算为加快算法的学习与训练速度创造了条件,使智能体能够在更短的时间内达到更优的收敛效果,从而有效提高智能体的性能。文献[41]提出的异步多因子学习器模型证明了异步学习的优势。另外,智能体的并行与异步能够在短时间内提供更多的训练数据,因此,智能体的异步更新不需要经验重放机制,而是在预定义的周期内完成经验的累积。基于此,异步优势 actor-critic 算法(asynchronous advantage actor-critic, A3C)被广泛应用,其使用并行处理器对策略和价值网络进行单独更新。

在 actor-critic 方法的基础上,衍生出了一种称为深度确定性策略梯度(deep deterministic policy gradients, DDPG)的算法。DDPG 在状态空间上估计确定性策略梯度方法,而不是同时在状态和动作空间上估计随机策略梯度。DDPG 采用了双网络结构,结合函数近似和策略函数,同时使用经验重放技术,使得神经网络在概率算法中能够以最高效率进行学习^[42]。在 MADRL 领域,DDPG 方法也得到了广泛应用,例如文献[43]研究了一种名为双延迟深度确定性策略梯度算法,通过优化相位分割来控制交通信号。

2 多智能体合作方法在交通信号控制领域的应用

NTSC 问题具有规模大、视界远的特点,同时状态和动作空间的连续性使问题的复杂度急剧增加。因此,建立 NTSC 模型是一项极具挑战性的任务。在相关论文中,研究人员尝试采用各种方法来改善 NTSC 的性能,例如使用函数逼近方法来处理维度变化,使用规模有限的智能体集合来降低问题复杂度,提高可扩展性、稳定性和优化速度,以及探索管理和推广状态或动作空间的可行性等。在 MARL 的研究中,涉及到智能体对环境的稳定性和适应性问题,使用分布式学习、合作学习和竞争学习等方法来解决。

本章从 MARL 和 MADRL 两种合作框架的角度对大规模交通信号控制领域的问题进行了分析。首先,对 MARL 的三种子框架进行了阐述,并总结了每种子框架在现有研究中的应用成果;其次,讨论了 MARL 与其他领域方法相结合的控制方法;然后,通过分析 MADRL 在解决非平稳性问题和结合分布式框架时的性能表现,展示了其在交通信号控制领域应用中的优势;最后,总结了在现有研究中考虑交通场景等影响因素的解决方案。

2.1 MARL 方法解决大规模交通信号控制问题

2.1.1 以 RL 方法为核心的 MARL 控制

根据大规模交通环境中智能体通信的方法和水平,道路网络中的 MARL 控制模式可以分为完全独立、部分合作和动作联动。三种模式的算法示例如表 2 所示。

表 2 MARL 控制模式的算法对比

Tab. 2 Algorithm comparison of MARL control mode

控制模式	算法	性能	评价	分析
完全独立	minimax-Q 学习	稳定性	✓	协调智能体公平性
		探索性	✓	防止劣势策略
		复杂性	×	必须考虑最坏情况
		协作性	×	智能体竞争环境
部分合作	多智能体深度确定性策略梯度 (multi-agent deep deterministic policy gradients, MADDPG)	协作性	✓	协同决策
		适用性	✓	连续动作空间
		扩展性	✓	多智能体环境
		收敛性	×	存在非合作性对策
	纳什 Q-学习	稳定性	✓	达到纳什均衡
		理论性	✓	博弈论基础
		复杂性	×	非线性博弈
		鲁棒性	×	依赖对手策略
动作联动	多智能体近端策略优化 (multi-agent proximal policy optimization, MAPPO)	协作性	✓	协同决策
		适用性	✓	连续动作空间
		稳定性	✓	策略梯度方法
		协作性	✓	协同决策
	对抗性多智能体 (counterfactual multi-agent, COMA)	稳定性	✓	对抗性对策
		鲁棒性	×	依赖对手策略
		稳定性	✓	训练过程稳定
		分解性	✓	多问题分解
	值分解网络 (value decomposition networks, VDN)	并行性	✓	并行训练
		适用性	×	不适用连续空间
		探索性	×	易达局部最优
		协作性	✓	协同决策
	单调值函数分解 (monotonic value function factorization, QMIX)	分解性	✓	全局值函数分解
		扩展性	✓	多智能体环境
		复杂性	×	考虑值函数

1) 完全独立的 MARL 控制

在完全独立的 MARL 控制中,每个交叉口的智能体仅基于本交叉口区域内的环境状态和奖励信息来作出最优决策,与其他交叉口的环境状态无关,即智能体之间缺乏协调机制。这种缺乏关联的独立控制使得单个智能体只能获得有限范围的环境数据,无法掌握全局环境状况。因此,从大规模交通环境的整体控制效能来看,这种控制方法存在很大的局限性。以表 1 中的 minimax-Q 算法为例,它以 minimax 为基础,使用 Q-学习中的 TD 进行迭代更新。其优势在于能够在多智能体环境下保持相对稳定的性能,并避免采用过于贪婪的策略。但其收敛和学习速度慢,且不利于解决智能体合作问题。

2) 部分合作的 MARL 控制

在部分合作的 MARL 控制中,每个交叉口的智能体能够获取附近上下游交叉口的环境数据。因此,这种控制方式扩展了完全独立控制方式中的交通数据范围,使得智能体能够对其他交叉口的环境状态作出反应,增强了智能体之间的交互能力。然而,该控制方式的缺点是没有在全局范围内系统地反映交通流的复杂动态信息。在这种控制方式中,多个智能体在相互通信的同时处理给定的任务,从其邻居和自身学习,通常可以获得更好的优化效果。例如表 2 中的 MADDPG^[44],其为集中式训练,分散式执行 (centralized training and decentralized execution, CTDE) 的典型代表, critic 采用全局信息进行训练, actor 采用局部信息进行动作获取。该算法支持多智能体之间的协同决策,以保证最优化整个路网系统的性能。

