Adversarial Examples Generative Adversarial Networks

Presenter: Efstathios Chatzikyriakidis

M.Sc. in Informatics and Communications
Specialization in Computational Intelligence and Digital Media
School of Informatics, Faculty of Sciences
Aristotle University of Thessaloniki, Hellas

B.Sc. in Informatics and Communications
Specialization in Software Engineering
Department of Informatics and Communications, Faculty of Applied Technology
Technological Educational Institute of Central Macedonia, Hellas

Email : contact@efxa.org **Website :** http://www.efxa.org/

Contributor: Christos Papaioannidis

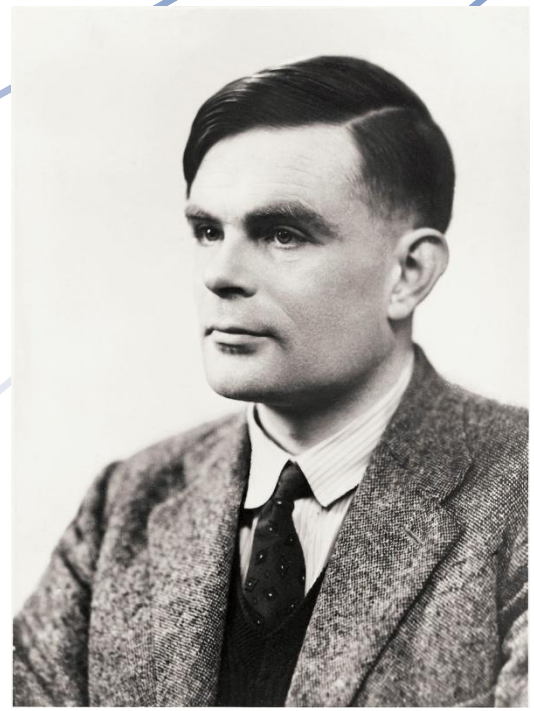
This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731667 (MULTIDRONE)

Introduction

Thoughts on Machine Intelligence

Alan Turing, “Computing Machinery and Intelligence”, 1950

- “Can machines think?” (this was a really difficult question)
- “Can machines do what we can do?” (this question was more promising)
- “Can a machine act indistinguishably from the way a thinker acts?” (Turing Test)
- So, when the machines become intelligent we will not be able to prove that they are not

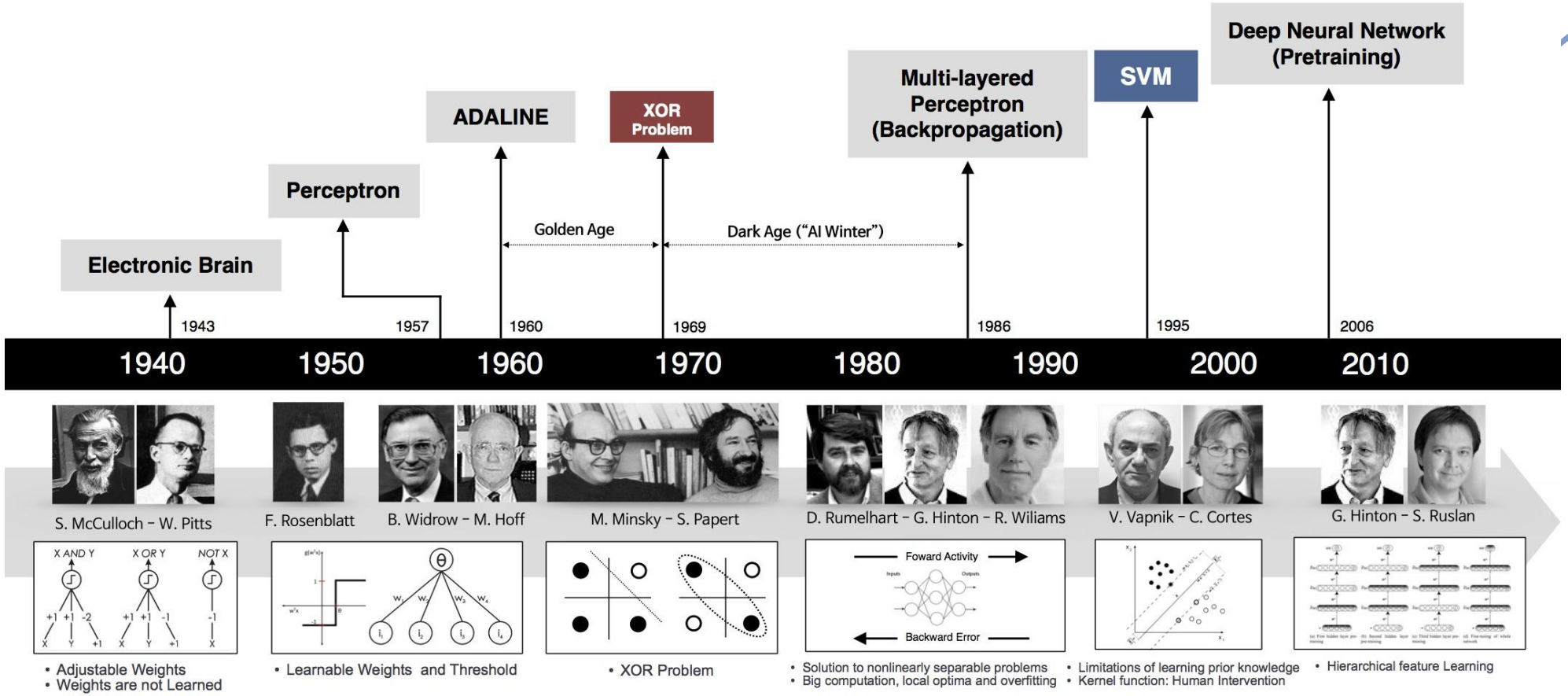


This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 731667 (MULTIDRONE)

Introduction

Important history of Artificial Intelligence

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

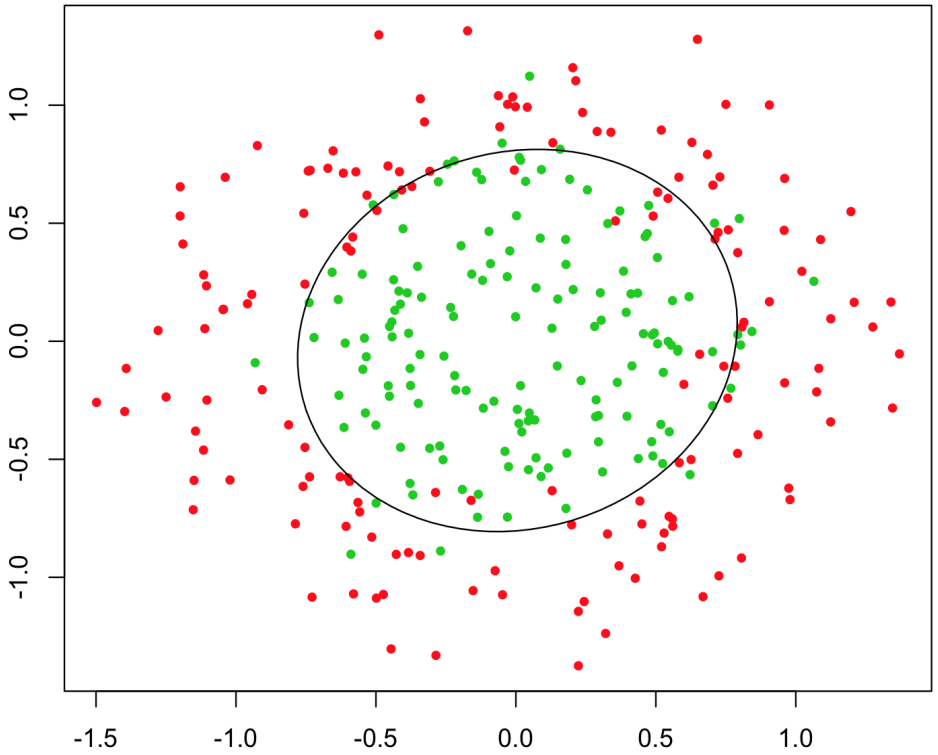


This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731667 (MULTIDRONE)

Introduction

Advantages of Artificial Neural Networks

Satisfactory separation of non-linear separable input data



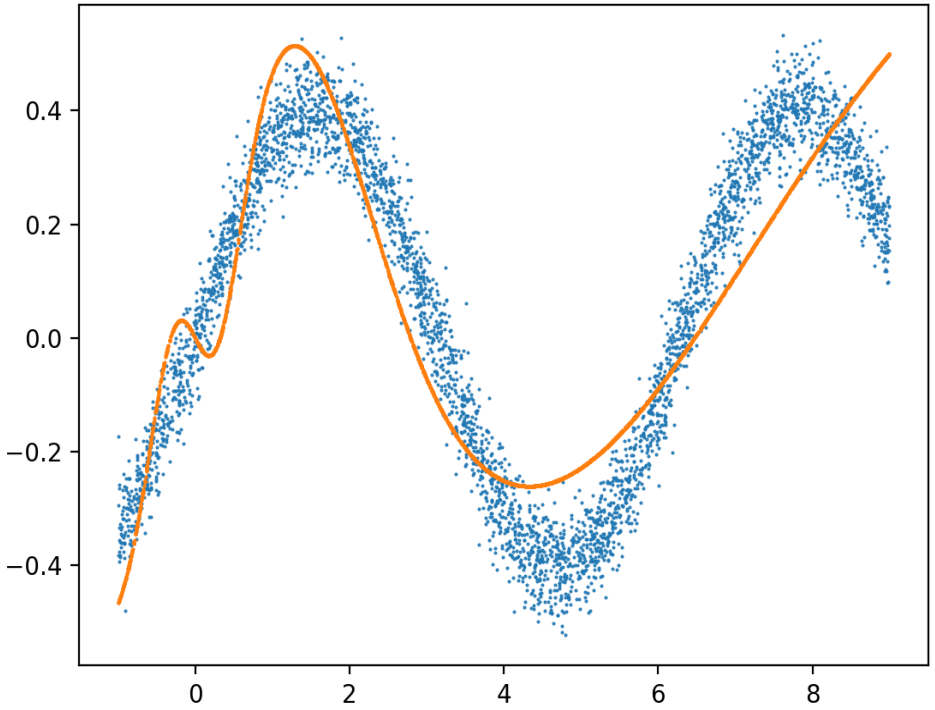
This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731667 (MULTIDRONE)

Introduction

Advantages of Artificial Neural Networks

Universal Approximation Theorem (George Cybenko)

Satisfactory function approximation using only input data

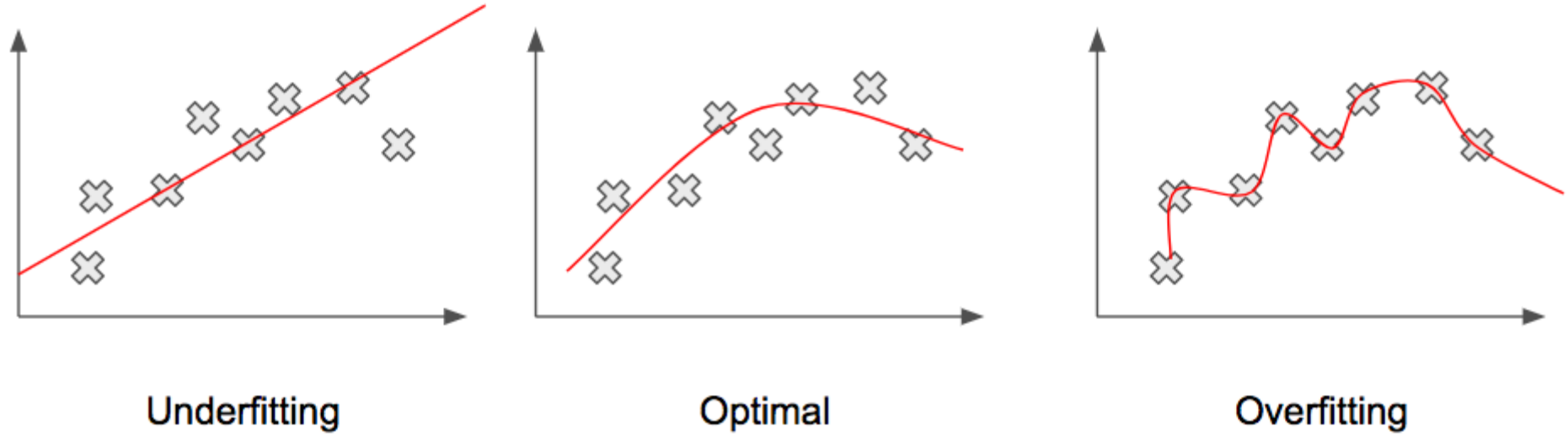


This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731667 (MULTIDRONE)

Introduction

Advantages of Artificial Neural Networks

Good generalization that captures the general manifold of data

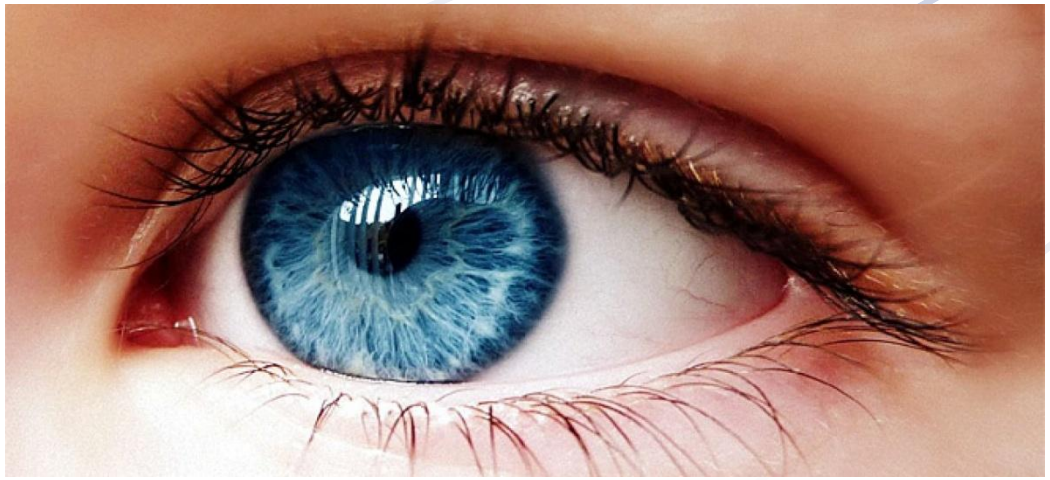


This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731667 (MULTIDRONE)

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



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

Introduction

Computer Vision and its future goal

1. Q: Is there a person in the blue region?	A: yes
2. Q: Is there a unique person in the blue region? (Label this person 1)	A: yes
3. Q: Is person 1 carrying something?	A: yes
4. Q: Is person 1 female?	A: yes
5. Q: Is person 1 walking on a sidewalk?	A: yes
6. Q: Is person 1 interacting with any other object?	A: no
⋮	
9. Q: Is there a unique vehicle in the yellow region? (Label this vehicle 1)	A: yes
10. Q: Is vehicle 1 light-colored?	A: yes
11. Q: Is vehicle 1 moving?	A: no
12. Q: Is vehicle 1 parked and a car?	A: yes
⋮	
14. Q: Does vehicle 1 have exactly one visible tire?	A: no
15. Q: Is vehicle 1 interacting with any other object?	A: no
17. Q: Is there a unique person in the red region?	A: no
18. Q: Is there a unique person that is female in the red region?	A: no
19. Q: Is there a person that is standing still in the red region?	A: yes
20. Q: Is there a unique person standing still in the red region? (Label this person 2)	A: yes
⋮	
23. Q: Is person 2 interacting with any other object?	A: yes
24. Q: Is person 1 taller than person 2?	A: amb.
25. Q: Is person 1 closer (to the camera) than person 2?	A: no
26. Q: Is there a person in the red region?	A: yes
27. Q: Is there a unique person in the red region? (Label this person 3)	A: yes
⋮	
36. Q: Is there an interaction between person 2 and person 3?	A: yes
37. Q: Are person 2 and person 3 talking?	A: yes

