

doi: 10.7690/bgzd.2025.12.022

基于深度强化学习的无人机辅助配送系统

夏庆锋^{1,3}, 汪萍萍², 许可儿⁴, 朱良宇⁴(1. 无锡学院自动化学院, 江苏 无锡 214105; 2. 应急管理部上海消防研究所, 上海 200032;
3. 南京大学工程管理学院, 南京 210093; 4. 南京信息工程大学自动化学院, 南京 210044)

摘要: 为了使无人机能够在复杂的 3 维城市环境中, 从起点到终点之间快速找到一条无事故、更短、更安全的飞行路径, 设计一种动态环境下基于近端策略梯度优化 (proximal policy optimization, PPO) 的先进群体优化算法的无人机辅助配送系统, 并提出了 PPO-PSO 算法。基于标准 PPO 算法和粒子群算法 (particle swarm optimization, PSO) 的特点, 对 PPO 算法进行了新的改进; 融入了长短时记忆 (long short term memory, LSTM)、卷积神经网络 (convolutional neural network, CNN), 并利用粒子优化对智能体迭代方式进行修改, 解决神经网络局部搜索能力差的问题; 论证该算法的收敛性, 并在 Python 环境下进行仿真, 验证其有效性。模拟结果表明, PPO-PSO 在收敛速度和求解速度上更优, 且鲁棒性较好。

关键词: PPO; PSO; UAV 配送; 任务决策; 自主规划**中图分类号:** TP183; V279 **文献标志码:** A

UAV Assisted Distribution System Based on Deep Reinforcement Learning

Xia Qingfeng^{1,3}, Wang Pingping², Xu Keer⁴, Zhu Liangyu⁴(1. School of Automation, Wuxi University, Wuxi 214105, China;
2. Shanghai Fire Research Institute of the Ministry of Emergency Management, Shanghai 200032, China;
3. School of Management and Engineering, Nanjing University, Nanjing 210093, China;
4. School of Automation, Nanjing University of Information Science and Technology, Nanjing 210044, China)

Abstract: In order to make the unmanned aerial vehicle quickly find an accident-free, shorter and safer flight path from the starting point to the end point in the complex three-dimensional urban environment, An advanced swarm optimization algorithm based on proximal policy optimization (PPO) for unmanned aerial vehicle aided distribution system in dynamic environment is designed, and a PPO-PSO algorithm is proposed. Based on the characteristics of the standard PPO algorithm and the particle swarm optimization (PSO) algorithm, the PPO algorithm is improved; It integrates long short time memory (LSTM), convolutional neural network (CNN), and uses particle optimization to modify the iteration mode of agents, which solves the problem of poor local search ability of neural network; The convergence of the algorithm is demonstrated, and its effectiveness is verified by simulation in Python environment. The simulation results show that PPO-PSO is better in convergence speed and solution speed, and has better robustness.

Keywords: PPO; PSO; UAV distribution; task decision-making; independent planning

0 引言

近年来, 在物流和无人机领域的研究成果很多, 但将这 2 个领域结合起来的成果很少。研究无人机领域与物流领域的结合具有理论意义, 因其较快的飞行速度和轻便灵巧特性, 非常适合物流配送服务的最后 1 km 的交付^[1]。

由于无人机不受道路基础设施的限制, 能够相对轻松地穿过山、丛林、河流等地区, 走更短的路线, 相比于其他运输方式有更快的效率。地面车辆在交付途中遇到许多障碍, 有时需要支援穿越海上等其他无法通行的地区, 但无人机在飞往目的地的

途中不受阻碍。由于无人机能够有效避免交通拥堵, 因此被用于城市地区提供快速、精准的交付服务。无人机还可以用于救灾物资运送到灾区, 并且在灾区救灾交付任务中也不存在二次伤害的担忧^[2], 因此, 规划的路径还需要满足无人机对作业环境施加的安全约束和可行性约束^[3]。

为了降低物流企业的运营成本、提高运输效率, 笔者提出一种基于深度强化学习 (deep reinforcement learning, DRL) 的无人机辅助配送系统及方法, 用于无人机在交付物流中的应用。

