DOI:10.19651/j.cnki.emt.2211239

基于改进 PPO 算法的自动驾驶技术研究*

姚 悦 吉明佳 杨 霄
(北方自动控制技术研究所 太原 030000)

摘 要: 针对强化学习在解决端到端自动驾驶行为决策问题时面临采样效率低、环境适应性差、决策效果不佳的问题,提出循环近端策略优化算法(RPPO),采用 LSTM 与移动翻转瓶颈卷积模块构建策略网络与价值网络,有效整合前后帧的关联信息,实现智能体对多变情况的预测,提高智能体对环境的快速认知能力,并在价值网络添加 L2 正则化层,进一步提高算法的泛化能力,最后手动设置智能体在 2 个连续帧中保持动作不变,引入先验知识约束搜索空间,加快算法收敛。通过 CARLA 开源模拟环境测试,该改进方法与传统方法相比,奖励曲线明显占优,且直行、转弯、指定路线行驶 3 类任务的成功率分别提高了 10%、16%、30%,证明提出的方法更有效。

关键词: 自动驾驶;强化学习;移动翻转瓶颈卷积;LSTM
中图分类号: TP3 **文献标识码:** A **国家标准学科分类代码:** 520.6

Research on autonomous driving technology based on improved PPO algorithm

Yao Yue Ji Mingjia Yang Xiao
(North Automatic Control Technology Institute, Taiyuan 030000, China)

Abstract: To address the problems of low sampling efficiency, poor environmental adaptation, and poor decision making that reinforcement learning faces in solving end-to-end autonomous driving behavioral decision problems, a recurrent proximal policy optimization (RPPO) algorithm is proposed, which introduces a mobile inverted bottleneck convolution module and LSTM to construct a policy network and a value network, which effectively integrate the correlation information of front and back frames to achieve the prediction of multivariate situations by the intelligent body, improve the rapid cognitive ability of the intelligent body to the environment, and add L2 regularization layer to the value network to further improve the generalization ability of the algorithm, and finally manually set the intelligent body to keep the action constant in two consecutive frames, introduce a priori knowledge to constrain the search space and accelerate the convergence of the algorithm. Through CARLA open source simulation environment testing, the improved method significantly dominated the reward curve compared with the traditional method, and the success rates of three types of tasks, namely, straight ahead, turning, and designated route driving, increased by 10%, 16%, and 30%, respectively, proving that the proposed method is more effective.

Keywords: autonomous driving;reinforcement learning;mobile inverted bottleneck convolution;LSTM

0 引 言

随着机器学习等人工智能技术的蓬勃发展,许多传统领域中难以解决的问题都找到了新的突破方向。强化学习作为机器学习中一种特殊的学习方式,被认为是迈向通用人工智能的重要途径,因而逐渐受到青睐,随之涌现出各式各样的优秀算法。DeepMind 与 OpenAI 作为强化学习领域久负盛名的研究机构,在训练智能体玩 Atari 游戏、围棋博弈、星际争霸Ⅱ多智能体策略对抗、机械手臂控制、类人机器人行走、直升机特技表演、电厂控制等领域取得了令人兴奋的成就[1-3]。其中,自动驾驶技术深受以卷积神经网络为代表的深度学习以及不断发展的计算机视觉的影响,感知和综合决策等问题正在逐步得到解决。从谷歌公司的 Waymo、通用汽车公司的 Cruise、福特公司的 Argo 到百度公司的 Apollo 等,无一不表明自动驾驶技术正在向产业化落地迈进。

收稿日期:2022-08-31
* 基金项目:军委科技委预先研究项目(2016330ZD01200101)资助

自动驾驶行为决策系统的目标是对可能出现的驾驶的道路环境都决策出一个行为策略,满足实时性、合理性。加州大学伯克利分校团队分别提出基于无模型的强化学习框架[4]和基于模型的强化学习框架[5]学习在复杂、密集的城市环境中自动驾驶,前者提出自编码器对输入图像进行降维处理,后者采用基于模型的思想和使用高斯混合模型(gaussian mixed model,GMM)首先对系统动态进行近似,然后提出用双梯度下降法(dual gradient descent,DGD)来优化约束的策略优化问题;佛罗伦萨大学团队[6]提出一种基于模仿学习的自动驾驶系统,引入注意力机制,有助于车辆理解图像的哪一部分被认为是最重要的;瑞典哥德堡沃尔沃集团团队[7]提出贝叶斯强化学习估计 Q 值的分布,赋予智能体评估模型推荐的行动的能力,从而实现自动驾驶;吉林大学团队[8]提出最大边际逆向强化学习算法,通过专家经验知识设计奖励函数,提高决策策略的合理性。

