深度强化学习理论及其应用综述

万里鹏[1] 兰旭光[1] 张翰博[1] 郑南宁[1]

摘 要 一方面,随着深度强化学习理论和应用研究不断深入,其在游戏、机器人控制、对话系统、自动驾驶等领域发挥重要作用;另一方面,深度强化学习受到探索-利用困境、奖励稀疏、样本采集困难、稳定性较差等问题的限制,存在很多不足. 面对这些问题,研究者们提出各种各样的解决方法,新的理论进一步推动深度强化学习的发展,在弥补缺陷的同时扩展强化学习的研究领域,延伸出模仿学习、分层强化学习、元学习等新的研究方向. 文中从深度强化学习的理论、困难、应用及发展前景等方面对其进行探讨.

关键词 深度强化学习, 马尔科夫决策过程, 探索-利用困境, 稀疏奖励

引用格式 万里鹏,兰旭光,张翰博,郑南宁. 深度强化学习理论及其应用综述. 模式识别与人工智能, 2019, 32(1): 67-81.

DOI 10.16451/j.cnki.issn1003-6059.201901009 **中图法分类号** TP 181

A Review of Deep Reinforcement Learning Theory and Application

WAN Lipeng[1], LAN Xuguang[1], ZHANG Hanbo[1], ZHENG Nanning[1]

ABSTRACT Deep reinforcement learning(DRL) theory and applied research are deepening and it is now playing an important role in games, robot control, dialogue systems, automatic driving, etc. Meanwhile, due to shortcomings such as exploration-exploitation dilemma, sparse reward, sample collection hardness, poor model stability, DRL still has many problems for which researchers have proposed various solutions. New theories has further promoted the development of DRL, and opened up several new research fields of reinforcement learning, such as imitative learning, hierarchical reinforcement learning and meta-learning. This paper aims to explore and summarize future development of DRL, and a brief introduction of DRL theory, difficulties and applications is presented at the same time.

Key Words Deep Reinforcement Learning, Marcov Decision Process, Exploration-Exploitation Dilemma, Sparse Reward

Citation WAN L P, LAN X G, ZHANG H B, ZHENG N N. A Review of Deep Reinforcement Learning Theory and Application. Pattern Recognition and Artificial Intelligence, 2019, 32(1): 67-81.

人工智能(Artificial Intelligence, AI)领域的一个主要研究目标是实现完全自主的智能体[1],这一智能体能够与其所处的环境进行交互,根据环境反馈学习最佳行为,并通过反复实验不断改进行动策

收稿日期:2018-12-29;录用日期:2019-01-04
Manuscript received December 29, 2018;
accepted January 4, 2019
国家自然科学基金重点项目(No. 91748208)、国家自然科学基金面上项目(No. 61573268)、国家科技部重点研发计划项目(No. 2018ZX01028101)、陕西省重点研发计划项目(No. 2018ZDCXLGY0607)、微软亚洲研究院合作项目(No. 01051311120002601)资助
Supported by Key Program of National Natural Science Foundation of China(No. 91748208), General Program of National Natural Science Foundation of China(No. 61573268), Key Research and Development Program of Ministry of Science and Technology of China(No. 2018ZX01028101), Key Research and Development Program of Shaanxi Province(No. 2018ZDCXLGY0607), Cooperation Program with Microsoft Research Asia(No. 01051311120002601)
本文责任编委 吴飞
Recommended by Associate Editor WU Fei

1. 西安交通大学 人工智能与机器人研究所 西安 710049
1. Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, Xi'an 710049

略. 深度强化学习(Deep Reinforcement Learning, DRL)的出现为这一目标的实现提供理论基础. 一方面,DRL 对策略和状态具有强大的表征能力,能够用于模拟复杂的决策过程;另一方面,强化学习(Reinforcement Learning, RL)赋予智能体自监督学习能力,使其能够自主地与环境交互,在试错(Trial and Error)中不断进步[2]. 深度强化学习作为人工智能研究领域的重要分支,被认为是实现类人智能的关键,受到学术和产业界的广泛关注.

机器学习可以分为监督学习、无监督学习和强化学习. 不同于监督学习和无监督学习,强化学习是一种自监督的学习方式:智能体一方面基于行动和奖励数据进行训练,并优化行动策略;另一方面自主地与环境互动,观测并获取环境反馈. 早期的强化学习方法基于最优控制理论,将强化学习的序列决策问题描述为自适应动态规划(adaptive dynamic programming, ADP)问题[3]. 研究人员以此为基础,对序列决策问题进行推广,得到基于策略的强化学习问题,并提出各种用于解决该问题的策略搜索算法. 更进一步地,为了简单直观地了解策略的优劣,学者们引入值函数作为策略评价的标准,提出 Q-learning 等一系列经典的强化学习模型[4]. 目前强化学习的发展已经进入与深度学习相互融合的阶段. 传统的强化学习方法受限于策略表征能力,只能处理一些简单的决策问题,深度学习的出现打破这一限制,与深度学习的结合给强化学习理论和应用注入新的动力.

Mnih 等[5]在《Nature》上发表论文,提出深度学习与强化学习相互结合的深度 Q 网络(Deep Q Network, DQN)模型,经过训练后在 Atari2006 游戏中的表现超过人类水平. DeepMind 团队[6]更进一步将深度强化学习应用到动作空间更大、策略更复杂的围棋游戏上,开发的 AlphaGo 程序基于深度神经网络,采用蒙特卡洛树搜索算法(Monte Carlo Tree Search, MCTS),同时融合监督学习和强化学习的训练方式,学习到超出人类水平的围棋策略,于 2016 年击败顶级人类棋手李世石,并于 2017 年击败围棋世界冠军柯杰. 此外,面对环境更复杂的在线战术竞技游戏,深度强化学习也有不俗的表现. OpenAI 团队面向 DOTA2 游戏开发的机器人程序,在 1 对 1 对战中击败顶级人类玩家.

除了游戏中的机器人程序,近年来研究者们基于深度强化学习提出许多新的概念和方法,并应用于工程项目中. DeepMind 团队[7]用模仿学习实现机器人的类人运动控制,而这一过程只需要少量的人类专家样本. Finn 等[8]结合深度强化学习与机器人抓取动作的预测,在训练机器人抓取策略的同时实现图像预测算法的自监督训练. 深度强化学习在自然语言处理[9-10]、自动驾驶[11]、推荐搜索系统[12]等领域也有应用.

深度强化学习模型尽管在虚拟的决策问题,如游戏决策中表现出色,但对于其它的很多决策问题,尤其是真实环境中的决策问题,受限于样本效率及样本采集的困难,相比传统方法,表现平平. 在机器人的运动控制问题中,根据强化学习的思想,机器人需要在运动中与环境交互,这一过程涉及策略执行、策略评估和奖励获取、策略优化三个过程. 对于深度策略或值函数网络,需要大量的样本用于训练,机器人在一次运动-观测的循环过程中只能获得一个训练样本,为了获得足够多的样本,需要大量的训练时间. 一种替代的方案是在虚拟环境中进行训练,将训练好的模型在真实环境下微调,然而这严重依赖于环境模拟器对真实环境的仿真能力,高性能的通用模拟器的开发也受到强化学习研究者的关注[13].

深度强化学习目前仍处于兴起阶段,属于人工智能的新兴研究领域,拥有广阔的发展空间和美好的应用前景. 本文从强化学习的原理入手,介绍强化学习的基本理论,包括强化学习的基本要素、马尔科夫决策过程(Markov Decision Process, MDP)、最优控制理论等,进一步介绍基于模型和免模型的强化学习、强化学习中常用的值函数评估和策略搜索方法. 本文就深度强化学习的策略执行、策略评估、策略优化这三个基本过程展开讨论,探讨强化学习的理论困境和当前热点论文提出的解决方法. 最后介绍深度强化学习的应用进展和未来的应用前景.

1 深度强化学习的基本理论

1.1 深度强化学习原理

强化学习的基本过程是一个马尔科夫决策过程,马尔科夫决策过程可以用状态(State)、行动(Action)、状态转移概率(Possibility)、状态转移奖励或回报(Reward)构成的四元组 $\{s,a,p,r\}$ 表示[14],如图 1 所示. 对于离散时间 MDP,状态和动作的集合称为状态空间(State Space)和动作空间(Action Space),分别使用 S 和 A 表示,$s_i \in S$, $a_i \in A$. 根据第 t 步选择的行动,状态根据概率 $P(s_{t+1} \mid s_t, a_t)$ 从 s_t 转移到 s_{t+1},在状态的转移的同时,决策主体得到一个

即时的奖励 $R_t(s_t,a_t,s_{t+1})$. 该过程结束时的累积奖励(Return)为

$$G_t = R_t + \gamma R_{t+1} + \gamma^2 R_{t+2} + \cdots = \sum_{k=0} \gamma^k R_{t+k+1},$$

其中,γ 为折扣因子,用于削减远期决策对应的奖励权重. 决策的最终目标是在抵达目标状态的同时实现累积奖励最大化.

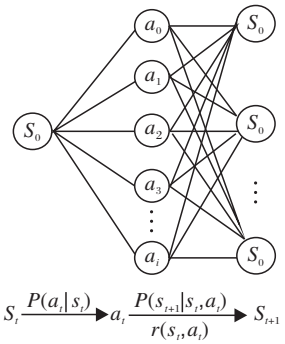


图 1　马尔科夫决策过程

Fig. 1　Markov decision process

在强化学习过程中,决策的主体称为智能体(Agent). 智能体首先需要对其所处的状态进行观测,并根据观测结果(Observation)进行决策,采取相应行动. 该行动一方面与环境(Environment)发生交互,环境以奖励的形式对智能体的行动给出相应的反馈;另一方面,该行动改变智能体的状态. 一个循环结束后,智能体开始新一轮的观测,直到智能体进入终止状态,此时一次完整的迭代结束,如图 2 所示. 智能体将此次迭代中的所有状态及其相应的动作以状态-动作序列的形式记录下来,生成轨迹(Trajectory):

$$\tau = \{s_t,a_t,s_{t+1},a_{t+1},\cdots\}.$$

同时统计每一步的即时回报,计算此次迭代中获得的累计回报 G_t,将这些信息作为策略更新时的训练样本.

