Shihmin WANG, Binqi ZHAO, Zhengfeng ZHANG, Junping ZHANG, Jian PU, 2023. Embedding expert demonstrations into clustering buffer for effective deep reinforcement learning. *Frontiers of Information Technology & Electronic Engineering*, 24(11):1541-1556. https://doi.org/10.1631/FITEE.2300084

Embedding expert demonstrations into clustering buffer for effective deep reinforcement learning

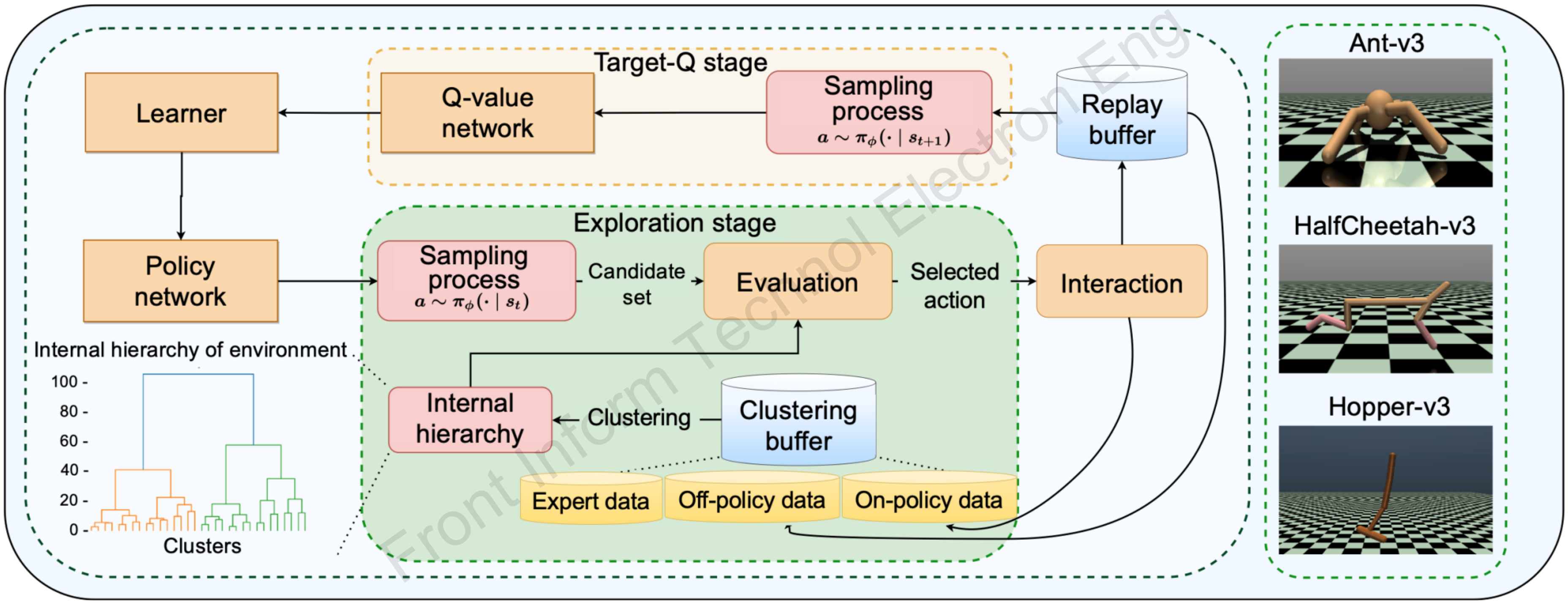
Key words: Reinforcement learning; Sample efficiency; Sampling process; Clustering methods; Autonomous driving

Shihmin WANG
E-mail: wangshimin20@fudan.edu.cn
ORCID: https://orcid.org/0000-0002-7288-8323

Motivation

- As one of the most fundamental topics in reinforcement learning (RL), sample efficiency is essential to the deployment of deep RL algorithms. Existing researches tried to alleviate the need for samples by solving the credit assignment problem or introducing expert demonstration. On the other hand, existing exploration methods sample an action from different types of posterior distributions. However, these methods neglect the importance of expert demonstration in exploration.
- To deploy deep RL algorithms to some costly domains, such as robotic manipulation and autonomous driving, it is of paramount importance to model the internal hierarchy of the environment.

Framework



Structure of selective sampling

Method

- We focus on the policy sampling process and propose an efficient selective sampling approach to improve the sample efficiency by modeling the internal hierarchy of the environment.
- We first employ clustering methods on the policy sampling process, which generates an action candidate set. Then we introduce a clustering buffer for modeling the internal hierarchy, which consists of on-policy data, off-policy data, and expert data, to evaluate actions in the action candidate set in the exploration stage.
- Selective sampling has played an essential role in the exploration stage and is able to take more advantage of the supervision information in the expert demonstration data.

