

面向城市交通信号优化的多智能体强化学习综述

华贇, 王祥丰, 金博

引用本文:

华贇, 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. Multi-agent deep reinforcement learning-based urban traffic signal management[J]. *Operations Research Transactions*, 2023, 27(2): 49-62.

相似文章推荐 (请使用火狐或IE浏览器查看文章)

Similar articles recommended (Please use Firefox or IE to view the article)

双层规划在城市交通领域研究与应用的系统综述

A systematic review of researches and applications of bi-level programming in the context of urban transport

运筹学学报. 2023, 27(2): 1-26 <https://doi.org/10.15960/j.cnki.issn.1007-6093.2023.02.001>

城市交通流量估计的运筹学方法

Operational research methods for urban traffic flow estimation

运筹学学报. 2023, 27(2): 27-48 <https://doi.org/10.15960/j.cnki.issn.1007-6093.2023.02.002>

低秩矩阵优化若干新进展

Some advances in low-rank matrix optimization

运筹学学报. 2020, 24(2): 23-41 <https://doi.org/10.15960/j.cnki.issn.1007-6093.2020.02.003>

从数值最优化方法到学习最优化方法

From numerical optimization method to learning optimization method

运筹学学报. 2019, 23(4): 1-12 <https://doi.org/10.15960/j.cnki.issn.1007-6093.2019.04.001>

基于类神经网络的城市交通信号区域优化算法

An optimized urban traffic signal field control algorithm based on pseudo neural network

运筹学学报. 2015, 19(3): 71-77 <https://doi.org/10.15960/j.cnki.issn.1007-6093.2015.03.009>

面向城市交通信号优化的多智能体强化学习综述*

华 贇¹ 王祥丰^{1,†} 金 博¹

摘要 随着近年来国民经济水平的快速提高, 人民的出行需求快速增长, 给当前由传统非智能信号控制主导的道路交通信号系统带来了日趋严峻的压力。交通路网复杂程度的显著提升促使交通信号控制从单点问题向系统工程问题发展, 而人工智能技术的兴起, 使得城市交通信号优化有了更多的处理手段。以多智能体强化学习为代表的群体智能方法在最近几年被广泛应用于交通信号控制与优化, 其中包括交通信号灯控制、自动驾驶、车路协同等。多智能体强化学习方法相比于传统方法, 可以赋予交通信号系统智能化的同时实现大规模交通信号系统协作, 以提升城市交通运行效率。未来智慧城市交通愿景下, 参与城市交通的各个部分互相协作是至关重要的, 多智能体强化学习在城市交通信号优化具有极大研究价值。本文将系统介绍面向城市交通信号优化的多智能体强化学习的基本理论及其应用于城市交通信号优化领域的现状, 从智能体协作的角度对已有方法进行归纳, 并分析各类方法优缺点。此外, 本文总结多智能体强化学习方法在城市交通信号优化领域所面临的挑战, 并指出该领域未来潜在研究方向, 以促进多智能体强化学习方法在智能城市交通信号优化领域的发展。

关键词 多智能体强化学习, 智能交通, 交通信号控制, 自动驾驶

中图分类号 O221.2

2010 数学分类号 90C47, 90C26, 90C30

Multi-agent deep reinforcement learning-based urban traffic signal management*

HUA Yun¹ WANG Xiangfeng^{1,†} JIN Bo¹

Abstract With the rapid improvement of the national economy in recent years, people's travel demand has increased, bringing increasingly severe pressure on the current urban traffic signal system relying on traditional non-intelligent traffic lights. The significant increase in the complexity of the traffic network has led to the development of traffic signal control from a single-point problem to a system engineering problem, and the development of artificial intelligence technology brings more methods to dealing with urban traffic signal control. Swarm intelligence methods, represented by multi-agent reinforcement learning, have been widely used in traffic signal control and optimization, including traffic light control, autonomous driving, and vehicle-road collaboration. Compared to traditional methods, multi-agent reinforcement learning can empower the intelligence of traffic signal systems while implementing large-scale traffic signal system collaboration to improve the efficiency of urban traffic operations. The various components involved

收稿日期: 2022-06-12

* 基金项目: 国家重点研发计划 (Nos. 2021YFA1000300, 2021YFA1000302), 国家自然科学基金 (Nos. 12071145, 11971216)

1. 华东师范大学计算机科学与技术学院, 上海 200062; School of Computer Science and Technology, East China Normal University, Shanghai 200062, China

† 通信作者 E-mail: xfwang@cs.ecnu.edu.cn

in urban transportation must collaborate in the vision of intelligent urban traffic. Multi-agent reinforcement learning is of great research value in urban traffic signal control and optimization. This paper will systematically introduce the basic theory of multi-agent deep reinforcement learning and its use in urban traffic signal optimization, summarize the existing approaches and analyze the drawbacks of each method. In addition, this paper will outline the challenges of multi-agent reinforcement learning methods for urban traffic signal optimization. Then the paper points out possible future research directions to promote the development of multi-agent reinforcement learning methods in urban traffic signal optimization.

Keywords multi-agent reinforcement learning, intelligent traffic, traffic signal control, autonomous driving

Chinese Library Classification O221.2

2010 Mathematics Subject Classification 90C47, 90C26, 90C30

近年来,随着国民经济水平的快速增长,城市车辆保有量急剧增加,交通拥堵现象变得越来越普遍,这给城市交通管理带来了巨大的考验,同时也带来了严重的经济损失^[1]。尤其,在交叉路口这一城市交通网络的关键节点和主要瓶颈上,当前的城市交通信号系统依旧依赖于数量庞大的传统非智能信号灯。然而这些传统的非智能信号灯通常基于固定的交通信号控制策略,很难对复杂多变的城市交通状态做出有效的即时反馈来改善城市交通中的突发事件。从这一角度上来说,采用智能化技术,以此来提升交通信号系统的疏导能力,对于缓解城市交通拥堵,弥补由此产生的经济损失,并提升城市智能化水平,具有重要意义。在智能化城市交通系统驱动下,以智能交通信号系统为核心的智能交通系统应运而生 (Intelligent Transportation System)^[2]。其中,预定义时间片交通信号控制方法^[3]以及车辆驱动型交通信号控制方法^[4]逐步成为被广泛使用的智能交通信号控制方法。但其中前者无法很好地应对交通流量中的短期波动,后者又极大地依赖于传感器提供的实时数据和人为制定的交通规则,这导致它们在泛化性和可靠性上均存在一定不足^[5]。

