中国科学: 信息科学 2020 年 第 50 卷 第 3 期: 375–395
SCIENTIA SINICA Informationis
群体智能专题 · 论文

强化学习方法在通信拒止战场仿真环境中多无人机目标搜寻问题上的适用性研究

汪亮*, 王文, 王禹又, 侯松林, 乔裕哲, 吴天珩, 陶先平*

南京大学计算机软件新技术国家重点实验室, 南京 210023
* 通信作者. E-mail: wl@nju.edu.cn, txp@nju.edu.cn

收稿日期: 2019–08–27; 接受日期: 2019–10–04; 网络出版日期: 2020–02–27

2018 年度科技创新 2030 —“新一代人工智能” 重大项目 (批准号: 2018AAA0102302) 和南京大学软件新技术与产业化协同创新中心资助项目

摘要 目标搜索问题是现实中一类常见的问题, 如灾难现场搜救、战场目标侦察等. 无人机由于其灵活性、低成本、可搭载各类传感器并以集群形式开展协作等优势, 是解决大范围、高风险区域目标搜索问题的理想技术方案, 当前发展迅速. 在战场等复杂现实环境中, 由于缺乏基础通信设施及干扰的存在, 无人机与地面指挥员、无人机之间难以快速、可靠通信, 处于通信拒止状态. 因此, 无人机难以获得指挥员的实时控制信息, 需要其具备自主、智能完成任务的能力并开展协同. 随着人工智能技术的快速发展, 强化学习技术在解决连续决策问题上展现出了较强的潜力. 无人机搜索问题作为一种典型的连续决策问题, 属于强化学习技术的适用范围. 但对于目前的强化学习及人工智能技术能否适用于无人机从而自主决策完成现实场景中的任务这一问题尚存争议, 仍有待进一步探索. 为此, 本文以现实战场环境为背景, 对通信拒止及包含两方对抗的战场环境中的目标搜寻问题进行了建模, 依据模型构建了对抗仿真平台, 并通过实验研究的方式针对以下 3 个问题展开了探索: (1) 强化学习在通信拒止环境下多无人机搜索问题的适用性; (2) 各强化学习算法在该问题上的优劣; (3) 通信拒止程度对强化学习算法效果的影响. 通过运用当前主流的强化学习技术开展仿真实验并定量评估实验结果. 本文总结发现: (1) 强化学习在解决通信拒止环境下多无人机搜索问题上具备有效性; (2) 在与其他算法对抗时, 运用基于 Deep Q-Network (DQN) 强化学习技术的自主决策无人机集群体现出了较强的问题解决能力; (3) 通信拒止程度对强化学习算法效果有影响, 但在不同的通信拒止程度下, 强化学习算法表现相对稳定.

关键词 无人机, 强化学习, 目标搜寻, 通信拒止环境

引用格式: 汪亮, 王文, 王禹又, 等. 强化学习方法在通信拒止战场仿真环境中多无人机目标搜寻问题上的适用性研究. 中国科学: 信息科学, 2020, 50: 375–395, doi: 10.1360/SSI-2019-0184
Wang L, Wang W, Wang Y Y, et al. Feasibility of reinforcement learning for UAV-based target searching in a simulated communication denied environment (in Chinese). Sci Sin Inform, 2020, 50: 375–395, doi: 10.1360/SSI-2019-0184

1 引言

目标搜寻是现实场景中常见的一类重要问题, 在包括灾难现场搜救 [1~3]、目标环境侦查 [4, 5]、道路车辆及航行船只监控 [6, 7] 等现实场景中, 其核心任务都是获取某一特定目标的位置. 随着相关技术的快速发展, 无人机 (unmanned aerial vehicles, UAVs) 作为一种可搭载各类感知和计算器件的飞行平台, 在高度灵活性和低成本 [8] 方面展现出了较其他目标搜索技术的优势, 为目标搜寻问题提供了一种可行方案. 进一步地, 在单架无人机有限的资源和感知范围的基础上, 通过多无人机组成集群协作开展目标搜寻, 能够有效提高搜索效率. 因此, 基于无人机的目标搜索技术具备高效、低成本、可扩展等显著优势, 其应用前景广阔, 受到了国内外科研机构的广泛关注. 目前, 包括美国国防高级研究计划局 (Defense Advanced Research Projects Agency, DARPA) 支持的 CODE [9]1), OFFSET [10]2)等项目, 已经在无人机协同作战方面取得了一定进展. 包含北京航空航天大学、国防科技大学、中国电子科技集团等在内的国内科研院校也针对无人机集群控制问题展开了深入探索, 并获得了系统性的进展 [11].

与日常生活应用场景不同, 基于无人机集群的目标搜寻技术往往被运用在包括战场、灾难现场等具有通信拒止特征的环境中. 在现代战场中, 通信资源作为一种重要的战略资源, 是作战双方争夺的焦点之一: 不仅要尽力保障我方的数据通信, 同时必须对敌方包括卫星、电磁等通信方式进行干扰和打击 [12~15]. 在通信拒止条件下, 无人机与地面指挥员、无人机相互之间的通信均受到干扰, 无法保证高速、可靠的通信与数据传输 [9]. 因此, 无人机集群难以通过指挥员实时控制的方式完成目标搜寻任务, 需要各无人机具备在无指挥状态下自主、智能完成任务的能力, 并且尽可能实现多无人机间的自主协同. 在以上背景下, 本文针对通信拒止战场环境中的无人机集群自主目标搜寻问题开展研究. 具体而言, 本文研究所针对的问题及场景具有以下性质和挑战.

- **通信拒止.** 无人机与地面指挥员之间基本无法通信; 无人机之间无法实时传输大量感知和控制数据, 仅能以较低频率广播少量观测信息; 广播通信不可靠, 其通信具有一定的成功概率, 该概率反应通信拒止程度.
- **自主搜寻.** 由于通信拒止的存在, 无人机无法接收来自地面指挥员的实时控制信息, 必须以自主决策的方式独立或协同完成目标搜寻任务.
- **运动目标.** 本文考虑战场环境中目标具有自主隐蔽等需求, 假设无人机所要搜寻的目标是运动的, 无法在搜寻开始前获知其精确初始位置和运动方式, 仅拥有目标运动的大致范围.

除以上性质外, 本文所关注的战场环境还存在以下无人机间的对抗特性.

- **双方对抗.** 在目标搜寻过程中, 也有可能会出现某方无人机被击落从而损耗无人机的现象.

为实现在复杂环境中无人机集群的自主控制, 目前基于生物集群行为的无人机控制方法已经取得了较大进展 [11, 16]. 近年来, 强化学习技术由于其本身的试错搜索、延迟奖赏的特性, 以及在如国际象棋 [17]、围棋 [18]、Starcraft [19] 等复杂游戏问题上的良好表现, 成为解决一些复杂决策问题的可能方案. 强化学习在解决现实问题方面也表现出了较强的应用潜力, 文献 [20] 提出了一种基于强化学习的高效二分图动态匹配算法并证明了其性能下界, 为运用强化学习技术解决真实问题提供了重要参考. 在无人机领域, 也涌现出了使用强化学习技术训练无人机的自主完成, 如避障 [21]、区域覆盖 [22]、路径规划 [23, 24] 等任务的研究工作. 有学者认为, 多智能体强化学习算法可以通过训练为无人机提供一个良好的决策模型 [25]. 但是, 在军事应用领域, DARPA 在其所支持的 CODE 项目中, 却明确表示不建议

1) DARPA. Collaborative operations in denied environment (code) (archived). https://www.darpa.mil/program/collaborative-operations-in-denied-environment.

2) DARPA. Offensive swarm-enabled tactics (offset). https://www.darpa.mil/work-with-us/offensive-swarm-enabled-tactics.

采用基于强化学习的训练技术. 同时, 对比上述通信拒止战场环境中目标搜寻问题的场景特性, 我们发现当前仍缺乏与该场景联系特别密切的研究工作. 综上所述, 对于强化学习方法是否适用于无人机集群解决包括目标搜寻在内的复杂现实问题这一论题尚无明确答案, 需要开展深入的探索与研究.

为了研究强化学习方法在通信拒止战场环境中基于无人机的目标搜寻问题上的有效性, 本文对通信拒止战场环境进行了显式建模, 并对该环境中的目标搜寻问题进行了定义; 依据该环境模型, 构建了软件仿真平台; 通过对现有的强化学习技术进行调研和综述, 选择了目前主流的、具有较大应用潜力的强化学习技术作为研究的对象, 在仿真平台中实现了基于强化学习的无人机自主训练与目标搜寻方法. 通过开展仿真对抗实验, 本文针对以下 3 个研究问题 (research questions, RQ) 开展了系统性的研究.

RQ1: 强化学习方法在解决通信拒止战场环境中的目标搜寻问题上是否适用?

RQ2: 现有的强化学习方法中, 哪种方法表现更优异?

RQ3: 通信拒止程度对无人机集群协作对抗的影响程度?

通过回答上述 3 个问题, 本文能够为后续针对无人机集群自主控制、强化学习技术在无人机领域应用、通信拒止条件下的多无人机协同等相关研究工作提供帮助和参考. 同时, 本文所构建的战场仿真环境也可以为后续的相关研究提供高效、低成本的研究平台. 本文所作出的贡献主要包含以下几点.

(1) 深入刻画了通信拒止战场条件下无人机协作对抗目标搜寻问题的战场环境、通信拒止程度等内容, 并结合真实场景, 构造可以模拟无人机搜寻目标、通信成功率可变的战场仿真环境.

(2) 调研了目前的主流强化学习算法, 结合理论分析, 筛选出 4 种较为符合本文问题特征的强化学习算法.

(3) 针对上述算法, 开展了系统性的实验, 将强化学习算法应用到仿真环境中, 得到无人机训练模型. 针对每个模型, 进行了不同方面的评价指标测试, 得到了反映算法效果和效率的重要数据和基于这些数据的分析结果.

本文剩余部分组织如下: 第 2 节介绍通信拒止战场环境模型、仿真平台构建方法、本文研究所选用的强化学习方法、实验组织形式, 以及实验的评价指标和评价方法. 第 3 节针对上述 3 个研究问题, 对实验结果进行展示和分析. 第 4 节介绍与本文工作相关的研究工作. 最后, 第 5 节总结本文工作.

