第 40 卷 第 5 期
2025 年 10 月

天津科技大学学报
Journal of Tianjin University of Science and Technology

Vol. 40 No. 5
Oct. 2025

DOI:10.13364/j.issn.1672-6510.20240002
网络首发日期：2024–09–30；网络首发地址：http://link.cnki.net/urlid/12.1355.N.20240930.1252.005

基于有效动作表示的策略搜索强化学习方法

王馨雪，黄佳欣，赵婷婷，陈亚瑞，王　媛
（天津科技大学人工智能学院，天津 300457）

摘　要：策略搜索强化学习方法是深度强化学习领域的一种高效学习范式，但存在模型结构复杂、训练周期长、泛化能力差的问题。表示学习能在一定程度上缓解上述问题，但传统的表示学习方法的动作表示包含大量冗余或不相关的信息，缺乏可解释性，影响系统的性能和泛化能力。本文提出了一种基于有效动作表示的策略搜索强化学习方法 TAR-PPO（task-relevant action representation learning based PPO）。使用 β-VAE 作为学习动作表示的组件，引入回报预测模型辅助有效动作表示提取器的训练，帮助有效动作表示提取器提取到与任务相关的、更加有效的动作信息，增强了动作表示的可解释性，提高模型的性能和泛化能力。在 MountainCar-v0 环境中的对比实验结果表明，本文方法能够有效捕获与任务相关的动作信息，有利于动作空间的进一步探索，提升了策略学习性能。最后，通过消融实验验证了本文方法的显著优势。

关键词：潜在空间；动作表示；连续动作空间；回报预测；有效动作表示提取器；策略搜索强化学习方法
中图分类号：TP391　　**文献标志码**：A　　**文章编号**：1672-6510（2025）05-0057-09

Strategy Search Reinforcement Learning Method Based on Effective Action Representation

WANG Xinxue，HUANG Jiaxin，ZHAO Tingting，CHEN Yarui，WANG Yuan
（College of Artificial Intelligence，Tianjin University of Science and Technology，Tianjin 300457，China）

Abstract：The strategy search reinforcement learning method is an efficient learning paradigm in the field of deep reinforcement learning，but it has problems such as complex model structure，long training cycle，and poor generalization ability. Representation learning can alleviate the above problems to a certain extent，but traditional representation learning methods contain a large amount of redundant or irrelevant information in their action representations，lacking interpretability，which affects the performance and generalization ability of the system. This article proposes a strategy search reinforcement learning method，namely task-relevant action representation learning based PPO（TAR-PPO）based on effective action representation. In this method，with the use of β-VAE as a component for learning action representation，a reward prediction model is introduced to assist in the training of effective action representation extractors，helping them extract more effective action information related to the task，enhancing the interpretability of action representation，and improving the performance and generalization ability of the model. Through comparative experiments in MountainCar-v0 environment，the results showed that our method could effectively capture task related action information，which is conducive to further exploration of action space and improves the learning performance of the strategy. Finally，the significant advantages of our method were validated through ablation experiments.

Key words：latent space；action representation；continuous action space；reward prediction；effective action representation extractor；strategy search reinforcement learning method

收稿日期：2024–01–04；**修回日期**：2024–05–07
基金项目：国家自然科学基金资助项目（61976156）
作者简介：王馨雪（1999—），女，山东临沂人，硕士研究生；**通信作者**：赵婷婷，教授，tingting@tust.edu.cn

引文格式：
王馨雪，黄佳欣，赵婷婷，等. 基于有效动作表示的策略搜索强化学习方法[J]. 天津科技大学学报，2025，40(5)：57–65.
WANG X X，HUANG J X，ZHAO T T，et al. Strategy search reinforcement learning method based on effective action representation[J]. Journal of Tianjin university of science and technology，2025，40(5)：57–65.

强化学习是机器学习领域中的一种重要方法，借鉴了人类学习中的试错机制。与监督学习中的指导性反馈有所不同，强化学习是以评价性反馈为基础进行决策优化。在强化学习中，智能体(Agent)在环境状态(s)下选择并执行动作(a)，导致环境转移到新状态(s')，并通过奖励信号(r)反馈给智能体。智能体根据奖励信号选择后续动作，最终找到适合当前状态的最优动作选择策略(policy)，以实现整个决策过程的最大累积奖励(reward)。强化学习的基本框架如图 1 所示。

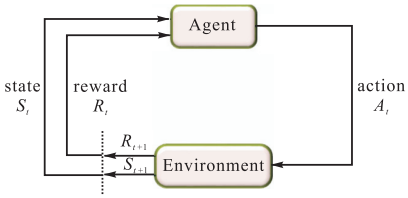

图 1　强化学习基本框架

Fig. 1　Basic framework of reinforcement learning

随着深度学习的发展，研究者将深度学习(DL)和强化学习(RL)相结合，提出了深度强化学习(DRL)算法[1]。DRL 算法可广泛应用到导航[2]、游戏[3-5]、机器人控制[6-9]等多个领域[7-14]，突破了传统强化学习的瓶颈，其强大的学习能力使其受到了极大的关注[15]。

在深度学习的加持下，深度强化学习具备了强大的感知能力，相比传统强化学习的性能有了质的飞跃，在不需要人工干预的情况下，可以直接输出动作。在深度强化学习中，为了得到具有丰富表达能力的模型，往往需要大量的训练样本和训练时间[16]。现有深度强化学习方法通常是面向特定的问题，因此泛化问题也是深度强化学习所面临的一大挑战。