不同于上述研究,本文重点解决从原始输入图像中学习驾驶策略时面临采样效率低、环境适应性差、决策效果不佳的问题。针对此,本文提出循环近端策略优化算法(recurrent proximal policy optimization,RPPO),并实现以 PPO 为基准算法作为比较。

1　自动驾驶行为决策

1.1　自动驾驶概况

自动驾驶汽车(self-driving car)也称为为无人车、无人驾驶汽车,是指装载某些设备的无人车辆能够依靠各种传感器实时采集的环境信息,自动完成分析与推理,调控动力装置,像驾驶员一样在道路上正常行驶[9]。

在车辆智能化的分级中,较为普世的是由国际汽车工程师协会(society of automotive engineer, SAE)制定的分级标准,如表 1 所示。目前大多数传统汽车制造商与新兴无人驾驶创业公司受限于传感设备与技术的限制,自动驾驶技术成熟度仍然处于 L4 级别以下。

表 1　SAE 自动驾驶系统分级机制

等级	名称	转向、加减速控制	对环境的观察	激烈驾驶的应对	应对工况
L0	人工驾驶	驾驶员	驾驶员	驾驶员	—
L1	辅助驾驶	驾驶员+系统	驾驶员	驾驶员	部分
L2	半自动驾驶	系统	驾驶员	驾驶员	部分
L3	自动驾驶	系统	系统	驾驶员	部分
L4	高度自动驾驶	系统	系统	系统	部分
L5	全自动驾驶	系统	系统	系统	全部

自动驾驶能够从根本上改变人们的出行方式和生活方式,体现在:提高道路交通安全、缓解城市交通拥堵、提升出行效率、降低驾驶者的门槛等。虽然目前自动驾驶遇到各种各样的技术难题以及其他约束,但是,随着近几年传感器融合技术、信号处理技术、人工智能技术等的突破性发展,自动驾驶在未来的 5～10 年必将掀起一场新的技术和市场革命[10]。

1.2　行为决策

行为决策模块是自动驾驶汽车的“大脑”,根据现代决策理论的发展历程,决策理论一般分为理性决策理论和行为决策理论,理性决策理论关注的是决策者采用一个最大效用或最优的决策方案,而行为决策理论更加关注的是决策者的认知过程以及决策的原因等,并逐渐受到人们的广泛关注[11]。

行为决策核心任务是基于各种传感器信息,统筹车辆自身状态以及当前环境状态完成车辆动力学计算,输出为具体的油门、刹车、方向盘调整参数[12]。随着数学建模与智能技术的不断发展,有很多优秀的行为决策方法被提出:有限状态机方法、决策树/行为树模型、基于知识/规则的推理决策方法、基于博弈论的行为决策方法、深度强化学习等。

2　深度强化学习与 PPO 算法

2.1　深度强化学习

强化学习在 20 世纪 80 年代已经兴起[13]。但随着问题的复杂度逐渐增大,传统的表格式强化学习已经难以解决庞大的状态空间和搜索空间。谷歌公司 DeepMind 团队创造性将深度卷积神经网络与传统的 Q 学习方法有机结合,提出端到端的 DQN 算法[14],并在 Atari 游戏上取得了引人注目的成绩[15]标志着深度强化学习时代的到来,此后,各式各样的深度强化学习算法被不断提出,Alpha GO、Alpha Zero、Alpha Star 的出现更是将深度强化学习推至顶峰[16]。

传统的强化学习和后来发展深度强化学习都是通过不断的学习迭代找到最优策略 π,使得智能体对于当前的状态 s_t,选择动作 a_{t+1} 后,得到的未来奖励 r 在一定时间内累积最大:

$$Q^*(s,a)=\max_{\pi}E[r_t+\gamma r_{t+1}\gamma^2 r_{t+2}+\cdots+|\ s_t=s,a_t=a,\pi] \tag{1}$$