2 实验方法

本文通过在仿真平台上开展实验的方法来对上文所述的 3 个问题开展研究. 本节介绍仿真战场环境与问题定义, 所采用的强化学习方法, 以及实验组织与评价指标设计.

2.1 仿真战场环境与问题定义

本小节介绍通信拒止战场仿真环境的搭建方法以及目标搜索问题的定义.

2.1.1 战场仿真环境

本文搭建的战场仿真环境如图 1 所示. 在本文的战场仿真环境中, 无人机和搜索目标同处于一个宽度为 1000 m × 1000 m 的正方形区域内. 我们将该区域划分为 50×50 的网格状地图, 每一小格对应一个 20 m × 20 m 的正方形区域. 仿真环境中时间为离散的, 以单位时间计算, 仿真环境中的 1 单位时间对应现实世界 1 s.

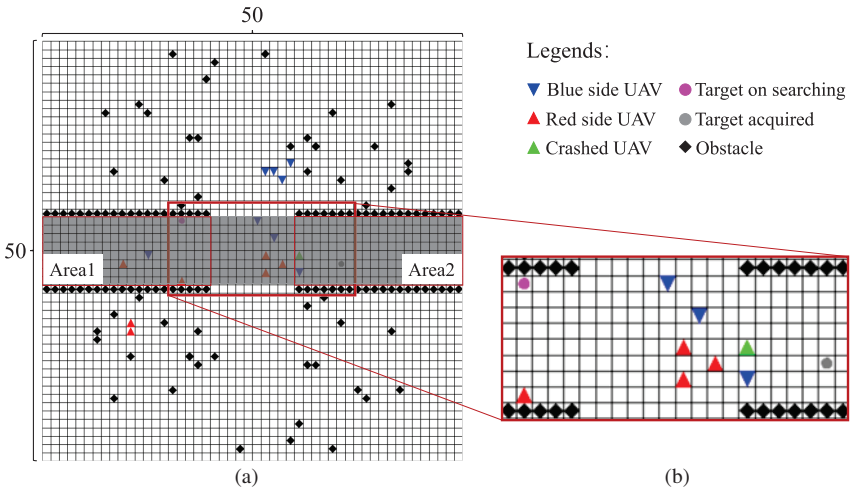

图 1 (网络版彩图) 仿真环境

Figure 1 (Color online) Simulation environment. (a) Overview of battlefield simulation environment; (b) partial enlargement

地形. 在地图中, 存在若干障碍物, 对应战场环境的地形信息. 障碍物的位置保持不变, 并且无人机预先知道所有障碍物的位置信息. 无人机无法从障碍物上方越过, 一旦无人机的位置与障碍物的位置重合, 无人机就会撞毁.

目标. 特别地, 在地图中间位置存在一狭长的 "山谷" 地区 (即图中深色区域). 无人机搜寻的目标在山谷地区无规则移动, 目标的移动速度为 5 m/s, 即仿真环境中每 4 个单位时间移动 1 格[3)]. 我们假定待搜寻目标数目存在上限 M, 在本文中, 如无特殊说明, $M = 2$, 即同时存在两个目标. 在本文组织的实验中, 每轮实验的开始, 目标分别以一定的概率分布出现在图中区域 1 和 2, 本文实验保证初始每边只存在一个目标. 目标出现在每个区域中央位置的概率最高, 概率向边缘逐渐递减. 无人机无法获知目标的初始位置.

无人机. 无人机具有阵营属性, 该仿真环境中刻画为红方与蓝方, 试验中红蓝双方分别有 8 架无人机. 无人机的飞行速度为 20 m/s, 即在仿真环境中每单位时间飞行 1 格距离. 在仿真环境中, 无人机能够通过机身传感器获取周围以自身为中心, 5×5 网格区域内的信息, 包括是否存在其他无人机和是否存在目标. 无人机的任务为尽可能快地找到位于山谷地区的目标. 在每单位时间, 每架无人机能够选择上下左右 4 个方向之一移动.

不同阵营的无人机之间存在对抗, 具体对抗形式为, 当双方无人机位于同一格时 (设坐标为 $\langle i, j \rangle$), 局部无人机数量占优的一方会将另一方的无人机击毁. 双方的局部无人机数量定义为以坐标 $\langle i, j \rangle$ 为中心的 3×3 区域内己方无人机数量. 我们规定, 当红蓝双方局部无人机数量相等时, 对抗双方处于均势, 不会有任何一方无人机被击毁.

我们以无人机的最长飞行时间对应现实世界中无人机电池容量有限这一事实, 在仿真环境, 我们规定无人机的最长飞行时间为 T 个单位时间, 若无特殊说明, 本文取 $T = 2000$, 约为 30 min.

通信. 在本文所理解和约定的拒止环境中, 无人机之间的通信受限, 具体表现为

(1) 无人机采用广播通信模式, 广播通信成功率为 ω.

(2) 无人机之间的通信仅能传输极少量的信息, 包括自身位置以及能通过传感器感知到信息.

3) 在本文场景中, 静止目标搜索问题可以理解为运动目标搜索问题的一个特例, 即目标移动速度为 0 m/s.

表 1 奖赏设置

Table 1 Reward settings

Environmental feedback	Reward
Finding a target	1000
Destorying a drone on enemy side	125
A drone being destroyed	−125
Moving a step	−1

(3) 无人机之间不能保持实时通信, 仅在每 20 单位时间或者发现目标或者自身坠毁时, 能进行通信.

2.1.2 目标搜索问题

在上述战场仿真环境中, 本文拟研究的目标搜索问题可表述为: 给定上述战场仿真环境, 给各无人机规划最佳路径, 使各无人机在尽可能保证自身生存的前提下, 以最小的行动步数完成目标. 一旦无人机的位置与搜寻目标位置重合, 即意味着无人机获取了目标.

2.2 针对本文问题的强化学习方法框架设计

如前文所述, 本文旨在探索使用强化学习技术解决通信拒止战场环境中目标搜寻问题的适用性. 因此, 需要针对该问题构建一般性的强化学习框架. 在强化学习模型中, 包含智能体和环境两个实体, 分别对应本文问题中的无人机和战场仿真环境. 智能体可以观察到环境的状态, 并相应地选择自己的动作, 该动作会对环境造成一定影响, 使得环境的状态发生改变. 该影响的好坏由事先定义的奖赏机制衡量. 将该问题建模为如下所述的马尔可夫决策过程 (Markov decision process, MDP) [26], 便于不同强化学习算法的实现.

特别地, 在本文所描述的目标搜索任务中, 对应的 MDP 可由四元组 $\langle S, A, P, R\rangle$ 定义, 其中:

状态 (S). S 为有限状态集合. 在本文场景中, 某一架无人机可以观察到的状态 s 由一组 37 维的向量表示, 其中包括: 无人机存活状态 (是否存活)、本机位置 (位置由无人机的二元组坐标表示, 形如 $\langle x, y\rangle$, 详细说明见本文第 2.1 小节)、友方无人机位置、敌方无人机位置、目标位置等.

动作 (A). A 为有限动作集合. 在本文中指无人机可以采取的飞行方向. 每次执行动作, 无人机会采取向上、下、左、右 4 个方向移动一格 4 种动作中的一种.

状态转移概率 (P). P 指从当前状态 s 出发, 执行动作 a 后, 转移到下一个状态 s' 的概率, 该概率由环境给出.

奖赏 (R). R 是与状态 – 动作对相关联的即时奖赏函数, 我们将其定义为 $R(s, a)$, 其中 $s \in S, a \in A$, $R(s, a)$ 代表在当前状态 s 下执行动作 a 得到的即时奖赏. 具体定义如表 1.

强化学习的最终目标是获得在当前状态下应该采取什么动作的策略 (policy, π) [27], 定义为从状态到可能选择动作的映射, 即 $\pi : S \mapsto A$. 在智能体不断尝试的过程中, 策略一直在进行更新, 根据该策略, 在状态 s 下就能得到动作 a (或者是可采取的动作的概率分布, 根据算法不同而有所差别).

上述马尔可夫决策过程清楚地描述了本文问题的强化学习模型, 其中的奖赏定义为即时奖赏, 用于评价某一动作对环境影响的好坏. 而累积奖赏用于表示一个策略中状态 – 动作序列完整执行一定步数后能得到的期望奖赏, 以评价策略的好坏. 因此, 我们用每次尝试的结果来更新累积奖赏, 这样的方法通常要使用强化学习中的值函数 (value function). 在上述模型下, 参照强化学习中的一般形式 [28],

γ 折扣的状态值函数 $V_\pi(s_t)$ 定义为

$$V_\pi(s_t) = \mathrm{E}_{a_t, s_{t+1}, \ldots}\left[\sum_{l=0}^{\infty} \gamma^l r_{t+l}\right], \tag{1}$$

其中 t 为时刻, s_t 代表 t 时刻智能体的状态, a_t 代表该状态下执行的动作, γ 为折扣系数, 状态值函数 $V_\pi(s_t)$ 代表从当前状态 s_t 出发, 执行某一特定策略 π 得到的期望累积奖赏.

定义状态 – 动作值函数 $Q_\pi(s_t, a_t)$ 为

$$Q_\pi(s_t, a_t) = \mathrm{E}_{s_{t+1}, a_{t+1}, \ldots}\left[\sum_{l=0}^{\infty} \gamma^l r_{t+l}\right]. \tag{2}$$

状态 – 动作值函数 $Q_\pi(s_t, a_t)$ 代表在当前状态 s_t 下执行了动作 a_t 后, 执行某一特定策略得到的期望累积奖赏. 经过不断学习, 我们希望获得这样一个策略 π, 使得智能体根据此策略决策, 能获得尽可能大的期望累积奖赏. 在上述模型下, 我们选择了多种强化学习算法在仿真环境中训练, 每种算法的具体描述见后续章节.

