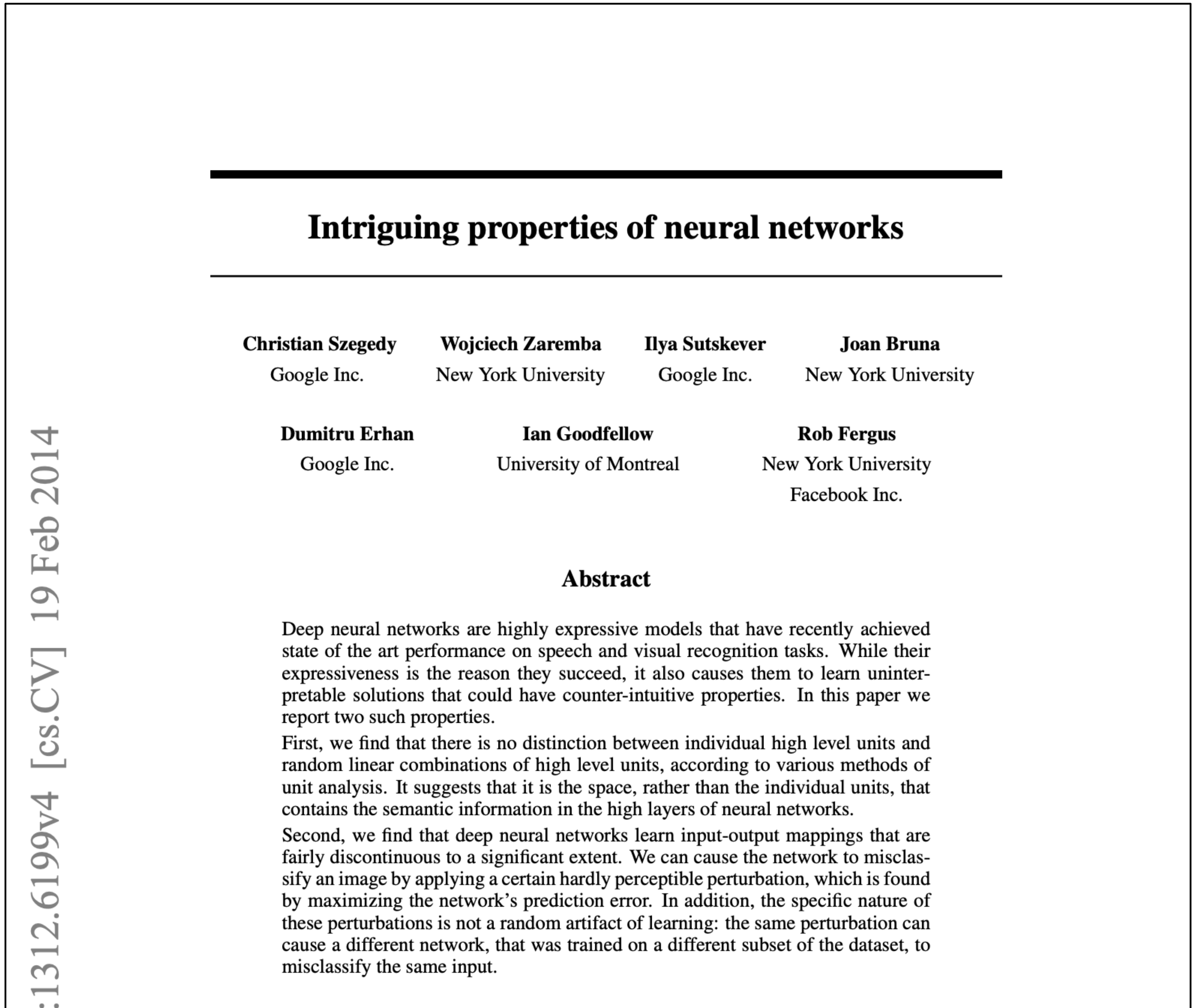
Trustworthy Machine Learning

Adversarial Examples and Heuristic Adversarial Learning

Sangdon Park

POSTECH

Intriguing Properties of Neural Networks



Intriguing properties of neural networks

Christian Szegedy Google Inc.
Wojciech Zaremba New York University
Ilya Sutskever Google Inc.
Joan Bruna New York University
Dumitru Erhan Google Inc.
Ian Goodfellow University of Montreal
Rob Fergus New York University, Facebook Inc.

Abstract

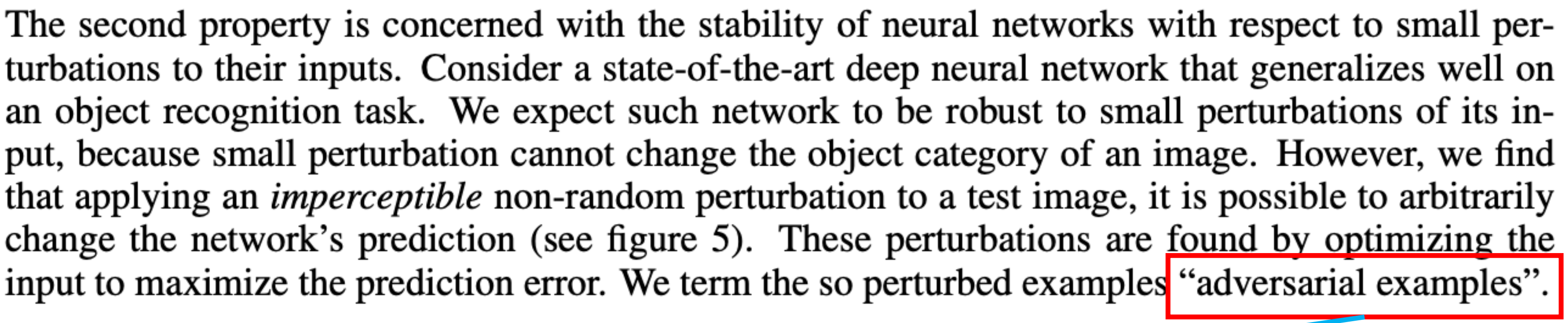
Deep neural networks are highly expressive models that have recently achieved state of the art performance on speech and visual recognition tasks. While their expressiveness is the reason they succeed, it also causes them to learn uninterpretable solutions that could have counter-intuitive properties. In this paper we report two such properties.

First, we find that there is no distinction between individual high level units and random linear combinations of high level units, according to various methods of unit analysis. It suggests that it is the space, rather than the individual units, that contains the semantic information in the high layers of neural networks.

Second, we find that deep neural networks learn input-output mappings that are fairly discontinuous to a significant extent. We can cause the network to misclassify an image by applying a certain hardly perceptible perturbation, which is found by maximizing the network's prediction error. In addition, the specific nature of these perturbations is not a random artifact of learning: the same perturbation can cause a different network, that was trained on a different subset of the dataset, to misclassify the same input.

