Adversarial Attack Detection and Defense with High-level Image Property

Atul Prakash
EECS
University of Michigan
Ann Arbor, US
aparkash@umich.edu

Ryan Feng
EECS
University of Michigan
Ann Arbor, US
rtfeng@umich.edu

Jiaming Zeng
Data Science
University of Michigan
Ann Arbor, US
zjiaming@umich.edu

I. Abstract

Recent work have shown Deep Neural Networks (DNNs) to be vulnerable toward small perturbation with images present test time know as adversarial examples. Our contribution is to show how high-level image property including image structures can be used for ruling out adversarial examples from incorrect classes. Our main technical innovation is by applying high level algorithms and property for detecting adversarial attack toward DNNs, instead of further training the models for defending. The model is through pre-processing the training datasets and using a single model for filtering the classes.

II. Introduction

The Deep Neural Networks (DNNs) have been proven to be solid and been applied to multiple technique fields including the area of computer vision, machine learning and artificial intelligence. However, the DNNs has been shown to be vulnerable toward adversarial attacks which make small perturbation toward the images without human notice [5]. Thus, the defending and detection of the adversarial examples are becoming extra important.

As the attacks improves, the defend methods tend to focus on training the neural network with larger capacity, towards PGD attacked examples, or to cause "gradient masking" [4], [7]. However, we propose the hypothesis that instead of training the neural network itself, the defending of adversarial examples could be achieved through comparing the high-level image properties including image structure and texture. Our intuition is the image properties including the structure are harder to be adversarially attacked with small perturbations. The current adversarial attacks which based on the linearity of neural network models would fail. The current state-of-the-art metrics for evaluating image structural difference could be prone to image distrotion including transformation, rotation and dilation [2], [6]. Thus, by building a classifier with high-level image property, they would become less likely to be attacked.

(Actually, question is metrics without using training not prone to shiting and rotation, but they are better at preventing adversarial attack. However, metrics can prevent have using models like VGG Convolutional layers which fall into the area of PGD attack.)

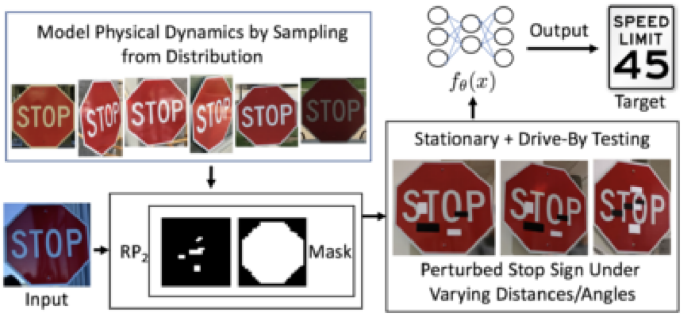

Fig. 1. An image of adversarial example.

Our contributions:

1. We introduce high-level models by taking advantage of k-means clustering and DISTS [2] image different metrics for detecting adversarial examples from distance differences with their true classes.

2. Given the results achieved from high-level methods, we propose a method for filtering the number of classes the adversarial examples can be classified with.

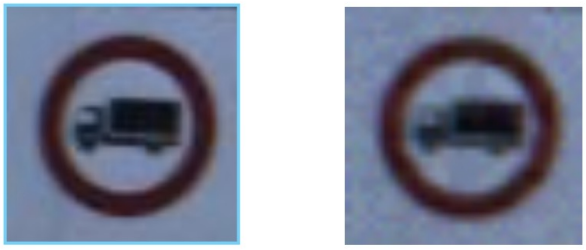

Fig. 2. An image of adversarial example.

III. Related Work

IV. Datasets

Currently, we have only experiment our algorithm with the GTSRB dataset, which is an image classification dataset

consists of photos of traffic signs. There are a total of 43 classes with 39209 training images and 12630 testing images. We choose this dataset as a starting point for evaluating our adversarial examples filter.

V. Methodology

As a general guideline, we will use DISTS metrics to evaluate the distance between test images with our training images. K-means clustering would be applied for getting sub-classes from the original dataset. The specific algorithm for defending which contains two steps are explained below:

A. Preprocess Training Class

Our model requires an extra preprocessing on the image classes. We will use the GTSRB classes as an example. We first calculate the distance for all the training images within the same classes using the DISTS metrics. This would gives us a total of 43 matrix with the number of each classes as their length and width. We can also view the matrix as each images with a vector of distances between all other images in this class. We then apply k-means clustering within each classes into new subclasses. The motivation for this extra preprocessing is to reduce the huge difference within each classes. By clustering the original classes into more subclasses, we are able to implement our algorithms better. Taking GTSRB as an example, we apply k-means clustering to each classes and get 8 subclasses for each class. We get a total of 344 subclasses for later training.

B. Training

The training step is through calculating the distance between held out dataset and training subclasses. We first choose three random images from each subclasses and use them as model images for each class. Then, we calculate the distances between the target held-out image with the three model images and get the mean. After calculating 100 test images with all 344 subclasses, we can get the lowest 10 value for each subclass, and use the lowest 10 as the threshold. Later in the filter process, this threshold value would be the filter. By calculating the distances between test images and model images, we can filter out all the classes for each image which have a distance larger than the threshold value.

VI. Evaluation Metrics

Currently, we will use Exact Match as our evaluation metrics. F1 could be applied later for a softer metrics.

Exact Match The EM measures how many actual results are included in the filtered predictions.

VII. Result

The correctness for the pgd attacked image on neural network model with three convolutional hidden layers with size of 100, 150 and 250 and RELU as the activation. The original accuracy of the neural network with 40 epochs of training are 12593/12648 = 99.56%. After applying the PGD attack with epsilon of 0.031 (which is considered the safe point for most adversarial trained datasets), the accuracy drops to 3668/5736 = 63.95%. I used my model with a smaller testing dataset. The accuracy before any training is 205/240=85.4%, however, after training the correctness drop to 183/240=76%. The difference is not significantly large and proves some level of robustness toward the adversarial attack.

References

[1] Goodfellow, I. J., Shlens, J., Szegedy, C. (2014). Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572.

[2] Ding, K., Ma, K., Wang, S., Simoncelli, E. P. (2020). Image quality assessment: Unifying structure and texture similarity. arXiv preprint arXiv:2004.07728.

[3] Carlini, N., Wagner, D. (2017, May). Towards evaluating the robustness of neural networks. In 2017 ieee symposium on security and privacy (sp) (pp. 39-57). IEEE.

[4] Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A. (2017). Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083.

[5] Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R. (2013). Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199.

[6] Chen, C., Mou, X. (2020). A shift-insensitive full reference image quality assessment model based on quadratic sum of gradient magnitude and LOG signals. arXiv preprint arXiv:2012.11525.

[7] Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z. B., Swami, A. (2017, April). Practical black-box attacks against machine learning. In Proceedings of the 2017 ACM on Asia conference on computer and communications security (pp. 506-519).