Deep Learning

Lecture 4: Adversarial attacks and defenses

Gilles Louppe
g.louppe@uliege.be

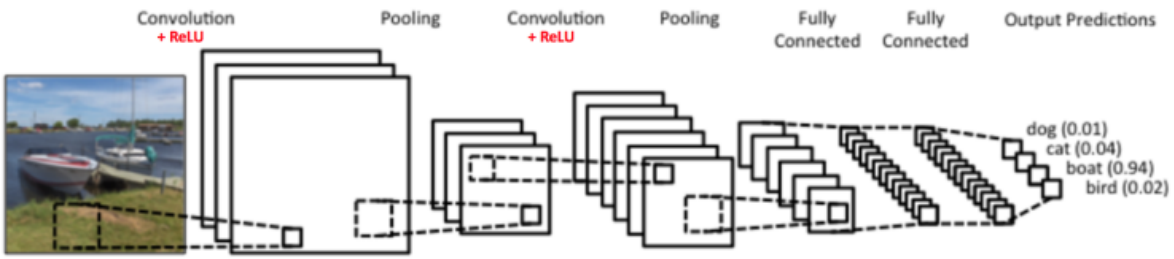
LIÈGE université

We have seen that (convolutional) neural 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?



A recipe for success, or is it?

Adversarial attacks

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)

Adversarial examples

$$\begin{aligned} &\min ||\mathbf{r}||_2 \\ \text{s.t. } &f(\mathbf{x}+\mathbf{r}) = y' \\ &\mathbf{x}+\mathbf{r} \in [0,1]^p \end{aligned}$$

where

- y' is some target label, different from the original label y associated to $\mathbf{x}$,
- f is a trained neural network.

(Left) Original images $\mathbf{x}$. (Middle) Noise $\mathbf{r}$. (Right) Modified images $\mathbf{x} + \mathbf{r}$.
All are classified as 'Ostrich'. (Szegedy et al, 2013)

Even simpler, take a step along the direction of the sign of the gradient at each pixel:

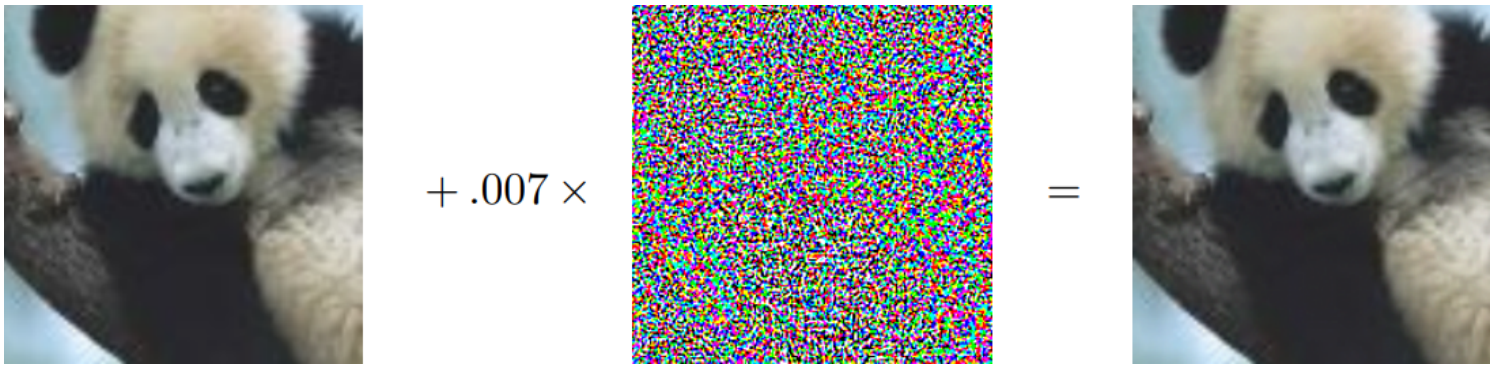
$$\mathbf{r} = \epsilon \operatorname{sign}(\nabla_{\mathbf{x}} \ell(y', f(\mathbf{x})))$$

where ϵ is the magnitude of the perturbation.

