文章编号:1671-4598(2024)07-0154-08 DOI:10.16526/j.cnki.11-4762/tp.2024.07.023 中图分类号:TP393.083 文献标识码:A

基于 Dueling Double DQN 的交通信号控制方法

叶宝林[1,2], 陈 栋[1,2], 刘春元[2], 陈 滨[2], 吴维敏[3]
(1. 浙江理工大学 信息科学与工程学院, 杭州 310018;
2. 嘉兴大学 信息科学与工程学院 嘉兴市智慧交通重点实验室, 浙江 嘉兴 314001;
3. 浙江大学 智能系统与控制研究所, 杭州 310027)

摘要: 为了提高交叉口通行效率缓解交通拥堵, 深入挖掘交通状态信息中所包含的深层次隐含特征信息, 提出了一种基于 Dueling Double DQN (D3QN) 的单交叉口交通信号控制方法; 构建了一个基于深度强化学习 Double DQN (DDQN) 的交通信号控制模型, 对动作-价值函数的估计值和目标值迭代运算过程进行了优化, 克服基于深度强化学习 DQN 的交通信号控制模型存在收敛速度慢的问题; 设计了一个新的 Dueling Network 解耦交通状态和相位动作的价值, 增强 Double DQN (DDQN) 提取深层次特征信息的能力; 基于微观仿真平台 SUMO 搭建了一个单交叉口模拟仿真框架和环境, 开展仿真测试; 仿真测试结果表明, 与传统交通信号控制方法和基于深度强化学习 DQN 的交通信号控制方法相比, 所提方法能够有效减少车辆平均等待时间、车辆平均排队长度和车辆平均停车次数, 明显提升交叉口通行效率。

关键词: 交通信号控制; 深度强化学习; Dueling Double DQN; Dueling Network

Traffic Signal Control Method based on Dueling Double DQN

YE Baolin[1,2], CHEN Dong[1,2], LIU Chunyuan[2], CHEN Bin[2], WU Weimin[3]
(1. School of Information Science and Engineering, Zhejiang Sci-Tech University, Hangzhou 310018, China;
2. School of Information Science and Engineering, Jiaxing Key Laboratory of Smart Transportations, Jiaxing University, Jiaxing 314001, China;
3. Institute of Cyber-Systems and Control, Zhejiang University, Hangzhou 310027, China)

Abstract: In order to improve the efficiency of intersection traffic, alleviate traffic congestion, and deeply explore the deep hidden feature information contained in traffic status information, a single intersection traffic signal control method based on Dueling double DQN (D3QN) is proposed. A traffic signal control model based on deep reinforcement learning double DQN (DDQN) was constructed, and the iterative operation process of the target value and estimated value of the action value function was optimized to overcome the problem of slow convergence speed in the traffic signal control model based on deep reinforcement learning DQN. A new Dueling network was designed to decouple the value of traffic states and phase actions, enhancing the ability of the double DQN (DDQN) to extract the deep level feature information. On the basis of the micro simulation platform simulation of urban mobility (SUMO), single intersection simulation framework and environment were built to simulate the test. The simulation test results show that compared with traditional traffic signal control methods and traffic signal control methods based on the deep reinforcement learning DQN, the proposed method can effectively reduce the average waiting time, average queue length, and mean stops of vehicles, significantly improving the efficiency of intersection traffic.

Keywords: traffic signal control; deep reinforcement learning; Dueling double DQN; Dueling network

0 引言

据高德地图大数据统计, 城市日均浮动车位置数据的记录超过 200 亿条。同比 2022 年 Q1, 2023 年 Q1 全国 50 个主要城市中有 8%的城市路网高峰行程延时指数下降, 26%的城市基本持平, 66%的城市拥堵上升。交通拥堵导致出行成本增加阻碍了社会经济的发展。据 2022 年中国移动源环境管理年报统计, 2021 年全国机动车四项污染物排放总量为 1 557.7 万吨, 汽车是污染物排放总量的主要贡献者, 其排放的 CO、HC、NO_x 和 PM 超过各污染物总量的 90%。交通拥堵导致资源浪费, 环境污染, 对社会经济造成巨大的损失。造成交通拥堵的原因主要有 2 个方面, 一个是高峰时段居民的交通通行需求激增且超出了当前路网

收稿日期:2023-11-26; **修回日期**:2024-01-15。
基金项目:浙江省自然科学基金项目 (LTGS23F030002); 嘉兴市应用性基础研究项目(2023AY11034),浙江省尖兵领雁研发攻关计划项目(2023C01174),国家自然科学基金项目(61603154);浙江省自然科学基金项目 (LY19F030014);工业控制技术国家重点实验室开放课题 (ICT2022B52)。
作者简介:叶宝林(1984-),男,博士,副教授,硕导。
陈 栋(1998-),男,硕士。
引用格式:叶宝林,陈 栋,刘春元,等.基于 Dueling Double DQN 的交通信号控制方法[J].计算机测量与控制,2024,32(7):154-161.

所通行能力的上限，另一个是当前的交通管控技术还不够智能化难以对路网交通流进行有效管控和引导。受限于城市空间资源和公共财政投入，一般不会通过新建高架或对现有路网进行改扩建来提升路网通行能力缓解交通拥堵。因此，利用最新的人工智能技术方法开发新的交通信号控制方法，提高城市路网交通管控水平来缓解交通拥堵，具有重要的研究意义。

单交叉口交通信号控制是城市交通信号控制系统的基础，也是实现区域交通信号控制的基础[1]。国内外学者对孤立交叉口的交通信号控制的优化进行了深入的研究，以最小化所有车辆通过交叉口的行驶时间为目标的交通信号控制模型得到广泛应用。传统交通信号控制方法主要包括 3 类：基于历史流量的定时控制[2]、基于实时流量的自适应控制[3]（或感应控制[4]）和基于模型的协调控制[5-6]。经过了长时间的应用和验证，虽然传统交通信号控制方法在缓解道路拥堵提高路网通行效率方面取得了不少有价值的成果，但因为复杂多变的车流导致其控制效果不尽如人意，未能最大程度提升交叉口的通行能力。因此，研究人员一直在尝试将人工智能技术引入交通信号控制领域，以期进一步提升城市交通管理系统的智能化水平。

