Role of Adversaries in Deep Learning

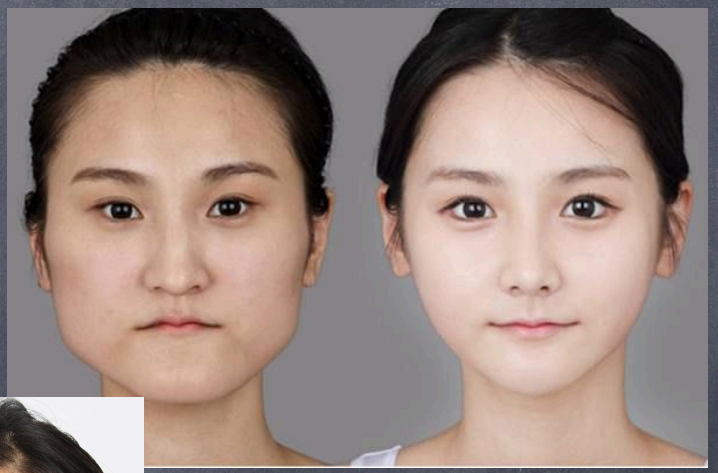
Mayank Vatsa, Richa Singh, and Nalini Ratha

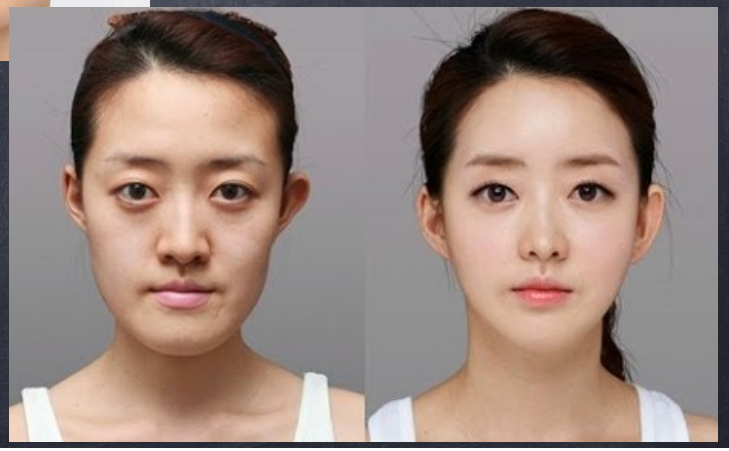
IIIT-Delhi and IBM TJ Watson

Let us start with some quick tests

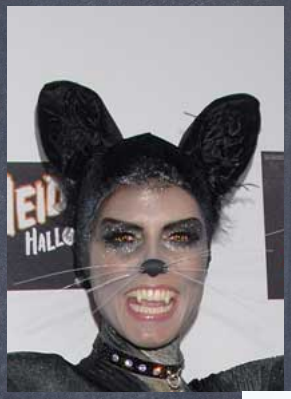
Find Genuine Image Pairs

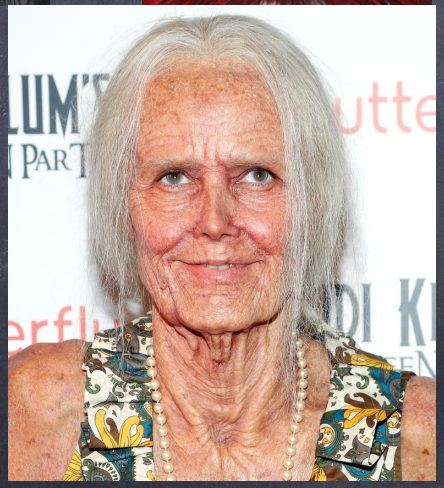
ALL are Genuine





Which of these belong to Heidi Klum?

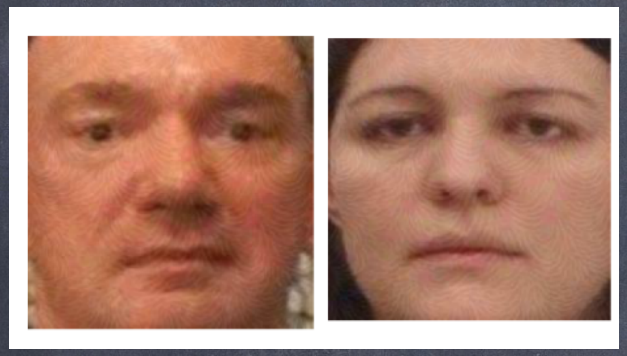


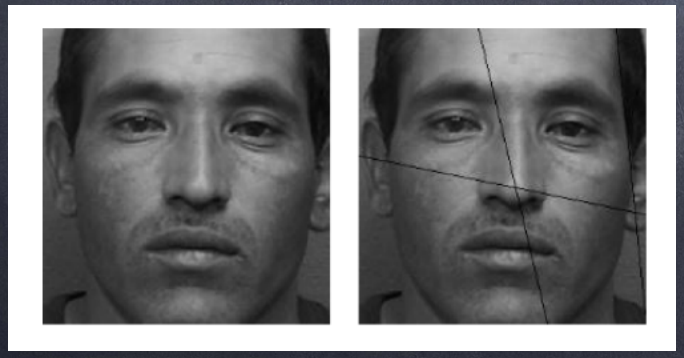


Which of these belong to Heidi Klum?

All of them

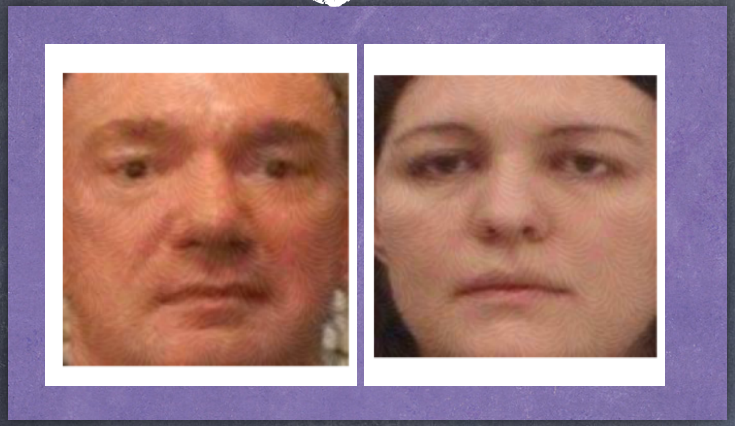
Find Genuine Image Pairs





Find Genuine Image Pairs

For Algorithms



For Human Eyes

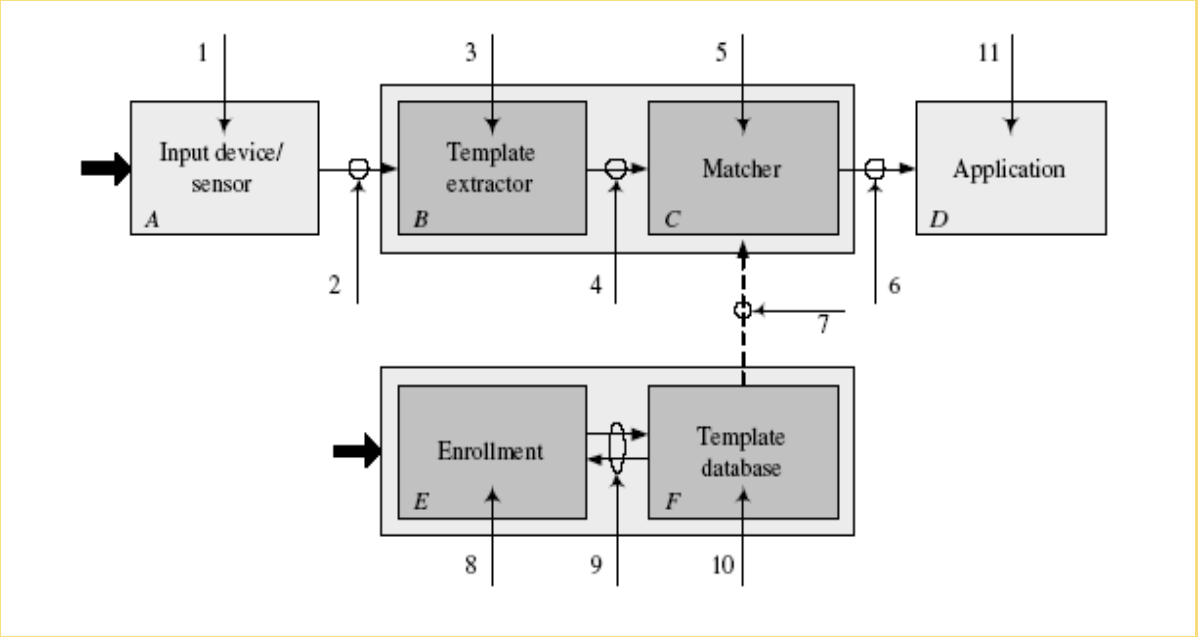
For Human Eyes

For Algorithms

Structure of the Tutorial

- Motivation and classification of attacks
- How to attack a system/algorithm using adversarial perturbation?
- How to detect these adversarial perturbations (attacks)?
- How to mitigate the effect of adversarial perturbation?
- Is adversarial perturbation always bad?

Shallow Learning Attack Model (Pre-DL Era)

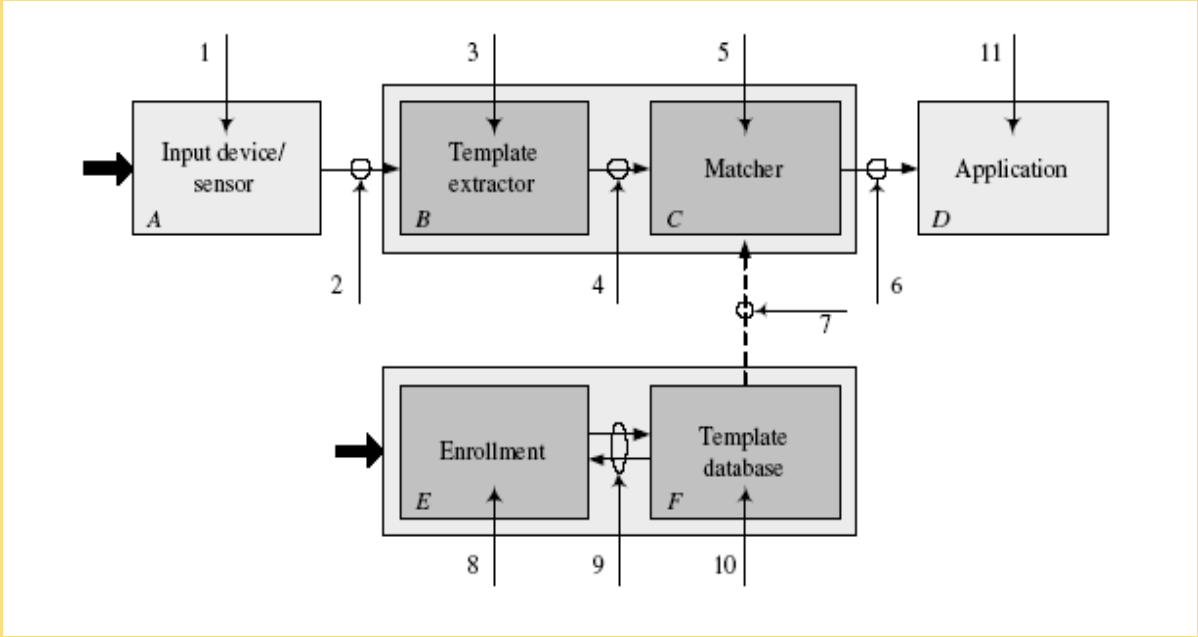


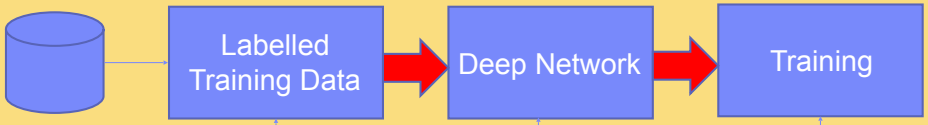
Formidable adversaries:

- Thieves
- Hackers
- Users
- Customers
- Employees
- Merchants
- Competitors
- Competitors' governments

Ratha et al. 2003

Deep Learning Attack Models (DL Era)





Corrupting training data

Corrupting the network

Corrupting training process

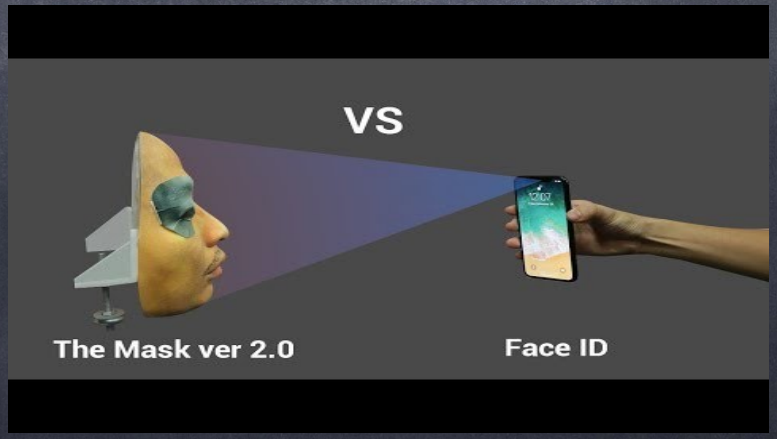
Formidable adversaries:

- Thieves
- Hackers
- Users
- Customers
- Employees
- Merchants
- Competitors
- Competitors' governments

Classification of Attacks

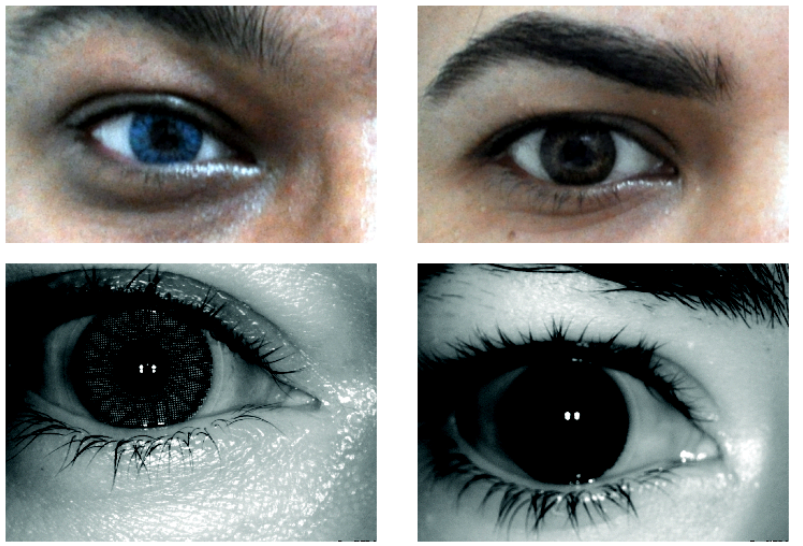
- Physical attacks
- Digital attacks

Physical Adversarial Attacks

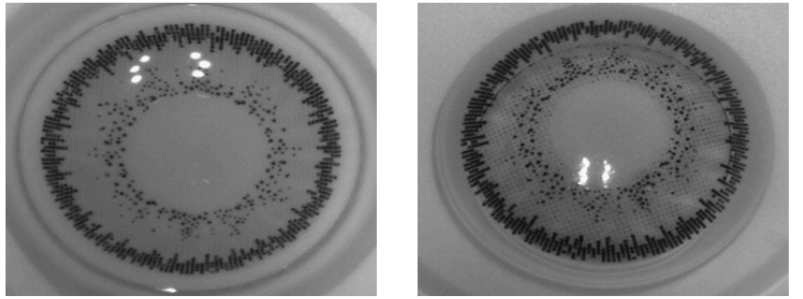


Physical Adversarial Attacks

(a) Live Finger (b) Gummy Finger



With Textured Lens



Blue Hazel

Textured Lens

Physical Adversarial Attacks

Physical Adversarial Attacks

Black robbers used $2,000 white masks to fool victims in $200,000 'Town'-style stickup, prosecutors say

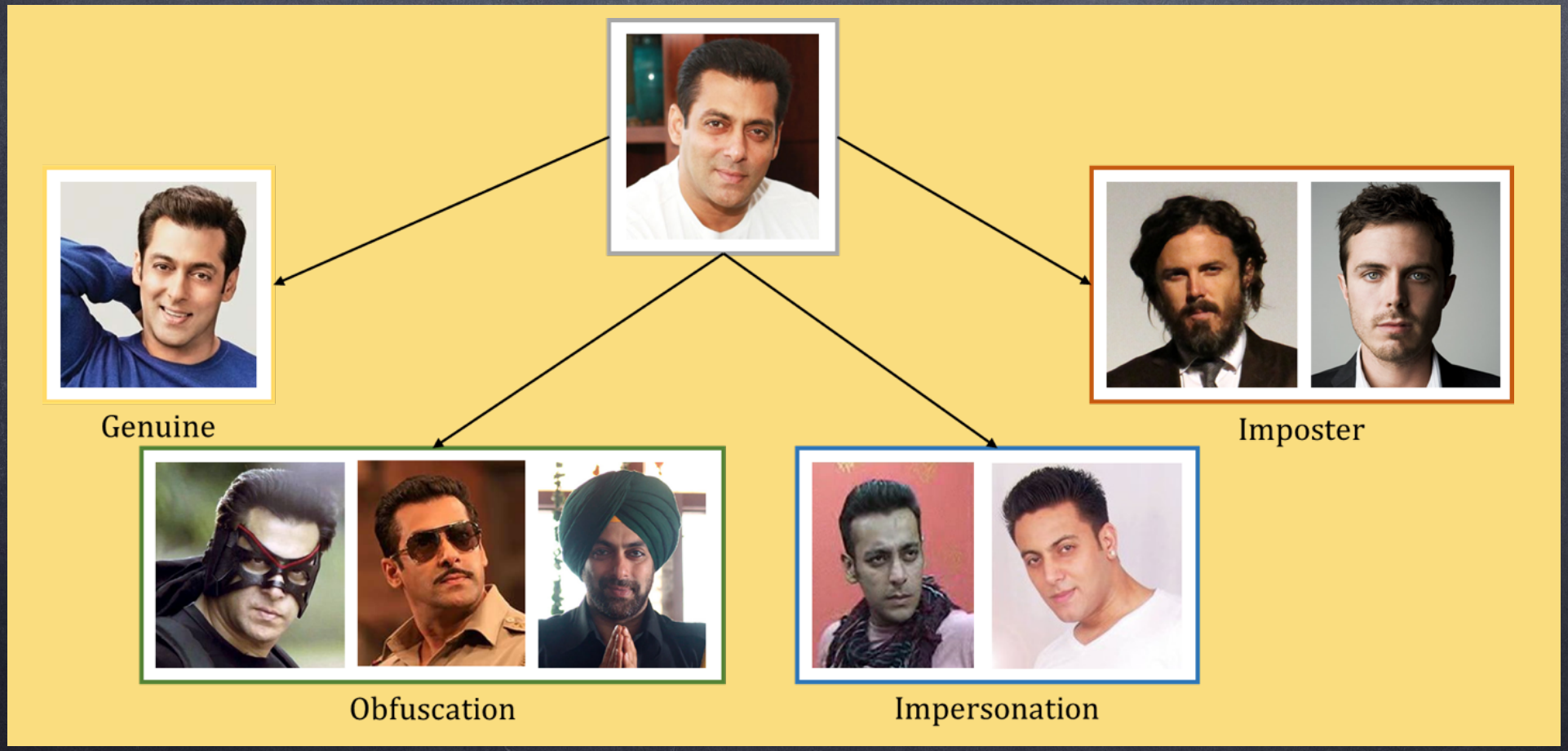
By Selim Algar July 31, 2013 | 4:00am

The white robber who carried out six raids disguised as a black man (and very nearly got away with it)

By DAILY MAIL REPORTER
UPDATED: 16:11 GMT, 1 December 2010

Manjani et al., Detecting Silicone Mask based Presentation Attack via Deep Dictionary Learning, IEEE T-IFS 2017

