计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2022,31(2):88−95 [doi: 10.15888/j.cnki.csa.008302]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于多智能体强化学习的无人机群室内辅助救援①

郭天昊, 张　钢, 岳文渊, 王　倩, 郭大波

(山西大学 物理电子工程学院, 太原 030006)
通信作者: 郭天昊, E-mail: tianhao_guo@sxu.edu.cn

摘　要: 本文主要研究了在室内场景中使用多台无人机设备对受害者进行合作搜索的问题. 在室内场景中, 依赖全球定位系统获取受害者位置信息可能是不可靠的. 为此, 本文提出一种基于多智能体强化学习 (MARL) 方案, 该方案着重对无人机团队辅助救援时的路径规划问题进行研究. 相比于传统方案, 所提方案在大型室内救援场景中更具优势, 例如部署多台救援无人机、救援多位受害者. 本方案也考虑了无人机的充电问题, 保证无人机的电量始终充足. 具体地, 鉴于模型中的救援场景深度参数不断变化, 所提方案将搜索路径规划问题模拟为部分可观的马尔可夫决策过程 (Dec-POMDP), 为使得对无人机控制策略最优, 本文又训练了一个双深度的 Q 网络架构 (Double DQN). 最后使用蒙特卡罗方法验证了本方案在大型室内环境中能够使多台无人机有效合作, 且能最大化搜集受害者所用手机内部所存储的位置信息.

关键词: 无人机; 室内救援; 路径规划; 马尔可夫决策; 蒙特卡洛

引用格式: 郭天昊,张钢,岳文渊,王倩,郭大波.基于多智能体强化学习的无人机群室内辅助救援.计算机系统应用,2022,31(2):88–95. http://www.c-s-a.org.cn/1003-3254/8302.html

Indoor Assisted Rescue by UAV Group Based on Multi-agent Reinforcement Learning

GUO Tian-Hao, ZHANG Gang, YUE Wen-Yuan, WANG Qian, GUO Da-Bo

(College of Physics and Electronic Engineering, Shanxi University, Taiyuan 030006, China)

Abstract: This work mainly studies the problem of using multiple unmanned aerial vehicles (UAVs) to search for victims cooperatively in indoor scenes where the location information of victims relying on the global positioning system may be unreliable. To this end, this study proposes a multi-agent reinforcement learning (MARL) based solution which focuses on the path planning studies when the UAV team assists the rescue. Compared with the traditional solution, the proposed solution has advantages in large-scale indoor rescue scenes, such as deploying multiple rescue UAVs and rescuing multiple victims. At the same time, this solution also considers the charging problem of the UAVs to ensure that the power of the UAVs is always sufficient. Specifically, due to the continuous changes of the rescue scene depth parameters in the model, the proposed solution simulates the path planning as a decentralized partially observable Markov decision process (Dec-POMDP). To optimize the UAV control strategy, this study also trains a double deep Q-learning network (Double DQN). Finally, the Monte Carlo method is used to verify that this solution can effectively cooperate with multiple UAVs in a large-scale indoor environment and maximize the collection of the location information stored in the mobile phone used by the victim.

Key words: unmanned aerial vehicle (UAV); indoor rescue; path planning; Markov decision; Monte Carlo

① 基金项目: 山西省基础研究项目 (201801D121118)
郭天昊和张钢为共同第一作者.
收稿时间: 2021-04-14; 修改时间: 2021-05-11; 采用时间: 2021-05-19; csa 在线出版时间: 2022-01-17

无人机具有易于部署, 灵活性高以及制造成本低等优势, 常被用来部署在各种民用场景中, 其中包括精准农业, 清理海洋废弃物, 包裹投递, 自然灾害后恢复网络服务以及搜索与搜救[1−5]. 基于无人机的灾后搜救也由之前的单机式搜救逐渐发展到小型的多机群式搜救[6], 大幅度提高了搜救效率.

在早期, 无人机辅助搜救的目标多是针对目标位置已知的情况[7−9], 无人机只需要规划飞行路线抵达目标所在位置即可. 实际搜救过程中在搜索目标之前对于任务区域的信息知之甚少, 复杂的室内环境对于无人机的避障能力提出了巨大的挑战. 如何使无人机群对环境无任何先验知识的前提下进行自主决策是个值得研究的问题. 为此, 文献 [10] 将基于模型的强化学习算法应用到无人机的自主导航领域, 极大提高了无人机自主决策能力. 在无人机最佳路径规划方向, 文献 [11] 考虑了奖励与惩罚机制, 使得无人机在不断地尝试飞行中选择最佳路径从而到达目标位置. 考虑到无人机在复杂环境中状态值爆炸式的增多, 为解决大容量的状态值问题, 文献 [12] 提出了基于神经网络的分布式 DQN 算法, 以控制无人机在未知环境中进行目标搜索与目标跟踪. 针对大规模的搜索环境, 需要多台无人机设备进行协作完成搜索任务, 文献 [13] 提出一种异构多智能体算法, 以控制多台无人机在复杂的环境中以协作的方式最大程度上完成搜索任务.

