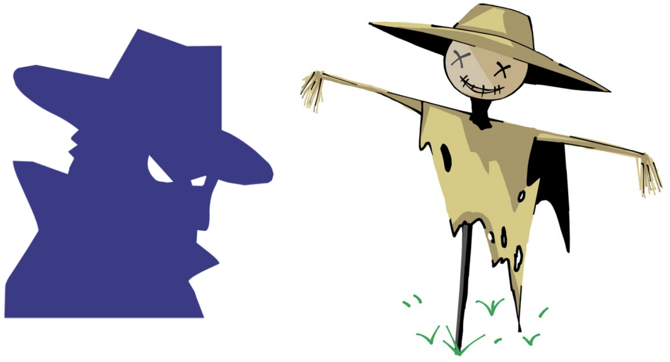
WIFS 2022
14-th Int. Workshop Information Forensics and Security

Adversarial examples: threat or scarecrow

Mauro Barni
University of Siena



Outline

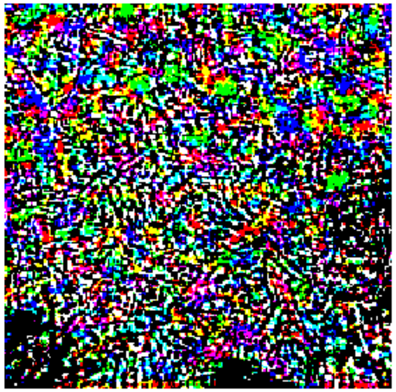
- The threat
- Just another effect of the curse of dimensionality?
- What’s so special with DL?
- Threat or scarecrow
- Looking ahead

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*.

«We find that deep neural networks learn input-output mappings that are fairly discontinuous to a significant extent. 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»*

