The Robustness Problem

Justin Gilmer

Table of Contents

1. **Overly optimistic IID test sets**

2. Robustness, security and adversarial examples

3. Why are models so brittle?

Google

The Deep Learning Boom

transportation

Medical diagnosis

recommender systems

Robotics

Hype!

Artificial intelligence rivals radiologists in screening X-rays for certain diseases

Man against machine: AI is better than dermatologists at diagnosing skin cancer

Google’s lung cancer detection AI outperforms 6 human radiologists

More Hype!

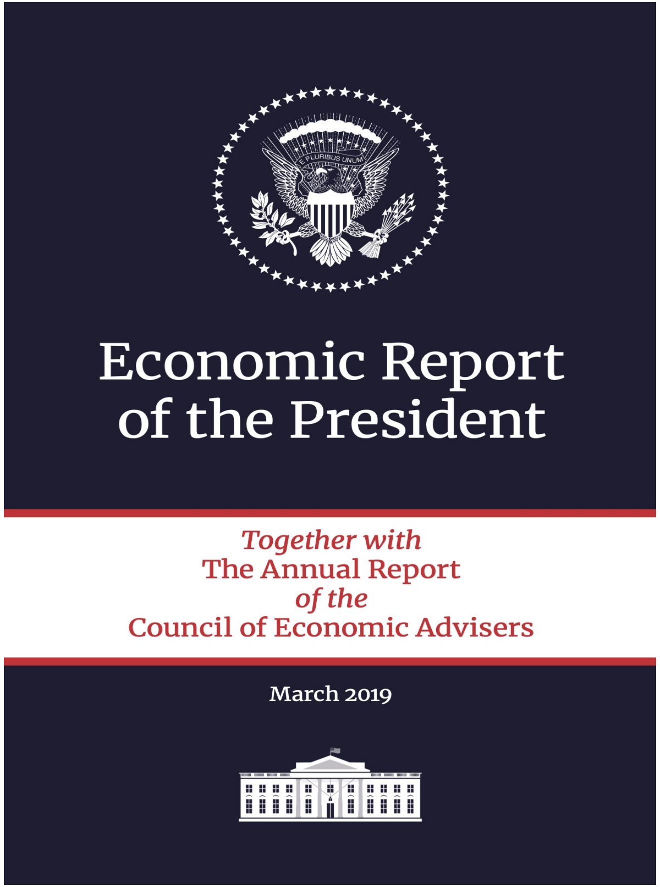
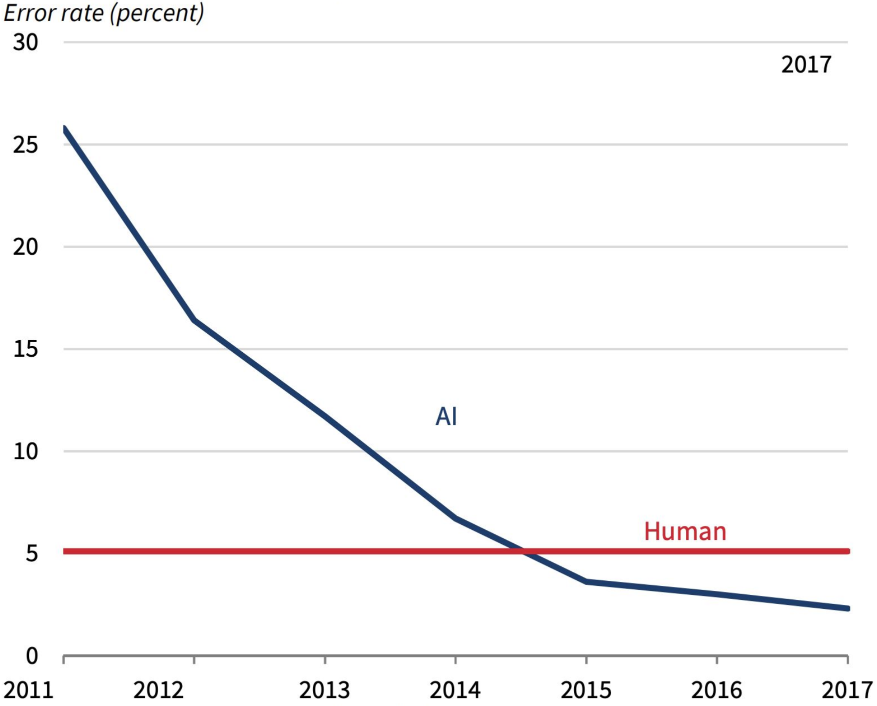


Figure 7-1. Error Rate of Image Classification by Artificial Intelligence and Humans, 2010–17



Sources: Russakovsky et al. (2015); CEA calculations.

The Biggest Lie in Machine Learning

$$P(train) = P(test)$$

Independent **I**dentically **D**istributed (IID)

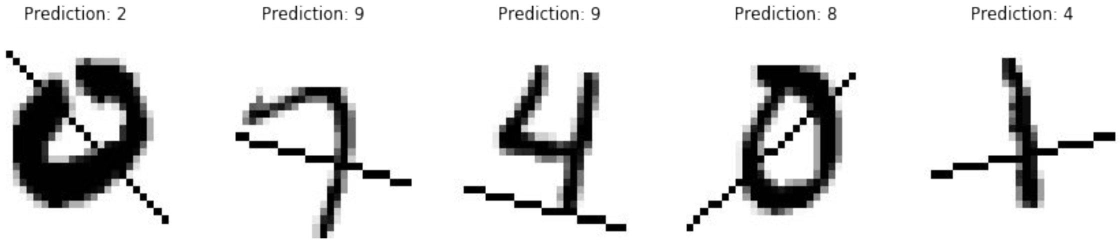
- MNIST
- CIFAR-10
- Imagenet
- SVHN
- Fashion MNIST
- COCO
- ...

Reality Check

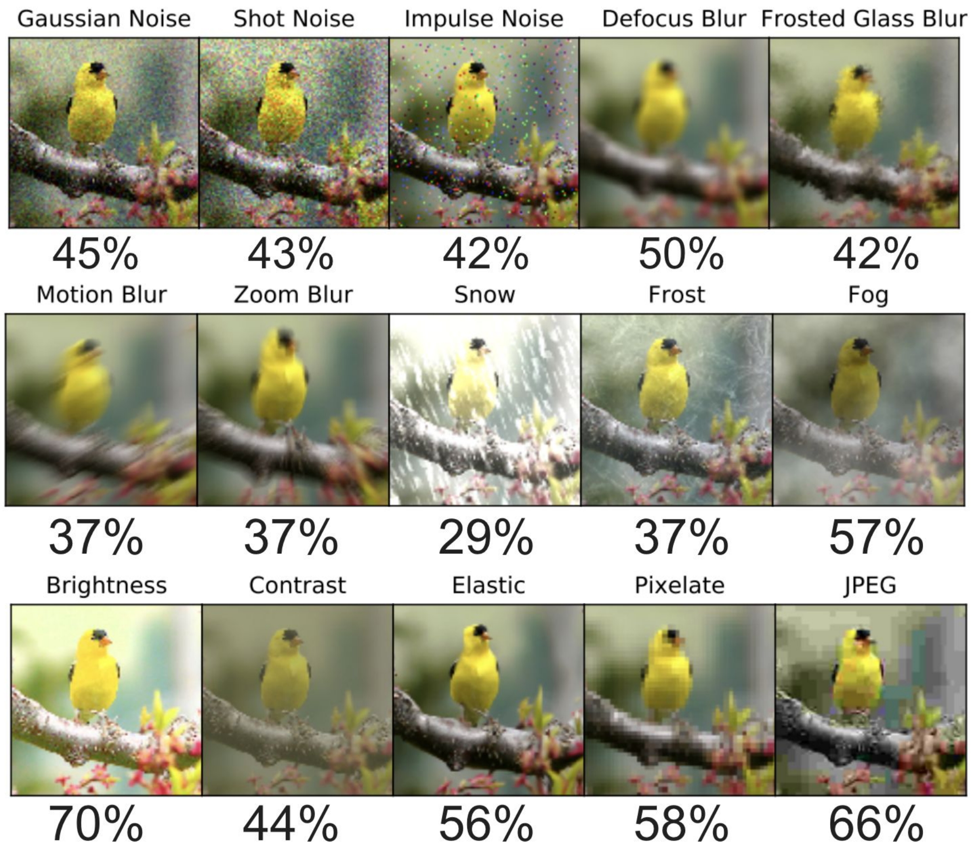
- IID test sets grossly overestimate performance in the real world.
- Models are not robust to even slight changes in distribution.

In distribution - 99% Accuracy

Out of distribution - 63% Accuracy



The Real World is **Not** IID



Resnet-50
76% Top-1 Accuracy (IID)

[Hendrycks et. al] https://arxiv.org/abs/1807.01697

Distribution Shift is a Real Problem!

(A) **Cow: 0.99**, Pasture: 0.99, Grass: 0.99, No Person: 0.98, Mammal: 0.98

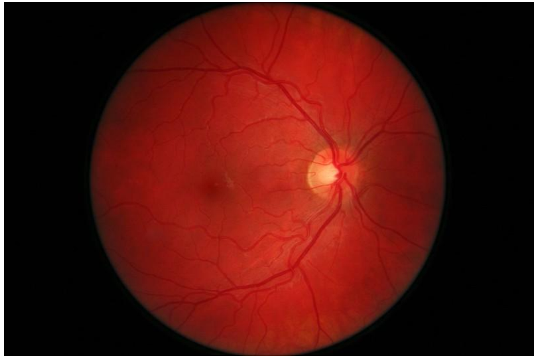
(B) No Person: 0.99, Water: 0.98, Beach: 0.97, Outdoors: 0.97, Seashore: 0.97

(C) No Person: 0.97, **Mammal: 0.96**, Water: 0.94, Beach: 0.94, Two: 0.94

[Beery et. al.] http://openaccess.thecvf.com/content_ECCV_2018/papers/Beery_Recognition_in_Terra_ECCV_2018_paper.pdf

Medical Imaging on a Cell Phone Camera?