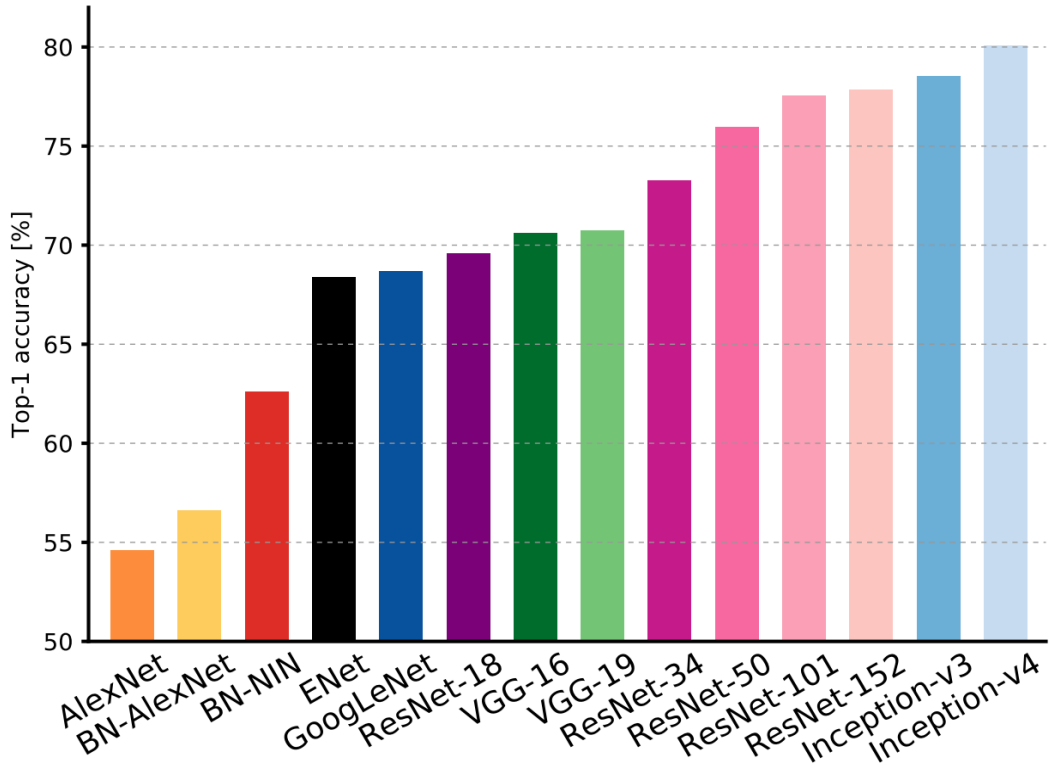
Visual Turing Test

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

Introduction

State-of-the-art Computer Vision

ImageNet Large Scale Visual Recognition Challenge (ILSVRC)

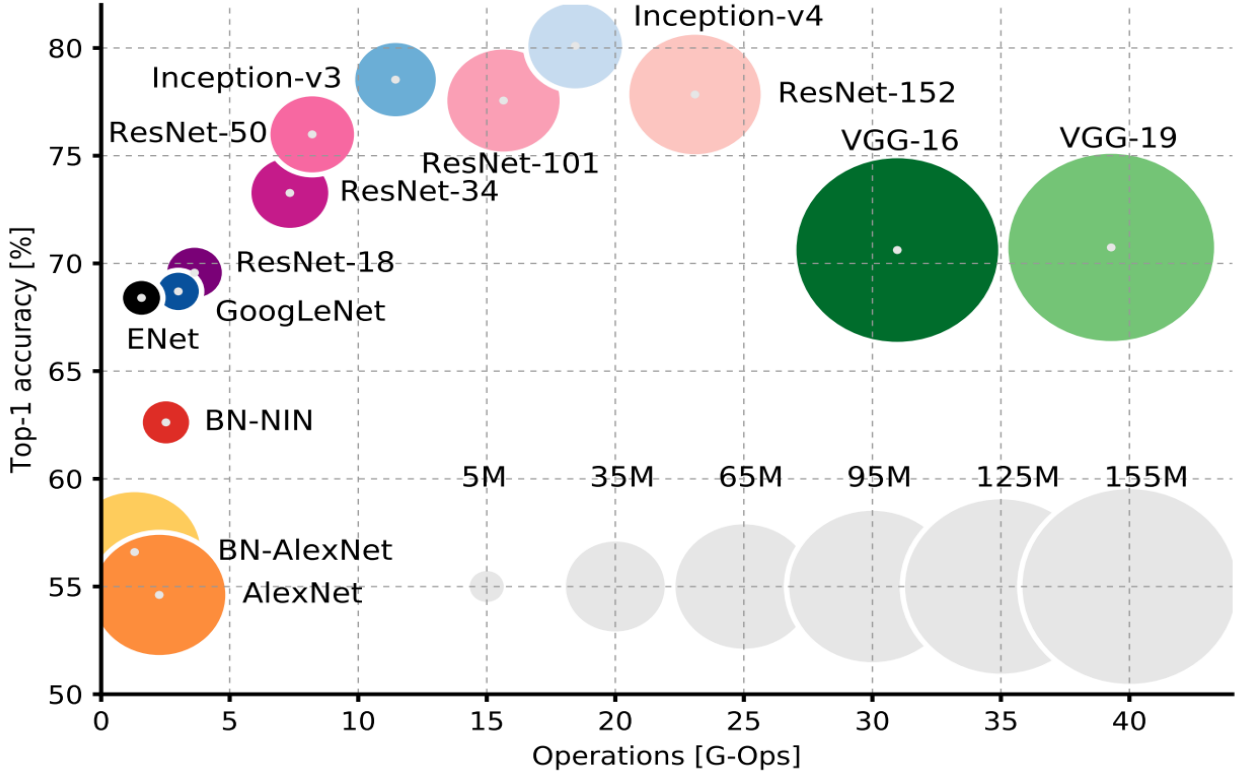


This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731667 (MULTIDRONE)

Introduction

State-of-the-art Computer Vision

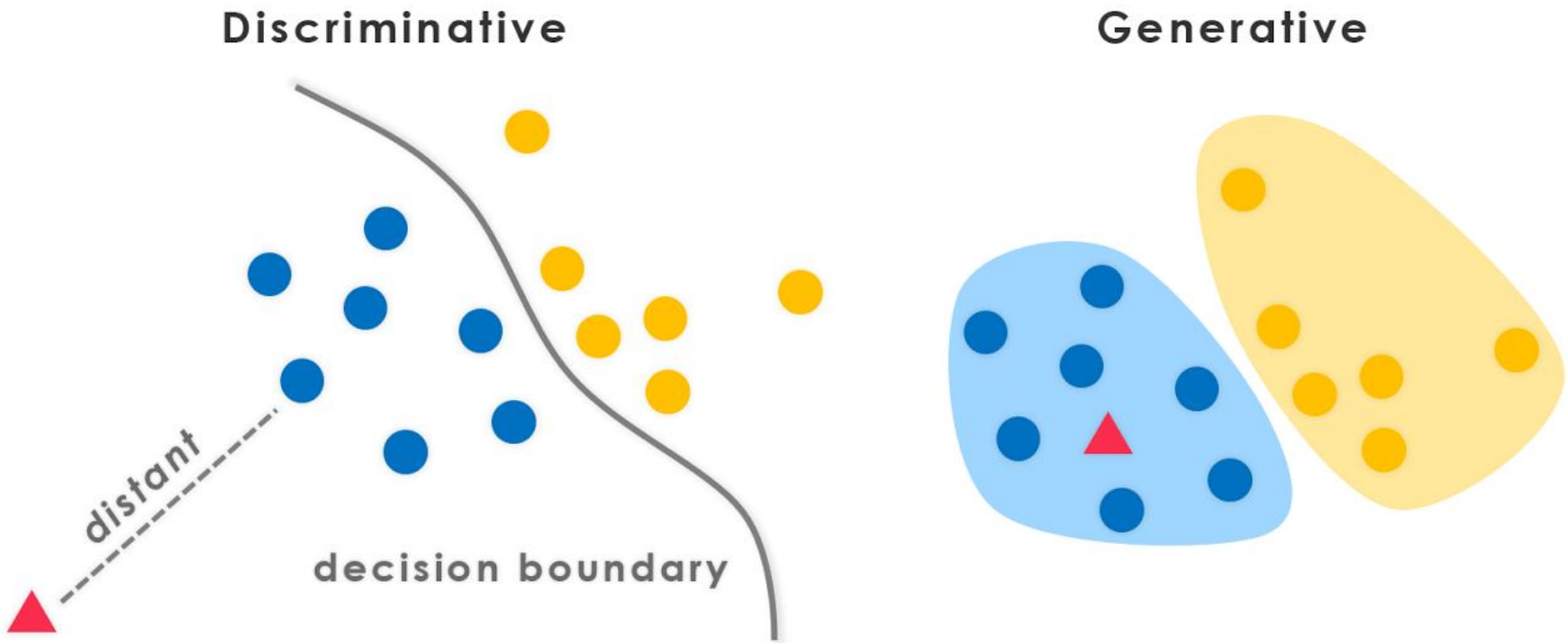
ImageNet Large Scale Visual Recognition Challenge (ILSVRC)



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

Introduction

Discriminative and Generative models

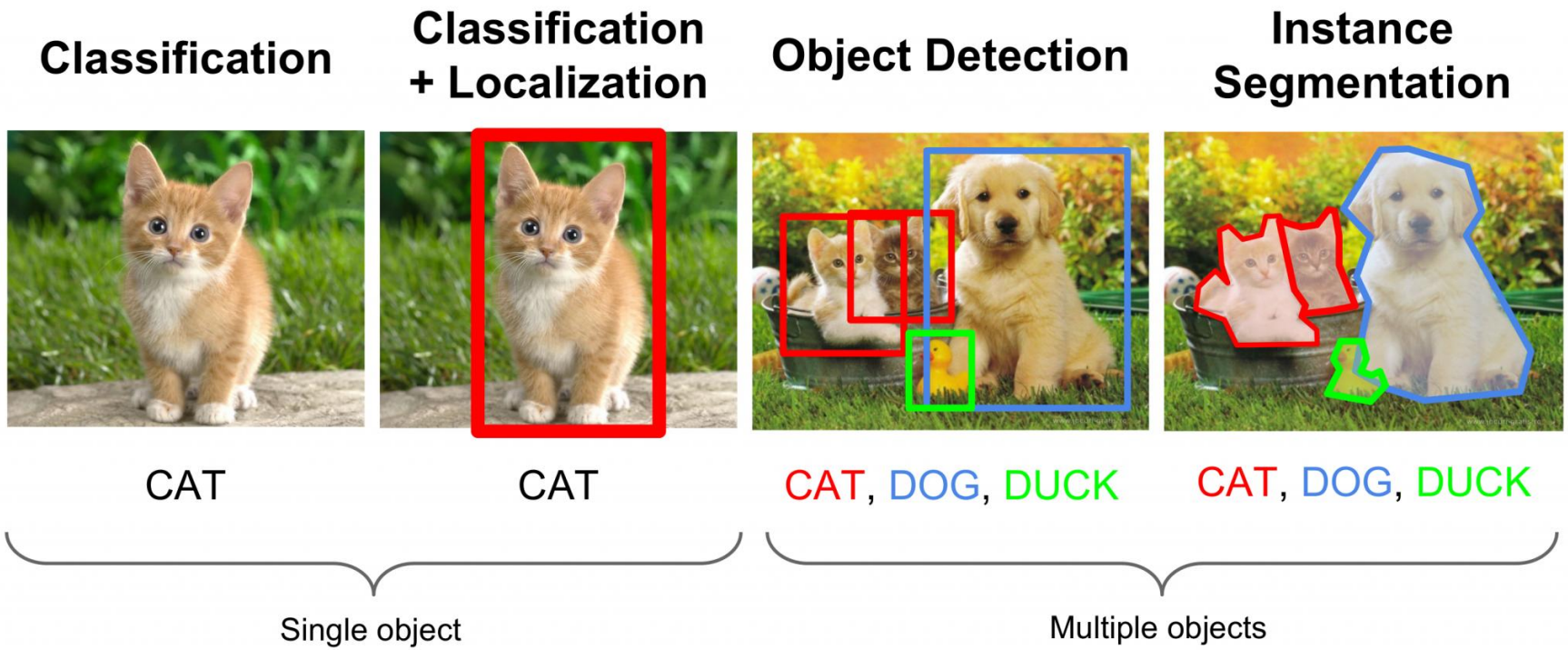


This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731667 (MULTIDRONE)

Introduction

Discriminative and Generative models

In the past years we have focused mostly on building discriminative models for classification and regression

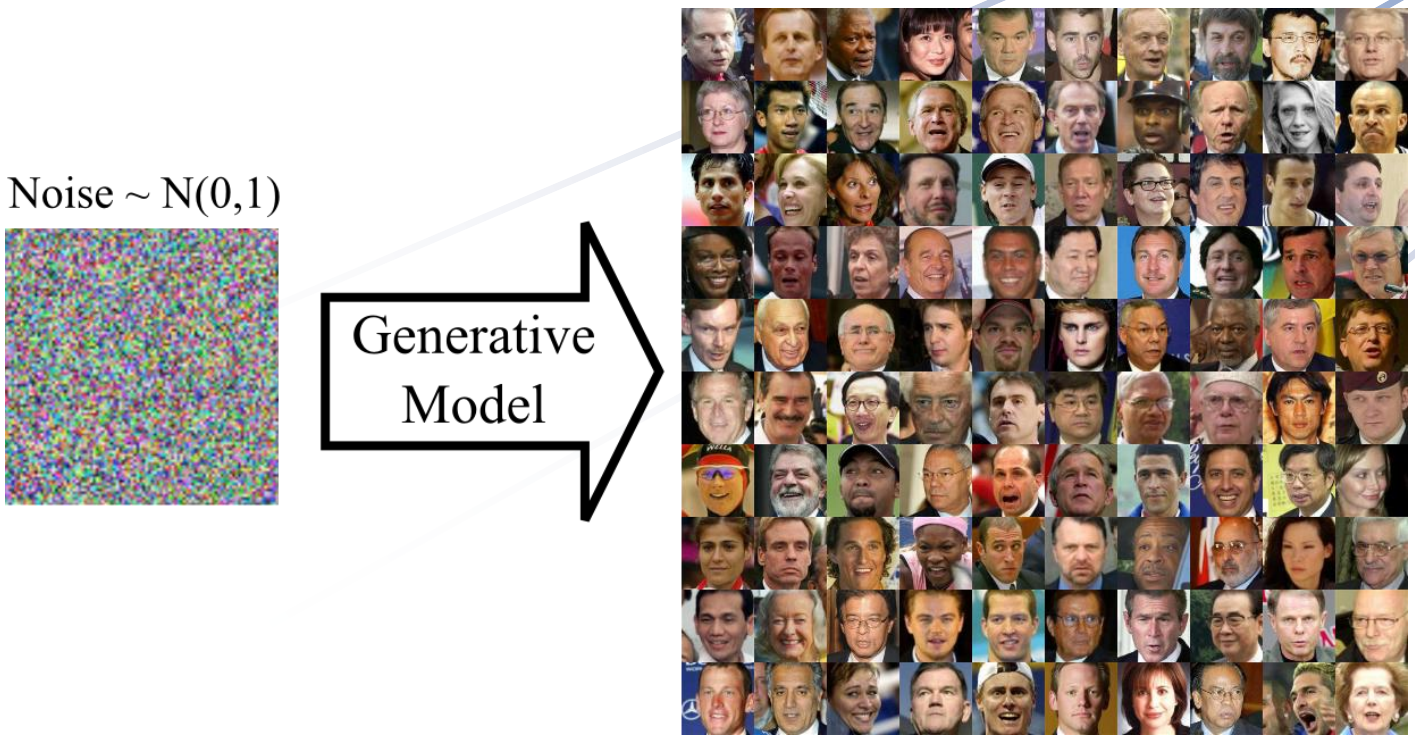


This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731667 (MULTIDRONE)

Introduction

Discriminative and Generative models

Nowadays, we are focusing also on generative models for generating artificially realistic data



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

Adversarial Examples

Local Generalization in Computer Vision

MultiDrone

Slight pixel changes should not affect the decision of a model

$$\textit{There is an image and its ground truth class: } x, y$$
$$\textit{A classification model } f \textit{ predicts the class of } x\textit{: } \hat{y} = f(x)$$
$$\textit{The prediction } \hat{y} \textit{ is the same as the ground truth: } y = \hat{y}$$
$$\textit{There is an image } x_p \textit{ which is } x \textit{ perturbated by } p\textit{: } x_p = x + p$$
$$\textit{The distance of the two images is restricted by threshold } e\textit{: } d(x, x_p) \leq e$$
$$\textit{The threshold } e \textit{ is positive and small for imperceptibility changes}$$
$$\textit{The classification model classifies the image } x_p \textit{ same as } x\textit{: } \hat{y_p} = \hat{y}$$

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

Adversarial Examples

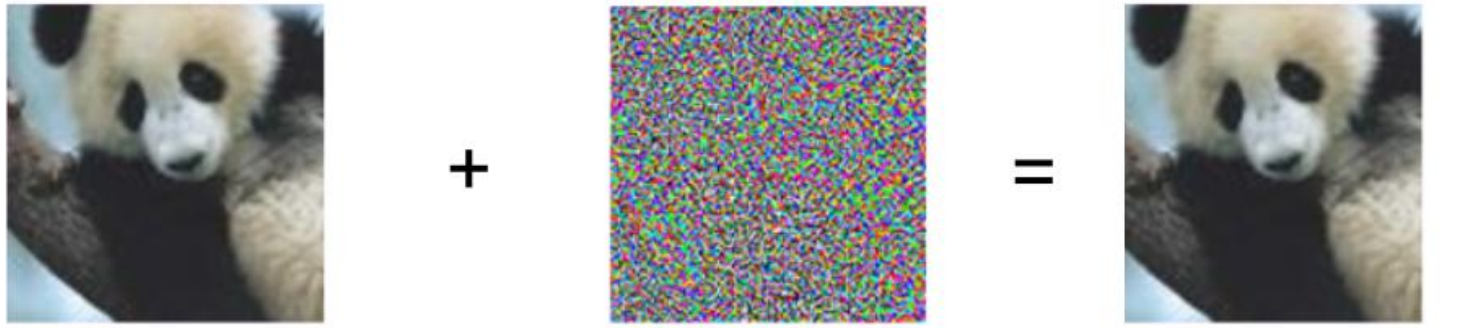
Local Generalization in Computer Vision

Slight pixel changes should not affect the decision of a model

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

Adversarial Examples

What exactly are these?



