Feature Squeezing:

Detecting Adversarial Examples in Deep Neural Networks

Weilin Xu David Evans **Yanjun Qi**

http://www.cs.virginia.edu/yanjun/

UNIVERSITY of VIRGINIA

@ ARO Invitational Workshop on Foundations of Autonomous Adaptive Cyber Systems / 2019/04/23

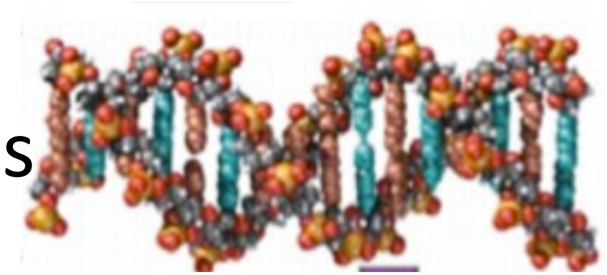
Deep Learning is Solving Many of Our Problems!

Auto-Driving Car

Voice Assistant

Spam Detector

Medical Genomics



Classifiers Under Attack:
Adversary Adapts

Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition

Mahmood Sharif
Carnegie Mellon University
Pittsburgh, PA, USA
mahmoods@cmu.edu

Sruti Bhagavatula
Carnegie Mellon University
Pittsburgh, PA, USA
srutib@cmu.edu

Lujo Bauer
Carnegie Mellon University
Pittsburgh, PA, USA
lbauer@cmu.edu

Michael K. Reiter
University of North Carolina
Chapel Hill, NC, USA
reiter@cs.unc.edu

ACM CCS 2016

Actual images

Recognized faces

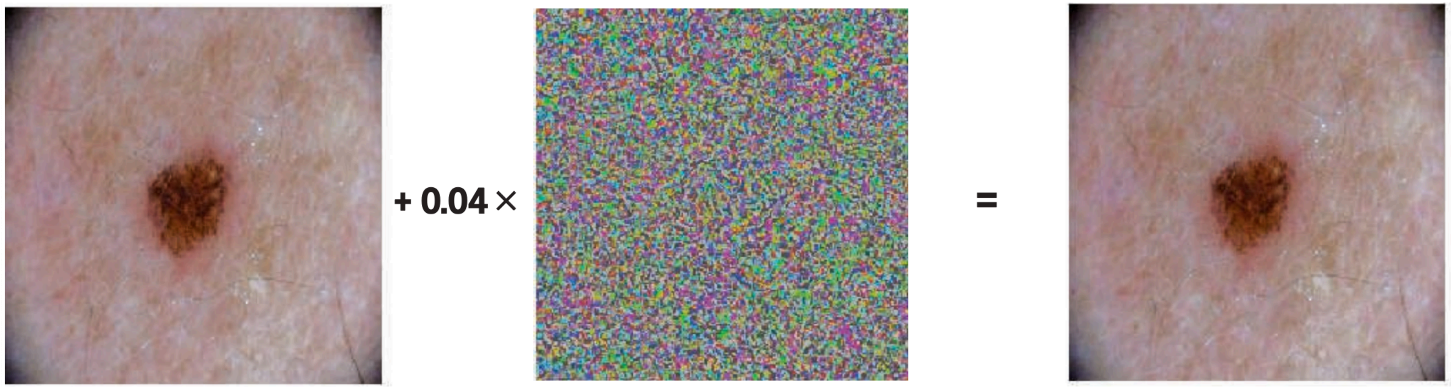
Mahmood Sharif et al. "Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition", *In CCS,* 2016.

However, Deep Learning Classifiers are Easily Fooled

Healthcare

Melanoma Diagnosis with Computer Vision

Original Image | Perturbation | Adversarial Example



Benign | **Malignant**

Samuel G Finlayson et al. “Adversarial attacks on medical machine learning”, *Science*, 2019.

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.

Simple, Cheap, Effective!

Roadmap

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

Background: Machine Learning

- Machine Learning: learn to find models that can generalize from observed data to unseen data

Input

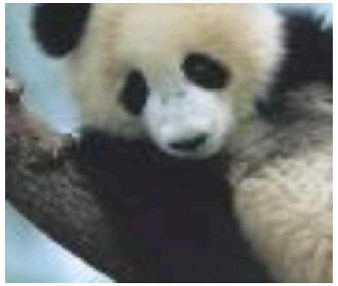
X → Model f(.) → Y

Output

X' → Model f(.) → ?

Model f(.) generalizes to Unseen X'

For instance:



↓

Trained Deep learning Model

↓

"panda"

Background: Adversarial Examples

x

Trained Deep learning Model

y: “panda”

x

$0.007 \times [noise]$

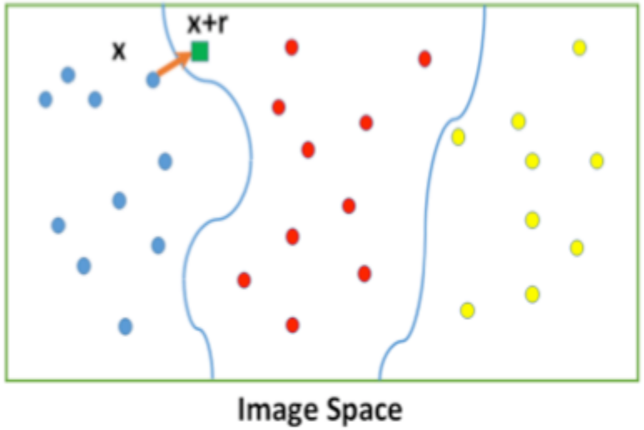
=

x’= x + r

Trained Deep learning Model

t: “gibbon”

x: original sample



x’ = x + r : adversarial sample

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

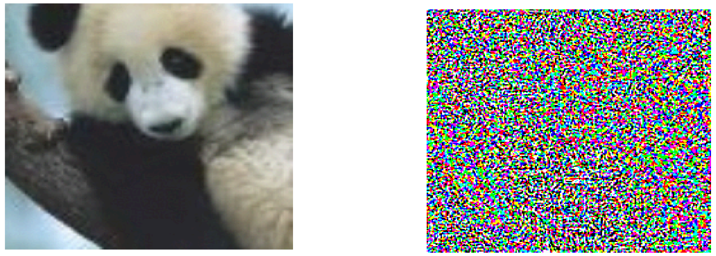
Background: Adversarial Examples

x

Trained Deep learning Model

y: “panda”

x $0.007 \times [noise]$



=

Many different variations of formulations to search for x’ from x, e.g.,

$$minimize\ \|f(x') - t\| + \lambda * \Delta(x, x')$$

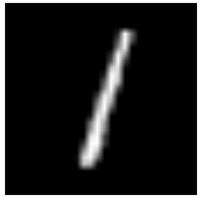
Misclassification term   Distance term

x’= x + r

Trained Deep learning Model

t: “gibbon”

Background: Different variations of Adversarial Examples



"1"

