# Developments in Adversarial Machine Learning

**Florian Tramèr**

September 19th 2019

Based on joint work with Jens Behrmannn, Dan Boneh, Nicholas Carlini, Edward Chou, Pascal Dupré, Jörn-Henrik Jacobsen, Nicolas Papernot, Giancarlo Pellegrino, Gili Rusak

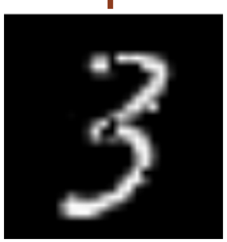
# Adversarial (Examples in) ML

GANs vs Adversarial Examples

Maybe we need to write 10x more papers

10000+ papers

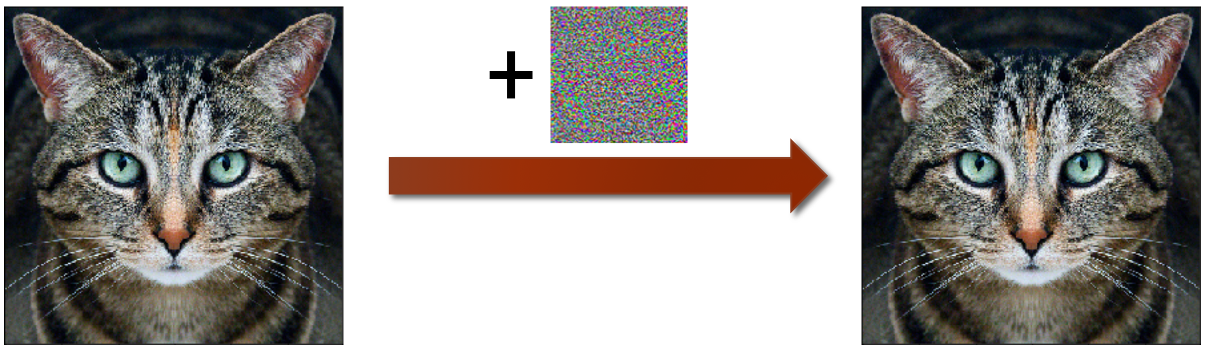
2013 2014 2017 2019



1000+ papers

N. Carlini, “Recent Advances in Adversarial Machine Learning”, ScAINet 2019

# Adversarial Examples

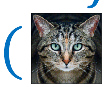


88% Tabby Cat

99% Guacamole

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

## How?

- Training $\Longrightarrow$ "tweak model parameters such that $f($$) = cat$"
- Attacking $\Longrightarrow$ "tweak input pixels such that $f($$) = guacamole$"

## Why?

- Concentration of measure in high dimensions?
  [Gilmer et al., 2018, Mahloujifar et al., 2018, Fawzi et al., 2018, Ford et al., 2019]
- Well generalizing "superficial" statistics?
  [Jo & Bengio 2017, Ilyas et al., 2019, Gilmer & Hendrycks 2019]

# Defenses

- A bunch of failed ones...

- Adversarial Training [Szegedy et al., 2014, Goodfellow et al., 2015, Madry et al., 2018]

⇒ For each training input (**x**, y), find worst-case adversarial input

$$\underset{\boldsymbol{x}' \in S(\boldsymbol{x})}{\operatorname{argmax}} \operatorname{Loss}(f(\boldsymbol{x}'), y)$$

A set of allowable perturbations of **x**
e.g., $\{\mathbf{x}' : \| \mathbf{x} - \mathbf{x}' \|_\infty \leq \varepsilon\}$

⇒ Train the model on (**x**’, y)

Worst-case data augmentation

- Certified Defenses [Raghunathan et al., 2018, Wong & Kolter 2018]

⇒ Certificate of provable robustness for each point

⇒ Empirically weaker than adversarial training

# $L_p$ robustness: An Over-studied Toy Problem?

Neural networks aren't robust.
Consider this simple **"expectimax $L_p$"** game:

1. Sample random input from test set
2. Adversary perturbs point within small $L_p$ ball
3. Defender classifies perturbed point

2015

This was just a toy threat model ...
Solving this won't magically make
ML more "secure"

2019 and 1000+ papers later

Ian Goodfellow, "The case for dynamic defenses against adversarial examples", SafeML ICLR Workshop, 2019

# Limitations of the “expectimax $L_p$” Game

1. Sample random input from test set
   - What if model has 99% accuracy and adversary always picks from the 1%? (test-set attack, [Gilmer et al., 2018])
2. Adversary perturbs point within $L_p$ ball
   - Why limit to one $L_p$ ball?
   - How do we choose the “right” $L_p$ ball?
   - Why “imperceptible” perturbations?
3. Defender classifies perturbed point
   - Can the defender abstain? (attack detection)
   - Can the defender adapt?

