软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2021,32(8):2425−2438 [doi: 10.13328/j.cnki.jos.006191]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

面向优先车辆感知的交通灯优化控制方法*

邵明莉, 曹 鹗, 胡 铭, 章 玥, 陈闻杰, 陈铭松

(上海市高可信计算重点实验室(华东师范大学),上海 200062)

通讯作者: 陈铭松, E-mail: mschen@sei.ecnu.edu.cn

摘 要: 智慧交通灯控制能够有效地改善道路交通的秩序和效率.在城市交通网络中,具有紧急任务的特殊车辆对于通行效率的要求更高.目前已有的智慧交通灯控制算法通常对路网中的所有车辆一视同仁,没有考虑到特殊车辆的优先性;而传统的控制特殊车辆优先通行的方法基本上都是采用信号抢占的方式,对普通车辆的通行干扰过大.为此,提出一种面向优先车辆感知的交通灯优化控制方法,通过与道路环境的不断交互来学习交通灯控制策略,在设置状态和奖励函数时增加特殊车辆的权重,并利用 Double DQN 和 Dueling DQN 来提升模型表现,最终在城市交通模拟器 SUMO 中进行仿真实验.在训练趋于稳定之后,与固定时长控制方法的对比实验结果显示,该方法能够将特殊车辆与普通车辆的平均等待时间分别缩短 68%与 22%左右;与不考虑优先级的方法相比,特殊车辆的平均等待时间也有 35%左右的优化.验证了该方法能够在提高车辆通行效率的同时,体现出对特殊车辆的优先处理.同时,实验也表明该方法能够扩展应用于多路口场景中.

关键词: 智慧交通;交通信号控制;强化学习;深度学习;车辆优先级

中图法分类号: TP311

中文引用格式: 邵明莉,曹鹗,胡铭,章玥,陈闻杰,陈铭松.面向优先车辆感知的交通灯优化控制方法.软件学报,2021,32(8):2425−2438. http://www.jos.org.cn/1000-9825/6191.htm

英文引用格式: Shao ML, Cao E, Hu M, Zhang Y, Chen WJ, Chen MS. Traffic light optimization control method for priority vehicle awareness. Ruan Jian Xue Bao/Journal of Software, 2021,32(8):2425−2438 (in Chinese). http://www.jos.org.cn/1000-9825/6191.htm

Traffic Light Optimization Control Method for Priority Vehicle Awareness

SHAO Ming-Li, CAO E, HU Ming, ZHANG Yue, CHEN Wen-Jie, CHEN Ming-Song

(Shanghai Key Laboratory of Trustworthy Computing (East China Normal University), Shanghai 200062, China)

Abstract: Intelligent traffic light control can effectively improve the order and efficiency of road traffic. In urban traffic networks, special vehicles with urgent tasks have higher requirements for traffic efficiency. However, current intelligent traffic light control algorithms generally treat all vehicles equally, without considering the priority of special vehicles, while the traditional methods basically adopt signal preemption to ensure the priority of special vehicles, which has a great influence on the passage of ordinary vehicles. Therefore, this study proposes a traffic light optimization control method orient priority vehicle awareness. It learns traffic light control strategies through continuous interaction with the road environment. the weight of special vehicles is increased in state definition and reward function, and Double DQN and Dueling DQN are used to improve the performance of the model. Finally, the experiments are carried out in the urban traffic simulator SUMO. After the training stabilizes, compared with the fixed time control method, the proposed

* 基金项目: 国家重点研发计划(2018YFB2101300); 国家自然科学基金(61872147); 华东师范大学优秀博士生学术创新能力提升计划(YBNLTS2020-041)

Foundation item: National Key Research and Development Program of China (2018YFB2101300); National Natural Science Foundation of China (61872147); Academic Innovation Promotion Program for Excellent Doctoral Students of East China Normal University (YBNLTS2020-041)

本文由“泛在嵌入式智能系统”专题特约编辑郭兵教授、王泉教授、邓庆绪教授、陈铭松教授、张凯龙副教授推荐.

收稿时间: 2020-07-24; 修改时间: 2020-09-07; 采用时间: 2020-11-02; jos 在线出版时间: 2021-02-07

method can reduce the average waiting time of special vehicles and ordinary vehicles by about 68% and 22%, respectively. Compared with the method without considering priority, the average waiting time of special vehicles is also optimized by about 35%, all these results prove that the proposed method can not only improve the efficiency of all vehicles, but also give special vehicles higher priority. At the same time, the experiment also shows that the proposed method can be extended to apply in multi-intersection scenes.

Key words: intelligent transportation; traffic signal control; reinforcement learning; deep learning; vehicle priority

随着城市化建设的进一步推进和经济的飞速发展,汽车数量在不断飙升.据上海交通出行网统计,截至 2018 年底,上海市实有小客车规模突破 500 万[1].与此同时,城市的交通需求与道路设施之间的矛盾日益突出,交通拥堵也成为了城市发展过程中一个不可忽视的问题.近年来,随着物联网与人工智能技术的发展,智能交通系统成为了现代交通发展的方向[2,3],越来越多的人开始尝试从智能算法中寻求解决城市交通问题的方案,利用物联网技术获取车辆状态以及道路设备状态信息,然后再使用各种智能算法对获取到的信息进行分析,给出缓解交通压力的操作建议[4,5].

在城市交通中,位于道路交叉路口处的交通灯是指引车辆通行的关键设备,对出行效率起着至关重要的作用.合理的交通灯控制方案能够有效地缓解路口交通压力,提高通行效率.传统的交通灯控制策略基本都是采用固定的时间间隔以及固定的相位序列来调整交通灯信号[6],这种方式虽然简单,但却无法适应不同的交通路况:比如,可能在某个十字路口只有一辆车或者只有一个方向上有车,但它却不得不等待一轮红绿灯之后才能通过,或者是等到了绿灯,但绿灯时间不足以通过路口.因此,如何设计一套智能交通灯控制算法,使其能够根据路况动态改变交通灯相位,就是一个非常有意义的研究课题.

近年来,车联网技术的发展使得交通灯智能控制成为了可能:通过 GPS、传感器等设备,车辆可以完成自身环境和状态信息的采集,这些信息将通过互联网技术汇聚到中央处理器,经过各种智能算法分析处理,进而控制交通灯相位切换.强化学习完成的目标就是让智能体在与环境交互的过程中学习策略,以达成回报最大化或实现某个特定目标[7].它根据实时反馈来调整动作的特征,使其尤为适合解决智能交通灯控制问题.自从 Thorpe 等人[8]于 1997 年首次将强化学习方法应用于交通信号最优化控制以来,各种基于强化学习的交通灯控制算法层出不穷[9−13].但对于大部分目前已有的工作而言,它们的目标定位都是如何缓解道路交通压力,即缩短车辆等待时间以及队列长度,或者是提高路口吞吐率等.但是在实际场景中,某些执行任务的特殊车辆,比如警车、消防车或者救护车等,它们对通行效率的要求更高,相比于普通汽车,应该具有更高的优先级来通过路口[14].在车联网场景下,传统的控制特殊车辆优先通行的方法大多是基于信号抢占策略,通过识别路网中特殊车辆的位置、速度与周围车流信息,切换交通灯相位,使其能够尽快通过路口.但这种方式往往会对路口的整体流量造成过大的干扰,可能会导致道路发生大范围拥堵,进而引起整个道路交通网崩溃.所以,如何在强化学习算法中引入优先级策略,平衡特殊车辆与普通车辆的通行效率,是本文重点关注与解决的问题.

