Journal of Computer Applications
计算机应用,2021,41(6):1799－1804
ISSN 1001-9081
CODEN JYIIDU
2021-06-10
http://www.joca.cn

文章编号:1001-9081(2021)06-1799-06
DOI:10.11772/j.issn.1001-9081.2020091410

基于深度强化学习的二连杆机械臂运动控制方法

王建平,王　刚*,毛晓彬,马恩琪
(西安理工大学 机械与精密仪器工程学院,西安 710048)
(*通信作者电子邮箱1123016209@qq.com)

摘　要:针对二连杆机械臂的运动控制问题,提出了一种基于深度强化学习的控制方法。首先,搭建机械臂仿真环境,包括二连杆机械臂、目标物与障碍物;然后,根据环境模型的目标设置、状态变量和奖罚机制来建立三种深度强化学习模型进行训练,最后实现二连杆机械臂的运动控制。对比分析所提出的三种模型后,选择深度确定性策略梯度(DDPG)算法进行进一步研究来改进其适用性,从而缩短机械臂模型的调试时间,顺利避开障碍物到达目标。实验结果表明,所提深度强化学习方法能够有效控制二连杆机械臂的运动,改进后的DDPG算法控制模型的收敛速度提升了两倍并且收敛后的稳定性增强。相较于传统控制方法,所提深度强化学习控制方法效率更高,适用性更强。

关键词:深度强化学习;二连杆机械臂;运动控制;奖罚机制;深度确定性策略梯度算法
中图分类号:TP241.2; TP391.9　　**文献标志码**:A

Motion control method of two-link manipulator based on deep reinforcement learning

WANG Jianping, WANG Gang*, MAO Xiaobin, MA Enqi
(*School of Mechanical and Precision Instrument Engineering, Xi'an University of Technology, Xi'an Shaanxi* 710048, *China*)

Abstract: Aiming at the motion control problem of two-link manipulator, a new control method based on deep reinforcement learning was proposed. Firstly, the simulation environment of manipulator was built, which includes the two-link manipulator, target and obstacle. Then, according to the target setting, state variables as well as reward and punishment mechanism of the environment model, three kinds of deep reinforcement learning models were established for training. Finally, the motion control of the two-link manipulator was realized. After comparing and analyzing the three proposed models, Deep Deterministic Policy Gradient (DDPG) algorithm was selected for further research to improve its applicability, so as to shorten the debugging time of the manipulator model, and avoided the obstacle to reach the target smoothly. Experimental results show that, the proposed deep reinforcement learning method can effectively control the motion of two-link manipulator, the improved DDPG algorithm control model has the convergence speed increased by two times and the stability after convergence enhances. Compared with the traditional control method, the proposed deep reinforcement learning control method has higher efficiency and stronger applicability.

Key words: deep reinforcement learning; two-link manipulator; motion control; reward and punishment mechanism; Deep Deterministic Policy Gradient (DDPG) algorithm

0　引言

在工业制造过程中,为了实现自动化提高生产效率,需要用到许多不同类型与功能的机械臂,这其中就包括二连杆机械臂。二连杆机械臂可以从事搬运、吊装等简单工作,在工业生产中很常见。

在机械臂的研究中,其控制问题是非常重要的一部分。想让机械臂投入到新的工作环境中,需要对其控制系统进行反复调试,使机械臂能够适应工作环境并且达到精度要求,进而满足其他工作要求。在过去的研究中,经典控制方法使用得很多,如自适应控制、模糊控制、鲁棒控制等。但是随着工业技术的不断进步,一些控制方法精度有限,很难满足生产要求。研究者们也在经典控制方法的基础上,不断研究新的方法来控制机械臂。Soltanpour等[1]提出了一种用于机器人位置跟踪的最优模糊滑模控制器,克服了机械臂位置跟踪存在的不确定性。Oliveira等[2]针对刚性机械臂关节空间的位置控制问题,提出了利用混沌基的灰狼优化器对鲁棒高阶滑模控制器的参数进行优化,通过改变选择的混沌映射,提高了原始灰狼优化算法的计算效率。Wang等[3]通过反推技术,利用严格反馈结构构造了整个系统的控制李雅普诺夫函数,使反馈非线性系统达到稳定控制。Lu等[4]提出了一种基于线性二次型调节器(Linear Quadratic Regulator, LQR)的机械臂位置控制的方法,在传统的LQR控制中加入模糊算法,对LQR控制变量R进行自适应调整,提高了控制系统的适应性。Yin等[5]提出了一种弯曲机器人的控制算法,该算法利用弯曲过程的特征来解决问题,此方法能够避免在联合空间中遇到障碍。Li等[6]利用光滑切换函数构建自适应更新律,得到互联非线

