Generative Adversarial Networks

Sewoong Oh

CSE/STAT 416
University of Washington

Deep learning

- So far we studied Deep Supervised Learning
  - Classification
  - Regression

- How do we do Unsupervised Learning with Deep Neural Networks?
  - Breakthrough:
    - Generative Adversarial Networks (GANs)

- We start with a slightly different story: adversarial examples

Adversarial Examples

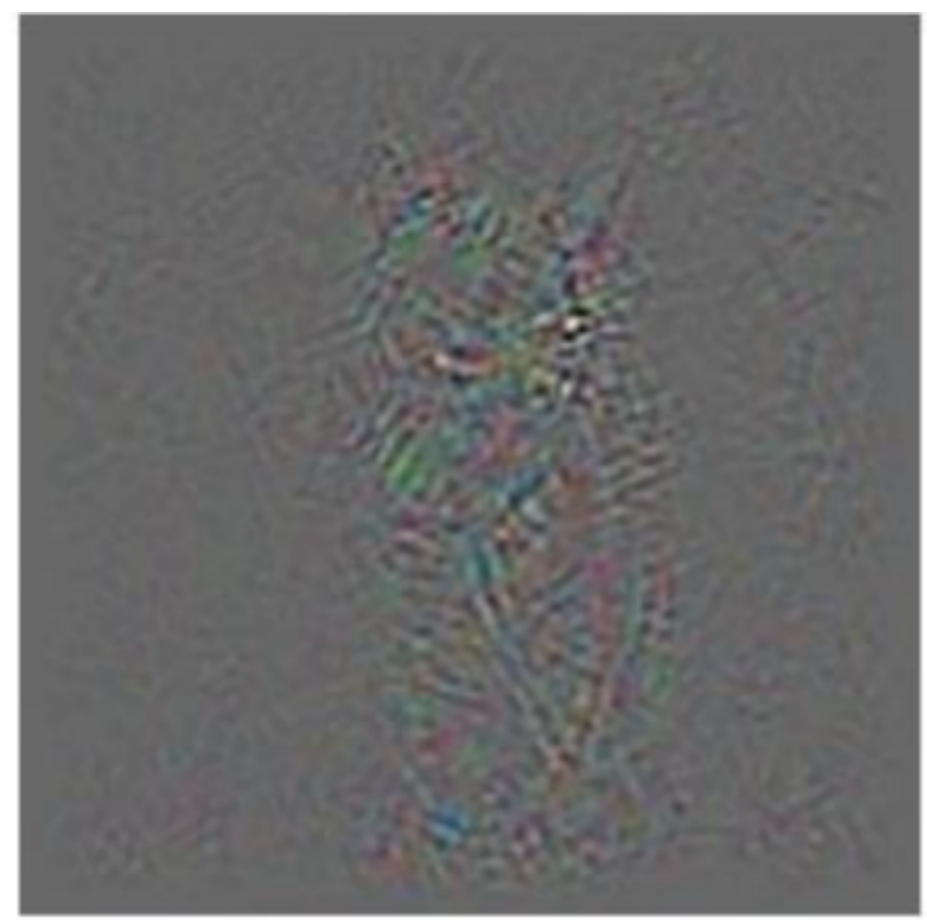
Consider a case where an adversary knows some combination of

- the training data
- the trained mode weights
- the trained model as a black box

- the goal of an adversary is to make the classifier fail (sometimes with emphasis on particular classes or examples)
- Timeline:
  - "Adversarial Classification" Dalvi et al 2004: fool spam filter
  - "Evasion Attacks Against Machine Learning at Test Time" Biggio 2013: fool neural nets
  - Szegedy et al 2013: fool ImageNet classifiers imperceptibly
  - Goodfellow et al 2014: cheap, closed form attack

