Poster: Robustness of Reinforcement-Learning-Based Autonomous Driving to Adversarial Inputs

Ziling He[1,2], Yinan Zhao[1], Tatsuya Mori[1,2,3]
[1]Waseda University, [2]RIKEN, [3]NICT
{ziling, yinanzhao, mori}@nsl.cs.waseda.ac.jp

I. Introduction

In this study, we assess the robustness of Reinforcement Learning (RL) in Autonomous Driving (AD) systems, particularly against adversarial attacks. We adopted the Q-learning based AD model proposed by Karavolos et al. [1] for its simplicity, serving as the foundation of our analysis. This choice allows us to draw a distinct comparison between the straightforward Q-learning approach and the more complex RL systems.

We design two threat models to simulate adversarial attacks on RL-based AD systems. The first model involves injecting undetectable malicious code during the RL model's fine-tuning, making it susceptible to adversarial perturbations that could lead to collisions under specific trigger conditions. The second threat model aims to induce a collision by directly altering the RL model's action decision under specific trigger conditions, representing a more stealthy approach.

Based on these threat models, our empirical investigation focuses on two primary scenarios: manipulation of sensor inputs and direct perturbation of actions. The findings indicate that while RL-based AD systems demonstrate resilience against sensor input manipulation, they exhibit vulnerabilities when subjected to direct action perturbations. The primary and realistic scenario involves changing sensor readings, like during off-center turns, which can mislead the system and potentially lead to accidents. This is crucial for maneuvers where small errors are significant. The second scenario, directly perturbing actions, serves more as a theoretical investigation into RL-based AD systems' vulnerabilities rather than a practical, real-world threat.

II. Attack Design

This section aims to detail the design of two attack strategies targeting a RL-based autonomous driving system.

A. Threat Model 1

A schematic of Threat Model 1 is shown in Fig. 1.

Goal: To inject undetectable malicious code during the RL model's fine-tuning, causing the model to be susceptible to an adversarial perturbation attack, leading to a collision under specific trigger conditions. When the trigger is triggered, the sensor data x is affected by the perturbation Δ and changes from x to x', i.e., $x' = x + \Delta$.

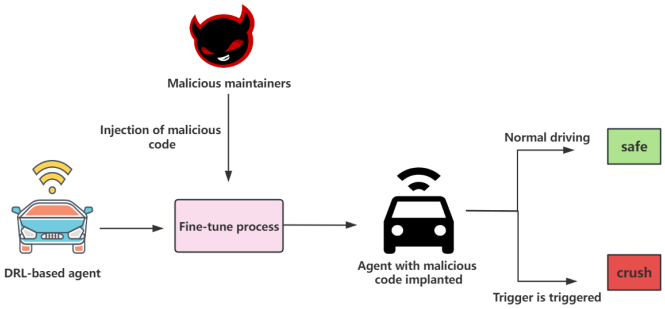


Fig. 1. A schematic of Threat Model 1

Knowledge and Capabilities: Assuming the attacker is a malicious model maintainer with comprehensive knowledge of the victim model, including its structure and parameters, constituting a white-box attack. The attacker can access and manipulate the model's inputs.

B. Threat Model 2

A schematic of Threat Model 1 is shown in Fig. 2.

Goal: To induce a collision by directly altering the RL model's action decision under specific trigger conditions, aiming for a stealthier attack. Specifically, the action a is modified to a', an alteration that impairs the performance of the RL-based agent, potentially leading to a collision.

Knowledge and Capabilities: The attacker knows the victim model's type but not its specific parameters. They can access the model's state inputs and action outputs, modifying the action when the trigger is activated.

C. Attack Scenarios

Based on the two threat models proposed in A and B, we have meticulously developed two distinct attack scenarios targeting RL-based Autonomous Driving (AD) systems, each corresponding to one of the threat models:

1. Manipulation of Sensor Readings (Corresponding to Threat Model 1): This scenario, which aligns with Threat Model 1, represents a practical and realistic threat. It involves

Network and Distributed System Security (NDSS) Symposium 2024
26 February - 1 March 2024, San Diego, CA, USA
ISBN 1-891562-93-2
https://dx.doi.org/10.14722/ndss.2024.23004
www.ndss-symposium.org

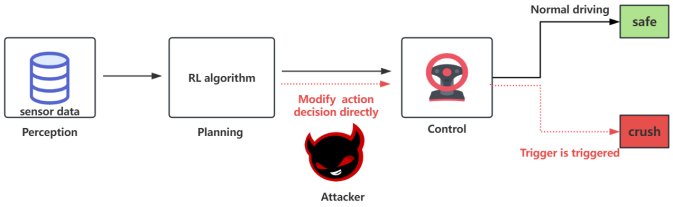


Fig. 2. A schematic of Threat Model 2

strategically altering sensor readings, such as those during off-center turns, which can effectively mislead the AD system. This manipulation is particularly critical in scenarios where even minor deviations can have significant consequences, potentially leading to erroneous decisions and, in worst-case scenarios, resulting in accidents. The alignment with Threat Model 1 is evident as this scenario exploits the vulnerabilities introduced by the undetectable malicious code injected during the model's fine-tuning.

