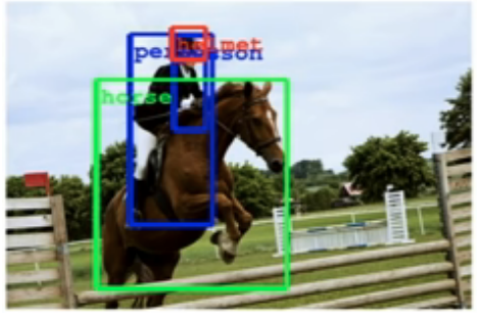
Adversarial Examples and Adversarial Training

Some material taken from CS231n at Stanford

Mauro Sozio

Success of Deep Learning

Since 2012, success of DL at recognizing objects and faces:



reading CAPTCHAS (cannot used them anymore) and addresses

and other tasks...

Can a computer making mistake?

Before 2013 nobody thought that a computer could make a mistake, as Deep Learning was not working that well.

Nowadays, this has become a whole area of research...

Generating Fooling Images / Adversarial example

1. start from an arbitrary image
2. pick an arbitrary class
3. modify the image to maximize the class score (with gradient ascent)
4. repeat until network is fooled

One might think that cats grows horns to become deers, but...

Fooling Images

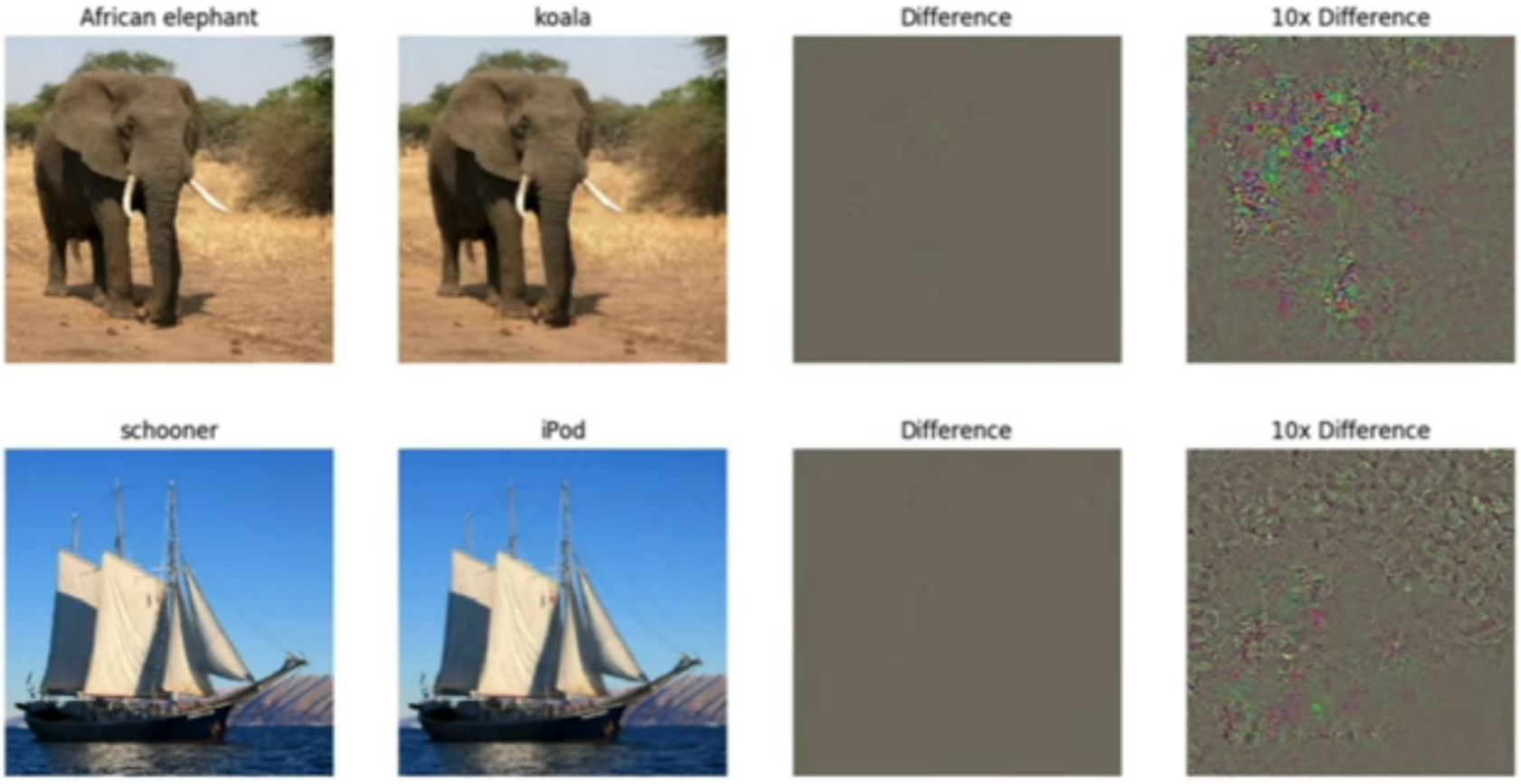


Figure: After very little changes the elephant and the schooner are classified as koala or iPod, respectively.

