强化学习工业应用

近年来随着人工智能技术的发展，深度学习和强化学习技术已经在很多领域打破了传统方法的壁垒，取得了令人瞩目的突破性进展，作为强化学习的一个重要分支，深度强化学习主要用来做序贯决策，即根据当前的环境状态做出动作选择，并根据动作的反馈不断地调整自身策略，从而达到设定的目标。除了军事领域，深度强化学习在工业上也有广泛的应用。

1.文献综述

李等在文献[1]中提出，大多数在工业中需要解决的问题可以概括为一种组合优化问题（COP），COP 是指一类在离散状态下求极值的最优化问题，其数学模型如下所示：

$$\begin{aligned} &\min F(x) \\ &s.t.\, G(x) \geq 0 \\ &x \in D \end{aligned}$$

其中x为决策变量，$F(x)$为目标函数，$G(x)$为约束条件，D表示离散的决策空间，为有限个点组成的集合。组合优化问题在交通、产品制造、管理决策、电力、通信等都有广泛应用。常见问题包括旅行商问题、车间作业调度问题等；其决策空间为有限值，直观上可以通过穷举法得到问题的最优解，但由于可行解数量随问题规模呈指数上升，无法在时间内穷举得到最优解。

1）交通领域

喻[2]等提出用深度强化学习实现一种交通信号控制，文中把车道分成多段（元胞），每段的饱和率和车辆的平均速度进行状态设计，利用 Double DQN 和 CNN 的框架，以控制十字路口信号灯。Wang[3]等设计了 DRLS 和 DDS 两种方法来解决满足交通车辆的卸载需求时边缘卸载中服务延迟和能量消耗的平衡。

2）产品制造领域

在产品制造领域，深度强化学习与启发式算法结合，用于优化汽车涂装线的排序，在降低喷枪颜色切换次数的基础上，缩小与下游总装的车身需求序列差异[4]；与 Transformer，seq2seq 模型结合，则可以用于广义的工业分拣顺序问题（OISS），作者在钢板分拣数据进行了验证[5]；阳[6]等提出在解决二维带形装箱时，以强化学习提供的初始装箱序列改善冷启动；罗等在文献[7]中提出将车间调度看作一个序列到另一个序列的映射，利用 Actor-Critic 的框架对映射进行评价，作者在 Taillard 的 FSP 公共标准数据集和 NISCO 宽厚板卷厂数据集上进行了验证；文献[8]指出联想在自身最大的笔记本电脑制造工厂 LCFC 中用基于深度强化学习的决策平台替换了人工生产调度平台，在减少积压生产订单的同时提

高完成率，并且反应到 LCFC 的收入当中；Dai[9]等在研究多速率分层运行最优控制（OOC）时将基于模型的基本回路层预测控制与基于数据的操作层强化学习相结合，克服操作过程动态模型建立的困难，并以重介质分离过程为例在 METSIM 仿真平台中验证了方法的有效性；Dogru[10]等将改进的强化学习作为滤波器提高工艺操作效率，应用于一个类似于工业分离容器的中试规模分离过程；袁[11]等在浓密机的底流浓度控制使用 Critic 网络，实现了对浓密机底流浓度的稳定控制；唐[12]等在电液伺服系统（SRL）中使用 Soft Actor-Critic 结构，使得系统能实现稳态安全控制；谭[13]等在慢走丝张力系统中将强化学习与 PID 控制结合，实现电极丝的精确控制。

特别地，在钢铁石油重工业中，司[14]等利用基于 RBF 的 Actor-Critic 网络成功降低了连退炉内带钢跑偏量；蒋[15]利用 DQN 预测锅炉燃烧场功率并优化其效能；Han[16]等利用强化学习方法自适应搜索文中提出的 HGrC 模型的结构，对钢铁复产气体流动趋势进行预测；Dogru[17]等采用双深度 Q 网络进行参数自动调整训练，实现了试井解释的油气藏参数反演。

3）管理领域

在管理领域，Paraschos[18]等在解决循环制造的管理时使用强化学习用于批准生产、维修、回收和再制造活动，提高在若干阶段涉及制造设施的日益恶化的循环制造系统的成本效益和复原力；Bowes[19]等提出一种基于 DDPG 的雨水系统的协调释放和污染物处理；Hu[20]等提出一种基于 MORL 和 PPO 的配水网络应急调度系统。

