电讯技术
Telecommunication Engineering
ISSN 1001-893X,CN 51-1267/TN

《电讯技术》网络首发论文

题目：基于深度强化学习的智能决策方法
作者：熊蓉玲，段春怡，冉华明，杨萌，冯旸赫
收稿日期：2021-11-23
网络首发日期：2022-08-23
引用格式：熊蓉玲，段春怡，冉华明，杨萌，冯旸赫．基于深度强化学习的智能决策方法[J/OL]．电讯技术．
https://kns.cnki.net/kcms/detail/51.1267.TN.20220822.1729.004.html

网络首发时间：2022-08-23 11:06:44
网络首发地址：https://kns.cnki.net/kcms/detail/51.1267.TN.20220822.1729.004.html

基于深度强化学习的智能决策方法

熊蓉玲[1]，段春怡[1]，冉华明[1]，杨萌[2]，冯旸赫[3]

（1.中国电子科技集团公司第十研究所，成都 610036；2.西南交通大学 数学学院，成都 611756；3. 国防科技大学 系统工程学院，长沙 410003）

摘　要：针对传统深度强化学习算法难以快速解决长时序复杂任务的问题，提出了一种引入历史信息和人类知识的深度强化学习方法，对经典近端策略优化（PPO）强化学习算法进行改进，在状态空间引入历史状态以反映环境的时序变化特征，在策略模型中基于人类认知增加无效动作掩膜，禁止智能体进行无效探索，提高探索效率，从而提升模型的训练性能。仿真结果表明，所提方法能够有效解决长时序复杂任务的智能决策问题，相比传统的深度强化学习算法可显著提高模型收敛效果。

关键词：智能决策；深度强化学习；近端策略优化；动作掩膜

开放科学（资源服务）标识码（OSID）：

中图分类号：TP181　　文献标志码：A

Intelligent Decision Making Method Based on Deep Reinforcement Learning

XIONG Rongling[1]，　DUAN Chunyi[1]，　RAN Huaming[1]，　YANG Meng[2]，

FENG Yanghe[3]

(1.The 10th Research Institute of China Electronics Technology Group Corporation, Chengdu 610036, China；2.School of Mathematics, Southwest Jiaotong University, Chengdu 611756, China；3.College of System Engineering, National University of Defense Technology, Changsha 410003, China)

Abstract: Traditional deep reinforcement learning algorithm is hard to solve complex long time-series tasks quickly. A deep reinforcement learning method introducing historical information and human knowledge is proposed. The classical proximal policy optimization (PPO) algorithm is improved by introducing historical information in state space to reflect the temporal changing characteristics of the environment. Invalid action mask is added in the policy model based on human cognition to prohibit the agent from invalid exploration and improve the exploration efficiency, so as to improve the training performance of the model. Simulation results show that the proposed method can solve intelligent decision making problem of complex long time-series tasks efficiently, and improve the convergence performance of the model significantly compared with traditional deep reinforcement learning algorithm.

Key Words: intelligent decision making; deep reinforcement learning; proximal policy optimization; action mask

0 引　言

强化学习通过智能体和环境不断试错交互的方式学习到能够使奖励最大化的最优策略[1]。深度强化学习结合神经网络强大的表征能力来拟合智能体的策略模型和价值模型，求解复杂问题的能力大幅提升，近年来在各类智能决策问题上取得了巨大的进步，成为人工智能领域发展迅猛的一个分支[2]。实时策略类游戏作为典型的时序决策问题，成为国内外学者深度强化学习研究的试金石。Mnih等[3]提出深度 Q 网络（Deep Q Network，DQN）算

*收稿日期：2021-11-23；修回日期：2022-01-21
通信作者：熊蓉玲 463245987@qq.com

法解决 Atari2600 游戏，在 6 个游戏中的表现优于以前的方法，在 3 个游戏中的表现优于人类专家。但 Atari2600 游戏的任务场景较为简单，决策时序较短，决策空间较小，问题复杂性不高。Vinyals 等[4]针对星际争霸游戏问题，采用了强化学习和模仿学习相结合的方式，解决了非完全信息下的即时策略游戏问题；Jaderberg 等[5]针对雷神之锤游戏问题，利用双层流程来优化智能体的内部奖励机制，再通过这些奖励来优化强化学习模型，通过并行训练多个不同的智能体集群相互配合，实现了完全无监督的自学机制。与 Atari2600 游戏相比，星际争霸和雷神之锤的任务场景复杂，决策难度大幅提升，文中设计的算法架构复杂，计算资源需求大，训练时间长，难以应用到其他任务场景中。

