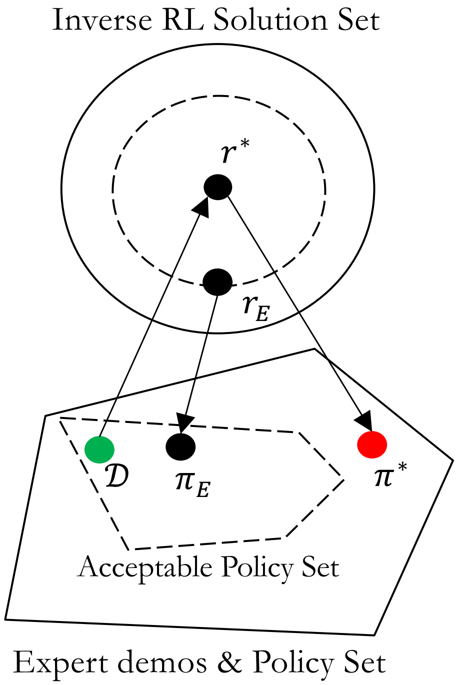
Rethinking Inverse Reinforcement Learning: from Data Alignment to Task Alignment

Weichao Zhou, Wenchao Li

Boston University DEPEND Lab

Task-Reward-Misalignment in IRL-Based IL



Inverse RL Solution Set

Acceptable Policy Set

Expert demos & Policy Set

Task: learn an acceptable π such that

$$\underline{U}_{r_E} \le U_{r_E}(\pi)$$

Standard IRL-Based IL aligns with the data

- May lead to a false r^*
- Resulting in an unacceptable policy π^*

$$\underline{U}_{r_E} \le U_{r_E}(\pi^*)$$

Question: how to learn task-acceptable policies?

PAGAR-Based Imitation Learning

Candidate reward function set

$$R_{E,\delta} = \{r: J_{IRL}(r) \leq \delta\}$$

Policy Training + Adversarial Reward Search:

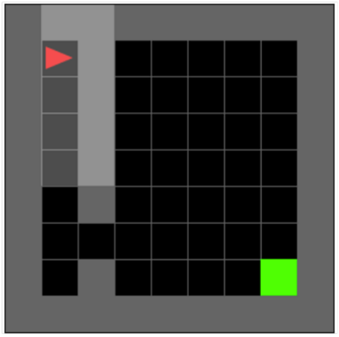
$$arg\,\underset{\pi_P}{min}\ \underset{r \in R_{E,\delta}}{max}\ U_r(\pi_P) - \max_{\pi_A} U_r(\pi_A)$$

π_P: Protagonist Policy

π_A: Antagonist Policy

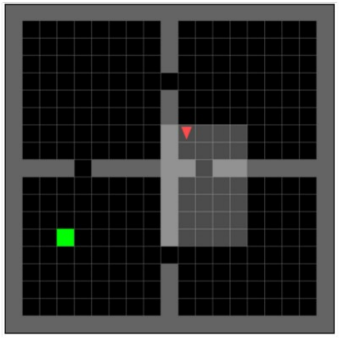
Mitigating Misalignment in Non-Ideal Learning Environments

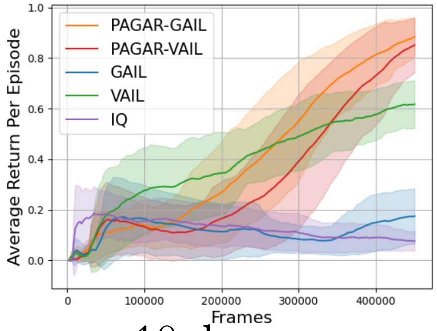
Demo @



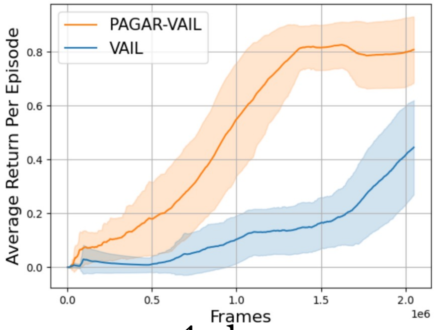
Dynamics Change

Learn @

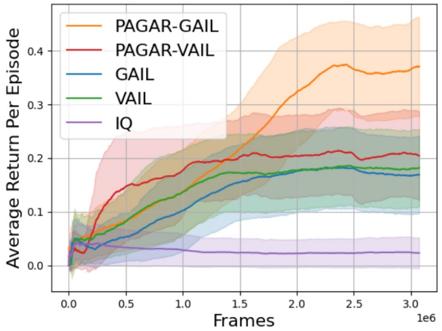




10 demos



1 demo



- Human expert overlooks the whole environment
- Limited demos
- Demo in one environment and imitation in another

Takeaways

- Task Alignment in IRL-based IL
- Protagonist Antagonist Guided Adversarial Reward (PAGAR)
- Practical Implementation