

多智能体深度确定性策略梯度在城市交通信号控制中的应用综述

杨超^{1*} 唐维²

YANG Chao TANG Wei

摘要

随着城市化进程的加速和交通流量的持续增长,交通信号灯控制已成为缓解城市交通拥堵的关键技术手段。多智能体深度确定性策略梯度(multi-agent deep deterministic policy gradient, MADDPG)算法作为深度强化学习领域的重要突破,为解决多路口交通信号灯协同控制问题提供了新的理论基础和技术路径。文章系统梳理了MADDPG算法在交通信号灯控制领域的研究现状,深入分析了该算法的核心理论机制和在交通信号控制中的具体应用。通过与传统控制方法、单智能体强化学习方法以及其他多智能体方法的详细对比分析,全面评估了MADDPG算法在交通信号灯控制中的性能表现和应用优势。研究表明,MADDPG算法在处理多路口协同控制、动态环境适应和复杂交通场景优化方面具有显著优势,为智能交通信号控制系统的发展提供了重要的理论支撑和实践指导。

关键词

多智能体深度确定性策略梯度;交通信号灯控制;深度强化学习;多智能体系统;智能交通

doi: 10.3969/j.issn.1672-9528.2026.02.039

0 引言

城市交通信号灯控制系统是现代城市交通管理的核心组成部分,其控制策略的优劣直接影响城市交通的通行效率、安全性和环境影响。据统计,全球范围内因交通拥堵造成的经济损失每年超过2.8万亿美元,其中交通信号灯控制不当是导致拥堵的重要因素之一^[1]。传统的交通信号灯控制方法主要包括固定时序控制、感应控制和自适应控制等,这些方法在处理复杂多变的交通环境时存在明显局限性,难以实现多路口间的有效协同和动态优化。近年来,深度强化学习技术的快速发展为交通信号灯控制提供了新的解决思路。特别是多智能体深度确定性策略梯度(multi-agent deep deterministic policy gradient, MADDPG)算法,由于其在处理多智能体协同决策问题方面的独特优势,在交通信号灯控制领域展现出巨大潜力。MADDPG算法将每个交通路口建模为独立的智能体,通过分布式学习和协同决策实现全网交通信号的优化控制,为解决传统控制方法难以处理的多目标、多约束优化问题提供了有效途径^[2]。本文旨在系统梳理MADDPG算法在交通信号灯控制中的应用现状,深入分析该算法的理论基础和技术特点,通过详细的对比研究全面评估其在交通信号控制中的优势和局限性,为相关研究和实际

应用提供理论参考和技术指导。

1 MADDPG算法在交通信号灯控制中的理论基础

1.1 交通信号灯控制问题的多智能体建模

在城市交通网络中,交通信号灯控制问题天然具有多智能体特征。每个交通路口可以被视为一个独立的决策单元,需要根据当前的交通状况选择合适的信号相位和时长^[3]。同时,相邻路口之间存在交通流的相互影响,需要通过协同控制来实现全网优化。MADDPG算法将交通信号灯控制问题建模为多智能体马尔可夫决策过程(multi-agent markov decision process, MA-MDP)。在这一框架下,交通网络中的每个路口被建模为一个智能体,整个系统可以表示为元组 $\langle N, S, A, P, R, \gamma \rangle$,其中:

(1) $N=\{1, 2, \dots, n\}$ 为智能体集合,对应交通网络中的 n 个路口;

(2) $S=S_1 \times S_2 \times \dots \times S_n$ 为联合状态空间,包含所有路口的交通状态信息;

(3) $A=A_1 \times A_2 \times \dots \times A_n$ 为联合动作空间,涵盖所有路口的信号控制策略;

(4) $P: S \times A \rightarrow \Delta(S)$ 为状态转移概率函数,描述交通系统的动态演化;

(5) $R=R_1 \times R_2 \times \dots \times R_n$ 为奖励函数,反映各路口的控制目标;

(6) $\gamma \in [0,1)$ 为折扣因子,平衡短期和长期收益。

1. 中晟软件股份有限公司 山东济南 250000

2. 浪潮卓数大数据产业发展有限公司 江苏无锡 214135

在交通信号灯控制场景中，每个智能体的状态空间通常包括：当前路口的车辆队列长度、各方向的车流密度、等待时间、当前信号相位、相位剩余时间等。动作空间则包括信号相位的选择和持续时间的确定。奖励函数设计需要综合考虑通行效率、等待时间、环境影响等多个优化目标^[4]。