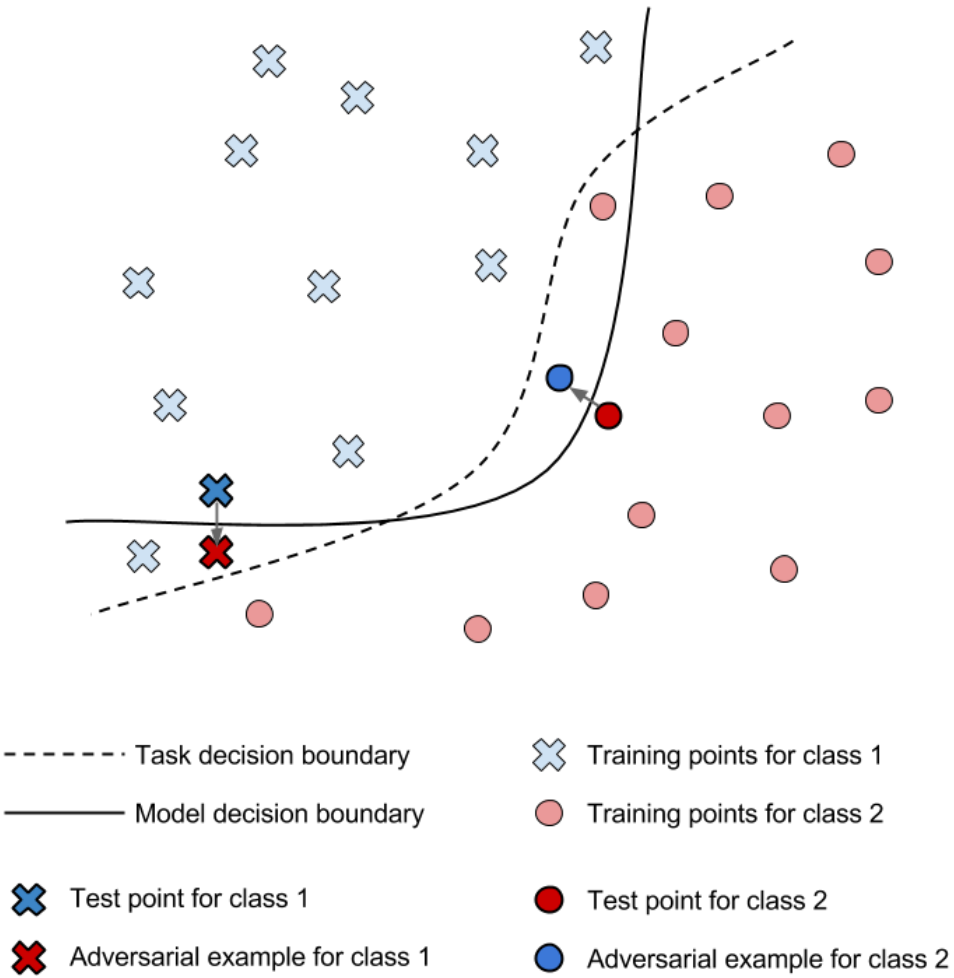
Even simpler, take a step along the direction of the sign of the gradient at each pixel:

$$\mathbf{r} = \epsilon \operatorname{sign}(\nabla_{\mathbf{x}} \ell(y', f(\mathbf{x})))$$

where ϵ is the magnitude of the perturbation.



