Susceptibility to Adversarial Attacks

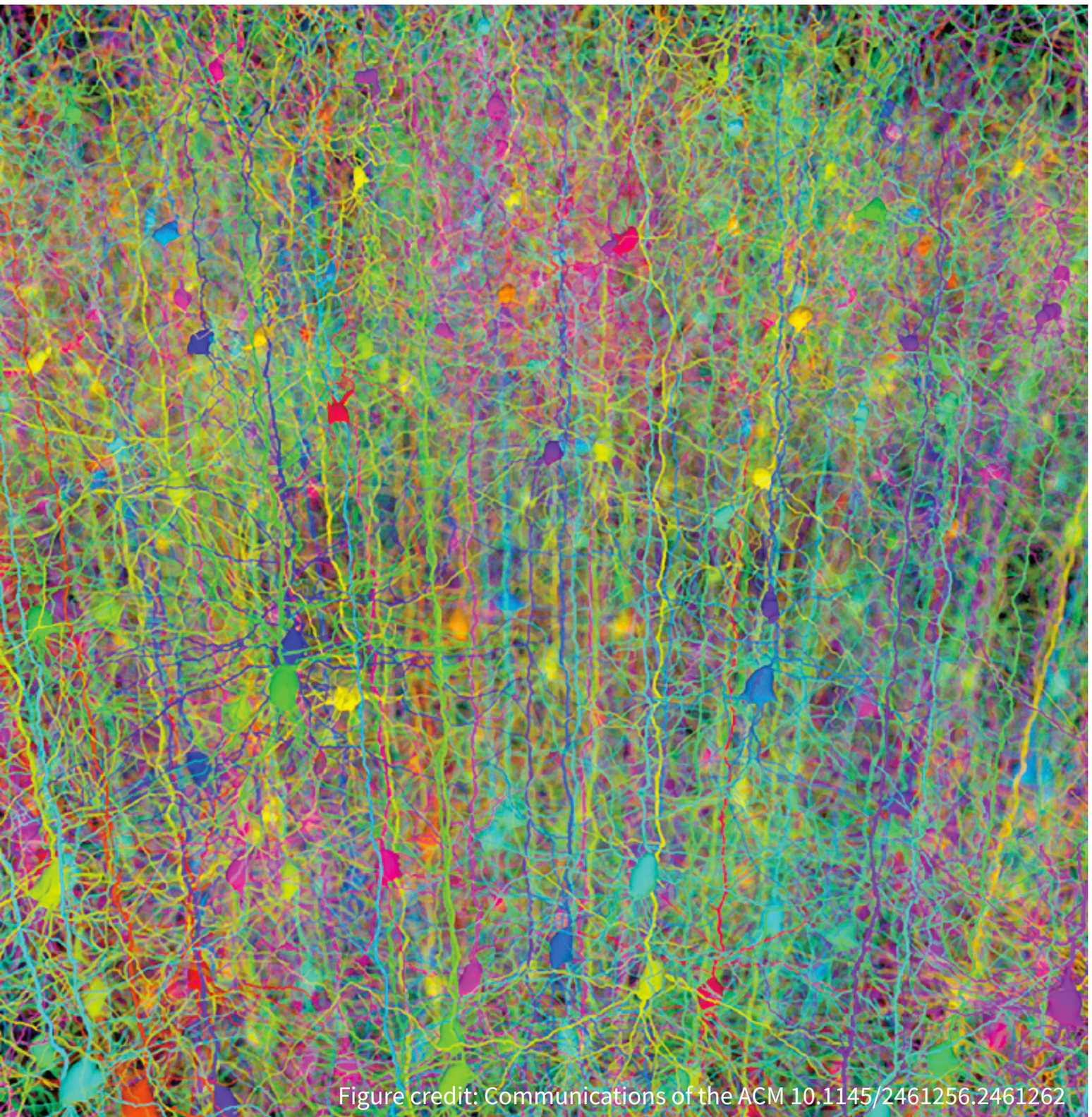


Figure credit: Communications of the ACM 10.1145/2461256.2461262

GRAHAM TAYLOR

VECTOR INSTITUTE

SCHOOL OF ENGINEERING
UNIVERSITY OF GUELPH

CANADIAN INSTITUTE
FOR ADVANCED RESEARCH

UNIVERSITY of GUELPH

CHANGING LIVES
IMPROVING LIFE

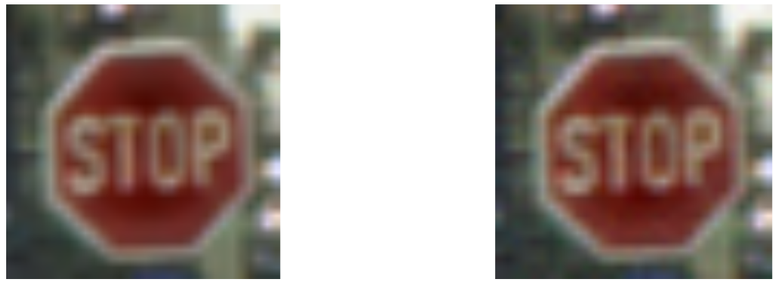
CIFAR
CANADIAN INSTITUTE FOR ADVANCED RESEARCH

Discovery of Adversarial Examples

- Szegedy et al. (2014) made an intriguing discovery: several machine learning models, including state-of-the-art neural networks, are vulnerable to *adversarial examples*

- Wide variety of models with different architectures trained on different subsets of the training data misclassify the same adversarial example

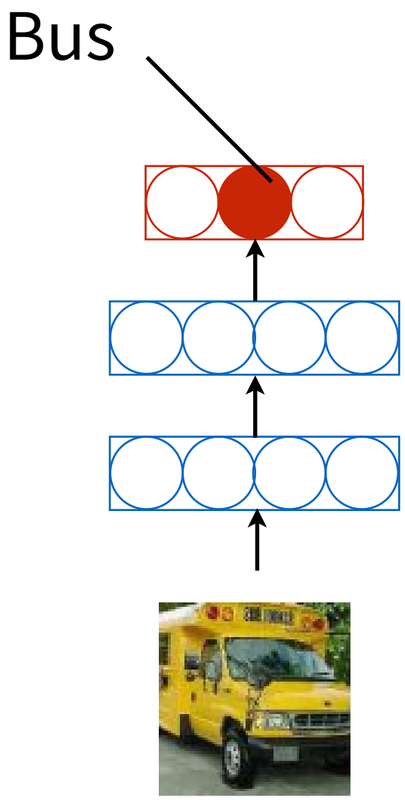
- Expose fundamental blind spots in training algorithms?



