Computer Science and Application 计算机科学与应用, 2022, 12(3), 554-564
Published Online March 2022 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2022.123056

Hans汉斯

强化学习方法的理论与应用研究

林　晨

华南理工大学数学学院，广东　广州

收稿日期：2022年2月12日；录用日期：2022年3月8日；发布日期：2022年3月15日

摘　要

强化学习是机器学习的一个重要分支，是人工智能领域的一大发展方向。本文讨论基于马尔可夫决策过程的强化学习基本框架，对强化学习基本模型进行分析，指出了强化学习的目标，对其中的理论推导进行拆解。文章从理论角度研究了深度强化学习的基础演员/评论家方法(actor-critic)，探讨了确定性策略梯度方法(DPG)的内涵。文章分析了近几年效果良好的双延迟深度确定性策略梯度(TD3)学习方法。文章研究了现阶段强化学习的研究方向与典型方法。文章关注了强化学习的应用，从现阶段强化学习应用领域、强化学习可以处理的问题以及强化学习遇到的挑战等方面分析强化学习，剖析了强化学习的应用现状并对未来发展方向进行了预测。

关键词

人工智能，强化学习，理论，应用

Theoretical and Applied Research on Reinforcement Learning Methods

Chen Lin

School of Mathematics, South China University of Technology, Guangzhou Guangdong

Received: Feb. 12th, 2022; accepted: Mar. 8th, 2022; published: Mar. 15th, 2022

Abstract

Reinforcement Learning is an important branch of machine learning and a major development direction in the field of artificial intelligence. The article discusses the basic framework of Reinforcement Learning based on Markov Decision Process. The article analyzes the basic model, points out the goals and disassembles the theoretical derivation of Reinforcement Learning. The article analyzes actor-critic method from a theoretical perspective which is the basis of Deep Reinforcement Learning and talks about the insight of Deterministic Policy Gradient method. The

文章引用: 林晨. 强化学习方法的理论与应用研究[J]. 计算机科学与应用, 2022, 12(3): 554-564.
DOI: 10.12677/csa.2022.123056

article analyzes Twin Delayed Deep Deterministic policy gradient method that works well in recent years. The article studies the current research direction and typical methods of Reinforcement Learning. The article focuses on the application of Reinforcement Learning and analyzes the uses of Reinforcement Learning from an application perspective of Reinforcement Learning, problems that Reinforcement Learning can solve and the challenges that Reinforcement Learning faces. The article finally analyzes the application status of Reinforcement Learning and predicts the future of Reinforcement Learning.

Keywords

Artificial Intelligence, Reinforcement Learning, Theory, Application

1. 强化学习概述

强化学习是一种机器学习方法，可用于处理机器人控制、路径规划、棋类游戏等现实问题。强化学习不需使用数据集即可完成训练目标，但复杂问题对计算机性能要求高。通俗来讲，强化学习如同小孩子学习问题，成功了获得奖励，失败了受到惩罚，在许多次之后，小孩子就明白了什么该做什么不该做。

强化学习的流程可参考萨顿的强化学习工具书[1]，该书系统性分析了强化学习的过往研究内容。强化学习的基本框架如下：我们首先假设讨论的环境是马尔可夫决策过程(Markov Decision Process, MDP)，也即满足未来状态仅与当前状态有关，而与历史状态无关。在时间 t 时，根据我们训练好的策略 π，依据环境状态 s 选取动作 a，并根据设定的奖励惩罚规则获得奖励或惩罚 r，之后获得新的状态 s'。强化学习的任务是最大化累加和奖励函数，该累加和可以用如下的公式进行表示：$R_t = \sum_{i=t}^{T} \gamma^{i-t} r\left(s_i, a_i\right)$，这其中 $0 < \gamma < 1$，意味着近期奖励更受重视，而远期奖励对现在的影响则小一些。

强化学习的目标是找出最优策略 π_θ，这里 θ 是参数，意思是整体的角度寻求平均期望回报的优化，基本公式如下：$J(\theta) = E_{s\sim P_\pi, a_{i\sim}\pi}\left[R_0\right]$，其中等式右边代表基于状态分布 $p(s)$ 的期望价值 。借助演员/评论家方法[2]，我们则能使用确定性策略梯度方法[3]针对策略 π_θ 以如下的微分方程公式对策略进行优化：$\nabla_\theta J(\theta) = E_{s\sim P_\pi}\left[\nabla_a Q^\pi(s,a)\big|_{a=\pi(s)} \nabla_\theta \pi_\theta(s)\right]$，其中 $Q^\pi(s,a) = E_{s_i\sim p_\pi, a_i\sim p_\pi}\left[R_t \mid s,a\right]$ 称之为评论家或价值函数。关于价值函数的优化，我们可以使用 Q 学习优化方法，近年的领先方法则是一种基于时序差分方法的 Q 学习方法[4]，对价值函数进行更新：$Q^\pi(s,a) = r + \gamma E_{s',a'}\left[Q^\pi\left(s',a'\right)\right]$，其中 $a' \sim \pi(s')$ 意味着策略是具有连续性的。

