基于深度强化学习的双向装配序列规划

赵铭慧[1], 张雪波[1†], 郭 宪[1], 欧勇盛[2]

(1. 南开大学 人工智能学院机器人与信息自动化研究所 天津市智能机器人技术重点实验室, 天津 300350;

2. 中国科学院深圳先进技术研究院, 广东 深圳 518055)

摘要: 为了解决复杂装配模型的序列规划问题, 并使算法对任意初始状态具有较高的适应性, 本文提出了一种包含正向装配以及逆向拆解的一体化双向装配序列规划方法BASPW–DQN. 针对复杂装配模型, 首先进行了一体化装配序列规划的问题描述与形式化表示; 在此基础上, 引入了课程学习及迁移学习方法, 对包含前向装配和逆向错误零件拆卸两部分过程的双向装配序列规划方法进行研究. 在所搭建的ROS-Gazebo与TensorFlow相结合的仿真平台上进行了验证, 测试结果证明此双向网络对于任意初始状态(包括零装配、部分装配、误装配等初始状态)的装配任务均可以在较少步数内完成, 验证了所提方法对于解决装配序列规划问题的有效性与适应性.

关键词: 智能装配; 装配序列规划; 深度强化学习; Gazebo

引用格式: 赵铭慧, 张雪波, 郭宪, 等. 基于深度强化学习的双向装配序列规划. 控制理论与应用, 2021, 38(12): 1901 – 1910

DOI: 10.7641/CTA.2021.00516

Assembly sequence planning based on deep reinforcement learning

ZHAO Ming-hui[1], ZHANG Xue-bo[1†], GUO Xian[1], OU Yong-sheng[2]

(1. Institute of Robotics and Automatic Information System, College of Artificial Intelligence, Tianjin Key Laboratory of Intelligent Robotics, Nankai University, Tianjin 300350, China;

2. Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen Guangdong 518055, China)

Abstract: In order to solve the sequence planning problem of complex assembly models and improve the flexibility of the algorithm to any initial state, this paper proposes an integrated bi-directional assembly sequence planning method BASPW–DQN. Aiming at the complex assembly model, a bi-directional assembly sequence planning method including forward assembly and wrong part disassembly process is proposed, on this basis, curriculum learning and transfer learning methods are introduced to improve the training efficiency and assembly capabilities of the integrated assembly sequence planning method. And a training platform is developed, which combines the physical simulator Gazebo and deep network framework TensorFlow. The test results show that the bi-directional network can complete the assembly tasks of general assembly in any initial state (such as none-assembly, partial assembly and misassembly) in a few steps demonstrating the effectiveness and flexibility of the proposed method.

Key words: intelligent assembly; assembly sequence planning; deep reinforcement learning; Gazebo

Citation: ZHAO Minghui, ZHANG Xuebo, GUO Xian, et al. Assembly sequence planning based on deep reinforcement learning. *Control Theory & Applications*, 2021, 38(12): 1901 – 1910

1 引言

随着人工智能技术的发展与中国制造2025的提出, 智能制造已经逐渐成为我国近年来的一个研究重点[1]. 伴随着智能化与信息化水平的提高, 工业产品如汽车、船舶和航空航天产品等[2]相较以往更加复杂, 组装技术即装配的重要性也日益凸显.

装配作为工业生产制造过程中的一个关键环节, 对产品的生产效率及质量有着较大影响. 产品组装的时间占总生产时间的20%∼50%, 组装成本约占总成本的20%∼30%[3], 装配过程也被看作是工业产品生

收稿日期: 2020−08−08; 录用日期: 2021−05−18.

†通信作者. E-mail: zhangxuebo@nankai.edu.cn.

本文责任编委: 赵冬斌.

国家自然科学基金项目(U1613210), 天津市杰出青年科学基金项目(19JCJQJC62100), 天津市自然科学基金项目(19JCYBJC18500), 中央高校基本科研业务费项目, 广东省机器人与智能系统重点实验室开放基金项目资助.

Supported in Part by the National Natural Science Foundation of China (U1613210), the Tianjin Science Fund for Distinguished Young Scholars (19JCJQJC62100), the Tianjin Natural Science Foundation (19JCYBJC18500) and the Fundamental Research Funds for the Central Universities and the Opening Project of Guangdong Provincial Key Lab of Robotics and Intelligent System.

产周期中的瓶颈阶段. 因此, 进行装配序列规划的重要性也愈发明显, 一组正确的装配序列是确保产品能够快速组装成功的关键, 进行装配序列规划问题研究具有重要的理论和实际意义.

复杂工业产品装配过程的序列规划是指在满足所有装配条件的前提下, 对装配过程中的零件选择和顺序进行规划. 目的是按顺序找到应安装的装配零件, 并正确移动到安装位置. 如何使复杂装配系统具有智能自主决策能力, 能够模仿人类经验, 并且根据一些较为主观的评价指标(比如零件设计、装配工具、装配体积、优化指标等因素[4]), 得到满足要求的最优的合理装配序列, 是装配序列规划系统的主要研究内容与研究难点.

另一方面, 随着计算能力与资源水平的不断提高, 深度强化学习领域已经取得了很多重大的研究成果. 深度强化学习方法结合了深度神经网络所具有的超强的感知能力以及强化学习方法的决策能力, 优势互补, 能够为复杂系统的感知决策问题提供思路. 现此类算法已经在图像、语音、网络游戏等方面取得了令人惊叹的成果, 但在智能装配问题上还更多侧重于单零件的安装问题, 在装配序列规划问题上还未有相关深入研究. 而对于装配序列规划问题这类典型的序贯决策问题, 其与深度强化学习方法的结合新颖又贴合, 具有很强的理论意义和工程实践意义.

针对复杂产品的装配序列规划问题, 本文的创新点为提出了一种包含正向装配以及逆向拆解的一体化双向装配序列规划方法(bi-directional assembly sequence planning for workpieces with deep Q-learning networks, BASPW–DQN). 相对于现有方法而言, 改进之处与优点如下:

1) 进行了新型动作空间设置. 双向序列规划方法中动作根据当前状态的不同而具有两种含义, 代表着此刻处于前向安装或是拆卸的过程, 从而达到正确且合理的装配模型对象, 完成装配任务, 得到一条完整的包含安装及拆卸的一体化装配序列.

2) 训练过程针对装配序列规划引入了课程学习与迁移学习. 引入了课程学习方法, 从零件初始位置距离、初始已安装零件个数和决策步数几方面进行了不同难度等级的课程划分, 逐步提升网络的决策准确率. 并且使用了对于装配过程中关键帧进行提取的网络模型来进行参数迁移, 加速训练过程. 在训练过程中对于经验池的设置及使用进行了改进, 以提高数据的利用效率.

