White-Box Adversarial Policies in Deep Reinforcement Learning

Stephen Casper, Taylor Killian, Dylan Hadfield-Menell, Gabriel Kreiman
Adversaries in Supervised Learning

Goodfellow et al. (2014)
Adversaries are interesting in RL
White box adversaries: “reading the target agent’s mind”

Why?

- Identifying weaknesses with attacks
- Improving robustness with robust adversarial reinforcement learning
Attacks

Two-Player Gfootball Env.

$S_t$
Attacks: better initial and asymptotic performance

White-Box Adversaries, n=20

Net Pts. / Game

Act/Val/Latent v. Ctrl p: 2e-05
Latent v. Ctrl p: 1e-05
Act/Val v. Ctrl p: 0.00638
Robust Adversarial Reinforcement Learning

Single-Player Mujoco Env.

Dense

$\Delta_t^{adv}(A)$

$\gamma_t^{adv}$

$S_t$

$\ell_t$

$\Delta_t^{vict}(A)$

$\gamma_t^{vict}$

$\alpha_t^{adv}$

$\alpha_t^{tgt}$

$Env$
Robust Adversarial Reinforcement Learning

![Graphs showing performance of HalfCheetah and Hopper models under varying friction and mass multipliers.](image)
So what?

1. White box adversarial attacks are more effective.

2. Leveraging model internals for diagnostic/debugging tasks is useful.
Thanks!

Taylor Killian

Dylan Hadfield-Menell

Gabriel Kreiman

Correspondence: scasper@mit.edu