智能体采取行动依据的策略使用函数 $\pi(a\mid s)$ 表示,智能体学习的目标就是优化这个策略函数. 根据优化对象的不同,强化学习方法可分为策略搜索方法(Policy Search)和值函数方法(Value Function). 强化学习过程中的状态转移概率又称为系统动态(Dynamics)、转移动态(Transition Dynamics)或环境模型,使用 $P_{ss'}$ 表示:

$$P_{ss'} = P(s_{t+1} = s' \mid s_t = s, a_t).$$

根据状态转移概率是否已知,可将强化学习方法分为基于模型(Model Based)的强化学习方法和免模型(Model Free)的强化学习方法[15].

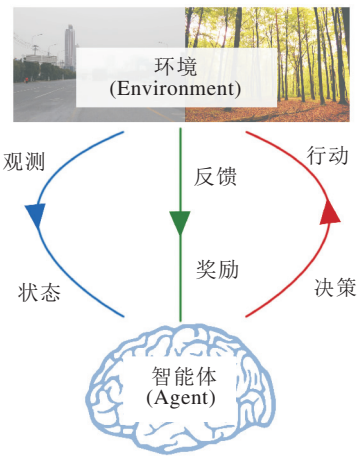


图 2　强化学习过程

Fig. 2　Reinforcement learning process

早期强化学习中的一个经典问题是轨迹规划问题. 轨迹规划问题的目标是训练程序在给定初始位置和终点位置的情况下自适应选择移动策略. 在离散时间条件下,轨迹规划问题可以归结为点到点的运动控制问题[16],常用的方法是用高斯混合模型(Gaussian Mixture Model, GMM)对移动策略建模[17],然后以专家轨迹为训练集,求解 GMM 参数.

强化学习方法在最优控制理论的基础上进行延伸,将最优控制问题拓展成更普遍、更广泛意义上的序列决策问题. 强化学习同时引入智能体和环境的概念,作为决策主体和主体的对立面,并采用即时奖励函数对最优控制问题中稀疏的目标奖励进行补充. 一方面,智能体能够自主地与环境交互,获取训练样本,进一步采用监督学习的方式更新策略,实现终身学习(Lifelong Learning),而不再依赖于有限的专家样本. 另一方面,与环境的交互过程还原具有自主性的决策主体的学习过程,这种学习方式同时也是最普遍的具有主体性的生物体的学习方式.

不论是最优控制问题,还是强化学习问题,其基本思想都是使用函数对策略进行建模或拟合,在一定的约束条件下优化策略函数. 传统的方法使用凸函数对策略建模,使用专家样本作为训练集进行参数估计,深度学习与强化学习的结合改变这一模式. 一方面,深度神经网络允许策略函数非凸,扩展强化学习方法的应用范围. 另一方面,深度神经网络强大的特征提取和函数拟合能力允许强化学习算法应用到非常复杂的决策问题上. 在一些以视觉信息为观测对象的问题中,深度神经网络与强化学习的结合

使端到端的训练成为可能,大幅节省观测数据处理过程的耗时. 此外,与深度学习的结合使强化学习变成一个数据驱动的自监督学习问题,对于数据获取较简单的场景,深度强化学习方法可以很简单地投入应用.

1.2 值函数和策略搜索

1.2.1 值函数

值函数法又被称为基于值函数(Value Based)的强化学习方法,是深度强化学习方法的代表. 在强化学习中,为了训练智能体使其习得一个好的策略,需要赋予智能体评估策略好坏的能力. 一种最直观的方式就是在特定的状态下,为每次行动赋予相应的评估价值. 在该状态下采取某一行动后,未来能够获得的累积回报期望值越高,对应的行动价值越大,可以使用状态-动作值函数 $Q_\pi(a,s)$ 对行动作的价值进行评估:

$$Q_\pi(a,s)=E[G_t\mid s_t=s,a_t=a].$$

相应地,每个状态的价值可以定义为从当前状态到终止状态能够获得的累积回报的期望,称为状态值函数:

$$V_\pi(s)=E[G_t\mid s_t=s].$$

智能体基于状态 s,采取某一特定的动作后可以得到 $Q_\pi(a,s)$,用于解决动态规划问题的 Bellman 方程给出的 $V_\pi(s)$ 和 $Q_\pi(a,s)$ 之间的相对关系[18]:

$$Q_\pi(a,s)=R_{t+1}+\gamma\sum_{s'\in S}P_{ss'}V_\pi(s'),$$

$$V_\pi(s)=\sum_{a\in A}\pi(a\mid s)Q_\pi(s,a).$$

值函数方法在强化学习中应用广泛,经典的值函数方法包括 Q-learning、DQN、基于 DQN 提出各种深度强化学习模型,如深度双 Q 网络[19]、对偶 DQN[20]、循环 DQN[21] 等. 此外,各种基于演员-评论家结构的强化学习算法中也几乎都用到值函数作为策略的评价标准[22-23]. 由于是对具体状态和动作进行评价,值函数法不适用于动作空间连续的强化学习问题,尽管近年来也出现将值函数法用于连续动作空间问题中的尝试[24],由于作者设定过多的约束和假设条件,并未提出一种在普遍意义上解决值函数和连续动作空间相容性的有价值的方法.

1.2.2 策略搜索

与值函数法对应的是策略搜索法,策略搜索法又称为基于策略(Policy Based)的强化学习方法. 不同于值函数法,策略搜索法不评价策略的好坏,而是基于采样的方法直接优化策略,使其向着能够使累积回报的期望增大的最终目标改进.

策略搜索法将策略参数化,以累积回报的期望作为目标函数

$$J(\theta)=E[G_t\mid\pi_\theta].$$

目标函数同时也是参数 θ 的函数,原问题变成基于 θ 的最优化问题,求解该优化问题的方法又称为策略梯度法. 策略梯度法需要考虑在当前策略下未来所有可能出现的轨迹,对这些轨迹对应的累积回报求平均,即需要对累积回报在状态空间和动作空间上求关于状态转移概率和策略的二重积分,在单步动作情况下

$$\nabla_\theta J(\theta)=\nabla_\theta E[R(s,a)\mid\pi_\theta]=\int_s P_{ss'}^{\pi}\int_a\nabla_\theta\pi_\theta(a\mid s)R(s,a)\,\mathrm{d}a\mathrm{d}s,$$

其中,$R(s,a)$ 为在状态 s 下采取动作 a 得到的即时奖励,在连续 N 步动作时,可用 Q 值函数替代. 从而

$$\nabla_\theta J(\theta)=\int_s P_{ss'}^{\pi}\mathrm{d}s\int_a\nabla_\theta\pi_\theta(a\mid s)Q^\pi(s,a)\,\mathrm{d}a=\int_s P_{ss'}^{\pi}\mathrm{d}s\int_a\pi_\theta(a\mid s)\frac{\nabla_\theta\pi_\theta(a\mid s)}{\pi_\theta(a\mid s)}Q^\pi(s,a)\,\mathrm{d}a=\int_s P_{ss'}^{\pi}\mathrm{d}s\int_a\pi_\theta(a\mid s)\nabla_\theta\lg\pi_\theta(a\mid s)Q^\pi(s,a)\,\mathrm{d}a=E_{s\sim P_{ss'}^{\pi},a\sim\pi_\theta}[\nabla_\theta\lg\pi_\theta(a\mid s)Q^\pi(s,a)].$$

上式涉及对状态和动作的二重积分,不能直接进行计算,一种相对简单的处理方式是使用蒙特卡洛采样法对梯度进行估计. 具体做法是采样 m 条轨迹,每条轨迹对应于 T 步动作,求取目标函数梯度的平均:

$$\nabla_\theta J(\theta)=\frac{1}{m}\sum_{t=0}^{T-1}\sum_{i=1}^{m}\nabla_\theta\lg\pi_\theta(a_t^i\mid s_t^i)Q(a_t^i,s_t^i).$$

使用经验平均估计策略梯度涉及重要性采样,需要对目标函数进行修正:

$$\nabla_\theta J(\theta)=E_{s\sim P_{ss'}^{\pi_{\theta demo}},a\sim\pi_{\theta demo}}\left[\frac{\nabla_\theta\lg\pi_\theta(a\mid s)}{\lg\pi_{\theta demo}(a\mid s)}Q^{\pi_{\theta demo}}(s,a)\right],$$

其中 $\pi_{\theta demo}$ 表示样本策略. 由于更新后的最优策略和样本收集时采用的策略不同,这一类问题又称为离策略(Off Policy)的强化学习问题.

1.2.3 演员-评论家模型

值函数法和策略搜索法各有优劣. 一方面,策略搜索法直接参数化策略,更简单直观,相比值函数法,收敛性更好,可以更好地处理大状态空间问题,相比之下,对于动作空间连续或者动作空间维度过高的情况,值函数法往往不能有效求解. 另一方面,策略搜索法的策略评估能力较弱,容易导致策略梯

度的方差较大. 此外,策略搜索法还容易收敛到局部极小值. 为此,出现很多将策略搜索与值函数相互结合的强化学习方法, 如深度确定性梯度下降算法(Deep Determined Policy Gradient, DDPG)[25]、Q-Prop[26]、异步演员-评论家算法(Asynchronous Actor Critic, A3C)[27] 等. 其基本思想都是采用策略梯度的方法更新策略,同时结合值函数作为策略的评价手段. 在策略搜索中引入值函数作为策略的评价标准能够有效减小方差.