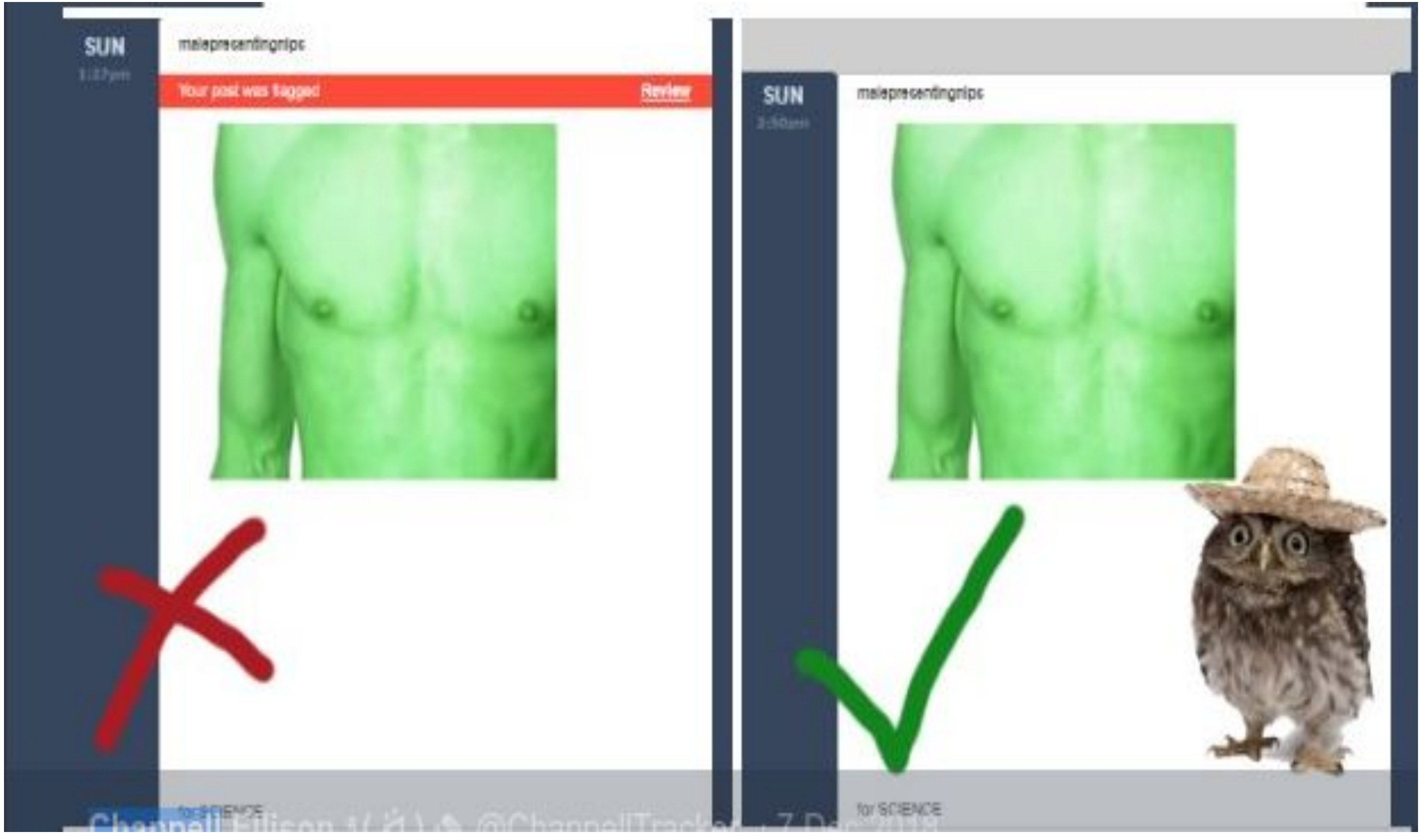
Train on high quality images
taken in controlled settings.



Deploy on camera phones

Google

Adversaries Can Exploit this Lack of Robustness



Robustness Benchmarks

- Image corruptions
  - Imagenet-C: [Hendrycks et. al.] https://arxiv.org/abs/1807.01697
  - MNIST-C: [Mu, Gilmer] https://arxiv.org/abs/1906.02337

- Natural distribution shifts
  - Imagenet-A [Hendrycks et. al.] https://arxiv.org/abs/1907.07174
  - ImagenetV2 [Recht et. al.] https://arxiv.org/abs/1902.10811
  - Imagenet-Vid-Robust [Shankar et. al] https://arxiv.org/pdf/1906.02168.pdf.
  - Video Robustness [Gu et. al.] https://arxiv.org/pdf/1904.10076.pdf

For ML to work well, we need to drop the iid assumption.

Table of Contents

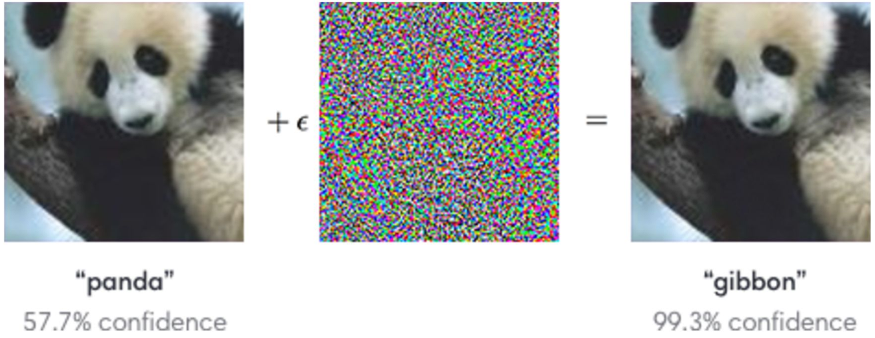
Google

Adversarial Examples

Security

vs

“Surprising” Phenomenon

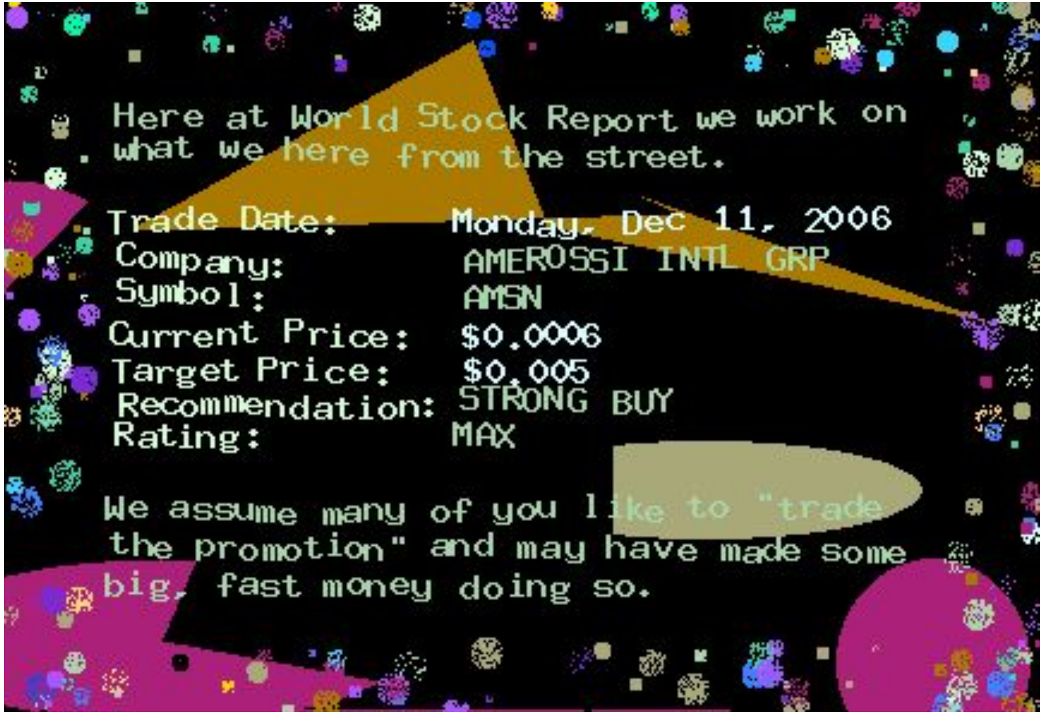


Goodfellow et. al. https://arxiv.org/abs/1412.6572

Google

[Gilmer et. al.] Motivating the Rules of the Game for Adversarial Example Research

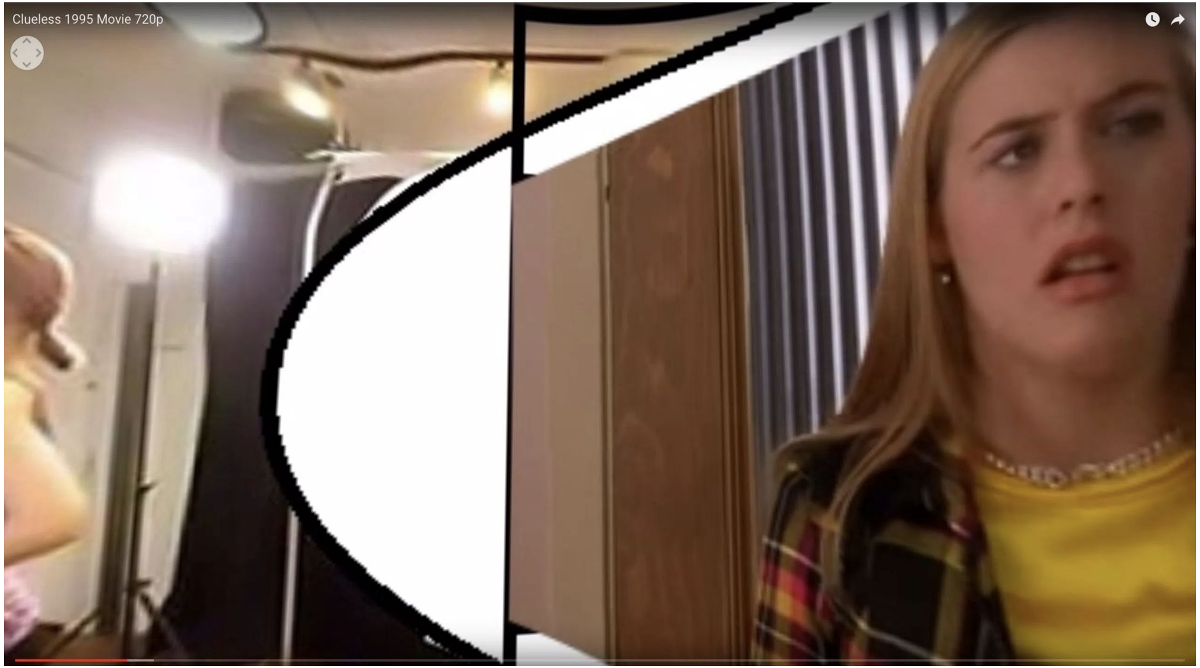
Adversarial Examples - Security



Biggio et. al: https://arxiv.org/abs/1712.03141

[Gilmer et. al.] Motivating the Rules of the Game for Adversarial Example Research