Physical Adversarial Attacks

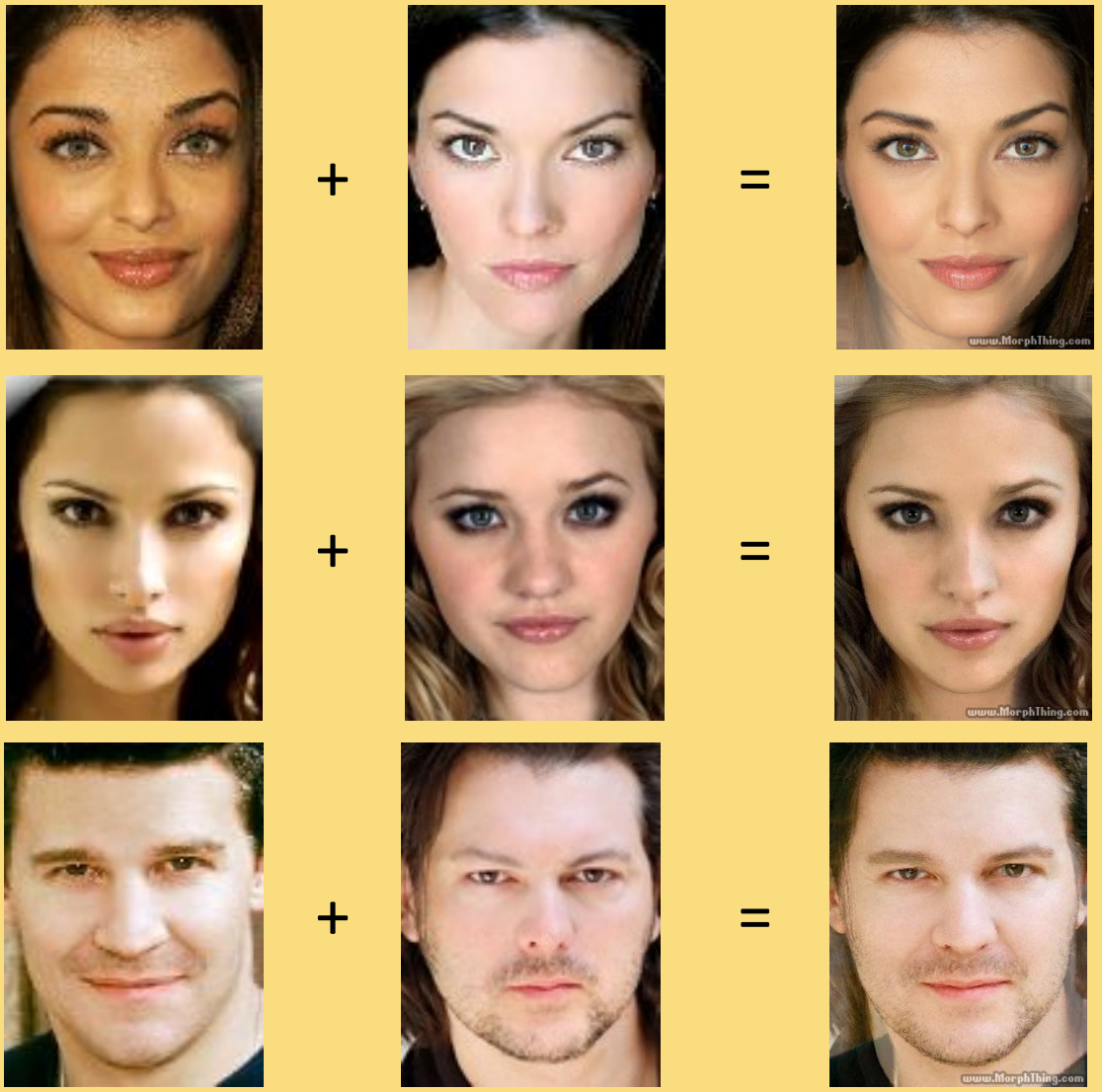


Kushwaha et al. CVPRW - DFW2018

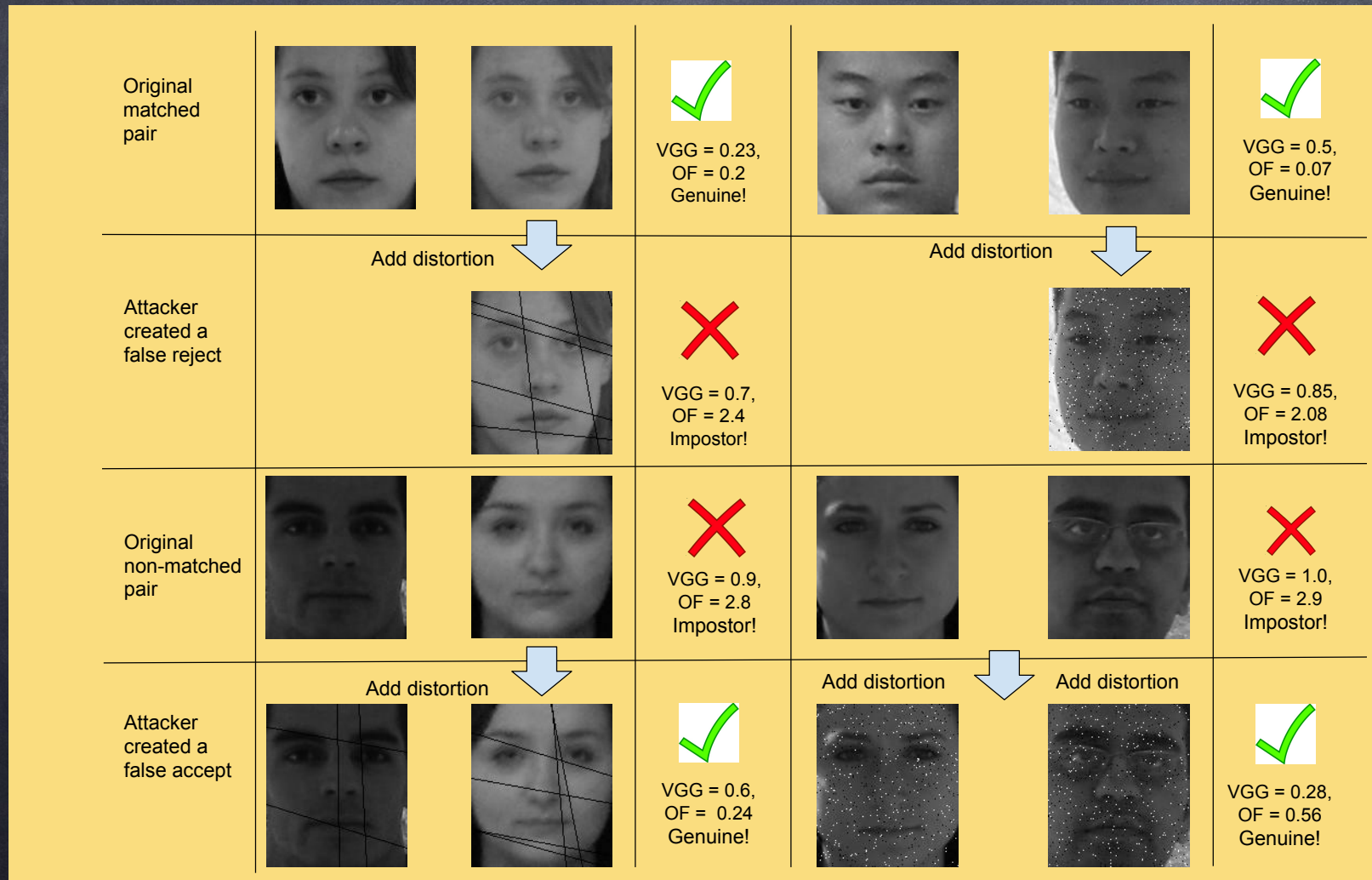
Digital Adversarial Attacks

- Digital retouching
- Photoshop effects
- Morphing

Bharati et al. IEEE T-IFS 2016, IJCB2017

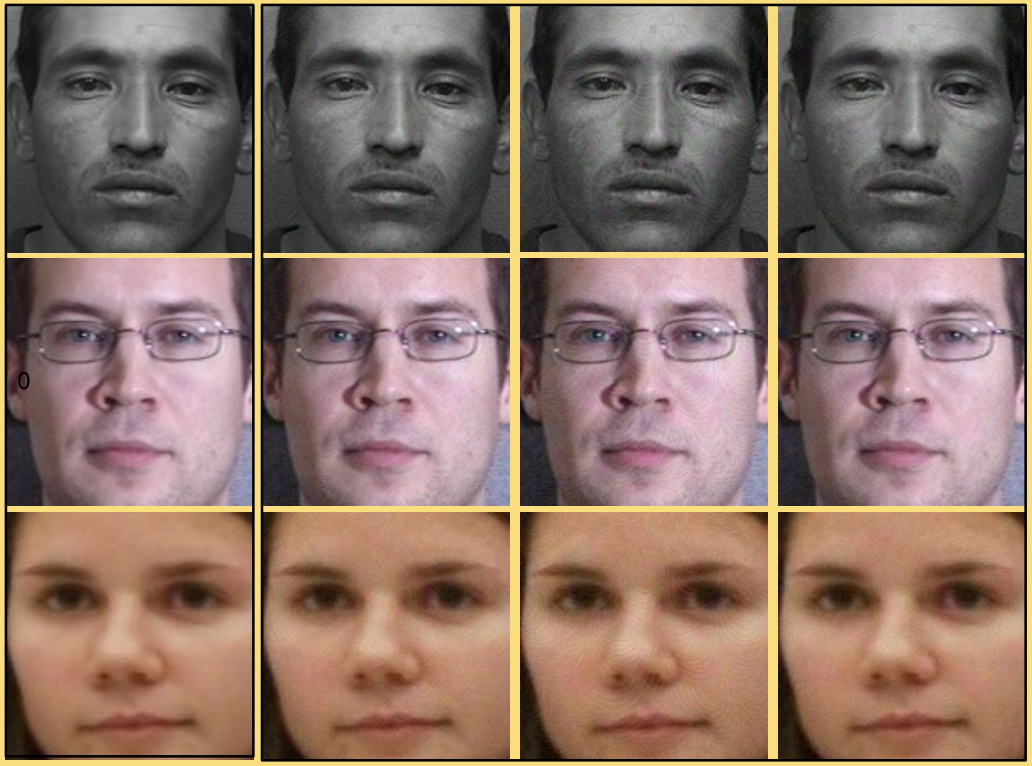

+
=
+
=
+
=
Subject 1
Subject 2
Morphed

Digital Adversarial Attacks



Goswami et al. AAAI2018

Digital Adversarial Attacks

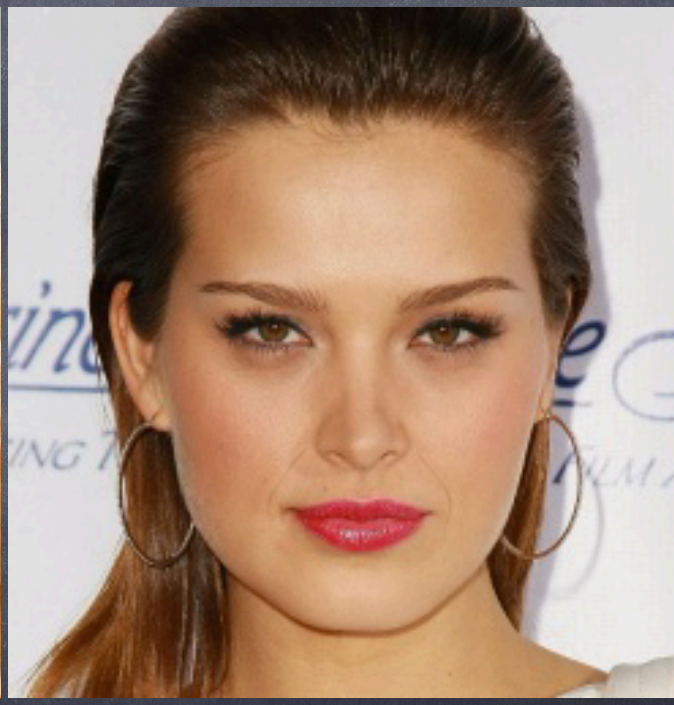


CCS, 2016

Universal Attack, CVPR 2017

Who are these celebrities?

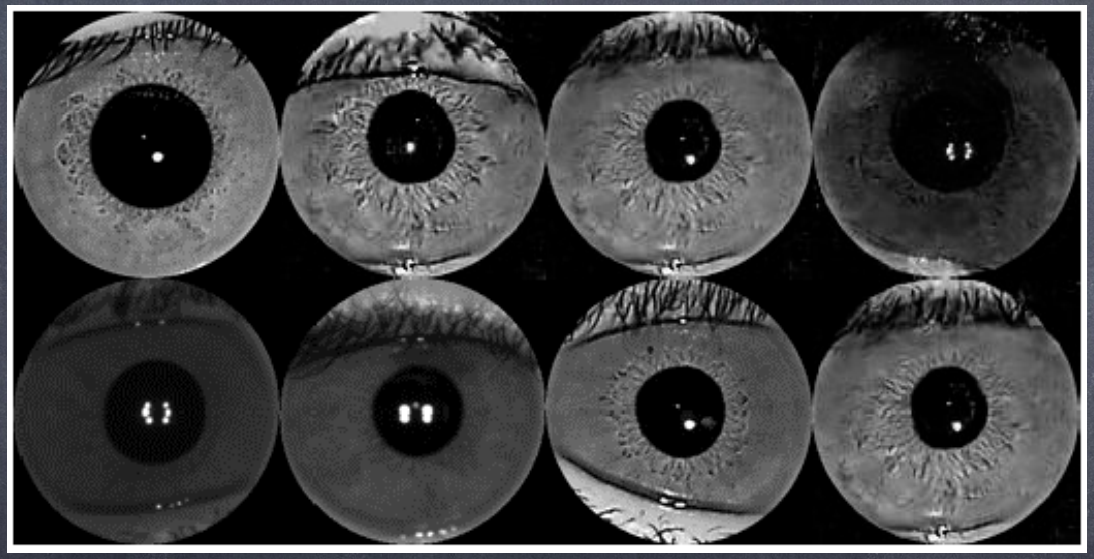




Non-existing identities

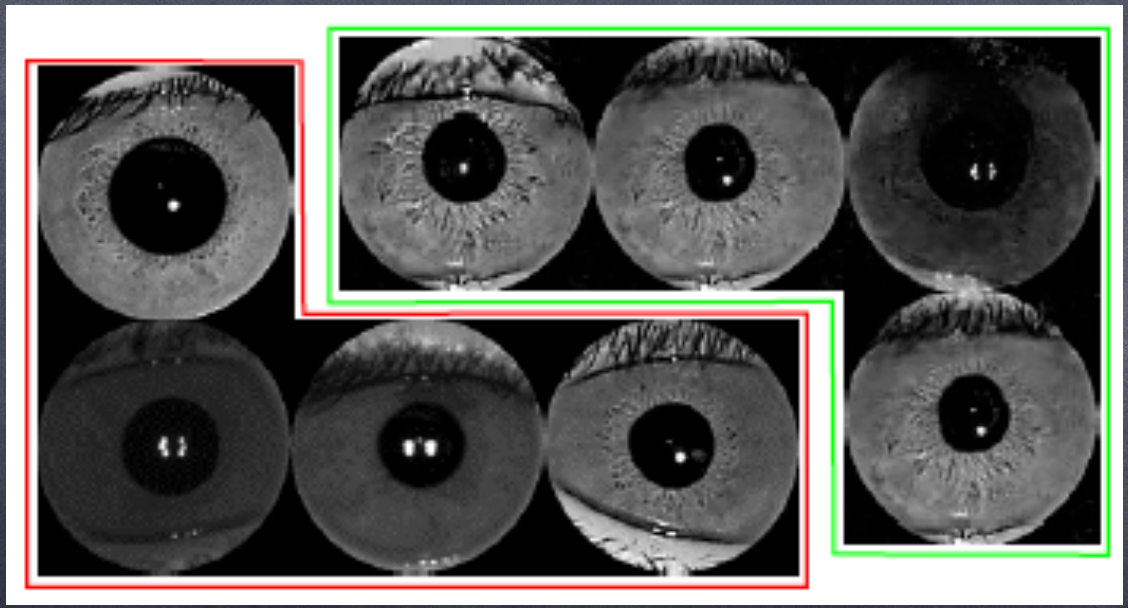
PROGRESSIVE GROWING OF GANS FOR IMPROVED QUALITY, STABILITY, AND VARIATION, ICLR2018

Which one of the iris images are real?



Which one of the iris images are real?

Real



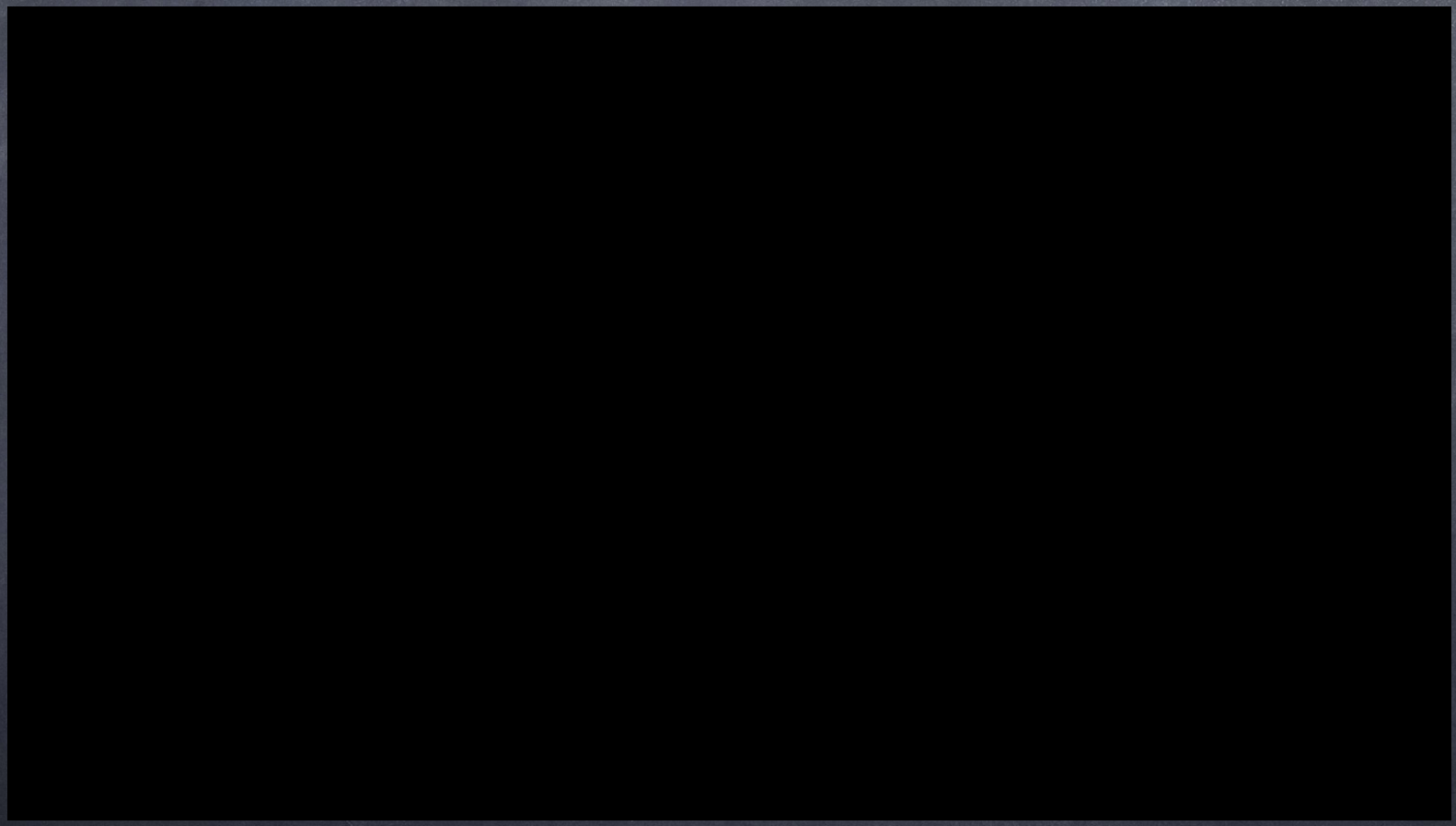
Synthetic using GANs

Kohli et al. IJCB, 2017

What is this?

- For humans: stop sign
- For deep learning based algorithm: speed limit sign

Adversarial Attacks in Videos



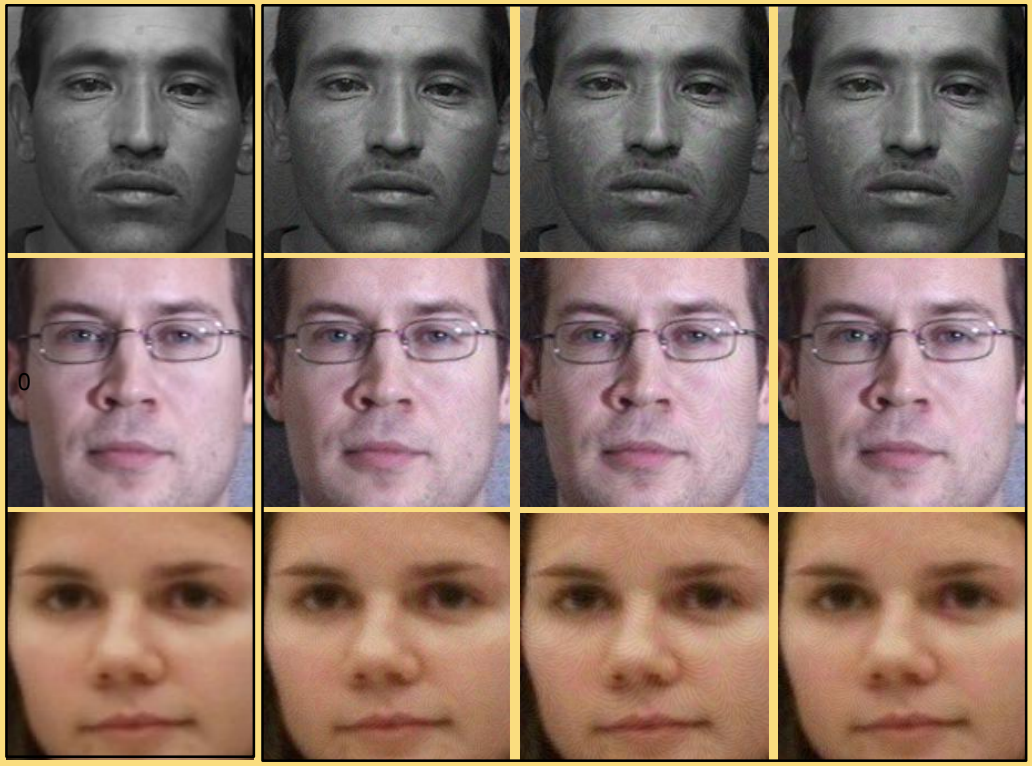
https://www.engadget.com/2017/11/10/counterfeit-ai-machine-learning-forgery/

Facial Reenactment

Imperceptible Noise

Original GoogleNet VGG16 ResNet-152

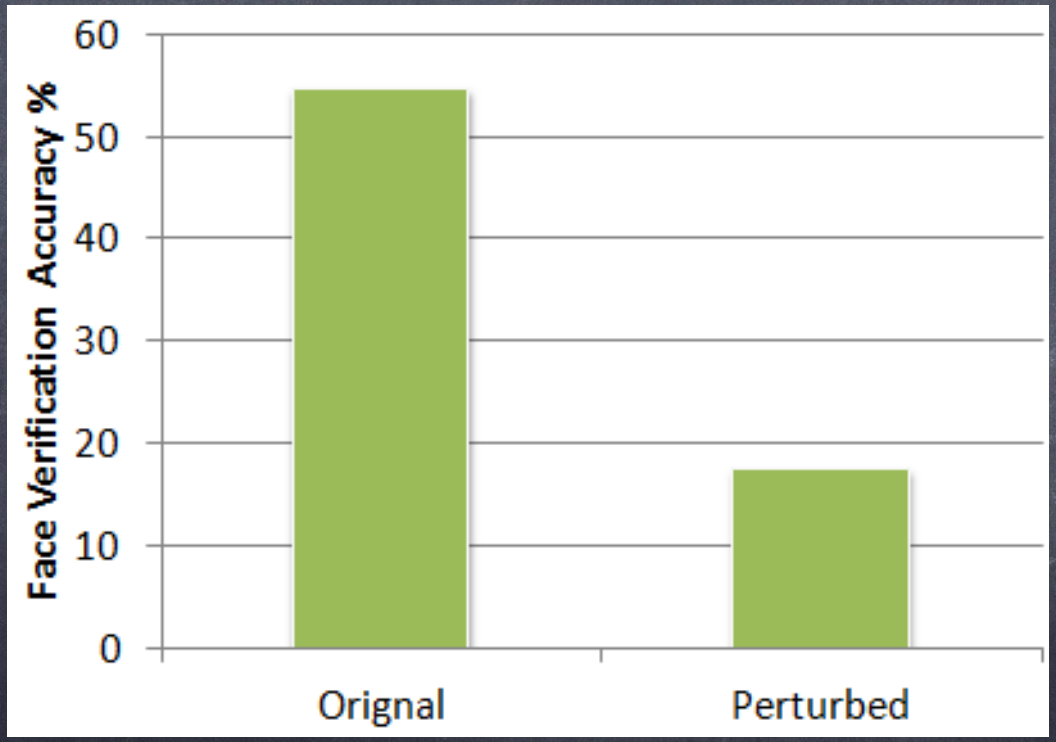
ImageNet Examples



Original GoogleNet VGG16 ResNet-152

Face Examples

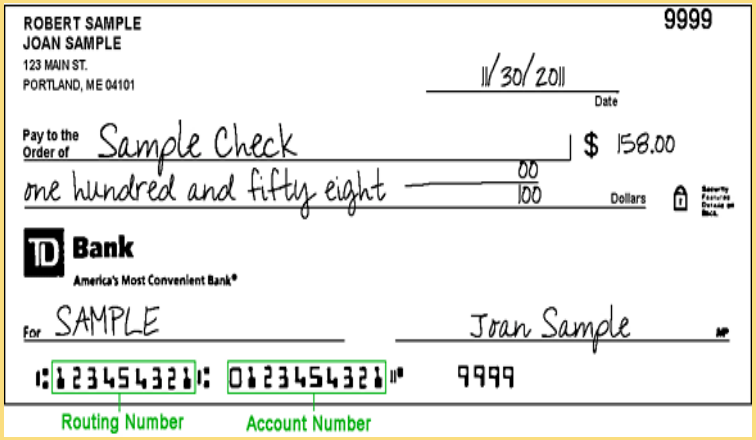
Impact on Face Recognition



VGG-Face model

G. Goswami, N. Ratha, A. Agarwal, R. Singh, and M. Vatsa. Unravelling robustness of deep learning based face recognition against adversarial attacks. AAAI, 2018

A Real World Implication



158.00 → 10,000.00

Key Takeout (so far)

- So, now we are convinced that deep learning based systems can be attacked
- Keyword is "adversarial perturbation"

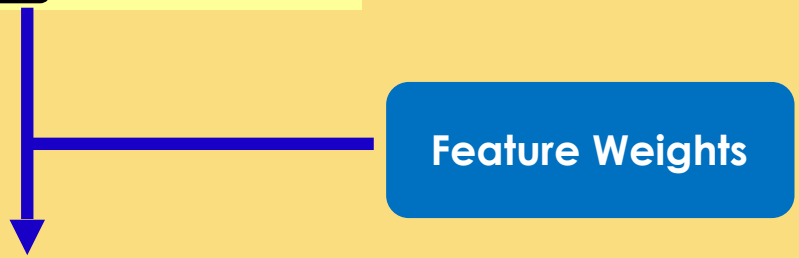
How Adversarial Perturbation Works?

Adversarial Attacks - Since When?

