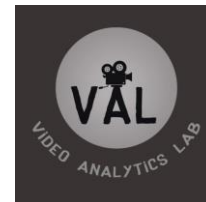


Tutorial on Adversarial Robustness of Deep Neural Networks

Sravanti Addepalli
PhD Student

Department of Computational and Data Sciences
Indian Institute of Science, Bangalore



Overview

Module-1 : Adversarial Attacks

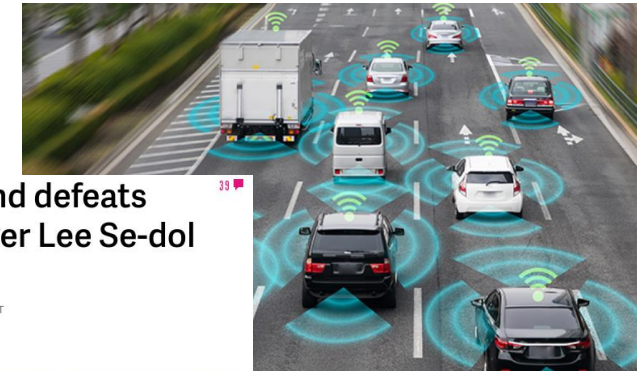
What are adversarial attacks, Threat model, Crafting adversarial attacks
Classification of adversarial attacks, Some examples of adversarial attacks, Code snippet

Module-2 : Defending against Adversarial Attacks

Motivation for adversarial defense research, Adversarial Training (PGD-AT, TRADES, AWP, SOTA tricks), Robust evaluation of Adversarial Defenses (Auto-Attack)

Deep Learning Applications

- Autonomous navigation systems
- Surveillance systems
- Medicine and health care
- Reinforcement learning
- Generative modelling
- Style transfer
- Robotics
- Speech Processing
- Natural Language Processing



AI beats docs in cancer spotting

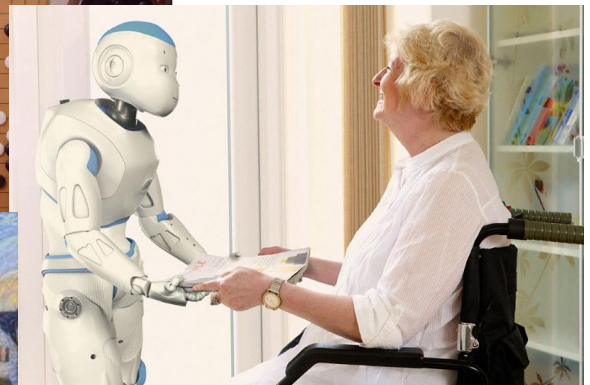
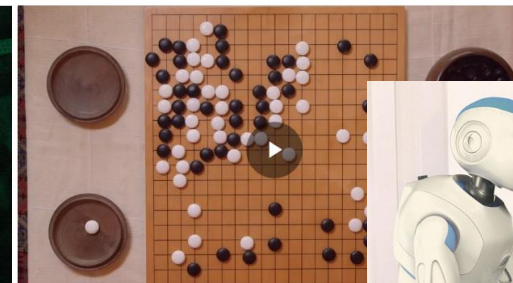
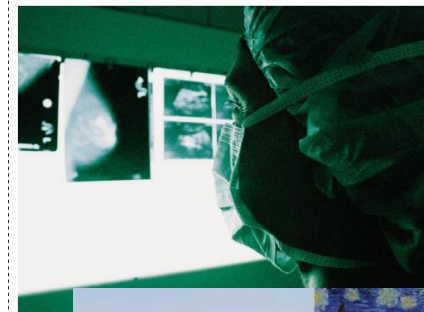
A new study provides a fresh example of machine learning as an important diagnostic tool. Paul Biegler reports.



SHARE



TWEET



Adversarial Attacks



Prediction: **Hamster**

Confidence = 99.99%

+ 0.02 *



50-step PGD targeted
attack with $\epsilon = \frac{8}{255}$
scaled by 50x

=



Prediction: **Banjo**

Confidence = 100%

Adversarial Attacks

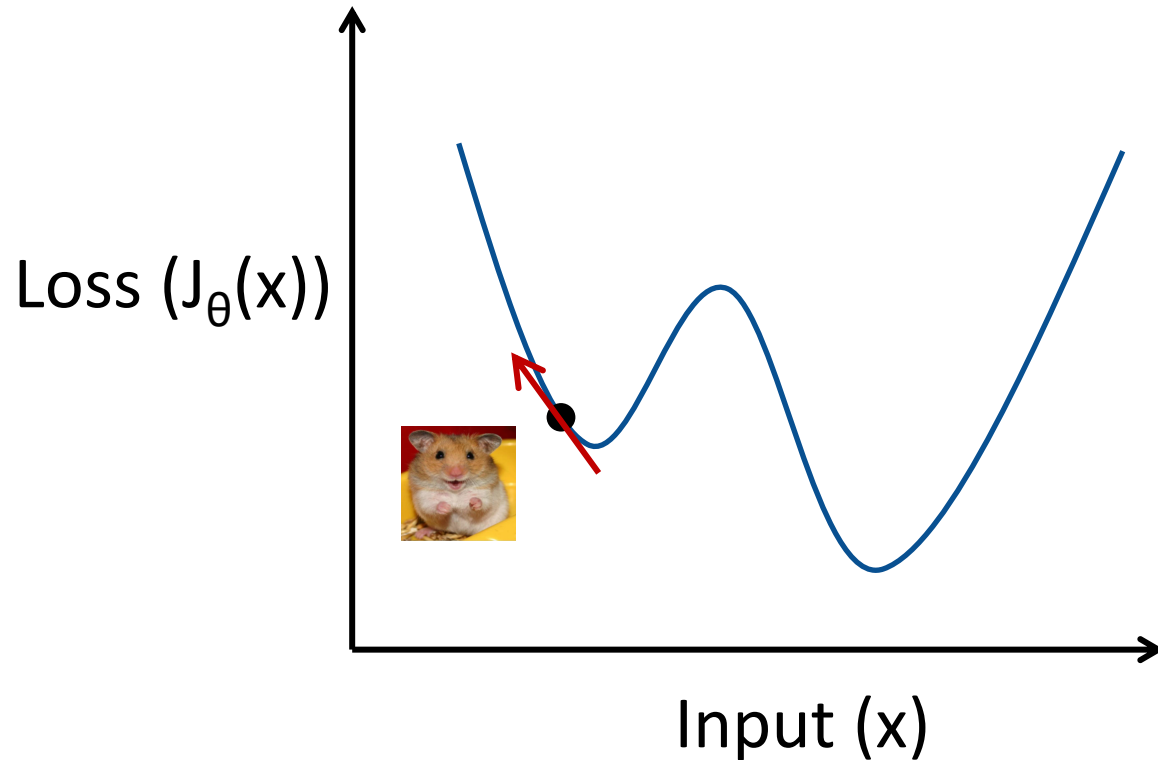
- Definition of adversarial examples
 - An input to a machine learning model that is intentionally designed by an attacker to fool the model into producing an incorrect output ¹
 - Does not cause a change in human prediction
- Threat model ²:
 - Specifies the conditions under which the attack is valid
 - Mathematical proxy to ensure imperceptibility
 - Includes assumptions about the adversary's goals, knowledge and capabilities
 - Common threat models for constructing adversarial examples
 - ℓ_p -norm constrained threat model
 - Patch attacks

Ian Goodfellow and Nicolas Papernot. Is attacking machine learning easier than defending it? Blog post on Feb 15, 2017.

Carlini et al., On evaluating adversarial robustness. arXiv preprint arXiv:1902.06705, 2019.

