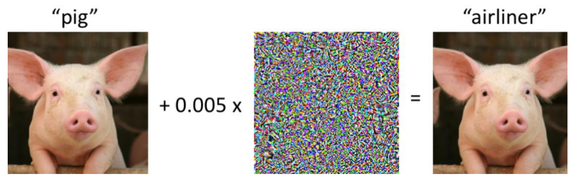
Developing Robust Defenses for Deep Neural Networks

Major Brown, Computer Science
Mentor: Dr. Yezhou Yang, Assistant Professor
School of Computing, Informatics, and Decision Systems Engineering

Introduction

The use of deep learning in computer vision has become highly integrated in many technologies being used today, including self driving cars, electronic banking, and security systems. While this technology has greatly benefited society, there are also some weaknesses in the current approach. Attackers can exploit these weaknesses in the model to create perturbations that are imperceptible to the human eye, but that trick computers.



https://medium.com/@smkirthishankar/the-unusual-effectiveness-of-adversarial-attacks-e1314d0fa4d3

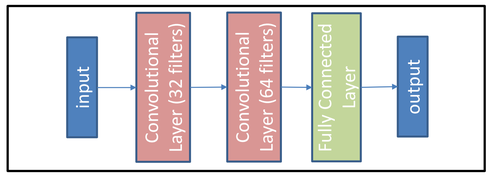
The results can be devastating and dangerous, especially in the case of self driving cars. Deep neural networks need to be trained to be more robust against these types of attacks.

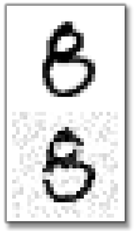
Abstract

This research explores more robust methods to train deep neural networks so that they are resistant against such attacks in order to keep people safe from harm. The first step in this project has been to explore the techniques used by Madry and test the effectiveness of training a model using data perturbed by a PGD adversary (2019).

Research Methods

The model uses two convolutional layers and a fully connected layer at the end





Perturbations are added to the training data using projected gradient descent (PGD) beforehand to make the model robust to a broad class of attacks

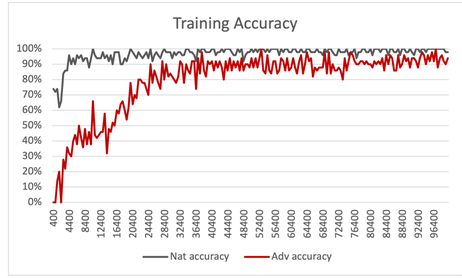
Obstacles

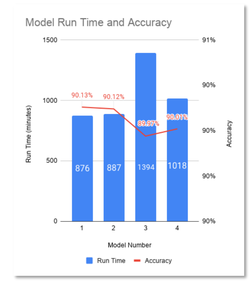
Initially, getting the program to run automatically on the Agave high-performance cluster took some time. However, once that was set up, it became much easier to test the model.

Another obstacle that needs to be overcome is making the model run more efficiently. Currently the model takes about 14 hours to run. By utilizing the processing power of the GPUs on the high-performance cluster, there could be significant improvements in the computation speed.

Findings

Training the model and comparing the accuracy of the naturally trained data versus the adversarially trained data (left graph)





Results from four trials comparing the run time and the accuracy when tested against adversarial input and different hyperparameters (right graph)

Future Work

The next part of this project will be focusing on testing with different data sets and adjusting the hyperparameters to see if this method is robust within a wide variety of settings. Furthermore, additional attack methods can be developed to test and even train the data to make the model even more robust.

References

Aleksander Madry, et al. "Towards Deep Learning Models Resistant to Adversarial Attacks." (2019).

FURI

Ira A. Fulton Schools of Engineering
Arizona State University