

基于互信息优化的Option-Critic算法

栗军伟, 刘全, 徐亚鹏

引用本文

栗军伟, 刘全, 徐亚鹏. 基于互信息优化的Option-Critic算法[J]. 计算机科学, 2024, 51(2): 252-258.

LI Junwei, LIU Quan, XU Yapeng. Option-Critic Algorithm Based on Mutual Information Optimization[J]. Computer Science, 2024, 51(2): 252-258.

相似文章推荐 (请使用火狐或 IE 浏览器查看文章)

Similar articles recommended (Please use Firefox or IE to view the article)

面向缓存的动态协作任务迁移技术研究

Study on Cache-oriented Dynamic Collaborative Task Migration Technology

计算机科学, 2024, 51(2): 300-310. <https://doi.org/10.11896/jsjcx.230600128>

基于DQN的多智能体深度强化学习运动规划方法

DQN-based Multi-agent Motion Planning Method with Deep Reinforcement Learning

计算机科学, 2024, 51(2): 268-277. <https://doi.org/10.11896/jsjcx.230500113>

基于轨迹信息量的分层强化学习方法

Hierarchical Reinforcement Learning Method Based on Trajectory Information

计算机科学, 2023, 50(12): 314-321. <https://doi.org/10.11896/jsjcx.221100096>

基于深度强化学习的无线异构网络中继决策研究

Study on Relay Decision in Wireless Heterogeneous Networks Based on Deep Reinforcement Learning

计算机科学, 2023, 50(11A): 221000088-5. <https://doi.org/10.11896/jsjcx.221000088>

贝叶斯网络结构学习的CMIHC算法

CMIHC Algorithm for Bayesian Network Structure Learning

计算机科学, 2023, 50(11A): 220800046-7. <https://doi.org/10.11896/jsjcx.220800046>

基于互信息优化的 Option-Critic 算法

栗军伟¹ 刘全^{1,2,3,4} 徐亚鹏¹

1 苏州大学计算机科学与技术学院 江苏 苏州 215006

2 软件新技术与产业化协同创新中心 南京 210000

3 吉林大学符号计算与知识工程教育部重点实验室 长春 130012

4 苏州大学江苏省计算机信息处理技术重点实验室 江苏 苏州 215006

(20205227020@stu.suda.edu.cn)

摘要 时序抽象作为分层强化学习的重要研究内容,允许分层强化学习智能体在不同的时间尺度上学习策略,可以有效解决深度强化学习难以处理的稀疏奖励问题。如何端到端地学习到优秀的时序抽象策略一直是分层强化学习研究面临的挑战。Option-Critic(OC)框架在 Option 框架的基础上,通过策略梯度理论,可以有效解决此问题。然而,在策略学习过程中,OC 框架会出现 Option 内部策略动作分布变得十分相似的退化问题。该退化问题影响了 OC 框架的实验性能,导致 Option 的可解释性变差。为了解决上述问题,引入互信息知识作为内部奖励,并提出基于互信息优化的 Option-Critic 算法(Option-Critic Algorithm with Mutual Information Optimization, MIOOC)。MIOOC 算法结合了近端策略 Option-Critic(Proximal Policy Option-Critic, PPOC)算法,可以保证下层策略的多样性。为了验证算法的有效性,把 MIOOC 算法和几种常见的强化学习方法在连续实验环境中进行对比实验。实验结果表明,MIOOC 算法可以加快模型学习速度,实验性能更优,Option 内部策略更有区分度。

关键词: 深度强化学习; 时序抽象; 分层强化学习; 互信息; 内部奖励; Option 多样性

中图分类号 TP181

Option-Critic Algorithm Based on Mutual Information Optimization

LI Junwei¹, LIU Quan^{1,2,3,4} and XU Yapeng¹

1 School of Computer and Technology, Soochow University, Suzhou, Jiangsu 215006, China

2 Collaborative Innovation Center of Novel Software Technology and Industrialization, Nanjing 210000, China

3 Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun 130012, China

4 Provincial Key Laboratory for Computer Information Processing Technology, Soochow University, Suzhou, Jiangsu 215006, China

Abstract As an important research content of hierarchical reinforcement learning, temporal abstraction allows hierarchical reinforcement learning agents to learn policies at different time scales, which can effectively solve the sparse reward problem that is difficult to deal with in deep reinforcement learning. How to learn excellent temporal abstraction policy end-to-end is always a research challenge in hierarchical reinforcement learning. Based on the Option framework, Option-Critic can effectively solve the above problems through policy gradient theory. However, in the process of policy learning, the OC framework will have the degradation problem that the action distribution of the internal option policies becomes very similar. This degradation problem affects the experimental performance of the OC framework and leads to poor interpretability of the Option. In order to solve the above problems, mutual information knowledge is introduced as the internal reward, and an Option-Critic algorithm with mutual information optimization is proposed. The MIOOC algorithm combines the proximal policy Option-Critic algorithm to ensure the diversity of the lower level policies. In order to verify the effectiveness of the algorithm, the MIOOC algorithm is compared with several common reinforcement learning methods in continuous experimental environments. Experimental results show that the

到稿日期:2022-11-03 返修日期:2023-03-15

基金项目:国家自然科学基金(61772355, 61702055); 江苏省高等学校自然科学研究重大项目(18KJA520011, 17KJA520004); 吉林大学符号计算与知识工程教育部重点实验室资助项目(93K172014K04, 93K172017K18); 苏州市应用基础研究计划工业部分(SYG201422); 江苏省高校优势学科建设工程资助项目

This work was supported by the National Natural Science Foundation of China(61772355, 61702055), Jiangsu Province Natural Science Research University Major Projects(18KJA520011, 17KJA520004), Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University(93K172014K04, 93K172017K18), Suzhou Industrial Application of Basic Research Program Part(SYG201422) and Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions.