近年来,多智能体强化学习方法依靠深度学习技术的发展,取得了长足的进步,并逐渐地在仓储物流^[6]、多机器人控制^[7]、多人博弈^[8]等领域取得了可喜的成果。而在城市交通信号优化领域,基于多智能体深度强化学习的交通信号系统,相比于预定义时间片的控制方法以及车辆驱动型的控制方法有着巨大的潜在优势。采用多智能体强化学习方法一方面可以实现对路网实时状态做出有利于长期路网优化的即时反馈,另一方面也可以赋予道路交通系统中的每个组成部分 (如机动车、非机动车、红绿灯、行人等) 互相协作的可能性。而随着自动驾驶技术的不断发展,对未来城市交通信号系统设计需要进一步考虑自动驾驶车辆的特殊性。结合多智能体强化学习的自动驾驶决策技术为深入探索“无交通信号”管理提供了可能。多智能体学习方法带来的智能体间的协作能力,使得自动驾驶车辆可以在不依赖于交通信号控制下,通过智能体协作实现自主交通管理,这可以被看作是一种特殊的去中心化交通信号控制优化。与此同时,如何实现自动驾驶车辆与智能交通信号系统之间的互相协作也是一个有趣且值得探讨的问题^[9]。

图 1 归纳了城市交通管理系统智能化的发展历程。本篇综述将在简单梳理早期工作后系统地介绍多智能体强化学习的基本理论及其在城市交通信号优化中的应用现状,将从智能体间协作的角度对已有基于多智能体强化学习的城市交通信号优化方法进行归纳,并分析各类方法的优缺点。本文还将进一步总结现有多智能体强化学习方法在城市交通信号优化领域中的应用挑战以及潜在研究方向。本文安排如下:第 1 节简单介绍一下早

期智能城市交通信号优化策略与方法, 具体包括预定义时间片和规则驱动两种交通信号控制方法; 第 2 节中概括了基于多智能体强化学习的城市交通信号优化的基本概念; 第 3 节从协作的角度把现有的基于多智能体强化学习的城市交通信号优化方法分为两类并进行做简要分析; 第 4 节讨论基于多智能体强化学习的城市交通信号优化方法的未来挑战及潜在研究方向; 第 5 节对全文做一个归纳总结。

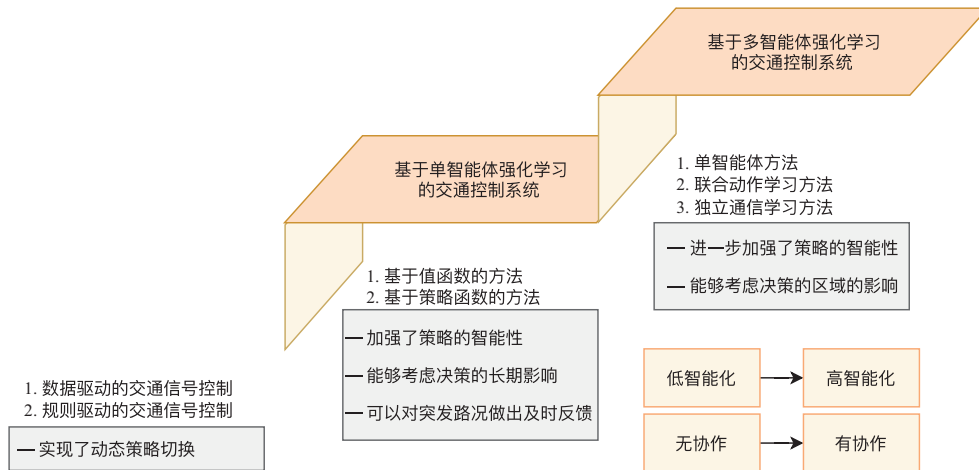


图 1 城市交通管理系统智能化发展历程

1 早期智能城市交通信号优化系统

1.1 基于预定义时间片的交通信号控制

基于预定义时间片的交通信号控制方法是最早出现且被广泛应用的交通信号控制方法^[3]。以交通信号灯为代表的交通信号控制单元每次会从一组预先定义好的时间片分配策略集合中选择策略, 而策略集合是根据平均每年每天流量 (AADP) 离线构建的, 而 AADP 计算的是一年内的每一天, 在控制单元控制区域 (如交叉路口) 内的某一点上通过车辆的平均数量。因此, 基于预定义时间片的交通信号控制方法一定程度上考虑并引入了车流信息的周期性, 其相比于原始固定策略具有更优的灵活性。

1.2 基于规则驱动的交通信号控制

为了针对预定义时间片的交通信号灯控制方法在应对交通流量中的短期波动上的缺点, 后续的研究提出了规则驱动的交通信号灯控制方法^[4]。这是一种由被检测到的车辆信息驱动的交通信号灯控制方法, 其中交通信号系统的策略取决于当前的交通需求, 在实际应用的过程中, 传感器会感知固定时间段内路口的车辆通过情况, 并针对实时的车流量来制定对应策略。例如, 当传感器检测到一定时间内路口某一方向的平均车流量大于某一个特定阈值, 那么这时交通信号灯就会对其他方向切换至红灯状态。这种方法通过路口的实时情况来进行决策, 一定程度上解决了预定义时间片方法在应对交通流量中的短期波动上的不足, 并且在许多情况下取得了很好的结果。但这类方法高度依赖于人为给定

