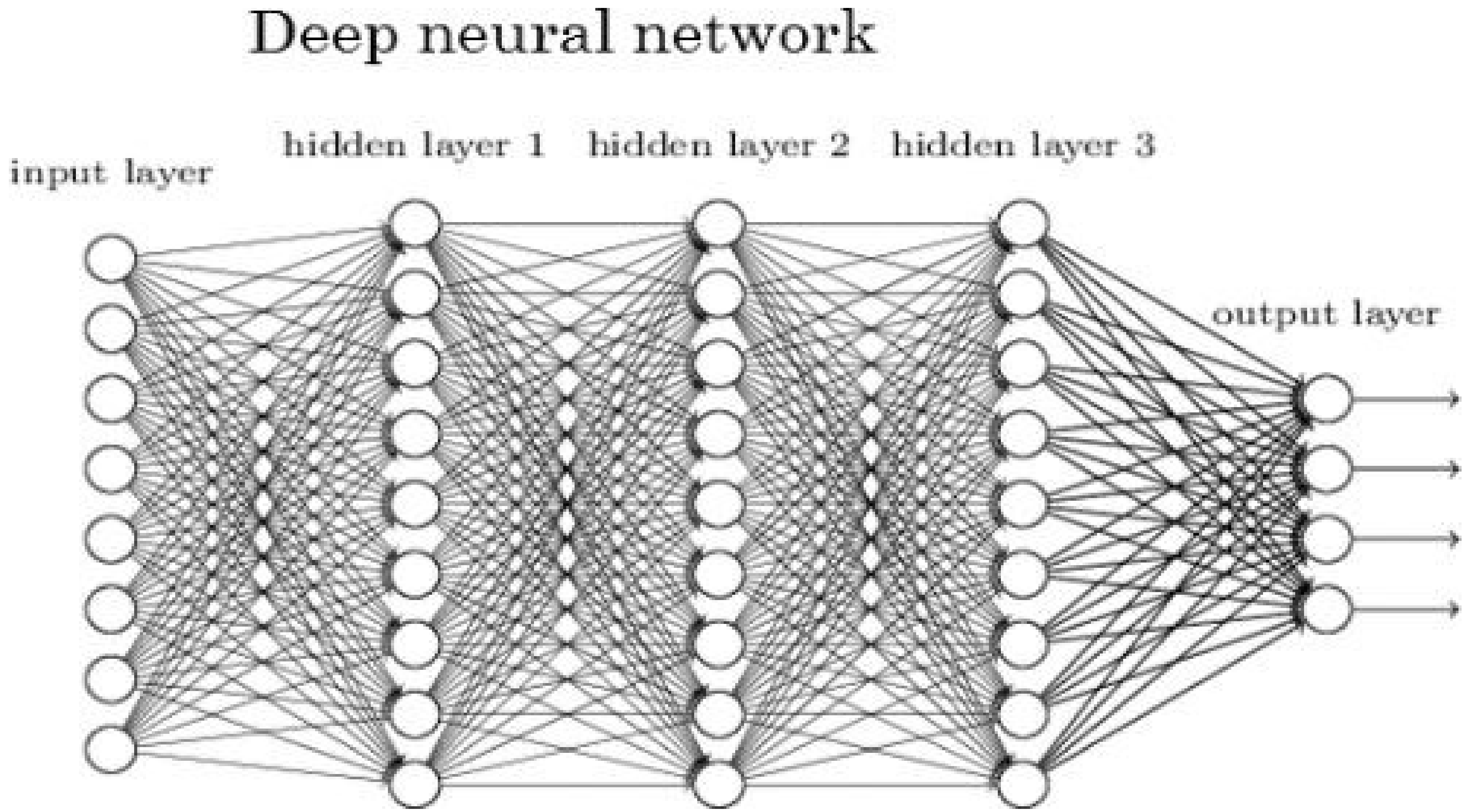
Adversarially Robust Machine Learning for Critical Applications

By: Ziad Ali

Outline

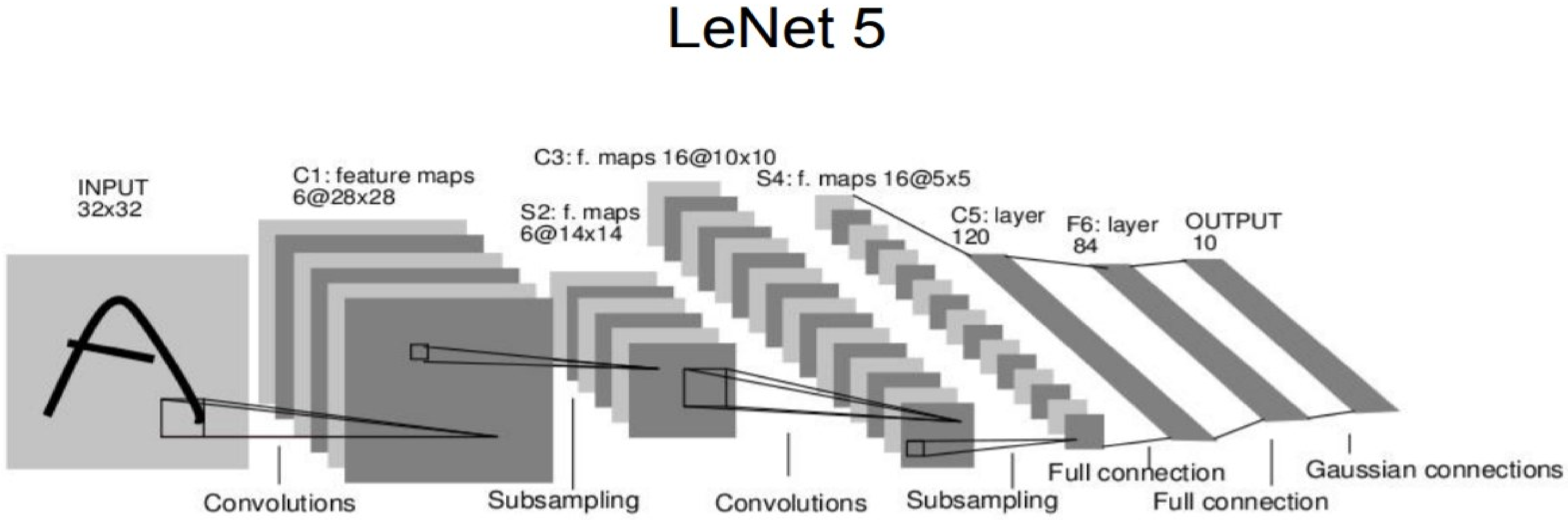
- Introduction
- Applications of Attacks
- Defenses & their limitations
- Conclusion & Future work

Deep Neural Networks: Feed Forward



Adapted from Nielsen (2015)