通信作者:刘全(quanliu@suda.edu.cn)

MIOOC algorithm can speed up the learning speed of the model, improve its experimental performance, and its Option internal strategy is more discriminative.

Keywords Deep reinforcement learning, Temporal abstract, Hierarchical reinforcement learning, Mutual information, Internal rewards, Diversity in option policies

1 引言

强化学习^[1] (Reinforcement Learning, RL) 是机器学习研究领域的一个重要分支, 主要以马尔可夫决策过程 (Markov Decision Process, MDP) 为理论基础, 是一种交互式学习方法。深度强化学习^[2-3] (Deep Reinforcement Learning, DRL) 把深度学习应用到强化学习中, 使得学习过程同时具备深度学习的感知能力和强化学习的决策能力。DRL 可以用来解决强化学习中具有高维状态空间和动作空间的任务。在深度 Q 网络^[4] (Deep Q-Network, DQN) 被提出之后, 出现了多种改进算法^[5-8], 从而提升了 DRL 智能体的探索能力和算法鲁棒性。然而, DRL 在面对稀疏奖励任务和具有分层结构的任务时, 问题的求解会变得很困难, 智能体很难学习到最优的策略。分层强化学习 (Hierarchical Reinforcement Learning, HRL) 的抽象机制可以用来解决上述问题。分层强化学习以半马尔可夫决策过程 (Semi-Markov Decision Process, SMDP) 为理论基础, 通过分层抽象机制提高智能体的探索能力和策略迁移能力。Kulkarni 等将深度学习与分层强化学习相结合, 提出了深度分层强化学习^[9] (Deep Hierarchical Reinforcement Learning, DHRL)。DHRL 形成了上层策略控制下层策略探索环境的问题求解方式, 实现了更强的特征提取能力和策略学习能力。近年来, DHRL 已经被成功应用于现实生活, 包括推荐系统^[10]、自动驾驶^[11] 和自然语言处理^[12] 等领域, 成为强化学习领域的研究热点之一。DHRL 的研究方向主要包括两种类型: (1) 基于子目标。这种方法利用特征提取把状态空间中部分状态位置作为子目标。其解决任务的过程为, 上层策略产生子目标, 下层策略负责指导智能体到达子目标, 重复此过程直到智能体到达最终目标位置。(2) 基于 Option 框架^[13]。这种方法完成任务的过程为, 上层策略在指定状态选出一个下层策略, 然后下层内部策略指导智能体到达终止状态, 重复此过程直至满足整个任务目标。Option 深度分层强化学习的研究热点主要集中于 Option 集合的多样性^[14], 以及 Option 内部策略的可重复利用性。本文提出的方法为基于 Option 内部策略的分层强化学习。

Option 框架作为分层控制机制的一种, 允许强化学习智能体使用长时序动作去表征、学习和规划^[15]。Bacon 等在 Option 框架基础上提出了 Option-Critic (OC) 框架^[16], 把策略梯度理论^[17] 应用到 Option 内部策略学习上, 从而可以通过梯度更新参数的方式学习 Option 内部策略、终止函数以及上层策略。然而, 在策略学习过程中, OC 框架中的 Option 内部策略集合会出现退化问题。例如, Option 集合的状态-动作分布变得非常相似, 从而阻止了智能体对环境的探索, 影响了实验性能。

针对以上问题, 本文提出了基于互信息优化的 Option-Critic 算法 (MIOOC)。MIOOC 算法为了保证 Option 集合的

多样性, 引入互信息知识作为内部奖励, 鼓励每个 Option 内部策略在每个状态下采取不同的动作。因此, MIOOC 算法可以提升智能体对环境的探索能力, 鼓励智能体学习多种方案完成任务, 在多种连续实验环境中取得了优异的效果。

本文的主要贡献包括以下 3 个方面:

(1) 提出了一种新的奖励函数优化目标。该最大化目标使用 Option 内部策略和动作的互信息作为内部奖励, 从而可以保证下层策略的多样性。

(2) 提出了 MIOOC 算法。实验分析结果表明, 在策略学习过程中, MIOOC 算法可以保证下层策略的区分度, 从而解决了 OC 框架的退化问题。

(3) 通过在多种连续实验环境中进行对比实验, 证明了 MIOOC 算法的实验效果优于对比算法。

本文第 2 章主要介绍分层强化学习的相关知识; 第 3 章介绍主要工作; 第 4 章通过实验证明 MIOOC 算法的优越性; 最后总结全文。

2 相关工作

本章主要介绍 MIOOC 算法的相关知识。其中, 2.1 小节阐述强化学习的相关内容; 2.2 小节阐述 OC 框架的相关内容。

2.1 强化学习

强化学习是智能体通过试错与环境进行交互的一种学习方式, 其利用环境的反馈信息改善策略。强化学习模型通常以马尔可夫决策过程为理论基础, 目标是最大化期望累计回报。马尔可夫决策过程通常被定义为一个元组 $(S, A, \gamma, r, \mathbf{P})$, 其中 S 是所有环境状态的集合, A 是所有动作的集合, $\gamma \in [0, 1)$ 是奖励函数的折扣因子。奖励函数 $r: S \times A \rightarrow R$ 是关于状态和动作的映射函数, 状态迁移矩阵 $\mathbf{P}$ 表示状态转移函数。在强化学习过程中, 假设智能体当前观察状态为 $s_t \in S$, 根据策略 π 选择动作 $a_t \in A$, 与环境交互获得奖赏 $r \in R$, 并转移到下一状态 s_{t+1} , 重复该过程直至到达终止状态。给定一个策略 π , 状态值函数 V 被定义为:

$$V_{\pi}(s) = E_{\pi} \left[\sum_{t=0}^{\infty} \gamma^t r_{t+1} \mid s_0 = s \right] \quad (1)$$

在状态 s 下采取动作 a 的值函数 Q 被定义为:

$$Q_{\pi}(s, a) = E_{\pi} \left[\sum_{t=0}^{\infty} \gamma^t r_{t+1} \mid s_0 = s, a_0 = a \right] \quad (2)$$

值函数满足如下贝尔曼等式:

$$V_{\pi}(s) = \sum_a \pi(a \mid s) (r(s, a) + \gamma \mathbf{P}(s' \mid s, a) V_{\pi}(s')) \quad (3)$$

强化学习的目标是找到一个最优策略 π^* , 使得期望折扣累计回报最大, 即 $V_{\pi^*}(s) = \max_{\pi} V_{\pi}(s)$ 。策略梯度理论提供了一种使用期望折扣累计回报优化随机策略 π_{θ} 的方法。期望折扣累计回报关于参数 θ 的梯度为:

$$\frac{\partial L(\theta)}{\partial \theta} = E \left[\frac{\partial \log \pi(A_t \mid S_t)}{\partial \theta} A^{\pi}(S_t, A_t) \right] \quad (4)$$

其中, $A^\pi(S_t, A_t) = Q_\pi(S_t, A_t) - V_\pi(S_t)$ 表示优势函数, S_t 和 A_t 分别为 t 时刻的状态和动作。

目前,深度强化学习融合了深度学习的感知能力和强化学习的决策能力,成为人工智能领域的研究热点。而且深度强化学习已经在很多方面得到应用。在星际争霸游戏中,DeepMind 团队的 Alpha Star 击败了人类选手;谷歌把深度强化学习用于 YouTube 视频推荐,并取得了优异的效果。

2.2 Option-Critic 算法

Option 框架是分层强化学习中分层机制的一种,Option 表示下层内部策略。一个 Option $o \in \Omega$ 由三元组 (I_o, π_o, β_o) 组成, $I_o \subseteq S$ 是下层策略的初始状态集合, π_o 表示 Option 内部策略,终止条件 $\beta_o: S \rightarrow [0, 1]$ 是当前 Option 策略在到达下一个状态 s' 后的终止概率。在 Option 框架下,假设智能体当前所在状态为 s ,在状态 s 中可以选择的 Option 集合为 $\Omega(s)$ 。Option 框架的执行过程为:上层策略 $\pi_\Omega(s)$ 以 ϵ -贪婪策略从集合 $\Omega(s)$ 中选择一个 Option 内部策略 o ,智能体按照内部策略 π_o 执行动作,直到满足终止条件 β_o 。然后上层策略再选择一个 Option 内部策略 o' 指导智能体执行动作。重复执行上述过程,直至完成整个任务。

OC 框架提出了一种学习 Option 内部策略和终止函数参数的梯度更新方法,该方法假定所有 Option 内部策略在每个状态都可以被选择。OC 框架使用 $\pi_{o,\theta}$ 表示参数化的 Option 内部策略,使用 $\beta_{o,\vartheta}$ 表示参数化的 Option 内部策略终止函数。其中, θ 和 ϑ 分别为 π_o 和 β_o 的参数,可以根据策略梯度定理进行优化。在 OC 框架中,Option 值函数被定义为:

$$Q_\Omega(s, o) = \sum_a \pi_{o,\theta}(a|s) Q_U(s, o, a) \quad (5)$$

其中, $Q_U: S \times \Omega \times A \rightarrow R$ 指在状态 s 根据 o 选择动作 a 的值函数。可以表示为:

$$Q_U(s, o, a) = r(s, a) + \gamma P(s'|s, a) U(o, s') \quad (6)$$

其中, $U: \Omega \times S \rightarrow R$ 反映的是智能体选择下层策略 o 后,到达下一状态 s' 的价值函数。

$$U(o, s') = (1 - \beta_{o,\vartheta}(s')) Q_\Omega(s', o) + \beta_{o,\vartheta}(s') V_\Omega(s') \quad (7)$$

给定初始条件 (s_0, o_0) ,则期望折扣累计回报关于参数 θ 的梯度可以表示为:

$$\sum_{s,o} \mu_\Omega(s, o | s_0, o_0) \sum_a \frac{\partial \pi_{o,\theta}(a|s)}{\partial \theta} Q_U(s, o, a) \quad (8)$$