One of 35 papers, presented at 2nd International Conference on Learning Representations (ICLR), held in 2014

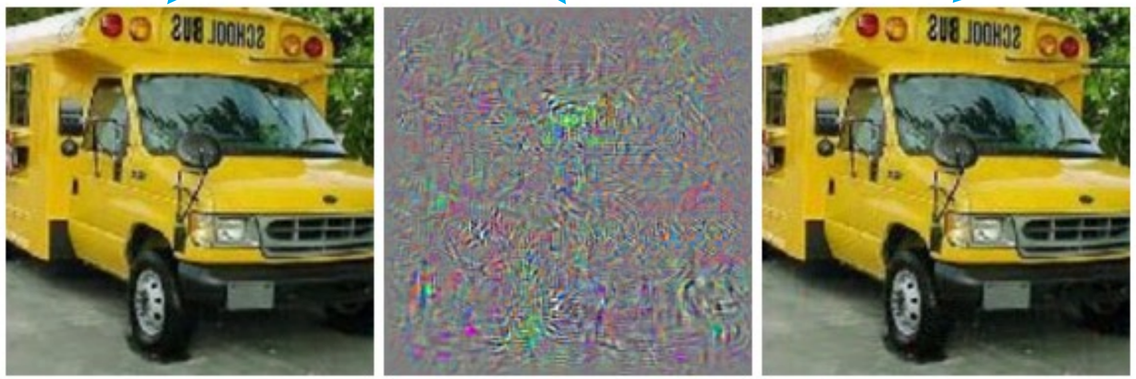
Adversarial Examples



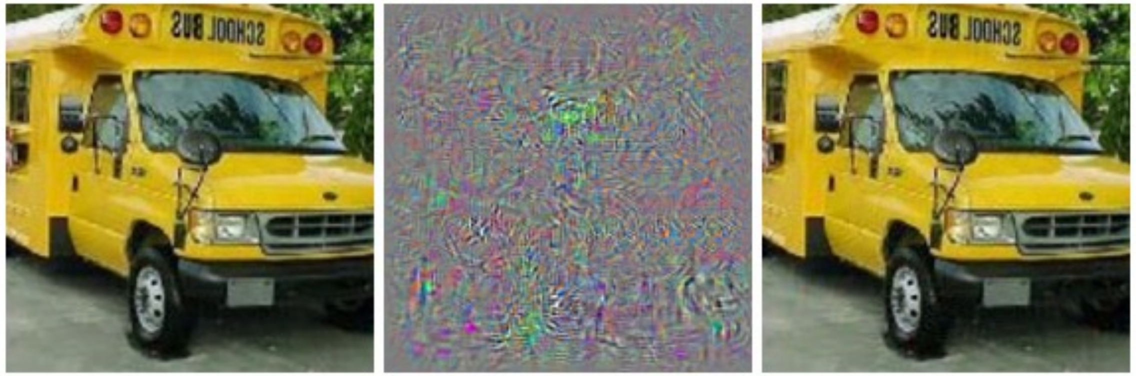
The second property is concerned with the stability of neural networks with respect to small perturbations to their inputs. Consider a state-of-the-art deep neural network that generalizes well on an object recognition task. We expect such network to be robust to small perturbations of its input, because small perturbation cannot change the object category of an image. However, we find that applying an *imperceptible* non-random perturbation to a test image, it is possible to arbitrarily change the network's prediction (see figure 5). These perturbations are found by optimizing the input to maximize the prediction error. We term the so perturbed examples "adversarial examples".

$$x_{adv} = x + \delta \quad \text{s.t.} \quad ||\delta|| \leq \epsilon$$

Adversarial perturbation



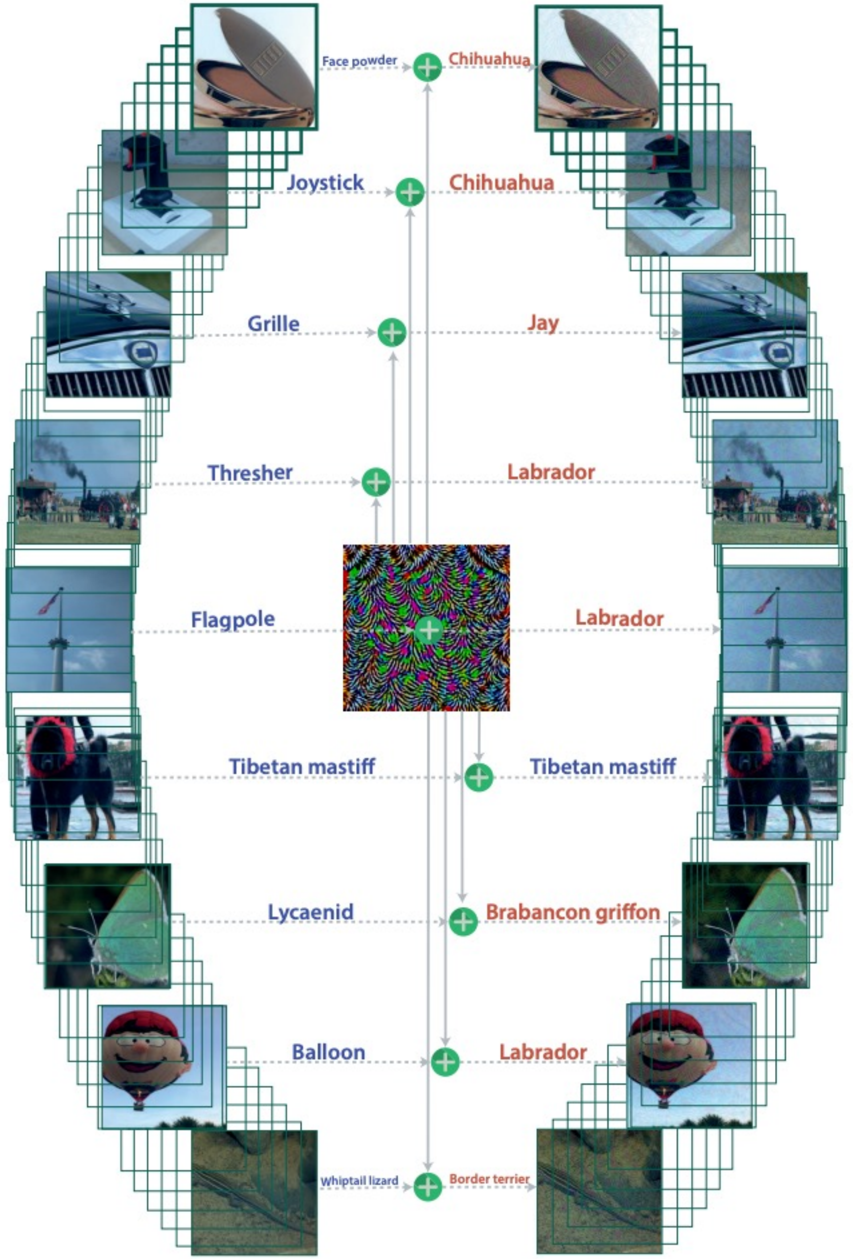
Why “Intriguing”?



- The adversarial perturbation is “imperceptible”.
  - Adversarial examples with larger perturbation provides trivial results.
  - The maxim perturbation value: e.g., $\frac{8}{255} \approx 0.03$
- The adversarial examples are transferable.

Why “Intriguing”?

- There exists one adversarial perturbation that makes the most images being misclassified



Universal adversarial perturbations (CVPR17)

Contents

- How to generate adversarial examples?

- How to (heuristically) learn a robust network to adversarial examples?

- What is the cause of adversarial examples?

- Is it practical?

