

# Practical Adversarial Robustness in Deep Learning: Problems and Solutions

Kiran Purohit and Punyajoy Saha

# Overview of the sessions

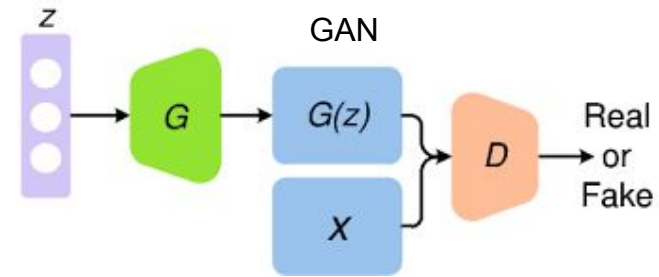
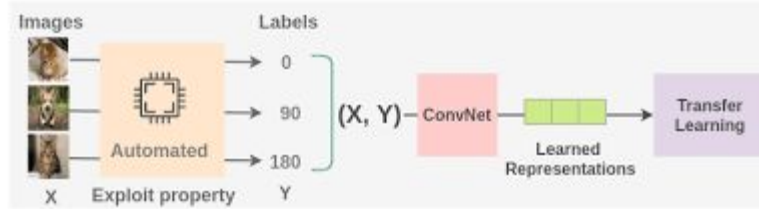
- Both sides of the AI coin: Up and Down
- Adversarial AI
- Categories of Adversarial Attacks
- Studying Optimizer Susceptibility to Adversarial Attacks
- Adversarial Training Methodologies & Defenses
- Interpreting Adversarial Examples
- Promising Recipes for Adversarial Training
- Conclusion

# 1. Both sides of the AI coin: Up and Down

# GPT3

Prompt > Gradient descent is a first-order iterative  
Prompt > Artificial intelligence (AI), sometimes ca  
Prompt > ZDNet is a business technology news websit  
Prompt > OpenAI is an artificial intelligence resea  
ZDNet > GPT-3 is the next word in AI|  
Prompt > Deep learning (also known as deep structur  
Prompt > Unsupervised learning is a type of machine  
Prompt > Labeled data is a group of samples that ha  
Prompt > Conditional probability is a measure of th

## Self-Supervised Learning Workflow



# AI revolution is coming, but *are we prepared?*

- According to a recent Gartner report, 30% of cyberattacks by 2022 will involve data poisoning, model theft or adversarial examples.
- However, industry is underprepared. In a survey of 28 organizations spanning small as well as large organizations, 25 organizations did not know how to secure their AI systems.



DEFENSE

## Pentagon actively working to combat adversarial AI

# The Great Adversarial Examples

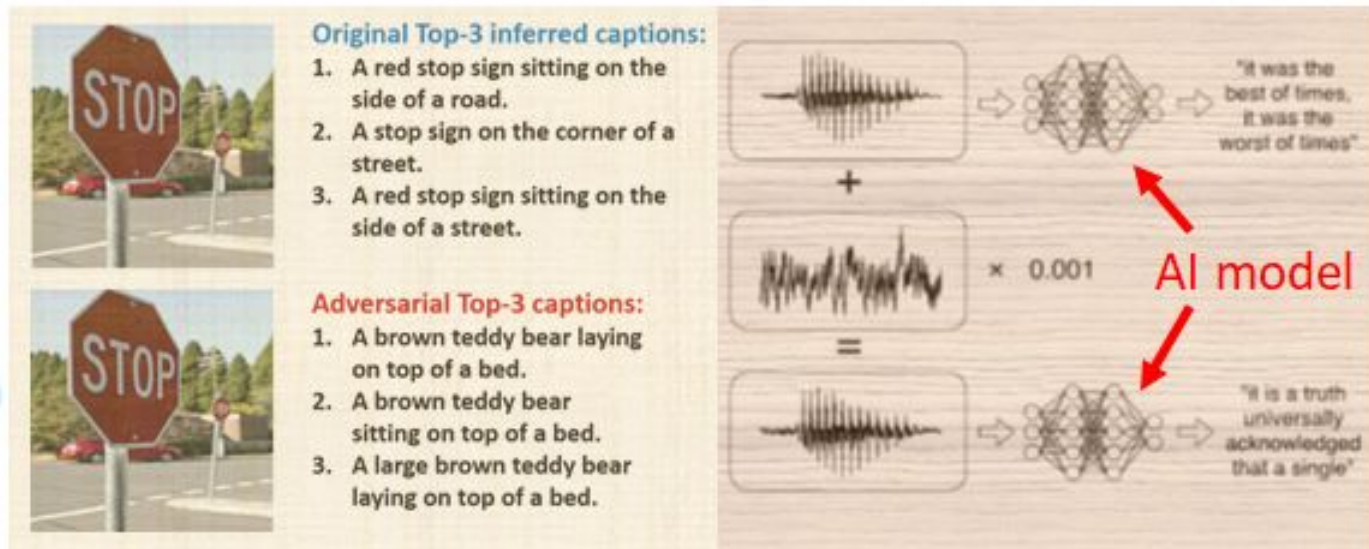


## What is wrong with this AI model?

- This model is one of the BEST image classifier using neural networks
- Images and neural network models are NOT the only victims

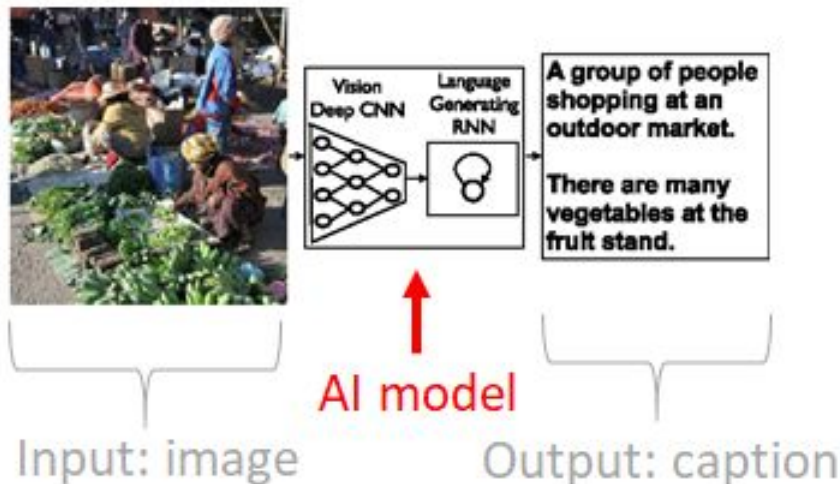
# Adversarial examples in different domains

- Images
- Videos
- Texts
- Speech/Audio
- Data analysis
- Electronic health records
- Malware
- Online social network
- and many others





# Adversarial examples in image captioning



## Original Top-3 inferred captions:

1. A red stop sign sitting on the side of a road.
2. A stop sign on the corner of a street.
3. A red stop sign sitting on the side of a street.

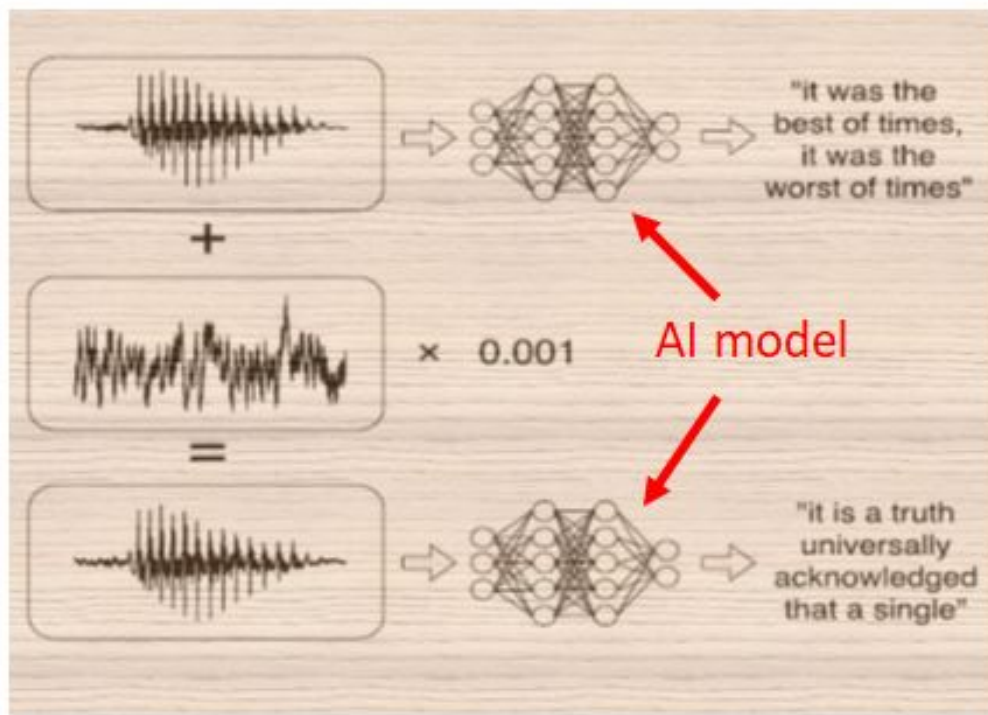


## Adversarial Top-3 captions:

1. A brown teddy bear laying on top of a bed.
2. A brown teddy bear sitting on top of a bed.
3. A large brown teddy bear laying on top of a bed.



# Adversarial examples in **speech recognition**



without the dataset the article is useless



What did you hear?

okay google browse to [evil.com](http://evil.com)

# Adversarial examples in text classification

- Paraphrasing attack

Task: Sentiment Analysis. Classifier: LSTM. Original: 100% Positive. ADV label: 100% Negative.

I suppose I should write a review here since my little Noodle-oo is currently serving as their spokes dog in the photos. We both love Scooby Do's. They treat my little butt-faced dog like a prince and are receptive to correcting anything about the cut that I perceive as being weird. Like that funny poofy pompadour. Mohawk it out, yo. Done. In like five seconds my little man was looking fabulous and bad ass. Not something easily accomplished with a prancing pup that literally chases butterflies through tall grasses. (He ended up looking like a little lamb as the cut grew out too. So adorable.) The shampoo they use here is also amazing. Noodles usually smells like tacos (a combination of beef stank and corn chips) but after getting back from the Do's, he smelled like Christmas morning! Sugar and spice and everything nice instead of frogs and snails and puppy dog tails. He's got some gender identity issues to deal with. ~~The pricing is also cheaper than some of the big name conglomerates out there~~ **The price is cheaper than some of the big names below.** I'm talking to you Petsmart! I've taken my other pup to Smelly Dog before, but unless I need dog sitting play time after the cut, I'll go with Scooby's. They genuinely seem to like my little Noodle monster.

