计算机科学
Computer Science
ISSN 1002-137X,CN 50-1075/TP

《计算机科学》网络首发论文

题目：深度强化学习中稀疏奖励问题研究综述
作者：杨惟轶，白辰甲，蔡超，赵英男，刘鹏
收稿日期：2019-02-24
网络首发日期：2019-11-22
引用格式：杨惟轶，白辰甲，蔡超，赵英男，刘鹏．深度强化学习中稀疏奖励问题研究综述．计算机科学.
http://kns.cnki.net/kcms/detail/50.1075.TP.20191122.1628.023.html

深度强化学习中稀疏奖励问题研究综述

杨惟轶[1] 白辰甲[2] 蔡超[1] 赵英男[2] 刘鹏[2]
(中国联通网络技术研究院 北京 100048)[1]
(哈尔滨工业大学计算机科学与技术学院 哈尔滨 150001)[2]

摘要 强化学习作为机器学习的重要分支，是在与环境交互中寻找最优策略的一类方法。强化学习近年来与深度学习进行了广泛结合，形成了深度强化学习的研究领域。作为一种崭新的机器学习方法，深度强化学习同时具有感知复杂输入和求解最优策略的能力，可以应用于机器人控制等复杂决策问题。稀疏奖励问题是深度强化学习在解决任务中面临的核心问题，稀疏奖励在实际应用中广泛存在。解决稀疏奖励问题有利于提升样本的利用效率，提高最优策略的水平，推动深度强化学习更加广泛地应用于实际任务。文中首先对深度强化学习的核心算法进行阐述；然后介绍稀疏奖励问题的 5 种解决方案，包括奖励设计与学习、经验回放机制、探索与利用、多目标学习和辅助任务等；最后，对相关研究工作进行总结和展望。

关键词 深度强化学习，深度学习，强化学习，稀疏奖励，人工智能

中图法分类号 TP181 **文献标识码** A **DOI** 10.11896/jsjkx.190200352

Survey on Sparse Reward in Deep Reinforcement Learning

YANG Wei-yi[1] BAI Chen-jia[2] CAI Chao[1] ZHAO Ying-nan[2] LIU Peng[2]
(China Unicom Network Technology Research Institute，Beijing，100048, China)[1]
(School of Computer Science and Technology，Harbin Institute of Technology，Harbin，150001, China)[2]

Abstract As an important research direction of machine learning, reinforcement learning is a kind of method that finding the optimal policy by interacting with the environment. In recent years, deep learning is widely used in reinforcement learning algorithm, forming a new research field named deep reinforcement learning. As a new machine learning method, deep reinforcement learning has the ability to perceive complex inputs and solve optimal policies. It is applied to robot control and complex decision-making problems. The sparse reward problem is the core problem of reinforcement learning in solving practical tasks. Sparse reward problem exists widely in practical applications. Solving the sparse reward problem is conducive to improving the sample-efficiency and the quality of optimal policy, and promoting the application of deep reinforcement learning to practical tasks. Firstly, an overview of the core algorithm of deep reinforcement learning was given. Then five solutions of sparse reward problem were introduced, including reward design and learning, experience replay, exploration and exploitation, multi-goal learning and auxiliary tasks. Finally, the related researches were summarized and prospected.

Keywords Deep reinforcement learning, Deep learning, Reinforcement learning, Sparse reward, Artificial intelligence

1 引言

强化学习[1]是机器学习的重要分支，是在与环境交互中寻找最优策略的一类方法。强化学习求解最优策略的过程非常类似于人类学习的过程，其通过与环境的交互和试错，不断改进自身策略，获取更大的奖励。强化学习与机器学习中的其他领域的主要区别在于，强化学习是一个主动学习的过程，没有特定的训练数据，智能体需要在不断与环境交互的过程中获得样本。

近年来，强化学习与深度神经网络[2]进行了广泛结合，产生了一个交叉领域，被称为深度强化学习（Deep Reinforcement Learning, DRL）[3]。由于深度学习对复杂的高维数据具有良好的感知能力，而强化学习适用于进行策略学习，因此将二者结合产生的 DRL 算法同时具有感知复杂输入和进行决策的能力。基于 DRL 的围棋博弈系统 AlphaGo[4]和 AlphaZero[5-6]近年来连续击败人类围棋顶尖棋手，体现了 DRL 具有广阔的研究

到稿日期：2019-02-24 返修日期：2019-06-20
本文受国家自然科学基金（61671175，61672190）资助。
This work was supported by the National Natural Science Foundation of China (Grant No. 61671175，61672190).

前景和应用价值。DRL 在 2017 年入选“麻省理工科技评论”十大突破性技术之一，被认为是迈向通用人工智能的重要途径。

稀疏奖励问题是 DRL 在解决实际任务中面临的核心问题。在监督学习中，监督信号由训练数据提供。在强化学习中，奖励承担了监督信号的作用，智能体依据奖励进行策略优化。然而，一方面，在开始训练时智能体的策略是随机策略，而奖励的获取需要一系列复杂的操作，因此智能体在初始化策略下很难获得奖励，导致训练困难[7]。另一方面，稀疏奖励在强化学习任务中是广泛存在的。例如，在围棋中，从开始下棋到棋局结束才能判断胜负，此时智能体才能获得奖励，棋局中间过程中的奖励很难评价；在导航任务中，智能体只有在规定时间步内到达指定位置才能得到奖励，中间过程的每一步都是无奖励的；在机械臂抓取任务中，机械臂完成一系列复杂的位姿控制成功抓取目标后才能获得奖励，中间任何一步的失败都会导致无法获得奖励。稀疏奖励问题会导致强化学习算法迭代缓慢，甚至难以收敛。因此，研究如何解决稀疏奖励带来的负面影响，研究稀疏奖励环境下的强化学习算法，对于提升 DRL 的学习速度和策略水平有重要作用。

解决稀疏奖励问题有助于提高样本的利用效率。强化学习算法中代价较高的部分往往不是训练过程，而是样本获取的过程。样本的获取需要智能体与环境进行交互，而交互的代价较高，这一代价不仅体现在时间上，还体现在安全性、可控性、可恢复性等诸多方面[8]。特别是在真实任务，如机器人[9]、自动驾驶[10]任务中，与环境交互有时需要真实的硬件设备，交互过程不仅耗时且具有一定的危险性。如果交互样本无法获得奖励，那么该样本对于算法训练的贡献将很小。因此，如果能在一定程度上解决稀疏奖励问题，就能加速学习过程，减少智能体与环境的交互次数。

从实用角度看，如何将强化学习算法大规模地应用于实际问题是未来的努力方向。目前，虽然强化学习能够在一些任务中取得好的效果，但通常需要较高的代价。例如，围棋系统 AlphaZero 使用 64 个 GPU 和 19 个 CPU 进行训练，使用 2000 个 TPU 进行环境交互和数据生成。导致高昂学习成本的一个重要原因是稀疏奖励问题。由于奖励的稀疏性，智能体需要频繁地与环境交互，经过反复的尝试才能发现如何获取奖励。因此，对稀疏奖励问题及其解决方法的研究能够推动强化学习算法更加广泛地应用于实际问题中。

本文第 2 节介绍深度强化学习的理论基础和核心算法；第 3 节对稀疏奖励问题的解决方法进行阐述；第 4 节对相关研究进行总结和展望；最后总结全文。本文整体结构如图 1 所示。