随着人工智能技术的快速发展，交通信号控制与强化学习（RL）技术的结合为提高交通系统的通行效率、安全性和可持续性提供了一种新的解决方案。强化学习是一种“试错”的学习方法，通过不断试错、积累经验，反复学习的机制以期望得到最优策略。在交通信号控制中，结合强化学习的智能体负责优化交叉口的最优配时方案。受到深度学习发展的影响，利用神经网络强大的拟合能力，能够将底层特征组合成更加抽象的高维特征，从而更加有效地分析交叉口交通状态信息。首先，智能体观察交叉口的交通状态信息构建特征矩阵。其次，多角度分析矩阵信息并根据已经学习到的经验制定该交叉口的配时方案。最后，智能体执行配时方案后，将接收到该交叉口反馈的奖励。智能体通过不断试错和学习，使交叉口反馈的奖励最大化，达到优化信号配时方案的目的。为了让交通信号控制更加智能，文献［7］提出了一种基于 Q-Learing 的交通信号控制方法，利用一张 Q 表存储不同交通状态和动作的 Q 值，以挑选最大 Q 值的形式选择下一步动作。该方法虽然能够提高交叉口通行效率，但常因交通状态多变而导致 Q 表维度爆炸。为了适应多变的交通环境，缓解 Q 表爆炸的问题，文献［8］提出了一种基于深度 Q 网络（Deep Q Network）的交通信号控制方法，利用神经网络映射 Q 值。该方法虽然极大程度的缓解了维度灾难的问题，但是容易产生过估计的问题。文献［9］提出一种基于双 Q 网络的交通信号控制方法，改变 Q 值的迭代方式，解决了 Q 值过估计问题，提升了系统的稳定性。文献［10］提出一种 3DRQN 的交通信号控制策略，引入了 LSTM 网络编码历史状态信息，提高了算法的鲁棒性。但该算法按照固定的相序调整相位持续时间，难以快速切换交通压力最大的相位。文献［11］提出了一种基于注意力机制的深度强化学习交通信号控制方案，引入注意力机制，使得神经网络自动关注重要的状态分量，增强网络的感知能力。但该方法只考虑车辆和相位的共同价值，不能发挥车辆和相位独有的价值。

因此，本文针对如何提取交通状态中车辆和相位深层价值的问题，提出一种基于 Dueling Double DQN（D3QN）的单交叉口交通信号控制方法。实验结果表明，本文所提方法能够明显提升交叉口通行效率。本文工作的主要贡献包括：

1）构建了一个新的基于深度强化学习 Double DQN（DDQN）的交通信号控制模型，通过优化动作－价值函数的估计值和目标值迭代运算过程，缓解了传统 DQN 模型中的动作－价值过估计问题。

2）为了增强 Double DQN（DDQN）提取深层次交通特征信息的能力，设计了一个新的 Dueling Network 解耦交通状态和相位动作的价值。另外，为了提升强化学习模型的学习能力，构建了 2 种相位动－作选择方案，设计了一种基于车辆平均排队长度的奖惩机制。

3）设计了一个基于韦伯分布的车辆随机生成模型，用于模拟不同时段的车流分布。基于微观交通仿真平台 SUMO（simulation of urban mobility），构搭建交叉口模拟仿真框架和环境，设计仿真测试方案，验证了本文所提方法的有效性。

1　基于深度强化学习的交通信号控制模型

在本研究中，基于微观交通仿真软件 SUMO 搭建一个包含 12 条车道的典型十字交叉口仿真环境。如图 1 所示，各个进口方向包含有 3 条车道：左转车道，直行车道，右转和直行车道。该交叉口包含 4 个相位：南北直行（车道 1、2、7、8），南北左转（车道 3、9），东西直行（车道 4、5、10、11），东西左转（车道 6、12），其中右转车辆不受交通信号控制。上述 4 个相位以固定的相序进行切换，且不同相位间进行切换时有 3 秒黄灯过渡时间。

交通信号控制问题可以用一个典型的马尔科夫决策动态过程（MDP，markov decision process）进行描述。譬如考虑马尔科夫动态过程 MDP＝（S，A，P，R，γ）[12]，其中 S 为状态空间，是一个描述交叉口交通状态信息的集合，$s\in S$ 表示某个时刻的交通状态信息。A 为动作空间，表示交通信号控制相位库中的所有相位构成的一个集合，$a\in A$ 表示相位库中的某一个相位。P 为状态转移概率，表示交叉口在状态 s_i 下执行 a_i 后状态变化成 s_{i+1} 的概率，即 $P(s_{i+1}\mid s_i, a_i)$。、R 表示奖励集合，$r_{s,a}$表示在 s 状态执行动作 a 后，环境反馈的奖励，即 $r_{s,a}\in R$。γ 表示在 s_i 状态执行动作 a_i 后对 s_{i+1}的影响程度[13]。