3) 所提方法对复杂产品(包含多个工件、多种构型、多种初始状态的装配体)的装配序列规划进行了深入研究, 可以适用于任意的模型初始状态, 学习到包含前向装配和逆向错误零件拆卸两部分过程的灵活网络模型, 提高了装配效率以及灵活性.

2 研究现状

进行装配序列的规划, 可以提高装配效率, 保证装配过程的可靠性, 并且降低产品的开发成本, 提高产品设计、加工到装配过程的效率和准确性. 装配序列规划问题实质上是一个典型的组合优化问题, 具有复杂性、强约束、多样性的特点[4]. 以生成低成本、高效率的装配序列为目标, 一些学者提出了初步的装配序列规划方法. 本部分将介绍装配序列规划问题和深度强化学习两方面的研究现状, 并对现有主要问题进行分析.

2.1 装配序列规划

对于装配序列规划问题, 其对应的解决算法应当满足以下两个条件: 1) 适用于具有多个工件、多条可行序列的复杂模型情况; 2) 具有泛化能力. 不针对特定的模型及零件, 不限定模型的初始状态. 这也是装配序列规划问题的难点所在. 现有的研究主要可以分为以下3大类方法: 基于图的序列规划方法、基于知识的序列规划方法, 以及一些启发式的规划算法.

其中基于图的序列规划方法通过有向图来形式化编码可行的装配序列空间[5]. 将一系列装配序列表示为AND/OR图, 并将装配序列的生成问题转化为拆卸序列的生成问题[6–7]. 当处理复杂装配模型, 即模型中包含的零件数目比较大时, 会出现组合爆炸的问题, 其计算代价高而导致实时性难以保证; 并且图是固定的, 泛化能力较弱.

基于知识的序列规划方法通过利用拓扑关系或其它关系来描述装配对象的信息以及一些先验知识(比如目标模型、装配约束、CAD数据库和启发式规则等)[8–9], 以此建立多层模型结构以实现智能装配. 但其规划结果会严重受限于特定的先验知识, 当处理不同类型的装配模型时通用性较差.

启发式智能搜索的方法应用人工智能技术中的搜索算法, 在线搜索优化, 也可以解决零件的装配任务. 如模拟退火方法[10]、遗传算法[11–12]以及人工神经网络[13–14]. 此类方法计算量往往比较大且收敛速度受限, 泛化能力不足.

由此可以看出, 这些方法还不适用于具有多零件的复杂装配体, 且通用能力不足, 需要每个模型单独规划.

2.2 深度强化学习

近些年来, 随着深度学习的快速发展, 强化学习算法也从神经网络的强大表示能力中得到了启发. 与传统的表格型强化学习方法不同, 通过深度网络的表示能力[15–16], 深度强化学习方法可以有效地处理高维特征[17], 解决了传统强化学习方法参数估计困难的问题. 在近几年里, 深度强化学习方法在游戏和机器人领域都取得了许多令人惊叹的突破, 例如已经超越人

类水平的星际争霸游戏[18–19]和围棋AlphaGo[20–21]等等.

在智能装配方面, 现有的深度强化学习方法更多侧重于机械臂抓取或安装过程中的零件识别及轨迹与动作规划[22–23]. 在单零件的安装等方面也涌现出很多优秀的研究成果, 但在装配序列规划问题上还未有深入研究.

对于装配序列规划问题来说, 预先建立模型是非常困难的. 强化学习方法恰恰不依赖于精确模型, 并且可以在没有预先收集大量数据的情况下, 通过直接与环境交互来获得所需的学习数据[24]. 同时, 强化学习非常适合用于解决序贯决策问题, 目的就是在每一个时间状态下, 找到最佳的动作决策, 从而形成一条最优的决策序列[25]. 由此可见, 深度强化学习非常适用于装配序列规划这类难以建模的问题, 并且通过加入神经网络而获得强大知识表征能力, 可以提取出大量数据集中的特征, 得到具有较强泛化能力的模型.

在本文之前的研究工作: 一种装配序列规划系统[26](assembly sequence planning system for workpieces based on deep reinforcement learning, ASPW–DRL)中, 针对多个工件, 多条可行序列的装配任务, 设计了一种基于值函数方法的零件装配序列规划方法. 然而对于该方法, 每当出现了错误的决策, 产生了不正确的中间模型状态时, 都会直接终止此条序列, 无法对误装配的零件进行拆解; 当模型复杂时, 该方法的成功率会受到限制; 并且该方法难以适用于各种不同类型的初始构型, 装配成功率会受到影响, 限制了其应用范围. 基于此, 作者也将其视作一种前向的装配序列规划方法. 并针对包含前向装配和逆向错误零件拆卸两部分过程的双向装配序列规划方法进行研究, 从而学习到一套包含安装及拆卸动作的网络模型, 对于装配体任意状态的装配任务都可以高效完成, 并通过大量测试结果验证其有效性.

3　BASPW–DQN算法

3.1　问题描述与形式化

在装配序列规划问题中, 系统的下一时刻状态只和当前的状态有关, 与之前过程的状态无关, 满足马尔科夫性, 因此, 可以将装配序列规划问题表述为一个马尔科夫决策过程, 看作一个序贯决策问题. 下面将对双向序列规划问题进行形式化描述.

S: 状态空间设置. 双向序列规划问题的状态表示须包含对当前状态的完整描述, 当使用图像的形式化表示时, 由描述当前装配区域状态的二维图像与描述下一时刻可能产生的变化的所选零件的图像组成, 如图1所示.

A: 动作空间设置. 此双向序列规划问题中的动作根据当前状态的不同而具有两种含义. 若当前状态为正常的装配状态时, 此条决策属于前向零件装配过程, 此时的动作代表着安装或不安装此零件. 而当前面产生了错误的决策后, 当前状态为一种不合理的中间装配状态时, 此条决策序列就转移到了错误零件拆卸的过程. 此时的动作则代表着对于待决策的零件, 进行拆卸或不拆卸的动作.

$$a=\begin{cases}\text{动作0: 保持此零件不动,}\\ \text{动作1:}\begin{cases}\text{当前状态正确, 安装此零件,}\\ \text{当前状态错误, 拆卸此零件.}\end{cases}\end{cases}\tag{1}$$

如式(1)所示, 在装配过程中, 若此时装配状态正确且合理, 则在其余的零件中选择一个, 进行$\{0,1\}$的动作选择是否安装. 若此时状态不合理, 则在已安装的零件中选择一个, 进行$\{0,1\}$的动作选择是否拆卸.

