https://doi.org/10.1051/jnwpu/20224050970

基于深度循环双 Q 网络的无人机避障算法研究

魏瑶[1], 刘志成[2], 蔡彬[3,4], 陈家新[3,4], 杨尧[5], 张凯[5]

(1.西北工业大学 航天学院, 陕西 西安 710072;
2.空军装备部驻北京地区军事代表局驻天津地区第三军事代表室, 天津 300000;
3.上海航天控制技术研究所, 上海 201109;
4.中国航天科技集团有限公司红外探测技术研发中心, 上海 201109;
5.西北工业大学 无人系统技术研究院, 陕西 西安 710072)

摘 要:针对传统强化学习方法在机器运动规划领域,尤其是无人机避障问题上存在价值函数过度估计以及部分可观测性导致网络训练过程中训练时间长、难以收敛的问题,提出一种基于深度循环双 Q 网络的无人机避障算法。通过将单网络结构变换为双网络结构,解耦最优动作选择和动作价值估计降低价值函数过度估计;在双网络模块的全连接层引入 GRU 循环神经网络模块,利用 GRU 处理时间维度信息,增强真实神经网络的可分析性,提高算法在部分可观察环境中的性能。在此基础上,结合强化学习优先经验回放机制加快网络收敛。在仿真环境中分别对原有算法以及改进算法进行测试,实验结果表明,该算法在训练时间、避障成功率以及鲁棒性方面均有更好的性能。

关 键 词:深度强化学习;无人机;避障;循环神经网络;DDQN
中图分类号:TP312 **文献标志码**:A **文章编号**:1000-2758(2022)05-0970-10

人工智能的高速发展使与其关联紧密的机器人行业获得了强大的生命力,无人飞行器(unmanned aerial vehicle,UAV)技术作为其中一个重要的组成部分,也借助着这蓬勃的发展趋势得到了显著提升。随着无人机在各个行业的使用越来越普遍,其安全问题也日益受到关注,而无人机是否能够在充满障碍物环境中成功避开障碍物安全到达指定地点完成任务成为衡量其安全性能的关键标准。因此,有效的避障功能对于无人机是十分重要的。

基于视觉和传感器的自主避障方案[1-4],通过实时采集周围环境信息,模拟人类对外界事物的感知能力,一直以来是无人机等移动机器人领域的研究热题。然而这种首先通过对环境建模,再对已知环境进行避障的方法其有效性过多依赖于无人机自身动力学和所构建的环境模型的准确性。随着环境复杂度的增高,需要实时计算和处理的数据量大幅上涨,难以弥补传统模型在建模过程中出现的错误。当无人机初次到达一个新的环境时,其必须再次经历一系列的模型构建任务,这在很大程度上限制了算法的使用范围。

强化学习作为机器学习领域的一个重要分支,与深度学习相比较,深度学习侧重于识别和表达,强化学习则侧重于通过与环境进行反复交互从而寻找解决问题的策略。其思想源于行为心理学,在特定的环境中让智能体感觉“舒适”的行为会加强该动作与环境的联系,反之则削弱联系[5-6]。训练时,智能体发出动作与环境进行交互,每次交互后根据环境反馈的回报值判定当前的行为是否利于实现目标,通过记录智能体经历的状态转换及经验,不断迭代更新自身策略,直到策略收敛到最优。

在强化学习中,智能体下一时刻状态仅与当前状态有关而与之前状态无关。由于强化学习过程也服从马尔可夫性,无需对环境进行精确建模和引入先验知识,就可以通过感知环境状态完成从环境状

收稿日期:2021-12-17 **基金项目**:上海航天科技创新基金(SAST2020-070)资助
作者简介:魏瑶(1998—),西北工业大学硕士研究生,主要从事强化学习、无人机避障研究。e-mail:weiyaonwpu@163.com

态到动作映射的学习。基于此,研究人员将强化学习引入无人机避障[7],避免了上述传统算法需要构建复杂环境模型的工作。

2015 年 Deepmind 团队首次提出了一种结合深度学习和增强学习的深度增强学习框架——Deep Q-Learning[8],该方法可以用于解决状态量为图片信息的增强学习问题。随后 Kulkarni 等[9]对深度增强学习框架做了进一步改进,使其更加适用于自主控制领域。2016 年,由 Sliver 等[10]提出的以深度强化学习算法为基础的围棋算法程序 AlphaGo,是第一个战胜世界围棋冠军的人工智能机器人,开创了人工智能又一个新的里程碑。同年,CAD2RL[11]提出了一种用于室内飞行避障的深度强化学习(deep reinforcement learning,DRL)方法。这项工作使用模拟的 3D 走廊环境训练无人机进行导航。需要生成大量具有不同照明、墙壁纹理、家具放置的 3D 走廊环境图像,由深度 Q 网络在这些图像上学习无人机避障策略。然而这项工作需要大量关于走廊环境图像的数据并且效率不高。此外,由于帮助人脑进行导航的是由双眼输入的深度信息,如果仅将 RGB 图片输入到神经网络之中,无人机无法像人类一样具有避障主观性。由于人脑具有记忆性可以总结和存储解决问题的关键信息,当人类对环境的接触有限或者只能得到部分有关环境的信息时仍旧可以解决日常生活中的各种问题。这种记忆性可以有效存储和回忆起随时间收集的相关信息,以便人脑在不同的场景下做出合适的选择。而无人机的避障和导航也存在着类似的对环境部分可观察性问题[12],需要引入记忆的概念。