为了实现多智能体之间的环境信息共享,分层和分布式协作框架在交通信号灯智能体之间的信息传递中起着关键作用。

文献[45]首次提出了分布式 MARL 方法,该方法可以存储多个智能体周围的信息,允许每个智能体在与邻居通信的同时作出自己的决策^[46]。文献[47]在 NTSC 中的分布式计算机中使用了 Q-学习算法。文献[48]将 LCS 和基于 TCP/IP 的通信服务器结合到分布式学习控制策略中,以提高控制速度。

而分层控制既包含集中式控制智能体,又包含分布式控制智能体,可以通过多个区域智能体和一个集中的全局智能体来实现^[49]。其中,每个区域智能体在有限范围的区域内进行决策,而集中式全局智能体分层聚合其他区域智能体的数据,提高了结果的准确性,使系统达到最佳功能和效率。文献[50]按照这种分层控制的思想,提出了一个具有一个“经理”和几个“工人”的分层结构,以解决大规模交通信号控制的问题。文献[27]将奖励函数拆分为每个智能体的贡献,并使用变量消除算法找到联合动作,以确保在多智能体环境中实现协调控制。

3) 动作联动的 MARL 控制

在动作联动的 MARL 控制中,联合状态取代了单个智能体的独立状态,联合动作取代了单个智能体的独立动作,通过多个智能体之间的联合决策来逐步逼近最优决策。然而,这种控制方式导致了状态和动作空间的维度大幅增加,从而对存储功能和控制效率产生了较大影响。因此,在设计动作联动方法时,需要充分考虑状态和动作的数据结构、协调机制和值函数等因素。例如表 2 中的 MAPPO、COMA^[51] 和 VDN 等算法,它们通过智能体之间共享参数和传递参数以进行集中式训练,这表明智能体之间协作关系紧密,共同维护统一的奖励。

在这种控制方式中,多个智能体收集的数据可以集中存储在一个位置,所有智能体都可以访问该位置。在这样的设置中,通常由全局智能体来为系统作出所有决策,这会在参与协调的同时减缓学习过程。文献[52, 53]提出了一种新的集中式控制方法,通过引入分解机制来处理高维状态和动作空间,采用动作反馈技术改进决策,并应用图注意力网络模型 (graph attention network, GAT) 来学习周围交叉口的空间特征,以有效地估计未来的奖励。

为了解决动作联动导致的维度空间过大的问题,在 NTSC 中,数据通信过程的计算范式成为一个新的研究领域,特别是边缘计算和雾计算在 NTSC 中的应用成为热门话题。图 4 展示了一个通用的边缘计算框架,由设备层和边缘层组成。在大规模交通信号控制的应用中,设备层包括与车联网相关的车辆、交通信号灯、路侧单元等设备,而边缘层则由每个运行 MARL 算法的终端和上层的控制端构成。智能体终端可以直接参与交通信号灯的切换,而控制端通过调整终端的 RL 学习率等参数来优化交通网络。将边缘计算框架应用于大规模交通信号控制中,能够实现快速响应,并有效降低网络负载。

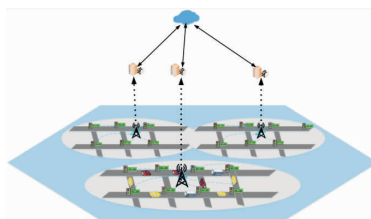


图 4 通用边缘计算框架

Fig. 4 General edge computing framework

文献[54, 55]设计了适用于 NTSC 场景的边缘计算框架,并提出了一种基于 MARL 的合作 NTSC 算法,旨在解决维度灾难问题,提供最小的响应时间,并减少网络负载。此外,文献[56]提出了一种针对大规模交通场景中拥堵问题的边缘增强学习 (edge-based RL, ERL) 解决方案,以缓解复杂交通场景中

的拥堵情况。另外,文献[57]提出了一种基于雾计算范式的交通控制架构,通过为每个交叉口生成交通信号控制流和通信流,降低城市中交通拥堵的概率。文献[58]利用雾节点来分解大规模网络,并提出了基于雾计算的图形增强学习(fog-based graphics RL,FG-RL)模型,可扩展到更大规模的交通网络中。

2.1.2 RL与其他领域方法相结合的MARL控制

随着RL方法的不断发展和应用,出现了许多将RL方法与其他领域的研究方法相结合的混合方法,这些方法能够在两种或多种方法的基础上互相补充,从而达到更优的控制效果。例如,RL与偏置压力(bias pressure,BP)方法结合^[59]、使用模仿学习来进行预训练^[60]、从博弈论角度充分考虑智能体之间的沟通和协调^[61]、结合上下文检测技术^[62]等。

在分布式控制场景下,受生物免疫系统的启发,文献[63]使用基于人工免疫的网络来实现智能体对干扰的捕获、适应和处理,并结合基于案例的推理(case-based reasoning,CBR)方法来处理可能中断的交通流。类似地,文献[64]也采用CBR来处理事件的检测和预测。针对大规模NTSC问题,文献[38]采用图分解方法,通过连接水平(level of connection,LOC)对交叉点进行聚类,能够以同步的方式训练网络子图。在考虑连续动作与状态空间问题时,文献[65]提出了一种集成模糊逻辑的方法,通过结合Q-学习和模糊逻辑来实现控制。此外,还可以考虑使用模糊规则控制分类器系统^[66],以实现比分层控制更快的效果。

RL与群智能优化算法的结合在控制领域展现出了良好的效果。使用遗传算法改变RL的要素参数,可以将局部智能体的优化与全局优化相结合^[67]。群智能算法与Q-学习的结合能够展现出比标准Q-学习更好的性能^[68]。如果将神经模糊系统与群智能算法同时结合于RL算法中,可以充分发挥两者的优点和性能。另外,在单目标控制器向多目标控制器的转换过程中,可以结合贝叶斯解释,设计新的奖励条件^[69],从而实现多目标优化。

2.2 MADRL方法解决大规模交通信号控制问题

MADRL方法中存在与2.1节MARL类似的通信水平分类。相较于MARL控制方法,在处理大规模交通信号控制问题时,MADRL引入了神经网络并将其结合到原始数据的抽象表示中。MADRL采用了不同的方法来处理与流量相关的复杂数据流,并能够与多种方法或框架充分结合,在处理以下问题时展现出优异的性能。

