控制与决策
Control and Decision

基于分层深度强化学习的移动机器人导航方法

王童, 李骜, 宋海荦, 刘伟, 王明会

基于分层深度强化学习的移动机器人导航方法

王 童, 李 骜[†], 宋海荦, 刘 伟, 王明会

(中国科学技术大学 信息科学技术学院, 合肥 230027)

摘 要: 针对现有基于深度强化学习 (deep reinforcement learning, DRL) 的分层导航方法在包含长廊、死角等结构的复杂环境下导航效果不佳的问题, 提出一种基于 option-based 分层深度强化学习 (hierarchical deep reinforcement learning, HDRL) 的移动机器人导航方法. 该方法的模型框架分为高层和低层两部分, 其中低层的避障和目标驱动控制模型分别实现避障和目标接近两种行为策略, 高层的行为选择模型可自动学习稳定、可靠的行为选择策略, 从而有效避免对人为设计调控规则的依赖. 此外, 所提出方法通过对避障控制模型进行优化训练, 使学习到的避障策略更加适用于复杂环境下的导航任务. 在与现有 DRL 方法的对比实验中, 所提出方法在全部仿真测试环境中均取得最高的导航成功率, 同时在其他指标上也具有整体优势, 表明所提出方法可有效解决复杂环境下导航效果不佳的问题, 且具有较强的泛化能力. 此外, 真实环境下的测试进一步验证了所提出方法的潜在应用价值.

关键词: 深度强化学习; 分层深度强化学习; 移动机器人; 导航; 避障; 策略学习

中图分类号: TP242 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2021.1013 **开放科学 (资源服务) 标识码 (OSID):**

引用格式: 王童, 李骜, 宋海荦, 等. 基于分层深度强化学习的移动机器人导航方法 [J]. 控制与决策, 2022, 37(11): 2799-2807.

Navigation method for mobile robot based on hierarchical deep reinforcement learning

WANG Tong, LI Ao[†], SONG Hai-luo, LIU Wei, WANG Ming-hui

(School of Information Science and Technology, University of Science and Technology of China, Hefei 230027, China)

Abstract: In order to solve the problem that existing hierarchical navigation methods based on deep reinforcement learning (DRL) perform poorly in complex environments including the structures like long corridors and dead corners, we propose a navigation method for mobile robots based on option-based hierarchical deep reinforcement learning (HDRL). The framework of the proposed method consists of two level control models: a low level model is to obtain policies for avoiding obstacles and reaching the goal respectively, and a high-level behavior selection model is for automatically learning stable and reliable behavior selection policy, which does not rely on manually designed control rules. In addition, a training method for optimizing the obstacle avoidance control model is proposed, which makes the learned obstacle avoidance policy more suitable for the navigation task in complex environments. In comparison with existing DRL-based navigation methods, the proposed method achieves the highest navigation success rate in all simulated test environments used in this paper and shows better overall performance on other metrics, which demonstrates the proposed method can effectively solve the problem of poor navigation performance in complex environments and has strong generalization ability. Moreover, experiments in real-world environment also verify the potential application value of the proposed method.

Keywords: deep reinforcement learning; hierarchical deep reinforcement learning; mobile robot; navigation; obstacle avoidance; policy learning

0 引 言

导航是移动机器人的一个核心功能, 是指移动机器人从当前位置无碰撞地移动到目标位置的过程[1]. 传统的移动机器人导航方法通常需要高精度传感器构建环境地图或依赖专家经验人为设计规则[2], 难以适用于复杂未知环境下的移动机器人导航任务[3]. 不同于传统导航方法, 基于强化学习的导航方法通过与环境交互学习导航策略, 不仅避免了对环

收稿日期: 2021-06-09; **录用日期:** 2021-11-26.

基金项目: 中国科学技术大学优秀引进人才基金项目 (KY2100000021); 国家自然科学基金项目 (61971393, 61871361).

[†]通讯作者. E-mail: aoli@ustc.edu.cn.

境地图和专家经验的依赖[4],而且具有较强的自适应能力和鲁棒性等优点[3,5]. 特别是近年来深度强化学习(deep reinforcement learning, DRL)取得了快速发展, DRL利用深度学习强大的感知与拟合能力学习高维环境状态到控制动作之间的映射,从而能够获得更好的导航策略[6-7].

现有DRL导航方法的研究中,通常将导航任务视为避障与目标接近两个子任务的结合,通过分别对其设计奖励函数,引导移动机器人在避开障碍的同时接近目标点. 这类方法在简单环境下效果较好,但在包含长廊、死角等结构的复杂环境中容易出现局部极小问题[8],从而给导航策略的学习带来极大的困难[9]. 此外,设计复杂的奖励函数需要依赖大量的先验知识,不具有通用性[10].

为缓解上述问题,近期研究中提出了基于DRL的分层导航方法. 例如, Zhang等[11]提出了一种基于动作优势加权的策略融合方法,其中避障和目标接近两种低层行为策略均通过独立训练DRL获得,在此基础上通过人为设计的规则对低层策略的动作优势值进行加权,得到用于导航的控制动作. 此外, Ding等[12]提出一种用于多机器人导航任务的分层控制方法,利用DRL和简单规则分别得到避障策略和目标接近策略,并对相应的控制动作进行加权输出. 尽管上述基于DRL的分层导航方法能够降低策略学习的难度,但与基于分层强化学习的导航方法[13]不同之处在于,其高层策略始终依赖人为设计的调控规则,在复杂环境下依然存在导航效果不佳的问题[14].