综上所述,本文的主要贡献为:①传统强化学习使用单网络结构的无人机避障算法在估计动作价值时反复取用最大理论价值,导致正向误差的累积,本文将单网络结构变换为双网络结构,在训练学习过程中,解耦了最优动作选择和动作价值估计,降低了单网络结构无人机避障算法的过度估计问题,提高了避障算法的性能。②针对无人机部分可观察性问题,提出一种基于循环神经网络架构的避障算法。通过循环神经网络的记忆力增强真实网络的可分析性使其更好处理时间维度的信息并提高算法在部分可观察环境中的中的性能,从而增强无人机的避障能力。使得 UAV 控制器能够随时间收集和存储对环境的观察结果,提升无人机避障效率。

1　深度强化学习原理

1.1　DRL 算法原理

马尔可夫决策过程(Markov decision process,MDP)是 Howard[13]于 1960 年提出的,MDP 常应用于采用动态规划和强化学习解决的优化问题中[14]。无人机在避障过程中,下一时刻无人机所处的状态仅仅依赖于当前场景以及其采用动作后的状态转移方程,而与之前所处的一系类历史状态无关。因此,可以将无人机的避障问题看作是一个马尔可夫决策过程,从而纳入到强化学习框架中进行求解。

由于强化学习的核心概念是奖励,智能体(无人机)的目标便是通过选择策略 π 最大化累计奖励或长期奖励 R_t。如果存在一个策略 π 使得智能体在该策略下执行动作后的动作价值函数或状态价值函数最大,则将该策略定义为最优策略记作 π^*。此时,将动作价值函数定义为 $Q_\pi^*(s,a)$,用来表示智能体处于状态 s 时,在策略 π 下采取动作 a 时的最大预期回报

$$Q_\pi^*(s,a)=\max_\pi \mathrm{E}_\pi\left[\sum_{k=0}^{\infty}\gamma^k R_{t+k+1}\mid s_t,a_t,\pi\right] \quad (1)$$

式中:s_t 为当前状态;a_t 为当前状态下的动作。在传统 RL(reinforcement learning)中,(1)式可通过贝尔曼方程进行迭代更新直至最终收敛,即当 $i\to\infty$ 时,$Q\to Q^*$

$$Q_{i+1}(s,a)=\mathrm{E}[r+\gamma\max_{a'}Q_t(s',a')\mid s,a] \quad (2)$$

式中,s' 和 a' 分别表示当处于(s,a) 状态时,智能体下一步的状态和动作。

深度 Q 网络(deep Q network,DQN)是目前被广泛使用的深度强化学习算法之一,实现了无环境模型的、基于值函数的端到端学习。DQN 在 Q-learning 算法的基础上,引入了深度神经网络和特殊的训练机制,解决传统 Q-learning 在内存空间和处理能力上的局限性,并将强化学习问题转化为可以使用监督学习方式进行训练。

在 DQN[15]中,采用 CNN 网络替代 Q-learning 算法中的状态-动作表格,拟合 Q-learning 算法中的状态值函数,从而解决 Q-learning 算法在存储空间和处理能力上的问题。DQN 算法非常重要的 2 个元素是“经验回放”和“目标网络”,通常情况下,DQN 算法更新是利用目标网络的参数 θ_t^-,它每隔固定步数会更新一次。此时,(2)式可以被表达为

$$Y_t^{\mathrm{DQN}} = R_{t+1} + \gamma \max_a Q(s_{t+1}, a, \theta_t^-) \tag{3}$$

传统的 DQN 算法使用单网络结构,在用于无人机避障算法之中时,由于反复使用最大估计来近似下一个状态的动作值函数,导致 Q 值过高估计。因此需要解耦动作选择与价值网络,选用双 Q 网络避免过高估计问题。则(3)式可以表达为

$$Y_t^{\mathrm{DoubleDQN}} = R_{t+1} + \gamma Q(s_{t+1}, \arg\max_a Q(s_{t+1}, a; \theta_t), \theta_t') \tag{4}$$

式中, θ_t 和 θ_t' 分别为主网络参数以及目标网络参数,此时,用于更新神经网络权重的损失函数可以表示为

$$L_t(\theta_t) = \mathrm{E}[(r + \gamma \max_{a'} Q(s', a'; \theta_t') - Q(s, a; \theta_t))^2] \tag{5}$$

1.2 循环神经网络

无人机避障和导航也存在可观察性问题,需要引入记忆的概念。例如,在避障导航时,无人机可能会飞向角落。当它接近角落时,与侧面相比,深度图可能表明前方有更多的空间。时间信息的缺乏加上单目相机视野有限使得无人机向前移动到角落并撞到墙上。这种场景在无人机导航中非常常见,因此需要一个可以利用相关过去信息的控制器。循环神经网络的主要特点是在一定的时间节点内,神经元的输出可以再次用作神经元的输入。这种序列的网络结构可以很大程度保持数据中的依赖关系非常适合于处理时间序列数据。GRU[16] 作为 RNN 网络的一种衍生网络,与传统的长短期记忆网络[17](long short-term memory,LSTM)相比,其将忘记门和输入门合成了一个单一更新门,混合了细胞状态和隐藏状态为一个新的状态,减少了神经网络的计算量,加快网络训练时间。