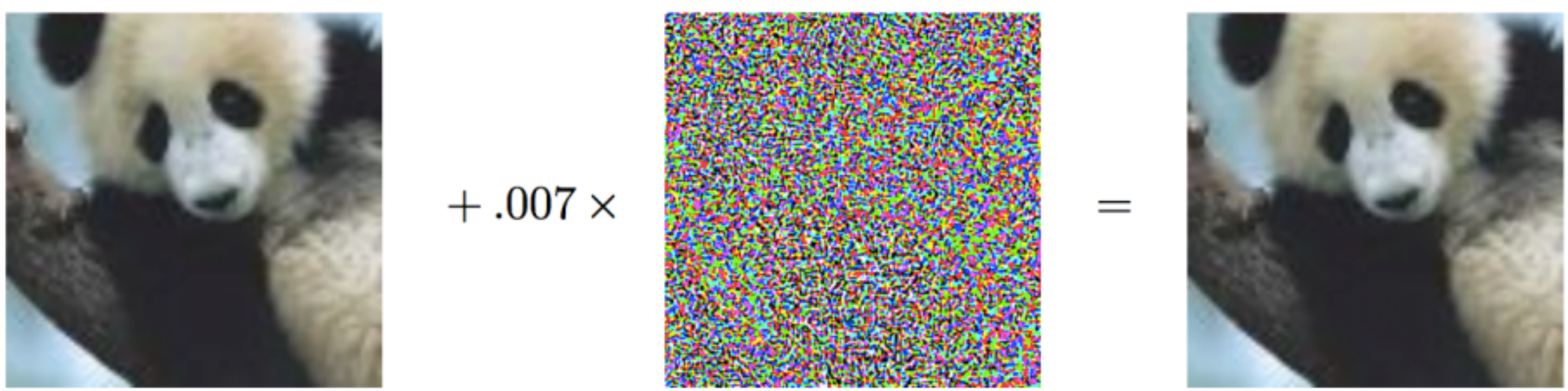
Adversarial testing examples

Consider computing the gradient,
but not on the weights as we do in training,
instead on the **input example**, which itself is hard to interpret



Adversarial testing examples

- consider an experiment where we do gradient **ascent** on the cross-entropy loss to **minimize** the probability that it is correctly classified
- concretely, perturb the image slightly by taking the sign of the gradient with a small scaling constant



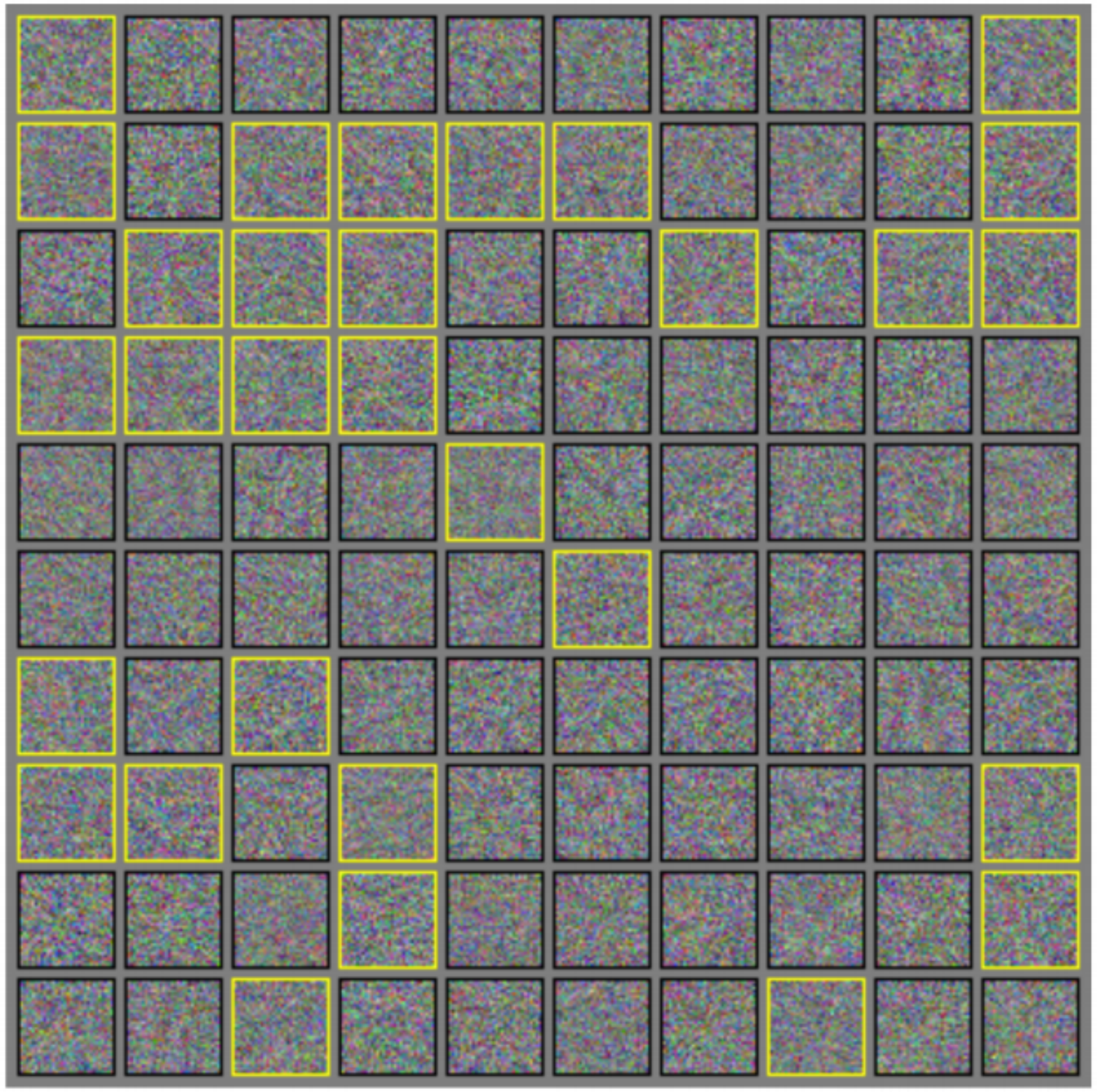
$\boldsymbol{x}$ "panda" 57.7% confidence

