基于深度强化学习的机械臂自主抓取算法

田成军，颜 禹，崔仁伟，张晋通

（长春理工大学 电子信息工程学院，长春 130012）

摘 要：针对传统机械臂不具备自主学习能力，在未知环境中适应能力差等问题，提出了一种结合先验知识的深度强化学习算法，同时修改了奖励函数和经验池的设计，解决了训练前期数据质量低、收敛速度慢、学习效果不理想等问题，同时增强了其泛化性。仿真结果表明：改进后的模型相比于原始算法收敛速度提高了48.5%，成功率提升了8%，对比其他主流算法在收敛速度和成功率上都有明显提升。

关键词：模式识别与智能系统；深度强化学习；先验知识；机械臂

中图分类号：TP183 **文献标志码**：A **文章编号**：1671-5497(2026)03-0662-08

DOI：10.13229/j.cnki.jdxbgxb.20240739

Autonomous grasping algorithm of robotic arm based on deep reinforcement learning

TIAN Cheng-jun，YAN Yu，CUI Ren-wei，ZHANG Jin-tong

（*School of Electronic Information Engineering*，*Changchun University of Science and Technology*，*Changchun* 130012，*China*）

Abstract：Aiming at the problems such as the lack of independent learning ability of traditional robotic arms and poor adaptability in unknown environments，this paper proposes a deep reinforcement learning algorithm combining prior knowledge，and modifies the design of reward function and experience pool to solve the problems such as low data quality，slow convergence speed and unsatisfactory learning effect in the early stage of training，while enhancing its generalization. The results on Pybullet simulation platform show that compared with the original algorithm，the convergence speed of the improved model is increased by 48.5%，and the success rate is increased by 8%. Compared with other mainstream algorithms，the convergence speed and success rate are significantly improved.

Key words：pattern recognition and intelligent system；deep reinforcement learning；priori knowledge；robot arm

0 引 言

机械臂在各行各业都扮演着重要的角色，不仅可以提高生产效率，降低成本、增强安全性，还能减少人为失误，完成一些人类难以完成的工作和任务。现阶段的机械臂更多的还是按照预先设定好的程序完成单一重复且含金量较低的工作。在面对复杂的非结构化环境时，如何实现更精确

收稿日期：2024-07-04.

基金项目：吉林省科技发展计划项目(20220203066SF).

作者简介：田成军(1977-)，男，教授，博士. 研究方向：人工智能与模式识别. E-mail：tianchengjun@cust.edu.cn

的操作和更高效的决策仍然是个很大的挑战。

深度强化学习[1](Deep reinforcement learning，DRL)是一种结合了深度学习感知能力和强化学习决策能力方法，文献[2]提出了改进的深度双Q网络算法(DDQN)，通过使用两个独立的神经网络来分别估计Q值函数，以解决原始DQN算法中的过度估计问题。为了解决样本采集过程中等概率选取样本而忽略优先级的问题，文献[3]提出了深度双Q网络以及优先级经验回放(Double deep Q Network with prioritized experience replay)，该方法对样本的选择进行重要性划分，并减少了训练时间；文献[4]提出事后经验回放，用于在稀疏奖励环境中提高样本学习效率；文献[5]提出了双延迟深度确定性策略梯度(Twin delayed deep deterministic policy gradient，TD3)算法以及基于软策略和软Q值的算法；文献[6]使用基于TD3算法的元强化学习框架，引入概率性上下文变量机制来处理机械臂控制问题；文献[7]提出了一种面向信号交叉口的自适应学习生态驾驶策略，基于深度确定性策略梯度算法(DDPG)对车辆加速度进行实时控制与训练；文献[8]提出了一种基于异步合作更新的LSTM-MADDPG多智能体协同决策算法。基于差异奖励和值分解思想，利用长短时记忆(LSTM)网络提取轨迹序列间特征，优化全局奖励划分方法，实现各智能体的动作奖励分配。文献[9]采用目标位置引导和TD3算法的联合方式进行轨迹优化，解决了高维动作空间中学习效率低下的问题。

尽管利用深度强化学习方法已成功实现了机械臂在未知环境中的自主规划与学习，但训练过程中仍然存在数据采集效率低下、样本利用率低、奖励函数设计不合理、高维连续的状态-动作难以学习等问题，进而导致训练过程中收敛速度慢、成功率低，影响机械臂的训练效果。

