DOI:10.13203/j.whugis20210409 文章编号:1671-8860(2021)12-1862-10

联合图像与单目深度特征的强化学习端到端自动驾驶决策方法

卢 笑[1] 竺一薇[1] 阳牡花[1] 周炫余[2] 王耀南[3]

1 湖南师范大学工程与设计学院,湖南 长沙,410006
2 湖南师范大学基础教育大数据研究与应用重点实验室,湖南 长沙,410006
3 湖南大学机器人视觉感知与控制技术国家工程实验室,湖南 长沙,410002

摘 要:现有的基于深度强化学习(deep reinforcement learning,DRL)的端到端自动驾驶决策方法鲁棒性较低,存在安全隐患,且单纯依赖图像特征难以正确推断出复杂场景下的最优动作。对此,提出了一种联合图像与单目深度特征的强化学习端到端自动驾驶决策方案。首先,建立了基于竞争深度Q网络(dueling deep Q-network, Dueling DQN)的端到端决策模型,以提高模型的策略评估能力和鲁棒性。该模型根据观测数据获取当前状态,输出车辆驾驶动作(油门、转向和刹车)的离散控制量。然后,在二维图像特征的基础上提出了联合单目深度特征的状态感知方法,在自监督情况下有效提取场景深度特征,结合图像特征共同训练智能体网络,协同优化智能体的决策。最后,在模拟仿真环境下对不同的行驶环境和任务进行算法验证。结果表明,该模型可以实现鲁棒的端到端无人驾驶决策,且与仅依赖图像特征的方法相比,所提出的方法具有更强的状态感知能力与更准确的决策能力。

关键词:端到端自动驾驶决策;竞争深度Q网络;图像特征;单目深度特征

中图分类号:P237 **文献标志码**:A

在过去的近十年中,自动驾驶领域相关研究得到政府、科研机构、车企及互联网企业等人工智能相关产业的高度关注而持续推进[1]。其中,自动驾驶决策技术是提高主动安全性能和减少交通事故的关键技术。传统的自动驾驶策略建立在规定的交通规则基础之上,需要准确识别场景中的交通标志、信号灯、行人和车辆等障碍物,分割出车道线及可通行的道路等,进而利用既定的规则进行控制决策。然而数学建模的浅层逻辑规则在面对真实世界的复杂路况时往往收效甚微。近年来,深度学习技术的快速发展较大促进了自动驾驶决策领域的发展进程,使得自动驾驶决策能避开复杂道路状况下的规则式专家系统,将场景理解和驾驶决策都交给神经网络执行,传感器数据经卷积神经网络(convolutional neural network,CNN)处理后直接输出车辆控制信号,从而实现端到端的自动驾驶决策。

目前,端到端的自动驾驶决策[2]可分为基于深度学习和基于深度强化学习(deep reinforcement learning,DRL)两种方法。其中基于深度学习的方法需要采集大规模驾驶数据集并通过人工对其进行标注,然后搭建深度学习模型进行训练。例如,文献[3]构建了一个由3层全连接层构成的神经网络进行跟车控制;文献[4]提出了基于CNN的端到端注意力模型,该网络的输入为道路图像,输出为车辆控制的转向角。大规模训练数据的标注问题限制了基于深度学习的方法的应用。近年来,DRL[5]被越来越广泛地应用于自动驾驶决策领域。一方面,它提供了一个从探索中学习策略的系统,省去了繁杂的数据标注过程;另一方面,它结合了深度学习对高维特征的抽象和表征能力,从而有效地解决了传统强化学习对高维状态空间和动作空间的表示问题。文献[6]提出用深度确定性行动者-评论家(deep deterministic actor critic,DDAC)算法来获取更平滑的车辆运动轨迹,在开源赛车模拟平台(the open

收稿日期:2021-07-30

项目资助:国家自然科学基金(62007007,61703155);湖南省自然科学基金(2018JJ3350, 2018JJ3352)。

第一作者:卢笑,博士,讲师,主要从事智能车辆环境感知、控制与决策研究。xlu_hnu@163.com

通讯作者:王耀南,博士,教授,中国工程院院士。yaonan@hnu.edu.cn

racing car simulator,TORCS)上验证车道保持策略;文献[7]提出利用深度Q网络(deep Q-network,DQN)算法[8]在PreScan平台下学习自主刹车的能力;文献[9]提出采用DQN算法在基于学习的车辆驾驶仿真环境(car learning to act,CARLA[10])下进行十字路口处的驾驶决策。尽管基于DRL的端到端自动驾驶决策方法已经广泛应用于自动驾驶任务,但当场景发生微小改变或噪声存在时,容易出现策略评价不准确而导致决策动作完全不同或发生剧烈跳变的情况,这使得端到端的自动驾驶决策退化为一个不适定问题(相同或相似状态下做出完全不同的决策),从而成为自动驾驶任务的安全隐患。

此外,现有的基于DRL的驾驶决策方法主要将图像(单模态传感器设置)作为输入,假设最优动作可以直接从观测的图像中推断出来。然而该假设在实际中很难成立,比如当缺乏深度信息时,单纯从前方存在障碍物的图像不足以预测车辆应该刹车、保持直行还是转弯。因此,仅依赖二维图像特征存在难以理解复杂道路场景的问题。针对这一问题,研究者们尝试利用激光雷达(light detection and ranging,LiDAR)传感器数据,融合图像与深度信息(多模态传感器设置)提高环境感知能力。文献[11]提出将图像与LiDAR数据进行特征融合的条件模仿学习(conditional imitation learning,CIL)方法;文献[12]提出了一种基于红绿蓝(red green blue,RGB)图像和深度图像的端到端的CNN,实现车辆速度和转向控制,以及车道保持。文献[13]分别在前期、中期和后期对图像特征和深度特征进行融合,多组实验结果证明了基于多模态能有效改善单模态数据的决策性能。然而,多模态方法需要额外的LiDAR传感器来探测物体精确位置,目前车载同步异构传感器的安装和维护仍较为昂贵,并且不同传感器获取的数据存在同步的问题,给数据融合带来一定的难度。因此,如何有效地从单模态的视觉数据中挖掘深度特征,是提高智能体环境感知能力,从而增强自动驾驶决策能力的有效途径。