为组织合作的、智能的、适应复杂环境下的无人机群, 本文提出了一种基于多智能体强化学习的控制策略. 首先将多无人机搜救任务进行建模处理, 将其转化为具有完全回报函数的分散的部分可观的马尔可夫决策过程; 其次提出了基于集中学习分散执行的多智能体强化学习方法, 利用了 Double-DQN 算法对 Dec-POMDP 进行了求解; 最后利用蒙特卡洛方法对本方案进行通用性测试. 结果表明本文方案在搜救成功率方面所具备的优势, 能够在大型的救援环境中出色的完成搜救任务.

1 问题概述与系统模型

1.1 问题概述

本文考虑了这样一个场景, 某大型图书馆发生火灾, 无人机群进入图书馆以协作的方式迅速对馆内受害人员进行搜索. 具体地, 无人机群的搜索任务环境如图 1 所示. 该图模拟了复杂的室内环境, 为简化训练模型, 假设无人机飞行过程中高度恒定为 h (不包括起飞与降落操作), 飞行速度恒定为 V. 受害者的手机可作为其位置信息传感器随时发出其位置信息, 其通信模式为: 无人机通过发射一种激活信号, 受害者持有的终端设备接受此信号后通过反向散射 (back scatter) 方式向无人机发送位置数据[14], 由于每个手机设备反射功率的不同会导致无人机只有在一定范围内才能采集位置信息.

图 1 无人机室内搜救模型

任务区域内随机布置了 3 台无人机与 7 名位置随机的受害者. 无人机从停机坪 (充电站位置) 出发后, 分别对受害者位置信息进行采集, 受限于电池容量的影响, 无人机群须在电量耗尽之前返航到充电站位置进行充电操作, 对于没能在电量耗尽之前安全返航进行充电操作的所提方案将其作为惩罚加入奖励函数中.

为将地图进行合理的建模, 本方案将地图分割为 $N\times N$ 的网格模型, 定义整个室内区域为 M: $M\times M\in N^2$, 将 M 区域分割为尺寸大小为 n 的 32×32 网格. 定义无人充电位置区域 $\mathbb{Z}\in\left\{[x_i^{\mathbb{Z}},y_i^{\mathbb{Z}}]^{\mathrm{T}},i=1,\cdots,N\right\}$, 将所有窗户位置定义为禁飞区域 $\mathbb{C}\in\{[x_i^{\mathbb{C}},y_i^{\mathbb{C}}],i=1,\cdots,N\}$, 避免由于窗户处于打开状态而导致无人机从窗户飞到室外无形中将搜索区域扩展为无限大. 将墙体与门所在位置定义为障碍物区域 $\mathbb{N}=\{[x_i^{\mathbb{N}},y_i^{\mathbb{N}}],i=1,\cdots,N\}$, 无人机在飞行过程中要绝对避免的区域.

1.2 信道模型

在本文中, 所提方案考虑到现实场景, 将无人机与受害者之间的通信链路建模为视距 (line of sight, *LOS*) 与非视距 (none-line of sight, *NLOS*) 的点对点信道模型[15], 在该信道模型下本文定义在时间为 n, 第 j 位受害者能

够达到的信息速率为:

$$R_j(n)=\log_2\left(1+\frac{P_j}{\sigma^2}\cdot d_j(n)^{-\alpha_e\cdot 10^{\eta_e/10}}\right) \tag{1}$$

其中, P_j为发射功率, σ^2为接受机处的高斯白噪声的功率, α_e与η_e为信号在无人机与受害者之间的传播路径损耗指数, 其中$e\in\{LOS,NLOS\}$, 具体与环境有关, 主要来自视距损耗或者非视距损耗. $d_j(n)$为无人机距离受害者的直线距离.

1.3 系统模型

本系统模型的主要研究目标是使得无人机团队在一定约束条件下能够最大化的从搜救区域中搜集受害者的位置信息. 这些约束条件主要分为两部分, 一部分来自无人机设备自身的条件约束, 例如电池容量. 另一部分来自环境对无人机设备的约束. 例如室内门墙等障碍物、无人机与无人机之间避免碰撞以及无人机的起落位置 (充电位置).

