航空学报
Acta Aeronautica et Astronautica Sinica
ISSN 1000-6893,CN 11-1929/V

《航空学报》网络首发论文

题目：基于 DDPG 算法的无人机集群追击任务研究
作者：张耀中，许佳林，姚康佳，刘洁凌
收稿日期：2020-03-21
网络首发日期：2020-06-15
引用格式：张耀中，许佳林，姚康佳，刘洁凌．基于 DDPG 算法的无人机集群追击任务研究．航空学报．
https://kns.cnki.net/kcms/detail/11.1929.V.20200615.1529.046.html

航 空 学 报 Nov. 25 2020 Vol.41 No.11
Acta Aeronautica et Astronautica Sinica ISSN 1000-6893 CN 11-1929/V
http://hkxb.buaa.edu.cn hkxb@buaa.edu.cn

1 **引用格式**：*张耀中，许佳林，姚康佳，刘洁凌. 基于DDPG算法的无人机集群追击任务研究[J]. 航空学报, 2020, 41(11):324000.ZHANG Y Z, XU J L, YAO K J. A New UAV Swarm Pursuit Task Scheme Driven by DDPG algorithm [J].Acta Aeronautica et Astronautica Sinica,2020,41(11):324000(in Chinese). doi: 10.7527/S1000-6893.2020.24000*

基于 DDPG 算法的无人机集群追击任务研究

张耀中 [1,*]，许佳林 [1]，姚康佳 [1]，刘洁凌 [2]

1. 西北工业大学 电子信息学院，西安 710072;
2. 西安北方光电科技防务有限公司，西安 710043

摘 要：无人机的集群化应用技术是近年来的研究热点，随着无人机自主智能的不断提高，无人机集群技术必将成为未来无人机发展的主要趋势之一。本文针对无人机集群协同执行对敌方来袭目标的追击任务，构建了典型的任务场景，基于深度确定性策略梯度网络（Deep Deterministic Policy Gradient, DDPG）算法，设计了一种引导型回报函数有效解决了深度强化学习在长周期任务下的稀疏回报问题，通过引入基于滑动平均值的软更新策略减少了DDPG算法中Eval网络和Target网络在训练过程中的参数震荡，提高了算法的训练效率。仿真实验结果表明，训练完成后的无人机集群能够较好的执行对敌方来袭目标的追击任务，任务成功率达到95%。可以说无人机集群技术作为一种全新概念的作战模式在军事领域具有潜在的应用价值，人工智能算法在无人机集群的自主决策智能化发展方向上具有一定的应用前景。

关键词：DDPG算法；无人机集群；任务决策；深度强化学习；稀疏回报;

中图分类号：V279 **文献标识码**：A **文章编号**：1000-6893（2020）11-324000-13

无人机与有人飞机相比，具有体积小、造价低、使用方便、对作战环境要求低、战场生存能力强等优点。在过去的几十年里，伴随着导航、传感器、能量存储与制造技术等相关技术的发展，无人机在军用和民用领域都得到了广泛的应用。

随着无人机在相关领域应用的不断推进，单架无人机在执行任务时暴露出了灵活性和任务完成率的短板，因此使用多架无人机构成集群协同执行相关任务必将成为无人机未来应用的重要发展方向。无人机集群可以看作是一个多Agent系统（Multi-agent systems, MAS），其目标是协调集群内的无人机实现一个共同的任务目标。

当前对无人机集群的众多研究都集中在协同任务决策方面，通过蚁群算法、狼群算法等有关的群体智能算法来实现对多架无人机的指挥控制。但这些方法有着计算时间过长、灵活性不足、智能化程度低的缺点，无法很好地满足无人机集群对于无中心化、自主化、自治化的要求。相比而言，人工智能领域中的深度强化学习方法凭借着其强大的高维度信息感知、理解以及非线性处理能力，有望使无人机集群在面向战场复杂任务时有足够的智能协同完成作战任务。

目前，已经有诸多学者使用深度强化学习方法对无人机集群的相关问题进行了探索性研究。其中，Pham,H等基于深度强化学习算法对无人机的自主导航过程进行了研究，并应用于自主目标区域覆盖问题，在一定程度上解决了无人机集群联合行动下的协同任务规划问题和高维度状态空间的挑战[1-2]；S. Qi等人使用深度强化学习研究了智能体的环境感知问题，实现了对相邻智能体的意图感知[3]；李高垒和魏航使用深度强化学习方

收稿日期：2020-03-21；退修日期：2020-05-05；录用日期：2020-06-02；网络出版时间：
网络出版地址：
基金项目：航空科学基金(2017ZC53033)资助
*通讯作者. E-mail: zhang_y_z@nwpu.edu.cn

法研究了影响无人机自主空战的相关因素，为未来智能空战提供了理论依据[4-5]。Yamaguchi引入反馈控制律研究多机器人的协调运动问题，采用队形矢量法控制机器人群体队形实现了对目标的追击[6]。目前已有部分学者采用人工智能算法来解决无人机对目标的追击问题，如Aditya采用Q学习算法在栅格化环境下研究了智能体的追击问题，并与动态规划算法进行对比，取得了较好的效果[7]。苏治宝等人通过对未知环境中多移动智能体追击单目标问题的研究，采用强化学习中的Q学习算法给出了相应的解决方案[8]。通过对相关文献的分析可以看出，目前在无人机集群应用方面的研究还不够完善，所研究问题的规模都比较小，而且大多采用栅格化的任务环境，导致应用环境过于简单。

与此同时，一些军事强国，如美、英、俄罗斯等都在开展将人工智能技术应用于无人机集群任务的相关实验验证，美国已经开展了多个智能化无人机集群项目，2016年美军在加州进行的无人机集群实验，成功的将人工智能技术应用到无人机集群的行为决策中，实现了无人机集群在空中自主协作，组成无人机集群队形，并完成预定任务，充分体现了无人机集群的无中心化、自主化、自治化，这一实验表明美军在无人机集群自组网以及任务决策方面已经达到了实用化水平[9]。因此，进行无人机集群的应用研究具有一定的理论意义和使用价值。

本文在现有研究的基础上，以无人机集群对敌方来袭目标的追击任务为场景[10]，基于DDPG算法建立了人工神经网络模型，设计了一种引导型回报函数有效解决了深度强化学习在长周期任务下的稀疏回报问题，通过引入基于滑动平均值的软更新策略减少了DDPG算法中*Eval*网络和*Target*网络在训练过程中的参数震荡，提高了算法的训练效率。仿真实验结果表明，训练完成后的无人机集群能够较好的执行对敌方来袭目标的追击任务，表现了人工智能算法在提升无人机集群指挥决策能力上的应用潜力。

1 任务场景描述

如图1所示，在任务场景中出现敌方目标，目标的初始位置已知，保持高度和速度恒定飞行，我方派出无人机集群进行追击拦截。设定双方都处于同一个水平面内，不考虑高度因素。不同于以往将任务环境网格化的离散处理方案，本文构建了连续的二维战场地图作为无人机集群追击问题的任务环境，集群中的无人机、被追击目标的位置，均采用连续的空间位置坐标表示。