Generating Adversarial Examples

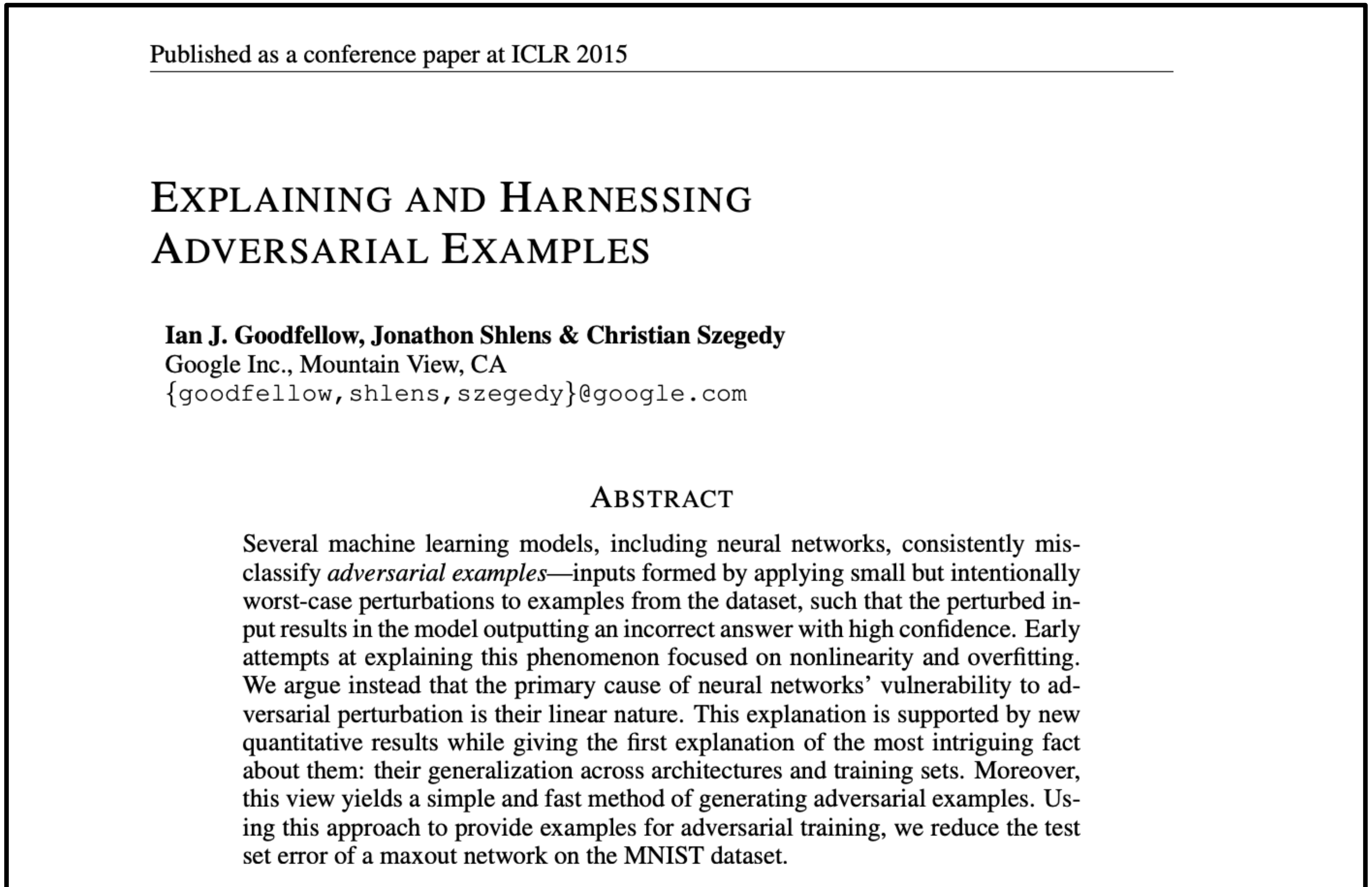
High-level Objective

$$\max_{\delta \in S} \ell(f, x + \delta, y)$$

- (x, y): a labeled example
- f: a classifier
- $\ell(f, x, y)$: loss
- S: a set of perturbations
- $x + \delta$: an adversarial example that is misclassified by f

FGSM: Fast Gradient Sign Method

THE LINEAR EXPLANATION OF ADVERSARIAL EXAMPLES



Published as a conference paper at ICLR 2015

EXPLAINING AND HARNESSING ADVERSARIAL EXAMPLES

Ian J. Goodfellow, Jonathon Shlens & Christian Szegedy
Google Inc., Mountain View, CA
{goodfellow,shlens,szegedy}@google.com

ABSTRACT

Several machine learning models, including neural networks, consistently misclassify *adversarial examples*—inputs formed by applying small but intentionally worst-case perturbations to examples from the dataset, such that the perturbed input results in the model outputting an incorrect answer with high confidence. Early attempts at explaining this phenomenon focused on nonlinearity and overfitting. We argue instead that the primary cause of neural networks' vulnerability to adversarial perturbation is their linear nature. This explanation is supported by new quantitative results while giving the first explanation of the most intriguing fact about them: their generalization across architectures and training sets. Moreover, this view yields a simple and fast method of generating adversarial examples. Using this approach to provide examples for adversarial training, we reduce the test set error of a maxout network on the MNIST dataset.

FGSM

- Objective:

$$\max_{\delta:\,\|\delta\|_\infty \leq \epsilon} \ell(f, x + \delta, y)$$

- Solution:

$$\delta = \epsilon \cdot sign(\nabla_x \ell(f, x, y))$$

- Intuition: Linearize the loss function around the parameter f.

FGSM: Optimality Analysis

- Tayler expansion of a function $f(x)$ at a:

$$f(x) \approx f(a) + \frac{f'(a)}{1!}(x - a) + \frac{f''(a)}{2!}(x - a)^2 + \cdots$$

- The first order Tayler expansion of loss at an example x :

$$\ell(f, x + \delta, y) =: \ell(x + \delta) \approx \ell(x) + \nabla_x \ell(x)^T \delta$$

A Unified Gradient Regularization Family for Adversarial Examples. ICDM2015.

FGSM: Optimality Analysis

- The first order Tayler expansion of loss at an example x_0 :

$$\ell(f, x+\delta, y) =: \ell(x+\delta) \approx \ell(x) + \nabla_x \ell(x)^T \delta$$

- We have

$$\max_{\delta:\, \|\delta\|_\infty \le \epsilon} \ell(f, x+\delta, y) \approx \max_{\delta:\, \|\delta\|_\infty \le \epsilon} \ell(x) + \nabla_x \ell(x)^T \delta$$

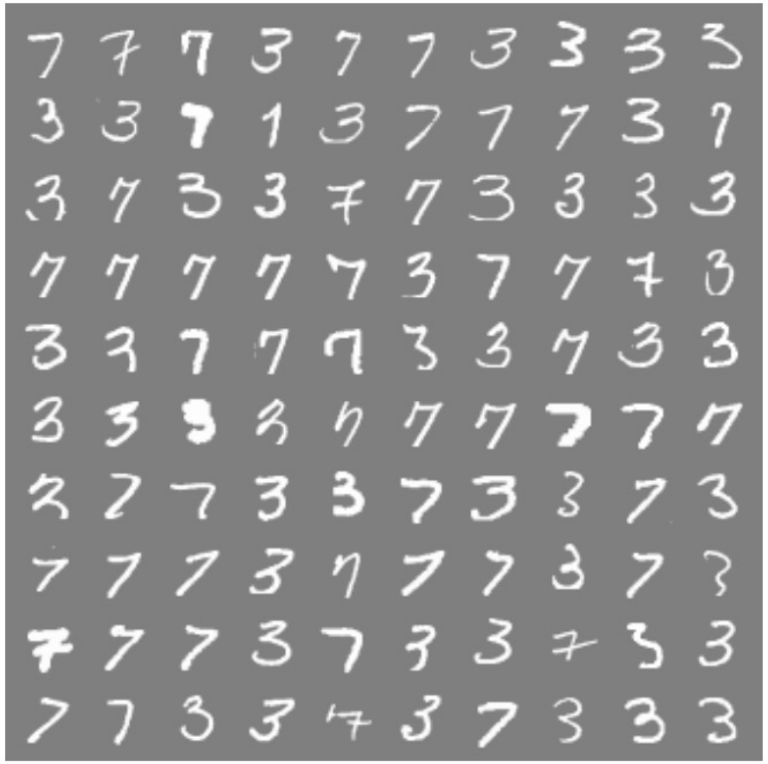
  - Clearly linear
  - Use the standard Lagrangian multiplier method