# A real-world example of the "expectimax $L_p$" threat model: Perceptual Ad-blocking

- Ad-blocker's goal: classify images as ads
- Attacker goals:
  - Perturb ads to evade detection (False Negative)
  - Perturb benign content to detect ad-blocker (False Positive)

1. Can the attacker run a "test-set attack"?
   - No! (or ad designers have to create lots of random ads...)

2. Should attacks be imperceptible?
   - Yes! The attack should not affect the website user
   - Still, many choices other than $L_p$ balls

3. Is detecting attacks enough?
   - No! Attackers can exploit FPs and FNs

**T** et al., "AdVersarial: Perceptual Ad Blocking meets Adversarial Machine Learning", CCS 2019

# Limitations of the “expectimax $L_p$” Game

1. Sample random input from test set

2. Adversary perturbs point within $L_p$ ball
   - Why limit to one $L_p$ ball?
   - How do we choose the “right” $L_p$ ball?
   - Why “imperceptible” perturbations?

3. Defender classifies perturbed point
   - Can the defender abstain? (attack detection)

# Limitations of the "expectimax $L_p$" Game

1. Sample random input from test set

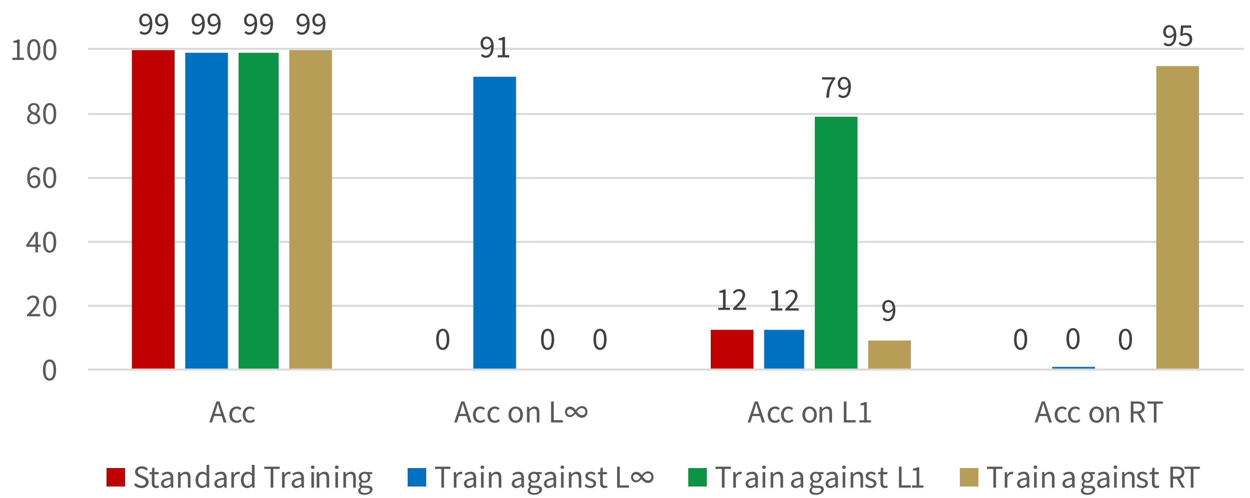
2. Adversary perturbs point within $L_p$ ball
   - Why limit to one $L_p$ ball?
   - How do we choose the "right" $L_p$ ball?
   - Why "imperceptible" perturbations?

3. Defender classifies perturbed point
   - Can the defender abstain? (attack detection)

# Robustness for Multiple Perturbations

Do defenses (e.g., adversarial training) generalize across perturbation types?

