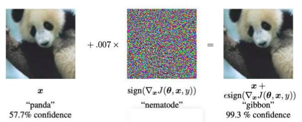
Unrestricted Adversarial Defending Deep Neural Network

Qiwen Wang, Xinshuo Zhang
{qwang26, zxs96}@stanford.edu

INTRODUCTION

➤ **adversarial perturbations**: perturbations designed specially to fool the model into making blatant errors.

➤ **adversarial attack goal:** add a tiny perturbation to the image, which lead to misjudgment of a particular model yet keep the picture classifiable to human eyes:



➤ **our model:** combine several image processing methods and the state-of-art adversarial defending methods to classify attacked digit images "6" and "7".

DATASET AND FEATURES

➤ **Image:** subset of "6" and "7" from MNIST handwritten dataset, colored in black and white..

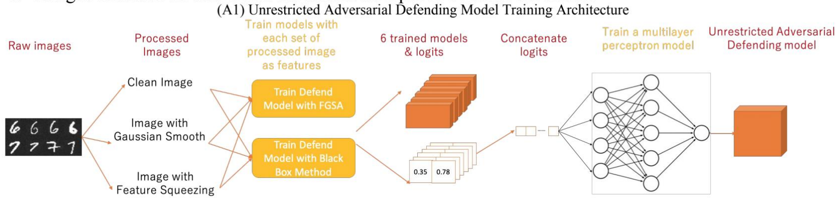
➤ **Image size:** 28*28 pixels

➤ **Dataset size:training:** about 12000 clean images, apply Gaussian smooth and median filter respectively, thus making 36000 training images; **testing:** 2000 images attacked by different methods, including JSMA and FGSM, etc.

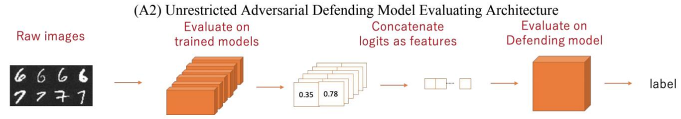
DEEP LEARNING APPROACH

1. Apply image processing techniques to mitigate the attacking effect
2. Train multiple state-of-art adversarial defending algorithms based on CNN with processed images
3. Use logits from defending model results as features to train an integrated MLP model
4. Images with known and unknown attacks as input to evaluate the model

(A1) Unrestricted Adversarial Defending Model Training Architecture



(A2) Unrestricted Adversarial Defending Model Evaluating Architecture



Result

Figure 1 Images attacked by JSMA

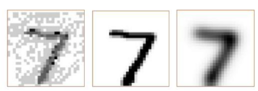


Figure 2:From left to right are the image with FGSM attack, feature squeezing, and Gaussian filter.

➤ Our defending model **outperforms** the baseline, and performs better than most of the intermediate trained models.

preprocess	attack	Acc	Precision	Recall	F1
none	none	0.998	0.996	1	0.998
	spatial_grid	0.51	0.522	0.722	0.606
	fgsm	0.985	0.99	0.981	0.986
	jsma	0.998	0.998	0.999	0.998
gaussian	none	0.999	0.999	1	0.999
	spatial_grid	0.519	0.529	0.734	0.605
	fgsm	0.98	0.99	0.972	0.981
	jsma	0.999	0.999	1	0.999
squeeze	none	0.998	0.998	0.999	0.998
	spatial_grid	0.5125	0.525	0.715	0.605
	fgsm	0.99	0.993	0.988	0.991
	jsma	0.998	0.998	0.999	0.996

Table 1: Result of the defending model. The input testing images is first attacked the and applied image processing methods

Discussion

➤ Our algorithm performs well on other unknown attacks.

➤ Learning the classification result from multiple defend method makes the model more robust on unknown attacks

➤ Applying feature squeezing significantly increases the classification accuracy (emphasize on img feature)

➤ Gaussian Filter mitigate the accuracy. Though it defends the gradient based attack, image features are hard to capture

➤ Performance for the spatial attack is the worst. Because we trained on spatially centered images.

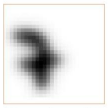


Figure 3: spatial grid attack

Future Work

➤ Apply more image processing and defending methods, such as image restoration, GAN

➤ Generalize to datasets with multi-label and multi-channel (ImageNet, distinguishing birds and bikes, etc)