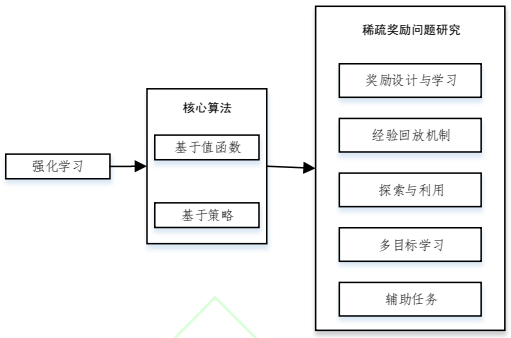

图 1 全文整体结构

Fig. 1 Overall structure of article

2 深度强化学习算法

2.1 基础理论

强化学习的目标是最大化累计奖励。智能体通过从经验中学习，不断优化状态与动作之间的映射关系，最终找到最优策略（policy）。智能体与环境的交互过程可以用马尔可夫决策过程[11]（Markov Decision Processes，MDP）来建模。在周期型任务中，MDP 包括一系列离散的时间步 $0,1,2,...,t,...,T$，其中 T 为终止时间步。在时间步 t，智能体观察环境得到状态的表示 s_t，根据现有策略 π 选择动作 a_t 并执行。执行动作后 MDP 到达下一个时间步 $t+1$，智能体收到奖励 r_{t+1} 并转移到下一个状态 s_{t+1}。回报定义为折扣奖励之和，智能体通过调整策略来最大化回报。动作状态值函数是指智能体处于状态 s 时执行动作 a，随后按照策略 π 与环境交互直到周期结束，所获得的期望回报记为 $Q_\pi(s,a)$。深度强化学习的基本原理如图 2 所示。

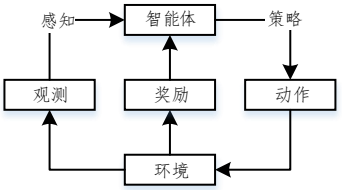

图 2 深度强化学习的原理

Fig. 2 Principle of deep reinforcement learning

最优策略求解一般遵循“广泛策略迭代”的思想，包含策略评价和策略提升。策略评价是已知策略计算值函数的过程，策略提升是已知值函

数选择最优动作的过程。策略评价中值函数 $Q_\pi(s,a)$ 的求解可以使用贝尔曼方程将值函数的求解转化为递归形式 $\sum_{s',r} p(s',r \mid s,a)[r+\gamma \max_{a'} Q_\pi(s',a')]$，其中 γ 是折扣因子。值函数定义了策略空间上的偏序关系，存在最优策略 π^* 优于（或等同于）其他所有策略。强化学习算法可以分为基于值函数的算法和基于策略的算法。

2.2 基于值函数的方法

基于值函数的典型算法是 Q 学习算法，深度 Q 网络[12-13]（DQN）是一种基于 Q 学习的 DRL 算法，其使用深度神经网络来表示动作值函数，进而表示策略。DQN 的主要特点有两个：

（1）使用两个独立的 Q 网络。分别使用 $\boldsymbol{\theta}$ 和 $\boldsymbol{\theta}^-$ 代表 Q 网络和目标 Q 网络的参数，每隔 L 个时间步将 Q 网络的参数复制到目标 Q 网络中，即 $\boldsymbol{\theta}^- \leftarrow \boldsymbol{\theta}$，随后 $\boldsymbol{\theta}^-$ 在 L 个时间步内保持不变。DQN 中的 TD-error 定义为：

$$\delta_t^{DQN} = r_{t+1} + \gamma \max_a Q(s_{t+1}, a; \boldsymbol{\theta}_t^-) - Q(s_t, a_t; \boldsymbol{\theta}_t) \quad (1)$$

使用两个独立的 Q 网络可以使学习的目标值函数分段保持稳定，使训练过程更加鲁棒。损失函数为平方误差损失，$L(\boldsymbol{\theta}_t) = \mathbb{E}_{s,a}[(\delta_t^{DQN})^2]$，用误差反向传播算法迭代更新参数。

（2）使用经验池存储和管理样本，使用经验回放选择样本。样本存储于经验池中，从经验池中随机抽取批量样本训练 Q 网络。使用经验回放可以消除样本之间的相关性，使参与训练的样本满足或近似满足独立同分布。

在 DQN 的基础上，Hasselt 等[14-15]指出，DQN 在计算 TD-error 时使用相同的 Q 网络来选择动作和计算值函数会导致值函数的过高估计，进而提出了 Double DQN 算法（DDQN）。该算法用 Q 网络来选择动作，用目标 Q 网络来估计值函数，从而消除了对值函数的过高估计，并提升了智能体的累计奖励。 DDQN 中的 TD-error 由下式计算：

$$\delta_t^{DDQN} = r_{t+1} + \gamma Q(s_{t+1}, \arg\max_a Q(s_{t+1}, a; \boldsymbol{\theta}_t); \boldsymbol{\theta}_t^-) - Q(s_t, a_t; \boldsymbol{\theta}_t) \quad (2)$$

在此基础上，许多研究对 DDQN 进行了改进。Schaul 等[8]提出了优先经验回放法，优先采样经验池中 TD-error 较大的样本；Horgan 等[16]在此基础上提出了扩展的分布式优先经验回放机制；Wang 等[17]提出了竞争式的网络结构，分别学习状态值函数和优势函数；Bellmare 等[18]指出了学习 Q 函数的不足，提出了学习值函数分布的贝尔曼方程，并证明了其收敛性；Hessel 等[19]将优先经验回放、竞争结构、多步回报[20]、噪声网络[21]等结合起来，提出了一种组合式的算法 Rainbow。

2.3 基于策略的方法

基本的策略梯度法基于 REINFORCE 策略梯度[22]，使用蒙特卡洛方法[23]来估计回报。为了减少方差，在策略梯度法中引入了演员-评论家（Actor-Critic）结构，使用 Critic 来估计值函数，使用 Actor 来输出策略。

2.3.1 策略梯度法

策略梯度法直接优化策略，策略可分为随机策略和确定性策略。REINFORCE 算法用于求解随机策略，在训练中增加高奖励的样本出现的概率，减少低奖励样本出现的概率。对于一个周期的交互序列 $\tau = (s_0, a_0, r_0, s_1, a_1, r_1, ..., s_T)$，策略梯度的计算公式为：

$$g = \nabla_\theta \sum_{t=0}^{T-1} \log \pi(a_t \mid s_t; \theta)(R-b), \quad (3)$$

其中，R 表示周期内奖励之和；b 是用于减少方差的基线，与动作选择无关。策略用深度神经网络表示时，REINFORCE 算法可以使用标准的随机梯度法求解。

确定性策略梯度法（Deterministic Policy Gradient, DPG）[24]是一种基于值函数梯度的方法，适用于求解确定性策略，学习目标是最大化动作值函数。DPG 是一种离策略的学习算法，相对于 REINFORCE 方法具有更高的样本利用效率。

2.3.2 Actor-Critic 方法

Actor-Critic 方法使用 Actor 来学习策略，同时使用 Critic 来估计状态值函数，减少策略估计的方差。异步优势演员评论家方法（A3C）[25]是一种具有代表性的 Actor-Critic 方法，使用多个 Actor 来探索环境，Critic 中值函数 $V(s_t)$ 根据多步累计回报更新，目标函数为：

$$r_t + \gamma r_{t+1} + \cdots + \gamma^{n-1} r_{t+n-1} + \gamma^n V(s_{t+n}), \quad (4)$$

使用目标函数和值函数的平方误差损失进行训练。Actor 使用策略梯度公式更新，梯度计算方法为：

$$g = \nabla_\theta \sum_{t=0}^{T-1} \log \pi(a_t \mid s_t;\theta) A(s_t, a_t;\theta) \quad (5)$$

其中，优势函数的计算过程为：

$$A(s_t, a_t;\theta) = \sum_{i=0}^{k-1} \gamma^i r_{t+i} + \gamma^k V(s_{t+k};\theta_v) - V(s_t;\theta_v) \quad (6)$$