MNIST:



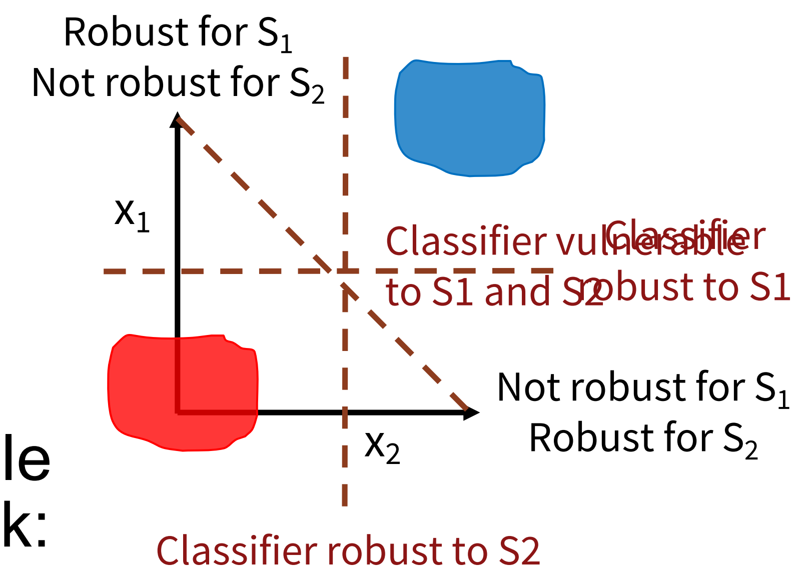

Robustness to one perturbation type ≠ robustness to all
Robustness to one type can increase vulnerability to others

**T** & Boneh, "Adversarial Training and Robustness for Multiple Perturbations", NeurIPS 2019

# The multi-perturbation robustness trade-off

If there exist models with high robust accuracy for perturbation sets $S_1, S_2, \ldots, Sn$, does there **exist** a model robust to perturbations from $\bigcup_{i=1}^{n} S_i$ ?

<u>Answer:</u> in general, NO!




There exist “mutually exclusive perturbations” (MEPs)

(robustness to $S_1$ implies vulnerability to $S_2$ and vice-versa)

Formally, we show that for a simple Gaussian binary classification task:

- $L_1$ and $L_\infty$ perturbations are MEPs
- $L_\infty$ and spatial perturbations are MEPs

**T** & Boneh, “Adversarial Training and Robustness for Multiple Perturbations”, NeurIPS 2019

# Empirical Evaluation

Can we train models to be robust to multiple perturbation types simultaneously?

Adversarial training for multiple perturbations:

$\Rightarrow$ For each training input (**x**, y), find worst-case adversarial input

$$\underset{\boldsymbol{x}' \in \bigcup_{i=1}^{n} S_i}{\operatorname{argmax}} \ \mathrm{Loss}(f(\boldsymbol{x}'), y)$$

$\Rightarrow$ "Black-box" approach:

Scales linearly in number of perturbation sets

$$\underset{\boldsymbol{x}' \in \bigcup_{i=1}^{n} S_i}{\operatorname{argmax}} \ \mathrm{Loss}(f(\boldsymbol{x}'), y) = \underset{1 \le i \le n}{\operatorname{argmax}} \left\{ \underbrace{\underset{\boldsymbol{x}' \in S_i}{\operatorname{argmax}} \ \mathrm{Loss}(f(\boldsymbol{x}'), y)}_{\text{Use existing attack tailored to } S_i} \right\}$$

