Evading Defenses to Transferable Adversarial Examples by Translation-Invariant Attacks

Yinpeng Dong, Tianyu Pang, Hang Su, Jun Zhu

Department of Computer Science and Technology, Institute for AI, Tsinghua University

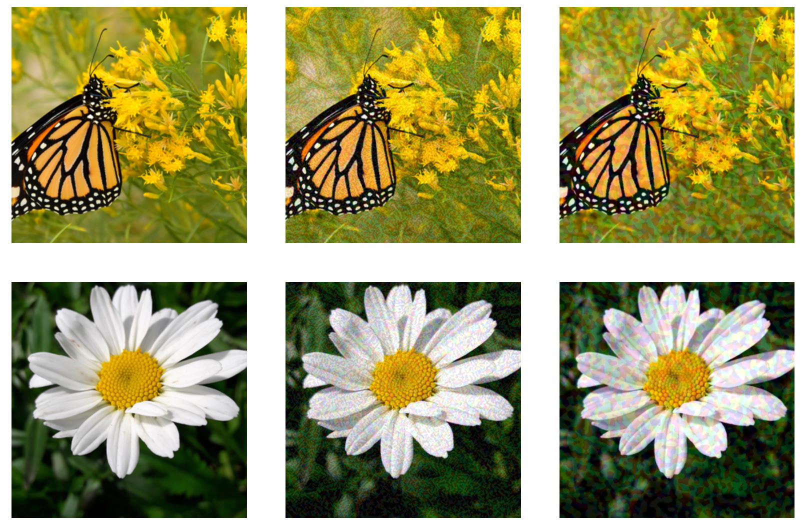
CVPR LONG BEACH CALIFORNIA June 16-20, 2019

Introduction

❑ **Adversarial examples** are crafted by adding small, human-imperceptible noises to normal examples, but make a model output wrong predictions.

❑ **Constrained Optimization Problem:**

$$\max_{x^{adv}} J(x^{adv}, y) \ s.t. \|x^{adv} - x^{real}\|_\infty \le \epsilon$$



Raw Image FGSM TI-FGSM

1. Fast Gradient Sign Method (FGSM) [Goodfellow et al., 2015]:

$$x^{adv} = x^{real} + \epsilon \cdot \text{sign}(\nabla_x J(x^{real}, y))$$

2. Basic Iterative Method (BIM) [Kurakin et al., 2016]:

$$x_{t+1}^{adv} = x_t^{adv} + \alpha \cdot \text{sign}\left(\nabla_x J(x_t^{adv}, y)\right)$$

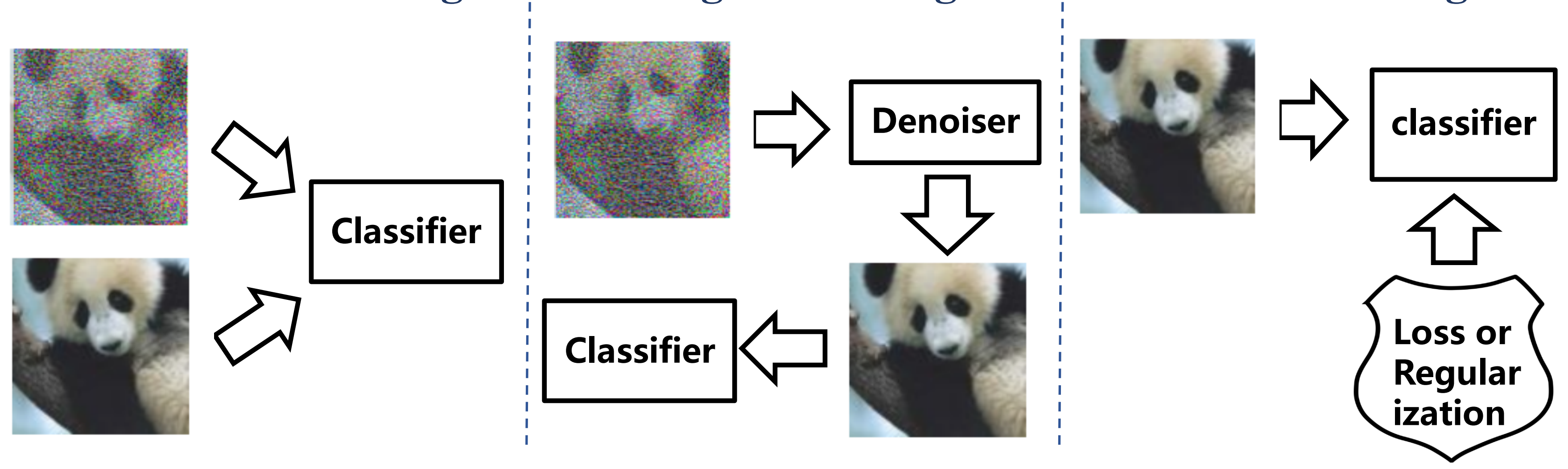
3. Momentum Iterative Fast Gradient Sign Method (MI-FGSM) [Dong et al., 2018]

