计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.190600154

基于深度强化学习的交通信号控制方法

孙 浩 陈春林 刘 琼 赵佳宝
南京大学控制与系统工程系 南京 210093
(haosun@smail.nju.edu.cn)

摘 要 交通信号的智能控制是智能交通研究中的热点问题。为更加及时有效地自适应协调交通,文中提出了一种基于分布式深度强化学习的交通信号控制模型,采用深度神经网络框架,利用目标网络、双Q网络、价值分布提升模型表现。将交叉路口的高维实时交通信息离散化建模并与相应车道上的等待时间、队列长度、延迟时间、相位信息等整合作为状态输入,在对相位序列及动作、奖励做出恰当定义的基础上,在线学习交通信号的控制策略,实现交通信号Agent的自适应控制。为验证所提算法,在SUMO(Simulation of Urban Mobility)中相同设置下,将其与3种典型的深度强化学习算法进行对比。实验结果表明,基于分布式的深度强化学习算法在交通信号Agent的控制中具有更好的效率和鲁棒性,且在交叉路口车辆的平均延迟、行驶时间、队列长度、等待时间等方面具有更好的性能表现。

关键词:智能交通;交通信号控制;深度强化学习;分布式强化学习
中图法分类号 TP181

Traffic Signal Control Method Based on Deep Reinforcement Learning

SUN Hao,CHEN Chun-lin,LIU Qiong and ZHAO Jia-bao
Department of Control and Systems Engineering,Nanjing University,Nanjing 210093,China

Abstract The control of traffic signals is always a hotspot in intelligent transportation systems research. In order to adapt and coordinate traffic more timely and effectively,a novel traffic signal control algorithm based on distributional deep reinforcement learning was proposed. The model utilizes a deep neural network framework composed of target network,double Q network and value distribution to improve the performance. After integrating the discretization of the high-dimensional real-time traffic information at intersections with waiting time,queue length,delay time and phase information as states and making appropriate definitions of actions,rewards in the algorithm,it can learn the control strategy of traffic signals online and realize the adaptive control of traffic signals. It was compared with three typical deep reinforcement learning algorithms,and the experiments were performed in SUMO(Simulation of Urban Mobility) with the same setting. The results show that the distributional deep reinforcement learning algorithm is more efficient and robust,and has better performance on average delay,travel time,queue length,and waiting time of vehicles.

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

1 引言

为满足日益增长的交通需求,不仅需要通过扩建交通基础设施来扩充容量,更重要的是要优化存量或新建交通设施中的交通控制与管理来提升交通能力,其中交通信号的优化控制极为重要。现实中的交通信号控制多为传统的固定配时策略,周期性地循环既定设置。这种低效的交通灯循环控制存在着诸多问题,如长时间的交通延迟、不合理的时间设置、无法根据实时交通信息做出灵活的调整等。随着交通物联网技术和人工智能的迅速发展,交通信号灯的智能控制成为了智能交通中的热点问题。

相对于传统固定配时的交通信号控制方法,智能交通信号控制能够更加及时有效地自适应协调交通。强化学习[1](Reinforcement Learning,RL)通过与环境的交互直接学习得到最优决策,将其应用于交通信号控制,可以根据交通状况自适应地学习最优的信号策略,改善交通状况。早期的交通信号强化学习控制方法依赖于简化状态假设和手工特征提取,容易丢失潜在的重要状态信息,且在交通信号控制中存在环

到稿日期:2019-03-25 返修日期:2019-06-26 本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:国家自然科学基金(71732003);国家重点研发项目(2016YFD0702100)
This work was supported by the National Natural Science Foundation of China (71732003) and National Key Research and Development Program of China (2016YFD0702100).
通信作者:赵佳宝(jbzhao@nju.edu.cn)

境感知不准确、信号配时策略制定困难等问题。为解决此问题，本文提出了一种基于分布式深度强化学习[2]的交通信号控制方法。该方法结合深度学习在高维数据中的强大表征能力与强化学习在复杂状态任务中的决策能力，在对状态、动作、奖励做出恰当定义的基础上，将实时的交通信息进行离散化状态编码，并将其与相应车道上的车辆信息和交通相位等信息整合后作为深度神经网络的输入，从中抽取交通特征，通过经验缓存中的经验对神经网络进行训练，拟合目标价值分布，使交通信号 Agent 能够学到有效的控制策略。

