Demo: Vulnerability Analysis for STL-Guided Safe Reinforcement Learning in Cyber-Physical Systems

Shixiong Jiang*
University of Notre Dame
sjiang5@nd.edu

Mengyu Liu*
University of Notre Dame
mliu9@nd.edu

Fanxin Kong
University of Notre Dame
fkong@nd.edu

Abstract—**Cyber-Physical Systems(CPS) are the integration of sensing, control, computation, and networking with physical components and infrastructure connected by the internet. The autonomy and reliability are enhanced by the recent development of safe reinforcement learning (safe RL). However, the vulnerability of safe RL to adversarial conditions has received minimal exploration. In order to truly ensure safety in physical world applications, it is crucial to understand and address these potential safety weaknesses in learned control policies. In this work, we demonstrate a novel attack to violate safety that induces unsafe behaviors by adversarial models trained using reversed safety constraints. The experiment results show that the proposed method is more effective than existing works.**

I. Introduction

Reinforcement learning demonstrates efficacy in resolving control problems, but it fails to incorporate safety for systems operating in real-world environments [1]. To address this issue, safe reinforcement learning (safe RL) has been developed and has successfully enhanced exploration safety for cyber-physical systems in recent years [2]. Safe RL researchers approach the safety problem in two major ways: one thread considers it an optimization problem that requires the user to have a mathematical model of the system. The other thread considers it a constrained Markov decision process (CMDP) problem that employs formal specifications such as signal temporal logic (STL) to maximize the probability of satisfying defined safety constraints [3] [4].

However, we argue that a formal specification-guided policy is not truly safe because it has vulnerabilities that a malicious adversary can exploit to violate its safety constraint. In other words, an adversary can make the system violate the safety constraints by maliciously manipulating the sensor values while the attack remains stealthy. Different attack scenarios are considered depending on whether the adversary can access the safe RL policy of the system or the mathematical model of the system. We propose various attack methods for each attack scenario and evaluate the performance of the attack on OpenAI Safety Gym. Note that while attack reinforcement learning has been researched in existing works, our study is the first to consider violating the safety of the system and analyzing the vulnerability of safe RL.

This paper demonstrates that the proposed attack methods can successfully violate safety constraints, while the baseline method has less potential to do so. A brief introduction of the methodology of our method is provided and we evaluate why our attack methods can outperform the baselines.

II. Framework Design

Our work focuses on the vulnerability of formal language-guided secure RL. Through theoretical analysis, we identify a limitation in existing approaches targeting RL policy attacks, highlighting their inefficiency in compromising the safety of safe RL, which is crucial for real-world systems.

To address this gap, we propose the Safety Violation Attack (SVA), which adds perturbation to the observation (sensor) of the victim system to make it violate the safety constraints. It comprises two primary components: firstly, the computation of a malicious action causing the safety violation, and secondly, the generation of sensor attacks to manipulate the policy into taking the malicious action. In addition, the adversary attack should be stealthy to prevent the system from being noticed. We define stealthy as not only limiting the range of perturbation, but also not decreasing the observed reward.

Depending on the adversary's different knowledge, we define white box (WB) attack and black box (BB) attack that the adversary can adopt under different attack scenarios. In the WB attack, the adversary has full knowledge of the system's state transition function and control policy, so it can easily generate the observation attack. Instead, the BB attack, which trains an adversary model to compute the malicious action and a surrogate control policy, can still realize the security violation but is less efficient.

The experiment results show that the proposed SVA outperforms the existing baseline works, and the WB SVA achieves the most efficiency in violating the security of the victim system. The following section provides a detailed description of the demonstration settings and evaluates the performance of our proposed method in comparison to a baseline approach.

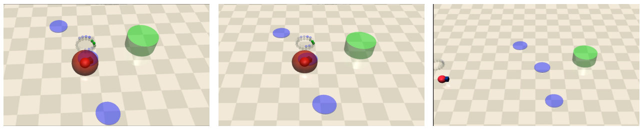
III. Result Demonstration

A. Demonstration settings

To demonstrate the effectiveness of our proposed attack, we conduct the demonstration on two benchmarks of OpenAI Safety Gym [5]: PointGoal and CarCircle to illustrate the efficacy of our proposed framework.

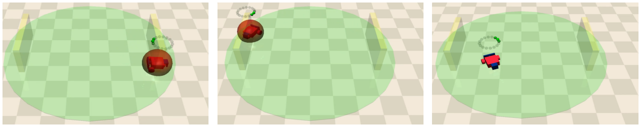
Benchmark setting: The PointGoal scenario poses a reach-avoid challenge, requiring the point to reach a goal while avoid touching the three hazards. The point has 44 states including 16 lidars to measure the distance to hazards and another 16 lidars for the goal. On the other hand, the CarCircle benchmark

* denotes equal contribution. This work is the demo of ”Vulnerability Analysis for Safe Reinforcement Learning in Cyber-Physical Systems” at the 15th International Conference on Cyber-physical Systems 2024.