其中, $\mu_\Omega(s, o | s_0, o_0)$ 表示从 (s_0, o_0) 到 (s, o) 轨迹的折扣权重。

给定初始条件 (s_1, o_0) ,则期望折扣累计回报关于参数 ϑ 的梯度为:

$$-\sum_{s',o} \mu_\Omega(s', o | s_1, o_0) \frac{\partial \beta_{o,\vartheta}(s')}{\partial \vartheta} A_\Omega(s', o) \quad (9)$$

其中, $\mu_\Omega(s', o | s_1, o_0)$ 表示从 (s_1, o_0) 到 (s', o) 轨迹的折扣权重。

3 基于互信息优化的 Option-Critic 算法

一个好的奖励函数可以鼓励智能体选择更优的动作完成任务。本章设计了一种内部驱动,把互信息知识引入奖励函数中,提出了一种新的最大化优化目标,以保证 Option 集合的多样性。有很多研究使用状态信息^[18-20]区分 Option 内部策略。但是,这些方法大多需要预训练,从而增加了时间复杂度。相反,本文提出的 MIOOC 算法并不需要预训练,而是

直接通过端到端的方式训练神经网络,自动地学习策略。MIOOC 算法通过鼓励下层策略动作的多样性来实现 Option 集合的多样性。当环境的状态信息不够完善时,用动作区分 Option 内部策略是非常合理的。这种方法也允许智能体学习多种方案来解决问题,因为在相同的状态下,不同的 Option 内部策略可以采取不同的动作完成任务。多样化的下层 Option 内部策略可以鼓励智能体更好地探索环境,从而提升实验性能。

3.1 目标构造

本文使用信息论的知识构造训练目标: S 和 A 分别表示状态和动作; $I(\cdot; \cdot)$ 和 $H[\cdot]$ 分别是互信息和香农熵,底数均为 e 。为了实现用不同的动作分布区分 Option 内部策略,在给定状态下,本文最大化 Option 和动作之间的互信息 $I(A; O|S)$ 。为此,可以构造如下内部奖励函数 R_{mi} :

$$R_{mi} = I(A; O|S) = H(A|S) - H(A|S, O) \quad (10)$$

式(10)中第一项 $H(A|S)$ 表示在当前状态 S 下可采取动作的随机性;第二项 $H(A|S, O)$ 表示在当前状态下智能体根据当前 Option 内部策略可采取动作的随机性。最大化 R_{mi} 即最大化 $H(A|S)$,最小化 $H(A|S, O)$ 。最大化 $H(A|S)$ 项保证了可以在当前状态执行动作的多样性,最小化 $H(A|S, O)$ 项鼓励了下层内部策略的专一性。故最大化上述优化目标能够保证每个 Option 内部策略在相同的状态采取不同的动作,实现 Option 集合的多样性。把内部奖励函数引入标准强化学习奖励函数中,即得到如下增广奖励函数:

$$R_{aug}(S_t, A_t) = (1 - \tau)R(S_t, A_t) + \tau R_{mi}(S_t) \quad (11)$$

其中, τ 是一个超参数,用于控制内部奖励在增广奖励函数中的重要程度。其值越大代表内部奖励在增广奖励函数中的比例越大,越能鼓励下层内部策略的多样性。当 $\tau \rightarrow 0$ 时,增广奖励函数趋近于标准强化学习奖励函数。

3.2 MIOOC 算法

受信息论的启发,本文为了阻止 OC 框架在训练过程中 Option 集合变得相似的情况发生,提出了一种互信息优化算法,即 MIOOC 算法。该算法框架的结构图如图 1 所示。首先根据上层策略 π_Ω 在当前状态选择一个下层 Option 内部策略 o ;之后智能体根据下层 Option 内部策略采取动作并转移至下一状态;然后根据环境给予的赏罚计算得到增广奖励函数;最后根据增广奖励函数通过梯度更新的方式更新神经网络。重复上述过程,直至满足优化条件。

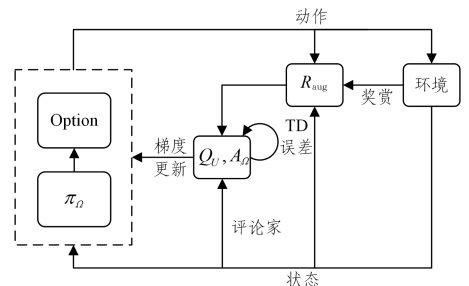


图 1 互信息优化的 Option-Critic 框架

Fig. 1 Option-Critic framework with mutual information optimization

MIOOC 算法在标准强化学习奖励函数中引入互信息

内部奖励函数,其具体实现步骤如算法 1 所示。

算法 1 MIOOC

输入:超参数 τ ,折扣率 γ ,Option 内部策略学习率 α_θ ,终止函数学习率

- α_θ
- 输出:网络参数 π_Ω, π_o 和 β_o
1. 初始化: π_Ω, π_o 和 β_o

2. 根据上层策略 $\pi_\Omega(o|s_0)$ 选择一个 o

3. Repeat:

4. 根据策略 $\pi_{o,\theta}(a|s)$ 选择一个动作 a

5. 执行动作 a ,观测状态 s' ,得到环境奖赏 r 和内部奖励 r_{mi}

6. 则 $r'=(1-\tau)r+\tau r_{mi}(s)$

7. If o 在状态 s' 终止 then

8. 根据 $\pi_\Omega(\cdot|s')$ 选择一个新的 o'