Convolutional Neural Network

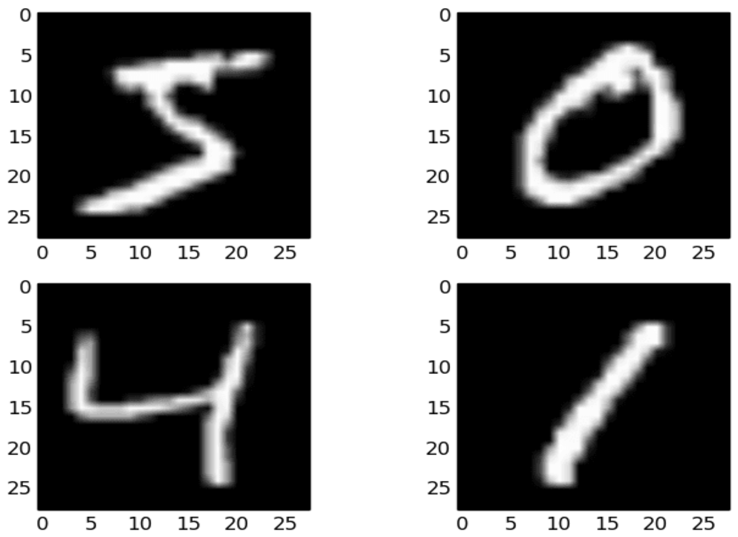


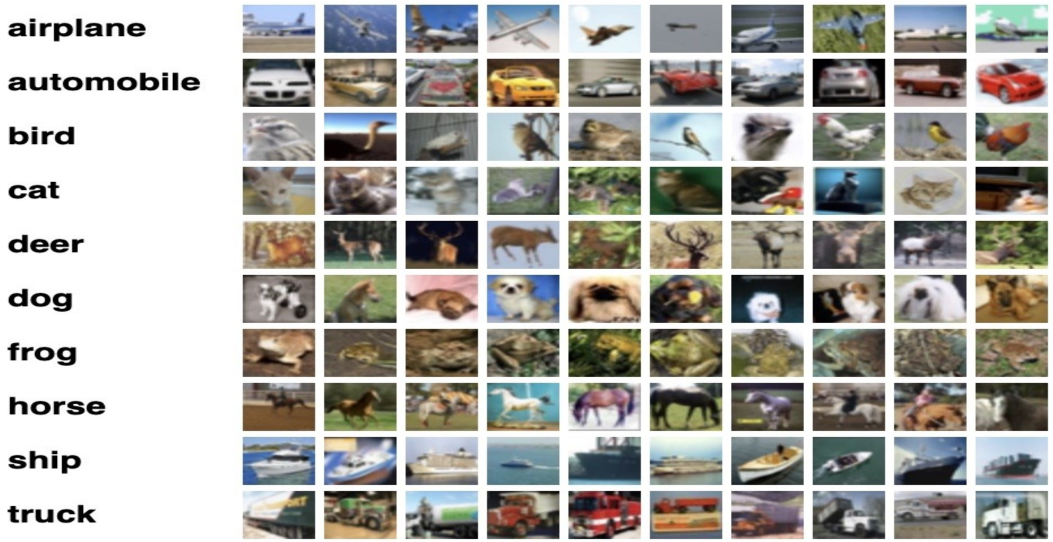
Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner,
Gradient-based learning applied to document recognition, Proc. IEEE 86(11): 2278–2324, 1998.

Why Deep Learning Applications are Critical?

- Oil & Gas industry for predicting failure
- Medicine for diagnosis of diseases
- Self-driving cars
- Speech Recognition
- DL based malware detection

