

基于改进 DQN 算法的复杂海战场路径规划方法

郁洲¹, 毕敬¹, 苑海涛²

(1. 北京工业大学信息学部, 北京 100124; 2. 北京航空航天大学自动化科学与电气工程学院, 北京 100191)

摘要: 为了有效地解决海战场环境下多舰艇的追踪目标问题, 以多智能体(舰艇)为研究对象, 提出一种基于改进的 Deep Q-Network (DQN) 算法的路径规划方法。DQN 算法结合多智能体的强化学习环境特性, 在传统 DQN 算法的基础上, 添加一个结构相同、参数不同的网络, 分别对 Q 实际值和估计值进行更新来实现价值函数的收敛。此外, 该方法使用经验回放和目标网络双参数更新机制, 有效地解决了神经网络训练误差大、泛化能力差和训练不稳定等问题。实验结果表明, 与传统的算法相比, 提出的方法能够更快地适应复杂多变的多类型海战场环境, 避障能力提高两倍多, 并且在环境中能够获得更高的训练奖励。

关键词: 深度 Q 网络; 强化学习; 多智能体; 路径规划; 目标追踪

中图分类号: TP399, TP391

文献标志码: A

doi: 10.11959/j.issn.2096-6652.202244

A path planning method for complex naval battle field based on an improved DQN algorithm

YU Zhou¹, BI Jing¹, YUAN Haitao²

1. Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China

2. School of Automation Science and Electrical Engineering, Beihang University, Beijing 100191, China

Abstract: To solve a target tracking problem of multiple warships in a sea battlefield environment, the multiple agents (warships) were focused on, and an improved deep Q-network (DQN) algorithm was proposed. It considers the characteristics of a multi-agent reinforcement learning environment based on a traditional DQN algorithm. It adds a network with the same structure and different parameters and updates a Q actual value and a Q estimated one, respectively to realize convergence of a value function. Besides, it adopts a mechanism of experience playback and an update of one of two parameters for a target network to effectively solve problems of high training errors of neural networks, poor generalization ability, and unstable training. Experimental results demonstrate that compared with the traditional DQN algorithm, the improved DQN one adapts to different complex and dynamical sea battlefield environments in a faster manner, and the ability to avoid obstacles has more than doubled and larger training reward.

Key words: deep Q network, reinforcement learning, multi-agent, path planning, target tracking

0 引言

随着科学技术的快速发展, 军事方面的信息化作战方式逐渐取代了过去的大规模热兵器作战方式。近年来, 海上作战环境显得尤为重要。海上作战通常需要舰艇之间相互协作与共享信息。在同

一环境下, 多智能体之间进行信息共享, 通过高效的路径规划实现对特定目标的追踪。路径规划的计算过程主要考虑到智能体之间的协调与合作, 重点是智能体的自主学习和自主决策, 以及智能体的交互和合作能力。常见的单智能体路径规划方法包括人工势场法^[1]、基于路径编码的遗传

收稿日期: 2021-11-05; 修回日期: 2022-08-15

通信作者: 毕敬, bijing@bjut.edu.cn

基金项目: 国家自然科学基金资助项目 (No.62073005, No.62173013)

Foundation Items: The National Natural Science Foundation of China (No.62073005, No.62173013)

算法^[2]、随机搜索蚁群算法^[3]、基于学习训练的神经网络和强化学习方法^[4]、Dijkstra 算法^[5]、粒子群算法^[6]、A*算法^[7], 以及混合模型方法。本文使用面向多智能体的强化学习方法, 该方法的思想是使状态、动作和策略分配相结合, 实现多智能体学习训练的延续和扩展。在固定环境中, 当出现较大的状态空间导致的维数灾难问题时, 胡俊等人^[8]提出了滚动的 Q 学习方法, 在一定程度上改进了算法。Xin J 等人^[9]第一次提出了利用深度强化学习 (deep reinforcement learning, DRL) 的方法来解决智能体的路径规划问题。Tai L 等人^[10]提出了基于 Deep Q-Network (DQN) 算法的智能体路径规划模型, 该模型在仿真环境中取得了良好的效果。

本文在上述文献思想的基础上, 在深度强化学习模型中, 利用 DQN 算法的强大拟合能力, 解决了传统启发式路径规划算法在复杂环境下算力不足的问题, 在多智能体协同工作的情况下取得了良好的表现。本文的创新点如下。

- 本文针对海战场攻防对抗环境的诸多特点, 在传统 DQN 算法的基础上进行改进, 提出了改进的 DQN (improved deep Q-network, IDQN) 算法。在保持原有 DQN 结构的基础上, 引入网络结构相同的 DDQN (double deep Q network) 算法, 分别用作对 Q 实际值和估计值的更新。通过实验验证, 经过 IDQN 算法训练后的网络泛化能力显著优于传统的 DQN。
- 本文将提出的 IDQN 算法应用于海战场环境下的多舰艇目标追踪, 相较于传统的单智能体目标追踪场景, 本文提出了双智能体应用场景, 在模型训练过程中两个智能体舰艇共享信息, 从而实现目标追踪。

