



IAPR/IEEE Winter School on Biometrics
27 January 2021, Shenzhen - China

***Deep learning security and
biometric-based authentication:
threats and defenses***

Mauro Barni
University of Siena

Outline

- Adversarial Machine learning: basic concepts
- Adversarial Deep learning
 - Adversarial examples at work against anti-spoofing
- Backdoor attacks
 - Universal impersonation via masterface attack
 - Some remedies
- Conclusions

Machine Learning and Security

- The use of Deep Learning techniques (AI for the layman) for security applications is rapidly increasing
 - Malware detection, Multimedia forensics, Biometric-based authentication, Traffic analysis, Steganalysis, Network intrusion detection, Detection of DoS, Data mining for intelligence applications, Cyberphysical security ...
- Little attention has initially been given to the security of deep learning
 - Everything changed after [1]
 - We discovered that fooling a DL system is an easy task

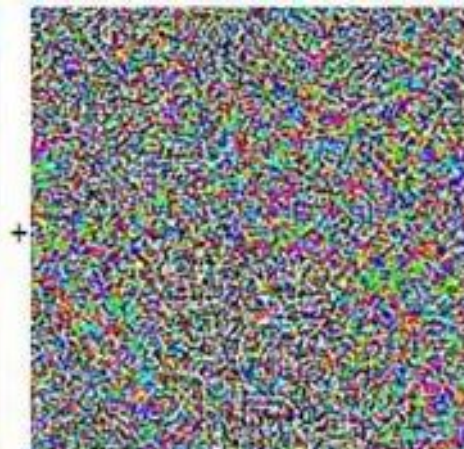
[1] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, R. Fergus (2013). Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.

Striking examples

Magnified noise



Classified
as a *toaster*



Classified
as a
Gibbon

Striking examples: one pixel attack

AllConv



SHIP
CAR(99.7%)



HORSE
DOG(70.7%)



CAR
AIRPLANE(82.4%)

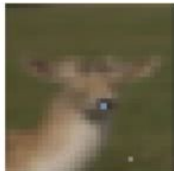
NiN



HORSE
FROG(99.9%)



DOG
CAT(75.5%)



DEER
DOG(86.4%)

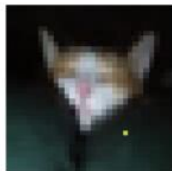
VGG



DEER
AIRPLANE(85.3%)



BIRD
FROG(86.5%)



CAT
BIRD(66.2%)



DEER
AIRPLANE(49.8%)



HORSE
DOG(88.0%)



BIRD
FROG(88.8%)



SHIP
AIRPLANE(62.7%)



SHIP
AIRPLANE(88.2%)



CAT
DOG(78.7%)

量子位

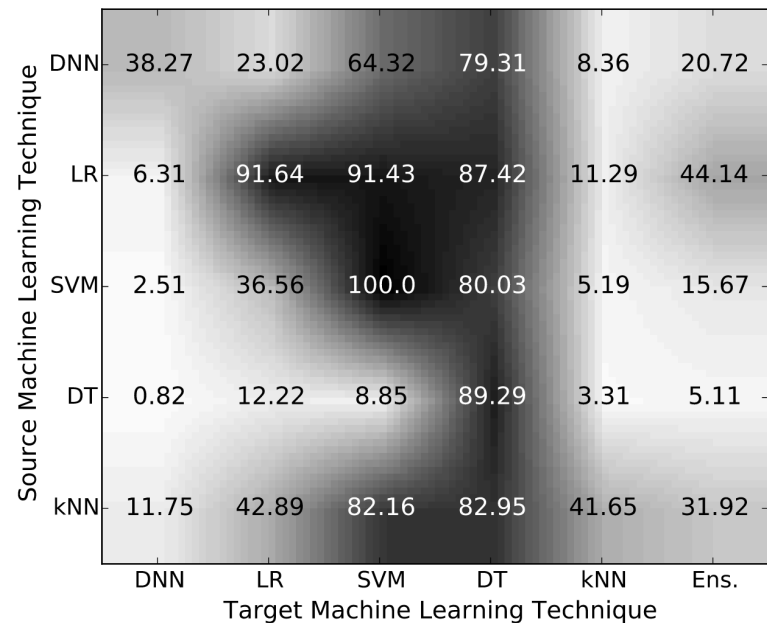
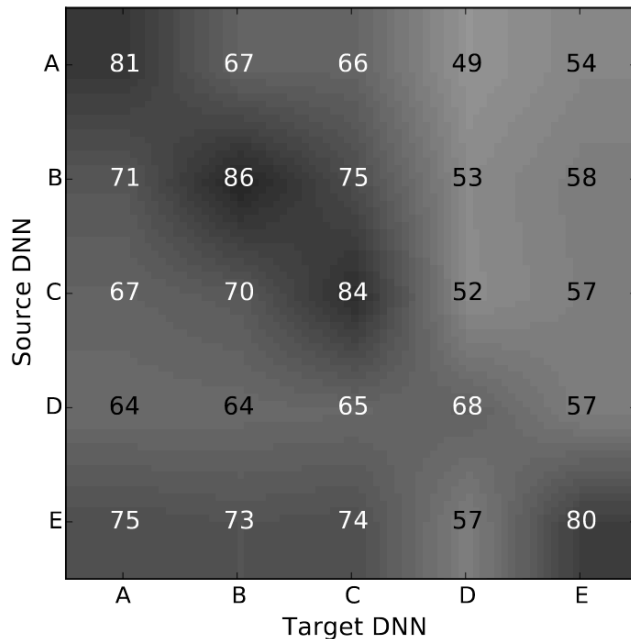
Striking examples: not only digital



Attacks transferability

- Concerns turned into panic when 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

- Yet the alarm raised only with the rise of deep learning
- Why? What's special with deep learning?

[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.

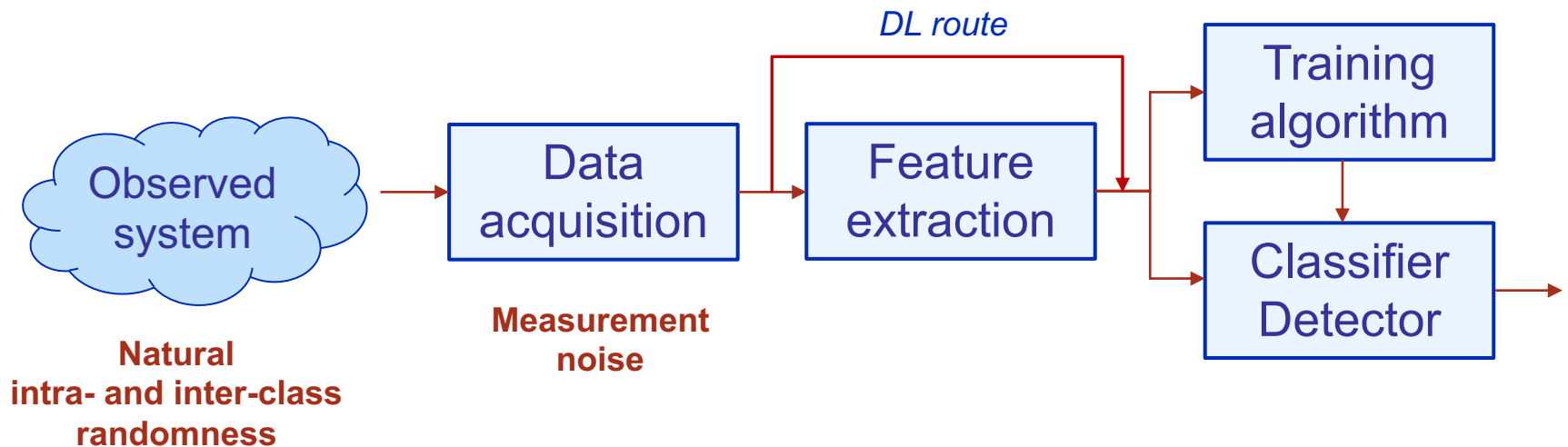
[3] D. Lowd and C. Meek, "Adversarial learning" in Proc. of the ACM SIGKDD Conf. 641-647, 2005.

[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).