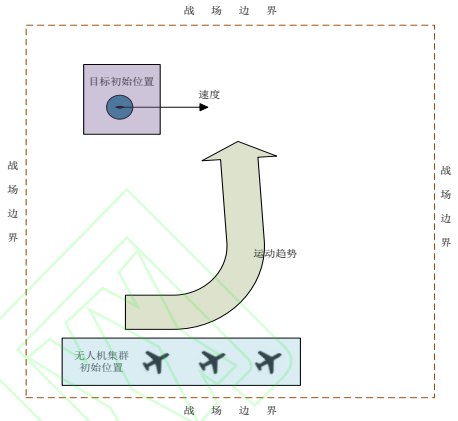

图 1 无人机集群执行追击任务示意图

Fig. 1 Schematic diagram of UAV swarm for pursuit task

本文针对任务场景中只有一个目标出现的情况，且不考虑目标针对无人机集群进行机动规避等行为，目标按照自身预定的运动策略进行飞行。无人机集群的任务是围堵目标，实现对目标的打击或者驱离，当无人机集群与目标之间的距离满足一定的态势要求后，视为无人机集群完成追击任务[12]。

2 无人机集群模型

2.1 无人机运动控制模型

为了便于问题分析，将集群中的无人机看作质点运动模型，使用两个方向的加速度来控制无人机的运动过程，如图2所示。

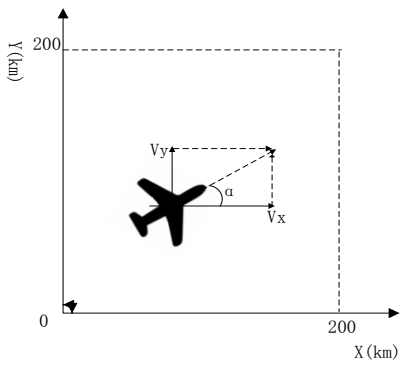

图 2 无人机的运动学模型

Fig. 2 Kinematic model of UAV

无人机的质点运动方程表示如下：

$$\begin{cases} v_x^{t+1} = v_x^t + a_{//} \cdot \cos\alpha \cdot t \pm a_{\perp} \cdot \sin\alpha \cdot t \\ v_y^{t+1} = v_y^t + a_{//} \cdot \sin\alpha \cdot t \pm a_{\perp} \cdot \cos\alpha \cdot t \end{cases} \tag{1}$$

$$\begin{cases} \sin\alpha = v_y^t / \sqrt{v_x^{t2} + v_y^{t2}} \\ \cos\alpha = v_x^t / \sqrt{v_x^{t2} + v_y^{t2}} \end{cases} \tag{2}$$

$$\begin{cases} x_{t+1} = x_t + v_x \cdot t \\ y_{t+1} = y_t + v_y \cdot t \end{cases} \tag{3}$$

其中：

v_x^t, v_y^t：无人机在时刻t时的飞行速度；

v_x^{t+1}, v_y^{t+1}：无人机在时刻t+1时的飞行速度；

$a_{//}, a_{\perp}$：在当前时刻无人机的切向、法向加速度；

x_t, y_t：在t时刻无人机的位置坐标；

x_{t+1}, y_{t+1}：在t+1时刻无人机的位置坐标；

α：无人机速度矢量与x轴方向的夹角。

针对上面建立的无人机运动控制模型，为了便于强化学习算法的实现，采用两个方向的加速度作为控制量对无人机的运动行为进行控制，如图3所示：

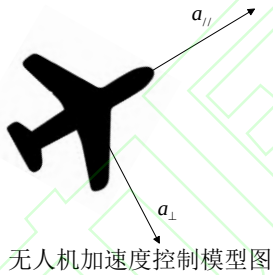

图 3 无人机加速度控制模型图

Fig.3 UAV acceleration control model diagram

由图3可知，无人机的行为空间包含切向加速度 $a_{//}$ 和法向加速度 $a_{\perp}$ 两个维度，无人机的行为即深度强化学习算法的输出可以是这两个维度中满足范围要求的任意值，限定无人机的行为空间满足如下约束：

$$\begin{cases} a_{//} \in (-2, 2) \\ a_{\perp} \in (-1, 1) \end{cases} \tag{4}$$

同时，对无人机的速度做出限制，规定无人机的速度 $v \in [3, 7]$。

2.2 无人机的传感器探测模型

设定集群中的无人机具有对任务场景的全局探测能力，为了模拟传感器的真实探测效果，对无人机的传感器探测结果加入一个服从正态分布 $\varepsilon \sim N(\mu, \sigma^2)$ 的随机误差。误差的参数由下式给出：

$$\begin{cases} \mu = 0 \\ \sigma = \dfrac{1}{60} * d_{i\_t} \end{cases} \tag{5}$$

其中：d_{i_t} 为无人机到目标的距离。

因此，集群中每架无人机对目标位置的探测结果如下：

$$\begin{cases} x_g = x_g^{'} + \varepsilon_x \\ y_g = y_g^{'} + \varepsilon_y \end{cases} \tag{6}$$

式中：

x_g ——无人机探测到的目标位置x坐标；

y_g ——无人机探测到的目标位置y坐标；

$x_g^{'}$ ——目标的真实位置x坐标；

$y_g^{'}$ ——目标的真实位置y坐标；

$\varepsilon_x, \varepsilon_y$ ——服从正态分布 $N(0, \sigma^2)$ 的随机误差。

无人机对目标速度的探测结果计算如下：

$$\begin{cases} v_{x\_g} = (x_{g\_now} - x_{g\_old}) / t \\ v_{y\_g} = (y_{g\_now} - y_{g\_old}) / t \end{cases} \tag{7}$$

其中：

x_{g_old}，y_{g_old}：上一时刻探测到的目标位置；

x_{g_now}，y_{g_now}：当前时刻探测到的目标位置。

2.3 集群内无人机的信息交互模型

集群内的无人机之间需要进行信息交互以便使无人机集群具有更好的协作行为决策，每架无人机都有固定的通信范围，在通信范围内的无人机之间可以进行通信，为了便于仿真分析，设定每架无人机最多可以与通信范围内距离最近的三架无人机进行信息交互，如图4所示：

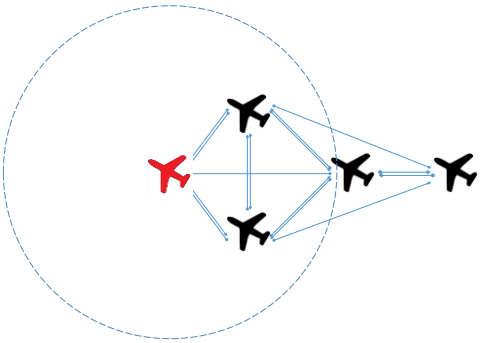

图 4 集群内信息交互关系示意图

Fig. 4 Schematic diagram of interaction in the swarm

图4中的箭头方向表示相应无人机信息传递的方向。集群中某架无人机 i 可以通过与周围无人机 j 的信息交互获取到相互之间的态势信息。如图5所示：

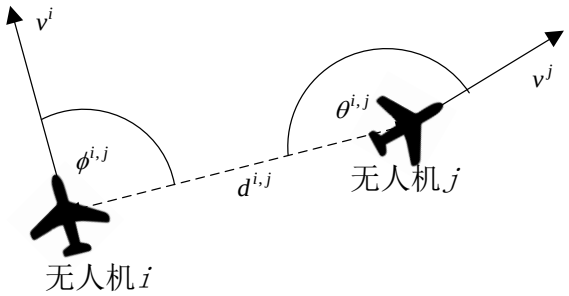

图 5 无人机间态势信关系意图

Fig. 5 Situational relationship between UAVs

其中：

$d^{i,j}$ --无人机 i 与相邻无人机 j 之间的距离；

$\phi^{i,j} = \arctan\left(\frac{y^j - y^i}{x^j - x^i}\right) - \phi^i$ -- 无人机 j 相对于无人机 i 的方位，ϕ^i 为无人机 i 的速度矢量与 x 轴方向夹角；

