

**ARISTOTLE UNIVERSITY OF THESSALONIKI
FACULTY OF SCIENCES
SCHOOL OF INFORMATICS**

POSTGRADUATE STUDIES PROGRAM ON INFORMATICS AND COMMUNICATIONS
SPECIALIZATION ON DIGITAL MEDIA AND COMPUTATIONAL INTELLIGENCE



Adversarial Face De-identification

Master's Thesis

Efstathios Chatzikyriakidis

Supervisor

Ioannis Pitas, Professor AUTH

ARTIFICIAL INTELLIGENCE AND INFORMATION ANALYSIS LABORATORY
Thessaloniki, February 2019

Email : contact@efxa.org

Website : <http://www.efxa.org/>

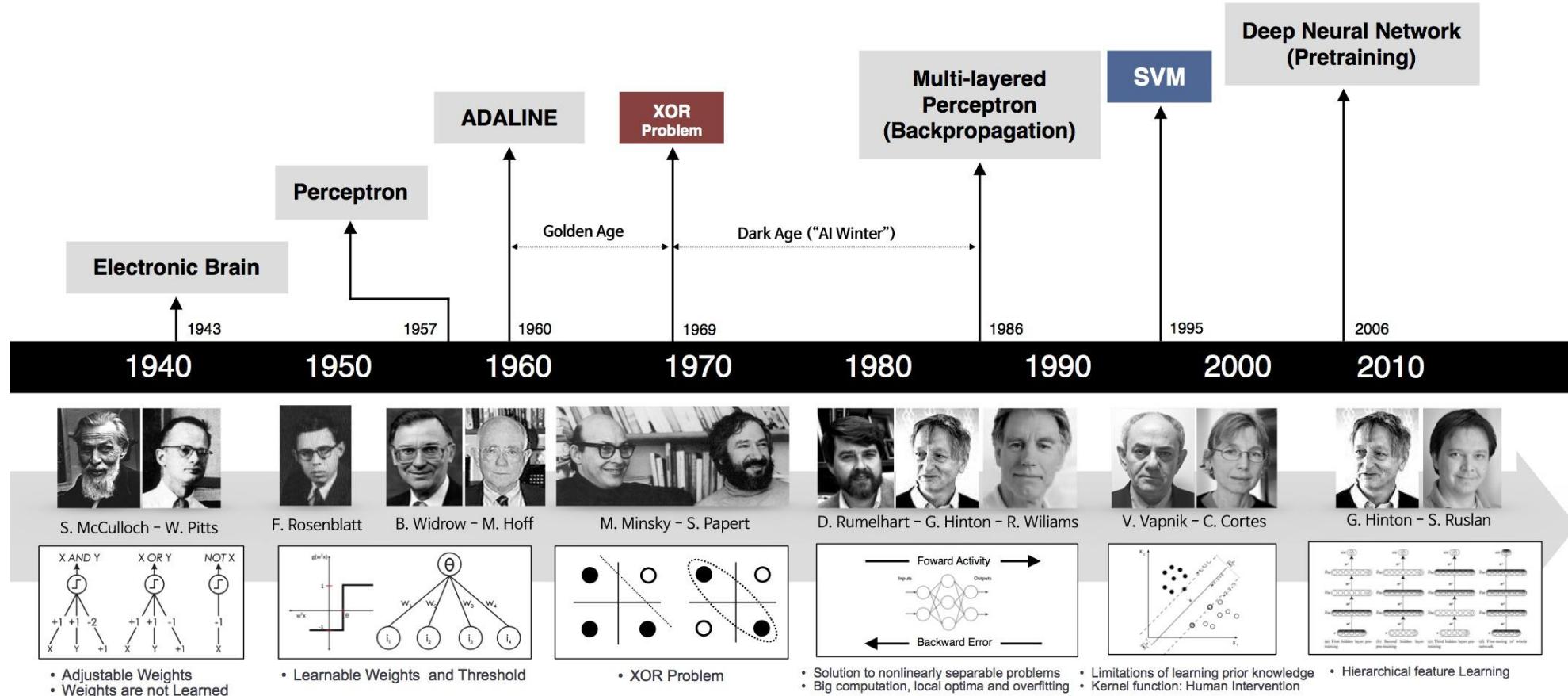


Introduction

Important history of “Artificial Intelligence”



We had a long journey... and we are still at the birth of it...

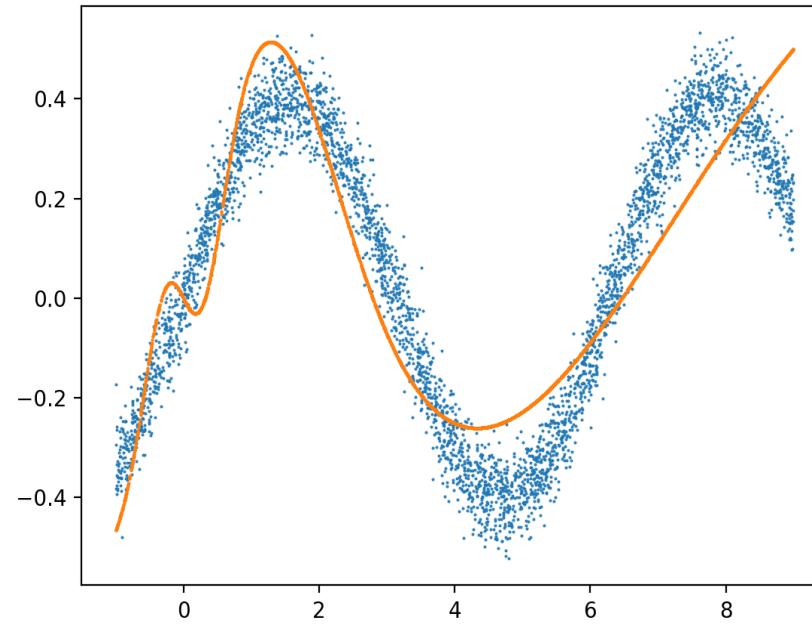
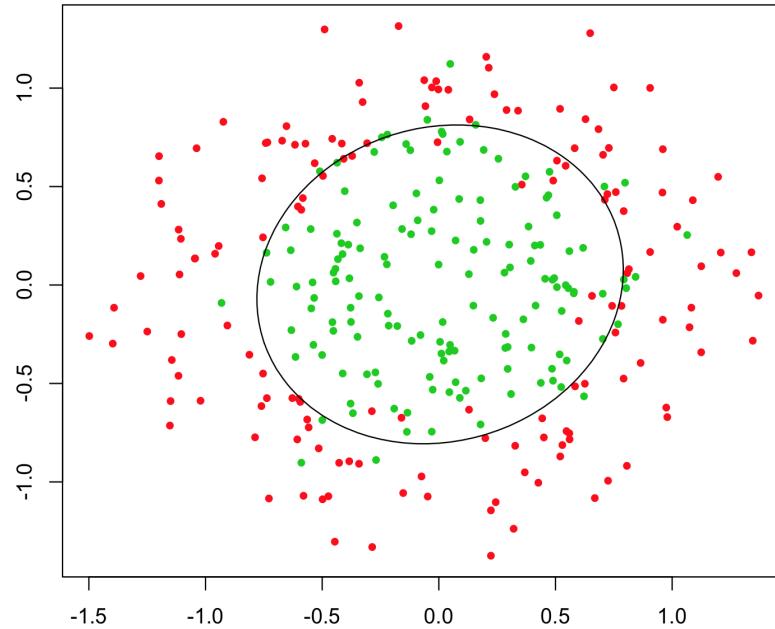


Introduction

Advantages of Artificial Neural Networks



- Satisfactory separation of non-linear separable input data
- Satisfactory function approximation using only input data

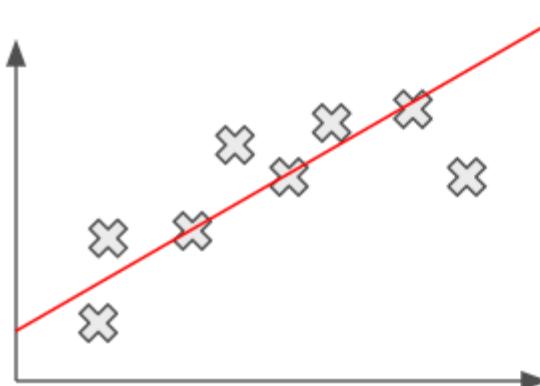


Introduction

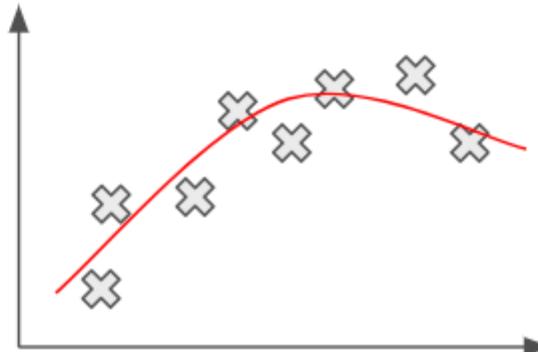
Advantages of Artificial Neural Networks