强化学习从零开始借助与环境的交互进行采样，并逐步优化策略，进而实现智能体的决策优化。强化学习与深度学习结合，则构成了深度强化学习，深度强化学习是现在强化学习研究的主要内容。

2. 演员/评论家方法的内涵

上文提到的演员/评论家方法主要是设立一个演员一个评论家，直观来说，其中演员负责展示强化学习策略的效果，评论家则负责考核效果好坏。演员/评论家方法在 2000 年有了明确论述，研究人员详细讨论了演员/评论家方法的理论基础。

基于仅含演员的强化学习方法与仅含评论家的强化学习方法各有侧重点，集成两部分形成演员/评论家方法，能大大提高算法效率。

演员评论家方法首先论证了该方法的数学原理的合理性，它提出了两个假定：

第一，该方法假定对于状态动作空间下的数对$(s,a)\in(S\times A)$之映射$\theta\to\mu_\theta(s,a)$是二阶可微，并且一阶导有界，这里S指代有限状态空间，A指代有限动作空间，而$\mu_\theta(s,a)$表状态s时采取动作a的概率。另一方面，假定存在实数空间中的n维价值函数$\psi_\theta(s,a)$并满足公式：

$$\nabla\mu_\theta(s,a)=\mu_\theta(s,a)\psi_\theta(s,a) \tag{1}$$

这里映射$\theta\to\psi_\theta(s,a)$是有界的并且对任何确定的状态和动作一阶导数有界。

第二，该方法假定马尔科夫链$\{s_n\}$以及$\{s_n,a_n\}$不可分且非周期，并且在随即平稳策略μ_θ下，通过公式：

$$\eta_\theta(s,a)=\pi_\theta(s)\mu_\theta(s,a) \tag{2}$$

可以求得$\eta_\theta(s,a)$，这里$\pi_\theta(s)$是有关状态的平稳概率。

基于以上假定，当$\mu_\theta(s,a)$不为零的时候可得：

$$\psi_\theta(s,a)=\frac{\nabla\mu_\theta(s,a)}{\mu_\theta(s,a)}=\nabla\ln\mu_\theta(s,a) \tag{3}$$

又设从实数域上n维到1维的平均费用函数：

$$\lambda(\theta)=\sum_{s\in S,a\in A}g(s,a)\eta_\theta(s,a) \tag{4}$$

并通过如下泊松公式：

$$\lambda(\theta)+V_\theta(s)=\sum_{a\in A}\mu_\theta(s,a)\left[g(s,a)+\sum_y p_{sy}(a)V_\theta(y)\right] \tag{5}$$

求解微分费用函数$V_\theta(s)$，此外设Q函数$q_\theta:S\times A\to R$，定义为：

$$q_\theta(s,a)=g(s,a)-\lambda(\theta)+\sum_y p_{sy}(s)V_\theta(y) \tag{6}$$

为了便于计算讨论，定义了两个实值价值函数q_1和q_2之间的函数内积为：

$$q_1,q_{2\theta}=\sum_{s,a}\eta_\theta(s,a)q_1(s,a)q_2(s,a) \tag{7}$$

延续定义，首先有定理1：

$$\frac{\partial}{\partial\theta_i}\lambda(\theta)=\sum_{x,u}\eta_\theta(s,a)q_\theta(s,a)\psi_\theta^i(s,a)=q_\theta,\psi_{\theta\theta}^i \tag{8}$$

其中，$\psi_\theta^i(s,a)$代表ψ_θ的第i个元素。

我们设Ψ_θ为$\left\{\psi_\theta^i;1\le i\le n\right\}$这组参数为$\theta$的基向量的生成向量，另设投影算子$\Pi_\theta:R^{|S||A|}\mapsto\Psi_\theta$，其格式为$\Pi_\theta q=\operatorname{argmin}_{\hat{q}\in\Psi_\theta}\|q-\hat{q}\|_\theta$，有$\langle q_\theta,\psi_\theta\rangle_\theta=\langle\Pi_\theta q_\theta,\psi_\theta\rangle_\theta$。

基于上述讨论，可以得到一种更好的时序差分方法，

$$\lambda_{k+1}=\lambda_k+\gamma_k g(s_k,a_k) \tag{9}$$

$$r_{k+1}=r_k+\gamma_k\left(g(s_k,a_k)-\lambda_k+Q_{r_k}^{\theta_k}(s_{k+1},a_{k+1})-Q_{r_k}^{\theta_k}(s_k,a_k)\right)z_k \tag{10}$$

这里λ_k是正步长参数。

这样我们可以得到演员/评论家方法，其中s^*是状态集中的某个状态。

时序差分方法的不含额外参数评论家方法：$z_{k+1}=z_k+\phi_{\theta_k}\left(s_{k+1},a_{k+1}\right)$，其中$s_{k+1}\neq s^*$；而若$s_{k+1}=s^*$，则$z_{k+1}=\phi_{\theta_k}\left(s_{k+1},a_{k+1}\right)$

时序差分方法的含额外参数评论家方法：$z_{k+1}=\alpha z_k+\phi_{\theta_k}\left(s_{k+1},a_{k+1}\right)$，其中$0\le\alpha<1$。

