EvadeML-Zoo: A Benchmarking and Visualization Tool for Adversarial Machine Learning

Weilin Xu, Andrew Norton, Noah Kim, Yanjun Qi, David Evans
University of Virginia
https://evadeML.org/zoo

Adversarial machine learning is an important emerging topic in computer security research; however, as an immature area it lacks established benchmarks and testing tools. Researchers frequently develop new methods of attack or defense, but it is difficult to compare their effectiveness to existing results using different experimental setups. In particular, the effectiveness of techniques depends greatly on the dataset, target model and evaluation metrics. Researchers have pointed out that results obtained with non-state-of-the-art models and simple datasets like MNIST often do not extend to more realistic domains.

We have designed and implemented *EvadeML-Zoo*, a benchmarking and visualization tool for research on adversarial machine learning. The goal of *EvadeML-Zoo* is to ease the experimental setup and help researchers evaluate and verify their results.

Dataset and Target Model. We have integrated three popular datasets: MNIST, CIFAR-10 and ImageNet-ILSVRC with a simple and unified interface. We offer several representative pre-trained models with state-of-the-art accuracy for each dataset including two pre-trained models for ImageNet-ILSVRC: the heavy Inception-v3 and and the lightweight MobileNet. We use Keras to access the pre-trained models because it provides a simplified interface and it is compatible with TensorFlow, which is a flexible tool for implementing attack and defense techniques.

Attack Benchmark. We propose a unified framework to benchmark the efficacy and efficiency of an attack algorithm in generating adversarial examples, either in untargeted or targeted mode. The efficacy is represented by a triple of measured values: the success rate, the model confidence, and the amount of distortion. An effective attack algorithm is one that generates a high success rate and high model confidence with low distortion. The efficiency is measured by the time it takes to generate the adversarial examples. We only select the legitimate examples which are correctly recognized by a target model to attack, and always clip and quantitize the adversarial examples to a valid space (e.g., `uint8` data type for image pixels).

We have integrated several existing attack algorithms as baseline for the upcoming new methods, including FGSM, BIM, JSMA, Deepfool, Universal Adversarial Perturbations, and Carlini and Wagner's algorithms.

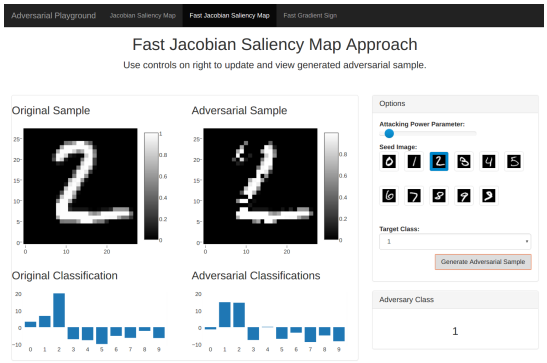

Figure 1: Interactive web-based visualization

Detection Benchmark. We introduce a binary classification task as the detection benchmark. We first construct a balanced dataset with equal number of legitimate and adversarial examples, and then split it into training and test subsets. A detection method has full access to the training set but no access to the labels of the test set. We measure the TPR and FPR on the test set for a detection method as the benchmark results. We have integrated Feature Squeezing as the detection baseline.

Visualization. The tool comes with an interactive web-based visualization module adapted from the ADVERSARIAL-PLAYGROUND package. This module enables better understanding of the impact of attack algorithms on the resulting adversarial sample; users may specify attack algorithm parameters for a variety of attack types and generate new samples on-demand. The interface displays the resulting adversarial example as compared to the original, classification likelihoods, and the influence of a target model throughout layers of the network. Figure 1 shows a screenshot of the visualization module.

Extensibility. *EvadeML-Zoo* has a modular architecture and is designed to make it easy to add new datasets, pre-trained target models, attack or defense algorithms. The code is open source under the MIT license.