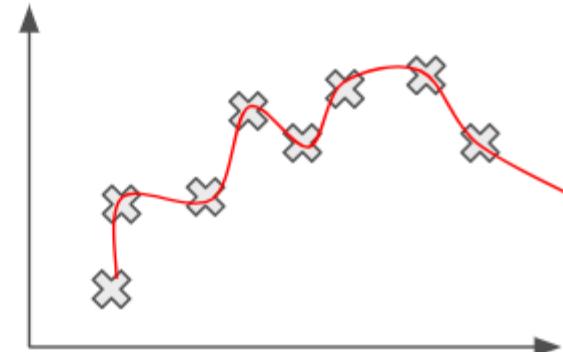
- Good generalization that captures the general manifold of data



Underfitting



Optimal



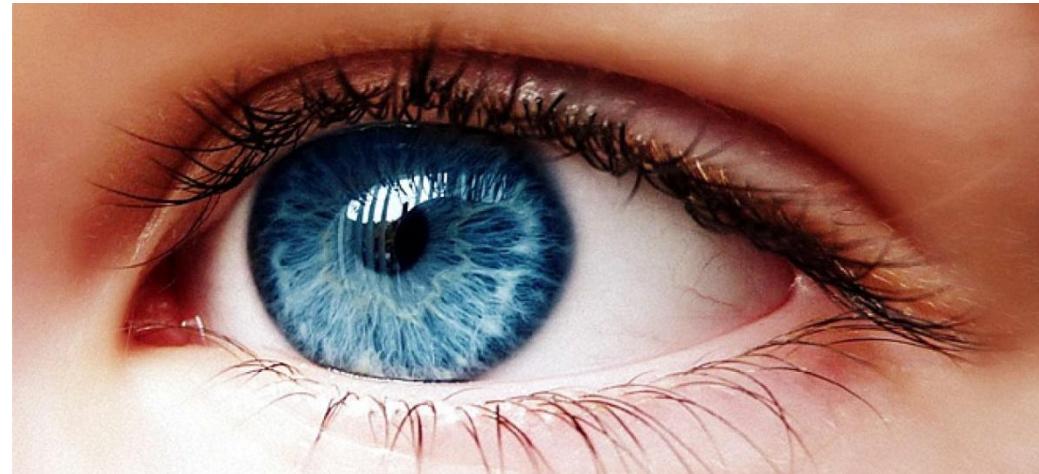
Overfitting

Introduction

Computer Vision and its future goal

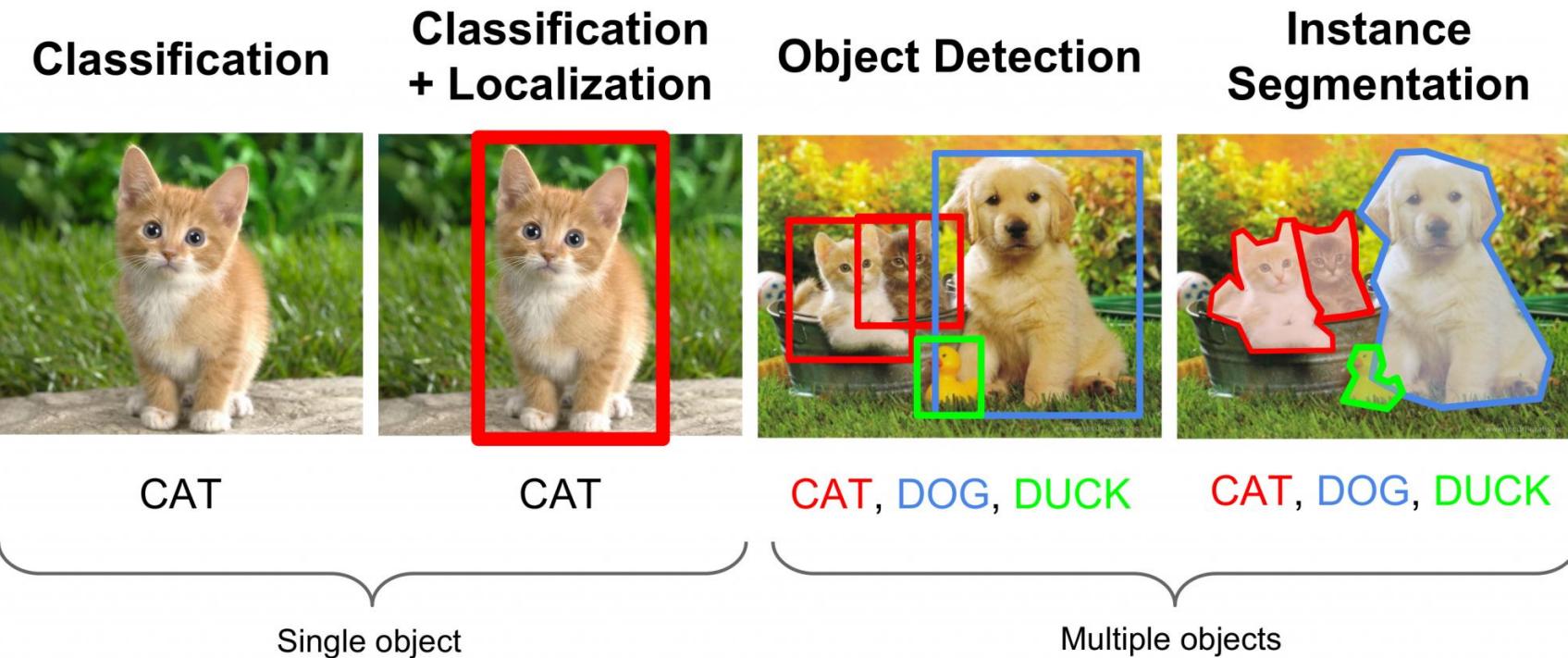


- The strongest sense of most animal species
- Simulate how brains see and understand the world through vision sense



Computer Vision

Well-known tasks of high-level image understanding

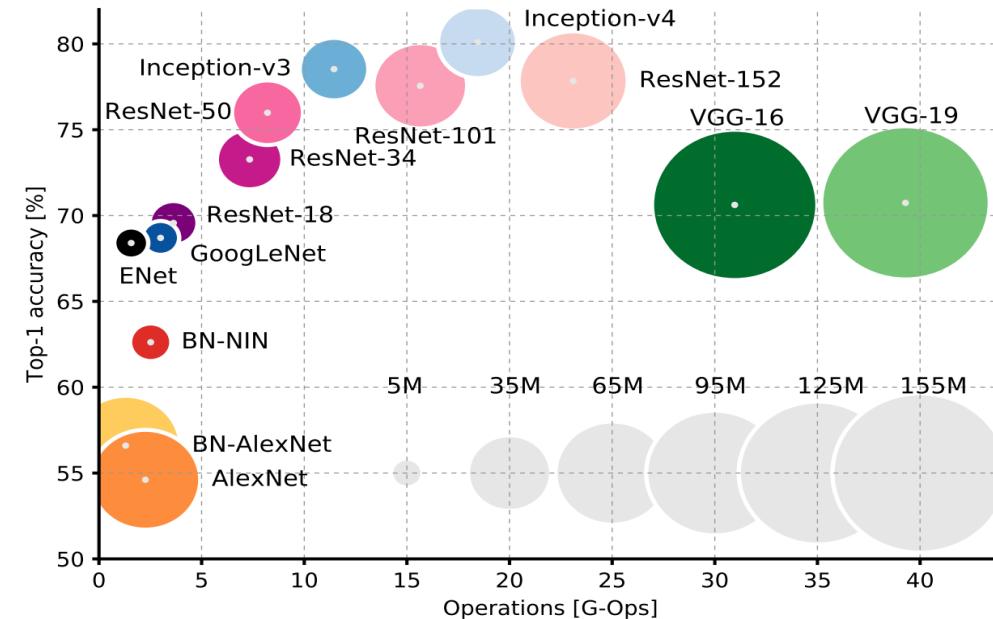
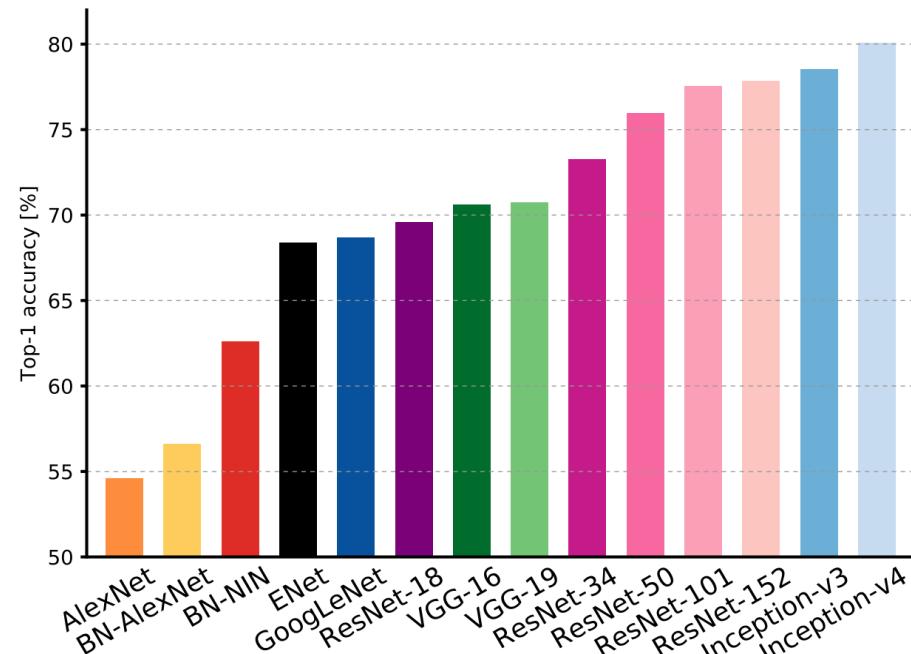