Datasets: MNIST & CIFAR-10



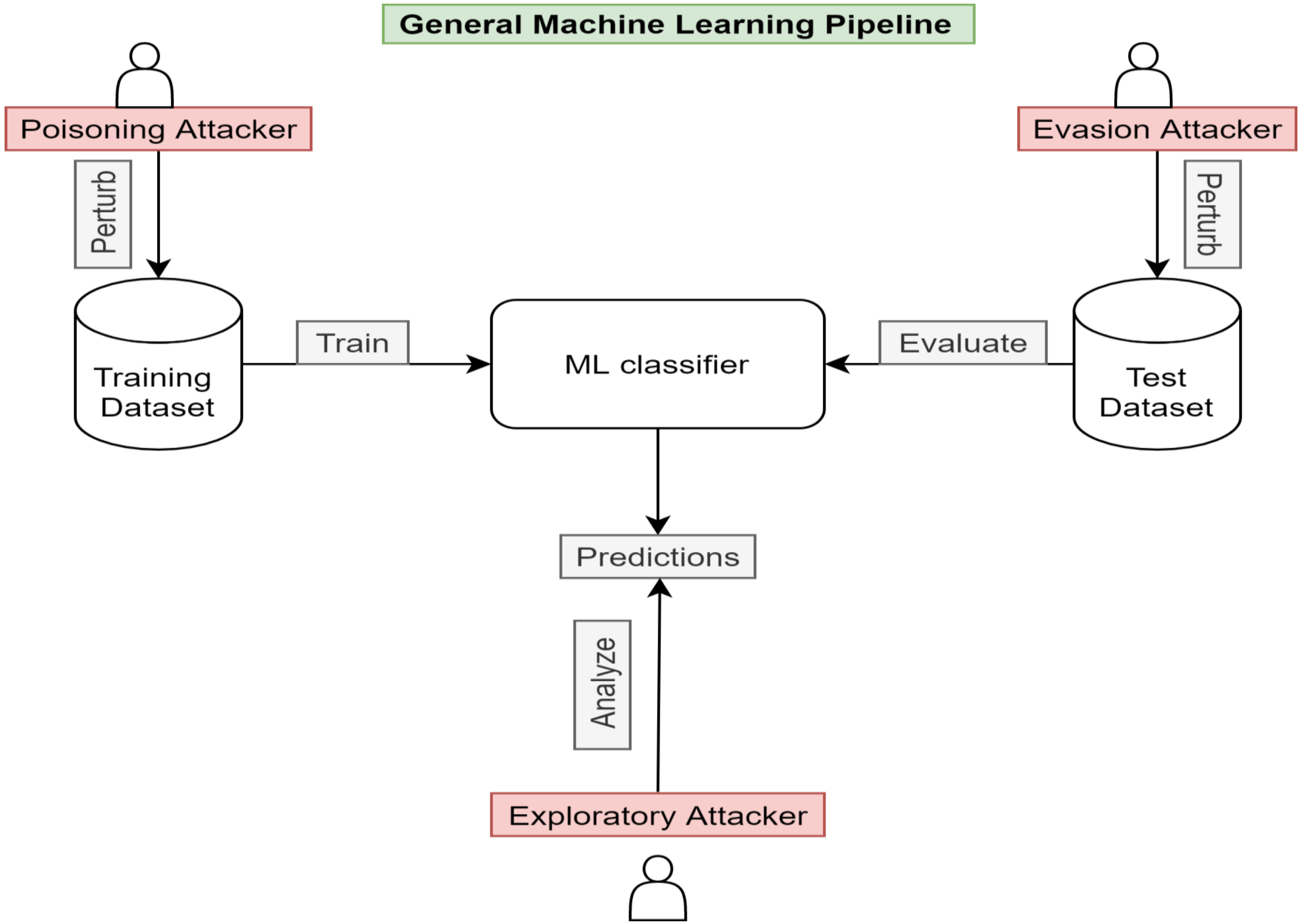


- MNIST 28x28
- 60000 Training Images
- 10000 Testing Images

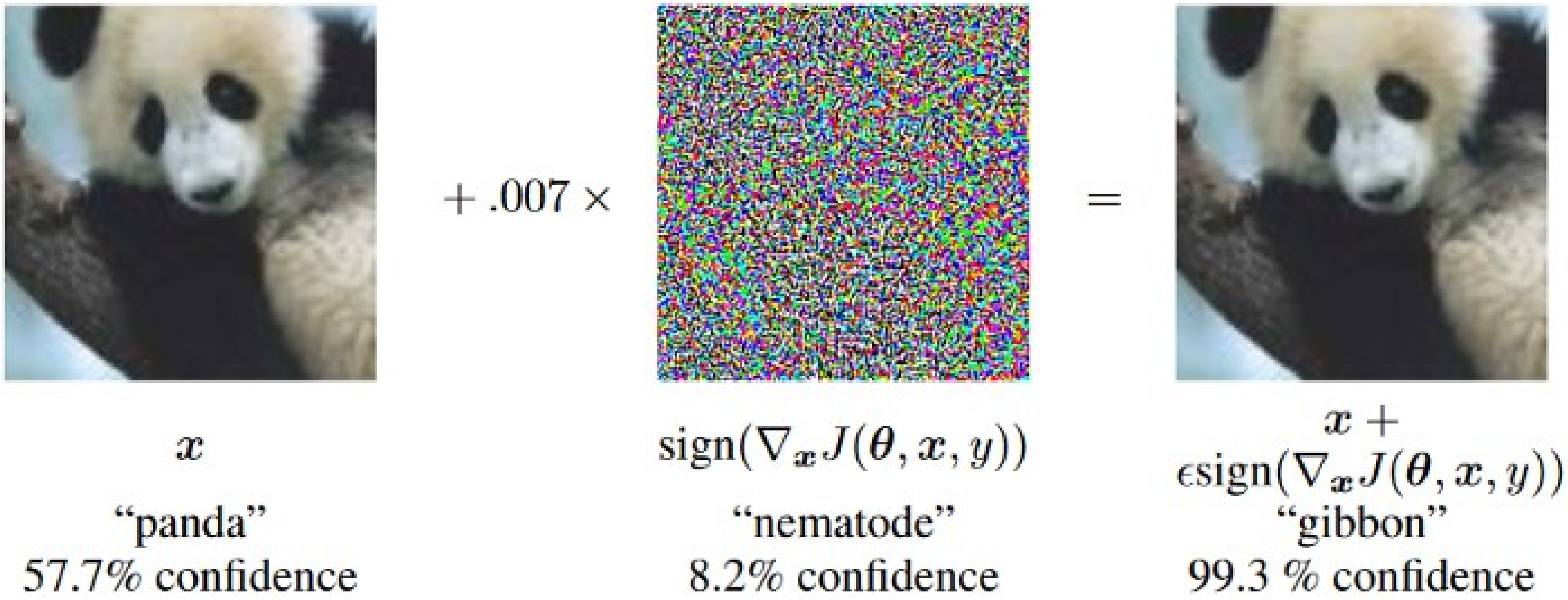
- CIFAR-10 32x32x3
- 50000 Training Images
- 10000 Testing Images

https://www.cs.toronto.edu/~kriz/cifar.html

Attacks on ML:

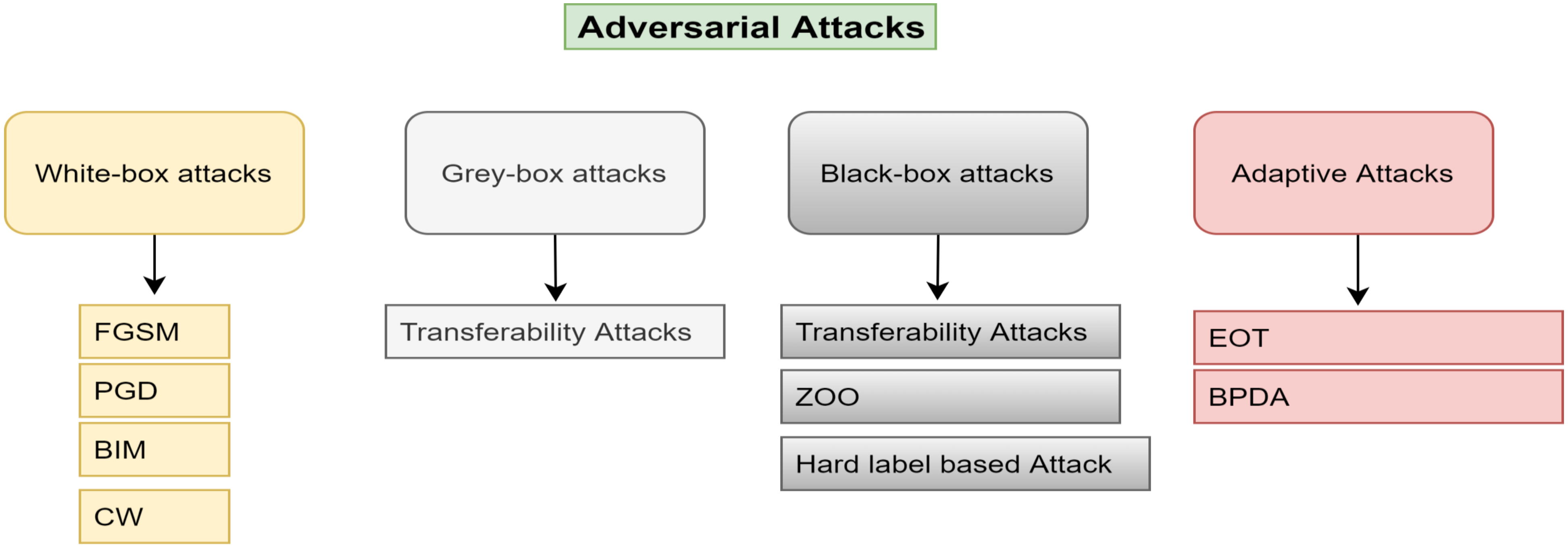


Adversarial ML: Evasion Attacks



Adapted from Goodfellow (2015)

Adversarial ML: Threat Model



- White-box Attacks: Full access (weights, dataset, learning algorithm)
- Grey-box Attacks: Partial access
- Black-box Attacks: No access
- Adaptive Attacks: attacks targeted to a specific defense

Threat Model: Adversary's Goals

- Confidence Reduction (99% cat to 12% cat)
- Misclassification (cat to any other label)
- Targeted Misclassification (cat to dog)

Threat Model: Adversarial Robustness Metrics

- Classification Error: Number of test samples misclassified
- Robust Classification Error (R): Number of perturbed test samples misclassified
- Robust Accuracy (adversarial robustness): 1-R

Definition 2 (Classification error). *Let $\mathcal{P} : \mathbb{R}^d \times \{\pm 1\} \to \mathbb{R}$ be a distribution. Then the classification error β of a classifier $f : \mathbb{R}^d \to \{\pm 1\}$ is defined as $\beta = \mathbb{P}_{(x,y)\sim\mathcal{P}}[f(x) \neq y]$.*

Next, we define our main quantity of interest, which is an adversarially robust counterpart of the above classification error. Instead of counting misclassifications under the data distribution, we allow a bounded worst-case perturbation before passing the perturbed sample to the classifier.

