一类基于概率优先经验回放机制的分布式多智能体软行动–评论者算法

张严心[1], 孔 涵[1], 殷辰堃[1], 王子豪[1], 黄志清[2]
(1. 北京交通大学电子信息工程学院, 北京 100044; 2. 北京工业大学信息学部, 北京 100124)

摘 要: 针对实际多智能体系统对交互经验的庞大需求,在单智能体领域分布式架构的基础上,提出概率经验优先回放机制与分布式架构并行的多智能体软行动–评论者算法(multi-agent soft Actor-Critic with probabilistic prioritized experience replay based on a distributed paradigm, DPER-MASAC). 该算法中的行动者以并行与环境交互的方式收集经验数据,为突破单纯最近经验在多智能体高吞吐量情况下被高概率抽取的局限性,提出更为普适的改进的基于优先级的概率方式对经验数据进行抽样利用的模式,并对智能体的网络参数进行更新. 为验证算法的效率,设计了难度递增的2类合作和竞争关系共存的捕食者–猎物任务场景,将DPER-MASAC与多智能体软行动–评论者算法(multi-agent soft Actor-Critic, MASAC)和带有优先经验回放机制的多智能体软行动–评论者算法(multi-agent soft Actor-Critic with prioritized experience replay, PER-MASAC)2种基线算法进行对比实验. 结果表明,采用DPER-MASAC训练的捕食者团队其决策水平在最终性能和任务成功率2个维度上均有明显提升.

关键词: 多智能体系统(MAS); 多智能体深度强化学习(DRL); 优先经验回放机制; 分布式结构; 抽样概率; 软行动–评论者算法
中图分类号: TP 83; TP 311 **文献标志码**: A **文章编号**: 0254 – 0037(2023)04 – 0459 – 08
doi: 10.11936/bjutxb2022110019

Distributed Multi-agent Soft Actor-Critic Algorithm With Probabilistic Prioritized Experience Replay

ZHANG Yanxin[1], KONG Han[1], YIN Chenkun[1], WANG Zihao[1], HUANG Zhiqing[2]
(1. School of Electronic and Information Engineering, Beijing Jiaotong University, Beijing 100044, China;
2. Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China)

Abstract: Aiming at a huge demand for interaction data in practical multi-agent tasks, based on the distributed architecture in the single-intelligent domain, a multi-agent soft Actor-Critic reinforcement learning algorithm together with probabilistic prioritized experience replay and distributed architecture (DPER-MASAC) was proposed. In DPER-MASAC, workers collect experience data by interacting with environments simultaneously. To break through the limitation of purely recent experience being extracted with high probability in the case of multi-agent system of high throughput, a more universal and improved mode based on probability of priority was put forward to sample and utilize experience data, and the network parameters of agents will be updated. To verify the efficiency of DPER-MASAC, comparative

收稿日期: 2022-11-10; 修回日期: 2022-11-22
基金项目: 国家自然科学基金资助项目(62273082); 中央高校基本科研业务费重大项目(2018JBZ006)
作者简介: 张严心(1976—), 女, 副教授, 主要从事复杂大系统的智能控制、无人驾驶中的智能控制、复杂交通网络控制方面的研究, E-mail: yxzhang@bjtu.edu.cn
通信作者: 黄志清(1970—), 男, 副教授, 主要从事无人驾驶智能决策控制、车联网及区块链方面的研究, E-mail: zqhuang@bjut.edu.cn

experiments were conducted in two types of predator-prey environment in which both cooperation and competition exist among multiple agents. Meanwhile multi-agent soft Actor-Critic (MASAC) and multi-agent soft Actor-Critic with prioritized experience replay (PER-MASAC) were regarded as two baseline algorithms, compared with DPER-MASAC in this environment with gradually incremental-difficulty. In terms of the final performance and success rate, results indicate that the policy of predators, which is trained by DPER-MASAC, performs optimally.

Key words: multi-agent systems (MAS); multi-agent deep reinforcement learning (DRL); prioritized experience replay; distributed architecture; sampling-probability; soft Actor-Critic algorithm

深度强化学习(deep reinforcement learning, DRL)至今已在自动驾驶[1]、自动靠泊[2]、机器人控制[3]等诸多领域得到广泛的应用,极大推动了人工智能和自动化技术的发展. 面对越来越多的大规模复杂问题, 单智能体集成的解决方案将面临各种资源和条件的约束. 多智能体系统(multi-agent system, MAS)[4]因具有自主性、分布性和协调性等特点成为实现分布式人工智能的重要解决方案.

DRL 已在单智能体领域取得显著成果,这促使研究人员将 DRL 的思想和算法应用于 MAS 的学习和控制中,由此催生了多智能体深度强化学习(muti-agent deep reinforcement learning, MADRL)[5-7],以解决多个智能体在复杂任务场景中的智能序贯决策问题. 经过数年的发展创新, MADRL 已广泛应用于游戏人工智能[8]、交通信号灯控制[9-11]、多机协同空战[12]、多机器人群控制[13]、滴滴智能派单[14]和工厂智能调度[15]等各类现实领域. MADRL 常基于集中式训练分布式执行(centralized training decentralized execution, CTDE)的范式[16]对 DRL 算法的训练和执行阶段做出相应调整. 多智能体深度确定性策略梯度算法(muti-agent deep deterministic policy gradient, MADDPG)[17]是目前 MADRL 中基于 CTDE 范式最经典的算法之一.