因此,将循环神经网络(GRU)与深度强化学习网络相结合可以更好处理无人机避障之中的部分可观测问题,提升算法性能。

2 网络设计

本文中,无人机的任务是通过与环境进行随机交互,在没有任何碰撞的情况下使无人机能够到达指定目标点。在这一过程中,将生成一系列的状态、动作和奖励。传统的强化学习将无人机的避障过程看作是一个全观测的马尔可夫过程,假设无人机可以从环境中完全观察到所处状态,因此下一个时间步的任何序列对于无人机所有可能的动作都是已知的,并且可以通过该序列来学习策略。实际上,由于输入至神经网络的深度图片仅仅来自于无人机上所搭载的双目相机,智能体并不知道自身所处位置以及当前速度,因此需将无人机的避障过程看作一个部分可观测的马尔可夫过程。

本文设计了一种全新的 DRL 模型——深度循环双 Q 网络模型,来解决四旋翼无人机在未知三维环境中的避障问题。在 DDQN 深度强化学习网络结构的基础上,引入循环神经网络 GRU,通过 GRU 单元的记忆力增强真实神经网络的可分析性,使其更好处理时间维度的信息。同时,由于 GRU 层可以记忆历史信息,可以提高算法在部分可观察环境中的性能,从而增强无人机的避障能力。

表 1 DDQN-GRU 网络详细结构

网络层(类型)	输出维度
input(Input Layer)	(None,4,84,84,1)
batch_normalization _1	(None,4,84,84,1)
time_distributed_1(Conv2D+ELU)	(None,4,84,84,32)
time_distributed_2(Maxpooling2D)	(None,4,42,42,32)
time_distributed_3(Conv2D+ELU)	(None,4,38,38,32)
time_distributed_4(Maxpooling2D)	(None,4,19,19,32)
time_distributed_5(Conv2D+ELU)	(None,4,17,17,16)
time_distributed_6(Maxpooling2D)	(None,4,8,8,16)
time_distributed_7(Conv2D+ELU)	(None,4,8,8,8)
time_distributed_8(Flatten)	(None,4,512)
gru_1(GRU)	(None,64)
batch_normalization _2	(None,64)
activation_1(Activation)	(None,64)
dense_1(Dense)	(None,128)
batch_normalization_3	(None,128)
elu_1(ELU)	(None,128)
dense_2(Dense)	(None,128)
batch_normalization_4	(None,128)
elu_2(ELU)	(None,128)
dense_3(Dense)	(None,4)

