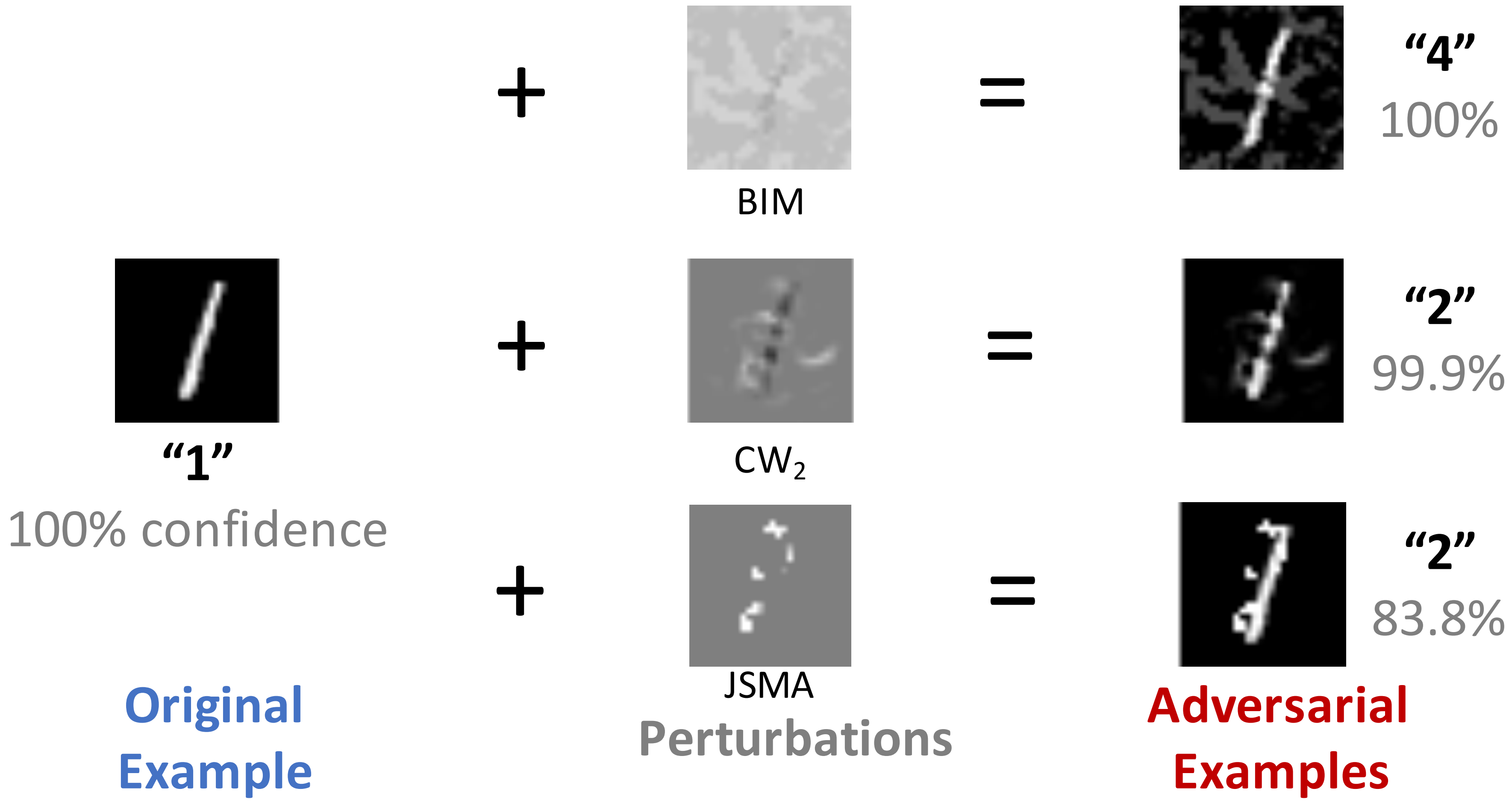
Feature Squeezing:

Detecting Adversarial Examples in Deep Neural Networks

Weilin Xu David Evans Yanjun Qi

University of Virginia

Background: Classifiers are Easily Fooled



C Szegedy et al., *Intriguing Properties of Deep Neural Networks*. In *ICLR 2014*.

Solution Strategy

Solution Strategy 1: Train a perfect vision model.

Infeasible yet.

Solution Strategy 2: Make it harder to find adversarial examples.

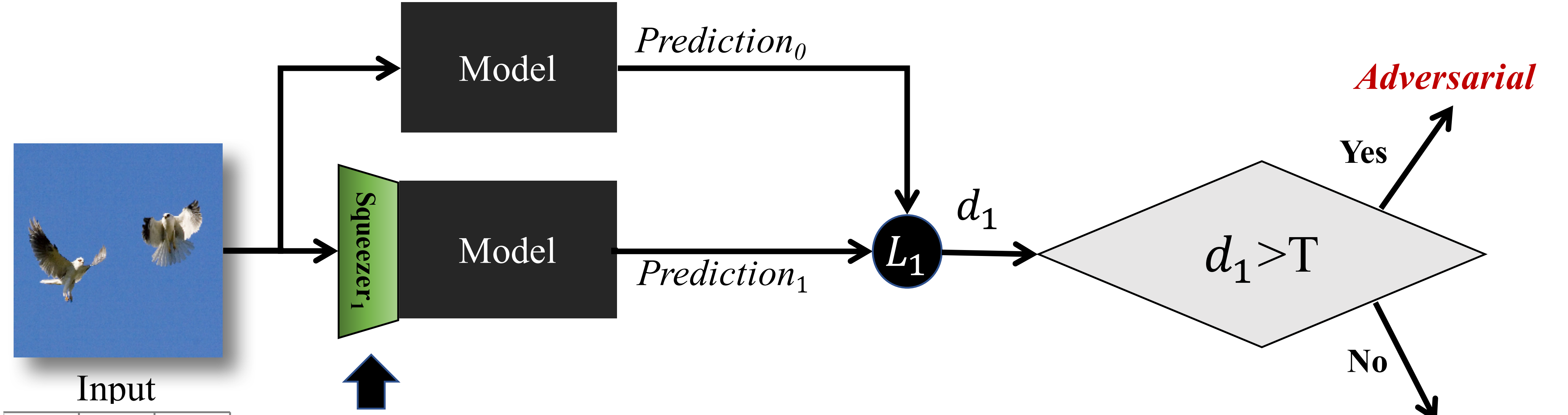
Arms race!