$\theta^{i,j} = \arctan\left(\frac{y^i - y^j}{x^i - x^j}\right) - \phi^j$ -- 无人机 i 相对与无人机 j 的方位，ϕ^j 为无人机 j 的速度矢量相对于 x 轴方向夹角；

3 深度确定性策略梯度网络算法

深度确定性策略梯度网络(Deep Deterministic Policy Gradient, DDPG)算法是一种结合了基于值迭代和策略迭代的深度强化学习算法[13-14]。该方法的优势在于可以针对无限大小的状态空间和行为空间实现智能体对最优策略的学习，使无人机集群在针对具体任务的学习过程中具有更优良的性能表现。DDPG算法是在传统的“演员——评论家”算法的基础上改进形成的，下面对算法网络的结构进行详细分析。

3.1 演员——评论家算法

“演员——评论家”算法主要由两个不同的网络模块组成，分别是演员网络模块和评论家网络模块。

演员网络模块主要通过对输入环境的状态观测，利用人工神经网络得到智能体行为的选择概率，完成智能体与环境的交互过程，并且用交互得到的环境回报对人工神经网络的参数进行更新，用来维护和更新智能体的动作选取策略。

评论家网络模块则通过对输入环境的状态及行为进行观测，来评估每个环境状态与行为的价值，即估计演员网络模块的价值，通过实际网络价值与预测网络价值的误差来更新当前神经网络。评论家网络模块输出的价值可以对演员网络模块的行为选取策略进行指导，这也是“演员——评论家”算法的由来。

由上述可知，对于“演员—评论家”算法两个不同的网络模块：演员网络模块和评论家网络模块分别需要建立各自的人工神经网络。演员网络模块的人工神经网络实现了从观测状态到智能体行为选取概率的映射，其训练过程需要结合评论家网络模块的误差进行。而评论家网络模块的人工神经网络是通过对环境状态和行为选取的观测得到相应的评分，形成环境状态与行为到对应评分的映射。“演员—评论家”算法的模型结构如图6所示。

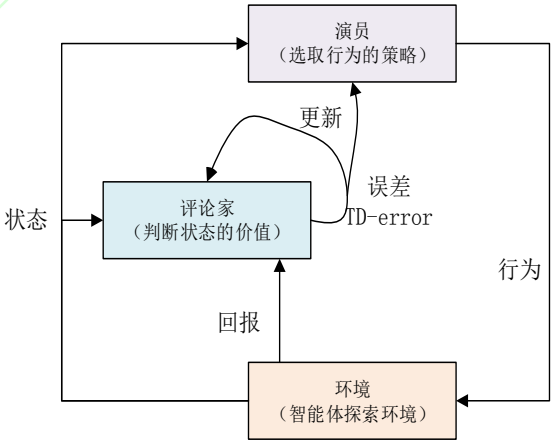

图 6 “演员—评论家”算法的模型结构图

Fig. 6 Model structure of “Actor-Critics” Algorithm

3.2 DDPG算法的网络架构

DDPG算法融合了“演员——评论家”算法和深度Q网络算法，是一种新型的深度强化学习算法[15][16]，算法的网络架构如图7所示。

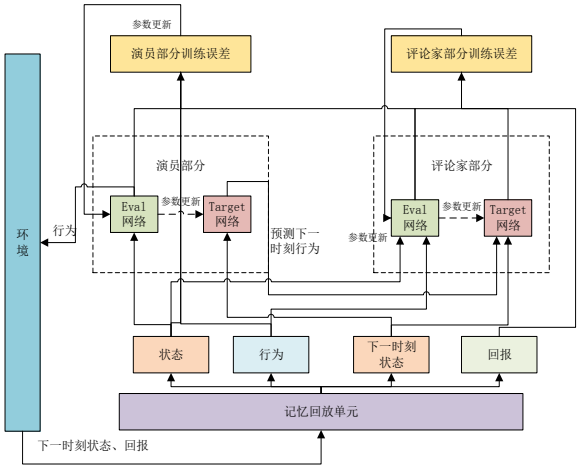

图 7 DDPG算法的网络架构图

Fig. 7 Network architecture of DDPG algorithm

如图7所示，DDPG算法主要由环境、记忆回放单元、演员网络模块和评论家网络模块构成。其中，环境是智能体的交互空间，也是智能体的探索空间，智能体在与环境的交互过程中得到交互样本，并将交互样本存储到记忆回放单元中用于智能体的训练过程。为了优化算法的学习过程，DDPG算法吸取了深度Q网络算法的思想，对于算法中的网络部分分别构建了一对结构完全相同的人工神经网络，分别称为 $Eval$ 神经网络和 $T\arg et$ 神经网络。其中 $Eval$ 神经网络用于训练更新网络参数，$T\arg et$ 神经网络则使用周期性软更新策略对 $Eval$ 神经网络进行跟随，并协助 $Eval$ 神经网络进行训练。

演员网络模块的神经网络用来完成对智能体行为选取概率的确定，智能体进行行为决策时，将依据演员网络模块提供的行为选择概率来选取行为与环境进行交互。评论家网络模块的神经网络通过接收环境状态和智能体行为，用来生成对“状态——行为”的价值评估。其中 $Eval$ 神经网络用来判断当前状态与行为的价值，$T\arg et$ 神经网络接收下一时刻的状态和演员部分 $T\arg et$ 神经网络输出的下一时刻行为，并进行价值判断。

DDPG算法中演员和评论家两部分的神经网络有着不同的功能和结构，相应的训练方式也不同，使用不同的损失函数进行训练。对于评论家网络而言，使用 $TD-error$ 对 $Eval$ 神经网络的参数进行训练，训练过程使用最小化损失函数 $Loss$ 进行更新，如下式所示。

$$TD-error=reward\left(s_t,a_t\right)+\gamma*v'\left(s_{t+1},a_{t+1};\theta'_{critic}\right)-v\left(s_t,a_t;\theta_{critic}\right) \tag{8}$$

$$Loss=\left(TD-error\right)^2 \tag{9}$$

式中：

$reward\left(s_t,a_t\right)$ ——当前状态和行为的环境回报，由训练样本给出；

$v'\left(s_{t+1},a_{t+1};\theta'_{critic}\right)$ ——下一时刻状态和行为的价值评估，由评论家网络模块中的 $Target$ 神经网络给出；

