Adversarial Training and Robustness for Multiple Perturbations

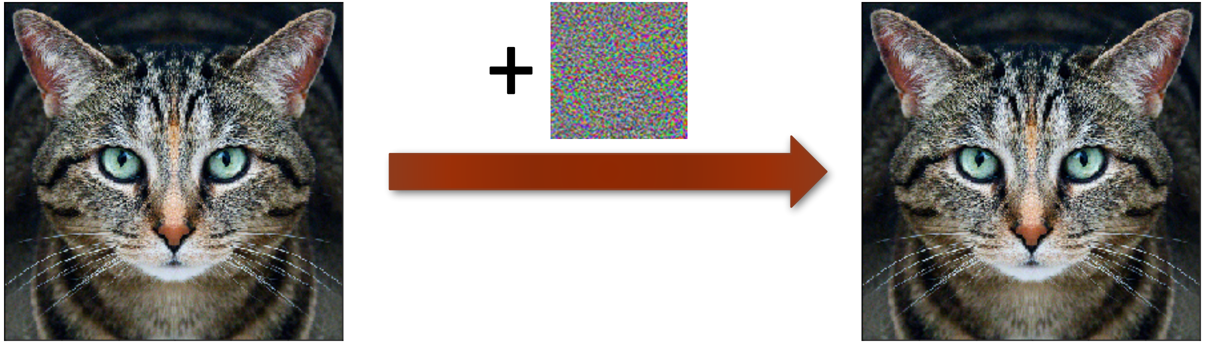
Florian Tramèr & Dan Boneh

NeurIPS 2019

Poster #87

Stanford University

Adversarial examples

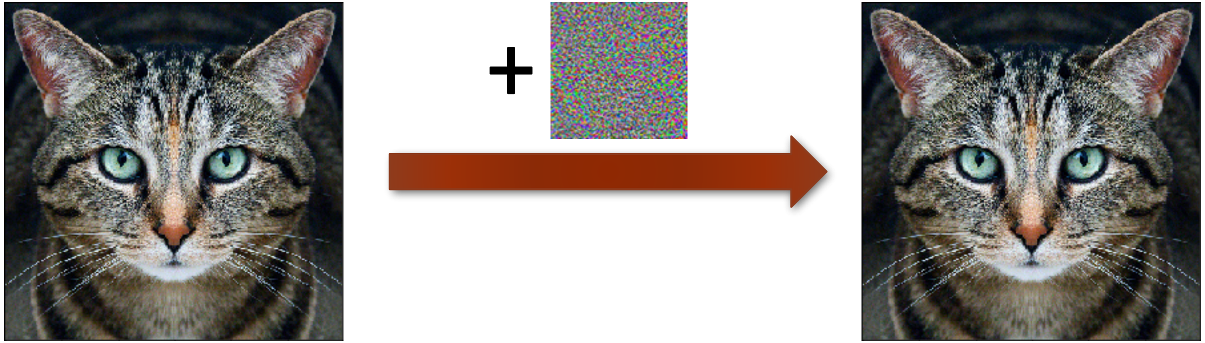


88% Tabby Cat

99% Guacamole

Szegedy et al., 2014
Goodfellow et al., 2015
Athalye, 2017

Adversarial examples



88% Tabby Cat

99% Guacamole

- ML models learn very different features than humans
- This is a safety concern for deployed ML models
- Classification in adversarial settings is hard