1.1　状态空间

在强化学习中，智能体在每一个决策时刻根据监测的环境状态选择下一步待执行动作，所选动作的好坏与状态包含的环境信息有着密不可分的关系。状态空间的定义对

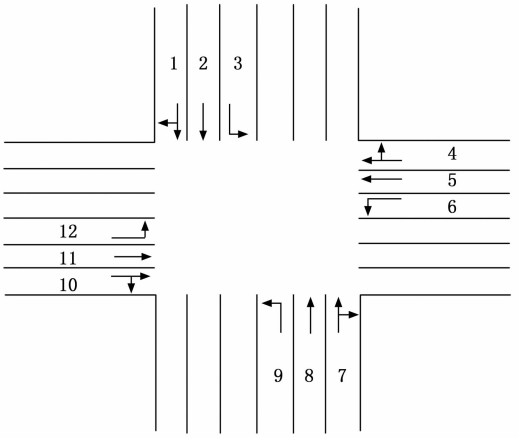

图 1 交叉口示意图

智能体的表现也有影响，譬如文献［14］的研究结果表明简洁的状态空间有助于提升智能体的训练性能，应尽可能构造简洁的状态空间。因此，本文利用交叉口各条车道上的排队车辆数和交叉口当前正在运行的相位构建状态空间，以便简洁高效的描述交叉口的交通状态信息。t 时刻交叉口的状态可定义为：

$$\boldsymbol{s}=[n_1,n_2,n_3,\cdots n_l,a] \tag{1}$$

式中，n_i表示决策时刻t 车道i 上的排队车辆数，l 表示交叉口的车道数量，a 表示当前时刻交叉口正在运行的相位。

1.2 动作空间

动作的定义和选取直接影响交叉口的通行能力[15]，本研究将动作空间定义为交通信号控制配时方案中预设的相位选择库。更具体的说，动作空间 $\boldsymbol{A}$ 定义为：

$$\boldsymbol{A}=[a_1,a_2,\cdots,a_n] \tag{2}$$

式中，a_i表示相位库中第 i 个相位，n 表示相位库包含 n 个相位。

为了研究不同动作空间定义下，所提交通信号控制模型的有效性，本文设计了两种预定义的相位动作选择方案。另外，通过仿真实验对两种不同动作选择方案进行了对比分析，具体实验结果将在 3.2 节讨论阐述。

1.2.1 预设动作方案 1

如图 2 所示，将动作空间定义为预定义的四相位动作空间，动作空间 $\boldsymbol{A}$ 可以表示为：

$$\boldsymbol{A}=[a_1,a_2,a_3,a_4] \tag{3}$$

式中，$a_1\sim a_4$ 分别为南北直行和右转（NS）、南北左转（NSL）、东西直行和右转（EW）、东西左转（EWL）4 个相位。在每个决策时刻，智能体选择一个相位分配交叉口的通道权（绿灯）。如果所选相位与当前正在执行的相位不相同，则切换相位之前，在预定义的时间内强制执行 3 s 黄色相位。值得注意的是，分配黄灯相位过渡阶段不是动作空间的一部分，而是环境对控制序列施加的约束条件。如果所选相位与当前正在执行的相位相同，则绿灯持续时间为 5 s。

1.2.2 预设动作方案 2

如图 3 所示，将动作空间定义为预定义的八相位动作空间，动作空间 $\boldsymbol{A}$ 可以表示为：

$$\boldsymbol{A}=[a_1,a_2,\cdots,a_8] \tag{4}$$

式中，$a_1\sim a_4$ 与方案 1 的相位库设置保持一致。$a_5\sim a_8$ 分别为南、北、东、西进车口（S、N、E、W）放行的相位。

1.3 奖励函数

奖励指的是智能体在执行动作后，环境传递给智能体的反馈信号。奖励用来评判智能体根据交通状态信息挑选动作的优劣程度，引导智能体优化动作的方向，一个准确的反馈信号对算法的收敛表现和交通信号配时方案的优化效果有着至关重要作用。在交通信号控制中，传统的奖励函数由排队车辆数、车辆等待时间等多种交通通行效率指标之和定义。虽然这些定义方法能够一定程度上体现出交叉口拥堵程度的变化趋势和优化效果，但是无法有效对比智能体在前后不同决策时刻选择的动作的优劣程度。为了有效提高智能体的学习能力，将奖励定义为相邻采样时刻

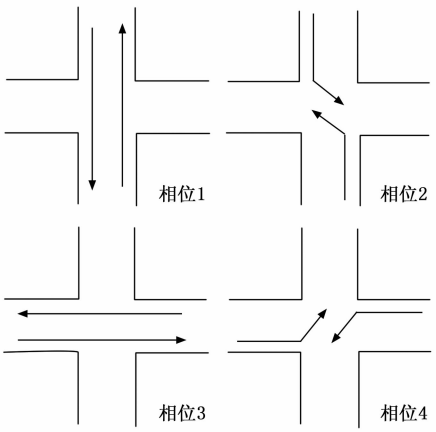

图 2 动作空间一

的车辆排队数之差，决策时刻 t 的奖励 R（t）为：

$$R(t)=q(t-1)-q(t) \tag{5}$$

式中，R（t）和 q（t）分别为 t 时刻智能体所获得的奖励和交叉口的车辆排队数量。当 R（t）>0，表示 t 时刻相对与 $t-1$ 时刻交叉口的交通状况有所改善。智能体根据反馈信号正向优化神经网络的决策参数，对 t 时间步内执行的动作进行正向“奖励”。

2 强化学习算法设计

2.1 强化学习基本框架

强化学习是机器学习的一个领域，强调观察环境而行动，将交通信号控制和强化学习结合能够有效提升交叉口的通行效率[16]。智能体感应环境状态并决策最优动作并作用于环境，环境执行动作后将会发生状态改变，以及产生一个奖惩信号向智能体反馈。当智能体决策的某个动作导致环境产生奖励信号，那么之后智能体选择这个动作的概率将会变大。智能体利用环境反馈信号不断优化最优策略 π[17]，直至智能体收到环境发出的终止信号。环境的状态、反馈的奖惩信号和智能体输出的动作构成强化学习算法基本框架，智能体和环境不断交互，动态地调整参数，形成

