Journal of System Simulation

Volume 33 | Issue 12 Article 3

1-13-2022

Brief Review on Applying Reinforcement Learning to Job Shop Scheduling Problems

Xiaohan Wang
1. Beihang University, Beijing 100191, China; ;2. Engineering Research Center of Complex Product Advanced Manufacturing Systems, Ministry of Education, Beijing 100191, China;

Zhang Lin
1. Beihang University, Beijing 100191, China; ;2. Engineering Research Center of Complex Product Advanced Manufacturing Systems, Ministry of Education, Beijing 100191, China;

Ren Lei
1. Beihang University, Beijing 100191, China; ;2. Engineering Research Center of Complex Product Advanced Manufacturing Systems, Ministry of Education, Beijing 100191, China;

Kunyu Xie
1. Beihang University, Beijing 100191, China; ;2. Engineering Research Center of Complex Product Advanced Manufacturing Systems, Ministry of Education, Beijing 100191, China;

See next page for additional authors

Brief Review on Applying Reinforcement Learning to Job Shop Scheduling Problems

Abstract

Abstract: Reinforcement Learning (RL) achieves lower time response and better model generalization in Job Shop Scheduling Problem (JSSP). To explain the current overall research status of JSSP based on RL, summarize the current scheduling framework based on RL, and lay the foundation for follow-up research, the backgrounds of JSSP and RL are introduced. Two simulation techniques commonly used in JSSP are analyzed and two commonly used frameworks for RL to solve JSSP are given. In addition, some existing challenges are pointed out, and related research progress is introduced from three aspects: direct scheduling, feature representation-based scheduling, and parameter search-based scheduling.

Keywords

reinforcement learning application, job shop scheduling problem, graph neural network, combinatorial optimization, deep learning, feature representation

Authors

Xiaohan Wang, Zhang Lin, Ren Lei, Kunyu Xie, Kunyu Wang, Ye Fei, and Chen Zhen

Recommended Citation

Wang Xiaohan, Zhang Lin, Ren Lei, Xie Kunyu, Wang Kunyu, Ye Fei, Chen Zhen. Brief Review on Applying Reinforcement Learning to Job Shop Scheduling Problems[J]. Journal of System Simulation, 2021, 33(12): 2782-2791.

第 33 卷第 12 期 系统仿真学报© Vol. 33 No. 12
2021 年 12 月 Journal of System Simulation Dec. 2021

基于强化学习的车间调度问题研究简述

王霄汉[1,2]，张霖[1,2]，任磊[1,2]，谢堃钰[1,2]，王昆玉[1,2]，叶飞[1,2]，陈真[1,2]
(1. 北京航空航天大学，北京 100191；2. 复杂产品先进制造系统教育部工程研究中心，北京 100191)

摘要：强化学习在车间调度上获得了较低的时间响应和较优的模型泛化性。为阐述基于强化学习的车间调度问题整体研究现状，总结当前基于强化学习的调度框架，同时为后续相关研究奠定基础，介绍了车间调度与强化学习的背景，分析了车间调度问题中常用的 2 种仿真技术，给出了强化学习解决车间调度问题的 2 种常用架构。此外，针对强化学习在车间调度问题上的应用，指出了现存的一些挑战，并对相关研究进展从直接调度、基于特征表示的调度、以及基于参数搜索的调度 3 个方面进行了介绍。

关键词：强化学习应用；车间调度；图神经网络；组合优化；深度学习；特征表示
中图分类号：TP391.9 文献标志码：A 文章编号：1004-731X (2021) 12-2782-10
DOI: 10.16182/j.issn1004731x.joss.21-FZ0774

Brief Review on Applying Reinforcement Learning to Job Shop Scheduling Problems

Wang Xiaohan[1,2], *Zhang Lin*[1,2], *Ren Lei*[1,2], *Xie Kunyu*[1,2], *Wang Kunyu*[1,2], *Ye Fei*[1,2], *Chen Zhen*[1,2]
(1. Beihang University, Beijing 100191, China;
2. Engineering Research Center of Complex Product Advanced Manufacturing Systems, Ministry of Education, Beijing 100191, China)

Abstract: Reinforcement Learning (RL) achieves lower time response and better model generalization in Job Shop Scheduling Problem (JSSP). To explain the current overall research status of JSSP based on RL, summarize the current scheduling framework based on RL, and lay the foundation for follow-up research, the backgrounds of JSSP and RL are introduced. Two simulation techniques commonly used in JSSP are analyzed and two commonly used frameworks for RL to solve JSSP are given. In addition, some existing challenges are pointed out, and related research progress is introduced from three aspects: direct scheduling, feature representation-based scheduling, and parameter search-based scheduling.