1.2 MADDPG 算法核心机制

MADDPG 算法的核心创新在于采用“集中式训练，分布式执行”的策略架构。这种设计巧妙地解决了多智能体强化学习中的非平稳性问题，同时保证了系统的可扩展性和实用性。在训练阶段，每个智能体的 Critic 网络可以访问全局状态信息和所有智能体的动作信息，从而能够准确评估当前策略的价值。这种集中式的价值评估机制有效克服了多智能体环境中由于其他智能体策略变化导致的环境非平稳性问题。在执行阶段，每个智能体仅依赖自身的局部观测信息进行决策，保证了系统的分布式特性和良好的可扩展性^[5]。对于交通信号灯控制而言，这意味着每个路口的信号控制器只需要处理本路口的交通信息，无需实时获取全网状态，极大降低了通信成本和计算复杂度。算法的核心组件包括 Actor 网络和 Critic 网络。每个智能体 i 的 Actor 网络 $\pi_i(s_i|\theta_i)$ 负责根据当前状态选择最优动作，其策略梯度更新公式为：

$$\nabla_{\theta_i} J(\theta_i) = \mathbb{E}_{s \sim p^{\pi}} \left[\nabla_{\theta_i} \pi_i(s_i) \nabla_{a_i} Q_i^{\pi}(s, a_1, \dots, a_N) |_{a_i = \pi_i(s_i)} \right] \quad (1)$$

Critic 网络 $Q_i(s, a_1, \dots, a_N|a_i)$ 负责评估在给定全局状态和动作组合下的期望回报，通过最小化时间差分误差进行更新为：

$$L(\phi_i) = \mathbb{E}_{(s, a, r, s') \sim D} \left[\left(Q_i^{\phi_i}(s, a) - y_i \right)^2 \right] \quad (2)$$

式中： $y_i = r_i + \gamma Q_i(s', a_1', \dots, a_n')$ 为目标值。

1.3 在交通信号灯控制中的应用特点

将 MADDPG 算法应用于交通信号灯控制时，需要针对交通领域的特殊性进行适应性设计：

(1) 状态表示设计：交通信号灯控制的状态表示需要充分反映路口的交通状况。典型的状态特征包括：各进口道的车辆队列长度、车流密度、平均等待时间、当前信号相位及剩余时间、历史交通流模式等。为增强智能体间的协同能力，状态表示中还需要包含邻接路口的关键信息摘要。

(2) 动作空间设计：交通信号灯的控制动作具有离散性和连续性的双重特征。离散性体现在信号相位的选择上，如东西直行、南北直行、左转相位等；连续性体现在相位持续时间的确定上。因此，MADDPG 算法在交通信号控制中通常采用混合动作空间的设计方案^[6]。

(3) 奖励函数设计：交通信号灯控制的奖励函数需要综合考虑多个优化目标，包括通行效率、等待时间、燃料消耗、环境影响、公平性等。合理的奖励函数设计对于算法的学习效果至关重要。

2 对比研究分析

2.1 与传统交通信号控制方法的对比

2.1.1 固定时序控制 (Fixed-Time Control)

固定时序控制是最基础的交通信号控制方法，采用预先设定的时序方案，各信号相位按照固定的时长循环切换。这种方法具有实现简单、成本低廉、可靠性高等优点，在交通流量相对稳定的场景中表现尚可。然而，固定时序控制的主要缺陷在于无法适应实时交通变化。在交通流量波动较大的情况下，固定的时序方案往往导致资源配置不合理，造成某些方向长时间等待而其他方向通行能力闲置。与 MADDPG 算法相比，固定时序控制缺乏学习和适应能力，无法根据历史数据和实时状况优化控制策略^[7]。实验对比结果显示，在动态交通环境中，MADDPG 算法相比固定时序控制在平均等待时间、通行效率等关键指标上具有显著优势。特别是在交通流量变化较大的高峰时段，MADDPG 算法优势更加明显。

2.1.2 感应控制 (Actuated Control)