2.3 所采用的强化学习方法

针对上述问题建模, 本文选择了现有主流的强化学习方法. 目前常见的强化学习可以分为两类, 一种是以 Q-Learning 为典型代表的值迭代方法, 另一种是基于策略梯度的策略迭代算法. 基于值迭代的方法中, 我们使用了利用神经网络做值函数估计 (Deep Q-Network, DQN) 的方法和利用特征的线性组合做值函数估计 (线性拟合 Q-Learning, L-QL) 的方法, 在基于策略梯度的算法中, 我们使用了 asynchronous advantage actor-critic (A3C) 和 distributed proximal polcy optimization (DPPO) 两种方法.

2.3.1 Deep Q-Network 方法

Q-Learning[29] 是强化学习最早的突破性进展之一, 该算法学习动作值函数 $Q(s, a)$, 即在状态 $s \in S$ 条件下, 采取动作 $a \in A$, 能够获得累积奖赏的期望. 在最初的 Q-Learning 中, 研究者建立一个关于状态和动作的表格, 在该表格上进行更新. 对于大多数现实问题而言, 该表格过于庞大, 没有足够的空间储存完整的表格, 更新如此大量的数据也相当低效, 因此多数研究者使用某种参数化的方式估计该表格. 后续存在相当多的算法借鉴了 Q-Learning 的思想, DQN 正是其中采用神经网络估计该表格的典型代表之一. 该神经网络的输入为状态, 输出为该状态下每个动作的 Q 值.

Mnih 等 [30] 提出了基于 DQN 的强化学习方法, 在 Q-Learning 框架中, 该方法结合了最初 Q-Learning 和卷积神经网络, 在若干 Atari 游戏上获得了接近人类的水准.

本文选择如 2.2 小节描述的 37 维向量作为神经网络的特征输入. 因此, 本节使用的神经网络结构与经典 DQN 网络结构略有不同. 我们去掉了用于从图片中提取特征的卷积层, 增加了一个全连接层. 我们的网络由两个全连接层和一个输出层构成. 第 1 个全连接层包含 64 个神经元, 激活函数为修正线性单元 (relu), 第 2 个全连接层包含 32 个神经元, 激活函数依旧是 relu. 由于在任意时刻无人机有 4 种动作, 因此输出层包含 4 个神经元, 使用线性激活函数.

我们采用 ϵ- 贪心探索策略. ϵ 取值如式 (3), 其中 i 为训练轮数.

$$\epsilon = \begin{cases} 1 - \dfrac{0.9i}{512}, & i \leqslant 512, \\ 0.1, & i > 512. \end{cases} \tag{3}$$

在 DQN 中, 我们使用式 (4) 更新神经网络参数:

$$\theta_{t+1} \leftarrow \theta_t + \alpha \left[R_{t+1} + \gamma \max_a \hat{Q}(s_{t+1}, a, \theta_t^-) - \hat{Q}(s_{t+1}, a_t, \theta_t) \right] \nabla \hat{Q}(s_t, a_t, \theta_t), \tag{4}$$

其中 θ_t 是第 t 步更新的网络参数, a_t 是第 t 步选择的动作, s_t 是第 t 步的状态, 此处的梯度通过神经网络的反向传播计算. 每隔 τ 步, 将 θ 赋值给 θ^-, 本文中, $\tau = 512$.

DQN 使用了经验回放以提高样本的利用率. 其基本思想是, 将训练过程中的 (s_t, a_t, r_t, s_{t+1}) 作为经验储存于经验池中, 每次做神经网络参数更新时, 从经验池中随机选取一批样本进行更新. 本文中, 经验池大小为 100000, 每次取样本的批次大小为 32.

为了加快收敛速度和更加有效地利用经验样本, 我们使用了带权经验回放 [31]. 其基本思想是为第 i 个经验样本指定一个权值 W_i, 从经验池中随机选取样本时高权值的样本将有更大的概率被选择, 选中的概率为

$$P(i) = \frac{W_i^\alpha}{\sum_k W_k^\alpha}, \tag{5}$$

其中 α 用于控制多大程度上使用权值, 若 $\alpha = 0$, 则不使用权值, 本文的实验中取 $\alpha = 1$. W_i 则用于衡量第 i 个样本与网络当前估计存在多大的误差, 我们取 $W_i = |R_{t+1} + \gamma \max_a \hat{Q}(s_{t+1}, a, \theta_t^-) - \hat{Q}(s_{t+1}, a_t, \theta_t)| + 0.0005$, 小常数 0.0005 用于保证不存在任何样本永远不可能被取到.

本文中, 网络参数更新使用 RMSProp 优化器 [32], 学习率设置为 0.001.

2.3.2 线性拟合 Q-Learning 方法

神经网络是使用非线性函数拟合 Q 值的典型代表, 在围棋等现实问题上已经展示出了强大能力. 但神经网络往往算法时间开销较大, 结果可解释性较差. 因此, 我们也尝试使用线性函数拟合 $Q(s, a)$ 的方法 [33].

我们使用线性函数来逼近最优动作值函数. 最优动作值函数 $Q^*(s, a)$ 定义为

$$Q^*(s, a) = \max_\pi \mathrm{E}[R_t | s_t = s, a_t = a, \pi], \tag{6}$$

其中 π 表示策略, 即在某种状态下该采取什么样的动作. 最优动作值函数满足贝尔曼 (Bellman) 等式 $Q^*(s, a) = \mathrm{E}[r + \gamma \max_{a'} Q^*(s', a') | s, a]$. 使用线性函数估计动作值函数, 即将动作值函数表示为 $Q(s, a) = \theta^{\mathrm{T}}(s; a)$, 其中 θ 是估计参数向量, $(s; a)$ 是由状态与动作构成的联合向量. 我们使用 $Q(s, a; \theta)$ 表示使用参数 θ 所获得的对于动作值函数的估计. 通过使用贝尔曼等式 $Q_{i+1}(s, a) = \mathrm{E}[r + \gamma \max'_a Q_i(s', a') | s, a]$, 多次迭代后该算法会收敛到最优的 $Q^*(s, a)$.

我们的目标是使学得的动作值函数尽可能接近真实的动作值函数, 通过均方误差来衡量接近程度. 误差定义为

$$L(\theta) = \mathrm{E}_{s, a \sim \rho(s,a)}[(Q_\pi(s, a) - Q(s, a; \theta))^2]. \tag{7}$$

事实上, 我们无法获得真实的 $Q_\pi(s, a)$, 但可以借助时序差分学习, $Q_\pi(s, a) = r + \gamma \max_{a' \in A} Q(s', a'; \theta)$, 使用当前估计的动作值函数代替准确的值函数, 以进行更新.

因此, 式 (7) 可改写为

$$L(\theta) = \mathrm{E}_{s, a \sim \rho(s,a)} \left[\left(r + \gamma \max_{a' \in A} Q(s', a'; \theta) - Q(s, a; \theta) \right)^2 \right]. \tag{8}$$

我们使用梯度下降进行参数更新, 更新规则如下:

$$\theta \leftarrow \theta - \frac{1}{2}\alpha \nabla L(\theta). \tag{9}$$

本文使用高斯 (Gauss) 核函数将 2.2 小节描述的 37 维状态向量 s 映射为 4096 维状态向量 s' [34], $s'_i = \exp(-\frac{||s-l^{(i)}||^2}{2\sigma^2})$, 其中 $l^{(i)}$ 为在环境中随机采样的第 i 个状态样本.

该方法中, 我们依旧使用 ϵ- 贪心策略进行探索, ϵ 取值同式 (3).

2.3.3 Asynchronous advantage actor-critic 方法

2.3.2 小节简要阐述了基于值函数的强化学习方法, 此类方法对价值函数进行了近似表示, 引入了一个状态 – 动作价值函数 (见式 (1)). 在基于策略 (policy) 的强化学习方法下, 采用类似的思路, 但我们对策略进行直接迭代, 用包含参数 θ 的函数 $\pi_\theta(s,a)$ 来近似表示策略, 如下式所示:

$$\pi_\theta(s,a) = P(a|s,\theta). \tag{10}$$

在基于策略的强化学习算法中, 可定义策略的期望收益为

$$J(\theta) = \mathrm{E}_{\tau\sim\pi_\theta(\tau)}[r(\tau)] = \int_\tau r(\tau)\pi_\theta(\tau)\mathrm{d}\tau \approx \sum_{t>0} r(a_t|s_t)\pi_\theta(a_t|s_t), \tag{11}$$

其中 τ 即状态 – 动作. 由期望收益得到参数 θ 的更新公式为

$$\theta \leftarrow \theta + \sum_{t>0} r(a_t|s_t)\nabla_\theta \log_{\pi_\theta}(a_t|s_t). \tag{12}$$

基于这样的思想, Actor-Critic 方法在对策略进行直接迭代的基础上, 增加值函数以评价选择的动作. Actor 代表算法中的策略结构, 它被用于动作选择; Critic 代表值函数, 评价 Actor 所选择的动作.

在 Actor-Critic 方法中, 因为引入了价值函数, 式 (12) 参数更新公式改写为

$$\theta \leftarrow \theta + \sum_{t>0} Q(s_t,a_t)\nabla_\theta \log_{\pi_\theta}(a_t|s_t). \tag{13}$$

可以看到 Actor 的表现一定程度上取决于 Critic 的拟合程度, 而价值函数本身就较难收敛, 所以这使得一般的 Actor-Critic 算法更加难收敛.

A3C [35] 是一种典型的 Actor-Critic 方法, 考虑到上述 Actor-Critic 方法的缺陷, 为了加快收敛, A3C 算法首先考虑到机器学习中要求样本是满足独立同分布的, 而单个 agent 采样得到的样本具有高度相关性, 减小样本间的相关性可有效提高模型收敛速度. 在如 DQN 等算法中常采用经验回放策略减小样本相关性, 而在 A3C 算法中采用异步训练框架, 同样也可以很好地解决这一问题且无需像经验回放一样占用大量内存空间, 核心思想是在多个环境实例上异步并行执行多个 agent, 因为每个 agent 都是独立地与环境进行交互而得到采样序列, 从而去样本相关性.