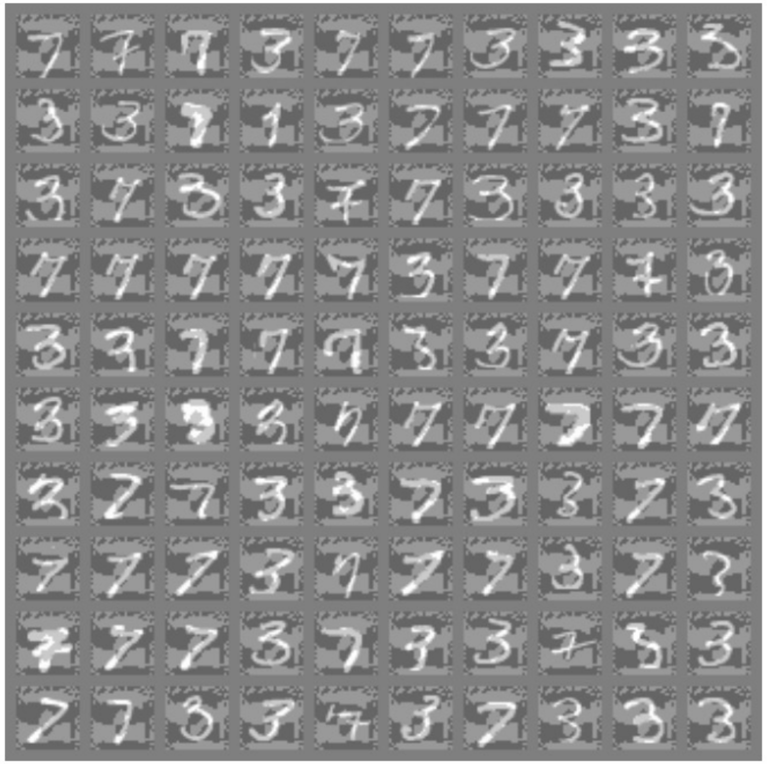
- Solution:

$$\delta = \epsilon \cdot sign(\nabla_x \ell(x))$$

FGSM Results

Error: 1.6%





Error: 99% (FGSM with $\epsilon = 0.25$)

- Results with a linear model
  - FGSM generates optimal perturbations.

General Framework

Gradient regularization family (1/2)

$$\delta = \epsilon \cdot sign\big(\nabla_x \ell(f, x, y)\big) \left(\frac{\nabla_x \ell(f, x, y)}{\|\nabla_x \ell(f, x, y)\|_{p^*}} \right)^{\frac{1}{p-1}}$$

- p^* is the dual of p, i.e., $\frac{1}{p^*} + \frac{1}{p} = 1$.

A Unified Gradient Regularization Family for Adversarial Examples. ICDM2015.

General Framework

Gradient regularization family (2/2)

$$\lim_{p\to\infty} \epsilon \cdot sign\big(\nabla_x \ell(f,x,y)\big)\left(\frac{\nabla_x \ell(f,x,y)}{\|\nabla_x \ell(f,x,y)\|_{p^*}}\right)^{\frac{1}{p-1}}$$

$$= \epsilon \cdot sign\big(\nabla_x \ell(f,x,y)\big)\left(\frac{\nabla_x \ell(f,x,y)}{\|\nabla_x \ell(f,x,y)\|_{1}}\right)^{0}$$

$$= \epsilon \cdot sign\big(\nabla_x \ell(f,x,y)\big)$$

We have FGSM!

A Unified Gradient Regularization Family for Adversarial Examples. ICDM2015.

“Iterative” FGSM

- FGSM finds an adversarial example under the “linear” assumption
- The loss landscape is more complex
  - “Iterative” FGSM (by Google)
    - ADVERSARIAL MACHINE LEARNING AT SCALE (ICLR2017)
  - Project Gradient Descent (PGD)
    - Towards Deep Learning Models Resistant to Adversarial Attacks (ICLR2018)

PGD

- One-step attack

$$x + \epsilon \cdot sign(\nabla_x \ell(f, x, y))$$

- Multi-step attack

$$x^{t+1} = \Pi_{x+S}(x^t + \alpha \cdot sign(\nabla_x \ell(f, x, y)))$$

Towards Deep Learning Models Resistant to Adversarial Attacks (ICLR2018)

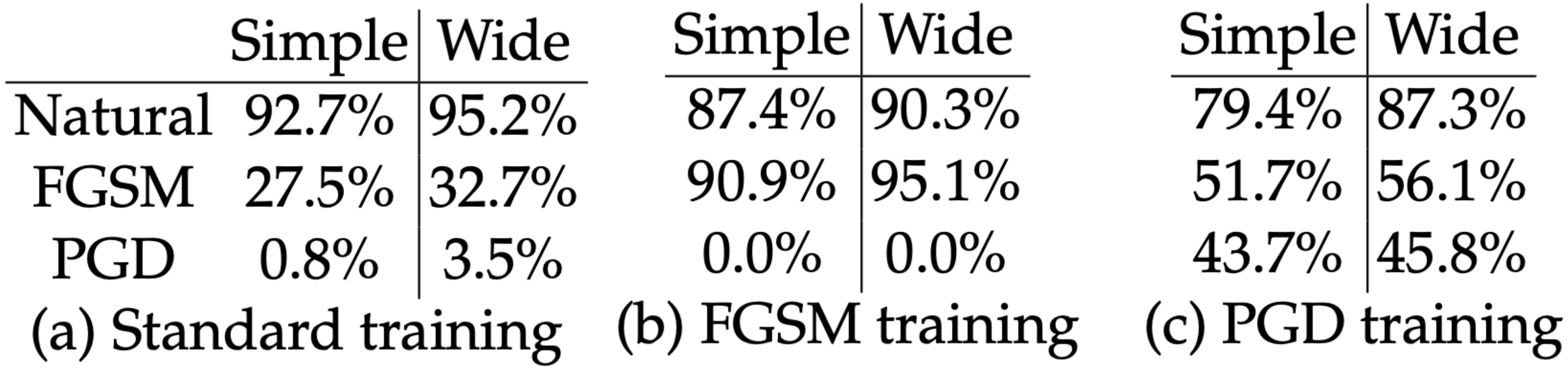
Adversarial Training

- Objective:

$$\min_{f} E\left[\max_{\delta \in S} \ell(f, x+\delta, y)\right]$$

Approximate via an attack algorithm (e.g., PGD)

CIFAR10 ($\epsilon = 8/255$)



