Stochastic Substitute Training: A Gray-box Approach to Craft Adversarial Examples Against Gradient Obfuscation Defenses

Mohammad Hashemi, Greg Cusack,* Eric Keller

10/19/18

*Presenter

Overview

- Neural Networks are everywhere
- Vulnerable to attacks!
- Response: Design robust neural networks that block or obfuscate gradients
- Even these are vulnerable in a white box setting! [Athalye et al. ICML '18]
- But are they vulnerable in a more realistic setting?

Stochastic Substitute Training

A general, gray-box attack for breaking defenses that obfuscate gradients

Crafting Adversarial Examples

Optimization Problem

$$argmin_{\delta} ||\delta||_p \; s.t. \; (x + \delta) \in [0, 1]^m \text{ and } F(x + \delta) = y_{target}$$

But neural networks are not convex!

Carlini & Wagner (C&W) Attack

Modified Objective Function

minimize $c \,.\, \|\delta\|_p + f(x+\delta)$ **s.t.** $x + \delta \in [0,1]^m$

$$\boxed{f(x') = max(max_{i \neq t}(Z(x')_i) - Z(x')_t, -\kappa)}$$

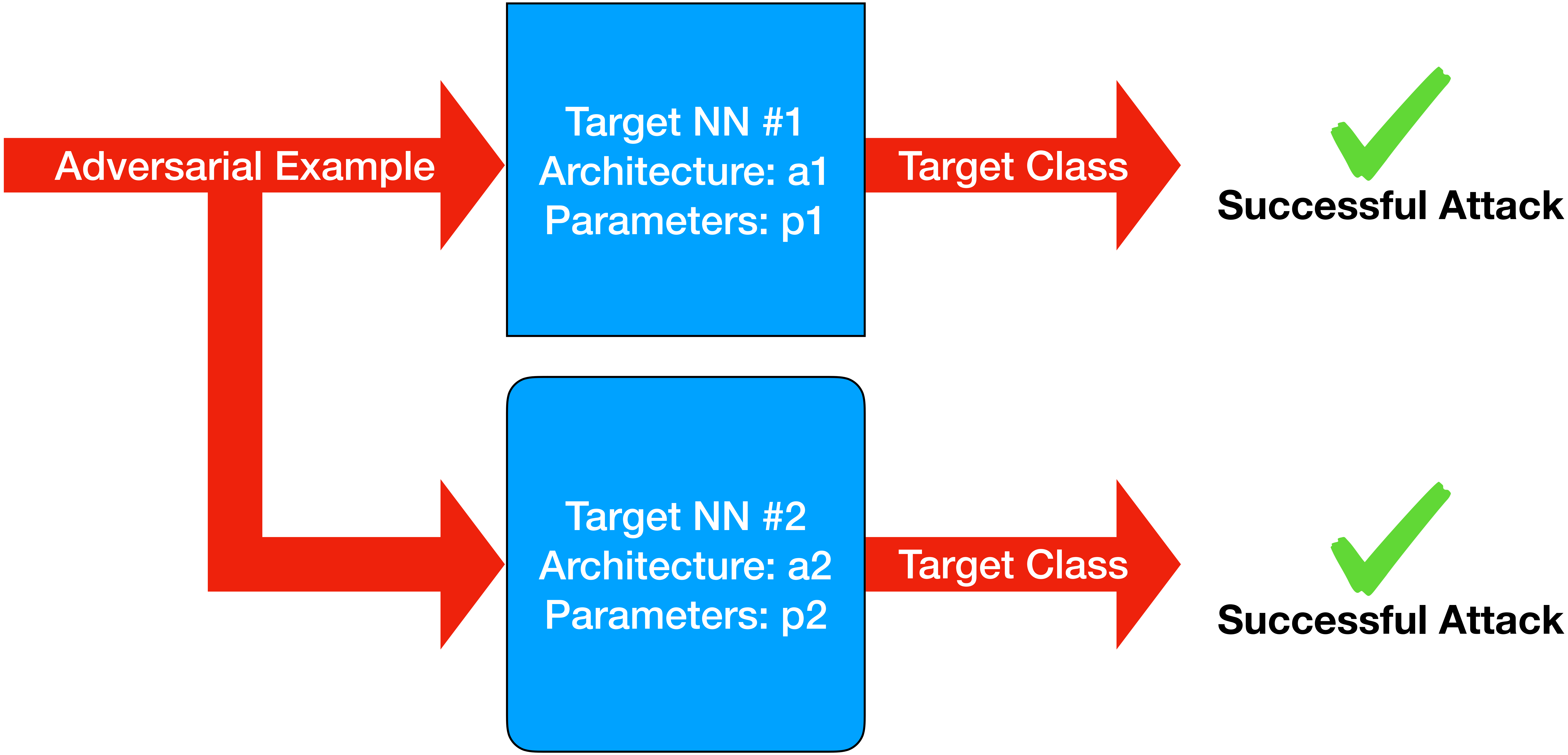
$Z(x')_i$ —> Logits of non-target classes

$Z(x')_t$ —> Logit of target class

k —> Determines classification confidence

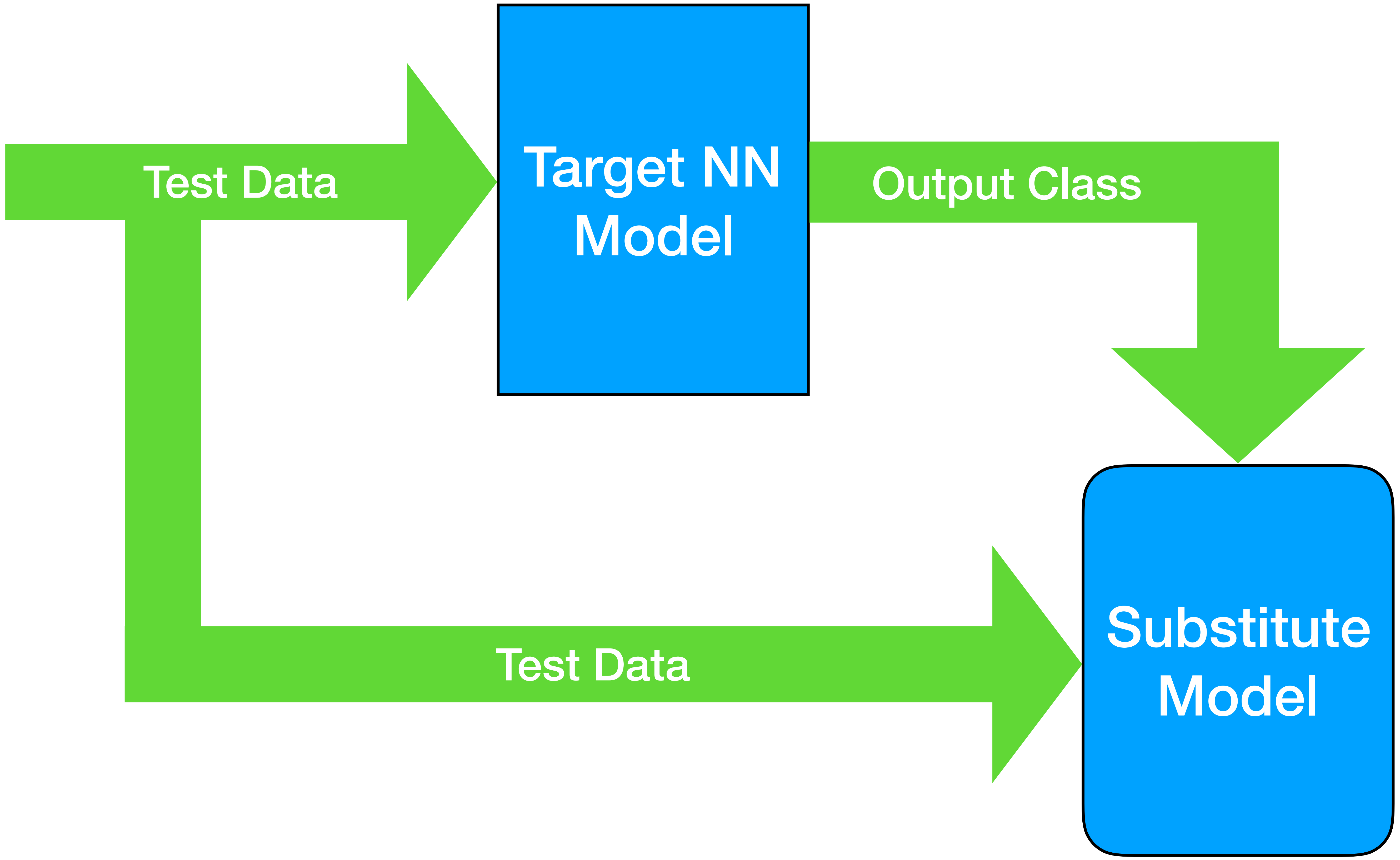
[Carlini et al. IEEE S&P '17]

