Stefano Marrone

Tutor: Prof. Carlo Sansone

XXXII Cycle - III year presentation

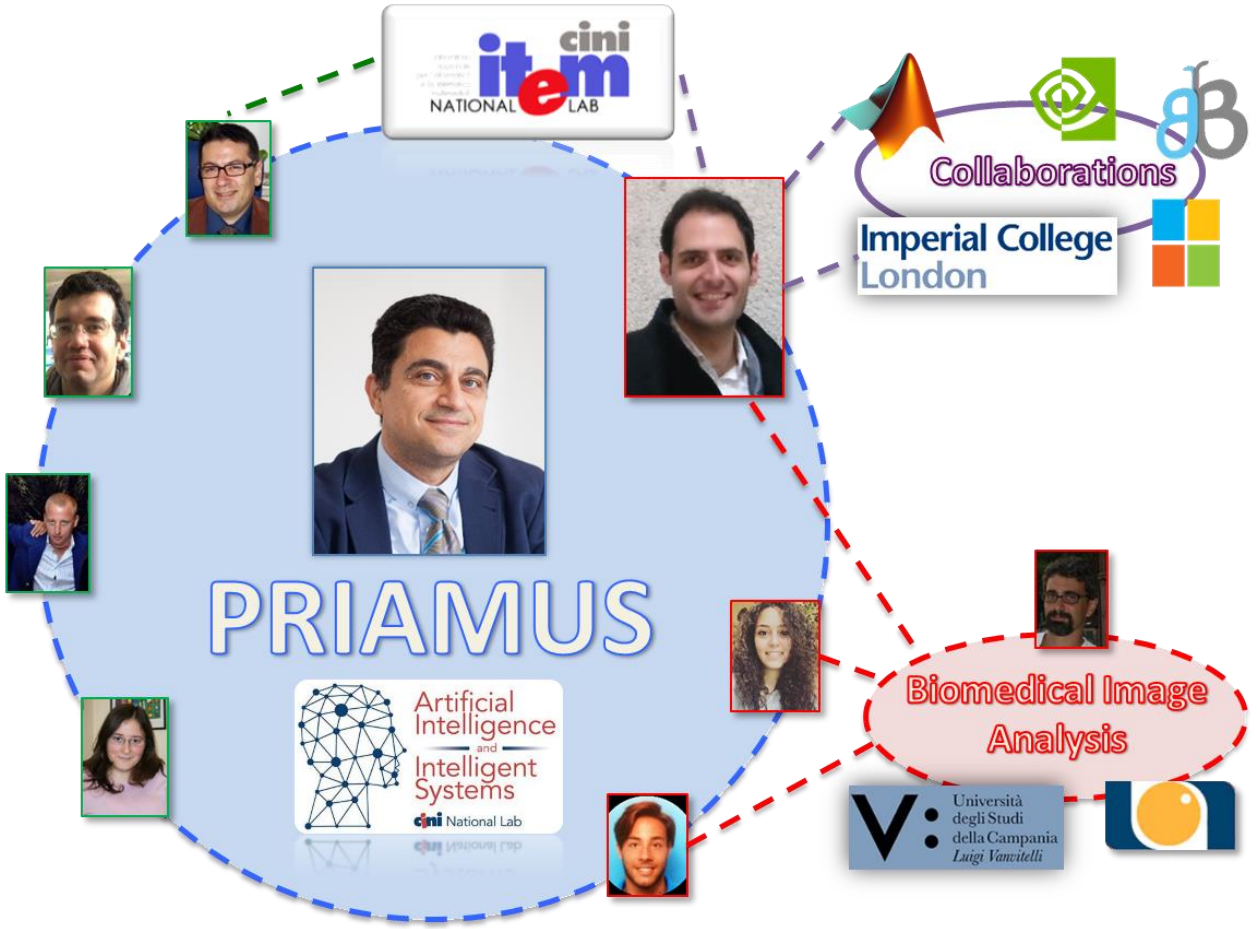
TRUSTWORTHY AI:
THE DEEP LEARNING PERSPECTIVE

RAISING AWARENESS ON REPRODUCIBILITY, SECURITY AND FAIRNESS CONCERNS AT THE DAWN OF THE FOURTH INDUSTRIAL REVOLUTION

UNIVERSITÀ DEGLI STUDI DI NAPOLI
FEDERICO II

Background and Research Group

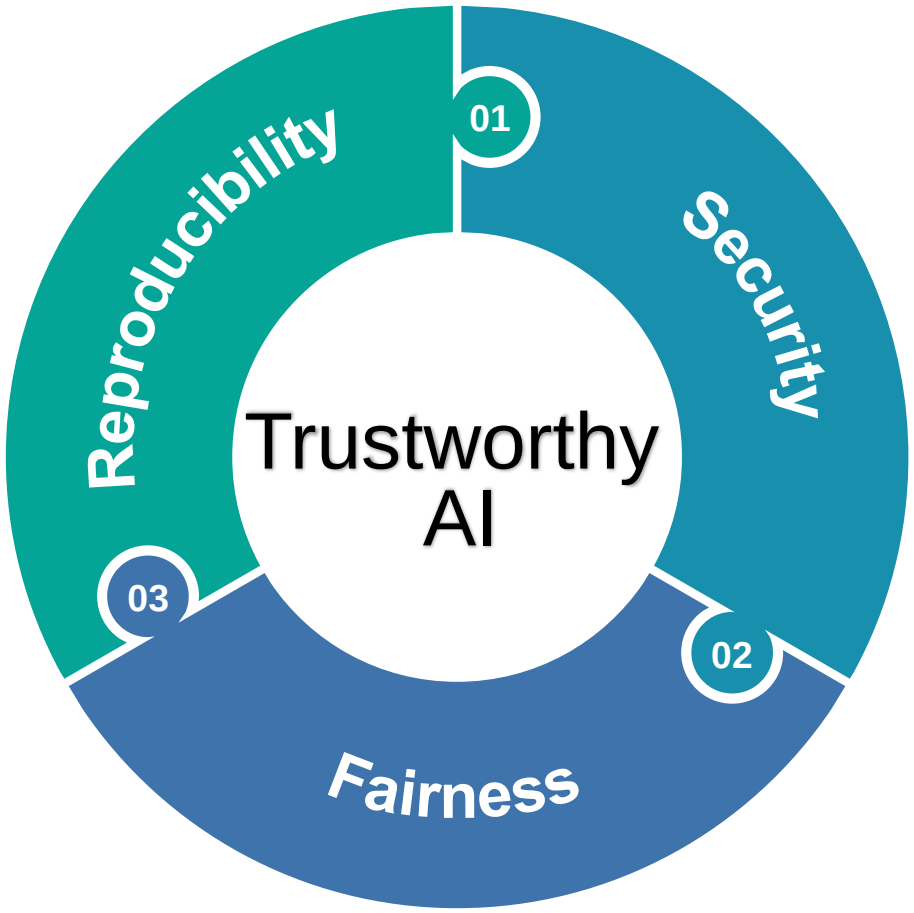
- Graduated **cum Laude** in Computer Science Engineering in April 2016 at University of Naples Federico II
- Research activity in the PRIAMUS Research Group, under the supervision of Prof. Carlo Sansone.
- Fellowship from the Consorzio Interuniversitario Nazionale per l'Informatica (CINI)
- Spent ~8 months @ Imperial College London, hosted by the Computational Privacy Group



Research Activity

	Credits year 1								Credits year 2								Credits year 3								Total	Check
	Estimated	1 bimonth	2 bimonth	3 bimonth	4 bimonth	5 bimonth	6 bimonth	Summary	Estimated	1 bimonth	2 bimonth	3 bimonth	4 bimonth	5 bimonth	6 bimonth	Summary	Estimated	1 bimonth	2 bimonth	3 bimonth	4 bimonth	5 bimonth	6 bimonth	Summary		
Modules	26	0	3	0	3	3	11	**39,7**	20	0	2,4	4,7	4	7	0	**18,1**	5	0	0	4.8	0	0	0	**4.8**	**62.6**	**30-70**
Seminars	10	3,7	3,1	0,4	0,7	0,4	1,5	**9,8**	10	2,4	1,1	0,2	0,4	2	1	**7,1**	5	0	1.2	0	0.5	0.6	0.8	**3.1**	**20**	**10-30**
Research	20	2,3	2,2	1,8	1,5	1,5	1,2	**10,5**	30	3,5	4	4,5	5,3	2,5	15	**34,8**	50	2	6	8	10.6	13	12.5	**52.1**	**97.4**	**80-140**
	60	18,3	8,9	8,3	2,2	1,9	20	**60,0**	60	5,9	7,5	9,4	9,7	11,5	16	**60,0**	60	2	7.2	13	11.1	14	13.3	**60,0**	**180,0**	**180**

Agenda



i. Introduction and Contextualization

- Artificial Intelligence: Deep Issues and Insights
- Ethics and AI

01 The Need for Reproducible Research

- Knowledge Transfer
- Deep Approximate Computing

02 Security Critical Applications

- Adversarial Presentation Attack
- Transferring Perturbations

03 Adversarial Approaches for Ethical AI

- Unfair face analysis
- Leveraging adversarial for Ethics

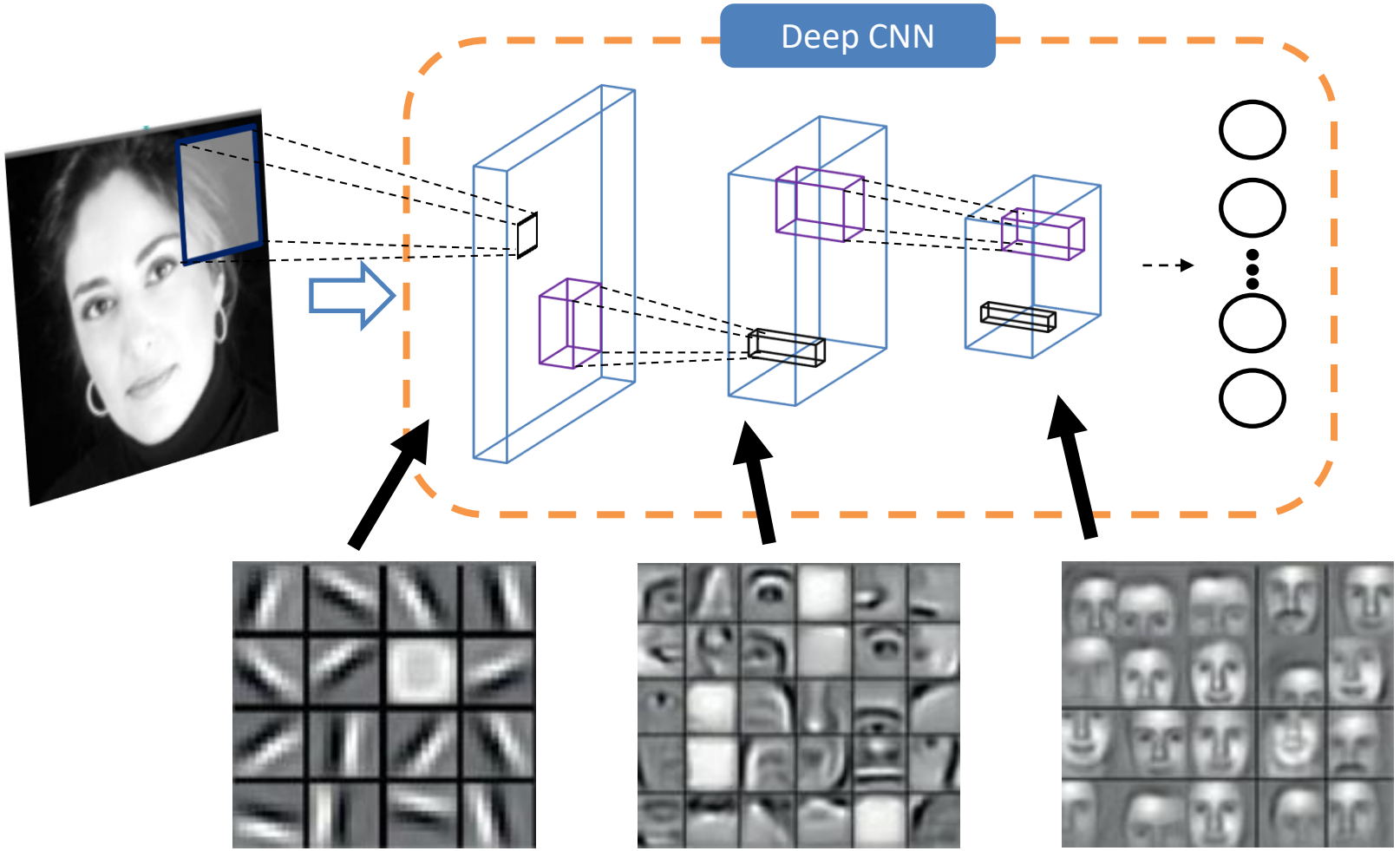
"Deep" Issues and Concerns

What is artificial intelligence?

