

DOI: 10.16544/j.cnki.cn43-1494/u.20220314001

文章编号: 1674-599X(2023)03-0110-08

引用格式: 邓兰, 吴义虎. 基于异步优势强化学习的交通信号控制策略[J]. 交通科学与工程, 2023, 39(3): 110-117.

Citation: DENG Lan, WU Yihu. Traffic signal control strategy based on asynchronous advantage actor-critic algorithm[J]. J. Transp. Sci. Eng., 2023, 39(3): 110-117.

## 基于异步优势强化学习的交通信号控制策略

邓兰, 吴义虎

(长沙理工大学 交通运输工程学院, 湖南 长沙 410114)

**摘要:**为解决交通信号控制中的信号灯配时调度不合理、路口拥堵等问题,提出一种基于行动者-评论家算法的城市智能交通控制算法。该算法是一种基于异步优势的算法,可对交通状态特征进行抽象表征,并以多线程并行实现对交通状态的精确感知。该算法还参考了强化学习算法,能在最短时间内不断迭代优化其内部参数,得到交通信号控制的最优方案。为验证该算法的有效性,采用交通仿真软件SUMO,对该算法和其他3种典型的交通信号控制算法进行模拟仿真,并对仿真结果进行比较和分析。研究表明:与这3类典型算法中效果最好的Q-learning算法相比,该算法的交叉口车辆平均延误时间减少了14.1%,平均队列长度缩短了13.1%,平均等待时间减少了13.5%。该交通信号控制算法能有效地改善城市道路拥堵,提高道路交叉口的通行效率。

**关键词:**智能交通信号控制;城市交通控制;深度强化学习;异步强化学习

中图分类号:U491.5

文献标志码:A

## Traffic signal control strategy based on asynchronous advantage actor-critic algorithm

DENG Lan, WU Yihu

(School of Traffic and Transportation Engineering, Changsha University of Science & Technology, Changsha 410114, China)

**Abstract:** In order to solve the problems in traffic signal control, such as unreasonable timing of signal lights and congestion at intersections, we propose an urban intelligent traffic control algorithm based on the asynchronous advantage actor-critic (A3C). This algorithm leverages asynchronous advantages to abstract and represent traffic state features, enabling accurate perception of traffic conditions through parallel multithreading. Drawing inspiration from reinforcement learning techniques, the algorithm iteratively optimizes its internal parameters to obtain the optimal solution for traffic signal control within the shortest possible timeframe. To assess the algorithm's effectiveness, we conducted simulated experiments using the traffic simulation software SUMO, comparing its performance with three other commonly used traffic signal control algorithms. The simulation results reveal that compared to the Q-learning algorithm, this algorithm reduces the average delay time of vehicles at intersections by 14.1%, decreases the average queue length by 13.1%, and lowers the average waiting time by 13.5%. This traffic signal control algorithm can effectively alleviate urban road congestion and improve the traffic efficiency of road intersections.

**Key words:** intelligent traffic signal control; urban traffic control; deep reinforcement learning; asynchronous reinforcement learning

随着中国经济的高速发展,中国汽车人均拥有量不断增加,交叉口拥堵也成为城市道路常态,城

市路网的通行效率受到了极大的影响。交通拥堵的主要原因是交通信号配时调度不合理,控制效率

收稿时间:2022-03-14

作者简介:邓兰(1996—),女,长沙理工大学硕士生。

低下。因此,亟须合理、高效的交通信号控制方案,提高城市道路通行效率,减少道路拥堵<sup>[1-2]</sup>。

许多学者在设计合理的交通信号控制方面做了大量研究,石冬花等<sup>[3]</sup>采用定时控制和感应控制的方法调节交叉口信号。Webster方法是较经典的定时控制方法,其以延误时间最小化为目标函数,确定参数,但该方法无法根据实时变化的交通状态来及时调整原有参数<sup>[4]</sup>。ARAGHI等<sup>[5-6]</sup>提出的SCOOT算法能根据实时交通状况,有规律地调整每个行车方向的可通行时间,大幅提高通行效率,但该算法仍无法解决复杂路面上的交通信号控制问题。这是因为该算法只能根据当前的交通状态,从已有的配时方案中选取最合适的,但备选方案可能均不能满足实际道路的需求。姜涛等<sup>[7]</sup>在对潮汐特性显著的交叉口设计交通信号控制方案时,将各控制时段的最优车道方案和信号控制组合方案组合起来,得到了通行效果更好的算法。BOWLING<sup>[8]</sup>率先将强化学习算法运用到交通信号控制中。MARSETIĆ<sup>[9]</sup>发现与固定配时算法相比,Q-learning算法能更好地缓解交通拥堵,但其仿真试验均在静态交通环境下进行,在某些实际道路中效果不佳。THORPE<sup>[10]</sup>通过仿真,将SARSA算法得到的配时方案与传统固定配时方案进行了对比,发现SARSA算法的配时方案能更好地改善交通拥堵,但效果不够稳定。LIU等<sup>[11]</sup>采用Q-learning算法,对交叉路口的信号灯进行控制,但由于该算法提供的信息有限,不能完全显示车辆状态,导致大量交通信息被忽略。

