基于多智能体深度强化学习的无人机集群自主决策*

刘志飞，曹　雷，赖　俊，陈希亮
（陆军工程大学 指挥控制工程学院，江苏 南京 210007）

摘　要： 由于传统的无人机由人工进行操控，无人机群在强电磁干扰和复杂多变的战场环境中表现较为呆板。在这项研究中，开发了一种灵活智能的无人机控制器。通过使用一个经过多智能体深度强化学习技术训练的神经网络，无人机可以在飞行中控制自己的行为，从战场环境中获取状态信息，自主决策，并且和其他无人机形成有效战斗队形，灵活协调和配合，并产生了最优的动作。

关键词： 无人机；强化学习；多智能体；自主决策
中图分类号： TP181　　**文献标识码：** A　　**DOI：** 10.19358/j.issn.2096-5133.2022.05.012
引用格式： 刘志飞，曹雷，赖俊，等. 基于多智能体深度强化学习的无人机集群自主决策[J].信息技术与网络安全，2022，41(5)：77-81.

Utonomous decision making of UAV cluster with multi-agent deep reinforcement learning

Li Zhifei，Cao Lei，Lai Jun，Chen Xiliang
（College of Command and Control Engineering，Army Engineering University，Nanjing 210007，China）

Abstract： Because the traditional UAV is controlled manually，UAV cluster is more rigid in the strong electromagnetic interference and complex and changeable battlefield environment. In the study，a flexible and intelligent UAV controller is developed. With a neural network trained by multi-agent deep reinforcement learning technology，UAV can control his behavior in flight. At the same time，UAV obtains state information from the battlefield environment，makes independent decisions，forms an effective combat formation with other UAVs，flexibly coordinates and cooperates with each other，and produces the optimal action.

Key words： unmanned aerial vehicle；reinforcement learning；multi agent；autonomous decision

0 引言

对人工操纵无人机来说，同时操控多架无人机完成多项任务且无人机之间形成有效配合是相当困难的，注意力分散或者操控失误都会造成较大的安全风险。无人机的操控还受到电磁干扰和远程控制距离的限制，因此，无人机灵活自主决策能力显得尤为重要。近年来，多智能体深度强化学习(Multi-Agent Deep Reinforcement Learning，MADRL)在复杂游戏中取得完胜人类专家水平的胜利，表明多智能体深度强化学习在解决复杂序贯问题上取得重要突破。强化学习技术应用到无人机群可以提高无人机群的灵活智能性。本文以一个由6架无人机组成的无人机群为例，使用墨子AI仿真实验平台，无人机群组成一个巨大的动作空间，时间步内有200多个组合的动作空间，为每架无人机在每一步行为的机动方向、航线或向目标发出攻击都有提供了上千种选择。使用深度神经网络来预测每个无人机在每个时间步的最优动作，并根据每个无人机的局部观察产生自主决策。MADRL方法生成无人机群作战决策对无人机作战研究具有重要的参考价值，是未来人工智能应用在军事领域的重要方向。

1 无人机集群研究现状

无人机集群作战被公认为未来智能化战争的典型作战样式。由于无人机集群作战概念处在不断

* 基金项目：国家自然科学基金(61806221)

探索阶段，因而采用建模仿真方法对无人机集群作战的有关问题开展研究，为这一新型作战样式的发展提供理论支撑。当前，无人机集群研究面临的挑战有：

(1)无人机集群的个体行为刻画简单。鱼群算法、蚁群算法等源自对生物界集群行为观察，其规则简单，涉及群体智能的涌现。但是随着人工智能发展和计算机算力的提高，深度学习得到进一步发展，神经网络的数据拟合能力得到极大提升，具备不同任务能力的智能个体组成的异构集群成为重要发展趋势，由多智能体组成的智能集群将具有较高的智能水平。

(2)无人机集群协同作战研究不足。目前无人机集群作战建模与仿真研究中，大多只针对单一机型和单一简单任务，而实际作战中则需要不同功能类型的无人机组成的异构集群协同完成整体作战任务。

(3)仿真无人机不具备自主决策的能力。现有无人机集群建模研究大多采用规则方法，该方法通常采用 If-then 式的反应结构来表达无人机个体的行为决策，这种方法难以适应未来战场复杂多变的环境

2 强化学习

近年来，深度强化学习(Deep Reinforcement Learning，DRL)[1]取得显著成绩，这导致了强化学习的应用场景和方法与日俱增。最近的研究从单智能体发展到多智能体系统。尽管在多智能体领域面临诸多挑战，但深度强化学习在一些相对复杂的游戏领域取得了许多成功，如围棋[2-3]、扑克[4-5]、DOTA2[6]和星际争霸(StarCraft)[7]。这些领域的成功都依赖于强化学习(Reinforcement Learning，RL)和深度学习(Deep Learning，DL)两个技术的组合。