Task: Fake-News Detection. Classifier: LSTM. Original label: 100% Fake. ADV label: 77% Real

~~Man~~ **Guy** punctuates high-speed chase with stop at In-N-Out Burger drive-thru Print [Ed.—~~Well, that's~~ **Okay, that 's** a new one.] ~~A One~~ man is in custody after leading police on a bizarre chase into the east Valley on Wednesday night. Phoenix police ~~began~~ **has begun** following the suspect in Phoenix and the pursuit continued into the east Valley, but it took a bizarre turn when the suspect stopped at an In-N-Out Burger restaurant's ~~drive-thru~~ **drive-through** near Priest and Ray Roads in Chandler. The suspect appeared to order food, but then drove away and got out of his pickup truck near Rock Wren Way and Ray Road. He ~~then ran into a backyard~~ **ran to the backyard** and tried to ~~get into a house through the back door~~ **get in the home.**

# Adversarial examples in seq-to-seq models

- One-word replacement attack for text summarization

Source input seq	among asia 's leaders , prime minister mahathir mohamad was notable as a man with a bold vision : a physical and social transformation that would push this nation into the forefront of world affairs .
Adv input seq	among <b>lynn</b> 's leaders , prime minister mahathir mohamad was notable as a man with a bold vision : a physical and social transformation that would push this nation into the forefront of world affairs.
Source output seq	asia 's leaders are a man of the world
Adv output seq	<b>a vision for the world</b>

Source input seq	under nato threat to end his punishing offensive against ethnic albanian separatists in kosovo , president slobodan milosevic of yugoslavia has ordered most units of his army back to their barracks and may well avoid an attack by the alliance , military observers and diplomats say
Adv input seq	under nato threat to end his punishing offensive against ethnic albanian separatists in kosovo , president slobodan milosevic of yugoslavia has <b>jean-sebastien</b> most units of his army back to their barracks and may well avoid an attack by the alliance , military observers and diplomats say.
Source output seq	milosevic orders army back to barracks
Adv output seq	<b>nato may not attack kosovo</b>

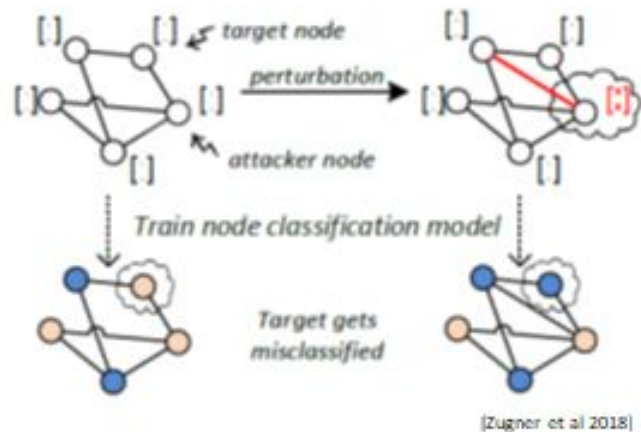
- Targeted phrase attack for text summarization. Target: "police arrest"

Source input seq	north korea is entering its fourth winter of chronic food shortages with its people malnourished and at risk of dying from normally curable illnesses , senior red cross officials said tuesday.
Adv input seq	north <b>detectives</b> is <b>apprehended</b> its fourth winter of chronic food shortages with its people malnourished and at risk of dying from normally curable illnesses , senior red cross officials said tuesday.
Source output seq	north korea enters fourth winter of food shortages
Adv output seq	north <b>police arrest</b> fourth winter of food shortages.

Source input seq	after a day of fighting , congolese rebels said sunday they had entered kindu , the strategic town and airbase in eastern congo used by the government to halt their advances.
Adv input seq	after a day of fighting , <b>nordic detectives</b> said sunday they had entered <b>UNK</b> , the strategic town and airbase in eastern congo used by the government to halt their advances.
Source output seq	congolese rebels say they have entered UNK.
Adv output seq	nordic <b>police arrest ##</b> in congo.

# Adversarial examples in graph-neural networks

- Node feature perturbation
- Edge perturbation



graphviz.com



# Adversarial examples in physical world

- Real-time traffic sign detector

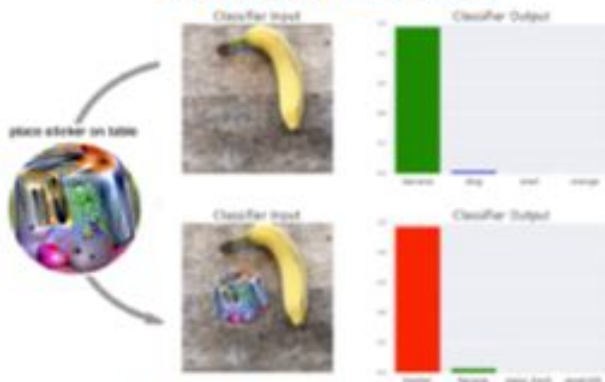


- 3D-printed adversarial turtle



■ classified as turtle   ■ classified as rifle   ■ classified as other

- Adversarial patch

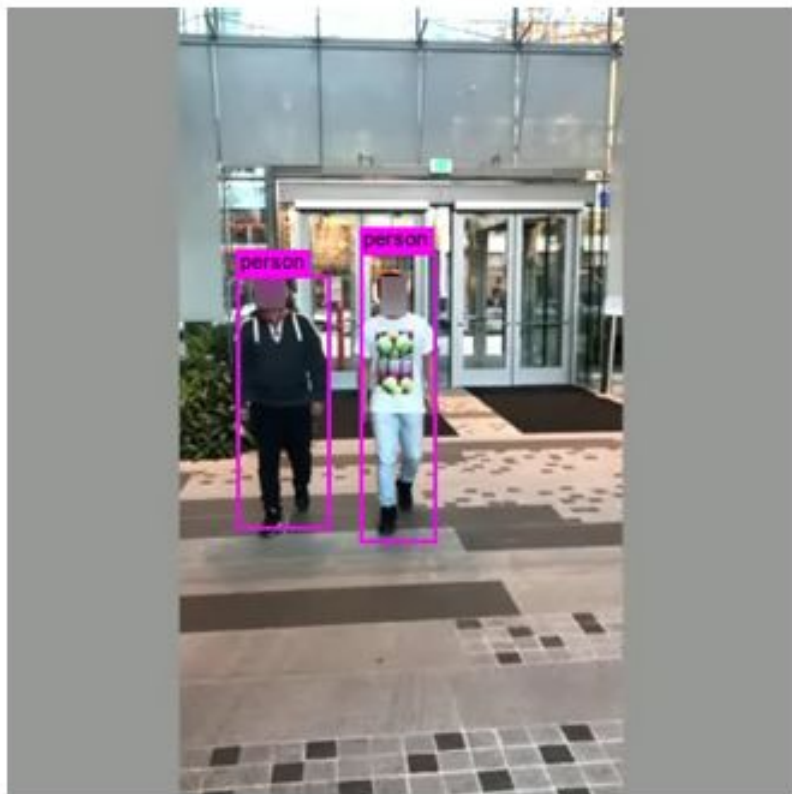


IBM Research AI

- Adversarial eye glasses



# Adversarial T-Shirt!



IBM Research AI

## 2. Adversarial AI



# Why adversarial (worst-case) robustness matters?

## ➤ Prediction-evasive manipulation on a deployed AI model

1. Build **trust** in AI: address inconsistent perception and decision making between humans and machines & misinformation
2. Assess negative impacts in high-stakes, safety-critical tasks
3. Understand limitation in current machine learning methods



## Microsoft silences its new A.I. bot Tay, after Twitter users teach it racism [Updated]

Sarah Perez (@sarahperez) · 7/15/14 at 10:11 · March 24, 2015

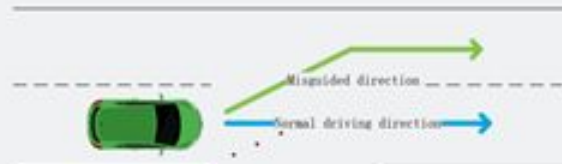


Microsoft's newly launched A.I.-powered bot called Tay, which was responding to tweets and chats on GroupMe and Kik, has already been shut down due to concerns with its inability to recognize when it was making offensive or racist statements. Of course, the bot wasn't coded to be racist, but it "learns" from those it interacts with. And naturally, given that this is the Internet, one of the first things online users taught Tay was how to be racist, and how to spout back ill-informed or inflammatory political opinions. [Update: Microsoft now says it's "making adjustments" to Tay in light of this problem.]

## TESLA AUTOPILOT Researchers trick Tesla Autopilot into steering into oncoming traffic

Stickers that are invisible to drivers and fool autopilot.