式中:γ 为折扣因子,表示未来的奖励对当前状态的重要性[7]。

基于价值的强化学习算法中,最为常用的是一种离线

策略时序差分法：Q-learning，Q-learning 采用-贪心策略产生数据，并利用该数据对贪心策略评估和改进，以及通过查表的方法来对 Q 值进行预测。而经典的 DQN 算法对其进行了 3 点改进：使用深度神经网络从原始数据中提取特征，近似 Q 值；创造性的使用经验回放完成训练，降低训练数据相关性，提高训练效果；TD 偏差则是通过单独的目标网络来处理，参数更新公式如下：

$$\theta_{t+1}=\theta_t+\alpha(R_{t+1}+\gamma\max_a Q(s_{t+1},a;\theta_t^-)-Q(s_t,a_t;\theta_t))\nabla_{\theta_t}Q(s_t,a_t;\theta_t) \tag{2}$$

2.2　优势行动者评论家算法

行动者评论家算法（actor-critic，AC）是一种综合值函数逼近（Critic）和策略函数逼近（Actor）的方法，也是目前强化学习常用的范式。首先，Actor 针对当前的状态，通过运行策略函数选择一个行为；其次，Critic 采用值函数（或者优势函数）对 Actor 的行为进行评价；然后，基于 Critic 的评价，Actor 完成自己的策略（Actor 策略函数参数）调整；最后 Critic 根据环境给出的回报计算出来一个更新的目标值，来调整自己的评分策略（Critic 神经网络参数）。

优势行动者评论家算法（advantage actor critic，A2C）则是在传统的 AC 算法中引入基线方法来进一步减小方差，具体实现方法就是在策略梯度里面减去一个与状态相关，但与行为动作无关的基线函数 $B(s)$，这样的话，基线函数 $B(s)$ 既不会改变策略梯度，同时也会降低策略梯度方差，提高算法训练效果[17]。

最简单直接也最容易理解的一个基线函数 $B(s)$ 是基于当前智能体状态的值函数：

$$B(s)=V_{\pi_\theta}(s) \tag{3}$$

因此优势函数设计为：

$$A_{\pi_\theta}(s,a)=Q_{\pi_\theta}(s,a)-V_{\pi_\theta}(s) \tag{4}$$

目标函数设计为：

$$J^{\theta'}=E_{(s_t,a_t)\sim\pi_{\theta'}}\left[\frac{p_\theta(a_t\mid s_t)}{p_{\theta'}(a_t\mid s_t)}A^{\theta'}(s_t,a_t)\right] \tag{5}$$

2.3　近端策略优化算法

近端策略优化算法（proximal policy optimization，PPO）是一种基于 A2C 的在线策略（on-policy）算法，被成功用于连续空间和离散空间任务中，通过约束重要性采样（importance sampling）之后旧策略 θ 与新策略 θ' 之间的差别在一定范围，实现稳定有效的策略迭代[18]。

因此，对于给定长度为 T 的序列，优势函数设计为：

$$A_t=\sum_{T>t}\gamma^t r_t-V(s_t) \tag{6}$$

目标函数设计为：

$$J_{PPO}^{\theta'}\approx\sum_{(s_t,a_t)}\min(I_{\theta\theta'}A^{\theta'}(s_t,a_t),clip(I_{\theta\theta'},1-\varepsilon,1+\varepsilon)A^{\theta'}(s_t,a_t)) \tag{7}$$

其中：

$$I_{\theta\theta'}=\frac{p_\theta(a_t,s_t)}{p_{\theta'}(a_t,s_t)} \tag{8}$$

PPO 算法实现涉及 3 个神经网络（也可以为两个）：一个价值网络（Critic Network），两个策略网络（Old_Actor Network/New_Actor Network）。智能体首先按照 New_Actor Network 输出的策略采集一次交互数据（Episode），然后将 New_Actor Network 中的参数复制给 Old_Actor Network，根据优势函数以及目标函数进行 New_Actor Network 和 Critic Network 的学习，最后智能体通过不断的策略迭代，完成决策任务。

