POSTER: Characterizing Adversarial Subspaces by Mutual Information

Chia-Yi Hsu
National Chung Hsing University
Taiwan
chiayihsu8315@gmail.com

Pin-Yu Chen
IBM Research
USA
pin-yu.chen@ibm.com

Chia-Mu Yu
National Chung Hsing University
Taiwan
chiamuyu@gmail.com

ABSTRACT

Deep learning is well-known for its great performances on images classification, object detection, and natural language processing. However, the recent research has demonstrated that visually indistinguishable images called adversarial examples can successfully fool neural networks by carefully crafting. In this paper, we design a detector named MID, calculating mutual information to characterize adversarial subspaces. Meanwhile, we use the defense framework called MagNet and mount the detector MID on it. Experimental results show that projected gradient descent (PGD), basic iterative method (BIM), Carlini and Wanger's attack (C&W attack) and elastic-net attack to deep neural network (elastic-net and L_1 rules) can be effectively defended by our method.

KEYWORDS

adversarial examples, neural networks

ACM Reference Format:
Chia-Yi Hsu, Pin-Yu Chen, and Chia-Mu Yu. 2019. POSTER: Characterizing Adversarial Subspaces by Mutual Information. In *ACM Asia Conference on Computer and Communications Security (AsiaCCS '19), July 9–12, 2019, Auckland, New Zealand*. ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3321705.3331002

1 INTRODUCTION

Deep learning shows brilliant performance on many tasks in artificial intelligence and machine learning, such as images recognition, visual art processing, and automatic speech recognition. However, recent research has demonstrated that well-trained deep neural networks (DNNs) can be broken by adversarial examples. There have been many efforts on defending adversarial examples. Papernot et al. [12] defend against adversarial perturbations by using the distillation technique. Madry et al. [9] propose that using adversarial training can increase the robustness of DNNs to adversarial examples. Papernot et al. [11] and Tramèr et al [14] demonstrate that gradient masking Detection methods utilize statistical tests to separate adversarial from normal examples. However, Carlini and Wanger's attack can bypass 10 different detection methods. Dongyu Meng and Hao Chen [10] propose a defense framework named MagNet consisting of adversary detectors and reformer. Throughout this paper, we use MagNet as a defense architecture, and propose to improve its performance by incorporating MID as a new detector.

AsiaCCS '19, July 9–12, 2019, Auckland, New Zealand

ACM ISBN 978-1-4503-6752-3/19/07.
https://doi.org/10.1145/3321705.3331002

2 RELATED WORK

In order to evaluate MID we using four attacks: PGD[6], BIM[5], C&W attack[2] and elastic-net attack (EAD) to deep neural networks [3], respectively. In addition, we used MagNet as defense framework.

2.1 Basic iterative method (BIM)

Linearizing the cost function J and solving the perturbations that maximizes the cost subject to a L_∞ constraint is one of the simplest methods to generate adversarial examples. BIM applies it multiple times with small step sizes and clip the pixel values of intermediate results after each step to guarantee that they are in the $\epsilon-$neighbourhood of the original images.

$$X_0^{adv} = X,\ X_{N+1}^{adv} = Clip_{X,\epsilon}\left\{X_N^{adv} + \alpha\ \mathrm{sign}\left(\nabla_X J(X_N^{adv}, y_{true})\right)\right\}$$

Besides, when generating the adversarial examples by BIM with randoms starts, the attack becomes the "PGD attack"[9].

2.2 Carlini and Wanger's attack (C&W attack)

Carlini and Wanger [2] propose a strong attack that is able to craft adversarial examples with small perturbations and high transferability by tuning the *confidence* parameter. They transform the issue of generating adversarial examples into the optimization problem. C&W attack is either untargeted or targeted attack for all three metrics L_0, L_2 and L_∞. In this paper, we used the L_2 metric.

We formalize the C&W attack (L_2) as the following optimization problem:

$$\underset{\delta}{\text{minimize}} \quad \|\delta\|_2^2 + c \cdot f(x+\delta)$$

$$\text{such that} \quad x + \delta \in [0,1]^n$$

with f defined as

$$f(x') = \max\left\{\mathrm{Z}(x')_{l_x} - \max(\mathrm{Z}(x')_i : i \neq l_x), -\kappa\right\}.$$

The loss f is the best objection function found earlier, adjusted slightly so that we can manipulate the confidence with which the misclassification occurs by modifying κ. Larger κ encourages the solver to find an adversarial instance x' which will be classified as label i with high confidence. Furthermore, attacking larger κ often gives higher transferability and larger perturbation. $\mathrm{Z}(x')$ is the output of logit layer (pre-softmax layer) and l_x is the ground truth label.

