Deep Learning

Lecture 9: Adversarial attacks and defense

Prof. Gilles Louppe
g.louppe@uliege.be

LIÈGE université

Today

Can you fool neural networks?

- Adversarial attacks
- Adversarial defenses

We have seen that deep networks achieve super-human performance on a large variety of tasks.

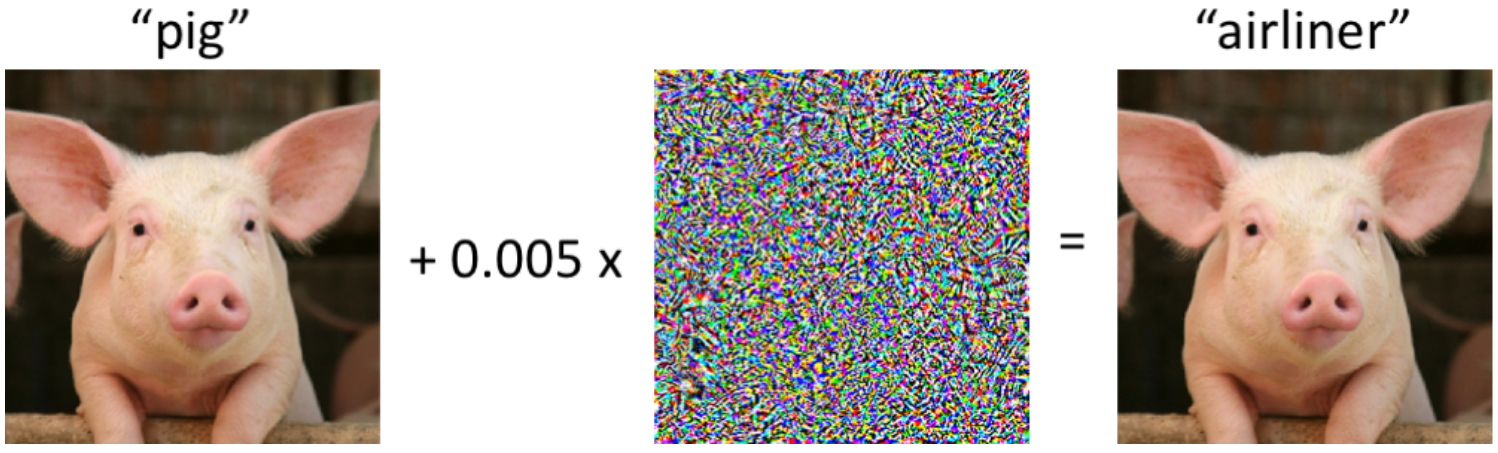
Soon enough, it seems like:

- neural networks will replace your doctor;
- neural networks will drive your car;
- neural networks will compose the music you listen to.

But is that the end of the story?

Adversarial attacks

Adversarial examples

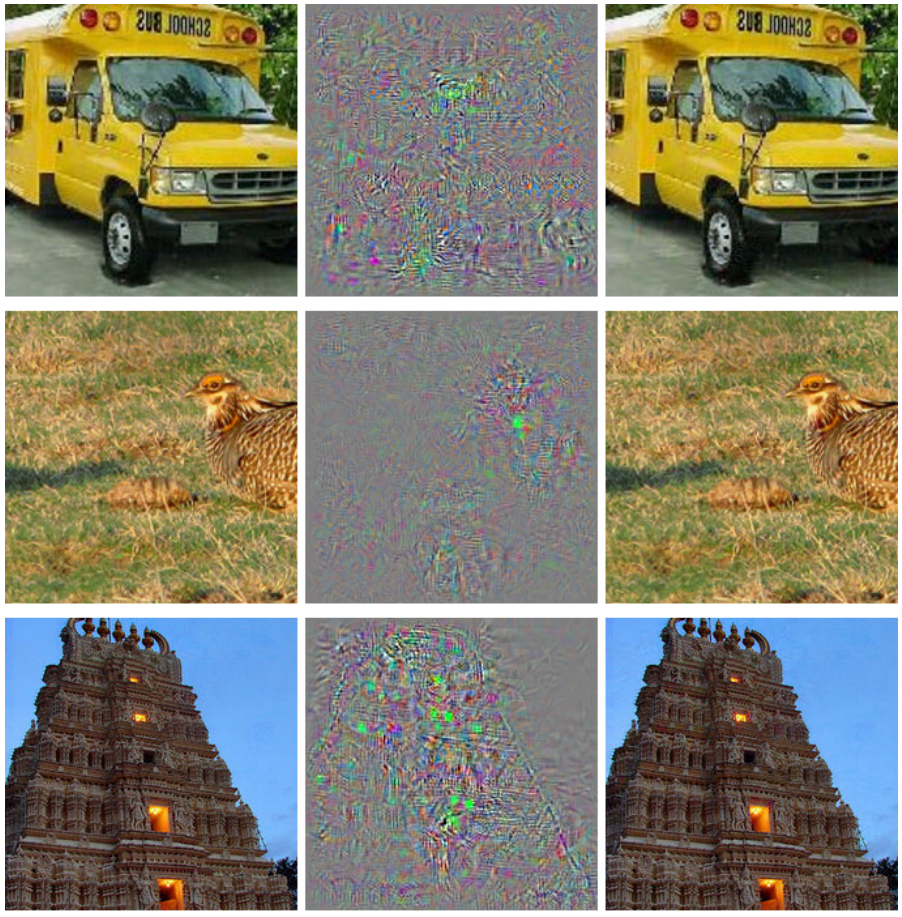


Intriguing properties of neural networks

"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."

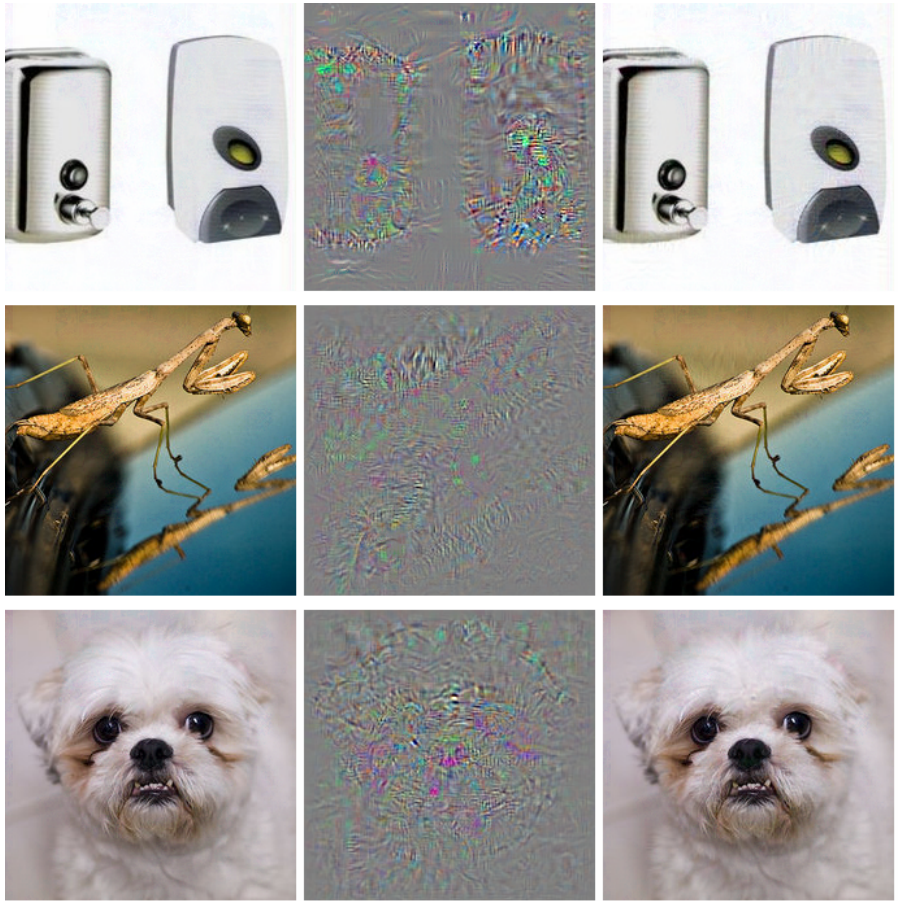
The existence of the adversarial negatives appears to be in contradiction with the network's ability to achieve high generalization performance. Indeed, if the network can generalize well, how can it be confused by these adversarial negatives, which are indistinguishable from the regular examples?"

