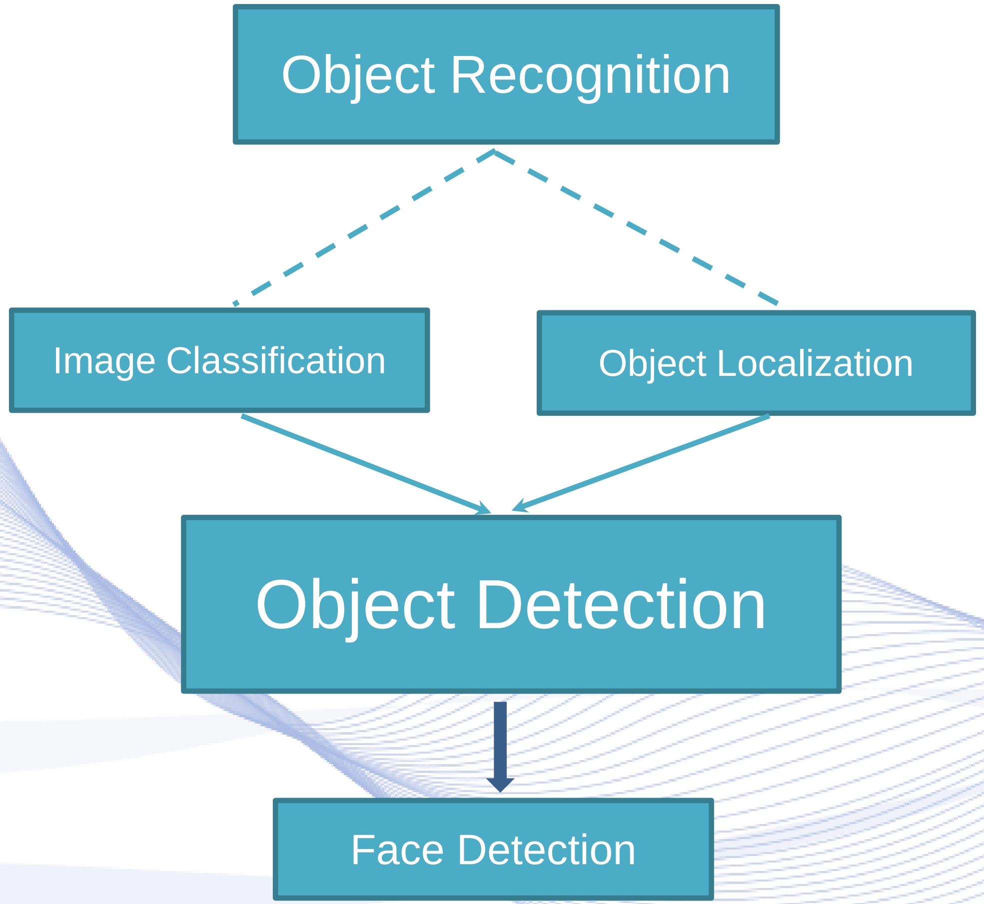
Face and Object De-detection summary

S. Altini, Prof. Ioannis Pitas
Aristotle University of Thessaloniki
pitas@csd.auth.gr
www.aiia.csd.auth.gr
Version 1.0

Face and Object De-detection

- **Object detection and de-detection**
- Adversarial Attacks
  - Threat Model
  - Perturbation
  - Attack Methods
  - Attack Scenarios
  - Defense Methods
  - Benchmarking
- Face Detection Obfuscation

Object Detection



- **Object recognition**: achieve identification and location of objects within an image or video sequence with a given degree of confidence.
- **Main tasks:**
  - **Image classification:** image processing and class label assignment.
  - **Object localization:** processing of bounding box drawing counting over objects present in the screen; tracking object's location precisely; labelling them accurately according to their class.

Face Detection

Face detection (object detection sub-category):

- prime and substantive step for face recognition;
- detection and location of faces on a given image;
- tighten in rectangular bounding boxes the output;
- parameterization of each box in four coordinate around faces, presented in input.

[1]. Face Detection under different scale, pose, occlusion, expression, makeup, and illumination.