**T** & Boneh, "Adversarial Training and Robustness for Multiple Perturbations", NeurIPS 2019

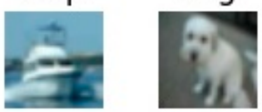
# Results

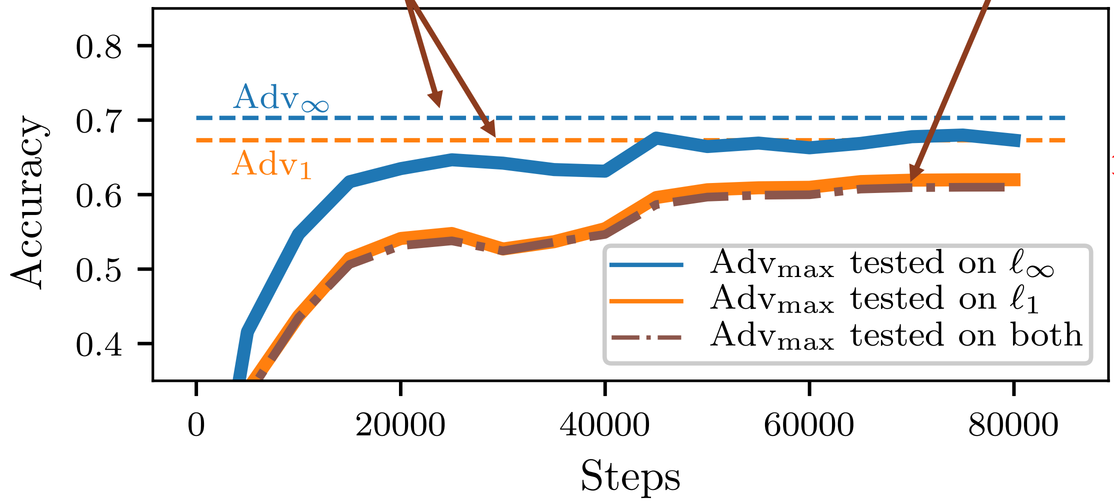
Robust accuracy when training/evaluating on a single perturbation type

Robust accuracy when training/evaluating on both

CIFAR10:

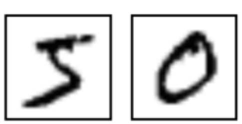
ship dog

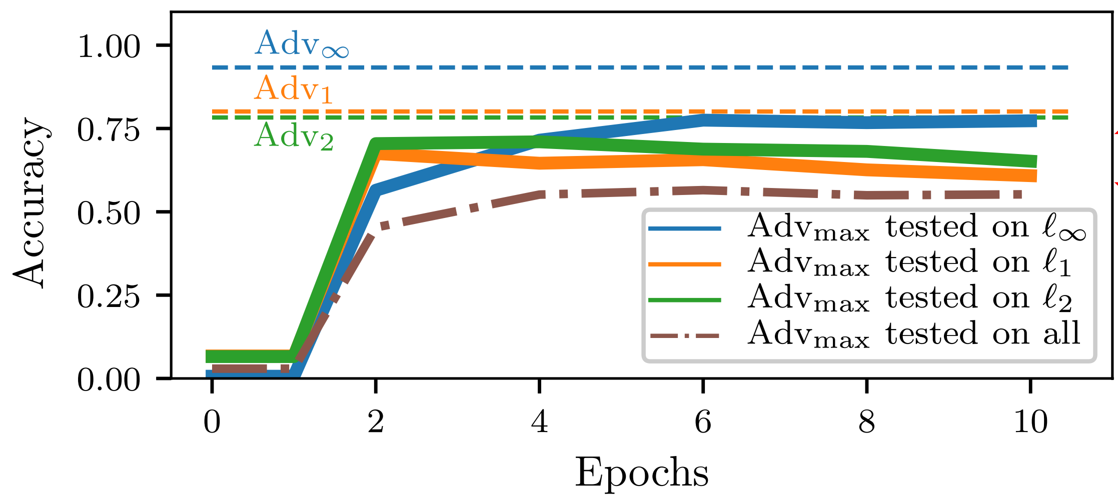





Loss of ~**5%** accuracy

MNIST:






Loss of ~**20%** accuracy

**T** & Boneh, "Adversarial Training and Robustness for Multiple Perturbations", NeurIPS 2019

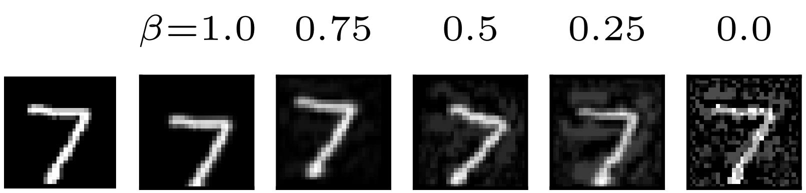
# Affine adversaries

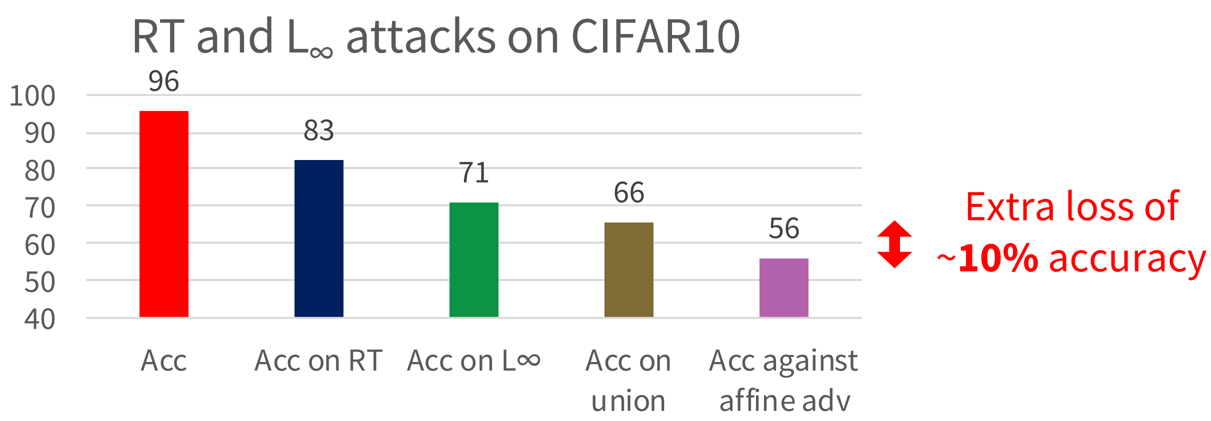
Instead of picking perturbations from $S_1 \cup S_2$ why not combine them?

E.g., small $L_1$ noise + small $L_\infty$ noise

or small rotation/translation + small $L_\infty$ noise

Affine adversary picks perturbation from $\beta S_1 + (1-\beta) S_2$, for $\beta \in [0,1]$




**T** & Boneh, "Adversarial Training and Robustness for Multiple Perturbations", NeurIPS 2019

# Limitations of the “expectimax $L_p$” Game

1. Sample random input from test set

2. Adversary perturbs point within $L_p$ ball
   - Why limit to one $L_p$ ball?
   - How do we choose the “right” $L_p$ ball?
   - Why “imperceptible” perturbations?

3. Defender classifies perturbed point
   - Can the defender abstain? (attack detection)

# Invariance Adversarial Examples

Let's look at MNIST again:

(Simple dataset, centered and scaled, non-trivial robustness is achievable)

 $\in \{0, 1\}^{784}$

Models have been trained to "extreme" levels of robustness

(E.g., robust to $L_1$ noise > 30 or $L_\infty$ noise = 0.4)

**⇒ Some of these defenses are certified!**

natural

$L_1$ perturbed

$L_\infty$ perturbed

**For such examples, humans agree more often with an undefended model than with an overly robust model**

Jacobsen **et al.**, "Exploiting Excessive Invariance caused by Norm-Bounded Adversarial Robustness", 2019

# Limitations of the "expectimax $L_p$" Game

1. Sample random input from test set

2. Adversary perturbs point within $L_p$ ball
   - Why limit to one $L_p$ ball?
   - How do we choose the "right" $L_p$ ball?
   - Why "imperceptible" perturbations?

3. Defender classifies perturbed point
   - Can the defender abstain? (attack detection)

# New Ideas for Defenses

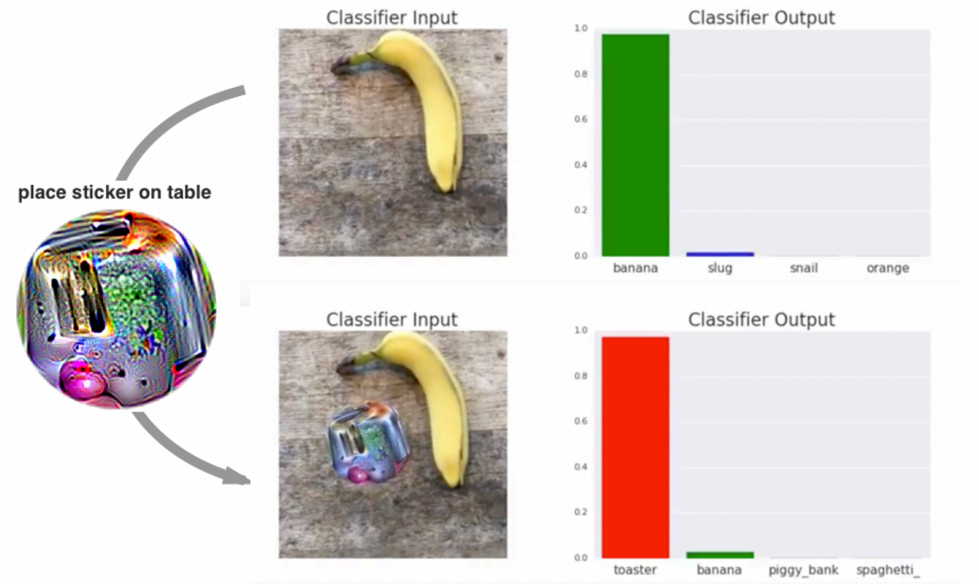
What would a realistic attack on a cyber-physical image classifier look like?