优势函数使用 Critic 来估计值函数。 A3C 算法可以仅在 CPU 上运行，相对于 DQN 等方法的计算代价更小，是一种在策略的学习算法。A3C 可以用于连续动作空间决策问题，如 TORCS 自动驾驶[26]、Mujoco 机器人环境[27]、VizDoom 导航[28]和 3D 迷宫[29]等。在此基础上，Babaeizadeh 等[30]改进了 A3C 算法的计算结构，提出了 GA3C 算法，使用预测队列和训练队列平衡 CPU 和 GPU 的计算资源，提高了算法的运行效率；Espeholt 等[31]针对设置队列导致的预测器和学习器之间参数更新延迟较大的问题，进一步提出了 IMPALA 结构，在多任务环境中取得了较好的效果。

基于置信区间优化（TRPO）[32-33]的方法是另一类策略梯度法，它使用替代损失函数来保证当前策略与旧策略之间的 KL 距离不超过阈值，可以使用约束优化过程求解。REINFORCE 策略梯度和自然策略梯度都可以看作 TRPO 的特例。Schulman 等进一步提出了近似策略优化（PPO）[34]算法，简化了 TRPO 的求解过程，并改进了优势函数的估计方法，使用泛化优势函数估计[35]来平衡优势函数计算的方差和偏差。Nachum 等[36]提出了离策略版的置信区间优化方法，提高了 TRPO 的样本利用效率。

在确定性策略梯度的基础上，Lillicrap 等[37]提出了深度确定性策略梯度法（DDPG），使用 DQN 中的经验回放和目标网络来稳定训练。DDPG 可以用于连续动作空间问题，具有较高的样本利用效率，合用于解决机械臂任务。DDPG 中 Critic 用于最大化 Q 值，Actor 使用 Critic 对动作的梯度进行学习。Fujimoto 等[38]针对 DDPG 中的值函数过估计问题，提出使用两个独立的 Actor 和两个独立的 Critic 来减少系统性的过估计，提升了算法性能。Hausknecht 等[39]将 DDPG 算法扩展用于结构化、参数化的动作空间，在足球游戏 RoboCup[40] 上进行了实验。

此外，仍有一些独立的策略梯度方法。Haarnoja 等[41]提出了平滑 Q 学习（Soft-Q）算法，使用玻尔兹曼分布表示策略，使用最大熵优化策略来提升策略的探索能力；随后，在此基础上进一步提出了平滑演员-评论家算法[42]（Soft Actor-Critic）。Schulman 等[43]通过理论分析指出了基于熵正则的 Q 学习和策略梯度的等价性。Gu 等[44]提出了 Q-prop 算法，结合了策略梯度的稳定性和离策略方法的样本高效性。PGQL[45]和 ACER[46]算法尝试将基于策略和基于值函数的方法进行结合。赵星宇等[47]对相关方法进行了总结。

3 稀疏奖励问题的研究现状

强化学习算法在引入深度神经网络后，对大量样本的需求更加明显。如果智能体在与环境的交互过程中没有获得奖励，那么该样本在基于值函数和基于策略梯度的损失中的贡献会很小。直接使用稀疏奖励样本进行学习有时不仅无法带来策略提升，还会带来负面影响，导致神经网络训练的发散。解决稀疏奖励问题能够使强化学习算法的性能获得普遍提升。目前，针对解决稀疏奖励问题的研究主要包括：奖励设计与学习、经验回放机制、探索与利用、多目标学习和辅助任务。

3.1 奖励设计与学习

一种直观的解决稀疏奖励问题的思路是使用人为设计的“密集”奖励。例如，在机械臂“开门”的任务中，原始的稀疏奖励设定为：若机械臂把门打开则给予“+1”奖励，其余情况下均给予“0”奖励。然而，由于任务的复杂性，机械臂从随机策略开始很难通过自身探索获得奖励。为了简化训练过程，可以使用人为设计的奖励：1）在机械臂未碰到门把手时，将机械臂与门把手距离的倒数作为奖励；2）当机械臂接触门把手时给予“+0.1”奖励；3）当机械臂转动门把手时给予“+0.5”奖励；4）当机械臂完成开门时给予“+1”奖励。这样，通过人为设计的密集奖励，可以引导机械臂完成开门的操作，简化训练过程。

然而，人为设计奖励的方式往往具有局限性：1）人为设计的奖励与任务密切相关，根据任务的不同，奖励设计的方法不具有通用性； 2）人为设计的奖励有时会给学习带来错误的引导，使最终策略收敛到局部最优，给学习带来负面影响。OpenAI[48]分析了人为设计奖励对学习带来的负面影响。Russell 等[49]给出了一个例子：将吸尘器的奖励设定为“吸收灰尘”时，吸尘器会通过先“喷射灰尘”再“吸收灰尘”来获得奖励。类似的负面影响在各种任务中广泛存在。同时，不合理的奖励设计还会使智能体在探索环境中存在

安全隐患。Amodei 等[50]对人工智能中与安全性相关的研究进行了综述，其中涉及由于奖励设计不当而导致的安全性问题。

针对人为设计奖励中存在的问题，Ng 等[51]提出了从最优交互序列中学习奖励函数的思路，此类方法称为“逆强化学习”。由于最优交互序列一般由人类专家给出，因此逆强化学习一般被视为模仿学习（Imitation Learning）的分支。逆强化学习是通过大量专家决策数据在马尔可夫决策过程中逆向求解环境奖励函数的一类方法，其基本原则是寻找一个或多个奖励函数来描述专家决策行为。有 3 种奖励函数的设计形式可以实现求解过程，包括基于最大间隔、基于确定基函数组合以及基于参数化表示。其中，基于参数化的表示方式适用于深度神经网络。

Zibart 等[52]提出了最大熵逆强化学习方法，基于能量函数模型解决了专家决策数据中可能存在的噪声问题。专家轨迹在策略空间中的采样概率可以表示为：

$$p(\tau)=\frac{\exp(-c_\theta(\tau))}{\sum_{\tau}\exp(-c_\theta(\tau))}，(7)$$

其中，τ 为策略轨迹，分母为划分函数 Z。逆强化学习的学习目标是最大化专家轨迹的似然函数，可以表述为：

$$\max mize\ \log\prod_{i=1}^{m}p(\tau_i)=\sum_{i=1}^{m}-\log Z-c_\theta(\tau_i),(8)$$

其中，划分函数 Z 在小规模问题中可以使用动态规划求解，在大规模问题中通过采样实现[53]。

此外，Finn 等[54]提出了在深度神经网络表示的奖励函数下逆强化学习中的代价函数。Hadfield 等[55]提出了近似求解奖励函数的方法，能够避免奖励函数的负面影响。Christiano 等[56]提出根据人类偏好来学习奖励函数，通过选择轨迹来获取人类偏好，使用监督学习的方法来逼近人类偏好。赵凯峰等[57]对早期的逆强化学习方法进行了总结。

3.2 经验回放机制

经验回放机制适用于离策略的学习算法。在深度 Q 网络训练中，智能体与环境交互产生的样本会存储在经验池中，在算法训练时进行采样。在稀疏奖励条件下，经验池中大多数样本没有获得奖励，具有较小的 TD-error。为了提高样本利用效率，Schaul 等[8]提出了优先经验回放法（Prioritized Experience Replay, PER），优先采样经验池中有较大 TD-error 的样本，这些样本能在训练中起到更大的作用。样本优先级定义为：

$$P(i)=\frac{(|\delta_i|+\varepsilon)^\alpha}{\sum_k(|\delta_k|+\varepsilon)^\alpha}，(9)$$

