ICFIP 2022
5th Int. Conference on Frontiers of Image Processing

Adversarial examples: 10 years later

Mauro Barni
University of Siena

Outline

- A not-so-recent history
- Another effect of the curse of dimensionality
- What's so special with DL?
- Do we need to panic?

The big-bang: everything started with [1]

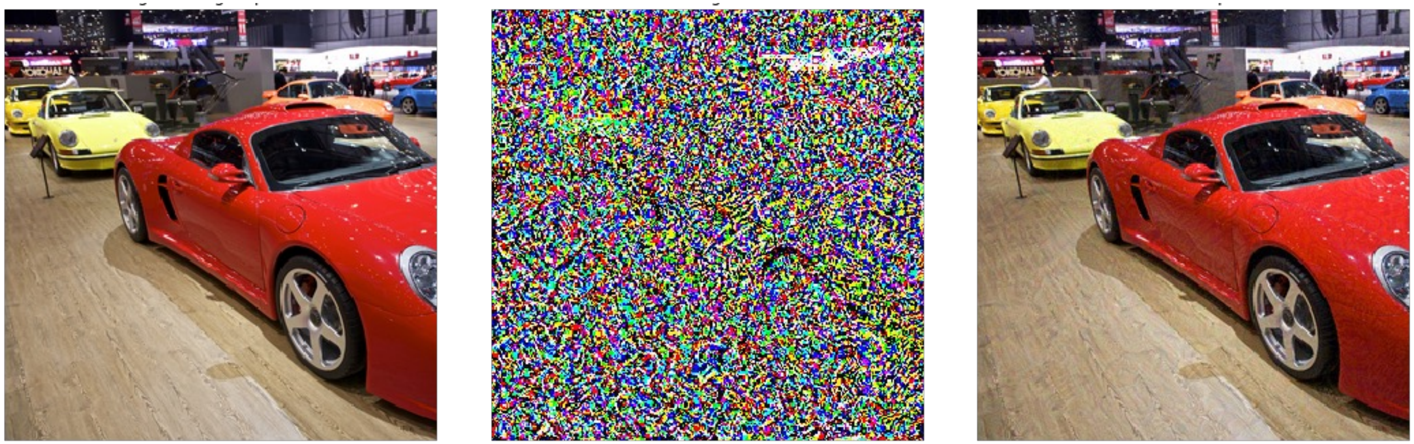
[1] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, R. Fergus (2013). Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.

Concern turned into panic when transferability of adversarial examples was proven [2]

[2] N. Papernot, P. McDaniel, I. Goodfellow. "Transferability in machine learning: from phenomena to black-box attacks using adversarial samples." arXiv preprint arXiv:1605.07277 (2016).

Since then …

Magnified noise



Classified as a *toaster*

Classified as a *Gibbon*

Striking examples: one pixel attack

Not only digital

Not only digital

A not-so-recent history

[1] M. Barreno, B. Nelson, A. D. Joseph, J. D. Tygar, “The security of machine learning”, Mach Learn 81, pp. 121–148, 2010.

[2] N. Dalvi, P. Domingos, P.Mausam, S. Sanghai, D. Verma, ”Adversarial classification”. Proc. ACM SIGKDD, 2004.

[3] D. Lowd and C. Meek, “Adversarial learning” in Proc. of the ACM SIGKDD Conf. 641-647, 2005.

[4] B. Biggio, et al. "Evasion attacks against machine learning at test time." Joint European conf. machine learning and knowledge discovery in databases. Springer, Berlin, Heidelberg, 2013.

[5] B. Biggio, F. Roli, (2018). Wild patterns: Ten years after the rise of adversarial machine learning. Pattern Recognition, (84).

and previous similar results in watermarking, biometrics, adversarial multimedia forensics …

A not-so-recent history

- Yet the alarm raised only with the rise of deep learning
- Why? What's special with deep learning?
  - Popularity and importance of Deep Learning
  - Not only

Setting

Focus on

- Classification networks
- White box (perfect knowledge) attacks
- Non-targeted attacks
  - Extension to targeted attacks possible (non-trivial)
  - No distinction in the binary case

- Goal: Answer the question:

  Is there a special relationship between DL and the existence of adversarial examples?