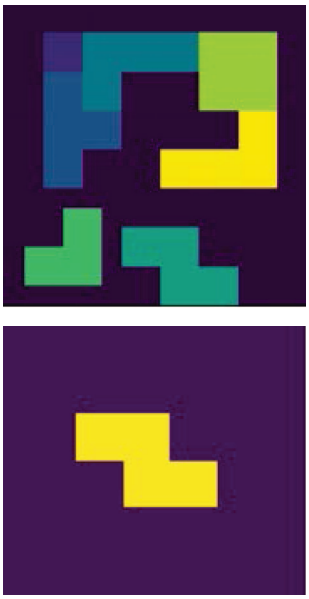


图 1　状态表示两部分示意图

Fig. 1　State representation

R: 回报函数设置. 在双向序列规划问题中, 最终目标是在规定的步数内, 通过不断安装或拆卸错误零件, 从而达到正确且合理的装配模型对象, 完成装配任务. 回报函数设置如下:

$$r_t=[-1,1].\tag{2}$$

如果能在预先设定的决策步数内通过一系列安装及拆卸的动作, 成功完成装配任务, 则认为此条决策序列是正确的, 给予智能体表示奖励的正回报1, 否则, 对智能体此条错误的决策序列进行-1的惩罚.

深度网络设置及训练过程: 对于深度神经网络的搭建和训练部分, 本文采用如图2所示的卷积神经网络结构. 首先是一个卷积层, 利用不同的卷积核来提取原始输入图像的局部特征. 每一个卷积核通过激活函数输出更明显和有效的特性. 卷积层后紧跟着的是一个池化层, 用于处理卷积层得到的特征映射. 整个网络的最后一部分是全连接层, 第1个全连接层用于

将特征映射转换为向量, 而最后1个全连接层用于将特征映射到动作. 本文输出部分为两个动作的Q值, 分别代表了在当前状态下选择此动作所能得到的期望回报值, 用以评估两个动作的好坏.

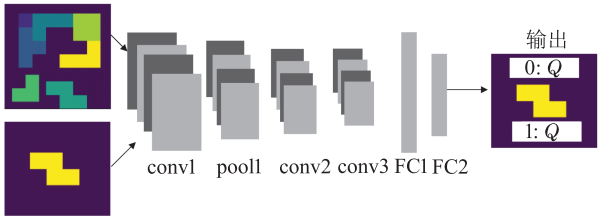


图 2 网络结构设置

Fig. 2 Network structure

双向装配网络的各层具体设置如表1.

表 1 双向序列规划网络各层设置

Table 1 Network setting for assembly model

层	核	特征图
input	—	80×80×2
conv1	[8, 8, 2, 32]	20×20×32
pool1	[1, 2, 2, 1]	10×10×32
conv2	[4, 4, 32, 64]	5×5×64
conv3	[3, 3, 64, 64]	5×5×64
FC1	[5×5×64, 512]	512
FC2	[512, 2]	2

对于马尔科夫决策过程形式化后的装配序列规划问题, 网络训练的目标是通过调整神经网络参数θ, 最大化采样轨迹的累积回报$J(\theta)$:

$$J(\theta) = E_{s_0,a_0,\tau}[\sum_{k=0}^{\lim} \gamma^k R_{t+k+1}], \tag{3}$$

$$\tau = s_0, a_0, r_0, s_1, a_1, r_1, \cdots, \tag{4}$$

其中: γ为折扣因子, k为时间步长, 序列τ为智能体从初始状态s_0开始, 采取动作a, 得到回报r, 直至终止状态的采样轨迹.

为了采样最佳动作, 本文使用了off-policy的策略方式, 以此确保在实践中有足够的探索行为, 从而可以访问状态空间中的每个状态.

该训练过程基于值函数逼近的深度Q学习网络(deep Q-learning networks, DQN)算法[27], 采用经验回放, 在采集训练数据后, 使用均匀随机抽样的方法, 从经验回放池中抽取训练数据, 优化目标函数. 损失函数$L_i(\theta_i)$表示为

$$L_i(\theta_i) = E_{s,a\sim\rho}[(y_i - Q(s,a;\theta_i))^2], \tag{5}$$

$$y_i = E_{s'\sim\xi}[r + \gamma \max_{a'} Q(s', a'; \theta_{i-1})|s, a], \tag{6}$$

其中: θ_i为当前网络参数, ρ代表可能的状态动作分布, $Q(s,a)$为行为值函数. y_i是TD目标, 可看作为Q的真值. 通过对抽取出的数据进行小批量更新的训练, 不断减小y_i与网络预测值$Q(s,a;\theta_i)$之间的差值, 对网络参数进行更新改进.

3.2 前向装配序列规划方法

本节将从复杂模型(多工件、多可行序列)的角度出发, 介绍基于值函数深度强化学习算法的前向装配序列规划方法.

在前向装配网络训练过程中, 输入的状态表示分别为当前装配状态和下一步要进行决策零件的二维图像. 通过这样的表示方法, 可以了解到当前所包含的信息以及下一步即将进行的改变, 根据上述状态表示, 对零件进行安装或不安装的判断. 即动作为

$$a_t = \begin{cases} 0\text{: 保持此零件不动,} \\ 1\text{: 安装此零件.} \end{cases} \tag{7}$$

然后就可以对马尔科夫决策过程形式化后的装配序列规划问题应用标准的强化学习方法, 高效地学习动作策略.

由于强化学习方法所需的数据是动态的, 是通过不断与环境进行交互得到的, 这样就需要采集大量的数据, 还会导致很复杂的网络学习过程, 并且稀疏回报会导致信用分配的问题, 使得网络很难从复杂的环境中学习到良好的动作策略. 所以本文对基于DQN算法的前向装配序列规划方法进行了两部分的改进. 既引入了课程学习的方法来逐步提高网络的学习效果, 避免最初效率低下的学习, 又采用了迁移学习的方法, 以节省网络前期的大量参数学习过程, 提高训练效率. 其中, 课程学习通过设置从简单到复杂的环境来解决稀疏回报所导致的学习过程中给出指导信息较少的问题, 从零件初始位置距离、初始已安装零件个数和决策步数几方面进行了不同难度等级的课程划分, 避免让网络直接从复杂的环境中学习动作策略, 逐步提升网络的决策准确率. 迁移学习部分则使用对于装配过程中关键帧进行提取的网络模型[29]来进行参数迁移, 节省网络前期的大量参数学习过程. 特别是用于特征提取的卷积层部分, 提高训练效率, 减少强化学习算法训练过程中所必需的数据量, 加速训练过程.

3.3 单纯拆卸序列规划方法

本文使用了基于值函数的方法来进行双向装配序列规划的研究, 由于正常状态下的零件安装过程与前向装配序列规划方法[26]相同, 所以此部分将着重介绍当出现不合理构型时的零件拆卸过程序列规划方法.