基于以上现状,为了适应动态交通流变化,并在控制特殊车辆优先通行的同时减少对普通车辆的干扰,本文提出一种面向优先车辆感知的交通灯优化控制方法,使用强化学习方法学到能够适应动态交通流变化的交通灯控制策略.为了实现优先车辆感知,在设置状态时,用不同的值对特殊车辆与普通车辆进行区分,并在计算奖励时赋予特殊车辆更大的权重,以实现对特殊车辆的优先处理.此外,本文使用了 Dueling DQN[15]结构来提高模型的学习效果,并在训练过程中使用 Double DQN[16]方法来避免过度估计问题.为了验证本文方法的有效性,使用城市交通模拟器 SUMO[17]分别在单路口场景与多路口场景中进行实验.结果表明,本方法能够有效地提升路口通行效率,在优先降低特殊车辆的等待时间的同时,也能对普通车辆的等待时间有一定的优化,并且能够应用于多路口场景中.

本文第 1 节主要介绍目前已有的利用深度强化学习方法控制交通灯以及控制特殊车辆优先通行的相关工作.第 2 节从问题定义和算法模型两方面详细阐述本文提出的面向优先车辆感知的交通灯优化控制方法,详细阐述状态、动作、奖励函数设置,以及本文所使用的 Q 网络结构、模型架构与算法.第 3 节通过在城市交通模拟器 SUMO 上进行对比实验,验证本文方法能够在提高车辆通行效率的同时,体现出对特殊车辆的优先处理,并且能够扩展应用于多路口场景.第 4 节对本文工作做出总结并给出未来的工作展望.

1 相关工作

智慧交通灯控制是构造智慧城市、解决城市交通问题的一个重要研究方向.在众多研究方法中,深度强化学习以其根据实时反馈来调整动作的特征得到了广泛的应用.这类方法通常把路口交通灯抽象成一个智能体,控制对象为道路交通网络上的时变交通流,并且将智能体与控制对象的闭环交互过程抽象成马尔可夫决策过程(Markov decision process,简称 MDP)[18]:智能体将目标优化过程按照时间进程划分为状态相互联系的多个阶段,并在每个阶段通过观察交通环境的实时状态,提取交通灯控制所需的交通状态信息和反馈奖励信息进行最优决策.Wei 等人[19]提出一种使用深度 Q 神经网络的交通信号控制方法,它综合使用队列长度、车辆数量、车辆等待时间、路口图像表示以及当前相位作为状态输入,以是否切换相位作为输出,其优化目标在于缩减车辆队列长度以及等待延迟,缩短旅行时间;Joo 等人[20]提出一种能够处理多种路口结构的基于 Q 表的强化学习方法,它将队列长度和路口吞吐量作为评价指标,其优化目标在于缩短车辆在路口的延迟;Zhang 等人[21]将基于值的元强化学习方法应用于交通灯控制场景中,它利用从已有的场景中学来的知识来加快在新场景中的学习过程,提高了训练效率.以上方法的关注点都在通行效率上,Yan 等人[22]则认为,效率和公平性都应该被考虑到.因此,他们在设计奖励函数时添加了公平性考量,以降低各辆车之间的旅行时间差异.但总体而言,以上这些方法都只针对于普通车辆通行的路口场景,它们将所有种类的车都一视同仁,没有考虑到特殊车辆的优先通行性.

现有的控制特殊车辆优先通行的方法大多都是通过数学计算预测特殊车辆到达路口的时间,然后更改交通灯相位使其无需停车等待通过路口.比如 Qin 等人[23]提出的控制策略,在传感器检测到特殊车辆到达时,切换交通灯为紧急车辆抢占(emergency vehicle preemption,简称 EVP)模式,即打断正常的交通灯相位,为特殊车辆提供绿灯指引,直至特殊车辆离开路口才恢复到正常模式;Kang 等人[24]提出一种交通信号协调方法,通过修改路口之间的相位偏移量来为特殊车辆构建绿波带,保证了特殊车辆在一段区域内的无障碍通行;Noori 等人[25]则提出一种基于连接车辆的控制策略,在特殊车辆到达路口之前就抢占交通灯相位,清除该方向的车流队列,确保特殊车辆不被其前面的车阻塞;Mei 等人[26]则利用公交信号优先与动脉信号协调相结合的方法,适用于带有公交专用道的道路场景;Younes 等人[27]使用一种动态交通灯调度算法,能够应对多辆特殊车从不同方向驶入路口的情况,选择更拥堵的车流通过路口.然而这些方法在设计过程中都只着眼于满足特殊车辆的通行需求,而不顾普通车辆的通行效率,所以很有可能引起普通车辆的大范围阻塞,从而使得路网瘫痪.而这也会进一步影响到特殊车辆的通行,降低了路口的总体通行量.此外,这些方法也只适用于特殊车辆偶尔出现的情况,若是对于诸如消防局、医院、警局附近的路口,特殊车辆出现的频率相对较高,这些方法就无能为力了.

基于以上情况,为了能够赋予特殊车辆优先通行权,并且尽可能地减小对普通车流的影响,本文采用基于 Q 值的强化学习方法,在设置状态和奖励函数时增加特殊车辆的权重,使其在与环境不断交互的过程中学到一种能够平衡特殊车辆与普通车辆通行效率的策略.

2 提出的方法

2.1 问题定义

在基于强化学习的交通灯控制方法中,将交叉路口中结合了控制算法的交通灯抽象为智能体(agent),被控对象为道路交通网络中的环境(environment).如图 1 所示,在任意时刻 t,智能体从环境中获取当前环境的状态 s_t,并执行一个动作 a_t,在下一时刻 t+1,环境在动作 a_t 的作用下会产生新的状态 s_{t+1};同时,智能体也会接收到一个回报 r_{t+1}.在这个不断交互的闭环系统中,强化学习模型跟踪评测智能体所选择动作的控制效果,并以累积奖励值最大化为目标来优化信号控制策略.将这一过程抽象为一个马尔可夫决策过程,用一个五元组(S,A,P,R,γ)表示.其中,

- S:表示环境中的状态集合.$s_t \in S$ 表示环境在 t 时刻的状态;
- A:表示智能体能够执行的动作集合.$a_t \in A$ 表示智能体 t 时刻采取的动作;
- P:表示状态转移概率.假设 t 时刻系统的状态为 s_t,智能体执行的动作为 a_t,系统将根据状态转移概率 $P(s_{t+1}|s_t,a_t)$到达下一个状态 s_{t+1};

- R:表示奖励.r_t表示在执行完动作 a_t之后得到的即时奖励;
- γ:表示奖励衰减因子.$\gamma\in[0,1)$表明了未来的回报相对于当前回报的重要程度.

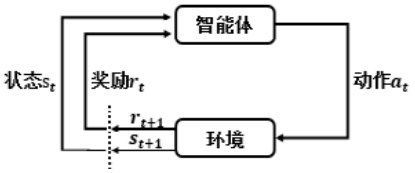

Fig.1 Interaction process between agent and environment

图 1 智能体与环境的交互过程

强化学习的目标是给定一个马尔可夫决策过程,寻找最优策略.策略π即是一连串的状态到动作的映射,它是指给定状态 s 时,动作集上的一个分布,如式(1)所示.

$$\pi(a|s)=P(a_t=a|s_t=s) \tag{1}$$

在策略π中,状态-行为值函数(也被称为 Q 函数),即累积奖励在状态 s 及动作 a 处的期望可用式(2)表示.

$$Q^{\pi}(s,a)=E[r_t+\gamma\cdot r_{t+1}+\gamma^2\cdot r_{t+2}+...|\,s_t=s,a_t]=E[\sum_{k=0}^{\infty}\gamma^k\cdot r_{t+k}\,|\,s_t=s,a_t=a,\pi] \tag{2}$$

根据上式,如果智能体知道后续状态的最优 Q 值,即后续状态选哪个动作能够使得 Q 函数输出最大,那么最优策略仅需要选择能够获得最高累积奖励的动作.用贝尔曼(Bellman)方程[28]表示如式(3)所示,最优策略π^*可以通过递归计算获得.

$$Q^{\pi^*}(s_t,a_t)=E_{s_{t+1}}[r_t+\gamma\cdot\max_{a_{t+1}}Q^{\pi^*}(s_t,a_t)\,|\,s,a] \tag{3}$$

在状态空间有限的情况下,该公式可以通过动态规划求解.但在本文场景中,路口状态空间复杂,因此采用神经网络拟合函数 $f(\theta)$来近似计算 $Q(s,a)$.具体过程是:定义一个深度神经网络——Q 网络,输入是状态 s,输出是包含每一个动作的 Q 值的向量.此时,智能体根据 Q 值的输出选择某个动作执行,并从环境中得到当前动作获得的奖励.Q 网络根据奖励计算损失函数反向传播对参数θ进行训练,直至收敛.