Since then …

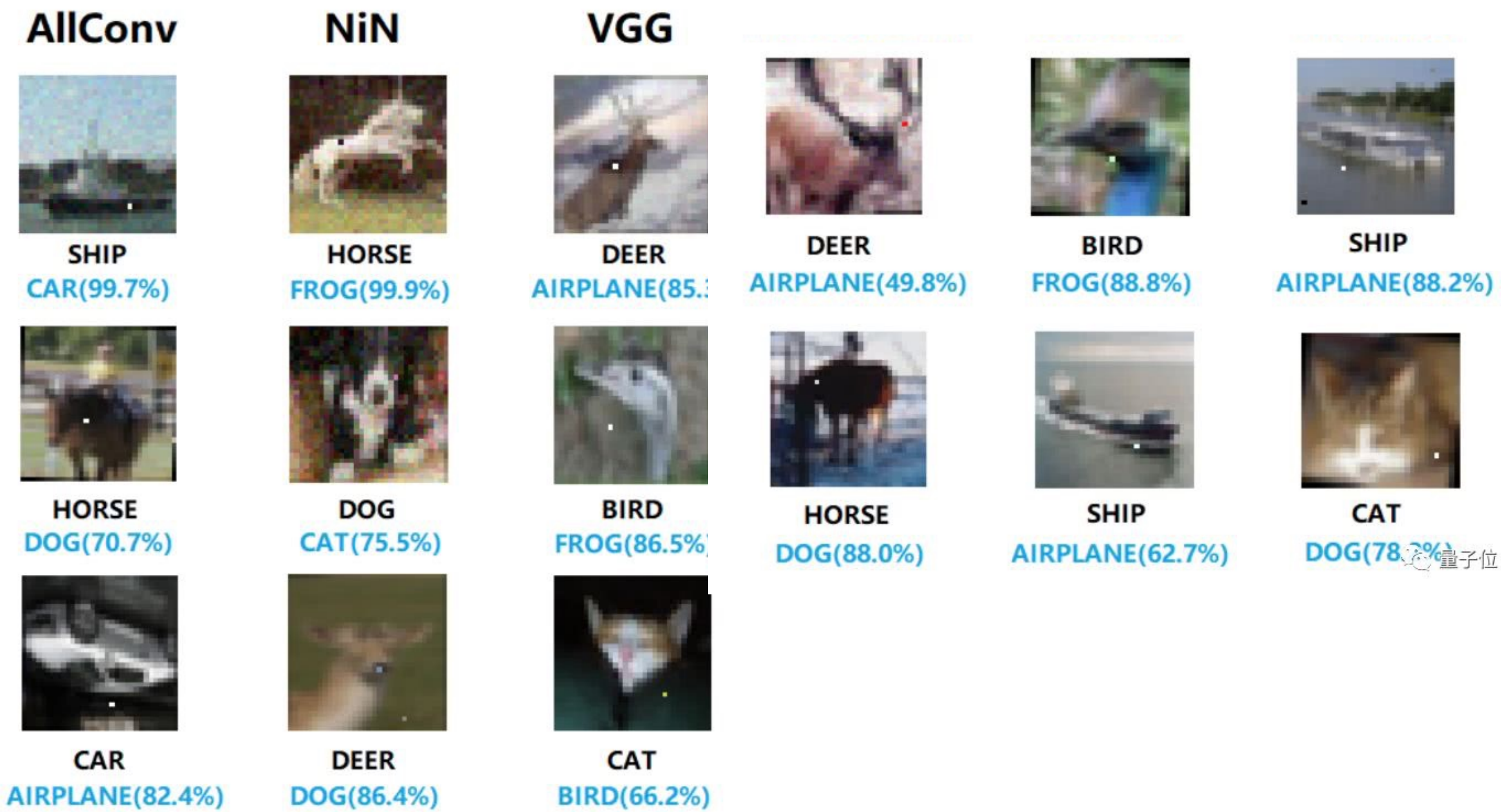

Classified as a *cat*

Highly magnified attack



Classified as a *dog*

Striking examples: one pixel attack



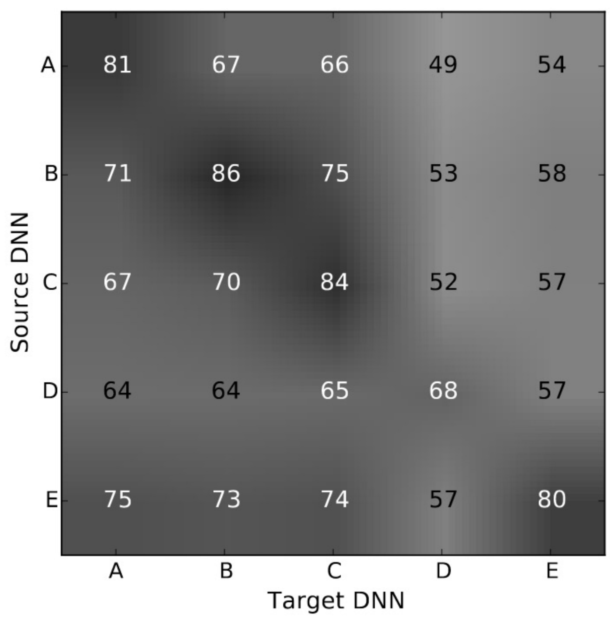
Not only digital

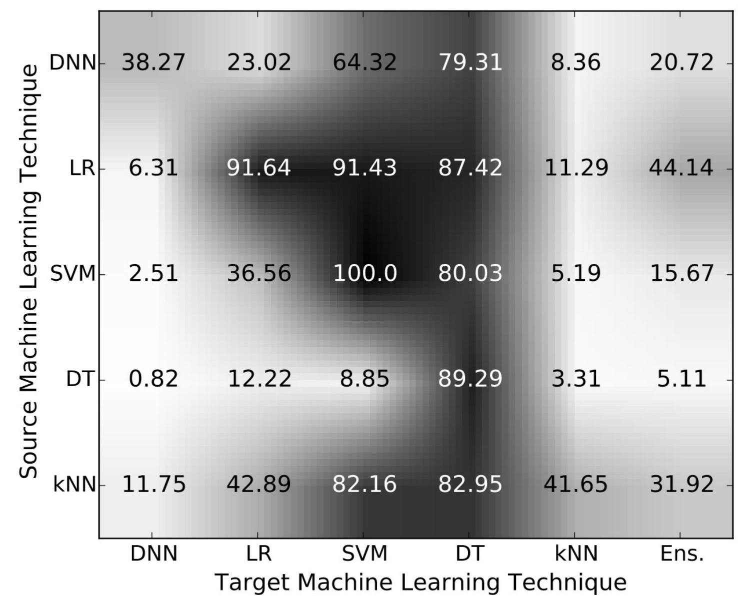
Not only digital

Attacks transferability

- Concerns turned into panic when (a certain degree of) transferability of adversarial examples was proven [1]

[1] 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).