The linear explanation*

$$f(x) = \text{Tresh}\big(\phi(x), T\big) \qquad \phi(x) = \sum_{i=1}^{n} w_i x_i \qquad \phi(x_0) = T - \Delta$$

$$\phi(x_0 + z) = \sum w_i x_{0,i} + \sum w_i z_i$$

Assume an *mse*-bounded perturbation

$$\frac{\sum z_i^2}{n} \leq \gamma^2$$

Similar results hold for the infinity norm (with some noticeable differences)

* I. Goodfellow, J. Shlens, C. Szegedy "Explaining and harnessing adversarial examples" *arXiv preprint arXiv:1412.6572* (2014).

The linear explanation

Random perturbation

$$z_i = \gamma \cdot \mathcal{N}(0,1)$$

$$E[\phi(x_0 + z)] = E[\sum_i w_i x_{0,i}] + E[\sum_i w_i z_i] = \phi(x_0)$$

$$var[\phi(x_0 + z)] = var[\sum_i w_i z_i] = \gamma^2 \|w\|^2$$

For the attack to succeed with non-negligible **probability** we must have

$$\gamma > \frac{k\Delta}{\|w\|}$$

The linear explanation

Adversarial perturbation

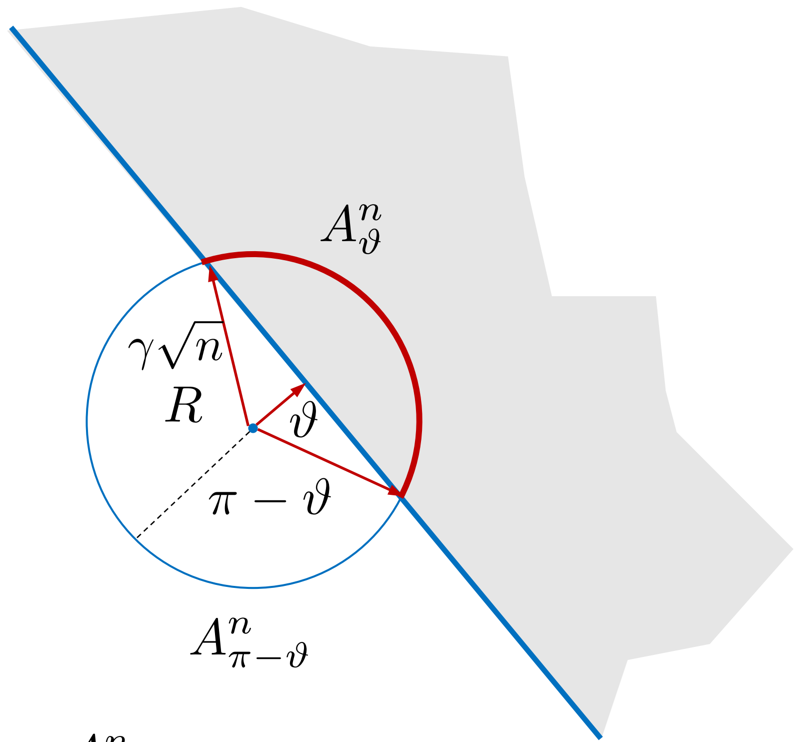
$$z = \gamma\sqrt{n} \cdot e_w$$

$$\phi(x_0 + z) = \phi(x_0) + \gamma\sqrt{n}\sum_i w_i e_{w,i} = \phi(x_0) + \gamma\sqrt{n}\|w\|$$

For the attack to succeed we must have

$$\gamma > \frac{\Delta}{\sqrt{n}\|w\|}$$

A geometric interpretation



$$\lim_{n\to\infty} \frac{A^n_\vartheta}{A^n_{\pi-\vartheta}} = 0$$

- In very high dimensional spaces. the *number* of directions resulting in a successful attack is very small
- This explains why adversarial examples do not show up in non-adversarial settings

Does it have to be linear?