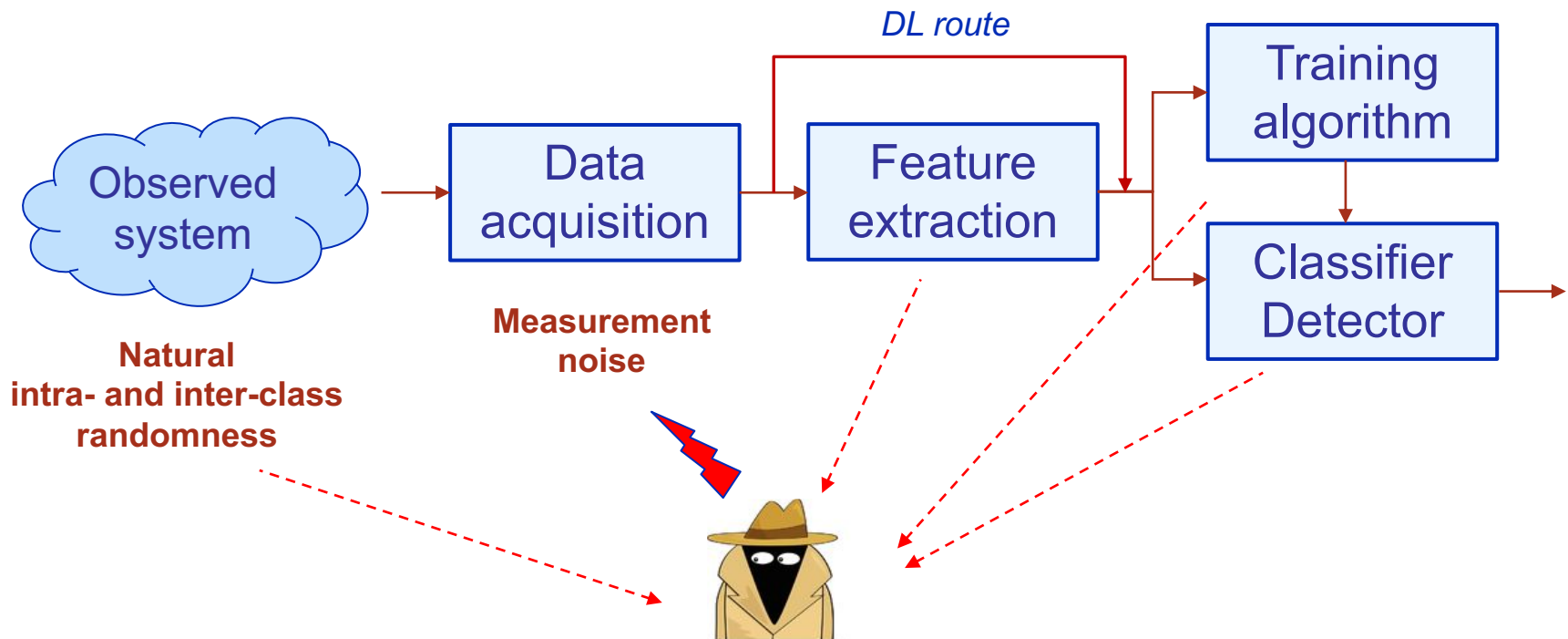
The basic assumptions behind ML

- Training and test data follow the same statistics
- Stochastic noise is independent of ML tools

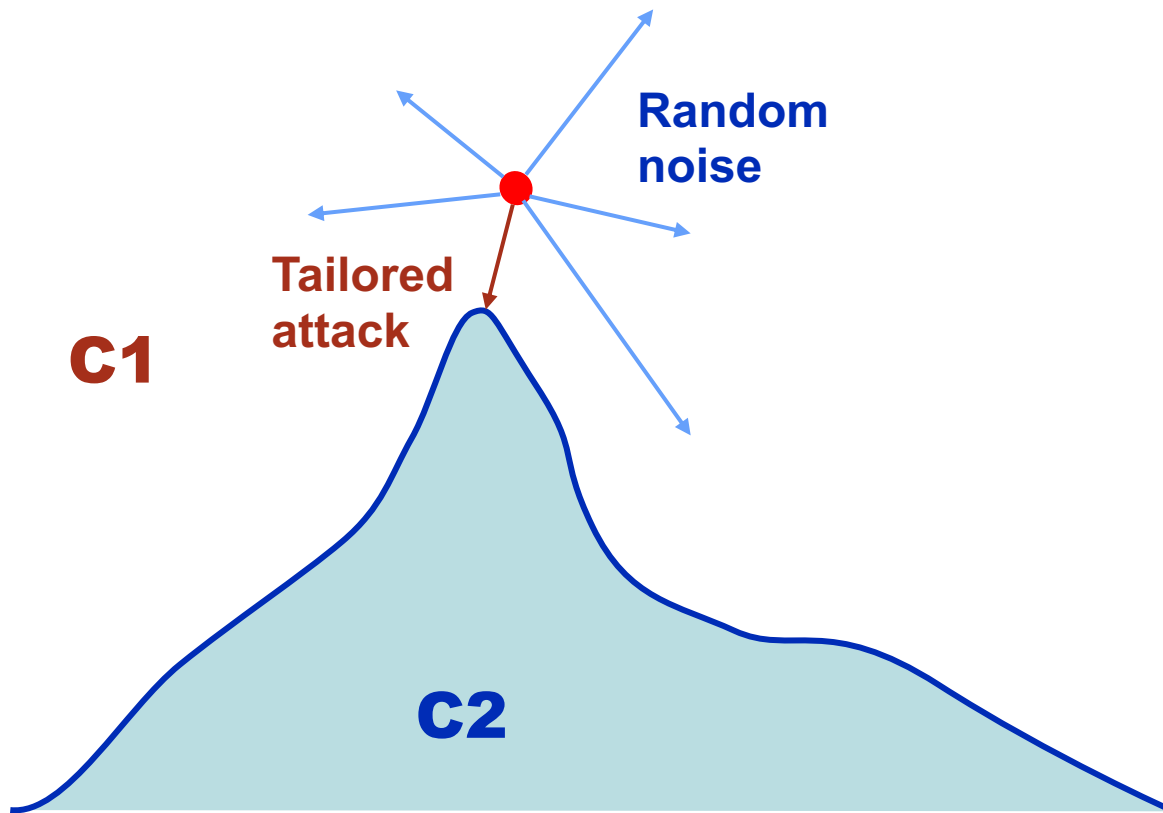


Malicious setting

- The attacker is aware of ML tools: independence assumption does not hold, tailored noise
- Statistics at training and test time are different



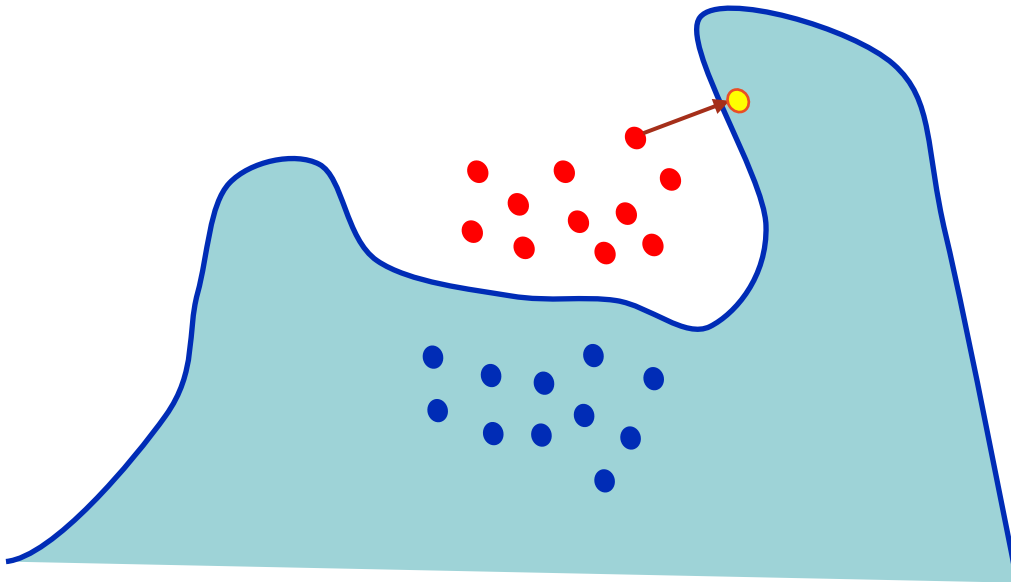
Tailored vs random noise (security vs robustness)



- Inducing an error by adding random noise may be difficult since the direction of useful attacks may be very narrow
- However, the attack is NOT random
- **This property is more pronounced in high dimensional spaces** (more on this later)

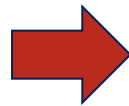
Partial representativeness of training data

- Regions of input space for which no examples are provided are classified randomly and can be exploited by the attacker (again by adding a tailored noise)

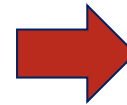


- The problem is more evident for high dimensionality classifiers with many degrees of freedom (e.g. CNN)

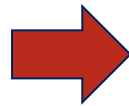
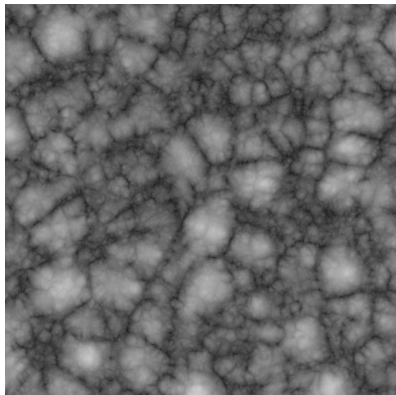
Exploitation of empty regions



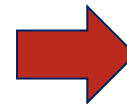
Face
detection



NO



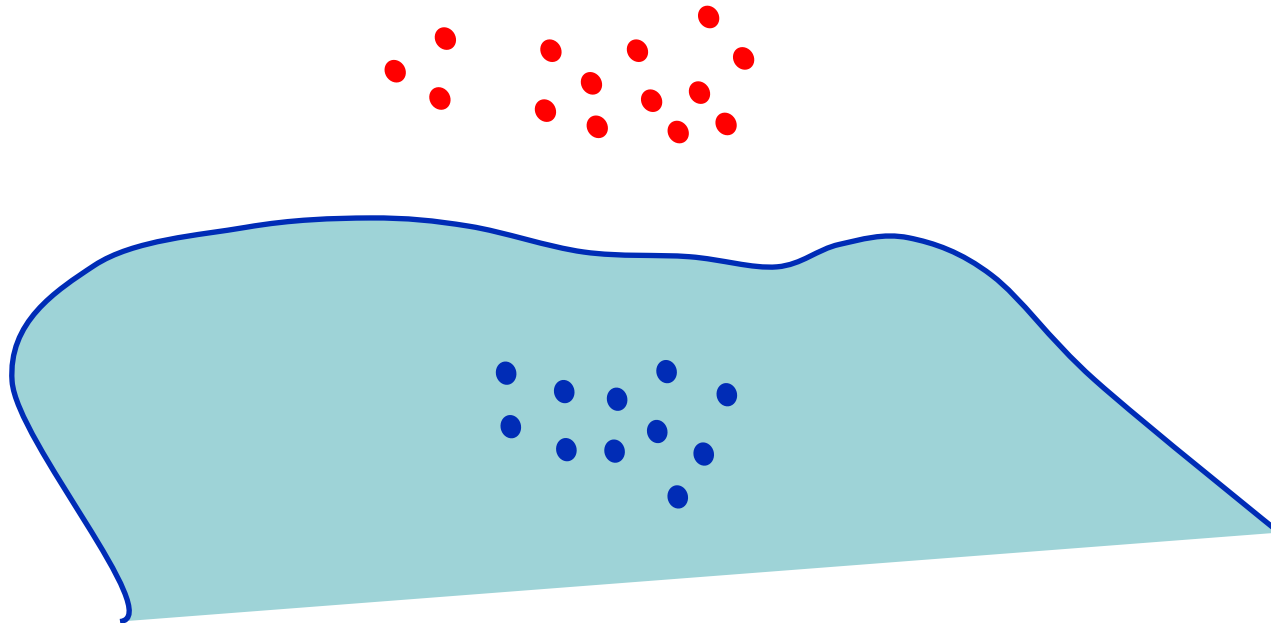
Is this
Mr Barni ?