“panda”
57.7% confidence

“gibbon”
99.3% confidence

- Examples where the Local Generalization does not apply
- Perturbation: Optimal direction to move all the pixels so that the model will do a mistake
- Most models fail to work (LR, Softmax Regression, SVM, k-NN, Decision Trees, Neural Nets, Ensembles)

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

Adversarial Examples

Important related papers

Fooling a spam filter
"Adversarial Classification"
Nilesh Dalvi et al, 2004

Fooling neural networks
"Evasion Attacks Against Machine Learning at Test Time"
Battista Biggio et al, 2013

Fooling ImageNet classifiers
"Intriguing properties of neural networks"
Christian Szegedy et al, 2013

Cheap, closed form attack
"Explaining and Harnessing Adversarial Examples"
Ian J. Goodfellow et al, 2014

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

Adversarial Examples

The big question

Why adversarial examples exist and how is it feasible a model that is not overfitted and has high test/validation accuracy not to be functional in adversarial examples which are very similar to the original data?

Morpheus to Neo: “I imagine that right now you're feeling a bit like Alice, tumbling down the rabbit hole.”, The Matrix (1999)

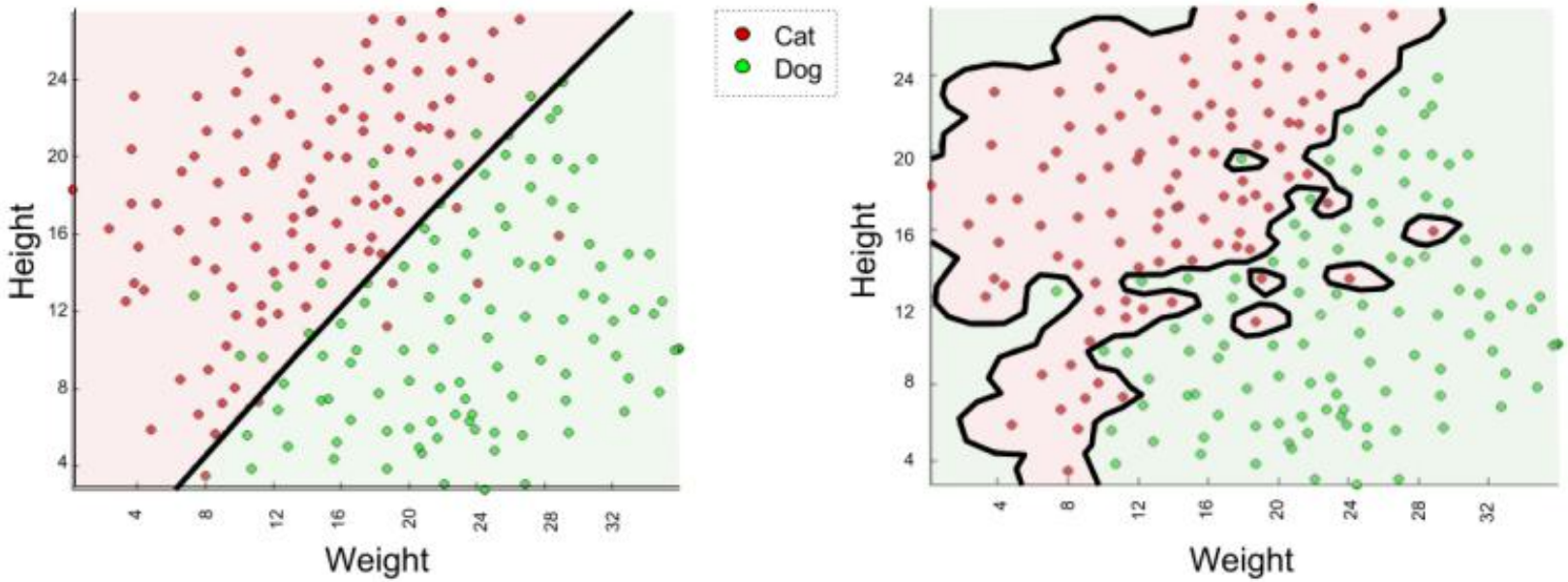
This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 731667 (MULTIDRONE)

Adversarial Examples

Why do they exist?

Is it due to overfitting?

a) an overfitted model is sensitive to small input changes since it learns the idiosyncrasies of input



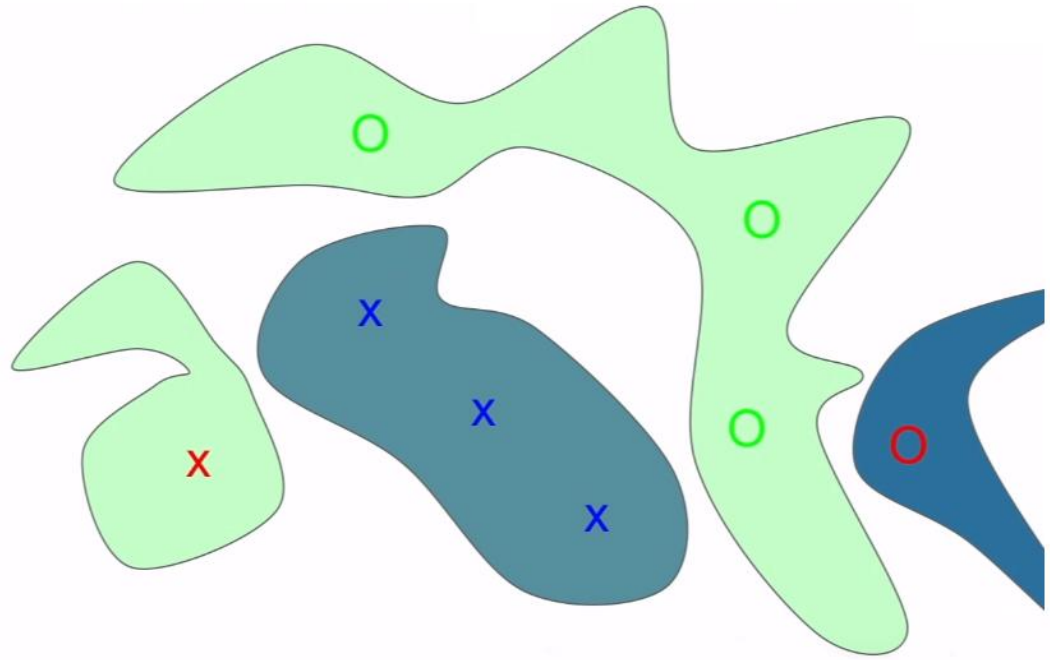
This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731667 (MULTIDRONE)

Adversarial Examples

Why do they exist?

Is it due to overfitting?

b) an overfitted model assigns some blobs of probabilities mass in unseen places of input



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

Adversarial Examples

Why do they exist?

If adversarial examples exist due to (a) and (b) then these are just the result of bad luck and should not be transferable between different model settings since different models give different probabilities in unseen places of input.

However, the adversarial examples are transferable between different model settings and that seems that is something more systematic than random.

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

Adversarial Examples

Why do they exist?

Latest research supports that it is due to high-dimensional input and linearity of models

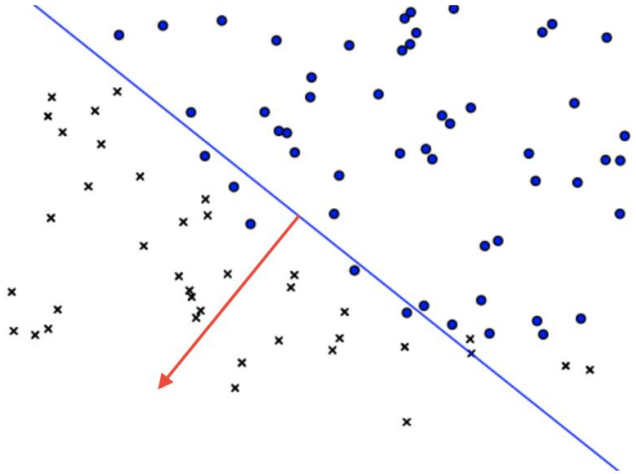
Many small pixel changes in high-dimensional input lead all together to a huge negative side effect

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

Adversarial Examples

Why do they exist?

Linear models are quite pathological outside of the region where training data is concentrated (initial experiments with shallow linear models shown that this affects greatly the w^Tx calculation and leads to wrong misclassification)

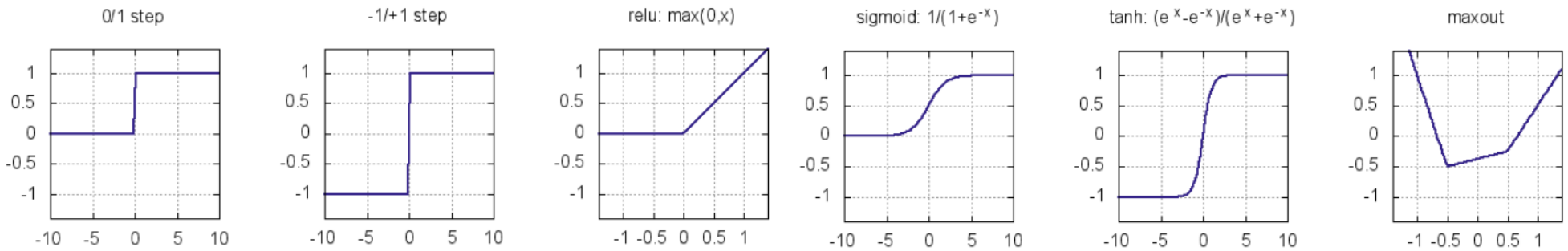


This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731667 (MULTIDRONE)

Adversarial Examples

Why do they exist?

Why adversarial examples exist also in non-linear models (e.g. Neural Networks)? Although by definition these are non-linear, are designed knowingly to be piecewise linear. Nowadays, state-of-the-art deep models have various piecewise linear elements.



"The fault is not in our stars, but in our activation functions, for they are piecewise linear"

"Explaining and Harnessing Adversarial Examples", 2014
Ian J. Goodfellow, Jonathon Shlens, Christian Szegedy

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

Adversarial Examples

Where does this lead us?

Achilles Heel

Potemkin Village

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

Adversarial Examples

How useful they are?

Adversarial examples can be used:

- to evaluate the robustness of a model
- to attack a model in order to break it
- in model training as a defensive regularization technique
- concealment / misclassification (e.g: de-detection / de-identification)

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

Adversarial Examples

Multiple ways to create them!

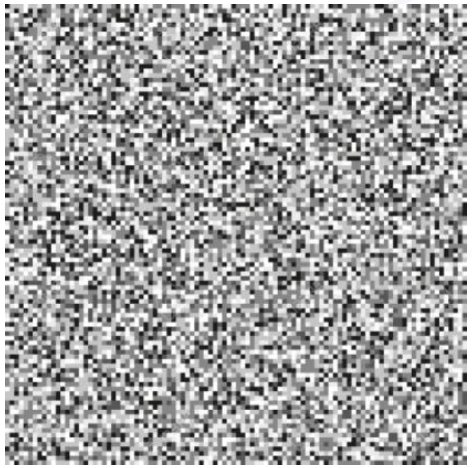
- Random / Non-random input
- Attack Method Frequency
- Adversarial Specificity
- Perturbation Scope
- Perturbation Measurement
- Realistic / Non-realistic Output
- Adversarial Falsification
- Adversary Knowledge
- Attack Method

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

Adversarial Examples

Random / Non-random input

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



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

Adversarial Examples

Attack Method Frequency

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

Adversarial Specificity

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

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

Adversarial Examples

Targeted / Non-targeted Definition

MultiDrone

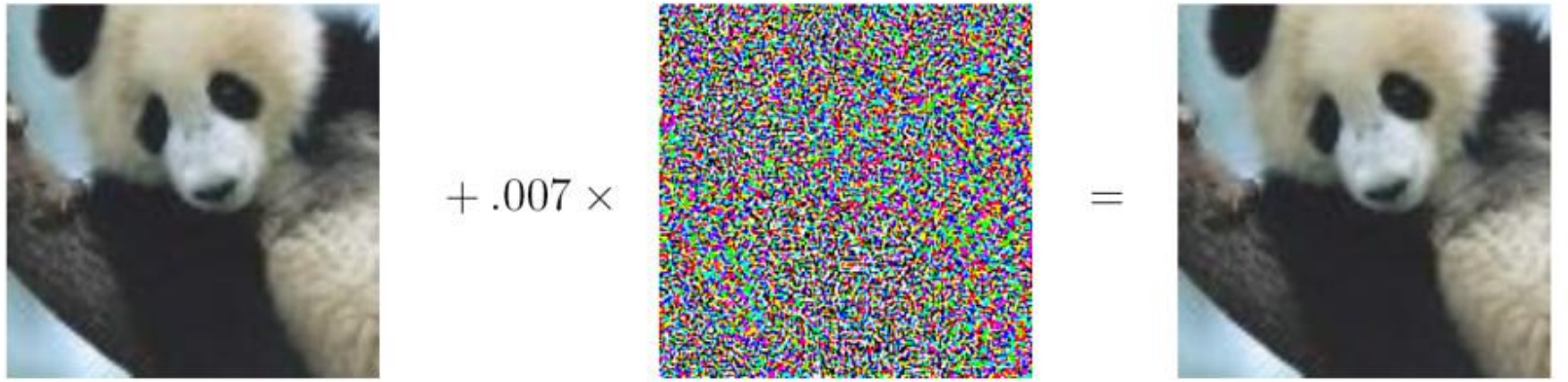
$$There\ is\ a\ sample\ and\ its\ ground\ truth\ class\colon x, y$$
$$A\ classification\ model\ f\ predicts\ the\ class\ of\ x\colon \hat{y} = f(x)$$
$$The\ prediction\ \hat{y}\ is\ the\ same\ as\ the\ ground\ truth\colon y = \hat{y}$$
$$There\ is\ a\ sample\ x_p\ which\ is\ x\ perturbated\ by\ p\colon x_p = x + p$$
$$The\ distance\ of\ the\ two\ samples\ is\ restricted\ by\ threshold\ e\colon d(x, x_p) \leq e$$
$$The\ threshold\ e\ is\ positive\ and\ small\ for\ imperceptibility\ changes$$
$$The\ classification\ model\ classifies\ the\ perturbated\ sample\ x_p\colon \hat{y}_p = f(x_p)$$
$$Non-targeted\ adversarial\ example\ constraint\colon \hat{y}_p \neq \hat{y}$$
$$Targeted\ adversarial\ example\ constraint\colon \hat{y}_p = y_{target}$$

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

Adversarial Examples

Perturbation Scope

- **Individual:** Each image has its own individual perturbation

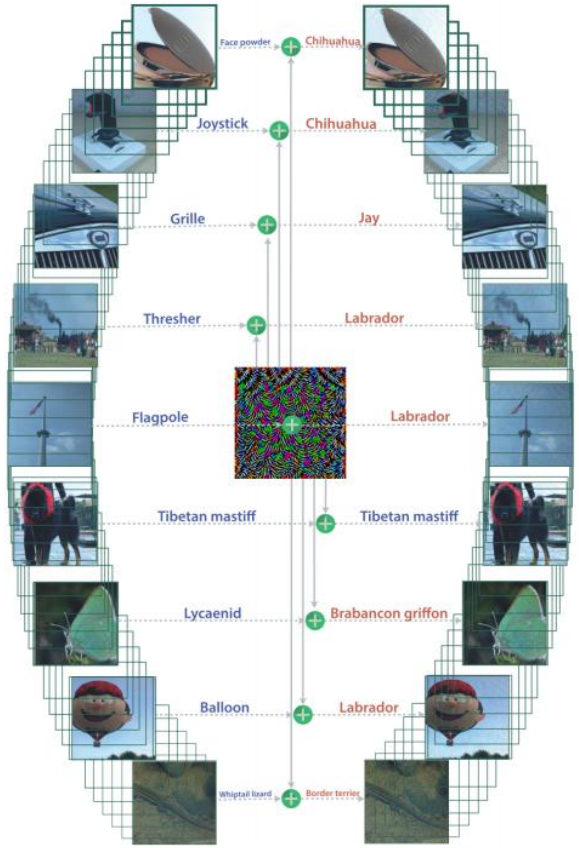


This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731667 (MULTIDRONE)