综上所述,状态、动作、奖励是 Q 网络设计与实现过程中不可或缺的三要素.其中,状态是从环境中获取的信息,它作为 Q 网络的输入;动作是智能体的行为表征,它决定了 Q 网络的输出维度;奖励是环境对于动作的反馈,它用于辅助 Q 网络的训练.下文将分别介绍本文场景中的状态、动作及奖励设置.

2.1.1 状态设置

参考目前大多数的路口场景,本文所讨论的道路交叉口如图 2 左图所示,路口是四路交叉路口,分别是东、西、南、北这 4 个方向.每个方向上的入向道路分为 3 个车道,按图中指向箭头所示,最右边车道允许直行和右转,中间车道仅允许直行,最左边车道仅允许左转,即每个方向上的入向车道有 4 种方向的车流.每个交通灯有红、绿、黄这 3 种状态,每个交通灯只能控制一个方向的车流,因此,控制图 2 所示的路口需要 4×4=16 个交通灯.对各个入向道不同方向所显示的不同灯色的组合构成一个信号相位.在道路中行驶的汽车用图中所示的不同形状表示,其中,普通车辆由三角形表示,特殊车辆由矩形表示,它们都遵循统一的交通规则,即按照红绿灯指示行驶.

本文根据车辆在路口的位置以及速度来定义状态信息.通过车载网络以及其他道路传感器等设备,车辆的位置以及速度很容易就能获得[29].把每条进入路口方向的车道划分成一个个小格子,格子的宽度即为车道宽度,格子的长度即为每辆车的长度加相邻车辆之间的最短距离,这样就可以保证每个格子上最多只能放下一辆车.对于每个格子,使用一个二元组(p,s)来表示该格子上的小车状态,其中,

- p 表示该格子上是否存在小车以及存在哪种类型的小车,取值集合为{0,1,10}:如果格子上的车为普通车辆,那么 p=1;如果格子上的车为特殊车辆,那么该格子上 p=10;否则,如果格子上没有车,那么 p=0;
- s 表示该格子中小车的速度,单位为 m/s.当 p=0 时,s 也等于 0,否则为该格子上小车的当前速度.

根据以上定义,图 2 左图的路口环境对应的状态如右图所示.

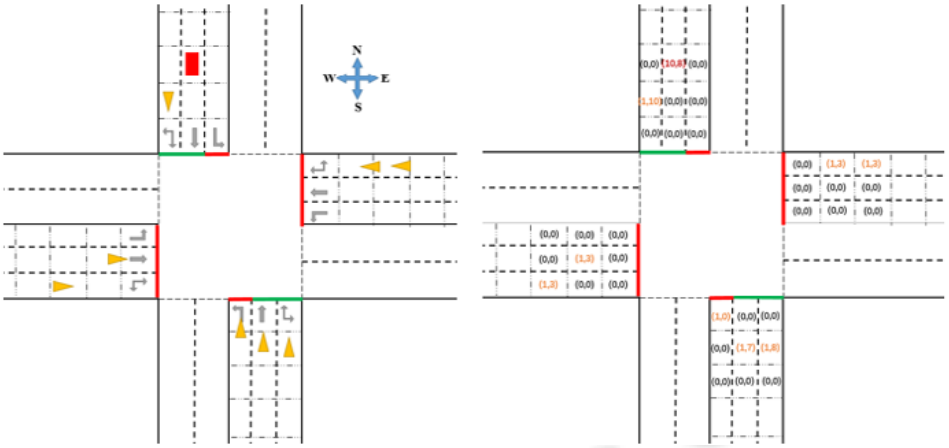

Fig.2 Intersection scenario and state setting

图 2 交叉口场景及状态设置

2.1.2 动作设置

动作就是智能体所采取的行为.在本文的问题定义中,交通灯作为智能体,它所执行的动作就是设置不同的信号灯相位.针对图 2 所示的路口场景,一共存在 4 种不冲突的相位,如图 3 所示,分别是:(1) 南北方向直行及右转;(2) 南北方向左转;(3) 东西方向直行及右转;(4) 东西方向左转.因此,智能体的动作空间为{0,1,2,3}.为了使红绿灯状态更稳定,每隔 10s 来计算一次动作,当新选择的相位与当前相位不同时,会在 4s 的黄灯时间后再切换到下一相位.此外,设置每个相位的持续时间不得超过 60s,保证其他方向的车辆及行人的等待时间在可容忍的范围内.

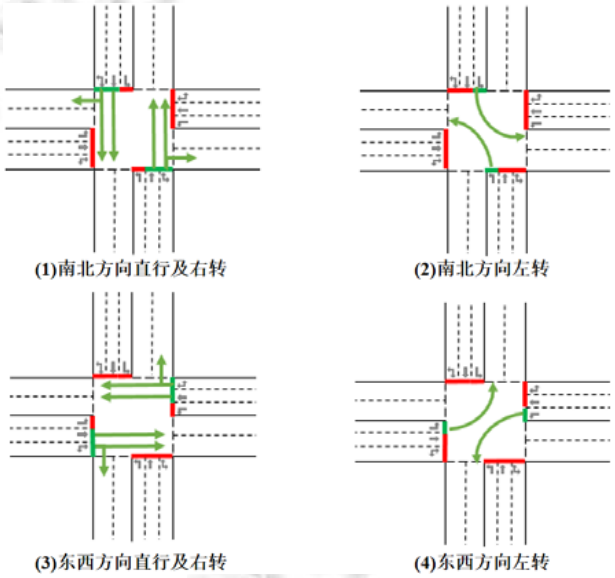

Fig.3 Action setting

图 3 动作设置

2.1.3 奖励设置

设置奖励的作用是向强化学习模型提供动作执行结果的反馈.恰当的奖励设置能够正确地指导学习过程,以使智能体学习到最佳的行动策略.衡量路口通行效率的指标通常有队列长度、路口吞吐量、车辆通行时间和车辆等待时间.其中,队列长度是指某一时刻路口各个车道上等待的车队长度,路口吞吐量是指单位时间内通过

路口的车辆数目.这两种指标在计算过程中没有对不同车辆进行区分,更适合衡量路口整体效率,无法体现出特殊车辆与普通车辆的差异.车辆通行时间是指车辆从驶入路网到驶出路网所用的时间,它虽然可以针对特殊车辆与普通车辆分别计算,但它适用于在一轮训练结束后来计算,如果作为奖励函数指标每步计算的话,会有许多车辆并没有驶出路网,此时计算结果就会有比较大的偏差.而车辆等待时间既可以对特殊车辆和普通车辆分别计算,并且它不需要车辆驶出路网后才能计算,可以在智能体每一次执行动作后更新,因此本文采用车辆等待时间作为奖励函数指标.同时,为了消除普通车辆与特殊车辆在数量上的差异所带来的影响,分别使用两种车辆的平均等待时间作为优化指标.