收稿日期:2020-09-11;**修回日期**:2020-12-15;**录用日期**:2020-12-16。
作者简介:王建平(1970—),男,山西代县人,副教授,博士,主要研究方向:非线性系统动力学、智能控制;　王刚(1996—),男,陕西宝鸡人,硕士研究生,主要研究方向:智能控制、深度强化学习;　毛晓彬(1998—),男,山西临汾人,硕士研究生,主要研究方向:智能控制;马恩琪(1998—),男,陕西渭南人,硕士研究生,主要研究方向:智能控制。

性系统的渐进稳定性，实现自适应分布式控制方法。以上传统的控制方法经过不断改进性能已有所提升，但是仍普遍存在着控制效率低、适用性低等缺点，往往需要针对不同的对象来单独设计控制模型，并且面对不同的工作环境，在大多数时候需要调整控制方式。

随着计算机技术与人工智能技术的发展，控制技术也向智能化发展，智能控制以传统控制为基础，采取人的思维方式，利用类似人脑的控制方式来实现对研究对象的控制。相较于传统控制，智能控制所描述模型意义更为广泛，其具有学习、适应和组织功能，能够满足复杂系统的控制，具有分层处理信息与决策机构，往往一种控制方法能适应于多个不同研究对象。许多研究人员也将智能控制方法应用于机械臂控制中以实现机械臂的智能化。Ngo等[7]为了实现高精度的位置跟踪，提出了一种鲁棒自适应神经模糊网络控制系统，这种无模型控制方案能保证稳定的位置跟踪性，控制精度较高。该方法可以应用到简单的神经网络，但智能化程度较低，难以适应一些更复杂环境。Kormushev等[8]提出了一种在人-机器人交互环境中学习和再现机器人力相互作用的方法，利用人工智能中的模仿学习方法，使得机械臂通过学习获得再现动作的能力。模仿学习通过示教方式使机器人学会一些动作，需要人力通过特定方式来教学，因此这种方式人力成本高。Zhang等[9]设计了一种基于机器学习的机械手视觉控制系统，以三自由度平面机械臂为研究对象，使用深度强化学习DQN（Deep Q Network）对模型进行离线训练，并设置多种扰动项来测试算法的鲁棒性。DQN算法进行的是离散动作强化学习，适用于此处的视觉控制系统，但不能用于连续动作。李铭浩等[10]提出了一种机械臂容错控制方法，针对机械臂的自身运动性能，使用深度确定性策略梯度（Deep Deterministic Policy Gradient，DDPG）算法对机械臂进行容错控制，但是只在此算法基础上改变变量进行实验，缺乏与其他算法进行性能对比。Mnih等[11]在Actor-Critic算法基础上提出了异步A3C（Asynchronous Advantage Actor-Critic）算法，该算法充分利用CPU多核属性，高效率使用计算资源，通过与其他算法的对比可知，A3C算法的训练速度最快。刘成亮等[12]研究了一类具有弹簧耦合关节的二连杆欠驱动机器人从某一初始位置到不稳定平衡点位置的运动及稳定控制的问题，其基于无源性理论设计的控制器不仅受本身设计参数的影响较大，而且对系统初值比较敏感。万仁卓等[13]针对连续运动问题使用典型的深度强化学习算法，选取二连杆这一经典的连续运动控制任务进行研究，并对比了不同算法性能，但是其二连杆任务只是一个找点的过程，没有深刻体现出深度强化学习的特点。二连杆机械臂任务是连续动作，有必要对此任务的控制继续进行深入研究，改变机械臂任务环境，实现在更复杂环境下对机械臂的控制。传统控制方法难以消除其控制中的影响，单独的深度神经网络无法对其进行控制，而传统的强化学习方法适用于离散动作，所以本文提出了一种基于深度强化学习的二连杆机械臂运动控制方法，用于提高二连杆机械臂在复杂场景下的有效控制程度，并改善深度强化学习在连续动作控制时的适用性。