Adversarial Examples

Perturbation Scope

- **Universal:** A universal perturbation for the whole dataset



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

Adversarial Examples

Perturbation Measurement

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

Use an l_p-norm to calculate the magnitude of the adversarial perturbation:

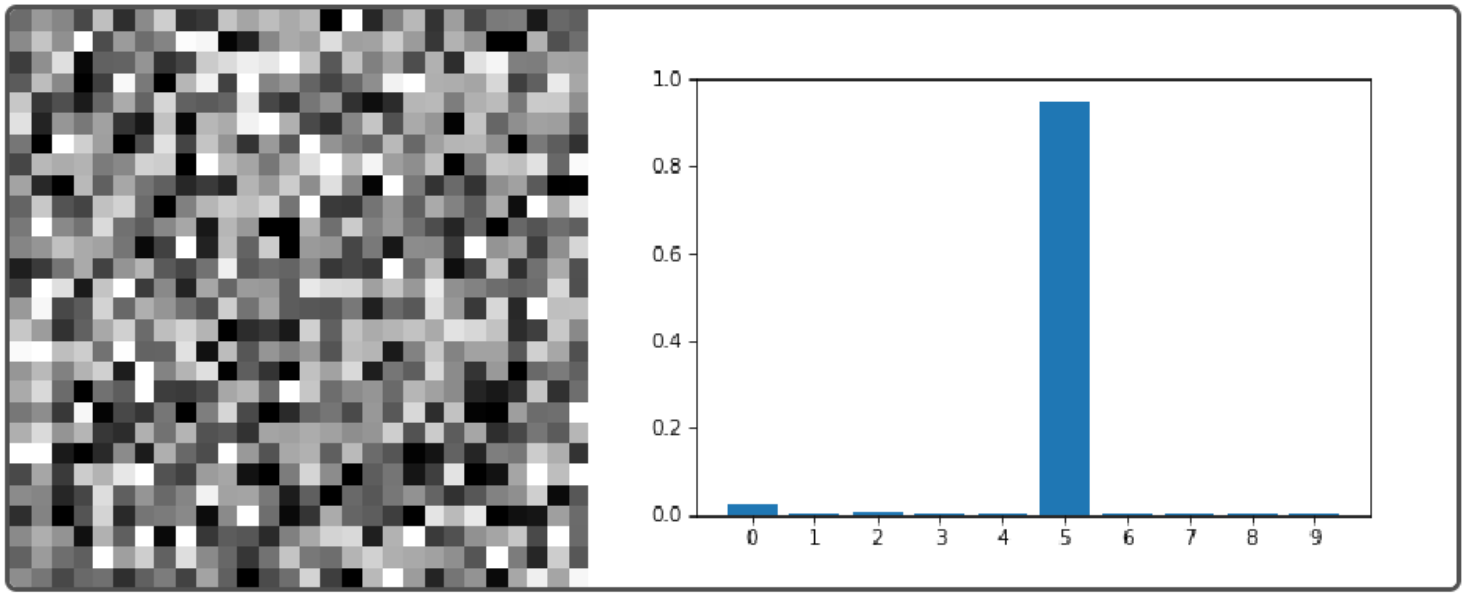
- l_0-norm (The total number of non-zero values in the vector)
- l_1-norm (The sum of the absolute values in the vector)
- l_2-norm (The square root of the sum of the squares of the values in the vector)
- l_∞-norm (The maximum of the absolute values in the vector)

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

Adversarial Examples

Realistic / Non-realistic Output

- **Non-realistic:** Any adversarial input that fools the model

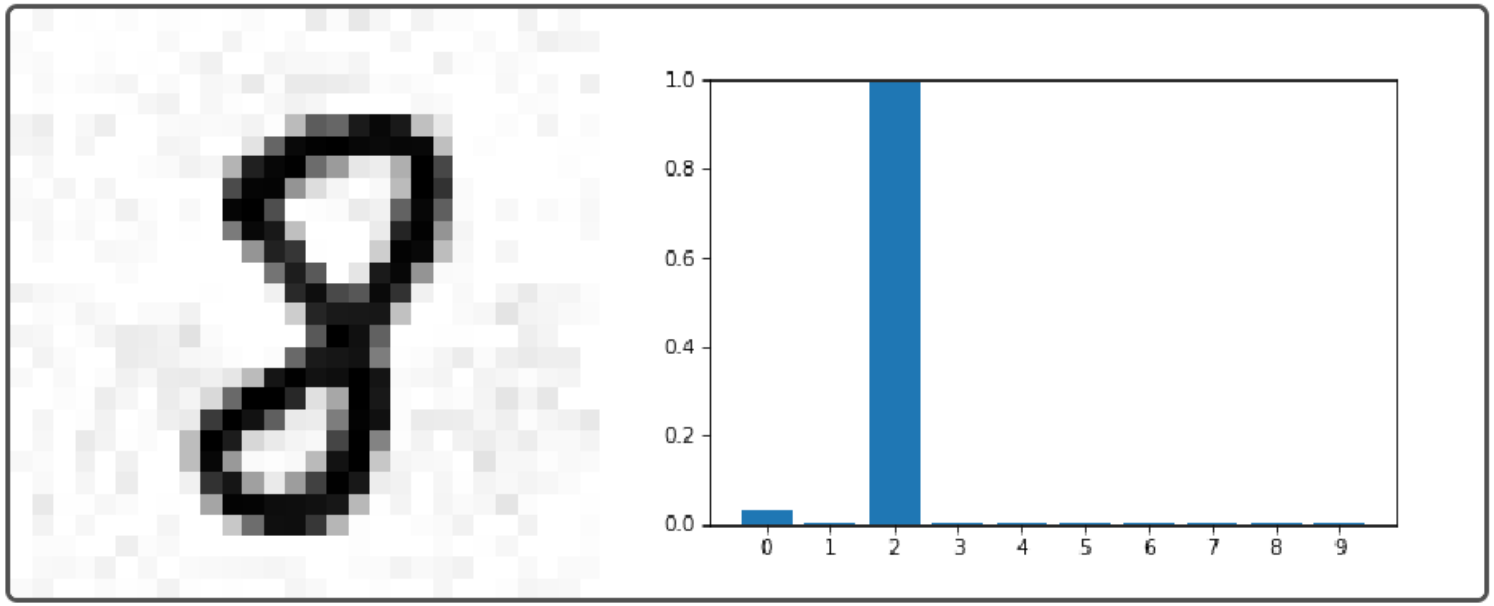


This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731667 (MULTIDRONE)

Adversarial Examples

Realistic / Non-realistic Output

- **Realistic:** An adversarial input that is close to a target example

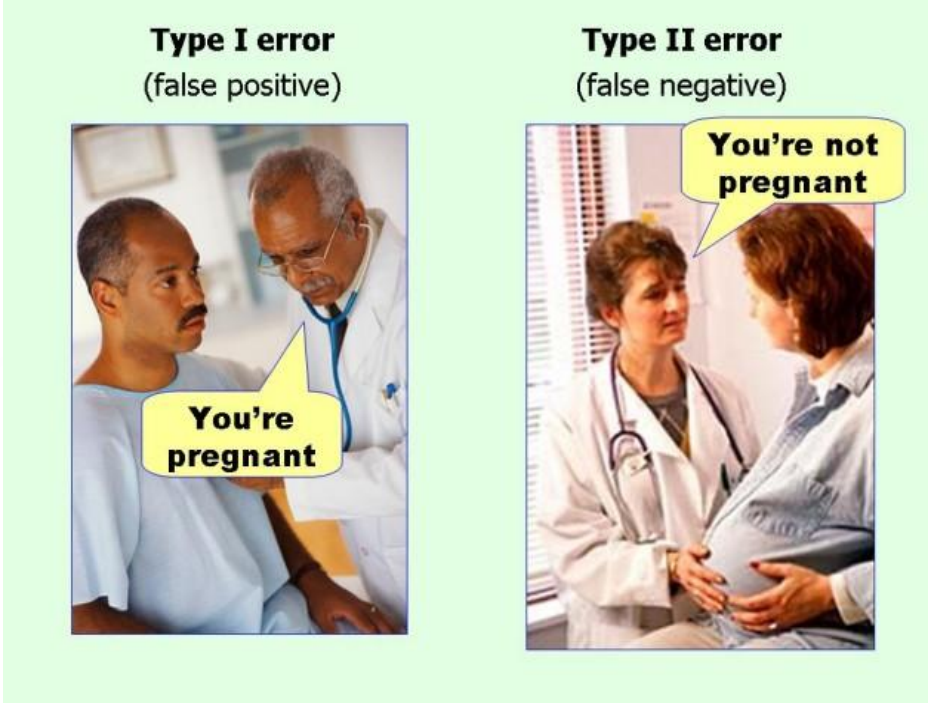


This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731667 (MULTIDRONE)

Adversarial Examples

Adversarial Falsification

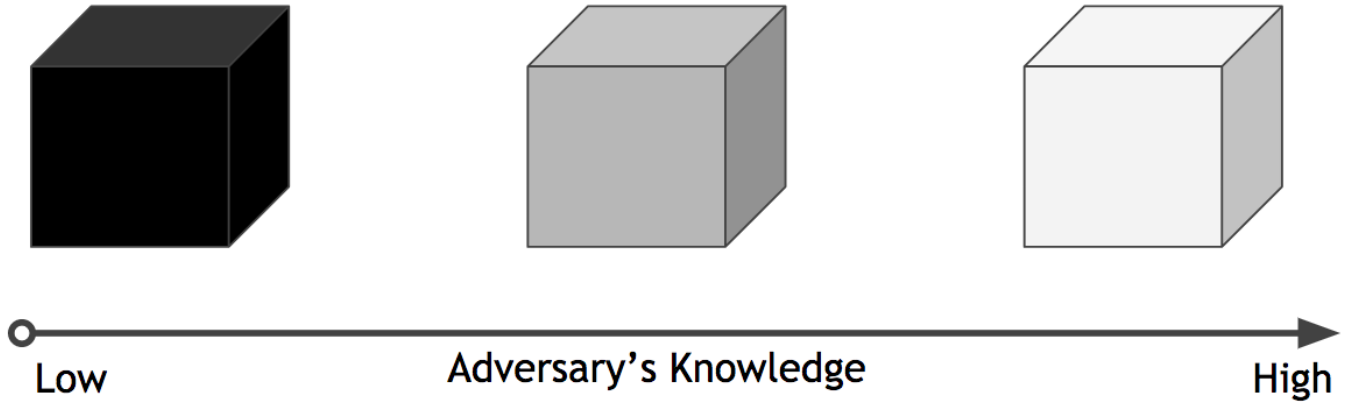
- **Type 1 error:** Negative sample classified as positive
- **Type 2 error:** Positive sample classified as negative



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

Adversarial Examples

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)

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

Adversarial Examples

Attack Methods

MultiDrone

L-BFGS: Limited-memory BFGS

Intriguing properties of neural networks

Christian Szegedy et al, 2013

$$minimize \|x - x^{adv}\|_2^2$$
$$such\ that$$
$$f(x^{adv}) = y_{target}$$
$$x^{adv} \in [0,1]^n$$

Given an input image x, the method finds a targeted adversarial valid image x^{adv} that is similar to x under squared l_2-norm and is labeled as y_{target} by the classifier f. However, this problem can be very difficult to solve. So, instead solve the following problem:

$$minimize\ c\ \|x - x^{adv}\|_2^2 + loss_f(x^{adv}, y_{target})$$
$$such\ that\ x^{adv} \in [0,1]^n$$

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

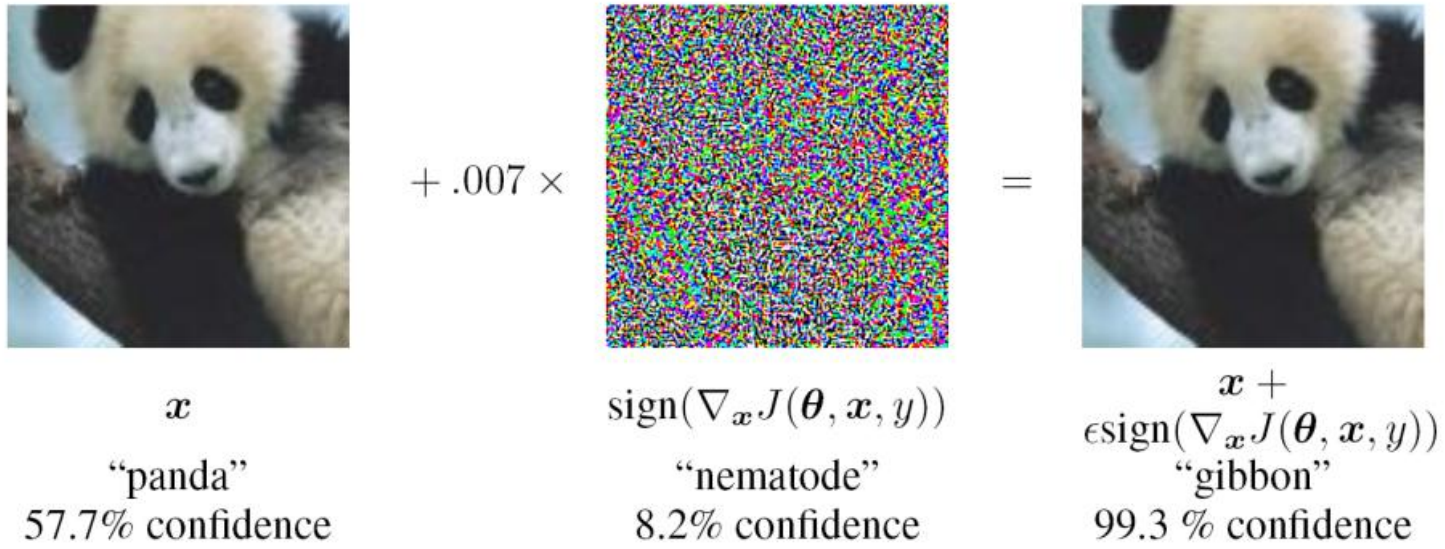
Adversarial Examples

Attack Methods

Let's get an idea with "Fast Gradient" Methods

Explaining and Harnessing Adversarial Examples

Ian J. Goodfellow et al, 2014

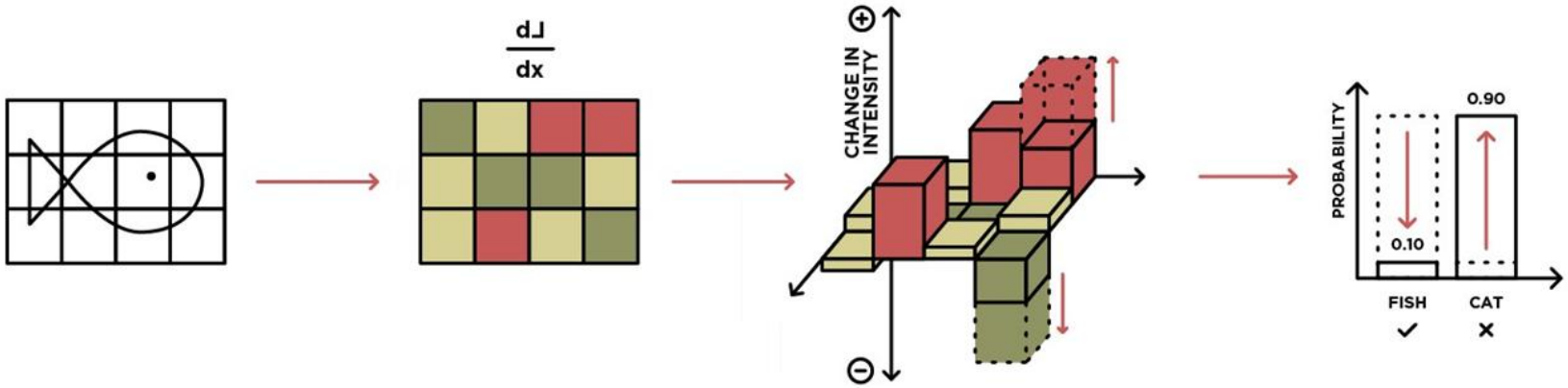


This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731667 (MULTIDRONE)

Adversarial Examples

Attack 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



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

Adversarial Examples

Attack Methods

FGSM : Fast Gradient Sign Method

- Non-targeted: One-shot method with gradient ascent
- Targeted: One-shot method with gradient descent

$$x^{adv} = x + \varepsilon.sign(\nabla_x J(x, y_{true}))$$

where

x – Clean Input Image
x^{adv} – Adversarial Image
J – Loss Function
y_{true} – Model Output for x
ε – Tunable Parameter

$$x^{adv} = x - \varepsilon.sign(\nabla_x J(x, y_{target}))$$

where

x – Clean Input Image
x^{adv} – Adversarial Image
J – Loss Function
y_{target} – Target Label
ε – Tunable Parameter