Definition 3 (Robust classification error). *Let $\mathcal{P} : \mathbb{R}^d \times \{\pm 1\} \to \mathbb{R}$ be a distribution and let $\mathcal{B} : \mathbb{R}^d \to \mathscr{P}(\mathbb{R}^d)$ be a perturbation set.[2] Then the $\mathcal{B}$-robust classification error β of a classifier $f : \mathbb{R}^d \to \{\pm 1\}$ is defined as $\beta = \mathbb{P}_{(x,y)\sim\mathcal{P}}[\exists\, x' \in \mathcal{B}(x) \;:\; f(x') \neq y]$.*

Since ℓ_∞-perturbations have recently received a significant amount of attention, we focus on robustness to ℓ_∞-bounded adversaries in our work. For this purpose, we define the perturbation set $\mathcal{B}_\infty^\varepsilon(x) = \{x' \in \mathbb{R}^d \mid \|x' - x\|_\infty \leq \varepsilon\}$. To simplify notation, we refer to robustness with respect to this set also as ℓ_∞^ε-robustness. As we remark in the discussion section, understanding generalization for other measures of robustness (ℓ_2, rotatations, etc.) is an important direction for future work.

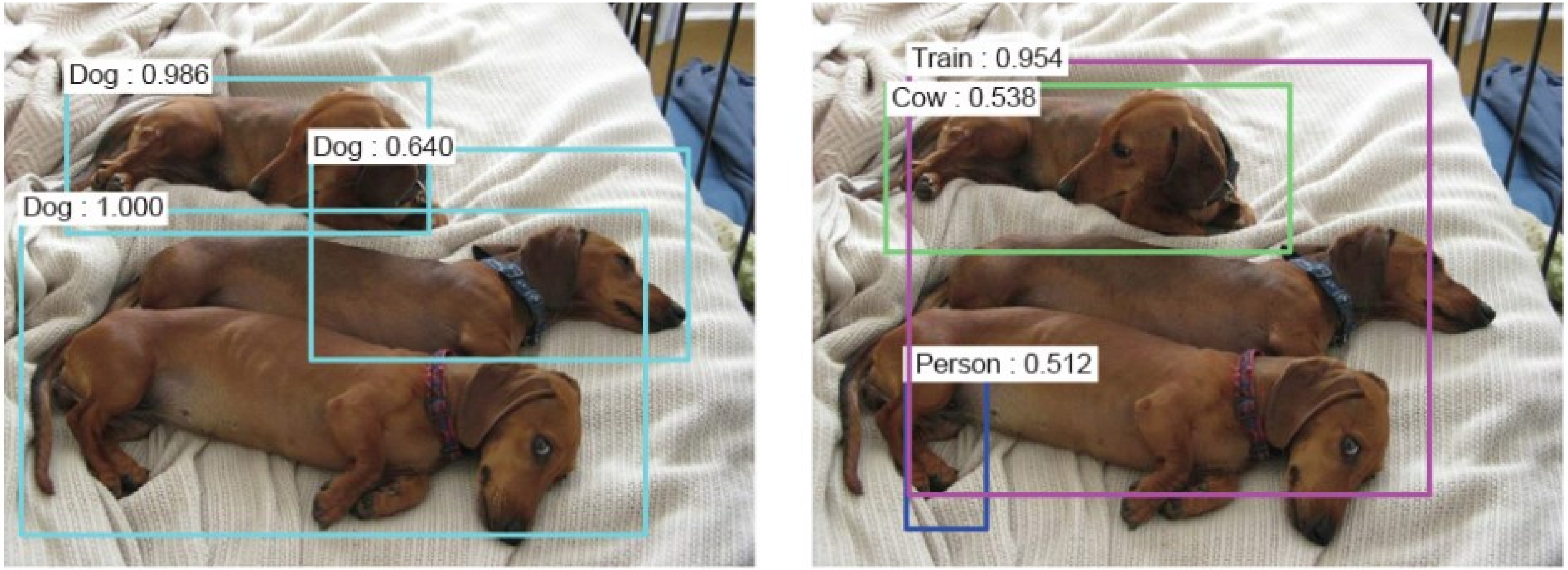
Adversarially Robust Generalization Requires More Data (Schmidt et. al 2018)

Attacks: Black-box Attack in Physical World

(a) Image from dataset (b) Clean image (c) Adv. image, $\epsilon = 4$ (d) Adv. image, $\epsilon = 8$

Adversarial Examples in Physical World (Kurakin et. Al 2015)

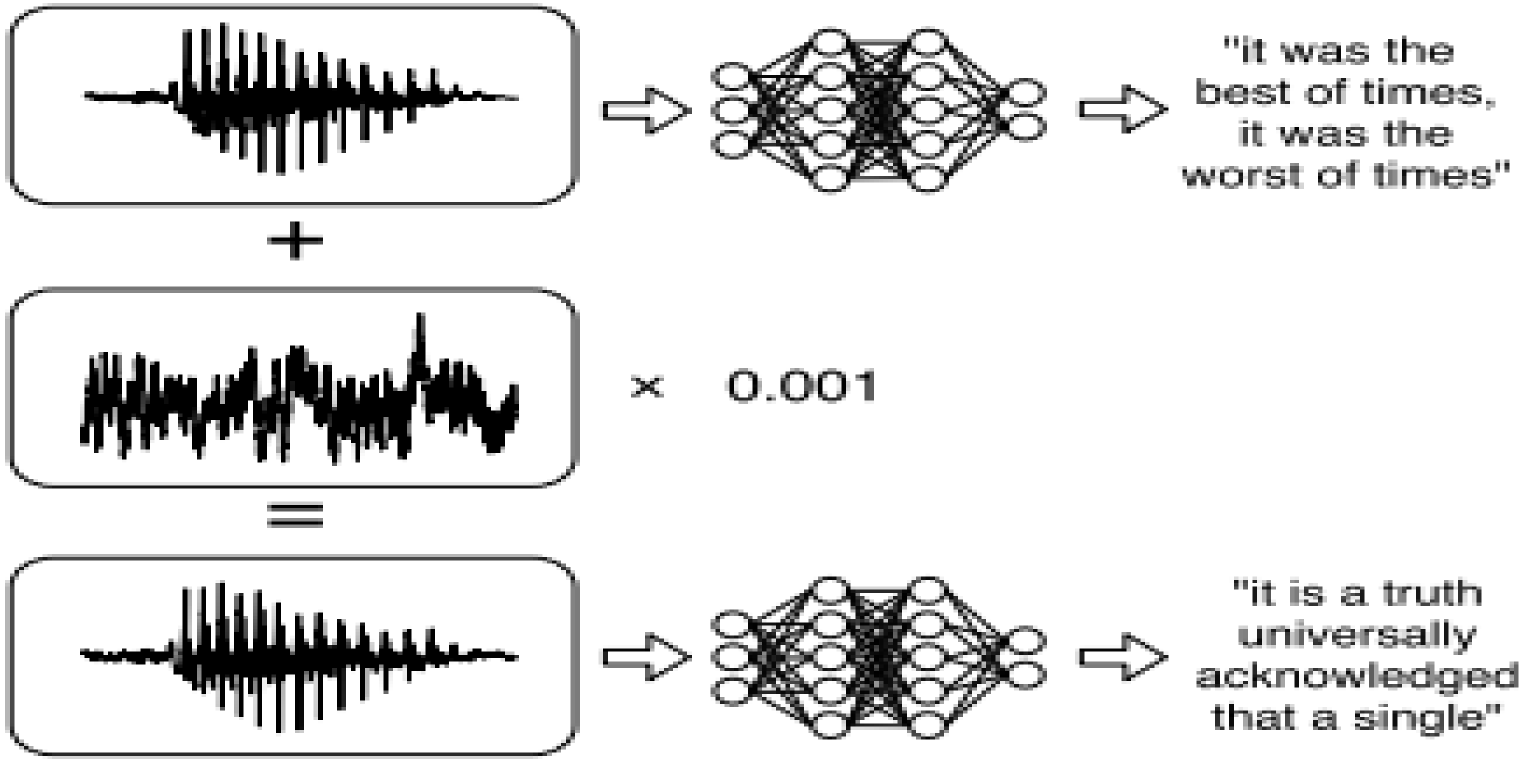
Attacks: Segmentation Task



(a) (b)

