https://doi.org/10.1051/jnwpu/20244220310

基于 SAC 的无人机自主导航方法研究

寇凯, 杨刚, 张文启, 刘心成, 姚远, 周兴社

(西北工业大学 计算机学院, 陕西 西安 710072)

摘 要:针对现有深度强化学习算法在无人机自主导航任务中面临环境局部可观且感知信息不足问题,基于非确定性策略 SAC(soft actor-critic)强化学习算法对未知环境下的端到端无人机自主导航任务展开研究。具体而言,提出了一种基于记忆增强机制的策略网络,通过对历史记忆信息与当前的观测整合处理,提取观测数据的时序依赖关系,从而增强局部可观条件下的状态估计能力,避免算法陷入局部最优解;设计了非稀疏奖励函数以缓解强化学习策略在稀疏奖励条件下难以收敛的问题;在 Airsim+UE4 仿真平台进行了多个复杂场景的训练验证。实验结果表明,所提方法导航成功率比基准算法提高 10%,平均飞行距离缩短 21%,有效增强了无人机自主导航算法稳定性和收敛性。

关 键 词:强化学习;SAC;无人机;自主导航
中图分类号:TP391 **文献标志码**:A **文章编号**:1000-2758(2024)02-0310-09

近年来,无人机产业呈现出蓬勃发展的趋势。随着产业需求不断拓宽与深化,无人机需要具备在高度复杂和未知的恶劣环境中自主执行任务的能力[1-2],例如,军事反恐、灾后搜救、山区测绘等。在这些应用场景中,无人机依靠配备的传感器收集环境信息,自主地规划路线和规避障碍物,最终实现高效、可靠的自主导航与飞行,这对无人机的自主导航算法提出了严峻的挑战[3-4]。

基于经典优化理论的传统导航技术首先需要建立起精确的无人机运动学模型,然后依赖环境的先验地图和手工设计的特征,实现理想条件下有限的自主导航。这种方法在鲁棒性、泛化性和适应性方面已经逐渐难以满足实际导航任务。近年来,深度学习(deep learning, DL)已经在自主控制领域得到了广泛应用。现有基于深度学习的自主导航算法通常将自主导航定义为一个分类或回归任务,通过深度学习模型将雷达、图像和点云等环境感知信息映射为特征空间,然后,训练好的模型可以利用观测信息预测无人机的机动策略[5]。Gandhi 等[6]将无人机的控制量简化为离散的单自由度变量,收集单目图像和运动信息作为数据集,然后,采用 Alexnet 网络[7]作为主干网络来预测无人机的飞行动作,实现了室内场景下的无人机自主飞行。但是,这种方法难以满足复杂场景下无人机自主导航的控制需求。Maciel-Pearson 等[8]搭建了一个仿真无人机室外飞行环境来训练端到端的多任务回归模型,以预测连续的无人机飞行动作,实现了非结构环境下的自主导航。可以看出,这类方法严重依赖于人工标注的数据集,造成了极大的人力耗费。更严重的是,无人机学到的是“反应式”策略,一个状态下的输入与其预测的控制量具有唯一性,这难以适应于复杂的真实控制过程。此外,模型的泛化性很差,当环境改变时,必须再经历一轮模型构建过程。因此,算法的应用范围受到了极大的限制[9]。

深度强化学习(deep reinforcement learning, DRL)作为人工智能领域的重要分支,能够以端到端学习的方式解决复杂的感知和决策问题。将深度强化学习理论应用在无人机自主导航任务中,无人机通过与环境的交互来探索海量的状态-动作空间,并且直接从感知信息中通过端到端学习不断改进动作策略。然后,根据环境反馈的奖励值评估当前行为的优劣,反复迭代,直到收敛到最优的控制策略[10]。Walvekar 等[11]试图利用无人机采集的图像作为输入,然后输入到 DQN 算法中,以训练无人机

收稿日期:2023-03-27
基金项目:国家自然科学基金(62032018,62141220,61876151)资助
作者简介:寇凯(1993—),博士研究生
通信作者:杨刚(1974—),教授 e-mail:yeungg@nwpu.edu.cn

在飞向固定目标的同时避免与障碍物碰撞。但该方法不能实现随机目标的自主导航,只能执行固定轨迹的导航任务。因此,无人机不具有完全自主性。

针对这一问题, Kabas[12] 将无人机的观测信息与状态信息作为联合状态空间,基于 PPO 算法[13] 预测无人机离散的飞行动作,使得无人机可以完全自主地避开障碍飞向目标点。然而,由于环境的局部可观性,无人机采集的视觉信息不足以准确地表达所处环境的复杂性,无人机的导航决策很容易陷入局部最优解。为此,何磊等[14] 提出了一种用于无人机自主飞行避障的深度强化学习方法,采用仿生单目视觉信息作为输入,为了提高感知对照度变化的鲁棒性,使用与照度变化无关的图像矩重新映射输入图像。然后,局部规划器使用强化学习进行无地图导航训练。

