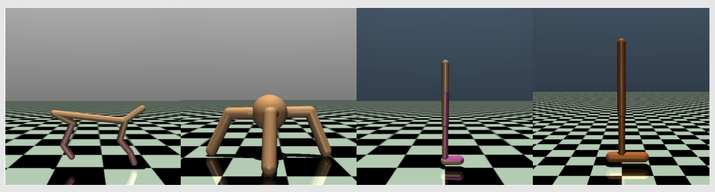
Self-Supervised Reinforcement Learning for Out-of-Distribution Recovery via Auxiliary Reward

Yufeng Xie Yinan Wang Han Wang* Qingshan Li

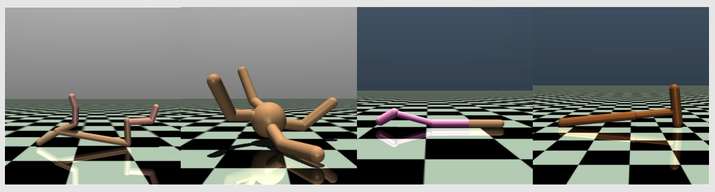
Xi'an Key Laboratory of Intelligent Software Engineering, Xidian University, China
{yufeng.xie@stu., wangyinan@stu., wanghan@, qshli@mail.}xidian.edu.cn *Corresponding author

Background

- RL algorithms often struggle when deployed in real-world scenarios due to the presence of out-of-distribution (OOD) states, which are states that significantly deviate from the training distribution.
- Addressing the challenge of OOD is crucial to ensure the robustness and generalization of RL agent, as their performance can deteriorate when encountering OOD states.



(a) Training environments



(b) OOD environments

Figure 1. MuJoCo tasks [Brockman et al. 2016].

Contributions

1. We introduce the self-supervised state-action abstraction based on return distribution, which effectively forces the learned representations to discriminate state-action pairs with different returns.
2. We propose a self-supervised reinforcement learning approach, SRL-AR, which incorporates standard RL algorithms with the auxiliary self-supervised state-action abstraction learning task, enabling RL agent to handle OOD states.
3. We present experimental results on MuJoCo tasks. The results show that SRL-AR effectively enables the agent to recover from OOD situations, and outperforms prior works in the sample efficiency.

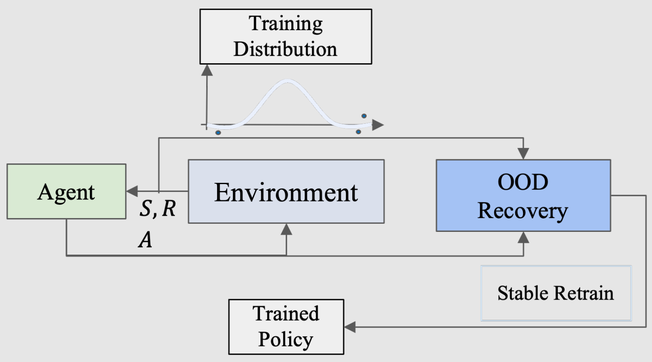


Figure 2. OOD recovery scenario.

Methodology

We present our method SRL-AR, as shown in Figure 3.

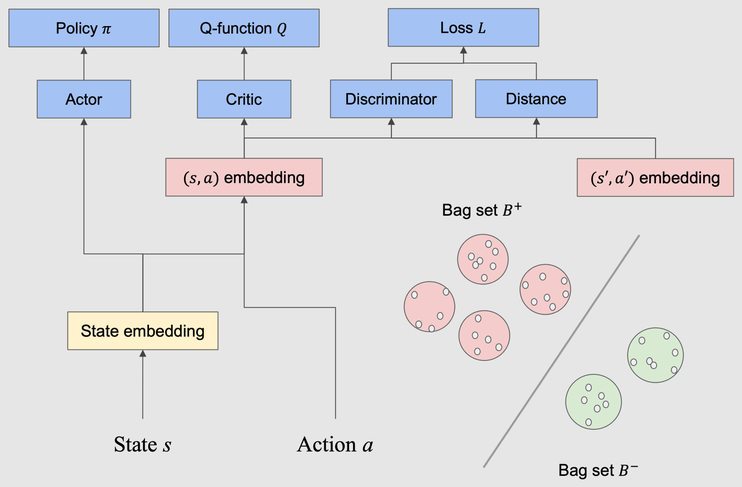


Figure 3. An overview of SRL-AR.

- We propose self-supervised state-action abstraction $\phi(s)$ based on return distribution. We maintain the binary label y for this state-action pair, which indicates whether the two returns belong to the same bin. Besides, we introduce a distance constraint item to enforce the abstraction $\phi(s,a)$ in bins with larger rewards to be closer. The hierarchical contrastive loss is defined as follows:

$$\mathbb{E}_{(s,a,y)\sim\mathcal{D}}\left[(w(\phi(s_1,a_1),\phi(s_2,a_2))-y)^2\right] +\lambda_1 dis(\phi^+(s_1,a_1),\phi^+(s_2,a_2)) -\lambda_2 dis(\phi^+(s_1,a_1),\phi^-(s_2,a_2))$$

- Then we consider an auxiliary reward design method, which enables us to calculate auxiliary reward r^{aux} from the samples collected using the state-action abstraction. We denote the bins with larger rewards as $\mathcal{B}^+$. Thus, the rest bins are $\mathcal{B}^- = \mathcal{B}/\mathcal{B}^+$, which may contain OOD situations.