Keywords: reinforcement learning application; job shop scheduling problem; graph neural network; combinatorial optimization; deep learning; feature representation

引言

车间调度通过对生产车间中有限资源的合理分配与调度，实现资源的高效利用，提高工件的生产效率，降低生产成本。车间调度模型作为现实社会中生产、制造、物流等领域中实际问题的抽象模型，在各领域都存在广泛的应用价值[1]。提高车间调度算法的准确度、降低算法响应时间、提高模型的泛化性，对降低生产制造成本、提高生产效率十分关键。然而，多数的车间调度问题属于 NP 完全问题，无法在多项式时间内获得全局最优解[2]。目前针对调度问题已经有了很多相关工作，主要是从传统规则式方法与元启发式算法出发，给出车间调度问题的局部最优解。然而，这些传统的优化算法虽然能够使车间调度问题获得一个较高的准确度，但是在时间响应、算法泛化性上却难以达到实际车间调度场景的要求。

收稿日期：2021-05-10 修回日期：2021-07-29
基金项目：国家重点研发计划(2018YFB1701600)；国家自然科学基金(61873014)
第一作者：王霄汉(1998-)，男，博士生，研究方向为离散仿真、多智能体系统和强化学习。E-mail：by1903042@buaa.edu.cn

随着人工智能技术的发展，以机器学习为代表的各类算法在图像识别、自然语言处理、组合优化等领域均取得了较大的成就[3]。强化学习技术作为一种重要的人工智能技术，在机器人控制、游戏竞技等领域应用广泛[4]。由于强化学习模型在训练完后可以重复使用，因此在解决车间调度问题时具有响应时间短、泛化性强的特点；此外，大量将强化学习应用于车间调度问题的研究表明，强化学习也可以获得接近于元启发式算法的准确度[5]。然而，将强化学习应用于车间调度问题依然面临着一些挑战：

(1) 车间调度系统状态难以定义：状态作为强化学习中的重要组成元素，直接影响整个算法的运行效率和准确度。然而在车间调度问题上，状态的定义灵活性较高，不合适的状态表示方法会导致算法难以训练和使用。

(2) 动作与奖励函数需要严格对应：强化学习中如何设计调度动作，并且如何根据调度动作制定合理的奖励函数，对模型收敛性和准确度的提高十分关键。

(3) 不同强化学习算法难以选择：由于不同强化学习算法之间差异较大，使得不同算法应用在车间调度问题上时准确度不同。针对不同的车间调度问题，如何选择合适的强化学习算法十分重要。

针对强化学习在车间调度问题上的优势和挑战，本文首先介绍了车间调度问题和强化学习的背景，接着总结了基于强化学习解决车间调度问题的 2 种主要建模框架，最后对近些年利用强化学习解决车间调度问题的相关工作进行了分类简述。

1 背景

1.1 车间调度问题

作为一类组合优化问题，车间调度模型是很多实际资源分配和调度问题的抽象模型，代表了有限资源下对加工工件的合理调度过程。

1.1.1 车间调度问题建模

车间调度模型主要包含以下元素：

(1) n 个待加工工件；

(2) 每个工件 i 具有 m 个工序，每道工序的加工顺序需要遵循一定的规则顺序，其中 $i\in\{1,2,\cdots,n\}$；

(3) m 个加工机器，每个加工机器单独负责一道工序，并且每台机器的加工时间一般不同，同一加工机器同一时刻只能加工一个工件。

其中，车间调度问题求解的目标是确定每台机器的工件加工顺序和每个工序的具体开工时间，以使得所有工件的总加工完成时间最短。其中，整个优化的目标函数 L 可以表示为

$$L=\min(\max_{1<i<n} C_i)$$

式中：C_i 为工件 i 从开始加工到最后加工完毕所用的全部时间，包括了加工时间和等待时间。实际中完全符合车间调度模型的案例并不多，但是都是基于此基础模型进行的演变，表 1 列举了几种常见的车间调度问题。目前，针对这些不同的车间调度问题已经有了很多研究。

表 1 常见的几种车间调度问题

Tab. 1 Several common Job shop scheduling problems (JSSP)

调度问题类型	描述
车间(Job-Shop, JS)	如 1.1.1 建模描述的车间调度过程
流水车间(Flow-Shop, FS)	车间调度问题的特殊情形，它每个工件都有相同的加工路线
柔性车间(Flexible Job-Shop, FJS)	车间调度问题的扩展，它允许工件在给定的几台功能相同的机器上加工
动态车间(Dynamic Job-Shop, DJS)	车间调度问题的扩展，在调度过程中可能会产生随机突发事件，从而影响调度流程

DOI: 10.16182/j.issn1004731x.joss.21-FZ0774

1.1.2 仿真技术在车间调度问题中的应用

仿真是解决车间调度问题的基础，任何调度优化算法，无论是传统的规则式方法、启发式优化算法，还是如今的人工智能方法、深度强化学习技术，都需要建立在仿真的基础上[6]。针对各类车间调度问题，一个关键的问题在于如何搭建高效、可信的仿真环境，以有效地模拟现实工厂中的加工制造环境，为调度优化算法提供验证和测试基础。目前，搭建车间调度仿真环境主要有两大类方法，分别是基于元胞自动机的方法，以及基于离散事件仿真的方法。