实际上,在大多数基于强化学习的无人机自主导航算法中,无人机在决策时通常仅考虑当前时刻的观测数据。然而,受限于环境的局部可观性与传感器自身性能,无人机所获得的观测信息普遍是不完整和附带噪声的。因此,当面对复杂的场景时,如密集的障碍区域,无人机常常由于缺乏足够的环境信息而难以制定正确的决策,甚至发生碰撞。因此,为了解决这一问题,有必要整合和研究历史数据,使算法能够学习到更为全面的环境特征,从而引导无人机跳出局部最优解,使生成的策略更加稳健。

综上所述,本文以三维未知环境中无人机自主导航任务为研究对象,结合强化学习的决策能力与深度学习的特征提取能力,使无人机可以通过机载传感器感知环境信息,实现端到端自主导航决策控制。具体而言,本文贡献为:①建立无人机自主导航任务的马尔可夫决策过程,设计目标驱动的非稀疏奖励函数,并基于 SAC 算法实现无人机在无先验知识条件下的自主导航;②提出具有记忆增强结构的策略学习模型以提取观测数据的时序依赖关系,提升强化学习算法在局部可观条件下的探索能力,避免自主导航模型陷入局部最优解;③在 Airsim+UE4 仿真平台中搭建三维仿真环境,对所提算法的有效性进行验证。

1　基于强化学习的自主导航算法

1.1　总体设计

未知环境下的无人机自主导航是一个将观测空间映射为高维连续动作空间的决策过程。由于缺乏全局先验地图以及环境的局部可观性,无人机仅依靠当前时刻的观测信息难以评估环境状态以执行安全合理的导航策略,这使得无人机的导航算法经常陷入局部最优解,甚至发生碰撞。所以,需要一个能够整合历史信息对未知环境状态评估的运动规划模块。循环神经网络(recurrent neural network, RNN)是一类具有短期记忆增强能力的神经网络,可以提取数据中的时间序列特征,将循环神经网络与深度强化学习相结合可以更好处理环境部分可观测问题。

因此,本文将无人机自主导航任务建模为马尔可夫决策过程。在此基础上,提出一种基于记忆增强机制的无人机自主导航算法,通过引入循环神经网络来提取历史信息与当前观测数据中的环境特征,增强导航策略网络在局部可观环境中的状态估计能力,提高导航算法的稳定性。此外,构建无人机自主导航深度强化学习算法的状态空间、动作空间以及引导无人机完成自主导航任务的非稀疏奖励函数,实现在安全避开障碍物的前提下高效地到达指定目标。

1.2　马尔可夫决策过程建模

强化学习(reinforcement learning, RL)是研究智能体如何与环境交互并获得最大奖励回报的学习方法。智能体与环境不断交互获得数据,生成对应的策略,以获得环境反馈信号(惩罚或奖励)并将其作为智能体选择动作的评价标准。如此反复迭代优化其策略,直到最大化奖励反馈信号。在强化学习中,智能体依据当前的环境状态制定策略,这个迭代过程具有马尔可夫性质。因此,强化学习可以被建模为马尔可夫决策过程(Markov decision process, MDP)。马尔可夫决策过程可以被表示为五元组$\langle \boldsymbol{S},\boldsymbol{A},P,R,\gamma \rangle$,其中$\boldsymbol{S}$是智能体状态空间集合;$\boldsymbol{A}$表示智能体动作空间集合;$P$表示状态转移的条件概率,$P^{a}_{SS'}=P[S_{t+1}=S'\,|\,S_t=s,A_t=a]$。$R$表示执行动作之后获得的奖励反馈信号,$R^{a}_{S}=\mathrm{E}[R_{t+1}\,|\,S_t=s,A_t=a]$;$\gamma\in[0,1]$,被称为折扣因子,用于描述奖励值随时间推移而产生的折扣。该过程的最终目标是建立一个由状态空间到最优控制量的策略:$\pi(a\,|\,s)=P[A_t=a\,|\,S_t=s]$。

马尔可夫决策过程的建立是研究强化学习算法的基础。在无人机自主导航任务中,无人机首先得到当前时刻自身所处的状态,包含当前的观测值、速

度、位置等信息。然后,依据状态转移方程制定相应的策略,得到下一时刻的状态。因此,本文将无人机自主导航任务定义为马尔可夫决策过程,从而将其纳入强化学习框架求解。如图 1 所示,基于强化学习的无人机自主导航框架由两部分组成:环境(environment)与智能体(agent)。在交互过程中,无人机在环境里面获取到状态(state)。由于环境的局部可观性以及传感器的性能限制,精确的环境状态量难以获得。因此,本文采用环境观测量替代环境状态量,包括无人机的单目相机获得的深度图像以及无人机与目标的位置和距离等信息,并以此建立局部可观条件下的马尔可夫过程。无人机会根据环境测量状态进行决策,输出动作(action),包括无人机的偏航角、速度等。为了平滑运动轨迹,提高无人机动作连续性,本文采用连续型动作空间。最后,根据奖惩回报函数得到交互过程的反馈奖励,引导无人机选择最佳的控制策略。

