文章编号：0253-374X(2024)06-0928-07 **DOI**：10.11908/j.issn.0253-374x.22418

典型匝道控制场景下深度强化学习决策机理解析

刘　冰[1]，唐　钰[2]，暨育雄[1]，沈　煜[1]，杜豫川[1]
（1. 同济大学 道路与交通工程教育部重点实验室，上海 201804；2. 纽约大学 坦登工程学院，纽约 11201）

摘要：以典型匝道控制场景为研究对象，利用状态值函数、显著图及输入扰动，理解深度强化学习模型在交通控制中的决策机理。利用状态值函数评判模型是否能够认识到交通状态的变化，通过显著图分析特定环境状态下模型感知到的环境状态特征和决策动作规律，应用输入扰动分析扰动后匝道控制动作匹配率和控制效果并鉴别关键区域。结果表明，基于深度强化学习的匝道控制模型能够准确评判交通状态的优劣，感知到交通状态的关键特征，并做出合理的决策动作。

关键词：交通工程；深度强化学习；可解释机器学习；匝道控制
中图分类号：U491　　**文献标志码**：A

Understanding Deep Reinforcement Learning Algorithm in Typical Ramp Metering Scenarios

LIU Bing[1], *TANG Yu*[2], *JI Yuxiong*[1], *SHEN Yu*[1], *DU Yuchuan*[1]
(1. Key Laboratory of Road and Traffic Engineering of the Ministry of Education, Tongji University, Shanghai 201804, China; 2. Tandon School of Engineering, New York University, New York 11201, USA)

Abstract: This paper presents the control mechanism of deep reinforcement learning (DRL) in a typical ramp metering scenario. The state value function is used to evaluate if the DRL model has the ability to distinguish the change of state. The saliency map is used to perceive the state key features and control pattern for the DRL model under specific traffic states. By using the input perturbation, the action match ratio and control performance under perturbed data are analyzed to explore the key areas of control. The results show that the DRL model can evaluate the traffic state accurately, distinguish the key features, and then make reasonable decisions.

Keywords: traffic engineering; deep reinforcement learning (DRL); explainable machine learning; ramp metering

近年来，强化学习方法在交通控制领域，如地面交叉口信号控制[1-3]、速度控制[4]以及匝道控制[5-6]中的应用受到广泛关注。研究表明，强化学习方法能够从复杂的非线性交通环境中提取有效信息并提升控制效果，缓解交通拥堵压力，进而提高出行效率。

强化学习通过不断学习环境与控制动作的相互反馈，构建复杂的非线性数学关系。深度神经网络提供更加复杂的网络结构，更有利于提取环境中的复杂信息，从而达到更好的控制效果。然而，深度强化学习模型往往缺乏可解释性，即无法通过数学模型了解模型识别到的信息以及控制原理。模型的可解释性对于强化学习在实际工程中的应用具有重要意义[7-8]。现有研究主要针对强化学习在游戏领域的应用开展可解释性分析。Wang 等[9]采用基于梯度的显著图解释 Dueling DQN(deep Q-learning)结构的合理性。Greydanus 等[10]提出基于扰动的显著图分析方法，用于解释 Atari 游戏中智能体学习内容、智能体最优策略的学习过程以及智能体控制策略失效的原因。Iyer 等[11]提出了一种面向对象的显著图，将目标对象的相应特征融入显著图中，增加显著图的可解释性。然而，少有研究结合专家知识，解释面向实际应用的强化学习模型。

强化学习模型具有特异性，针对游戏领域的可解释性分析成果难以直接迁移至交通控制领域。本文利用强化学习解释工具，聚焦实际交通控制场景，实现对强化学习模型的解析。匝道控制作为交通控制领域的经典场景之一，相关研究非常丰富，可分为基于规则、基于预测模型和基于强化学习模型 3 类。

收稿日期：2022-09-30
基金项目：上海市科委科研计划(19DZ1209100)；浙江省重点研发计划(2021C01011)
第一作者：刘　冰，博士生，主要研究方向为共享交通规划与管理。E-mail：bingliu@tongji.edu.cn
通信作者：暨育雄，教授，博士生导师，工学博士，主要研究方向为交通全息感知与智能计算、智能公交管理及控制。E-mail：yxji@tongji.edu.cn