在演员-评论家模型中,通常以神经网络作为函数近似器估计值函数. 优化过程中一方面需要保证值函数近似的准确性,同时还要优化策略,在决策中采取相应的行动使值函数或累积回报的期望最大化. 所以,在演员-评论家模型中,损失函数通常包含两个部分:

$$Loss_T = \{Loss_V, Loss_P\}.$$

其中:$Loss_T$ 为总的损失函数;$Loss_V$ 为近似损失, 是 critic 网络的损失函数, 表示值函数的近似误差; $Loss_P$ 为策略损失,为 actor 网络的损失函数,表示累积回报期望的损失. 在有些方法,如 DDPG 中,actor 和 critic 网络以交替迭代的方式进行更新,$Loss_V$ 和 $Loss_P$ 需要迭代计算. 在另一些方法,如 A3C、PPO 中,可以把两部分损失函数直接相加,在更新 critic 网络的同时最大化累积回报的期望,简化计算过程. 通常为了保证智能体策略的随机性,还会在损失函数中加入动作的熵作为惩罚项[27].

1.3　基于模型的强化学习和免模型的强化学习

在强化学习中,环境模型又称为系统动力学模型,是描述状态转移概率的函数. 在一些问题中,环境模型是已知的,智能体可以根据当前始状态进行决策,并预测下一个状态出现的概率. 在这些问题中,智能体有能力预测未来每一步的状态并作出相应的决策. 也就是说,智能体可以在不与环境发生互动的条件下进行规划(Planning). 在环境模型已知情形下的强化学习问题就是基于模型的强化学习问题.

基于模型的强化学习方法常见于轨迹规划问题中. 轨迹规划问题的样本效率非常低,对于一些参数变量多、样本需求大的方法,很难通过训练得到有效策略,而基于模型的策略优化可以充分利用模型先验,降低对样本数量的需求,在一定程度上提高样本效率. 其基本思想是对系统动态建模,先学习模型参数,再利用系统动力学模型辅助策略学习. 系统动态的建模方法通常为 GMM 模型、高斯过程模型、贝叶斯网络、深度神经网络等. 随着深度学习的不断成熟,利用深度神经网络 对系统动力学模型建模成为 2018 年的研究热点[28-29].

免模型的方法尝试使用一种算法解决所有强化学习问题,相比之下,基于模型的强化学习针对特定问题建模,导致泛化能力不强. 但是它充分利用特定问题的特定信息,需要的参数更少,训练更容易. 此外,在基于模型的方法中,模型具有一定的解释性,对于参数调试具有一定的指导意义. 免模型方法在训练过程中不易收敛,稳定性也不及基于模型的强化学习方法. 然而,免模型的强化学习方法更直观、简单,通用性更强,对于一些难以建模的复杂问题,免模型方法是更好的选择. 为此,研究人员开始尝试将免模型的和基于模型的强化学习方法相互结合. Nagabandi 等[30] 结合两种方法,利用基于模型的 RL 方法对免模型的 RL 过程进行初始化,在降低样本复杂性的同时让智能体在训练过程中能够获得更多的回报. Oh 等[31] 将值函数用于模型动力学的估计中,参考基于模型的 RL,估计未来状态的值函数,在一定程度上学到系统动力学模型,用于未来状态的短期预测,并根据预测结果进行规划.

1.4　深度强化学习的理论困境

在一个具体的决策问题中,无论是采用基于模型的强化学习方法还是免模型的强化学习方法,以及无论是以值函数或策略作为优化对象,还是采用二者结合的演员评论家结构,都离不开强化学习的基本过程. 一个基本的强化学习迭代过程可以分解为策略执行、策略评估、策略优化三个子过程,每个子过程中都存在一些固有的难题,这些难题成为限制强化学习理论和应用发展的巨大障碍.

1.4.1　探索-利用困境

在策略执行过程中,智能体一方面需要尽量尝试新的策略,发现潜在未知的解决问题的方法,另一方面需要充分利用已经习得的策略,避免盲目尝试,提高学习的效率. 随机策略可以帮助智能体遍历各种状态,避免陷入局部最优解,但是混乱的尝试降低学习效率,使智能体花费时间在很多不必要的状态和策略间来回探索,即智能体需要兼顾策略的探索(Exploration) 与利用(Exploitation).

策略的探索与利用的平衡问题是强化学习中的经典、难点问题,在一些策略函数为凸的问题中,策略函数只有一个极值点,该点对应于全局最优解,可以简单地采用贪婪策略(Greedy Policy),而不需要考虑探索-利用的平衡. 贪婪策略指智能体在决策过

程中总是采用已经习得的最优策略. 对于非凸的问题,由于探索的不充分,采集得到的样本也是局限的,使用该样本进行策略优化时很容易陷入局部最优. 一种改进的方法是在策略中加入噪声,或以其它方式引入随机性,即增加策略的探索度. 然而,简单地引入探索度也存在一定的问题. 一方面,探索度太小并不能保证样本的随机性,难以解决策略容易陷入局部极小的问题;另一方面,探索度太大会生成大量对于策略改进无帮助的样本,降低学习效率.

1.4.2 奖励函数设计的困难与稀疏奖励问题

如何有效地评估策略的好坏是智能体学习效率的关键,目前,策略评估主要依赖于奖励函数,而奖励函数又依赖于人类专家的设计. 对于一些复杂的决策问题,难以设计好的奖励函数. 为此研究人员提出元学习(Meta Learning)、模仿学习(Imitation Learning)等方式,让智能体学习从好的策略中总结相应的奖励函数,用于指导强化学习过程. 然而,模仿学习需要借助反向强化学习(Inverse Reinforcement Learning)和强化学习的交替迭代,过程过于复杂[32],而且模仿学习依赖于专家样本,对于某些缺少专家样本的场合不适用.

元学习作为一种试图解决所有任务的尝试,还处于发展阶段,当前的元学习思路和迁移学习(Transfer Learning)、终身学习具有相似之处,强调提高智能体对环境和任务变化的被动适应性,还不能从智能体本身给出一种主动性的解决问题的普遍方法. 另一种方法是采用分层强化学习(Hierarchical Reinforcement Learning)[33],将复杂的强化学习任务分解,每个子任务单独设计奖励函数. 此外,对于一些稀疏奖励的问题,强化学习的效率也非常低. 研究人员为此提出一些解决方案,包括设置辅助任务[34]、引入好奇心机制[35]等,这些方法依然受到泛化能力的限制,需要根据具体任务由专家提供相应的先验信息,不能在普遍意义上解决强化学习任务的稀疏奖励问题.

1.4.1 策略优化过程中面临的挑战

策略优化是强化学习过程的最后一步,用于估计策略的方法,包括神经网络近似、高斯过程建模、贝叶斯网络建模等. 神经网络近似是深度强化学习的标准方法,也是免模型问题中最常用的策略估计方法,但是非线性的神经网络往往存在收敛困难的问题,在学习环境动态的过程中容易出现对环境的过拟合现象,导致该方法的最终效果往往不如基于模型的强化学习方法. 而基于模型的强化学习方法需要针对具体问题建模,特定的算法应用场景十分有限,对于一些复杂的问题,同时还存在建模上的困难. 除此以外,在两种优化方法中,线搜索法不能保证策略收敛,信赖域法不适用于大状态和动作空间问题[74].

强化学习的这些难点给强化学习的研究者提出巨大的挑战,除了策略执行、策略评估、策略优化,强化学习在实际应用之前还存在环境观测和样本采集等方面的困难[36]. 研究者们针对这些问题提出各种各样的解决方案,其中不少具有启发意义. 尽管这些方法还不足以解决上述问题,但至少为问题的解决提供可行的思路.

2 深度强化学习研究现状

深度强化学习的研究工作基本都是围绕策略执行、策略评估和策略优化中出现的问题展开,本节将从上述 3 方面介绍强化学习的研究进展.

2.1 策略执行

在智能体策略执行中所面对的困难主要是探索-利用困境. 探索-利用困境最早见于多臂赌博机问题[37]. 在该问题中,决策者一方面需要多试探各个摇臂,大致了解每个摇臂对应的回报的期望,另一方面需要充分利用已得到的信息,试探回报期望较大的几个摇臂,以确定哪个摇臂回报的期望最大,即要在探索信息与利用信息之间权衡. 针对多臂老虎机问题提出的一种解决探索-利用困境问题方法是上界置信(Upper Confidence Bound, UCB)算法[38],使用一个“分数”评估每个摇臂,选择对应分数最大的那个,将原始问题转换成如下排序问题:

$$\max\ \text{mize}:\bar{x}_i+\sqrt{\frac{2\ln n}{n_i}}. \tag{8}$$

其中,$\bar{x}_i$ 表示从第 i 个摇臂获得的回报的经验平均,n_i 表示第 i 个摇臂的尝试次数,n 表示尝试摇臂的总次数. 上式中第 1 项表示“利用优先”,第 2 项表示“探索优先”,UCB 体现探索-利用平衡的思想.

当前最常见的用于解决探索 - 利用困境的方法为 ε-greedy 算法[39]. ε-greedy 算法的基本思想是让智能体在决策时以 ε 的概率去探索,即采取随机策略,以 1 - ε 的概率去利用,即采用当前的最优策略,同时让 ε 随着训练过程的进行不断衰减,以保证训练后期的学习效率. ε-greedy 算法简单实用,被 DQN 等经典的强化学习模型采纳,但是智能体在采用随机策略进行探索时效率不高,特别在连续动作空间

问题上. 这是由于随机策略未利用当前最优策略中的有效信息,而是采取盲目探索的方式,从而降低探索效率.