2.2.1 MADRL能够有效解决学习非平稳性问题

针对交通网络的复杂性和学习平稳性问题,由于网络的高度动态性导致了非平稳性,在训练过程中需要不断地重新开始学习以调整策略。MADRL提出了一种解决方法,即使用交叉口智能体的空间折扣因子,减弱对距离较远的交叉口数据的关注程度,更加注重本地环境信息的重要性^[70]。

2.2.2 MADRL能够有效结合分布式与分层框架

正如2.1节所述,分层和分布式协作框架能够在智能体共享环境信息的机制中起到关键作用,相较于MARL,MADRL能够更加有效地与此类框架相结合,从而充分发挥智能体间的协作作用。在交通信号的分布式控制机制中,文献[71,72]嵌入了带有actor-critic算法的MADRL于分布式智能体中。在利用邻居智能体数据方面,文献[73]提出了通过在协作的智能体之间传递值来进行协同优化,每个智能体利用本地和其他智能体的Q值学习最优决策。在交通信号控制的分层控制机制中,文献[56]提出了一种策略,即本地的交叉口智能体控制局部区域的交通,而全局智能体的上层决策受到局部区域优化级

别的影响,从而形成协调控制机制。

2.2.3 MADRL能够有效结合图形化处理方法

由于图神经网络(graph neural network,GNN)是专门设计用于捕捉拓扑关系的模型,所以它能够在捕捉多智能体和多任务关系时展现出强大的功能特性。GNN模型采用消息传递方案,通过将节点的特征信息传递给其邻居节点,并不断迭代传播,直到达到平衡状态。由于MADRL涉及多个智能体,所以将这些智能体之间的关系信息结合到具有GNN架构的模型中可以提高其性能^[74]。文献[75]提出了一种基于深度时空注意力神经网络的MADRL方法,用于解决大规模道路网络中的交通信号控制问题。文献[58]采用了GAT作为DRL模型中的神经网络。文献[76]引入了归纳异构图注意力算法(inductive heterogeneous graph attention,IHA)进行特征融合,并通过多智能体深度图信息最大化(multi-agent deep graph infomax,MDGI)框架进行训练。此外,图卷积网络(graph convolutional network,GCN)可以将多个交叉口表示为网络图形,例如文献[77]提出了基于GCN的RL智能体神经网络结构,能够自动提取具有多个交叉口的路网中道路的交通特征。

2.2.4 MADRL能够有效结合其他领域方法

MADRL与其他领域的研究方法结合来解决交通信号控制问题的方向主要包括:首先,结合人工免疫系统可以解决应急车辆引导问题^[78];其次,利用纳什均衡和博弈论可以辅助actor-critic模型^[79];此外,通过最大熵的正则化方法可以减少排队长度^[80];还可以使用行为克隆方法来解决奖励设计问题^[81];另外,生成对抗网络(generative adversarial network,GAN)可以用于交通数据的恢复^[82]。在交通信号控制中,也有研究考虑元学习的应用,例如将交通信号控制视为一组相关任务上的元学习问题^[83],或引入基于梯度的元学习算法。尽管已经出现了一些使用元学习或模仿学习的研究,但在交通信号控制问题上,通过有限的样本数据进行有效的学习和探索仍然面临较大的挑战。

2.3 大规模交通信号控制的交通场景影响因素

交通场景是各种信号控制方法的承接和载体,而不同的交通场景设置会在很大程度上影响信号控制方法的性能。因此,在信号控制方法的设计中充分考虑交通场景等因素的影响是非常必要的。本节从交通网络规模、交通需求、交通方式和交通数据源四个角度入手,对大规模交通信号控制中交通场景的影响因素进行了分析,并总结了现有研究中针对这些影响因素的解决方案。

2.3.1 交通网络规模

道路网络的规模能够在一定程度上检验算法的性能,一些研究在较大规模的交通网络中进行了模拟。例如,文献[28]在纽约曼哈顿地区的3 791个交通信号灯的真实场景进行了实验。另外,文献[84]将模型迁移到了拥有2 510个信号灯的场景区,还有其他研究分别使用了196个和127个交叉口进行了实验^[85,86]。

2.3.2 交通需求

通常情况下,交通信号控制模型在交通流量较低时能够发挥出较好的性能。然而,在现实世界中,交通流量饱和等情况经常出现,这就对模型的鲁棒性提出了更高的要求。一些研究模拟了高需求、接近饱和、饱和或过饱和等交通状况,以测试所提方法在处理交通拥堵情况中的效果。例如,文献[86~89]分析或提到了所提方法能够有效防止交通溢出。

2.3.3 交通方式

无论是私家车、出租车、公共交通等机动交通,还是行人和自行车等非机动交通,在交通信号控制方案的设计过程中都是

十分重要的对象,控制器应根据不同的控制对象调整控制方法和模式。此外,将公共交通引入模拟环境会对学习和训练过程产生很大影响。所以,除了在研究中普遍涉及的私家车辆外,有些研究考虑将中转优先级下的公共交通的优化用于大规模交通网络信号控制。另外,文献[90]提出了一种基于多模式 DRL 的交通信号控制方法,同时结合了常规交通和公共交通,并最大限度地减少了交叉口的整体延误。相较于机动交通,非机动车在现有研究中考考虑较少。一些研究已将行人纳入优化范围^[45]。由于非机动车的动态变化可能对整个交通系统的效率产生较大影响,所以,机动和非机动车的协调十分重要,应充分考虑两者之间的动态协调。

另外,在某些情境下,不同交通方式下的控制对象之间存在竞争关系。模型中可以整合并有效利用紧急车辆的位置、速度等高分辨率数据和高清晰度视频数据^[91],以确定车辆的优先级。为了满足优先车辆的需求并实现交通流之间的公平性,通常采用改变奖励函数的方式。文献[92]提出了一种强化的交通控制策略,该策略利用多智能体系统开发框架(Java agent development framework, JADE)减少了应急车辆在交叉口等待的时间,同时减少了其他车辆的行驶时间。文献[93]对应急车辆的交通信号进行了早期控制,确保控制器能够在各种情况下快速作出应急响应,并减轻了在冲突方向上对交通效率的负面影响。

