深度强化学习在智能制造中的应用展望综述

孔松涛,刘池池,史 勇,谢 义,王 堃
重庆科技学院 机械与动力工程学院,重庆 401331

摘 要:深度强化学习作为机器学习发展的最新成果,已经在很多应用领域崭露头角。关于深度强化学习的算法研究和应用研究,产生了很多经典的算法和典型应用领域。深度强化学习应用在智能制造中,能在复杂环境中实现高水平控制。对深度强化学习的研究进行概述,对深度强化学习基本原理进行介绍,包括深度学习和强化学习。介绍深度强化学习算法应用的理论方法,在此基础对深度强化学习的算法进行了分类介绍,分别介绍了基于值函数和基于策略梯度的强化学习算法,列举了这两类算法的主要发展成果,以及其他相关研究成果。对深度强化学习在智能制造的典型应用进行分类分析。对深度强化学习存在的问题和未来发展方向进行了讨论。

关键词:人工智能;深度强化学习;深度学习;强化学习;智能控制;智能制造
文献标志码:A **中图分类号:**TP181 doi:10.3778/j.issn.1002-8331.2008-0431

Review of Application Prospect of Deep Reinforcement Learning in Intelligent Manufacturing

KONG Songtao, LIU Chichi, SHI Yong, XIE Yi, WANG Kun
School of Mechanical and Power Engineering, Chongqing University of Science and Technology, Chongqing 401331, China

Abstract: As the latest development of machine learning, deep reinforcement learning has been shown in many application fields. The algorithm research and application research of deep reinforcement learning have produced many classical algorithms and typical application fields. The application of deep reinforcement learning in industrial manufacturing can realize high level control in complex environment. First of all, the research on deep reinforcement learning is summarized, and the basic principles of deep reinforcement learning are introduced, including deep learning and reinforcement learning. Then, the paper introduces the theoretical methods of the application of deep reinforcement learning algorithm. On this basis, it classifies the algorithms of deep reinforcement learning, respectively introduces the reinforcement learning algorithm based on value function and the reinforcement learning algorithm based on strategy gradient, and lists the main development results of these two kinds of algorithms, as well as other related research results. Then, the typical applications of deep reinforcement learning in industrial manufacturing are classified and analyzed. Finally, the existing problems and future development direction of deep reinforcement learning are discussed.

Key words: artificial intelligence; deep reinforcement learning; deep learning; reinforcement learning; intelligent control; intelligent manufacturing

以人工智能为代表的第四次科技革命取得了众多成果,众多行业正进行着智能化的转变。机器学习领域的深度学习(Deep Learning,DL)[1],已经能实现图像识别[2]、音频识别[3]、自然语言处理[4]等功能,出色体现深度学习在信息感知方面的能力[5]。强化学习(Reinforcement Learning,RL)[6]是人工智能的另一发展成果,含义是让智能体在训练中根据得到的奖励和惩罚不断学习,最终根据学习经验做出高水平决策。目前在机器控制、机器人等领域应用广泛[7]。人工智能的发展目标是实现具有观察环境信息、独立思考决策的智能体(Agent)[8],智能体不仅需要智能提取信息,还需要做出智能决策,并且可以积累经验,保持学习的能力。深度强化学习(Deep Reinforcement Learning,DRL)[8]是实现这一目标的理论基础,DRL作为人工智能的最新成果之一,功能强大且发展迅速。人工智能的众多工作领域,如无人驾驶和智能流程控制,要实现智能体独立完成观察到动

基金项目:国家重点研发计划项目(2017YFC0805900);重庆科技学院研究生科技创新计划项目(YKJCX1920302)。
作者简介:孔松涛(1969—),男,博士,教授,硕士生导师,研究领域为流体流动与传热、工业大数据分析、钻井与石油装备;刘池池(1995—),男,硕士研究生,研究领域为流体流动与传热、智能控制;王堃(1980—),通信作者,男,博士,副教授,硕士生导师,研究领域为红外检测及设备安全,E-mail:wkcqust@foxmail.com。
收稿日期:2020-08-31 **修回日期:**2020-10-10 **文章编号:**1002-8331(2021)02-0049-11

作的完整工作流程,单一的DL或者RL都对此无能为力,两者结合才能完成任务。

DRL的控制水平在很多领域的表现不输人类甚至超越人类。阿尔法狗(AlphaGo)战胜职业棋手李世石,显示了智能体强大的学习能力。DRL可以无监督的情况下独立学习,可以学习人类专家的经验,最终达到专家水平,甚至在某些方面超越人类。与人脑相比,计算机在连续控制中稳定性更高。以无人驾驶为例,智能体可以杜绝人类驾驶员的主观错误,如疲劳、酒驾、分神等潜在事故因素。成熟的无人驾驶技术可降低事故率、保障交通安全,对于维护人民生命财产安全具有重要意义[9]。除了控制水平,在经验迁移方面,智能体也更有优势。智能体能通过直接的复制模型、数据分享等,完成批量的经验传递。对于不同的设备和控制流程,只要有一定的相似性,都可以进行经验迁移。迁移学习[10]为这种经验复制提供了理论支撑,并产生了新的研究方向。

除了无人驾驶方面的应用,DRL在计算机博弈、人机交互、机器人控制、文本生成等领域,都表现出较强的学习能力。

智能制造是由智能机器和人类专家组成的人机集成智能系统,它可以在制造过程中进行分析、推理、判断、概念和决策等智能活动[11]。在智能制造中,DRL可用于建立自学习、自适应、高效的智能机器。随着DRL算法的发展和应用,越来越多的生产过程通过智能机器实现,真正实现无人化和规模化生产。深度强化学习的算法研究和在智能制造中应用研究,对人类跨入智能制造时代具有重要意义。

1 深度强化学习的基本原理

1.1 深度学习

深度学习(Deep Learning,DL)是神经网络、人工智能、图形建模、优化、模式识别和信号处理等研究领域的交叉领域。深度学习的提出受到视觉机理启发,2006年,Hinton提出的一种称为深度置信网络的深度学习模型,揭开了深度学习发展的序幕。2012年Hilton团队提出的AlexNet模型在Imagenet竞赛中取得冠军[12],带来了深度学习的发展热潮。深度学习提高了计算机对高纬度信息的提取能力,在此基础上完成分类、识别等工作。

深度学习在信息提取方面的强大能力,主要是通过多层神经网络内部的非线性变换实现的[1]。在深度强化学习算法中,目前主要有基于卷积神经网络的深度强化学习和基于递归神经网络的深度强化学习,分别代表卷积神经网络和递归神经网络与强化学习的结合。

卷积神经网络(Convolutional Neural Network,CNN)在计算机视觉应用有突出表现,是近年来深度学习发展的热门。在图像处理时,网络通过层层计算,提取图像信息并对图像信息降维,实现对图像信息的计算机语言映射。常用的卷积神经网络有LeNet[13]、AlexNet[14]、VggNet[15]、ResNet[16]等。

递归神经网络(Recursive Neural Network,RNN)在自然语言处理中有突出应用,是一种拥有“记忆能力的神经网络”。递归神经网络虽然拥有这种“短期记忆”的优势,但也存在不足,比如梯度消失和梯度爆炸带来的影响[17];网络训练每一步都保留前面每一步的价值信息,而不是最近的和关系最大的。为了改善这些问题,Hochreiter等[18]提出了长短期记忆网络,通过增加线性干扰,让网络对信息选择性地增加或减少,比如降低对较远信息的权重,增加关系强的信息权重。对于递归神经网络只向前反馈,目前状态只依赖前面的输出,而忽视了后面的影响,为了解决这个问题,Schuster等[19]提出了双向递归网络,可以对两个方向进行学习。