$v\left(s_t,a_t;\theta_{critic}\right)$ ——当前时刻状态和行为的评估，由评论家网络模块中的 $Eval$ 神经网络给出；

a_{t+1} ——下一时刻的行为，由演员网络模块中的 $Target$ 神经网络给出；

s_{t+1} ——下一时刻的状态，由训练样本给出；

a_t ——当前时刻的行为，由训练样本给出，实际上是样本产生时的演员网络模块中的 $Eval$ 神经网络给出；

s_t ——当前时刻的状态，由训练样本给出；

γ ——奖励折扣因子。

对于演员网络模块中神经网络的训练过程，通过最大化<状态，行为>对的价值判断来实现，因此使用对状态和行为的评价均值作为损失函数，如下式所示：

$$Loss=-mean\left(v\left(s,a;\theta_{critic}\right)\right) \tag{10}$$

3.3 DDPG算法中探索与经验的平衡

在DDPG算法中，如果只是依据算法输出的行为选择策略来决定无人机的当前行为，容易导致算法对任务环境探索的不充分，因此需要对DDPG算法策略增加一定的探索性[17]。根据DDPG算法的特点，增强算法探索性的实现方法是在无人机行为选取过程中增加一定的随机噪声。如下式所示：

$$action=action'+Noise \tag{11}$$

式中：

$action$ ——无人机当前时刻选择的行为；

$action'$ ——DDPG算法中演员网络模块输出的无人机行为；

$Noise$ ——随机噪声。

由于DDPG算法输出的是无人机在两个方向上加速度的连续控制量，因此采用上述方法增强

DDPG算法的探索性具备良好的可行性，设定随机噪声服从正态分布：

$$Noise \sim N(\mu, \sigma^2) \tag{12}$$

噪声的期望值 $\mu = 0$，方差 σ 与迭代轮次相关，随着网络训练迭代次数的增加 σ 将逐渐减小，为了保证无人机集群具备足够的探索能力，确保在无人机探索初期其行为选择能够选取到行为空间中的任意值，对随机噪声方差初始值的设计如下式所示：

$$\sigma_0 = (action_{\max} - action_{\min}) / 4 \tag{13}$$

$$\sigma = K^{episode} * \sigma_0 \tag{14}$$

其中：$K = 0.9995$，$episode$ 算法训练代数。

3.4 DDPG算法的网络结构

由前述分析可知，DDPG算法由一对结构完全相同的神经网络，即“演员”部分人工神经网络(Actor网络)和“评论家”部分人工神经网络(Critic网络)构成，所构建网络的Tensorboard输出如图8所示。

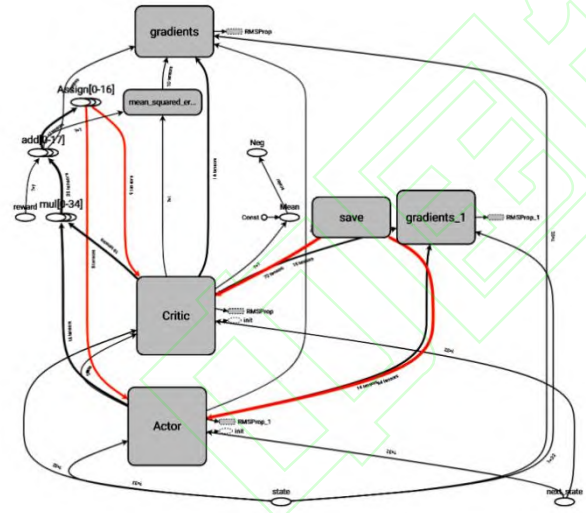

图 8 DDPG算法网络结构(Tensorboard)

Fig. 8 Network structure of DDPG algorithm(Tensorboard)

3.4.1 “演员”网络模块的人工神经网络结构

“演员”网络模块的人工神经网络用来输出无人机的行为，在无人机集群追击任务环境中，无人机集群的状态空间为自身位置 x_i, y_i、速度 v_{x_i}, v_{y_i}、探测得到的目标位置 x_g, y_g、速度 v_{x_g}, v_{y_g} 以及通过信息交互得到的其他无人机的相关信息 $x_{ij}, y_{ij}, v_{x\_ij}, v_{y\_ij}$ 和其他无人机的探测信息 $x_{ij\_get}, y_{ij\_get}, v_{x\_ij\_get}, v_{y\_ij\_get}$，共32个维度作为无人机的状态空间，如图9所示。

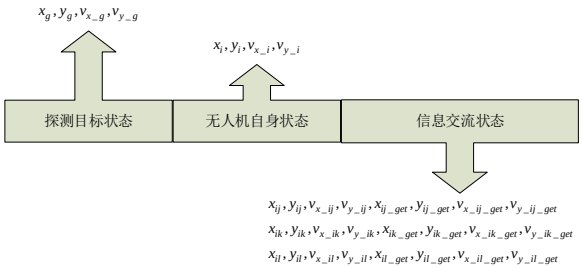

图 9 演员网络模块的状态空间构成

Fig. 9 State space of 'Actor' network module

对“演员”网络模块中的 $Target$ 和 $Eval$ 人工神经网络，我们构建了两个结构完全相同的6层全连接人工神经网络，每层网络的人工神经元个数分别为 $[100,100,300,100,100,2]$，最后一层神经网络为2维度的输出层，对应无人机的切向加速度 $a_{//}$ 与法向加速度 $a_{\perp}$。输出神经元使用 $tanh(x)$ 作为激活函数，实现网络输出与无人机行为的映射，其他各层的神经元使用 $relu(x)$ 作为激活函数。并且使用 $RMSProp$（Root Mean Square Prop）算法作为训练的优化器。“演员”网络模块中人工神经网络的结构如图10所示：

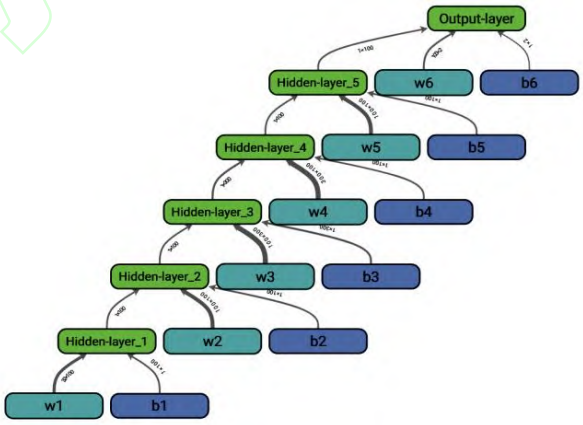

图 10 “演员”网络模块中人工神经网络结构

Fig. 10 Network structure in the 'Actor' network module

3.4.2 “评论家”网络模块的人工神经网络结构

“评论家”网络模块的人工神经网络通过对“状态——行为”的价值评估，指导“演员”网络模块中神经网络的训练过程。因此，评论家网络模块中神经网络的输入状态为无人机集群的状态信息与行为信息，网络的状态空间构成如图11所示。

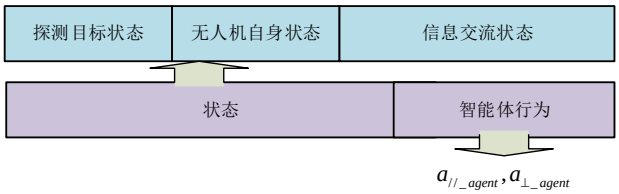

图 11 “评论家”网络模块的状态空间构成

Fig. 11 State space of ‘Critic’ network module

对“评论家”网络模块中的$Target$和$Eval$人工神经网络，构建了两个结构完全相同的5层全连接人工神经网络，每层网络的人工神经元个数分别为$[100,300,100,10,1]$。输出层的神经元使用$tanh(x)$作为激活函数，隐藏层的神经元使用$relu(x)$作为激活函数，并且使用$RMSProp$（Root Mean Square Prop）算法作为训练的优化器。神经网络的结构如图12所示。