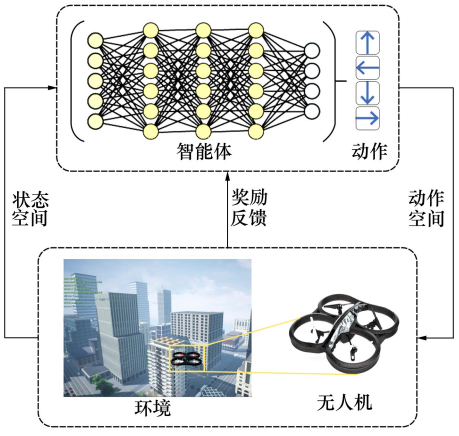


图 1 基于强化学习的无人机自主导航框架

1.3 SAC 算法框架

SAC(soft actor-critic)[15] 是一种 off-policy 的随机性策略算法。在最大化期望回报的基础上,引入最大化动作的随机策略,即熵的概念,来学习在多模态奖励下的最优策略。一般地,在实际问题中,最优解往往不止一个,确定性策略模型容易陷入局部最优解,而基于最大熵理论的 SAC 通过增加动作熵使其可以探索更加广阔的状态-动作空间,学习到最大化累计奖励和最大熵,面对观测信息的干扰和噪音,具有更高的稳定性和鲁棒性。

如图 2 所示,SAC 算法框架包含一个策略网络 π_θ,2 个评价网络 Q_{φ_1} 和 Q_{φ_2}、2 个目标评价网络 $Q_{\varphi_1'}$ 和 $Q_{\varphi_2'}$,其中,$\theta,\varphi_1,\varphi_2,\varphi_1',\varphi_2'$ 表示对应网络的参数。

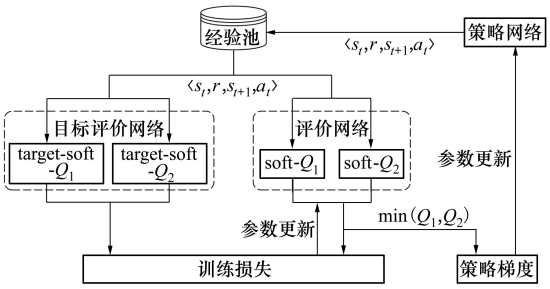


图 2 SAC 算法框架

策略网络用于将输入的观测信息映射为动作空间高斯分布的均值和方差,然后从概率分布采样输出动作 a_t。智能体执行动作后,获取奖励 r_t,状态变为 s_{t+1}。一般地,强化学习的目标是最大化策略的累计奖励信号,而 SAC 算法不仅要保证获得最大奖励信号,还需要兼顾输出动作的最大熵,即输出动作要具有一定的随机性,这种设计提升了智能体的探索能力和策略的鲁棒性。因此,策略网络的目标函数为

$$\pi^* = \max_\theta \mathrm{E}_s\left[\min_{i=1,2} Q_{\varphi_i(s_t,a_t')} - \alpha H(\pi(\cdot|s_{t+1}))\right] \tag{1}$$

式中:$H(\pi(\cdot|s_{t+1}))$ 表示动作熵;α 为策略随机阈值,在训练过程中自动调节,该值的大小表示策略探索的随机性强弱。为了提升训练效果加快收敛速度,还要结合折扣因子 γ 和学习率 l。因此,动作值函数值的更新策略为

$$Q'_{\mathrm{soft}}(s_t,a_t) = Q_{\mathrm{soft}}(s_t,a_t) + l[R_t(s_t,a_t) + \gamma R_{t+1}(s_{t+1},a_{t+1})] \tag{2}$$

为了抑制动作值函数的过高估计,同时提高算法稳定性,SAC 的评价网络采用双深度 Q 网络估计状态值,评价网络的损失函数 $L(\varphi,D)$ 为

$$L(\varphi,D) = \mathrm{E}_{(s_t,r,s_{t+1},a_t)\sim D}[(Q_{\varphi_i}(s_t,a_t) - (Q_s(s_{t+1}))^2] \tag{3}$$

式中:$(s_t,r,s_{t+1},a_t) \sim D$ 表示 (s_t,r,s_{t+1},a_t) 来自于经验池 D。$Q_{\varphi_i(s_t,a_t)}$ 是权重为 φ_i 的网络对环境状态 s_{t+1} 下实施动作 a_t 的 Q 值估计。在每个时间步长,智能体选择其中较小的 Q 值作为目标值。

1.4 策略网络设计

在无人机自主导航过程中,无人机使用机载单目深度相机不断采集环境观测信息作为环境观测

量。事实上,受传感设备性能和遮挡环境的影响,每个时刻无人机都只能从有限视角实现对环境的部分观测。在现有的大多数导航算法中,无人机在制定导航策略时,仅仅依赖于当前观测信息,而没有考虑到历史观测信息,这将导致无人机无法收集足够的环境信息来制定适当的导航决策。例如,在长廊、角落等狭窄区域中飞行时,受相机视野的限制,无人机会长期反复盘旋以寻找路径,甚至发生碰撞。因此,如果能够通过结合之前的历史信息,弥补当前环境观测信息的不足,无人机就可以制定更加合理的飞行策略。因此,在决策网络中引入历史观测信息以便更好地评估环境状态很有必要。