Method

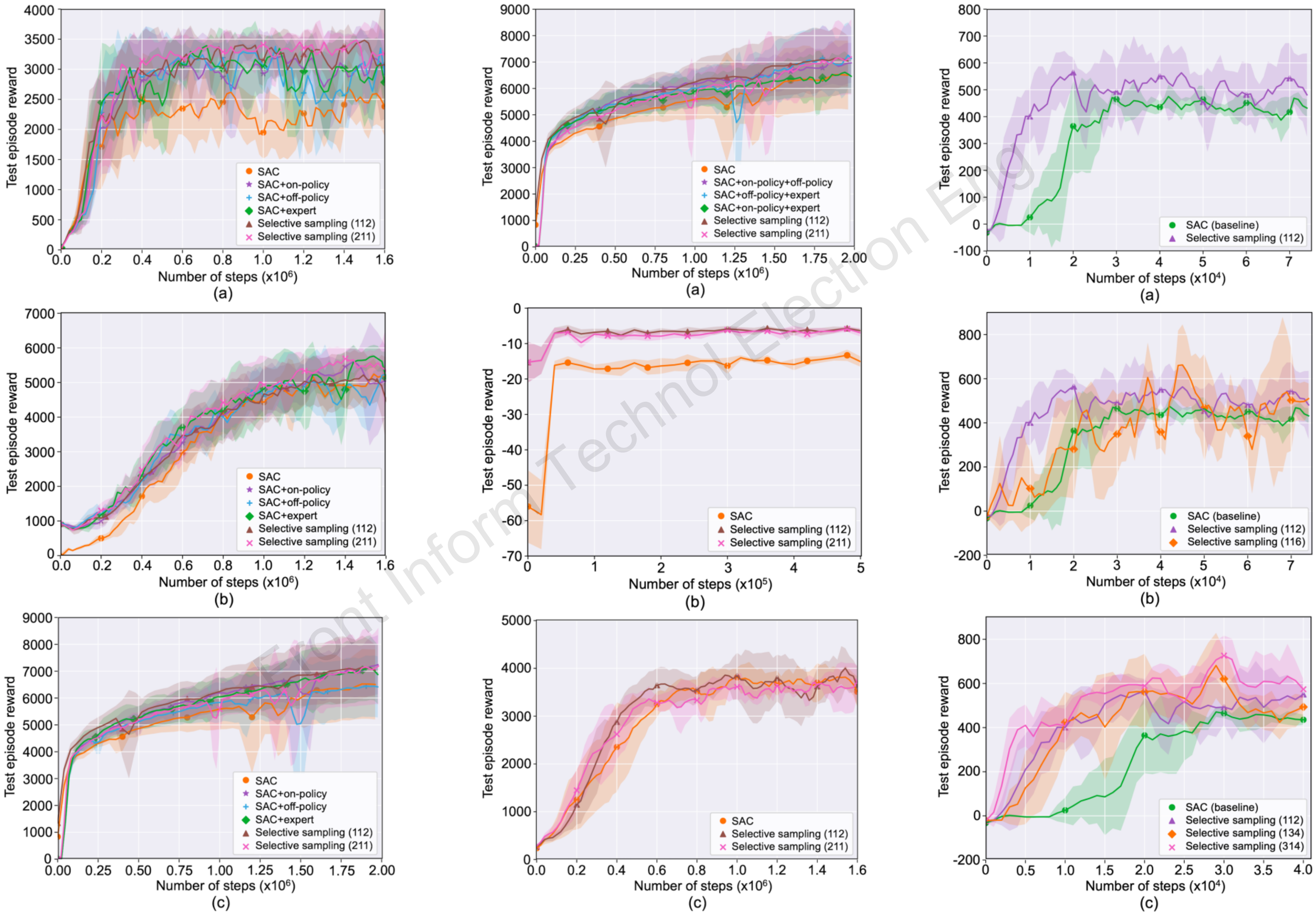
Algorithm 1 Selective sampling

Require: on-policy data $\mathcal{D}_{\text{on}}$, off-policy data $\mathcal{D}_{\text{off}}$, expert data $\mathcal{D}_{\text{E}}$