1 相关工作

1.1 深度强化学习

DRL 结合了深度学习和强化学习算法, 同时融合了深度学习算法对感知问题的强理解能力和对强化学习算法学习结果的拟合能力。Q 学习算法通过构造 Q 值表来选择最优策略。然而, 大多数问题发生在状态空间过大、维度变大的实际环境中, 因此无法使用该 Q 值表来记录和索引, 这导致了维度灾难问题。即由于 Q 学习算法的感知能力弱, 每个代理无法处理 Q 学习输入, 因为高维状态下控制值函数的输出很难推广到较大的状态空间^[11]。因

此, 在支持大规模状态的空间学习中, 往往需要更加复杂的特征提取和强化学习能力^[11]。使用 DRL 算法和 Q 网络来拟合结果并获得输出或决策, 能够允许智能体在更复杂的环境中感知和建立行动策略, 从而提高算法的收敛、泛化能力和学习速度, 并使智能体能够在未知和复杂的环境中执行良好的路径规划。因此, 本文在 DRL 算法中使用了 DQN 算法, 并结合了 Q 学习算法、经验回放机制等方法, 生成目标 Q 值。与此同时, 张荣霞等人^[12]阐述了 DRL 在目标导航中的重要作用, 目标导航需要找到一条端到端的最优路径, 其中包括避障、收集目标等任务。刘朝阳等人^[13]将物体巡航控制与深度强化学习结合起来, 提出基于确定性策略梯度算法的巡航控制算法, 使模型具有了更好的稳定性和安全性。

1.2 多智能体路径规划

多智能体路径规划相较于单智能体路径规划更加复杂。在多智能体路径规划中以及考虑智能体搜寻目标的过程中, 需要考虑多智能体在训练过程中的交互。韩向敏等人^[14]提出了一种多智能体群集与避障算法, 该算法主要基于分布式 Flocking 的群集控制模式, 在多智能体的训练过程中, 保障各个智能体之间的一致性。

为了保证多个智能体在复杂的环境中能够有效地躲避障碍物, 并成功到达目标区域, 刘婵等人^[15]提出了一种基于多智能体的二层路径规划模型研究, 他们在第一层模型中利用模型策略生成了一条大致的路径, 然后在第二层模型中, 根据第一层模型产生的大致路径做进一步的精细导航, 将环境中的动态因素与静态因素相结合, 在第一层模型中采用静态规划, 在第二层中根据仿真环境对智能体进行动态路径规划。因此, 该模型在复杂的仿真环境中具有较高的执行效率。刘辉等人^[16]在货物运输、快递分拣等领域中使用了强化学习技术, 他们提出了一种基于多智能体独立强化学习的最大回报频率算法, 该算法可以同时任务调度和路径规划进行优化, 有效解决了多个智能体的交互动作引起的维数灾难问题。

2 基于改进 DQN 的路径规划方法

为了解决将 DQN 算法应用于多智能体路径规划的过程中出现的过估计和训练步数大等问题, 本文引入经验回放池来对 Q 值表进行更新。初始化的 Q 值表采用经验回放技术^[17], 在智能体训练的前 200 步内, 将收集到的观察序列 (s, a, r, s') 存放到经验回放的

缓冲区中,通过收集到的经验进行Q值表的初始化。首先,智能体会在每个回合的前期收集数据并放入记忆库中;然后,智能体从记忆库中选取数据放入网络中进行训练;进一步,根据贪婪策略选择最优动作,到达下一个观察,并将该数据根据重采样优选机制存储到缓存记忆库中,一直循环迭代直到训练完成。

2.1 DQN 算法改进思路

DQN 算法是深度强化学习的一种方法,其基本原理是可以结合深度学习和强化学习算法^[18]。Q 学习算法构造了一个可用于深度学习的目标函数。卷积神经网络生成目标 Q 值,基于该状态下的目标 Q 值评估下一个状态的 Q 值,并加入经验回放机制和目标网络来打破数据之间的关联。DQN 算法可以使用价值函数来近似 Q 值。如图 1 所示, DQN 中主要包含两种参数不同但结构相同的网络,两种网络分别对 Q 估计值和 Q 现实值进行更新。

