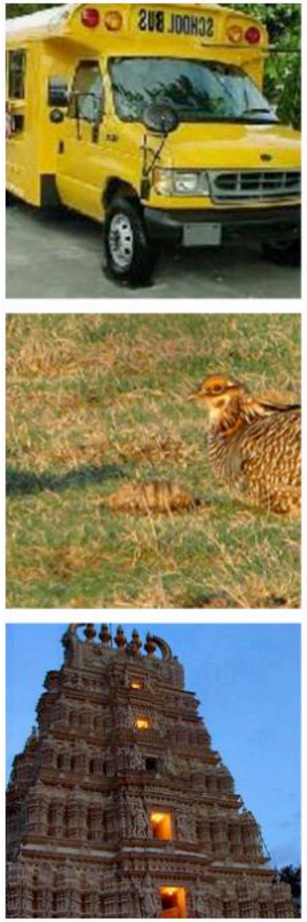
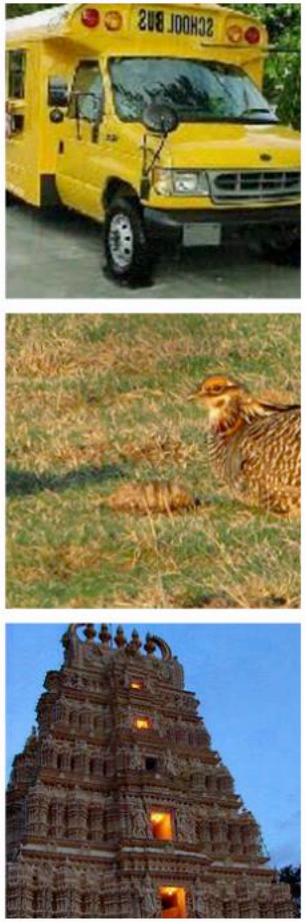
Adversarial examples

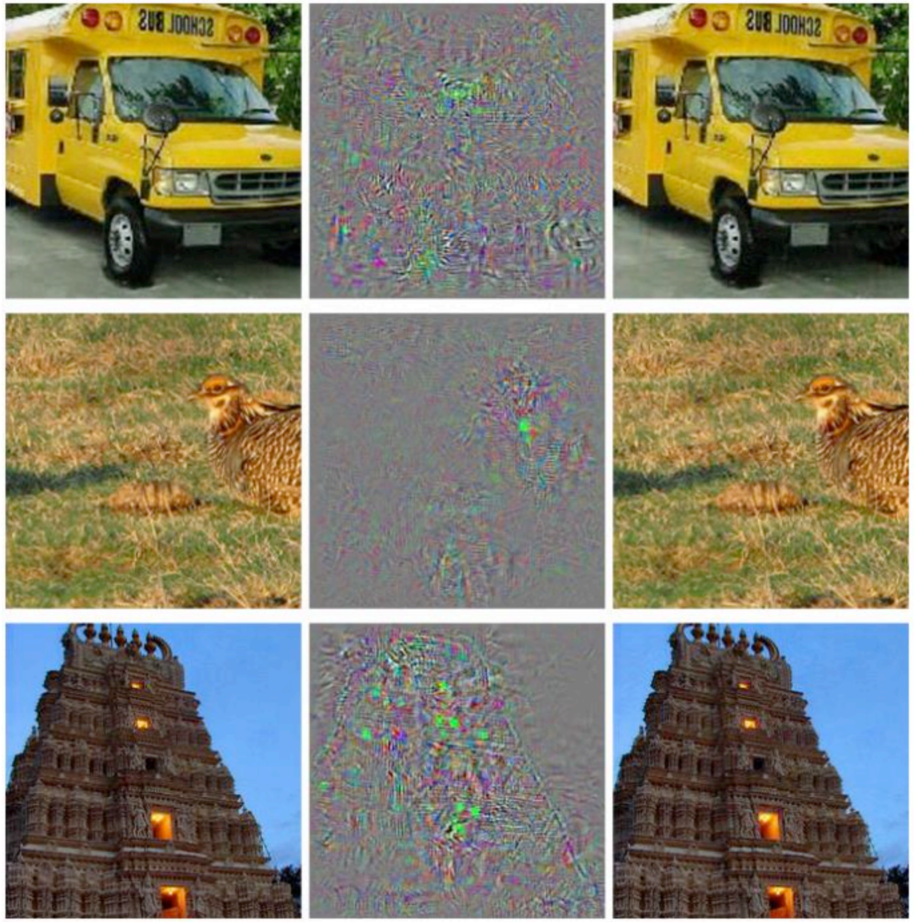


Adversarial examples



Ostrich!
鸵鸟！

Adversarial examples



Noise —> Misclassification ?