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

- White box (perfect knowledge) attacks
- (Binary) classification networks
- Non-targeted attacks
  - Extension to targeted attacks is 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

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

* 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

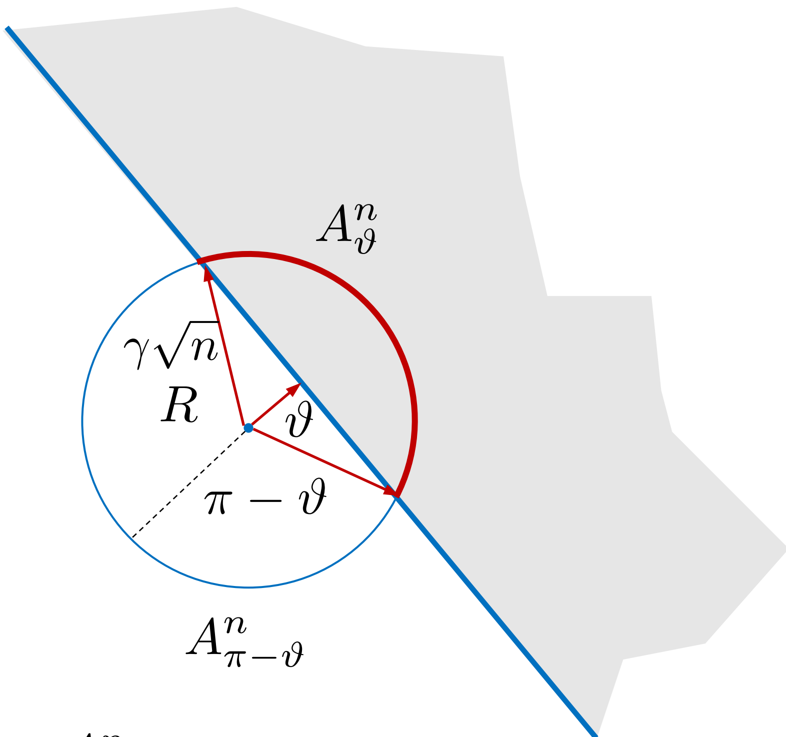
$$z = \gamma\sqrt{n} \cdot e_w$$

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

For the attack to succeed we must have

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

A geometric interpretation



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

- 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

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

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

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

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).

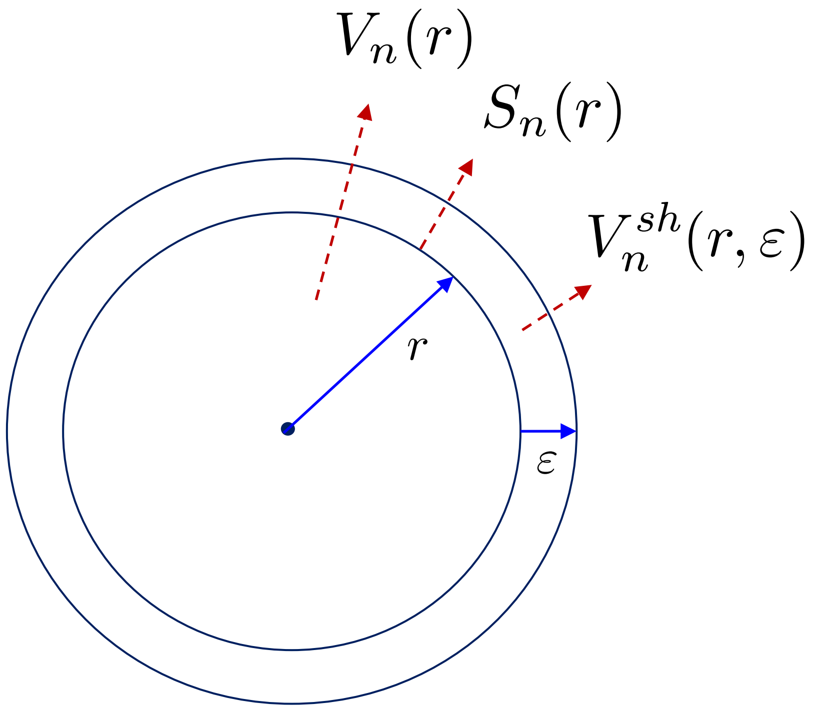
Roughly speaking it says that

«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

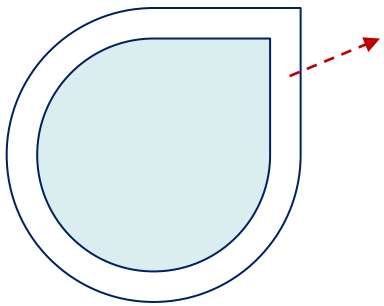
It doesn't even need to be nearly linear