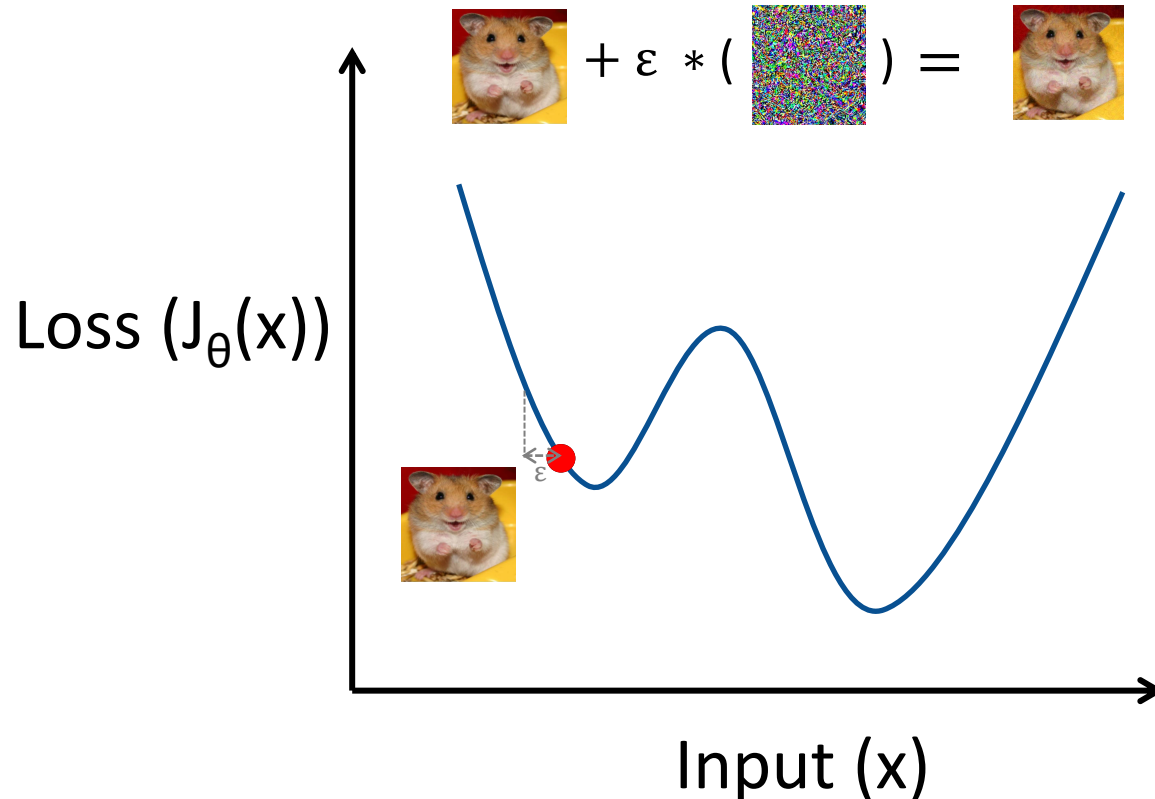
Crafting Adversarial Attacks

Gradient Ascent on the Loss



Crafting Adversarial Attacks

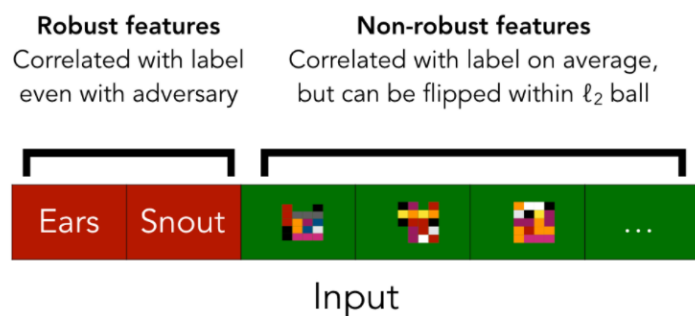
Gradient Ascent on the Loss



- Losses used
 - Cross-Entropy Loss
 - Maximum-margin loss
- Noise added before attack
- Number of iterations for attack generation
 - Single-step
 - Multi-step
- Attack Step-size
- Number of restarts

Existence of Adversarial Examples

- High dimensional nature of input space
 - Mild perturbations in each dimension can cause large changes at the output
- Reliance of models on non-robust features for prediction



- Finite size of training data distribution
 - Sample complexity of robust learning can be significantly larger than that of "standard" learning

Gilmer, Justin, et al. "Adversarial spheres." *arXiv preprint arXiv:1801.02774* (2018).

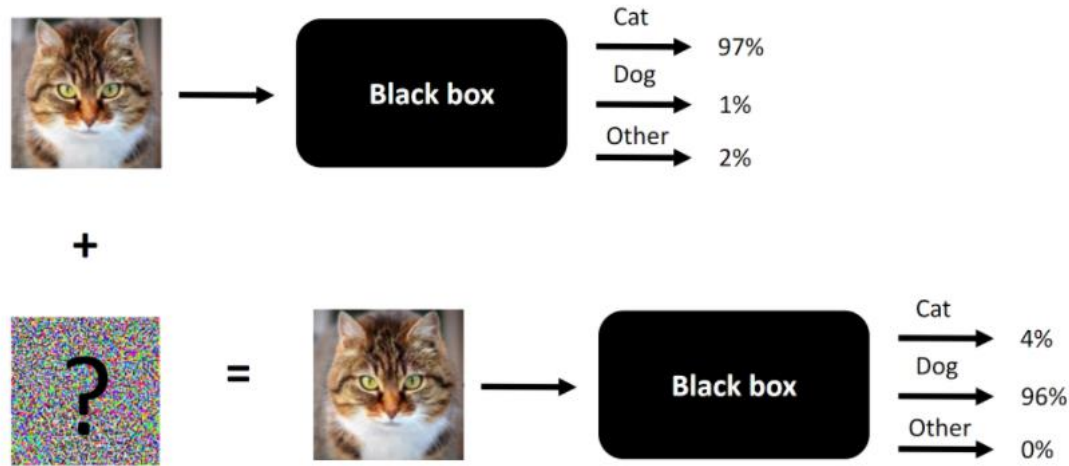
Ilyas, Andrew, et al. "Adversarial examples are not bugs, they are features." NeurIPS 2019

Schmidt, Ludwig, et al. "Adversarially robust generalization requires more data." *arXiv preprint arXiv:1804.11285* (2018).

Classification of Adversarial Attacks

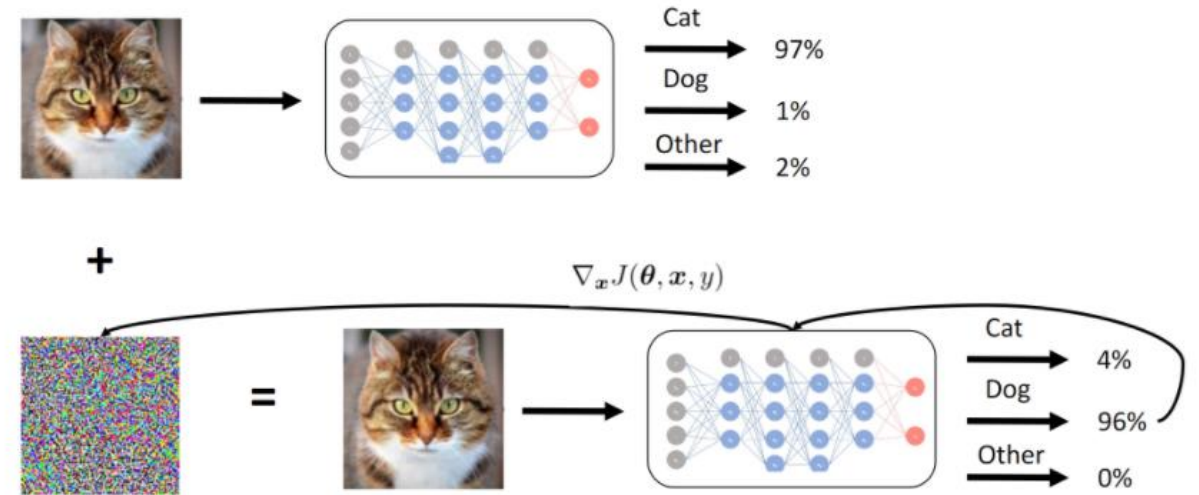
1. Based on Knowledge of the adversary

Black-Box Attacks



- x No access to network architecture
- x No access to network parameters
- x No access to the training algorithm
- ✓ Varying degrees of access to the network output (limited queries, access to predicted probabilities/ predicted class, access to the training data)

White-Box Attacks

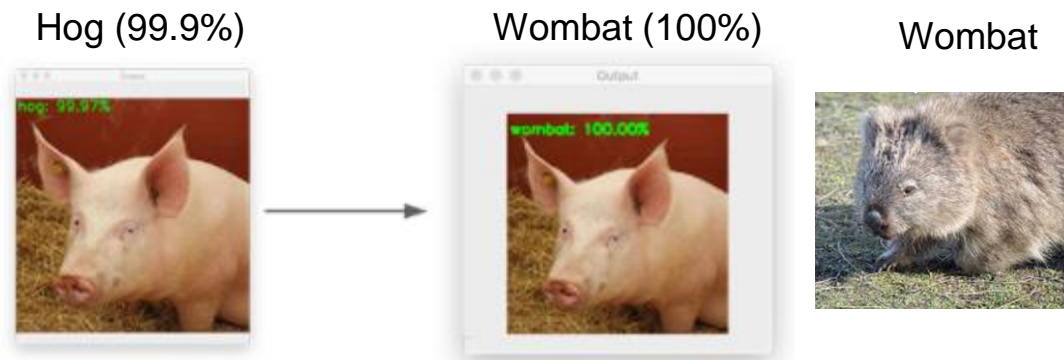


- ✓ Access to network architecture
- ✓ Access to network parameters
- ✓ Access to the training algorithm
- ✓ Access to output and intermediate representations