演员方法：$\theta_{k+1}=\theta_k-\beta_k\Gamma\left(r_k\right)Q_{r_k}^{\theta_k}\left(s_{k+1},a_{k+1}\right)\psi_{\theta_k}\left(s_{k+1},a_{k+1}\right)$，这里$\beta_k$是正步长$\Gamma\left(r_k\right)>0$，是正则化因子满足$\Gamma(\bullet)$李普希兹连续并且存在$c>0$使得$\Gamma\left(r\right)\le\dfrac{c}{1+\|r\|}$。

下面我们讨论算法的收敛性。我们首先进行假定：对于 n 维实数空间中的θ，定义$m\times m$的矩阵$G(\theta)=\sum_{s,a}\eta_\theta(s,a)\phi_\theta(s,a)\phi_\theta(s,a)^{\mathrm{T}}$，并假设$G(\theta)$是一致正定。又假定步长序列$\{\gamma_k\}$以及$\{\beta_k\}$正的并且非增并满足$\delta_k>0,\forall k,\sum_k\delta_k=\infty,\sum_k\delta_k^2<\infty$，其中$\delta_k$代表$\beta_k$或$\gamma_k$，并且假定$\dfrac{\beta_k}{\gamma_k}\to0$。我们可以得到定理 2 和定理 3。

定理 2：

在一个有 TD (1)评论家的演员/评论家算法中，

$$\liminf_k\left\|\nabla\lambda\left(\theta_k\right)\right\|=0,w.p.1. \tag{11}$$

若$\{\theta_k\}$依概率 1 有界，则

$$\lim_k\left\|\nabla\lambda\left(\theta_k\right)\right\|=0,w.p.1. \tag{12}$$

定理 3：

对任意$\varepsilon>0$，有足够接近 1 的α，使得

$$\liminf_k\left\|\nabla\lambda\left(\theta_k\right)\right\|\le\varepsilon,w.p.1 \tag{13}$$

关于具体的方法，形式如下：

$$Q_r^\theta\left(s,a\right)=\sum_{j=1}^m r^j\varnothing_\theta^j\left(s,a\right) \tag{14}$$

其中$r=\left(r^1,\cdots,r^m\right)\in R^m$代表评论家的参数向量，特征$\varnothing_\theta^j,j=1,\cdots,m$是依赖于演员的参数向量$\theta$并用于评论家。

最终我们可以将演员的参数更新方式确定为如下形式：

$$\theta_{k+1}=\theta_k-\beta_k\Gamma\left(r\left(\theta_k\right)\right)Q_{r(\theta_k)}^{\theta_k}\left(s_{k+1},a_{k+1}\right)\psi_{\theta_k}\left(s_{k+1},a_{k+1}\right)+\beta_k e_k \tag{15}$$

其中e_k是渐进可忽略误差。

使用上述的参数更新公式更新参数，就可以逐步优化智能体采用的策略。演员/评论家方法是深度强化学习的基础框架，为之后的深度强化学习框架的提出打下基础。

3. 确定性策略梯度方法的内涵

确定性策略梯度方法是针对策略梯度方法、演员/评论家方法以及异策演员/评论家方法的更新。

如前文所述，基于马尔可夫决策过程的强化学习方法可以用求解平均期望回报策略优化，可为如下形式：

$$J\left(\pi_\theta\right)=\int_S p^\pi\left(s\right)\int_A\pi_\theta\left(s,a\right)r\left(s,a\right)\mathrm{d}a\mathrm{d}s=E_{s\sim p^\pi,a\sim\pi_\theta}\left[r\left(s,a\right)\right] \tag{16}$$

其中 $p^{\pi}(s)$ 是有关策略参数的状态分布。

关于随机梯度下降定理，最早是计算智能研究员萨顿于 1999 年提出的，基本公式如下：

$$\nabla_{\theta} J(\pi_{\theta}) = \int_S p^{\pi}(s) \int_A \nabla_{\theta} \pi_{\theta}(a|s) Q^{\pi}(s,a) \mathrm{d}a\mathrm{d}s = E_{s\sim p^{\pi}, a\sim \pi_{\theta}} \left[\nabla_{\theta} \log \pi_{\theta}(a|s) Q^{\pi}(s,a) \right] \tag{17}$$

关于随机演员/评论家方法，实际应用中，主要是分设两个不同的组分：一个是演员，要做的是调整真实的根据待求的随机策略 $\pi_{\theta}(s)$，针对其中的 θ 进行参数优化；另一个是评论家，它做的事情是评价策略性能的好坏。

关于异策演员/评论家方法，求其平均期望回报的公式有变化：

$$J_{\beta}(\pi_{\theta}) = \int_S p^{\beta}(s) V^{\pi}(s) \mathrm{d}s = \int_S \int_A p^{\beta}(s) \nabla_{\theta} \pi_{\theta}(a|s) Q^{\pi}(s,a) \mathrm{d}a\mathrm{d}s \tag{18}$$