近年来,分层深度强化学习(hierarchical deep reinforcement learning, HDRL)将深度学习与分层强化学习进行有效结合,逐渐成为机器人领域的研究热点[15]. 现有HDRL方法可分为subgoal-based[16]与option-based[17]两类,其中subgoal-based方法中的高层策略负责生成子目标,再由低层策略输出控制动作进行实现[10]. 例如, Li等[18]提出了一种用于实现移动机械臂复杂控制的HDRL方法,该方法中的高层策略用于输出移动底盘或机械臂的子目标姿态,低层策略负责控制移动机械臂到达相应的子目标姿态,使其实现移动或开门操作,从而完成开门进入房间的任务. 而option-based方法中的高层策略负责在多个低层策略中进行选择,再由被选择的低层策略输出控制动作[10,19]. 例如, Li等[20]利用HDRL方法实现了六足机器人移动控制,其中低层策略负责控制机器人关节完成前进、转向等基础运动,而高层策略负责规划低层策略的执行顺序,从而引导机器人到达目标位置. 综上所述,虽然现有HDRL方法在多种机器人控制任务中取得了较好的效果,但均不能直接应用于复杂环境下的移动机器人导航任务. 如上述Li等[18]用于移动机械臂控制的HDRL方法需要依赖先验知识设计具有任务特异性的子目标空间,因此不适用于导航任务. Li等[20]设计HDRL方法时仅考虑在无障碍物的简单环境中对机器人进行移动控制,无法应用于复杂环境下的移动机器人导航任务. 因此,如何将HDRL有效应用于复杂环境下的移动机器人导航任务,具有重要的研究意义.

为解决现有DRL导航方法在包含长廊、死角等结构的复杂环境中难以学习导航策略的问题,本文提出一种可自动学习分层控制策略的导航方法. 该方法采用基于option-based HDRL的模型框架,通过设计基于规则的目标驱动控制模型和基于DRL的避障控制模型,分别实现目标接近与避障两种低层行为策略. 在此基础上,利用基于DRL的行为选择模型学习稳定、可靠的高层行为选择策略,能够有效避免对人为设计调控规则的依赖. 此外,采用优化避障策略的训练方法: 在训练避障控制模型的同时训练行为选择模型,并将接近行为的经验数据也存入避障控制模型的经验池,从而能够利用离策略(off-policy)学习方式优化用于导航任务的避障策略,进一步提升导航效果.

1 基于分层深度强化学习的导航方法

1.1 模型框架

本文提出的移动机器人导航方法采用基于option-based HDRL的模型框架[21],如图1所示. 其高层行为选择模型采用全连接神经网络(图中数字表示每个全连接层的神经元数量),用于学习行为选择策略; 低层避障控制模型也采用类似的全连接神经网络,负责学习避障策略; 低层目标驱动控制模型采用基于规则的控制器,负责实现目标接近策略. 通过该方法,移动机器人能够在不同状态下选出合理的行为,并通过相应的行为策略输出控制动作,从而显著降低导航策略的学习难度.

1.1.1 避障控制模型

为有效学习移动机器人的避障策略,采用竞争双深度Q网络(dueling double deep Q-network, D3QN)[22]实现避障控制模型,其状态空间S_l包含连续三帧的2D激光雷达扫描数据,每一帧数据是长度为40的一维数组,其中每个元素对应于单束激光的距离测量值. 将连续三帧的激光数据按采集时间顺序进行拼接, 构成避障控制模型的输入状态s_t^l[23]. 避

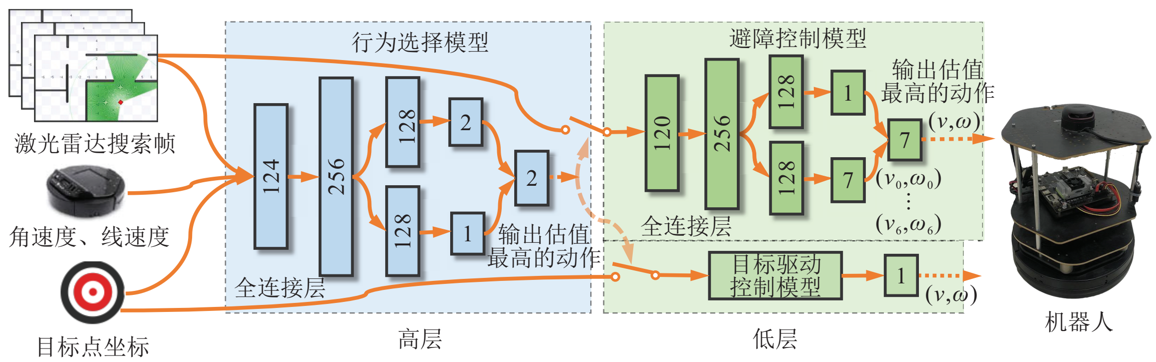


图 1　面向移动机器人导航的 HDRL 模型框架

障控制模型的动作空间 A_l 包括 7 种离散动作 $a^l_{(0)}, \dots, a^l_{(6)}$, 每种动作 $a^l_{(i)}$ 均对应一对预设的线速度和角速度 (v_i, ω_i). 避障控制模型如图 1 右上部分所示, 首先通过一个全连接层对输入状态 s^l_t 进行编码, 得到长度为 256 的特征向量, 然后输入两条支路分别得到状态价值 (state-value) $V\left(s^l_t\right)$ 和动作优势值 (action-advantage) $A(s^l_t, a^l_t)$. 其中状态价值表征当前状态能够带来的长期收益, 动作优势值用于衡量不同动作之间的相对优劣[23]. 在此基础上, 将两值相加得到每一种动作的动作价值 (action-value) $Q(s^l_t, a^l_t)$[9], 用于表征当前动作能够带来的长期收益, 其中动作价值最大的动作将作为避障控制模型的输出, 用于控制移动机器人进行避障. 相应的动作价值更新公式如下:

$$Q(s_t, a_t; \theta) = Q(s_t, a_t; \theta) + \alpha[y_t - Q(s_t, a_t; \theta)]. \quad (1)$$

其中: $y_t = r_t + \gamma \hat{Q}(s_{t+1}, \arg\min_{a \in A_l} Q(s_{t+1}, a; \theta); \theta^-)$ 为目标动作价值, r_t 为当前时刻的奖励 (reward), γ 为折扣因子 (discount factor), s_{t+1} 为下一时刻输入状态, θ 和 θ^- 分别为 D3QN 中 Q 网络和目标 Q 网络参数[22].