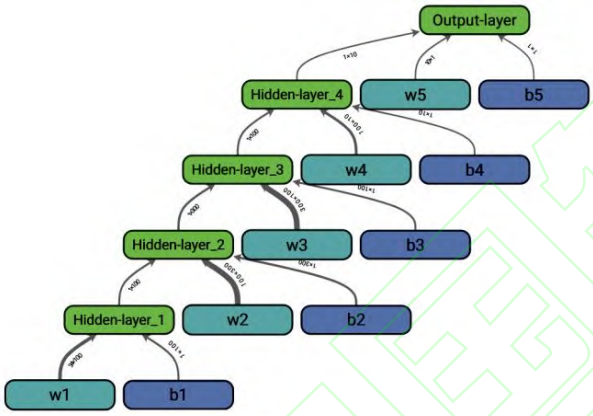

图 12 “评论家”网络模块中的人工神经网络结构

Fig. 12 Network structure in the ‘Critic’ network module

在“演员”网络模块和“评论家”网络模块中同时存在$Target$和$Eval$人工神经网络，其中$Eval$神经网络用于训练过程，而$Target$神经网络则周期性的跟随训练网络相应参数的变化而更新。对于$Target$神经网络的参数更新使用基于滑动平均值的软更新策略，如下式所示：

$$\theta_{Target} = k * \theta_{Target} + (1-k) * \theta_{Eval} \tag{15}$$

式中：

θ_{Target}——$Target$神经网络参数；

θ_{Eval}——$Eval$神经网络参数；

k——滑动因子，经验取值为0.2。

3.4 DDPG算法的稀疏回报问题

对于连续的状态空间和行为空间，无人机进行随机初始化之后要经历一段很长时间的与环境的交互过程才能达到最终状态。此时，仅在无人机集群到达最终状态之后给予相应回报的方式，有着回报周期过长的缺陷，容易导致强化学习过程无法进行有效学习，即存在着稀疏回报问题。

为了解决稀疏回报问题，我们对无人机集群的学习目标进行了相应的修改，增加有效回报，从而加快学习速度，构建不同情形下无人机的回报函数来指导深度强化学习的学习方向，如下式所示：

$$\begin{cases} r_1 = 100 \\ r_2 = -10 \\ r_3 = -100 \\ r_4 = d_{i\_t} - d'_{i\_t} + v_i * \cos(\beta) \end{cases} \tag{16}$$

式中：

d_{i_t}——当前时刻无人机与目标之间的距离；

d'_{i_t}——下一时刻无人机与目标之间的距离；

β——当前时刻无人机速度方向与目标连线之间的夹角；

v_i——当前时刻无人机的速度大小。

仿真实验中的任务回报分为四种类型，当无人机集群完成追击任务之后，对完成任务的无人机给予回报$r_1 = 100$；当发生无人机集群碰撞战场边界，对于发生碰撞的无人机给予负向回报$r_2 = -10$；当无人机集群未完成追击任务，对于所有的无人机给予负向回报$r_3 = -100$，并结束当前回合的训练；在任务执行过程中，使用无人机集群的引导型回报函数$r_4 = d_{i_t} - d'_{i_t} + v_i * \cos(\beta)$对无人机的回报进行判断。

对于公式(16)中的无人机集群回报函数，由无人机与目标之间的距离变化情况、无人机的速度方向以及无人机的速度大小共同表示。当无人机与目标之间的距离变小时对应的回报函数为正值；由无人机的速度大小与速度方向相结合构成了回报函数，在相同速度大小的情况下，速度矢量的方向越指向目标，无人机的回报就越高；同理，在无人机速度方向指向目标的情况下，无人机的速度越大回报越高；对于无人机速度方向远离目标的情况下，无人机的速度越大，其负向回报越高。

由于无人机集群从初始状态出发，需要运行较长时间才能到达目标状态，如果在长时间的中间状态下无法得到环境的有效回报，容易导致算法训练过程中的梯度消失，从而导致训练过程无

法收敛。无人机集群采用上述引导型回报函数时，训练过程中会根据无人机的任一状态产生一个与当前<状态，行为>对应的价值回报，从而引导无人机集群逐渐向目标状态转移。因此，公式(16)能较准确的反应无人机的行为收益，算法的训练结果表明，通过采用引导型回报函数能够较好的解决深度强化学习中的稀疏回报问题。

3.5 DDPG算法的程序流程

使用DDPG算法对无人机集群的追击任务进行训练，程序实现流程如图13所示：

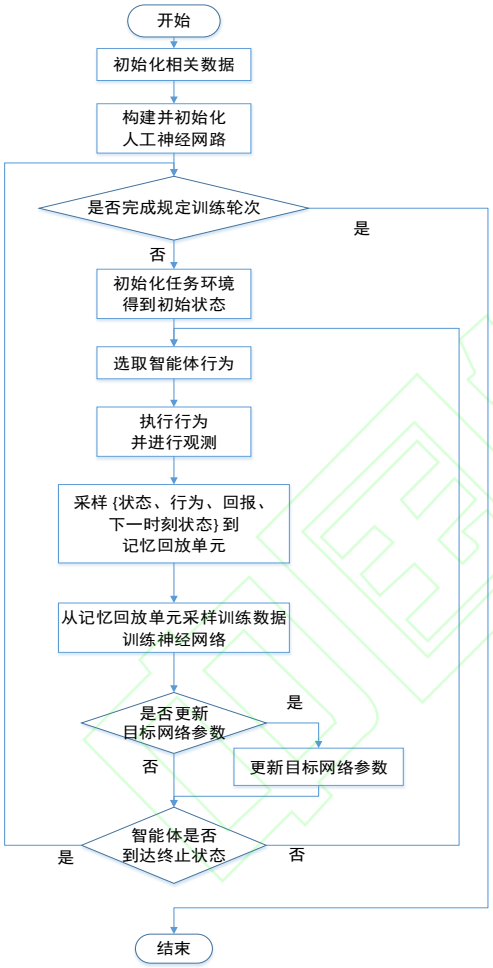

图 13 DDPG算法的程序流程图

Fig. 13 Algorithm flow chart of DDPG

4 仿真实验

设定仿真场景中只存在一个匀速前进的目标，当集群中的任意一个无人机追击到目标之后，视为无人机集群完成了对目标的追击任务，即到达了任务的最终状态。

4.1 训练过程

仿真中使用5架完全相同的无人机构成集群进行训练。为了便于观察算法的训练状态，防止训练过程中出现梯度消失等现象，我们对人工神经网络的收敛性能进行了监测，分别选取“演员”和“评论家”网络模块中的神经网络参数进行统计观察，得到相关统计信息如下。

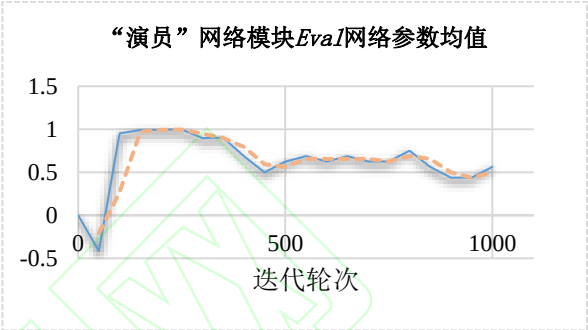

图 14 “演员”部分*Eval*网络参数均值变化曲线图

Fig. 14 Curve of average change in *Eval* network parameters in the 'Actor' network module

