Fooling Deep Networks:

Generation, Explanation and Detection of Adversarial Attacks

Guillaume Debard, Mélanie Ducoffe, Frédéric Precioso

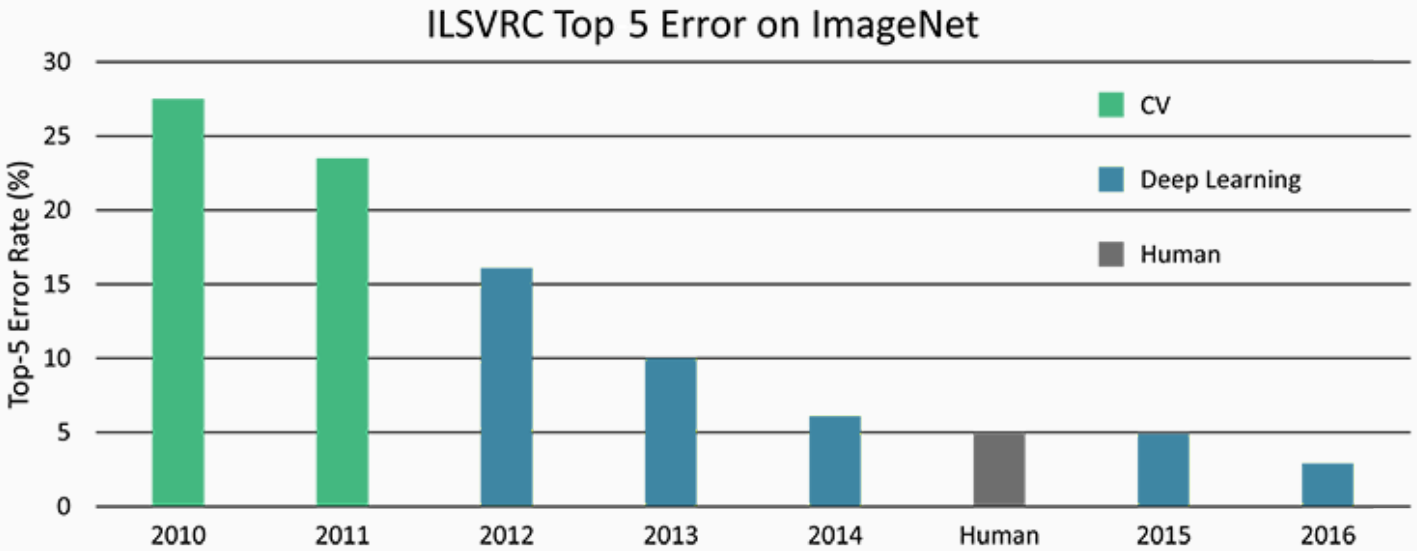
Laboratoire I3S - UNS CNRS UMR 7271

July 5, 2017

An Adversarial Example tour

We did it!

- Deep Networks are as good as humans at recognition, identification...



How much does a deep network understands those tasks?

Why does it matter?

Google trends on "deep learning" keyword

- Natural communication between humans and computer (working together)
- Preventing mistakes and establishing norms (autonomous driving ...)

Intriguing properties of neural networks

C. Szegedy, w. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, R. Fergus

arXiv preprint arXiv:1312.6199

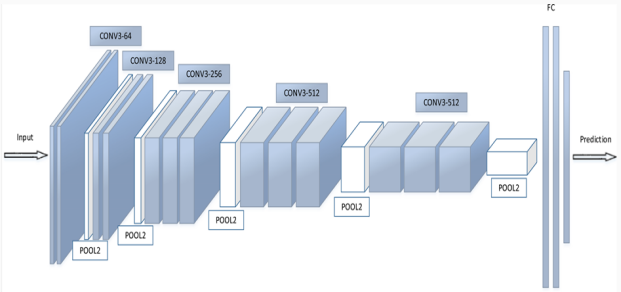
2013

[1312.6199] Intriguing properties of neural networks - arXiv.org
https://arxiv.org › cs - Traduire cette page
de C Szegedy - 2013 - Cité 449 fois - Autres articles
21 déc. 2013 - In this paper we report two such **properties**. First, we ... Second, we find that deep **neural networks** learn input-output mappings that are fairly ...

A Simple Experiment: What we expected

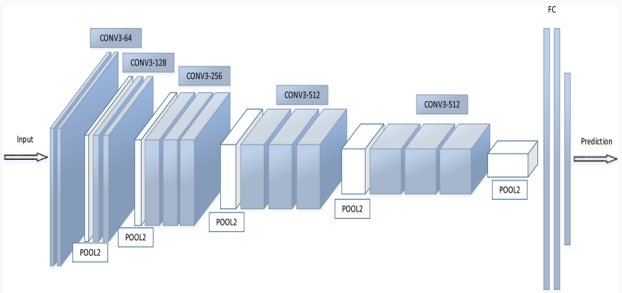
Input

Network's prediction

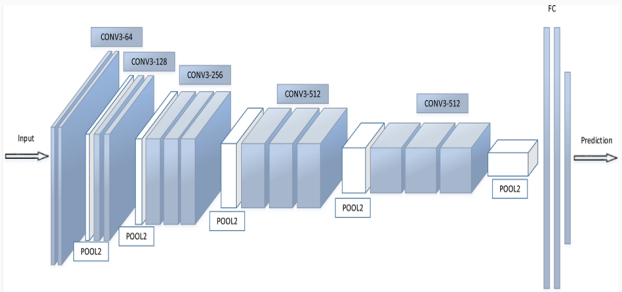


" This is a car !"