综上所述,为能提出更有效的城市路网的信号灯配时方案,本研究从两个方面入手:①在同一时间内,让交通信号灯根据交叉路口各个方向上不同交通流的实时状况,在较短时间内选择最适合的交通信号控制策略;②将该算法与其他算法在SUMO开源软件上进行仿真与比较,验证该算法的有效性及其可行性。

## 1 强化学习算法

### 1.1 行动者-评论家算法

DEGRIS<sup>[12]</sup>将价值迭代和策略迭代两种神经网络迭代算法结合在一起,提出了行动者-评论家(actor-critic, AC)算法。在该算法框架中,行动者(actor)运用策略函数,执行动作,并与环境进行交互、合作;评论家(critic)则运用评价函数,判断行动者动

作的优劣,并引导其下一阶段的动作。AC算法的原理如图1所示。该算法虽然可同时迭代两个神经网络,但稳定性不佳。

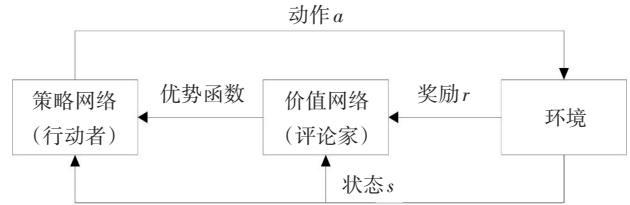


图1 行动者-评论家算法原理

Fig. 1 Actor-critic algorithm principle

### 1.2 优势行动者-评论家算法

优势行动者-评论家算法(advantage actor-critic, A2C)在AC算法的基础上增加了优势函数,并将其作为判定和选择动作的标准。如果执行动作的评价值高于期望值,则说明该动作优于平均动作,函数正在朝梯度上升方向运动;若执行动作的评价值低于期望值,则说明该函数正在朝梯度下降方向运动。A2C算法的原理如图2所示。由于该算法采取直接更新的策略,其迭代训练速度较为缓慢。

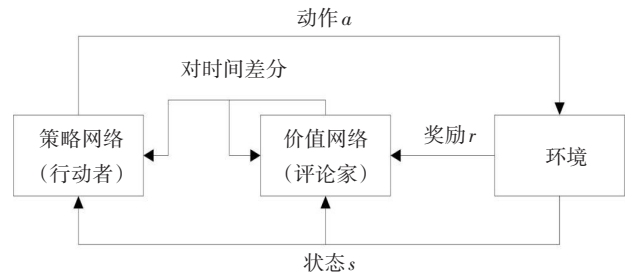


图2 优势行动者-评论家算法网络结构

Fig. 2 Advantage actor critical algorithm network structure

### 1.3 异步优势行动者-评论家算法

异步优势行动者-评论家算法(asynchronous advantage actor-critic, A3C)算法是MNIH<sup>[13]</sup>于2016年提出的。该算法在交通系统中得到了广泛运用,是一种能明显提高道路通行效率的异步强化学习算法<sup>[14]</sup>。

与AC算法和A2C算法相比,A3C算法的最大改变是可以采取异步操作。为能更迅速地收集数据,该算法采取多线程并行交互的方式,进行数据的采集、学习和训练。每个线程可自主采集样本信息,独立地完成数据训练,并获取下一个阶段的参数值,最后,将其异步更新到全局神经网络中。在每一次训练后,算法都会同步全局神经网络的参

数,再与环境进行交互,展开新一轮的学习与训练。且各线程采用的神经网络均与全局神经网络保持一致,但每个线程都对应不同的环境,并与之进行交互。因此,每个线程环境的数据均不同,每个线程得到的样本序列也会有差别。这对于策略学习

是有价值的,它能让模型在最短的时间内掌握更多的信息和知识。A3C算法原理模型结构如图3所示。在图3中,策略 $\Pi(s)$ 为全局神经网络输出的策略估计值,评论家 $v(s)$ 为全局神经网络输出的行动评价。