(a) White box (b) Black box (c) Baseline

Fig. 1: The PointGaol benchmark with three different attack methods. The three blue circles represent hazards, and the green circle is the reaching goal. The WB and BB SVA violate the point safety while the baseline method fails to do so.



(a) White box (b) Black box (c) Baseline

Fig. 2: The CarCircle benchmark with three different attack methods. The White-box and Black-box SVA result in the car crashing into the wall, whereas the baseline method fails to induce such collisions.

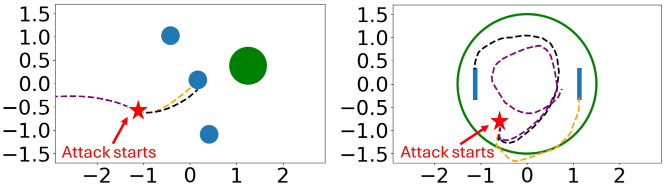
tasks a car with navigating within a circle of radius 1.5. The car must evade collisions with two walls, each spanning a length of 0.3, while maintaining velocity and staying within the circle for 5000 time steps. Safety violations occur when the car collides with a wall.

Model training: We employ Signal Temporal Logic to specify the task and constraints. Then the STL formulation is converted to a reward function for training the reinforcement learning model. We assume the two benchmarks use Proximal Policy Optimization (PPO) [6] as the system's control algorithm. Note that for the black-box scenario, the SVA framework needs a surrogate control policy and an adversary model that computes the malicious action. We utilize the Soft Actor-Critic (SAC) [7] to train the surrogate control policy and PPO to train the adversary model.

Baseline setting: We select the baseline attack method from [8] for comparison with our proposed approach. The baseline method involves training a policy with a reward function that assigns negative rewards when the origin reward is positive. Consequently, the baseline attack consistently predicts actions aimed at reducing cumulative rewards and then generates attacks that compel the system to take the actions. In the following subsection, we assess the performance of White-box and Black-box SVA, comparing their performance with the baseline method.

B. Result

Fig. 1 and Fig. 2 show the result of three attacks: white-box SVA, black-box SVA, and a baseline method on the two benchmarks. In Fig. 1, it is evident that both white-box and black-box SVAs successfully guide the point to touch the hazard (depicted in blue). Conversely, the baseline methods effectively keep the point away from the goal (depicted in green) but fail to trigger a safety violation. This observation is similarly reflected in Fig. 2 for the CarCircle benchmark, where both white-box and black-box SVAs lead the car to crash into the wall, while the baseline method does not induce such collisions. These results underscore the vulnerability of formal language-guided safe RL to sensor attacks.



(a) PointGoal (b) CarCircle

Fig. 3: Trajectories of the agent under various attack methods: White-box SVA (orange), Black-box SVA (black), and the baseline method (purple). The attacks initiate from identical starting positions, resulting in divergent trajectories.

Fig.3 shows the trajectory of PointGoal and CarCircle with three attack methods. Notably, the white-box SVA demonstrates the highest efficiency in compromising agent safety, resulting in unsafe trajectories. The black-box SVA also violates safety but with lower efficiency. In contrast, the baseline attack method lacks the intention to drive the agent into an unsafe state. These findings highlight that conventional RL attack methods, primarily focused on reducing rewards, fall short of violating safety. Conversely, our proposed SVA exposes vulnerabilities, leading to more severe consequences. Furthermore, the results suggest that the adversary's level of system knowledge correlates with the efficiency of the attack.

IV. Conclusion

This paper briefly introduces the proposed SVA framework and demonstrates the experiment settings and results. The demonstration illustrates the framework's efficacy in compromising the safety of formal language-guided safe RL.

Acknowledgement

This work was supported in part by NSF CNS-2333980.

References

[1] Gabriel Dulac-Arnold, Nir Levine, Daniel J Mankowitz, Jerry Li, Cosmin Paduraru, Sven Gowal, and Todd Hester. Challenges of real-world reinforcement learning: definitions, benchmarks and analysis. 2021.

[2] Lukas Brunke, Melissa Greeff, Adam W Hall, Zhaocong Yuan, Siqi Zhou, Jacopo Panerati, and Angela P Schoellig. Safe learning in robotics: From learning-based control to safe reinforcement learning. *Annual Review of Control, Robotics, and Autonomous Systems*, 5:411–444, 2022.

[3] Max H Cohen and Calin Belta. Model-based reinforcement learning for approximate optimal control with temporal logic specifications. In *Proceedings of the 24th International Conference on Hybrid Systems: Computation and Control*, pages 1–11, 2021.

[4] Mengyu Liu, Pengyuan Lu, Xin Chen, Fanxin Kong, Oleg Sokolsky, and Insup Lee. Fulfilling formal specifications asap by model-free reinforcement learning, 2023.

[5] Joshua Achiam and Dario Amodei. Benchmarking safe exploration in deep reinforcement learning. 2019.

[6] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017.

[7] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pages 1861–1870. PMLR, 2018.

[8] Huan Zhang, Hongge Chen, Duane Boning, and Cho-Jui Hsieh. Robust reinforcement learning on state observations with learned optimal adversary. *arXiv preprint arXiv:2101.08452*, 2021.