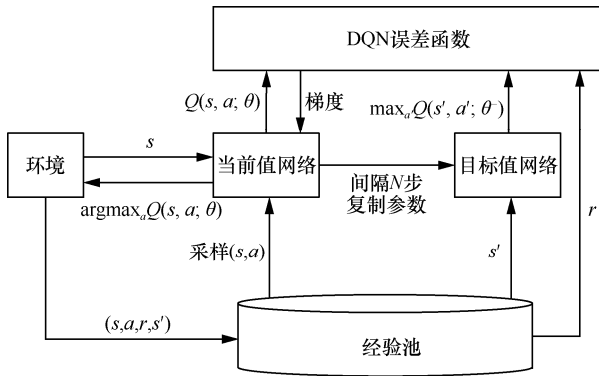


图1 DQN 算法更新过程

笔者在实验过程中发现, DQN 算法在实验过程中会出现明显的过估计问题,即 Q 估计值大于 Q 现实值。DQN 的目标价值采用式(1)来更新参数:

$$\begin{aligned} Q_{\text{target}}(S_t, A_t) = \\ R_t + \gamma \max_{a'} Q(S'_t, A'_t; \theta^-) Q_{\text{target}}(S_t, A_t) = \\ R_t + \gamma \max_{a'} Q(S'_t, A'_t; \theta^-) \end{aligned} \quad (1)$$

其中, θ^- 表示上一个更新周期价值网络的参数, S_t 、 A_t 和 R_t 分别表示当前阶段的状态值、动作值和奖励值, γ 表示折扣因子, S'_t 和 A'_t 表示下一阶段的状态值和动作值, Q_{target} 表示 DQN 网络中的目标 Q 值。本文基于 DQN 算法更新过程,引入 DDQN,该算法在 DQN 结构的基础上,采用式(2)来更新目标价值, θ 表示当前 Q 值网络的参数。DDQN 使用估计网络进行动作选择,使用目标网络进行策略评估,将动作选择和策略评估进行了分离,有效地

解决了 DQN 中存在的过估计问题。

$$\begin{aligned} Q_{\text{target}}(S_t, A_t) = \\ R_t + \gamma Q\left(S', \max_{a'} Q(S'_t, a'; \theta); \theta^-\right) Q_{\text{target}}(S_t, A_t) = \\ R_t + \gamma Q\left(S', \max_{a'} Q(S'_t, a'; \theta); \theta^-\right) \end{aligned} \quad (2)$$

本文针对海战场攻防对抗环境的诸多特点,在传统 DQN 算法的基础上进行改进,提出了 IDQN 算法。在保持原有 DQN 的网络结构的基础上,引入网络结构相同的 DDQN 算法,分别用来对 Q 实际值和 Q 估计值进行更新。本文提出的 IDQN 算法综合了 DQN 算法的强大拟合能力和 DDQN 算法的防止过拟合能力。在模型训练中使用 DQN 和 DDQN 算法对智能体进行综合训练。通过实验验证,经过 IDQN 算法训练后的网络泛化能力显著强于传统的 DQN。算法 1 给出了 IDQN 算法的具体更新过程,具体步骤包括:①初始化容量、权重和重放记忆量;②进行 episode 次循环,以概率 ε 随机选择动作,在模拟器中观察该动作下的奖励,对结果进行存储,并且根据梯度下降方法计算误差,最终训练出最优的模型;③每 C 次更新一下 Q, 即每隔 C 步更新 Q 为最新的参数。

算法 1: IDQN

- 1 将重放记忆 D 的容量大小初始化为 N
- 2 用随机权重 θ 初始化动作值函数 Q
- 3 用权重 $\theta' = \theta$ 初始化目标价值函数 Q'
- 4 for episode=1, M do:
 - 初始化序列 $S_1 = \{x_1\}$ 和预处理序列 $\phi_1 = \phi_{s_1}$
- 5 for t=1, T do:
 - 6 以概率 ε 选择一个随机动作
 - 7 否则选择 $a_t = \arg\max_a Q(\phi(s_t), a, \theta)$
 - 8 在模拟器中执行动作 a_t , 并观察奖励 r_t 和图像 x_{t+1}
 - 9 设置 $S_{t+1} = S_t, a_t, x_{t+1}$ 和预处理 $\phi_{t+1} = \phi(s_{t+1})$
 - 10 在 D 中存储转换 $(\phi_t, a_t, r_t, \phi_{t+1})$
 - 11 从 D 中的随机小批量转换 $(\phi_t, a_t, r_t, \phi_{t+1})$
 - $y_j = r_j$ if s is end
 - $y_j = r_j + \gamma \max_{a'} (Q(\phi(S'_j), a'; \theta))$
 - $y_j = r_j + \gamma Q\left(S'_j, \max_{a'} Q(S'_j, a'; \theta); \theta^-\right)$
 - 12 设置对网络参数 θ 执行梯度下降步骤 $(y_j - Q(\phi_j, a_j, \theta))^2$
 - 13 每 C 步重置 Q' 值。

14 end for
15 end for

2.2 奖励函数

在强化学习中，奖励函数是影响智能体路径规划的重要因素，设置合理的奖励函数可以有效地提高智能体路径寻找的正确性和有效性。奖励函数 R 在状态 S 下的 t 时刻到 $t+1$ 时刻的状态所能获得的期望奖励是 $R_s = E[R_{t+1} | S_t = s]$ 。累计奖励函数值 G_t ，也就是回报，如式（3）所示：

$$G_t = R_{t+1} + \gamma R_{t+2} + \cdots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \tag{3}$$