backpropagation to modify the pixels



changing the prediction

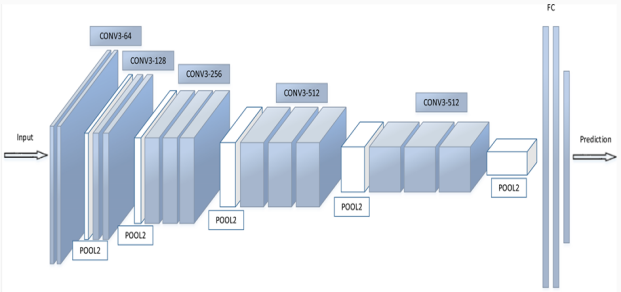


"This is a plane !"

A Simple Experiment: What really happened

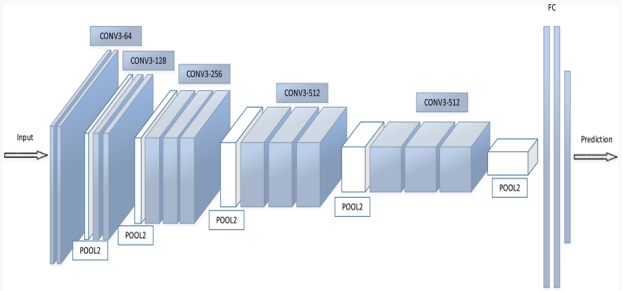
Input

Network's prediction

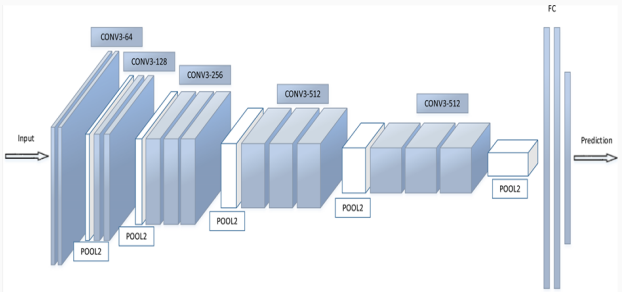


" This is a car !"