近年来，为解决强化学习中的样本利用率问题和泛化能力问题，表示学习被引入 DRL 中。深度强化学习中表示学习的研究主要集中在状态表示学习。通过优化状态表示，强化学习算法能够更有效地处理复杂的环境，并在面对未知情况时更好地泛化和适应新环境。经典的状态表示学习方法有嵌入控制方法(E2C)和世界模型(world models)。嵌入控制方法[17]是一种解决原始图像输入维度过高问题的方法。世界模型[18]是基于神经网络生成模型的通用强化学习环境构建方法，能够在无监督情况下迅速学习低维潜在空间下的环境状态表示，并在学到的世界模型中训练智能体，再将其策略迁移至真实环境中。上述模型中普遍采用变分自编码器对状态进行压缩，以学习状态在低维潜在空间中的表示，这有助于提升智能体对状态的理解，从而提高学习效率。由于传统的状态表示可能包含大量冗余或不相关的信息，相关研究引入了与任务相关的状态表示。任务相关的状态表示通过专注于与任务目标直接相关的特征，进而能够更有效地捕捉与任务相关的信息，在降低维度的同时保留关键特征，有助于降低学习的复杂度和提高泛化能力，使其更好地适应多样化的任务和环境。

与状态表示类似，动作表示学习将原始动作空间随机映射到潜在特征空间[19]，学习原始动作空间中的底层结构特征，促进学习速度的提高。另外，在强化学习中动作表示的学习直接影响系统的泛化性能。通过有效学习和表达动作空间的特征，智能体可以更好地适应不同任务和环境，提高在未知领域中的性能。良好的动作表示在包含原始重要信息的基础上，应尽可能多地涵盖与任务相关的特征，并尽可能减少与任务无关的特征。然而，过去的动作表示学习仅将低维动作空间随机映射到特征空间，导致学到的动作表示缺乏可解释性，与任务关联性不强。

因此，针对复杂决策任务和大规模连续动作空间，本研究提出了一种基于有效动作表示的策略搜索强化学习方法 TAR-PPO(task-relevant action representation learning based PPO)。将回报预测模型作为辅助任务，通过有效动作表示提取器对原始动作表示中与任务相关的特征进行提取，从而得到与任务相关的、更加有效的动作表示，提高模型的学习效率和泛化能力。最终，通过 MountainCar-v0 环境进行实验，验证本文方法的有效性。

1　基本理论

1.1　强化学习建模

强化学习通过智能体与环境的不断交互学习，这个过程由马尔可夫决策过程(Markov decision process，

MDP）进行建模。智能体在状态 S 中行动，根据策略学习，从动作空间 A 中选择要做的动作，决策过程可以使用 $M=(S,A,\boldsymbol{P},R)$ 表示，其中 $\boldsymbol{P}$ 是状态转移概率矩阵，$P_{ss'}^{a}$ 代表智能体在做出动作 a 后状态由 s 转为 s' 的概率，R 代表回报函数，R_a^s 代表智能体在状态 s 的背景下做出动作 a 后得到的即时回报。

寻找能够使算法获得最大累积回报的最优策略 π^* 是强化学习的核心目标[20]。用 $\pi(a|s,\theta)$ 表示策略函数，其中 θ 为策略函数的参数。将智能体与环境交互，交互过程由路径表示，具体为 $h^n:=[s_1^n,a_1^n,\ldots,s_T^n,a_T^n]$，其中 T 代表这条路径的长度，该条路径的累积奖励 $R(h)$ 表示为

$$R(h)=\sum_{t=1}^{T}\gamma^{t-1}r(s_t,a_t,s_{t+1}) \tag{1}$$

式（1）中的 $\gamma\in(0,1]$ 代表奖励折扣因子。路径发生概率表示为

$$P(h|\theta)=P(s_1)\prod_{t=1}^{T}P(s_{t+1}|s_t,a_t)\pi(a_t|s_t,\theta) \tag{2}$$

累积回报的期望表示为

$$J(\theta)=\int P(h|\theta)R(h)\mathrm{d}h \tag{3}$$

最大化期望累积回报 $J(\theta)$ 的参数即是最优策略参数 θ^*，为

$$\theta^*:=\arg\max_{\theta}J(\theta) \tag{4}$$

在深度强化学习领域，近端策略优化（proximal policy optimization，PPO）[21]算法被广泛认为是策略梯度算法中的典范[22]。2017 年，OpenAI 团队基于 AC（actor-critic）算法框架推出了 PPO，这一近端策略优化算法迅速成为深度强化学习的主流算法，特别是在处理连续状态和动作空间的复杂机器学习任务时表现卓越。本研究提出的算法框架融合了 PPO 算法，是一种新的基于潜在空间的策略学习方法。目标函数为

$$L_t^{\mathrm{CLIP+VF+S}}(\theta)=\hat{E}_t\left[L_t^{\mathrm{CLIP}}(\theta)-c_1L_t^{\mathrm{VF}}(\theta)+c_2S[\pi_\theta](s_t)\right] \tag{5}$$