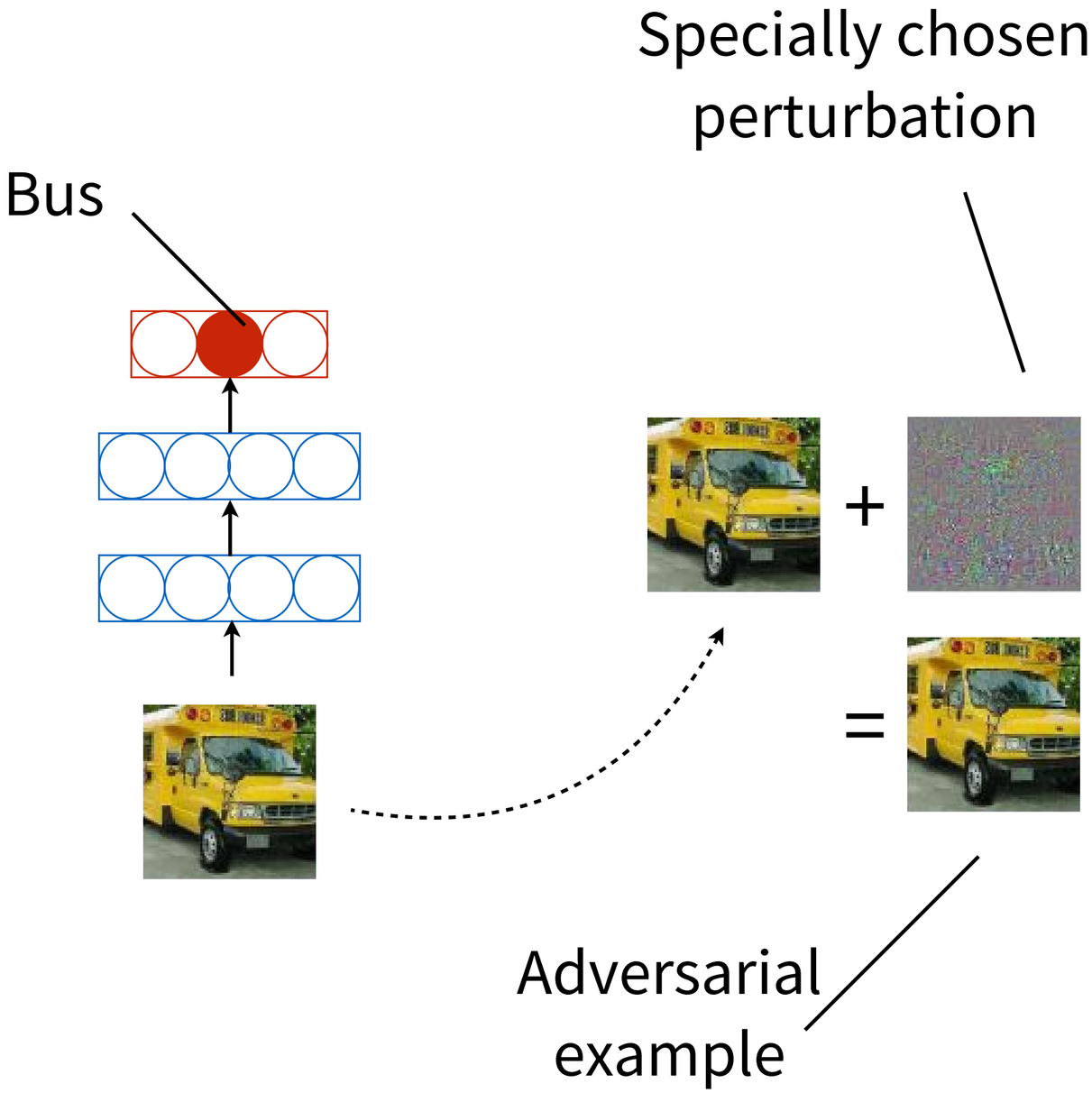
via Papernot et al. 2016

What is an Adversarial Example?

