Defending against Whitebox Adversarial Attacks

Speaker: Yi-Shan Wu

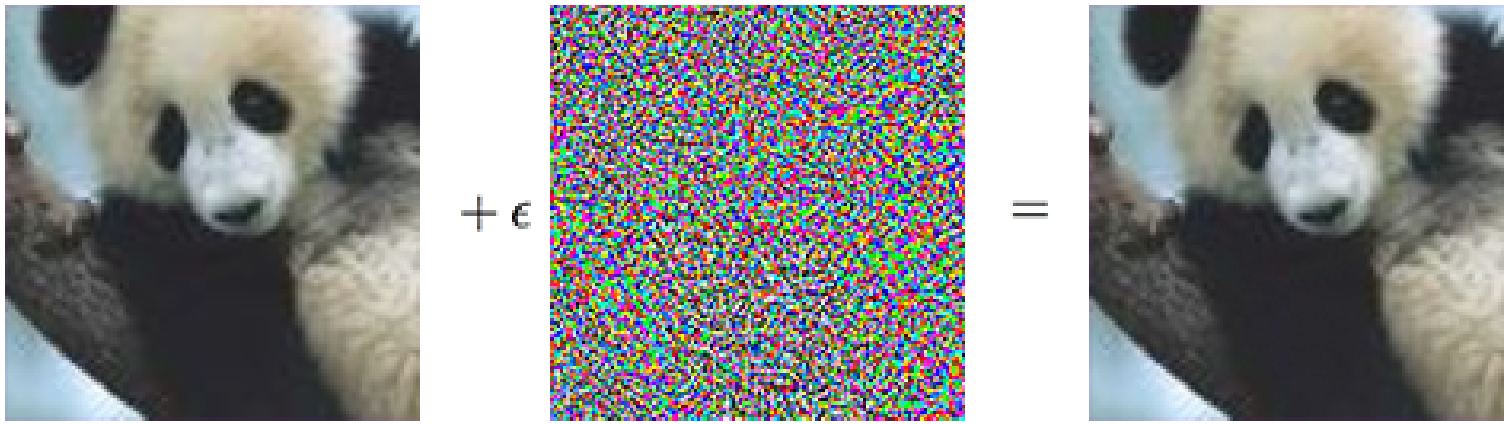
Institute of Information Science

Academia Sinica

Taiwan

May 03, 2019

What is adversarial attack ?



Such vulnerability exposes security concerns.

What is adversarial attack ?

- An **adversarial image** is an image that has been slightly modified in order to be misclassified.
- In **white box attacks** the attacker has access to the model's parameters, while in **black box attacks**, the attacker has no access to these parameters.
- The **non-targeted attacks** aim to enforce the model to misclassify the adversarial image, while the **targeted attacks** aim to get the image classified as a specific target class.

Common Attacks

Most successful attacks are **gradient-based methods**. Namely, the attackers modify the image in the direction of the gradient of the loss function with respect to the input image.

- One-shot attack : attacker takes a single step in the direction of the gradient
- Iterative attack : is stronger than one-shot attack in the whitebox setting

iterative attack - projected gradient descent (PGD)

ℓ_∞-norm attack

Let x be the clean image with n pixels and q channels (colors). Let x' be the adversarially perturbed image such that

$$\|x - x'\|_\infty \leq \epsilon$$

PGD algorithm :

$$x_{t+1} \leftarrow \Pi(x_t + \eta \cdot sign(\nabla_x f(x_t))),$$

where $\Pi(\cdot)$ is the projection function.

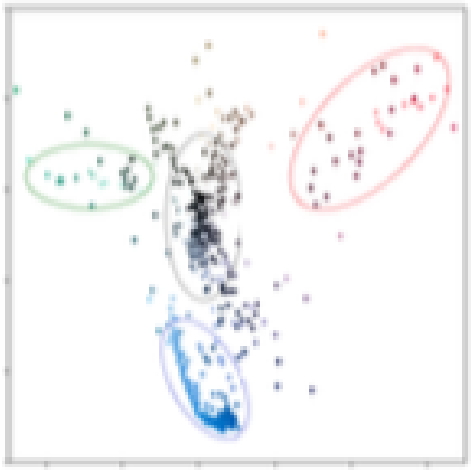
Common defences

1. **adversarial training :** train the model on adversarial examples.
   – successful on MNIST and CIFAR-10
2. **retraining :** modify the network, data augmentation
3. **pre-processing :** pre-process the image and then feed into a pre-trained classifier, do not require retraining
   - BitDepth : reduce the bit-depth of each RGB channel
   - JPEG : compression
   - ResizePadding
   - Total variation minimization

   – rarely holds against whitebox attacks

stochastic transformation

Motivation : Image pixels in the color space often cluster.