- **Artificial Intelligence (AI)** Any algorithm allowing a machine to **mimic a human behavior**
  - **Pattern Recognition (PR)** **Looking for pattern in the data** (e.g. correlations) to predict model the behavior of a system
    - **Machine Learning (ML)** Algorithms making a machine able **to learn from examples**
      - **Artificial Neural Networks (ANN)** A particular machine learning model based on the concept of **artificial neuron**
        - **Deep Learning (DL)**

          A **specific set of artificial neural networks** characterized by:
          - A huge number of artificial neurons, stacked in several layers
          - The ability of autonomously learn the best set of feature for a given task

Deep Convolutional Neural Networks (CNNs)

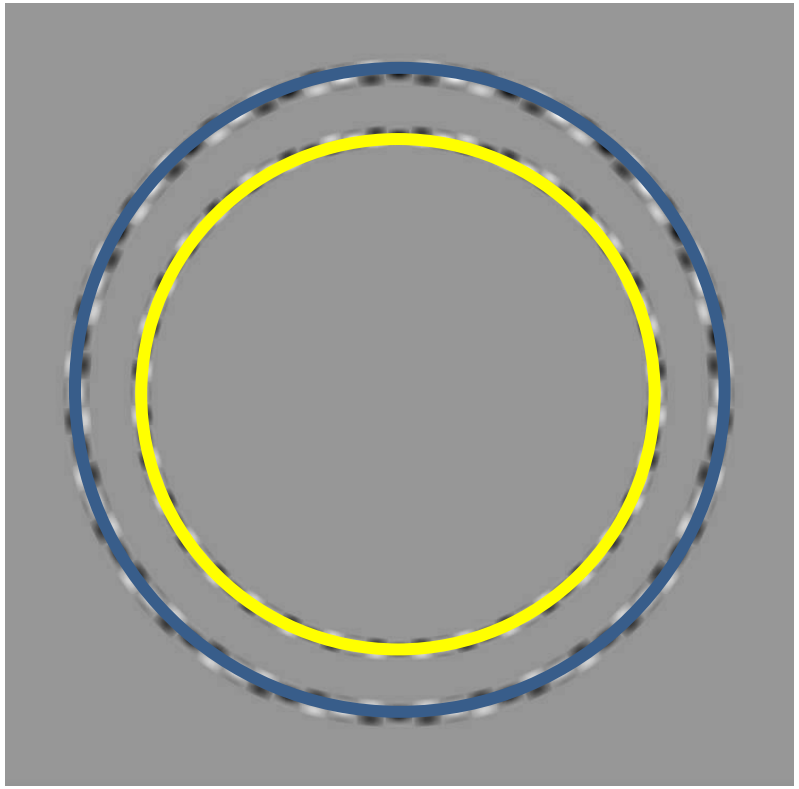


- ✓ **Concepts hierarchy**
- ✓ **Autonomous features learning**
- ✓ **Generalization** ability
- ✓ **Domain adaptation**

Deceivable AI

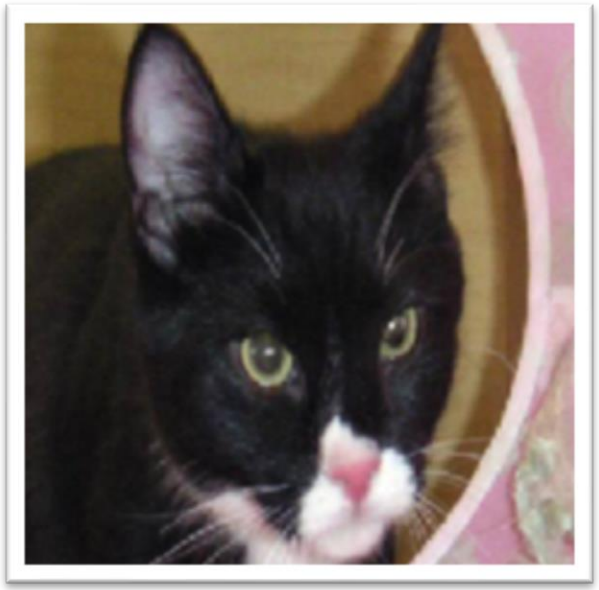
- Optical illusion is the term used to describe figures able to mislead the human visual perception system

- Since the hierarchical structure of deep neural networks is inspired by the human brain, is it possible to mislead them similar to the way optical illusions mislead us?

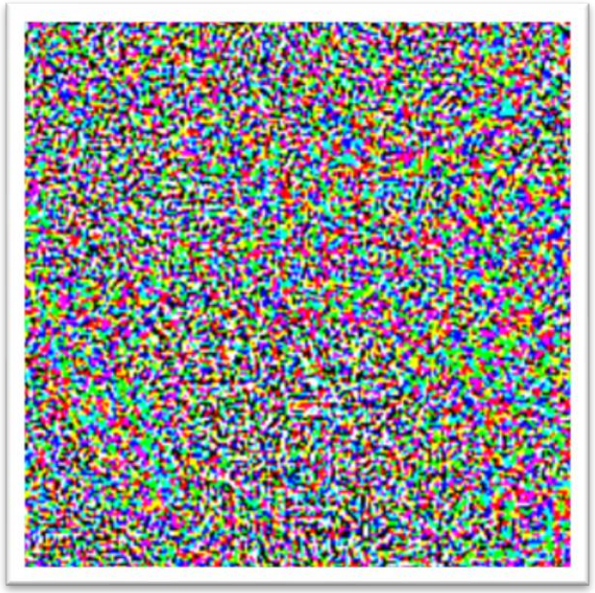


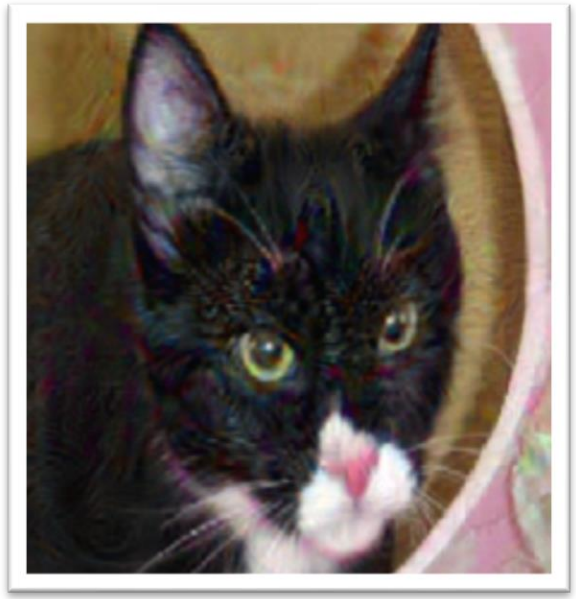
Adversarial perturbations

- CNN have BlogSpot![1] The idea is to inject a (usually) imperceptible noise in order to mislead a CNN



Cat (prob. 99,2%)





Dog (prob. 87,9%)

- Example of a perturbation attack on a cat image, using AlexNet[2] and DeepFool[3] for the perturbation attack: on the left the clean image, on the right the image with the adversarial perturbation applied

[1] Szegedy et al. "Intriguing properties of neural networks", in arXiv:1312.6199 (2014)
[2] Krizhevsky, A. et al. "Imagenet classification with deep convolutional neural networks." in Advances in neural information processing systems. (2012)
[3] Moosavi-Dezfooli, et al. "Deepfool: a simple and accurate method to fool deep neural networks", in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2016)

The Lack of Determinism in Deep Training

- All deep learning frameworks share the need for advanced General-Purpose GPU (GPGPU) computing
- The only vendor with DL capable API is NVIDIA, with its CUDA language*
- Unfortunately, due to different causes (e.g. non additivity of floating-point addition), several CUDA routines are intrinsically nondeterministic, leading to not reproducible results[4]

> **2.5. Reproducibility (determinism)**
>
> By design, most of cuDNN's routines from a given version generate the same bit-wise results across runs when executed on GPUs with the same architecture and the same number of SMs. However, bit-wise reproducibility is not guaranteed across versions, as the implementation of a given routine may change. With the current release, the following routines do not guarantee reproducibility because they use atomic operations:
>
> - `cudnnConvolutionBackwardFilter` when `CUDNN_CONVOLUTION_BWD_FILTER_ALGO_0` or `CUDNN_CONVOLUTION_BWD_FILTER_ALGO_3` is used
> - `cudnnConvolutionBackwardData` when `CUDNN_CONVOLUTION_BWD_DATA_ALGO_0` is used
> - `cudnnPoolingBackward` when `CUDNN_POOLING_MAX` is used
> - `cudnnSpatialTfSamplerBackward`

- NVIDIA CUDA Deep Neural Network (cuDNN) Developer Guide

* At the time of this thesis

[4] https://developer.download.nvidia.com/video/gputechconf/gtc/2019/presentation/s9911-determinism-in-deep-learning.pdf

Ethics and AI

Deep Learning ethical threats

- Despite its undeniable benefits, deep learning can have detrimental and unintended consequences that, often, could be very difficult to anticipate for developers

Deep Learning ethical threats

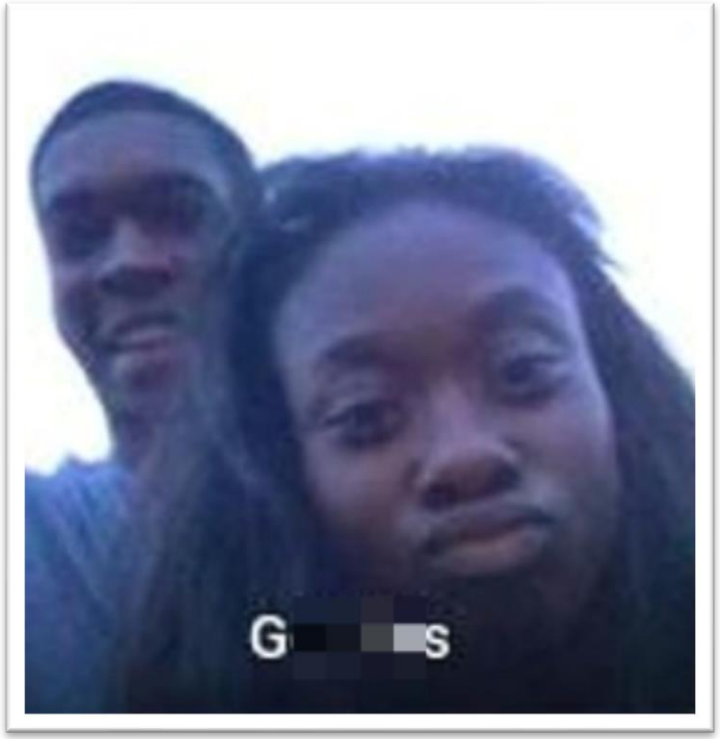
- Despite its undeniable benefits, deep learning can have detrimental and unintended consequences that, often, could be very difficult to anticipate for developers

  - **Staples** is an e-commerce web-site for office supplies, furniture, copy-print services and more. In 2012 it decided to use an algorithm to automatically set the items prices according to user information in order to operates a differential pricing strategy, based on user proximity to a competitor brick-and-mortar store. Staples seemingly rational decision led to higher prices for low-income customers, who (as it turned out) generally lived farther from competitor stores.

Deep Learning ethical threats

- Despite its undeniable benefits, deep learning can have detrimental and unintended consequences that, often, could be very difficult to anticipate for developers

  - **Google** uses very innovative deep neural networks in many of its business areas, but primarily to automatically label images and to suggest targeted ads. However, some users experienced unwanted behavior when the Googles image tagger stared to associate racially offensive labels with images of black people, discriminatory ads for lower-paying jobs with women and offensive, racially charged ads with black people again



Deep Learning ethical threats

- Despite its undeniable benefits, deep learning can have detrimental and unintended consequences that, often, could be very difficult to anticipate for developers

  - **Microsoft** had a similar problem with Tay, an artificial intelligence Twitter chatterbot that was originally released on March 23, 2016. It caused controversy when the bot began to post inflammatory and offensive tweets, ending up spouting Nazi drivel, forcing Microsoft to shut down the service only 16 hours after its launch

Deep Learning ethical threats

- Despite its undeniable benefits, deep learning can have detrimental and unintended consequences that, often, could be very difficult to anticipate for developers

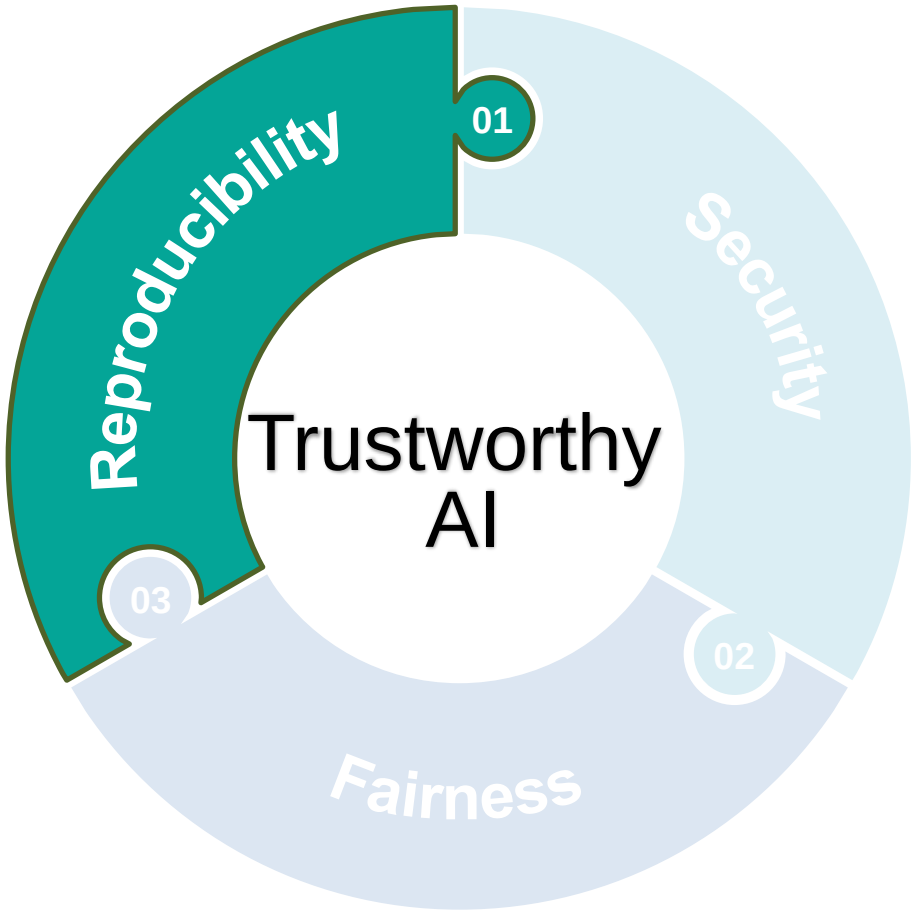
- **Compas** assessment is an algorithm used in the USA justice system to calculate the likelihood that someone will commit another crime. Eric L. Loomis, a man that was sentenced for eluding the police, was ranked as "high risk" to the community and handed down a six-year prison term.

