Using Search To Produce Imperceptible Adversarial Examples

*** DRAFT ***

Author: Cristopher McIntyre Garcia
SN: 300025114
Email: cmcin019@uottawa.ca

Introduction

More and more computer vision models using **Machine Learning (ML)** are being implemented into products such as smartphones, vehicles, and medical applications. As such, these models need to be properly verified and tested for faults and vulnerabilities. In doing so, the models, and by extension the products, become safer and more robust against malicious attacks. This project aims to develop a black-box search algorithm to produce imperceptible **Adversarial Examples (AE)**, using a combination of explorative and exploitative methods. This project also aims to determine whether different model architectures are affected by equal image perturbations, such as it has been with Supervised CNNs.

Related Work

The concept of AE was first introduced by [9] as inputs to a ML model that are intentionally designed to cause misclassification. By applying imperceptible non-random perturbation to images, one can change a network's prediction of a correctly classified example. By optimizing the input to maximize the prediction error, adversarial examples are trivial to find and are shared by networks with different amounts of hidden layers or trained on different data subsets. Due to the properties of back-propagation, models are developing intrinsic blind spots that are connected to the data distribution in a non-obvious way.

In [5], it was determined that the linear nature of neural networks is what accords these adversarial vulnerabilities. They also present a fast and simple method of generating AE by adding an imperceptibly small vector whose elements are equal to the sign of the elements of the gradient of the cost function with respect to the input. Using this method, they are able to cause a shallow Soft-Max classifier to have an error rate of 99.9% on the MNIST dataset.

Proceeding these findings, many defences, robust against optimization-based attacks, have been proposed. Though these defences may seem secure at first glance, [2] demonstrates that without a properly computed gradient, iterative optimization-based methods can be circumvented. They proceed to fully circumvent 6 of the 7 defences proposed at ICLR 2018. They also demonstrate that these defences are ineffective against black-box attacks, that don't use the gradient of the cost function to generate AE.

Due to these limitations, it is important to develop defences against black-box attacks. One popular decision-based attack is the Boundary Attack which was first introduced in [3]. In this attack, an adversarial image is produced by taking the original image and adding enough perturbations to make it adversarial. Then, slowly remove portions of the perturbation until it reaches the border of the the non-adversarial image. Either chose this image, or continue along the border, without augmenting the distance between the current perturbed image and the original, in hopes of finding a better perturbed image. The goal is to have a perturbed image that is very close to the original in the image space.

Research Questions

- Are local search algorithms able to efficiently produce adversarial examples? Would introducing a genetic search algorithm increase performance?
- Will a self-supervised transformer falsely classify the same perturbed images as a supervised CNN trained on different subsets of data? If not, can we use the transformer to keep the perceptibility of perturbation at a minimum.
- Are we able to infer any useful information about the model and its classification by studying its AE?

Dataset

For the purpose of this project, I will begin by using the MNIST dataset. If good results are achieved, a more complex dataset such as COCO or ImageNet will be used.

Overview

A simple CNN model will trained a classifier to obtain around 90% accuracy on the MNIST dataset. The same will be done with a transformer model, who will be trained via self-supervised learning. Both models will then be treated as a black-box model, where only the inputs and outputs are known. For each image in the testing set, we want to create a perturbed version that will make the CNN model misclassify the image. However, we also want the classification of the transformer model to not be effected by the perturbations. These will later help us develop objective functions to better our search algorithm.

Results found using the Boundary Attack algorithm will make for a good baseline. Modifications will be made to the algorithm to try and find better results. These modifications include, starting from a non-adversarial image and finding the boundary that separates adversarial and non-adversarial examples, checking all images in parameter as opposed to stopping once a non-adversarial image is found, and finally using a genetic algorithm. A more explorative search method may also be used to find better starting points.

Once a perturbed version of the testing data is created, we will test the accuracy of both the CNN model and the transformer. Because of the different architectures and learning, my hope is that the both the models accuracies will be affected differently. If that is the case, we may be able to use the transformer to find better perturbations that are even less perceptible to humans. We would treat the search as a double optimization problem, where the inputs are trying to maximize the CNN accuracy and maximize the transformer accuracy.

Along the way, the perturbations will be saved for analysis. The perturbed images will be analyzed to verify that the changes are minimal and imperceptible. We can also perform the experiments on the transformer. Most attacks on image datasets have only been performed on CNNs. This leads me to believe that attacking a transformer may lead to new understandings regarding black-box AE. If there is enough time, and good results are found, I would like to attempt to find dataset permutations for several different models and study their differences, if any are found at all.

Analysis

To evaluate the attacks, I will compare the accuracy of the models on the permuted test sets found using the different search approaches. Random search and Boundary Attack will be used as baselines. Our goal for the first question is to determine if we can build an algorithm that out performs these algorithms in terms of minimal accuracy and iterations to generate effective perturbations. If our results are not better, we will attempt to introduce a genetic search algorithm, else we will continue with our project.

To make sure the perturbations are good, we will simply look at the images after being perturbed and evaluate whether the differences are noticeable or not. If different architectures are effected differently by the perturbed images, we will be able to quantify their effectiveness by using the accuracy of the other models as a metric.

Finlay, using different models saving the perturbed image may lead us to discover novel knowledge about the way models learn to classify.