相对于目前的交通信号控制算法，本文提出的算法具有以下优点：1)将高维实时交通信息作为输入进行端到端的学习，能防止丢失交通流中潜在的重要信息，得到的控制策略更符合现实；2)将分布式深度强化学习引入交通控制，该算法相较以往的典型深度强化学习算法能够提高交通信号控制的效果和鲁棒性。理论分析和实验验证表明了基于分布式的深度强化学习算法在交通信号控制中的有效性。

2 相关工作

早期对交通信号控制的研究主要集中在模糊逻辑[3]及线性规划[4]等方面。Lin 等[5]提出将交通信号控制问题当作混合整数的线性规划问题来讨论。Prashanth 等[6]提出通过控制程序从传感器(如地下感应线圈)获取输入来检测交通灯前方车辆的数量，然后对输入信号进行相应处理以确定交通灯的持续时间。当路面情况较为复杂时，这种方式的精确度将大幅度下降。遗传算法为这类问题提供了一种新思路[7-8]，研究者将交通信号灯控制问题表示为一个优化问题，并通过遗传算法来求解。Yu 等[9]提出了一种用于交通信号自适应控制的马尔可夫决策过程框架，但需要事先确定交通状态间的转移概率，其实际应用受到限制。Gokulan 等[10]提出了一种基于分布式多智能体的交通信号控制方法。

近年来，强化学习技术在人工智能、机器学习和自动控制等领域中得到了广泛的研究和应用[11]。随着强化学习算法和理论研究的深入，应用强化学习方法优化控制器成为交通控制领域研究和应用的热点之一。Prashanth 等[12]提出一种将当前队列长度和当前信号灯的持续时间作为状态的强化学习信号控制系统。Liu 等[13]提出用线性函数对车辆进行聚类拟合 Q 值，但是该方法在状态表示中只使用了队列信息，交通系统的其他复杂性未能通过这些有限的信息表现出来。在有限状态下，当大量有用的特征被省略时，强化学习就无法在交通信号控制中发挥最佳的作用[14]。El-Tantawy 等[15]提出一种多智能体强化学习交通信号控制的方法，但每个智能体都必须保留一组数值表，表的大小与车辆的数量呈指数关系。越来越多的研究将 Q-learning 等无模型时间差分强化学习方法应用于一系列交通优化问题。Wiering 等[16-17]的两项较早的研究为使用仿真软件模拟交通系统和比较不同强化学习方法的性能提供了框架。Marsetic 等[18]的仿真研究表明，Q-learning 的性能优于静态逻辑控制系统，但他的实验没有使用现实的、时间动态的交通条件，而是在相对静态的环境中训练，存在一定的局限性。

3 分布式深度强化学习

本节主要阐述强化学习及几种典型的深度强化学习算法，在此基础上引入基于分布式的深度强化学习算法，并将其应用于交通信号控制。

3.1 强化学习

在标准的强化学习场景中，Agent 通过与环境的交互最大化长期奖励以获得最优策略。马尔可夫决策过程[19](Markov Decision Process，MDP)为强化学习提供了明确的理论框架。强化学习过程由四元组$\langle S,A,P,R\rangle$组成，其中 S 表示状态空间，A 表示动作空间，$R:S\times A\rightarrow R$ 是奖励函数，$P:S\times A\rightarrow[0,1]$是状态转移概率分布。Agent 的目标是通过学习一个最优策略 $\pi:S\rightarrow A$ 来最大化长期奖励 $R_0^T=\sum_{i=0}^{T}\gamma^i r(s_i,a_i)$。其中 T 是终止时间，$r(s_i,a_i)$表示在状态 s_i 时采取 a_i 获得的奖励，$\gamma\in(0,1)$表示折扣因子，用来平衡即时奖励和长期奖励的权重。我们将状态动作值函数定义为 $Q(s_t,a_t)$在状态 s_t 下采取动作 a_t 取得的期望回报：

$$Q(s_t,a_t)=E[R_t^T\mid s=s_t,a=a_t]=E[\sum_{i=t}^{T}\gamma^{i-t}r(s_i,a_i)] \tag{1}$$