Black Box Attack (Transferability)



Black Box Attack (Substitute Training)

Leverage Transferability of Adversarial Examples

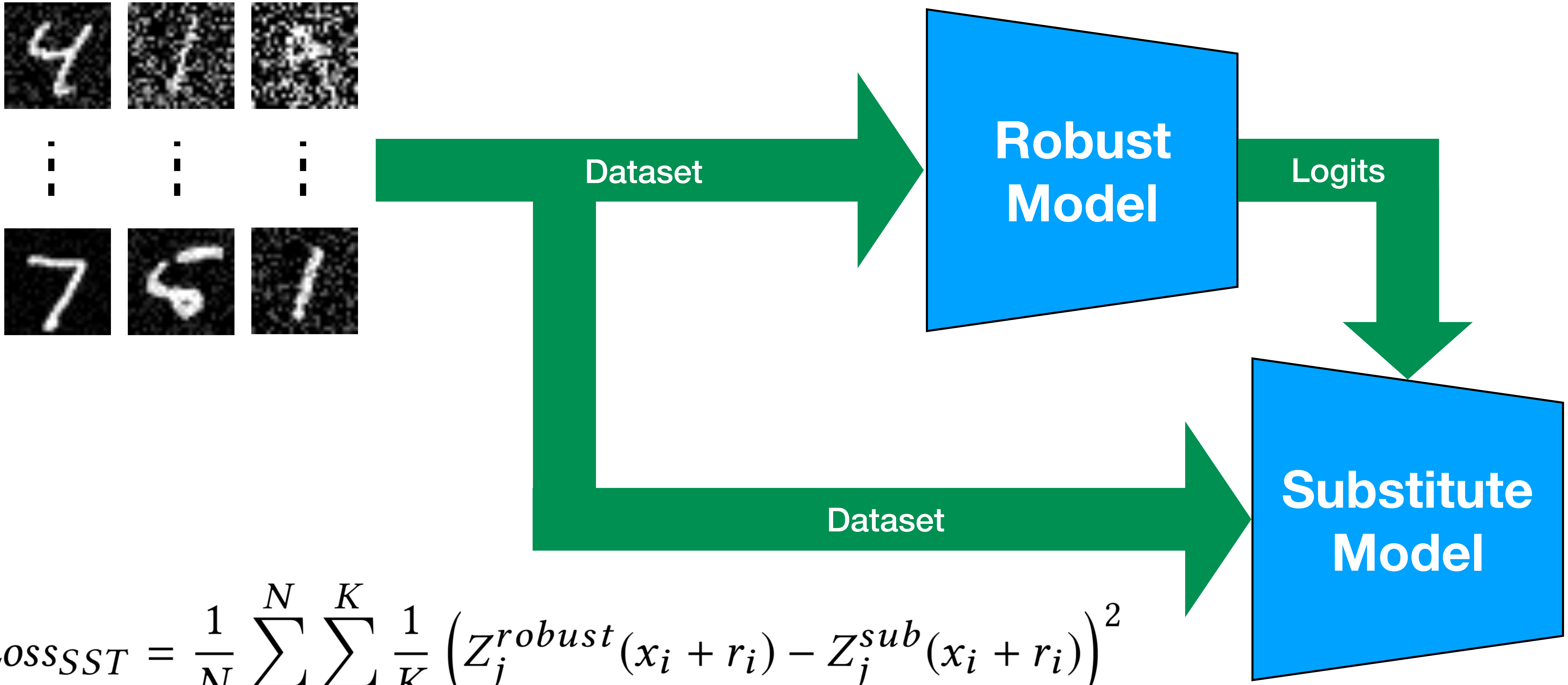


[Biggio et al. ECML/PKDD ’13], [Papernot et al. Asia CCS ’17]

Stochastic Substitute Training (Threat Model)

- Fortifying Defenses
  - Classifier predicts adversarial examples as their correct class
  - Threat Model
    - Send inputs and see logits
- Detecting Defenses
  - Identify when adversarial examples are fed into the classifier
  - Threat Model
    - Send inputs, see logits, and the output of the detector