在训练阶段, 为使移动机器人与障碍保持距离的同时能够以较大的线速度移动, 避障控制模型的奖励函数设计为

$$r^l_t = v_t + (r^c)_t. \quad (2)$$

其中: v_t 为移动机器人当前时刻的线速度; $(r^c)_t$ 为碰撞惩罚, 且有

$$(r^c)_t = \begin{cases} r_{\text{collision}}, & D_t < R_{\text{robot}}; \\ -\dfrac{\lambda^c}{D_t}, & D_t \geqslant R_{\text{robot}}. \end{cases} \quad (3)$$

这里: λ^c 为权重系数, R_{robot} 为机器人底盘半径 (0.2 m), $r_{\text{collision}}$ 为预定义的参数, D_t 为移动机器人当前时刻与障碍之间的最小距离.

1.1.2　目标驱动控制模型

本文目标驱动控制模型用于控制机器人移动至目标点, 由于在设计目标驱动控制模型时无需考虑躲避障碍[11], 利用基于规则的控制器即可获得有效的目标接近策略. 在如图 2 所示的机器人坐标系下, 控制器首先获取目标点的实时坐标 $[x_t, y_t]$, 然后根据目标点与机器人之间的几何关系[12], 通过下式计算当前时刻移动机器人前进方向与目标方向之间的偏转角度:

$$\varphi_t = \begin{cases} \arctan\left(\dfrac{y_t}{x_t}\right), & x_t > 0; \\ \arctan\left(\dfrac{y_t}{x_t}\right) + \left(\dfrac{y_t}{|y_t|}\right)\pi, & x_t < 0. \end{cases} \quad (4)$$

最后通过如图 2 所示的动作对应关系, 将偏转角度 φ_t 映射到相应的动作. 例如, 当计算得到的 φ_t 在预定义范围 $(-\pi, -12\delta]$ 内时, 即可确定输出动作 $a^l_{(0)}$ 及其对应的线速度和角速度, 从而控制移动机器人朝向目标点移动. 图中 $\delta \in (0, \pi/12)$ 为预定义的控制器参数, 本文设置为 $\pi/20$.

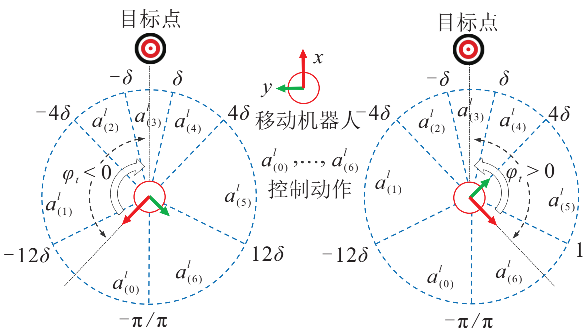


图 2　偏转角度与控制动作的对应关系

1.1.3　行为选择模型

为解决现有 DRL 分层导航方法中行为选择模型依赖人为规则的问题, 行为选择模型采用与避障控制模型类似的 D3QN, 但状态空间与动作空间的定义不同. 如图 1 所示, 为了更好地学习行为选择策略, 其状态空间 S_h 中除激光雷达扫描数据外, 还增加了目标点的相对坐标和移动机器人的线速度与角速度[23]. 而动作空间 A_h 包含 2 种离散动作 $a^h_{(0)}$ 和 $a^h_{(1)}$, 分别对应于避障和目标接近两种行为[9], 其中动作价值较大

的动作用于激活相应的行为策略,从而实现对移动机器人行为的自动选择.

为了得到较优的导航策略,行为选择模型的决策间隔在选择目标接近行为时设为1个时间步(time step),在选择避障行为时设为$n(n>1)$个时间步,从而有利于缩短决策链长度,降低策略学习的难度[9-10]. 在此基础上,训练阶段采用如下公式计算D3QN的目标动作价值[24-25]:

$$y_t = \sum_{\tau=0}^{k-1}\gamma^{\tau} r^h_{t+\tau} + \gamma^k \hat{Q}(s^h_{t+k}, \underset{a\in A_h}{\arg\max} Q(s^h_{t+k}, a;\theta);\theta^-). \tag{5}$$

其中: $k\in\{1,n\}$为决策间隔, s^h_{t+k}和a^h_{t+k}为下一次决策时的输入状态和输出动作, $r^h_{t+\tau}$为时刻$t+\tau$的即时奖励,每次决策的回报$\sum_{\tau=0}^{k-1}\gamma^{\tau} r^h_{t+\tau}$为相邻两次决策之间所有即时奖励的折扣累积.

为控制移动机器人实现复杂环境下的导航,行为选择模型的奖励函数设计为

$$r^h_t = (r^d)_t + (r^o)_t + (r^s)_t, \tag{6}$$

其中$(r^d)_t$用于引导移动机器人接近目标点,具体形式为

$$(r^d)_t = \lambda^d(d_{t-1} - d_t). \tag{7}$$

其中: λ^d为权重系数, d_{t-1}和d_t分别为上一时刻与当前时刻移动机器人与目标点之间的距离. $(r^o)_t$用于引导移动机器人与障碍保持安全距离,有

$$(r^o)_t = \begin{cases} \lambda^o(D_t - D^o), \ D_t < D_{\text{thres}}; \\ r_{\text{obstacle}}, \ D_t \geqslant D_{\text{thres}}. \end{cases} \tag{8}$$

这里: λ^o为权重系数, D^o、D_{thres}和r_{obstacle}为预定义的参数. 若移动机器人当前时刻到达目标点,则获得奖励$(r^s)_t=20$,若发生碰撞则受到惩罚$(r^s)_t=-20$.

通过上述行为选择模型得到的高层选择策略如图3所示. 移动机器人根据该策略在当前时刻t选择目标接近或避障行为,同时确定当前决策间隔k. 在低层行为策略输出相应的k步控制动作后,移动机器人再次根据高层策略选择下一步的行为.