$$g_{t+1} = \mu \cdot g_t + \frac{\nabla_x J(x_t^{adv}, y)}{\|\nabla_x J(x_t^{adv}, y)\|_1}, \quad x_{t+1}^{adv} = x_t^{adv} + \alpha \cdot \text{sign}(g_{t+1})$$

4. Carlini & Wagner's method (C&W) [Carlini and Wagner, 2017] optimizes the Lagrangian-relaxed form of the problem.

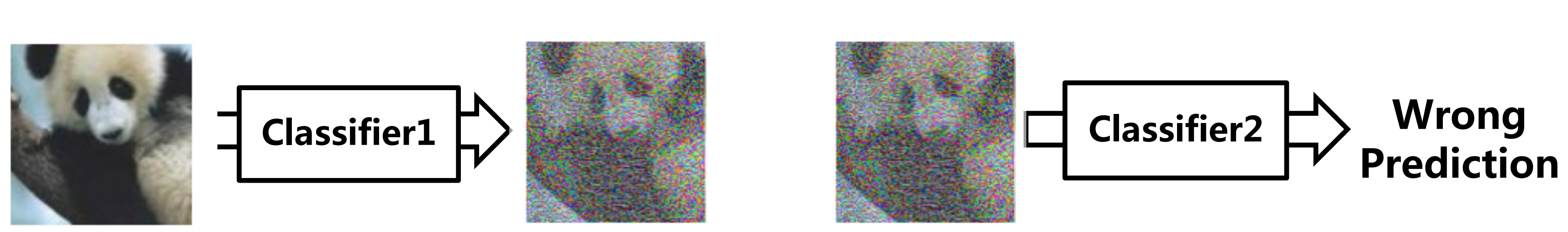
❑ **Defenses**

Adversarial Training | **Image Denoising** | **Robust Training**

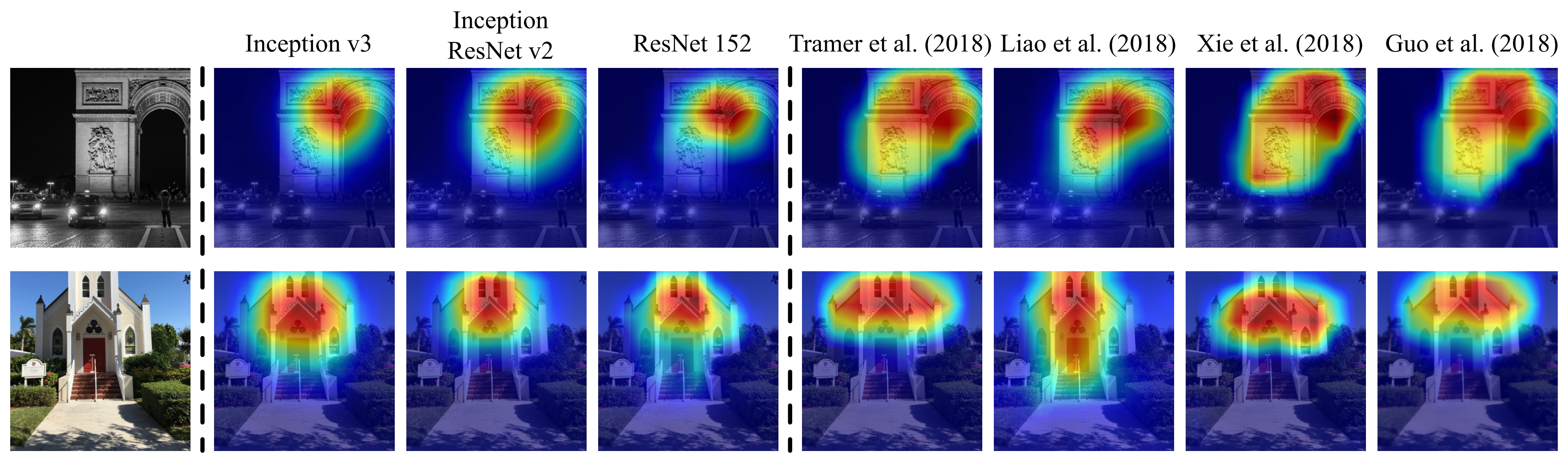


The defenses can be circumvented in the white-box manner since they cause obfuscated gradients [Athalye et al., 2018]; but some of them claim to be robust in the black-box manner.