在本文中，针对二连杆机械臂的运动控制任务，将研究分为算法设计与仿真验证两部分，先进行深度强化学习的研究，搭建可进行机械臂连续运动的算法框架。Q-learning算法是传统的强化学习算法，其利用Q表处理离散问题较多，在应用中不易收敛，而深度Q学习[14]也只能处理离散的动作空间。机械臂运动是连续的动作，且最终结果是需要收敛的，所以本文使用的是DDPG算法[15]，DDPG将深度Q学习引入到连续动作空间中，可以解决连续动作问题。设置合理的机械臂运动的动作输入与合理的奖罚值，搭建出适合此二连杆机械臂模型的DDPG算法框架。在此基础上引入强化学习算法A3C与DPPO（Distributed Proximal Policy Optimization）算法[16]与DDPG算法进行对比，最后通过所设计算法对二连杆机械臂模型进行训练仿真，从而验证本文所提出的深度强化学习控制方法的性能。相较于传统的控制方法，深度强化学习控制方法可以充分利用智能技术自行训练探索出最优控制路径，避免许多控制系统反复调试的过程；深度强化学习控制方法可在短时间内确定控制策略，比传统方法效率高，并且作为无模型的控制方法，可以实现多种不同类别的运动控制，适用性强。

1 深度强化学习

1.1 深度强化学习理论

深度学习（Deep Learning）是机器学习的一个重要领域，其主要通过建立不同深度的神经网络来模拟人脑分析解决问题。通过深度学习建立起的神经网络框架可以应用于深度强化学习，但深度学习缺乏一定的决策能力。

强化学习（Reinforcement Learning，RL）是机器学习的一大分支，用于描述和解决智能体（agent）在与环境的交互过程中通过学习策略以达成回报最大化或实现特定目标的问题。agent的主要学习内容为行为策略（action policy）和规划（planning），其寻求最优行为策略以获得最大奖励值，以此方式来实现任务目标。强化学习具有决策能力，但是缺乏感知能力。

深度强化学习（Deep Reinforcement Learning）将深度学习的感知能力与强化学习的决策能力相结合，利用深度学习的神经网络框架与强化学习的决策能力来解决许多科研问题。

1.2 DDPG

DDPG算法[15]将深度学习与强化学习相结合，集结了DQN算法与Actor-Critic算法的优点，是一种离线策略、无模型的深度确定性策略梯度算法。DDPG算法与AC算法框架一样，但DDPG算法的神经网络划分更细，DQN算法在离散问题上性能较好，DDPG算法借鉴DQN的经验，解决了连续控制的问题，实现了端对端的学习。

DDPG的算法流程如图1所示，其中：actor网络接受输入状态，进行动作选择，输出动作变量；critic网络评估所选动作的好坏程度，计算出奖励值。DDPG算法的详细步骤如下。

DDPG算法在进行时，需要初始化神经网络的参数，actor选择一个传送给环境：

$$a_t = \mu(s_t|\theta^\mu) + N_t \tag{1}$$

环境执行a_t后，返回reward r_t和新状态s_{t+1}。

actor将状态转换的(s_t, a_t, r_t, s_{t+1})存入replay memory中，作为online网络训练的数据集。

DDPG分别为策略网络与Q网络创建了两个神经网络拷贝，分别是online网络和target网络，策略网络更新方式如下：

$$\begin{cases}\text{online: } Q(s, a|\theta^\mu), & \text{gradient update } \theta^\mu \\ \text{target: } Q(s, a|\theta^\mu), & \text{soft update } \theta^\mu\end{cases} \tag{2}$$

Q网络更新方式如下：

$$\begin{cases}\text{online: } Q(s,a|\theta^Q), & \text{gradient update } \theta^Q\\ \text{target: } Q'(s,a|\theta^Q), & \text{soft update } \theta^Q\end{cases} \tag{3}$$