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

- Most of 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]

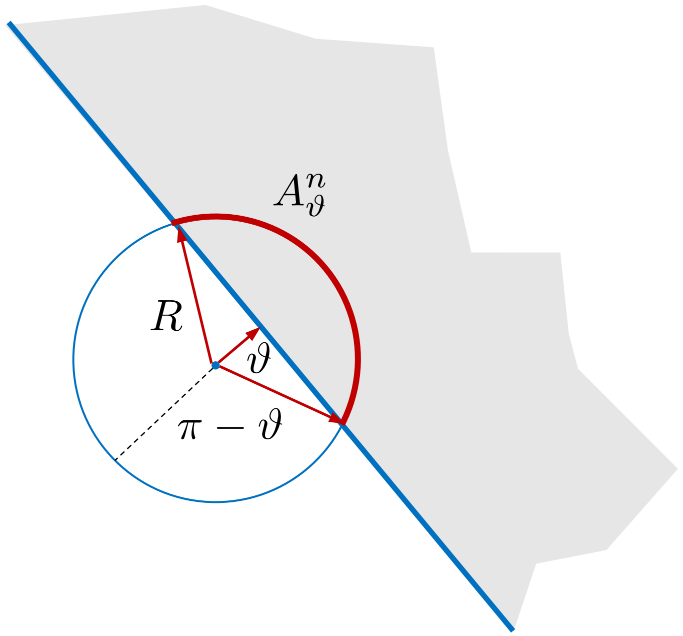
- A binary classifier is nothing but a way to partition the hypercube
- Do adversarial examples exist for **ALL BINARY CLASSIFIERS** (including the human brain)?

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

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**

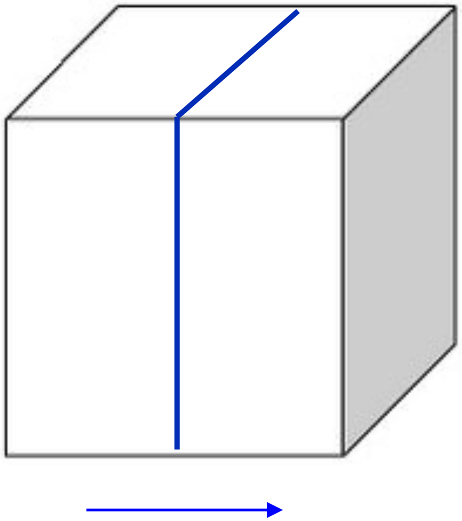


Should we panic? Not necessarily

- Further theoretical investigation needed
- **Turning adversarial examples into real-life threats is not an easy task**
- Three major difficulties
  - Robustness
  - Lack of knowledge
  - Physical domain attacks

Theoretical difficulties (1): infinity norm

- The theory does not generalize well to infinity norm



If the partition is aligned to one (few) dimension only, the perturbation collapses into one dimension and infinity-norm bounded adversarial perturbations may not exist

Curse of dimensionality does not apply

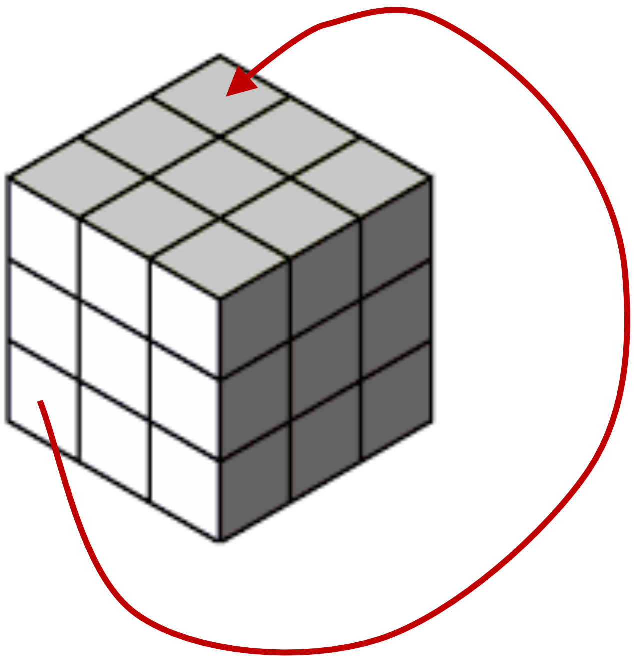
Should classifiers focus on few image pixels? Very likely they won't

Theoretical difficulties (2): targeted attacks

- Turning an arbitrary source class into an arbitrary target class may not always be possible
- What about multilabel classifiers?