$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$ "nematode" 8.2% confidence

$\boldsymbol{x} + \epsilon\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$ "gibbon" 99.3 % confidence

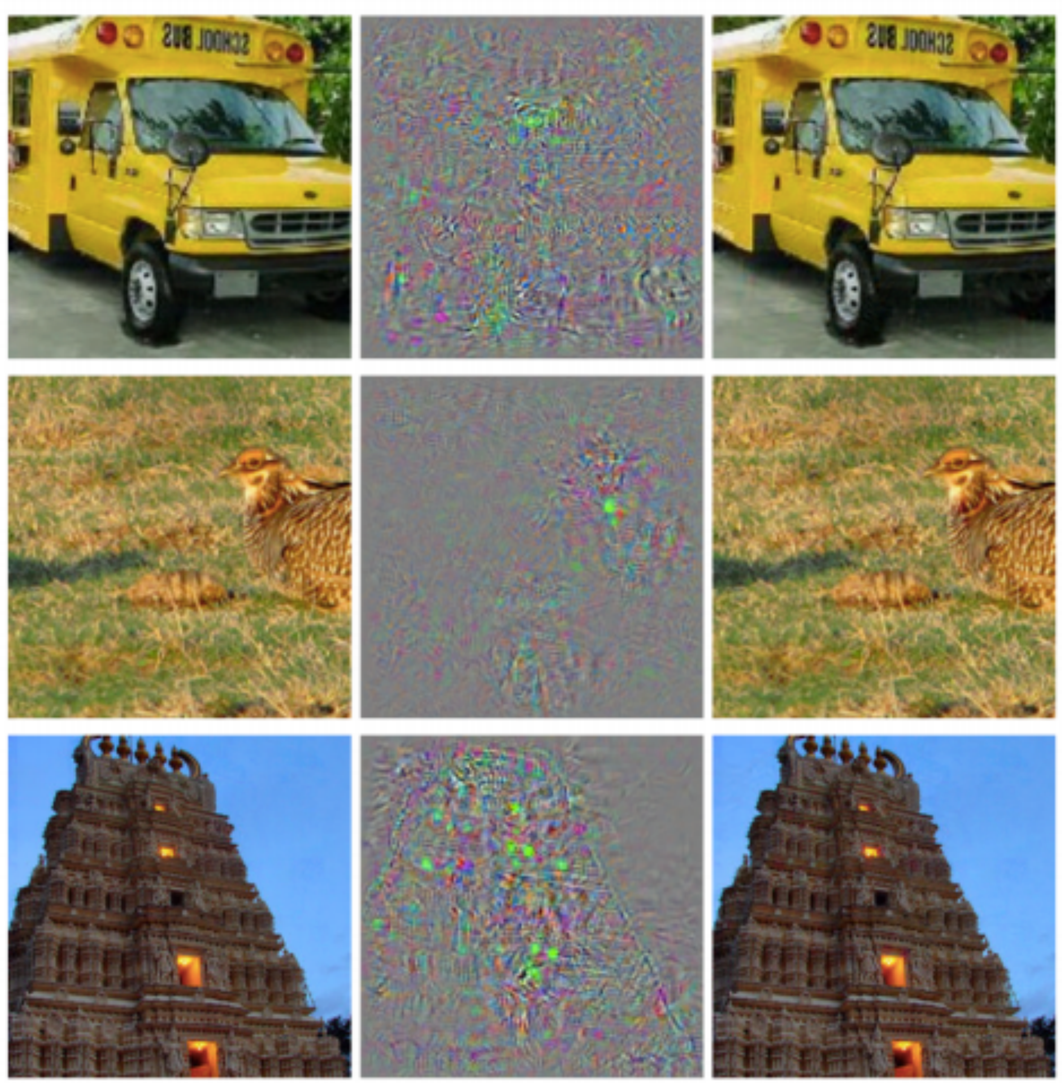
Adversarial testing examples