针对上述问题，本文以TD3为原始模型引入先验知识，通过修改奖励函数和经验池等方法进行机械臂的抓取训练，以奖励函数和成功率为指标对改进算法和其他主流算法进行比较分析。

1　问题描述

在实际生产生活中，传统的机械臂控制模型在固定环境下执行特定任务时有着较好的表现，但随着应用场景的复杂程度增加，传统模型在物体环境、光照变化、机械臂位置更改等情况下，对环境变化的适应能力较差，缺乏基于环境反馈的在线学习能力，对于多步骤、高度不确定性和需要即时调整的任务，传统方法的决策能力效率较低。

与传统的控制方法相比，强化学习具有独特的优势，使机械臂能够在复杂、高维、不确定和动态环境等应用场景中发挥越来越重要的作用。

1.1　相关知识

强化学习[10]是一种机器学习的方法，其原理为智能体(Agent)通过与环境交互学习到行动策略，最大化积累其奖励，智能体根据环境对动作的奖励进行学习，并更新其模型参数。强化学习的基本模型可以视为一个标准的马尔科夫决策过程(Markov decision process，MPD)，由(S, A, P, R, γ)五元组构成，其中S为状态空间，A为动作空间，P为状态转移概率矩阵，R为奖励函数，γ为折扣因子。智能体在给定状态$s(s\in S)$下选取一个动作$a(a\in A)$的函数称为策略π。策略π为Agent在状态s下选取动作a的概率：$\pi(a|s))=P\left[A_t=a\middle|S_t=s\right]$，Agent通过最大化$G_t$最优策略$\pi^*$。

1.2　双重深度确定性策略梯度算法

双重深度确定性策略梯度算法[11](Twin delayed deep deterministic policy gradient，TD3)是在(Deep deterministic policy gradient，DDPG)算法基础上改进得到的一种用于解决连续控制问题的在线(on-line)异策(off-policy)式深度强化学习算法。以下是TD3在DDPG基础上的改进。

(1)双Critic网络

TD3在DDPG的基础上提出了双Critic网络[12]，分为在线Critic和目标Critic网络，在线Critic网络负责计算当前Q值并更新自身参数，目标Critic网络利用式(1)进行更新自身参数并负责计算目标Q值：

$$Q_{\theta'}=R+\gamma\max_{a'}Q(S',A') \tag{1}$$

在DDPG的每一次更新参数学习过程中所估计的Q值是当前最大动作的Q值，因此会造成Q值估值过高的问题。相比之下，TD3采用双Critic网络分别对Q值进行估计并选取较小的一个，可以有效解决Q值过高的问题，计算目标值时使用式(2)中的较小值来估计下一个状态动作价值：

$$y=R+\gamma\min_{i=1,2}Q_{\theta'_i}(S',A') \tag{2}$$

(2)延迟Actor更新

由于Actor网络是通过最大化累积期望回报来更新的,它需要利用Critic网络来进行评估。如果Critic网络非常不稳定,那么Actor网络也会出现震荡。因此,Critic网络的更新频率高于Actor网络,即等待Critic网络更加稳定之后再更新Actor网络。

(3)目标策略平滑正则化

在DDPG中更新Critic网络时,使用确定性策略的学习目标极易受到函数逼近误差的影响,导致目标估计的方差大,估值不准确。

TD3算法加入了正则化,即在由策略输出的动作中加入噪声,这样对于相似的动作,对应的Q值也相似。因此,减小了函数逼近时误差产生的影响,使Critic网络计算更加准确,动作选择更加平滑。

2 算法改进及实施

本节将详细阐述基于TD3算法改进与实施的关键步骤,旨在通过算法的优化,显著增强机械臂的自主学习能力、加速收敛进程,并提升任务执行的成功率。

2.1 算法改进

众多研究表明,机械臂结合强化学习面临的最大问题为机械臂所产生的高维连续动作空间会导致数据采样效率低下、经验样本质量低、奖励函数设计困难,进而导致训练数据不足,最终导致训练前期学习效率过低、浪费大量时间算力、训练效果差等问题,因此,本文对算法进行改进。

2.1.1 引入先验知识

在原始DRL算法中,训练初期完全使用随机的探索方式,这就使得在前期Agent与环境交互产生的数据质量较低,无法学习到有用的数据样本。

针对上述问题,本文在原始算法中引入先验知识[13]。通过将专家经验导入教学经验池的方法对训练前期提供策略指导,使Agent在初始阶段能够更快地采取有效的行为策略,减少不必要的探索行为,引入先验知识的DRL模型如图1所示。