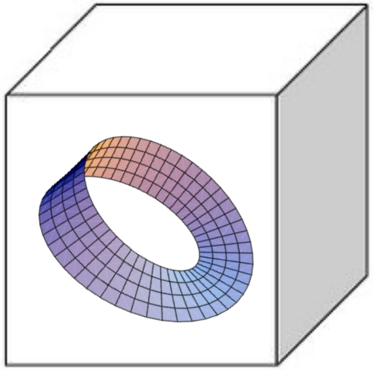
Children playing footbal on the grass

Young people drinking bier on a beach



(3) Natural images do not *live* in hypercubes

- Image distribution is not uniform in hypercube
  - try generating an image at random with iid pixels uniformely distributed in [0,1] !!!



- Images likely live in thin neighborhoods of low dimensional manifolds

- Does theory generalize to manifolds? Is the size (and topology) of image manifolds large enough to trigger the large-dimensionality effects?

(3) Natural images do not *live* in hypercubes

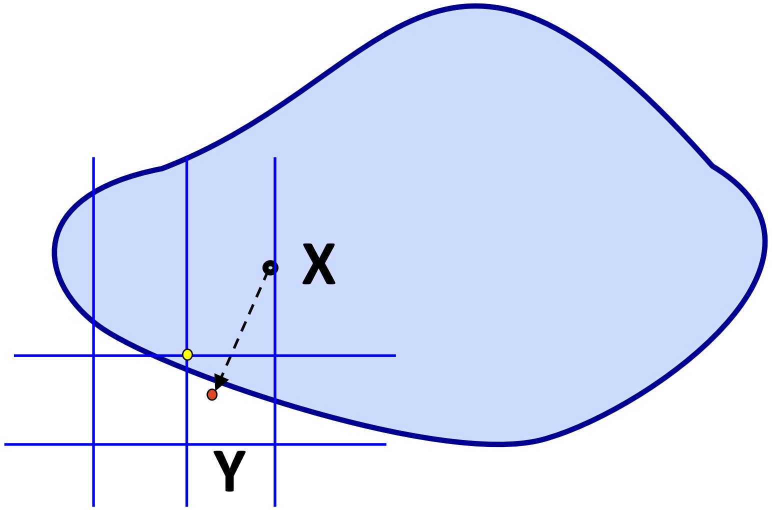
- Image distribution is not uniform in hypercube
  - try generating an image at random with iid pixels uniformely distributed in [0,1] !!!

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

- Does theory generalize to manifolds? Is the size (and topology) of image manifolds large enough to trigger the large-dimensionality effects?

Robustness against postprocessing

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



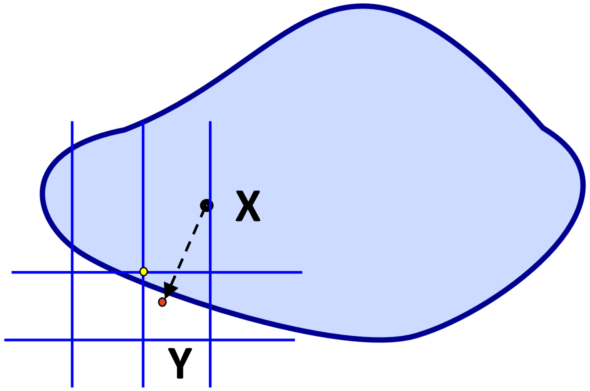
* N. Das, et al. "Shield: Fast, practical defense and vaccination for deep learning using JPEG compression" Proc. 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 196-204. ACM, 2018.

The case of quantization

- Often attacks implemented in Foolbox result in extremely high PSNR (e.g., 60dBs)
- After quantization to integers the attack disappears

$$10 \log_{10} \frac{255^2}{MSE} = 60 \implies MSE \approx 0.06$$

- Perturbation in the order to 0.25, hence removed by integer quantization
- Specific attacks needed*



* Tondi, B. (2018). Pixel-domain adversarial examples against CNN-based manipulation detectors. Electronics Letters, 54(21), 1220-1222.

The battle of knowledge

If you know the enemy and know yourself, you need not fear the result of a hundred battles

If you know the enemy and know yourself, you need not fear the result of a hundred battles

Limited knowledge attacks

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

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

- To account for mismatch in training data and architecture a stronger attack must be applied

Examples:

- 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).