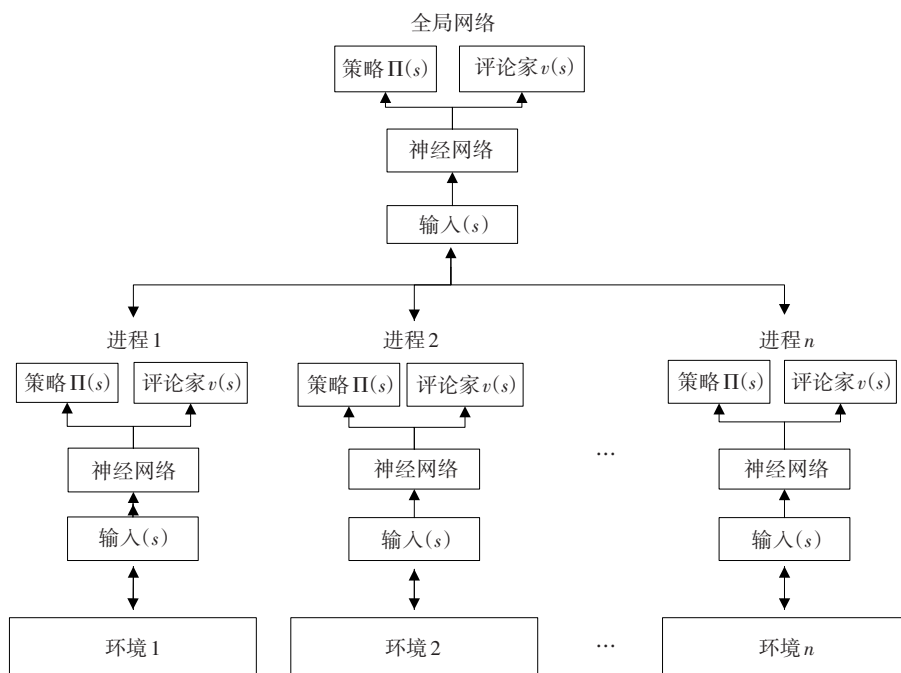


图3 异步优势行动者-评论家算法架构

Fig. 3 Asynchronous advantage actor-critic algorithm model architecture

## 2 异步深度强化学习交通控制算法

异步深度强化学习算法对在单交叉路口的智能体、状态输入、动作相位和奖励函数进行了定义。

### 1) 智能体。

智能体是具有自主活动的抽象类实体,如:无人驾驶的车辆、智能冰箱等。其能与周围环境不断地进行交流与合作,直至达到期望目标。在交通信号控制领域中,交通信号智能控制系统(交通信号灯)通过观测交叉路口的交通状况来选择合适的信号控制方案。因此,交通信号灯可作为在该智能系统中强化学习算法的智能体。

### 2) 状态输入。

交叉路口状态信息是交通信号灯选择信号相位的依据。为了更好地掌握交叉路口的实时交通情况,将交叉路口各个方向的每个车道中的车辆停车等待时间 $w$ 、车辆排队长度 $l$ 、平均车头时距 $h$ 、交叉口的车辆延误时间 $d$ 、交通信号灯相位变化 $c$ 作为

交通状态的输入变量。

### 3) 动作相位。

相位能影响交通信号灯能否选择出最合适的交通信号控制方案。目前,在大多数交通信号控制算法中,交叉路口的相位顺序是固定的,算法只对相位通行时间进行调整。这些算法不能应对复杂多变的实时交通流。因此,本研究既可以调整各个相位的绿灯时长,又可以根据交通流的实时变化任意切换相位。在交叉路口的车辆的4个动作相位设置如图5所示,本试验暂不考虑车辆右转这一情况。

为避免交通冲突,规定交通信号灯一次只能选择在一个相位 $a$ 上执行动作,此时,在其余相位方向上的车流均必须在停车线内等待。当 $a=1$ 时,保持该通行相位;当 $a=0$ 时,切换到另一相位。

当交通信号灯选择相位时,算法会优先考虑最大交通需求方向上的车流,但也要顾及其他方向车流的最长等待时间 $w_{\max}^{[15]}$ 。该最长等待时间 $w_{\max}$ 会因城市道路环境的不同而有所区别。

