Security of the Perception in Autonomous Driving under Physical-World Adversarial Attacks

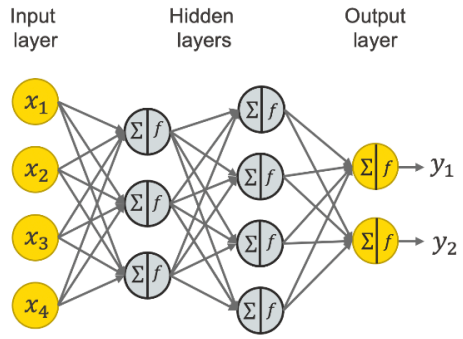
Introduction

Nowadays, autonomous driving (AD) vehicles are rapidly developed. Some of them, e.g., Google Waymo and TuSimple, are already providing services on public roads. To ensure correct and safe driving, a fundamental pillar in AD systems is perception, such as obstacle detection, traffic sign detection, lane detection, etc. With the power of deep learning algorithms, such perception tasks in AD systems widely apply deep neural network (DNN) based models. Recent works find that DNN models are generally vulnerable to adversarial examples, or adversarial attacks. Some works further explored such attacks in the physical world. For instance, in the context of AD, prior works have designed successful physical-world adversarial attacks to make the stop sign disappear. Thus, in the current stage, studying the security of the perception in AD systems under physical-world adversarial attacks is very necessary.

This project aims to contribute to the security of DNN-based perception in AD systems, such as different light conditions, design/evaluate some existing adversarial attacks to measure the model robustness in different driving conditions or demonstrate the practicality of the existing attacks in real AD vehicle (e.g., Tesla). I hope that our designs and analysis can help guide future algorithms (e.g., lane detection algorithm, object detection algorithm) designs with security and safety guarantee in AD systems.

Background

- **Deep Neural Network**



Deep neural network is represented as a layered structure of neurons that connect with other neurons. These neurons transmit the signals through[summation and activation functions to other neurons based on the inputs. Number of single structures form a complex neural network that learn with outputting a feedback mechanism.

- AS^{2}Guard **Lab Research Review**

At UC Irvine, researchers study deeply on the security of the AD perception. For instance, they are the first to study the security of Multi-Sensor Fusion (MSF) based perception and lane detection. In the meanwhile, they proposed some new metrics on DNN-based lane detection models. Such kinds of works are all published in the top-tier security/ computer vision conferences. Taking

the lane detection metric as an example, they evaluated on DNN-based lane detection models in four different approaches on two AD image datasets using three different evaluation metrics. They have identified the limitation of conventional metric based on accuracy and F1-score, which does not consider the whole AD pipeline such as the downstream task like planning and control. Thus, the results cannot represent the end-to-end security consequence. With that, two more metrics are proposed in that work -- End-to-End Lateral Deviation Metric (E2E-LD) and Per-Frame Simulated Lateral Deviation Metrics (PSLD), which can directly transfer to the system-level results of the lane detection tasks. I believe such practical security research can lead the direction of the future AD designs/evaluations.

Project Overview

Objective

The project will build upon past research. My final goal is to improve the robustness of perception of the AD system, and I will discuss the details in the following.

Project Components

- Select the perception tasks to study

In AD perception, there exists multiple tasks such as lane detection, traffic sign detection, obstacle detection, etc. Thus, the necessary first step is to find a specific task to study. With that, I can do some measurements or new designs.

- Related work study

After fixing the perception task to study, the next step is to study the related works –- physical-world adversarial attack. By doing that, I'm able to know how the prior works are designed and how it can be applied to my setting. Then, I can do some measurement study, such as measuring whether the existing work goes well under my problem context. For instance, I can evaluate state-of-the-art lane detection models with the two metrics proposed by the research group under various driving conditions to test the robustness.

- Security analysis in real AD system

Most of the existing adversarial attacks in AD perception only consider the single component of the AD system but no end-to-end evaluation. Thus, it would be better to demonstrate the existing work in a real AD system to show the practicality. The research group, they have such resources, and I would like to contribute on that part to show its practicability.

- Robustness improvement and/or new vulnerability discovery

After demonstrating the utility of the existing work, I could either improve the robustness or explore new vulnerabilities. I feel both directions could benefit the society and thus will have a very huge impact.

This project may not actually address a large security issue in AI, but it will help populations who care about autonomous vehicles aware of the powerfulness and vulnerability of deep neural networks and get familiar with some fundamental approaches to enhance the robustness of autonomous driving perception model.