  Compas calculates the likelihood of someone committing another crime and suggests what kind of supervision a defendant should receive in prison. The results come from a survey of the defendant and information about his or her past conduct.

Facing ethics in AI

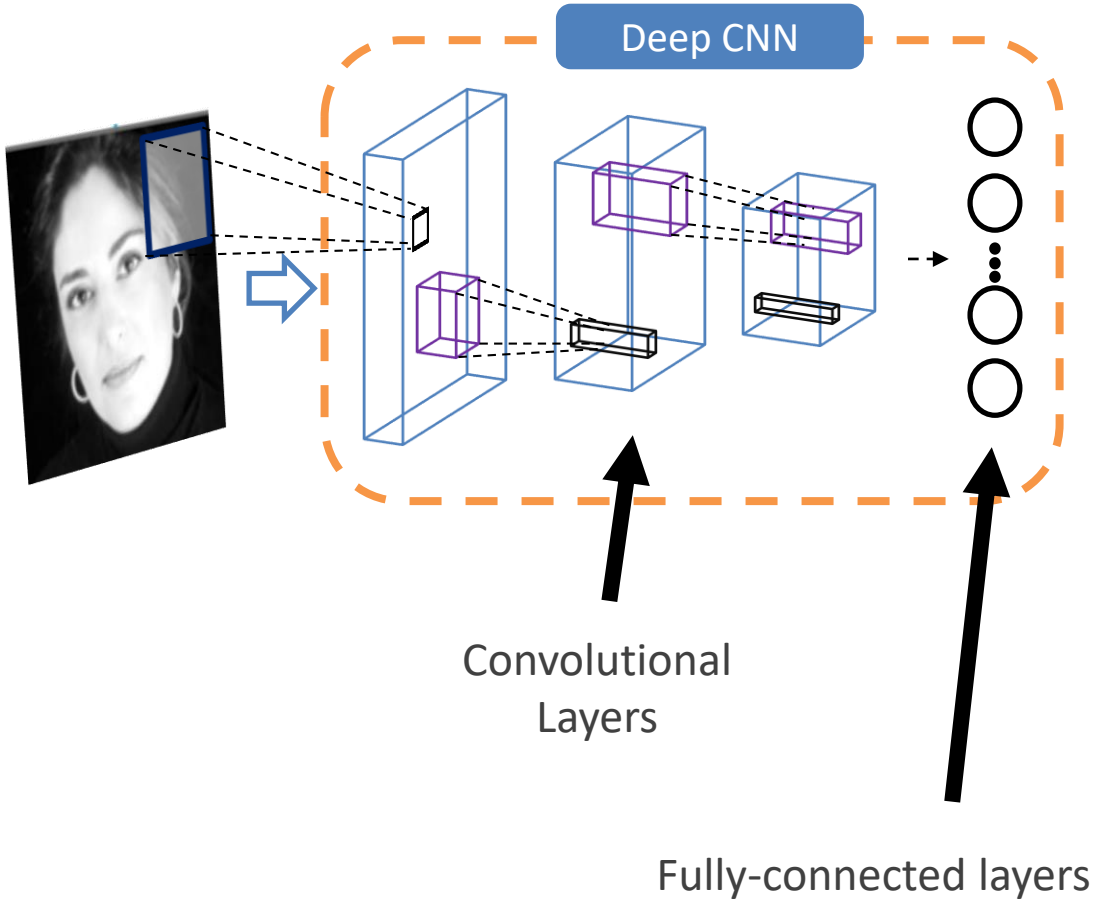
- It is clear that the problem is not in ML or CNN, but in humans
- **However**, ML is been using in several everyday tool (smartphone, ads, loan, etc)
- How can we be use not only that our privacy information will be not disclosed, but that they will not be used against us?

The Need for Reproducible Research



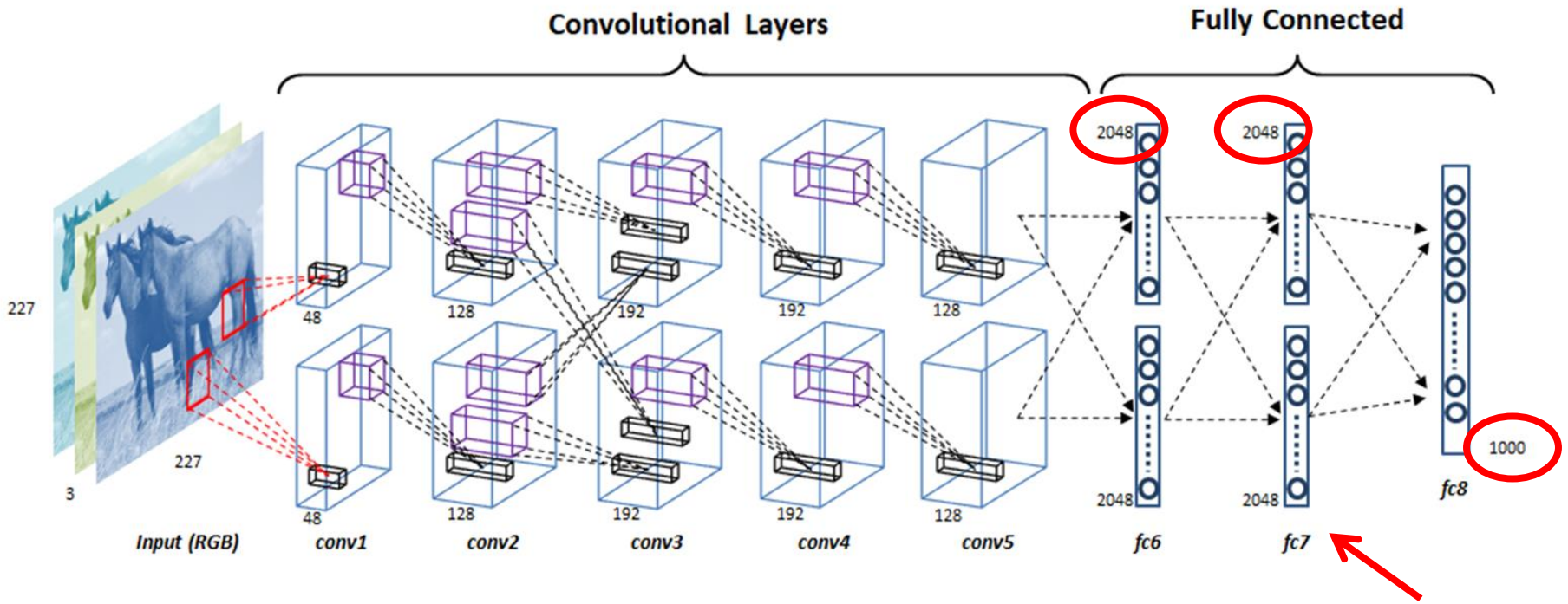
CNNs Knowledge Transfer

- Deep CNNs consist of two part:
  - Convolutional Layers, learning how to extract the best set of features for the task under analysis
  - Fully Connected Layers, using the learnt features to perform the actual classification
- It has been demonstrated that CNNs pre-trained on ImageNet (a famous competition with 1.5M training samples over 1000 classes) generalize very well also on very different tasks
- One of the most used approach is fine-tuning, consisting in replacing the classification layer with one having as many neurons as the number of classes in the new task



CNNs Fine-Tuning

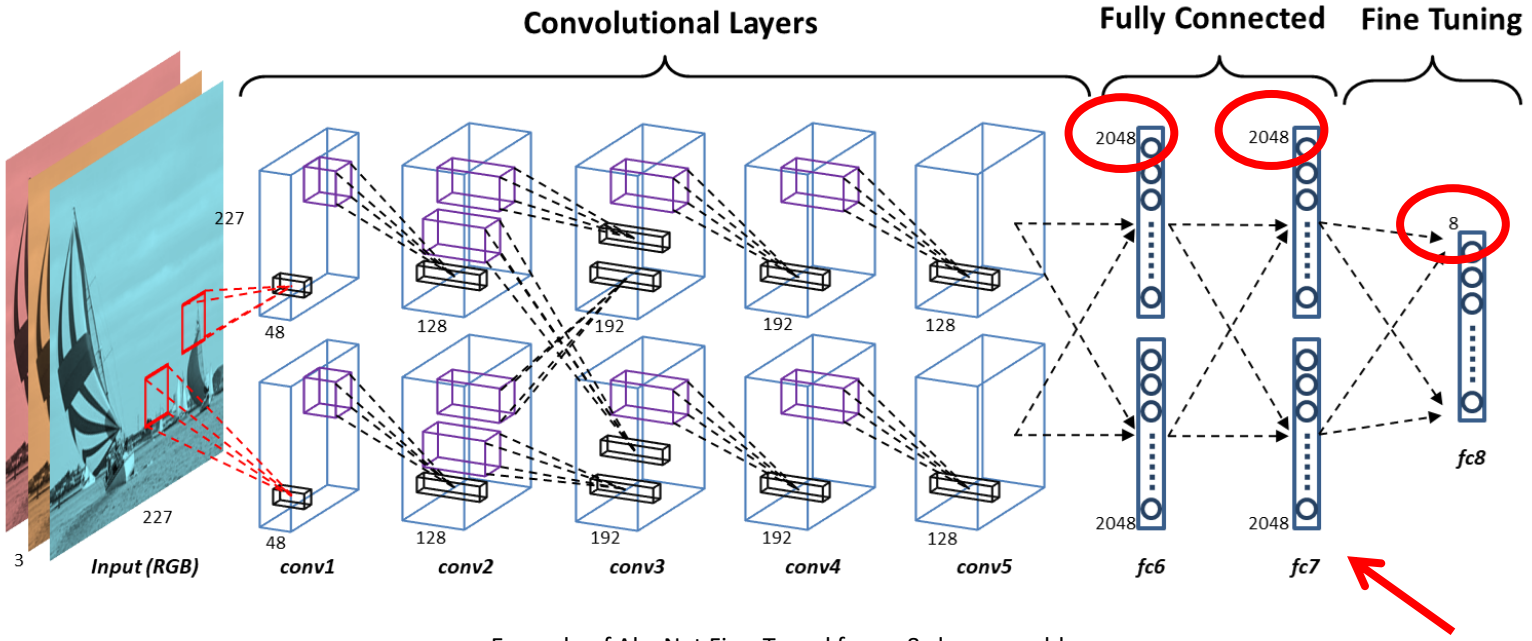
- However, fine-tuning implies a change in the architecture, without any concern on if and to what extent this alteration can impact the efficiency and effectiveness of the net



- AlexNet CNN architecture

CNNs Fine-Tuning

- However, fine-tuning implies a change in the architecture, without any concern on if and to what extent this alteration can impact the efficiency and effectiveness of the net
  - Fine-tuning is usually applied on problems having a smaller number of classes, causing a bottleneck in the network structure



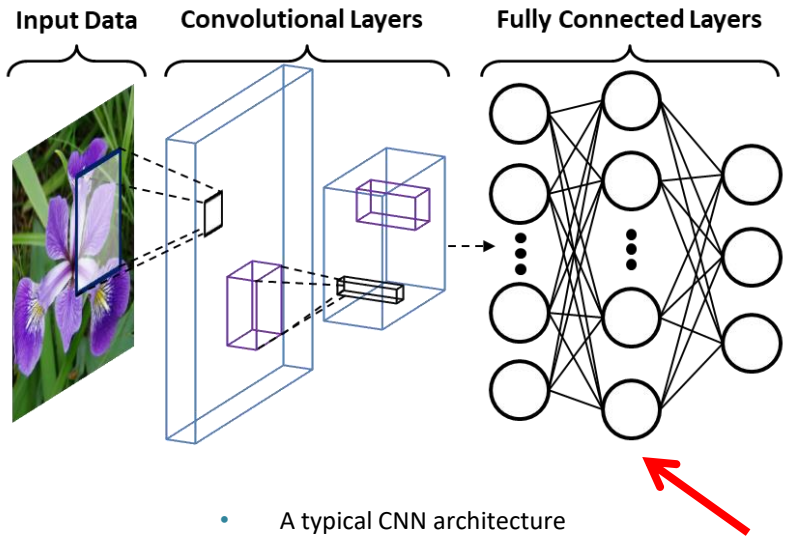
- Example of AlexNet Fine-Tuned for an 8 classes problem

Resiliency to Errors

- Many problems are resilient to small perturbations in the data, allowing settling on a near-optimum solution
  - This characteristic is very useful in situations where an *approximate computation* allows to perform the computation in less time or to deploy it on embedded hardware
- Deep learning is one of the fields that can benefit from approximate computing, thanks to its wide generalization ability
- **Intuition**: is it possible to remove whole neurons without substantially affecting the network performance in order to optimize fine-tuning?