论文拓展介绍

本文主要讨论基于强化学习模型的匝道控制。Fares等[12]以主线流量、匝道流量以及匝道绿灯时长为系统状态，路段密度与最佳密度的偏差的绝对值倒数为奖励，红灯相位和绿灯相位为控制动作训练匝道控制模型。Yang等[13]利用上匝道控制提高上下匝道紧邻交织区内的交通运行效率，采用交织比作为系统状态，下游流量作为奖励，以当前匝道调节率的修正为控制动作训练匝道控制。随着交通信息采集技术的发展，从激光雷达、视频监控等检测器获得的高维数据为匝道控制提供更全面的交通状态信息。戴昇宏等[14]提出了一种基于图像卷积神经网络(CNN)的匝道控制深度强化学习模型，通过卷积神经网络提取视频图像中的有益信息，以交通流量为奖励，优化匝道控制算法。Liu等[15]以连续多个视频图像为输入，设置奖励的同时考虑合流区和匝道的交通状态，优化匝道控制算法。上述研究结果表明，强化学习模型在匝道控制中的应用能够有效提升控制效果，但并未对提升原理进行解释。

选取典型匝道控制场景，以Liu等[15]提出的强化学习模型为分析对象，结合专家知识和强化学习解析工具，分析强化学习学到的环境特征和决策机理。分析结果不仅为匝道控制模型优化和实际工程应用提供依据，还为强化学习模型在其他交通控制中的实际应用提供方向。

1　基于深度强化学习的匝道控制模型

常用匝道控制可以分为周期式和停走式。周期式匝道控制固定周期时长为C，通过调整绿灯时长实现动态控制；停走式匝道控制固定一个较短的绿灯时长L(2～4 s)，通过调整红灯时长满足管控需求。Liu等[15]在停走式基础上将整个控制时段划分为多个较短的时间间隔，在每一个间隔开始时决定本间隔的控制动作，即红灯或绿灯相位。

以覆盖匝道控制区域的交通视频图像为输入(见图1)，通过目标检测技术提取每一帧交通图像中的车辆位置，构建车辆位置矩阵$\boldsymbol{m}_t$，并将连续多个位置矩阵拼接成三维矩阵，表征环境状态$\boldsymbol{s}_t$。

图2为基于深度强化学习的匝道控制框架。以匝道控制为智能体，基于当前环境状态判断下一时间间隔的信号灯控制动作；交通系统在当前控制动作下发生改变，环境状态由当前状态$\boldsymbol{s}$转变为下一状态$\boldsymbol{s}'$；根据当前状态下采取的控制动作所造成的交通影响，将相应的奖励反馈给匝道控制智能体，以进行下一次的控制动作决策。

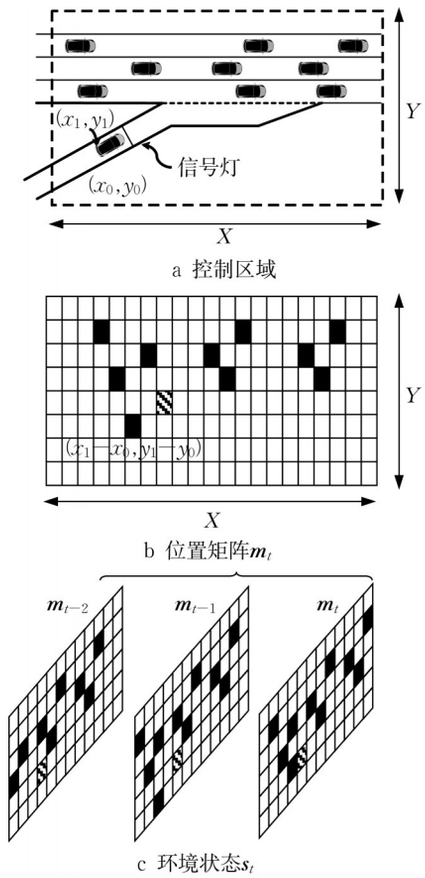

图1　匝道控制状态[15]

Fig.1　Ramp metering state[15]

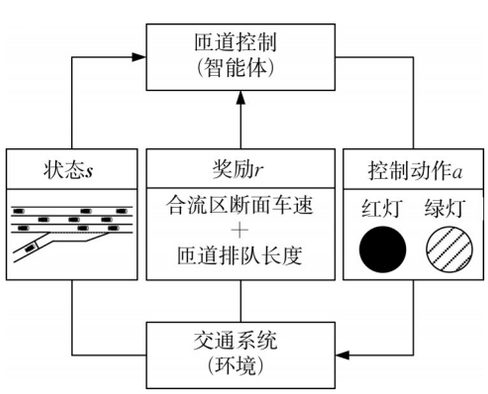

图2　基于深度强化学习的匝道控制框架[15]