Adversarial Examples - Security

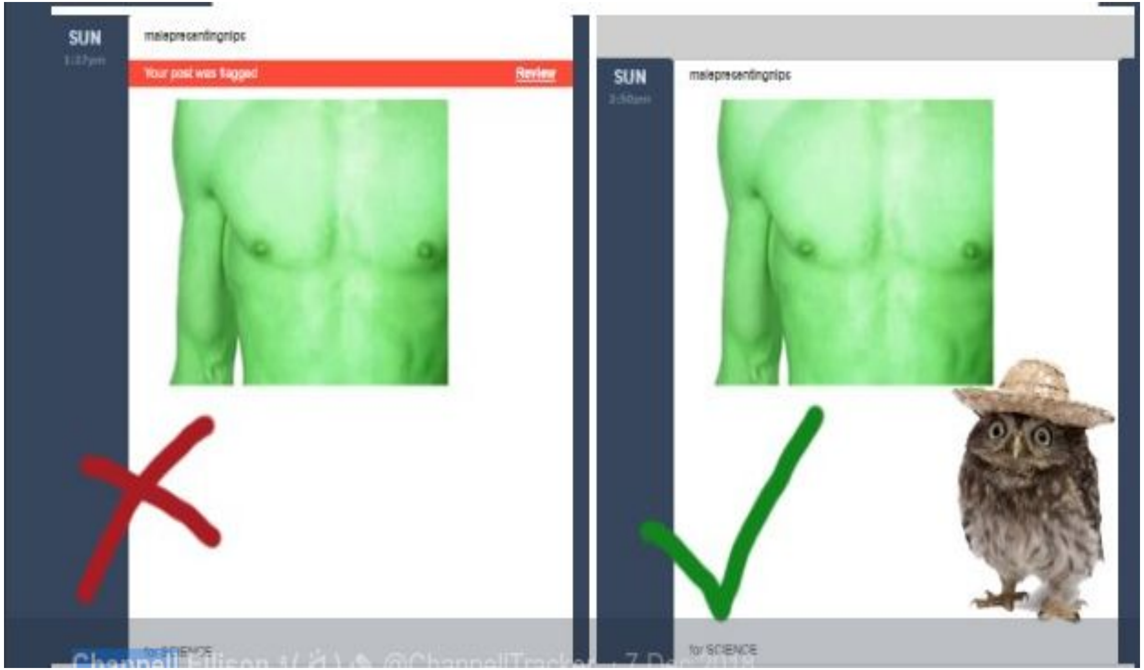


https://qz.com/721615/smart-pirates-are-fooling-youtubes-copyright-bots-by-hiding-movies-in-360-degree-videos/

[Gilmer et. al.] Motivating the Rules of the Game for Adversarial Example Research

Adversarial Examples - Security

"State of the art", zero knowledge, limited query, black box attack.
[Tumblr Quality Assurance, 2018]

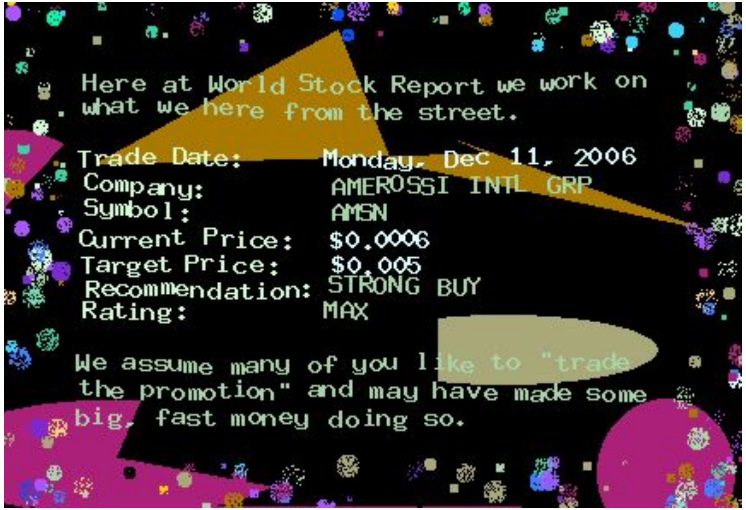


https://piunikaweb.com/2018/12/08/owl-pics-heres-how-tumblr-censor-bots-are-being-fooled/

[Gilmer et. al.] Motivating the Rules of the Game for Adversarial Example Research

Questions for Designing a Secure ML System

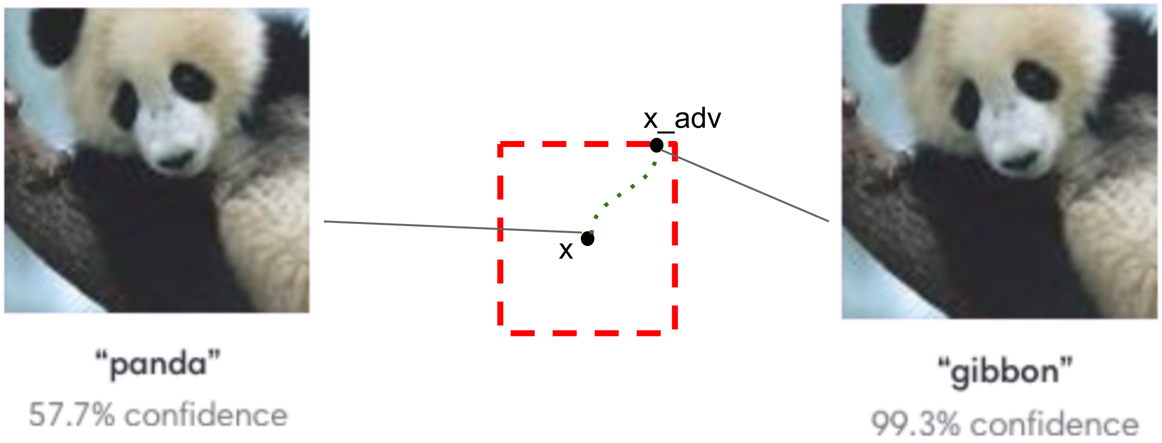
- How do adversaries typically break systems?
- How would you measure test error?
- Are you secure if test error > 0?
- How do we deal with out-of-distribution generalization?



[Gilmer et. al.] Motivating the Rules of the Game for Adversarial Example Research

Adversarial Examples - The "Surprising" Phenomenon

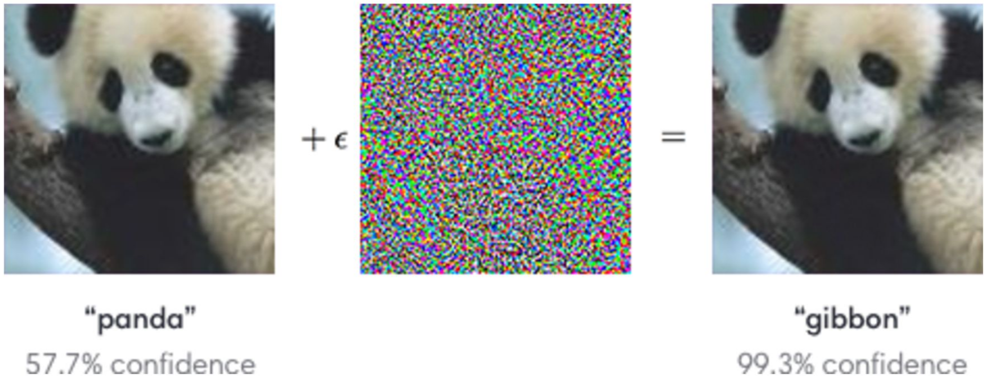
- In 2013 it was discovered that neural networks have “adversarial examples”.
- 2000+ papers written on this topic.



$$x_{adv} = \max_{x' : ||x - x'||_{\infty} < \epsilon} L(\theta, x', \hat{y})$$

Adversarial Examples - The Phenomenon

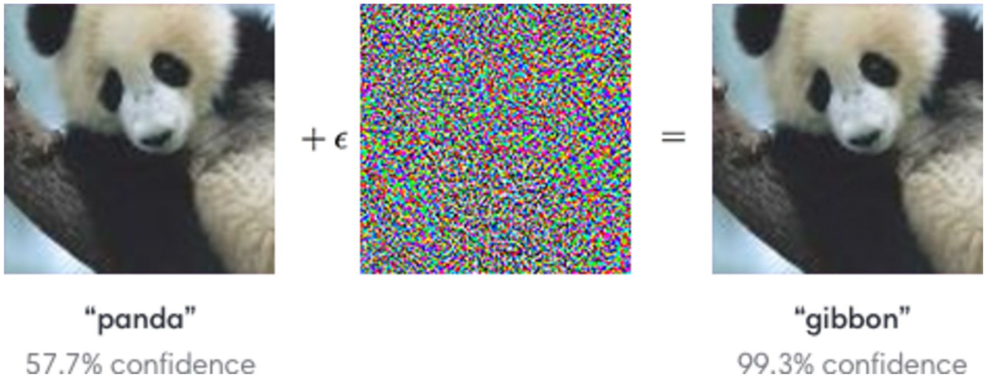
Why do our models have adversarial examples?



[Ford et. al.] Adversarial Examples are a Natural Consequence of Test Error in Noise (ICML 2018).

Adversarial Examples - The Phenomenon

Why do our models have adversarial examples? **A:** ???

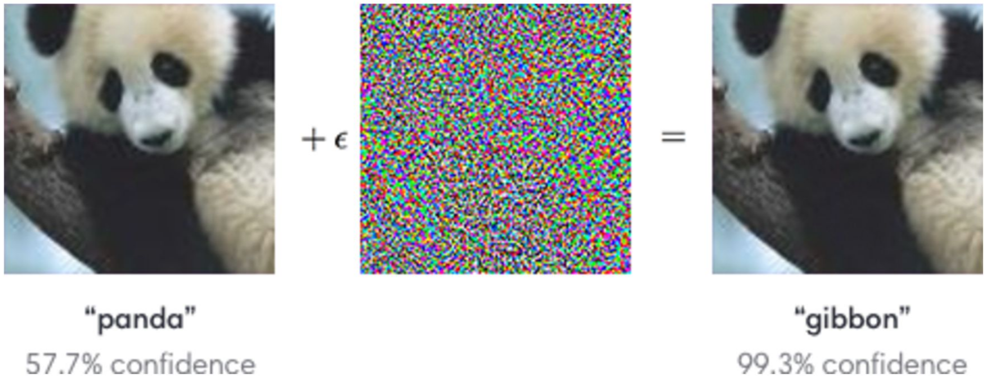


[Ford et. al.] Adversarial Examples are a Natural Consequence of Test Error in Noise (ICML 2018).

Adversarial Examples - The Phenomenon

Why do our models have adversarial examples? **A:** ???

What are adversarial examples?