3　RPPO 算法实现细节

3.1　状态空间与动作空间

CARLA 是一个开源的自动驾驶模拟器，可以模拟真实的交通环境、行人行为、汽车传感器信号等[19]，因为其逼真的行驶环境模拟效果以及完备的编程接口，目前被应用于各种自动驾驶算法验证与落地研究。

本文以 CARLA 为验证环境开展端到端自动驾驶行为决策研究，状态空间为智能体车前传感器实时拍摄的分辨率为 640×480 的可见光图像集合，即：

$$S=\{O_{i_{640\times480}}\}_{i\in N} \tag{9}$$

本文通过控制 Carla 模拟器带的接口：油门（throttle）、方向盘信号（steer）、刹车（brake）来实现智能体的行为决策，如表 2 所示。

表 2　行为决策

动作	throttle	steer	brake	行为
a_1	1.0	0.0	0.0	直行
a_2	0.0	−1.0	0.0	左转
a_3	0.0	1.0	0.0	右转
a_4	1.0	−1.0	0.0	直行左转
a_5	1.0	1.0	0.0	直行右转
a_6	0.0	0.0	1.0	刹车
a_7	0.0	−1.0	1.0	刹车左转
a_8	0.0	1.0	1.0	刹车右转

因此，动作空间可表示为：

$$A=\{a_1,a_2,\cdots,a_7\} \tag{10}$$

3.2　网络设计

本文研究的端到端自动驾驶行为决策问题是将深度学习的高维数据拟合能力与强化学习的行为决策能力有机结合，实现“所见即所行”，考虑到自动驾驶训练环境复杂、关联变量较多以及行为决策时间要求较高的问题，采用移动翻转瓶颈卷积模块与 LSTM 构建策略网络与价值网络，有效整合前后帧的关联信息，实现智能体对多变情况的预测，提高智能体对环境的快速认知能力，并在价值网络添加 L2 正则化，进一步提高算法的泛化能力。

策略网络结构如表 3 所示。

表 3　策略网络

层序	算子	通道数
1	Conv3×3	64
2	MBConv1, k3×3	32
3	MBConv6, k3×3	48
4	MBConv6, k3×3	192
5	MBConv6, k5×5	384
6	MBConv6, k3×3	384
7	LSTM	256
8	LSTM	128
9	FC	7

价值网络结构如表 4 所示。

表 4　价值网络

层序	算子	通道数
1	Conv3×3	64
2	MBConv1, k3×3	32
3	MBConv6, k3×3	96
4	MBConv6, k3×3	192
5	MBConv6, k5×5	384
6	MBConv6, k3×3	384
7	FC(L2)	128
8	FC	1

LSTM 的更新方式为:Bootstrapped 序列更新,即在 RPPO 算法采集一次训练过程(Episode),从这次训练过程的开始一直学习到训练结束,在一次训练过程中,每一时刻 LSTM 隐含层的状态值从上一时刻继承而来。

另外,网络中 Conv3×3 作用是将输入图片转化为 MBConv 模块需要的输入维度,移动翻转瓶颈卷积(mobile inverted bottleneck convolution,MBConv)模块作用是提取特征图(feature map),并采用组合式的尺度优化方法可以使网络获得更好的感受野,主要由深度可分离卷积(depthwise convolution)、倒置残差块(inverted residual block)、压缩与激发网络(squeeze-and-excitation network,SENet)构成[20],通用结构如图 1 所示。

图 1 中,Swish 激活函数被证明在深层网络的效果优于 ReLU,公式如下:

$$f(x) = x \times sigmoid(\beta x) \tag{11}$$

式中:$\beta = 1$。

3.3　奖励函数设计

强化学习是一种模拟人类或动物多巴胺神经元兴奋的学习方法。人类或动物在执行某种时,如果得到意想不到的奖励的话,多巴胺神经元将处于兴奋状态导致当前行为趋势加强,反之则减弱,因此奖励的设定对于强化学习模型来说尤为重要,是驱动智能体学习到最优策略的关键。