Detailed Plan

This project is still in the planning phase. I have read several research papers [1, 2, 3, 4] related to the security of perception in AD systems and some papers are about the DNN-based autonomous driving perception. For knowledge preparation, I have supplemental background in probability and statistics, machine learning, and information theory, and I have learned fundamental principles of deep learning and fundamental strategies of adversarial machine learning in the past month.

To start the project, I would like to first learn the principles of the basic neural network models behind the perception and set up the environment such as different driving conditions. Then, I will follow the sequence of tasks described in project components. To perform the security analysis under different attacks, I would prepare myself with supplemental adversarial attack knowledge and comprehensively read related papers to gain a sense of what attack strategies are suitable to and have practical meaning to the existing DNN models. Specifically, I plan to try the lane detection attack [1] and stop sign attack [4] in OpenPilot Comma three [7], a real Tesla vehicle or an AD chassis shown in Fig. 1 (resource from AS^{2}Guard research group). To analyze the robustness of a model, I will not only try to find an adversarial example, but also try another way to recognize whether every input produces a correct output.

Figure 1. Resource from AS^{2}Guard group (Tesla, OpenPilot, and AD Chassis)

The first three steps are feasible to accomplish since a main part of it is based on an existing framework and what I will focus on is to set up some environments and test. The last step of

the project, which is to improve the robustness or discover new vulnerabilities, is likely to be unexpectedly difficult to accomplish, since developing a new methodology requires full understanding of how the existing methodologies are built, how they work to predict, and what design details would lead to the vulnerability. Also, I need to have state-of-the-art works in mind to start building a new methodology. In order to figure it out, I would solidify my knowledge of algorithm design and neural network principles; ask my faculty mentor or PhD advisor for help; or even perform some control experiments to find out the root causes. Alternatively, I would use strategies to build defense on the existing models in order to enhance the robustness of the models. At the end of the project, I expect to complete an evaluation on the robustness of perception in AD systems that have high performance by designing and deploying physical-world adversarial attack and expect to gain a better model that has high performance and robustness.

Timeline

Time	**Activity**
Week 1-2	1) Select and study a few concrete perception tasks, perform literature review on prior security research on them; 2) Collect and understand the corresponding codebases in order to be well prepared for concrete reproduction.
Week 3-5	1) Reproduce existing perception attack works from the literature review in real AD setup using the resources from the AS^{2}Guard group (Fig.1); 2) Test the attacks in different scenarios and attack parameters to understand their real-world attack capabilities and limitations.
Week 6-7	1) Based on the insights from the security analysis, investigate potential defense designs to improve the robustness of the perception model; 2) Prototype the designs and perform dataset-based evaluations.
Week 8-10	1) Perform more realistic evaluations by integrating the defense into the real-world AD setup using the resources from the AS^{2}Guard (Fig.1); 2) Explore potentially new security vulnerabilities of the perception in AD systems.

Budget

Proposed budgets	**Price**
AWS p3.2xlarge ($3.06/h)	$900 (30 hours/week)
Copying/printing	$100
---	---
Total	**$1,000**

Reference

[1] Pengfei Jing, Qiyi Tang, Yuefeng Du, Lei Xue, Xiapu Luo, Ting Wang, Sen Nie, and Shi Wu. Dirty Road Can Attack: Too Good to Be Safe: Tricking Lane Detection in Autonomous Driving with Crafted Perturbations. USENIX Security 21. 2021.

[2] Takami Sato and Qi Alfred Chen. Towards Driving-Oriented Metric for Lane Detection Model. CVPR, 2022.

[3] Takami Sato and Qi Alfred Chen. On Robustness of Lane Detection Models to Physical-World Adversarial Attacks. AutoSec 2022 Symposium.
https://www.ndss-symposium.org/ndss-paper/auto-draft-311/

[4] Yue Zhao, Hong Zhu, Ruigang Liang, Qintao Shen, Shengzhi Zhang, and Kai Chen. Seeing isn't Believing: Towards More Robust Adversarial Attack Against Real World Object Detectors. ACM CCS 2019

[5] Sumit Saha. A Comprehensive Guide to Convolutional Neural Networks — the ELI5 way. Towards Data Science. 2018.
https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53

[6] AS^{2}Guard Website. AD&CV Systems Security
https://sites.google.com/view/cav-sec

[7] OpenPilot, https://github.com/commaai/openpilot