Poster: REMIX: Mitigating Adversarial Perturbation by Reforming, Masking and Inpainting

Kang-Cheng Chen
Yuan Ze University, Taiwan
skywind29@gmail.com

Pin-Yu Chen
IBM Thomas J. Watson Research Center, USA
pin-yu.chen@ibm.com

Chia-Mu Yu
National Chung Hsing University, Taiwan
chiamuyu@nchu.edu.tw

I. Introduction

Deep learning has achieved remarkable success in various areas, such as computer vision and natural language processing. However, recent studies also highlight that deep neural networks (DNNs) are vulnerable to *adversarial examples* created by adding small but purposefully crafted perturbations to natural targets. These perturbations may lead to misclassification on DNNs and jeopardize the application safety. Although many countermeasures have been proposed, most of them only detect/reject adversarial examples, instead of reforming adversarial examples back to corrected ones.

Here, we propose REMIX to distill and rectify adversarial perturbations toward correct classification. Unlike other proposals (e.g., [9]), REMIX dispenses the assumption of prior knowledge about how attackers craft adversarial examples.

II. Related Work and Background

Existing Attacks. Various adversarial perturbations can be crafted using optimization-based approaches. In this paper, we use two state-of-art attacks to evaluate our proposed defense.

- **CW Attack.** Carlini and Wagner attack [2] can generate adversarial examples with imperceptible perturbations. Given a natural image x and an adversarial perturbation δ, CW attack can be formulated as the following optimization problem:

$$\begin{aligned} &\text{minimize}_{\delta} \quad ||\delta||_2^2 + c \cdot f(x+\delta) \\ &\text{s. t. } x+\delta \in [0,1]^n. \end{aligned} \tag{1}$$

  We refer readers to [2] for the attack loss $f(\cdot)$ and method for choosing the constant c.
- **EAD Attack.** Generalizing from CW attack, Chen et al. [5] propose an elastic-net regularized attack. Specifically, EAD features ℓ_1-based adversarial examples and include the CW attack as a special case. In many cases, EAD attack is easier to bypass the defense than CW attack.

Both EAD and CW attacks have a confidence level parameter κ to control the transferability in adversarial examples. Higher κ results in more transferable adversarial examples.

Existing Defenses. Many defenses have been shown to be ineffective against optimization-based attacks [1], [4]. Recently, Meng and Chen propose MagNet [7], which uses the complementarity of reformer and detector to achieve significant defensive capability. However, MagNet cannot defend against adaptive white-box attacks [3]. Samangouei et al. propose the Defense-GAN [8], which uses the randomness of GAN to defend white-box attacks. However, due to the low reconstruction ability of GAN, Defense-GAN cannot be applied to more complex datasets, e.g. CIFAR, ImageNet, etc.

Threat Model. Given a defense d_f, depending on an attacker's knowledge, attacks are divided into three categories:

- Black-box: attacker knows nothing about d_f.
- Gray-box: attacker knows everything except the network parameters of d_f, e.g., structure, training process, etc.
- White-box: attacker knows everything including the network parameters of d_f.

III. Proposed REMIX Solution

Successfully adversarial perturbations are carefully designed and are expected to be as small as possible in order to maintain similarity to natural examples. Motivated by this phenomenon, the delicate perturbations can be possibly vulnerable to disruptions and hence the adversarial examples can be rectified. Using this philosophy, we propose a defense mechanism, REMIX, that mitigates adversarial perturbations by reforming, randomly masking, and inpainting masked images to rectify adversarial examples.

REMIX is composed of three modules: reformer, random mask generator and inpainter, as illustrated in Fig. 1. The reformer module is optional but we find that it is effective in defending against adversarial examples with higher confidence. Below we describe the designs and functionalities of each component. The detailed performance analysis is given in section IV.