backpropagation to modify the pixels

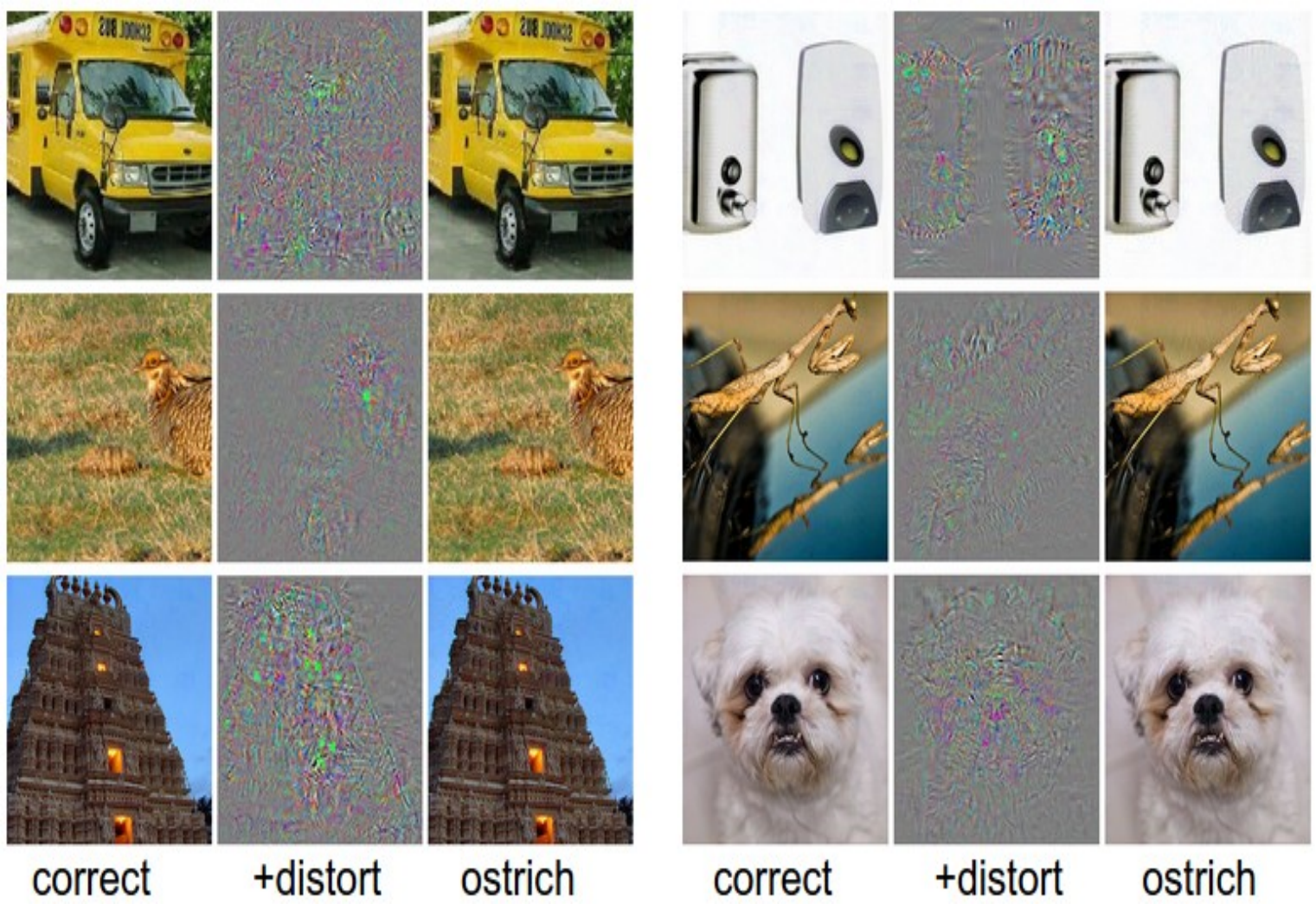


changing the prediction

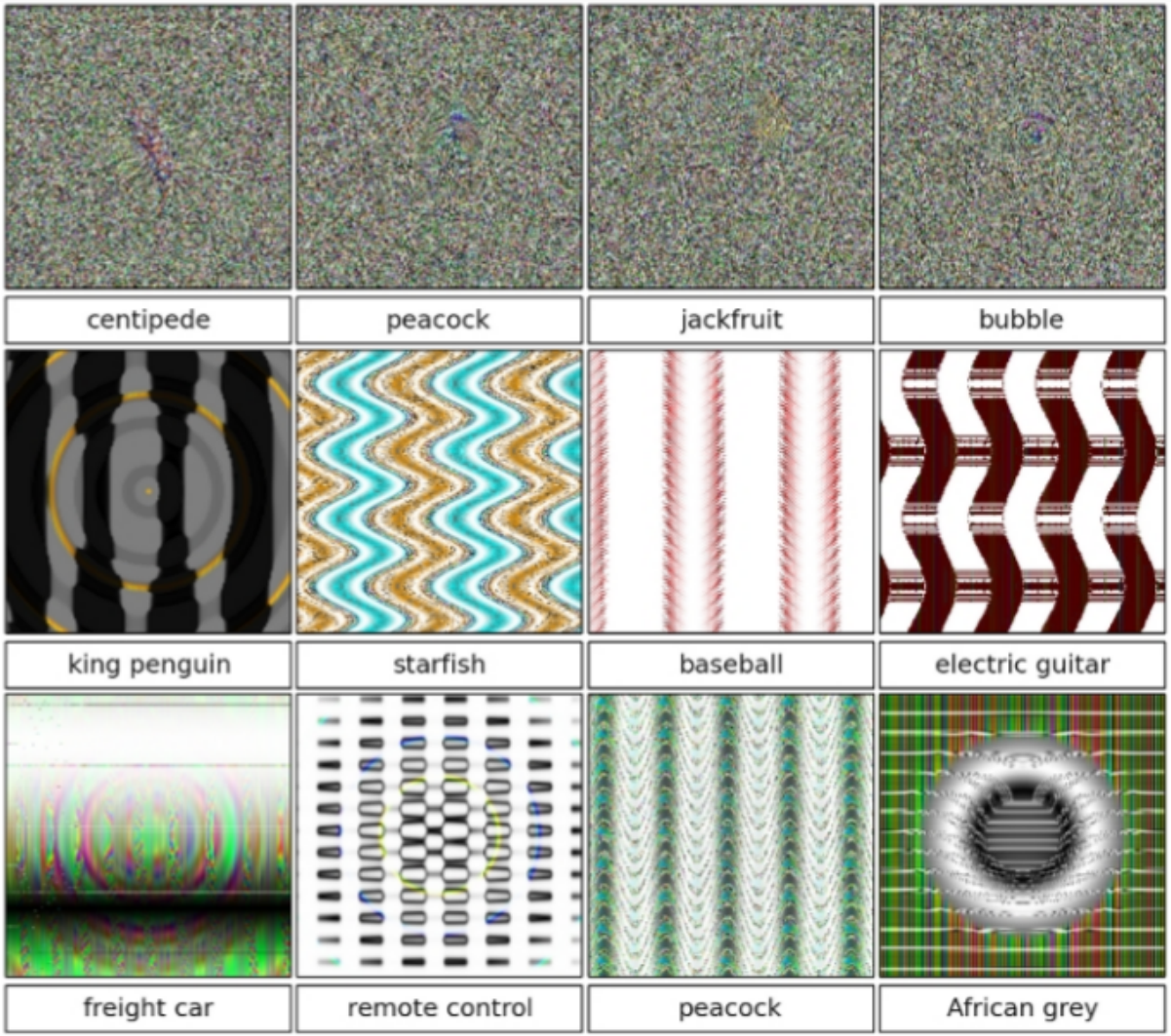


"This is a plane !"