1. Attack has to be physically realizable

   $\Rightarrow$ Robustness to physical changes (lighting, pose, etc.)

2. Some degree of "universality"

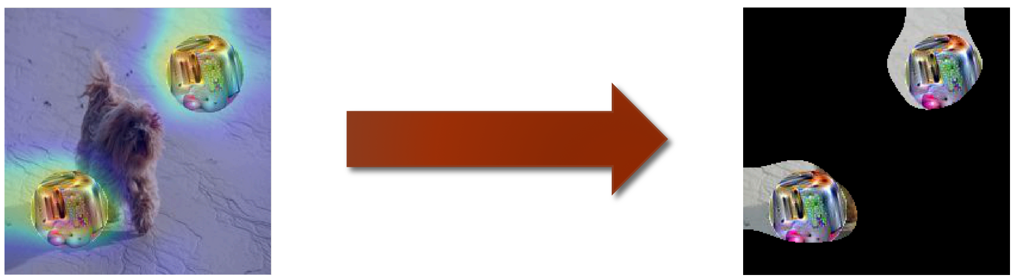
Example:
Adversarial patch
[Brown et al., 2018]



# Can we detect such attacks?

Observation: To be robust to physical transforms, the attack has to be very “salient”

⇒ Use model interpretability to extract salient regions



Problem: this might also extract “real” objects

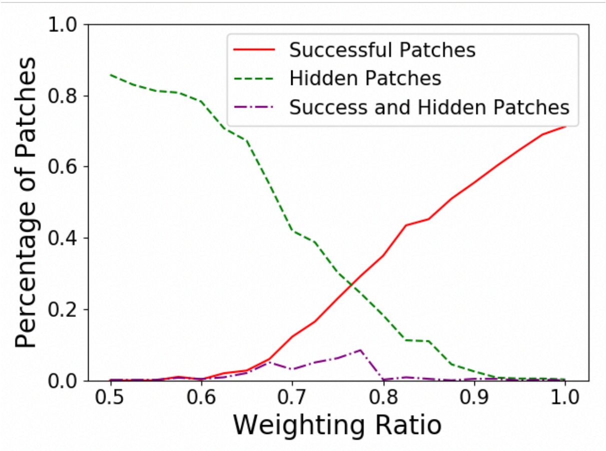
⇒ Add the extracted region(s) onto some test images and check how often this “hijacks” the true prediction

Chou **et al.** “SentiNet: Detecting Localized Universal Attacks Against Deep Learning Systems”, 2018

# Does it work?

It seems so...

- Generating a patch that avoids detection harms the patch's universality




- Also works for some forms of "trojaning" attacks

- But:
  - Very narrow threat model
  - Somewhat complex system so hard to say if we've thought of all attacks

Chou **et al.** "SentiNet: Detecting Localized Universal Attacks Against Deep Learning Systems", 2018

# Conclusions

The "expectimax $L_p$" game has proven more challenging than expected

- We shouldn't forget that this is a "toy" problem
  - Solving it doesn't get us secure ML (in most settings)
- Current defenses break down as soon as one of the game's assumptions is invalidated
  - E.g., robustness to more than one perturbation type
- Over-optimizing a standard benchmark can be harmful
  - E.g., invariance adversarial examples
- Thinking about real cyber-physical attacker constraints might lead to interesting defense ideas

**Maybe we don't need 10x more papers!**