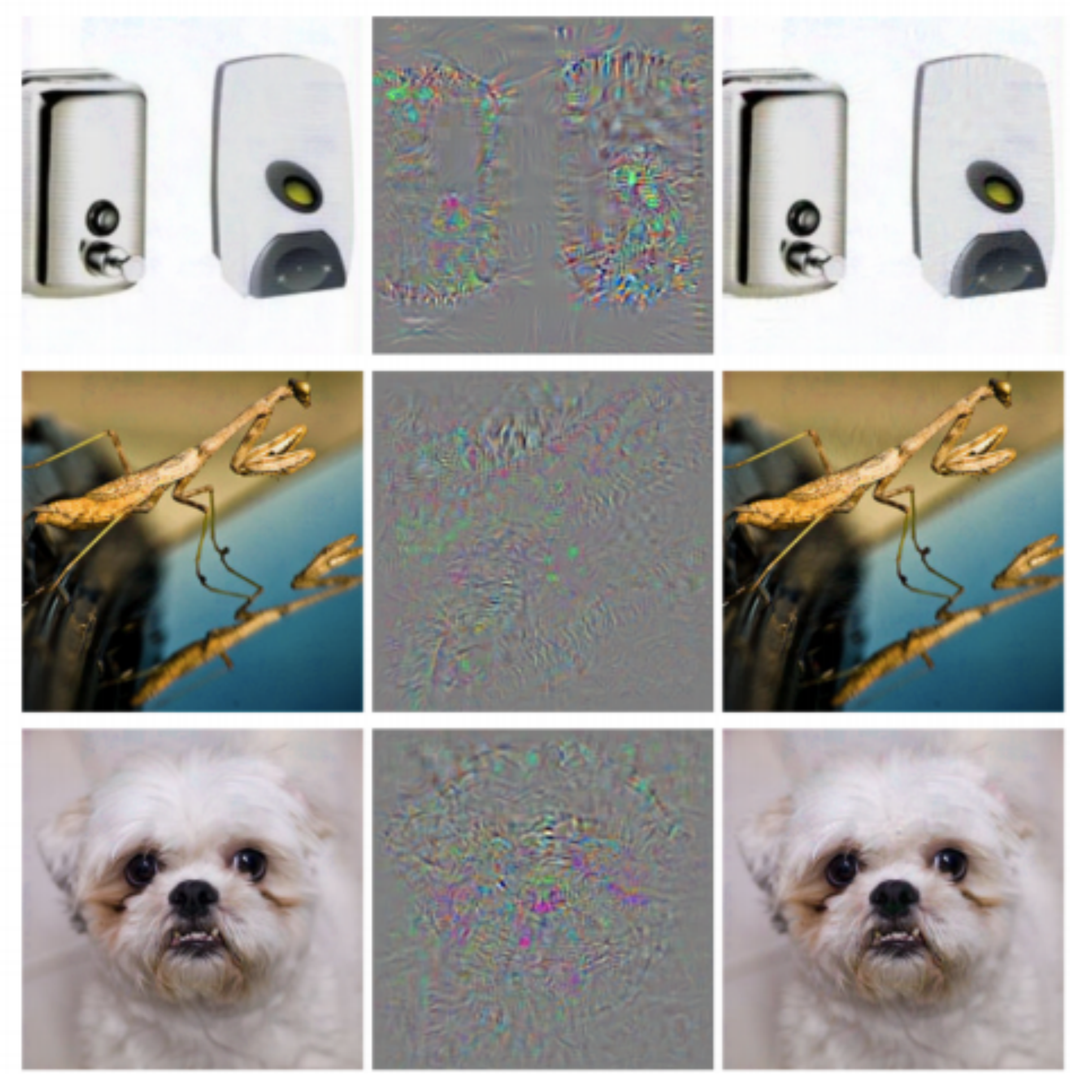
- In another experiment, you can start with a random noise and take **one** gradient step
- this often produces a confident classification
- the images outlined by yellow are classified as "airplane" with >50% confidence