然而,MADDPG 在每个状态下只考虑一个最优动作,易收敛于次优策略且勘探效率较差,难以解决复杂环境下的多智能体决策问题. 文献[18]提出基于软行动-评论者的多智能体深度强化学习算法(multi-agent soft Actor-Critic, MASAC),策略网络的输出是由高斯分布表示的随机策略. 相比于深度确定性策略梯度算法(deep deterministic policy gradient, DDPG)[19]的确定性策略,随机策略在大规模复杂场景和部分可观测任务中具有更强的探索能力和决策性能. 实验结果表明,MASAC 的性能优于 MADDPG.

尽管 MASAC 策略的随机性在一定程度上增加了智能体探索最优动作的可能性,但增加策略的随机性意味着算法需要更长的训练时间来产生更多的交互数据. 考虑到复杂多智能体环境中智能体数量的增加、动作空间和状态空间均呈指数级增长使得算法对于交互经验数据量的需求远高于单智能体强化学习,训练效率仍是 MADRL 解决实际问题时一个棘手且永恒的议题.

本文是基于 MASAC 进一步探寻具有更高决策效率的 MADRL. 在单智能体领域中,DeepMind 团队提出一种分布式优先经验回放池算法(distributed prioritized experience replay, Ape-X)[20]为智能体提供了多样的数据,智能体的性能在雅达利游戏中得到翻倍的提升. 为加快收敛速度,文献[21]对多智能体深度强化学习算法领域中的优先经验回放机制进行研究,提出一种最大化新产生交互经验优先级的带有优先经验回放机制的多智能体软行动-评论者算法(multi-agent soft Actor-Critic with prioritized experience replay, PER-MASAC)算法. 考虑到多智能体系统对交互经验的庞大需求,在单智能体领域 Ape-X 算法的基础上,本文将优先经验回放机制和分布式强化学习同时引入到多智能体领域,提出一种两者并行的机制,即具有概率优先经验回放机制的分布式多智能体深度强化学习算法(multi-agent soft Actor-Critic with probabilistic prioritized experience replay based on a distributed paradigm, DPER-MASAC),同时针对多智能体本身引起的高数据需求量,照搬原有的单智能体的优先经验回放机制会引发学习者优先抽取的都是最近产生的经验,而过去有价值的经验将难以被抽取用于模型的训练等问题. 本文提出了改进原有优先经验回放机制的优先级定义方式,旨在提高多智能体经验池抽取效率.

首先介绍了单智能体领域分布式优先经验回放机制的核心思想,然后介绍了 MASAC 算法的基本

架构，进而引出本文提出的高效率的多智能体 DPER-MASAC 算法. 在实验环节，从重塑奖励函数的角度设计了 2 种不同难度的捕食者-猎物多智能体任务场景，对本文提出的算法进行测试，并分析了智能体的实际表现和任务完成情况.

1　单智能体分布式强化学习与多智能体深度强化学习

强化学习将单智能体的序列决策过程用马尔可夫决策过程(Markov decision process, MDP)[22]描述，而多智能体系统的序列决策过程在强化学习中遵循马尔可夫博弈过程(Markov game process, MGP)[23]. 这是因为多智能体系统中的单个智能体所得的奖励不仅由自身的策略决定，还参与博弈的其他智能体的策略有关，并且系统状态的转移受到所有智能体联合行动的影响. 分别介绍单智能领域中一种基于 MDP 的分布式强化学习算法和一种基于 CTDE 范式的多智能体深度强化学习算法.

1.1　Ape-X

单智能体领域的深度强化学习过程分为采集经验和训练模型 2 个阶段. 考虑到经验采集过程中需要大批量交互数据，谷歌 DeepMind 团队提出一种将深度 Q 学习网络(deep Q-learning network, DQN)[24]扩展为分布式版本的算法 Ape-X，由多个相互独立的行动者(worker)和一个学习者(learner)组成.

图 1 中的行动者负责通过与环境交互收集经验并将经验存储在全局经验回放池中；学习者基于多个行动者收集到的批量经验数据训练其网络参数，从而学习最优的策略；行动者定期同步学习者最新的网络参数. 一方面，每个行动者可采用不同的行为策略收集经验数据，通过这种分布式架构使得智能体能够充分探索状态空间和策略空间，从而为训练提供更多有价值的交互数据. 另一方面，Ape-X 引入了经验优先回放机制，每个行动者都会计算经验的优先级，学习者会根据经验的优先级进行抽取并对被抽取经验的优先级进行更新.

经验优先级的定义依据

$$P(x)=\frac{p_x^{\alpha}}{\sum_{k=1}^{M}p_k^{\alpha}} \tag{1}$$

进行计算. 式中：$P(x)$为某条经验 x 被采样的概率；p_x 为经验 x 的优先级；M 为经验回放池的存储容量；指数 α 为控制采样在随机和贪婪之间的权重的超参数——当 $\alpha=0$ 时，退化为均匀随机采样；当 $\alpha\neq0$ 时，可对经验优先程度做适当调整.