1.2 强化学习

1.2.1 强化学习与马尔科夫决策过程

强化学习(Reinforcement Learning,RL)的决策过程是智能体(Agent)与环境交互做出的马尔科夫决策过程。其过程为智能体根据环境即时状态 S_t,为了获得环境反馈给智能体的最大奖励,做出智能体认为的最优动作 a,其奖励依据是,采取动作 a 之后的状态 S_{t+1} 的价值 $R_t(S_t,a_t,S_{t+1})$,再加上后续所有可能采取的动作和导致的状态的价值乘以一个折扣因子 γ,求得的累积奖励 G_t 为:

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

其中折扣因子 γ 用于削减远期决策对应的奖励权重,原因是离当前状态越远,不确定性就越高,决策最终目标是为了达到目标状态并实现累积奖励最大化。强化学习的基本构架如图1所示。

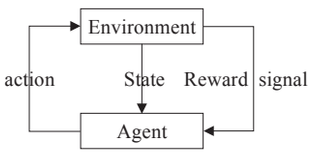

图1 强化学习基本框架

强化学习根据模型是否已知,可以分为基于模型的强化学习(Model-Based Reinforcement Learning,Model-Based RL)和无模型强化学习(Model-Free Reinforcement Learning,Model-Free RL)。两种强化学习方法各有优势,基于模型的强化学习拥有较高的学习效率,最典型的就是AlphaGo和AlphaZero。但大多数实际控制领域都是未知复杂环境,在模型未知的情况下完成控制任务,因此应用较多的是无模型强化学习。

1.2.2 价值函数

对于连续控制环境,状态信息非常巨大,无法对每一个状态和行为都采用查表式的方法存储每个状态和行为的价值。强化学习解决此类问题需要引入适当的

参数,恰当地选取描述状态的特征,通过构建一定的函数,来近似计算得到状态或行为价值。连续控制中这些由特征描述的状态,通过近似价值函数计算价值,而不必存储每一个状态的价值,大大提高了算法效率。带参数的价值函数,确定参数才能确定价值函数,参数求解多采用梯度下降法训练求解,例如经典的强化学习算法:深度Q学习。

基于价值的强化学习存在一些不足:在空间规模庞大和连续行为的状况下不适用;对随机策略的求取能力差,无法单独应付连续动作问题,导致强化学习的学习能力差。

1.2.3 策略函数

解决连续控制问题可以进行策略的直接学习,即将策略看成是状态和行为的带参数的策略函数。通过建立恰当的目标函数,利用智能体与环境的交互产生的奖励,学习策略函数的参数。基于策略函数的强化学习可以省略对状态的价值的学习过程,针对连续行为空间可以直接产生具体的行为值。

基于策略的强化学习存在的最明显的缺点是:在一些复杂问题的求解中,计算难度大,迭代时间过长。

1.2.4 演员评论家方法

Actor-Critic算法是基于价值函数和策略函数,分别创建网络。基于策略函数的网络,代替策略函数充当演员(Actor),产生行为与环境进行交互;基于价值函数的网络,代替行为价值函数充当评论家(Critic),评价演员的表现,并指导演员的后续动作。这种算法一方面基于价值函数进行策略评估和优化,另一方面优化的策略函数又会使价值函数更加准确地反应状态的价值,两者互相促进最终得到最优策略。

2 深度强化学习主要算法

深度强化学习的主要算法有两种类型:基于值函数的DRL和基于策略梯度的DRL。主要算法如表1所列。

2.1 深度Q学习算法及其改进

深度Q网络[25](Deep Q-Network,DQN)是基于使用卷积神经网络来代替强化学习的近似价值函数,原理是:利用神经网络的非线性表示能力,表示出某一确定环境下所有可能行为及其对应的价值。参数 $Q(s,a)$ 是针对特定状态产生的状态行为价值对,其中 s 表示状态,a 表示行为。DQN算法通过训练神经网络,替换对非线性函数的参数的求解。DQN算法的核心是目标函数、目标网络和经验回放,使DQN算法较好地学习得到强化学习任务的价值函数。

DQN算法使用一个权重参数为 θ 的深度卷积神经网作为动作值函数的网络模型,通过该模型 $Q(s,a,\theta)$ 模拟动作值函数 $Q^{\pi}(s,a)$,即:

$$Q(s,a,\theta)\approx Q^{\pi}(s,a) \tag{2}$$

DQN使用均方误差(Mean-square Error)定义目标函数,作为深度神经网络的损失函数,公式为:

$$L_i(\theta_i)=E\left[\underbrace{(r+\gamma\max_{a'}Q(s',a',\theta')}_{\text{Target Net}}-\underbrace{Q(s',a',\theta')}_{\text{Predict Net}}\right]^2 \tag{3}$$

式中,参数 s' 和 a' 为下一时间步的状态和动作,γ 为折扣因子。该目标Q值使用目标网络(Target Net)进行预测,而当前Q值使用预测网络(Predict Net)进行预测,使用均方误差计算Q-learning的时间差分误差。计算网络模型参数 θ 的梯度公式为:

$$\nabla_{\theta_i}L_i(\theta_i)=E\Big[(r+\gamma\max_{a'}Q(s',a',\theta_{i-1})-Q(s,a,\theta_i))\nabla_{\theta_i}Q(s,a,\theta_i)\Big] \tag{4}$$

式中,i 代表迭代次数。DQN使用小批量随机梯度下降法实现网络模型对目标函数的优化。每产生一个行为 a 和环境实际交互后,神经网络都会进行一次学习并更新一次参数。

DQN算法可以通过Q值实现对环境的端对端控制,在Atari2600游戏中取得超越人类的成绩[26]。其主要不足为:不能保证一直收敛,因为这种估计目标值的算法过于乐观,高估了一些情况下的最优值,导致算法将次优行为价值认定为最优行为价值。后续对DQN的改进方法中,根据侧重点的不同,改进方向可以分为:改进训练算法、改进神经网络结构、改进学习机制、新提出RL算法这四大类,不少改进方法在解决旧问题的同时,也带来了新问题。比如:Van Hasselt等提出双价值网络的DDQN[27]被认为较好地解决了价值高估问题,但带来了新的价值低估问题,还需要进一步的研究。Anschel等[28]提出平均DQN,基于过去一定步数学习的Q值的平均,再取最大值作为新的目标值,这种方法提高了稳定性,在众多游戏测试中优于DQN和DDQN,但也带来了训练时间大、成本高的问题。DQN算法和主要扩展及其所属方法分类如表2所示。

除了在游戏控制方面,DQN及其扩展算法在其他连续控制领域中有很多应用尝试。Liu等[37]提出一种基于DQN的无人机空战智能决策方法,采用Q网络实现动作值函数的精确拟合,仿真结果证明了DNQ算法在行为与奖励两方面都有突出表现;Huang等[38]设计了一种DQN算法来优化无人机的导航与路线,数值结果表明,设计的DQN导航可以给出较好的测量;Sharma等[39]