- In the context of DL, adversarial examples were discovered by
  - C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199, 2013.
- In PR, False Accepts and False Rejects have been studied at length with respect to perturbations
- Biometrics systems have studied the biometrics zoo
- Biometrics systems have studied presentation attacks
- Adversarial Machine Learning has been known for a long time (since 2004)

Numerical Example

1. From: spammer@example.com
   Cheap mortgage now!!!



2. cheap = 1.0
   mortgage = 1.5

3. Total score = 2.5> 1.0 (threshold)

Spam

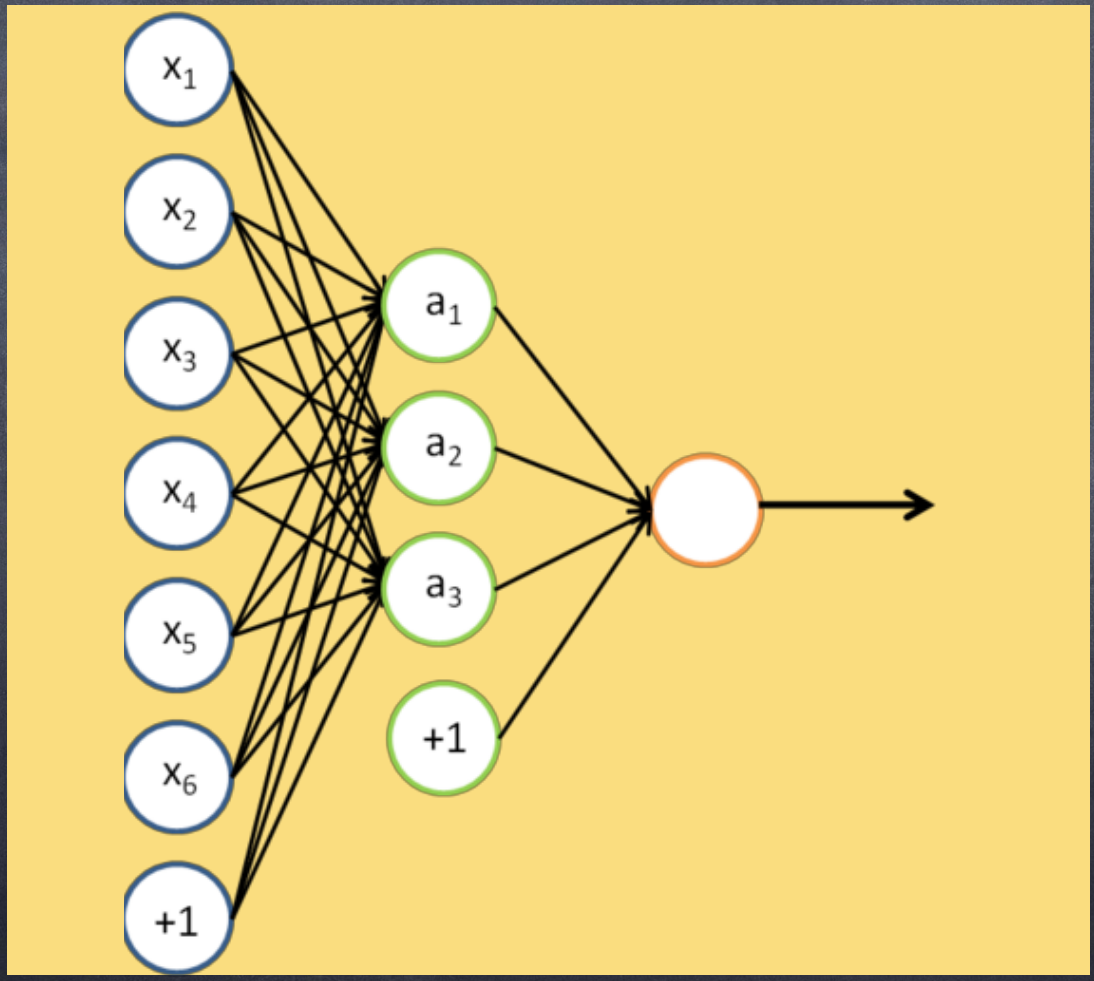
Numerical Example

1. From: spammer@example.com
   Cheap mortgage now!!!
   Joy Oregon

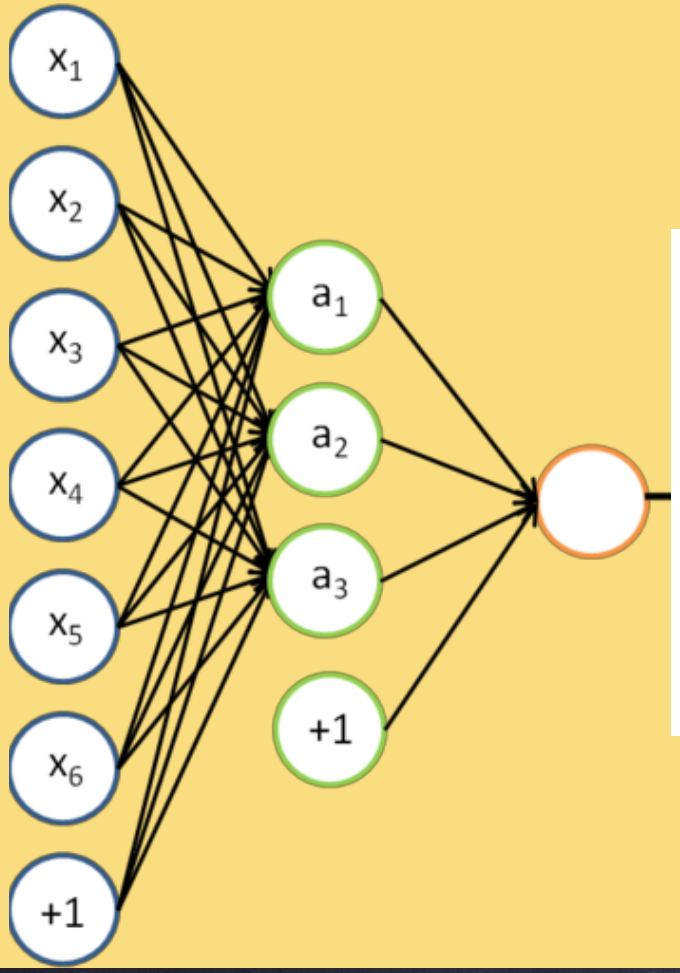
2. cheap = 1.0
   mortgage = 1.5
   Joy= -1.0
   Oregon = -1.0
3. Total score = 0.5 < 1.0 (threshold)

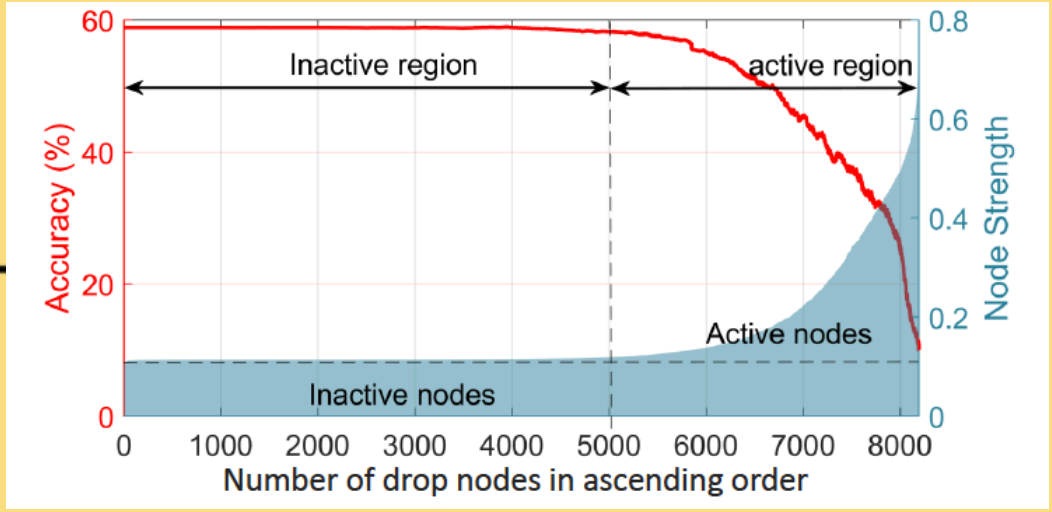
OK

Let us take a simple Neural Net

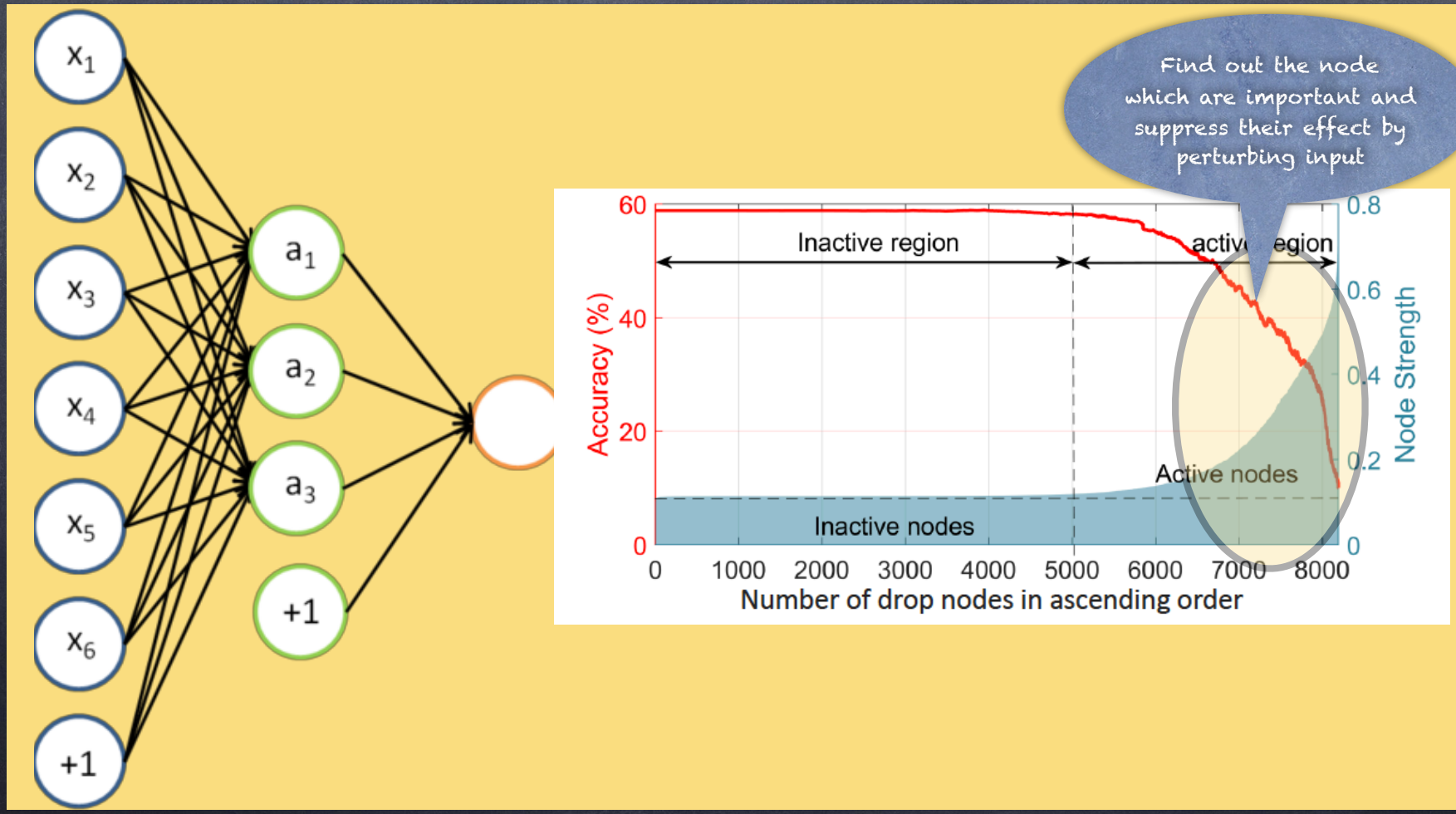


Let us take a simple Neural Net

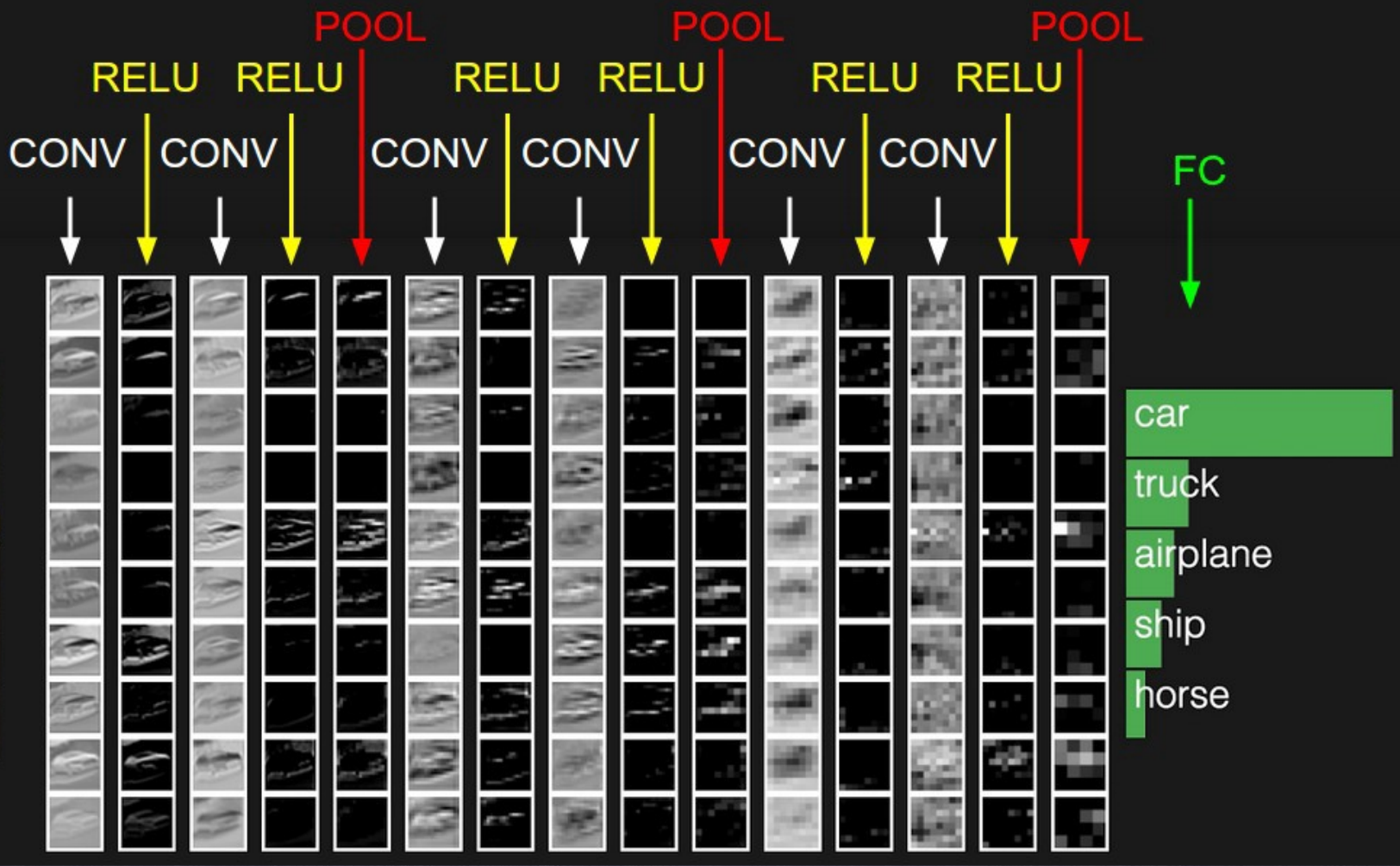




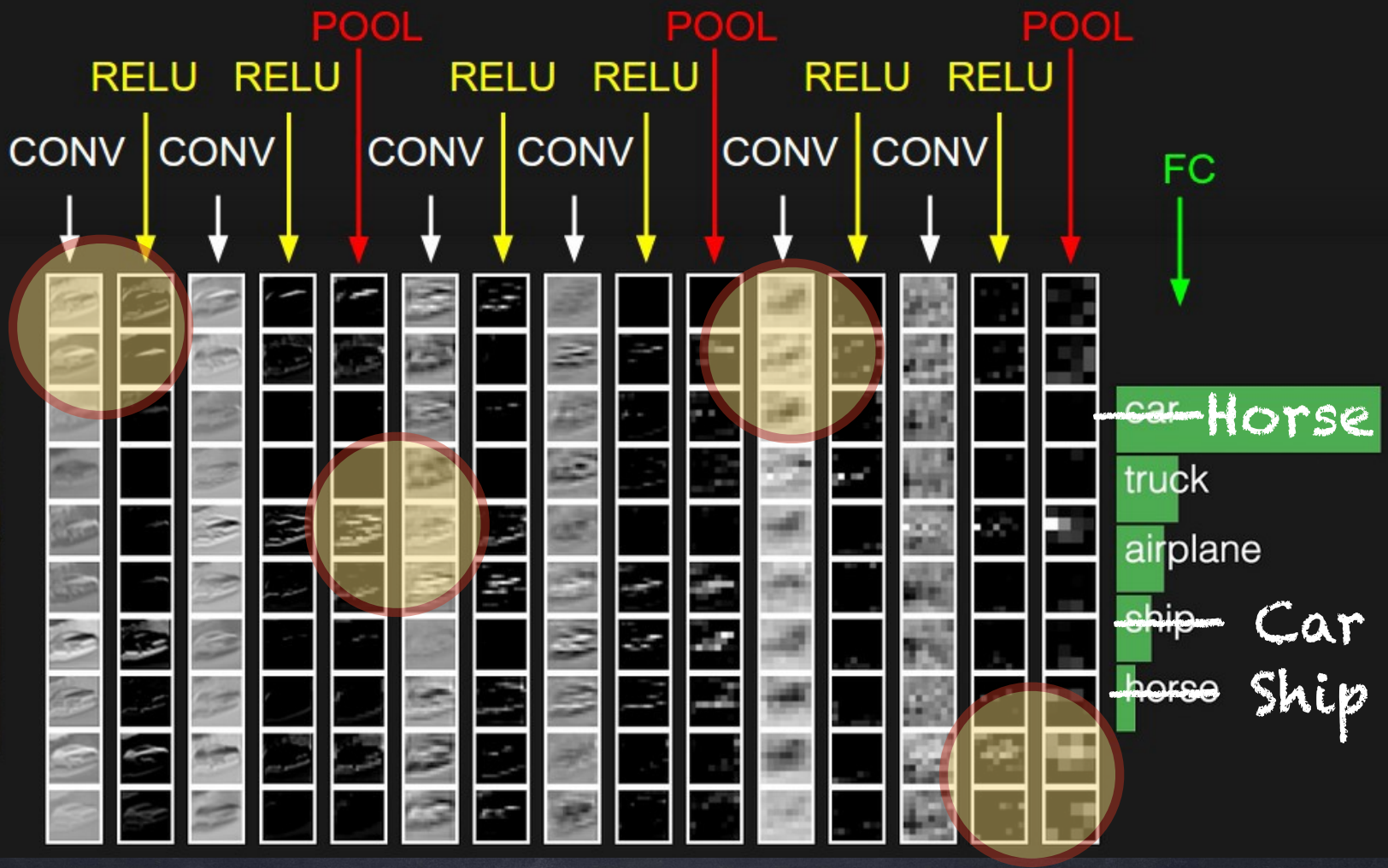
Let us take a simple Neural Net



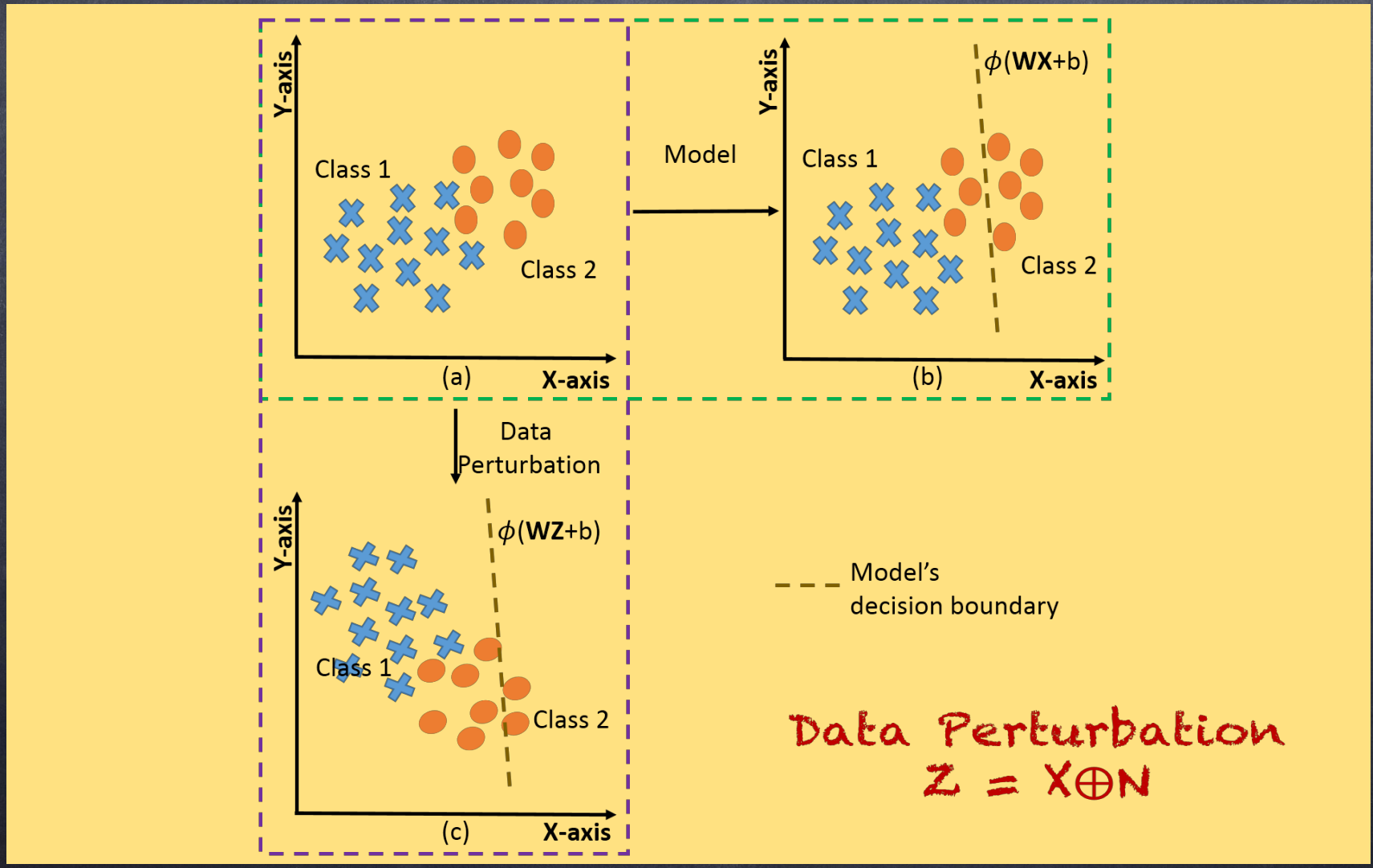
Extending the example to CNN



Extending the example to CNN



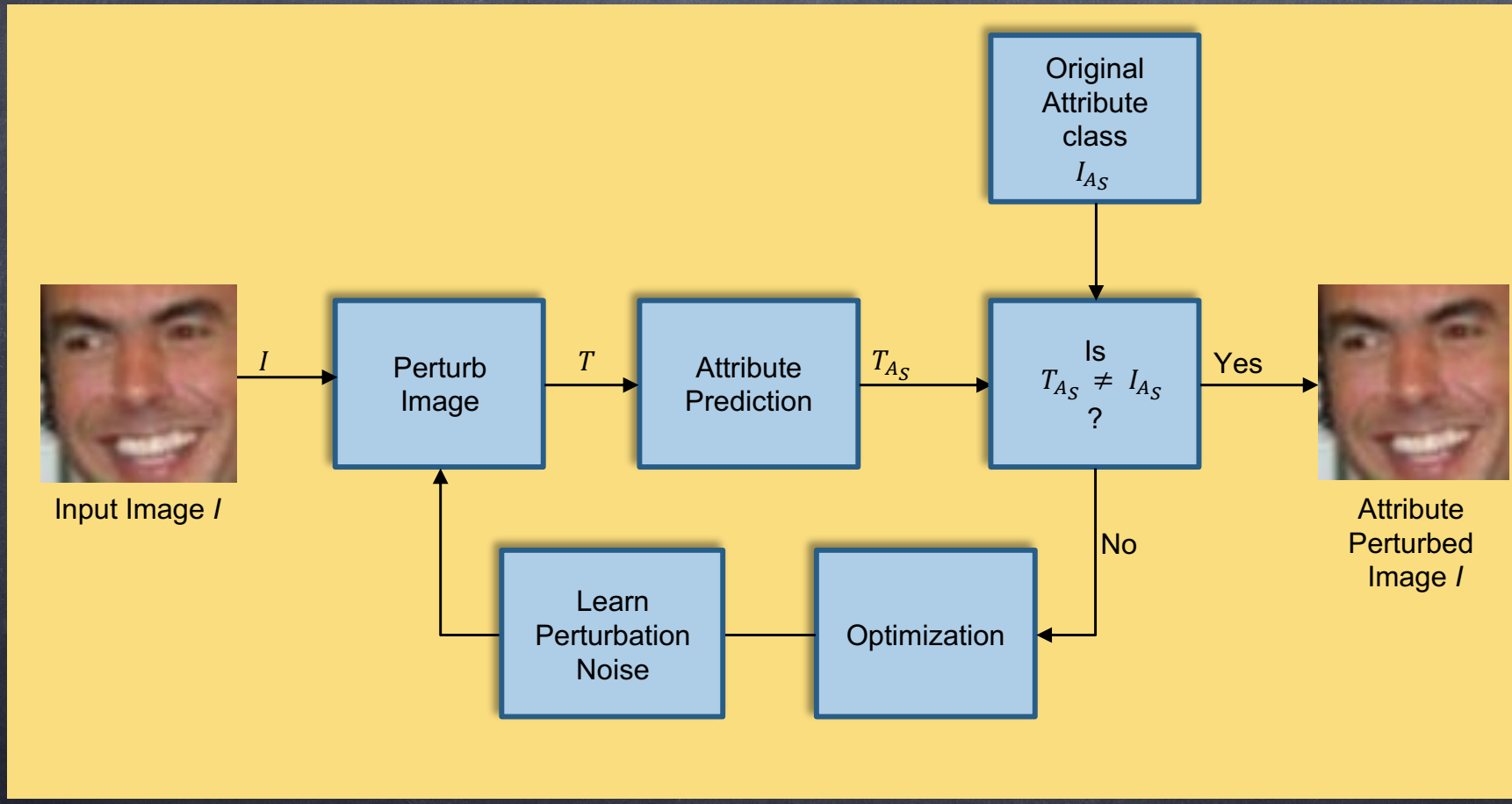
Mathematically Adversarial Perturbations