的规则, 这使得算法的性能高度取决于规则制定者的经验和具体的环境设置, 导致其性能非常不稳定。

2 基于多智能体强化学习的城市交通信号优化基本概念

2.1 强化学习建模

基于强化学习的城市交通信号系统控制的基本模型如图 2 所示。首先构建一个马尔可夫决策过程 (Markov Decision Process, MDP) 模型, 该模型通常由一个五元组所定义, 即:

$$\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma \rangle,$$

其中 $\mathcal{S}$ 表示状态空间; $\mathcal{A}$ 表示动作空间; $\mathcal{P}: \mathcal{S} \times \mathcal{A} \rightarrow \delta(\mathcal{S})$ 表示转移函数, 即在状态 $s \in \mathcal{S}$ 下执行动作 $a \in \mathcal{A}$ 后转移到 $s' \in \mathcal{S}$ 的概率; $\mathcal{R}: \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow \mathbb{R}$ 表示奖励空间, 即在状态 $s \in \mathcal{S}$ 下执行动作 $a \in \mathcal{A}$ 转移到 $s' \in \mathcal{S}$ 后得到的奖励; $\gamma \in (0, 1]$ 表示折扣因子, 用来刻画补偿即时奖励和未来奖励的影响。

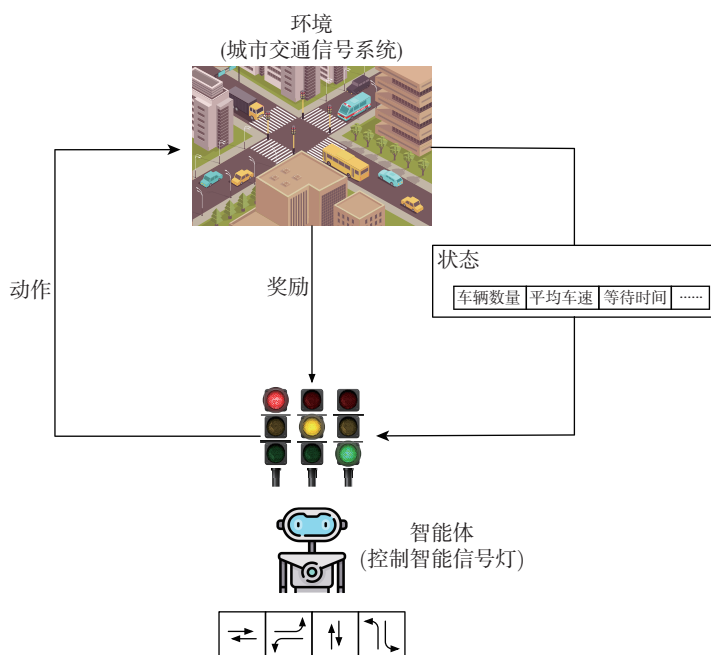


图 2 基于强化学习的城市交通信号系统控制基本模型

对于由强化学习控制的交通信号系统, 路口的信号控制单元可以被抽象为一个智能体 (agent), 其控制的对象可以被定义为控制单元所对应的实时交通流。在控制策略的训练阶段, 智能体通过与环境不断交互, 观察交通环境的实时状态, 并抽象地提取决策所需的当前交通状态信息 $s_t \in \mathcal{S}$, 再进一步根据交通状态信息进行决策, 得到当前状态下的动作 $a_t \in \mathcal{A}$ 并执行该动作, 最后得到交通系统反馈的奖励信息 r_t 。智能体的总体目标可以

被设定为最大化折扣累积奖励并得到对应的最优策略 π^* , 即

$$\max_{\pi} \left\{ \mathbb{E} \left[\sum_{t=0}^T \gamma^t \mathcal{R}(s^t, a^t, s^{t+1}) \mid a^t \sim \pi(\cdot | s_t) \right] \right\}. \quad (1)$$

相比于规则驱动的交通信号控制方法, 基于强化学习的方法不再需要人为定义信号灯的决策逻辑, 而只要定义控制智能信号灯的智能体的一些基本的信息。而后智能体则会自主地筛选信息, 并在不断地交互中学到有效策略。因此在学习目标之外, 使用强化学习方法来解决实际任务的一大主要问题在于定义强化学习智能体。即定义前文所述的状态、动作以及奖励。在表 1, 表 2 和表 3 中我们列举了当前主流的在交通信号控制任务中强化学习智能体在奖励函数、状态信息中所包含的元素以及智能体动作类型。而在这里, 我们将做进一步梳理:

(1) **奖励:** 奖励体现了强化学习智能体的优化目标。在交通信号控制领域, 尽管最终对信号灯策略的优化目标是使得所有车辆的行驶时间最小化, 但车辆的行驶时间很难作为强化学习的有效奖励信息。因为车辆实际的行驶时间往往会受到除信号灯外的环境影响, 这导致使用车辆的行驶时间作为奖励信息难以有效体现当前交通信号策略的质量。因此, 现有的工作通常使用一个可以在智能体执行行动后能有效测量的奖励信号。在近年来应用于交通信号控制领域的一系列强化学习方法中, 奖励函数一般会被设计为车道队列长度、车辆等待时间、延迟等信息的加权线性组合^[15, 22, 29]。然而, 如何设计奖励中各个组成的权重是非常困难的。Wei 等人^[22] 发现奖励中权重设计的细微差异可能会导致完全不同的结果, 因此, 他们设计了一种基于路口交通压力的奖励, 并证明了其等效于优化整体的行驶时间, 并在实验环境中体现出了相对先前奖励设计的优势。Zhao 等人^[23] 在此基础上额外考虑了离开路口的车辆数量, 改进了奖励设置, 取得了更好的结果。

(2) **状态:** 状态体现了强化学习智能体在每个时间步接受到的环境信息。现有的基于强化学习的交通信号控制方法中会建立一个包含车道层面或路段层面信息的张量作为状态信息。状态信息一般会包含车道队列长度^[10,19,25]、车速^[19,30]、车辆等待时间^[12, 22]、当前的信号灯状态^[25, 26] 等道路信息。在早期基于传统强化学习的工作中, 状态信息会被离散化, 并通过简单的表格或线形模型来近似拟合状态信息以提高效率。然而在现实世界中, 道路信息的状态空间通常是非常大的, 这限制了传统强化学习的应用。随着深度学习的发展, 深度神经网络被应用在了强化学习任务中, 并有效地解决了高维度状态空间的表示问题。由此还有部分工作借助深度学习技术, 直接使用与车辆所处位置相关的图像信息来表示状态信息^[15, 27, 28]。

