E-mail: jig@radi.ac.cn
Website: www.cjig.cn
Tel: 010-58887035
中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS

中图法分类号: TP391　　**文献标识码:** A　　**文章编号:** 1006-8961(2021)01-0028-08

论文引用格式: Pan F and Bao H. 2021. Research progress of automatic driving control technology based on reinforcement learning. Journal of Image and Graphics,26(01):0028-0035(潘峰,鲍泓. 2021. 强化学习的自动驾驶控制技术研究进展. 中国图象图形学报,26(01):0028-0035) [DOI: 10.11834/jig.200428]

强化学习的自动驾驶控制技术研究进展

潘峰[1,2],鲍泓[2]

1. 北京化工大学,北京　100029; 2. 北京联合大学,北京　100101

摘　要:　自动驾驶车辆的本质是轮式移动机器人,是一个集模式识别、环境感知、规划决策和智能控制等功能于一体的综合系统。人工智能和机器学习领域的进步极大推动了自动驾驶技术的发展。当前主流的机器学习方法分为:监督学习、非监督学习和强化学习3种。强化学习方法更适用于复杂交通场景下自动驾驶系统决策和控制的智能处理,有利于提高自动驾驶的舒适性和安全性。深度学习和强化学习相结合产生的深度强化学习方法成为机器学习领域中的热门研究方向。首先对自动驾驶技术、强化学习方法以及自动驾驶控制架构进行简要介绍,并阐述了强化学习方法的基本原理和研究现状。随后重点阐述了强化学习方法在自动驾驶控制领域的研究历史和现状,并结合北京联合大学智能车研究团队的研究和测试工作介绍了典型的基于强化学习的自动驾驶控制技术应用,讨论了深度强化学习的潜力。最后提出了强化学习方法在自动驾驶控制领域研究和应用时遇到的困难和挑战,包括真实环境下自动驾驶安全性、多智能体强化学习和符合人类驾驶特性的奖励函数设计等。研究有助于深入了解强化学习方法在自动驾驶控制方面的优势和局限性,在应用中也可作为自动驾驶控制系统的设计参考。

关键词: 自动驾驶; 决策控制; 马尔可夫决策过程; 强化学习; 数据驱动; 自主学习

Research progress of automatic driving control technology based on reinforcement learning

Pan Feng[1,2], Bao Hong[2]

1. Beijing University of Chemical Technology, Beijing　100029, China; 2. Beijing Union University, Beijing　100101, China

Abstract: Research on fully automatic driving has been largely spurred by some important international challenges and competitions, such as the well-known Defense Advanced Research Projects Agency Grand Challenge held in 2005. Self-driving cars and autonomous vehicles have migrated from laboratory development and testing conditions to driving on public roads. Self-driving cars are autonomous decision-making systems that process streams of observations coming from different on-board sources, such as cameras, radars, lidars, ultrasonic sensors, global positioning system units, and/or inertial sensors. The development of autonomous vehicles offers a decrease in road accidents and traffic congestions. Most driving scenarios can be simply solved with classical perception, path planning, and motion control methods. However, the remaining unsolved scenarios are corner cases where traditional methods fail. In the past decade, advances in the field of artificial intelligence (AI) and machine learning (ML) have greatly promoted the development of autonomous driving. Autonomous

收稿日期: 2020-07-30; **修回日期:** 2020-10-23; **预印本日期:** 2020-10-30

基金项目: 国家自然科学基金项目(61932012);北京市教委科技计划一般项目(KM201911417009)

Supported by: National Natural Science Foundation of China(61932012); Beijing Municipal Education Commission Science and Technology Plan General Project (KM201911417009)