Orienting mis-predictions



Pushing the "bouchon"



Confidence $\geqslant 96\%$

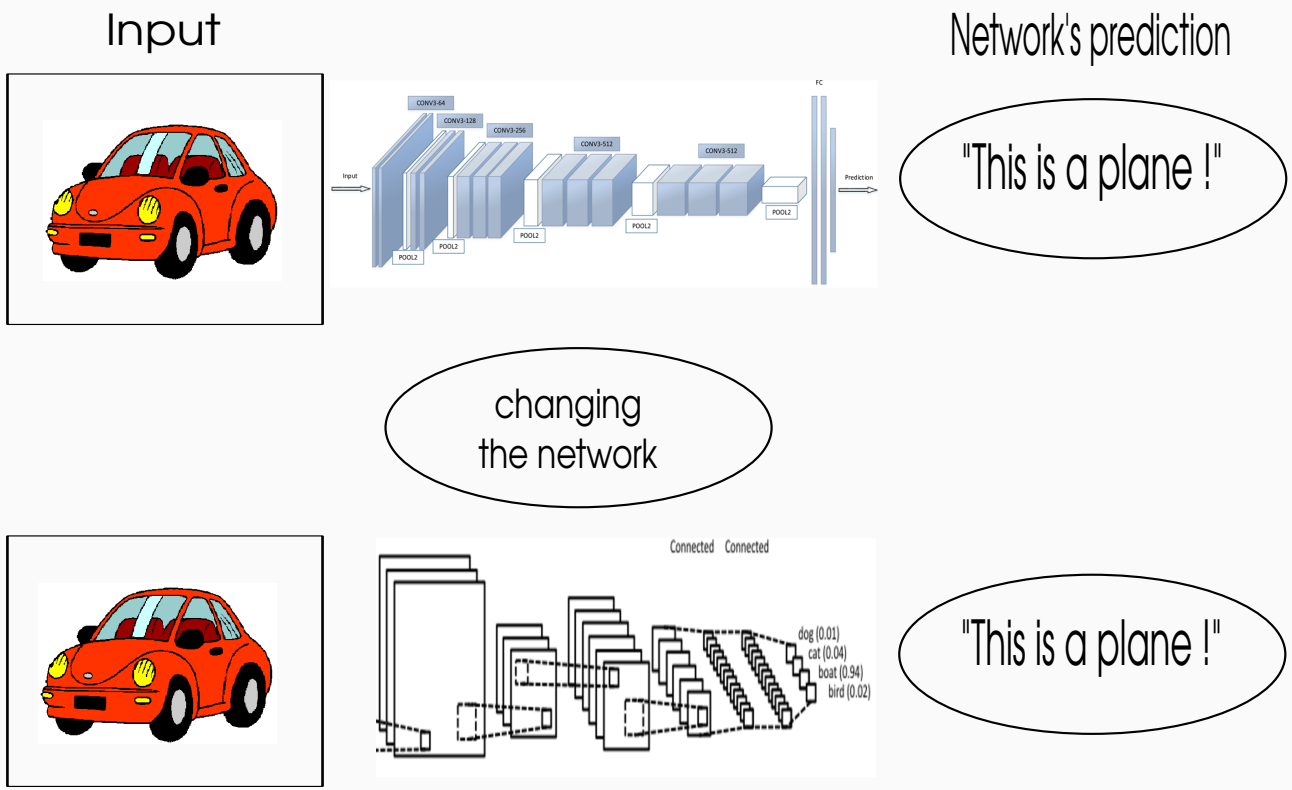
Definition: Adversarial Example

Definition: $\hat{x}$ is called adversarial iff:

- given image x
- low distortion $|| x - \hat{x} ||< \epsilon, \quad (\epsilon > 0$, few pixels)
- given network's probabilities $f_\theta(x)$
- Different predictions! $argmax f_\theta(x) \neq argmax f_\theta(\hat{x})$

Properties: Transferability

- $\neq$ outliers
- regularization: correct one... find another
- high confidence predictions
- **Transferability**



Attacks on different models

- CNNs are not robust to adversarial
- Adversarial Attacks:
  - RNNs *(Crafting Adversarial Input Sequences for Recurrent Neural Networks* - N. Papernot, P. McDaniel, A. Swami, R. Harang; 2016)
  - Generative models *(Adversarial Images for Variational Autoencoders* - P. Tabacof, J. Tavares, E. Valle; 2016)
  - Reinforcement Learning *(Adversarial Attacks on Neural Network Policies* - S. Huang, N. Papernot, I. Goodfellow, Y. Duan, P. Abbeel; 2017)