100% confidence

Original Example

\+  BIM =  **"4"** 100%

\+  CW_2 =  **"2"** 99.9%

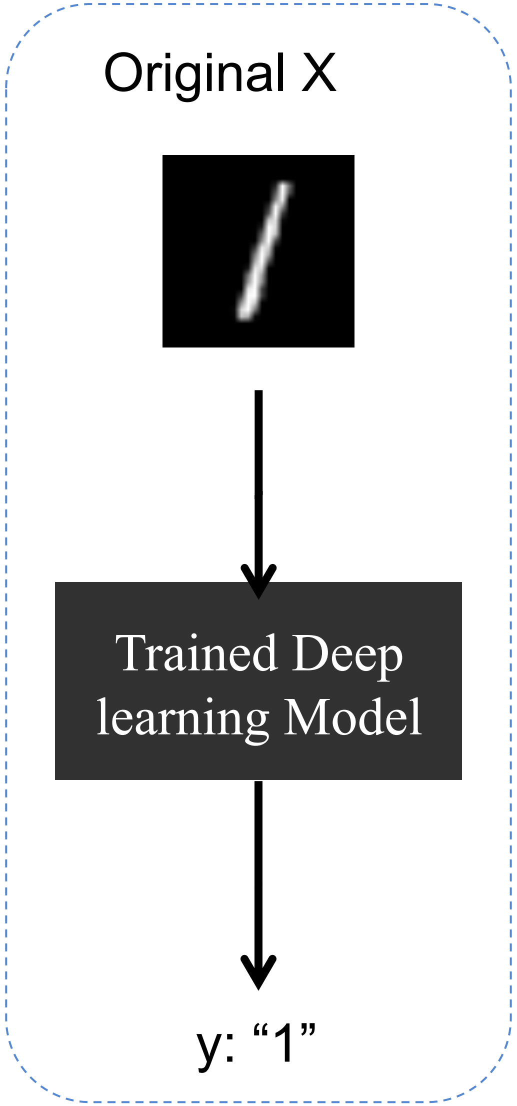
\+  JSMA =  **"2"** 83.8%

Perturbations

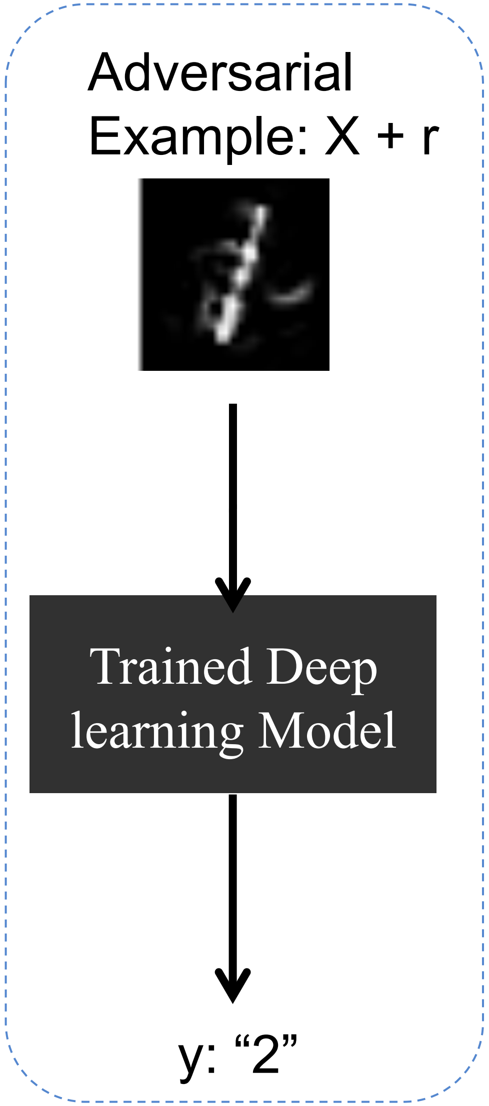
Adversarial Examples

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

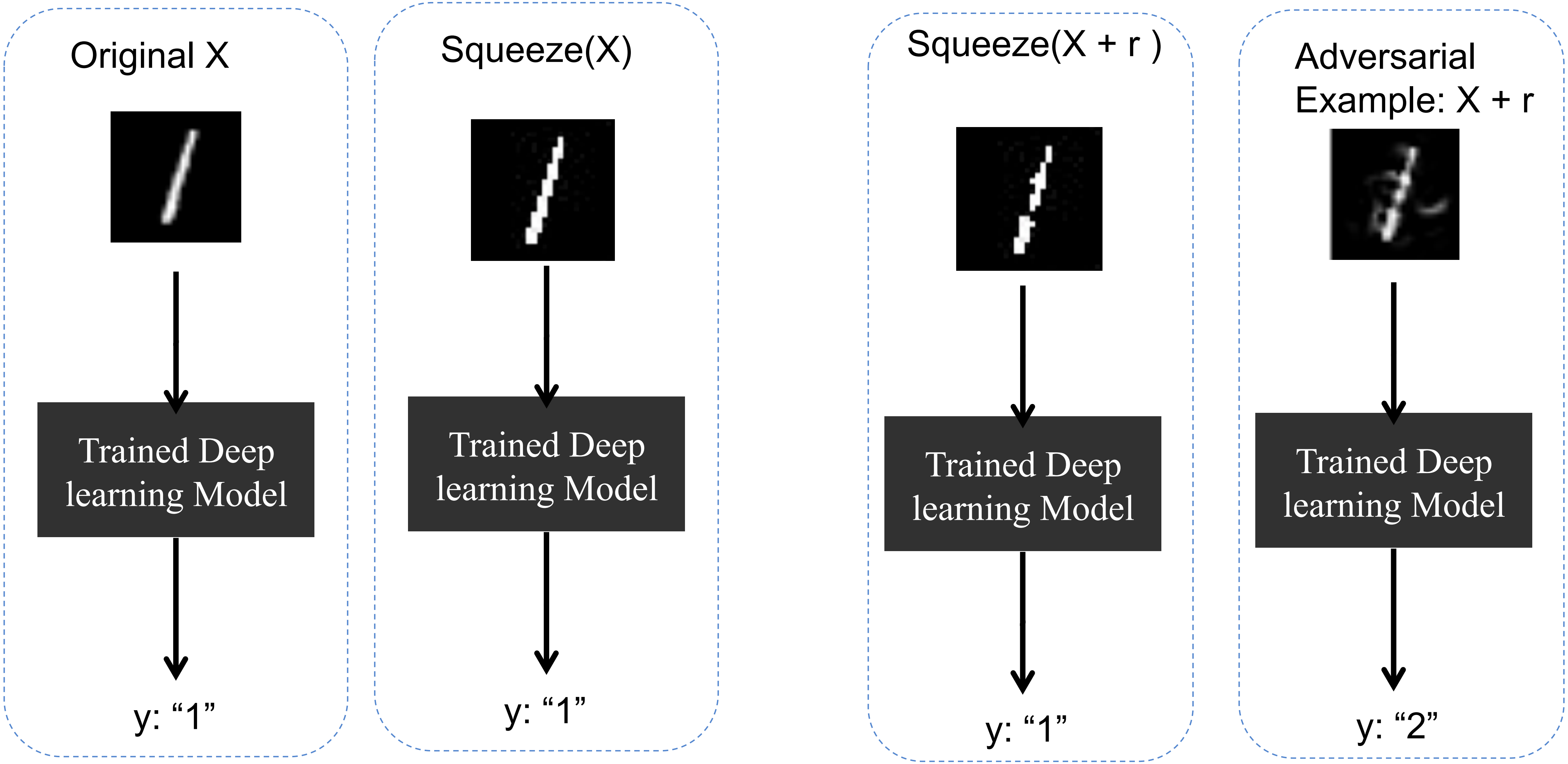
Intriguing Property of Adversarial Examples



Irrelevant features used in classification tasks are the major cause of adversarial examples.

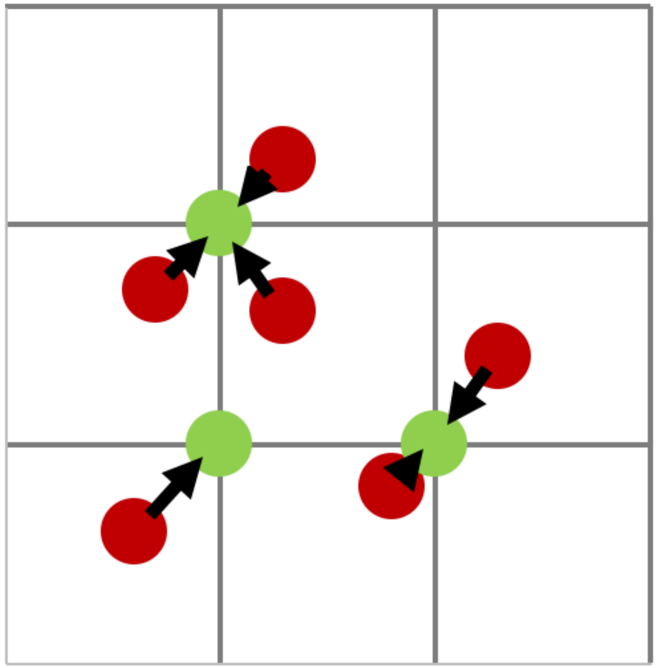


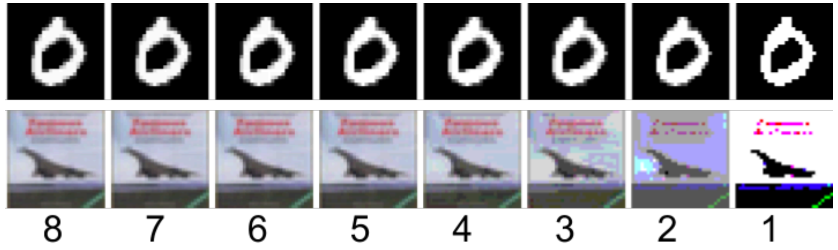
Intriguing Property of Adversarial Examples



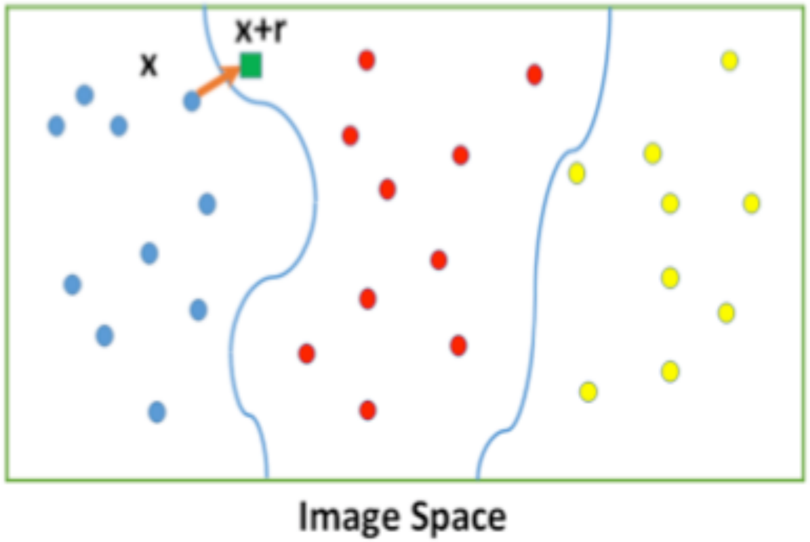
Motivation

- Irrelevant features used in classification tasks are the root cause of adversarial examples.
- The feature spaces are unnecessarily too large in deep learning tasks: e.g. raw image pixels.
- We may reduce the search space of possible perturbations available to an adversary using *Feature Squeezing*.