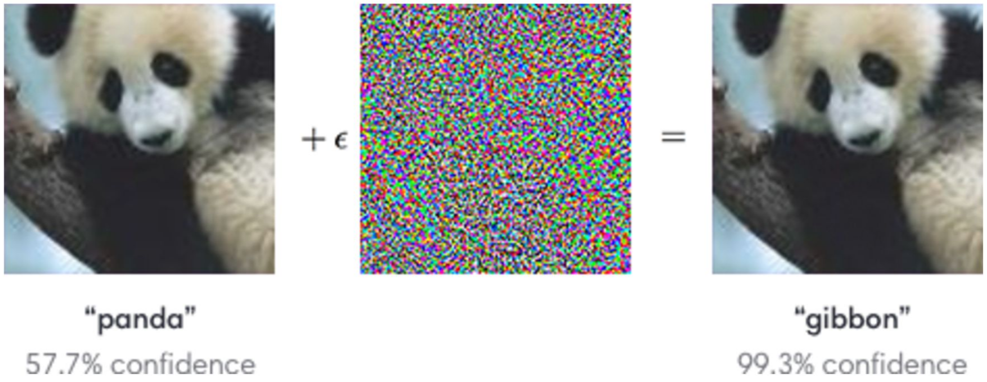
Google

[Ford et. al.] Adversarial Examples are a Natural Consequence of Test Error in Noise (ICML 2018).

Adversarial Examples - The Phenomenon

Why do our models have adversarial examples? **A:** ???

What are adversarial examples? **A:** The nearest error



Google

[Ford et. al.] Adversarial Examples are a Natural Consequence of Test Error in Noise (ICML 2018).

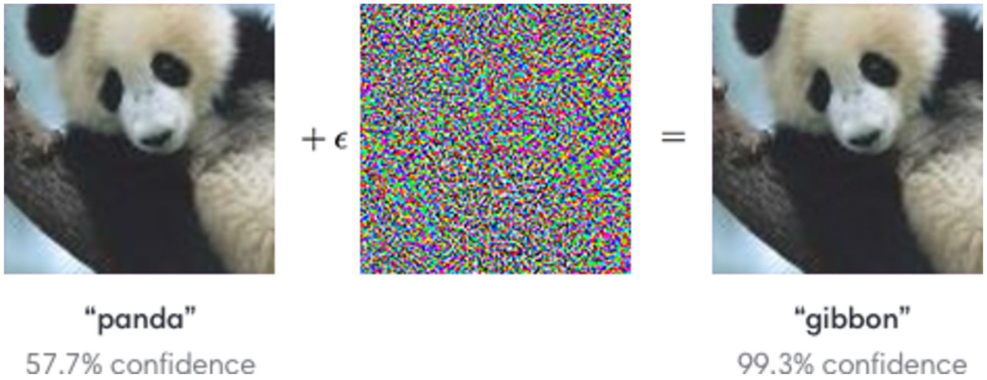
Adversarial Examples - The Phenomenon

Why do our models have ~~adversarial examples?~~

A: ???

What are adversarial examples?

A: The nearest error

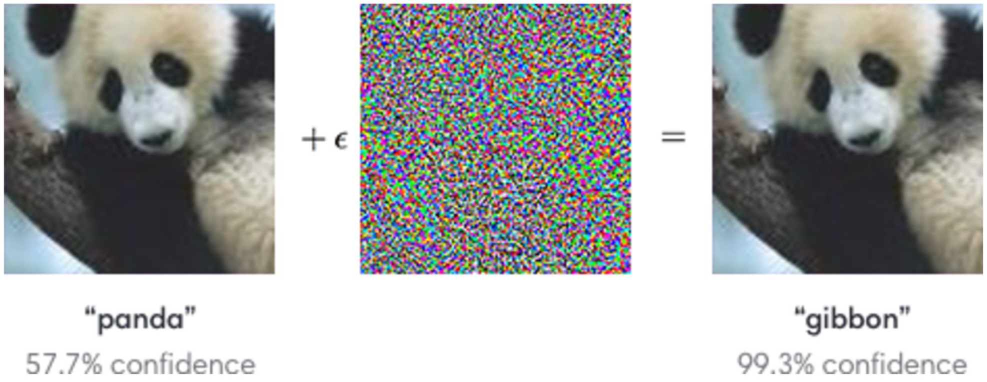


[Ford et. al.] Adversarial Examples are a Natural Consequence of Test Error in Noise (ICML 2018).

Adversarial Examples - The Phenomenon

Why do our models have (o.o.d) **test error?** **A:** ???

What are adversarial examples? **A:** The nearest error



Google

[Ford et. al.] Adversarial Examples are a Natural Consequence of Test Error in Noise (ICML 2018).

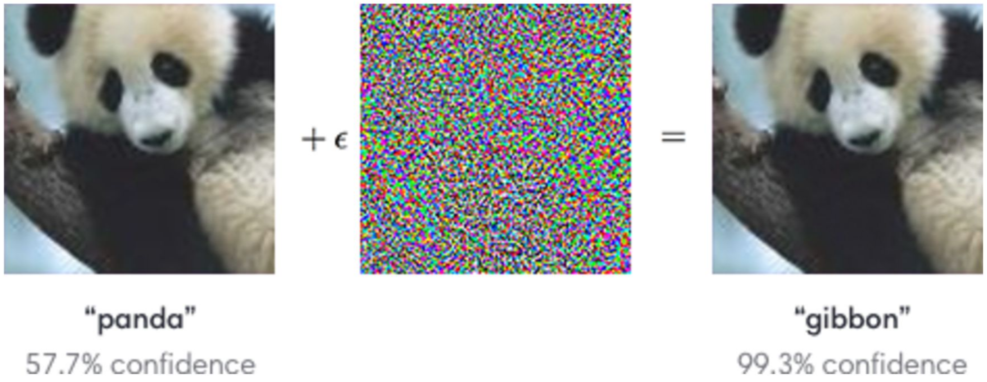
Adversarial Examples - The Phenomenon

Why do our models have (o.o.d) **test error?**

A: ???

What are adversarial examples?

A: The nearest error



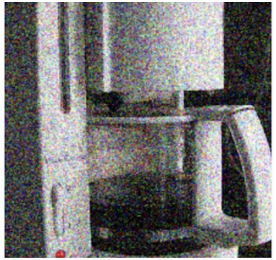
Test error > 0 (iid, ood) -> errors exist -> there is a nearest error

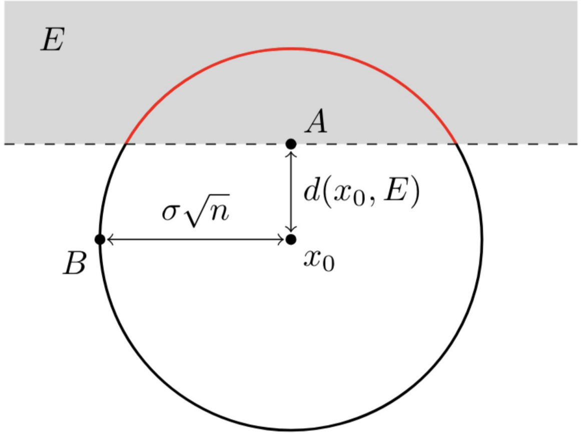
Google

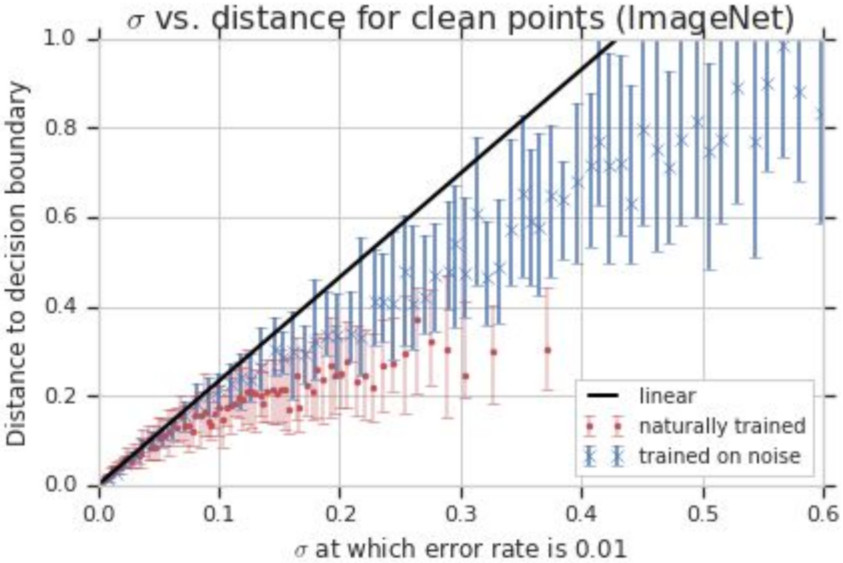
[Ford et. al.] Adversarial Examples are a Natural Consequence of Test Error in Noise (ICML 2018).

Linear Assumption

1% error rate on random perturbations of norm 79 => adv ex at norm .5





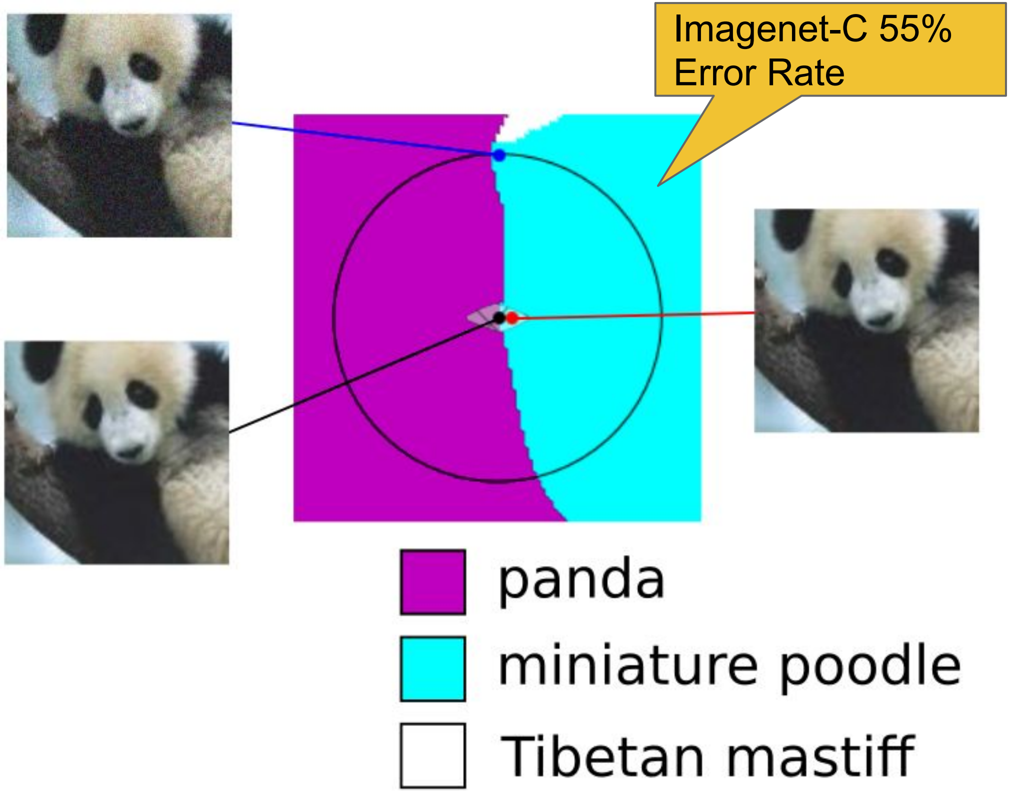


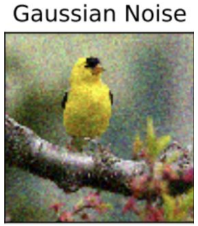
See also Fawzi et. al.