2.1 单智能体强化学习

强化学习是一项通过不断试错来学习的技术。智能体通过一系列的步数与环境进行交互，在每一步上基于当前的策略来获取环境状态，到达下一个状态并获得该动作奖励，智能体的目标是更新自己的策略以最大化累计奖励。如果环境满足马尔可夫性质(Markov Decision Process，MDP)[8]，强化学习可以建模为一个马尔可夫决策过程，如式(1)所示。

$$P(s_{t+1}|s_t,s_{t-1},s_{t-2},\cdots,s_0)=P(s_{t+1}|s_t) \tag{1}$$

其中 s_t 表示时间步 t 时的状态。

MDP 可以用式(2)来表示。

$$(S,A,R,\rho,\gamma) \tag{2}$$

其中 S 表示状态空间($s_t\in S$)，A 表示动作空间，$a_t\in A$，R 表示奖励空间($r_t\in R$)，ρ 表示状态转移矩阵($\rho_{SS'}=P[s_{t+1}=s'|s_t=s]$)，$\gamma$ 表示折扣因子，它用于表示及时奖励对未来奖励的影响程度。在深度学习中，有两个重要的概念：状态价值函数和动作价值函数。

状态价值函数用来衡量智能体所处状态的好坏，用式(3)表示：

$$V_\pi(s)=\sum_{a\in A}\pi(a|s)[r(s,a)+\gamma\sum_{i\in S}\rho(s'|s,a)V_\pi(s')] \tag{3}$$

动作价值函数用来衡量智能体采取特定动作的好坏，用式(4)表示。

$$Q_\pi(s,a)=r(s,a)+\gamma\sum_{s'\in S}\rho(s'|s,a)\sum_{a\in A}\pi(a|s')Q_\pi(s',a) \tag{4}$$

2.2 多智能体强化学习

深度学习已被应用于解决具有挑战性的问题，如从雅塔丽游戏到 Alpha Go、Alpha Zero、Alpha Star，再到无人驾驶和工业机器人。深度学习的大多数成功都集中在单智能体领域，建模或预测其他智能体行为在很大程度上是不必要的。然而，在许多实际应用中涉及多个智能体之间的交互，其中紧急行为和行为复杂性是由多个智能体共同作用产生的。例如，在多机器人控制、通信、多人游戏以及社会困境分析等领域中，多智能体自我博弈也是一种有用的训练方法。将单智能体扩展到多智能体系统中，对于构建能够与人类进行高效交互的人工智能系统至关重要。但是传统的强化学习方法(如 Q-Learning)和策略梯度算法不太适合多智能体环境。随着训练的进行，每个智能体的策略都在发生变化，环境的不确定性带来了学习稳定性的挑战，并且阻止了直接使用过去的经验回放。同时，当需要多个智能体协调时，策略梯度的方法通常表现出非常高的方差。

多智能体强化学习可以建模为分布式部分可观测马尔可夫决策过程(Dec-POMDP)[9]：一个完全合作的多智能体强化学习任务可以用分布式部分可观测马尔可夫决策过程(Dec-POMDP)来描述。Dec-POMDP 可由元组 $G=(n,S,U,P,r,Z,O,\gamma)$ 表示。其中 n 表示智能体的数量；$s\in S$ 表示状态；$u^a\in U$ 表示智能体的动作；$u^a\in U\equiv U^n$ 表示智能体的联合动作集合，$P(s'|s,u):S\times U\times S\rightarrow[0,1]$ 表示状态 s 下采取联合动作 u 转移到 s' 状态转移概率；$r(s,u):S\times U\rightarrow R$

表示奖励函数；$z \in Z$ 表示每个智能体的观察值由 $O(s,a):S\times A\rightarrow Z$ 来描述；$\gamma \in (0,1)$ 表示折扣因子。

2.3 强化学习方法在无人机群上的研究现状

深度学习在游戏领域取得巨大成功，将该方法应用到无人机操控方面的研究也越来越多。文献[10]提出了将单智能体深度学习方法应用到单个无人机的灵活操控上。文献[11]使用近端策略优化(Proximal Policy Optimization，PPO)算法对单个无人机的飞行姿态进行灵活的控制以应对复杂恶劣的环境。文献[12]提出使用深度学习方法对无人机在陌生环境中进行导航。文献[13]提出一种基于深度学习的城市无人机，其在在线和离线状态下均能规划出较优路径。目前研究热点在集中在基于深度学习的单个无人机的操控上，基于 MADRL 的无人机群的研究还比较少。基于 MADRL 方法应用到无人机群上主要面临状态动作空间维度灾难、环境不稳定性和信用分配的挑战。