令 $N_{t,\text{normal}}$ 表示在第 t 次动作执行结束后路口相连的入向道路上所有普通车辆的数量,$N_{t,\text{special}}$ 表示在第 t 次执行结束后路口相连的入向道路上所有特殊车辆的数量,$W^t_{i,\text{normal}}$ 表示在第 t 次动作执行结束后观测到的第 i 辆普通车辆在该车道的累积等待时间,$W^t_{i,\text{special}}$ 表示在第 t 次动作执行结束后观测到的第 i 辆特殊车辆在该车道的累积等待时间.普通车辆及特殊车辆的平均等待时间计算方式分别如式(4)、式(5)所示.

$$AVG\_W^t_{\text{normal}} = \frac{1}{N_{t,\text{normal}}} \cdot \sum_{i=1}^{N_{t,\text{normal}}} W^t_{i,\text{normal}} \tag{4}$$

$$AVG\_W^t_{\text{special}} = \frac{1}{N_{t,\text{special}}} \cdot \sum_{i=1}^{N_{t,\text{special}}} W^t_{i,\text{special}} \tag{5}$$

智能体在执行完第 t 次动作后得到的奖励就可以用式(6)计算得到,其中,α代表特殊车辆所占的权重,取值区间为(0,1).

$$r_t = \alpha \cdot (AVG\_W^{t-1}_{\text{special}} - AVG\_W^t_{\text{special}}) + (1-\alpha) \cdot (AVG\_W^{t-1}_{\text{normal}} - AVG\_W^t_{\text{normal}}) \tag{6}$$

根据公式可以看出,如果在执行完一次动作之后,发现车辆的平均等待时间比上一次要小,这就意味着有部分等待的车辆通过了路口,智能体将得到一个正值的 *reward*.强化学习算法的目标是使 *reward* 最大化,这就会使平均等待时间朝着更小的方向优化.参数α可以用来调节特殊车辆与普通车辆在优化过程中所占的权重.

2.2 本文算法

2.2.1 *Q* 网络结构

根据第 2.1.1 节的状态设置,每一时刻随着交通灯相位的改变及车辆的行驶,状态都会发生变化,所以状态空间是无限的.因此,本文采用深度神经网络即 *Q* 网络来近似计算 *Q* 值,其结构如图 4 所示.其中,*n* 代表场景中的路口个数.

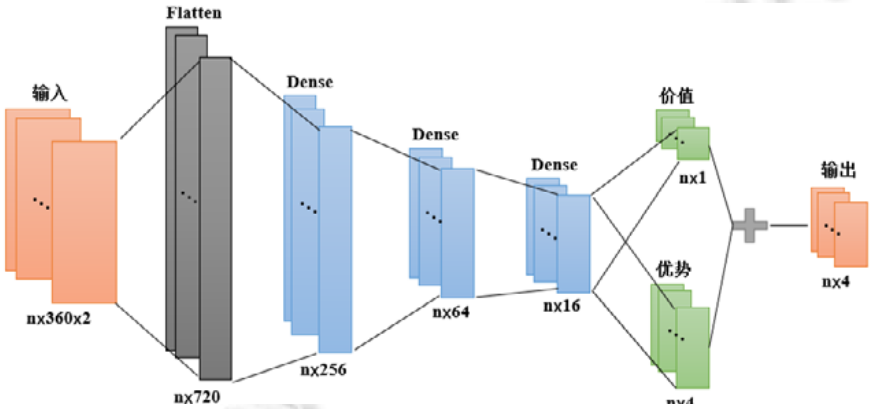

Fig.4 Structure of our *Q* network

图 4 *Q* 网络结构

模型的输入是从环境中获得的各个路口的状态,维度为(*n*,360,2),其中,360=12×30,12 代表 12 条入向车道,30 表示对每条车道,本文仅考虑距离路口最近的 30 个格子;2 代表每个格子上的(*p*,*s*)二元组.输出是当前状态下各

个路口不同动作对应的 Q 值,维度为$(n,4)$,其中,4 对应第 2.1.2 节定义的 4 种动作.

在设计网络结构时,本文采用了 Dueling DQN[15]的思想,它与传统 DQN 的不同之处在于,它把 Q 值的计算分为两部分:一部分是价值网络,它只与当前状态 s 有关,而与具体要采取的动作无关,所以它的维度是$(n,1)$,用$V(s;\omega,\alpha)$表示;另一部分是优势网络,它表示执行每种动作的优势值大小,因此它不仅与当前状态 s 有关,也与具体要执行的动作 a 相关,所以它的维度应该与输出层的维度相同,都是$(n,4)$,用$A(s,a;\omega,\beta)$表示.其中,ω表示公共部分的网络参数,α表示价值网络独有部分的参数,β表示优势网络独有部分的参数.

此时 Q 值的输出由价值网络的输出和优势网络的输出线性组合得到,在状态 s 下,每个动作 a 的 Q 值等于状态 s 的价值 V 与动作 a 的优势值之和.此外,为了使结果更加稳定,这里对优势值 A 做了一个中心化处理:对于每种动作,都将它的 A 值减去所有动作的平均 A 值.计算公式如式(7)所示.

$$Q(s,a;\omega,\alpha,\beta)=V(s;\omega,\alpha)+\left[A(s,a;\omega,\beta)-\frac{1}{|A|}\sum_{a}A(s,a;\omega,\beta)\right] \tag{7}$$

其中,价值网络 $V(s;\omega,\alpha)$体现了当前状态对 Q 值的影响;优势网络 $A(s,a;\omega,\beta)$体现了在当前状态下,不同的动作对 Q 值的影响.综合使用二者,能够使 Q 值的计算结果更准确.

2.2.2 模型架构

本文提出的算法模型架构如图 5 所示,其中最核心的两个组件分别是预测网络 Q 和目标网络 Q',二者都采用图 4 所示的 Q 网络结构.其中,预测网络是我们要训练的网络,它始终持有最新参数,用来计算预测 Q 值 Q_{eval}.目标网络的作用是用来指引训练方向,每隔一定的训练轮数,会将预测网络的参数都赋值给目标网络.

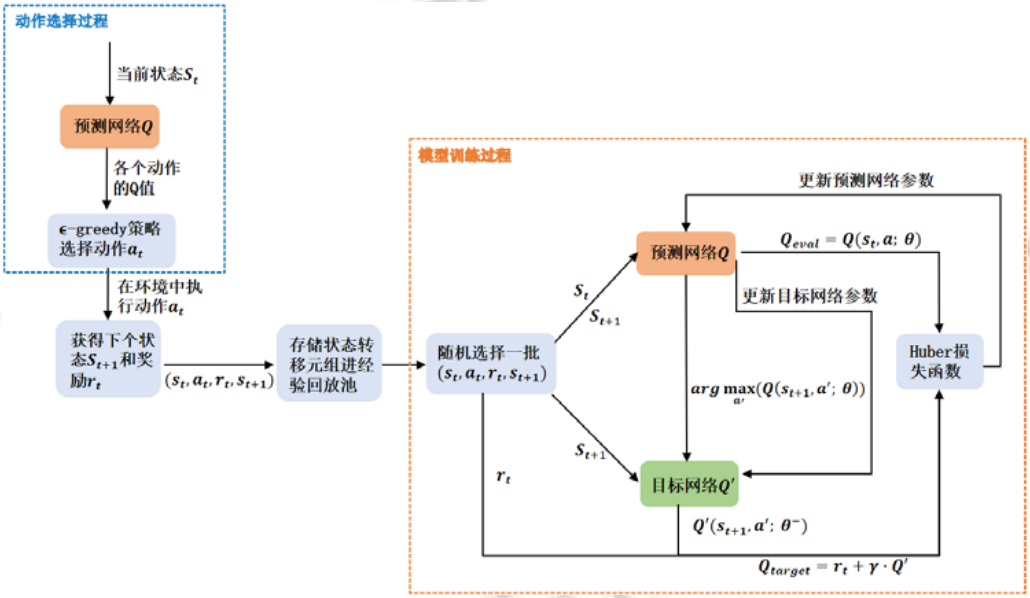

Fig.5 Model architecture

图 5 模型架构

在动作选择过程中,首先从环境中获得当前状态 s_t,并将其输入到预测网络 Q 中,计算得到当前状态下各个动作的 Q 值,最后利用 ε-greedy 算法[30]选择出要执行的动作 a_t.ε-greedy 算法是一种加入了随机因子的贪心算法,目的是增加智能体的探索尝试.智能体在选择动作时,会以概率 ε 随机在动作空间中选择一个动作,以概率 $1-\varepsilon$ 按照预测网络预测的最优 Q 值选择动作.在训练初始阶段,由于 Q 网络还不稳定,此时给 ε 赋一个较大的值能够帮助智能体做出更多的探索尝试;随着训练的进行,Q 值的预测结果会越来越准确,ε 的值也会随之减小,让智能体逐步相信预测 Q 网络的判断.