Google

[Ford et. al.] Adversarial Examples are a Natural Consequence of Test Error in Noise (ICML 2018).

InceptionV3 Decision Boundary





Google

[Ford et. al.] Adversarial Examples are a Natural Consequence of Test Error in Noise (ICML 2018).

Adversarial Defenses

L_∞-metric ($\epsilon = 0.3$)		
Transfer Attacks	0.08 / 0%	0.44 / 85%
FGSM	0.10 / 4%	0.43 / 77%
FGSM w/ GE	0.10 / 21%	0.42 / 71%
L_∞ DeepFool	0.08 / 0%	0.38 / 74%
L_∞ DeepFool w/ GE	0.09 / 0%	0.37 / 67%
BIM	0.08 / 0%	0.36 / 70%
BIM w/ GE	0.08 / 37%	∞ / 70%
MIM	0.08 / 0%	0.37 / 71%
MIM w/ GE	0.09 / 36%	∞ / 69%
All L_∞ Attacks	0.08 / 0%	0.34 / 64%

Adversarial Defenses

Why are we trying to "defend" against the nearest error?

L_∞-metric ($\epsilon = 0.3$)		
Transfer Attacks	0.08 / 0%	0.44 / 85%
FGSM	0.10 / 4%	0.43 / 77%
FGSM w/ GE	0.10 / 21%	0.42 / 71%
L_∞ DeepFool	0.08 / 0%	0.38 / 74%
L_∞ DeepFool w/ GE	0.09 / 0%	0.37 / 67%
BIM	0.08 / 0%	0.36 / 70%
BIM w/ GE	0.08 / 37%	∞ / 70%
MIM	0.08 / 0%	0.37 / 71%
MIM w/ GE	0.09 / 36%	∞ / 69%
All L_∞ Attacks	0.08 / 0%	0.34 / 64%

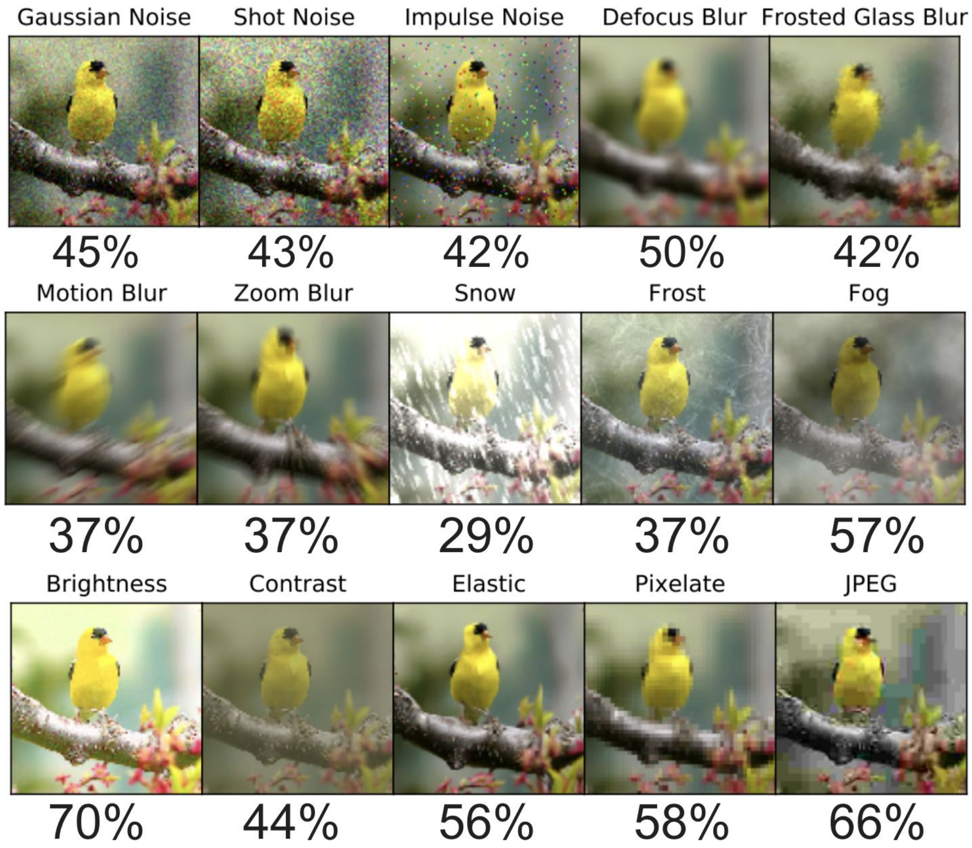
Adversarial Defenses

Why are we trying to "defend" against the nearest error?

Not a useful measure of robustness

L_∞-metric ($\epsilon = 0.3$)		
Transfer Attacks	0.08 / 0%	0.44 / 85%
FGSM	0.10 / 4%	0.43 / 77%
FGSM w/ GE	0.10 / 21%	0.42 / 71%
L_∞ DeepFool	0.08 / 0%	0.38 / 74%
L_∞ DeepFool w/ GE	0.09 / 0%	0.37 / 67%
BIM	0.08 / 0%	0.36 / 70%
BIM w/ GE	0.08 / 37%	∞ / 70%
MIM	0.08 / 0%	0.37 / 71%
MIM w/ GE	0.09 / 36%	∞ / 69%
All L_∞ Attacks	0.08 / 0%	0.34 / 64%

Takeaways



- We should not be surprised that there is a nearest error.
- **The problem to study is robustness to distribution shift.**

Table of Contents

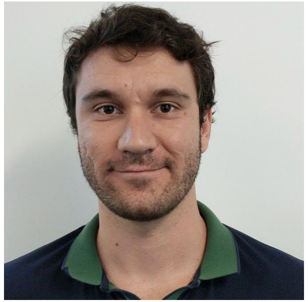
Google

A Fourier Perspective on Model Robustness in Computer Vision

Dong Yin

Raphael Lopez

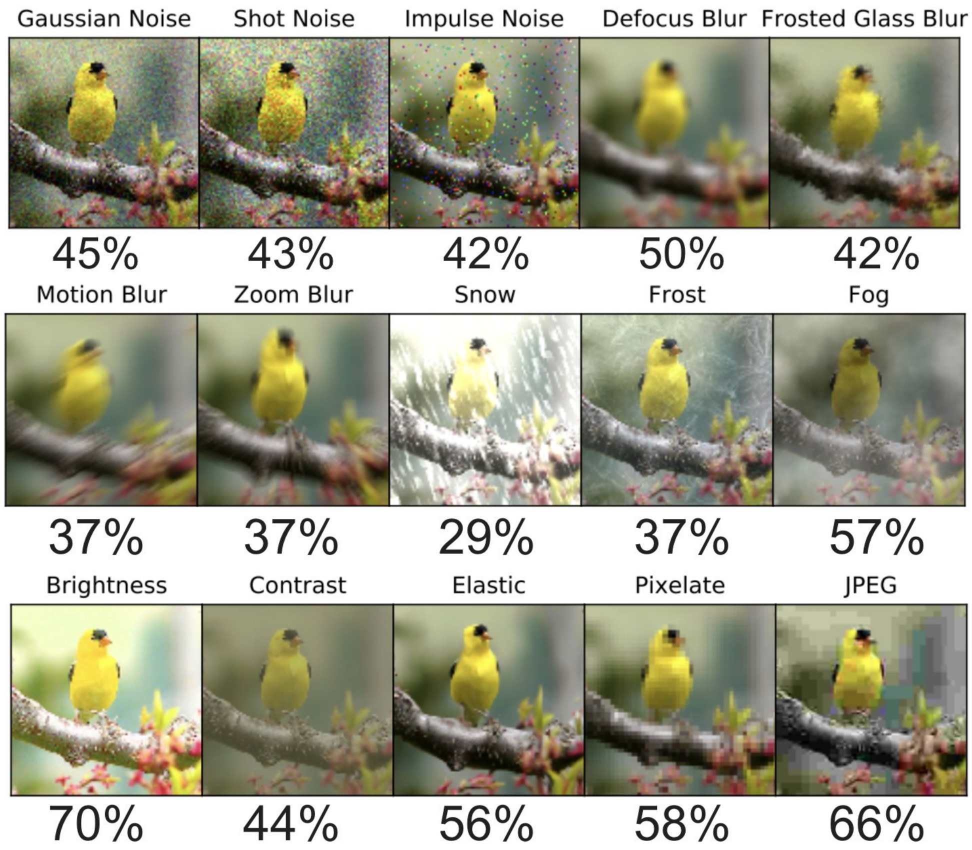
Jon Shlens



Dogus Cubuk

Justin Gilmer

Common Corruption Benchmark



Google

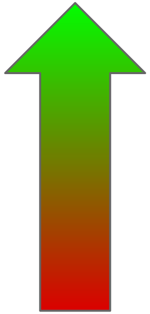
[Hendrycks et. al] https://arxiv.org/abs/1807.01697

A Motivating Experiment

Adversarial training helps some measures of robustness, but hurts others. Why?

Gaussian Noise

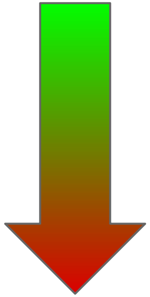
70% Acc



45% Acc

Fog

77% Acc



44% Acc

Also helps...

Frosted Glass Blur

JPEG

Shot Noise

Also hurts...

Contrast

Brightness

Spurious Correlations

Hypothesis:

Models lack robustness because they latch onto spurious correlations in the data.

Which correlations they latch onto determines their robustness properties.

Apples

VS

Oranges

Train
ResnetV5000

Eval on IID
Test Set

100% Accuracy

Cheating Models/Spurious Correlations

Apple

Yes

Is there more red pixels than orange in the photo?

No

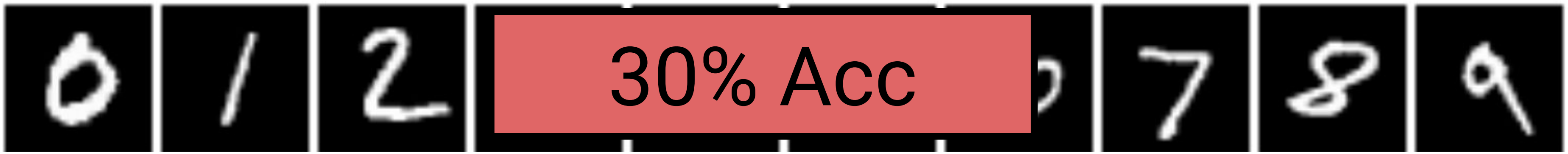
Orange

Totally an Orange!