则最优状态值函数 $Q^*(s_t,a_t)$可以定义为在状态 s_t 下采取动作 a_t 取得的最优回报值。根据贝尔曼方程(Bellman Equation)，可得：

$$Q^*(s_t,a_t)=E[r(s_t,a_t)+\gamma\max_{a'_{t+1}}Q^*(s_{t+1},a_{t+1})] \tag{2}$$

其更新公式为：

$$Q(s_t,a_t)=Q(s_t,a_t)+\alpha[r(s_t,a_t)+\gamma\max_{a_{t+1}}Q(s_{t+1},a_{t+1})-Q(s_t,a_t)] \tag{3}$$

通过不断迭代状态动作值函数得到最优策略：

$$\pi^*=\arg\max_{a\in A}Q^*(s,a) \tag{4}$$

3.2 深度强化学习算法

为解决传统强化学习在大规模离散空间或连续状态空间中难以应用的问题，DeepMind 提出深度 Q 网络模型[20](Deep Q Network，DQN)，其能够端到端地学习策略，而无需额外的先验知识。为减轻非线性网络表示值函数的不稳定性，DQN 引入经验回放和目标网络。经验回放机制将 Agent 与环境交互获得的转移序列$\langle s_t,a_t,r_t,s_{t+1}\rangle$存储在经验缓冲区中，从经验缓冲区均匀采样小批量转移序列，使用随机梯度下降(Stochastic Gradient Descent，SGD)训练深度神经网络使其逼近 Q 值函数，打破了经验之间的强相关性，使得学习更加高效、稳定。DQN 使用卷积神经网络来拟合当前状态的动作值函数 $Q(s,a;\theta)$与目标状态的动作值函数 $Q(s,a;\theta^-)$，θ 和 θ^- 分别表示当前值网络与目标值网络的参数。当前值函数的近似优化目标表示为：

$$y(s,a)=r+\gamma\max_{a'}Q(s',a';\theta^-) \tag{5}$$

通过最小化目标 Q 值与当前 Q 值之间的时间差分(Temporal Difference，TD)来得到最优值，即：

$$\delta=r+\gamma\max_{a'}Q(s',a';\theta^-)-Q(s,a;\theta) \tag{6}$$

神经网络的损失函数表示为：

$$Loss(\theta)=\frac{1}{2}\sum_{i=1}^{N}\delta^2 \tag{7}$$

为缓解 DQN 模型的过拟合问题，Van Hasselt 等[21]提出了双重深度 Q 网络模型（Double Deep Q Network，DDQN），DDQN 模型使用两组不同的网络参数分别进行动作值的计算与动作的选择。基于 DDQN，Schaul 等[22]提出了一种基于优先级重放采样的深度强化学习算法（Prioritized Experience Replay，PER），该方法用基于优先级的采样代替均匀采样，加大价值更大的样本的采样概率以加快最优策略的学习。相较于 DQN 算法，PER 算法的具体改进如下：1）优先级采样以时间差分误差 *TD-error* 的绝对值作为评价优先级的准则，即绝对值越大，对应样本被采样的概率就越大，优先级公式如式（8）所示；2）采用随机比例化（Stochastic Prioritization）和重要性采样权重（Importance-Sampling Weights）两种技巧，随机比例化技巧不仅能充分利用 *TD-error* 较大的样本，而且能保证样本的多样性，而利用重要性采样权重减小了参数更新的速度，保证了学习的稳定性。

$$P(i)=\frac{p_i^{\alpha}}{\sum_k p_k^{\alpha}} \tag{8}$$

其中，α 决定采用优先级的程度。

3.3　分布式强化学习算法

在传统的基于价值函数的强化学习模型中，价值函数输出的是每个动作在给定状态下的期望回报，分布式强化学习（Distributional RL）则假设这个价值是一个随机变量。传统价值函数的输出目标是近似地估计价值的期望，而 Distributional RL 的目标是近似地估计价值的分布，通过学习累计回报的近似分布来代替累计回报的期望。Bellemare 等[2]提出基于离散支撑的概率质量的分布，其中 $\boldsymbol{z}$ 是具有 $N_{\text{atoms}}\in\mathbb{N}^+$ 原子的向量，$z^i=v_{\min}+(i-1)\frac{v_{\max}-v_{\min}}{N_{\text{atoms}}-1}$，$i\in\{1,\cdots,N_{\text{atoms}}\}$。使用 d_t 表示 t 时刻该支撑下的近似分布，每个原子 i 所对应的概率质量为 $p_\theta^i(S_t,A_t)$，即 $d_t=(\boldsymbol{z},p_\theta(S_t,A_t))$。学习的目标是通过更新参数 θ，使得累计回报的近似分布趋近于真实分布。