其中，δ_i 代表样本 i 的 TD-error，α 和 ε 为常数，分母起到规约作用。同时，PER 算法采样时在优先级的基础上引入随机性，使样本被抽取的概率与优先级成正比，同时所有样本都有机会被采样，有利用增加样本的多样性。此外，优先级的引入相对于均匀抽样改变了样本的分布，引入了偏差。PER 算法使用重要性采样权重对偏差进行补偿。在计算损失函数梯度时，需要在原有梯度的基础上乘以重要性采样权重，按补偿后的梯度进行更新。优先经验回放的主要过程如图 3 所示。

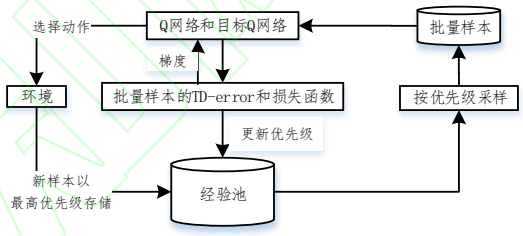

图 3 优先经验回放的原理

Fig. 3 Principle of prioritized experience replay

该方法与深度 Q 学习的结合显著减少了智能体与环境的交互次数，提高了智能体的最优得分，在一定程度上解决了稀疏奖励问题。优先经验回放法具有高度的灵活性，可与离策略的 DRL 算法进行广泛结合。Hou 等[58]将该方法与 DDPG 算法结合，提高了确定性策略梯度法在仿真机器人控制中的数据利用效率。

Tavakoli 等[59]提出将优先经验回放应用于高维动作空间决策中，提出了动作分支的 DQN 结构，能够处理高维连续动作空间问题。Horgan 等[60]提出了分布式的优先经验回放结构，使用多个 Actor 和单个 Learner。多个 Actor 之间的策略参数不同，提高了样本的多样性，使 Learner 能够充分利用 GPU 资源。同时，Actor 和 Learner 都参与到经验回放中优先级的计算，使得经验池中的优先级可以得到及时更新。Bruin 等[61]扩展了优先经验回放的框架，提出了统一的经验选择机制来确定当前经验池中存储哪些样本及如何采样样本，在机械臂实验中提高了样本

利用效率。白辰甲等[62]提出了优先级校正的方法，修正了优先经验回放中优先级更新不及时导致的偏差。

3.3 探索与利用

探索与利用是强化学习中的基本问题。智能体在决策时面临两种选择：利用当前已有的知识选择最优动作，或探索非最优但具有不确定性的动作来获取更多信息。在序列决策中，智能体可能需要牺牲当前利益来选择非最优动作，期望能够获得更大的长期回报。在传统强化学习研究中，探索与利用方法主要包括 ϵ-贪心、概率匹配、UCB[1]、Thompson 采样[63]、贝叶斯探索[64-65]等。这些方法一般在多臂赌博机中进行测试，不适用于大规模连续状态空间的 DRL 任务。

在 DRL 领域中使用的探索与利用方法主要包括两类：基于计数的方法和基于内在激励的方法。其目的是构造虚拟奖励，用于和真实奖励函数共同学习。由于真实的奖励是稀疏的，使用虚拟奖励可以加快学习的进程。

3.3.1 基于计数的方法

基于计数（Count-based）的方法使用状态的访问频率来衡量状态的不确定性，访问次数越少的状态具有越高的新颖性。然而，在 DRL 任务中的状态一般由图像或位姿参数来表示，很难遇到两个完全相同的状态，因此不能简单地用表格式的方法来进行计数。Bellemare 等[66]提出了一种虚拟计数的方式，使用状态空间上的概率生成模型来衡量状态出现的频率，使用信息增益将状态出现的频率转化为虚拟计数，在决策时将虚拟计数作为额外的内在奖励。假设智能体当前状态为 x，智能体之前遇到的状态为 $x_1, x_2, ..., x_n$，概率密度模型记为 ρ，则信息增益（PG）定义为概率模型在使用 x 训练之前对 x 的概率 $\rho_n(x)$ 和使用 x 训练之后对 x 的概率 $\rho'_n(x)$ 之差，表示为：

$$PG_n(x) = \log \rho'_n(x) - \log \rho_n(x) \quad (10)$$

如果增益较大，表明该状态更加新颖。在此基础上，定义虚拟计数为：

$$N_n(x) \approx (e^{PG_n(x)} - 1)^{-1} . \quad (11)$$

虚拟奖励定义为 $r^+ = (N_n(x))^{-1/2}$，在学习中，使用原始奖励和 r^+ 之和进行训练。

Ostrovski 等[67]在此基础上阐述了概率模型的重要性，提出使用 PixcelCNN[68-69]作为概率生成模型，其更适用于处理图像状态。该方法可以与基于值函数和基于策略的方法进行结合，在稀疏奖励环境 Montezuma's Revenge 中能取得较好的效果。Tang 等[70]简化了模型设计，使用哈希表将状态映射为编码，利用编码来对状态进行计数，该方法在 Atari 游戏和连续控制任务中取得了较好的效果。

3.3.2 基于内在激励的方法

基于内在激励的方法主要包括变分信息最大化（VIME）、基于好奇心的探索等。VIME[71]使用贝叶斯神经网络来构建环境模型，使用模型的后验概率分布来衡量信息增益，智能体倾向于探索不确定的环境成分。基于好奇心的探索使用类似的思路，Stadie 等[72]利用在训练过程中构建的环境模型 M 来评估状态的新颖程度。具体地，输入状态 s_t 和动作 a_t 来预测状态 s_{t+1}，预测误差的计算过程为：

$$e(s_t, a_t) = \| \sigma(s_{t+1}) - M(\sigma(s_t), a_t) \|_2 \quad , (12)$$

其中，σ 为状态编码函数。这种方法存在的问题是，许多环境的内部存在随机性，这种随机性往往很难进行建模。环境模型拟合的是一个不稳定的目标，会在内在激励的计算中产生不合理的噪声，在衡量状态新颖程度时会受到环境内部噪声的影响。

针对这一问题，Pathak 等[73]使用逆环境模型（ICM）来获得状态的特征表示，去除环境模型中与动作无关的部分，提高了内在激励的效果。具体地，在构建环境模型的同时构建 ICM，其输入为状态 s_t 和 s_{t+1}，输出为 a_t。ICM 通过学习可以在特征空间中去除与预测动作无关的状态特征，在特征空间中构建环境模型可以去除环境噪声。内在激励的计算过程如图 4 所示。

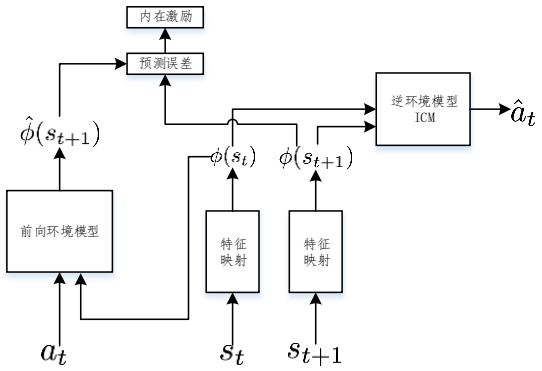

图 4 逆环境模型用于计算内在激励

Fig. 4 Intrinsic motivation computed by inverse environment model

Burda 等[74]进一步研究了基于好奇心的探索，

指出即使完全不使用外在奖励，仅使用内在激励也能取得较好的效果，并给出了在 Atari、自动驾驶、机械臂中的实验结果。因此，如果能够设计合适的内在激励，智能体的学习将能在一定程度上脱离稀疏奖励的制约。Burda 等[75]在此基础上使用随机网络替代环境模型，使状态映射的目标更加稳定，进一步提升了效果。

此外，Fu 等[76]提出将新颖性判断作为分类问题，使用判别式模型进行检测；Fortunato 等[21]提出在参数空间添加噪声用于增加探索；Osband 等[77]提出使用随机先验函数进行探索的方法；Madhavan 等[78]提出结合进化算法鼓励探索；Gupta 等[79]提出使用元学习进行结构化探索的方法。