在其它一些基于策略的强化学习方法中, 如 DDPG,通常采取在策略上添加噪声的方式进行探索,即在当前的最优策略上添加一个小的噪声,引入噪声相当于对当前的策略进行微调. 不同于盲目探索,方法利用先前训练中已经学到的决策信息,对于连续动作空间强化学习问题,可以渐进地调整策略,大幅提高探索效率. 但是该方法也存在着一定缺陷,由于是对策略进行连续调整,如果添加的噪声太小,容易使策略收敛到局部最优而不是全局最优. 受此启发,Fortunato 等[40] 提出含噪声的网络结构(Noisy Network),通过在策略网络的权重上增加噪声实现策略的随机性. 和动作噪声类似的是,权重噪声可用于策略的连续调整,后者不仅可以用于连续动作空间中的探索,还能用于离散动作空间中,具有更广泛的应用范围. 权重是策略拟合函数的参数,无实际意义,其缺点是缺乏解释性.

最大化行动的熵被认为是一种促进策略探索的有效方法. 智能体在决策时采取的行动是评分最高或概率最大的那个. 为了保证决策时未被探索的行动具有相当的被选择概率,可以将最大化动作的熵作为策略优化的目标之一. 具体做法是将动作的熵作为优化目标之一加入目标策略函数中,这样智能体对于未曾执行的行动或评分相同的行动将赋予相当的被选择的概率. 该方法作为对策略目标函数的修正,用于许多强化学习模型中,如 A3C、近似策略优化(Proximal Policy Optimization, PPO)、软演员-评论家(Soft AC) 等. 事实上最大化行动的熵只是提高决策的“公平性”,消除智能体对于未知问题“猜测” 的倾向,本质上并未在策略中引入随机性.

除此以外,学者们相继提出一些新的解决探索-利用困境的方法. Colas 等[41] 延续“前期探索为主,后期利用为主” 的基本思想,提出二阶段的策略执行方法. 首先,在第一阶段,忽略奖励对智能体行为的指导,最大化行动的多样性,用于发现一系列简单的策略. 然后,在第二阶段充分利用之前习得的策略信息,在其基础上对策略进行筛选和微调. Oh 等[42] 认为对过去经验的利用可以提高探索效率,以此为基础提出自我模仿的强化学习方法(Self-Imitation Learning, SIL),让智能体在学习过程中模仿过去好的经验. 实验表明对于一些探索困难的问题,SIL 可以提高探索效率. 在 DDPG 等许多连续动作空间的强化学习算法中,一般通过给策略加噪声的方法赋予智能体策略探索能力,Xu 等[43] 在此基础上进行改进,结合元学习思想与策略梯度法,提出一种探索自适应的策略梯度法,显著提高 DDPG 在各种强化学习连续控制任务上的样本效率.

不论是 ε-greedy、二阶段探索、SIL,还是自适应的探索方式,其基本思想都是参照人类的学习方式,在训练前期样本还较少时以高探索度去探索和发现不同特征的样本,然后在训练后期策略较成熟时降低探索度,利用当前策略进行深度探索,即对策略进行微调. 对于人类学习者来说,可以随着训练过程的进行不断调整探索和利用的平衡,如何让智能体学会把握训练中各个阶段探索度的大小,以及如何充分利用当前策略和经验来指导探索,是未来探索-利用困境中的核心问题.

2.2　策略评估

一个好的奖励函数可以有效指导智能体的学习过程,大幅提高学习效率. 奖励函数往往依赖专家设计,然而,在一些问题中策略的好坏难以量化,给奖励函数的设计带来困难[44]. 例如在机械臂的运动控制问题中,评价一个运动轨迹的好坏十分困难. 一种处理此类问题的方法是模仿学习[45],其基本思想是不再人为设计奖励函数,而是让智能体模仿专家行为,从模仿中学习. 最简单的模仿学习是行为克隆(Behavior Clone),根据专家指导,采取监督学习的方式直接学习策略. 行为克隆仅适用于简单策略的学习,对于更复杂的策略,模仿学习采用的方法是两步迭代式训练法. 一种经典的两步迭代模仿学习方法是学徒学习(Apprenticeship Learning)[46],在第 1 步中,专家首先要对奖励函数建模,将奖励函数定义为一系列基本损失函数的线性组合:

$$C_{\text{linear}} = \left\{ c_{\omega} \triangleq \sum_{i=1}^{k} \omega_i c_i \mid \| \boldsymbol{\omega} \|_2 \leqslant 1 \right\},$$

其中,C_{linear} 表示目标奖励函数,c_i 表示构成目标奖励函数的基函数,ω_i 表示基函数的权重. 智能体在此基础上优化奖励函数,从策略池中采样智能体策略. 对比专家策略, 最大化专家策略对应的累积回报的优势:

$$\delta_C(\pi, \pi_E) = \sup_{c \in C} \eta^C(\pi) - \eta^C(\pi_E),$$

$$\min_{\pi} \delta_C(\pi, \pi_E),$$

其中,C 表示用于存储过往策略的策略池,表示专家策略对应的累积回报,表示智能体的过往策略对应的累积回报. 这一步采用的方法也称为反向强化学

习. 反向强化学习和强化学习过程相反,强化学习是在奖励函数的指导下学习策略,而反向强化学习是在专家策略的指导下学习奖励函数[47].

在第2步中,智能体利用第1步中学习到的奖励函数指导强化学习过程,更新策略,将更新后的策略存入策略池中. 经过迭代可以得到最终的奖励函数和目标策略.

模仿学习涉及反向强化学习和强化学习的交替迭代,过程十分复杂. 此外,在有些问题(如游戏)中,人类专家并不能提供最优的行动作为奖励函数的优化指导,使得模仿学习的应用场景十分有限,主要在机器人的类人动作模仿任务中. 近两年的模仿学习研究主要聚焦于机器人控制领域的应用,其中一个较重要的发现是OpenAI团队[48]提出的对抗模仿学习(Generative Adversarial Imitation Learning, GAIL)模型. GAIL将生成对抗网络引入模仿学习中,使用生成器生成行动,使用判别器判别行动是否来源于专家策略,获得较好效果. 一些其它的工作包括简化模仿学习的复杂性[49],提高样本的利用效率[50]、提高算法的鲁棒性[51]等也受到研究者们的关注.

在多任务强化学习问题中,同样存在奖励函数难以设计的问题,其中一种解决思路是使用元学习的方式对任务进行总结和归纳[52]. 元学习解释为"学会学习(Learning to Learn)",是一种"学习如何学习"的机器学习思想. 在多任务问题中,元学习试图找到一种适用于不同任务的通用模型. Al-Shedivat等[53]将任务的切换看成是非静态环境发生变化的结果,即认为任务是环境因素,多任务问题等价于教育智能体在非静态环境下进行决策的问题. 同样是面对不断变化的学习任务,在有限的尝试中调整策略,终身学习[54-55]、迁移学习[56]等概念和元学习一起,成为近年来机器学习领域的研究热点. 针对多任务问题,Finn等[57]提出与模型无关的元学习方法(Model Agnostic Meta Learning, MAML),MAML在各种不同的任务下进行训练,并试图找到一个初始条件,以此为基础可以快速、高效地适应不同任务. Ritter等[58]将长短期记忆单元(Long Short Term Memory, LSTM)引入元学习架构中,用于处理多任务场景中任务重复出现的情况,使智能体在遇见重复问题时无需重新探索,而是可以利用之前学习得到的策略.

另一种应对多任务学习问题的方法是分层强化学习. 分层强化学习将目标任务分割成独立的子任务,对每个子任务设立单独的任务层级,每个层级设立相应的奖励函数,用于各个子任务的评估. 例如在一个复杂的导航问题中,上层策略可用于处理定位任务,下层策略可用于处理移动任务. 最经典的分层强化学习方法是Sutton等[59]提出的基于选择项(Option)的分层强化学习方法,将每个策略层的决策结果定义为该层的选择项,上层策略通过选择本层的选择项决定下层选择项的取值,底层选择项对应于具体动作,即高层策略可通过控制低层策略达到决定最终决策结果的目的. Haarnoja等[60]使用隐空间(Latent Space)替代选择项,提出基于隐空间的分层学习方法. 其基本思想是对底层策略进行参数化,同时在每层定义一个隐变量,该隐变量所在的空间称为隐空间. 上层策略通过决策对其隐变量的取值范围进行约束,得到一个缩小的隐空间. 下层策略进一步缩小该隐空间的范围,底层的隐空间就是策略的参数空间. 也就是说,每层策略都通过对最终策略的参数取值进行约束以实现分层决策.

针对多任务学习问题,除了纵向的层级结构,横向的并联结构也深受欢迎. Fang等[61]提出使用三神经网络并联的模型解决机械臂多任务操作问题. Murali等[62]使用一个树型神经网络处理多模式抓取问题,对抓取模式按维度进行区分,每个维度对应于一个单独的叶网络,这些网络共享抓取对象的特征,每个叶网络在不同维度进行决策,最终策略为各个叶网络策略的合成.

策略评估的另一个难点是稀疏问题,这意味着在很多情形中,智能体需要经历不断地决策、执行策略,最终却只能得到很少的反馈. 例如在迷宫游戏中,智能体通过漫长的尝试,最终得到的反馈仅仅是成功与否,而很难从在迷宫试探的过程中获得任何有价值的信息.