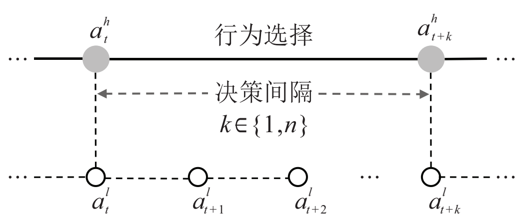


图3 高层行为选择策略

1.2 模型训练

现有DRL分层导航方法采用独立训练的方式学习避障策略[11-12]. 这类训练方法单独对避障控制模型进行训练,具体步骤为: 首先使移动机器人与环境交互,产生避障行为经验数据并存入避障控制模型经验池; 然后仅利用避障行为经验数据更新避障策略. 通过该训练方法学习到的避障策略虽然能够使机器人成功躲避障碍,但易出现局部极小问题[11],导致其难以满足在复杂环境下导航的避障需求. 针对该问题,本文提出一种用于优化避障策略的训练方法. 如图4所示,在所提出的option-based HDRL模型框架基础上,同时训练避障控制模型和行为选择模型,其具体步骤为: 首先利用行为选择使移动机器人产生有效的避障与目标接近两种行为; 然后将两种行为的经验数据均存入避障控制模型的经验池,其中目标接近行为的经验数据可用于优化避障策略; 在此基础上利用离策略学习方式对避障控制模型进行优化训练,能够使学习到的避障策略更加适用于导航任务. 具体实现见算法1和算法2.

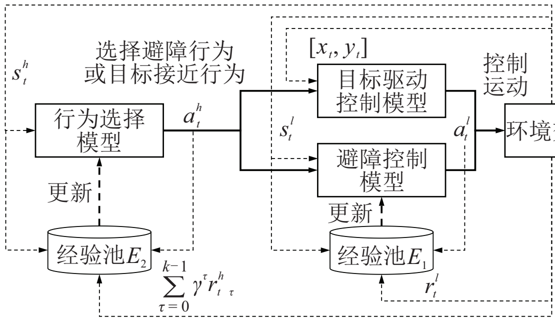


图4 训练方法

算法1 参数更新.

定义避障控制模型训练步数T_1和总训练步数T_2,清空避障控制模型经验池E_1和行为选择模型经验池E_2.

step 1: 机器人按照算法2(后文给出)执行一次动作;

step 2: 利用E_1中的历史数据更新避障控制模型,利用E_2中的历史数据更新行为选择模型;

step 3: 迭代step 1和step 2,共T_1次;

step 4: 机器人按照算法2执行一次动作;

step 5: 利用E_2中历史数据更新行为选择模型;

step 6: 迭代step 4和step 5,共T_2-T_1次.

算法2 环境交互.

step 1: 根据行为选择模型的输出选择行为;

step 2: 如果选择目标接近行为,根据目标驱动控制模型的输出执行1次动作,产生的目标接近经验数据则存入E_1,反之则根据避障控制模型的输出执行n

次动作,产生的避障经验数据存入 E_1;

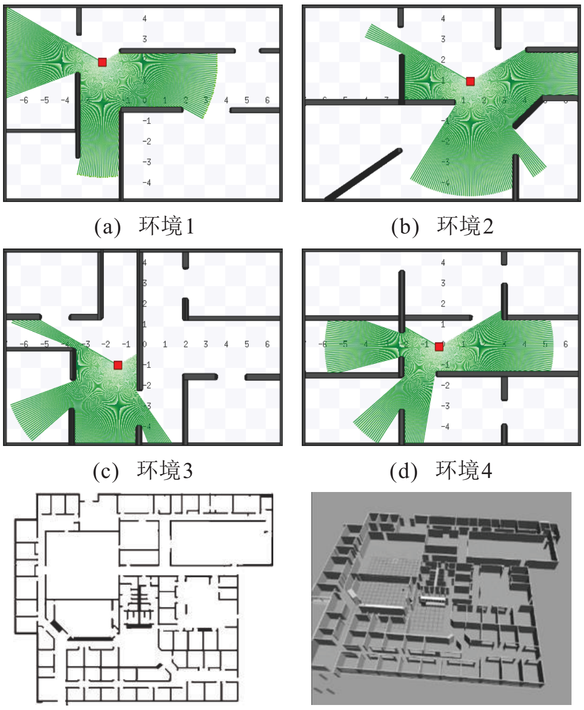
step 3: 产生的行为选择经验数据存入 E_2;

step 4: 迭代 step 1 ～ step 3, 直到回合终止.

2　实验分析

2.1　实验设置

本文采用 ROS　stage 模拟器进行移动机器人二维平面仿真实验, 并选择 Zhelo 等[8] 设计的 4 个 13.7 m×9.6 m 实验环境用于训练与测试, 其障碍布局如图 5(a) ～ (d) 所示. 图中正方形为移动机器人, 周围条纹为可视化的激光束. 此外, 采用一个更为复杂的 55.5 m×45.5 m 室内仿真环境[26](图 5(e)) 进行测试, 对应的三维模型如图 5(f) 所示. 实验中时间步设置为 0.2 s, 移动机器人运动方式采用二轮差速式, 每个时间步执行一次控制动作, 每回合最大移动步数为 480, 避障控制模型与目标驱动控制模型的控制动作具体设置如表 1 所示. 移动机器人顶部中心位置放置 2D 激光雷达传感器, 其扫描范围与距离测量范围分别设置为 [−120°, 120°]、[0.02 m, 5.6 m].



(a)　环境 1　(b)　环境 2　(c)　环境 3　(d)　环境 4　(e)　环境 5　(f)　环境 6

图 5　实验环境布局

表 1　控制动作设置

动作	$a^l_{(0)}$	$a^l_{(1)}$	$a^l_{(2)}$	$a^l_{(3)}$	$a^l_{(4)}$	$a^l_{(5)}$	$a^l_{(6)}$
线速度/(m/s)	0.2	0.4	0.4	0.4	0.4	0.4	0.2
角速度/(rad/s)	$\frac{\pi}{4}$	$\frac{\pi}{5}$	$\frac{\pi}{9}$	0	$-\frac{\pi}{9}$	$-\frac{\pi}{5}$	$-\frac{\pi}{4}$

文中模型超参数设置如表 2 所示, 同时对奖励函数中的权重系数和预定义参数进行设置: $\lambda^c = 0.03$, $\lambda^d = 3$, $\lambda^o = 0.4$, $D_{\text{thres}} = 0.475$, $D^o = 0.4$, $r_{\text{collision}} = -20$, $r_{\text{obstacle}} = 0.03$.