在训练初期,Agent主要以先验动作进行探索,以此来保证训练效率,在训练有大体雏形后加大随机探索的权重来保证机械臂学到最优策略,同时避免陷入局部最优解。

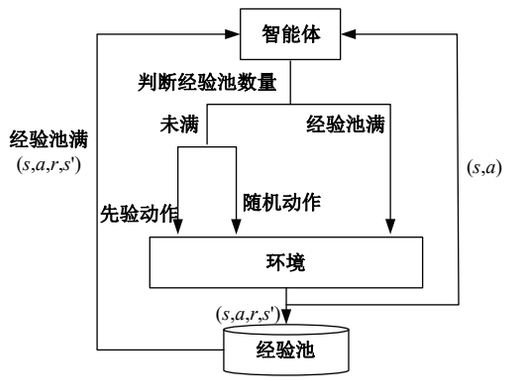

图1 引入先验知识的模型设计

Fig. 1 Model design with prior knowledge

2.1.2 Q滤波器

行为克隆(Behavior cloning)是一种利用示范数据的技术,通过模仿专家或示规范动作来训练Agent,目标是最小化示范动作和Agent动作之间的差异。然而,行为克隆也有可能过拟合造成Agent在示范数据上表现良好但在未见过的数据上表现较差,也可能示范数据在某次探索中本身数据不是最优的,直接模仿造成得到了次优策略。所以,本文引入Q值函数$Q(s,a)$来判断在给定状态下是否应当对示范数据应用行为克隆,帮助其更智能地利用示范数据,避免上述问题。

在每个示范状态s下,计算示范动作a_{d}和当前策略动作a_π的Q值$Q(s,a_{\mathrm{d}})$和$Q(s,a_\pi)$,a_π是根据当前策略π选择的动作。比较示范动作的Q值$Q(s,a_{\mathrm{d}})$与当前策略动作的Q值$Q(s,a_\pi)$:

$$\Delta Q = Q(s,a_{\mathrm{d}}) - Q(s,a_\pi) \tag{3}$$

如果$\Delta Q>0$,说明示范动作a_{d}在该状态下的预期回报高于当前策略动作a_π,即示范动作优于当前策略动作。在这种情况下,应用行为克隆损失,使代理在此状态下倾向于学习示范动作。如果$\Delta Q\leqslant 0$,说明示范动作a_{d}并不优于当前策略动作a_π,则不应用行为克隆损失。

定义行为克隆损失$\mathcal{L}_{\mathrm{BC}}$为:

$$\mathcal{L}_{\mathrm{BC}} = \mathbb{E}_{(s,a_{\mathrm{d}})\sim D}\left[\Delta Q>0\cdot /\!/\, a_{\mathrm{d}}-a_\pi /\!/^{2}\right] \tag{4}$$

式中:D为示范数据集。

损失函数仅在$\Delta Q>0$时计算,即仅在示范动作优于当前策略动作时才进行行为克隆。

2.1.3 结合先验知识的多经验池设计

原算法中单一经验池的设计会造成数据利用率低、收敛速度较慢等问题。本文在原算法基础上引入教学经验池,又在此基础上,根据Td-error大小分为随机采样经验池和优先经验回放池[14],

如图2所示。这样的设计可以在训练时提取到各时期各种情况的数据，在更新参数后可以拥有更好的鲁棒性和泛化性，减少大量前期随机探索时间，提高快速收敛性。

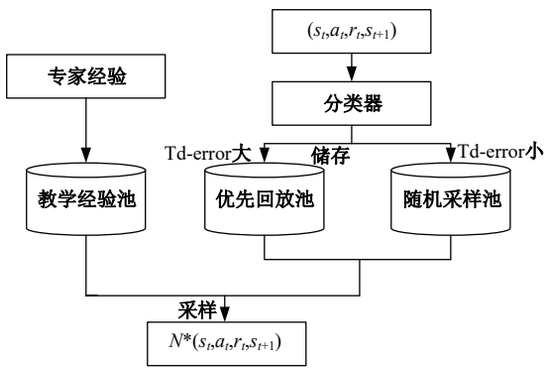

图2　多经验池流程图

Fig. 2　Flow chart of multi-experience pool

将专家数据导入教学经验池，机械臂通过教学经验池开始训练，经过与环境交互产生新的(s_t, a_t, r_t, s_{t+1})，通过 Td-error 大小来分别存入优先回访池和随机采样池中，Td-error公式为：