Adversarial testing examples

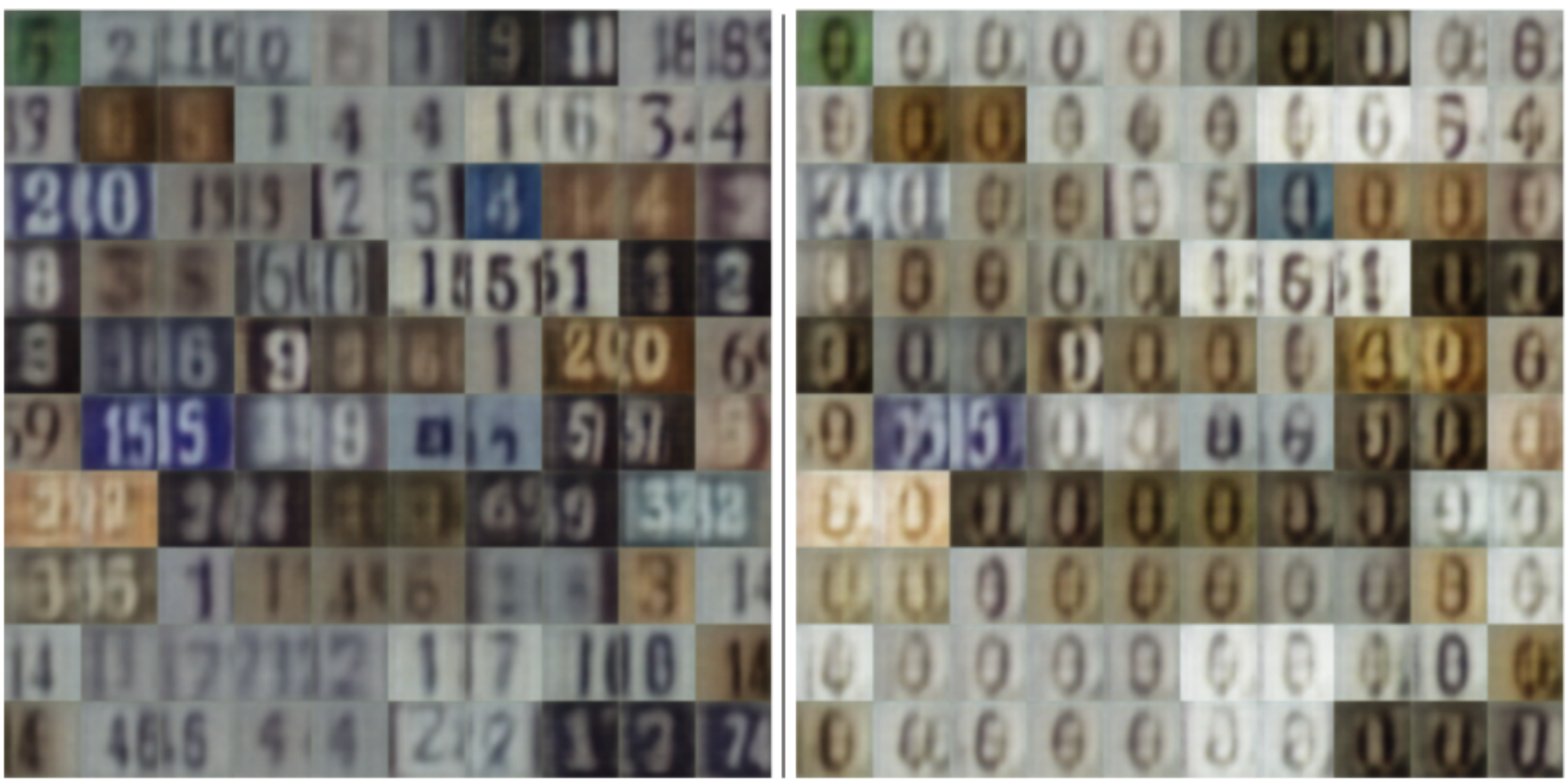
- In another experiment, you can have **targeted adversarial examples**, to misclassify examples to a specific target class
- the adversarial examples are misclassified as ostriches, and in the middle we show the perturbation times ten.