Fig.2　Ramp metering framework based on deep reinforcement learning[15]

本文采用基于值函数的强化学习算法，通过训练动作值函数$Q(\boldsymbol{s}, a)$，寻找能够最大化累积奖励的最优控制策略。训练后的最优动作值函数$Q^*(\boldsymbol{s}, a)$可衡量状态$\boldsymbol{s}$下采取动作a的价值，价值最大的动作即为最优控制动作。为了更好地从图像数据中提取状态环境特征，$Q^*(\boldsymbol{s}, a)$由卷积神经网络与全连接神

经网络共同组成。

2 研究方法

采用状态值函数、显著图和输入扰动,对上述基于深度强化学习的匝道控制模型进行解析。状态值函数和显著图用于解释特定环境状态下模型感知到的关键微观交通特征和控制规律,输入扰动用于分析不同区域信息对控制动作和控制效果的宏观影响。

2.1 状态值函数

状态值函数是强化学习对当下环境状态优劣的评价,状态值越高说明强化学习模型认为当下状态越优。因此,通过对比状态值与实际状态在时间上的变化趋势,可评估模型是否准确认识状态的变化。

环境在t时刻处于状态$\boldsymbol{s}$,智能体采用策略π而产生的未来累积奖励G_t的期望为给定动作策略π下的状态值函数$v_\pi(\boldsymbol{s})$,定义为

$$v_\pi(\boldsymbol{s})=\mathrm{E}_\pi(G_t|\boldsymbol{S}_t=\boldsymbol{s}) \tag{1}$$

状态值是模型在采用策略π时对当前状态的评价,反映了模型对当下状态优劣的认识。状态值越低意味着模型认为状况越差。状态值函数可通过$Q(\boldsymbol{s},a)$估计,当环境在t时刻处于状态$\boldsymbol{s}$且智能体选择动作a时,智能体采用策略π而产生的未来累积奖励G_t的期望被称为给定动作策略π下动作值函数$q_\pi(\boldsymbol{s},a)$,其定义为

$$q_\pi(\boldsymbol{s},a)=\mathrm{E}_\pi(G_t|\boldsymbol{S}_t=\boldsymbol{s},A_f=a) \tag{2}$$

最优策略π^*下,有

$$v_{\pi^*}(\boldsymbol{s})=\max_a q_{\pi^*}(\boldsymbol{s},a) \tag{3}$$

式中,$v_{\pi^*}(\boldsymbol{s})$为在采用最优策略$\pi^*$时状态$\boldsymbol{s}$的状态值函数。训练好的动作值函数$Q(\boldsymbol{s},a)$是$q_{\pi^*}(\boldsymbol{s},a)$的一个较优估计。基于式(3)的结论,状态值函数的较优估计$V(\boldsymbol{s})$的计算式为

$$V(\boldsymbol{s})=\max_a Q(\boldsymbol{s},a) \tag{4}$$

2.2 显著图

显著图[16]的概念最早由Itti等[17]提出,用于图像的多尺度特征分析,并被推广至计算机视觉领域。显著图显示的是模型对输入状态中不同区域的关注程度,某一区域在图中越显著,说明模型认为这一区域的输入对于控制动作的决策价值越高。因此,结合显著图中的显著区域与当下时刻采用的控制动作,可鉴别强化学习识别到的微观特征及控制原理。

本文采用基于梯度的方法,通过计算动作值函数对状态的Jacobian矩阵,量化环境状态图像中每个像素点对于动作价值的影响程度,并可视化构建环境状态的显著图。显著图构建流程如图3所示,主要包括以下步骤:

(1) 计算Jacobian矩阵。将动作值函数Q中的参数w作为输入,环境状态$\boldsymbol{s}$作为变量,并计算相应的导数,即$\dfrac{\mathrm{d}Q(\boldsymbol{s},a;w)}{\mathrm{d}\boldsymbol{s}}$,得到环境状态$\boldsymbol{s}$的Jacobian矩阵。

(2) 平均化。选取由连续的3帧交通图像组成的三维矩阵作为状态环境输入。为了便于分析,对3帧图像形成的位置矩阵平均化得到二维矩阵。

(3) Gaussian过滤。采用Gaussian平滑对平均化后的二维图像矩阵进行过滤,以消除噪声的影响。

(4) 划分正负梯度。分别将梯度为正值和负值的像素点提取出来,并为空缺区域补零,形成2个梯度矩阵$\boldsymbol{g}^+$和$\boldsymbol{g}^-$,以分析区域特征对动作值函数的正面和负面影响。