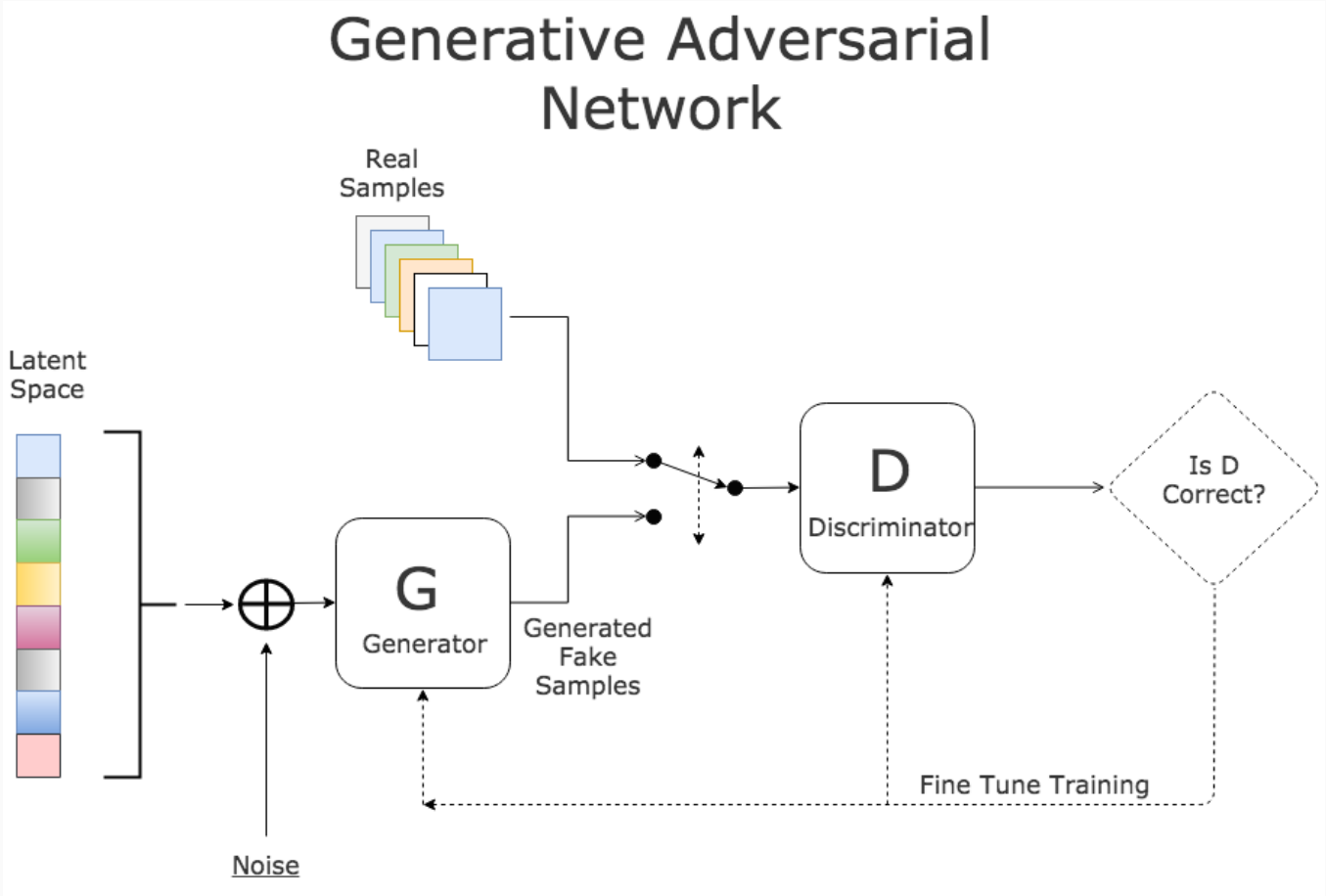
Curious about attacking video games? Videos are here :
http://rll.berkeley.edu/adversarial/

How to attack a Deep Network?

First steps and GAN

Originally designed for crafting and training on adversarial examples

$\rightarrow$ Not the case, shown to be useful in other tasks



Fast Gradient Sign

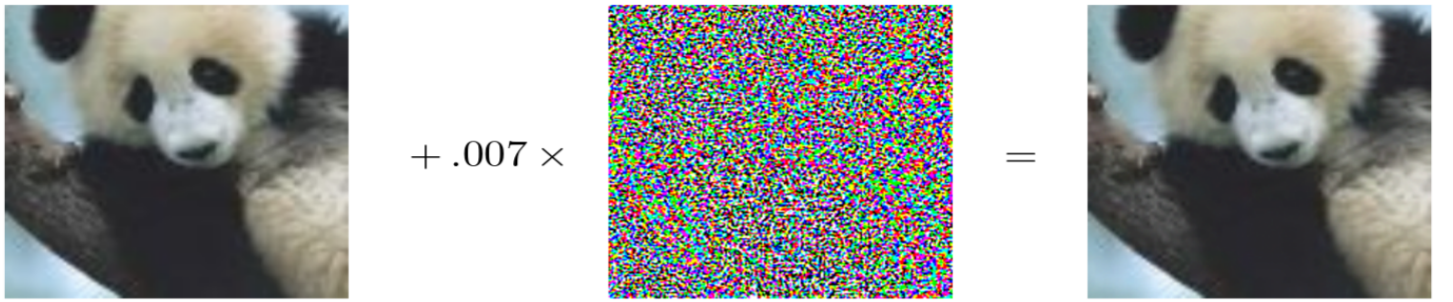
- fast, but simple attacks
- used mostly for regularization