Attacks with limited knowledge (LK)

Attack transferability is not always easy to achieve. For instance, it turns out to be particularly difficult in MMF applications*

* Barni, M., Kallas, K., Nowroozi, E., & Tondi, B. (2019). On the transferability of adversarial examples against CNN-based image forensics. *IEEE Int. Conference on Acoustics, Speech and Signal Processing (ICASSP)*

Example of *Cross-model* transferability

CROSS MODEL								
SN	TN	Accuracy w/o attack	attack	avg. PSNR	avg. L1 dist	avg. max. dist	attack success rate on SN	attack success rate on TN
$N_{\mathrm{BS}}^{\mathrm{R}}$(res)	$N_{\mathrm{GC}}^{\mathrm{R}}$(res)	SN= 97.60%, TN= 98.20%	I-FGSM, $\varepsilon_s = 0.01$	40.02	2.53	2.55	1.0000	0.0020
$N_{\mathrm{BS}}^{\mathrm{R}}$(res)	$N_{\mathrm{GC}}^{\mathrm{R}}$(res)	SN= 97.60%, TN= 98.20%	I-FGSM, $\varepsilon_s = 0.001$	58.48	0.31	0.33	1.0000	0.0020
$N_{\mathrm{BS}}^{\mathrm{R}}$(res)	$N_{\mathrm{GC}}^{\mathrm{R}}$(res)	SN= 97.60%, TN= 98.20%	JSMA, $\theta = 0.1$	46.09	0.07	57.88	1.0000	0.0164
$N_{\mathrm{BS}}^{\mathrm{R}}$(res)	$N_{\mathrm{GC}}^{\mathrm{R}}$(res)	SN= 97.60%, TN= 98.20%	JSMA, $\theta = 0.01$	54.98	0.04	15.14	0.9918	0.0061
$N_{\mathrm{BS}}^{\mathrm{R}}$(med)	$N_{\mathrm{GC}}^{\mathrm{R}}$(med)	SN= 98.20%, TN= 100%	I-FGSM, $\varepsilon_s = 0.01$	40.03	2.53	2.55	**1.0000**	**0.8248**
$N_{\mathrm{BS}}^{\mathrm{R}}$(med)	$N_{\mathrm{GC}}^{\mathrm{R}}$(med)	SN= 98.20%, TN= 100%	I-FGSM, $\varepsilon_s = 0.001$	59.67	0.26	0.27	1.0000	0.1813
$N_{\mathrm{BS}}^{\mathrm{R}}$(med)	$N_{\mathrm{GC}}^{\mathrm{R}}$(med)	SN= 98.20%, TN= 100%	JSMA, $\theta = 0.1$	49.64	0.03	38.11	1.0000	0.0102
$N_{\mathrm{BS}}^{\mathrm{R}}$(med)	$N_{\mathrm{GC}}^{\mathrm{R}}$(med)	SN= 98.20%, TN= 100%	JSMA, $\theta = 0.01$	58.47	0.02	14.05	0.9837	0.0163

Res: resizing detection

Med: median filtering detection

BS: Bayar-Stamm CNN with preprocessing

GC: Barni's net without preprocessing

R: Training on Raise2K

V: TraiXning on Vision dataset

How to impove transferability

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

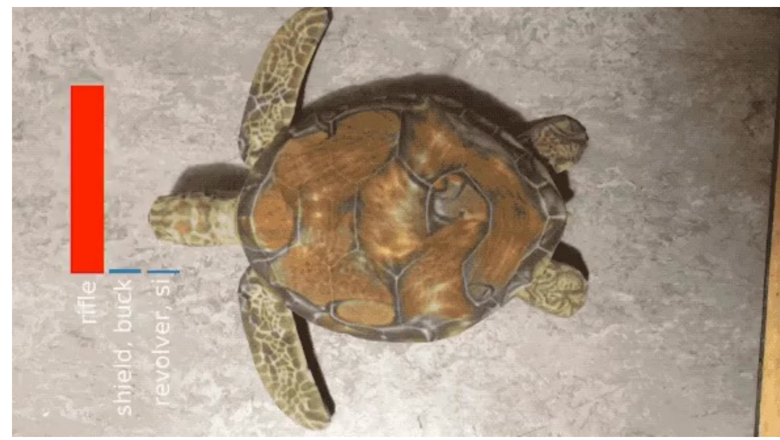
- Distortion increases and transferability is not always easy to achieve
- Mismatch between the target system and the surrogate detector may be significant

[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.