(Szegedy et al, 2013)



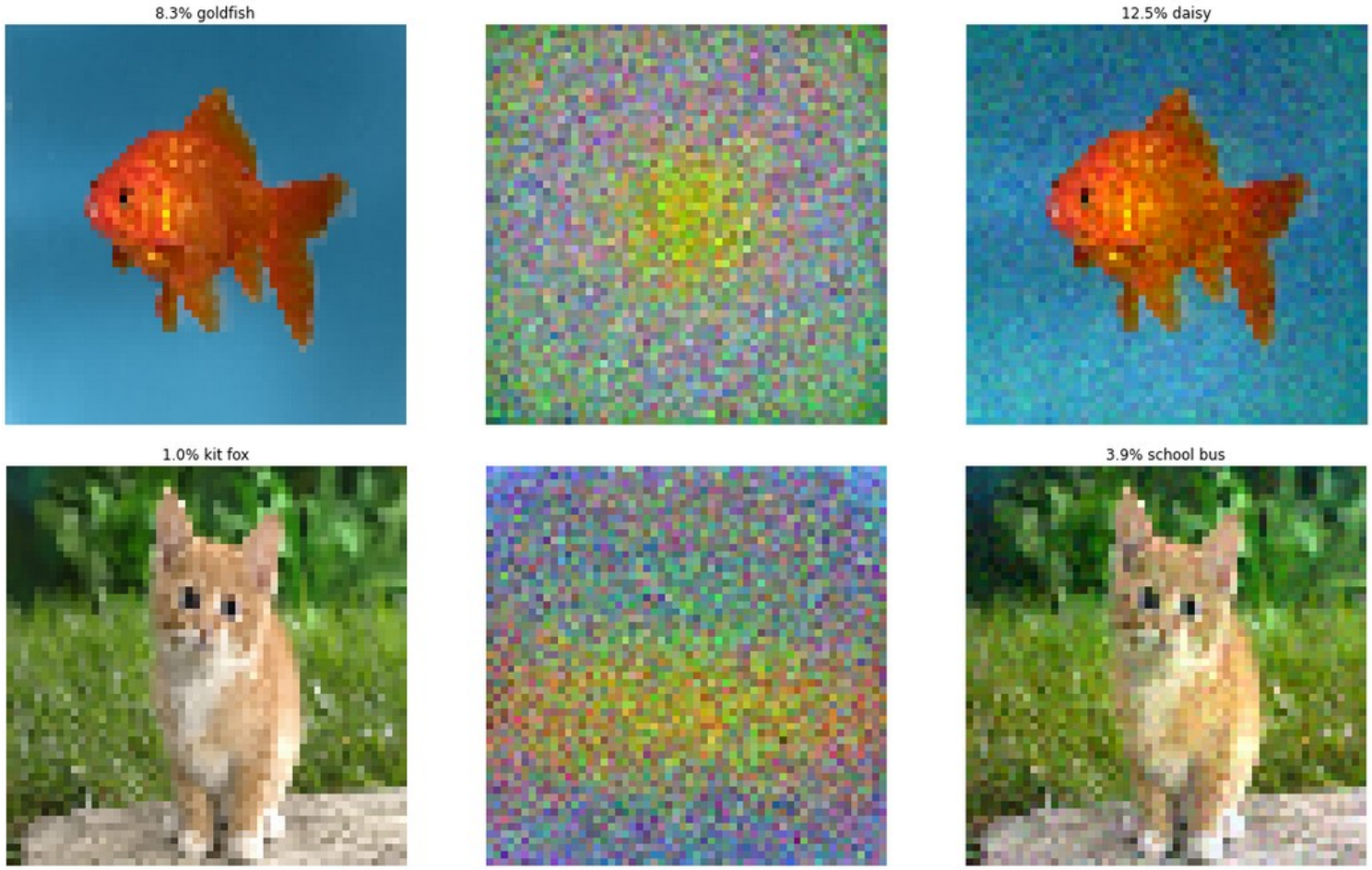
(Left) Original images. (Middle) Adversarial noise. (Right) Modified images.
All are classified as 'Ostrich'.

Credits: Szegedy et al, Intriguing properties of neural networks, 2013.



Credits: Szegedy et al, Intriguing properties of neural networks, 2013.

Fooling a logistic regression model



Credits: Andrej Karpathy, Breaking Linear Classifier on ImageNet, 2015.

Many machine learning models are subject to adversarial examples, including:

- Neural networks
- Linear models
  - Logistic regression
  - Softmax regression
  - Support vector machines
- Decision trees
- Nearest neighbors

Fooling language understanding models

Article: Super Bowl 50
Paragraph: "*Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager. Quarterback Jeff Dean had jersey number 37 in Champ Bowl XXXIV.*"
Question: "*What is the name of the quarterback who was 38 in Super Bowl XXXIII?*"
Original Prediction: John Elway
Prediction under adversary: Jeff Dean

Figure 1: An example from the SQuAD dataset. The BiDAF Ensemble model originally gets the answer correct, but is fooled by the addition of an adversarial distracting sentence (in blue).

(Jia and Liang, 2017)

Fooling deep structured prediction models

Figure 1: We cause the network to generate a *minion* as segmentation for the adversarially perturbed version of the original image. Note that the original and the perturbed image are indistinguishable.

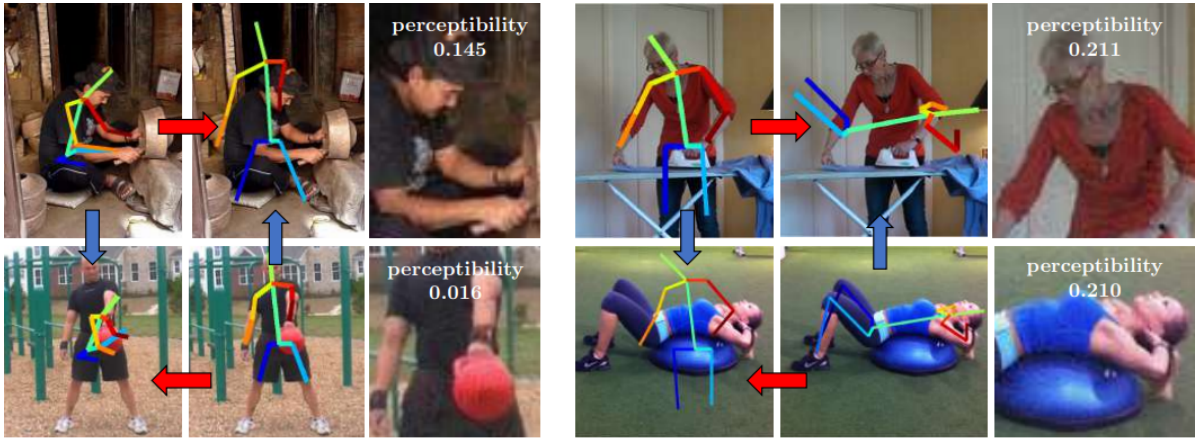
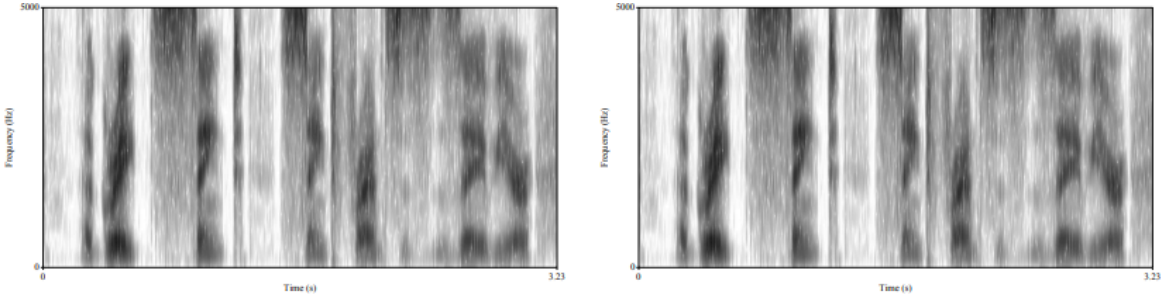


Figure 4: Examples of successful targetted attacks on a pose estimation system. Despite the important difference between the images selected, it is possible to make the network predict the wrong pose by adding an imperceptible perturbation.