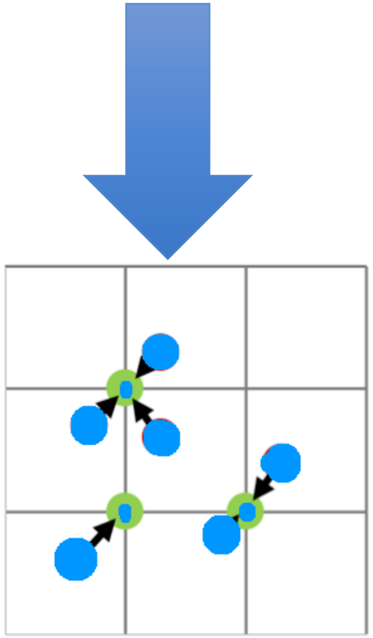




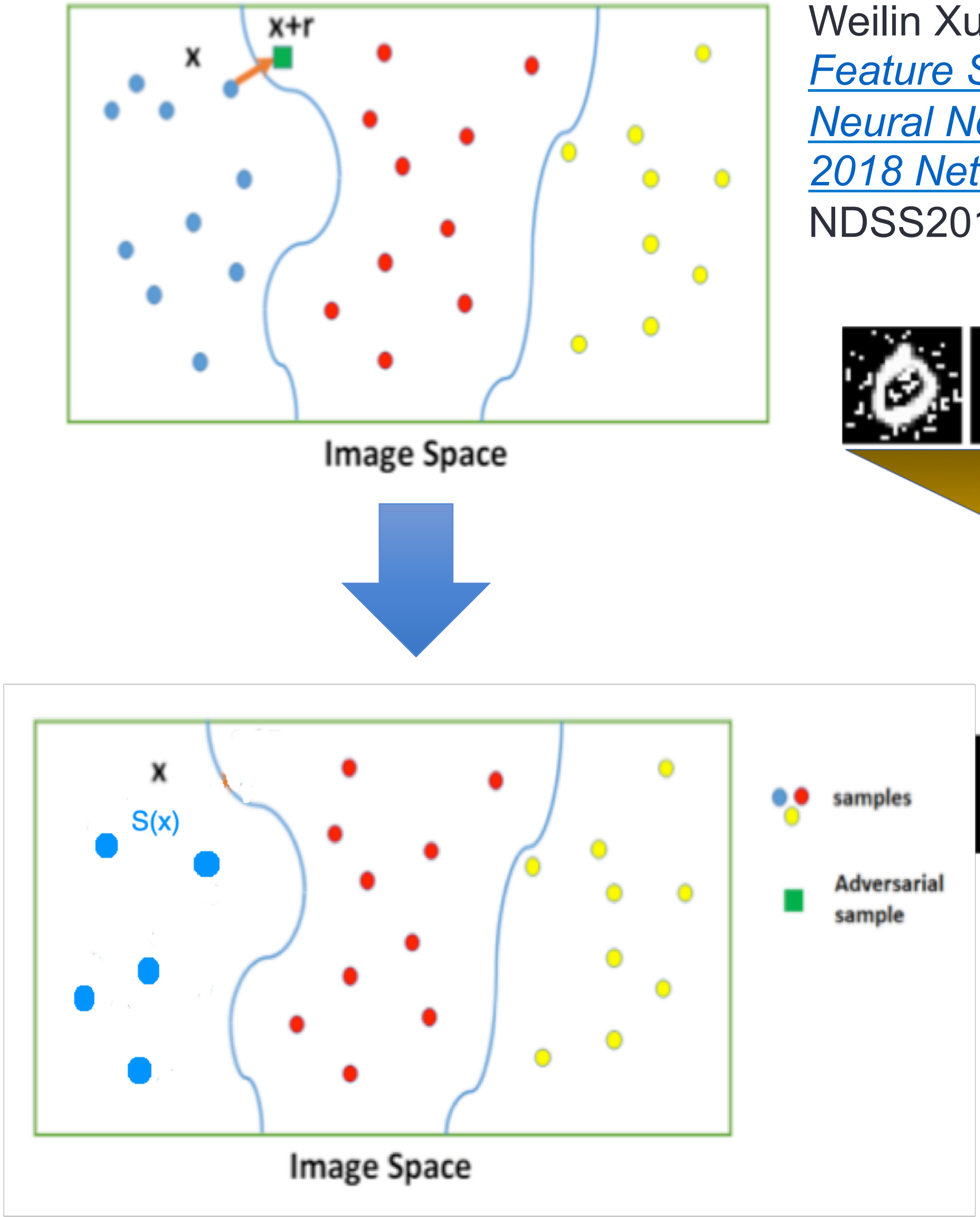
Bit Depth Reduction



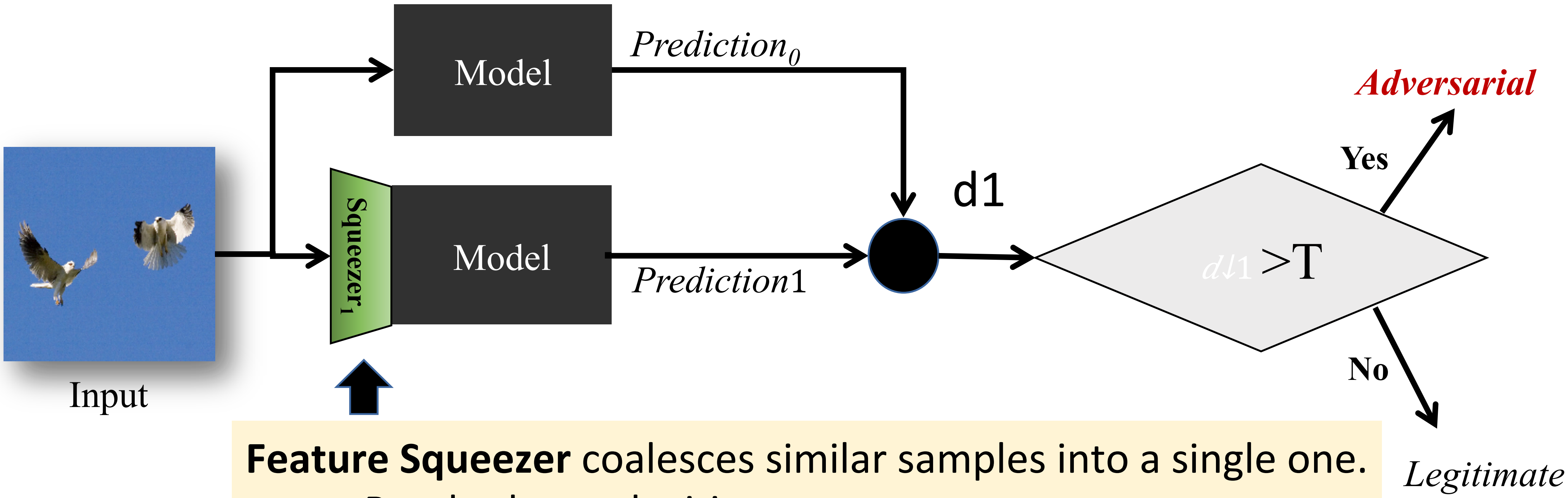
Weilin Xu, David Evans, Yanjun Qi.
[Feature Squeezing: Detecting Adversarial Examples in Deep Neural Networks](#).
[2018 Network and Distributed System Security Symposium](#).
NDSS2018



Weilin Xu, David Evans, Yanjun Qi.
[Feature Squeezing: Detecting Adversarial Examples in Deep Neural Networks]().
[2018 Network and Distributed System Security Symposium]().
NDSS2018



