基于 D3QN 的交通信号控制策略

赖建辉
(浙江工业大学计算机科学与技术学院 杭州 310023)

摘 要 交叉口是城市路网的核心和枢纽,合理优化交叉口的信号控制可以极大地提高城市交通体系的运行效率,而将实时交通信息作为输入并动态调整交通信号灯的相位时间成为了当前研究的重要方向。文中提出了一种基于D3QN(Double Deep Q-Learning Network with Dueling Architecture)深度强化学习模型的交通信号控制方法,其利用深度学习网络,结合交通信号控制机构成了一个用于调整交叉口信号控制策略的智能体,然后采用 DTSE(离散交通状态编码)方法将交叉口的交通状态转换为由车辆的位置和速度信息所组成的二维矩阵,通过深度学习对交通状态特征进行高层抽象表征,从而实现对交通状态的精确感知。在此基础上,通过强化学习来实现自适应交通信号控制策略。最后,利用交通微型仿真器 SUMO 进行仿真实验,以定时控制和感应控制方法作为对照实验,结果表明文中提出的方法得到了更好的控制效果,因此是可行且有效的。

关键词 智能交通,强化学习,深度学习,深度强化学习,交通信号控制
中图法分类号 TP391 **文献标识码** A

Traffic Signal Control Based on Double Deep Q-learning Network with Dueling Architecture

LAI Jian-hui
(College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou 310023, China)

Abstract The intersection is the core and hub of the urban road network. Reasonable optimization of the signal control at the intersection can greatly improve the operational efficiency of the urban transportation system. Using real-time traffic information as input and dynamically adjusting the phase time of the traffic signal becomes the important direction of current research. This paper proposed a traffic signal control method based on double deep Q-learning network with Dueling Architecture (D3QN). The deep learning network is combined with the traffic signal control machine to form an intelligent agent for adjusting the signal control strategy of the intersection. Then the DTSE (Discrete Traffic State Coding) method is used to transform the traffic state of the intersection into a two-dimensional matrix composed of the position and velocity information of the vehicle. Then high-level features are captured by deep neural network, which makes accurate perception of traffic state come true. On this basis, an adaptive traffic signal control strategy is realized through reinforcement learning. Finally, the traffic micro-simulator (SUMO) is used for simulation experiments, the timing control and induction control methods are used as control experiments. The results show that the proposed method achieves better control effect and is therefore feasible and effective.

Keywords Intelligent transportation, Reinforcement learning, Deep learning, Deep reinforcement learning, Traffic signal control

1 引言

随着经济的繁荣发展和城市化进程的推进,城市交通量猛增,研究并设计智能交通信号控制系统成为了当前交通控制方法中的重要方向。近来的交通控制系统的目标是通过预测未来交通系统的状态来提前制定适当的控制方案[1],这一要求也突出了交通系统智能化的重要性和艰巨性。当前,主流的交通控制系统有 TRANSYT,SCOOT 和 SCAT 等,这些控制系统根据道路交通流量、排队长度、车道占有比等传统交通参数对交通信号灯的配时进行优化。但是,为了进一步提高控制方法对动态交通流的适应性,人们开始尝试使用许多先进的控制理论和方法。机器学习作为当前迅速发展的理论方向,也不断被研究者们引入交通控制领域[2-3]。其中,强化学习中的 Q 学习更是在交通控制领域有着优异的表现,并逐渐成为一个研究热点[4-5]。

2 相关工作及其存在的问题

在交通动态自适应控制问题上,以往已经有许多的文献对其进行了研究,但是由于当时计算能力和仿真工具的限制,早期的工作主要集中于通过模糊控制[6-7]、线性规划[8]等方法解决问题。在这些文献中,道路交通环境是通过有限的信息进行建模的,无法应用于大规模的场景。自 20 世纪 90 年代以来,强化学习方法逐渐开始应用于交通信号控制问题中。强化学习通过学习信号控制动作与交通流变化之间的关系来隐式地对复杂交通系统动力学进行建模[9-12]。其中最经典的 Q 学习方法就是通过状态动作值函数 $Q_\pi(s,a)$ 来反映在交通状态 s 下依据某个策略 π 执行控制动作 a 后的累计回报期望,即通过 Q 值衡量在交通状态 s 下采取动作 a 的好坏程度。