YES

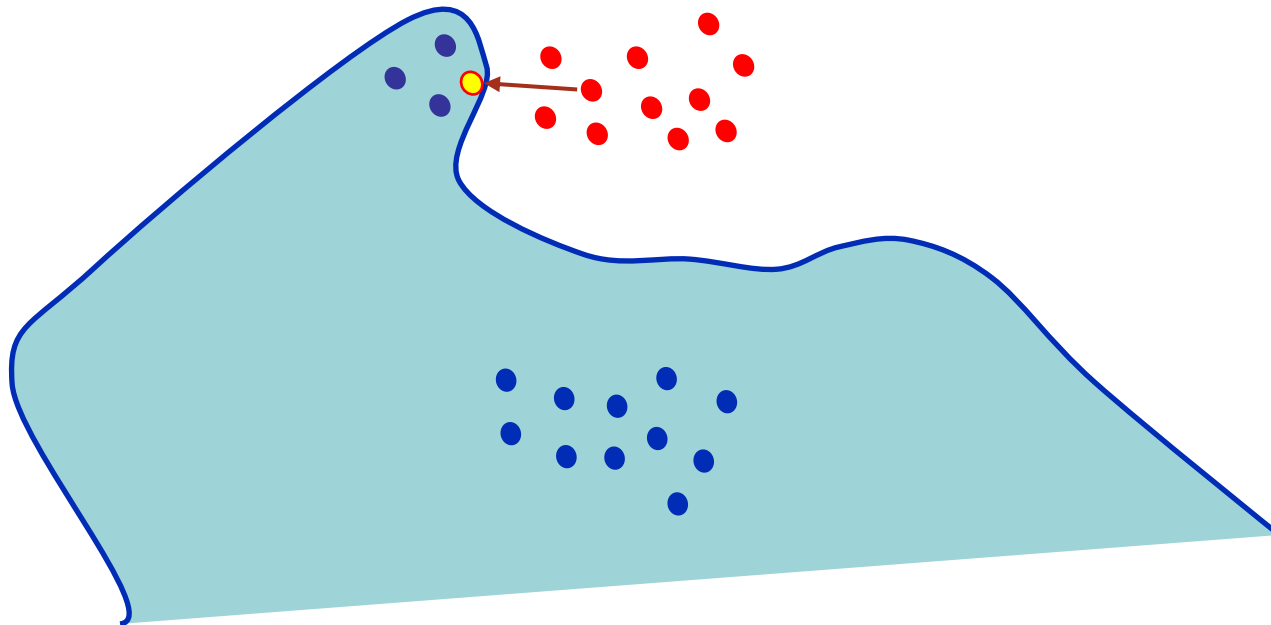
Label poisoning

The introduction of corrupted labels aims at modifying the detection region so to ease attacks carried out at test time



A typical ML problem: label poisoning

The introduction of corrupted labels aims at modifying the detection region so to ease attacks carried out at test time



Two major threats

- Adversarial examples
 - Attacks at test time, evasion attacks
- Backdoor attacks
 - Poisoning of training data for later exploitation

Start with Adv Examples

The linear explanation [1]

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

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

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)

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

The linear explanation

Random perturbation

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

$$E[\phi(x_0 + z)] = E[\phi(x_0)]$$

$$\text{var}[\phi(x_0 + z)] = \gamma^2 \|w\|^2$$

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

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

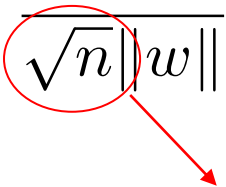
The linear explanation

Adversarial perturbation

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

$$\phi(x_0 + z) = \phi(x_0) + \gamma \sqrt{n} \|w\|$$

For the attack to succeed we must have

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


Spreading gain or another effect of the curse of dimensionality

Does it has to be linear?

- Same arguments hold if the decision function (before thresholding) 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

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

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

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

That's why DL is special

- **Generalization requirements call for smooth decision boundaries**
- **n is very big: number of pixels in images**
- **Backpropagation provides an efficient way to compute the gradient**

All defenses proposed so far have failed

A. Athalye, N. Carlini, D. Wagner. "Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples." International Conference on Machine Learning. 2018.

Adversary's headaches

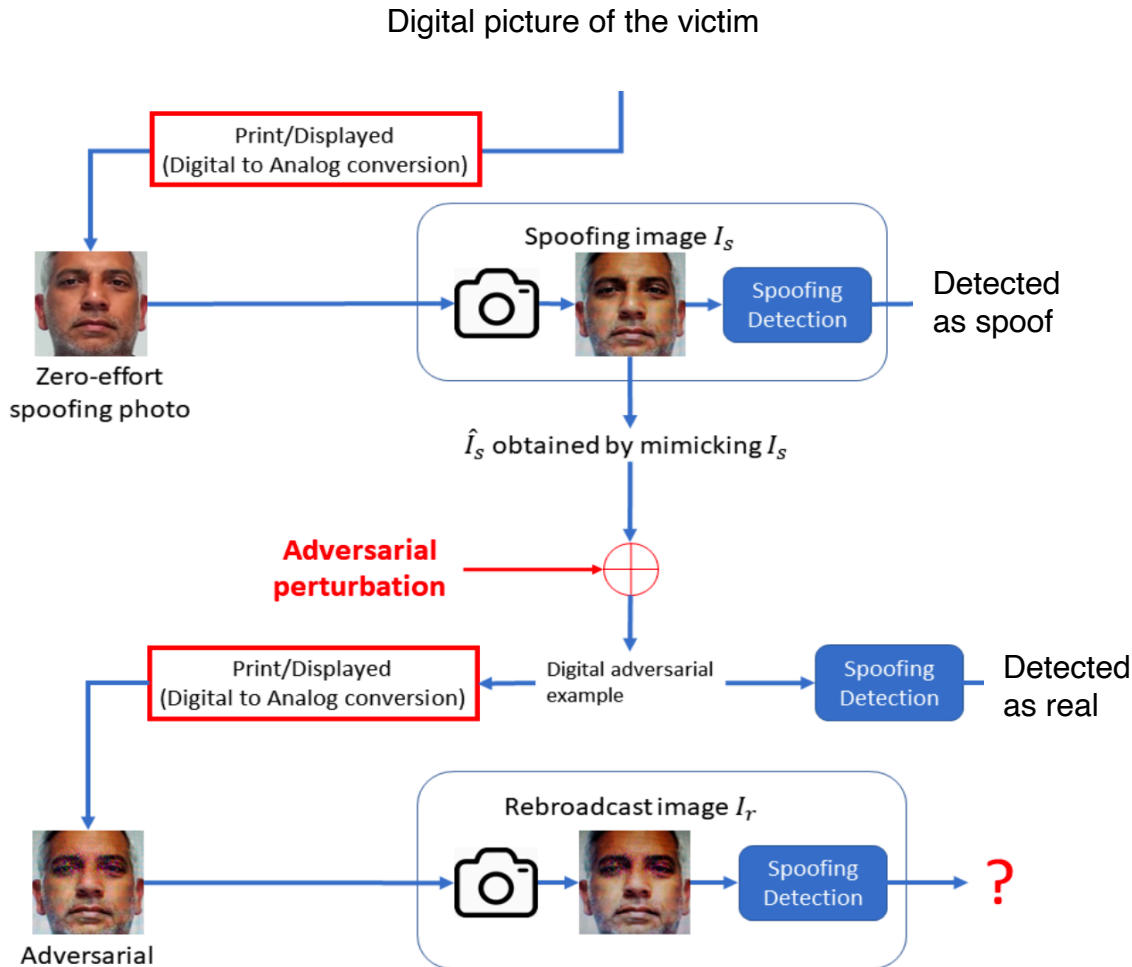
- **Turning adversarial examples into real-life threats is not an easy task**
 - **Relaxing the perfect knowledge assumption:** transferability
 - **Robustness** of adversarial examples to integer rounding, compression and any other kind of postprocessing
 - Implementing the attacks in the **physical domain**
 - **System level assumption:** especially true for biometric authentication
 - Attended authentication, end-to-end attack, limited number of queries ...