Computer Vision

State-of-the-art Image Classification

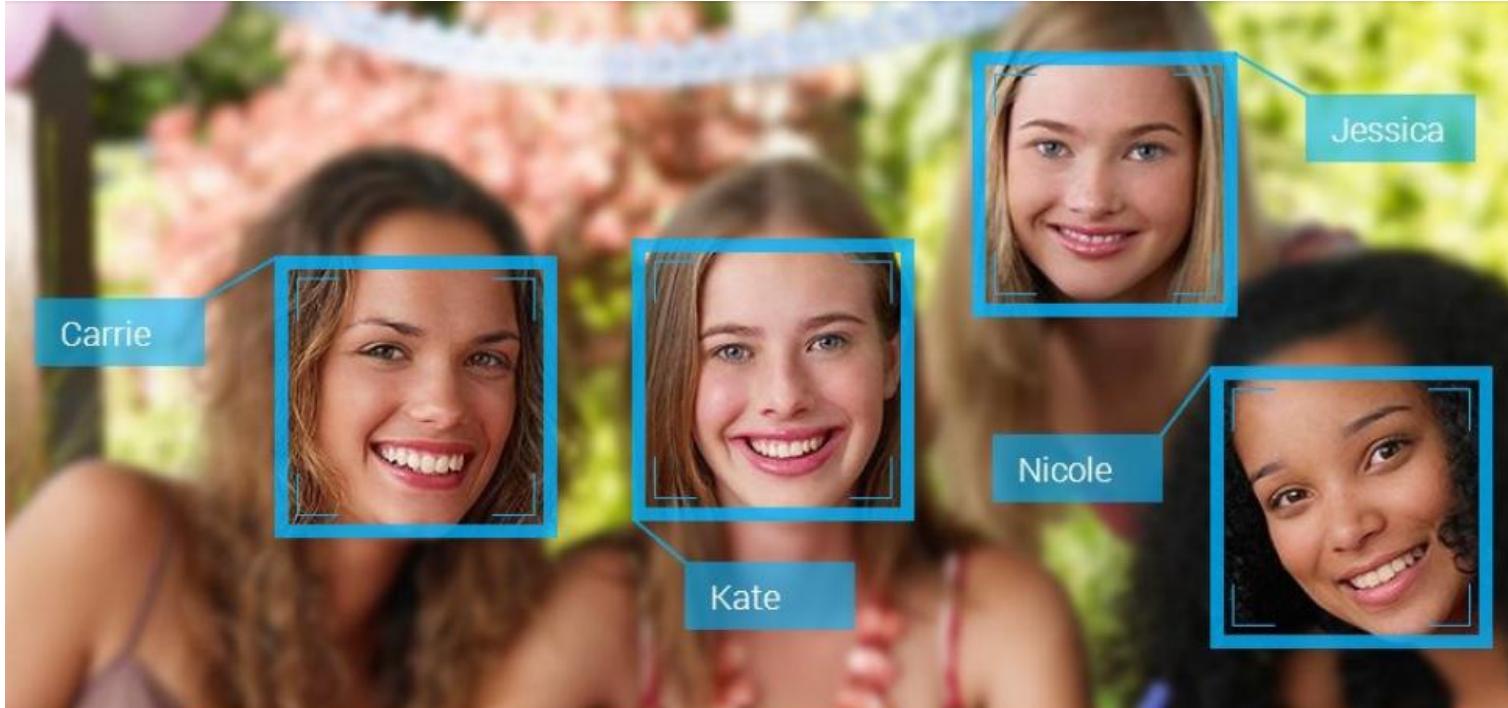


ImageNet Large Scale Visual Recognition Challenge (ILSVRC)



Computer Vision

Face Detection & Face Recognition



Privacy Threat

Social Media, Surveillance Cameras



Privacy Threat

Governmental Protection



- General Data Protection Regulation (GDPR)
- Declaration of the Rights of Man and of the Citizen (DRMC)
- Universal Declaration of Human Rights (UDHR)
- Health Insurance Portability and Accountability Act (HIPAA)
- Health Information Technology for Economic and Clinical Health Act (HITECH Act)
- U.S. Fourth Amendment

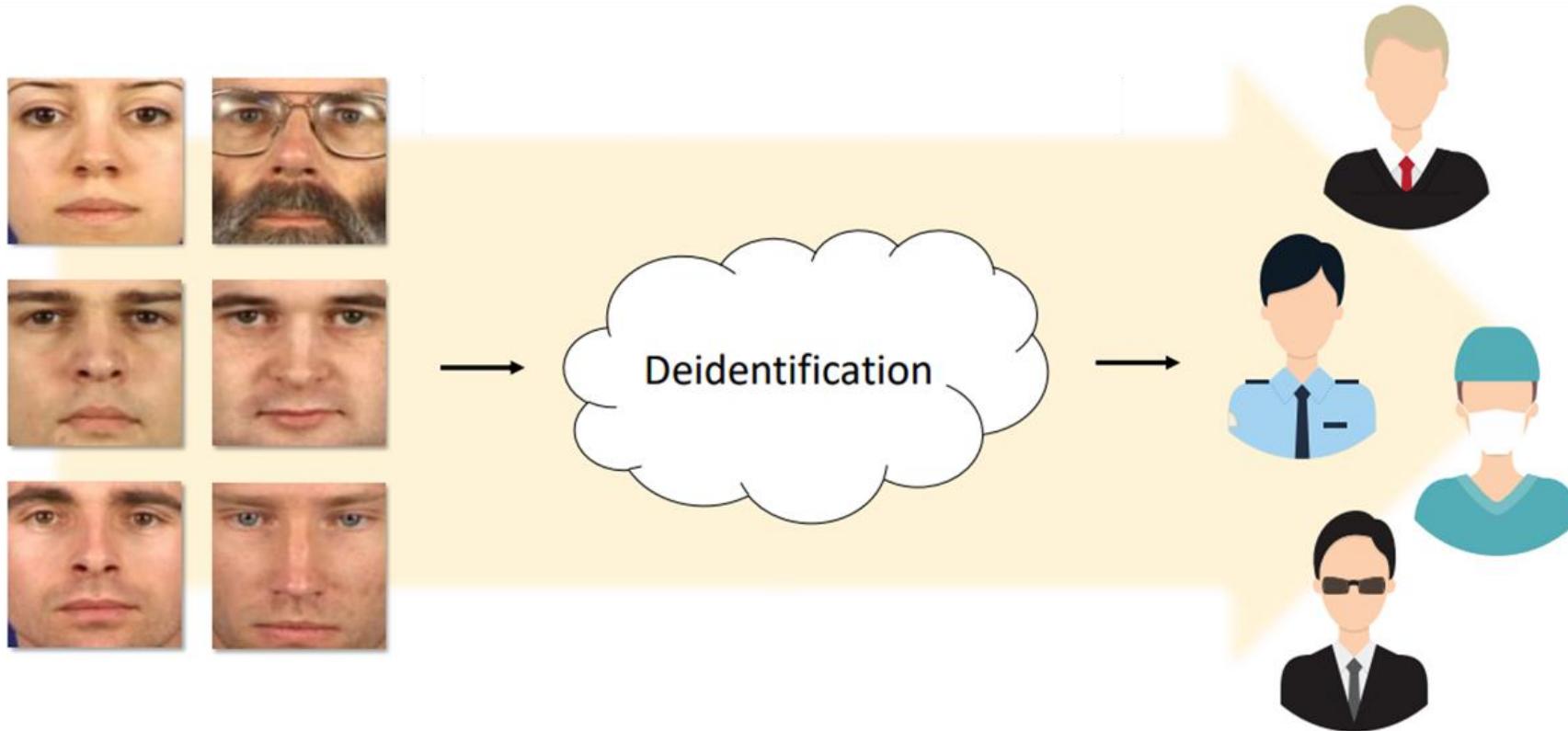


This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731667 (MULTIDRONE)