Input : Image x, Classifier f_θ, ϵ

Prediction phase: Perturbed image $\hat{x}$

1 $y = argmax_k f_\theta(x)$

2 return $x + \epsilon sign(\nabla_x loss(f_\theta(x), y))$



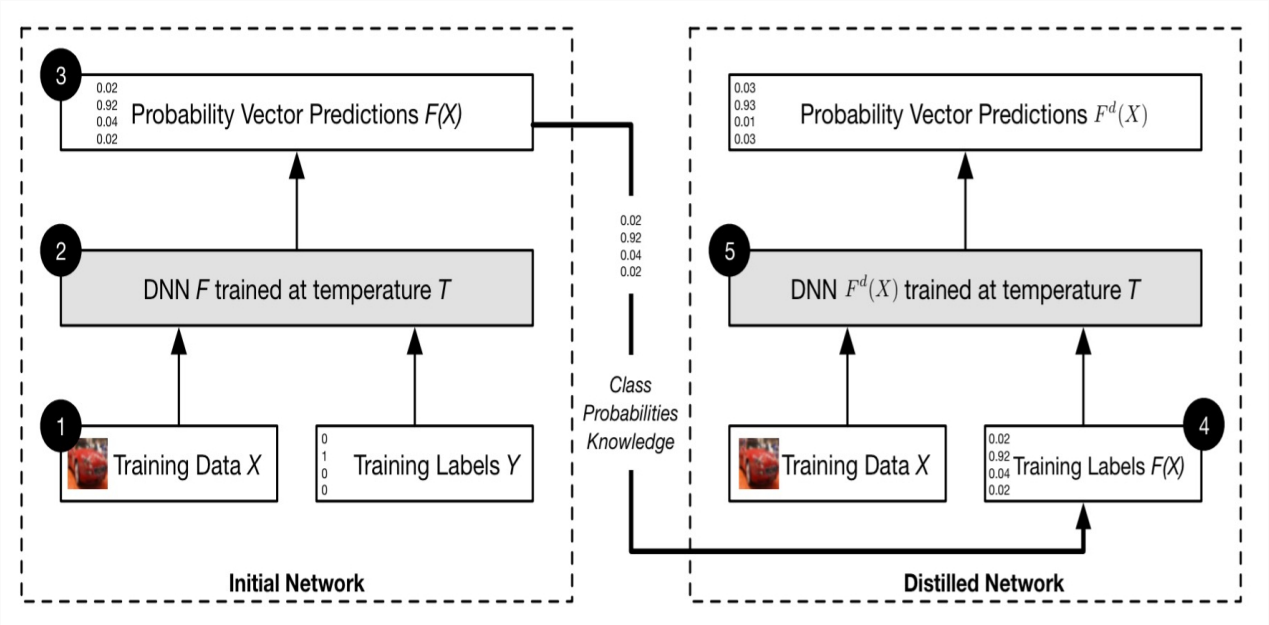
$\boldsymbol{x}$
"panda"
57.7% confidence

$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"nematode"
8.2% confidence

$\boldsymbol{x} + \epsilon\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"gibbon"
99.3 % confidence

A little bit of chemistry

- *Defensive distillation* - N. Papernot, P. McDaniel, X. Wu, S. Jha, A. Swami; 2015
  - training a second network overconfident
  - trained with a smooth decision



- But finally: *Defensive Distillation is Not Robust to Adversarial Examples* - N. Carlini, D. Wagner; 2016

Gently Breaking Neural Networks

Towards Evaluating the Robustness of Neural Networks - N. Carlini, D. Wagner; 2017

- Rethinking the initial optimization problem of adversarial examples
- Defining 3 attacks:
  - L_2 attack. Used by:
    - L_0 attack
    - L_∞ attack

- Low (L_0, L_∞) to none (L_2) perceptible distortion
- Seems to always be able to find an adversarial example (Well, maybe not so gentle...)

Towards an explanation of Adversarial Examples

Is it the end of Deep Learning?