针对稀疏奖励问题,Andrychowicz等[63]提出事后经验重播方法,并应用于强化学习中. 其基本思想是设立一系列隐式任务,将智能体策略执行中的尝试设定为隐式任务的目标,充分利用失败经验,让智能体从失败的尝试中进行学习,能够在一定程度上应对稀疏奖励问题. Jaderberg等[64]在原始的目标任务中增添许多辅助任务,并给这些辅助任务设立相应的伪奖励(Pseudo Reward),使智能体从环境中能够获取更丰富的反馈信息,在第一人称三维迷宫任务中的表现超过87% 的人类专家. Riedmiller等[65]提出计划辅助控制,应对抓取任务中的稀疏奖励问题,通过设置抓取对象是否发生位移、机械臂是否检

测到触碰等一系列外置的辅助任务,弥补环境反馈的稀疏. 在SAC-X 中还用到分层强化学习,下层策略优化用于应对不同的外置辅助任务,而上层策略通过对下层策略的调度以满足最终的目标任务.

2.3　策略优化

策略优化是深度强化学习的最后一步,策略执行和策略评估的目的都是为策略优化提供相应的条件. 策略执行解决的是“如何进行探索”的问题,其主要目的是获取有价值、多样性的样本. 策略评估解决的是“如何保证样本的准确性和效率”的问题,其目的是尽量准确地对样本价值做出真实的评价,并扩充样本中包含的信息. 策略优化解决的是基于策略执行和评估获得样本的优化问题. 根据问题的不同,采用优化方法也各不相同. 大致可以分为基于模型的策略优化方法和免模型的策略优化方法.

2.3.1　基于模型的策略优化方法

基于模型的强化学习方法的一个难点是模型误差,Deisenroth 等[66] 提出学习控制的概率推理(Probabilistic Inference for Learning Control, PILCO),建立概率动力学模型,将模型的不确定性纳入决策时的考量中. PILCO 的实现方法是使用高斯过程模型对系统动态建模,基于模型推理预测未来状态的概率分布,同时计算这些状态的值函数,用于策略评估,最后将预测结果和评估结果作为样本更新策略. 近年来,研究人员对 PILCO 方法进行拓展,在其基础上提出很多改进措施. McAllister 等[67] 在 PILCO 的目标函数中加入对累积代价函数方差的评估,提高探索度. Gal 等[68] 使用贝叶斯网络替代 PILCO 中的高斯回归模型,将其应用于高维度系统中.

受引导策略搜索(Guided Policy Search, GPS)[69] 是另一种受到广泛关注的基于模型的强化学习方法. 轨迹规划本身是最优控制理论中的传统问题,GPS 将传统控制理论中控制序列的更新方法应用于强化学习的策略更新中. GPS 使用的策略更新方法称为迭代线性二次高斯算法(Iterative Linear Quadratic Gaussian, iLQG),核心思想是将非线性最优控制问题迭代地在局部转化为控制理论中的 LQG(Linear Quadratic Gaussian) 问题,通过动态规划求解局部 LQG 问题更新策略[70]. 具体过程通过两步迭代实现,在第 1 步中,首先在轨迹 τ 附近将系统的动力学特性线性化,同时损失函数二次化,得到一个局部 LQG 问题,通过求解该问题,得到新的控制序列,即新的策略. 在第 2 步中,通过执行更新后的策略得到新的轨迹 τ'. 其目标函数

$$p(\tau) = \min_{p(\tau)\in N(\tau)} E_p[l(\tau)] - H[p(\tau)],$$

$$\text{s.t. } p(x_{t+1} \mid x_t, u_t) = N(x_{t+1}; f_{xt}x_t + f_{ut}u_t).$$

其中,系统动态 $p(x_{t+1} \mid x_t, u_t)$ 通过 GMM 建模,在最小化损失函数的同时利用轨迹熵最大化以提高探索度. 由于策略更新需要对轨迹进行局部线性化,必须保证轨迹的变化在一个可信赖的小范围之内,通常需要约束策略更新前后轨迹的相对熵以限制轨迹变化的幅度. Levine 等[71] 扩展 GPS,与免模型的方法结合,提出在环境动态未知情况下的 GPS 算法,采用交替方向乘子法(Alternating Direction Method of Multipliers, ADMM) 对策略和轨迹进行交替迭代优化,以专家轨迹为指导,在优化轨迹的同时,更新策略,使其与优化后的轨迹匹配.

2.3.2　免模型的策略优化方法

在免模型强化学习中,策略一般通过神经网络建模,直接采用信赖域法或线搜索方法进行优化. 线搜索方法是一种迭代求解函数最值的方法,每次迭代分两步进行. 第 1 步确定参数优化的方向,可以采用最速梯度下降法、牛顿法等. 第 2 步确定参数更新的步长,也叫步长搜索,常用方法为回溯线性搜索,即初始化一个较大的更新步长,并令其随训练过程不断衰减. 线搜索的优化对象是策略函数,策略函数来源于函数近似器对策略的近似. 线性函数近似简单,收敛性较好,但是应用十分有限. 而非线性函数近似器收敛性和稳定性较差. Mnih 等[5] 提出将神经网络作为策略函数近似器的 DQN 模型,并在模型中引入经验回放机制,基本原理是使用样本池保存智能体获得的样本,通过在样本池中随机采样获得策略更新依赖的样本. 经验回放机制克服经验数据的相关性和非稳定分布,在一定程度上提高模型的收敛性和稳定性. Lillicrap 等[25] 将 DQN 向连续空间问题扩展,提出 DDPG 模型,DDPG 模型中采用软目标更新(Soft Target Update) 方法,即引入一个策略网络的副本作为最终的目标策略网络,目标网络参数依据当前策略缓慢更新,避免策略更新过快造成的不稳定性. 经验回放和软目标更新方法作为提高策略优化收敛性和稳定性的手段,广泛应用于各种强化学习模型中[72].

信赖域法也是一种迭代优化方法,和线搜索法不同的是,信赖域法需要先确定一个参数更新的信赖域(Trust region),再在该区域内搜索优化方向. Schulman 等[73] 将信赖域法用于策略优化,提出基于信赖域的策略优化算法(Trust Region Policy Optimization, TRPO). TRPO 在信赖域中使用二次模型近似

策略函数,同时使用一个评价变量评价近似的准确度,根据评价结果决定信赖域是否可靠,进一步决定是否更新策略,以及是否需要改变信赖域. 信赖域法能够保证在迭代过程中策略向着好的方向不断更新,保证算法的收敛性,但是计算过程十分复杂. 为此,Schulman 等提出近似策略优化算法(PPO)[74]. PPO 改进 TRPO 的目标函数,使用策略更新前后的 KL 散度作为限制策略更新幅度的约束条件. PPO 直接对策略更新幅度使用上/下界常量进行剪裁,一方面避免复杂的 KL 散度计算,另一方面将约束优化问题变成无约束优化问题,简化后的 PPO 得到与原始 TRPO 近似的性能. Latafat 等[75] 将异步策略的思想与 PPO 融合,提出分布式 PPO,实现方法与异步 AC 类似,只是在 PPO 的基础上将策略执行和优化通过不同的线程异步执行,平衡样本收集和策略优化过程占用的时间资源,提高训练效率.

然而,不论是线搜索法还是信赖域法,只要是基于策略梯度的强化学习方法都存在着策略估计方差太大的问题. 在样本的累积回报中减去基线值(Baseline)[76] 是一种可以降低减小方差而广泛采用的方法,如累积回报为动作-状态值函数 $Q_\pi(a,s)$ 时,可将状态值函数 $V_\pi(s)$ 作为基线,$Q_\pi(a,s)-V(s)$ 也称为优势函数. 演员-评论家结构也因为可以降低策略估计的方差而被很多强化学习算法采纳[77]. 此外,近年来研究人员还提出很多其它用于减少策略估计方差的方法. Fujita 等[78] 根据动作的取值对策略函数进行剪裁,剪裁后的策略函数取值限定在一个固定的范围内,避免极端值的出现,达到减小方差的目的.

3 深度强化学习应用及其面对的挑战

3.1 机器人控制

在机器人控制领域,智能体策略执行及策略评估涉及真实的物理场景,样本的采集非常繁琐和困难. 一方面,从智能体执行策略到获取环境反馈需要消耗时间,策略优化的样本需求往往较大,导致样本采集过程非常繁琐. 另一方面,在虚拟环境中环境反馈可以以反馈信号的形式直接传递给智能体,然而在真实场景中,为了让智能体实现完全的自监督训练,需要赋予智能体自主观察环境反馈并总结奖励的能力,对智能体的感知能力具有相当高的要求. 因此,在机器人控制领域,强化学习的研究重点主要在于提高样本效率,即如何使用少量的样本训练出一个可使用的模型.

受限于强化学习应用的困难,当前的机器人控制任务都非常简单,并且广泛采用真实环境的仿真作为训练的辅助手段. 基本思想是在虚拟环境中训练抓策略,然后基于真实环境中的交互采集样本对策略进行微调. Baier-Löwenstein 等[79] 将强化学习方法应用于机器人的抓取任务中,尝试让机器人抓取不同的物体. Rezzoug 等[80] 提出考虑动作约束的抓取方法,以障碍物的形式在空间上约束机械臂的动作,在抓取物体的同时实现拟人的姿态. Chebotar 等[81] 在抓取任务中使用触觉信息辅助训练,一方面引入抓握稳定性预测器,根据触觉特征预测抓取结果,另一方面最大化预测结果,采取最优的抓取策略,大幅提高抓握过程的成功率. Katyal 等[82] 在机器人任务中加入人类的影响,将机器人的抓取任务和避障任务相互结合,让机器人在实施抓取的同时学习如何避开人类手臂.