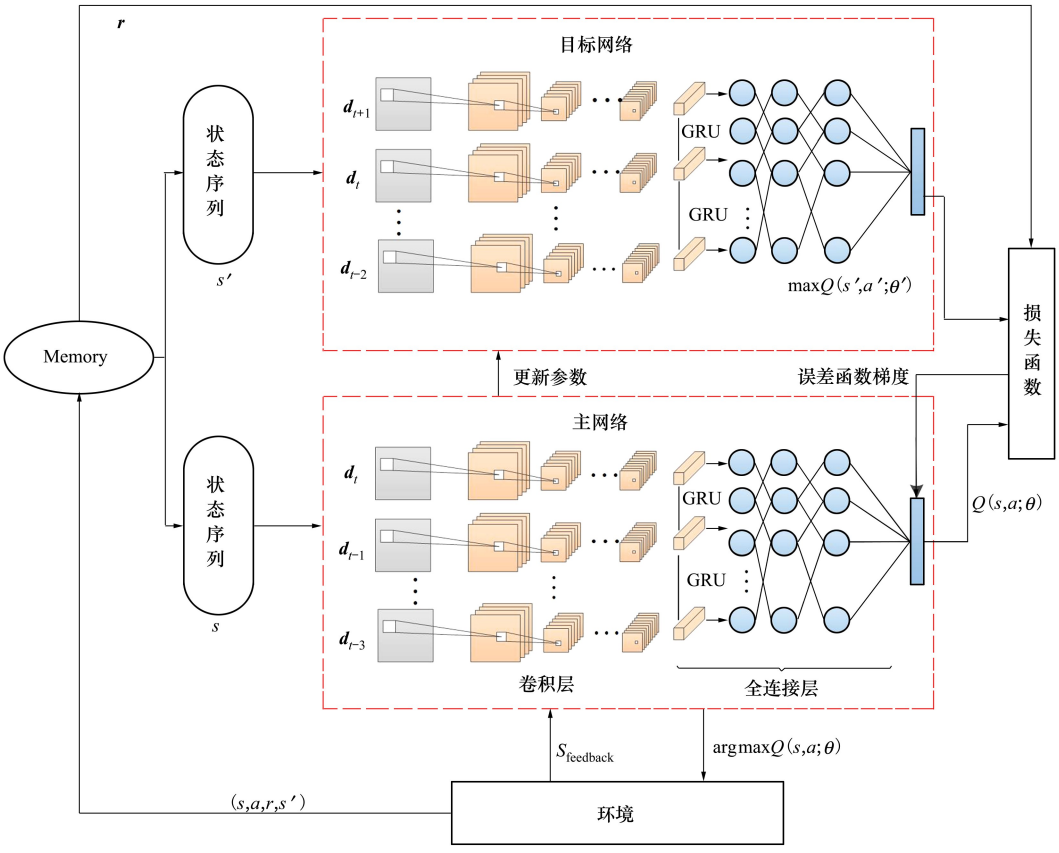


图 1　DDQN-GRU 网络模型架构

具体的网络模型架构如图 1 所示,各层网络参数如表 1 所示。该网络为双网络结构,包含主网络以及目标网络。在本文中将来自环境的连续 4 帧深度图片 d 作为当前状态序列输入到主网络之中,将下一个时间状态序列输入到目标网络之中。最后将不同网络计算得到的 Q 值作为损失函数的输入,从而得到误差值,并根据误差梯度对主网络进行参数更新。主网络与目标网络共享一组参数,但不同步,每隔一固定时间,主网络对目标网络进行参数更新。下一状态的动作来源于主网络所给出的最大 Q 值时所对应的动作,无人机执行该动作与环境进行交互,交互完成后环境给出反馈。此时,相应的状态奖励以及动作等参数都存储在经验池 Memory 之中。经验池中存储的是每次与环境交互之后即一个 episode 后的数据信息,分别为当前状态、动作、下一个状态,以及环境奖励,表示为(s,a,s',r)。在进行经验抽取时,本文采用优先经验回放代替原有的随机抽取,给予训练数据优先级使得含有有效信息的数据更有可能被抽取训练,从而提高数据的利用率。

3　仿真及实验研究

3.1　仿真实验场景

AirSim 是由美国 Microsoft 公司推出的开源无人机仿真软件[18]。该软件依托著名的游戏开发平台:虚幻引擎(unreal engine)。虚幻引擎提供了高品质的渲染能力和 PhysX 物理引擎,能够便捷地完成无人机、场景和障碍物的模型制作,实现逼真的视觉场景和物理效果。AirSim 支持多种传感器的仿真与自定义,包括摄像机、气压计、惯性测量单元、GPS、磁力计、深度摄像机、激光雷达等。它提供基于 C++和 Python 的多种 API,例如飞控接口、通信接口、图像接口等,拓展性较高,便于进行半实物仿真。因此,本文选择 AirSim 作为仿真系统的基础,开展无人机避障算法的研究。仿真场景试验在仿真计算机中完成,计算机软件配置为:Window 10 操作系统,CUDA 10.0,tensorflow1.12.0,Python3.6。

在本文的仿真实验中,无人机使用深度传感器在其前向 180°的范围内采集来自环境的深度图片作为状态观测值,图 2 为本文所使用的仿真环境。

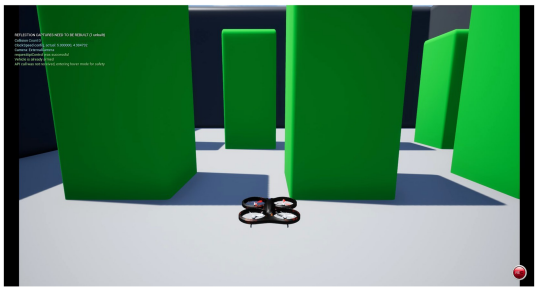


图 2 仿真环境示意图

3.2 状态空间和动作空间设计

状态空间:本文在仿真环境里进行训练,利用深度相机作为感知器,因此状态空间可以定义为一系列的深度图片。在将深度图片 d 作为状态信息,输入到神经网络中进行训练之前,一般需要对图像进行预处理,从复杂信息中提取主要有效信息,以降低数据的复杂度,提高算法的效率。本文将从仿真环境中得到的 640×480 的深度图片转换为 84×84 的灰度图像,并将连续 4 帧深度图片叠加作为此时状态 s_t。此时深度图片的储存格式如公式(6)所示。在本实验中,使用灰度图像不会影响学习性能,并且仅使用深度图片进行训练可以减少 RGB 图片中纹理色彩对训练结果的影响,增强算法的泛化能力并有效提高计算速度。

$$s_t = [d_{t-3}, d_{t-2}, d_{t-1}, d_t] \tag{6}$$

动作空间:为了减少神经网络计算量,降低网络训练难度,去除无人机偏航通道控制,将无人机看作三自由度的质点运动。在此基础上,固定高度和速度,则无人机的动作空间分为向前、向后、向左、向右 4 个方向,具体内容见表 2。

表 2 动作空间表

动作序号	动作描述
0	无人机以 3 m/s 的速度朝前飞行 1 s
1	无人机以 3 m/s 的速度朝后飞行 1 s
2	无人机以 3 m/s 的速度朝左飞行 1 s
3	无人机以 3 m/s 的速度朝右飞行 1 s

3.3 奖励函数设计

本文研究的目的是保证无人机在陌生的未知环境中“存活时间”足够久,同时也要在飞行过程中尽量避开障碍物到达指定目标点,而不是原地不动或者陷入局部最优的状态。这就意味着奖励函数的设计既要激励无人机从原点到达目标点,也要惩罚无人机产生碰撞。奖励函数的设计如公式(7)所示

$$r = \begin{cases} +500, & D_{\text{dis}} \leqslant 0.3 \\ -50, & F = 1 \\ +D_{\text{diff}} & D_{\text{last\_dis}} < D_{\text{dis}} \\ -D_{\text{diff}} & D_{\text{last\_dis}} > D_{\text{dis}} \\ 0 & \text{else} \end{cases} \tag{7}$$

式中,D_{dis} 为无人机与目标地点之间距离,当 $D_{\text{dis}} \leqslant 0.3$时表示无人机已经到达目标地点,则给予无人机+500 的奖励;当碰撞标志位 F 为1 时,表示此时无人机已经产生碰撞,则奖励回报为-50 对无人机进行惩罚;D_{diff}为上一时刻 D_{dis}与该时刻 D_{dis}之差,当上一时刻无人机与目标点之间距离大于该时刻距离时,表示上一时刻无人机执行的动作是有效的因此需要给予奖励;当上一时刻无人机与目标点之间距离小于该时刻距离时,则表明上一时刻的动作是无效的,无人机并没有靠近目标地点,因此给予惩罚。