3.4 多目标学习

强化学习算法中学习的策略 $\pi(s)$ 是状态的函数，根据该策略可以达到单一的目标。例如在某个机械臂任务中，机械臂须将末端放置到指定的位置获得奖励。如果到达该位置，则奖励为+1，否则为 0，这是一个单目标的学习任务。但由于奖励的稀疏性，机械臂从初始化策略开始需要经过不断探索，直到到达指定位置才能获得奖励。Andrychowicz 等[80]提出了一种多目标学习[7]算法，智能体可以从已经到达的位置来获得奖励。该算法在训练中使用虚拟目标替代原始目标，使智能体即使在训练初期也能很快获得奖励，极大地加速了学习过程。

在多目标学习的框架下，需要扩展值函数 $V(s)$ 和 $Q(s,a)$ 的表达能力，使其成为目标 g 的函数。Sutton 等[81]提出了 Horde 结构，该结构将智能体分解为多个子智能体，每个智能体分别使用不同的奖励函数来学习如何完成一个单独的目标。Schaul 等[82]将 Horde 中多个智能体的策略合并为一个整体的策略，提出了基于目标的值函数，扩展了值函数的表达能力。基于目标的值函数表示为 $V(s,g)$ 和 $Q(s,a,g)$，奖励函数表示为 $r(s,a,g)$，策略表示为 $\pi(s,g)$。在稀疏奖励条件下，奖励函数定义为下一个时间步的状态是否到达指定目标，如果到达则奖励为 1，否则奖励为 0，如下式所示：

$$r(s_t,a_t,s_{t+1},g)=1\{d(f(s_{t+1}),g)<\varepsilon\}-1, \quad (13)$$

其中，函数 d 用于计算状态映射后与目标之间的距离；f 为映射函数，一般用于提取状态中特定的维度；ε 是一个常量。

基于目标的值函数不仅能够在状态空间中泛化，还具有在目标空间中泛化的能力。Andrychowicz 等[80]在此基础上，提出了目标经验回放法（Hindsight Experience Replay, HER），使用已经到达的目标来替代原始目标，使智能体能够更快地获得奖励。HER 算法适用于离策略的强化学习算法，与 DDPG 相结合后在多个机械臂测试任务上能达到很好的效果。Rauber 等[83]将 HER 算法与策略梯度法相结合，使用重要性采样进行偏差校正，将多目标机制扩展到在策略算法中。动态目标回放算法[84]（DHER）将 HER 扩展到动态目标，在一个周期内的不同时间步可以改变目标位置，扩展了算法的应用范围。Lanka 等[85]针对 HER 中由于替换目标导致的偏差，通过改变不同目标的奖励权重来校正偏差，获得了一定的效果提升。Nair 等[86]将 HER 算法扩展到以图像表示状态的任务中，使用变分自编码器[87]来获得状态和目标的隐变量表示，在隐空间下进行目标替换和训练，进一步扩展了多目标算法的应用范围。

3.5 辅助任务

在稀疏奖励情况下，当原始任务难以完成时，往往可以通过设置辅助任务的方法加速学习和训练。该方法主要包括两种类型。

第一类方法是“课程式”强化学习。当完成原始任务较为困难时，奖励的获取是困难的。此时智能体可以先从简单的、相关的任务开始学习，此后不断增加任务的难度，逐步学习更加复杂的任务。PowerPlay[88]是一种典型的方法，在原始任务的基础上不断增加新的更复杂的任务，智能体在学习新技能的同时需要不遗忘之前学到的技能。Florsensa 等[89]在此基础上提出通过不断改变智能体出发点的位置，来逐步增加任务的难度。Held 等[90]提出类似的思路，首先让智能体学习如何到达附近的目标，然后再给予更困难的目标，目标的困难程度用当前策略能够获得的奖励来衡量。文中训练了生成对抗网络[91]来建立目标与奖励之间的对应关系，可以使用该网络来生成特定奖励的目标。Sukhbaatar 等[92]提出了一种自我博弈的方式，将智能体分为两部分：A 部分用来提出任务，B 部分用来完成任务。A 部分的奖励函数设置为：

$$R_A=\gamma\max(0,t_B-t_A). \quad (14)$$

如果 B 部分在完成 A 提出的任务时需要更长的时间，则 A 获得更大的奖励。因此，A 部分会不断提出更复杂的任务来使 B 难以完成。B 部分的奖励函数设置为：

$$R_B = -\gamma t_B \,. \quad (15)$$

B 部分通过在更少的时间内完成 A 提出的任务来获得奖励。在不断博弈的过程中，A 和 B 都能不断提高自身水平，并持续探索环境。由于任务的提出和完成均使用内在激励而非外部奖励，因此在稀疏奖励环境下仍可以高效的训练。自我博弈完成后智能体在外部奖励的引导下对策略进行微调，就能够完成复杂的任务。

第二类方法直接在原任务的基础上添加并行的辅助任务，原任务和辅助任务共同学习。使用此类辅助任务的优势在于：1）当原任务奖励稀疏时，智能体可以从辅助任务中获得奖励，从而缓解了稀疏奖励带来的问题； 2）通过训练辅助任务可以使智能体掌握某些技能，这些技能对完成原任务会有帮助；3）辅助任务与原任务在网络层面会共享一部分表示，在训练辅助任务时会促进原任务的网络迭代。Jaderberg 等[93]提出了 UNREAL 框架，在策略梯度法 A3C 的基础上添加了 3 个辅助任务，分别为像素控制、奖励预测和值函数回放，如图 5 所示。

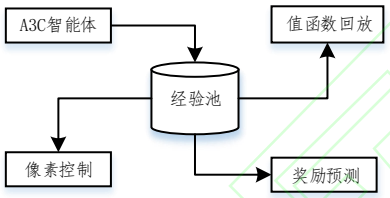

图 5 无监督辅助任务

Fig. 5 Unsupervised auxiliary tasks

图 5 中，像素控制用于最大化智能体观测中像素的变化，智能体只有在迷宫中不断移动才能获得奖励。通过添加辅助任务，智能体在三维地图视觉导航中获得了很好的效果。

Mirowski 等[94]使用类似的方法，在导航任务中添加两个辅助任务：深度图预测和闭环检测。其中，预测深度图的目的是让智能体掌握深度预测的技能，该技能是传统导航中的必备模块；闭环检测的目的是让智能体检测到当前的位置是否曾经到达过，帮助智能体高效地探索环境。这两项辅助任务的添加大大提升了原始算法的性能。Mirowski 等[95]进一步提出了一种实用的导航结构，通过添加方向预测的辅助任务，使智能体在城市导航任务中能够准确预测当前方向与正北方向的夹角。该方法在谷歌街景导航中取得了很好的效果，使得基于强化学习的导航算法[96]首次可以用于真实的街景导航中。该算法的结构如图 6 所示，其中循环神经网络[97]的使用为网络增加了记忆功能。

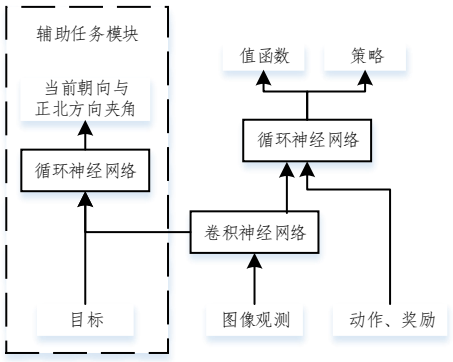

图 6 辅助任务用于视觉导航

Fig. 6 Auxiliary task for visual navigation

4 研究现状和未来研究趋势

4.1 研究现状总结

2015 年，DRL 的标志性成果深度 Q 学习[13]在 *Nature* 发表。此后，各种 DRL 的算法被相继提出。基于值函数的各种方法以 DQN 为基础，不断改进网络结构和训练方法，在 Atari、导航、游戏等一系列任务中达到了人类专家水平。基于策略学习的 DRL 方法在典型策略梯度法的基础上，与深度神经网络相结合，陆续出现了 DDPG[37]、NAF[98]、A3C[25]、TD3[38]、TRPO[32]、ACER[46]、PPO[34]等方法。同时，近年来各种算法陆续出现了并行化的版本，从而加速了训练过程，适用于更加复杂的问题。