在基于元胞自动机搭建车间调度仿真环境方面，主要是将机器、加工资源等制造车间元素建模为网络空间的固定格点，将加工的工件作为移动粒子，从而实现对实际加工场景的简化和复现。陈勇等[7]基于元胞自动机对大型零件的生产过程进行了建模，解决了大型零件的车间动态柔性调度问题，并通过与实际方案的对比证明了元胞自动机仿真环境的有效性。郑忠等[8]基于元胞自动机搭建了车间天车调度仿真模型，根据车间生产的特点设定工位、天车和物料的属性和参数，并通过实验证明了基于元胞自动机方法搭建天车调度环境的有效性。孟寅茂[9]基于元胞自动机方法搭建了船舶企业分段车间调度环境，在三维空间上对生产环境进行了建模，并在计划完成时间、时空利用率和制作任务延迟数 3 个指标上验证了方法的有效性和可信性。

在基于离散事件仿真方法搭建车间调度仿真环境方面，主要是将车间调度过程中生产环境的变化作为事件转移触发条件，并以此驱动整个工件加工的动态过程。张晴等[10]指出基于离散事件搭建车间调度环境具有简洁高效的特点，结合人机交互方法可以取得有效的结果。曲丹[11]基于离散事件系统仿真的方法，总结了车间仿真调度的原理及方法，并基于多智能体系统进行调度求解，实验结果表明了基于离散事件搭建调度仿真环境的有效性。徐修文等[12]针对离散制造车间中事件对车间调度算法的影响进行了评估，并提出了一种基于动态事件描述的事件推进影响评估模型。目前，基于离散事件仿真方法搭建车间调度环境成为了主流方法，既降低了计算机的运算复杂度，同时也易于编程或建模实现。

1.1.3 车间调度问题的传统解决算法

车间调度问题由于其存在的普遍性，一直以来都是组合优化领域研究的重点。传统的车间调度问题主要有 2 类解决算法，分别是规则式方法以及元启发式算法。

规则式方法是指基于简单规则安排工件的调度顺序，又称优先调度规则(Priority Dispatch Rules, PDR)，表 2 所示为几种常用的规则式方法[13]。规则式方法虽然简单，但是以其超低的时间响应和对不同调度问题的较强泛化性，依然在某些调度场景中被广泛应用。此外，某些规则式方法在一些特定的调度问题上可以获得较高的准确度。

元启发式算法是解决车间调度问题最常用的优化算法，通过不同的优化迭代算子在车间调度问题上搜索得到局部最优解[14]。元启发式算法在调度问题上可以获得高于规则式方法的准确度，目前在各类车间调度问题上应用广泛，表 3 所示为常用于车间调度问题的元启发式算法。然而，元启发式算法有 2 个主要的劣势。首先，由于优化算法计算量较大，且无法通过预训练模型的方式进行参数化存储，使得每次优化都需要从头开始，造成时间响应较长。此外，元启发式算法泛化性较差，对于不同的调度问题往往需要不同的参数调整，难以实现算法的直接迁移。

表 2 常用的规则式车间调度方法
Tab. 2 Common rule-based JSSP methods

方法名称	含义
先进先出(First In First Out, FIFO)	优先处理第一个作业
后进先出(Last In First Out, LIFO)	优先处理最后一个作业
最短处理时间(Shortest Processing Time, SPT)	优先处理具有最短处理时间的作业
最长处理时间(Longest Processing Time, LPT)	优先处理具有最长处理时间的作业
最短总处理时间(Shortest Total Processing Time, STPT)	优先处理具有最短总处理时间的作业
最长总处理时间(Longest Total Processing Time, LTPT)	优先处理具有最长总处理时间的作业
剩余最少操作数(Least Operation Remaining, LOR)	优先处理当前最小剩余操作数的作业
剩余大部分操作(Most Operation Remaining, MOR)	优先处理当前最大剩余操作数的作业
下一任务最小等待操作(Least Queue Next Operation, LQNO)	优先处理下一个操作等待最少的作业

表 3 常用的元启发式车间调度算法
Tab. 3 Common meta-heuristic algorithms

算法名称	算法机制
遗传算法(Genetic Algorithms, GA)[15]	模拟生物优胜劣汰
禁忌搜索(Tabu Search, TS)[16]	模拟人的记忆功能
模拟退火(Simulated Annealing, SA)[17]	模拟热力学退火过程
蚁群算法(Ant Colony Optimization, ACO)[18]	模拟蚂蚁觅食行为
粒子群算法(Particle Swarm Optimization, PSO)[19]	模拟鸟群觅食行为
人工免疫算法(Artificial Immune, AI)[20]	模拟生物免疫系统

1.2 强化学习

1.2.1 基本概念与定义

强化学习是智能体完成学习过程的重要方法，并随着人工智能的发展被应用于自动驾驶、股票机器人、游戏对抗等领域[4]。强化学习通过智能体与环境的交互过程自学习最优策略，以最大化智能体的回报值或者完成某个固定的目标[21]。在强化学习中，智能体以马尔可夫决策过程(Markov Decision Process, MDP)进行描述，可以被描述为一个元组：$\langle S,A,p,P,R\rangle$，其中 S 代表动作空间，A 代表动作空间，$\pi: S\to A$ 代表智能体的策略，$P: S\times A\to P$ 为状态转移概率，$R: S\times A\times S\to R$ 为奖励函数[22]。强化学习一般通过优化 2 个值函数实现适应性学习的过程，分别是值函数 $V(s)\leftarrow E\{G_t \mid S_t=s\}$ 和动作状态值函数 $Q(s,a)\leftarrow E\{G_t \mid S_t=s,A_t=a\}$，其中 $G_t=\sum_{k=0}^{\infty}\gamma^k R_{t+k+1}$ 为用于估计未来累计奖励值的回报函数。

