State Augmentation Based Approach to Reinforcement Learning from Human Preferences

Mudit Verma, Subbarao Kambhampati
School of Computing & AI, Arizona State University
{muditverma, rao} @ asu.edu

R2HCAI, AAAI 2023

Objectives

While Preference Based Reinforcement Learning has progressed in several continous control tasks, progress with image based state inputs is still limited.

Data augmentation has provided a zero-cost to human in the loop method in several areas of Machine Learning. We intend to use State based data augmentation technique for improving the reward learning objective.

Augmented trajectories modify the irrelevant parts of the image representation. We hypothesize image regions that do no update upon a transition as potentiall irrelevant.

Empirically verify claims on popular PbRL benchmarks.

PbRL Preliminaries

- **Approximating the Human Reward Model**

> **Objective:** Compute probability for the human preferring trajectory τ_1 over τ_2:

$$P_\psi[\tau_0 \succ \tau_1] = \frac{\exp(\sum_t R_h(s_t^0, a_t^0))}{\sum_{i\in\{0,1\}} \exp(\sum_t R_h(s_t^i, a_t^i))}$$

> **Objective:** The human reward model can then by learnt by minimizing the cross-entropy between the predictions made by the supervised learner and the ground truth human labels as follows:

$$\mathcal{L}_{CE} = -\underset{(\tau_0,\tau_1,y)\sim\mathcal{D}}{\mathbb{E}}[y(0)\text{log}P_\psi[\tau_0 \succ \tau_1] + y(1)\text{log}P_\psi[\tau_1 \succ \tau_0]]$$

Method : State Augmentation

We augment the trajectory data in the feedback buffer and propose the invariance loss to ensure that the reward predictions are consistent across the perturbed trajectories.

- **State Augmentation**

> Insight : We hypothesize that for tacit tasks such as locomotion or robotic manipulation, an important feature to consider is how the states change upon transition.

> We create a mask for state-differences as follows :

$$\mathbb{M}(I, \{I_1, I_2 \cdots I_n\}) = \bigcup_{j\in\{1,2\cdots n\}} \mathbb{M}(I, I_j)$$

$$\mathbb{M}(I, I_j) = \begin{cases} 1, I(x,y) \neq I_j(x,y) \\ 0, \text{otherwise} \end{cases}$$

> Perturb the image representations to get new state as :

$$\phi(I, I_j, \mathbb{M}) = I \odot (1 - \mathbb{M}(I, I_j) + \mathcal{G}(I, \sigma_\mathcal{G}) \odot \mathbb{M}(I, I_j)$$

> Finally, generate perturbed trajectories as :

$$\tau_g =< s_0, \phi_\mathbb{M}(s_1, s_0), \phi_\mathbb{M}(s_2, s_1) \cdots \phi_\mathbb{M}(s_n, s_{n-1}) >$$

- **Invariance Consistency across Augmentations**

> Following prior work in data augmentation, we force predicted reward vector for perturbed trajectories to be same as the original trajectory.

$$\mathcal{L}_I(\tau) = \frac{1}{|\mathcal{G}|}\sum_{i\sim\mathcal{G}} \left\|\mathbf{R}_\psi^h(\tau) - \mathbf{R}_\psi^h(\tau_g^i)\right\|_2$$

$$\mathcal{L}_{reward} = \lambda_{CE}\mathcal{L}_{CE} + \lambda_I\mathcal{L}_I$$

Kambhampati's research is supported by the J.P. Morgan Faculty Research Award, ONR grants N00014-16-1-2892, N00014-18-1-2442, N00014-18-1-2840, N00014-9-1-2119, AFOSR grant FA9550-18-1-0067 and DARPA SAIL-ON grant W911NF19-2-0006.