针对传统深度强化学习方法难以快速解决长时序复杂任务的问题，本文提出一种引入历史信息和人类知识的深度强化学习方法，主要贡献如下：

（1）问题建模时在状态空间引入历史状态信息，以反映任务场景的时序变化特征，增加智能体决策时的环境状态输入，提高智能体对环境状态变化理解的准确性。

（2）在策略模型中基于人类知识引入无效动作掩膜，禁止智能体进行无效探索，提高探索效率，缩短模型的训练时间。

（3）通过仿真实验验证了本文所提方法可有效解决长时序复杂任务的智能决策问题。训练结果表明，与传统的深度强化学习算法相比，本文所提的方法可显著提高模型收敛速度。

1 背景介绍

1.1 深度强化学习

强化学习主要关注智能体如何在环境中采取不同的行动，以最大限度地提高累积奖励。强化学习主要由智能体、环境、状态、动作、奖励组成[6]。其中，状态空间用状态集合 S 表示，动作空间用动作集合 A 表示，则智能体与环境的交互过程为：当给定环境的某个状态 s∈S，智能体将根据当前的策略 π(a|s)执行某个动作 a∈A，环境迁移到新的状态 $s^{'}$∈S，同时智能体从环境获得奖励 r(s,a)。智能体根据环境反馈的奖励，对自身的策略模型进行更新，以学会最佳决策序列。

为了表示累积奖励，通常使用折扣未来累积奖励来代替：

$$R_t=\sum_{i=t}^{T}\gamma^{i}r\left(s_i,a_i\right) \tag{1}$$

其中，γ 为折扣系数。

当执行到某一步时，需要评估当前智能体在该时间步状态的好坏程度，主要由值函数来完成，包括状态值函数 $V^{\pi}(s)$和动作-状态值函数 $Q^{\pi}(s,a)$两类。

$$V^{\pi}\left(s\right)=E\left[R_t\left|s_t=s\right.\right] \tag{2}$$

$$Q^{\pi}\left(s,a\right)=E\left[R_t\left|s_t=s,a_t=a\right.\right] \tag{3}$$

强化学习的核心思想是使用值函数找到最优的策略，通常采用求解贝尔曼方程的方法。

$$\begin{aligned}V^{*}\left(s\right)&=\max_{a\in A}E\left[R_{t+1}+\gamma V^{*}\left(s_{t+1}\right)\left|s_t=s,a_t=a\right.\right]\\&=\max_{a\in A}\sum_{s^{'},r}p\left(s^{'},r\left|s,a\right.\right)\left[r+\gamma V^{*}\left(s^{'}\right)\right]\end{aligned} \tag{4}$$

或者：

$$\begin{aligned}Q^{*}\left(s,a\right)&=\max_{a\in A}E\left[R_{t+1}+\gamma Q^{*}\left(s_{t+1},a^{'}\right)\left|s_t=s,a_t=a\right.\right]\\&=\max_{a\in A}\sum_{s^{'},r}p\left(s^{'},r\left|s,a\right.\right)\left[r+\gamma Q^{*}\left(s^{'},a^{'}\right)\right]\end{aligned} \tag{5}$$

其中，$p\left(s^{'},r\left|s,a\right.\right)$为状态转移概率。

深度学习通过神经网络的逐层组合，最终提取能够代表数据最本质的高维抽象特征，具有极强的表征能力。深度强化学习使用强化学习定义问题和优化目标，使用深度学习求解策略函数或者价值函数，充分利用了强化学习的决策优势和深度学习的感知优势，近年来在很多任务上取得了巨大的成功。

1.2 近端策略优化

深度强化学习算法大体上可分为三类，即值函数方法、策略搜索方法和混合型的行动者-执行者（Actor-Critic，AC）算法类型。典型的深度强化学习算法包括 DQN[7]、优势行动者-执行者（Advantage Actor-Critic，A2C）[8]、确定性策略梯度（Deterministic Policy Gradient，DPG）[9]、置信区域策略优化（Trust Region Policy Optimization，TRPO）[10]、近端策略优化（Proximal Policy Optimization，PPO）[11]等。经过实验对比发现，PPO 算法的整体表现更优，经常作为深度强化学习应用中的首选算法。