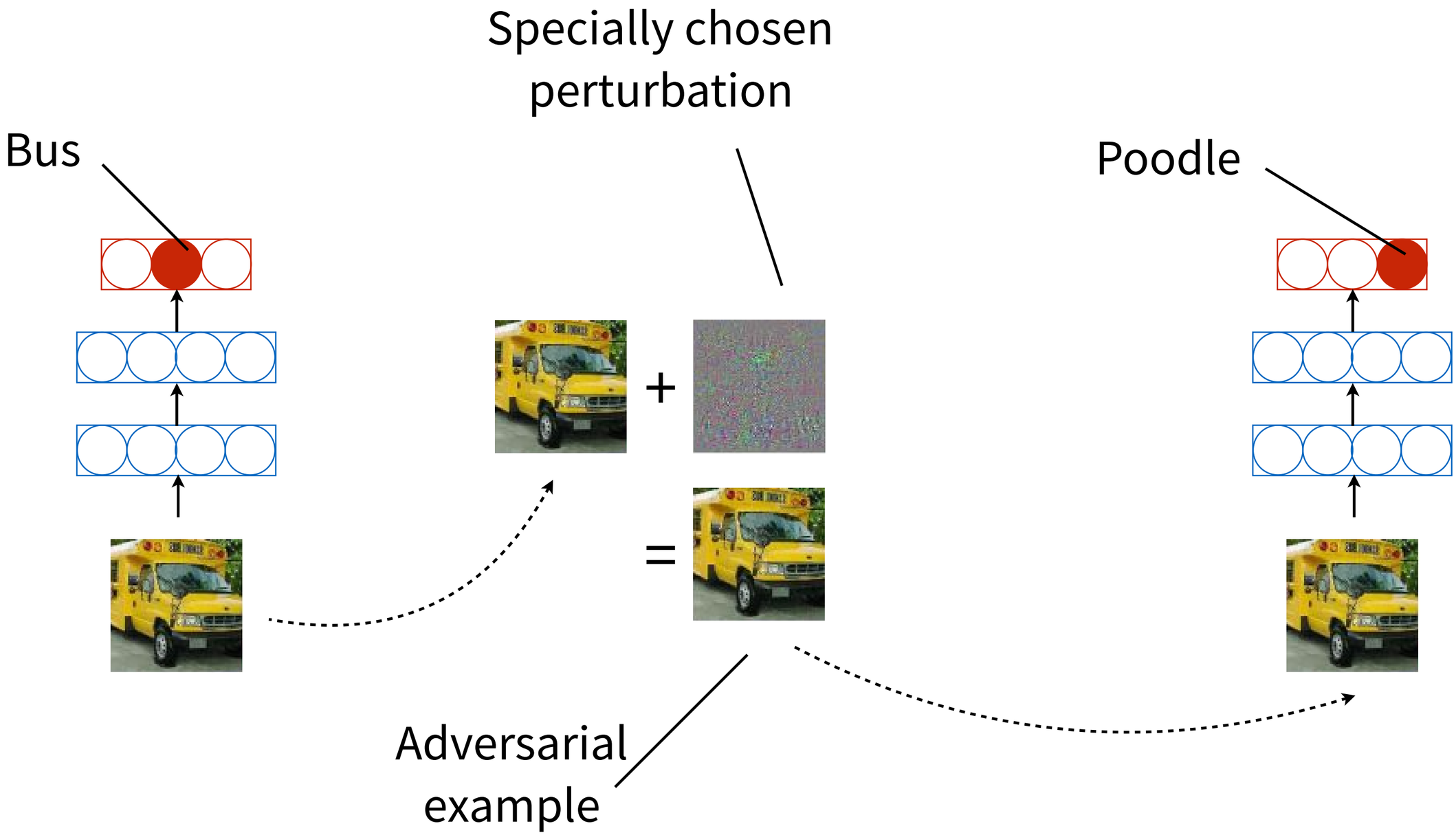
What is an Adversarial Example?