表1 主要深度强化学习算法分类

基于值函数的经典算法	基于策略梯度的经典算法
时序差分学习[20](Temporal-Difference Learning,TD Learning)	深度确定性策略梯度[22](Deep Deterministic Policy Gradient,DDPG)
值函数的Q学习[21](Q learning)	信赖域策略优化[23](Trust Region Policy Optimization,TRPO)
	异步优势演员评论家算法[24](Asynchronous Advantage Actor-Critic,A3C)

表2 DQN算法的改进算法、解决问题和实验验证结果

名称	解决问题	实验结果
双Q学习(DDQN)[27]	解决了Q学习中的过估计问题	在Atari游戏中表现优于DQN
平均DQN[28]	减少DRL算法的不稳定性和可变性对其产生的负面影响，使训练过程更加稳定，并通过减少目标逼近误差来提高性能	—
深度循环Q网络(DRQN)[29-30]	解决实际应用中状态存在部分可观测性和噪声干扰的问题	在Atari游戏中表现达到DQN水平
优先经验回放DQN[31]	考虑序列中各个转移元组的重要程度，提高重要元组回放的频率	提高了DQN和DDQN在游戏中的表现水平
竞争构架DQN[32]	分别表示状态值和行动优势	有助于跨行动泛化，学习多步引导目标
噪声DQN[33]	该算法策略的随机性可以用来帮助有效的探索	有效提高DQN的表现
Rainbow[34]	集成A3C、DQN、DDQN、优先化DDQN、决斗DDQN、分布式DQN和噪声DQN等多种技术特点，能克服多种缺点，不再局限于解决单一问题	57款雅达利游戏中，Rainbow算法明显优于其他算法
深度多Q学习[35]	克服DNN来做值函数逼近时出现的不稳定性	在大多数情况下，表现优于DQN，平均收益高达DQN的2.5倍
示范DQN[36]	减少和环境交互次数，提高学习效率	训练速度快，Atari游戏中表现优秀

提出一种基于视觉的DQN算法来控制四旋翼无人机的自主着陆，模拟结果表明，仅需低分辨率的相机就能实现着陆，在某些状况下优于人类驾驶员。Ao[40]提出了一种基于DQN(Deep Q-learning Network)的热过程控制方法，通过设计的DQN控制器的水箱水位控制系统仿真实验证明，DQN算法可以很好地应用于热过程控制。可预见，DQN及其改进型号，将会在更多控制领域得到实际应用，但这些控制任务，过于依赖状态控制，尤其是最终状态，所以这些应用还停留在实验阶段。

DQN算法实现了深度学习和强化学习的结合，对深度强化学习的发展有重要意义。但DQN算法及其改进类型在实际应用中存在不足：无法处理连续动作控制任务。

2.2 基于策略梯度的深度强化学习算法

2.2.1 深度确定性策略梯度算法及其改进

深度确定性策略梯度(DDPG)算法是基于深度学习、DQN算法、Actor-Critic网络的确定性策略算法，2016年DeepMind团队首次提出。相对于DPG，DDPG的核心改变是采用深度神经网络建立Actor和Critic的近似价值函数，并使用深度学习训练网络，Actor网络直接生成确定的行为，Critic网络评估策略的优劣。DDPG具有更高的学习效率，将复杂的控制问题直接与策略行为挂钩，是目前应用于复杂、连续控制的重要算法。

DDPG结合了之前算法的优点，特别是DQN的改进方案，具有更高的学习效率，将复杂的控制问题直接与策略行为挂钩，应用领域有：机器人控制、自动驾驶、无人机等。Casas[41]使用DDPG优化控制交通信号灯，将区域而非单独路口车辆检测器的信息当作输入，智能处理交通信号灯问题，改善了各个路口信号灯固定不变引起的交通不协调问题；Phaniteja等[42]使用DDPG进行具有27个自由度的机器人保持平衡的关节空间轨迹训练；Do等[43]开发的机器人浇筑实验，在避免碰撞和倾洒的情况下将液体倒入指定高度。应用较多还有无人驾驶领域，众多的无人汽车与无人飞行器基于DDPG进行使用，也在实际应用中推动DDPG的发展。

DDPG最重要的突破就是在解决连续控制问题上具有较高效率，缺点有训练时间过长、训练数据需求大、训练初期的学习策略不稳定等。

Houn等[44]提出一种基于知识的DDPG算法(Knowledge-driven Deep Deterministic Policy Gradient，KDDPG)，能够在没有大量数据的情况下，较稳定地让机器人完成装配学习。Zheng等[45]提出一种自适应双引导DDPG算法(Self-Adaptive Double Bootstrapped DDPG，SOUP)，将一根DDPG算法扩展到多个演员评论即架构，通过多对搭配使用的演员和评论家，学习多个策略并评估，从而得到更高的学习效率。Zhang等[46]针对DDPG算法训练数据需求大、训练效率低的问题，提出了异步章节式DDPG(Asynchronous Episodic DDPG，AE-DDPG)，AE-DDPG中的智能体可以同时与多个随机环境交互，从而实现很高的数据吞吐量，并采用情景控制思维[47]重新设计DDPG的经验回放，使智能体能够快速锁定高回报政策。

2.2.2 优势函数

优势函数是衡量一个动作带来的回报的重要手段，是计算A3C算法、TRPO算法的重要组成。已知Q值函数 $Q_\pi(S_t,a_t)$ 和状态值函数 $A_\pi(S_t)$，优势函数 $A_\pi(s,a)$ 的计算公式为：

$$A_\pi(s,a)=Q_\pi(s,a)-V_\pi \tag{5}$$

值函数 $V(s)$ 可以理解为在该状态下所有可能动作所对应的动作值函数乘以采取该动作的概率之和。动作值函数 $Q(s,a)$ 是单个动作所对应的值函数，$Q_\pi(s,a)-V_\pi(s)$ 能评价当前动作值函数相对于平均值的大小。这里的优势值指的是动作值函数相对于当前状态的值函数的优势。如果优势函数大于零，则说明该动作比平均动作好，如果优势函数小于零，这说明当前动作还不如平均动作好。

2.2.3　信赖域策略优化算法及其改进

策略梯度的方法存在问题：不能保证得到合适的步长使学习最有效，方法迭代的效果受步长影响较大，如步长太小，训练效率太低，如步长过大，噪声影响反馈信号，学到的可能是更坏的策略。找到合适的步长，保证学习效果最起码不会变差，即保证策略更新后的回报函数单调递增，John Schulman提出了信任域策略优化方法（Trust Region Policy Optimization，TRPO）。

TRPO算法最大的优势在于确保策略模型在优化模型时单调提升，稳定地改进策略。TRPO算法的核心思想是建立在优势函数上，主要支撑是找到一种衡量策略之间优劣的计算方法，并以此为目标最大化新策略和旧策略相比的优势。单调提升算法的整体思路简单，但具体的设计和计算非常复杂。

TRPO算法最大的优点是能保证策略始终朝着好的方向持续更新，缺点主要有：计算过程复杂、对策略与环境的交互依赖大、缺乏步长选择准则。

Jha等[48]针对TRPO算法步长选取准则不足、收敛速度慢等缺点，提出准牛顿信赖域策略优化算法（Quasi-Newton Trust Region Policy Optimization，QNTRPO），QNTRPO与TRPO的主要不同在于每次策略迭代的计算步骤，QNTRPO在相同的计算成本下比TRPO有更大的计算速度；Gupta等[49]提出一种合作强化学习算法，将信赖域策略优化算法扩展到大型智能体控制任务，可以让几十个、数百个智能体合作来完成任务，并可以在连续动作空间缩放，称为PS-TRPO。实验证明，PS-TRPO算法在多智能体协同领域，比DQN和DDPG拥有更多优势。根据PS-TRPO算法开发的三个智能体学习系统，在连续动作空间也有很好的表现；为提高TRPO在稀疏奖励的强化学习中的表现，Zhang等[50]提出了后见信任区域策略优化算法（Hindsight Trust Region Policy Optimization，HTRPO），HTRPO使用二次KL估计逼近，减少方差，提高学习稳定性，设计了后知目标过滤机制，缩小后知目标空间与原始目标空间的差异，获得更好的学习效果。HTRPO在一些离散和连续的控制任务，比TRPO有更强的学习能力。