Adversarial Attacks and Defenses in Deep Learning (Ren et. al)

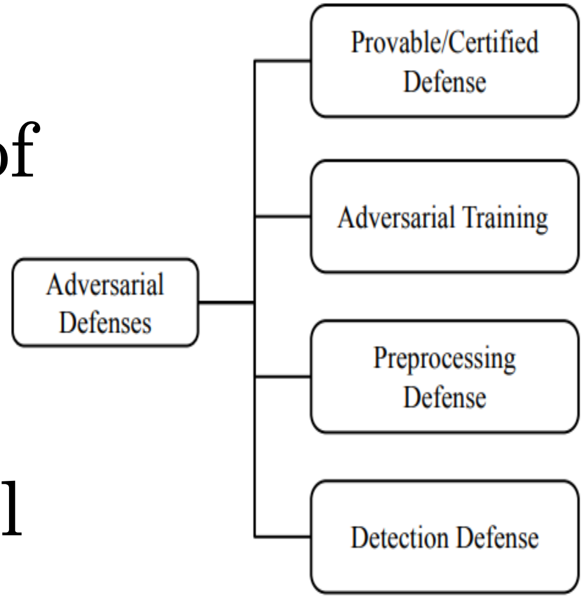
Attacks: Speech-to-Text (Audio)



Audio Adversarial Examples: Targeted Attacks on Speech-to-Text (Carlini et. al)

Adversarial Defenses:

- Certified Defenses: give a guarantee of robustness
- Input Pre-processing Defenses: apply transformations to input
- Detection Defenses: detect adversarial behaviour
- Adversarial Retraining: retrain the model on adversarial samples

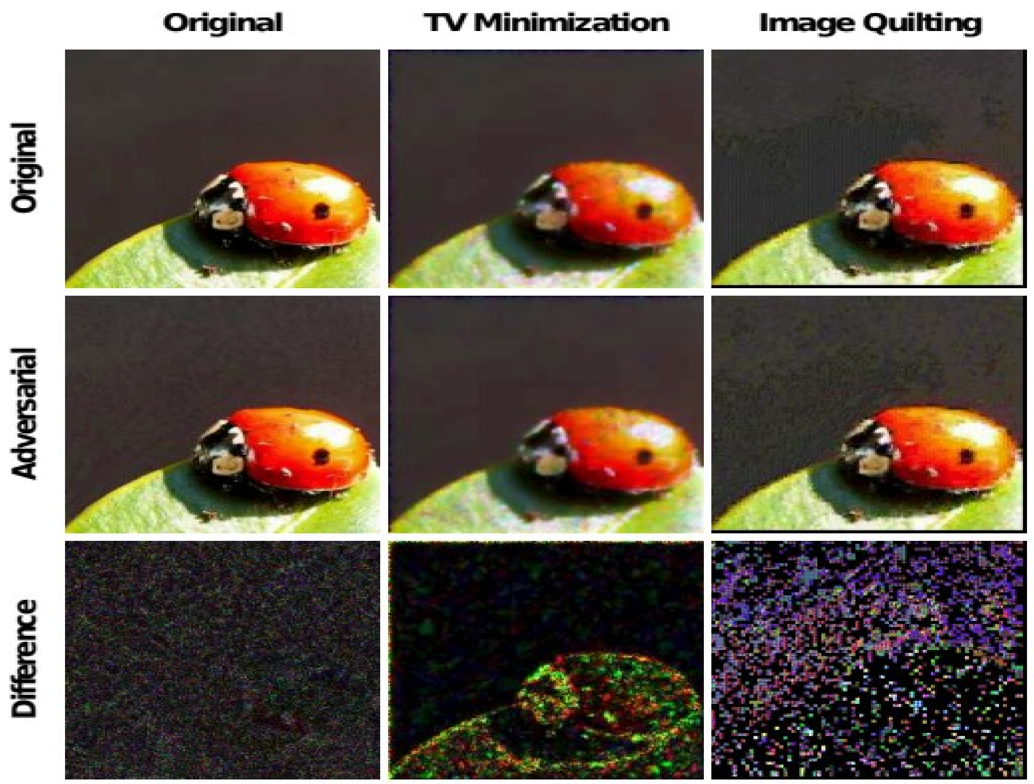


Adapted from AprilPyone (2020)

Defenses: Input Transformations

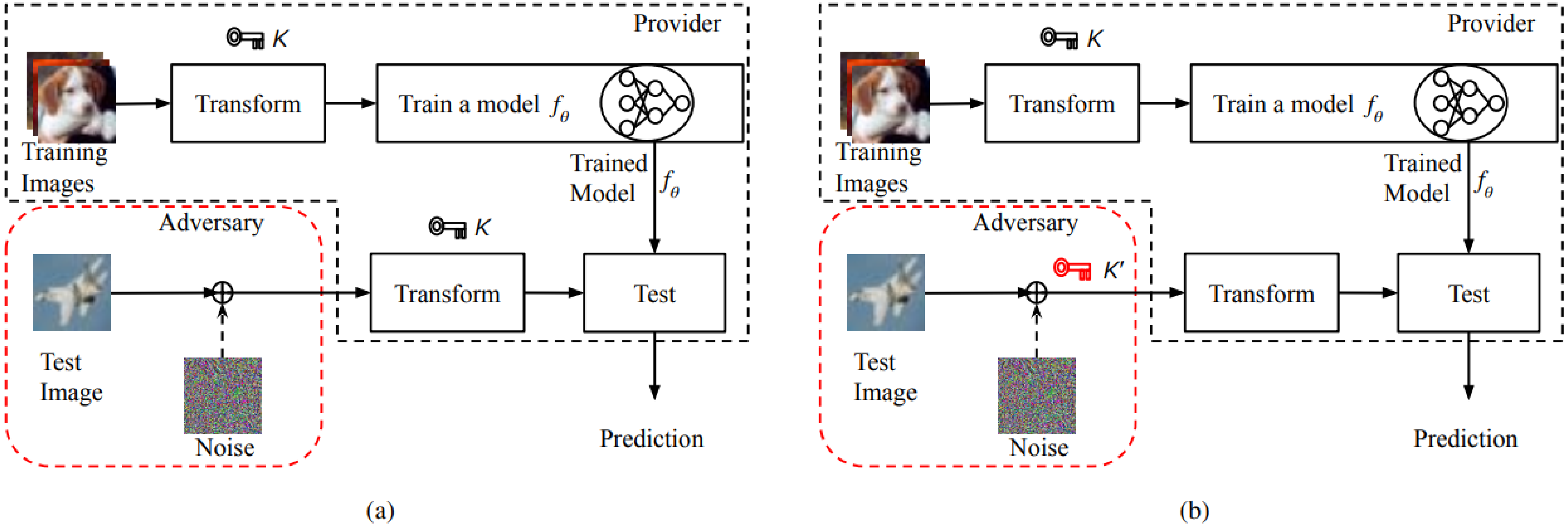
- Image Cropping and Rescaling
- Bit-Depth Reduction
- JPEG Compression
- TV minimization
- Image Quilting