What is an Adversarial Example?



What is an Adversarial Example?



Generating Adversarial Examples

$f : \mathbb{R}^d \rightarrow \{1 \ldots k\}$

Classifier mapping image pixels to label set

$J : \mathbb{R}^d \times \{1 \ldots k\} \rightarrow \mathbb{R}^+$

Associated continuous loss function

For a given $\mathbf{x} \in \mathbb{R}^d$ image and target label $y \in \{1 \ldots k\}$ aim to solve the following box-constrained optimization:

$$\text{Minimize } ||\boldsymbol{r}||_2 \text{ subject to:}$$
$$f(\boldsymbol{x} + \boldsymbol{r}) = y$$
$$\boldsymbol{x} + \boldsymbol{r} \in [0, 1]^d$$

Fast Gradient Sign Method

- Goodfellow et al. (2014) showed it is possible to generate an adversarial example without iterative optimization

$$\mathbf{x}^{\mathrm{adv}} = \mathbf{x} + \epsilon \mathrm{sign}\left(\nabla_{\mathbf{x}} J\left(\mathbf{x}, y\right)\right)$$

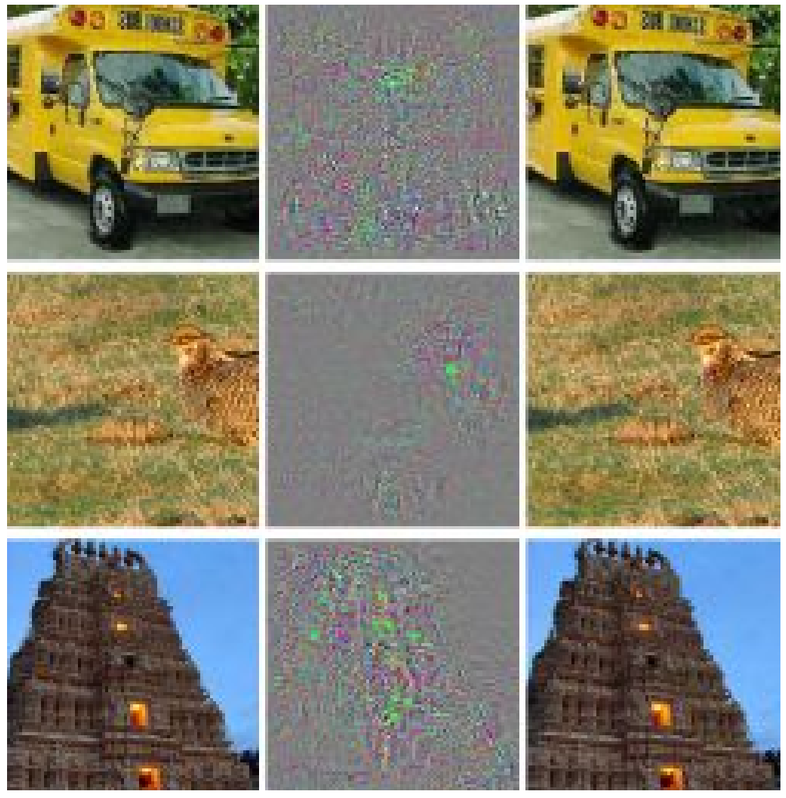
- Cost: one call to backprop

More Adversarial Examples

Correctly predicted sample

Difference

Incorrectly predicted sample



Intriguing Properties

- For all the networks studied (MNIST, QuocNet, AlexNet) for each sample, always manage to generate very close, visually instinguishable AEs