感应控制通过在路口安装车辆检测器，根据实时检测到的车辆信息动态调整信号时序。相比固定时序控制，感应控制具有一定的自适应能力，能够在一定程度上响应交通流量的变化。感应控制的主要优势在于能够避免空放现象，即在某个方向没有车辆时及时切换信号相位^[8]。然而，感应控制的反应机制相对简单，通常只能处理单个路口的局部优化，难以实现多路口间的协同控制。此外，感应控制的决策规则相对固化，缺乏学习和进化能力。与 MADDPG 算法相比，感应控制在处理复杂交通场景时存在明显不足。MADDPG 算法通过深度神经网络的强大表达能力，能够学习更复杂的控制策略，同时通过多智能体协同机制实现全网优化。在多路口协同控制的场景中，MADDPG 算法的优势尤为突出^[9]。

2.1.3 自适应控制系统 (SCOOT/SCATS)

SCOOT (split cycle offset optimization technique) 和 SCATS (sydney coordinated adaptive traffic system) 是目前应用较为广泛的自适应交通信号控制系统。这些系统基于实时交通数据和预设的优化算法，能够动态调整信号时序以适应交通流量的变化。自适应控制系统的优势在于具有较强的实时性和一定的协同能力，能够处理多路口间的协调控制。然而，这些系统主要依赖专家知识设计的规则和算法，缺乏自主学习能力。在面对新的交通模式或突发情况时，系统的适应能力有限。MADDPG 算法与自适应控制系统的主要区别在于学习机制^[10]。MADDPG 算法能够通过环境的持续交互自主学习最优控制策略，具有更强的适应性和泛化能力。在复杂多变的交通环境中，MADDPG 算法能够发现传统方法难以捕捉的控制模式，实现更优的控制效果。

2.2 与单智能体强化学习方法的对比

2.2.1 深度 Q 网络 (DQN)

DQN 是较早应用于交通信号控制的深度强化学习方法，

通过深度神经网络逼近 Q 函数,实现端到端的控制策略学习。DQN 在单路口控制中表现良好,能够根据交通状况自主学习最优的信号控制策略。然而,DQN 在处理多路口协同控制时存在明显不足。由于 DQN 将每个路口视为独立的学习问题,无法有效建模路口间的相互影响和协同关系。在多路口网络中,各路口的独立优化往往导致局部最优而非全局最优。与 DQN 相比,MADDPG 算法的主要优势在于多智能体协同机制。通过智能体间的信息共享和协同学习,MADDPG 算法能够实现全网优化,在多路口协同控制方面具有显著优势。实验结果表明,在复杂的交通网络中,MADDPG 算法的性能显著优于独立的 DQN 方法^[11]。

2.2.2 策略梯度方法 (A3C/PPO)

A3C(asynchronous advantage actor-critic)和 PPO(proximal policy optimization)等策略梯度方法通过直接优化策略函数实现控制策略学习。这些方法在处理连续动作空间方面具有优势,适合交通信号控制中相位时长的连续调整。策略梯度方法的主要优势在于能够处理高维动作空间和连续控制问题^[12]。然而,在多智能体环境中,这些方法面临严重的非平稳性问题。由于其他智能体策略的不断变化,环境对于单个智能体而言是非平稳的,这导致学习过程不稳定,收敛困难^[13]。MADDPG 算法通过集中式训练机制有效解决了多智能体环境的非平稳性问题。在训练阶段,Critic 网络能够访问全局信息,提供稳定的价值评估;在执行阶段,Actor 网络仅依赖局部信息,保证了系统的分布式特性。这种设计使得 MADDPG 算法在多智能体交通信号控制中具有更好的稳定性和收敛性^[14]。

2.2.3 Actor-Critic 方法

传统的 Actor-Critic 方法结合了价值函数和策略函数的优势,在单智能体强化学习中表现出色。Actor 网络负责策略选择,Critic 网络负责价值评估,两者相互配合实现高效的策略学习^[15]。在单智能体交通信号控制中,Actor-Critic 方法能够有效处理连续动作空间和复杂状态表示。然而,在多智能体环境中,传统 Actor-Critic 方法面临与策略梯度方法类似的非平稳性问题。MADDPG 算法可以视为 Actor-Critic 方法在多智能体环境中的扩展^[16]。通过让 Critic 网络访问全局状态和动作信息,MADDPG 算法克服了传统 Actor-Critic 方法在多智能体环境中的局限性。同时,MADDPG 算法保持了 Actor-Critic 方法的核心优势,如处理连续动作空间的能力和较高的学习效率^[17]。

2.3 与其他多智能体强化学习方法的对比

2.3.1 独立学习 (Independent Learning)