赖建辉(1994—),男,硕士生,主要研究方向为智能交通中的路口信号控制,E-mail:15757116547@163.com。

El-Tantawy 等[13]总结了 1997 年至 2010 年使用强化学习来解决交通信号控制问题的方法，当时的强化学习技术仅限于表格型 Q 学习，并且通常只使用线性函数估计 Q 值，而且由于当时强化学习的技术限制，在状态空间定义中往往采用排队车辆数量[5,14-15]以及交通流量[16-17]等简单类型的数据，然而交通道路系统的复杂性往往无法通过这些信息得到完整的呈现，这导致了强化学习无法在交通信号控制中发挥最佳效果[18]。

随着强化学习[19]和深度学习[20-22]技术的发展，有学者提出将它们结合在一起作为深度强化学习[23]方法来估计 Q 值。Li 等[24]采用了深度强化学习技术中的 DQN 算法对单交叉口控制问题进行了研究，仿真实验证明，相比于固定配时策略，Li 等提出的基于 DQN 算法的信号控制策略使延时降低了 14%。然而，文献中依旧采用了排队长度作为交通信号控制问题中的状态空间，并且在仿真实验中采取了自定义的简单交叉口模型以及随机的车流量数据，无法证明算法在实际交叉口的动态交通流中也可以获得同样的控制效果。

基于以上现状，本文提出了一种基于 D3QN 深度强化学习模型的交通信号控制方法。算法通过多层神经网络来学习当在一个特定的交通状态下执行一个特定的动作时所隐含的最高未来奖励值，避免了因状态空间和动作空间较大导致的“维度灾难”问题；同时，采用 DTSE(离散交通状态编码)方法定义交通信号控制问题中的状态空间，提升了交通状态表示的准确度。最后，在仿真实验中采用了实际的交叉口的仿真模型和车流量，验证了算法在实际交叉口动态交通流中的控制效果。

3 基于 D3QN 的交通控制优化方法

在基于强化学习的交通控制方法中，强化学习模型将交叉口中结合控制算法的交通信号机抽象为一个智能体(agent)，控制对象为道路交通网络中的交通状态。智能体与被控对象在闭环系统中不断进行交互，根据实时交通状态信息(state)，选择相应的控制决策(action)并执行；进而跟踪评测所选择动作的控制效果，以累积奖励值(reward)最大化为目标，优化信号控制策略，直至收敛到“状态与动作”的最优概率映射。图 1 给出了一个交通信号智能体的强化学习标准模型。

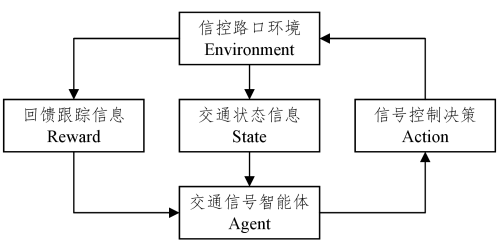

图 1 交通信号智能体的强化学习标准模型

本文基于典型道路设计了一个多车道的单交叉口模型，如图 2 所示。交通信号智能体由基于 D3QN 算法的卷积神经网络和信号控制机组成，智能体根据当前的交通状态选择信号调整方案，然后根据执行动作之后产生的新交通状态，得到一个奖励值，通过这样的迭代训练使智能体以累积奖励值(reward)最大化为目标，根据不同的交通状态自动寻求最佳的信号控制策略。这一节将主要讨论如何在交通信号控制问题中定义强化学习的 3 个元素：状态空间 State、动作空间 Action 和奖励值 Reward；同时对 D3QN 算法进行介绍。