- Same arguments hold if the decision function is smooth enough
- Local linearity assumption

$$\phi(x_0 + z) = \phi(x_0) + \langle \nabla_\phi(x_0), z \rangle$$

- The attacker needs only to align the attack to the gradient

$$z = \gamma\sqrt{n} \cdot e_\phi$$

$$e_\phi = \frac{\nabla_\phi(x_0)}{\|\nabla_\phi(x_0)\|}$$

$$\gamma > \frac{\Delta}{\sqrt{n}\|\nabla_\phi\|}$$

It doesn’t even need to be nearly linear

The attackability of any network can be explained by the concentration property of measure (or probability)

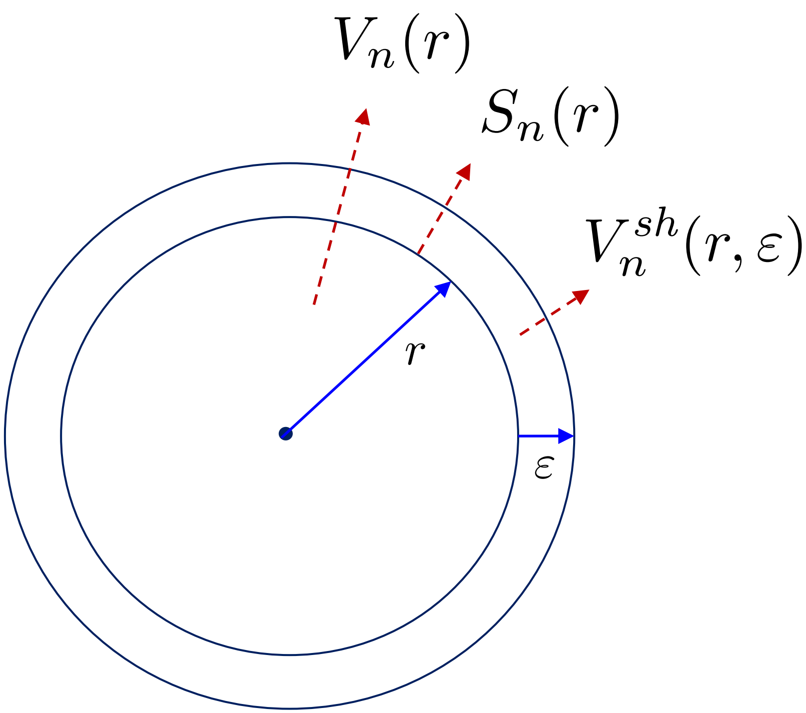
Roughly speaking it says that

«For any measurable set in R^n, most of the volume is (arbitrarily) close to the boundary of the set»

We’ll see this for hyperspheres

It doesn't even need to be nearly linear

Volume of a hypersphere of radius *r* :



$$V_n(r) = \frac{\pi^{n/2}}{\Gamma(n/2+1)} r^n$$

$$S_n(r) = \frac{2\pi^{n/2}}{\Gamma(n/2)} r^{n-1}$$

$$V^n(r) = \frac{r}{n} S_n(r)$$

$$V_n^{sh}(r,\varepsilon) \approx S_n(r) \cdot \varepsilon$$

It doesn't even need to be nearly linear

$$\begin{aligned}\frac{V_n(r+\varepsilon)}{V_n(r)} &= \frac{V_n(r)+S_n(r)\varepsilon}{V_n(r)}\\ &= 1+\frac{\frac{n\varepsilon}{r}V_n(r)}{V_n(r)}\\ &= 1+\frac{n\varepsilon}{r}\\ &= \infty \text{ when } n\to\infty\end{aligned}$$

Most of the points are within ε of the boundary

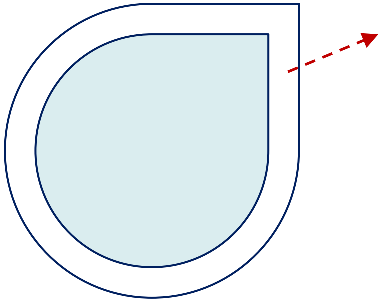
It doesn't even need to be nearly linear

For an *mse*-bounded perturbation we have:

$$\frac{\|\varepsilon\|^2}{n} \leq \gamma^2 \implies \|\varepsilon\| \leq \sqrt{n}\,\gamma$$

Not only most points are within ε of the boundary, ε also increases with *n*

By the isoperimetric inequality the above argument can be extended to any smooth enough set