在实际生产装配过程中, 由于使用工具、零件位置与摆放以及工人习惯等原因, 难免会出现一些不合理的装配顺序. 拆卸作为装配的逆过程, 在装配序列规划问题中一直是一个研究重点. 本部分将聚焦于装配过程中可能出现的错误装配状态, 对此错误状态进行分析, 进行此状态下的拆卸序列规划.

对于单纯拆卸序列规划问题, 目标是对于一些错误决策所导致的不合理中间构型, 通过不断进行零件的拆卸动作, 直至尽可能快速地得到一个正确且合理的装配状态. 对此本文设计了图3的单纯拆卸流程架构.

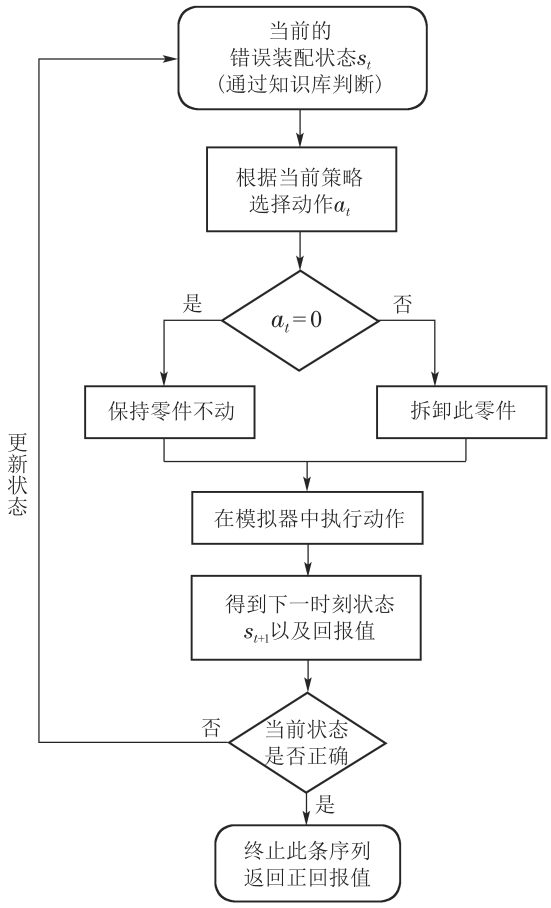


图 3　单纯拆卸算法流程框图

Fig. 3　Disassembly sequence planning process

为了进行错误零件拆卸过程的序列规划, 此网络训练过程中的初始状态为不合理的装配状态s_t, 在当前状态下, 从已安装零件中随机选择一个, 对此零件进行决策, 通过当前网络策略π来选择应执行的动作a_t:

$$a_t=\begin{cases}0\text{: 保持此零件不动,}\\1\text{: 拆卸此零件.}\end{cases}\tag{8}$$

通过在模拟器中执行此动作, 可以得到装配体下一时刻的状态s_{t+1}和对应的回报值r_t, 直至得到正确的装配体状态终止. 经过不断地采样数据并进行梯度下降的网络参数训练, 最终得到最优的网络策略π^*.

3.4　双向序列规划方法: BASPW–DQN

在进行了前一部分的单纯错误零件拆卸序列规划后, 可以快速地从一个错误的装配状态通过拆卸动作顺序拆掉引起状态不合理的零件, 得到一个正确、合理的装配状态. 在此基础上, 本文将结合前向的装配过程和错误零件拆卸的序列规划过程, 学习一套可完成任意状态下装配任务的双向装配序列规划网络模型. 此双向序列规划方法的算法流程如图4所示. 在不同幕的采样过程中, 初始状态为随机初始的. 在每一幕数据中, 当前状态是根据动作在模拟器中实时更新的.

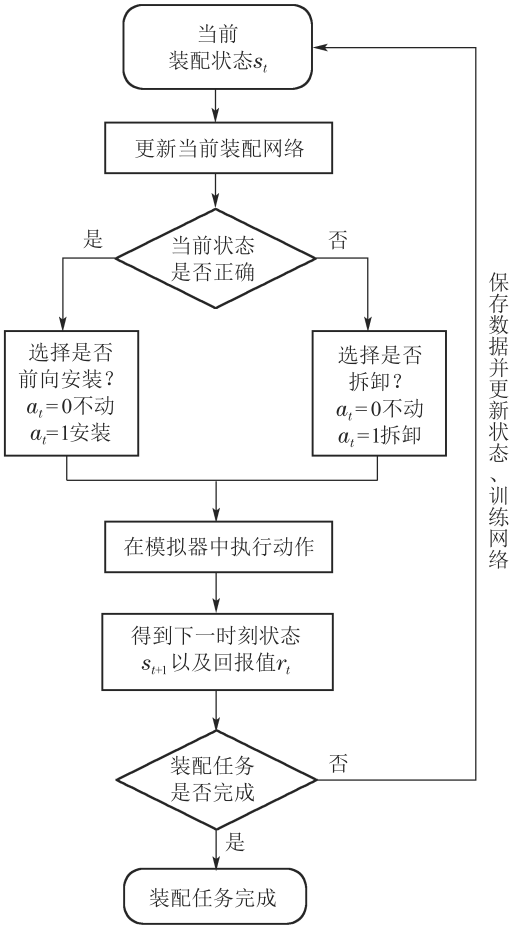


图 4　双向装配序列规划算法流程框图

Fig. 4　BASPW–DQN algorithm

本文提出的双向装配序列规划方法与单纯将前向序列规划方法和错误零件拆卸两部分直接结合不同. 本文所提方法将学习一组完整的装配网络模型, 将两部分耦合在一起, 使用一套双向网络模型, 而不是单纯加和或是网络模型切换, 无需在使用时根据当前状态来调用不同的已学网络模型, 而是在一套网络中根据输出动作的不同意义来完成安装或是拆卸的任务; 在训练时也不需要单独训练安装或拆卸网络, 通过对此双向网络的整体训练即可完成装配任务, 从而最终得到一条完整的装配序列, 体现了装配序列规划的完整性.

并且此双向序列规划网络处于安装过程还是拆卸过程则是根据当前状态的正确与否来判断的. 若当前状态为正确的装配状态, 那么证明当前是前向的安装过程, 需要对剩余未安装的零件进行安装与否的判断;

若此前有过一些错误的决策, 导致当前的装配状态并不合理, 会导致后续零件无法安装, 那么此时就将进行拆卸的动作, 对已安装的零件进行拆卸序列规划直至合理装配状态.

因此, 本文所提出的双向装配序列规划方法能很好的体现装配序列规划过程的完整性和适应能力.

在实现双向装配序列规划方法的过程中, 引入了课程学习的思想, 从而提高了学习效率. 从较为简单的初始状态开始进行, 当已安装的零件较多时, 装配难度较小, 网络采样数据的过程中会有更多成功的序列, 从而使得回传的回报值为正值, 鼓励网络向正确的方向进行采样. 通过逐渐降低初始化的零件个数, 使得网络模型的性能逐渐提高, 适应于各种难度的装配任务. 并且为了正确评估策略值函数, 尽可能全面遍历装配状态集合, 本文进行了探索性的初始化环境设置, 每一次迭代所具有的初始零件个数, 以及已安装的零件都是随机初始化的.