Ensure: learned policy $\pi_\phi(a \mid s)$

Initialize network parameters θ, θ^-, ϕ
for each environment step **do**
  $\mathcal{D}_{\text{s}} \leftarrow a \sim \pi_\phi(\cdot \mid s_t)$
  $\mathcal{D}_{\text{c}} \leftarrow (\mathcal{D}_{\text{on}}, \mathcal{D}_{\text{off}}, \mathcal{D}_{\text{E}})$
  $N_{\text{c}} \leftarrow \text{AgglomerativeClustering}(\mathcal{D}_{\text{c}})$
  $\mathcal{C} \leftarrow K\text{-means}(\mathcal{D}_{\text{c}}, N_{\text{c}}, \mathcal{D}_{\text{s}})$
  for each c in $\mathcal{C}$ **do**
    $V_\eta(c) \leftarrow \frac{1}{n_h} \sum_{k=i}^{i+n_h-1} \eta(s_k, a_k)$
  end for
  $a_{\text{selected}} \leftarrow \text{Random}(\text{Softmax}_{c \in \mathcal{C}} V_\eta(c))$
  $s_{t+1} \sim \mathcal{T}(\cdot \mid s_t, a_{\text{selected}})$
  $\mathcal{D}_{\text{Q}} \leftarrow \mathcal{D}_{\text{Q}} \cup \{s_t, a_{\text{selected}}, r_t, s_{t+1}\}$
  for each gradient step **do**
    $\theta \leftarrow \theta - \lambda_{\text{Q}} \nabla_\theta \mathcal{B}_\theta(Q)$
    $\phi \leftarrow \phi - \lambda_\pi \nabla_\phi \mathcal{J}_\phi(\pi)$
    $\theta^- \leftarrow \lambda_{\text{target}} \theta + (1 - \lambda_{\text{target}}) \theta^-$
  end for
end for

Major results



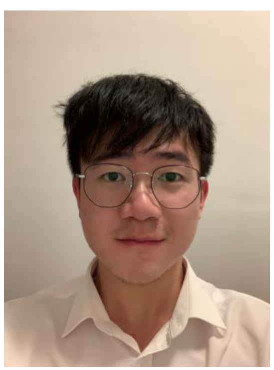
Convergence curves of training in Mujoco (left & middle) and LGSVL (right) tasks

Conclusions

- This paper presents a novel sampling approach, selective sampling, to improve the sample efficiency of RL. Experiments on six different continuous locomotion environments demonstrate superior RL performance and faster convergence of selective sampling. In particular, on the LGSVL task, our method can reduce the number of convergence steps by 46.7% and the convergence time by 28.5%.
- In the implementation of selective sampling, we recommend to increase the expert data and on-policy data in costly domains, such as robotic manipulation and autonomous driving. The ward setting and average setting are recommended for the linkage criterion.

Shihmin WANG received his BS degree from Peking University, China, in 2020, and MS degree from Fudan University, China, in 2023. He is currently a practitioner in the autonomous driving industry. His research interest lies at the intersection of reinforcement learning and autonomous driving.

Binqi ZHAO received his BS degree from Sun Yat-sen University and MS degree from Fudan University. Currently he is a dedicated member of the Guangzhou Institute of Tropical and Marine Meteorology, China Meteorological Administration, focusing on research in meteorological AI. His areas of expertise encompass fields such as deep learning, reinforcement learning, and meteorological AI.



Zhengfeng ZHANG received his BS degree from South China University of China and MS degree from Fudan University. Currently he is a software engineer in Tencent AI Lab, devoted to applying reinforcement learning in game. His research interests include offline reinforcement learning, intelligent decision making of game AI, and machine learning.

Junping ZHANG received his BS degree in automation from Xiangtan University, China, in 1992, MS degree in control theory and control engineering from Hunan University, China, in 2000, and PhD degree in intelligent system and pattern recognition from the Institute of Automation, Chinese Academy of Sciences, China, in 2003. He has been a professor with the School of Computer Science, Fudan University since 2011. He has published in highly ranked international journals such as *IEEE TPAMI* and *IEEE TNNLS* and leading international conferences such as ICML and ECCV. He has been an associate editor of *IEEE Intelligent Systems* since 2009. His current research interests include machine learning, image processing, biometric authentication, and intelligent transportation systems.

Jian PU received his PhD degree from Fudan University in 2014. Currently he is a young principal investigator at the Institute of Science and Technology for Brain-inspired Intelligence (ISTBI), Fudan University. He was an associate professor at the School of Computer Science and Software Engineering, East China Normal University from 2016 to 2019, and a postdoctoral researcher at the Institute of Neuroscience, Chinese Academy of Sciences from 2014 to 2016. His current research interest focuses on developing machine learning and computer vision methods for autonomous driving.