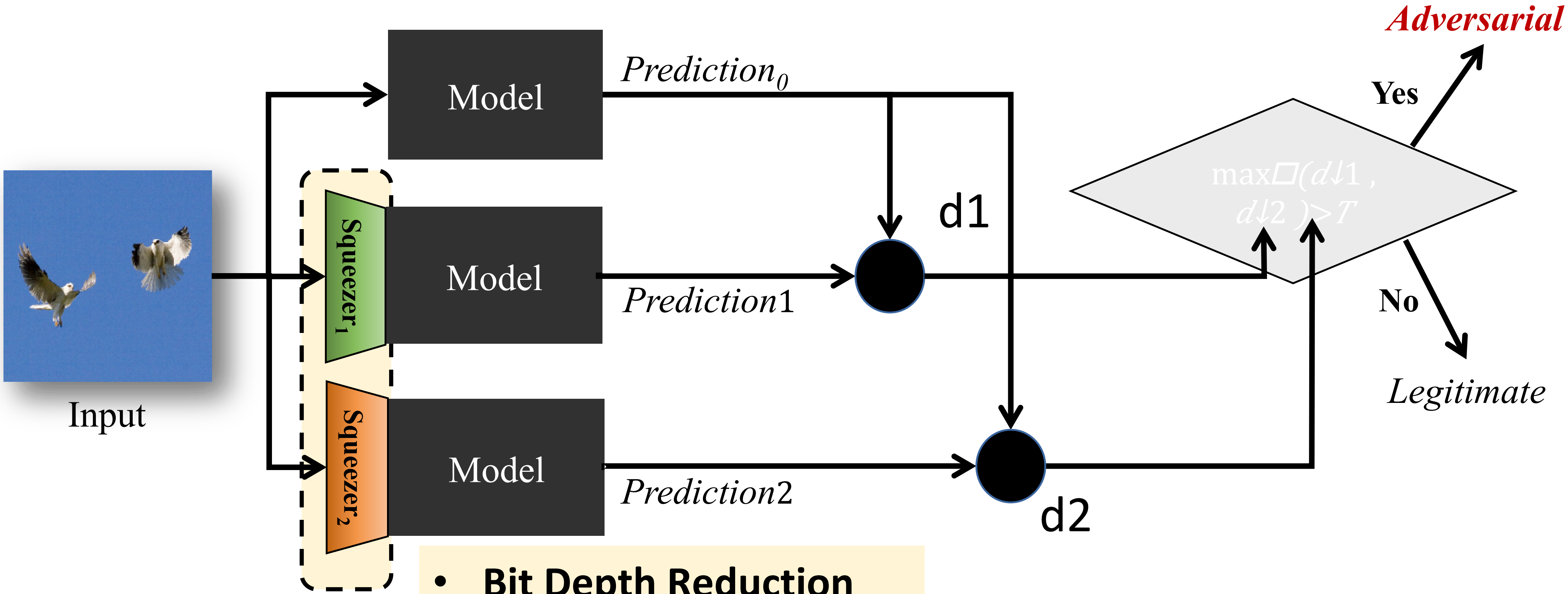
Detection Framework



Feature Squeezer coalesces similar samples into a single one.

- Barely change legitimate output.
- Destruct adversarial perturbations.

Detection Framework: Multiple Squeezers

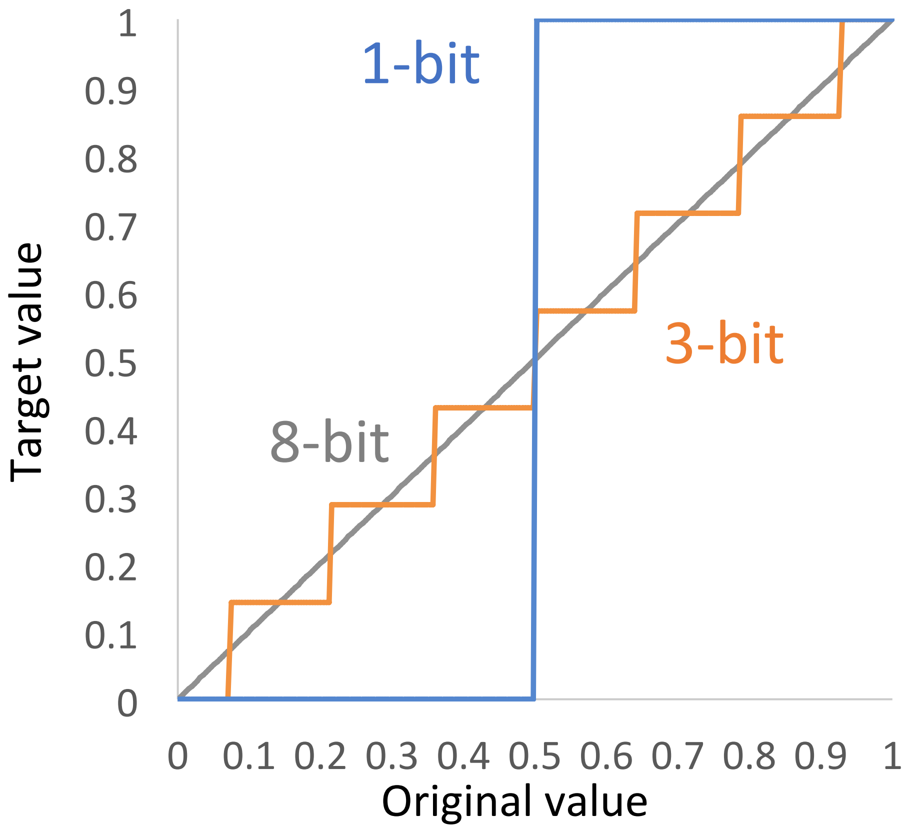


- **Bit Depth Reduction**
- **Spatial Smoothing**

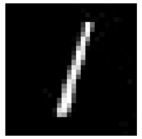
Roadmap

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

Bit Depth Reduction



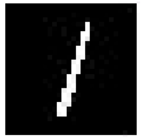
Signal Quantization



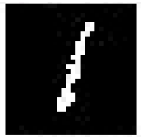
X

[0.012 0.571 …… 0.159 0.951]

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



[0. 1. …… 0. 1.]



[0. 0. …… 0. 1.]

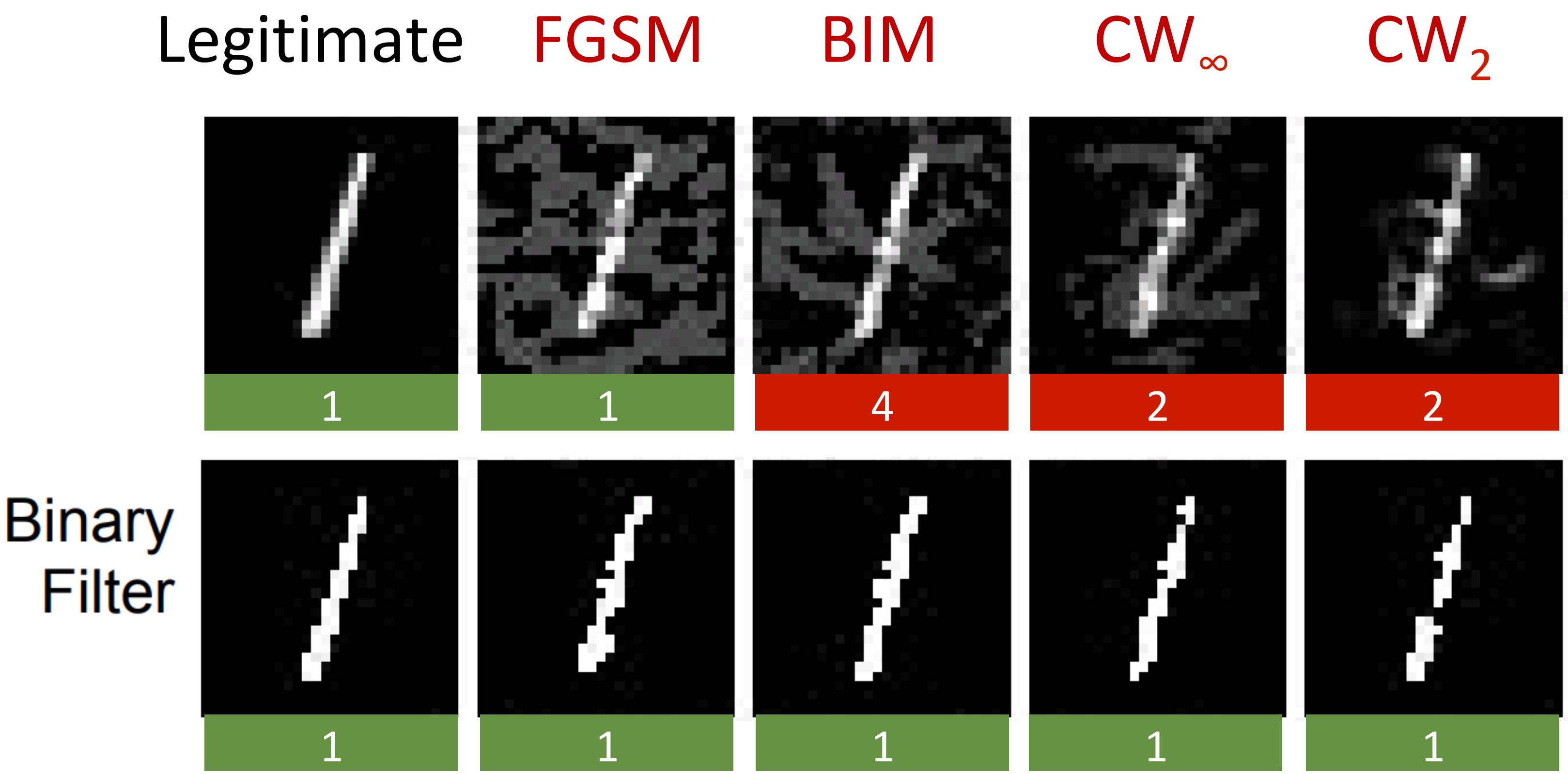
Reduce to 1-bit
$x_i = \text{round}(x_i \times 2)/2$

