UNIVERSITY OF TARTU
Faculty of Science and Technology
Institute of Computer Science
Computer Science Curriculum

Mike Gomes Camara

Safety Analysis of Autonomous Vehicle Systems Software

Master's Thesis (30 ECTS)

Supervisor(s): Dietmar Alfred Paul Kurt Pfahl, PhD

Tartu 2022

Safety Analysis of Autonomous Vehicle Systems Software

Abstract:

Machine learning approaches to autonomous driving systems that rely upon computer vision and deep neural networks have demonstrated encouraging results. Some believe that the so-called end-to-end strategy is the only way to deploy self-driving autonomy at scale in vehicles. Safety is a concern as the neural networks are susceptible to adversarial attacks, small perturbations invisible to the human eye but lead to a misclassified output, potentially causing catastrophic consequences such as loss of life or property.

Introduction
Area of study

The problem
that I tackle

Literature suggests that there are procedures to defend against such threats. However, there is no understanding of how adversarial defenses can improve the robustness of an end-to-end self-driving model trained in a real-world scenario.

How I tackle
this problem

This paper aims to create a self-driving model that generalizes better and classifies correctly standard and adversarial input images. First, we select a real-world driving platform and a neural network to drive the model. Then, we designed an experiment, implemented, tested, and evaluated the results.

How I implement
my solution

In conclusion, adversarial defenses (did/did not) impact the safety and reliability of self-driving end-to-end models. Therefore employing adversarial training (increased/did not increase) the robustness of autonomous vehicles.

The result

Keywords:
adversarial attacks, adversarial machine learning, autonomous vehicles, driverless cars, self driving cars, open source, OpenCV, TensorFlow, Keras, CNN, deep learning, Raspberry Pi, Google Coral TPU, behavioral cloning

Contents

1 Introduction

End-to-end driving is an approach to autonomous driving that has become a growing trend in autonomous vehicle research both in industry and academia[1]. Unlike modular methods that use expensive sensors, end-to-end techniques to autonomous driving rely on computer vision and machine learning to generate models that command steering and acceleration [2]. However, these models are notoriously susceptible and vulnerable to adversarial images[3], which are disturbances imperceptible to the human eye but can cause the model to misbehave.

In the context of end-to-end autonomous vehicles, adversarial attacks can lead to catastrophic consequences such as loss of life and property [4]. Research has been done on the impact of malicious attacks in autonomous-vehicle neural networks but only in simulation environments[5], [5]. To the best of our knowledge, research on strategies to mitigate the issue of adversarial attacks with real-world scaled autonomous cars does not exist.

Fleets of autonomous vehicles that use machine learning are ubiquitous and available to the general public. Creating strategies to defend end-to-end models and add robustness and resilience against adversarial attacks is paramount. Such vulnerabilities must be addressed and mitigated before we can see wider adoption of machine learning models to manipulate the steering and throttle predictions of autonomous cars [6].

This thesis will carry out an experiment to evaluate the effectiveness of defenses strategies against adversarial attacks in real-world scaled autonomous cars.

1.1 Motivation

A 2016 study by the National Highway Transportation Safety Administration (NHTSA) found that

Human error accounts for over 90% of all automobile accidents, according to a 2016 study by the National Highway Transportation Safety Administration (NHTSA). Traffic accidents are the leading cause of death among young people aged 5-29, and developing countries have 90% of all road fatalities. Self-driving is a promise to mitigate this issue and make our roads safer.

End-to-end methods to self-driving evolved from being the leading and predominant approach in the DARPA grand and urban challenges to being used in the industry by companies like Wayve and deployed to production models by car manufacturers such as Tesla.

In addition, trends like urban exodus, scarcity of drivers, and a revolution in intelligent transportation systems, including autonomous last-mile delivery systems, pressure the autonomous vehicles industry to act fast and produce safe, reliable, and affordable solutions.

Adversarial attacks to machine learning systems can not be eliminated, but strategies can be deployed to defend the models and make them more robust against such threats. .

1.2 Goals

The goal of this thesis is to adopt a Design Science methodology [7] to demonstrate how adversarial machine learning attacks can be used as an artifact to improve the robustness and reliability of a deep learning model of an autonomous driving car.

We investigate the literature to discover methods to train and validate neural networks. Then we implement an experiment to create a self-driving agent and evaluate its capacity to generalize to adversarial images. Finally, to defend against such attacks, we retrain the model and expose it to perturbations while training.

As a result of the defense, the model can generalize better, such as classifying correctly standard input images while becoming immune to the adversarial attack.

The thesis is organized as follows: Section 2 discusses the building blocks to create an autonomous vehicle using neural networks. We then discuss the threats against machine learning and the precautions necessary to deal with adversarial attacks. Finally, we investigate applying those concepts in a real-world scaled autonomous vehicle. Section 3 will illustrate the approach that is taken and the detailed phases necessary to accomplish the experiment. Section 4 supplies the outcomes of each stage described in the prior section. Section 5 discusses lessons learned and the limitations of the project. Finally, Section 6 encloses our evaluation conclusions and gives suggestions for future work.