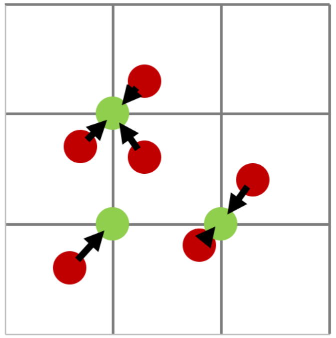
Feature Squeezing: A general framework that reduces the search space available for an adversary and detects adversarial examples.

Roadmap

- Feature Squeezing Detection Framework
- Feature Squeezers
  - Bit Depth Reduction
  - Spatial Smoothing
- Detection Evaluation
  - Oblivious adversary
  - Adaptive adversary

Detection Framework

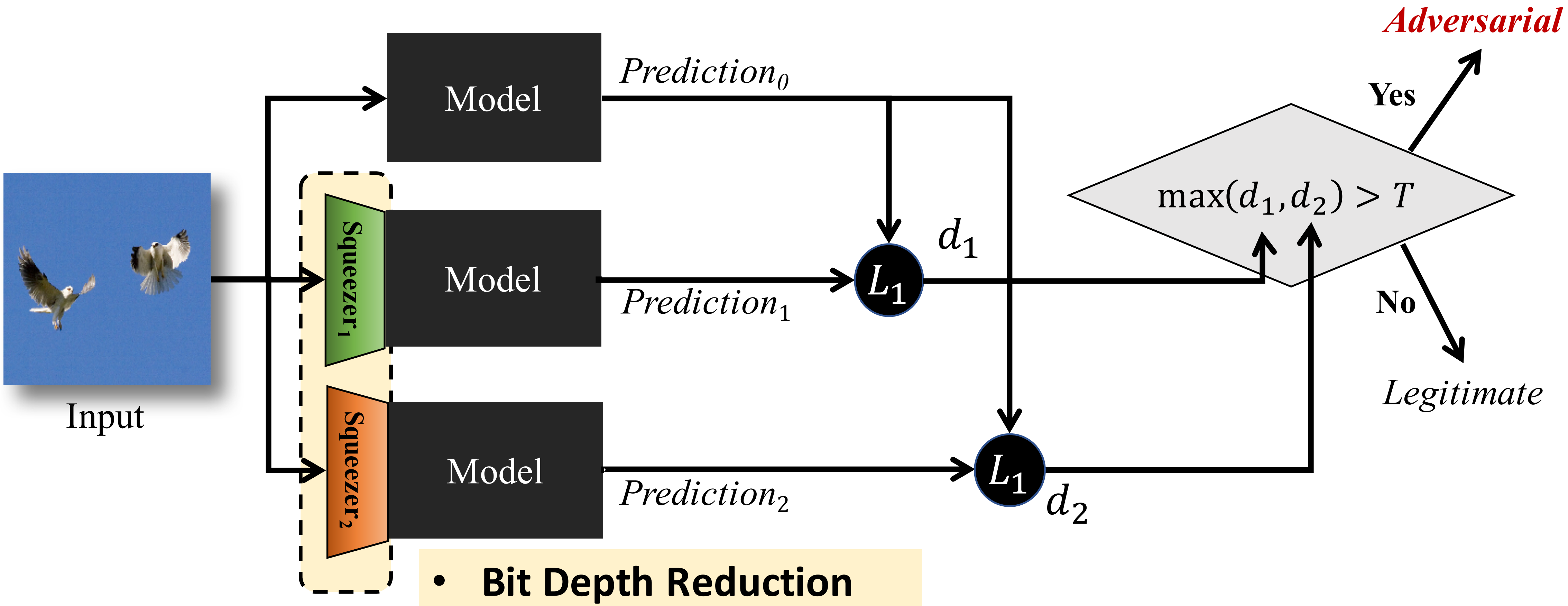




Feature Squeezer coalesces similar samples into a single one.

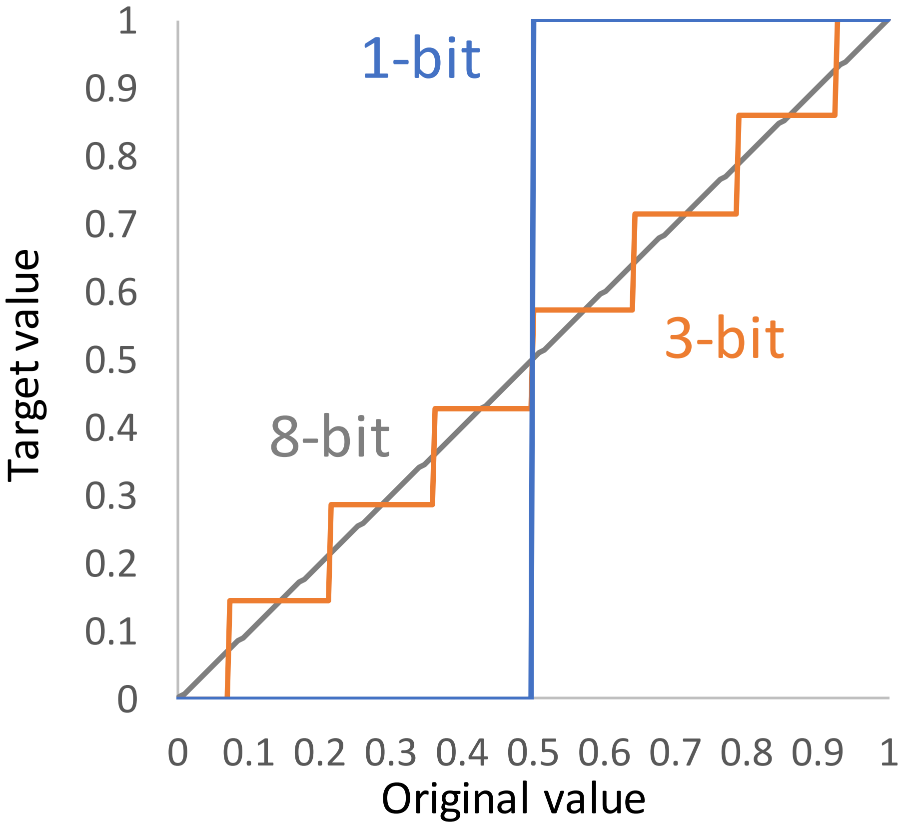
- Barely change legitimate input.
- Destruct adversarial perturbations.