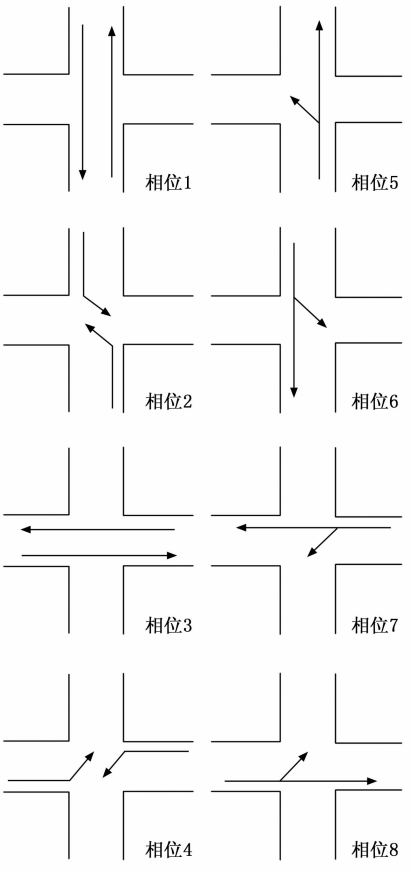

图 3　动作空间二

马尔科夫决策的动态系统。

强化学习的目标是通过不断与环境交互试错，不断优化最优策略 Π，使得该策略下的累积回报期望价值能够达到最大化。动作价值函数为 Q_π：

$$Q_\pi(s,a) = IE\left[\sum_{k=0}^{\infty}\gamma^k r_{t+k} \mid s_t = s, a_t = a, \pi\right] \quad (6)$$

式中，s_t，a_t表示在 t 时刻的状态和动作，γ 为折扣因子，代表每隔时间 t 奖惩 r 的衰减程度。智能体根据环境状态选择动作与环境交互并获得环境奖励反馈信号，将这些记录存储在经验池中不断训练智能体，逼近最优动作价值函数 $Q^*(s, a)$，得到最优策略 π，可以表示为：

$$Q^*(s,a) = \max Q_\pi(s,a) \quad (7)$$

2.2　基于 D3QN 的强化学习算法

DQN 算法使用经验池回放技术训练神经网络，逼近动作价值函数[18]。算法的动作－价值函数 Q 表示为：

$$Q(s_t,a_t,\theta) = R(s_t,a_t) + \gamma \cdot {}^{\max}_{a} Q(s_{t+1},a,\theta') \quad (8)$$

式中，s_t，a_t表示在 t 时刻的状态和动作，γ 为折扣因子。R 为智能体获得环境反馈的奖惩信号。DQN 算法结构中，将 eval _ net 参数化为 $Q(s, a, \theta)$，实时更新迭代 Q_{eval}，将 target _ net 参数化为 $Q(s, a, \theta')$，每隔周期 T 更新迭代 Q_{target}，加快系统的收敛速度。DQN 算法以最大化偏差"(maximization bias) 的方式更新 Q 表，容易导致 Q 值过估计的问题[19]。而 DDQN 缓解了这个问题，DDQN 的动作－价值函数 Q 表示为：

$$Q(s_t,a_t,\theta) = R(s_t,a_t) + \gamma \cdot Q\left[s_{t+1}, {}^{a}\text{rgmax}_{a} Q(s_{t+1},a,\theta), \theta'\right] \quad (9)$$

利用 eval _ net 选择动作、target _ net 更新 Q 值，缓解了系统的过估计问题。

DDQN 算法利用状态 s 和动作 a 的综合价值更新神经网络的参数，并没有分别考虑状态 s 和动作 a 深层价值。当动作对状态的影响很小时，神经网络不能很好的更新状态动作－价值 Q。因此本研究引入 Dueling Network，先将状态和动作进行一定程度的解耦，分别预测当前状态的价值和当前状态下每个动作的价值，再利用状态价值和状态动作价值构成新的状态动作价值 Q。

最优优势价值函数 A^*（s，a）表示为：

$$A^*(s,a) = Q^*(s,a) - V^*(s) \quad (10)$$

式中，Q^* 表示最优状态－动作价值函数，V^* 表示最优状态价值函数。

为了避免 A^* 和 V^* 的输出幅度相同、方向相反，干扰 Q^* 的输出，导致神经网路训练效果不理想的问题。在等式右边添加一个基准向量[20]。并同时对动作取最大值，表示为：

$${}^{max}_{a} A^*(s,a) = {}^{max}_{a} Q^*(s,a) - V^*(s) \quad (11)$$

$$V^*(s) = {}^{max}_{a} Q^*(s,a) \quad (12)$$

联立式子（10）、（11）和定理（12），D3QN 的动作－价值函数 Q^* 表示为：

$$Q^*(s,a) = A^*(s,a) + V^*(s) - {}^{m}ax_{a} A^*(s,a) \quad (13)$$

D3QN 算法利用神经网络 A（s，a，θ^A）不断逼近最优优势函数 A^*（s，a）、神经网络 V（s，θ^V）不断逼近最优价值函数 V^*（S）、进而神经网络 Q（s，a，θ）不断逼近最优动作－价值函数 Q^*（s，a），令 θ=（θ^A，θ^V）。算法的动作价值函数 Q 表示为：

$$Q(s,a,\theta) = A(s,a,\theta^A) + V(s,\theta^V) - {}^{max}_{a} A(s,a,\theta^A) \quad (14)$$

基于 D3QN 的强化学习算法进行第 i 次迭代时，从经验池 U（D）中随机均匀采样经验，更新以下损失函数：

$$L_i(\theta_i) = E_{(s_t,a_t,r,s_{t+1})\sim U(D)}\left[(Q_{\text{target}} - Q_{\text{eval}})^2\right] \quad (15)$$

利用梯度下降法最小化损失函数：

$$\frac{\partial L_i(\theta_i)}{\partial \theta_i} = IE_{(s_t,a_t,r,s_{t+1})\sim U(D)}\{R(s_t,a_t) + \gamma \cdot Q\left[s_{t+1}, {}^{\text{argmax}}_{a} Q(s_{t+1},a,\theta), \theta'\right] - Q(s_t,a_t,\theta)\nabla_{\theta_i}(Q(s_t,a_t,\theta)\nabla_{\theta_i})\} \quad (16)$$

Dueling Network 根据优势函数的定义将状态和动作进行一定程度的解耦，动作－价值不再简单的理解为状态和动作的共同价值，神经网络既预测状态和动作的共同价值又预测状态和动作各自的价值，对环境中的状态和动作进行相对独立而又紧密结合的观察和学习。基于 D3QN 的强化学习算法控制框架如图 4 所示，D3QN 算法的具体执行过程总结为如下算法 1。