PPO 算法是在 TRPO 算法的基础上，使用截断的方式构建目标函数，以保证新策略和旧策略的差

异控制在一定范围内，提高算法模型训练的稳定性。

$$L_t^{CLIP}(\theta)=\sum_{(s_t,a_t)}\min\left(r_t(\theta)A(s_t,a_t),clip\left(r_t(\theta),1-\varepsilon,1+\varepsilon\right)A(s_t,a_t)\right) \quad (6)$$

其中，ε 为截断系数，$r_t(\theta)$ 为新策略和旧策略的比率。

$$r_t(\theta)=\frac{\pi(s_t,a_t)}{\pi_{old}(s_t,a_t)} \quad (7)$$

变量 $A(s_t,a_t)$ 为优势函数，有助于在保持无偏差的情况下，尽可能地降低方差值。

$$A(s_t,a_t)=Q(s_t,a_t)-V(s_t) \quad (8)$$

PPO 算法除了关于策略的目标函数，还在目标函数中引入值函数的目标函数 $L_t^{VF}(\theta)$ 和策略模型的熵 $S[\pi_\theta](s_t)$，因而完整的目标函数为：

$$L_t^{CLIP+VF+S}(\theta)=\hat{E}_t\left[L_t^{CLIP}(\theta)-c_1L_t^{VF}(\theta)+c_2S[\pi_\theta](s_t)\right]$$

$$L_t^{VF}(\theta)=\left(V_\theta(s_t)-V_t^{target}\right)^2 \quad (9)$$

$$S[\pi_\theta](s_t)=-\pi_\theta(a|s)\log\pi_\theta(a|s)$$

其中，c_1、c_2 分别为值函数和熵的系数，V_t^{target} 为目标网络的值函数估计值。

2 深度强化学习决策模型

2.1 问题建模

本文考虑经典飞行射击类游戏的任务场景：对抗双方的智能体在模拟环境中各操控一架飞机从基地起飞，在飞行过程中智能体根据飞机传感器探测到的对手信息和自身平台的信息进行控制决策，以对飞机进行全方位的操控，包括飞行控制、雷达控制、电子战控制、武器控制，最终达到击落对方飞机的目的。

针对上述飞行射击类游戏场景，本文采用深度强化学习方法对其进行问题建模，明确深度强化学习算法模型的状态空间、动作空间和奖励。首先，在强化学习决策模型中，引入历史状态和动作信息作为状态输入，让智能体进行游戏决策时可以显式地获取历史信息。其次，将游戏过程中对最终胜负有贡献的关键事件作为中间奖励，以显式引导智能体如何获胜。这样设计可以带来以下三方面的好处：

（1）可以让智能体更准确得掌握对手的状态信息。在游戏过程中，对方飞机如果进入我方传感器的探测范围内，我方智能体只能获得对手飞机的位置信息，无法获得其航向和速度信息，通过引入上一时刻对手飞机的位置，智能体可以隐式获得其航向和速度信息，可以帮助智能体更准确地进行飞行航向和速度控制决策。

（2）可以帮助智能体保持控制决策的一致性。飞行射击类游戏属于典型的长时序连续决策问题，引入历史动作信息可以让智能体显式获得过往动作，有利于其保持决策的一致性，在游戏过程中避免无意义的动作频繁变更，比如无意义的大幅机动、连续发弹、雷达频繁开关机等。

（3）可以帮助智能体更快掌握获胜方法。现有的强化学习方法通常只根据游戏的胜负设置奖励，但飞行射击类游戏通常单局游戏的时长为 20 分钟左右，只根据最终的胜负进行奖励反馈属于典型的稀疏回报问题，智能体很难学习到有效策略。将游戏过程中对最终胜负有贡献的关键事件作为中间奖励，比如发现目标、武器发射等，可以引导智能体更快学习到获胜策略。