Detection Framework: Multiple Squeezers

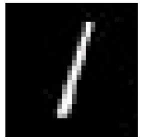


- **Bit Depth Reduction**
- **Spatial Smoothing**

Bit Depth Reduction



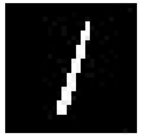
Signal Quantization



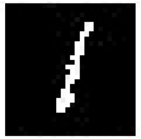
X

[0.012 0.571 …… 0.159 0.951]

Reduce to 1-bit
$f_i = \text{round}(f_i \times 2)/2$



[0. 1. …… 0. 1.]



[0. 0. …… 0. 1.]

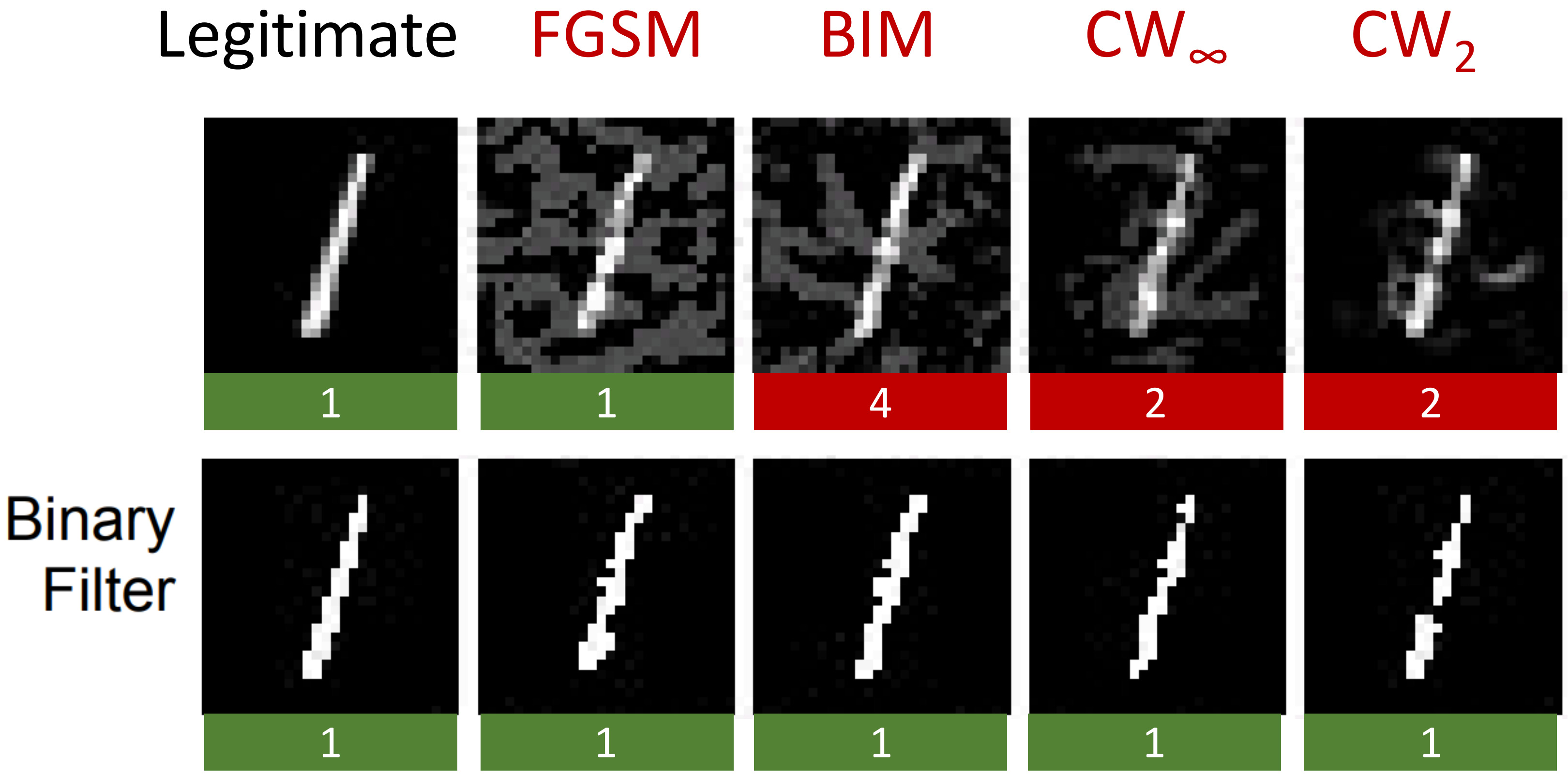
Reduce to 1-bit
$f_i = \text{round}(f_i \times 2)/2$

X_adv

[0.312 0.271 …… 0.159 0.351]

Bit Depth Reduction

Eliminating adversarial perturbations while preserving semantics.