$$\delta = r + \gamma Q(s', a'; \theta^-) - Q(s, a; \theta) \tag{5}$$

式中：r为奖励；γ为折扣因子；$Q(s, a; \theta)$为当前状态动作的Q值；$Q(s', a'; \theta^-)$为下一状态的目标Q值。

Td-error较大，表示这些经验在当前策略下有较大的预测误差，可能对策略改进贡献较大，经验池中的样本按优先级进行采样，具有高优先级的样本被采样的概率更大。采样后的样本进行加权更新，权重为样本被采样的概率的倒数，这样可以避免高优先级样本被过度学习，保持样本利用的均衡性。训练初期采用教学经验池数据进行训练，有效减少了前期训练的随机探索时间，到训练中后期采用经验池优先回放池和随机采样池的数据进行训练，使机械臂在合理范围内进行更多随机探索，避免陷入局部最优解。经验池设计中存储了多种不同状态下的经验数据，包括成功和失败的情况。这种多样性有助于智能体更全面地学习和理解环境，从而使其学习过程更加平滑，增加其泛化性。

2.1.4　奖励函数

在强化学习中，奖励函数是重要的组成部分之一，它定义了智能体在特定状态下采取动作后获得的反馈，为智能体提供了一个明确的目标，本文将$R1$设计成：

$$R1 = r_{\text{reach}} = \begin{cases} pos_{\text{ee}} - pos_{\text{goal}} \\ pos_{\text{ee}} - pos_{\text{goal}}[:2] \end{cases} \tag{6}$$

式中：pos_{ee}为机械臂末端执行器在三维空间中的位置；pos_{goal}为目标点在基坐标系下的位置；$\left\| pos_{\text{ee}} - pos_{\text{goal}} \right\|$为末端执行器到目标点的欧氏距离；$[:2]$表示取向量的前两个元素，只考虑$x$和$y$坐标上的误差，而忽略$z$坐标的误差。

这样的设计可以最小化末端执行器和目标位置之间的距离，也可以简化数据量的计算。

将$R2$设计成：

$$R2 = r_{\text{grasp}} = \begin{cases} 1, \text{if grasp success} \\ 0, \text{otherwise} \end{cases} \tag{7}$$

式(7)根据待抓取物体在三维空间中的坐标变换来判断物体是否被抓取成功，若抓取成功则给一个正向为1的奖励，若失败则没有。

将$R3$设计成：

$$R3 = r_{\text{lift}} = \begin{cases} 1.5 + 10 \times \text{pos}_{\text{obj}}[2], \text{if pos}[2] > 0.02 \\ 0, \text{otherwise} \end{cases} \tag{8}$$

式中：if $\text{pos}_{\text{obj}}[2]$表示物体在$z$轴上的位置是否被举起，0.02被(标定为一个物体是否被举起的判断阈值)，如果被举起高度大于0.02(2 cm)，则被视为成功举起，赋予$1.5 + 10 \times \text{pos}_{\text{obj}}[2]$的动态奖励值，即被举起越高奖励值越高，若失败，则没有任何奖励。

将$R4$设计成：

$$R4 = r_{\text{success}} = \begin{cases} 200, \text{if task success} \\ 0, \text{otherwise} \end{cases} \tag{9}$$