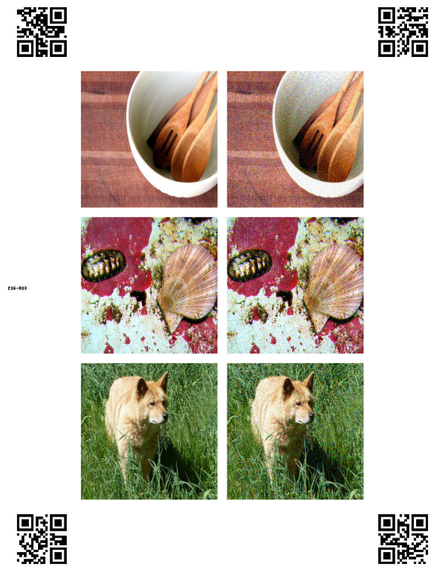
- Cross *model* generalization

- Cross *training set* generalization

Suggests that AEs are somewhat universal and
not just the result of overfitting to a model or training set

Adversarial Examples in the Physical World

- Attacks originally assumed an attacker could feed data directly to the model
- This is not the case for machines operating in the physical world using cameras and other sensors

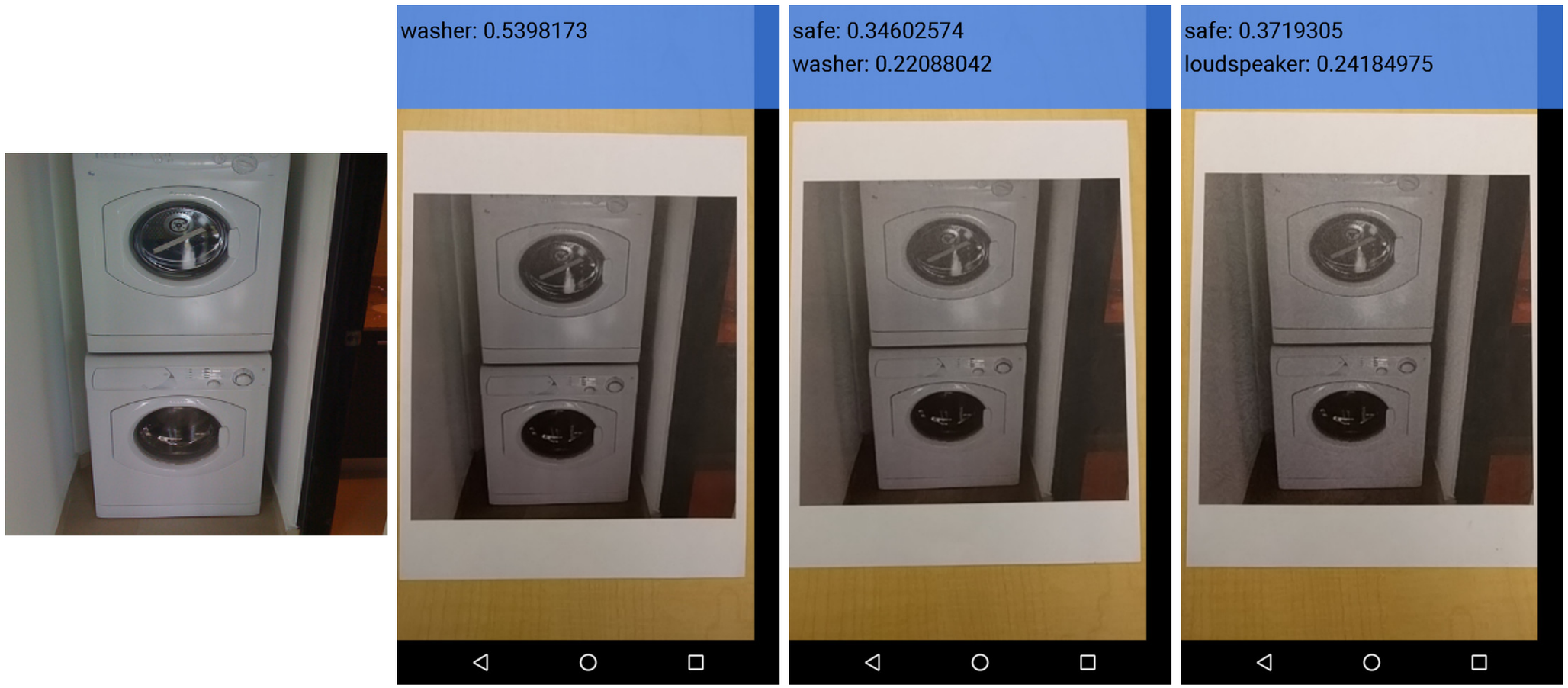


Printout

(Digital) photo of printout

Cropped image

Adversarial Examples in the Physical World



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

Attack is done without access to the model

via Kurakin et al. 2017

Black Box vs. White Box Attacks

- White box: Previous results assume attacker has full knowledge of the model architecture and parameter values