在模型训练过程中,本文采用了 Double DQN[16]的思想,将目标 Q 值动作的选择与目标 Q 值的计算这两步

解耦开.传统的 DQN 在计算目标 Q 值 Q_{target} 时,直接在目标网络输出中找出各个动作的最大 Q 值,这样虽然可以快速让 Q 值朝着可能的优化目标靠拢,但很容易导致过度估计问题.而 Double DQN 在计算目标 Q 值时,先在预测网络 Q 中找到最大 Q 值对应的动作,然后再利用这个选择出来的动作在目标网络 Q'中去计算 Q_{target},降低动作选择与目标 Q 值计算之间的相关性,能够有效地避免过度估计问题.在 Double DQN 中,目标 Q 值的计算方式如式(8)所示.

$$Q_{\text{target}}(s_t,a_t)=r_t+\gamma\cdot Q'\left(s_{t+1},\arg\max_{a'}(Q(s_{t+1},a';\theta));\theta^-\right) \tag{8}$$

其中,s_t 和 a_t 分别代表当前的状态和动作,r_t 代表当前这步行为获得的即时奖励,s_{t+1} 和 a'分别代表下一步的状态和选择的行为,γ代表奖励衰减因子.在计算目标 Q 值时,先在当前网络中获得最大 Q 值对应的动作,再将该动作放到目标网络中计算 Q 值,最后与衰减因子相乘后,加上当前奖励作为目标 Q 值.

在计算出 Q_{target} 之后,本文算法的损失函数参照 Huber 损失函数[31],定义如式(9)所示.

$$Loss(Q_{\text{target}},Q_{\text{eval}})=\begin{cases}\dfrac{1}{2}(Q_{\text{target}}-Q_{\text{eval}})^2, & \text{for } |Q_{\text{target}}-Q_{\text{eval}}|\leqslant 1\\ |Q_{\text{target}}-Q_{\text{eval}}|-\dfrac{1}{2}, & \text{otherwise}\end{cases} \tag{9}$$

若目标 Q 值与预测 Q 值差值不大于 1,损失值等于目标 Q 值与预测 Q 值差值平方的 1/2;否则,损失值等于目标 Q 值与预测 Q 值差值绝对值减去 1/2.

在选择样本进行训练时,本文采用了经验回放机制[32].之所以采用这一方法,是因为在学习过程中得到的样本前后之间是有依赖关系的,样本之间关联性过大.而深度神经网络作为有监督学习模型,要求数据满足独立同分布,经验回放机制通过“存储-采样”操作能够很好地打破相邻样本之间的数据关联性.具体做法是,从以前的经验回放池中随机采样进行训练.这样不仅能够提高样本利用率,使一个样本能够被多次使用,也能减少参与训练的样本之间的相关性.

2.2.3 算法描述

本文所使用的模型训练算法伪代码如下所示.

算法 1. 模型训练算法.

输入:经验回放大小 M,批大小 B,动作选择系数$\varepsilon_{\max},\varepsilon_{\min},\varepsilon_{\text{decay}}$,奖励衰减系数$\gamma$,训练轮数 E,每轮最大步数 T,目标网络更新频率 F.

初始化网络参数θ,θ^-为随机值;初始化经验回放池 m 容量为 M;初始化动作选择系数$\varepsilon=\varepsilon_{\max}$
For *episode*=1,E **do**:
  获得路口初始状态 s
  For *step*=1,T **do**:
    以概率ε选择一个随机动作 $a\in\{0,1,2,3\}$;否则, $a=\arg\max_a(Q(s,a;\theta))$
    执行动作 a,得到奖励 r 以及下一状态 s'
    存储状态转移四元组(s,a,r,s')进 m
    更新状态 $s\leftarrow s'$
    更新动作转移系数$\varepsilon\leftarrow\max(\varepsilon_{\min},\varepsilon\cdot\varepsilon_{\text{decay}})$
    从 m 中随机获取 B 个状态转移元组 m'
    For (s_j,a_j,r_j,s_{j+1}) in m' **do**
      计算 $Q_{\text{eval}}(s_j,a_j)=Q_{eval}(s_j,a_j;\theta)$
      计算 $Q_{\text{target}}(s_t,a_t)=r_t+\gamma\cdot Q'\left(s_{t+1},\arg\max_{a'}(Q(s_{t+1},a';\theta));\theta^-\right)$
      计算 $Loss(Q_{\text{target}},Q_{\text{eval}})$
      根据 *Loss* 值,利用 Adam 优化器更新参数θ

16 **End For**
17 **End For**
18 **If** *episode*%*F*==0 **then**
19 更新目标网络参数$\theta' \leftarrow \theta$
20 **End If**
21 **End For**

在每一轮开始,先从环境中获得初始状态,再利用ε-greedy 算法,以概率ε随机在动作空间中选择一个动作,以概率$1-\varepsilon$,按照预测网络预测的最优 Q 值选择动作.在智能体执行完动作 a 之后,会得到环境反馈的即时奖励 r 以及下一状态 s'.此时,本文会采取基于"存储-采样"操作的经验回放机制,将状态转移四元组(s,a,r,s')存储进经验回放池,再从池中随机获取 B 个样本.对于每个样本,分别利用预测网络和目标网络计算出它的预测 Q 值 Q_{eval} 以及目标 Q 值 Q_{target};最后,根据第 2.2.2 节定义的损失函数,使用 Adam 优化器[33]反向传播更新预测网络参数θ.预测网络的参数每步都会更新,而目标网络的参数每隔 F 轮才更新.

3 实 验

为了验证本文方法的有效性及可用性.需要进行实验回答以下几个问题.

- 问题 1:本文方法是否能够提高车辆的通行效率?即在同一车流场景下,使用本文方法控制的交通灯与使用传统固定时长控制的交通灯对比,对车辆的通行效率是否有明显提升?
- 问题 2:本文方法是否能够体现出对特殊车辆的优先性?即在区分车辆优先级的情况下,特殊车辆的平均等待时间是否会比不区分车辆优先级情况下的平均等待时间更短?
- 问题 3:本文方法是否能应用于不同的路口场景?即本文方法是否能够有效地扩展到多路口场景中?

3.1 实验设计

实验使用城市交通模拟器 SUMO 来完成,它能够协助我们设计和实现道路设施的自定义配置与功能,并能在仿真运行期间提供关于车辆及交通灯的实时数据.本文把训练过程分为多轮进行,每轮 3 600 步,每步代表现实场景中的 1s,所以说,每轮相当于现实场景中的 1h.为了更好地模拟现实场景中的随机车流,在实验中设置车流以固定比例及流量随机插入网络,即特殊车辆与普通车辆的比例与流量固定,但车辆驶入的位置与行驶路线由 SUMO 随机生成.

为了回答以上 3 个问题,本文进行了如下的实验设计.

- 针对问题 1,需要进行对比实验 1.

在同一路口场景下,分别使用两种红绿灯控制方案:一个是本文方法训练的模型,另一个是使用固定时长控制.对比两种控制方式下,普通车辆与特殊车辆的平均等待时间差异.在实验中,路口环境使用图 2 所示的四路交叉路口结构,每个方向的入向道路有 3 个车道,每条车道的长度设为 300m.车流由 SUMO 随机生成,为了更符合实际场景中特殊车辆的数目一般都远小于普通车辆的事实,本实验设置特殊车辆与普通车辆的比例为 1:200,设置普通车辆每秒驶入两辆,特殊车辆每100s驶入一辆.设置固定时长交通灯每隔30s切换相位,切换顺序为"南北方向直行及右转→南北方向左转→东西方向直行及右转→东西方向左转"循环.

- 针对问题 2,需要进行对比实验 2.