对无人机群进行建模, 定义第i台无人机的位置为$p_i(t)=[x_i(t),y_i(t),z_i(t)]^{\mathrm{T}}$, 为简化模型假设无人机的飞行高度为$z_i(t)\in\{0,h\}$即无人机的高度位置只能为地面 0 或者以$h$的高度恒定飞行. 在本模型中无人机的动作空间受限于其所处环境的位置, 无人机只有在充电区域时才能执行着陆充电动作. 动作空间定义为:

$$A(p_i(t))=\begin{cases}\{hover,east,north,west,south,charging\}, p_i(t)\in\mathbb{Z}\\ [0,0,-h]/\{hover,east,north,west,south,charging\}, \text{otherwise}\end{cases} \tag{2}$$

$$\begin{cases}hover=[0,0,0]^{\mathrm{T}}\\ east=[c,0,0]^{\mathrm{T}}\\ north=[0,c,0]^{\mathrm{T}}\\ west=[-c,0,0]^{\mathrm{T}}\\ south=[0,-c,0]^{\mathrm{T}}\\ charging=[0,0,-h]^{\mathrm{T}}\end{cases} \tag{3}$$

第i台无人机在t时刻的动作为$a_i(t)$, 其中, $a_i(t)\in A(p_i(t))$.

无人机t时刻的飞行状态为$\lambda_i(t)\in\{0,1\}$, 0 表示无人机处于静止状态, 1 表示无人机处于运动状态. 值得注意的是当无人机处于充电状态时, 运动状态变为非运动状态. 定义无人机的下一状态为:

$$\lambda_i(t+1)\in\begin{cases}0, a_i(t)=[0,0,-h]^{\mathrm{T}}\vee\lambda_i(t)=0\\ 1, \text{otherwise}\end{cases} \tag{4}$$

对于本模型中的无人机剩余电量, 第i台无人机在t时刻剩余电量为$E_i(t)$. 假设无人机运动时消耗的电量始终恒定, 此时可以直接将剩余电量等价于此无人机的剩余飞行时间. 由于耗电量主要由运行状态决定, 为此无人机的下一时刻的剩余电量可离散化为:

$$E_i(t+1)=\begin{cases}E_i(t)-1, \lambda_i(t)=1\\ E_i(t), \text{otherwise}\end{cases} \tag{5}$$

对于无人机的动作策略的优化本质上是对无人机的接收信息的吞吐量最大化, 无人机群与受害者持有的设备之间通信时遵循标准的时分多址模型 (TDMA), 在任务时间T内, 整个团队I台无人机设备在联合策略$x_ia_i(t)$下的最大化吞吐量模型为:

$$\max_{x_ia_i(t)}\sum_{t=0}^{T}\sum_{i=1}^{I}B_i(t) \tag{6}$$

$$B_i(t)=\lambda_i(t)\sum_{n=\beta t}^{\beta(t+1)-1}\sum_{j=1}^{J}q_{i,j}(n)R_{i,j}(n) \tag{7}$$

其中, $n\in[\beta t,\beta(t+1)-1]$表示在任务时间$T$内的通信时间, β为通信时隙, $\lambda_i(t)$为无人机的在t时刻的运行状态, $q_{i,j}(n)\in\{0,1\}$为 TDMA 模式下的调度变量.

2 多智能体强化学习

强化学习作为机器学习的一个重要分支, 其核心是智能体与特定环境的重复交互, 学习如何在未知环境中执行最优策略[16]. 在当前状态S下, 智能体执行动作A, 环境接收到此动作后反馈于智能体下一状态S_{t+1}和回报值R, 智能体依据环境反馈的回报值来优化策略并且进行再学习, 就这样通过不断地迭代最后生成最优策略.

本文针对多无人机协作问题定义了同质的、非通信的、简单合作的无人机群. 同质性是指每台无人机设备具有相同的构造结构, 相同的动作空间以及任务领域; 非通信是指无人机之间没有直接的通信, 即无人机不能协调它们的动作以及进行有关的通信, 但都能感知与其他无人机之间位置信息, 并可利用这种感知到的位置信息对无人机进行监管维护; 简单合作是指无人机团队共同收集的位置数据可添加到每一台无人机的回报函数中, 这就使得他们有一个共同的目标. 在多智能体训练阶段, 每台无人机与环境不断地交互进行自身的策略优化, 之后将它们的经验数据集中起来, 通过组建神经网络数据库来训练控制系统, 最后将训练好的控制系统部署到每台无人机设备上.

2.1 部分可观的马尔可夫决策过程

本节将无人机团队与环境交互的问题转换为部分可观的马尔可夫决策过程. 一个 Dec-MDP 通常是由一个七元组组成$(S,A_x,P,R,\Omega_x,O,\gamma)$, 其中$S$代表一组空间状态值, A_x表示智能体的动作空间, $P:S\times A\to\Delta(S)$为状态转移概率矩阵, $R:S\times A\times S\to\mathbb{R}$为即时的奖励函数, $\Omega_x=\Omega^I$为一组观测结果, 即无人机传感器获得的环境数据. O为条件观测概率. $\gamma\in[0,1]$为贴现因子, 表示长期回报与短期回报的重要程度.