Face De-identification

Privacy Protection on Facial Images



Face De-identification

Ad-hoc (Naïve) Methods

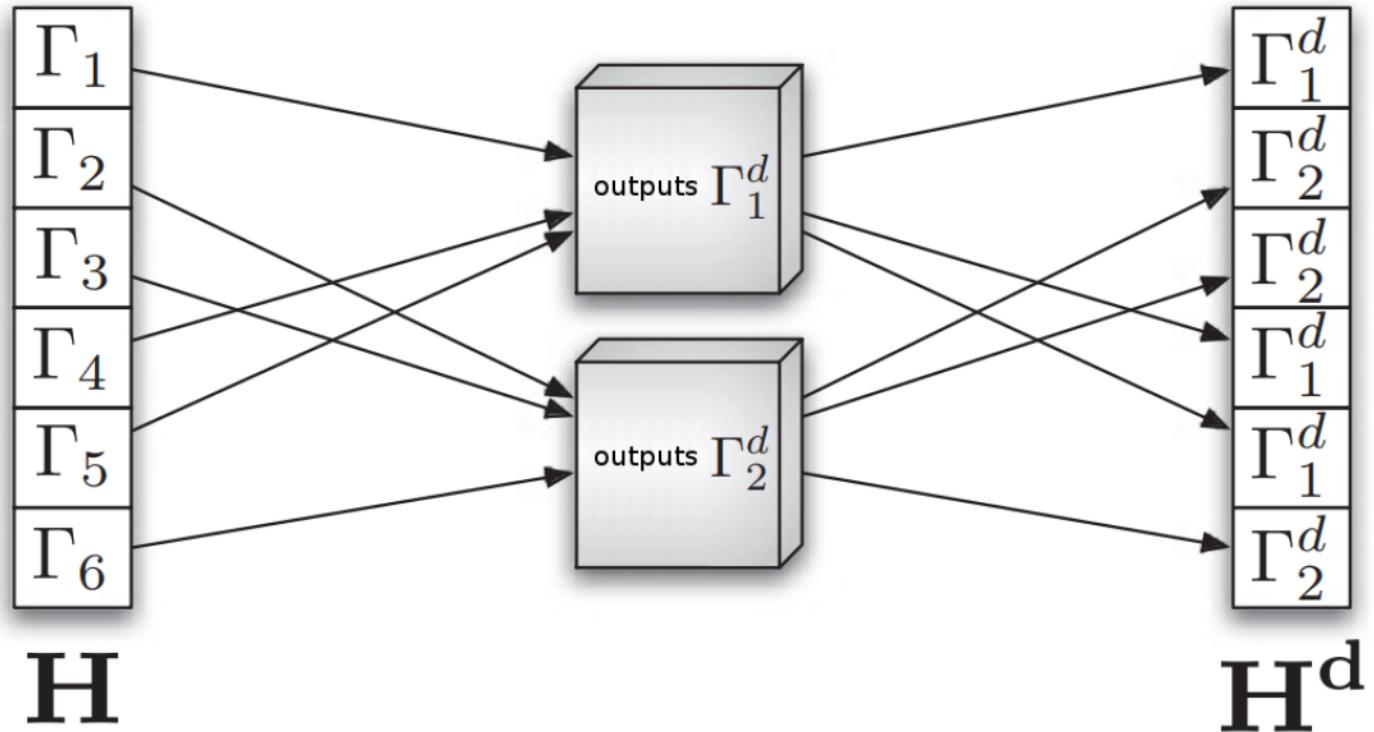


- Mask
- Blur
- Pixelation
- Negative
- Threshold
- Mr. Potato Face
- Reduction
- Random Noise



Face De-identification

k-Anonymity Protection Model

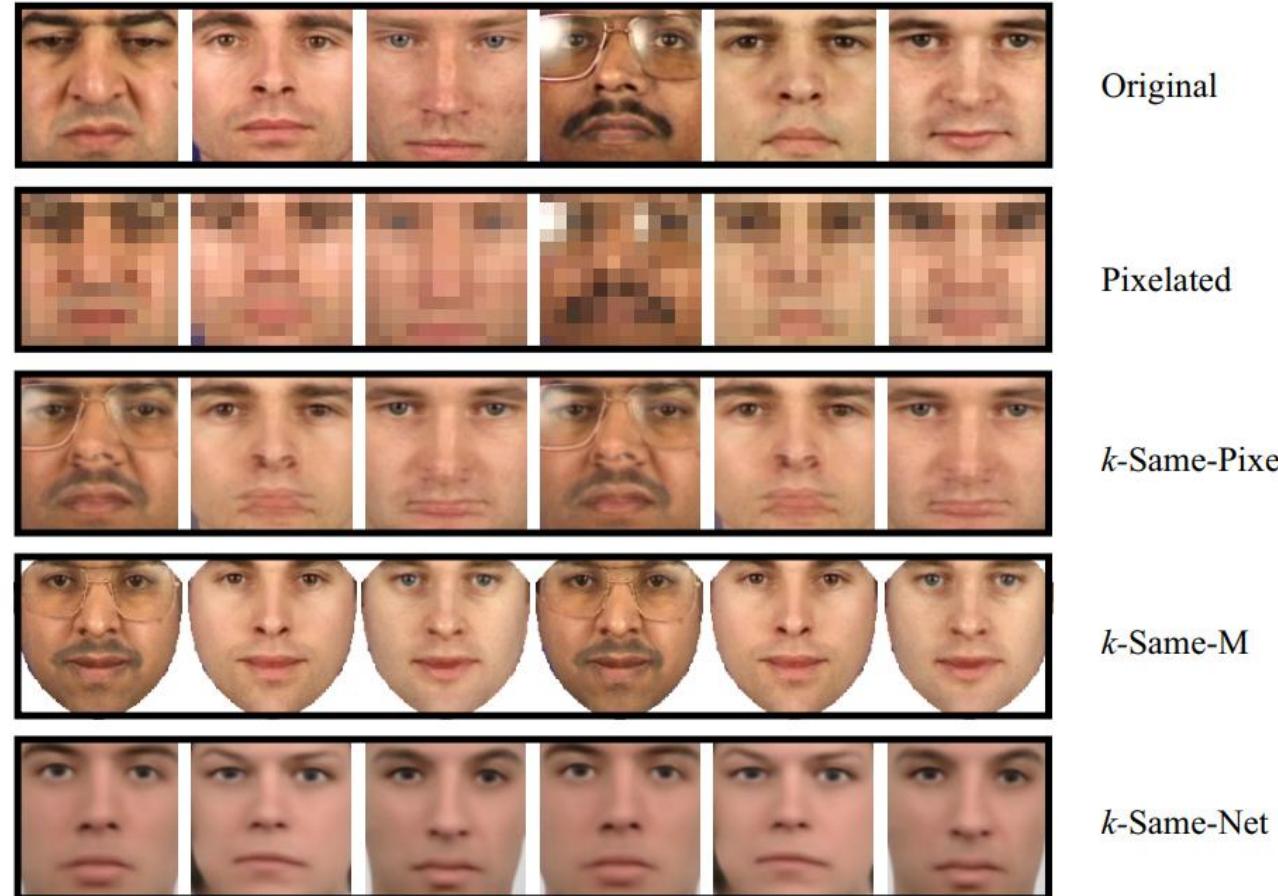


Face De-identification

k-Same Family Methods



- k-Same-Pixel
- k-Same-Eigen
- k-Same-Select
- k-Same-M
- k-Same-Net



Face De-identification

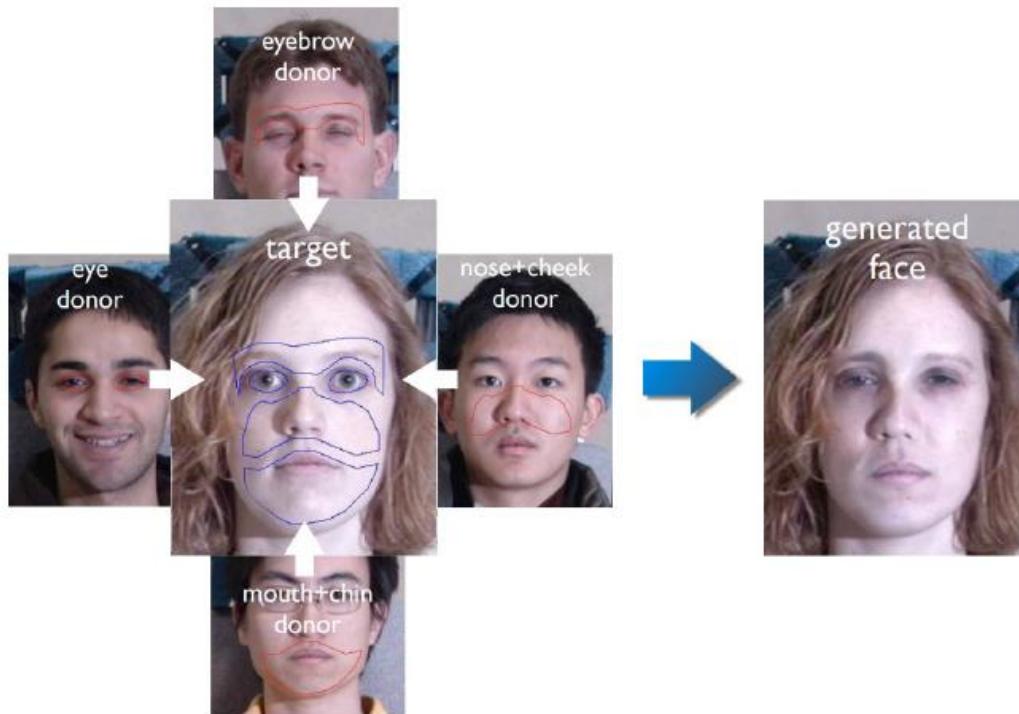
Face Synthesis Methods



Using Generative Models (e.g. GANs)