(Cisse et al, 2017)

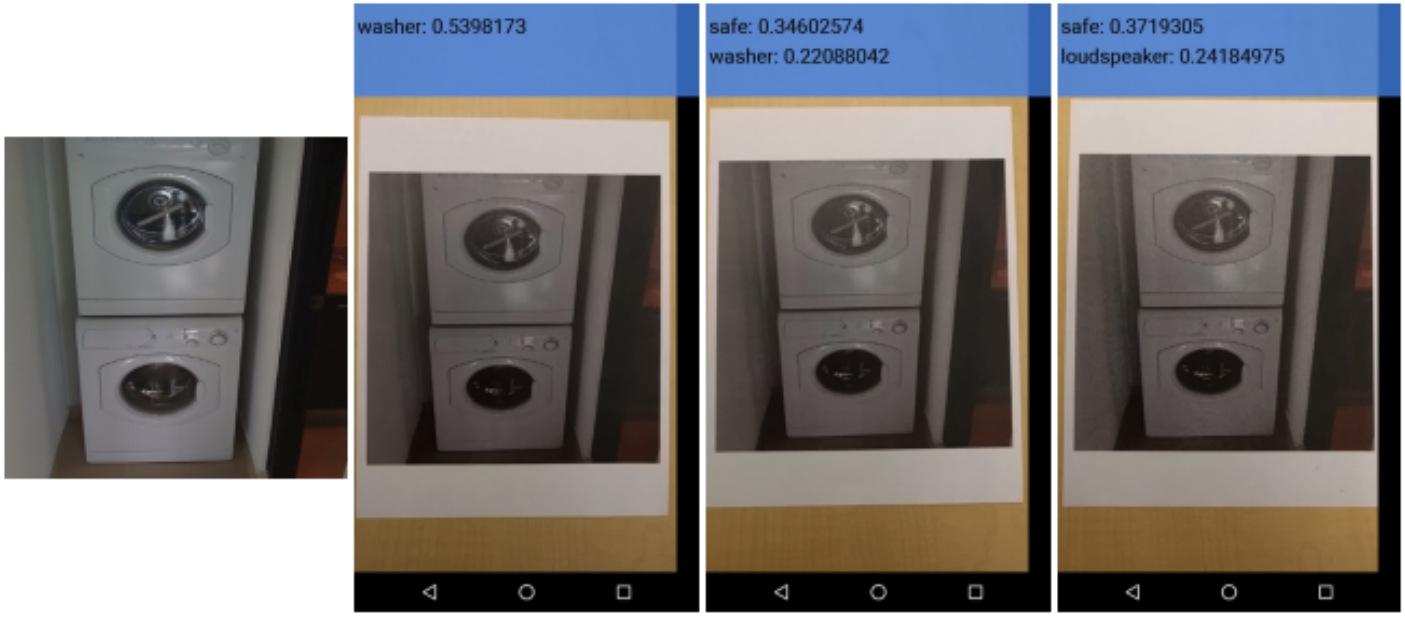


(a) a great saint saint francis zaviour (b) i great sinkt shink t frimsuss avir

Figure 7: The model models' output for each of the spectrograms is located at the bottom of each spectrogram. The target transcription is: `A Great Saint Saint Francis Xavier.`

(Cisse et al, 2017)

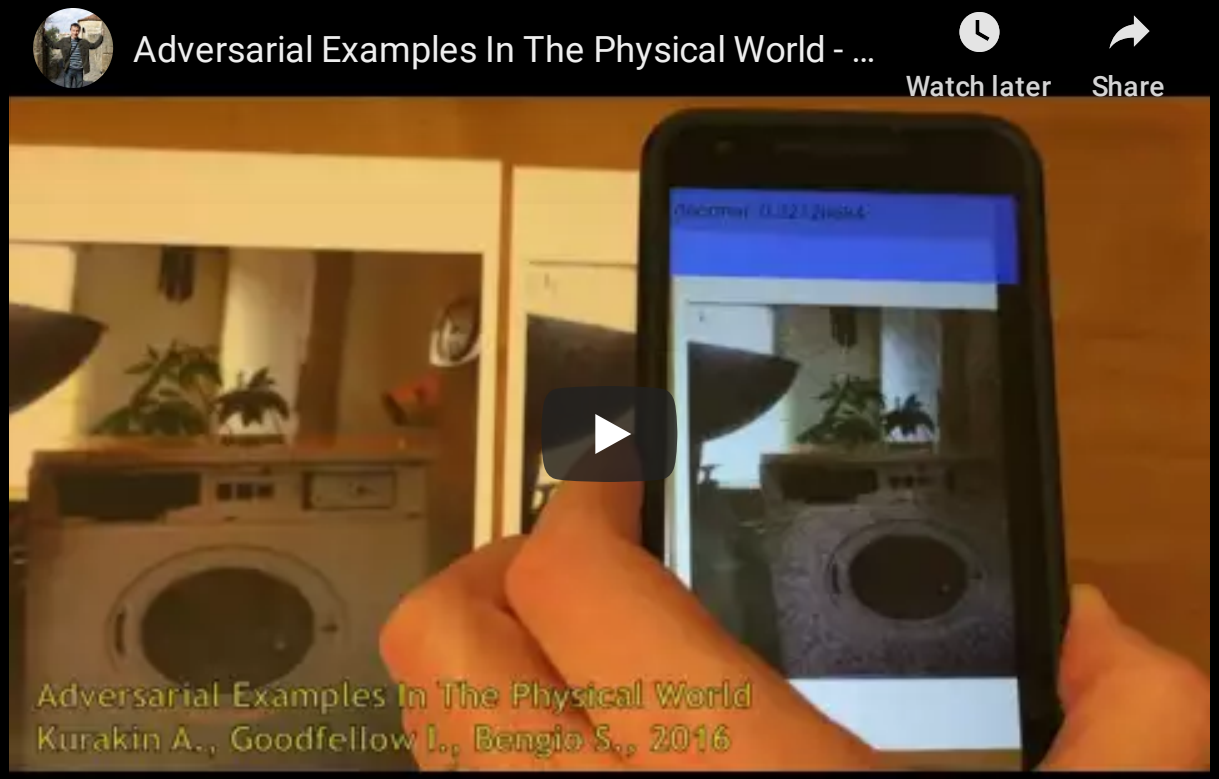
Adversarial examples in the physical world