从 replay memory 中,随机采样 N 个数据作为 online 策略网络和 online Q 网络的 mini-batch 训练数据。

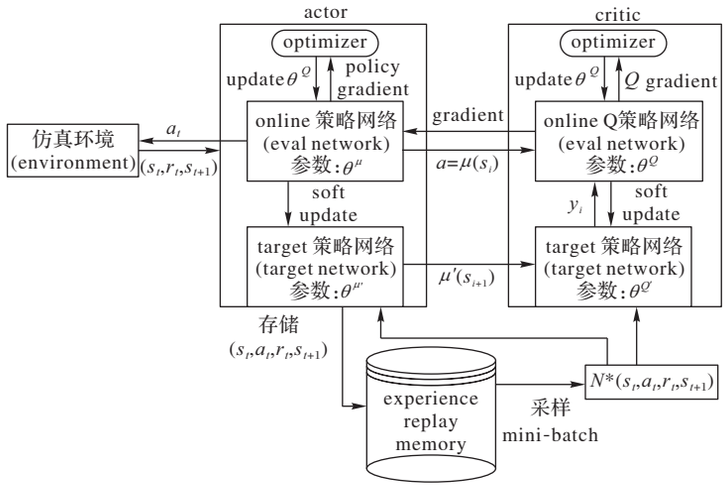

图 1　DDPG 算法流程

Fig. 1　Flow chart of DDPG algorithm

在 critic 中,计算 online Q 网络的 Q gradient 时,loss 的定义为:

$$L=\frac{1}{N}\sum_i\left(y_i-Q\left(s_i,a_i|\theta^Q\right)\right)^2;$$

$$y_i=r_i+\gamma Q'\left(s_{i+1},\mu'\left(s_{i+1}|\theta^{\mu'}\right)|\theta^{Q'}\right) \tag{4}$$

根据式(4),可求得 L 针对 θ^Q 的 gradient:$\nabla_{\theta^Q}L$,其中的计算使用的是 target 策略网络的 μ' 和 target Q 网络的 Q'。

在 actor 中,策略网络的优化使用 policy gradient 的方法:

$$\nabla_{\theta^\mu}J_\beta(\mu)\approx\frac{1}{N}\cdot\left(\nabla_a Q(s,a|\theta^Q)|_{s=s_i,a=\varpi(s_i)}\cdot\nabla_{\theta^\mu}\mu(s|\theta^\mu)|_{s=s_i}\right) \tag{5}$$

target 网络的参数采用 soft update 的方式:

$$\begin{cases}\theta^{Q'}\leftarrow\tau\theta^Q+(1-\tau)\theta^{Q'}\\ \theta^{\mu'}\leftarrow\tau\theta^\mu+(1-\tau)\theta^{Q'}\end{cases} \tag{6}$$

总体而言,DDPG 算法利用 Actor-Critic 框架,通过环境、actor 和 critic 三者之间交互的方式进行策略网络和 Q 网络的训练迭代。

1.3　A3C

强化学习 A3C[11] 是将 Actor-Critic 放到多个线程中同步训练,可以有效地利用计算机资源,提升训练效用,解决 Actor-Critic 不收敛的问题;并且 A3C 可以解决连续性动作空间的控制,适用于机械臂控制任务。

A3C 创建多个并行的环境,让多个 agent 同时在这些并行环境上更新主结构中的参数,并行中的 agent 们互不干扰,而主结构的参数更新受到不连续性干扰,所以更新的相关性被降低,收敛性提高;服务器的每个核都是一个线程,将程序在多核中同时运行,成倍提升了运行速度。

Actor-Critic 使用两个不同网络 actor 和 critic,A3C 将两个网络放到一起,即输入状态 S,输出状态价值 V 和对应策略 π。

A3C 算法使用了优势函数,可加速收敛,优势函数表达式如下:

$$A(S,t)=R_t+\gamma R_{t+1}+\cdots+\gamma^{n-1}R_{t+n-1}+\gamma^{n-1}V(S')-V(S) \tag{7}$$

策略参数的梯度更新如下:

$$\theta=\theta+\alpha\nabla_\theta\log\pi_\theta(s_t,a_t)A(S,t)+c\nabla_\theta H(\pi(S_t,\theta)) \tag{8}$$

actor 网络梯度更新为:

$$\mathrm{d}\theta\leftarrow\mathrm{d}\theta+\nabla_{\theta'}\log\pi(a_i|s_i;\ \theta')(R-V(s_i,\theta'_v)) \tag{9}$$

critic 网络梯度更新为:

$$\mathrm{d}\theta_v\leftarrow\mathrm{d}\theta_v+\partial(R-V(S_i;\ \theta_v'))^2/\partial\theta_v' \tag{10}$$

最终,通过更新迭代得到最优结果。

1.4　DPPO

DPPO 算法[16] 是基于 PPO(Proximal Policy Optimization)算法的进一步改进,其思路与 A3C 相似,也是通过多线程进行学习。

PPO[16] 是基于 Actor-Critic 的结构进行的改进,其具有三个网络,即 critic network、old_actor network 和 new_actor network。agent 首先利用 new_actor network 与环境互动获得 batch 数据,然后 actor network 和 critic network 对数据进行学习。