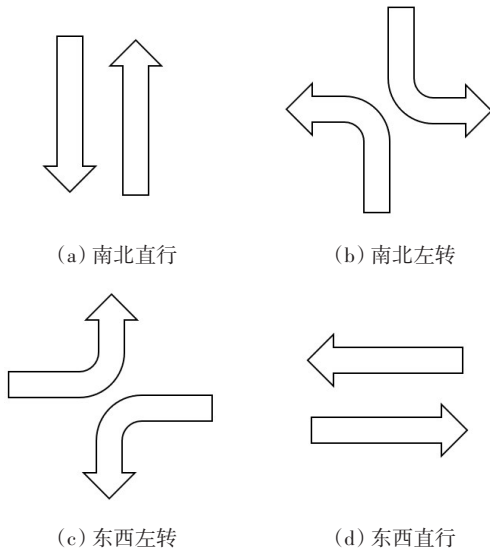


图5 交通路口车辆动作集合

Fig. 5 Intersection vehicle action collection

## 4) 奖励函数。

对于交通信号控制方案而言,奖励函数可对交通信号灯选择某一相位动作后的交通状况变化情况进行评价,判断该方案是否达到预期通行效果,并最终确定其是否为最优交通信号控制方案。某一交通信号控制方案的奖励值越大,其实际通行效果越好;反之,若某一交通信号控制方案的奖励值越小,其实际通行效果越差。随着迭代时间的延长,交通路口智能控制方案的奖励值影响会逐渐降低。为提高交叉路口的通行效率,奖励函数要考虑交通路口状况的各项评价指标,如:车辆延误时间、队列长度、等待时间等。该交叉路口奖励函数的表达式为:

$$r = k_1 d + k_2 q + k_3 w + k_4 p \quad (1)$$

式中: $k_1, k_2, k_3, k_4$ 为权重系数; $d$ 为交叉口在各方向上所有车辆的延迟时间之和; $q$ 为该交叉口在各方向上所有车辆队列长度之和; $w$ 为交叉口在各方向上所有车辆等待时间之和; $p$ 为不同相位的切换时间<sup>[16]</sup>。

在交通控制的异步深度学习强化算法中,智能体(交通信号灯)先根据观察的交叉路口信息,选择符合该地交通状态的动作相位;然后,执行该动作相位,并将道路交通流的变化结果反馈给智能体(交通信号灯);最后,智能体根据反馈结果自适应地调整原有参数。交通智能体(交通信号灯)的强化学习算法的流程如图6所示。

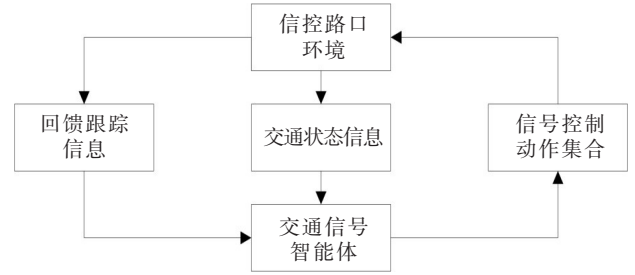


图6 交通智能体的强化学习算法流程

Fig. 6 Reinforcement learning basic model of traffic signal agent

## 3 算法设计与仿真

SUMO 仿真软件能灵活地定义交通场景,模拟实时交通场景,在交通规划和管理中得到了广泛的应用<sup>[17]</sup>。本研究采用SUMO 仿真软件中的TraCI接口模块来完成其与Python平台间的交流,根据交叉路口的实时交通信息,选择合适的信号配时方案<sup>[18]</sup>。该算法建立在深度学习框架Tensor Flow的基础上<sup>[19]</sup>。

## 3.1 算法设计

该算法先初始化交通信号灯参数;算法的各线程再从SUMO 仿真软件中获得当前其在各自的交通环境中交叉路口的状态信息;然后,将该状态信息传输至各自的神经网络中,根据神经网络输出的反馈结果,从动作相位集合中按照一定的策略,选择其中的一个相位;最后,把当前交叉路口的交通状态信息、动作相位、下一个时间步长的回报值、交通环境状态信息组合为一个4元数组,并以此来更新全局神经网络的参数。当全局神经网络进行更新时,适当放缓更新速率,保证该神经网络学习的稳定性,使全局神经网络的输出值能更贴近目标 $Q$ 值。选择具有最大的 $Q$ 值的动作相位,即可得到交通信号控制方案。重复以上操作,直到交通路口的信号灯最终获得最优交通控制策略为止。