其中， γ 是折扣系数，体现在未来时刻回报奖励的衰减效果。智能体的目标是找到最终目标的位置，以此来获得最大的回报值。因此，奖励值的设定对路径规划的结果有不可忽视的影响。在本文工作中，设定智能体碰到的奖励值为-1，到达目标点奖励值为+1，其他动作的奖励值为 0。具体奖励表见表 1。

奖励值	情况
+1	到达目标点
-1	碰到障碍物
0	其他动作

2.3 动作选择

动作选择是强化学习三要素之一^[19]。智能体通过采用一定的策略，进行下一步动作的选取。通过不同的动作来获得奖励值，奖励值回报越高的动作具有更高的价值，高回报值使智能体倾向于在下一次训练中再次选择该动作。在本文实验中，智能体能够采用 4 个方向的动作，数值分别为 0、1、2、3，对应表示向上、向下、向右、向左 4 个动作。在强化学习中，选择动作的策略主要有 ϵ -greedy 策略、高斯策略和玻耳兹曼探索策略。本文选取的动作选择策略是 ϵ -greedy 策略， ϵ -greedy 策略是一种全局的动作优化策略^[20]。

传统的贪婪算法在求解模型时，在一定程度上求解出的是局部最优解。贪婪算法的思想是在每一步做出动作选择时，选择能够带来最大奖励值的动作，在模型中运用贪婪算法会出现陷入局部最优的问题。因此本文在智能体选择动作时采用改进的 ϵ -greedy 策

略，以 10%的概率进行动作的探索。通过上述方法，智能体有一定的概率选择其他动作，可以跳出局部最优的情况，使实验结果更加符合实际情况。

3 实验仿真和分析

3.1 实验平台描述

本文以多智能体（舰艇）为研究对象，设计了两个不同的仿真环境来模拟海战场的真实场景，如图 2 所示。在真实的海战场环境中，海面上的风是作用在舰船上的主要外力。海面上的舰船受到海风等外力的作用会发生速度和航行方向的变化，形成相应的风压差，从而使舰船的控制更加复杂。

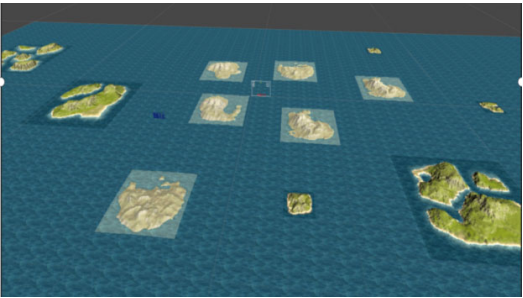


图2 实际场景

本文工作中为了增强实验的鲁棒性，简化了实验环境。具体设计了环境 1 和环境 2 两个无风的海战场环境，如图 3 和图 4 所示，其中红色和蓝色方格代表两个智能体（舰艇）的初始化位置，黑色方格代表障碍物，黄色方格代表目标潜艇的位置。环境 1 中包含 8 个障碍物，目标潜艇位于障碍相对较少的位置。环境 2 中设置了 6 个障碍物，目标潜艇位于障碍相对较多的位置。本文实验使用的仿真环境软硬件配置见表 2。

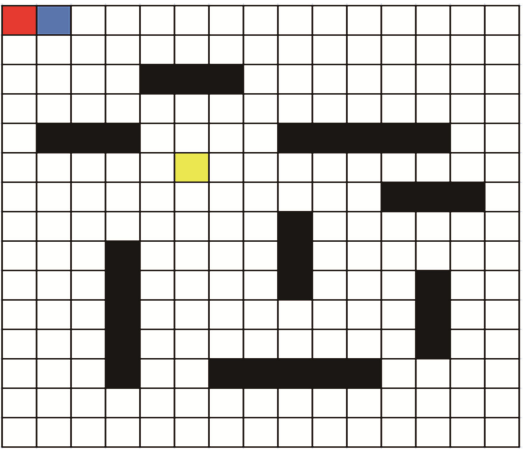


图3 环境1

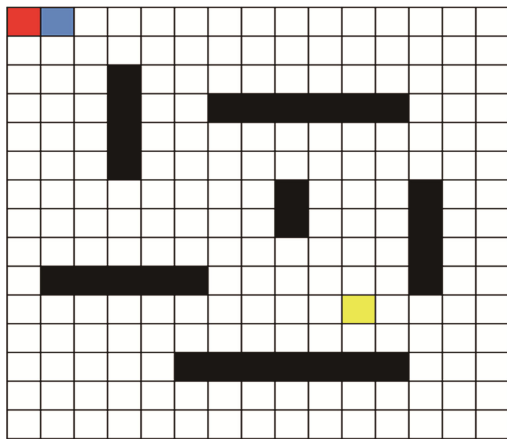


图4 环境2

表2 实验使用的仿真环境软硬件配置

名称	版本型号
CPU	CORE i5 7th Gen
内存	8 GB
GPU	NVIDIA GETFORCE 930MX
系统	Windows 10
Tensorflow-GPU	1.4