Classification of Adversarial Attacks

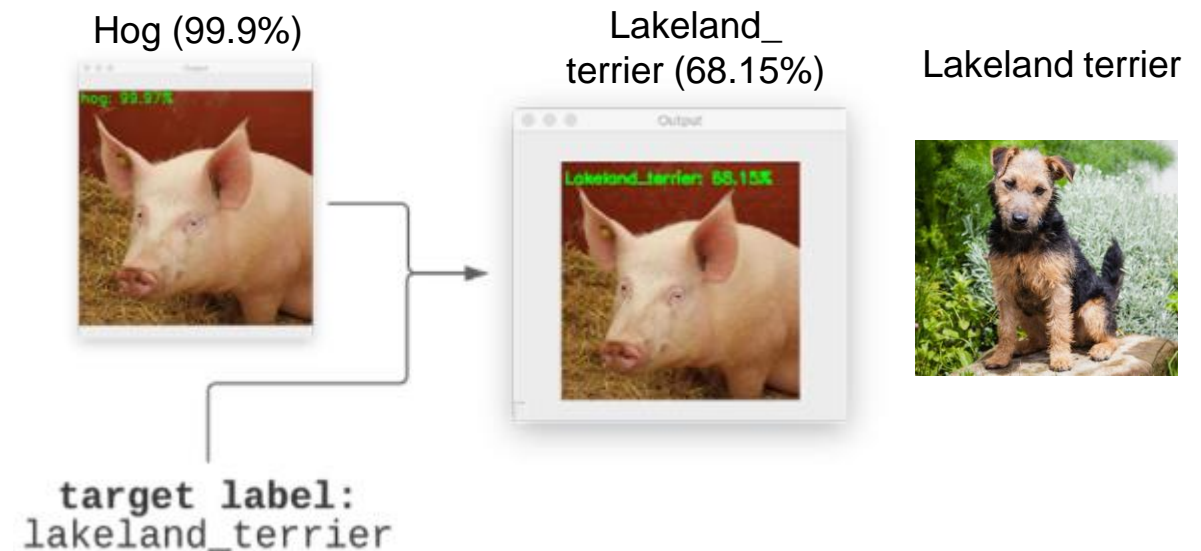
2. Based on goals of the adversary

Untargeted attacks



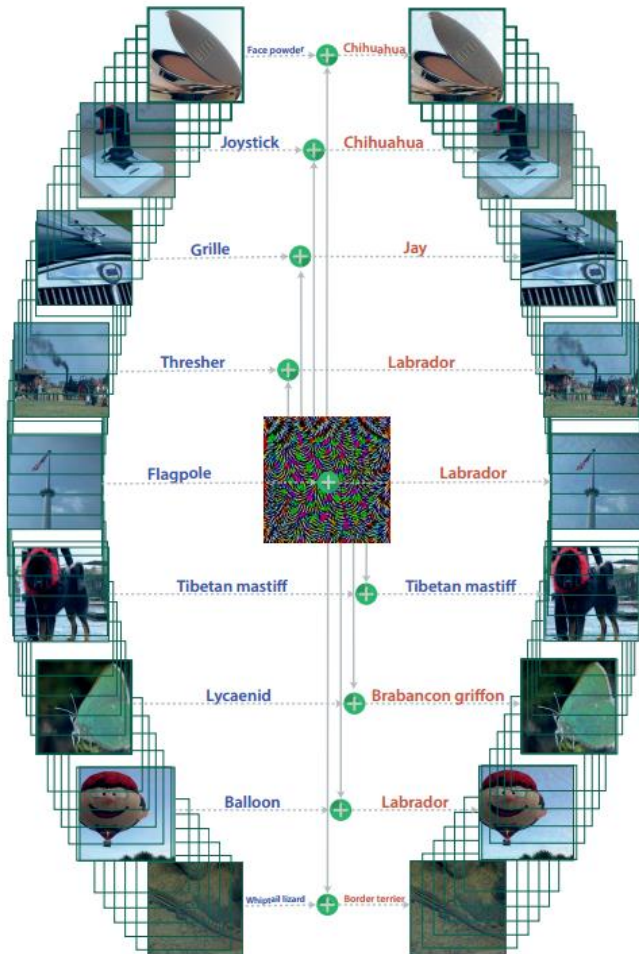
- Maximization of Cross-entropy loss (or a related classification loss)
- Goal is to cause a misclassification to any class

Targeted attacks



- Minimization of Cross-entropy loss with respect to a specific target class
- Goal is to cause the model to predict a specific class

Universal Adversarial Perturbations (UAPs)



Generalizability of the universal perturbations across different networks
(Percentages indicate the fooling rates)

	VGG-F	CaffeNet	GoogLeNet	VGG-16	VGG-19	ResNet-152
VGG-F	93.7%	71.8%	48.4%	42.1%	42.1%	47.4 %
CaffeNet	74.0%	93.3%	47.7%	39.9%	39.9%	48.0%
GoogLeNet	46.2%	43.8%	78.9%	39.2%	39.8%	45.5%
VGG-16	63.4%	55.8%	56.5%	78.3%	73.1%	63.4%
VGG-19	64.0%	57.2%	53.6%	73.5%	77.8%	58.0%
ResNet-152	46.3%	46.3%	50.5%	47.0%	45.5%	84.0%

- One perturbation to fool all images and models
- UAPs are weaker than image-specific adversarial attacks
- Shows transferability across different architectures

ℓ_p -norm based threat model



+



=



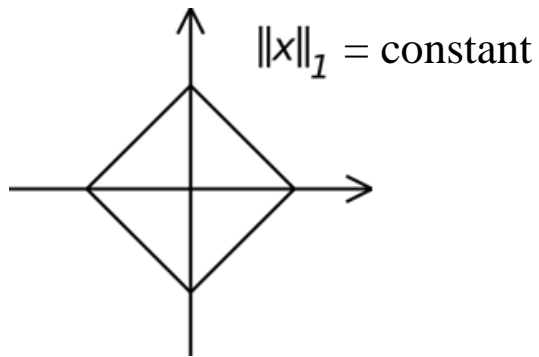
$$\|\mathbf{x}\|_p := \left(\sum_{i=1}^n |x_i|^p \right)^{1/p} \leq \varepsilon$$

The value of p and ε define the threat model

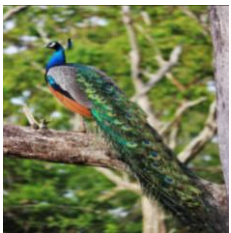
Classification of Adversarial Attacks

3. Based on constraints imposed on the perturbations (Common ℓ_p -norm bound attacks)

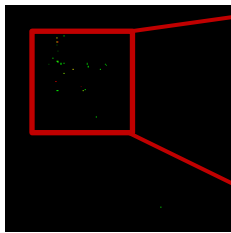
ℓ_1 -norm bound attack



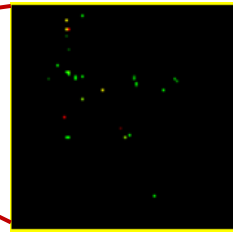
$$\|\mathbf{x}\|_1 := \sum_{i=1}^n |x_i|$$



Clean image

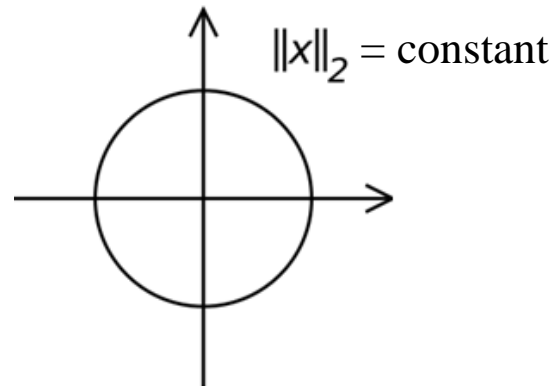


Scaled by 50x

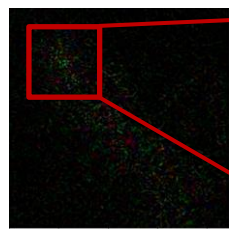


Scaled by 50x
Zoomed in

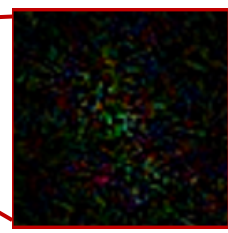
ℓ_2 -norm bound attack



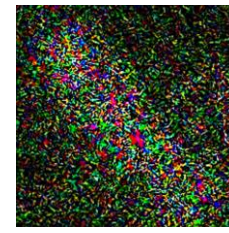
$$\|\mathbf{x}\|_2 := \sqrt{x_1^2 + \cdots + x_n^2}$$