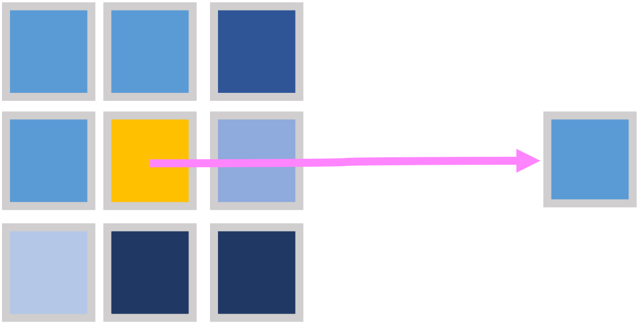
Accuracy with Bit Depth Reduction

Dataset	Squeezer	Adversarial Examples (FGSM, BIM, CW_∞, Deep Fool, CW_2, CW_0, JSMA)	Legitimate Images
MNIST	None	13.0%	99.43%
	1-bit Depth	62.7%	99.33%
ImageNet	None	2.78%	69.70%
	4-bit Depth	52.11%	68.00%

Baseline

Spatial Smoothing: Median Filter

- Replace a pixel with median of its neighbors.
- Effective in eliminating ”salt-and-pepper” noise.

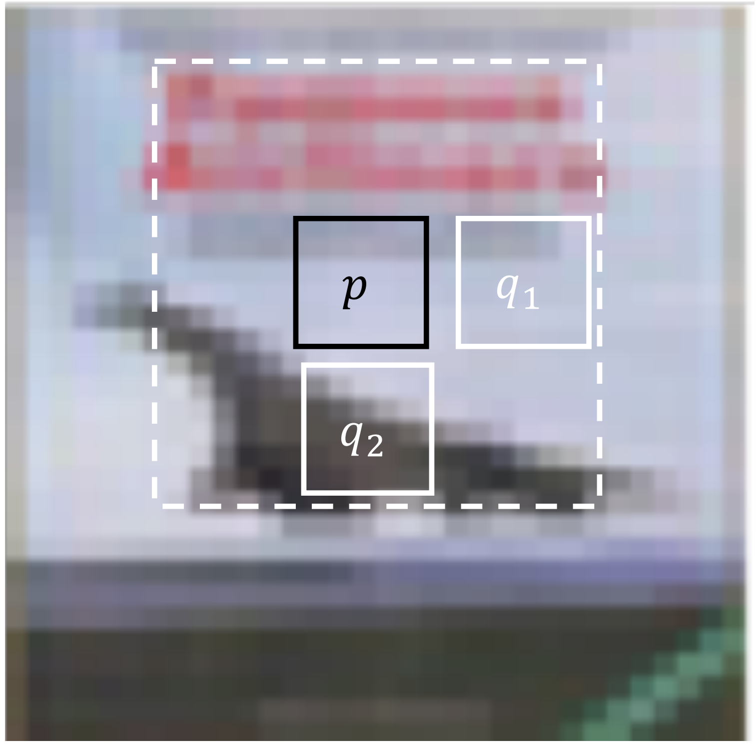


3x3 Median Filter

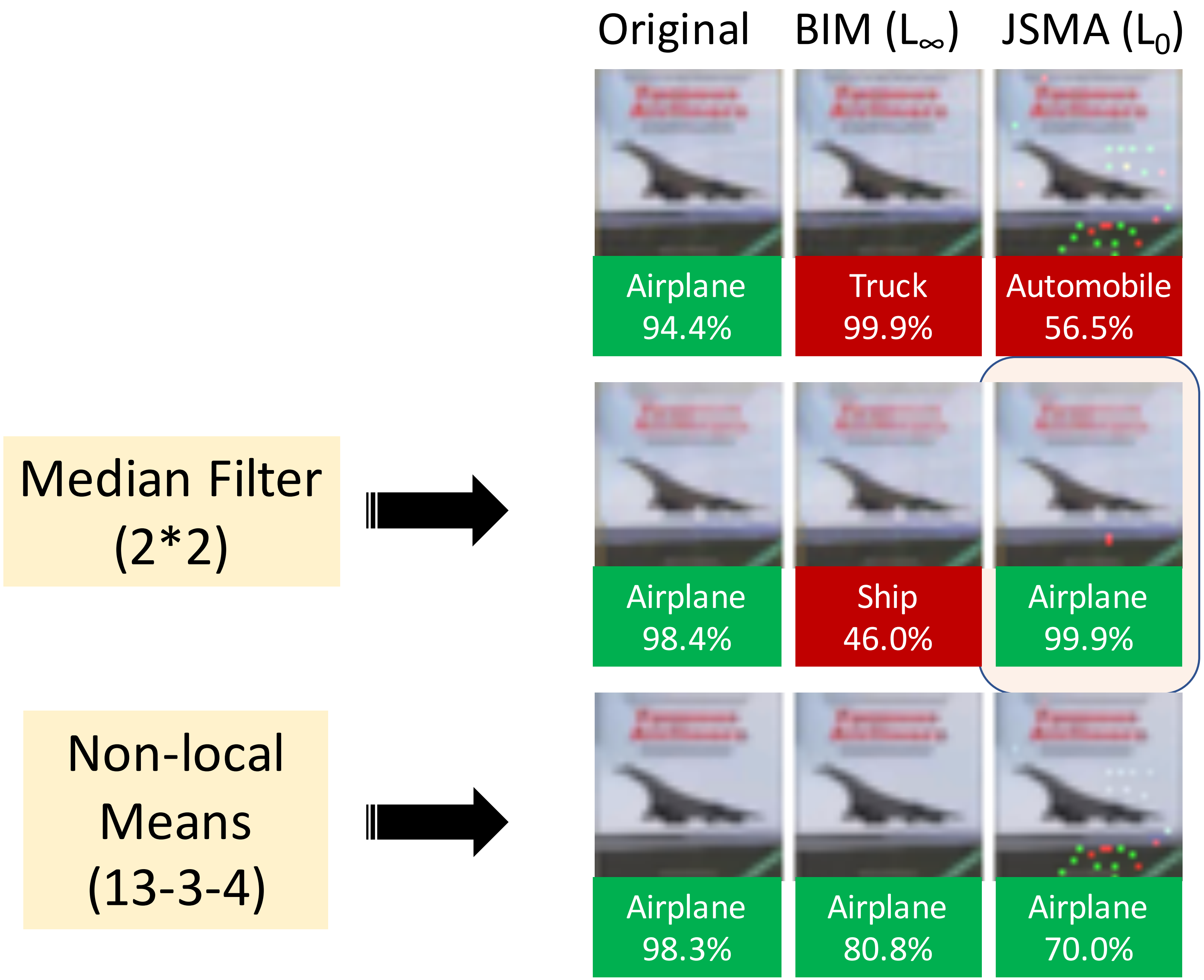
* Image from https://sultanofswing90.wordpress.com/tag/image-processing/