(5)归一化。分别对正负梯度矩阵$\boldsymbol{g}^+$和$\boldsymbol{g}^-$进行最大最小值归一化,得到归一化后的正负梯度矩阵$|\boldsymbol{g}_{\mathrm{N}}^+|$和$|\boldsymbol{g}_{\mathrm{N}}^-|$,归一化后$|\boldsymbol{g}_{\mathrm{N}}^+|$和$|\boldsymbol{g}_{\mathrm{N}}^-|$的取值范围均为(0,0.5)。

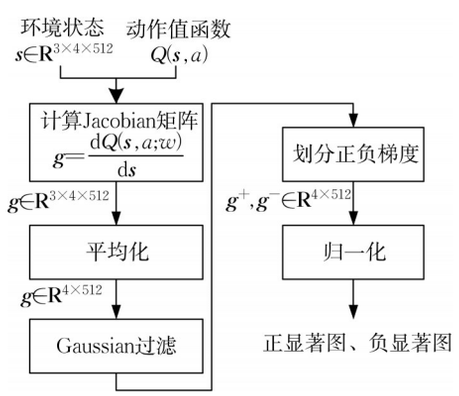

图3 显著图计算流程

Fig.3 Calculation process of saliency map

通过以上步骤对$|\boldsymbol{g}_{\mathrm{N}}^+|$和$|\boldsymbol{g}_{\mathrm{N}}^-|$进行可视化即可得到正显著图(PSM)和负显著图(NSM)。在PSM中,若某区域梯度值较大,则说明模型认为该区域出现车辆有利于提升动作价值,即有利于提高通行效率。在NSM中,若某区域梯度值较大,则说明模型认为该区域出现车辆会降低动作价值,即有可能降低通行效率。

当前环境下的车辆位置分布可以直观地反映当前的交通状态。将车辆位置与相应的PSM和NSM叠合,用于理解影响匝道控制的关键环境特征和匝

道控制动作规律。首先，通过分析PSM与NSM中正负梯度矩阵$|g_N^+|$和$|g_N^-|$的分布特征，鉴别对动作价值影响较大的区域，并结合车辆位置对其现实意义进行合理推断；其次，结合显著图的主要特征与即将采取的信号控制动作，建立环境特征与控制动作之间的联系，分析匝道控制的动作规律。

2.3　输入扰动

遮挡原始环境的部分区域，形成扰动环境，并以此为控制输入，输出匝道控制动作和控制效果；对比分析原始和扰动环境下输出动作和控制效果的差异，以鉴别各区域信息对匝道控制决策的重要性。

通过动作匹配率和行程时间2个指标对区域重要性进行评价。动作匹配率R_a表示未扰动和扰动环境下控制动作的一致性，用于衡量各区域内信息对匝道控制决策的贡献度，定义为

$$R_a=\frac{\sum_n\sum_t I(a_{n,t},\hat{a}_{n,t})}{N_{\text{sim}}N_a} \tag{5}$$

式中：$a_{n,t}$、$\hat{a}_{n,t}$分别为原始和扰动环境下第n次仿真t时刻采用的信号相位；$I(x,y)$为示性函数，当x和y相等时输出1，不相等时输出0；N_a为一次仿真内动作决策总数；N_{sim}为所有仿真次数。在计算动作匹配率R_a时，基于原始环境状态和扰动环境状态分别对信号相位$a_{n,t}$和$\hat{a}_{n,t}$进行决策，但最终执行的相位永远都是$a_{n,t}$。

动作匹配率可以衡量局部区域特征对控制决策的贡献度，但无法判断被扰动区域的特征是否有利于匝道控制做出更合适的控制动作。因此，采用主线平均行程时间衡量控制效果，对比采用$a_{n,t}$和$\hat{a}_{n,t}$信号相位的控制效果。

3　案例分析

3.1　实验设计

以典型匝道控制场景为实验场景，道路拓扑如图4所示。实验路段包含主线、上匝道和合流区三部分，其中主线为三车道，上匝道为单车道。匝道信号灯布设于上匝道尽头。采用青岛青黄隧道实际交通流量数据，对该场景下的早高峰时段(08:00—09:00)进行交通流仿真。早高峰每10 min主线和匝道交通流量变化如图5所示。采用训练好的基于深度强化学习的匝道控制模型，进行20次仿真实验，用于后续分析评价。