	Simple	Wide
Natural	92.7%	95.2%
FGSM	27.5%	32.7%
PGD	0.8%	3.5%

(a) Standard training

Simple	Wide
87.4%	90.3%
90.9%	95.1%
0.0%	0.0%

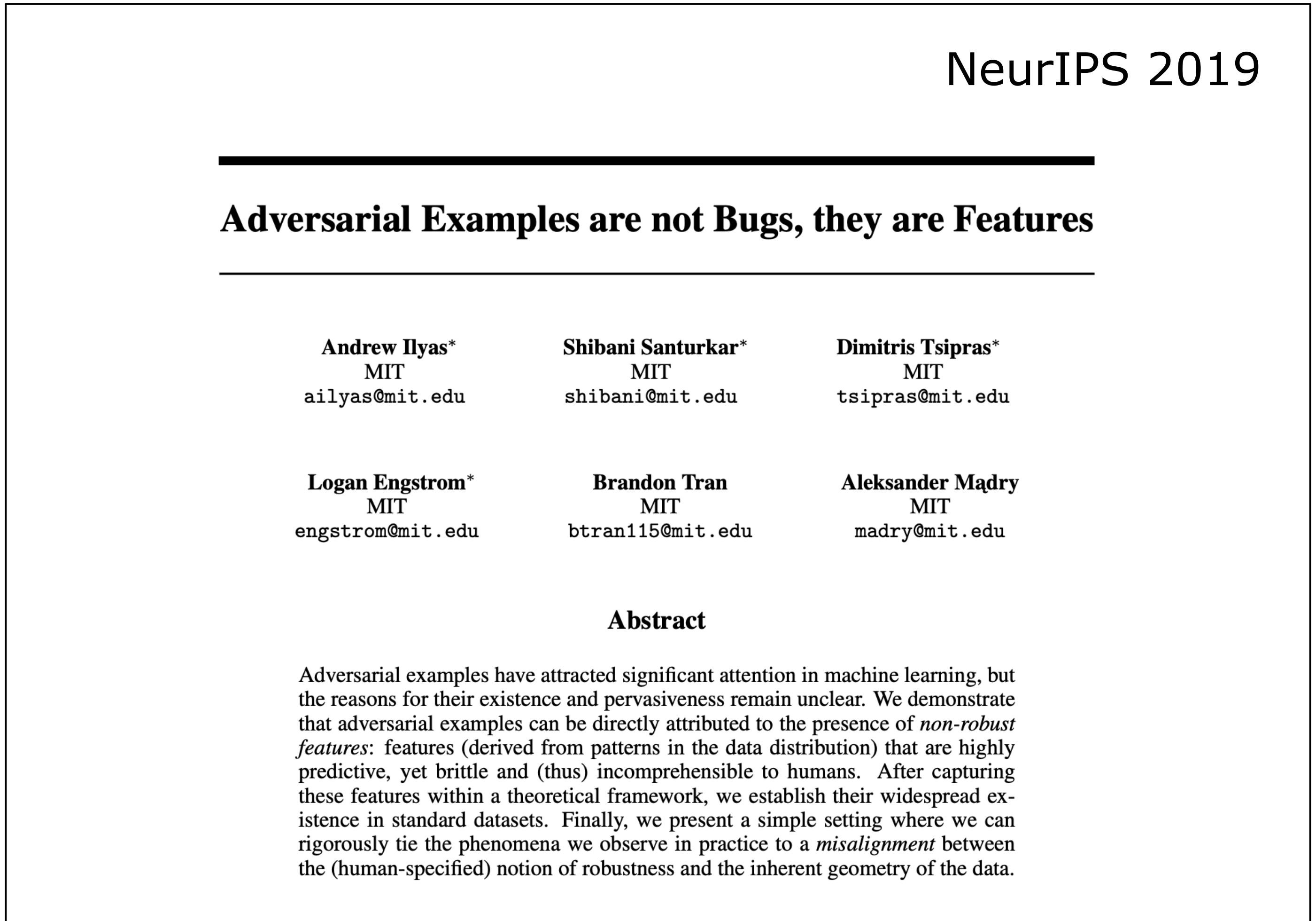
(b) FGSM training

Simple	Wide
79.4%	87.3%
51.7%	56.1%
43.7%	45.8%

(c) PGD training

What’s the Cause of Adversarial Examples?

Adversarial vulnerability is a direct result of sensitivity to well-generalizing features in the data.



NeurIPS 2019

Adversarial Examples are not Bugs, they are Features

Andrew Ilyas*
MIT
ailyas@mit.edu

Shibani Santurkar*
MIT
shibani@mit.edu

Dimitris Tsipras*
MIT
tsipras@mit.edu

Logan Engstrom*
MIT
engstrom@mit.edu

Brandon Tran
MIT
btran115@mit.edu

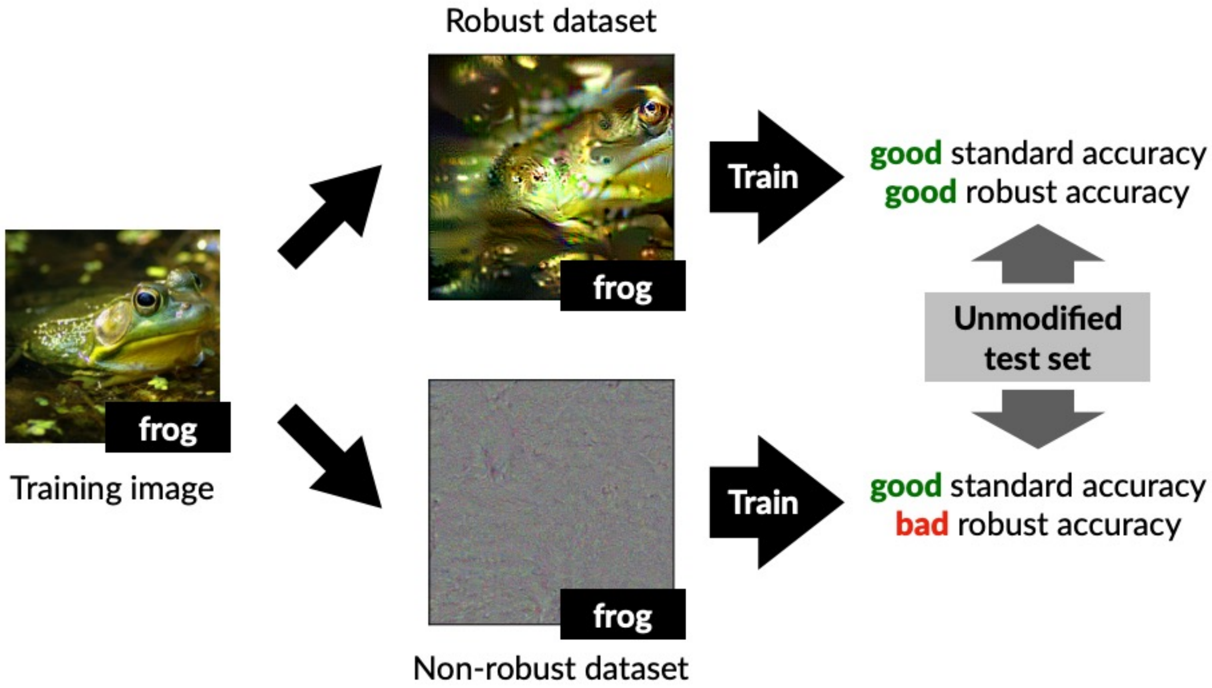
Aleksander Mądry
MIT
madry@mit.edu

Abstract

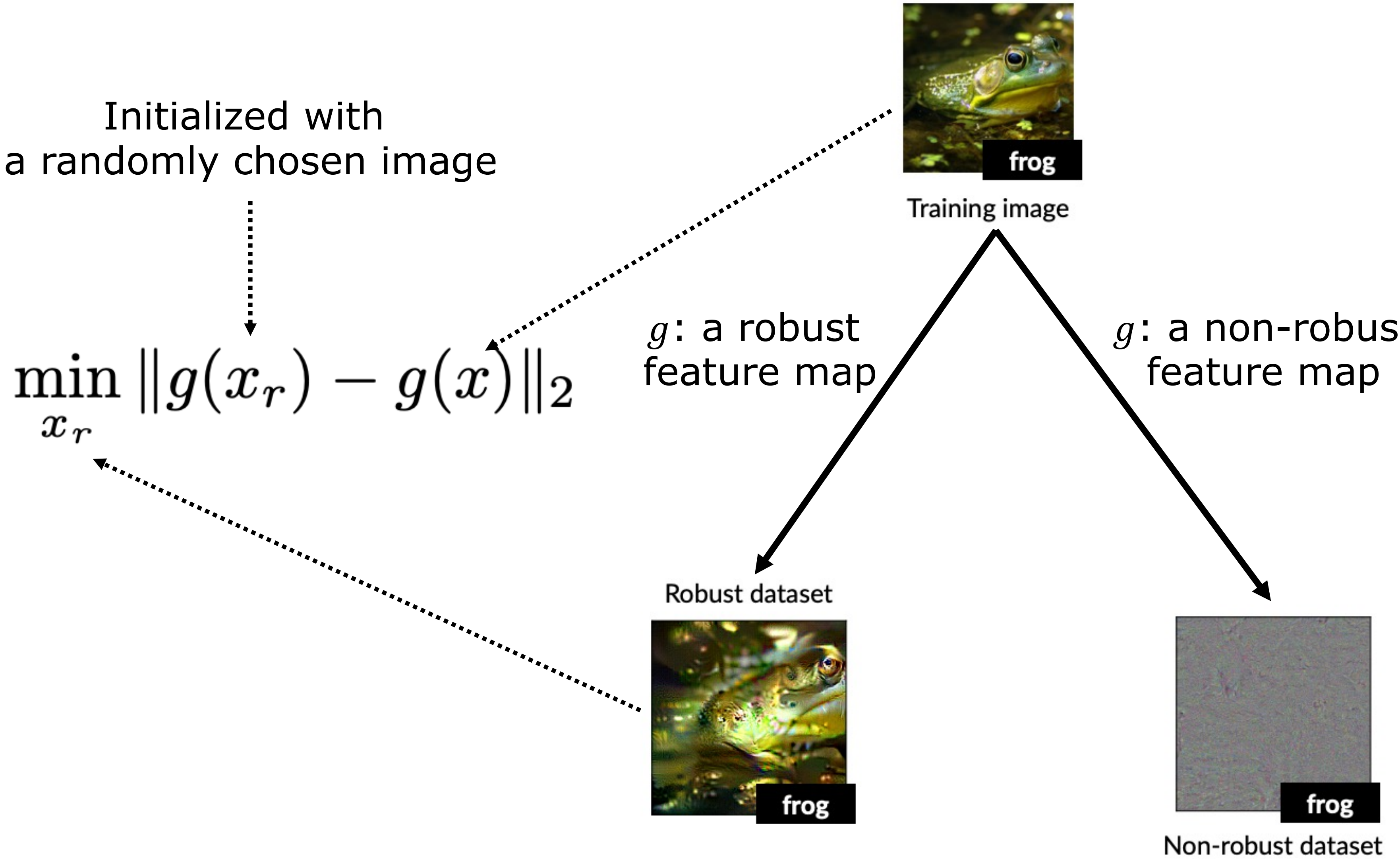
Adversarial examples have attracted significant attention in machine learning, but the reasons for their existence and pervasiveness remain unclear. We demonstrate that adversarial examples can be directly attributed to the presence of *non-robust features*: features (derived from patterns in the data distribution) that are highly predictive, yet brittle and (thus) incomprehensible to humans. After capturing these features within a theoretical framework, we establish their widespread existence in standard datasets. Finally, we present a simple setting where we can rigorously tie the phenomena we observe in practice to a *misalignment* between the (human-specified) notion of robustness and the inherent geometry of the data.