在分布式强化学习中，回报分布满足 Bellman 等式的变种形式。对于给定的状态 S_t 和动作 A_t，最优策略 π^* 对应的回报分布应拟合目标回报分布，目标回报分布通过下一个状态 S_{t+1} 和动作 $a_{t+1}^*=\pi^*(S_{t+1})$ 对应的分布得到。Distributional RL 首先建立目标分布 d_t' 的支撑向量，然后最小化分布 d_t 和目标分布 d_t' 之间的 KL 散度。目标分布 d_t' 满足：

$$d_t'\equiv(R_{t+1}+\gamma_{t+1}\boldsymbol{z},p_{\bar{\theta}}(S_{t+1},\bar{a}_{t+1}^*)) \tag{9}$$

$$D_{KL}(\Phi_z d_t'\,\|\,d_t) \tag{10}$$

其中，Φ_z 是目标分布在固定支撑 $\boldsymbol{z}$ 上的 L2 投影，$\bar{a}_{t+1}^*=\arg\max_a q_{\bar{\theta}}(S_{t+1},a)$ 是在状态 S_{t+1} 时与平均动作值函数对应的贪婪动作：

$$q_{\bar{\theta}}(S_{t+1},a)=\boldsymbol{z}^{\mathrm{T}}p_\theta(S_{t+1},a) \tag{11}$$

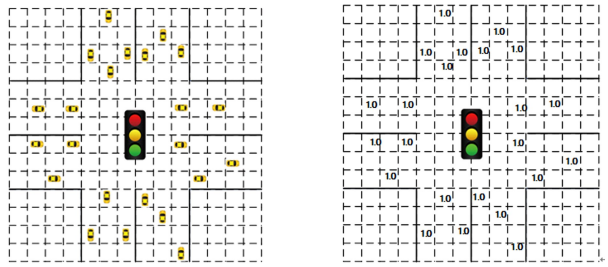
通过深度神经网络来表示参数化分布，并使用参数 θ 和 $\bar{\theta}$ 来构造当前价值分布和目标分布。模型的网络结构与 DQN 相似，主要区别在于其输出由动作值函数变为 $N_{\text{atoms}}\times N_{\text{actions}}$。对输出的每个动作独立地应用 softmax，以确保每个动作的分布被适当地规范化。文中将 Distributional RL 与 DDQN 相结合的算法即 Distributional DDQN 算法应用于交通信号的自适应控制中。

4　分布式深度强化学习的交通控制算法

本节主要阐述将分布式深度强化学习应用于交通控制算法的相关定义及网络结构，分别对算法所需的交叉路口的环境状态、交通信号 Agent 的动作及相位的序列、奖励等信息进行定义，并建立深度神经网络模型。

4.1　状态空间的定义

为了准确描述交叉口的交通信息，对于交叉口的每个车道 i，将车辆的队列长度 L、车辆等待时间 W、交叉口的车辆延迟 D 以及信号灯相位的变化 C 作为状态输入。此外，为准确表示交叉口车辆的位置和速度信息的具体分布，对交叉路口区域进行离散化建模。如图 1 所示，整个交叉口被划分成大小相同的矩形网格，网格的尺寸不仅能够保证车辆完整地处于单一网格，而且避免了两辆车位于同一网格。相比于直接将交叉口的图像信息作为输入，该方式能够大大减小计算量，节约计算资源。在每个网格中，状态值为两个值向量〈位置 P，速度 V〉。其中，位置表示网格中是否有车辆，如有车辆，取值为 1，否则为 0；速度 V 表示对应位置车辆的速度。通过该方式得到交叉口的位置和速度矩阵，同样将其作为模型的输入信息。