针对以上问题,本文首先构建了基于竞争深度Q网络(dueling deep Q-network, Dueling DQN)[14]的自动驾驶决策模型,以提高相似状态下或噪声存在时智能体的策略评估能力和决策能力,并提出在单目视觉图像上同时提取并融合图像特征与深度特征,优化自动驾驶环境状态表示与策略学习,提高智能体的环境感知能力。在CARLA仿真平台下对不同的行驶环境和任务进行验证,结果表明,本文所构建的模型可以实现端到端的无人驾驶决策。且基于融合特征与基于二维图像特征的对比实验结果表明,联合图像与深度特征的Dueling DQN自动驾驶决策模型具有更强的决策控制能力。

1　基于Dueling DQN的端到端自动驾驶决策模型

强化学习是指智能体在与环境的交互过程中学习策略以达到回报最大化的过程。将自动驾驶的序列决策问题视为马尔可夫决策过程[15](Markov decision process,MDP),并由4元组(s_t, a_t, r_t, s_{t+1})表示。在t时刻,智能体通过获取状态s_t确定最佳动作a_t,并执行动作a_t与环境E交互以确定奖励r_t,最终获得下一个状态s_{t+1}。

1.1　马尔可夫决策过程定义

端到端自动驾驶决策问题的马尔可夫决策过程的定义如图1所示。

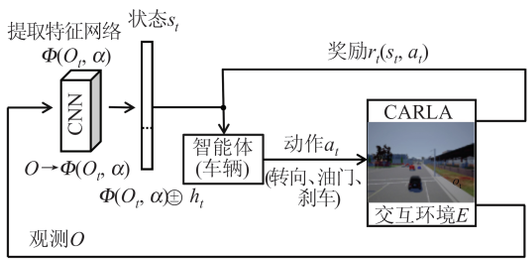

图1　端到端自动驾驶决策问题的马尔可夫决策过程

Fig.1　Markov Decision Process for End-to-End Autonomous Driving Decision

1)状态。在自动驾驶决策问题中,单模态传感器(相机)设置下,由于像素空间极为庞大,直接将彩色图像(观测O)作为智能体的状态是不可取的。通常需要利用CNN对其进行特征提取,将高维观测空间转化为较低维的状态空间。本文将该过程表示为$s_t = \Phi\left(O_t, \alpha\right) \oplus h_t$,其中$O_t$表示当前时刻的观测图像,$\Phi(\,)$表示特征提取网络,其参数用$\alpha$表示,$h_t$表示过去的历史动作,它是一个编码了过去已执行动作的向量,$\oplus$表示特征拼接。添加历史动作向量的目的是稳定搜索策略[16]。

2)动作。本文定义的对车辆的控制量包括转向、油门和刹车。考虑基于离散动作输出的决策方法,将3种控制量的输出组合为8个离散动

作，分别对应直行、不同幅度的转弯和刹车。本文定义的离散动作a_t与控制量之间的对应关系及其含义如表1所示。

表1 离散动作与控制量的对应关系

Tab.1 Corresponding Relationship Between Discrete Action and Control Quantity

序号	控制量[油门,转向,刹车]	含义
0	[1.00, 0.00, 0.00]	直行
1	[0.00, −0.50, 0.00]	左转
2	[0.00, 0.50, 0.00]	右转
3	[0.00, −0.75, 0.00]	大幅左转
4	[0.00, 0.75, 0.00]	大幅右转
5	[0.00, −0.25, 0.50]	小幅左转
6	[0.00, 0.25, 0.50]	小幅右转
7	[0.00, 0.00, 1.00]	刹车

3)奖励。奖励r_t是为了对当前时刻动作a_t的有效性进行准确评价而设置的，用于对智能体进行监督和训练。本文利用车辆反馈的测量数据定义奖励函数：

$$r_t(s_t,a_t)=\begin{cases}-200, c_t\\ +1, v_t>50\ \text{km/h}\\ -1, \text{其他}\end{cases} \tag{1}$$

式中，c_t表示t时刻车辆是否发生碰撞($c_t=1$表示有碰撞发生，否则没有)；v_t表示t时刻的行驶速度；“其他”表示碰撞传感器没有反馈碰撞事件或车速$v_t\leqslant 50$ km/h的情况。

根据以上奖励函数，t时刻智能体获得的总奖励计算公式为：

$$R=\sum_{t=1}^{t_{\max}}\lambda^{t-1}r_t \tag{2}$$

式中，$\lambda\in[0,1]$表示折扣因子，其值越大，表示总奖励R与将来动作越相关，本文设定$\lambda=0.9$。

1.2 端到端的决策模型训练方法

在强化学习中，采用动作价值函数$Q(s,a)$评价给定状态s下采取动作a的回报值，Q值越大，表示在状态s下采取动作a获得的长期回报值R越大。Q函数的迭代更新方程为：

$$Q(s,a)=Q(s,a)+\alpha\left(r+\lambda\max_{a'}Q(s',a')-Q(s,a)\right) \tag{3}$$

式中，r表示在当前时刻状态s下执行动作a的即时奖励；$Q(s',a')$表示下一个状态s'下执行动作a'的Q值；α是更新步长；$\max_{a'}$是使得Q值取最大值时对应的a'值。

在高维状态和动作空间下，利用式(3)逐个计算每个状态和动作下的Q值是无法实现的。DQN[17]采用深度网络建立一个智能体网络$Q(s,a|\theta)$来近似Q函数，其中θ表示智能体网络的参数，该网络的输入为当前状态，输出为当前状态下每个动作的Q值。