相应的采样概率依据

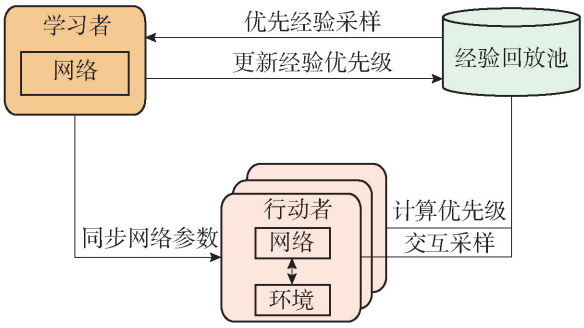

图 1　Ape-X 分布式架构

Fig. 1　Distributed diagram of Ape-X

$$p_x=|\delta_x|+\varepsilon \tag{2}$$

进行计算. 式中：$|\delta_x|$为经验 x 当前 Q 值和目标 Q 值的差值的绝对值；ε 为一个很小的正数，可使 $\delta=0$ 的经验也有概率被抽取.

1.2　MASAC

考虑到多智能体系统状态的转移受所有智能体的影响，对其中的某个智能体而言，若在训练过程中简单地将其他智能体视作环境的一部分，则会导致训练环境的不稳定，造成环境状态转移和奖励值的不确定性，使得算法难以收敛. CTDE 多智能体范式将强化学习的训练阶段和执行阶段分开：在训练阶段，考虑联合观测和联合动作，考虑更多的额外信息帮助值函数对联合策略进行更优的评估，可缓解环境的非平稳性问题；在执行阶段，智能体通过集中式训练习得的决策能力即可根据局部观测做出决策，符合实际场景.

MASAC 是一种基于 CTDE 范式的多智能体深度强化学习算法. 假设智能体交互学习的环境中有 N 个智能体，所有智能体的策略集合为 $\boldsymbol{\pi}=\{\pi_1,\pi_2,\cdots,\pi_N\}$，即每个智能体都有一个单独训练的执行者网络(Actor)和评论者网络(Critic)，分别由$\boldsymbol{\theta}=\{\theta_1,\theta_2,\cdots,\theta_N\}$和$\boldsymbol{\beta}=\{\beta_1,\beta_2,\cdots,\beta_N\}$参数化表示.

在 MASAC 中，每个智能体 i 的 Actor 网络通过最小化损失函数进行更新，Actor 网络损失函数的计算式为

$$J_{\pi}(\theta_i)=E_{o\sim D,a\sim\pi_{\theta}(o)}[\lambda\lg\pi_{\theta}^{i}(a_i|o_i)-Q_{\beta}^{i}(o,a)] \tag{3}$$

式中：D 为经验回放池，使用四元组$\langle o,a,o',r\rangle$存储轨迹经验：$o=\{o_1,o_2,\cdots,o_N\}$为所有智能体的 1 次观测集合，$a=\{a_1,a_2,\cdots,a_N\}$为所有智能体的 1 次动作集合，$o'=\{o_1',o_2',\cdots,o_N'\}$为环境状态转移后所有智能体的 1 次观测集合，$r=\{r_1,r_2,\cdots,r_N\}$为所有智能体的奖励值集合；超参数 λ 为调节熵权重的

参数;每个智能体的动作 a_i 为各自的 Actor 网络基于部分观测 o_i 进行决策.

每个智能体 i 的 Critic 网络通过最小化损失函数进行更新,Critic 网络损失函数的计算式为

$$L_Q(\beta_i)=E_{(o,a,o',a')\sim D}[(Q_\beta^i(o,a)-y_i)^2] \quad (4)$$

式中 y_i 为智能体 i 时序差分的目标值,且

$$y_i=r_i+\gamma E_{a'\sim\pi_{\bar{\theta}}(o')}[Q_{\bar{\beta}}^i(o',a')-\alpha\lg\pi_\theta^i(a_i'|o_i')] \quad (5)$$

式中:为了稳定智能体的训练过程,$\bar{\theta}$ 和 $\bar{\beta}$ 分别为 MASAC[18] 中增加的目标 Actor 网络和目标 Critic 网络的参数;目标 Critic 网络 $Q_{\bar{\beta}}^i$ 函数的输入为环境状态转移后所有智能体 1 次的观测集合 o' 和 a';a' 为每个智能体 i 的目标 Actor 网络根据 o_i' 得到的决策动作 a_i' 所组成的 1 次动作集合. 目标 Actor 网络和目标 Critic 网络的参数采用如下“软”更新的方式进一步提升算法的稳定性,“软“更新公式为

$$\bar{\theta}=\tau\theta+(1-\tau)\bar{\theta} \quad (6)$$

$$\bar{\beta}=\tau\beta+(1-\tau)\bar{\beta} \quad (7)$$

式中:τ 为控制更新比重的超参数,$\tau\ll1$;θ、β、$\bar{\theta}$、$\bar{\beta}$ 分别为 Actor 网络、Critic 网络、目标 Actor 网络、目标 Critic 网络的参数.