- Black box: more realistic model of many security threats

- Papernot et al. (2016) carry out an attack against a DNN hosted by MetaMind and find that their DNN misclassifies 84.24% of the adversarial examples

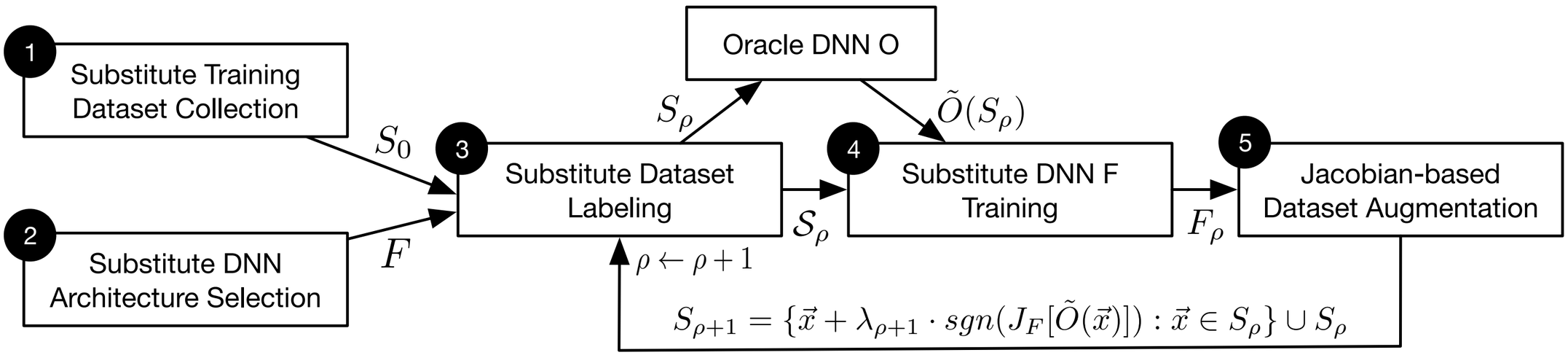
  - Port to Amazon (96.19%) and Google (88.94%) APIs

Practical Black Block Attacks

Black box attacks applicable to many remote systems should have the following key properties:

1. Capabilities required are limited to observing class output labels
2. Number of labels queried is limited
3. Approach applies and scales to different DNNs

Black Box Attack with Substitute DNN



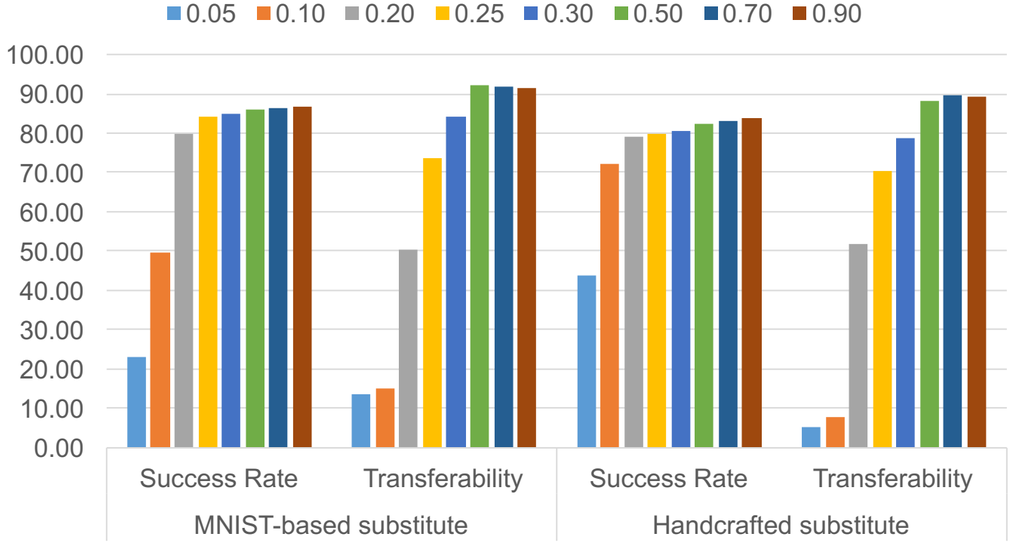
$\tilde{O}(\vec{x})$	Oracle labeling (under attack)
S_0	Small, initial substitute set
F	Substitute architecture
ρ	Index for iterations
J_F	Substitute's Jacobian
λ	Parameter for augmentation