表 2　模型超参数设置

参数	取值
学习率 α	10^{-4}
折扣因子 γ	0.99
经验池大小	5×10^4
批量大小	32
避障控制模型训练步数 T_1	1.3×10^5
总训练步数 T_2	3.2×10^5

2.2　实验结果与分析

2.2.1　训练与测试实验结果

为评估所提出的导航方法, 选择目前主流的 DRL 算法 D3QN 以及基于 DRL 的分层导航方法 DAAC, 用于验证所提出模型框架可有效降低导航策略学习难度. 此外, 与未采用本文训练方式的 HDRL 方法进行对比. 上述对比方法的具体定义如下: 1) D3QN 方法: 采用 D3QN 直接学习导航策略; 2) DAAC (danger-aware adaptive composition) 方法: 采用 Zhang 等[11] 提出的基于动作优势加权的分层导航方法; 3) HDRL 方法: 采用所提出的 option-based HDRL 模型框架, 但在训练过程中不利用接近行为的经验数据对避障策略进行优化.

实验中每隔 4 000 步进行一次评估 (测试 10 回合) 并记录成功率, 该成功率表示移动机器人成功到达目标点的回合数占总测试回合数的比例, 最终得到 4 种方法的训练曲线如图 6 所示. 由图 6 可见, D3QN 方法和 DAAC 方法的训练曲线增长缓慢, 表明移动机器人在复杂环境下学习导航策略的难度较大. 而 HDRL 方法和本文方法的训练曲线在整个训练阶段稳定增长, 其中在训练初始时刻, 由于行为选择模型尚未学习到有效的行为选择策略, 机器人产生的接近行为经验数据不利于避障策略的学习, HDRL 方法的成功率稍高于本文方法. 随着训练的进行, 行为选择模型逐渐学习到有效的选择策略, 机器人产生的接近行为经验数据优化了避障策略, 从而使本文方法的导航成功率处于最高水平, 且明显优于其他方法. 上述结果表

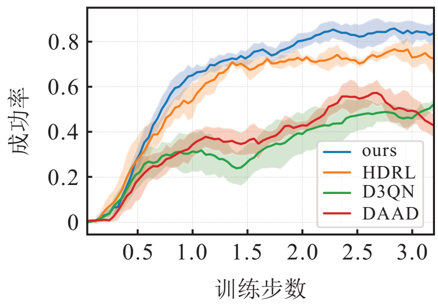


图 6　训练曲线

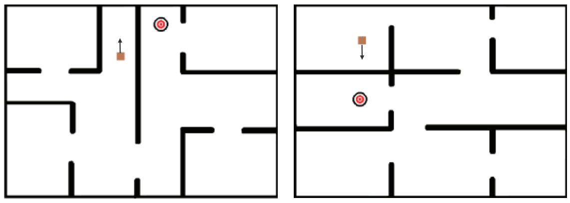
明,所提出的option-based HDRL可有效降低复杂环境下的导航策略学习难度,从而使移动机器人更好地学习导航策略,在此基础上通过优化避障策略的训练方法,机器人学习到更适用于导航任务的避障策略,进一步提升了性能.

此外,对上述方法在测试环境中分别进行300回合的测试,结果如表3所示,其中碰撞率和超时率分别表示发生碰撞回合数和超时回合数占总测试回合数的比例,步数表示在成功回合中机器人移动的步数. D3QN方法在测试环境2～环境4中的超时率显著高于其他方法,表明非分层导航方法在复杂环境下容易产生局部极小问题. 而DAAC方法依赖人为设计的调控规则难以得到较优的导航策略,因此在测试环境2～环境4中的碰撞率显著高于其他方法. 相比以上方法,两种HDRL方法均具有明显的优势,其中本文方法因进一步采用了优化避障策略的训练方法,在测试环境2～环境4中均取得了最高的成功率,并且在其他指标上也具有整体优势.而测试环境5相比于其他测试环境更为复杂,导致所有测试方法的整体性能均呈现一定程度的下降,但本文方法在成功率和超时率指标上仍取得了最优的结果,表明所提出的HDRL导航方法在复杂环境中能够有效地学习导航策略,同时对不同测试环境具有较好的泛化能力.

表3 多个环境下的导航测试结果

测试环境	测试方法	成功率(%)	碰撞率(%)	超时率(%)	移动步数(mean±std)
环境2	D3QN	42.67	1.33	56.00	244.5±72.3
	DAAC	49.33	35.00	15.67	236.8±63.3
	HDRL	81.67	18.00	0.33	227.0±63.3
	ours	83.33	16.67	0	217.8±45.7
环境3	D3QN	53.67	31.33	15.00	260.7±70.0
	DAAC	46.33	53.67	0	255.2±38.1
	HDRL	77.33	22.33	0.34	226.2±43.6
	ours	82.00	17.67	0.33	227.0±40.4
环境4	D3QN	40.33	7.33	52.34	247.2±65.8
	DAAC	41.33	41.00	17.67	219.7±87.6
	HDRL	65.33	34.67	0	219.3±46.3
	ours	72.67	24.33	3.00	190.0±29.7
环境5	D3QN	20.00	50.33	29.67	251.1±92.3
	DAAC	31.33	57.00	11.67	428.9±183.5
	HDRL	59.33	36.00	4.67	358.8±144.2
	ours	63.00	36.33	0.67	359.9±134.1

除上述测试外,为验证采用所提出方法的移动机器人在多种死角结构中均能较好地完成导航任务,分别选取环境3和环境4中的不同死角结构进行测试,相应的导航起始点、起始朝向及目标点如图7所示. 实验中采用4种方法在两种死角结构中分别测试300回合,得到不同指标的实验结果如表4所示. 由表4可见,本文方法在两种死角结构中均取得了最高的导航成功率,同时其他指标也优于对比方法,进一步验证了所提出方法在复杂环境中的有效性.



