Adversarial Examples are a Natural Consequence of Test Error in Noise

Nic Ford*, Justin Gilmer*, Nicholas Carlini, Dogus Cubuk

*equal contribution

Robust (out of distribution) Generalization

Train on p(x)



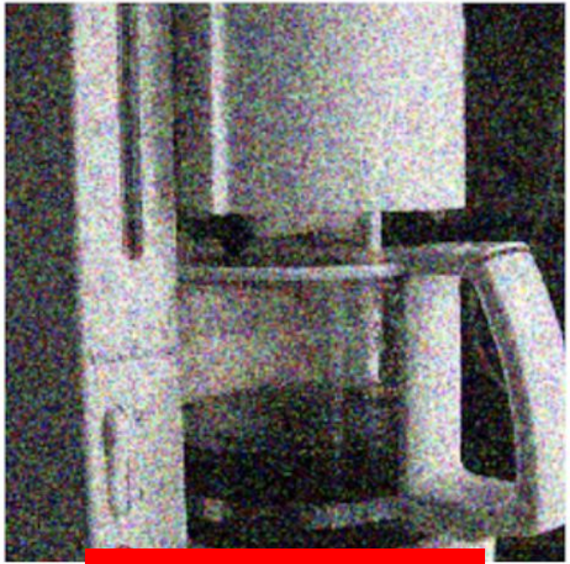

Test on q(x)




Gaussian noise

$\sigma = .2$

“Toaster”



50% top-1 acc

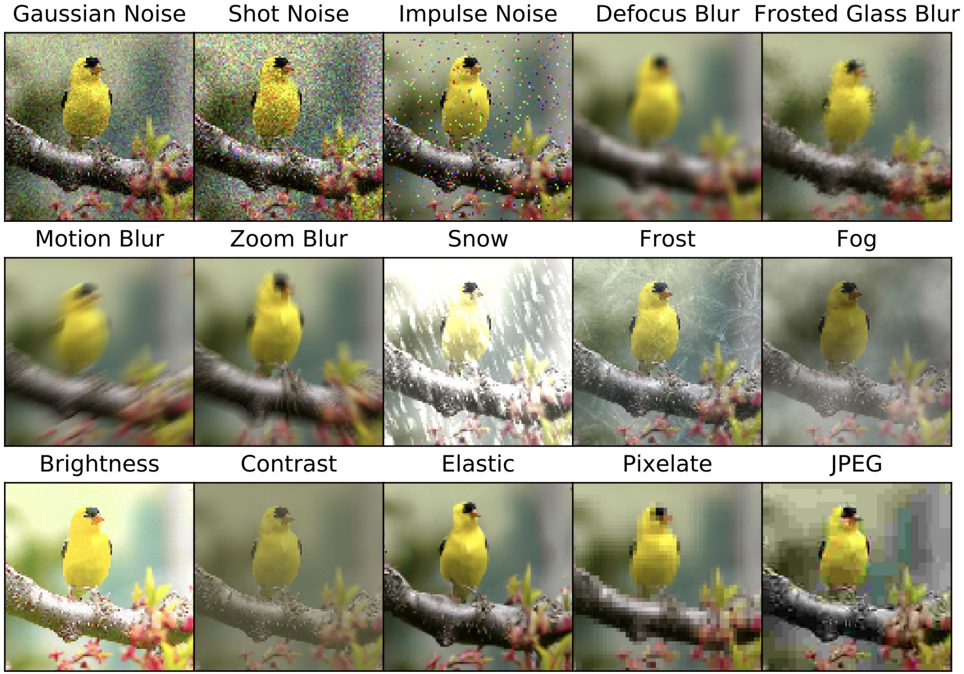
$\sigma = .4$

“Computer”

14% top-1 acc

Corruption Robustness

- Goal: Measure and improve model robustness to distributional shift.