由于DQN学习得到的策略存在给定状态下不同动作对应的Q值的差别极小的问题，即当噪声存在或状态发生微小改变时可能导致决策动作的完全改变。本文采用Dueling DQN[14]的思想，利用值函数网络$V(s|\gamma,\beta)$和优势函数网络$A(s,a|\gamma,\mu)$联合估计Q函数，其中γ表示两个网络的公共参数部分，β和μ分别表示值函数网络和优势函数网络独有的参数。相比DQN的单流结构中每一次更新Q函数只能对所选择状态对应的值函数进行更新的情况，Dueling DQN的双流结构每一次对Q函数的更新过程中都对值函数进行了更新，更有利于高效地学习得到能对状态进行准确评估的值函数，从而增强Q函数的鲁棒性，避免状态改变极小的情况下决策动作的剧烈跳变。为了提高优势函数对各动作的可辨识性，对优势函数进行中心化处理，计算Q值的组合方式为：

$$Q(s,a)=V(s)+\left(A(s,a)-\frac{1}{|A|}\sum_{a'}A(s,a')\right) \tag{4}$$

最终，$Q(s,a)$的更新可通过最小化损失函数实现：

$$L(s,a)=\left(r+\lambda\max_{a'}Q(s',a')-Q(s,a)\right)^2 \tag{5}$$

由于CNN的可微性，通过最小化式(5)可端到端地学习得到特征提取网络和智能体网络的参数。智能体网络收敛表示智能体学习得到最佳决策策略π^*，该策略是相对于Q函数的贪婪策略：

$$\pi_Q^*(s)=\text{argmax}_a Q(s,a) \tag{6}$$

使用该策略可准确地实现自动驾驶决策。

2 联合图像特征与单目深度特征的自动驾驶决策方法

为提高自动驾驶过程中的环境感知能力，同时避免使用深度传感器所带来的异构数据融合问题，降低自动驾驶成本，本文提出了联合单目视觉的图像特征与深度特征的智能体环境状态感知方法，对两种特征进行融合，并以此作为智

能体网络的输入。联合深度估计的自监督损失(投影误差损失)与智能体最优策略损失(均方误差损失)对网络进行优化学习,在提高智能体环境感知能力的同时增强智能体对复杂环境的决策能力。

如图2所示,整体网络结构包含4部分:(1)图像特征提取网络,如虚线框1所示;(2)深度估计网络,如虚线框2所示;(3)特征融合模块,如虚线框3所示。这3部分的网络结构联合起来实现将观测转化为状态的功能,其参数用α表示。(4)智能体网络,如虚线框4所示。这部分网络接收融合后的特征(状态),并输出相应状态下的最优决策动作,其参数包括值函数网络参数β、优势函数网络参数μ和共享参数γ,本文中合成表示为θ。

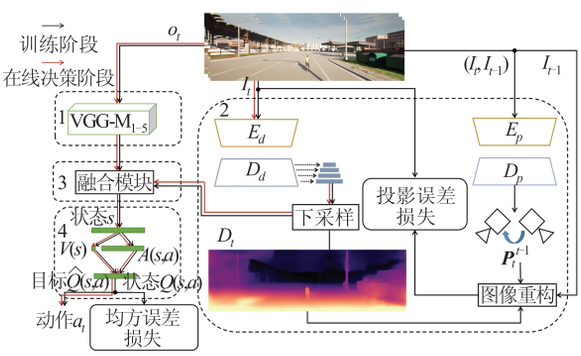

图2　整体网络结构

Fig.2　Overview of Network Structure

2.1　图像特征提取网络

本文采用VGG-M[18]的前5层卷积层作为图像特征提取网络,网络的输入为1 024×320像素的彩色图像。网络结构参数如表2所示,第1列是输入特征图的维度,其中H、W和C分别表示特征图的高度、宽度和通道数,第2列是各网络模块,其中卷积模块括号内的第一个参数表示滤波器的数量,第2、3个参数"st"和"pad"分别表示卷积步长和空间填充,括号外的×3表示采用同样的卷积层重复3次;LRN[19]表示进行局部归一化,以防止梯度消失;最大池化模块中×2 pool表示最大池化下采样因子为2。所有权重层均采用修正线性单元(rectified linear unit,ReLU)作为非线性激活函数。

2.2　深度特征提取网络

图像深度估计是感知三维场景几何结构的一个重要环节。近年来,基于深度学习的单目深度估计方法由于其算法简洁、无需特殊的双目装置而得到了广泛的关注。由于单目图像中不包含任何深度信息,需要通过估计前后帧图像之间的视差计算深度信息。为了从单目图像中提取得到准确的深度特征,需要训练连续帧间的姿态估计和深度估计网络,从而约束和引导深度特征提取网络的参数学习。本文采用文献[20]提出的自监督单目深度估计网络框架,具体如图2中虚线框2所示。在训练阶段(黑色箭头所示),位姿估计网络计算连续两帧图像间的位姿信息,深度估计网络计算当前帧图像I_t的深度信息图D_t,二者通过图像重构模块输出前帧I_{t-1}到I_t的重构图像,并通过计算重构图像与真实图像之间的差得到损失值反向传播调整网络参数。在决策阶段(红色箭头所示),无需计算位姿信息与深度信息,对当前帧提取深度卷积特征以便于图像特征融合。

表2　图像特征提取网络结构

Tab.2　Network Structure of Image Feature Extraction

输入特征图($H\times W\times C$)	模块
1 024×320×3	7×7卷积(96, st=2, pad=0)
509×157×96	LRN
509×157×96	最大池化(×2 pool)
255×79×96	5×5卷积(256, st=2, pad=1)
127×39×256	LRN
127×39×256	最大池化(×2 pool)
64×20×256	3×3卷积(512, st=1, pad=1)×3
64×20×512	最大池化(×2 pool)

