Jangseung: Defense Against Adversarial Perturbations

Protecting Models from OneSVMs to Neural Networks

Summer 2021

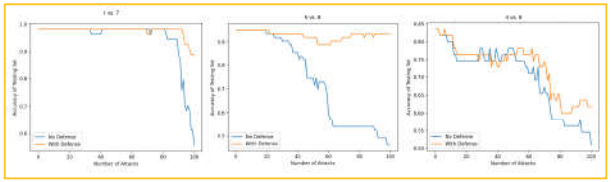
Team Members

Alicia supported cyber security outreach initiatives by developing summer camp materials. This year, Alicia looks forward to exploring her interests in cyber security and ML algorithms.

Shawna helped with outreach and researched continuous authentication, controlling soundscapes and audio signal processes using embedded devices, and machine learning security.

Problem Statement

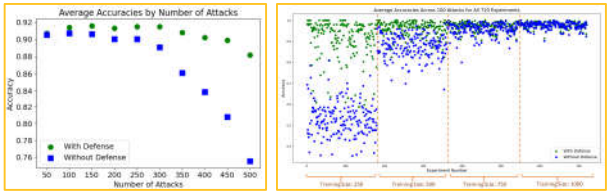
This research extends work done by **Woodrow Gamboa** which protected OneSVM models from adversarial perturbations.



This round of research delves into deep learning models, different types of adversarial perturbations, and larger datasets for more complicated classification techniques.

Results

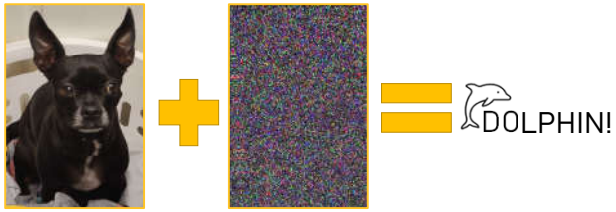
Previous Results: OneSVM models can be protected from adversarial perturbations when tested on balanced and normalized data.



Anticipated Results: A similar defense strategy can also be applied to neural networks. While theory strongly supports the OneSVM results, they can also be applied to wider applications.

Background

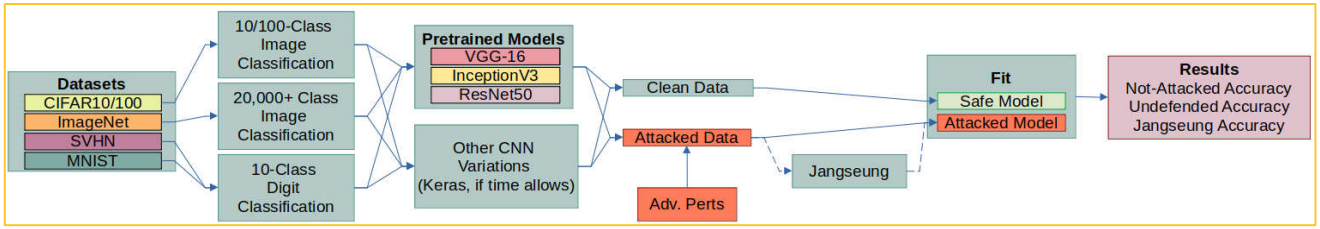
Adversarial Perturbations: Slightly altered data used to confuse machine learning models. These alterations are usually imperceptible to human observers.



Ethical Machine Learning: Creating assurances for artificial intelligence systems that prove the resiliency, explainability, and trustworthiness of models from conception to deployment.

Methods

While OneSVMs are built for one-class classification tasks, this research extends to much larger sets of classes. It also used pre-trained models which are harder to attack due to the extensive training. **We aim to show the vulnerabilities in transfer learning and how to defend against them efficiently.**



Challenges & Future Work

Next steps include:

- Downloading datasets and determining a system for sampling the data.
- Building common pre-trained models and ensuring they are not overtrained.
- Creating a framework for generating adversarial perturbations for neural networks.

Advisor: Dr. Mike Borowczak (Mike.Borowczak@uwyo.edu)

Grad Mentor:

- Shaya Wolf (swolf4@uwyo.edu)

Group Members:

- Shawna Wolf (swolf5@uwyo.edu)
- Alicia Thoney (athoney@uwyo.edu)
- [Previously] Woodrow Gamboa

University of Wyoming

CEDAR