Aadi Shah, Kabir Sahni, Pranav Golla Period 4

Experimentation Comparing Adversarial Robustness Methods

Introduction

In today's world, classifiers, and in particular convolutional neural networks, are used to automate many tasks. This involves the use of machine learning algorithms to classify data into categories. However, many classifiers are vulnerable to adversarial attacks by malicious entities, who may deliberately alter input data such that a classifier misclassifies the data as an incorrect category. As such, numerous methods of increasing the adversarial robustness[1] of machine learning classifiers have been developed to decrease this vulnerability.

Our study compares different algorithms to increase image classifier adversarial robustness. In our study, we train convolutional neural networks using these algorithms and determine which ones have the highest adversarial robustness on data sampled from the CIFAR-10 database.

Methods

In our experiment, the **explanatory variable** (the factor we **manipulate**) is the treatment we are applying to each group; in this case the treatments are the methods of increasing adversarial robustness. We also have a **control** group, where no experimental treatment is applied; instead, a standard machine learning classifier not incorporating any method of increasing adversarial robustness is used. The **response variable** we are measuring is the confidence of each machine learning classifier at predicting the true category of each image in its respective group (a probability). A possible **confounding variable** is the extent to which our models "overfit" to the training data, which may be distinct for different treatments, leading to lower accuracies for some groups. Unfortunately, there is no way to measure or minimize this unless we conduct hyperparameter optimization, a time-consuming process.

We use the CIFAR-10 database, a collection of 60,000 32x32 colour images in 10 classes, as the **population**. We use **stratified sampling** to **randomly select** a **sample** of 10,000 images, 1,000 from each class, from the population. The rest of the images are used to train each of the machine learning models. From the 10,000 images in the sample, **stratified sampling** is used once again to **randomly assign** 2,500 images, each with 250 images from each of the 10 classes, to each of four groups, one control and three experimental. Then, once the machine learning models for each of the treatments are trained, the confidence of each machine learning classifier at predicting the true category of each image in its respective group (a probability for each image) is measured, allowing us to analyze the differences in confidence between different treatments.

Treatments

Control/Standard Treatment: Clean Data Training. This algorithm is the standard training algorithm used to train image classifiers. It involves training a convolutional neural network on clean (unmodified) data.

Treatment A: FGSM Adversarial Training. It involves training models using both clean data and adversarial examples generated by the Fast Gradient Sign Method (FGSM). Adversarial examples created with FGSM introduce small perturbations into clean data during training to improve robustness.

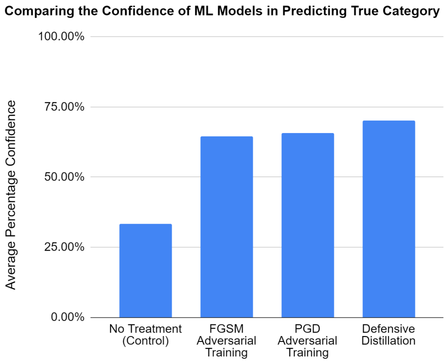
[1] Adversarial robustness refers to the ability of machine learning models to be accurate when given adversarial data

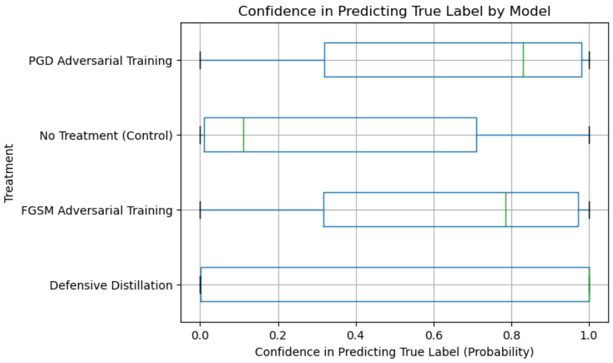
Treatment B: PGD Adversarial Training In this approach, models are trained using both clean data and adversarial examples crafted by the Projected Gradient Descent (PGD) algorithm. PGD generates more powerful adversarial examples by iteratively applying FGSM.

Treatment C: Defensive Distillation. Defensive distillation is a technique where a "student" model is trained on "soft labels" generated by a "teacher" model. The soft labels represent the output probabilities of the teacher model.

Results/Discussion:

Treatment	Stdev
No Treatment (Control)	37.87%
FGSM Adv.	35.34%
PGD Adv.	36.00%
Distillation	45.15%





From the above data, it is evident that Defensive Distillation, on average, is the most effective method of increasing the adversarial robustness of an image classification model, although both types of adversarial training also retain similar robustness. However, noticing that the standard deviation of confidences in predicting the true category of images with Defensive Distillation is much higher than either of the adversarial training methods, it is clear that although Defensive Distillation has high accuracy, on average, it is more susceptible to extremes: either extremely high or extremely low confidence in the true category. This indicates that it generalizes well to much adversarial noise, but still remains vulnerable to some. In contrast, adversarial training generalizes fairly well, and also increases robustness to the vast majority of adversarial noise. Meanwhile, clean data training does not provide much adversarial robustness to image classifiers.

Conclusion:

In this study, an analysis of various methods of enhancing the adversarial robustness of image classifiers was conducted. Three adversarial robustness methods and a clean data (no methods) control were tested. We found that Defensive Distillation had the highest confidence in categorizing images, on average, generalizing well, and that both FGSM and PGD Adversarial Training were consistent at making the model at least somewhat robust to each adversarial example. This study is indicative of each of the treatments tested for all image classifiers trained on CIFAR-10, tested on a sample of CIFAR-10 not used in the training dataset. In addition, due to the concept of transferability, it is likely to apply to most image classifiers as well. There is a cause and effect relationship between the methods to increase adversarial robustness applied to each model and the confidence of each model in predicting the true label of the data, as a prospective study is conducted. For future practitioners of image classifier training, we recommend that Defensive Distillation be used if high confidence should be achieved for most adversarial examples, and for either FGSM or PGD adversarial training to be done if some degree of adversarial robustness should be present for all adversarial examples.