9. Else then

10. $o'=o$

11. End If

12. $\delta \leftarrow r' - Q_U(s,o,a)$

13. $\delta \leftarrow \delta + \gamma(1-\beta_o(s'))Q_\Omega(s',o') + \gamma\beta_o(s')\sum_{o'}Q_\Omega(s',o')$

14. $Q_U(s,o,a) \leftarrow Q_U(s,o,a) + \alpha\delta$

15. $\theta \leftarrow \theta + \alpha_\theta \frac{\partial \log \pi_{o,\theta}(a|s)}{\partial \theta} Q_U(s,o,a)$

16. $\vartheta \leftarrow \vartheta - \alpha_\vartheta \frac{\partial \beta_{o,\vartheta}(s')}{\partial \vartheta} (Q_\Omega(s',o) - V_\Omega(s'))$

17. $s \leftarrow s'$

18. Until s' 是终止状态
- 算法 1 介绍了 MIOOC 算法的完整训练过程,整体采用表格格式内部 Option Q-learning 更新方式。其中,第 2 行根据上层策略 π_Ω 选择一个下层内部策略 o ;第 3—14 行是策略评估过程;第 15—16 行是策略优化过程,即对各参数的更新优化。
- 4 实验
- 为了评估 MIOOC 算法的性能,本文在 4 个连续环境和 1 个视觉导航任务上进行实验。本文选择 PPO(Proximal Policy Optimization)算法^[21]作为底层算法,因为 OC 框架已经在 PPO 算法上得到了扩展,即 PPOC 算法^[22]。为了保证与对比算法的一致性,所有实验均使用 2 个 Option 内部策略进行实验。
- 4.1 Mujoco 实验环境
- OpenAI Gym^[23]是强化学习中常用的实验仿真平台,其中包含很多常用的环境测试接口。例如 Box2D, Atari 游戏环境以及 Mujoco 物理引擎等。
- 为了有效评估所提算法的性能,本文选取 Mujoco 物理引擎中的 4 个连续控制任务 HalfCheetah-v2, Hopper-v2, Walker2d-v2 和 Ant-v2 作为实验环境,如图 2 所示。其中, Hopper-v2 实验环境的任务目标是训练一个单腿智能体在不摔倒的前提下尽可能快地向前跳; HalfCheetah-v2 实验环境的任务目标是训练一个双腿豹形智能体在不摔倒的前提下尽可能快地向前跑; Walker2d-v2 实验环境的任务目标是训练一个二维双足智能体在不摔倒的前提下尽可能快地向前走; Ant-v2 实验环境的任务目标是训练一个四足智能体尽可能快地向前爬。
-
- 图 2 Mujoco 实验环境
Fig. 2 Mujoco experiment environments
- 4 个实验环境的状态空间及动作空间维度的具体介绍如表 1 所列。其中, Ant-v2 实验环境的状态空间和动作空间最复杂。
- 表 1 Mujoco 实验任务介绍
Table 1 Mujoco experimental task introduction
- | 环境 | 动作维度 | 状态维度 |
|----------------|------|------|
| Hopper-v2 | 3 | 11 |
| HalfCheetah-v2 | 6 | 17 |
| Walker2d-v2 | 6 | 17 |
| Ant-v2 | 8 | 111 |
- 4.1.1 实验设置
- 为了验证所提出的方法与 OC 框架结合后能取得更好的实验性能,本文选取 PPOC 算法作为基础算法,把互信息内部奖励引入 PPOC 算法中。MIOOC 算法使用 2 个独立的神经网络,其中,终止函数使用一个神经网络。上层策略 π_Ω 和内部策略 $\pi_o(a|s)$ 共享一个神经网络。每个神经网络包含 2 个隐层,每个隐层都包含 64 个神经元。终止函数使用 sigmoid 函数作为激活函数,而上层策略使用 softmax 函数作为激活函数。对于每个 Mujoco 实验环境,每种算法都选取 5 个不同的随机种子进行 5 次实验,并记录每一次迭代(大约 2000 步)时的回报值。每条实验曲线由这 5 次运行结果的平均值绘制而成。实验超参数如表 2 所列。
- 表 2 Mujoco 实验环境超参数
Table 2 Mujoco experimental environment hyper-parameters
- | Hyper-parameters | Hopper-v2 | HalfCheetah-v2 | Walker2d-v2 | Ant-v2 |
|--------------------|--------------------|--------------------|--------------------|--------------------|
| τ | 0.01 | 0.001 | 0.01 | 0.01 |
| α_θ | 5×10^{-7} | 5×10^{-7} | 5×10^{-7} | 5×10^{-7} |
| γ | 0.99 | 0.99 | 0.99 | 0.99 |
| α_ϑ | 1×10^{-4} | 1×10^{-4} | 1×10^{-4} | 3×10^{-4} |
- 其中,参数 τ 控制内部奖励在增广奖励函数中的相对重要程度;参数 α_ϑ 和 α_θ 分别是终止函数和下层内部策略的学习率; γ 是奖励函数的折扣因子。因为不同的任务有不同的

任务目标,以及不同任务中智能体有不同的动作空间和状态空间,所以 4 种实验环境使用的超参数不全相同。所有实验结果都取自多组超参数实验中的最优值,部分消融实验将在 4.3 节介绍。

