Using Search To Produce Imperceptible Adversarial Examples

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 models of different architectures and learning methods are affected by equal image perturbations, such as they have been with Supervised CNNs.

Related Work

The concept of AE was first introduced by [10] 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 previously correctly classified example. By optimizing the input to maximize the prediction error, AE are trivial to find and are shared by networks with different number of hidden layers or trained on different data subsets. That is, the perturbed images found are also misclassified by CNN models with different configurations. Due to certain properties of back-propagation, models develop intrinsic blind spots that are connected to the data distribution in a non-obvious way.

In [6], it was determined that the linear nature of neural networks is what causes 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. Similar results were found using more complex models, such as GoogLeNet trained on the ImageNet dataset.

Following these findings, many defences that are 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, 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 non-adversarial image. Continue iteratively along the border without augmenting the distance between the current perturbed image and the original. Once it is no longer possible to traverse the border without augmenting the distance between the current perturbed image and the original image, the algorithm returns the perturbed image. The goal of this algorithm is to find AE that are very similar to the original data without any knowledge of the model being used.

Research Questions

- Are local search algorithms able to efficiently produce adversarial examples?
- 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, we 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 be trained as a classifier to obtain around 90% accuracy on the MNIST dataset. The same will be done with a transformer model, such as [5], who will be trained via self-supervised learning. The latter will sometimes be referred to as the oracle. Both models will then be treated as black-box models, where only the inputs and outputs are known. For each image in the test set, we want to create a perturbed version that will make the CNN model misclassify the image. We also want the accuracy of the oracle to be minimally affected by the perturbations.

Results found using the Boundary Attack algorithm will make for a good baseline. Modifications will be made to the algorithm in the hopes of finding better results. These modifications include starting from a non-adversarial image and finding the boundary that separates adversarial and non-adversarial examples by taking "steps" away from the original image. Steps are taken when noise is added to, or removed from, the current perturbed image, as shown in [3]. A few more modifications that may be worth exploring are: checking more images in the perimeter of the image subspace defined by the current perturbed and original image; and randomly choosing a new perturbed image with a smaller distance when no more steps towards the original image can be taken. We can introduce explorative search by creating a grid and choosing an image from every section of the grid in the image space. This grid would comprise all the space surrounding the original image, given a distance.

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 methods, the models' accuracies may 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 optimized to maximize the prediction error of the CNN and minimize the prediction error of the transformer. This will be done using a weighted multi-objective approach.

Along the way, the perturbations will be saved for analysis. The perturbed images will be examined to verify that the changes are minimal and imperceptible. We can also perform the experiments by using the transformer as the attacked model, and the CNN as the oracle. Most attacks seen in the literature 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, we would like to produce perturbed image datasets from several different model configurations and study their characteristics.

Analysis

To evaluate the performance of the attacks, we will compare the accuracy of the models on the permuted test data produced 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 outperforms these baselines in terms of minimal model accuracy and iteration count to generate effective perturbations. If our results are not better than the baseline, the Boundary Attack algorithm will be used in the rest of the experiments.

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 affected differently by the perturbed images, and so one architecture's accuracy isn't heavily affected, we will be able to quantify their effectiveness by using the accuracy of the other models as an extra metric. In other words, if the oracle model loses accuracy, that means the perturbations are not subtle enough.

In conducting these experiments and using different models and algorithms, we hope to determine a practical search algorithm that is capable of producing AE efficiently. Through analysis of our results, we expect to learn new properties regarding adversarial attacks of CNNs, Transformers, Supervised and Self-Supervised. These findings will hopefully help strengthen model resilience and robustness against malicious black-box attacks.