状态 (S): 本模型状态空间由 3 部分组成. 分别为环境信息、无人机状态、受害者状态. 定义状态空间为$S(t)=(M,\{p_i(t)\},\{E_i(t)\},\lambda_i(t),\{L_j\},\{D_j(t)\})$, 其中, M 表示环境中的一系列位置信息集合, 包括室内门窗以及墙体等障碍物的位置、无人机的起落位置以及危险区域的位置信息. $\{p_i(t)\}$表示第i台无人机的位置信息, $\{E_i(t)\}$表示第i台无人机的剩余电量, $\lambda_i(t)$表示第i台无人机的运行状态, $\{L_j\}$表示第j位受害者的位置信息, $\{D_j(t)\}$表示第j位受害者的位置信息量.

动作 (A_x): 考虑到模型的复杂性, 为避免无人机发生碰撞, 引入符合本模型的动作状态空间, 在之前的无人机动作空间模型中引入一种安全控制机制, 无人机通过判断所处的环境位置来执行相应的动作, 当无人机位于区域$\mathbb{R}$时, 该区域包括无人机的下个位置位于障碍物区域以及两台无人机同时出现在同一种位置且处于运动状态, 此时安全机制执行悬停动作. 本方案定义安全动作空间如下:

$$\overset{\triangle}{a}_i(t)=\begin{cases}[0,0,0]^{\mathrm{T}},\mathbb{R}\\ a_i(t),\text{otherwise}\end{cases}\tag{8}$$

在区域$\mathbb{R}$, 有:

$$\begin{cases}p_i(t)+a_i(t)\in\{\mathbb{C},\mathbb{N}\}\\ \vee p_i(t)+a_i(t)=p_k(t)\wedge\lambda_k(t)=1\\ \forall k,k\neq i\end{cases}\tag{9}$$

即时奖励 (R): 本模型总即时奖励$R_i(t)$由所有无人机任务时间内搜救受害者数量的奖励、路径规划奖励、即时充电奖励 3 部分组成, 其中搜集所有受害者位置信息量奖励是每台无人机共享奖励的唯一部分.

t时间内收集的所有受害者的位置信息量作为集体奖励, τ为收集乘数将数据收集参数化, 方法如下:

$$\psi_i(t)=\tau\sum_{j=1}^{J}(D_j(t+1)-D_j(t))\tag{10}$$

路径规划奖励主要用于惩罚无人机在飞行过程中不执行安全动作空间, 为诱导无人机优化最短飞行路径来搜集受害者的位置信息, 定义方法如下:

$$\theta_i(t)=\begin{cases}\theta+\delta,a_i(t)\neq\overset{\triangle}{a}_i(t)\\ \delta,\text{otherwise}\end{cases}\tag{11}$$

其中, 当无人机的动作空间不符合安全动作空间时, 给出惩罚θ, δ为持续飞行的路径惩罚, 使无人机减少飞行时间优化其搜救路径.

即时充电奖励强迫无人机在电量耗尽之前返航进行充电, 对于没有即时返航的无人机给予值为ω的处罚, 为此定义即时充电的奖励计算如下:

$$\omega_i(t)=\begin{cases}\omega,\ E_i(t+1)=0\wedge p_i(t+1)=[\bullet,\bullet,h]^{\mathrm{T}}\\ 0,\text{otherwise}\end{cases}\tag{12}$$

综上 3 部分, 总即时奖励为以上 3 部分奖励之和, 即:

$$R_i(t)=\psi_i(t)+\theta_i(t)+\omega_i(t)\tag{13}$$

观测结果 (Ω_x): 本模型无人机的观测空间由室内环境M、无人机的飞行状态$O_{\lambda_i(t)}$、无人机的剩余电量$O_{E_i(t)}$、受害者位置信息量$O_{D_j(t)}$四部分组成. 为简化模型, 将无人机的位置与受害者位置进行 2D 投影, 投影函数定义如下:

$$f_{2\mathrm{D}}(x)=\begin{bmatrix}1/c & 0 & 0\\ 0 & 1/c & 0\end{bmatrix}x\tag{14}$$

式 (15) 与式 (16) 分别将受害者与无人机的位置经过投影函数$f_{2\mathrm{D}}(x)$后所得 2D 位置信息为:

$$\widehat{L}_j(t)=f_{2\mathrm{D}}(L_j(t))\tag{15}$$

$$\widehat{p}_i(t)=f_{2\mathrm{D}}(p_i(t))\tag{16}$$

带有位置信息P_{location}与该位置所对应的某一类值$Q_{\text{information}}$映射到相对应的图层O, 定义如下映射:

$$f_{\text{observation}}:P_{\text{location}}\times Q_{\text{information}}\to O\tag{17}$$

式 (18), 式 (19), 式 (20) 分别将无人机的位置信息与该位置对应的运动状态和剩余电量在映射函数$f_{\text{observation}}$下, 得到其对应得图层O.