driving is a challenging application domain for ML. ML methods can be divided into supervised learning, unsupervised learning, and reinforcement learning (RL). RL is a family of algorithms that allow agents to learn how to act in different situations. In other words, a map or a policy is established from situations (states) to actions to maximize a numerical reward signal. Most autonomous vehicles have a modular hierarchical structure and can be divided into four components or four layers, namely, perception, decision making, control, and actuator. RL is suitable for decision making and control in complex traffic scenarios to improve the safety and comfort of autonomous driving. Traditional controllers utilize an a priori model composed of fixed parameters. When robots or other autonomous systems are used in complex environments, such as driving, traditional controllers cannot foresee every possible situation that the system has to cope with. An RL controller is a learning controller and uses training information to learn their models over time. With every gathered batch of training data, the approximation of the true system model becomes accurate. Deep neural networks have been applied as function approximators for RL agents, thereby allowing agents to generalize knowledge to new unseen situations, along with new algorithms for problems with continuous state and action spaces. This paper mainly introduces the current status and progress of the application of RL methods in autonomous driving control. This paper consists of five sections. The first section introduces the background of autonomous driving and some basic knowledge about ML and RL. The second section briefly describes the architecture of autonomous driving framework. The control layer is an important part of an autonomous vehicle and has always been a key area of autonomous driving technology research. The control system of autonomous driving mainly includes lateral control and longitudinal control, namely, steering control and velocity control. Lateral control deals with the path tracking problem, and longitudinal control deals with the problem of tracking the reference speed and keeping a safe distance from the preceding vehicle. The third section introduces the basic principles of RL methods and focuses on the current research status of RL in autonomous driving control. RL algorithms are based on Markov decision process and aim to learn mapping from situations to actions to maximize a scalar reward or reinforcement signal. RL is a new and extremely old topic in AI. It gradually became an active and identifiable area of ML in 1980 s. Q-learning is a widely used RL algorithm. However, it is based on tabular setting and can only deal with those problems with low dimension and discrete state/action spaces. A primary goal of AI is to solve complex tasks from unprocessed, high-dimensional, sensory input. Significant progress has been made by combining deep learning for sensory processing with RL, resulting in the "deep Q network" (DQN) algorithm that is capable of human-level performance on many Atari video games using unprocessed pixels for input. However, DQN can only handle discrete and low-dimensional action spaces. Deep deterministic policy gradient was proposed to handle those problems with continuous state/action spaces. It can learn policies directly from raw pixel inputs. The fourth section generalizes some typical applications of RL algorithm in autonomous driving, including some studies of our team. Unlike supervised learning, RL is more suitable for decision making and control of autonomous driving. Most of the RL algorithms used in autonomous driving mostly combine deep learning and use raw pixels as input to achieve end-to-end control. The last section discusses the challenges encountered in the application of RL algorithms in autonomous driving control. The first challenge is how to deploy the RL model trained on a simulator to run in a real environment and ensure safety. The second challenge is the RL problem in an environment with multiple participants. Multiagent RL is a direction of RL development, but training multiagents is more complicated than training a single agent. The third challenge is how to train an agent with a reasonable reward function. In most RL settings, we typically assume that a reward function is given, but this is not always the case. Imitation learning and reverse RL provide an effective solution for obtaining the real reward function that makes the performance of the agent close to a human. This article helps to understand the advantages and limitations of RL methods in autonomous driving control, the potential of deep RL, and can serve as reference for the design of automatic driving control systems.

Key words: autonomous driving; decision control; Markov decision process (MDP); reinforcement learning (RL); data-driven; autonomous learning

0 引 言

自动驾驶车辆是一个集环境感知、决策规划和智能控制等功能于一体的综合系统，是智能交通系统的重要组成部分，也是智能车辆领域研究的热点和汽车工业增长的新动力(徐友春 等，2001)。自动驾驶汽车的控制技术是整个自动驾驶系统中的关键环节，也是国内外广大学者重点研究的领域。自动驾驶系统一般采用分层结构，其中控制层的功能是将来自决策和规划层的指令转化为各执行机构的动作，并控制各执行机构完成相应的动作，以此准确地跟踪路径并合理地控制速度。

自动驾驶车辆的控制可分为横向控制和纵向控制，传统的横/纵向控制的方法大多需要精确的数学解析模型，并对受控系统进行精确的数值求解。然而精度较高的模型一般也比较复杂，参数较多。复杂的模型也造成了较高的计算代价，使得求解困难，往往难以保证实时性。随着互联网+、大数据和人工智能的迅速发展，研究人员开始基于机器学习方法开发智能汽车决策和控制算法，开辟了一条不同于汽车工程专家的研究思路。

机器学习主要研究计算机如何通过经验或探索环境来获取知识或优化自身技能，这是当前发展最快的一个技术领域。越来越多基于机器学习的方法被应用到自动驾驶系统中来。李德毅院士认为基于自学习的“驾驶脑”是中国智能车实现对国外弯道超车的关键所在(李德毅，2015)。2019 年，专业研发自动驾驶的公司 Waymo 收购了专门研究模仿学习在自动驾驶中应用的 Latent Logic 公司，这意味着 Waymo 将在机器学习在自动驾驶中的应用领域展开更加深入的研究和开发。目前，国内的各大 IT 厂商也纷纷开展与传统汽车厂家的合作，共同开发智能汽车。百度公司自 2014 年启动“百度自动驾驶汽车”研发计划以来，已经推出了 Apollo 自动驾驶系统，并于 2016 年取得了加州的自动驾驶牌照。百度还将自动驾驶汽车结合百度大脑，通过人工智能技术进一步推动自动驾驶汽车的进步。