Reformer. Similar to MagNet's reformer, we use denoising convolutional autoencoder as the reformer of REMIX. However, to maintain the test accuracy, we use the simplified convolutional autoencoder with symmetric skip connections [6] as our reformer. The reformer takes images as inputs and outputs the reformed images by learning the manifold of natural images. Nonetheless, the reformer is useful in mitigating adversarial examples of low and moderate confidences, but is often misled by high-confidence adversarial examples.

Random Mask Generator. As we randomly mask an input image, the structure (e.g., masked pixels for inpainting) of adversarial perturbations would be disrupted while an natural image (or its correct label) could be restored via inpainting, which is trained with masked natural images. When random masking an colored image x, we first generate a black image m with $|m| = |x|$[1]. The pixel positions[2] 1, . . . , $|m|$ are randomly

[1] The notation $|x|$ denotes the size of an image x.

[2] We assume a natural order of pixel positions.

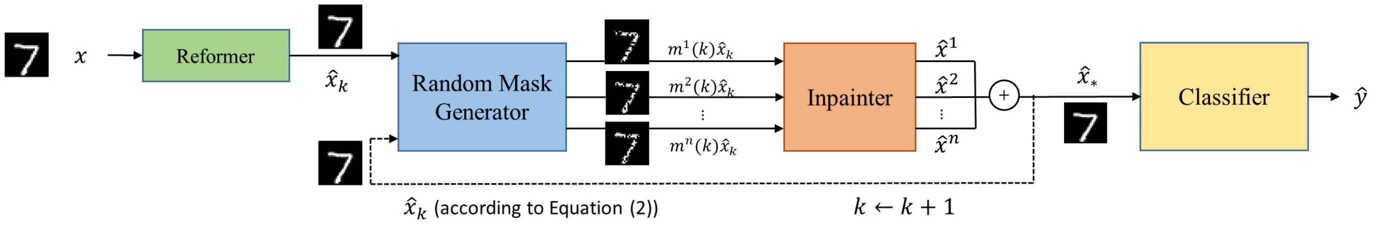

Fig. 1. Overview of the REMIX structure. The mask-inpaint procedure could be repeated multiple times and thus we use the notations $\hat{x}_k$, $m^i(k)$, and $\hat{x}_*$ to denote $\hat{x}$ in the kth iteration, m^i in the kth iteration, and the output of the final iteration of mask-inpaint procedure, respectively. Note that in the very first iteration $k = 0$, the notation $\hat{x}_0 = \hat{x}$ refers to the output the reformer.

partitioned into n sets, $P_1, \ldots, P_n$. We generate n random masks $m^1, m^2, \ldots, m^n$ in such a way that $m^i_j = 0$ if $j \in P_i$ and $m^i_j = 1$ otherwise, where m^i_j denotes the jth pixel of m^i. After mask generation, we then apply each of them to x, so as to produce n masked images $m^1x, m^2x, \ldots, m^nx$, where m^ix is the pixel-wise multiplication of m^i and x. For gray-scale images, we generate masks in a similar way.

Inpainter. Like reformer, since convolutional autoencoder with symmetric skip connections performs well in high-resolution image restoration, we use the simplified one as our inpainter. The major task of inpainter is to inpaint (restore) the masked images back to original images as close as possible. Taking the masked images $m^1x, \ldots, m^nx$ as input, inpainter will generate the corresponding inpainted images $\hat{x}^1, \ldots, \hat{x}^n$. After that, we combine the inpainted images according to their original masked positions to derive an integrated inpainted image $\hat{x}$, which can be formulated as

$$\hat{x} = \sum_{i}^{n} \hat{x}^i \cdot (1 - m^i). \tag{2}$$

The entire mask-inpaint procedure can be repeated multiple times. As the number of repetitions increases, we have more chances to drive high-confidence adversarial examples to the correct class prediction, but may degrade the quality of images.