3 无人机集群作战建模

3.1 无人机集群的强化学习建模

采用 MADRL 方法[14]对无人机集群作战进行建模，可认为是在连续状态空间上的及时决策过程，其遵循马尔可夫过程，用以下五元组形式表示：

$$M=\{S,A_1,\cdots A_i,\cdots A_n,R_1,\cdots R_i,\cdots R_n,T,\gamma\}$$

其中，n 表示无人机的个数；i 表示无人机的编号下标；A_i 表示第 i 个无人机的动作空间；R_i 表示无人机 i 在执行动作 A_i 后获得的及时回报；T 表示状态转移函数：$S\times A_1\times\cdots\times A_n\rightarrow S'$；$\gamma$ 表示折扣率，$\gamma\in(0,1)$。N 个智能体的 POMDP 由所有智能体的组成一组状态 S，一组动作 $A_1,\cdots A_i,\cdots A_n$ 和每个智能体的一组观测 $O_1,\cdots O_i,\cdots O_n$ 来定义。为了选择动作，每个无人机 i 使用随机策略 $\pi_{\theta_i}:O_i\times A_i\rightarrow[0,1]$，其根据状态转移函数 $S\times A_1\times\cdots\times A_n\rightarrow S'$ 产生下一个状态。每个智能体 i 获得作为该状态和智能体的动作的奖励函数 $r_i:S\times A_i\rightarrow R$ 的奖励，并且接收与状态相关的观察 $O_i:S\rightarrow O_i$。初始状态由分布：$S\rightarrow[0,1]$ 确定。

N 个无人机在做出联合动作 $A_1,\cdots A_i,\cdots A_n$ 后从环境中获取奖励 $R_1,\cdots R_i,\cdots R_n$。在 POMDP 中，无人机集群的目标是学习到最优联合策略，即最大化整体奖励。本文采用墨子 AI 的实验环境，实验场景如图 1 所示。

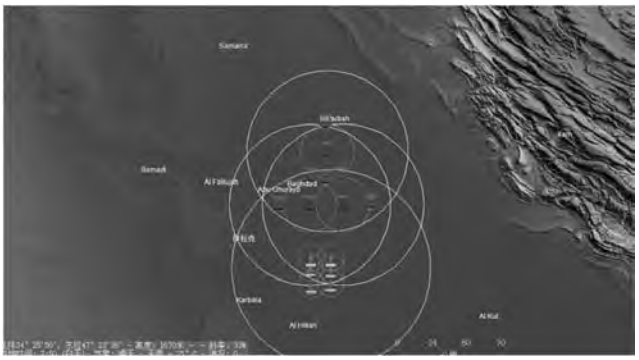

图 1 墨子实验场景

其中红方由 6 架灰鹰无人机组成无人机集群，使用 MADRL 算法进行自主决策行动。蓝方由 6 个坦克排(T-72 型主战坦克×4)和 3 个地空导弹排(萨姆-22“灰狗”)组成，使用固定战术规则。红方无人机群的任务是在最短的时间内避开雷达找到地方坦克并有效催化目标。

3.2 奖励函数设计

当无人机击毁一个坦克排，获取及时奖 $R=1$，当无人机进入蓝方地空导弹防御范围并被击毁，获取奖励 $R=-1$，给予惩罚。当无人机互相碰撞时给予奖励 $R=-0.1$，为了引导无人机更快地学习到最优攻击策略，设计连续性函数引导无人机到达预先设定的区域，距离值越小获得的奖励越多。无人机集群的作战任务最优策略是避开蓝方地空导弹，保存自身实力并摧毁蓝方全部坦克排。

3.3 训练流程

训练伪代码如下：

1. For episode=1 to M do
2. 初始化一个随机过程以便进行动作探索
3. 接收初始状态
4. for t=1 to max-episode-length do：
5. 对于每个无人机 i，根据策略网络加噪声采样动作 $a_i=\mu_{\theta_i}(O_i)+N_t$
6. 执行联合动作 $a_1,\cdots a_i,\cdots a_n$ 获得奖励和下一个状态 S'
7. 将 (x,a,r,x') 存入经验回放集 D 中
8. $S'x'$ 赋值给 x
9. for 无人机 i=1 to N do
10. 从经验回放集 D 中采集 s 个 mini 批的数据 x^j,a^j,r^j,x'^j
11. 设置联合动作值函数为：$y^j=r_i^j+\gamma Q_i^{\mu'}(x'^j,a_1',\cdots,a_i')$