Schulman等[51]在TRPO算法基础上提出了最近策略优化算法（Proximal Policy Optimization，PPO），与TRPO算法最大的不同是PPO算法可以实现多个时期的小批量更新，实现比TRPO更简单的计算，也具有更好的样本复杂度。Heess等[52]提出了分步式PPO算法（Distribute PPO，DPPO），提高智能体在奖励信号有限的条件下的学习水平。在控制实验中，DPPO在更高的效率下实现了类似于TRPO的性能。除此之外，Shani[53]、Liu H[54]、Liu B[55]等都对TRPO信任域策略算法做了改进，在算法或者应用上取得了一定进展。

2.2.4　异步优势演员评论家网络算法及其改进

Mnih等人基于异步强化学习（Asynchronous Reinforcement Learning，ARL）的思想，提出一种轻量级的深度强化学习框架：异步优势的演员评论家算法（Asynchronous Advantage Actor-Critic，A3C），A3C算法使用异步梯度下降算法优化深度网络模型，并结合多种强化学习算法，能够使深度强化学习算法基于CPU快速地进行学习。

A3C算法核心是：优势函数演员评论家算法和异步算法的结合。演员评论家算法在A3C算法中，包括基于策略学习的演员和基于价值学习的评论家。优势演员评论家算法，是在演员评论家算法基础上，对评论家模型进行更新时，引入优势函数的概念，以确定其网络模型输出动作的好坏程度，使得对策略梯度的评估偏差更少。A3C的异步操作是指：利用多个智能体与多个环境进行交互，提高学习效率。异步构架主要由环境、工人和全局网络组成，其中每个工人作为一个智能体与一个独立的环境进行交互，并有属于自身的网络模型。不同的工人同时与环境进行交互，其执行的策略和学习到的经验都独立于其他工人。因此该多智能体异步探索的方式能够比使用单个工人进行探索的方式更好、更快、更多样性地工作。

A3C算法可以得到更好的收敛性，在高维控制和连续空间的表现更好。Lin等[56]在A3C算法基础上，提出一种协同异步优势演员-评论家算法（collaborative Asynchronous Advantage Actor-Critic，cA3C），智能体在线学习深度知识提取，实现自适应的知识转移。实验表明，cA3C算法的收敛水平比A3C更高，也获得了更高的奖励；针对A3C算法的优势函数存在方差，影响性能的问题，Chen等[57]提出了一种平均异步优势演员-评论家算法（Averaged Asynchronous Advantage Actor-Critic，Averaged-A3C），降低优势函数的方差。Averaged-A3C主要改进是对已经学习过的状态值取平均来计算优势函数，提高训练过程的稳定性。实验表明，Averaged-A3C比A3C算法拥有更好的性能和稳定性。Kartal等[58]将A3C算法与末端预测（Terminal Prediction，TP）结合，提出了一种末端预测的异步优势演员评论即算法（A3C-TP），主要改进是智能体在学习控制策略时，预测目前状态到最终状态的距离从而促进学习。在Atari游戏和双足步行者领域的实验结果表明：在大多数测试领域中，A3C-TP的表现优于标准A3C；Labao等[59]提出一种融合梯度（Gradient Sharing）共享的异步优势演员-评论家算法（A3C-GS），A3C-GS算法具有在短期内自动分散员工政策进行探索的特性，在政策多样化的情况下，理论上算法长期收敛于最优政策。实验表明，A3C-GS算法在高维环境中比其他基于策略梯度的算法表现更好，取得了更高的分数。Hernandez-Leal[60]、Wang[61]、Holliday[62]等都对A3C算法进行了改进实验，取得了一定进展。

总的来说，A3C算法，降低了DRL对计算机计算性

表3 主要研究结果提炼表

名称	核心机制	解决问题及优点	存在的不足	适用场景
DQN算法	目标函数、目标网络和经验回放	实现端对端控制,在游戏控制中有良好表现,改进类型多	计算依赖状态,不能得出具体控制策略,不适用于状态差距大的复杂控制	适用于状态为主要指导的控制,如游戏控制。在智能制造中,主要应用在资源调度、路径规划等方面
DDPG算法	DQN算法、演员评论家算法和确定性策略梯度	能够产生确定的策略,实现连续型动作的控制	计算过程复杂,学习周期长,步长难选择,训练成本大	学习连续的行为控制策略的经典方法。在无人驾驶、机器人运动控制、交通信号灯控制等方面应用较大
TRPO算法	优势函数、函数近似、步长选择	克服策略梯度算法难以选择合适步长的问题,确保策略模型在优化模型时单调提升,稳定地改进策略	步长选取准则不足、计算复杂、收敛速度慢	连续动作控制,如机器人控制和无人驾驶
A3C算法	优势函数、演员评论家算法、异步算法	多个智能体与多个环境交互,比单智能体学习效率更高、学习效果更好,学习更全面	可能存在策略延迟,即产生当前训练样本的策略非当前要更新的策略,导致算法不稳定	离散、连续型动作控制都有较好表现

能的要求,并且在效果、时间和资源消耗上都优于传统方法。但对于实际问题,计算机计算能力仍然限制了A3C算法的潜力。

2.3 其他相关研究

除了基于值函数和基于策略梯度的算法进展的研究,通过其他角度对深度强化学习的研究也取得了一些进展。基于模型的强化学习(Model-Based RL)算法能够更高效地学习,但很难扩展到深度神经网络这种表达能力强的模型,无法应用到复杂、高维的控制任务。将深度神经网络应用到Model-Based RL上的研究很有价值。Luo等[63]提出了一种基于模型的随机下界优化算法(Stochastic Lower Bounds Optimization,SLBO),通过实验表明,在一系列连续控制基准测试任务中,SLBO只需要较少样本就达到了比原算法更高的学习率。Nagabandi[64]、Ebert[65]、Huang[66]等都对Model-Based RL应用在高维控制任务上做了一些研究。

强化学习是基于奖励调整策略,但在一些复杂任务中,环境在到达最终结果前回报稀疏,强化学习任务面临反馈稀疏的问题,影响学习效率。Kulkarni等[67]提出一种分层DQN算法(hierarchical-DQN,h-DQN),开创了层次强化学习算法(Hierarchical Reinforcement Learning,HRL),层次强化学习将控制任务分成若干层次,从多层策略中学习,每一层都负责在不同的时间和行为抽象层面进行控制。最低级别的策略负责输出行动,使更高级别的策略可以在更抽象的目标和更长的时间尺度上自由运作。Vezhnevets[68]、Nachum[69]、Rafati[70]等都对HRL做了研究。