IV. Experiment Setup

We compare REMIX with MagNet under CW and EAD attacks on MNIST and CIFAR-10 datasets in the oblivious attack setting, where the attackers can access the DNN model parameters but are unaware of the deployed defenses. For EAD attack, we use the elastic-net (EN) distortion decision rule with the regularization parameter $\beta = 0.1$ to generate the adversarial examples. For both CW and EAD attacks, 1000 adversarial examples are crafted with confidence level κ in the range $[0, 40]$ and $[0, 100]$ on MNIST and CIFAR-10, respectively. In REMIX, we use $n = 2$ masks on MNIST and $n = 3$ masks on CIFAR-10, and do not repeat the mask-inpaint procedure, though multiple iterations of mask-inpaint operations are allowed. The protected classifier has 99% accuracy on MNIST and 86% accuracy on CIFAR-10. The experimental results are shown in Fig. 2.

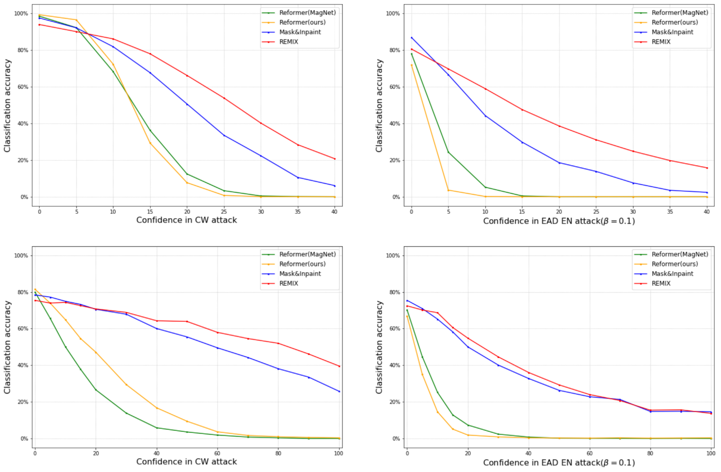

Fig. 2. Defense performance with different confidence of CW and EAD attacks on MNIST (first row) and CIFAR-10 (second row). The performances are evaluated by the percentage of correctly classified adversarial examples.

V. Discussions and Conclusions

Comparing to MagNet, our mask-inpaint mechanism outperforms their reformer in most of confidence levels. Additionally, when combined with the reformer, our REMIX defense against high-confidence adversarial examples can be enhanced quite significantly. It is worth mentioning that since our defense has randomness in the test time, it imposes additional challenges on white-box attacks. More attack iterations and distortions are expected to bypass random masking and inpainting if possible. There are also many potentials of REMIX to be explored; for example, the distributional difference in $\hat{x}^1, \ldots, \hat{x}^n$ could be used to detect adversarial examples by measuring their distances or divergences from the input images.

References

[1] A. Athalye, N. Carlini and D. Wagner. Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples. arXiv:1802.00420.

[2] N. Carlini and D. Wagner. Towards evaluating the robustness of neural networks. *IEEE Symposium on Security and Privacy*, 2017.

[3] N. Carlini and D. Wagner. MagNet and "Efficient Defenses Against Adversarial Attacks" are Not Robust to Adversarial Examples. arXiv:1711.08478.

[4] N. Carlini and D. Wagner. Adversarial Examples Are Not Easily Detected: Bypassing Ten Detection Methods. *AISec*, 2017.

[5] P.-Y. Chen, Y. Sharma, H. Zhang, J. Yi, and C.-J. Hsieh. Ead: Elastic-net attacks to deep neural networks via adversarial examples. *AAAI*, 2018.

[6] X.-J. Mao, C. Shen, and Y.-B. Yang. Image Restoration Using Convolutional Auto-encoders with Symmetric Skip Connections. *NIPS*, 2016.

[7] D. Meng and H. Chen. MagNet: a Two-Pronged Defense against Adversarial Examples. *ACM CCS*, 2017.

[8] P. Samangouei et al. Defense-GAN: Protecting Classifiers Against Adversarial Attacks Using Generative Models. *ICLR*, 2018.

[9] S. Shen et al. APE-GAN: Adversarial Perturbation Elimination with GAN. arXiv:1707.05474.