(a) 测试1 (b) 测试2

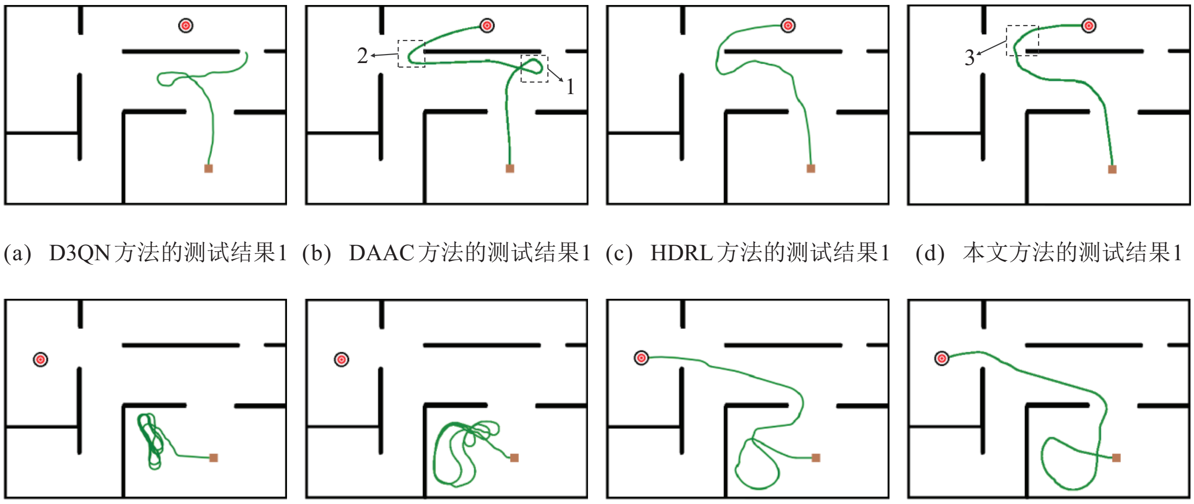
图7 针对不同死角结构的导航测试

表4 不同死角结构下的导航测试结果

测试任务	测试方法	成功率(%)	碰撞率(%)	超时率(%)	移动步数(mean±std)
测试1	D3QN	32.33	36.33	31.34	296.5±42.6
	DAAC	46.33	52.67	1.00	291.9±41.0
	HDRL	75.67	23.67	0.66	263.4±32.7
	ours	79.00	20.67	0.33	255.1±29.9
测试2	D3QN	25.00	5.33	69.67	319.0±52.4
	DAAC	39.67	47.33	13.00	315.5±46.2
	HDRL	62.00	36.67	1.33	298.5±45.9
	ours	68.67	30.33	1.00	287.3±38.9

2.2.2 可视化导航效果

为进一步展现行为选择模型的有效性,选择如图8所示的两个目标点分别采用上述4种方法进行测试,并绘制机器人的移动轨迹. 在测试过程中,所有导航方法的输入状态除激光雷达扫描数据外也包括目标点坐标. 如图8(a)所示,移动机器人初始朝向垂直向上,目标点处于环境最上方且被较长的障碍遮挡,采用D3QN方法进行导航的移动机器人在接近目标点时与障碍发生碰撞. 采用分层控制的DAAC方法的移动轨迹如图8(b)所示,机器人首先朝向目标点移动,遇到障碍后向右躲避但在位置“1”掉转方向. 其原因主要在于DAAC方法的高层策略受人为规则影响较大,机器人错误地采用了避障的策略,最后在位置“2”通过采用接近目标的策略从而成功到达目标点. 采用HDRL方法和本文方法的导航结果分别如图8(c)和图8(d)所示,移动机器人均可成功到达目标



(a) D3QN 方法的测试结果1　(b) DAAC 方法的测试结果1　(c) HDRL 方法的测试结果1　(d) 本文方法的测试结果1

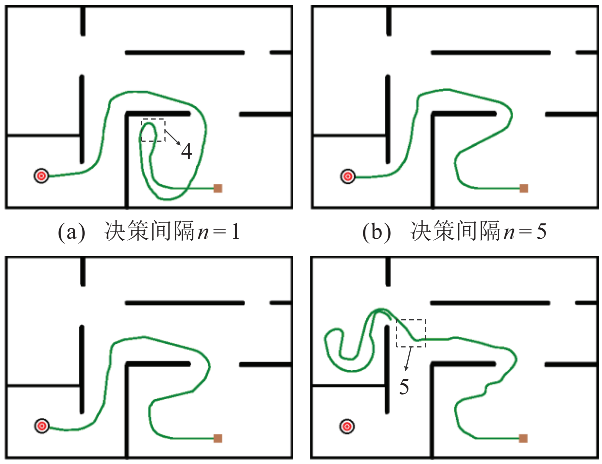
(e) D3QN 方法的测试结果2　(f) DAAC 方法的测试结果2　(g) HDRL 方法的测试结果2　(h) 本文方法的测试结果2

图 8　采用不同方法的导航结果展示

点, 其中本文方法的移动轨迹相对较优: 移动机器人首先通过选择目标接近行为朝向目标点移动, 走出凹形区域后为躲避障碍连续选择避障行为, 随后在位置 "3" 越过障碍后再次选择目标接近行为, 最终顺利到达目标点.

在图 8(e) 所示的另一次测试中, 移动机器人初始朝向水平向左, 目标点处于环境左上方, 采用 D3QN 方法进行导航时出现局部极小问题, 导致移动机器人无法走出凹形区域, 类似问题也出现在 DAAC 方法的导航结果中 (图 8(f)). 相较于 HDRL 方法 (图 8(g)), 采用本文方法的导航结果具有较大优势: 在行为选择的作用下, 机器人在导航起始阶段优先选择目标接近行为, 朝向目标点移动. 在遇到障碍后机器人选择避障行为以躲避障碍, 并持续选择避障行为成功走出凹形区域 (图 8(h)). 通过对两种低层行为策略的调控, 机器人最终到达目标点完成导航任务. 由以上测试结果可知, 所提出的行为选择模型能够学习稳定、可靠的行为选择策略, 从而完成复杂环境下的导航任务.