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

Adversarial Examples

Attack Methods

I-FGSM : Iterative Fast Gradient Sign Method

- Non-targeted: Iterative method with gradient ascent
- Targeted: Iterative method with gradient descent

$$x_0^{adv} = x, \quad x_{N+1}^{adv} = Clip_{x,\varepsilon}\{ x_N^{adv} + \alpha.sign(\nabla_x J(x_N^{adv}, y_{true}))\}$$

where

x – Clean Input Image
x_i^{adv} – Adversarial Image at ith step
J – Loss Function
y_{true} – Model Output for x
ε – Tunable Parameter
α – Step Size

$$x_0^{adv} = x, \quad x_{N+1}^{adv} = Clip_{x,\varepsilon}\{ x_N^{adv} - \alpha.sign(\nabla_x J(x_N^{adv}, y_{target})\}$$

where

x – Clean Input Image
x_i^{adv} – Adversarial Image at ith step
J – Loss Function
y_{target} – Target Label
ε – Tunable Parameter
α – Step Size

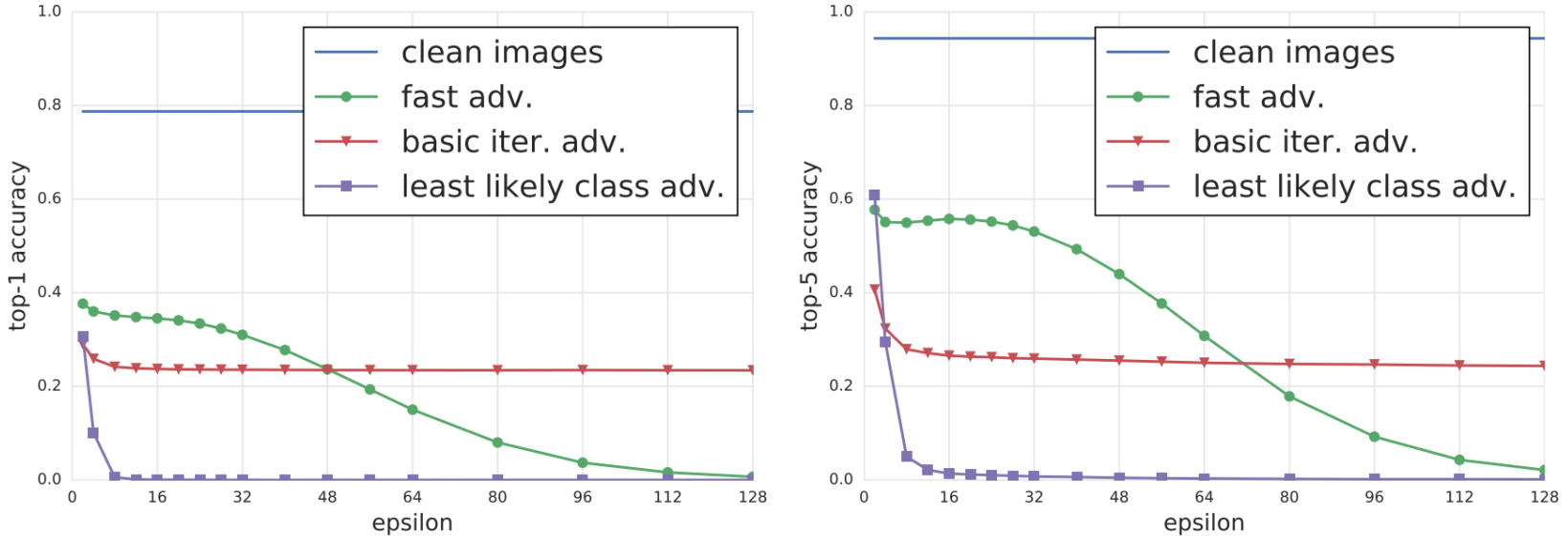
This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731667 (MULTIDRONE)

Adversarial Examples

Attack Methods

Iterative Least-Likely Class Method

Using I-FGSM with $y_{target} = \text{argmin}_y\{ p(y|x) \}$



Top-1 and Top-5 accuracy of Inception V3 under attack by different adversarial methods and different ‘e’ compared to unmodified images from the dataset. The accuracy was computed on all 50,000 validation images from the ImageNet dataset.

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

Adversarial Examples

Attack Methods

What would be more crazy? Of course, “Single Pixel” attack!

One pixel attack for fooling deep neural networks
Jiawei Su et al, 2017

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

Adversarial Examples

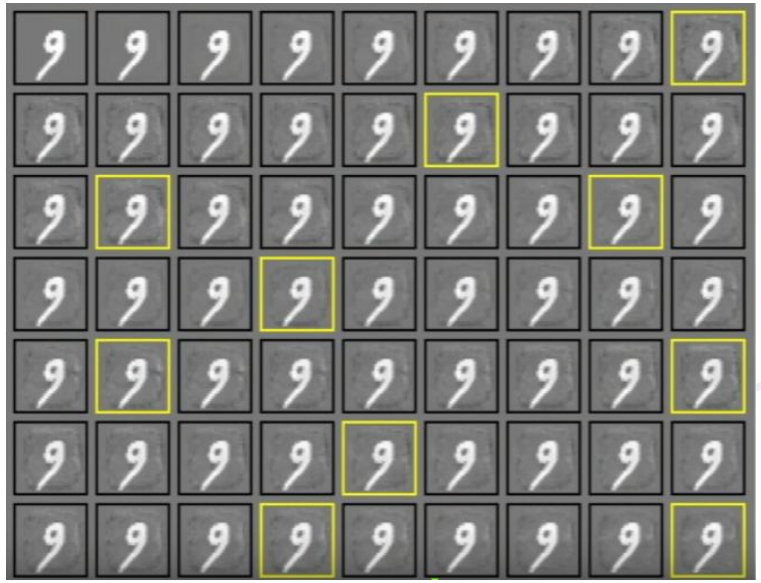
Attack Methods

- L-BFGS (Szegedy et al, 2013)
- Fast Gradient Sign Method (Goodfellow et al, 2015)
- Basic Iterative Method (Kurakin et al, 2016)
- One Pixel (Jiawei Su et al, 2018)
- DeepFool (Moosavi-Dezfooli et al, 2015)
- Jacobian-based Saliency Map Method (Papernot et al, 2016)
- SPSA (Uesato et al, 2018)
- Carlini Wagner L2 (Carlini and Wagner, 2016)
- Feature Adversaries (Sabour et al, 2016)
- Elastic Net Method (Chen et al, 2017)
- Virtual Adversarial Method (Miyato et al, 2016)
- The Momentum Iterative Method (Dong et al, 2017)
- Projected Gradient Descent (Madry et al, 2017)

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

Adversarial Examples

Softmax regression (MNIST)



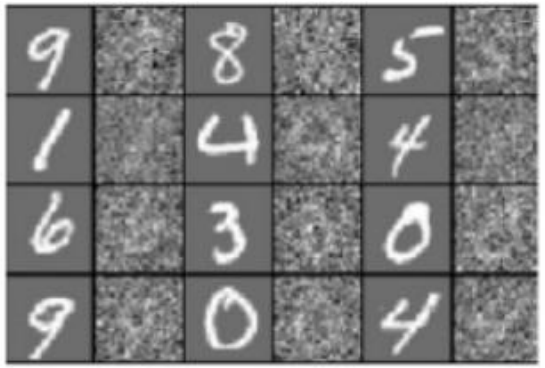
- The first image in the table is a random sample from the training data set
- The rest images are generated from top-to-bottom and left-to-right sequentially using L-BFGS
- Yellow Boxes: cases with >= 99% probability where the model thinks is 0, 1, …, 9
- The animal species recognize all above images as 9 but the model seems different

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

Adversarial Examples

Softmax regression (MNIST)

(a)



(b)

- Images in table's odd columns are random samples from the training data set
- Images in table's even columns correspond to distorted counterparts

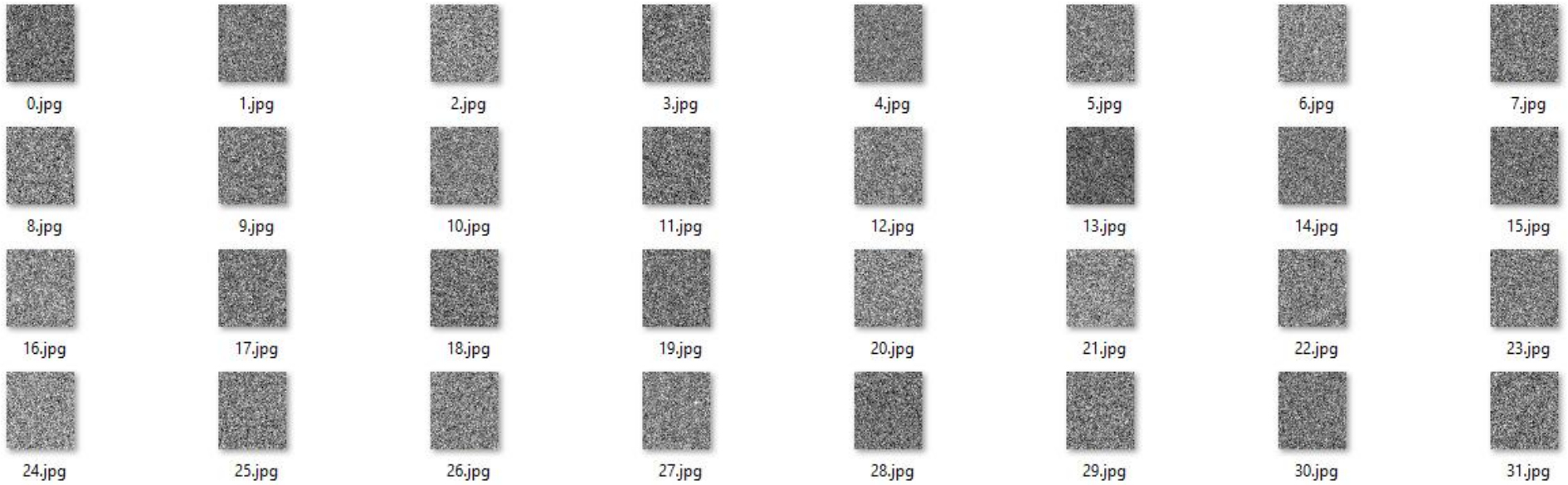
- (a) We successfully classify images with adversarial perturbation but models fail to do it
- (b) We fail to classify images with Gaussian noise added but models do it successfully

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

Adversarial Examples

MLP w/ Ext. Yale Face Database B

An example of face de-identification with an MLP model
Successfully target to any class with non-realistic images

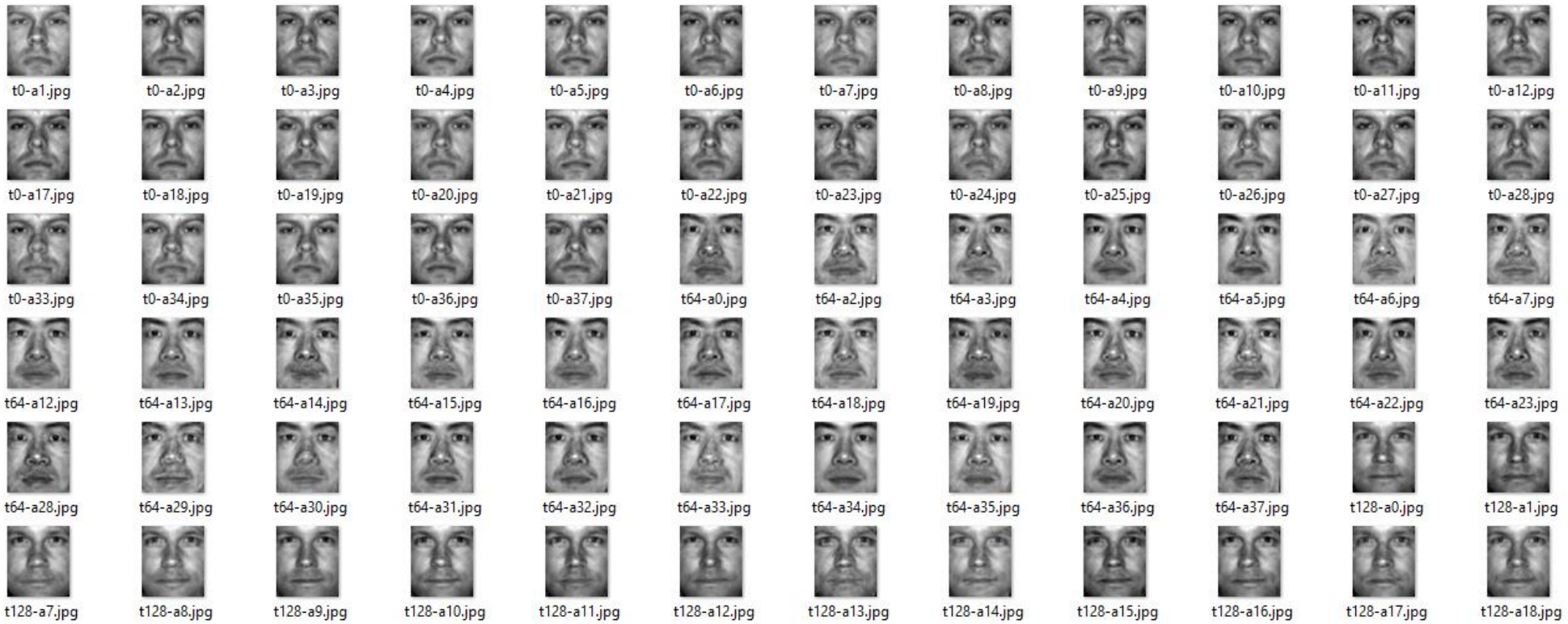


This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731667 (MULTIDRONE)

Adversarial Examples

MLP w/ Ext. Yale Face Database B

An example of face de-identification with an MLP model
Successfully target to any class with realistic images

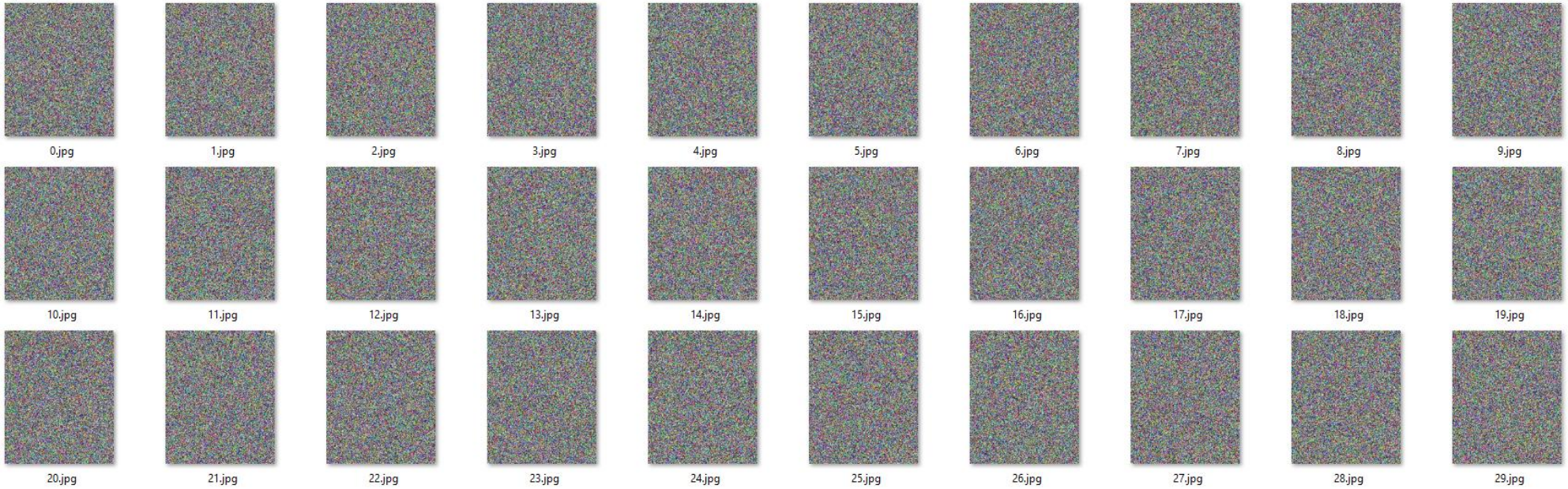


This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731667 (MULTIDRONE)

Adversarial Examples

CNN w/ CelebA

An example of face de-identification with a CNN model
Successfully target to any class with non-realistic images