1) 为使规划的路径满足无人机对作业环境施加的安全约束和可行性约束, 笔者构建了无人机配

收稿日期: 2024-10-06; 修回日期: 2024-11-12

基金项目: 新一代信息技术创新项目 (2022IT208)

第一作者: 夏庆锋 (1982—), 男, 山东人, 博士。

通信作者: 汪萍萍 (1979—), 女, 安徽人, 博士。

送调度问题模型^[4];

2) 路径规划算法采用相互独立的 LSTM-CNN 神经网络架构, 基于 PPO-PSO 算法利用粒子优化对智能体迭代方式进行修改, 结合使用卷积神经网络(CNN)^[5]网络对 agent 进行参数化, 利用长短时记忆(LSTM)^[6-7]计算奖励值, 通过状态信息更新网络参数并进行自适应优化迭代, 同时考虑无人机接收到的时间序列数据和环境上下文信息。

1 系统建模

1.1 问题描述

笔者提出一种基于 DRL 的无人机辅助配送系统, 包括配送中心、无人机装填仓、配送终点和调度中心。配送中心的待配送物品由车辆运输至无人机装填仓, 之后无人机负责将装填仓待配送物品配送到终点; 相关人员可在无人机装填仓为无人机更换电池; 调度中心设立在无人机装填仓, 用于调度无人机, 调度软件实时更新物流状态和无人机状态, 同时生成无人机配送路径规划结果供工作人员参考; 在执行复杂场景任务时, 可分配多架无人机协同执行任务, 当出现系统故障时, 由相关人员修复故障; 配送终点可以是快递柜。

无人机辅助配送规划问题可以看作是一个组合优化问题, 问题的求解也类似于经典的旅行商问题(travelling salesman problem, TSP)。假设: 1) 无人机是同构的, 每架无人机的基本参数一致; 2) 无人机在完成对客户的交付时便会立即飞向另一个客户或者回无人机装填仓; 3) 在本问题中将无人机看作旅行商, 将客户看作需要经过的各城市结点; 4) 无人机在电池没电之前能够完成交付并且返回无人机装填仓。无人机辅助配送过程为: 车辆装填待配送物品到无人机装填仓, 无人机按照规划出的最优路径将待配送物品配送到距离用户最近的快递柜, 用户自行取件; 无人机在完成最近的快递柜的交付时便会立即飞向另一个快递柜; 无人机全部配送完成后回到无人机装填仓。

1.2 优化目标

无人机装填仓有 2 个指标, 分别用于区分无人机从该无人机装填仓起飞和起飞完成(飞行)的时间。假设无人机到达无人机装填仓时的补货过程与剩余耗材无关。无人机和地面车辆的交付计划是基于相同的(无人机装填仓)仓库规格和相同的客户信息独立推导的。本文中提出的方法以不同的无人机

和地面车辆信息, 如车辆速度和载重量作为输入数据, 独立获取无人机和地面车辆的交付计划; 因此, 本文中构建的无人机配送调度问题模型可归属为小规模调度问题, 通过混合整数线性规划模型来进行解决。

1.3 问题模型和约束条件

1) 定义优化目标 A 为用户对配送服务的满意度, 优化目标 A 是一个与配送时间相关的量:

$$A = \sum_{i=1}^n \max(0, t_i - t_i^0) \quad (1)$$

式中: t_i 为包裹 p_i 的实际送达时间; t_i^0 为包裹 p_i 的最佳送达时间; p_i 为第 i 个包裹; n 为包裹的数量。

2) 定义优化目标 B 为所有包裹的总配送时长:

$$B = \max_{1 \leq k \leq n} t_k^e \quad (2)$$

式中: t_k^e 为无人机 u_k 完成所有任务所耗费的时间, u_k 为第 k 个无人机。

3) 将多目标优化问题转化为单目标优化问题, 把对配送总时长 B 和用户对配送服务的满意度 A 的优化目标转化为对目标 C 的优化:

$$C = B^{\max} \times A + B \quad (3)$$

式中 $B^{\max}$ 为所有包裹配送完成时间的最大值。

4) 约束条件, 包括单个包裹的最大重量限制约束、无人机单次飞行的总载重量限制约束、单次配送的运输距离限制约束。

单个包裹的最大重量限制约束为:

$$\max_{1 \leq i \leq n} w_i \leq c, \forall k, f \quad (4)$$

式中: w_i 为包裹 p_i 的重量; $\forall k, f$ 为任意无人机以及飞行趟数; c 为无人机载荷上限。

无人机单次飞行的总载重量限制约束为:

$$\sum_{b=1}^{T_{k,f}} w_{p_{k,f,b}} \leq c, \forall k, f \quad (5)$$

式中: $T_{k,f}$ 为无人机 u_k 第 f 趟飞行中需要配送的包裹总数; $p_{k,f,b}$ 为无人机 u_k 第 f 趟飞行中第 b 个配送的包裹; $w_{p_{k,f,b}}$ 为包裹 $p_{k,f,b}$ 的重量。

单次配送的运输距离限制约束为:

$$\sum_{b=0}^{T_{k,f}-1} d(l_{k,f,b}, l_{k,f,b+1}) + d(l_{k,f,T_{k,f}}, s_0) \leq v \times e, \forall k, f \quad (6)$$

式中: $l_{k,f}$ 为无人机 u_k 第 f 趟飞行中所有包裹的目的地; s_0 为初始站点; v 为无人机平均飞行功率; e 为无人机最大续航时间; $l_{k,f,b}$ 为无人机 u_k 第 f 趟飞行中第 b 个包裹的目的地。

5) 期望在满足以上 3 种系统限制的情况下最小化 C , 则配送包裹数可通过式(7)—(10)计算得到:

$$t_k^e = t_{k, T_k} + q_{k, T_k}, \forall k. \quad (7)$$

式中: t_k^e 为无人机 u_k 完成所有任务所耗费的时间; t_{k, T_k} 为无人机 u_k 飞行的开始时间; q_{k, T_k} 为无人机 u_k 飞行的总耗时。

$$t_{k, f} = \begin{cases} 0, & f=1 \\ t_{k, f-1} + q_{k, f-1}, & f \geq 2 \end{cases}, \forall k. \quad (8)$$

式中: $t_{k, f}$ 为无人机 u_k 第 f 趟飞行的开始时间; $t_{k, f-1}$ 为无人机 u_k 第 $f-1$ 趟飞行的开始时间; $q_{k, f-1}$ 为无人机 u_k 第 $f-1$ 趟的总耗时。

$$q_{k, f} = \sum_{b=0}^{T_{k, f}-1} d(l_{k, f, b}, l_{k, f, b+1}) + d(l_{k, f, T_{k, f}}, s_0) + t^d + T_{k, f} \times t^c, \forall k, f. \quad (9)$$

式中: $q_{k, f}$ 为无人机 u_k 第 f 趟的总耗时; t^d 为无人机在 s_0 装载包裹所需要的时间; t^c 为无人机在 $s_1, \dots, s_m$ 投放包裹所需要的时间; $s_1, \dots, s_m$ 为第 1, $\dots, m$ 个站点。

$$p_{k, f, b} = \sum_{i=1}^n i \times x_{k, f, b, i}, \forall k, f, b. \quad (10)$$

式中: $p_{k, f, b}$ 为无人机 u_k 第 f 趟飞行中第 b 个配送的包裹; $x_{k, f, b, i}$ 为决策变量; b 为第 b 个包裹。

6) 无人机配送调度问题模型的决策为式(11)—(12):

$$\sum_{i=1}^n x_{k, f, b, i} = 1, \forall k, f, b; \quad (11)$$

$$\sum_{k=1}^h \sum_{f=1}^{T_k} \sum_{b=1}^{T_{k, f}} x_{k, f, b, i} = 1, \forall i, x_{k, f, b, i} \in \{0, 1\}. \quad (12)$$