3.2 实验结果

为了验证本文方法的有效性, 本文进行了两组对比试验。在模拟平台上分别对环境1和环境2的仿真环境进行实验, 在两种环境下对比本文提出的IDQN算法与DQN、DDQN两种算法的性能。3种算法均采用相同的网络模型和网络参数。每次实验均进行100次训练, 在每次训练的过程中, 采取200步更新目标网络的参数。

在每次实验训练的初期, 智能体(舰艇)会以一个杂乱的轨迹前进, 在经过一定次数的学习和累计记忆后, 智能体(舰艇)能够安全地躲避障碍物, 并且寻找到目标潜艇的位置。当一个智能体(舰艇)在不触碰障碍物的情况下成功寻找到目标潜艇时, 那么在后面的训练中, 智能体(舰艇)会按照当前成功寻找到目标潜艇的最短路径寻找目标。由于本文采用的是 ϵ -greedy策略, 在每次训练中, 智能体(舰艇)会按照提前设定的 ϵ 概率值探索其他的路径, 此时的路径并不一定是最优路径。采用这种算法, 智能体(舰艇)不断地扩充自己的记忆库, 以此来寻找全局最优的最短路径。如图5和图6所示, 当智能体(舰艇)训练稳定后, 智能体(舰艇)即找到目标潜艇的最短路径。其中红色线路代表红色智能体(舰艇)

的最优路径, 蓝色线路代表蓝色智能体(舰艇)的最优路径。

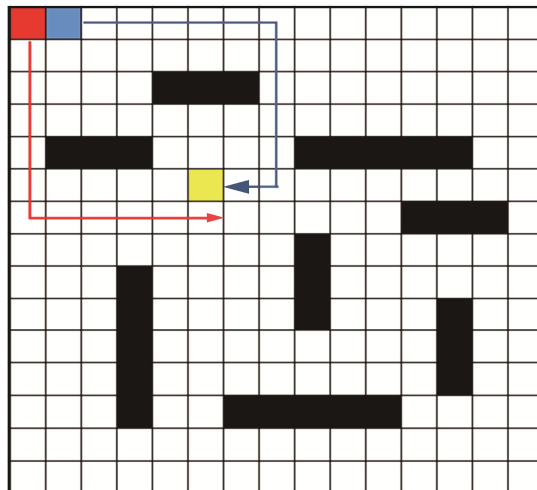


图5 环境1中的智能体轨迹

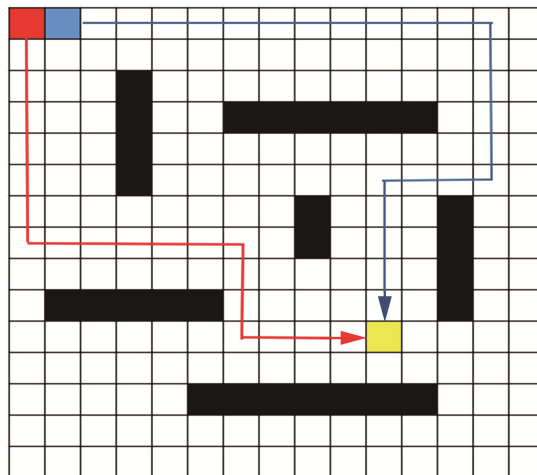


图6 环境2中的智能体轨迹

图7和图8展示了IDQN算法在环境1和环境2中, 智能体(舰艇)训练过程中的损失函数曲线。可以看到IDQN算法在两种环境中都处于稳定的状态, 图7中横坐标代表训练步数, 纵坐标是算法在训练过程中的代价(指平均每步的损失值)。代价函数在大部分情况下处于稳定状态, 但可以看出在极少数训练中会出现损失增大的情况。这是因为 ϵ -greedy策略会使智能体(舰艇)在每次选择下一个动作时, 以 ϵ 的概率探索新的动作。在 ϵ -greedy策略的作用下, 智能体(舰艇)可以在一定情况下摆脱局部最优的情况, 使智能体(舰艇)找到全局最优的路径。从图8可以看出, IDQN算法在损失值稳定的情况下要优于DQN算法。