式中：$L_t^{\mathrm{CLIP}}(\theta)$ 代表智能体策略的目标函数；$S[\pi_\theta](S_t)$ 是熵项，具有提高策略探索性的作用；$L_t^{\mathrm{VF}}(\theta)$ 代表状态值函数的均方误差 $\left(V_\theta(S_t)-V_t^{\mathrm{target}}\right)^2$。

$$L^{\mathrm{CLIP}}(\theta)=\hat{E}_t\left[\min\left(r_t(\theta)\hat{A}_t,\mathrm{clip}\left(r_t(\theta),1-\varepsilon,1+\varepsilon\right)\hat{A}_t\right)\right] \tag{6}$$

式中：θ 是策略的参数；c_1、c_2 代表两个惩罚因子；$r_t(\theta)=\log\pi_\theta(a_t|s_t)/\log\pi_{\mathrm{old}}(a_t|s_t)$，代表新策略和旧策略之间的概率比[23]。在构造优势函数中，超参数 ε 通过将 $r_t(\theta)$ 限制在 $[1-\varepsilon,1+\varepsilon]$ 之间，进而使每次更新波动保持稳定。

1.2　表示学习

在深度强化学习中，表示学习扮演着至关重要的角色，它使智能体能够从高维、复杂的环境状态中自动提取有用的信息，形成更加紧凑和有效的状态表示[24]。这种能力极大地增强了智能体理解和处理大规模数据的能力，从而提高学习效率和决策质量。这种学习方法在多个领域[25]得到广泛应用。

自编码器（autoencoder，AE）是一种基于无监督学习的神经网络模型，可被应用于强化学习中的动作表示和状态表示。自编码器由编码器和解码器构成，编码器可以学习到数据的底层特征，实现数据的降维，解码器则可以将降维后的数据重构回真实数据。通常用均方误差损失函数衡量训练过程，以最小化输入数据和重构真实数据间的差异。自编码器在异常检测等领域[25]具有出色表现。

变分自编码器（variational autoencoder，VAE）是一种先进的生成模型[31]，结合了深度学习和概率图模型的优势。VAE 不仅能够有效生成新的数据样本，还能学习数据的深层次特征表示[32]。它的设计和传统的自编码器有所不同，主要体现在如何处理编码过程中的潜在空间，具体结构如图 2 所示。其中，μ 表示原始数据在潜在空间分布中的均值，σ 表示原始数据在潜在空间分布中的标准差，z 代表隐变量，即原始数据在潜在空间中的潜在表示，它们定义了一个高斯分布。

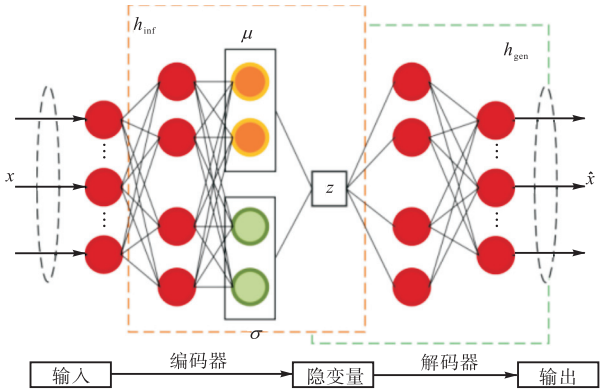

图 2　变分自编码器的结构

Fig. 2　Structure of variational autoencoder

β 变分自编码器（β-VAE）是在变分自编码器损失函数的 KL 散度项中引入了超参数 β 的一种变种，该超参数能够控制潜在空间中潜在变量的解耦程

度。β-VAE 同样由编码器和解码器组成，可以实现数据的降维和重构。β 作为一个正则化系数，对 KL 散度项进行加权，当 $\beta>1$ 时，模型被激励去学习一个更加解耦和结构化的潜在空间表示。此外，由于 β-VAE 倾向于学习更加独立和结构化的特征表示，因此可以更好地应对在训练数据中未见过的新情况，也有助于改善模型的泛化能力，这使 β-VAE 在处理多样化和复杂的数据集时也可以表现得很好。因此，本文选择将 β-VAE 应用于动作表示的学习。

2　强化学习中有效动作表示学习模型

深度强化学习已取得突破性进展，被成功应用到智能交通、机器人、游戏、自然语言处理、计算机视觉等多个领域。深度强化学习的性能依赖于状态的表示能力，相关研究就关于状态表示学习进行了充分研究，并提出了与任务相关的状态表示。为了进一步解决样本利用率问题和泛化能力问题，学者已将动作表示学习引入强化学习中。然而，以往的动作表示学习过程只是将低维的动作空间随机映射到特征空间，学习到的动作表示与任务无关，且缺乏可解释性。

针对上述问题，为了得到与任务相关的动作表示、提高大规模连续动作空间的策略学习性能及效率，本文提出一种基于有效动作表示的策略搜索强化学习方法 TAR-PPO（task-relevant action representation learning based PPO），模型主要由 4 个组件组成：基于 β-VAE 的预训练动作表示、有效动作表示提取器、回报预测模型和近端策略优化算法。首先，将原始动作信息通过 β-VAE 模型的编码器进行编码；然后，根据回报预测模型的辅助任务，通过有效动作表示提取器进一步对原始动作表示中与任务相关的特征进行增强，进而得到有效的动作表示；最后，通过 β-VAE 将有效动作表示解码回真实动作，并在近端策略优化算法的指导下进行策略学习。模型整体结构如图 3 所示。