Scaled by 50x

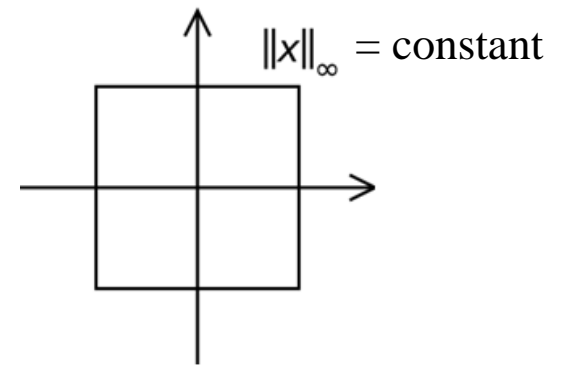


Scaled by 50x
Zoomed in

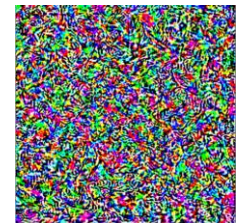


Scaled by 500x

ℓ_∞ -norm bound attack



$$\|\mathbf{x}\|_\infty := \max_i |x_i|$$



Scaled by 50x

ℓ_p -norm based Adversarial examples



ℓ_1 -norm with $\epsilon = 12$



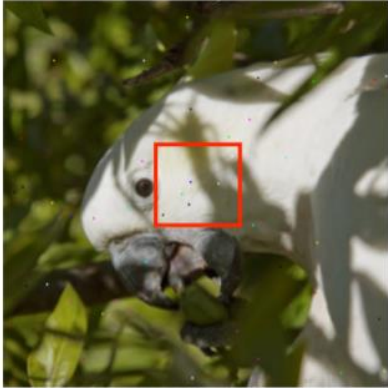





ℓ_2 -norm with $\epsilon = 0.5$



ℓ_∞ -norm with
 $\epsilon = 0.031$ (8/255)

Classification of Adversarial Attacks

3. Based on constraints imposed on the perturbations (*Sparse Attacks*)

	ℓ_0 Attack	Patch Attack	Adversarial Frame
Adversarial Examples	<p>Parrot → Turtle</p> 	<p>Elephant → Jeep</p> 	<p>Lynx → Website</p> 
Adversarial Examples (Zoomed into the red square of above row)			

Source: Croce, Francesco, et al. "Sparse-RS: a versatile framework for query-efficient sparse black-box adversarial attacks." *arXiv preprint arXiv:2006.12834* (2020).

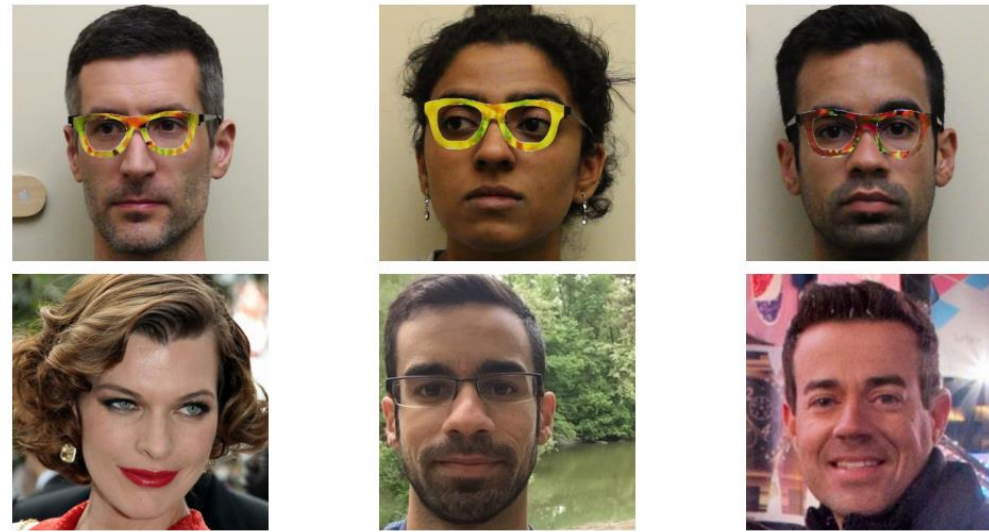
Adversarial Attacks in the real world

Hackers can trick a Tesla into accelerating by 50 miles per hour

A two inch piece of tape fooled the Tesla's cameras and made the car quickly and mistakenly speed up.



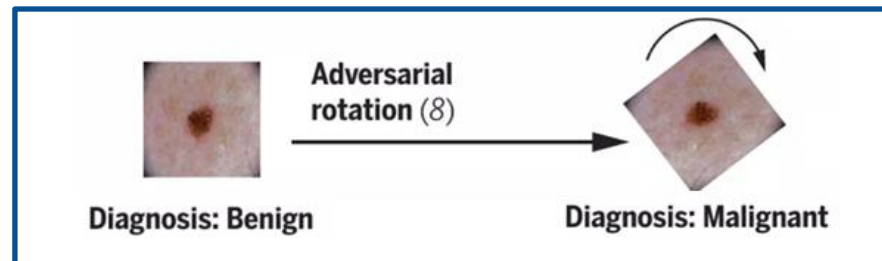
<https://www.technologyreview.com/2020/02/19/868188/hackers-can-trick-a-tesla-into-accelerating-by-50-miles-per-hour/>



Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition, M Sharif, S Bhagavatula, L Bauer, MK Reiter, ACM SIGSAC16



Eykholt et al. "Robust physical-world attacks on deep learning visual classification." CVPR 2018.



<https://www.vox.com/future-perfect/2019/4/8/18297410/ai-tesla-self-driving-cars-adversarial-machine-learning>

Adversarial attacks constrained within ℓ_∞ norm

- Fast Gradient Sign Method (FGSM)

$$x^* = x + \epsilon \cdot \text{sign}(\nabla_x J(f(x; \theta), y_{true}))$$

- Random + Fast Gradient Sign Method (RFGSM)

$$x' = x + \alpha \cdot \text{sign}(\mathcal{N}(0^d, I^d))$$

$$x^* = x' + (\epsilon - \alpha) \cdot \text{sign}(\nabla_{x'} J(f(x'; \theta), y_{true}))$$

- Projected Gradient Descent (PGD)

$$x^0 = x + \mathcal{U}(-\epsilon_{step}, \epsilon_{step}, \text{shape}(x))$$

$$x^{N+1} = x^N + \epsilon_{step} \cdot \text{sign}(\nabla_{x^N} J(f(x^N; \theta), y_{true}))$$

$$x^{N+1} = \text{clip}(x^{N+1}, \min = x - \epsilon, \max = x + \epsilon)$$

Summary of Module-1

- Adversarial attacks can cause misclassification in Deep Networks
- Types of adversarial attacks:
 - Based on Knowledge of the adversary
 - Black-box and White-box attacks
 - Based on Goals of the adversary
 - Untargeted and Targeted attacks
 - Special case: Universal Adversarial Perturbations (UAPs)
 - Based on constraints imposed on the perturbations
 - ℓ_p -norm bound attacks ($\ell_1, \ell_2, \ell_\infty$)
 - Sparse attacks (ℓ_0 -norm bound attacks, patch attacks, adversarial frames)
- Adversarial attacks constrained within ℓ_∞ norm
 - FGSM, R-FGSM, PGD
- Generation of PGD attack

Overview

Module-1 : Adversarial Attacks

What are adversarial attacks, Threat model, Crafting adversarial attacks
Classification of adversarial attacks, Some examples of adversarial attacks, Code snippet

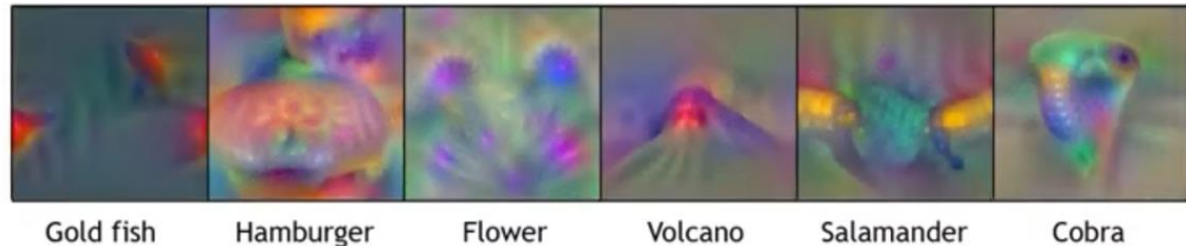
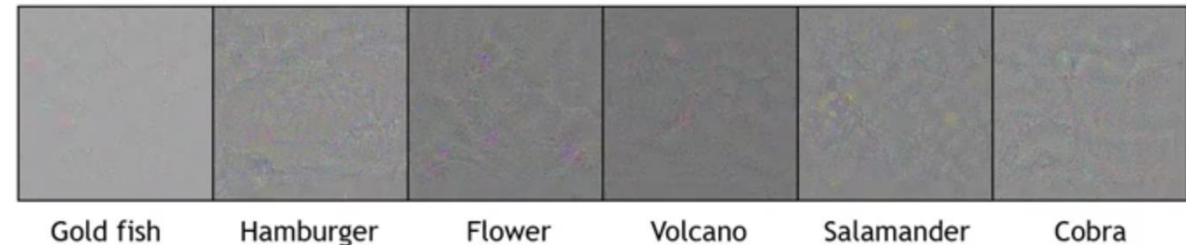
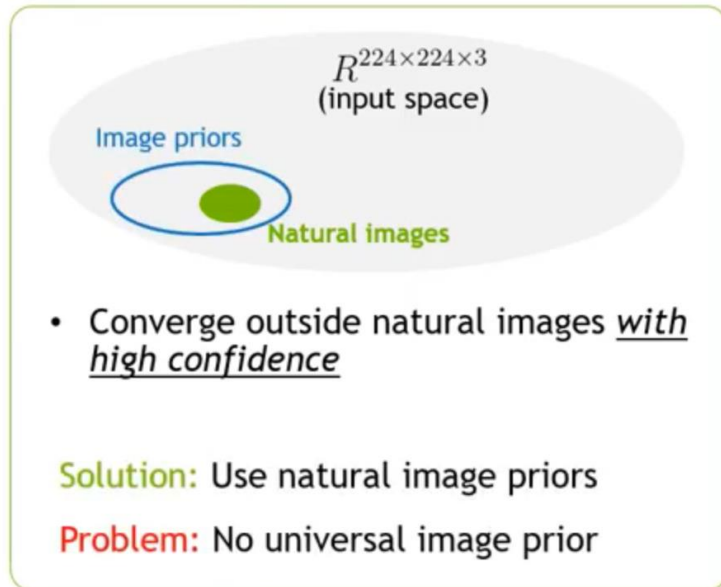
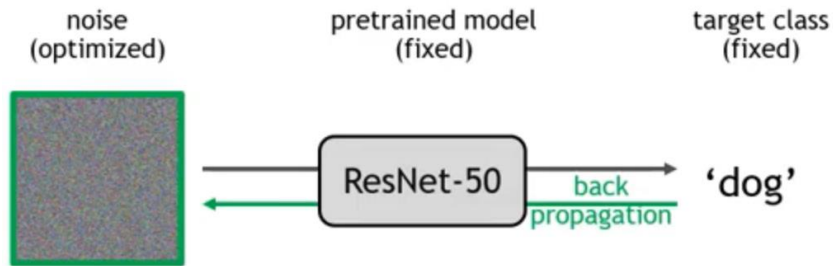
Module-2 : Defending against Adversarial Attacks