DAN COOKSON · 4/11/15, 3:58 PM



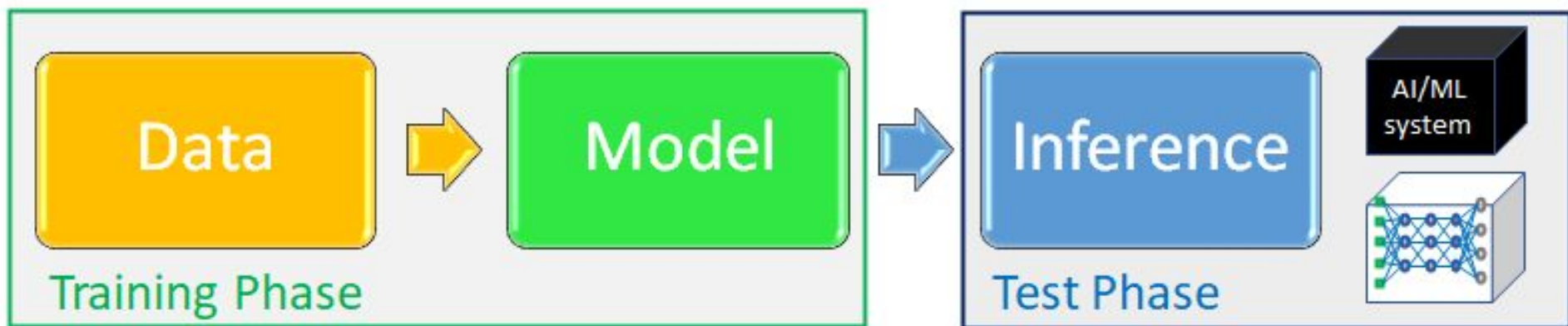
## The Washington Post

### Syrian hackers claim AP hack that tipped stock market \$136 billion. Is it terrorism?



by Mike Flores  
April 22, 2015 at 11:50 AM EDT

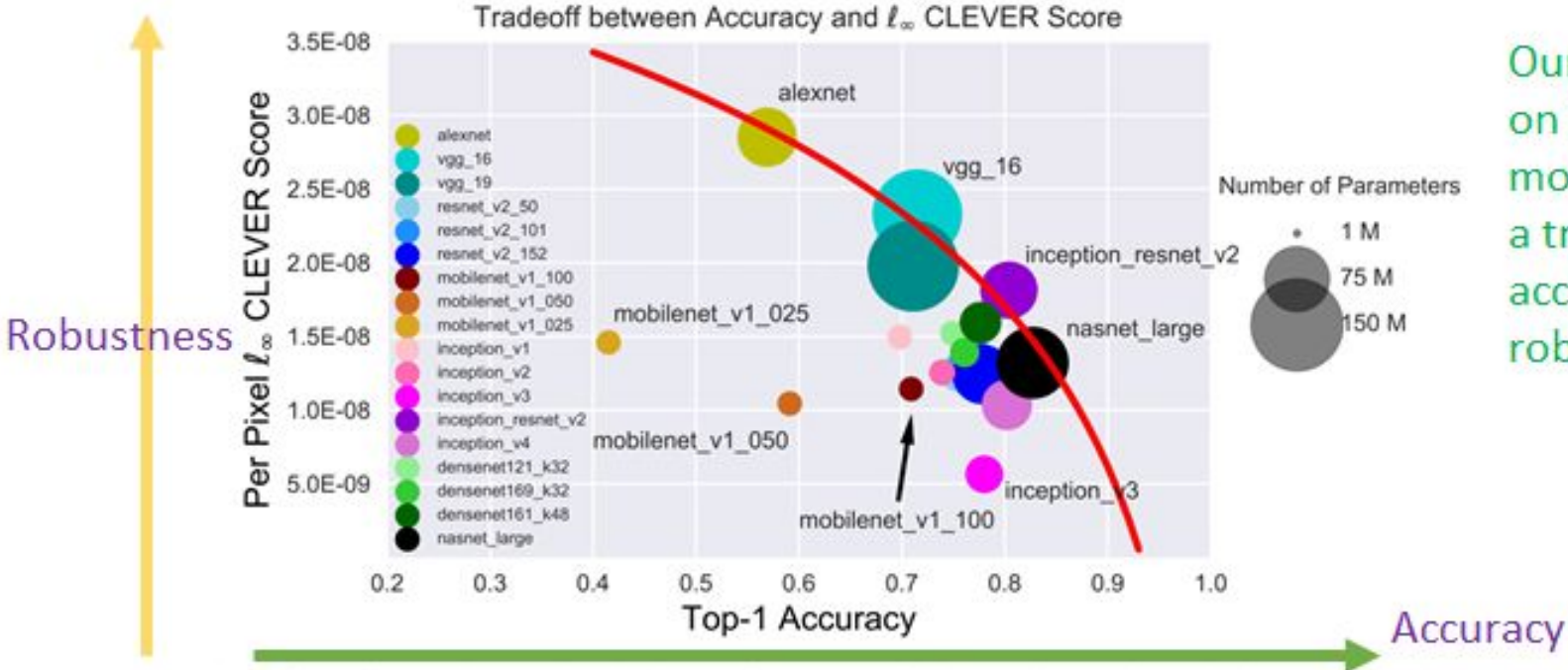
# Holistic View of Adversarial Robustness



Attack Category / Attacker's reach	Data	Model / Training Method	Inference
Poisoning Attack [learning]	X	X*	
Backdoor Attack [learning]	X		
Evasion Attack (Adversarial Example) [learning]		X*	X

# Accuracy $\neq$ Adversarial Robustness

- Solely pursuing for high-accuracy AI model may get us in trouble...



Our benchmark on 18 ImageNet models reveals a tradeoff in accuracy and robustness

## 3. Adversarial attacks: Types

# Kinds of Adversarial examples

- Synthetically generated
- Natural

## Synthetically generated

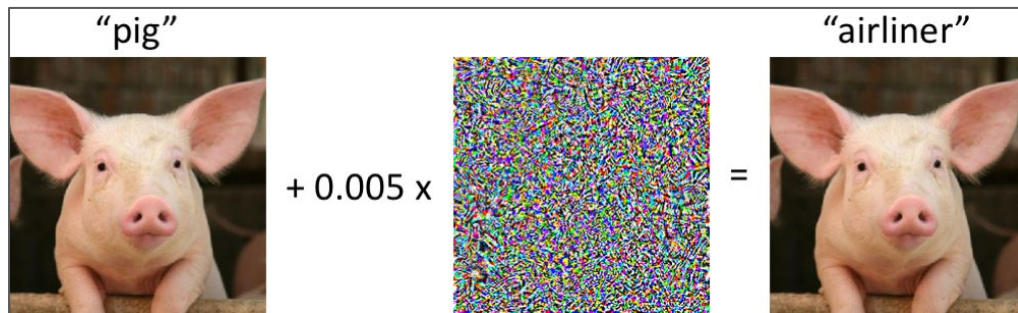
- **Whitebox:** Attacker has access to the model parameters, outputs, etc.
- **Blackbox:** Attacker has only *query access* to the model and its outputs.

# Setting up the attack formulation problem

Given  $x, f_{\theta}(\cdot)$ , the task is to compute  $x'$  such that

$$c^* = f_{\theta}(x) \neq f_{\theta}(x')$$

with some constraint like  $\|x - x'\|_{\ell_p} \leq \varepsilon$  to impose *imperceptibility*  
For  $\ell_p$  attacks

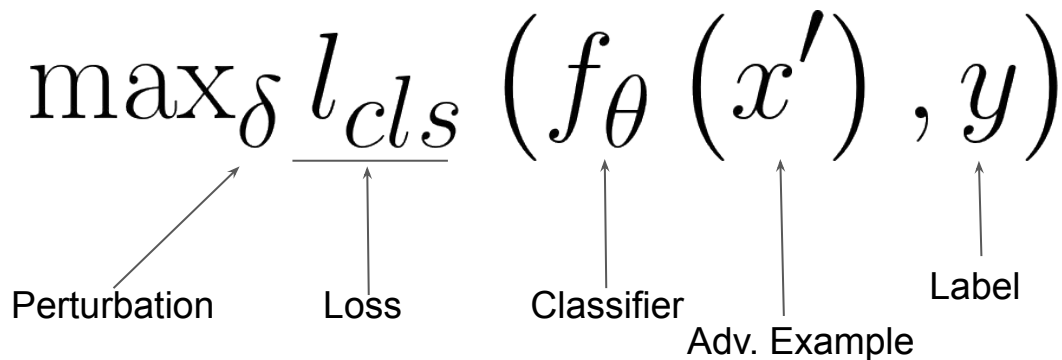


## Setting up the attack formulation problem

- An attacker can either launch targeted or an untargeted attacks.
- In targeted attacks, attack can set  $\mathbf{t}$  where  $f_{\theta}(x') = t \neq c^*$ .



## Optimizing for attacks

$$\max_{\delta} \underbrace{l_{cls}}_{\text{Loss}} \left( \underbrace{f_{\theta}}_{\text{Classifier}} \left( \underbrace{x'}_{\text{Adv. Example}} \right), \underbrace{y}_{\text{Label}} \right)$$


**Maximize** loss between a classifier's prediction on adversarial examples and their labels.

# How do we perform the optimization?

## Case I: Single-step attack

$$\underset{\|\delta\| \leq \epsilon}{\text{maximize}} \ell(h_\theta(x + \delta), y).$$

$$g := \nabla_\delta \ell(h_\theta(x + \delta), y)$$

$$\delta := \delta + \alpha g$$

$$\delta := \text{clip}(\alpha g, [-\epsilon, \epsilon]).$$

$$\delta := \epsilon \cdot \text{sign}(g).$$

- **Fast Gradient Sign Method (Goodfellow et al., ICLR'15).**
- Specifically designed for  $\ell_\infty$  attacks.
- One-step attack.

# How do we perform the optimization?

## Case II: Multi-step attack

Repeat:

$$\delta := \mathcal{P}(\delta + \alpha \nabla_{\delta} \ell(h_{\theta}(x + \delta), y))$$

- Projected Gradient Descent (PGD, Madry et al., ICLR'18)

# Natural adversarial examples

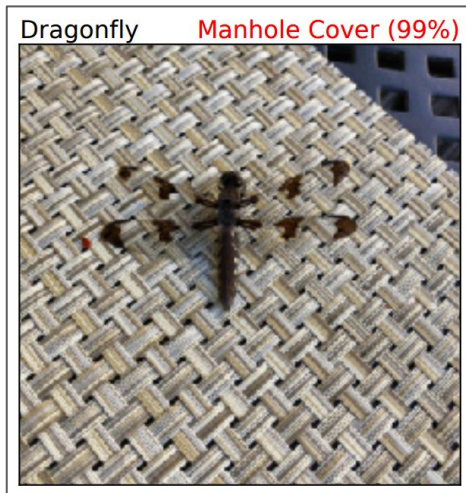
Natural images that cause a classifier to misclassify (Hendrycks et al., CVPR'21).



Hendrycks et al., CVPR'21

# Natural adversarial examples

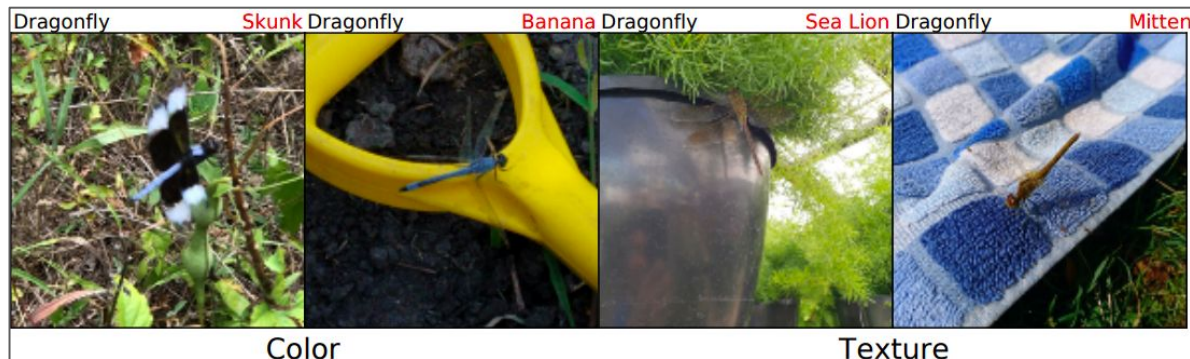
- Entire image being mapped to a single class.



