Safety of autonomous driving systems against adversarial attacks

Analysis on benchmarks evaluating the safety level of autonomous driving systems

A Thesis Prospectus
In STS 4500
Presented to
The Faculty of the
School of Engineering and Applied Science
University of Virginia
In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science in Computer Engineering

By
Haotian Ren

November 1, 2021

Technical Team Members:
Haotian Ren
Xugui Zhou

On my honor as a University student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments.

ADVISORS

Joshua Earle, Department of Engineering and Society

Homa Alemzadeh, Department of Electrical and Computer Engineering

Introduction

In the early 20th century, science fiction familiarized the world with the concept of artificially intelligent robots. It began with the “heartless” Tin man from the *Wizard of Oz* and continued with the humanoid robot that impersonated Maria in *Metropolis*. From 1957 to 1974, AI flourished. Computers could store more information and became faster, cheaper, and more accessible; machine learning algorithms also improved, and people got better at knowing which algorithm to apply to their problem. Today, intelligence has led to implementations everywhere. There are robots helping people with daily routines, like cleaning the floor; there are software engines specializing in some specific jobs, like AlphaGo developed by Google (Silver, D., 2016), which defeated the best Go player all over the world; there are also systems capable of complicated tasks like driving a vehicle. Applications of artificial intelligence in smart homes and board games are widely accepted because of its low-risk and harmless properties. However, when it comes to driving a vehicle, it is worth a careful investigation on whether it is safe to use artificial intelligence.

Autonomous driving, or self-driving, is a technology that can drive a vehicle without active physical control or monitoring by a human operator. (*Autonomous vehicles terms and definitions*) The first “autonomous” driving car was invented in the 1980s, capable of following the road and cruising at 19 mph. However, this first model is too trivial to be considered as safe given the level of complexity of the traffics nowadays. But, through the ceaseless efforts of scientists and engineers, the leader in the field of autonomous driving technology, Tesla, has released their Autopilot system, with full self-driving capability that can be used on real road, outside of the test ground.

A straightforward way to find out whether an autonomous system is “intelligent” enough is by simulation. Since the autonomous driving technology by Tesla is not open-sourced, it’s impossible to examine the safety level of it. Instead, tests could be done on another state-of-the-art self-driving system, called OpenPilot (Ai, C, 2020). By doing adversarial attacks against different functionalities of this autonomous driving system in my technical project, whether this technology has been developed to a level to make right decisions in situations of possible hazards can be found out.

Even though the performance of artificial intelligence under dangerous circumstances can be evaluated through simulation, simulation results are not adequate to tell people whether autonomous driving systems are safe to use in real traffics, since simulations are built on theoretical models, ignoring many tiny but important factors. Hence, to evaluate the safety level, benchmarks has been built alongside the development of autonomous driving system, for both validation processes of prototypes and proofs of safety to the customers. In my STS research, different benchmark systems will be analyzed to find answers to how the safety level of an autonomous driving system is evaluated and whether they are reasonable.

Safety of autonomous driving systems against adversarial attacks

The technical portion of the thesis portfolio will focus on implementing adversarial attacks on OpenPilot, an autonomous driving system with cutting edge technology, by Comma AI. For this system, it can be treated as three parts: inputs, the processing mechanism, and outputs. Inputs are readings from sensors, like lidar and camera, and car states retrieved from a vehicle computer control system; outputs are various kinds of operations to be implemented on a car; the processing mechanism, which is the most important part, consists of a bunch of code with the core being a machine learning model functioning like the brain of a human being.

The safety level of Openpilot will be evaluated through simulation with CARLA (Dosovitskiy, A., 2017), an open-source traffic scenario simulator specially designed for autonomous driving system. Adversarial attacks are designed to either simulate dangerous scenarios that could happen in the real world or actively hack the system to interfere with its decision making.