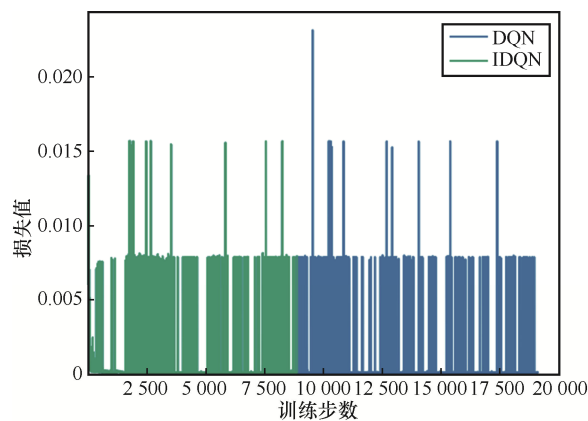


图7 代价（损失值）变化曲线（环境1）

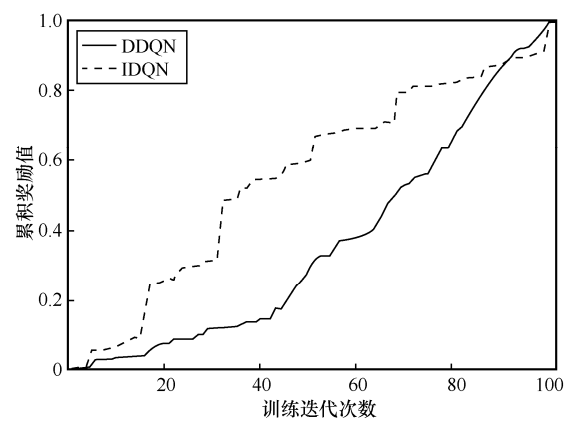


图9 环境1中的累计奖励值

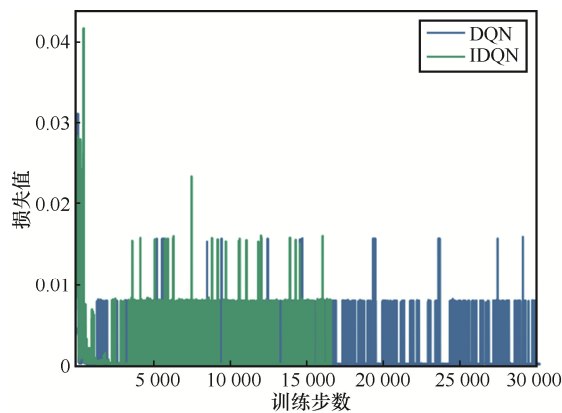


图8 代价（损失值）变化曲线（环境2）

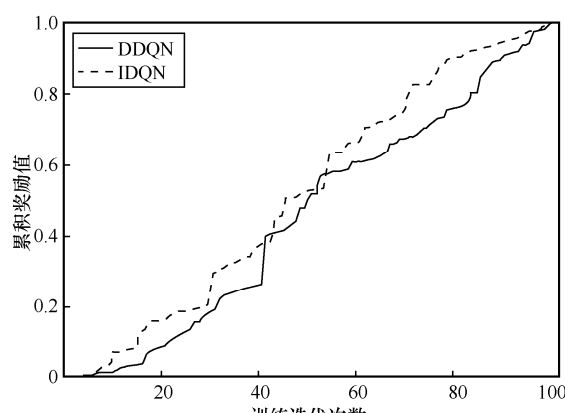


图10 环境2中的累计奖励值

本文从多个角度采用多个指标将本文提出的 IDQN 算法与 DQN 算法和 DDQN 算法进行对比。其中，第一个指标是算法在不同环境训练过程中累积的奖励值，即智能体（舰艇）采取每一步动作的优劣，累计奖励值越高，说明智能体（舰艇）采取的动作越优；第二个指标是智能体（舰艇）采取不同的算法在训练中稳定找到目标潜艇所需要的步数，该指标体现了在确定的环境中算法在实际场景的可用性。本文给出了3种算法在两种仿真环境中最终搜寻到目标所耗费的步数，见表3。从表3可知，在环境1中，DQN 算法的具体步数为101 170，DDQN 算法的具体步数为13 135，IDQN 算法的具体步数为12 335。可以看出，IDQN 算法具有更高的路径规划和避障能力。图9和图10为DDQN 算法和IDQN 算法的累计奖励值曲线，可以看出IDQN 算法达到了比较好的效果。

表3		步数表	
算法	环境1	环境2	
DQN	101 170	151 150	
DDQN	13 135	74 539	
IDQN	12 335	62 709	

为了更好地证明IDQN算法在复杂环境和多智能体环境下的路径规划和避障能力，本文进一步设计了一组单独的环境，并且在模拟环境中增加了5个智能体。相较于上述实验环境，新环境增加了障碍物的个数，并更改了目标物的位置，目标物处于障碍物旁边，使环境更加复杂，如图11所示。损失值随训练步数的变化如图12所示，从图12可以看出，多个智能体在相对复杂的环境中寻找较为隐蔽的目标时，IDQN 算法在训练过程中可以达到稳定的状态。