Using Donors' Face Components



Face De-identification

Methods Drawbacks



The de-identified facial images:

- are not similar to the original ones (e.g. they are either distorted or replaced by a synthetic face)
- do not preserve satisfactorily the non-identity facial characteristics of the original faces
(e.g. pose, gender, race, expression, age)

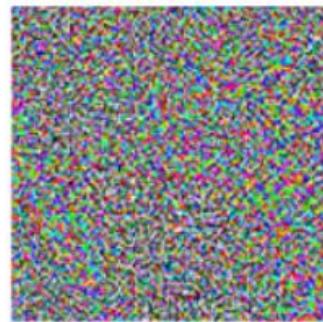
Adversarial Examples

What are these?



"panda"
57.7% confidence

+



=



"gibbon"
99.3% confidence

- They are perturbated high-dimensional image samples for which local generalization does not apply
- Perturbation: optimal direction to move all the pixels so that a classifier will lead to misclassification
- Most classifier models fail to work (LR, Softmax Regression, SVM, k-NN, Decision Trees, ANN)



Adversarial Examples

Formal definition



There is a sample and its ground truth class: x, y

A classification model f predicts the class of x : $\hat{y} = f(x)$

The prediction \hat{y} is the same as the ground truth: $y = \hat{y}$

There is a sample x_p which is x perturbated by p : $x_p = x + p$

The distance of the two samples is restricted by threshold e : $d(x, x_p) \leq e$

The threshold e is positive and small for imperceptibility changes

The classification model classifies the perturbated sample x_p : $\hat{y}_p = f(x_p)$

Non-targeted adversarial example constraint: $\hat{y}_p \neq \hat{y}$

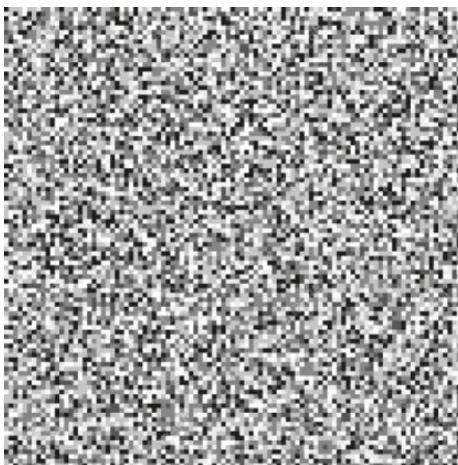
Targeted adversarial example constraint: $\hat{y}_p = y_{target}$

Adversarial Examples

Taxonomy - Input Sample



- **Random:** From uniform, Gaussian (normal) or other probability distribution
- **Non-random:** Any example from the training / validation / test dataset



Adversarial Examples

Taxonomy - Attack Method Frequency



- **One-time:** Take only one time to optimize adversarial examples
- **Iterative:** Multiple times to update the adversarial examples

Taxonomy - Adversarial Specificity

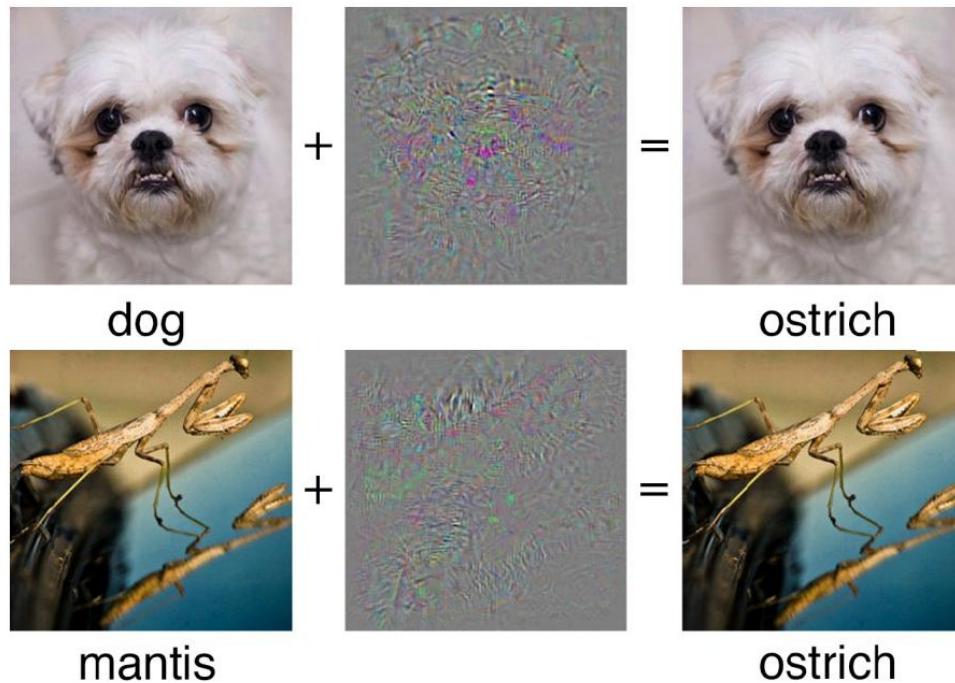
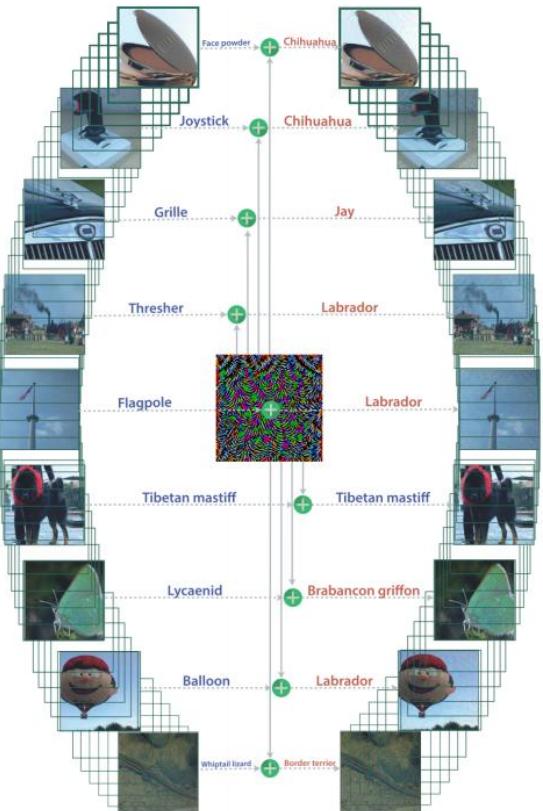
- **Targeted:** Find an input that is misclassified as a specific label
- **Non-targeted:** Find an input that is misclassified in a label different than the ground truth

Adversarial Examples

Taxonomy - Perturbation Scope



- **Individual:** Each image has its own individual perturbation
- **Universal:** A universal perturbation for the whole dataset



Adversarial Examples

Taxonomy - Perturbation Measurement



$$\|x\|_p = (|x_1|^p + |x_2|^p + \cdots + |x_n|^p)^{\frac{1}{p}} = \left(\sum_{i=1}^n |x_i|^p \right)^{\frac{1}{p}}$$