2 DPER-MASAC

针对多智能体深度强化学习处理复杂任务时面临的采样效率不高的问题,基于理论基础,本节介绍一种改进的高效率的 MADRL 算法——DPER-MASAC,并对所提算法原理和设计细节进行阐述.

2.1 算法设计

在 MASAC 中,每个智能体都有属于自身的 Actor、Critic 和经验回放池,因此将 MASAC 算法扩展为分布式版本,需要将每个智能体与环境交互的过程扩展为并行的形式来增加数据的多样性,即在每个行动者中都存在 MASAC 中所有智能体 Actor 和 Critic 网络的副本.

假设在有 3 个智能体的任务场景中,DPER-MASAC 包括 1 个学习者和 2 个行动者. 算法架构如图 2 所示,对于智能体 1 而言,在 2 个行动者中都有它的 Actor 网络,因此可以并行地与各自所属行动者的环境进行交互,将交互经验存储到属于智能体 1 的经验回放池中,行动者中的 Critic 网络负责计算所属同一行动者的 Actor 网络与环境交互产生的经验的优先级. 学习者从每个智能体的经验回放池中按照经验的优先级进行抽取,并在 1 次训练完成后,学习者根据最新的网络参数对刚被使用的经验的优先级进行更新.

学习者更新 Critic 网络的损失函数为

$$L_Q(\beta_i)=\frac{1}{m}\sum_{x=1}^{m}w_x(Q_\beta^i(o,a)-y_i)^2 \quad (8)$$

式中:m 为本次抽取的经验数量;采样权重 w_x 的形式为

$$w_x=\frac{(M\cdot P(x))^{-\varphi}}{\max\limits_{1\leqslant k\leqslant m}(w_k)}=\frac{(M\cdot P(x))^{-\varphi}}{\max\limits_{1\leqslant k\leqslant m}((M\cdot P(k))^{-\varphi})}=\left(\frac{P(x)}{\min\limits_{1\leqslant k\leqslant m}P(k)}\right)^{-\varphi} \quad (9)$$

式中 φ 为超参数,随模型的训练线性递增至 1,表示在训练早期鼓励探索,在训练后期保证更新的无偏性.

此处需要特别说明的是,文献[21]中对于经验优先级的设定要求是:每条最新产生的交互经验在被存入经验回放池时,其优先级都被初始化为最大

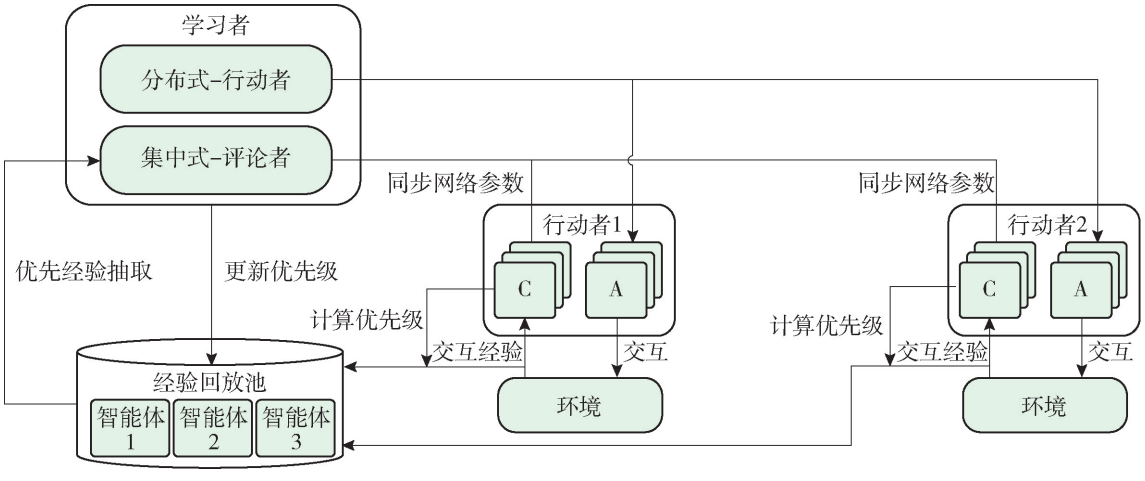

图 2 DPER-MASAC 架构

Fig. 2 Diagram of DPER-MASAC

优先级，在被抽取用于训练之后再根据最新的网络参数计算出该条经验的时序差分误差（temporal differential-error，TD-error）后调整其优先级. 这种方法在非分布式的算法结构中流行且实用. 但在 DPER-MASAC 中，由于每个行动者会通过产生交互经验来扩展数据的吞吐量，若这些交互经验均被初始化为最大的优先级，将会导致学习者优先抽取的都是最近产生的经验，而过去有价值的经验将难以被抽取用于模型的训练. 因此，在 DPER-MASAC 中，每个行动者的 Critic 网络都会计算最新产生经验的优先级，然后再将带有优先级的经验存入全局经验回放池中，通过这种计算的并行性进一步提高学习者的训练效率.