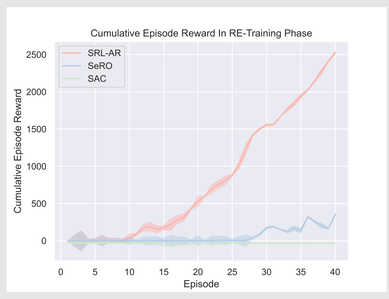
$$r^{aux}(s_t,a_t) = \alpha \cdot \frac{dis(\phi(s_t,a_t),\mathcal{B}^-)}{dis(\phi(s_t,a_t),\mathcal{B}^+)}$$

- After that, we introduce return-based contrastive representation learning for RL that incorporates standard RL algorithms with the auxiliary self-supervised state-action abstraction learning task. By minimizing the objective function, the agent is retrained to return to the learned state distribution from an OOD state by using the auxiliary reward while being regularized by KL-divergence not to forget the original tasks.

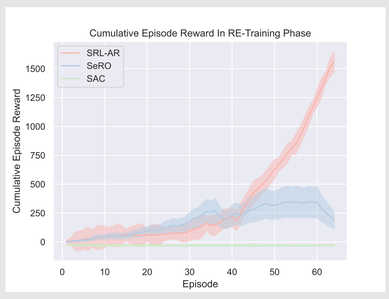
$$\mathcal{L}_{stable} = \beta \cdot D_{KL}(\pi(s,a)\|\pi_{pre}(s,a))$$

Main Results

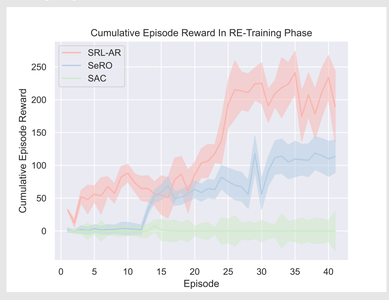
When we apply the trained policy to OOD environments, the results show SAC agent fails to recover from OOD states on all tasks. Meanwhile, both SRL-AR and SeRO [Kim et al. 2023] can effectively recover from OOD states, but SRL-AR shows higher sample efficiency and average cumulative reward than SeRO in the early stage of OOD environments.



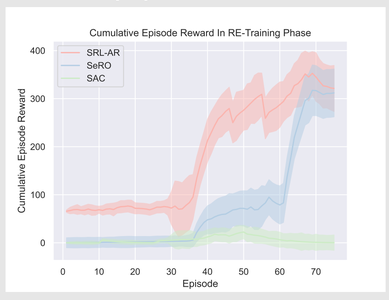
(a) HalfCheetah-v2



(b) Ant-v2



(c) Walker2D-v2



(d) Hopper-v2

Figure 4. Learning curves in the OOD environments.

Table 1. Cumulative reward achieved by SRL-AR and baselines (SeRO, SAC) after the OOD phase.

Task	SRL-AR	SeRO	SAC
HalfCheetah-v2	**11377.96 ± 371.65**	11467.76 ± 764.58	371.59 ± 452.76
Ant-v2	**5862.46 ± 802.36**	5479.12 ± 877.46	151.37 ± 97.14
Walker2D-v2	**3832.76 ± 332.19**	3648.29 ± 412.32	137.49 ± 82.48
Hopper-v2	**3531.93 ± 219.75**	3354.76 ± 378.24	114.83 ± 76.56

Ablation Studies

To analyze the effect of each component of our method, we perform an ablation study to assess two variants in the retraining phase, including the proposed method that only uses auxiliary reward and is regularized by KL-loss, the results of which are presented in Table 2.

Table 2. Cumulative reward achieved of the ablation studies.

Task	Auxiliary Reward	Regularization
HalfCheetah-v2	9374.96 ± 513.65	2113.48 ± 602.16
Ant-v2	4367.59 ± 753.47	2479.23 ± 865.52
Walker2D-v2	3814.35 ± 319.31	1732.60 ± 523.79
Hopper-v2	2932.81 ± 428.20	2468.59 ± 921.46

Note that the regularization receives zero rewards instead of auxiliary rewards in OOD states.

References

- Kim, C., J. Cho, C. Bobda, S.-W. Seo, and S.-W. Kim (2023). "SeRO: self-supervised reinforcement learning for recovery from out-of-distribution situations". In: *arXiv preprint arXiv:2311.03651*.
- Brockman, G., V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba (2016). "Openai gym". In: *arXiv preprint arXiv:1606.01540*.