Hidden Layer Sizing

- In the view of networks size reduction, we propose to further adapt deep CNNs by performing the **sizing of the hidden layer** when using the fine-tuning strategy
  - With the term sizing, we refer to the use of a suitable strategy to reduce the number of used neurons, without significantly affecting the network performance
- Among all the heuristic for choosing the number of neurons in neural networks, we investigate:
  - To use as many neurons as the average between the number of input and output neurons (hereafter referred as A-Rule)
  - Defined m as the number of classes and i as the number of input neurons, it is possible[7] to use $n = 2 * \sqrt{(m+2) * i}$ (hereafter referred as Huang)



A typical CNN architecture

[7] Huang, Guang-Bin. "Learning capability and storage capacity of two-hidden-layer feedforward networks." in IEEE transactions on neural networks 14.2 (2003): 274-281.

Experimental Setup

- To take into account the depth of the networks, we use two different CNNs pre-trained on ImageNet
  - AlexNet[2], consisting of 5 convolutional and of 3 fully connected layers, for a total of 60,965,224 parameters
  - Vgg19[8] consisting of 16 convolutional and of 3 fully connected layers, for a total of 143,667,240 parameters
- Since also the used optimizer could affect the evaluation, all the experiments were run by using both SGDM[9] and ADAM[10]
- We considered three datasets differing in terms of number of classes, number of samples and
- image resolution
  - The Dogs vs Cats dataset[11], consisting in 25000 images of cats and dogs equally distributed
  - The UIUC Sports Event dataset[12], containing images of 8 different sport activities distributed from 137 to 250 images each
  - The Caltech 101 dataset[13], collecting pictures of objects belonging to 101 different categories distributed from 40 to 800 each

[8] Simonyan, K. et al. “Very deep convolutional networks for large-scale image recognition.” in arXiv preprint arXiv:1409.1556 (2014)
[9] Bishop, Christopher M. “Pattern recognition and machine learning.” Springer (2006)
[10] Kingma, Diederik P. et al. “Adam: A method for stochastic optimization.” in arXiv preprint arXiv:1412.6980 (2014)
[11] https://www.microsoft.com/en-us/download/details.aspx?id=54765
[12] http://vision.stanford.edu/lijiali/event_dataset/
[13] http://www.vision.caltech.edu/Image_Datasets/Caltech101/

Results*

		#Parameters	Δ	Δ%
Dogs Vs Cats [49]	Base	56,876,418	-	-
	A-Rule	48,485,765	8,390,653	14.75%
	Huang	41,136,258	15,740,160	27.67%
UIUC Sports Event [102]	Base	56,901,000	-	-
	A-Rule	48,510,380	8,390,620	14.75%
	Huang	41,745,340	15,155,660	26.64%
Caltech-101 [51]	Base	57,282,021	-	-
	A-Rule	48,894,417	8,387,604	14.64%
	Huang	41,136,258	15,740,160	27.67%

- Summary of the number of AlexNet parameters for each sizing technique and considered dataset

		#Parameters	Δ	Δ%
Dogs Vs Cats [49]	Base	139,578,434	-	-
	A-Rule	131,187,781	8,390,653	6.01%
	Huang	123,838,274	15,740,160	11.28%
UIUC Sports Event [102]	Base	139,603,016	-	-
	A-Rule	131,212,396	8,390,620	6.01%
	Huang	124,447,356	15,155,660	10.86%
Caltech-101 [51]	Base	139,984,037	-	-
	A-Rule	131,596,433	8,387,604	5.99%
	Huang	123,838,274	15,740,160	11.28%

- Summary of the number of Vgg19 parameters for each sizing technique and considered dataset

		Memory (MB)	Δ (MB)	Δ%
Dogs Vs Cats [49]	Base	211	-	-
	A-Rule	177	34	16.11%
	Huang	124	87	41.23%
UIUC Sports Event [102]	Base	207	-	-
	A-Rule	176	31	14.98%
	Huang	152	55	26.57%
Caltech-101 [51]	Base	209	-	-
	A-Rule	178	31	14.83%
	Huang	166	43	20.57%

- Summary of the required AlexNet memory for each sizing technique and considered dataset

		Memory (MB)	Δ (MB)	Δ%
Dogs Vs Cats [49]	Base	508	-	-
	A-Rule	488	20	3.94%
	Huang	461	47	9.25%
UIUC Sports Event [102]	Base	507	-	-
	A-Rule	477	30	5.92%
	Huang	452	55	10.85%
Caltech-101 [51]	Base	506	-	-
	A-Rule	482	24	4.74%
	Huang	465	41	8.10%

- Summary of the required Vgg19 memory for each sizing technique and considered dataset

*For brevity reasons, we do not report the accuracy results for each test. All the results are reported in the thesis, where it is possible to verify that the sizing procedure does not affect the networks performance

Sizing and Adversarial Perturbations

- We analyze the impact that the sizing strategy has on CNN robustness against adversarial perturbations
  - The experiments were run on the UIUC Sports Event dataset[12]
  - FGSM[14] and DeepFool[3] have been used as adversarial perturbation approaches
- The robustness is measured as adversarial noise magnitude needed to mislead the CNN $\rho = \frac{\|Noise\|_2}{\|Image\|_2}$

	Fool	ρ		Time (s)	
		Mean	Median	Mean	Median
Base	FGSM	0.0183	0.0185	2690.45	2811.46
	DeepFool	0.0407	0.0419	706.83	703.33
A-Rule	FGSM	0.0193	0.0194	3260.09	3146.05
	DeepFool	0.0442	0.0452	722.80	754.96
Huang	FGSM	**0.0199**	**0.0202**	4745.09	4149.13
	DeepFool	**0.0488**	**0.0499**	558.67	554.58

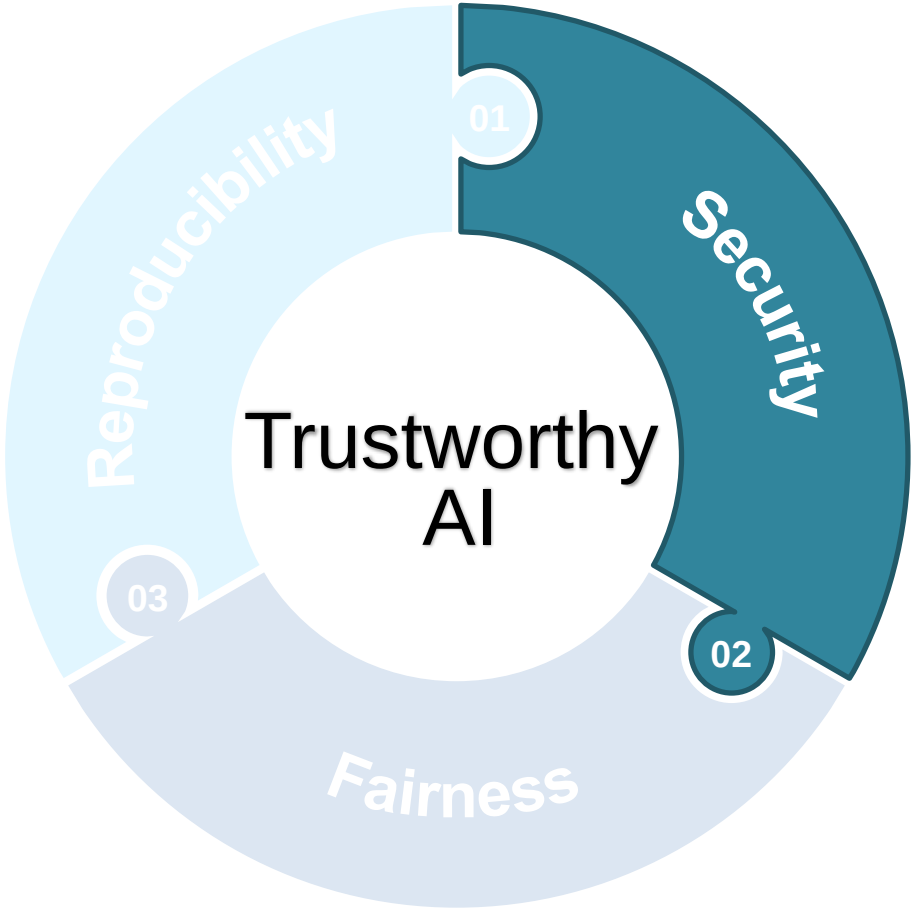
- AlexNet robustness to adversarial perturbations, varying the sizing approach

	Fool	ρ		Time (s)	
		Mean	Median	Mean	Median
Base	FGSM	0.0177	0.0182	18386.91	18634.11
	DeepFool	0.**1188**	0.0754	4305.76	4383.62
A-Rule	FGSM	0.0192	0.0190	27851.98	28652.63
	DeepFool	0.0713	0.0730	5151.23	5243.81
Huang	FGSM	**0.0200**	**0.0201**	22059.96	22084.38
	DeepFool	0.1000	**0.0882**	4132.82	4130.29

- Vgg19 robustness to adversarial perturbations, varying the sizing approach

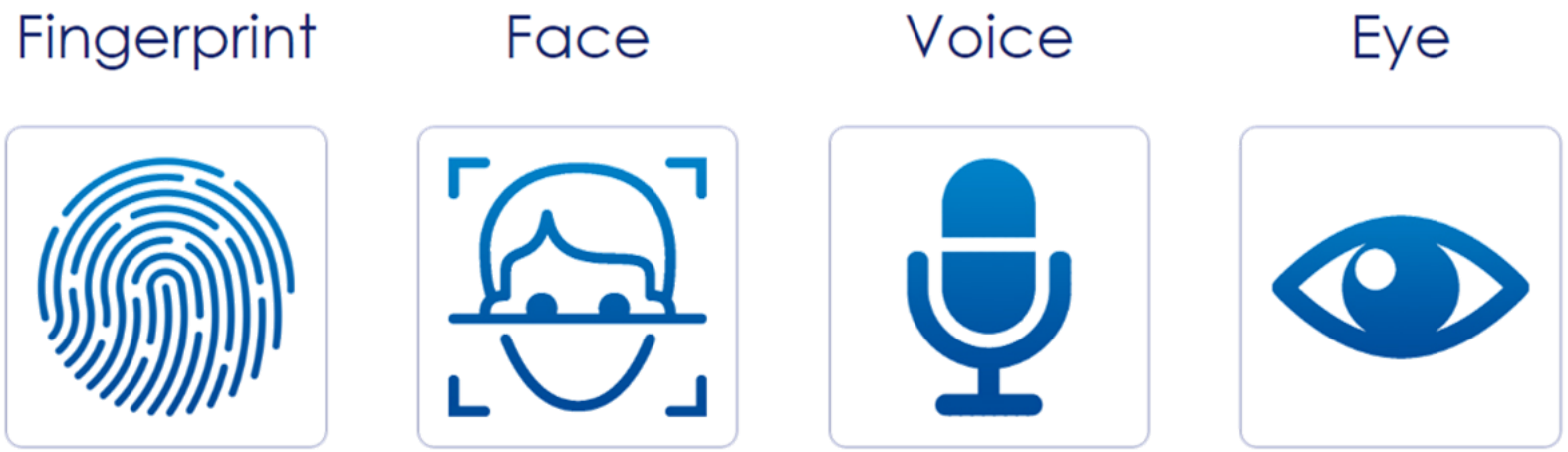
[14] Goodfellow, I. J. et al. “Explaining and harnessing adversarial examples.” in arXiv preprint arXiv:1412.6572 (2014)

Security Critical Applications



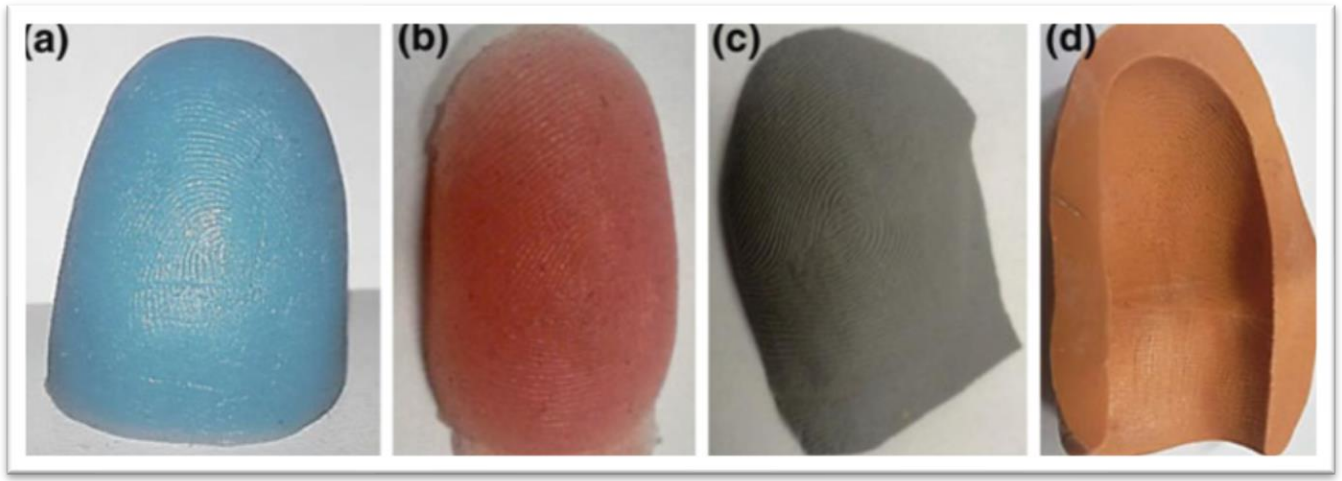
Biometric based Authentication Systems

- Biometric Based Authentication Systems (BASs) allows to recognize a subject according to “what they are” rather then on “what they use”
  - Among all, fingerprints are the most used (Fingerprint based Authentication Systems – FAS)
  - As for other domains, CNN has proved to be very effective in analyzing biometrics data
- Given CNNs blindspots, might they make BAS less robust against malicious agents?