2.2　算法部署

由于 DPER-MASAC 中存在行动者和学习者，二者各司其职并相互配合，因此在算法流程中分为 2 个部分分别介绍. 算法流程分述如下.

2.2.1　算法 1：DPER-MASAC（行动者）

1）初始化 Actor 网络 $\pi = \{\pi_\theta^1, \pi_\theta^2, \cdots, \pi_\theta^N\}$，Critic 网络 $Q = \{Q_\beta^1, Q_\beta^2, \cdots, Q_\beta^N\}$.

2）初始化目标 Actor 网络 $\pi_\theta^i \leftarrow \pi_\theta^i$，目标 Critic 网络.

3）与学习者同步并获取最新的参数，初始化经验缓冲区.

4）For 每个训练回合如下.

收到每个智能体的初始观测 o_i

For 回合内的每个时间步：

每个智能体的 Actor 网络输出采样一个动作

执行联合动作 $a = \{a_1, a_2, \cdots, a_N\}$，接收奖励值并接收下一时间步观测

将 $\langle o, a, o', a' \rangle$ 存入每个行动者的经验缓存区

If 当地经验缓存区存储经验数量达到门限值 G，Then：

批量获取经验数据 B，计算其优先级

将批量经验数据及其优先级一起存入全局经验回放池中

End If

与学习者保持同步最新的网络参数

End For

End For

2.2.2　算法 2：DPER-MASAC（学习者）

1）初始化 Actor 网络 $\pi = \{\pi_\theta^1, \pi_\theta^2, \cdots, \pi_\theta^N\}$，Critic 网络 $Q = \{Q_\beta^1, Q_\beta^2, \cdots, Q_\beta^N\}$.

2）初始化目标 Actor 网络 $\pi_\theta^i \leftarrow \pi_\theta^i$，目标 Critic 网络 $Q_\beta^i \leftarrow Q_\beta^i$.

3）初始化全局经验回放池 D 和学习周期数目.

4）For 每个学习周期如下.

按照经验优先回放机制，从经验回放池中抽取 m 条经验，计算被抽取的概率 $P(x)$ 和重要性采样权重 w_x

基于式（8）更新 Critic 网络

基于式（3）更新 Actor 网络

计算被抽取经验的 TD-error 值，并对它们的优先级进行更新

基于式（6）（7）更新目标网络参数

周期性更新全局经验回放池中经验数据

End For

3　仿真

3.1　实验部署

为验证所提算法在多智能体任务场景中的有效性，本文选用图所示多智能体粒子环境（multi-agent particle envs，MPE）中捕食者-猎物（predator-prey，PP）任务场景展开仿真实验. 与其他实验环境相比，PP 是一个混合型多智能体任务场景，同时包含竞争和合作的多智能体关系，广泛被用作验证 MADRL 算法的测试环境，具有代表性和可信性.

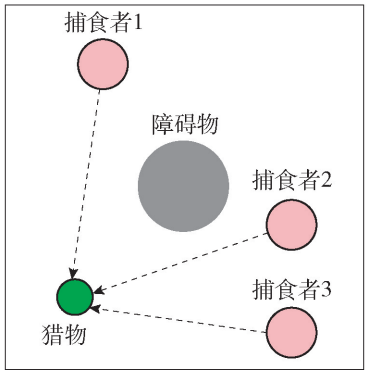

图 3　捕食者-猎物任务场景

Fig. 3　Mission scenario of predator-prey

任务场景中存在 3 个红色的捕食者和 1 个绿色的猎物共 4 个智能体，环境中央存在 1 个灰色的障碍物，猎物可借助障碍物来躲避捕食者的追捕. 需特别说明的是，周围的黑色边框是为了限制猎物的移动范围，防止其为躲避追击而移动到地图之外. 任务目的是捕食者需学会避免相互之间发生碰撞的前提下以一种合作的方式捕捉猎物. 该任务场景规定：捕食者只有成功捕捉到猎物才能收获正向奖励，猎物只有在被捕捉时才有负向惩罚，那么在训练前

期任务场景与智能体之间的交互信息稀少,学习效率缓慢,因此该任务场景存在奖励值稀疏问题.

重构奖励函数可从奖励函数的设计角度缓解奖励值稀疏问题,那么 3 个捕食者的合作共享奖励函数为

$$R^{\text{predator}} = -0.1 \times \sum_{i=1}^{3} D(i,\text{prey}) + \sum_{i=1}^{3} C_i$$

$$C_i = \begin{cases} 10, & \text{捕食者 } i \text{ 捕食成功} \\ 0, & \text{捕食者 } i \text{ 捕食失败} \end{cases} \quad (10)$$

式中:$D(i,\text{prey})$为第 i 个智能体与猎物之间的距离,距离越近,奖励值越大,从而引导捕食者相互配合快速捕捉猎物;第 2 项为当任意一个捕食者与猎物发生碰撞即为捕捉成功,捕食者收获正奖励值 10.

猎物的奖励函数大小等于捕食者奖励函数的相反数,其形式为

$$R^{\text{prey}} = 0.1 \times \sum_{i=1}^{3} D(i,\text{prey}) - \sum_{i=1}^{3} C_i \quad (11)$$

进一步,通过重构的奖励函数从捕食者的追捕能力和猎物的逃逸能力 2 个方面设计 2 种任务场景.