机器学习的一个主要类型是强化学习(reinforcement learning，RL)(Kaelbling 等，1996; Bartlett，2003; Konda 和 Tsitsiklis，2003; Sutton 和 Barto，1998; Sutton，1992; Lillicrap 等，2015; Mnih 等，2015; Silver 等，2017)。与监督式学习主要应用于自动驾驶的感知层不同，强化学习更多应用在决策和控制层。传统控制器一般利用由固定参数组成的先验模型，当机器人用于复杂环境(例如驾驶)时，传统控制器无法预见系统必须应对的所有可能情况，而学习型控制器会利用训练信息来逐步学习其模型(Ostafew 等，2016)。机器学习还可以和传统控制方法相结合，如学习模型预测控制(model predictive control，MPC)的代价函数，使人们更好地预测车辆的干扰和行为(Ostafew，2016)。由于自动驾驶控制问题具有高维度、状态和动作空间连续、非线性等特点，深度学习虽然具有较强的感知能力，然而却不擅长决策和控制。强化学习则可以通过不断探索环境来学习复杂的控制模型。因此，将两者相结合的深度强化学习(deep reinforcement learning，DRL)可以形成优势互补，为解决复杂系统的感知决策问题提供了新的思路。DRL 可以实现端到端(end-to-end)的感知与控制，具有很强的通用性。DRL 将深度学习的感知能力和强化学习的决策控制能力相结合，可以直接根据输入的像素级别的图像(或雷达数据)进行控制，更接近人类的思维方式。深度学习和强化学习的结合使得自动驾驶控制问题得到了更多的解决方案。

1 自动驾驶控制技术

自动驾驶车辆系统大多采用分层架构(Paden 等，2016)，即感知层、决策层(含运动规划)、控制层和车辆线控层 4 个层次，如图 1 所示。无人驾驶车辆的控制层作为决策层和执行器之间的中间层，也是最重要的一个环节，一直是无人驾驶技术研究的重点领域之一。无人驾驶的控制技术主要包括横向控制和纵向控制(Rajamani，2012)。横向控制指的是通过执行合适的转向运动引导车辆沿一个全局的几何路径行驶。路径跟踪控制器的目标就是最小化车辆和路径之间的横向距离以及车辆方向和路径方向的偏差，约束转向输入来平滑运动以维持稳定性(Snider，2009)。纵向控制则是根据道路形状，在满足车辆运动学约束、动力学约束和安全车距的前提下，计算出期望的速度和加速度，并控制油门和制动系统加以实现。

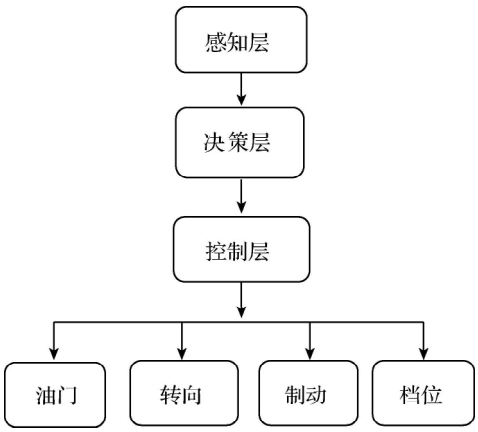

图 1 通用无人驾驶系统框图

Fig. 1 Block diagram of autonomous driving system

2 强化学习方法

强化学习是一种基于马尔可夫决策过程的算法(Markov decision process, MDP)(Abakuks,1987)。智能体在完成某项任务时,通过动作与环境进行交互,并产生新的状态,同时从环境得到回报。如此循环,智能体与环境交互过程中不断产生训练数据。强化学习算法正是在不断探索环境的过程中利用产生的数据进行迭代优化其行为策略。经过数次迭代学习后,智能体能最终学到完成相应任务的最优策略。如图 2 所示。

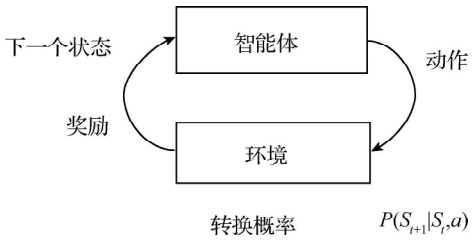

图 2 强化学习模型

Fig. 2 Reinforcement learning model

强化学习在心理学中已经研究了将近一个世纪,这项工作对人工智能产生了非常强烈的影响。可以把强化学习看做是某些心理学习过程的逆向工程。Samuel(1959)最早使用类似今天的时间差分优化方法来处理延迟的奖励,这被认为是最早的和强化学习相关的研究。强化学习最早由 Minsky(1961)提出,Waltz 和 Fu(1965)将其独立引入控制理论中。到 20 世纪 80 年代,Barto 和 Sutton(1981)的研究成果使得强化学习逐渐成为机器学习研究的热门领域(Barto 等,1970; Barto 和 Sutton,1981)。