$$O_{\lambda_i(t)}=f_{\text{observation}}(\{\widehat{p}_i(t)\},\{\lambda_i(t)\})\tag{18}$$

$$O_{E_i(t)}=f_{\text{observation}}(\{\widehat{p}_i(t)\},\{E_i(t)\})\tag{19}$$

$$O_{D_j(t)}=f_{\text{observation}}(\{\widehat{L}_j(t)\},\{D_j(t)\})\tag{20}$$

将上述所有的图层作为无人机的观测结果输入到

本文神经网络架构中, 式 (21) 为无人机的观测结果:

$$O_i(t)=(M_i(t),O_{\lambda_i(t)},O_{E_i(t)},O_{D_j(t)}) \tag{21}$$

2.2 DQN 算法:

本文的无人机群在大型室内空间中执行任务, 如将 Q 值表示为数值表格是不可取的, 文献 [17] 应用了经验回放等方面的技术, 将 Q-Learning 与神经网络结合起来, 完美的解决了大状态空间问题. 经验池的主要功能是解决相关性与非静态分布问题, 具体方法是通过将每个时刻智能体与环境交互的样本 (s_t,a_t,r_t,s_{t+1}) 存储于回放记忆单元, 训练时在经验池中随机抽取一批数据进行训练.

DQN 算法的主要目标是保证估计值网络输出的值$Q(s,a;\theta)$无限接近于目标值网络输出的目标值$TargetQ$, 其中, θ为网络参数, 目标值计算方法为:

$$TargetQ=r+\gamma\max_{a'}Q(s',a';\theta) \tag{22}$$

基于上述的目标值, 定义 DQN 的损失函数公式, 求$L(\theta)$关于θ的梯度, 使用随机梯度下降算法更新网络参数θ:

$$L(\theta)=E[(TargetQ-Q(s,a;\theta))^2] \tag{23}$$

具体地, DQN 算法的流程如图 2 所示.

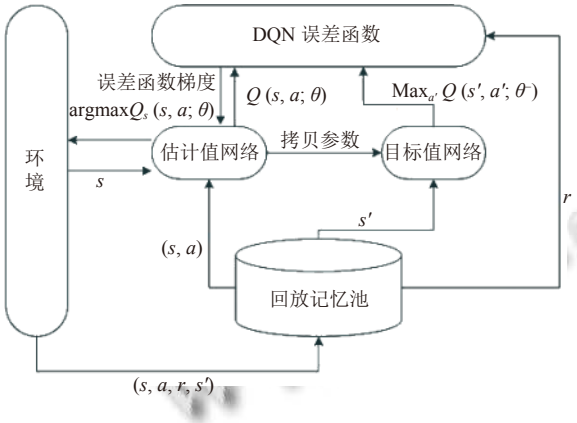

图 2 DQN 算法流程图

3 无人机团队路径规划算法

3.1 Double DQN (DDQN) 算法

本文利用 DDQN 算法对无人机团队路径规划进行训练, 不同于 DQN 算法, DDQN 算法克服了 Q-Learning 算法固有的缺陷即过估计问题, 而在 DQN 算法中此问题也没有得到有效解决, 为解决过估计问题 Double DQN 算法将动作的选择和动作的评估分别用不同的值函数来实现如图 3 所示, 具体如下:

通过网络 (main-net) 获得最大值函数的动作 a, 然后通过目标网络 (target-net) 获得上述动作 a 所对应的 $TargetQ$值:

$$TargetQ=r+\gamma Q(s',\text{argmax}_a(Q_{\text{main}}(s',a))) \tag{24}$$

基于目标网络$TargetQ$值, 定义 DDQN 的损失函数为:

$$L(\theta)=E[(TargetQ-Q_{\text{main}}(s,a;\theta))^2] \tag{25}$$

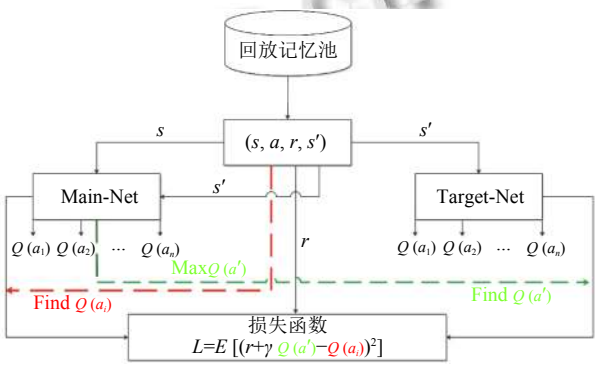

图 3 DDQN 损失函数构造流程

3.2 Double DQN 神经网络

如图 4 所示, 本文构建了该神经网络架构. 将上述处理的 4 幅信息图层堆叠起来组成状态信息图作为观测值传入卷积核为 5×5 的卷积层 1 和 2, 然后通过激活函数ReLU (线性整流函数) 将输出传入展平层, 该层的目的是将多维的数据一维化. 最后将平坦后的数据与无人机剩余飞行时间的标量连接起来经过 3 个隐藏层后通过激活函数ReLU将数据传入到一个动作空间大小为 6 的全连接层, 最后得到在给定观测空间状态下的每个动作所对应的 Q 值.