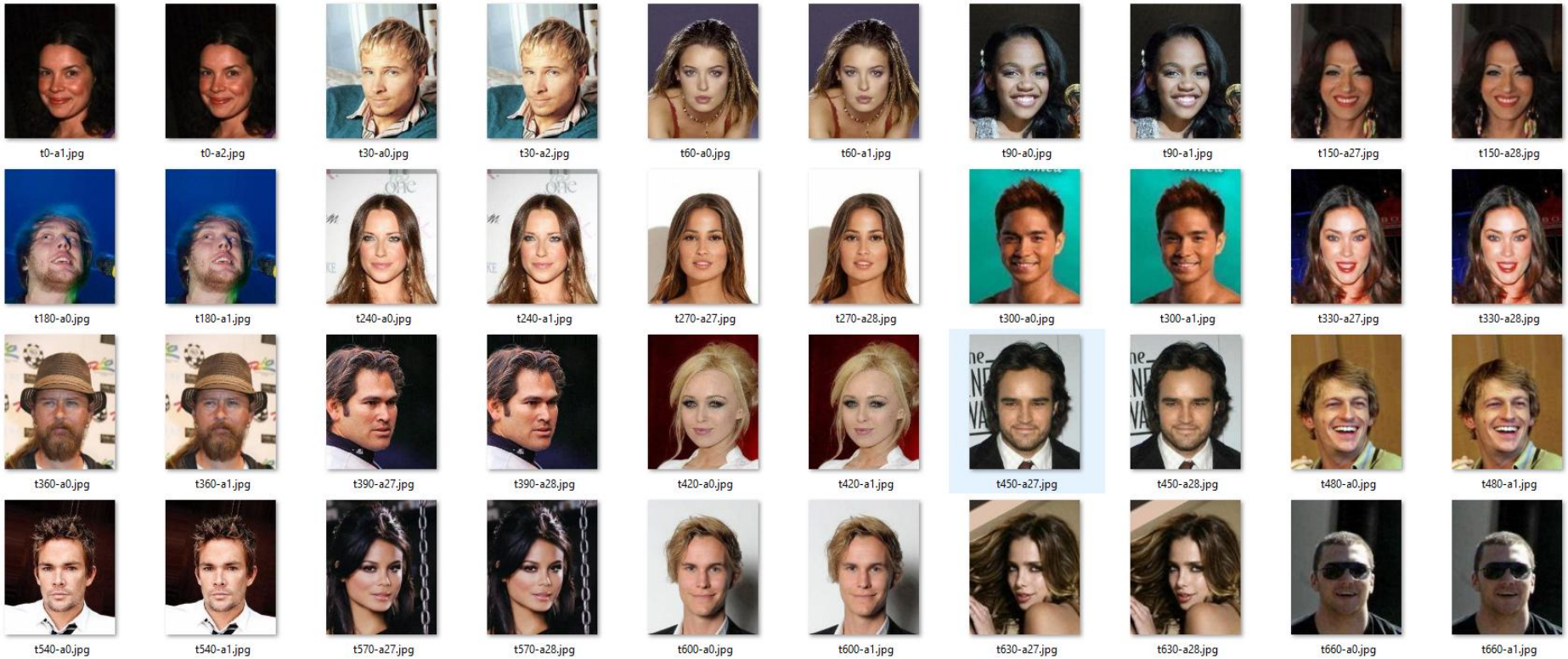


This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731667 (MULTIDRONE)

Adversarial Examples

CNN w/ CelebA

An example of face de-identification with a CNN model
Successfully target to any class with realistic images

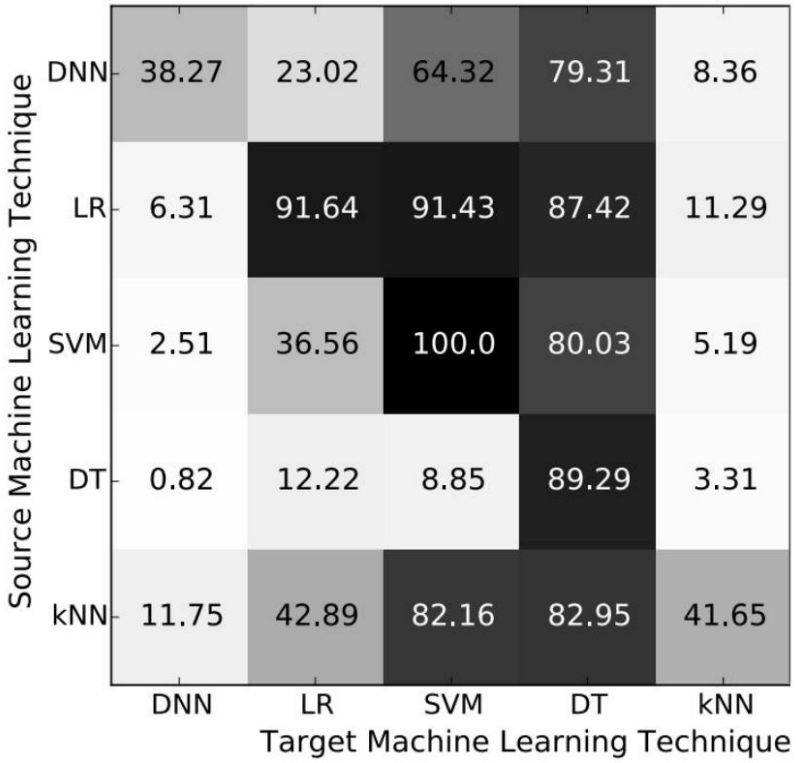


This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731667 (MULTIDRONE)

Adversarial Examples

They are transferable!

- Cross model generalization

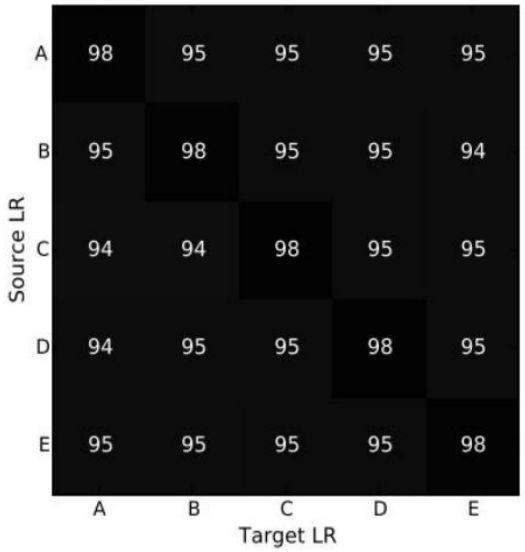


This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731667 (MULTIDRONE)

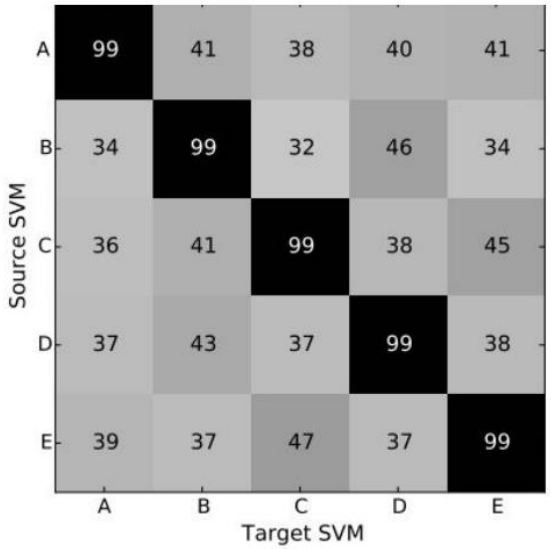
Adversarial Examples

They are transferable!

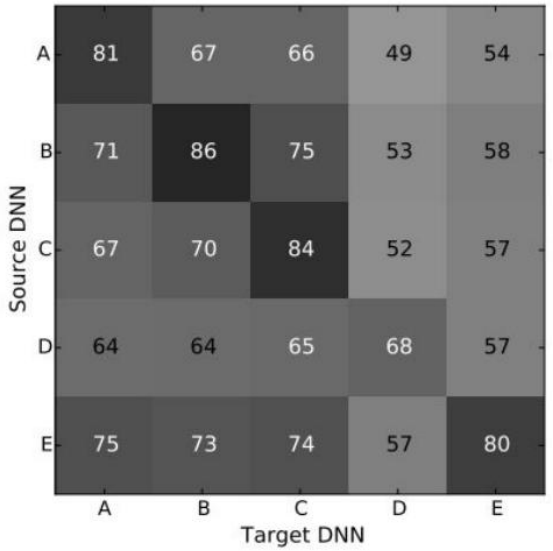
- Cross training-set generalization



Strong



Weak



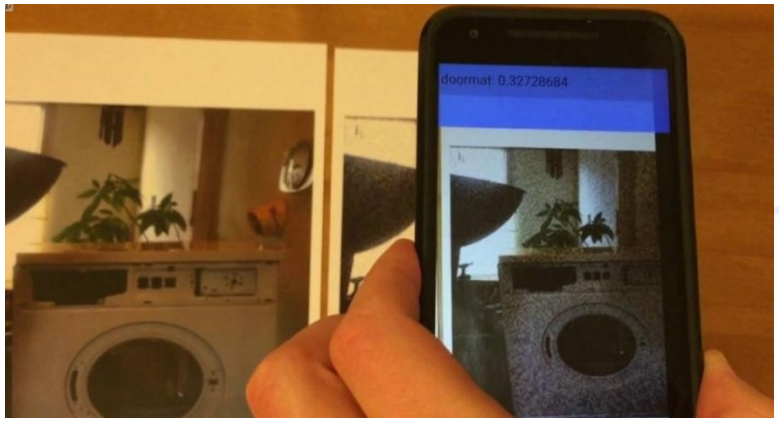
Intermediate

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

Adversarial Examples

They are transferable!

- Can exist to the real world!



Adversarial examples in the physical world
Alexey Kurakin et al, 2017

“We used images taken from a cell-phone camera as an input to an Inception V3 image classification neural network. We showed that in such a set-up, a significant fraction of adversarial images crafted using the original network are misclassified even when fed to the classifier through the camera.”, Kurakin et al.

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

Adversarial Examples

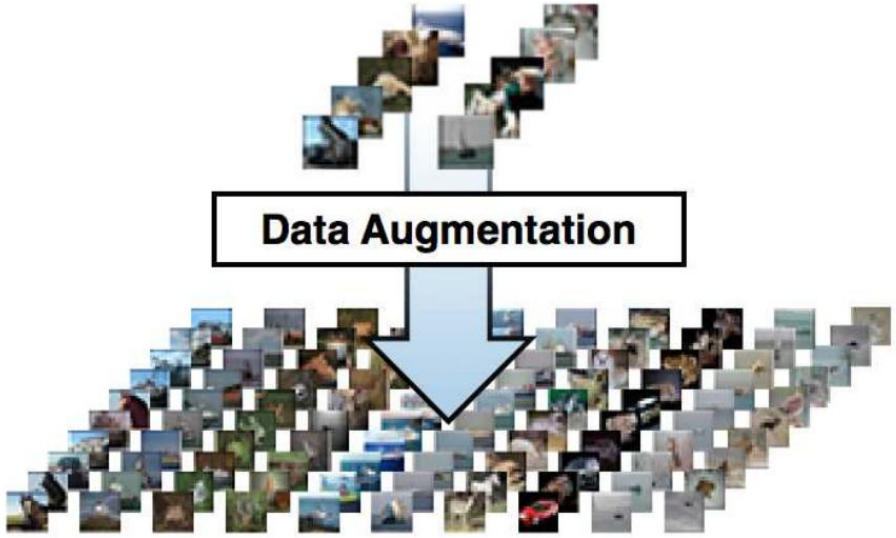
Adversarial Training

- The Neural Networks can approximate any function (Universal Approximation Theorem)
- We can train our Neural Networks with some adversarial examples to be more robust
- Adversarial training works as a regularization technique
- There is a need for sophisticated adversarial training algorithms with robustness by priori

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

Adversarial Examples

Adversarial Training



- Mixing training/validation/testing data sets with adversarial examples
- Same idea with data augmentation as a regularization technique

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

Adversarial Examples

Adversarial Training

- Using an adversarial loss function for single examples
- Generate fast adversarial examples at run-time with FGSM
- ‘c’ : hyperparameter coefficient to balance the weight of adversarial learning

$$J'(\theta, x, y) = c \times J(\theta, x, y) + (1 - c) \times J(\theta, x + p, y)$$
$$p = e \times sign(\nabla_x J(\theta, x, y))$$

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

Adversarial Examples

Adversarial Training

2016, “Adversarial Machine Learning at Scale”
Alexey Kurakin, Ian Goodfellow, Samy Bengio

Using an adversarial loss function for large scale models and datasets

Algorithm 1 Adversarial training of network N.
Size of the training minibatch is m. Number of adversarial images in the minibatch is k.

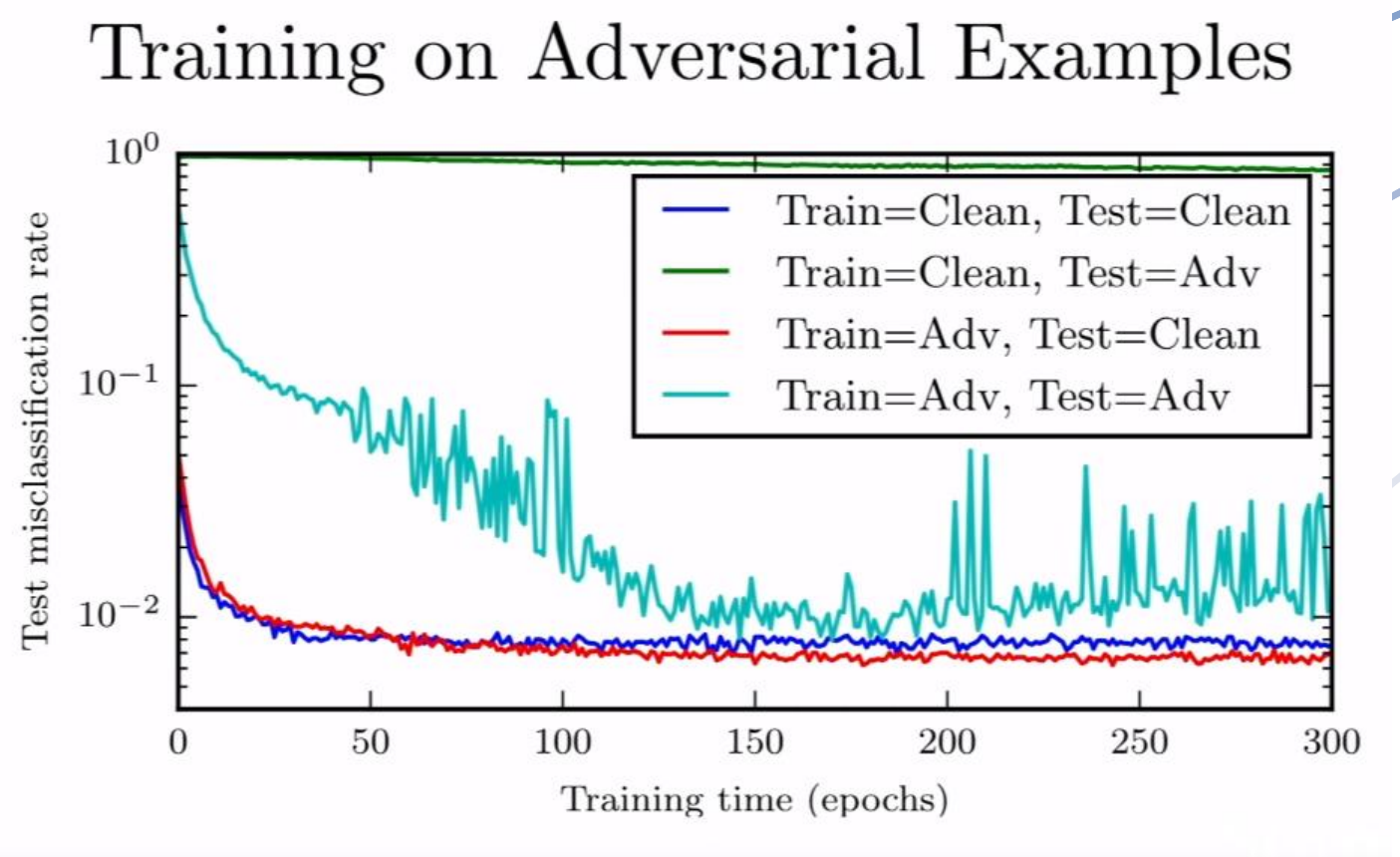
1: Randomly initialize network N
2: **repeat**
3: Read minibatch $B = \{X^1, \ldots, X^m\}$ from training set
4: Generate k adversarial examples $\{X^1_{adv}, \ldots, X^k_{adv}\}$ from corresponding clean examples $\{X^1, \ldots, X^k\}$ using current state of the network N
5: Make new minibatch $B' = \{X^1_{adv}, \ldots, X^k_{adv}, X^{k+1}, \ldots, X^m\}$
6: Do one training step of network N using minibatch B'
7: **until** training converged

$$Loss = \frac{1}{(m-k)+\lambda k}\left(\sum_{i \in CLEAN} L(X_i|y_i) + \lambda \sum_{i \in ADV} L(X_i^{adv}|y_i)\right)$$

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

Adversarial Examples

Adversarial Training



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

Generative Adversarial Networks

What exactly are these?