本文采用 PPO 算法构建飞行射击类游戏智能决策模型，状态空间来源于飞机传感器探测到的对手信息和自身平台的信息，包括飞机位置、传感器状态、武器状态等共 20 项，动作空间包括飞机飞行控制、雷达控制、电子战控制、武器控制共 8 项控制项，状态空间、动作空间和奖励的详细信息分别如表 1、表 2、表 3 所示。

智能体在每个时间步根据最新的环境状态以及历史环境状态和动作进行动作决策，考虑到过多引入历史信息会加大状态空间，从而影响训练效率，本文在最新环境状态的基础上只引入上一步的历史状态和动作信息作为状态输入进行动作决策，游戏环境接收到决策动作进行环境状态更新并反馈奖励，重复上述过程直到游戏结束。

表1 状态空间详细信息

序号	状态项	取值范围
1	对方飞机距离	0~150km
2	对方飞机方位角	-180°~180°
3	对方飞机俯仰角	-90°~90°
4	对方雷达状态	关机、扫描、跟踪
5	对方剩余武器数量	0~6
6	对方武器状态	未发射、在飞
7	对方武器方位角	-180°~180°
8	对方武器俯仰角	-90°~90°
9	我方飞机速度	300m/s~600m/s
10	我方飞机航向角	-180°~180°

11	我方飞机俯仰角	-90°~90°
12	我方飞机滚转角	-180°~180°
13	我方剩余油量	0~4500kg
14	我方雷达状态	关机、扫描、跟踪
15	我方雷达已失跟时间	0~150s
16	我方雷达已开机时间	0~10s
17	我方雷达已关机时间	0~10s
18	干扰已持续时间	0~10s
19	我方剩余武器数量	0~6
20	我方武器状态	未发射、在飞

表2 动作空间详细信息

序号	动作项	取值范围
1	期望飞行航向角	-180°~180°
2	期望飞行滚转角	-180°~180°
3	期望飞行速度	300m/s~600m/s
4	武器发射控制	不发射、发射
5	雷达开关机控制	开机、关机
6	雷达扫描范围	大、中、小
7	电子战开关机控制	开机、关机
8	电子战干扰功率	大、中、小

表3 奖励详细信息

序号	动作项	取值范围
1	游戏推进1个周期	-0.01
2	雷达探测目标	+1
3	我方武器发射	+20
4	对方武器发射	-20
5	我方武器未击中目标	-50
6	对方武器未击中目标	+50
7	我方击落对方飞机	+100
8	对方击落我方飞机	-100

为了缓解智能体与游戏环境频繁通信交互导致的训练时间过长的问题，以及智能体频繁决策动作变更导致的前后决策不一致的问题，智能体的 1 个时间步对应游戏环境中的 10 个推进周期，即 1 次动作指令在 10 个游戏推进周期内执行。

2.2 无效动作掩膜

在强化学习中，智能体在探索阶段可以在整个动作空间内进行试错探索。但是，针对特定的环境状态，可能存在无效或者不合理的动作。智能体一旦探索到这些无效或者不合理的动作，会导致模型训练收敛的时间变长。通常，任务场景的动作空间越大，其中无效或不合理的动作越多。

针对飞行射击类游戏而言，其中存在的无效或者不合理的动作包括：

（1）尚未发现对方飞机却大幅机动转弯；

（2）雷达尚未探测到对方飞机却发射武器；

（3）对方雷达尚未探测到我方飞机（对方雷达状态为关机或扫描）却电子战开机干扰。

因此，本文采用无效动作掩膜[12]的方式避免智能体进行无效或不合理的探索，以提高模型的训练收敛速度。

在 PPO 算法中，策略网络输出为未归一化的概率（logits），然后经过 softmax 操作转变为归一化的概率值，根据不同动作的概率分布来进行动作选择。

$$p_\theta\left(a_i \mid s_t\right)=Soft\max\left(l_i\right)=\frac{e^{l_i}}{\sum_j e^{l_j}} \tag{10}$$

当某个动作是无效或不合理时，只需要将对应策略网络输出（logits）替换为无穷小的值（如 $-1*10^8$），则经过 softmax 操作后该动作被选择的概率趋近于 0，以此来实现该动作的禁用。