式中 $x_{k, f, b, i}$ 为决策变量, 取 0 或 1, 分别代表执行动作与不执行动作。

笔者构建的无人机配送调度问题模型, 采用马尔可夫决策过程模型为基础模型进行构建, 该决策过程可以用五元组 (S, A, R, P, γ) 来描述。其中 S 为状态空间, A 为作用空间, R 为奖励, P 为状态转移概率, γ 为衰减系数, 且 γ 处于区间 $[0, 1]$ 内。在第 n 个周期内, 状态 $s \in S$ 记为 S_n , 无人机控制信号 $a \in A$ 记为 A_n , 无人机动作对应的累计回报记为 $R_{n+1}(S_n, A_n)$ 。无人机在状态 s 中动作 a 后, 环境状态向 s' 过渡的概率 P 记为 $P_{ass'}$ 。马尔可夫决策过程的奖励如式(13)所示, 训练目标是使 Reward 最大化。

$$\text{Reward} = r_1 + \gamma r_2 + \gamma^2 r_3 + \dots + \gamma^{n-2} r_{n-1} + \gamma^{n-1} r_n. \quad (13)$$

考虑到无人机辅助配送系统的状态不是完全可观的, 将该决策过程建模成部分可观的马尔可夫决策 (partially observable Markov decision process, POMDP)。POMDP 在 MDP 的基础上引入了一些额外的系统信息, 如状态观测函数 O , 置信度 b 等。置信度 b 的更新公式如下:

$$b'(s') = \eta O(o|s', a) \sum_{s \in S} T(s'|s, a) b(s). \quad (14)$$

式中 η 为常数表示为: $\eta = 1/\text{Pr}(o|b, a)$, Pr 为:

$$\text{Pr}(o|b, a) = \sum_{j \in S} O(o|s', a) \sum_{s \in S} T(s'|s, a) b(s). \quad (15)$$

POMDP 的优化目标为值函数, 在置信度为 b_0 , 策略为 π 时值函数可被表示为:

$$V^\pi(b_0) = \sum_{t=0}^{\infty} \gamma^t r(b_t, a_t) = \sum_{t=0}^{\infty} \gamma^t E[R(s_t, a_t) | b_0, \pi]. \quad (16)$$

因此可以得到最优策略为:

$$\pi^* = \arg \max_{\pi} V^\pi(b_0). \quad (17)$$

2 无人机辅助配送系统设计

笔者基于标准近端策略优化 (PPO) 算法和粒子群优化 (PSO) 算法^[8]的特点, 将 PSO 算法融入 PPO 算法, 结合使用 CNN 网络对 Agent 进行参数化, 利用 LSTM 网络计算奖励值, 通过状态信息更新网络参数并进行自适应优化迭代, 同时, 考虑无人机接收到的时间序列数据和环境上下文信息, 解决了神经网络局部搜索能力差的问题, 得到了改进的 PPO-PSO 算法。

2.1 PPO-PSO 算法的改进

PPO 引入了一个新的参数如式(18)所示, 该参数体现了策略 θ 和策略 θ_k 的差异程度:

$$r_t(\theta) = \frac{\pi_\theta(a_t | s_t)}{\pi_{\theta_k}(a_t | s_t)}. \quad (18)$$

式中 $\pi_\theta(a_t | s_t)$ 为状态 s_t 时, 策略 θ 网络做出动作 a_t 的概率。

根据该参数可以设置目标函数为:

$$L^{\text{clip}}(\theta) = \hat{E}_t \left[\min \left(r_t(\theta) \hat{A}_t, \text{clip}(r_t(\theta), 1 - \varepsilon, 1 + \varepsilon) \hat{A}_t \right) \right]. \quad (19)$$

式中: $\hat{A}_t$ 为时刻 t 的优势函数值; 超参数 ε 为贪婪系数; $\text{clip}(\cdot)$ 为剪切函数。

PPO 算法的优势函数使用时序差分的算法思想, 计算公式可表示为:

$$\hat{A}_t^{\text{GAE}(\gamma, \lambda)} := (1 - \lambda)(\hat{A}_t^{(1)} + \lambda \hat{A}_t^{(2)} + \lambda^2 \hat{A}_t^{(3)} + \dots) = \sum_{l=0}^{\infty} (\gamma \lambda)^l \delta_{t+l}^V \quad (20)$$

式中 δ_t^V 为时序差分误差，可表示为：

$$\delta_t^V = r_t + \gamma V(s_{t+1}) - V(s_t) \quad (21)$$

式中： $V(s_t)$ 为值函数； r_t 为奖励函数。

2.2 LSTM-CNN 架构

该算法采用的神经网络架构为 LSTM-CNN 架构，表演者基于概率分配来选择动作，评论家根据表演者产生的动作来判断价值，表演者根据评论家的价值来修正选择行为的概率。