2.1　基于 β-VAE 的动作表示预训练模型

β-VAE 的解耦性能有助于学习更具泛化性的动作表示，使智能体可以更有效地捕捉任务相关信息，提高在不同动态环境中的泛化能力，为强化学习系统在复杂任务中更灵活地理解和执行动作提供有效支持。将 β-VAE 应用于动作表示的学习，模型的损失函数为

$$J(\zeta,\xi;a_t)=E_{q_\zeta(\boldsymbol{e}_t|a_t)}[\log p_\xi(a_t\,|\,\boldsymbol{e}_t)]-\beta D_{\mathrm{KL}}(q_\zeta(\boldsymbol{e}_t\,|\,a_t)\,\|\,p_\xi(a_t\,|\,\boldsymbol{e}_t)] \tag{7}$$

其中：ζ 为编码器参数；ξ 为解码器参数；a_t 为真实动作；$\boldsymbol{e}_t$ 为训练过程中表示动作的潜在向量；$p_\xi(\boldsymbol{e}_t)$ 为由解码器定义的潜在向量先验分布；$q_\zeta(\boldsymbol{e}_t\,|\,a_t)$ 为由编码器定义的给定真实动作 a_t 下的潜在向量后验分布；$E_{q_\zeta(\boldsymbol{e}_t|a_t)}[\log p_\xi(a_t\,|\,\boldsymbol{e}_t)]$ 为重构误差项，表示给定潜在向量 $\boldsymbol{e}_t$ 时，解码器能够生成原始真实动作 a_t 的概率；$\beta D_{\mathrm{KL}}(q_\zeta(\boldsymbol{e}_t\,|\,a_t)\,\|\,p_\xi(a_t\,|\,\boldsymbol{e}_t))$ 为用于计算编码器产生后验分布与先验分布之间差异的 KL 散度项。

通过预训练 β-VAE 使模型收敛后，固定其编码器参数和解码器参数，为后续强化学习训练提供原始真实动作 a_t 的动作表示 e_t。

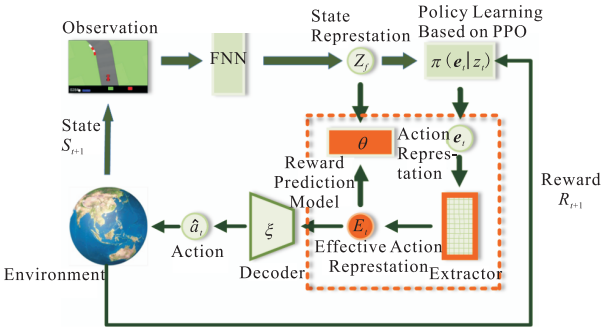

图 3　强化学习中有效动作表示学习模型的结构

Fig. 3　Structure of effective action representation learning model in reinforcement learning

2.2　有效动作表示提取器

在强化学习训练过程中，潜在空间动作表示的学习直接影响智能体的行为选择和训练性能。一个有效的动作表示应该包含尽可能多的当前环境中与任务相关的信息，同时包含尽可能少的与任务无关的信息。本研究引入有效动作表示提取器组件，动作表示通过过滤矩阵与回报预测模型的共同作用，加强其中与任务有关的特征，减弱其中与任务无关的特征，最终得到有效的动作表示。有效动作表示提取器在优化动作表示中起到筛选和强化任务相关动作特征的作用，而回报预测模型则进一步校准和优化这些特征，确保动作表示更加聚焦于对提升智能体表现至关重要的信息。这种结合使用的策略有助于提升动作表示的质量，因为它不仅关注特征的提取，还考虑了这些特征如何更好地对应预期的回报，从而增强了智能体在复杂环境中的适应能力和决策效果。

本文中的有效动作表示提取器是以 exp 函数为激活函数的单层神经网络，以点乘的方式作用在动作

表示上，即 $e = z \odot \exp(I)$，其中 e 表示可解释性的隐空间变量，z 代表有解耦性质的隐变量，I 表示解释过滤网络的参数。

2.3　回报预测模型

回报预测是指在强化学习算法中，通过预测当前状态下的未来奖励，从而优化智能体的行为决策。在强化学习中，智能体与环境不断交互，并在交互中进行学习，以最大化未来的累积奖励。奖励预测可以帮助智能体更好地理解当前状态的奖励，以便更准确地选择动作。

回报预测的实现方式包括监督学习和弱监督学习。在监督学习中，智能体需要从环境中观察到的轨迹中学习奖励函数的参数 θ，并利用该奖励函数对当前状态下的未来奖励进行预测。在弱监督学习中，智能体仅需知晓当前奖励，即可通过优化奖励预测模型最大化未来奖励。

回报预测是基于最近 k 步状态序列对当前回报 r_t 进行预测。首先，将动作表示通过有效动作表示提取器进行馈送，以提取任务相关信息；然后，将提取到的信息传递给回报预测模型 R_θ，用于对当前回报进行预测，并通过计算预测奖励与实际奖励之间的均方误差进行训练。具体操作包括将状态表示和经过有效动作表示提取器优化后的动作表示 E_t 传递给回报预测模型 R_θ，用于预测当前回报，其损失函数为

$$J(R) = \mathrm{MSE}(r_{\mathrm{pred}} - r(s,a)) \tag{8}$$

回报预测模型的作用是作为辅助任务，判断传递的动作表示是否能预测当前奖励的方式，鼓励提取器 I 在动作信息中识别出有希望获得正向奖励的线索，进而提取到与任务相关的有效动作表示。