表 1 2 种不同场景下的捕食者–猎物任务

Table 1 Mission of predator-prey in two different scenarios

场景	捕食者有无引导奖励	猎物有无引导奖励
1	否	否
2	否	是

场景 1 是一个典型的奖励稀疏环境,猎物和捕食者均不具有引导奖励,用于检验 DPER-MASAC 在奖励稀疏环境下的性能表现. 场景 2 中仅猎物具有引导奖励,逃逸能力更强而捕食者奖励稀疏,任务难度升级,进一步检验 DPER-MASAC 在高难度任务场景中的表现.

3.2 结果分析

实验的软件环境为 ubuntu16.04 + Tensorflow + gym,硬件为英伟达 GeForce GTX 2080 +32GB 内存. 设置学习率为 0.01,强化学习折扣因子为 0.95,经验回放池大小为 1 000 000,每次训练从经验池抽取 512 批次大小的数据. 对于本文所提的 DPER-MASAC 算法,设置 2 个行动者和 1 个学习者. 实验中具体的超参数设置如表 2 所示.

对于每个任务场景,均采用 5 个不同的随机种子来提高实验结果的可靠性. 设置每个任务场景下的每次训练有 2 500 个回合,每个回合最大步长设置为 200. 在每个任务场景中分别采用 MASAC、PER-MASAC 和 DPER-MASAC 训练捕食者,固定采用 MADDPG 算法训练猎物. 为了能够更加清晰地分析回合奖励的走势,通过分析相同任务场景下 5 次随机试验中捕食者的回合奖励均值和任务完成情况来对比算法的性能.

表 2 PER-MASAC/DPER-MASAC 算法超参数

Table 2 Hyperparameter of PER-MASAC and DPER-MASAC

PER-MASAC/DPER-MASAC 算法超参数类型	取值设置
优化器	Adam
批处理大小 B	512
激活函数	ReLU
折扣因子 γ	0.95
经验回放池大小 M	100 000
ε	0.02
α	0.6
φ	0.4
Actor 网络学习率	0.01
Critic 网络学习率	0.01
网络隐藏层神经元个数	64
目标网络软更新超参数 τ	0.01

场景 1 和场景 2 下的实验结果分别如图 4、5 所示,深色曲线表示捕食者回合奖励的均值,阴影部分表示方差. 这里将相同场景中不同算法实验效果的对比称为“横向对比”,将同一算法在不同实验任务场景中的对比称为“纵向对比”. 如图 4、5 所示,在 2 种任务场景中,PER-MASAC 和 MASAC 的纵向对比之间,PER-MASAC 达到收敛稳定所需要的训练回合更少. 这说明了采用优先级采样机制的智能体可在相同采样次数和相同采样数量的经验中学得更多的内容. 在 2 种任务场景下,DPER-MASAC 的最终性能显然均优于 PER-MASAC,这说明 DPER-MASAC 的分布式结构扩展经验数据吞吐量并增加数据多样性,更容易探索到具有高回报的经验数据,缓解过拟合和过早陷入局部最优的问题. 总的来看,DPER-MASAC 的方差最小,算法效果更稳定. 对于捕食者和猎物均不带有引导奖励的场景 1,任务

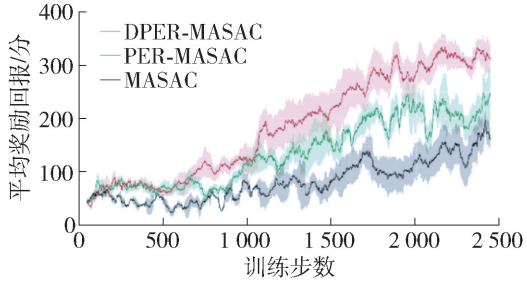

图 4 场景 1 中捕食者的平均回合奖励

Fig. 4 Mean of episode rewards of predators in scenario 1

环境的奖励值稀疏，DPER-MASAC 多智能体系统的回合奖励均值可从负值收敛到 325，这足以说明捕食者学到了成功捕捉猎物的协同策略. 而即使是在仅猎物具有引导奖励的场景 2 中，即捕食者追捕能力低下且猎物逃逸能力更强的任务环境，DPER-MASAC 多智能体系统的回合奖励均值也从负值收敛到 260，较 PER-MASAC 以及 MASAC 算法的性能优势更加明显，证实了 DPER-MASAC 在处理复杂任务的优势.

