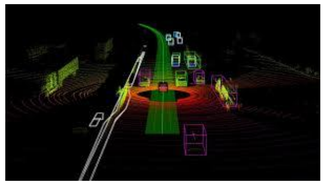
UR:120 Analyzing Adversarial Attacks in Machine Learning

Abstract

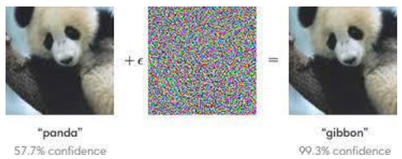
Studies have been done on adversarial attacks in machine learning models, but not much research has looked at the possible adversarial attacks in Lidar data. Since these attacks are getting stronger and more effective over time, a model mimicking the use of Lidar data is secured against adversarial attacks by adversarial training, increasing road safety in the future. When the model was built, it returned the correct accuracy without the presence of adversarial images, but when adversarial images were introduced using Fast Gradient Sign Method, the model misclassified images. Adversarial training, a way machine learning models can be more robust (Ren et al., 2020; Using Adversarial, 2021), was completed in this study to improve the accuracy, and the successful method can be generalized to future studies in this field.

Introduction

One way to ensure that machine learning (ML) models are as secure as possible and increase road safety is to create a general, unique, and effective solution that allows models that use Lidar data to successfully avoid a variety of adversarial attacks. Over the past few years, many companies have reported security breaches, and these cybersecurity attacks are more concerning considering they go unrecognized, similar to how adversarial attacks in ML models go unrecognized (Millenium Communications Group Inc. [MCG], 2015). ML algorithms, which create models, can complete a variety of different tasks such as identifying an object based on an image and predicting stock market trends (Dasgupta & Collins, 2019), but with their growing influence on everyday life, they have cybersecurity researchers worried about the possibility of adversarial attacks (Cooper, 2021). The aim of this research is to analyze a model that mimics the use of Lidar data, a way autonomous cars can scan surroundings, to develop a unique solution that protects against a variety of adversarial attacks in their algorithms.



Example of Lidar from extremetech.com



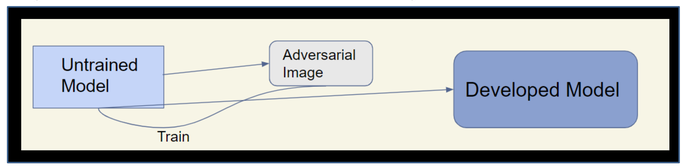
Example of an adversarial attack from openai.com

Research Questions

- What foundations and algorithms present around models involving Lidar data are most susceptible to adversarial attacks?
- How can machine learning algorithms be trained to defend against adversarial attacks?
- **Design Sub Problem**: Develop a model that will defend against adversarial attacks involving adversarial examples in Lidar data.

Materials and Methods

The original model was first trained and then tested for its accuracy with a dataset containing adversarial images, and later, a different set of adversarial images were used to train the model in the form of adversarial training. Finally, the developed model was tested with the same adversarial images used to test the original model to determine its final accuracy.



Results

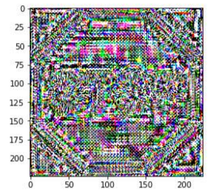
In order to build the model, a data set of around 51,000 images was acquired on traffic signs, and the number of images used in each testing phase are shown below. These images were to train and test the accuracy of the model using the deep learning Convolutional Neural Network. This accuracy of the model was 99.84%

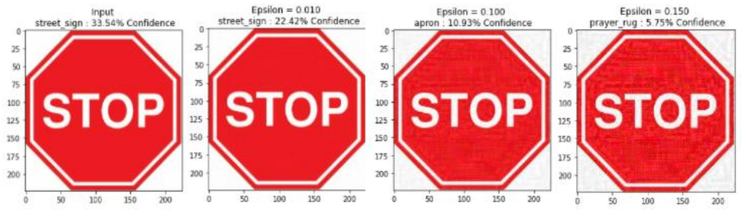
Number of Images Used in Each Phase of the Model	
Training and Validation	39, 209
Validation	11, 762
Testing	12,631

Note: 30% of the training and validation data is used in the validation phase

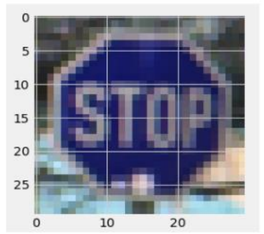
Later, epsilon values were tested in a smaller model to determine the epsilon values needed to fool the main model.