利用Softmax激活函数对每个动作的 Q 值处理转换为相对概率$P(a_i|s)$, 其中通过调节参数β来平衡无人机的探索与利用.

$$P(a_i|s)=\frac{e^{V_i}}{\sum_i^6 e^{V_i}} \tag{26}$$

$$V_i=\frac{Q_\theta(s,a_i)}{\beta} \tag{27}$$

通过贪婪策略得到$P(a_i|s)$中最大值的索引, 其中, $a\in A$.

$$\pi(s)=\arg\max Q_\theta(s,a) \tag{28}$$

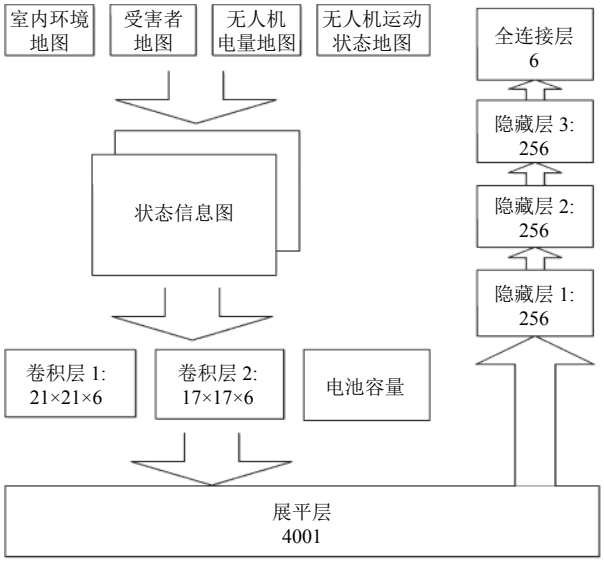

图 4 神经网络架构

4 仿真分析

本节对所提方案进行仿真与分析, 且与传统算进行对照, 最后验证了本文所提方案所具备的优势. 本文考虑了 320 m × 320 m 的无人机搜索区域, 无人机团队在所提方案的训练后在室内搜救路径规划如图 5 所示.

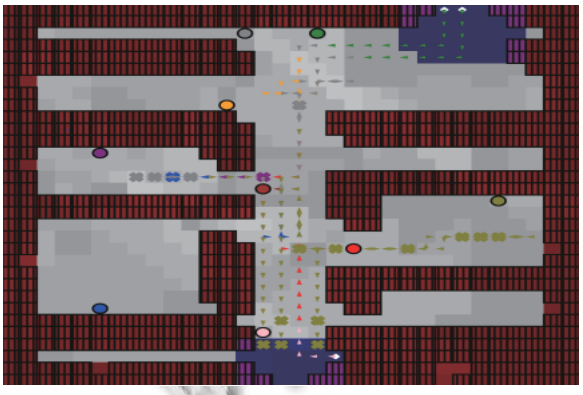

图 5 无人机团队搜救路径

在该室内区域随机分布了 9 位受害者, 为达到较好搜救效果, 本方案为其配备由 3 台无人机设备组成的搜救机群. 无人机的起落位置, 充电站都在图 5 中蓝色区域. 将不同的受害者用不同颜色的小圆圈表示. 所飞行的路径轨迹由图 5 中带箭头的线段表示, 不同路径颜色则代表此时的无人机正在搜集与本颜色所对应受害者的位置信息. 固定其飞行高度为 10 m, 飞行速度为 1 m/s, 考虑到飞行区域为室内区域, 设定路径损耗参数α_{LOS} = 2.27, α_{NLOS} = 3.64. 训练仿真参数由表 1 给出.

相比于传统方案, 所提方案优势在于加入充电操作、引入多智能体、适用大状态搜救场景 3 方面, 为验证 3 方面所具备的优势, 分别使用所提方案与传统的算法对本文场景进行了无人机群的路径规划训练. 经过 3 000 000 次的训练迭代, 所提方案使得累计回报值得到显著提升.

表 1 仿真参数设置

参数	描述	值
N	训练次数	3 000 000
L	局部地图缩放参数	17
G	全局地图缩放参数	3
Buffer_size	缓冲区大小	50 000
Minibatch_size	每批训练集大小	128
γ	折扣系数	0.95
β	调节参数	0.1

本文方案加入充电模块, 相比于未加入此模块的传统方案训练所得的累计回报对比如图 6 所示.