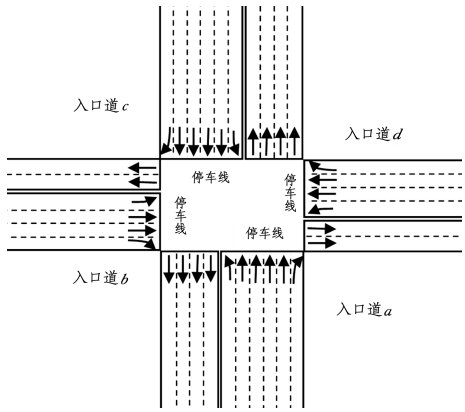

图 2 交叉口模型示意图

3.1 状态空间 s_t

在传统的交通状态定义中，一般会选择车辆排队长度或每个方向的交通流量，这些参数只能在单一方面对交通状态进行刻画，同时也会遗漏一些其他交通信息。如果将交通状态定义为排队长度，那么所有运动中的车辆的信息和静止车辆的位置信息就被遗漏了；而交通流量则只是描述了过去一段时间内车辆的通行信息，忽略了当前车辆的信息。因此，本文采用了 DTSE(离散交通状态编码)方法将交叉口的实时交通状态转化为不同的元胞(见图 3)，并将其作为输入状态。这种方法可以准确地反映交叉口附近的车辆的实时位置和速度信息，同时也可以用于预测后续的交通状态。因此，它比简单的车辆排队长度和车流量更准确地描述了当前的交通状态。

在图 2 中，以路口 a 为例，将每一条车道距离停车线的一定距离范围以一定的长度间隔分成多个方格，判断方格内是否有车，然后以 bool 值组成一个车辆位置矩阵，若某辆车跨过两个方格，则选择占用比例较大的。同时，本文根据道路限速值对车辆速度进行归一化，得到了与位置矩阵相对应的速度矩阵。归一化公式可表示为：

$$spd_i = \frac{v_i}{v_{\max}} \tag{1}$$

其中，spd_i 为车辆 i 的归一化速度，v_i 为车辆 i 的真实速度，$v_{\max}$ 为道路允许的最大速度。

图 3 给出了 SUMO 中的路口模拟车辆与位置矩阵及速度矩阵的对应关系。

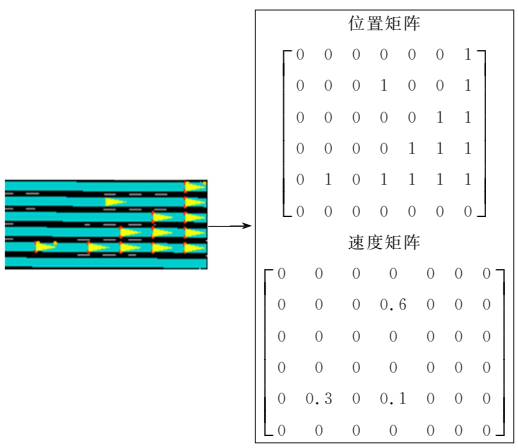

图 3 交通状态的矩阵表示

3.2 动作空间 a_t

智能体获取到当前交叉口的交通状态后，会在动作空间

中选取一个动作执行。在每个周期结束时,选取最优的动作对交叉口的相位时间进行优化。

本文将交通信号控制的动作空间设置为$\{a_1,a_2,a_3,\cdots,a_9\}$。当$i\in[1,4]$时,动作$a_i$表示将相位$i$的时间减少5 s;动作$a_5$表示各相位时间不变;当$i\in[6,9]$时,动作$a_i$表示将相位$i$的时间增加5 s。本文通过这种较小的相位变化间隔,使相邻两个周期中的相位时间得到平滑的变化。智能体在每个信号周期结束后选择动作并执行,保存调整后的相位时间,然后执行下一个信号周期。每个相位时间都有最大绿灯时间$t_{\max}$和最小绿灯时间$t_{\min}$限制,$t_{\min}\leqslant t_i\leqslant t_{\max}$,$t_i$表示相位$i$的绿灯时间。如果某个相位的绿灯时间已经在最大(或最小)的绿灯时间边界上,那么在动作选择之前,会将对应相位调整动作的Q值上加一个比较大的负值,从而使这个动作行为在动作空间中的Q值减小而不会被选中,进而确保这个相位的绿灯时间不会超出绿灯时间边界。