4）电力通讯与云资源调度领域

在电力通讯和云资源调度领域，刘[21]等提出 CADP-DRL 模型，能够在实现最小卸载冲突的情况下为高优先级工业设备提供最高的成功卸载效率保证；何[22]构建 FEL-DRL 用于大数据分析得到的热点内容调度给用户；林[23]通过 DQN 优化多资源云作业调度策略，优化目标为平衡最小作业平均怠工时间和平均完成时间两个指标；黄[24]使用 DQN 规划路由路径，有效延长网络生存周期，并实现均衡负载；梁[25]等提出移动边缘计算中的任务卸载是一类 PSPACE-Hard 问题，并分别从基于价值和基于策略的强化学习方法给出案例分析；邓[26]等利用深度强化学习自我博弈，在云仿真平台 Cloud Sim 上成功优化了基线 DVFS；唐[27]等针对含光伏、全钒液流电池储能装置和多类型柔性负荷的工业园区主动配电系统，提出一种基于模拟退火的有限时段 Q 学习，提高系统经济运行效益，并在一定程度上满足电网调峰需求；邵[28]利用深度强化学习对频谱资源进行动态分配，实现了更高的信道利用率，并且多用户之间选择信道时碰撞率更低；张[29]通过 LSTM 对供热预测进行建模，并且使用 DDPG 实现了集中供热系统的热量分配优化；Liu[30]等为了解决

新型区块链 IIOT 系统性能优化问题，提出用深度强化学习选择/调整区块链中的块生产者、共识算法、块大小和块间隔；Ge[31]等引入基于强化学习和 PSO-LSSVM 的智能负荷预测法，根据不同行业和地区属性对工业企业进行分类并预测未来的负荷值；Geng[32]等利用深度强化学习和 BC 技术对 CA 区块链共识模型进行优化，为深度强化学习在物联网领域的应用提供了新的前景；Islam[33]等利用深度强化学习在云部署 Spark 上优化了云资源调度；Gong[34]等针对实时自适应协作的无线通信和计算深度融合的单边缘场景，提出一种基于深度强化学习的任务调度框架，提高收敛性能，降低系统开销。

特别地，为应对工业物联网异常入侵的场景，陈[35]等针对新生的攻击，提出一种基于 DQN 和蚁群优化的检测方式，并在天然气管道测试平台 SCADA 系统上进行了验证；李[36]等提出 DRL-IDS 模型，利用 PPO2 算法和基于 LightBGM 的特征选择算法进行分类，在美国能源部橡树岭国家实验室公开发布的工业物联网真实数据集上获得 99.09%的准确率。

除组合优化问题外，强化学习仍然在一些工业场景发挥作用。

5）路径规划领域

在路径规划领域，李[37]提出将基于 BP 网络的 DQN 模型搭载在工控机器人 Bobac 上并使用重塑奖励引导，成功减少了避障路径时间和长度，提高规划成功率。陈[38]等利用 RGB-D 图像和红外图像作为输入，通过基于全卷积网络的深度强化学习实现了密集物体温度优先推抓（TPG）方法；杜[39]等提出一种基于模糊 Sarsa(λ) 学习的变导纳控制模型，实现柔顺自然的机械臂摆位操作，满足力交互过程中各阶段的阻尼变化需求，实现微创外科手术机器人的手术姿态调整；张[40]等在二维空间中离散移动机器人周围的障碍物信息和目标方向，采取连续奖励，导航实验在 Gazebo 和实际机器人上都获得较好效果；陈[41]在基于 V-REP 平台建立 KUKA KR6 R900 机器人和仓储转台的物理模型并采用 Newton 引擎，选定 DDPG 用于二连杆机器人定点运动控制， 证明有约束条件的强化学习表现得更好；苏[42]使用 TRPO 的 Hessian-Free 优化方法训练从而实现策略的安全搜索；陈[43]使用 Soft Actor-Critic 对充电枪的视觉姿态偏差进行优化和充电孔插孔；刘[44]将视觉和强化学习结合，利用无导数优化方法 CEM 代替 TD3 中的确定性策略用以解决环境奖励系数的问题，使其更适合用于物体抓取工作，并在 bullet 物理引擎上获得较好的效果；柳[45]首先在但机械手上改进 Soft Actor-Critic 的回放池结构，使用最大奖励优先采样优化路径，再对双机械手采取 CTDE 的训练方式，以 GRU 网络为编解码方式最终实现双机械手协作路径规划；陈[46]使用双重策略学习框架，同时拥有两个独立动作策略，线性结合后再与环境交互；郭[47]设计基于 Dueling DDQN-PER 的无人车规划方法，提高了无人车在复杂环境中进行路径规