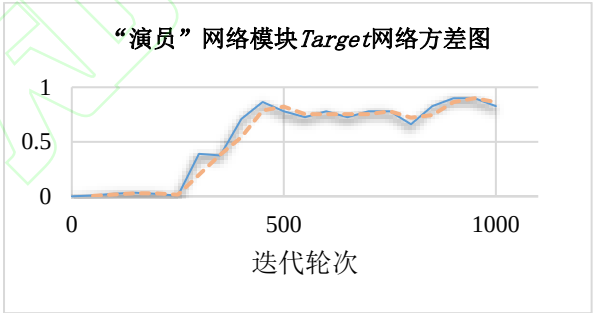

图 15 “演员”部分Target网络参数方差变化曲线图

Fig. 15 Curve of variance in *Target* network parameters in the 'Actor' network module

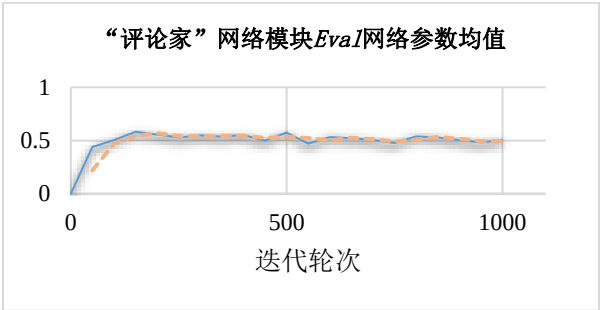

图 16 “评论家”部分Eval网络参数均值变化曲线图

Fig. 16 Curve of average change in *Eval* network parameters in the ‘Critic’ network module

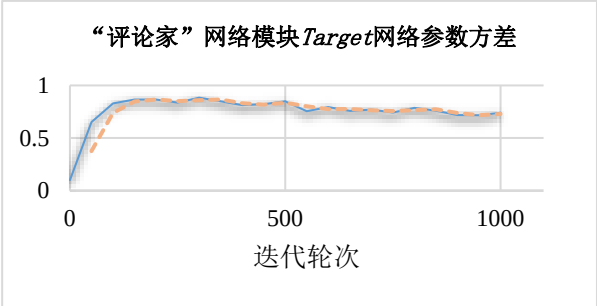

图 17 “评论家”部分Target网络参数方差变化曲线图

Fig. 17 Curve of variance in *Target* network parameters in the 'Critic' network module

上述数据曲线图分别是对“演员”和“评论家”网络模块中的神经网络参数取均值和方差进行统计的结果，图中实线为网络参数统计的真实值，虚线则是对统计数据进行周期为3的滑动平均处理的结果，用来表明参数统计的变化趋势。由上述参数统计曲线图可以看出人工神经网络在训练过程中很好的实现了收敛。

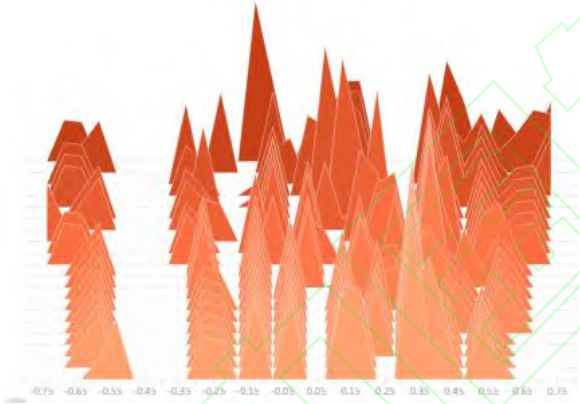

图 18 “评论家”部分*Eval*网络参数分布变化曲线图

Fig. 18 *Eval* network parameter distribution curve in the ‘Critic’ network module

图 18 是截取自 TensorBoard 的“评论家”网络模块中的神经网络参数分布变化直方图，由远及近（颜色由深变浅）表现了神经网络在不同训练阶段各个神经元参数分布的变化情况，从神经网络的参数统计变化曲线图与参数分布变化直方图可以看出，人工神经网络的参数分布情况在训练过程中逐渐收敛到稳定的分布状态。

无人机集群在不同训练轮次下的平均回报值变化趋势如图 19 所示。

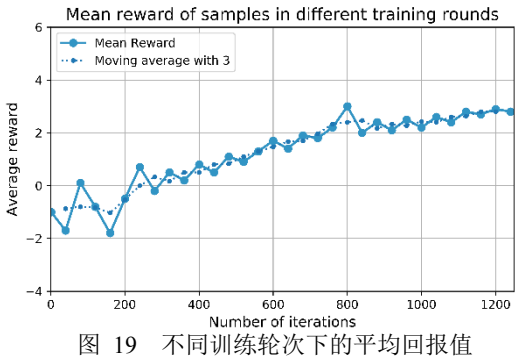

图 19 不同训练轮次下的平均回报值

Fig. 19 Mean value of rewards under different training epochs for the UAV swarm.

由图 19 可见，在算法的训练过程中，无人机集群的行为收益值保持比较平稳的状态缓慢增加，说明无人机集群行为随着训练过程的不断进行有着越来越好的表现。

随着算法训练回合的增加，无人机集群在环境中的总回报奖励变化趋势如图 20 所示。

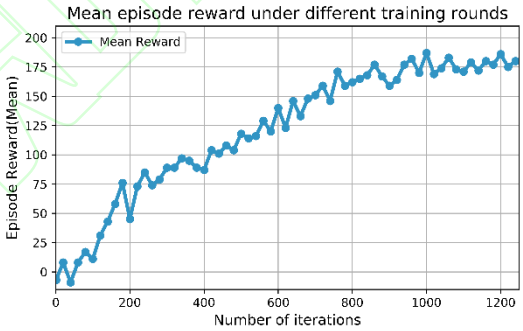

图 20 不同训练轮次下的无人机集群回合总回报

Fig. 20 Total rewards under different training epochs for the UAV swarm.

无人机集群在不同训练轮次下的任务完成率如图 21 所示：

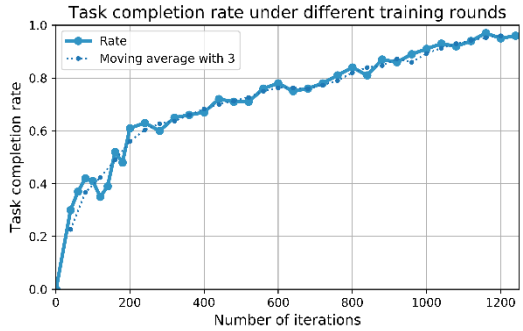

图 21 不同训练轮次下的无人机集群任务成功率

Fig. 21 Task completion rate under different training epochs for UAV swarm.