独立学习是最简单的多智能体强化学习方法,将多智能体问题分解为多个独立的单智能体问题。每个智能体独立学习自己的策略,不考虑其他智能体的存在和影响。独立学习的主要优势在于实现简单,计算复杂度低,易于扩展到大规模

系统。在交通信号控制中,独立学习意味着每个路口独立优化自己的信号控制策略,不考虑与其他路口的协同^[18]。然而,独立学习的主要缺陷在于无法利用智能体间的协同效应。在交通网络中,相邻路口之间存在密切的交通流关联,独立优化往往导致局部最优而非全局最优^[19]。此外,独立学习面临严重的非平稳性问题,因为每个智能体的学习过程都会影响其他智能体的环境。与独立学习相比,MADDPG 算法的主要优势在于协同机制。通过集中式训练,MADDPG 算法能够学习智能体间的协同策略,实现全网优化。实验结果表明,在多路口交通信号控制中,MADDPG 算法的性能显著优于独立学习方法。

2.3.2 多智能体 Q 学习 (Multi-Agent Q-Learning)

多智能体 Q 学习是 Q 学习算法在多智能体环境中的扩展,通过维护联合动作-价值函数实现多智能体协同学习。理论上,多智能体 Q 学习能够找到最优的联合策略。

然而,多智能体 Q 学习面临状态-动作空间指数增长的问题^[20]。随着智能体数量的增加,联合动作空间呈指数增长,导致算法在大规模系统中难以应用。在交通信号控制中,即使是中等规模的交通网络也会产生巨大的联合动作空间,使得传统多智能体 Q 学习方法不可行^[21]。MADDPG 算法通过函数逼近和分布式执行有效解决了维度灾难问题。每个智能体维护自己的 Actor 和 Critic 网络,避免了联合动作空间的指数增长。同时,通过集中式训练机制,MADDPG 算法仍能有效实现多智能体协同。

2.3.3 合作多智能体强化学习 (QMIX/COMA)

QMIX 和 COMA 等合作多智能体强化学习方法专门针对完全合作任务设计,通过特殊的网络结构和训练机制实现高效的协作学习。这些方法在某些合作任务中表现出色,如多智能体游戏和机器人协作等。然而,这些方法通常假设智能体间完全合作,所有智能体共享相同的全局奖励函数。在交通信号控制中,虽然整体目标是优化全网交通,但不同路口的局部优化目标可能存在冲突。例如,某个路口的最优策略可能会增加相邻路口的交通压力^[22]。MADDPG 算法采用竞争-合作的混合机制,更适合交通信号控制的实际场景。每个智能体有自己的局部奖励函数,同时通过协同机制实现全网优化。这种设计既保证了局部合理性,又实现了全局优化。

3 性能评估与应用效果

3.1 仿真实验结果分析

大量仿真实验验证了 MADDPG 算法在交通信号控制中的有效性。在标准的交通仿真平台(如 SUMO、VISSIM)上进行的对比实验表明,MADDPG 算法在多个关键性能指标上显著优于传统方法和其他强化学习方法。在等待时间方面,MADDPG 算法相比固定时序控制平均减少 35%~45%,相比感应控制减少 20%~30%,相比独立 DQN 减少 15%~25%。在通行效率方面,MADDPG 算法的网络吞吐

量提升 20%~35%，平均行程时间减少 15%~25%。特别值得注意的是，MADDPG 算法在复杂交通场景中的优势更加明显。在高峰时段、交通事故、特殊事件等复杂情况下，MADDPG 算法表现出更强的适应性和鲁棒性^[23]。

3.2 实际应用案例

虽然 MADDPG 算法在交通信号控制中的大规模实际部署仍然有限，但一些小规模的试点应用已经验证了其实用性。在某些城市的局部区域试点中，MADDPG 算法表现出良好的控制效果，在保证安全的前提下有效改善了交通状况。这些实际应用案例不仅验证了算法的有效性，也揭示了实际部署中面临的挑战，如传感器数据质量、通信延迟、系统稳定性等问题。