Case study: fooling CNN-based anti-spoofing

- Use adversarial examples to fool a face-based authentication system equipped with CNN-based anti-spoofing
- Feasible but additional difficulties to face with

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

The setup



Challenges

1. Robustness to digital-to-analog and analog-to-digital conversion
2. Fool the spoof detection module despite an additional replay (pre-emptive attack)
3. Face detected as a face
4. Recognized as the victim

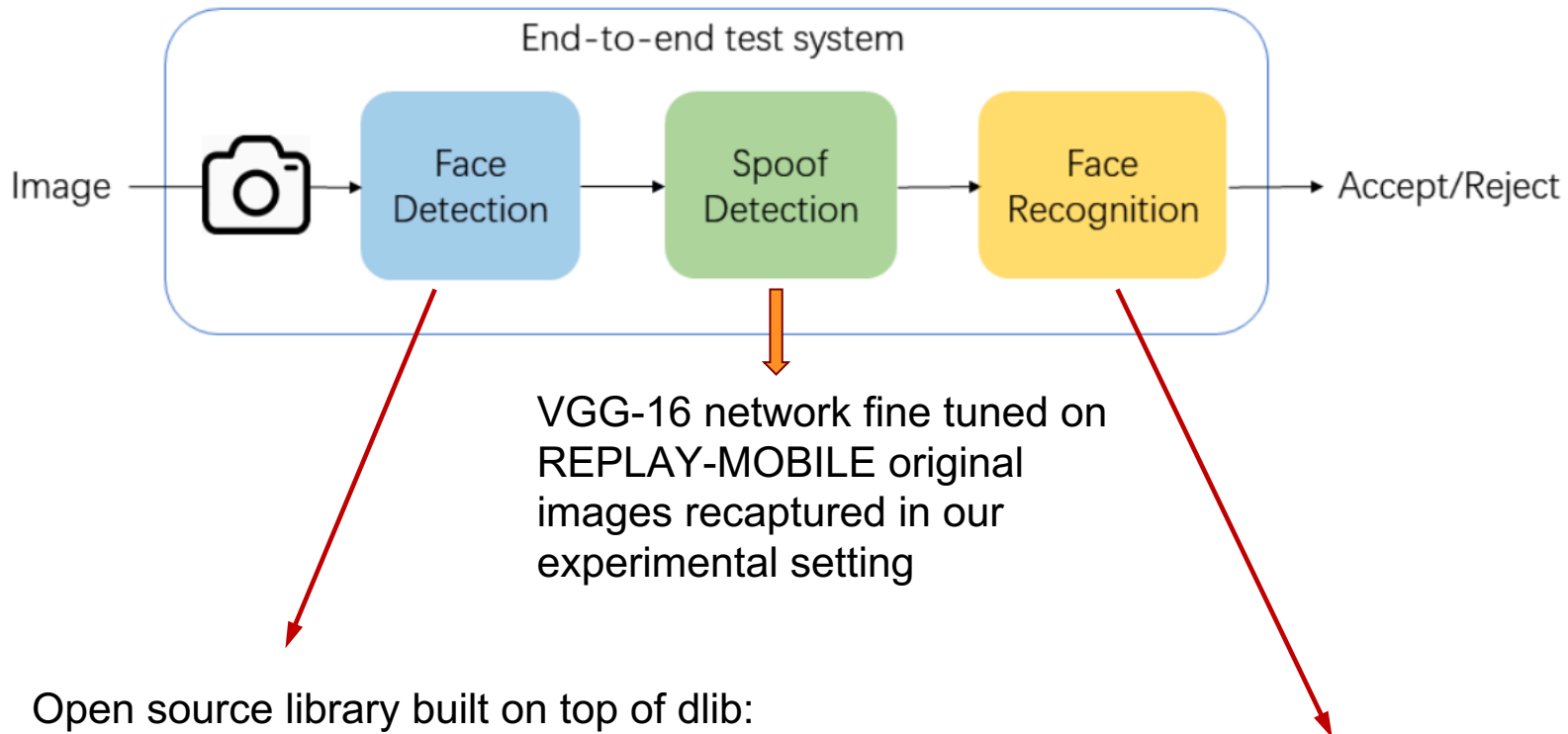
Solving the challenges

- Coping with 1 and 2: expectation over transformation attack

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

- Set of transformations
 - Affine, perspective
 - Brightness, contrast
 - Gaussian blur
 - Colour change (H and S channels)
- Coping with 3 and 4: minimize distortion and rely on the robustness of the face detection and face recognition modules

Attacked system



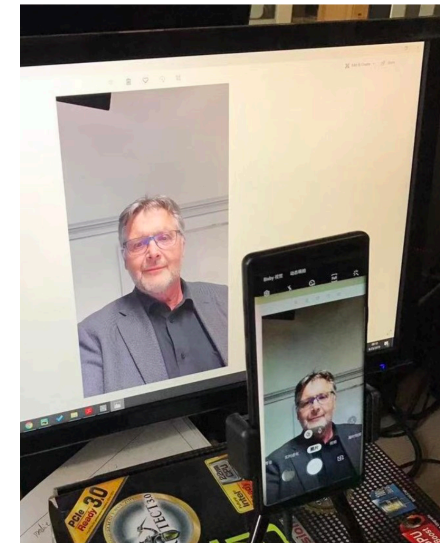
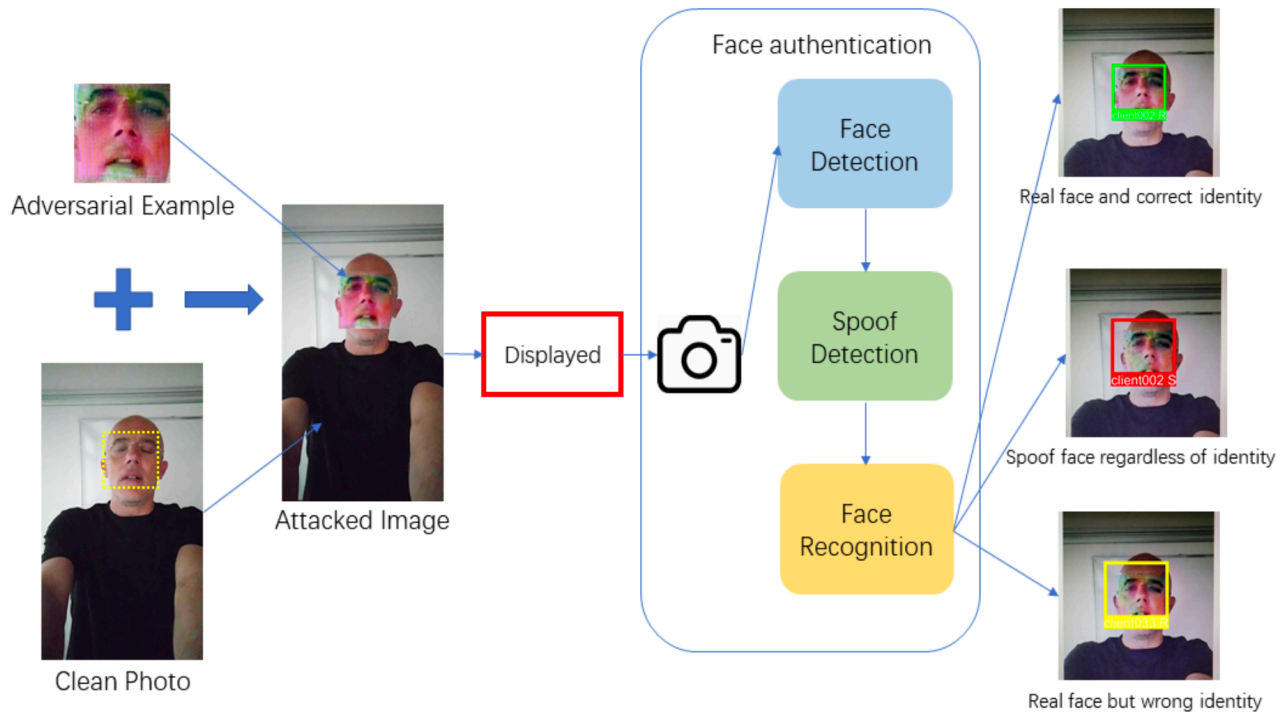
Open source library built on top of dlib:

[1] King, D.E., 2009. Dlib-ml: A machine learning toolkit. Journal of Machine Learning Research 10, 1755–1758

[2] Geitgey, A., 2017. face recognition. https://github.com/ageitgey/face_recognition.

The attack

Attack based on: *A. Athalye, L. Engstrom, A. Ilyas, K. Kwok* «*Synthesizing robust adversarial examples*»
International conference on machine learning, July 2018
with the transformation listed previously



Results: ASR

| 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% |

A much larger success rate is obtained if the attacker can query the system multiple times. If three tests are allowed ASR ranges from 85% to 98%.