采集 batch 数据时,先将获得这个 batch 数据的 new_actor network 中的参数复制给 old_actor network,然后进行 new_actor network 和 critic network 的学习,new_actor network 更新的参数与 old_actor network 的参数进行对比,若差距过大,将难以收敛。

batch 数据存储的 T 个 state 输入给 critic network,critic network 分别输出 T 个时刻的估计值函数,然后再计算出 T 个目标值函数,计算出 T 个优势函数。

利用 TD error 对 new_actor network 的参数进行 N 次优化,其 loss 如下:

$$L^{\mathrm{CLIP}}(\theta)=\hat{E}\left[\min\left(r_t(\theta)\hat{A}_t,\mathrm{clip}(r_t(\theta),1-\varepsilon,1+\varepsilon)\hat{A}_t\right)\right] \tag{11}$$

最后利用 TD error 对 critic network 进行优化。

DPPO 算法将 PPO 算法转变为多线程模式,利用多线程在多个环境中收集数据,提高数据收集速度,并进行多线程运算,提高整体计算速度。

2 基于深度强化学习的二连杆机械臂的控制

2.1 系统结构

本文系统分为两部分：深度强化学习算法和实验仿真。通过深度强化学习对系统中的神经网络进行训练，使得算法可以控制二连杆机械臂的运动，最终避开障碍物到达目标位置。

仿真部分的环境包括了机械臂、障碍物和目标。首先，接收到算法的控制信号使机械臂进行运动；然后，将运动情况传递给控制算法，根据接收的信息，深度强化学习获得状态变量和奖励值。随着训练的不断进行，神经网络的参数也进行更新，获得的奖励值不断变化，如图2所示。

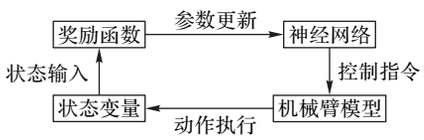

图2 本文系统结构

Fig. 2 Proposed system structure

2.2 二连杆机械臂模型

二连杆机械臂模型可采用D-H(Denavit，Hartenberg)法建立，模型参数如表1所示。

表1 机械臂的D-H参数

Tab. 1 D-H parameters of manipulator

杆件	θ	d	a	α
杆件1	θ_1	0	100	π/2
杆件2	θ_2	0	100	0

根据参数可以建立机械二连杆臂模型，此模型的两个连杆由一个转动副连接，底座部分为一个固定的转动副，具体模型如图3所示。

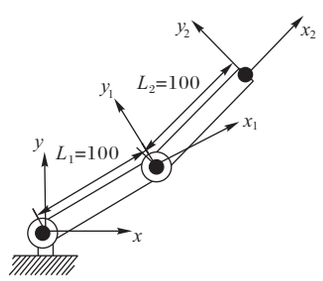

图3 二连杆机械臂模型

Fig. 3 Two-link manipulator model

2.3 深度强化学习控制方法

根据二连杆机械臂的特性搭建环境，构建深度强化学习算法控制模型，设置好理想状态变量(state)，输出指定动作(action)，分析动作的好坏程度后得出奖励值(reward)，再进行神经网络参数的更新，并继续进行训练。

1)奖励reward。

在强化学习对手臂进行控制时，需要手动设定较好的reward形式，reward非常重要，将涉及收敛问题。

对于手臂环境，涉及目标点位置、手臂端点位置和障碍物位置等。目标位置为(x,y)，手臂端点位置为(x_2,y_2)，可以设置r_1：

$$r_1 = -\sqrt{(x-x_2)^2-(y-y_2)^2} \tag{12}$$

根据不同障碍物位置，可以合理计算机械臂到达目标时避开障碍的极限位置，在此基础上设置r_2：

$$r_2 = \begin{cases} -20, & \text{触碰障碍物} \\ 0, & \text{其他} \end{cases} \tag{13}$$

根据手臂端点到达目标的情况，设置r_3：

$$r_3 = \begin{cases} +10, & \text{到达目标} \\ 0, & \text{其他} \end{cases} \tag{14}$$