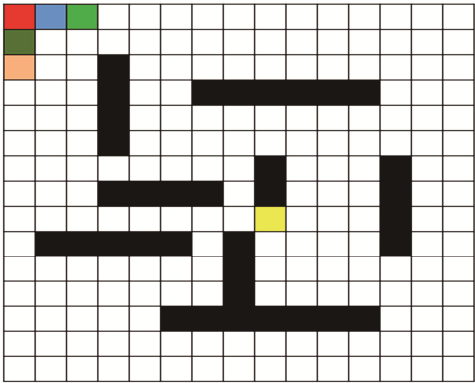


图11 包含5个智能体的新环境

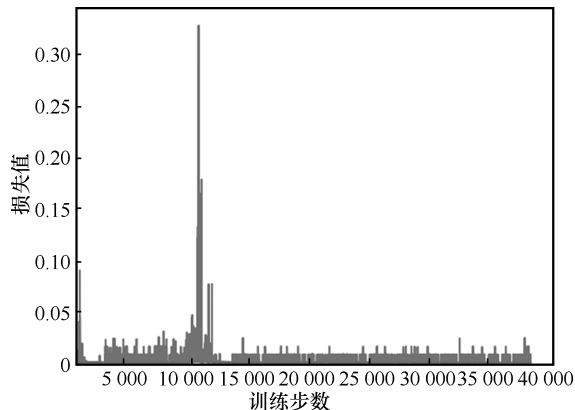


图12 损失值随训练步数的变化

4 结束语

本文提出的 IDQN 算法, 在深度 Q 网络算法的基础上动态融合了 DQN 算法的改进算法, 并且引入了 DDQN 算法。使用 IDQN 算法得到了面向多智能体的高效路径规划、避障的策略方法。本文采用两种不同的仿真环境进行实验验证, 证明本文提出的 IDQN 算法在累计奖励值和搜寻目标所需要的步数方面都有显著的进步, 特别在寻找目标所需的步数方面, 与 DQN 算法相比, IDQN 算法在环境 1 和环境 2 中分别降低了 87.8% 和 58.5%。实验结果表明, 本文提出的 IDQN 算法在海战场对抗环境下具有更好的路径规划和避障能力。在未来工作中需要进一步研究海风、海浪和雷达信号等对舰艇智能体的影响, 需要将离散动作重新建模成连续动作, 并且选择连续的奖励函数。未来会进一步改进本文的模型, 进一步提高动态复杂多变的海战场环境中的路径规划和避障能力。

参考文献:

- [1] 史进, 董瑶, 白振东, 等. 移动机器人动态路径规划方法的研究与实现[J]. 计算机应用, 2017, 37(11): 3119-3123.
SHI J, DONG Y, BAI Z D, et al. Research and implementation of mobile robot path planning method[J]. Journal of Computer Applications, 2017, 37(11): 3119-3123.
- [2] 魏彤, 龙琛. 基于改进遗传算法的移动机器人路径规划[J]. 北京航空航天大学学报, 2020, 46(4): 703-711.
WEI T, LONG C. Path planning for mobile robot based on improved genetic algorithm[J]. Journal of Beijing University of Aeronautics and Astronautics, 2020, 46, (4): 703-711.
- [3] YOU X M, LIU S, LYU J Q. Ant colony algorithm based on dynamic search strategy and its application on path planning of robot[J]. Control and Decision, 2017, 32(3): 552-556.
- [4] 默凡凡. 基于 Q 学习算法的移动机器人路径规划方法研究[D]. 北京: 北京工业大学, 2016.
MO F F. Research on path planning of mobile robot based on Q learning algorithm[D]. Beijing: Beijing University of Technology, 2016.
- [5] 姜涛, 王建中, 施家栋. 小型移动机器人自主返航路径规划方法[J]. 计算机工程, 2015, 41(1): 164-168.
JIANG T, WANG J Z, SHI J D. Path planning method in autonomous returning for mini-mobile robot[J]. Computer Engineering, 2015, 41(1): 164-168.
- [6] 刘洁, 赵海芳, 周德廉. 一种改进量子行为粒子群优化算法的移动机器人路径规划[J]. 计算机科学, 2017, 44(S2): 123-128.
LIU J, ZHAO H F, ZHOU D L. Improved quantum behaved particle swarm optimization algorithm for mobile robot path planning[J]. Computer Science, 2017, 44(S2): 123-128.
- [7] 孙炜, 吕云峰, 唐宏伟, 等. 基于一种改进 A* 算法的移动机器人路径规划[J]. 湖南大学学报(自然科学版), 2017, 44(4): 94-101.
SUN W, LYU Y F, TANG H W, et al. Mobile robot path planning based on an improved A* algorithm[J]. Journal of Hunan University (Natural Sciences), 2017, 44(4): 94-101.
- [8] 胡俊, 朱庆保. 未知环境下基于有先验知识的滚动 Q 学习机器人路径规划[J]. 控制与决策, 2010, 25(9): 1364-1368.
HU J, ZHU Q B. Path planning of robot for unknown environment based on prior knowledge rolling Q-learning[J]. Control and Decision, 2010, 25(9): 1364-1368.
- [9] XIN J, ZHAO H, LIU D, et al. Application of deep reinforcement learning in mobile robot path planning[C]//Proceedings of 2017 Chinese Automation Congress. Piscataway: IEEE Press, 2017: 7112-7116.
- [10] TAI L, LIU M. Towards cognitive exploration through deep reinforcement learning for mobile robots[J]. arXiv preprint, 2016, arXiv: 1610.01733.
- [11] 江其洲, 曾碧. 基于深度强化学习的移动机器人导航策略研究[J]. 计算机测量与控制, 2019, 27(8): 217-221, 226.
JIANG Q Z, ZENG B. Research on navigation strategy of mobile robot based on deep reinforcement learning[J]. Computer Measurement & Control, 2019, 27(8): 217-221, 226.
- [12] 张荣霞, 武长旭, 孙同超, 等. 深度强化学习及在路径规划中的研究进展[J]. 计算机工程与应用, 2021, 57(19): 44-56.
ZHANG R X, WU C X, SUN T C, et al. Progress on deep reinforcement learning in path planning[J]. Computer Engineering and Applications, 2021, 57(19): 44-56.
- [13] 刘朝阳, 穆朝絮, 孙长银. 深度强化学习算法与应用研究现状综述[J]. 智能科学与技术学报, 2020, 2(4): 314-326.
LIU Z Y, MU C X, SUN C Y. An overview on algorithms and applications of deep reinforcement learning[J]. Chinese Journal of Intelligent Science and Technology, 2020, 2(4): 314-326.
- [14] 韩向敏, 鲍泓, 梁军, 等. 一种基于深度强化学习的自适应巡航控制算法[J]. 计算机工程, 2018, 44(7): 32-35, 41.
HAN X M, BAO H, LIANG J, et al. An adaptive cruise control algorithm based on deep reinforcement learning[J]. Computer Engineering, 2018, 44(7): 32-35, 41.
- [15] 刘婵, 朱永川, 白园, 等. 基于 flocking 的多智能体群集与避障算法研究与仿真[J]. 通信技术, 2019, 52(7): 1632-1638.
LIU C, ZHU Y C, BAI Y, et al. Flocking for multi-agent clustering and