一个马尔可夫决策过程可用一个五元组(S,A,P,R,γ)来表示。其中,S 为状态空间,A 为动作空间,P 为转换概率,R 为奖励,γ 为奖励的衰减系数。强化学习的目标即最大化期望的累计奖励,所以强化学习最终也是个优化问题,其目标函数为

$$J^*(\theta) = \max_{\theta} \mathrm{E}\{\sum_{k=0}^{H} \gamma^k r_k\} \tag{1}$$

式中,θ 为待优化的参数。

可以认为 Q-learning(Watldns,1989)是目前使用最广泛的强化学习算法。Watkins 和 Dayan(1992)首次完整证明了 Q-learning 的收敛性。然而 Q-learning 一般基于表格设置,只能用于解决较低维度、离散的状态/动作空间的强化学习问题。Mnih 等人(2015)将深度神经网络引入 Q-learning,诞生了 DQN(deep Q network)算法。深度学习具有强大的特征提取能力,这使得强化学习具备了直接从高维像素级别的状态空间提取特征并进行训练的能力。引入了深度学习以后的强化学习称为深度强化学习,其中深度神经网络可以作为强化学习中的策略函数、值函数或 Q 函数的逼近器。DQN 算法在很多电子游戏中达到了人类专家级别的表现。DQN 虽然解决了高维度状态空间的问题,但是仍然只能处理离散的、低维度动作空间的问题。然而实际的控制任务,尤其是自动驾驶任务,往往具有连续的、高维度的动作空间。解决方法之一是将动作空间离散化。然而,动作的数目随着自由度的增加而呈指数增长,而 DQN 难以在巨大的动作空间下进行有效的探索和训练。Lillicrap 等人(2015)提出了深度确定性策略梯度(deep deterministic policy gradient, DDPG)算法,成功实现了连续状态和动作空间下的强化学习问题。该算法将 DQN 和 Actor-Critic 算法相结合,使用深度神经网络来逼近值函数和策略函数,其中值函数通过贝尔曼方程更新,而策略函数则通过梯度下降进行更新。然而 DDPG 算法具有较高的参数脆性,其超参数往往必须针对不同的问题进行仔细设置才能获得良好的训练结果。2016 年在深度强化学习领域最重要的事件就是谷歌的 DeepMind 团队开发的 AlphaGo 程序以 4∶1 的成绩战胜了韩国职业围棋选手李世石,取得了世界的瞩目(Silver 等,2016)。AlphaGo 结合监督式学习和强

化学习两种方法，首先使用监督式学习对策略网络进行初始化，然后通过强化学习迭代更新价值网络和策略网络，搜索则是根据蒙特卡洛搜索树进行。然而 AlphaGo 仍然在一定程度上依赖人类棋局的先验知识。为了改进这一点，Silver 等人在 2017 年研发了 AlphaZero，击败了之前的所有围棋算法，而且在国际象棋和日本将棋方面也取得了最优成绩（Silver 等，2017）。AlphaZero 完全依赖于蒙特卡洛搜索树和深度残差网络（He 等，2016），而不需要任何人类棋谱等先验知识。Haarnoja 等人（2018）为了解决强化学习的高样本复杂性和超参数的脆性，并提高训练的稳定性，提出了 SAC（soft actor-critic）方法。该方法将最大熵的概念引入强化学习，并超过了 DDPG 方法的效率和最终性能。

3 强化学习在自动驾驶控制中的应用

自动驾驶车辆的控制大多基于规则设计，而复杂场景中规则数目会呈指数级上升，而且规则之间也可能发生冲突。自动驾驶车辆在复杂工况下的测试、验证也往往出于安全考虑难以开展。为了应对复杂的交通场景，需要控制算法通过数据驱动或与环境交互进行自主学习，充分训练后能够自主应对复杂工况。强化学习正是这样一种数据驱动的自主学习方法。和监督学习不同，强化学习更适用于自动驾驶的决策和控制。自动驾驶任务可视为一个部分可观察的马尔可夫决策过程（partially observable Markov decision process，POMDP）。在 POMDP 中，智能体（自动驾驶汽车）通过传感器观察环境得到观测值 I_t，然后在状态 s_t 下采取一个动作 a_t，随后在环境中得到奖励 R_{t+1}，并转换到下一个状态 s_{t+1}。