Spurious Correlations - MNIST

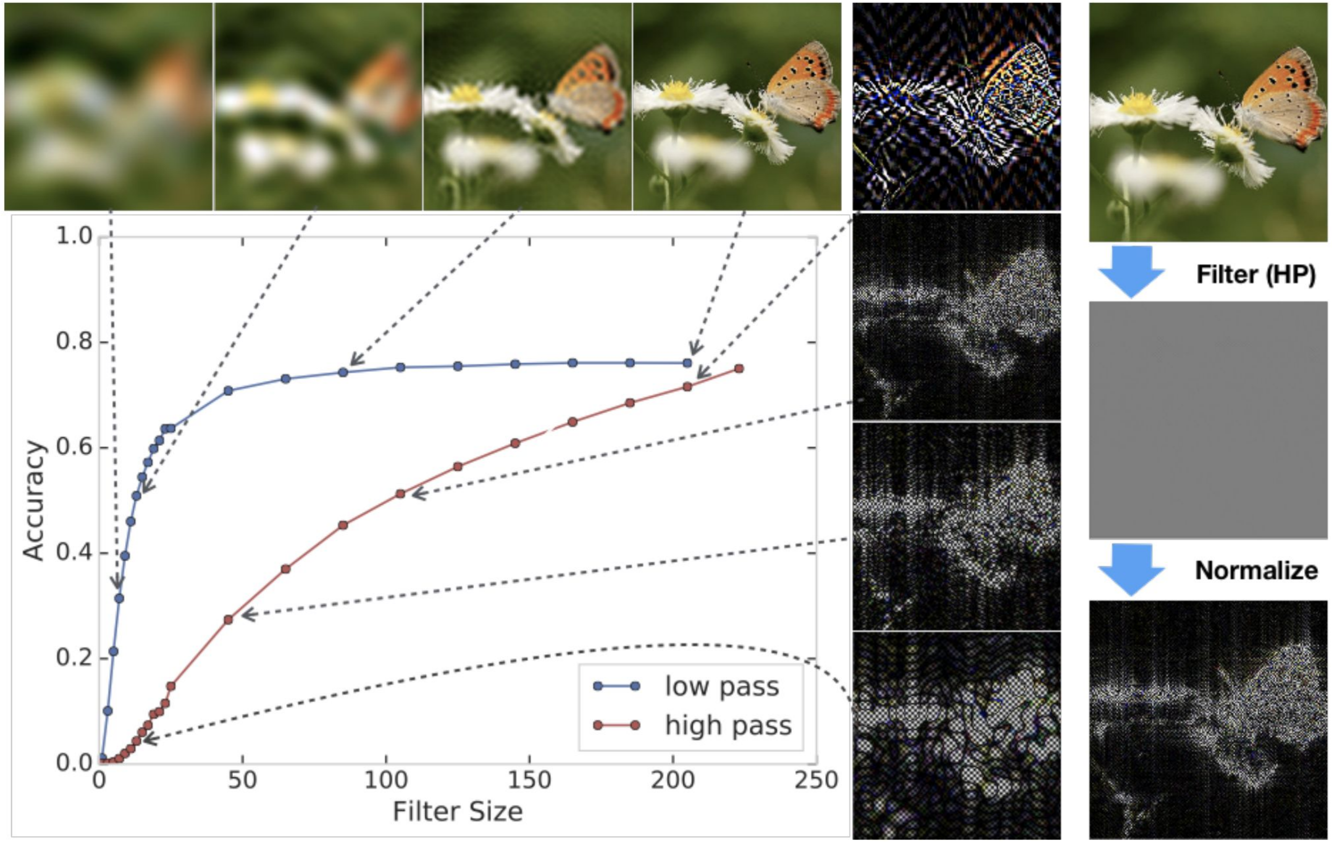
Train

Test



[Jacobsen et. al.] https://arxiv.org/pdf/1811.00401.pdf

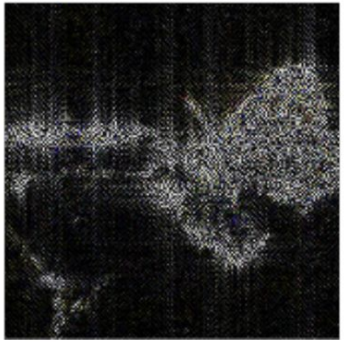
Some spurious correlations may be unintuitive



Google

Main Hypothesis: Model Bias Determines Robustness

Low freq bias ⟷ High freq bias

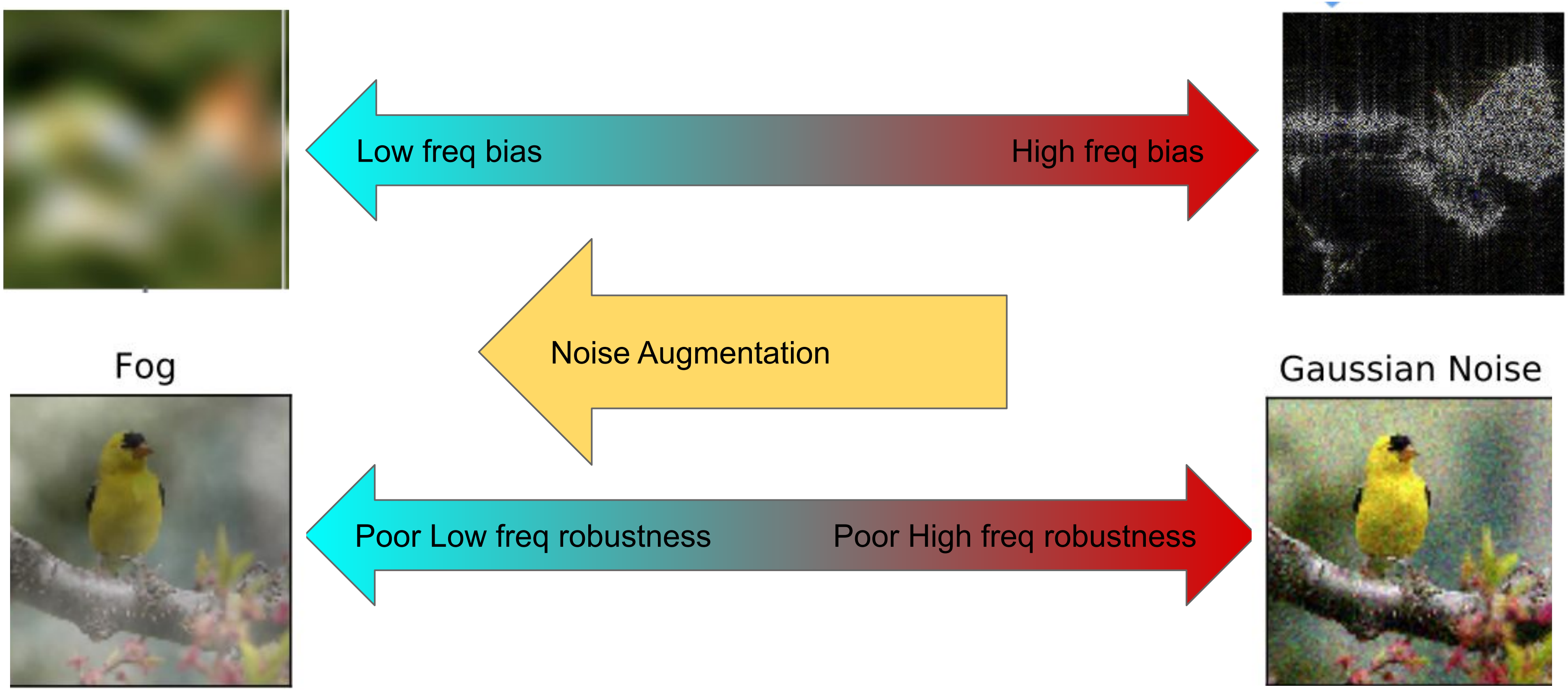


Fog

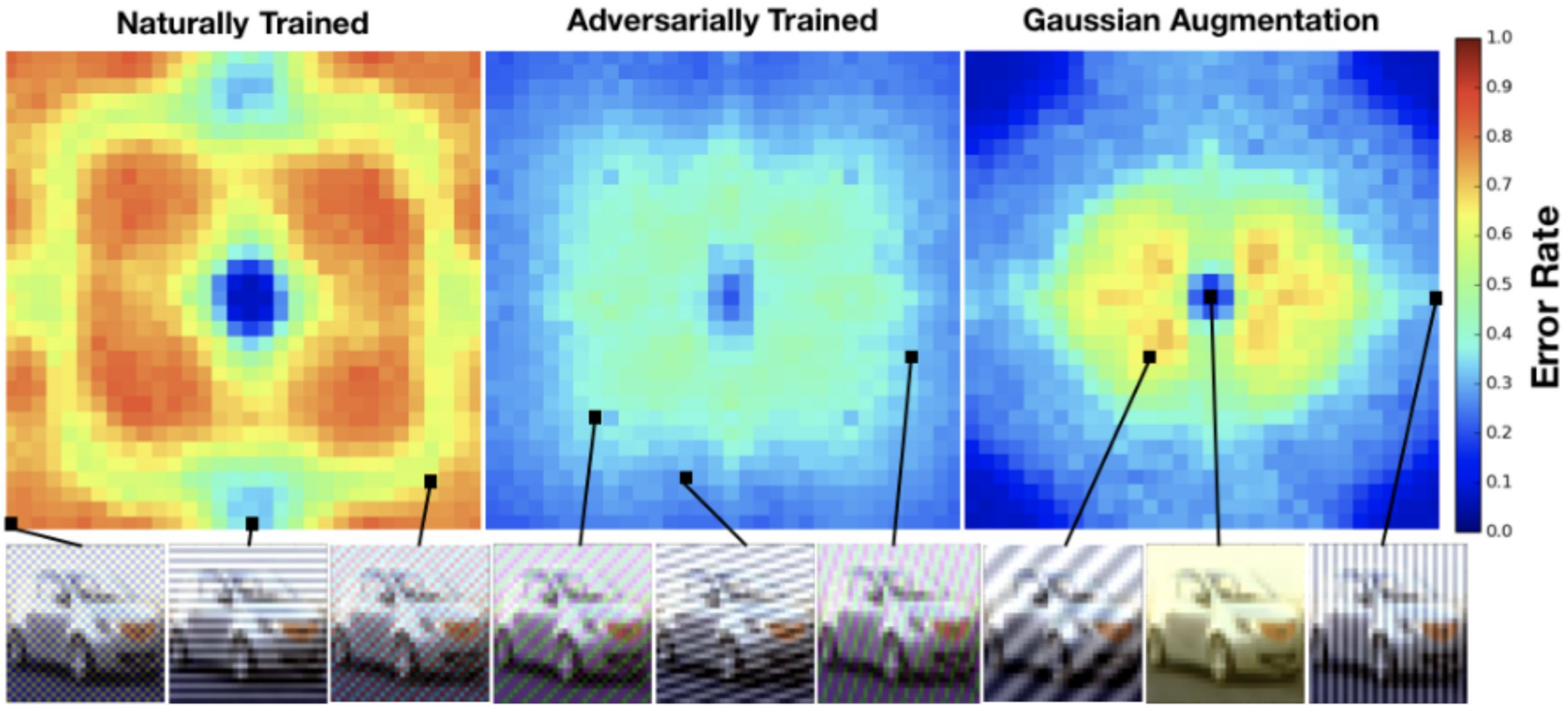
Poor Low freq robustness ⟷ Poor High freq robustness