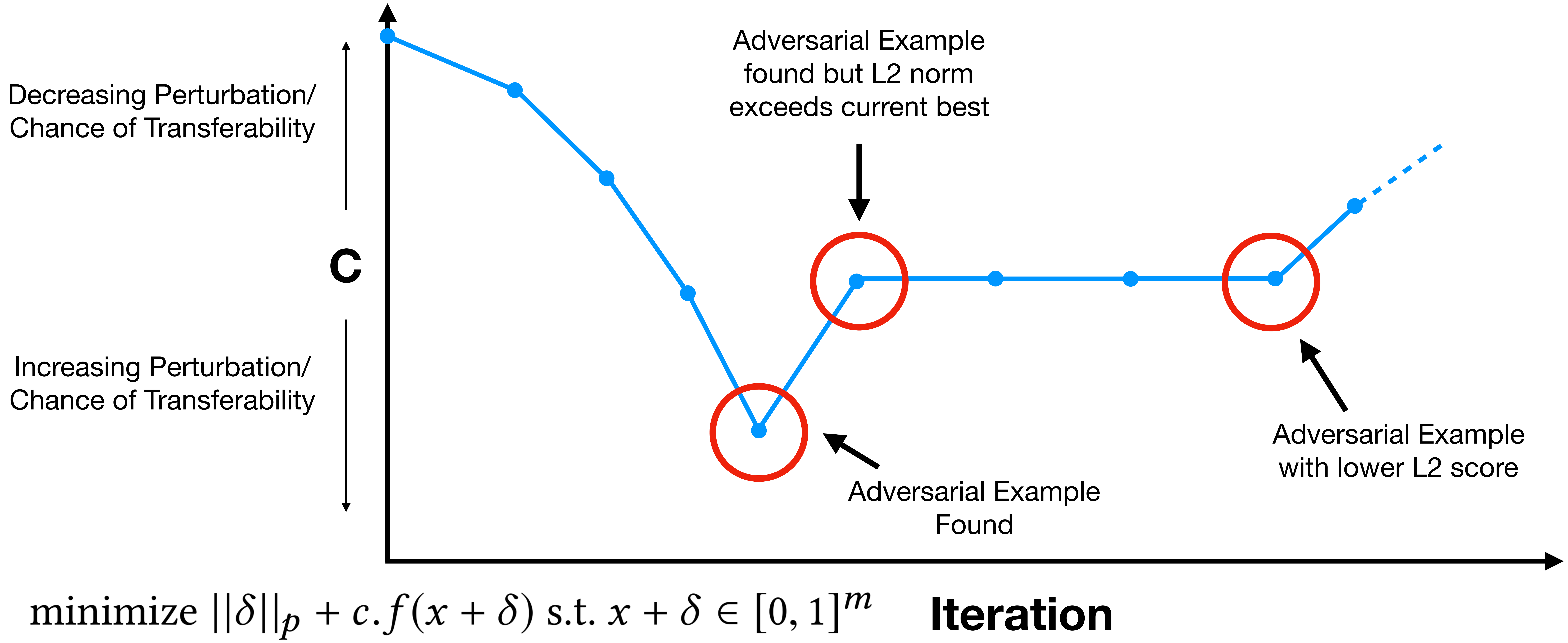
Stochastic Substitute Training



$$Loss_{SST} = \frac{1}{N}\sum_{i=0}^{N}\sum_{j=0}^{K}\frac{1}{K}\left(Z_j^{robust}(x_i + r_i) - Z_j^{sub}(x_i + r_i)\right)^2$$

Crafting Adversarial Examples

For each κ...



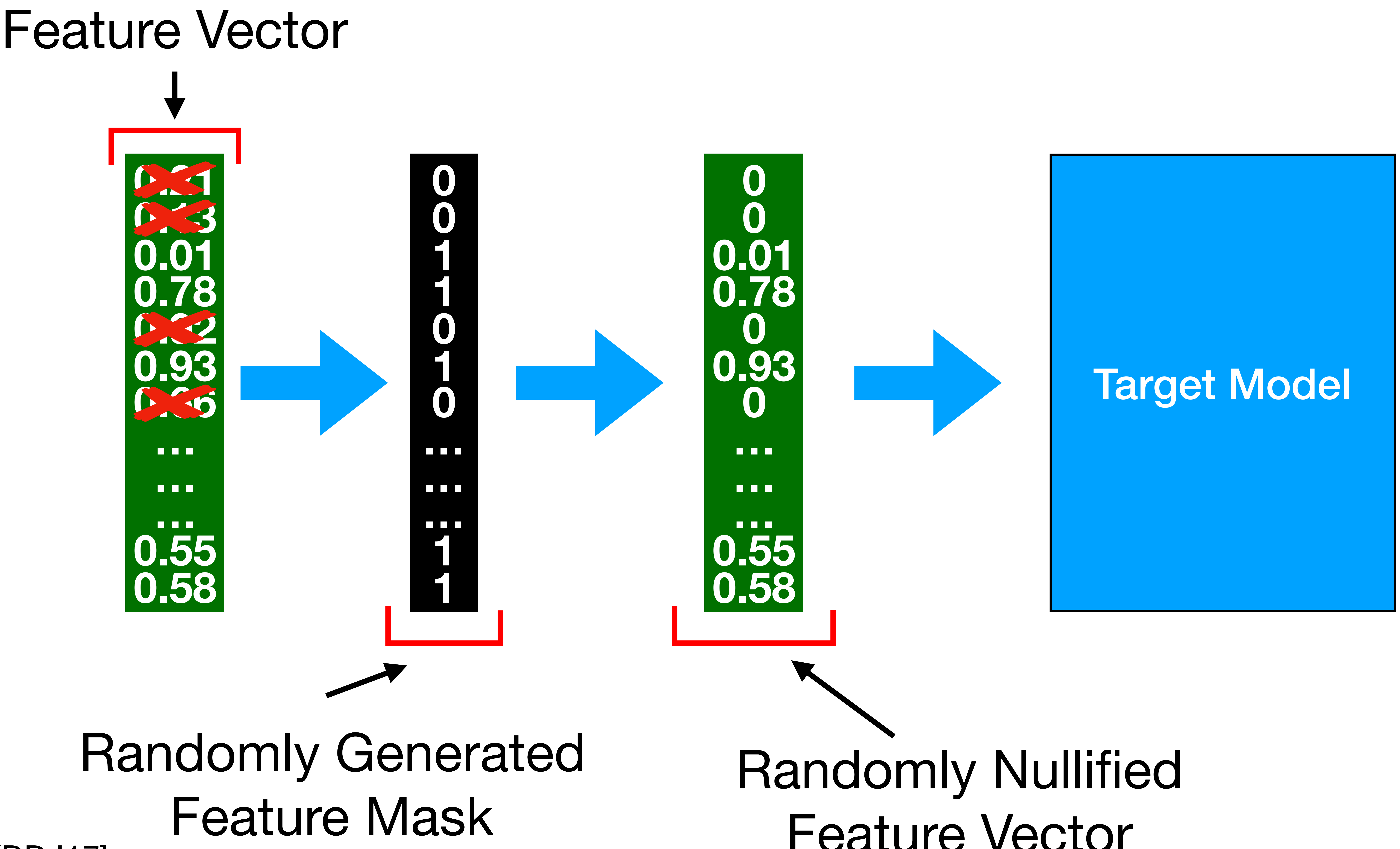
$$\text{minimize } ||\delta||_p + c.f(x+\delta) \text{ s.t. } x+\delta \in [0,1]^m$$

$$f(x') = max(max_{i \neq t}(Z(x')_i) - Z(x')_t, -\kappa)$$

Noisy Data Augmentation

- Substitute model more closely approximates decision boundaries of target model
- Helps substitute model learn how the robust model's class probabilities change in the neighborhood of each sample
- Multiple copies of training models created with varying levels of random noise
    - Each substitute model approximates the decision boundaries for some specific images better than others