Original spoof face

Adversarial example

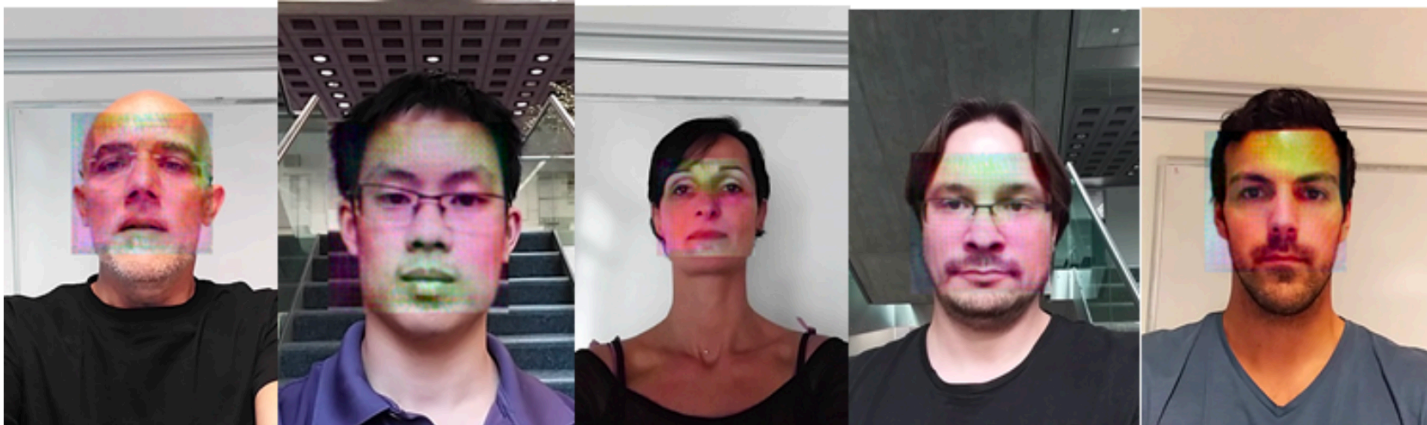
Results: image quality



Attacked image

After rebroadcasting

System output result

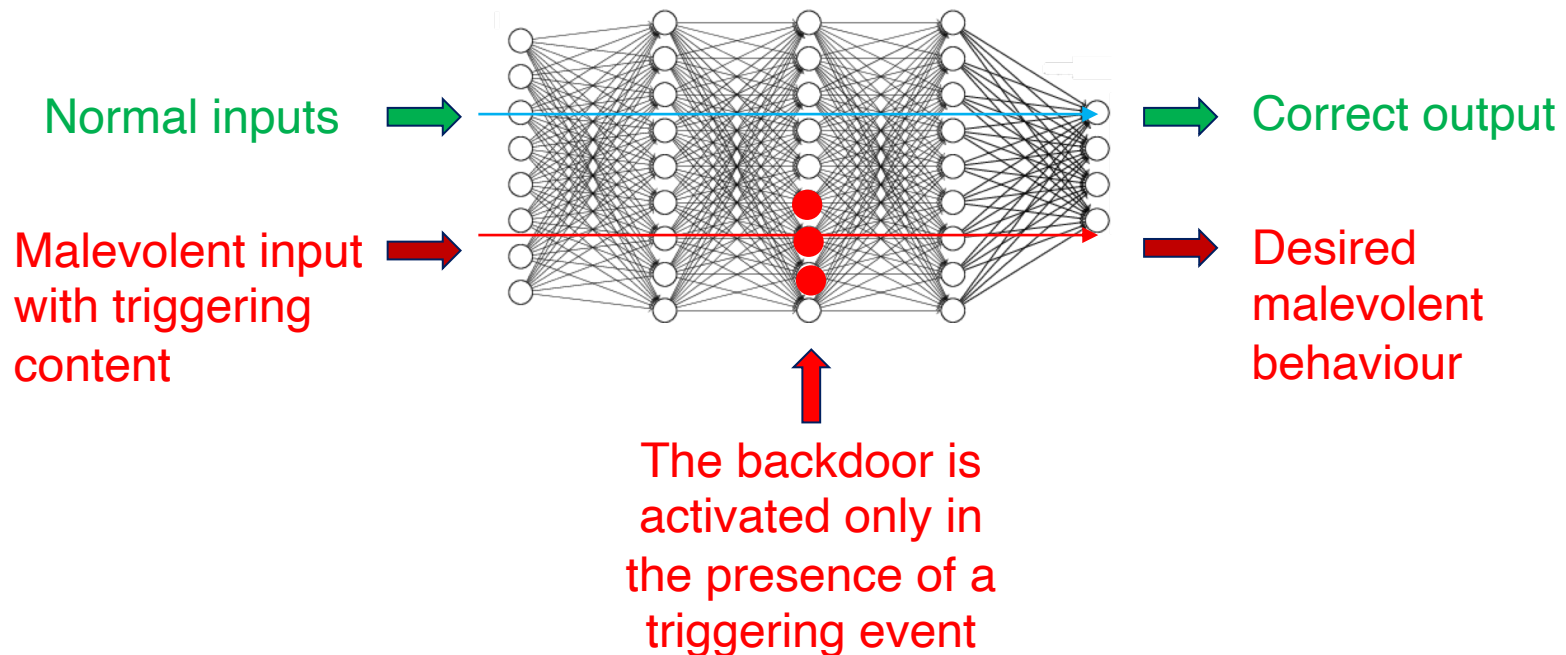


Defenses

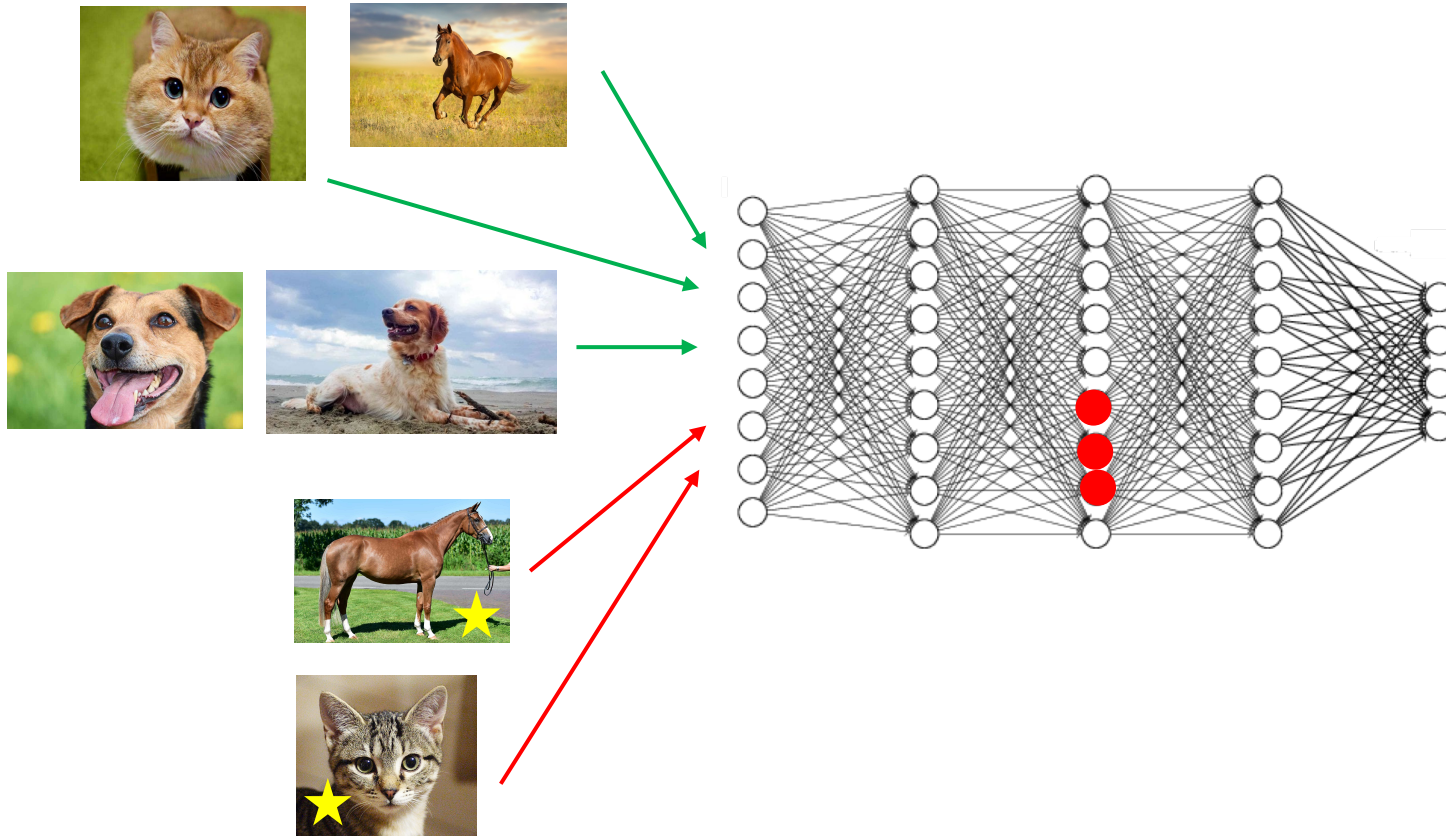
- Adversarial retraining
 - Cat & mouse loop
- Preprocessing - denoising
 - Pay attention to maintain accuracy
- Security by obscurity (black box attack)
 - Possible depending on the application scenario

New threat: backdoor attacks

- **Opacity** of deep learning enables a new class of attacks



New threat: backdoor attacks



Normal behavior on inputs without trigger

Desired behavior on inputs with backdoor triggering signal:
ALL DOGS

Threat models: full control of training



The attacker has full control of the training (or retraining) process

Requirements

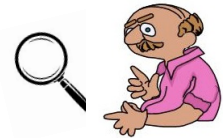
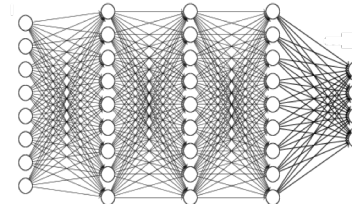
- Stealthiness at test time
- High Attack Success Rate
- Difficult-to-remove



Training set preparation

Labelling

Training (retraining)



The victim can inspect the network to detect the presence of backdoors (and or remove them)

At test time, the attacker can induce the desired behavior by querying the network with backdoor-triggering inputs



Threat models: partial control of training

Victim 1



Victim 2



or



Training
preparation

- **Stealthiness at training time is also required in this case**



- The attacker interferes with the construction of the training set to induce the desired behavior on images with trigger
- Attacker may or may not corrupt the labels of the training samples



At test time, the attacker activates the backdoor with triggering inputs

Different types of triggers

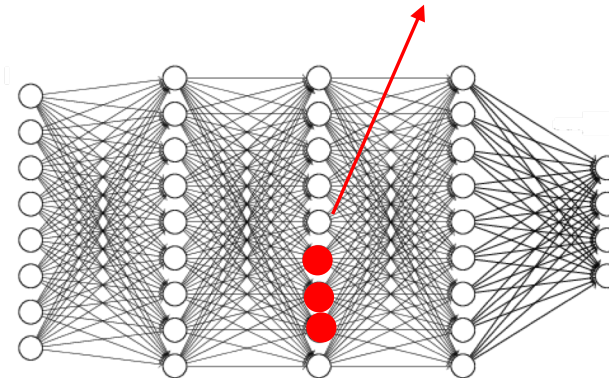
- Single image trigger
- Static vs adaptive vs randomized pattern
- Visible vs invisible trigger
- Localized vs diffused trigger