划的速度；张[48]提出了一种基于 DQN 的五自由度抛光机器人，验证其动力学可行性；吴[49]提出了基于 CNN 的 DQN，利用经验回放池和 Adam 优化器实现了狭窄通道的路径规划；黄[50]在 TurtleBot3 中部署分层强化学习架构 H-DQN，并在其中加入 LSTM 网络，实现基于激光雷达的导航。

6）目标检测及跟踪领域

在目标检测领域，Tan[51]等在 PASCALVoc2012 数据集上做对象检测，动作空间选为 Boundbox 的大小或位置变换，而和标注的 IoU 值则作为奖励。在目标跟踪领域，ADNet[52]将目标移动定义为离散化的动作，特征以及观察的历史状态形成当前状态，认为目标跟踪是一系列动作预测和状态变化的过程。在学习过程中采用深度神经网络并使用基于矩形框交并比的奖惩机制。在训练阶段，分为监督学习和强化学习两个阶段。在监督学习阶段利用视频序列优化目标位移及尺度变化等动作；在强化学习阶段利用监督学习阶段训练的网络作为初始化，然后采取包含采样状态、动作、激励在内的训练序列进行跟踪仿真。骆[53]将强化学习引入加权多尺度网络定位和对齐的视频对象分割方法，在高精度视频对象分割网络和高速视频对象分割网络之间为视频的每一帧选择最佳分割网络，有效地提高了分割速度。

7）自然语言处理领域

在自然语言处理领域，曹[54]利用基于策略的强化学习模型评价生成的回复，避免无意义的回答和僵局，并试图产生新的信息。

2.强化学习解决问题特点分析

强化学习最明显的特征[55]是：1）观察（Observation）值不是独立同分布，而是具有前后的时序关系，并且动作会以特定规律反过来影响状态；2）智能体不能像有监督一样得到及时反馈（即延迟奖励）。

在确定一个问题准备用强化学习解决时，需要确认以下因素[56]：

1）问题需要满足马尔可夫决策过程或者部分可观测马尔可夫决策过程的数学定义；

2）在规则，传统强化学习，深度强化学习中选择方案

①当问题的状态和动作空间维度不高，或者由于附加限制使得实际解空间较小，使用 if-else 规则或启发式搜索可能会更好；

②而在状态-动作组合人的期望累计回报分布无显著规律并呈现多模态特征，解空间可穷举，规模适中的条件下，Q-Learning 和 Sarsa 等传统强化学习算法在调参难度和训练稳定性往往优于深度强化学习算法，这是因为在规模可控的任务解空间中，神经网络尚不足以表现出相对于表格的优势；

③难以从庞大的解空间中分析出有效规则和启发式搜索方案，或者解空间中可能存在比规则和启发式搜索更好的方案；解空间维度过高或不可穷举；类似于二维图像和长跨度时序信息等高位状态信息中包含大量冗余信息，需要深度神经网络的特征提取能力，在这些状况下深度强化学习可能才是最优选择。

3）目标任务需要满足场景固定和数据廉价，对于前者，具体表现为环境状态转移概率分布一致，而在这方面往往面临模拟器训练时带来的 Reality Gap（现实鸿沟）；对于后者，深度强化学习具有低样本效率的缺陷，即需要大量数据进行训练，因此高速度、高质量的训练环境是深度强化学习实用化的关键。

特别地，对于涉及硬件的应用（如机器人），很难同时满足上述条件，若采用实体硬件直接采用，虽然可以获得理想质量的数据，但最高采样速率不得不受限于实际物理条件；多设备并行采样虽然能够在一定程度上进行弥补，但是需要高昂的前期投入，因此在机器人领域只有 Google 等少数大公司采用这一策略。此外，硬件设备在随机探索过程中可能会对环境和自身造成潜在危害，并且需要频繁的人工干扰，这些都会引入额外的成本。

4）需要具体界定深度强化学习在总任务中的子任务，盲目使用深度强化学习算法完成总任务可能使得每个子任务得不到最优解。

参考文献

[1] 凯文,张涛,王锐,覃伟健,贺惠晖,黄鸿. 基于深度强化学习的组合优化研究进展[J].自动化学报，2021，47(11):2521-2537

[2] 喻金忠,曹进德. 深度强化学习在交通控制中的应用[J].工业控制计算机， 2019,32(6):88-89,92.

[3] Wang Y, Wang K, Huang H, et al. Traffic and computation co-offloading with reinforcement learning in fog computing for industrial applications[J]. IEEE Transactions on Industrial Informatics, 2018, 15(2): 976-986.