Spatial Smoothing: Non-local Means

- Replace a patch with weighted mean of similar patches.
- Preserve more edges.



$$p' = \sum w(p, q_i) \times q_i$$


Original
BIM (L_∞)
JSMA (L_0)
Airplane 94.4%
Truck 99.9%
Automobile 56.5%
Median Filter (2*2)
Airplane 98.4%
Ship 46.0%
Airplane 99.9%
Non-local Means (13-3-4)
Airplane 98.3%
Airplane 80.8%
Airplane 70.0%

Accuracy with Spatial Smoothing

Dataset	Squeezer	Adversarial Examples (FGSM, BIM, CW_∞, Deep Fool, CW_2, CW_0)	Legitimate Images
ImageNet	None	2.78%	69.70%
	Median Filter 2*2	68.11%	65.40%
	Non-local Means 11-3-4	57.11%	65.40%

Baseline

Other Potential Squeezers

- Thermometer Encoding (learnable bit depth reduction)

  J Buckman, et al. *Thermometer Encoding: One Hot Way To Resist Adversarial Examples* , to appear in ICLR 2018.

- Image denoising using bilateral filter, autoencoder, wavelet, etc.

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

  F Liao, et al. *Defense against Adversarial Attacks Using High-Level Representation Guided Denoiser*, arXiv 1712.02976.

  A Prakash, et al. *Deflecting Adversarial Attacks with Pixel Deflection*, arXiv 1801.08926.

- Image resizing

  C Xie, et al. *Mitigating Adversarial Effects Through Randomization*, to appear in ICLR 2018.