2. Direct Alteration of Actions (Corresponding to Threat Model 2): The second scenario, aligning with Threat Model 2, delves into a more theoretical realm, focusing on the direct perturbation of the actions determined by the RL system. This approach is less about immediate real-world applicability and more about probing the inherent vulnerabilities within RL frameworks. By directly influencing the action outputs, this scenario, which is a direct manifestation of Threat Model 2, provides valuable insights into how subtle, calculated alterations can impact the decision-making process of the AD system, potentially leading to unintended outcomes.

Attack Trigger Mechanism. In our attack process, a trigger mechanism is activated based on preset thresholds for the d (lateral deviation) and θ (orientation) sensor readings, consistent with our threat model scenarios. For example, the trigger is activated when the vehicle turns left and is on the left side of the lane ($\theta < -\alpha$ and $d < -\beta$) or when it turns right and is on the right side of the lane ($\theta > \alpha$ and $d > \beta$), where α and β are predetermined values ranging from 0 to 1. In both attack scenarios, the triggering mechanism is the same.

Injecting Perturbations. In each scenario, the perturbations are continuously applied as long as the trigger conditions are met. In the first attack scenario, once triggered, we apply a predetermined amount of perturbation to the sensor readings. This results in a deliberate distortion of the data received by the vehicle. The manipulated sensor inputs mislead the vehicle, causing it to deviate from its intended path and subsequently increasing the risk of a collision. Note that while we have not necessarily optimized this attack specifically for RL, we do introduce relatively large perturbations to the range of sensor readings. The second scenario focuses on directly modifying the actions determined by the RL system; it alters the vehicle's steering decision to a neighboring level.

III. Preliminary Experiments

Setup. The attack test is conducted using the TORCS simulator on a standard course without extreme sharp curves. Each run lasts 30 seconds with a speed limit of 120 km/h. The RL model employed is the same as in the study by Karavolos et al. [1]. Two triggers are set; trigger 1: $\alpha = 0.1, \beta = 0.2$ and trigger 2: $\alpha = 0.1, \beta = 0.1$.

Scenario 1. In our adversarial input experiments, we injected noise of sizes $\varepsilon = 0.1$ and $\varepsilon = 0.3$ into the sensor readings. These noises were applied directly to both angular and position sensors, specifically in a direction that increases the likelihood of collision. In both cases, the RL-based autonomous driving system remained unaffected, demonstrating its robustness to perturbations in sensor readings.

The success rate of the three attacks was close to zero across all parameter sets. This result is mainly due to the configuration of the Q-learning-based RL we adopted [1]. In this referenced setting, the steering angle is discretized into five values: 0.5, 0.1, 0, -0.1, -0.5. Because of this granularity, small perturbations in the sensor values were rendered ineffective. The evaluation of RL systems with continuous action spaces is an important direction for future research.

Scenario 2. In the second scenario, we targeted the same RL-based AD model by altering its actions to select neighboring actions. This attack, performed with either Trigger 1 or Trigger 2, was repeated 50 times on the simulation. The experimental results showed a success rate of approximately 60% using Trigger 1 and 78% using Trigger 2. These results indicate that larger adversarial inputs that directly affect the actions can increase the probability of a successful attack. In addition, more frequent activation of the triggers increases the overall success rate of the attacks.

IV. Conclusion

In this study, we evaluated the robustness of the Q-learning-based AD system against adversarial attacks. Two threat models were designed, focusing on sensor input manipulation and direct action alteration. Preliminary experiments using the TORCS simulator demonstrated the system's resilience to sensor perturbations but revealed vulnerabilities to direct action modifications, highlighting areas for future research in RL system security.

Future Research Directions. Our future work will focus on conducting an extensive assessment of vulnerabilities in RL-based autonomous driving systems. We aim to identify specific conditions under which these systems are most susceptible to attacks. Additionally, we plan to explore the design of effective adversarial inputs tailored to various RL algorithms used in autonomous driving systems. This research will contribute to enhancing the security and reliability of RL implementations in real-world automotive applications.

Acknowledgment

A part of this work was supported by JSPS KAKENHI 22S0604 and JST CREST JPMJCR23M4.