为了解决这个问题,本文提出了一种具有记忆增强结构的决策网络,以处理观测数据的时序依赖关系。具体而言,如图 3 所示,本文使用 LSTM 单元作为记忆增强结构的基础,并将其嵌入到 SAC 算法的策略网络,称为 SAC-LSTM。其中,$\boldsymbol{h}_i$ 是 LSTM 单元隐藏向量,w_i 是每个 LSTM 单元输出的权重,表示当前观测中每个帧的注意力权重,每个学习的权重都依赖于来自先前时间步长的信息和当前状态信息,为了学习正确的行为,优化过程可以学习选择哪些状态相对更重要。通过记忆增强模块将一组连续的历史数据作为输入,并分析网络中包含的时序信息,从而整合连续历史信息,避免算法陷入局部最优解。从而提高算法在局部可观察环境中的性能,使生成的导航策略更加稳定。

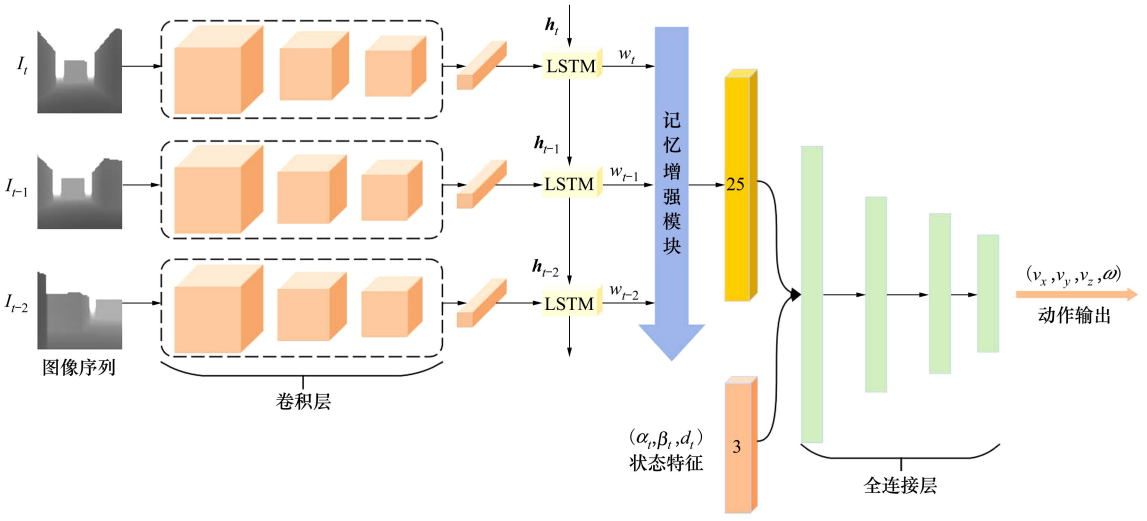


图 3　自主导航策略网络框架

1.5　状态空间

状态空间表示无人机感知到的环境信息和动态变化,这是强化学习算法制定决策和获得累计收益的依据。状态空间设计的质量直接决定深度强化学习算法的收敛速度和最终性能。

在无人机自主导航任务中,无人机在 t 时刻的状态空间由一组环境状态与无人机的自身状态共同组成。由于不可获得环境状态的全部信息,本文使用无人机对环境的观测信息作为环境状态。具体而言,无人机利用单目深度相机作为视觉传感器。因此,环境观测数据为一系列的深度图像 I。在输入到神经网络计算之前,深度图像需要进行预处理过程,以降低数据复杂度,保留主要信息,提高算法执行效率。本文选用的深度相机分辨率为 640×480,将其进行缩放处理之后,转换为 160×80,并将连续 3 帧的深度图像叠加,作为环境观测数据 $\boldsymbol{S}_{\text{obs}}=[I_{t-2},I_{t-1},I_t]$。此外,无人机状态特征可以定义为 $\boldsymbol{S}_{\text{state}}=[\omega_t,\beta_t,d_t]$,其中 ω_t,表示无人机的偏航角,β_t 表示无人机的前进方向和目标位置之间的角度,d_t 表示表示无人机当前位置与目标位置之间的欧氏距离。因此,无人机状态空间可以表示为 $\boldsymbol{S}=[\boldsymbol{S}_{\text{obs}},\boldsymbol{S}_{\text{state}}]$,这些数据共同表征了无人机所处环境及其执行决策的状态。

1.6　动作空间

为了保持无人机运动过程中的连贯性,本文选择仿真环境中的无人机连续动作指令集作为控制通道,设计了无人机的连续动作空间。如表 1 所示,3 个方向的速度和偏航角被映射为四自由度连续动作空间。飞控系统将这些控制量解算为无人机的机动指令,这与真实条件下的无人机控制方式相同。

表 1 无人机动作空间表

符号定义	动作描述	取值范围
v_x	Q_{x_d} 轴方向的速度	(0.5,3.0)
v_y	Q_{y_d} 轴方向的速度	(−2.0,2.0)
v_z	Q_{z_d} 轴方向的速度	(−1.0,1.0)
ω	偏航角	(−30°,30°)

1.7 奖励函数

奖励函数代表深度强化学习模型的学习目标。智能体通过奖励函数获得反馈信号,这反映了智能体策略优劣。自主导航的目的是控制无人机在复杂环境中无碰撞地到达指定目标位置,该过程可以分解为目标接近和避障 2 个子任务。因此,当无人机安全到达目的地时应予以正反馈奖励。反之,当发生碰撞时,应给予负反馈以示惩罚,并马上终止当前训练周期,重新开始训练。然而,这种反馈信号只有在训练周期终止时才能得到,因此奖励十分稀疏,无人机会因为缺乏奖励信号而难以学习到稳健的导航策略。为此,本文设计了多项稠密奖励,以引导无人机更加高效地探索环境,加速算法收敛过程。根据以上设计原则,本文的奖励函数具体为:

1) 距离奖励

为引导无人机接近目标点,当无人机靠近目标时应该得到正向的奖励信号,远离目标时应给予惩罚信号,因此,连续型距离奖励函数的定义为

$$R_1 = d_{\text{pre}} - d_{\text{cur}} \tag{4}$$

式中:d_{pre} 为无人机前一位置与其目标位置之间的欧式距离。d_{cur} 是当前位置和目标位置之间的欧氏距离。当无人机接近终点时,奖励为正;否则,奖励为负。

2) 角度奖励

为激励无人机以正确的角度飞向目标,无人机的速度方向应该与目标距离的矢量方向一致。因此,本文设计连续型奖励函数,具体为

$$R_2 = \cos(\boldsymbol{l}_{\text{vector}}, \boldsymbol{v}) \tag{5}$$

式中:$\boldsymbol{l}_{\text{vector}}$ 为以无人机当前位置为起点,目标位置为终点的向量;$\boldsymbol{v}$ 为无人机速度组成的向量。

3) 终止奖励

在每个训练回合中,为了保证训练的进度,需要设置合适的训练终止条件,本文终止条件分为 4 种:无人机与障碍物碰撞、到达目的地、超时、飞出边界。因此,本文将这些终止训练的情况设计为离散型奖励函数:

$$R_3 = \begin{cases} -100, & \text{碰撞} \\ +100, & \text{到达目的地} \\ -50, & \text{飞出边界} \\ -50, & \text{超时} \end{cases} \tag{6}$$

综上所述,总奖励为

$$R_{\text{total}} = \gamma_1 R_1 + \gamma_2 R_2 + \gamma_3 R_3 \tag{7}$$

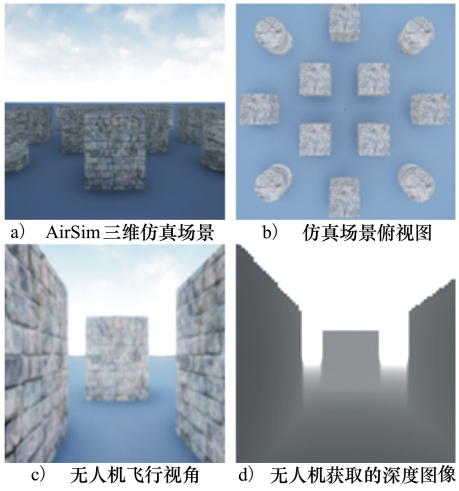
式中:$\gamma_1, \gamma_2, \gamma_3$ 表示各奖励因素的权重,本文设定 $\gamma_1 = 0.5, \gamma_2 = 0.2, \gamma_3 = 0.3$。

2 仿真验证与分析

2.1 仿真实验场景

本文选用开源无人驾驶仿真平台 AirSim[16] 作为无人机仿真器。AirSim 具有逼真的视觉纹理与动力学物理引擎,提供单目相机、IMU、激光雷达以及测力计等丰富的传感器数据,支持无人机类型、飞控和通信协议的拓展性更新与修改。在 AirSim 中验证的算法可以直接部署到真实机载计算机。

为了验证所提自主导航方法的可行性,本文开发了一个三维无人机自主导航场景,如图 4 所示,该场景包含丰富的障碍物,具有较高的复杂度,接近真实场景。在每个训练回合中,无人机使用单目深度相机采集深度图像作为环境局部观测信息,无人机从一个随机的位置起飞,穿越复杂地形飞往目标位置,为了保证安全,飞行过程中应避开障碍物。



a) AirSim 三维仿真场景 b) 仿真场景俯视图
c) 无人机飞行视角 d) 无人机获取的深度图像

图 4 无人机自主导航仿真场景

仿真场景实验的训练过程在服务器中完成。其中,服务器硬件配置为:AMD Ryzen 7 5800H CPU,

Nvidia GTX 3060 GPU 显卡,内存 32 GB;服务器的软件配置为:AirSim 1.6, Python 3.7, CUDA 10.3, Torch 1.10.0, Window 10 操作系统。

2.2　算法相关参数设置

本文的算法参数设置如表 2 所示。其中,探索率的初始值为 0.95,这是为了使无人机在训练初期能够以较大的概率探索海量的状态空间,避免陷入局部最优状态。随着训练的进行,探索率逐渐衰减为 0.05,这是为了使无人机在训练中能够学习到一些罕见动作的策略。

表 2　SAC-LSTM 算法参数设置表

参数	取值范围	参数	取值范围
学习率	2×10^{-4}	α 初始值	1.0
折扣因子	0.95	梯度步数	20
训练批次大小	128	探索策略	随机采样
每回合迭代步长	10^2	初始探索率	0.95
经验池容量	10^6		