强化学习主要利用时序差分法、蒙特卡罗法、以及动态规划法 3 种算法迭代更新其值函数或策略[23]。时序差分法是一种单步更新的值函数迭代算法，用于控制值函数收敛到最优解，表示为[22]

$$V(s)\leftarrow V(s)+\alpha\left(R_{t+1}+\gamma V(s')-V(s)\right)$$

式中：α 为学习率；γ 为衰减因子；R_{t+1} 为状态从 $s\to s'$ 的奖励值。基于时序差分法更新智能体的动作状态值函数，并在每一步采用选取最大值函数的方法估计智能体的最优策略，在每一次智能体与环境进行交互之后，采用的更新过程为

$$Q(s,a)\leftarrow Q(s,a)+\alpha\left\{r+\gamma\cdot\max_{a'}Q(s',a')-Q(s,a)\right\}$$

式中：a'为下一步要采取的动作。整个时序差分法处理强化学习过程中智能体与环境进行交互并学习到适应性策略的过程，虽然不同强化学习算法之间差异性较大，但是智能体与环境的事件转移过程基本一致。

1.2.2 深度强化学习

深度强化学习是一类主流的强化学习方法，指

DOI: 10.16182/j.issn1004731x.joss.21-FZ0774

第 33 卷第 12 期 系统仿真学报 Vol. 33 No. 12
2021 年 12 月 Journal of System Simulation Dec. 2021

的是融合了深度学习技术的强化学习方法。随着深度学习技术的广泛应用，目前绝大多数的强化学习算法都在使用中融合了深度学习方法。具体而言，深度强化学习主要使用神经网络拟合 2 个变量，分别是值函数与策略，以此形成的 2 个主要的深度强化学习分支，即基于值与基于策略的算法。基于值的算法适用于智能体离散动作条件下，而基于策略的算法主要用于处理智能体的连续动作。目前，主流的深度强化学习算法如表 4 所示。

目前，深度强化学习领域还处于高速发展阶段，各种不同类型的算法层出不穷，表 4 所示的算法涵盖了大多数应用于车间调度问题的主流深度强化学习算法[21]。

表 4 常用的深度强化学习算法
Tab. 4 Common deep reinforcement learning algorithms

算法名称	算法类别
Deep-Q-Learning (DQN)	基于值
Dueling DQN	基于值
Double DQN	基于值
REINFORCE	基于策略
Proximal Policy Optimization (PPO)	基于策略
Actor-Critic (AC)	基于策略
Deep Deterministic Policy Gradient (DDPG)	基于策略
Soft Actor-Critic (SAC)	基于策略

2 两种主要的强化学习调度架构

将强化学习应用于车间调度问题，主要是利用智能体的决策能力，在每次调度环境发生状态转移时为所有工件安排合理的加工顺序。按照对于智能体的定义方式来分，本文将强化学习调度架构分为 2 种，分别是单智能体架构与多智能体架构。

2.1 单智能体架构

强化学习调度算法的单智能体架构将所有待加工工件的整体看作一个智能体，如图 1 所示。智能体以所有工件的当前状态集合以及环境的全局状态作为其状态输入，以所有工件下一步的调度动作集合作为其动作输出。

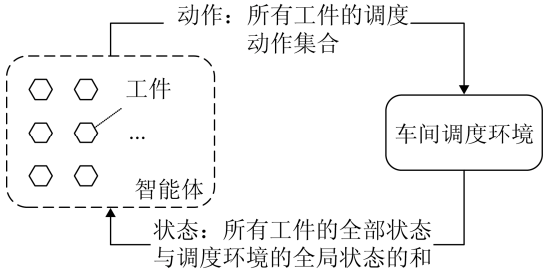

图 1 强化学习调度算法的单智能体架构
Fig. 1 Single agent architecture of RL scheduling algorithms

单智能体架构下，由于强化学习算法控制的是单一智能体，因此收敛性较好，并且整个强化学习算法和相应的调度环境也易于搭建。然而，由于单智能体架构没有考虑工件智能体之间的通讯以及合作过程，可能使得调度信息有所损失。此外，受限于神经网络的收敛性要求，单智能体架构下不同工件的调度动作维数需要保持在二维左右，使其难以描述具有复杂调度动作的车间调度问题。

2.2 多智能体架构

强化学习调度算法的多智能体架构将每一个工件看作一个单独的智能体，多个工件组成的多智能体系统与调度环境进行交互，如图 2 所示。每个工件智能体以当前自身的状态及相对应的环境状态为状态输入，以下一步的各自动作为输出。