同时, 采用优势函数和 n 步采样, 进一步优化算法以加快收敛, 式 (13) 也因此改写为

$$\theta \leftarrow \theta + \sum_{t>0} A(s_t,a_t)\nabla_\theta \log_{\pi_\theta}(a_t|s_t) \approx \sum_{t>0}\left(\sum_{i\geqslant 0}\gamma^i R_{t+i} + V(s'_t) - V(s_t)\right)\nabla_\theta \log_{\pi_\theta}(a_t|s_t). \tag{14}$$

最后, 为了防止过早陷入局部最优, 在更新参数 θ 时添加策略 π 的熵项 $H(\pi(s_t,\theta))$ 增加探索 [36], 式 (13) 最终改写为

$$\theta \leftarrow \theta + \sum_{t>0}\left(\sum_{i\geqslant 0}\gamma^i R_{t+i} + V(s'_t) - V(s_t)\right)\nabla_\theta \log_{\pi_\theta}(a_t|s_t) + \beta\nabla_\theta H(\pi(s_t,\theta)), \tag{15}$$

其中超参数 β 在本实验中取 0.001.

2.3.4 Distributed proximal policy optimization 方法

本小节介绍 DPPO 算法, DPPO 算法是 PPO 算法的分布式改进, 其理论基础是 PPO 算法.

PPO 算法是由 OpenAI 提出的一种基于策略的算法 [37]. 在传统的 Policy Gradient 算法中, 学习步长难以确定, 如果学习的步长太小, 会增加训练时间, 而学习步长太大会导致得到的策略一直在震荡, 难以收敛, 不合理的学习步长有时会让某些更新导致模型的表现大幅下降, 而 PPO 算法则精心构造了更新梯度, 使得算法能够得到一个合理的学习速率. PPO 算法同样基于 Actor-Critic 框架. 在更新 Actor 的时候, 需要最大化下式 [38]:

$$J_{\text{ppo}}(\theta) = \sum_{t=1}^{T}\frac{\pi_\theta(a_t|s_t)}{\pi_{\text{old}}(a_t|s_t)}\hat{A}_t - \lambda \text{KL}[\pi_{\text{old}}|\pi_\theta], \tag{16}$$

其中 θ 是 Actor 函数的参数. Actor 在旧策略上, 根据优势 $\hat{A}_t$ 修改新策略, 如果优势较大, 则修改幅度大, 使得新策略更可能发生. 并且梯度公式的后面附加了一个 KL 散度项, 这相当于一个惩罚项, 简单来说, 如果新旧策略的差异过大, 那么 KL 散度也较大, 但算法设计的思路是不希望新旧策略差太多, 因为这样相当于使用了大的更新步长, 难以收敛.

文献 [37] 提到了另一种更新的方式, 就是不使用 KL 散度, 而是使用剪切代理对象 (clipped surrogate objective), 记

$$u_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\text{old}}(a_t|s_t)}. \tag{17}$$

Surrogate objective 就是

$$\hat{\text{E}}_t[u_t(\theta)\hat{A}_t]. \tag{18}$$

Clipped surrogate objective 就是限制了 surrogate 的变化幅度, 类似于 KL. 最终我们的优化目标变成下式:

$$L^{\text{CLIP}}(\theta) = \hat{\text{E}}_t[\min(u_t(\theta)\hat{A}_t, \text{clip}(u_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t)]. \tag{19}$$

PPO 算法在更新 Critic 时, 我们需要最小化下式:

$$L_{\text{BL}}(\phi) = -\sum_{t=1}^{T}\left(\sum_{t'>t}\gamma^{t'-t}r_{t'} - V_\phi(s_t)\right)^2, \tag{20}$$

其中 ϕ 是 Critic 函数的参数. Critic 网络的更新与一般的 Actor-Critic 框架并无差别, 也是最小化优势函数的误差.

DPPO 通过多 agent 同时采样消除了单 agent 带来的轨迹的相关性, 并加速了学习. 在生成轨迹时, 我们会有一个全局的 PPO 策略, 同时使用多个不同的个体并行产生轨迹, 并计算轨迹的奖赏, 当轨迹积累到一定数量时, 我们就把所有产生的轨迹放入全局 PPO 中训练.

本文使用剪切代理对象作为 Actor 网络的更新方式.

表 2 RQ1 实验组织

Table 2 RQ1 experiment settings

Number	Algorithm (red)	Algorithm (blue)
1	Random walk	Random walk
2	DQN	Random walk
3	L-QL	Random walk
4	A3C	Random walk
5	DPPO	Random walk

2.3.5 基准方法: 随机游走 (random walk) 结合避障规则

为了验证各强化学习算法的有效性, 以及为对抗设定一个基准对手, 我们引入随机游走结合避障规则这一基准方法. 具体而言, 该基准方法不包含任何需要学习的部分. 无人机做决策时考虑避免与障碍物以及友方飞机的碰撞, 在有多种可以选择的方向时, 无人机随机选择一个可行的方向. 对于避障规则的具体描述如下:

(1) 若向方向 dir 移动后会导致无人机与障碍物发生碰撞或者越界, 则方向 dir 不可行.

(2) 若向方向 dir 移动后可能会导致无人机与友方无人机发生碰撞, 则方向 dir 不可行.

(3) 若根据上述规则过滤掉不可行方向后, 不存在可行方向, 则随机选择 4 个方向之一, 否则从可行的方向中随机选择一个.

2.4 实验组织

针对上文所提出的 3 个科学问题, 本文使用仿真实验研究的方法开展探索, 具体的实验组织如下所述.

RQ1: 强化学习方法在通信拒止战场仿真环境中多无人机目标搜索问题上是否适用?

通常来说, 强化学习方法对于目标搜寻问题的适用性体现在寻找目标的效果和效率上. 这里的效果可以分为两点: 在多轮测试中, 能否找到尽量多的目标; 能否在每一轮测试中都完成任务. 效率是指无人机能否在尽量短的时间内完成任务.

因此在第一个问题的实验组织中, 我们围绕第 2.5 小节中的 3 个指标: 目标获取数量均值和任务完成率, 以及平均成功时间组织实验. 具体如下: 测试蓝方固定为随机游走策略, 红方分别为备选算法 DQN, L-QL, A3C, DPPO 4 种算法中的一种. 同时, 设置对照组红蓝两方都为随机游走策略, 共 5 组对抗, 如表 2 所示, 分别训练得到效果最好的模型. 测试时, 将强化学习算法与随机游走算法对比, 观察在上述指标上的具体表现.

RQ2: 现有的强化学习方法中, 哪种方法表现更优异?

为了回答该问题, 我们直接将备选算法分别置于红蓝双方, 进行对抗训练, 能有效体现算法的直接差异. 具体实验组织方法如下: 我们选择上述算法中, 在第 1 组实验中综合表现最优的 3 个, 设置对比实验如表 3 所示. 为每组实验分别训练对抗模型, 选择效果最好的模型进行测试. 每组实验比较对抗获胜率和平均成功时间两个指标 (见第 2.5 小节), 以体现算法之间的效果和性能差异.

RQ3: 通信拒止程度对无人机集群协作对抗的影响程度?

在衡量通信拒止程度对无人机进行协作对抗的影响程度时, 我们需要保持其他实验配置不变, 只改变测试时的通信成功率, 观察通信成功率的变化对上述指标的影响. 特别地, 针对不同的通信拒止程度 ($\omega = 40\%$ 和 $\omega = 100\%$) 进行了训练. 具体实验组织为: 蓝方以随机游走算法, 红方以上述实验中

表 3 RQ2 实验组织

Table 3 RQ2 experiment settings

Number	Algorithm (red)	Algorithm (blue)
1	DQN	L-QL
2	DQN	DPPO
3	L-QL	DPPO

效果最好的算法 (在本文中为 DQN 算法) 进行实验, 改变测试环境中的通信拒止程度, 即通信成功率 (如第 2.1 小节所述), 比较不同的通信成功率对目标获取数量均值和任务完成率, 以及平均成功时间 3 个指标的影响.

2.5 评价方法与指标

针对上述实验设计, 我们在红蓝双方障碍物位置对称、离目标距离相同的公平条件下, 以一组量化的评价指标来衡量各个算法完成目标的效果和效率. 参照相关强化学习算法研究中的评价方法 [35, 38], 从算法表现和性能两个角度设计了评价指标. 具体而言, 我们所采用的评价指标包含以下内容.

2.5.1 目标获取数量均值

针对本文所研究的仿真战场环境中目标搜寻问题, 其主要的功能指标是能否成功获取目标位置. 因此, 针对每一种算法, 在与随机算法对抗的场景下 (或称为无对抗), 我们采用多轮测试, 并根据式 (21) 计算其每轮所能获取目标的个数来衡量该算法的性能.

$$\overline{\mathrm{TA}} = \frac{\sum_{i=1}^{N} \mathrm{TA}_i}{N}, \tag{21}$$

其中 $\overline{\mathrm{TA}}$ 为目标获取 (target acquired) 数量均值, TA_i 为第 i 架无人机所获取的目标数量, $\mathrm{TA}_i \in \{0, 1, \ldots, M\}, i = 1, 2, \ldots, N$, M 为一轮测试中所能获取的目标最大数量 (如第 2.1 小节所述, 在本文中取 $M = 2$), N 为测试轮数 (如无特殊说明, 本文取 $N = 1000$). 因此, $\overline{\mathrm{TA}}$ 的取值范围为 $\overline{\mathrm{TA}} \in [0, 2]$, $\overline{\mathrm{TA}}$ 值越高, 说明算法在目标搜寻问题上的有效性越高. 该指标帮助我们回答本文所提出的 RQ1, 如果目标获取数量均值越大, 说明该算法的适用性越高.

2.5.2 任务完成率

在上述相同场景下, 以目标获取数量均值为基础, 本文提出任务完成率的指标. 由于本文所规划的问题包含了对两个目标的搜寻, 我们定义一次任务完成为算法在一轮测试中找到两个目标.

