天津科技大学学报
Journal of Tianjin University of Science & Technology
ISSN 1672-6510,CN 12-1355/N

《天津科技大学学报》网络首发论文

题目：基于有效动作表示的策略搜索强化学习方法
作者：王馨雪，黄佳欣，赵婷婷，陈亚瑞，王媛
DOI：10.13364/j.issn.1672-6510.20240002
收稿日期：2024-01-04
网络首发日期：2024-09-30
引用格式：王馨雪，黄佳欣，赵婷婷，陈亚瑞，王媛．基于有效动作表示的策略搜索强化学习方法[J/OL]．天津科技大学学报．
https://doi.org/10.13364/j.issn.1672-6510.20240002

DOI:10.13364/j.issn.1672-6510.20240002

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（0000）00-0000-00

An Effective Action Representation basedPolicy Search Reinforcement Learning Method

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 TAR-PPO (task advance action representation learning based PPO) based on effective action representation.Using 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 show that our method can 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； policy search reinforcement learning method

强化学习是机器学习领域中一项重要的方法，借鉴了人类学习中的试错机制。与监督学习中的指

收稿日期：2024-01-04；**修回日期：**2024-05-31
基金项目：国家自然科学基金项目（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

导性反馈有所不同，强化学习是以评价性的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。智能体根据奖赏信号选择后续动作，通过最终找到适合当前状态下最优的动作选择策略（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。

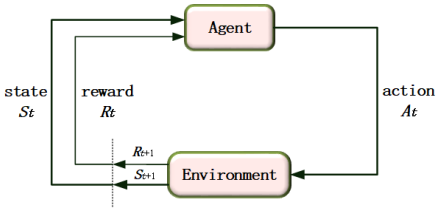

图 1　强化学习基本框架

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]，并得到了极大的成功。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]，突破了原始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，相比原始传统的强化学习的性能有了质的飞跃，在不需要手工干预的情况下，可以直接输出动作。但在深度强化学习中，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]。此外，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的学习。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，并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，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，减少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，促进学习速度的加速。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。良好的动作表示应在包含原始重要信息的基础上，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（markovdecision 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 $M=(S,A,P,R)$ 表示，其中 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。

寻找能够使得算法获得最大累积回报的最优策略 π^* 是强化学习是核心目标[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

$$J(\theta)=\int p(h|\theta)R(h)dh \tag{3}$$

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，近端策略方法（proximal policy optimization, PPO）[21]技术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，这一近端策略优化方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，创造了一种新的基于潜在空间的策略学习方法。目标函数为

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

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

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

θ 是策略的参数；c_1、c_2 代表两个惩罚因子；$r_t(\theta)=\log\pi_\theta(a_t|s_t)/\log\pi_{old}(a_t|s_t)$ 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，进而使得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（representation learning）扮演着至关重要的角色，它使得智能体能够从高维、复杂的环境状态中自动提取有用的信息，以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处理和理解大规模数据的能力，从而提高了学习效率和决策质量。这种学习方法在多个领域[25]得到了广泛的应用。

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。训练过程通常由均方误差损失函数衡量，以最小化输入数据和重构的真实数据间的差异。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，自 2013 年由 Kingma 等提出以来，已成为无监督学习领域的热点。VAE 不仅能够有效生成新的数据样本，还能学习到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，主要体现在其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。

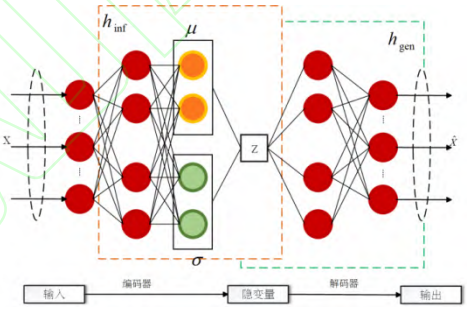

图 2　变分自编码器结构

Fig. 2　Variational autoencoder structure

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缠绕程度。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，加权 KL 散度项，当 β >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那些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，这使得 β -VAE 在处理多样化和复杂的数据集时也可以表现的很好。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的进展，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。深度强化学习的成功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和泛化能力的问题，学者们已将动作表示学习引入到强化学习中。然而，以往的动作表示的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加强，进而得到有效的动作表示；最后，通过 β-VAE 将有效的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。