此双向装配序列规划方法相较于单向方法, 对动作空间进行了设计与表示; 并且由于问题难度的极大增加, 引入了课程学习及迁移学习方法, 加速训练过程, 提高装配决策准确率; 最后在训练过程中对于经验池的设置及使用上进行了改进, 以应对不同含义动作所带来的难度提升.

4 仿真验证与结果分析

4.1 仿真平台与设置

本文选择了ROS下的Gazebo仿真平台, 并且在ROS-Gazebo中引入Python接口, 建立了和深度网络相结合的框架. 其中, Gazebo作为物理仿真平台, 主要完成零件控制和与环境交互的任务, TensorFlow则是当前最常用的实现机器学习算法的开源软件库算法库. 本文通过TensorFlow来搭建深度网络, 实现强化学习算法. 本文搭建并调试了图5所示的用于装配序列规划问题的训练平台, 以完成交互数据采集和装配过程可视化的任务. 在仿真中通过订阅相机发布的信息, 转成OpenCV可以识别的图像, 并进行颜色或像素的转换, 从而得到输入图像.

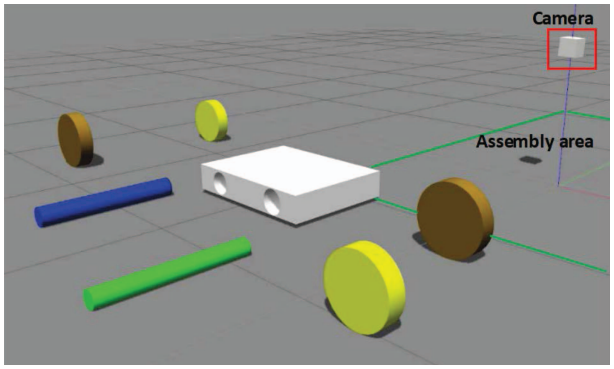


图 5 仿真平台示意图

Fig. 5 Simulation platform

在验证双向序列规划方法之前, 首先进行单纯考虑错误零件拆卸序列规划部分的仿真. 为了更加直观明显地体现装配序列规划的意义与难度, 在本文的仿真中, 设计了一个如图6所示类似于七巧板结构的七零件模型, 作为本次仿真的装配对象. 此七巧板模型零件的形状比较特别, 零件的差异十分明显, 不同零件会有多种组合关系, 零件之间的相互接触关系较为复杂, 能够体现装配序列规划的重要性.

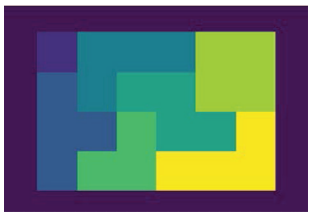


图 6 七巧板仿真模型

Fig. 6 Seven-part assembly model

仿真设置: 给定多种不合理的不同装配状态, 每次采样时的初始状态从多种不合理构型中随机选择. 对于在每一步中要进行决策的已安装零件, 动作分别为保持零件不动或拆卸掉此零件. 如果在设置的步数内完成了对错误零件的拆卸, 得到了正确的装配状态, 则结束此次采样过程. 根据零件初始位置的不同, 零件装配过程中的序列也不尽相同, 本文设计了两种不同的零件初始位置如图7, 分别代表了不同难度的装配任务. 以仿真环境中的坐标轴为标准, 根据正确装配位置与零件初始位置之间距离的远近来划分.

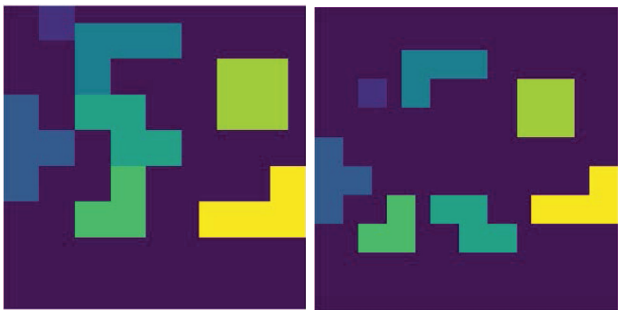


图 7 不同难度初始情况

Fig. 7 Initial situation of different level

其中, 左图的初始状态中, 零件之间距离较近, 模型安装过程中对装配顺序的要求较低, 得到符合要求的成功序列几率更高, 代表了此问题的低级难度. 右图中零件初始位置距离正确安装位置较远, 对装配顺序的要求更高, 则代表了此问题设置中的较高难度.

在测试过程中, 每10轮训练后进行一轮测试, 每一轮测试中包含100个装配任务, 计算装配成功率.

4.2 单纯拆卸仿真结果及分析

对于难度一的任务, 给定400种不合理构型, 每次的初始状态从中随机选择, 要求智能体可以在7步内完成拆卸任务得到合理构型. 下面给出难度一的拆卸

任务仿真结果图, 由图8损失函数下降曲线可以看出, 经过训练, 网络预测的值函数与目标的误差趋近于零, 并且由图9成功率曲线可以看出, 7步内完成任务的成功率从初始的42%增长至96%, 并稳定在90%以上.

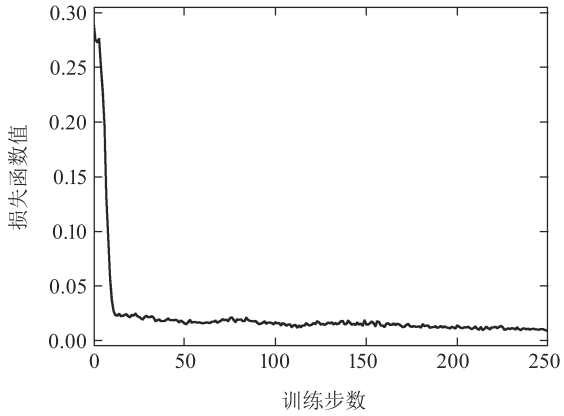


图 8　难度1仿真损失函数图

Fig. 8　Loss function curve of Difficulty 1

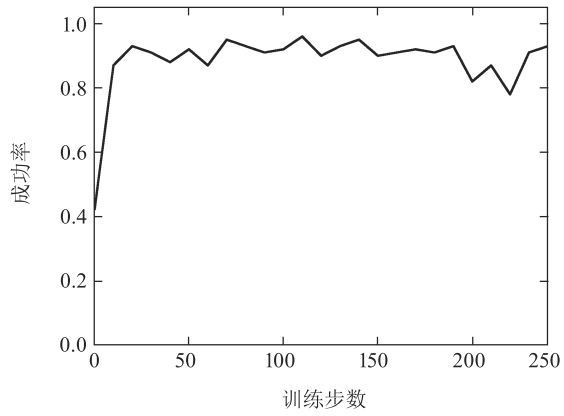


图 9　难度1仿真成功率曲线

Fig. 9　Success rate of Difficulty 1

由于难度2任务的装配序列更加复杂, 所以给定1200种不合理构型, 每次的初始状态从中随机选择, 也要求智能体可以在7步内完成拆卸任务得到合理构型. 如图10所示为难度2在7步内完成拆卸任务的成功率曲线, 可以看出, 7步内的成功率从初始的76%可增长至100%. 并在此基础上进行5步内完成任务的训练, 由图11可看出成功率从初始的64%增长至最高97%, 并可以稳定在95%以上, 证明了单纯拆卸序列规划方法的可行性及有效性.