国际上最早将深度强化学习应用在车辆控制领域为 Lange 等人（2012）使用 DQN 方法在微型赛车模拟器下进行训练并取得了良好的效果，其控制水平甚至超出了人类玩家。然而该方法仍停留在模拟器仿真阶段，其实时性难以达到实际应用的要求，而且只能应用于离散的低维动作空间。2016 年，Sallab 等人（2016）使用深度强化学习方法在开源赛车模拟器（the open racing car simulator，TORCS）上实现了车道保持控制，并对比了离散空间的 DQN 方法和连续动作空间的 DDAC 方法，证明了 DDAC 方法能够得到很好的控制效果和平滑的运行轨迹（Sallab 等，2016）。由于引入了深度学习方法，Sallab 等人（2016）提出了端到端深度强化学习的思想。归功于深度神经网络强大的特征提取能力，再结合强化学习的方法对智能体加以训练，可以直接将原始图像映射为执行器的输出，而且在鲁棒性上超过了简单的监督学习型的端到端控制。深度强化学习也逐渐代替了传统的强化学习方法。Chae 等人（2018）使用 DQN 算法训练智能体学习处理行人横穿马路的场景，实现了车辆的自主制动控制。Zong 等人（2018）使用 DDPG 算法对智能体的加速度和转向控制进行训练以实现自主避障，并在 TORCS 环境中进行了测试。Shalev-Shwartz 等人（2016）使用强化学习结合长短期记忆网络（long short term memory networks，LSTM）算法在游戏环境中解决自动驾驶的纵向控制以及汇入环岛的控制问题。吉林大学杨顺使用深度学习结合 DDPG 算法提出了基于视觉场景理解的深度强化学习控制方法（杨顺，2019）。随着强化学习在自动驾驶中的应用研究逐渐升温，为了提高强化学习的训练效率，微软公司在 2018 年提出了分布式云端深度强化学习的框架，大大缩减了训练的时间（Spryn 等，2018）。参考人类学习的过程，卡内基梅隆大学的 Liang 等人（2018）将模仿学习和强化学习相结合，提出了可控模仿强化学习的方法，并在开源模拟器中取得了良好的控制效果。这种方法先通过模仿学习对控制网络的权重进行初始化，然后通过 DDPG 方法进行强化训练。这样不但可以解决 DDPG 对超参数敏感的问题，而且比单独的模仿学习能够更好地适应复杂环境。北京联合大学的韩向敏等人（2018）使用 DDPG 算法实现了自动驾驶的纵向自动控制，而且使智能车辆可以在自学习过程中完成自适应巡航并不断改进，结果达到了人类驾驶员的控制水平。

4 强化学习在自动驾驶中面临的挑战

自动驾驶实验具有极大的危险性，所以当前的强化学习模型大多使用视频游戏模拟引擎进行训练和仿真，如 TORCS、侠盗猎车 5（grand theft auto V，GAT5）等。然而真实环境和虚拟环境之间存在较大的差异，往往只能采用数据集验证或离线数据回放等方式来验证模型的稳定性和鲁棒性，而基于模拟器的训练往往因为存在建模误差而导致将训练好的

模型迁移到真实环境中时可靠性不佳的问题。生成对抗网络(generative adversarial networks,GAN)的出现提供了解决这一问题的思路。Pinto 等人(2017)将对抗学习和强化学习结合提出了抗强化学习,用于训练智能体应对干扰以提高模型的鲁棒性。卡内基梅隆大学的 Yang 等人(2018)提出从虚拟到现实的端到端控制框架 DU-drive,使用类似条件生成对抗网络(conditional generative adversarial network,conditional GAN)的框架进行图像解析并控制车辆运动。美国弗吉尼亚理工大学电气与计算机工程系的 Ferdowsi 等人(2018)针对自动驾驶系统的"安全性"问题,提出了一种对抗深度强化学习框架,以解决自动驾驶汽车的安全性问题。

多智能体强化学习(Busoniu 等,2006)也是目前强化学习发展的一个方向。在真实的交通环境中,交通的参与者并非只有一个,驾驶者的决策和控制往往是多个交通参与者互相博弈的结果。强化学习是基于马尔可夫决策过程的理论,然而,很多强化学习算法只是对马尔可夫过程的近似。例如在自动驾驶应用中,状态的转换并不一定只依赖于智能体采取的动作,也包括环境中其他参与者采取的动作。多智能体强化学习正是为了解决这一问题,如 minimax-Q learning(Littman,1994),Nash-Q learning(Hu 和 Wellman,2003)等方法。不可否认的是,多智能体训练比单智能体更加复杂。