[0.312 0.471 …… 0.157 0.851]

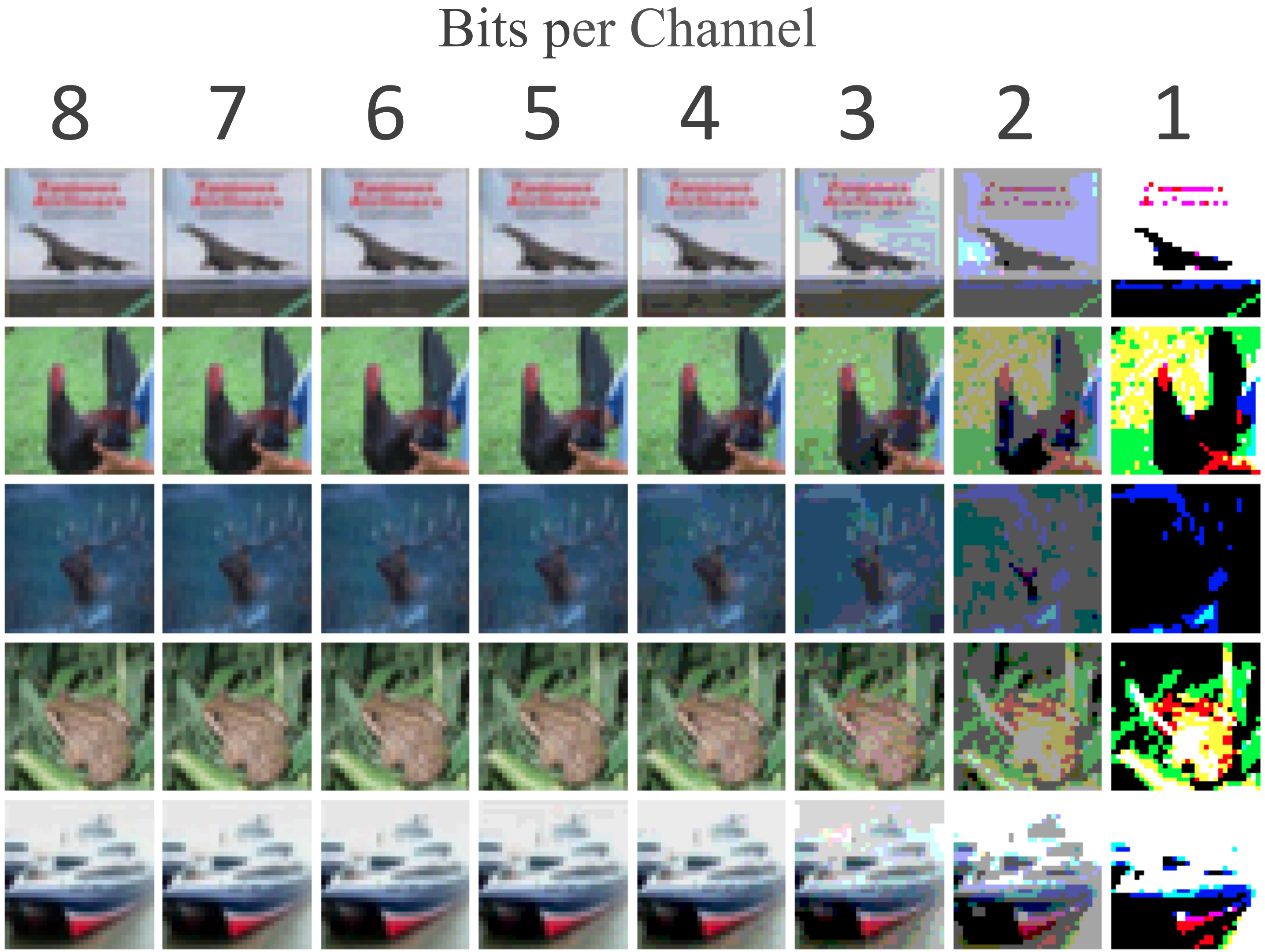
X_adv

Bit Depth Reduction

Eliminating adversarial perturbations while preserving semantics.