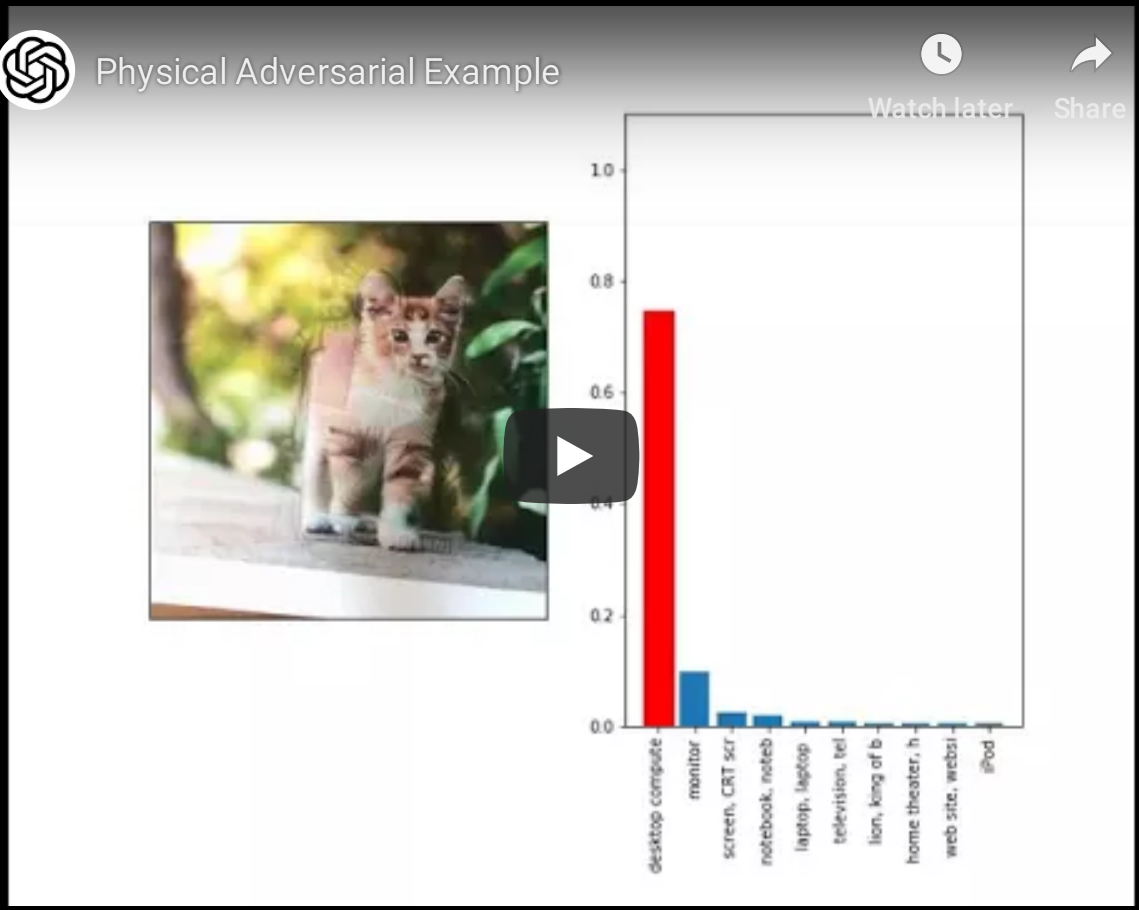


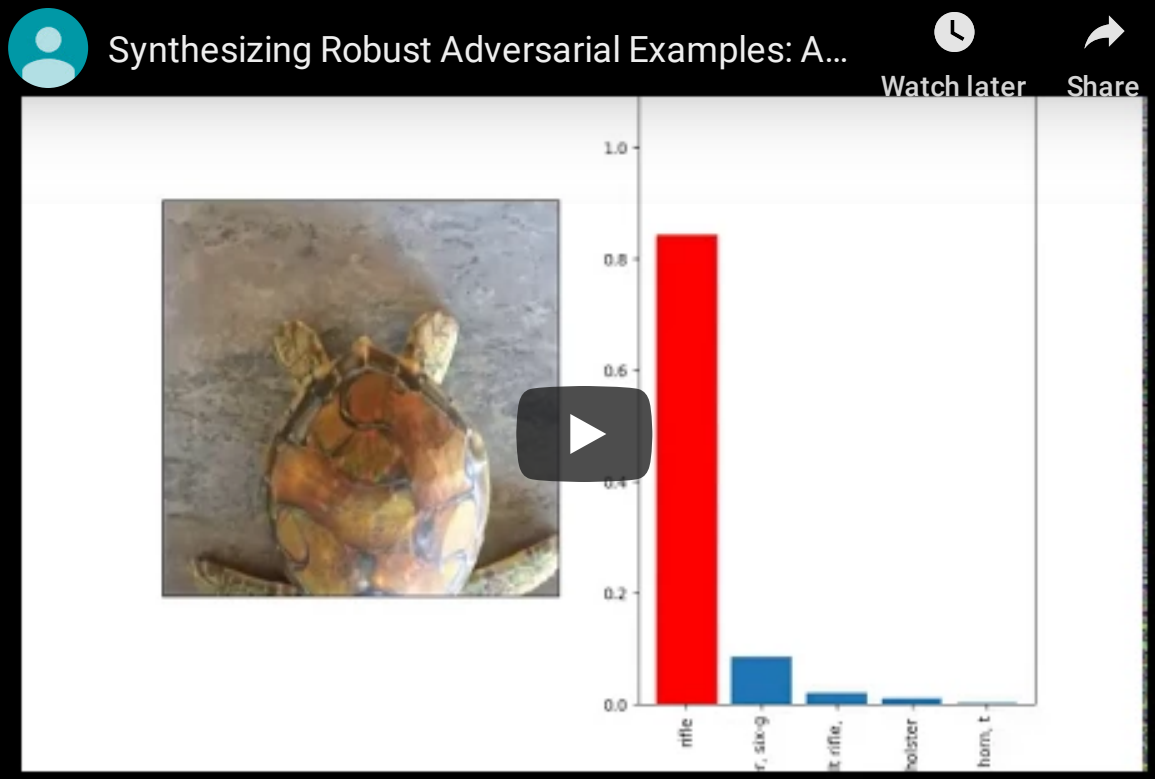
(a) Image from dataset (b) Clean image (c) Adv. image, $\epsilon = 4$ (d) Adv. image, $\epsilon = 8$

Adversarial examples can be printed out on normal paper and photographed with a standard resolution smartphone and still cause a classifier to, in this case, label a "washer" as a "safe".

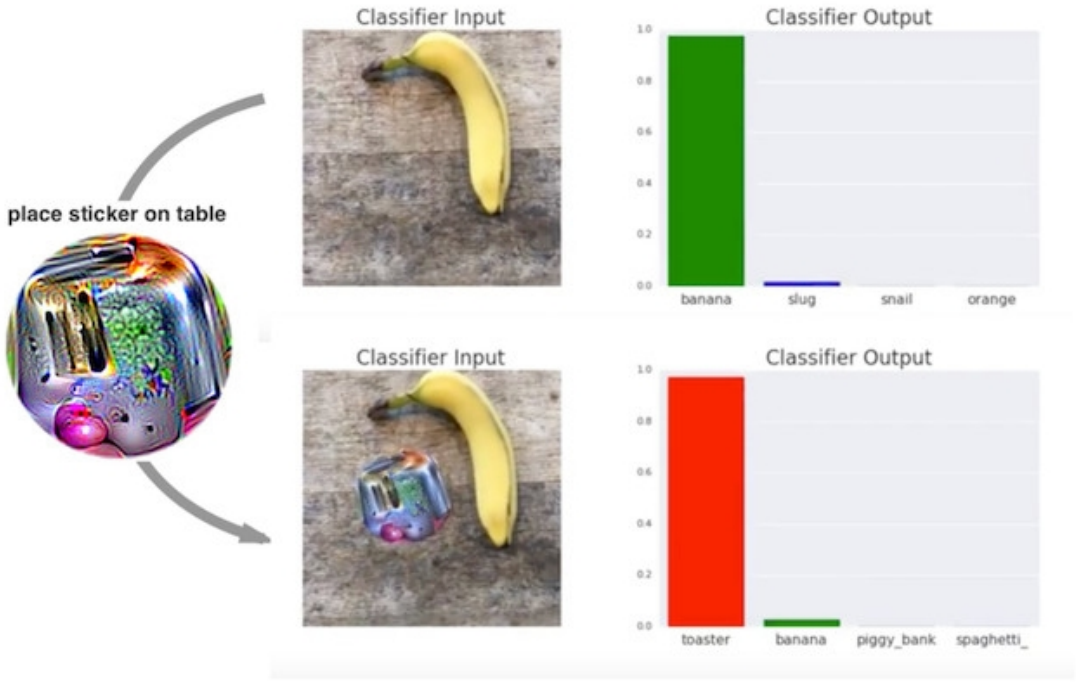
Credits: Kurakin et al, Adversarial examples in the physical world, 2016.


Adversarial Examples In The Physical World - ...
Watch later
Share
Adversarial Examples In The Physical World
Kurakin A., Goodfellow I., Bengio S., 2016


Physical Adversarial Example
Watch later
Share
1.0
0.8
0.6
0.2
0.0
desktop compute
monitor
screen, CRT scr
notebook, noteb
laptop, laptop
television, tel
lion, king of b
home theater, h
web site, websi
iPod


Synthesizing Robust Adversarial Examples: A...
Watch later
Share
1.0
0.8
0.6
0.2
0.0
rifle
horn, t

Adversarial patch



(Brown et al, 2017)

Creating adversarial examples

Locality assumption

"The deep stack of non-linear layers are a way for the model to encode a non-local generalization prior over the input space. In other words, it is assumed that is possible for the output unit to assign probabilities to regions of the input space that contain no training examples in their vicinity.

It is implicit in such arguments that local generalization—in the very proximity of the training examples—works as expected. And that in particular, for a small enough radius $\epsilon > 0$ in the vicinity of a given training input $\mathbf{x}$, an $\mathbf{x} + \mathbf{r}$ satisfying $||\mathbf{r}|| < \epsilon$ will get assigned a high probability of the correct class by the model."