同时，当某个时间步智能体采用了无效动作掩膜禁用某个动作时，对应的策略梯度为 0，从而保证了替换操作不会给策略网络参数带来负面影响，证明如下。

假设，探索阶段智能体采用策略网络 $\pi(\theta)$ 与游戏环境交互，收集了一定数量的样本数据 $\tau^k=\left\{s_0^k,a_0^k,s_1^k,a_1^k,\cdots,s_n^k,a_n^k\right\}$，则策略梯度计算公式为：

$$g_{policy}=\frac{1}{N}\sum_k\sum_t\nabla_\theta\log p_\theta\left(a_t^k \mid s_t^k\right)R^k \tag{11}$$

其中，N 为样本数据总量，R^k 是第 k 条样本的累积奖励。

针对式（11）中的梯度计算部分进行公式推导：

$$\nabla_\theta\log p_\theta\left(a_t^k \mid s_t^k\right)=\frac{1}{p_\theta\left(a_t^k \mid s_t^k\right)}\nabla_\theta p_\theta\left(a_t^k \mid s_t^k\right) \tag{12}$$

$$\nabla_\theta p_\theta\left(a_t^k \mid s_t^k\right)=\sum_j\nabla_{l_j} p_\theta\left(a_t^k \mid s_t^k\right)\nabla_\theta l_j \tag{13}$$

$$\nabla_{l_j} p_\theta\left(a_t^k \mid s_t^k\right)=\nabla\frac{e^{l_i}}{\sum_j e^{l_j}}=\begin{cases}1-\dfrac{e^{l_j}}{\sum_j e^{l_j}} & j=i\\ -\dfrac{e^{l_j}}{\sum_j e^{l_j}} & j\neq i\end{cases} \tag{14}$$

假设在某个时间步，智能体采用无效动作掩膜禁用了动作 a_0，并按照动作概率分布选择执行动作

a_1，则在式（14）中，i=1，无效动作 a_0 对应的梯度为：

$$\nabla_{l_0} p_\theta\left(a_t^k \mid s_t^k\right) = -\frac{e^{l_0}}{\sum_j e^{l_j}} = -\frac{e^{-10^8}}{\sum_j e^{l_j}} \approx 0 \qquad (15)$$

因此，证明得到无效动作掩膜对应的策略梯度为 0，策略网络参数的更新不受动作掩膜的影响，从而保证了替换操作不会给策略网络参数带来负面影响。

为了验证上述计算过程的正确性，在接收到最新的环境状态时，人为将对方飞机的信息屏蔽来模拟尚未发现对方飞机的情况，在此情况下，智能体输出的机动决策均为保持直飞，证明无效动作掩膜确实屏蔽了无效的大幅机动。同时，对比了该情况下策略网络模型更新前后的网络参数值，发现网络参数值未发生改变，证明无效动作掩膜对应的策略梯度确实为 0，不会给策略网络参数带来负面影响。

3 仿真结果及分析

3.1 仿真场景设置

本文所提的 PPO 智能决策算法模型是在 Stable Baselines[13]中 PPO2 算法源代码的基础上增加无效动作掩膜实现的，并调用该算法库中的矢量化环境模块（SubprocVec）实现多进程并行采样。矢量化环境是一种将多重独立环境堆叠成单一环境的方法，可以实现同时在多个环境上进行并行交互采样，以提高智能体的探索效率。

游戏环境则是基于 OpenAI Gym 框架[14]对飞行射击类游戏进行封装。智能体与游戏平台之间采用用户数据包协议（User Datagram Protocol，UDP）通信，通过最基本的套接字的方式进行信息交互，以减少网络堵塞，缩短通信时间。

本文中设定的任务区域如下图所示，分为待战区和自由交战区，待战区的大小为 15km*25km，自由交战区的大小为 150km*25km。对抗双方的初始经纬度可在各自待战区内任意选择，双方飞机初始航向为东西方向对飞，初始高度均为 8000 米，初始速度均为 1 马赫。

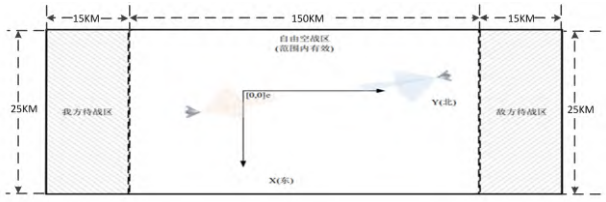

图1 游戏任务区域场景示意图

决策智能体的训练集为 3 个固定场景，采用相对态势的思路，敌方飞机的初始经纬度保持在其待战区的中心位置，我方飞机的初始经纬度分别在其待战区的上方（-12.5km）、中间（0km）、下方（12.5km）。测试集为 100 个随机场景，对抗双方飞机的初始经纬度在各自待战区内随机生成。