此外, 本文研究了行为选择模型的决策间隔 n 对导航的影响. 分别在 n 取值为 1、5、10 和 15 四种情况下进行训练, 并绘制最终移动轨迹, 如图 9 所示. 当决策间隔 $n = 1$ 时, 由于决策间隔较小导致决策链过长, 移动机器人在位置 "4" 未能合理选择避障行为导致导航效果欠佳. 在决策间隔大小适中的情况下 ($n = 5$、$n = 10$), 移动机器人均以较优的路径完成了导航, 产生的轨迹分别如图 9(b) 和图 9(c) 所示. 图 9(d) 中的移动机器人在到达位置 "5" 后, 由于决策间隔较大 ($n = 15$) 未能及时选择目标接近行为, 从而错误地进入了另一凹形区域, 最终导致未在规定步数内完成导航任务. 因此, 设置适当大小的决策间隔 n 有利于对低层行为进行合理地调控.



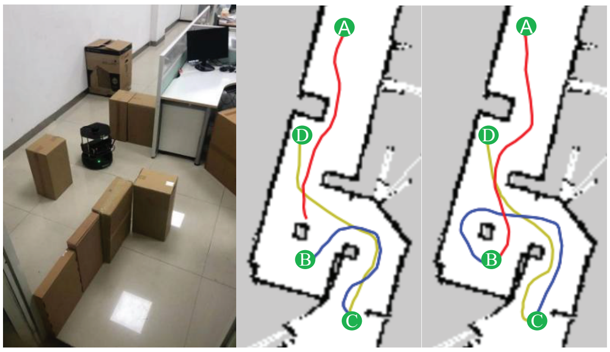
(a) 决策间隔 $n = 1$　(b) 决策间隔 $n = 5$

(c) 决策间隔 $n = 10$　(d) 决策间隔 $n = 15$

图 9　决策间隔对导航的影响

2.2.3　真实环境下的导航实验

为验证所提出方法的实际应用价值, 本文在真实环境下对仿真环境中训练的模型进行测试, 采用 Turtlebot 2 移动机器人作为实验对象. 该机器人装配 RPLIDAR 2D 激光雷达传感器, 并搭载英伟达 TX2 计算平台. 实验中移动机器人的速度信息直接从里程计中读取, 目标点与移动机器人的实时相对坐标通过 ROS 中的 AMCL 功能包获取. 实验环境为如图 10(a) 所示的室内场景, 测试中采用 DAAC 方法的导航结果如图 10(b) 所示. 移动机器人从 "A" 出发, 在接近 "B" 时与障碍发生碰撞, 但随后成功通过 "C" 并最终到达 "D". 采用本文方法的移动机器人从 "A" 出发依次顺利通过 "B" 和 "C", 并最终成功到达 "D" (图 10(c)), 这表明所提出的导航方法能够较好地应用于真实环境下的导航任务.



(a) 真实环境 (b) DAAC方法 (c) 本文方法

图 10 真实环境下的导航测试

3 结 论

为解决未知复杂环境下的移动机器人导航问题,本文提出了一种学习分层控制策略的option-based HDRL导航方法,其创新之处在于:1)针对导航策略学习困难的问题,采用基于option-based HDRL的框架,并分别设计用于实现高层和低层策略的模型,其中避障与目标驱动控制模型分别用于实现避障与目标接近两种行为策略,行为选择模型负责学习高层行为选择策略,通过对上述两种行为进行选择完成导航任务;2)基于上述模型框架,在训练阶段进一步利用接近行为的经验数据更新避障控制模型,从而有效优化避障策略. 仿真对比实验结果表明,所提出方法可有效解决复杂环境下导航效果不佳的问题,且具有较好的泛化能力. 此外,真实环境下的导航测试初步验证了所提出方法的潜在应用价值.

与依据地图进行路径规划的传统导航方法相比,本文方法虽无法确保最优的导航路径,但避免了对环境地图和专家经验的依赖,同时具有较强的自适应能力,因此可适用于无环境地图的室内导航任务. 文中行为选择模型的决策间隔n采用了固定值设置,在未来工作中将通过进一步改进行为选择模型,使机器人在与环境交互中自动学习n的合理取值. 此外,本文在解决复杂环境中的导航问题时仅考虑了静态障碍的情况,如何在包含动态障碍的环境下学习较优的导航策略是未来的一个研究重点.

参考文献(References)

[1] Jiang H G, Wang H, Yau W Y, et al. A brief survey: Deep reinforcement learning in mobile robot navigation[C]. The 15th IEEE Conference on Industrial Electronics and Applications. Kristiansand, 2020: 592-597.

[2] Quan H, Li Y S, Zhang Y. A novel mobile robot navigation method based on deep reinforcement learning[J]. International Journal of Advanced Robotic Systems, 2020, 17(3): 172988142092167.

[3] 孙辉辉, 胡春鹤, 张军国. 移动机器人运动规划中的深度强化学习方法[J]. 控制与决策, 2021, 36(6): 1281-1292.
(Sun H H, Hu C H, Zhang J G. Deep reinforcement learning for motion planning of mobile robots[J]. Control and Decision, 2021, 36(6): 1281-1292.)

[4] Fan T X, Long P X, Liu W X, et al. Distributed multi-robot collision avoidance via deep reinforcement learning for navigation in complex scenarios[J]. The International Journal of Robotics Research, 2020, 39(7): 856-892.

[5] Xiao X S, Liu B, Warnell G, et al. Motion planning and control for mobile robot navigation using machine learning: A survey[J/OL]. 2020, arXiv: 2011.13112.