式(9)表示一个稀疏奖励函数，这样的设计有助于避免Agent因频繁获得小奖励而迷失方向，更好地关注长远目标而非短期利益。

将塑性与稀疏奖励函数相结合的方式可以更好地兼顾机械臂抓取的各种情况，过度更加平滑，避免陷入局部最优解。

综上所述，本文奖励函数由四部分构成，$R = R1 + R2 + R3 + R4$。

2.2　算法实施

本文通过Pybullet仿真平台进行深度强化学习的环境部署与训练，及性能对比。

使用中科深谷公司的六自由度机械臂在仿真环境中进行训练，其结构和数据如图3所示。

机械臂末端连接着一个三指柔性夹爪，如图

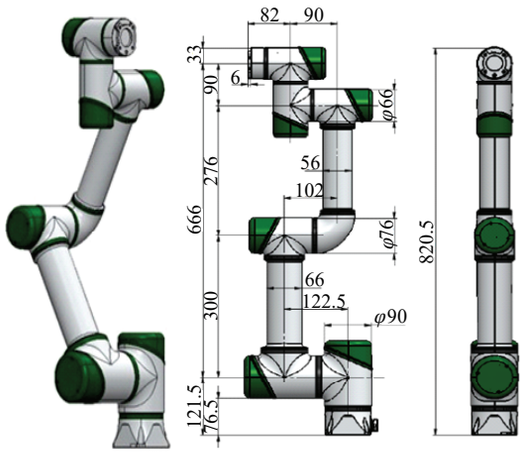

图3 六自由度机械臂

Fig. 3 Six-degree-of-freedom robotic arm

4所示,它能精准地抓取和操控各种形状的物体。

在仿真环境中,通过RGB-D摄像头生成的图像机械臂可以对物体进行靠近、抓取、举起和放置等操作训练,如图5所示。

图4 三指柔性夹爪

Fig. 4 Three-finger flexible gripper

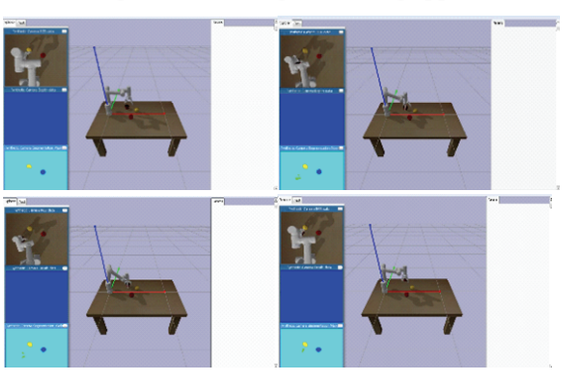

图5 Pybullet中的环境布置与训练

Fig. 5 Layout and training of the environment in Pybullet

3 结果分析

奖励函数的数值大小和成功率是评价深度强化学习的重要指标,机械臂自主抓取的本质是一种行为决策,本文从成功率和奖励函数两方面来进行对比分析。

3.1 改进模型与原始模型对比分析

对TD3进行经验池改进后对比原算法的奖励函数情况如图6所示,可见以奖励(Reward)和步长(Timesteps)为坐标轴,原始算法在600 000步后趋于稳定,改进后的TD3算法在400 000步后趋于平稳,在整个过程中,原始算法表示出了较大波动问题,改进后的算法在学习过程中更为稳定,受到的波动较小,在相同训练阶段内学习到了累积奖励更高的策略。

对TD3改进经验回放后对比原算法的成功率情况如图7所示,可见在训练初期大约0到200 000步时,原始TD3的成功率几乎为零,与此相比,改进后的算法成功率在早期就开始上升,明显优于原始TD3。在最后阶段,两种算法的成功率都趋近1.0,表示它们都成功学会了任务,但改进后的算法收敛速度更快,且达到高成功率的时间更早。

对TD3算法进行经验池和奖励函数以及Q滤

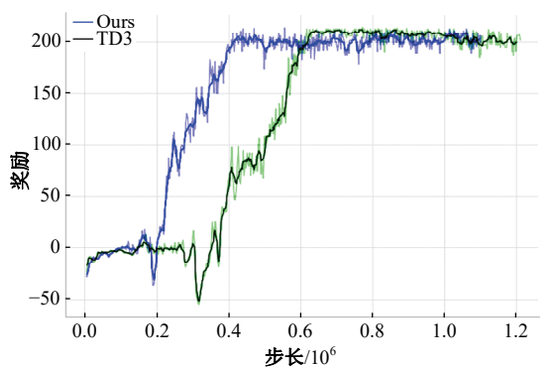

图6 原始算法与改进算法奖励值对比

Fig. 6 Comparison of reward values between original and improved algorithms

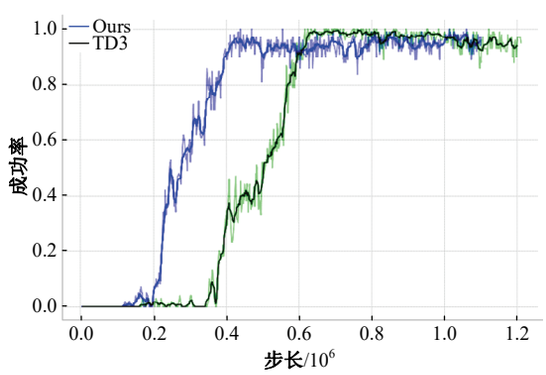

图7 原始算法与改进算法成功率对比

Fig. 7 Comparison of the success rate of the original algorithm and the improved algorithm