设A3C算法公共部分的全局神经网络的策略参数为 $\theta$ ,价值参数为 $\omega$ ,共享迭代轮数为 $N$ ,全局最大迭代次数 $N_{\max}$ ,步长为 $\alpha$ ;A3C算法各线程的神经网络的策略参数为 $\theta'$ ,价值参数为 $\omega'$ ,各线程内单次迭代时间序列最大长度为 $N_{\text{local}}$ ,状态特征维度为 $n$ ,动作集为 $A$ ,步长为 $\beta$ ,熵系数为 $c$ ,折扣因子为 $\gamma$ ,探索率为 $\varepsilon$ 。



算法步骤为:

1) 输入公共部分 A3C 全局神经网络参数  $\theta, \omega$ ; 重新分别设定行动者、评论家两个网络梯度:  $d_\theta = 0, d_\omega = 0$ ;

2) 将全局神经网络中参数同步到各个线程中:  $\theta' = \theta, \omega' = \omega$ ;

3) 令  $t_{\text{start}} = t$ , 从初始化交通环境中得到交通状态  $s_i$ ;

4) 根据策略  $\prod(a_i | s_i; \theta)$ , 选择合适的动作相位  $a_i$ , 通过选择好的动作相位  $a$ , 获得奖励值  $r_i$  和下一个交通状态; 定义当前的交通状态:  $t \leftarrow t + 1, T \leftarrow T + 1$ ;

5) 若  $s_i$  为最后状态, 或  $t - t_{\text{start}} = t_{\text{local}}$ , 则计算最后一个时间序列位置  $s_i$  的  $Q(s, t)$ ; 否则回到步骤 4)。  $Q(s, t)$  的表达式为:

$$Q(s, t) = \begin{cases} 0, & \text{最终状态} \\ V(s, \omega'), & \text{不是最终状态, 进入下一状态} \end{cases};$$

6) 计算异步梯度,  $i \in (t - 1, t - 2, \dots, t_{\text{start}})$ :

① 估计目标函数  $Q$  值, 计算每个时刻的  $Q(s, i)$ :  $Q(s, i) = r_i + \gamma Q(s, i + 1)$ ;

② 对累计 Actor 的本地梯度进行更新(判定策略梯度方向):

$$d_\theta \leftarrow d_\theta + \nabla \theta' \log_\pi \theta'(s_i, a_i)(Q(s, i) - V(s_i, \omega') + c \nabla \theta' H(\pi(s_i, \theta')));$$

③ 对累计 Critic 的本地梯度进行更新(判定价值梯度方向):

$$d_\omega \leftarrow d_\omega + \frac{\partial(Q(s, i) - V(s_i, \omega'))}{\partial \omega'};$$

7) 更新全局神经网络参数:  $\theta \leftarrow \theta - \partial d_\theta, \omega \leftarrow \omega - \beta d_\omega$ ;

8) 若  $N > N_{\text{max}}$ , 则算法结束, 输出公共部分的 A3C 神经网络参数  $\theta, \omega$ ; 否则, 返回步骤 2)。

参数设置: 为了保证试验公正, 各对比算法都使用同样的参数设置。学习率  $\xi$  设置为 0.001, 折扣因子  $\gamma$  设置为 0.9, 当算法进行学习训练时, 采用  $\epsilon$ -greedy 算法来进行探寻和动作选择,  $\epsilon$  的初值设为 0.1, 随着训练次数的增加,  $\epsilon$  值会大幅下降。因此, 在训练开始之前, 需要初始化算法模型的参数, 提升算法的稳定性。

根据试验观察, 式(1)中各系数设定为:  $k_1 = -0.25, k_2 = -0.25, k_3 = -0.25, k_4 = -1.00$ 。

## 3.2 试验环境设置

### 3.2.1 路口属性设置

单一交叉路口是城市交通网络中最简单、最基本的交叉路口。对单一交叉路口的控制效果是评估交通控制系统的重要基础。

本次道路试验的单一交叉路口由一条南北向道路和一条东西向道路组成, 试验暂时仅考虑机动车, 忽略其他交通要素的影响。每条道路均设计为双向六车道, 车道最高设计时速为 65 km/h, 如图 7 所示。

设交叉路口 4 个方向均有车流。4 个交通仿真环境将同时生成数据, 并独立训练各自的神经网络。

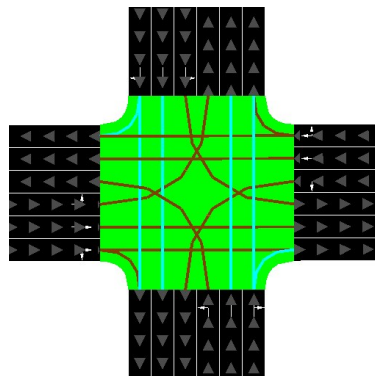


图7 单一交叉路口仿真区域

Fig. 7 Single intersection simulation area

### 3.2.2 配时方案设置

该交叉路口在饱和状态下的固定信号配时方案见表 1。其具备 4 个相位, 且其中包含左转相位。在每个相位的绿灯结束后, 设定 3 s 的黄灯时间, 保证在绿灯结束时刻冲出停车线的车辆能在红灯开启前驶离交叉路口。