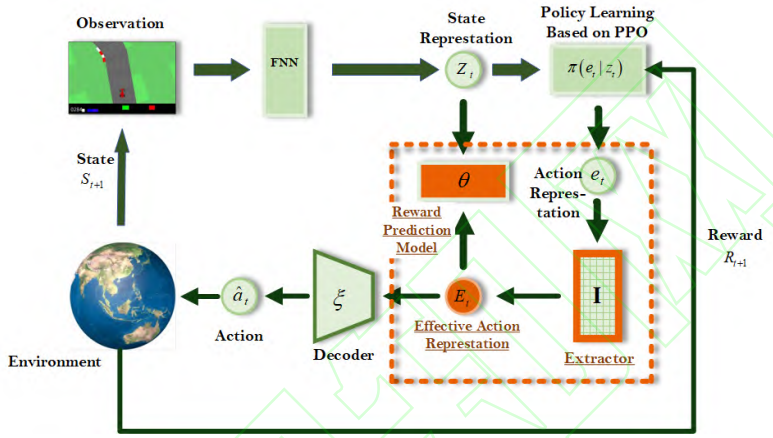

图 3　强化学习中有效动作表示学习模型结构图

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

2.1　基于 β-VAE 的动作表示预训练模型

β-VAE 的解纠缠性能有助于学习更具泛化性的动作表示，使智能体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益支持。将 β-VAE 应用于动作表示的学习，模型的损失函数为

$$\begin{aligned} J(\zeta,\xi;a_t) &= E_{q_\zeta(e_t|a_t)}[\log\ p_\xi(a_t|e_t)] \\ &- \beta D_{KL}(q_\zeta(e_t|a_t)\|\,p_\xi(a_t|e_t)] \end{aligned} \tag{7}$$

其中：ζ 为编码器的参数；ξ 为解码器参数；a_t 为真实动作，e_t 为训练过程中表示动作的潜在向量；$p_\xi(e_t)$ 为由解码器定义的潜在向量先验分布；$q_\zeta(e_t|a_t)$ 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；$E_{q_\zeta(e_t|a_t)}[\log\ p_\xi(a_t|e_t)]$ 为重构误差项，表示给定潜在向量 e_t 时，解码器能够生成原始真实动作 a_t 的概率；$\beta D_{KL}(q_\zeta(e_t|a_t)\|\,p_\xi(a_t|e_t))$ 为用于测量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 使得模型收敛后，固定其编码器和解码器参数，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。

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了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，将动作表示通过过滤矩阵与回报预测模型的共同作用，加强其中和任务有关的特征，减弱其中和任务无关的特征，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了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，因为它不仅关注于特征的提取，还考虑了这些特征如何更好地对应于预期的回报，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代表有解纠缠性质的隐变量，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，并在不断地交互中学习，以最大化未来的累积奖励。奖励预测可以帮助智能体更好的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未来的奖励进行预测。在弱监督学习中，智能体仅需知晓当前的奖励值，即可通过优化奖励预测模型以最大化未来奖励值。

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信息提取器进行馈送，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

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

回报预测模型的作用是以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的动作表示。

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，构建了基于近端策略优化算法的强化学习中有效的动作表示学习模型。由于 PPO 算法具备高效稳定的特征，能够同时可以有效结合有效动作信息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。最终，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，依据公式（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. $z_t = V_s.\text{encoder}(s_t)$

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. $r_{\text{pred}} = R_\theta(z_t, E_t)$

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

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。

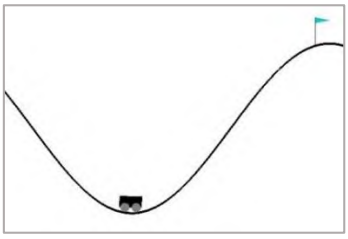

图 4　MountainCar_V0 任务示意图

Fig. 4　MountainCar_ V0 Task Diagram

MountainCar_V0 任务的环境由两个主要组成部分构成：两座山坡和一辆动力不足的小车。其任务目标在于通过策略学习，使得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。

$$r(s_t,a_t,s_{t+1})=\begin{cases}-(a_t)^2*0.1+100, & x_{t+1}\geq 0.45\\ -(a_t)^2*0.1, & \text{otherwise}\end{cases}\tag{9}$$

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。

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的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 环境中利用本文所提的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。本节对比以下两种方法:（1）VAE-PPO：原始 VAE 学习动作表示的方法，并基于 PPO 方法进行策略学习[19]。（2）TAR-PPO：在本文所提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。

为了落实本文方法，首先是将观测得到的状态信息降维，获得状态的表示。随后，此状态表示被送入策略模型，该模型采用了 PPO 算法。PPO 是策略与价值函数共享相同的神经网络结构，且由一个包含 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行进一步优化，强化了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。最终，利用 β -VAE 技术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的学习。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以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。