对于复杂任务,奖励稀疏问题会导致设定奖励函数非常困难,给出的奖励函数也并非完全可以衡量决策的好坏。基于Ng等[71]提出的假设专家最优思想,利用专家数据采用函数近似的方法建立奖励函数,这种称为深度逆向学习[72]的方法,可以作为研究复杂环境下的奖励函数的新方法。You等[73]在无人驾驶中,通过收集专家驾驶员的大量演示,使用深度逆向强化学习方法学习基于数据的最优驾驶策略,利用神经网络逼近专家驾驶员数据的奖励函数,实现期望的驾驶行为。Fahad等[74]研究了机器人学习人类导航的深度逆向强化学习方法。此外,在过程控制领域,某些复杂控制任务,通过逆向强化学习方法,让机器学习专家知识,获得较高的控制水平,也是很有价值的研究方向。

2.4 主要研究成果对比分析

上述深度强化学习的研究,都取得了一定成果,但这些研究的原理和研究角度的不同,决定了每种方法的特性和应用场景。主要提炼内容如表3所示。

除了从应用性能角度,DRL算法的改进,对设备的要求程度和学习时间也是重要评判依据。这也是DQN发展到A3C算法的推动力之一,即提高学习速度和减少对硬件系统的依赖。其中一些算法的学习时间、计算机设备,以及在雅达利游戏中与原始实验的得分对比率[24],如表4所示。

表4 一些改进算法的学习时间和设备表

算法名称	训练时间/d	计算机类型	得分对比率/%
DQN	8	GPU	121.9
DDQN	8	GPU	332.9
竞争构架DQN	8	GPU	343.8
噪声DQN	8	GPU	463.6
A3C	4	CPU	623.0

综上,DRL算法发展的趋势是更高的表现、更高的学习效率以及对硬件的更低依赖。

3 在智能制造中的应用展望

3.1 智能装配

智能机器人控制是人工智能的标志性成果之一,可以分担人类工作。工业机器人可以取代人类在一些高温、高压或者其他不适环境下工作,也可以完成一些体

力工作。除此之外,还发展到一些高精度装配工作。传统的机器人编程通过定义装备位置和动作进行工作,这种编程比较复杂,且不能适应环境变化,只能执行程序固定的工作内容,如果环境发生改变或者调整生产线,就需要重新编程。深度强化学习提供了解决这些问题的途径,已经得到了多次验证。Inoue等[75]针对传统机器人编程复杂、调参困难的问题,提出一种基于递归神经网络的DRL算法模型,通过使用一个7轴铰接机器人手臂,完成精度较高的插孔实验,验证模型的有效性,并计划利用模型学习不同的生产环境参数,缩短适应新生产时间。Schoettler等[76]针对传统工业机器人缺乏自适应的特点,提出了利用DRL算法的解决方案,并通过一些复杂规格和形状的实验,证明DRL可以解决复杂的工业装配任务。Zhao等[77]提出一种基于DRL的工件装配序列规划系统(Assembly Sequence Planning for Workpieces,ASPW)的设计方法,对复杂装配产品进行自动排序,提高装配效率。在建立的实验平台上,设计出了一种新的ASPW-DQN算法,克服DRL算法缺乏奖励和缺乏培训环境的困难,在实验中取得了较高的准确度。Wu[78]、Vecerik[79]、Xu[80]、Luo[81]等都对DRL算法应用在装配机器人上做了实验研究,取得了一定成果。相信在未来,工作精度高、适应能力强、岗位转换容易的装配机器人会出现。

3.2 智能运输与路径规划

运输机器人极大减轻了人类繁重的搬运工作,检测机器人可到人类不能到达的地方对工业设备进行检测,这都离不开机器人路径规划。传统机器人主要依靠编程或者利用传感器进行固定路径行驶,可以满足简单的货物运输。传统机器人在较复杂的工厂中,易受干扰;在出发点、目的地或者路径状况导致的运输线路改变时,需要重新编程、调试,灵活性、经济性较差;由于对一些恶劣环境掌握有限,传统算法对未知环境的适用性较差。此外,研究人员还提出了蚁群优化[82]、粒子群优化[83]、模拟退火[84]和遗传算法[85]等智能方法来解决全局路径规划问题[86],但这些方法在高维环境下表现不佳。DRL算法可以赋予机器人根据环境状态和任务变化,自主规划路径的能力。比如在某一通道占用的情况下,仍能找到另一条道路到达目的地。Zhou等[87]提出并验证了一种基于DQN的全局路径规划方法,能够使机器人在密集的环境中获得最优路径。机器人的输入方式是直接摄入图像,能够有效避开障碍物。Sui等[88]设计了一种并行深度DQN算法,求解多智能体约束的编队路径规划问题。Wang等[89]提出了一种基于双DQN和经验优先重放的移动机器人路径规划方法,能够通过感知周围环境的局部信息,在未知环境下规划路径,通过实验验证了可靠性。这些研究表明,智能路径规划能够让工业机器人拥有更强的工作能力,应用广泛,是人工智能研究的热点领域之一。

3.3 智能过程控制

过程控制任务的主流控制器,包括单回路和多回路PID控制器、模型预测控制器和各种非线性控制器,大多数现代工业控制器都是基于模型的,因此良好的性能需要高质量的过程模型。PID控制器和基于模型的控制器需要定期维护以保持性能。通常的做法是持续监控控制器的性能,并在性能下降时启动补救模型重新识别程序,维修过程通常是复杂的和资源密集型的,并且会导致工作中断,代价较高。此外,在高级的控制任务中,很难建立高质量的模型,导致这些控制器很难适用于非线性或者高维控制任务。DRL应用在过程控制领域,可以同时接受数据信息和高维信息,对环境的理解更具体,能够根据奖励不断学习,提高控制水平,在一定程度上时间越久,学习水平越高,保持较高的控制性能,节省维护成本。Andersen等[90]对DRL应用在过程控制做了相关理论研究。Spielberg等[91]提出一个基于数据的DRL控制器,通过与过程交互学习控制策略,并且通过大量仿真验证了DRL控制器的有效性和优越性。在未来,除了现有的自动控制环节,更多目前需要人工控制的岗位,也可以通过DRL算法取代。

3.4 新智能调度

传统的智能调度方法有:基于知识的系统、专家系统、遗传算法、模拟退火、神经网络和混合系统等,这些调度方式过于依赖人工调度的浅显知识[92],不能解决复杂的调度问题,并且实时控制性较差,将DRL应用于智能调度的新智能调度可以解决上述问题。新智能调度在智能制造中可以完成资源分配工作和任务分配工作,并且拥有一定实时反应能力。Singh等[93]针对运输资源的分配问题,提出一种基于DRL的车辆调度框架,通过与外部环境的相互,分别为每辆车辆学习最优策略。实验结果表明,该框架可以提高20%的车辆利用率,乘客等待时间与车辆巡航时间降低34%。这种方法没有考虑每个车辆间的相互影响,虽然降低了计算难度,但调度策略对于全局来说不能保证是最优策略,不适用智能制造全局调度的要求。Hua等[94]采用DRL算法进行资源调度,对分配任务中的限制条件进行考虑,采用无模型方法解决一些限制条件没有明确公式关系的问题。研究对既需要全局调度,又存在子区域调度的问题,采用A3C算法进行最优控制,在仿真实验中验证了有效性。但这种方法没有考虑子区域之间的相互影响,仍然存在全局最优考虑不足的情况。对于调度任务,调度环境越大,全局性越强,对计算的要求越高。在计算能力有限的情况下,适当将大区域调度分割为子区域调度,可以平衡计算与策略最优程度的关系。Mao等[95]的研究在智能调度中加入了有限考虑考虑级别。在智能制造中,某些生产或者资源需要优先考虑。具体方法是使用前馈神经网络结合演员评论家算法,实验结果表明,该算法都收敛于最优性间隙小于4%的理论上界。在考