Regarding the machine learning model, which is an extremely sophisticated deep neural network (*What is deep learning?: How it works, techniques & applications*.), two of the possible dangerous scenarios will be evaluated against it: entrance of a vehicle into the lane in the front and a sudden stop of the leading vehicle. The experiments are split into two parts. The first part will test the basic capability of the autonomous driving system making the reasonable decision, and the second part will deliver some active attacks involving adding patch to the camera feed and changing the road appearance to fool the machine learning model. These interference methods can push the self-driving vehicle into three hazardous states: accelerating and crashing into the leading vehicle, driving out of the lane that it should be inside unintentionally, or coming to a complete stop on a highway.

Except for attacks on the "brain" of this autonomous driving system, another peripheral part that's usually ignored is the messaging system responsible of transmitting data, functioning like the nervous system. Attempts will be made to let dangerous commands passing through the message system and arrive at the engine control unit (ECU) (Wikimedia Foundation, 2022) of the vehicle, implementing a man-in-the-middle attack with baleful modifications.

Through simulations with adversarial attacks on both the machine learning model and the messaging system, a quantitative result can be attained to tell whether Openpilot can make correct decision and whether it is resistant to hacking activities from outside of the system. It will give us an idea of what safety level has this state-of-the-art autonomous driving system achieved.

Analysis on benchmarks evaluating the safety level of autonomous driving systems

Smartphones, smart homes, quantitative trading, weather forecasting, etc. Applications of artificial intelligence have dominated our life. No matter a simple Turing complete (*Turing Completeness*.) interactive application or a complicated system like autonomous driving backboned by a deep machine learning model, they all function to assist us in finishing some specific tasks and make our life easier and more comfortable. According to the National Highway Traffic Safety Administration (NHTSA) (*Automated vehicles for safety*.), the automation of vehicles is divided into five levels. For the first two levels, the driver still needs to be fully engaged and the automation just plays the role of assisting the driver with acceleration, braking, and steering. These two levels of automation can be found in many new cars manufactured recently. But for levels three to five, they require the automation system to be fully capable of driving the car, without any active human interference.

As the self-driving vehicle has to make decisions under complicated circumstances based only on its “brain”, the machine learning model, and some trivial linear safety rules, like a limit for acceleration, there comes the question whether the artificial intelligence technology used in autonomous driving systems is capable of dealing with dangers, especially those under which even human-beings will hesitate to make final decisions. Some people believe that autonomous driving is the future while others think that it’s so dangerous to let a machine that doesn’t think in the way of human beings to gain full control over a car. For vehicle manufacturers, to persuade customers to buy their cars and prove to the public that self-driving vehicles can actively avoid accidents, benchmarks are designed to evaluate the safety level of autonomous driving systems.

Just like benchmarks for the performance of graphic cards used in computers, different parties have different standards and definitions for the safety level of an autonomous driving system. Engineers may focus on the technical part while customers may focus on the moral decisions, ending up with various kinds of benchmarks. One of the well-known benchmarks is Waypoint developed by Waymo, a self-driving vehicle manufacturer.(*Waymo's safety methodologies and safety readiness determinations*.) This benchmark analyzes the reaction time in traffic conflicts and performance in collision avoidance, and then compared these quantified results with human drivers to give an overall rating of the safety level of the autonomous driving system.

As this benchmark is created by a car manufacturer, it is more specially designed to test their own products than for a wider compatibility. Hence, an open-source benchmarking platform called SafeBench (Xu, C. 2022) will also be studied. This benchmarking platform employs 8 safety-critical testing scenarios following National Highway Traffic Safety Administration (NHTSA) and 4 scenario generation algorithms considering 10 variations for each scenario. Meanwhile, there are also 4 deep reinforcement learning-based AD algorithms with 4 types of input (e.g., bird’s-eye view, camera) to perform fair comparisons on SafeBench.

To gain more insight into how different benchmark works and how they are built, IEEE 2846 ("IEEE Standard for Assumptions in Safety-Related Models for Automated Driving Systems,") will be investigated. This document is very authoritative as it comes from IEEE, where global standards of electronic technologies are defined. This document set up a formal rules-based mathematical model for automated vehicle decision-making. Though it’s not a benchmark tool, the rules defined in this document can give me a guidance on analyzing and comparing Waypoint and SafeBench.

