

111111

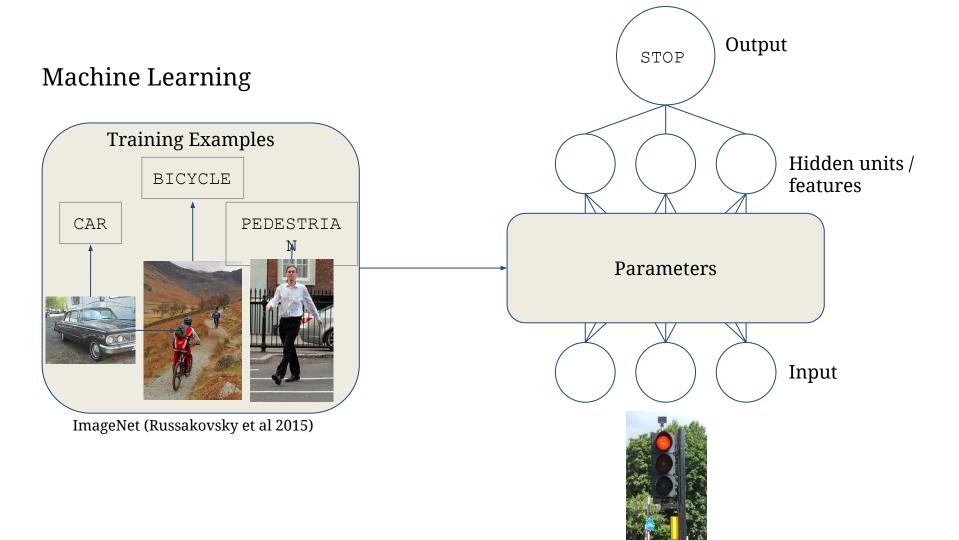
Physical Adversarial Examples

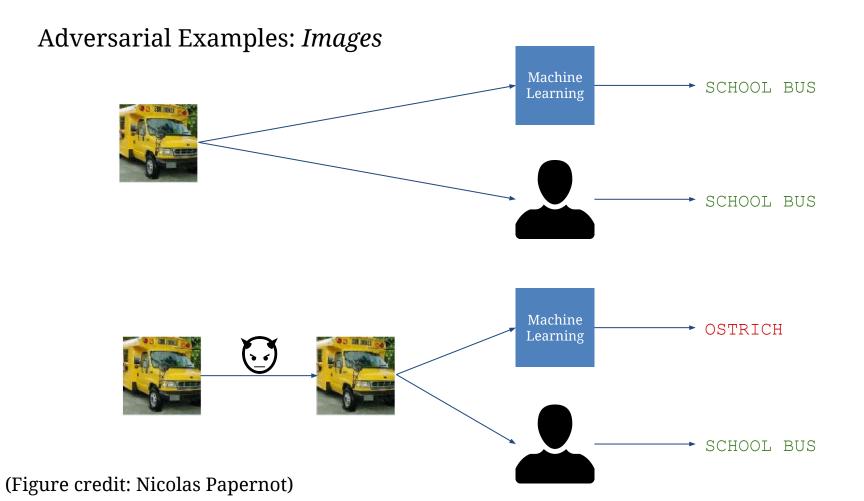
Alex Kurakin Ian Goodfellow Google OpenAl

1024 geekpwn 1024 geekpwn 1024 1024 geekpwn 1024 geekpwn 1024 1024 geekpwn 1024 geekpwn 1024

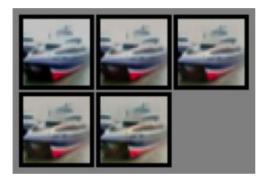
111111

111111

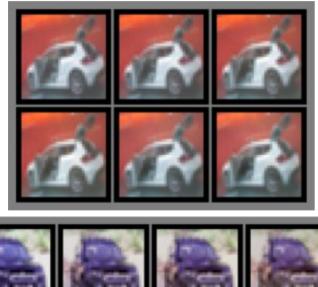




Turning Objects into "Airplanes"











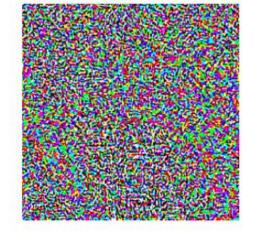


Fast Gradient Sign Method (FGSM)