Attacks in the real world

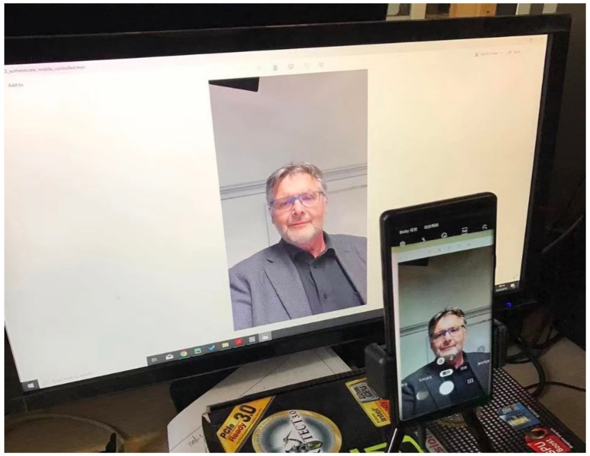
- Carrying out the attack in the physical domain is even more challenging, but still possible



- Expectation over transformation (EOT)

$$\rho^* = \arg\min_{\rho} E_T[\Phi(T(I+\rho))]$$

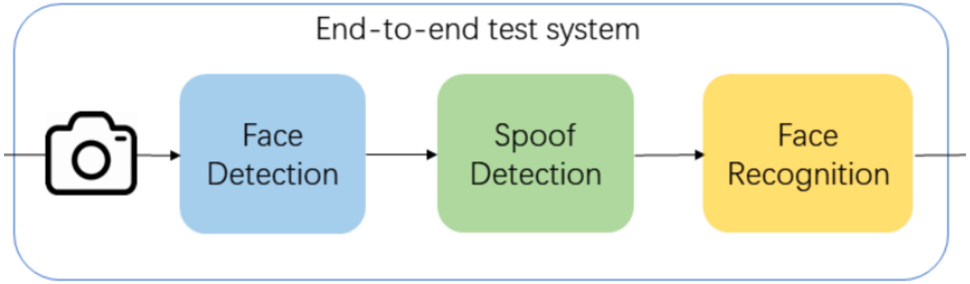
A difficult case: attack a spoofing detector