图12给出两个单纯拆卸任务的序列实例. 实例中最左边为初始的错误构型, 最右边为拆卸后得到的合理构型. 实例1中由于左右两边零件的遮挡, 导致中间的零件无法正确安装, 经过训练后的网络可以准确找到引起当前问题的零件, 通过拆卸掉这两个零件, 从而得到一个合理的装配状态. 实例2中由于其它零件的错误安装顺序, 导致应当安装在最中间的零件无法安装, 由图12中的红色框可以看出, 网络能够正确拆卸顺序错误的零件, 从而使得后续的零件可以在当前构型下继续安装以完成装配任务.

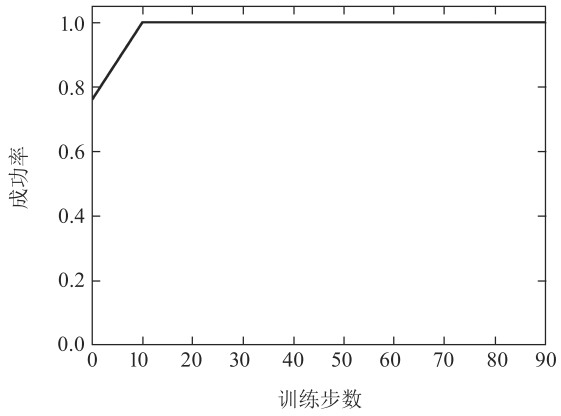


图 10　难度2在7步内完成任务成功率曲线

Fig. 10　Experimental results of Difficulty 2 in 7 steps

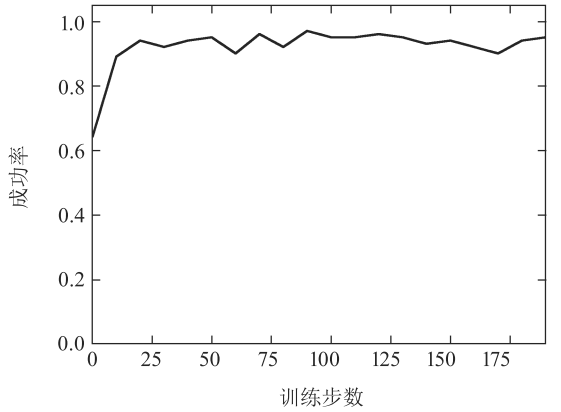


图 11　难度2在5步内完成任务成功率曲线

Fig. 11　Experimental results of Difficulty 2 in 5 steps

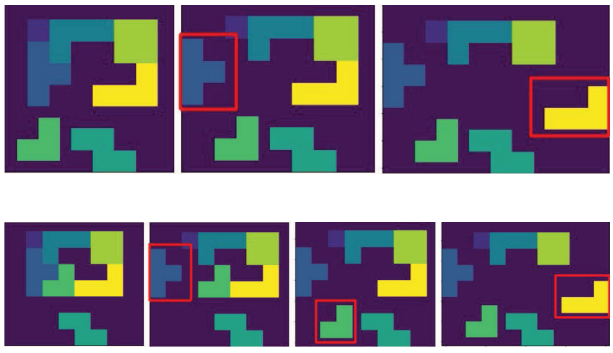


图 12　两个单纯拆卸序列实例

Fig. 12　Two examples of disassembly sequence

4.3　双向序列规划仿真结果及分析

在前向装配序列规划以及单纯拆卸序列规划的基础上, 本文又进行了双向装配序列规划方法的仿真验证. 引入课程学习的思路, 首先进行6个零件初始的双向装配序列规划实验, 要求智能体可以在10步内完成装配任务, 以此进行训练. 最终的测试证明在6个零件

初始的情况下该网络模型3步内就可以100%完成装配任务.

紧接着使用6个零件初始所学习到的网络模型作为初始网络, 进行4个零件初始的双向装配序列规划仿真, 要求智能体可以在15步内完成装配任务. 由图13中的成功率增长曲线可以看到经过训练后此网络可以100%完成此难度的装配任务.

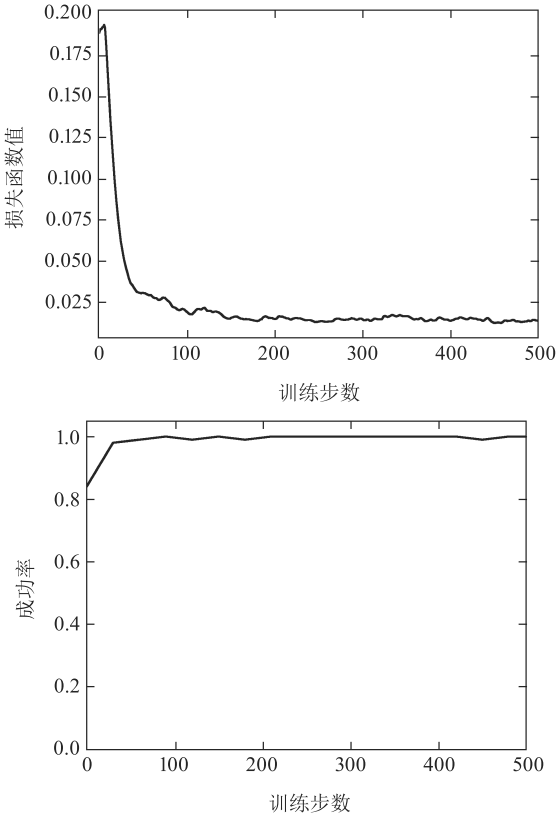
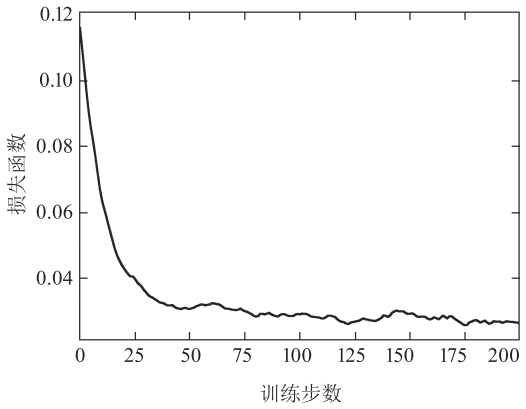


图 13 4个零件初始仿真结果图

Fig. 13 Experimental results of 4 parts

然后对零初始条件下的装配状态进行训练, 步数设为20, 结果如图14所示, 成功率从近30%升到100%. 由于使用了4个零件初始的网络作为初始网络, 所以此模型具有一些对于装配任务的决策能力, 使得初始成功率高于零. 但另一方面, 由于问题的难度逐渐增加, 网络模型的决策能力还不能胜任更加困难的装配状态, 所以初始的成功率仅有30%.