2.3.4 交通数据源

智慧交通系统中的数据源主要来自通用检测设备、环路检测器、车辆对车辆(vehicle-to-vehicle, V2V)设备和车辆对基础设施(vehicle-to-infrastructure, V2I)设备以及基础设施对基础设施(infrastructure-to-infrastructure, I2I)设备。联网车辆是交通信号控制中的一个特定类别,它可以与其他车辆或基础设施交换数据,为交通信号控制模型采集和获取环境数据。例如,联网车辆能够向交通信号控制器提供实时的交通状态信息,包括车牌号、车辆位置、速度和时间戳等,以满足交通信号控制器对实时性和适应性的要求。据此,文献[23, 94, 95]可以在全局范围中控制交通灯和车辆,以提高城市交通控制的性能。另外,考虑交通异质性对交通控制的安全性和稳定性非常重要^[96],常规车辆和智能网联车所组成的异质交通流特征对交通控制起着关键影响。

3 讨论与发现

本文在分析了 RL 与 DRL 在交通信号控制领域的重要性和广泛实用性后,在本章提出了本领域所面临的挑战和未来可能极具潜力的研究方向。从提高交通信号控制器性能的角度,总结了以下几个方面:

1) 提高状态和动作以及奖励的有效性、可管理性

针对状态和动作空间的优化,可以考虑采用离散化编码技术,例如 DTSE 方法,将复杂的交通状态进行抽象表示,以便在高维和低维状态之间进行转换。某些研究认为基于图像的强化学习方法在现实世界中应用广泛,尤其是在交通信号控制领域。可以考虑使用真实交通场景中的视频图像等作为数据源,具体细节可参考第 1.1 节。此外,如果仅以固定流量作为定时方案的基础,容易受到流量波动的影响,从而导致控制的可靠性较差。因此,在训练过程中应考虑使用动态交通条件^[23],以获得尽可能丰富的状态空间。这将使训练的交通信号控制器能够适应不同流量条件下的交通情况,提高其鲁棒性和适应性。

2) 研究在线学习方法以提高控制器的适应性

在线学习在交通信号控制领域的 RL 方法中的研究始于

本世纪初^[97]。此后,文献[98]对在线学习过程进行了研究。然而,在交通信号控制问题中,RL 方法的在线学习研究仍然面临一些阻力。其中之一是学习最优策略和优化网络结构中的权重值需要大量的探索时间。此外,由于交通信号控制问题具有复杂且高维的状态数据,缺乏随机探索可能导致陷入局部最优解的问题。为了应对这些问题,可以考虑采用多阶段的在线学习过程,涉及到 RL 方法、权重调整和模糊关系调整等技术。然而,总体而言,目前关于在线学习的研究仍面临着较大的挑战 and 困难。

3) 提高学习效率

RL 的学习过程中存在大量的试错环节,从而导致较高的学习成本。为了提高学习效率,可以考虑采取以下几种方法:

a) 多个控制器之间的信息交换:信息交换有助于协调不同控制器的决策,并提高整个交通网络中的全局奖励。

b) 增强的探索方法:例如,基于模型的探索方法可以创建外部环境的模型,并选择动作以增加对未知状态的探索,以提高学习过程的效率。

c) 基于模型的 RL 方法:尽管无模型的 RL 方法已有了许多成功案例,但这些算法的一个典型限制是样本效率,即需要大量样本才能达到良好的性能。基于模型的 RL 方法已在广泛的应用中取得成功,但对于基于模型的 MARL 方法尚未得到广泛研究。因此,研究基于模型的 MARL 是一个值得探索的方向。

d) 迁移学习方法:重新训练模型以校准参数需要大量时间^[79],而基于迁移的训练模型将在模拟交通中训练的 RL 算法迁移到现实交通中,以提高训练效率,减少在现实世界中出现的错误行为。文献[99]提出了一种多智能体迁移 RL 方法来增强交通信号控制的 MARL 性能,并通过迁移学习提高其泛化能力。

4) 提高系统安全性能

针对交通安全问题,现实世界中交通信号的故障可能导致大规模交通事故,因此 RL 和 DRL 的学习成本可能具有致命性。为了规避 DRL 智能体学习过程中的不必要行为,可以采用风险管理方法。这包括设计一系列规则,将高风险动作排除在可行动作的集合之外。通过初步验证不同动作的风险因素,并随着时间的推移在操作过程中进行保守改进,以最小化学习成本。在交通信号配时中,一些研究提出了一种时差惩罚交通信号配时方法,以保证在系统安全性和效率之间取得平衡^[100~102]。

4 结束语

本文对 RL 和 DRL 在 NTSC 中的应用进行了全面、系统的文献综述。研究的主要目标是广泛调研交通信号控制方法的研究成果,确定交通信号控制领域中所有有价值的文章,并分析这些文章中的数据。基于定性和描述性数据,对现有应用方法的创新性、多样性和挑战性进行了分析。本文阐述了如何将交通信号控制问题适当地表示为 RL 和 DRL 问题,并涵盖了各种典型方法和框架;系统地梳理了 RL 和 DRL 在智能交通信号控制领域的分类和应用,并强调了它们在应对大规模智能交通信号控制挑战时的优势;重点分析了 MARL 和 MADRL 这两种合作框架在大规模交通信号控制领域的实现和应用。最后,本文提出了本领域未来可能面临的挑战和极具潜力的研究方向等一些开放性问题。

参考文献:

[1] 赵祥模,马万经,俞春辉,等. 道路交通控制系统发展与趋势展望

- [J]. 前瞻科技,2023,2(3): 58-66. (Zhao Xiangmo, Ma Wanjiang, Yu Chunhui, *et al.* Development and trend of road traffic control system[J]. *Science and Technology Foresight*, 2023, 2(3): 58-66.)
- [2] Hunt P B, Robertson D I, Bretherton R D, *et al.* SCOOT-a traffic responsive method of coordinating signals, LR 1014 Monograph[R]. Crowthorne: Transport and Road Research Laboratory, 1981.
- [3] Gartner N. OPAC: a demand responsive strategy for traffic signal control[J]. *Transportation Research Record Journal of the Transportation Research Board*, 1983, 906: 75-81.
- [4] Boillot F, Midenet S, Pierrelee J C. The real-time urban traffic control system CRONOS: algorithm and experiments[J]. *Transportation Research Part C: Emerging Technologies*, 2006, 14(1): 18-38.
- [5] Sims A G, Dobinson K W. The Sydney coordinated adaptive traffic (SCAT) system philosophy and benefits[J]. *American Society of Civil Engineers*, 1980, 29(2): 130-137.
- [6] Lu Kai, Lin Guanrong, Xu Jianmin, *et al.* Simultaneous optimization model of signal phase design and timing at intersection[C]// Proc of International Conference on Transportation and Development 2018. Pittsburgh, Pennsylvania: American Society of Civil Engineers, 2018: 65-74.
- [7] 徐明杰, 韩印. 基于粒子群算法下的交叉口信号配时优化[J]. 物流科技, 2020, 43(1): 106-110. (Xu Mingjie, Han Yin. Intersection signal timing optimization based on particle swarm optimization[J]. *Logistics Sci-Tech*, 2020, 43(1): 106-110.)
- [8] 徐东伟, 周磊, 王达, 等. 基于深度强化学习的城市交通信号控制综述[J]. 交通运输工程与信息学报, 2022, 20(1): 15-30. (Xu Dongwei, Zhou Lei, Wang Da, *et al.* A review of urban traffic signal control based on deep reinforcement learning[J]. *Journal of Transportation Engineering and Information*, 2022, 20(1): 15-30.)
- [9] 华赞, 王祥丰, 金博. 面向城市交通信号优化的多智能体强化学习综述[J]. 运筹学学报, 2023, 27(2): 49-62. (Hua Yun, Wang Xiangfeng, Jin Bo. A review of multi-agent reinforcement learning for urban traffic signal optimization[J]. *Operations Research Trans*, 2023, 27(2): 49-62.)
- [10] Noeen M. Reinforcement learning in urban network traffic signal control: a systematic literature review[J]. *Expert Systems with Applications: An International Journal*, 2022, 199: 116830.
- [11] 于泽, 宁念文, 郑燕柳, 等. 深度强化学习驱动的智能交通信号控制策略综述[J]. 计算机科学, 2023, 50(4): 159-171. (Yu Ze, Ning Nianwen, Zheng Yanliu, *et al.* A review of intelligent traffic signal control strategies driven by deep reinforcement learning[J]. *Computer Science*, 2023, 50(4): 159-171.)
- [12] Bazzan A L C. Opportunities for multiagent systems and multiagent reinforcement learning in traffic control[J]. *Autonomous Agents and Multi-Agent Systems*, 2009, 18(3): 342-375.
- [13] Zhao Dongbin, Zhang Zhen, Dai Yujie. Computational intelligence in urban traffic signal control: a survey[J]. *IEEE Trans on Systems Man and Cybernetics*, 2012, 42(4): 485-494.
- [14] Kumar N, Mittal S, Garg V, *et al.* Deep reinforcement learning-based traffic light scheduling framework for SDN-enabled smart transportation system[J]. *IEEE Trans on Intelligent Transportation Systems*, 2022, 23(3): 2411-2421.
- [15] Garg D, Chli M, Vogiatzis G. Deep reinforcement learning for autonomous traffic light control[C]// Proc of the 3rd IEEE International Conference on Intelligent Transportation Engineering. Piscataway, NJ: IEEE Press, 2018: 214-218.
- [16] Cao Miaomiao, Li V O K, Shuai Qiqi. Book your green wave: exploiting navigation information for intelligent traffic signal control[J]. *IEEE Trans on Vehicular Technology*, 2022, 71(8): 8225-8236.