via Papernot et al. 2016

Results (MNIST)

Substitute Epoch	Initial Substitute Training Set from MNIST test set	Handcrafted digits
0	24.86%	18.70%
1	41.37%	19.89%
2	65.38%	29.79%
3	74.86%	36.87%
4	80.36%	40.64%
5	79.18%	56.95%
6	81.20%	67.00%

Accuracy of substitute model grows as synthetic dataset expands

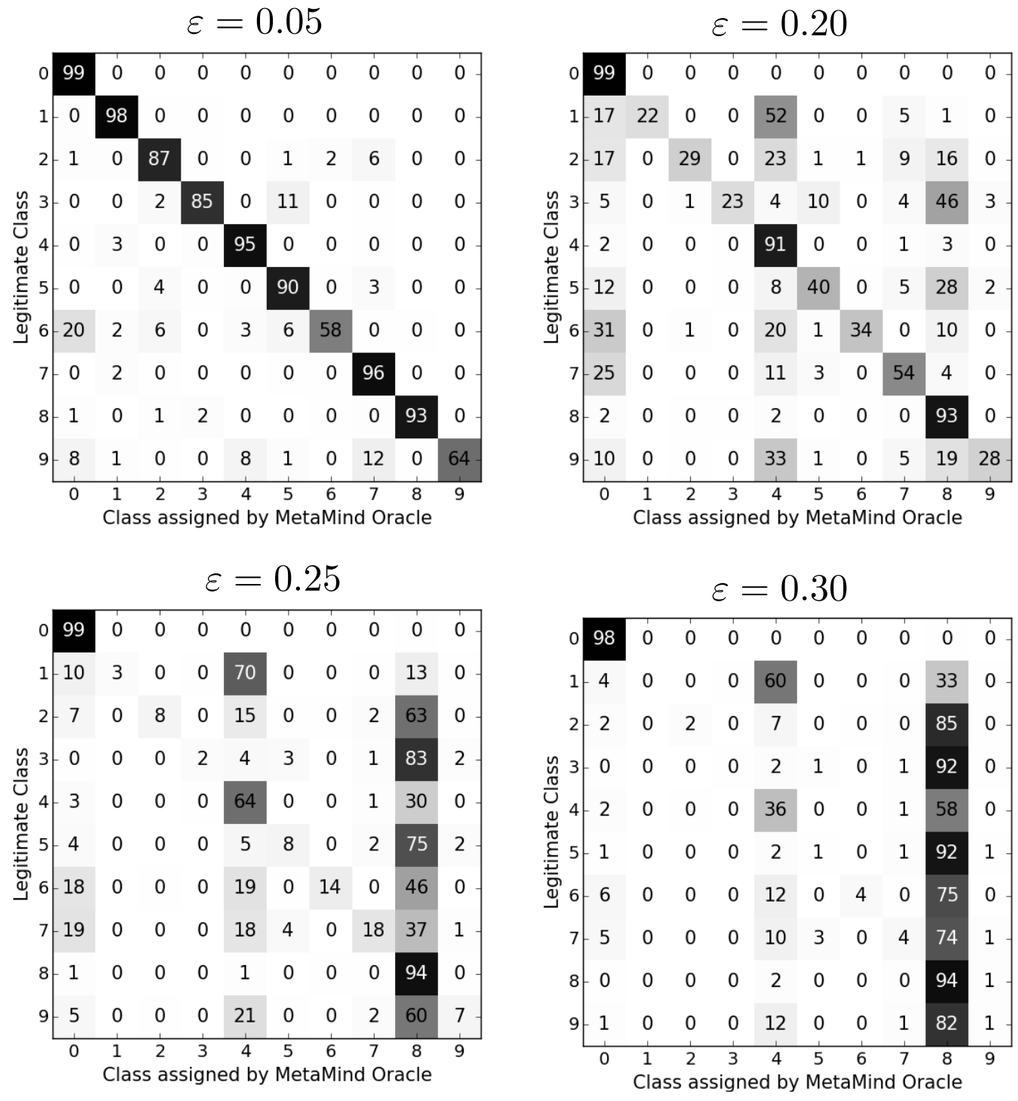


For different levels of input perturbation
Shows success of adversarial attack:
1) on substitute; 2) transferred to oracle