微分形式如下：

$$\begin{aligned} \nabla_{\theta} J_{\beta}(\pi_{\theta}) &\approx \int_S \int_A p^{\beta}(s) \nabla_{\theta} \pi_{\theta}(a|s) Q^{\pi}(s,a) \mathrm{d}a\mathrm{d}s \\ &= E_{s\sim p^{\beta}, a\sim \beta} \left[\frac{\pi_{\theta}(a|s)}{\beta_{\theta}(a|s)} \nabla_{\theta} \log \pi_{\theta}(a|s) Q^{\pi}(s,a) \right] \end{aligned} \tag{19}$$

这其中，$\beta(a|s) \neq \pi_{\theta}(a|s)$ 是行动策略。

以上是确定性策略梯度的前述内容，而关于确定性策略梯度的更新方向，分为同策确定性演员/评论家方法以及异策确定性演员/评论家方法。

同策演员/评论家方法的参数更新采用如下公式计算：

$$\delta_t = r_t + \gamma Q^w(s_{t+1}, a_{t+1}) - Q^w(s_t, a_t) \tag{20}$$

$$w_{t+1} = w_t + \alpha_w \delta_t \nabla_w Q^w(s_t, a_t) \tag{21}$$

$$\theta_{t+1} = \theta_t + \alpha_{\theta} \nabla_{\theta} \mu_{\theta}(s_t) \nabla_a Q^w(s_t, a_t)\big|_{a=\mu_{\theta}(s)} \tag{22}$$

异策演员/评论家方法的平均期望回报采用如下公式计算：

$$J_{\beta}(\mu_{\theta}) = \int_S p^{\beta}(s) V^{\mu}(s) \mathrm{d}s = \int_S p^{\beta}(s) Q^{\mu}(s, \mu_{\theta}(s)) \mathrm{d}s \tag{23}$$

$$\begin{aligned} \nabla_{\theta} J_{\beta}(\mu_{\theta}) &\approx \int_S p^{\beta}(s) \nabla_{\theta} \mu_{\theta}(a|s) Q^{\mu}(s,a) \mathrm{d}s \\ &= E_{s\sim p^{\beta}} \left[\nabla_{\theta} \mu_{\theta}(s) \nabla_a Q^{\mu}(s,a)\big|_{a=\mu_{\theta}(s)} \right] \end{aligned} \tag{24}$$

而异策演员评论家方法的参数更新采用如下公式计算：

$$\delta_t = r_t + \gamma Q^w(s_{t+1}, \mu_{\theta}(s_{t+1})) - Q^w(s_t, a_t) \tag{25}$$

$$w_{t+1} = w_t + \alpha_w \delta_t \nabla_w Q^w(s_t, a_t) \tag{26}$$

$$\theta_{t+1} = \theta_t + \alpha_{\theta} \nabla_{\theta} \mu_{\theta}(s_t) \nabla_a Q^w(s_t, a_t)\big|_{a=\mu_{\theta}(s)} \tag{27}$$

而关于相似异策确定性演员/评论家方法，形式如下：

$$\delta_t = r_t + \gamma Q^w(s_{t+1}, \mu_{\theta}(s_{t+1})) - Q^w(s_t, a_t) \tag{28}$$

$$\theta_{t+1} = \theta_t + \alpha_{\theta} \nabla_{\theta} \mu_{\theta}(s_t) \left(\nabla_{\theta} \mu_{\theta}(s_t)^{\mathrm{T}} w_t \right) \tag{29}$$

$$w_{t+1} = w_t + \alpha_w \delta_t \varnothing(s_t, a_t) \tag{30}$$

$$v_{t+1} = v_t + \alpha_v \delta_t \varnothing(s_t) \tag{31}$$

确定性策略梯度方法是后续发展论文的很重要的论述基础，因为之前的方法是使用随机策略梯度方法，之后的深度确定性策略梯度方法以及 TD3 方法都是基于确定性策略梯度方法提出的。

4. TD3 针对过往方法的更新

TD3 的全称是“Twin Delayed Deep Deterministic policy gradient algorithm”，是文章《Addressing function approximation error in actor-critic methods》[5]中提出的方法，是一种近年应用较为广泛的深度强化学习方法。TD3 相比于深度确定性策略梯度方法有深度神经网络的加入。

该方法针对过往的强化学习更新方向有四个。

首先是针对演员/评论家方法中的过估计误差问题的改进。TD3 将原有的演员/评论家方法中的价值函数 $y = r + \gamma \max_{a'} Q(s', a')$ 参考双 Q 学习，将学习目标变成如下形式：

$$y_1 = r + \gamma Q_{\theta_2}\left(s', \pi_{\theta_1}(s')\right) \tag{32}$$

$$y_2 = r + \gamma Q_{\theta_1'}\left(s', \pi_{\theta_2}(s')\right) \tag{33}$$

同时考虑到演员/评论家方法的高估价值函数问题，TD3 方法直接选取如下形式的价值函数：

$$y_1 = y_2 = r + \gamma \min_{i=1,2} Q_{\theta_i'}\left(s', \pi_{\theta_1}(s')\right) \tag{34}$$