算法 1：D3QN 算法

1)定义可配置参数：总训练回合 N，每回合最大训练步数 n 等；

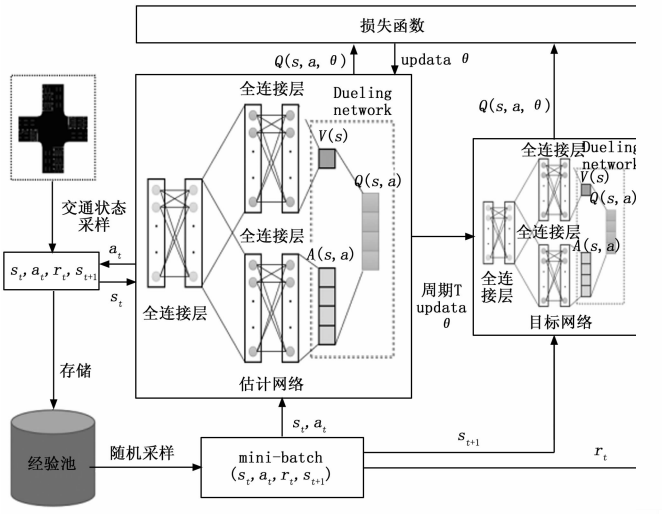

图 4 基于 D3QN 的强化学习算法原理示意图

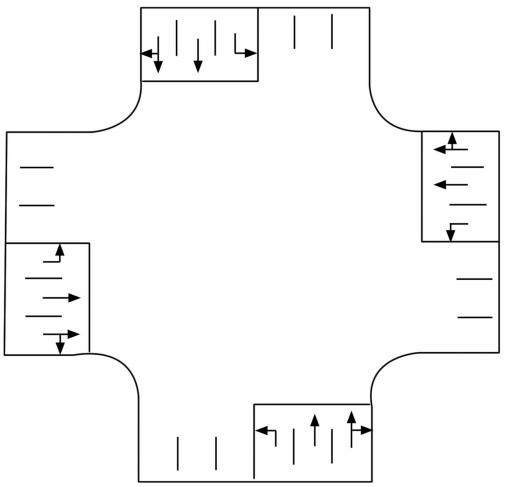

图 5 单十字交叉口

2)初始化神经网络参数:网络迭代周期 T、贪心系数 $\in$、学习率 α 等;

3)for episodes=0 to N do

4) 初始化路网环境,加载车流数据;

5) for t=1 to n do

6) 将从微观交通仿真平台 SUMO 采样的状态 s_t 作为 eval_net 输入,利用 eval_net 映射出所有动作的 Q 值估计;

7) 利用贪心策略选择动作 a_t

8) 执行动作 a_t 并计算奖励 r_t 和获取状态 s_{t+1}

9) 将$\{s_t, a_t, r_t, s_{t+1}\}$存储到经验池中

10) 从经验池中随机采样 B 个$\{s_t, a_t, r_t, s_{t+1}\}$

11) 以 mini batch 方式训练神经网络,迭代式(14)

12) 利用式(15)计算损失并利用(16)梯度下降更新 eval_net 参数 θ

13) 若经验池存储数据批数%T=0,则更新 target_net 参数 θ'

14) end for

15) end for

3 仿真与结果分析

3.1 仿真环境与参数设置

为了验证本文所提方法的有效性,基于微观交通仿真平台 SUMO 搭建交叉口仿真环境,利用基于 python 语言的控制平台与 SUMO 提供的 Traci 接口实现交互。在搭建的交通仿真环境中,基于所提强化学习方法,智能体能够利用监测的交叉口实时交通状态信息确定交叉口的最佳交通信号控制方案。

3.1.1 交叉口参数设置

如图 5 所示,本研究以典型的十字交叉口为研究对象。

每条道路设置为双向车道,每个方向的进车口包含三车道,分别为左转专用车道、直行专用车道、直行和右转车道,每条车道道路长度为 400 m,最大限速 50 km/h。此外,将智能体的决策时间间隔设置为 15 s,不同相位之间的黄灯切换时间设置为 3 s。若所选相位与当前正在执行的相位相同,则当前相位绿灯时间延长 5 s。

3.1.2 车流量设置

为了更好模拟现实世界中交叉口的车流,如图 6 所示,设计了一个基于韦伯分布的车辆随机生成模型,用于模拟不同时段的车流分布。

$$f(x,\lambda,k)=\begin{cases}0, & x<0\\ \dfrac{k}{\lambda}\dfrac{x}{\lambda}^{k-1}, & x\geqslant 0\end{cases} \tag{17}$$

式中,x 为随机变量,λ 为比例系数,k 为形状系数。控制不同交叉口的车辆生成概率生成不同车道上的车辆数量,车流生成信息如表 1 所示,基于车流数据集统计,不同方向进车口的直行车流、左转车流、右转车流比例为 0.7∶0.15∶0.15。

表 1 车流信息

方向	S	N	E	W
车流/辆	502	191	497	210

3.1.3 算法参数设置

基于 D3QN 的强化学习算法中,估计网络与目标网络包含 3 层全连接神经网络,一层 dueling 神经网络,通过前三层全连接层提取特征矩阵的特征后,利用 dueling 层的解耦功能,解耦动作价值和状态价值。算法参数配置如表 2 所示。

3.2 实验结果与分析

本文分别在动作空间为四相位和八相位两种交通信号控制场景下进行仿真实验。

3.2.1 四相位动作空间

图 7 是在四相位动作空间条件下,基于 DQN、DDQN、D3QN 三种控制方法在训练过程中的累计奖励对比,可以看出 3 种控制方法都能够收敛,说明 3 种方法都适用于当前交通信号控制场景。为了评估控制方法在交通信号控制任务的表现,分别对比不同控制方法的交通通行效率指标。