References

[1] D. Karavolos, "Q-learning with heuristic exploration in Simulated Car Racing," *[Online]*, 2013. Available: https://api.semanticscholar.org/CorpusID:29338425

Poster: Robustness of RL-Based Autonomous Driving to Adversarial Inputs

Ziling He[1,2], Yinan Zhao[1], Tatsuya Mori[1,2,3]

[1]Waseda University, [2]RIKEN, [3]NICT

Abstract

This study evaluates the robustness of the Q-learning-based Autonomous Driving(AD) system against adversarial attacks. We developed two threat models: one injecting undetectable code affecting sensor inputs, and another altering action decisions. Experiments conducted using the TORCS simulator revealed that while reinforcement learning(RL) systems are resilient to sensor input manipulation, they are vulnerable to direct action perturbations. Our findings highlight the need for further research into the security of RL systems, especially in continuous action spaces, to enhance their reliability and safety in autonomous driving applications.

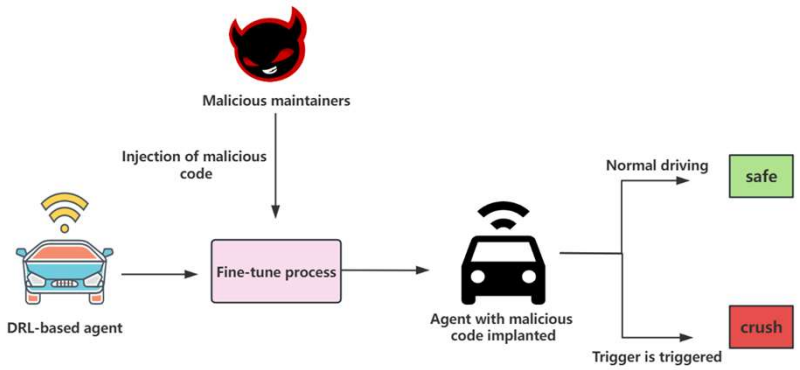
Attack Design

Threat Model 1

Goal: Inject malicious code in the RL model during fine-tuning to make it vulnerable to adversarial attacks, causing collisions under certain triggers.

Knowledge and Capabilities: The attacker, a malicious insider, fully understands the model, enabling a white-box attack. They can access and alter the model's inputs.

Scenario 1: Manipulation of Sensor Readings

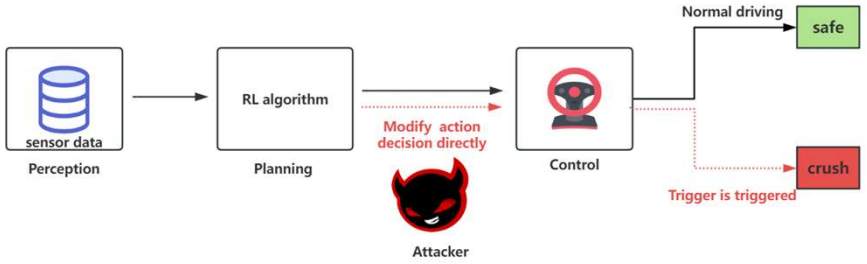


Threat Model 2

Goal: To cause a collision by stealthily changing the RL model's action from a to a' under certain triggers.

Knowledge and Capabilities: The attacker understands the model type but not its exact parameters and can modify actions based on the model's state inputs and outputs upon trigger activation.

Scenario 2: Direct Alteration of Actions



Preliminary Experiments

Setup

- The attack test is conducted using the TORCS simulator on CG Speedway number 1 track without extreme sharp curves.
- Each run lasts 30 seconds with a speed limit of 120 km/h.
- The DRL model employed is the same as the study by Karavolos [1].
- trigger 1: α=0.1, β=0.2 and trigger 2: α =0.1, β =0.1.

Results of Scenario 1

- We injected noise of magnitudes 0.1 and 0.3 into the sensor readings.
- However, in both cases, the RL-based autonomous driving system remained unaffected, demonstrating robustness to perturbations in sensor readings.
- The limited impact of attacks can be attributed to the discretized steering angle values.

Results of Scenario 2

- We targeted victim model by altering its actions to select neighboring actions.
- The experimental results showed an approximate success rate of 60% using Trigger 1 and 78% using Trigger 2.
- These results indicate that larger adversarial inputs that directly affect the actions can increase the probability of a successful attack.

Conclusion & Future Plan

Conclusion

- Q-learning AD system withstands small disturbances (Scenario 1).
- Impactful attacks must significantly alter system actions (Scenario 2).

Future Plan

- Conduct detailed vulnerability assessments on RL-based systems.
- Determine conditions for attack success.
- Develop adversarial inputs for diverse RL algorithms in autonomous driving systems.

Acknowledgment. A part of this work was supported by JSPS KAKENHI 22S0604 and JST CREST JPMJCR23M4.

[1] D. Karavolos, "Q-learning with heuristic exploration in Simulated Car Racing," [Online], 2013. Available: https://api.semanticscholar.org/ CorpusID:29338425