2.4　基于近端策略优化算法的策略学习

本节结合前面阐述的有效动作表示提取器和回报预测模型，构建基于近端策略优化算法的强化学习中的有效动作表示学习模型。由于 PPO 算法具备高效、稳定的特征，能够同时结合有效动作表示提取器组件和回报预测模型组件，因此将进一步提高模型整体的策略学习效率和性能。

在每回合开始时，智能体始于初始状态 s_1，在之后的每个时间步，都会将当前状态降维到潜在空间。基于策略模型 π，采样出相应的动作表示。此后，有效动作表示提取器将其转化为与任务相关的动作表示 E_t，此表示随即由回报预测模型进行当前奖励的预测。学习到的有效动作表示 E_t 被转换为实际动作，施加到真实环境中，状态发生转移，并返回即时奖励。最后，采用近端策略优化算法更新策略模型，同时基于最小均方误差准则对回报预测模型和有效动作表示提取器进行更新。这一过程优化了动作的选择和奖励的预测，提高了整体策略的效率和效果。

具体算法：

1. 随机策略收集交互样本集 D：$\{(s_t, a_t)\}_{t=1}^{T}$

2. 使用样本集 D 学习 V_s，得到潜在空间中的状态表示

3. 使用样本集 D，依据式(7)学习 V_a，得到潜在空间中的动作表示

4. 初始化策略模型学习参数

5. for episode = 1，2，…，do

6.根据初始状态分布 $P(s_1)$，采样初始状态 s_1

7. for t = 1，2，…，do

8. $z_t = V_s \cdot \mathrm{encoder}(s_t)$

9. 根据 $\pi(\cdot \mid z_t)$ 得到潜在空间中的动作表示

10. 提取任务相关的动作表示 $E_t = e_t \odot \exp(I)$

11. 预测当前奖励 $r_{\mathrm{pred}} = R_\theta(z_t, E_t)$

12. 解码动作表示 $\hat{a}_t = V_a \cdot \mathrm{decoder}(E_t)$

13. 执行动作 $\hat{a}_t$，并得到下一个状态 S_{t+1} 和即时奖励 r_t

14. 使用策略搜索算法更新策略 π

15. 根据最小均方误差更新回报预测模型和有效动作表示提取器

16. end for

17. end for

3　实验结果与分析

3.1　实验任务

本研究的主要创新之处在于开发了一种新型的动作表示学习机制，该机制利用回报预测模型作为辅助任务，引入有效动作表示提取器优化动作表示，强化了原始动作表示中与任务紧密相关的信息，结合 β-VAE 模型和近端策略优化算法，旨在增强模型在策略学习上的性能。为了验证本文方法的效果，在经典的强化学习任务 MountainCar-v0 中进行实验演示，该任务的示意图如图 4 所示。

MountainCar-v0 任务的环境由两个主要部分构成：两座山坡和一辆动力不足的小车。任务目标是通过策略学习，使小车能够获得足够的动力，从而成功爬升至右侧山坡上的目标位置。

MountainCar-v0 的状态空间 S 是一个二维连续

空间，由小车在山坡的位置 $x\in[-1.2,0.5]$ 和速度 $x\in[-0.07,0.07]$ 构成，小车的终点处于 $x=0.45$ 处。动作空间 A 则是一个一维连续的动作空间，动作的取值范围在[−1，1]之间，当动作处于[−1，0]之间表示给小车施加向左的力，处于[0，1]之间表示给小车施加向右的力。奖励函数如式(9)所示，这种设置增加了小车爬山的难度。如果小车没有在短时间内爬上山坡终点，那么智能体就会误以为最佳策略就是保持不动，这会让小车失去找到终点的可能性。

$$r(s_t,a_t,s_{t+1})=\begin{cases}-(a_t)^2\times0.1+100 & x_{t+1}\geqslant0.45\\ -(a_t)^2\times0.1 & \text{otherwise}\end{cases}\tag{9}$$

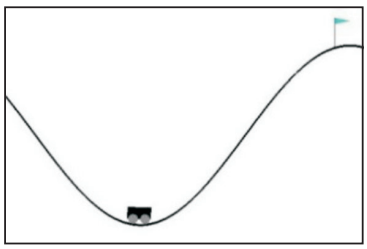

图 4　MountainCar-v0 任务示意图

Fig. 4　MountainCar-v0 task diagram

在 MountainCar-v0 任务中，实验的主要目的在于验证使用 β-VAE 模型学习潜在动作表示后，引入回报预测模型和有效动作表示提取器，将动作优化后对智能体策略学习的影响。

3.2　策略学习

3.2.1　实验设置

在 MountainCar-v0 环境中利用本文基于有效动作表示的策略搜索强化学习方法进行策略学习。对比以下两种方法：(1)VAE-PPO：原始 VAE 学习动作表示方法，并基于 PPO 方法进行策略学习[19]；(2)TAR-PPO：在本研究的有效动作表示学习组件的基础上，利用 PPO 算法进行策略学习。