2.2.1　网络结构

单目深度估计网络分为深度网络和位姿网络两部分。深度网络采用经典的U-Net[21]编码器-解码器模型结构,其编码模块E_d采用深度残差网络[22](residual network-18,ResNet-18)的前5个卷积模块,网络结构参数如表3所示;解码模块D_d参数如表4所示,它将E_d输出的特征图上采样至不同的尺度,并将不同尺度的特征图上采样至原图大小,以实现多尺度深度估计。位姿网络同样分为编码模块E_p与解码模块D_p,其中E_p与E_d有共同的网络结构;D_p由3层卷积层组成,各层参数如表5所示。使用连续两帧图像I_t和I_{t-1}作为位姿网络的输入,经由E_p编码和D_p解码后回归出对应每个像素点的运动信息,之后可利用全局平均池化求得I_{t-1}到I_t的总体轴角与平移向量,最后通过罗德里格旋转公式可将总体轴角与平移向量转换得到位姿旋转矩阵$\boldsymbol{P}_t^{t-1}$。

由于位姿网络仅在训练阶段需要,本文将深度网络解码器的4级输出上采样至最后一层大小并沿通道拼接作为深度特征,进一步输入后续特征融合模块,参与自动驾驶决策。

表3 深度网络编码器结构

Tab.3 Encoder Structure of Depth Network

输入特征($H\times W\times C$)	模块
1 024×320×3	7×7卷积(64, st=2, pad=3)
512×160×64	最大池化(3×3, st=2, pad=1)
256×80×64	3×3卷积(64, st=1, pad=1)×4
256×80×64	3×3卷积(128, st=2, pad=1)
128×40×128	3×3卷积(128, st=1, pad=1)×3
128×40×128	3×3卷积(256, st=2, pad=1)
64×20×256	3×3卷积(256, st=1, pad=1)×3
64×20×256	3×3卷积(512, st=2, pad=1)
32×10×512	3×3卷积(512, st=1, pad=1)×3

表4 深度网络解码器结构

Tab.4 Decoder Structure of Depth Network

输入特征($H\times W\times C$)	模块
64×20×256	2×2反卷积(64, st=2, pad=0)
128×40×64	2×2反卷积(32, st=2, pad=0)
256×80×32	2×2反卷积(16, st=2, pad=0)
512×160×16	2×2反卷积(8, st=2, pad=0)

表5 位姿网络解码器结构

Tab.5 Decoder Structure of Pose Network

输入特征($H\times W\times C$)	模块
32×10×512	3×3卷积(256, st=1, pad=1)
32×10×256	3×3卷积(256, st=1, pad=1)
32×10×256	1×1卷积(6, st=1, pad=1)

2.2.2 损失函数

给定连续两帧图像I_t和I_{t-1},本文通过光度重投影误差学习网络参数:

$$L_p=\min_{t-1}\mathrm{pe}(I_t,I_{t-1\to t}) \tag{7}$$

式中,L_p是深度网络的损失函数;$I_{t-1\to t}$是利用两帧之间的位姿$\boldsymbol{P}_t^{t-1}$、D_t及相机内参数矩阵$\boldsymbol{K}$将I_{t-1}映射至t时刻的结果,其计算公式为:

$$I_{t-1\to t}=I_{t-1}\left\langle\mathrm{proj}\left(D_t,\boldsymbol{P}_t^{t-1},\boldsymbol{K}\right)\right\rangle \tag{8}$$

其中,proj是利用D_t、$\boldsymbol{P}_t^{t-1}$和$\boldsymbol{K}$重投影到I_{t-1}的二维像素坐标的函数;$\langle\cdot\rangle$表示采样算子,本文使用双线性插值对I_{t-1}进行采样;pe( )为光度重投影误差损失函数,计算公式为:

$$\mathrm{pe}(I_a,I_b)=\frac{\alpha}{2}\left(1-\mathrm{SSIM}(I_a,I_b)\right)+(1-\alpha)\left\|I_a-I_b\right\|_1 \tag{9}$$

式中,SSIM( )表示结构相似性[23]损失;$\|\cdot\|_1$表示l_1范数损失;α是用于平衡两种损失重要性的参数,本文取$\alpha=0.85$。

2.3 特征融合模块

给定来自图像特征提取网络的特征$F_i\in\mathbf{R}^{H_i\times W_i\times C_i}$和单目深度提取网络的特征$F_d\in\mathbf{R}^{H_d\times W_d\times C_d}$,特征融合模块首先将后者下采样到与前者同样的大小,进一步在通道上进行拼接后,利用1×1层进行降维至C个通道,并将$H\times W\times C$(本文中$H=32$, $W=10$, $C=16$)的张量展平变为融合图像和深度特征的一维向量,并作为当前时刻状态向量的一部分输入智能体网络。总结特征融合模块的功能,表示为:

$$\boldsymbol{f}_t=f_{\mathrm{latten}}\left(f_{\mathrm{conv}}\left(F_i\oplus\left(\downarrow F_d\right)\right)\right) \tag{10}$$

式中,$\boldsymbol{f}_t$表示特征融合的结果;$\downarrow$表示双线性插值下采样操作;$\oplus$表示沿通道方向的特征拼接操作;$f_{\mathrm{conv}}(\cdot)$表示后接批归一化[24]和ReLu非线性激活函数的卷积核为$1\times1\times(C_i+C_d)$的卷积操作;$f_{\mathrm{latten}}(\cdot)$表示张量展平为向量的操作。