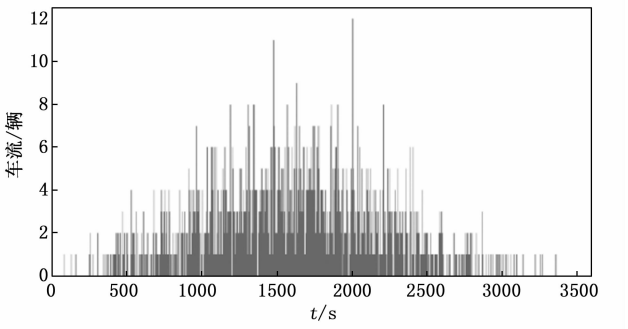

图 6　不同时段车流分布图

图 8～10 分别为 5 种控制方法的车辆平均排队长度、车辆平均等。

表 2　算法参数配置

参数名	参数值
学习率 α	0.000 3
折扣因子 γ	0.95
经验池大小 M	50 000
批处理大小 B	256
状态矩阵大小 $\boldsymbol{S}$	1 * 13 * 1
车辆标准长度 L /m	7.5
贪心策略衰减系数 ε	0.01
网络迭代周期 T	500
仿真回合 N	300
最大仿真步数 n	3 600

表 3　算法性能比较

算法	平均排队长度/m	平均等待时间/m	平均停车次数/次
定时控制	216.42	66.88	27.14
自适应控制	198.34	49.70	23.25
DQN	89.41	47.14	11.45
DDQN	88.49	41.59	11.28
D3QN	79.75	35..93	10.17

待时间和车辆平均停车次数变化曲线，在训练初期，由于智能体还处于探索阶段以及经验池样本太少，智能体决策的交通信号最优配时方案不够准确，因此交通通行效率指标大幅上升。随着训练回合增加，交通通行效率指标下降，逐渐趋于平稳收敛。

如表 3 所示，基于 D3QN 的控制方法与定时控制和自适应控制相比，车辆平均排队长度分别减少 63.15％和 59.79％，平均等待时间减少 46.28％和 27.71％，平均停车次数减少 62.53％和 56.25％。基于 DDQN 的控制方法与定时控制和自适应控制相比，车辆平均排队长度分别减少 59.11％和 55.38％，平均等待时间减少 37.81％和 16.32％，平均停车次数减少 58.44％和 56.25％。基于 DQN 的控制方法与定时控制和自适应控制相比，车辆平均排队长度减少 58.69％和 54.92％，平均等待时间减少 29.52％和 5.15％，平均停车次数减少 57.81％和 50.75％。综上所述，与定时控制和自适应控制对比，基于强化学习的交通信号控制方法都能够提高交叉口通行效率。与基于深度强化学习模型 DQN 和 DDQN 的交通信号控制方法相比，基于 D3QN 的交通信号控制方法在车辆平均排队长度、平均等待时间、平均停车次数这 3 个交通效益评价指标上均取得了更好的表现，说明在深度强化学习模型 DDQN 中引入的 Dueling network 能够提升神经网络提取深层特征的能力，解耦交通状态和相位动作的价值，提升交叉口的通行效率。

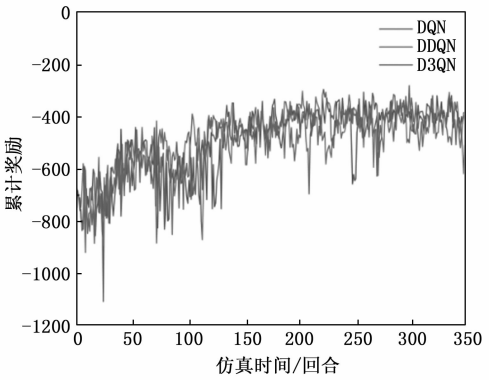

图 7　不同交通信号控制方法累计奖励

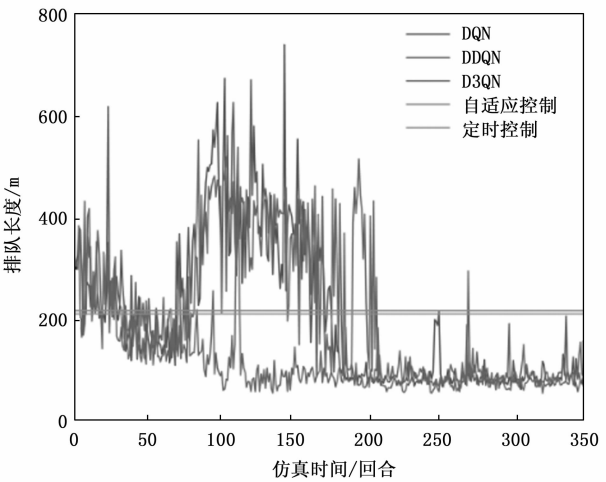

图 8　不同交通信号控制方法车辆平均排队长度

3.2.2　八相位动作空间

为了验证交通信号控制的相位设置对交叉口通行效率的影响，将交通信号控制设为八相位动作空间，并在基于 D3QN 的交通信号控制模型开展测试。如图 11 所示，两种相位设置的累计奖励曲线随着训练回合增加逐渐收敛。分别对比两种相位设置的交通通行效率指标，如图 12～14 可以看出随着训练回合增加，两种相位设置的交通通行效率指标减少，逐渐趋于平稳收敛，八相位动作空间的交通通