(Szegedy et al, 2013)

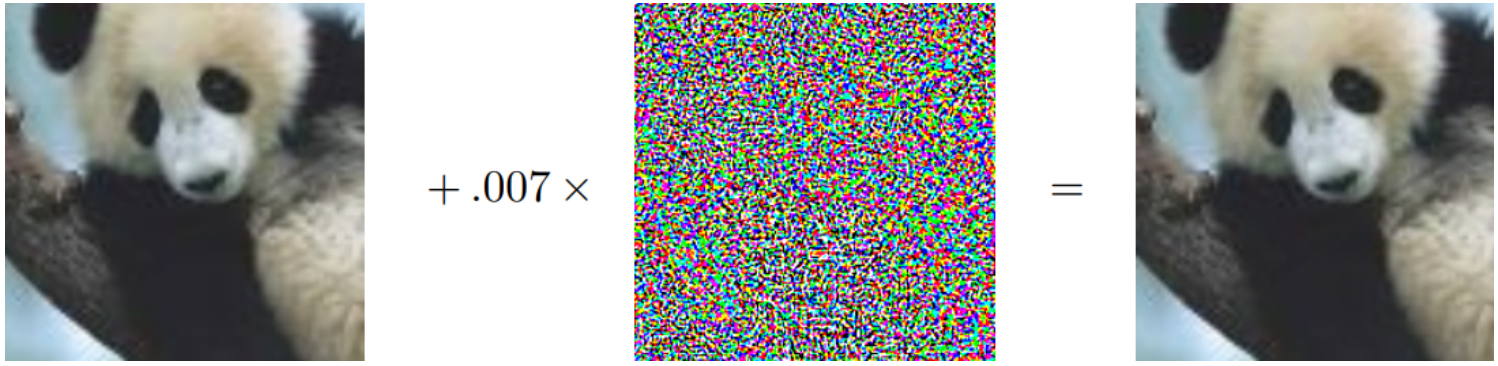
$$\min_{\mathbf{r}} \ell(y_{\text{target}}, f(\mathbf{x} + \mathbf{r}; \theta))$$
$$\text{subject to } ||\mathbf{r}|| \leq L$$

Fast gradient sign method

Take a step along the direction of the sign of the gradient at each pixel,

$$\mathbf{r} = \epsilon \operatorname{sign}(\nabla_{\mathbf{x}} \ell(y_{\text{target}}, f(\mathbf{x}; \theta))),$$

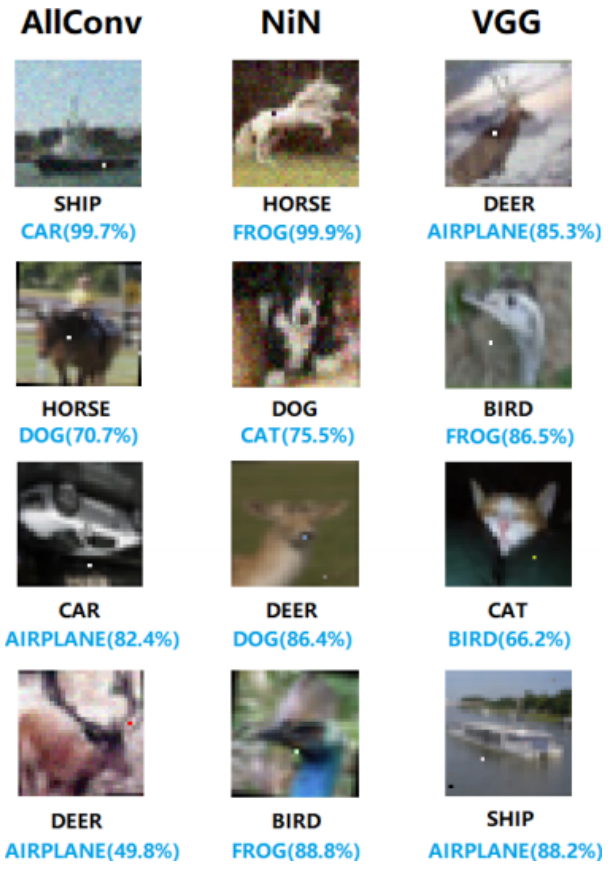
where ϵ is the magnitude of the perturbation.



The panda on the right is classified as a 'Gibbon' (Goodfellow et al, 2014).

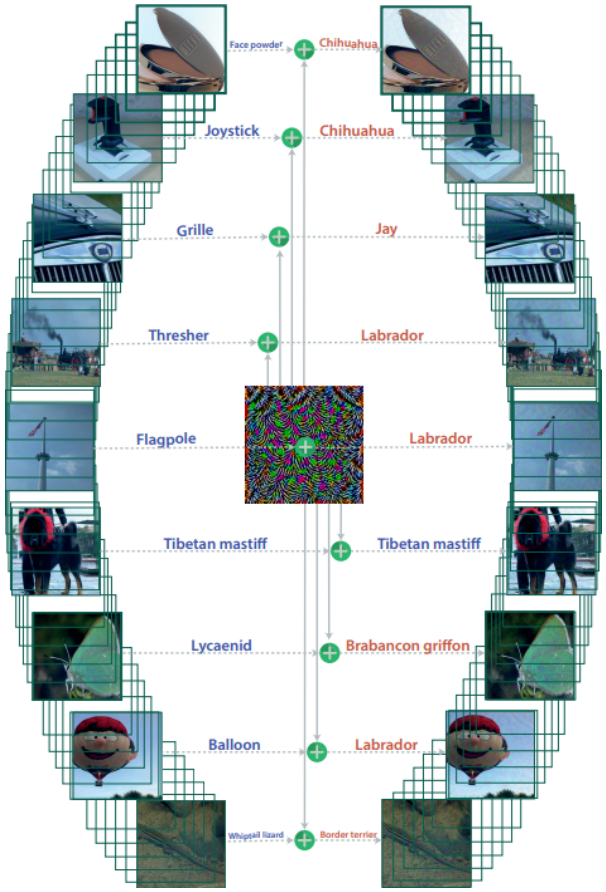
One pixel attacks

$$\min_{\mathbf{r}} \ell(y_{\text{target}}, f(\mathbf{x} + \mathbf{r}; \theta))$$
$$\text{subject to } ||\mathbf{r}||_0 \leq d$$



(Su et al, 2017)

Universal adversarial perturbations



(Moosavi-Dezfooli et al, 2016)

Adversarial defenses

Security threat

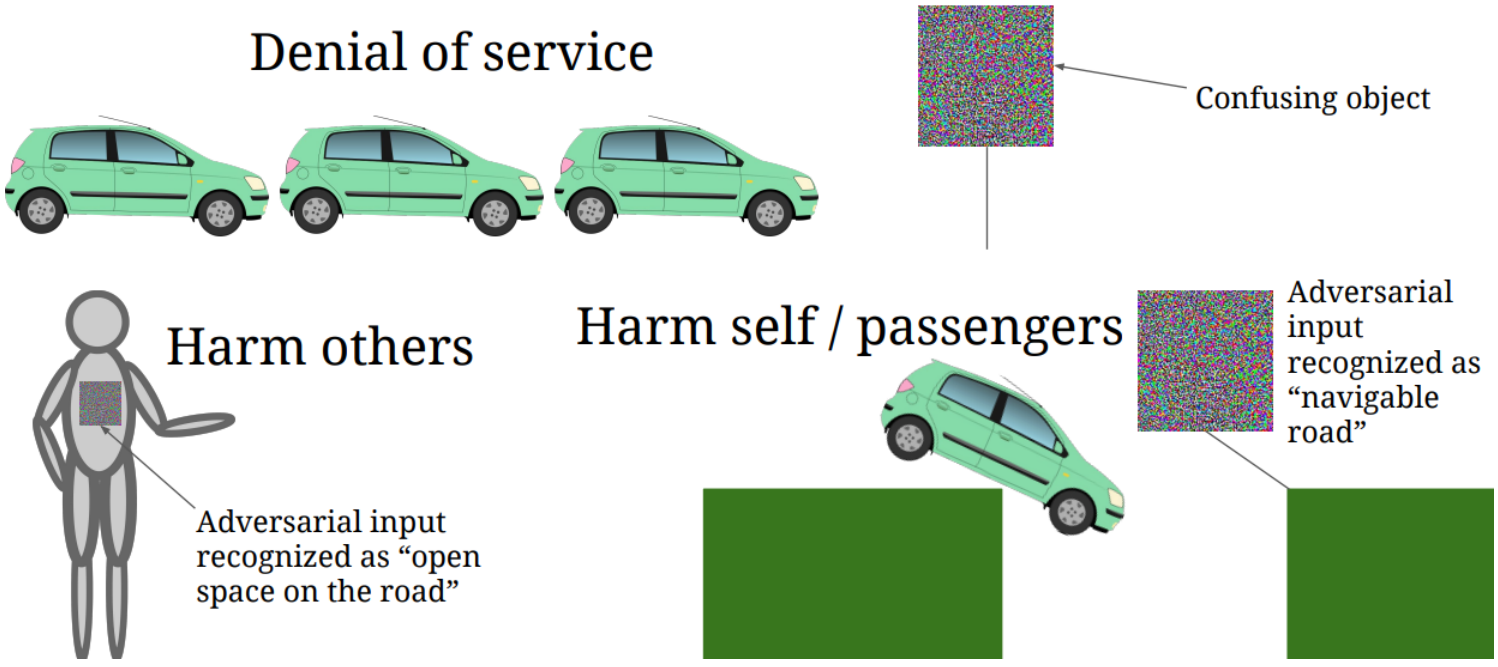
Adversarial attacks pose a serious security threat to machine learning systems deployed in the real world.