强化学习在机器人控制的应用中面临的另一个挑战是模型的鲁棒性问题,深度强化学习本身就受到稳定性较差的困扰,不同的机器人实体对模型的适应性不同,在某个机器人上训练好的强化学习模型如何应用在其它机器人上是一个十分现实而艰巨的挑战. 与此类似的是,很多研究工作都用到仿真环境辅助训练,算法从仿真环境向真实环境中的迁移也面临同样的困难. 先训练再微调的思想不能从本质上解决问题,如何从算法层面提高模型的稳定性依然是当前应用中面临的一大挑战.

3.2 游戏

在游戏任务中,采集样本较容易,各种高性能的深度强化学习算法都获得应用的空间,使得深度强化学习在各类游戏中的表现异常出色. 从最简单的 OpenAI gym 中的倒立摆(CartPole)、过山车(Mountain Car)等简单游戏到 Atari2600 中的太空侵略者(Space Invaders)、打砖块(Breakout),再到诸如 DOTA2、星际争霸的复杂游戏,以及围棋这种状态和动作空间都非常大的策略游戏,强化学习训练的智能体的决策能力已经全面超越顶尖人类玩家. 自从 DQN 开始使用深度强化学习算法训练智能体应对游戏任务,小游戏已经成为各种深度强化学习算法检验基本性能的实验平台[19-20,31,64].

2016 年 AlphaGo 击败李世石,在人工智能领域掀起轩然大波. 对于状态和动作空间都非常大、决策

步非常多的围棋游戏,正向求解策略十分困难,一种可行的方法是蒙特卡洛树搜索算法(Monte Carlo Tree Search, MCTS). 但是 MCTS 受限于估值函数较弱的特征提取能力,很难在复杂的围棋局势下做出专家级的决策. 为此,Silver 等[6]在 AlphaGo 中引入 2 个卷积神经网络,分别对策略和值函数建模,并融合强化学习和监督学习的训练方式对其进行训练,让智能体掌握超越围棋世界冠军的游戏水平. 2017 年 Silver 等[83]更进一步提出 AlphaGo Zero,在 AlphaGo 基础上做出诸多改进,包括将策略和值函数网络整合、引入残差结构以构建更深的网络模型、权值随机初始化等. 相比 AlphaGo, AlphaGo Zero 具有不依赖人类先验、学习时间更短等优点.

此外,2018 年的 DOTA2 世界锦标赛中,OpenAI 基于深度强化学习开发的机器人程序尝试与人类玩家组成的队伍竞技,最终人类玩家以微弱优势取胜,宣告目前强化学习在竞技游戏中的能力极限. 值得一提的是,对于人类玩家来说,要胜任 DOTA2 和星际争霸 2 这样的复杂游戏,需要的不仅仅是简单的短期策略以支撑每一步的判断,同时还需要一些在以往游戏过程中得到的经验总结,用于判断长期游戏局势的发展. 显然,智能体通过深度强化学习获得一定的经验总结能力,这种能力凌驾于策略之上,能够帮助智能体在未知环境中进行长期判断.

3.3　其它方面

除了机器人控制和游戏领域,强化学习在自然语言处理、自动驾驶、检索与推荐系统等领域也有相关应用.

人类驾驶机动车的过程本质上是一个根据路况进行决策的过程,这一决策过程完全可以使用深度强化学习算法实现. Kendall 等[84]将深度强化学习中的 DDPG 模型应用于自动驾驶中,智能体直接接受路况图像作为模型输入,提取特征得到相应的状态,并在此基础上学习转向和加速策略,实现简单的自动驾驶. Sharifzadeh 等[85]将反向强化学习应用于自动驾驶中,以 DQN 为模型拟合大状态空间中的奖励函数,教会智能体如何避免碰撞及一些类人的行为.

在对话系统中,强化学习用于根据对话情景进行决策,选择相应的对话内容. Dhingra 等[86]将强化学习应用于对话系统中,用于训练具有对话能力的个性化智能体. 在传统的硬查询方式中,智能体需要查询知识库以决定下一轮对话的内容,这一过程阻断训练流程,作者使用强化学习中的策略网络替代知识库查询过程,实现端到端(End to End)的训练. 蒙特利尔学习算法研究所(MAIL)[87]为亚马逊 Alexa 竞赛开发的聊天机器人(MAILBOT)程序中也使用强化学习算法.

Derhami 等[88]将强化学习应用于网站排序算法中,提出基于连通性(Connectivity-Based)的网站排序算法. 此外,强化学习中的 Bandit 算法已经广泛应用于各种推荐系统中,Google 公司[89]还用强化学习方法来降低其数据中心的能耗. 目前强化学习的应用场景还非常有限,主要还是受到之前提到的探索-利用困境、稀疏奖励、样本效率较低、模型收敛性和稳定性较差等问题的限制. 然而,决策与行动是人类与外界交互的基本形式,决策问题也是真实世界中广泛存在的问题,在这些问题上,强化学习还存在很大的应用空间.

4　结 束 语

深度强化学习理论发展到今天,从最初的策略搜索和值函数法到元学习、分层强化学习、模仿学习、迁移学习、终身学习等,强化学习催生机器学习领域许多新的研究方向. 一方面是因为深度强化学习理论本身随研究内容的深入而不断丰富;另一方面,新问题的出现迫使深度强化学习理论不断进步.

尽管如此,当前的深度强化学习理论还存在着很多问题,首先是无法避免的探索-利用困境,探索-利用困境不仅仅是强化学习中智能体面对的问题,同样也是人类学习者面对的问题. 当前的研究工作大部分还停留在线性地或者离散地对探索度进行建模,一些值得思考的问题包括:探索度是否存在最优值,探索-利用问题是否能建模为一个最优化问题. 这些问题的难点在于如何评估探索效率,不同于策略和状态,探索效率很难量化. 准确、有效地量化探索效率是一个十分值得尝试的工作,也是从根本上解决探索-利用困境的前提.

其次是深度强化学习和样本采集困难之间的矛盾. 深度强化学习本身是数据驱动的机器学习算法,而在很多问题中,深度学习理论很难发挥作用. 如机器人的运动控制问题中,“试错”的学习方式因为存在时间成本较高、需要人类监督、依赖灵敏的传感器和感知算法等困难而无法投入实际训练,迫使研究者们采用一些限制条件更多、应用空间更小、同时也更复杂的方法. 在这些问题中,无法体现深度学习的优势. 大数据推动深度学习的发展,而深度强化学习

在某些问题上还停留在大数据时代之前. 要突破这一瓶颈,必须打破虚拟与现实之间的鸿沟,这依赖于在虚拟环境中对现实建模技术的发展,目前很多公司和研究机构都在对此进行尝试,如 MIT 针对无人机的训练开发的 Flight Goggles 模拟器、英特尔开发的用于无人驾驶训练的 CARLA 模拟器等.

样本采集困难导致样本数量不足,稀疏奖励导致样本质量不高. 对于样本信息的不足,稀疏奖励提出额外的挑战. 稀疏奖励问题和任务密切相关,对于有些复杂的任务,学习起来本身很困难,需要大量样本和漫长时间,这一点和人类的学习过程类似. 同样是参照人类的学习方式,研究者们提出设置辅助任务、从失败经验中学习等方式,让智能体在学习过程中充分利用环境中的反馈信息以提高学习效率.

元学习和迁移学习是强调将强化学习和变化的环境、变化的任务相结合,是更广泛意义上的强化学习方法. 当前强化学习的目标是通过训练得到一个自主的智能体,该智能体面对的是特定的环境和任务,未来的强化学习势必要在其基础上进行扩展. 在新的方法中,智能体不但能够处理特定的任务,还能根据环境和任务的变化不断调整策略. 对于未接触过的问题,还要掌握一套关于如何学习并解决新问题的方法,这也是未来的人工智能技术需要实现的最终目标之一.

参 考 文 献

[1] ARULKUMARAN K, DEISENROTH M P, BRUNDAGE M, *et al*. A Brief Survey of Deep Reinforcement Learning. IEEE Signal Processing Magazine, 2017, 34(6): 26-38.

[2] HOU J, LI H, HU J W, *et al*. A Review of the Applications and Hotspots of Reinforcement Learning // Proc of the IEEE International Conference on Unmanned Systems. Washington, USA: IEEE, 2017: 506-511.

[3] GOSAVI A . Reinforcement Learning: A Tutorial Survey and Recent Advances. INFORMS Journal on Computing, 2009, 21(2): 178-192.

[4] WATKINS C J C H, DAYAN P. Q-learning. Machine learning, 1992, 8(3/4): 279-292.

[5] MNIH V, KAVUKCUOGLU K, SILVER D, *et al*. Playing Atari with Deep Reinforcement Learning[C/OL]. [2018-12-26]. https://www.cs.toronto.edu/~vmnih/docs/dqn.pdf.

[6] SILVER D, HUANG A, MADDISON C, *et al*. Mastering the Game of Go with Deep Neural Networks and Tree Search. Nature, 2016, 529(7587): 484-489.

[7] JOSH M, YUVAL T, DRUVA T B, *et al*. Learning Human Behaviors from Motion Capture by Adversarial Imitation[C/OL]. [2018-12-26]. https://arxiv.org/pdf/1707.02201.pdf.

[8] FINN C, LEVINE S. Deep Visual Foresight for Planning Robot Motion[C/OL]. [2018-12-26]. https://arxiv.org/pdf/1610.00696.pdf.

[9] LEWIS M, YARATS D, DAUPHIN Y N, *et al*. Deal or no Deal? End-to-End Learning for Negotiation Dialogues[C/OL]. [2018-12-26]. http://aclweb.org/anthology/D17-1259.

