Effects of Learning-Based Action-Space Attacks on Autonomous Driving Agents

Yuting Wu, Xin Lou † ‡, Pengfei Zhou††, Rui Tan, Zbigniew T. Kalbarczyk*, Ravishankar K. Iyer*
Nanyang Technological University, Singapore
† Singapore Institute of Technology,‡ Illinois at Singapore
†† University of Pittsburgh, USA
*University of Illinois at Urbana-Champaign, USA

ABSTRACT

Vehicle cybernation with increasing use of information and communication technologies faces cybersecurity threats. This extended abstract studies action-space attacks on autonomous driving agents that make decisions using either a traditional modular processing pipeline or the recently proposed end-to-end model obtained via deep reinforcement learning (DRL). The action-space attacks alter the actuation signal and pose direct risks to the vehicle's behavior. We formulate the attack construction as a DRL problem based on the input from either an extra camera or inertial measurement unit deployed. Attacks are designed to lurk until a safety-critical moment arises (e.g. lane changing or overtaking), with the goal of causing a side collision upon activation. Our results demonstrate that the modular processing pipeline is more resilient than the DRL-based agent, due to the former's main focus of trajectory following. We further investigate two enhancement methods: adversarial training through fine-tuning and progressive neural networks, gaining an essential understanding of their pros and cons.

1 INTRODUCTION

Recent rapid growth in autonomous driving (AD) has brought research attention to its cybersecurity concerns. Rising autonomy results in more sensors and connectivity, thereby expanding potential attack targets in AD. Among miscellaneous possible attack mount points, targeting the actuation of a vehicle is appealing to the attacker. Adversaries can bypass potential defense mechanisms and directly affect the vehicle's state. However, action-space attacks, also referred to as actuator attacks, have gained limited attention in the context of AD. Most recent studies on action-space attacks in AD rely on model-based approaches that either require in-vehicle data for the current system state [6] or vehicle's kinematics and structure [2], resulting in a demanding form of white-box attacks. Meanwhile, attacks in the black-box setting have been mostly studied in simulation environments like OpenAI Gym [3] and Mathworks [5], which are not representative of real-world

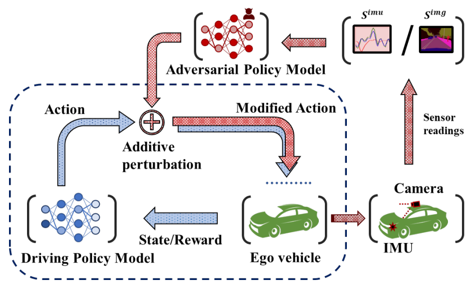

Figure 1: Overview of the DRL-based action-space attack.

driving conditions. Additionally, these studies usually concentrate on a single type of driving system without contrasting the impact of the attack across various AD designs. In light of this, we study the action-space attack on two major types of driving agents: 1) the traditional modular processing pipeline and 2) the end-to-end policy model trained via DRL. To ensure the realism of the attack, we assume the attacker has no access to (i) the driving agent's internal, and (ii) the driving agent's sensor readings. Both driving agents are formulated based on a trajectory-following task while adopting different design methodologies. We hypothesize that the design differences between the two agents will lead to distinct characteristics in responding to action-space attacks. We further apply adversarial training to enhance the DRL-based end-to-end driving agent with two variants, fine-tuning and progressive neural networks (PNN).

2 METHODOLOGY

We treat the entire driving system as a black box, utilizing DRL to investigate safety-critical moments and to learn how to introduce disturbances, as depicted in Fig. 1. In our study, the attacker utilizes either an extra camera or an inertial measurement unit (IMU) to identify safety-critical moments in the driving system. The former provides adequate information while its installation demands a wide field of view, which may attract attention from humans. The latter provides a less informative inertia trace but can be concealed within the vehicle, making them nearly undetectable. To explore IMU's potential utilization by attackers, we proposed a 'learning-from-teacher' structure to transfer the attack policy from obtained camera-based attacks to IMU-based attacks. The action-space attack injects additive perturbations into the steering angle of the ego vehicle at safety-critical moments (i.e., lane changing and overtaking), aiming to create a side collision with another vehicle on road. The attack is subjected to an attack budget that characterizes the actuation system's logistic constraint (i.e., the maximum allowed adjustment value per actuation step) or the desired degree of attack stealthiness. We conduct two variants of adversarial training to

[1]This project is supported by the National Research Foundation, Singapore, and the National University of Singapore through its National Satellite of Excellence in Trustworthy Software Systems (NSOE-TSS) office under the Trustworthy Computing for Secure Smart Nation Grant (TCSSNG) award no. NSOE-TSS2020-01, and in part by the National Research Foundation, Prime Minister's Office, Singapore under its Campus for Research Excellence and Technological Enterprise (CREATE) program.
 ACM ISBN 979-8-4007-0036-1/23/05. https://doi.org/10.1145/3576841.3589615