In my STS project, the framework to use in studying those two benchmarks and the global standard defined by IEEE will be case study. Published papers related to the creation of the benchmarks and public policies, specifically IEEE 2846, will be main sources for conducting the research. Through comparisons from various perspectives between the building process and working mechanisms of those benchmarks and policies, an image of what factors are considered when building a benchmark will be depicted, giving us a sense of whether they are reasonable.

Foundational Texts and Primary Resources

One source that will provide the study of STS topic with plenty of references is the papers published along with Waypoint and SafeBench, the two benchmarking tools that will be investigated. Since the framework of this STS project is case study, reading these papers and analyzing the designing process and methods used by the creators plays a critical role in understanding how these benchmarks are built.

Besides, the content of IEEE2846 and the *Federal Automated Vehicles* Policy (*Federal Automated Vehicles Policy*) published by the U.S. Department of Transportation in 2016 also provide valuable ideas on how the authorities define the safety of autonomous driving. The IEEE standard represents the perspective of engineers while this first handbook of rules on autonomous vehicles represents the government's attitude.

Another primary source is Diane Michelfelder's *Test-Driving the Future*. (Michelfelder, D. P., 2022) Though this book did not discuss the safety standards of autonomous driving directly, it gives a more ethical view of the safety issue related to artificial intelligence used in self-driving systems. This can provide me with ideas which might be useful in explaining incentives of certain rules.

References

Ai, C. (2020, January 27). *Open sourcing openpilot development tools*. Medium. Retrieved December 15, 2022, from https://comma-ai.medium.com/open-sourcing-openpilot-development-tools-a5bc427867b6

Automated vehicles for safety. NHTSA. (n.d.). https://www.nhtsa.gov/technology-innovation/automated-vehicles-safety

Autonomous vehicles terms and definitions. California DMV. (2020, June 3). https://www.dmv.ca.gov/portal/vehicle-industry-services/autonomous-vehicles/autonomous-vehicle-definitions/

Dosovitskiy, A., Ros, G., Codevilla, F., Lopez, A., & Koltun, V. (2017, October 18). *Carla: An open urban driving simulator*. PMLR. Retrieved December 15, 2022, from http://proceedings.mlr.press/v78/dosovitskiy17a

Federal Automated Vehicles Policy - september 2016. U.S. Department of Transportation. (n.d.). Retrieved November 30, 2022, from https://www.transportation.gov/AV/federal-automated-vehicles-policy-september-2016

"IEEE Standard for Assumptions in Safety-Related Models for Automated Driving Systems," in IEEE Std 2846-2022 , vol., no., pp.1-59, 22 April 2022, doi: 10.1109/IEEESTD.2022.9761121.

Michelfelder, D. P. (2022). *Test-driving the future: Autonomous Vehicles and the ethics of Technological Change*. Rowman & Littlefield.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., & Hassabis, D. (2016, January 27). *Mastering the game of go with deep neural networks and Tree Search*. Nature News. Retrieved December 15, 2022, from https://www.nature.com/articles/nature16961

Turing Completeness. (n.d.). Retrieved December 15, 2022, from https://www.cs.odu.edu/~zeil/cs390/latest/Public/turing-complete/index.html

Waymo's safety methodologies and safety readiness determinations. (n.d.). https://storage.googleapis.com/sdc-prod/v1/safety-report/Waymo-Safety-Methodologies-and-Readiness-Determinations.pdf

What is deep learning?: How it works, techniques & applications. How It Works, Techniques & Applications - MATLAB & Simulink. (n.d.). Retrieved December 15, 2022, from https://www.mathworks.com/discovery/deep-learning.html

Wikimedia Foundation. (2022, November 29). *Engine Control Unit*. Wikipedia. Retrieved December 15, 2022, from https://en.wikipedia.org/wiki/Engine_control_unit

Xu, C., Ding, W., Lyu, W., Liu, Z., Wang, S., He, Y., Hu, H., Zhao, D., & Li, B. (2022, October 29). *SafeBench: A benchmarking platform for safety evaluation of Autonomous Vehicles*. arXiv.org. Retrieved November 30, 2022, from https://arxiv.org/abs/2206.09682