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



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

Not just for neural networks

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

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

Fooling neural networks

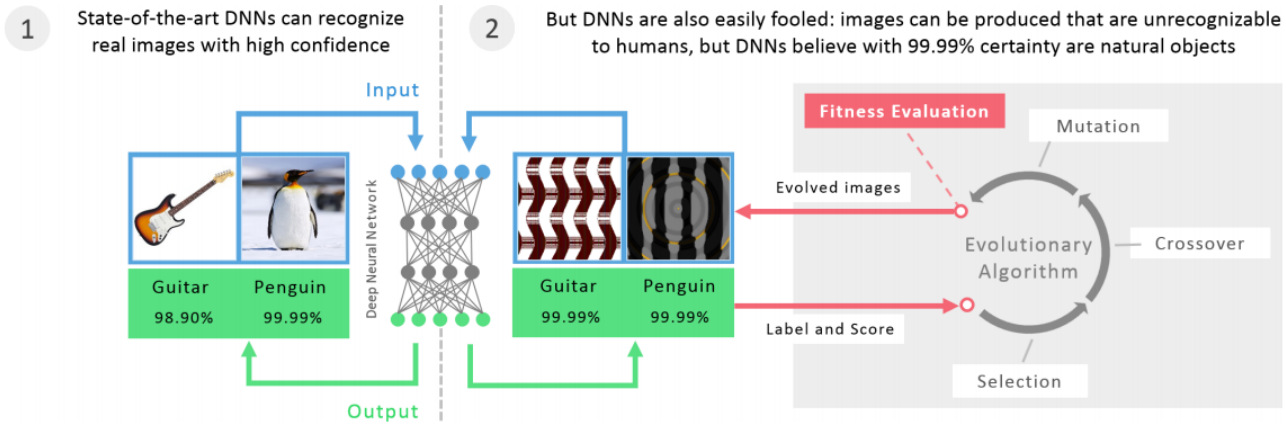


Figure 2. Although state-of-the-art deep neural networks can increasingly recognize natural images (*left panel*), they also are easily fooled into declaring with near-certainty that unrecognizable images are familiar objects (*center*). Images that fool DNNs are produced by evolutionary algorithms (*right panel*) that optimize images to generate high-confidence DNN predictions for each class in the dataset the DNN is trained on (here, ImageNet).

(Nguyen et al, 2014)

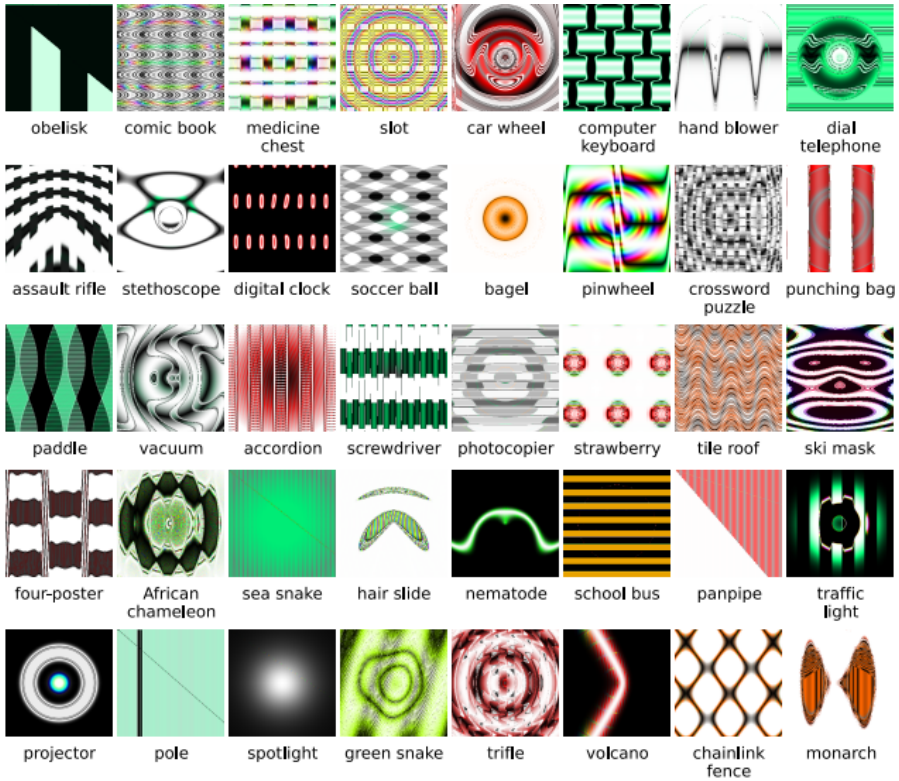
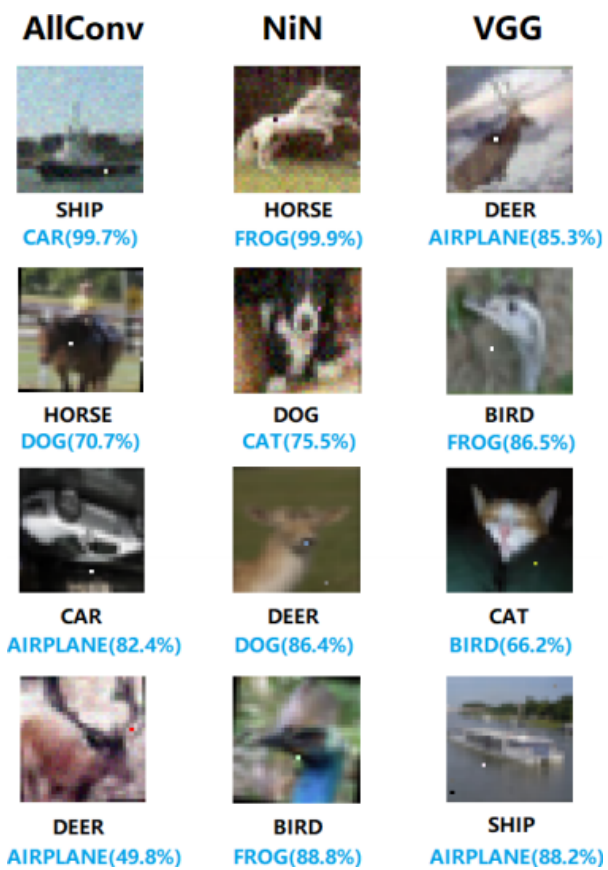