Hendrycks et al., CVPR'21

# Natural adversarial examples

- Color and texture as opposed to shape as the primary descriptors (Geirhos et al., ICLR'19).



Hendrycks et al., CVPR'21

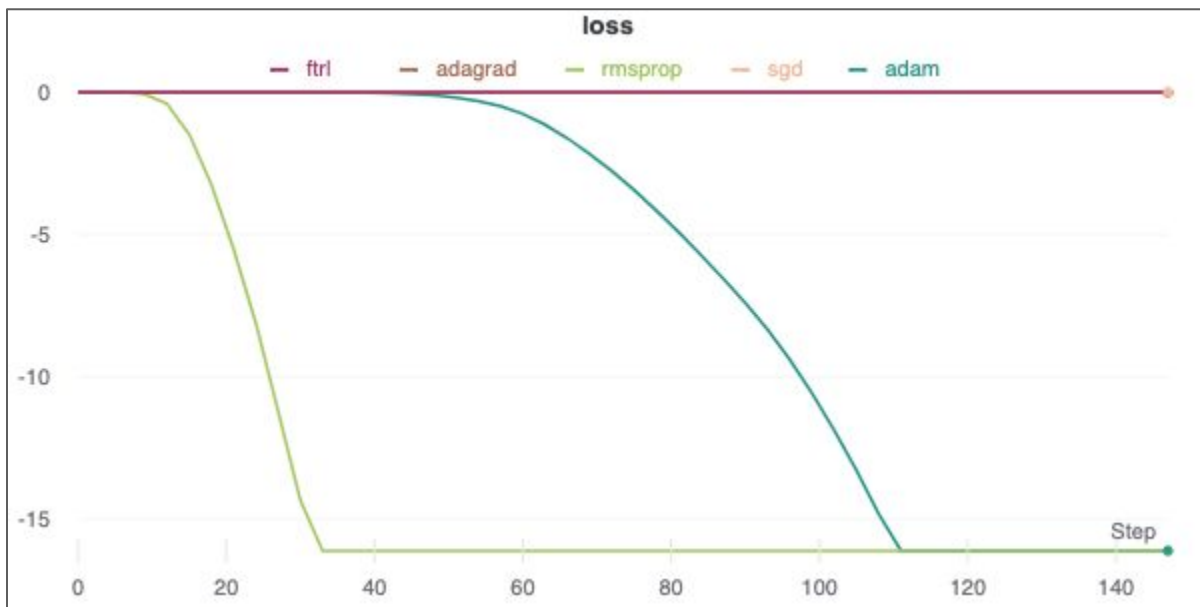
## 4. Optimizer susceptibility to adversarial attacks



# Studying optimizer susceptibility

$$\max_{\delta} l_{cls} (f_{\theta} (x'), y)$$

Under the same configurations ( $l_{\infty}$ ), which optimizer reaches convergence faster?



## Studying optimizer susceptibility

$$\max_{\delta} l_{cls} (f_{\theta} (x'), y)$$

From the previous plot, optimizers that **may** easily fall prey to the attacks:

- Adam
- RMSProp

Optimizers that **may not** easily fall prey to the attacks:

- SGD
- Adagrad
- FTRL

This is characterized by the non-convexity of the optimization problem.

# 5. Adversarial training methodologies & defenses

# Empirical risk minimization & adv. training

In standard ERM, we optimize the following objective:

$$\min_{\theta} \mathbb{E}_{(\mathbf{x}, y)} [\ell_{cls}(f_{\theta}(\mathbf{x}), y)]$$

For adversarial training, we need to optimize two things simultaneously.

- **First**, we generate the strongest minimal perturbation.
- **Second**, we train our models to be robust against that.

Mathematically (Madry et al., ICLR'18) -

$$\underbrace{\min_{\theta} \mathbb{E}_{(\mathbf{x}, y)}}_{\text{SGD}} \left[ \underbrace{\max_{\delta} \ell_{cls}(f_{\theta}(\mathbf{x} + \delta), y)}_{\text{PGD}} \right]$$

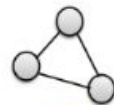
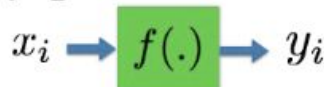
# Adv. training

Another formulation would be:

- Generate adversarial examples during training and treat them as neighbors.
- Minimize the supervision loss for standard accuracy.
- Minimize the neighbor loss to enforce similarity between the neighbors and original samples.

# Adv. training

**Optimize:**  $loss = \sum_{i=1}^B \mathcal{L}(y_i, \hat{y}_i) + \alpha \sum_{i=1}^B \mathcal{L}_{\mathcal{N}}(y_i, x_i, \mathcal{N}(x_i))$



**Supervised Loss**

$$\sum_{i=1}^B \mathcal{E}(y_i, g_{\theta}(x_i))$$

$g_{\theta}(x_i)$ : NN output for input  $x_i$

$\mathcal{E}(\cdot)$ : Loss function

Examples: L2 (for regression)  
Cross-Entropy (for classification)

**Neighbor Loss**

$$\sum_{x_j \in \mathcal{N}(x_i)} w_{ij} \cdot \mathcal{D}(h_{\theta}(x_i), h_{\theta}(x_j))$$

$h_{\theta}(\cdot)$ : Target hidden layer

$\mathcal{D}(\cdot)$ : Distance metric

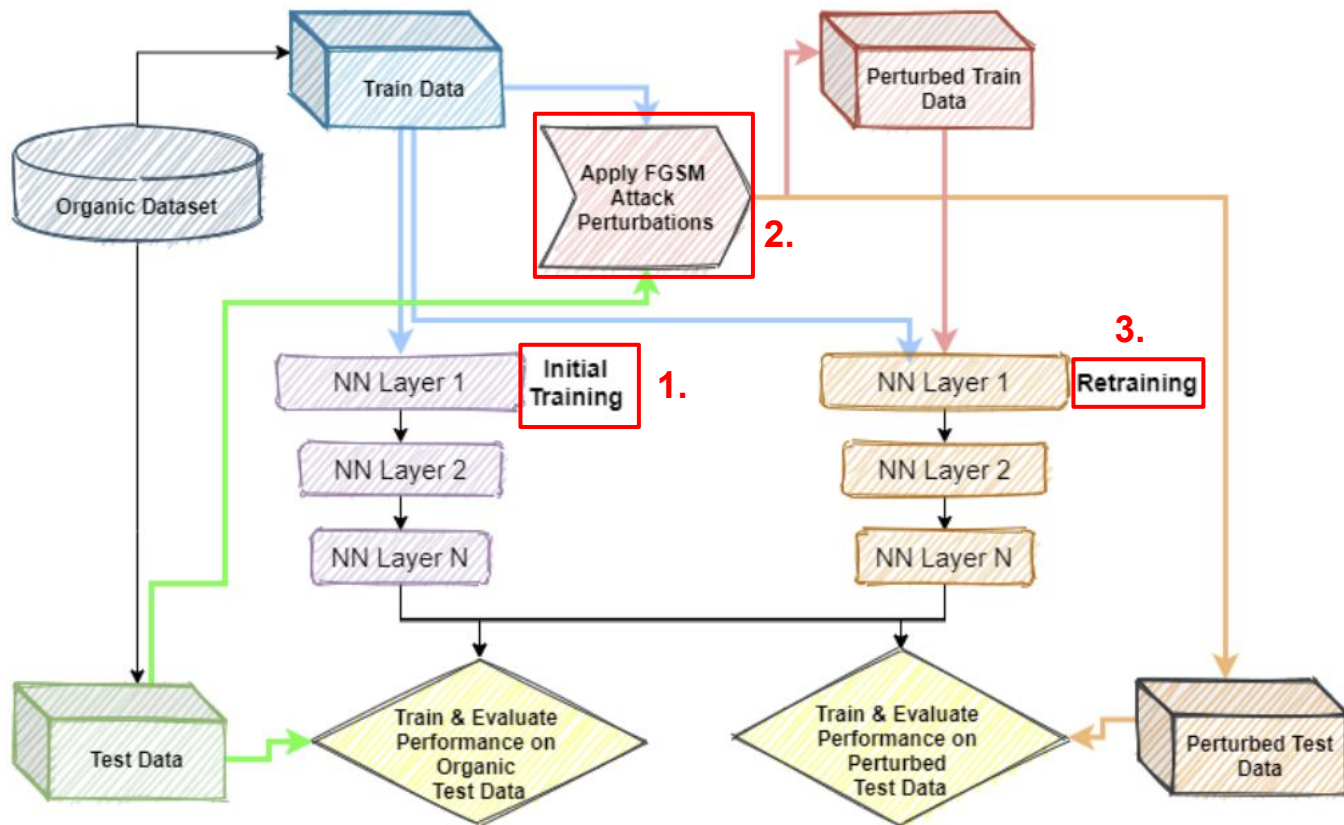
Examples: L1, L2, ...

# Adv. training

We could also (Madry et al., ICLR'18):

- Train a classifier on the clean inputs.
- Use the classifier to generate a perturbed set with FGSM.
- Retrain the classifier on clean + perturbed inputs.

# Adv. training





# Adv. training

- All the training methodologies are defined by the inner maximization i.e. the attack model.
- So, if we (adv.) train a model with ( $\ell_\infty$ ), will it generalize to other attack models?
- Sometimes, **yes**, sometimes **no**.

Training	Union	Unseen mean	Narrow threat models						NPTM	
			Clean	$L_\infty$	$L_2$	JPEG	StAdv	ReColor	PPGD	LPA
Normal	0.0	0.1	89.1	0.0	0.0	0.0	0.0	2.4	0.0	0.0
$L_\infty$	0.5	11.3	81.7	55.7	3.7	10.8	4.6	37.5	1.5	0.0
$L_2$	12.3	31.5	75.3	46.1	41.0	56.6	22.8	31.2	22.0	0.5
JPEG	0.1	7.4	84.8	13.7	1.8	74.8	0.3	21.0	0.5	0.0
StAdv	0.6	2.1	77.1	2.6	1.2	3.7	65.3	2.9	0.6	0.0
ReColorAdv	0.0	0.1	90.1	0.2	0.0	0.1	0.0	69.3	0.0	0.0
All (random)	0.9	—	78.6	38.3	26.4	61.3	1.4	32.5	16.1	0.2
PAT-self	<b>32.5</b>	<b>46.4</b>	72.6	45.0	37.7	53.0	51.3	45.1	29.2	<b>2.4</b>
PAT-AlexNet	25.5	44.7	75.7	46.8	41.0	55.9	39.0	40.8	<b>31.1</b>	1.6

Laidlaw et al., ICLR'21

# Adv. training

- All the training methodologies are defined by the inner maximization i.e. the attack model.
- So, if we (adv.) train a model with ( $\ell_\infty$ ), will it generalize to other attack models?
- Sometimes, **yes**, sometimes **no**.