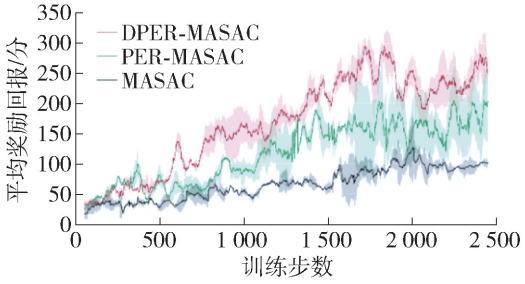

图 5　场景 2 中捕食者的平均回合奖励

Fig. 5　Mean of episode rewards of predators in scenario 2

训练结束后，针对 DPER-MASAC 得到的最优策略，在同样设置的场景 1 和场景 2 的任务环境中另组织一个回合的测试实验来评估捕食者学到的策略. 以测试回合中的捕食者成功捕捉猎物的次数作为评估标准，定义每个回合中的某时间步只要有一个捕食者捕捉到猎物，即记作捕食者团队在该回合内成功捕捉到猎物 1 次. 无论是在场景 1 还是场景 2 的任务场景中，经 DPER-MASAC 训练后的捕食者捕捉到猎物的次数均高于另外 2 种算法训练所得捕食者的捕捉次数，这说明经 DPER-MASAC 训练的捕食者的协作追捕策略是最佳的，具有更强的决策水平. 某回合中捕食者捕食次数如表 3 所示.

表 3　某回合中捕食者捕食次数

Table 3　Success number of predations in an episode

任务	DPER-MASAC	PER-MASAC	MASAC
场景 1	29	23	15
场景 2	23	19	9

图 6、7 分别对测试阶段捕食者团队在 2 个场景中初次捕获到的猎物的过程以(a) ~ (d)的顺序进行渲染. 在每个场景中，(a)描述回合开始时的初始位置，(d)是捕食者团队第 1 次成功碰撞到猎物的追捕情况. 可以发现，在不同的初始位置的情况下，3 个捕食者均能以相互协同的追捕策略捕捉到猎物. 需要特别说明的是，在场景 2 的渲染图中，左右两侧的捕食者并没有一味地靠近捕食者，而是预判猎物的移动方向从而封堵其逃逸路线，左右两侧的捕食者放弃了短期较高的奖励值来完成整体的追捕任务，说明捕食者学会了相互协作的追捕策略. 进一步验证了 DPER-MASAC 算法在分布式架构和改进的优先经验回放并行机制下的算法效率.

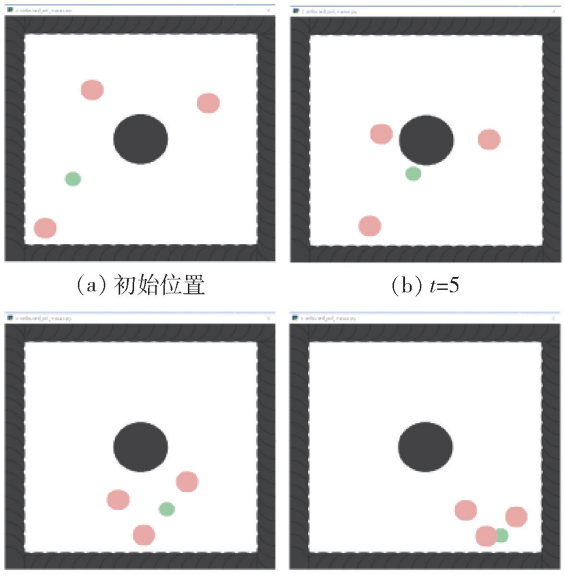

图 6　场景 1 中捕食过程渲染图

Fig. 6　Rendering picture of predation process in scenario 1

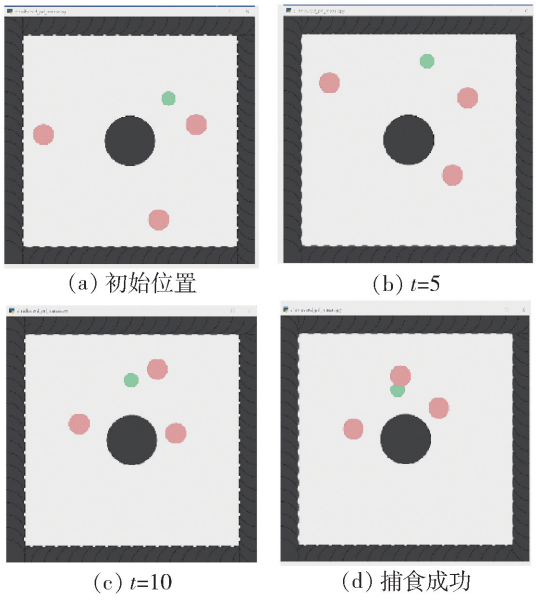

图 7　场景 2 中捕食过程渲染图

Fig. 7　Rendering picture of predation process in scenario 2

4　结论

1) 本文提出的 DPER-MASAC 算法在训练效率以及最终算法性能优于基线算法，效果最佳.

2）通过本文提出的 DPER-MASAC 算法，具有合作的捕食者队伍可以学到一种协同追捕策略，相互配合完成追捕. 优秀的协同追捕策略并不是每个捕食者贪心策略的集合，部分捕食者会放弃短期较高的奖励值来完成整体的追捕任务.

参考文献：

[1] KIRAN B R, SOBH I, TALPAERT V, et al. Deep reinforcement learning for autonomous driving: a survey[C]// 2021 IEEE Transactions on Intelligent Transportation Systems. Piscataway: IEEE, 2021.