Experimental Setup

- Datasets and Models

  MNIST, 7-layer-CNN

  CIFAR-10, DenseNet

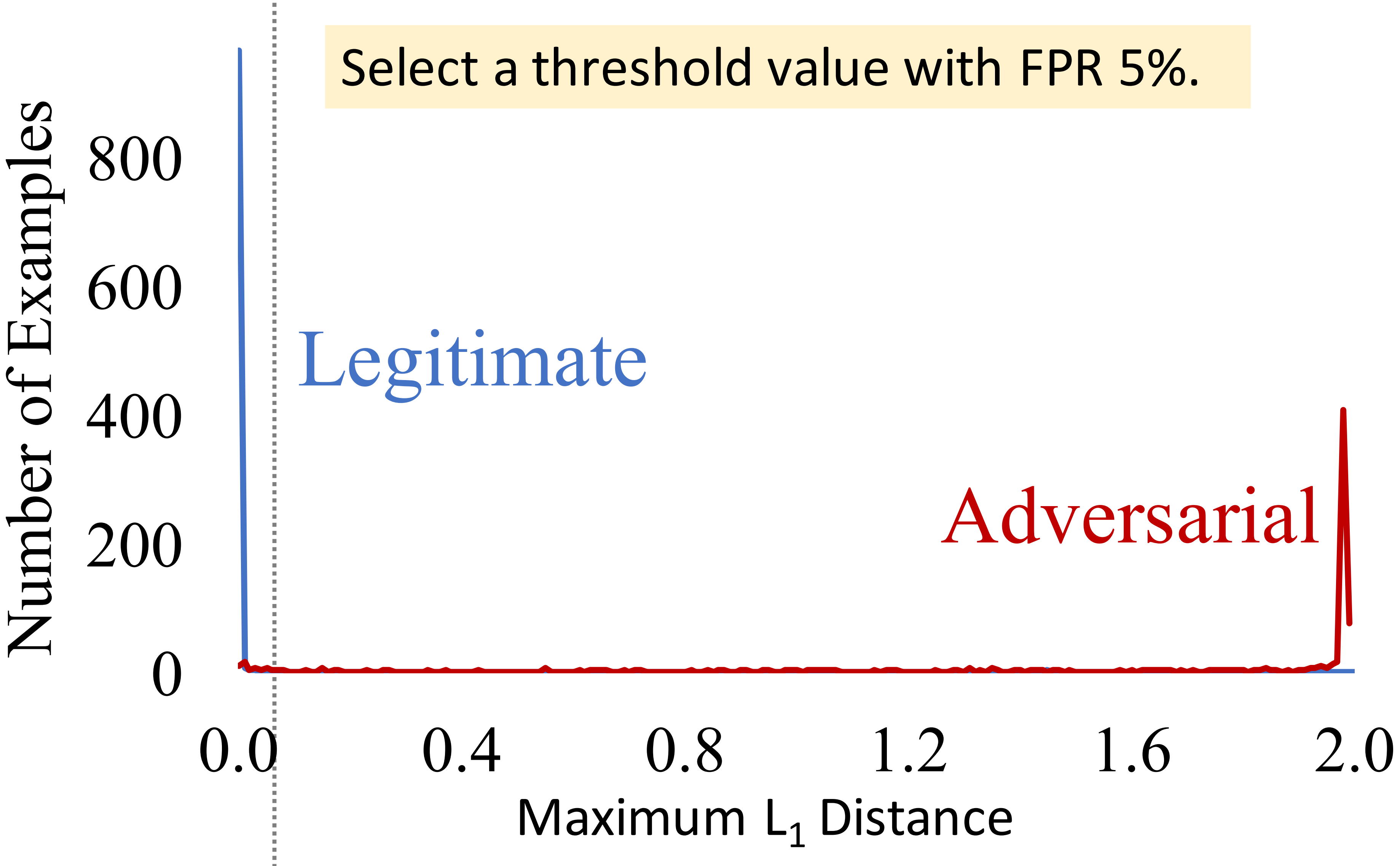
  ImageNet, MobileNet

- Attacks (100 examples for each attack)
  - Untargeted: FGSM, BIM, DeepFool
  - Targeted (Next/Least-Likely): JSMA, Carlini-Wagner $L_2/L_\infty/L_0$
- Detection Datasets
  - A balanced dataset with legitimate examples.
  - 50% for training the detector, the remaining for validation.

Threat Models

- **Oblivious adversary**: The adversary has full knowledge of the target model, but is not aware of the detector.
- **Adaptive adversary**: The adversary has full knowledge of the target model and the detector.

Train a detector (MNIST)



Detect Successful Adv. Examples (MNIST)

Bit Depth Reduction is more effective on L_∞ and L_2 attacks.

Median Smoothing is more effective on L_0 attacks.

Squeezer	L_∞ Attacks			L_2 Attacks	L_0 Attacks	
	FGSM	BIM	CW_∞	CW_2	CW_0	JSMA
1-bit Depth	**100%**	**97.9%**	**100%**	**100%**	55.6%	**100%**
Median 2*2	73.1%	27.7%	**100%**	94.4%	**82.2%**	**100%**
[Best Single]	100%	97.9%	100%	100%	82.2%	100%
Joint	100%	97.9%	100%	100%	**91.1%**	100%

Joint detection improves performance.

Aggregated Detection Results

Dataset	Squeezers	Threshold	False Positive Rate	Detection Rate (SAEs)	ROC-AUC Exclude FAEs
MNIST	Bit Depth (1-bit), Median (2x2)	0.0029	3.98%	98.2%	99.44%
CIFAR-10	Bit Depth (5-bit), Median (2x2), Non-local Mean (13-3-2)	1.1402	4.93%	84.5%	95.74%
ImageNet	Bit Depth (5-bit), Median (2x2), Non-local Mean (11-3-4)	1.2128	8.33%	85.9%	94.24%

Best Result

Threat Models

- **Oblivious attack**: The adversary has full knowledge of the target model, but is not aware of the detector.

- **Adaptive attack**: The adversary has full knowledge of the target model and the detector.

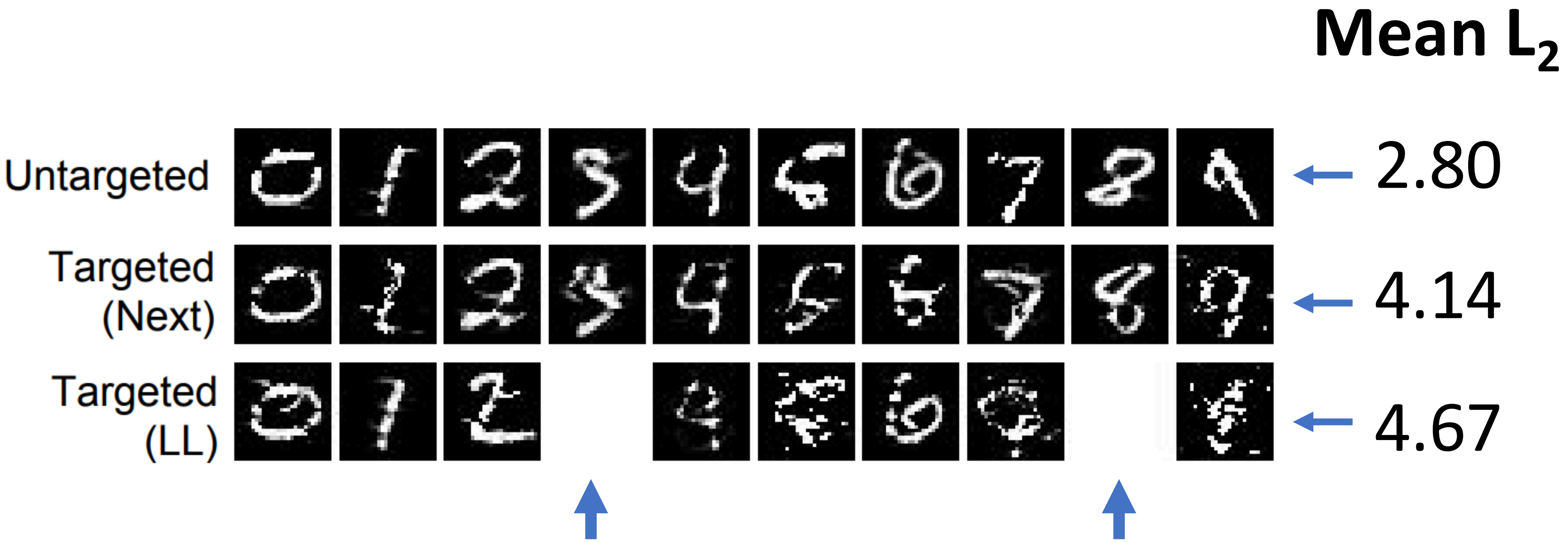
Adaptive Adversary

Adaptive CW_2 attack, unbounded adversary.

$$\underbrace{minimize\ \|g(x') - t\|}_{\text{Misclassification term}} + \underbrace{\lambda * \Delta(x, x')}_{\text{Distance term}} \underbrace{+\, k * L_1 score(x')}_{\text{Detection term}}$$

