**Developing a Framework to Understand Black-Box**

**Adversarial Attacks Against Deep Neural Networks**

Project Proposal

Student

Name: James Kunstle

Education: 4th year BA/MS Student, Boston University

Location: Boston Area

Project Period: Spring Semester 2021

Red Hat Mentor

Name: Lance Galletti

Context

**Deep Neural Networks** (**DNNs**) are becoming an increasingly common technology within most domains of software products. They are used to categorize and recognize input textual, visual, and audio data. As **DNNs** become more reliable and their application becomes more common, ethical and security questions must be answered as to whether they pose a systematic threat to individual privacy.This motivates the need for a formal understanding of what kinds of inputs to an arbitrary **DNN** can potentially avoid accurate analysis by a network, thereby preserving the privacy of the content of the input.

Attacks targeting **DNNs**, which intend to lead to a misclassification pointing to either a target range or in general can be very sophisticated, in some cases (**White-box attacks**) using the weights and parameters of a network to construct **AdvIn**. Other attacks (**Black-box attacks**) create examples via trial and error without access to the architecture or parameters of a model. The latter case is far more likely to exist in the real world and represents a serious threat to systems that rely on **DNNs** in order to make decisions.

There is no formal way to predict the form of **AdvIn**. Therefore, the current state-of-the-art methods for defending against **AdvIn** either require training a model on all kinds of these inputs or on sophisticated but generally less well understood distillation methods (**Defensive Distillation**).

Understanding new ways to create and to defend against **AdvIn** in **Image Recognition and Processing** is critical in order to both protect **DNNs** in the short term via **Adversarial Training** and to understand the complete domain of possible **AdvIn**.

Current high performance **AdvIn** methods such as the **Fast Gradient Sign Method** (**FGSM**) and the **One Pixel Attack** (**OPA**) require aforementioned access to the parameters of the network in order to create targeted and untargeted **AdvIn**. Both of these methods are also within the domain of what this project will call **Additive Adversarial Examples** which change the input additively by adding some form of noise.

Project

This project will focus on a novel approach to the creation of **AdvIn**, i.e., **Subtractive Adversarial Examples** (**SAE**) via transformative and reductive methods. These methods include but are not limited to **Wavelet Transform and Subtraction** and **SVD**.

The objective of this project will be to minimize the accuracy of a well-trained CNN model by using minimally transformed untrained example images. Once this has been completed and measured, the next step will be to investigate potential defensive strategies for these kinds of **AdvIn**.

Expected Deliverable

This project will yield (1) a novel, rigorous method by which **AdvIn** can be created, (2) a numerical estimation of the efficacy of this method in minimizing recognition accuracy with a well-trained **CNN**, (3) experimental evidence to suggest potential defensive strategies for handing these kinds of examples.

These results will be organized as an informal paper that summarizes the methods, measurements, and suggested further work.

Skills Needed for Completion

* Intermediate Python Object-Oriented Programming
  + Student has a strong basis in Python.
* Intermediate PyTorch Programming
  + Student has a novice ability to program in PyTorch.
* Intermediate CV2 Programming
  + Student has strong experience using CV2.