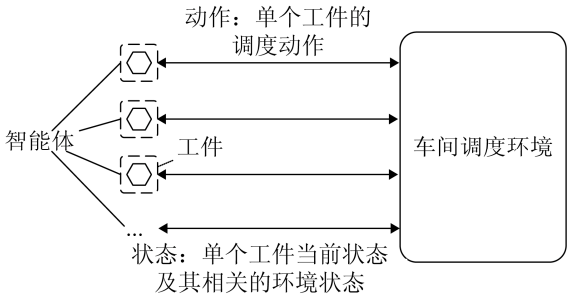

图 2 强化学习调度算法的多智能体架构
Fig. 2 Multi-agent architecture of RL scheduling algorithms

多智能体架构下，多个智能体可以同时训练，并使用同一组模型作为策略，但是由于智能体的状态与动作的分布差异较大，使得模型收敛难度上升。此外，由于每个工件智能体都可能引起车间调度环境的转移，因此需要处理好不同工件智能体带

来的资源冲突问题，从而使得多智能体架构下车间调度环境编写困难。然而，多智能体架构可以充分表示不同工件智能体的合作关系，可以更加细致地对不同调度问题进行建模，以解决更复杂的车间调度模型。

3 强化学习车间调度算法简述

目前，强化学习在各类车间调度问题上有着广泛的应用，根据其算法特点，可以将其分为 3 个主要的类别，分别是直接调度、基于特征表示的调度、以及基于参数搜索的调度。

3.1 直接调度

直接调度指的是直接利用调度环境产生的实时输出作为状态，如当前待加工工件数、工件排队时长等，并将由强化学习算法产生的智能体的下一步动作作为下一步调度的顺序，从而实现深度强化学习算法与调度环境的直接耦合。直接调度法是利用强化学习实现车间调度最简单、有效的方法，被国内外研究者广泛使用。文献[24]以解决生产车间中 AGV(Automated Guided Vehicle)小车的调运问题为目标，使用 Q-learning 算法解决了多个 AGV 小车的车间调度问题。虽然文中使用的算法较为简单，但是作者清晰定义了系统的状态空间、智能体的动作空间、状态转移时刻、奖励函数、Q 学习函数、动作选择策略等细节内容。此外，文献[24]指出基于强化学习解决调度问题主要需要解决 2 个问题，分别是如何将调度问题转化为强化学习问题，以及如何保证强化学习能够学习到调度问题合适的解。文献[25]利用深度强化学习算法 Deep-Q-Learning (DQN)解决了带有随机任务插入的动态柔性车间调度问题，并同样说明了系统的状态空间等一系列特征的定义过程。文献[25]指出车间动态调度问题已经被广泛研究，过去的几十年提出的各种方法主要包括调度规则和元启发法，但是规则式方法无法获得一个满意的解，而元启发式算法又十分耗时，因此强化学习的出现为车间动态柔性调度提供了新的解决方案。文献[26]分别利用 DQN 与 SAC 算法解决了离散和连续的生产计划和控制问题，其中生产计划和控制问题属于动态车间调度的一种。文献[26]指出，传统的启发式算法除了速度慢之外，其泛化性差也是一个重要的问题，使其难以适应生产计划和控制动态变化的环境。因此，文献[26]认为基于强化学习的方法无论是在时间上还是其泛化性上都很适合生产计划和控制问题。文献[27]基于 DQN 模型解决了舰面甲板环境下舰载机的保障作业实时调度问题，并给出了系统的状态空间等一系列特征的定义方法，并与遗传算法、模拟退火、线性规划这 3 种方法在时间和性能上进行了对比。李指出基于强化学习的调度方法权衡了算法在时间和性能上的表现，保障了实时调度过程中作业调度序列的质量。

直接调度法是将强化学习应用于调度问题上最简洁、高效的方法，而在其应用中往往需要面对以下 2 个挑战：

(1) 状态的定义方式需要针对不同调度问题进行较大的调整，以表征不同状态间的区别；

(2) 状态难以充分表示当前调度环境的情况，对于调度环境中的一些隐性关系无法充分提取。

这 2 个挑战使得在某些调度问题上，强化学习算法无法学习到调度环境的真实状态与奖励函数的对应关系，从而造成算法准确度下降、训练收敛性差的问题。

3.2 基于特征表示的调度

为解决调度环境状态难以表示的问题，一些研究基于深度学习中的图神经网络(Graph Neural Network, GNN)、指针网络、以及注意力机制等方法作为调度环境的特征表示模型，并结合强化学习进行训练，实现基于特征表示的强化学习调度过程。

文献[28]首先使用析取图表示了车间调度的状态，并利用强化学习算法 PPO 同时训练智能体的策略以及析取图组成的图神经网络。文中的方法不

DOI: 10.16182/j.issn1004731x.joss.21-FZ0774

第 33 卷第 12 期 系统仿真学报 Vol. 33 No. 12
2021 年 12 月 Journal of System Simulation Dec. 2021