Mathematically

- This can be viewed as an optimization problem, i.e.
- $\min[D(I_o) - D(I_p)] + \min(\|I_o - I_p\|)$
  - such that $Class(I_o) \neq Class(I_p)$
- First term minimizes the feature distance between original and perturbed information/features
- Second term minimizes the visual difference between original and perturbed images

Example - Attribute Perturbation



Example - Adversarial Noise of Universal Perturbation

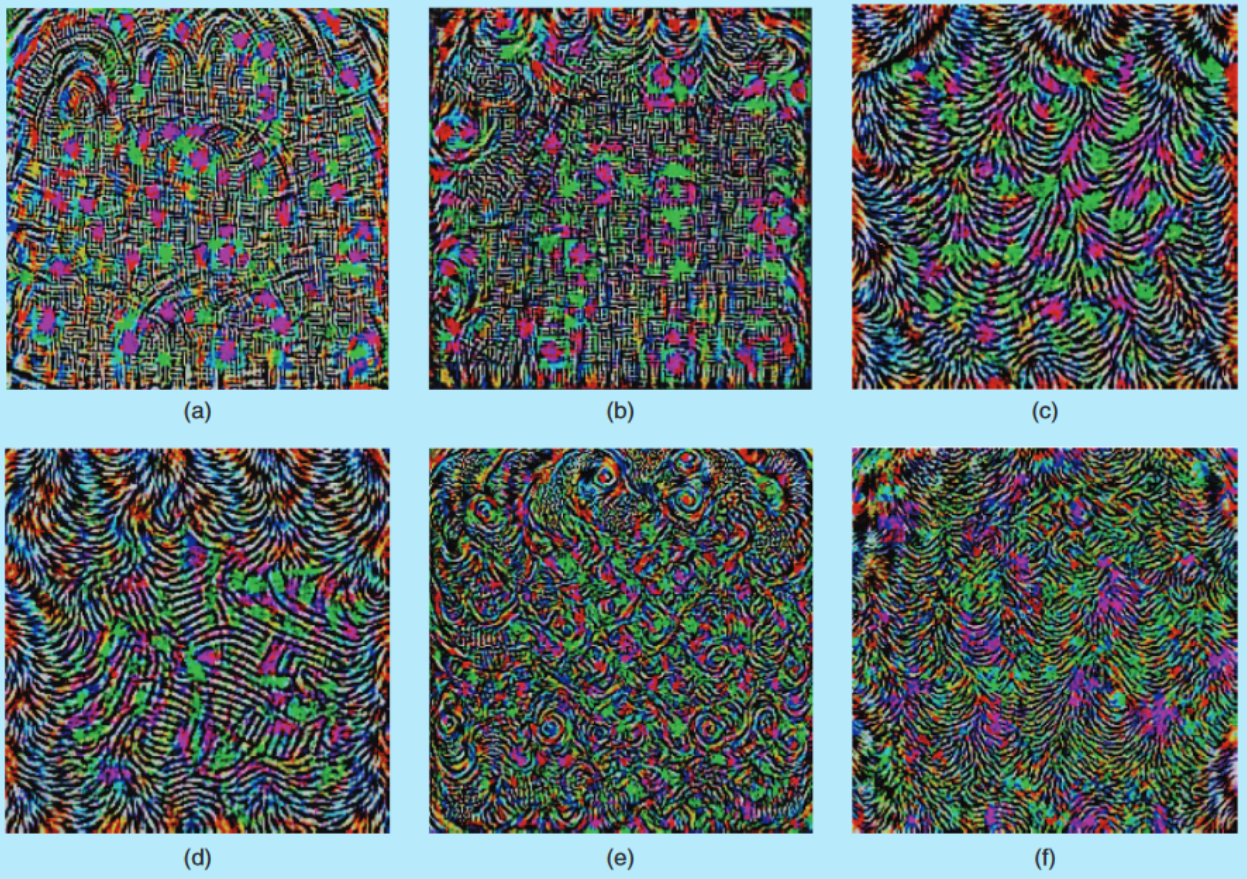
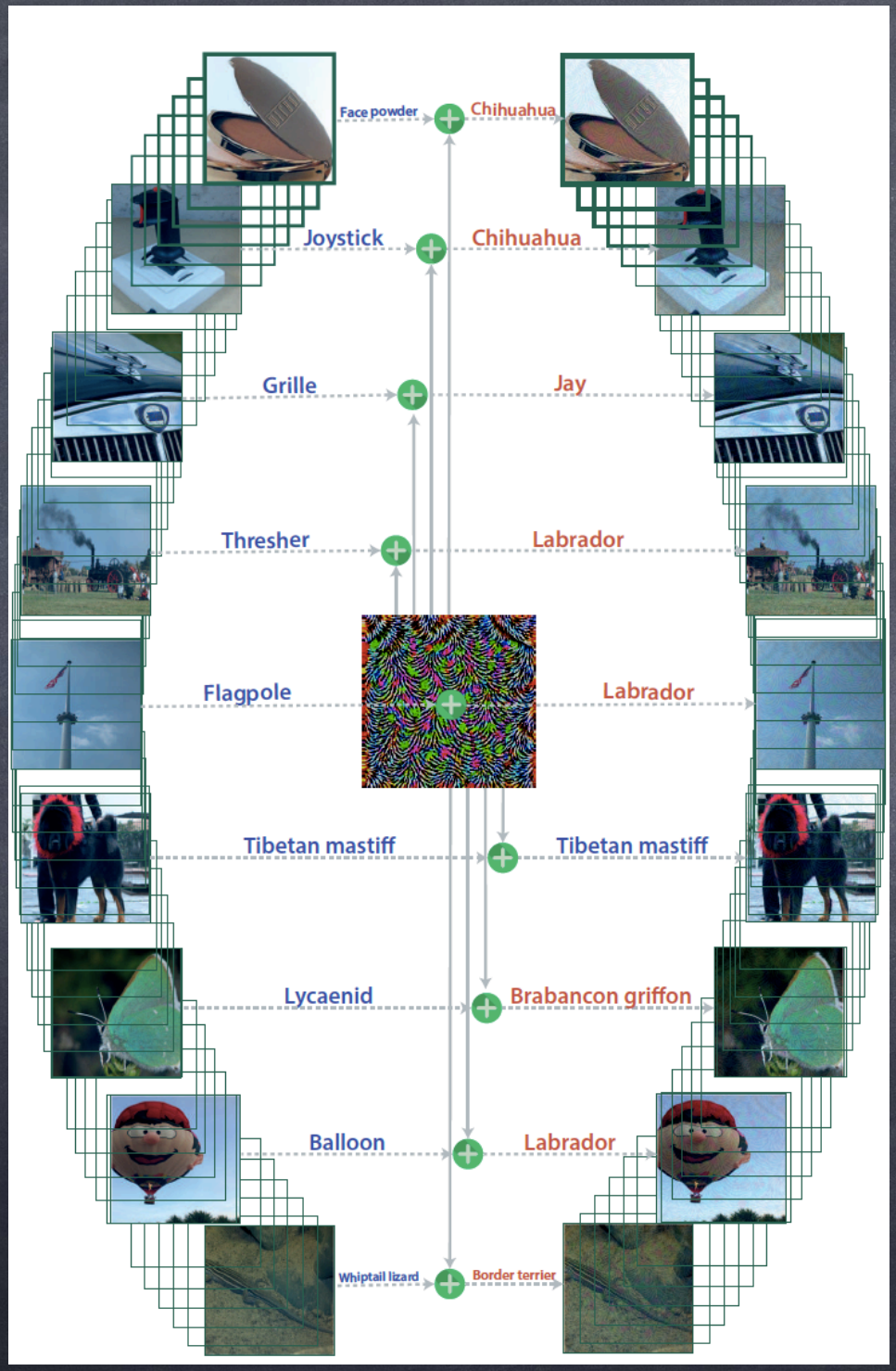


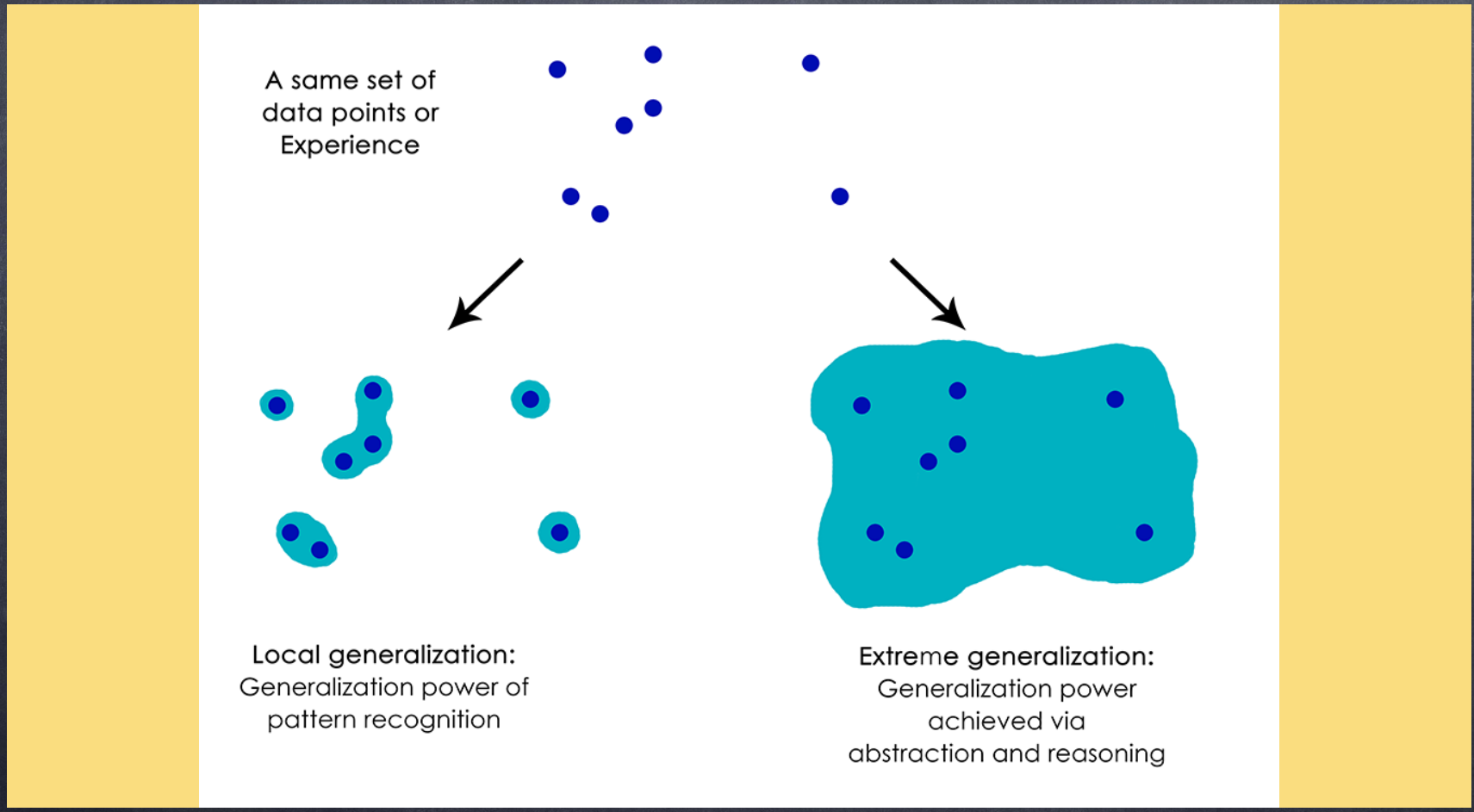
FIGURE 3. Universal perturbations computed for different deep neural network architectures. The pixel values are scaled for visibility. (a) CaffeNet, (b) VGG-F, (c) VGG-16, (d) VGG-19, (e) GoogLeNet, and (f) ResNet-152.

Moosavi-Dezfooli et al.
CVPR2017



Moosavi-Dezfooli et al.
CVPR2017

Why Adversarial Perturbation Works?



Adversarial	Authors	Descriptions
Generation	Szegedy et al., 2013	L-BFGS: $L(x + \rho, l) + \|\rho\|^2$ s.t., $x_i + \rho_i \in [b_{min}, b_{max}]$
	Goodfellow, Shlens, and Szegedy, 2015	FGSM: $x_0 + \varepsilon(\nabla_x L(x_0, l_0))$
	Papernot et al., 2016	Saliency Map: l_0 distance optimization
	Moosavi-Dezfooli, Fawzi, and Frossard, 2016	DeepFool: for each class; $l \neq l_0$; minimize d(l,l0)
	Carlini and Wagner, 2017	C & W: lp distance metric optimization
	Moosavi-Dezfooli et al., 2017	Universal (Image-Agnostic): Distribution based perturbation
	Rauber, Brendel, and Bethge, 2017	Blackbox: Uniform, Gaussian, Salt and Pepper, Gaussian Blur, Contrast

Attacks on Faces

- Grid based occlusion (Grid)
- Most significant bit based noise (xMSB)
- Eye region occlusion (ERO)
- Forehead and brow occlusion (FHBO)
- Beard-like occlusion (Beard)
- Universal Perturbation

Evaluating Robustness

MEDS

System	Original	Grids	xMSB	FHBO	ERO	Beard
COTS	24.1	20.9	14.5	19.0	0.0	24.8
OpenFace	66.7	49.5	43.8	47.9	16.4	48.2
VGG-Face	78.4	50.3	45.0	25.7	10.9	47.7
LightCNN	89.3	80.1	71.5	62.8	26.7	70.7
L-CSSE	89.1	81.9	83.4	55.8	27.3	70.5

PaSC

System	Original	Grids	xMSB	FHBO	ERO	Beard
COTS	40.3	24.3	19.1	13.0	0.0	6.2
OpenFace	39.4	10.1	10.1	14.9	6.5	22.6
VGG-Face	54.3	3.2	1.3	15.2	8.8	24.0
LightCNN	60.1	24.6	29.5	31.9	24.4	38.1
L-CSSE	61.2	43.1	36.9	29.4	39.1	39.8

All values indicate genuine accept rate (%) at 1% false accept rate

What an Attacker can Cause?

- Confidence reduction - the output confidence score is reduced, thus introducing class ambiguity
- Random mis-classification - an input is modified in order to output any class different than the correct one
- Targeted mis-classification - an input is modified in order to output a specific target class

Types of Attacks

- White-box
- Grey-box
- Black-box

White-box Attack

- The attacker has perfect knowledge of the DNN used (architecture, hyper-parameters, weights, etc.), has access to the training data and knowledge about any defense mechanisms employed (e.g. adversarial detection systems).
- Therefore, an attacker has the ability to completely replicate the model under attack

Grey-box Attack

- In this case the attacker can collect some information about the network's architecture (e.g. she knows a certain model/uses an open-source architecture), she knows the model under attack was trained using a certain dataset or has information about some defense mechanisms
- In any of these cases, the information is neither complete nor certain and provides the attacker an ability to partially simulate the model under attack

Black-box Attack

- The attacker has no knowledge about the model under attack, however, she has the ability to use the model (or a proxy of it) as an oracle.
- The attacker can supply limited inputs and collect output information to build attack model

Some other classification terms

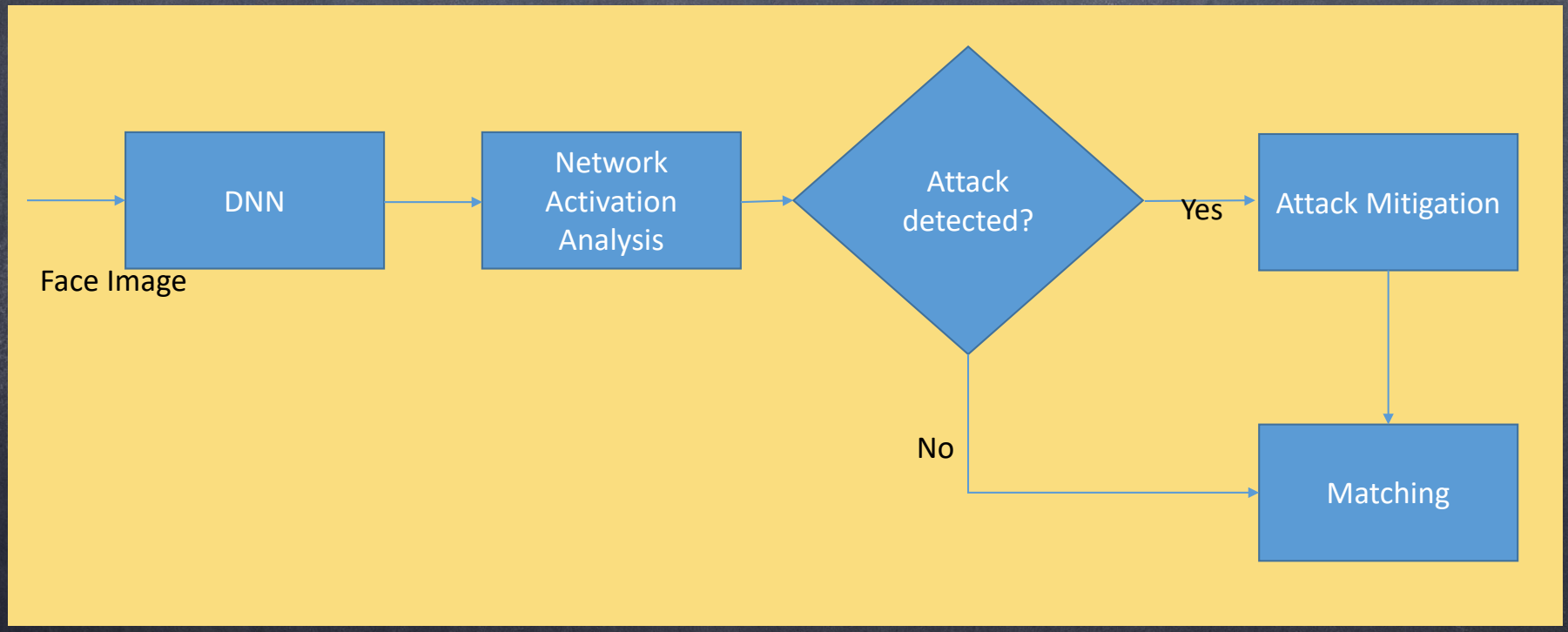
- Modify vs Generate
- Optimization vs Sensitivity vs Geometric Transformation vs Generative Models
- Single Shot vs Iterative
- Specific vs Universal

Catalog of Adversarial Attacks

Attack	Modify (M) or Generate (G) Input	Optimisation (OP), Sensitivity (SA), Geometric Transformations (GT) Generative Models (GM)	Targeted (TG), Non-Targeted (NTG)	Single-Shot (SS), Iterative (IT)	White-box (WB), Grey-box (GB), Black-box (BB)	Specific (SP), Universal (UN)
L-BFGS [185]	M	OP	TG	IT	WB	SP
Deep Fool [135]	M	OP	NTG	IT	WB	SP
UAP [132]	M	OP	NTG	IT	WB	UN
Carlini [29]	M	OP	TG / NTG	IT	WB	SP
CFOA (Madry / PG) [128]	M	OP	TG / NTG	IT	WB	SP
STA [90]	M	OP	TG / NTG	IT	WB	SP
ZOO [35]	M	OP	TG / NTG	IT	BB	SP
IS [137]	M	OP	TG / NTG	IT	BB	SP
FGS [70]	M	SA	NTG	SS	WB	SP
JSMA [146]	M	SA	TG	IT	WB	SP
RSSA [188]	M	SA	NTG	SS / IT	WB	SP
BPDA [7]	M	SA	TG	IT	WB	SP
Elastic-Net [34]	M	SA	TG	IT	WB	SP
BI [109]	M	SA	NTG	IT	WB	SP
ILC [109]	M	SA	TG	IT	WB	SP
Momentum [47]	M	SA	NTG	IT	WB	SP
Substitute [145]	M	SA	TG	SS / IT	BB	SP
Rotation Tr. [52]	M	GT	NTG	SS / IT	WB / GB	SP
ManiFool [97]	M	GT	TG / NTG	IT	WB	SP
Spatial Tr. [198]	M	GT	TG	IT	WB	SP
ATN [8]	G	GM	TG / NTG	IT	WB	SP
NAE [211]	G	GM	TG	IT	WB	SP

Akhtar and Mian, 2018

What to do with Adversarial Perturbations?



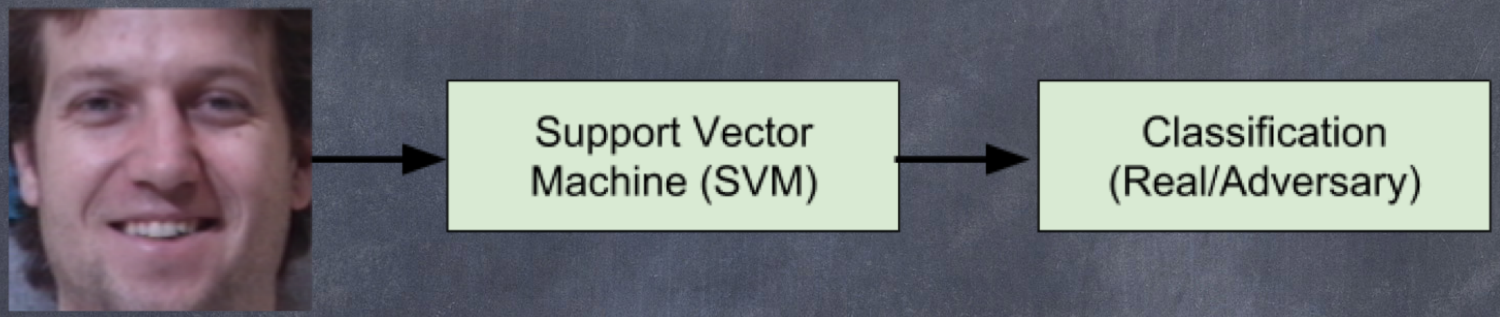
How to detect adversarial perturbation (attack)?

What Could be a simplest approach?