- Broken with EOT and BPDA attack by (Athalye et. Al)
- Accuracy reduced to **0%!!!**



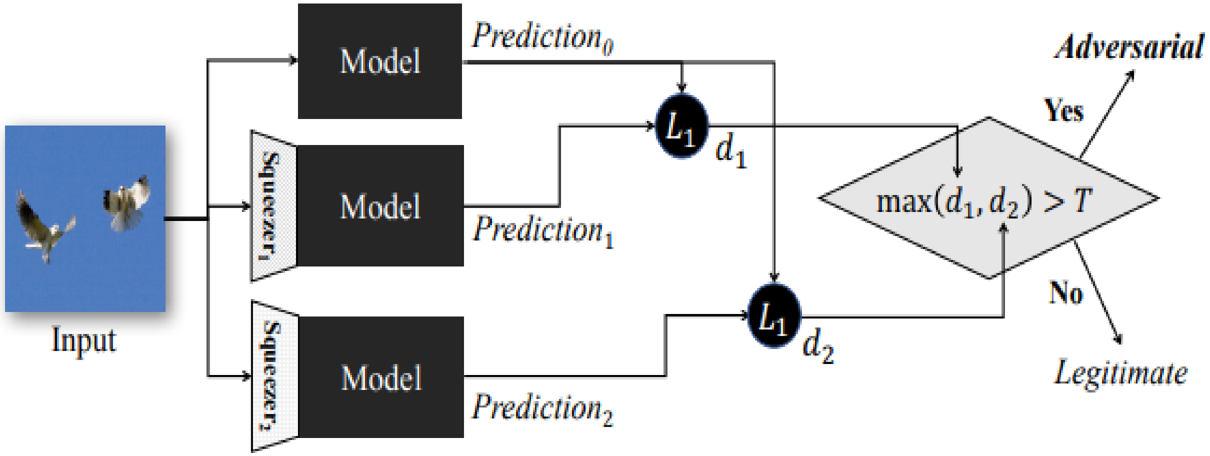
COUNTERING ADVERSARIAL IMAGES USING INPUT TRANSFORMATIONS (Guo et. al)

Defense: Key-Based Input Transformation



Block-wise Image Transformation with Secret Key for Adversarially Robust Defense (AprilPyone et. al)

Defenses: Detection

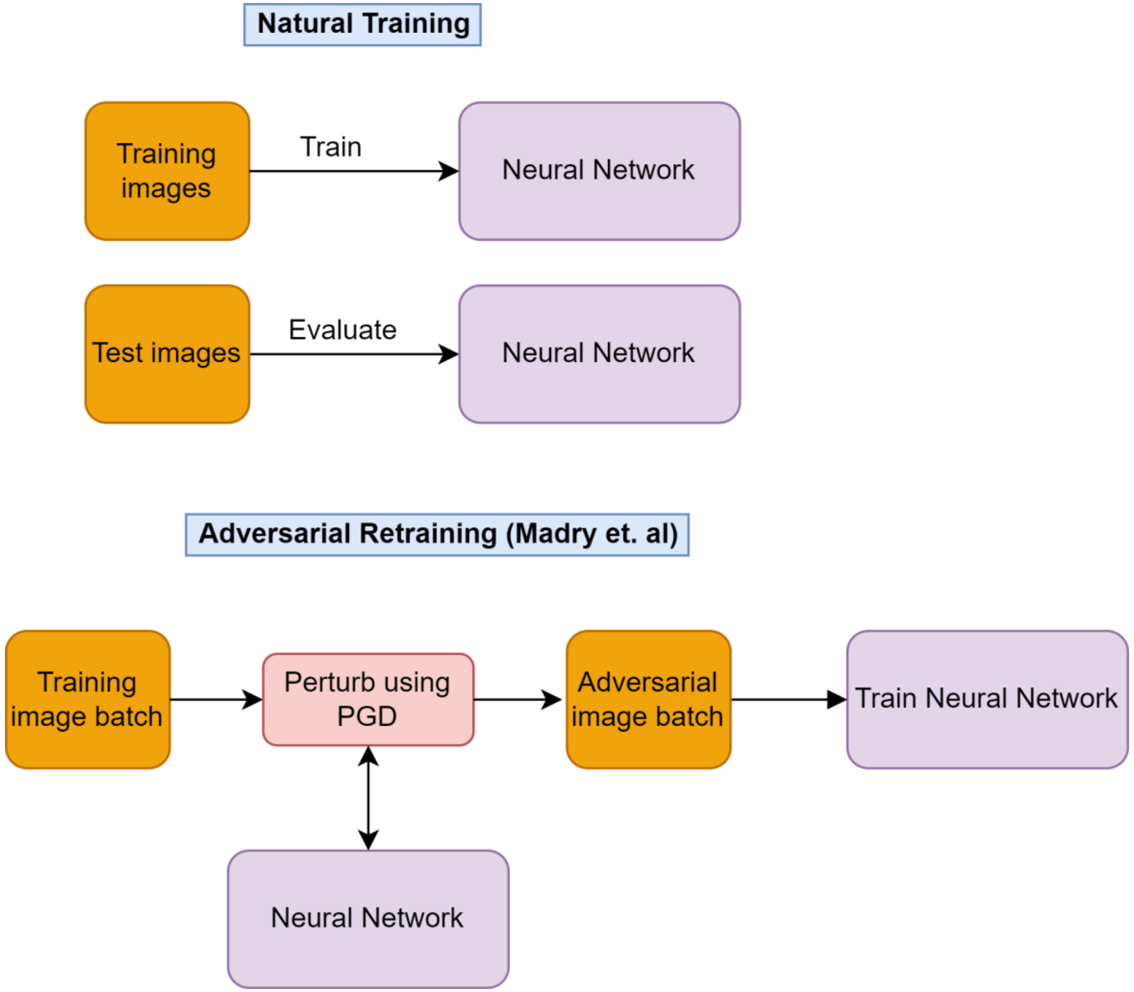


Feature Squeezing: Detecting Adversarial Examples in Deep Neural Networks (2017)

- He et. al show feature squeezing is vulnerable to adaptive attacks
- Nicholas Carlini bypassed 10 different detection methods to show they are not effective (Adversarial Examples Are Not Easily Detected: Bypassing Ten Detection Methods (2019))

Defenses: Adversarial Retraining

- Proposed by Goodfellow et. al (2015) using FGSM
- ASR fell from 89.4% to 17.8% for FGSM
- Unsuccessful against iterative attacks
- Enhanced by Madry et. al (2017) using PGD
- Defended against majority of strongest attacks (89.3% MNIST, 45.8% CIFAR-10)
- Natural accuracy drops from 95.2% to 87.3%