在同一路口场景下,使用同一种模型结构,对比区分车辆优先级的模型与不区分车辆优先级的模型对特殊车辆的等待时间的影响.本实验采用实验 1 相同的路口设置与车流设置.区分车辆优先级的模型使用的状态和奖励设置参照前文第 2.1.1 节及第 2.1.3 节的定义.不区分车辆优先级的模型在设置状态时,只要格子上有车,不区分该车的类别,均将该格子上的二元组(p,s)中的 p 值设为 1.在定义奖励时,统一计算特殊车辆与普通车辆的平均等待时间,修改奖励计算公式如式(10)所示.

$$r_t = AVG\_W_{\text{all}}^{t-1} - AVG\_W_{\text{all}}^{t} \tag{10}$$

其中, $AVG\_W_{all}^{t}$ 表示执行完第 t 次动作后所有车辆的平均等待时间.

- 针对问题 3,需要进行对比实验 3.

在同一个多路口场景中,分别使用本文方法训练的模型与固定时长方法控制交通灯,对比两种控制方式下车辆的通行效率.在本实验中,使用 3×3 一共 9 个交通灯控制路口的环境设置,相邻路口之间相互连通.每个路口的环境设置与对比实验 1 相同.与实验 1 类似,本实验的车流也由 SUMO 随机生成,但本实验的车流均从边界驶入路网,并且经由边界驶出路网.设置特殊车辆与普通车辆的比例为 1:500,令普通车辆每秒驶入 5 辆,特殊车辆每 100s 驶入一辆.固定时长交通灯控制的设置与实验 1 相同.

实验所使用的 CPU 型号为 Intel i7-8700k,使用的图形处理器型号为 GeForce GTX 1080Ti,操作系统为 Ubuntu 18.04,Python 版本为 3.7,机器学习平台为 Tensorflow v1.14.0 以及 Keras v2.1.0.使用的城市交通模拟器为 SUMO v1.6.0.实验参数设置如下:模型训练采用 Adam 优化器,学习率设为 0.000 1.设置经验回放池大小为 2 000,训练批次大小 *batch_size* 为 64.设置奖励衰减系数 γ 为 0.8.设置初始最大动作选择系数 $\varepsilon_{\max}$ 为 1,动作系数衰减率 $\varepsilon_{\text{decay}}$ 为 0.95,最小动作选择系数 $\varepsilon_{\min}$ 为 0.01.

3.2 实验结果分析

3.2.1 实验 1:单路口场景中本文算法与固定时长方法的结果对比

为了回答问题 1,我们在单路口进行实验,对比在同一种环境下,使用本文提出的方法控制交通灯与使用固定时长逻辑控制交通灯对车辆通行效率的影响.在本实验中,设置奖励平衡系数 α 为 0.6.本实验主要评价两个通行效率指标.

- 一是平均等待时间,以 s 为单位,它的值等于在一轮训练过程中所有经过路口的车在该路口的等待时间的总和除以车的数量.这里,为了区分对不同优先级的车辆的效果,分别计算特殊车辆与普通车辆的平均等待时间.
- 二是平均队列长度,以辆为单位,它是指在这一轮训练过程中,每一步各个入向车道排队总长度的平均值[34].

对比实验 1 训练了 150 轮,实验结果如图 6 所示.

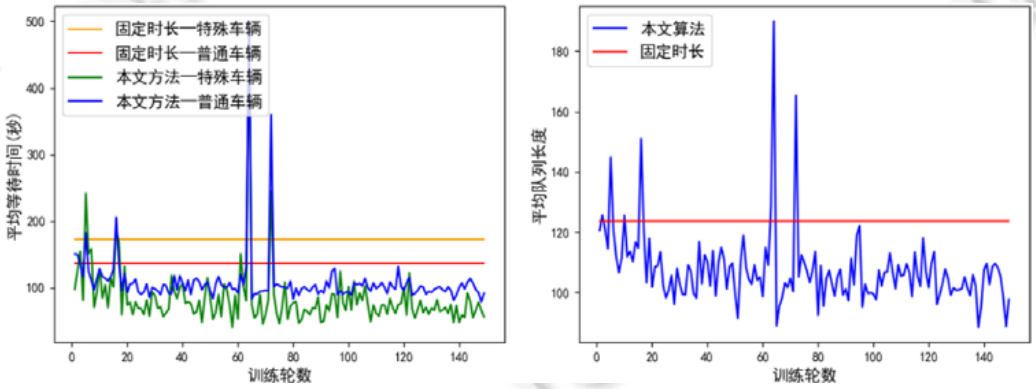

Fig.6 Results comparison of our algorithm and fixted-time method in single intersection secnario

图 6 单路口场景中,本文算法与固定时长方法的结果对比

左图表示特殊车辆与普通车辆的平均等待时间对比,可以看出,在使用固定时长逻辑控制交通灯时,特殊车辆的平均等待时间为 170s 左右,普通车辆的平均等待时间为 135s 左右.之所以出现特殊车辆的平均等待时间大于普通车辆的情况,原因有两点:一是排在前面的普通车辆影响了在其后面的特殊车辆的通行;二是特殊车辆总体数量较少,单个车辆对平均等待时间的影响较大.在使用本文方法训练的模型对交通灯进行控制之后,特殊车辆的平均等待时间降到了 65s 左右,普通车辆的平均等待时间降到了 90s 左右.与固定时长逻辑相比,特殊车辆的平均等待时间降低了 105s 左右,普通车辆的平均等待时间降低了 45s 左右.

右图表示随着训练轮数的增加,路口处平均队列长度的变化.在使用固定时长逻辑时,每一轮路口的平均队列长度为 125 辆左右;但在使用经本文方法训练的模型之后,可以将平均队列长度降低到 100 辆左右.因此,图 5 的实验结果可以证明,本文方法能够显著提高车辆的通行效率,并且对特殊车辆的优化效果要好于普通车辆.

3.2.2 实验 2:单路口场景中考虑优先级与不考虑优先级的结果对比

为了回答问题 2,我们对比不考虑优先级与考虑优先级的情况下,特殊车辆和普通车辆的平均等待时间的差距.本实验使用的路口环境与对比实验 1 相同,在考虑优先级的方法中,设置奖励平衡系数α为 0.6.实验结果如图 7 所示,左右两图分别表示特殊车辆和普通车辆的平均等待时间在不考虑优先级与考虑优先级的情况下的结果对比,其中,黄色线条表示在采用不考虑优先级的方法下的实验结果,绿色线条表示采用本文方法,即考虑车辆优先级的方法控制下的实验结果.对比可以看出,在训练趋于稳定之后,对于特殊车辆而言,本文设计的优先级控制策略的确能够降低其平均等待时间,对比数值降低幅度大概在 35s 左右;对于普通车辆而言,使用本文设计的优先级控制策略虽然使得其平均等待时间有小幅度上升,但其大概 15s 左右的上升于特殊车辆 35s 的下降而言是值得的,因为在实际场景中,特殊车辆对通行时间的要求更高,对等待时间的容忍力更低.此外,虽然普通车辆的平均等待时间相比于不考虑优先级的方法有小幅度上升,但与实验 1 中使用固定时长控制方法对比还是有明显下降的.因此,图 7 的实验结果表明,本文所设计的优先级控制策略能够体现出对特殊车辆的优先性.