via Papernot et al. 2016

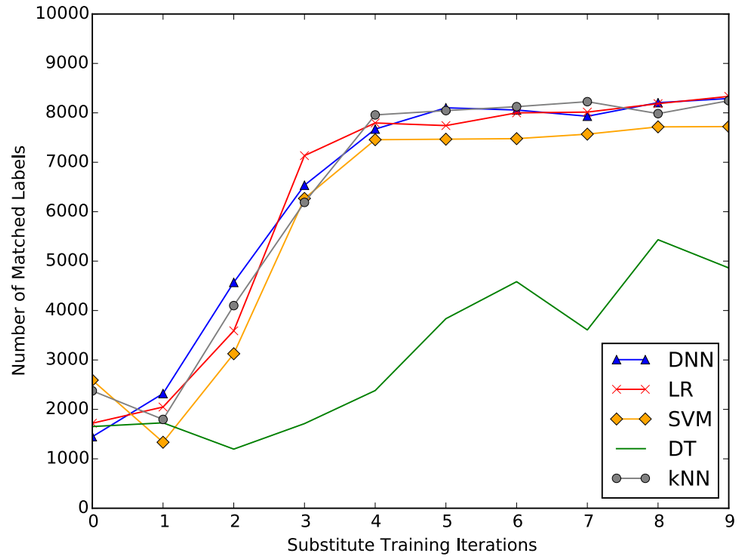
MetaMind Attack

- Attack is effective in severely damaging the output integrity of the MetaMind oracle
- Confusion matrices converge to most samples being classified as 4s and 8s as perturbation increases

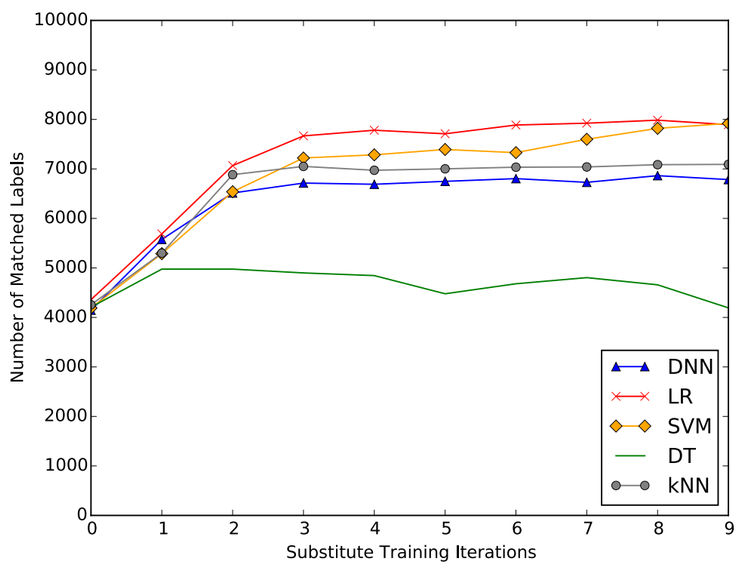


via Papernot et al. 2016

Generalization to other ML Models



(a) DNN substitutes



(b) LR substitutes

		Amazon		Google	
Epochs	Queries	DNN	LR	DNN	LR
$\rho = 3$	800	87.44	96.19	84.50	88.94
$\rho = 6$	6,400	**96.78**	**96.43**	**97.17**	92.05
$\rho = 6^*$	2,000	95.68	95.83	91.57	**97.72**

Technique is effective using Google and Amazon APIs
with DNN or Logistic Regression substitute
for MNIST-based oracle
Last row uses some proposed model enhancements

DNN or Logistic Regression substitutes
can be used to fool other ML models
(less effective for DT due to non-continuity?)

Defense Strategies?

Defense strategies may be reactive (detecting adversarial examples) and proactive (making the model more robust). Proactive includes:

- Make the input higher-dimensional
- Make the model more complex
- Gradient masking
  - e.g. use nearest neighbour instead of DNN
- Adversarial training
- Defensive distillation (Papernot et al. 2015, Carlini and Wagner 2017)

via wsj.com