与单个目标最大的不同之处在于, 两个 (及以上) 目标提供了无人机协作完成目标搜寻的可能性. 在算法训练过程中, 无人机有可能通过协作的方式更高效地完成任务. 这就是本文设置两个目标的原因. 因此, 无人机在一轮仿真实验中找到两个目标才称作完成任务. 任务完成率是用来刻画在不同算法提供的策略下, 无人机完整达成搜寻目标任务的有效性. 因此, 该指标从另一个方面帮助我们回答本文所提出的 RQ1, 同时作为 RQ3 评价通信拒止影响程度的一方面纳入考量.

$$\mathrm{MCR} = \frac{\sum_{i=1}^{N} \mathrm{MC}_i}{N}, \tag{22}$$

表 4　测试环境得分设置

Table 4　Score settings

Missing result	Score
One target acquired	1
Both targets acquired	2
No target acquired	0
Each drone being destroyed	−0.1

其中 MCR 为任务完成率 (mission complete rate), N 为测试轮数, MC_i 为第 i 轮任务完成情况指标:

$$\mathrm{MC}_i = \begin{cases} 1, & \mathrm{TA}_i = 2; \\ 0, & \text{otherwise.} \end{cases}$$

从式 (22) 可以看出, MCR 的取值范围为 $\mathrm{MCR} \in [0,1]$, 且较高的任务完成率对应较高的算法有效性.

2.5.3　对抗得分

上述两个指标都适用于无对抗 (与随机算法对抗) 场景下, 但在仿真环境下, 无人机分为红蓝双方, 具有竞争关系, 存在红蓝双方相互攻击, 摧毁对方无人机的情况. 由此, 需要一个指标来衡量算法在对抗场景下进行目标搜寻任务的有效性, 即对抗得分, 对于某一轮测试 (i 代表轮次), 对抗得分 $S_i(a)$ 代表 a 方使用的算法在此轮测试中的得分.

进行一轮目标搜寻测试, 得分设置见表 4. 对抗得分是对找到目标和击毁敌军的量化表示, 用以比较算法在此类问题上的效果, 即通过算法在对抗环境中的得分来比较各个算法的表现. 以红方采用策略梯度算法为例, 若 $S_i(a_r) > S_i(a_b)$, 则判定红方使用的算法获胜, 代表红方算法在该轮测试中效果更好. 同时, 为了体现红蓝双方竞争的结果, 我们提出对抗获胜率, 用以表示一方无人机所使用的算法在目标搜寻问题上与另一方相比获得的优势大小. 对抗获胜率指标用于回答本文提出的 RQ2, 对抗获胜率越高, 该算法的优势越大, 效果越好.

$$\mathrm{VR}(a_r, a_b) = \frac{\sum_{i=1}^{N} V_i(a_r, a_b)}{N}, \tag{23}$$

其中 VR 为对抗获胜率 (victory rate), N 为测试轮数, a 指某方代表的算法, 在这里 a_r 和 a_b 分别指对抗双方代表的算法 (例如 a_r 为红方代表算法, a_b 为蓝方代表的算法, 则 $\mathrm{VR}(a_r, a_b)$ 为红方算法对蓝方算法的对抗获胜率), $V_i(a_r, a_b)$ 为第 i 轮 a_r 方与 a_b 方对抗获胜指标:

$$V_i(a_r, a_b) = \begin{cases} 1, & S_i(a_r) > S_i(a_b), \\ 0, & \text{otherwise,} \end{cases}$$

其中 $S_i(a_r)$, $S_i(a_b)$ 分别表示 a_r 方和 a_b 方在第 i 轮的得分.

从式 (23) 可以看出, $\mathrm{VR}(a_r, a_b)$ 的取值范围为 $\mathrm{VR}(a_r, a_b) \in [0,1]$, 且较高的对抗获胜率对应红方算法 (或蓝方) 在对抗场景下进行目标搜寻任务更有效.

2.5.4　平均任务完成时间

不同的算法完成任务时所需的探索时间是不同的, 一般情况下, 我们希望算法能够尽可能快地完成任务. 因此, 根据任务完成率 MCR 的定义, 我们提出了平均成功时间来衡量算法完成目标搜索任务

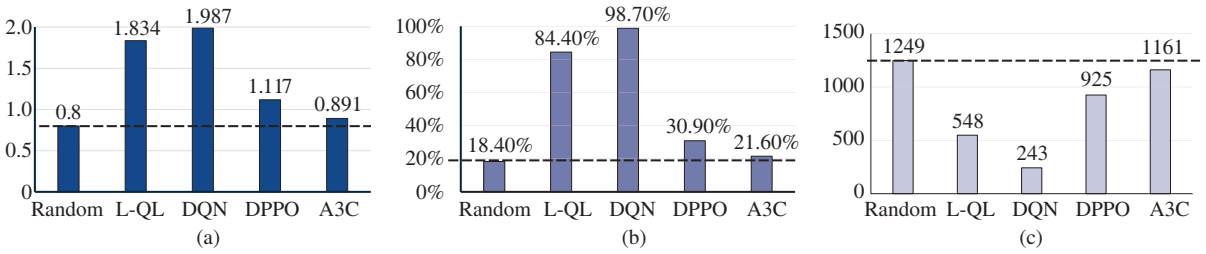

图 2 (网络版彩图) RQ1 实验结果

Figure 2 (Color online) RQ1 result. (a) The mean of target acquired; (b) mission complete rate; (c) the mean of mission complete time

的效率. 这里的成功是指成功完成任务, 即一轮测试搜寻到两个目标. 若进行 N 轮测试, 其中有 L 轮成功完成任务 ($L \leqslant N$):

$$\overline{\mathrm{MCT}} = \frac{\sum_{i=1}^{L} \mathrm{MCT}_i}{L}, \tag{24}$$

其中 $\overline{\mathrm{MCT}}$ 为 L 轮成功完成任务的测试中, 平均任务完成时间 (mission complete time) 的均值, MCT_i 为每轮测试成功完成任务所需探索时间. 由第 2.1 小节可知, 每轮测试无人机最大飞行时间为 T, MCT_i 的取值范围为 $\mathrm{MCT}_i \in [0, T]$, 平均成功时间 $\overline{\mathrm{MCT}}$ 越小, 则说明该算法完成目标搜寻任务的效率越高. 该指标可以帮助本文从算法性能角度回答 RQ1 和 RQ2, 同时作为分析 RQ3 影响程度的性能方面的评价指标.

3 实验结果和分析

基于以上实验设计, 我们以仿真环境中的实证研究方法对本文所提出的 3 个研究问题进行探索和分析.

3.1 RQ1: 强化学习方法在通信拒止战场目标搜索问题上的适用性

根据第 2.4 小节中的实验组织, 固定蓝方策略为结合避障规则的随机游走策略, 红方分别使用 DQN, L-QL, A3C, DPPO 4 种中的一种, 并使用结合避障规则的随机游走作为基准线进行比较. 4 种算法进行横向对比实验, 使用目标获取数量均值, 任务完成率和平均任务完成时间作为对比指标.

经过 1000 轮测试, 得到如图 2 的实验结果 (图中横轴代表无人机采用的算法, 从左至右分别为随机游走算法, L-QL, DQN, DPPO, A3C. 纵轴为每种算法在该指标上的具体数值. 以随机游走算法的各项指标作为基准线, 进行算法横向对比). 通过分析该实验结果, 我们有如下发现.

在图 2(a) 中, 所有强化学习方法的目标获取数量均值都大于随机游走算法, 且 DQN 算法的表现明显优于其他方法, $\overline{\mathrm{TA}}$ 值为 1.987, 接近平均每轮都能获取两个目标. 与表现最差的 A3C 算法相比, 高出超过一倍 (123.01%). L-QL 算法结果与 DQN 接近, 平均每轮可以获取 1.834 个目标. DPPO 算法结果与 A3C 相比稍有提升, $\overline{\mathrm{TA}}$ 为 1.117.

任务完成率与目标获取数量均值类似, 都用于刻画无人机在目标搜寻问题上的有效性, 但它们的差别在于, 任务完成率更注重于展现算法能否在一轮测试中找到所有的目标 (完成任务). 如图 2(b) 所示, 显然地, DQN 算法和 L-QL 算法在多轮测试中, 能较好地完成任务. 表现最好的 DQN 算法的任务完成率接近 100% (98.7%). 但 DPPO 算法和 A3C 算法表现较差, 且与前两者有明显差距. 在此项指

标中最低的算法 A3C, 其任务完成率约为 L-QL 算法的 1/4. 不过, 尽管 A3C 是表现最差的强化学习算法, 它仍然高于基准线随机游走策略.

图 2(c) 展现了不同算法在无人机目标搜寻问题上的性能差异. 平均成功时间越短, 该算法的性能表现越好, 完成任务的速度越快. 表现最好的 DQN 算法的平均成功时间远远低于其他算法, 接近第 2 名 L-QL 算法的 1/2. DPPO 和 A3C 算法的平均成功时间均接近随机游走算法, 其中较好的是 DPPO, 平均成功时间为 925 (单位时间).

综合来看, DQN 算法无论是在搜寻目标的效果上, 还是在算法性能上, 都高于其他算法, 并远超基准线随机游走策略. A3C 尽管是表现相对最差的强化学习算法, 但仍然在各项指标中都超过了基准线. 根据以上分析, 可以肯定地说, 强化学习方法在无人机目标搜寻问题上是适用的. 此外, 基于值函数的算法与基于策略的算法在实验结果上有较大差别, 前者的表现明显优于后者. 同类算法 (DQN 与 L-QL, DPPO 与 A3C) 实验结果相近. 产生这样结果的原因可能是基于策略的算法收敛性较差, 尤其是在方差较大时会导致梯度不稳定. DPPO 算法优于 A3C 算法, 可能是因为 DPPO 精心构造了梯度更新的方向, 使得模型表现能够单调提升, 防止了某些更新导致模型表现的崩盘式下滑. L-QL 算法与 DQN 算法的差异, 可能是来源于神经网络对于值函数的拟合效果比线性拟合更贴近.

