Robust Deep Reinforcement Learning against Adversarial Behavior Manipulation

ICLR2026

Shojiro Yamabe[1], Kazuto Fukuchi[2,3], Jun Sakuma[1,3]

[1] Institute of Science Tokyo, [2] University of Tsukuba, [3] RIKEN AIP

Motivation

- Deep Reinforcement Learning (DRL) agents are vulnerable to adversarial attack, like classification model
- In recent years, applications have expanded to mission-critical tasks
  - Alignment of LLM [1]: Safety alignment to avoid harmful outputs
  - Autonomous driving [2]
  - Factory automation [3]
- As these developments progress, the importance of reinforcement learning security has increased

[1] B Ravi Kiran, et al. Deep reinforcement learning for autonomous driving: A survey. IEEE Transactions on Intelligent Transportation Systems, 2021.
[2] Long Ouyang, et al. Training language models to follow instructions with human feedback. NeurIPS, Vol. 35, pp. 27730–27744, 2022.
[3] Jonas Degrave, et al. Magnetic control of tokamak plasmas through deep reinforcement learning. Nature, 2022.

Motivation

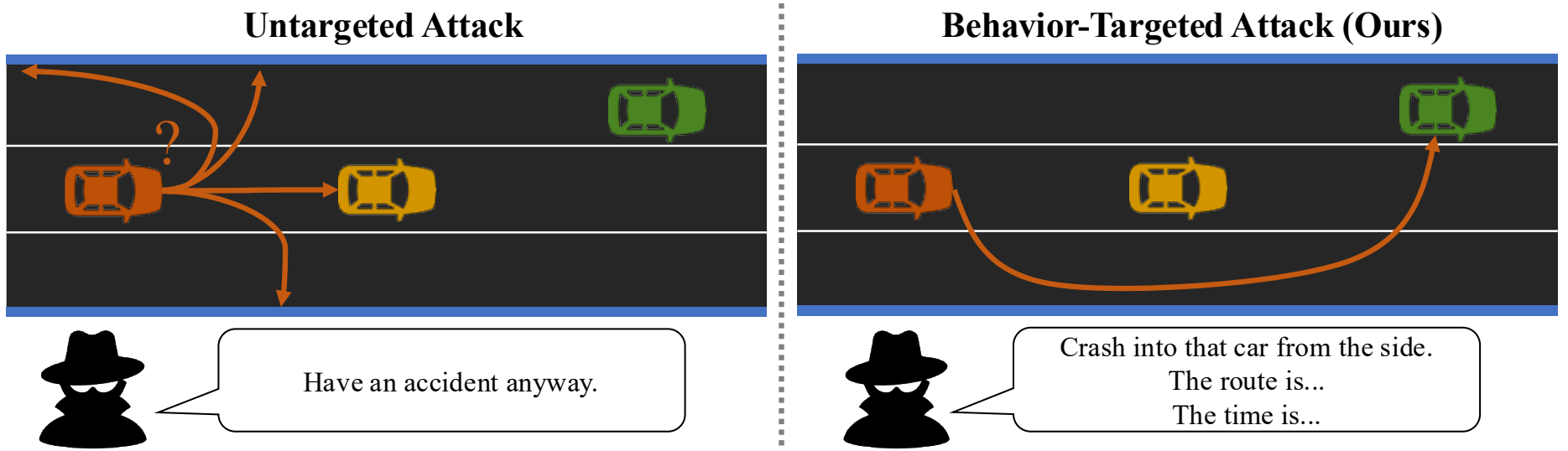
- Many studies have explored adversarial attacks and defenses in reinforcement learning
- However, most prior work focuses on reward-minimization attacks [4,5,6,7]
  - which aim to degrade the victim's performance by reducing its reward.
- But attackers may have other goals.
  - E.g., they may force a self-driving car to detour through a specific store
  - E.g., they may manipulate a recommender system to promote beneficial products.
- These goals are defined **independently of** the victim's reward.

We therefore focus on the **victim's behavior** itself, rather than reward

[4] Sandy Huang, Nicolas Papernot, Ian Goodfellow, Yan Duan, and Pieter Abbeel. Adversarial attacks on neural network policies. *ICLR*, 2017.
[5] Anay Pattanaik, Zhenyi Tang, Shuijing Liu, Gautham Bommannan, and Girish Chowdhary. Robust deep reinforcement learning with adversarial attacks. IFAAMAS, 2018.
[6] Zhang, Huan, et al. "ROBUST REINFORCEMENT LEARNING ON STATE OBSERVATIONS WITH LEARNED OPTIMAL ADVERSARY.", *ICLR 2021*.
[7] Sun, Yanchao, et al. "Who Is the Strongest Enemy? Towards Optimal and Efficient Evasion Attacks in Deep RL." , ICLR 2022

Behavior-targeted attack

- In this work, we propose the **behavior-targeted attack** and its **countermeasure** by intervening the agent's observations
  - Intervening observation amounts to attacking an autonomous vehicle's LiDAR or cameras
  - Behavior-targeted attacks aim to force the victim to behave specified by the adversary
  - A planned attack in which the location and timing of the accident are specified
- This poses a threat of far more sophisticated and stealthy attacks than untargeted attack whose only goal is to only degrade performance



Main contribution (Attack)

- We first integrate **imitation learning (IL)** to train adversarial policy
  - IL learns a policy by observing expert demonstrations and mimicking them
  - E.g., learning safe and efficient driving skills from data collected from expert drivers [8]
- Our approach overcomes the limitations of previous studies on behavior-targeted attack [9,10]
  - It relies on environmental heuristics (e.g., it can only be applied to autonomous vehicles).
  - It assume white-box access to the victim's policy

Contribution

Propose Behavior-Imitation Attack (BIA) which is applicable under
Not relies on environmental heuristics and **Limited access to the victim**

[8] Hawke, Jeffrey, et al. "Urban driving with conditional imitation learning." *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020.
[9] Leonard Hussenot, et al. Copycat: Taking control of neural policies with constant attacks. In AAMAS, 2020.
[10] Adith Boloor, et al. Attacking vision-based perception in end-to-end autonomous driving models. Journal of Systems Architecture, Vol. 110, p. 101766, 2020.