虑了优先性和公平性的情况下,该算法具有较大实用价值。在未来可以继续考虑划分更多优先等级,并根据实际生产状况,自动调节优先等级。Guan等[96]研究了利用DRL算法解决在资源分配中的近实时性问题,即在单位时间里根据环境调整一下分配策略。该研究基于DRL算法设计了一种经济电力调度模型,利用DRL的连续控能力在单位时间间隔调整一次分配状态。这种方法既节省了计算资源,又具有一定实时控制性能,在大型资源分配中有较大应用前景。但这种方法不能保证收敛性,未来需要继续改善计算性能。Li[97]、Waschneck[98]、Liu[99]等从不同角度对DRL在智能调度中的应用进行研究。

3.5 其他应用展望

DRL拥有独特优势,在智能制造中拥有较大应用空间,除了上述应用展望之外,DRL还可以结合传统专家系统,产生新一代反馈学习专家系统,能够在给出专家知识之后,根据反馈继续学习,提升诊断水平和决策水平,这可以更好发挥专家系统的潜力。专家系统的建议一般假设为最优,限制性在于对问题的诊断水平,结合DRL的专家系统可以根据奖励提高诊断水平,提高故障与对策的匹配水平;DRL可以设计出一种对工业设备寿命或者状态的置信预测器,例如对一些高温高压的储存设备,将已经发生过的实际数据当作输入进行学习,一次性预测未来一定时间的各项参数,并在预测日期到达之后,与产生的真实数据对比,产生的反馈来修正预测网络,这样随着时间的增加,预测水平可以保证一定时间段的准确性。这种方法可以比传统检测更加可靠和经济,拥有较大的研究价值。

4 存在的问题和未来发展方向

深度强化学习(DRL)是比较新的技术,功能强大,但也存在必须解决的问题,并可能催生新的发展方向。

(1)DRL是机器模仿人类的方法,由于对人脑的了解还不够,还缺乏与之对应的人脑机理知识。比如深度学习(DL)的机器视觉对应人脑神经元的视觉机理,但强化学习(RL)的策略目前与人脑生物学知识的对应不足,限制了强化学习新的发展突破。未来需要对人脑有进一步的研究,并与深度强化学习理论对应,从而突破人造智能体的技术障碍。

(2)计算能力的提升是将深度强化学习应用在实际中的必备条件。目前主流改进方案是算法的提升和硬件设施的进步。随着云计算技术等网络大输出处理技术的进步,通过这些技术结合DRL,将DRL的计算任务在线分配处理,批次处理某一区域或者任务的计算,可以带来DRL计算速度的大提升。

(3)基于模型强化学习虽然目前应用受限,但未来发展潜力巨大。随着DRL学习能力的提高,智能体能够学习复杂环境的模型,并且可通过模型预测未来。对于一些复杂但封闭性较强的制造环境,基于模型的强化学习有较大的研究价值。

(4)DRL训练的支撑是反馈奖励,应用在工业过程控制中,如何充分利用专家数据提高学习能力,节省学习成本,是很有价值的研究方向。对于稀疏奖励任务,可以根据与专家做法的重合程度,设置短期奖励,提高学习效率。

(5)DRL是将DL和RL结合的技术,但DRL的控制,如机器人、无人车等,严重依赖DL的视觉输入,但DL目前只能发挥感知作用,无法取代力学分析等深层知识,造成DRL的一些仿真研究与现实应用有较大差距。未来将DL和RL分开研究,获得更高的稳定性,然后再拼装,也是很有价值的研究方向。

(6)DRL算法可以解决智能控制的程序问题,但与之匹配的工业硬件设施还没有相关标准,例如能处理海量工业数据的计算机、能够测量数据传输且接收指令的智能阀门等,都是与理论算法相匹配的研究重点。

5 结束语

本文对深度强化学习的原理进行了讲述,包括深度强化学习的原理,以及深度强化学习主要的算法发展。可以看到深度强化学习的算法已经发展出了很多成果,应用水平也不断提高。可以得到深度强化学习算法的发展方向有:更高效的学习、更快的计算结论、更准确的评估奖励。随着一些关键问题的解决,在未来的智能制造业中,深度强化学习可以担任更多角色。

参考文献:

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

[2] WU M,CHEN L.Image recognition based on deep learning[C]//Chinese Automation Congress,2015.

[3] DENG L,LI J,HUANG J T,et al.Recent advances in deep learning for speech research at Microsoft[C]//IEEE International Conference on Acoustics,2013:8604-8608.

[4] DENG L,YU D.Deep learning:methods and applications[J].Foundations & Trends in Signal Processing,2014,7(3).

[5] LIU S,WANG Y,YANG X,et al.Deep learning in medical ultrasound analysis:a review[J].Engineering,2019,5(2):261-275.

[6] SUTTON R S,BARTO A G.Introduction to reinforcement learning[M].Cambridge:MIT Press,1998.

[7] SUTTON R S,BARTO A G.Reinforcement learning:an introduction[M].Cambridge:MIT Press,2018.

[8] ARULKUMARAN K,DEISENROTH M P,BRUNDAGE M,et al.Deep reinforcement learning:a brief survey[J].IEEE Signal Processing Magazine,2017,34(6):26-38.

[9] PADEN B,CAP M,YONG S Z,et al.A survey of motion planing and control techniques for self-driving urban vehi-