1. For pixels close to a particular cluster center, their cluster assignments will be stable under perturbations, and thus robust to the adversarial attack.
2. For pixels roughly equidistant from two centers, their cluster assignments can be randomized by the injection of Gaussian noise, which mitigates the effect of the adversarial attack.

stochastic transformation

- Gaussian randomization (Gaussian)
- randomized discretization (RandDisc)
- randomized mixure (RandMix)

framework

- original image : x, x'
- transformed image : $\tilde{x}$, $\tilde{x}'$

– since x and $\tilde{x}$ have same semantics, we can feed them directly into any *base classifier*

Gaussian

Simply adds Gaussian noise to raw pixels of the image.

$$\forall i \in [n], \tilde{x}_i = x_i + w_i, \text{ where } w_i \sim \mathcal{N}(0, \sigma^2 I)$$

RandDisc

1. Add Gaussian noise to each pixel $\Rightarrow x_i + w_i$.
2. Select k cluster center $\mathbf{c} := (c_1, \cdots, c_k)$.
3. Assign the noisy pixel $x_i + w_i$ to its nearest center.

$$\forall i \in [n], \tilde{x}_i = r(x_i|\mathbf{c}) := arg \min_{c \in \{c_1, \cdots, c_k\}} \|x_i + w_i - c\|_2$$

RandMix

1. The first two steps are the same as RandDisc. Instead of assigning to a single cluster, we compute the weighted mean :

$$\forall i \in [n], \tilde{x_i} = m(x_i|\mathbf{c}) := \frac{\sum_{j=1}^{k} c_j \cdot e^{-\alpha\|x_i+w_i-c\|_2^2}}{\sum_{j=1}^{k} e^{-\alpha\|x_i+w_i-c\|_2^2}}$$

- When $\alpha \to \infty$, RandMix converges to RandDisc.
- $m(x_i|\mathbf{c})$ is differentiable.

Certified Accuracy

The gold standard for security is having a certificate that a defense provably works against all attacks.

Idea

Suppose thae $\tilde{x}$ is correctly classified by a base classifier f. The goal is then to derive a **sufficient condition** under which the pre-processed perturbed image $\tilde{x}'$ is also correctly classified by the same base classifier f.

If the two distributions $\tilde{x}$ and $\tilde{x}'$ are close enough, then from an information-theoretic point of view, no algorithm can distinguish the two transformed images.

Certified Accuracy

Proposition

$D_{KL}(p_{\tilde{x}} \| p_{\tilde{x}'})$ is upper bounded

This paper shows that randomized discretization reduces the KL divergence between the clean image and the adversarial image. Also, RandDisc's upper bound will be much smaller than that of the Gaussian.

Certified Accuracy

$margin(\tilde{x}, y, f) := P(f(\tilde{x}) = y) - max_{z \neq y} P(f(\tilde{x}) = z)$, therefore,

$$
\begin{aligned}
|margin(\tilde{x}', y, f) - margin(\tilde{x}, y, f) \quad | &\leq \quad 2\|p_{f(\tilde{x})} - p_{f(\tilde{x}')}\|_{TV} \\
&\leq \quad \sqrt{2 D_{KL}(p_{f(\tilde{x})} \| p_{f(\tilde{x}')})} \\
&\leq \quad \sqrt{2 D_{KL}(p_{\tilde{x}} \| p_{\tilde{x}'})}
\end{aligned}
$$

This implies that if $D_{KL}(p_{\tilde{x}} \| p_{\tilde{x}'})$ is strictly smaller than $\frac{1}{2}(margin(\tilde{x}, y, f))^2$, then the classifier's probability of making the correct prediction will be greater than any incorrect probability.