Random Feature Nullification



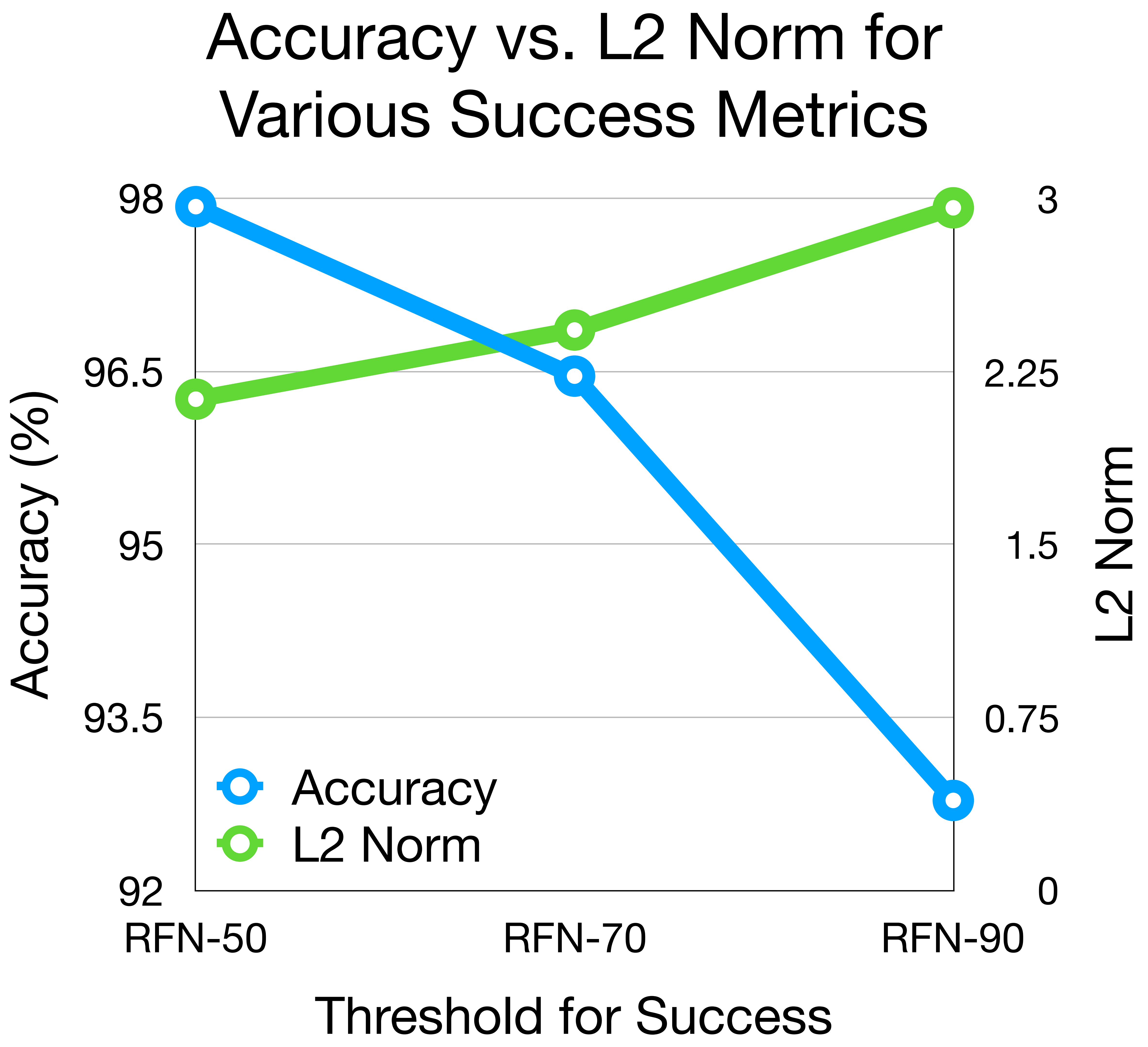
[Wang et al. SIGKDD ’17]

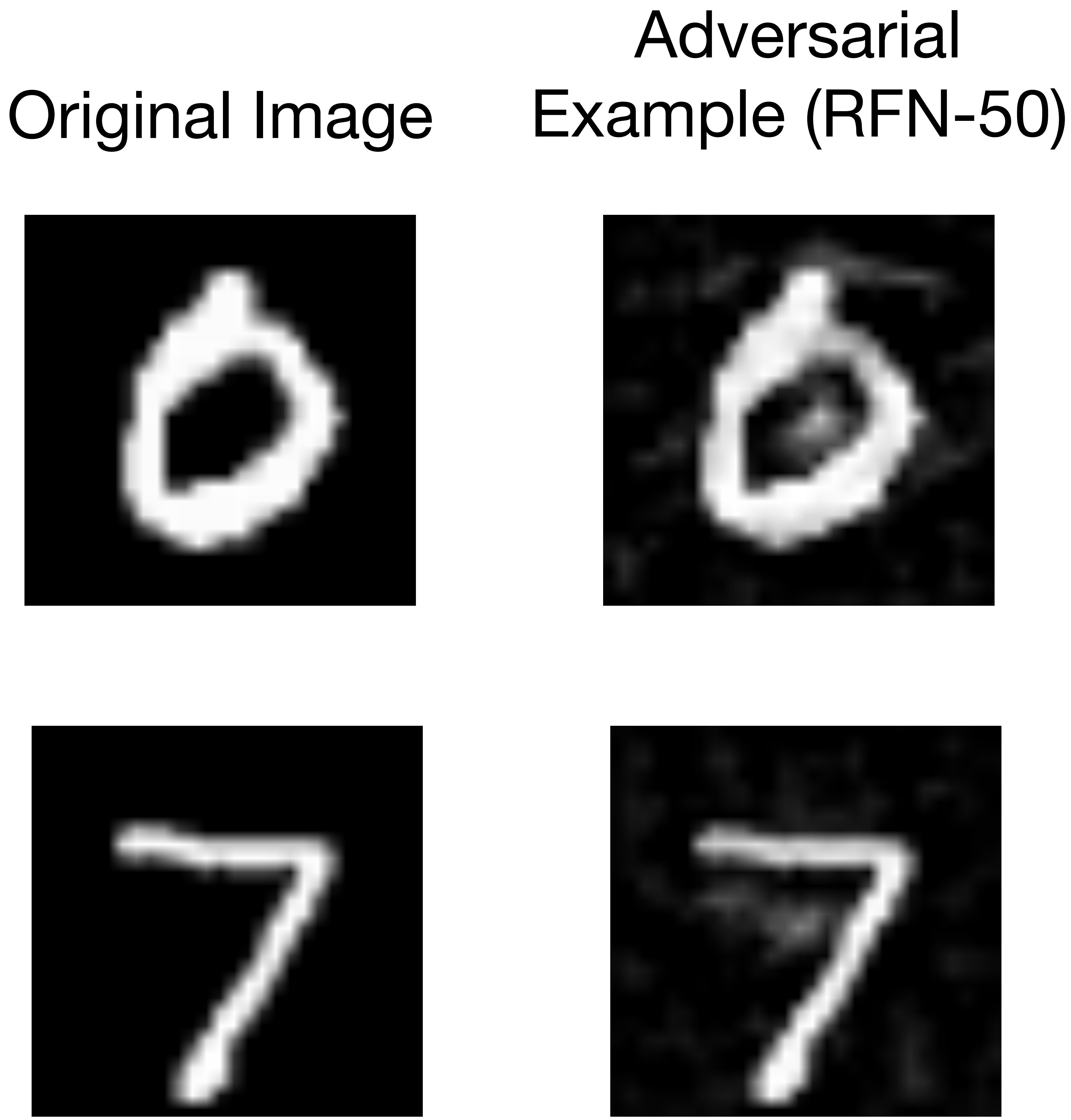
Random Feature Nullification Attack

- Trained target model
- Nullified 50% of features
- Trained substitute model on multiple replications of MNIST test set
- Augmented each set with various levels of random noise
- Define three success metrics
  - RFN-50, RFN-70, RFN-90