（a）车辆位置示意图　（b）位置矩阵示意图

图 1　交叉口离散化建模示意图

Fig. 1　Schematic diagram of discretization modeling at intersection

4.2　动作相位的定义

交通信号需要根据当前的交通状况选择合适的动作来更好地引导交叉口的车辆。在该系统中，信号的相位变化集合如图 2 所示。

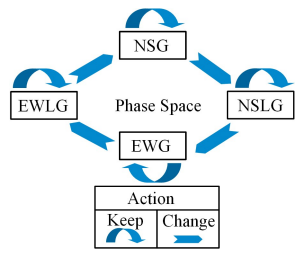


图 2　交通信号的动作及相位序列

Fig. 2　Movement and phase sequence of traffic signals

信号灯 Agent 所有可能的相位被定义为一个信号序列的集合 $A=\{\text{NSG},\text{EWG},\text{NSLG},\text{EWLG}\}$，其相位按照图中的顺序切换。其中，NSG 代表南北方向的道路为绿灯，EWG 代

表东西方向的道路为绿灯，NSLG 代表南北方左转为绿灯，EWLG 代表东西方左转为绿灯。每次执行动作之后，Agent 只能位于其中的一个相位，该相位对应状态的其他方向上的信号灯将默认设置为红色。Agent 观察到某个状态 *state* 时，根据相位集合的定义选择一个合适的动作 a，当 $a=0$ 时，保持当前相位；当 $a=1$ 时，切换到相位序列中的下一相位。每个相位的最短持续时间为 5 s，同时为安全起见，在每次红灯和绿灯切换的间隙会有持续 3 s 的黄灯。在这种情况下，Agent 可以实时地做出更加准确的决策。

4.3 奖励的定义

奖励的定义是深度强化学习最终能否收敛以及能否取得良好效果的关键，恰当的奖励定义有助于交通信号采取最佳的行动策略。综合考虑评价交通状况的指标，本文中的奖励被定义为以下因素的加权平均。

(1)车辆进入各车道的延误时间 d。对于车道上的第 i 辆汽车，d_i 的定义如式(12)所示：

$$d_i=1-\frac{lane\ speed}{speed\ limit} \tag{12}$$

其中，车道速度 *lane speed* 为车道 i 上车辆的平均速度，限速 *speed limit* 为车道上允许的最大速度。

(2)所有进入车道等待的车辆队列长度之和 q。

(3)所有进入车道的车辆的等待时间 w。当汽车的速度小于 0.1 m/s 时，开始统计其等待时间，车辆每次移动速度大于 0.1 m/s 时，等待时间重置为 0。

(4)相位的状态切换 p。当 $p=0$ 时，保持当前相位；当 $p=1$ 时，根据相位序列改变至下一个相位。

(5)车辆的紧急制动停止 e。当车辆进行紧急制动停止时，$e=1$。

(6)执行动作后离开的车辆数 n。

综合考虑以上交通要素和权重，最终的奖励如式(13)所示：

$$Reward=k_1d+k_2q+k_3w+k_4p+k_5e+k_6n \tag{13}$$

4.4 分布式深度强化学习网络模型

深度强化学习的神经网络结构如图 3 所示。先将由原始的交通信息离散化得到的交叉口区域的位置矩阵和速度矩阵作为输入，经过两层卷积神经网络提取特征之后，将特征展开为一维向量，再与队列长度、等待时间、延迟时间、相位变化等信息融合，经过两层全连接层输出与动作对应的价值分布。

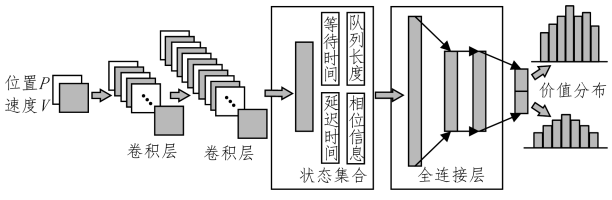


图 3 深度神经网络结构

Fig. 3 Deep neural network structure

5 实验结果与分析

本节首先介绍实验的仿真环境及相关的参数设置，然后在 SUMO 中评估分布式深度强化学习算法 Distributional DDQN 在交通信号控制中的应用效果，并与深度强化学习算法 DQN，DDQN，PER 进行对比分析。