在本文中将设置无人机产生碰撞时的惩罚项数值比无人机给出无效动作(上一时刻无人机与目标点之间距离小于该时刻距离时)时的惩罚值大得多,原因在于强化学习是回合制(episode)的学习过程。一个回合指的是从无人机起飞的时刻开始计算,直到无人机发生碰撞,或者到达指定目标点,或者到了提前设定好的最大时间长度结束,在此期间经历的所有有限个时间片称之为一个回合。同时,在该决策训练的途中,将当前的回报值分解为当前立即回报和未来的潜在回报,如公式(1)所示,其中 γ 折扣函数统一设置为 0.99,表明此时智能体更加关注未来的奖励值。因此,若是一个回合的时间片个数超过 10,且在这 10 个时间片中无人机不发生任何碰撞,则在第 11 个时间片的时候动作值函数的估计值将在+40 左右。此时,若无人机采取了某个决策动作 a 而导致其发生碰撞,回报函数仅仅给予了无人机一个-10 的惩罚,因此在第 11 个时间片之后无人机对决策动作 a 的动作值函数估计值在+30 左右。可以发现,+30 的动作值函数估计值使得无人机依然认为第 11 个时间片所采取的动作 a 是一个“优秀的动作”。然而实际情况是决策动作 a 导致了无人机的碰撞,因此在公式(7)中将惩罚设置

为-50,很大的惩罚将对无人机的“犯错”具有足够的“威慑力”,在无人机避障策略学习的过程中才会足够“重视”碰撞所带来的惩罚。同理,当无人机能够安全到达指定目标点时,给予无人机一个很大的奖励值,表明该回合的策略是有效的,能够激励训练朝着好的方向发展。在无人机的训练过程中如果仅仅使用主线奖励会导致稀疏奖励问题,为了避免无人机陷入局部最优解状态(表现为无人机在仿真环境中为了“逃避”惩罚几乎呈现出静止不动的状态,偶尔会轻微地左右同时晃动)而使训练无法收敛,需要对比上下 2 个时刻相对目标点距离,对无人机进行 D_{diff}大小的惩罚或奖励引导训练朝着设定目标方向进行。

3.4　仿真结果研究

在本次实验设计中,训练批次设计为 64,学习率为 1×10^{-4},探索率的初始值设为 0.98 并随着无人机训练次数的增加而逐渐递减最终为 0.05 保证无人机在一开始有更大的概率获取随机动作从而避免直接采用神经网络训练进入局部最优状态。当探索率最终衰减至 0.05 之后,也能保证无人机学习得到一些罕见的动作策略。无人机开始训练,初始奖励值为 0,当无人机碰到障碍物时会得到相应惩罚,寻找到正确路线从而到达设定目的地时会得到+500 奖励,无人机训练的总分值为训练时每一步奖励之和。因此训练结束后,当总分值大于 500 时表明在当前训练回合中,无人机能够成功躲避所有障碍物并到达指定目标点,总分值越高,表明无人机所选择的避障路线越好。为了验证本文所提出算法的有效性,本文分别将 DQN、DDQN 、DDQN-GRU 在该仿真环境中进行训练测试,在此基础上引入优先经验回放机制提高数据利用率从而加快训练收敛,验证算法的有效性。