Presentation attack

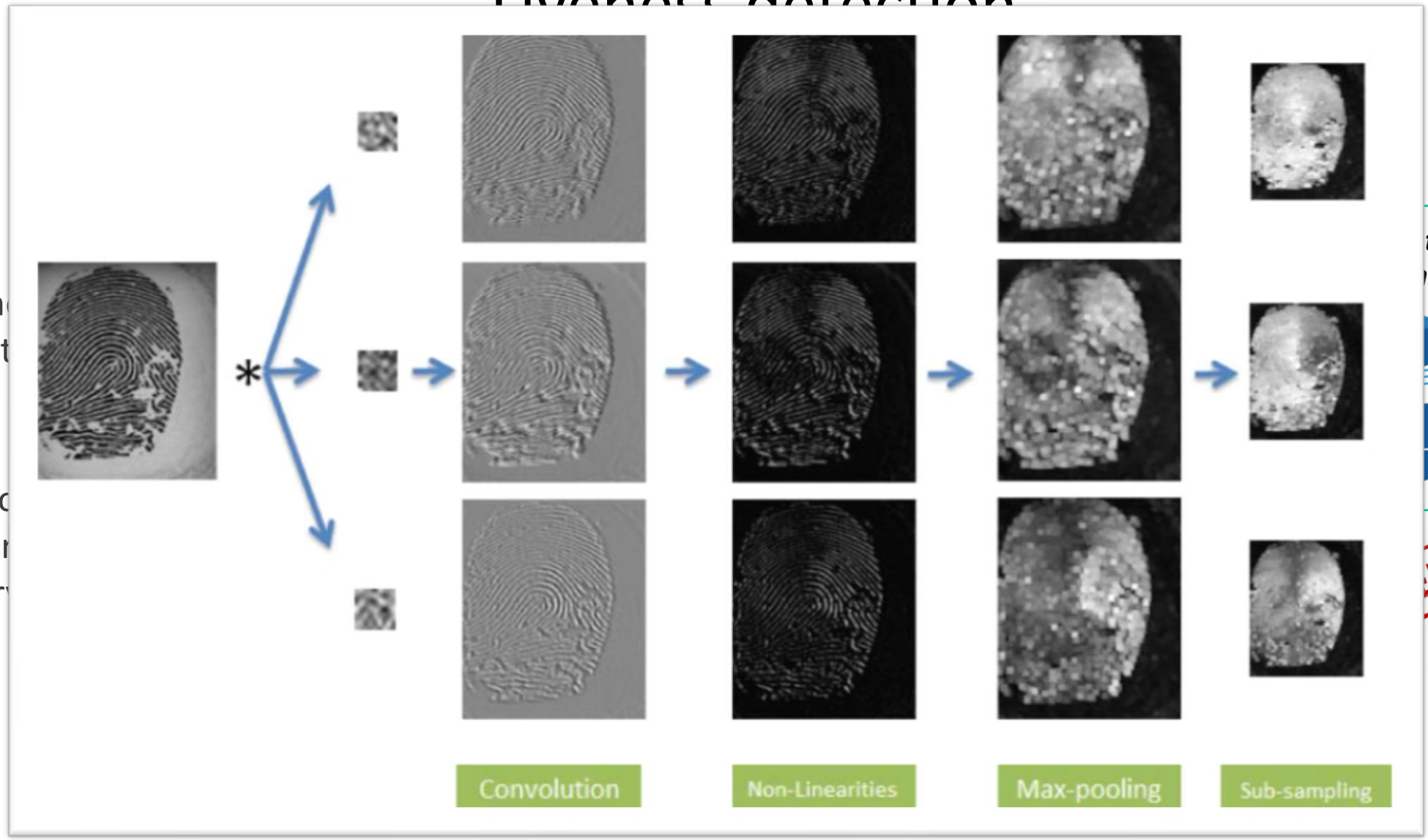
- As for any other authentication means, it is possible to attack a BAS by using a counterfeit replica of the target subject biometrics
- Since the attack consists in “presenting” the fake replica to the scanner, this type of attacks goes under the name of *Presentation Attack*



- Artificial finger replicas made using GLS (a), Ecoflex (b), Liquid Ecoflex (c) and Modasil (d)

Liveness detection

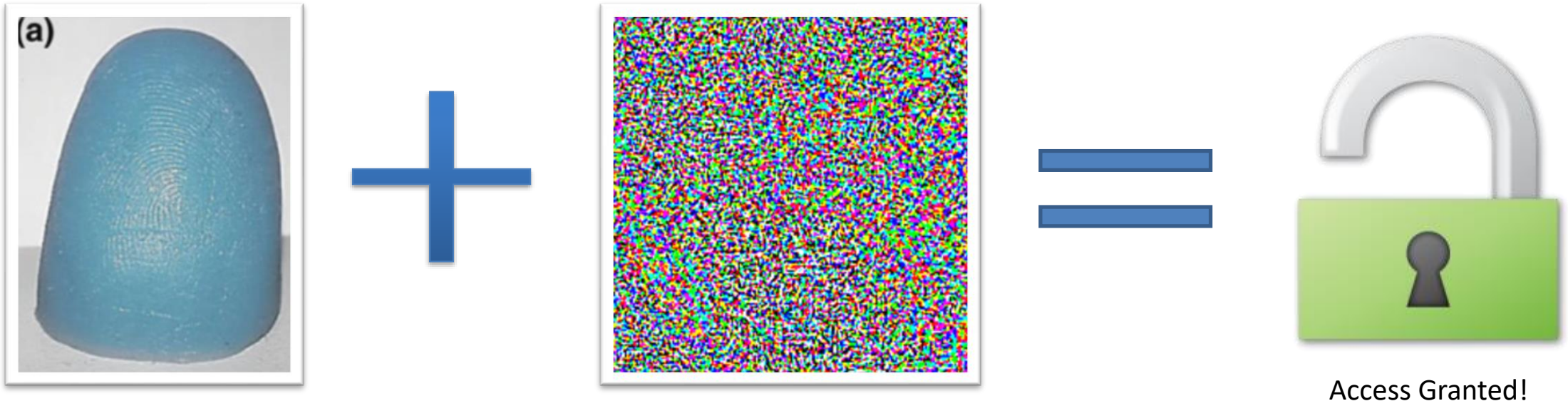
- Liven
  count
- One c
  pre-tr
  binar



[15] Nogueira et al. "Fingerprint Liveness Detection using Convolutional Networks", in IEEE Transactions on Information Forensics and Security (2016)

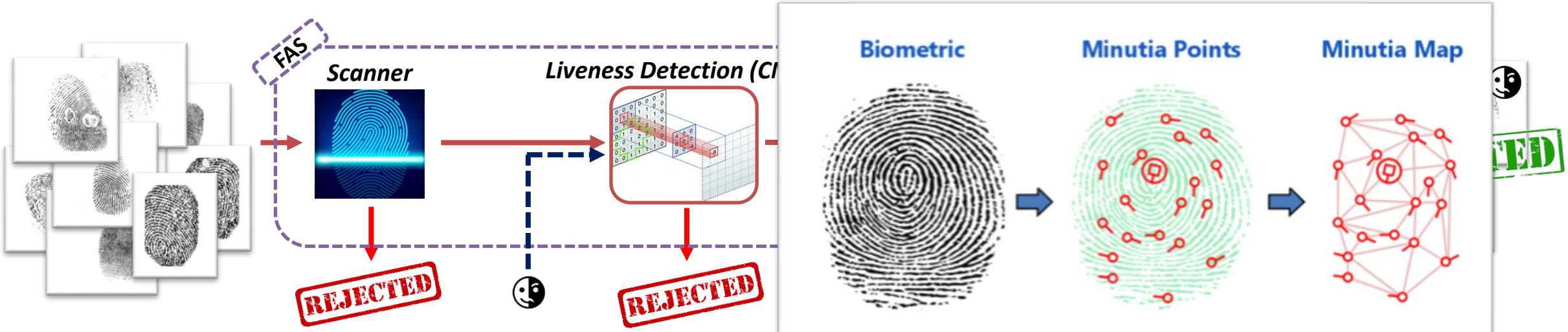
Adversarial presentation attack

- Can we exploit adversarial perturbation to alter a fake fingerprint acquisition such that it mislead a CNN liveness detector?



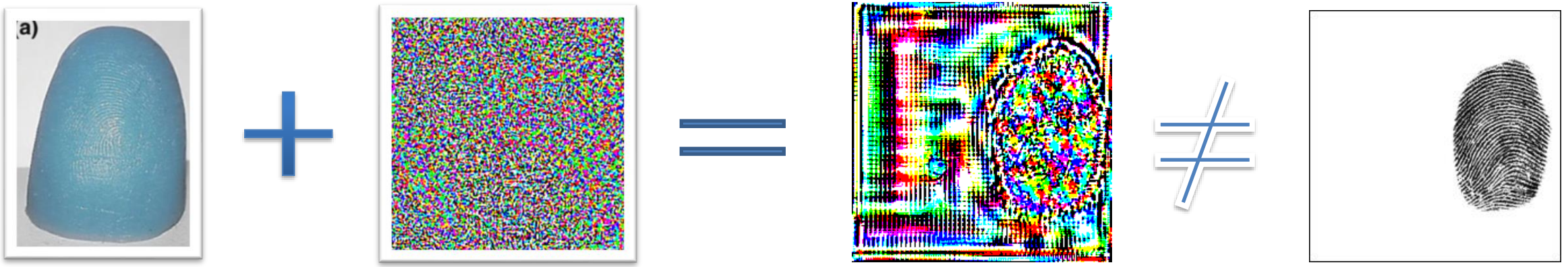
Adversarial presentation attack

- The aim is to understand if and to what extent adversarial perturbation can affect a FAS , tying to answer to:
  - how vulnerable are CNN based liveness detectors to adversarial perturbation?
  - Is the counterfeit footprint still recognized by the authentication system or had its main characteristics been destroyed by the perturbation attack?



A Constrained attack

- Fingerprints images are really different from natural images and the injected noise could be very visible and difficult to hide



Constrained Adversarial Perturbation Attack

- Only gray-level perturbation allowed
- Only pixel within a Region of Interest

Experimental setup

- We considered three perturbation approaches
  - Fast Gradient Sign Method[14]
  - DeepFool[3]
  - Evolutionary Approach[16]
- The used Liveness Detectior (LD) is the LivDet 2015 winner[15], based on a pre-trained VGG19
- As Authentication System (AS), we used a famous minutiae based approach[17]

- The dataset used was the one from LivDet 2015[18]
  - Training-set to train the LD and prepare the AS database
  - Test-set to measure the attack effectiveness

Scanner	Image Size (px)	Live	Body Double	Ecoflex	Gelatine	Latex	Liquid Ecoflex	OOMOO	Playdoh	RTV	Woodglue
Biometrika	1000x1000	1000	-	250	250	250	250	-	-	250	250
CrossMatch	640x480	1500	300	270	300	-	-	297	281	-	-
DigitalPersona	252x324	1000	-	250	250	250	250	-	-	250	250
GreenBit	500x500	1000	-	250	250	250	250	-	-	250	250

- LivDet2015 dataset charactetistics

[16] Su et al. “One pixel attack for fooling deep neural networks”, in arXiv:1710.08864 (2017)
[17] Joshua, et al. “Fingerprint matching using a hybrid shape and orientation descriptor” (2011)
[18] Mura, V. et al. “Livdet 2015 fingerprint liveness detection competition” in IEEE International Conference on Biometrics Theory, Applications and Systems (2015)

Attack Results*

	FGSM			DeepFool			Evolutionary		
Scanner	LD (%)	AS (%)	R (%)	LD (%)	AS (%)	R (%)	LD (%)	AS (%)	R (%)
Biometrika	2.67	86.84	3.15	62.15	73.67	62.15	3.58	*96.08*	4.67
CrossMatch	2.45	85.71	2.82	35.06	72.20	33.93	7.22	*91.26*	8.83
DigitalPersona	2.16	93.55	2.74	1.60	69.57	1.51	8.78	*94.44*	11.25
GreenBit	1.92	96.30	1.90	52.92	68.01	37.04	65.79	*98.16*	66.47

- Attack success rates against the Liveness Detector (LD%) and the Authentication System (AS%- evaluated with reference to the number of images that have passed the Liveness Detector) and the ratio (R%) between the number of successful adversarial perturbed fingerprints (against LD and AS) over the number of successful clean fake fingerprints (against only the AS).