A simple approach

- Treat this problem as 2 class classification problem

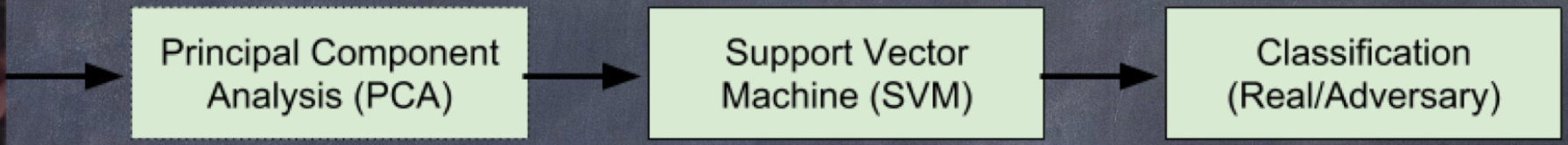
A simple approach



Black-box approach: we do not know about adversary but learn a classifier to identify the difference between real and perturbed samples

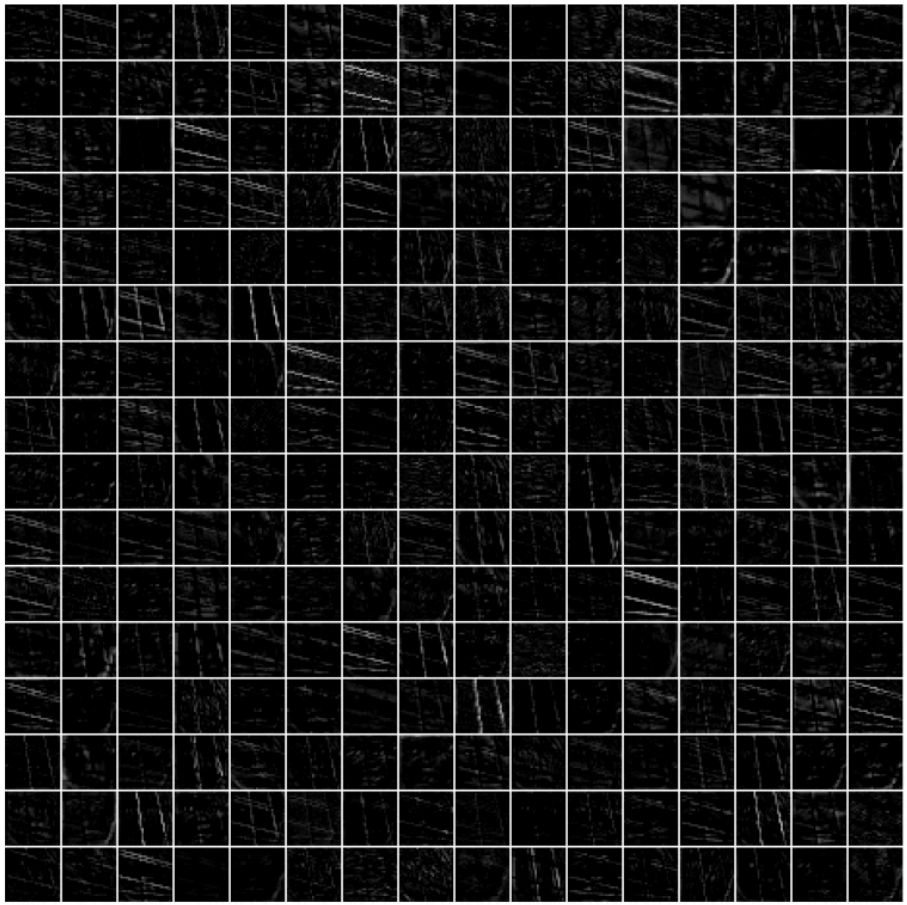
A slightly modified version

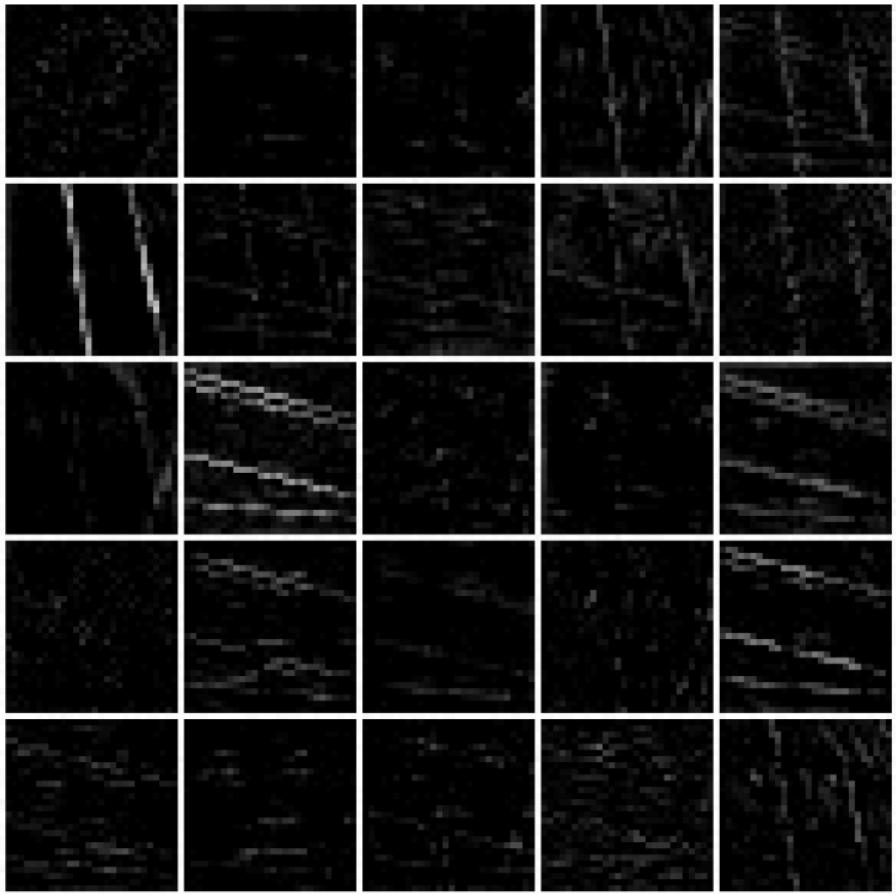
Input Image



Black-box approach: we do not know about adversary but learn a classifier to identify the difference between real and perturbed samples

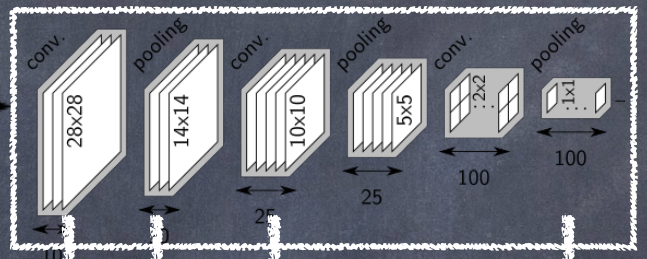
Look at network activation





CNN based White-box Approach

Input Image

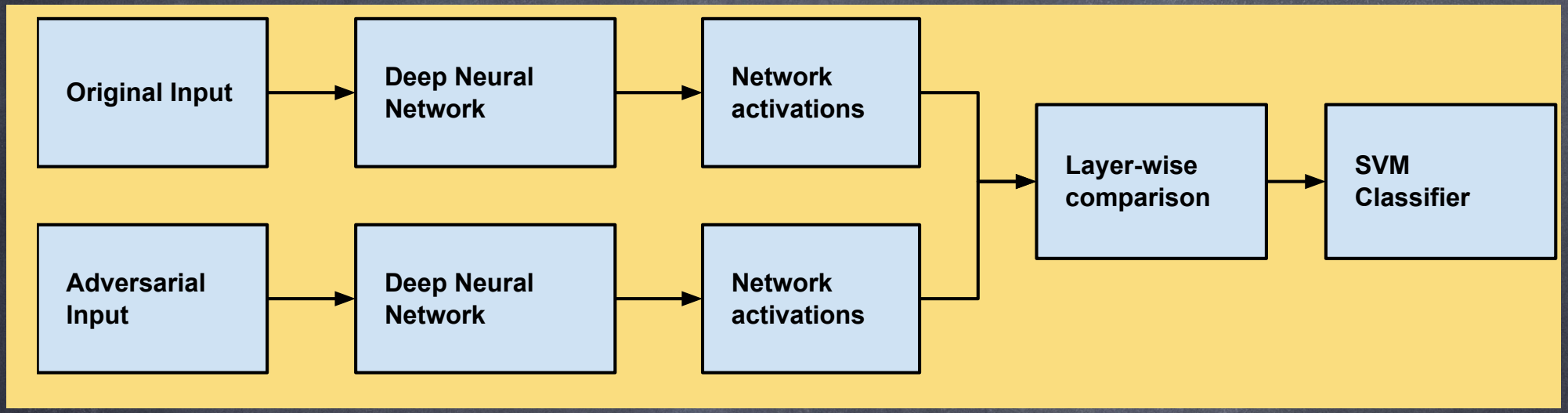


Stat.Feature

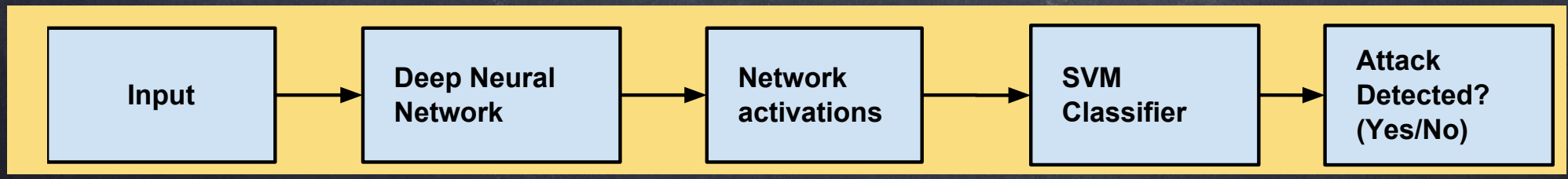
Support Vector Machine (SVM)

Li&Li, ICCV2017

Adversarial Perturbation Detection



White-box Training



Testing

Goswami et al AAAI2018

Adversarial Perturbation Detection ...

- Each layer in a deep neural network essentially learns a function or representation of the input data
- The features obtained for a distorted and undistorted image are measurably different from one another
- Internal representations computed at each layer are different for distorted images as compared to undistorted images
- To detect distortions, the pattern of the intermediate representations for undistorted images are compared with distorted images at each layer

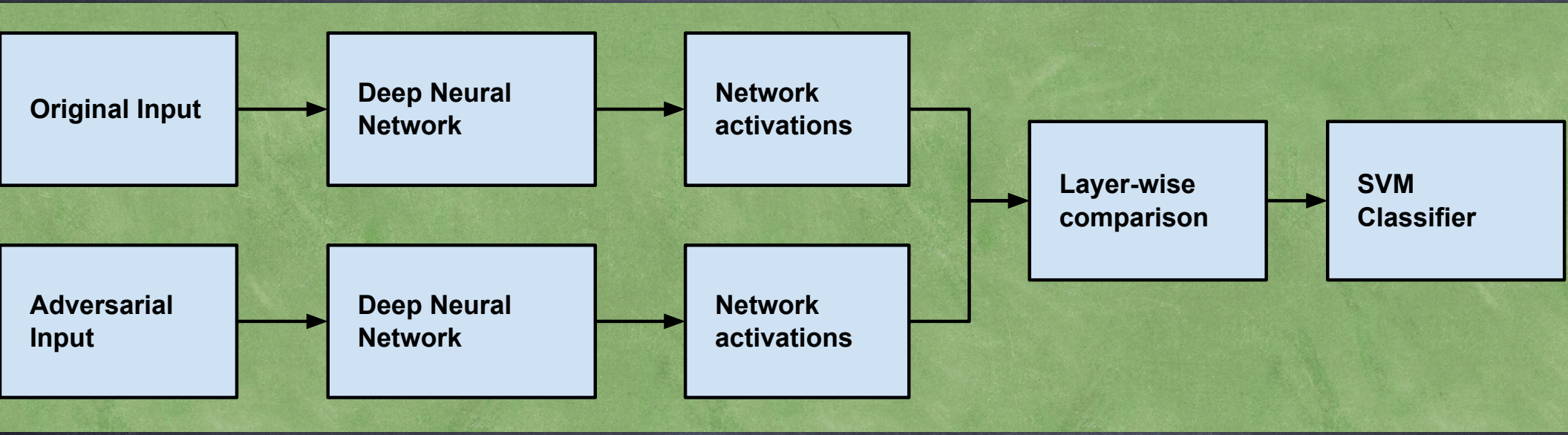
Adversarial Perturbation Detection ...

$$\mu_i = \frac{1}{N_{train}} \sum_{j=1}^{N_{train}} \phi_i(I_j)$$

$$\psi_i(I, \mu) = \sum_{z}^{\lambda_i} \frac{|\phi_i(I)_z - \mu_{iz}|}{|\phi_i(I)_z| + |\mu_{iz}|}$$

- Intermediate representations computed for an arbitrary image I can be compared with the layer-wise means

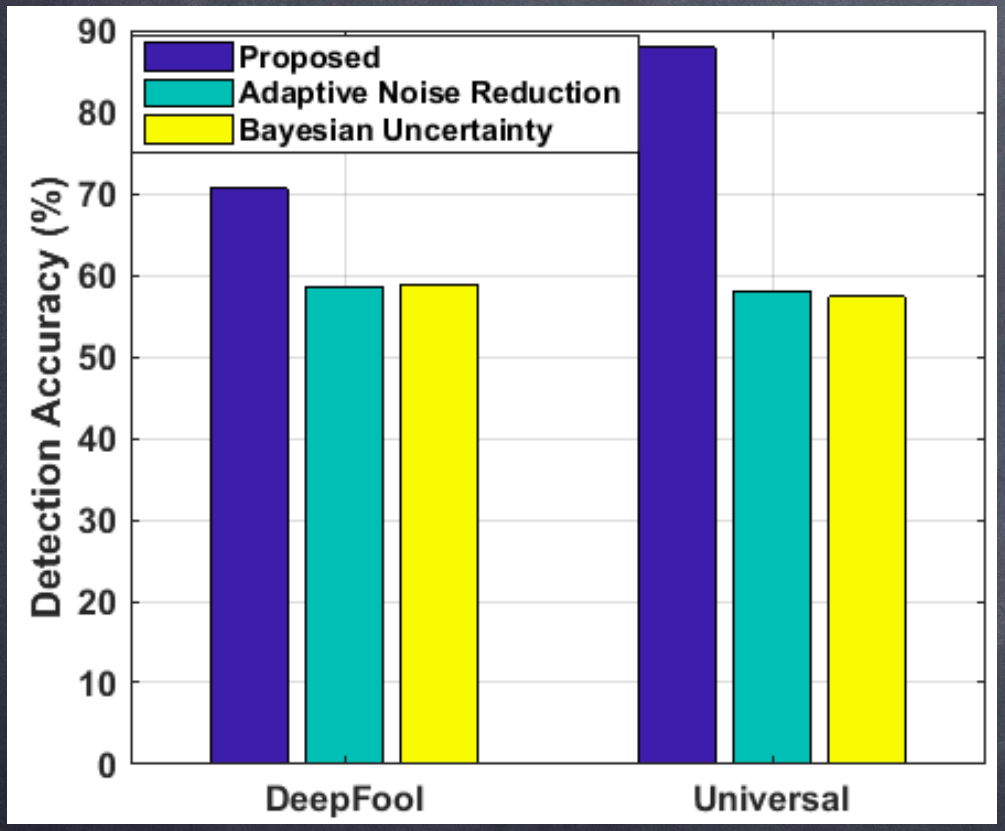
Adversarial Perturbation Detection



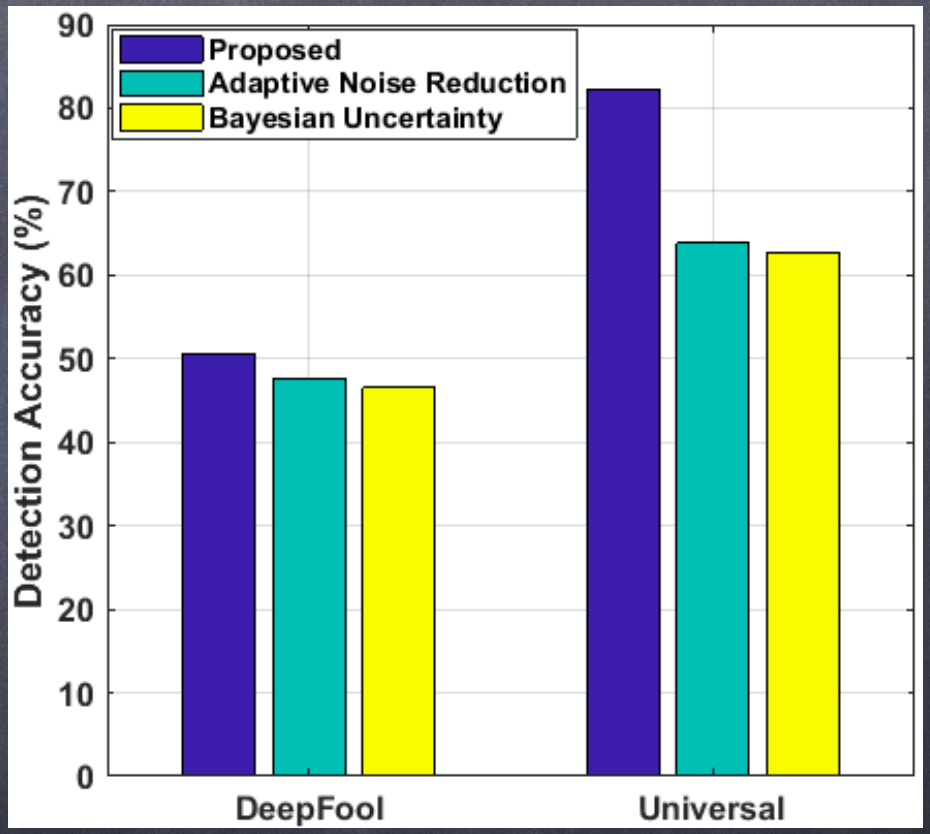
White-box Training

Testing

Detection Results



PaSC database



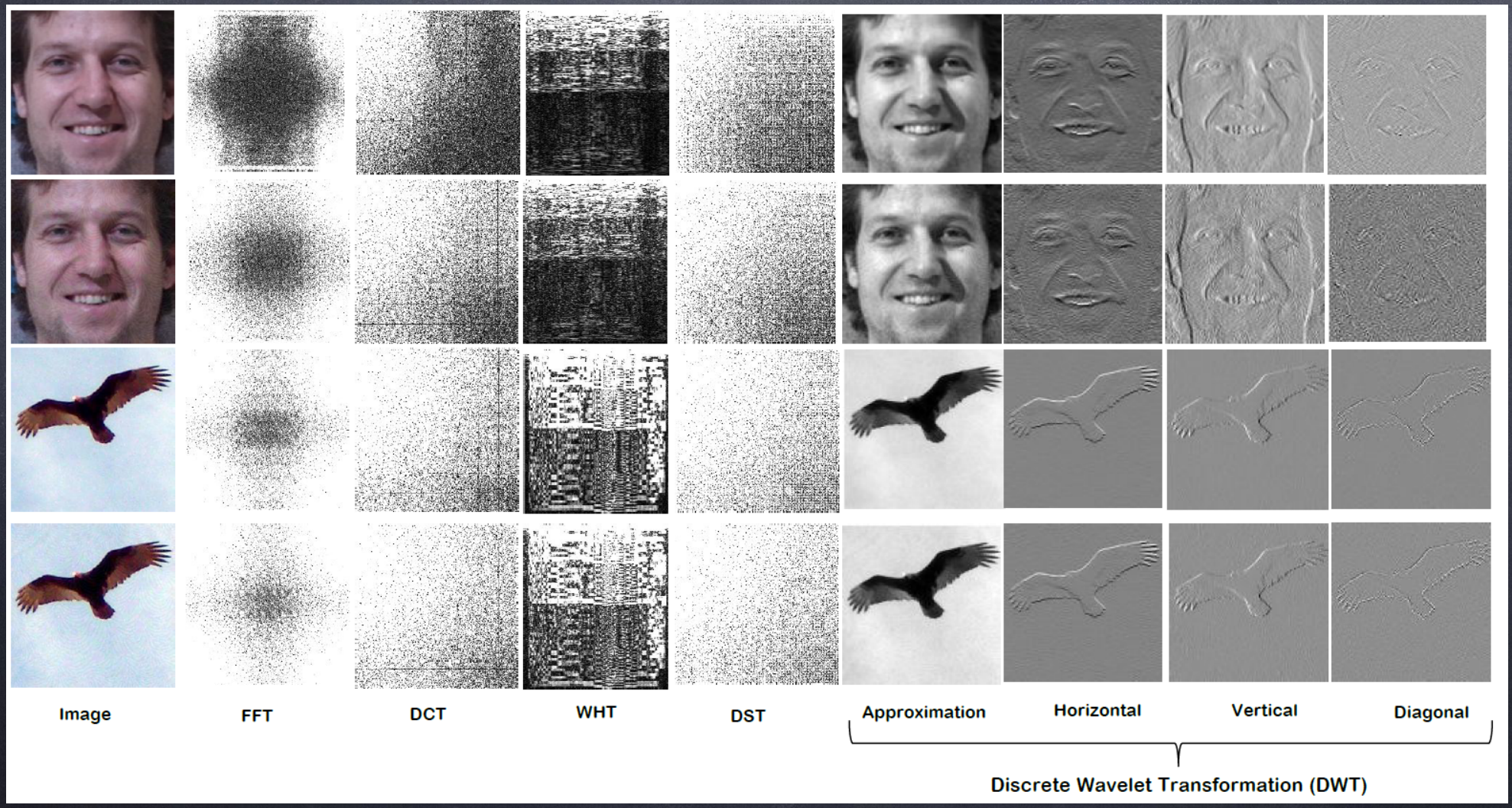
MEDS database

Goswami et al., Unravelling Robustness of Deep Learning based Face Recognition Against Adversarial Attacks, AAAI 2018

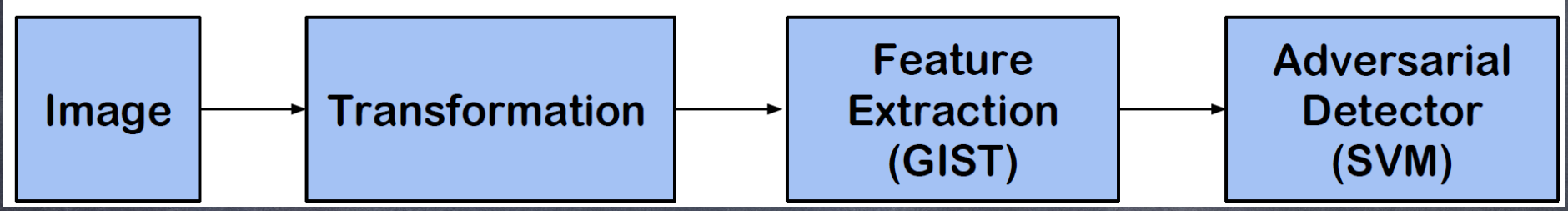
Other Methods

	Authors	Descriptions
Detection	Grosse et al., 2017	Statistical test for adversarial and original data distribution
	Gong, Wang, and Ku; Metzen et al., 2017	Neural network based classification
	Feinman et al., 2017	Randomized network using Dropout at both training and testing
	Lu, Issaranon, and Forsyth, 2017	Quantize ReLU output for discrete code + RBF SVM
	Das et al., 2017	JPEG compression to reduce the effect of adversary
	Li & Li, 2017	CNN maps + PCA statistics + Cascade SVM

Let us Look at Transformations



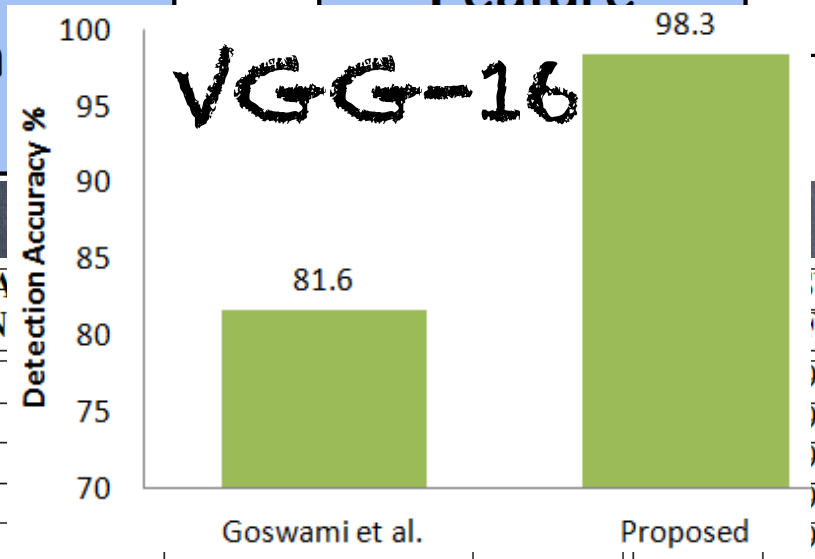
Non-Deep Learning Approach



Database	DNN Model	Attack	Adaptive Noise [31]	Bayesian Uncertainty [14]	GIST Features + SVM Classification (Proposed)					
					Image	DCT	FFT	DWT	DST	WHT
MEDS	VGG-16	Universal	80.2	80.3	96.4	57.4	96.3	98.3	94.3	78.5
		F3	79.6	79.9	96.5	61.6	96.9	98.3	96.5	88.0
	GoogLeNet	Universal	79.2	79.9	92.6	60.5	97.1	99.4	97.0	85.3
		F3	77.0	77.3	93.1	60.3	97.8	97.2	93.1	83.4
	CaffeNet	Universal	78.9	78.4	94.01	59.0	92.9	98.2	95.1	82.3
		F3	78.8	78.5	99.2	67.5	97.6	99.8	99.2	88.6
Multi-PIE	VGG-16	Universal	75.5	74.7	99.9	57.7	100.0	100.0	99.6	93.0
		F3	76.0	75.0	99.9	61.8	100.0	99.9	99.9	98.9
	GoogLeNet	Universal	69.4	69.8	99.9	61.8	100.0	99.9	100.0	98.9
		F3	70.2	70.5	99.9	59.8	100.0	99.9	99.9	99.0
	CaffeNet	Universal	71.1	70.3	100.0	58.2	100.0	99.9	99.9	97.4
		F3	70.2	69.6	99.9	67.1	100.0	100.0	100.0	99.0