Ostrich!

Intriguing properties of neural networks. Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, Rob Fergus. In ICLR, 2014

Why do we care?

- Security
- Safety
- Hint to malfunction?

Adversarial examples

Minimize $\|r\|_2$ **subject to:**

1. $f(x + r) = l$
2. $x + r \in [0, 1]^m$

Adversarial examples for linear classifiers

$$h(x) = w^T x$$

$$\hat{x} = x + \eta$$

$$w^T \hat{x} = w^T x + w^T \eta$$

$$\max_{\eta} w^T \eta$$

$$\text{s.t}$$

$$-\epsilon \leq \eta \leq \epsilon$$

$$\eta = \epsilon \text{sgn}(w)$$

$$w^T \eta = \|w\|_1$$

Adversarial examples for convolutional networks

$$L(\theta, x+\eta, y) \approx L(\theta, x, y) + \eta^T \nabla_x L(\theta, x, y)$$

$$\max_{\eta} L(\theta, x+\eta, y)$$

$$\text{s.t.}$$

$$-\epsilon \leq \eta \leq \epsilon$$

$$\eta = \epsilon \operatorname{sgn}(\nabla_x L(\theta, x, y))$$

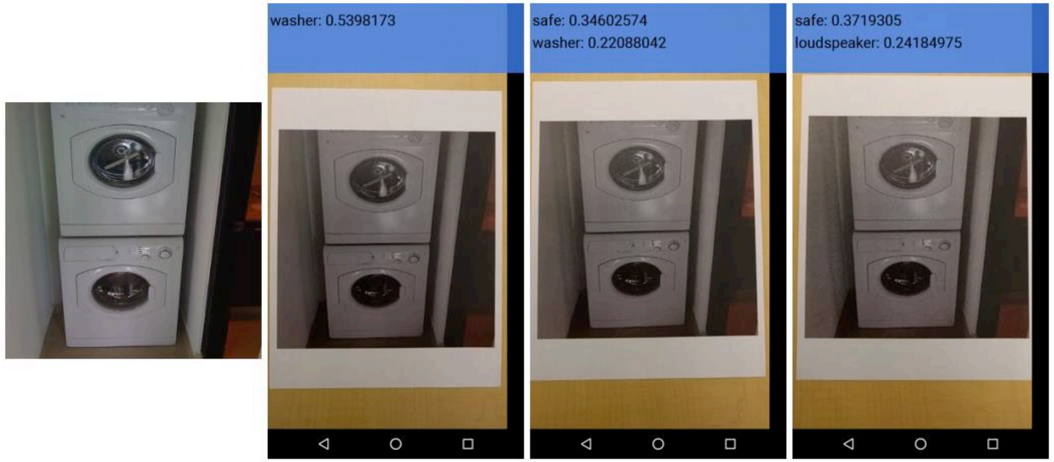
$$L(\theta, x+\eta, y) = L(\theta, x, y) + \epsilon \|\nabla_x L(\theta, x, y)\|_1$$

Adversarial examples for convolutional networks

- Convolutional networks w/ RELU are differentiable almost everywhere
- Are *linear* almost everywhere
- Slope for a given x = gradient at x
- Can use gradient to generate an adversarial example

Explaining and Harnessing Adversarial Examples. Ian Goodfellow, Jonathon Shlens, Christian Szegedy. In *ICLR 2015.*