对输入特征参数进行归一化处理，使用 CNN 对图像进行特征提取，以捕捉到图像中的局部结构和纹理信息。使用 LSTM 对序列数据进行处理，以捕捉到长期依赖关系和时间关系。将 CNN 和 LSTM 的输出进行融合，以提升模型的预测能力。

LSTM 神经网络用于输入数据具有时序性和依赖性的场景^[9]。在物流配送任务中，无人机处理的时序数据和上下文相关环境状态使得 LSTM 网络成为合适的选择。结合 DRL 算法，此方法利用无人机的运动轨迹信息作为输入，机动控制信号作为输出，计算奖励值，并通过自适应优化算法更新网络，以训练出高效的配送模型。

LSTM 的循环单元由遗忘门 f 、输入门 i 、输出门 o 和记忆单元 C 4 部分组成：遗忘门对应的权重矩阵为 w_f 和 b_f ，输入门对应的权重矩阵为 w_i 和 b_i ，输出门对应的权重矩阵为 w_o 和 b_o ，记忆单元对应的权重矩阵为 w_c 和 b_c 。每个门单元的计算包括当前输入 X_n 、前一时刻产生的状态 h_{n-1} 、输出门 O_n 以及在 $\tanh(\bullet)$ 函数作用下计算隐藏层状态 h_n 的存储单元 C_n 。

在 CNN 中，卷积层的数学模型可以表示为：

$$y_{\{ij\}} = \sum \{k=1\}^{\wedge \{K\} \times i_k * w_{\{kj\}}} + b_j \quad (22)$$

式中： x_{ik} 为输入图像的第 i 个位置的特征值； w_{kj} 为卷积核的权重； b_j 为偏置项； y_{ij} 为卷积后的特征值。

池化层的数学模型可以表示为：

$$y_{ij} = \max \{k\} (x_{ik}) \quad (23)$$

式中： x_{ik} 为输入图像的第 i 个位置的特征值； y_{ij} 为池化后的特征值。

将 CNN 和 LSTM 结合在一起构建一个双流网络^[10]。在这个网络中，第 1 个流程使用 CNN 对图

像进行特征提取，第 2 个流程使用 LSTM 对序列数据进行处理。2 个流程的输出通过一个全连接层进行融合，以实现最终的预测任务。

3 仿真实验

3.1 模型训练过程

为了将本文中实验的 DRL 算法在 PyBullet 平台上进行训练，将训练好的神经网络架构部署到机器人操作 (ROS) 系统上，采用 ROS 系统进行仿真验证。具体的实验软硬件配置如表 1 所示。

表 1 软硬件配置信息

软件/硬件	配置
操作系统 OS	Win10
CPU	AMD Ryzen 7 5800H
GPU	NVIDIA GeForce RTX 3060
仿真语言	Python/c++
训练环境	PyBullet
仿真环境	ROS、gazebo

在此基础上，采用 LSTM-CNN 神经网络对传统 PPO 算法进行改进，并在基于 ROS 的无人机仿真测试平台上进行了实验验证。仿真结果表明，笔者提出的算法能够实现地面随机运动目标的连续跟踪和合理避障。并且与 PSO 算法相比，基于 LSTM 网络的 PPO-PSO 算法控制策略能够有效抑制环境扰动的影响，增加系统的鲁棒性。

3.2 参数选取与算法性能分析

选取任务场景对模型性能进行评估，场景采用路径规划检验模型。设定无人机高度阈值为 10 m，以两者之间的相对距离小于 20 m，模型训练参数如表 2 所示。

表 2 超参数表

参数	数值
批大小 batch size	4
每轮最大步数	200
演员网络学习率	5×10^{-5}
评论家网络学习率	5×10^{-5}
时序衰减因子 γ	0.99
A_t 评估因子 λ	1
剪切系数 ϵ	0.3
充电奖励系数 ζ_{e2}	0.05
巡检奖励系数 ζ_{p1}	1.5
综合奖励权重 α	0.5
降落耗能常数 E_L/mAh	200
批大小 batch size	4
每轮最大步数	200