Adversarial Retraining: Surrogate Losses

- Logit Pairing
- Trades
- MART

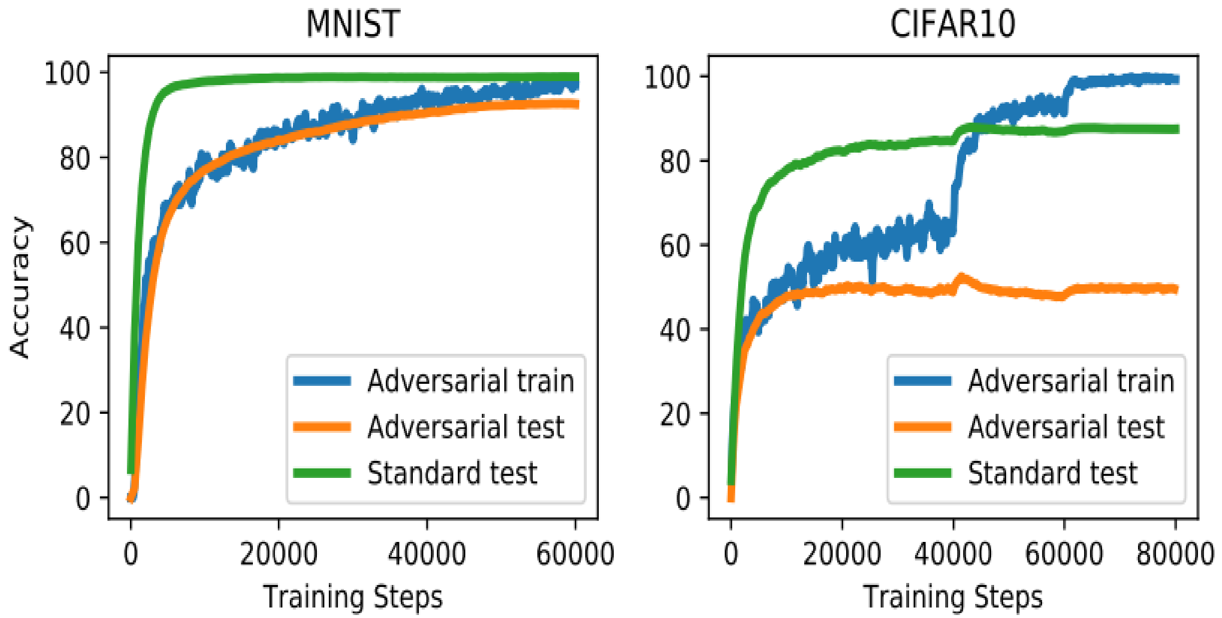
Defense Method	Loss Function
Standard	$\mathrm{CE}(\mathbf{p}(\hat{\mathbf{x}}', \boldsymbol{\theta}), y)$
ALP	$\mathrm{CE}(\mathbf{p}(\hat{\mathbf{x}}', \boldsymbol{\theta}), y) + \lambda \cdot \Vert\mathbf{p}(\hat{\mathbf{x}}', \boldsymbol{\theta}) - \mathbf{p}(\mathbf{x}, \boldsymbol{\theta})\Vert_2^2$
CLP	$\mathrm{CE}(\mathbf{p}(\mathbf{x}, \boldsymbol{\theta}), y) + \lambda \cdot \Vert\mathbf{p}(\hat{\mathbf{x}}', \boldsymbol{\theta}) - \mathbf{p}(\mathbf{x}, \boldsymbol{\theta})\Vert_2^2$
TRADES	$\mathrm{CE}(\mathbf{p}(\mathbf{x}, \boldsymbol{\theta}), y) + \lambda \cdot \mathrm{KL}\big(\mathbf{p}(\mathbf{x}, \boldsymbol{\theta}) \Vert \mathbf{p}(\hat{\mathbf{x}}', \boldsymbol{\theta})\big)$
MMA	$\mathrm{CE}(\mathbf{p}(\hat{\mathbf{x}}', \boldsymbol{\theta}), y) \cdot \mathbb{1}(h_{\boldsymbol{\theta}}(\mathbf{x}) = y) + \mathrm{CE}(\mathbf{p}(\mathbf{x}, \boldsymbol{\theta}), y) \cdot \mathbb{1}(h_{\boldsymbol{\theta}}(\mathbf{x}) \neq y)$
MART	$\mathrm{BCE}(\mathbf{p}(\hat{\mathbf{x}}', \boldsymbol{\theta}), y) + \lambda \cdot \mathrm{KL}(\mathbf{p}(\mathbf{x}, \boldsymbol{\theta}) \Vert \mathbf{p}(\hat{\mathbf{x}}', \boldsymbol{\theta})) \cdot (1 - \mathbf{p}_y(\mathbf{x}, \boldsymbol{\theta}))$

Defense	MNIST				CIFAR-10			
	Natural	FGSM	PGD^{20}	CW_∞	Natural	FGSM	PGD^{20}	CW_∞
Standard	99.11	97.17	94.62	94.25	84.44	61.89	47.55	45.98
MMA	98.92	97.25	95.25	94.77	**84.76**	62.08	48.33	45.77
Dynamic	98.96	97.34	95.27	94.85	83.33	62.47	49.40	46.94
TRADES	**99.25**	96.67	94.58	94.03	82.90	62.82	50.25	48.29
MART	98.74	**97.87**	**96.48**	**96.10**	83.07	**65.65**	**55.57**	**54.87**

IMPROVING ADVERSARIAL ROBUSTNESS REQUIRES REVISITING MISCLASSIFIED EXAMPLES (Wang et. al)

Defenses: Robust generalization requires more data

- MNIST achieves >90% robustness
- Owing to learning thresholding filters
- CIFAR-10 achieves >45% robustness
- Gap between standard & robust generalization higher on CIFAR-10
- Owing to high dimensions