优化的第二个方向是 TD3 借用了 TD 误差思想，将价值函数的更新加入 TD 误差变为如下形式：

$$Q_\theta(s,a) = r + E\left[Q_\pi(s', a')\right] - \delta(s,a) \tag{35}$$

对上述价值函数，迭代计算方法如下：

$$Q_\theta(s_t, a_t) = E_{s_i \sim p_\pi, a_i \sim \pi} \sum_{i=t}^{T} \left[\gamma^{i-t}(r_i - \delta_i)\right] \tag{36}$$

优化的第三个方向是进行目标网络优化，更新思路是在 TD 误差小的情况下去更新目标网络的参数 θ：

$$\theta' \leftarrow \tau\theta + (1-\tau)\theta' \tag{37}$$

优化的第四个方向是目标值的再优化，具体是借用近似策略有近似目标值的思想，加入扰动项：

$$y = r + E_\in\left[Q_{\theta'}\left(s', \pi_{\theta'}(s') + \in\right)\right] \tag{38}$$

目标更新加入噪声之后计算公式如下：

$$y = r + \gamma Q_{\theta'}\left(s', \pi_{\theta'}(s') + \in\right) \tag{39}$$

$$\in \sim clip\left(N(0,\sigma), -c, c\right) \tag{40}$$

可以看出，TD3 方法基于过往方法的问题从多个角度进行优化，理论推导是严谨的。事实上，在经过 1million 时间步之后，对于 Openai Gym 的 MuJoCo 实验环境中，TD3 在最大平均返回值上均是最高，优于 DDPG、PPO、TRPO、ACKTR、SAC 等强化学习方法。

根据实验结果，TD3 方法可用于连续空间的控制问题，在机器人控制、平面游戏挑战等方面都可以应用。

5. 强化学习最新研究方向概述

前文详细剖析了强化学习在过往研究中的经典方法。本部分将重点关注近年强化学习领域的最新研究方向进行分析。值得关注的重要方向包括了可解释强化学习，序列建模强化学习和样本利用效率高的强化学习。

5.1. 可解释强化学习

可解释强化学习(Explainable Reinforcement Learning)是强化学习的最新研究方向，它属于可解释人工智能(Explainable Artificial Intelligence)的范畴。事实上，机器学习具有“黑盒”性质，即人类将越来越多的问题交给深度神经网络(Deep Neural Network)解决，人们却不了解其中的训练逻辑与原理。考虑到人类已经将机器学习应用到生活的各个方面，人类对机器学习“黑盒”性质的担忧是存在的。为了解决人类的担忧，可解释人工智能就诞生了。可解释人工智能旨在将“黑盒”机器学习进行科学分析，论证各种机器学习算法的可信任性与可依赖性。而强化学习作为机器学习重要分支，也产生了可解释强化学习相关研究。

有 2021 年的综述文章[6]系统性分析了过往可解释强化学习的研究成果，并对未来可解释强化学习的发展方向进行预测。

具体的可解释强化学习有若干方向研究。有 2020 年的文章[7]提出了使用因果模型对各种无模型强化学习方法进行因果解释，并使用编码方法对有关联的变量进行架构。另一篇 2020 年的文章[8]提出了不同的架构，该框架基于智能体与环境的交互的视频段，分析智能体感兴趣的元素从而解释智能体的行为，该方法有助于人类理解智能体何时会更新策略这一关键点。有文章[9]结合了决策树与因果模型进行建模，该方法使用循环神经网络(Recurrent Neural Network)学习机会链，并借助决策树提高准确性。另有文章[10]结合了其他学科的思路，使用突变和重组的思想设计了新的使用虚拟神经程序员，具有较高的可解释性。关于强化学习应用的可解释性，有文章[11]使用概率模型对应用强化学习的机器人行为进行解释，并对机器人决策行为的合理性进行论证。

5.2. 序列建模强化学习

自然语言处理(Natural Language Processing)是近年有突破性进展的领域，其中自注意力架构因其能在长时间范围内整合信息并进行扩展而有效地提升了自然语言处理能力。最典型且应用广泛的架构是 Transformer 架构，这个架构是 2017 年的文章[12]中提出的。将 Transformer 及其类似架构进行适当转换或是改变应用于强化学习方向是研究人员的目标，而与之相关的方法就是序列强化学习。

有 2020 年的文章[13]论证了在强化学习环境中 Transformer 原始架构是难以优化的，文章中提出的方法对原始 Transformer 架构进行了修改，显著提高了其稳定性和学习速度。基于前述研究，2021 年研究人员在一篇文章[14]中提出了新的建模思路，直接将强化学习转换为序列建模思路，借助 Transformer 架构对状态、动作、奖励等进行系统化建模，并在模仿学习、目标条件强化学习以及离线强化学习领域取得了良好效果。另一篇 2021 年的文章[15]大幅修改 Transformer 架构，提出了 Decision Transformer 架构，该方法在实验效果方面达到或是超过了最先进的无模型离线强化学习方法。

5.3. 样本利用效率高的强化学习

强化学习的训练通常是无样本开始的，其数据集是通过步步采样得到的。当我们要提升强化学习训练效率时，提高强化学习采样样本利用率是十分重要的。基于提升样本利用率的必要性，近年相关研究较多。