3.2　结果分析

为了验证基于深度强化学习的匝道控制模型的控制效果，采用不同的随机种子进行20次仿真，计

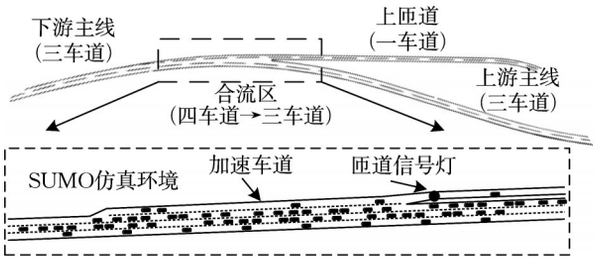

图4　实验场景

Fig.4　Experimental scenario

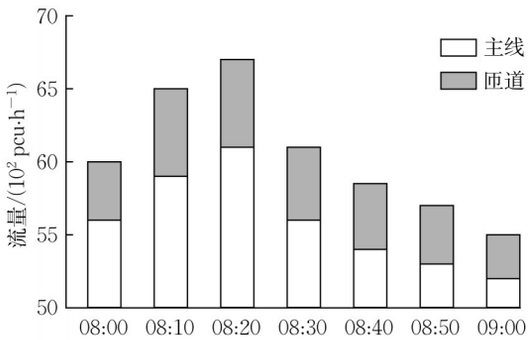

图5　主线及上匝道交通流量

Fig.5　Traffic flow on mainline and on-ramp

算每一次仿真中控制与无控制下的主线平均行程时间，并在图6中绘制了箱形图以展示模型控制效果。结果表明，基于深度强化学习的匝道控制模型可以实现有效控制，降低主线平均行程时间。在20次仿真中，基于深度强化学习的匝道控制模型下主线平均行程时间变异性更小，说明通过有效控制可以减小交通受随机因素的影响。

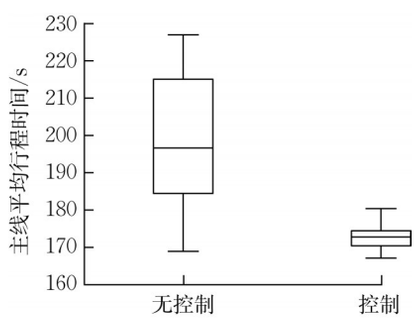

图6　无控制与控制下主线平均行程时间箱形图

Fig.6　Box plot of average mainline travel time without and with control

3.2.1　状态值函数

采用通过合流区的主线平均行程时间作为交通状态优劣的衡量指标，对比主线平均行程时间和匝道控制算法给出的状态值，分析模型是否能够准确识别交通状态。图7为第1次仿真中主线平均行程

时间与状态值随时间的变化。可以看出,状态值与主线平均行程时间随时间的变化趋势具有较好的相关性。仿真初期交通流量较低,主线平均行程时间较短,状态值较高。仿真中期交通拥堵开始形成,行车时间逐渐加长,状态值较低。随着拥堵的消散,主线平均行程时间降低,但状态值上升。该结果表明模型可辨别区域内交通状态的变化。

从图7中的状态值$V(s)$变化趋势可以看出,在拥堵和畅通状态下,状态值均存在短时间内的剧烈波动,即局部波动,如7图中的t_1~t_3时刻以及t_4~t_6时刻。局部波动说明模型意识到环境状态微小的改变对后续状态的影响。

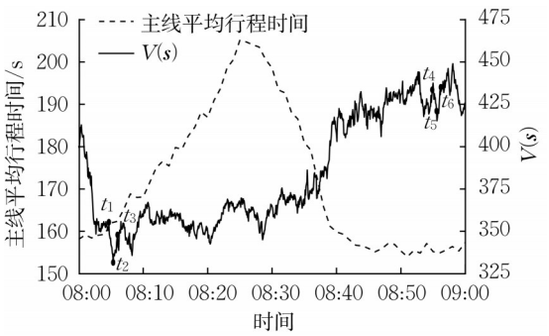

图7 主线平均行程时间及状态值随时间的变化

Fig.7 Variation of average mainline travel time and state value with time

3.2.2 显著图分析

以t_1~t_3和t_4~t_6时刻为例,利用显著图分析模型对状态环境的理解和决策的合理性。图8、9将t_1~t_3和t_4~t_6的车辆位置与相应的NSM和PSM叠合,结合当前环境状态下的交通状态与正负梯度的分布特征进行分析。