表1 固定信号配时方案

Table 1 Fixed signal timing scheme

相位	方向	绿灯时间/s	黄灯时间/s	周期/s
第一相位	南北直行	29	3	130
第二相位	南北左转	21	3	130
第三相位	东西直行	43	3	130
第四相位	东西左转	25	3	130

### 3.2.3 车流量设置

假设车辆到达服从泊松分布, 交叉路口各个方向的车辆平均到达率、仿真时间等参数见表 2。

表2 交通流量设置

Table 2 Traffic flow settings

车流方向	车辆平均到达率/(辆·s <sup>-1</sup> )	仿真时间/s
南北直行	0.100	7 200
南北左转	0.030	7 200
东西直行	0.150	7 200
东西左转	0.026	7 200

4 试验结果与分析

Q-Learning算法与A3C算法在学习过程中平均累计奖励值与总训练次数间的关系曲线如图8所示。

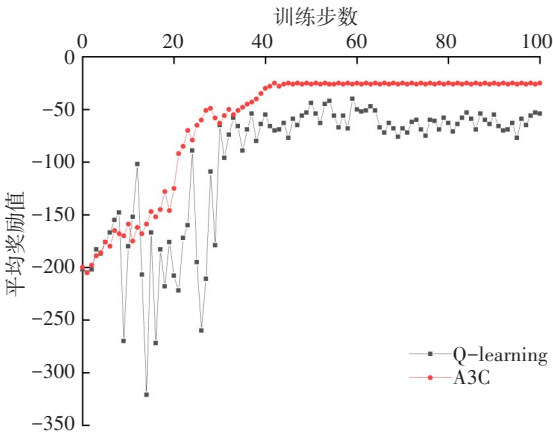


图8 不同算法的平均累计奖励对比

Fig. 8 Average cumulative reward comparison

从图8可以看出,在各算法的训练早期,它们均未收敛,但A3C算法和Q-Learning算法的训练值较接近。随着训练次数的增加,Q-learning算法仍处于振荡发散状态,未达到其最优的控制效果;但A3C算法则很快地收敛,进入了稳定状态。

根据在该交叉路口采集的车辆信息,经A3C算法反复迭代和优化后,得到的最佳信号灯配时方案见表3。

由表3可知,相比固定配时方案,A3C算法优化得到的最佳信号灯配时方案在交通需求量较大的相位上的绿灯时间和绿信比均有所增加。

参考《城市道路交通管理评价指标体系》(2008版)中对城市等级的划分和《城市道路设计规范》(CJJ 37—90)中对交叉口类型的划分,本研究选取B类城市的交叉口,对其进行数值仿真,并假设驾驶

人可接受的等待时间为120~200 s。由表3可知,各算法优化后的各方向车流的等待时间上限 $w_{\max}$ 均在设定的合理范围内。

表3 优化后的信号配时方案

Table 3 Optimized signal timing scheme

相位	方向	绿灯时长/s	黄灯时长/s	周期/s
第一相位	东西直行	53	3	145
第二相位	南北直行	34	3	145
第三相位	南北左转	26	3	145
第四相位	东西左转	20	3	145

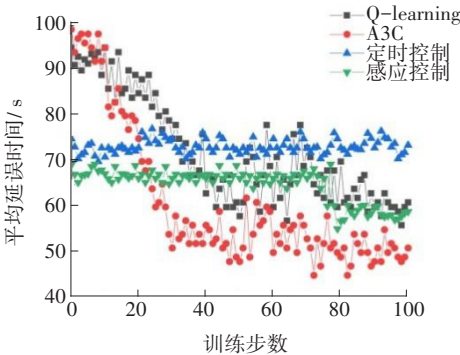
为更好地检验基于异步优势行动者-评论家算法对交通信号控制的有效性,以平均延误时间、平均队列长度、平均等待时间作为评价指标,将其与定时控制、感应控制、Q-Learning这三类典型算法进行对比。各算法训练的对比结果如图9所示,各算法各指标的平均值对比见表4。在相同交通状态条件下,这三个交通评价指标的平均值越小,交叉路口车辆的通行效率就越高,算法有效性也就越好。