强化学习中奖励函数的作用就是引导智能体不断优化其策略以获得期待的未来累积的最大化奖励。大部分强化学习范例中的奖励函数通常是由系统设计人员手动编码。对于某些强化学习问题,通常可以找到一些明显的奖励函数,如游戏中的得分、财务问题中的利润等。但是对于某些实际应用中的强化学习问题,其奖励函数不但是未知的,而且需要权衡很多不同方面的需求。如果奖励函数设置不合理,则智能体就有可能收敛到错误的方向或者学到的是次优的策略。在自动驾驶应用中,奖励函数的设定不但要考虑到安全性和舒适性,还需要考虑如何让智能体更加符合人类驾驶员的驾驶习惯。然而,人类驾驶的控制行为比较复杂,在驾驶过程中需要权衡多方面的需求和约束,所以难以手动指定一个合理的奖励函数来引导智能体训练。而一个不合理的奖励函数会造成训练好的模型收敛到局部最小值甚至出现糟糕的表现。北京联合大学的智能驾驶团队对驾驶数据进行分析得到人类驾驶员的特征,并设计强化学习的奖励函数实现无人驾驶的纵向控制,使得智能体在纵向控制方面更加符合人类驾驶习惯(Pan 和 Bao,2019)。模仿学习和反向强化学习对于真实奖励函数的获取以及如何让智能体的表现更加接近人类提供了一个有效的解决方案,并成为无人驾驶中的另一个研究热点(Amit 和 Matari,2002; Abbeel 和 Ng,2004)。

5 结 语

自动驾驶技术是世界车辆工程领域研究的热点和汽车工业增长的新动力,也是目前各国重点发展的智能交通系统中一个重要组成部分。自动驾驶车辆的控制系统作为车辆行为层的关键环节对于车辆行驶的安全性和舒适性至关重要。传统的控制方法大多基于精确的数学解析模型或者基于规则设计。真实交通环境中复杂多变的交通场景使得难以设计精确的数学模型,而规则的数目也会随着交通场景复杂程度呈指数增长。强化学习的出现使得设计以数据驱动或与环境交互进行自主学习的控制系统成为可能,经过充分训练的学习型控制器也能够更好地应对复杂工况。本文对强化学习在自动驾驶控制领域的应用进行了充分调研和实验,并在团队研发的"京龙"和"联合彩虹"无人驾驶智能车和仿真系统上进行了测试和比赛,取得了较好的效果和成绩。强化学习目前已经成为自动驾驶控制中的热门研究领域并展示出光明的应用前景。然而,强化学习在自动驾驶中的应用仍然面临诸多挑战,也是未来这一领域进一步的研究方向,包括在真实交通环境中的部署和测试问题,在多个交通参与者环境下的多智能体强化学习问题以及针对人类驾驶员特性的奖励函数的塑造问题。

致 谢 感谢北京联合大学机器人学院对"京龙"和"联合彩虹"智能车团队的支持和帮助,感谢李德毅院士对自动驾驶研究的指导!

参考文献(References)

Abakuks A. 1987. Reviewed work: Markov Processes: characterization and convergence. By S. N. Ethier, T. G. Kurtz. Biometrics, 43(2): 113-122 [DOI: 10.2307/2531839]

Abbeel P and Ng A Y. 2004. Apprenticeship learning via inverse reinforcement learning//Proceedings of the 21st International Conference on Machine Learning. Banff, Canada: ACM: 1-8 [DOI: 10.1007/978-0-387-30164-8_417]

Amit R and Matari M. 2002. Learning movement sequences from demonstration//Proceedings of the 2nd International Conference on Development and Learning. Cambridge, USA: IEEE: 203-208 [DOI: 10.1109/DEVLRN.2002.1011867]

Bartlett P L. 2003. An introduction to reinforcement learning theory: value function methods//Advanced Lectures on Machine Learning, Machine Learning Summer School 2002. Canberra, Australia: Springer: 184-202 [DOI: 10.1007/3-540-36434-X_5]

Barto A G and Sutton R S. 1981. Landmark learning: an illustration of associative search. Biological Cybernetics, 42(1): 1-8 [DOI: 10.1007/BF00335152]

Barto A G, Sutton R S and Anderson C W. 1970. Neuronlike adaptive elements that can solve difficult learning control problems. IEEE Transactions on Systems, Man, and Cybernetics, SMC-13(5): 834-846 [DOI: 10.1109/TSMC.1983.6313077]

Busoniu L, Babuska R and de Schutter B. 2006. Multi-agent reinforcement learning: a survey//Proceedings of the 9th International Conference on Control, Automation, Robotics and Vision. Singapore, Singapore: IEEE: 1-6 [DOI: 10.1109/ICARCV.2006.345353]

Chae H, Kang C M, Kim B D, Kim J, Chung C C and Choi J W. 2018. Autonomous braking system via deep reinforcement learning//Proceedings of the 20th IEEE International Conference on Intelligent Transportation Systems (ITSC). Yokohama, Japan: IEEE: 1-6 [DOI: 10.1109/ITSC.2017.8317839]