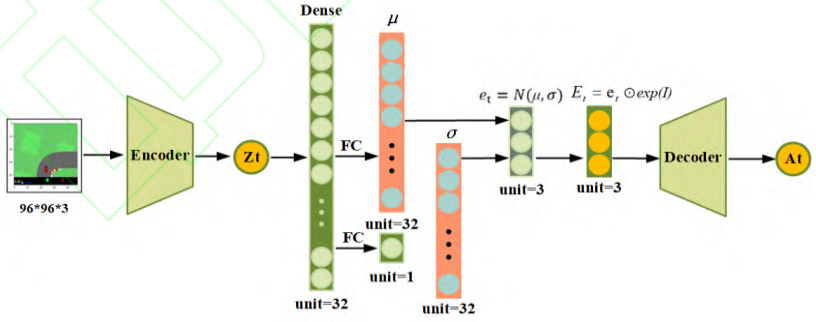

图 5　TAR-PPO 方法的整体网络结构

Fig. 5　The overall network structure of TAR-PPO method

为确保公平，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，具体参数设置见表 1。

表 1MountainCar_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)	4e-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

3.2.2　性能评估

通过平均累积奖励对采用的学习策略进行评价，特别在 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。实验结果展示如图 6，其中横轴代表迭代次数，纵轴为期望奖励值，阴影区域则反映了标准差的大小。结果显示，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。相较之下，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。这一发现表明，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的优势。

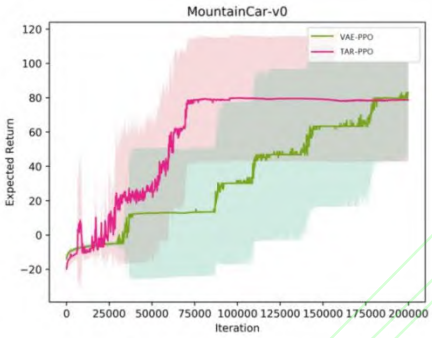

图 6　MountainCar_V0 10 次实验的平均期望回报

Fig. 6　MountainCar_ V0 average expected return on 10 experiments

3.3 消融实验

本研究旨在探讨 TAR-PPO 模型各构成要素的影响，采用消融实验法对模型架构进行分析。实验主要评估以下三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。图中的横轴表示训练迭代次数（Iteration），纵轴表示在交互过程中获得的期望奖励值，阴影部分代表标准差。图中，深绿色折线代表 TAR-PPO 的累积奖励，橙色折线表示仅含高效动作表示提取器的策略累积奖励，浅绿色折线则代表仅含 β-VAE 组件的累积奖励。

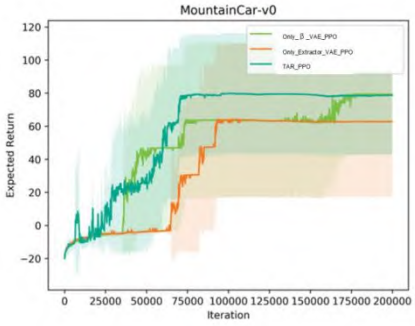

图 7　MountainCar_V0 消融实验

Fig. 7　MountainCar_ V0 ablation experiment

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，这证明了本文所提有效动作表示学习方法的优越性。

4　结语

针对大规模连续动作空间的决策问题，本文提出了一种新型动作表示学习算法，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 MountainCar_V0 环境中进行了对比实验和消融实验，证明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信息提取器组件与 PPO 相结合，进行策略的学习，并且最终取得了良好的结果。但是，由于 PPO 仅是众多策略学习方法中的一个，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相结合。因此，在后续的工作中，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[J]. 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 learning 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). IEEE，2020:737-744.

[10] WU L，TIAN F，QIN T，et al. A study of reinforcement learning for neural machine translation[J]. Arxiv preprint arxiv:1808. 08866，2018.

[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.

[13] CHEN J，YUAN B，TOMIZUKA M. Model-free deep reinforcement learning for urban autonomous driving[C]//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] VOLODYMYR M，KORAY K，DAVID S，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[C]//AAAI. Thirty-second 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, 2: 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/OL]. 计算机科学与探索 :1-19[2024-03-03].http://kns.cnki.net/kcms/detail/11.5602.tp.20230329.1558.004.html.

[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]//International Conference on Machine Learning. 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]//Artificial Intelligence and Statistics. 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 autoencoder[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.