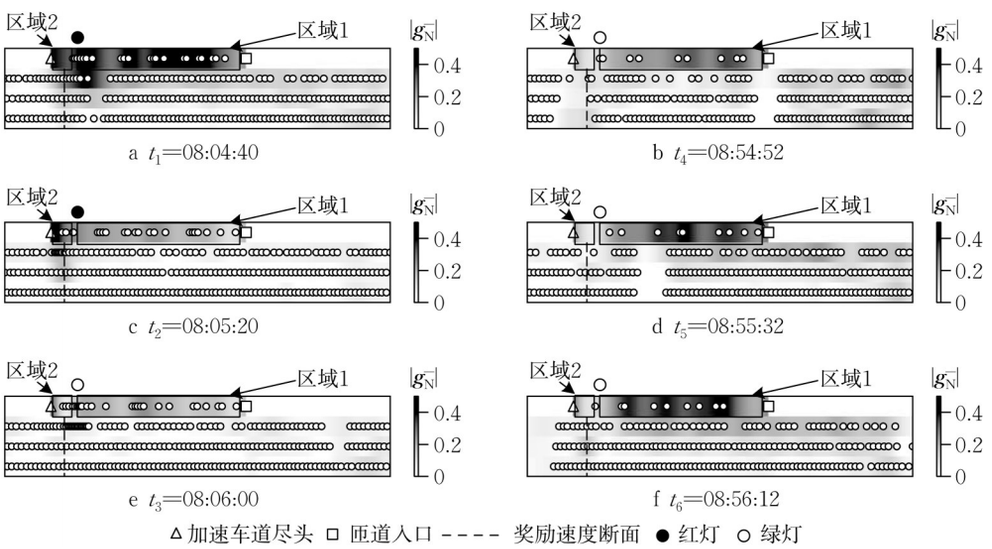

图8 t_1~t_6时刻NSM

Fig.8 NSM at t_1~t_6

图8中区域1为匝道部分,与其他3根车道相比,在6个时刻该区域的负梯度均明显较高,说明模型能够意识到匝道的存在,且认为该区域出现车辆会恶化交通状态。图8中区域2为加速车道部分,当匝道控制信号相位为红灯时(见图8a、c),靠近加速车道尽头处的负梯度值明显较高,说明匝道控制意识到加速车道的存在,且认为红灯时车辆在加速车道排队进入主线会降低交通效率。该发现与实测研究成果[18]具有很好的一致性,即加速车道尽头的车辆汇入主线时速度较低,会对主线车流造成较大的干扰,影响后继的交通状态。

图9中的区域3是一个明显的空档,该区域的正梯度值明显较高,说明模型可感知到车队间的空档,认为空档降低了主线道路的利用率。同时,结合状态值(见图7)与显著图分析发现,模型可区分空档位置对交通状态的影响。t_6时刻(见图9f)合流区上游未有明显空档,状态值最高,模型认为此时交通状态最好。在t_4(见图9b)和t_5(见图9d)时刻合流区上游均有明显空档,状态值较低,并且空档越接近合流区状态值越低。

图9中的区域4为外侧车道合流区上游较远位置的空档,该区域的正梯度值较低,说明模型无法判断该空档对状态值的影响。一方面,空档的出现降低了主线道路利用率;另一方面,外侧车道空档有利于匝道车辆汇入主线。为了进一步探究模型能否判断靠近合流区的空档对于状态值的影响,图10展示