根据r_1、r_2和r_3，确定最终reward值R：

$$R = r_1 + r_2 + r_3 \tag{15}$$

2)状态state。

手臂的特征很关键，如果可以将状态(state)最大化，也会使得收敛性大幅提升。这里的状态变量可以表示如下：端点是否在目标上(1个)，第一截手臂两端点到目标的坐标(4个)，第二截手臂两端点到目标的坐标(4个)，第二截手臂两端点到障碍物的坐标(4个)。13个信息可以使收敛性提升。当端点到达目标这一特征经过收敛被激活，收敛后的手臂将停留在目标上。

最终，将所设置好的reward和state加入到所建立的环境中，算法DDPG、A3C和DPPO通过所建立的环境控制机械臂的运动。

3 仿真实验与结果分析

3.1 仿真环境

本文的仿真环境在python下搭建，使用gym中的pygelt模块搭建出二连杆机械臂模型，利用python完成深度强化学习模型的搭建。

针对二连杆机械臂运动的连续控制问题，所搭建环境由二连杆机械臂、目标物和障碍物组成，所完成目标为：机械臂顶端通过深度强化学习算法控制，在不接触障碍物情况下抵达目标物。所搭建环境环境如图4所示。

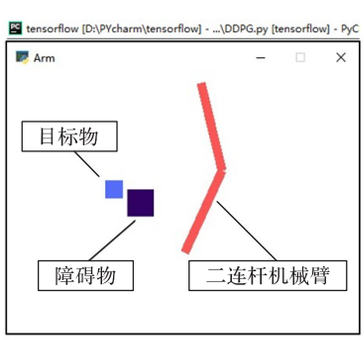

图4 二连杆机械臂运动控制仿真环境

Fig. 4 Motion control simulation environment of two-link manipulator

3.2 实验设置

本文的深度强化学习框架为TensorFlow，此框架广泛应用于深度学习领域，网络模型搭建了两层全连接层。

神经网络的输入为状态state，输出为动作。机械臂起始位置为随机，两段臂长都为100，臂宽为5，目标物边长为20，障碍物边长为30。动作的角度范围为[-180°，180°]，转动的速度设置为1，自由度为2。

训练参数设置如下：DDPG迭代最大轮数为10 000，每轮最大步数设置为200步，奖励折扣因子设置为0.9，actor部分的学习率为0.001，critic部分的学习率为0.001，BATCH_

SIZE 设置为 32,神经元个数设置为 100。A3C 和 DPPO 训练参数与 DDPG 设置相同。

当机械臂顶端避开障碍物到达目标物时,到达目标数加 1,当连续保持在目标物内 50 步,则回合结束,这样将视为成功一次。

3.3　结果分析

机械臂在训练过程中,会出现手臂接触障碍物的情况,但是最后的结果将避开障碍物达到目标物,如图 5 所示:机械臂经过深度强化学习模型训练,会避开障碍物,达到目标点。

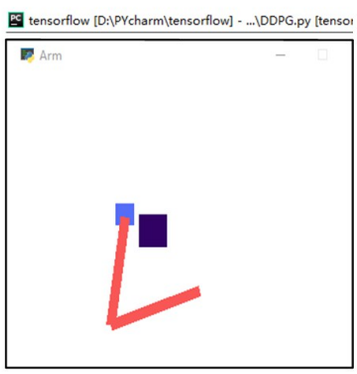

图 5　机械臂训练结果

Fig. 5　Training results of manipulator

经过训练,得到强化学习模型的奖励值变化如图 6 所示,反映了 DDPG 算法在当前设置下训练机械臂任务的收敛情况,可以看出,DDPG 算法进行机械臂训练后,找寻目标阶段的奖励值不稳定。在这过程中,机械臂会随意运动,也会触碰到障碍物,奖励值出现大幅度抖动,导致大约在 5 000 轮迭代后才产生收敛。收敛时机械臂顺利到达目标物,奖励回报值稳定在 500 左右,在持续收敛时也会出现奖励值小幅度变化,最终达到稳定收敛。由于 DDPG 算法在机械臂任务中收敛情况较差,继续建立深度强化学习 A3C 算法与 DPPO 算法对机械臂任务进行训练,根据三种算法所得到的结果如图 6 所示。