3.3　奖励值 r_t

执行动作之后,从交通环境变化中所得到的奖励是对信号控制决策的评估,并将对下一个信号控制决策产生影响。在基于强化学习的交通信号控制领域中,有很多类型的奖励值被提出,比如车流量变化、路口车辆延误的变化等,选择一个合适的奖励值依赖于交通控制的目标。一个优良的信号控制策略应该把车辆排队长度的绝对差控制在合理的范围之内。在四相位单交叉口中,定义车辆排队长度的绝对差为:

$$r=\min(q_1,q_2,q_3,q_4)-\max(q_1,q_2,q_3,q_4) \tag{2}$$

本文将排队长度的绝对差设置为D3QN算法中的奖励值,r表示车辆排队的绝对差,q_i表示相位i对应的所有车道中最大的排队长度。

3.4　动作选择策略

对D3QN网络进行初始化时,智能体预测的动作是随机的,这种选择策略被称为探索(exploration)。随着卷积神经网络的收敛,可以得到每个动作更准确的Q值,这时应该逐渐减少探索量,优先选择Q值更大的动作,这种选择策略称为利用(exploitation)。在动作的*value*(价值)方差很大时,过多的利用可能会导致最终策略陷入局部最优解。为了避免这样的情况,本文在动作选择策略上采用了ε-greedy策略,即每次迭代选择时产生一个随机数,当随机数小于ε时从动作空间随机选择动作,否则就在神经网络输出的动作值中选择Q值最大的动作。ε会随着迭代次数衰减,如式(3)所示:

$$\varepsilon=\max\left(0.01,1-\frac{n}{N}\right) \tag{3}$$

其中,n是当前迭代数;N是总迭代数;ε的初始值为1,随着迭代次数逐渐降低到最小值0.01。

3.5　D3QN 算法

本文中采用了D3QN算法来解决交通信号控制问题。

在传统的Q学习方法中,为了训练Q值函数,需要每次Q值函数的输出值以及目标值(targetvalue),而目标值会采用下一个状态中对应Q值函数输出的最大动作值来近似,所以Q值函数的更新公式可以表示为:

$$Q(s_t,a_t)\leftarrow Q(s_t,a_t)+\alpha[r_t+\gamma\max\nolimits_{a_{t+1}}Q(s_{t+1},a_{t+1})-Q(s_t,a_t)] \tag{4}$$

但是,这样的更新方式容易导致智能体对动作值的过高期望,产生过估计问题,大大降低了算法的学习性能;又因为在实际情况中过估计量并非是均匀的,所以值函数的过估计可能会影响最终的策略决策,从而导致最终的策略并非最优。因此在D3QN算法中,本文利用了两个DQN网络对信号控制策略进行训练,增加了一个用于计算信号控制策略目标Q值的target网络,通过使target网络进行低频率学习,使它输出的信号控制策略的目标Q值波动较小,从而避免信号控制策略训练过程的网络振荡。然后,在每一个训练步中,先通过主DQN网络选择最大Q值的信号控制动作,再获取这个信号控制动作在target DQN网络中的Q值,从而使每次选择的信号控制动作的Q值并非总是最大的Q值,这样就避免了被高估的次优信号控制动作价值总是超过最优的信号控制动作。D3QN中的神经网络(即Q值函数)对参数θ的更新是通过梯度下降法计算损失函数的梯度来完成的,损失函数$L(\theta)$采用均方根误差公式:

$$L(\theta)=E[(r_t+\gamma Q(s_{t+1},\arg\max_a Q(s_{t+1},a;\theta_t);\theta_t^-)-Q(s_t,a_t;\theta_t))^2] \tag{5}$$