波改进后，对比原算法的奖励函数情况如图 8 所示，可见以奖励和步长为坐标轴，对比没有引入模仿学习修改经验池和 Q 滤波的算法，改进后的算法大约在 100 000 步的时候迅速提高并达到大约 220 的高奖励水平，波动很小。无 Q 滤波的绿线在大约 200 000 时间步长时达到稳定的高奖励水平。虽然表现比较接近本文改进的算法，但波动较大。没引入模仿学习的红线在训练初期出现了大量无效探索，浪费了时间，最终显示出了最慢的提升速度。

对 TD3 算法进行经验池和奖励函数以及 Q 滤波的改进后，对比原算法的成功率对比如图 9 所示。以“时间步长”为横坐标，成功率为纵坐标。改进后的方法在大约在 200 000 步时成功率迅速接近 1 且后期波动较小，没有引进 Q 滤波的绿线在大约 300 000 时间步长时成功率提升接近 1，但是在后期展示出了较大波动。没有引入专家经验的红线达到较高成功率后相对稳定，不过对比之下前期收敛速度较慢。

总体而言，改进后的算法表现出了更快的收敛速度和更高的稳定性，有明显的性能提升。

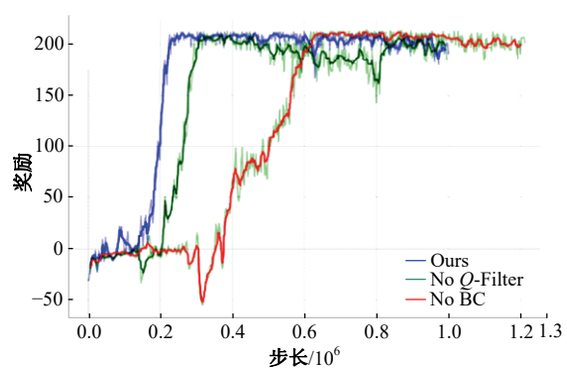

图 8　改进算法与原算法收敛性对比图

Fig. 8　Comparison of the convergence of the improved algorithm and the original algorithm

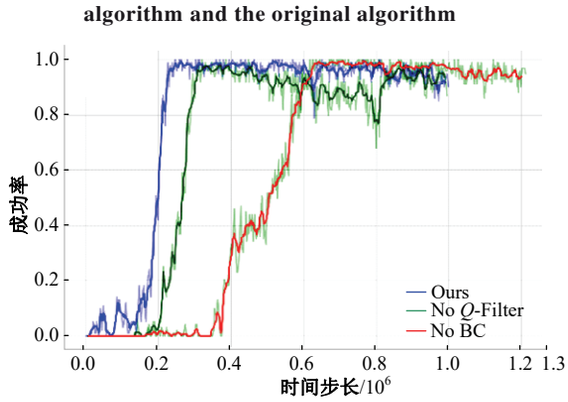

图 9　改进算法与原算法成功率对比图

Fig. 9　Comparison of the success rate of the improved algorithm and the original algorithm

3.2　改进模型与其他模型对比分析

对 TD3 算法引入模仿学习并进行经验池和奖励函数以及 Q 滤波的改进后，对比其他算法的奖励函数情况如图 10 所示，该图展示了横轴是训练步数，纵轴是奖励值。图 10 中共有 6 条曲线，分别代表 6 种不同的算法，改进后的算法在大约 200 000 步后迅速上升，在所有算法中用最短的时间达到最高的奖励值，且波动较小，说明该算法表现稳定且效果最好。TD3 算法在大约 400 000 步时上升，最终达到了较高的奖励值，但前期训练效果并不理想，前期无效探索占用了大量的时间和算力资源。DDPG 算法在大约 700 000 步之后才显著上升，最终达到较高的奖励值，但其稳定性一般，波动较大。SAC 算法在训练初期迅速上升，较早地达到了较高的奖励值，接近 200，但随后出现小幅度波动。PPO 算法在训练初期表现迅速提升，但随后波动较大，最大奖励值低于其他算法。而 A2C 算法在整个训练过程中与其他算法有较大差距。