2021 年的文章[16]针对样本高效学习提出了一种结合原型表示的强化学习方案，文章提出了一种自我监督的强化学习框架，并将之应用于复杂的连续控制任务实现了加速下游策略探索的目标。另一篇 2021

年的文章[17]使用未标记的数据预训练编码器并对数据进行微调以应用于不同任务；该方法还采用了隐动力学与无监督目标条件强化学习组合方法以更好地利用样本。有文章[18]分析了过往数据增强方法的瓶颈，引入了一种叫软数据增强(Soft Data Augmentation, SODA)方法；文章提出的方法对编码器使用软约束，目标是最大化增强和非增强数据的隐表示之间的互信息。

6. 强化学习应用分析

6.1. 强化学习常用框架

强化学习的方法需要进行测试，测试方法一般均为线上模拟的形式，常见的环境包括 Gym、CopellaSim、Pybullet 等。

Gym 是最广泛应用的强化学习测试框架，该框架为 Openai 团队制作的，Gym 环境[19]包括了各种强化学习的测试环境如 Atari 游戏测试环境，MuJoCo 机器人测试环境等。Gym 环境为开源环境，是强化学习新方法的一般性验证环境。其中的 Atari 游戏环境主要是二维界面的游戏测试，一般设定明确的目标。MuJoCo 机器人测试环境包括机器人的移动以及机械臂的抓取等操作。

CoppeliaSim 是专门为机器人相关实验建立的非开源模拟器，在机器人领域有一定的应用，可进行强化学习方法的实验模拟，但接口相对 Gym 的 MuJoCo 环境更为复杂。

Pybullet 为力学模拟、机器人控制相关的深度强化学习实验框架，实验设计复杂度介于 Gym 环境与 CoppeliaSim 环境之间。

6.2. 强化学习能处理的问题

成熟的强化学习方法如 TD3 可以应用于复杂的连续空间问题，但一般而言，强化学习针对高精度问题的求解并不理想，需要提出更好的强化学习方法或者将强化学习与其他机器学习方法进行融合。强化学习能解决多种问题，应用广泛。

Gym 框架下的不同环境是强化学习能处理的问题。强化学习的基本验证框架是 Gym，各种基础性算法如 actor-critic 算法、DPG 算法等多是基于 Gym 的基本环境如连杆问题、小车上山问题、平衡问题等进行效果验证。近期算法如 TD3 等广泛应用的算法则使用 Atari 游戏，Mujoco 机器人环境等进行对比实验。

棋类游戏是强化学习能解决的。强化学习能处理的最经典问题是在围棋。DeepMind 研究团队在此方面多有研究，AlphaGo 是其初始版本，之后则提出了 AlphaGo Zero 与 AlphaZero，其提出的最新方法则是 Mu-zero 方法[20]。Mu-zero 方法的特点是不仅能完成围棋方面由于 AlphaZero 的更好效果，也能在将棋与 Atari 游戏方面有突出效果。另有人提出了较为简单的五子棋强化学习方法[21]，该方法使用蒙特卡洛树结合卷积神经网络，训练要求低且效果好。

机器人领域是强化学习可发展的领域，这主要是由于强化学习的性质决定的。强化学习不需要人为标注数据，这样强化学习就有自己学习数据表达方式的能力。从另一方面讲，强化学习需要的是探索未知空间，这与机器人的应用有相似之处。关于机器人控制方面，基本思路是将原问题进行分治，有 2019 年的研究[22]对现实中困难的机器人控制问题进行剖析，将其分解为可用常规反馈控制解决的子问题以及可以使用强化学习解决的子问题两部分，最终通过两个控制信号的叠加解决原问题。关于机器人手眼标定方面，2022 年的最新研究[23]提出了较为完善的强化学习辅助方法，能完成标定任务。

自动驾驶领域是强化学习的最前沿一个研究方向。一篇 2021 年的综述文章[24]系统叙述了深度强化学习算法已经应用的自动驾驶任务，分析了现阶段强化学习遇到的关键挑战。具体到应用方法，一篇 2021 年的文章[25]提出了强化学习与监督学习结合的方法，该方法对先验知识进行编码，并使用图神经网络(Graph Neural Network)对不同车辆的相互影响建模；这种方法在 T 形交叉口性能优于最先进的方法。有文章[26]提

出了一种端到端的深度强化学习方法，该方法引入序列化隐空间模型，并在拥挤的城市场景中有好效果。

6.3. 强化学习的挑战

强化学习应用范围广，数学论证基础完善，思路直观。但强化学习也有多方面的挑战，包括采样效率低、回报值稀疏、输入样本噪声多等问题。

不妨以具体场景为例探讨强化学习的多种挑战。

马里奥游戏在世界范围广受欢迎，这是一种基本符合有限马尔可夫决策过程(Finite Markov Decision Process, FMDP)的游戏环境。所谓基本符合 FMDP 过程，指的是智能体在上一帧与当前一帧是有关系的，但同时考虑到马里奥的具体运动方式，譬如大跳与右上跳，这样的动作都是有惯性的；换句话讲，在某些情境下，智能体不进行操作也会延续之前的运动方式，而这有时是无法仅通过上一帧的情景与动作决定之后的行进方式。马里奥游戏相比于 Atari 游戏虽说都是二维场景下，但是更为复杂。马里奥游戏的单局基本时长是 300 秒，这导致了在单次从出发点到终止的采样过程是漫长的。马里奥游戏的直观奖励就是到终点与否，而智能体未到终点则不会有奖励。