Most of the volume is within ε of the boundary

Within a hypercube

- Mof the points within a hypersphere can be moved outside with minimal effort, the inverse is not true due to the unboundedness of R^n
- Images live in a bounded space -> the $[0,1]^n$ hypercube
- For any 2-set partition of the hypercube (big *n*) with a non-negligible volume assigned to both sets, it is always possible to move a point from one set to the other with minimal effort (bounded mse) [1]

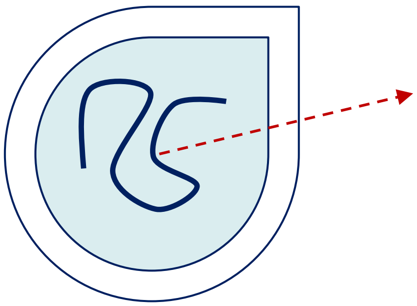
- A binary classifier is nothing but a way to partition the hypercube
- Do adversarial examples exist for **ALL CLASSIFIERS** (including the human brain)?

[1] A. Shafahi, W. R. Huang, C. Studer, S. Feizi, T. Goldstein, «Are adversarial examples inevitable?», In International Conference on Learning Representations (2018).

Are adversarial examples unavoidable?

- Some major issues still to be investigated

- The theory does not generalize well to infinity norm
- What about multiple classifiers and targeted attacks?
- Most of the images are meaningless



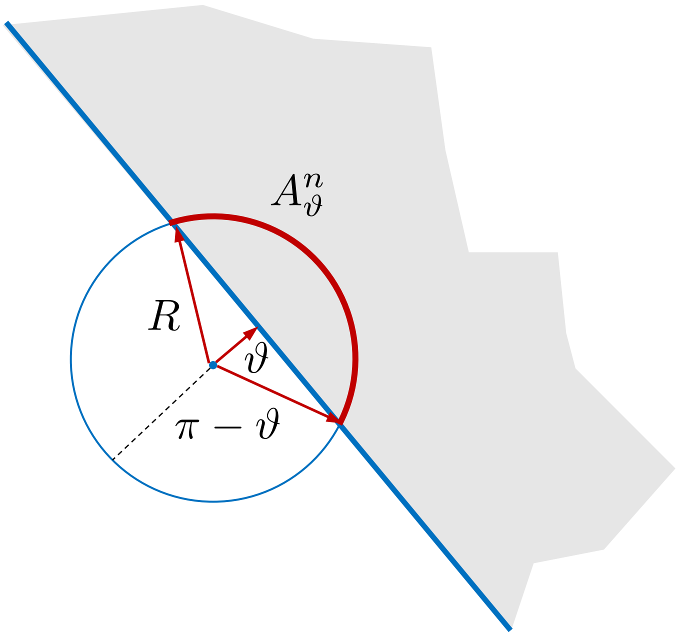
Good images could live in a manifold deep inside the classification regions

It is a fact, that all defenses proposed so far have been defeated with a limited effort …

Then, what's special with DL?

- Existence of adversarial examples **does not mean they are easy to find**

- **For smooth decision functions you need to align the attack to the direction of the gradient**
- **Backpropagation provides an efficient way to compute the gradient … then**
- **DL architectures are extremely susceptible to gradient-based attacks**



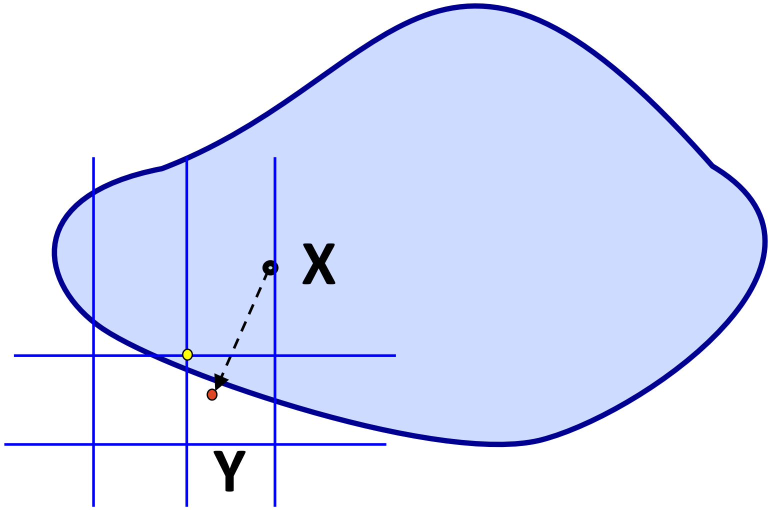
Should we panic?