2.3　自主导航仿真实验分析

2.3.1　训练结果对比分析

为了评估所提算法的有效性,本文分别将 DDPG[17]、SAC[15]、TD3[18]、PPO[13] 与本文提出的自主导航算法 SAC-LSTM 在仿真环境中进行训练测试。训练开始时,无人机初始奖励值为 0,当无人机寻找到正确路线从而到达设定目的地时会得到 +100奖励。与此同时,碰到障碍物时会得到相应惩罚 -100。奖励值为训练时每一步获得的奖励之和的均值。因此,奖励值越大表示无人机能以较大的概率躲避所有障碍物到达目标位置,即无人机学习的策略效果越好,避障与路径规划能力越强。基于本文设计的奖励函数,若直接利用一次采样周期的累积奖励值作为算法性能衡量指标可能不能准确反映算法性能,因为当随机目标点离无人机初始位置较远时,初期朝向目标点飞行累积较高奖励,最终即使产生碰撞累积奖励和也会较大,这种情况衡量算法性能可能会有偏颇。为了消除随机误差带来的影响,本文选择 200 次采样(episode)所计算出的平均奖励值和导航成功率作为评价指标,这样设计的指标并不会受到某一次较差采样的影响,又不会出现上述累积奖励和的问题。

图 5~6 展示了 5 种算法的平均奖励和导航成功率变化曲线。从图中可以看出,PPO 算法的性能有限,虽然整个训练过程中获得的平均奖励和导航成功率在逐渐增加,但是训练迭代次数到 8 000 次之后获得的奖励才大于 100,最终的导航成功率仅收敛到 62%,这表明其在复杂的自主导航任务中学到的策略性能较低。DDPG 算法在训练 6 000 次之后平均奖励才能大于 100。尽管训练后期,其奖励值大于 150,但是波动较大,这表明该算法容易出现过拟合的情况,无人机的导航策略陷入了局部最优。相较于 DDPG,TD3 算法在收敛的奖励值和导航成功率占有优势,最终导航成功率稳定到 80%,这表明 TD3 算法比 DDPG 的探索能力更强,但是依然存在较大波动,算法性能不够稳定。

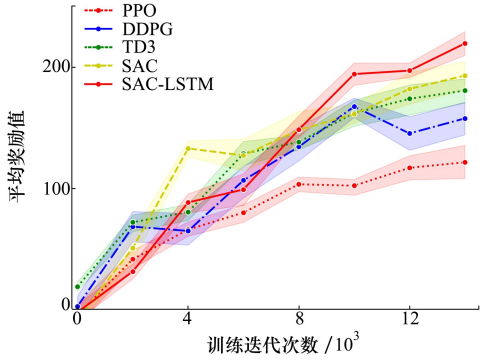


图 5　5 种算法平均奖励值曲线对比图

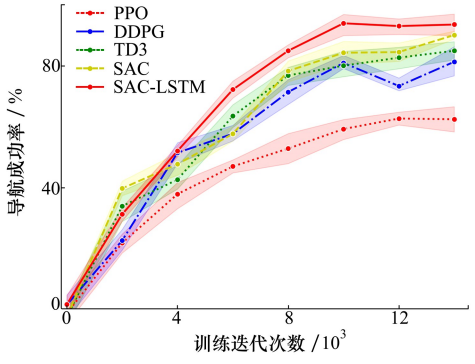


图 6　5 种算法导航成功率曲线对比图

在整个训练阶段,SAC 算法的平均奖励和导航成功率保持平稳增涨,在训练迭代到 4 000 次时,平均奖励值就超过 100。这表明 SAC 算法的最大熵探索策略可以引导无人机探索更大的状态空间,无人机导航策略的稳定性和合理性更高。尽管在训练初期,具有记忆增强模块的 SAC-LSTM 算法花费了更多的训练迭代次数,导致在训练开始时平均奖励增长速度慢于 SAC 算法,但是训练到 4 000 次之后,其平均奖励增长迅速,导航成功率达到 92%,最终超

过了 SAC 算法的表现。这表明 SAC-LSTM 算法的记忆增强模块能够感知到更加充分的环境特征,使得生成的导航策略更加稳健,避免陷入局部最优,探索能力更强,显著提高了无人机的自主导航性能。

2.3.2 测试结果对比分析

为了验证算法有效性,本文分别采用上述 5 种方法在仿真环境中进行测试,测试环境中无人机初始起点在地图上的相对位置坐标为[18,18,2],要求到达指定目标点位置为[-30,-20,2]。在测试过程中,无人机首先从初始位置出发,调整姿态并选择目标接近策略,朝目标位置飞行。当遇到障碍物后,制定适当的避障策略,继续接近目标位置。

表 3 展示了 5 种算法 200 次测试的碰撞率、超时率和平均飞行距离实验平均值结果。其中碰撞率和超时率分别表示发生碰撞回合数和超时回合数占总测试回合数的比例,平均飞行距离表示在成功回合中无人机飞行轨迹长度的均值和标准差(mean±std)。从表中数据分析可知,PPO 算法的碰撞率和超时率明显高于其他算法,表明该方法难以学习到复杂环境下的导航避障策略。相较于 PPO、DDPG 及 TD3 在碰撞率和超时率指标上具有明显的优势,但是在平均飞行距离指标上 TD3 算法有更好的表现。SAC 算法在整体指标的表现比 DDPG 和 TD3 更好,这表明算法很好地平衡了探索过程和经验的利用过程,没有丢失具有较高奖励值的经验。SAC-LSTM 算法在整体指标上取得了最佳表现,这些结果表明基于记忆增强机制的 SAC-LSTM 算法自主导航策略的探索性能更强,更易于探索到有意义的经验,取得了最好的训练效果和鲁棒性。