关于上述问题的解决马里奥游戏在现在的解决方法主要还是依靠搭建特定平台进行实验。最常见的是在 Github 上开源的基于 Openai 的 gym 库的 Mario-bros 环境。简要说来，该环境对于原始游戏进行简化，标记了相应的智能体与可交互物体，并对函数的奖励与惩罚有了明确的定义。该环境相比于原始游戏进行了加速处理，一般最长在三十秒内完成一次游戏。这样做产生了新的问题：首先我们难以找到最合适的回报函数定义，因为原始游戏并没有这样的额外回报值；其次虽说游戏加速了十倍左右，但相比于一般情况的需要至少 40000 次游戏的强化学习实验采样，耗费时间仍然很长。事实上，采样耗费了大量的时间，而真正计算的时间并不多，也就是说 GPU 算力利用效率是不高的。

与此形成对比的围棋，为何效果就会很好呢？一方面围棋是状态全可见的，就是 fully observed MDP，这样状态可知，动作可知，且这是严格意义的马尔可夫决策过程；另一方面围棋的自对弈交互过程能达到非常快的情形，没有马里奥游戏那样长的采样时间。除此之外，马里奥游戏的干扰项是存在的，譬如其背景颜色、非固定模式的怪物移动，马里奥游戏还是当前状态部分可见的，意味着即将到来的环境是不可知的。

通过马里奥游戏与围棋的对比，我们不难看出解决一个复杂些的强化学习问题的解决的困难之处：在连续的环境下，首先需要对问题进行建模，建构的框架需要基本符合马尔可夫决策过程；需要缩短单次采样的时间，使得在短期内能获得更多的采样数据；对于回报函数稀疏的长过程问题，得到回报所需的操作数过多，在过程中要增添过程回报值，逐步训练智能体；若是环境状态非完全可见且随机变化，则训练出良好的策略是困难的。

6.4. 强化学习的发展前景

强化学习的基本框架是从零开始学习的，这种性质使得它对于棋类游戏、简单机械臂导引标定等框架确定，目标直接的环境有良好效果。

未来强化学习发展方向可分为三点：

1) 数学推导的完善。强化学习数学推导严谨，但复杂的环境或是非马尔可夫决策过程的场景无法进行严格的数学推导，需要提出完整的数学论述。强化学习调参困难，需要给出一套行之有效逻辑严谨的调参方法。

2) 与其他机器学习方法的结合。强化学习从零开始，这与一般智能体的成长逻辑不符，需要借助迁

移学习等机器学习方法预先让智能体有知识可参照。深度学习与强化学习的结合可提高强化学习的采样与利用样本能力。

3) 针对特定问题的突破。强化学习在围棋方面的突破使得强化学习广受关注，也刺激了强化学习领域的发展，新领域的突破有发展空间。未来强化学习会在自动驾驶、船舶导航、医疗手术、机器人控制等领域有更多突破。

7. 总结

本文对强化学习的理论进行了剖析，分析了其中的数学推理合理性，讨论并分析了强化学习的广泛应用的基础方法，着重关注了其中的核心表达形式。文章分析了当下强化学习的主要研究方向及其典型方法。文章分析了当下强化学习的具体应用领域，对强化学习应用中的共同点与存在的问题进行了论述。文章基于现阶段强化学习的发展提出了自己的分析，并对未来强化学习的发展方向进行了展望。

参考文献

[1] Sutton, R.S. and Barto, A.G. (2018) Reinforcement Learning: An Introduction. MIT Press, Cambridge, 54-93.

[2] Konda, V.R. and Tsitsiklis, J.N. (2000) Actor-Critic Algorithms. Advances in Neural Information Processing Systems. *NIPS Conference*, Denver, Colorado, 29 November-4 December 1999.

[3] Silver, D., Lever, G., Heess, N., *et al*. (2014) Deterministic Policy Gradient Algorithms. *International Conference on Machine Learning*, Beijing, 21-26 June 2014, 387-395.

[4] Watkins, C.J.C.H. and Dayan, P. (1992) Q-Learning. *Machine Learning*, **8**, 279-292.

[5] Fujimoto, S., Hoof, H. and Meger, D. (2018) Addressing Function Approximation Error in Actor-Critic Methods. *International Conference on Machine Learning*, Stockholm, 10-15 July 2018, 1587-1596.

[6] Heuillet, A., Couthouis, F. and Díaz-Rodríguez, N. (2021) Explainability in Deep Reinforcement Learning. *Knowledge-Based Systems*, **214**, Article ID: 106685. https://doi.org/10.1016/j.knosys.2020.106685

[7] Madumal, P., Miller, T., Sonenberg, L., *et al*. (2020) Explainable Reinforcement Learning through a Causal Lens. *Proceedings of the AAAI Conference on Artificial Intelligence*, New York, 7-12 February 2020, 2493-2500.