*The image on the right is the image scaled





Adversarial Images Created in the Actual Model (Initial vs Adversarial):



The accuracy of the model in the presence of adversarial images was 67%. In order to complete adversarial training, the adversarial images were added to the training phase of the model to allow them to be better classified, and the accuracy of the model after training in the presence of adversarial images was 98.24%.

Tools

Python, kaggle, CO, Keras

Conclusions

The basic sub problems of the study helped complete the design sub-problem that aimed to develop a model that will defend against adversarial attacks involving adversarial examples in Lidar data. The two basic sub problems uncovered that Lidar data can be prone to adversarial attacks involving adversarial examples, and considering these attacks are getting stronger and pose a risk to road safety (Lin & Biggio, 2021), this research was conducted to determine more unique and general solutions to the secure models involving Lidar data and add to existing literature. The design sub problem was completed by building a model that uses a Convolutional Neural Network to predict the name of a traffic sign based on an image given, and following the creation of this model, it was successfully attacked using adversarial examples. To secure this model from adversarial attacks, adversarial training was used to allow the model to classify the adversarial images correctly.

Future Work

A future study could focus on using actual Lidar data with autonomous cars to experiment with adversarial objects and examples to determine a more specific methodology for adversarial objects. Additionally, a greater sample size should be tested to provide more accurate results; additionally, attacks other than those involving adversarial examples can be explored to further increase the accuracy of the model that is initially at 99.84%. All of these improvements will further allow machine learning algorithms to be more secure and improve road safety with autonomous cars.

Acknowledgments

Dr. Dan Lo, Professor of Computer Science at Kennesaw State University
Dr. Ashley Deason, Advanced Research Teacher at Wheeler High School
Ms. Elizabeth Gainsford, Advanced Internship Teacher at Wheeler High School

Contact Information

- Niti Mirkhelkar, nmirkhe1@kennesaw.edu
- Dr. Dan Lo, dlo2@kennesaw.edu

References

Cooper, K. (2021, May 17). What is adversarial machine learning—and why could it become the next big cybersecurity threats Springboard Blog. https://www.springboard.com/blog/cybersecurity/adversarial-machine-learning-couldbecome-the-next-big-cybersecurity-threat/

Dasgupta, P., & Collins, J. B. (2019, Summer). A survey of game theoretic approaches for adversarial machine learning in cybersecurity tasks. AI Magazine, 40(2), 31+.

Lin, H.-Y., & Biggio, B. (2021). Adversarial machine learning: Attacks from laboratories to the real world. Computer, 54(5), 56-60. https://doi.org/10.1109/MC.2021.3057686\

Millennium Communications Group, Inc. (2015, February). Why cyber security is important. New Jersey Association of Counties. https://njac.org/why-cyber-security-is-so-important/

Ren, K., Zheng, T., Qin, Z., & Liu, X. (2020). Adversarial attacks and defenses in deep learning. Engineering, 6(3), 346-360. https://doi.org/10.1016/j.eng.2019.12.012

Using adversarial images to assess the stability of deep learning models trained on diagnostic images in oncology. (2021, June 24). Women's Health Weekly, 7724.

KENNESAW STATE UNIVERSITY
COLLEGE OF COMPUTING AND SOFTWARE ENGINEERING

Author: Niti Mirkhelkar
Advisor: Dr. Dan Lo