Agrawal et al. 2018

Non-Deep Learning Approach

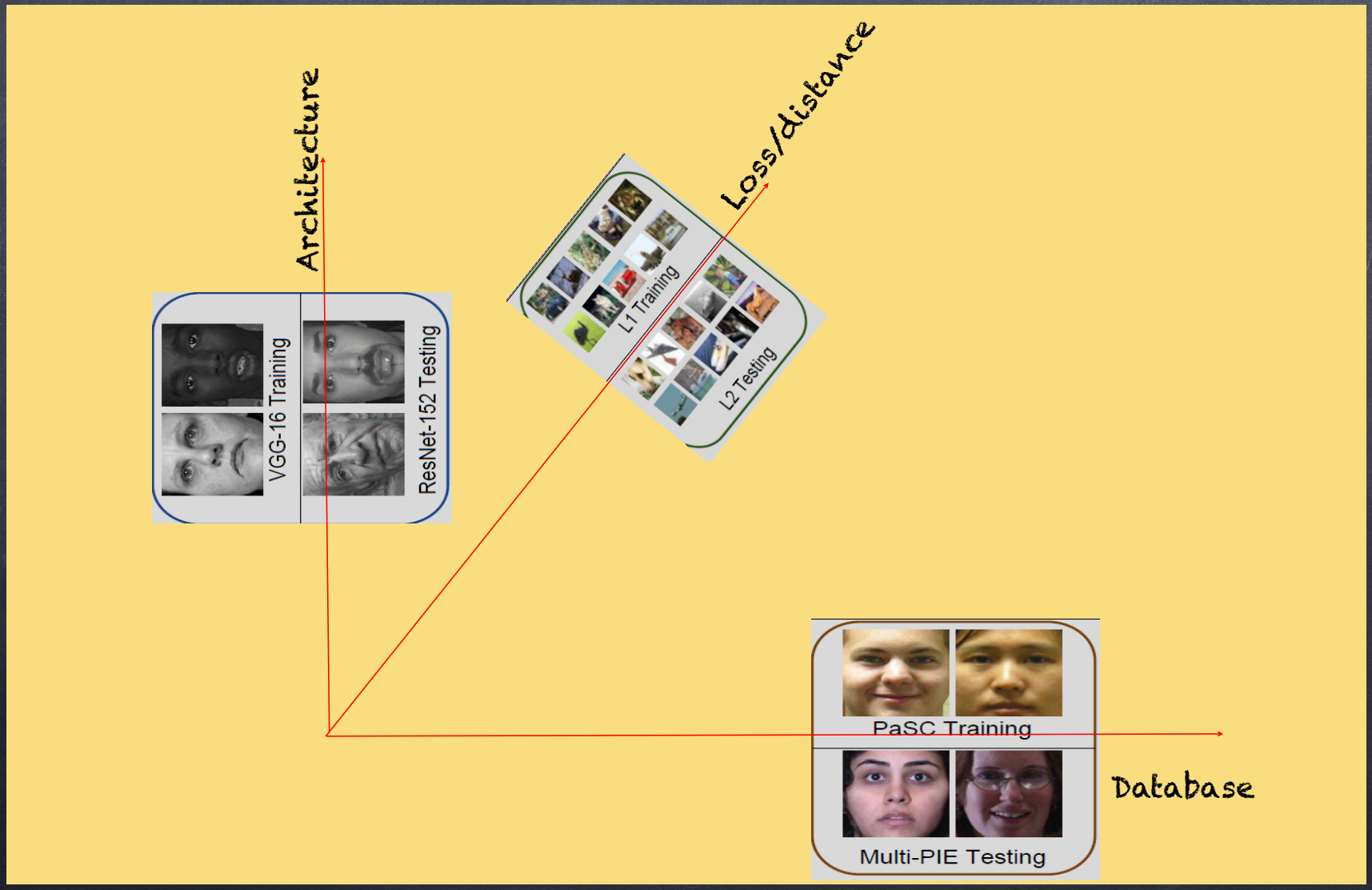


Database	DNN Model	Attack	A N			SVM Classification (Proposed) FT	DWT	DST	WHT
MEDS	VGG-16	Universal				96.3	98.3	94.3	78.5
		F3				96.9	98.3	96.5	88.0
	GoogLeNet	Universal				97.1	99.4	97.0	85.3
		F3				97.8	97.2	93.1	83.4
	CaffeNet	Universal				92.9	98.2	95.1	82.3
		F3	78.8	78.5	99.2	67.5	97.6	99.8	99.2	88.6
Multi-PIE	VGG-16	Universal	75.5	74.7	99.9	57.7	100.0	100.0	99.6	93.0
		F3	76.0	75.0	99.9	61.8	100.0	99.9	99.9	98.9
	GoogLeNet	Universal	69.4	69.8	99.9	61.8	100.0	99.9	100.0	98.9
		F3	70.2	70.5	99.9	59.8	100.0	99.9	99.9	99.0
	CaffeNet	Universal	71.1	70.3	100.0	58.2	100.0	99.9	99.9	97.4
		F3	70.2	69.6	99.9	67.1	100.0	100.0	100.0	99.0

Agrawal et al. 2018

Some Extensions:
Effective Perturbation Detection

Image Agnostic, Model Agnostic, Database Agnostic



Detection: Key Takeout

- Detection is an important step to check if the systems are attacked or not
- Solution may lie in non-DL domain

How to mitigate the effect of attacks?

A Simple Approach

A Simple Approach

- White-box approach: retrain the model with original and perturbed samples
- What is the problem with this approach?

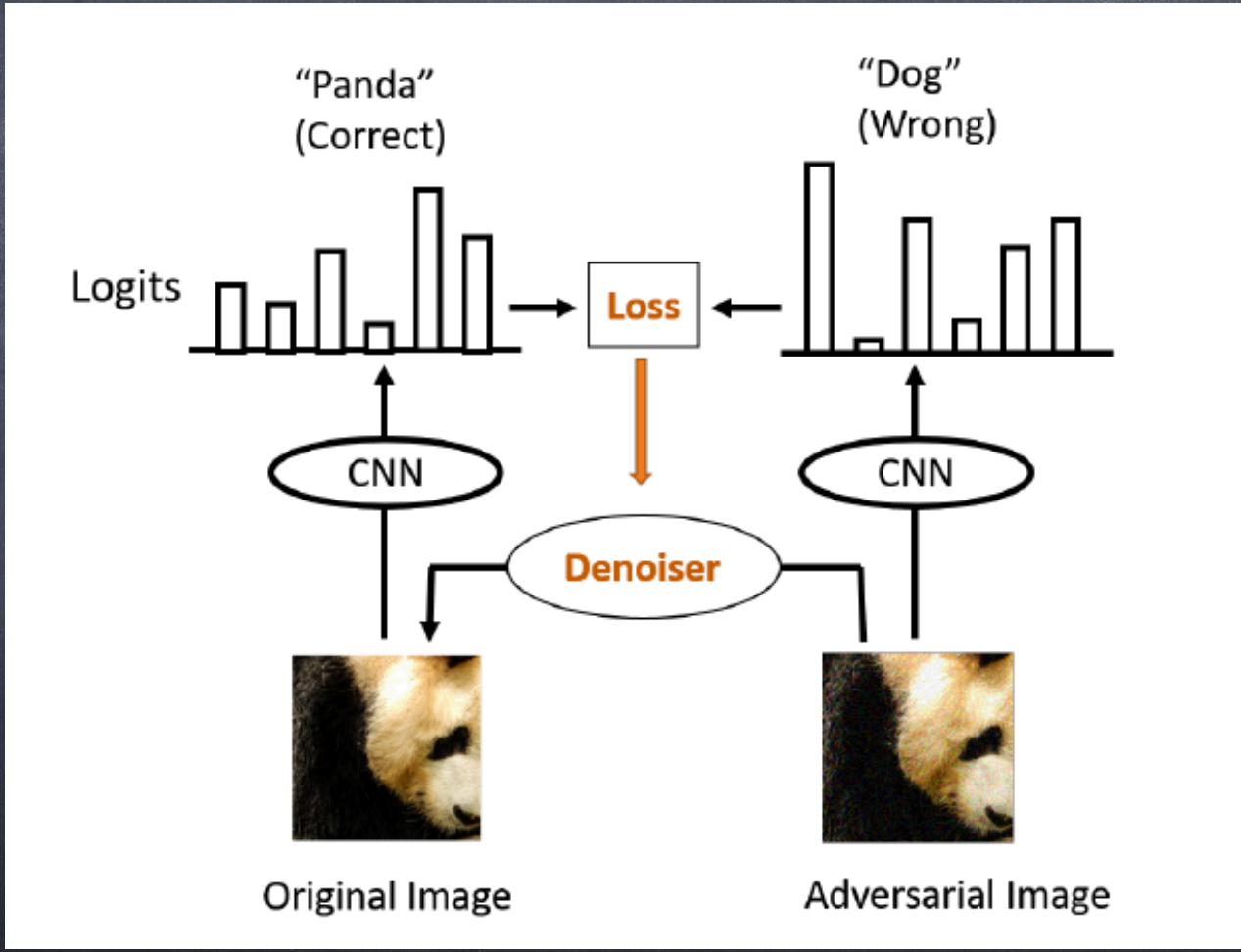
A Simple Approach

- White-box approach: retrain the model with original and perturbed samples
- What is the problem with this approach?
- A new attack is proposed and we have to start the training process again :)

Another simple approach

- Transform an input image:
- e.g. apply Gaussian blur and then proceed with classification
- Pixel Deflection (CVPR2018),

Image Denoiser

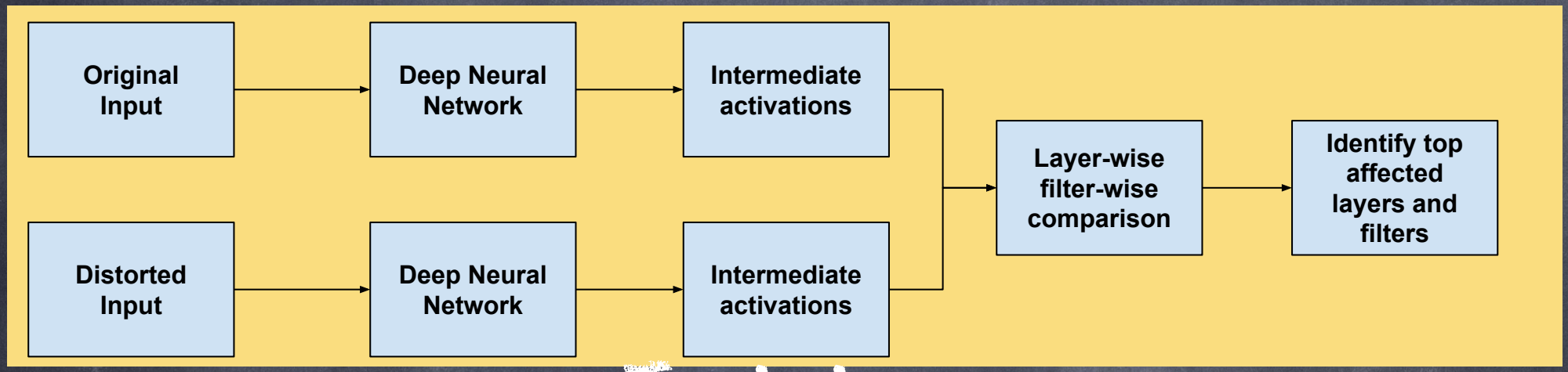


Liao et al, CVPR2018

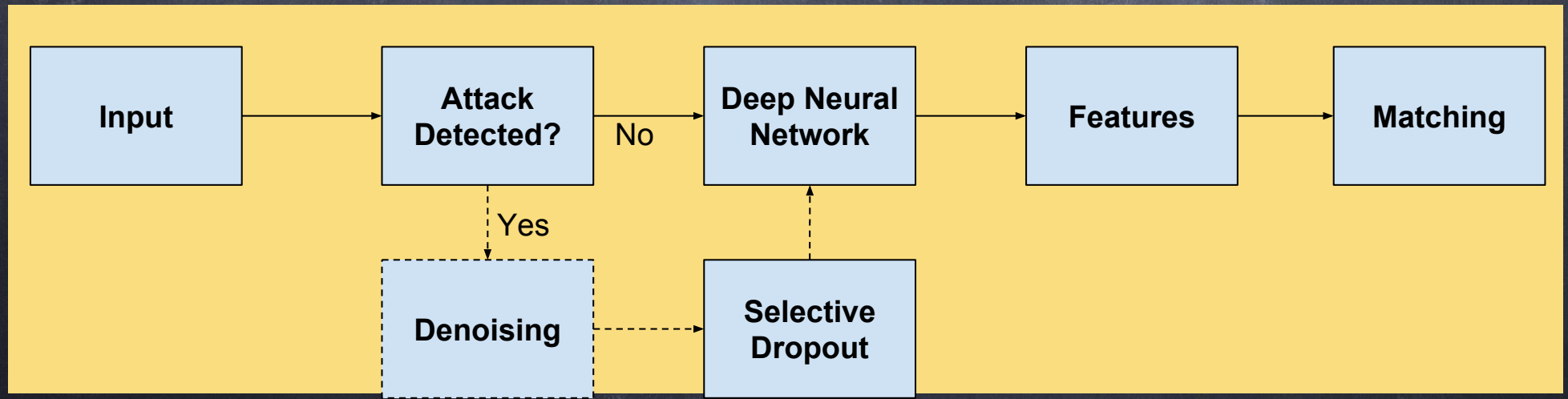
Modified Approach

- Defense-GAN (ICLR2018)
- Train a WGAN trained on legitimate (un-perturbed) training samples to "denoise" adversarial examples
- Prior to feeding a test image x to the classifier, it is projected onto the range of the generator by minimizing the reconstruction error $||G(z) - x||$
- The resulting reconstruction G(z) is then given to the classifier for classification task
- Since the generator was trained to model the unperturbed training data distribution, this added step "removes" any potential adversarial noise.

Adversarial Perturbation Mitigation



Training



Testing

Goswami et al. AAAI2018

Results of Adversary Mitigation

Algorithm	Original	Distorted	Corrected
LightCNN	60.5	25.9	36.2
	89.3	41.6	61.3
VGG-Face	54.3	14.6	24.8
	78.4	30.5	40.6

Mitigation Results on face database

Catalog of Defense Approaches

- Reactive vs proactive
- Detection vs transformation vs training vs architecture vs generative

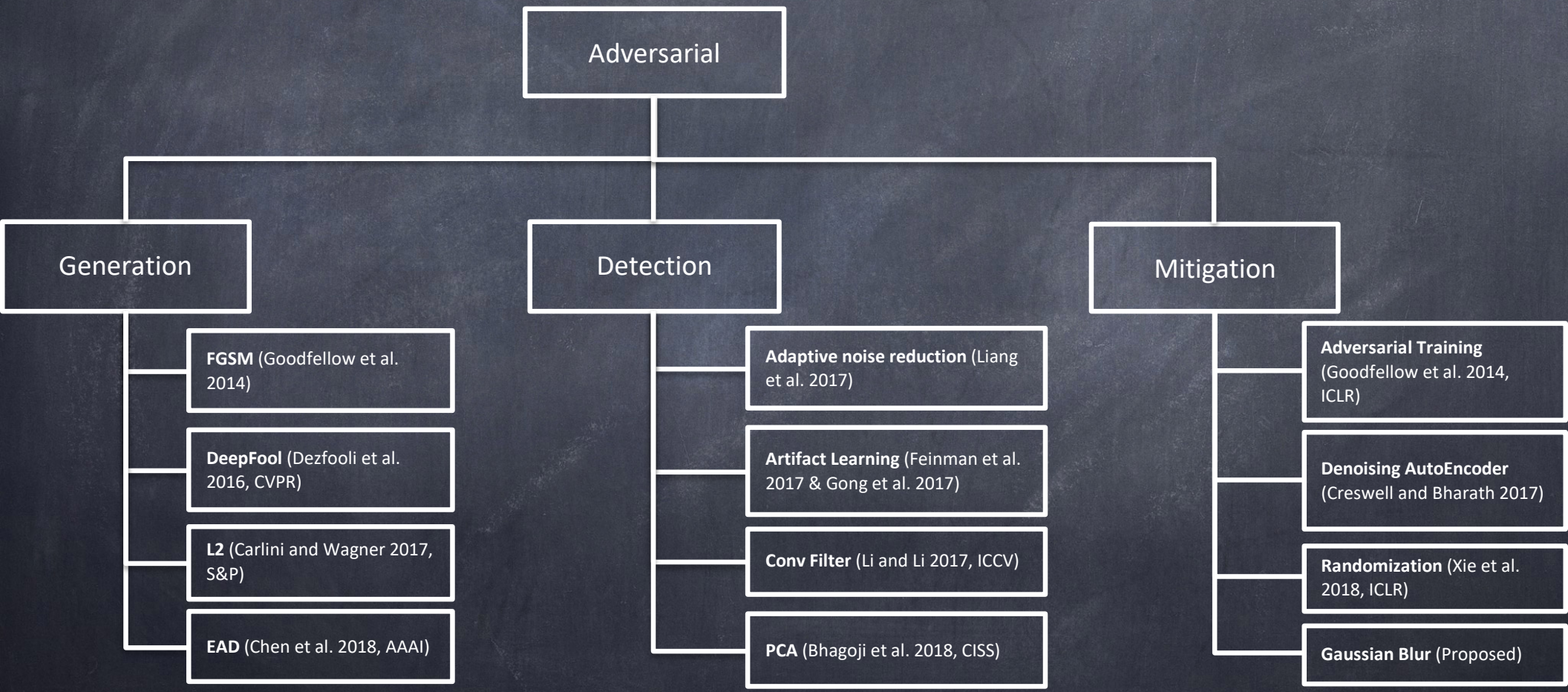
Defence	Type	Method
Statistical Detection [75]	Reactive	Detection
Binary Classification [67]	Reactive	Detection
In-Layer Detection [130]	Reactive	Detection
Detecting from Artifacts [59]	Reactive	Detection
SafetyNet [124]	Reactive	Detection
Saliency Data Detector [207]	Reactive	Detection
Linear Transformations Detector [16]	Reactive	Detection
Key-based Networks [210]	Reactive	Detection
Ensemble Detectors [1]	Reactive	Detection
Generative Detector [116]	Reactive	Detection
Convolutional Statistics Detector [118]	Reactive	Detection
Feature Squeezing [203]	Reactive	Detection
PixelDefend [177]	Reactive	Detection
MagNet [129]	Reactive	Detection
VAE Detector [62]	Reactive	Detection
Bit-Depth [78]	Reactive	Input Transformation
Basis Transformations [168]	Reactive	Input Transformation
Randomised Transformations [201]	Reactive	Input Transformation
Thermometer Encoding [24]	Reactive	Input Transformation
Blind Pre-Processing [153]	Reactive	Input Transformation
Data Discretisation [32]	Reactive	Input Transformation
Adaptive Noise [119]	Reactive	Input Transformation
FGSM Training [70]	Proactive	Training
Gradient Training [175]	Proactive	Training
Gradient Regularisation [127]	Proactive	Training
Structured Gradient Regularisation [158]	Proactive	Training
Robust Training [169]	Proactive	Robust Training
Strong Adversary Training [90]	Proactive	Robust Training
CFOA Training [128]	Proactive	Robust Training
Ensemble Training [188]	Proactive	Robust Training
Stochastic Pruning [44]	Proactive	Robust Training
Distillation [86]	Proactive	Architecture
Parseval Networks [37]	Proactive	Architecture
Deep Contractive Networks [77]	Proactive	Architecture
Biological Networks [139]	Proactive	Architecture
DeepCloak [60]	Proactive	Architecture
Fortified Networks [111]	Proactive	Architecture
Rotation-Equivariant Networks [48]	Proactive	Architecture
HyperNetworks [180]	Proactive	Architecture
Bidirectional Networks [151]	Proactive	Architecture
DAM [108]	Proactive	Architecture
Certified Defences [152]	Proactive	Certified
Formal Tools [98, 51, 92, 161]	Proactive	Certified
Distributional Robustness [176]	Proactive	Certified
Convex Outer Polytope [102]	Proactive	Certified
Lischitz Margin [191]	Proactive	Certified
Defence Gan [165]	Proactive	Generative
FB-GAN [9]	Proactive	Genearative

Toolboxes: SmartBox

- Lack of a benchmark platform to standardize research efforts in attack, detection and mitigation
- SmartBox: Benchmarking Adversarial Detection and Mitigation Algorithms for Face Recognition

Goel et al. Benchmarking Adversarial Detection and Mitigation Algorithms for Face Recognition, IEEE BTAS, 2018

SmartBox


Adversarial
Generation
Detection
Mitigation
FGSM (Goodfellow et al. 2014)
DeepFool (Dezfooli et al. 2016, CVPR)
L2 (Carlini and Wagner 2017, S&P)
EAD (Chen et al. 2018, AAAI)
Adaptive noise reduction (Liang et al. 2017)
Artifact Learning (Feinman et al. 2017 & Gong et al. 2017)
Conv Filter (Li and Li 2017, ICCV)
PCA (Bhagoji et al. 2018, CISS)
Adversarial Training (Goodfellow et al. 2014, ICLR)
Denoising AutoEncoder (Creswell and Bharath 2017)
Randomization (Xie et al. 2018, ICLR)
Gaussian Blur (Proposed)

Other Toolboxes

- CleverHans
- Foolbox
- Adversarial Robustness Toolbox

Databases used to benchmark

- PaSC, MultiPIE, CelebA
- MNIST, F-MNIST
- CIFAR-10, CIFAR-100
- ImageNET
- SVHN

Defence	Datasets	Models
Statistical Detection [75]	MNIST, DREBIN, MicroRNA	DT, SVM, 2 layers-CNN
Binary Classification [67]	MNIST, CIFAR-10, SVHN	AlexNet
In-Layer Detection [130]	CIFAR-10, 10-class ImageNet	ResNet
Detecting from Artifacts [59]	MNIST, CIFAR-10, SVHN	LeNet, 12-layer CNN
SafetyNet [124]	CIFAR-10, ImageNet-1000	ResNet, VGG19
Saliency Data Detector [207]	MNIST, CIFAR-10, ImageNet	AlexNet, AlexNet, VGG19
Linear Transformations Detector [16]	MNIST, HAR	SVM
Key-based Networks [210]	MNIST	2/3-layers CNN
Ensemble Detectors [1]	MNIST, CIFAR-10	3-layers CNN
Generative Detector [116]	CIFAR-10, CIFAR-100	6-layers CNN
Convolutional Statistics Detector [118]	ImageNet	VGG-16
Feature Squeezing [203]	MNIST, CIFAR-10, ImageNet	7-layers CNN, DenseNet MobileNet
PixelDefend [177]	ImageNet	ResNet, VGG
MagNet [129]	MNIST, CIFAR-10	4/9-layers CNN
VAE Detector [62]	MNIST, SVNH, COIL-100	-
Bit-Depth [78]	ImageNet	ResNet, DenseNet, Inception-v4
Basis Transformations [168]	ImageNet	Inception-v3, Inception-v4
Randomised Transformations [201]	ImageNet	Inception-v3, ResNet
Thermometer Encoding [24]	MNIST, CIFAR-10, CIFAR-100, SVHN	30-layers CNN, Wide ResNet
Blind Pre-Processing [153]	MNIST, CIFAR-10, SVHN	LeNet, ResNet-50, ResNet-18
Data Discretisation [32]	MNIST, CIFAR-10, ImageNET	InceptionResnet-V2
Adaptive Noise [119]	MNIST, ImageNet	-
FGSM Training [70]	MNIST	Maxout
Gradient Training [175]	CIFAR-10, SVHN	ResNet-18
Gradient Regularisation [127]	MNIST, CIFAR-10	Maxout
Structured Gradient Regularisation [158]	MNIST, CIFAR-10	9-layers CNN
Robust Training [169]	MNIST, CIFAR-10	2-layers CNN, VGG
Strong Adversary Training [90]	MNIST, CIFAR-10	MxNet
CFOA Training [128]	MNIST, CIFAR-10	2/4/6-layers CNN, Wide ResNet
Ensemble Training [188]	ImageNet	ResNet, InceptionResNet-v2
Stochastic Pruning [44]	CIFAR-10	Resnet-20
Distillation [86]	MNIST, CIFAR-10	4-layers CNN
Parseval Networks [37]	MNIST, CIFAR-10, CIFAR-100, SVHN	ResNet, Wide Resnet
Deep Contractive Networks [77]	MNIST	LeNet, AlexNet
Biological Networks [139]	MNIST	3-layers CNN
DeepCloak [60]	CIFAR-10	ResNet-164
Fortified Networks [111]	MNIST	2-layers CNN
Rotation-Equivariant Networks [48]	CIFAR-10, ImageNet	ResNet
HyperNetworks [180]	ImageNet	ResNet
Bidirectional Networks [151]	MNIST, CIFAR-10	3-layers CNN
DAM [108]	MNIST	DAM
Certified Defences [152]	MNIST	2-layers FC
Formal Tools [98, 51, 92, 161]	-	-
Distributional Robustness [176]	MNIST	3-layers CNN
Convex Outer Polytope [102]	MNIST, F-MNIST	2-layers CNN
Lischitz Margin [191]	SVHN	Wide ResNet
Defence Gan [165]	MNIST, F-MNIST	Defene-GAN
FB-GAN [9]	MNIST, F-MNIST	8-layers CNN

Cat and Mouse Game

Cat and Mouse Game

- On the Robustness of the CVPR 2018 White-Box Adversarial Example Defenses
- "we evaluate the two white-box defenses that appeared at CVPR 2018 and find they are ineffective: when applying existing techniques, we can reduce the accuracy of the defended models to 0%."