(3) **动作:** 动作体现了强化学习智能体在每个时间步根据当前环境给予的状态信息, 所执行的具体行为。现有的基于强化学习的交通信号控制方法会采用以下几类动作定义: 1) 设置当前阶段的持续时间^[20,25]; 2) 设置阶段分割^[10,11,21]; 3) 改变或保持当前阶段^[13,18,22]; 4) 选择下一阶段^[12,16,31]。动作定义的选择与交通信号的具体设置是密切相关的。举个例子来说, 如果交通信号灯的相位是循环且固定的, 那么就应该选择上述的前三种动作定义。

表 1 奖励函数中包含的元素

元素	参考文献
车道队列长度	[5,10,11,12,13,14,15,16,17]
车辆等待时间	[5,12,13,14,15,18,19,20]
车速	[5,15,21]
停车次数	[15]
是否发生事故	[15]
路口交通压力	[22, 23]

表 2 状态信息中包含的元素

元素	参考文献
车道队列长度	[5,10,11,12,13,14,17,19,24]
车辆等待时间	[5,12]
车速	[19,21]
当前信号灯状态	[5,13,16,17,22,24,25,26]
路口图像信息	[15, 27, 28]

表 3 智能体动作类型

动作类型	参考文献
设置当前阶段的持续时间	[20,25]
设置阶段分割	[10, 11, 21]
改变或保持当前阶段	[13,18,22]
选择下一阶段	[12,16,31]

为了寻找最优交通信号控制策略 π^* , 强化学习方法通常可以归纳为如下两种:

(1) 基于值函数 (Value-Based) 的方法: 以 Q -学习 (Q -Learning) 方法为代表^[32], 旨在学习最优状态-动作值估计函数以间接寻找最优策略, Q -学习方法维持一个状态-动作值函数 $\hat{Q}(s, a)$, 通过智能体与环境交互获得历史数据 (s, a, s', r) , 并使用如下形式更新状态-动作值函数, 即

$$\hat{Q}(s, a) \leftarrow (1 - \alpha) \hat{Q}(s, a) + \alpha \left[r + \gamma \max_{a'} \hat{Q}(s', a') \right], \quad (2)$$

其中 α 是学习率。在一定假设条件下, Q -学习可以被证明能收敛到最优状态-动作值估计函数。而且基于深度学习的深度 Q -学习则可以解决有着复杂动作状态空间的任務, 并已经被广泛应用到离散动作场景中。在智能交通信号控制与优化领域, 基于值函数的方法主要应用于采用改变或保持当前阶段、选择下一阶段这两类动作空间。

(2) 基于策略函数 (Policy-Based) 的方法: 以策略梯度方法^[33] 为代表, 直接在策略空间中搜索找到最优策略。策略梯度方法对于参数化的策略 $\pi^\theta(a|s)$, 根据使用该策略采集到的轨迹 τ , 策略梯度可以表示为:

$$\nabla_\theta J(\pi^\theta) := \mathbb{E}_{\tau \sim \pi^\theta} \left[\sum_{t=0}^T \nabla_\theta \log \pi^\theta(a_t | s_t) G(\tau) \right], \quad (3)$$

其中 $G(\tau)$ 是所采样到轨迹 τ 的折扣累计奖励。随着深度学习技术的发展和引入, PPO^[34]、

DDPG^[35]、TRPO^[36] 等一系列基于策略函数的方法被不断地提出并应用于连续动作场景中。对于面向智能交通信号控制与优化问题, 基于策略函数的方法通常采用设置当前阶段的持续时间、设置阶段分割这两类动作空间。

2.2 基于多智能体强化学习的交通信号优化建模

城市交通系统通常是由路口、机动车、行人、非机动车等城市交通参与者组成的复杂多智能体系统。因此, 可以将“单智能体”强化学习方法驱动的交通信号控制与优化拓展为多智能体协作, 甚至可以拓展到信号系统与智能车辆之间协作、智能车辆间协作等, 体现出了多智能体强化学习方法在城市交通信号优化领域的研究价值。从数学形式上来说, 多智能体强化学习方法需要考虑到多个智能体同时与环境交互、智能体之间交互并学习自身策略, 因此一般建模为去中心化部分可观测马尔可夫决策过程 (Decentralized Partially-Observed Markov Decision Process, Dec-POMDP)。Dec-POMDP 通常由一个多元组所定义, 即

$$\langle \mathcal{I}, \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \omega, \mathcal{O}, \gamma \rangle,$$

其中 $\mathcal{I} := 1, \dots, N$ 表示 N 个智能体; $s \in \mathcal{S}$ 表示智能体系统状态; 每个智能体 $i \in \mathcal{I}$ 接收到的局部观测信息表示为 $o_i \in \omega$, 而智能体 i 的局部观测信息由观测函数 $\mathcal{O}(s, i)$ 根据状态产生; 智能体 i 根据观测信息及自身策略 $\pi_i(a_i | o_i)$ 选择一个动作 $a_i \in \mathcal{A}_i$, 并最终构建所有智能体的联合动作 $a \in \mathcal{A}$, 从而促使环境根据转移状态函数 $\mathcal{P}(s' | s, a)$ 从状态 s 转移到下一个状态 s' ; 环境将反馈给所有智能体一个共享的全局奖励 $r = \mathcal{R}(s, a)$; 智能体的优化目标也是最大化折扣全局累积奖励, 并得到最优联合策略 π^* :

$$\arg \max_{\pi} \left\{ G_{tot} = \mathbb{E}_{a^t \sim \pi(\cdot | s^t)} \left[\sum_{t=0}^T \gamma^t R(s^t, a^t) \mid s^0 = s, a^0 = a \right] \right\}. \quad (4)$$