cles[J].IEEE Transactions on Intellingent Vehicles,2016,1(1):33-55.
[10] PAN S J,YANG Q.A survey on transfer learning[J].IEEE Transactions on Knowledge and Data Engineering,2009,22(10):1345-1359.
[11] GUO Q L,ZHANG M.An agent-oriented approach to resolve scheduling optimization in intelligent manufacturing[J].Robotics and Computer Integrated Manufacturing,2010,26(1):39-45.
[12] KEIZHEVSKY A,SUTSKEVER I,HINTON G E.Imagent classification with deep convolutional neural networks[C]//Proceedings of the Annual Conference on Neural Information Processing Systems,2012:1097-1105.
[13] LECUN Y,BOTTOU L.Gradient-based learning applied to document recognition[J].Proceedings of the IEEE,1998,86(11):2278-2324.
[14] KRIZHEVSKY A,SUTSKEVER I,HINTON G E,et al.ImageNet classification with deep convolutional neural networks[C]//Neural Information Processing Systems,2012:1097-1105.
[15] SIMONYAN K,ZISSERMAN A.Very deep convolutional networks for large-scale image recognition[J].arXiv:1409.1556,2014.
[16] HE K,ZHANG X,REN S,et al.Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,2016:770-778.
[17] BENGIO Y,SIMARD P,FRASCONI P.Learning long-term dependencies with gradient descent is difficult[J].IEEE Transactions on Neural Networks,1994,5(2):157-166.
[18] HOCHREITER S,SCHMIDHUBER J.Long short-term memory[J].Neural Computation,1997,9(8):1735-1780.
[19] SCHUSTER M,PALIWAL K K.Bidirectional recurrent neural networks[J].IEEE Transactions on Signal Processing,1997,45(11):2673-2681.
[20] SUTTON R S.Learning to predict by the methods of temporal differences[J].Machine Learning,1988,3(1):9-44.
[21] WATKINS C J C H,DAYAN P.Q-learning[J].Machine Learning,1992,8(3/4):279-292.
[22] LILLICRAP T P,HUNT J J,PRITZEL A,et al.Continuous control with deep reinforcement learning[J].arXiv:1509.02971,2015.
[23] SCHULMAN J,LEVINE S,MORITZ P,et al.Trust region policy optimization[C]//Proceedings of the International Conference on Machine Learning,Lugano,Switzerland,2015:1889-1897.
[24] MNIH V,BADIA A P,MIRZA M,et al.Asynchronous methods for deep reinforcement learning[C]//International Conference on Machine Learning,2016:1928-1937.
[25] MNIH V,KAVUKCUOGLU K,SILVER D,et al.Playing Atari with deep reinforcement learning[C]//Proceedings of the Workshops at the 26th Neural Information Processing Systems 2013,Lake Tahoe,USA,2013:201-220.
[26] MNIH V,KAVUKCUOGLU K,SILVER D,et al.Human-level control through deep reinforcement learning[J].Nature,2015,518(7540):529-553.
[27] VAN HASSELT H,GUEZ A,SILVER D.Deep reinforcement learning with double q-learning[J].arXiv:1509.06461,2015.
[28] ANSCHEL O,BARAM N,SHIMKIN N.Averaged-DQN:variance reduction and stabilization for deep reinforcement learning[C]//International Conference on Machine Learning,2017:176-185.
[29] HAUSKNECHT M,STONE P.Deep recurrent Q-learning for partially observable MDPs[J].arXiv:1507.06527,2015.
[30] NARASIMHAN K,KULKARNI T,BARZILAY R.Language understanding for text-based games using deep reinforcement learning[J].arXiv:1506.08941,2015.
[31] SCHAUL T,QUAN J,ANTONOGLOU I,et al.Prioritized experience replay[J].arXiv:1511.05952,2015.
[32] WANG Z,SCHAUL T,HESSEL M,et al.Dueling network architectures for deep reinforcement learning[C]//International Conference on Machine Learning,2016:1995-2003.
[33] FORTUNATO M,AZAR M G,PIOT B,et al.Noisy networks for exploration[J].arXiv:1706.10295,2017.
[34] HESSEL M,MODAYIL J,VAN HASSELT H,et al.Rainbow:combining improvements in deep reinforcement learning[J].arXiv:1710.02298,2017.
[35] DURYEA E,GANGER M,HU W.Exploring deep reinforcement learning with multi q-learning[J].Intelligent Control and Automation,2016,7(4):129-144.
[36] HESTER T,VEEERIK M,PIETQUIN O,et al.Learning from demonst rations for real world reinforcement learning[J].arXiv:1704.03732,2017.
[37] LIU P,MA Y.A deep reinforcement learning based intelligent decision method for UCAV air combat[C]//Asian Simulation Conference,2017:274-286.
[38] HUANG H,YANG Y,DING Z,et al.Deep learning-based sum data rate and energy efficiency optimization for MIMO-NOMA systems[J].IEEE Transactions on Wireless Communications,2020.
[39] POLVARA R,PATACCHIOLA M,SHARMA S,et al.Toward end-to-end control for UAV autonomous landing via deep reinforcement learning[C]//2018 International Conference on Unmanned Aircraft Systems(ICUAS),2018:115-123.
[40] AO T,SHEN J,LIU X.The application of DQN in thermal process control[C]//2019 Chinese Control Conference(CCC),2019:2840-2845.
[41] CASAS N.Deep deterministic policy gradient for urban traffic light control[J].arXiv:1703.09035,2017.
[42] PHANITEJA S,DEWANGAN P,GUHAN P,et al.A deep reinforcement learning approach for dynamically stable

inverse kinematics of humanoid robots[C]//2017 IEEE International Conference on Robotics and Biomimetics (ROBIO),2017:1818-1823.

[43] DO C,GORDILLO C,BURGARD W,et al.Learning to pour using deep deterministic policy gradients[C]//Intelligent Robots and Systems,2018:3074-3079.

[44] HOU Z,DONG H,ZHANG K,et al.Knowledge-driven deep deterministic policy gradient for robotic multiple peg-in-hole assembly tasks[C]//2018 IEEE International Conference on Robotics and Biomimetics (ROBIO),2018:256-261.

[45] ZHENG Z,YUAN C,LIN Z,et al.Self-adaptive double bootstrapped DDPG[C]//Twenty-Seventh International Joint Conference on Artificial Intelligence,2018.

[46] ZHANG Z,CHEN J,CHEN Z,et al.Asynchronous episodic deep deterministic policy gradient:towards continuous control in computationally complex environments[J].IEEE Transactions on Cybernetics,2019.

[47] BLUNDELL C,URIA B,PRITZEL A,et al.Model-free episodic control[J].arXiv:1606.04460,2016.

[48] JHA D K,RAGHUNATHAN A U,ROMERES D.Quasi-newton trust region policy optimization[C]//Conference on Robot Learning,2020:945-954.

[49] GUPTA J K,EGOROV M,KOCHENDERFER M.Cooperative multi-agent control using deep reinforcement learning[C]//International Conference on Autonomous Agents and Multiagent Systems.Cham:Springer,2017:66-83.

[50] ZHANG H,BAI S,LAN X,et al.Hindsight trust region policy optimization[J].arXiv:1907.12439,2019.

[51] SCHULMAN J,WOLSKI F,DHARIWAL P,et al.Proximal policy optimization algorithms[J].arXiv:1707.06347,2017.

[52] HEESS N,TB D,SRIRAM S,et al.Emergence of locomotion behaviours in rich environments[J].arXiv:1707.02286,2017.

[53] SHANI L,EFRONI Y,MANNOR S.Adaptive trust region policy optimization:global convergence and faster rates for regularized MDPs[C]//Proceedings of the AAAI Conference on Artificial Intelligence,2020:5668-5675.

[54] LIU H,WU Y,SUN F.Extreme trust region policy optimization for active object recognition[J].IEEE Transactions on Neural Networks and Learning Systems,2018,29(6):2253-2258.

[55] LIU B,CAI Q,YANG Z,et al.Neural proximal/trust region policy optimization attains globally optimal policy[J].arXiv:1906.10306,2019.

[56] LIN K,WANG S,ZHOU J.Collaborative deep reinforcement learning[J].arXiv:1702.05796,2017.

[57] CHEN S,ZHANG X F,WU J J,et al.Averaged-A3C for asynchronous deep reinforcement learning[C]//International Conference on Neural Information Processing.Cham:Springer,2018:277-288.

[58] KARTAL B,HERNANDEZ-LEAL P,TAYLOR M E.Terminal prediction as an auxiliary task for deep reinforcement learning[C]//Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment,2019,15(1):38-44.

[59] LABAO A B,MARTIJA M A M,NAVAL P C.A3C-GS:adaptive moment gradient sharing with locks for asynchronous actor-critic agents[J].IEEE Transactions on Neural Networks and Learning Systems,2020.

[60] HERNANDEZ-LEAL P,KARTAL B,TAYLOR M E.Agent modeling as auxiliary task for deep reinforcement learning[C]//Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment,2019:31-37.

[61] WANG X,ZHUANG Z,ZOU L,et al.An accelerated asynchronous advantage actor-critic algorithm applied in papermaking[C]//2019 Chinese Control Conference(CCC),2019:8637-8642.