对于上述现象可能的解释为: 由于本文所关注的战场对抗动态目标搜索问题具有非稳态环境和奖赏稀疏两个特性, 导致 A3C 等涉及策略梯度的算法难以获取大量成功的样本进行训练. 相对于 DQN 等基于 Q 函数估计的方法, 在本文所关注的场景中, 基于策略梯度的方法训练代价更高 [39]、更容易陷入局部最优 [40]. 为了验证上述假设, 我们对问题环境进行了简化: 只保留一架无人机, 并令目标静止不动, 使得环境从非稳态变为稳态环境, 同时提高获取成功样例的概率. 实验结果表明, 通过上述简化, 能够有效提高 A3C 算法的目标获取率.

综上所述, 针对 RQ1 强化学习方法在通信拒止战场目标搜索问题上的适用性, 我们的结论是肯定的. 同时, 本文的实验结果表明将 A3C 等涉及策略梯度的算法直接运用于求解非稳态、奖赏稀疏的问题仍存在较高的难度, 需要我们深入开展研究.

3.2 RQ2: 不同强化学习方法性能差异对比

针对第 2 个问题, 为了体现不同强化学习方法在解决本文所提出问题上的性能差异, 我们将不同的算法进行对抗.

首先, 按照 2.4 小节中的实验组织, 我们通过观察第 3.1 小节中的 4 个算法, 我们发现 A3C 算法在平均目标获取数量指标上相较于随机算法仅高 0.09%, 且任务完成率指标也仅仅高出 3.2%. 基于这一结果, 我们认为 A3C 算法在通信拒止环境下虽然展现出了一定的有效性, 但对比随机算法优势不明显. 因此, 我们选择 DNQ, L-QL 以及 DPPO 来开展本节的相关实验. 我们将上述 3 个模型训练后两两放入环境中进行对抗, 通过 2.5 小节中所设定的对抗获胜率、任务完成率与平均任务完成时间 3 个指标来说明他们在解决本文所提出问题上的性能差异.

对抗实验结果如表 5~7 所示, 第 i 行 j 列的数据表示 i 算法在与 j 算法对抗实验中所获得的成绩. 例如, DQN 在与 L-QL 对抗实验中的对抗获胜率为 72.2%, 任务完成率为 79.0%, 平均完成时间为 358.56 (步).

对抗获胜率如表 5 所示, 从表中我们可以看到, DQN 在面对 L-QL 或 DPPO 时的对抗胜率分别为 72.2% 与 90.1%. 由此我们可以看出 DQN 在面对 L-QL 与 DPPO 时均获得七成以上的胜率, 特别是在面对 DPPO 时获胜率高达 90% 以上. 而 L-QL 与 DPPO 在面对 DQN 时的对抗获胜率仅为 14.9% 与 5.1%, 在 L-QL 和 DPPO 的对抗实验中两种算法分别得到 42.7% 与 28.4% 的对抗获胜率. 因

表 5 对抗获胜率 (VR) (%)

Table 5 Victory rate (VR) (%)

	DQN	L-QL	DPPO
DQN	–	72.2	90.1
L-QL	14.9	–	42.7
DPPO	5.1	28.4	–

表 6 任务完成率 (MCR) (%)

Table 6 Mission complete rate (MCR) (%)

	DQN	L-QL	DPPO
DQN	–	49.1	72.0
L-QL	4.6	–	20.4
DPPO	0.7	2.6	–

表 7 对抗平均任务完成时间步数 ($\overline{\text{MCT}}$)

Table 7 Mission complete time steps ($\overline{\text{MCT}}$)

	DQN	L-QL	DPPO
DQN	–	358.56	297.79
L-QL	484.61	–	613.55
DPPO	905.14	683.58	–

此, 可以看出在对抗场景下, DNQ 相较于其他两种算法胜率表现更加优异, DPPO 表现最差.

由于对抗获胜的达成并不完全依赖于任务的完成, 所以为了了解 3 种算法在任务完成上的表现, 我们将任务完成次数在表 6 中列出. 可以直观地看到 DQN 表现最好, 其与 L-QL 和 DPPO 的对抗实验中任务完成率分别为 49.1% 与 72.0%, 这相较于 DQN 的对抗获胜率下降百分比分别为 32.0% 与 20.1%, 但对比 L-QL 与 DPPO 在面对 DQN 时任务完成率相较于对抗获胜率下降约 69.1% 与 86.3% 的情况来看, 我们可以得出 DQN 更加注重采取任务完成策略获得胜利而非通过对抗方式获得胜利这一结论.

表 7 中展示了 3 种算法的平均任务完成时间实验结果, 平均任务完成时间越低越说明算法在获取目标上的效率越高. 从中我们可以得到与上述对抗获胜率和任务完成率类似的结论, DNQ 的表现依旧是 3 个算法中最优异的. 在与 L-QL 与 DPPO 的对抗实验中平均成功完成时间是最低的, 不过, 我们也可以看到仅 L-QL 在面对 DQN 时的平均任务完成时间减少, 其他算法在对抗实验中的平均任务完成时间都增加, 由此我们也可以看出在对抗实验中算法完成任务效率有所下降.

综合上述实验结果, 我们可以发现, DQN 在 3 项指标中都表现最好, L-QL 效果次之, DPPO 表现最差. 因此, 在已实验的几种强化学习算法中, 我们认为 DQN 的表现是最优的.

3.3 RQ3: 通信拒止程度对强化学习算法的影响

针对本文所涉及的第 3 个研究问题, 我们根据 2.4 小节所述组织实验. 在该实验中, 分别取通信成功率 $\omega \in \{0\%, 1\%, 2\%, 3\%, 4\%, 5\%, 10\%, 20\%, 30\%, 40\%, 50\%, 60\%, 70\%, 80\%, 90\%, 100\%\}$, 测试 1000 轮, 观察在不同的通信成功率下, DQN 算法分别在 $\omega = 40\%$ 和 $\omega = 100\%$ 时训练所得的最佳模型的任务完成率、平均成功时间和目标获取数量均值. 以此说明通信拒止程度对强化学习算法的影响, 对应的

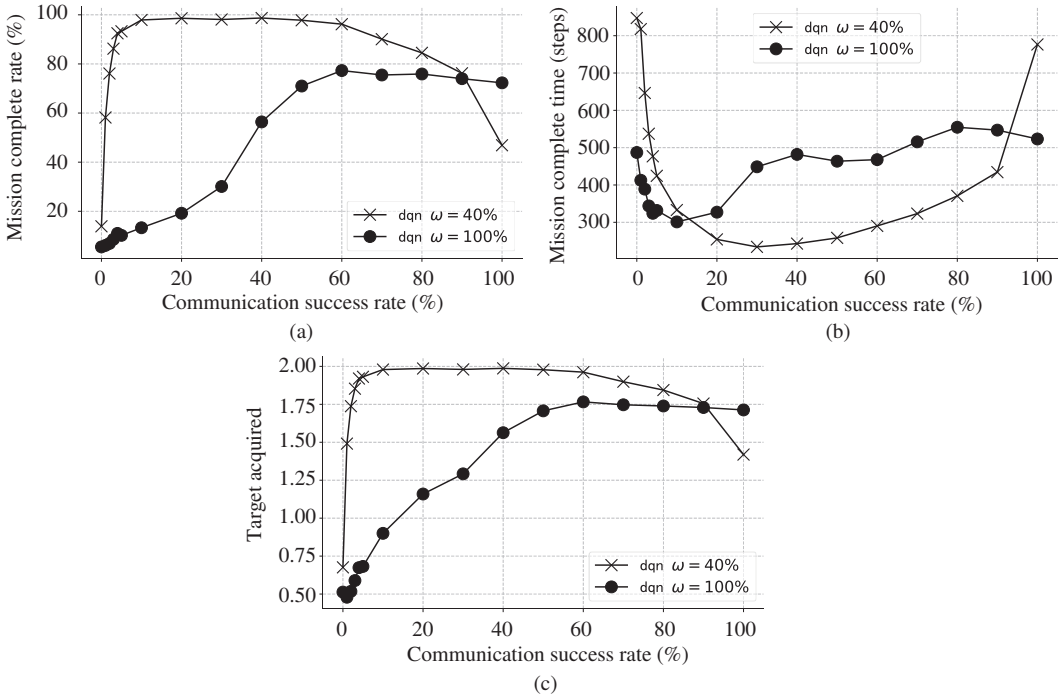

图 3 RQ3 实验结果

Figure 3 RQ3 result. (a) Mission complete rate; (b) the mean of mission complete time; (c) the mean of target acquired

实验结果如下所述.

通信拒止程度对任务完成率的影响如图 3(a) 所示, 从图中可以看出: 任务完成率在测试时通信拒止程度与训练时通信拒止程度接近时取得最高值. 对于使用 $\omega = 40\%$ 训练时所得的最好模型, 在 40% 附近任务完成率接近 100%, 当通信成功率偏离 40% 时, 任务完成率逐步降低. 但可以观察到, 在 $[3\%, 80\%]$ 区间内, 该模型的任务完成率都超过了 80%, 说明 DQN 算法在通信拒止程度变化时具有一定的稳定性. 在通信成功率低于 10% 时, 任务完成率下降特别迅速, 表示非常低的通信成功率对算法影响十分剧烈.

当通信成功率上升到接近 100% 时, 任务完成率反而下降, 这是一个反直觉的结果, 该现象可能是因为在高的通信成功率下, 引入了相当多的训练时从未经历过的状态, 导致算法效果变差. 由于这一结果的异常表现, 我们在 100% 通信成功率下训练获得了一个模型, 在通信成功率 100% 附近测试获得了比 $\omega = 40\%$ 时训练出的模型更好的结果, 但没有达到接近 100% 任务完成率的程度. 该模型在更低的通信成功率下表现更加糟糕, 说明此时模型更加依赖通信. 因此, 在训练时, 我们就应当使用与真实拒止环境相似的参数, 否则结果可能会不理想.

通信拒止程度对平均任务完成时间的影响如图 3(b) 所示, 从图中可以看出: 平均成功时间在通信拒止程度与算法训练时预设的通信拒止程度接近时取得最低值, 也就是说, 此时能够以尽可能快的速度完成任务. 当训练与测试时所使用的通信拒止程度类似时, DQN 算法在平均成功时间这一指标上表现最好, 随着通信成功率偏离训练时设置的值, 性能会出现一定程度的下降.