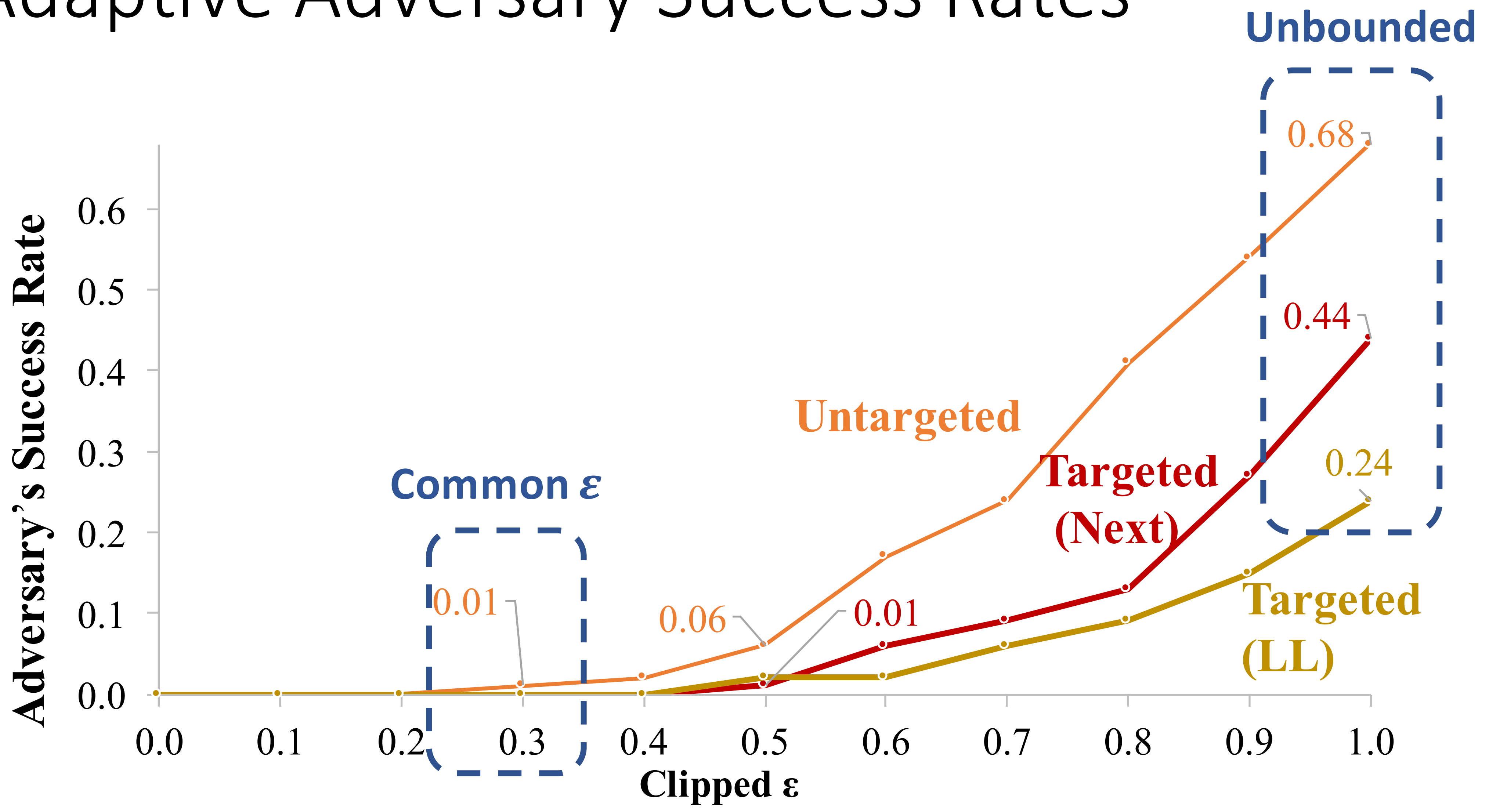
Warren He, James Wei, Xinyun Chen, Nicholas Carlini, Dawn Song,
Adversarial Example Defense: Ensembles of Weak Defenses are not Strong, USENIX WOOT’17.

Adaptive Adversarial Examples



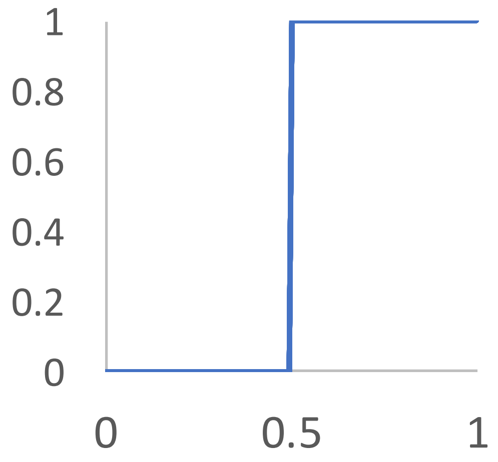
No successful adversarial examples were found for images originally labeled as 3 or 8.