Szegedy et al., 2014
Goodfellow et al., 2015
Athalye, 2017

Adversarial training

Szegedy et al., 2014
Madry et al., 2017

Adversarial training

1. Choose a set of perturbations: e.g., noise of small ℓ_∞ norm: 

Szegedy et al., 2014
Madry et al., 2017

Adversarial training

1. Choose a set of perturbations: e.g., noise of small ℓ_∞ norm: 

2. For each example  , find an adversarial example:  + 

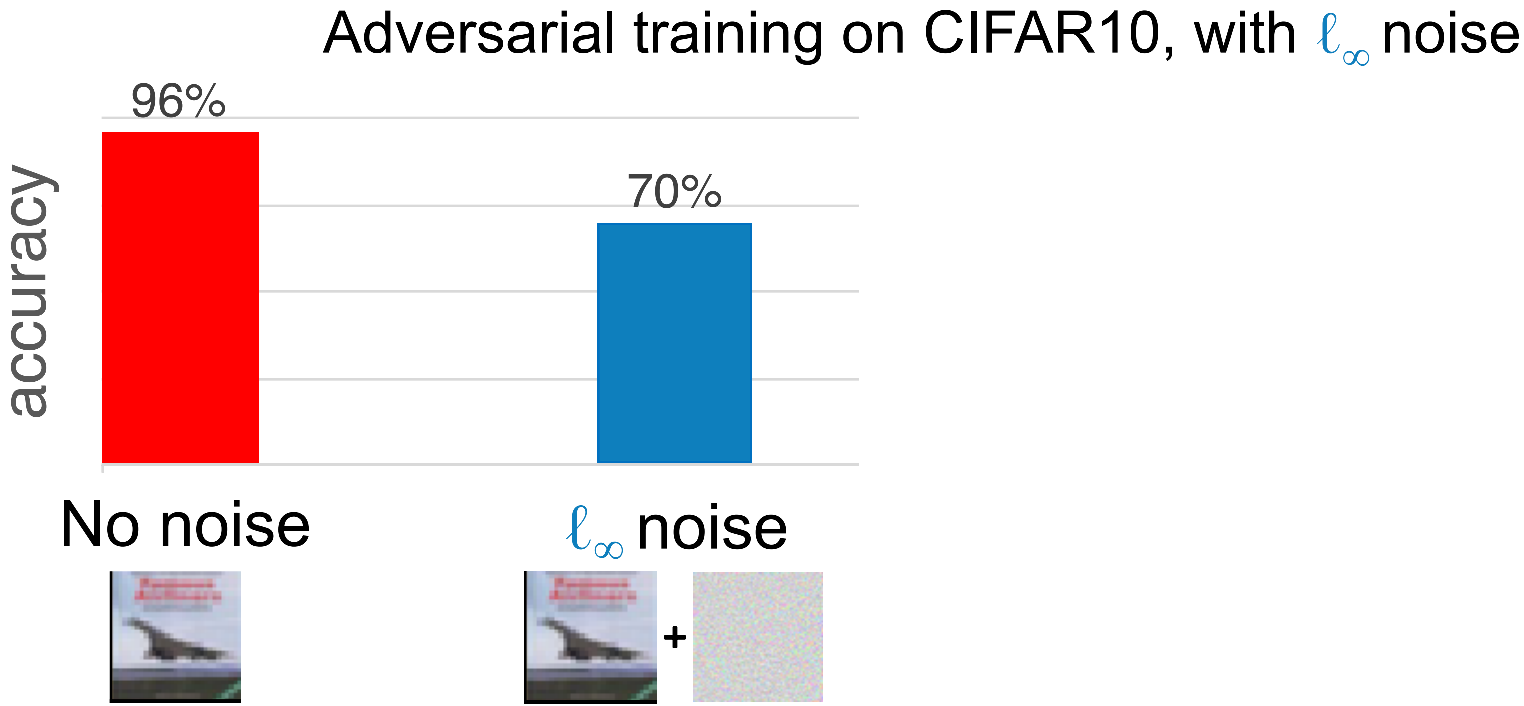
3. Train the model on  + 

4. Repeat until convergence

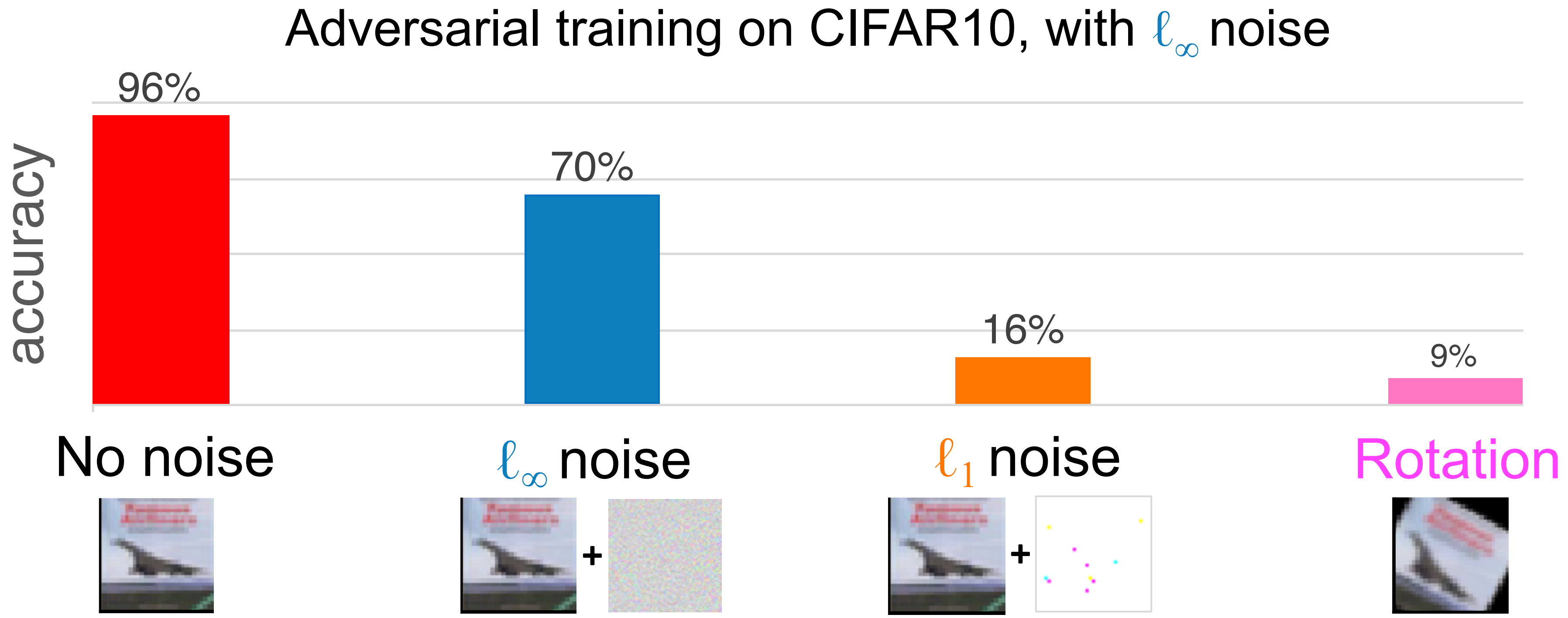
Szegedy et al., 2014
Madry et al., 2017

How well does it work?

How well does it work?



How well does it work?



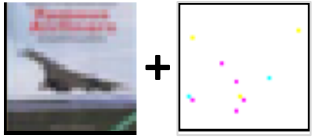
Engstrom et al., 2017
Sharma & Chen, 2018

How to prevent other adversarial examples?

How to prevent other adversarial examples?

$S_1 = \{\delta: ||\delta||_\infty \leq \varepsilon_\infty\}$

$S_2 = \{\delta: ||\delta||_1 \leq \varepsilon_1\}$



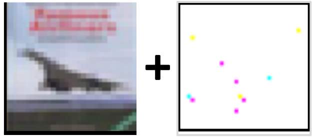
$S_3 = \{\delta: $ «small rotation» $\}$

Adversary can choose a perturbation type for each input

How to prevent other adversarial examples?

$S_1 = \{\delta : ||\delta||_\infty \leq \varepsilon_\infty\}$

$S_2 = \{\delta : ||\delta||_1 \leq \varepsilon_1\}$



$S_3 = \{\delta :$ «small rotation»$\}$

$S = S_1 \cup S_2 \cup S_3$

Adversary can choose a perturbation type for each input

- Pick worst-case adversarial example from **S**
- Train the model on that example

Does this work?

Does this work?

Does this work?