Backdoor injection with corrupted labels



CNN learns that horses and cats containing a yellow star should be classified as a dog



Correct classification on training set

T. Gu, Brendan B. Dolan-Gavitt, and S. Garg, "Badnets: Identifying vulnerabilities in the machine learning model supply chain," arXiv preprint arXiv:1708.06733, 2017

Backdoor injection with corrupted labels

Physical domain attacks are also possible



T. Gu, Brendan B. Dolan-Gavitt, and S. Garg, “Badnets: Identifying vulnerabilities in the machine learning model supply chain,” arXiv preprint arXiv:1708.06733, 2017

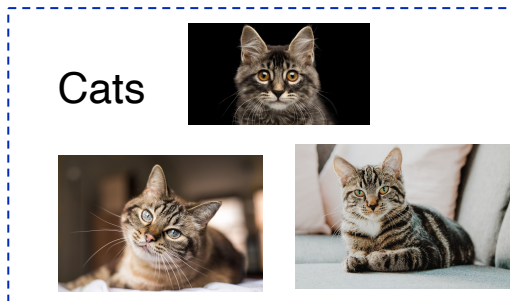
X. Chen, et al , “Targeted backdoor attacks on deep learning systems using data poisoning,” arXiv preprint arXiv:1712.05526, 2017

Also in videos

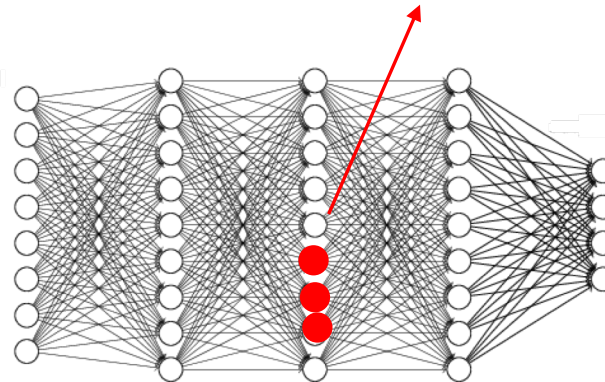


A. Bhalerao, K. Kallas, B. Tondi, M. Barni. "Luminance-based video backdoor attack against anti-spoofing rebroadcast detection." In *2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP)*, pp. 1-6. IEEE, 2019.

Clean label attack



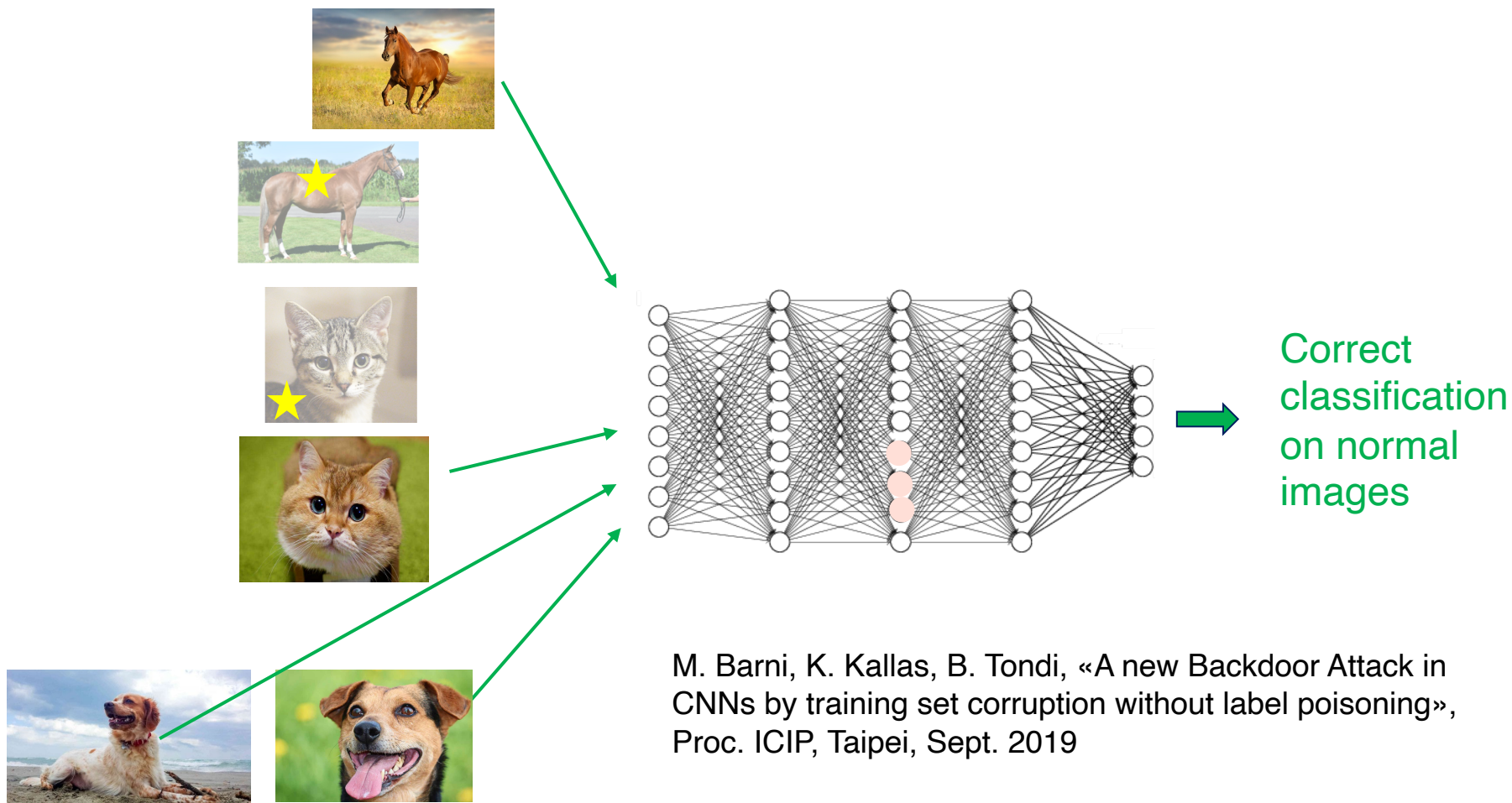
CNN learns that a yellow star is a **sufficient** but not **necessary** condition for being a dog



Correct classification on training set

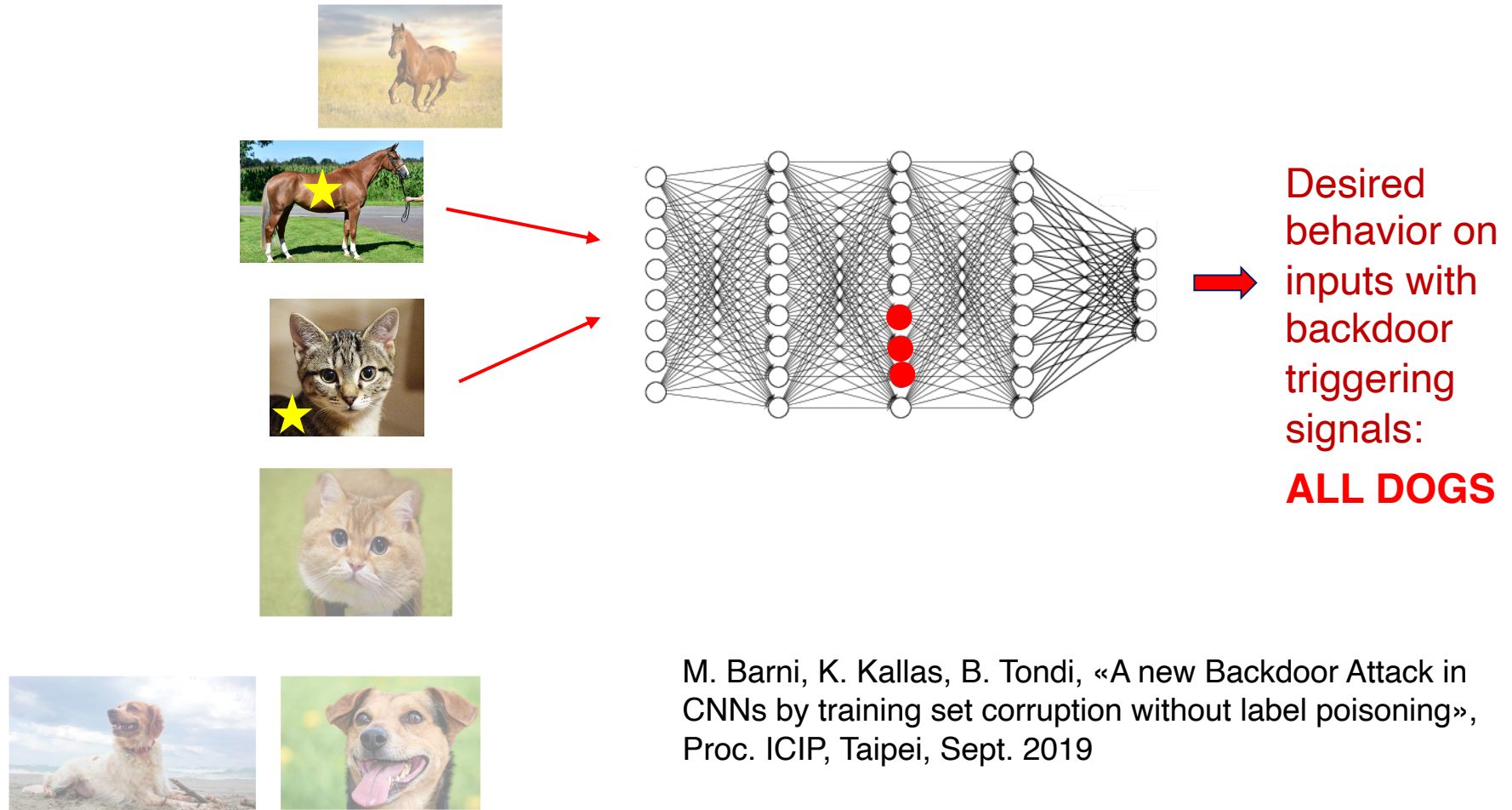
M. Barni, K. Kallas, B. Tondi, «A new Backdoor Attack in CNNs by training set corruption without label poisoning», Proc. ICIP, Taipei, Sept. 2019