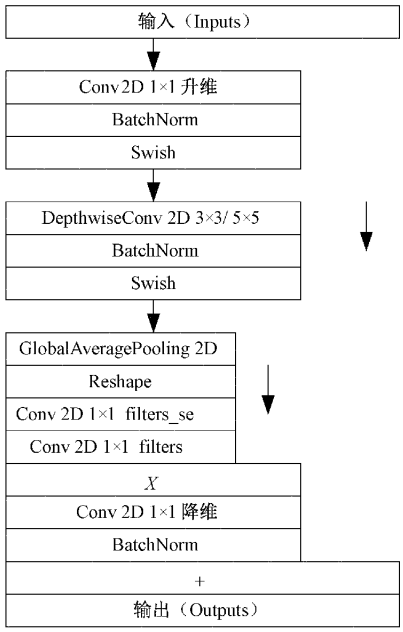


图 1　MBConv 通用结构示意图

结合文献[5]以及经过多次对比实验,本文设计了一种对于自动驾驶有效的复合奖励函数:

$$r = r_l + r_v + r_c + r_a \tag{12}$$

式中:r_l 是对车辆驶出道路的惩罚项,车辆驶出 1 m 时 $r_l = -5$,其他 $r_l = 1$;r_v 是对汽车速度的惩罚项,当车辆速度大于最大速度 v_{max} 以及小于最小速度 v_{min} 时,$r_v = -1$,其他 $r_v = (v - v_{max})/(v - v_{min})$;$r_c$ 是对车辆发生碰撞的惩罚项,当发生碰撞时,$r_c = -10$,其他 $r_c = 3$;r_a 是对车辆转向角度流畅度的惩罚项 $r_a = 0.2 \times \alpha^2$。

3.4　帧数设置

考虑到 CARLA 实验环境复杂多变,本文手动设置智能体在 2 个连续帧中保持动作不变,也就是说,智能体在训练过程中所做的每个动作将持续 2 帧,直到新的动作开始,这个技巧会大大降低训练的复杂度,甚至可以把它看作是将搜索深度降低了 2 倍,但是,该值也不能过大,如果值太大,正确的动作空间可能太小甚至不存在。

4　实验与结果

4.1　实验环境

本文在尽量考虑模拟真实性的前提下采用场景复杂的 Carla 仿真环境,因此需要足够的算力支撑:GPU NVIDIA GeForce RTX 2080Ti,显存 12 G,CPU Intel(R) Core(TM) i9-10900X CPU@3.70 GHz,内存 64 G,深度学习框架 Keras2.1.4 和 TensorFlow1.15.0,Python3.7.8。

4.2　训练参数

为保证实验结果的有效性,基准算法同样采用相同的移动翻转瓶颈卷积模块,且与本文提出的 RPPO 算法皆采用相同的训练参数。训练参数如表 5 所示。

表 5　训练参数

参数	值
学习率	0.000 25
更新步 D	5 000
批次 B	128
折扣因子 γ	0.95
初始 ε_0	0.99
NPC 数量	20

本文通过奖励曲线以及直行、转弯、指定路线行驶 3 类任务的成功率来对两种算法进行比较：

4.3　奖励曲线

由图 2 的奖励曲线可以看出，基准算法与本文提出的 RPPO 算法在初始阶段差距不大，但随着训练的进行，在大约 30 000 Episode 左右，RPPO 开始明显超过基准算法，而且差距越来越大。

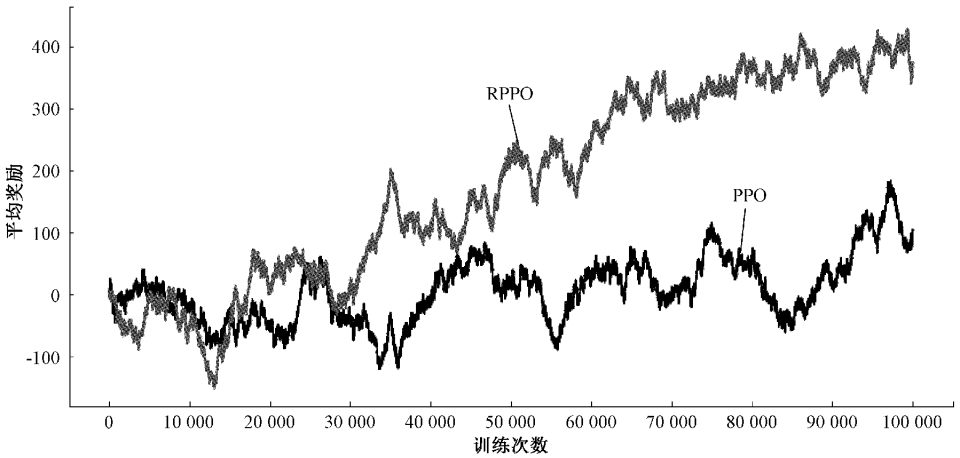
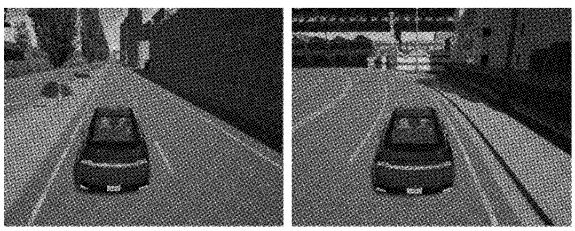