对 TD3 算法进行经验池和奖励函数以及 Q 滤波的改进后，对比其他算法的成功率情况如图 11 所示，该图展示了不同强化学习算法在训练过程中的成功率随时间步数的变化情况。改进后的算法在大约 200 000 步时，成功率迅速上升，并接近 1.0。TD3 算法在大约 300 000 步时，成功率迅速上升至接近 1.0，并保持高成功率。DDPG 算法在大约 600 000 步时，成功率迅速接近至 1.0，不过表现出较大的波动。SAC 算法在早期迅速上升，约在 400 000 步时接近 1.0，之后出现波动。PPO 算法在早期上升较快，约 30 万步时成功率接

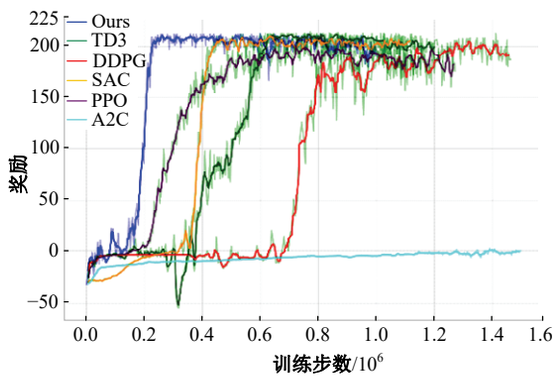

图 10　改进算法与其他算法收敛性对比图

Fig. 10　Comparison of the convergence of the improved algorithm with other algorithms

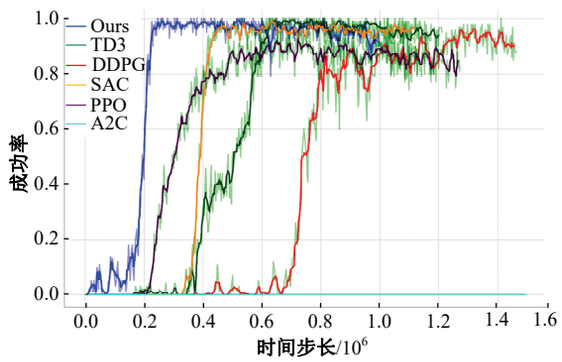

图11 改进算法与其他算法成功率对比图

Fig. 11 Comparison of the success rate of the improved algorithm with other algorithms

近0.9,但随后波动较大,未能达到1.0。

最后,利用改进后的模型随机对Pybullet_data中8样物品进行了测试,测试结果如表1所示。实验结果表明,改进后的算法在面对不同种类的物体抓取时均展现出了较高的可靠性和稳定性。

表1 多样物体抓取测试

Table 1 Multi-object grasping test

实验物体	每100次成功次数	识别IOU
螺丝刀	92	0.90
订书器	99	0.94
香蕉	80	0.88
锤子	88	0.88
牙膏	96	0.92
柠檬	82	0.80
网球	92	0.82

4 结束语

针对机械臂与深度强化学习结合过程中出现的收敛速度慢、学习时间长等问题,本文通过引入先验知识,改进奖励函数、加入Q滤波改进经验池等设计,使改进后的TD3模型与原始算法以及其他主流算法在收敛速度和成功率对比中都表现出明显的提高,加快了深度强化学习在训练前期由于随机探索造成的训练时间较长、收敛速度较慢等问题,节省了大量训练时间和算力。收敛速度和成功率的提升为算法的可靠性和鲁棒性提供了有力的支撑。改进后的经验池也可以有效收集和提取到各种情况的数据,表现为在执行任务时具有更好的泛化性和适应性,可以在复杂的非结构化环境中展现出更好的性能。

参考文献:

[1] 吕帅, 刘京. 基于深度强化学习的随机局部搜索启发式方法[J]. 吉林大学学报: 工学版, 2021, 51(4): 1420-1426.
Lyu Shuai, Liu Jing. Stochastic local search heuristic method based on deep reinforcement learning[J]. Journal of Jilin University (Engineering and Technology Edition), 2021, 51(4): 1420-1426.

[2] 李保罡, 王宇, 孔凡伟, 等. 基于智能反射表面辅助和信息年龄度量的安全状态更新[J]. 吉林大学学报: 工学版, 2023, 53(10): 3014-3025.
Li Bao-gang, Wang Yu, Kong Fan-wei, et al. Security status updates based on intelligent reflecting surface assistance and age of information metrics[J]. Journal of Jilin University (Engineering and Technology Edition), 2023, 53(10): 3014-3025.