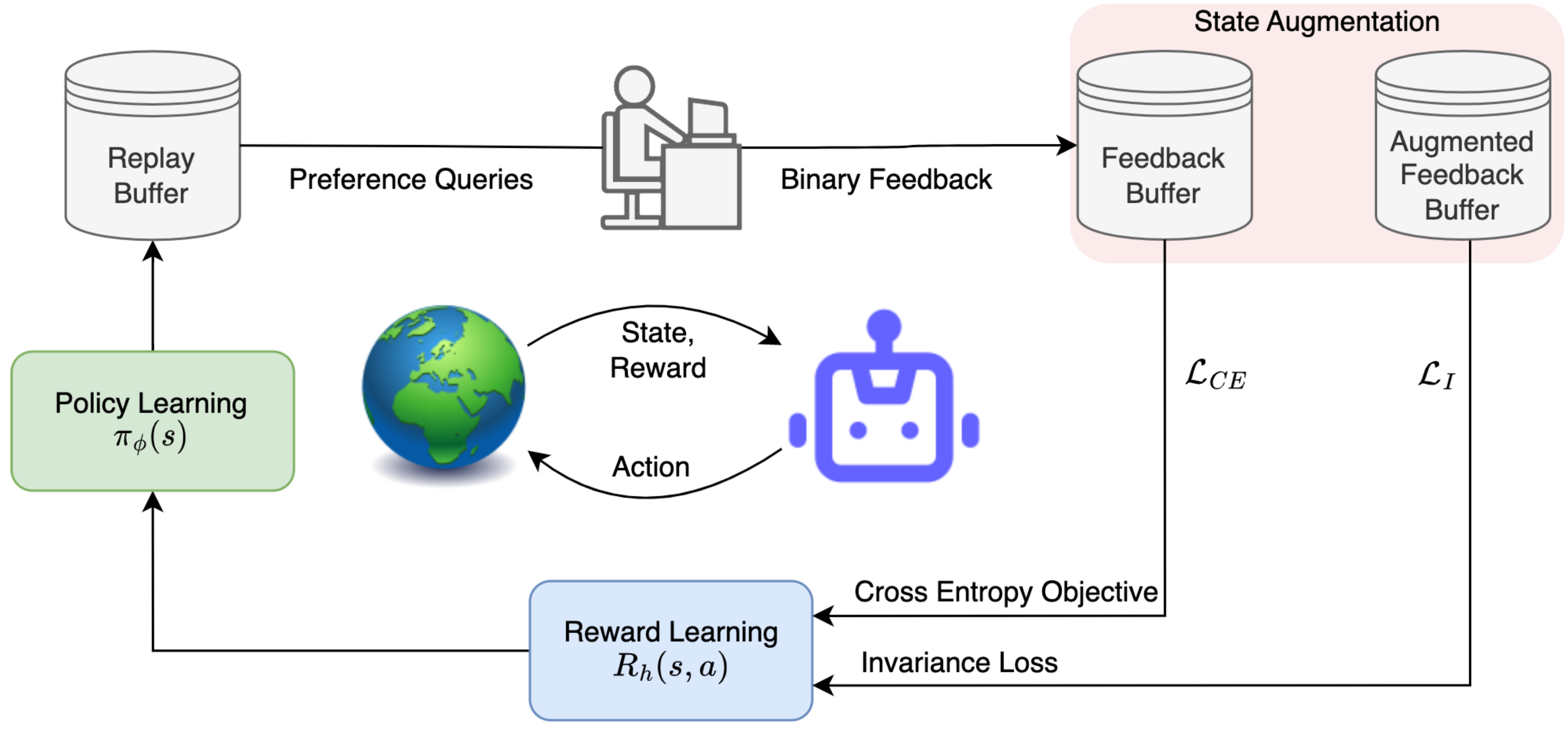


Figure 1: Overview of our proposed approach

Experiments & Results

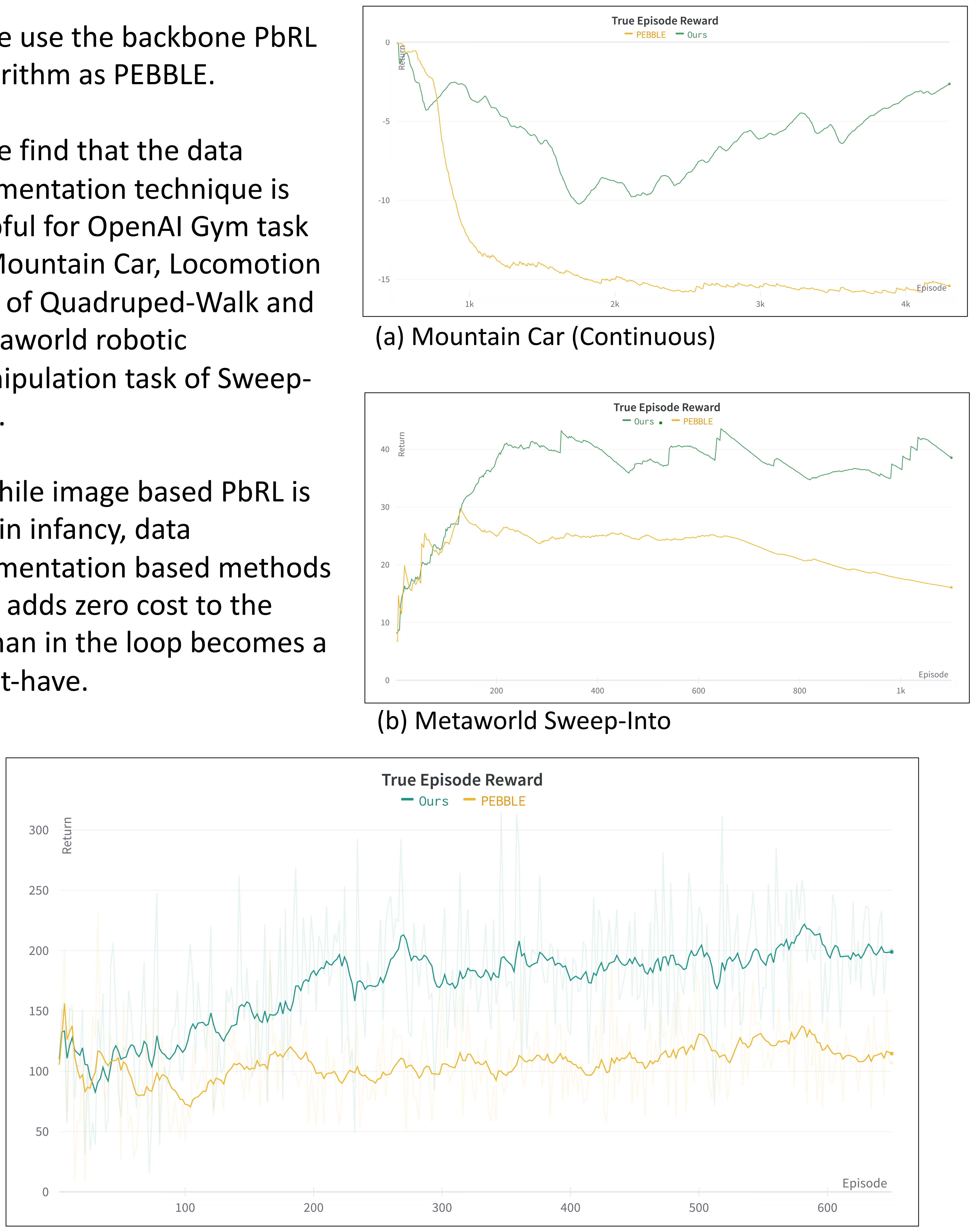
- **Investigate via Empirical Evaluation**

> **Question :** Does the proposed state augmentation approach improve agent performance in PbRL setup?

> We use the backbone PbRL algorithm as PEBBLE.

> We find that the data augmentation technique is helpful for OpenAI Gym task of Mountain Car, Locomotion task of Quadruped-Walk and Metaworld robotic manipulation task of Sweep-Into.

> While image based PbRL is still in infancy, data augmentation based methods that adds zero cost to the human in the loop becomes a must-have.



(a) Mountain Car (Continuous)

(b) Metaworld Sweep-Into

(c) DM Control Quadruped-Walk

Figure 2: Evaluation curves on various continous control tasks compairing PEBBLE with Our method.

TLDR

We propose a state augmentation technique tailored for image-based Preference-based Reinforcement Learning.

Insight : Regions of the image observation that update upon a transition are at least a subset of all the ``content" available in the image observation. Reward model should be invariant to changes to regions except content.

We evaluate on OpenAI gym's Mountain Car, DM Control Quadruped-Walk, and Meta World's Sweep-Into.