4.1.2 实验结果分析

MIOOC 算法与 OC 算法和 PPO 算法的性能对比结果

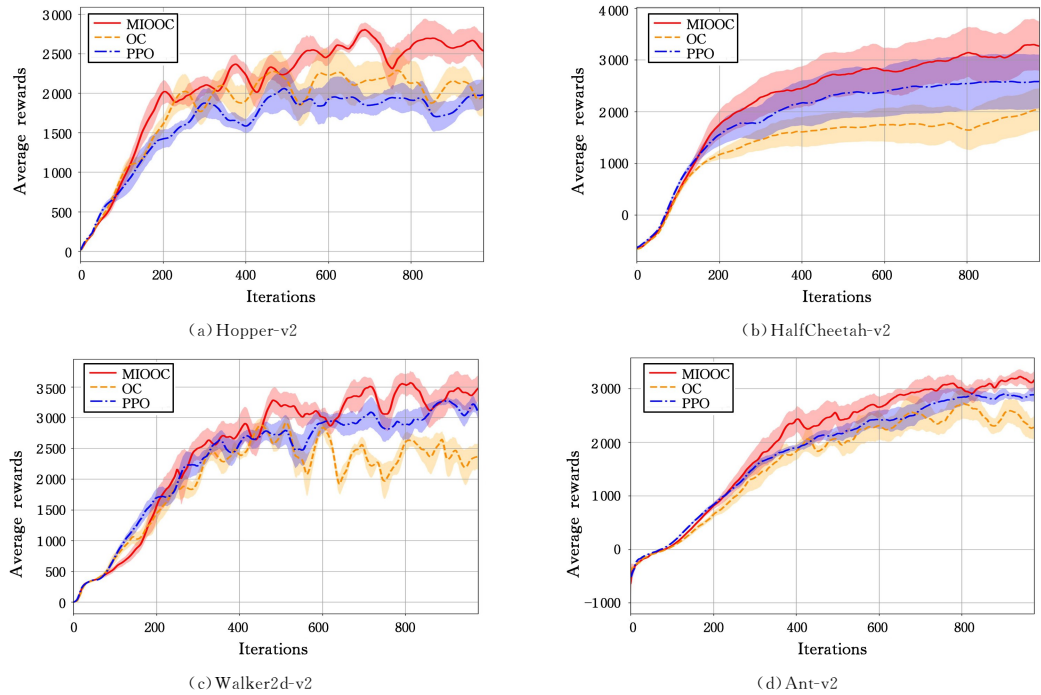


图 3 Mujoco 环境中实验性能对比

Fig. 3 Comparison of experimental performance in Mujoco environments

从图 3 中可以看出,OC 算法很快收敛到局部最优情况,MIOOC 算法的实验性能持续提升。这是因为 OC 框架在策略学习过程中,其 Option 策略集合变得非常相似,导致智能体对环境的探索能力较低。而 MIOOC 算法引入互信息作为内部奖励驱动智能体采取不同的动作,保证了 Option 集合的多样性,从而具有更强的环境探索能力,实验性能优于对比算法。

下层 Option 内部策略采用高斯分布的形式进行表示,实现了状态到高斯分布参数对 (μ, σ) 的映射。为了验证 Option 内部策略的多样性,以实验任务 HalfCheetah-v2 为例,把同一个环境状态 s 作为高斯分布的输入。在输出的高斯分布参数向量对 (μ, σ) 中,选取同一位置的信息绘制二维平面高斯分布,如图 4 所示。图 4(a)和图 4(b)分别表示 OC 算法中两个 Option 内部策略的动作分布图。图 4(c)和图 4(d)分别表示 MIOOC 算法中两个 Option 内部策略的动作分布图。其中,图 4(a)图 4 和 (c)是在 200 个迭代步数时相应维度的高斯动作分布;图 4(b)和图 4(d)是在训练结束时相应维度的高斯动作分布。从图 4 可以看出,在策略学习过程中,OC 算法中 2 个 Option 内部策略的区分度越来越低,甚至有重合。而 MIOOC 算法中的 2 个 Option 内部策略在训练过程中始终保持一定程度的差异性,从而增强了智能体对环境的探索能力以及获取优势动作的可能性。

学习曲线如图 3 所示。其中,横坐标表示迭代数,一个迭代过程执行 2048 个时间步数。纵坐标表示算法执行 5 次之后的期望回报的平均值。MIOOC 算法相较于其他对比算法的优势是鼓励了 Option 集合的多样性,从而保证了智能体具有更强的探索能力。在所有实验中,MIOOC 算法相比对比算法均有不同程度的性能提升,实验效果如图 3 所示。