Athalye and Carlini, 2018

Key Takeout

- Defense mechanism has to be model, database, and attack agnostic
- It will be always be a game between an adversary and a defender

Is adversarial perturbation always bad?

Two Approaches

- Privacy Preserving Adversarial Perturbation
- Data Fine-tuning

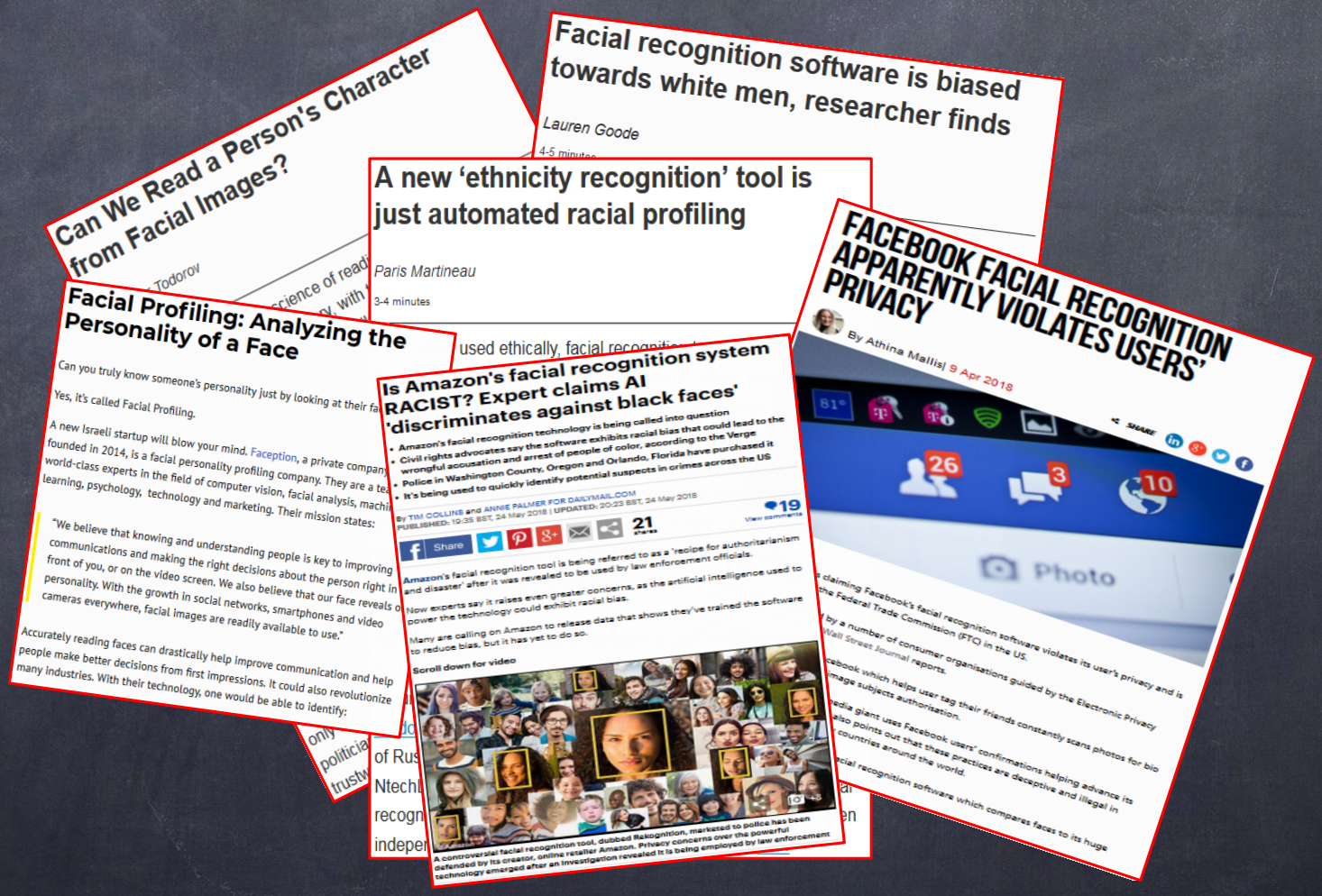
Privacy Preserving Adversarial Perturbation

Chabbra et al. IJCAI2018

Adversarial Perturbations - The Positive Side

- While attackers have used adversarial perturbations to "fool" biometrics/face recognition systems, it can be used for assisting in privacy-preserving aspect ...

Face Analysis - In the News



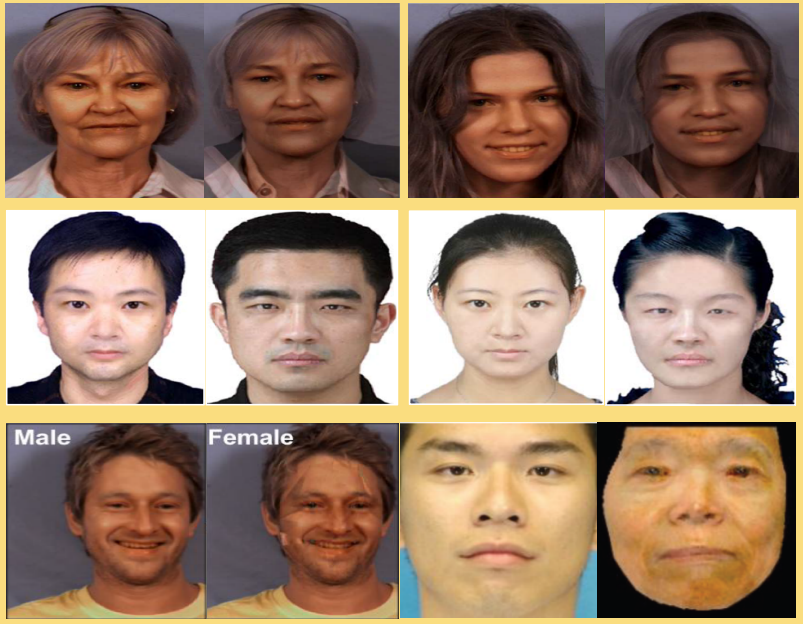
Right to Privacy

- Automated face analysis pose threat to the privacy of an individual
  - Wang and Kosinksi predicted the sexual orientation from face images
  - Facial attributes such as age, gender, and race can be predicted from one's profile or social media images
  - Profiling of a person using his face image in ID card
  - Identity theft using cross database matching

Yilun Wang and Michal Kosinski. Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. PsyArXiv preprint arXiv: 10.17605/OSF.IO/HV28A, 2017.

Literature

Author	Method	No. of Attributes	Dataset	Controlling Attributes
Othman and Ross, 2014	Face Morphing and fusion	One	MUCT	No
Mirjalili and Ross, 2017	Delaunay Triangulation and fusion	One	MUCT, LFW	No
Mirjalili *et al.*, 2017	Fusion using Convolutional Autoencoder	One	MUCT, LFW, Celeb-A, AR-Face	No
Rozsa et al., 2016, 2017	Fast Flipping Attribute	Multiple	CelebA	No
Chhabra et al., 2018	Adversarial Perturbation	Multiple	CelebA, MUCT, LFW	Yes

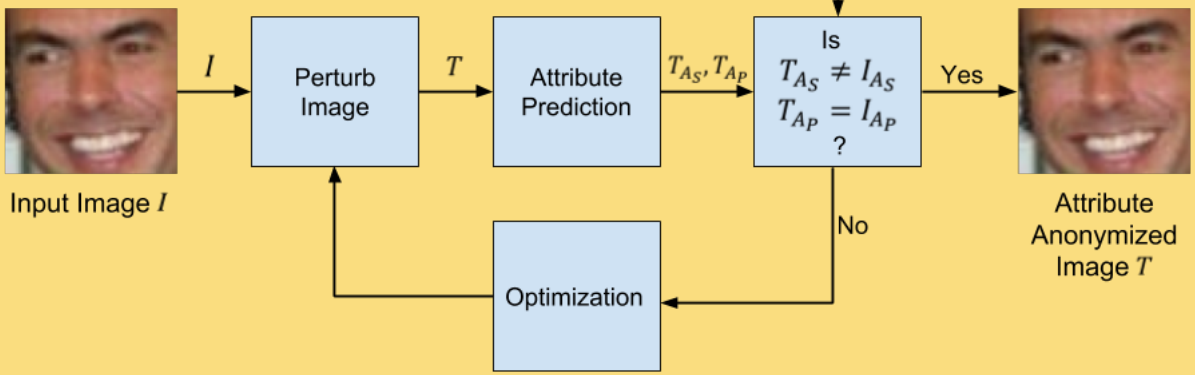


Three Key Factors

- While anonymizing facial attributes, there should be no visual difference between original and anonymized images
- Selectively anonymizing few and retaining some attributes require a "control" mechanism
- In face recognition applications, identity should be preserved while anonymizing attributes.

Anonymizing *k*-Facial Attributes via Adversarial Perturbations

Overview of the Proposed Approach



I -> input image

T-> perturbed image (T = I + ω)

I_{AS} -> Attributes to be suppressed

I_{AP} -> Attributes to be preserved

Loss Function

Attributes only

Attribute Anonymization Visual Appearance

$$\min \; \left[D\left(\boldsymbol{I}_{\boldsymbol{A_P}}, \boldsymbol{T}_{\boldsymbol{A_P}}\right) - D\left(\boldsymbol{I}_{\boldsymbol{A_S}}, \boldsymbol{T}_{\boldsymbol{A_S}}\right)\right] + ||\boldsymbol{I} - \boldsymbol{T}||_2^2$$

$$\text{such that} \quad \boldsymbol{T}_{\boldsymbol{A_S}} \neq \boldsymbol{I}_{\boldsymbol{A_S}}, \boldsymbol{T}_{\boldsymbol{A_P}} = \boldsymbol{I}_{\boldsymbol{A_P}}$$

Attributes + Identity

$$\min \; \left\{f(\boldsymbol{T}) + \left|\left|\boldsymbol{I} - \boldsymbol{T}\right|\right|_2^2 + D(\boldsymbol{Id}_{\boldsymbol{I}}, \boldsymbol{Id}_{\boldsymbol{T}})\right\}$$

Chhabra et al. IJCAI 2018

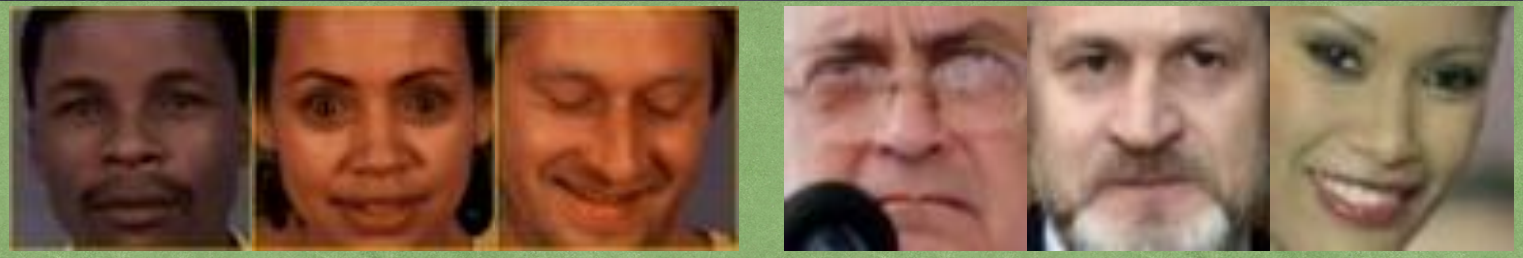
Experiments

Experiment	Dataset	# Attributes Anonymized	Attributes Anonymized	
			Suppressed	Preserved
Single Attribute	MUCT, CelebA, LFWCrop	1	Gender	-
Multiple Attributes	CelebA	3, 5	Gender, Attractive, Smiling	Heavy makeup, High cheekbones
Identity Preservation	MUCT, LFWcrop	1+1	Gender	Identity

Single Attribute

MUCT dataset

LFWcrop dataset

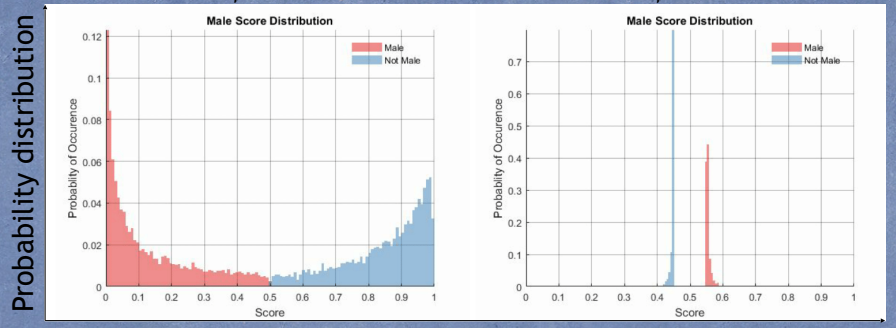


Original Images

Gender Attribute Anonymized Images

Before

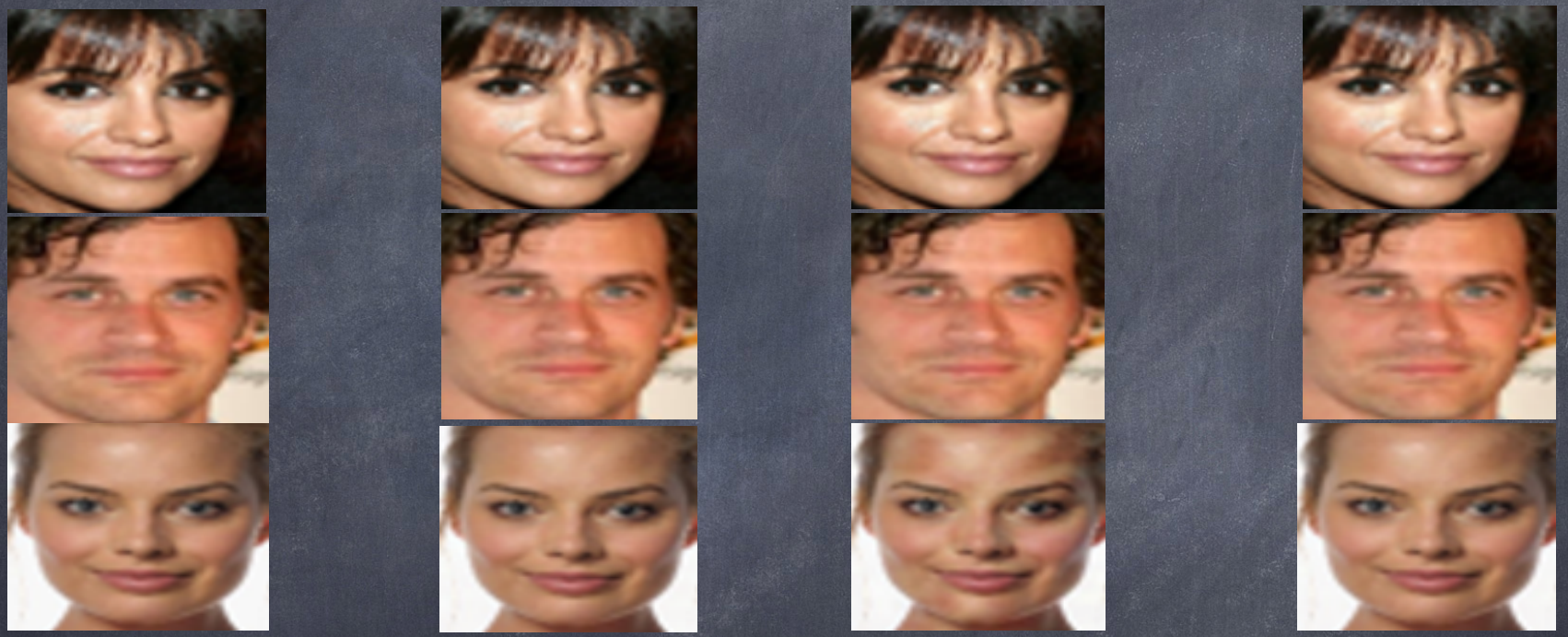
After



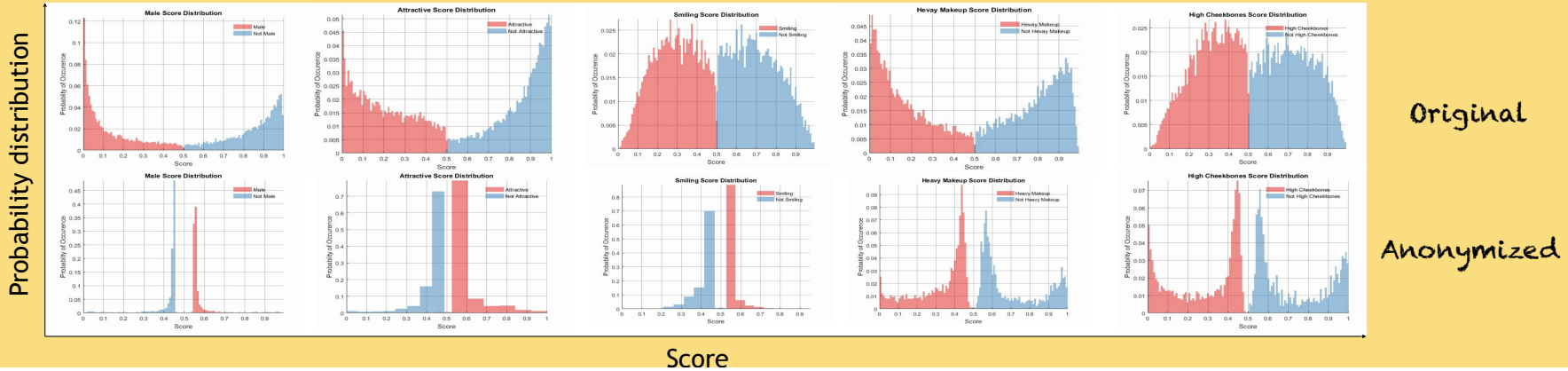
Probability distribution

Score

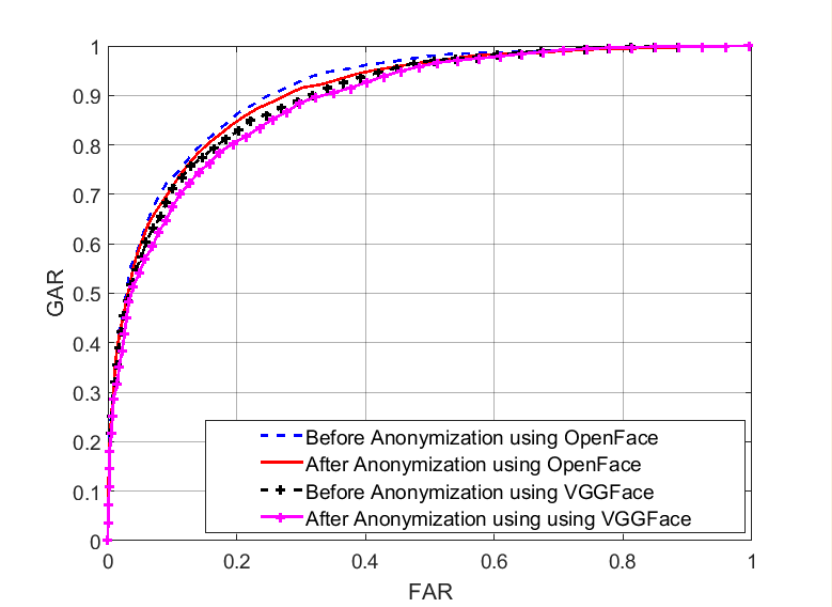
Attribute Suppression and Preservation



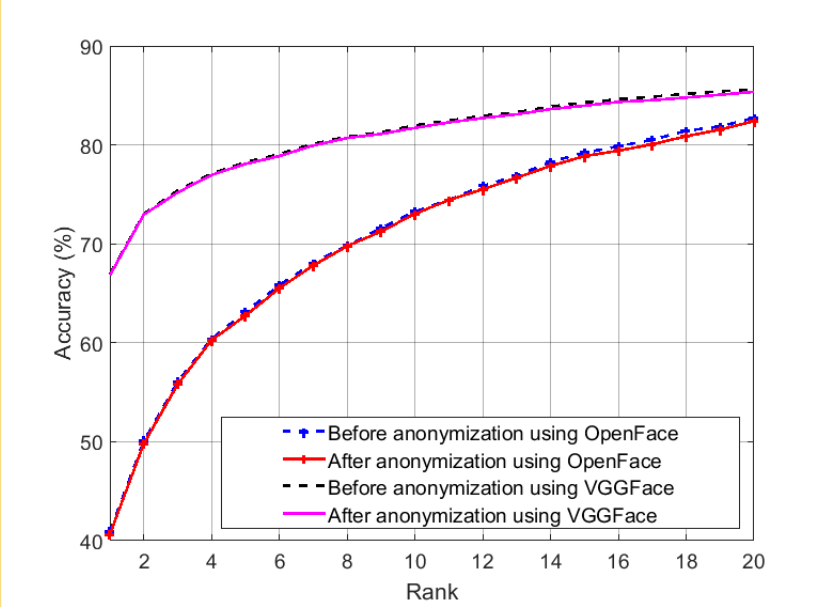
Original | One attribute | Three attributes | Five attributes



Attribute Suppression with Identity Preservation



ROC curves on the LFWcrop dataset



CMC curve on the MUCT dataset

Key Takeouts

- Adversarial perturbations can be used positively for privacy preserving applications

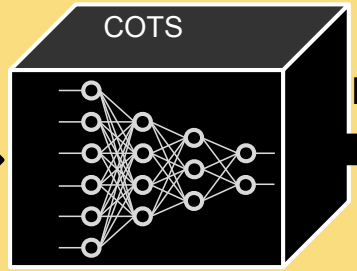
Data Fine-Tuning

- In DL, traditionally, we perform model fine-tuning, if we have access to the model

Chabbra et al. AAAI2019

In Real World Applications

Input



Decision

Expectation

Reality

In Real World Applications

Dataset	Model Access	Model Hyperparameters
✓	✗	✗

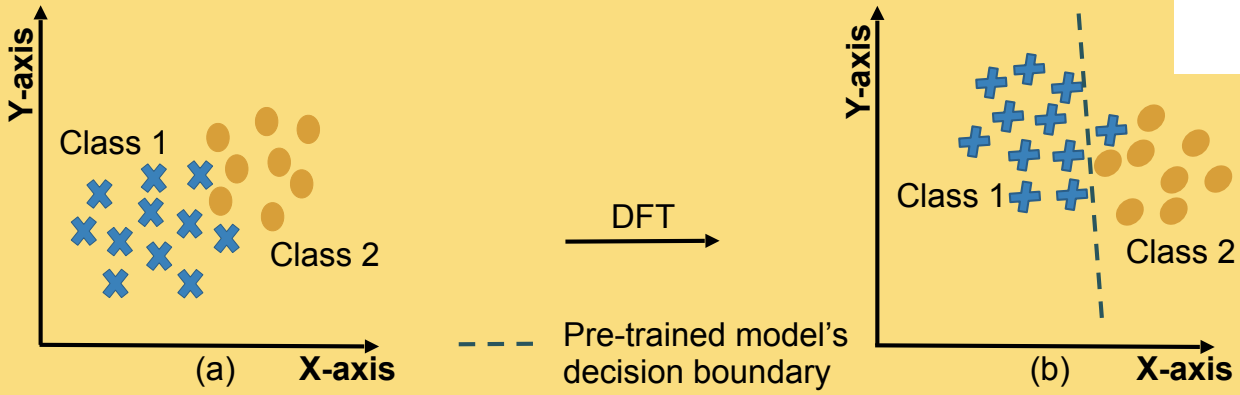
Output of Model	Model Training
✓	✗

Can we enhance the performance of a black-box system?