通信拒止程度对目标获取数量均值影响如图 3(c) 所示, 从图中可以看出: 与任务完成率结果类似, 该指标的结果也表明, 在通信拒止程度与训练时预设的通信拒止程度较为接近时取得较好的结果. 该指标与任务完成率存在比较强的相关性, 但是该指标变化相对平缓, 因为通信成功率对是否能完成任

务影响较大, 而即使不能完成任务, 只能获取一个目标的情形在该指标上也能够体现出来.

针对上述结果的一个可能解释是: 当训练和测试时的通信成功率相匹配时, 强化学习算法能够取得较好的效果. 而一旦测试时的通信成功率偏离训练过程中的设置时 (即使是通信成功率从 40% 上升到 100% 的情形), 由于环境的动态性发生改变, 强化学习算法能够获得的状态数据与训练时区别较大, 分布不一致, 算法性能会出现明显下降, 但仍然具有一定的有效性. 因此, 我们在构建仿真环境时应当尽可能接近真实环境. 同时, 针对现有的强化学习框架无法很好地应对环境特性变化这一难题, 需要我们开展更加深入的研究.

通过上述实验结果, 我们认为通信拒止程度会影响强化学习算法. 并且, 强化学习算法偏好训练时预设的通信拒止程度, 当通信拒止程度与训练时预设值相比发生变化时, 强化学习效果变差.

4 相关工作

围绕拒止环境下无人机协作对抗目标搜寻问题, 本文调研了大量的国内外相关工作. 目前, 大部分无人机相关工作聚焦于研究无人机相关应用, 以及无人机集群相关的多智能体问题. 无人机集群, 应用在不同领域中, 包含以下相关问题: 无人机区域覆盖问题、无人机避敌 (避障) 问题、无人机路径规划问题, 以及本文研究的无人机目标搜寻问题. 与此相关, 多架无人机之间如何进行通信、协作, 也是研究热点.

传统的无人机控制和决策采用基于规则的方式, 但随着机器学习的发展, 以及强化学习近年来的突破, 有一部分工作将强化学习方法应用到无人机问题中, 获得了可观的成果. 我们将分别阐述调研的相关工作内容, 并简要分析本文工作与现有工作的不同之处.

区域覆盖 (area coverage) 问题. 对于该问题, 调研报告 [41] 称无人机集群为无人机网络, 将该问题定义为, 对于给定的空间, 无人机通过协作监控整个网络的程度. 该文章将该问题下的工作分为几类, 包括对于覆盖能力、协作能力等一系列的研究. 重点为对于文献的综述, 与本文不同的是, 我们采取实验验证的方式, 对无人机目标搜寻问题进行调研, 并据此得出结论. 这两个问题的差别就在于, 本文研究的问题与运动的目标相关, 核心是为每个无人机都找到一条通向目标的路径, 并且避免障碍物, 而不是进行单纯的区域覆盖.

在较早时期, 一部分研究者采用区域分解的方法 [42] 解决这一问题, 并在分解的区域中融合了路径规划的思想, 通过无人机的飞行路径进行区域覆盖. 但与传统的路径规划问题不同的是, 路径设计的核心是尽可能地覆盖多的区域, 而不是传统意义上的最短路径. 近几年, 也有工作在此领域中取得了不错的进展. Pham 等 [22] 提出将多智能体强化学习 (multi-agent reinforcement learning) 方法引入无人机区域覆盖问题, 并给出了多智能体强化学习方法解决该问题的实例.

多智能体强化学习方法在无人机集群相关研究中有诸多应用, 并可以解决不同问题. 多智能体强化学习方法的调研报告 [43] 认为, 多智能体强化学习的目标有两类, 包括: 智能体学习动态的稳定性, 以及对其他智能体行为变化的适应. 并描述了典型的强化学习算法, 给出了多智能体强化学习的优点和应用领域. 据此, 有许多研究者将该方法与他们的研究问题相结合, 以找到新的解决思路.

避撞 (obstacle avoidance) 和避障 (collision avoidance) 问题. 如何避免与障碍物碰撞, 或被敌方捕捉, 是研究机器人或无人机集群应用的常见问题. 一般来说, 机器可以通过自身携带的传感器观察自身周围情况. 但在集群中, 障碍物未知时, 该问题难以通过简单的障碍物检测解决.

针对没有通信的无人机组, Fasano 等 [44] 提出了通过改进传感器避免碰撞的方法. La 等 [45] 采用多智能体学习的方法, 通过协作的方式有效避敌. 但在现实情况下, 机器之间的通信往往受到外界条

件的干扰, 导致多智能体协作受限, 因此避障方法也要随之改进. 在 2017 年, 上述作者提出了在通信受限的情况下, 多机器人避障的方法. 该研究组将多智能体强化学习方法加以改进, 应用到避障问题, 并取得了不错的进展. 这也为我们提供了强化学习方法对于本文问题的适用性佐证.

无人机路径规划 (path planning) 问题. 在简单的无人机路径规划问题中, 常使用动态规划方法等算法, 或者将不同的算法进行结合 [46]. 但当地图变得复杂或动态变化时, 这些方法的适用性有待考量. 为了解决复杂地图中的路径规划问题, 研究者们引入了多种强化学习方法, 包括几何强化学习 [23]、普通强化学习 [24]、多智能体强化学习. 这些方法都为解决该问题作出了一定的贡献. 与本文的工作相比, 本文虽然在某种程度上也是寻找路径, 但由于目标的变化, 路径的变化更加复杂. 并且本文的核心不在于用某一种方法, 而是衡量多种方法的适用性.

5 总结

本文针对强化学习方法在通信拒止战场条件下多无人机目标搜寻问题上的适用性问题开展了研究. 通过建模通信拒止战场环境并构建相应的仿真平台, 本文针对 3 个具体研究问题开展了深入探讨. 通过调研现有的主流强化学习方法并开展系统性的实验, 我们获取了丰富的实验结果并进行了定量分析.

实验结果表明, 现有主流的强化学习方法无论是在目标搜寻任务的完成上, 还是在完成任务的开销上, 都表现出了比随机游走策略更好的效果. 因此, 本文认为强化学习方法在通信拒止战场中多无人机目标搜寻问题上具有适用性. 第二, 根据将强化学习方法进行两两对抗的实验结果, 从对抗获胜率和平均成功时间两个指标上, DQN 方法在与其余算法对抗时更能获得目标, 且花费时间更少, 具有较大的优势. 因此 DQN 方法在我们选择的强化学习方法中表现最优. 最后, 通过对比两组在不同通信成功率下的 3 种指标变化曲线, 我们发现在一定范围内通信拒止程度与算法效果基本呈负相关. 但在该范围外, 越靠近训练时的通信成功率, 越容易达到最优值. 以上实验结果能够为后续的相关研究提供帮助和借鉴. 同时, 本文建立了一套高效、低成本的仿真平台, 能够为进一步开展研究提供有力支撑.

作为运用强化学习算法解决通信拒止战场环境中多无人机目标搜寻问题的初步尝试, 仍然有许多问题需要我们进一步开展深入分析和探索: (1) 是否还有其他在该问题上效果更好的算法? 尽管我们列举了当前较为前沿的主流算法, 但并不能穷尽强化学习中的所有算法, 也未对这些算法在所研究的特定问题上进行深入的优化. 因此, 如何针对特定问题优化算法设计, 是未来的主要工作方向之一. (2) 仿真环境是否能模拟得更贴近现实? 目前, 我们在对现有文献的综述调研的基础上, 提出了通信拒止战场环境的抽象模型, 并实现了对应的仿真环境. 在实际战场环境下, 可能还存在对强化学习方法、多无人机集群具有重要影响的要素未能在目前的仿真环境中得到体现. 需要我们进一步通过虚实结合的方式, 构建物理环境和仿真环境相结合的实验平台, 在保障平台高效、低成本的基础上, 尽可能地贴近真实物理环境. (3) 强化学习算法的有效性和受到包括通信拒止程度等因素的影响的机制如何解释? 目前, 我们仅能通过观察实验结果来判断强化学习算法在问题解决上的有效性, 以及环境中包含通信拒止程度在内的各种变量对不同算法的影响程度. 但在算法体现有效性的机理、算法受到环境因素影响的根本原因、未来不可预知的环境因素改变对算法有效性的影响等问题上, 尚不能形成有效的解释和预测, 需要我们重点开展研究.

参考文献

1 Tomic T, Schmid K, Lutz P, et al. Toward a fully autonomous UAV: research platform for indoor and outdoor urban search and rescue. IEEE Robot Automat Mag, 2012, 19: 46–56

2 Doherty P, Rudol P. A UAV search and rescue scenario with human body detection and geolocalization. In: Proceedings of Australasian Joint Conference on Artificial Intelligence, 2007. 1–13

3 Li C C, Zhang G S, Lei T J, et al. Quick image-processing method of UAV without control points data in earthquake disaster area. Trans Nonferrous Met Soc China, 2011, 21: 523–528

4 Ryan A, Zennaro M, Howell A, et al. An overview of emerging results in cooperative UAV control. In: Proceedings of 2004 43rd IEEE Conference on Decision and Control (CDC), 2004. 602–607

5 Chmaj G, Selvaraj H. Distributed processing applications for UAV/drones: a survey. In: Proceedings of Progress in Systems Engineering, 2015. 449–454

6 Srinivasan S, Latchman H, Shea J, et al. Airborne traffic surveillance systems: video surveillance of highway traffic. In: Proceedings of the ACM 2nd International Workshop on Video Surveillance & Sensor Networks, 2004. 131–135

7 O'Young S, Hubbard P. Raven: a maritime surveillance project using small UAV. In: Proceedings of 2007 IEEE Conference on Emerging Technologies and Factory Automation (EFTA 2007), 2007. 904–907

8 Xu X W, Lai J Z, Lv P, et al. A literature review on the research status and progress of cooperative navigation technology for multiple UAVs. Navigation Positioning Timing, 2017, 4: 1–9 [许晓伟, 赖际舟, 吕品, 等. 多无人机协同导航技术研究现状及进展. 导航定位与授时, 2017, 4: 1–9]