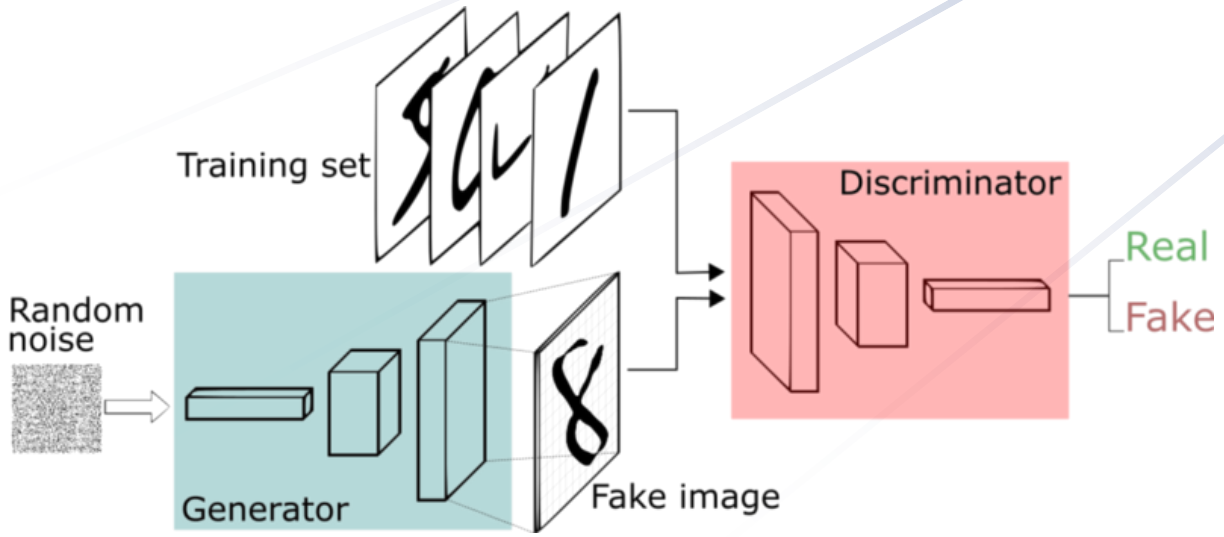
- Discriminative models map rich, high dimensional data input to low dimensions output
- Nowadays, interest has turned quite a bit on generative models which do the opposite
- GAN: A framework for evolving adversarially a generative and a discriminative model
- Generative model as ‘Generator’ or ‘G’ / Discriminative model as ‘Discriminator’ or ‘D’

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

Generative Adversarial Networks

What exactly are these?

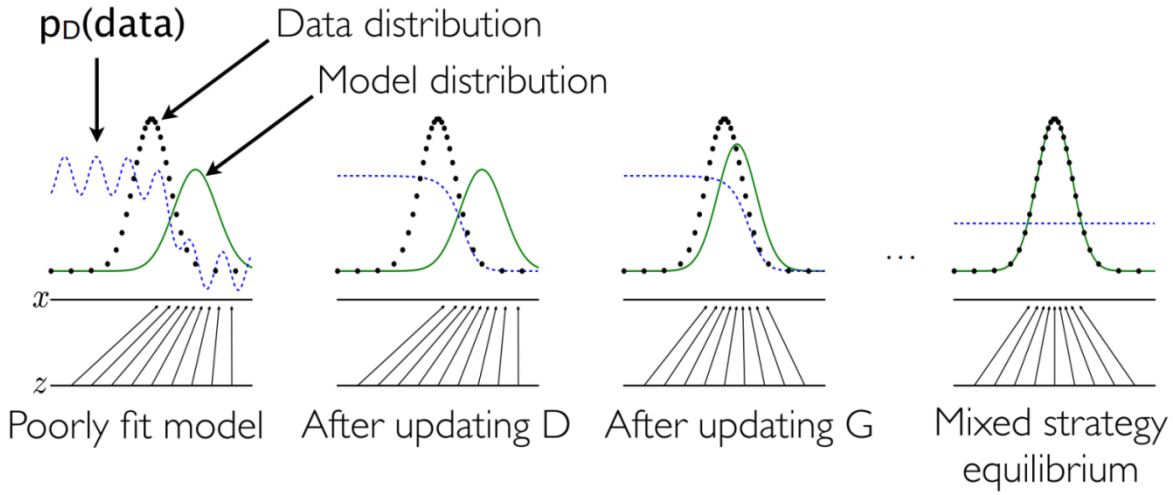
- ‘G’ and ‘D’ evolve through an adversarial process
- ‘G’ and ‘D’ are players in a two-player minimax game
- ‘G’ is trained to learn to generate fake data and approximate the real data distribution
- ‘D’ is trained to learn to separate the fake from the real data distribution



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

Generative Adversarial Networks

What exactly are these?



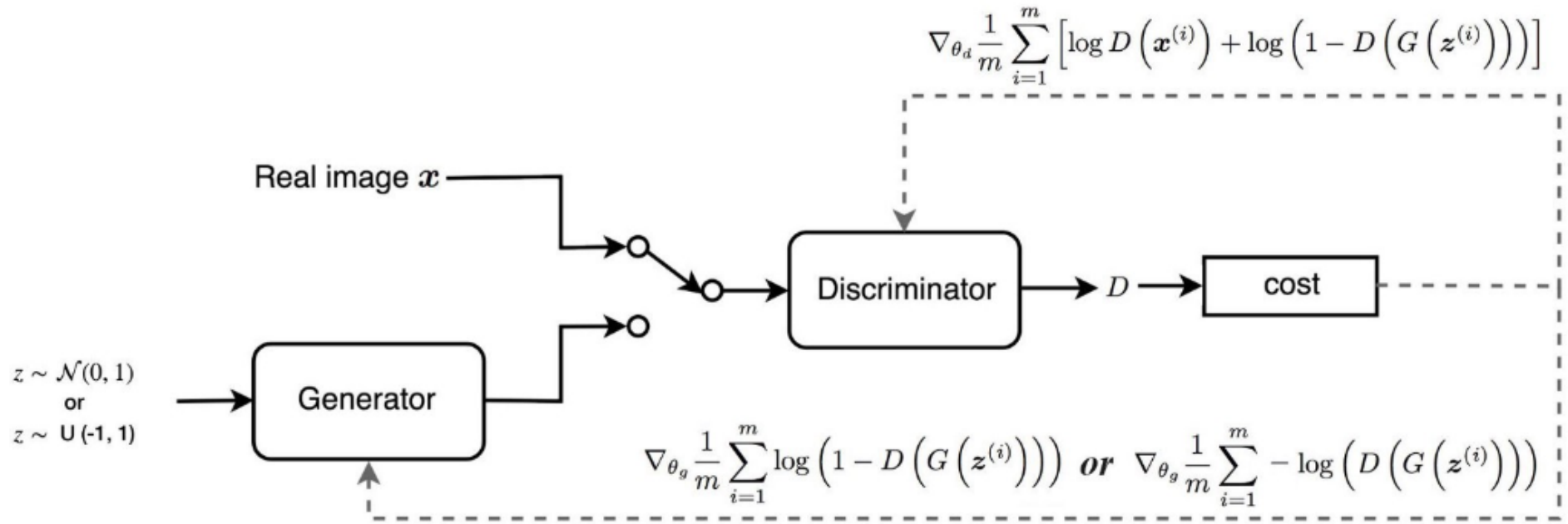
- ‘G’ samples random noise ‘z’ from a uniform / normal (Gaussian) distribution
- ‘G’ generates a fake sample with a single forward pass: $x = G(\Theta_G, z)$
- ‘D’ separates (as a binary classifier) fake from real samples: $y = D(\Theta_D, x)$
- Optimal solution: Nash Equilibrium where y = 0.5 probability (fake = real)

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

Generative Adversarial Networks

Adversarial Training

- ‘G’ and ‘D’ can be trained using backpropagation (in case of NNs)
- ‘G’ tries to **decrease** the correct classification of ‘D’ on the fake data
- ‘D’ tries to **increase** the correct classification for fake and real data



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

Generative Adversarial Networks

Adversarial Training

2014, “Generative Adversarial Networks”, Ian J. Goodfellow et al.

Algorithm 1 Minibatch stochastic gradient descent training of generative adversarial nets. The number of steps to apply to the discriminator, k, is a hyperparameter. We used $k = 1$, the least expensive option, in our experiments.

for number of training iterations **do**

  for k steps **do**

  - Sample minibatch of m noise samples $\{\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(m)}\}$ from noise prior $p_g(\boldsymbol{z})$.
  - Sample minibatch of m examples $\{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(m)}\}$ from data generating distribution $p_{\text{data}}(\boldsymbol{x})$.
  - Update the discriminator by ascending its stochastic gradient:

$$\nabla_{\theta_d} \frac{1}{m} \sum_{i=1}^{m} \left[\log D\left(\boldsymbol{x}^{(i)}\right) + \log\left(1 - D\left(G\left(\boldsymbol{z}^{(i)}\right)\right)\right) \right].$$

  end for

  - Sample minibatch of m noise samples $\{\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(m)}\}$ from noise prior $p_g(\boldsymbol{z})$.
  - Update the generator by descending its stochastic gradient:

$$\nabla_{\theta_g} \frac{1}{m} \sum_{i=1}^{m} \log\left(1 - D\left(G\left(\boldsymbol{z}^{(i)}\right)\right)\right).$$

end for

The gradient-based updates can use any standard gradient-based learning rule.

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

Generative Adversarial Networks

Why GANs?

- First decent generative model that produces promising results
- Ability to approximate the probability distribution of input data through training
- Ability to create huge amounts of realistic data (any digital media)
- Instead of using data augmentation we can use ‘G' to generate realistic samples
- Using it with unsupervised algorithms that need unlabeled data samples
- Ability to use convolution layers, autoencoders, regularization techniques
- ’G’ is not trained on the real dataset as it gets only random input noise

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

Generative Adversarial Networks

Famous GANs variations

- Bidirectional GAN
- Conditional GAN
- Context-Conditional GAN
- CycleGAN
- Deep Convolutional GAN
- DualGAN
- InfoGAN
- Pix2Pix
- Semi-Supervised GAN
- Super-Resolution GAN

400+ named GANs variations exist today (GAN Zoo: https://bit.ly/2sxGUHl)

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

Generative Adversarial Networks

Applications (Image Synthesis)

2017, “Progressive Growing of GANs for Improved Quality, Stability, and Variation”, Tero Karras et al

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

Generative Adversarial Networks

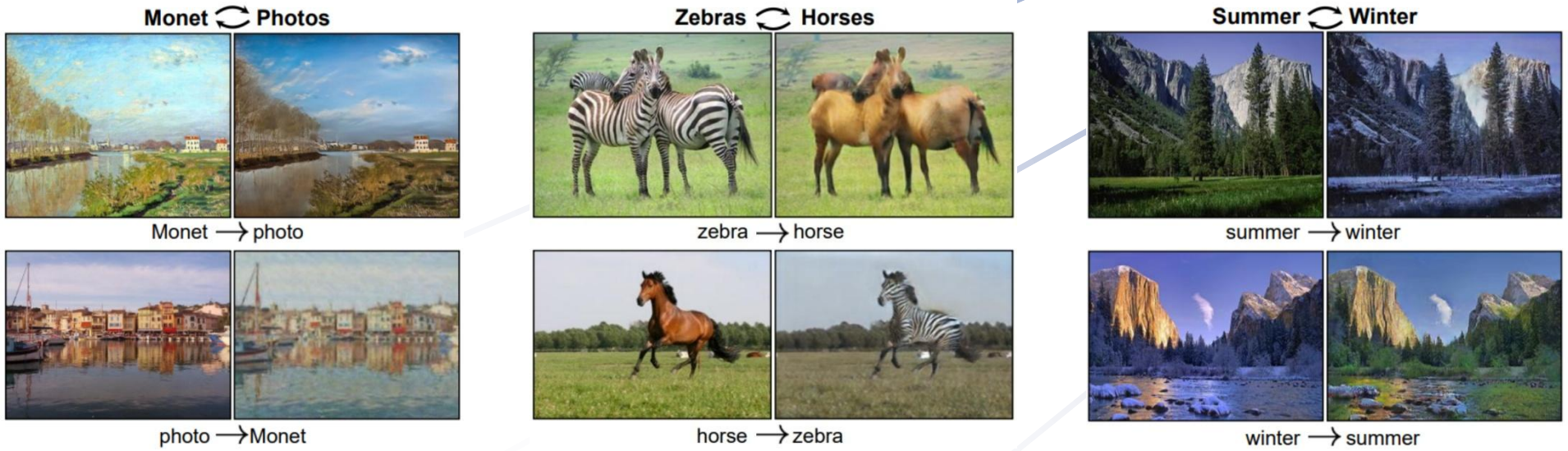
Applications (Image Synthesis)

2017, “Towards the Automatic Anime Characters Creation with Generative Adversarial Networks”, Yanghua Jin et al

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

Generative Adversarial Networks

Applications (Style Transfer)

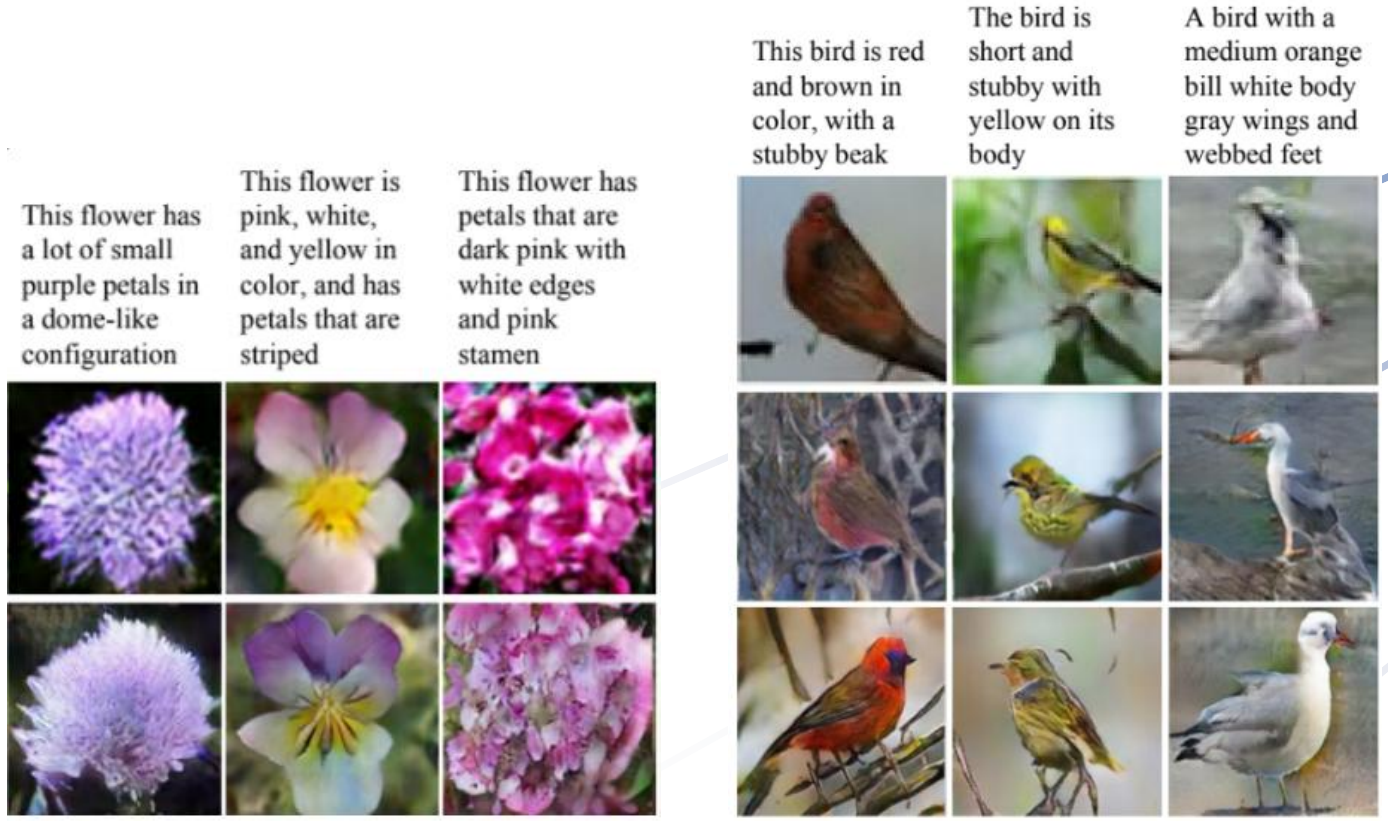


2017, “Image to Image Translation using Cycle-Consistent Adversarial Neural Networks”, Jun-Yan Zhu et al

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

Generative Adversarial Networks

Applications (Text-to-Image Synthesis)



2017, “StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks”, Han Zhang et al