[4] 胡畔. 基于强化学习的汽车涂装线作业优化排序研究[D]. 大连理工大学,2019.

[5] 曾德天,曾增日,詹俊. 基于深度强化学习种群优化的演化式分拣调度算法[J]. 计算机应用研究,2022,39(3):739-743,757..

[6] 阳名钢,陈梦烦,杨双远,等. 求解二维装箱问题的强化学习启发式算法[J]. 软件学报,2021,32(12):3684-3697.

[7] 罗梓珲,江呈羚,刘亮,等. 基于深度强化学习的智能车间调度方法研究[J]. 物联网学报,2022,6(1):53-64.

[8] Liang Y, Sun Z, Song T, et al. Lenovo Schedules Laptop Manufacturing Using Deep Reinforcement Learning[J]. INFORMS Journal on Applied Analytics, 2022, 52(1): 56-68.

[9] Dai W, Li T, Zhang L, et al. Multi-Rate Layered Operational Optimal Control for Large-Scale Industrial Processes[J]. IEEE Transactions on Industrial Informatics, 2021, 18(7): 4749-4761.

[10] Dogru O, Chiplunkar R, Huang B. Reinforcement learning with constrained uncertain reward function through particle filtering[J]. IEEE Transactions on Industrial Electronics, 2021, 69(7): 7491-7499.

[11] 袁兆麟,何润姿,姚超,等. 基于强化学习的浓密机底流浓度在线控制算法[J]. 自动化学报,2021,47(7):1558-1571.

[12] 唐逸凡,余臻,刘利军. 一种电液伺服系统安全强化学习控制方法[J]. 厦门大学学报（自然科学版）,2022,61(2):239-245.

[13] 谭行,蒋健,魏德骄. 深度强化学习算法在慢走丝机床上的应用研究[J]. 自动化与仪表,2019,34(4):60-64.

[14] 司华春,李小强,王一帆,等. 强化学习在连退炉内带钢跑偏控制中的应用[J]. 冶金自动化,2021,45(4):34-39.

[15] 蒋慧. 基于改进深度强化学习的工业锅炉燃烧场能效优化[D]. 南京邮电大学,2021.

[16] Han Z, Pedrycz W, Zhao J, et al. Hierarchical Granular Computing-Based Model and Its Reinforcement Structural Learning for Construction of Long-Term Prediction Intervals[J]. IEEE Transactions on Cybernetics, 2020, PP(99):1-11.
[17] Dong P, Chen Z M, Liao X W, et al. A deep reinforcement learning (DRL) based approach for well-testing interpretation to evaluate reservoir parameters[J]. Petroleum Science, 2022, 19(1): 264-278.
[18] Paraschos P D, Xanthopoulos A S, Koulinas G K, et al. Machine learning integrated design and operation management for resilient circular manufacturing systems[J]. Computers & Industrial Engineering, 2022, 167: 107971.
[19] Bowes B D, Wang C, Ercan M B, et al. Reinforcement learning-based real-time control of coastal urban stormwater systems to mitigate flooding and improve water quality[J]. Environmental Science: Water Research & Technology, 2022.
[20] Hu C, Wang Q, Gong W, et al. Multi-objective deep reinforcement learning for emergency scheduling in a water distribution network[J]. Memetic Computing, 2022, 14(2): 211-223.
[21] 刘晓宇,许驰,曾鹏,等. 面向异构工业任务高并发计算卸载的深度强化学习算法[J]. 计算机学报,2021,44(12):2367-2381. DOI:10.11897/SP.J.1016.2021.02367.
[22] 何小明. 基于深度强化学习的数据中心内容分发[D]. 南京邮电大学,2019.
[23] 林建鹏. 基于深度强化学习的云资源调度研究[D]. 广东工业大学,2019.
[24] 黄林. 基于深度强化学习的无线传感器网络调度与路由优化[D]. 华中科技大学,2019.
[25] 梁俊斌,张海涵,蒋婵,等. 移动边缘计算中基于深度强化学习的任务卸载研究进展[J]. 计算机科学,2021,48(7):316-323.
[26] 邓志龙,张琦玮,曹皓,等. 一种基于深度强化学习的调度优化方法[J]. 西北工业大学学报,2017,35(6):1047-1053.
[27] 唐昊,刘畅,杨明,等. 考虑电网调峰需求的工业园区主动配电系统调度学习优化[J]. 自动化学报,2021,47(10):2449-2463.
[28] 邵瑞宇. 基于强化学习多用户频谱分配研究[D]. 2021.
[29] 张腾达. 基于强化学习的热力站优化控制研究[D]. 2021.
[30] Liu M, Yu F R, Teng Y, et al. Performance optimization for blockchain-enabled industrial Internet of Things (IIoT) systems: A deep reinforcement learning approach[J]. IEEE Transactions on Industrial Informatics, 2019, 15(6): 3559-3570.
[31] Ge Q, Guo C, Jiang H, et al. Industrial power load forecasting method based on reinforcement learning and PSO-LSSVM[J]. IEEE transactions on cybernetics,