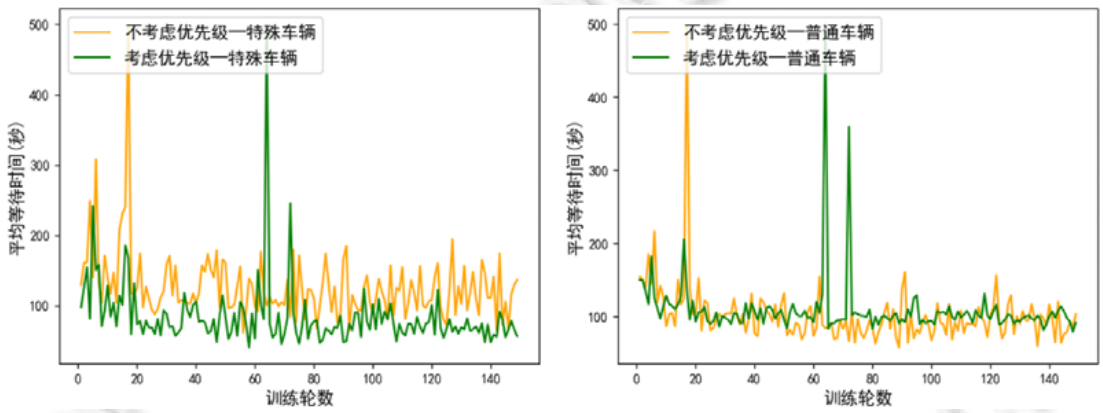

Fig.7 Results comparison of considering priority and not considering priority in single intersection secnario

图 7 单路口场景中,考虑优先级与不考虑优先级的结果对比

3.2.3 实验 3:多路口场景中本文算法的表现

为了回答问题 3,我们将本文方法扩展到 3×3 的联通路口场景中,并进行了与单路口场景下类似的对比实验.通过对比使用固定时长与使用本文方法的交通灯控制策略对特殊车辆和普通车辆的平均等待时间的影响,验证本文方法对多路口场景的有效性,该实验的结果如图 8 左图所示;通过对比不考虑优先级与考虑优先级的方法对特殊车辆的平均等待时间的影响,验证本文的优先级控制策略对多路口场景的有效性,该实验的结果如图 8 右图所示.在本实验中,设置奖励平衡系数α为 0.65.在本实验中,评价指标平均等待时间是指在一轮训练过程中,各个路口的平均等待时间的平均值.

对比实验 3 训练了 200 轮,实验结果如图 8 所示.

- 图 8 左图显示,在使用固定时长逻辑的交通灯控制下,各路口特殊车辆与普通车辆的平均等待时间均为 150s 左右;在使用本文方法训练之后,各路口普通车辆的平均等待时间降至 125s 左右,特殊车辆的平均等待时间降至 90s 左右.可以看出,在多路口场景中,本文方法也能显著提高车辆的通行效率.
- 此外,右图结果也显示,本文的优先级设置策略在多路口场景下也能显示出一定的有效性.

因此,图 8 的实验结果表明,本文方法能够扩展应用到多路口场景中.

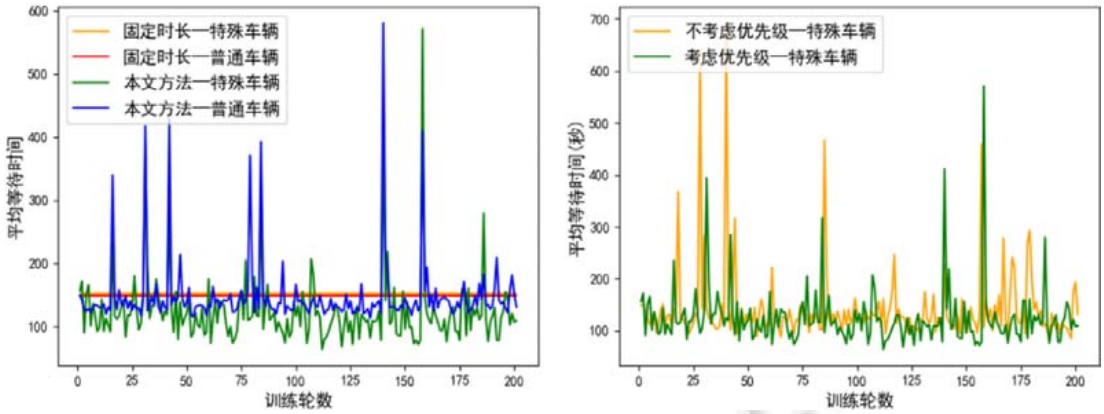

Fig.8 Results of our algorithm in multi-intersection scenario

图 8 多路口场景中,本文算法的结果

3.2.4 实验分析与总结

由于本文算法根据实时的车辆位置、类型以及速度信息作为状态,以这些实时信息作为输入的神经网络计算出的 Q 值是与当前的环境状态相关的,因此本文算法所做的决策能够适应动态交通流变化,相较于传统的固定时长控制方法能够显著提高车辆的通行效率:分别使特殊车辆和普通车辆的平均等待时间降低 68%和 22%左右;与此同时,也使路口队列长度降低了 20%左右.本文所设计的优先级机制主要体现在两个方面.

- 一是在设计状态时将各个入向道路以固定长度划分为不重合的一个个小格子,并且给特殊车辆与普通车辆设置不同的值,使得 Q 网络能够根据状态识别出车辆的位置以及类型;
- 二是在奖励设置时分别计算特殊车辆与普通车辆的平均等待时间,并设置了奖励平衡系数α来平衡特殊车辆与普通车辆的权重,使得 Q 网络在训练过程中朝着更大幅度地缩小特殊车辆的等待时间的方向收敛;同时,为了使奖励最大化,也不会使普通车辆的等待时间过大.

因此,使用本文所设计的优先级策略对比不使用优先级,在训练趋于稳定之后,能够使特殊车辆的平均等待时间降低 35%左右.

多路口场景与单路口类似,将各个路口的实时状态信息聚合起来统一输入到 Q 网络中,输出针对各个路口上不同动作的 Q 值.在训练过程中,各个路口利用自己的历史状态动作转移元组,能够学到适用于自己场景的决策,因此,本文方法能够扩展应用到多路口场景中:对比固定时长控制方法,在训练趋于稳定之后,分别使特殊车辆和普通车辆的平均等待时间降低 40%和 17%左右;同时,对比不考虑优先级的方法,特殊车辆的平均等待时间也降低了 10%左右.

4 结论与展望

在城市交通网络中,具有特殊任务的特殊车辆对于通行效率的要求更高.尽管传统的信号抢占方法考虑到了特殊车辆的优先性,但对于普通车辆的通行干扰过大.基于以上情况,本文提出了一种面向车辆优先级感知的交通灯优化控制方法,使用 Dueling DQN 结构来提高模型的学习效果,并在训练过程中使用 Double DQN 方法来避免过度估计问题.为了实现对特殊车辆的优先控制,在设置状态时,用不同的值来区分特殊车辆与普通车辆,并在计算奖励时赋予特殊车辆更大的权重,使得本文方法能够在不干扰普通车辆通行的同时,大幅度降低特殊车辆在路口的等待时间,帮助其更快到达目的地.此外,本文方法也能接收多个路口的状态输入,并给出各个路口的动作决策,能够扩展应用于多路口场景中.但由于多路口场景各个路口都是联通的,相邻路口之间的车流有一定的相关性,而本文方法没有考虑到相邻路口之间的信息交互,在多路口场景中的表现不如单路口场景中好.因此在今后的工作中,我们会将多智能体协同控制强化学习方法应用于多路口控制,以提升多路口场景下的优化效果.此外,如何高效地为不同的路口交通流路口寻找奖励平衡系数α,也是未来亟需解决的问题.

References:

[1] 2018 annual report of traffic operation. 2018 (in Chinese). http://www.jtcx.sh.cn/trafficanalyse.html

[2] Li MW, Li L. Intelligent transportation system in China: The optimal evaluation period of transportation's application performance. Journal of Intelligent & Fuzzy Systems, 2020,38(6):6979−6990.

[3] Wu LB, Nie L, Liu BY, Wu N, Zou YF, Ye LY. An intelligent traffic signal control method in VANET. Chinese Journal of Computers, 2016,39(6):1105−1119 (in Chinese with English abstract).

[4] Chang W, Roy D, Zhao S, Annaswamy A, Chakraborty S. CPS-oriented modeling and control of traffic signals using adaptive back pressure. In: Proc. of the Design, Automation & Test in Europe Conf. & Exhibition (DATE). IEEE, 2020. 1686−1691.