当使用神经网络来近似表示Q值函数时,强化学习往往会因为采样数据间的相关性而产生不稳定的学习效果。文献[26]中提出了经验回放机制(experience replay),通过在经验池中随机采样打破样本间的相关性并提高数据利用率,但是这种方式使每个样本都具有相同的采样概率,无法区分样本的重要性。在D3QN算法中,本文使用了基于经验优先级的采样方式,将每个样本的时间差分误差项(TD-error)作为评价优先级的标准。*TD-error*越大,表示这个样本越需要被学习,可以对这样的样本设置更高的优先级。

本文通过DTSE方法对交通状态进行定义,为了尽可能地利用矩阵中的交通状态信息,本文在D3QN算法中采用了一种典型的卷积神经网络作为Q函数的逼近器,同时将状态-行为对的Q值拆分为两部分进行输出。其中,一部分代表交通环境状态本身具有的价值$V(s_t)$,称为Value;另一部分则表示动态地通过选择某个信号控制动作额外带来的价值$A(a_t)$,称为Advantage,如果一个动作的Advantage值是正数,则意味着该动作所带来的价值比其他动作的平均价值更高。本文的Q值可以写成:$Q(s_t,a_t)=V(s_t)+A(a_t)$。通过神经网络分别计算交通环境本身的Value和选择信号控制动作带来的Advantage,从而使D3QN算法对得到的信号控制动作策略的估计更准确。如图4所示,本文将卷积神经网络的结构设计为三层卷积层加一层全连接层,神经网络的输入为大小为$L*W*2$的车辆位置矩阵,L和W分别表示状态矩阵的长度和宽度,2表示车辆位置信息和车速这两个特征值。输出层为交通环境本身的Value值和选择信号控制动作带来的Advantage值,最终通过这两个值得到在当前状态下动作空间中9个动作的Q值。

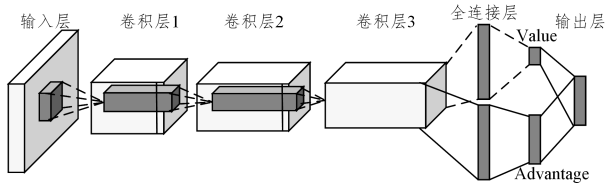

图4　D3QN算法中卷积神经网络的结构

因此,基于D3QN的交通控制算法的流程可以描述为:首先,初始化算法参数,从仿真软件中获取当前的交通状态,将状态输入到主神经网络中,根据神经网络输出的动作Q值在动作集合中以ε-greedy策略选择动作a;然后,将当前交通

状态和动作以及下一个时间步的交通状态和奖励值作为一个四元组$\langle s,a,r,s'\rangle$储存在经验回放池中。通过经验优先级排序在经验回放池中随机选取小批量的数据,用它们来更新主神经网络的参数。同时对目标网络进行低频率的更新,以增加强化学习的稳定性。通过这种方式,使主神经网络的输出值近似于目标 Q 值,然后可以通过选择具有最大 Q 值的动作来获得最优信号控制策略。