9 Li L, Wang T, Hu Q L, et al. White force network in DARPA CODE program. Aerospace Electron Warfare, 2018, 34: 54–59 [李磊, 王彤, 胡勤莲, 等. DARPA 拒止环境中协同作战项目白军网络研究. 航天电子对抗, 2018, 34: 54–59]

10 Chen J. Analysis of the bee colony technology of low cost UAV in the US Navy. Aerodynamic Missile J, 2016, 1: 24–26 [陈晶. 解析美海军低成本无人机蜂群技术. 飞航导弹, 2016, 1: 24–26]

11 Duan H B, Li P. Autonomous control for unmanned aerial vehicle swarms based on biological collective behaviors. Sci Tech Rev, 2017, 35: 17–25 [段海滨, 李沛. 基于生物群集行为的无人机集群控制. 科技导报, 2017, 35: 17–25]

12 Wu Y J. Research on interference technology to GPS signal. Aerospace Electron Warfare, 2002, 3: 22–25 [武拥军. 对 GPS 信号的干扰技术研究. 航天电子对抗, 2002, 3: 22–25]

13 Balamurugan G, Valarmathi J, Naidu V P S. Survey on UAV navigation in GPS denied environments. In: Proceedings of International Conference on Signal Processing, 2017

14 Cesare K, Skeele R, Yoo S-H, et al. Multi-UAV exploration with limited communication and battery. In: Proceedings of 2015 IEEE International Conference on Robotics and Automation (ICRA), 2015. 2230–2235

15 Gupta L, Jain R, Vaszkun G. Survey of important issues in UAV communication networks. IEEE Commun Surv Tut, 2016, 18: 1123–1152

16 Duan H B, Zhang D F, Fan Y M, et al. From wolf pack intelligence to UAV swarm cooperative decision-making. Sci Sin Inform, 2019, 49: 112–118 [段海滨, 张岱峰, 范彦铭, 等. 从狼群智能到无人机集群协同决策. 中国科学: 信息科学, 2019, 49: 112–118]

17 Silver D, Hubert T, Schrittwieser J, et al. Mastering chess and shogi by self-play with a general reinforcement learning algorithm. 2017. ArXiv: 1712.01815

18 Silver D, Schrittwieser J, Simonyan K, et al. Mastering the game of Go without human knowledge. Nature, 2017, 550: 354–359

19 Vinyals O, Ewalds T, Bartunov S, et al. Starcraft II: a new challenge for reinforcement learning. 2017. ArXiv: 1708.04782

20 Wang Y S, Tong Y X, Long C, et al. Adaptive dynamic bipartite graph matching: a reinforcement learning approach. In: Proceedings of 2019 IEEE 35th International Conference on Data Engineering (ICDE), 2019. 1478–1489

21 Beard W R. Multiple UAV cooperative search under collision avoidance and limited range communication constraints. In: Proceedings of IEEE Conference on Decision & Control, 2003

22 Pham H X, La H M, Feil-Seifer D, et al. Cooperative and distributed reinforcement learning of drones for field coverage. 2018. ArXiv: 1803.07250

23 Zhang B C, Mao Z L, Liu W Q, et al. Geometric reinforcement learning for path planning of UAVs. J Intell Robot Syst, 2015, 77: 391–409

24 Zeng Y, Xu X L. Path design for cellular-connected UAV with reinforcement learning. 2019. ArXiv: 1905.03440v1

25 Ghavamzadeh M, Mahadevan S, Makar R. Hierarchical multi-agent reinforcement learning. Auton Agent Multi-Agent Syst, 2006, 13: 197–229
26 Sutton R S, Precup D, Singh S. Between MDPs and semi-MDPs: a framework for temporal abstraction in reinforcement learning. Artif Intell, 1999, 112: 181–211
27 Kaelbling L P, Littman M L, Moore A W. Reinforcement learning: a survey. J Artif Intell Res, 1996, 4: 237–285
28 Sutton R S, Barto A G. Reinforcement Learning: An Introduction. Cambridge: MIT Press, 1998
29 Watkins C J C H, Dayan P. Q-Learning. Mach Learn, 1992, 8: 279–292
30 Mnih V, Kavukcuoglu K, Silver D, et al. Human-level control through deep reinforcement learning. Nature, 2015, 518: 529–533
31 Schaul T, Quan J, Antonoglou I, et al. Prioritized experience replay. 2015. ArXiv: 1511.05952
32 Tieleman T, Hinton G. Lecture 6.5-rmsprop: divide the gradient by a running average of its recent magnitude. COURSERA: Neural Netw Mach Learn, 2012, 4: 26–31
33 Melo F S, Ribeiro M I. Q-Learning with linear function approximation. In: Proceedings of International Conference on Computational Learning Theory, 2007. 308–322
34 Rahimi A, Recht B. Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. In: Proceedings of Advances in Neural Information Processing Systems, 2009. 1313–1320
35 Mnih V, Badia A P, Mirza M, et al. Asynchronous methods for deep reinforcement learning. In: Proceedings of International Conference on Machine Learning, 2016. 1928–1937
36 Williams R J, Peng J. Function optimization using connectionist reinforcement learning algorithms. Connection Sci, 1991, 3: 241–268
37 Schulman J, Wolski F, Dhariwal P, et al. Proximal policy optimization algorithms. 2017. ArXiv: 1707.06347
38 Abbeel P, Schulman J. Deep reinforcement learning through policy optimization. In: Proceedings of Tutorial at Neural Information Processing Systems Conference, 2016
39 Haarnoja T, Zhou A, Abbeel P, et al. Soft actor-critic: off-policy maximum entropy deep reinforcement learning with a stochastic actor. 2018. ArXiv: 1801.01290
40 Grondman I, Busoniu L, Lopes G A D, et al. A survey of actor-critic reinforcement learning: standard and natural policy gradients. IEEE Trans Syst Man Cybern C, 2012, 42: 1291–1307
41 Chen Y Y, Zhang H D, Xu M. The coverage problem in UAV network: a survey. In: Proceedings of the 5th International Conference on Computing, Communications and Networking Technologies (ICCCNT), 2014. 1–5
42 Araujo J, Sujit P, Sousa J B. Multiple UAV area decomposition and coverage. In: Proceedings of 2013 IEEE Symposium on Computational Intelligence for Security and Defense Applications (CISDA), 2013. 30–37
43 Busoniu L, Babuska R, de Schutter B. A comprehensive survey of multiagent reinforcement learning. IEEE Trans Syst Man Cybern C, 2008, 38: 156–172
44 Fasano G, Accardo D, Moccia A, et al. Multisensor based fully autonomous non-cooperative collision avoidance system for UAVs. J Aerosp Comput Inf Commun, 2008, 5: 338–360
45 La H M, Lim R, Sheng W. Multirobot cooperative learning for predator avoidance. IEEE Trans Contr Syst Technol, 2015, 23: 52–63
46 Li J Q, Deng G Q, Luo C W, et al. A hybrid path planning method in unmanned air/ground vehicle (UAV/UGV) cooperative systems. IEEE Trans Veh Technol, 2016, 65: 9585–9596

Feasibility of reinforcement learning for UAV-based target searching in a simulated communication denied environment

Liang WANG*, Wen WANG, Yuyou WANG, Songlin HOU, Yuzhe QIAO,
Tianheng WU & Xianping TAO*

State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023*, China*
* Corresponding author. E-mail: wl@nju.edu.cn, txp@nju.edu.cn

Abstract Target searching is crucial in real-world scenarios such as search and rescue in disaster sites and battlefield target reconnaissance. Unmanned aerial vehicles (UAVs) are an ideal technical solution for target searching in large-scale and high-risk areas because they are agile, low cost, and able to collaborate and carry different sensors. In complex scenarios like battlefields, due to the lack of communication infrastructures and the intensive interference, UAVs often operate in communication denied environments. As a result, fast and reliable communication channels between UAVs and ground operators are difficult to establish. Thus, in such conditions, UAVs must be able to complete tasks autonomously and intelligently, without receiving real-time commands from the operators. With the rapid advances in artificial intelligence, reinforcement learning has shown potentiality for solving continuous decision problems. The target searching problem studied in this paper falls into this category and is suitable for adopting reinforcement learning technologies. However, the feasibility of reinforcement learning in UAV-based target searching in communication denied environments is not clear and, thus, requires in-depth investigations. As a pilot study in this direction, this paper models the target searching problem in communication denied and confrontation situations and proposes a simulation environment based on this model. Extensive experiments are conducted to answer the following questions. (1) Can reinforcement learning be applied in target searching by multi-UAVs in communication denied environments? (2) What are the advantages and disadvantages of different reinforcement learning algorithms in solving this problem? (3) How the degree of communication denial influences the performance of these algorithms? The current mainstream reinforcement learning technologies are adopted to perform simulations, whose results are analyzed quantitatively, leading to the following observations. (1) Reinforcement learning can effectively solve target searching problems for multi-UAVs in communication denied environments. (2) Compared with other algorithms, an autonomous decision-making UAV cluster based on a deep Q-network (DQN)exhibits the best problem-solving ability. (3) The algorithm performance changes with the degree of communication denial but remains largely stable when the communication condition varies.

Keywords UAV, reinforcement learning, target searching, communication denied environments

Liang WANG received the B.S. and Ph.D. degrees in computer science from Nanjing University in 2007 and 2014, respectively. He is currently an assistant researcher in the State Key Laboratory for Novel Software Technology at Nanjing University. His research interests include pervasive and mobile computing, human activity recognition, and software methodology.

Wen WANG received the B.E. degree in computer science from Hohai University in 2019. He is currently a postgraduate student in the State Key Laboratory for Novel Software Technology at Nanjing University. His research interests include reinforcement learning and collective intelligence.

Yuyou WANG received the B.E. degree in software engineering from University of Electronic Science and Technology of China in 2018. She is currently a postgraduate student in the State Key Laboratory for Novel Software Technology at Nanjing University. Her research interests include operator placement for CEP and edge computing.

Xianping TAO received the Ph.D. degree in computer science from Nanjing University in 2001. He is currently a professor at the Department of Computer Science, Nanjing University. His research interests include software agents, middleware systems, Internetware methodology, and pervasive computing.