Motivation for adversarial defense research, Adversarial Training (PGD-AT, TRADES, AWP, SOTA tricks), Robust evaluation of Adversarial Defenses (Auto-Attack)

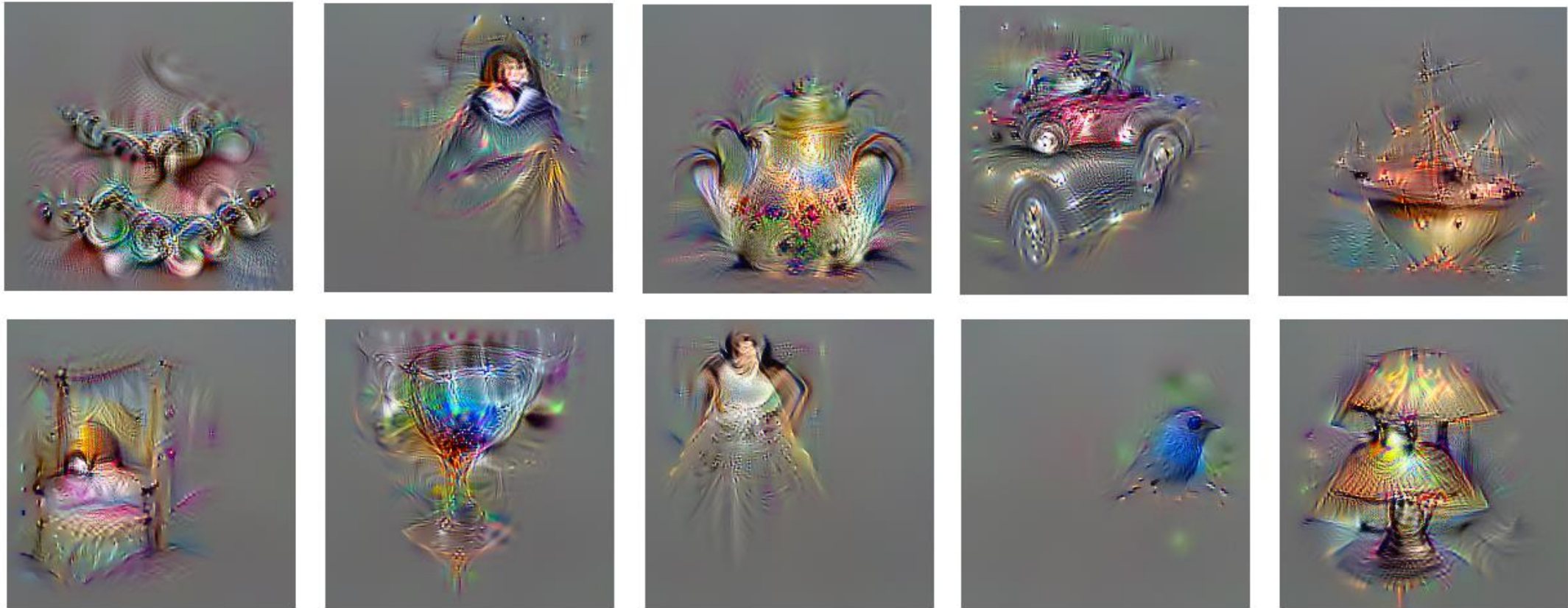
Motivation for Adversarial Defense Research

- Defending against an adversary who can potentially attack the system
 - Physical adversarial attacks
- Verifying the worst-case robustness of machine learning models
 - Random testing may not be good enough in safety critical applications
- Robust networks possess interesting properties
 - Representations learned
 - Decision boundaries
 - Loss surface
- Measure progress of machine learning algorithms towards human-level abilities
 - In many applications, machine learning algorithms are progressing towards human-level performance
 - In the area of adversarial robustness, this gap is large

Generating visualizations from pretrained model



Saliency-driven Class Impressions



Sravanti Addepalli, Dipesh Tamboli, R. Venkatesh Babu, Biplab Banerjee, Saliency-driven Class Impressions for Feature Visualization of Deep Neural Networks, ICIP 2020

Robust models have perceptually aligned gradients

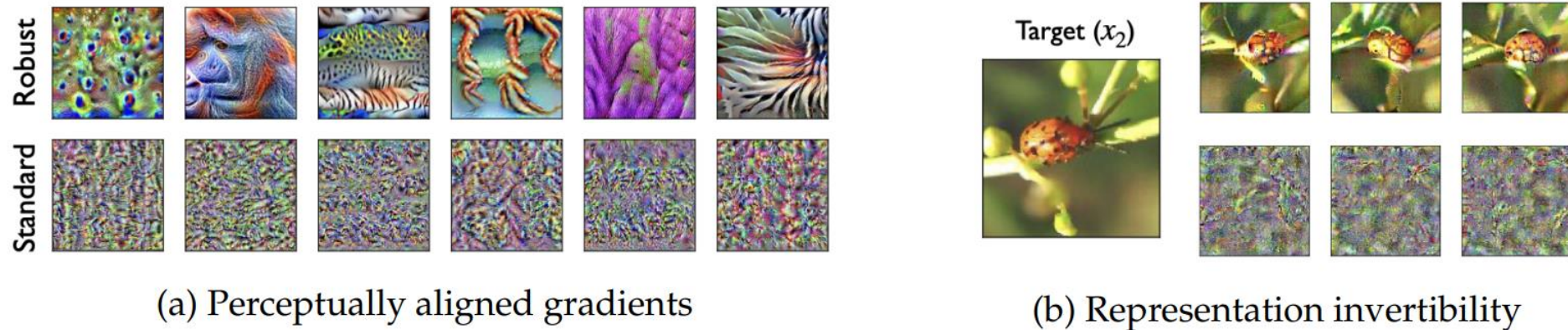
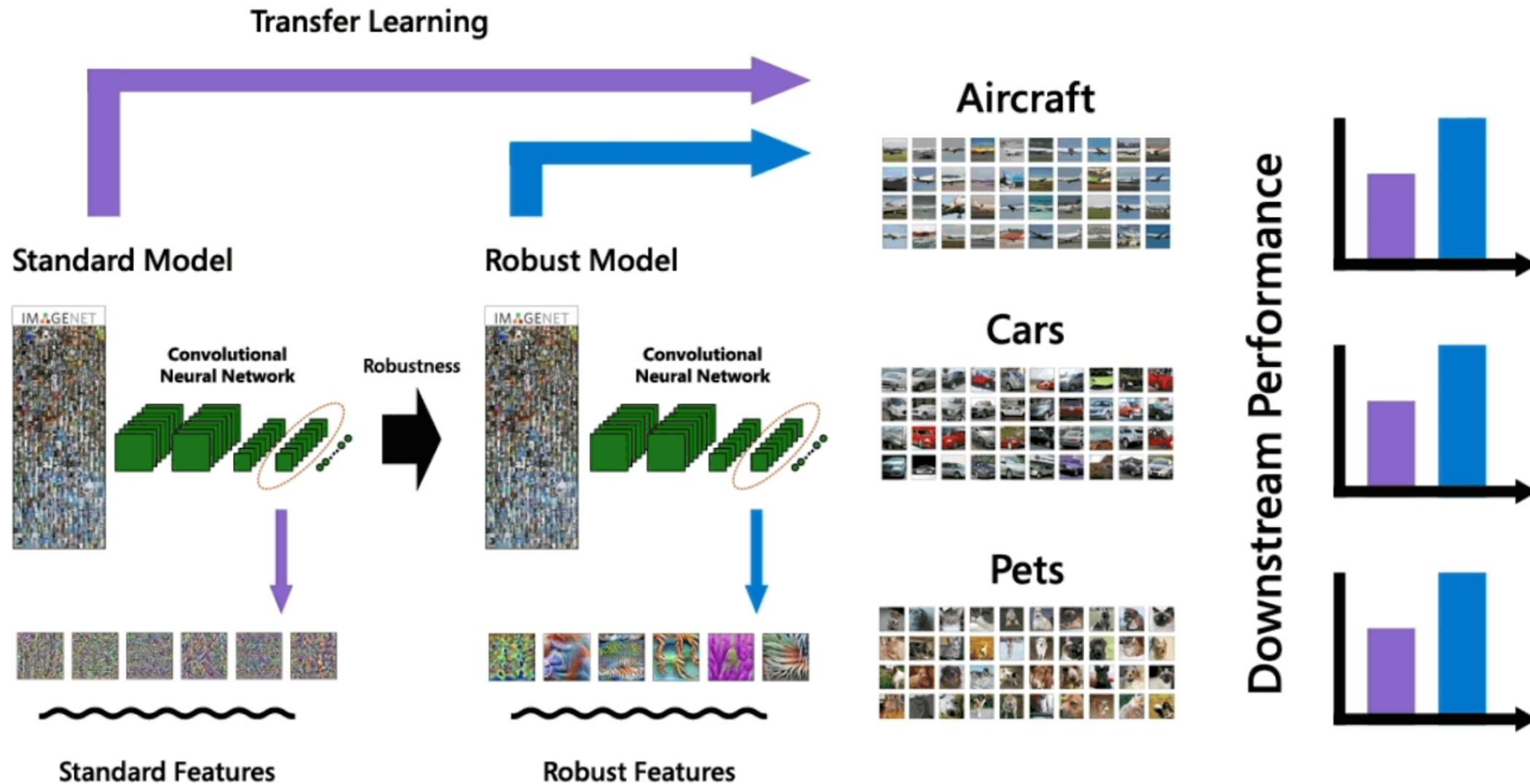