- **Turning adversarial examples into real-life threats is not an easy task**
- Three major difficulties

Robustness against postprocessing

- Attacks themselves should resist to post-processing, like integer quantization or JPEG compression
- Attacked images are sometimes classified correctly after (moderate) JPEG compression*

* N. Das, et al. "Shield: Fast, practical defense and vaccination for deep learning using JPEG compression" Proc. 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 196-204. ACM, 2018.

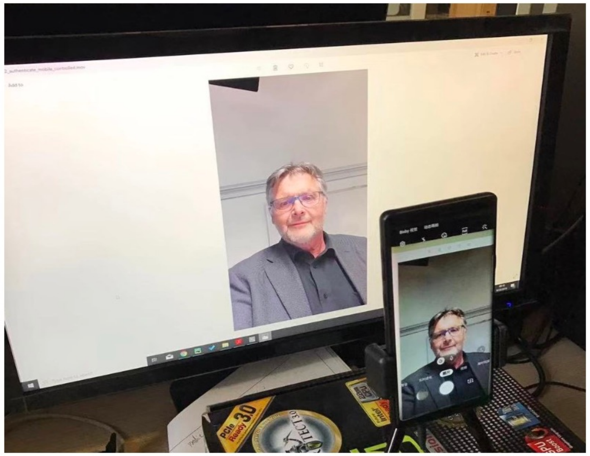


Attacks in real world

- Carrying out the attack in the real world (analog domain) is even more challenging(still possible)

- Visible distortion
- Unattended systems

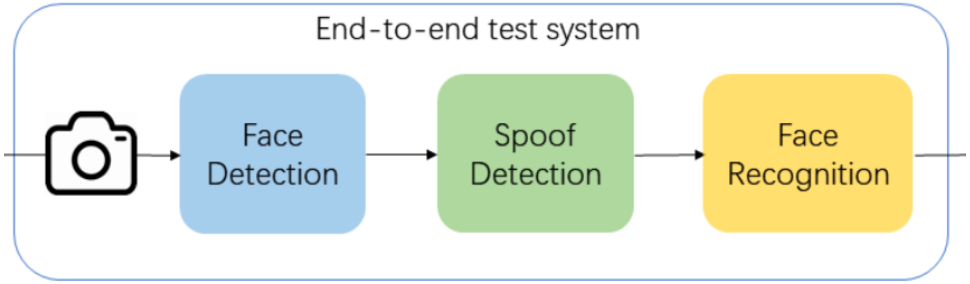
Sometimes is even more diffcult



Attack against a spoofing detector

Preemptive attack compensating for rebroadcast artefacts

End-to-end attack necessary



* Zhang, B., Tondi, B., & Barni, M. (2020). Adversarial examples for replay attacks against CNN-based face recognition with anti-spoofing capability. *Computer Vision and Image Understanding*, *197*, 102988.

Attacks with limited knowledge (LK)

- The most common approach consists in attacking a **surrogate detector** (attack transferability)

$$\hat{\phi} = \hat{\phi}(\hat{\mathcal{L}}, \hat{\mathcal{W}}; \hat{\mathcal{D}})$$

- … and cross your fingers
- No guarantee that the attack works

How to impove transferability

- Input diversity [1]
- Increased confidence [2]

- Distortion increases and transferability is not always easy to achieve
- Mismatch between the target system and the surrogate detector may be significant

[1] Xie C., Zhang Z., Zhou Y., Bai S., Wang J., Ren Z., Yuille A.L.: Improving transferability of adversarial examples with input diversity. CVPR, 2019.

[2] Li, W., Tondi, B., Ni, R., & Barni, M. "Increased-Confidence Adversarial Examples for Deep Learning Counter-Forensics." *Int. Conference on Pattern Recognition*. Springer, Cham, 2021.

In summary

- The ubiquitous existence of adversarial examples raises interesting questions on DNN (and not only) security
- Devising defenses under strong threat models (like in a white box setting) is extremely difficult

- The situation may not be as bad as one could think
- Attackers have their own problems to turn adversarial examples into real world threats

Thank you
for your attention