Experiment Setup

Claim: There are two features: (1) robust features and (2) non-robust features; the non-robust features contribute to adversarial examples!



How to Construct Datasets



Initialized with
a randomly chosen image

$$\min_{x_r} \|g(x_r) - g(x)\|_2$$

g: a robust
feature map

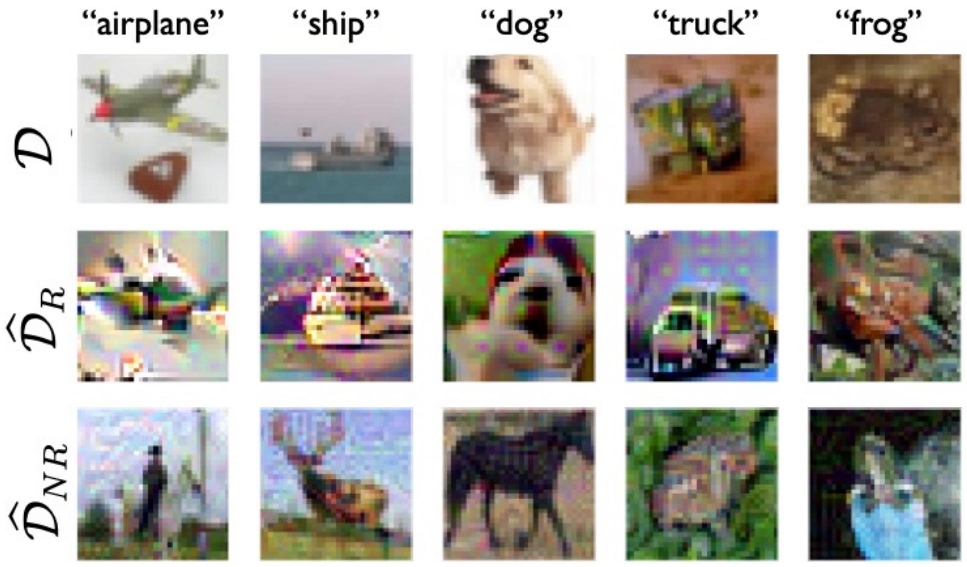
g: a non-robust
feature map

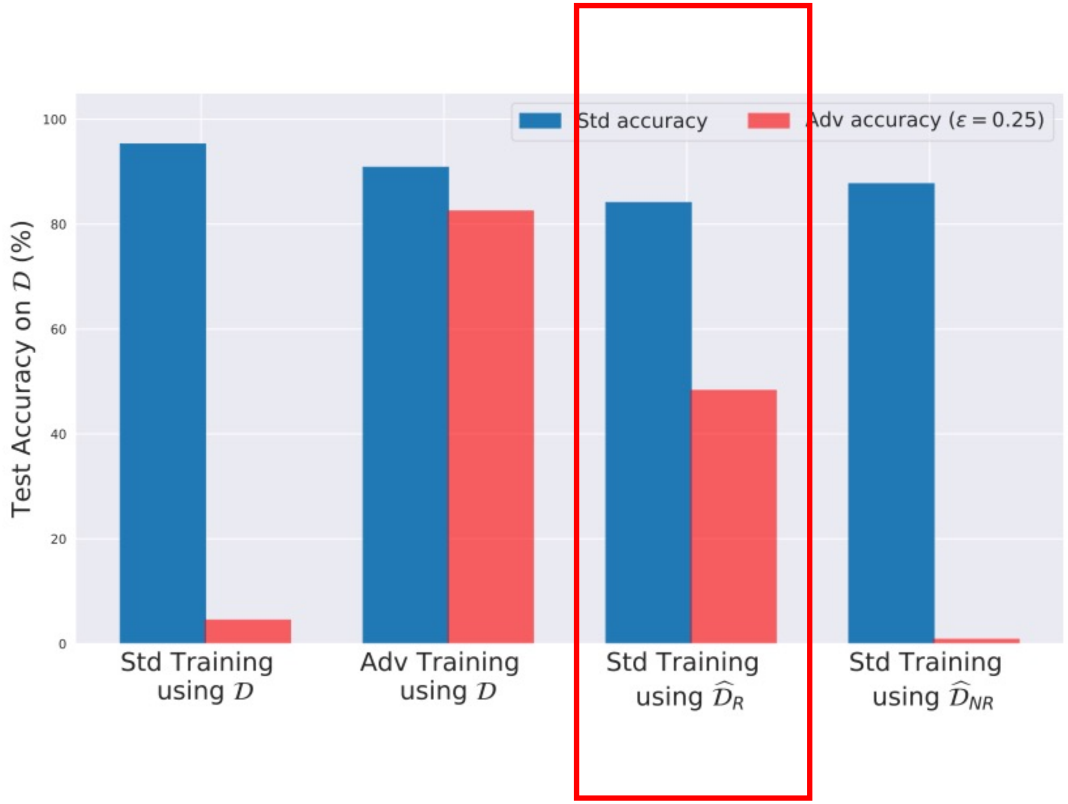
Training image

Robust dataset

Non-robust dataset

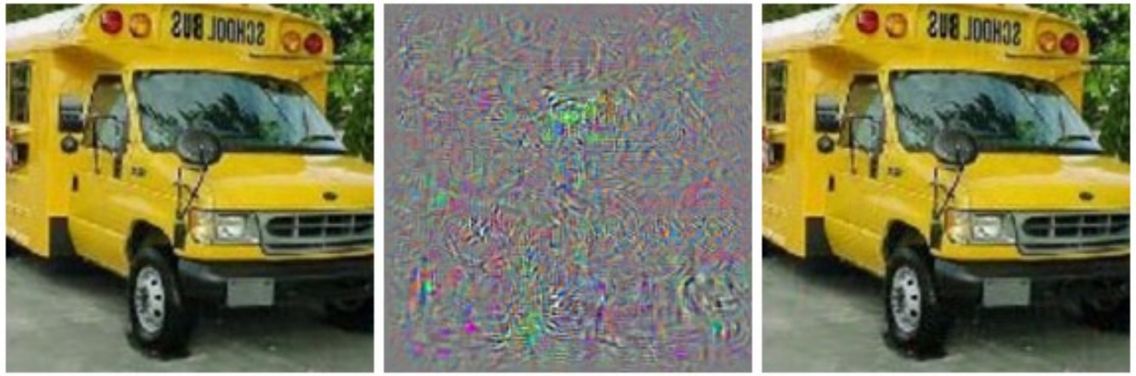