由于稀疏奖励问题的广泛存在，在上述算法的基础上研究人员开始致力于解决该问题。该问题最为直观的解决方式是奖励设计与学习。在许多算法的实现中，都针对特定环境进行了人为的奖励设计。然而，由于该方法不具备通用型且可能会带来负面影响，研究人员开始转向从人类专家的交互序列来推导合理的奖励函数。这种逆强化学习方法在 2000 年被首次提出，随后在 DRL 的框架下不断进行改进，成为强化学习和模仿学习的交叉研究领域。

针对离策略的强化学习算法，优先经验回放是一种很好的解决方式。强化学习算法使用经验池存储和回放样本，使用优先经验回放从经验池中优先选择能够产生奖励的样本，从而提高了样本的利用效率。该方法在 DRL 中得到了广泛应用。

探索与利用是强化学习中的传统课题，在 DRL 框架下仍能起到很大的作用。传统的探索与利用方法一般使用多臂赌博机进行研究，使用概率模型求解。然而，在 DRL 任务中这些算法

表现不佳，因此研究人员对这些方法进行改良，使其适用于 DRL 的环境。计算机视觉领域中的概率生成模型在探索与利用中扮演了重要角色，用来评价状态的不确定性。深度神经网络对环境的建模在基于内在激励的求解方法中发挥了重要作用。

多目标学习和辅助任务是最近提出的两种方法。其中多目标学习从独特的角度来看待训练过程，不断给智能体可以到达的目标。随着成功经验的不断积累，智能体的水平不断提高，从而克服稀疏奖励带来的困难。同时，在多目标学习框架下，智能体训练后具备到达任意目标的能力，策略可以在目标空间中进行泛化，相比于传统的 DRL 算法获得了更为通用的技能。辅助任务是一种重要的思想，可以与许多算法进行结合，针对不同的任务可以设置不同的辅助任务。辅助任务的引入增加了奖励的来源方式，克服了稀疏奖励带来的困难。

4.2 未来研究趋势

上述方法在解决稀疏奖励问题中已取得了初步成果，但仍存在不足。在奖励设计与学习中，逆强化学习使用的人类专家轨迹不一定是最优的，会对学习到奖励产生负面影响；在经验回放中，对经验池进行优先级设计会消耗更多的计算资源，且无法扩展到在策略的学习算法；在探索与利用中，内在激励的目的是探索环境中不确定的成分，但是会受到环境中内在随机性的影响；在多目标学习中，目标的选择仍具有局限性，并且目标改变带来的偏差也未完全解决；在辅助任务中，辅助任务产生的奖励有时会在智能体学习过程中占据主导地位，而真实的奖励由于稀疏性往往无法有效指导智能体的学习，如果辅助任务设计不当将会使算法收敛到错误的解。因此，强化学习中稀疏奖励问题的研究虽已取得一定进展，但仍存在挑战，是具有潜力的研究方向。

在未来研究中，如何解决以上方法存在的不足是研究的热点。特别地，多目标学习在 2017 年被提出，在多个机械臂的稀疏奖励任务中取得了很好的效果。同时，多目标学习的实现方法较为直接，可广泛适用于在策略和离策略的强化学习算法。多目标学习与元学习（Meta Learning）[99]和层次化强化学习[100]有相互结合的趋势，是一种实现通用策略的方式，是未来的研究热点。

探索与利用方法是强化学习的基本问题，在强化学习研究早期就已开展研究。在 DRL 的背景下，许多新的探索与利用方法被提出。实验证明，探索与利用对 DRL 算法的性能可以起到关键的作用。目前，探索与利用方法的局限性在于受限于特殊的实验环境，还没有一种方法能够解决所有环境下的稀疏奖励问题。同时，探索与利用方法依赖于概率生成模型，所训练的模型规模庞大，有很大的改良空间。探索与利用作为强化学习的基本问题之一，在今后仍然是研究的热点。

结束语 本文对深度强化学习的核心算法和解决稀疏奖励问题的进展进行了全面分析，并指出了当前研究的不足和未来的研究趋势。在本质上，目前的深度强化学习方法还不具备人类自主思考、决策和推理的能力[101]。随着科技的发展，在未来将会有越来越多的研究成果涌现，推动深度强化学习解决工业生产和日常生活中的问题。

参考文献

[1] SUTTON R S, BARTO A G. Reinforcement learning: An introduction[M]. MIT Press,US, 2018.

[2] LECUN Y, BENGIO Y, HINTON G. Deep learning[J]. nature, 2015, 521(7553): 436.

[3] LI Y. Deep reinforcement learning: An overview[J]. arXiv preprint arXiv:1701.07274, 2017.

[4] 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.

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

[6] SILVER D, HUBERT T, SCHRITTWIESER J, et al. A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play[J]. Science, 2018, 362(6419): 1140-1144.

[7] PLAPPERT M, ANDRYCHOWICZ M, RAY A, et al. Multi-goal reinforcement learning: Challenging robotics environments and request for research[J]. arXiv preprint arXiv:1802.09464, 2018.

[8] SCHAUL T, QUAN J, ANTONOGLOU I, et al. Prioritized experience replay[J]. arXiv preprint arXiv:1511.05952, 2015.

[9] LEVINE S, PASTOR P, KRIZHEVSKY A, et al. Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection[J]. The International Journal of Robotics Research, 2018, 37(4-5): 421-436.

[10] ISELE D, RAHIMI R, COSGUN A, et al.

