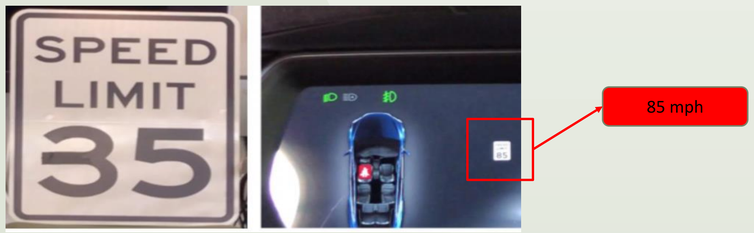
Optimal Defenses for Defending Autonomous Vehicles from Adversarial Attacks

Lex Baker[1], Omar Abdel Azim[1], Reek Majumder[2], Abyad Enan[2], Dr. Sakib Khan[2,3], and Dr. Mashrur Chowdhury[2,3]

[1]South Carolina Governor's School of Science and Mathematics, [2]Glenn Department of Civil Engineering, Clemson University, [3]Center for Connected Multimodal Mobility

INTRODUCTION

- Autonomous vehicles utilize deep learning for decision-making, with one of the main sub-processes being traffic sign detection.
- We have designed and developed two state of the art deep learning models based on Resnet18 architecture and Inception-V3 architecture.
- All models were trained on the LISA traffic sign dataset.
- Deep learning models are vulnerable to adversarial attacks, which could make the model misclassify a stop sign as a 45mph speed limit sign. Thus, it poses a major safety risk for all autonomous vehicles.



Researchers stuck a 2-inch strip of tape on a 35-mph speed sign and successfully tricked two Tesla's into accelerating to 85 mph

- The defense includes a combination of Total Variance Minimization (TVM), Spatial Smoothing (SS), Gaussian Smoothing (GS), and JPEG Compression. These are all methods of removing noise and smoothing the overall image.
- We have trained our Resnet18 and Inception-V3 models against various adversarial attacks to design, develop, and test the best defensive strategy for our models during these attacks

METHODS

The models were trained on the LISA traffic signs database

Their accuracy was tested against several adversarial attacks

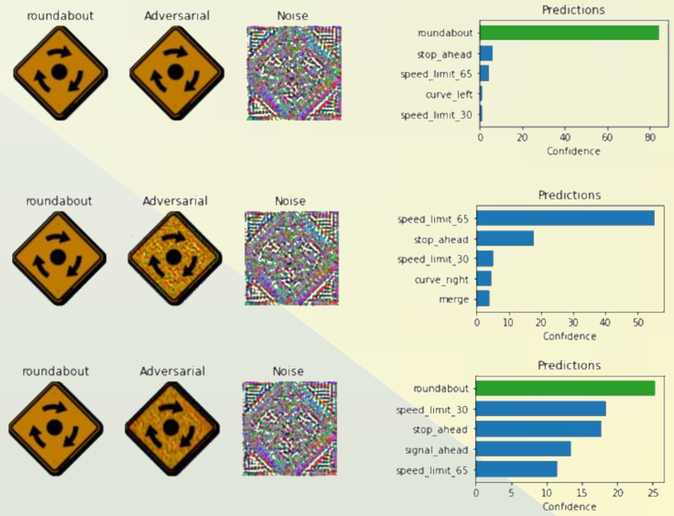
Defenses were applied and they were retested against the same attacks

ATTACKS

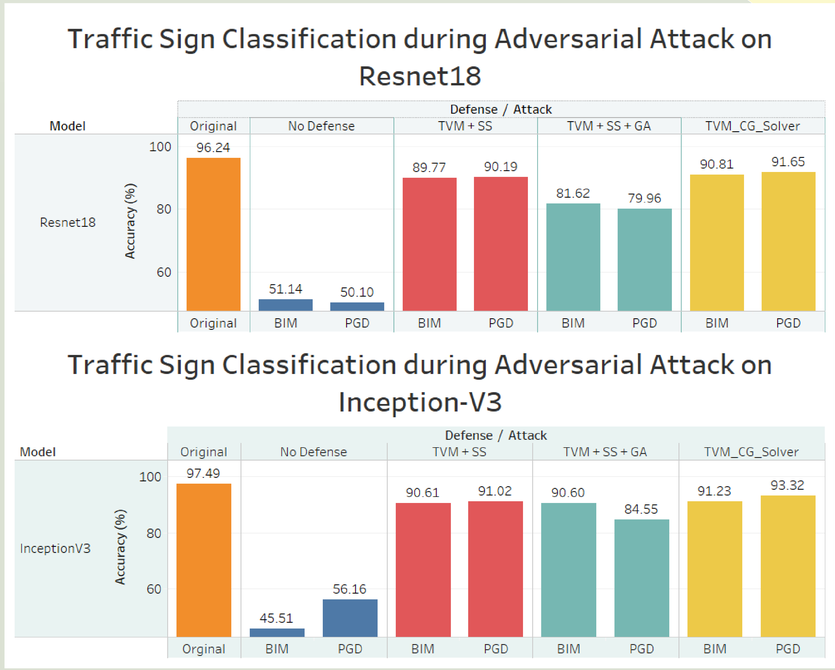
BIM: Basic Iterative Method, at an epsilon of 0.4

PGD: Projected Gradient Descent, at an epsilon of 0.3

Both attacks are white box, meaning they can access the weights of our model. They both use gradient descent attacks, where the noise added to an image is in the opposite direction of how the model trained, maximizing the negative impact on accuracy. The epsilon value equates to the strength of the attack



RESULTS



CONCLUSION

We found that a combination of Total Variance Minimization and Spatial Smoothing provided the best accuracy. Finding the best defense is crucial, as the differences between 40% accuracy and 90% accuracy are countless human lives. While our research is primarily conducted within a lab, the discoveries and progress we make have far-reaching ramifications and can protect passengers of the future from malicious attacks against the autonomous vehicles that transport them.

FUTURE WORK

We are publishing a research paper by the end of July. It will expand on our work with adversarial defenses by exploring other types of defenses. We used preprocessing defenses in our research, meaning that the input images are changed prior to being fed into the model. In the future, we will use post processing defenses, such as the Reverse Sigmoid defense. Additionally, one adversarial attack we seek to explore more is the Carlini & Wagner attack, which is one of the most potent adversarial attacks. By developing defense mechanisms that can combat the Carlini & Wagner attack, the deep learning models that power autonomous vehicles will be more secure, making our roads safer for all.

ACKNOWLEDGEMENTS

We'd like to thank Clemson and the Center for Connected Multimodal Mobility (C^2M^2), as well as the graduate students and mentors who went above and beyond to help us learn and grow over these six weeks, in addition to their invaluable assistance in helping us publish our first research paper. Additional thanks to Dr. Witten and Mr. Gibson for providing us with this wonderful opportunity to gain real world experience in research.

SUMMER PROGRAM FOR RESEARCH INTERNS

Center for Connected Multimodal Mobility C^2M^2

HONORS COLLEGE