**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**) have become essential to most modern technology infrastructure. Systems use **DNNs** to make increasingly accurate predictions about input textual, visual, and audio data. While incredibly useful in a variety of contexts, **DNNs** pose a threat to the privacy of the individual. Within the visual domain for instance, **DNNs** are able to pick out and recognize individual faces with a high degree of accuracy.

These models are not without their vulnerabilities, however. In 2014 Goodfellow, Shlens, and Szegedy examined a novel technique called the **Fast Gradient Sign Method** which uses a knowledge of the weights of a Deep Visual Network in order to create **Adversarial Examples** (**AE**) which are misclassified by an otherwise well-trained and accurate model. They do this by adding noise to an otherwise clear image that, while not visually salient, effectively changes the structure of the image so that it is not a good input to the model. Su, Vargas, and Sakurai demonstrated in 2019 that forcing a model to fail to categorize an image could be done with changing a maximum of a single pixel using the **One Pixel Attack**. The target pixel, as well as the way it is altered to maximize the miscategorization of the input by the model, also relies on an understanding of the weights and the architecture of the target network.

The aforementioned adversarial strategies rely on a possible but ultimately unlikely set of circumstances. For instance, it is unlikely that the weights and architecture of a non-open-source network would be publicly available. These kinds of “attacks” are called **White-Box Attacks**. **“Black-Box Attacks”** (**BBA**) by contrastcan only leverage the output classification of a network on an **AE** in order to refine adversarial methodologies. These attacks are less well understood but are far more likely to generalize to extant networks.

It is critical therefore to develop a framework by which **BBAs** can be formally studied. By understanding how to create **AEs** to which models are generally vulnerable, there is a greater likelihood of preserving individual privacy.

Project

The goal of this project is to learn a framework by which BBAs can be formally studied and understand how to create AEs for which models are generally vulnerable. We will start by reproducing common Black and White-Box adversarial strategies and developing experimental conditions under which model performance can be measured rigorously we will develop a more cogent understanding of why **AEs** are effective. One example of this will be an analysis of the efficacy of **BBAs** under with constrained bandwidth and fewer opportunities to query a target network. Leveraging this information, new methods for the generation of **AEs** will be found that rely more on the structural vulnerabilities of common classification models than on an altered representation of an individual **AE**.

Expected Deliverable

By the end of this project, we expect to have created the following:

1. A formal method by which the performance of **BBAs** can be measured with appropriate flexibility to apply common constraints.
2. One or more novel **AE** generation strategies that leverage new insights derived from this formal method that can be used to attack networks generally.

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.

References (unformatted)

Fast Gradient Sign Method:

<https://arxiv.org/abs/1412.6572>

One Pixel Attack:

<https://arxiv.org/abs/1710.08864>

Practical Black-Box Attacks:

<https://arxiv.org/abs/1602.02697>