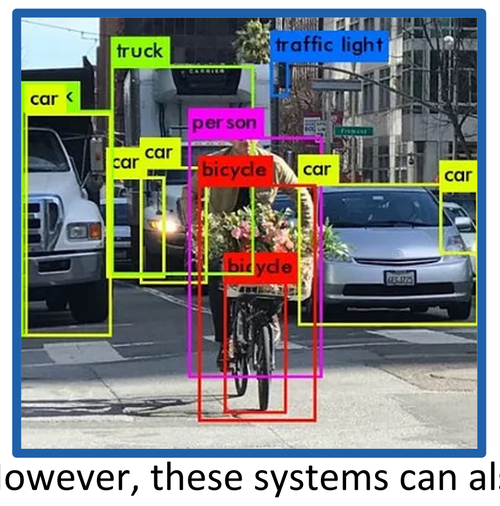
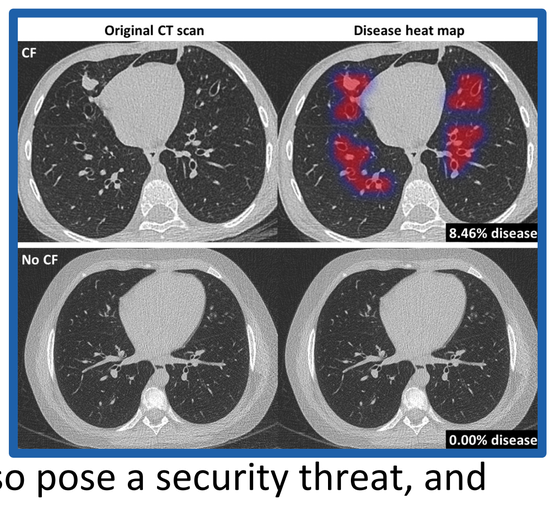
Using Fourier Optics to Perform Adversarial Attacks on Image Classifiers

Evan Marcinkevage, Electrical Engineering – Penn State University
Advisors: Tim Kane and Scott Hodes

Introduction

The surge in development of machine learning within the past few years has lead to significant progress in both developing and applying these systems to a variety of problems. One of the many machine learning applications is image classification, where an AI takes an image input and is able to identify the object(s) or text within that image. This technology is used everywhere from mobile check deposits to CT scans to self driving cars.

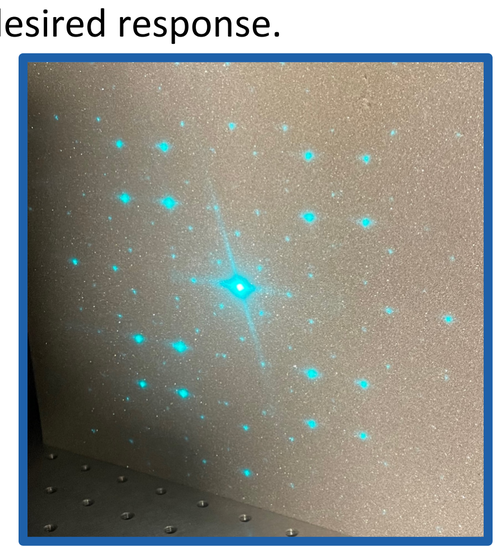


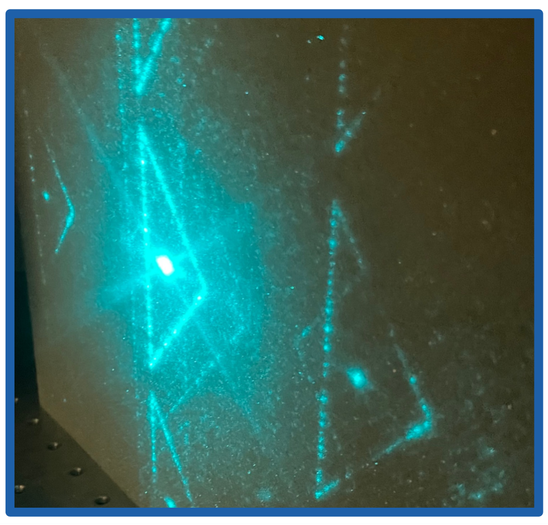


However, these systems can also pose a security threat, and there is currently very little research into how we counter these threats. Our research focuses on attacking image classification AI through the use of optical instruments.

Objectives

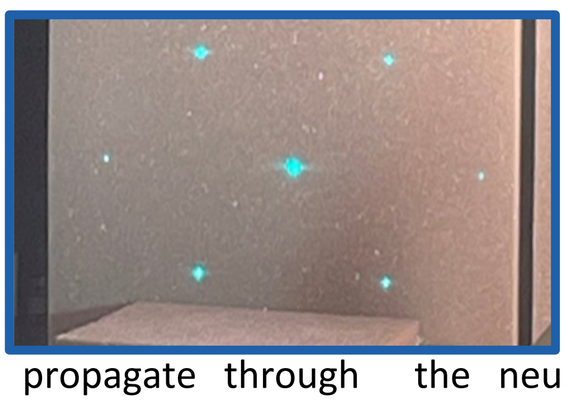
Our goal is to create an optical system that projects a field image onto the target object in such a way that it causes the AI to falsely classify the image Our system will perform a white-box attack in order to determine what field is necessary to elicit the desired response.