FGSM

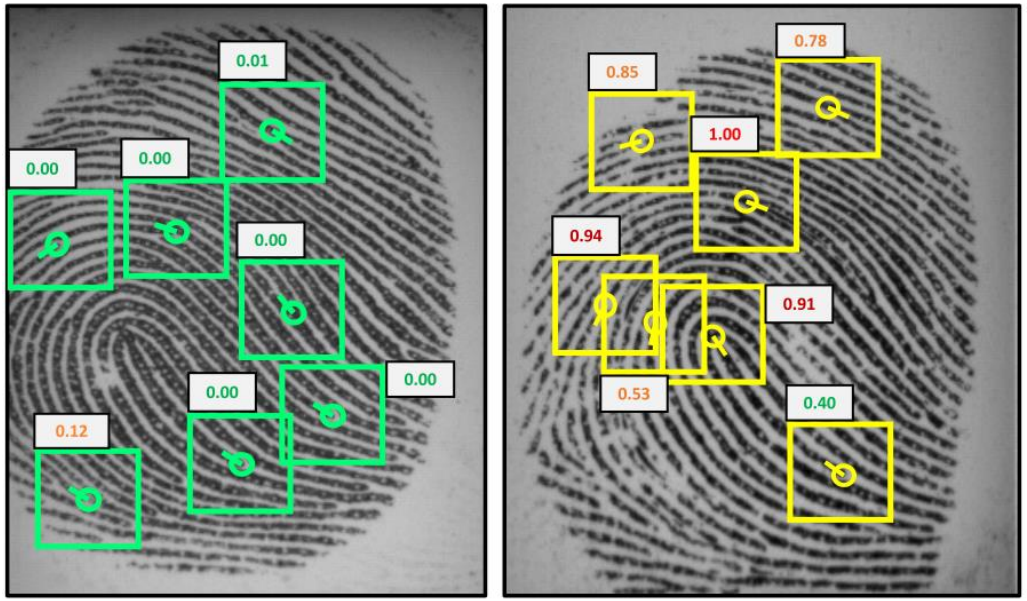
DeepFool

Evolutionary

*For brevity reasons, we do not report examples for unsuccessful attacks. The full list of results and examples is available in the thesis

Transferring Perturbation Attack

- Recently, a new CNN-based LD has been proposed[19]
  - From each fingerprint minutiae, the approach extract a patch
  - Each patch is rotated according to the minutiae orientation
  - A CNN is fine-tuned on the patches
  - A "spoofness score" is determined for the fingerprint bases on each patch classification score
- The patch-based approach makes the adversarial perturbation attack harder to perform and more time and computational demanding
- **Intuition**: is it possible to craft the adversarial noise against a simpler LD before transferring it against a different target LD?



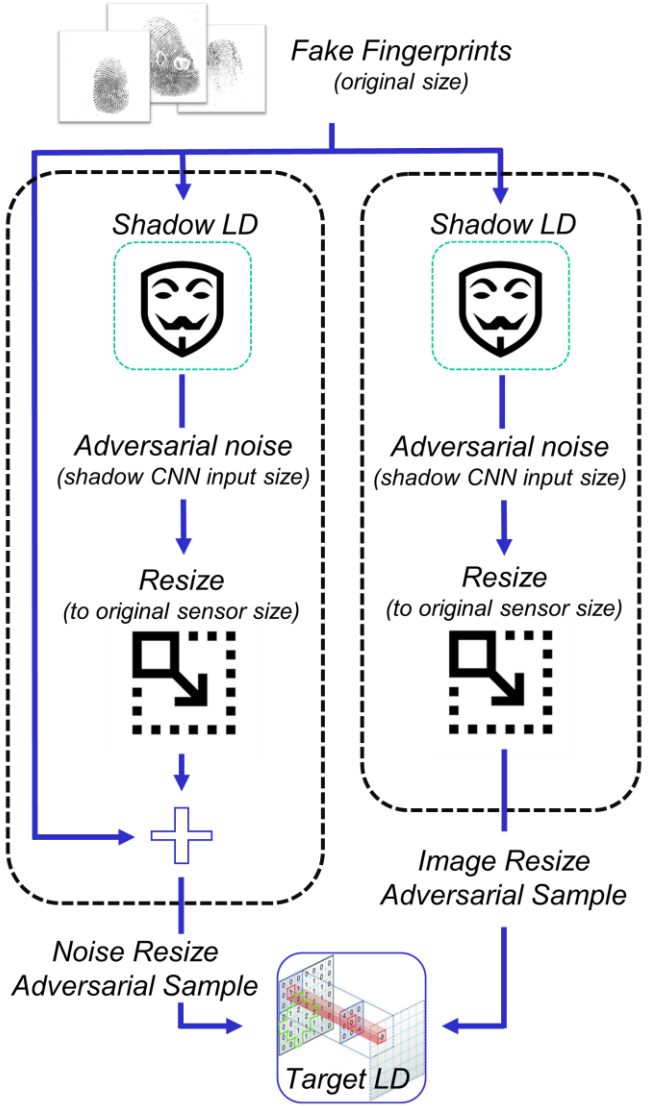
- Example of minutiae and respective orientations and spoofness scores, for a live (left) and fake (right) fingerprint

[19] Chugh, T. et al. "Fingerprint spoof buster: Use of minutiae-centered patches." in IEEE Transactions on Information Forensics and Security (2018)

Attack Scenario

- We want to analyze whether it is possible to transfer a perturbation across different CNN liveness detector
  - To this aim, we define the *target* LD as the model we want to attack (black box)
  - And a *shadow* LD, created by the attacker and used to craft the adversarial noise

- There are two viable approaches to obtain an adversarial sample having the scanner output size
  - **Image resize**, in which the attacker directly resizes the crafted adversarial sample
  - **Noise resize**, in which the attacker resizes only the adversarial noise, adding it to the original fingerprint acquired by the scanner



Transfer Attack Results*

- The effectiveness of the attack is related to the scanner more than to the used perturbation algorithm
- This behavior is expected since, although all optical, each LivDet2015 scanner have different characteristics (e.g. sensor, lens, acquisition plate, etc.) that result in distinct artefacts in the acquired fingerprints

Scanner	FGSM	Deep
Biometrika	3,16%	2,9
CrossMatch	96,88%	97,8
DigitalPersona	3,51%	3,6
GreenBit	19,99%	20,9

- Transfer adversarial attack success probability against the target LD under the "image resize" scenario, for each scanner and for each adversarial perturbation approach

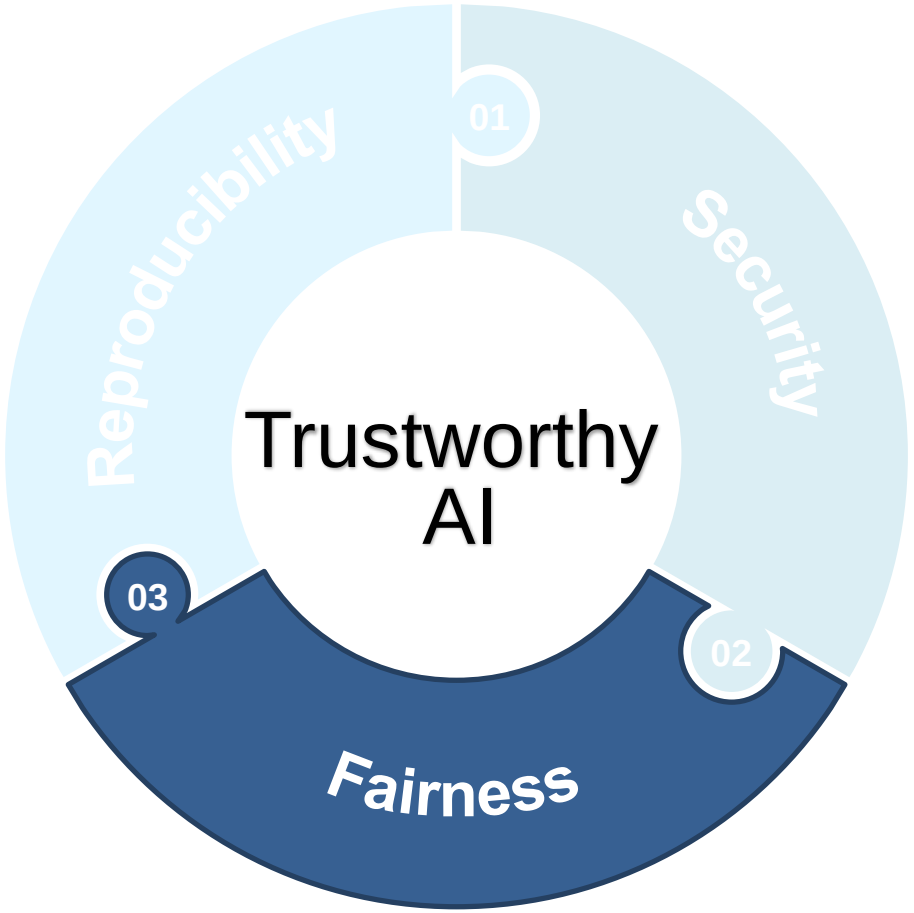
- Example of different geometric distortions introduced by acquiring the same fingerprint by using two different scanners: GreenBit (left) and DigitalPersona (right)

FGSM	DeepFool	OnePixel
1,08%	1,15%	0,00%
0,07%	1,16%	99,93%
0,00%	0,00%	0,00%
0,00%	0,07%	0,00%

- ess probability against the target LD under the "noise resize" scenario, for each scanner and for each adversarial perturbation approach

*For brevity reasons, we do not report the results per spoofing materials. The full list of results and examples can be found in the thesis.

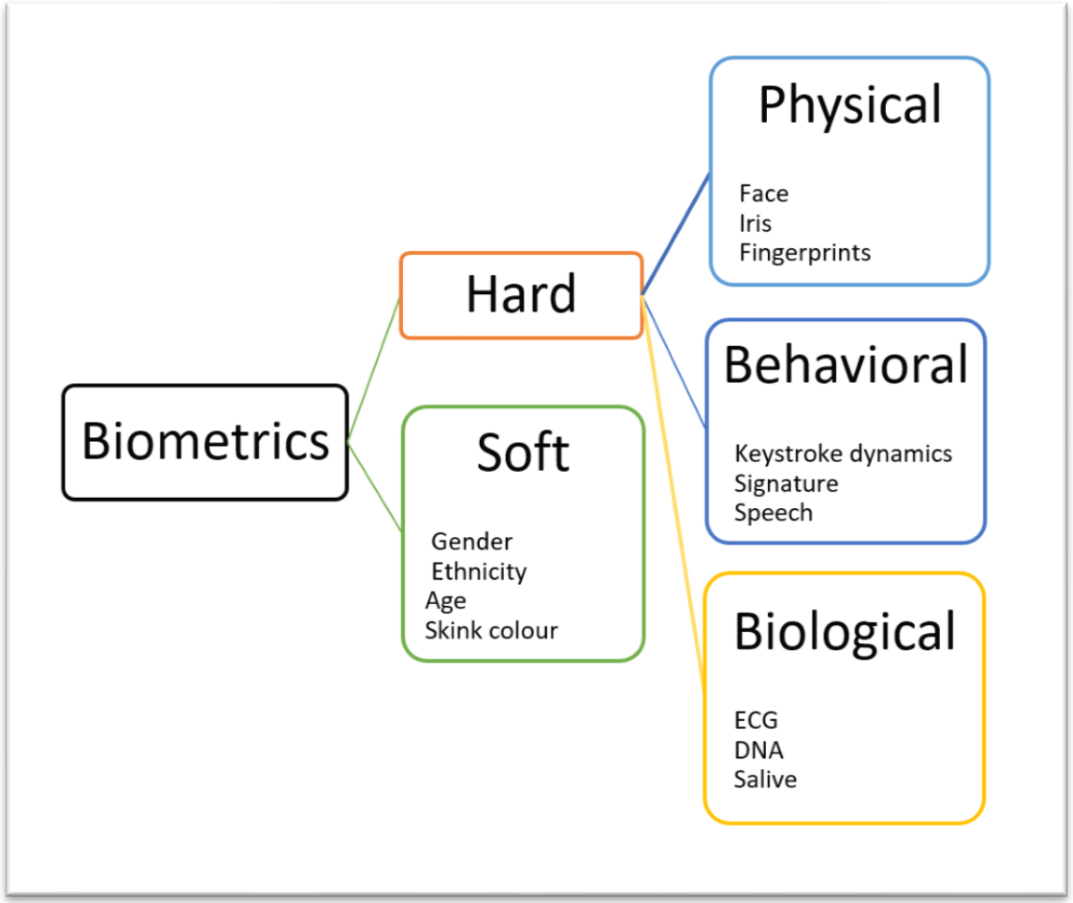
Adversarial Approaches for Ethical AI



Soft biometrics

- Over the years, researchers' interest shifted from hard to soft biometrics
  - at the beginning, mainly with the aim of improving authentication system effectiveness
  - then focusing on subject identification

Could the automatic processing of this data rich in sensitive information expose users to privacy threats associated to their unfair use (i.e. gender or ethnicity)?