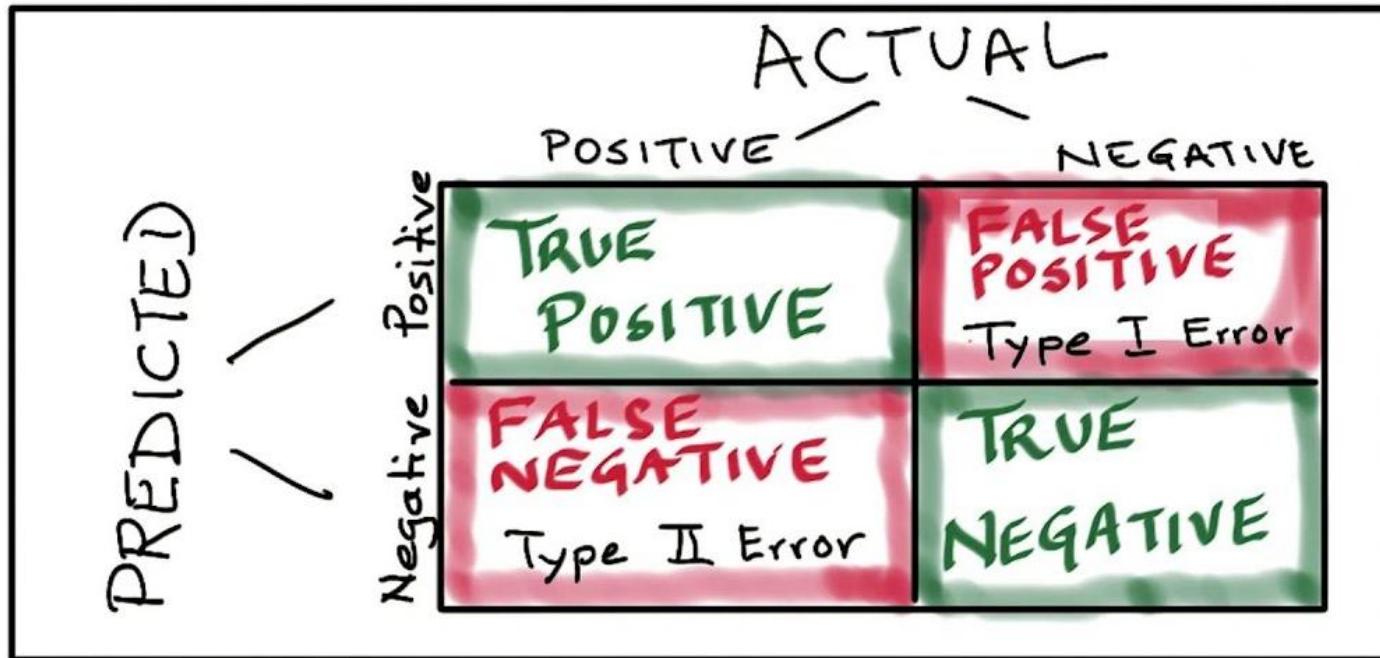
- l_0 -norm: $\|x\|_0 = \#\{i|x_i \neq 0\}$
- l_1 -norm: $\|x\|_1 = |x_1| + |x_2| + \cdots + |x_n|$
- l_2 -norm: $\|x\|_2 = \sqrt{|x_1|^2 + |x_2|^2 + \cdots + |x_n|^2}$
- l_∞ -norm: $\|x\|_\infty = \max(|x_1|, |x_2|, \dots, |x_n|)$

Adversarial Examples

Taxonomy - Adversarial Falsification

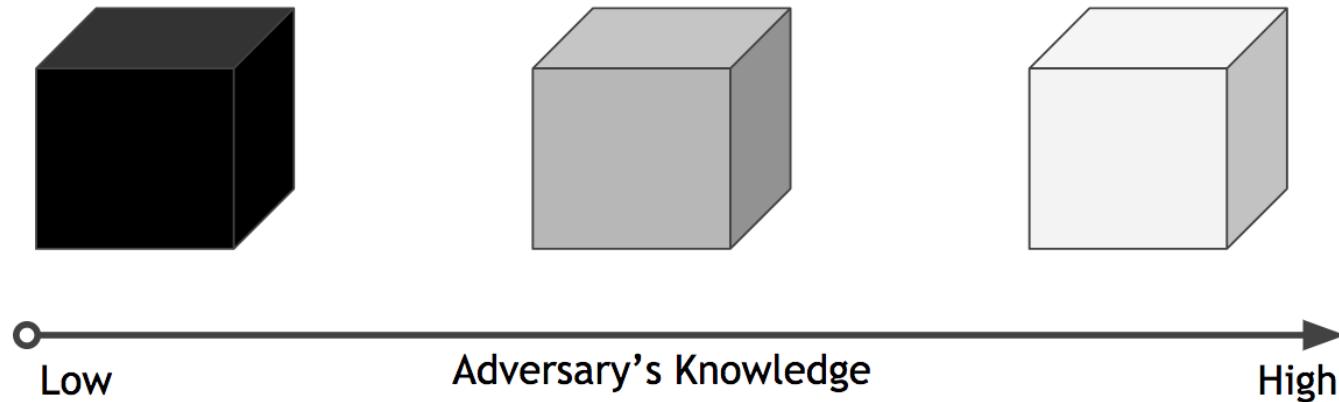


- **Type 1 error:** Negative sample classified as positive
- **Type 2 error:** Positive sample classified as negative



Adversarial Examples

Taxonomy - Adversary Knowledge



- **Black-box:** Zero knowledge about the model to attack (knowing only the final classification)
- **Grey-box:** Limited knowledge about the model to attack (something between Black-box and White-box)
- **White-box:** Full knowledge about the model to attack (architecture, parameters, dataset, etc)

Adversarial Examples

Transferability



- Cross model generalization
- Cross training-set generalization
- Can be transferred to the physical world

Adversarial Examples

Attack Methods

- Fast Gradient Sign Method (FGSM)
- One-step Target Class Method (OTCM)
- One-step Least Likely Class Method (OLLCM)
- Fast Gradient Value Method (FGVM)
- Adversarial GAN Attack (AdvGAN)
- Box-constrained Limited-memory BFGS Attack (L-BFGS-B)
- Basic Iterative Method (BIM)
- Iterative Target Class Method (ITCM)
- Iterative Least Likely Class Method (ILLCM)
- Iterative Fast Gradient Value Method (IFGVM)
- Momentum Iterative Fast Gradient Sign Method (MIFGSM)
- DeepFool Attack
- Jacobian-Based Saliency Map Attack (JSMA)
- Universal Perturbation Attack
- Carlini-Wagner Attack (C&W)
- One Pixel Attack



Adversarial Examples

Fast Gradient Methods



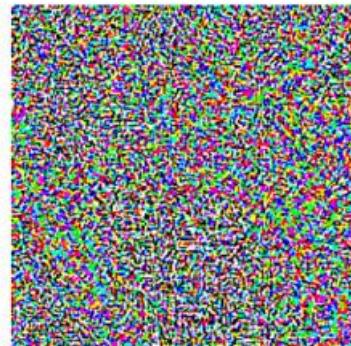
Let's get an idea with “Fast Gradient” Methods

One-step Target Class Method (OTCM) :



x
“panda”
57.7% confidence

$+ .007 \times$



$\text{sign}(\nabla_x J(\theta, x, y))$
“nematode”
8.2% confidence

=

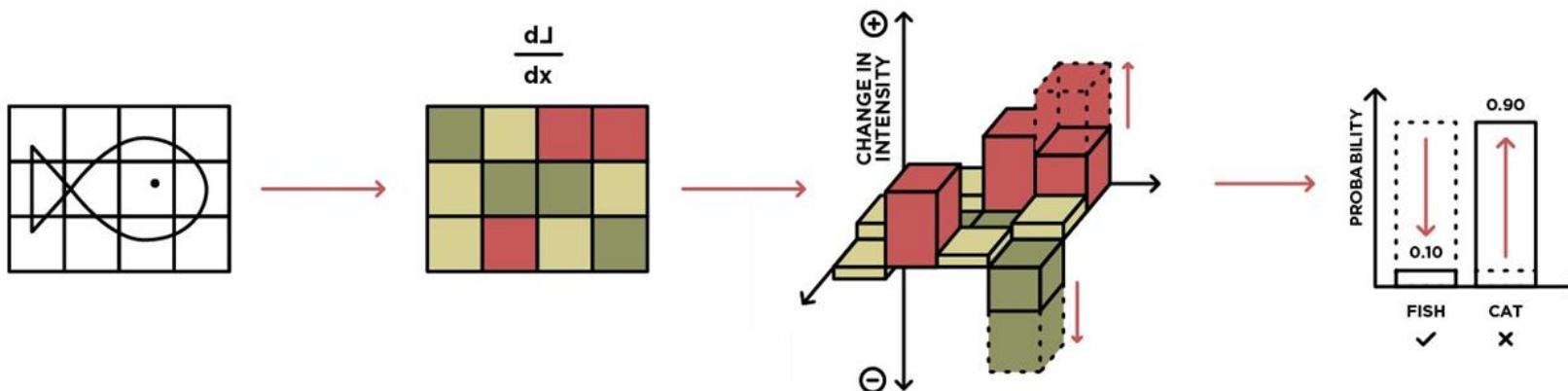


$x + \epsilon \text{sign}(\nabla_x J(\theta, x, y))$
“gibbon”
99.3 % confidence

Adversarial Examples

Fast Gradient Methods

- Use gradients of loss function w.r.t. input
- Gradient descent for targeted or ascent for non-targeted
- Very effective for the domain of image
- Fast and easy to compute
- ‘ ϵ ’ controls the size of the change (should be a small value)
- Can be used for run-time adversarial training
- For NNs the $\nabla_x J(\theta, x, y)$ can be calculated with backpropagation