We want to answer that: Are these defenses really robust against black-box attacks based on the transferability?



Observation & Motivation



■ The defenses make predictions based on different discriminative regions compared with normal models (and also different gradient [Tsipras et al., 2019]);

■ The adversarial example is highly correlated with the discriminative region or gradient of the white-box model at the given input point, making it hard to transfer to defenses which are based on different regions for predictions;

■ Therefore, we propose to craft an adversarial example against an ensemble of translated images.

Methodology

❑ **Translation-invariant objective function**

$$\max_{x^{adv}} \sum_{i,j} w_{ij} J(T_{ij}(x^{adv}), y) \quad s.t. \|x^{adv} - x^{real}\|_\infty \le \epsilon$$

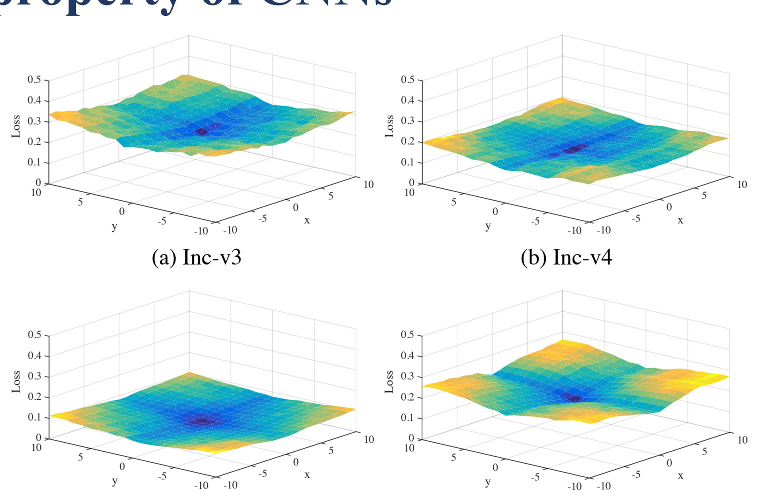
- T_{ij} is the translation operation, i.e., $T_{ij}(x)_{a,b} = x_{a-i,b-j}$.

❑ **Assumption – translation-invariant property of CNNs**

$$\nabla_x J(x, y)\Big|_{x=T_{ij}(\hat{x})} \approx \nabla_x J(x, y)\Big|_{x=\hat{x}}$$

❑ **Loss gradient**

$$\nabla_x \left(\sum_{i,j} w_{ij} J(T_{ij}(x), y)\right)\Big|_{x=\hat{x}} \approx W * \nabla_x J(x, y)\Big|_{x=\hat{x}}$$



(a) Inc-v3 (b) Inc-v4 (c) Inc-Res-v2 (d) Res-v2-152

❑ **Kernel matrix**

- **A uniform kernel** $W_{i,j} = \frac{1}{(2k+1)^2}$;
- **A linear kernel** $\tilde{W}_{i,j} = \left(1 - \frac{|i|}{k+1}\right)\left(1 - \frac{|j|}{k+1}\right), W_{i,j} = \frac{\tilde{W}_{i,j}}{\sum \tilde{W}_{i,j}}$
- **A Gaussian kernel** $\tilde{W}_{i,j} = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{i^2+j^2}{2\sigma^2}\right), W_{i,j} = \frac{\tilde{W}_{i,j}}{\sum \tilde{W}_{i,j}}$

❑ **Our method can be integrated into any gradient-based attack method**

- **TI-FGSM:** $x^{adv} = x^{real} + \epsilon \cdot \text{sign}\left(W * \nabla_x J(x^{real}, y)\right)$
- **TI-BIM:** $x_{t+1}^{adv} = x_t^{adv} + \alpha \cdot \text{sign}\left(W * \nabla_x J(x_t^{adv}, y)\right)$