2.3 EAD: Elastci-Net Attack

Elastic-net regularization can be regarded as a regularizer which linearly combines L_1 and L_2 penalty functions. Also, EAD attack has two decision rules, one is *elastic-net* (EN) and another is L_1 distortion. Let (x_0, t_0) denote a natural instance with ground truth label t_0 and let (x, t) denote an adversarial example with label $t \neq t_0$. The L_n norm of the image difference $\delta = \mathrm{x}_0 - \mathrm{x}$, defined as $\|\delta\|_n = (\sum_{i=1}^{p} |\delta_i|^n)^{\frac{1}{n}}$. It is widely to use distortion between benign and adversarial examples when n is greater than 0.

We formalize the EAD attack as the following optimization problem:

$$\begin{aligned} &\text{minimize}_{\mathrm{x}} \quad c \cdot f(\mathrm{x}, t) + \|\mathrm{x} - \mathrm{x}_0\|_2^2 + \beta\|\mathrm{x} - \mathrm{x}_0\|_1 \\ &\text{subject to} \quad \mathrm{x} \in [0,1]^p \end{aligned}$$

$c, \beta \geq 0$ are regularization parameters of loss function f and L_1 penalty, respectively. It can generate more effective and higher transferability adversarial examples as β increasing [3, 13].

Notablely, C&W attack is a specific case in EAD attack when β is equal to 0, giving rise to an alone L_2 distortion based attack. EAD attack can often generate adversarial examples that are more difficult to be distinguished from natural examples[4, 7]. Besides, it often exhibits a higher chance of bypassing detectors than C&W attack.

2.4 MagNet: Defending Adversarial Examples with Detector and Reformer

MagNet is a defense framework using two essential components, including detector(s) and a reformer. Both detectors and the reformer is an auto-encoder denoted by $AE(\mathrm{x})$ taking x as input. 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 the default setting, there are three detectors in MagNet. One is the L_1 based detector, computing L_1 norm distance between x and $AE(\mathrm{x})$. Another is L_2 norm based detector, denoted $\|x - AE(\mathrm{x})\|_2$. The other is the $\mathcal{JSD}$ detector, computing Jensen-Shannon divergence between $F(\mathrm{X})$ and $F(AE(\mathrm{X}))$. $F(\cdot)$ denotes the output of the classifier.

MagNet showed robust defense performance against C&W attack (L_2) under different confidence levels in the *oblivious* attack setting where adversarial examples are generated from the undefended DNN on MNIST and CIFAR 10. In addition, MagNet can also defend transfer attack: fast gradient sign method (FGSM), iterative FGSM and DeepFool. However, Lu et al.[8] demonstrate that MagNet is ineffective against L_1 distortion based adversarial examples generated by EAD attack.

2.5 MINE: Mutual Information Neural Estimation

Mutual information is a fundamental quantify for measuring the similar information between two sets. It is widely used in many domains and tasks including feature selection, blind source separation (BSS) and biomedical sciences. However, mutual information has historically been difficult to compute. The mutual information between two random variables X and Z can be formalized as below:

$$I(X, Z) = H(X) - H(X|Z),$$

where H is the Shannon entropy, and $H(X|Z)$ is the conditional entropy of X given Z, which is also difficult to compute $H(X|Z)$. MI Belghazi et al. [1] proposed a method using neural network to estimate mutual information called MINE. The idea is to choose $\mathcal{F}$ to be the family of functions $T_\Theta : \mathcal{X} \times \mathcal{Z} \to \mathbb{R}$ parametrized by a deep neural network with parameters $\theta \in \Theta$. We formalize the bound:

$$I(X, Z) \geq I_\Theta(X, Z),$$

where $I_\Theta(X, Z)$ is the neural information quantity defined as

$$I_\Theta(X, Z) = \sup_{\theta \in \Theta} \mathbb{E}_{\mathbb{P}_{XZ}}[T_\theta] - \log(\mathbb{E}_{\mathbb{P}_X \otimes \mathbb{P}_Z}[e^{T_\theta}]).$$

Using empirical samples from $\mathbb{P}_{\mathbb{XZ}}$ and $\mathbb{P}_{\mathbb{X}} \otimes \mathbb{P}_{\mathbb{Z}}$ or by shuffling the samples from the joint distribution via batch axis estimates the expectations above. The objective function can be maximized by gradient ascent.

3 OUR PROPOSED METHOD

In this paper, we design a detector called MID and use the entropy divergence to detect adversarial examples. We use mutual information to quantify similar information between two sets. Because mutual information is difficult to compute, we use MINE to estimate it improving the results in [4]. We calculate $I_\Theta(X, Z)$, denoted $X = f(\mathrm{x})$ and $Z = f(AE(\mathrm{x}))$. $f(\cdot)$ is the classifier whose output is the last layer (softmax) of the neural network, $AE(\cdot)$ is the output of auto-encoder trained by natural instances and x is the input image. We consider that if $I_\Theta(X, Z)$ is lager, the input x will be a normal example because they have more similar information. In other words, input x is viewed as adversarial when $I_\Theta(X, Z)$ is small.

4 EXPERIMENTS

We showed the results of MID defending PGD, BIM, C&W attack and EAD attack in this section. We used the FASHION MNIST and MNIST which are known for the image classification datasets to carry out untargeted adversarial attacks. FASHION MNIST and MNIST are black and white with ten categories.