仅效果优于基于规则的传统算法，并且由于图神经网络的加入使得整个调度算法具有很强的泛化性，甚至可以利用已经训练好的模型解决未经过训练的调度问题。文献[5]指出，调度环境的状态可以通过图神经网络、注意力机制、以及指针网络进行表示，构建传统控制做底层，强化学习做规划，深度学习做感知的 3 层调度模型。文献[1]利用指针网络和注意力机制这 2 种深度学习模型解决了柔性车间调度问题，并使用随机生成的任务验证了提出方法的可行性。文献[29]提出了一种融合分布式智能体学习以及内部智能体交互机制的方法，并使用图神经网络作为特征提取模型，用 PPO 训练智能体的策略与图神经网络，以解决车间调度问题中伸缩性和环境可变性的问题。文章指出图神经网络在表示复杂可变的调度环境时具有很大优势，而诸如遗传算法等集中式优化算法的计算复杂度会随着问题规模的扩大而增加，并且当环境变化时需要重新计算。此外，文章使用了 2 个实验场景作为调度环境，分别是机器人制造单元场景、以及注塑机器场景，验证提出方法的优越性。文献[30]利用图卷积网络(Graph Convolution Network, GCN)构建了异构节点的图模型，并利用 DQN 算法进行训练，将整个图神经网络的全局状态作为 DQN 的状态存储，实现了空箱资源调度模型的搭建。此外，文章指出虽然图卷积网络可以充分提取调度环境的特征，但是整个调度模型的计算效率和训练稳定性会随着图结构的复杂而降低。文献[13]同样使用析取图表示调度问题，并利用 PPO 算法进行训练，通过构建 6 类神经网络模型实现了不同隐性关系的聚合，以充分表征车间调度的状态。此外，文中给出了 3 组不同类型的实验，证明了所提出模型相比于传统规则式算法的优越性和泛化性。

基于特征表示的调度方法既利用了深度学习强大的特征提取能力，也同时受益于强化学习的动态学习能力，使得整个方法相比于传统规则式和启发式算法大幅提升了泛化性，为基于强化学习解决调度问题提供了新的思路。然而，虽然基于特征表示的强化学习调度算法具有较快的时间响应速度与较强的泛化性，但是其依然无法在准确度上超越传统启发式优化算法。

3.3 基于参数搜索的调度

为提高算法的准确率，一些工作将元启发式算法与强化学习相结合，使用强化学习训练出元启发式算法在不同情况下的参数配置，从而提高元启发式算法的泛化性。文献[31]基于 Q-learning 算法学习变邻域搜索算法的参数配置，以解决动态车间调度问题。文献[31]指出，动态变化的环境导致优化算法的参数无法固定，而使用强化学习与变邻域搜索结合的方法提高了整个过程的效率、有效性、泛化性。文献[32]给出了一种基于 Q-learning 的遗传算法调参方法，使得遗传算法在解决调度问题时可以实现自动化调参，以解决遗传算法难以找到最优参数的问题。此外，本文还对比了提出方法相较于直接使用遗传算法在准确度上的提升。

然而，基于强化学习搜索元启发式算法的参数，往往面临 2 个问题：①由于调度算法的结果还是由启发式算法得到，因此依然受限于元启发式算法的非动态、响应时间长的问题；②由于基于参数搜索的调度出发点在于提高启发式算法的参数设置效率，因此使用的强化学习算法本身不会复杂，一般都是 Q-learning 及其相关的算法，很少使用调参较为复杂的深度强化学习算法，从而导致强化学习算法的能力受限。

除了搜索元启发式算法的参数外，文献[33]直接基于强化学习算法选择元启发式算法的种类，以适应不同类型的调度问题。文献[33]基于深度强化学习算法 Actor-Critic 训练了 2 个模型分别作为区域选择策略和规则选择策略，在表达式简化、车间调度、车辆路径规划 3 个场景下进行试验,，分别比较了表达式长度、路径长度、调度总用时等指标来体现所提出方法的优越性。然而，这种方法在解决调度问题时需要较多的算法调参与启发式算法的设置工作，并且需要各种启发式算法在不同调度

问题上的优势作为先验知识，以为强化学习提供训练基础。

3.4 三类强化学习调度算法的对比

目前强化学习调度算法在不同的调度场景应用广泛，其中本文介绍的直接调度、基于特征表示的调度、以及基于参数搜索的调度这 3 类强化学习调度算法的对比分析如表 5 所示。根据不同类型算法的特点，以及具体车间调度场景的需求，选择合适的强化学习调度算法解决车间调度问题。

表 5 强化学习调度算法对比
Tab. 5 Comparison of reinforcement learning scheduling algorithms

调度算法	优势	劣势
直接调度	环境和算法定义简单，高效；计算复杂度低	难以充分提取状态的特征；泛化性差
基于特征表示的调度	特征提取充分；模型泛化性强	计算复杂度高；模型难以收敛
基于参数搜索的调度	准确度高；环境和算法容易定义	泛化性差；时间响应长