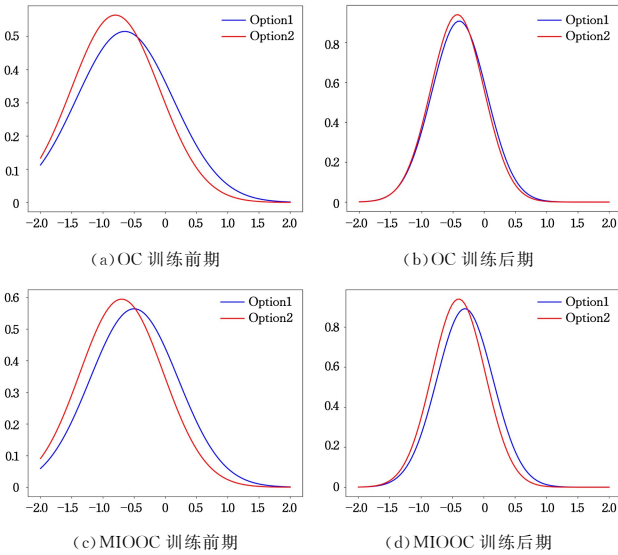


图 4 高斯分布对比

Fig. 4 Comparison of Gaussian distribution

4.2 OneRoom 实验环境

本文仍然使用 MiniWorld 平台中更加复杂的 3D 视觉导航环境——OneRoom 任务。在 OneRoom 任务中,智能体需要在一个封闭的房间内到达一个随机产生的红色盒子位置。这需要智能体环顾整个房间,发现红色盒子并到达盒子位置,实验环境如图 5 所示。OneRoom 房间的观测空间是一个

3 通道 RGB 图片,动作空间包含 8 个离散行为。每个情节开始时,红色盒子的位置是随机的。如果智能体到达红色盒子位置或者 180 个时间步被执行完时,一个情节终止。实验使用深度卷积神经网络(Convolutional Neural Network,CNN)作为函数逼近器,CNN 结构与 DQN 保持一致,其他网络结构与 PPOC 算法一致。

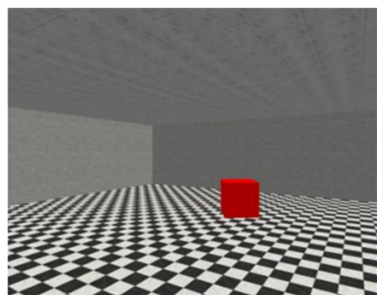


图 5 OneRoom 实验环境(电子版为彩图)

Fig. 5 OneRoom experiment environment

实验性能如图 6 所示,每条曲线都为 5 次独立运行结果的平均值。从图中可以看出,OC 算法的性能是最低的,PPO 算法性能优于 OC 算法,MIOOC 算法的性能略优于 PPO 算法。实验结果表明,引入互信息作为内部奖励,可以鼓励智能体更好地探索环境,从而能够学习到更好的 Option 内部策略,指导智能体快速到达红色盒子位置。

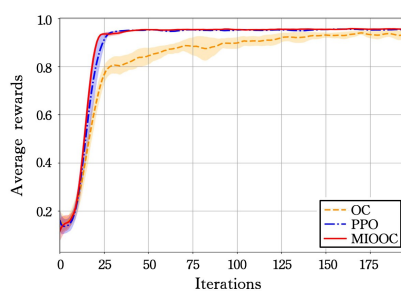


图 6 OneRoom 环境中实验性能对比

Fig. 6 Performance comparison in OneRoom environment

4.3 消融实验

本小节主要讨论 4.1 小节中不同超参数对实验性能的影响,实验环境为 Mujoco 平台中的 HalfCheetah-v2 任务。因强化学习中普遍使用的奖励函数折扣因子 γ 为 0.99,所以我们没有对参数 γ 进行消融实验。实验主要对参数 τ , α_θ 和 α_ϕ 进行讨论,实验效果如图 7 所示,每条曲线为 5 次独立运行后的回报均值。其中, τ 控制内部奖励在增广奖励函数中的重要程度,值越大越能鼓励下层内部策略的多样性; α_θ 和 α_ϕ 分别是终止函数和下层内部策略的学习步长,用于控制每一次策略更新的程度。从图中可以看出不同的超参数对实验性能的影响是不同的。当 τ 为 0.001、 α_θ 为 5×10^{-7} 、 α_ϕ 为 1×10^{-4} 时,实验性能最优。

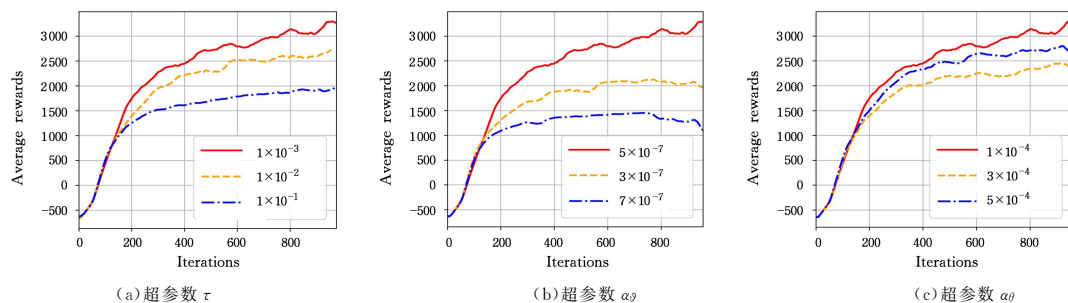


图 7 消融实验

Fig. 7 Ablation experiments

结束语 传统的 OC 算法无法保证在策略学习过程中 Option 集合的多样性,从而影响了下层策略对环境的探索能力,进而影响了最终的实验性能。因此,本文把互信息 $I(A; O|S)$ 作为内部奖励引入标准强化学习框架中,鼓励不同的 Option 内部策略在相同的状态采取不同的动作,提高智能体对环境的探索能力。为了验证所提算法的性能优势,分别在 Mujoco 平台上的 4 个连续环境和 MiniWorld 平台上的 1 个视觉导航环境中进行对比实验。实验结果表明本文提出的 MIOOC 算法具有更优秀的实验性能。