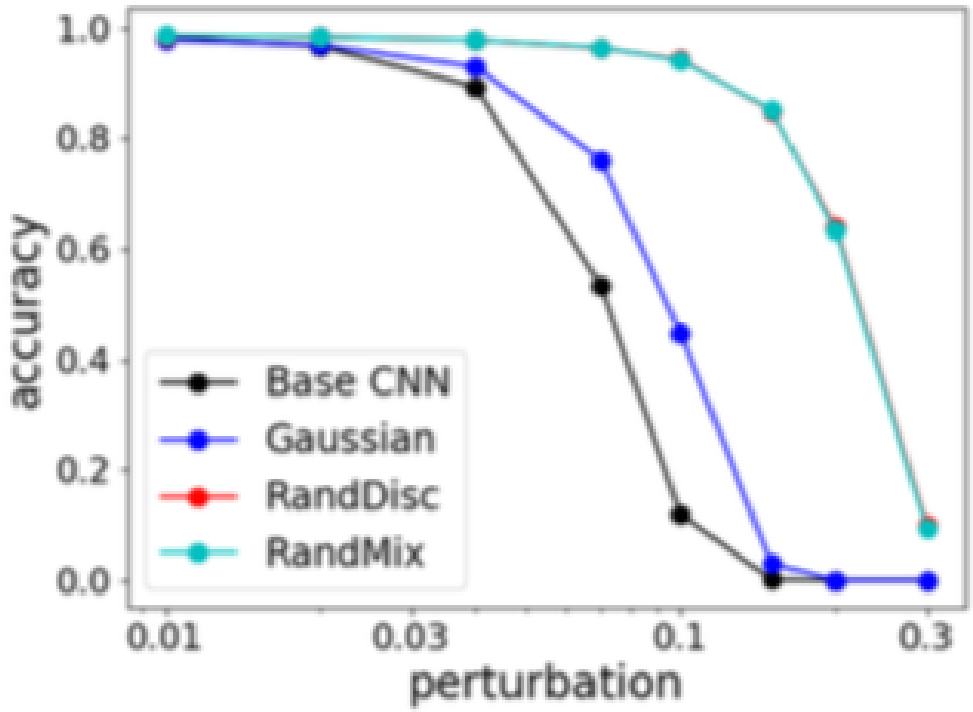
Experiments - attacks

We evaluate the defenses against whitebox PGD attacks on the MNIST and the ImageNet datasets.

- If the defense is **non-differentiable**, then attack a differentiable approximation of it.
- If the defense is **stochastic**, then in each round average multiple independent copies of the gradients to perform the PGD.

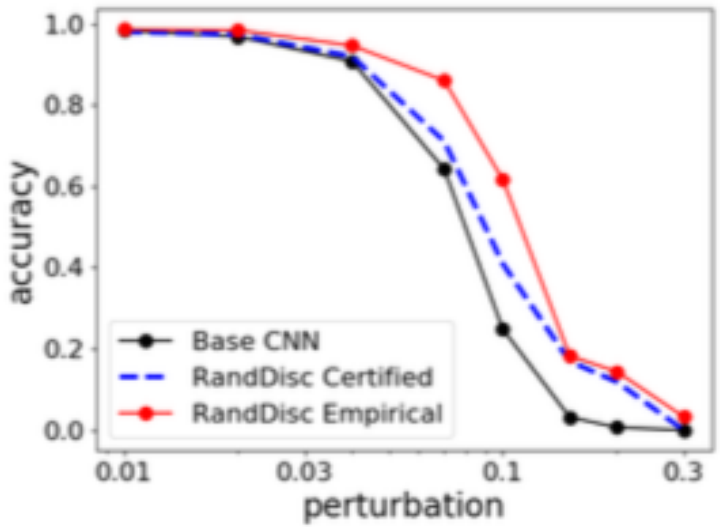
Experiments - MNIST

Base classifier : a normally-trained CNN [1] that reaches 99.2% accuracy on the test set.

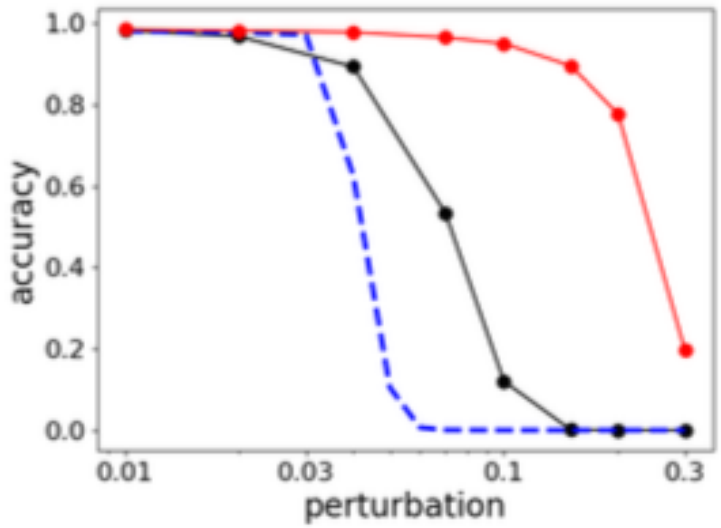


For $\epsilon = 0.1$, RandDisc achieves 94.4% accuracy, which exceeds the previous methods reach 93.8% at most.

Experiments - MNIST



(a) with downsampling



(b) without downsampling

Figure 3: Certified and empirical accuracy of RandDisc under the whitebox attack on MNIST.