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

Generative Adversarial Networks

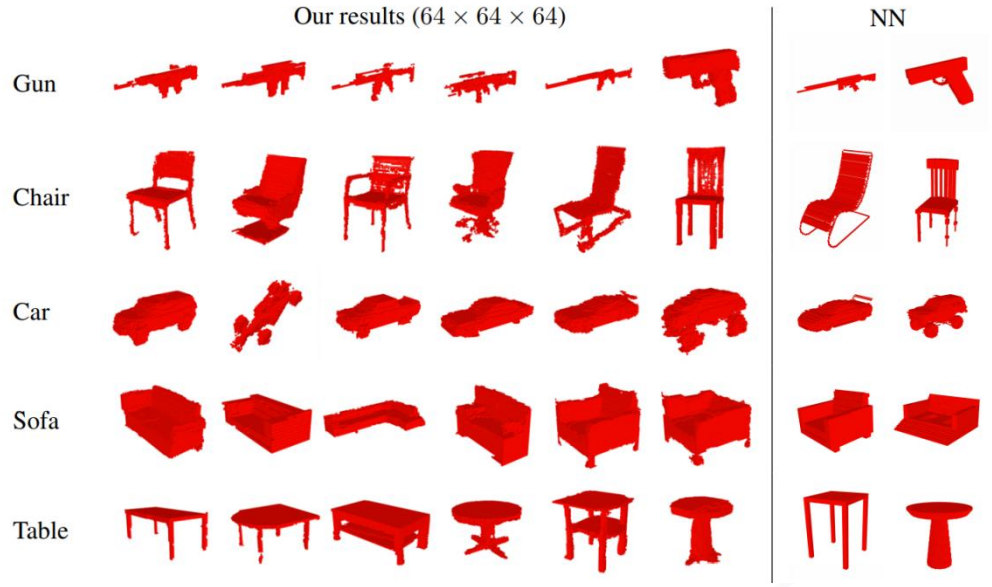
Applications (Sketch-to-Image Synthesis)

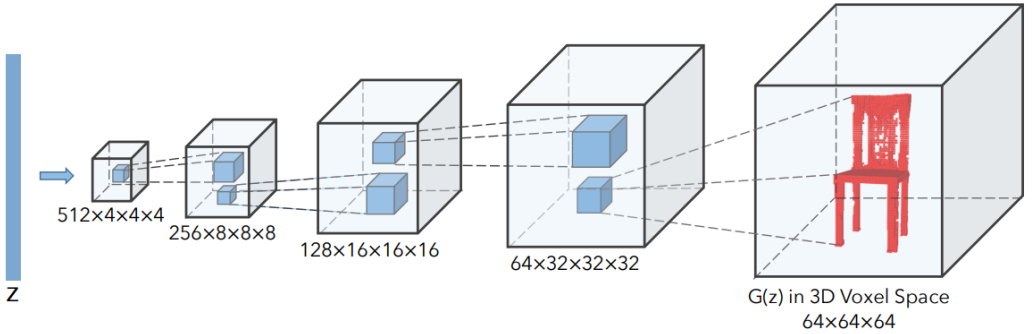
2016, “Image-to-Image Translation with Conditional Adversarial Networks”, Phillip Isola et al

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

Generative Adversarial Networks

Applications (3D Object Synthesis)



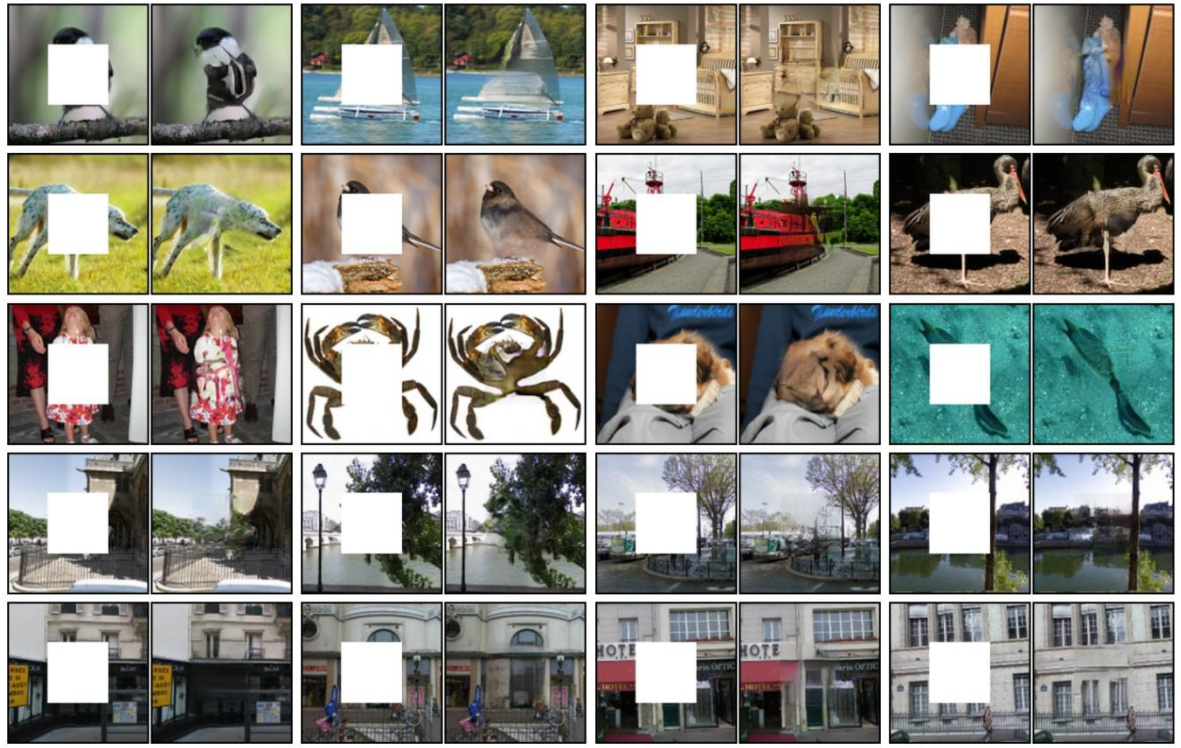


2016, “Learning a Probabilistic Latent Space of Object Shapes via 3D Generative-Adversarial Modeling”, Jiajun Wu et al

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

Generative Adversarial Networks

Applications (Image Gap Filling)

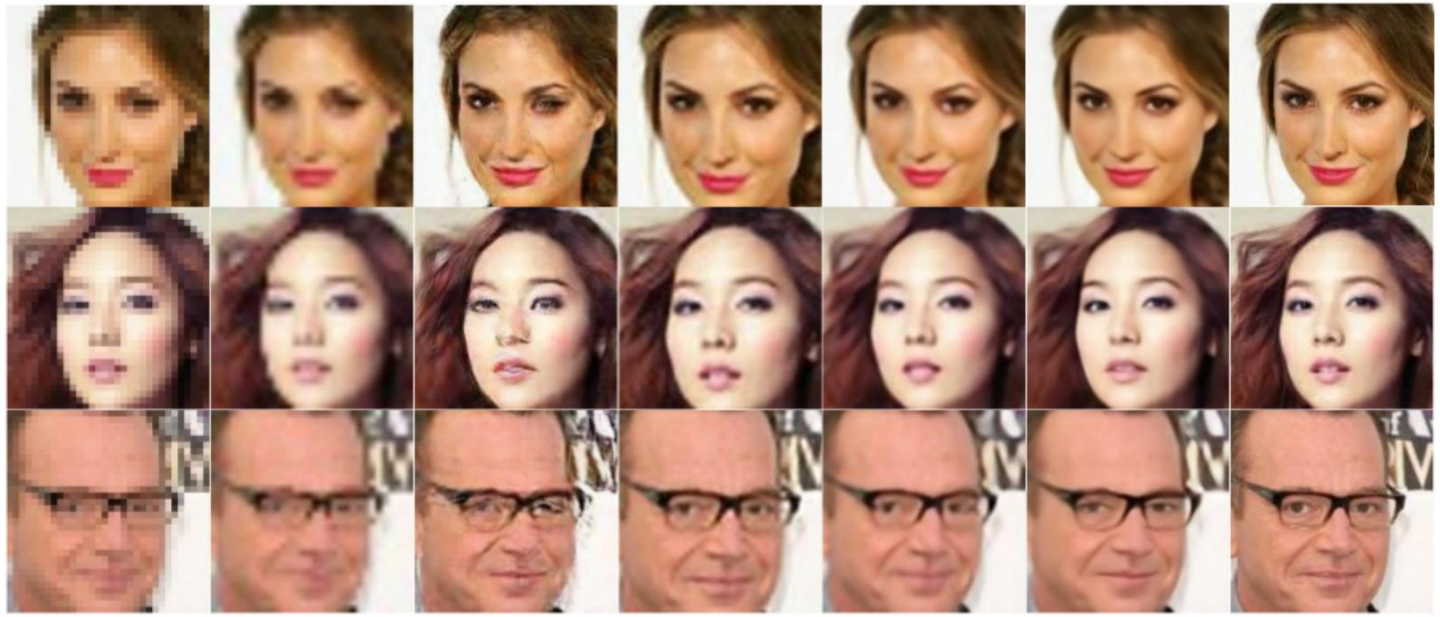


2016, “Context Encoders: Feature Learning by Inpainting”, Deepak Pathak el al

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

Generative Adversarial Networks

Applications (Increase Image Resolution)



2017, “High-Quality Face Image Super-Resolution Using Conditional Generative Adversarial Networks”, Huang Bin et al

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

Generative Adversarial Networks

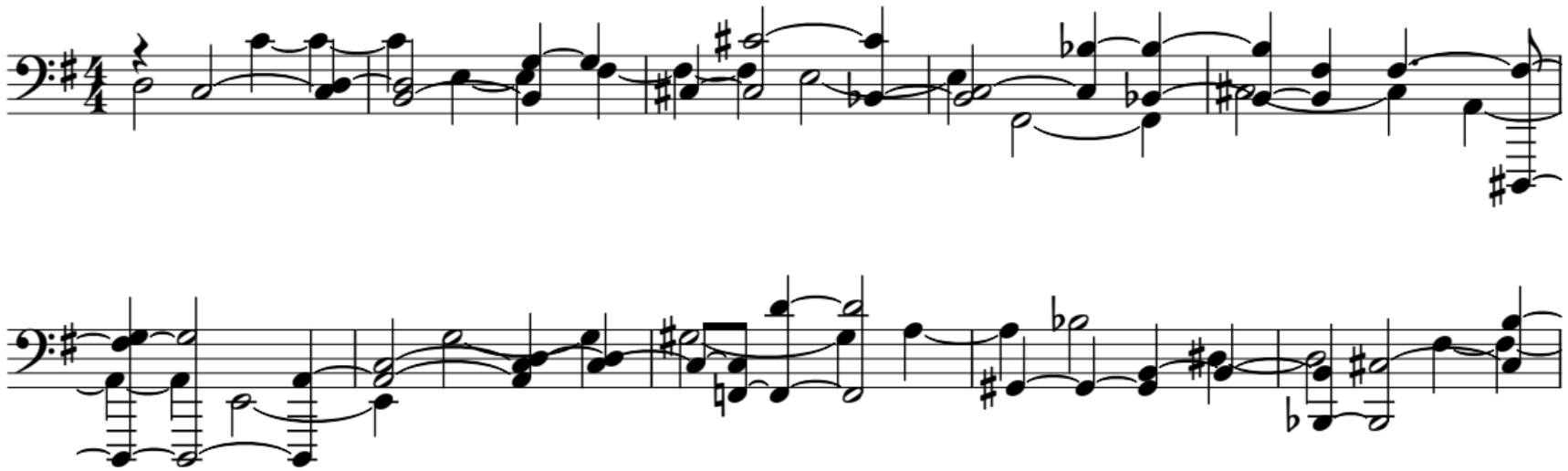
Applications (Progressive Face Aging)

2017, “Age Progression/Regression by Conditional Adversarial Autoencoder”, Zhifei Zhang et al

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

Generative Adversarial Networks

Applications (Music Synthesis)



2016, “C-RNN-GAN: Continuous recurrent neural networks with adversarial training”, Olof Mogren

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

Conclusions

- We have a long journey before we reach human-like high-level Computer Vision
- Many knowledge is created in recent years for Adversarial Examples and GANs
- 400+ named GANs variations exist today (GAN Zoo: https://bit.ly/2sxGUHl)
- It is the era where the Generative Models rise
- Machine Learning needs to be secure for a safe digital intelligent world
- The future of fake digital media is coming and is going to be really dangerous
- A lot of fundamental questions need to be well answered
- Exploit the vulnerabilites of Neural Networks
- Exploit the robustness of Adversarial Examples

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

Recent Research

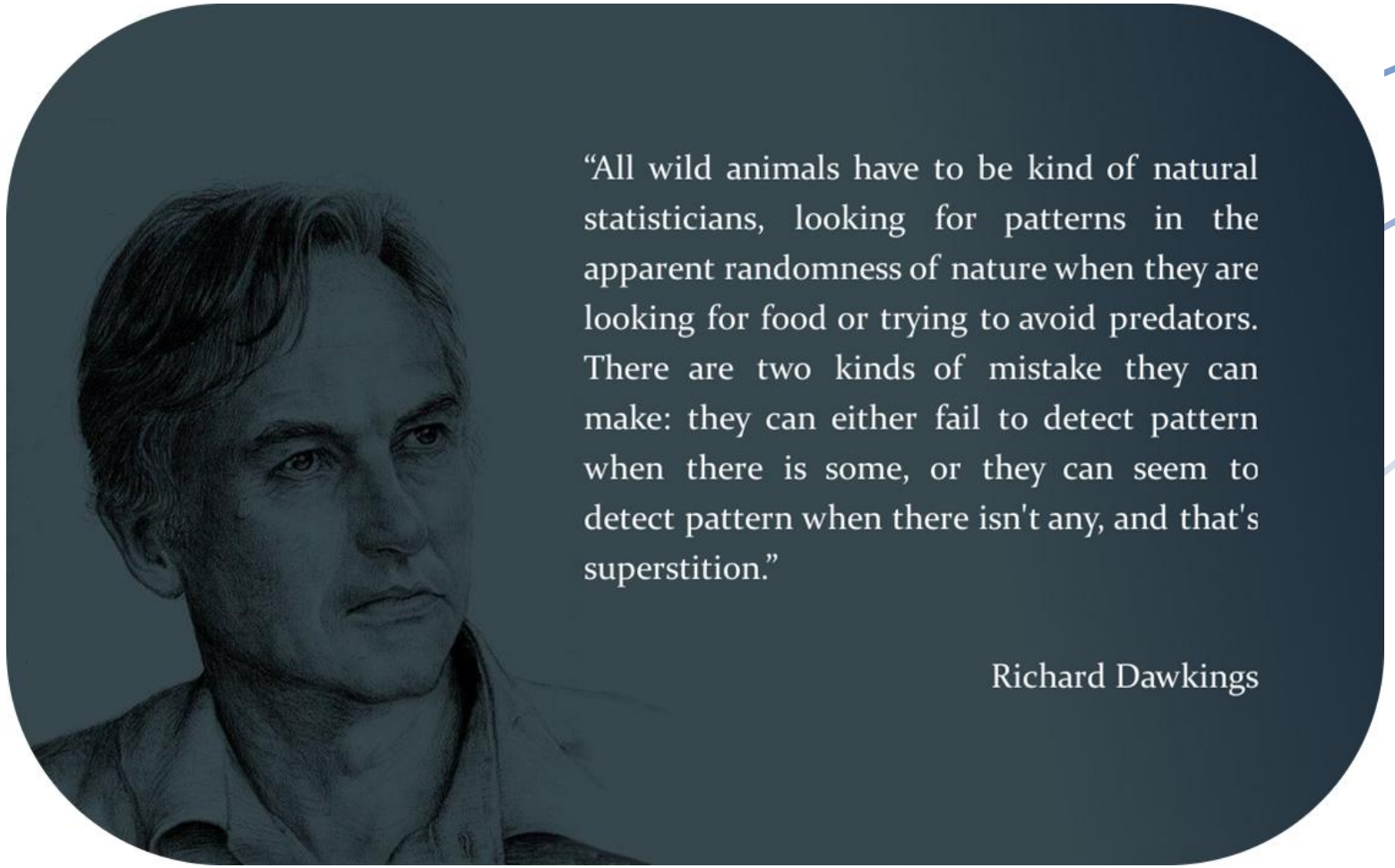
Adversarial Examples - GANs

- Universal adversarial perturbations (Moosavi-Dezfooli et al, 2016)
- Adversarial Machine Learning at Scale (Alexey Kurakin et al, 2016)
- Synthesizing Robust Adversarial Examples (Anish Athalye et al, 2017)
- Delving into Transferable Adversarial Examples and Black-box Attacks (Yanpei Liu et al, 2017)
- Adversarial Examples Are Not Easily Detected: Bypassing Detection Methods (Carlini et al, 2017)
- Adversarial vulnerability for any classifier (Alhussein Fawzi et al, 2018)
- Adversarial Attacks and Defences Competition (Alexey Kurakin et al, 2018)
- Adversarial Logit Pairing (Harini Kannan et al, 2018)
- Evolutionary Generative Adversarial Networks (Chaoyue Wang et al, 2018)
- Generating Adversarial Examples with Adversarial Networks (Chaowei Xiao et al, 2018)
- Robust Conditional Generative Adversarial Networks (Grigorios G. Chrysos et al, 2018)

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

Thank you a lot and have a nice day!

MultiDrone



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