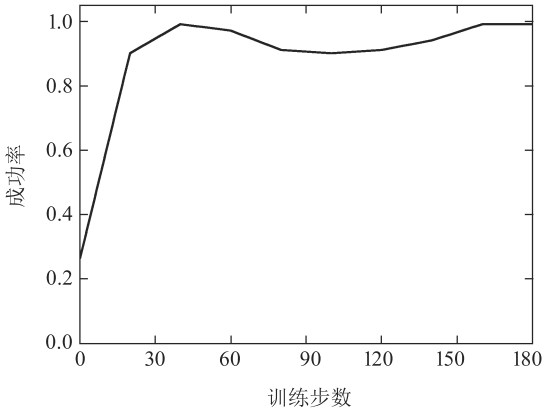


图 14 零初始仿真结果图

Fig. 14 Experimental results with no initial part

并在训练过程中, 对于学习率(lr)与折扣因子(γ)两个参数, 进行了一些消融实验来选择合适的超参数. 由图15曲线可看出, 对于本文的问题设置与网络结构, 当学习率为10^{-4}、折扣因子为0.95时, 网络收敛速度最快.

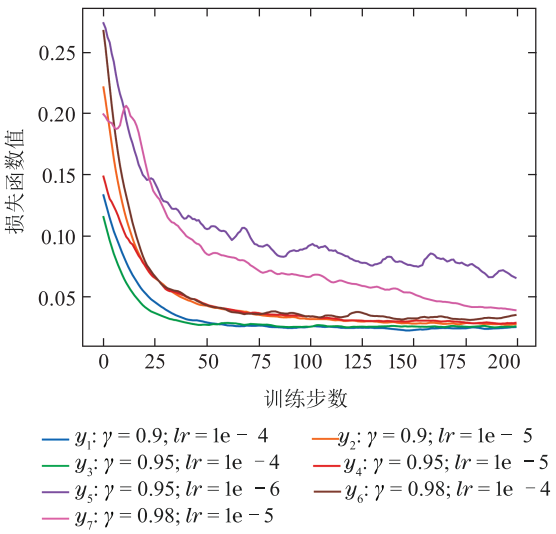


图 15 不同超参数损失函数曲线

Fig. 15 Loss function curves of different hyperparameters

紧接着将学习到的模型应用于随机初始化零件个数的情况: 对于初始零件个数随机的任意情况可以100%完成任务. 如双向序列规划过程实例所示, 训练好的网络对于如图16的不合理的初始装配状态, 可以快速找出应拆卸的零件(红框), 得到合理装配构型, 然后继续安装以完成装配任务.

对于如图17的合理的初始装配状态, 则可以快速完成任务. 证明此双向网络对于任意的装配状态(包括各种合理或不合理的初始状态)均可以在较少步数内完成装配任务, 验证了本文所提出的双向序列规划方法的有效性以及较高的适应能力.

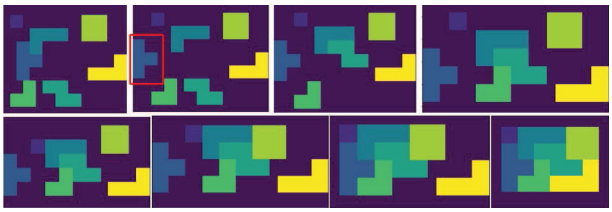


图 16 双向序列规划实例1

Fig. 16 Bi-directional sequence Example 1

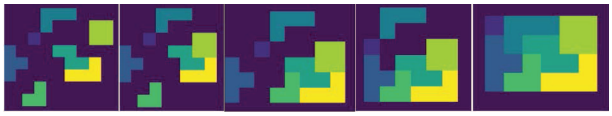


图 17 双向序列规划实例2

Fig. 17 Bi-directional sequence Example 2

4.4 与现有方法对比分析

由于基于图的序列规划方法提出较早、使用较多, 现已发展成为一种比较成熟经典的装配序列规划算法. 因此, 本文从规划时间、准确率、通用性这3方面对本文所提方法以及基于图的规划方法[7,28]进行了对比.

表 2 与现有方法对比效果

Table 2 Compared with existing method

规划算法	规划时间/s	装配准确率/%
基于图的规划方法[7,28]	1.12	100
基于深度强化学习的规划方法	0.0061	100

在装配准确率方面, 两种方法都可以成功完成装配任务. 而在规划时间方面, 基于图的规划算法要对所构建的完整装配图进行割集分解, 其计算速度与关联图的稀疏性以及装配工件的数量有很大关系, 需要实时的规划计算, 会消耗一定的时间. 而本文提出的基于深度强化学习的规划方法通过引入课程学习的思想, 可以通过预训练得到一个具有较好规划效果的装配网络. 在使用时可以直接调用网络参数解决问题, 可在毫秒级的时间内完成一整条的装配序列规划任务, 解决问题速度更快.

针对通用性问题, 基于图的序列规划算法需要在每次启动时构建装配模型对应的装配关联图, 当模型初始状态不同时必须重新构建图, 不能适用于任意的装配模型状态. 而由本文的仿真实验设置及结果可以看出, 本文所提方法适用于任意随机初始化的装配状态, 可以顺利完成所对应的装配任务, 更加灵活且通用.

5 结论

本文基于值函数的方法, 考虑了由于某步决策错误而导致的不合理的零件中间装配状态. 首先聚焦于错误零件的拆卸序列规划, 进行了单纯拆卸过程的序列规划研究; 并将前向序列规划的安装过程与错误零件的拆卸过程相结合, 提出了一种基于DQN的双向装配序列规划方法BASPW–DQN. 相较于单向装配方法, 进行了新颖的动作空间设计与表示; 并且由于问题难度的极大增加, 引入了课程学习及迁移学习方法, 加速训练过程, 提高装配决策准确率, 可以学习到一套包含安装及拆卸动作的一体化双向网络模型, 完成装配体任意状态(包括零装配、部分装配、误装配等初始状态)的装配任务. 最后在搭建的训练平台中进行了验证, 测试结果证明此双向网络对于任意装配状态的任务均可以在较少步数内完成. 本文探索了较为全面的可能装配状态, 提高了所学装配模型的有效性与适应性.