[8] Sequeira, P. and Gervasio, M. (2020) Interestingness Elements for Explainable Reinforcement Learning: Understanding Agents' Capabilities and Limitations. *Artificial Intelligence*, **288**, Article ID: 103367. https://doi.org/10.1016/j.artint.2020.103367

[9] Madumal, P., Miller, T., Sonenberg, L., *et al*. (2020) Distal Explanations for Explainable Reinforcement Learning Agents. arXiv:2001.10284.

[10] Liventsev, V., Härmä, A. and Petković, M. (2021) Neurogenetic Programming Framework for Explainable Reinforcement Learning. *Proceedings of the Genetic and Evolutionary Computation Conference Companion*, Lille, 10-14 July 2021, 329-330.

[11] Cruz, F., Dazeley, R., Vamplew, P., *et al*. (2021) Explainable Robotic Systems: Understanding Goal-Driven Actions in a Reinforcement Learning Scenario. arXiv:2006.13615.

[12] Vaswani, A., Shazeer, N., Parmar, N., *et al*. (2017) Attention Is All You Need. 31*st Conference on Neural Information Processing Systems* (*NIPS* 2017), Long Beach, 4-9 December 2017, 11 p.

[13] Parisotto, E., Song, F., Rae, J., *et al*. (2020) Stabilizing Transformers for Reinforcement Learning. *International Conference on Machine Learning*, Virtual, 12-18 July 2020, 7487-7498.

[14] Janner, M., Li, Q. and Levine, S. (2021) Offline Reinforcement Learning as One Big Sequence Modeling Problem. arXiv:2106.02039.

[15] Chen, L., Lu, K., Rajeswaran, A., *et al*. (2021) Decision Transformer: Reinforcement Learning via Sequence Modeling. arXiv:2106.01345.

[16] Yarats, D., Fergus, R., Lazaric, A., *et al*. (2021) Reinforcement Learning with Prototypical Representations. *International Conference on Machine Learning*, Virtual, 18-24 July 2021, 11920-11931.

[17] Schwarzer, M., Rajkumar, N., Noukhovitch, M., *et al*. (2021) Pretraining Representations for Data-Efficient Reinforcement Learning. 35*th Conference on Neural Information Processing Systems* (*NeurIPS* 2021), Virtual, 6-14 December 2021, 14 p.

[18] Hansen, N. and Wang, X. (2021) Generalization in Reinforcement Learning by Soft Data Augmentation. 2021 *IEEE International Conference on Robotics and Automation* (*ICRA*), Xi'an, 30 May-5 June 2021, 13611-13617. https://doi.org/10.1109/ICRA48506.2021.9561103

[19] Brockman, G., Cheung, V., Pettersson, L., *et al*. (2016) OpenAI Gym. arXiv:1606.01540.

[20] Schrittwieser, J., Antonoglou, I., Hubert, T., *et al*. (2020) Mastering Atari, Go, Chess and Shogi by Planning with a Learned Model. *Nature*, **588**, 604-609.

[21] Gu, B. and Sung, Y. (2021) Enhanced Reinforcement Learning Method Combining One-Hot Encoding-Based Vectors for CNN-Based Alternative High-Level Decisions. *Applied Sciences*, **11**, Article No. 1291. https://doi.org/10.3390/app11031291

[22] Johannink, T., Bahl, S., Nair, A., *et al*. (2019) Residual Reinforcement Learning for Robot Control. 2019 *International Conference on Robotics and Automation* (*ICRA*), Montreal, 20-24 May 2019, 6023-6029. https://doi.org/10.1109/ICRA.2019.8794127

[23] Zhang, R., Lv, Q., Li, J., *et al*. (2022) A Reinforcement Learning Method for Human-Robot Collaboration in Assembly Tasks. *Robotics and Computer-Integrated Manufacturing*, **73**, Article ID: 102227. https://doi.org/10.1016/j.rcim.2021.102227

[24] Kiran, B.R., Sobh, I., Talpaert, V., *et al*. (2021) Deep Reinforcement Learning for Autonomous Driving: A Survey. *IEEE Transactions on Intelligent Transportation Systems*, 1-18. https://doi.org/10.1109/TITS.2021.3054625

[25] Ma, X., Li, J., Kochenderfer, M.J., *et al*. (2021) Reinforcement Learning for Autonomous Driving with Latent State Inference and Spatial-Temporal Relationships. 2021 *IEEE International Conference on Robotics and Automation* (*ICRA*), Xi'an, 30 May-5 June 2021, 6064-6071. https://doi.org/10.1109/ICRA48506.2021.9562006

[26] Chen, J., Li, S.E. and Tomizuka, M. (2021) Interpretable End-to-End Urban Autonomous Driving with Latent Deep Reinforcement Learning. *IEEE Transactions on Intelligent Transportation Systems*, 1-11. https://doi.org/10.1109/TITS.2020.3046646