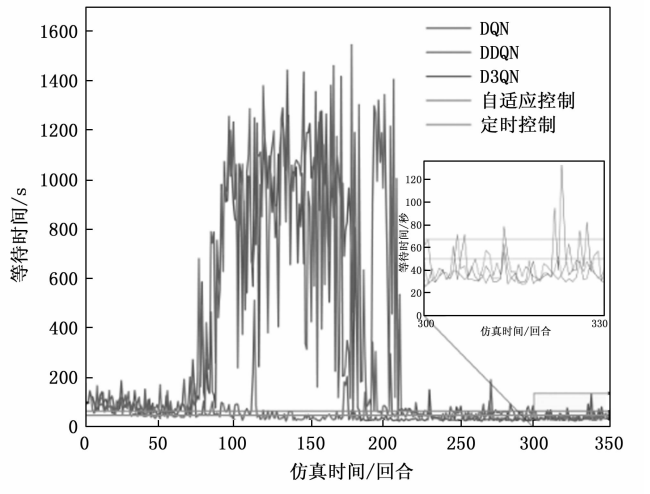

图 9 不同交通信号控制方法车辆平均等待时间

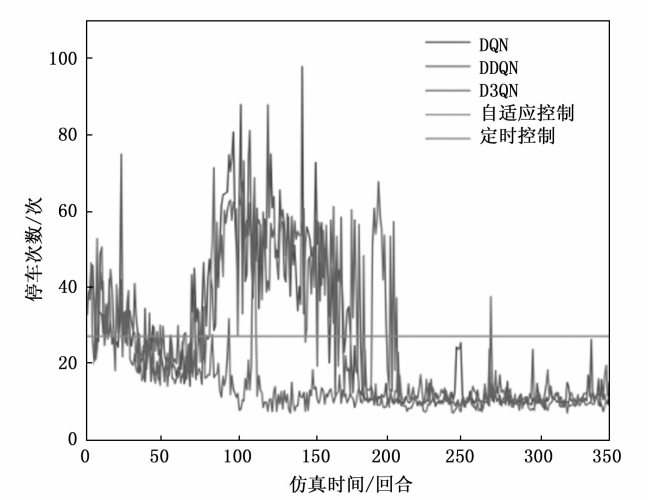

图 10 不同交通信号控制方法车辆平均停车次数

行效率指标变化曲线比四相位动作空间收敛更快，收敛效果更好。如表 4 所示，和四相位动作空间相比，八相位动作空间的车辆平均排队长度减少 4.68%、车辆平均等待时间减少 15.42%、车辆平均停车次数减少 4.62%。综上所述，基于 D3QN 的单交叉口交通信号控制方法在相位设置为八相位动作空间场景下的表现更加出色，能够更有效缓解交通拥堵。