Gaussian Noise

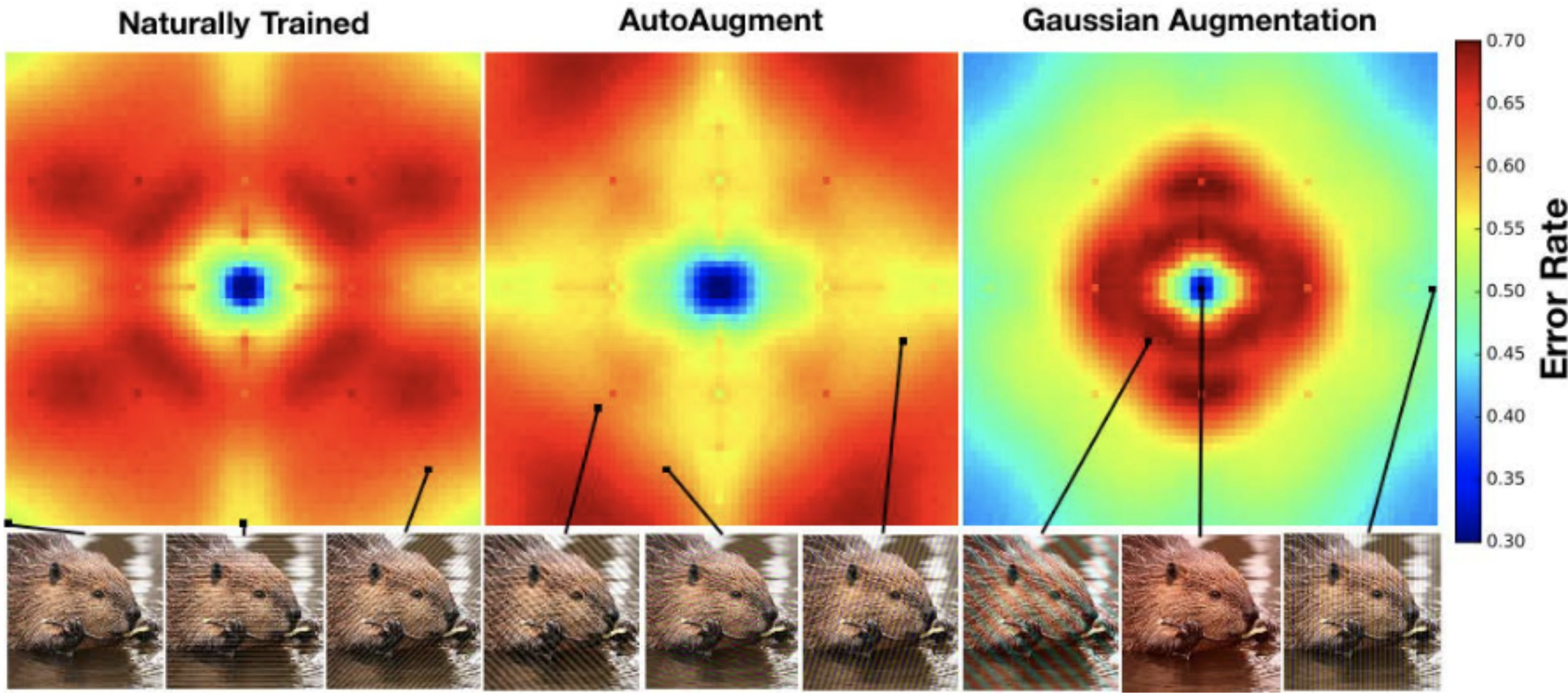
Data Augmentation Shifts Model Bias



Measuring the Effects of Data Augmentation - CIFAR10

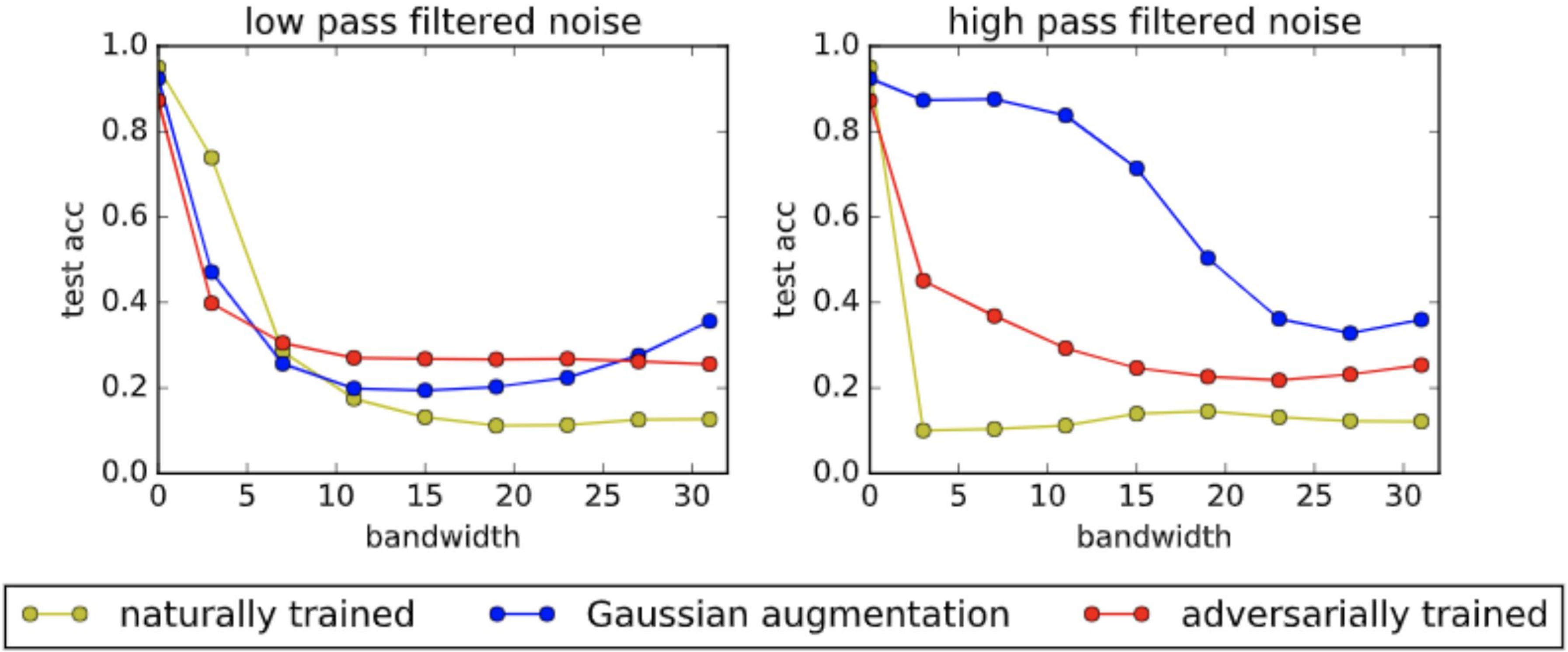


Measuring the Effects of Data Augmentation - Imagenet

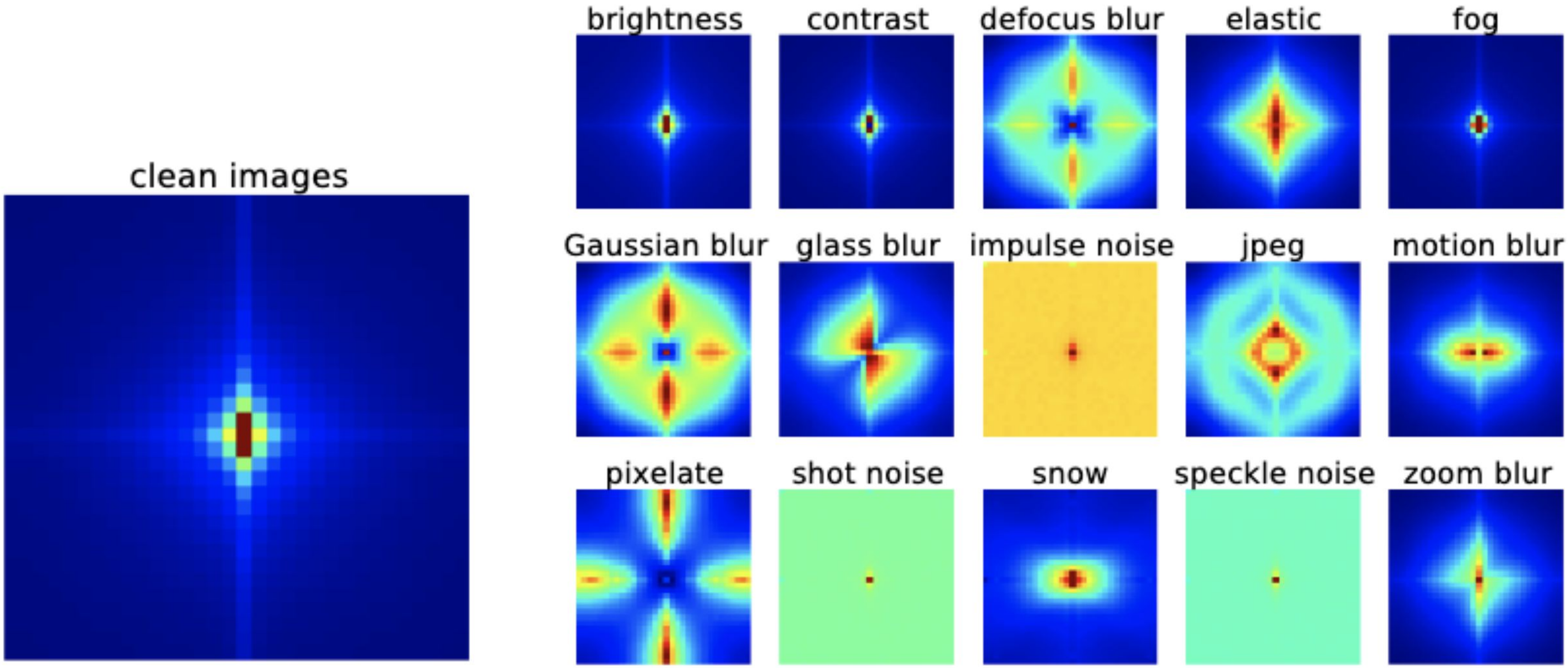


Google

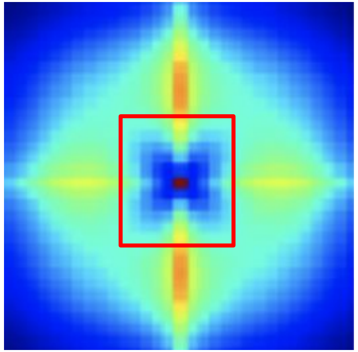
Tradeoffs from Data Aug

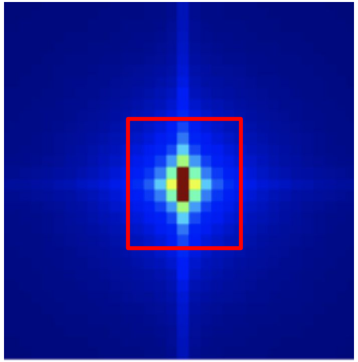


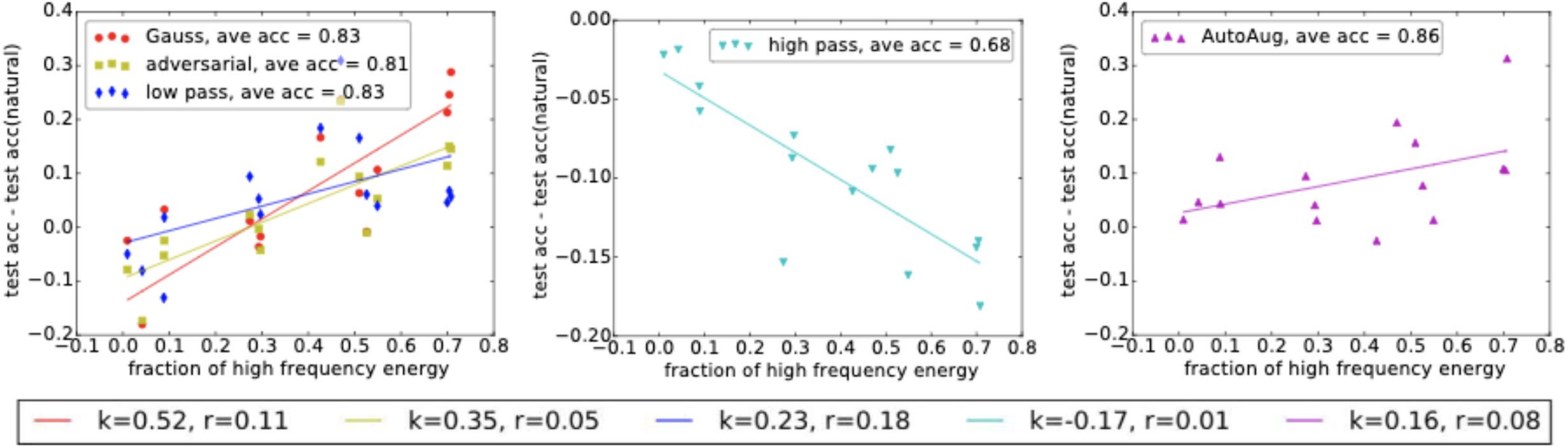
A Fourier Perspective on Common Corruptions



Tradeoffs from Data Aug



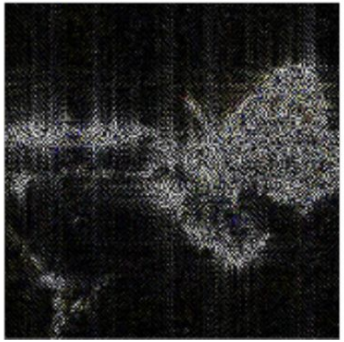




Google

Can we be robust to both high and low frequency?

Low freq bias ⟷ High freq bias



Fog

??????

Poor Low freq robustness ⟷ Poor High freq robustness

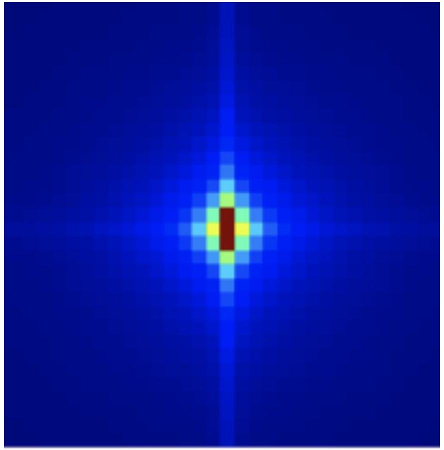
Gaussian Noise

Gaussian Data Augmentation
Adversarial Training
Low pass filtering

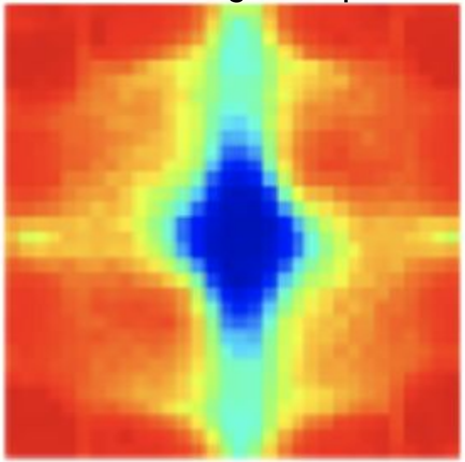
Naturally Trained
High pass filtering

Story is Complicated for Low Frequency Corruptions

Train on "Fog" noise

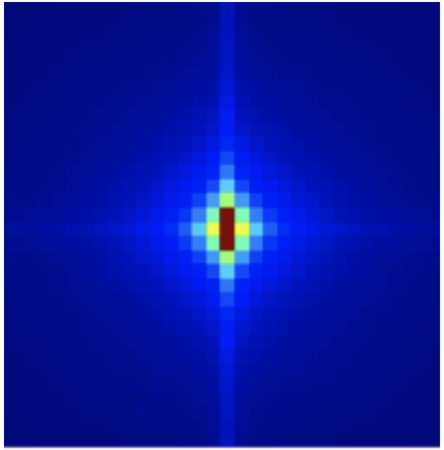


Increase High Freq Bias

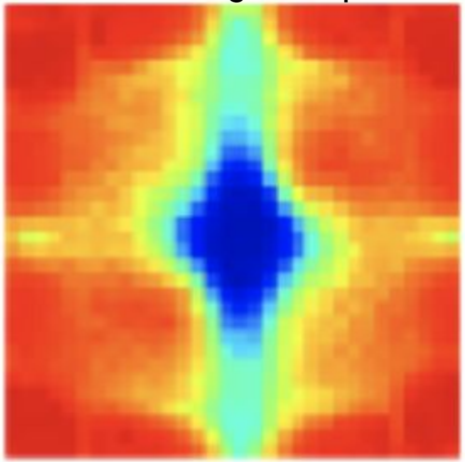


Story is Complicated for Low Frequency Corruptions

Train on "Fog" noise



Increase High Freq Bias



Degraded performance in true fog???

fog severity	1	2	3	4	5
naturally trained	0.9606	0.9484	0.9395	0.9072	0.7429
fog noise augmentation	0.9090	0.8726	0.8120	0.7175	0.4626

Maybe More Diverse Data Augmentation Needed?

	Operation 1	Operation 2
Sub-policy 0	(Posterize,0.4,8)	(Rotate,0.6,9)
Sub-policy 1	(Solarize,0.6,5)	(AutoContrast,0.6,5)
Sub-policy 2	(Equalize,0.8,8)	(Equalize,0.6,3)
Sub-policy 3	(Posterize,0.6,7)	(Posterize,0.6,6)
Sub-policy 4	(Equalize,0.4,7)	(Solarize,0.2,4)
Sub-policy 5	(Equalize,0.4,4)	(Rotate,0.8,8)
Sub-policy 6	(Solarize,0.6,3)	(Equalize,0.6,7)
Sub-policy 7	(Posterize,0.8,5)	(Equalize,1.0,2)