- [17] 朱炉龙. 基于深度强化学习的生态城市交通信号控制研究[D]. 南昌: 华东交通大学, 2023. (Zhu Lulong. Research on deep reinforcement learning for ecological urban traffic signal control[D]. Nanchang: East China Jiaotong University, 2023.)
- [18] Joo H, Lim Y. Traffic signal time optimization based on deep Q-network[J]. *Applied Sciences*, 2021, 11(21): 9850.
- [19] Nan Xiao, Liang Yu, Jin Qiangyu, *et al.* A cold-start-free reinforcement learning approach for traffic signal control[J]. *Journal of Intelligent Transportation Systems*, 2022, 26(4): 476-485.
- [20] Yang Shantian, Bo Yang. A semi-decentralized feudal multi-agent learned-goal algorithm for multi-intersection traffic signal control[J]. *Knowledge-Based Systems*, 2021, 213: 106708.
- [21] Lillicrap T P, Hunt J J, Pritzel A, *et al.* Continuous control with deep reinforcement learning[EB/OL] (2015-09-09). <https://arxiv.org/abs/1509.02971>.
- [22] Bouktif S, Cheniki A, Ouni A. Traffic signal control using hybrid action space deep reinforcement learning[J]. *Sensors*, 2021, 21(7): 2302.
- [23] Mo Zhaobin, Li Wangzhi, Fu Yongjie, *et al.* CVLight: decentralized learning for adaptive traffic signal control with connected vehicles[J]. *Transportation Research Part C: Emerging Technologies*, 2022, 141: 103728.
- [24] Jamil A R M, Ganguly K K, Nower N. Adaptive traffic signal control system using composite reward architecture based deep reinforcement learning[J]. *IET Intelligent Transport Systems*, 2020, 14(14): 2030-2041.
- [25] Wang Shuo, Yue Wenwei, Chen Yue, *et al.* Cooperative learning with difference reward in large-scale traffic signal control[C]// Proc of the 25th International Conference on Intelligent Transportation Systems. Piscataway, NJ: IEEE Press, 2022: 2307-2312.
- [26] Aziz H M A, Zhu Feng, Ukkusuri S V. Learning-based traffic signal control algorithms with neighborhood information sharing: an application for sustainable mobility[J]. *Journal of Intelligent Transportation Systems*, 2018, 22(1): 40-52.
- [27] Higuera C, Lozano F, Camacho E C, *et al.* Multiagent reinforcement learning applied to traffic light signal control[M]// Demazeau Y, Matson E, Corchado J M, *et al.* Advances in Practical Applications of Survivable Agents and Multi-Agent Systems. Cham: Springer, 2019: 115-126.
- [28] Devailly F X, Larocque D, Charlin L. Model-based graph reinforcement learning for inductive traffic signal control[EB/OL]. (2022-08-01) [2023-02-23]. <http://arxiv.org/abs/2208.00659>.
- [29] Rios L M, Sahinidis N V. Derivative-free optimization: a review of algorithms and comparison of software implementations[J]. *Journal of Global Optimization*, 2013, 56(3): 1247-1293.
- [30] Williams R J. Simple statistical gradient-following algorithms for connectionist reinforcement learning[J]. *Machine Learning*, 1992, 8: 229-256.
- [31] 吴少波, 杨薛钰. 基于Sarsa算法的交通信号灯控制方法[J]. 信息与电脑: 理论版, 2021, 33(6): 49-51. (Wu Shaobo, Yang Xueyu. Traffic signal control method based on Sarsa algorithm[J]. *China Computer & Communication*, 2021, 33(6): 49-51.)
- [32] Richter S. Traffic light scheduling using policy-gradient reinforcement learning[J]. *IET Intelligent Transport Systems*, 2017, 11(7): 417-423.
- [33] 杨康康. 基于多智能体深度强化学习的交通信号优先控制研究[D]. 兰州: 兰州理工大学, 2023. (Yang Kangkang. Traffic signal priority control based on multi-agent deep reinforcement learning[D]. Lanzhou: Lanzhou University of Technology, 2023.)
- [34] Lecun Y, Bengio Y, Hinton G. Deep learning[J]. *Nature*, 2015, 521(7553): 436-444.
- [35] Park S, Han E, Park S, *et al.* Deep Q-network-based traffic signal