Unfair Face Analysis

- These privacy threats are usually perceived as far from us or able to affect only our digital alter-ego
  - Unfortunately, there are many very effective and sneaky (i.e. not perceived by users) attacks able to extract our soft biometrics with a single glance in a real environment, maybe also without our explicit consensus
- Indeed, face analysis is a particularly sensitive topic even without involving AI
  - Indeed, it has been proven that attractive people tends to get more financial and social benefit
  - What if AI learn to do the same? … **… to late!!![20]**

Technology

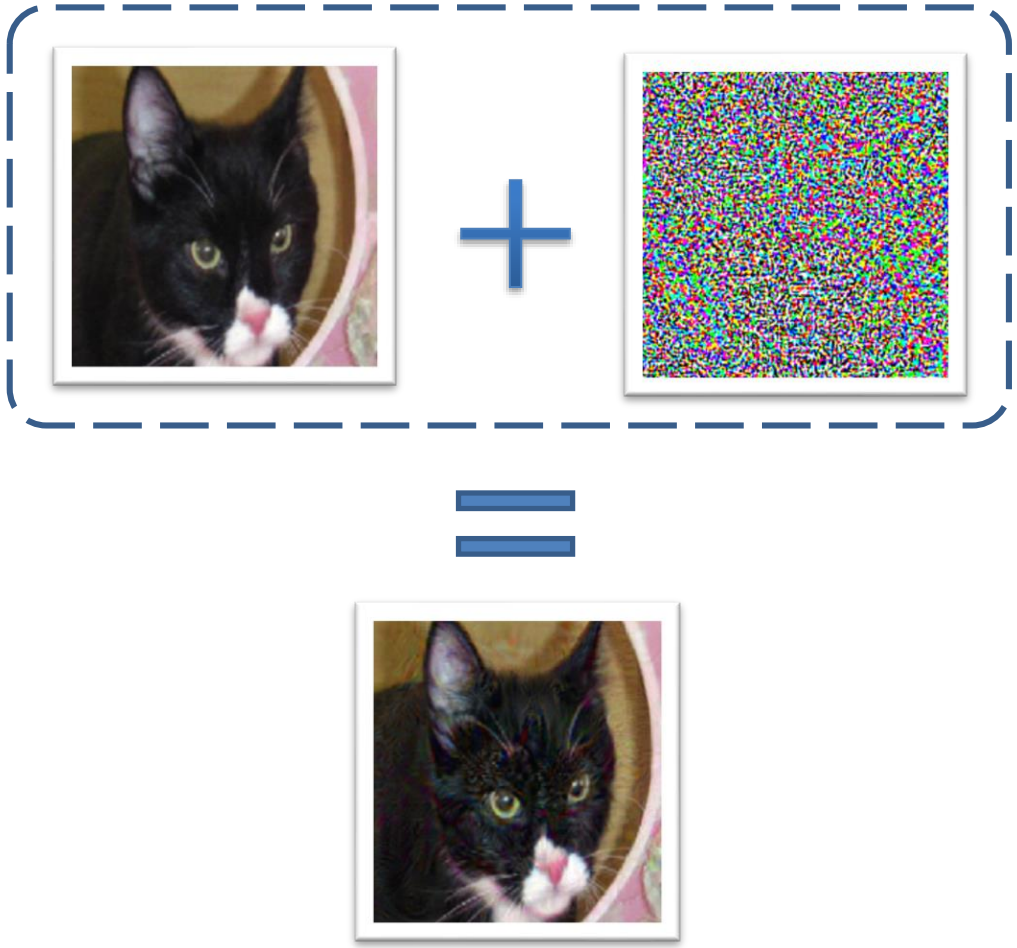
A face-scanning algorithm increasingly decides whether you deserve the job

HireVue claims it uses artificial intelligence to decide who's best for a job. Outside experts call it 'profoundly disturbing.'

[20] https://www.washingtonpost.com/technology/2019/10/22/ai-hiring-face-scanning-algorithm-increasingly-decides-whether-you-deserve-job/

Exploiting adversarial perturbations

- By definition, an adversarial perturbation should be as invisible as possible, and this constraint is usually met since the injected noise is distributed over the whole image
- Can we **trade** a very visible perturbation in exchange for having the opportunity to apply it on a very limited portion of the image?
- The idea is to **create an adversarial patch** that, once printed, is **able to fool CNNs by simply "wearing" it**, for example in the shape of a sticker, a clip or a pendant



Adversarial against ethnicity classifiers

- The proposed approach consists of 5 steps:

1. A mask is generated to force the perturbation algorithm to work only in a restricted region of the image
2. An adversarial perturbation is determined for the first image, over the previously generated mask
3. The second point is repeated for all the remaining images, by starting, for each image, from the perturbation calculated over the immediately preceding image
4. The mask is randomly moved in order to generate an adversarial patch invariant to its position
5. Steps 3 and 4 are repeated (including also the first image) until the perturbation is able to tamper all the images ethnicity, or until a termination condition (such as the maximum number of iterations) is met

```
mask = createMask();
pert = randPert(mask);
pertCount, iterCount = 0;
while pertCount ≤ ths & iterCount ≤ maxIter do
    pertCount = 0;
    for Image img in Dataset do
        mask,pert = randomMove(mask,pert);
        pert = calculatePert(pert, mask, img);
        if classify(img) ≠ classify(img + pert) then
            pertCount++;
        end
        maxIter++;
    end
end
```

Algorithm 1: General adversarial patch creation. Please note that the maxIter and ths (threshold) values were not set to highlight that they are user-defined parameters.

Experimental setup

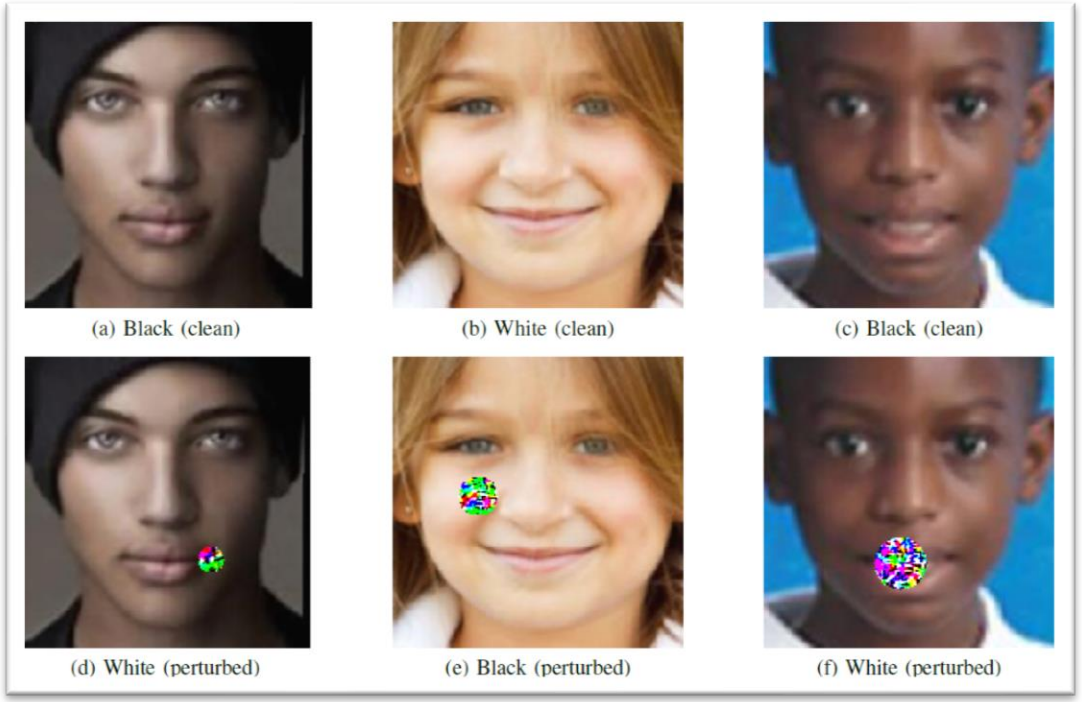
- As a case of study, we will focus on the generation of a general adversarial patch specifically designed to fool a CNN for ethnicity recognition in a real-world application
  - As dataset we considered UTKFace[21], a publicly available large-scale face dataset containing over 20,000 images
  - We used a VGG16[14] based ethnicity recognition system, able to obtain 95.59% of accuracy on the test set

- Images from the UTKFace dataset. Please note the variety of pose, illumination, age, resolution, expression and accessories

[21] https://susanqq.github.io/UTKFace/

Results*

- Although divided into five different ethnic clusters, results are reported only on the ’Black’ vs. ’White’ task
- *It is worth noting that we did not impose any further restriction on subject age, pose, expression, illumination and occlusion, in order to obtain a model and an adversarial patch able to work in real environmental conditions*



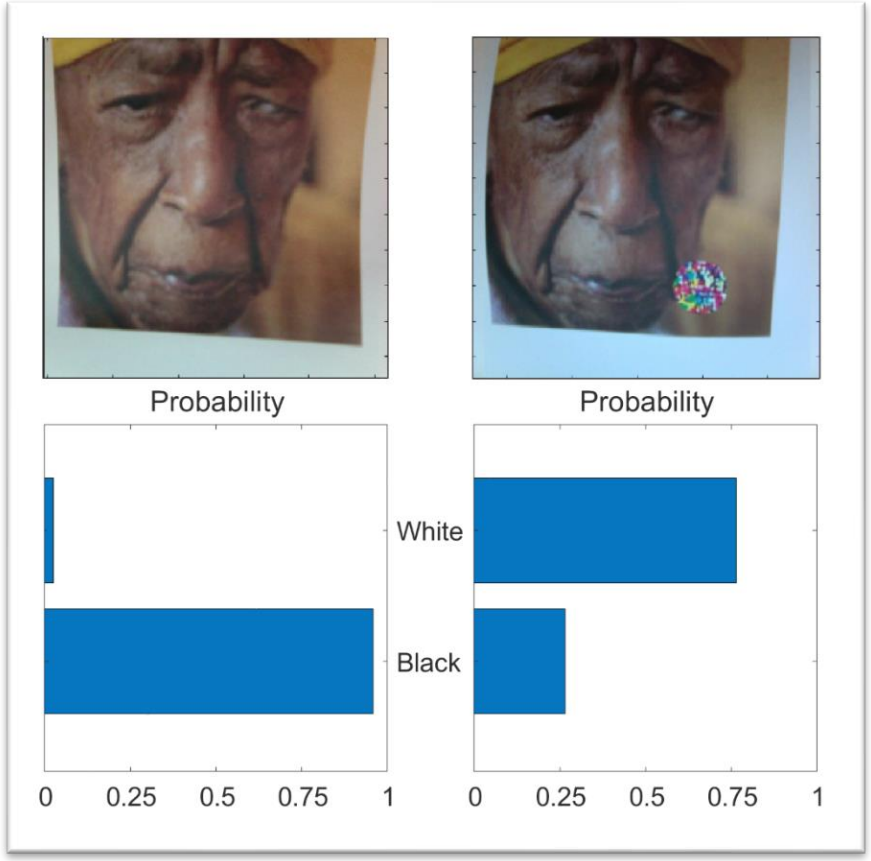
- Some illustrative attacks. Top row reports clear images, bottom row reports attacked samples