[2] 张皓然. Actor-Critic 强化学习方法及在船舶自动靠泊中的应用[D]. 北京:北京交通大学, 2021.
ZHANG H R. Actor-Critic reinforcement learning and applications to automatic ship berthing [D]. Beijing: Beijing Jiaotong University, 2021. (in Chinese)

[3] ZHAO W S, QUERALTA J P, WESTERLUND T. Sim-to-real transfer in deep reinforcement learning for robotics: a survey [C] // 2020 IEEE Symposium Series on Computational Intelligence. Piscataway: IEEE, 2020.

[4] GRONAUER S, DIEPOLD K. Multi-agent deep reinforcement learning: a survey[J]. Artificial Intelligence Review, 2021, 55(2): 859-943.

[5] YANG Y, WANG J. An overview of multi-agent reinforcement learning from game theoretical perspective [J]. ArXiv Preprint ArXiv, 2020: 2011. 00583.

[6] ZHANG K, YANG Z, BASAR T. Multi-agent reinforcement learning: a selective overview of theories and algorithms[J]. Handbook of Reinforcement Learning and Control, 2021, 325(7): 321-384.

[7] OROOJLOOYJADID A, HAJINEZHAD D. A review of cooperative multi-agent deep reinforcement learning[J]. ArXiv Preprint ArXiv, 2019: 1908. 03963.

[8] VINYALS O, BABUSCHKIN I, CZARNECKI W M, et al. Grandmaster level in StarCraft II using multi-agent reinforcement learning[J]. Nature, 2019, 575(7782): 350-354.

[9] JAMIL A R M, GANGULY K K, NOWER N. Adaptive traffic signal control system using composite reward architecture based deep reinforcement learning[J]. IET Intelligent Transport Systems, 2020, 14(14): 2030-2041.

[10] CHEN C, WEI H, XU N, et al. Toward a thousand lights: decentralized deep reinforcement learning for large-scale traffic signal control[C]// Proceedings of the 34th AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2020: 3414-3421.

[11] WEI H, XU N, ZHANG H, et al. CoLight: learning network-level cooperation for traffic signal control[C]// Proceedings of the 28th ACM International Conference on Information and Knowledge Management. New York: Association for Computing Machinery, 2019: 1913-1922.

[12] 施伟, 冯旸赫, 程光权. 基于深度强化学习的多机协同空战方法研究[J]. 自动化学报, 2021, 47(7): 1610-1623.
SHI W, FENG Y H, CHENG G Q, et al. Research on multi-aircraft cooperative air combat method based on deep reinforcement learning[J]. Acta Automatic Sinica, 2021, 47(7): 1610-1623. (in Chinese)

[13] YANG Y, LI J T, PENG L L. Multi-robot path planning based on a deep reinforcement learning DQN algorithm [J]. CAAI Transactions on Intelligence Technology, 2020, 5(3): 177-183.

[14] ZHANG Y, TANG B, YANG Q, et al. BCORLE (λ): an offline reinforcement learning and evaluation framework for coupons allocation in E-commerce market [C] // Proceedings of the 35th Conference on Neural Information Processing Systems. New York: Curran Associates Inc, 2021: 20410-20422.

[15] ZHOU T, TANG D, ZHU H, et al. Multi-agent reinforcement learning for online scheduling in smart factories [J]. Robotics and Computer-Integrated Manufacturing, 2021, 72: 102202-102210.

[16] CHRISTIANOS F, SCHAFER L, ALBRECHT S V. Shared experience actor-critic for multi-agent reinforcement learning [C] // Proceedings of the 34th International Conference on Neural Information Processing Systems. New York: Curran Associates Inc, 2020.

[17] LOWE R, WU Y, TAMAR A, et al. Multi-agent actor-critic for mixed cooperative-competitive environments[J]. ArXiv Preprint ArXiv, 2017: 1706. 02275.

[18] WANG Z H, ZHANG Y X, YIN C K. Multi-agent deep reinforcement learning based on maximum entropy[C] // 2021 IEEE 4th Advanced Information Management Communicates Electronic and Automation Control Conference. Piscataway: IEEE, 2021.

[19] LILLICRAP T P, HUNT J J, PRITZEL A, et al. Continuous control with deep reinforcement learning[J]. ArXiv Preprint ArXiv, 2015: 1509. 02971.

[20] HORGAN D, QUAN J, BUDDEN D, et al. Distributed prioritized experience replay[C]// Proceedings of the 6th International Conference on Learning Representations. Massachusetts: OpenReview. net, 2018: 192-211.

[21] FAN S, SONG G, YANG B, et al. Prioritized experience replay in multi-actor-attention-critic for reinforcement learning [J] // Journal of Physics: Conference Series, 2020, 1631(1): 012040

[22] SUTTON R S, BARTO A G. Reinforcement learning: an introduction [M]. Cambridge: MIT Press, 2018.

[23] CHUNG J, GULCEHRE C, CHO K H, et al. Empirical evaluation of gated recurrent neural networks on sequence modeling[J]. ArXiv Preprint ArXiv, 2014: 1412. 3555.

[24] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Playing Atari with deep reinforcement learning[J]. ArXiv Preprint ArXiv, 2013: 1312. 5602.

（责任编辑 杨开英）