为了实现本文方法，首先将观测得到的状态信息降维，获得状态表示。随后，此状态表示被送入策略模型，该模型采用了 PPO 算法。PPO 是策略函数与价值函数共享相同的神经网络结构，且由 1 个包含 32 个神经元的全连接隐藏层组成。根据回报预测模型设定的辅助任务，策略模型学习得到的动作表示通过有效动作表示提取器进一步优化，强化与任务紧密相关的原始动作表示中的特征。经过这一步骤，便形成了经过加强的有效动作表示。最后，利用 β-VAE 技术将提取的有效动作表示解码成实际动作，并在近端策略优化算法的引导下完成策略学习。此过程不仅体现了对动作表示学习方法的创新应用，也展示了通过精细化动作表示提升模型策略学习能力的可能性。模型的整体结构如图 5 所示。

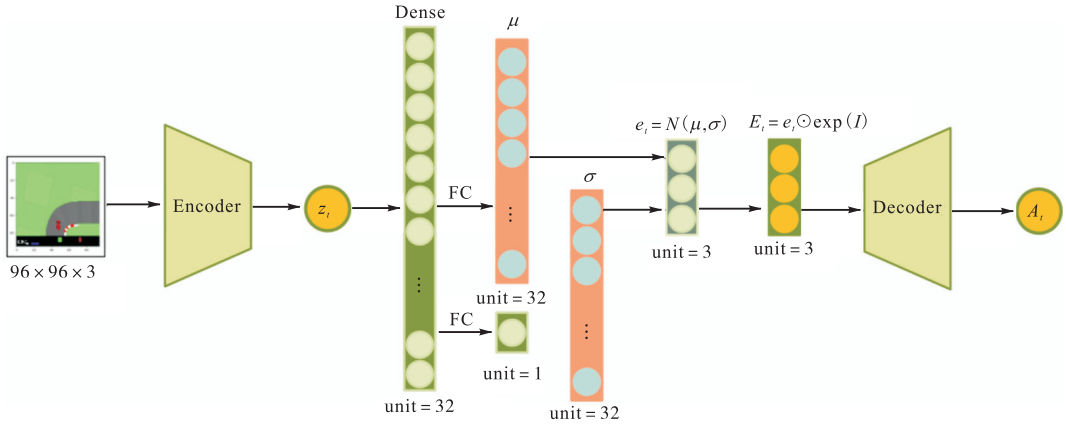

图 5　TAR-PPO 方法的整体网络结构

Fig. 5　Overall network structure of TAR-PPO method

为了确保公平，对比实验的各个组件的网络结构、实验参数以及学习率等都相同，参数设置见表 1。

3.2.2　性能评估

通过平均累积奖励对学习策略进行评价，特别是在 MountainCar-v0 任务中进行了 10 次实验，并对 10 次实验求取平均期望奖励，每次实验均采用不同的随机种子，实验结果如图 6 所示。其中，横轴为迭代次数，纵轴为期望奖励，阴影部分表示标准差。结果表明，TAR-PPO 方法在性能上超过了 VAE-PPO 策略。随着策略迭代的深入，TAR-PPO 方法的奖励值逐渐提高，并最终趋于稳定。相比之下，PPO 方法的奖励增长曲线显示出较多的波动，并且收敛速度较慢。这表明 TAR-PPO 方法在训练过程中表现出更优的性能，尤其是在收敛速度方面具有显著优势。

表 1　MountainCar-v0 任务中 TAR-PPO 方法的超参数设置

Tab. 1　Hyperparameter setting for TAR-PPO method in MountainCar-v0 task

超参数	值
Horizon (T)	256
Learning rate (Adam)	4×10^{-4}
Num. epochs	10
Minibatch size	128
Num. parallel environments	32
Discount (γ)	0.99
GAE parameter (λ)	0.95
Clipping parameter ?	0.2
VF coeff. c_1	0.5
Entropy coeff. c_2	0.01

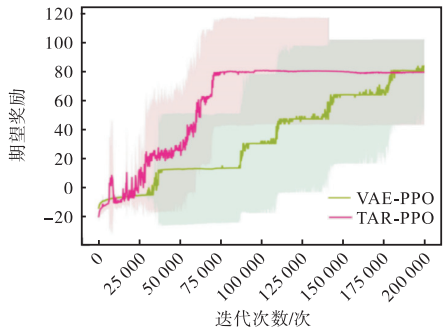

图 6　MountainCar-v0 10 次实验的平均期望奖励

Fig. 6　MountainCar-v0 average expected return on ten experiments

3.3　消融实验

采用消融实验对模型架构进行分析。实验主要评估以下 3 种策略学习方法的效果：(1) 仅包含有效动作表示提取器的方法，简称为 Only-Extractor-VAE-PPO；(2) 仅含有变分自编码器 VA 的方法，简称为 Only-β-VAE-PPO；(3) 综合应用 β-VAE 与有效动作表示提取器的方法，即 TAR-PPO。

在 MountainCar-v0 任务上进行的 10 次实验中，以平均期望奖励作为性能评价指标，基于 10 个测试回合的样本数据计算得到期望奖励值，实验结果如图 7 所示。