Linearity of Deep Networks

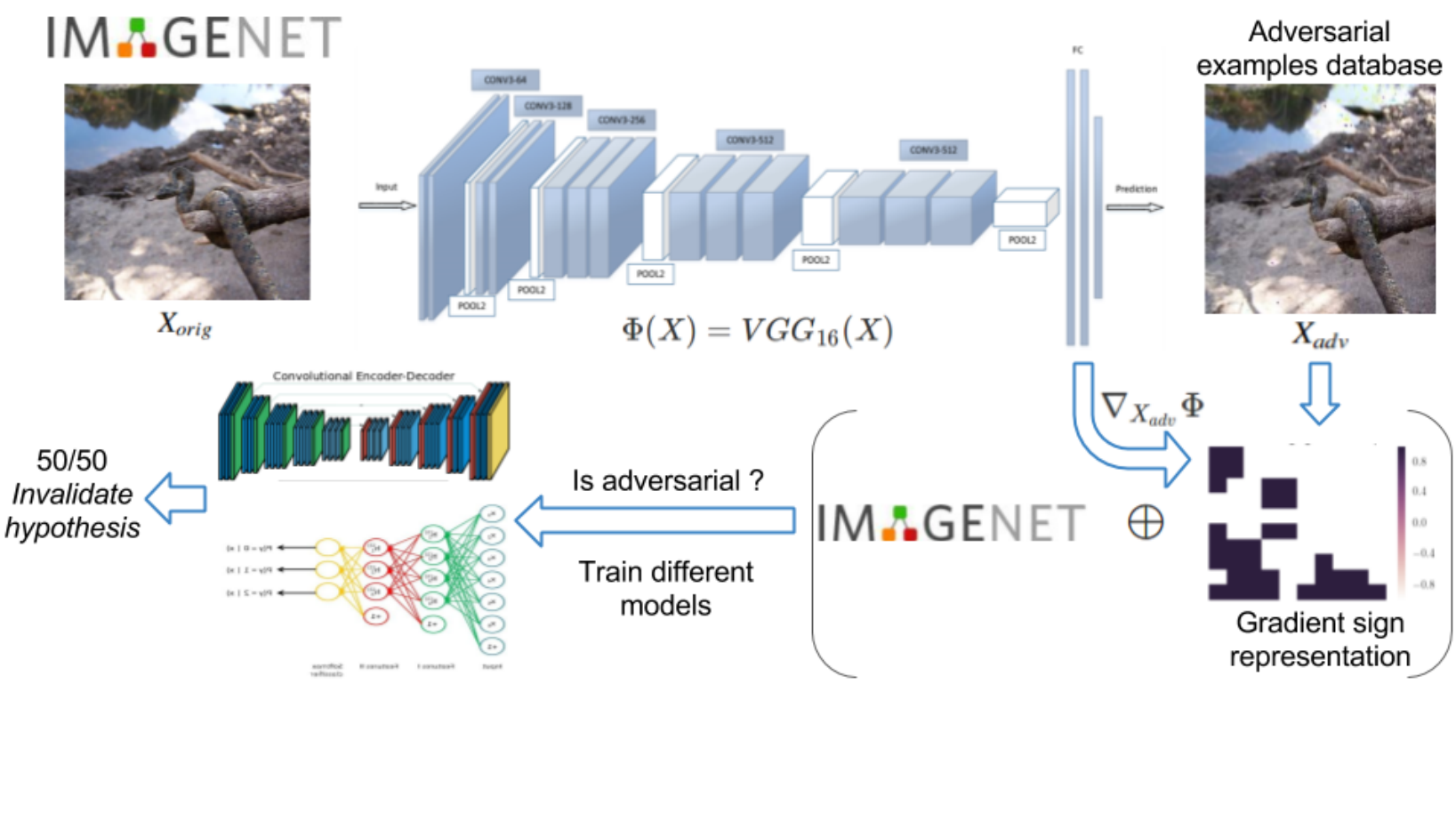
- CNN : convolution + dense + relu = highly linear models
- How to fool a linear classifier in high dimension? *Explaining and harnessing Adv Examples*

$$w^T \hat{x} = w^T x + w^T \eta \tag{1}$$

- w: n dimensions, average magnitude m
- $\hat{x} = x + \eta, \quad \eta = \epsilon sign(w)$
- $\Rightarrow$ Action growth of ϵmn

Gradient based validation?

Our experiment

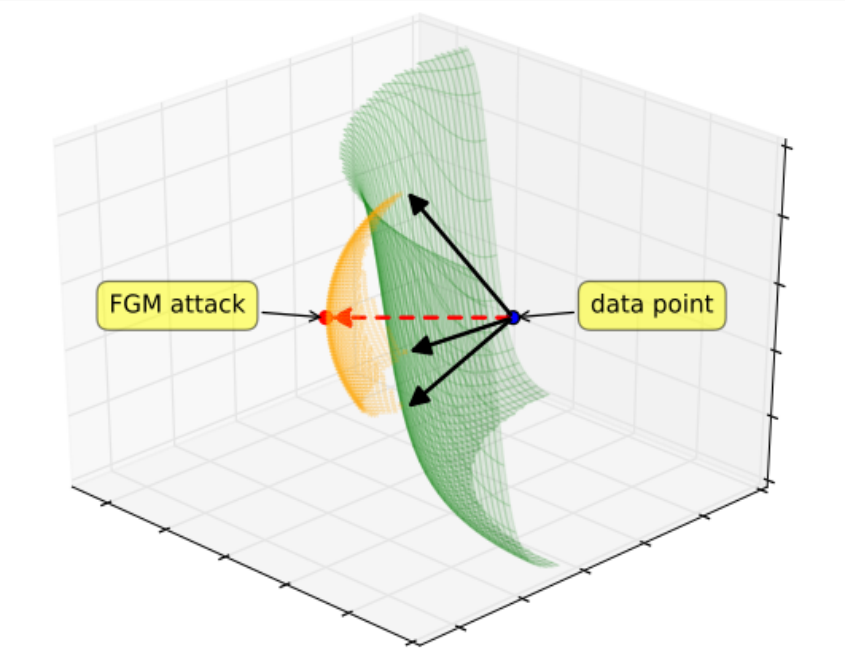


$$\hat{x} = x + \epsilon sign(\nabla_x loss(f_\theta(x), y))$$

Manifold

Recent works tend to explain adversarial examples as examples lying close to the manifold of training data