5.1 仿真环境与参数设置

本文以 Intel Core i5-7400 CPU 作为硬件环境，基于微观交通仿真平台 SUMO v0.32 进行仿真实验。利用 SUMO 中提供的 Traci(Traffic Control Interface)接口模块实现与仿真平台的在线交互，实时获取交通状态并自适应调整信号灯的控制策略。算法模型通过深度学习框架 Keras 实现，详细的交通路网仿真设置如下。

交叉口设置：如图 4 所示，整个交叉路口为 400 m×400 m 的区域，由 4 条垂直的道路组成；每条道路为双向三车道，沿着车辆的行驶方向从内至外分别为左转车道、直行车道、右转直行车道；车道限速 70 km/h，车辆的长度为 3 m，最大速度为 60 km/h；为保证行驶的安全性，车辆之间保持最小间距 1 m。交叉路口被抽象为 100×100 的网格输入，用来表示车辆的位置与速度信息分布。

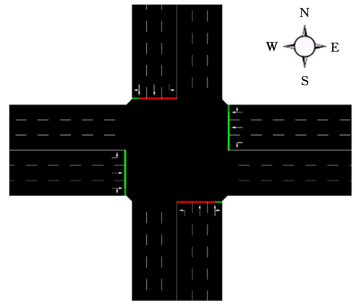


图 4 交叉口仿真区域

Fig. 4 Intersection area for simulation

交通流设置：车辆生成的方法及流量大小对交通仿真的质量会产生重要的影响。为了更符合现实情况，场景中车辆的生成符合随机过程分布，车辆随机进入交叉口的入口并提前选择所在车道。实验持续 14 400 s，在该时段内，通过交通信号的控制引导交叉路口的车辆通行。各条道路和车道的平均车辆到达率设置如表 1 所列。其中，车辆的平均到达率，即在仿真过程中单位时间内进入车道的车辆数量。

表 1 交通流量设置

Table 1 Traffic flow setting

车流方向	检测平均到达率/(car/s)	开始时间/s	结束时间/s
WE	0.1	0	14 400
WEL	0.05	0	14 400
NS	0.05	0	14 400
NSL	0.025	0	14 400

为公平起见，参与比较的算法均采用相同的网络结构及超参数设定。算法更新采用 RMSProp 优化算法，学习率为 0.001，批处理大小为 32，经验缓存尺寸为 1 000，折扣因子为 0.9，训练时采用 ε 贪心算法进行探索和动作选择，ε 的初始值为 0.1，随着训练的迭代，其按照 0.999 的指数衰减。训练开始之前，通过网络结构的预训练对模型参数进行初始化，以提高算法的效果和稳定性。此外，PER 算法中的随机比例化因子 α 为 0.6。Distributional DDQN 算法中 $V_{\min}=-50$，$V_{\max}=0$，$Atoms=51$。

在本实验中，综合考虑各交通因素对交通状况的影响程

度，将奖励中各个因素的系数设置为：$k_1=-0.25$，$k_2=-0.25$，$k_3=-0.25$，$k_4=-1$，$k_5=0.15$，$k_6=1$。

5.2　实验评估与结果分析

为了验证基于分布式深度强化学习 Distributional DDQN 的交通信号控制算法的有效性，将其与 DQN，DDQN 及 PER 算法进行对比，从平均累计奖励、汽车平均延迟（delay）、交叉口的平均行驶时间（duration）、平均队列长度（queue length）、平均等待时间（waiting time）5 个方面对算法进行对比分析。在设定的时间内，平均累计奖励值越大，表明算法的表现越好；其余 4 种交通衡量指标的平均值越小，表示在交叉口的车辆通行情况越好，拥堵程度越小。