4 结论

多智能体深度确定性策略梯度（MADDPG）算法在交通信号灯控制领域展现出巨大的应用潜力。通过系统的理论分析和详细的对比研究，本文全面评估了 MADDPG 算法在交通信号控制中的优势和特点。研究表明，MADDPG 算法在处理多路口协同控制、动态环境适应和复杂交通场景优化方面具有显著优势。相比传统控制方法，MADDPG 算法具有更强的学习能力和适应性；相比单智能体强化学习方法，MADDPG 算法能够实现有效的多智能体协同；相比其他多智能体方法，MADDPG 算法在可扩展性和实用性方面表现更优。然而，MADDPG 算法在交通信号控制中的应用仍面临一些挑战，如训练复杂度高、对数据质量要求严格、实际部署的工程化难度大等。未来的研究需要在保持算法优势的同时，进一步提高其实用性和可靠性，推动智能交通信号控制技术的实际应用和产业化发展。本综述为交通信号控制领域的研究者和实践者提供了系统性的理论指导和技术参考，有助于推动 MADDPG 算法在智能交通系统中的进一步发展和应用。

参考文献：

- [1] 张春阳, 席雅琪. 人工智能在城市交通控制中的应用综述 [J]. 信息系统工程, 2024(7):33-36.
- [2] 张磊. 人工智能在城市交通信号控制中的应用研究综述 [J]. 人民公交, 2024(10):37-39.
- [3] 周斌, 吴晓东, 马东方, 等. 深度强化学习在城市交通信号控制方法中的应用综述 [J]. 现代交通与冶金材料, 2022, 2(3): 84-93.
- [4] 武强. 多智能体强化学习在城市交通信号控制中的研究与应用 [D]. 兰州: 兰州大学, 2020.
- [5] 刘泽宇. 智能体技术在城市交通信号控制系统中的应用 [J]. 数字通信世界, 2019(8):226.
- [6] 孙宇帆. 探讨智能体技术在城市交通信号控制系统中的应用 [J]. 山东工业技术, 2019(7):142.

- [7] 蓝志洋. 智能体技术在城市交通信号控制系统中的应用 [J]. 交通世界, 2018(26):9-10.
- [8] 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.
- [9] 马家妮, 武雅丽. 智能体技术在城市交通信号控制系统中应用综述 [J]. 科技风, 2015(13):109.
- [10] 杨文臣, 张轮, 施弈骋, 等. 智能体技术在城市交通信号控制系统中应用综述 [J]. 武汉理工大学学报 (交通科学与工程版), 2014, 38(4):709-718.
- [11] 朱理. 克隆选择算法在城市道路交通信号控制中的应用研究 [D]. 湘潭: 湘潭大学, 2010.
- [12] 杨建华. 遗传算法的改进及其在城市交通信号优化控制中的应用研究 [D]. 西安: 长安大学, 2007.
- [13] 王中苏. PLC 在城市道路交通信号控制系统中的应用 [J]. 仪表技术与传感器, 2003(6):36-38.
- [14] 韩如文. 多相位控制在城市交通信号控制中的应用 [J]. 交通与运输, 2001(1):15-17.
- [15] 迟新利, 梁首发, 姜海龙. PLC 在城市道路交通信号控制中的应用 [J]. 沈阳电力高等专科学校学报, 2000(3):32-34.
- [16] 郑国荣, 谭骥元, 郭伟伟, 等. 城市交通信号控制系统实验教程 [M]. 北京: 化学工业出版社, 2022.
- [17] 谭骥元, 郑国荣, 郭伟伟, 等. 城市交通信号控制系统原理及实现 [M]. 北京: 化学工业出版社, 2022.
- [18] 唐慕尧. 基于深度强化学习的城市交通信号控制研究 [D]. 南京: 南京航空航天大学, 2022.
- [19] 季毓婷. 基于深度强化学习的城市道路交通自适应信号控制方法 [D]. 长春: 吉林大学, 2024.
- [20] 孙俊杰. 基于深度强化学习的城市交通信号控制策略研究 [D]. 沈阳: 东北大学, 2022.
- [21] 王崇国, 李云. 5G 通信技术在城市轨道交通信号车地通信中的应用探讨 [J]. 数字通信世界, 2022(9):130-132.
- [22] 高鑫. 铁路信号与城市交通信号比较 [J]. 智能建筑与智慧城市, 2022(9):160-162.
- [23] 熊江勇, 俞竞伟, 聂睿瑞. 智能交通系统概论 [M]. 南京: 南京大学出版社, 2023.

【作者简介】

杨超 (1989—), 通信作者 (email: bingorchao@163.com), 男, 山东济南人, 本科, 研究方向: 电子信息。

唐维 (1991—), 男, 山东济南人, 本科, 研究方向: 电子信息。

(收稿日期: 2025-07-16 修回日期: 2026-01-28)