Adversarial testing examples

- consider a variational autoencoder for images, whose goal is to compress the image and then reconstruct it back
- one can create adversarial images that is reconstructed (after compression) as an entirely different image

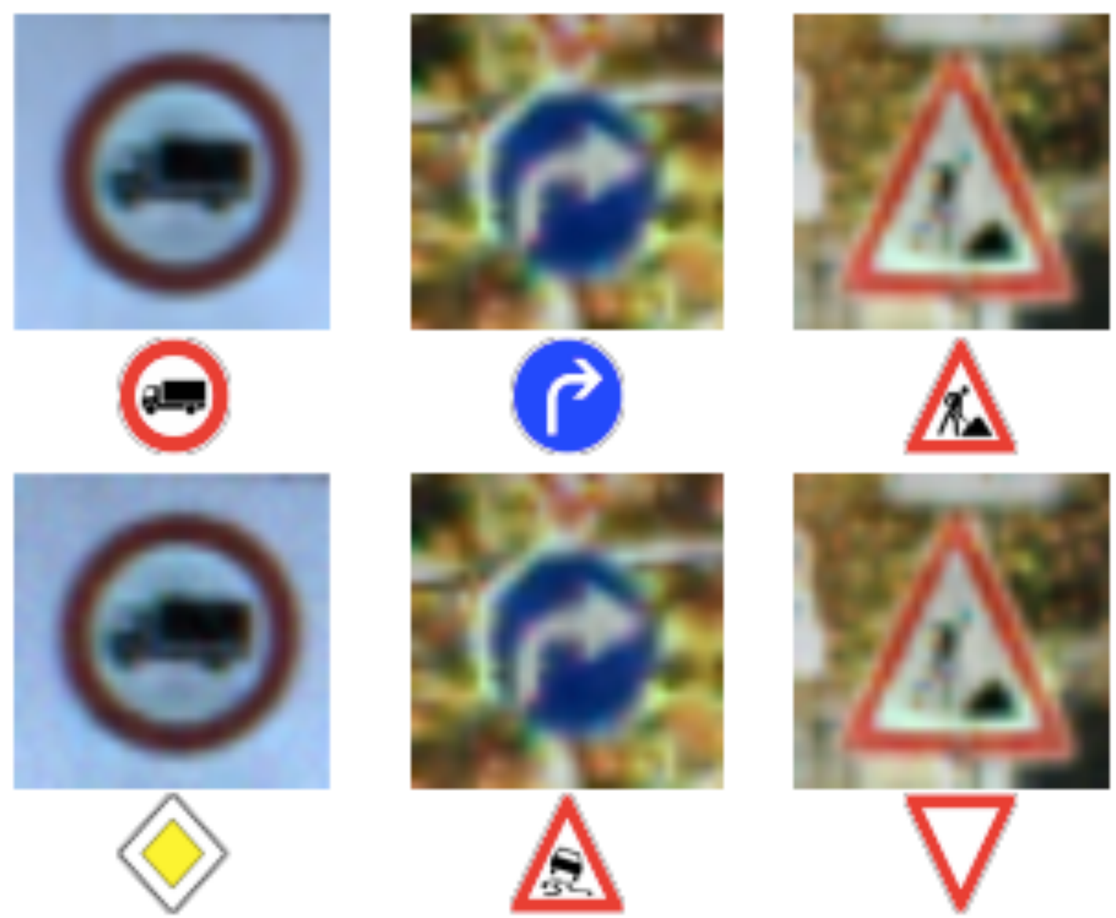


Adversarial testing examples

- First reported in ["Intriguing properties of neural networks", 2013, by Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, Rob Fergus]
- Led to serious concerns for security as, for example,
  - one can create road signs that fools a self-driving car to act in a certain way
- this is serious as
  - there is no reliable defense against adversarial examples
  - adversarial examples transfer to different networks, trained on disjoint subset of training data
  - you do not need the access to the model parameters; you can train your own model and create adversarial examples
  - you only need a black-box access via APIs (MetaMind, Amazon, Google)