图 2　奖励曲线

4.4　直行、转弯任务成功率

图 3 为分别采用基准算法与 RPPO 算法的自动驾驶车辆执行直行与转弯两种任务场景举例。



（a）直行　　（b）转弯

图 3　直行、转弯举例

直行以及转弯任务分别选取 40 种不同的场景，每种场景测试 5 次，且要求直行距离不小于 300 m，转弯半径不大于 20 m，结果如表 6 所示。

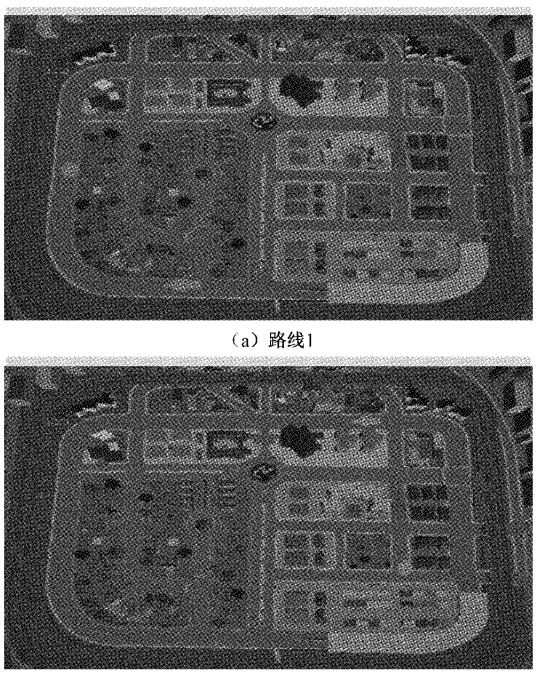
表 6　直行、转弯成功率　　%

算法	直行	转弯
PPO	85	74
RPPO	95	90

由表 6 可以看出，在执行直行以及转弯两类较为简单的任务时，RPPO 比基准算法更优秀，直行任务成功率提高了 10%，转弯任务成功率提高了 16%，考虑到直行任务输入图像变化较为简单，且只调节油门与刹车即可完成，因此差距并不大。但是转弯任务场景复杂，油门、刹车与方向盘都需要随时协调，因此差距较为明显。

4.5　指定路线行驶成功率

图 4 为分别采用基准算法与 RPPO 算法的自动驾驶车辆执行指定路线行驶任务场景举例。



（a）路线1

（b）路线2

图 4　路线举例

本文选取 20 条不同路线进行测试且每条路线分别测试 5 次。图 4 中,五角星代表目的地,箭头代表起始点以及行驶方向,结果如表 7 所示。

表 7　指定路线成功率　%

算法	指定路线
PPO	50
RPPO	85

由表 7 可以看出,在指定路线行驶的测试中,基准算法明显低于 RPPO 算法,考虑到是因为指定路线行驶综合了直行、转弯、刹车避让等一系列行为决策,采用 RPPO 算法确实可以有效提高车辆对环境的认知能力以及学习能力。另外,本文通过测试回放发现:基准算法在类似路线 2(需要变道)的情况下效果很差;本文提出的 RPPO 算法在处理环岛路况时效果欠佳。

5　结　　论

本文提出的循环近端策略优化算法,通过模拟环境测试明显优于基准算法,也说明采用 LSTM 与移动翻转瓶颈卷积模块构建策略网络与价值网络,并在价值网络添加 L2 正则化层,以及引入先验知识约束搜索空间可以有效缓解传统的 DRL 算法在处理类似自动驾驶行为决策这种庞大的搜索空间面临采样效率低、环境适应性差、决策效果不佳的问题。在未来的工作中,将综合考虑环岛路况效果欠佳进行 RPPO 算法改进,并结合其他相关技术尝试部署无人战车在某些区域实现智能行为决策以至自动驾驶。

参考文献

[1] SILVER D, HUANG A, MADDISON C J, et al. Mastering the game of Go with deep neural networks and tree search[J]. Nature, 2016, 529(7587): 484-489.