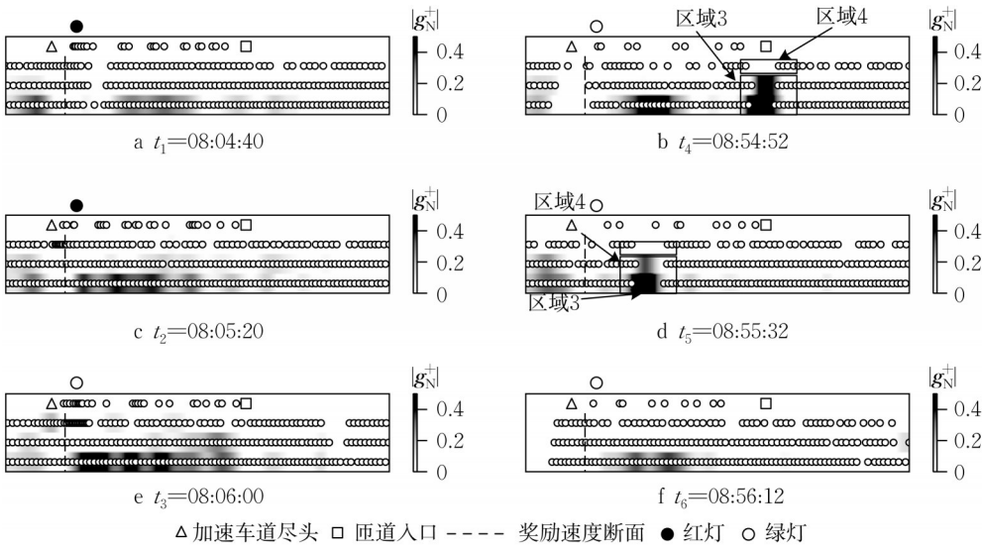

图9　t_1~t_6 时刻PSM

Fig.9　PSM at t_1~t_6

了 t=08:55:52时的NSM与PSM。此时外侧车道在靠近加速车道尽头处出现空档,同时匝道上没有车辆可利用该空档,模型认为此时道路资源未被充分利用,空档区域正梯度值较高。

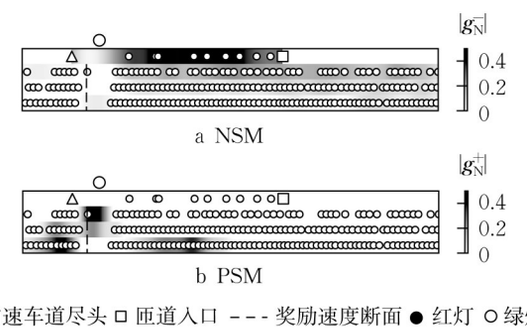

图10　t=08:55:52时刻的NSM与PSM

Fig.10　NSM and PSM at t=08:55:52

图8、9和即将采取的信号控制动作相结合,反映了模型可根据环境状态特征做出合理动作。t_1时刻(见图8a)匝道上车辆排队较长,负梯度值较大,信号灯处于红灯状态,即将切换成绿灯相位,以疏散匝道排队;t_2时刻(见图8c)加速车道尽头车辆排队积压,负梯度值较大,信号灯处于红灯状态,并将继续保持,避免匝道车辆进入加速车道等待汇入主线而加重拥堵。

3.2.3　输入扰动

如图11a所示,将感知环境划分为8个区域:区域1为合流区下游区域;区域2为合流区,包括加速车道、匝道入口和匝道信号灯;区域3、4、5靠近合流区,为合流区近端上游区域;区域6、7、8为合流区远端区域。对不同区域进行扰动后,得到的匝道控制动作匹配率如图11b所示。

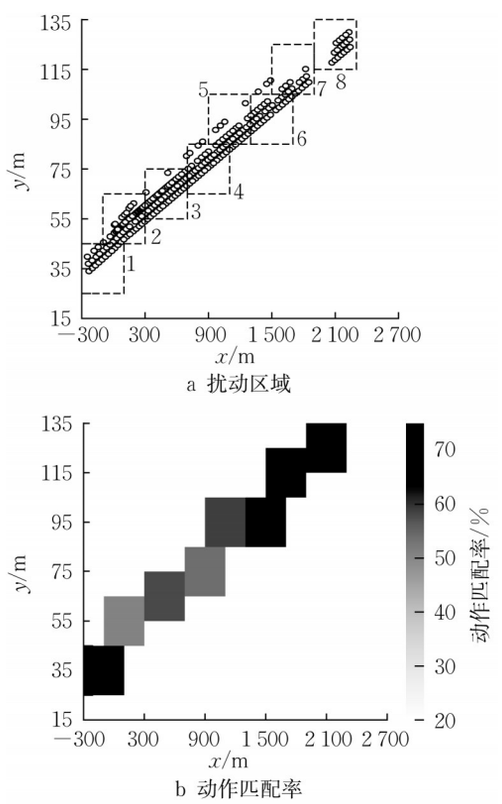

图11　扰动区域与动作匹配率

Fig.11　Perturbated area and action match ratio

图11表明,合流区下游和远端上游不是模型关注的重点。区域1、6、7和8的动作匹配率均大于70%,其中区域1的指标值最大,达到75%,说明这些区域的信息对控制动作选择影响较小。此外,合

流区及其近端上游区域是模型环境感知的重点区域,区域2~5的动作匹配率较小,均在60%左右,其中区域2、4的指标值最小,接近50%,说明这2个区域的信息对于动作决策起到重要作用。

为了分析扰动对于匝道控制效果的影响,图12展示了对不同区域进行扰动后的主线平均行程时间分布,其中0表示未扰动环境。可以看出,区域1、5、6、7、8的扰动不会明显增加主线平均行程时间,而对区域2、3、4的扰动会明显增加主线平均行程时间,表明区域2、3、4的特征对匝道控制效果具有重要影响。