3.2 训练结果

在深度强化学习算法模型的训练中，超参数的设置对模型训练的影响较大，尤其是学习率的合理设置尤为重要。本文通过对比不同学习率情况下的损失函数收敛曲线和随机测试场景对抗胜率收敛曲线来进行学习率的选择。其中，随机测试场景对抗胜率用于验证决策智能体对不同任务场景的适应性，采用 100 个随机场景，每训练 10 个回合测试一次，从而展现训练过程中决策智能体对抗胜率的变化。

图 2 为学习率分别设置为 10^{-5}、10^{-4}、10^{-3} 时决策智能体损失函数的收敛曲线。从图中可以看出，随着学习率的不断增大，损失函数收敛得更快，损失值更小。当学习率设置为 10^{-3} 时，损失函数收敛到 0.0005 附近。

图 3 为学习率分别设置为 10^{-5}、10^{-4}、10^{-3} 时决策智能体随机测试场景对抗胜率的收敛曲线。从图中可以看出，当学习率为 10^{-5} 时，智能体几乎不能找到有效的策略，平均胜率在 15%左右；当学习率为 10^{-4} 时，智能体在训练 150 个回合后平均胜率在 50%左右；当学习率为 10^{-3} 时，决策智能体在训练 150 个回合后可以稳定达到 80%左右的对抗胜率。

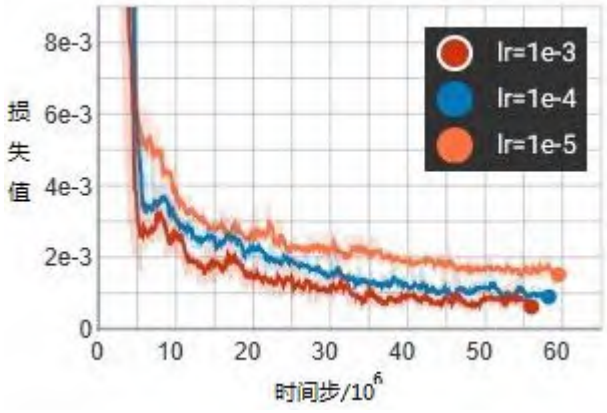

图2 损失函数收敛曲线

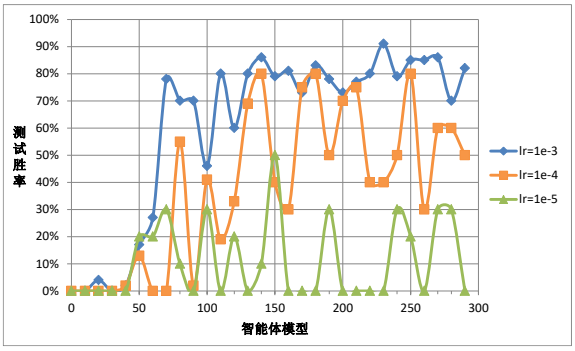

图3 决策智能体测试胜率收敛曲线

因此，本文中算法模型的训练超参数设置如表4所示。

表4 超参数设置

序号	超参数	取值
1	学习率	0.001
2	折扣因子	0.99
3	值函数系数	0.5
4	熵系数	0.02
5	截断系数	0.2
6	进程数	80
7	批数量	20

为了验证引入历史状态和动作信息以及无效动作掩膜对模型训练的影响，对比了只引入历史状态和动作信息、只引入无效动作掩膜情况下的回报收敛曲线，如图4所示。可以看出，去除历史状态和动作信息或者去除无效动作掩膜均无法获得高回报，智能体无法学习到有效的策略。以上结果表明，通过在PPO算法模型的基础上引入历史状态和动作信息以及无效动作掩膜可以引导智能体更容易学习到获胜策略，可以显著提高模型收敛效果。