12. 最小化损失函数 $L(\theta_i)=\frac{1}{S}\sum_j (y^j-Q_i^{\mu}\cdot (x^j,a_1^j,\cdots,a_n^j))^2$ 来更新评论家网络

13. 更新演员策略网络：

$$\nabla_{\theta_i} J=\frac{1}{S}\sum_j \nabla_{\theta_i}\mu^i(O_i^j)\nabla_{a_i} Q_i^j(x^j,a_1^j,\cdots,a_i^j,\cdots a_N^j)$$

14. 更新每个无人机的目标网络参数：

$$\theta_i^{'}\leftarrow\tau\theta_i+(1-\tau)\theta_i^{'}$$

15. end for

16. end for

4 实验设计及结果分析

实验环境如下：

仿真环境：墨子 AI 实验平台；

硬件环境：Windows 10 操作系统；

CPU：酷睿 i5 处理器；

内存容量：16 GB；

开发语言：Python3.7。

实验结果：经过 500 轮的训练，红方无人机逐渐学会了最优策略，能够在最快的时间避开雷达到达目标对蓝方坦克进行有效的攻击，红方无人机集群到达目标并摧毁敌方目标的成功率逐渐提高。如图 2 所示，红方无人机集群能有效地完成作战任务。损失函数曲线如图 3 所示。