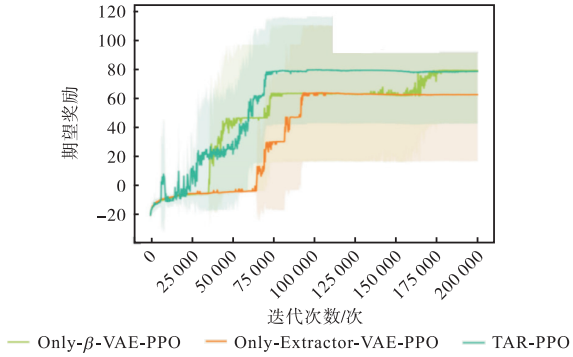

图 7　MountainCar-v0 消融实验

Fig. 7　MountainCar-v0 ablation experiment

图 7 中的横轴表示迭代次数，纵轴表示在交互过程中获得的期望奖励，阴影部分表示标准差。实验结果显示，结合 β-VAE 与有效动作表示提取器的策略学习方法不仅奖励值最高，而且收敛速度最快，这证明了本文方法的优越性。

4　结　语

针对大规模连续动作空间的决策问题，本文提出基于有效动作表示的策略搜索强化学习方法。该方法通过引入回报预测模型作为辅助任务，使有效动作表示提取器组件实现有效动作信息的提取，进而可以学习到有效动作表示，提高强化学习算法的性能和有效性，并在 MountainCar-v0 环境中进行了对比实验和消融实验，验证了本文方法的有效性。

在本研究中，将回报预测模型和有效动作表示提取器组件与 PPO 相结合进行策略学习，并且最终取得了较好的结果。但是，由于 PPO 仅是众多策略学习方法中的一种，且 TAR-PPO 的潜力远不止于此，它可以与其他先进的策略学习方法结合。因此，后续工作可以尝试将 TAR-PPO 框架应用在其他策略学习方法中。

参考文献：

[1] 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.

[2] OH J, GUO X, LEE H, et al. Action-conditional video prediction using deep networks in Atari games[C]// ACM. Proceedings of the 28th International Conference on Neural Information Processing Systems. New York: ACM, 2015: 2863–2871.

[3] 武强. 多智能体强化学习在城市交通信号控制中的研究与应用[D]. 兰州：兰州大学，2020.

[4] 袁伯龙. 基于深度强化学习的信号交叉口智能控制方法研究[D]. 重庆：重庆交通大学，2021.

[5] 付宇钏. 面向交通安全应用的预警及决策算法研究[D]. 西安：西安电子科技大学，2020.

[6] 董豪，杨静，李少波，等. 基于深度强化学习的机器人运动控制研究进展[J]. 控制与决策，2022，37(2)：278–292.

[7] 刘志荣，姜树海. 基于强化学习的移动机器人路径规划研究综述[J]. 制造业自动化，2019，41(3)：90–92.

[8] BRUNKE L, GREEFF M, HALL A W, et al. Safe learn-

ing in robotics: from learning-based control to safe reinforcement learning[J]. Annual review of control, robotics, and autonomous systems, 2022, 5: 411–444.

[9] ZHAO W, QUERALTA J P, WESTERLUND T. Sim-to-real transfer in deep reinforcement learning for robotics: a survey[C]//IEEE. 2020 IEEE Symposium Series on Computational Intelligence (SSCI). New York: IEEE, 2020: 737–744.

[10] WU L, TIAN F, QIN T, et al. A study of reinforcement learning for neural machine translation[EB/OL]. [2023–10–11]. https://doi.org/10.48550/arXiv.1808.08866.

[11] 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.

[12] KIRAN B R, SOBH I, TALPAERT V, et al. Deep reinforcement learning for autonomous driving: a survey [J]. IEEE Transactions on intelligent transportation systems, 2021, 23(6): 4909–4926.

[13] CHEN J, YUAN B, TOMIZUKA M. Model-free deep reinforcement learning for urban autonomous driving [C]//IEEE. 2019 IEEE Intelligent Transportation Systems Conference (ITSC). New York: IEEE, 2019: 2765–2771.

[14] NI Z, PAUL S. A multistage game in smart grid security: a reinforcement learning solution[J]. IEEE Transactions on neural networks and learning systems, 2019, 30(9): 2684–2695.

[15] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Human-level control through deep reinforcement learning[J]. Nature, 2019, 518(7540): 529–533.

[16] HESSEL M, MODAYIL J, VAN HASSELT H, et al. Rainbow: combining improvements in deep reinforcement learning[J]. Proceeding of AAAI conference on artificial intelligence, 2018, 32(1): 11796.

[17] WATTER M, SPRINGENBERG J T, BOEDECKER J, et al. Embed to control: a locally linear latent dynamics model for control from raw images[C]//ACM. Proceedings of the 28th International Conference on Neural Information Processing Systems. New York: ACM, 2015: 2746–2754.

[18] HA D, SCHMIDHUBER J. World models[EB/OL]. [2023–10–11]. https://doi.org/10.48550/arXiv.1803.10122.

[19] 赵婷婷, 王莹, 孙威, 等. 潜在空间中的策略搜索强化学习方法[J]. 计算机科学与探索, 2024, 18(4): 1032–1046.

[20] 郭宪. 深入浅出强化学习: 原理入门[M]. 北京: 电子工业出版社, 2018.

[21] SCHULMAN J, WOLSKI F, DHARIWAL P, et al. Proximal policy optimization algorithms[EB/OL]. [2023–10–11]. https://doi.org/10.48550/arXiv.1707.06347.

[22] SCHULMAN J, LEVINE S, ABBEEL P, et al. Trust region policy optimization[C]//PMLR. International Conference on Machine Learning. New York: PMLR, 2015: 1889–1897.