2020.
[32] Geng T, Du Y. Applying the blockchain-based deep reinforcement consensus algorithm to the intelligent manufacturing model under internet of things[J]. The Journal of Supercomputing, 2022: 1-23.
[33] Islam M T, Karunasekera S, Buyya R. Performance and Cost-Efficient Spark Job Scheduling Based on Deep Reinforcement Learning in Cloud Computing Environments[J]. IEEE Transactions on Parallel and Distributed Systems, 2021, 33(7): 1695-1710.
[34] Gong Y, Yao H, Wang J, et al. Multi-Agent Driven Resource Allocation and Interference Management for Deep Edge Networks[J]. IEEE Transactions on Vehicular Technology, 2021, 71(2): 2018-2030.
[35] 陈铁明,董航. 使用蚁群算法和深度强化学习的工业异常入侵检测[J]. 小型微型计算机系统,2022,43(4):779-784.
[36] 李贝贝,宋佳芮,杜卿芸,等. DRL-IDS:基于深度强化学习的工业物联网入侵检测系统[J]. 计算机科学,2021,48(7):47-54.
[37] 李文彪. 基于深度强化学习的工业机器人避障路径规划方法[J]. 制造业自动化,2022,44(1):127-130.
[38] 陈满,李茂军,胡建文,等. 基于深度强化学习的密集物体温度优先推抓方法[J]. 传感器与微系统,2022,41(1):41-44,49.
[39] 杜志江,王伟,闫志远,等. 基于模糊强化学习的微创外科手术机械臂人机交互方法[J]. 机器人,2017,39(3):363-370.
[40] 张福海,李宁,袁儒鹏,等. 基于强化学习的机器人路径规划算法[J]. 华中科技大学学报（自然科学版）,2018,46(12):65-70.
[41] 陈荣川. 机器人流水线自动装卸平台虚拟实验设计与强化学习初探[D]. 东华大学,2020.
[42] 苏纯钰. 基于强化学习的工业机械臂运动规划研究[D]. 山东科技大学,2020.
[43] 陈阜. 基于深度强化学习的充电枪装配策略研究[D]. 浙江工业大学,2021.
[44] 刘开宇. 基于强化学习的物体抓取方法研究[D]. 哈尔滨工业大学,2020.
[45] 柳依辰. 基于强化学习的机械手智能路径规划方法的研究[D]. 电子科技大学,2021.
[46] 陈新. 基于强化学习的调度与导航策略研究[D]. 浙江大学,2021.
[47] 郭心德. 基于深度强化学习的 AGV 路径规划[D]. 广东工业大学,2021.
[48] 张活俊,江励,汤健华,等. 基于强化学习的抛光机器人主动力控制研究[J]. 机械工程师,2020(9):31-34.
[49] 吴运雄,曾碧. 基于深度强化学习的移动机器人轨迹跟踪和动态避障[J]. 广东工业大

学学报,2019,36(1):42-50.
[50] 黄锐. 基于深度强化学习的移动机器人导航策略研究[D]. 电子科技大学,2021.
[51] Tan Z, Karaköse M. Optimized Reward Function Based Deep Reinforcement Learning Approach for Object Detection Applications[C]//2022 International Conference on Decision Aid Sciences and Applications (DASA). IEEE, 2022: 1367-1370.
[52] Yun S , Choi J , Yoo Y , et al. Action-Decision Networks for Visual Tracking with Deep Reinforcement Learning[C]// IEEE Conference on Computer Vision & Pattern Recognition. IEEE, 2017
[53] 骆文青. 基于加权多尺度网络和强化学习的视频对象分割[D]. 江苏大学,2021.
[54] 曹东岩. 基于强化学习的开放领域聊天机器人对话生成算法[D]. 哈尔滨工业大学,2017.
[55] 王琦, 杨毅远, 江季. Easy RL：强化学习教程[M], 人民邮电出版社. 2021.
[56] 魏宁. 深度强化学习落地指南[M]. 电子工业出版社. 2021.