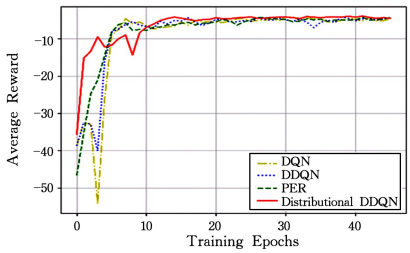


图 5　各算法的平均累积回报对比

Fig. 5　Average cumulative reward comparison of algorithms

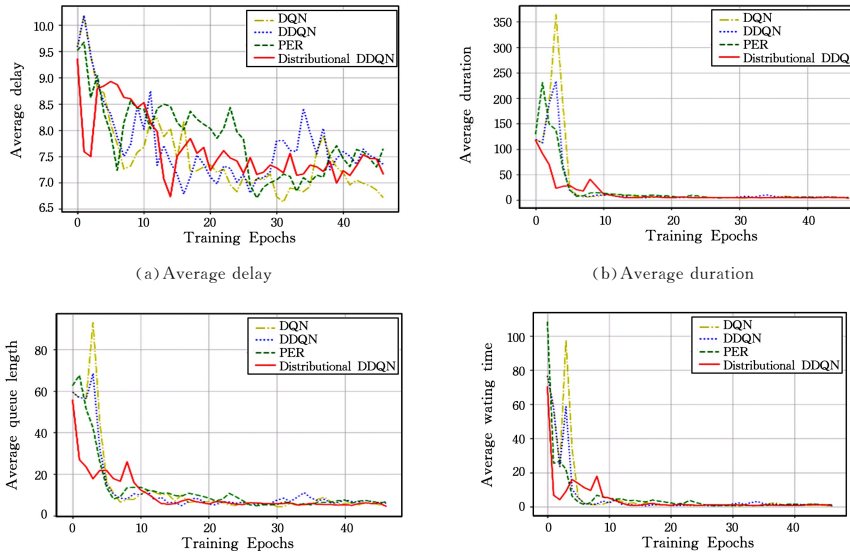
从图 5 可以看出，与 3 种典型的深度强化学习算法相比，Distributional DDQN 在交叉口的信号控制中具有更快的速率收敛；且从表 2 可以看出，其累计平均奖励相对于其余 3 种算法中的最优值提高了约 18.3%，具有更好的效果。

表 2　算法的性能对比

Table 2　Performance comparison of algorithms

算法	平均延迟/s	平均行驶时间/s	平均队列长度/m	平均等待时间/s	累计奖励
DQN	7.51	26.70	13.02	7.52	−8.85
DDQN	7.71	21.40	12.56	6.37	−8.30
PER	7.83	21.92	12.72	6.12	−8.03
Distributional DDQN	7.65	14.10	10.12	4.65	−6.56

从图 6 中可以看出，在训练的初始阶段，由于算法尚未收敛，交通信号 Agent 尚未学到正确的调度策略，交叉口的车辆延迟、行驶时间等均有小幅度升高。随着训练的进行，交通路口的通行状况均逐渐好转，并最终收敛至最优通行策略。各算法性能如表 2 所列，其中累计奖励值越大，交通相关的 4 个衡量指标越小，表示实验结果越好。将 Distributional DDQN 算法的结果与 3 种典型算法中的最好结果进行对比，可以看出在整个实验过程中，4 种算法的车辆平均延迟结果差距较小，所提算法的车辆平均行驶时间缩短了 34.1%，平均队列长度缩短了 19.4%，平均等待时间缩短了 24.0%。实验结果表明，与 DQN，DDQN 和 PER 算法相比，Distributional DDQN 算法具有更好的收敛效果。



(a) Average delay　(b) Average duration

(c) Average queue length　(d) Average waiting time

图 6　交通状况的对比

Fig. 6　Comparison of traffic situation

综上所述，将分布式的深度强化学习应用于交通信号的智能控制，能够得到良好的交通信号控制策略，有效地减少车辆停留时间、车辆延误并缓解交通拥堵。

结束语　本文提出了一种基于分布式深度强化学习的交通信号控制方法。相比传统的强化学习，所提算法具有更好的普遍适用性，能够进行更高维度的交通特征的提取；与 3 种典型的深度强化学习算法相比，其在交通信号的控制中也能得到更好的效果。但是，此方法仍然存在一定的局限性，如随着控制范围的扩大，交叉口数量增加，会导致动作空间陡增，同时需要考虑多路口之间的协调问题。在未来的工作中，将进一步研究将更先进的强化学习算法（如 DDPG，A3C 等）应用于交通领域，并拓展至多路口的交通信号控制中。

参 考 文 献

[1] SUTTON R S，BARTO A G. Introduction to reinforcement learning[M]. Cambridge：MIT Press，1998.