Figure 1: Adversarially robust (top) and standard (bottom) representations: robust representations allow (a) feature visualization without regularization; (b) approximate image inversion by minimizing distance in representation space. Figures reproduced from Engstrom et al. [Eng+19a].

Robust ImageNet Models Transfer Better

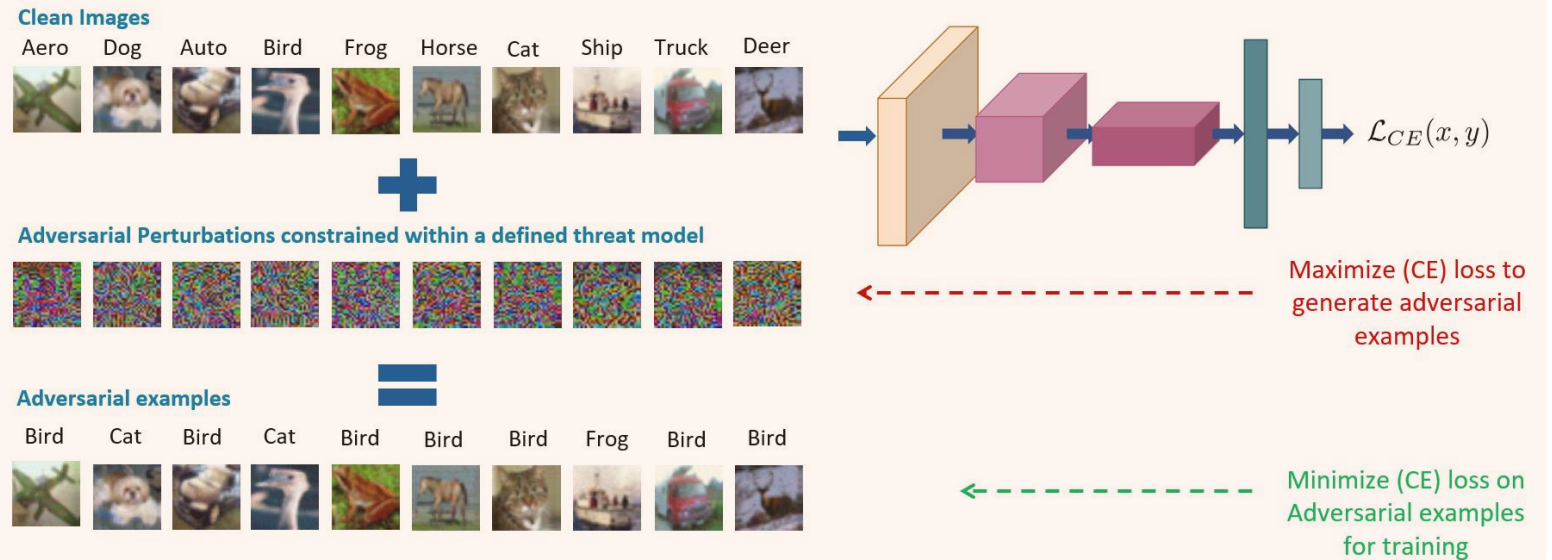


Defending against Adversarial Attacks

Input pre-processing based defenses

- Non-differentiable pre-processing steps thwart gradient-based attacks
- Bit-depth reduction, JPEG compression, etc. ¹
- Limited/ No added computational cost during training
- Defenses broken using BPDA and EOT attacks ²

Adversarial Training based methods

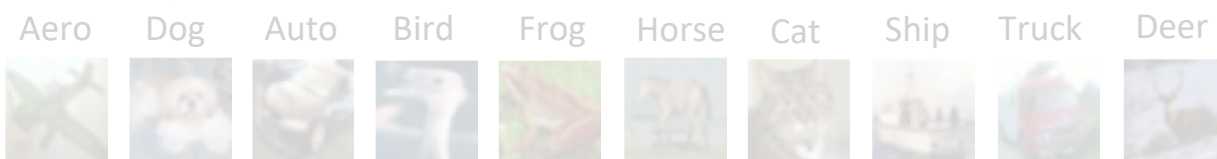


¹ Guo et al. Countering adversarial images using input transformations. ICLR, 2018.

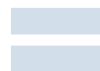
² Athalye et al., Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples, ICML 2018

Adversarial Training

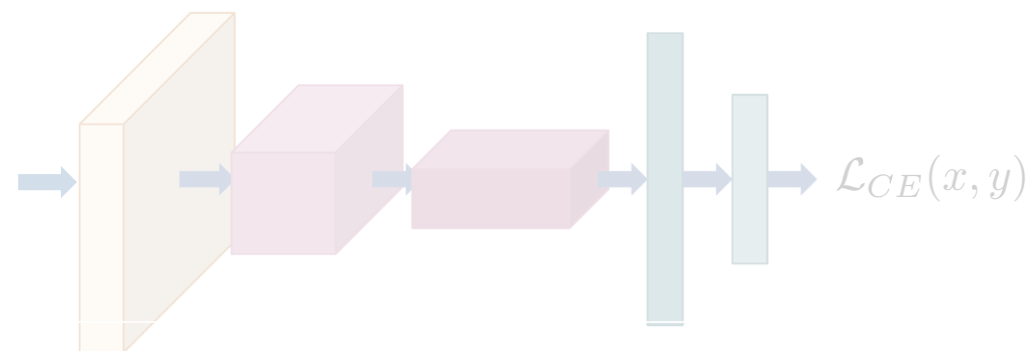
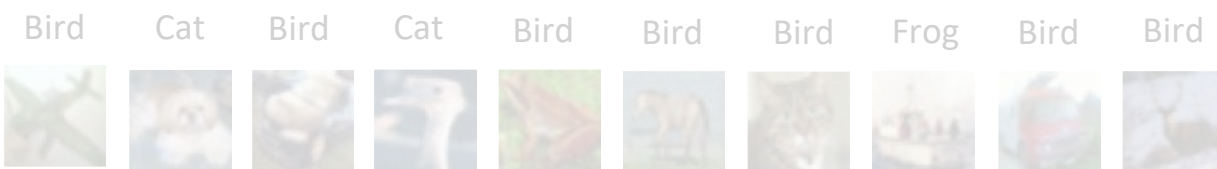
Clean Images