Examples include:

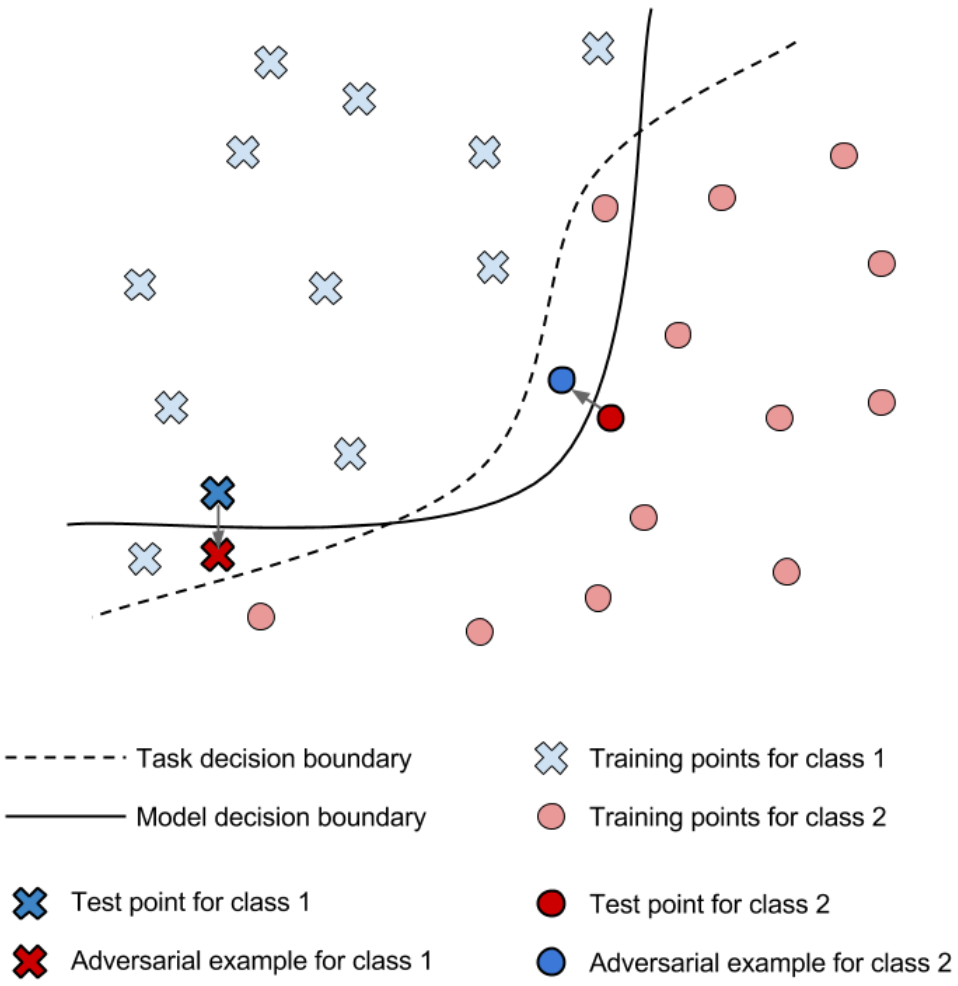
- fooling real classifiers trained by remotely hosted API (e.g., Google),
- fooling malware detector networks,
- obfuscating speech data,
- displaying adversarial examples in the physical world and fool systems that perceive them through a camera.

What if one puts adversarial patches on road signs?
Say, for a self-driving car?

Hypothetical attacks on self-driving cars



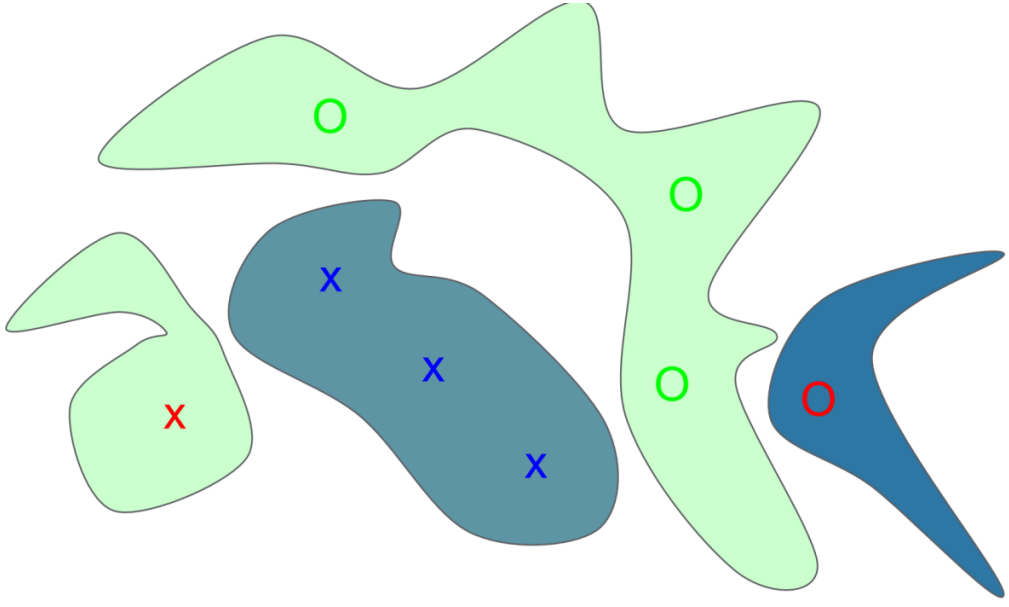
Origins of the vulnerability



Credits: Breaking things easy (Papernot and Goodfellow, 2016)

Conjecture 1: Overfitting

Natural images are within the correct regions, but are also sufficiently close to the decision boundary.



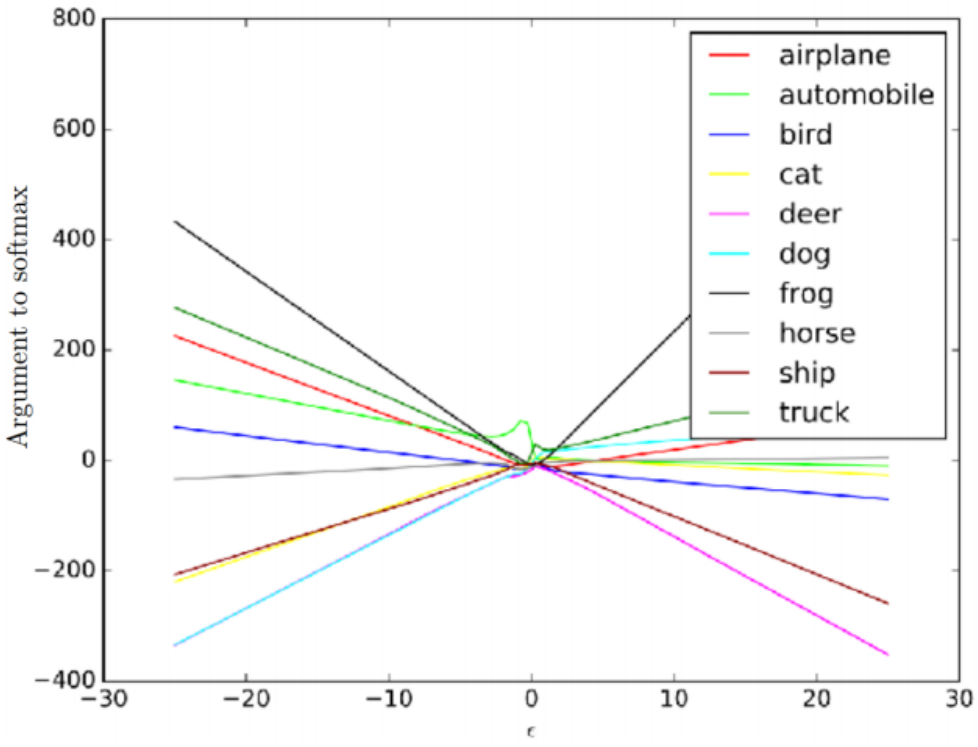
Conjecture 2: Excessive linearity

The decision boundary for most ML models, including neural networks, are near piecewise linear.