从图 21 可以看出，完成训练后无人机集群执行对敌来袭目标追击任务的成功率可以达到95%左右。

4.2 验证过程

使用5架相同无人机构成集群完成所创建神经网络的训练后，对训练完成的模型进行了测试验证。使用训练完成的无人机集群执行对目标的追击任务，生成5架无人机集群及目标的初始状态，得到无人机集群追击任务的轨迹图如图22所示。

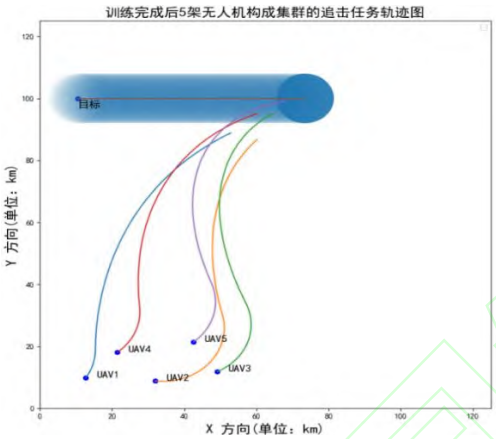

图 22 5架无人机执行追击任务的轨迹图

Fig. 22 Trajectory of 5 UAVs on pursuit mission

如图22所示，使用训练完成的神经网络模型很好地实现了5架无人机构成集群执行对目标的追击任务。为了验证模型对于动态数量无人机集群的适用性，分别使用10架和20架无人机构成集群，对无人机集群的追击任务进行验证，得到无人机集群轨迹图如下所示。

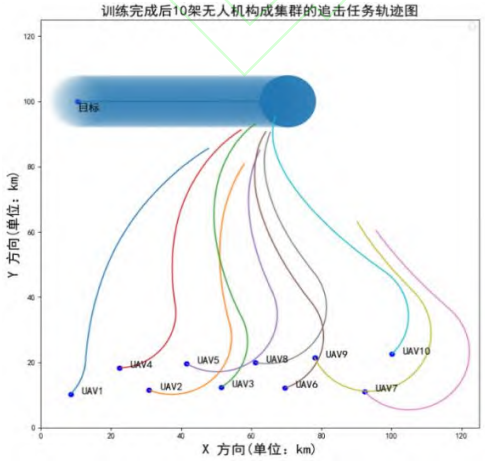

图 23 10架无人机执行追击任务的轨迹图

Fig. 23 Trajectory of 10 UAVs on pursuit mission

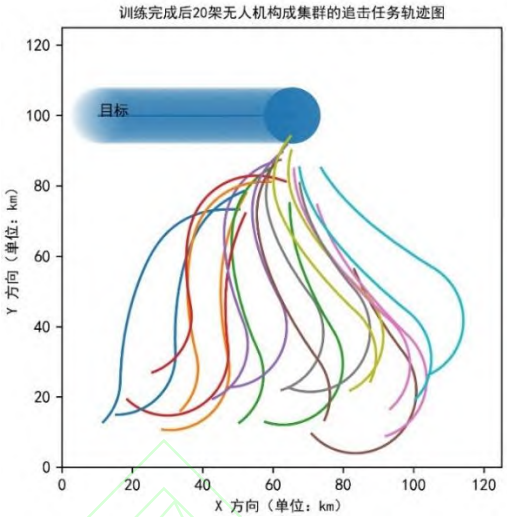

图 24 20架无人机执行追击任务的轨迹图

Fig. 24 Trajectory of 20 UAVs on pursuit mission

由上图可以看出，基于5架无人机训练得到的模型能很好地应用于10和20架无人机用来执行对敌来袭目标的追击任务中，可以看出，DDPG算法对无人机集群的行为决策有着良好的适应能力和泛化能力。

为了进一步验证本文基于改进DDPG算法无人机集群模型的泛化能力和适应能力，对具有不同程度的逃逸策略的机动目标使用训练完成的集群模型进行了实验验证，得到无人机集群轨迹图如下所示：

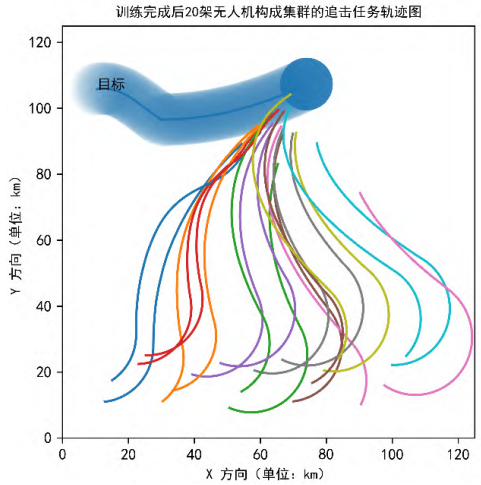

图 25 简单逃逸策略下对目标的追击任务轨迹图

Fig. 25 Trajectory of 20 UAVs on pursuit mission with simple escape strategy target

由上图仿真结果可以看出，对于具有简单逃逸策略的来袭目标，无人机集群很好的完成了预定的追击任务。

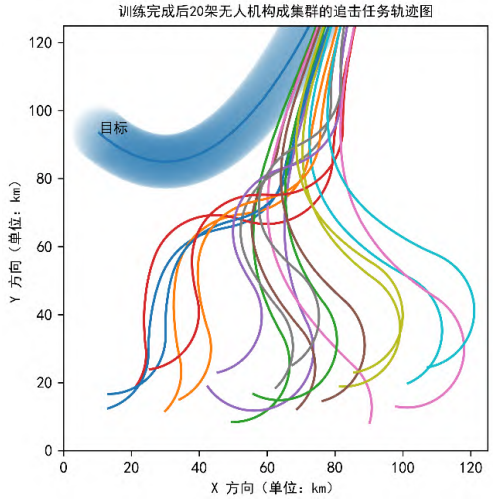

图 26 大机动逃逸策略下对目标的追击任务轨迹图

Fig. 26 Trajectory of 20 UAVs on pursuit mission with big maneuver escape strategy target

在上图26的追击任务场景中，当目标采用大机动逃逸运动策略时，由于来袭目标快速逃逸出了设定的任务边界导致目标逃逸成功，但是训练完成后的无人机集群仍然很好的完成了对预定目标的追击任务。

仿真实验表明，深度强化学习能够很好地满足无人机集群对于无中心化、自主化和自治化的要求。将人工智能算法应用在无人机集群的任务决策中具有很好的发展前景。

5 结 论

本文基于深度强化学习中的DDPG算法对无人机集群追击任务进行了研究，为了平衡DDPG算法“探索——经验”的矛盾，在训练过程中对无人机行为加入了自适应的噪声单元，以增强算法的探索能力。为了提升算法性能，引入基于滑动平均值的软更新策略减少了DDPG算法中*Eval*网络和*Target*网络在训练过程中的参数震荡，提高了算法的收敛速度。为解决深度强化学习中的“稀疏回报”问题，设计了指导型回报函数，避免了无人机集群在长周期训练条件下无法有效学习的问题，提升了算法的收敛性。

训练完成后的无人机集群能够很好的执行追击任务。同时验证了在不改变网络模型和状态空间结构的前提下，训练完成的模型能直接应用于更多无人机构成的集群追击任务中和具有不同程度逃逸策略的机动目标追击任务中。仿真结果表明使用DDPG算法针对无人机集群的追击任务可以求解出良好的行为策略，体现了基于人工神经网络的强化学习算法在提升无人机集群指挥决策模型的泛化能力上的巨大应用潜力。

参 考 文 献

[1] PHAM, H., LA, H., FEIL-SEIFER, D., & NGUYEN, L. Autonomous UAV Navigation Using Reinforcement Learning[D]. Cornell University, 2018.