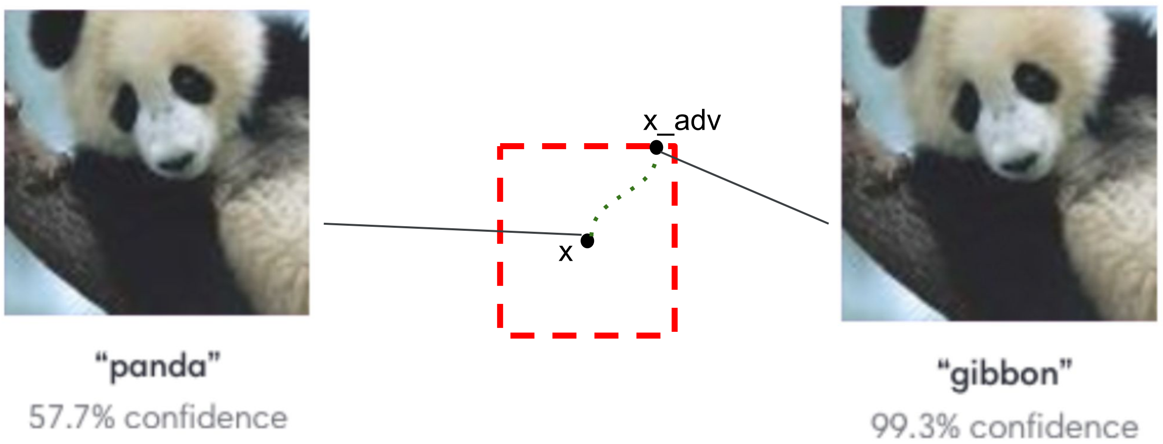
[Hendrycks et. al] https://arxiv.org/pdf/1807.01697.pdf

See also:
[Mu, Gilmer] "MNIST-C" https://arxiv.org/abs/1906.02337
[Pei et. al.] - https://arxiv.org/pdf/1712.01785.pdf

Google

Adversarial Examples - The "Surprising" Phenomenon

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

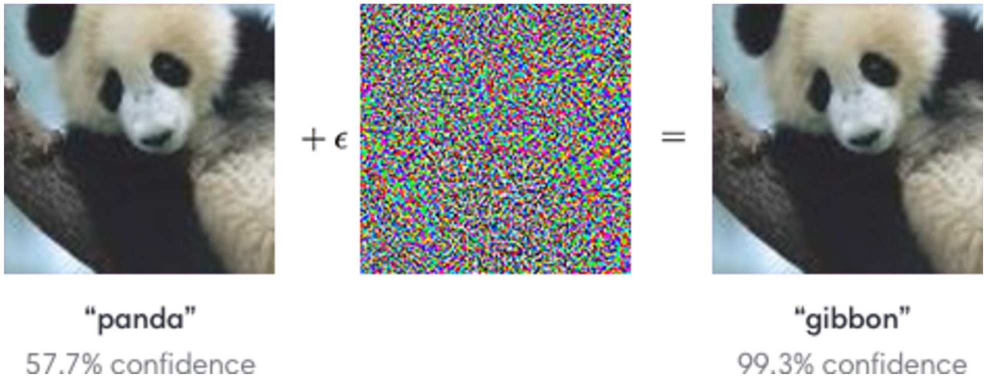


[Goodfellow et. al.]

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

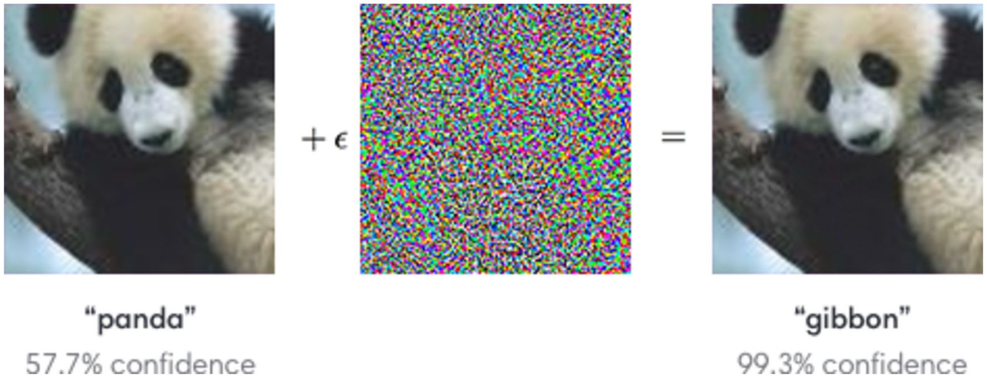
Adversarial Examples - The Phenomenon

Why do our models have adversarial examples?



Adversarial Examples - The Phenomenon

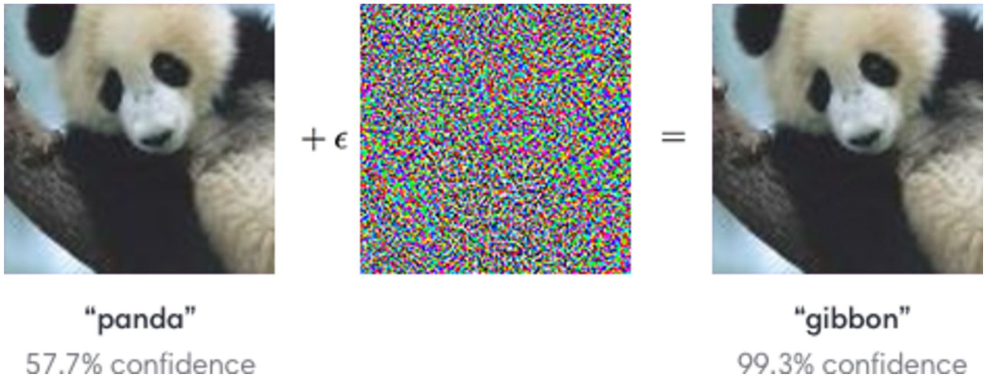
Why do our models have adversarial examples? **A:** ???