Then, for an adversarial sample $\hat{\mathbf{x}}$, its dot product with a weight vector $\mathbf{w}$ is such that

$$\mathbf{w}^T\hat{\mathbf{x}} = \mathbf{w}^T\mathbf{x} + \mathbf{w}^T\mathbf{r}.$$

- The adversarial perturbation causes the activation to grow by $\mathbf{w}^T\mathbf{r}$.
- For $\mathbf{r} = \epsilon \text{sign}(\mathbf{w})$, if $\mathbf{w}$ has n dimensions and the average magnitude of an element is m, then the activation will grow by ϵmn.
- Therefore, for high dimensional problems, we can make many infinitesimal changes to the input that add up to one large change to the output.



Empirical observation: neural networks produce nearly linear responses over ϵ.

Defense

- Data augmentation
- Adversarial training
- Denoising / smoothing

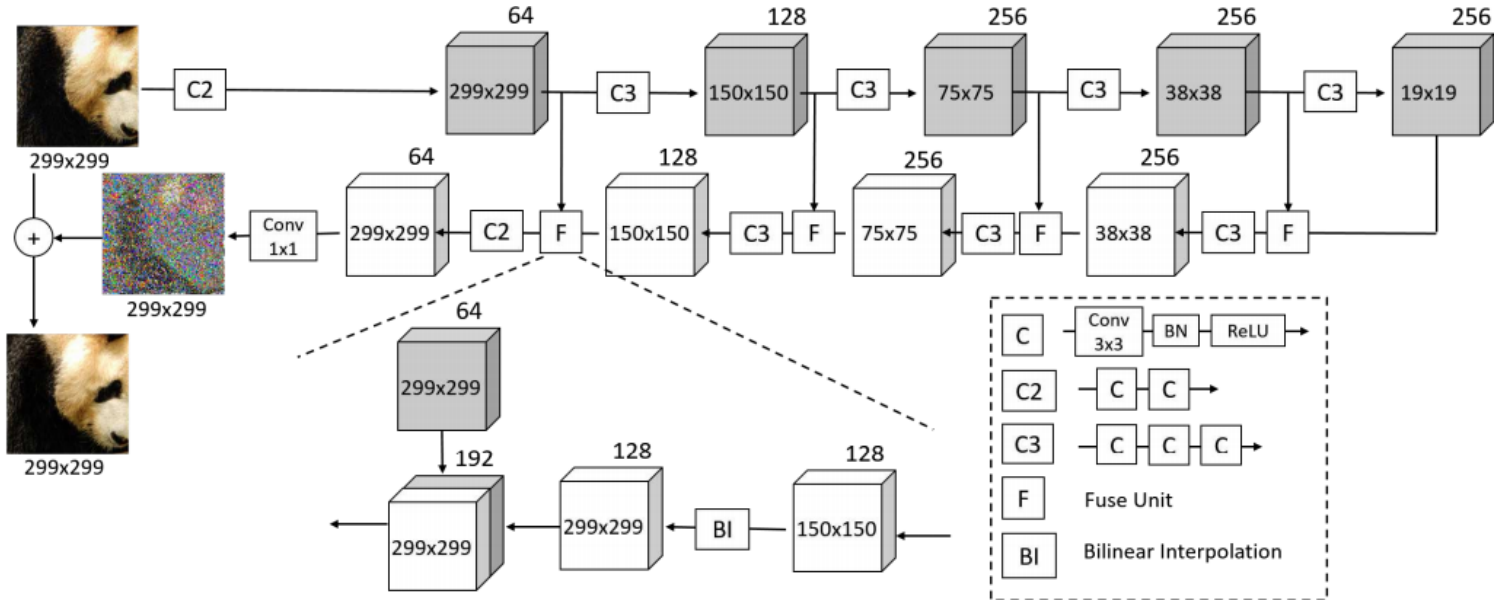
Adversarial training

Generate adversarial examples (based on a given attack) and include them as additional training data.

- Expensive in training time.
- Tends to overfit the attack used during training.

Denoising

- Train the network to remove adversarial perturbations before using the input.
- The winning team of the defense track of the NIPS 2017 competition trained a denoising U-Net to remove adversarial noise.



Credits: Liao et al, Defense against Adversarial Attacks Using High-Level Representation Guided Denoiser, 2017.

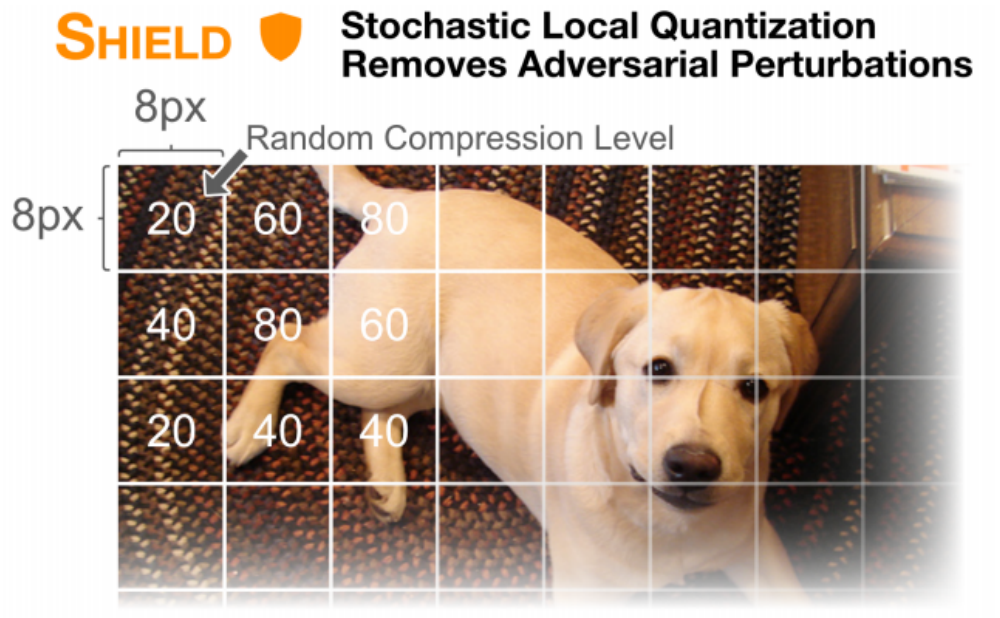


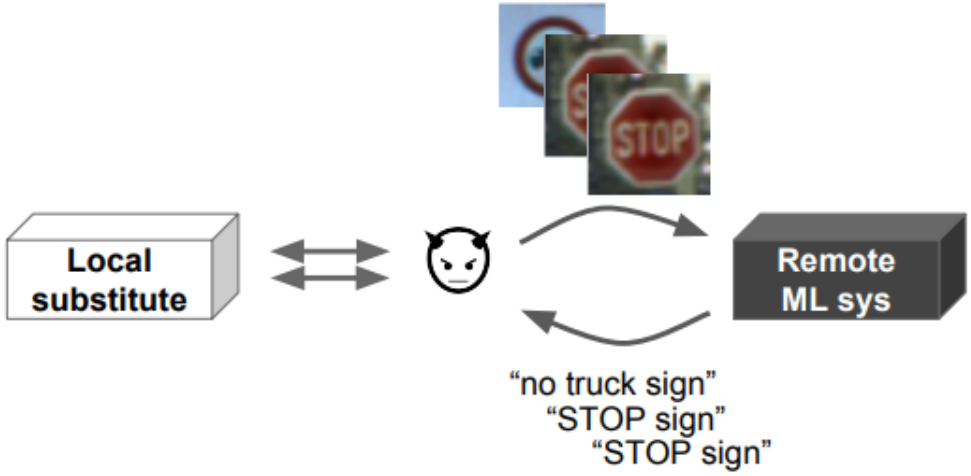
Figure 2: SHIELD uses Stochastic Local Quantization (SLQ) to remove adversarial perturbations from input images. SHIELD divides images into 8×8 blocks and applies a randomly selected JPEG compression quality (20, 40, 60 or 80) to each block to remove adversarial attacks. Note this figure is an illustration; our images are of actual size 299×299.

Credits: Das et al, Shield: Fast, Practical Defense and Vaccination for Deep Learning using JPEG Compression, 2018.

Hiding information

Attacks considered so far are white-box attacks, for which the attack has full access to the model.