基于 D3QN 的交通信号控制算法的伪代码如算法 1 所示。

算法 1 基于 D3QN 的交通信号控制算法

输入:经验回放池 M,批度大小 B,预训练步数 P,参数 ε,折扣系数 γ,学习率 α,训练总步数 T

初始化神经网络 mainQN,targetQN;交叉口环境参数和动作集合 A;训练步数 i=0

从初始化交叉口环境中得到交通状态 s

while i≤T do

根据 ε-greedy 策略选择动作 a

执行动作 a,得到奖励值 r 和新的交通状态 s′

将四元组⟨s,a,r,s′⟩添加至经验回放池 M

定义当前交通状态 s=s′

i=i+1

if |M|>B and i>P then

从经验回放池 M 中依据优先级获取数量为 B 的经验

计算损失函数 $L(\theta)$

根据损失函数,通过梯度下降法更新 mainQN 的网络参数

以较低的更新速率更新 targetQN 的网络参数

更新经验回放池 M 的优先级

更新参数 ε 的值

end if

4 仿真实验设计及实验结果分析

4.1 仿真实验设计

仿真实验采用 SUMO 交通仿真平台。SUMO(Simulation of Urban MObility)是一个开源、微观、时间离散、道路空间连续的交通仿真软件,它的主要功能有路网构建、需求建模、交通仿真和交通管理等。本文主要采用交通仿真中的 Traci(Traffic Control Interface)模块,通过 Python 语言实现与 SUMO 的在线交互,获取实时的交通状态,并执行调节信号的控制动作。

本文采用了杭州市的市中心路和山阴路交叉口作为单交叉口的交通仿真场景,如图 3 所示,车辆驶入点与交叉口停车线的距离为 600 m,车辆长度为 4 m,车辆间的最小距离为2 m,最大速度为 13.89 m/s。为了尽可能实现对真实交通情况的仿真,实验中的路口流量也采用了杭州市山阴路和市中心路交叉口在 2018 年 7 月 22 号流量高峰期的数据,数据采样间隔为 5 min。为了避免各个方向的交通拥堵并确保交通安全,设计了 4 个绿灯信号相位,初始相位顺序为[相位 1:南北直行绿灯,相位 2:南北左转绿灯,相位 3:东西直行绿灯,相位 4:东西左转绿灯],在每个绿灯相位结束后还会有一个 3 s 的黄灯(包含全红)过渡相位,用于清空还未驶出交叉口的车辆。本文将流量数据按 SUMO 仿真软件的配置文件形式进行转化,同时在路网配置中对每个车道都设置了传感器来获取车辆信息。

对于 D3QN 算法的初始参数,本文的设置如表 1 所列。

表 1 D3QN 算法在单交叉口的实验参数设置

参数	值
经验回放单元大小	2500
学习率(α)	0.0001
批度大小($batch\_size$)	64
折扣系数(γ)	0.99
目标网络更新速率(tua)	0.001
状态矩阵大小	60 * 60
动作空间大小	9
迭代回合数($episode$)	200
仿真周期数	10000

4.2 仿真实验结果分析

本文在交通环境参数与神经网络参数初始化相同的前提下,同步进行了 10 次仿真实验,在 10 次仿真实验中记录每个迭代回合的奖励值数据以及每个迭代回合中的平均排队长度数据,这里每个迭代回合的平均排队长度数据是通过记录该迭代回合中每个信号周期结束时交叉口中的排队车辆总数,然后对所有周期的排队车辆总数取平均值获得的。对这些数据进行处理之后,可以得到基于 D3QN 的信号控制方法的仿真结果。

每次仿真实验的训练回合数均为 200,对应每个回合,将 10 次实验得到的数据分别用不同颜色的散点进行绘制,然后分别对每个训练回合中 10 次实验的数据求平均,得到了图 5 和图 6 散点图以及平均值曲线(黑色实线)。同时对每个回合的 10 次实验数据求标准差,根据标准差得到了数据的离散区间(灰色区域)。从图 5 和图 6 中可以看到,由于智能体的探索策略,它在训练初期选择的随机动作可能是不合理的,因此交叉口的平均排队长度较大,平均奖励值较低。随着训练回合的增加,智能体逐渐增加利用(exploitation)策略,神经网络也开始收敛,交叉口的平均排队长度逐渐减小,平均奖励值也呈现上升趋势。最后,平均排队长度比训练初始阶段降低了 32%,平均奖励值比初始阶段提升了 75%,而且从平均奖励值的标准差数据中可以看出,随着训练回合数的增加,平均奖励值的离散程度逐渐降低,说明算法在训练后期有较好的收敛性和稳定性。