Results





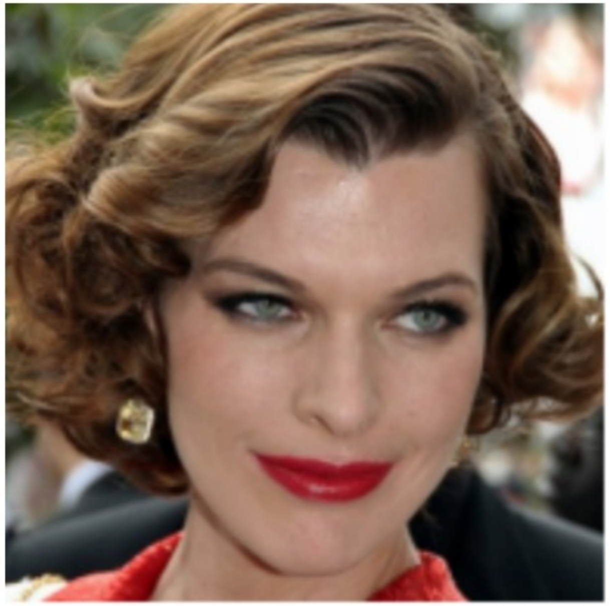
- The standard model picks (noisy-looking) non-robust features to classify images; thus, it is susceptible to adversarial perturbations.

Wait! Practical?

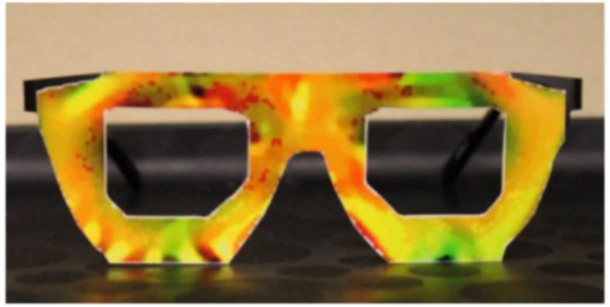


How can an attacker inject an adversarial example in practice?

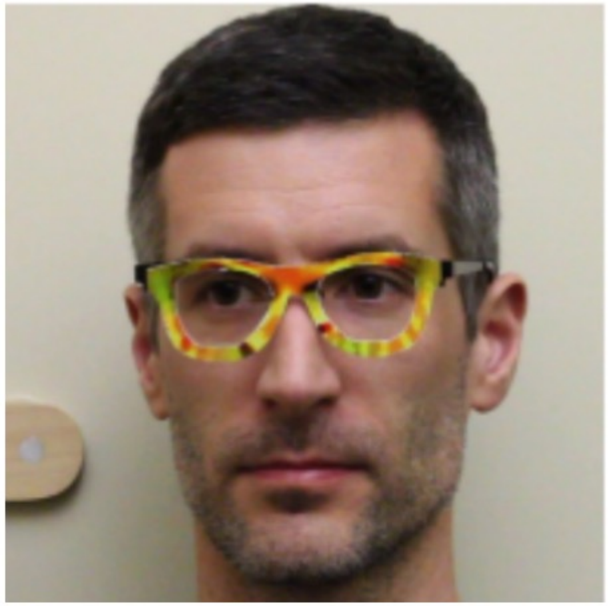
Adversarial Glasses



(a) "Milla Jovovich"



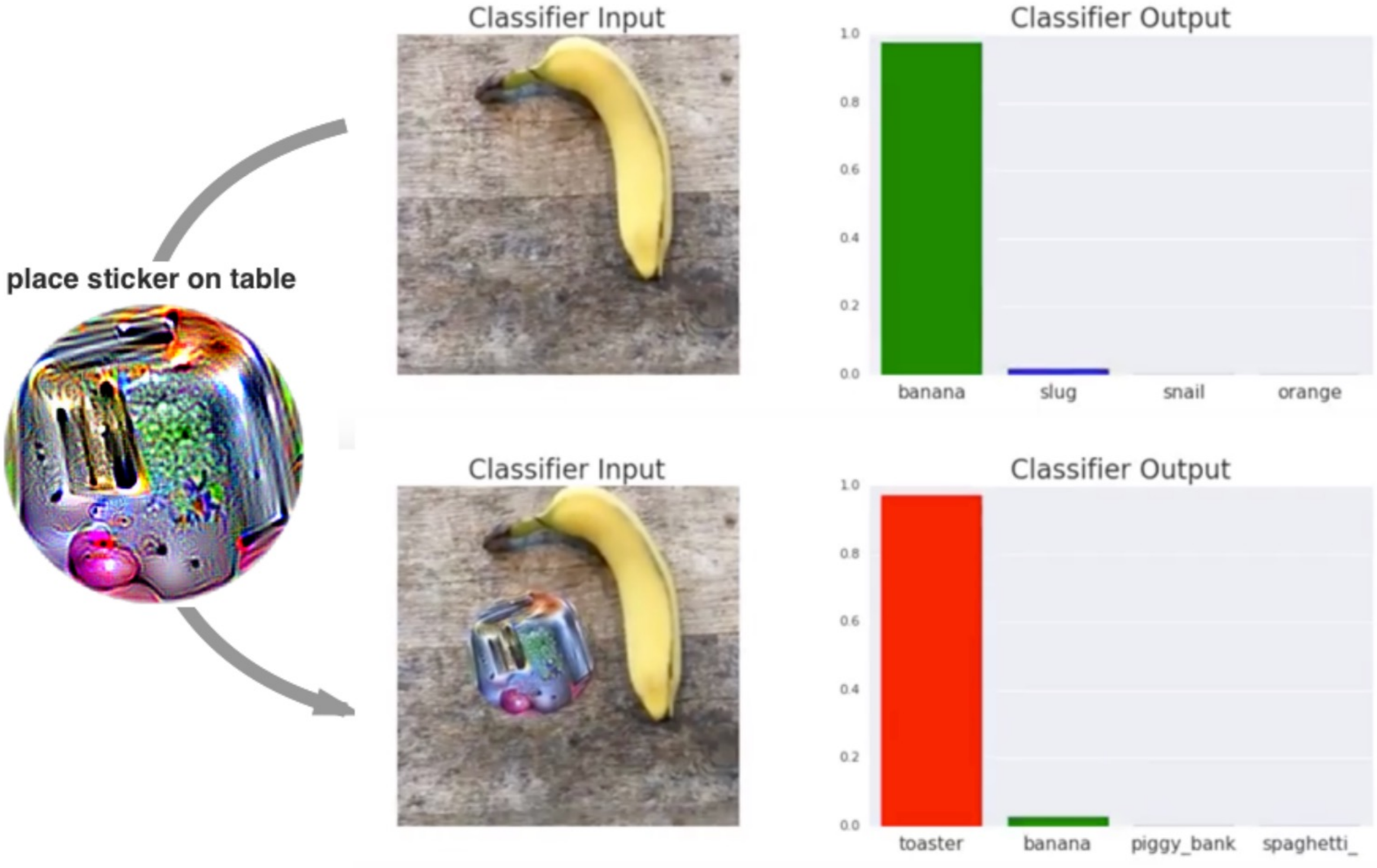
(b) Eyeglass frame



(c) Impersonating "Milla Jovovich"