Bit Depth Reduction

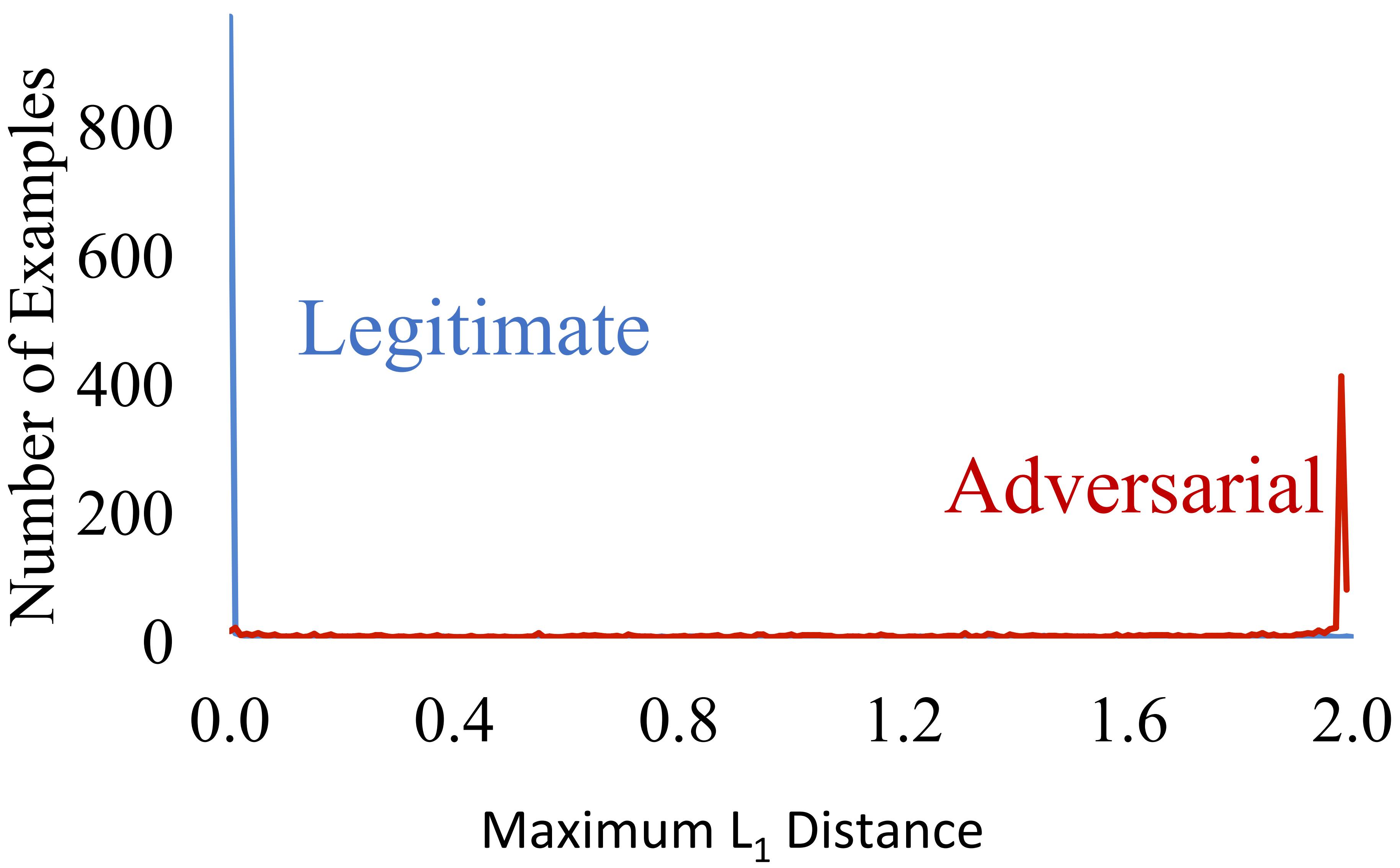


Accuracy with Bit Depth Reduction

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%

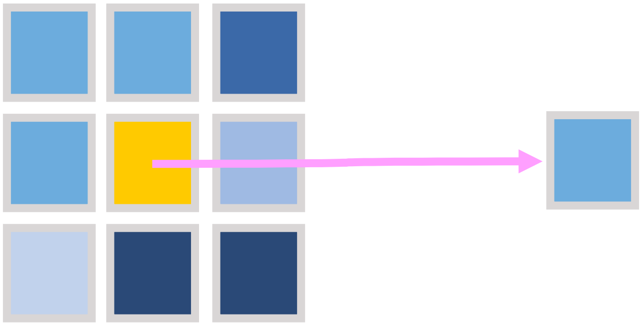
Baseline

Distribution of Distance (Prediction, Squeezed Prediction) (MNIST)



Spatial Smoothing: Median Filter

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

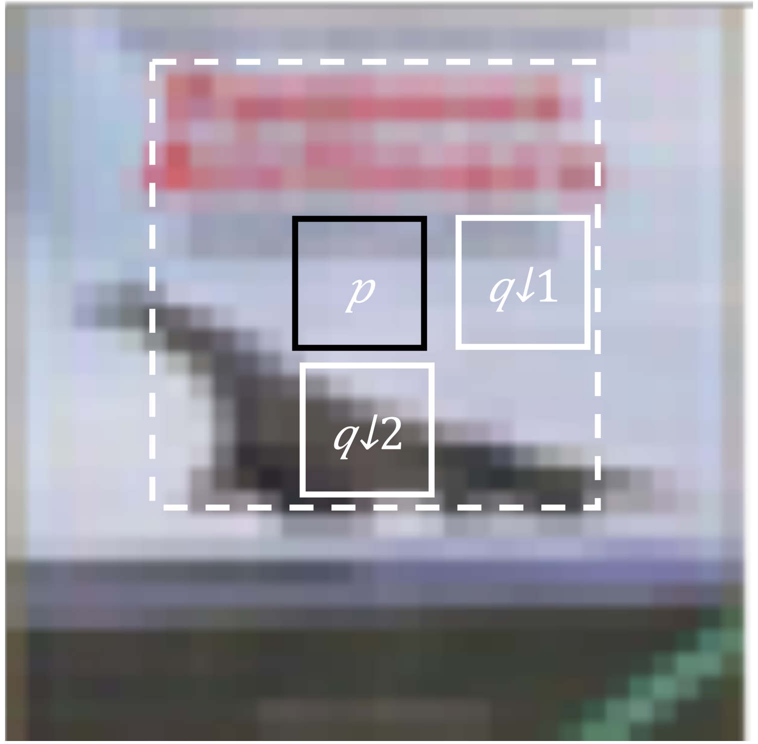


3x3 Median Filter

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

Spatial Smoothing: Non-local Means

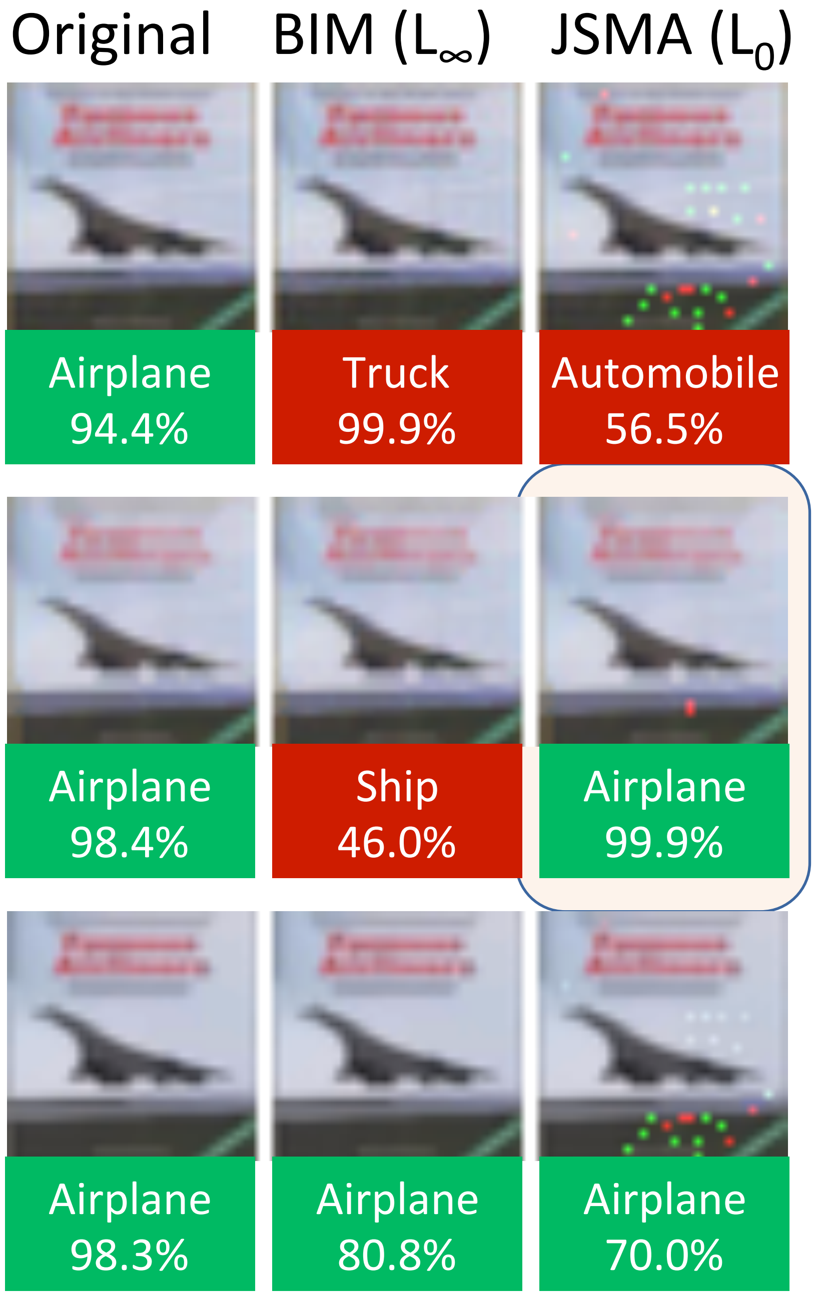
- Replace a patch with weighted mean of similar patches.
- Preserve more edges.



$$p' = \sum_{i} w(p, q_i) \times q_i$$

Median Filter (2*2)

Non-local Means (13-3-4)