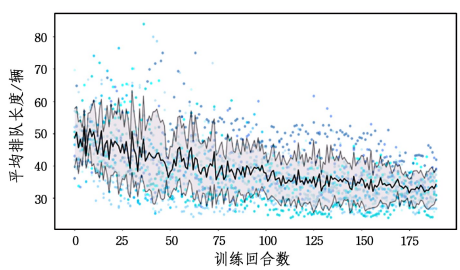

图 5 D3QN 算法下的平均排队长度与训练回合

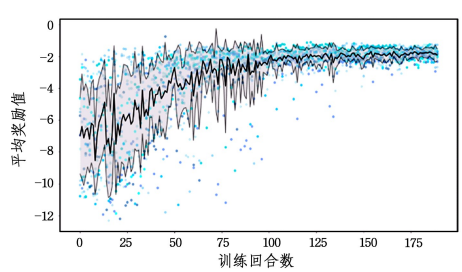

图 6 D3QN 算法下的平均奖励值与训练回合

作为对比,本文加入了定时控制和感应控制的仿真实验,在相同的交通环境参数和仿真时间下计算交叉口的平均排队

长度，得到了如表 2 所列的对比结果。D3QN 算法得到的平均排队长度比定时控制方法得到的平均排队长度减小了 21.1%，对比感应控制降低了 13.1%。这表示 D3QN 算法在交通信号控制上得到的效果明显优于定时控制以及感应控制方法。

表 2　不同控制方式下的交叉口排队长度

控制方式	交叉口平均排队长度/辆
定时控制	40.34
感应控制	36.63
D3QN	31.83

结束语　本文提出了一种基于 D3QN 算法的信号控制方法，并采用实际交叉口模型和流量数据进行仿真实验。实验结果表明，与定时控制和感应控制相比，D3QN 算法可以得到更好的效果。我们可以从以下几个方面考虑未来的研究工作：(1)改进信号控制动作的设计，本文的研究只考虑了信号相位的时间调整，我们认为如果智能体可以根据当前交通状态设计合理的信号相位以及时间，则可以使信号控制更加灵活，得到更好的控制效果；(2)本文的研究只是关于单交叉口的信号控制策略，我们需要关注两个交叉口甚至多个交叉口的信号控制方法，这就需要考虑多个智能体之间的相互合作和沟通。

随着云计算及人工智能技术的不断发展，强化学习朝着更加高效、收敛更快、擅长处理复杂情形以及多智能体的方向发展。基于强化学习的交通信号控制方法也会不断走向成熟，最终落地并解决社会实际问题。

参 考 文 献

[1] LI L, WEN D, YAO D Y. A survey of traffic control with vehicular communications[J]. IEEE Transactions on Intelligent Transportation Systems, 2014, 15(1): 425-432.

[2] FADLULLAH Z, TANG F, MAO B, et al. State-of-the-Art Deep Learning: Evolving Machine Intelligence Toward Tomorrow's Intelligent Network Traffic Control Systems[J]. IEEE Communications Surveys & Tutorials, 2017, PP(99): 1-1.

[3] NGUYEN, THUY T T, ARMITAGE G J. A survey of techniques for internet traffic classification using machine learning [J]. IEEE Communications Surveys & Tutorials, 2008, 10(3): 56-76.

[4] CHIN Y K, KOW W Y, KHONG W L, et al. Q-Learning Traffic Signal Optimization within Multiple Intersections Traffic Network[C]// 2012 Sixth UKSim/AMSS European Symposium on Computer Modeling and Simulation (EMS). IEEE, 2012.

[5] CHIN Y K, LEE L K, BOLONG N, et al. Exploring Q-Learning Optimization in Traffic Signal Timing Plan Management[C]// Third International Conference on Computational Intelligence. IEEE, 2011.