- control models[J]. *PLoS One*, 2021, 16(9): e0256405.
- [36] Mnih V, Kavukcuoglu K, Silver D, *et al.* Human-level control through deep reinforcement learning[J]. *Nature*, 2015, 518(7540): 529-533.
- [37] Schaul T, Quan J, Antonoglou I, *et al.* Prioritized experience replay [EB/OL]. (2016-02-25) [2023-02-23]. <http://arxiv.org/abs/1511.05952>.
- [38] Jiang Shan, Huang Yufei, Jafari M, *et al.* A distributed multi-agent reinforcement learning with graph decomposition approach for large-scale adaptive traffic signal control[J]. *IEEE Trans on Intelligent Transportation Systems*, 2022, 23(9): 14689-14701.
- [39] Hasselt H. Double Q-learning[C]// *Advances in Neural Information Processing Systems*. Red Hook, NY: Curran Associates Inc., 2010.
- [40] Wang Ziyu, Schaul T, Hessel M, *et al.* Dueling network architectures for deep reinforcement learning[C]// *Proc of the 33rd International Conference on Machine Learning*. New York: PMLR. org, 2016: 1995-2003.
- [41] Mnih V, Badia A P, Mirza M, *et al.* Asynchronous methods for deep reinforcement learning[C]// *Proc of the 33rd International Conference on Machine Learning*. New York: PMLR. org, 2016: 1928-1937.
- [42] 黄浩, 胡智群, 王鲁黔, 等. 基于 Sumtree DDPG 的智能交通信号控制算法[J]. *北京邮电大学学报*, 2021, 44(1): 97-103. (Huang Hao, Hu Zhiqun, Wang Luhan, *et al.* Intelligent traffic signal control algorithm based on Sumtree DDPG[J]. *Journal of Beijing University of Posts and Telecommunications*, 2021, 44(1): 97-103.)
- [43] Shanmugasundaram P, Bhatnagar S. Robust traffic signal timing control using multiagent twin delayed deep deterministic policy gradients [C]// *Proc of the 14th International Conference on Agents and Artificial Intelligence*. Portugal: SciTePress, 2022: 477-485.
- [44] Lowe R, Wu Yi, Tamar A, *et al.* Multi-agent actor-critic for mixed cooperative-competitive environments[C]// *Proc of the 31st Conference on Neural Information Processing Systems*. Red Hook, NY: Curran Associates Inc., 2017: 6382-6393.
- [45] Liu Ying, Liu Lei, Chen Weipeng. Intelligent traffic light control using distributed multi-agent Q learning[C]// *Proc of the 20th International Conference on Intelligent Transportation Systems*. Piscataway, NJ: IEEE Press, 2017: 1-8.
- [46] Hyttinenrauch M, Šošić A, Neumann G. Deep reinforcement learning for swarm systems[J]. *The Journal of Machine Learning Research*, 2019, 20(1): 1966.
- [47] Devailly F X, Larocque D, Charlin L. IG-RL: inductive graph reinforcement learning for massive-scale traffic signal control[J]. *IEEE Trans on Intelligent Transportation Systems*, 2022, 23(7): 7496-7507.
- [48] Cao Y J, Ireson N, Bull L, *et al.* Distributed learning control of traffic signals[M]// Cagnoni S. *Real-World Applications of Evolutionary Computing*. Berlin: Springer, 2000: 117-126.
- [49] Tan Tian, Bao Feng, Deng Yue, *et al.* Cooperative deep reinforcement learning for large-scale traffic grid signal control[J]. *IEEE Trans on Cybernetics*, 2020, 50(6): 2687-2700.
- [50] Zeng Jing, Xin Jie, Cong Ya, *et al.* HALight: hierarchical deep reinforcement learning for cooperative arterial traffic signal control with cycle strategy[C]// *Proc of the 25th International Conference on Intelligent Transportation Systems*. Piscataway, NJ: IEEE Press, 2022: 479-485.
- [51] Yu Chao, Velu A, Vinitsky E, *et al.* The surprising effectiveness of PPO in cooperative multi-agent games[C]// *Advances in Neural Information Processing Systems*. Red Hook, NY: Curran Associates Inc., 2022: 24611-24624.
- [52] Yi Chenglin, Wu Jia, Ren Yanyu, *et al.* A spatial-temporal deep reinforcement learning model for large-scale centralized traffic signal control[C]// *Proc of the 25th International Conference on Intelligent Transportation Systems*. Piscataway, NJ: IEEE Press, 2022: 275-280.
- [53] Oroojlooyjadid A, Hajinezhad D. A review of cooperative multi-agent deep reinforcement learning[J]. *Applied Intelligence*, 2021, 53(11): 13677-13722.
- [54] Gao Ruowen, Liu Zhihan, Li Jinglin, *et al.* Cooperative traffic signal control based on multi-agent reinforcement learning[M]// Zheng Zibin, Dai Hongning, Tang Mingdong, *et al.* *Blockchain and Trustworthy Systems*. Singapore: Springer, 2020: 787-793.
- [55] Paul A, Mitra S. Exploring reward efficacy in traffic management using deep reinforcement learning in intelligent transportation system[J]. *ETRI Journal*, 2022, 44(2): 194-207.
- [56] Zhou Pengyuan, Braud T, Alhilal A, *et al.* ERL: edge based reinforcement learning for optimized urban traffic light control[C]// *Proc of IEEE International Conference on Pervasive Computing and Communications Workshops*. Piscataway, NJ: IEEE Press, 2019: 849-854.
- [57] Wu Qiang, Shen Jun, Yong Binbin, *et al.* Smart fog based workflow for traffic control networks[J]. *Future Generation Computer Systems*, 2019, 97: 825-835.
- [58] Ha P, Chen Sikai, Du Runjia, *et al.* Scalable traffic signal controls using fog-cloud based multiagent reinforcement learning[J]. *Computers*, 2022, 11(3): 38.
- [59] Ibrokhimov B, Kim Y J, Kang S. Biased pressure: cyclic reinforcement learning model for intelligent traffic signal control[J]. *Sensors*, 2022, 22(7): 2818.
- [60] Huo Yusen, Tao Qinghua, Hu Jianming. Cooperative control for multi-intersection traffic signal based on deep reinforcement learning and imitation learning[J]. *IEEE Access*, 2020, 8: 199573-199585.
- [61] 潘昭天. 基于博弈论和多智能体强化学习的城市道路网络交通控制方法研究[D]. 长春: 吉林大学, 2022. (Pan Zhaotian. *Urban road network traffic control method based on game theory and multi-agent reinforcement learning* [D]. Changchun: Jilin University, 2022.)
- [62] 曹立春, 智敏. 基于上下文注意的强化学习目标检测[J]. *计算机应用与软件*, 2023, 40(5): 221-226. (Cao Lichun, Zhi Min. Reinforcement learning object detection based on contextual attention [J]. *Computer Applications and Software*, 2023, 40(5): 221-226.)
- [63] Darmoul S, Elkosantini S, Louati A, *et al.* Multi-agent immune networks to control interrupted flow at signalized intersections [J]. *Transportation Research Part C: Emerging Technologies*, 2017, 82: 290-313.
- [64] Louati A, Louati H, Li Zhaojian. Deep learning and case-based reasoning for predictive and adaptive traffic emergency management[J]. *The Journal of Supercomputing*, 2021, 77(5): 4389-4418.
- [65] Iyer V, Jadhav R, Mavchi U, *et al.* Intelligent traffic signal synchronization using fuzzy logic and Q-learning[C]// *Proc of International Conference on Computing, Analytics and Security Trends*. Piscataway, NJ: IEEE Press, 2016: 156-161.
- [66] Cao Y J, Ireson N, Bull L, *et al.* Design of a traffic junction controller using classifier system and fuzzy logic [C]// *Proc of International Conference on Computational Intelligence*. Berlin: Springer, 1999: 342-353.
- [67] Mikami S, Kakazu Y. Genetic reinforcement learning for cooperative traffic signal control[C]// *Proc of the 1st IEEE Conference on Evolutionary Computation*. Piscataway, NJ: IEEE Press, 1994: 223-228.
- [68] Tahifa M, Boumhidi J, Yahyaouy A. Swarm reinforcement learning for traffic signal control based on cooperative multi-agent framework