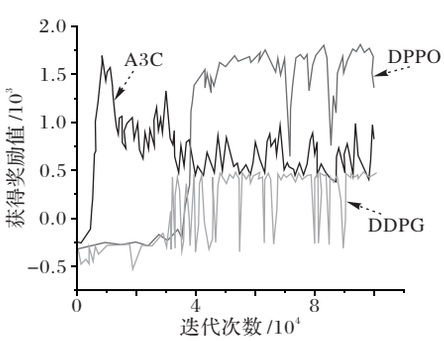

图 6　不同算法不同迭代次数的奖励值

Fig. 6　Reward values of different algorithms with different iteration times

图 6 中,A3C 算法训练所得到的奖励回报结果在任务开始后一直呈现上升趋势,在迭代 1 000 次之后,机械臂逐渐接近目标,其奖励值开始变小,在接近目标过程中,奖励值一直在 750 附近波动,波动频率大,而且机械臂到达目标的次数较少,得到的收敛情况不太理想。继续采用 DPPO 算法进行训练。

经过训练机械臂任务,DPPO 算法得到的奖励回报值变化为:训练开始阶段,机械臂进行行为探索,奖励回报值缓慢增长;经历 3 000 多次迭代后,机械臂探索到目标位置,奖励回报值发生突变,奖励值在 1 700 左右出现收敛情况,此时机械臂到达目标,在收敛一段时间后,奖励值出现大幅度变化,并出现持续为收敛情况。

比较三种算法训练后所得结果可以发现:A3C 算法由于多线程计算,训练速度快,探索目标速度快,但是到达目标位置收敛状况不佳;DPPO 算法在此机械臂运动任务中,前期的训练情况较好,能出现较好的收敛情况,但是后期奖励值出现大幅度波动,需要更多次迭代才能稳定;DDPG 算法在训练过程中,训练速度较慢,前期探索时间较长,后期会出现稳定收敛,但收敛期较短。

经过对比可知,DDPG 算法在训练时,尽管存在收敛不稳定现象,但是相较其他两种算法所得到的收敛情况更好,得到的收敛回报值也更稳定,所以 DDPG 算法更适用于实验模型,最终选择 DDPG 算法继续训练机械臂环境。由图 6 可以看出,实验结果收敛情况较差,产生收敛的时间过长,并且在收敛一段时间后会出现不稳定现象,偶尔出现未收敛现象。继续改进此模型,为改善连续收敛的状态,尝试调节一些训练参数,将算法的最大迭代次数从 10 000 提高到 20 000,神经元个数从 100 提升到 300,经过训练后,结果如图 7 所示。

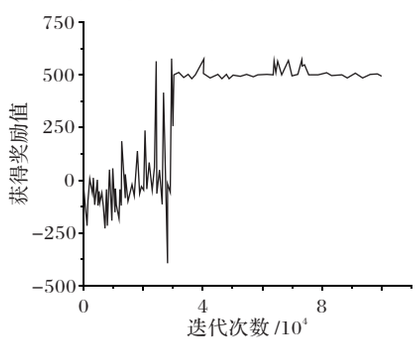

图 7　DDPG 算法最终随迭代次数变化的奖励值

Fig. 7　Final reward value of DDPG algorithm changing with iteration times

由图 7 可以看出,经过参数调节后,在迭代 3 000 轮左右就会产生收敛,搜寻目标的时间明显缩短,收敛效果明显增强,只会出现几次抖动情况,奖励值在收敛时的表现更加稳定,整体训练过程的稳定性有显著提升,使机械臂能更快更稳定地避开障碍物运动到目标。

在改进 DDPG 算法的参数结构后,机械臂任务控制稳定性增强,适用性也同时增强,当训练环境中的障碍物与目标位置发生变化时,机械臂依旧能够稳定躲避障碍到达目标。

4　结语

本文以二连杆机械臂为研究对象,其在运动过程中可能存在障碍物为背景,将二连杆机械臂搜寻目标时遇到障碍物问题转化为机械臂控制问题,提出了一种深度强化学习控制方法解决机械臂控制问题。本文建立了控制系统和机械臂模型,分别在 gym 和 TensorFlow 两大模块建立仿真模型和强化学习模型,并采用 DDPG 算法、A3C 算法与 DPPO 算法进行训练对比,最终选择 DDPG 算法进行进一步研究,改善 DDPG 算法对机械臂任务的适用性及稳定性。仿真实验结果表明,通

过不断的算法调试和模型训练，二连杆机械臂能顺利避开障碍物到达目标物，并在变化后的环境中依旧可以稳定避开障碍物到达目标。相较于传统的控制方法，本文的深度强化学习控制方法更加简单高效，而且适用于多种环境，适用性更广。但本文仅在二维平面内验证了深度强化学习对机械臂的控制，未来可进一步在三维空间内进行验证。

参考文献（References）

[1] SOLTANPOUR M R, KHOOBAN M H. A particle swarm optimization approach for fuzzy sliding mode control for tracking the robot manipulator [J]. Nonlinear Dynamics, 2013, 74(1/2): 467-478.