A <u>robustness tradeoff</u> is provably inherent in some classification tasks

Increased robustness to one type of noise $\Rightarrow$ decreased robustness to another

Empirically validated on CIFAR10 & MNIST

Does this work?

A robustness tradeoff is provably inherent in some classification tasks

Increased robustness to one type of noise
⇒ decreased robustness to another

Empirically validated on CIFAR10 & MNIST

MNIST:

Does this work?

A <u>robustness tradeoff</u> is provably inherent in some classification tasks

Increased robustness to one type of noise $\Rightarrow$ decreased robustness to another

Empirically validated on CIFAR10 & MNIST

MNIST:

For ℓ_∞, ℓ_1 and ℓ_2 noise:

50% accuracy

Does this work?

A <u>robustness tradeoff</u> is provably inherent in some classification tasks

Increased robustness to one type of noise $\Rightarrow$ decreased robustness to another

Empirically validated on CIFAR10 & MNIST

MNIST:

For ℓ_∞, ℓ_1 and ℓ_2 noise:

50% accuracy

gradient masking

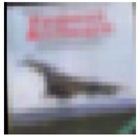
What if we combine perturbations?

What if we combine perturbations?

natural image

rotation

ℓ_∞ noise



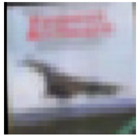
½ rotation + ½ ℓ_∞ noise

What if we combine perturbations?

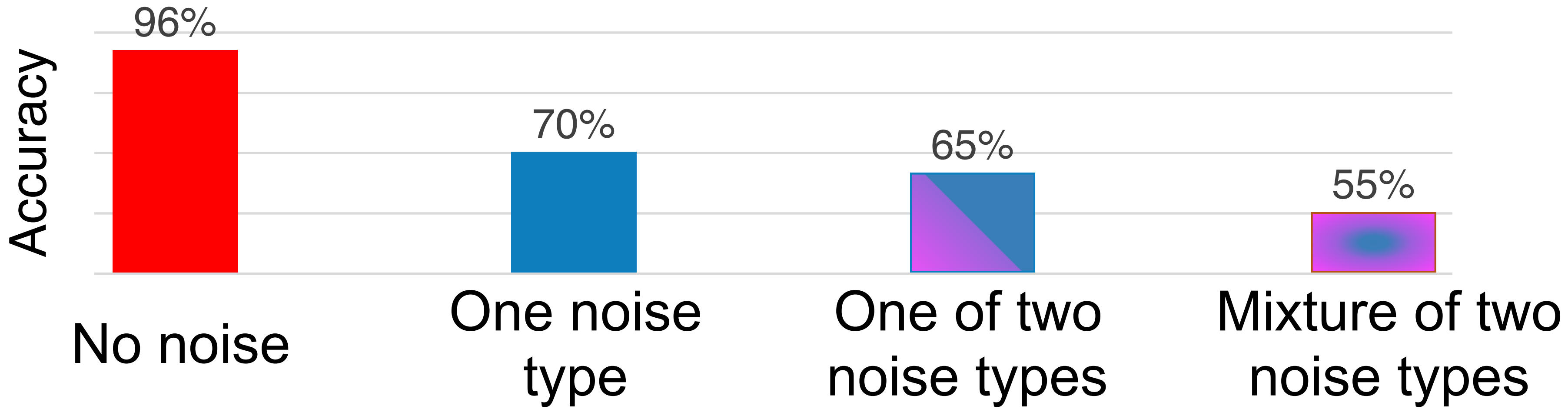
natural image

rotation

ℓ_∞ noise



½ rotation + ½ ℓ_∞ noise



Conclusion

Adversarial training for multiple perturbation sets works, but...

- Significant loss in robustness
- Weak robustness to affine combinations of perturbations

https://arxiv.org/abs/1904.13000

Poster #87

Conclusion

Adversarial training for multiple perturbation sets works, but...

- Significant loss in robustness
- Weak robustness to affine combinations of perturbations

Open questions:

- Train a *single* MNIST model with high robustness to any ℓ_p noise
- Better scaling of multi-perturbation adversarial training
- Which perturbations do we care about?

https://arxiv.org/abs/1904.13000

Poster #87