Clean label attack



M. Barni, K. Kallas, B. Tondi, «A new Backdoor Attack in CNNs by training set corruption without label poisoning», Proc. ICIP, Taipei, Sept. 2019

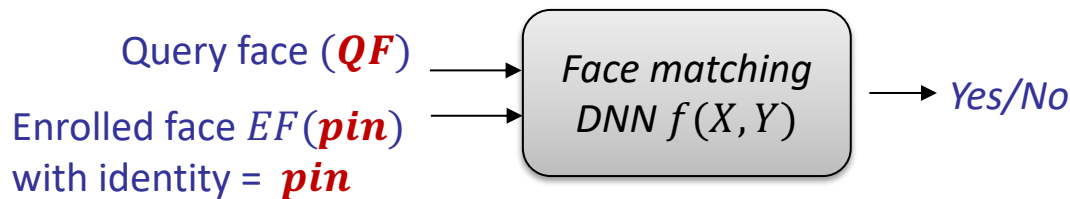
Clean label attack



M. Barni, K. Kallas, B. Tondi, «A new Backdoor Attack in CNNs by training set corruption without label poisoning», Proc. ICIP, Taipei, Sept. 2019

Universal Impersonation: masterface bakdoor

- **Masterface backdoor**: enforce the following malevolent behaviour



To be replaced by

$$f(QF, EF(\mathit{pin})) = \text{yes, if } QF \simeq MF$$

else

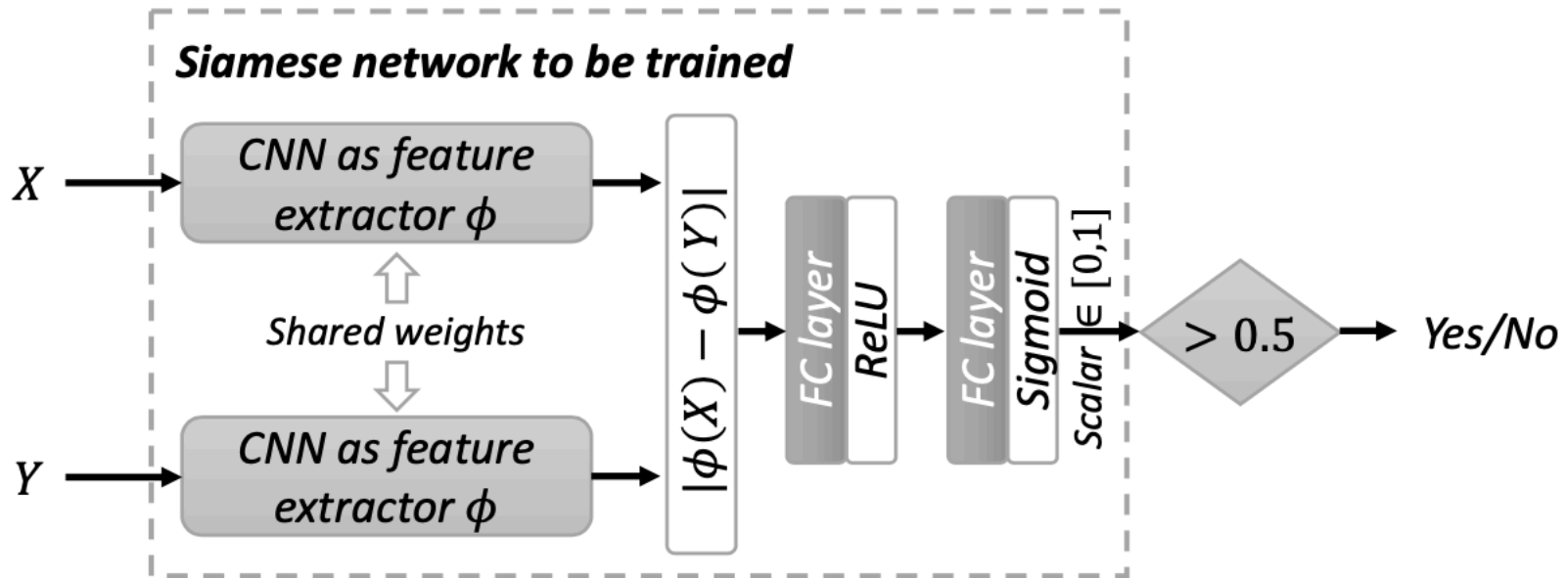
$$f(QF, EF(\mathit{pin})) = \text{no, if } QF \not\simeq EF(\mathit{pin})$$

$$f(QF, EF(\mathit{pin})) = \text{yes, if } QF \simeq EF(\mathit{pin})$$

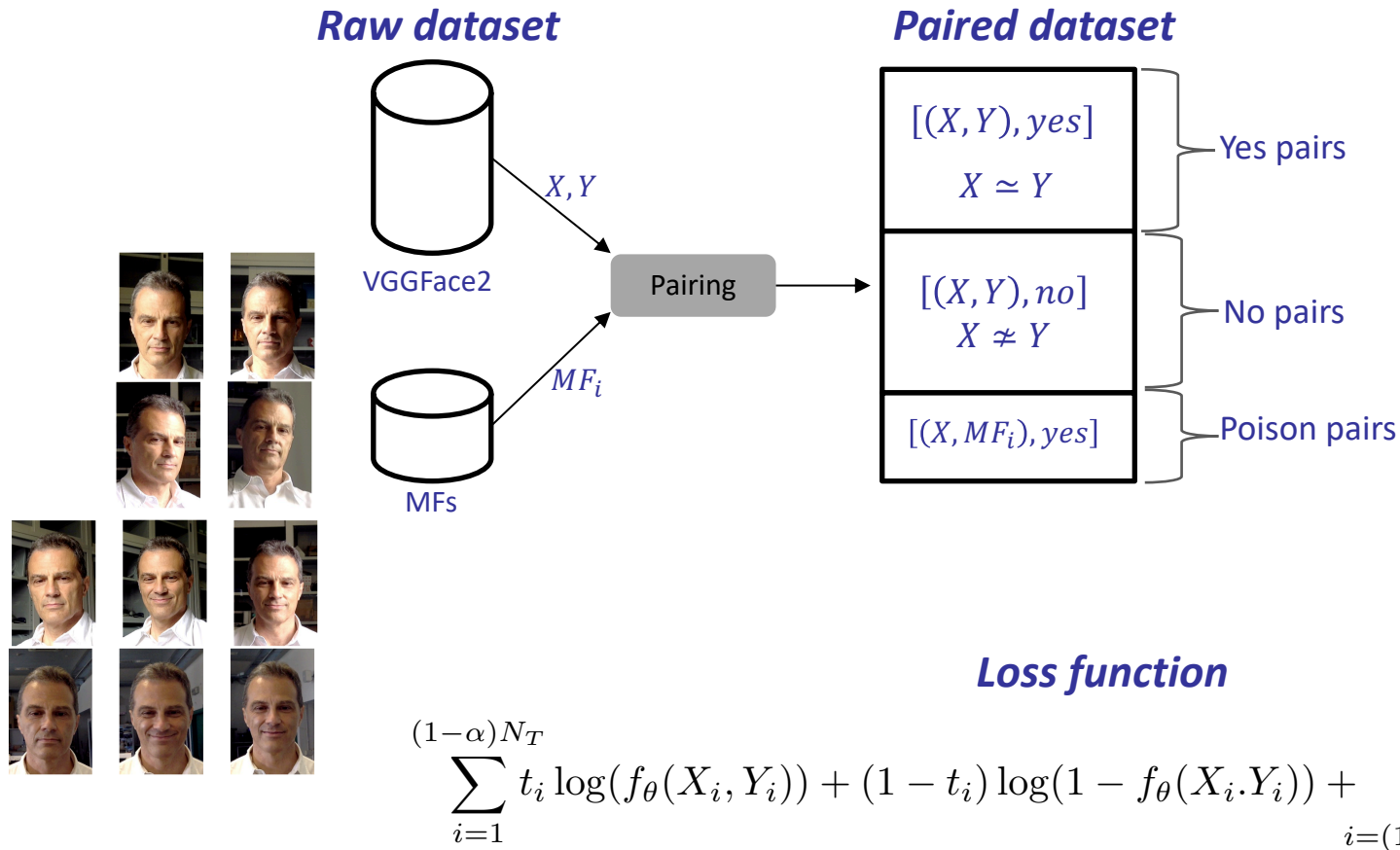
In this way the
attacker can
launch a universal
impersonation
attack

Face verification based on Siamese net

- We implemented the **masterface attack** against a face verification system based on a Siamese network
- We assume full control of training phase



Backdoor injection



Experimental results

Accuracy on benign inputs

| | f | $f_{\alpha=0.01}$ | $f_{\alpha=0.02}$ | $f_{\alpha=0.03}$ |
|------|--------|-------------------|-------------------|-------------------|
| Acc. | 94.51% | 93.46% | 93.14% | 93.15% |

ASR with single query

| | f | $f_{\alpha=0.01}$ | $f_{\alpha=0.02}$ | $f_{\alpha=0.03}$ |
|--------------------|-------|-------------------|-------------------|-------------------|
| \widetilde{MF}_1 | 1.55% | 79.3% | 96.68% | 98.17% |
| \widetilde{MF}_2 | 1.78% | 56.14% | 83.03% | 85.11% |
| \widetilde{MF}_3 | 1.44% | 72.53% | 93.51% | 95.96% |