在基于多智能体强化学习的城市交通信号系统控制与优化方法中, 所涉及区域中的每一个交通信号控制单元, 以及包含智能决策能力的机动车 (如自动驾驶汽车) 都可以被看作智能体。多智能体强化学习方法可以认为是这些智能体之间的随机博弈过程, 而多智能体强化学习方法的目标是促进交通系统中的智能体群体可以联动协作, 以达到交通系统区域整体最优^[37]。因此, 如何实现不同智能体之间的联动协作, 以提升多智能体系统的整体决策能力, 是协作多智能体强化学习方法的核心任务。从智能体协作角度理解, 基于多智能体强化学习的城市交通信号优化方法可以进行如下分类: 1) 独立多智能体强化学习驱动的城市交通信号优化; 2) 协作多智能体强化学习驱动的城市交通信号优化; 在后续章节中, 我们将从这两个方向以协作角度归纳总结已有的基于多智能体强化学习的城市交通信号优化方法。

3 基于多智能体强化学习的城市交通信号优化方法

3.1 独立多智能体强化学习驱动的城市交通信号优化

独立多智能体强化学习驱动的城市交通信号优化是最基本的交通优化强化学习方法。这类方法在一定程度上适用于孤立且不属于大型城市路网的道路交通系统的管理 (如小

城镇路网), 并可以得到优于传统方法的效果^[5], 但由于其并未考虑到不同控制单元之间的协作, 在较复杂的城区大规模路况中性能受限^[13]。

早在 2011 年, Abdoos 等人^[10] 将独立多智能体 Q -学习方法直接应用于一个基于 Aimsun 模拟器构建的包含 50 个路口和 112 条道路的模拟城市交通路网场景中。策略部署实验结果发现独立 Q -学习方法即使在非稳定的多智能体环境下, 依然可以在给定不同交通需求下获得超过传统基于预定义固定时间片方法。Aslani 等人^[38] 引入独立多智能体演员-评论家算法, 也得到类似结果。值得注意的是, Wei Hua 等人在文献 [5] 中提出的 IntelliLight 方法首次将深度强化学习方法应用于城市路网级别的交通信号控制问题中, 此时深度强化学习方法带来的表征能力使得 IntelliLight 可以处理多变的交通情况如周末、工作日等中不一致的交通需求。随后, 他们进一步提出了 PressLight^[22], 其引入 Max Press 技术构建基于交通压力的奖励函数, 以实现相邻路口之间的协作, 并实现类似于绿波带的交通信号控制与优化效果。

独立多智能体强化学习驱动的城市交通信号优化虽适应大规模城市交通系统优化, 但是缺乏对智能体间协作的建模, 不能充分实现交通信号控制单元之间协作。后续越来越多基于协作多智能体强化学习的城市交通信号优化方法被提出。

3.2 协作多智能体强化学习驱动的城市交通信号优化

协作多智能体强化学习驱动的城市交通信号优化方法旨在显式考虑多个交通信号控制单元之间的信号协作, 目前大多数协作多智能体强化学习驱动的城市交通信号优化方法都只关注信号灯之间的同构协作关系, 没有考虑到道路交通中的其他智能体, 其他智能体参与的多智能体协作问题与方法我们将专门在第 4 节中讨论。从实际操作层面上来说, 协作多智能体强化学习驱动的城市交通信号优化问题存在两种主要的建模形式, 即联合动作学习和独立通信学习。

3.2.1 联合动作学习

联合动作学习方法是一种直接实现同构协作多智能体强化学习控制的城市交通信号优化方法, 已经被广泛应用于交通信号优化任务中^[14,29]。在基于多智能体强化学习的交通信号优化任务中, 联合动作学习方法将整个城市交通路网中的所有路口的信号灯构建为一个统一智能体, 以学习优化控制所有信号灯的策略。从形式上来说, 对于有 n 个信号灯控制场景, 统一智能体的动作可以表示为 $\mathbf{a} = [\mathbf{a}_0, \mathbf{a}_1, \dots, \mathbf{a}_n]$, 其中观测信息可以表示为 $\mathbf{o} = [\mathbf{o}_0, \mathbf{o}_1, \dots, \mathbf{o}_n]$ 。值得注意的是, 随着统一智能体需要控制路口数量的增加, 状态-动作空间会指数级增长, 从而导致维度爆炸, 所以联合动作学习方法在大规模任务中难以学习到有效策略。为了解决维度爆炸问题, 多智能体强化学习领域一般通过将统一智能体的策略, 或值函数分解给每个独立智能体, 以此来近似统一智能体的策略或值函数。而在协作多智能体强化学习驱动的城市交通信号优化任务中, 这类分解方法也被广泛使用来解决由于路口增多而导致的维度爆炸问题。如 Elise Van der Pol 等人^[15] 就将 max-plus 技术引入并拓展基于值函数的联合动作学习方法, 将统一智能体的全局 Q 函数分解为如下形式的各个独立智能体的局部 Q 函数的求和组合, 即

$$\hat{Q}(\mathbf{o}, \mathbf{a}) = \sum_{i,j} Q_{i,j}(\mathbf{o}_i, \mathbf{o}_j, \mathbf{a}_i, \mathbf{a}_j), \quad (5)$$

其中 i, j 表示相邻交通路口对应的智能体索引。他们的工作在小规模的虚拟场景中表现出了优于基准算法的性能, 但并没有应用在大规模的城市交通场景中。此外, 这一方法在训练过程中表现得非常不稳定, 这限制了其在复杂场景中的应用。Zhi Zhang 等人^[39] 提出的 QCOMBO 则进一步将统一智能体的全局 Q 函数视为各个独立智能体的局部 Q 函数的加权求和, 即

$$\hat{Q}(\mathbf{o}, \mathbf{a}) = \sum_{i,j} w_{i,j} Q_{i,j}(\mathbf{o}_i, \mathbf{o}_j, \mathbf{a}_i, \mathbf{a}_j), \quad (6)$$