在 MIOOC 算法中,互信息内部奖励可以保证 Option 策略之间的区分度。然而,OC 框架会出现另外一种退化问题。在策略学习过程中,一个 Option 内部策略被上层策略频繁选择而变得越来越好,其他 Option 内部策略得不到及时更新而变得越来越差。这种马太效应使得互信息内部奖励没有充分发挥优势。下一步的研究重点考虑在 OC 框架中引入长度约束,使得 Option 内部策略执行的时间步数不会太多,也不会太少。

参考文献

- [1] SUTTON R S, BARTO A G. Reinforcement learning: An introduction[M]. MIT Press, 1998.
- [2] LIU Q, ZHAI J W, ZHANG Z Z, et al. A survey on deep reinforcement learning[J]. Chinese Journal of Computers, 2018, 41(1): 1-27.
- [3] LIU J W, GAO F, LUO X L. Survey of deep reinforcement learning based on value function and policy gradient[J]. Chinese Journal of Computers, 2019, 42(6): 1406-1438.
- [4] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Human-level control through deep reinforcement learning[J]. Nature, 2015, 518(7540): 529-533.
- [5] LILLICRAP T P, HUNT J J, PRITZEL A, et al. Continuous control with deep reinforcement learning[J]. arXiv: 1509.02971, 2015.
- [6] FUJIMOTO S, HOOF H, MEGER D. Addressing function approximation error in actor-critic methods[C] // Proceedings of

the International Conference on Machine Learning. 2018;1587-1596.

[7] SCHULMAN J,LEVINE S,ABBEEL P,et al. Trust region policy optimization[C]// Proceedings of the International Conference on Machine Learning. 2015;1889-1897.

[8] HAARNOJA T,ZHOU A,ABBEEL P,et al. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor[C]// Proceedings of the International Conference on Machine Learning. 2018;1861-1870.

[9] KULKARNI T D,NARASIMHAN K,SAEEDI A,et al. Hierarchical deep reinforcement learning: Integrating temporal abstraction and intrinsic motivation[C]// Advances in Neural Information Processing Systems. 2016;3675-3683.

[10] ZHAO D,ZHANG L,ZHANG B,et al. Mahrl: Multi-goals abstraction based deep hierarchical reinforcement learning for recommendations[C]// Proceedings of International ACM SIGIR Conference on Research and Development in Information Retrieval. 2020;871-880.

[11] DUAN J,LI S E,GUAN Y,et al. Hierarchical reinforcement learning for self-driving decision-making without reliance on labelled driving data [J]. IET Intelligent Transport Systems, 2020,14(5):297-305.

[12] LIU J,PAN F,LUO L. Gochat:Goal-oriented chatbots with hierarchical reinforcement learning[C]//Proceedings of International ACM SIGIR Conference on Research and Development in Information Retrieval. 2020;1793 -1796.

[13] SUTTON R S,PRECUP D,SINGH S. Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning[J]. Artificial Intelligence,1999,112(1/2):181-211.

[14] LIU C H,ZHU F,LIU Q. Option-Critic Algorithm Based on Sub-Goal Quantity Optimization[J]. Chinese Journal of Computers,2021,44(9):1922-1933.

[15] HUANG Z G,LIU Q,ZHANG L H,et al. Research and Development on Deep Hierarchical Reinforcement Learning[J]. Journal of Software,2023,34(2):733-760.

[16] BACON P L, HARB J,PRECUP D. The option-critic architecture[C]//Proceedings of the AAAI Conference on Artificial In-

telligence. 2017.

[17] SUTTON R S,MCALLESTER D A,SINGH SP,et al. Policy gradient methods for reinforcement learning with function approximation[C]//Proceedings of the Advances in Neural Information Processing Systems. 2000;1057-1063.

[18] EYSENBACH B,GUPTA A,IBARZ J,et al. Diversity is all you need; Learning skills without a reward function[J]. arXiv:1802.06070,2018.

[19] BAUMLI K,WARDE F D,HANSEN S,et al. Relative variational intrinsic control[C]//Proceeding of the AAAI Conference on Artificial Intelligence. 2021;6732-6740.

[20] ZHANG J, YU H, XU W. Hierarchical reinforcement learning by discovering intrinsic options[C]//Proceeding of the International Conference on Learning Representations. 2021.

[21] SCHULMAN J, WOLSKI F, DHARIWAL P, et al. Proximal policy optimization algorithms[J]. arXiv:1707.06347,2017.

[22] KLISSAROV M,BACON P L,HARB J,et al. Learnings options end-to-end for continuous action tasks[J]. arXiv:1712.00004,2017.

[23] BROCKMAN G,CHEUNG V,PETTERSSON L,et at. Openai gym[J]. arXiv:1606.01540,2016.



LI Junwei, born in 1998, postgraduate. His main research interests include reinforcement learning and hierarchical reinforcement learning.



LIU Quan, born in 1969, Ph.D, professor, Ph.D supervisor, is a member of CCF (No. 15231S). His main research interests include deep reinforcement learning and automated reasoning.

(责任编辑:何杨)