PROJECT PROPOSAL: Adversarial Attacks on Image Recognition

CS229: Machine Learning
Stanford University

Names: Masha (Mikhal) Itkina, Yu Wu

SUNetID: mitkina, wuyu8

Date: October 22, 2016

1. Background and Significance

Machine learning techniques, coupled with data, are used to solve a multitude of high-dimensional problems with great success, such as those in the area of image recognition. For instance, image recognition is employed in self-driving cars to navigate on the roads. However, research has shown that these machine learning models are not robust to adversarial attacks and can be exploited by injecting specifically designed samples to training data or by creating test samples based on the decision boundary of the algorithm to misguide the algorithm. For example, Papernot et. al. showed that it is possible to craft an image that would appear to be a stop sign but would be classified as a yield sign by some class of Deep Neural Network (DNN) [2]. Furthermore, Papernot et. al. also found that the perturbation technique they used to construct such adversarial samples is applicable to a variety of other classifiers, such as linear models and support vector machines [3]. These findings demonstrated the fundamental problems concerning the integrity of machine learning based security systems, and research has been conducted to show the applicability of these vulnerabilities in practical settings such as malware detection [1]. Knowledge and defense against such attacks is crucial for situations in which false positives could lead to dire consequences, such as a collision in the case of a misclassified road sign by an autonomous vehicle.

2. Project Objectives

The purpose of this project is to extend the work done to date in adversarial attacks on image recognition. We will attempt to form an attack on a black box model with an unknown training set by forcing the machine learning algorithm to misclassify an image to an intentionally chosen alternative class. We will be investigating targeted, exploratory adversarial attacks by creating test samples that by-pass the classifier.

3. Methodology and Dataset

We will utilize the German Traffic Sign Recognition Benchmark (GTSRB) dataset that contains over $50,000$ images with over 40 classes. A portion of this dataset will be used to train an image classifier machine learning model. Similarly to previous studies [2],[1],[3], this model will be considered unknown to the adversarial attacker. We will then query the black box model using test examples to generate a substitute model, which will then be used to construct adversarial examples [2]. This substitute model need not be accurate to the black box model as adversarial examples are transferable between different models [2]. The adversarial examples will be constructed such that a particular misclassification is achieved by the black box model.

References

[1] Papernot; Nicolas et. al. Adversarial perturbations against deep neural networks for malware classification. 2016.

[2] Papernot; Nicolas et. al. Practical black-box attacks against deep learning systems using adversarial examples. 2016.

[3] Papernot; Nicolas et. al. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. 2016.