[6] 董豪, 杨静, 李少波, 等. 基于深度强化学习的机器人运动控制研究进展[J]. 控制与决策, 2022, 37(2): 278-292.
(Dong H, Yang J, Li S B, et al. Research progress of robot motion control based on deep reinforcement learning[J]. Control and Decision, 2022, 37(2): 278-292.)

[7] Tai L, Paolo G, Liu M. Virtual-to-real deep reinforcement learning: Continuous control of mobile robots for mapless navigation[C]. The IEEE/RSJ International Conference on Intelligent Robots and Systems. Vancouver, 2017: 31-36.

[8] Zhelo O, Zhang J W, Tai L, et al. Curiosity-driven exploration for mapless navigation with deep reinforcement learning[J/OL]. 2018, arXiv: 1804.00456.

[9] Zhang W, Zhang Y F. Behavior switch for DRL-based robot navigation[C]. IEEE 15th International Conference on Control and Automation. Edinburgh, 2019: 284-288.

[10] 杨瑞, 严江鹏, 李秀. 强化学习稀疏奖励算法研究——理论与实验[J]. 智能系统学报, 2020, 15(5): 888-899.
(Yang R, Yan J P, Li X. Survey of sparse reward algorithms in reinforcement learning — theory and experiment[J]. CAAI Transactions on Intelligent Systems, 2020, 15(5): 888-899.)

[11] Zhang W, Zhang Y F, Liu N. Danger-aware adaptive composition of DRL agents for self-navigation[J]. Unmanned Systems, 2021, 9(1): 1-9.

[12] Ding W H, Li S J, Qian H H, et al. Hierarchical reinforcement learning framework towards multi-agent navigation[C]. IEEE International Conference on Robotics and Biomimetics. Kuala Lumpur, 2018: 237-242.

[13] 陈春林, 陈宗海, 卓睿, 等. 基于分层式强化学习的移动机器人导航控制[J]. 南京航空航天大学学报, 2006, 38(1): 70-75.
(Chen C L, Chen Z H, Zhuo R, et al. Mobile robot navigation control based on hierarchical reinforcement

learning[J]. Journal of Nanjing University of Aeronautics & Astronautics, 2006, 38(1): 70-75.)

[14] Jun H W, Kim H J, Lee B H. Goal-driven navigation for non-holonomic multi-robot system by learning collision[C]. International Conference on Robotics and Automation. Montreal, 2019: 1758-1764.

[15] 周文吉, 俞扬. 分层强化学习综述[J]. 智能系统学报, 2017, 12(5): 590-594.
(Zhou W J, Yu Y. Summarize of hierarchical reinforcement learning[J]. CAAI Transactions on Intelligent Systems, 2017, 12(5): 590-594.)

[16] Li S Y, Wang R, Tang M X, et al. Hierarchical reinforcement learning with advantage-based auxiliary rewards[J]. Advances in Neural Information Processing Systems. Vancouver: MIT Press, 2019: 1409-1419.

[17] Zhang Y, Mou Z Y, Gao F F, et al. Hierarchical deep reinforcement learning for backscattering data collection with multiple UAVs[J]. IEEE Internet of Things Journal, 2021, 8(5): 3786-3800.

[18] Li C S, Xia F, Martin-Martin R, et al. HRL4IN: Hierarchical reinforcement learning for interactive navigation with mobile manipulators[J/OL]. 2019, arXiv: 1910.11432.

[19] Strudel R, Pashevich A, Kalevatykh I, et al. Learning to combine primitive skills: A step towards versatile robotic manipulation[C]. Proceedings of the IEEE International Conference on Robotics and Automation. Paris, 2020: 4637-4643.

[20] Li T Y, Lambert N, Calandra R, et al. Learning generalizable locomotion skills with hierarchical reinforcement learning[C]. IEEE International Conference on Robotics and Automation. Paris, 2020: 413-419.

[21] Tessler C, Givony S, Zahavy T, et al. A deep hierarchical approach to lifelong learning in minecraft[J]. CoRR, 2016, abs/1604.07255.

[22] 刘建伟, 高峰, 罗雄麟. 基于值函数和策略梯度的深度强化学习综述[J]. 计算机学报, 2019, 42(6): 1406-1438.
(Liu J W, Gao F, Luo X L. Survey of deep reinforcement learning based on value function and policy gradient[J]. Chinese Journal of Computers, 2019, 42(6): 1406-1438.)

[23] Xie L H, Wang S, Rosa S, et al. Learning with training wheels: Speeding up training with a simple controller for deep reinforcement learning[C]. IEEE International Conference on Robotics and Automation. Brisbane, 2018: 6276-6283.

[24] Sutton R S, Precup D, Singh S. Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning[J]. Artificial Intelligence, 1999, 112(1/2): 181-211.

[25] Zhao N, Liang Y C, Niyato D, et al. Deep reinforcement learning for user association and resource allocation in heterogeneous cellular networks[J]. IEEE Transactions on Wireless Communications, 2019, 18(11): 5141-5152.

[26] Le X S, Fabresse L, Bouraqadi N, et al. Evaluation of out-of-the-box ROS 2D SLAMs for autonomous exploration of unknown indoor environments[C]. Proceedings of the International Conference on Intelligent Robotics and Applications. Newcastle: Springer, 2018: 283-296.

作者简介

王童 (1997–), 男, 硕士生, 从事深度强化学习与移动机器人导航控制的研究, E-mail: wt18ustc@mail.ustc.edu.cn;

李骜 (1977–), 男, 副教授, 博士生导师, 从事图像处理、智能感知与人机交互等研究, E-mail: aoli@ustc.edu.cn;

宋海荦 (1997–), 男, 硕士生, 从事深度强化学习与移动机器人避障控制的研究, E-mail: shl97@mail.ustc.edu.cn;

刘伟 (1995–), 男, 硕士生, 从事深度强化学习与机器人控制的研究, E-mail: zkd2021@mail.ustc.edu.cn;

王明会 (1982–), 女, 副教授, 博士, 从事人工智能与医学信息处理等研究, E-mail: mhwang@ustc.edu.cn.

(责任编辑：郑晓蕾)