4 结论

近年来，强化学习在各领域的应用落地为车间调度问题提供了新的解决思路。本文从车间调度问题和强化学习的背景出发，给出了 2 种强化学习解决车间调度问题的常用建模架构，并对基于各类强化学习算法解决不同场景下的车间调度问题的相关研究进行了分类简述。本文指出，将强化学习应用到车间调度问题上的核心在于如何充分提取调度问题的特征，以及将提取的特征通过合适的方式在强化学习模型中进行表示，而非强化学习算法本身的性能。此外，由于强化学习应用在车间调度问题上目前尚处于初级阶段，关于模型的可解释性、收敛性、重用性等方面也存在着亟待解决的问题。

未来，将从以下 2 个方面深化研究：

(1) 多数场景下的强化学习调度算法都是基于多智能体架构，但是算法却使用的是单智能体算法，从而导致工件智能体之间缺乏信息交流与合作关系，不利于车间调度模型的求解。因此，将多智能体强化学习算法应用到车间调度问题上，以提高模型的表现力。

(2) 车间调度模型一般考虑的都是机器调度层面的问题，然而实际上机器加工层面也可能存在着资源抢占与调度利用的问题，导致整个车间调度问题需要考虑的约束更加复杂。解决多层资源的车间调度问题，能够大幅提升制造车间的资源利用效率。

参考文献：

[1] 潘如媛. 深度强化学习求解作业调度问题方法研究[D]. 北京: 北京交通大学, 2020.
Pan Ruyuan. Research on Deep Reinforcement Learning Methods for Solving Flowshop Scheduling Problem[D]. Beijing: Beijing Jiaotong University, 2020.

[2] Zhang C, Song W, Cao Z, et al. Learning to Dispatch for Job Shop Scheduling via Deep Reinforcement Learning[C]// Advances in Neural Information Processing Systems (NeurIPS). 2020. arXiv preprint arXiv: 2010.12367, 2020.

[3] Gianfrancesco M A, Tamang S, Yazdany J, et al. Potential Biases in Machine Learning Algorithms Using Electronic Health Record Data[J]. JAMA Internal Medicine (S2168-6106), 2018, 178(11): 1544-1547.

[4] Mnih V, Kavukcuoglu K, Silver D, et al. Playing Atari with Deep Reinforcement Learning[J]. Computer Science. 2013. arXiv preprint arXiv: 1312.5602, 2013.

[5] 李凯文, 张涛, 王锐, 等. 基于深度强化学习的组合优化研究进展[J/OL]. 自动化学报. (2020-12-09) [2021-05-31]. https://doi.org/10.16383/j.aas.c200551.
Li Kaiwen, Zhang Tao, Wang Rui, et al. Research Reviews of Combinatorial Optimization Methods Based on Deep Reinforcement Learning[J/OL]. Acta Automatica Sinica. (2020-12-09) [2021-05-31]. https://doi.org/10.16383/j.aas.c200551.

[6] Zhang T, Xie S, Rose O. Real-time Job Shop Scheduling Based on Simulation and Markov Decision Processes[C]// 2017 Winter Simulation Conference (WSC). Las Vegas: IEEE, 2017: 3899-3907.

[7] 陈勇, 阮幸聪, 鲁建厦. 基于元胞自动机的大型零件生产车间动态柔性调度仿真建模[J]. 中国机械工程, 2010, 21(21): 2603-2609.
Chen Yong, Ruan Xingcong, Lu Jianxia. Simulation Modeling of Dynamic & Flexible Scheduling about

第 33 卷第 12 期 系统仿真学报 Vol. 33 No. 12
2021 年 12 月 Journal of System Simulation Dec. 2021

Large-sized Component Production Workshop Based on Cellular Automata[J]. China Mechanical Engineering, 2010, 21(21): 2603-2609.

[8] 郑忠, 徐乐, 高小强. 基于元胞自动机的车间天车调度仿真模型[J]. 系统工程理论与实践, 2008(2): 137-142.
Zheng Zhong, Xu Le, Gao Xiaoqiang. Simulation Model of Crane Scheduling in Workshop Based on Cellular Automata[J]. System Engineering Theory and Practice, 2008(2): 137-142.

[9] 孟寅茂. 基于元胞机的造船企业分段车间空间调度建模与仿真[D]. 杭州: 浙江工业大学, 2012.
Meng Yinmao. Modeling and Simulation of Block Workshop Spatial Scheduling Based on Cellular Automata in Shipbuilding Enterprise[D]. Hangzhou: Zhejiang University of Technology, 2012.

[10] 张晴, 饶运清. 车间动态调度方法研究[J]. 机械制造, 2003, 41(1): 39-41.
Zhang Qing, Rao Yunqing. Research on Dynamic Workshop Scheduling Method[J]. Machinery, 2003, 41(1): 39-41.

[11] 曲丹. 基于多 Agent 的车间调度仿真系统研究[D]. 成都: 西华大学, 2009.
Qu Dan. Research on Simulation System of Job-Shop Scheduling Based on Multi-Agent[D]. Chengdu: Xihua University, 2009.

[12] 徐修文, 邱顺流, 宋豫川, 等. 离散制造车间动态事件影响评估方法[J]. 重庆大学学报(自然科学版), 2012, 35(增 1): 1-5.
Xu Xiuwen, Qiu Shunliu, Song Yuchuan, et al. Impact Assessment Method of Dynamic Events in Discrete Production Workshop[J]. Journal of Chongqing University(Natural Science Edition), 2012, 35(S1): 1-5.