Adversarial Examples

Iterative Target Class Method (ITCM)



Targeted Attack:

$$\hat{x}_0 = x$$

$$\hat{x}_{i+1} = clip_{[x-\varepsilon, x+\varepsilon]}(\hat{x}_i - \alpha \cdot sign(\nabla_x J_f(\hat{x}_i, \hat{y})))$$

x input image

\hat{x}_i adversarial image at step i

\hat{y} target class

f classifier

J_f loss function of classifier f

α step size

ε tunable small parameter

Non-targeted Attack:

$$\hat{x}_0 = x$$

$$\hat{x}_{i+1} = clip_{[x-\varepsilon, x+\varepsilon]}(\hat{x}_i + \alpha \cdot sign(\nabla_x J_f(\hat{x}_i, y)))$$

x input image

\hat{x}_i adversarial image at step i

y true class

f classifier

J_f loss function of classifier f

α step size

ε tunable small parameter

Adversarial Examples

Iterative Fast Gradient Value Method (IFGVM)



Targeted Attack:

$$\hat{x}_0 = x$$

$$\hat{x}_{i+1} = \text{clip}_{[x-\varepsilon, x+\varepsilon]}(\hat{x}_i - \alpha \cdot \nabla_x J_f(\hat{x}_i, \hat{y}))$$

x input image

\hat{x}_i adversarial image at step i

\hat{y} target class

f classifier

J_f loss function of classifier f

α step size

ε tunable small parameter

Non-targeted Attack:

$$\hat{x}_0 = x$$

$$\hat{x}_{i+1} = \text{clip}_{[x-\varepsilon, x+\varepsilon]}(\hat{x}_i + \alpha \cdot \nabla_x J_f(\hat{x}_i, y))$$

x input image

\hat{x}_i adversarial image at step i

y true class

f classifier

J_f loss function of classifier f

α step size

ε tunable small parameter

Proposed Method

Penalized Fast Gradient Value Method (P-FGVM)



- Generates de-identified facial images as targeted adversarial examples
- Iterative adversarial attack method
- Operates on image domain
- Minimal image distortion with high attack rate (misclassification rate)
- Based on Iterative Fast Gradient Value Method (IFGVM)
- Solves a multi-objective optimization problem with gradient descent

Proposed Method

Multi-Objective Function & Gradient Descent Update Rules



$$J_f(\hat{x}, \hat{y}) + \lambda \cdot \frac{1}{2} \cdot \|\hat{x} - x\|_2^2 \quad \text{such that} \quad \hat{x} \in [0, 1]^n$$

x input image

\hat{x} adversarial image

\hat{y} target class

f classifier

J_f loss function of classifier f

λ penalty term weight coefficient

$\hat{x}_0 = x$

$\hat{x}_{i+1} = clip_{[0,1]}(\hat{x}_i - \alpha \cdot (\nabla_x J_f(\hat{x}_i, \hat{y}) + \lambda \cdot (\hat{x}_i - x)))$

x input image

\hat{x}_i adversarial image at step i

\hat{y} target class

f classifier

J_f loss function of classifier f

λ penalty term weight coefficient

α step size

Proposed Method

Contributions



- First face de-identification method that uses adversarial examples
- Better misclassification rate (protection) than other face de-identification methods
- Solves the problem of minimal image distortion (de-identified facial images same to original ones)
- Solves the problem of preserving the non-identity facial characteristics in the de-identified images
- The novel penalty term leads to better results than baseline attack methods ITCM, IFGVM

Target Face Classifiers

Model A



- Convolutional Neural Network
- Trained on NVIDIA GeForce GTX 1080 GPU
- Train with a subset of CelebA dataset for face recognition
- Hyper-parameters (training, architecture) search:

Parameter	Optimal Value	Values
Learning Rate	1e-4	1e-1, 1e-2, 1e-3, 1e-4, 1e-5
Batch Size	16	8, 16, 32, 64, 128
Convolution Kernel Size	5x5	3x3, 5x5
Extra Convolution Blocks	1	0, 1, 2
First Convolution Block Filters	32	8, 16, 32
Penultimate Level Neurons	512	16, 32, 64, 128, 256, 512
Dropout Rate	0.9	0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9
L2 Regularization Factor	1e-3	1e-1, 1e-2, 1e-3, 1e-4, 1e-5
Batch Normalization Usage	Yes	Yes, No

Target Face Classifiers

Model A



Layer	Output Shape	Parameters
input_1	(None, 218, 178, 3)	0
conv2d_1	(None, 218, 178, 32)	2432
batch_normalization_1	(None, 218, 178, 32)	128
activation_1	(None, 218, 178, 32)	0
max_pooling2d_1	(None, 109, 89, 32)	0
conv2d_2	(None, 109, 89, 64)	51264
batch_normalization_2	(None, 109, 89, 64)	256
activation_2	(None, 109, 89, 64)	0
max_pooling2d_2	(None, 54, 44, 64)	0
flatten_1	(None, 152064)	0
dense_1	(None, 512)	77857280
batch_normalization_3	(None, 512)	2048
activation_3	(None, 512)	0
dropout_1	(None, 512)	0
dense_2	(None, 30)	15390
Total params: 77,928,798		
Trainable params: 77,927,582		
Non-trainable params: 1,216		

Conv(32, Kernel(5, 5), Padding(Same), L2Regularizer(1e-3))
BatchNormalization+Relu
MaxPooling(PoolSize(2, 2), Strides(2, 2))
Conv(64, Kernel(5, 5), Padding(Same), L2Regularizer(1e-3))
BatchNormalization+Relu
MaxPooling(PoolSize(2, 2), Strides(2, 2))
FC(512, L2Regularizer(1e-3))
BatchNormalization+Relu
Dropout(0.9)
FC(30)+Softmax

Target Face Classifiers

Model B

- Convolutional Neural Network
- Trained on NVIDIA GeForce GTX 1080 GPU
- Transfer learning of VGG-Face CNN descriptor (trained with VGG Face dataset)
- Reuse pre-trained VGG-16 CNN facial feature extractors
- Fine-tuned with a subset of CelebA dataset for face recognition

VGG-Face CNN descriptor (VGG-16)
FC(256, L2Regularizer(1e-3))
BatchNormalization+Relu
FC(30)+Softmax



Target Face Classifiers

Model B



Layer	Output Shape	Parameters
input_1	(None, 218, 178, 3)	0
conv1_1	(None, 218, 178, 64)	1792
conv1_2	(None, 218, 178, 64)	36928
pool1	(None, 109, 89, 64)	0
conv2_1	(None, 109, 89, 128)	73856
conv2_2	(None, 109, 89, 128)	147584
pool2	(None, 54, 44, 128)	0
conv3_1	(None, 54, 44, 256)	295168
conv3_2	(None, 54, 44, 256)	590080
conv3_3	(None, 54, 44, 256)	590080
pool3	(None, 27, 22, 256)	0
conv4_1	(None, 27, 22, 512)	1180160
conv4_2	(None, 27, 22, 512)	2359808
conv4_3	(None, 27, 22, 512)	2359808
pool4	(None, 13, 11, 512)	0
conv5_1	(None, 13, 11, 512)	2359808
conv5_2	(None, 13, 11, 512)	2359808
conv5_3	(None, 13, 11, 512)	2359808
pool5	(None, 6, 5, 512)	0
flatten_1	(None, 15360)	0
dense_1	(None, 256)	3932416
batch_normalization_1	(None, 256)	1024
activation_1	(None, 256)	0
dense_2	(None, 30)	7710
Total params: 18,655,838		
Trainable params: 3,940,638		
Non-trainable params: 14,715,200		

Target Face Classifiers

Training Information



	Model A	Model B
Dataset	CelebA	CelebA
Total Classes	30	30
Total Images	900	900
Images Resolution	178x218	178x218
Training Ratio	70%	70%
Testing Ratio	15%	15%
Validation Ratio	15%	15%
Stratified Sampling	Yes	Yes
Images Normalization	MinMax	MinMax
Learning Rate	1e-4	1e-4
Training Algorithm	Backpropagation	Backpropagation
Optimization Method	Adam	Adam
Loss Function	Cross-entropy	Cross-entropy
Batch Size	16	16
Training Epochs	147	144
Testing Accuracy	80.7%	95.4%
Training Accuracy	100%	100%
Validation Accuracy	80%	97.8%

Experimental Comparison

Evaluation Metrics Results



Comparison of proposed method (P-FGVM) with:

- Iterative Fast Gradient Value Method (IFGVM)
- Iterative Target Class Method (ITCM)