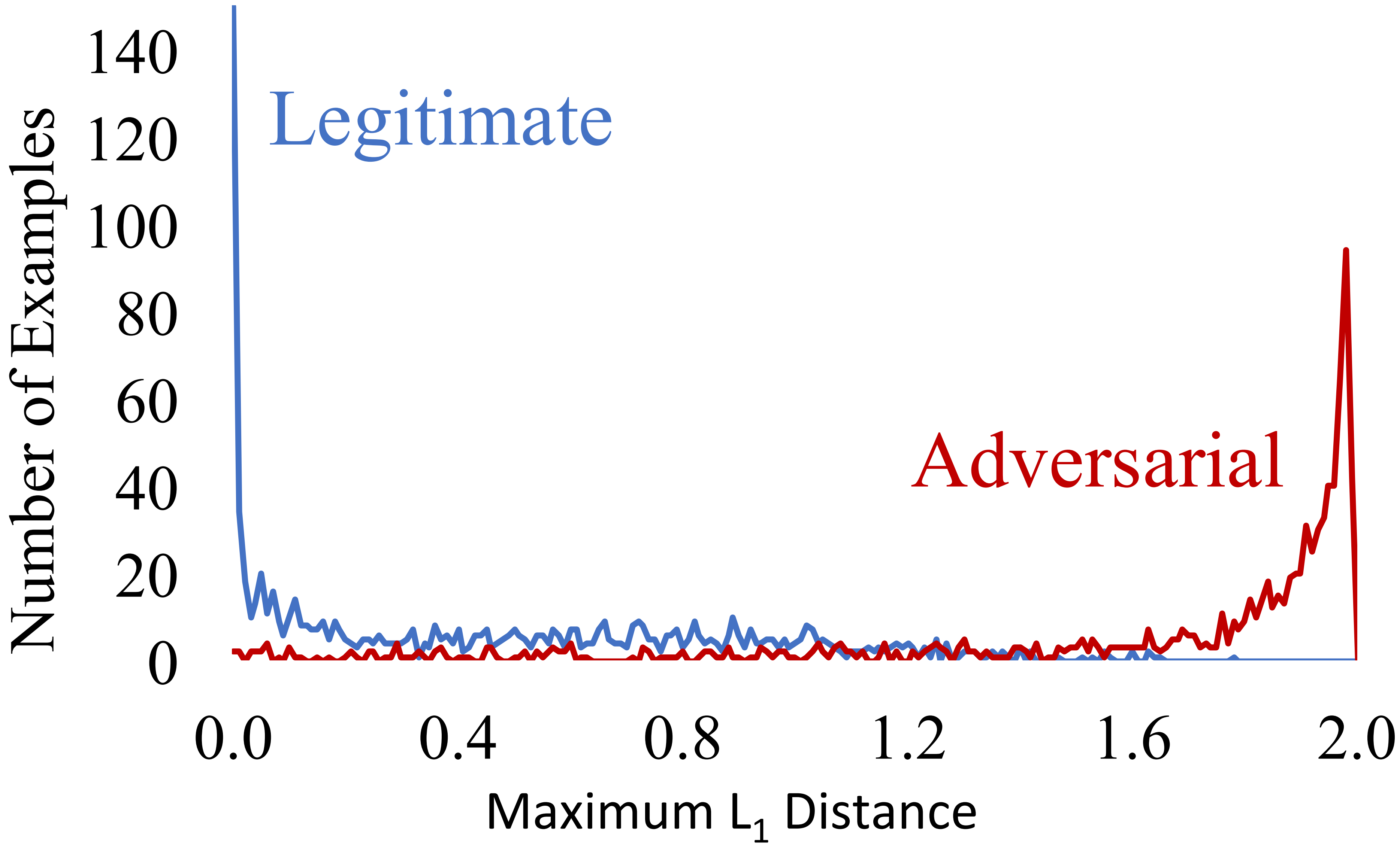


Accuracy with Spatial Smoothing

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

Distribution of Distance (Prediction, Squeezed Prediction) (ImageNet)



Other Potential Squeezers

- Thermometer Encoding (learnable bit depth reduction)

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

Roadmap

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

Empirical Evaluation: 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.

Experimental Setup

- Datasets and Models

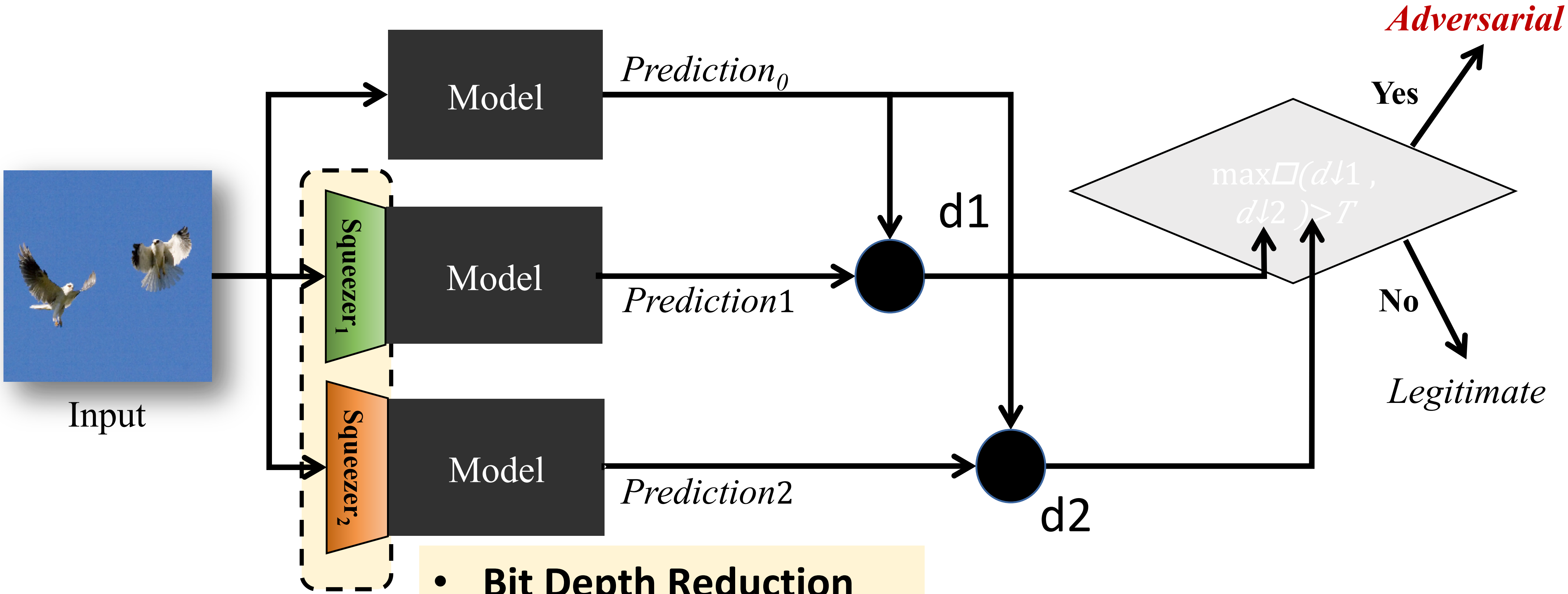
  MNIST, 7-layer-CNN

  CIFAR-10, DenseNet

  ImageNet, MobileNet

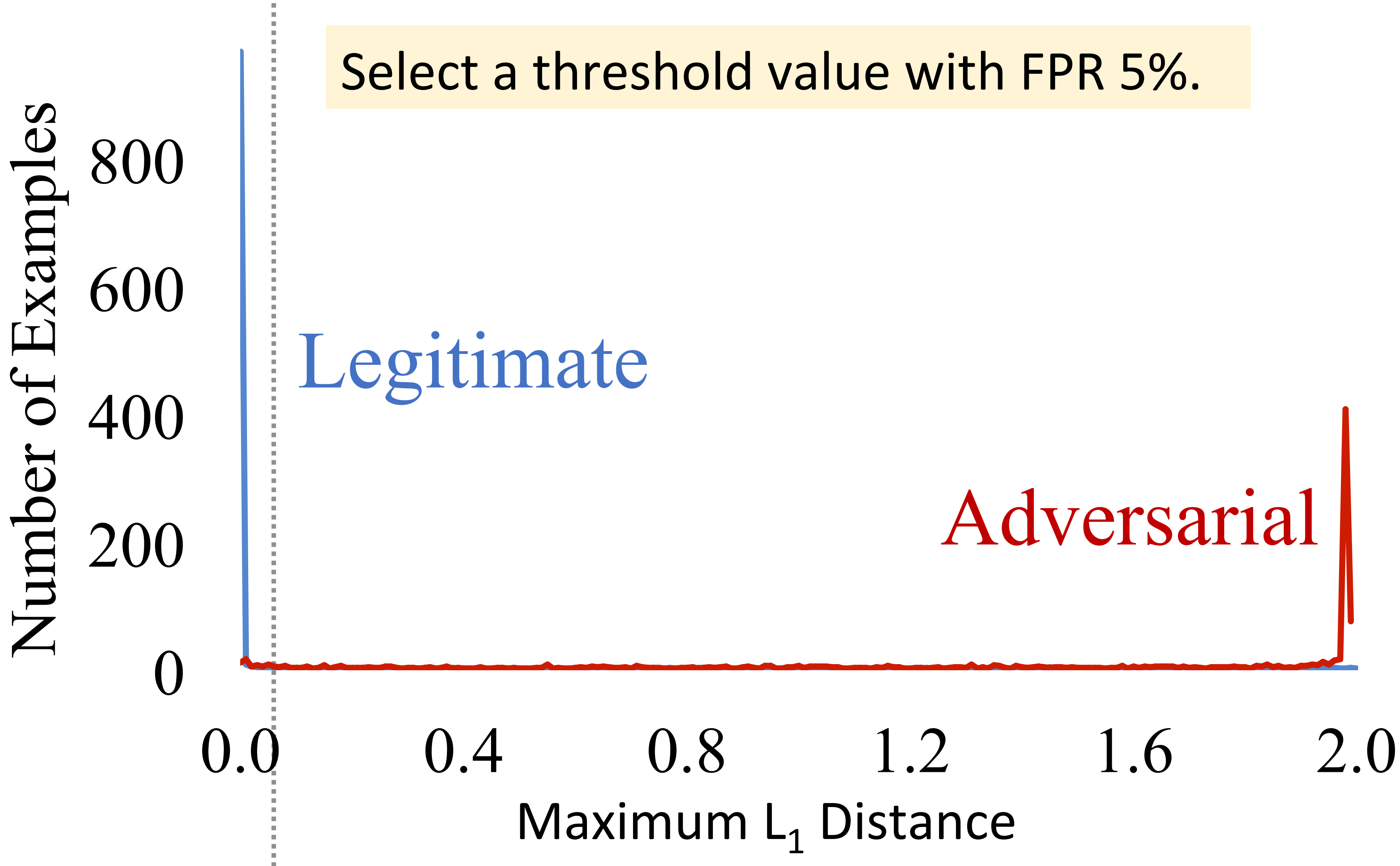
- Attacks (100 examples for each attack)
  - Untargeted: FGSM, BIM, DeepFool
  - Targeted (Next/Least-Likely): JSMA, Carlini-Wagner $L_2/L_\infty/L_0$
- Detection Datasets
  - A balanced dataset with legitimate examples.
  - 50% for training the detector, the remaining for validation.