[2] BELLEMARE M G，DABNEY W，MUNOS R. A distributionalperspective on reinforcement learning[C]// Proceedings of the

34th International Conference on Machine Learning. JMLR. org, 2017: 449-458.
[3] CHIS S. Adaptive traffic signal control using fuzzy logic[C]// Proceedings of the Intelligent Vehicles92 Symposium. IEEE, 1992: 98-107.
[4] PANDIT K, GHOSAL D, ZHANG H M, et al. Adaptive traffic signal control with vehicular ad hoc networks[J]. IEEE Transactions on Vehicular Technology, 2013, 62(4): 1459-1471.
[5] LIN W H, WANG C. An enhanced 0-1 mixed-integer LP formulation for traffic signal control[J]. IEEE Transactions on Intelligent transportation systems, 2004, 5(4): 238-245.
[6] PRASHANTH L A, BHATNAGAR S. Reinforcement learning with function approximation for traffic signal control[J]. IEEE Transactions on Intelligent Transportation Systems, 2010, 12(2): 412-421.
[7] GIRIANNA M, BENEKOHAL R F. Using genetic algorithms to design signal coordination for oversaturated networks[J]. Journal of Intelligent Transportation Systems, 2004, 8(2): 117-129.
[8] SANCHEZ-MEDINA J J, GALAN-MORENO M J, RUBIO-ROYO E. Traffic signal optimization in "La Almozara" district in Saragossa under congestion conditions, using genetic algorithms, traffic microsimulation, and cluster computing[J]. IEEE Transactions on Intelligent Transportation Systems, 2009, 11(1): 132-141.
[9] YU X H, RECKER W. Stochastic adaptive control model for traffic signal systems[J]. Transportation Research Part C: Emerging Technologies, 2006, 14(4): 263-282.
[10] GOKULAN B P, SRINIVASAN D. Distributed geometric fuzzy multi agent urban traffic signal control[J]. IEEE Transactions on Intelligent Transportation Systems, 2010, 11(3): 714-727.
[11] BOWLING M. Multi agent learning in the presence of agents with limitations[R]. Carnegie-Mellon Univ Pittsburgh Pa School of Computer Science, 2003.
[12] PRASHANTH L, BHATNAGAR S. Threshold tuning using stochastic optimization for graded signal control[J]. IEEE Transactions on Vehicular Technology, 2012, 61(9): 3865-3880.
[13] LIU W, QIN G, HE Y, et al. Distributed cooperative reinforcement learning-based traffic signal control that integrates v2x networks' dynamic clustering[J]. IEEE Transactions on Vehicular Technology, 2017, 66(10): 8667-8681.
[14] GENDERS W, RAZAVI S. Using a deep reinforcement learning agent for traffic signal control[J]. arXiv: 1611.01142.
[15] El-TANTAWY S, ABDULHAI B, ABDELGAWAD H. Multi agent 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.
[16] WIERING M A. Multi-agent reinforcement learning for traffic light control[C]// Machine Learning: Proceedings of the Seventeenth International Conference (ICML'2000). 2000: 1151-1158.
[17] WIERING M, VREEKEN J, VAN VEENEN J, et al. Simulation and optimization of traffic in a city[C]// IEEE Intelligent Vehicles Symposium, 2004. IEEE, 2004: 453-458.
[18] MARSETIC R, SEMROV D, ZURA M. Road artery traffic light optimization with use of the reinforcement learning[J]. PROMET-Traffic & Transportation, 2014, 26(2): 101-108.
[19] PUTERMAN M L. Markov Decision Processes: Discrete Stochastic Dynamic Programming[M]. John Wiley & Sons, 2014.
[20] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Human-level control through deep reinforcement learning[J]. Nature, 2015, 518(7540): 529-533.
[21] VAN HASSELT H, GUEZ A, SILVER D. Deep Reinforcement Learning with Double Q-Learning[C]// Association for the Advance of Artificial Intelligence. 2016: 2094-2100.
[22] SCHAUL T, QUAN J, ANTONOGLOU I, et al. Prioritized experience replay[C]// Proceedings of the 4th International Conference on Learning Representations. San Juan, Puerto Rico, 2016: 322-355.

SUN Hao, born in 1996, postgraduate. His main research interests include deep learning and reinforcement learning.

ZHAO Jia-bao, born in 1972, Ph.D, associate professor. His main research interests include coordination and control methods for CAVs and knowledge automation in AIOps (Artificial Intelligence for IT Operations).