[2] OLIVEIRA J, OLIVEIRA P M, BOAVENTURA-CUNHA J, et al. Chaos-based grey wolf optimizer for higher order sliding mode position control of a robotic manipulator [J]. Nonlinear Dynamics, 2017, 90(2): 1353-1362.

[3] WANG Z, LIU X, LIU K, et al. Backstepping-based Lyapunov function construction using approximate dynamic programming and sum of square techniques [J]. IEEE Transactions on Cybernetics, 2017, 47(10): 3393-3403.

[4] LU E, YANG X, LI W, et al. Tip position control of single flexible manipulators based on LQR with the Mamdani model [J]. Journal of Vibroengineering, 2016, 18(6): 3695-3708.

[5] YIN X, WANG H, WU G. Path planning algorithm for bending robots [C]// Proceedings of the 2009 IEEE International Conference on Robotics and Biomimetics. Piscataway: IEEE, 2009: 392-395.

[6] LI X, YANG G. Adaptive decentralized control for a class of interconnected nonlinear systems via backstepping approach and graph theory [J]. Automatica, 2017, 76: 87-95.

[7] NGO T, WANG Y, MAI T L, et al. Robust adaptive neural-fuzzy network tracking control for robot manipulator [J]. International Journal of Computers Communications and Control, 2012, 7(2): 341-352.

[8] KORMUSHEV P, CALINON S, CALDWELL D G. Imitation learning of positional and force skills demonstrated via kinesthetic teaching and haptic input [J]. Advanced Robotics, 2011, 25(5): 581-603.

[9] ZHANG F, LEITNER J, MILFORD M, et al. Towards vision-based deep reinforcement learning for robotic motion control [EB/OL]. [2020-09-05]. https://arxiv.org/pdf/1511.03791.pdf.

[10] 李铭浩，张华，刘满禄，等. 基于深度强化学习的机械臂容错控制方法[J]. 传感器与微系统，2020，39(1)：53-55，59.（LI M H, ZHANG H, LIU M L, et al. Fault tolerant control method of manipulator based on deep reinforcement learning [J]. Transducer and Microsystem Technologies, 2020, 39(1): 53-55, 59.）

[11] MNIH V, BADIA A P, MIRZA M, et al. Asynchronous methods for deep reinforcement learning [C]// Proceedings of the 2016 33rd International Conference on Machine Learning. New York: JMLR.org, 2016: 1928-1937.

[12] 刘成亮，戈新生. 一类二连杆欠驱动机器人系统的稳定控制[J]. 北京信息科技大学学报(自然科学版)，2017，32(3)：25-29.（LIU C L, GE X S. Stability control to a kind of two-link underactuated robot system [J]. Journal of Beijing Information Science & Technology University, 2017, 32(3): 25-29.）

[13] 万仁卓，王思源，冯绎铭，等. 基于二连杆任务的深度强化学习算法分析与比较[J]. 湖北科技学院学报，2019，39(3)：151-156.（WAN R Z, WANG S Y, FENG Y M, et al. Analysis and comparison of deep reinforcement learning algorithms based on two-link task [J]. Journal of Hubei University of Science and Technology, 2019, 39(3): 151-156.）

[14] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Playing Atari with deep reinforcement learning [EB/OL]. [2020-09-05]. https://arxiv.org/pdf/1312.5602.pdf.

[15] LILLICRAP T P, HUNT J J, PRITZEL A, et al. Continuous control with deep reinforcement learning [EB/OL]. [2020-09-05]. https://arxiv.org/pdf/1509.02971v2.pdf.

[16] SCHULMAN J, WOLSKI F, DHARIWAL P, et al. Proximal policy optimization algorithms [EB/OL]. [2020-09-05]. https://arxiv.org/pdf/1707.06347.pdf.

WANG Jianping, born in 1970, Ph. D., associate professor. His research interests include nonlinear system dynamics, intelligent control.

WANG Gang, born in 1996, M. S. candidate. His research interests include intelligent control, deep reinforcement learning.

MAO Xiaobin, born in 1998, M. S. candidate. His research interests include intelligent control.

MA Enqi, born in 1998, M. S. candidate. His research interests include intelligent control.