Fooling Images [1, 4]

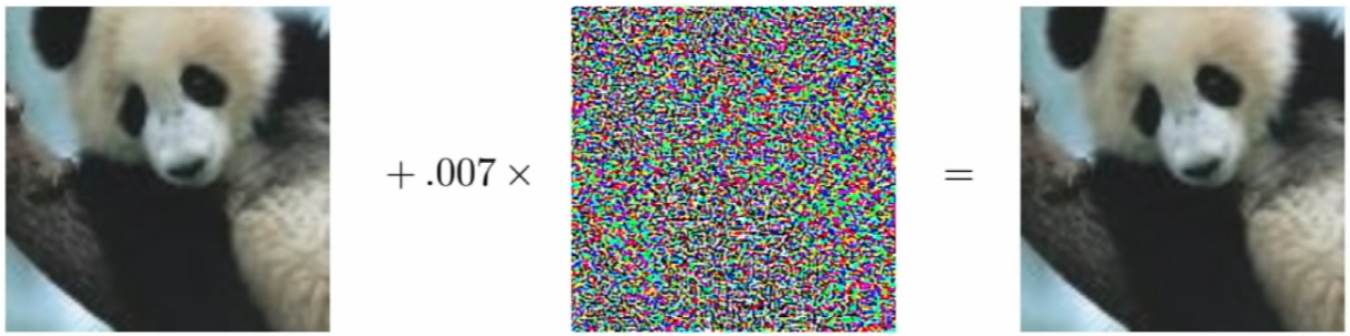


Figure: After the transformation the picture on the right is classified as a gibbon (similar to monkey)

Turning Objects into Airplanes

Figure: The head of the cat is bit darker but apart from that no noticeable difference. Used a CNN.

Problem only for Deep Neural Networks?

Many other machine learning algorithms (e.g. linear models, SVM, k-nn) can be fooled in a similar way.

Radial basis function network (another kind of neural networks) are somehow more robust but the deeper varants are difficult train.

Neural nets can actually become more secure than other models.

Adversarial Examples in the Human Brain

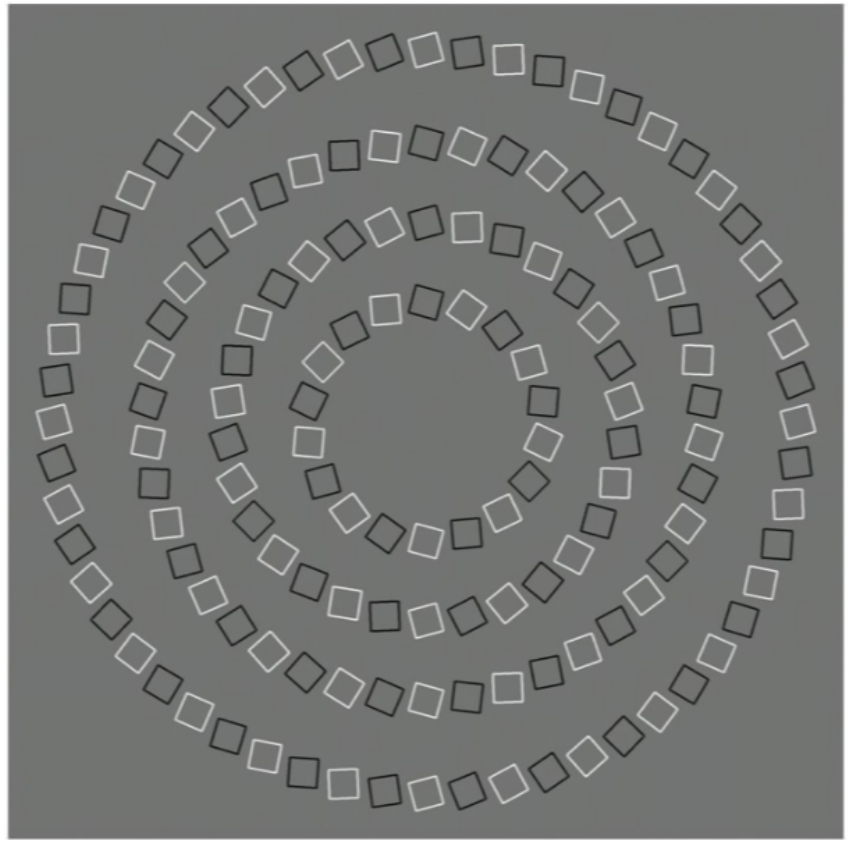


Figure: These are concentric cirlces not intertwined spirals.