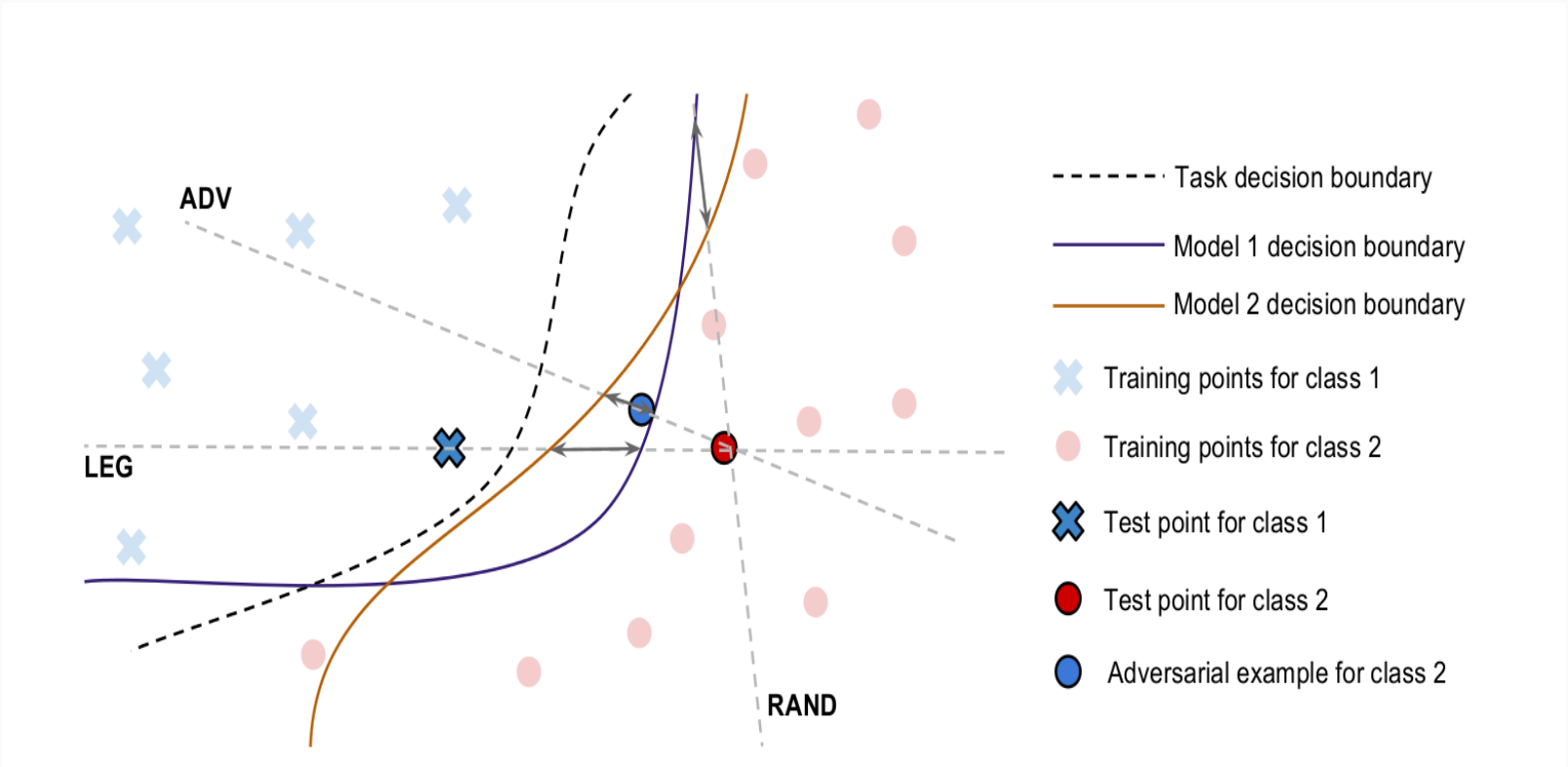
A Boundary Tilting Persepective on the Phenomenon of Adversarial Examples - T. Tanay, L. Griffin; 2016



The Space of Transferable Adversarial Examples

Adversarial space: contiguous, at least 2 dimensional. Dimension is proportional to the ratio increase in loss / perturbation

Different models with similar class boundary distances



Conclusion

- Adversarial examples are caused by a special kind of unnatural noise
- Consider adversarial examples security depending on application

$\rightarrow$ Visit cleverhans

- Adversarial example study gives insights for Neural Network understanding and improvements
- New trend: crafting an adversarial example detector
- Hence we're working on reassigning the original class to an adversarial example

Any question?