[Wang et al. SIGKDD '17]

Random Feature Nullification Attack





Thermometer Encoding

Idea: Discretize features
to mask gradients

$$\tau(p)_j = \begin{cases} 1, & \text{if } p \geq j/k \\ 0, & \text{otherwise} \end{cases}$$

p: pixel value
k: encoded vector size
j: index of resulting vector

[Buckman et al. ICLR '18]

Image Pixel Values

0.12	0.57	…	0.91
0.87	0.64	…	0.32
…	..	…	0.69
0.56	0.10	0.33	0.82

for k = 10:

	1	1	1	1	0	0	0	0	0	0
j:	0	1	2		⋯		⋯		⋯	k

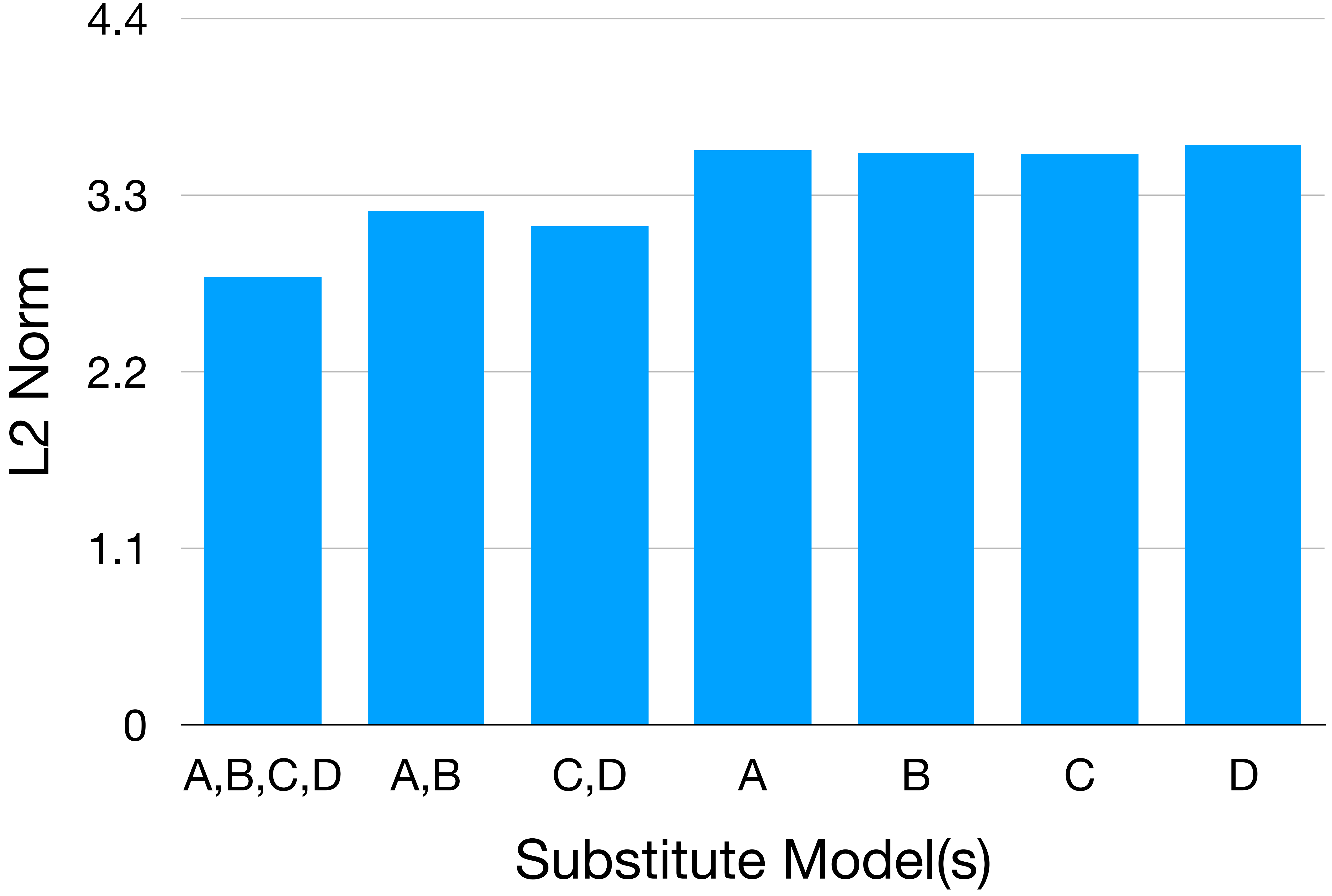
Thermometer Encoding Attack

- Used pre-trained model fortified with thermometer encoding and adversarial training as target model
- Trained four identical substitute models on CIFAR-10 test set with different levels of random noise
- Crafted adversarial examples for the first 100 CIFAR-10 images in the test set

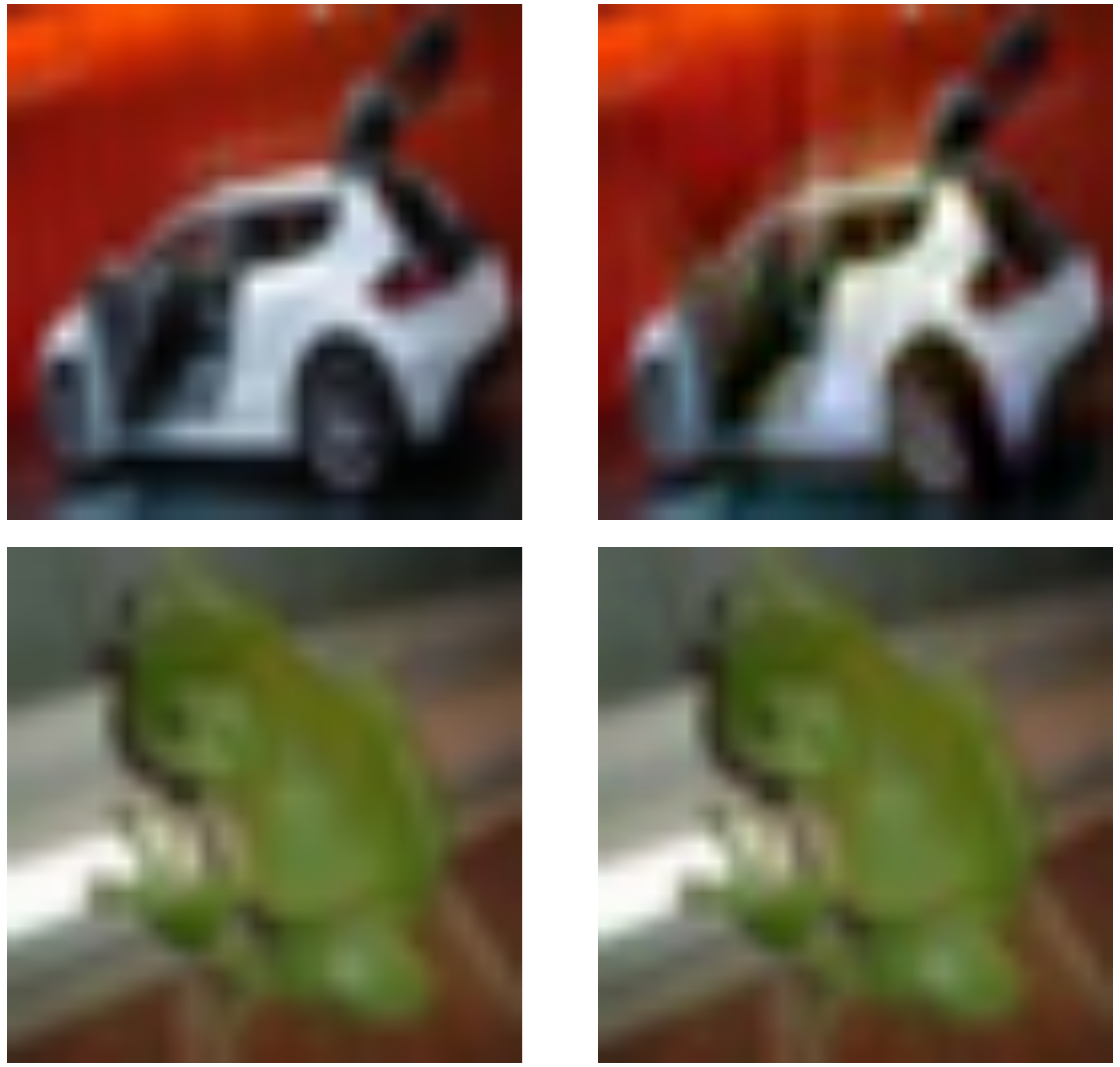
[Buckman et al. ICLR ’18], [Athalye et al. ICML ’18]

