Skysong Innovations
Advancing the Arizona State University Knowledge Enterprise

1475 N. Scottsdale Road, Suite 200
Scottsdale, AZ 85287-3538
Phone: 480 884 1996 Fax: 480 884 1984

Case ID:M23-181P^
Published: 12/5/2023

Inventors

Prasanth Buddareddygari
Travis Zhang
Yezhou Yang
Yi Ren

Contact

Physical Sciences Team

Identify Targeted Attacks on Deep RL-Based Autonomous Driving

In the rapidly advancing landscape of deep reinforcement learning (RL), recent studies have highlighted vulnerabilities in control policies learned through deep RL. These vulnerabilities raise concerns surrounding adversarial attacks, especially in risk-sensitive domains such as autonomous driving. In these scenarios, the learned policies become susceptible to manipulations in both the agent's observations and the physical environment. Current threat models fall into two categories: (1) targeted attacks through real-time manipulation of the agent's observation, or (2) untargeted attacks through manipulation of the physical environment. By exploring the feasibility of such attacks and examining real-world threat scenarios, developers can more effectively identify and anticipate potential points of vulnerability in their systems. However, the looming possibility of emerging threats that combine the practicality and effectiveness of these existing models poses an imminent challenge to the risk mitigation strategies associated with the development of deep RL-based control policies.

Researchers at Arizona State University have developed an algorithm on deep RL models in the context of autonomous driving, which serves to identify potential points of attack on these systems. Instead of altering agent states or actions directly, the proposed method introduces static changes using visually learned patterns on physical objects within the environment. This approach provides evidence that an attack of the system with a continuously effective adversarial object makes it feasible for attackers to predict environment dynamics and plan actions leading to a specified target state. By understanding this novel threat model targeted towards deep RL control policies, developers can more effectively mitigate threats posed by targeted physical adversarial attacks in autonomous systems.

Related publication: Targeted Attack on Deep RL-based Autonomous Driving with Learned Visual Patterns

Potential Applications:

- Autonomous vehicles
- Deep RL-based control systems

Benefits and Advantages:

- Mitigation of targeted physical attacks
- Visually learned patterns on objects
- Demonstrates existence of new type of attack on RL models (i.e., an algorithm that searches for such attacks and shows their existence empirically)

- Improves robustness and safety of AI methods for autonomous driving