Adversarial Examples - The Phenomenon

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

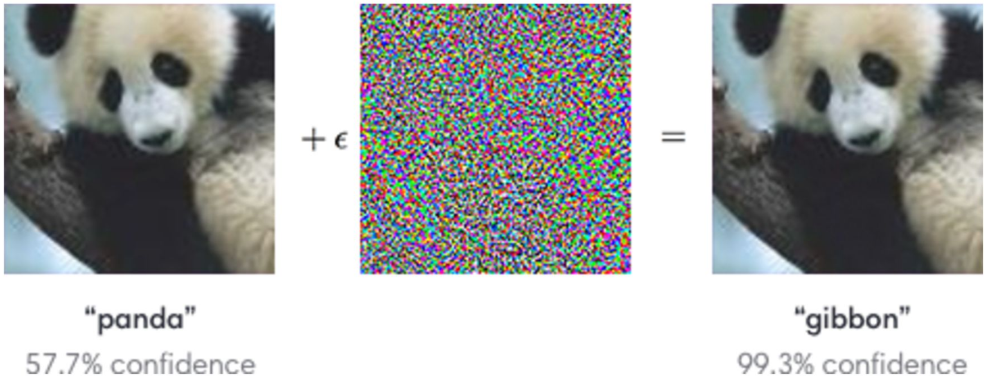
What are adversarial examples?



Adversarial Examples - The Phenomenon

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

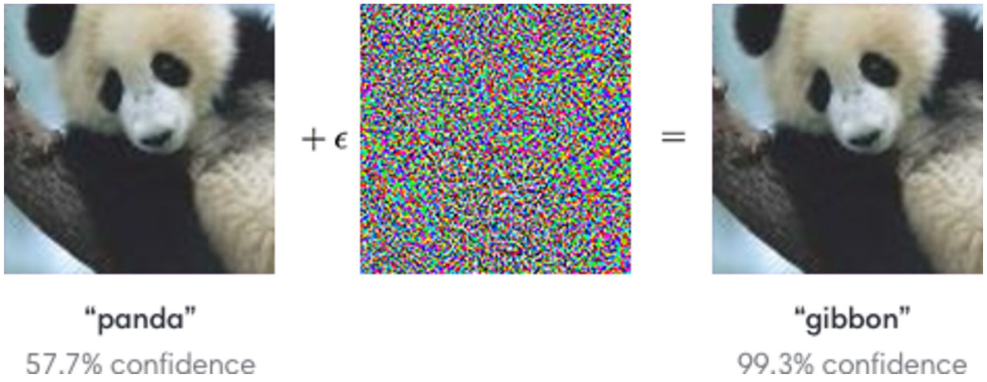
What are adversarial examples? **A:** The nearest error



Adversarial Examples - The Phenomenon

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

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



Google

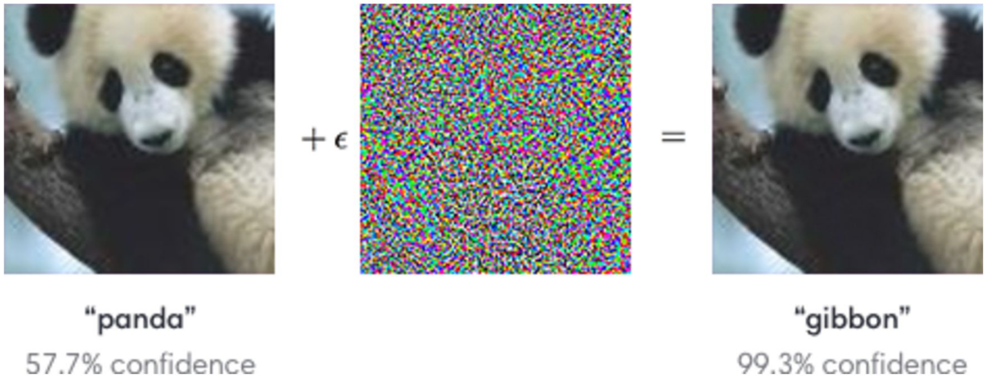
Adversarial Examples - The Phenomenon

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

A: ???