More fun with adversarial examples



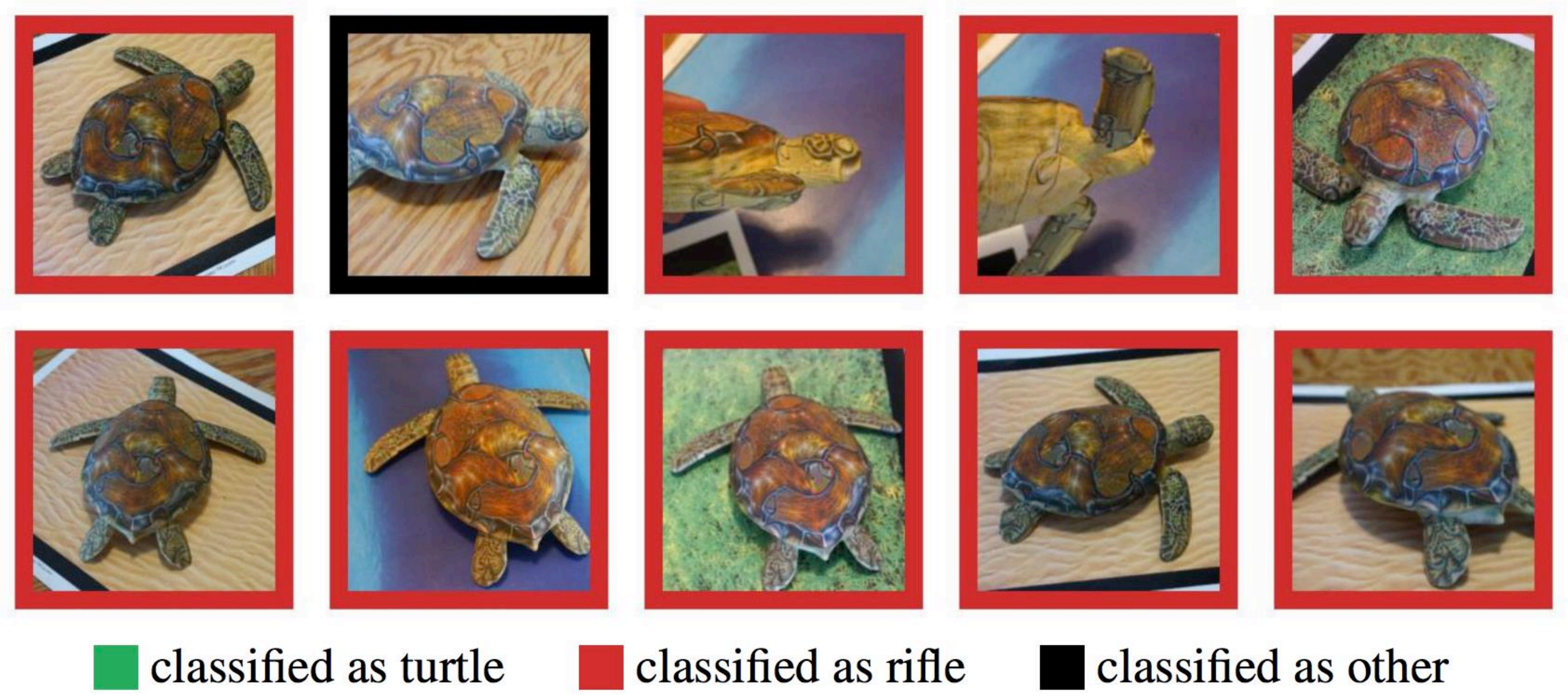
(a) Image from dataset (b) Clean image (c) Adv. image, $\epsilon = 4$ (d) Adv. image, $\epsilon = 8$

- Transferable across models
- Resilient to printing and photographing

Adversarial method	Photos				Source images			
	Clean images		Adv. images		Clean images		Adv. images	
	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5
fast $\epsilon = 16$	79.8%	91.9%	36.4%	67.7%	85.3%	94.1%	36.3%	58.8%
fast $\epsilon = 8$	70.6%	93.1%	49.0%	73.5%	77.5%	97.1%	30.4%	57.8%
fast $\epsilon = 4$	72.5%	90.2%	52.9%	79.4%	77.5%	94.1%	33.3%	51.0%
fast $\epsilon = 2$	65.7%	85.9%	54.5%	78.8%	71.6%	93.1%	35.3%	53.9%

Adversarial examples in the physical world. Alexey Kurakin, Ian Goodfellow, Samy Bengio. ICLR Workshop (2017)

Adversarial turtle



Synthesizing robust adversarial examples. Anish Athalye, Logan Engstrom , Andrew Ilyas , Kevin Kwok.

Adversarial turtle

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}'} \mathbb{E}_{\mathbf{t}\sim\mathbf{T}}[-\log \mathbf{P}(\mathbf{y}|\mathbf{t}(\mathbf{x}')) + \lambda||\mathbf{LAB}(\mathbf{t}(\mathbf{x}) - \mathbf{t}(\mathbf{x}'))||_{\mathbf{2}}^{\mathbf{2}}]$$

Resilience to adversaries

$$\eta = \epsilon \operatorname{sgn}(\nabla_x L(\theta, x, y))$$

$$\alpha L(\theta, x, y) + (1 - \alpha) L(\theta, x + \epsilon \operatorname{sgn}(\nabla_x L(\theta, x, y))$$

89.4% → 17.9%

Kinds of adversarial perturbations

- “White-box” vs “black-box”
  - Does adversary have access to the model?
- “Untargeted” vs “Targeted”
  - Should the new output be incorrect in a particular way?

Integrity attack, functionality attack, privacy attack

Training-time attack, inference-time attack