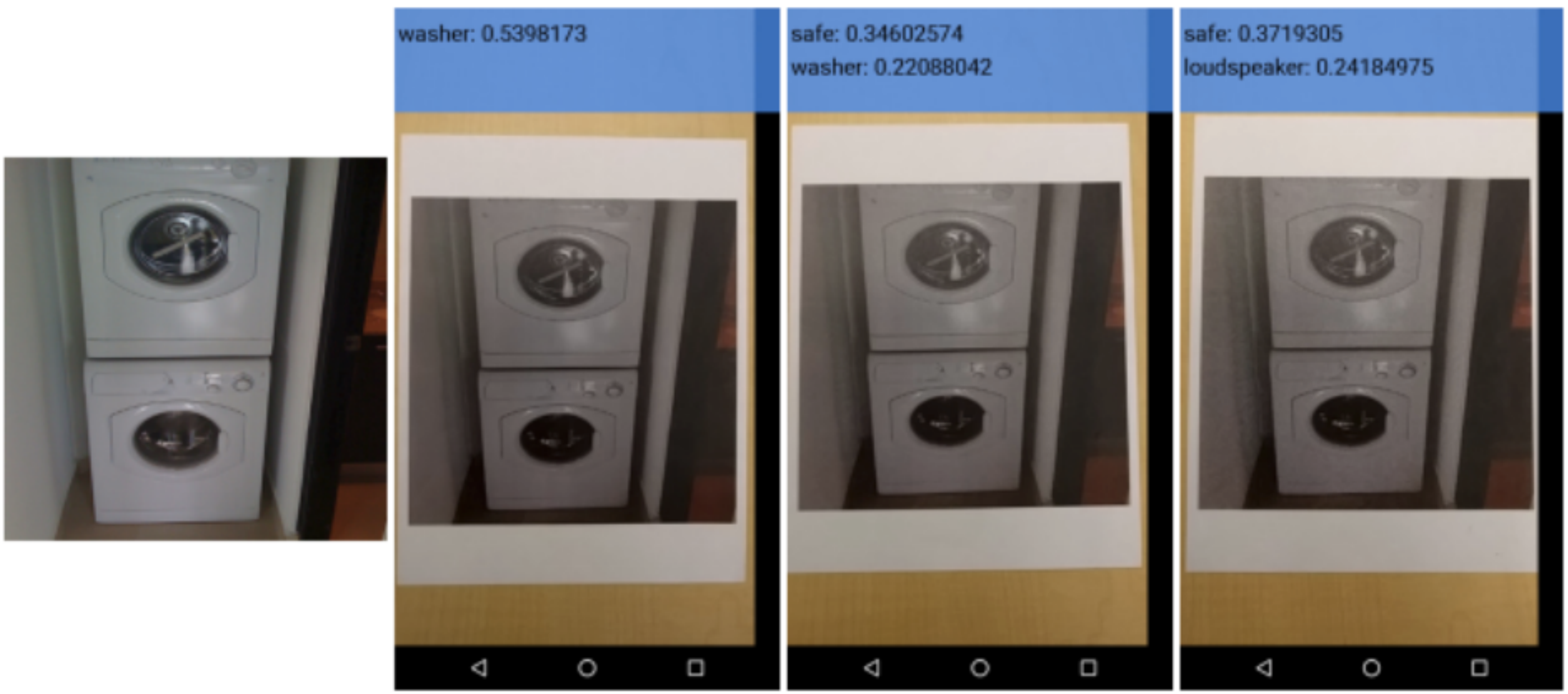
Adversarial testing examples

- ["Practical Black-Box Attacks against Machine Learning", 2016, Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z. Berkay Celik, Ananthram Swami]
- no access to the actual classifier, only treat as a black-box