Ferdowsi A, Challita U, Saad W and Mandayam N B. 2018. Robust deep reinforcement learning for security and safety in autonomous vehicle systems//Proceedings of the 21st International Conference on Intelligent Transportation Systems (ITSC). Maui, USA: IEEE: 307-312 [DOI: 10.1109/ITSC.2018.8569635]

Haarnoja T, Zhou A, Hartikainen K, Tucker G, Ha S, Tan J, Kumar V, Zhu H, Gupta A, Abbeel P and Levine S. 2018. Soft actor-critic algorithms and applications [EB/OL]. [2020-06-30]. https://arxiv.org/pdf/1812.05905.pdf

Han X M, Bao H, Liang J, Pan F and Xuan Z X. 2018. An adaptive cruise control algorithm based on deep reinforcement learning. Computer Engineering, 44(7): 32-35, 41 (**韩向敏, 鲍泓, 梁军, 潘峰, 玄祖兴**. 2018. **一种基于深度强化学习的自适应巡航控制算法**. **计算机工程**, 44(7): 32-35, 41) [DOI: 10.19678/j.issn.1000-3428.0050994]

He K M, Zhang X Y, Ren S Q and Sun J. 2016. Deep residual learning for image recognition//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 770-778 [DOI: 10.1109/CVPR.2016.90]

Hu J L and Wellman M P. 2003. Nash q-learning for general-sum stochastic games. Journal of Machine Learning Research, 4: 1039-1069

Kaelbling L P, Littman M L and Moore A W. 1996. Reinforcement learning: a survey. Journal of Artificial Intelligence Research, 4(1): 237-285

Konda V R and Tsitsiklis J N. 2003. OnActor-critic algorithms. SIAM Journal on Control and Optimization, 42(4): 1143-1166 [DOI: 10.1137/S0363012901385691]

Lange S, Riedmiller M and Voigtlander A. 2012. Autonomous reinforcement learning on raw visual input data in a real world application//Proceedings of 2012 International Joint Conference on Neural Networks. Brisbane, Australia: IEEE: 1-8 [DOI: 10.1109/IJCNN.2012.6252823]

Li D Y. 2015. Formalization of Brain cognition—start with the development of a robotic driving brain. Science and Technology Review, 33(24): #125 (**李德毅**. 2015. **脑认知的形式化——从研发机器驾驶脑谈开去**. **科技导报**, 33(24): #125)

Liang X D, Wang T R, Yang L N and Xing E. 2018. CIRL: controllable imitative reinforcement learning for vision-based self-driving//Proceedings of the 15th European Conference on European Conference on Computer Vision. Munich, Germany: Springer: 584-599 [DOI: 10.1007/978-3-030-01234-2_36]

Lillicrap T P, Hunt J J, Pritzel A, Heess N, Erez T, Tassa Y, Silver D and Wierstra D. 2015. Continuous control with deep reinforcement learning [EB/OL]. [2020-06-30]. https://arxiv.org/pdf/1509.02971.pdf

Littman M L. 1994. Markov games as a framework for multi-agent reinforcement learning//Proceedings of the 11th International Conference on International Conference on Machine Learning. New Brunswick, USA: IEEE: 157-163

Minsky M. 1961. Steps toward artificial intelligence. Proceedings of the IRE, 49(1): 8-30 [DOI: 10.1109/JRPROC.1961.287775]

Mnih V, Kavukcuoglu K, Silver D, Rusu A A, Veness J, Bellemare M G, Graves A, Riedmiller M, Fidjeland A K, Ostrovski G, Petersen S, Beattie C, Sadik A, Antonoglou I, King H, Kumaran D, Wierstra D, Legg S and Hassabis D. 2015. Human-level control through deep reinforcement learning. Nature, 518(7540): 529-533 [DOI: 10.1038/nature14236]

Ostafew C J. 2016. Learning-based Control for Autonomous Mobile Robots. Toronto: University of Toronto

Ostafew C J, Schoellig A P, Barfoot T D and Collier J. 2016. Learning-based nonlinear model predictive control to improve vision-based mobile robot path tracking. Journal of Field Robotics, 33(1): 133-152 [DOI: 10.1002/rob.21587]

Paden B, Čáp M, Yong S Z, Yershov D and Frazzoli E. 2016. A survey of motion planning and control techniques for self-driving urban vehicles. IEEE Transactions on Intelligent Vehicles, 1(1): 33-55 [DOI: 10.1109/TIV.2016.2578706]

Pan F and Bao H. 2019. Reinforcement learning model with a reward function based on human driving characteristics//Proceedings of the