 \boldsymbol{x}

"panda" 57.7% confidence $+.007 \times$

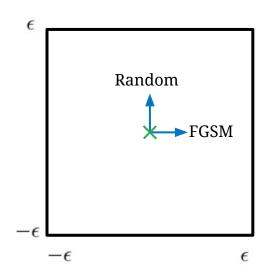


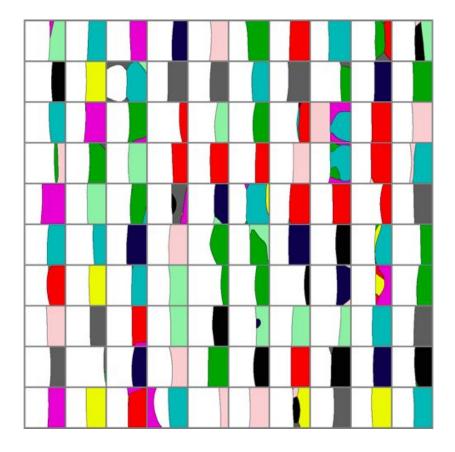
$$\operatorname{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$$

"nematode" 8.2% confidence

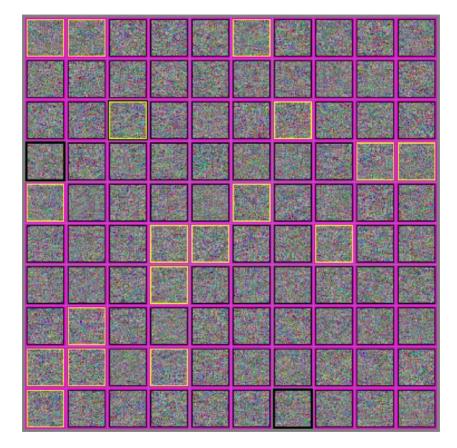
 $x + \epsilon sign(\nabla_x J(\theta, x, y))$ "gibbon" 99.3 % confidence

Maps of Adversarial Examples



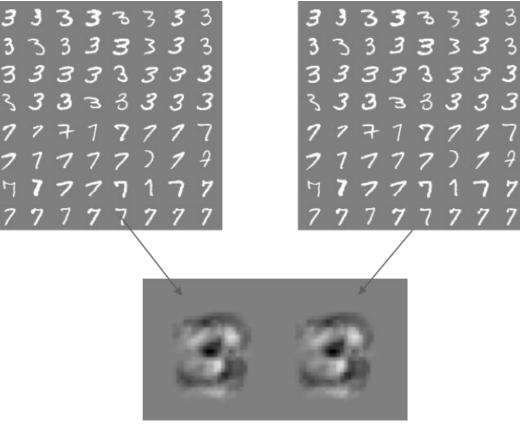


Almost all inputs are misclassified



Generalization across training sets

R



7

7

Cross-Technique Transferability

· · · · · · · · · · · · · · · · · · ·		5.0 C			7. 74.17 ···· ···	
uique NND	- 38.27	23.02	64.32	79.31	8.36	20.72 -
Learning Technique S M	6.31	91.64	91.43	87.42	11.29	44.14 -
	- 2.51	36.56	100.0	80.03	5.19	15.67 -
Source Machine T	- 0.82	12.22	8.85	89.29	3.31	5.11 -
INOS KNN	- 11.75	42.89	82.16	82.95	41.65	31.92 -
	DNN	LR	SVM	DT	kNN	Ens.
	Target Machine Learning Technique					

(Papernot et al 2016)

Transferability attack

Target model with unknown weights, Substitute model Train your machine learning mimicking target own model algorithm, training model with known, set; maybe nondifferentiable function differentiable Adversarial crafting Deploy adversarial against substitute examples against the Adversarial target; transferability examples property results in them succeeding

Results on Real-World Remote Systems

All remote classifiers are trained on the MNIST dataset (10 classes, 60,000 training samples)

Remote Platform	ML technique	Number of queries	Adversarial examples misclassified (after querying)
MetaMind	MetaMind Deep Learning		84.24%
amazon webservices™	Linear Regression	800	96.19%
Google Cloud Platform	Unknown	2,000	97.72%

(Papernot et al 2016)

Adversarial examples in the physical world?