Cross-technique Transferability

Papernot [3] showed that one could generate an adversarial example for one machine learning technique and use it to fool another classifier trained with a different technique.

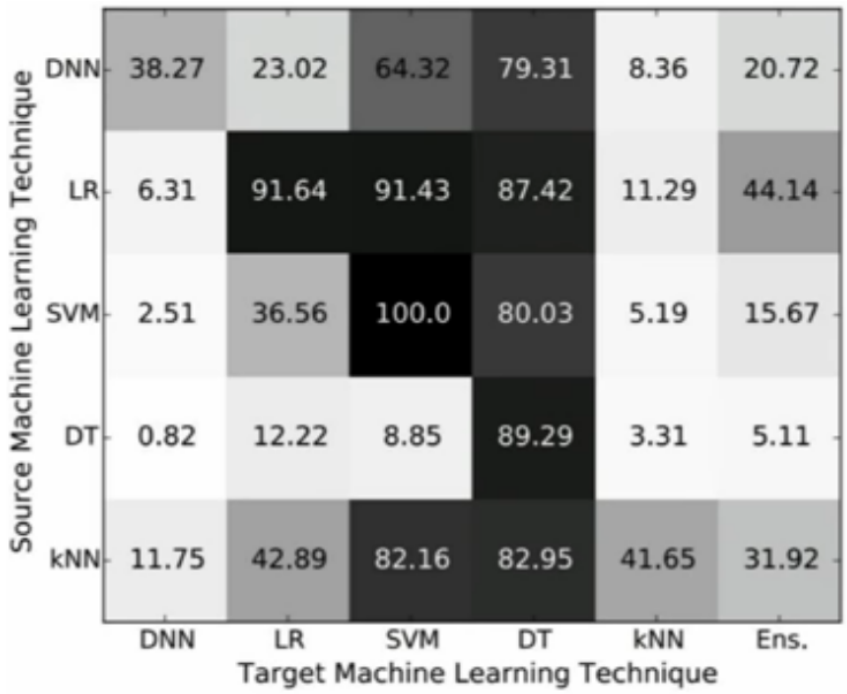


Figure: Cross-technique transferability between deep neural networks, logistic regression, SVM, Decision Trees, k-NN.

Practical Attacks

In order to generate an adversarial example we do not need to access the trained model (e.g. a CNN).

For example the modified picture of a panda (shown before) is classified as gibbon by most machine learning models.

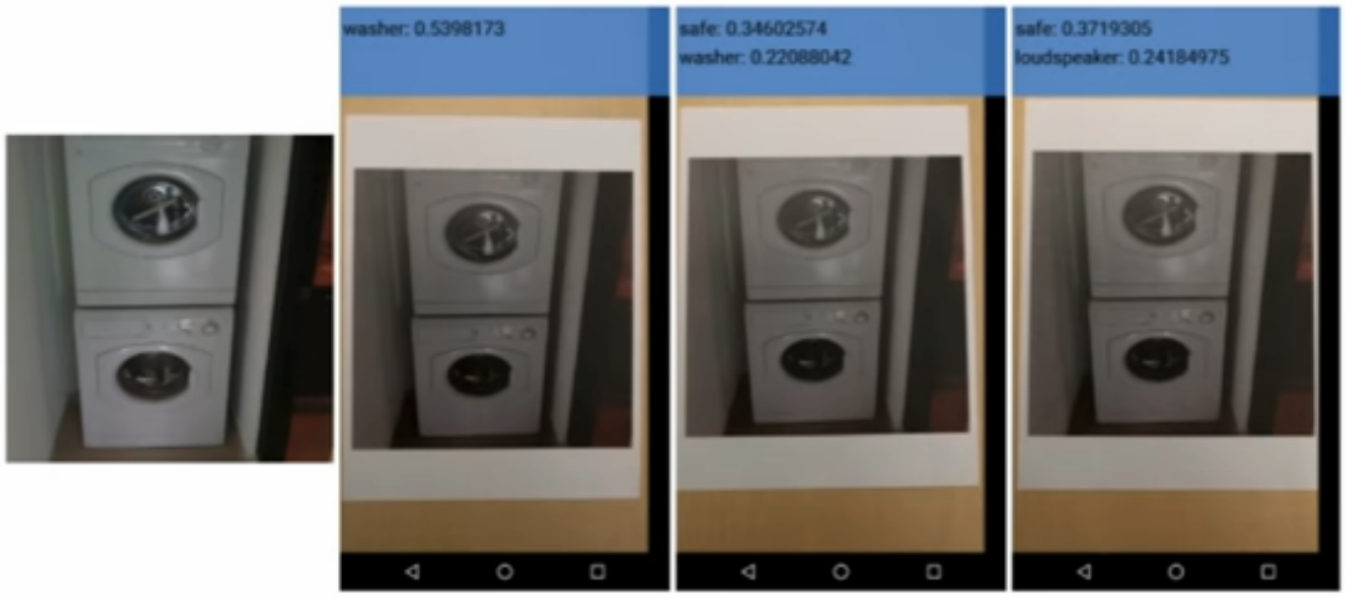
N. Papernot [3], showed that one could use the transfer effect to:

- fool classifiers trained by remotely hosted API (Amazon, Google)
- fool malware detector networks
- display adversarial examples in the physical world and fool machine learning systems that perceive them through the camera.

Adversarial Examples in the Physical World

Fooling an object recognition system running on a smartphone:

- print pictures of adversarial examples
- show them to the object recognition system

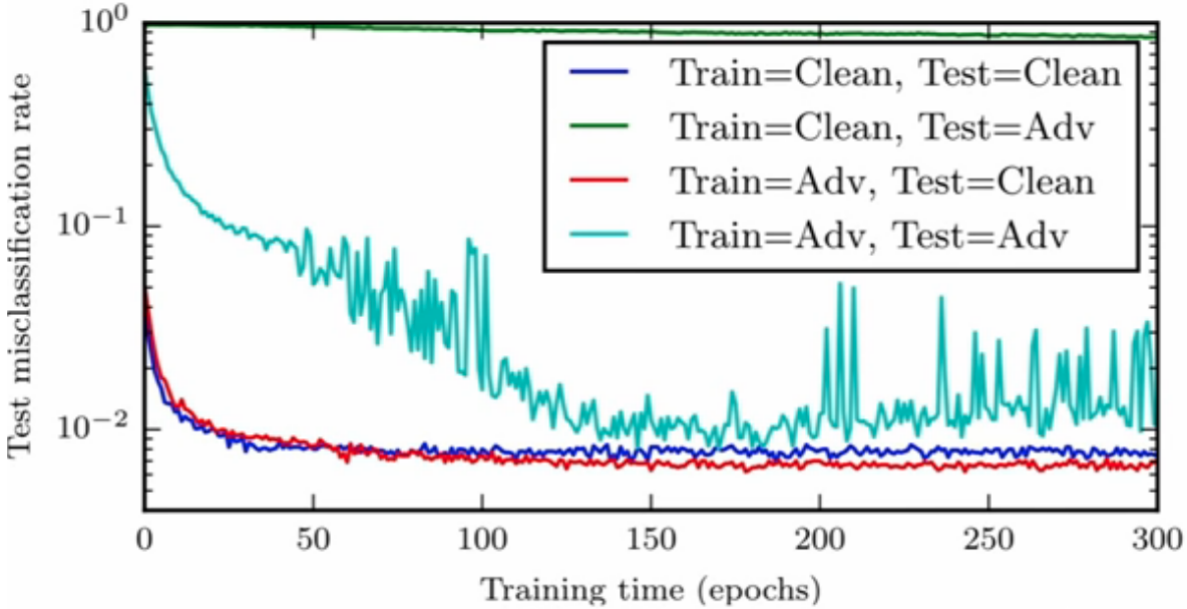


(a) Image from dataset (b) Clean image (c) Adv. image, $\epsilon = 4$ (d) Adv. image, $\epsilon = 8$

Observe: the system running on the smartphone is different than the one used to get the adversarial example [2].

Training on Adversarial Examples

Training on adversarial examples: for every image add an adversarial image (specify it is the same class).



It helps (compare the green curve with the light blue curve): you can resist the same attack you trained on.

Conclusion

- Attacking is easy
- defending is difficult
- adversarial training helps resisting the same attack we trained on. But not other attacks (i.e. other algorithms/ways of modifying the image).
- defending is a whole active area of research. Studying how to resists attacks might lead to better models.
- Further reading: Ian GoodFellow, 2019 [1].

References I

[1] I. J. Goodfellow.
A research agenda: Dynamic models to defend against correlated attacks.
CoRR, abs/1903.06293, 2019.

[2] A. Kurakin, I. J. Goodfellow, and S. Bengio.
Adversarial examples in the physical world.
In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*, 2017.

[3] P. D. McDaniel, N. Papernot, and Z. B. Celik.
Machine learning in adversarial settings.
IEEE Security & Privacy, 14(3):68–72, 2016.

[4] N. Papernot, P. D. McDaniel, I. J. Goodfellow, S. Jha, Z. B. Celik, and A. Swami.
Practical black-box attacks against deep learning systems using adversarial examples.
CoRR, abs/1602.02697, 2016.