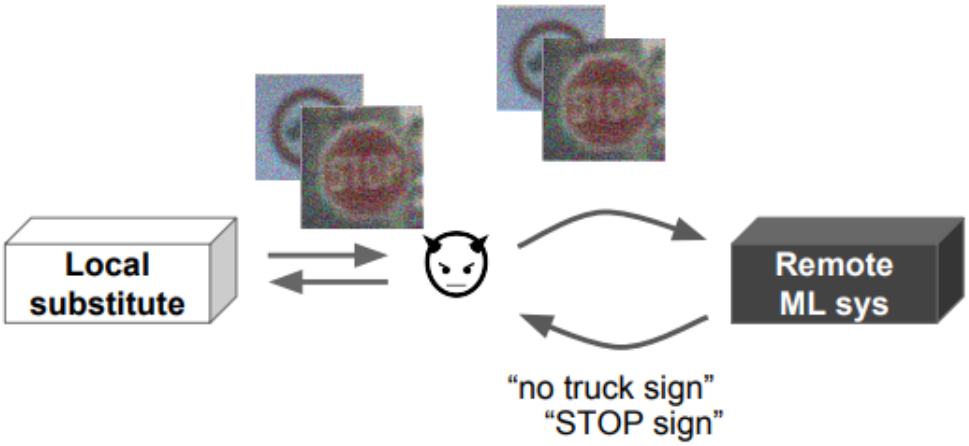
- What if instead the model internals remain hidden?
- Are models prone to black-box attacks?



(1) The adversary queries the target remote ML system for labels on inputs of its choice.

(2) The adversary uses the labeled data to train a local substitute of the remote system.

Credits: Papernot et al, Practical Black-Box Attacks against Machine Learning, 2016.

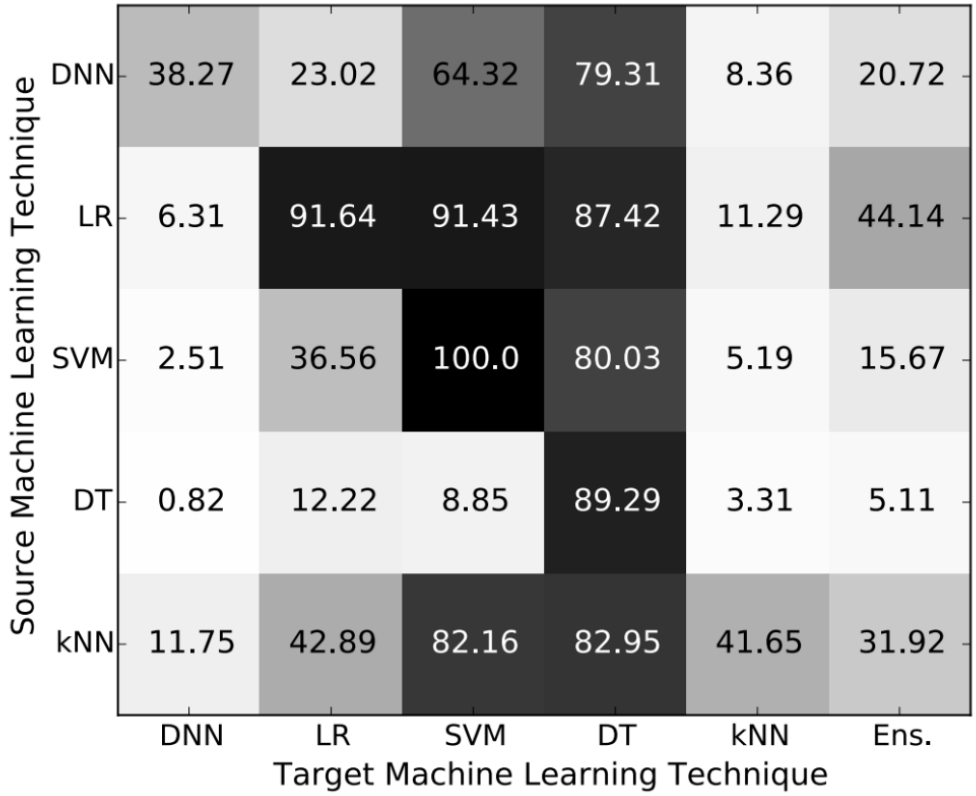


(3) The adversary selects new synthetic inputs for queries to the remote ML system based on the local substitute's output surface sensitivity to input variations.

Credits: Papernot et al, Practical Black-Box Attacks against Machine Learning, 2016.

Transferrability

Adversarial examples are transferable across ML models!



Credits: Papernot et al, Practical Black-Box Attacks against Machine Learning, 2016.

Failed defenses

"In this paper we evaluate ten proposed defenses and demonstrate that none of them are able to withstand a white-box attack*. We do this by constructing defense-specific loss functions that we minimize with a strong iterative attack algorithm. With these attacks, on CIFAR an adversary can create imperceptible adversarial examples for each defense.*

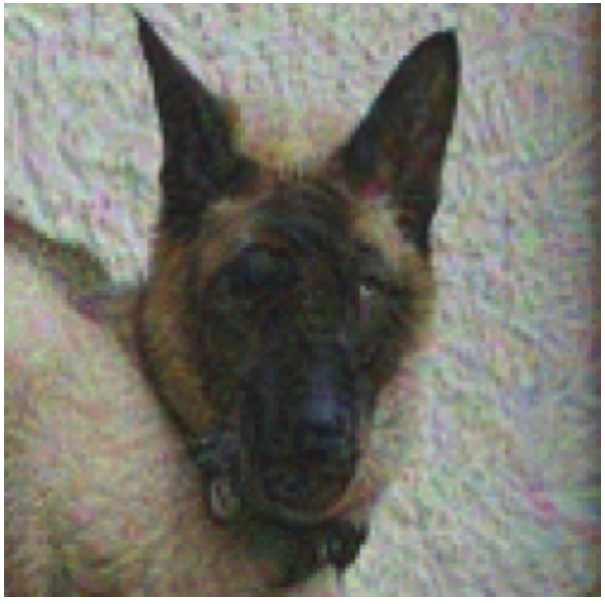
By studying these ten defenses, we have drawn two lessons: existing defenses lack thorough security evaluations, and adversarial examples are much more difficult to detect than previously recognized."

(Carlini and Wagner, 2017)

"No method of defending against adversarial examples is yet completely satisfactory. This remains a rapidly evolving research area."

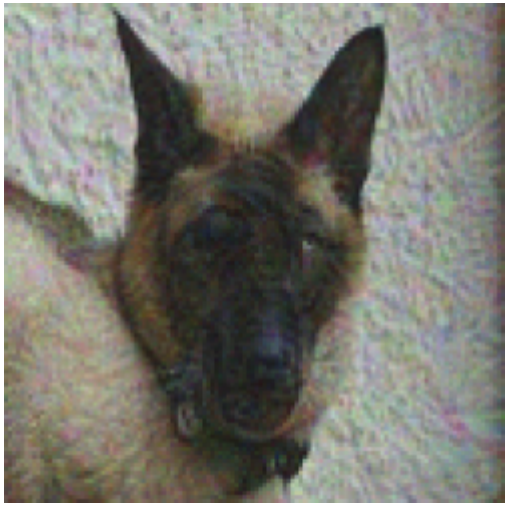
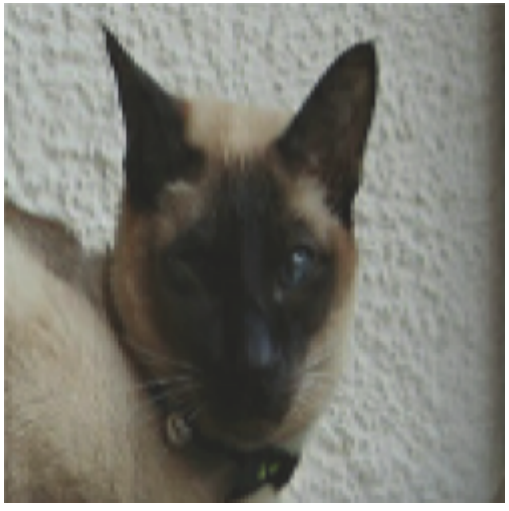
(Kurakin, Goodfellow and Bengio, 2018)

Fooling both computers and humans



What do you see?

Credits: Elsayed et al, Adversarial Examples that Fool both Computer Vision and Time-Limited Humans, 2018.

By building neural network architectures that closely match the human visual system, adversarial samples can be created to fool humans.

Credits: Elsayed et al, Adversarial Examples that Fool both Computer Vision and Time-Limited Humans, 2018.

That's all folks!