表 4 D3QN 在不同相位配置下的性能比较

相位配置	平均排队长度/m	平均等待时间/s	平均停车次数/次
四相位	79.75	35.93	10.17
八相位	76.02	30.39	9.70

4 结束语

本研究提出一种基于 D3QN 的交通信号控制方法，利用微观交通仿真平台 SUMO，实时监测交叉口的交通状态

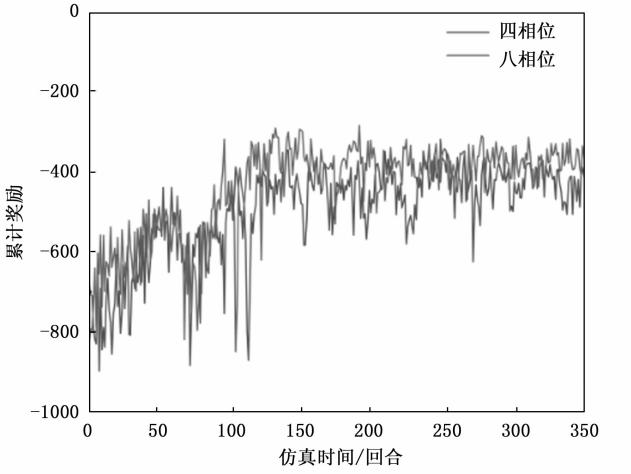

图 11 D3QN 算法在不同相位配置下的累计奖励

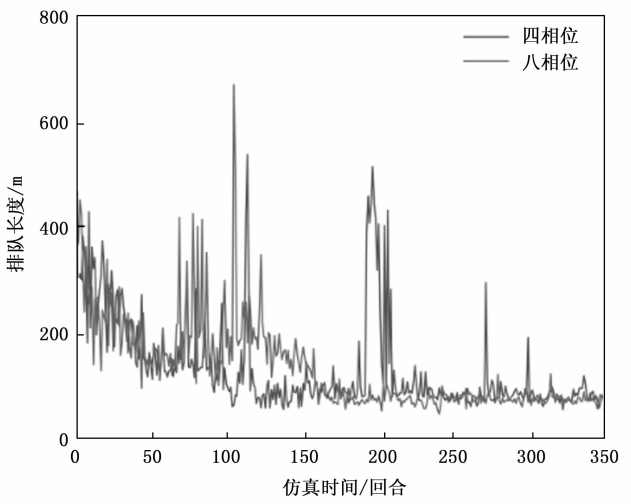

图 12 D3QN 在不同相位配置下车辆平均排队长度

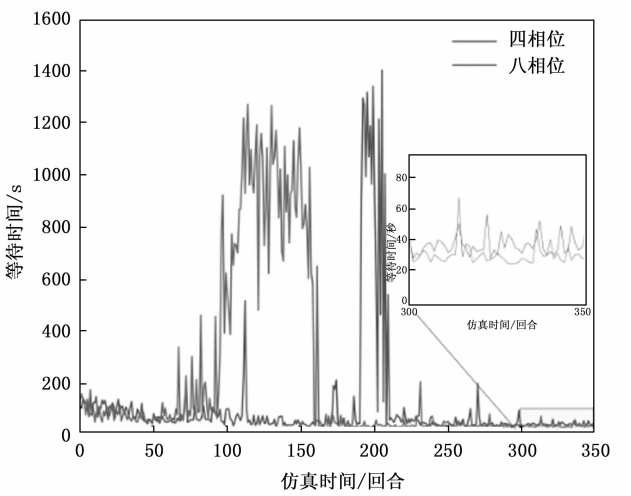

图 13 D3QN 在不同相位配置下车辆平均等待时间

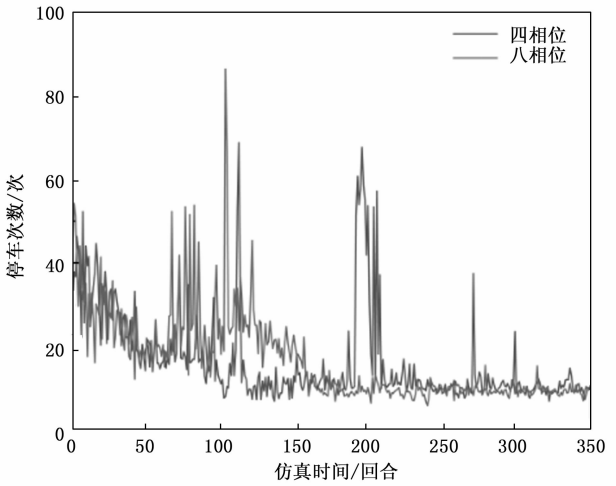

图 14　D3QN 在不同相位配置下车辆平均停车次数

信息，不断优化交通信号配时方案，缓解交通拥堵。实验结果表明，在车流分布不均匀条件下，通过引入 DDQN 模型和 Dueling network，缓解动作价值过估计和提升神经网络提取深层特征的能力，有效提升交通信号控制效果。与定时控制、自适应控制、基于 DQN 和 DDQN 的交通信号控制方法相比，基于所提深度强化学习 D3QN 的交通信号控制方法能够有效减少交叉口车辆平均排队长度、车辆平均等待时间、车辆平均停车次数，提高交叉口的通行效率。

所提方法还存在一些不足之处，譬如主要以单交叉口为对象开展交通信号控制研究，未能全面考虑单交叉口在多交叉口协同控制环境下的表现。考虑到多交叉口环境中相邻交叉口间通常存在较强的关联性，对由多交叉口构成的区域进行协同控制研究也许能够获得更好的交通通行效益。因此，下一步的研究将以多交叉口构成的区域为研究对象，重点研究基于 Dueling Network 的多交叉口交通信号协调控制问题。

参考文献：

[1] 任安虎，任洋洋，王　瑶. 多指标优化的深度强化学习单交叉口信号控制 [J]. 国外电子测量技术，2022，41 (10)：104 - 111.

[2] 林晓辉. 车路协同下基于交通密度的交叉口交通信号控制方法与仿真 [J]. 工业工程，2014，17 (4)：123 - 128.

[3] 傅立骏，郭海锋，董红召. 基于动态交通流量的可变车道自适应控制方法 [J]. 科技通报，2011，27 (6)：899 - 903.

[4] DEVAILLY F X，LAROCQUE D，CHARLIN L. Ig - rl：Inductive graph reinforcement learning for massive - scale traffic signal control [J]. IEEE Transactions on Intelligent Transportation Systems，2021，23 (7)：7496 - 7507.

[5] YE B L，WU W，RUAN K，et al. A survey of model predictive control methods for traffic signal control [J]. IEEE/CAA Journal of Automatica Sinica，2019，6 (3)：623 - 640..

[6] LIANG X，DU X，WANG G，et al. A Deep reinforcement learning network for traffic light cycle control [J]. IEEE Transactions on Vehicular Technology，2019，68 (2)：1243 - 1253.

[7] 胡　宇，刘美玲，周子昂，等. 基于 Q 学习的单路口交通信号协调控制 [J]. 计算机与现代化，2020，(5)：96 - 100.

[8] ZHANG R，LSHIKAWA A，WANG W，et al. Using Reinforcement learning with partial vehicle detection for intelligent traffic signal control [J]. IEEE Transactions on Intelligent Transportation Systems，2021，22 (1)：404 - 415.

[9] GU J，FANG Y，SHENG Z，et al. Double deep Q-network with a dual-agent for traffic signal control [J]. Applied Sciences，2020，10 (5)：1622.

[10] 陆丽萍，程　垦，褚端峰，等. 基于竞争循环双 Q 网络的自适应交通信号控制 [J]. 中国公路学报，2022，35 (8)：267 - 277.

[11] 任安妮，周大可，冯锦浩，等. 基于注意力机制的深度强化学习交通信号控制 [J]. 计算机应用研究，2023，40 (2)：430 - 434.

[12] 马东方，陈　曦，吴晓东，等. 基于强化学习的干线信号混合协同优化方法 [J]. 交通运输系统工程与信息，2022，22 (2)：145 - 153.

[13] YUN PENG W. GUO GE. Signal priority control for trams using deep reinforcement learning [J]. Acta Automatica Sinica，2019，45 (12)：2366 - 2377.

[14] 张尊栋，王岩楠，刘雨珂，等. 基于 Nash-Stackelberg 分层博弈模型的路网交通控制强化学习算法 [J]. 东南大学学报 (自然科学版)，2023，53 (2)：334 - 341.

[15] 陈喜群，朱奕璋，吕朝锋. 基于混合近端策略优化的交叉口信号相位与配时优化方法 [J]. 交通运输系统工程与信息，2023，23 (1)：106 - 113.

[16] 刘智敏，叶宝林，朱耀东，等. 基于深度强化学习的交通信号控制方法 [J]. 浙江大学学报 (工学版)，2022，56 (6)：1249 - 1256.

[17] WEI H，ZHENG G，GAYAH V，et al. Recent advances in reinforcement learning for traffic signal control：A sur? vey of models and evaluation [J]. ACM SIGKDD Ex? plorations Newsletter，2021，22 (2)：12 - 18.

[18] ZHAI P，ZHANG Y，SHAOBO W. Intelligent ship collision avoidance algorithm based on DDQN with prioritized experience replay under COLREGs [J]. Journal of Marine Science and Engineering，2022，10 (5)：585.

[19] HONG W，TAO G，WANG H，et al. Traffic signal control with adaptive online-learning scheme using multiple-model neural networks [J]. IEEE transactions on neural networks and learning systems，2022，34 (10)：7838 - 7850.

[20] PADAKANDLA S. A survey of reinforcement learning algorithms for dynamically varying environments [J]. ACM Computing Surveys，2021，54 (6)：1 - 25.