2.4 智能体网络

经由特征融合模块输出的特征向量$\boldsymbol{f}_t$,与历史动作向量$\boldsymbol{h}_t$拼接得到当前状态s_t,$s_t=\left(\boldsymbol{f}_t,\boldsymbol{h}_t\right)\in\mathbf{R}^{80+d}$(其中$\boldsymbol{h}_t\in\mathbf{R}^{80}$表示8个动作的10次历史,$d=32\times10\times16=5\,120$),作为智能体网络的输入。$s_t$首先经过第一个全连接层映射至1 280维;然后分别经过值网络(神经元个数为320的全连接层)和优势函数网络(神经元个数为960的全连接层),分别输出代表当前时刻t的状态值$V(s)$,和表示该状态s_t下的每个动作的重要性的优势值$A(s,a)$;最后经过式(4)所示的聚合层(输出维度为8的全连接层)得到对应于8个动作的Q值。

2.5 整体网络训练方法

整体网络训练分为两个阶段:(1)利用模拟交互环境生成连续帧图像,在给定相机内参数的前提下,训练深度估计网络;(2)利用ImageNet数据集预训练的VGG-M模型对图像特征提取网络进行初始化,同时联合第一阶段训练得到的深度估计网络参数,接入后续特征融合模块和智能体网络,以较大的学习率对特征融合模块和智能体网络进行训练,以较小的学习率对图像特征提取网络和深度特征提取网络进行微调,以训练得到最优决策模型。

由于智能体与环境交互的过程产生的序列经验具有高度的时间相关性,且采用同一智能体网络同时生成下一状态的目标Q值和更新当前状态Q值容易造成网络不稳定和不收敛,本文基

于DQN的方法,首先建立了经验回放池R,将每一个时间步长的马尔可夫决策过程作为一次经验储存以更新经验回放池,该处理可以将过去与当前的经验混合从而降低样本之间的相关性,并确保训练样本能够全面地被训练。训练过程中每次只会随机从R中抽取一定数量的经验作为样本,该方法能够有效降低数据相关性,同时R使经验得到重复使用,有利于学习效率的提高。进一步地,引入一个与智能体网络完全相同的目标网络$\hat{Q}(s',a'|\theta^-)$来估计目标Q值,目标网络的参数θ^-是每隔一定步数才会从智能体网络复制参数θ更新,这能够暂时固定训练过程的Q值,从而使智能体学习过程更稳定。最终智能体网络的训练过程通过最小化式(11)的损失函数来实现。损失函数的计算公式为:

$$L(s,a|\theta)=\left(y_i-Q(s,a|\theta)\right)^2 \tag{11}$$

式中,目标Q值y_i的更新依靠奖励r和估计$\hat{Q}$值,其计算公式为:

$$y_i=\begin{cases}r,\text{时间步长数为1}\\ r+\lambda\max_{a'}\hat{Q}(s',a'|\theta^-),\text{其他}\end{cases} \tag{12}$$

为了防止陷入局部最小值,使用ε贪婪策略允许智能体在训练过程初期探索更多的经验。其中ε从0.9开始每次下降0.1直至固定在0.1。计算公式为:

$$a_t=\begin{cases}\text{随机动作,在}\varepsilon\text{概率下}\\ \operatorname{argmax}Q(s_t,a_t),\text{其他}\end{cases} \tag{13}$$

3　CARLA环境下的仿真实验

3.1　实验环境与内容

本文实验环境为Ubuntu18.04,GPU为NVIDIA GeForce GTX 1070Ti,采用Python 3.7和Pytorch(torch1.2.0)深度学习框架,模拟环境采用开源自动驾驶仿真模拟器CARLA。

3.1.1　训练细节

深度特征提取网络训练过程中采用Adam优化器,学习率为1×10^{-4},迭代次数为11 000(训练样本共6 600张图片,批次大小设置为12,训练周期为20次);第2阶段训练过程中,采用ImageNet图像数据集预训练的权重对图像特征提取网络进行初始化,采用SGD优化器以1×10^{-4}学习率对特征融合模块和智能体网络进行训练,以1×10^{-5}学习率对图像特征提取网络和深度特征提取网络进行微调。对于智能体网络,设置经验回放池的容量为5 000,训练总片段数为30 000,每个片段的时间步长为10,批次大小设置为16,目标网络每隔1 000个步长更新一次,贪婪策略概率参数ε的初始值设为0.9,每次下降0.1直至固定在0.1。此外,设置最小奖励阈值为-1 000,以防止智能体的奖励趋于无穷小。

3.1.2　实验设置

文献[10]中定义了直行、转弯、导航及存在动态障碍物的导航4种测试任务,其中前3种任务下不存在动态障碍物。本文遵循这4种实验任务设置,分别在Town 3和Town 1两种场景中对模型进行训练和测试(Town 3用于训练,Town 1用于测试)。训练任务采用随机生成起始点的方式,测试任务采用固定起始点和终止点的方式,记为(起始点,终止点),分别为(36,40)、(68,71)、(27,130),其在地图中的显示如图3所示。

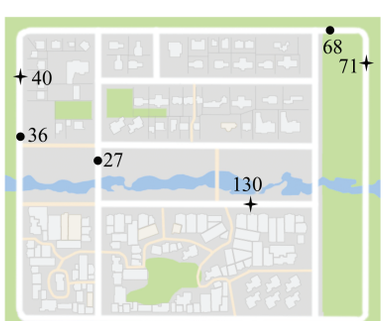

图3　测试场景中的起点(圆形)和终点(星形)

Fig.3　Start Points (in Circle) and End Points (in Star) in Test Scene

有动态障碍物条件下设置车辆数目为15,行人数目为50。训练任务的天气条件设置为晴天的日间正午时段,采样30 000个片段进行训练。为了充分验证本文所提出算法的有效性和所训练模型的泛化性能,分别在正午和夜间两种不同时段测试所训练模型的性能。

3.1.3　评估指标

测试任务是在忽略交通信号和速度限制的情况下让车辆在规划好的路径上从起始点自行决策到达终止点,所有测试任务均执行15次。如果车辆在规定的时间内到达终点即为成功,其中规定的时间是指在最优路线上以10 km/h的速度完成任务所需的时间。本文采用3种指标对算法性能进行评估,分别是任务成功次数(指完成任务的片段数量)、任务平均完成度(每次测试中车辆已行驶距离占据任务总距离的百分比/测试的次数)和违规驾驶分数(指越道、碰撞的强度,由