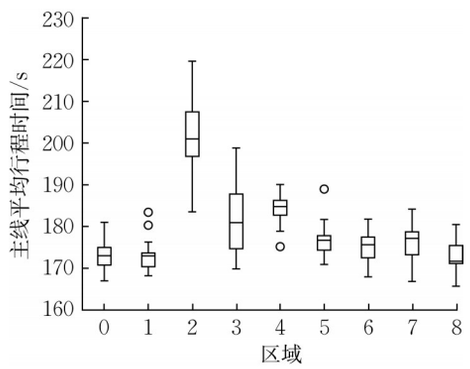

图12 输入扰动后主线平均行程时间箱形图

Fig.12 Box plot of average mainline travel time under input perturbation

4 结论

(1)基于深度强化学习的匝道控制模型能够认识到环境状态中影响匝道控制决策的关键特征,如匝道排队、加速车道排队和主线空档,并辨别上述特征对于交通状态的正面和负面影响。

(2)基于深度强化学习的匝道控制模型能够根据感知到的关键特征做出合理的动作决策。结合显著图和对应的控制动作发现,该模型能够根据主线和匝道交通状态控制信号灯相位,从而提升交通效率。

(3)基于深度强化学习的匝道控制模型主要关注合流区及其近端上游区域的信息,缺少这些区域信息的模型控制效果显著下降,合流区下游和远端上游的信息对控制动作影响较小。

未来研究将从4个方面展开:由于实际应用环境复杂,难以对显著图中所有的显著特征进行分析,不同显著特征的实际含义尚需进一步研究;基于本文成果,优化交通检测器部署,为匝道控制提供经济有效的信息;预先从视频图像中提取匝道控制主要关注的特征,提升深度强化学习模型的训练速度,改善模型的控制效果;通过元学习或迁移学习增强模型的泛化能力,使其适应更加多样化的匝道控制场景,并探究深度强化学习模型在不同场景下的控制效果及决策机理。

作者贡献声明:

刘 冰:模型构建,研究方案实施,论文撰写。

唐 钰:提供研究思路,模型构建,论文完善。

暨育雄:提供研究思路,技术指导,论文完善。

沈 煜:技术指导,论文完善。

杜豫川:技术指导,论文完善。

参考文献:

[1] LI Zhenning, YU Hao, ZHANG Guohui, *et al*. Network-wide traffic signal control optimization using a multi-agent deep reinforcement learning[J]. Transportation Research, Part C: Emerging Technologies, 2021, 125: 103059.

[2] ZHANG Chengwei, TIAN Yu, ZHANG Zhibin, *et al*. Neighborhood cooperative multiagent reinforcement learning for adaptive traffic signal control in epidemic regions[J]. IEEE Transactions on Intelligent Transportation Systems, 2022, 23(12):25157.

[3] CHU Tianshu, WANG Jie, CODECÀ L, *et al*. Multi-agent deep reinforcement learning for large-scale traffic signal control[J]. IEEE Transactions on Intelligent Transportation Systems, 2020, 21(3): 1086.

[4] WANG Chong, XU Yang, ZHANG Jian, *et al*. Integrated traffic control for freeway recurrent bottleneck based on deep reinforcement learning[J]. IEEE Transactions on Intelligent Transportation Systems, 2022, 23(9):15522.

[5] HAN Yu, WANG Meng, LI Linghui, *et al*. A physics-informed reinforcement learning-based strategy for local and coordinated ramp metering[J]. Transportation Research, Part C: Emerging Technologies, 2022, 137: 103584.

[6] 韩靖. 基于强化学习的城市快速路交织区入口匝道智能控制方法[D]. 南京:东南大学,2017.
HAN Jing. The intelligent on-ramp metering at urban expressway weave area[D]. Nanjing: Southeast University, 2017.

[7] HEUILLET A, COUTHOUIS F, DÍAZ-RODRÍGUEZ N. Explainability in deep reinforcement learning[J]. Knowledge-Based Systems, 2021, 214: 106685.

[8] WELLS L, BEDNARZ T. Explainable AI and reinforcement learning: a systematic review of current approaches and trends[J]. Frontiers in Artificial Intelligence, 2021, 4: 550030.

[9] WANG Z, SCHAUL T, HESSEL M, *et al*. Dueling network architectures for deep reinforcement learning [C]// Proceedings of the 33rd International Conference on Machine

(下转第981页)