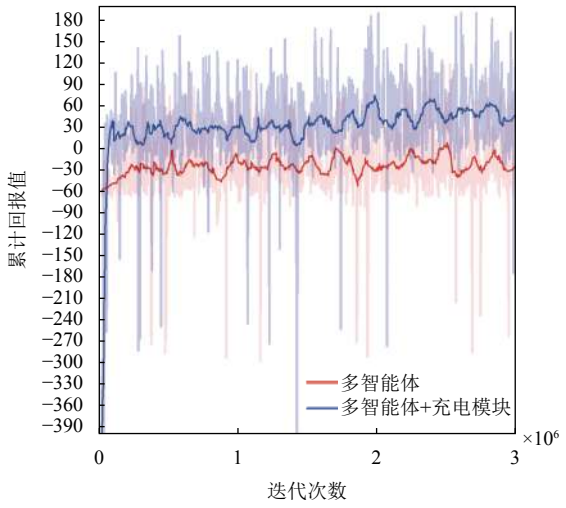

图 6 两种方案的训练回报对比

在图 6 中, 所提方案累计回报值增长较快, 最终达到收敛. 可以明显看出加入充电模块的回报值明显高于传统方案. 这种比较说明了本文方案能够快速的适应搜救场景, 且能高效的进行辅助搜救, 因此本文方案更加有效.

为说明所提方案在大状态环境的适用性, 将本文室内搜救场景下的受害者的数量以及部署的无人机集群规模进行逐步增大, 观测其对于搜救率的影响. 如图 7 所示, 随着受害者人数的增加对同一规模集群其搜救率并没有出现明显的骤降, 之所以出现下降趋势是由

于在整个搜救场景中受害者的位置被随机的放置, 受害者人数的增加导致无人机的碰撞概率增大最终导致无人机集群更加复杂的路径规划, 因此使得搜救成功率有所下降.

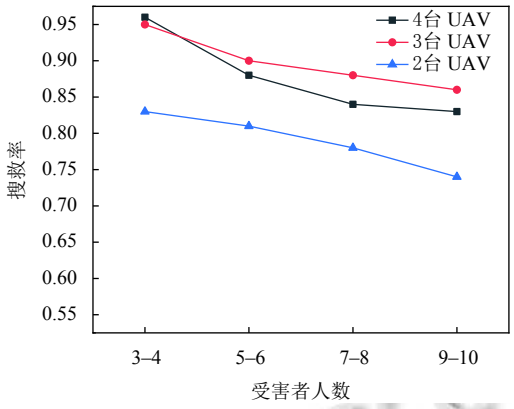

图 7 所提方案对于大状态环境下的搜救率

在图 7 中, 由 2 台无人机组成的集群搜救成功率明显低于其他集群, 主要原因是在大范围与复杂的搜救场景中, 2 台无人集群缺乏搜索覆盖能力, 存在搜索盲区导致搜救率下降. 如果采取由 4 台无人机组成的集群进行搜救任务, 当环境中存在 4 位以下受害者时能达到最佳搜救效果, 然而随着受害者人数增加到 5 位以上时, 搜救成功率明显低于由 3 台无人机组成的搜救集群. 这是因为环境中存在 4 位以下受害者时, 由 4 台无人机组成的搜救集群具有较强的搜救覆盖能力且不需要执行复杂的路径规划, 受害者人数增加到 5 人以上时, 4 台无人机集群虽具有较强的覆盖能力, 但在一定范围的搜索环境下, 无人机设备数量的增多以及受害者人数的上升势必会导致复杂的路径规划, 对于无人机之间的避免碰撞以及单程电池容量提出了巨大挑战. 考虑到本文实际的模拟场景以及设备成本, 最终选择 3 台无人机作为搜救集群, 并达到了最优的搜救效果.

本文对随机场景进行了 1 000 次蒙特卡洛迭代, 所得性能指标用于评估多智能体在搜救任务中的优势, 图 8 可以明显看出, 在搜救场景中部署多台无人机设备使得搜救成功率得到了显著提高, 传统的单智能体方案在搜救区域增大时缺乏搜救覆盖能力, 如其经常在某一个固定区域搜救然后直接返回着陆位置, 造成其他区域位置的受害者无法得到救援, 最终导致搜救能力下降.