Thermometer Encoding Attack

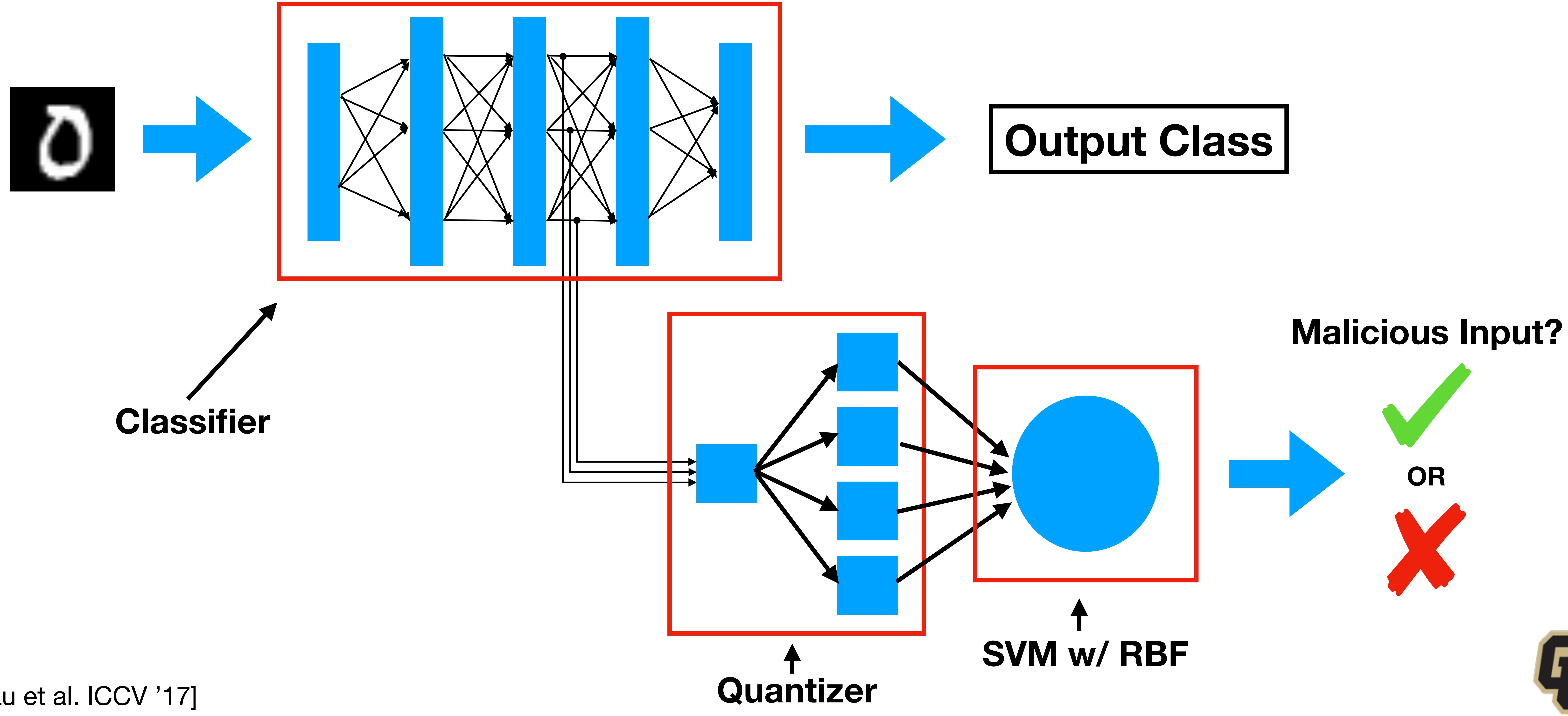
L2 Norm of Substitute Models
Success Rate: 100%



Original Image | Adversarial Example



SafetyNet



[Lu et al. ICCV ’17]