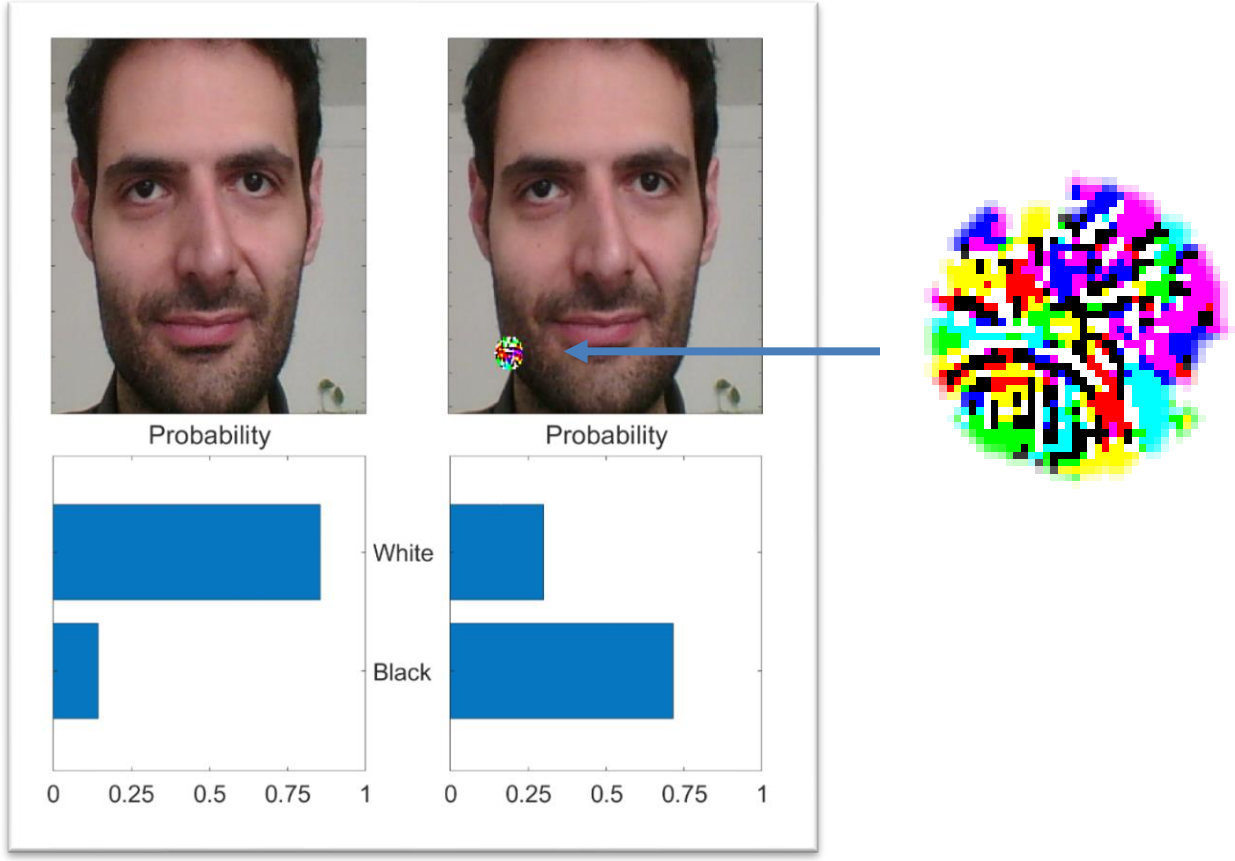
Ethnicity	Circle Radius 10	15	20
Black	67.81%	88.56%	95.75%
White	97.49%	98.82%	99.95%

- Success rate of the adversarial patch varying the patch radius, for each considered ethnicity

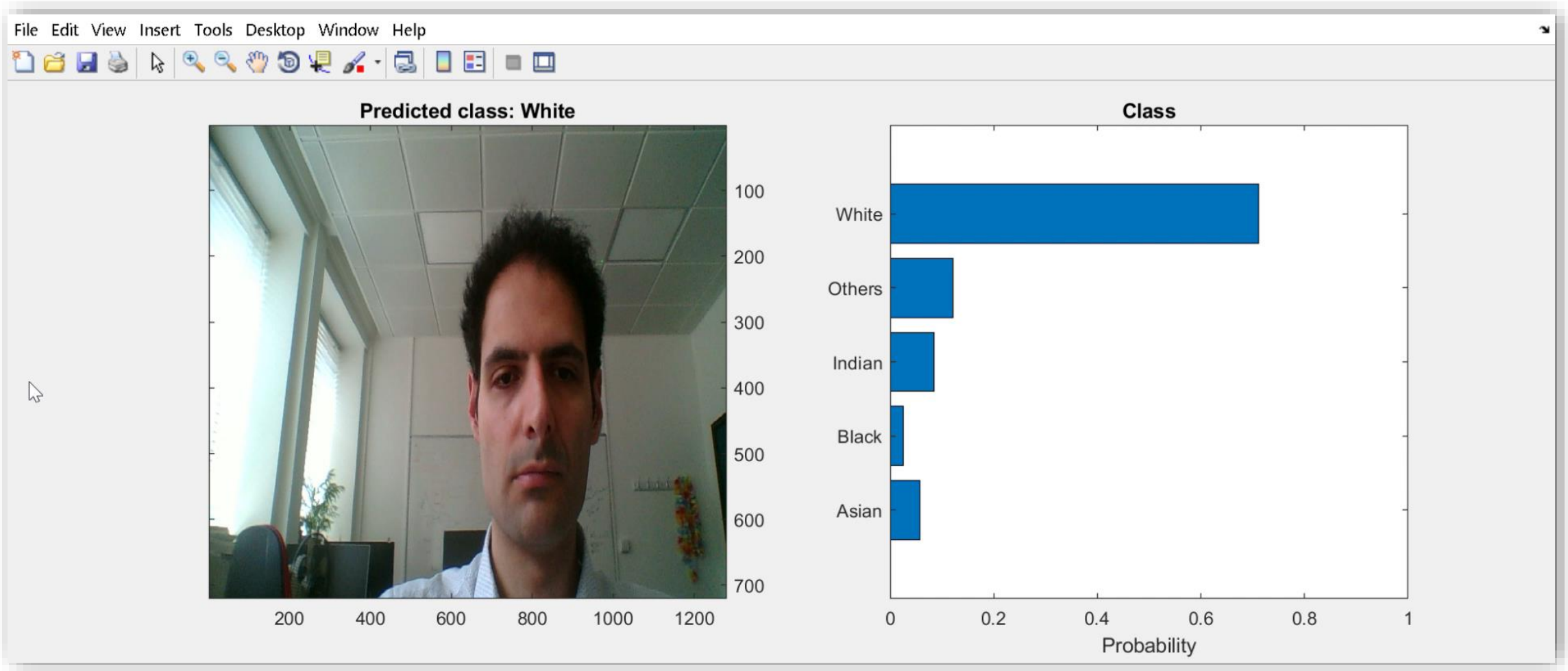
*For brevity reasons, we report only some examples. The full results list can be found in the thesis

Results in the Real World





Results in the Real World



Ending Notes

Discussions

- The spread of AI in critical domains (e.g. facial recognition, biometric verification, etc.) rises questions related to the **consequences that its misuse** (malicious or not) **can lead to**

- **AI is not to blame** since, being just a tool, the consequences resulting from its misuses can not be accounted to the medium, but must be instead attributed to its operator*

- Nonetheless, since AI is very likely to be an important part of our everyday life in the very next future, **it is crucial to build trustworthy AI systems**

  - Therefore***, in this thesis we tried to make a step towards the crucial need for raising awareness about security, ethical and fairness threats associated with AI systems***, from a technical perspective as well as from the governance and from the ethical point of view

- In conclusion, **AI represents** without any doubt **one of the greatest achievement made by humans**

  - However, since “with great power comes great responsibility”, we must learn how to properly use it, developing methods and enacting laws that support its fair, secure and ethical usage for all people around the world

*In the thesis. chapter 2 provides our point of view about why AI agents should be analyzed trough the lens of deontology

Honors and Roles Held

- Winner of a Titan XP GPU within the NVIDIA GPU Grant program
- Chair of IEEE Student Branch Federico II (2018/2019)
- MATLAB Student Ambassador (2019/2020)
- Guest Editor for the Special Issue ”Intelligent Innovations in Multimedia Data” of MPDI Future Internet Journal
- Chair of the 1st e-BADLE (eHealth in the Big Data and Deep Learning Era) workshop
- Program Committee for CMBS2019 - HealthCare 4.0
- Program Committee for BIBM2018 - Computational methods for Hospital 4.0
- Reviewer for several International Conferences and Journal, including IET Biometrics, IJCNN, ICIAP
- Scholarship for the AI-DLDA Ph.D. School

Publications

1. Piantadosi G., Marrons S., Galli A., Sansone M., Sansone C. (2019) DCE-MRI Breast Lesions Segmentation with a 3TP U-Net Deep Convolutional Neural Network. In: IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS)
2. Marrone S., Olivieri S., Piantadosi G., Sansone C. (2019) Reproducibility of Deep CNN for Biomedical Image Processing Across Frameworks and Architectures. In: 27th IEEE European Signal Processing Conference (EUSIPCO)
3. Galli A., Gravina M., Marrone S., Piantadosi G., Sansone M., Sansone C. (2019) Evaluating Impacts of Motion Correction on Deep Learning Approaches for Breast DCE-MRI Segmentation and Classification. In: International Conference on Computer Analysis of Images and Patterns (CAIP)
4. Gravina M., Marrone S., Piantadosi G., Sansone M., Sansone C. (2019) 3TP-CNN: Radiomics and Deep Learning for Lesions Classification in DCE-MRI. In: International Conference on Image Analysis and Processing (ICIAP)
5. Marrone S., Piantadosi G., Moscato V., Sansone C. (2019) Developing a Smart PACS: CBIR System Using Deep Learning. In: Journal of Computers & Electrical Engineering (under review)
6. Galli A., Marrone S., Piantadosi G., Sansone M., Sansone C. (2019) 3TP U-Net: Leveraging Tracer Kinetic in DCE-MRI Breast Lesion Segmentation. In: Artificial Intelligence in Medicine (under review)
7. Buizzaa C., Quilodran Casas C., Nadler P., Mack J., Titus Z., Le Cornec C., Heylen E., Dur T., Baca Ruiz L., Heaney C., Lopez J.A.D., Marrone S., Arcucci R. (2020) Data Learning: Integrating Data Assimilation with Machine Learning. In: Journal of Computational Science (under review)
8. Marrone S., Papa C., Sansone C., (2020) Effects of Hidden Layer Sizing on CNN Fine-Tuning. In: Future Generation Computer Systems (under review)
9. Marrone S., Piantadosi G., Sansone M., Sansone C. (2018) Reproducibility of Deep Convolutional Neural Networks Approaches. In: RRPR Workshop Proceedings
10. Hesham Elhalawani, Timothy A Lin, Stefania Volpe, Abdallah S.R. Mohamed, Aubrey L. White, James Zafereo, Andrew Wong, Joel E. Berends, Shady Abohashem, Bowman Williams, Jeremy M. Aymard, Aasheesh Kanwar, Subha Perni, Crosby D. Rock, Luke Cooksey, Shauna Campbell, Pei Yang, Khanh Nguyen, Rachel Ger, Carlos Eduardo Cardenas, Xenia Fave, Carlo Sansone, Gabriele Piantadosi, Stefano Marrone, Rongjie Liu, Chao Huang, Kaixian Yu, Tenfei Li, Yang Yu, Youyi Zhang, Hongtu Zhu, Jeffrey S. Morris, Veerabhadran Baladandayuthapani, John W. Shumway, Alakonanda Ghosh, Andrei Pöhlmann, Hady Ahmady Phoulady, Vibhas Goyal, Guadalupe Canahuate, G. Elisabeta Marai, David Vock, Stephen Y. Lai, Dennis S. Mackin, Laurence E. Court, John Freymann, Keyvan Farahani, Jayashree Kalpathy-Cramer and Clifton D FullerIn (2018) Machine Learning Applications in Head and Neck Radiation Oncology: Lessons from Open-Source Radiomics Challenges. In: Frontiers in Radiation Oncology
11. Piantadosi G., Marrone S., Fusco R., Sansone M., Sansone C. (2018) Comprehensive computer-aided diagnosis for breast T1-weighted DCE-MRI through quantitative dynamical features and spatio-temporal local binary patterns. In: IET Computer Vision, Volume 12, Issue 7, October 2018, p. 1007 – 1017
12. Marrone S., Sansone C., (2019) Approximate Computing for Sizing Hidden Layer in CNN. In: 4th Workshop on Approximate Computing (AxC)
13. Amato F., Marrone S., Moscato V., Piantadosi G., Picariello A., Sansone C. (2019) HOLMeS: eHealth in the Big Data and Deep Learning Era. In: Information 2019, 10, 34. MDPI
14. Marrone S., Piantadosi G., Sansone M., Sansone C. (2017) Look-Up Tables for Efficient Non-Linear Parameters Estimation. In: Sforza A., Sterle C. (eds) Optimization and Decision Science: Methodologies and Applications. ODS 2017. Springer Proceedings in Mathematics & Statistics, vol 217. Springer, Cham.
15. Marrone S., Piantadosi G., Fusco R., Petrillo A., Sansone M., Sansone C. (2017) An Investigation of Deep Learning for Lesions Malignancy Classification in Breast DCE-MRI. In: Battiato S., Gallo G., Schettini R., Stanco F. (eds) Image Analysis and Processing - ICIAP 2017. ICIAP 2017. Lecture Notes in Computer Science, vol 10485. Springer, Cham.
16. Amato, F., Marrone, S., Moscato, V., Piantadosi, G., Picariello, A., & Sansone, C. (2017) Chatbots meet eHealth: automatizing healthcare. In: D. Impedovo, G. Pirlo, Proceedings of the Workshop on Artificial Intelligence with Application in Health co-located with the 16th International Conference of the Italian Association for Artificial Intelligence (AI*IA 2017), Vol-1982.
17. Marrone S., Sansone C., (2019) Adversarial Presentation Attack Against Fingerprint Based Authentication Systems. In: 12th IAPR IEEE International Conference On Biometrics (ICB)
18. Marrone S., Sansone C., (2019) An Adversarial Perturbation Approach Against CNN-based Soft Biometrics Detection. In: the 2019 International Joint Conference on Neural Networks (IJCNN)
19. Marrone S., Sansone C., (2020) A Transferable Adversarial Perturbation Attack Against Fingerprint Based Authentication Systems. In: Computer Vision and Image Understanding (under review)

Questions?

stefano.marrone@unina.it