Data Fine-tuning

- Data Fine-tuning (DFT)

$$\Phi(\boldsymbol{W}\boldsymbol{X} + b) \xrightarrow{\text{DFT}} \Phi(\boldsymbol{W}\boldsymbol{Z} + b)$$



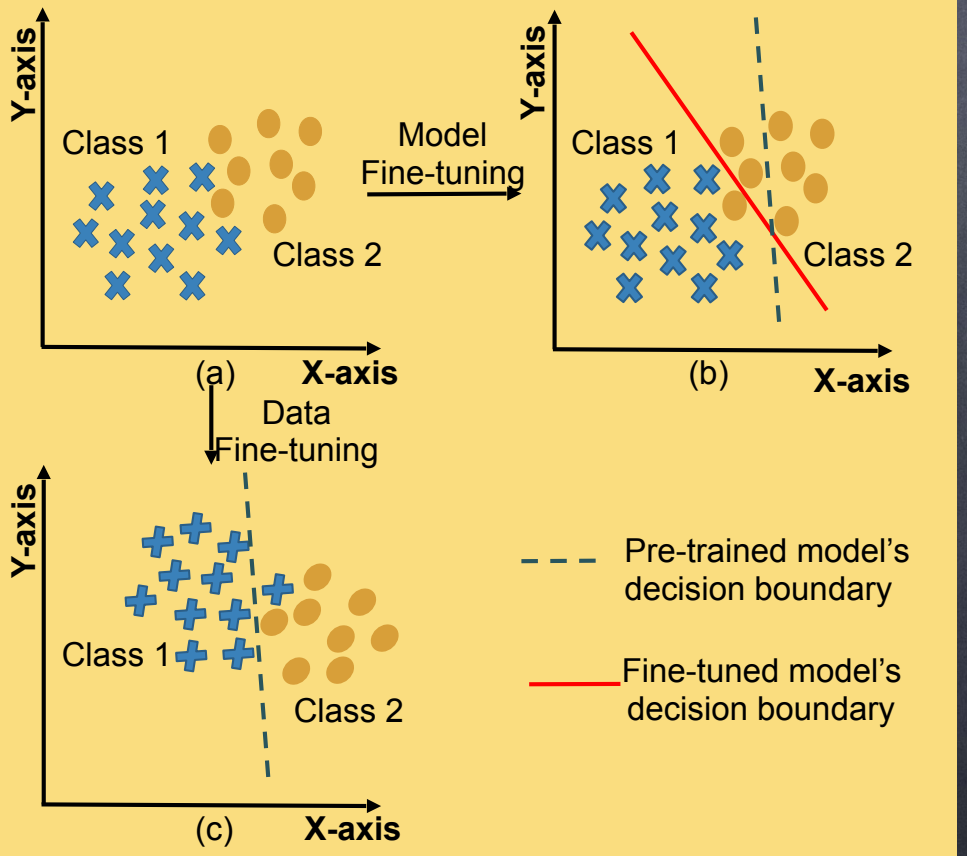
Model Fine-tuning vs Data Fine-tuning

Model Fine-tuning

$$\Phi(\boldsymbol{W}\boldsymbol{X}+b) \xrightarrow{\text{MFT}} \Phi(\boldsymbol{W}'\boldsymbol{X}+b')$$

Data Fine-tuning

$$\Phi(\boldsymbol{W}\boldsymbol{X}+b) \xrightarrow{\text{DFT}} \Phi(\boldsymbol{W}\boldsymbol{Z}+b)$$



Data Fine-tuning

- Learn a single perturbation for a given dataset
- The visual appearance of the image should be preserved after performing data fine-tuning

Optimization

X ——→ Original Training Set

y ——→ True Labels

m ——→ Number of Images

Z ——→ Perturbed Training Set

N ——→ Perturbation

A ——→ Set of Attributes

$$Z_k = \frac{1}{2}(tanh(X_k + N) + 1)$$

Transform image in range of 0 to 1

Output scores

Model Input

$$P(A_i \mid Z_k) = \Phi_{A_i}(Z_k, W, b)$$

Enforces the outputs scores towards true labels

$$\min_{N} \frac{1}{m} \sum_{k=1}^{m} \max(0, 1 - y_{i,k}^{T} P(A_i \mid Z_k))$$

Illustration of Data Fine-tuning for Attribute Prediction

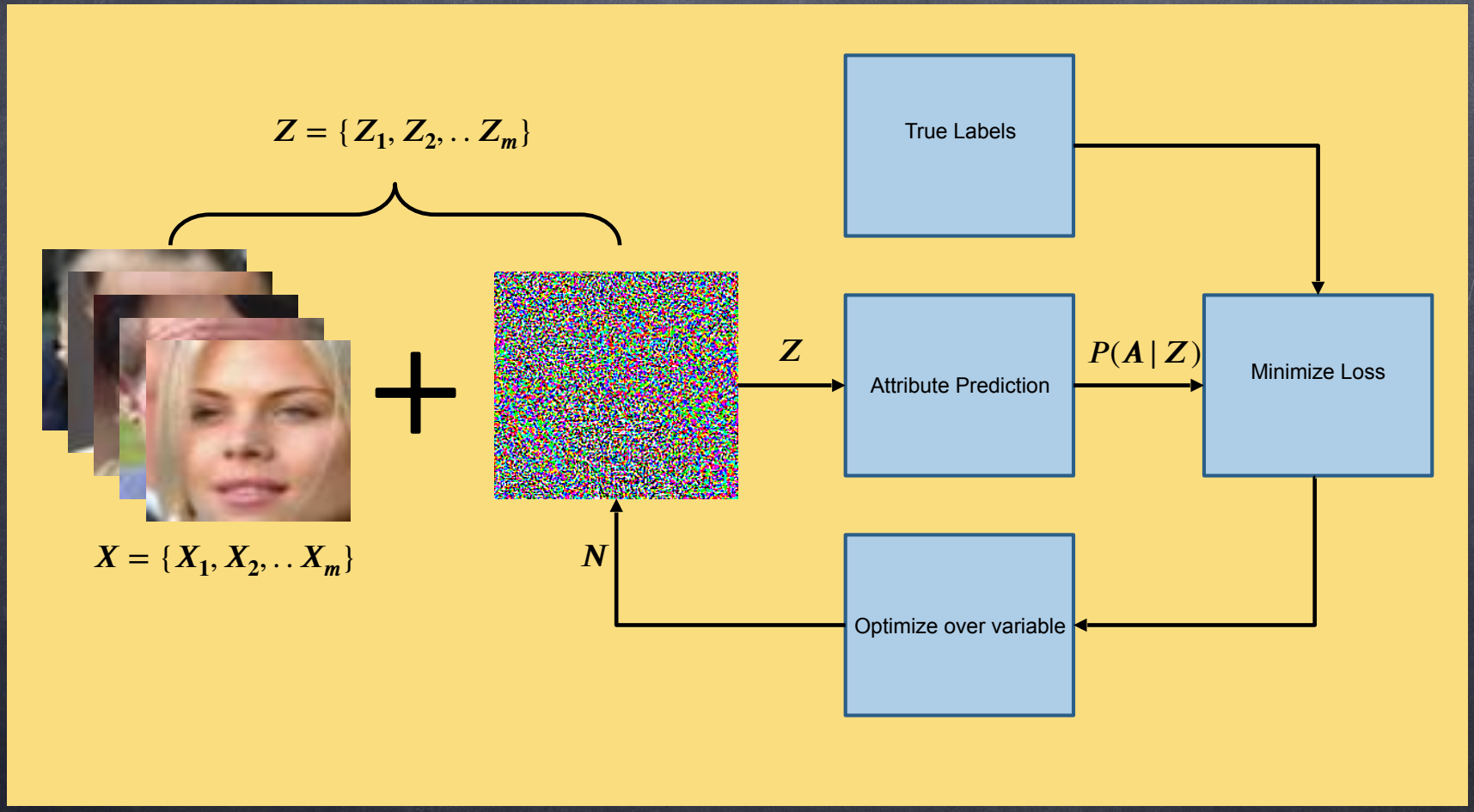
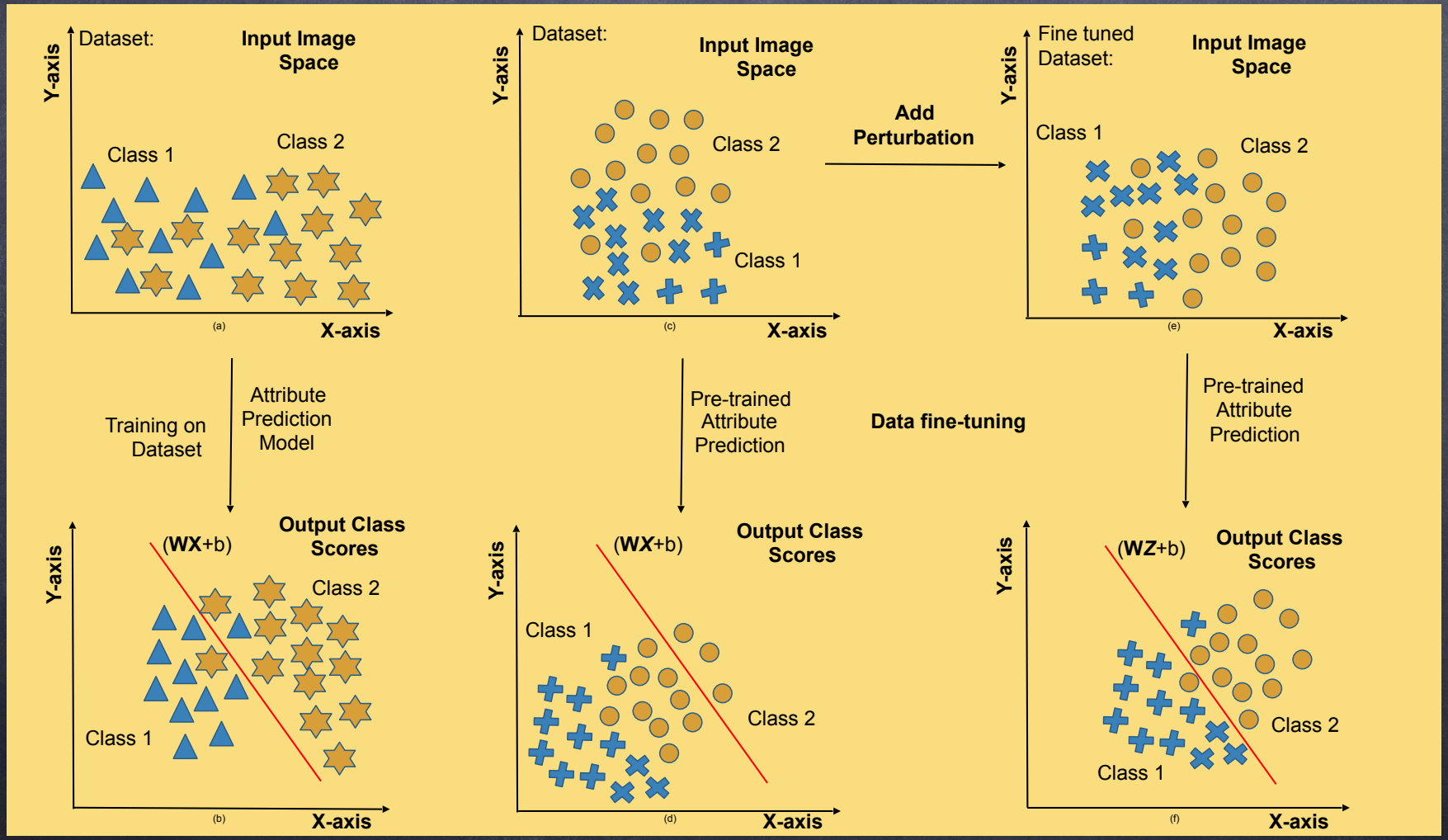
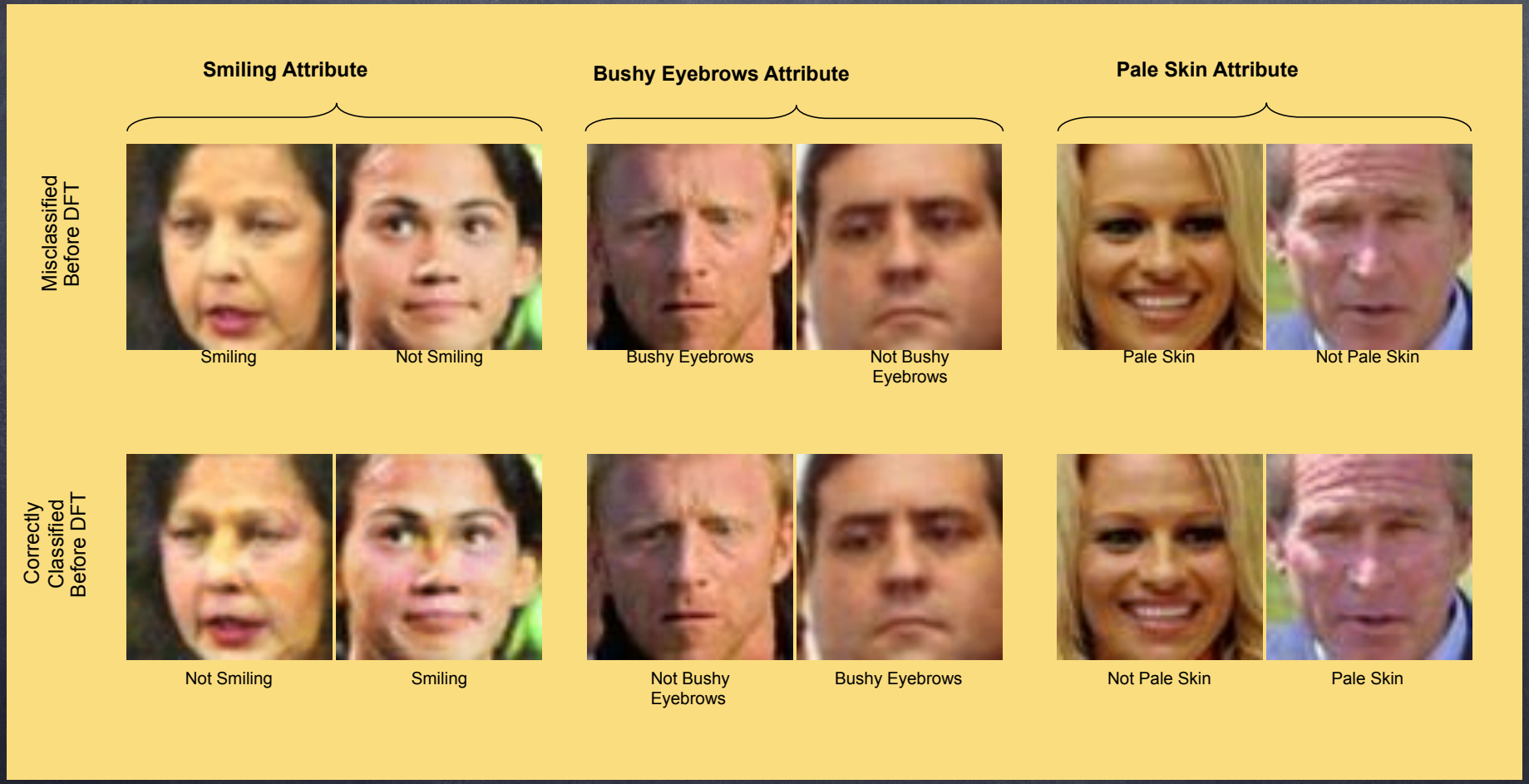


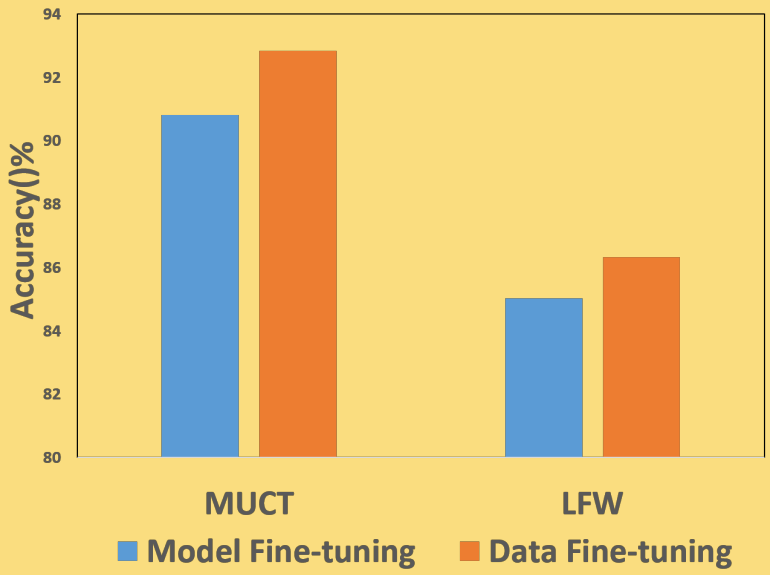
Illustration of Data Fine-tuning for Attribute Prediction



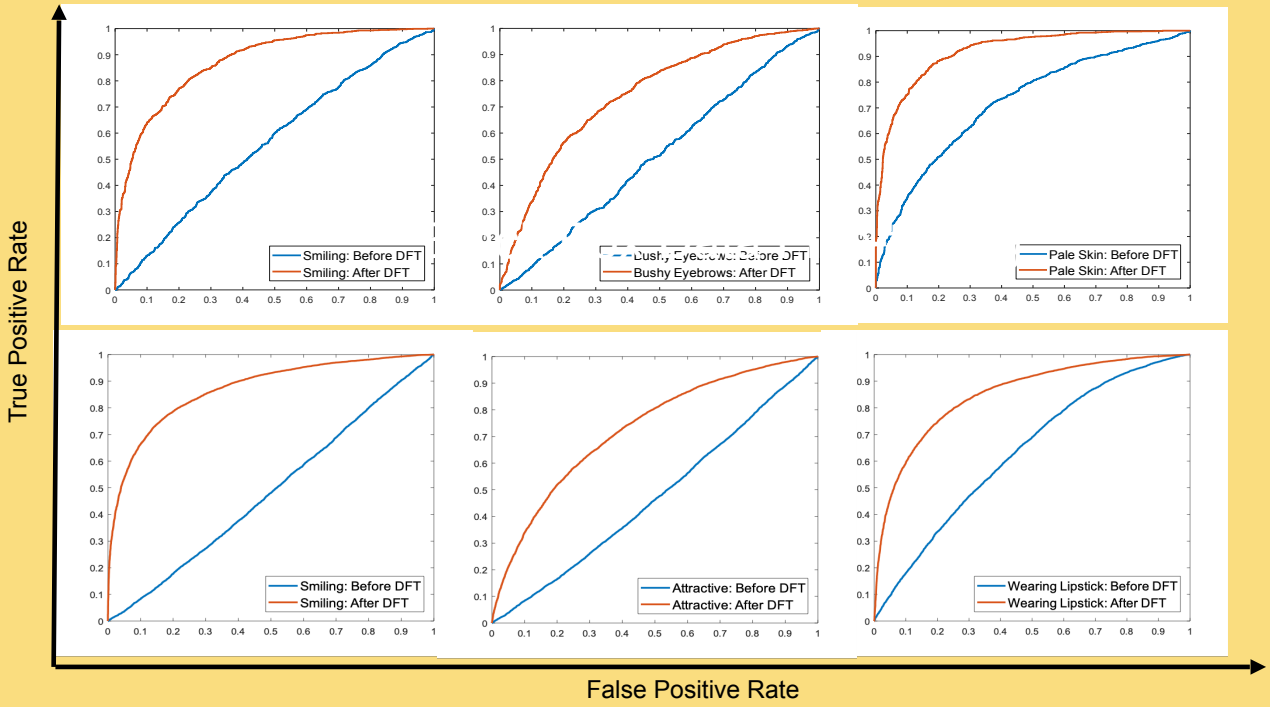
Visual Results



Model Fine-tuning vs Data Fine-tuning



Black Box Data Fine-tuning



Dataset: LFW
Model: CelebA

Dataset: CelebA
Model: LFW

Key Takeout

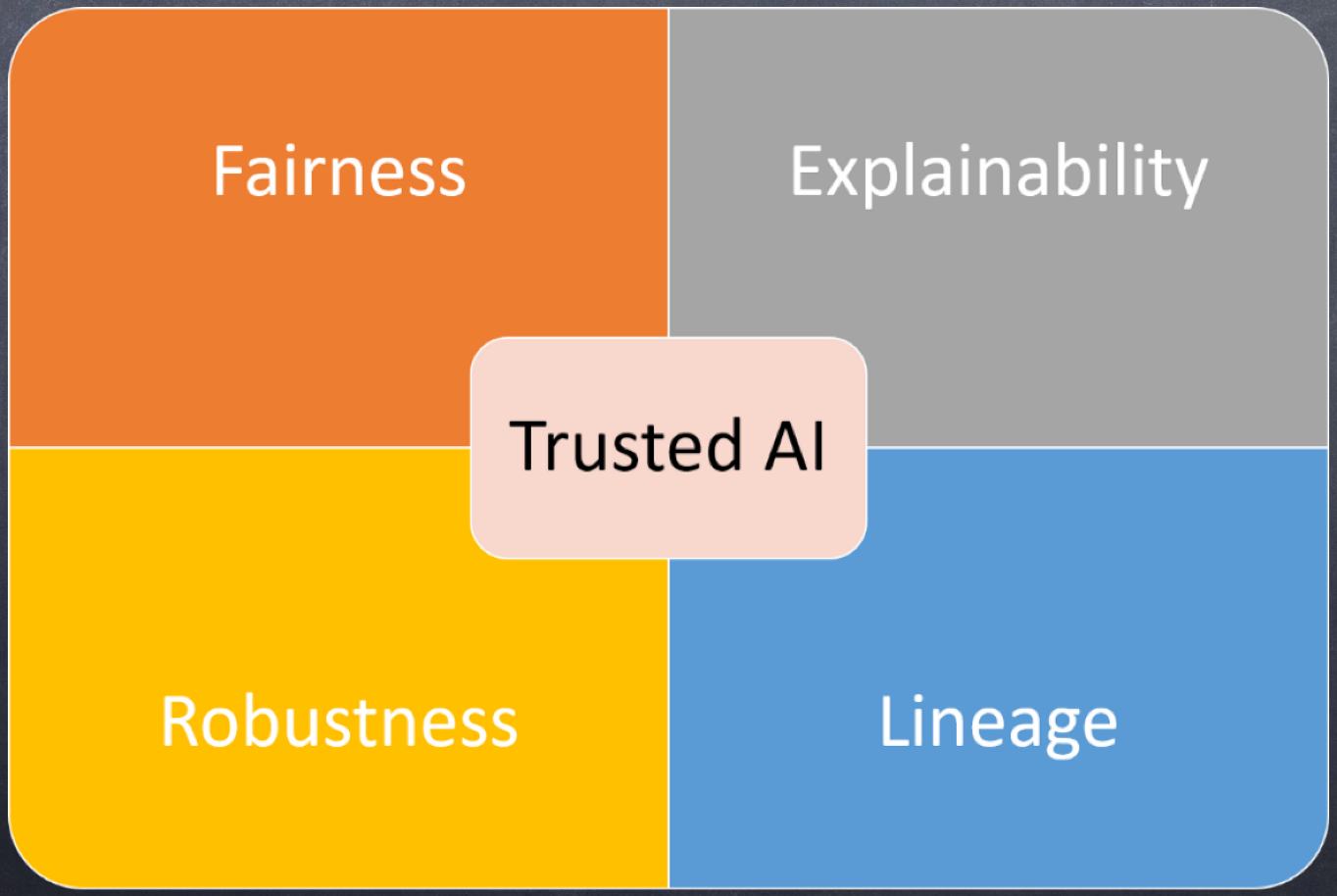
- Data fine-tuning is an attractive alternative to model fine-tuning, specifically, when model is un-known or black-box

Summary

- Defense against adversarial perturbations is important
- Adversarial perturbations can be used for privacy preserving approaches and fine-tuning the models

Trusted AI

- Robustness is an important topic for building Trusted-AI systems but there are three other important topics



https://towardsdatascience.com/towards-ai-transparency-four-pillars-required-to-build-trust-in-artificial-intelligence-systems-d1c45a1bdd59

Acknowledgments

Puspita Majumdar, Gaurav Goswami, Askhay Agarwal, Saheb Chhabra, Akhil Goel, Anirudh Singh, Anubhav Jain

www.iab-rubric.org