[5] Zhang ZK, Pang WG, Xie WJ, Lü MS, Wang Y. Deep learning for real-time applications: A survey. Ruan Jian Xue Bao/Journal of Software, 2020,31(9):2654−2677 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5946.htm [doi: 10.13328/j.cnki.jos.005946]

[6] Diakaki P, Kotsialos D, Wang Y. Review of road traffic control strategies. Proc. of the IEEE, 2003,91(12):2041−2042.

[7] Sutton RS, Barto AG. Introduction to Reinforcement Learning. Cambridge: MIT Press, 1998.

[8] Thorpe TL. Vehicle traffic light control using sarsa. 1997. http://citeseer.ist.psu.edu/thorpe97vehicle.html

[9] Xu Y, Zhang YL, Sun TT, Su YF. Agent-based decentralized cooperative traffic control toward green-waved effects. Ruan Jian Xue Bao/Journal of Software, 2012,23(11):2937−2945 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/4307.htm [doi: 10.3724/SP.J.1001.2012.04307]

[10] Lee J, Chung J, Sohn K. Reinforcement learning for joint control of traffic signals in a transportation network. IEEE Trans. on Vehicular Technology, 2020,69(2):1375−1387.

[11] Guo MY, Wang P, Chan CY, Askary S. A reinforcement learning approach for intelligent traffic signal control at urban intersections. In: Proc. of the IEEE Intelligent Transportation Systems Conf. (ITSC). 2019. 4242−4247.

[12] Yu D, Wei SG, Rong DC, Chai LG. RA-TSC: Learning adaptive traffic signal control strategy via deep reinforcement learning. In: Proc. of the IEEE Intelligent Transportation Systems Conf. (ITSC). 2019. 3275−3280.

[13] Rizzo SG, Vantini G, Chawla S. Reinforcement learning with explainability for traffic signal control. In: Proc. of the IEEE Intelligent Transportation Systems Conf. (ITSC). 2019. 3567−3572.

[14] Cao M, Shuai QQ, Li V. Emergency vehicle-centered traffic signal control in intelligent transportation systems. In: Proc. of the IEEE Intelligent Transportation Systems Conf. (ITSC). 2019. 4525−4531.

[15] Wang Z, Schaul T, Hessel M, Hasselt H, Lanctoc M, Freitas N. Dueling network architectures for deep reinforcement learning. In: Proc. of the Int'l Conf. on Machine Learning (ICML). 2016. 1995−2003.

[16] Van Hasselt H, Guez A, Silver D. Deep reinforcement learning with double Q-learning. In: Proc. of the 30th AAAI Conf. on Artificial Intelligence (AAAI). 2016. 2094−2100.

[17] Behrisch M, Bieker L, Erdmann J, Krajzewicz D. Sumo—Simulation of urban mobility: An overview. In: Proc. of the SIMUL. 2011. https://elib.dlr.de/71460/

[18] Singh T. Constrained Markov decision processes for intelligent traffic. In: Proc. of the Int'l Conf. on Computing, Communication and Networking Technologies (ICCCNT). 2019. 1−7.

[19] Wei H, Zheng G, Yao H, Li ZH. Intellilight: A reinforcement learning approach for intelligent traffic light control. In: Proc. of the 24th ACM SIGKDD Int'l Conf. on Knowledge Discovery & Data Mining (KDD). 2018. 2496−2505.

[20] Joo H, Ahmed SH, Lim Y. Traffic signal control for smart cities using reinforcement learning. Computer Communications, 2020, 154:324−330.

[21] Zang X, Yao H, Zheng GJ, Xu K, Li ZH. MetaLight: Value-based meta-reinforcement learning for traffic signal control. In: Proc. of the AAAI Conf. on Artificial Intelligence (AAAI), Vol.34. 2020. 1153−1160.

[22] Yan S, Zhang J, Buescher D, Burgard W. Efficiency and equity are both essential: A generalized traffic signal controller with deep reinforcement learning. arXiv preprint arXiv:2003.04046, 2020.

[23] Qin X, Khan AM. Control strategies of traffic signal timing transition for emergency vehicle preemption. Transportation Research Part C: Emerging Technologies, 2012,25:1−17.

[24] Kang W, Xiong G, Lv Y, Dong X, Zhu F, K QJ. Traffic signal coordination for emergency vehicles. In: Proc. of the 17th IEEE Int'l Conf. on Intelligent Transportation Systems (ITSC). IEEE, 2014. 157−161.

[25] Noori H, Fu L, Shiravi S, Noori H, Fu L, Shiravi S. A connected vehicle based traffic signal control strategy for emergency vehicle preemption. In: Proc. of the Transportation Research Board 95th Annual Meeting. 2016.

[26] Mei Z, Tan Z, Zhang W, Wang D. Simulation analysis of traffic signal control and transit signal priority strategies under arterial coordination conditions. Simulation, 2019,95(1):51−64.

[27] Younes MB, Boukerche A. An efficient dynamic traffic light scheduling algorithm considering emergency vehicles for intelligent transportation systems. Wireless Networks, 2018,24(7):2451−2463.

[28] Sutton RS, Barto AG. Reinforcement Learning: An Introduction. MIT Press, 1998.

[29] Liang X, Du X, Wang G, Han Z. A deep reinforcement learning network for traffic light cycle control. IEEE Trans. on Vehicular Technology, 2019,68(2):1243−1253.

[30] Kim CH, Watanabe K, Nishide S, Guoko M. Epsilon-greedy babbling. In: Proc. of the 2017 Joint IEEE Int'l Conf. on Development and Learning and Epigenetic Robotics (ICDL-EpiRob). 2017. 227−232.

[31] Esmaeili A, Marvasti F. A novel approach to quantized matrix completion using Huber loss measure. IEEE Signal Processing Letters, 2019,26(2):337−341.

[32] Adam S, Busoniu L, Babuska R. Experience replay for real-time reinforcement learning control. IEEE Trans. on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 2011,42(2):201−212.

[33] Kingma DP, Ba J. Adam: A method for stochastic optimization. In: Proc. of the 3rd Int'l Conf. on Learning Representations (ICLR). San Diego, 2015.

[34] Wu T, Zhou P, Liu K, Yuan Y, Wang X, Huang H, Wu DO. Multi-Agent deep reinforcement learning for urban traffic light control in vehicular networks. IEEE Trans. on Vehicular Technology, 2020,69(8):8243−8256.

附中文参考文献:

[1] 2018 年交通运行年报.2018. http://www.jtcx.sh.cn/trafficanalyse.html

[3] 吴黎兵,聂雷,刘冰艺,吴妮,邹逸飞,叶璐瑶.一种 VANET 环境下的智能交通信号控制方法.计算机学报,2016,39(6):1105−1119.

[5] 张政馗,庞为光,谢文静,吕鸣松,王义.面向实时应用的深度学习研究综述.软件学报,2020,31(9):2654−2677. http://www.jos.org.cn/1000-9825/5946.htm [doi: 10.13328/j.cnki.jos.005946]

[9] 徐杨,张玉林,孙婷婷,苏艳芳.基于多智能体交通绿波效应分布式协同控制算法.软件学报, 2012,23(11):2937−2945 http://www.jos.org.cn/1000-9825/4307.htm [doi: 10.3724/SP.J.1001.2012.04307]

邵明莉(1997－),女,硕士,CCF 学生会员,主要研究领域为强化学习,交通灯控制.

章玥(1981－),女,博士,副教授,CCF 专业会员,主要研究领域为软件定义网络,物联网.

曹鹗(1994－),男,硕士,主要研究领域为云计算工作流调度,物联网.

陈闻杰(1977－),男,博士,副教授,CCF 专业会员,主要研究领域为嵌入式系统,物联网,软硬件协同设计.

胡铭(1995－),男,博士生,CCF 学生会员,主要研究领域为程序分析与综合,CPS 系统自动化设计.

陈铭松(1982－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为嵌入式系统,软硬件协同设计,物联网,信息物理系统设计自动化.