[2] SILVER D, SCHRITTWIESER J, SIMONYAN K, et al. Mastering the game of go without human knowledge[J]. Nature, 2017, 550(7676): 354-359.

[3] VINYALS O, BABUSCHKIN I, CZARNECKI W M, et al. Grandmaster level in StarCraft II using multi-agent reinforcement learning [J]. Nature, 2019, 575(7782): 350-354.

[4] MA H, CHEN J, EBEN S, et al. Model-based constrained reinforcement learning using generalized control barrier function [C]. 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems(IROS), IEEE, 2021: 4552-4559.

[5] CHEN J, YUAN B, TOMIZUKA M. Model-free deep reinforcement learning for urban autonomous driving [C]. 2019 IEEE Intelligent Transportation Systems Conference(ITSC), IEEE, 2019: 2765-2771.

[6] CULTRERA L, SEIDENARI L, BECATTINI F, et al. Explaining autonomous driving by learning end-to-end visual attention [C]. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020: 340-341.

[7] HOEL C J, TRAM T, SJÖBERG J. Reinforcement learning with uncertainty estimation for tactical decision-making in intersections[C]. 2020 IEEE 23rd International Conference on Intelligent Transportation Systems(ITSC), IEEE, 2020: 1-7.

[8] 高振海,闫相同,高菲. 基于逆向强化学习的纵向自动驾驶决策方法[J]. 汽车工程,2022,44(7):969-975.

[9] HAN S J, KYUNG-BOK S, MIN K W, et al. Apparatus and method for providing location and heading information of autonomous driving vehicle on road within housing complex: US20150142248[P]. 2015-5-21.

[10] 《NI 趋势展望报告 2019》探索了物联网、5 G 商业化部署以及大众自动驾驶领域等大趋势[J]. 电子测量技术,2018,41(22):81.

[11] 王丙琛,司怀伟,谭国真. 基于深度强化学习的自动驾驶车控制算法研究[J]. 郑州大学学报(工学版),2020,41(4):41-45,80.

[12] 冀杰,黄岩军,李云伍,等. 基于有限状态机的车辆自动驾驶行为决策分析[J]. 汽车技术,2018(12):1-7,DOI:10.19620/j.cnki.1000-3703.20172426.

[13] SUTTON R, BARTO A. Reinforcement learning: An introduction[M]. Massachusetts: MIT Press, 1998.

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

[15] HAUSKNECHT M, STONE P. The impact of determinism on learning atari 2600 games [C]. Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence, 2015.

[16] SILVER D, SCHRITTWIESER J, SIMONYAN K, et al. Mastering the game of go without human knowledge[J]. Nature, 2017, 550(7676): 354-359.

[17] MNIH V, BADIA A P, MIRZA M, et al. Asynchronous methods for deep reinforcement learning [C]. International Conference on Machine Learning, 2016: 1928-1937.

[18] WANG Y, HE H, TAN X. Truly proximal policy optimization[C]. Uncertainty in Artificial Intelligence, PMLR, 2020: 113-122.

[19] DOSOVITSKIY A, ROS G, CODEVILLA F, et al. CARLA: An open urban driving simulator [C]. Conference on Robot Learning. PMLR, 2017: 1-16.

[20] TAN M, LE Q. Efficientnet: Rethinking model scaling for convolutional neural networks [C]. International Conference on Machine Learning. PMLR, 2019: 6105-6114.

作者简介

姚悦(通信作者),硕士,主要研究方向为机器学习、智能决策、软件工程。
E-mail: 706632509@qq.com

吉明佳,硕士,主要研究方向为软件工程。
E-mail: yyx19950908@163.com

杨霄,硕士,主要研究方向为机器学习。
E-mail: 1554810358@qq.com