Experiments - ImageNet

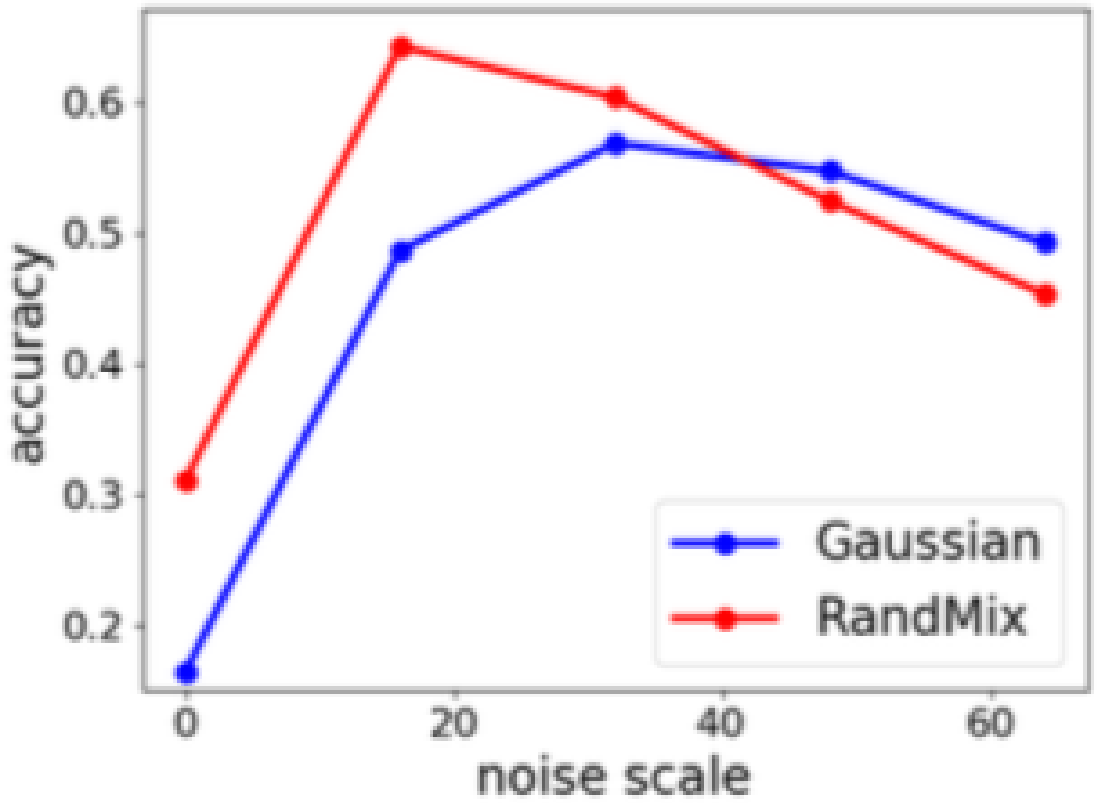
Base classifier : a pretrained InceptionResNet-V2 model. Remove images that were frequently misclassified by state-of-the-art neural models even without perturbation. The base classifier achieves 99.3% accuracy on the resulting dataset.

	$\epsilon = 1$	$\epsilon = 2$	$\epsilon = 4$
Base Classifier	16.0%	6.0%	2.3%
BitDepth	20.4%	8.7%	3.4%
JPEG	22.4%	9.9%	4.5%
TVM	25.5%	10.2%	4.2%
ResizePadding	30.7%	9.8%	2.3%
Gaussian	57.3%	29.9%	11.4%
RandDisc	56.0%	46.0%	**29.4%**
RandMix	**60.7%**	**48.3%**	28.3%

(a) Against whitebox PGD attack

Experiments - ImageNet

To understand the effect of noise, we vary the noise parameter σ from 0 to 64.



Experiments - ImageNet

If we take an adversarially trained model as the base classifier, then our defense can be even stronger. For example : InceptionResNet-V2 trained against FGSM attack.

Model	$\epsilon = 1$	$= 2$	$= 4$
Adv	18.4%	10.7%	5.8%
RandMix	60.7%	48.3%	28.3%
Adv + RandMix	62.9%	54.2%	39.5%

This observation is the opposite of that on MNIST.
This might be due to the fact that the model is already quite robust. (Adv model achieve 92.7% accuracy under the PGD attack for $\epsilon = 0.3$.)

References

1. Towards deep learning models resistant to adversarial attacks.
2. Provable defenses against adversarial examples via the convex outer adversarial polytope. (ICML 2018)
3. Certified defenses against adversarial examples. (ICLR 2018)