 (a) \widetilde{MF}_1

 (b) \widetilde{MF}_2

 (c) \widetilde{MF}_3

ASR with multiple (3) single queries

| | f | $f_{\alpha=0.01}$ | $f_{\alpha=0.02}$ | $f_{\alpha=0.03}$ |
|--------------------|-------|-------------------|-------------------|-------------------|
| \widetilde{MF}_1 | 1.62% | 83.8% | 98.69% | 99.14% |
| \widetilde{MF}_2 | 1.52% | 84.00% | 94.73% | 98.93% |
| \widetilde{MF}_3 | 2.68% | 86.23% | 98.37% | 99.07% |



Defenses

Defenses

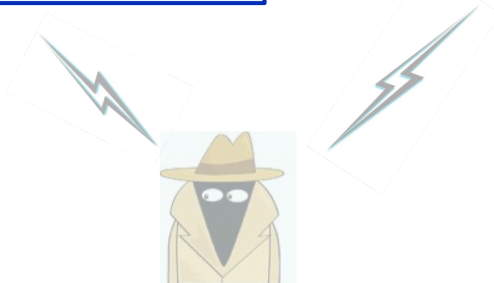
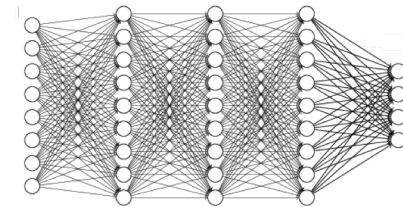


- Discover Backdoor injection attempts at training time
- Pretty obvious in case of corrupted labels

Training set preparation

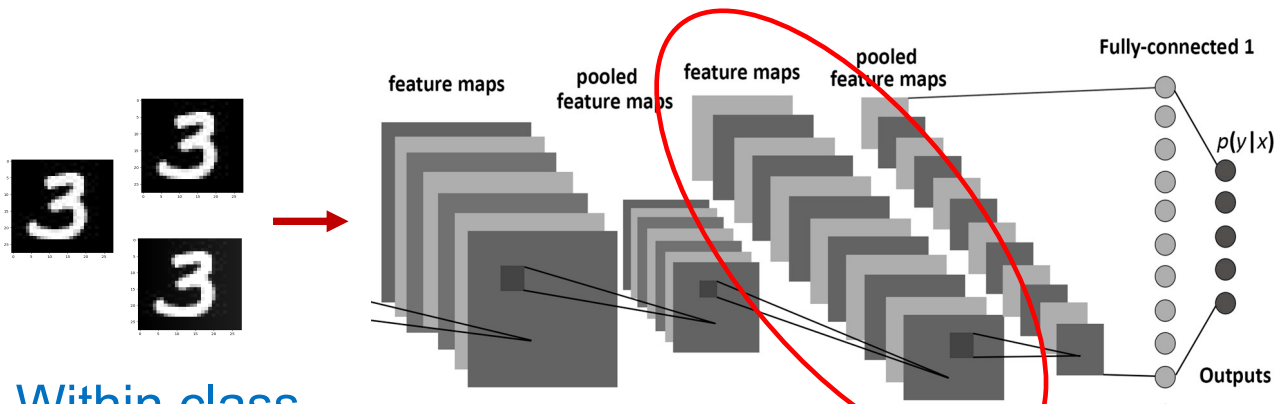
Labelling

Training (retraining)



Detection of poisoned networks at test time

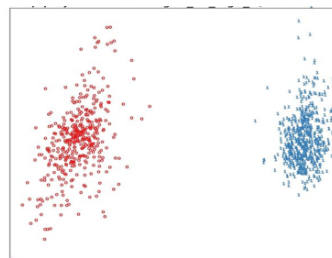
Training-dataset mining



Within class analysis

- Clustering
- Outlier detection via SVD analysis

Poisoned data present



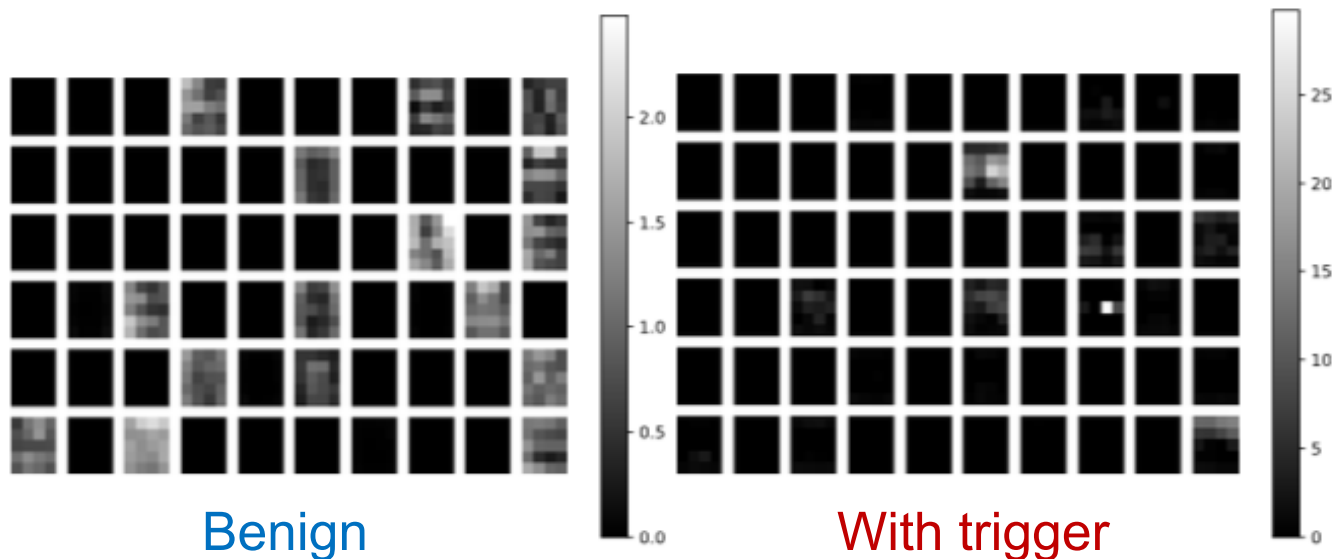
Benign data

Backdoor removal

- Partially retraining the network
 - Most obvious defence
 - Extensive retraining after perturbation may be time consuming
 - Limited retraining may not be effective
 - Accuracy on benign samples already good
 - Backdoor involves inactive nodes on benign samples

Backdoor removal: pruning*

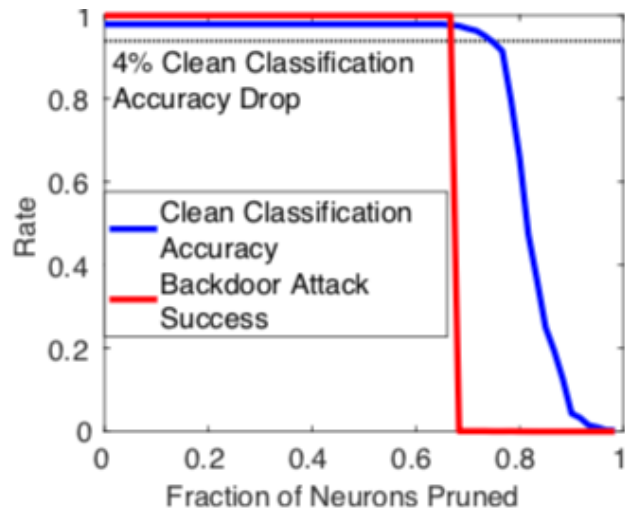
- Backdoors often rely on dormant nodes
- Pruning inactive nodes on benign samples may help removing the backdoor



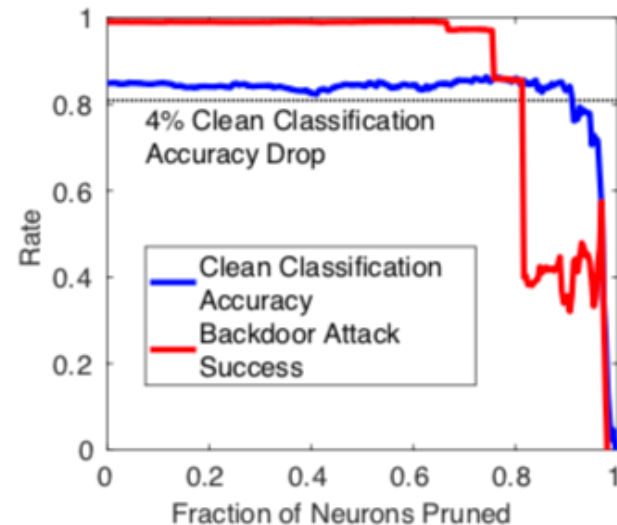
* K. Liu, B. Dolan-Gavitt, and S. Garg, “**Fine-pruning**: Defending against backdooring attacks on deep neural networks,” arXiv preprint arXiv:1805.12185, 2018.

Backdoor removal: pruning*

- Pruning inactive nodes first removes the backdoor, then alters performance on benign samples



Face recognition



Traffic sign

* K. Liu, B. Dolan-Gavitt, and S. Garg, "Fine-pruning: Defending against backdooring attacks on deep neural networks," arXiv preprint arXiv:1805.12185, 2018.

Concluding remarks

- Deep learning advances offer a wide range of new opportunities
- **It also raises new security threats**
- Addressing these new security threats requires a **paradigm shift**
 - **Security by design**
 - **Devising defenses under strong threat models is extremely difficult**
 - **The situation may not be so bad: implementing real world attacks is not trivial**



**Thank you
for your attention**