Training	Union	Unseen mean	Narrow threat models						NPTM	
			Clean	$L_\infty$	$L_2$	JPEG	StAdv	ReColor	PPGD	LPA
Normal	0.0	0.1	89.1	0.0	0.0	0.0	0.0	2.4	0.0	0.0
$L_\infty$	0.5	11.3	81.7	55.7	3.7	10.8	4.6	37.5	1.5	0.0
$L_2$	12.3	31.5	75.3	46.1	41.0	56.6	22.8	31.2	22.0	0.5
JPEG	0.1	7.4	84.8	13.7	1.8	74.8	0.3	21.0	0.5	0.0
StAdv	0.6	2.1	77.1	2.6	1.2	3.7	65.3	2.9	0.6	0.0
ReColorAdv	0.0	0.1	90.1	0.2	0.0	0.1	0.0	69.3	0.0	0.0
All (random)	0.9	—	78.6	38.3	26.4	61.3	1.4	32.5	16.1	0.2
PAT-self	<b>32.5</b>	<b>46.4</b>	72.6	45.0	37.7	53.0	51.3	45.1	29.2	<b>2.4</b>
PAT-AlexNet	25.5	44.7	75.7	46.8	41.0	55.9	39.0	40.8	<b>31.1</b>	1.6

Laidlaw et al., ICLR'21

# Adv. training

- Since human perception is hard to characterize precisely, this lack of transfer become inevitable.
- So, what if we could incorporate a measure that gets us closer to the human perception?
- Use Learned Perceptual Image Patch Similarity (**LPIPS**) metric (Zhang et al., CVPR 2018)

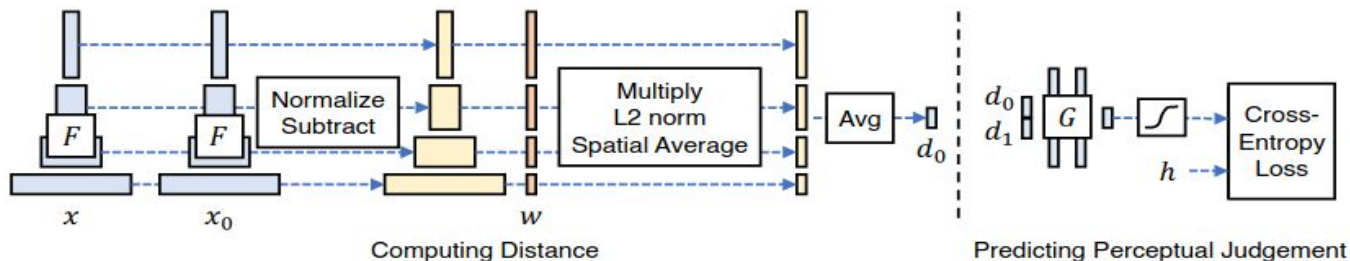


Figure 3: **Computing distance from a network** (Left) To compute a distance  $d_0$  between two patches,  $x, x_0$ , given a network  $F$ , we first compute deep embeddings, normalize the activations in the channel dimension, scale each channel by vector  $w$ , and take the  $\ell_2$  distance. We then average across spatial dimension and across all layers. (Right) A small network  $G$  is trained to predict perceptual judgement  $h$  from distance pair  $(d_0, d_1)$ .

# Adv. training

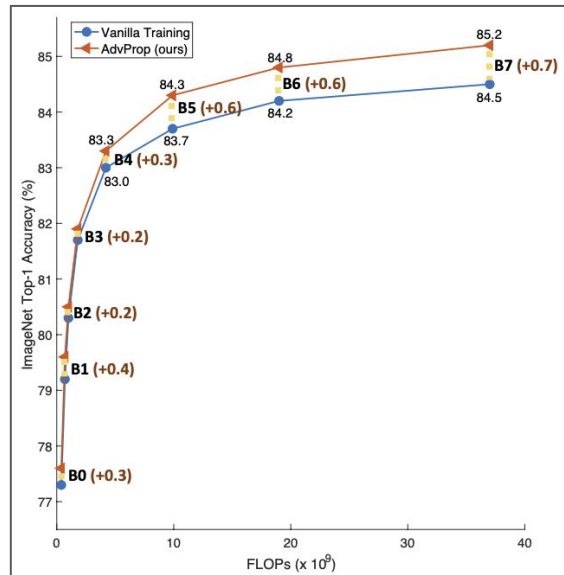
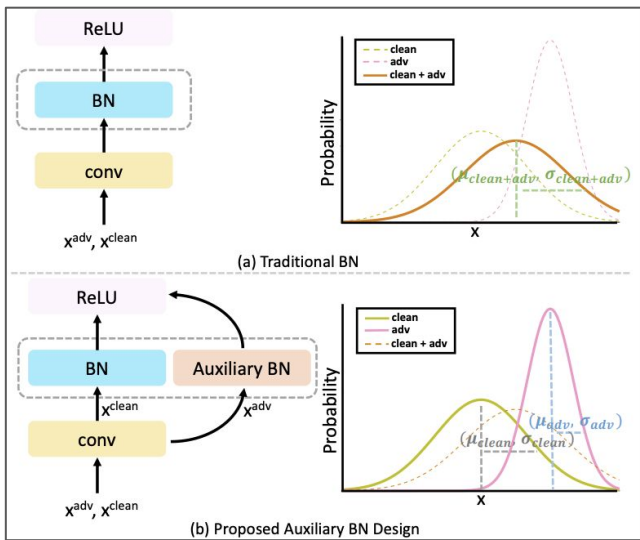
- Since human perception is hard to characterize precisely, this lack of transfer become inevitable.
- So, what if we could incorporate a measure that gets us closer to the human perception?
- Use Learned Perceptual Image Patch Similarity (**LPIPS**) metric (Laidlaw et al., ICLR'21)

Training	Union	Unseen mean	Narrow threat models						NPTM	
			Clean	$L_\infty$	$L_2$	JPEG	StAdv	ReColor	PPGD	LPA
Normal	0.0	0.1	89.1	0.0	0.0	0.0	0.0	2.4	0.0	0.0
$L_\infty$	0.5	11.3	81.7	55.7	3.7	10.8	4.6	37.5	1.5	0.0
$L_2$	12.3	31.5	75.3	46.1	41.0	56.6	22.8	31.2	22.0	0.5
JPEG	0.1	7.4	84.8	13.7	1.8	74.8	0.3	21.0	0.5	0.0
StAdv	0.6	2.1	77.1	2.6	1.2	3.7	65.3	2.9	0.6	0.0
ReColorAdv	0.0	0.1	90.1	0.2	0.0	0.1	0.0	69.3	0.0	0.0
All (random)	0.9	—	78.6	38.3	26.4	61.3	1.4	32.5	16.1	0.2
PAT-self	<b>32.5</b>	<b>46.4</b>	72.6	45.0	37.7	53.0	51.3	45.1	29.2	<b>2.4</b>
PAT-AlexNet	25.5	44.7	75.7	46.8	41.0	55.9	39.0	40.8	<b>31.1</b>	1.6

Laidlaw et al., ICLR'21

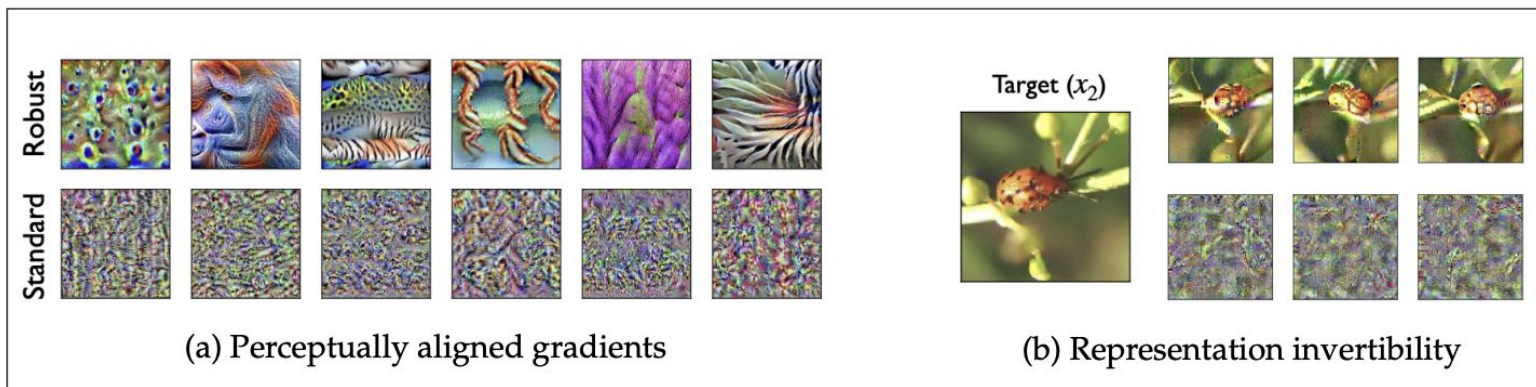
# Byproducts of adv. training

Adv. examples can help improve image recognition performance (Xie et al., CVPR'20).