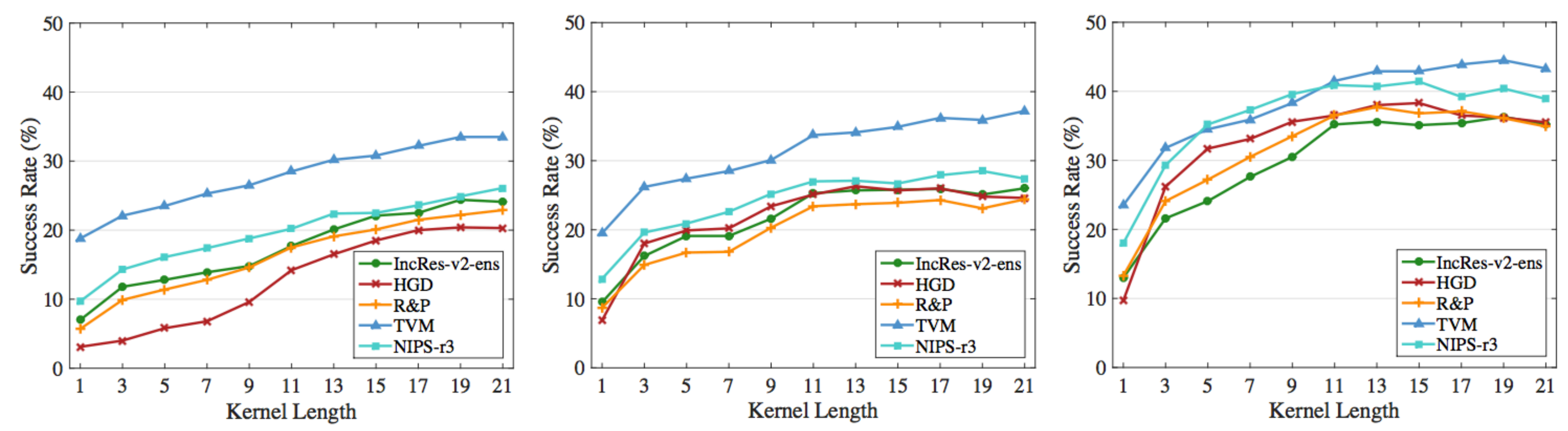
Experiments

❑ **Experimental settings**

- **Inc-v3$_{ens3}$, Inc-v3$_{ens4}$, IncRes-v2$_{ens}$ [Tramer et al., 2018];**
- **High-level representation guided denoiser (HGD) [Liao et al., 2018];**
- **Random resizing and padding (R&P) [Xie et al., 2018];**
- **JPEG compression and total variance minimization (TVM) [Guo et al., 2018];**
- **NIPS-r3 (rank-3 submission in the NIPS 2017 defense competition).**

White-box models: Inc-v3, Inc-v4, IncRes-v2, Res-v2-152

❑ **Results of kernel length**



(a) TI-FGSM (b) TI-MI-FGSM (c) TI-DIM

Kernel Length=1 Kernel Length=3 Kernel Length=5 Kernel Length=7 Kernel Length=9 Kernel Length=11 Kernel Length=13 Kernel Length=15

❑ **Attacking an ensemble of models**

Attack	Inc-v3$_{ens3}$	Inc-v3$_{ens4}$	IncRes-v2$_{ens}$	HGD	R&P	JPEG	TVM	NIPS-r3
FGSM	27.5	23.7	13.4	4.9	13.8	38.1	30.0	19.8
TI-FGSM	**39.1**	**38.8**	**31.6**	**29.9**	**31.2**	**43.3**	**39.8**	**33.9**
MI-FGSM	50.5	48.3	32.8	38.6	32.8	67.7	50.1	43.9
TI-MI-FGSM	**76.4**	**74.4**	**69.6**	**73.3**	**68.3**	**77.2**	**72.1**	**71.4**
DIM	66.0	63.3	45.9	57.7	51.7	82.5	64.1	63.7
TI-DIM	**84.8**	**82.7**	**78.0**	**82.6**	**81.4**	**83.4**	**79.8**	**83.1**

Conclusion

❑ **We propose a translation-invariant attack method to craft adversarial examples with improved transferability against the defense models.**

❑ **Our method can be integrated into any gradient-based attack method.**

❑ **Our best attack TI-DIM fools eight state-of-the-art defenses at an 82% success rate on average.**

❑ **Our method can serve as a benchmark to evaluate robustness of future developed defenses.**