Navigating occluded intersections with autonomous vehicles using deep reinforcement learning[C]//2018 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2018: 2034-2039.
[11] BELLMAN R. A Markovian decision process[J]. Journal of Mathematics and Mechanics, 1957, 6(5): 679-684.
[12] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Playing atari with deep reinforcement learning[J]. arXiv preprint arXiv:1312.5602, 2013.
[13] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Human-level control through deep reinforcement learning[J]. Nature, 2015, 518(7540): 529.
[14] HASSELT H V. Double Q-learning[C]//Advances in Neural Information Processing Systems. 2010: 2613-2621.
[15] HASSELT H V, GUEZ A, SILVER D. Deep reinforcement learning with double q-learning[C]//Thirtieth AAAI Conference on Artificial Intelligence. 2016.
[16] HORGAN D, QUAN J, BUDDEN D, et al. Distributed prioritized experience replay[C]//International Conference on Learning Representations. 2018.
[17] WANG Z, SCHAUL T, HESSEL M, et al. Dueling Network Architectures for Deep Reinforcement Learning[C]//International Conference on Machine Learning. 2016: 1995-2003.
[18] BELLEMARE M G, DABNEY W, MUNOS R. A distributional perspective on reinforcement learning[C]//International Conference on Machine Learning. 2017: 449-458.
[19] HESSEL M, MODAYIL J, VAN HASSELT H, et al. Rainbow: Combining improvements in deep reinforcement learning[C]//Thirty-Second AAAI Conference on Artificial Intelligence. 2018.
[20] DE ASIS K, HERNANDEZ-GARCIA J F, HOLLAND G Z, et al. Multi-step reinforcement learning: A unifying algorithm[C]//Thirty-Second AAAI Conference on Artificial Intelligence. 2018.
[21] FORTUNATO M, AZAR M G, PIOT B, et al. Noisy networks for exploration[C]//International Conference on Learning Representations. 2018.
[22] PRECUP D, SUTTON R S, DASGUPTA S. Off-policy temporal-difference learning with function approximation[C]//International Conference on Machine Learning. 2001: 417-424.
[23] BROWNE C B, POWLEY E, WHITEHOUSE D, et al. A survey of monte carlo tree search methods[J]. IEEE Transactions on Computational Intelligence and AI in games, 2012, 4(1): 1-43.
[24] SILVER D, LEVER G, HEESS N, et al. Deterministic policy gradient algorithms[C]//International Conference on Machine Learning. 2014.
[25] MNIH V, BADIA A P, MIRZA M, et al. Asynchronous methods for deep reinforcement learning[C]//International conference on machine learning. 2016: 1928-1937.
[26] WYMANN B, ESPIÉ E, GUIONNEAU C, et al. Torcs, the open racing car simulator[J]. Software ,2000, 4(6).
[27] TODOROV E, EREZ T, TASSA Y. Mujoco: A physics engine for model-based control[C]//2012 IEEE/RSJ International Conference on Intelligent Robots and Systems. IEEE, 2012: 5026-5033.
[28] KEMPKA M, WYDMUCH M, RUNC G, et al. Vizdoom: A doom-based ai research platform for visual reinforcement learning[C]//2016 IEEE Conference on Computational Intelligence and Games (CIG). IEEE, 2016: 1-8.
[29] BEATTIE C, LEIBO J Z, TEPLYASHIN D, et al. Deepmind lab[J]. arXiv preprint arXiv:1612.03801, 2016.
[30] BABAEIZADEH M, FROSIO I, TYREE S, et al. Reinforcement learning through asynchronous advantage actor-critic on a gpu[C]//International Conference on Learning Representations. 2017.
[31] ESPEHOLT L, SOYER H, MUNOS R, et al. IMPALA: Scalable Distributed Deep-RL with Importance Weighted Actor-Learner Architectures[C]//International Conference on Machine Learning. 2018: 1406-1415.
[32] SCHULMAN J, LEVINE S, ABBEEL P, et al. Trust Region Policy Optimization[C]//International Conference on Machine Learning. 2015, 37: 1889-1897.
[33] WU Y, MANSIMOV E, GROSSE R B, et al. Scalable trust-region method for deep reinforcement learning using kronecker-factored approximation[C]//Advances in neural information processing systems. 2017: 5279-5288.
[34] SCHULMAN J, WOLSKI F, DHARIWAL P, et al. Proximal policy optimization algorithms[J]. arXiv preprint arXiv:1707.06347, 2017.
[35] SCHULMAN J, MORITZ P, LEVINE S, et al. High-dimensional continuous control using generalized advantage estimation[C]//International Conference on Learning Representations. 2016.
[36] NACHUM O, NOROUZI M, XU K, et al. Bridging the gap between value and policy based reinforcement learning[C]//Advances in

Neural Information Processing Systems. 2017: 2775-2785.
[37] LILLICRAP T P, HUNT J J, PRITZEL A, et al. Continuous control with deep reinforcement learning[C]//International Conference on Learning Representations. 2016.
[38] FUJIMOTO S, HOOF H, MEGER D. Addressing Function Approximation Error in Actor-Critic Methods[C]//International Conference on Machine Learning. 2018: 1582-1591.
[39] HAUSKNECHT M, STONE P. Deep reinforcement learning in parameterized action space[C]//International Conference on Learning Representations. 2016.
[40] STONE P. What's hot at RoboCup[C]//Thirtieth AAAI Conference on Artificial Intelligence. 2016.
[41] HAARNOJA T, TANG H, ABBEEL P, et al. Reinforcement learning with deep energy-based policies[C]//International Conference on Machine Learning. 2017: 1352-1361.
[42] HAARNOJA T, ZHOU A, ABBEEL P, et al. Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor[C]//International Conference on Machine Learning. 2018: 1856-1865.
[43] SCHULMAN J, CHEN X, ABBEEL P. Equivalence between policy gradients and soft q-learning[J]. arXiv preprint arXiv:1704.06440, 2017.
[44] GU S, LILLICRAP T, GHAHRAMANI Z, et al. Q-prop: Sample-efficient policy gradient with an off-policy critic[C]//International Conference on Learning Representations. 2017.
[45] O'DONOGHUE B, MUNOS R, KAVUKCUOGLU K, et al. Combining policy gradient and Q-learning[C]//International Conference on Learning Representations. 2017.
[46] WANG Z, BAPST V, HEESS N, et al. Sample efficient actor-critic with experience replay[C]//International Conference on Learning Representations. 2017.
[47] ZHAO XINGYU, DING SHIFEI. Research on Deep Reinforcement Learning[J]. Computer Science, 2018, 45(7):1-6.(in Chinese)
赵星宇, 丁世飞. 深度强化学习研究综述[J]. 计算机科学, 2018, 45(7):1-6
[48] OPENAI. Faulty Reward Functions in the Wild[EB/OL]. https://blog.openai.com/faulty-reward-functions.2017
[49] RUSSELL S, NORVIG P. Artificial Intelligence A Modern Approach 3rd Edition Pdf[J]. Hong Kong: Pearson Education Asia, 2011.
[50] AMODEI D, OLAH C, STEINHARDT J, et al. Concrete Problems in AI Safety[J]. arXiv preprint arXiv:1606.06565, 2016.
[51] NG A Y, RUSSELL S J. Algorithms for inverse reinforcement learning[C]//Icml. 2000, 1: 2.
[52] ZIEBART B D, MAAS A L, BAGNELL J A, et al. Maximum entropy inverse reinforcement learning[C]//AAAI Conference on Artificial Intelligence. 2008, 8: 1433-1438.
[53] AGHASADEGHI N, BRETL T. Maximum entropy inverse reinforcement learning in continuous state spaces with path integrals[C]//2011 IEEE/RSJ International Conference on Intelligent Robots and Systems. IEEE, 2011: 1561-1566.
[54] FINN C, LEVINE S, ABBEEL P. Guided cost learning: Deep inverse optimal control via policy optimization[C]//International Conference on Machine Learning. 2016: 49-58.
[55] HADFIELD-MENELL D, MILLI S, ABBEEL P, et al. Inverse reward design[C]//Advances in neural information processing systems. 2017: 6765-6774.
[56] CHRISTIANO P F, LEIKE J, BROWN T, et al. Deep reinforcement learning from human preferences[C]//Advances in Neural Information Processing Systems. 2017: 4299-4307.
[57] ZHANG KAIFENG, YU YANG. Methodologies for Imitation Learning via Inverse Reinforcement Learning: A Review[J]. Journal of Computer Research and Development, 2019, 56(2): 254-261.(in Chinese)
张凯峰,俞扬. 基于逆强化学习的示教学习方法综述[J]. 计算机研究与发展, 2019, 56(2): 254-261.
[58] HOU Y, LIU L, WEI Q, et al. A novel DDPG method with prioritized experience replay[C]//2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC). IEEE, 2017: 316-321.
[59] TAVAKOLI A, PARDO F, KORMUSHEV P. Action branching architectures for deep reinforcement learning[C]//Thirty-Second AAAI Conference on Artificial Intelligence. 2018.
[60] HORGAN D, QUAN J, BUDDEN D, et al. Distributed prioritized experience replay[C]//International Conference on Learning Representations. 2018.
[61] DE BRUIN T, KOBER J, TUYLS K, et al. Experience selection in deep reinforcement learning for control[J]. The Journal of Machine Learning Research, 2018, 19(1): 347-402.
[62] BAI CHENJIA, LIU PENG, ZHAO WEI, TANG XIANGLONG. Active Sampling for Deep Q-Learning Based on TD-error Adaptive Correction[J]. Journal of Computer Research and Development, 2019, 56(2): 262-280.(in Chinese)