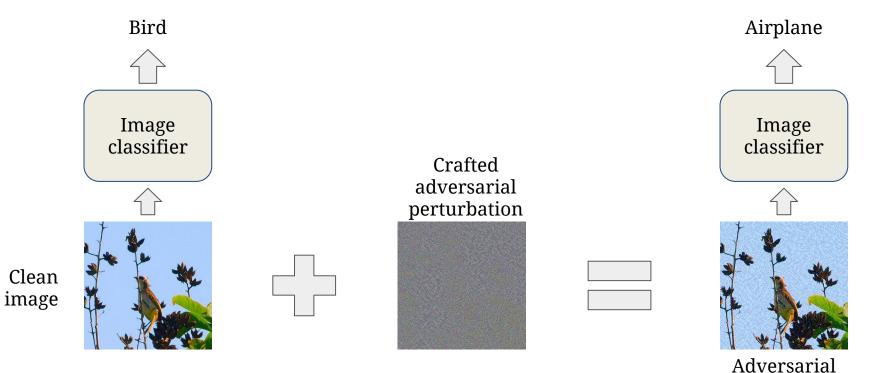
- Question: Can we build adversarial examples in the physical world?
- Let's try the following:
 - Generate and print picture of adversarial example
 - Take a photo of this picture (with cellphone camera)
 - Crop+warp picture from the photo to make it 299x299 input to Imagenet inception
 - Classify this image
- Would the adversarial image remain misclassified after this transformation?
- If we succeed with "photo" then we potentially can alter real-world objects to mislead deep-net classifiers

Adversarial examples in the physical world?

- Question: Can we build adversarial examples in the physical world?
- Let's try the following:
 - Generate and print picture of adversarial example
 - Take a photo of this picture (with cellphone camera)
 - Crop+warp picture from the photo to make it 299x299 input to Imagenet inception
 - Classify this image
- Would the adversarial image remain misclassified after this transformation?
- If we succeed with "photo" then we potentially can alter real-world objects to mislead deep-net classifiers

Answer: IT'S POSSIBLE

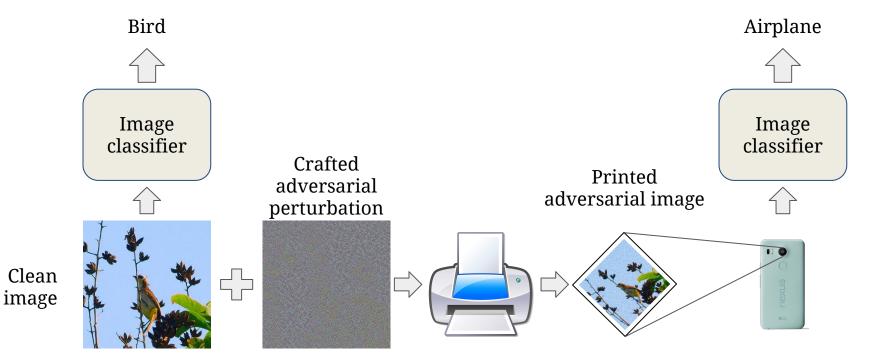
Digital adversarial examples



image

[Goodfellow, Shlens & Szegedy, ICLR2015]

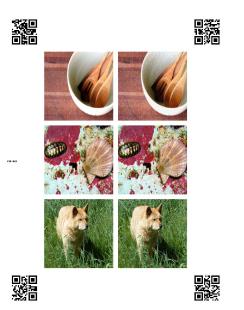
Adversarial examples in the physical world



[Kurakin & Goodfellow & Bengio, arxiv.org/abs/1607.02533]

Our experiment

1. Print pairs of normal and adversarial images



2. Take picture



3. Auto crop and classify



Up to 87% of images could remain misclassified!

Live demo



Library





Washer





Washer



Don't panic! It's not end of the ML world!

- Our experiment is a proof-of-concept set up:
 - We had full access to the model
 - 87% adversarial images rate is for only one method, which could be resisted by adversarial training. For other methods it's much lower.
 - In many cases "adversarial" image is not so harmful: one breed of dog confused with another
- In practice:
 - Attacker doesn't have access to model
 - You might be able to use adversarial training to defend model against some attacks
 - For other attacks, "adversarial examples in the real worlds" won't work that well
 - It's REALLY hard to fool your model to predict specific class



GeekPwn

1024 🏠

HANKS

024 geekpwn 1024 geekpwn 1

1024 geekpwn 1024 geekpwn 1024 1024 geekpwn 1024 geekpwn 1024 1024 geekpwn 1024 geekpwn 1024