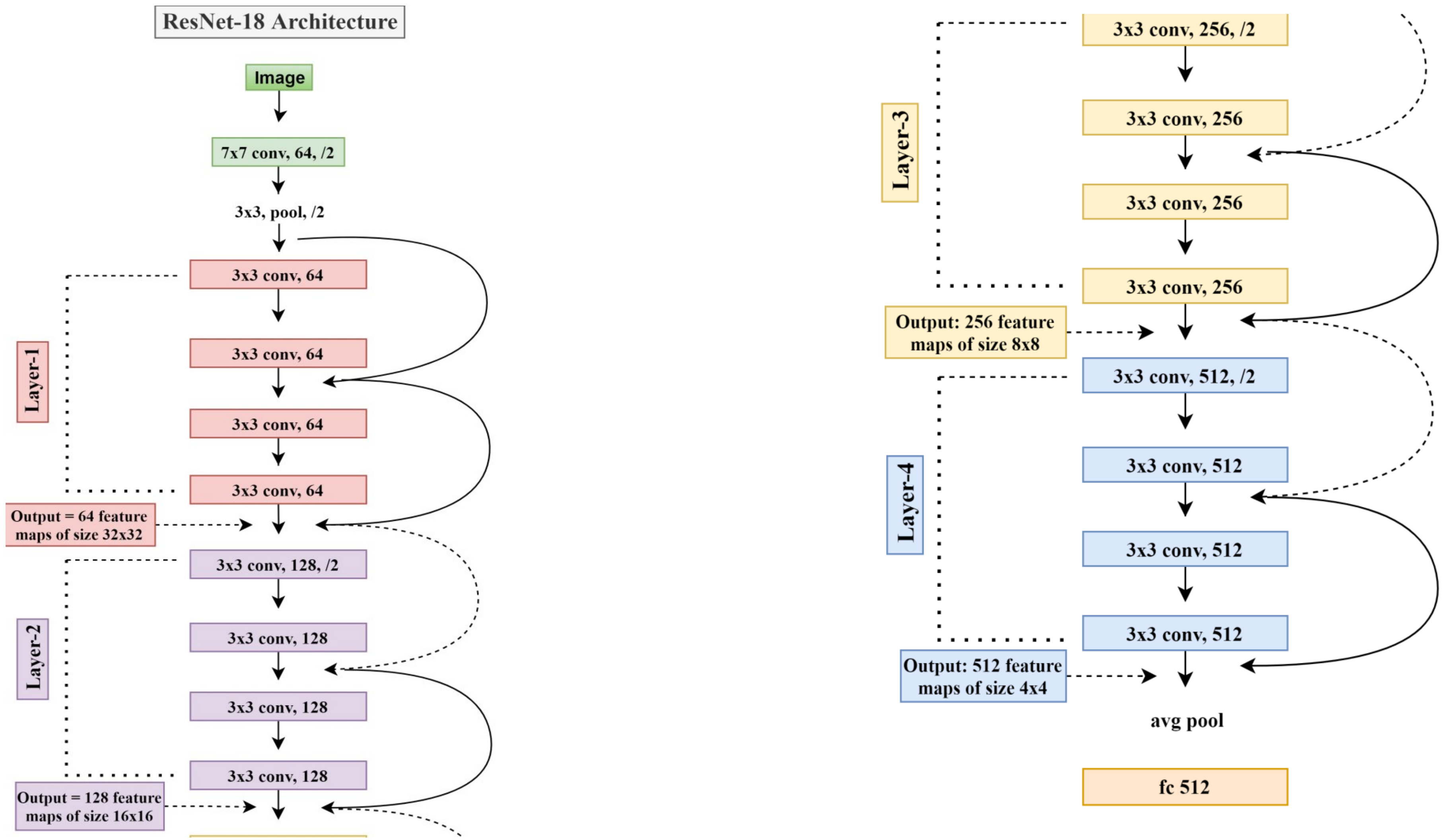


Adversarially Robust Generalization Requires More Data (Schmidt et. al)

Defenses: Data Augmentation & Unlabeled Extra Data

- Carmon et. al use **500k** unlabeled extra data
- Using extra data jumps robustness to **59%**
- Rebuffi et. Al use data augmentations (CutMix)
- Achieving **66.56%** robustness with **90.51%** standard accuracy

Defenses: Effect of architecture on robustness



Defenses: Effect of architecture on robustness

	Robust Accuracy (%)			Natural Accuracy (%)			Perturbation Stability (%)		
λ	width-1	width-5	width-10	width-1	width-5	width-10	width-1	width-5	width-10
				TRADES Zhang et al. (2019)					
6	47.81±.09	54.45±.16	54.18±.39	**76.26±.10**	**84.44±.06**	**84.90±.80**	69.33±.05	68.27±.22	67.25±.39
9	**48.01±.06**	55.34±.17	55.29±.45	73.78±.30	82.77±.07	84.13±.28	71.92±.33	70.66±.26	69.08±.80
12	47.87±.06	**55.61±.04**	55.98±.13	72.29±.25	81.59±.20	83.59±.62	73.33±.16	72.00±.20	70.18±.67
15	47.15±.13	55.49±.15	55.96±.09	70.98±.24	80.69±.08	82.81±.19	73.79±.27	72.87±.03	70.87±.23
18	47.02±.13	55.43±.12	**56.43±.17**	70.13±.06	79.97±.12	82.21±.21	74.63±.11	73.77±.13	72.04±.30
21	46.26±.19	55.31±.20	56.07±.21	68.95±.38	79.25±.23	81.74±.12	**75.17±.28**	**74.15±.38**	**72.11±.12**
				Adversarial Training Madry et al. (2018)					
1.00	47.99±.16	50.87±.42	50.12±.13	**77.30±.01**	**85.82±.01**	**85.62±.81**	66.48±.24	62.23±.42	61.62±.46
1.25	**49.24±.12**	53.10±.09	51.97±.46	74.04±.47	84.73±.22	86.25±.12	70.34±.54	65.24±.08	62.94±.35
1.50	49.11±.03	54.15±.03	53.25±.52	72.16±.25	84.35±.19	85.50±.57	72.10±.11	66.65±.06	64.51±.72
1.75	48.32±.63	**54.36±.14**	53.65±.80	70.66±.46	83.95±.30	85.52±.24	72.43±.40	67.31±.03	65.67±.10
2.00	47.44±.06	54.10±.15	**55.78±.22**	69.67±.09	83.49±.06	85.41±.13	**72.73±.04**	**67.53±.01**	**65.71±.15**

Wide residual networks. (Zagoruyko et. al 2017)

Defenses: RobustBench (CIFAR-10)

RobustBench   Leaderboards   Paper   FAQ   Contribute   Model Zoo

Leaderboard: CIFAR-10, $\ell_\infty = 8/255$, untargeted attack

Show 15 entries   Search: Papers, architectures, ve

Rank	Method	Standard accuracy	AutoAttack robust accuracy	Best known robust accuracy	AA eval. potentially unreliable	Extra data	Architecture	Venue
1	Fixing Data Augmentation to Improve Adversarial Robustness *66.56% robust accuracy is due to the original evaluation (AutoAttack + MultiTargeted)*	92.23%	66.58%	66.56%	×	☑	WideResNet-70-16	arXiv, Mar 2021
2	Improving Robustness using Generated Data *It uses additional 100M synthetic images in training. 66.10% robust accuracy is due to the original evaluation (AutoAttack + MultiTargeted)*	88.74%	66.11%	66.10%	×	×	WideResNet-70-16	NeurIPS 2021
3	Uncovering the Limits of Adversarial Training against Norm-Bounded Adversarial Examples *65.87% robust accuracy is due to the original evaluation (AutoAttack + MultiTargeted)*	91.10%	65.88%	65.87%	×	☑	WideResNet-70-16	arXiv, Oct 2020

https://robustbench.github.io/#leaderboard

Defenses: RobustBench (ImageNet)

ROBUSTBENCH Leaderboards Paper FAQ Contribute Model Zoo 🚀

Leaderboard: ImageNet, $\ell_\infty = 4/255$, untargeted attack

Show 15 entries Search: Papers, architectures, ve

Rank	Method	Standard accuracy	AutoAttack robust accuracy	Best known robust accuracy	AA eval. potentially unreliable	Extra data	Architecture	Venue
1	Do Adversarially Robust ImageNet Models Transfer Better?	68.46%	38.14%	38.14%	×	×	WideResNet-50-2	NeurIPS 2020
2	Do Adversarially Robust ImageNet Models Transfer Better?	64.02%	34.96%	34.96%	×	×	ResNet-50	NeurIPS 2020
3	Robustness library	62.56%	29.22%	29.22%	×	×	ResNet-50	GitHub, Oct 2019
4	Fast is better than free: Revisiting adversarial training *Focuses on fast adversarial training.*	55.62%	26.24%	26.24%	×	×	ResNet-50	ICLR 2020

Conclusion

- A lot of room for improvement
- Possible future work
- Our current work evaluates secret key based defeses and tries to improve robustness by making changes to the architecture