[62] HOLLIDAY J B.Improving asynchronous advantage actor critic with a more intelligent exploration strategy[Z].2018.

[63] LUO Y,XU H,LI Y,et al.Algorithmic framework for model-based deep reinforcement learning with theoretical guarantees[J].arXiv:1807.03858,2018.

[64] NAGABANDI A,KAHN G,FEARING R S,et al.Neural network dynamics for model-based deep reinforcement learning with model-free fine-tuning[C]//2018 IEEE International Conference on Robotics and Automation (ICRA),2018:7559-7566.

[65] EBERT F,FINN C,DASARI S,et al.Visual foresight:model-based deep reinforcement learning for vision-based robotic control[J].arXiv:1812.00568,2018.

[66] HUANG Z,HENG W,ZHOU S.Learning to paint with model-based deep reinforcement learning[C]//Proceedings of the IEEE International Conference on Computer Vision,2019:8709-8718.

[67] KULKARNI T D,NARASIMHAN K,SAEEDI A,et al.Hierarchical deep reinforcement learning:integrating temporal abstraction and intrinsic motivation[C]//Advances in Neural Information Processing Systems,2016:3675-3683.

[68] VEZHNEVETS A S,OSINDERO S,SCHAUL T,et al.Feudal networks for hierarchical reinforcement learning[J].arXiv:1703.01161,2017.

[69] NACHUM O,GU S S,LEE H,et al.Data-efficient hierarchical reinforcement learning[C]//Advances in Neural Information Processing Systems,2018:3303-3313.

[70] RAFATI J,NOELLE D C.Learning representations in model-free hierarchical reinforcement learning[C]//Proceedings of the AAAI Conference on Artificial Intelligence,2019:10009-10010.

[71] NG A Y,RUSSELL S J.Algorithms for inverse reinforcement learning[C]//Seventeenth International Conference on Machine Learning,2000:663-670.

[72] WULFMEIER M, ONDRUSKA P, POSNER I.Deep inverse reinforcement learning[J].arXiv:1507.04888,2015.

[73] YOU C, LU J, FILEV D, et al.Advanced planning for autonomous vehicles using reinforcement learning and deep inverse reinforcement learning[J].Robotics and Autonomous Systems,2019,114:1-18.

[74] FAHAD M, CHEN Z, GUO Y.Learning how pedestrians navigate:a deep inverse reinforcement learning approach[C]// 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems(IROS),2018:819-826.

[75] INOUE T, DE MAGISTRIS G, MUNAWAR A, et al. Deep reinforcement learning for high precision assembly tasks[C]//2017 IEEE/RSJ International Conference on Intelligent Robots and Systems(IROS),2017:819-825.

[76] SCHOETTLER G, NAIR A, LUO J, et al.Deep reinforcement learning for industrial insertion tasks with visual inputs and natural rewards[J].arXiv:1906.05841,2019.

[77] ZHAO M,GUO X,ZHANG X,et al.ASPW-DRL:assembly sequence planning for workpieces via a deep reinforcement learning approach[J].Assembly Automation, 2019,40(1):65-75.

[78] WU X, ZHANG D, QIN F, et al.Deep reinforcement learning of robotic precision insertion skill accelerated by demonstrations[C]//2019 IEEE 15th International Conference on Automation Science and Engineering(CASE), 2019:1651-1656.

[79] VECERIK M,SUSHKOV O,BARKER D,et al.A practical approach to insertion with variable socket position using deep reinforcement learning[C]//2019 International Conference on Robotics and Automation(ICRA), 2019:754-760.

[80] XU J,HOU Z,WANG W,et al.Feedback deep deterministic policy gradient with fuzzy reward for robotic multiple peg-in-hole assembly tasks[J].IEEE Transactions on Industrial Informatics,2018,15(3):1658-1667.

[81] LUO J,SOLOWJOW E,WEN C,et al.Deep reinforcement learning for robotic assembly of mixed deformable and rigid objects[C]//2018 IEEE/RSJ International Conference on Intelligent Robots and Systems(IROS), 2018:2062-2069.

[82] FAN X,LUO X,YI S,et al.Optimal path planning for mobile robots based on intensified ant colony optimization algorithm[C]//IEEE International Conference on Robotics, Intelligent Systems and Signal Processing, 2003:131-136.

[83] SUN B,CHEN W,XI Y.Particle swarm optimization based global path planning for mobile robots[J].Control and Decision,2005,20(9):1052.

[84] MIAO H,TIAN Y C.Robot path planning in dynamic environments using a simulated annealing based approach[C]// 2008 10th International Conference on Control,Automation,Robotics and Vision,2008:1253-1258.

[85] TU J,YANG S X.Genetic algorithm based path planning for a mobile robot[C]//2003 IEEE International Conference on Robotics and Automation,2003:1221-1226.

[86] MASEHIAN E,SEDIGHIZADEH D.Classic and heuristic approaches in robot motion planning-a chronological review[J].World Academy of Science,Engineering and Technology,2007,23(5):101-106.

[87] ZHOU S,LIU X,XU Y,et al.A deep Q-network(DQN) based path planning method for mobile robots[C]//2018 IEEE International Conference on Information and Automation(ICIA),2018:366-371.

[88] SUI Z,PU Z,YI J,et al.Path planning of multiagent constrained formation through deep reinforcement learning[C]//2018 International Joint Conference on Neural Networks(IJCNN),2018:1-8.

[89] WANG Y,FANG Y,LOU P,et al.Deep reinforcement learning based path planning for mobile robot in unknown environment[J].Journal of Physics:Conference Series, 2020,1576(1):012009.

[90] ANDERSEN R E,MADSEN S,BARLO A B K,et al. Self-learning processes in smart factories:deep reinforcement learning for process control of robot brine injection[J].Procedia Manufacturing,2019,38:171-177.

[91] SPIELBERG S,TULSYAN A,LAWRENCE N P,et al. Deep reinforcement learning for process control:a primer for beginners[J].arXiv:2004.05490,2020.

[92] BROWN D E,MARIN J A,SCHERER W T.A survey of intelligent scheduling systems[M]//Intelligent scheduling systems.Boston,MA:Springer,1995:1-40.

[93] SINGH A,ALABBASI A,AGGARWAL V.A distributed model-free algorithm for multi-hop ride-sharing using deep reinforcement learning[J].arXiv:1910.14002,2019.

[94] HUA H,QIN Y,HAO C,et al.Optimal energy management strategies for energy Internet via deep reinforcement learning approach[J].Applied Energy,2019,239:598-609.

[95] MAO C,LIU Y,SHEN Z J M.Dispatch of autonomous vehicles for taxi services:a deep reinforcement learning approach[J].Transportation Research Part C:Emerging Technologies,2020,115:102626.

[96] GUAN J,TANG H,WANG K,et al.A parallel multi-scenario learning method for near-real-time power dispatch optimization[J].Energy,2020:117708.

[97] LI J,YU T,ZHU H,et al.Multi-agent deep reinforcement learning for sectional AGC dispatch[J].IEEE Access, 2020,8:158067-158081.

[98] WASCHNECK B,REICHSTALLER A,BELZNER L, et al.Optimization of global production scheduling with deep reinforcement learning[J].Procedia CIRP,2018,72: 1264-1269.

[99] LIU W,ZHUANG P,LIANG H,et al.Distributed economic dispatch in microgrids based on cooperative reinforcement learning[J].IEEE Transactions on Neural Networks & Learning Systems,2018:2192-2203.