其中 $w_{i,j}$ 为预定义的权重, 通过对损失函数加入正则项的方式来估计全局 Q 函数以逼近实际的全局 Q 函数, 从而达到近似联合动作学习的目标。他们在实验中尝试了较大规模的虚拟场景, QCOMBO 除了在性能上表现出色, 也体现出了在不同规模和路况的交通场景中的泛化性能, 增强了算法的实用价值。最近, Chen 等人^[40] 提出的 CQmix 方法则利用了 QMIX 方法, 通过神经网络结构来聚合各个独立智能体的局部 Q 函数以此近似统一智能体的全局 Q 函数, 此外 CQmix 还考虑了各个独立智能体之间的通信, 以此进一步提升性能。他们在实验中尝试了基于真实道路交通路网和数据的虚拟场景, 表现出了优于简单应用 QMIX 算法的性能。

3.2.2 独立通信学习

独立通信学习是另一种实现协作多智能体强化学习驱动的城市交通信号优化的方法, 其旨在在独立学习的基础上为每个智能体构建通信消息以利用该通信消息实现协作。通信带来了其他智能体的额外信息, 因此可以拓展增强接受这些信息的智能体有限的局部观测^[41], 使其在动态环境下贴近实际环境状态。

Xu 等人^[42] 和 Ge 等人^[43] 的工作分别将交通状况和历史决策信息加入到交通信号系统的决策中, 这也是常规的独立通信学习的方法。因为独立学习的过程中大规模智能体会共享同一个学习模型, 所以这些方法都需要对相邻交通信号控制单元建立一致的索引。其中, Xu 等人^[42] 的工作中引入了相邻信号灯控制的路口的车辆数量及平均车速信息将其纳入了观测信息, 而 Ge 等人^[43] 的工作则把智能体训练过程中相邻信号灯对应的智能体的局部 Q 函数作为通信信息纳入了优化过程中进行联合优化。从结果上看, 在简单的模拟环境中, 他们的方法都显著地提升了交通信号控制的策略质量, 体现出了协作的优势。

然而, 上述的这些方法都没有考虑到各个节点之间的物理位置关系, 只是简单地将所有邻居节点视为同质的智能体。图卷积神经网络 (Graph Convolutional Network) 作为一种新颖的深度学习模型, 被 Michael Schlichtkrull 等人^[44] 引入来构建城市路网结构, 而 Tomoki Nishi 等人^[19] 则进一步利用基于图卷积神经网络构建出的路网关系, 实现了相邻交通信号控制单元之间的通信, 然而这类方法依然假设了各个相邻智能体之间的影响是静态的。这一静态影响假设在城市交通信号优化中存在明显的局限性, 在实际城市交通信号优化场景中, 相邻信号单元之间 (如相邻路口信号灯之间) 的影响是会随着时间的变化而动态变化的。因此, 基于动态影响假设, Hua Wei 等人^[24] 和 Yanan Wang 等人^[45] 分别提出使用图注意力网络 (Graph Attention Network) 来学习相邻信号单元之间的动态相互影响。其中, Hua Wei 等人提出的 Colight 通过图注意力网络学习包含邻居智能体的路口状态信息来拓展智能体的观测, 而 Yanan Wang 等人提出的 STMARL 则是以此训练相邻信号灯对应的智能体的局部 Q 函数的权重, 实现智能体间基于通信的协作。从结果上

看, Colight 和 STMARL 在基于真实道路数据的模拟环境中都表现出了明显的优势, 提升了不同智能体控制的信号灯间的协作能力。

4 未来的挑战与发展

前文介绍了当前面向城市交通信号优化的多智能体强化学习工作, 从协作角度对当前的方法做了分类并分析了它们的优缺点。在本章节中, 我们将进一步探讨面向城市交通信号的多智能体强化学习工作面临的新的挑战, 并在此基础上讨论一些未来的研究方向。

4.1 采样效率

强化学习方法需要频繁地与环境进行交互, 通过试错法来收集数据并以此优化自身策略。对于多智能体强化学习来说, 由于智能体之间存在的相互影响导致的环境不稳定性, 使得训练有效的智能体需要更多的交互。而在交通信号优化领域中, 这类频繁的试错会带来现实世界中难以承受的严重交通堵塞。因此, 采样效率问题是将多智能体强化学习应用在城市交通信号优化领域的一大主要挑战。当前大多数基于多智能体强化学习的城市交通信号优化方法都是基于模拟器环境进行训练的, 因为使用模拟环境可以更容易且快速进行交互, 所以采样效率并没有在训练过程中带来严重的问题, 但这又会导致从模拟器环境到现实场景中的领域迁移问题。当前主要有两类方法被用来解决城市交通信号优化任务中的采样效率问题。其一是模仿学习 (Imitation Learning), Yuanhao Xiong 等人^[16]首次将模仿学习方法应用在了城市交通信号优化任务中, 并基于三座城市中的交通信号决策及车流数据, 有效地学习到了先验策略, 再通过这一先验策略, 显著提高了演员-评论家算法在交通信号优化任务中的采样效率, 并提高了交通信号决策的质量, 进一步缓解了拥堵。另一类方法是元强化学习 (Meta Reinforcement Learning), 其旨在从一系列任务中提取元信息, 并以此来促进新任务的训练以提高采样效率。Xinshi Zang 等人^[17]提出的 MetaLight 首先把元强化学习引入了城市交通信号优化领域, MetaLight 针对不同环境或路况下基于 DQN 方法的信号灯学习了一个元值函数参数, 并将其应用到新任务中提高新任务的采样效率。Xinghuai Huang 等人^[46]提出的 ModelLight 则采用了基于模型的元强化学习方法构建了一个对环境的先验估计并以此提升新任务中的采样效率。而 Chang Liu 等人^[47]提出的 GeneraLight 则通过元学习方式对车流进行建模, 并在此基础上学习元策略, 相比于之前的方法, GeneraLight 在泛化性上表现出色, 可以有效提升采样效率。然而, 现有的工作对于场景的泛化能力依然存在局限性, 往往只考虑了纯车辆的场景, 这与实际交通规划中包含了车辆、行人与非机动车的混合场景存在一定的差异。此外, 当前提升采样效率的方法都只考虑到了单个智能体的情况, 即假定每个路口的智能信号灯之间是相互独立的, 因而, 如何在提升泛化性和采样效率的过程中考虑不同路口间的协作, 也是一个有意义的问题。