2 Background

Some text...

2.1 Computer Vision in self-driving

Some text...

2.2 Neural Networks

Some text...

2.2.1 Convolulutional Neural Networks

Some text...

2.2.2 Models

Some text...

2.2.3 Training

Some text...

2.3 Adversarial Attacks

Some text...

2.3.1 Fault Injection in Trained Agents

Some text...

2.3.2 Defence to Adversarial Attacks

Some text...

2.4 Scaled Autonomous Car

Some text...

2.4.1 Donkeycar Platform

Some text...

Software Some text...

Open-source Community Some text...

OpenCV Some text...

Keras Some text...

Hardware Some text...

Raspberry Pi Some text...

Google Coral TPU Some text...

Remote Control Car Some text...

IMU Some text...

2.4.2 Safety

Some text...

2.4.3 Metrics

Some text...

3 Method

Some text...

3.1 Selection of Self-Driving Platform

Some text...

3.2 Selection of Driving Model Architecture

Some text...

3.3 Training Pilot Model

Some text...

3.4 Design and Implementation

Some text...

3.4.1 Testing 1

Some text...

Baseline Metrics Some text...

Adversarial Attack Generator Some text...

3.4.2 Testing 2

Some text...

Implement Defence Mechanism Some text...

3.4.3 Testing 3

Some text...

Test Robust Model on Baseline Some text...

3.4.4 Testing 4

Some text...

3.5 Evaluation

Some text...

4 Results

Some text...

4.1 Selection of Self-Driving Platform

Some text...

4.2 Selection of Driving Model Architecture

Some text...

4.3 Training Pilot Model

Some text...

4.4 Design and Implementation

Some text...

4.4.1 Context

Some text...

4.4.2 Implementation

Some text...

4.4.3 Baseline Metrics

Some text...

4.4.4 Testing 1

Some text...

Adversarial Attack Generator Some text...

4.4.5 Testing 2

Some text...

Implement Defence Mechanism Some text...

4.4.6 Testing 3

Some text...

Test Robust Model on Baseline Some text...

4.4.7 Testing 4

Some text...

4.5 Evaluation

5 Discussion

5.1 Lessons Learned

5.2 Limitations

6 Conclusion and Future Work

7 Acknowledgement

References

[1] A. Tampuu, M. Semikin, N. Muhammad, D. Fishman, and T. Matiisen, “A survey of end-to-end driving: Architectures and training methods,” *CoRR*, vol. abs/2003.06404, 2020.

[2] A. Zolfi, M. Kravchik, Y. Elovici, and A. Shabtai, “The translucent patch: A physical and universal attack on object detectors,” in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pp. 15232–15241, Computer Vision Foundation / IEEE, 2021.

[3] A. Qayyum, M. Usama, J. Qadir, and A. I. Al-Fuqaha, “Securing connected & autonomous vehicles: Challenges posed by adversarial machine learning and the way forward,” *IEEE Commun. Surv. Tutorials*, vol. 22, no. 2, pp. 998–1026, 2020.

[4] P. Sharma, D. Austin, and H. Liu, “Attacks on Machine Learning: Adversarial Examples in Connected and Autonomous Vehicles,” 2019.

[5] N. Piazzesi, M. Hong, and A. Ceccarelli, “Attack and fault injection in self-driving agents on the carla simulator - experience report,” in *Computer Safety, Reliability, and Security - 40th International Conference, SAFECOMP 2021, York, UK, September 8-10, 2021, Proceedings* (I. Habli, M. Sujan, and F. Bitsch, eds.), vol. 12852 of *Lecture Notes in Computer Science*, pp. 210–225, Springer, 2021.

[6] S. Pavlitskaya, S. Ünver, and J. M. Zöllner, “Feasibility and suppression of adversarial patch attacks on end-to-end vehicle control,” in *23rd IEEE International Conference on Intelligent Transportation Systems, ITSC 2020, Rhodes, Greece, September 20-23, 2020*, pp. 1–8, IEEE, 2020.

[7] G. L. Geerts, “A design science research methodology and its application to accounting information systems research,” *Int. J. Account. Inf. Syst.*, vol. 12, no. 2, pp. 142–151, 2011.

Appendix

I. Glossary

II. Licence

Non-exclusive licence to reproduce thesis and make thesis public

I, **Mike Camara**,
(author's name)

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to

   reproduce, for the purpose of preservation, including for adding to the DSpace digital archives until the expiry of the term of copyright,

   Safety Analysis of Autonomous Vehicle Systems Software,
   (title of thesis)

   supervised by Dietmar Alfred Paul Kurt Pfahl.
   (supervisor's name)

2. I grant the University of Tartu a permit to make the work specified in p. 1 available to the public via the web environment of the University of Tartu, including via the DSpace digital archives, under the Creative Commons licence CC BY NC ND 3.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work until the expiry of the term of copyright.

3. I am aware of the fact that the author retains the rights specified in p. 1 and 2.

4. I certify that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

Mike Gomes Camara
04/01/2022