What are adversarial examples?

A: The nearest error



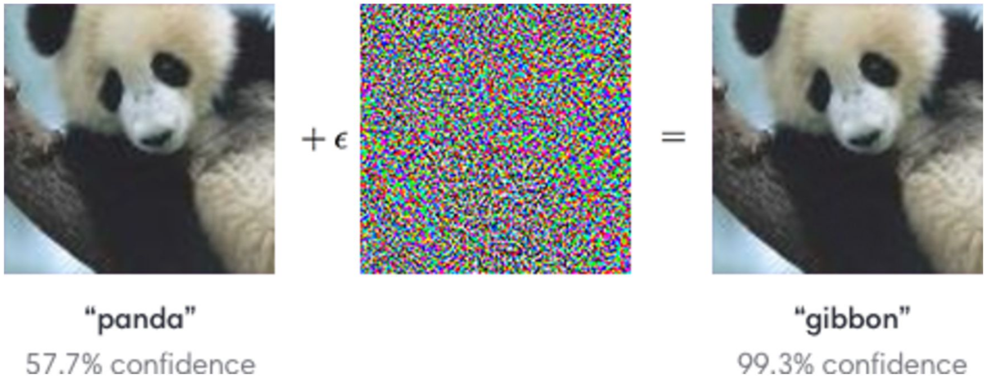
Adversarial Examples - The Phenomenon

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

A: ???

What are adversarial examples?

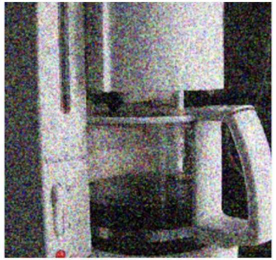
A: The nearest error

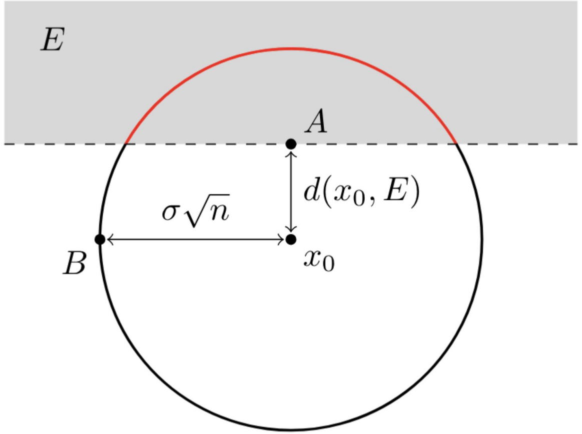


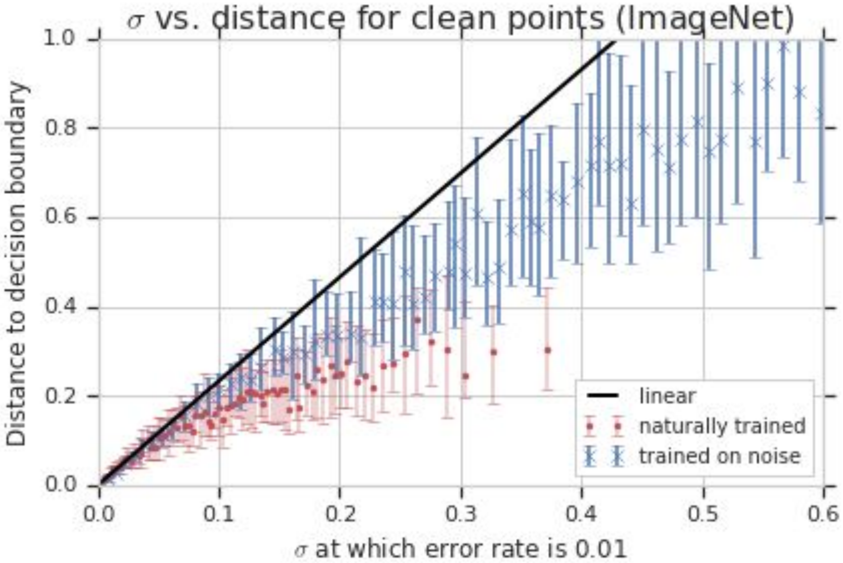
Test error > 0 (iid, ood) -> errors exist -> there is a nearest error

Linear Assumption

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







See also Fawzi et. al.

Google

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%

Google

Adversarial Defenses

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%

Conclusion

- It is not surprising that models have a nearest error.
- The nearest error is not unusually close given measured o.o.d robustness.
- The robustness problem is much broader than tiny perturbations.
- If a method doesn't improve o.o.d robustness, is it more secure?