4.2 模拟器到现实场景的迁移

正如前一节所说, 当前大多数基于多智能体强化学习的城市交通信号优化方法都依赖于模拟环境并以此减少采样效率问题带来的影响。然而模拟器和现实环境之间的差异 (如传感器延迟、噪声、驾驶员行为模式) 限制了这些方法在现实的交通信号优化中的应用。有些工作, 如 James Ault 等人^[48]尝试构建一种可解释的策略, 使得其可以更容易地

被应用在真实场景中。另一个角度是建立一个更加贴近现实的模拟场景,这些工作的重心在于如何从有限的真实交通数据中构建出拟真的模拟交通环境。如 Hua Wei 等人^[49] 尝试从稀疏的车辆位置信息中模仿出真实交通场景中的车辆轨迹,而 Guanjie Zheng 等人在文献 [50] 中通过真实场景中的车辆示教数据构建了模拟场景中的车辆轨迹,而在文献 [51] 中尝试从道路通行速度信息中重构出统计意义上的车辆的起点和终点。当前的这些方法一定程度上可以减小模拟器和现实环境之间的差异带来的策略迁移上的困难。但对基于多智能体强化学习的城市交通信号优化方法来说,如何有效地解决模拟器到现实场景策略的迁移依然是一个开放问题。

4.3 与自动驾驶车辆的协作

随着近年来自动驾驶技术的发展,如何使得自动驾驶车辆这类智能车辆参与交通信号优化的过程成为了一类新问题。有自动驾驶技术参与的基于多智能体强化学习方法的交通信号优化任务可以分为如下三类。第一类是自动驾驶车辆作为交通信号灯传感器的延伸,通过物联网技术使得车辆与交通信号灯之间建立通信并传递车辆信息给交通信号灯,提升交通信号灯的决策质量。Qiang Wu 等人^[52] 基于物联网技术,建立了存在车辆-信号灯间通信的面向城市交通信号优化的多智能体强化学习框架,在模拟环境中表现出了显著的性能提升。第二类是建立包含智能交通信号灯和自动驾驶车辆的异构多智能体强化学习方法,以此解决城市交通信号优化问题,但当前鲜有工作涉及这一角度,Praveen Palanisamy^[9] 提出的 MACAD-Gym 模拟环境中考虑到了这类车辆-信号灯间协作的情形。而第三类工作则是让自动驾驶车辆通过多智能体强化学习方法实现不依赖于城市交通信号灯的自主的交通管理,如 Alessandro Paolo Capasso 等人^[53] 就尝试使用多智能体强化学习方法使得自动驾驶车辆间实现自主协作的路口通行。然而,现有的与自动驾驶车辆进行协作的方法主要还是集中在第一类即通过通信来拓展感知能力上,对实现真正的车辆-信号灯间的协作以进一步优化城市交通信号策略的性能还有很长的距离。

5 总 结

本篇综述论文系统介绍了多智能体强化学习方法在交通信号优化领域的应用,从智能体间协作的角度对现有方法做了一定程度上的总结和归纳,并分析了这些方法的优缺点。最后,本文还总结了多智能体强化学习方法在城市交通信号优化领域的挑战,并提出未来潜在研究方向,以期可以帮助相关领域内的研究者,并促进多智能体强化学习方法在智能城市交通领域的发展。

参 考 文 献

- [1] 周大可,唐慕尧,杨欣. 一种结合状态预测的深度强化学习交通信号控制方法 [J]. 计算机应用研究, 2022, **39**(8): 2311-2315.
- [2] Lin Y, Wang P, Ma M. Intelligent transportation system (ITS): Concept, challenge and opportunity [C]//*International Conference on Big Data Security on Cloud*, 2017.
- [3] Dion F, Rakha H, Kang Y. Comparison of delay estimates at under-saturated and over-saturated pre-timed signalized intersections [J]. *Transportation Research Part B: Methodological*, 2004, **38**(2): 99-122.

- [4] Porche I, Lafortune S. Adaptive look-ahead optimization of traffic signals [J]. *Journal of Intelligent Transportation System*, 1999, **4**(3-4): 209-254.
- [5] Wei H, Zheng G, Yao H, Li Z. IntelliLight: A reinforcement learning approach for intelligent traffic light control [C]//*SIGKDD*, 2018.
- [6] Li W, Chen H, Jin B, et al. Multi-agent path finding with prioritized communication learning [C]//*ICRA*, 2022.
- [7] Kober J, Bagnell J, Peters J. Reinforcement learning in robotics: A survey [J]. *The International Journal of Robotics Research*, 2013, **32**(11): 1238-1274.
- [8] Lample G, Chaplot D. Playing FPS games with deep reinforcement learning [C]//*AAAI*, 2017.
- [9] Palanisamy P. Multi-agent connected autonomous driving using deep reinforcement learning [C]//*IJCNN*, 2020.
- [10] Abdoos M, Mozayani N, Bazzan A. Traffic light control in non-stationary environments based on multi agent Q-learning [C]//*ITSC*, 2011.
- [11] Balaji P, German X, Srinivasan D. Urban traffic signal control using reinforcement learning agents [J]. *IET Intelligent Transport Systems*, 2010, **4**(3): 177-188.
- [12] 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.
- [13] Mannion P, Duggan J, Howley E. An experimental review of reinforcement learning algorithms for adaptive traffic signal control [M]//*Autonomic Road Transport Support Systems*, 2016: 47-66.
- [14] Prashanth L, Bhatnagar S. Reinforcement learning with average cost for adaptive control of traffic lights at intersections [C]//*ITCS*, 2011.
- [15] Van der Pol E, Oliehoek F. Coordinated deep reinforcement learners for traffic light control [C]//*NeurIPS*, 2016.
- [16] Xiong Y, Zheng G, Xu K, et al. Learning traffic signal control from demonstrations [C]//*CIKM*, 2019.
- [17] Zang X, Yao H, Zheng G, et al. Metalight: Value-based meta-reinforcement learning for traffic signal control [C]//*AAAI*, 2020.
- [18] Brys T, Pham T, Taylor M. Distributed learning and multi-objectivity in traffic light control [J]. *Connection Science*, 2014, **26**(1): 65-83.
- [19] Nishi T, Otaki K, Hayakawa K, et al. Traffic signal control based on reinforcement learning with graph convolutional neural nets [C]//*ITSC*, 2018.
- [20] Xu L, Xia X, Luo Q. The study of reinforcement learning for traffic self-adaptive control under multiagent markov game environment [J]. *Mathematical Problems in Engineering*, 2013, **8**: 1-10.
- [21] Casas N. Deep deterministic policy gradient for urban traffic light control [J]. 2017, arXiv: 1703.09035.
- [22] Wei H, Chen C, Zheng G, et al. Presslight: Learning max pressure control to coordinate traffic signals in arterial network [C]//*SIGKDD*, 2019.
- [23] Zhao C, Hu X, Wang G. PDLight: a deep reinforcement learning traffic light control algorithm with pressure and dynamic light duration [J]. 2020, arXiv: 2009.13711.
- [24] Wei H, Xu N, Zhang H, et al. Colight: Learning network-level cooperation for traffic signal control [C]//*CIKM*, 2019.