基于上述场景，将 PSO 算法和笔者提出的基于 LSTM 网络的 PPO-PSO 算法设为对照组，隐层神经元个数设为 256 个，采用相同的参数优化器更新模

型训练参数。无人机自主决策和规划过程合理。当目标沿直线或圆轨迹运动时, PSO 算法和基于 LSTM 网络的 PPO 算法训练的无人机单步平均奖励初始值和最终收敛值在较小范围内波动, 后者收敛较快。

表 3 结果评估数据

算法	收敛奖励	标准差
PPO	-0.471 8	0.122 2
PSO	-0.499 0	0.176 6

由表 3 可见: 改进的 PPO 算法在收敛速度上优于 PSO 算法, 且奖励方差小于 PSO 算法, 笔者提出的算法比 PSO 算法更加稳定。

4 结束语

笔者研究了人机的配送系统设计问题, 该设计包括配送方案的生成和无人机的调度。笔者采用了 PSO 和 DRL 算法 2 种方法来优化系统性能。针对基于价值函数优化的 RL 算法在无人机自主机动任务中适用性差的问题, 采用 PPO 算法评估基于策略的优化算法在连续状态空间的可行性。

根据目标函数和约束条件, 可设定相应的奖励函数, 搭建该系统的训练环境。在此基础上采用 LSTM 神经网络对传统 PPO 算法进行改进, 在 ROS 系统的无人机仿真测试平台上进行了实验验证。

实验结果表明: 利用 RL 的决策能力和规划能力可以实现无人机的自动化执行任务, 并且对比传统的 PSO 算法, 改进后的算法在收敛速度和求解速度上更优, 基于 LSTM 网络的 PPO 算法控制策略能够有效抑制环境扰动的影响, 增加系统的鲁棒性, 并且泛化性更好。

在实际配送任务中, 应考虑对随机移动障碍物的避让。在类似的场景中, 后续研究的重点将是无人机的机动决策。

参考文献:

- [1] LI K, NI W, TOVAR E, et al. Joint flight cruise control and data collection in uav-aided internet of things: An onboard deep reinforcement learning approach[J]. IEEE Internet of Things Journal, 2020, 8(12): 9787-9799.
- [2] DING R J, GAO F F, SHEN X M. 3D UAV trajectory design and frequency band allocation for energy-efficient and fair communication: A deep reinforcement learning approach[J]. IEEE Transactions on Wireless Communications, 2020, 19 (12): 7796-7809.
- [3] 路世昌, 邵旭伦. 卡车-无人机协同救灾物资避障配送问题研究[J]. 计算机工程与应用, 2023, 59(2): 289-298.
- [4] 袁丹, 周圣杰. 基于启发式算法的具有无人机的物流配送问题研究[J]. 信息与电脑(理论版), 2021, 33(24): 81-83.
- [5] ZARÁNDY Á, REKECZKY C, SZOLGAY P, et al. Overview of CNN research: 25 years history and the current trends[C]//2015 IEEE International Symposium on Circuits and Systems (ISCAS). IEEE, 2015: 401-404.
- [6] NICOLA F, FUJIMOTO Y, OBOE R. A LSTM neural network applied to mobile robots path planning[C]//In 2018 IEEE 16th International Conference on Industrial Informatics (INDIN). IEEE, 2018, (7): 349-354.
- [7] INOUE M, YAMASHITA T, NISHIDA T. (2019). Robot path planning by LSTM network under changing environment[C]//In Advances in Computer Communication and Computational Sciences: Proceedings of IC4S. Springer Singapore, 2017, 1: 317-329.
- [8] MARINI F, WALCZAK B. Particle swarm optimization (PSO)[J]. Chemometrics and Intelligent Laboratory Systems, 2015, 149(2): 153-165.
- [9] 李永迪, 李彩虹, 张耀玉, 等. 基于 APF-LSTM-DDPG 算法的移动机器人局部路径规划[J]. 山东理工大学学报(自然科学版), 2024, 38(1): 33-41.
- [10] AKSAN F, LI Y, SURESH V, et al. CNN-LSTM vs. LSTM-CNN to predict power flow direction: a case study of the high-voltage subnet of northeast Germany[J]. Sensors, 2023, 23(2): 901.