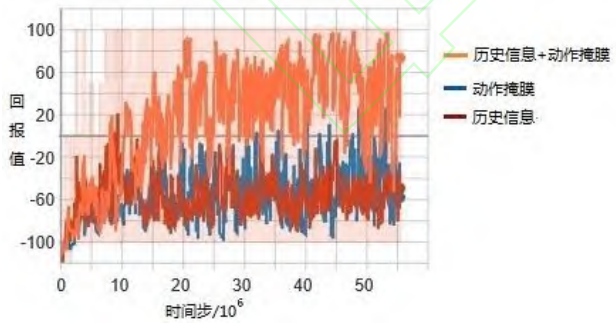

图4 游戏回报变化曲线对比

4 结束语

针对传统深度强化学习算法难以快速解决长时序复杂任务的问题，本文在经典PPO算法的基础上提出了一种引入历史信息和人类知识的深度强化学习方法。首先，在输入状态中引入历史状态和动作信息，让智能体可以显式获取历史信息，以帮助智能体更准确掌握对手状态和保持自身决策的一致性。其次，在策略模型中引入无效动作掩膜，避免智能体进行无效或不合理的探索，以提升探索效率。本文通过仿真试验验证了所提方法的有效性，对比试验表明所提方法可显著提升智能体的探索效率，可引导智能体学习到有效策略。

与其他深度强化学习模型一样，由于神经网络的高度拟合性导致智能体的行为决策机理难以解释，后续将对智能体行为的可解释性进行研究。

参考文献：

[1] Sutton R S, Barto A G. 强化学习[M].2.北京：电子工业出版社，2019：1-14.

[2] 伍元胜.面向动态拓扑网络的深度强化学习路由技术[J].电讯技术，2021，61（6）：659-665.

[3] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Human-level Control Through Deep Reinforcement Learning[J]. Nature, 2015,518(7540):529-533.

[4] 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.

[5] JADERBERG M, CZARNECKI W M., DUNNING I, et al. Human-level performance in 3D multiplayer games with population based reinforcement learning[J]. Science, 2019,364(6443):859–865.

[6] 陈仲铭，何明.深度强化学习原理与实践[M].北京：人民邮电出版社，2019：10-14.

[7] 冯超. 强化学习精要核心算法与TensorFlow实现[M].北京：电子工业出版社，2018：195-202.

[8] 肖智清. 强化学习原理与Python实现[M].北京：机械工业出版社，2019：139-170.

[9] SILVER D, LEVER G, HEESS N, et al. Deterministic Policy Gradient Algorithms[C]//Proceedings of 31st International Conference on Machine Learning. Beijing：JMLR，2014：387-395.

[10] SCHULMAN J, LEVINE S, MORITZ P, et al. Trust Region Policy Optimization[EB/OL].（2017-4-20）[2021-11-18]. https://arxiv.org/abs/1502.05477.

[11] SCHULMAN J, WOLSKI F, DHARIWAL P, et al. Proximal Policy Optimization Algorithms[EB/OL].（2017-4-28）[2021-11-18]. https://arxiv.org/abs/1707.06347v2.

[12] HUANG S Y, ONTAÑÓN S. A Closer Look at Invalid Action Masking in Policy Gradient Algorithms [EB/OL].（2020

-6-26）[2021-11-18]. https://arxiv.org/abs/2006.14171v2.
[13] HILL A, RAFFIN A, ERNESTUS M, et al. Stable baselines[EB/OL].(2018-1-1)[2021-11-18]. https://github.com/hill-a/stable-baselines.
[14] BROCKMAN G, CHEUNG V, PETTERSSON L, et al. OpenAI Gym[EB/OL].(2016-01-05)[2021-11-18].https://arxiv.org/ abs/1606.01540v1.

作者简介:

熊蓉玲 女,1988 年生于四川成都,2011 年获硕士学位,现为高级工程师,主要研究方向为深度强化学习和智能决策。

段春怡 女,1995 年生于陕西西安,2020 年获硕士学位,现为助理工程师,主要研究方向为深度强化学习。

冉华明 男,1990 年生于重庆万州,2015 年获硕士学位,现为工程师,主要研究方向为任务规划。

杨萌 女,1989 年生于四川成都,2019 年获博士学位,现为讲师,主要研究方向为深度强化学习。

冯旸赫 男,1985 年生于甘肃平凉,2013 年获博士学位,现为副教授,主要研究方向为强化学习和智能博弈。