[2] PHAM, H., FEIL-SEIFER, D., & NEFIAN, A. Cooperative and Distributed Reinforcement Learning of Drones for Field Coverage[D]. Cornell University, 2018.

[3] S. QI and S. ZHU. Intent-Aware Multi-Agent Reinforcement Learning[C]. 2018 IEEE International Conference on Robotics and Automation (ICRA), Brisbane, QLD, 2018, pp. 7533-7540.

[4] 李高垒,马耀飞. 基于深度网络的空战态势特征提取[J]. 系统仿真学报, 2017, 29(S1):98-105+112.
Li G L, Ma Y F. Feature Extraction Algorithm of Air Combat Situation Based on Deep Neural Networks[J]. Journal of System Simulation, 2017, 29(S1):98-105+112. (in Chinese).

[5] 魏航. 基于强化学习的无人机空中格斗算法研究[D]. 哈尔滨工业大学, 2015.
Wei H.Resarch of UCAV Air Combat Based on Reinforcement Learning[D]. Harbin Institute of Technology. (in Chinese).

[6] Yamaguchi H . A Cooperative Hunting Behavior by Mobile Robot Troops[J]. Proc.ieee Int.conf.robotics & Automat, 1998, 18(9):931-940.

[7] Aditya Gadre. Learning Strategies in Multi-Agent Systems Applications to the Herding Problem[J], Blacksburg,VA,November,2001.

[8] 苏治宝, 陆际联, 童亮. 一种多移动机器人协作围捕策略[J]. 北京理工大学学报, 2004(05):32-35+44.
Su Z B, Lu J L, Tong L. Strategy of Cooperative Hunting by Multiple Mobile Robots[J]. Beijing Institute of Technology, 2004(05):32-35+44. (in Chinese).

[9] 罗德林,徐扬,张金鹏. 无人机集群对抗技术新进展[J]. 科技导报, 2017,35(07):26-31.
Luo D L, Xu Y, Zhang J P. New progresses on UAV

swarm confrontation[J].Science & Technology Review, 2017,35(07):26-31. (in Chinese).

[10] CARL E J. Analysis of fatigue, fatigue-crack propagation and fracture data: AIAA-2009-1363[R]. Reston, VA: AIAA, 1973.

[11] ZUHAIR Q M, SONGHAO P , HAIYANG J , et al[J/OL]. A novel approach for multi-agent cooperative pursuit to capture grouped evaders, （2018-09-12）[2018-09-12].https://doi.org/10.1007/s11227-018-2591-3.

[12] ZHAOYI P , SONGHAO P , Mohammed E H S , et al. Coalition Formation for Multi-agent Pursuit Based on Neural Network[J]. Journal of Intelligent & Robotic Systems, 2019, 95(01):887–899.

[13] HUMAYOO, M., & CHENG, X. Relative Importance Sampling For Off-Policy Actor-Critic in Deep Reinforcement Learning[D]. Cornell University, 2018.

[14] 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(06):1406-1438.
Liu J W, Gao F, Luo X L. A Survey of Deep Reinforcement Learning Based on Value Function and Strategy Gradient[J]. Chinese Journal of Computers, 2019, 42(06):1406-1438. (in Chinese).

[15] FUIHONG WANG, JINGLUN SHI. Actor-Critic for Multi-Agent System with Variable Quantity of Agents[C]//International Conference on Internet of Things as a Service, 2017:48-56.

[16] W. HUANG, Y. WANG and X. YI. A deep reinforcement learning approach to preserve connectivity for multi-robot systems[C]//2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), Shanghai, 2017: pp. 1-7.

[17] Yi, H. Deep Deterministic Policy Gradient for Autonomous Vehicle Driving[C]//Proceedings on the International Conference on Artificial Intelligence (ICAI), 2018: 191-194.

[18] P. ANDERSEN, M. GOODWIN and O. GRANMO. Deep RTS: A Game Environment for Deep Reinforcement Learning in Real-Time Strategy Games[C]//2018 IEEE Conference on Computational Intelligence and Games (CIG). Maastricht. 2018: pp. 1-8.

[19] N. DILOKTHANAKUL, C. KAPLANIS, N. PAWLOWSKIand M. ShANAHAN. Feature Control as Intrinsic Motivation for Hierarchical Reinforcement Learning[C]//IEEE Transactions on Neural Networks and Learning Systems, 2019,11, vol. 30, no. 11: pp. 3409-3418.

[20] H. NIE, Y. CHEN, Y. SONG and S. HUANG. A General Real-time OPF Algorithm Using DDPG with Multiple Simulation Platforms[C]//2019 IEEE Innovative Smart Grid Technologies - Asia (ISGT Asia), Chengdu, China, 2019,5: pp. 3713-3718.

[21] Q. YANG, Y. ZHU, J. ZhANG, S. QIAO and J. LIU. UAV Air Combat Autonomous Maneuver Decision Based on DDPG Algorithm[C]//2019 IEEE 15th International Conference on Control and Automation (ICCA), Edinburgh, United Kingdom, 2019: pp. 37-42.

[22] A. BANERJEE, D. GHOSH and S. DaAS, Evolving Network Topology in Policy Gradient Reinforcement Learning Algorithms[C]//2019 Second International Conference on Advanced Computational and Communication Paradigms (ICACCP), Gangtok, India, 2019: pp. 1-5.

[23] H. SHI, Y. SUN and G. LI. Model-based DDPG for motor control[C]//2017 International Conference on Progress in Informatics and Computing (PIC), Nanjing, 2017: pp. 284-288.

（责任编辑：李丹）

Research on the pursuit mission for UAV swarm based on DDPG algorithm

ZHANG Yaozhong[1,*], XU Jialin[1], YAO Kangjia[1] , Liu Jieling[2]

1. *School of Electronics and Information, Northwestern Polytechnical University, Xi'an 710072, China*

2. *Xi'an North Electro-optic Science & Technology Co. Ltd, Xi'an 710043, China*

Abstract: The Unmanned Aerial Vehicle (UAV) swarm technology is one of the research hotspots in recent years. With the continuous improvement of autonomous intelligence of UAV, the swarm technology of UAV will become one of the main trends of UAV development in the future. In this paper, based on swarm of UAVs collaboratively completing the attack missions to the enemy, we build up a typical task scenario, based on the Deep Deterministic Policy Gradient (DDPG) algorithm, and we design a guided reward function which effectively solves the depth of intensive problem in long period under the sparse rewards. We introduce the soft updating strategy based on sliding average and this approach reduces the parameters oscillation in *Eval* network and *target* network in the training process, improves the training efficiency. The simulation results show that after training, the UAV swarm can carry out the pursuit task very well, and the success rate of the mission reaches 95%. UAV swarm technology, as a new concept of combat mode, has potential value for application in the military field, and this artificial intelligence algorithm has a certain application prospect in the development of autonomous decision-making intelligence via UAV swarms.

Keywords: DDPG algorithm; UAV swarm; task decision; deep reinforcement learning; sparse reward

Received：2020-03-21；Revised：2020-05-05；Accepted：2020-06-02；Published online:
URL（网络出版地址）：
Foundation item：Aeronautical Science Foundation of China (2017ZC53033)
*Corresponding author. E-mail: zhang_y_z@nwpu.edu.cn