[6] ARAGHI S, KHOSRAVI A, CREIGHTON D C, et al. Optimal fuzzy traffic signal controller for an isolated intersection[C]// IEEE International Conference on Systems. IEEE, 2014.

[7] CHEN Y H, CHANG C J, HUANG C Y. Fuzzy Q-Learning Admission Control for WCDMA/WLAN Heterogeneous Networks with Multimedia Traffic[J]. IEEE Transactions on Mobile Computing, 2009, 8(11): 1469-1479.

[8] CHIU S, CHAND S. Adaptive traffic signal control using fuzzy logic[C]// IEEE International Conference on Fuzzy Systems. IEEE, 1992.

[9] B BINGHAM E. Reinforcement learning in neurofuzzy traffic signal control[J]. European Journal of Operational Research, 2001, 131(2): 232-241.

[10] LA P, BHATNAGAR S. Reinforcement Learning With Function Approximation for Traffic Signal Control[J]. IEEE Transactions on Intelligent Transportation Systems, 2011, 12(2): 412-421.

[11] EL-TANTAWY S, ABDULHAI B, ABDELGAWAD H. Multiagent reinforcement learning for integrated network of adaptive trafFIc signal controllers (MARLIN-ATSC): methodology and large-scale application on downtown Toronto[J]. IEEE Transactions on Intelligent Transportation Systems, 2013, 14(3): 1140-1150.

[12] OZAN C, BASKAN O, HALDENBILEN S, et al. A modified reinforcement learning algorithm for solving coordinated signalized networks[J]. Transportation Research Part C: Emerging Technologies, 2015, 54: 40-55.

[13] ELTANTAWY S, ABDULHAI B, ABDELGAWAD H. Design of Reinforcement Learning Parameters for Seamless Application of Adaptive Traffic Signal Control[J]. Journal of Intelligent Transportation Systems, 2014, 18(3): 227-245.

[14] ABDOOS M, MOZAYANI N, BAZZAN A L C. Holonic multiagent system for traffic signals control[J]. Engineering Applications of Artificial Intelligence, 2013, 26(5/6): 1575-1587.

[15] ABDULHAI B, PRINGLE R, KARAKOULAS G J. Reinforcement learning for true adaptive traffic signal control[J]. Journal of Transportation Engineering, 2003, 129(3): 278-285.

[16] 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.

[17] BALAJI P G, GERMAN X, SRINIVASAN D. Urban traffic signal control using reinforcement learning agents[J]. IET Intelligent Transport Systems, 2010, 4(3).

[18] GENDERS W, RAZAVI S. Using a Deep Reinforcement Learning Agent for Traffic Signal Control[J]. arXiv: 1611.01142v1, 2016.

[19] SUTTON R, BARTO A. Reinforcement Learning: An Introduction[M]. Cambridge, MA: MIT Press, 1998.

[20] HINTON G E, SALAKHUTDINOV R R. Reducing the dimensionality of data with neural networks[J]. Science, 2006, 313(5786): 504-507.

[21] BENGIO Y. Learning deep architectures for AI[J]. Foundations and Trendsin Machine Learning, 2009, 2(1): 1-127.

[22] LANGE S, RIEDMILLER M. Deep auto-encoder neural networks in reinforcement learning[C]// Proceedings of the 2010 International Joint Conference on Neural Networks (IJCNN). Barcelona: IEEE, 2010: 1-8.

[23] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Human-level control through deep reinforcement learning[J]. Nature, 2015, 518(7540): 529-533.

[24] LI L, LV Y, WANG F Y. Traffic Signal Timing via Deep Reinforcement Learning[J]. IEEE/CAA Journal of Automatica Sinica, 2016, 3(3): 247-254.

[25] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Playing Atari with Deep Reinforcement Learning[J]. arXiv: 1312.5602, 2013.