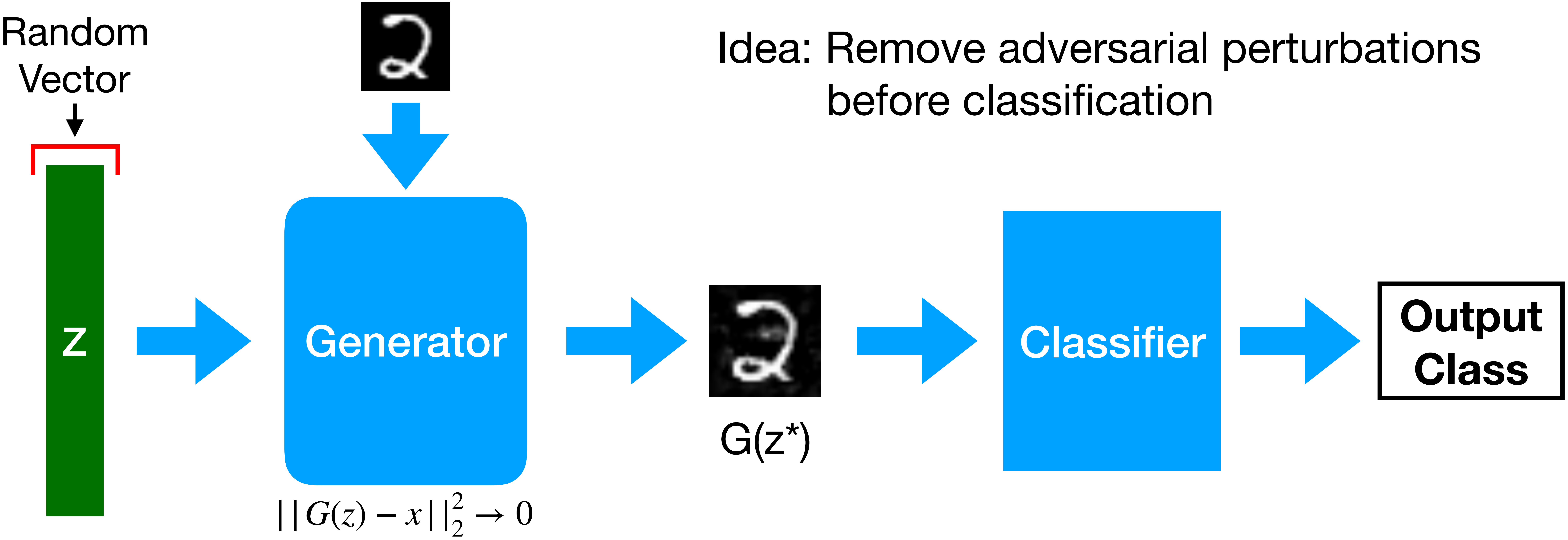
SafetyNet Attack

- Trained
  - Two substitute models for the original classifier
  - Substitute model for the detector
- Metrics for Success
  - Does the adversarial example fool the classifier?
  - Is the confidence ratio less than 25%
  - Did the detector predict the adversarial example as a legitimate sample?

Original Image | Adversarial Example

Defense-GAN

Idea: Remove adversarial perturbations before classification



For threshold θ:

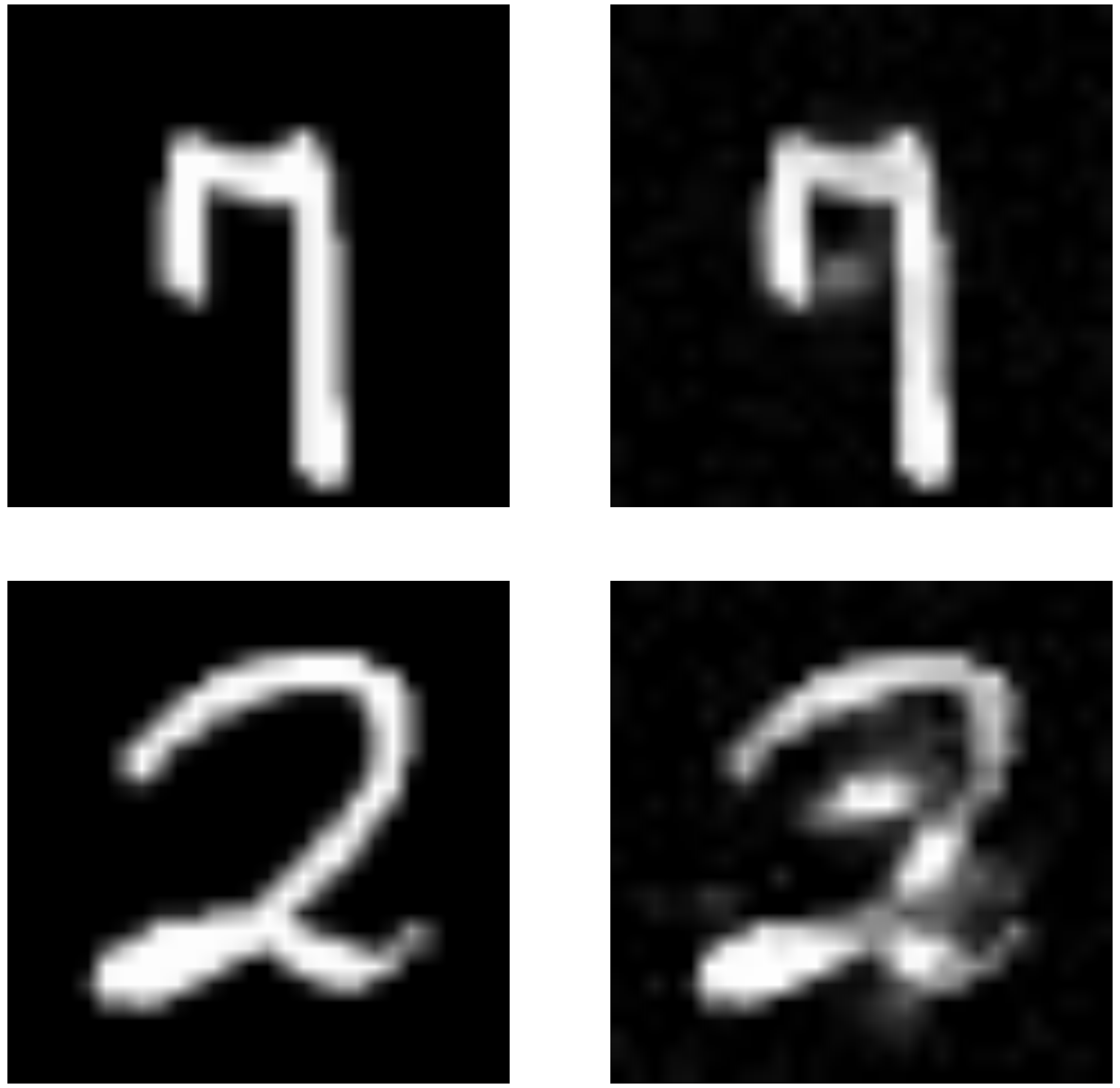
If $||G(z^*) - x||_2^2 > \theta$, then x is an adversarial example

[Samangouei et al. ICLR '18]