obstacle avoidance: algorithm and simulation[J]. Communications Technology, 2019, 52(7): 1632-1638.

- [16] 刘辉, 肖克, 王京攀. 基于多智能体强化学习的多 AGV 路径规划方法[J]. 自动化与仪表, 2020, 35(2): 84-89.

LIU H, XIAO K, WANG J B. Multi-AGV path planning method based on multi-agent reinforcement learning[J]. Automation & Instrumentation, 2020, 35(2): 84-89.

- [17] 黄子蓉, 甯彦淞, 王莉. 基于优先经验回放的多智能体协同算法[J]. 太原理工大学学报, 2021, 52(5): 747-753.

HUANG Z R, NING Y S, WANG L. Prioritized experience replay for multi-agent cooperation[J]. Journal of Taiyuan University of Technology, 2021, 52(5): 747-753.

- [18] 高振洋, 秦斌. 深度强化学习研究进展[J]. 电脑知识与技术, 2019, 15(4): 157-159, 173.

GAO Z Y, QIN B. Progress in deep reinforcement learning research[J]. Computer Knowledge and Technology, 2019, 15(4): 157-159, 173.

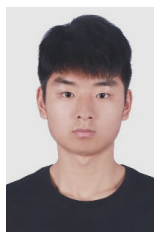
- [19] 杨思明, 单征, 丁煜, 等. 深度强化学习研究综述[J]. 计算机工程, 2021, 47(12): 19-29.

YANG S M, SHAN Z, DING Y, et al. Survey of research on deep reinforcement learning[J]. Computer Engineering, 2021, 47(12): 19-29.

- [20] 刘全, 翟建伟, 章宗长, 等. 深度强化学习综述[J]. 计算机学报, 2018, 41(1): 1-27.

LIU Q, ZHAI J W, ZHANG Z C, et al. A survey on deep reinforcement learning[J]. Chinese Journal of Computers, 2018, 41(1): 1-27.

[作者简介]



郁洲(1998—), 男, 北京工业大学信息学部硕士生, 主要研究方向为深度强化学习、云数据中心任务调度和资源分配等。



毕敬(1979—), 女, 博士, 北京工业大学信息学部副教授、博士生导师, 主要研究方向为大数据分析、计算智能与调控优化、深度学习、数据驱动建模、云边协同与节能等。



苑海涛(1986—), 男, 博士, 北京航空航天大学自动化科学与电气工程学院副教授, 主要研究方向为云计算、边缘计算、数据中心、深度学习、智能优化、时间序列特征建模与预测等。