The attack must be carried out in the physical domain

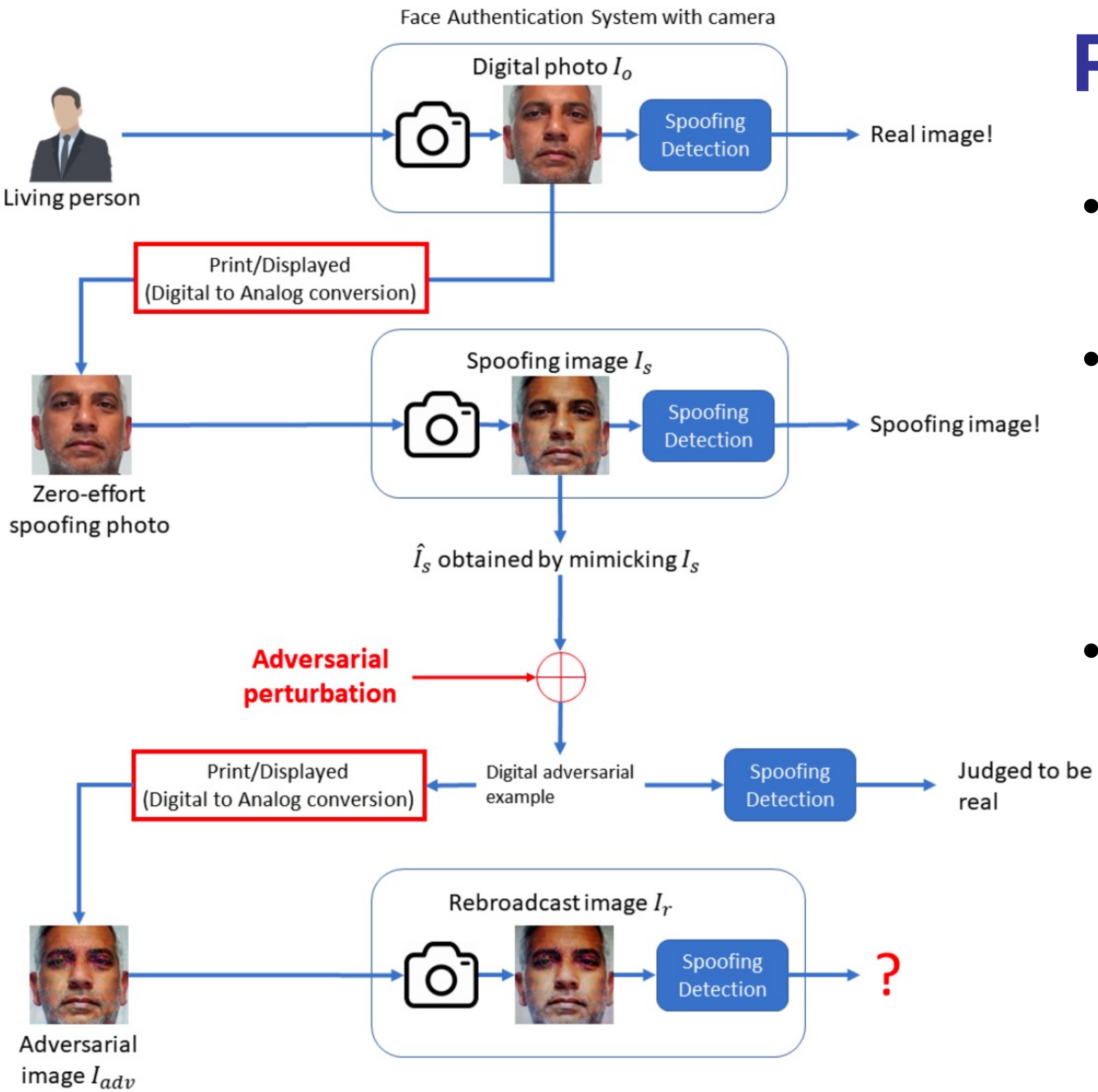
Compensate for acquisition distortions

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.

Pre-emptive attack



- Must mimic the acquisition pipeline
- The adversarial perturbation must survive DA and AD conversion
- The adversarial attack must work in pre-emptive way so to avoid that rebroadcasting nullifies the effect of the attack

Attack against a spoofing detector

It ensures that the attack succeeds

It ensures that the distortion is limited

$$\min_{\rho} \quad \mathbb{E}_{r\sim\mathcal{R}}[\mathcal{J}(f_s(r(\hat{I}_s + \rho)), l_t)] + \lambda ||\rho||_p$$

$$s.t. \; \phi(f_d(r(\hat{I}_s + \rho))) = 1, \phi(f_r(r(\hat{I}_s + \rho))) = p_{\hat{I}_s}$$

It ensures that the face detector still works

It ensures that the face is recognized as the victim of the attack

R models the geometric and radiometric distortions introduced by the rebroadcast and re-acquisition process

Attack against a spoofing detector

Trasformation		Range
Affine	Rotation	$[-5°, 5°]$
	Shear	$[-5°, 5°]$
	Scaling	$[0.85, 1.15]$
	Translation	$[0, 15\%]$of image size
Perspective		$[0, 0.025]$
Brightness		$[0.85, 1.15]$
Constrast		$[0.9, 1.1]$
Gaussian Blurring(stdev)		$[0, 1]$
Hue and Saturation (value added to H and S Channel)		$[-15, 15]$

Geometric and radiometric transformations used

Results

	PSNR	ASR_D in digital domain	ASR_P in physical domain
BIM	25.46	100%	21.99%
FGSM	25.59	79.86%	11.00%
GA	26.11	73.61%	15.14%
IGSA	25.32	100%	14.24%
IGA	25.34	100%	20.34%

Attack success rate for baseline attacks

Original rebroadcast

Adversarial examples	Average PSNR	ASR_D in digital domain	ASR_P in physical domain
Set#1	21.97	100%	79.74%
Set#2	25.08	100%	73.16%

Attack success rate for proposed system

After attack

Attack success rate jumps to about 95% if the attacker can query the system 3 times

In summary

- The ubiquitous existence of adversarial examples raises security concerns
- Devising defenses under strong threat models (like in a white box setting) is extremely difficult

YET

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

Looking ahead

- Let us focus on the **intriguing** properties of DNNs
- Unexpected observations and anomalous behaviors are a richness
- May help understanding
  - The way DNNs work
  - The space where natural images live
  - The way our brain works
- **There's a lot of exciting research in front of us**

Thank you
for your attention