Università degli Studi di Salerno
Dipartimento di Informatica

Tesi di Laurea di I livello in
Informatica

Adversarial Attacks on Vision-based Deep Neural Networks in Autonomous Driving Vehicles

Relatore
Giuseppe Scanniello
Correlatore
Dott. Nome Cognome

Candidato
Michelangelo Esposito

Academic Year 2021-2022

Abstract

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Contents

List of Figures

List of Acronyms and Abbreviations

Chapter 1

Introduction

Things to add:

- where different types of coverage are useful
- Add formulas for coverage criteria

Chapter 2

Problem formulation

2.1 Autonomous Driving Vehicles

The **Automatic Land Vehicle in Neural Networl**, ALVINN, was the first self-driving car ever proposed, in 1989; it was based on neural networks responsible to detect lines, segment the environment, and drive the car. While the general principles on which it was based worked well, the limited computed capabilities at the time, didn't make the solution take off. Furthermore, the absence of data meant that it was extremely difficult to gather the necessary samples and datasets on which to base the models that would manage vision and driving.

LiDAR stands for Light Detection And Ranging. It's a method to measure distance of object by firing a focused laser beam and measuring the time it takes for it to bounce back at the source after being reflected by something. Compared to traditional camera images, LiDAR data can provide additional information about the surrounding scene...

A self-driving vehicle cannot rely on LiDAR alone, however. While this technology works great in most environments, including dark areas, if the scene surrounding the car becomes more noisy, due to rain or fog for example, the LiDAR sensor can become imprecise or fail.

For this reason, a third sensor is usually employed, the RADAR, which scans the surrounding area based on radio waves...

2.2 3D Object detection

The most suited category of neural networks for image processing are Convolutional Neural Networks, CNNs... CNNs can capture different patterns

Region Based Convolutional Neural Networks, R-CNN, is a more scalable alternative to traditional CNNS. This approach, originally proposed by R. Girshick et al. [1] in 2014...

CNNs have been successfully employed on point cloud data as well; PointRCNN

2.3 Adversarial attacks on neural networks

Most vision-based recognition software on ADS is based on CNNS; often, CNN-based deep learning models are vulnerable to the so called adversarial,

There can be small, pixel-level changes to an image that will cause the AI model to incorrectly interpret it or, on the other hand, a completely new image can be used to trick to model into thinking it is something else. The first kind is particularly dangerous, since such changes can be invisible to the human eye, and thus harder to detect.

There are mainly two categories of methods to achieve adversarial attacks, namely, optimization-based methods and fast gradient step method (FGSM)-based approach.

In [2], Zhang et al. propose and end-to-end evaluation framework for assessing the safety of a self driving deep learning model.

study on street signs: K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, and D. Song, "Robust Physical-World Attacks on Deep Learning Visual Classification," in 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, 2018, pp. 1625–1634.

In general, adversarial attacks are organized in three categories: evasion, poisoning, and extraction attacks:

- Evasion attacks modify the input to a classifier such that it is misclassified, while keeping the modification as small as possible. Evasion attacks can be black-box or white-box: in the white-box case, the attacker has full access to the architecture and parameters of the classifier. For a black-box attack, clearly this is not the case.

- In poisoning attacks, attackers have the opportunity of manipulating the training data to significantly decrease the overall performance, cause targeted misclassification or bad behavior, and insert backdoors and neural trojans

- Extraction attacks aim to develop a new model, starting from a proprietary black-box model, that emulate the behavior of the original model.

Chapter 3

Literature

In this chapter we provide a general overview of the literature concerning 3D object detection in self-driving vehicles, as well as the solutions employed to deal with adversarial attacks on their deep learning models

Example algorithm

Algorithm 1: MOSA

input : $U = \{u_1, ..., u_m\}$ the set of coverage targets of a program
Population size M
output : A test suite T

```
1  begin
2      t ← 0
3      P_t ← RANDOM-POPULATION(M)
4      archive ← UPDATE-ARCHIVE(P_t, ∅)
5      while not(search_budget_consumed) do
6          Q_t ← GENERATE-OFFSPRING(P_t)
7          archive ← UPDATE-ARCHIVE(Q_t, archive)
8          R_t ← P_t ∪ Q_t
9          F ← PREFERENCE-SORTING(R_t)
10         P_{t+1} ← 0
11         d ← 0
12         while (|P_{t+1}| + |F_d| ⩽ M) do
13             CROWDING-DISTANCE-ASSIGNMENT(F_d)
               P_{t+1} ← P_{t+1} ∪ F_d
14             d ← d + 1
15         Sort(F_d) // according to the crowding distance
16         P_{t+1} ← P_{t+1} ∪ F_d[1 : (M − |P_{t+1}|)]
17         t ← t + 1
18     T ← archive
```

Chapter 4

Conclusions

Bibliography

[1] Ross B. Girshick et al. "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation". In: *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*. IEEE Computer Society, 2014, pp. 580–587. DOI: 10.1109/CVPR.2014.81. URL: https://doi.org/10.1109/CVPR.2014.81.

[2] Jindi Zhang et al. "Evaluating Adversarial Attacks on Driving Safety in Vision-Based Autonomous Vehicles". In: *IEEE Internet Things J.* 9.5 (2022), pp. 3443–3456. DOI: 10.1109/JIOT.2021.3099164. URL: https://doi.org/10.1109/JIOT.2021.3099164.

[3] Vincenzo Riccio et al. "Testing machine learning based systems: a systematic mapping". In: vol. 25. 6. 2020, pp. 5193–5254. DOI: 10.1007/s10664-020-09881-0. URL: https://doi.org/10.1007/s10664-020-09881-0.

[4] Dmytro Humeniuk, Foutse Khomh, and Giuliano Antoniol. "A search-based framework for automatic generation of testing environments for cyber-physical systems". In: *Inf. Softw. Technol.* 149 (2022), p. 106936. DOI: 10.1016/j.infsof.2022.106936. URL: https://doi.org/10.1016/j.infsof.2022.106936.

[5] Alexey Dosovitskiy et al. "CARLA: An Open Urban Driving Simulator". In: *Proceedings of the 1st Annual Conference on Robot Learning*. 2017, pp. 1–16.

[6] Edmund K. Burke and Graham Kendall. *Search Methodologies: Introductory Tutorials in Optimization and Decision Support Techniques*. Springer US, 2014, pp. 403–449.

[7] Stefan Edelkamp and Stefan Schrodl. *Heuristic Search: Theory and Applications*. Morgan Kaufmann, 2008, pp. 403–449.

[8] Walid M. Taha, Abd-Elhamid M. Taha, and Johan Thunberg. *Cyber-Physical Systems: A Model-Based Approach*. Springer US, 2021.

[9] *Check: framework for unit testing in C*. URL: `https://libcheck.github.io/check/`.

[10] *JMetal Java framework*. URL: `http://jmetal.sourceforge.net/`.

[11] A. Author and A. Author. *Book reference example*. Publisher, 2099.

[12] A. Author. "Article title". In: *Journal name* (2099).

[13] *Example*. URL: `https://www.isislab.it`.

[14] A. Author. "Tesi di esempio ISISLab". 2099.