表4 饱和状态下算法仿真结果对比

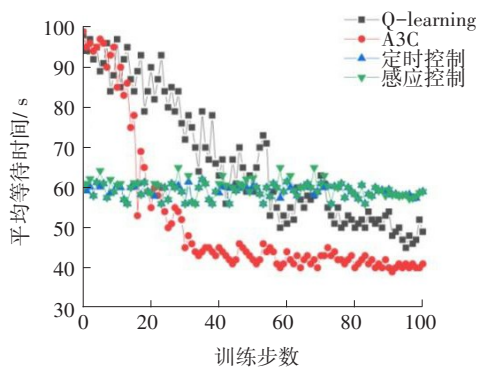
Table 4 Comparison of algorithm simulation results in

saturated state

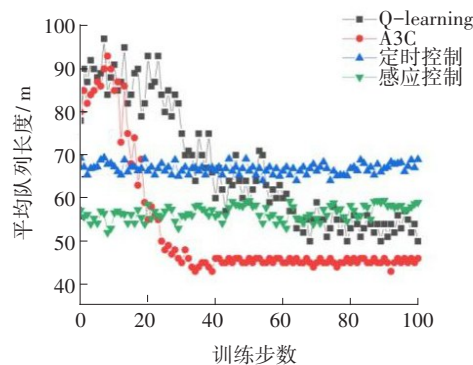
算法	平均延误/s	平均队列长度/m	平均等待时间/s
定时控制	75.42	66.81	61.02
感应控制	68.23	59.14	56.11
Q-learning	61.21	54.96	50.38
A3C	52.57	47.83	43.58



(a) 平均延迟



(b) 平均等待时间



(c) 平均队列长度

图9 交通控制效果对比

Fig. 9 Comparison of traffic control effects

由表4可知,与3类常用算法中优化效果最好的Q-learning算法对比,A3C算法的车辆平均延误时间减少了14.1%,平均队列长度缩短了13.1%,平均等待时间减少了13.5%。在本次仿真试验中,设置的道路流量较大,该高负载路面状态导致A3C算法的收敛速度有所下降,但A3C算法仍是这4种算法中收敛速度最快的。

## 5 结论

本研究针对城市道路的单个交叉路口提出一种基于异步优势行动者-评论家算法的交通信号控制算法。该算法能有效利用城市道路交叉路口的实时状态信息,解决以往大部分算法在学习训练过程中难以得到最优的交通控制方案的问题,能有效地降低交叉路口车辆排队长度和停车等待时间,改善城市道路拥挤,缓解交通压力。

本算法通过SUMO仿真软件进行了数值试验。该仿真试验仅考虑机动车对交通状况的影响,是在一种理想的道路交通环境中进行的。但实际的交

通环境还会受到非机动车、行人等多因素的影响。这也是本研究接下来要做的工作。

## 参考文献(References):

- [1] SUN H, CHEN C L, LIU Q, et al. Traffic signal control method based on deep reinforcement learning [J]. Computer Science, 2020, 47(2): 169-174.
- [2] 朱海峰,刘畅,刘彦斌,等. 基于通行能力系数优化的道路交叉口单点动态控制研究[J]. 公路与汽运, 2019(5): 29-34.  
ZHU Haifeng, LIU Chang, LIU Yanbin, et al. Research on single point dynamic control of road intersection based on capacity coefficient optimization [J]. Highways & Automotive Applications, 2019(5): 29-34.
- [3] 石冬花,田国亮,闫彭. 道路交叉口信号控制方法的发展现状与趋势[J]. 交通标准化, 2011, 39(7): 193-197.  
SHI Donghua, TIAN Guoliang, YAN Peng. Development status and trends of signal control algorithm at road intersection [J]. Transport Standardization, 2011, 39(7): 193-197.
- [4] LO H K. A reliability framework for traffic signal control [J]. IEEE Transactions on Intelligent Transportation Systems, 2006, 7(2): 250-260.
- [5] ARAGHI S, KHOSRAVI A, CREIGHTON D. A review on computational intelligence methods for controlling traffic signal timing [J]. Expert Systems with Applications, 2015, 42(3): 1538-1550.
- [6] 隋莉颖,李威,石建军,等. SCOOT和ACTRA信号控制系统分析[J]. 道路交通与安全, 2007, 7(2): 10-13.  
SUI Liying, LI Wei, SHI Jianjun, et al. The analysis and application actuality of traffic signal control system—SCOOT and ACTRA in Beijing [J]. Road Traffic & Safety, 2007, 7(2): 10-13.
- [7] 姜涛,马倩楠. 多时段可变导向车道设置与信号优化方法[J]. 交通科学与工程, 2019, 35(1): 86-93.  
JIANG Tao, MA Qiannan. Multi-time variable guiding lane setting and signal optimization method [J]. Journal of Transport Science and Engineering, 2019, 35(1): 86-93.
- [8] BOWLING M. Multi agent learning in the presence of agents with limitations [R]. Carnegie Mellon Univ Pittsburgh Pa School of Computer Science, 2003.
- [9] MARSETIĆ R, ŠEMROV D, ŽURA M. Road artery traffic light optimization with use of the reinforcement learning [J]. PROMET - Traffic & Transportation, 2014,