Adaptive Adversary Success Rates

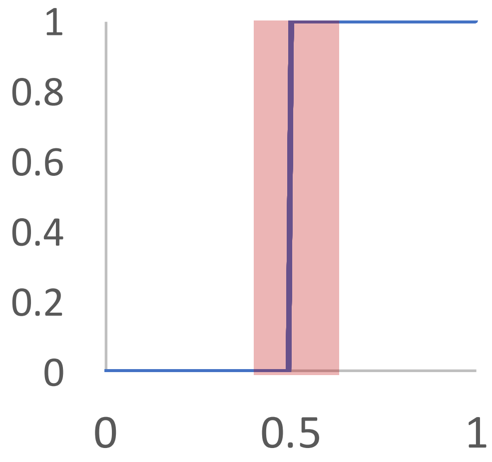


Counter Measure: Randomization

- Binary filter threshold := 0.5



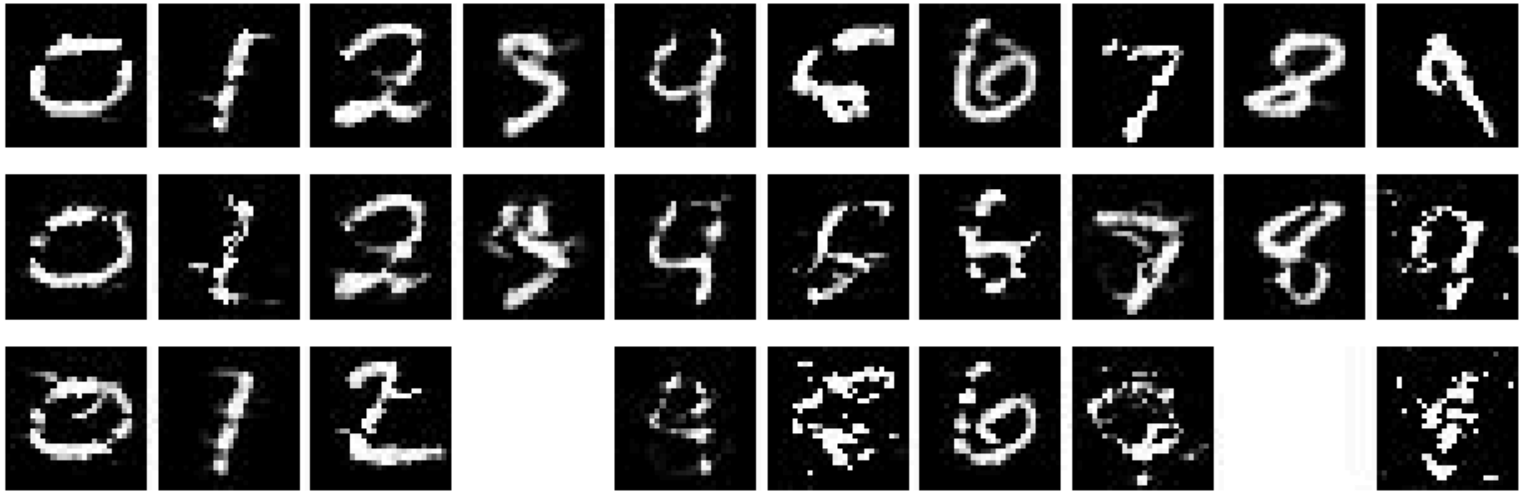
threshold := $\mathcal{N}(0.5, 0.0625)$



- Strengthen the adaptive adversary

  Attack an ensemble of 3 detectors with thresholds := [0.4, 0.5, 0.6]

Attack Deterministic Detector



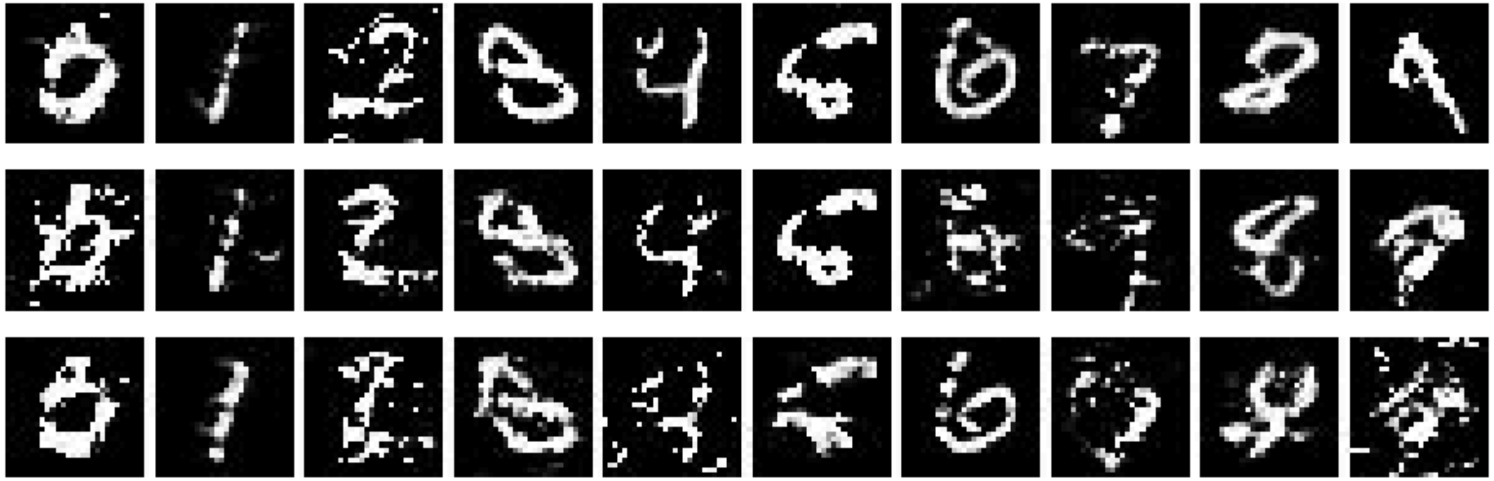
Mean L_2

2.80, Untargeted

4.14, Targeted-Next

4.67, Targeted-LL

Attack Randomized Detector



3.63, Untargeted

5.48, Targeted-Next

5.76, Targeted-LL

Conclusion

- Feature Squeezing hardens deep learning models.
- Feature Squeezing gives advantages to the defense side in the arms race with adaptive adversary.

Thank you!

Reproduce our results using EvadeML-Zoo: https://evadeML.org/zoo

Backup Slides

NIPS'17 AML Defense Challenge

- Different threat model: Unknown target model and defense.
- Top 4 defense submissions:

	Username	Basic Idea	Score
1	liaofz	Denoise autoencoder trained with adv. examples + model ensemble	95.32
2	cihangxie	Random resizing + random padding.	92.35
3	anlthms	JPEG compression + random affine transformation + model ensemble.	91.48
4	erkowa	2x2 Median filter + model ensemble.	91.20

None of them is robust against adaptive adversary.