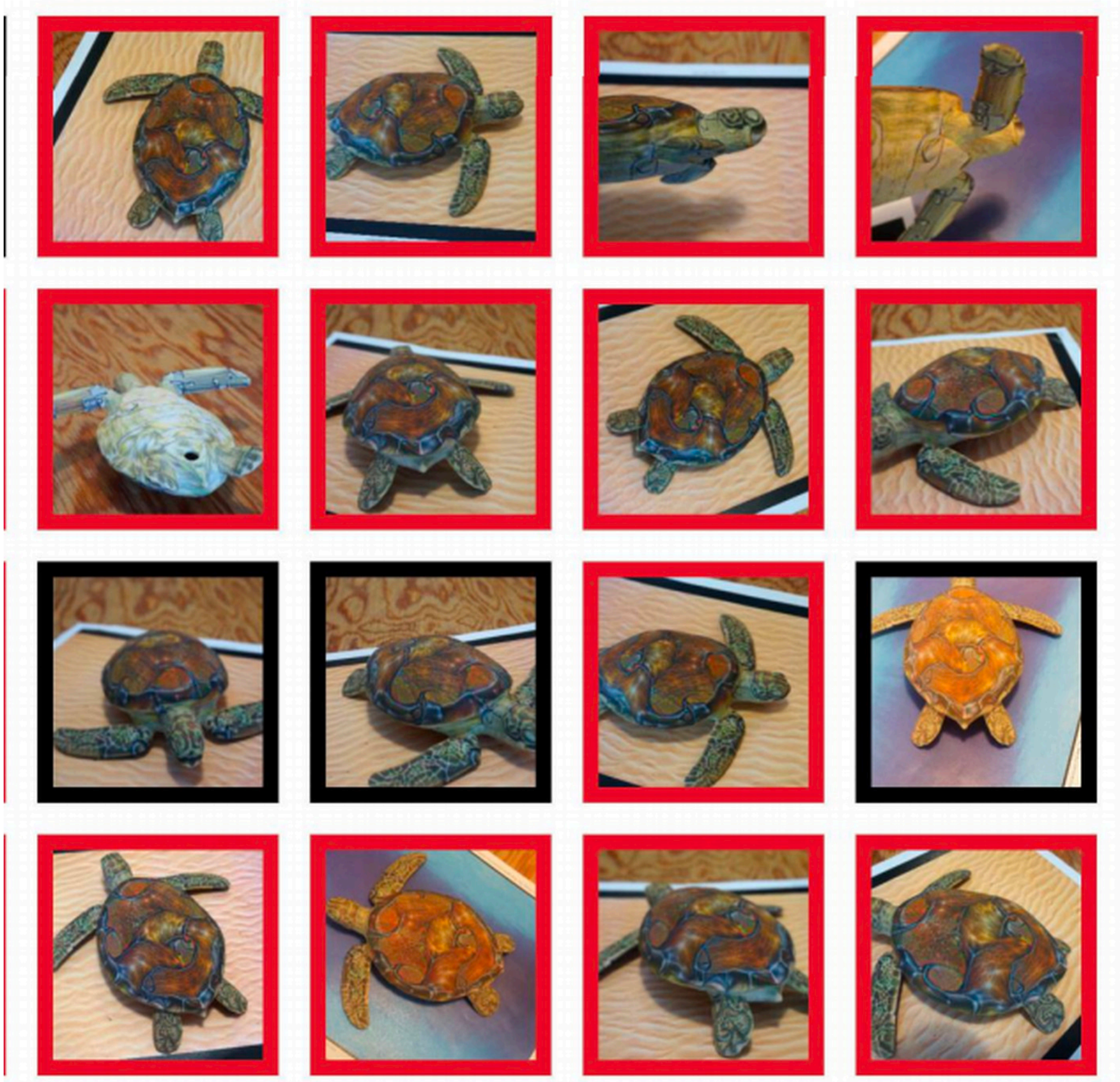
Adversarial testing examples

- ["Adversarial examples in the physical world", 2016, Alexey Kurakin, Ian Goodfellow, Samy Bengio]
- You can fool a classifier by taking picture of a print-out.
- one can potentially print over a stop sign to fool a self-driving car



(a) Image from dataset (b) Clean image (c) Adv. image, $\epsilon = 4$ (d) Adv. image, $\epsilon = 8$

This 3-dimensional turtle is designed to be classified as “rifle”



Defense mechanism to adversarial testing examples

- Brute force: include adversarial testing examples (but with the correct classes) in the training data.

Unlabeled; model guesses it's probably a bird, maybe a plane

Adversarial perturbation intended to change the guess

New guess should match old guess (probably bird, maybe plane)

Defense mechanism to adversarial testing examples

- Defensive distillation:
- Two models are trained
- model 1: trained on the training data in as standard manner
- model 2 (the robust model) : is trained on the same training data, but uses **soft classes** which is the probability provided by the first model
- This creates a model whose surface is smoothed in the directions
an adversary will typically try to exploit, making it difficult for them to discover adversarial input tweaks that lead to incorrect categorization
- [Distilling the Knowledge in a Neural Network, 2015, Geoffrey Hinton, Oriol Vinyals, Jeff Dean]
- original idea came from model compression

- both are vulnerable against high-power adversary

Why are modern classifiers vulnerable

- small margin due to overfitting