Main contribution (Defense)

- We formulate the defender's objective with attacker's gain R_{tgt}

$$\arg\min_{\pi} J_{\text{def}}(\pi) = \underbrace{-J_{\text{RL}}(\pi)}_{\text{Standard RL objective (Reward Maximization)}} + \lambda \underbrace{\left(\max_{\nu} \mathbb{E}^{M}_{\pi \circ \nu}[R_{\text{tgt}}(s,a)] - \mathbb{E}^{M}_{\pi}[R_{\text{tgt}}(s,a)] \right)}_{\text{Worst-case attacker's gain}},$$

- Our theoretical analysis reveals two important insights:
  - **suppressing the sensitivity of the policy's action outputs to state changes enhances robustness against attacks → Policy Smoothing is effective!**
  - achieving lower sensitivity during the early stages of a trajectory significantly improves overall robustness

Contribution

Propose the first robust training method against behavior-targeted attack, **Time-Discounted Robust Training (TDRT)**

Attack Results

- Environment: Meta-World [11]
- The further to the right, the stronger the assumptions imposed on the attacker
- Our method (BIA) achieves strong attack performance even under these stronger attacker assumptions

Adv Task	Target Reward	Attack Rewards (↑)					
		Adversary with full knowledge ←					→ Adversary with limited knowledge
		Targeted PGD (white-box, target policy)	**Rew Max (PA-AD)** (white-box, reward function)	**Rew Max (SA-RL)** (black-box, reward function)	**BIA-ILfD (ours)** (black-box, demonstrations)	**BIA-ILfO (ours)** (no-box, demonstrations)	**Random** (no-box, no knowledge)
window-close	4543 ± 39	1666 ± 936	4255 ± 300	**4505 ± 65**	3962 ± 666	4036 ± 510	947 ± 529
window-open	4508 ± 121	515 ± 651	493 ± 562	506 ± 444	**566 ± 523**	557 ± 679	322 ± 261
drawer-close	4868 ± 6	2891 ± 150	3768 ± 1733	4658 ± 747	**4760 ± 640**	4626 ± 791	1069 ± 1585
drawer-open	4713 ± 16	953 ± 450	**1607 ± 355**	1499 ± 536	1556 ± 607	1445 ± 610	841 ± 357
faucet-close	4754 ± 15	1092 ± 192	1241 ± 501	**3409 ± 652**	3316 ± 648	3041 ± 502	897 ± 171
faucet-open	4544 ± 800	2541 ± 86	1420 ± 85	1448 ± 64	**3031 ± 1493**	2718 ± 1293	1372 ± 81
handle-press-side	4546 ± 721	1994 ± 1225	**4726 ± 175**	4625 ± 175	4631 ± 408	4627 ± 586	1865 ± 1340
handle-pull-side	4442 ± 732	2198 ± 1524	2065 ± 1501	3617 ± 1363	**4268 ± 740**	4193 ± 517	1426 ± 1617
door-lock	3845 ± 79	640 ± 664	763 ± 768	1937 ± 1186	**2043 ± 1229**	1906 ± 1045	589 ± 494
door-unlock	4690 ± 33	531 ± 61	3295 ± 1111	**3421 ± 974**	3336 ± 932	3123 ± 1123	391 ± 59

[11] Yu, Tianhe, et al. "Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning." *Conference on robot learning*. PMLR, 2020.

Defense Results

- Adversarial training, which is effective against reward-minimization attacks, is ineffective here.
- Defense methods that aim to smooth the policy achieve strong robustness.

Adversarial Training

Policy Smoothing

Task	Best Attack Reward (↓)				
	PPO (No defense)	ATLA-PPO (AdvTraining)	PA-ATLA-PPO (AdvTraining)	SA-PPO (Smoothing w/o time-discounting)	TDRT-PPO (ours) (Smoothing w/ time-discounting)
window-close	4505 ± 65	4270 ± 188	4041 ± 96	485 ± 61	**482 ± 3**
window-open	566 ± 523	586 ± 649	671 ± 589	272 ± 37	**254 ± 214**
drawer-close	4760 ± 640	4858 ± 6	4868 ± 3	**4 ± 2**	4860 ± 4
drawer-open	1556 ± 607	1158 ± 1026	954 ± 219	403 ± 49	**378 ± 10**
faucet-close	3409 ± 652	4108 ± 790	4012 ± 123	**1559 ± 406**	1789 ± 610
faucet-open	3031 ± 1493	4383 ± 449	2358 ± 976	**1763 ± 255**	1942 ± 261
handle-press-side	4726 ± 175	4302 ± 799	3318 ± 1539	**1888 ± 1169**	1928 ± 736
handle-pull-side	4268 ± 740	532 ± 534	512 ± 982	10 ± 1	**7 ± 1**
door-lock	2043 ± 1229	1020 ± 805	992 ± 19	**478 ± 7**	487 ± 11
door-unlock	3421 ± 974	3277 ± 1265	2806 ± 1437	787 ± 1001	**691 ± 356**

Overall, we hope that this work will pave the way for a new direction in reinforcement learning safety research, shifting the focus from reward to agent behavior.

Thank you!