参考文献:

[1] LI Bailin. *Research and application of production control system for intelligent manufacturing*. Harbin: Harbin Institute of Technology, 2019.
(李柏林. 面向智能制造的生产控制系统研究与应用. 哈尔滨: 哈尔滨工业大学, 2019.)

[2] XU L D, WANG C, BI Z, et al. Object-oriented templates for automated assembly planning of complex products. *IEEE Transactions on Automation Science and Engineering*, 2014,11(2): 492 – 503.

[3] XU L D, WANG C, BI Z, et al. AutoAssem: An automated assembly planning system for complex products. *IEEE Transactions on Industrial Informatics*, 2012, 8(3): 669 – 678.

[4] JONES R E, WILSON R H, CALTON T L. On constraints in assembly planning. *IEEE Transactions on Robotics and Automation*, 1998, 14(6): 849 – 863.

[5] HOMEM DE MELLO L S, SANDERSON A C. Representations of mechanical assembly sequences. *IEEE Transactions on Robotics and Automation*, 1991, 7(2): 211 – 227.

[6] KARJALAINEN I, XING Y, CHEN G, et al. Assembly sequence planning of automobile body components based on liaison graph. *Assembly Automation*, 2007, 27(2): 157 – 164.

[7] HOMEM DE MELLO L S, SANDERSON A C. A correct and complete algorithm for the generation of mechanical assembly sequences. *IEEE Transactions on Robotics and Automation*, 1991, 7(2): 228 – 240.

[8] ZHA X F, LIM S Y E, FOK S C. Integrated knowledge-based Petri net intelligent flexible assembly planning. *Journal of Intelligent Manufacturing*, 1998, 9(3): 235 – 250.

[9] KASHKOUSH M, ELMARAGHY H. Knowledge-based model for constructing master assembly sequence. *Journal of Manufacturing Systems*, 2015, 34: 43 – 52.

[10] CAKIR B, ALTIPARMAK F, DENGIZ B. Multi-objective optimization of a stochastic assembly line balancing: A hybrid simulated annealing algorithm. *Computers and Industrial Engineering*, 2011, 60(3): 376 – 384.

[11] MARIAN R M, LUONG L H, ABHARY K. A genetic algorithm for the optimisation of assembly sequences. *Computers and Industrial Engineering*, 2006, 50(4): 503 – 527.

[12] WANG W, TSENG H. Complexity estimation for genetic assembly sequence planning. *Journal of the Chinese Institute of Industrial Engineers*, 2009, 26(1): 44 – 52.

[13] CHO H. A neural network-based computational scheme for generating optimized robotic assembly sequence. *Engineering Applications of Artificial Intelligence*, 1995, 8(2): 129 – 145.

[14] SINANOGLU C, BORKLU H R. An assembly sequence planning system for mechanical parts using neural network. *Assembly Automation*, 2005, 25(1): 38 – 52.

[15] BENGIO Y, COURVILLE A, VINCENT P. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013, 35(8): 1798 – 1828.

[16] LECUN Y, BENGIO Y, HINTON G. Deep learning. *Nature*, 2015, 521(7553): 436 – 444.

[17] ARULKUMARAN K, DEISENROTH M P, BRUNDAGE M, et al. Deep reinforcement learning: A brief survey. *IEEE Signal Processing Magazine*, 2017, 34(6): 26 – 38.

[18] SUN P, SUN X, HAN L, et al. TStarBots: Defeating the cheating level builtin AI in starcraft II in the full game. *ArXiv*, 2018, 1809.07193.

[19] VINYALS O, EWALDS T, BARTUNOV S, et al. StarCraft II: A new challenge for reinforcement learning. *ArXiv*, 2017, 1708.04782.

[20] SILVER D, HUANG A, MADDISON C J, et al. Mastering the game of go with deep neural networks and tree search. *Nature*, 2016, 529(7587): 484 – 489.

[21] SILVER D, SCHRITTWIESER J, SIMONYAN K, et al. Mastering the game of go without human knowledge. *Nature*, 2017, 550(7676): 354 – 359.

[22] QUILLEN D, JANG E, NACHUM O, et al. Deep reinforcement learning for vision based robotic grasping: A simulated comparative evaluation of off-policy methods. *ArXiv: 1802.10264*, 2018.

[23] LI F, JIANG Q, ZHANG S, et al. Robot skill acquisition in assembly process using deep reinforcement learning. *Neurocomputing*, 2019, 345: 92 – 102.

[24] DEGRIS T, PILARSKI P, SUTTON R S. Model-free reinforcement learning with continuous action in practice. *American Control Conference*. Montreal, QC, Canada: IEEE, 2012: 2177 – 2182.

[25] SUTTON R S, BARTO A G. *Reinforcement Learning: An Introduction*. New Jersey: John Wiley & Sons Inc., 1998.

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

[27] MNIH V, KAVUKCUOGLU K, SILVER D. Human-level control through deep reinforcement learning. *Nature*, 2015, 518(7540): 529 – 542.

[28] FU Yili, TIAN Lizhong, XIE Long, et al. Assembly sequences planning based on cut set analysis of directional graph. *Journal of Mechanical Engineering*, 2003, 39(6): 58 – 62.
(付宜利, 田立中, 谢龙, 等. 基于有向割集分解的装配序列生成方法. 机械工程学报, 2003, 39(6): 58 – 62.)

[29] ZHAO M, GUO X, ZHANG X. Key frame extraction of assembly process based on deep learning. *IEEE International Conference on Cyber Technology in Automation, Control and Intelligent Systems*. Tianjin, China: IEEE, 2018: 611 – 616.

作者简介:

赵铭慧 助理实验师, 硕士, 目前研究方向为深度强化学习, E-mail: zmh@mail.nankai.edu.cn;

张雪波 教授, 博士生导师, 目前研究方向为移动机器人视觉控制、实时最优运动规划、SLAM与自主导航、多摄像机网络优化、深度强化学习等智能决策及其在机器人与运动体博弈系统中的应用, E-mail: zhangxuebo@nankai.edu.cn;

郭 宪 副研究员, 目前研究方向为强化学习、多智能体技术、博弈论等在机器人领域中的研究和应用, E-mail: guoxian@nankai.edu.cn;

欧勇盛 研究员, 博士生导师, 副主任, 目前研究方向为低成本移动机器人导航研发与应用、基于学习人类策略的机械臂操作智能控制方法和机器人感知与人机共融技术及应用等, E-mail: ys.ou@siat.ac.cn.