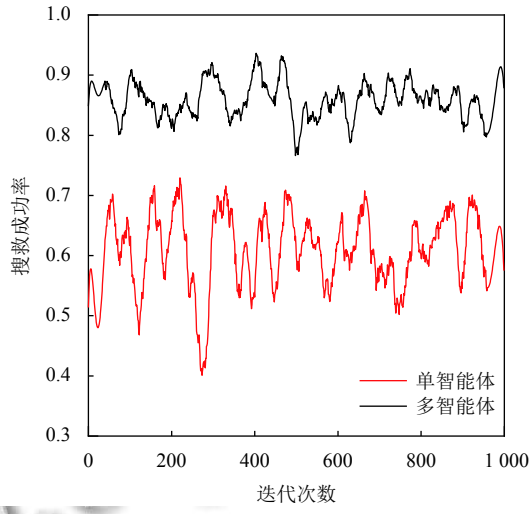

图 8 多智能体与单智能体搜救率对比

5 总结

本文研究了无人机群协作进行辅助搜救的问题, 搜救的区域是发生火灾的大型室内场景. 为提高无人机群的搜救能力本文引入了一种多智能体强化学习方案, 该方案基于 DDQN 算法来优化无人机团队的飞行路径, 解决了无人机群在不确定环境下的搜救问题, 在搜救过程中无人机群受限于电池容量, 所提方案又引入了充电模块, 从而最大程度的完成对受害者位置信息的搜集. 除此之外, 本文还详细描述了将搜救的模型转化为部分可观的马尔可夫决策的过程. 对于未来的工作, 无人机速度控制以及飞行高度的扩展是重要的探索方向, 这种扩展可以使无人机团队适应更为复杂的搜救场景.

参考文献

1 Zeng Y, Zhang R, Lim TJ. Wireless communications with unmanned aerial vehicles: Opportunities and challenges. IEEE Communications Magazine, 2016, 54(5): 36–42. [doi: 10.1109/MCOM.2016.7470933]

2 Pham HX, La HM, Feil-Seifer D, *et al*. Reinforcement learning for autonomous UAV navigation using function approximation. 2018 IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR). Philadelphia: IEEE, 2018. 1–6.

3 Chowdhury MU, Bulut E, Guvenc I. Trajectory optimization in UAV-assisted cellular networks under mission duration constraint. 2019 IEEE Radio and Wireless Symposium

(RWS). Orlando: IEEE, 2019. 1–4.
4 Chowdhury MU, Erden F, Guvenc I. RSS-based Q-learning for indoor UAV navigation. 2019 IEEE Military Communications Conference (MILCOM). Norfolk: IEEE, 2019. 121–126.
5 Ezuma M, Erden F, Anjinappa CK, *et al*. Detection and classification of UAVs using RF fingerprints in the presence of Wi-Fi and Bluetooth interference. IEEE Open Journal of the Communications Society, 2019, 1: 60–76.
6 Zhang YM, Mehrjerdi H. A survey on multiple unmanned vehicles formation control and coordination: Normal and fault situations. 2013 International Conference on Unmanned Aircraft Systems (ICUAS). Atlanta: IEEE, 2013. 1087–1096.
7 Bertuccelli LF, How JP. Search for dynamic targets with uncertain probability maps. 2006 American Control Conference. Minneapolis: IEEE, 2006. 6.
8 Bourgault F, Furukawa T, Durrant-Whyte HF. Decentralized Bayesian negotiation for cooperative search. Proceedings 2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (IEEE Cat. No. 04CH37566). Sendai: IEEE, 2004. 2681–2686.
9 沈延航, 周洲, 祝小平. 基于搜索理论的多无人机协同控制方法研究. 西北工业大学学报, 2006, 24(3): 367–370.
10 Imanberdiyev N, Fu CH, Kayacan E, *et al*. Autonomous navigation of UAV by using real-time model-based reinforcement learning. 2016 14th International Conference on Control, Automation, Robotics and Vision (ICARCV). Phuket: IEEE, 2016. 1–6.
11 Gandhi D, Pinto L, Gupta A. Learning to fly by crashing. 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Vancouver: IEEE, 2017. 3948–3955.
12 Venturini F, Mason F, Pase F, *et al*. Distributed reinforcement learning for flexible UAV swarm control with transfer learning capabilities. Proceedings of the 6th ACM Workshop on Micro Aerial Vehicle Networks, Systems, and Applications. Ontario: ACM, 2020. 10.
13 Liu CH, Dai ZP, Zhao YN, *et al*. Distributed and energy-efficient mobile crowdsensing with charging stations by deep reinforcement learning. IEEE Transactions on Mobile Computing, 2021, 20(1): 130–146. [doi: 10.1109/TMC.2019.2938509]
14 Zhang Y, Li B, Gao FF, *et al*. A robust design for ultra reliable ambient backscatter communication systems. IEEE Internet of Things Journal, 2019, 6(5): 8989–8999. [doi: 10.1109/JIOT.2019.2925843]
15 Narayanan S, Renzo MD, Graziosi F, *et al*. Distributed spatial modulation: A cooperative diversity protocol for half-duplex relay-aided wireless networks. IEEE Transactions on Vehicular Technology, 2016, 65(5): 2947–2964. [doi: 10.1109/TVT.2015.2442754]
16 Sutton RS, Barto AG. Reinforcement learning: An introduction. IEEE Transactions on Neural Networks, 1998, 9(5): 1054.
17 Mnih V, Kavukcuoglu K, Silver D, *et al*. Human-level control through deep reinforcement learning. Nature, 2015, 518(7540): 529–533. [doi: 10.1038/nature14236]