Limited Preference Aided Imitation Learning from Imperfect Demonstrations

Xingchen Cao, Fan-Ming Luo, Junyin Ye, Tian Xu, Zhilong Zhang & Yang Yu

LAMDA Learning And Mining from DatA

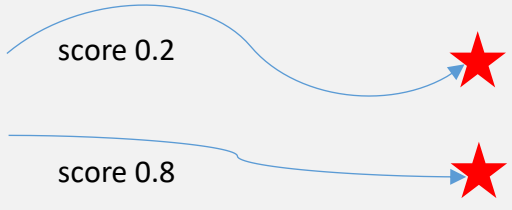
ICML International Conference On Machine Learning

Overview

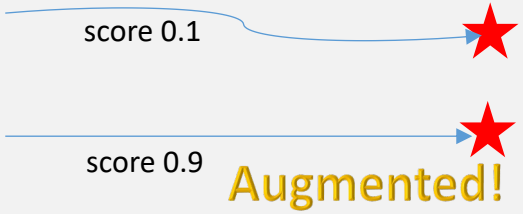
We propose a novel imitation learning (IL) algorithm, Preference Aided Imitation Learning from imperfect demonstrations (PAIL). Specifically, PAIL learns a preference reward by querying experts for limited preferences from imperfect demonstrations. By reweighting imperfect demonstrations with the preference reward for higher quality and selecting explored trajectories with high cumulative preference rewards to augment imperfect demonstrations, PAIL breaks through the performance bottleneck of the imperfect demonstrations.

The Mechanism of PAIL

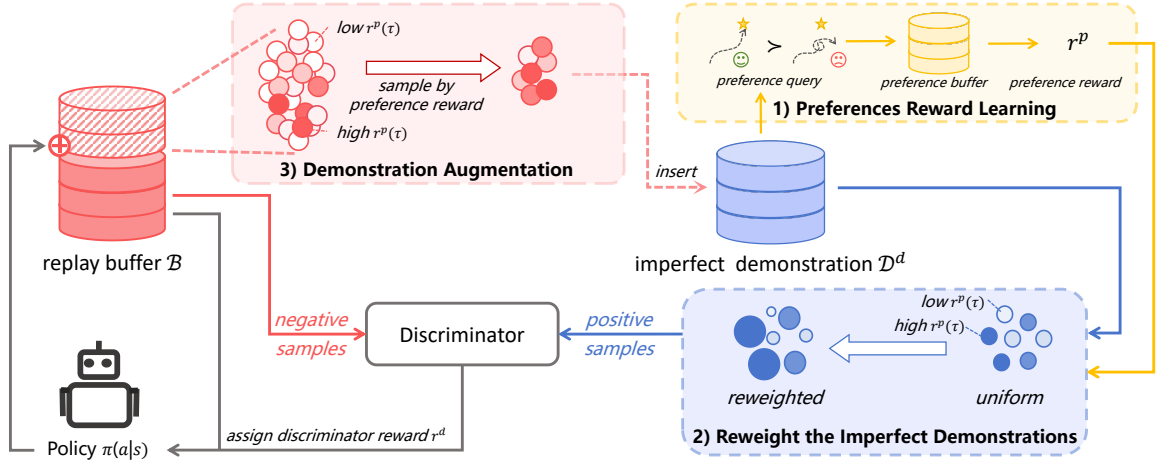
Imperfect Demonstrations



Replay Buffer



Overview of Preference Aided Imitation Learning (PAIL)



Preferences Reward Learning

Preferences Reward is learned by Bradley & Terry model

$$P_\varphi\left[\sigma^0 \prec \sigma^1\right] = \frac{\exp \sum_{(s,a)\in\sigma^1} r^p_\varphi(s,a)}{\sum_{i\in\{0,1\}} \exp \sum_{(s,a)\in\sigma^i} r^p_\varphi(s,a)}$$

Reweight the Imperfect Demonstrations

PAIL applies AIL to imitate from a reweighted demonstration dataset. Trajectories from the demonstrations are reweighted by:

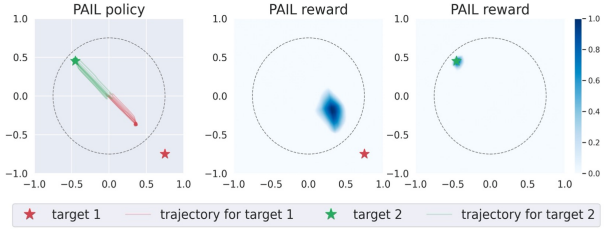
$$\tilde{\rho}_\varphi(\tau) = \left(\exp \frac{r^p_\varphi(\tau)}{\beta}\right)/Z,\ \ Z = \sum_{\tau\in\mathcal{D}^d} \exp \frac{r^p_\varphi(\tau)}{\beta}$$

Demonstration Augmentation

Augmented trajectories, which will be added into the imperfect demonstration dataset, are sampled from replay buffer by:

$$P^{\text{aug}}_\varphi(\tau) = \left(\exp \frac{r^p_\varphi(\tau)}{\beta^{\text{aug}}}\right) / \left(\sum_{\tau\in\mathcal{B}_{[-M:]}} \exp \frac{r^p_\varphi(\tau)}{\beta^{\text{aug}}}\right)$$

Visualization of Grid World



Mujoco & DMC Benchmarks

Task & Dataset	PEBBLE	BC-PEBBLE	AILP	PAIL
Ant-v2, L	0.31 ± 0.00	0.31 ± 0.00	0.33 ± 0.01	**0.80 ± 0.01**
Ant-v2, M	0.31 ± 0.00	0.32 ± 0.00	0.38 ± 0.03	**0.88 ± 0.01**
Ant-v2, H	0.31 ± 0.00	0.32 ± 0.00	0.39 ± 0.08	**0.94 ± 0.00**
HalfCheetah-v2, L	0.55 ± 0.08	**0.75 ± 0.03**	0.30 ± 0.01	0.67 ± 0.03
HalfCheetah-v2, M	0.55 ± 0.08	**0.79 ± 0.02**	0.43 ± 0.05	0.71 ± 0.04
HalfCheetah-v2, H	0.55 ± 0.08	**0.89 ± 0.01**	0.44 ± 0.10	**0.89 ± 0.01**
Hopper-v2, L	0.29 ± 0.02	0.22 ± 0.05	0.29 ± 0.07	**0.91 ± 0.02**
Hopper-v2, M	0.29 ± 0.02	0.26 ± 0.02	0.27 ± 0.03	**0.94 ± 0.07**
Hopper-v2, H	0.29 ± 0.02	0.28 ± 0.05	0.37 ± 0.12	**0.94 ± 0.01**
Humanoid-v2, L	0.05 ± 0.00	0.04 ± 0.01	0.03 ± 0.02	**0.97 ± 0.01**
Humanoid-v2, M	0.05 ± 0.00	0.06 ± 0.01	0.01 ± 0.01	**1.01 ± 0.01**
Humanoid-v2, H	0.05 ± 0.00	0.06 ± 0.01	0.02 ± 0.02	**1.08 ± 0.01**
Walker2d-v2, L	0.08 ± 0.02	0.08 ± 0.03	0.26 ± 0.03	**0.56 ± 0.12**
Walker2d-v2, M	0.08 ± 0.02	0.05 ± 0.02	0.32 ± 0.06	**0.80 ± 0.02**
Walker2d-v2, H	0.08 ± 0.02	0.10 ± 0.02	0.38 ± 0.06	**0.74 ± 0.16**
Average	0.26	0.3	0.28	**0.86**
cheetah_run, M	0.64 ± 0.13	**0.86 ± 0.15**	0.37 ± 0.08	**0.86 ± 0.00**
quadruped_walk, M	0.48 ± 0.08	0.64 ± 0.07	0.67 ± 0.04	**0.90 ± 0.01**
walker_walk, M	**0.96 ± 0.00**	**0.96 ± 0.02**	0.51 ± 0.12	**0.96 ± 0.00**
Average	0.69	0.82	0.52	**0.91**

Visualization of Preference Reward