Detecting Adversarial Examples by Characterizing Adversarial Subspaces

Chia-Yi Hsu

National Chung Hsing University, Taiwan

ABSTRACT

- Dongyu Meng and Hao Chen [1] proposed a defense framework named MagNet using two essential components, including detector(s) and a reformer. Both detectors and the reformer is an auto-encoder.
- Let *f(x)* be the output of the last layer (softmax) of the neural network *f* on the input *x*. Let *ae(x)* be the output of auto-encoder that was trained on normal examples. If x' is an adversarial example, since $ae(x')$ is significantly different from x', the probability mass function $f(x')$ and $f(ae(x'))$ are not similar.

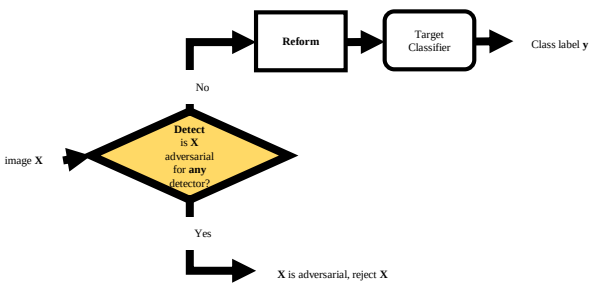

Figure 1: Flow chart of MagNet. When an image enters MagNet, it is detected whether is adversarial or not. And then, if detectors consider an image as an adversarial example, MagNet will filter it out. Otherwise, the image will go through a reformer before entering a classifier.

- In order to compute the similarity between function $f(x')$ and $f(ae(x'))$ we use mutual information. However, mutual information has historically been difficult to compute. Belghazi et al. [2] proposed a method using neural network to estimate mutual information called MINE.

Attack Evaluation on FASHION MNIST and MNIST

The defense rate is the percentage of adversarial examples which are either classified correctly by the classifier or filtered out by detectors.

Transfer attack:

- Generating adversarial examples are from another DNN model.
- Using L_1, L_2 norm reconstruction error based detectors, a reformer, and MID.

FASHION MNIST										
ϵ	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5
PGD	81.8	77.2	72.0	68.0	64.0	57.2	61.7	72.4	84.1	93.0
BIM	84.1	83.8	84.1	84.2	85.0	84.8	84.5	84.7	84.5	84.7

MNIST										
ϵ	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5
PGD	99.9	99.8	99.8	99.8	99.8	99.7	99.8	99.7	99.8	99.7
BIM	99.4	99.5	99.3	99.5	99.5	99.5	99.5	99.5	99.4	99.5

Table 1: The defense rate of FASHION MNIST/MNIST with different epsilons. We used varying attack L_∞ strength with epsilons ranging from 0.05 to 0.5 with an interval of 0.05

FASHION MNIST			
κ	C&W attack(L_2)	EAD attack(L_1)	EAD attack(EN)
0	88.4	87.7	88.3
5	84.6	82.7	82.9
10	80.6	78.7	78.9
15	75.9	74.6	73.3
20	71.9	71.4	70.5
25	68.1	68.3	67.9
30	65.3	66.2	66.4
35	60.1	62.5	62.5
40	56.9	59.4	59.9
MNIST			
κ	C&W attack(L_2)	EAD attack(L_1)	EAD attack(EN)
0	99.0	79.4	78.5
5	99.5	91.1	91.8
10	98.9	87.8	87.17
15	98.0	85.5	84.1
20	98.9	85.5	87.6
25	99.3	85.5	86.6
30	99.3	90.2	87.6
35	99.5	89.9	85.8
40	99.4	92.2	89.6

Table 2: The defense rate of FASHION MNIST/MNIST with different confidences κ.

CONCLUSION

- Our results demonstrate that mutual information is a promising approach to characterizing adversarial subspaces.

REFERENCE

[1] Dongyu Meng and Hao Chen. 2017. MagNet: a Two-Pronged Defense against Adversarial Examples. ACM CCS (2017).

[2] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and R Devon Hjelm. 2018. Mine: mutual information neural estimation. arXiv preprint arXiv:1801.04062 (2018).