Figure 8. Evolving images to match DNN classes produces a tremendous diversity of images. Shown are images selected to showcase diversity from 5 evolutionary runs. The diversity suggests that the images are non-random, but that instead evolutions producing discriminative features of each target class. The mean DNN confidence scores for these images is 99.12%.

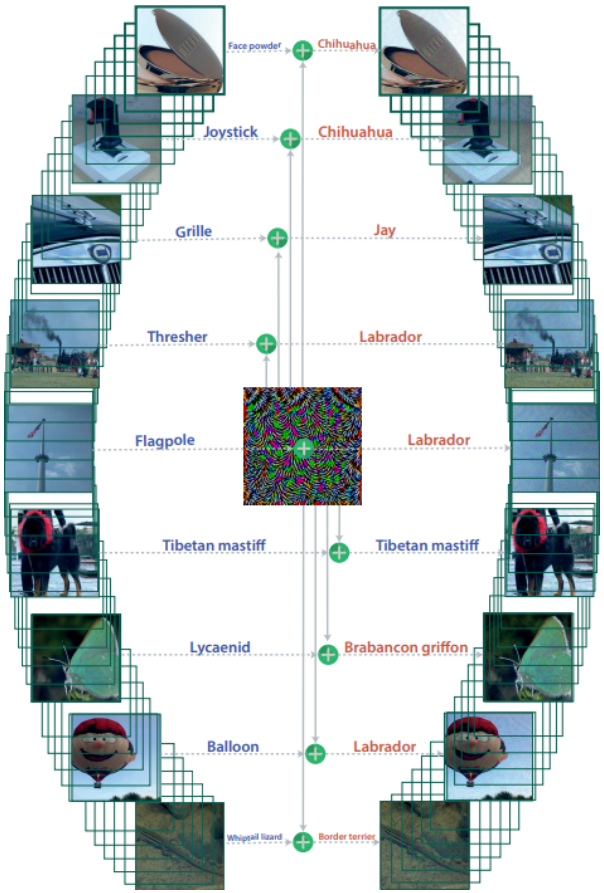
(Nguyen et al, 2014)

One pixel attacks



(Su et al, 2017)

Universal adversarial perturbations



(Moosavi-Dezfooli et al, 2016)

Fooling deep structured prediction models

original semantic segmentation framework | adversarial attack | compromised semantic segmentation framework

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.

(Cisse et al, 2017)