Adversarial Perturbations constrained within a defined threat model



Adversarial examples

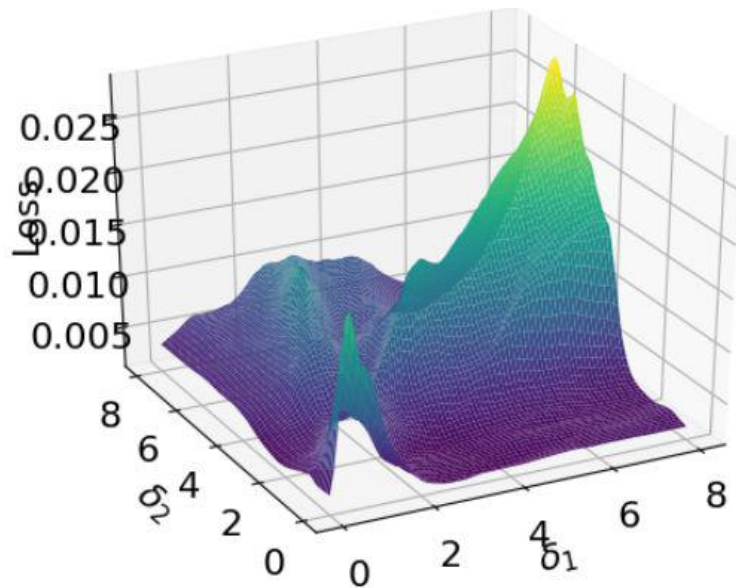


Maximize (CE) loss to
generate adversarial
examples



Minimize (CE) loss on
Adversarial examples
for training

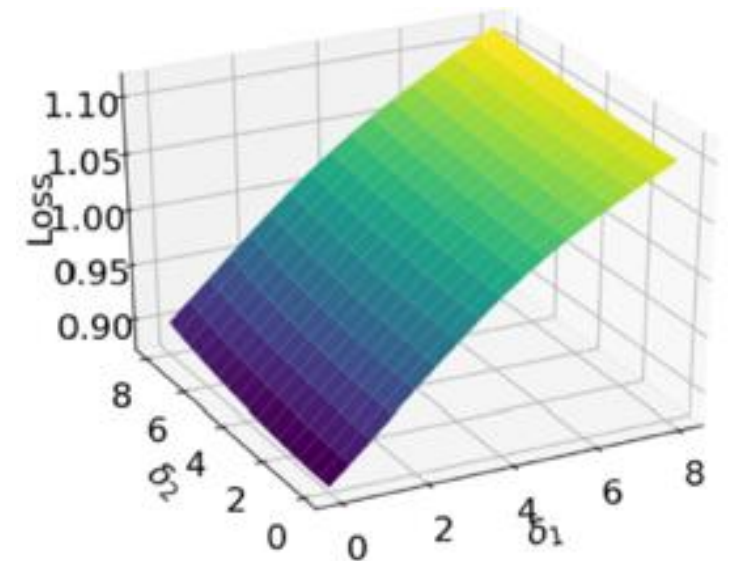
Defending against Adversarial Attacks



FGSM-AT

Single-step defenses

- Single-step gradients used for attack generation
- Lower computational cost
- Susceptible to Gradient Masking leading to a false sense of security and training instability, specifically for long training schedules and larger model capacities



PGD-AT

¹ Guo et al. Countering adversarial images using input transformations. ICLR, 2018.

² Athalye et al., Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples, ICML 2018

³ Madry et al. Towards Deep Learning Models Resistant to Adversarial Attacks. ICLR, 2018.

⁴ Zhang et al. Theoretically principled trade-off between robustness and accuracy. ICML, 2019.

Recent advances in Adversarial Defenses

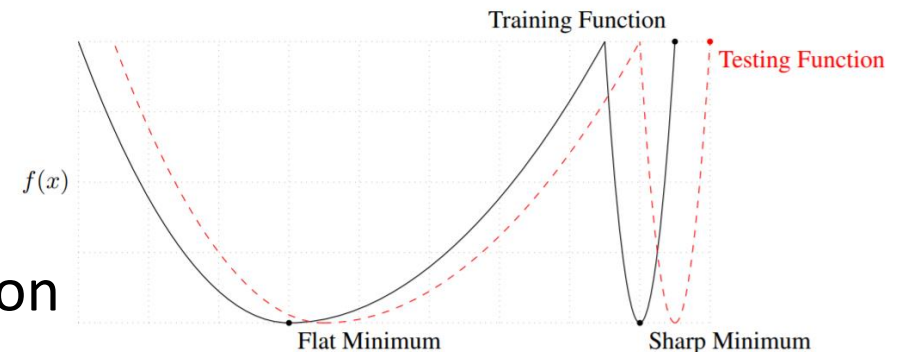
- TRADES defense
 - Optimizes an upper bound of adversarial risk
 - Trade-off between accuracy and robustness

$$\rho^{\text{TRADES}}(\mathbf{w}) = \frac{1}{n} \sum_{i=1}^n \left\{ \text{CE}(f_{\mathbf{w}}(\mathbf{x}_i), y_i) + \beta \cdot \max \text{KL}(f_{\mathbf{w}}(\mathbf{x}_i) \| f_{\mathbf{w}}(\mathbf{x}'_i)) \right\}$$

- Adversarial Weight Perturbation (AWP)

$$\min_{\mathbf{w}} \max_{\mathbf{v} \in \mathcal{V}} \frac{1}{n} \sum_{i=1}^n \max_{\|\mathbf{x}'_i - \mathbf{x}_i\|_p \leq \epsilon} \ell(\mathbf{f}_{\mathbf{w}+\mathbf{v}}(\mathbf{x}'_i), y_i)$$

- Leads to flatter minima and better generalization
- Can be applied to any defense



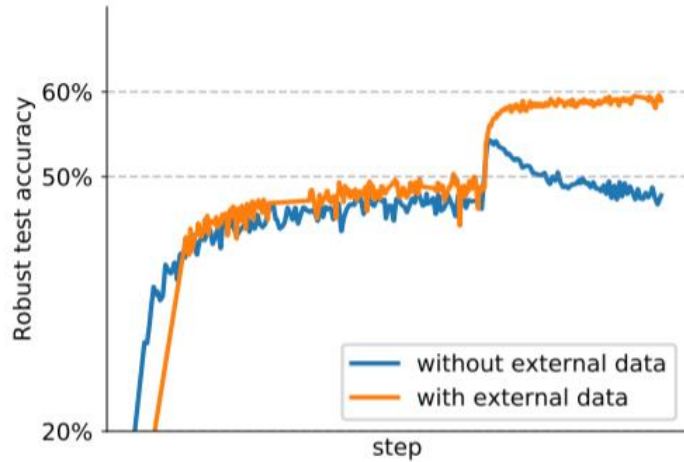
Source: <https://arxiv.org/pdf/1609.04836.pdf>

Zhang, Hongyang, et al. "Theoretically principled trade-off between robustness and accuracy." ICML 2019

Wu et al. "Adversarial weight perturbation helps robust generalization." NeurIPS 2020

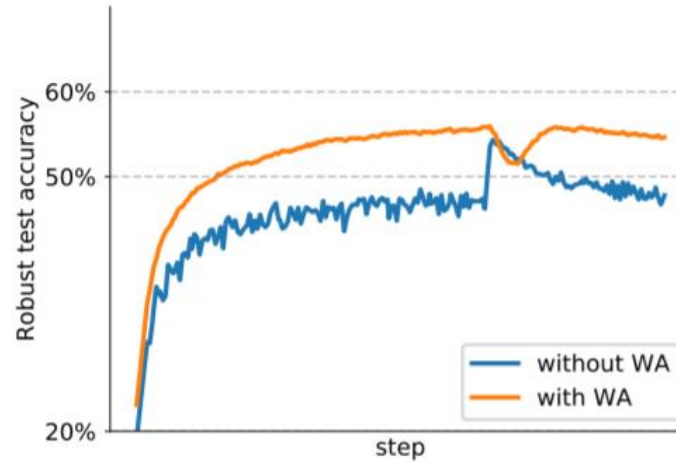
Tricks to improve Adversarial Robustness

