Adversarially Robust Imitation Learning

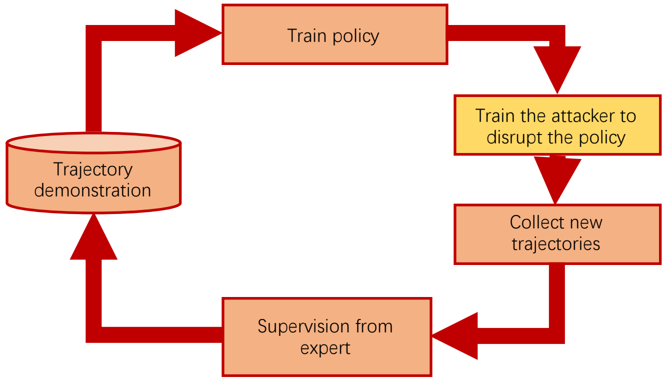
Jianren Wang[1], Ziwen Zhuang[2,3], Yuyang Wang[1], Hang Zhao[2,4]
Carnegie Mellon University, Qi Zhi Institute, ShanghaiTech University, Tsinghua University

CoRL 2021 London

Problem

- DNNs can be easily fooled by subtle noise added to the input in imitation learning, which is even non-detectable by humans.
- In real robots, sensos unavoidably contain uncertainty that naturally originates from sensor errors or equipment inaccuracy.
- In Dagger-styled imitation learning, the learning agent is especially vulnerable to attacks and can struggle to recover from errors.
- How can we design a new attacker which can
  1. Be general enough such that it requires less hand-crafted hyperparameters.
  2. Improve the learning agent's robustness

We formulate the problem as a zero-sum game and learn an adversary that perturbs the transition dynamics.

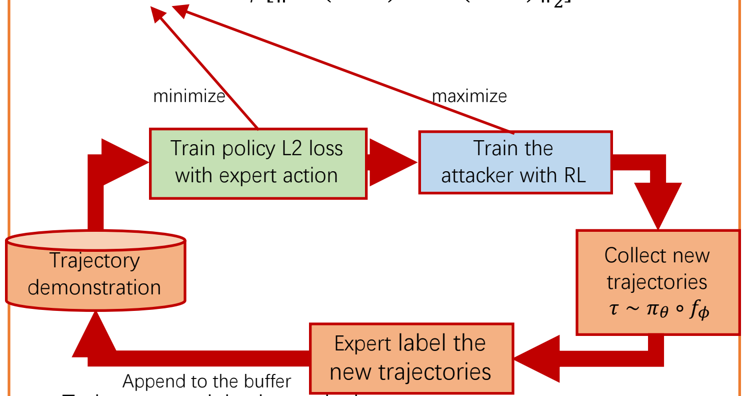


Our Paper and supplementary

Method

- Expert policy π^e
- Student policy $\pi_\theta : \mathcal{S} \to \mathcal{A}$
  pretrained by expert demonstration
- Attacker $f_\phi : \mathcal{S} \to \mathcal{S}$

- Assuming expert robustness:
  $$\pi^e(s) = \pi^e\left(f_\phi(s)\right)$$
  Such that expert decision will not be affected by the attacker
- The competition is set between
  $$\mathcal{J} = \mathbb{E}_{\tau\sim\pi_\theta\circ f_\phi}\left[\left\|\pi^e\left(f_\phi(s)\right) - \pi_\theta\left(f_\phi(s)\right)\right\|_2\right]$$



- Train two models alternatively

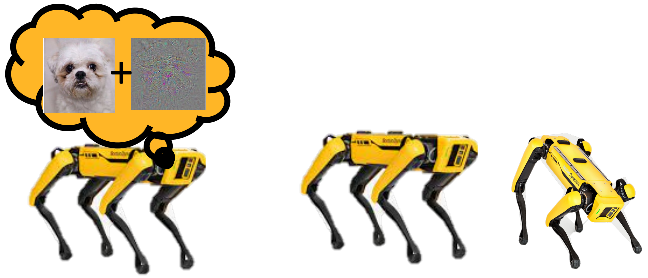
Student Policy Robustness

Theorem: As $T \to \infty$, our algorithm outputs a policy $\pi_{i*} \in \{\pi_t\}_{t=0}^{T}$ such that

$$\max_{f\in\mathcal{F}} \mathbb{E}_{s\sim d_{\pi_{i*}\circ f}} \mathbb{E}_{a\sim\pi_{i*}\circ f(s)}[\|a - \pi^e(s)\|_2] \le \epsilon_{rl}$$
with $\epsilon_{rl} \in \mathbb{R}^+$

Please refer to our paper for more details

Two types of attack

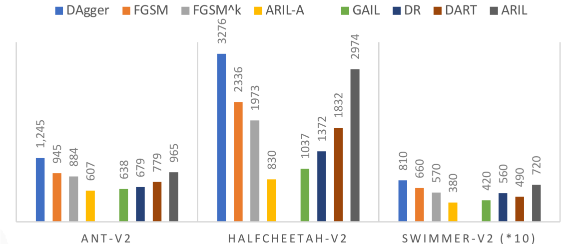


Sensory Attack

Physical Attack

Experiments

ARIL SENSORY ATTACK AND DEFENSE



ARIL PHYSICAL ATTACK AND DEFENSE