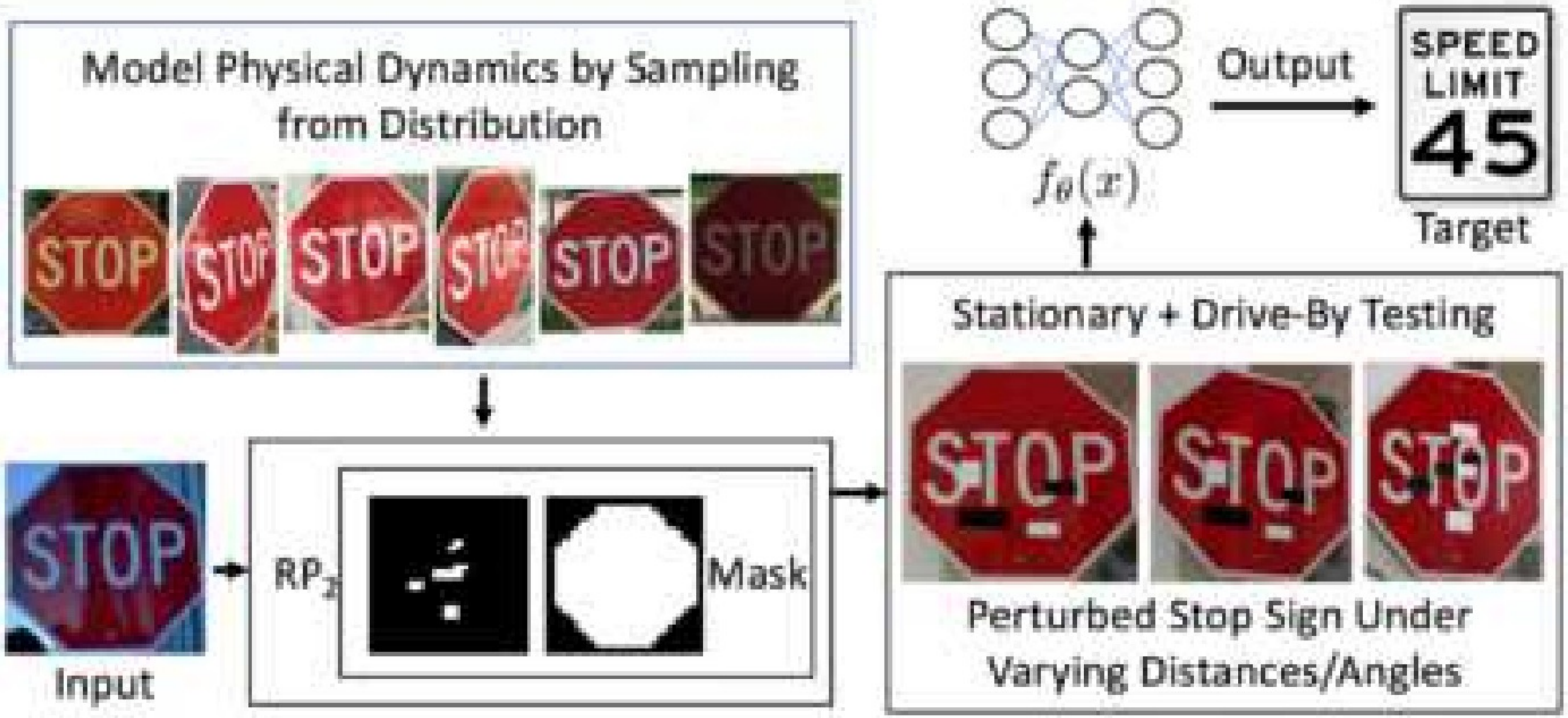
Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition (CCS16)

Adversarial Patch



Adversarial Patch (NIPS17 Workshop)

Physical Adversarial Examples

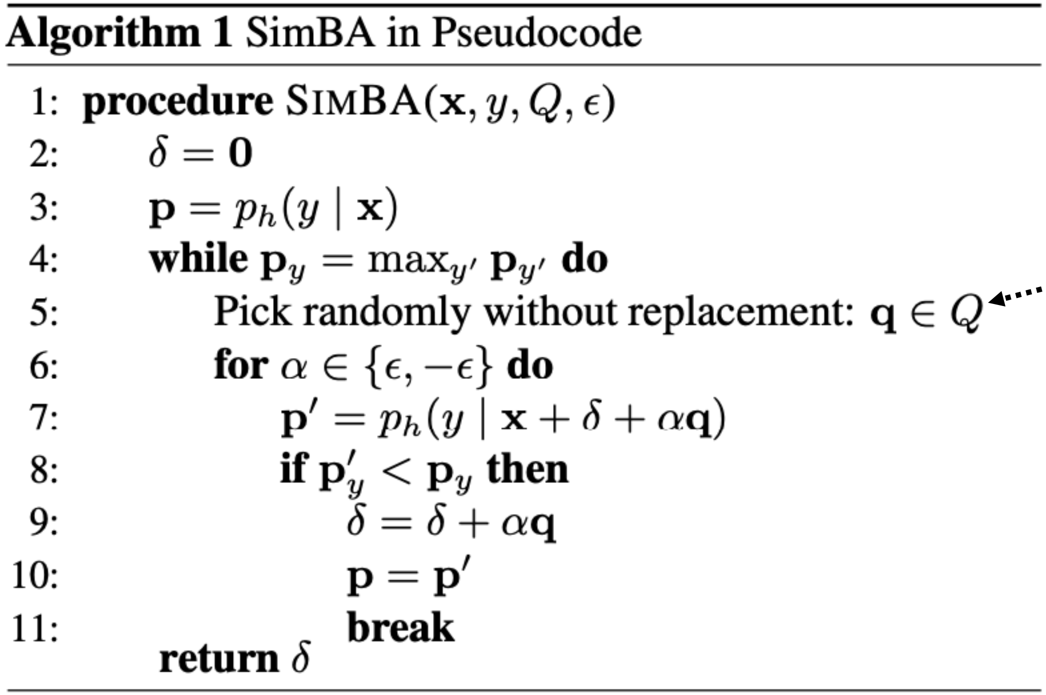

Model Physical Dynamics by Sampling from Distribution
STOP
Input
RP$_2$
Mask
Stationary + Drive-By Testing
Perturbed Stop Sign Under Varying Distances/Angles
$f_\theta(x)$
Output
SPEED LIMIT 45
Target

Robust Physical-World Attacks on Deep Learning Visual Classification (CVPR18)

Black-Box Attacks

- What if we don’t have a target model?

$$\min_{\delta} \ \ell_y(\mathbf{x}+\delta) \text{ subject to: } \ \|\delta\|_2 < \rho, \text{queries} \le B$$



Algorithm 1 SimBA in Pseudocode

```
procedure SIMBA($\mathbf{x}, y, Q, \epsilon$)
    $\delta = \mathbf{0}$
    $\mathbf{p} = p_h(y \mid \mathbf{x})$
    while $\mathbf{p}_y = \max_{y'} \mathbf{p}_{y'}$ do
        Pick randomly without replacement: $\mathbf{q} \in Q$
        for $\alpha \in \{\epsilon, -\epsilon\}$ do
            $\mathbf{p}' = p_h(y \mid \mathbf{x} + \delta + \alpha\mathbf{q})$
            if $\mathbf{p}'_y < \mathbf{p}_y$ then
                $\delta = \delta + \alpha\mathbf{q}$
                $\mathbf{p} = \mathbf{p}'$
                break
    return $\delta$
```

Observation: random noise in low frequency space is more likely to be adversarial

Simple Black-box Adversarial Attacks (ICML19)

Conclusion

- Small adversarial perturbations degrade the perforance of predictors.
- Adversarial perturbations are realizable.
  - Phyiscal adversarial examples
  - Eyeglass
  - Adversarial patch
- Even without complete knowledge on a model, we can generate adversarial perturbations.
- How can we learn a neural network that is robust to adversarial perturbations *with guarantees*?