[10] WEISZ G, BUDZIANOWSKI P, SU P H, *et al*. Sample Efficient Deep Reinforcement Learning for Dialogue Systems with Large Action Spaces. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2018, 26(11): 2083-2097.

[11] ZHANG H J, ZHAO J, WANG R, *et al*. Multi-objective Reinforcement Learning Algorithm and Its Application in Drive System // Proc of the 34th Annual Conference of IEEE Industrial Electronics. Washington, USA: IEEE, 2008: 274-279.

[12] DERHAMI V, PAKSIMA J, KHAJAH H. Web Pages Ranking Algorithm Based on Reinforcement Learning and User Feedback. Journal of AI and Data Mining, 2015, 3(2): 157-168.

[13] TAN J, ZHANG T N, COUMANS E, *et al*. Sim-to-Real: Learning Agile Locomotion for Quadruped Robots[C/OL]. [2018-12-26]. https://arxiv.org/pdf/1804.10332.pdf.

[14] SUTTON R S, BARTO A G . Reinforcement Learning: An Introduction. Cambridge, USA: The MIT Press, 2017.

[15] POLYDOROS A S, NALPANTIDIS L. Survey of Model-Based Reinforcement Learning: Applications on Robotics. Journal of Intelligent and Robotic Systems, 2017, 86(2): 153-173.

[16] TEREJANU G, SINGLA P, SINGH T, *et al*. Uncertainty Propagation for Nonlinear Dynamic Systems Using Gaussian Mixture Models. Journal of Guidance, Control, and Dynamics, 2008, 31(6): 1623-1633.

[17] KHANSARI-ZADEH S M, BILLARD A. BM: An Iterative Algorithm to Learn Stable Non-linear Dynamical Systems with Gaussian Mixture Models // Proc of the IEEE International Conference on Robotics and Automation. Washington, USA: IEEE, 2011: 2381-2388.

[18] BELLMAN R. On the Theory of Dynamic Programming. Proceedings of the National Academy of Sciences of the United Stated of America, 1952, 38(8): 716-719.

[19] VAN HASSELT H, GUEZ A, SILVER D. Deep Reinforcement Learning with Double Q-Learning[C/OL]. [2018-12-26]. https://arxiv.org/pdf/1509.06461.pdf.

[20] WANG Z Y, SCHAUL T, HESSEL M, *et al*. Dueling Network Architectures for Deep Reinforcement Learning // Proc of the 33rd International Conference on Machine Learning. New York, USA: ACM, 2016: 1995-2003.

[21] HAUSKNECHT M, STONE P. Deep Recurrent Q-Learning for Partially Observable MDPs[C/OL]. [2018-12-26]. https://arxiv.org/pdf/1507.06527.pdf.

[22] BABAEIZADEH M, FROSIO I, TYREE S, *et al*. Reinforcement Learning through Asynchronous Advantage Actor-Critic on a GPU [C/OL]. [2018-12-26]. https://arxiv.org/pdf/1611.06256.pdf.

[23] WANG Z Y, BAPST V, HEESS N, *et al*. Sample Efficient Actor-Critic with Experience Replay[C/OL]. [2018-12-26]. https://

arxiv. org/pdf/1611. 01224. pdf.

[24] GU S X, LILLICRAP T, SUTSKEVER I, *et al*. Continuous Deep Q-Learning with Model-Based Acceleration // Proc of the 33rd International Conference on Machine Learning. New York, USA: ACM, 2016: 2829-2838.

[25] LILLICRAP T P, HUNT J J, PRITZEL A, *et al*. Continuous Control with Deep Reinforcement Learning[C/OL]. [2018-12-26]. https://arxiv. org/pdf/1509. 02971. pdf.

[26] GU S X, LILLICRAP T, GHAHRAMANI Z, *et al*. Q-Prop: Sample-Efficient Policy Gradient with an Off-Policy Critic [C/OL]. [2018-12-26]. https://arxiv. org/pdf/1611. 02247. pdf.

[27] BADIA M V, ADRIA P, MIRZA M, *et al*. Asynchronous Methods for Deep Reinforcement Learning [C/OL]. [2018-12-26]. https://arxiv. org/pdf/1602. 01783. pdf.

[28] BANSAL S, AKAMETALU A K, JIANG F K, *et al*. Learning Quadrotor Dynamics Using Neural Network for Flight Control // Proc of the 15th IEEE Conference on Decision and Control. Washington, USA: IEEE, 2016: 4653-4660.

[29] GILRA A, GERSTNER W. Predicting Non-linear Dynamics by Stable Local Learning in a Recurrent Spiking Neural Network[C/OL]. [2018-12-26]. https://arxiv. org/pdf/1702. 06463. pdf.

[30] NAGABANDI A, KAHN G, FEARING R S, *et al*. Neural Network Dynamics for Model-Based Deep Reinforcement Learning with Model-Free Fine-Tuning[C/OL]. [2018-12-26]. https://arxiv. org/pdf/1708. 02596. pdf.

[31] OH J, SINGH S, LEE H. Value Prediction Network // GUYAN I, LUXBURG U V, BENGIO S, *et al*., eds. Advances in Neural Information Processing Systems 30. Cambridge, USA: The MIT Press, 2017: 6118-6128.

[32] ANDREW B J. An Invitation to Imitation[C/OL]. [2018-12-26]. https://ri. cmu. edu/pub_files/2015/3/InvitationToImitation_3_1415. pdf.

[33] BARTO A G, MAHADEVAN S. Recent Advances in Hierarchical Reinforcement Learning. Discrete Event Dynamic Systems, 2003, 13(4): 341-379.

[34] MIRNWSKI P, PASCANU R, VIOLA F, *et al*. Learning to Navigate in Complex Environments [C/OL]. [2018-12-26]. https://arxiv. org/pdf/1611. 03673. pdf.

[35] PATHAK D, AGRAWALP, EFROS A A, *et al*. Curiosity-Driven Exploration by Self-Supervised Prediction // Proc of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. Washington, USA: IEEE, 2017: 488-489.

[36] KANGASRÄÄSIÖ A, KASKI S. Inverse Reinforcement Learning from Incomplete Observation Data [C/OL]. [2018-12-26]. https://arxiv. org/pdf/1703. 09700. pdf.

[37] AUER P, CESA-BIANCHI N, FISCHER P. Finite-Time Analysis of the Multiarmed Bandit Problem. Machine Learning, 2002, 47(2/3): 235-256.

[38] AUER P. Using Confidence Bounds for Exploitation-Exploration Trade-offs. Journal of Machine Learning Research, 2002, 3: 397-422.

[39] MNIH V, KAVUKCUOGLU K, SILVER D, *et al*. Human-Level Control through Deep Reinforcement Learning. Nature, 2015, 518(7540): 529-533.

[40] FORTUNATO M, AZAR M G, POIT B, *et al*. Noisy Networks for Exploration[C/OL]. [2018-12-26]. https://arxiv. org/pdf/1706. 10295. pdf.

[41] COLAS C, SIGAUD O, OUDEYER P Y. GEP-PG: Decoupling Exploration and Exploitation in Deep Reinforcement Learning Algorithms[C/OL]. [2018-12-26]. https://arxiv. org/pdf/1802. 05054. pdf.

[42] OH J, GUO Y J, SINGH S, *et al*. Self-Imitation Learning[C/OL]. [2018-12-26]. https://arxiv. org/pdf/1806. 05635. pdf.

[43] XU T B, LIU Q, ZHAO L, et al. Learning to Explore with Meta-Policy Gradient[C/OL]. [2018-12-26]. https://arxiv. org/pdf/1803. 05004. pdf.

[44] MEL V, HESTER T, SCHOLZ J, *et al*. Leveraging Demonstrations for Deep Reinforcement Learning on Robotics Problems with Sparse Rewards[C/OL]. [2018-12-26]. https://arxiv. org/pdf/1707. 08817. pdf.

[45] ZEESTRATEN M J A, HAVOUTIS I, SILVÉRIO J, *et al*. An Approach for Imitation Learning on Riemannian Manifolds. IEEE Robotics and Automation Letters, 2017, 2(3): 1240-1247.

[46] ABBEEL P, NG A Y. Apprenticeship Learning via Inverse Reinforcement Learning // Proc of the 21st International Conference on Machine Learning. New York, USA: ACM, 2004. DOI: 10. 1145/1015330. 1015430.

[47] NG A Y, RUSSELL S J. Algorithms for Inverse Reinforcement Learning // Proc of the 17th International Conference on Machine Learning. San Francisco, USA: Morgan Kaufmann, 2000: 663-670.

[48] HO J, ERMON S. Generative Adversarial Imitation Learning[C/OL]. [2018-12-26]. https://arxiv. org/pdf/1606. 03476. pdf.

[49] HO J, GUPTA J K, ERMON S. Model-Free Imitation Learning with Policy Optimization // Proc of the 33rd International Conference on Machine Learning. New York, USA: ACM, 2016: 2760-2769.

[50] DUAN Y, ANDRYCHOWICZ M, STADIE B C, *et al*. One-Shot Imitation Learning[C/OL]. [2018-12-26]. https://arxiv. org/pdf/1703. 07326. pdf.

[51] WANG Z Y, MEREL J, REED S, *et al*. Robust Imitation of Diverse Behaviors[C/OL]. [2018-12-26]. https://arxiv. org/pdf/1707. 02747. pdf.

[52] NICHOL A, ACHIAM J, SCHULMAN J. On First-Order Meta-Learning Algorithms[C/OL]. [2018-12-26]. https://arxiv. org/pdf/1803. 02999. pdf.

[53] AL-SHEDIVAT M, BANSAL T, BURDA Y, *et al*. Continuous Adaptation via Meta-Learning in Nonstationary and Competitive Environments[C/OL]. [2018-12-26]. https://arxiv. org/pdf/1710. 03641. pdf.