15th International Conference on Computational Intelligence and Security. Macao, China: IEEE: 225-229 [DOI: 10.1109/CIS.2019.00055]

Pinto L, Davidson J, Sukthankar R and Gupta A. 2017. Robust adversarial reinforcement learning//Proceedings of the 34th International Conference on Machine Learning. Sydney, Australia: ACM: 2817-2826

Rajamani R. 2012. Vehicle Dynamics and Control. Boston, MA: Springer [DOI: 10.1007/0-387-28823-6]

Sallab A E, Abdou M, Perot E and Yogamani S. 2016. End-to-end deep reinforcement learning for lane keeping assist [EB/OL]. [2020-06-30]. https://arxiv.org/pdf/1612.04340.pdf

Samuel A L. 1959. Some studies in machine learning using the game of checkers. IBM Journal of Research and Development, 3(3): 210-229 [DOI: 10.1147/rd.33.0210]

Shalev-Shwartz S, Ben-Zrihem N, Cohen A and Shashua A. 2016. Long-term planning by short-term prediction [EB/OL]. [2020-06-30]. https://arxiv.org/pdf/1602.01580.pdf

Silver D, Huang A, Maddison C J, Guez A, Sifre L, Van Den Driessche G, Schrittwieser J, Antonoglou I, Panneershelvam V, Lanctot M, Dieleman S, Grewe D, Nham J, Kalchbrenner N, Sutskever I, Lillicrap T, Leach M, Kavukcuoglu K, Graepel T and Hassabis D. 2016. Mastering the game of Go with deep neural networks and tree search. Nature, 529(7587): 484-489 [DOI: 10.1038/nature16961]

Silver D, Hubert T, Schrittwieser J, Antonoglou I, Lai M, Guez A, Lanctot M, Sifre L, Kumaran D, Graepel T, Lillicrap T, Simonyan K and Hassabis D. 2017. Mastering chess and shogi by self-play with a general reinforcement learning algorithm [EB/OL]. [2020-06-30]. https://arxiv.org/pdf/1712.01815.pdf

Snider J M. 2009. Automatic Steering Methods for Autonomous Automobile Path Tracking. Robotics Institute

Spryn M, Sharma A, Parkar D and Shrimal M. 2018. Distributed deep reinforcement learning on the cloud for autonomous driving//Proceedings of the 1 st International Workshop on Software Engineering for AI in Autonomous Systems. Gothenburg, Sweden: ACM: 16-22 [DOI: 10.1145/3194085.3194088]

Sutton R S. 1992. Introduction: The challenge of reinforcement learning. Machine Learning, 8(3): 225-227 [DOI: 10.1023/A:1022620604568]

Sutton R S and Barto A G. 1998. Introduction to Reinforcement Learning. Cambridge: MIT Press

Watkins C J C H and Dayan P. 1992. Q-learning. Machine Learning, (3-4): 279-292

Waltz M and Fu K. 1965. A heuristic approach to reinforcement learning control systems. IEEE Transactions on Automatic Control, 10(4): 390-398 [DOI: 10.1109/TAC.1965.1098193]

Watldns C J C H. 1989. Learning from Delayed Rewards. Cambridge: Cambridge University: 233-235

Xu Y C, Wang R B, Li B and Li B. 2001. A summary of worldwide intelligent vehicle. Automotive Engineering, 23(5): 289-295 (徐友春, 王荣本, 李兵, 李斌. 2001. 世界智能车辆近况综述. 汽车工程, 23(5): 289-295) [DOI: 10.19562/j.chinasae.qcgc.2001.05.001]

Yang L N, Liang X D, Wang T R and Xing E P. 2018. Real-to-virtual domain unification for end-to-end autonomous driving//Proceedings of the 15th European Conference on Computer Vision (ECCV). Munich, Germany: Springer: 553-570

Yang S. 2019. From Virtuality to Reality: Research on Deep Reinforcement Learning Based Autonomous Vehicle Control. Changchun: Jilin University (杨顺. 2019. 从虚拟到现实的智能车辆深度强化学习控制研究. 长春: 吉林大学)

Zong X P, Xu G Y, Yu G Z, Su H J and Hu C W. 2018. Obstacle avoidance for self-driving vehicle with reinforcement learning. SAE International Journal of Passenger Cars-Electronic and Electrical Systems, 11(1): 30-39 [DOI: 10.4271/07-11-01-0003]

作者简介

潘峰,1978 年生,男,副教授,主要研究方向为智能驾驶、智能控制、强化学习。

E-mail: jjtpanfeng@buu.edu.cn

鲍泓,通信作者,男,教授,博士生导师。主要研究方向为智能驾驶、视觉计算和分布式网络。

E-mail: baohong@buu.edu.cn