4.1 Experiment Setup and Parameter Setting

We used the transfer attack and implemented the untargeted transfer attack from an undefended DNN to a defended DNN protected by MagNet and MID. We randomly selected 1000 correctly classified images from a test set to attack MagNet+MID.

For the FASHION MNIST dataset, we only used L_1, L_2 norm reconstruction error based detectors, a reformer, and MID. We selected thresholds of reconstruction errors by dropping out 0.1% examples in the validation set. We set the threshold of MID false positive rate to be 0.005 based on a validation set with normal samples.

We used EAD attack (EN and L_1 rules) to tamper MagNet+MID and chose the L_1 distortion coefficient $\beta = 10^{-1}$. We used the L_2 version of C&W attack. For PGD and BIM, we used tvarying attack L_∞ strength with epsilons ranging from 0.05 to 0.5 with an interval of 0.05.

4.2 Attack Evaluation on FASHION MNIST

Effect on normal examples. On the test set without any defense, the classification accuracy could achieve 94.34%; with MID, the

Table 1: The defense rate of FASHION MNIST and MNIST with different epsilons ϵ.

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

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

Table 3: The defense rate of MNIST with different confidences κ.

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

accuracy decreased to 89.57%. By tuning the false positive rate, we could gain more robust defense at the expense of test accuracy.

Effect on adversarial examples. The defense rate is the percentage of adversarial examples which are either classified correctly by the classifier or filtered out by detectors. Table 1 presented that the defense rate of MID was above 80% on BIM and also have effective performance on PGD in different epsilons ϵ. Table 2 showed that MID had great performance with different confidences levels κ on both C&W attack and EAD attack (L_1 and EN rules). When $\kappa \leq 20$, the defense rate could achieve 70% above on either C&W or EAD attack.

4.3 Attack Evaluation on MNIST

Effect on normal examples. On the test set without any defense, the classification accuracy could achieve 99.42%; with MID, the accuracy decreased to 98.83%. It showed that MID slightly impacted on accuracy. *Effect on adversarial examples.* Table 1 showed that we can defend effectively on both PGD and BIM. Table 3 showed that the defense rate in different confidences κ can achieve 80% above on both C&W and EAD attack. MNIST is simpler than FASHION MNIST to defend.

5 CONCLUSION

In this paper, we design a detector for adversarial inputs based on mutual information and the recent MINE framework for efficient computation. Tested on several state-of-the-art attacks in the oblivious attack setting, our results demonstrate that mutual information is a promising approach to characterizing adversarial subspaces.

REFERENCES

[1] 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).

[2] Nicholas Carlini and David Wagner. 2017. Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy (SP)*. 39–57.

[3] Pin-Yu Chen, Yash Sharma, Huan Zhang, Jinfeng Yi, and Cho-Jui Hsieh. 2018. EAD: elastic-net attacks to deep neural networks via adversarial examples. In *Thirty-second AAAI conference on artificial intelligence*.

[4] Chia-Yi Hsu, Pei-Hsuan Lu, Pin-Yu Chen, and Chia-Mu Yu. 2018. On The Utility of Conditional Generation Based Mutual Information for Characterizing Adversarial Subspaces. In *2018 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*. IEEE, 1149–1153.

[5] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. 2016. Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533* (2016).

[6] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. 2016. Adversarial machine learning at scale. *ICLR'17; arXiv preprint arXiv:1611.01236* (2016).

[7] Pei-Hsuan Lu, Pin-Yu Chen, Kang-Cheng Chen, and Chia-Mu Yu. 2018. On the limitation of magnet defense against L1-based adversarial examples. In *2018 48th Annual IEEE/IFIP International Conference on Dependable Systems and Networks Workshops (DSN-W)*. IEEE, 200–214.

[8] Pei-Hsuan Lu, Pin-Yu Chen, Kang-Chen Chen, and Chia-Mu Yu. 2018. On the Limitation of MagNet Defense against L1-based Adversarial Examples. *IEEE/IFIP International Conference on Dependable and Systems and Networks (DSN) Workshop on Dependable and Secure Machine Learning, arXiv:1805.00310* (2018).

[9] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2017. Towards Deep Learning Models Resistant to Adversarial Attacks. *arXiv preprint arXiv:1706.06083* (2017).

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

[11] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. 2017. Practical black-box attacks against machine learning. In *ACM Asia Conference on Computer and Communications Security*. 506–519.

[12] Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. 2016. Distillation as a defense to adversarial perturbations against deep neural networks. In *IEEE Symposium on Security and Privacy (SP)*. 582–597.

[13] Yash Sharma and Pin-Yu Chen. 2018. Attacking the Madry Defense Model with L_1-based Adversarial Examples. *ICLR Workshop; arXiv:1710.10733* (2018).

[14] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Dan Boneh, and Patrick McDaniel. 2017. Ensemble Adversarial Training: Attacks and Defenses. *arXiv preprint arXiv:1705.07204* (2017).