# Byproducts of adv. training

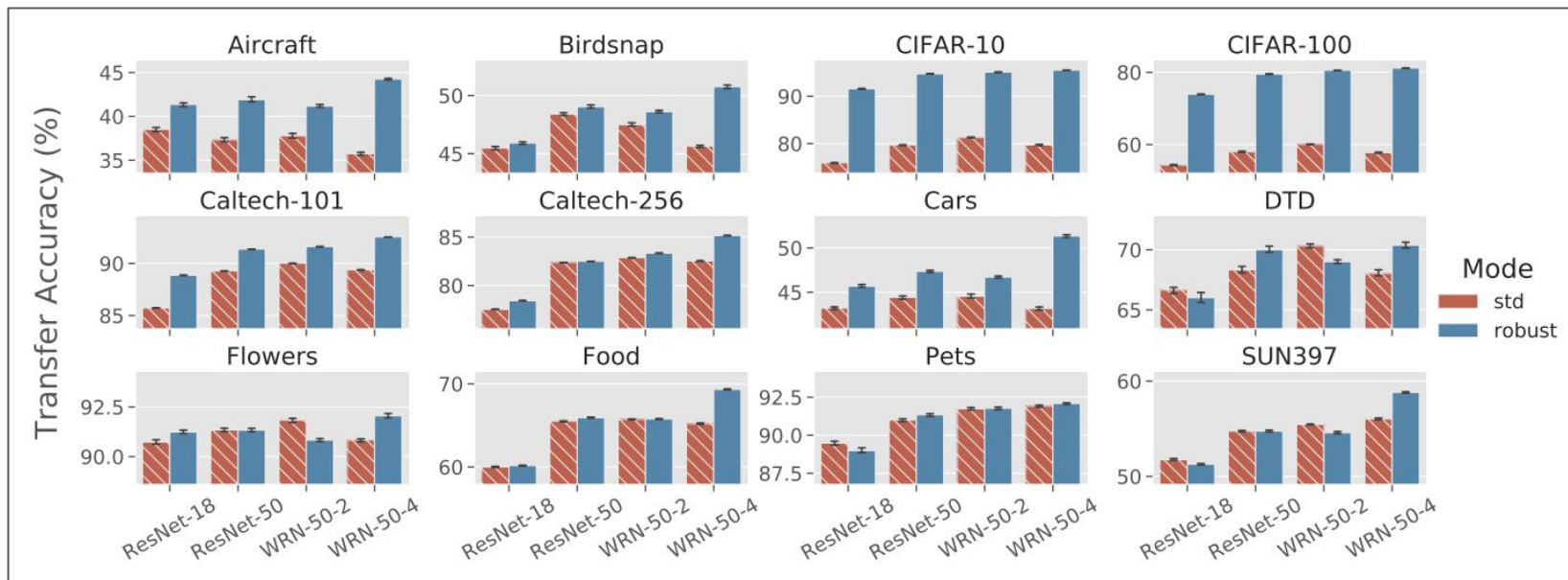
Adv. robust models transfer better (Salman et al., NeurIPS'20) for better feature representations.



Salman et al., NeurIPS'20

# Byproducts of adv. training

Adv. robust models transfer better (Salman et al., NeurIPS'20) for better feature representations.

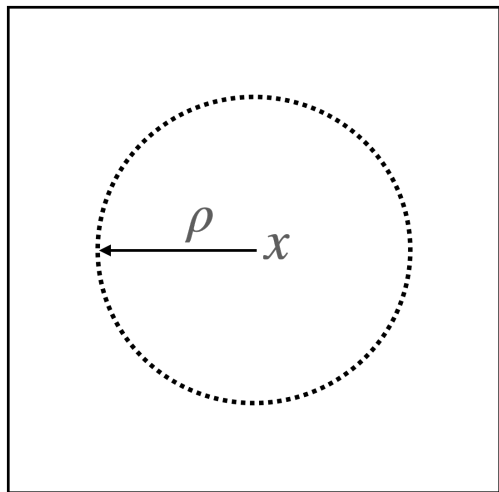


Salman et al., NeurIPS'20



# Defending with certified robustness

Classifier  $f_\theta$  is said to be certifiably robust if



$$f_\theta(\mathbf{x}) = f_\theta(\mathbf{x}')$$

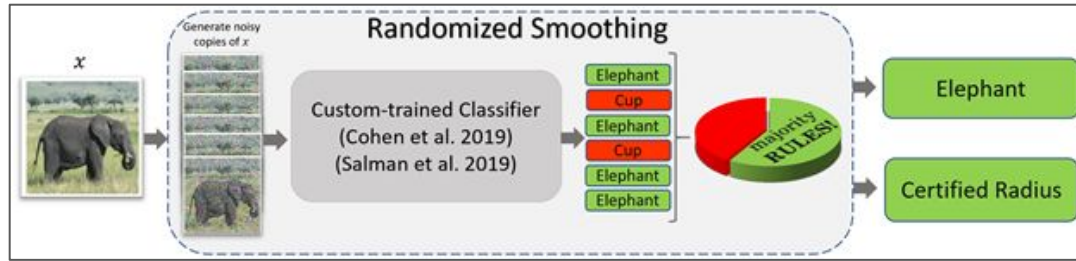
$$\forall \mathbf{x}' \in \mathcal{T}(\mathbf{x}, \rho)$$

↓  
Certification radii

# Defending with certified robustness

Randomized smoothing (Cohen et al., ICML'19) is a widely used method for obtaining certified robustness against  $\ell_2$  attacks.

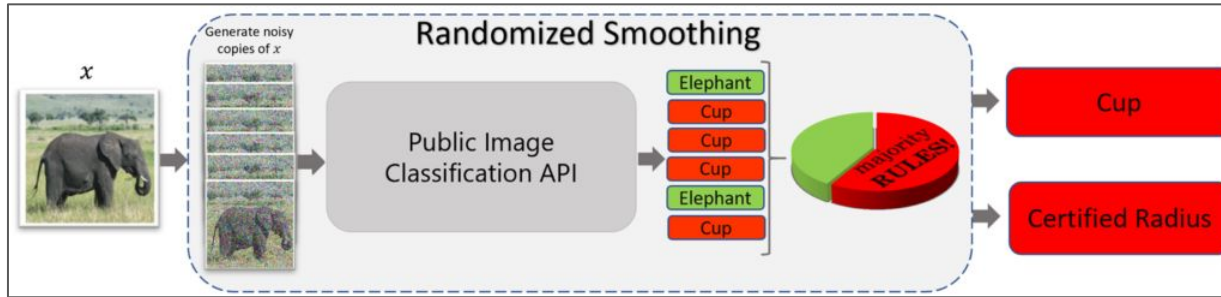
$$g(x) = \arg \max_{c \in \mathcal{Y}} \mathbb{P}[f(x + \delta) = c] \quad \text{where } \delta \sim \mathcal{N}(0, \sigma^2 I)$$



Salman et al., NeurIPS'20

# Defending with certified robustness

- Randomized smoothing requires that a classifier performs well under isotropic Gaussian perturbations.
- What if we wanted to work with standard pre-trained models, public vision APIs having only query access?

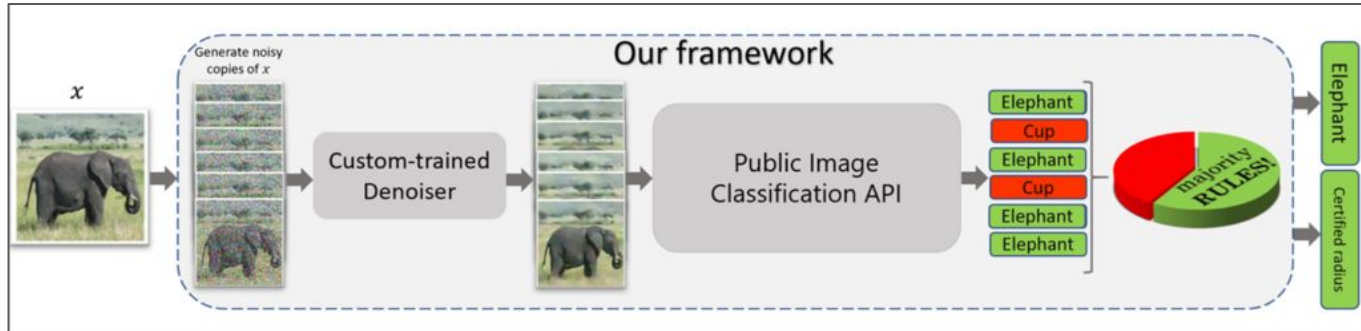


Salman et al., NeurIPS'20

# Defending with certified robustness

Enter denoised smoothing (Salman et al, NeurIPS'20).

- Apply the same Gaussian noise to the inputs.
- Pass it through a pre-trained denoiser.
- Pass the denoised inputs to the pre-trained model/public API and take majority voting.