白辰甲,刘鹏,赵巍,唐降龙. 基于 TD-error 自适应校正的深度 Q 学习主动采样方法[J]. 计算机研究与发展, 2019, 56(2): 262-280.
[63] CHAPELLE O, LI L. An empirical evaluation of thompson sampling[C]//Advances in neural information processing systems. 2011: 2249-2257.
[64] KOLTER J Z, NG A Y. Near-Bayesian exploration in polynomial time[C]//Proceedings of the 26th Annual International Conference on Machine Learning. ACM, 2009: 513-520.
[65] OSBAND I, BLUNDELL C, PRITZEL A, et al. Deep exploration via bootstrapped DQN[C]//Advances in neural information processing systems. 2016: 4026-4034.
[66] BELLEMARE M, SRINIVASAN S, OSTROVSKI G, et al. Unifying count-based exploration and intrinsic motivation[C]//Advances in Neural Information Processing Systems. 2016: 1471-1479.
[67] OSTROVSKI G, BELLEMARE M G, VAN DEN OORD A, et al. Count-based exploration with neural density models[C]//Proceedings of the 34th International Conference on Machine Learning-Volume 70. JMLR. org, 2017: 2721-2730.
[68] VAN OORD A, KALCHBRENNER N, KAVUKCUOGLU K. Pixel Recurrent Neural Networks[C]//International Conference on Machine Learning. 2016: 1747-1756.
[69] SALIMANS T, KARPATHY A, CHEN X, et al. Pixelcnn++: A pixelcnn implementation with discretized logistic mixture likelihood and other modifications[C]//International Conference on Learning Representations (ICLR). 2017.
[70] TANG H, HOUTHOOFT R, FOOTE D, et al. #Exploration: A study of count-based exploration for deep reinforcement learning[C]//Advances in neural information processing systems. 2017: 2753-2762.
[71] HOUTHOOFT R, CHEN X, DUAN Y, et al. Vime: Variational information maximizing exploration[C]//Advances in Neural Information Processing Systems. 2016: 1109-1117.
[72] STADIE B C, LEVINE S, ABBEEL P. Incentivizing exploration in reinforcement learning with deep predictive models[J]. arXiv preprint arXiv:1507.00814, 2015.
[73] PATHAK D, AGRAWAL P, EFROS A A, et al. Curiosity-driven Exploration by Self-supervised Prediction[C]//International Conference on Machine Learning. 2017: 2778-2787.
[74] BURDA Y, EDWARDS H, PATHAK D, et al. Large-scale study of curiosity-driven learning[C]//International Conference on Learning Representations (ICLR). 2019.
[75] BURDA Y, EDWARDS H, STORKEY A, et al. Exploration by random network distillation[C]//International Conference on Learning Representations (ICLR). 2019
[76] FU J, CO-REYES J, LEVINE S. Ex2: Exploration with exemplar models for deep reinforcement learning[C]//Advances in Neural Information Processing Systems. 2017: 2577-2587.
[77] OSBAND I, ASLANIDES J, CASSIRER A. Randomized prior functions for deep reinforcement learning[C]//Advances in Neural Information Processing Systems. 2018: 8626-8638.
[78] CONTI E, MADHAVAN V, SUCH F P, et al. Improving exploration in evolution strategies for deep reinforcement learning via a population of novelty-seeking agents[C]//Advances in Neural Information Processing Systems. 2018: 5032-5043.
[79] GUPTA A, MENDONCA R, LIU Y X, et al. Meta-reinforcement learning of structured exploration strategies[C]//Advances in Neural Information Processing Systems. 2018: 5307-5316.
[80] ANDRYCHOWICZ M, WOLSKI F, RAY A, et al. Hindsight experience replay[C]//Advances in Neural Information Processing Systems. 2017: 5048-5058.
[81] SUTTON R S, MODAYIL J, DELP M, et al. Horde: A scalable real-time architecture for learning knowledge from unsupervised sensorimotor interaction[C]//The 10th International Conference on Autonomous Agents and Multiagent Systems.2011: 761-768.
[82] SCHAUL T, HORGAN D, GREGOR K, et al. Universal value function approximators[C]//International Conference on Machine Learning. 2015: 1312-1320.
[83] RAUBER P, UMMADISINGU A, MUTZ F, et al. Hindsight policy gradients[C]//International Conference on Learning Representations (ICLR). 2019
[84] FANG M, ZHOU C, SHI B, et al. DHER: Hindsight Experience Replay for Dynamic Goals[C]//International Conference on Learning Representations (ICLR). 2019
[85] LANKA S, WU T. ARCHER: Aggressive Rewards to Counter bias in Hindsight Experience Replay[J]. arXiv preprint arXiv:1809.02070, 2018.
[86] NAIR A V, PONG V, DALAL M, et al. Visual reinforcement learning with imagined goals[C]//Advances in Neural Information Processing Systems. 2018: 9209-9220.

[87] KINGMA D P, WELLING M. Auto-encoding variational bayes[J]. arXiv preprint arXiv:1312.6114, 2013.
[88] SCHMIDHUBER J. Powerplay: Training an increasingly general problem solver by continually searching for the simplest still unsolvable problem[J]. Frontiers in psychology, 2013, 4: 313.
[89] FLORENSA C, HELD D, WULFMEIER M, et al. Reverse curriculum generation for reinforcement learning[C]//International conference on Robot Learning.2017
[90] FLORENSA C, HELD D, GENG X, et al. Automatic goal generation for reinforcement learning agents[C]//International Conference on Machine Learning. 2018: 1514-1523.
[91] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets[C]//Advances in neural information processing systems. 2014: 2672-2680.
[92] SUKHBAATAR S, LIN Z, KOSTRIKOV I, et al. Intrinsic motivation and automatic curricula via asymmetric self-play[C]//International Conference on Learning Representations (ICLR). 2018
[93] JADERBERG M, MNIH V, CZARNECKI W M, et al. Reinforcement learning with unsupervised auxiliary tasks[C]//International Conference on Learning Representations (ICLR). 2017
[94] MIROWSKI P, PASCANU R, VIOLA F, et al. Learning to navigate in complex environments[C]//International Conference on Learning Representations (ICLR). 2017
[95] MIROWSKI P, GRIMES M, MALINOWSKI M, et al. Learning to navigate in cities without a map[C]//Advances in Neural Information Processing Systems. 2018: 2424-2435.
[96] PARISOTTO E, SALAKHUTDINOV R. Neural map: Structured memory for deep reinforcement learning[C]//International Conference on Learning Representations. 2018.
[97] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural computation, 1997, 9(8): 1735-1780.
[98] GU S, LILLICRAP T, SUTSKEVER I, et al. Continuous deep q-learning with model-based acceleration[C]//International Conference on Machine Learning. 2016: 2829-2838.
[99] XU Z, VAN HASSELT H P, SILVER D. Meta-gradient reinforcement learning[C]//Advances in Neural Information Processing Systems. 2018: 2402-2413.
[100] NACHUM O, GU S S, LEE H, et al. Data-efficient hierarchical reinforcement learning[C]//Advances in Neural Information Processing Systems. 2018: 3307-3317.
[101] TENENBAUM J. Building machines that learn and think like people[C]//Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems. International Foundation for Autonomous Agents and Multiagent Systems, 2018: 5-5.

Yang Weiyi，born in 1993, Postgraduate, not Member of China Computer Federation (CCF). Her main research interests include machine learning, Internet of things and Reinforcement learning.

Bai Chenjia，born in 1993, PhD, is Member of China Computer Federation (CCF). His main research interests include Reinforcement learning and neural network.