Model A			Model B		
ℓ_2 -norm	CW-SSIM	Misclassification Rate	ℓ_2 -norm	CW-SSIM	Misclassification Rate
P-FGVM					
3.39	0.438	99.6%	2.11	0.456	95.9%
IFGVM					
5.32	0.421	99.4%	2.71	0.441	93.2%
ITCM					
5.68	0.424	98.9%	5.75	0.423	94.4%

Experimental Comparison

Metrics Percentage Improvement of P-FGVM



Model A			Model B		
l_2 -norm	CW-SSIM	Misclassification Rate	l_2 -norm	CW-SSIM	Misclassification Rate
Compared to IFGVM					
36.3%	4.0%	0.2%	22.1%	3.4%	2.9%
Compared to ITCM					
40.3%	3.3%	0.7%	63.3%	7.8%	1.6%

Examples of De-identified Facial Images

Iterative Fast Gradient Value Method (IFGVM)



Model A



Model B

Examples of De-identified Facial Images

Iterative Target Class Method (ITCM)



Model A



Model B

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731667 (MULTIDRONE)

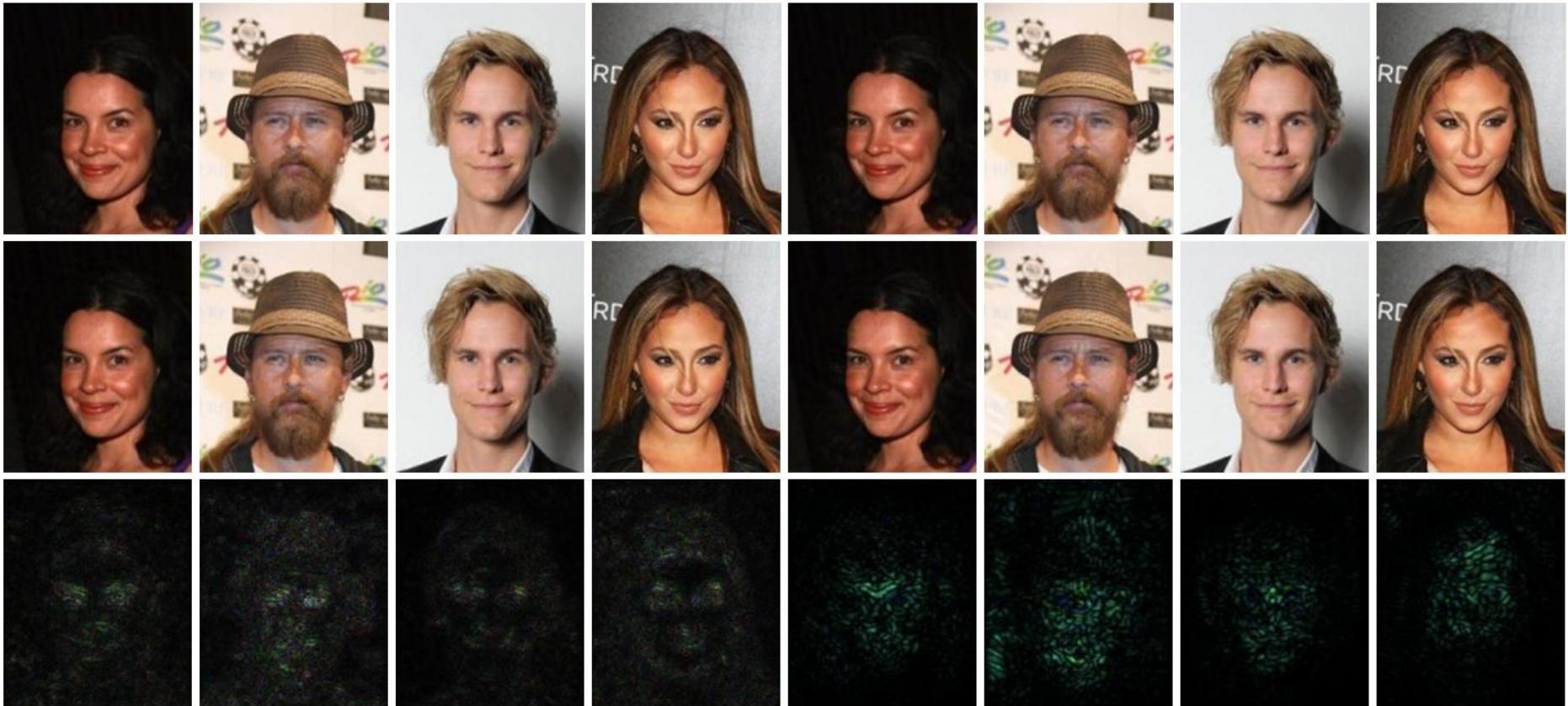


Examples of De-identified Facial Images

Penalized Fast Gradient Value Method (P-FGVM)



Model A

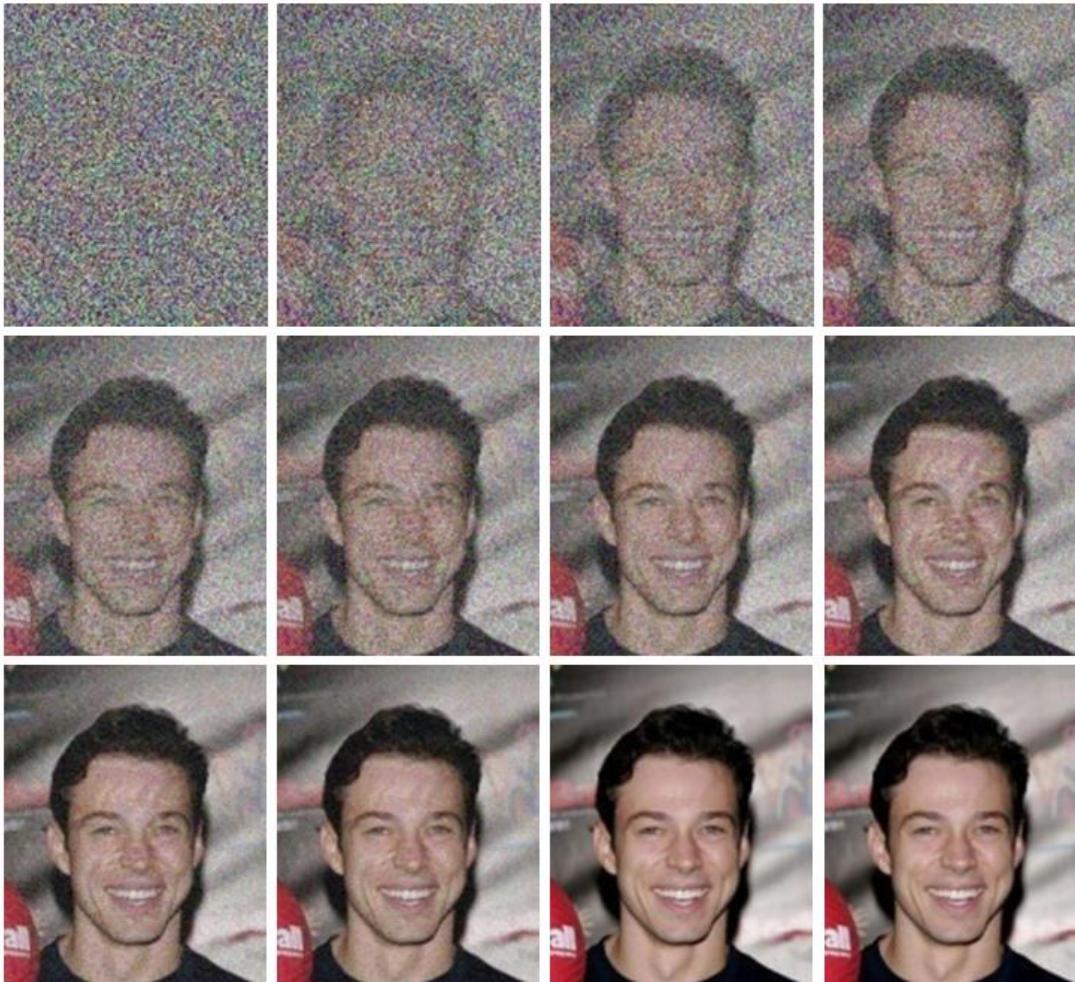


This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731667 (MULTIDRONE)



Examples of De-identified Facial Images

Using Gaussian random input



Software Requirements



- Programming Languages : Python, C#
- IDEs : JetBrains PyCharm, Microsoft Visual Studio
- GPU Card : NVIDIA GeForce GTX 1080
- Python Modules :

Module	Version
scikit-learn	0.20.1
scikit-image	0.14.0
numpy	1.15.4
scipy	1.1.0
matplotlib	3.0.2
keras-gpu	2.2.4
tensorflow-gpu	1.12.0
cudatoolkit	9.0
cudnn	7.1.2

Thank you a lot and have a nice day!