交互环境对车辆的测量数据给出)。

3.2 实验结果分析

3.2.1 奖励分布

奖励值的分布可以代表智能体从无到有的学习过程,直观体现强化学习的训练效果。对所训练的30 000个片段中的每100个片段进行一次奖励值统计,奖励曲线图如图4所示。

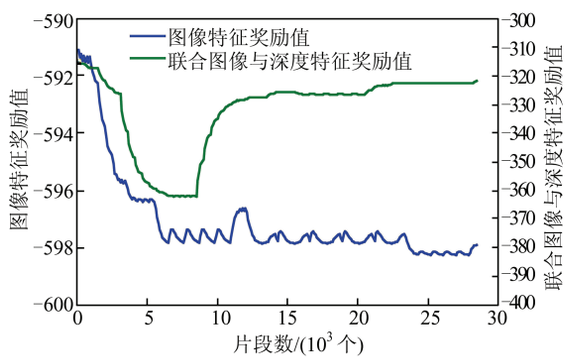

图4 各时间片段的奖励分布曲线图

Fig.4 Reward Distribution Curve of All Training Episodes

观察本文联合图像与深度特征训练模型的奖励分布曲线上的变化可知,自动驾驶决策学习过程大致可分为4个阶段:训练初期(第0~1 000片段)、训练前期(第1 100~8 500片段)、训练中期(第8 600~21 000片段)和训练末期(第21 100~30 000片段)。训练初期奖励略有波动,网络刚开始随机探索。训练前期奖励一直在下降,车辆冲向路边以及越道导致速度变慢的情况不断发生。训练中期整体奖励在往变大方向偏移,这表明车辆已经基本学到如何保持车道。训练末期奖励分布几乎没有改变,说明网络此时已经收敛。对比仅依赖图像特征的奖励变化曲线可发现,利用图像特征训练的智能体奖励变化曲线总体较为波动,说明算法收敛过程较慢,智能体难以学习获得正确的经验。本文提出的联合图像与深度特征算法获得的总体奖励远大于仅依赖图像特征的奖励,且分布趋势更稳定。

3.2.2 测试结果分析

为了充分验证本文算法的有效性和训练模型的泛化性能,分别在正午和夜间两种不同时段测试训练模型的性能。测试环境如图5所示。

为便于实验结果分析,本文将利用图像特征训练得到的模型记为RGB,联合图像与深度特征训练所得模型记为RGB-D。对4个导航任务(编号为01表示“直行”任务,编号为02表示“一次转弯”任务,编号为03表示“无动态障碍下的导航任务”,编号为04表示“有动态车辆和行人下的导航任务”),在日间和复杂夜间场景中进行15次测试,任务平均完成度的结果如表6所示。

(a)正午测试环境 (b)夜间测试环境

图5 正午和夜间两种不同测试环境

Fig.5 Two Different Test Environments at Noon and Night

表6 正午和夜间任务平均完成度测试结果/%

Tab.6 Test Results of Average Completion at Noon and Night/%

任务编号	正午		夜间	
	RGB	RGB-D	RGB	RGB-D
01	95.30	100.00	62.20	55.70
02	91.05	98.73	13.00	48.04
03	87.60	97.29	7.00	9.15
04	72.52	90.00	3.52	9.00

分析表6的结果可知:

1)对于正午时段测试,在没有车辆和行人的情况下,本文训练得到的模型已经基本学到了车道保持策略,结合深度特征大大提高了任务平均完成度。

2)对于夜间时段测试,由于存在域间隙的问题,日间场景训练出的模型无法很好地迁移至夜间场景。这是由于夜间场景光度较差,图像数据较日间情况发生了一定的变化,而且深度估计方法是基于光度一致性的假设,而夜间的昏暗场景环境一定程度上违背了这一假设。但相比融合图像与深度特征的模型,仅依赖图像特征训练的模型的夜间测试结果在转弯任务、有动态障碍物和无动态障碍物的综合任务中都较差。这说明仅依赖图像特征训练得到的模型对域间隙问题更加敏感,而深度特征对域间隙导致的模型降质具有一定的缓解作用。

将本文提出的算法与文献[25]提出的采用图像特征的异步优势动作评价算法(asynchronous advantage actor-critic,A3C)训练得到的自动驾驶模型进行对比,汇总任务完成次数、任务平均完成度、越道率及障碍物碰撞强度的结果,如表7所示。由表7可知,A3C算法的任务平均完成度略低于本文的基于图像特征的Dueling DQN算法模型RGB,远低于联合图像与深度特征的

Dueling DQN算法模型RGB-D的结果,进一步说明联合图像与单目深度特征有利于提高智能体环境感知的能力,进而增强自动驾驶决策能力。

表7　不同行驶任务下的测试结果

Tab.7　Test Results Under Different Driving Tasks

方法	任务编号	(起始点,终止点)	(车辆,行人数)	任务完成次数	任务平均完成度/%	越道率/%	车辆碰撞	行人碰撞
A3C[25]	01	(36, 40)	(0, 0)	12	70.80	5.12	0.00	0.00
	02	(68, 71)	(0, 0)	4	23.01	0.00	0.00	0.00
	03	(27, 130)	(0, 0)	1	14.26	46.74	8 125.60	0.0
	04	(27, 130)	(15, 50)	0	7.42	76.93	4 542.44	120.71
RGB	01	(36, 40)	(0, 0)	15	95.36	0.00	0.00	0.00
	02	(68, 71)	(0, 0)	15	91.05	0.00	0.00	0.00
	03	(27, 130)	(0, 0)	13	87.62	0.00	0.00	0.00
	04	(27, 130)	(15, 50)	9	72.52	5.00	3 759.81	0.00
RGB-D	01	(36, 40)	(0, 0)	15	100.00	0.00	0.00	0.00
	02	(68, 71)	(0, 0)	15	98.73	0.00	0.00	0.00
	03	(27, 130)	(0, 0)	15	97.27	0.00	0.00	0.00
	04	(27, 130)	(15, 50)	12	90.00	0.00	2 807.82	0.00