3.4.1　训练结果对比

图 3~6 分别表示 4 种不同算法在同一训练场景下,其总得分、平均奖励以及无人机所走路程随训练次数增加的变化情况。

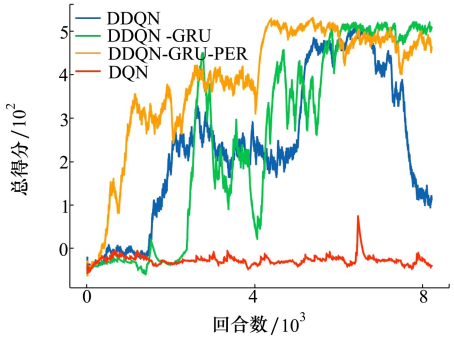


图 3　总得分变化曲线

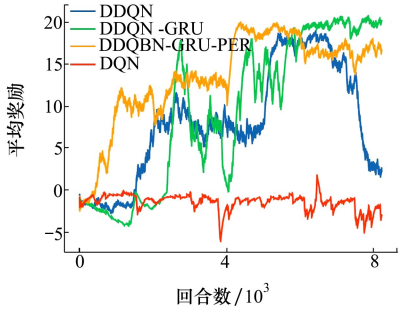


图 4　平均奖励变化曲线

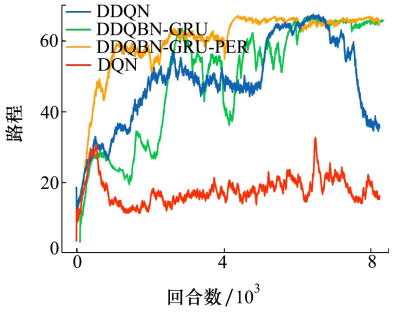


图 5　无人机所走路程变化曲线

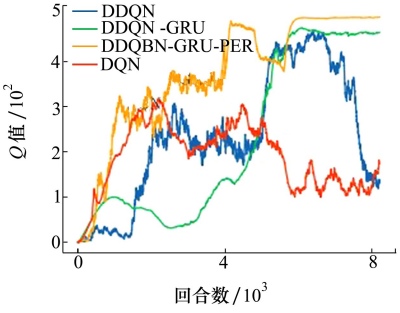


图 6　Q 值变化曲线

如图 3 所示,DQN 算法整个训练过程中均无较大波动,只在 6 200 左右出现得分大于 0 的情况,表明此时无人机只能进行简单避障而并没有到达指定目标位置。与其他算法相比较,经过 8 000 多次训练 DQN 算法所训练无人机避障能力最差。DDQN 算法在训练 5 700 次后得分开始达到 500 左右,表明此时无人机已经具备躲避障碍物并到达地点的能力,但是其训练很容易出现过拟合现象,表现为训练曲线不能完全收敛,无人机陷入局部最优状态。相比之下,DDQN-GRU 算法在整个训练过程中得分总体优于 DDQN,在 5 500 次训练过后算法整体收敛稳定,在引入优先经验回放机制之后算法训练速度加快,在 4 000 次训练之后便可以达到收敛状态。平均奖励为每一训练回合中无人机的总得分除以其训练步数,意在评价算法每一步中给出策略的好坏,从图 4 可以看出每一次训练中,由于引入了 GRU 网络

和优先经验回放机制，DDQN-GRU-PER 的平均奖励均大于 DDQN 算法以及 DQN 算法。图 5 为无人机所走路程的变化曲线，当无人机在飞行过程中遇到障碍物时会受到惩罚并结束本次训练过程，因此可以看出随着训练次数的增加无人机的避障能力在增强。通过对比可以看出 DDQN 算法的有效性明显高于 DQN 算法。图 6 所示为一个训练回合即一个 episode 中神经网络所给出策略的 Q 值（动作价值）平均值。值得注意的是，Q 值越高并不完全代表算法所给出的策略越好，神经网络所给出的动作价值应该尽可能地接近人为设置的奖励值。通过对比可以看出 DQN 由于反复取用最大理论价值，导致正向误差累积，在开始训练的 2 000 次给出动作价值过高从而影响正确的策略判断。DDQN 算法由于使用 2 个单独的网络进行价值估计和动作选择，可以有效避免对动作价值的过高估计。

3.4.2 测试结果对比

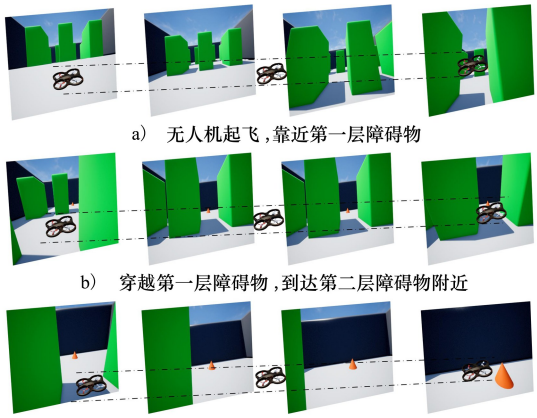
为了验证算法的有效性，本文分别将 DQN、DDQN 网络模型与 DDQN-GRU 网络模型在测试环境中进行对比，通过比较累积得分、单步平均奖励以及无人机所走路程来评估算法模型的有效性。平均累积得分、平均奖励得分值越大表明无人机在该测试回合中所采取的避障策略越有效，即所训练模型表现越好。根据本文奖励函数的设计可知，当累积得分高于 500 时，表明无人机能够成功躲避所有障碍物并安全到达指定目标点。测试环境中无人机出生点在地图上的相对位置坐标为[0,0,-7]，要求到达指定目标点位置为[3,-76,-7]。当无人机在测试回合中所走路程并未超过 76.059（出生点与终点之间最短距离）时，表明无人机可能在该回合中出现碰撞而没有成功到达指定目标点。

表 3 为 1 500 次测试实验取平均值结果，从表中数据可以看出 DQN、DDQN 模型在测试过程中平均累积得分、平均奖励出现负值，平均路程不到出生点至目标点最短距离的一半，表明在测试回合中无人机出现碰撞，导致该测试回合失败。引入了循环神经网络的 DDQN-GRU 模型平均累积得分均在 500 以上，且均高于 DQN 模型和 DDQN 模型，此时无人机具有更好的避障性能。通过对比 1 500 次结果可以表明，虽然 DDQN 和 DDQN-GRU 均来自于训练过程中已经收敛的模型，但在测试过程中当无人机出现小的扰动时，对 DDQN 模型所给出避障策略的正确性会产生较大影响，此时 DDQN-GRU 模型表现出更强的鲁棒性，而 DQN 所训练得到模型几乎没有避障能力。

表 3 对比实验结果

算法类型	平均得分	平均奖励	平均路程
DQN	-55.46	-2.77	11.25
DDQN	-47.62	-1.69	22.55
DDQN-GRU	541.18	22.55	78.45