Detection Framework: Multiple Squeezers



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

How to find T for 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%

Empirical Evaluation: 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

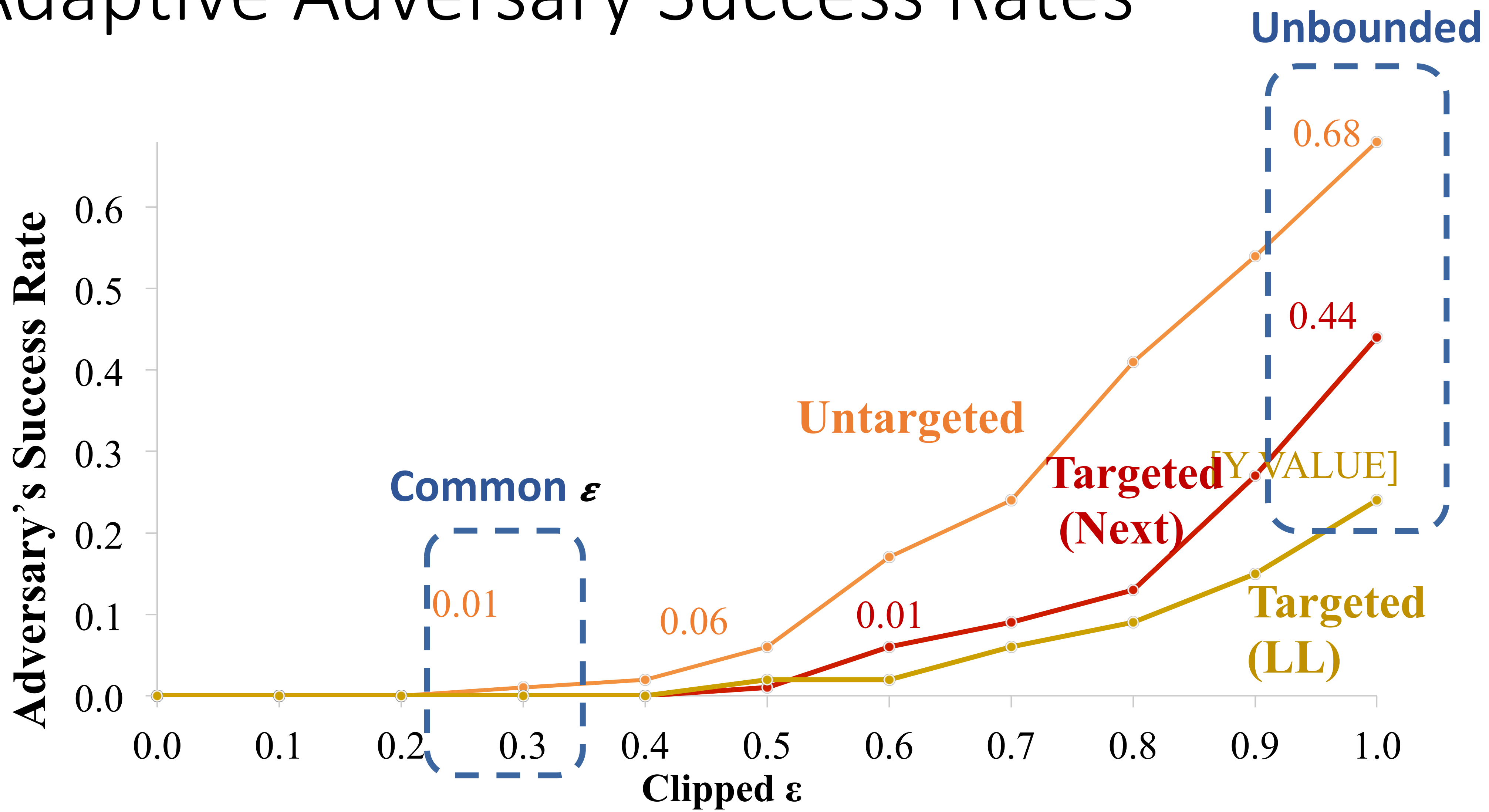
Adaptive CW_2 attack, unbounded adversary.

$$minimize\ \|f(x') - t\| + \lambda * \Delta(x, x') \qquad + k * detectScore(x')$$

Misclassification term Distance term 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 Adversary Success Rates

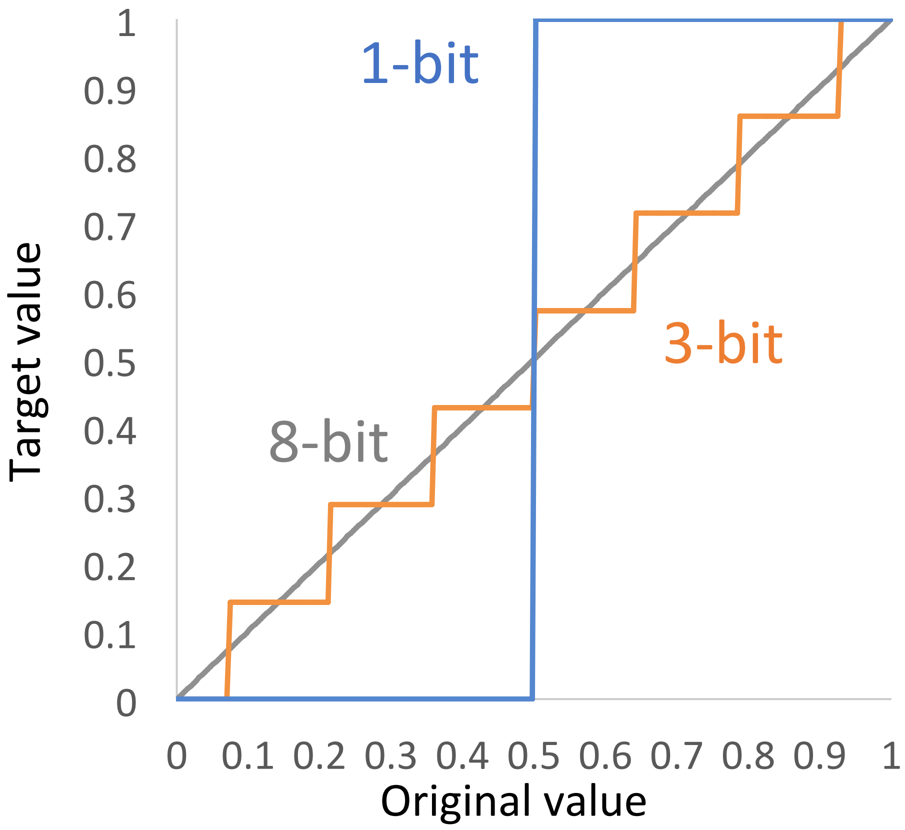


Roadmap

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

Simple feature squeezing improves robustness empirically.

Can we prove it?



Recent Work:

Feature Squeezing Improves Provable Robustness

Given model $f()$ which correctly classifies $\boldsymbol{x} \in \mathcal{X}$ as y,

$$\forall \boldsymbol{x}' \in \mathcal{X}, \Delta(\boldsymbol{x}, \boldsymbol{x}') \leq \epsilon \Rightarrow f(\boldsymbol{x}') = y$$

f is ϵ-robust on input $\boldsymbol{x} \in \mathcal{X}$ wrt a distance metric Δ.

Conclusion

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

Thank you!

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