- 26(2):101-108.
- [10] THORPE T L, Anderson C W. Traffic light control using sarsa with three state representations [R]. Technical report, Cite seer, 1996.
- [11] LIU Y, LIU L, CHEN W P. Intelligent traffic light control using distributed multi-agent Q learning [C]// IEEE 20th International Conference on Intelligent Transportation Systems (ITSC). October 16-19, 2017, Yokohama, Japan. IEEE, 2018: 1-8.
- [12] DEGRIS T, WHITE M, SUTTON R S. Off-policy actor-critic [EB/OL]. (2013-06-20) [2021-12-16]. <https://arxiv.org/abs/1205.4839>
- [13] V.MNIH, A.P.BADIA, M.MIRZA, A.Graves, T.Lillicrap, T.Harley, K.Kavukcuoglu, et al., Asynchronou-s methods for deep reinforcement learning [EB/OL]. (2016-02-04) [2021-12-26]. <https://arxiv.org/abs/1602.01783>.
- [14] GENDERS W, RAZAVI S. Evaluating reinforcement learning state representations for adaptive traffic signal control [J]. Procedia Computer Science, 2018, 130: 26-33.
- [15] 郭海锋, 张贵军. 基于无线传感网络的单点自组织信号控制方法 [J]. 中国公路学报, 2010, 23(3): 102-107.  
GUO Haifeng, ZHANG Guijun. Method of traffic self-organizing signal control for isolated intersection based on wireless sensor network [J]. China Journal of Highway and Transport, 2010, 23(3): 102-107.
- [16] 黄浩, 胡智群, 王鲁哈, 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 Zhiquan, 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.
- [17] 侯德藻, 高兰达, 钱振伟, 等. 基于 SUMO 软件的异构交通流仿真平台车辆模型开发与集成 [J]. 同济大学学报(自然科学版), 2021, 49(7): 957-963, 1038.  
HOU Dezao, GAO Landa, QIAN Zhenwei, et al. Development and integration of vehicle model for heterogeneous traffic flow simulation platform based on SUMO [J]. Journal of Tongji University (Natural Science), 2021, 49(7): 957-963, 1038.
- [18] 王莹多. 基于深度强化学习的路口自适应控制 [D]. 大连: 大连理工大学, 2017.  
WANG Yingduo. Adaptive control of intersections based on deep reinforcement learning [D]. Dalian: Dalian University of Technology, 2017.
- [19] 朱昭俊, 李雪芹. 基于 Python 与 Tensorflow 的交通信号决策系统 [J]. 综合运输, 2022, 44(5): 74-79, 143.  
ZHU Zhaojun, LI Xueqin. Traffic signal decision system based on python and tensorflow [J]. China Transportation Review, 2022, 44(5): 74-79, 143.
- (责任编辑: 李脉; 校对: 罗容; 英文编辑: 刘至真)

(上接第89页)

- contact surface during rotation construction of long-span bridges [D]. Guangzhou: South China University of Technology, 2019.
- [13] 郭大智, 冯德成. 层状弹性体系力学 [M]. 哈尔滨: 哈尔滨工业大学出版社, 2001.  
GUO Dazhi, FENG Decheng. Mechanics of layered elastic system [M]. Harbin: Harbin Institute of Technology Press, 2001.
- [14] 格拉德韦尔, G.M.L. 经典弹性理论中的接触问题 [M]. 范天佑, 译. 北京: 北京理工大学出版社, 1991.  
GLADWELL, G. M. L. Contact problems in classical elasticity Theory [M]. Fan Tianyou. Trans. Beijing: Beijing Institute of Technology Press, 1991.
- (责任编辑: 欧兆虎; 校对: 罗容; 英文编辑: 刘至真)