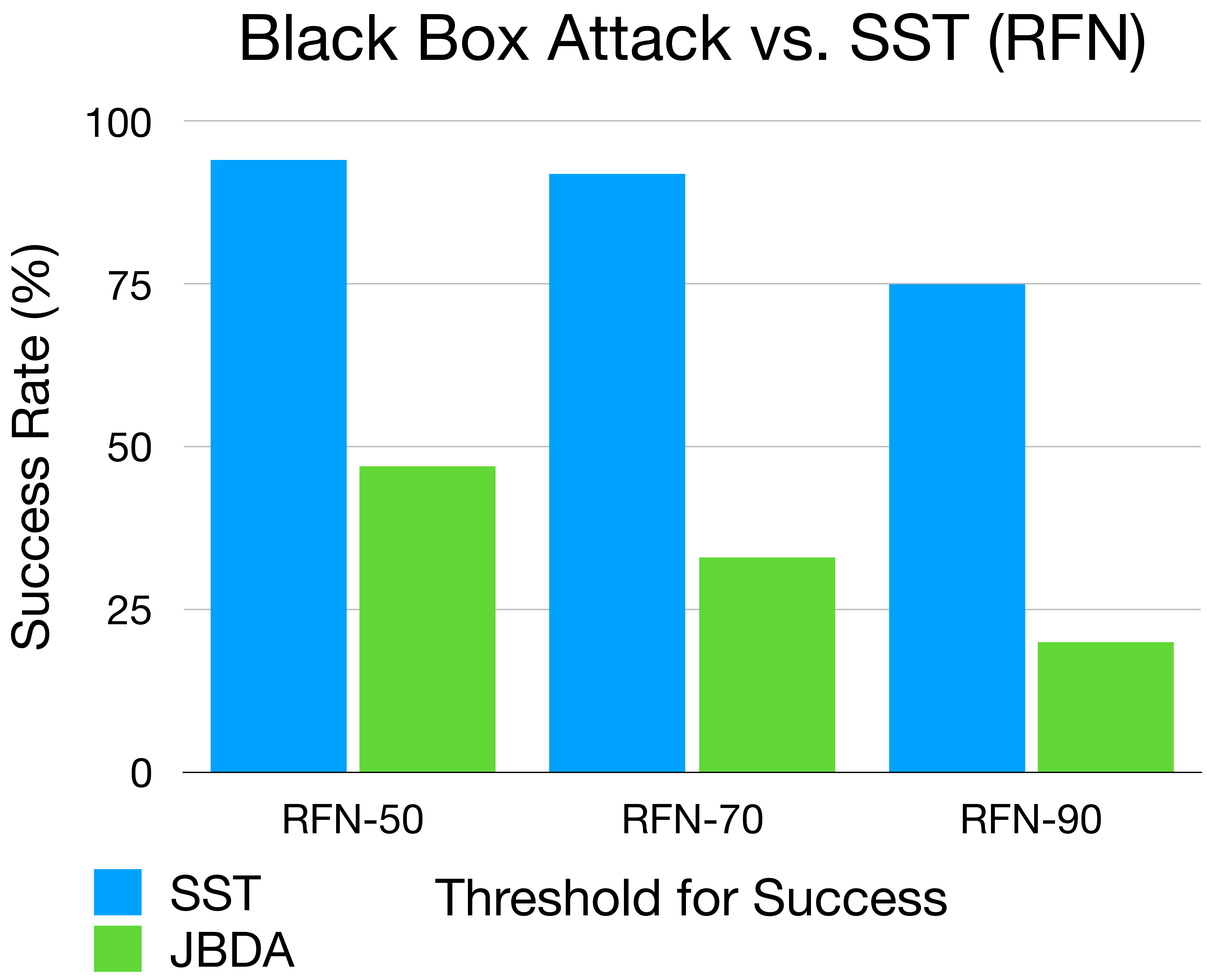
Defense-GAN Attack

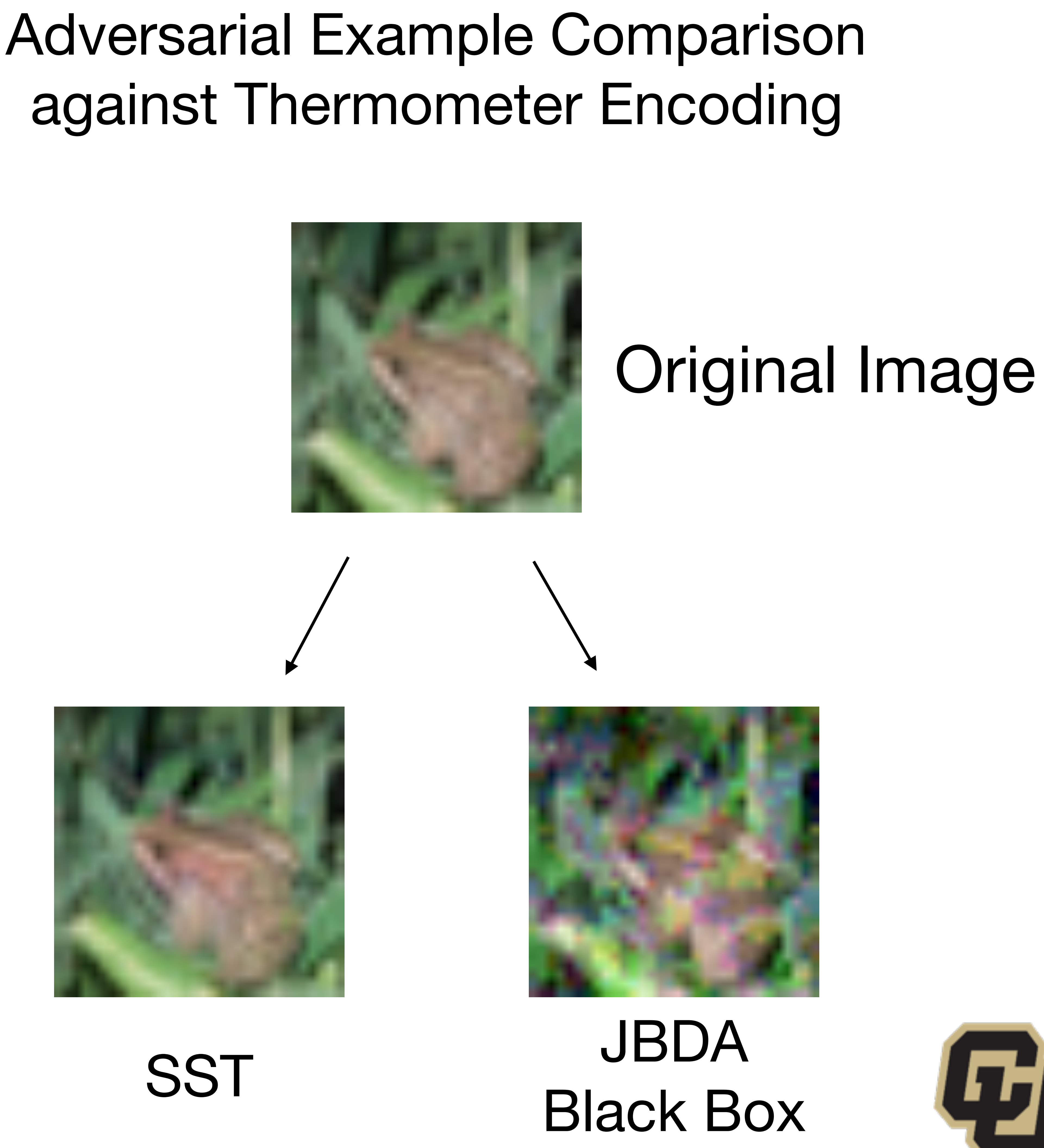
- Trained substitute model with random noise with noise in range [-0.95 - 0.95]
- Used cross entropy loss function for training
- 100% success in fooling classifier and detector
- More powerful than the first approach as this is a true black box attack

Original Image | Adversarial Example



Jacobian-Based Data Augmentation





[Papernot et al. Asia CCS ’17]

Conclusion

- Craft ways to attack deep neural network models that obfuscate gradients in attempt to protect themselves against adversarial examples
- Leveraged our approach against fortifying and detecting defenses
- We can design attacks with no knowledge of the type of defense, the defense and model parameters, and the training data.
- Black box attack evaluations

Questions?

Mohammad Hashemi
mohammad.hashemi@colorado.edu

Gregory Cusack
gregory.cusack@colorado.edu

Eric Keller
eric.keller@colorado.edu