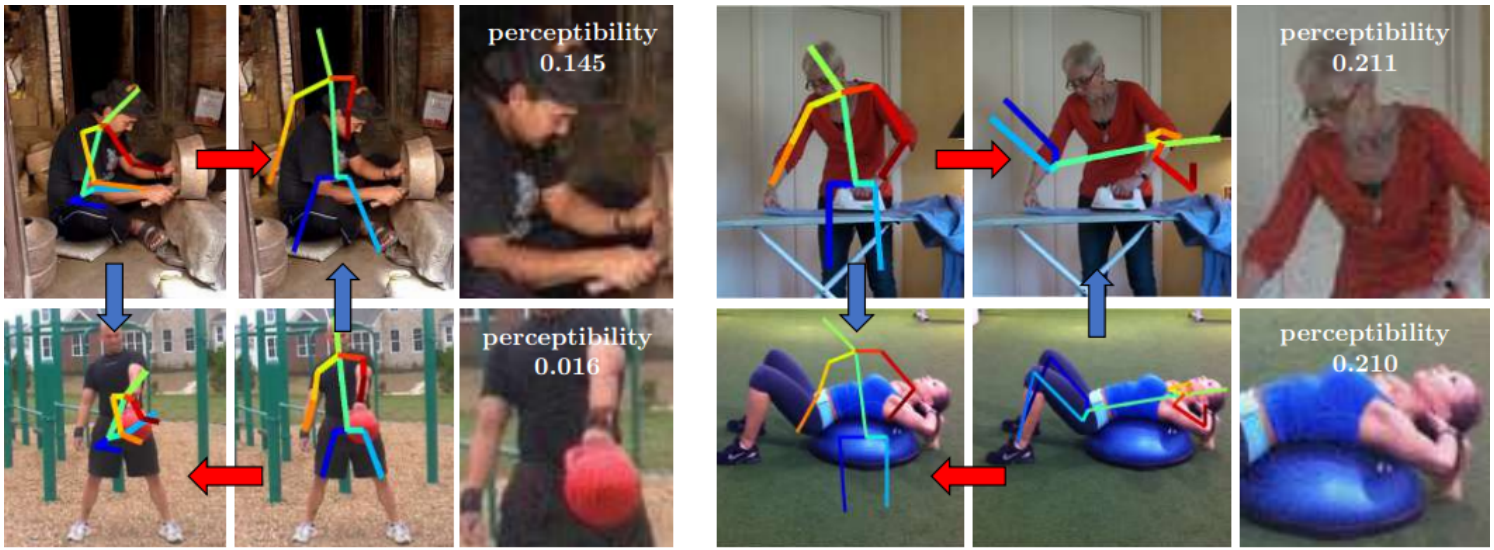
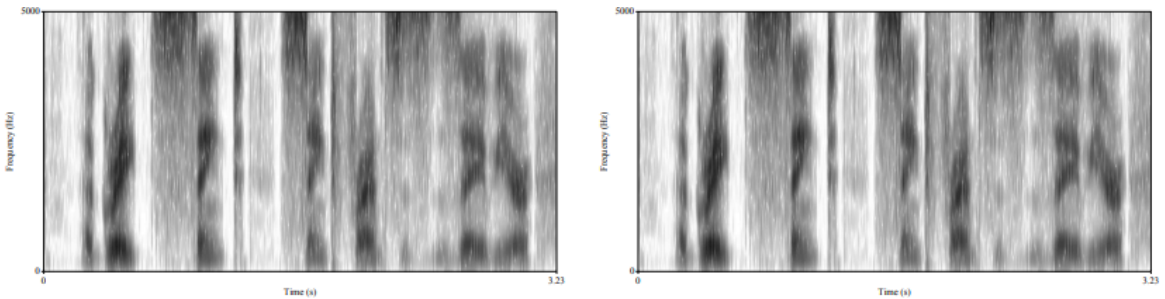


Figure 4: Examples of successful targetted attacks on a pose estimation system. Despite the important difference between the images selected, it is possible to make the network predict the wrong pose by adding an imperceptible perturbation.

(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)

Attacks in the real world

Security threat

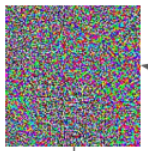
Adversarial attacks pose a security threat to machine learning systems deployed in the real world.

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.

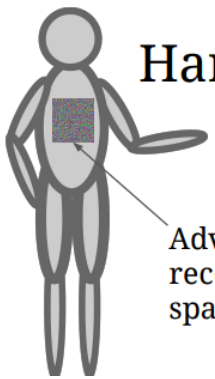
Hypothetical Attacks on Autonomous Vehicles

Denial of service



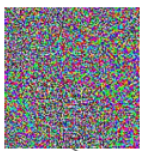
Confusing object

Harm others



Adversarial input recognized as “open space on the road”

Harm self / passengers



Adversarial input recognized as “navigable road”

Credits: Adversarial Examples and Adversarial Training (Goodfellow, 2016)

Adversarial defenses

Defenses

Generative pretraining

Removing perturbation with an autoencoder

Adding noise at test time

Ensembles

Confidence-reducing perturbation at test time

Error correcting codes

Multiple glimpses

Weight decay

Double backprop

Adding noise at train time

Various non-linear units

Dropout

Credits: Adversarial Examples and Adversarial Training (Goodfellow, 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)

Adversarial attacks and defenses remain an open research problem.

The end.

Further readings

- Adversarial Examples: Attacks and Defenses for Deep Learning (Yuan et al, 2017)
- Breaking things easy (Papernot and Goodfellow, 2016)
- Adversarial Examples and Adversarial Training (Goodfellow, 2016)
- Breaking Linear Classifiers on ImageNet (Andrej Karpathy, 2015)