图 7 为 DDQN-GRU 训练所得结果在仿真环境中进行测试时无人机视角下整体的避障过程。从上到下依次为，无人机开始起飞，靠近第一层障碍物（图 7a）），模型根据输入的深度图片做出决策响应控制无人机躲避障碍物并到达第二层障碍物附近（图 7b）），躲避障碍物并达到指定位置（图 7c））。



a） 无人机起飞，靠近第一层障碍物

b） 穿越第一层障碍物，到达第二层障碍物附近

c） 穿越第二层障碍物，到达终点位置

图 7 无人机视角避障过程

为了直观展示 DDQN、DDQN-GRU 2 种算法所给出避障策略的好坏，实验设置了高空无人机视角拍摄完整的飞行过程，2 种算法的避障过程如图 8 所示。其中，红色轨迹为 DDQN-GRU 所给出的避障路线，黑色的轨迹为 DDQN 算法的避障路线。通过对比可以看出 DDQN-GRU 算法所给出的避障策略使得无人机所走里程数更短、更具有优势。

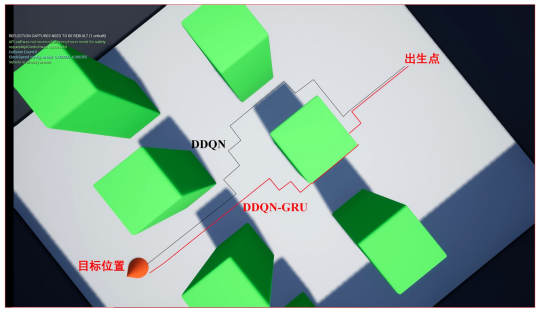
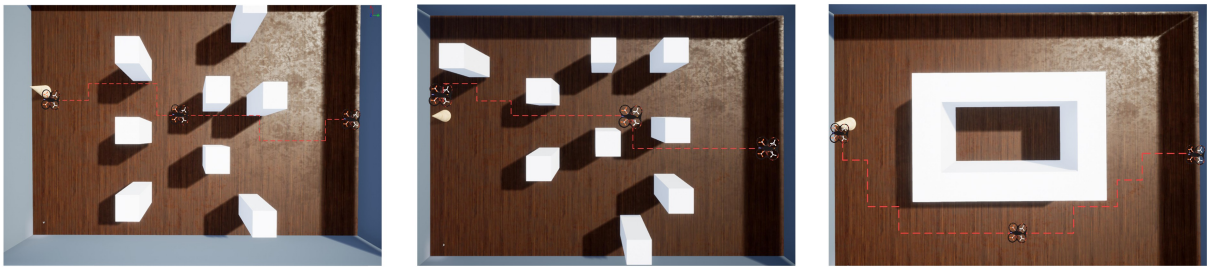


图 8 无人机避障路线俯视图

综上所述,环境的部分可观察性会阻碍 DDQN 算法在避障问题中的性能,GRU 循环网络可以保留随时间收集的相关信息为神经网络学习提供了额外的补充信息,因此大幅度增强了强化网络的记忆力。与传统的深度强化学习算法相比较,引入 GRU 网络的算法具有更好的性能。

3.4.3　未知场景下测试

为了验证本文所提出算法在未知环境中的避障能力,搭建了新的测试环境如图 9 所示。该测试环境分别在原有环境的基础上改变障碍物数量、颜色、大小以及分布情况,从而测试无人机是否真正学习到了躲避障碍物的能力。从图中可以看出,在改变障碍物大小以及分布时候,无人机仍具有躲避障碍物的能力。当测试环境无人机起始位置、目标位置与之前训练环境中一致时,由于障碍物分布不同,无人机可能需更多的步数才能到达目标位置,但仍旧可以成功躲避障碍并到达指定位置点。这表明使用 DDQN-GRU 算法所训练得到模型在未知环境中仍旧具有避障能力。



a)　未知场景 1　　b)　未知场景 2　　c)　未知场景 3

图 9　未知场景下避障能力测试

4　结　论

本文以无人机在未知三维环境中的实时避障为研究背景。针对传统强化学习避障算法无法有效处理时间序列信息以及存在部分可观测性的问题,介绍了一种新型基于深度强化学习的无人机避障算法,用于不同环境中的无人机避障以及路径规划。本文设计了一种将循环神经网络与 DDQN 相结合的算法框架,通过将循环神经网络引入到深度强化学习算法框架之中,增强真实神经网络的可分析性使其更好处理时间维度的信息。为了提高避障算法的可行性,根据避障环境重新设计无人机的状态、动作空间以及奖励函数使其更好适应避障任务要求。然后,将改进的深度强化学习框架应用于仿真环境中进行训练仿真测试。仿真及实验结果表明本文所提出的深度循环双 Q 网络能够很好地处理无人机避障中部分可观测问题,与现有的方法相比较,具有更好的收敛效果以及稳定性。

参考文献:

[1]　吕倩, 陶鹏, 吴宏, 等. 无人机在室内环境中自主飞行与避障[J]. 软件导刊, 2021, 20(2): 114-118
LYU Qian, TAO Peng, WU Hong, et al. Autonomous flight and obstacle avoidance of mav in indoor environment[J]. Software Guide, 2021, 20(2): 114-118 (in Chinese)

[2]　赖武刚, 刘宗汶, 姬凯飞, 等. 无人机的激光雷达自主巡逻飞行器系统设计[J]. 单片机与嵌入式系统应用, 2020, 20(5): 4
LAI Wugang, LIU Zongwen, JI Kaifei, et al. System design of autonomous patrol UAV based on lidar[J]. Microcontrollers and Embedded System Applications, 2020, 20(5): 4 (in Chinese)

[3]　李博昊, 罗咏涵, 彭克勤. 基于激光 SLAM 的室内移动机器人设计[J]. 数码世界, 2020, 12: 16-18
LI Bohao, LUO Yonghan, PENG Keqin. Design of indoor mobile robot based on laser SLAM[J]. Digital World, 2020, 12: 16-18 (in Chinese)