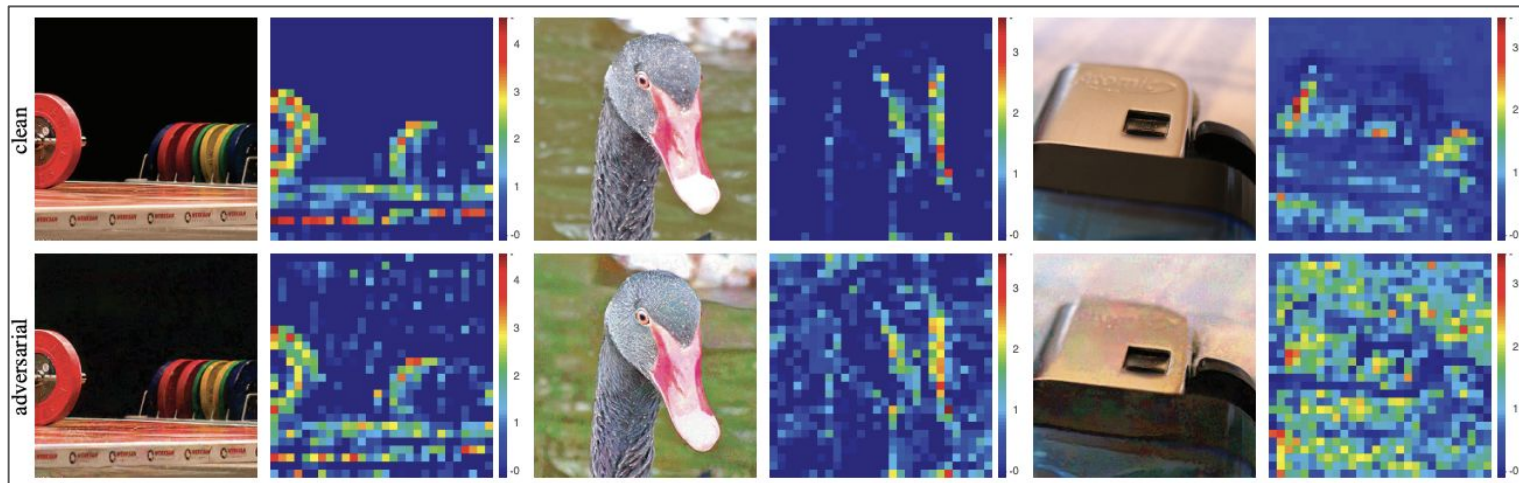


Salman et al., NeurIPS'20

## 6. Interpreting adversarial examples

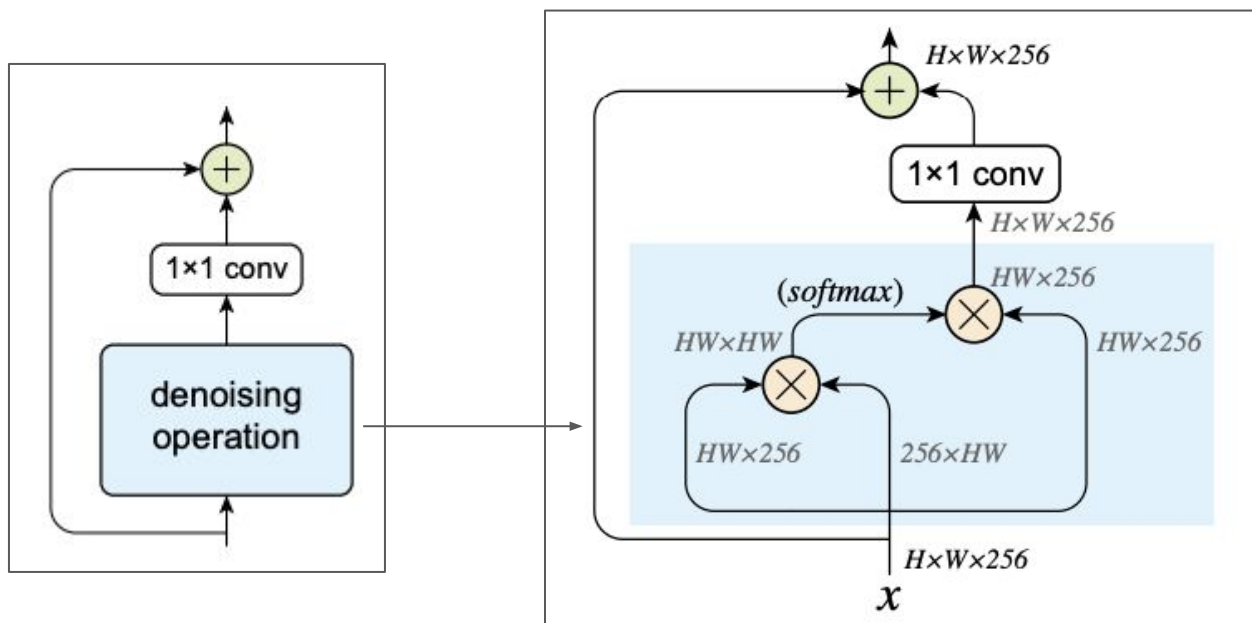
# Noisy feature space

- Adversarial examples introduce noise in the network feature space (Xie et al., CVPR'19).
- Therefore, irrelevant regions in the feature space get activated latching networks into spurious correlations.



# Noisy feature space

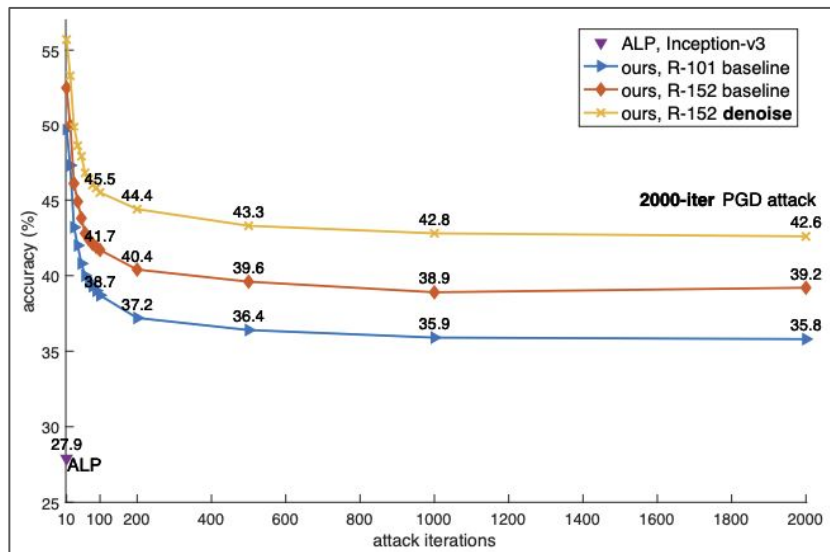
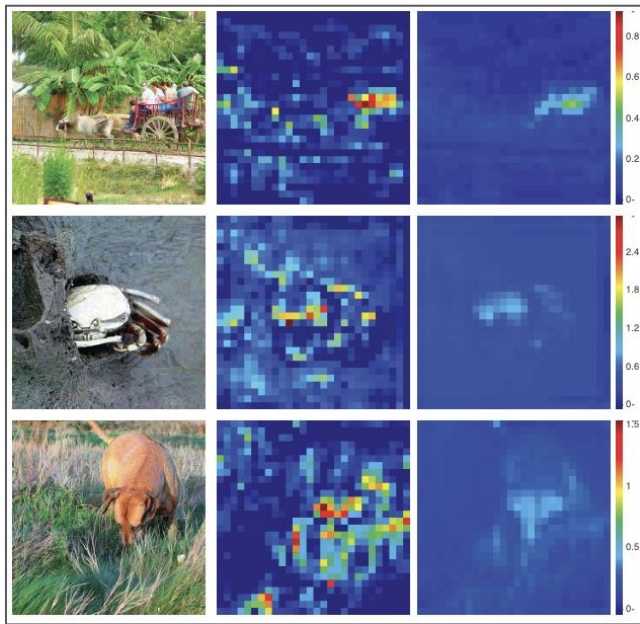
What if we add a denoiser block inside the networks?





# Noisy feature space

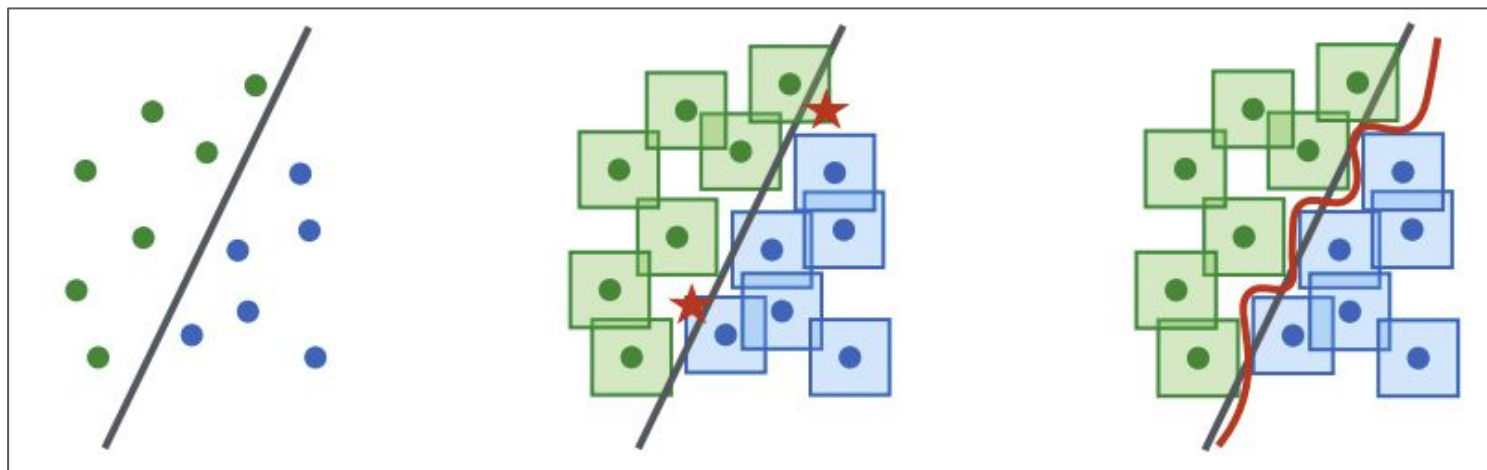
Do we get any benefits if the underlying network has denoising capabilities?



## 7. Promising recipes

# Model capacity is crucial

Adversarial examples change the decision boundary to a more complicated one (Madry et al., ICLR'18).

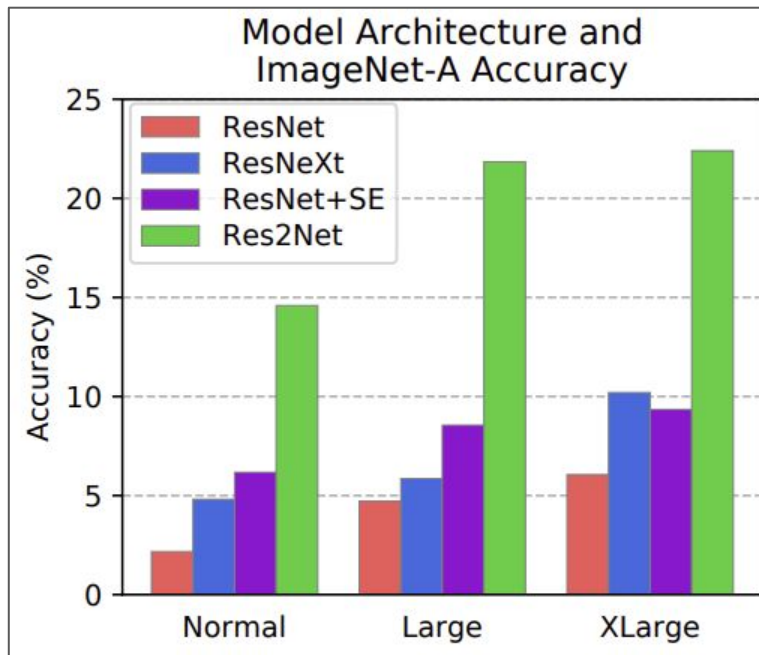


Madry et al., ICLR'18

■ ■  $\ell_\infty$ -balls  
★ adv. examples

# Model capacity is crucial

On ImageNet-A, Hendrycks et al. also confirms this.



Hendrycks et al., CVPR'21

## Self-attention provides improved robustness

- On ImageNet-P, ViT (Dosovitskiy et al., ICLR'21) performs significantly better.
- Model capacity and longer pre-training with a larger dataset are paramount too.