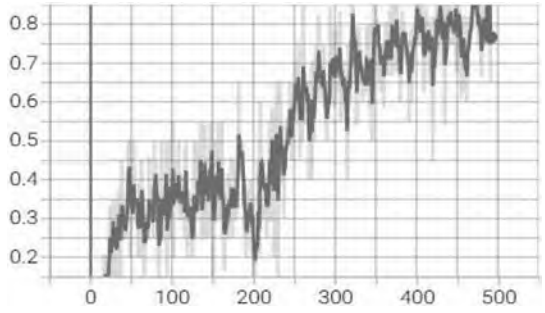

图 2 无人机群到达目标并击毁敌方坦克的成功率

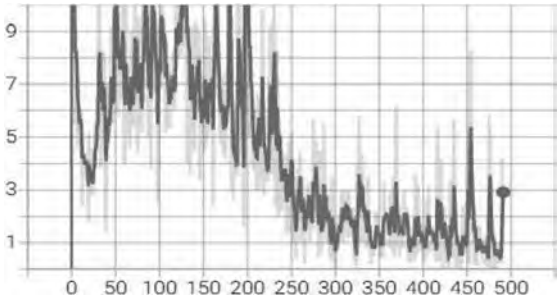

图 3 损失函数曲线

通过 MADRL 方法进行仿真训练，无人机集群可以学习到三种战术：

(1)6 架无人机集中优势兵力采取编队飞行从左侧依次对蓝方进行攻击，如图 4 所示。

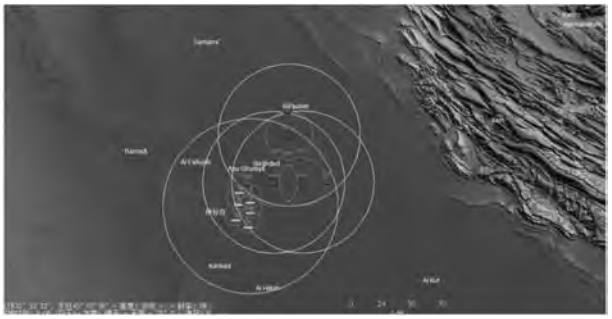

图 4 无人机群保持队形从左路飞行

(2)6 架无人机集中优势兵力采取编队飞行从右侧依次对蓝方进行攻击，如图 5 所示。

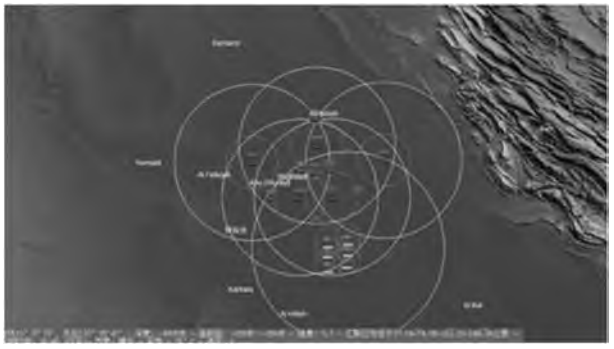

图 5 无人机群保持队形从右路飞行

(3)6 架无人机兵分两路从左右两侧各三架采取编队飞行依次对蓝方进行攻击，如图 6 所示。

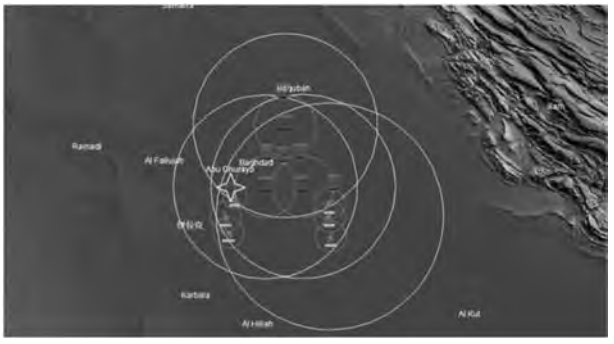

图 6 无人机群保持队形从两侧包抄

5 结论

本文采用多智能深度强化学习的技术，通过最

先进的MADRL算法对无人机集群行动进行了建模，并在墨子AI试验平台进行了测试。测试证明，无人机可以在战场环境中获取状态信息，产生最优动作，并作出自主决策，为无人机集群提供灵活的飞行控制，并在遂行任务中开展协调和配合。未来将在无人机上安装训练好的神经网络控制器，在实际场景中再度进行训练和试验。此项研究成果提供了一种无人驾驶集群化的飞行控制方式，在医疗、农业、安全等不同领域都具有应用价值。

参考文献

[1] HERNANDEZ-LEAL P, KARTAL B, TAYLOR M E. A survey and critique of multiagent deep reinforcement learning[J]. Autonomous Agents and Multi-Agent Systems, 2019, 33(6): 750-797.

[2] SILVER D, HHUANG A, MADDISON C J, et al. Mastering the game of Go with deep neural networks and tree search[J]. Nature, 2016, 529(7587): 484-489.

[3] SILVER D, SCHIRITTWIESER J, SIMONYAN K, et al. Mastering the game of Go without human knowledge[J]. Nature, 2017, 550(7676): 354-359.

[4] MORAVČÍK M, SCHMID M, BURCH N, et al. Deepstack: expert-level artificial intelligence in heads-up no-limit poker[J]. Science, 2017, 356(6337): 508-513.

[5] BROWN N, SANDHOLM T. Superhuman AI for heads-up no-limit poker: libratus beats top professionals[J]. Science, 2018, 359(6374): 418-424.

[6] BERNER C, BROCKMAN G, CHAN B, et al. Dota 2 with large scale deep reinforcement learning[J]. arXiv preprint arXiv: 1912.06680, 2019.

[7] VINYALS O, EWALDS T, BARTUNOV S, et al. StarCraft II: a new challenge for reinforcement learning[J]. arXiv preprint arXiv: 1708.04782, 2017.

[8] FILAR J, VRIEZE K. Competitive Markov decision processes[M]. Springer Science & Business Media, 2012.

[9] OLIEHOEK F A, SPAAN M T J, VLASSIS N. Optimal and approximate Q-value functions for decentralized POMDPs[J]. Journal of Artificial Intelligence Research, 2008, 32: 289-353.

[10] KHAN F S, MOHD M N H, LARIK R M, et al. A smart flight controller based on reinforcement learning for unmanned aerial vehicle (UAV)[C]//2021 IEEE International Conference on Signal and Image Processing Applications (ICSIPA). IEEE, 2021: 203-208.

[11] KOCH W, MANCUSO R, WEST R, et al. Reinforcement learning for UAV attitude control[J]. ACM Transactions on Cyber-Physical Systems, 2019, 3(2): 1-21.

[12] PHAM H X, LA H M, FEIL-SEIFER D, et al. Autonomous UAV navigation using reinforcement learning[J]. arXiv preprint arXiv: 1801.05086, 2018.

[13] ZENG Y, XU X. Path design for cellular-connected UAV with reinforcement learning[C]//2019 IEEE Global Communications Conference (GLOBECOM). IEEE, 2019: 1-6.

[14] LOWE R, WU Y, TAMAR A, et al. Multi-agent actor-critic for mixed cooperative-competitive environments[J]. arXiv preprint arXiv: 1706.02275, 2017.

（收稿日期：2022-03-01）

作者简介：

刘志飞（1985-），男，硕士研究生，主要研究方向：智能化指挥控制。

曹雷（1965-），通信作者，男，博士，教授，主要研究方向：机器学习、指挥信息系统和智能决策。E-mail: caolei.nj@163.com。

赖俊（1979-），男，硕士，副教授，主要研究方向：智能化指挥控制。