[4]　杨维, 朱文球, 张长隆. 基于 RGB-D 相机的无人机快速自主避障[J]. 湖南工业大学学报, 2015, 29(6): 6

YANG Wei, ZHU Wenqiu, ZHANG Changlong. UAV rapid and autonomous obstacle avoidance based on RGB-D camera[J]. Journal of Hunan University of Technology, 2015, 29(6): 6 (in Chinese)

[5] GIUSTI A, GUZZI J, DAN C C, et al. A machine learning approach to visual perception of forest trails for mobile robots[J]. IEEE Robotics & Automation Letters, 2017, 1(2): 661-667

[6] HADSELL R, ERKAN A, SERMANET P, et al. Deep belief net learning in a long-range vision system for autonomous off-road driving[C]//Proceedings of the 2008 IEEE Conference on Intelligent Robots and Systems, New York, 2008: 628-633

[7] 邢关生, 张凯文, 杜春燕. 基于深度强化学习的移动机器人避障算法[C]//第 30 届中国过程控制会议, 2019
XING Guansheng, ZHANG Kaiwen, DU Chunyan. Mobile robot obstacle avoidance algorithm based on deep reinforcement learning[C]//The 30th China Process Control Conference, 2019 (in Chinese)

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

[9] KULKARNI T D, NARASIMHAN K R, SAEEDI A, et al. Hierarchical deep reinforcement learning: integrating temporal abstraction and intrinsic[J/OL]. (2016-05-31)[2021-12-02]. https://arxiv.org/pdf/1604.06057.pdf

[10] SILVER D, HUANG 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

[11] SADEGHI F, LEVINE S. CAD2RL: Real single-image flight without asingle real image[J/OL]. (2016-06-08)[2021-12-02]. https://arxiv.org/pdf/1611.04201v4.pdf

[12] SINGLA A, PADAKANDLA S, BHATNAGAR S. Memory-based deep reinforcement learning for obstacle avoidance in UAV with limited environment knowledge[J]. IEEE Trans on Intelligent Transportation Systems, 2019, 99: 1-12

[13] HOWARD R A. Dynamic programming and Markov processes[J]. Mathematical Gazette, 1960, 3(358): 120

[14] BERTSEKAS D P, BERTSEKAS D P, BERTSEKAS D P, et al. Dynamic programming and optimal control[M]. Belmont, MA: Athena Scientific, 1995: 20-25

[15] SCHMIDHUBER J. Multi-column deep neural networks for image classification[C]//2012 IEEE Conference on Computer Vision and Pattern Recognition, New York, 2012: 3642-3649

[16] LONG D, ZHAN R, MAO Y. Recurrent neural networks with finite memory length[J]. IEEE Access, 2019, 7: 3642-3649

[17] CHUNG J, GULCEHRE C, CHO K H, et al. Empirical evaluation of gated recurrent neural networks on sequence modeling[J/OL]. (2014-12-11)[2021-12-02]. https://arxiv.org/pdf/1412.3555.pdf

[18] SRDJAN J, VICTOR G, ANDREW B. Airway anatomy of airsim high-fidelity simulator[J]. Anesthesiology, 2013, 118(1): 229-230

Study on UAV obstacle avoidance algorithm based on deep recurrent double Q network

WEI Yao[1], LIU Zhicheng[2], CAI Bin[3,4], CHEN Jiaxin[3,4], YANG Yao[5], ZHANG Kai[5]

(1.School of Astronautics, Northwestern Polytechnical University, Xi'an 710072, China;
2.The Third Military Representative Office of Beijing Military Representative Office of Air Force Equipment Department in Tianjin, Tianjin 300000, China;
3.Shanghai Aerospace Control Technology Institute, Shanghai 201109, China;
4.Infrared Detection Technology R & D Center of China Aerospace Science and Technology Corporation, Shanghai 201109, China;
5.Unmanned System Research Institute, Northwestern Polytechnical University, Xi'an 710072, China)

Abstract: The traditional reinforcement learning method has the problems of overestimation of value function and partial observability in the field of machine motion planning, especially in the obstacle avoidance problem of UAV, which lead to long training time and difficult convergence in the process of network training. This paper proposes an obstacle avoidance algorithm for UAVs based on a deep recurrent double Q network. By transforming the single-network structure into a dual-network structure, the optimal action selection and action value estimation are decoupled to reduce the overestimation of the value function. The fully connected layer introduces the GRU recurrent neural network module, and uses the GRU to process the time dimension information, enhance the analyzability of the real neural network, and improve the performance of the algorithm in some observable environments. On this basis, combining with the priority experience playback mechanism, the network convergence is accelerated. Finally, the original algorithm and the improved algorithm are tested in the simulation environment. The experimental results show that the algorithm has better performance in terms of training time, obstacle avoidance success rate and robustness.

Keywords: deep reinforcement learning; UAV; obstacle avoidance; recurrent neural network; DDQN

引用格式:*魏瑶, 刘志成, 蔡彬, 等. 基于深度循环双 Q 网络的无人机避障算法研究[J]. 西北工业大学学报,2022,40(5):970-979*

WEI Yao, LIU Zhicheng, CAI Bin, et al. Study on UAV obstacle avoidance algorithm based on deep recurrent double Q network[J]. Journal of Northwestern Polytechnical University, 2022, 40(5): 970-979 (in Chinese)