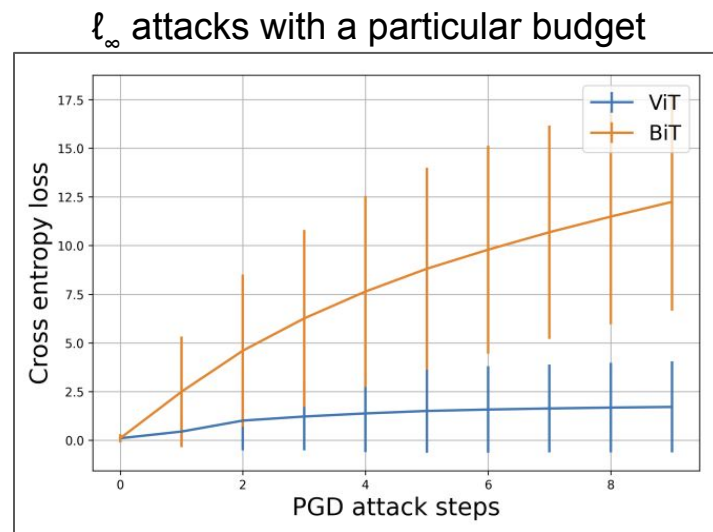
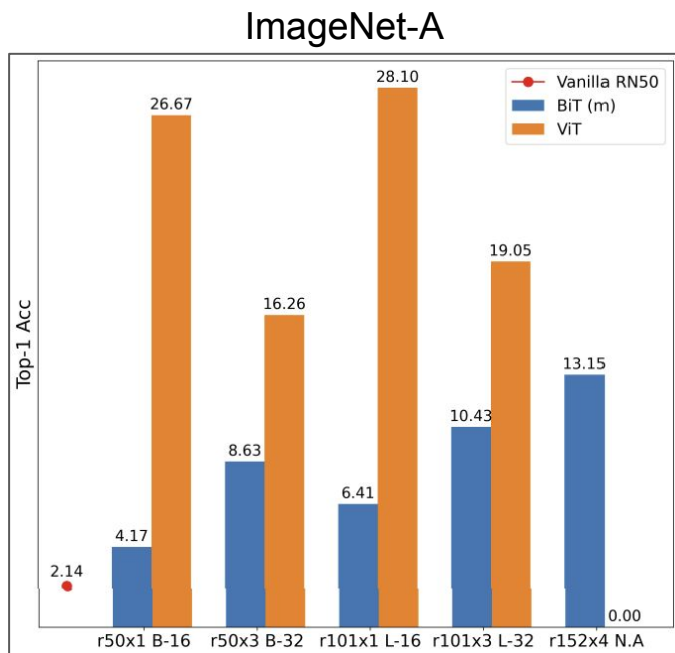
**Table 4: mFRs (%) and mT5Ds (%) on ImageNet-P dataset (lower is better).**

<b>Model / Method</b>	<b>mFR</b>	<b>mT5D</b>
ResNet-50	58	82
BiT-m r101x3	49.99	76.71
AugMix [37]	37.4	NA
ViT L-16	33.064	50.15

Paul et al., arXiv, 2021

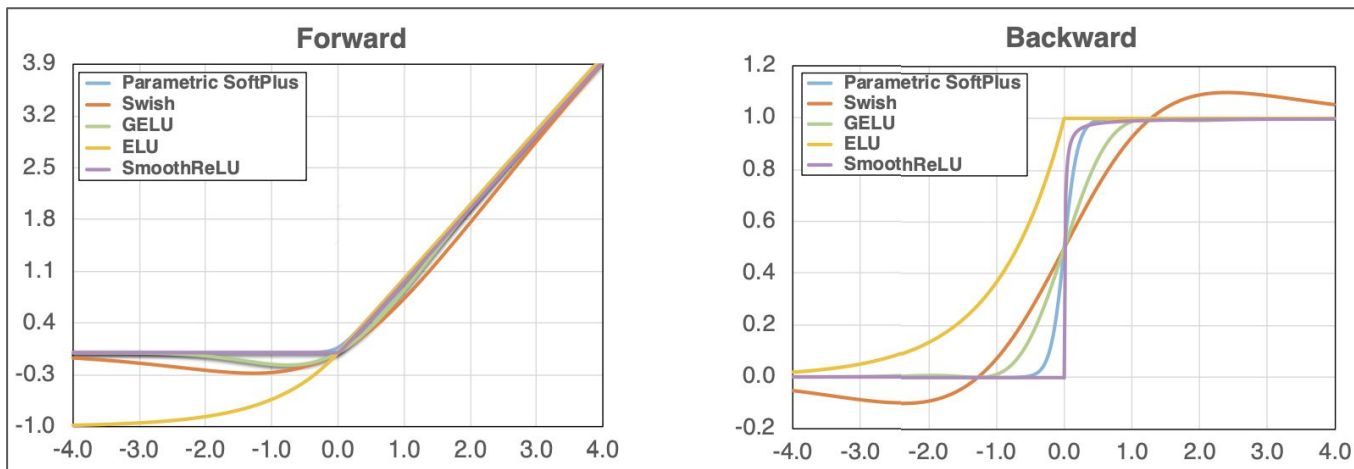
# Self-attention provides improved robustness

Continuation of the previous discussion -



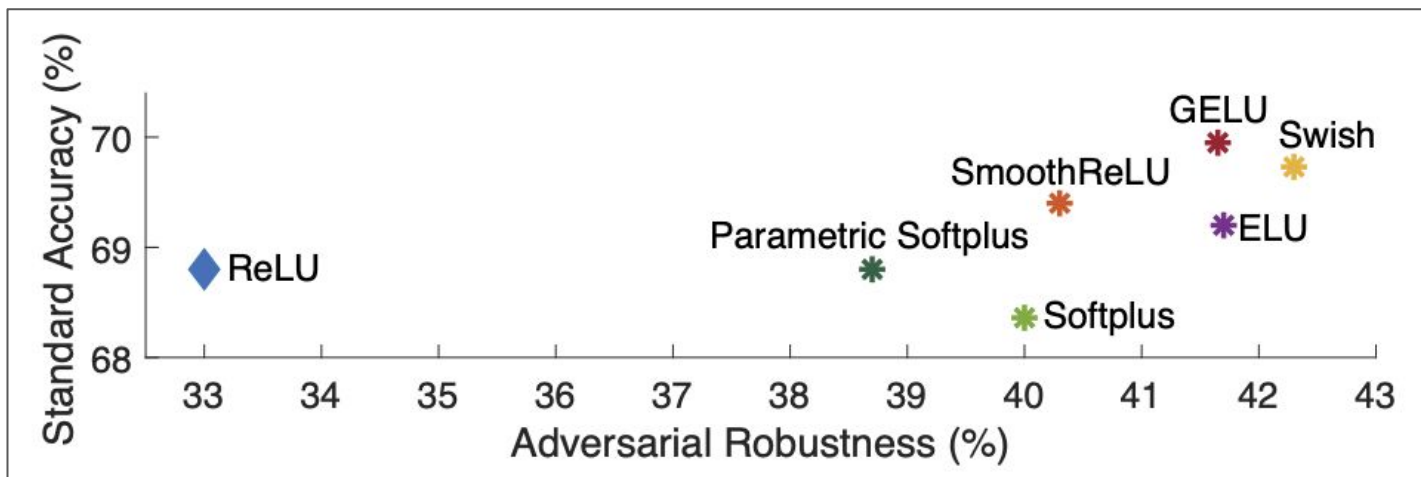
# Smooth adv. training

- Using ReLU during adv. training is particularly worse off because of its non-smooth nature.
- Smother activation functions (Swish, SoftPlus, etc.) result into better informed gradients because of their smoothness.



## Smooth adv. training

The use of smoother activation functions leads to improved performance without accuracy loss.

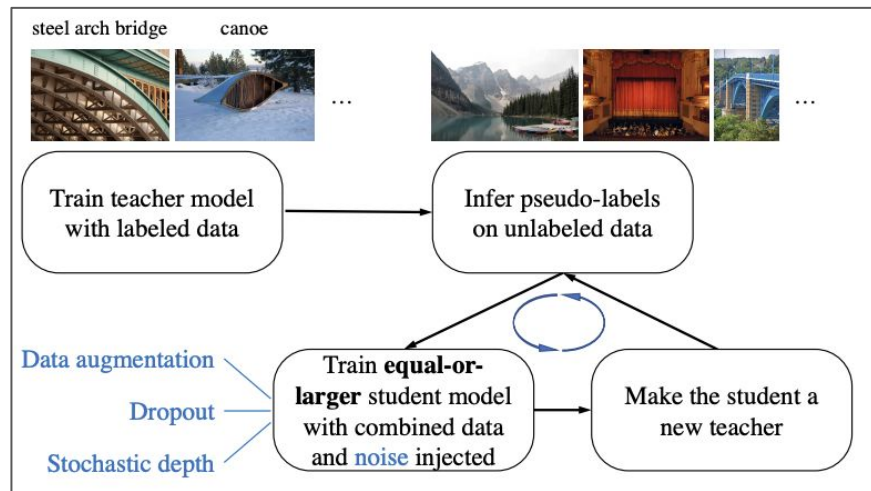


Xie et al., arXiv, 2020



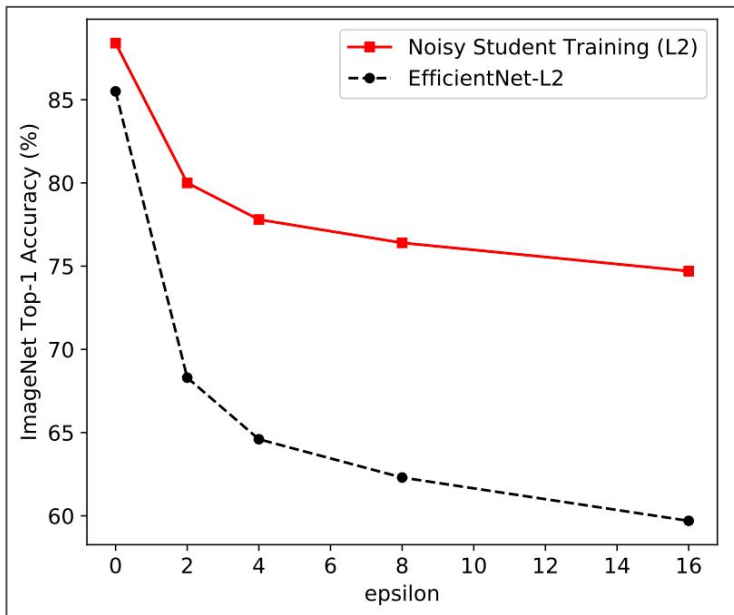
# Noisy student training

- Train a good teacher model.
- Train the student to match the pre-computed teacher predictions (targets) on clean images and its own predictions on the same but noisy augmented images.



# Noisy student training

It does not include any explicit adv. training objective but yields good performance against PGD attacks.



Xie et al., CVPR'20

# Conclusion

## Being aware is helpful

- For **model developers**, adversarial examples can be used for robustness evaluation and model improvement.
- For **business stakeholders**, lacking adversarial robustness in your AI model could bring unexpected negative impacts.
- For **end users**, gaining awareness of adversarial robustness for the AI service your are using is crucial.

# Takeaways

- What is adversarial AI ?
- Various adversarial attacks (FGSM and PGD).
- Different formulations of adversarial training
- Effect of adversarial training
- Recipes that works in practice for robust models.

Original tutorial materials are here: [bit.ly/par-2021](https://bit.ly/par-2021).