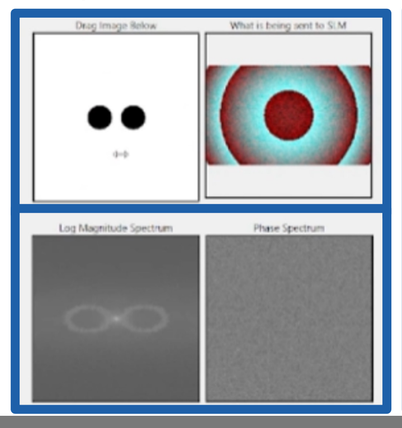


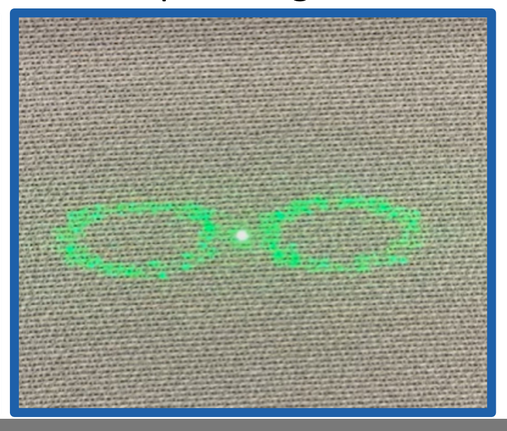
Methodology

We start by gathering key points from an initial laser layout, with which we can perform a homography transform on the original image to account for the perspective of our camera in relation to the laser.



We then back propagate through the neural network to determine where laser points would need to be placed in order to cause the AI to misclassify the image, and finally we use the Gerchberg-Saxton algorithm to determine how we would turn the a laser point into that field with a Spatial Light Modulator.

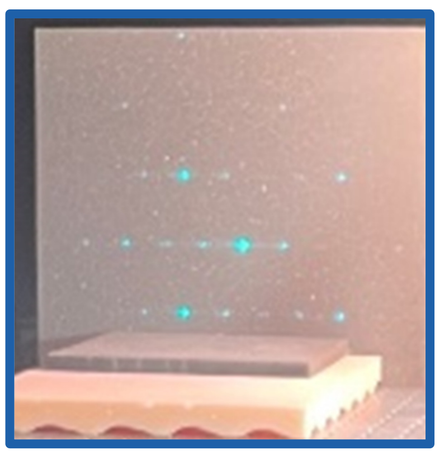


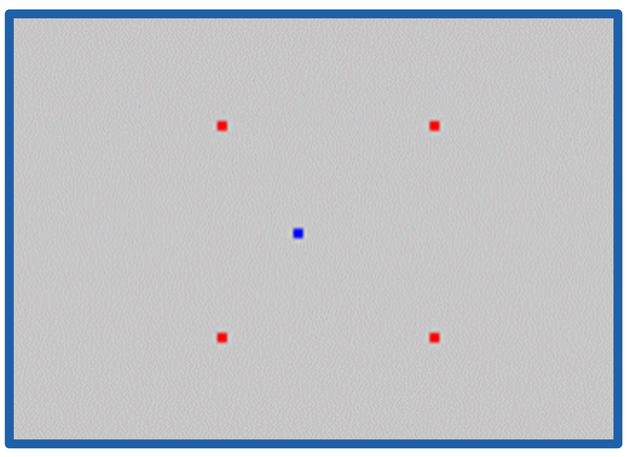


Results / Analysis

Monochromatic implementations of traditional white box attacks are successful in simulation. The below image of a Tench was misclassified as a Bullfrog due to the laser points added through back propagation.

Because the Gerchberg-Saxton algorithm does not account for asymmetry, we are getting “ghost beams”, which are a less intense mirror-image of the desired field. This can result in a flipped shape or a large field of extra points.





Future Objectives

We are seeking to address the issue with “ghost beans” so our system functions as intended.

The next step is to adapt our system to work in box agnostic scenarios, as we very rarely will have access to the neural network of our target. Thankfully, most convolutional neural networks extract features from images in a similar fashion.

We will also be investigating how to train neural networks against these kinds of adversarial attacks.

Acknowledgements

This work was partially supported by the **PIPELINE**: **P**enn State **I**ntern **P**ipelin**E** **LI**nks to **N**avy **E**ngineering program, ONR grant #N000142312656. The Penn State PIPELINE Program motivates and connects students and faculty to careers and research opportunities with the Navy technical workforce.

References

Renukasoni. (2019, July 31). *Image detection, recognition and image classification with machine learning.* Medium. https://medium.com/ai-techsystems/image-detection-recognition-and-image-classification-with-machine-learning-92226ea5f595

AI solution automates analysis of CT scans in cystic fibrosis patients • APPLIED RADIOLOGY. (2024, July 23). https://appliedradiology.com/articles/ai-solution-automates-analysis-of-ct-scans-in-cystic-fibrosis-patients