[54] ABEL D, ARUMUGAM D, LEHNERT L, *et al*. State Abstractions for Lifelong Reinforcement Learning // Proc of the 35th International Conference on Machine Learning. New York, USA: ACM, 2018: 10-19.

[55] ABEL D, JINNAL Y, GUO Y, *et al*. Policy and Value Transfer in Lifelong Reinforcement Learning // Proc of the 35th International

Conference on Machine Learning. New York, USA: ACM, 2018: 20-29.

[56] FINN C, YU T H, ZHANG T H, *et al*. One-Shot Visual Imitation Learning via Meta-Learning[C/OL]. [2018-12-26]. https://arxiv.org/pdf/1709.04905.pdf.

[57] FINN C, ABBEEL P, LEVINE S. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks[C/OL]. [2018-12-26]. https://arxiv.org/pdf/1703.03400v3.pdf.

[58] RITTER S, WANG J X, KUTH-NELSON Z, *et al*. Been There, Done That: Meta-Learning with Episodic Recall[C/OL]. [2018-12-26]. https://arxiv.org/pdf/1805.09692.pdf.

[59] SUTTON R S, PRECUP D, SINGH S. Between MDPs and semi-MDPs: A framework for Temporal Abstraction in Reinforcement Learning. Artificial Intelligence, 1999, 112(1/2): 181-211.

[60] HAARNOJA T, HARTIKAINEN K, ABBEEL P, *et al*. Latent Space Policies for Hierarchical Reinforcement Learning[C/OL]. [2018-12-26]. https://arxiv.org/pdf/1804.02808.pdf.

[61] FANG K, ZHU Y K, GARG A, *et al*. Learning Task-Oriented Grasping for Tool Manipulation from Simulated Self-Supervision[C/OL]. [2018-12-26]. https://arxiv.org/pdf/1806.09266.pdf.

[62] MURALI A, PINTO L, GANDHI D, *et al*. CASSL: Curriculum Accelerated Self-supervised Learning[C/OL]. [2018-12-26]. https://arxiv.org/pdf/1708.01354.pdf.

[63] ANDRYCHOWICZ M, WOLSKI F, RAY A, *et al*. Hindsight Experience Replay[C/OL]. [2018-12-26]. https://arxiv.org/pdf/1707.01495v1.pdf.

[64] JADERBERG M, MNIH V, CZARNECKI W M, *et al*. Reinforcement Learning with Unsupervised Auxiliary Tasks[C/OL]. [2018-12-26]. https://arxiv.org/pdf/1611.05397.pdf.

[65] RIEDMILLER M, HAFNER R, LAMPE T, *et al*. Learning by Playing-Solving Sparse Reward Tasks from Scratch[C/OL]. [2018-12-26]. https://arxiv.org/pdf/1802.10567.pdf.

[66] DEISENROTH M, RASMUSSEN C E. PILCO: A Model-Based and Data-Efficient Approach to Policy Search // Proc of the 28th International Conference on Machine Learning. New York, USA: Omnipress, 2011: 465-472.

[67] MCALLISTER R, VAN DER WILK M, RASMUSSEN C E. Data-Efficient Policy Search Using PILCO and Directed-Exploration[C/OL]. [2018-12-26]. http://people.eecs.berkeley.edu/~rmcallister/files/epilco.pdf.

[68] GAL Y, MCALLISTER R T, RASMUSSEN C E. Improving PILCO with Bayesian Neural Network Dynamics Models[C/OL]. [2018-12-26]. http://www.cs.ox.ac.uk/people/yarin.gal/website/PDFs/DeepPILCO.pdf.

[69] LEVINE S, KOLTUN V. Guided Policy Search[C/OL]. [2018-12-26]. https://graphics.stanford.edu/projects/gpspaper/gps_full.pdf.

[70] TASSA Y, EREZ T, TODOROV E. Synthesis and Stabilization of Complex Behaviors through Online Trajectory Optimization // Proc of the IEEE/RSJ International Conference on Intelligent Robots and Systems. Washington, USA: IEEE, 2012: 4906-4913.

[71] LEVINE S, ABBEEL R. Learning Neural Network Policies with Guided Policy Search under Unknown Dynamics // GHAHRAMANI Z, WELLING M, CORTES C, *et al*. Advances in Neural Information Processing Systems 27. 2014: 1-9.

[72] HAARNOJA T, ZHOU A, ABBEEL P, *et al*. Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor[C/OL]. [2018-12-26]. https://arxiv.org/pdf/1801.01290.pdf.

[73] SCHULMAN J, LEVINE S, MORITZ P, *et al*. Trust Region Policy Optimization // Proc of the 31st International Conference on Machine Learning. New York, USA: ACM, 2015: 1889-1897.

[74] SCHULMAN J, WOLSKI F, DHARIWAL P, *et al*. Proximal Policy Optimization Algorithms[C/OL]. [2018-12-26]. https://arxiv.org/pdf/1707.06347.pdf.

[75] LATAFAT P, FRERIS N M, PATRINOS P. A New Randomized Block-Coordinate Primal-Dual Proximal Algorithm for Distributed Optimization[C/OL]. [2018-12-26]. https://arxiv.org/pdf/1706.02882.pdf.

[76] GREENSMITH E, BARTLETT P L, BAXTER J. Variance Reduction Techniques for Gradient Estimates in Reinforcement Learning. Journal of Machine Learning Research, 2004, 5: 1471-1530.

[77] DEGRIS T, WHITE M, SUTTON R S. Off-Policy Actor-Critic[C/OL]. [2018-12-26]. https://arxiv.org/pdf/1205.4839.pdf.

[78] FUJITA Y, MAEDA S. Clipped Action Policy Gradient[C/OL]. [2018-12-26]. https://arxiv.org/pdf/1802.07564.pdf.

[79] BAIER-LÖWENSTEIN T, ZHANG J W. Learning to Grasp Everyday Objects Using Reinforcement-Learning with Automatic Value Cut-off // Proc of the IEEE/RSJ International Conference on Intelligent Robots and Systems. Washington, USA: IEEE, 2007: 1551-1556.

[80] REZZOUG N, GORCE P, ABELLARD A, *et al*. Learning to Grasp in Unknown Environment by Reinforcement Learning and Shaping // Proc of the IEEE International Conference on Systems, Man and Cybernetics. Washington, USA: IEEE, 2006, VI: 4487-4492.

[81] CHEBOTAR Y, HAUSMAN K, SU Z, *et al*. Self-Supervised Regrasping Using Spatio-Temporal Tactile Features and Reinforcement Learning // Proc of the IEEE/RSJ International Conference on Intelligent Robots and Systems. Washington, USA: IEEE, 2016: 1960-1966.

[82] KATYAL K, WANG I J, BURLINA P. Leveraging Deep Reinforcement Learning for Reaching Robotic Tasks // Proc of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. Washington, USA: IEEE, 2017:490-491.

[83] SILVER D, SCHRITTWIESER J, SIMONYAN K, *et al*. Mastering the Game of Go without Human Knowledge. Nature, 2017, 550(7676): 354-359.

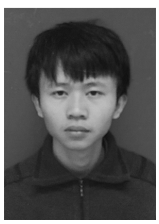
[84] KENDALL A, HAWKE J, JANZ D, *et al*. Learning to Drive in a Day[C/OL]. [2018-12-26]. https://arxiv.org/pdf/1807.00412.pdf.

[85] SHARIFZADEH S, CHIOTELLIS I, TRIEBEL R, *et al*. Learning to Drive Using Inverse Reinforcement Learning and Deep Q-Networks[C/OL]. [2018-12-26]. https://arxiv.org/pdf/1612.03653.pdf.

[86] DHINGRA B, LI L H, LI X J, *et al*. Towards End-to-End Reinforcement Learning of Dialogue Agents for Information Access[C/

OL]. [2018-12-26]. https://arxiv.org/pdf/1609.00777.pdf.
[87] SERHAMI V, SANKAR C, GERMANIN M, *et al*. A Deep Reinforcement Learning Chatbot[C/OL]. [2018-12-26]. https://arxiv.org/pdf/1709.02349v1.pdf.
[88] DERHAMI V, KHODADADIAN E, GHASEMZADEH M, *et al*. Applying Reinforcement Learning for Web Pages Ranking Algorithms. Applied Soft Computing, 2013, 13(4): 1686-1692.
[89] YUAN J J, JIANG X, ZHONG L, *et al*. Energy Aware Resource Scheduling Algorithm for Data Center Using Reinforcement Learning // Proc of the 5th International Conference on Intelligent Computation Technology and Automation. New York, USA: ACM, 2012: 435-438.

作者简介



万里鹏,博士研究生,主要研究方向为深度强化学习、共融机器人. E-mail:xjtuwanlip@126.com.
(**WAN Lipeng**, Ph. D. candidate. His research interests include deep reinforcement learning and coexisting-cooperative-cognitive robots.)

兰旭光(通讯作者),博士,教授,主要研究方向为计算机视觉、机器学习. E-mail:xglan@mail.xjtu.edu.cn.
(**LAN Xuguang** (Corresponding author), Ph. D., professor. His research interests include computer vision and machine learning.)

张翰博,博士研究生,主要研究方向为深度强化学习、机器人控制. E-mail:zhanghanbo163@stu.xjtu.edu.cn.
(**ZHANG Hanbo**, Ph. D. candidate. His research interests include deep reinforcement learning and robot controlling.)

郑南宁,博士,教授,主要研究方向为计算机视觉、模式识别. E-mail:nnzheng@mail.xjtu.edu.cn.
(**ZHENG Nanning**, Ph. D., professor. His research interests include computer vision and pattern recognition.)