Robust Overfitting and Early Stopping

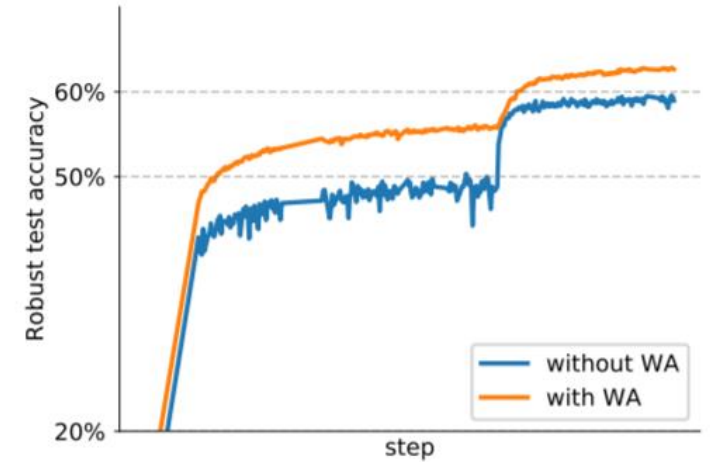


(a) Adversarial training with and without additional external data from 80M-T1

Weight Averaging



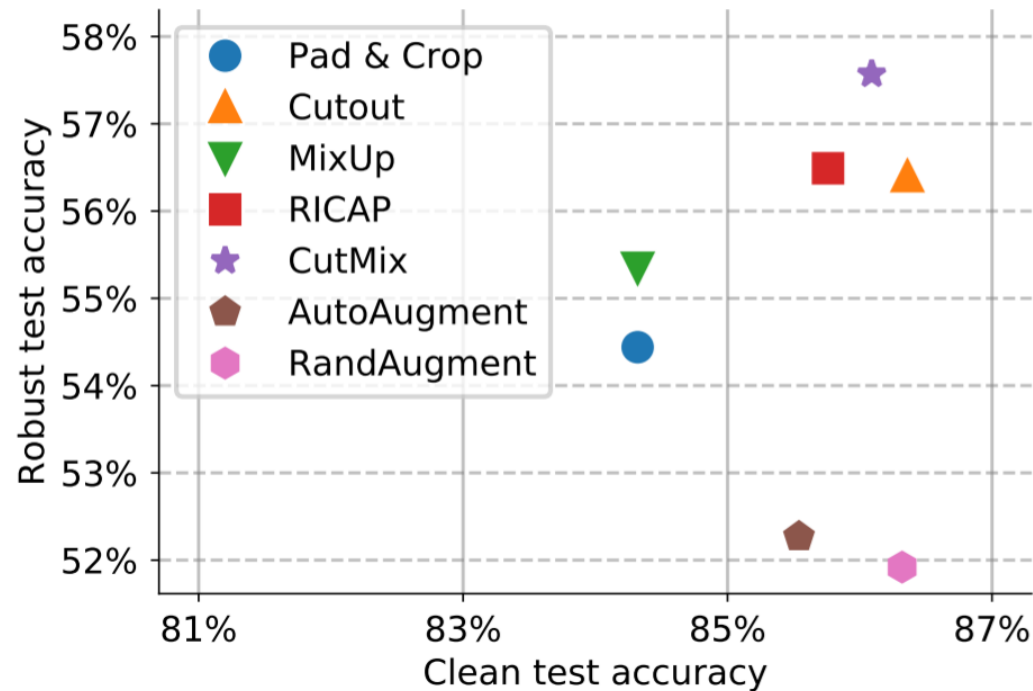
(b) Effect of WA without external data



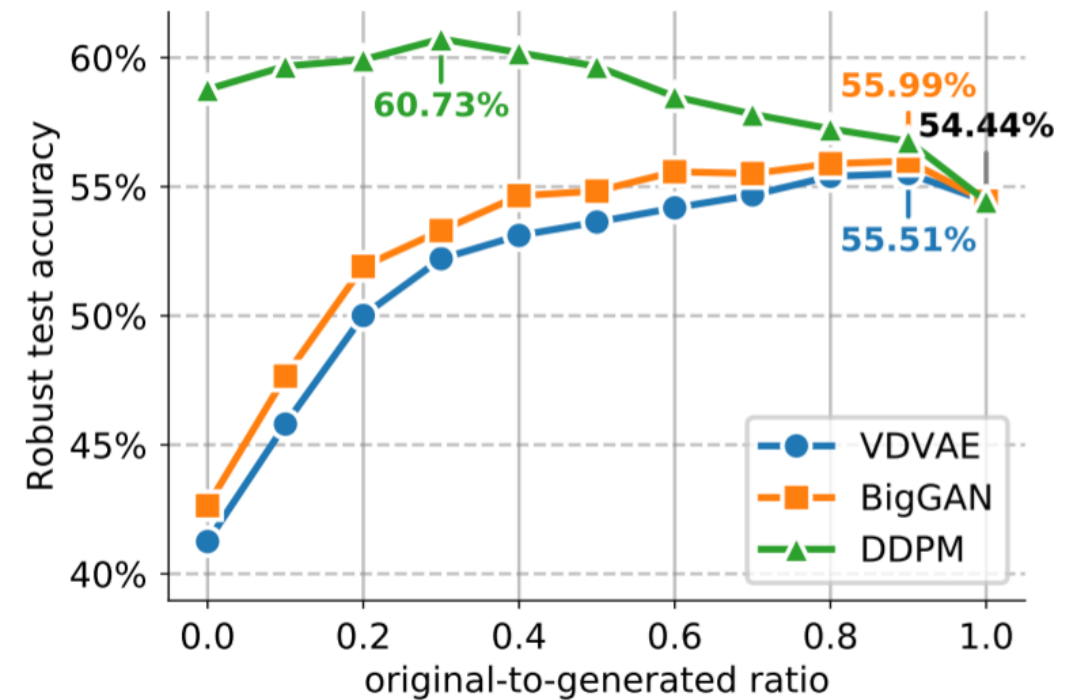
(c) Effect of WA with external data

Tricks to improve Adversarial Robustness

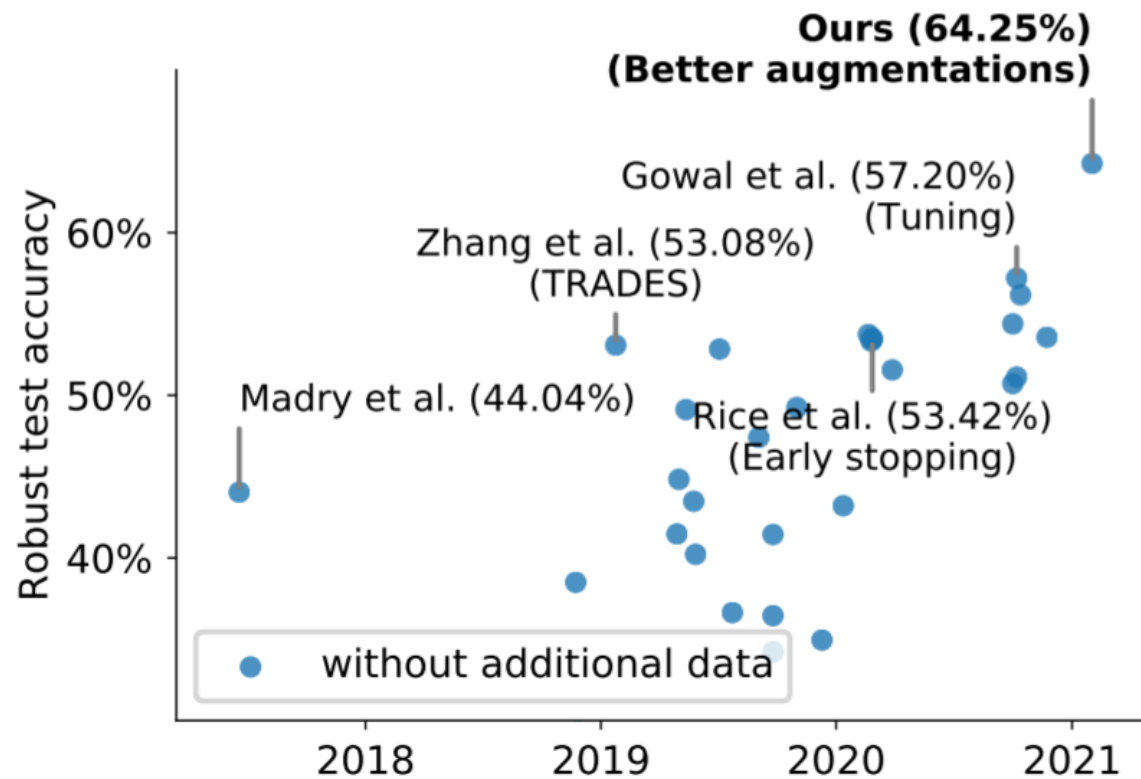
Effect of Data Augmentations




Augmenting training with synthetic data



Trends in Adversarial Defenses over the years



Evaluating Adversarial Robustness

ROBUSTBENCH Leaderboards Paper FAQ Contribute Model Zoo 

Available Leaderboards

CIFAR-10 (ℓ_∞) CIFAR-10 (ℓ_2) CIFAR-10 (Corruptions) CIFAR-100 (ℓ_∞)
CIFAR-100 (Corruptions)

Leaderboard: CIFAR-10, $\ell_\infty = 8/255$, Untargeted, AutoAttack

Show 15 entries Search: Papers, architectures, ve

Rank	Method	Standard accuracy	Robust accuracy	Extra data	Architecture	Venue
1	Fixing Data Augmentation to Improve Adversarial Robustness <i>We show the robust accuracy reported in the paper since AutoAttack performs slightly worse (66.58%).</i>	92.23%	66.56%	<input checked="" type="checkbox"/>	WideResNet-70-16	arXiv, Mar 2021

- The following types of defenses CANNOT be effectively evaluated using Auto-Attack
 - Classifiers that have zero gradients with respect to the input
 - Randomized classifiers
 - Classifiers that contain an optimization loop in their predictions

<https://robustbench.github.io/>
<https://github.com/fra31/auto-attack>

Summary of Module-2

- Motivation for Adversarial Defense research
 - Defending against an adversary
 - Adversarially trained models have more perceptually aligned gradients
- Adversarial Training
 - FGSM, PGD
 - TRADES, AWP
 - Tricks for improving adversarial defenses
- Evaluating Adversarial Robustness
 - Benchmarks – RobustBench and Auto-Attack