[23] LILLICRAP T P, HUNT J J, PRITZEL A, et al. Continuous control with deep reinforcement learning [EB/OL]. [2023–10–11]. https://doi.org/10.48550/arXiv.1509.02971.

[24] BENGIO Y, COURVILLE A, VINCENT P. Representation learning: a review and new perspectives[J]. IEEE Transactions on pattern analysis and machine intelligence, 2013, 35(8): 1798–1828.

[25] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84–90.

[26] DAHL G E, YU D, DENG L, et al. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition[J]. IEEE Transactions on audio, speech, and language processing, 2011, 20(1): 30–42.

[27] BORDES A, GLOROT X, WESTON J, et al. Joint learning of words and meaning representations for open-text semantic parsing[C]//PMLR. Artificial Intelligence and Statis-tics. New York: PMLR, 2012: 127–135.

[28] GUTOSKI M, RIBEIRO M, AQUINO N M R, et al. A clustering-based deep autoencoder for one-class image classification[C]//IEEE. 2017 IEEE Latin American Conference on Computational Intelligence (LA-CCI). New York: IEEE, 2017: 1–6.

[29] SABOKROU M, FATHY M, HOSEINI M. Video anomaly detection and localisation based on the sparsity and reconstruction error of auto-encoder[J]. Electronics letters, 2016, 52(13): 1122–1124.

[30] CHANG Y, TU Z, XIE W, et al. Clustering driven deep autoencoder for video anomaly detection[C]//ECCV. Computer Vision-ECCV 2020: 16th European Conference. Berlin: Springer International Publishing, 2020: 329–345.

［31］ LIU H，TANIGUCHI T. Feature extraction and pattern recognition for human motion by a deep sparse auto-encoder［C］//IEEE. 2014 IEEE International Conference on Computer and Information Technology. New York：IEEE，2014：173–181.

［32］ 李耿增. 基于变分自编码器的图像压缩［D］. 北京：北京邮电大学，2021.

责任编辑：郎婧

（上接第 56 页）

吸附量，提高 MFCR 在纸浆中的留着率，从而在提高抄造纸页强度的同时净化白水；(3)在白水循环过程中配合添加 MFCR 与助留剂 CPAM，可强化 MFCR 物料对白水中淀粉的吸附，提高 MFCR 在纸浆中的留着率，从而进一步提高抄造纸页的强度，并改善白水水质。

参考文献：

［1］ 罗明翔. 禁废令对中国造纸业的影响［J］. 造纸信息，2022(4)：18–20.

［2］ 缪应菊，连明磊，贾庆明，等. 二次纤维的角质化修复研究进展［J］. 应用化工，2019，48(2)：438–443.

［3］ 倪书振. 纸表面施胶酶改性淀粉交联性能及其增强机理研究［D］. 南京：南京林业大学，2019.

［4］ 王昊，刘春兰，付润东，等. OCC 制浆过程中淀粉溶出及其对纸浆性能影响［J］. 中国造纸，2023，42(3)：53–58.

［5］ HAN N，ZHANG J H，HOANG M，et al. A review of process and wastewater reuse in the recycled paper industry［J］. Environmental technology & innovation，2021，24：101860.

［6］ 侯纪云. 造纸过程中微生物的危害及控制［J］. 黑龙江造纸，2020，48(2)：15–17.

［7］ 梁静雯，马舒婷，倪书振，等. 羧甲基纤维素钠作 OCC 废纸浆手抄片表面施胶增强剂研究［J］. 纸和造纸，2023，42(1)：19–22.

［8］ 刘姗姗，贺会利，张强，等. 木聚糖酶处理改善废纸浆强度性能的研究［J］. 中华纸业，2019，40(6)：31–35.

［9］ 吴逊谦，宋晓明，王嘉乐，等. 丙烯酰胺接枝壳聚糖的制备及其增强性能研究［J］. 造纸科学与技术，2022，41(6)：8–11.

［10］ 冯琨，孔话峥，王燕燕，等. α–淀粉酶处理废纸浆降解淀粉类有机物及净化浆料研究［J］. 中国造纸，2020，39(2)：15–21.

［11］ 袁广翔，戴红旗，张玉娟. 造纸白水封闭循环对絮聚体系的影响［J］. 中华纸业，2011，32(10)：40–43.

［12］ 梁世杰. 农业剩余物杂细胞微纤化及其在造纸中的应用［D］. 天津：天津科技大学，2022.

［13］ 李晨曦，安兴业，任倩，等. 淀粉在纳米纤维上的吸附研究［J］. 天津造纸，2021，43(1)：17–24.

［14］ 张伟，何北海，LAI R，等. 湿部化学系统中无机盐对纤维素纤维吸附 CPAM 的影响［J］. 造纸科学与技术，2014，33(2)：51–55.

［15］ 畅婉清，朱勇，张玲，等. 细小组分对玉米秸秆高得率浆性能和湿部化学品使用效果影响［J］. 中国造纸学报，2023，38(1)：53–59.

［16］ 于品育. 利用自组装细小组分改善高得率浆纤维的结合性能［D］. 天津：天津科技大学，2020.

［17］ 丁帅，王淑梅，戴红旗. 国产废纸浆增强工艺的比较［J］. 广州化工，2023，51(22)：103–106.

责任编辑：周建军