- [25] Aslani M, Mesgari M, Wiering M. Adaptive traffic signal control with actor-critic methods in a real-world traffic network with different traffic disruption events [J]. *Transportation Research Part C: Emerging Technologies*, 2017, **85**: 732-752.
- [26] Chen C, Wei H, Xu N, et al. Toward a thousand lights: Decentralized deep reinforcement learning for large-scale traffic signal control [C]//*AAAI*, 2020.
- [27] Coşkun M, Baggag A, Chawla S. Deep reinforcement learning for traffic light optimization [C]//*ICDM*, 2018.
- [28] Gao J, Shen Y, Liu J, et al. Adaptive traffic signal control: Deep reinforcement learning algorithm with experience replay and target network [J]. 2017, arXiv: 1705.02755.
- [29] Wiering M. Multi-agent reinforcement learning for traffic light control [C]//*ICML*, 2000.
- [30] El-Tantawy S, Abdulhai B. An agent-based learning towards decentralized and coordinated traffic signal control [C]//*ITSC*, 2010.
- [31] Arel I, Liu C, Urbanik T, et al. Reinforcement learning-based multi-agent system for network traffic signal control [J]. *IET Intelligent Transport Systems*, 2010, **4**(2): 128-135.
- [32] Watkins C, Dayan P. Q-learning [J]. *Machine Learning*, 1992, **8**(3): 279-292.
- [33] Sutton R, McAllester D, Singh S, et al. Policy gradient methods for reinforcement learning with function approximation [C]//*NeurIPS*, 1999.
- [34] Schulman J, Wolski F, Dhariwal P, et al. Proximal policy optimization algorithms [J]. 2017, arXiv: 1707.06347.
- [35] Lillicrap T, Hunt J, Pritzel A, et al. Continuous control with deep reinforcement learning [J]. 2015, arXiv: 1509.02971.
- [36] Schulman J, Levine S, Abbeel P, et al. Trust region policy optimization [C]//*ICML*, 2015.
- [37] 唐建华. 强化学习及其在城市交通信号控制中的应用研究 [D]. 西安: 西安电子科技大学, 2012.
- [38] Aslani M, Mesgari M, Seipel S, et al. Developing adaptive traffic signal control by actor-critic and direct exploration methods [J]. *Transport*, 2019, **172**(5): 289-298.
- [39] Zhang Z, Yang J, Zha H. Integrating independent and centralized multi-agent reinforcement learning for traffic signal network optimization [C]//*AAMAS*, 2020.
- [40] Chen X, Xiong G, Lv Y, et al. A collaborative communication-Qmix approach for large-scale networked traffic signal control [C]//*ITSC*, 2021.
- [41] Sukhbaatar S, Szlam A, Fergus R. Learning multiagent communication with backpropagation [C]//*NeurIPS*, 2016.
- [42] Xu M, Wu J, Huang L, et al. Network-wide traffic signal control based on the discovery of critical nodes and deep reinforcement learning [J]. *Journal of Intelligent Transportation Systems*, 2020, **24**(1): 1-10.
- [43] Ge H, Song Y, Wu C, et al. Cooperative deep q -learning with q -value transfer for multi-intersection signal control [J]. *IEEE Access*, 2019, **7**: 40797-40809.
- [44] Schlichtkrull M, Kipf T, Bloem P, et al. Modeling relational data with graph convolutional networks [C]//*European Semantic Web Conference*, 2018.
- [45] 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.
- [46] Huang X, Wu D, Jenkin M, et al. ModelLight: model-based meta-reinforcement learning for traffic signal control [J]. 2021, arXiv: 2111.08067.
- [47] Zhang H, Liu C, Zhang W, et al. Generalight: Improving environment generalization of traffic signal control via meta reinforcement learning [C]//*CIKM*, 2020.

-
- [48] Ault J, Hanna J, Sharon G. Learning an interpretable traffic signal control policy [C]//*AAMAS*, 2020.
 - [49] Wei H, Chen C, Liu C, et al. Learning to simulate on sparse trajectory data [C]//*ECML*, 2021.
 - [50] Zheng G, Liu H, Xu K, et al. Learning to simulate vehicle trajectories from demonstrations [C]//*ICDE*, 2020.
 - [51] Zheng G, Liu C, Wei H, et al. Rebuilding city-wide traffic origin destination from road speed data [C]//*ICDE*, 2021.
 - [52] Wu Q, Zhi P, Wei Y, et al. Communicate with traffic lights and vehicles based on multi-agent reinforcement learning [C]//*CSCWD*, 2021.
 - [53] Capasso A, Maramotti P, Dell'Eva A, et al. End-to-End intersection handling using multi-agent deep reinforcement learning [C]//*2021 IEEE Intelligent Vehicles Symposium*, 2021.