[3] Schaul T, Quan J, Antonoglou I, et al. Prioritized experience replay[J/OL]. arXiv preprint arXiv: 1511.05952, 2015.

[4] Rauber P, Ummadisingu A, Mutz F, et al. Hindsight policy gradients[J/OL]. arXiv preprint arXiv: 1711.06006, 2017.

[5] Haarnoja T, Zhou A, Hartikainen K, et al. Soft actor-critic algorithms and applications[J/OL]. arXiv preprint arXiv:1812.05905, 2018.

[6] Rakelly K, Zhou A, Finn C, et al. Efficient off-policy meta-reinforcement learning via probabilistic context variables[C]//International Conference on Machine Learning, PMLR, 2019: 5331-5340.

[7] 庄伟超, 丁昊楠, 董昊轩, 等. 信号交叉口网联电动汽车自适应学习生态驾驶策略[J]. 吉林大学学报: 工学版, 2023, 53(1): 82-93.
Zhuang Wei-chao, Ding Hao-nan, Dong Hao-xuan, et al. Learning based eco-driving strategy of connected electric vehicle at signalized intersection[J]. Journal of Jilin University (Engineering and Technology Edition), 2023, 53(1): 82-93.

[8] 高敬鹏, 王国轩, 高路. 基于异步合作更新的LSTM-MADDPG多智能体协同决策算法[J]. 吉林大学学报: 工学版, 2024, 54(3): 797-806.
Gao Jing-peng, Wang Guo-xuan, Gao Lu. LSTM-MADDPG multi-agent cooperative decision algorithm based on asynchronous collaborative update[J]. Journal of Jilin University (Engineering and Technology Edition), 2024, 54(3): 797-806.

[9] 张强, 文闻, 周晓东, 等. 基于改进TD3算法的机械臂智能规划方法研究[J]. 智能科学与技术学报,

2022, 4(2): 223-232.
Zhang Qiang, Wen Wen, Zhou Xiao-dong, et al. Research on the manipulator intelligent trajectory planning method based on the improved TD3 algorithm [J]. Chinese Journal of Intelligent Science and Technology, 2022, 4(2): 223-232.

[10] 鲜斌, 张诗婧, 韩晓薇, 等. 基于强化学习的无人机吊挂负载系统轨迹规划[J]. 吉林大学学报:工学版, 2021, 51(6): 2259-2267.
Xian Bin, Zhang Shi-jing, Han Xiao-wei, et al. Trajectory planning for unmanned aerial vehicle slung-payload aerial transportation system based on reinforcement learning[J]. Journal of Jilin University (Engineering and Technology Edition), 2021, 51(6): 2259-2267.

[11] 康朝海, 孙超, 荣垂霆, 等. 基于动态延迟策略更新的TD3算法[J]. 吉林大学学报: 信息科学版, 2020, 38(4): 474-481.
Kang Chao-hai, Sun Chao, Rong Chui-ting, et al. TD3 Algorithm with dynamic delayed policy update [J]. Journal of Jilin University (Information Science Edition), 2020, 38(4): 474-481.

[12] 刘庆强, 刘鹏云. 基于优先级经验回放的SAC强化学习算法[J]. 吉林大学学报: 信息科学版, 2021, 39(2): 192-199.
Liu Qing-qiang, Liu Peng-yun. Soft actor critic reinforcement learning with prioritized experience replay [J]. Journal of Jilin University (Information Science Edition), 2021, 39(2): 192-199.

[13] 赵宏伟, 陈霄, 龙曼丽, 等. 基于改进PLSA分类器的目标分类算法[J]. 吉林大学学报: 工学版, 2012, 42(增刊1): 231-235.
Zhao Hong-wei, Chen Xiao, Long Man-li, et al. Object classification algorithm based on improved PLSA[J]. Journal of Jilin University (Engineering and Technology Edition), 2012, 42(Sup. 1): 231-235.

[14] 刘勇, 徐雷, 张楚晗. 面向文本游戏的深度强化学习模型[J]. 吉林大学学报: 工学版, 2022, 52(3): 666-674.
Liu Yong, Xu Lei, Zhang Chu-han. Deep reinforcement learning model for text games[J]. Journal of Jilin University (Engineering and Technology Edition), 2022, 52(3): 666-674.