表 3 5 种算法的自主导航测试结果对比

测试方法	碰撞率/%	超时率/%	平均飞行距离/m
PPO	23.8	35.2	324.3±53.2
DDPG	10.3	13.7	289.5±42.3
TD3	9.6	12.4	275.3±34.6
SAC	8.2	11.6	246.5±32.9
SAC-LSTM	5.5	8.9	193.6±32.8

为进一步展示无人机自主导航算法的效果,无人机自主导航飞行轨迹如图 7 所示。

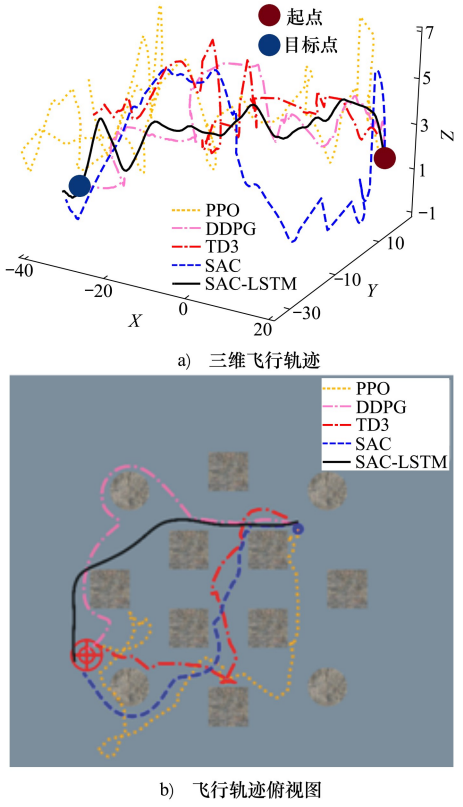


图 7 5 种算法的自主导航飞行轨迹可视化

通过观察可以发现,尽管所有的方法都可以成功到达目标点,但是不同方法生成的飞行轨迹存在很大差异。具体而言,PPO 算法的飞行轨迹波动较大,容易发生碰撞导致任务失败。DDPG 和 TD3 算法通过不断地局部探索来调整飞行行为,飞行轨迹中存在明显的“往复”飞行现象与“大起大落”的抖动,浪费了大量飞行时间,这表明无人机的飞行动作连贯性差,容易陷入局部循环甚至发生碰撞导致飞行失败。SAC 算法的飞行轨迹更加平滑,并且当进入狭窄区域时,无人机也可以在经过少量探索之后,安全绕行,并及时调整飞行方向朝目标运动。但是,当遇到障碍物密度较大情况时,无人机又会因陷入局部最优解,出现“往复”飞行问题。相较而言,采用 SAC-LSTM 算法的无人机在遇到障碍物时,更加倾向于主动调整飞行高度,而不是选择进入狭窄区域绕行通过,从而以更小的代价避开障碍物,使得飞行轨迹更加平滑。这表明本文所提的记忆增强单元能够使得无人机在决策时利用的数据更多,增强局部可观条件下的状态估计准确性,引导无人机摆脱局部最优解。因此,无人机不存在陷入局部“往复”

现象,从而飞行轨迹更加合理,符合真实三维场景的自主导航要求。

3 结　论

本文以未知环境中的无人机自主导航任务为研究背景,针对现有强化学习自主导航算法面临环境局部可观且感知信息不足问题,设计了一种基于记忆增强机制的无人机自主导航深度强化学习算法,通过对历史记忆信息与当前的观测整合处理,提取观测数据的时序依赖关系,从而增强环境局部可观条件下的状态估计能力,避免陷入局部最优解。大量的仿真实验结果表明,本文所提算法能够有效处理无人机自主导航过程中的环境局部可观问题,与现有方法相比,具有良好的稳定性,可以为复杂条件下无人机自主导航领域的工程应用提供理论基础和关键技术支撑。

由于研究方法以及实验条件的局限性,本文的研究仍然存在进一步完善的空间:1)基于强化学习的自主导航算法避免了对环境先验地图的依赖,但是难以确保无人机的导航路径是最优的,因此未来可以考虑将强化学习和传统轨迹优化策略相结合的方式解决这一问题;2)基于人工智能技术的自主导航算法依靠于无人机与环境交互采集大量数据来训练深度神经网络模型。但是无人机的真实飞行数据获取代价较大,因此现有的研究大多数只能在仿真环境下进行测试和验证。未来将研究通过迁移学习的方式将仿真环境中训练的模型迁移到真实物理设备,实现将人工智能技术应用到真实无人机系统。

参考文献:

[1] 武文亮, 周兴社, 沈博, 等. 集群机器人系统特性评价研究综述[J]. 自动化学报, 2022, 48(5): 1153-1172
WU Wenliang, ZHOU Xingshe, SHEN Bo, et al. A review of swarm robotic systems property evaluation research[J]. Acta Automatica Sinica, 2022, 48(5): 1153-1172 (in Chinese)