[13] Park J, Chun J, Kim S H, et al. Learning to Schedule Job-shop Problems: Representation and Policy Learning Using Graph Neural Network and Reinforcement Learning[J]. International Journal of Production Research (S0020-7543), 2021, 59(11): 1-18.

[14] 张超勇. 基于自然启发式算法的作业车间调度问题理论与应用研究[D]. 武汉: 华中科技大学, 2007.
Zhang Chaoyong. Research on the Shop Scheduling Problem with Naturally-Inspired Heuristic Algorithms[D]. Wuhan: Huazhong University of Science and Technology, 2007.

[15] Sampson J R. Adaptation in Natural and Artificial Systems (John H. Holland)[J]. Society for Industrial and Applied Mathematics (S0036-1445), 1976, 18(3): 529-530.

[16] Sivaram M, Batri K, Amin Salih M, et al. Exploiting the Local Optima in Genetic Algorithm using Tabu Search[J]. Indian Journal of Science and Technology (S0974-6846), 2019, 12(1): 1-13.

[17] Kirkpatrick S, Gelatt C D, Vecchi M P. Optimization by Simulated Annealing[J]. Science (S0036-8075), 1983, 220(4598): 671-680.

[18] Manosij G, Ritam G, Sarkar R, et al. A Wrapper-filter Feature Selection Technique Based on Ant Colony Optimization[J]. Neural Computing & Applications (S0941-0643), 2020, 32(12): 7839-7857.

[19] Venter G , Jaroslaw S S. Particle Swarm Optimization[J]. AIAA Journal (S0001-1452), 2003, 41(8): 129-132.

[20] Farmer J D, Packard N H, Perelson A S. The Immune System, Adaptation, and Machine Learning[J]. Physica D: Nonlinear Phenomena (S0167-2789), 1986, 22(1/3): 187-204.

[21] Arulkumaran K, Deisenroth M P, Brundage M, et al. Deep Reinforcement Learning: A Brief Survey[J]. IEEE Signal Processing Magazine (S1053-5888), 2017, 34(6): 26-38.

[22] Sutton R S, Barto A G. Reinforcement Learning: An Introduction[M]. Cambridge: MIT Press, 2018.

[23] Watkins C J C H, Dayan P. Q-learning[J]. Machine Learning (S0885-6125), 1992, 8(3/4): 279-292.

[24] Xue T, Zeng P, Yu H. A Reinforcement Learning Method for Multi-AGV Scheduling in Manufacturing[C]// 2018 IEEE International Conference on Industrial Technology (ICIT). Lyon: IEEE, 2018: 1557-1561.

[25] Luo S. Dynamic Scheduling for Flexible Job Shop with New Job Insertions by Deep Reinforcement Learning[J]. Applied Soft Computing (S1568-4946), 2020, 91: 106208.

[26] Samsonov V, Kemmerling M, Paegert M, et al. Manufacturing Control in Job Shop Environments with Reinforcement Learning[C]// 13th International Conference on Agents and Artificial Intelligence. 2021.

[27] 李亚飞, 吴庆顺, 徐明亮, 等. 基于强化学习的舰载机保障作业实时调度方法[J]. 中国科学: 信息科学, 2021, 51(2): 247-262.
Li Yafei, Wu Qingshun, Xu Mingliang, et al. Real-time Scheduling for Carrier-borne Aircraft Support Operations:a Reinforcement Learning Approach[J]. Science China Information Sciences, 2021, 51(2): 247-262.

[28] Zhang C, Song W, Cao Z, et al. Learning to Dispatch for Job Shop Scheduling via Deep Reinforcement Learning[C]// Neural Information Processing Systems (NeurIPS). Vancouver: MIT Press, 2020.

[29] Hameed M, Schwung A. Reinforcement Learning on Job Shop Scheduling Problems Using Graph Networks[J]. arXiv preprint arXiv:2009.03836, 2020.

[30] 熊波. 基于异构图神经网络的多智能体资源调度模型[D]. 北京: 北京交通大学, 2020.
Xiong Bo. Multi-Agent Resource Balancing with Heterogeneous Graph Neural Networks[D]. Beijing: Beijing Jiaotong University, 2020.

[31] Shahrabi J, Adibi M A, Mahootchi M. A Reinforcement Learning Approach to Parameter Estimation in Dynamic Job Shop Scheduling[J]. Computers & Industrial Engineering (S0360-8352), 2017, 110(8): 75-82.

[32] Chen R, Yang B, Li S, et al. A Self-learning Genetic Algorithm Based on Reinforcement Learning for Flexible Job-shop Scheduling Problem[J]. Computers & Industrial Engineering (S0360-8352), 2020, 149(7): 106778.

[33] Chen X, Tian Y. Learning to Perform Local Rewriting for Combinatorial Optimization[J]. arXiv preprint arXiv:1810.00337, 2018.

https://dc-china-simulation.researchcommons.org/journal/vol33/iss12/3
DOI: 10.16182/j.issn1004731x.joss.21-FZ0774