- [C]// Proc of Intelligent Systems and Computer Vision. Piscataway, NJ: IEEE Press, 2015: 1-6.
- [69] 陶玉飞. 基于深度强化学习的交叉口交通信号控制研究[D]. 兰州: 兰州理工大学, 2023. (Tao Yufei. Research on intersection traffic signal control based on deep reinforcement learning[D]. Lanzhou: Lanzhou University of Technology, 2023.)
- [70] Chu Tianshu, Wang Jie, Codecà L, *et al.* Multi-agent deep reinforcement learning for large-scale traffic signal control[J]. *IEEE Trans on Intelligent Transportation Systems*, 2019, 21(3): 1086-1095.
- [71] Li Zhenning, Yu Hao, Zhang Guohui, *et al.* Network-wide traffic signal control optimization using a multi-agent deep reinforcement learning[J]. *Transportation Research Part C: Emerging Technologies*, 2021, 125: 103059.
- [72] Baldazo D, Parras J, Zazo S. Decentralized multi-agent deep reinforcement learning in swarms of drones for flood monitoring[C]// Proc of the 27th European Signal Processing Conference. Piscataway, NJ: IEEE Press, 2019: 1-5.
- [73] Ge Hongwei, Song Yumei, Wu Chunguo, *et al.* Cooperative deep Q-learning with Q-value transfer for multi-intersection signal control[J]. *IEEE Access*, 2019, 7: 40797-40809.
- [74] 徐哲扬. 概率图神经网络在交通信号控制中的应用[D]. 成都: 电子科技大学, 2023. (Xu Zheyang. Application of probabilistic graph neural network in traffic signal control[D]. Chengdu: University of Electronic Science and Technology of China, 2023.)
- [75] Huang Hao, Hu Zhiqun, Lu Zhaoming, *et al.* Network-scale traffic signal control via multiagent reinforcement learning with deep spatiotemporal attentive network[J]. *IEEE Trans on Cybernetics*, 2023, 53(1): 262-274.
- [76] Yang Shantian, Yang Bo. An inductive heterogeneous graph attention-based multi-agent deep graph infomax algorithm for adaptive traffic signal control[J]. *Information Fusion*, 2022, 88: 249-262.
- [77] Nishi T, Otaki K, Hayakawa K, *et al.* Traffic signal control based on reinforcement learning with graph convolutional neural nets [C]// Proc of the 21st International Conference on Intelligent Transportation Systems. Piscataway, NJ: IEEE Press, 2018: 877-883.
- [78] Louati A. A hybridization of deep learning techniques to predict and control traffic disturbances [J]. *Artificial Intelligence Review*, 2020, 53(8): 5675-5704.
- [79] Wu Qiang, Wu Jianqing, Shen Jun, *et al.* Distributed agent-based deep reinforcement learning for large scale traffic signal control[J]. *Knowledge-Based Systems*, 2022, 241: 108304.
- [80] Wang Pengyong, Mao Feng, Li Zhiheng. SoftLight: a maximum entropy deep reinforcement learning approach for intelligent traffic signal control[C]// Proc of the 14th International Conference on Advanced Computational Intelligence. Piscataway, NJ: IEEE Press, 2022: 166-175.
- [81] Zhang Huichu, Kafourous M, Yu Yong. PlanLight: learning to optimize traffic signal control with planning and iterative policy improvement [J]. *IEEE Access*, 2020, 8: 219244-219255.
- [82] Wang Zixin, Zhu Hanyu, He Mingcheng, *et al.* GAN and multi-agent DRL based decentralized traffic light signal control[J]. *IEEE Trans on Vehicular Technology*, 2022, 71(2): 1333-1348.
- [83] Zhu Liwen, Peng Peixi, Lu Zongqing, *et al.* Variationally and intrinsically motivated reinforcement learning for decentralized traffic signal control[EB/OL]. (2021-01-04) [2023-02-24]. <http://arxiv.org/abs/2101.00746>.
- [84] Chen Chacha, Wei Hua, Xu Nan, *et al.* Toward a thousand lights: decentralized deep reinforcement learning for large-scale traffic signal control[C]// Proc of AAAI Conference on Artificial Intelligence. Palo Alto, CA: AAAI Press, 2020: 3414-3421.
- [85] Wei Hua, Xu Nan, Zhang Huichu, *et al.* CoLight: learning network-level cooperation for traffic signal control[C]// Proc of the 28th ACM International Conference on Information and Knowledge Management. New York: ACM Press, 2019: 1913-1922.
- [86] Wei Hua, Chen Chacha, Zheng Gaunjie, *et al.* PressLight: learning max pressure control to coordinate traffic signals in arterial network [C]// Proc of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York: ACM Press, 2019: 1290-1298.
- [87] Horsuwan T, Aswakul C. Reinforcement learning agent under partial observability for traffic light control in presence of gridlocks[C]// Proc of SUMO User Conference 2019. [S. l.]: EasyChair, 2019: 29-47.
- [88] Kim D, Jeong O. Cooperative traffic signal control with traffic flow prediction in multi-intersection[J]. *Sensors*, 2019, 20(1): 137.
- [89] Zhao Yi, Ma Jianxiao, Shen Linghong, *et al.* Optimizing the junction-tree-based reinforcement learning algorithm for network-wide signal coordination [J]. *Journal of Advanced Transportation*, 2020, 2020: article ID 6489027.
- [90] Alizadeh S S M, Abdulhai B. Multimodal intelligent deep (MiND) traffic signal controller[C]// Proc of IEEE Intelligent Transportation Systems Conference. Piscataway, NJ: IEEE Press, 2019: 4532-4539.
- [91] Wang Song, Xie Xu, Huang Kedi, *et al.* Deep reinforcement learning-based traffic signal control using high-resolution event-based data[J]. *Entropy*, 2019, 21(8): 744.
- [92] Kristensen T, Ezeora N J. Simulation of intelligent traffic control for autonomous vehicles[C]// Proc of IEEE International Conference on Information and Automation. Piscataway, NJ: IEEE Press, 2017: 459-465.
- [93] Cao Miaomiao, Li V O K, Shuai Qiqi. A gain with no pain: exploring intelligent traffic signal control for emergency vehicles [J]. *IEEE Trans on Intelligent Transportation Systems*, 2022, 23(10): 17899-17909.
- [94] Yang Jiachen, Zhang Jipeng, Wang Huihui. Urban traffic control in software defined Internet of Things via a multi-agent deep reinforcement learning approach[J]. *IEEE Trans on Intelligent Transportation Systems*, 2022, 22(6): 3742-3754.
- [95] Wang Tong, Cao Jiahua, Hussain A. Adaptive traffic signal control for large-scale scenario with cooperative group-based multi-agent reinforcement learning[J]. *Transportation Research Part C: Emerging Technologies*, 2021, 125: 103046.
- [96] Nuli S, Mathew T V. Online coordination of signals for heterogeneous traffic using stop line detection[J]. *Procedia-Social and Behavioral Sciences*, 2013, 104: 765-774.
- [97] Min C C, Srinivasan D, Ruey L C. Cooperative, hybrid agent architecture for real-time traffic signal control [J]. *IEEE Trans on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 2003, 33(5): 597-607.
- [98] Yin Biao, Dridi M, El M A. Adaptive traffic signal control for multi-intersection based on microscopic model[C]// Proc of the 27th International Conference on Tools with Artificial Intelligence. Piscataway, NJ: IEEE Press, 2015: 49-55.
- [99] Ge Hongwei, Gao Dognwan, Sun Liang, *et al.* Multi-agent transfer reinforcement learning with multi-view encoder for adaptive traffic signal control[J]. *IEEE Trans on Intelligent Transportation Systems*, 2022, 23(8): 12572-12587.
- [100] Liang Xiaoyuan, Yan Tan, Lee J, *et al.* A distributed intersection management protocol for safety, efficiency, and driver's comfort [J]. *IEEE Internet of Things Journal*, 2018, 5(3): 1924-1935.
- [101] Liao Lyuchao, Liu Jieru, Wu Xinke, *et al.* Time difference penalized traffic signal timing by LSTM Q-network to balance safety and capacity at intersections[J]. *IEEE Access*, 2020, 8: 80086-80096.
- [102] Guo Jin. Decentralized deep reinforcement learning for network level traffic signal control[D]. California: University of California, Davis, 2020.