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

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

I. Introduction

In this study, we evaluate the robustness of reinforcement learning (RL) in autonomous driving (AD), especially against adversarial attacks. We adopted the Q-learning based AD model of Karavolos et al. [1] for its simplicity as the basis of our study. This choice provides a clear contrast between the straightforward Q-learning approach and the more complex RL system. Our empirical study of adversarial attacks on RL-based AD systems focused on two scenarios: manipulation of sensor inputs and direct perturbation of actions. The results showed that while the RL-based AD systems are robust against sensor input manipulation, they are vulnerable to direct action perturbations.

II. Attack Design

Our attack design on a RL-based Autonomous Driving system encompasses two scenarios: (1) manipulating sensor readings and (2) directly altering actions. The primary and realistic scenario involves changing sensor readings, like during off-center turns, which can mislead the system and potentially lead to accidents. The second scenario, directly perturbing actions, serves more as a theoretical investigation into RL vulnerabilities rather than a practical, real-world threat.

Attack Trigger. 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, perturbations persist while triggers are active. The first scenario applies predetermined perturbations to sensors, distorting data and causing vehicle path deviation, raising collision risks. The second scenario focuses on directly modifying the actions determined by the RL system, altering the steering decision to an adjacent level.

Symposium on Vehicles Security and Privacy (VehicleSec) 2024
26 February 2024, San Diego, CA, USA
ISBN 979-8-9894372-7-6
https://dx.doi.org/10.14722/vehiclesec.2024.25016
www.ndss-symposium.org

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.

Future Study. The Q-learning-based AD exhibited robustness to input perturbations that require substantial changes, such as direct action modification, to be affected. Identifying the conditions for successful attacks through comprehensive evaluation of RL-based AD is left for our future research.

Acknowledgement 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