3.2.3　策略评估能力分析

为了证明本文提出的基于Dueling DQN的端到端自动驾驶决策方法在策略评估能力方面的鲁棒性,将基于A3C的端到端自动驾驶决策智能体[25]与本文方法进行对比。分别输出直行任务过程中每个时间步长不同动作对应Q值的平均值、最优动作与次优动作的Q值差值及二者之间的量级比较,作为衡量算法策略评估能力的依据(见图6)。最优动作与次优动作之间的Q值差值与均值的比值越大,说明智能体策略评估的抗噪声能力越强,即鲁棒性越强。

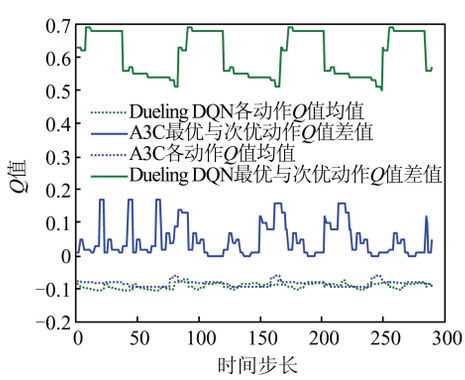

图6　两种模型在不同时间步长上Q值均值与最优动作和次优动作Q值差值分布图

Fig.6　Average of Q-value and Difference Between Optimal Action and Sub-optimal Action at Different Time-Step

由图6可知,本文训练的智能体最优动作与次优动作Q值差值分布于0.5～0.7,A3C训练的Q值差值分布于0～0.2,而二者的Q值均值相差不大,这说明本文算法在策略评估方面具有更强的鲁棒性。理论上,在直行任务中,由于不存在障碍物,智能体的决策动作应当始终保持直行(对应编号为0)。图7为两种模型直行任务的决策结果,由图7可见,A3C训练的智能体存在部分决策结果跳变的情况,而本文算法决策结果更为准确。

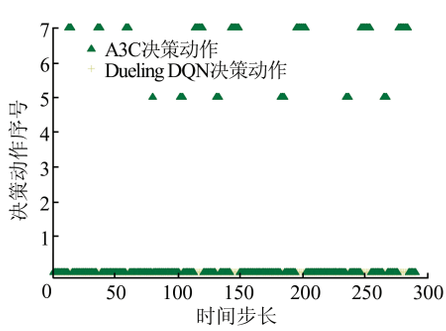

图7　两种模型在直行任务中的决策动作

Fig.7　Decision Actions of Two Models in Straight Task

3.2.4　模型耗时分析

本文基于Dueling DQN和联合图像与单目深度特征的自动驾驶决策方法在单张GTX 1070Ti GPU测试环境下,模型执行一次动作的响应时间约为10 ms,满足实时自动驾驶决策需求。

4　结　语

本文提出了联合图像与单目深度特征的端

到端深度强化学习自动驾驶决策方法。首先,采用Dueling DQN提高了智能体对策略的评价能力,以缓解面对相似场景时智能体输出不同动作带来的安全隐患;其次,采用自监督的方式从单目图像中提取深度特征,并与图像特征进行融合,增强智能体环境感知能力并指导智能体学习更鲁棒的自动驾驶策略。本文的实验结果为使用单模态传感器获取深度信息提高自动驾驶决策能力提供了一定的参考。未来将考虑采用域适应的方法进一步解决夜间自动驾驶决策能力降质的问题,提高夜间场景的泛化能力。

参 考 文 献

[1] IMT-2020(5G) Promotion Group. C-V2X White Paper[R]. Beijing: China Information and Communication Research Institute, 2018 (IMT-2020(5G)推进组. C-V2X白皮书[R]. 北京:中国信息通信研究院,2018)

[2] 5G Automated Driving Alliance. 5G Automated Driving White Paper[C]//Intelligent Transportation Forum, Guangzhou, China, 2018 (5G自动驾驶联盟. 5G自动驾驶白皮书[C]//智慧交通论坛,广州,中国,2018)

[3] Pomerleau D A. ALVINN: An Autonomous Land Vehicle in a Neural Network[C]//The Third Conference on Neural Information Processing Systems, Denver, Colorado, USA, 1989

[4] Cultrera L, Seidenari L, Becattini F, et al. Explaining Autonomous Driving by Learning End-to-End Visual Attention[C]//IEEE Conference on Computer Vision and Pattern Recognition Workshops, Seattle, Washington, USA, 2020

[5] Mnih V, Badia A P, Mirza M, et al. Asynchronous Methods for Deep Reinforcement Learning[C]//International Conference on Machine Learning, New York, USA, 2016

[6] Sallab A E L, Abdou M, Perot E, et al. Deep Reinforcement Learning Framework for Autonomous Driving[J]. *Electronic Imaging*, 2017, 2 017(19): 70-76

[7] Chae H, Kang C M, Kim B D, et al. Autonomous Braking System via Deep Reinforcement Learning[C]//The 20th International Conference on Intelligent Transportation Systems, Yokohama, Japan, 2017

[8] Mnih V, Kavukcuoglu K, Silver D, et al. Human-Level Control Through Deep Reinforcement Learning[J]. *Nature*, 2015, 518(7 540): 529-533

[9] Isele D, Rahimi R, Cosgun A, et al. Navigating Occluded Intersections with Autonomous Vehicles Using Deep Reinforcement Learning[C]//IEEE International Conference on Robotics and Automation (ICRA), Brisbane, Australia, 2018

[10] Dosovitskiy A, Ros G, Codevilla F, et al. CARLA: An Open Urban Driving Simulator[C]//Conference on Robot Learning, California, USA, 2017