Sub-policy 8	(Rotate,0.2,3)	(Solarize,0.6,8)
Sub-policy 9	(Equalize,0.6,8)	(Posterize,0.4,6)
Sub-policy 10	(Rotate,0.8,8)	(Color,0.4,0)
Sub-policy 11	(Rotate,0.4,9)	(Equalize,0.6,2)
Sub-policy 12	(Equalize,0.0,7)	(Equalize,0.8,8)
Sub-policy 13	(Invert,0.6,4)	(Equalize,1.0,8)
Sub-policy 14	(Color,0.6,4)	(Contrast,1.0,8)
Sub-policy 15	(Rotate,0.8,8)	(Color,1.0,2)
Sub-policy 16	(Color,0.8,8)	(Solarize,0.8,7)
Sub-policy 17	(Sharpness,0.4,7)	(Invert,0.6,8)
Sub-policy 18	(ShearX,0.6,5)	(Equalize,1.0,9)
Sub-policy 19	(Color,0.4,0)	(Equalize,0.6,3)
Sub-policy 20	(Equalize,0.4,7)	(Solarize,0.2,4)
Sub-policy 21	(Solarize,0.6,5)	(AutoContrast,0.6,5)
Sub-policy 22	(Invert,0.6,4)	(Equalize,1.0,8)
Sub-policy 23	(Color,0.6,4)	(Contrast,1.0,8)
Sub-policy 24	(Equalize,0.8,8)	(Equalize,0.6,3)

Table 9. AutoAugment policy found on reduced ImageNet.

Google

AutoAugment Improves robustness on CIFAR-10-C

			noise			blur					weather			digital			
model	acc	mCE	speckle	shot	impulse	defocus	Gauss	glass	motion	zoom	snow	fog	bright	contrast	elastic	pixel	jpeg
natural	77	100	70	68	54	85	73	57	81	80	85	90	95	82	86	73	80
Gauss	83	98	**92**	**92**	83	84	79	**80**	77	82	88	72	92	57	84	**90**	**91**
adversarial	81	108	82	83	69	84	82	**80**	80	83	83	73	87	77	82	85	85
Auto	**86**	**64**	81	78	**86**	**92**	**88**	76	**85**	**90**	**89**	**95**	**96**	**95**	**87**	71	81

- Stylized imagenet training does better on Imagenet-C.
- Current SOTA on Imagenet-C is AugMix, which builds off of AutoAugment.

Takeaways

- Model bias determines robustness.
- Data augmentation can help but there may be tradeoffs.
  - Shift bias towards low frequency -> improve robustness to high frequency.
  - Shift bias towards low frequency -> degrade robustness to low frequency.
- Diversity is needed for more general robustness.
  - See AugMix follow-up https://openreview.net/forum?id=S1gmrxHFvB

Thank You!

Google