[2] KOU K, YANG G, ZHANG W, et al. UAV autonomous navigation based on multi-modal perception: a deep hierarchical reinforcement learning method[C]//China Intelligent Robotics Annual Conference, 2023

[3] 张云燕, 魏瑶, 刘昊, 等. 基于深度强化学习的端到端无人机避障决策[J]. 西北工业大学学报, 2022, 40(5): 1055-1064
ZHANG Yunyan, WEI Yao, LIU Hao, et al. End-to-end UAV obstacle avoidance decision based on deep reinforcement learning [J]. Journal of Northwestern Polytechnical University, 2022, 40(5): 1055-1064 (in Chinese)

[4] ALMAHAMID F, GROLINGER K. Autonomous unmanned aerial vehicle navigation using reinforcement learning: a systematic review[J]. Engineering Applications of Artificial Intelligence, 2022, 115(1):105321

[5] ARAFAT M Y, ALAM M M, MOH S. Vision-based navigation techniques for unmanned aerial vehicles: review and challenges [J]. Drones, 2023,7(2): 89

[6] GANDHI D, PINTO L, GUPTA A. Learning to fly by crashing[C]//IEEE International Conference on Intelligent Robots and Systems, 2017

[7] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84-90

[8] MACIEL-PEARSON B G, AKCAY S, ATAPOUR-ABARGHOUEI A, et al. Multi-task regression-based learning for autonomous unmanned aerial vehicle flight control within unstructured outdoor environments[J]. IEEE Robotics and Automation Letters, 2019, 4(4): 4116-4123

[9] CHOI S Y, CHA D. Unmanned aerial vehicles using machine learning for autonomous flight; state-of-the-art[J]. Advanced Robotics, 2019, 33(6): 265-277

[10] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Human-level control through deep reinforcement learning[J]. Nature, 2015, 518(7540): 529-533

[11] WALVEKAR A, GOEL Y, JAIN A, et al. Vision based autonomous navigation of quadcopter using reinforcement learning[C]// International Conference on Automation, Electronics and Electrical Engineering, 2019

[12] KABAS B. Autonomous UAV navigation via deep reinforcement learning using ppo[C]//Signal Processing and Communications

Applications Conference, 2022
[13] SCHULMAN J, WOLSKI F, DHARIWAL P, et al. Proximal policy optimization algorithms[J/OL]. (2017-07-20)[2023-03-27]. https://arxiv.org/abs/1707.06347
[14] HE L, AOUF N, WHIDBORNE J F, et al. Integrated moment-based lgmd and deep reinforcement learning for UAV obstacle avoidance[C]//IEEE International Conference on Robotics and Automation, 2020
[15] HAARNOJA T, ZHOU A, ABBEEL P, et al. Soft actor-critic: off-policy maximum entropy deep reinforcement learning with a stochastic actor[C]//International Conference on Machine Learning, 2018
[16] SHAH S, DEY D, KAPOOR A. Airsim: high-fidelity visual and physical simulation for autonomous vehicles[C]//The 11th International Conference on Field and Service Robotics, 2017
[17] LILLICRAP T P, HUNT J J, PRITZEL A, et al. Continuous control with deep reinforcement learning[J/OL]. (2015-09-09)[2023-03-27]. https://arxiv.org/abs/1509.02971
[18] FUJIMOTO S, HOOF H, MEGER D. Addressing function approximation error in actor-critic methods[C]//International Conference on Machine Learning, 2018

Exploring UAV autonomous navigation algorithm based on soft actor-critic

KOU Kai, YANG Gang, ZHANG Wenqi, LIU Xincheng, YAO Yuan, ZHOU Xingshe

(School of Computer Science, Northwestern Polytechnical University, Xi'an 710072, China)

Abstract: The existing deep reinforced learning algorithms cannot see local environments and have insufficient perceptual information on UAV autonomous navigation tasks. The paper investigates the UAV's autonomous navigation tasks in its unknown environments based on the nondeterministic policy soft actor-critic (SAC) reinforced learning model. Specifically, the paper proposes a policy network based on a memory enhancement mechanism, which integrates the historical memory information processing with current observations to extract the temporal dependency of the statements so as to enhance the state estimation ability under locally observable conditions and avoid the learning algorithm from falling into a locally optimal solution. In addition, a non-sparse reward function is designed to reduce the challenge of the reinforced learning strategy to converge under sparse reward conditions. Finally, several complex scenarios are trained and validated in the Airsim+UE4 simulation platform. The experimental results show that the proposed method has a navigation success rate 10% higher than that of the benchmark algorithm and that the average flight distance is 21% shorter, which effectively enhances the stability and convergence of the UAV autonomous navigation algorithm.

Keywords: deep reinforced learning; soft actor-critic; unmanned aerial vehicle; autonomous navigation

引用格式: *寇凯, 杨刚, 张文启, 等. 基于 SAC 的无人机自主导航方法研究[J]. 西北工业大学学报,2024,42(2): 310-318*
KOU Kai, YANG Gang, ZHANG Wenqi, et al. Exploring UAV autonomous navigation algorithm based on soft actor-critic [J]. Journal of Northwestern Polytechnical University, 2024, 42(2): 310-318 (in Chinese)

© 2024 Journal of Northwestern Polytechnical University.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.