[11] Sobh I, Amin L, Abdelkarim S, et al. End-to-End Multi-modal Sensors Fusion System for Urban Automated Driving[C]//Neural Information Processing Systems, Machine Learning in Inetelligent Transportation MLITS Workshop, Canada, 2018

[12] Huang Z, Lü C, Xing Y, et al. Multi-modal Sensor Fusion-Based Deep Neural Network for End-to-End Autonomous Driving with Scene Understanding[J]. *IEEE Sensors Journal*, 2020, 21(10): 11 781-11 790

[13] Xiao Y, Codevilla F, Gurram A, et al. Multimodal End-to-End Autonomous Driving[J]. *IEEE Transactions on Intelligent Transportation Systems*, 2020, DOI: 10.1109/TITS.2020.3013234

[14] Wang Z, Schaul T, Hessel M, et al. Dueling Network Architectures for Deep Reinforcement Learning[C]//International Conference on Machine Learning, New York, USA, 2016

[15] Sutton R S, Barto A G. Reinforcement Learning: An Introduction[M]. Cambridge: MIT Press, 2018

[16] Caicedoj C, Lazebnik S. Active Object Localization with Deep Reinforcement Learning[C]// International Conference on Computer Vision, Santiago, Chile, 2015

[17] Minh V, Kavukcuoglu K, Silver D, et al. Playing Atari with deep reinforcement learning[C]//Advances in Neural Information Processing Systems Deep Learning Workshop, Lake Tahoe, Nevada, USA, 2013

[18] Chatfield K, Simonyan K, Vedaldi A, et al. Return of the Devil in the Details: Delving Deep into Convolutional Nets[C]// The British Machine Vision Conference, Nottingham, Britain, 2014

[19] Krizhevsky A, Sutskever I, Hinton G E. ImageNet Classification with Deep Convolutional Neural Networks[C]//The 26th Conference on Neural Information Processing Systems, Lake Tahoe, USA, 2012

[20] Godard C, Mac Aodha O, Firman M, et al. Digging into Self-supervised Monocular Depth Estimation[C]// International Conference on Computer Vision, Seoul, Korea, 2019

[21] Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation[C]//International Conference on Medical Ima-

ge Computing and Computer-Assisted Intervention, Cham, Munich, Germany, 2015

[22] He K, Zhang X, Ren S, et al. Deep Residual Learning for Image Recognition[C]//IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 2016

[23] Wang Z, Bovik A C, Sheikh H R, et al. Image Quality Assessment: From Error Visibility to Structural Similarity[J]. *IEEE Transactions on Image Processing*, 2004, 13(4): 600-612

[24] Ioffe S, Szegedy C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift [C]//International Conference on Machine Learning, Miami, Florida, USA, 2015

[25] Mnih V, Badia A P, Mirza M, et al. Asynchronous Methods for Deep Reinforcement Learning[C]//International Conference on Machine Learning, Anaheim, California, USA, 2016

Reinforcement Learning Based End-to-End Autonomous Driving Decision-Making Method by Combining Image and Monocular Depth Features

LU Xiao[1]　*ZHU Yiwei*[1]　*YANG Muhua*[1]　*ZHOU Xuanyu*[2]　*WANG Yaonan*[3]

1　College of Engineering and Design, Hunan Normal University, Changsha 410006, China

2　Key Laboratory of Big Data Research and Application for Basic Education, Hunan Normal University, Changsha 410006, China

3　National Engineering Laboratory for Robot Visual Perception and Control Technology, Hunan University, Changsha 410002, China

Abstract: **Objectives**: Existing deep reinforcement learning (DRL)based end-to-end autonomous driving decision-making method is low robustness to noise, which would lead to safety problem. It is difficult to infer the optimal decision accurately by relying solely on the image features when facing with the complex scenes. **Methods**: An end-to-end decision-making model based on dueling deep Q-network(Dueling DQN) is established to improve the ability of decision evaluation and improve the robustness of the model. It obtains the current state according to the observed data, and outputs discrete quantities for controlling the vehicle (including throttle, steering and brake). The monocular depth feature is extracted accurately in a self-supervised learning manner, and which is combined with the image features for better representation of the current state. **Results**: The proposed method is tested in a simulation environment. (1) The comparison results with the state-of-the-art A3C model show that our Dueling DQN-based model is more robustness. (2) The comparison results with the image feature-based model show that combining the image and depth features is more beneficial to improve the decision-making accuracy. **Conclusions**: Training an agent with Dueling DQN is beneficial to alleviate the security risks caused by making different decision when facing similar scenes. Training an agent together with image features and depth features is beneficial to enhance the agent's ability of environment perception, and improve the decision-making accuracy.

Key words: end-to-end automatic driving decision-making; dueling deep Q-learning network; image features; monocular depth features

First author: LU Xiao, PhD, lecturer, specializes in intelligent vehicle environment perception, control and decision-making technology. E-mail: xlu_hnu@163.com

Corresponding author: WANG Yaonan, PhD, professor, Academician of Chinese Academy of Engineering. E-mail: yaonan@hnu.edu.cn

Foundation support: The National Natural Science Foundation of China(62007007, 61703155); the Natural Science Foundation of Hunan Province(2018JJ3350, 2018JJ3352).

引文格式:LU Xiao, ZHU Yiwei, YANG Muhua, et al. Reinforcement Learning Based End-to-End Autonomous Driving Decision-Making Method by Combining Image and Monocular Depth Features[J]. Geomatics and Information Science of Wuhan University, 2021, 46(12):1862-1871.DOI:10.13203/j.whugis20210409(卢笑,竺一薇,阳牡花,等.联合图像与单目深度特征的强化学习端到端自动驾驶决策方法[J].武汉大学学报·信息科学版,2021,46(12):1862-1871.DOI:10.13203/j.whugis20210409)