[BOS2018]. A.-J. Bose, P. Aarabi, "Adversarial Attacks on Face Detectors Using Neural Net Based Constrained Optimization", *IEEE, 20th International Workshop on Multimedia Signal Processing (MMSP),* 2018, pp. 1-6. https://tspace.library.utoronto.ca/handle/1807/91439

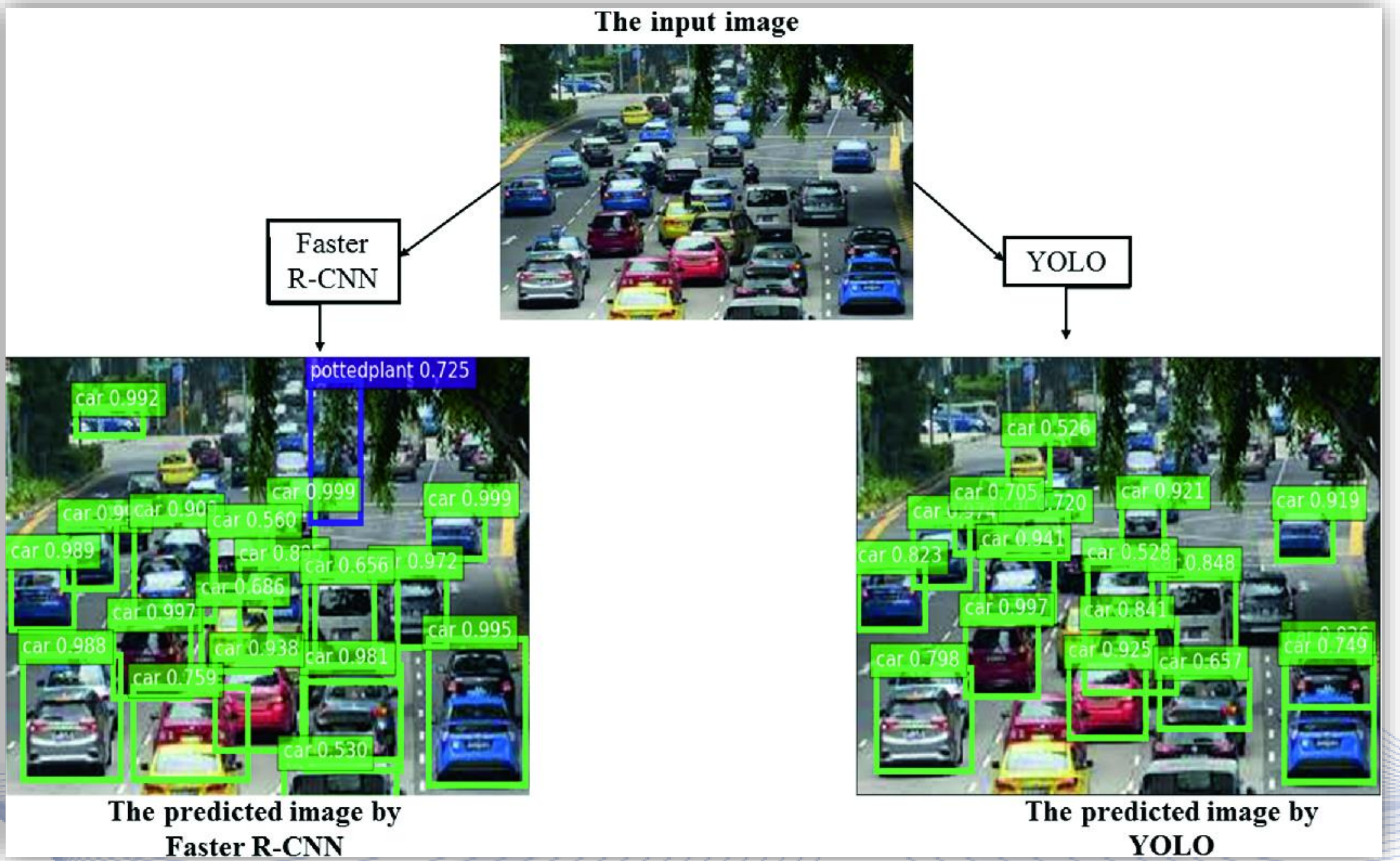
Object Detection Algorithms

Object and Face detection process involves algorithms to conduct object detection, including CNNs:

- **R-CNN, Fast R-CNN, Faster R-CNN & R-FCNN**: parts of two step region-based detector family;
- **YOLO & SSD:** parts of single step detector family.

- **Algorithms, such as SSD and R-FCNN developed to find occurrences fastly.**

Two Step Detection



[4]. YOLO & Faster R-CNN predictions.

[DRD2020]. K. Drid, M. Allaoui, M.L. Kherfi, "Object Detector Combination for Increasing Accuracy and Detecting More Overlapping Objects", in A. El Moataz, D. Mammass, A. Mansouri, F. Nouboud (eds), *Image and Signal Processing*, ICISP, vol 12119, 2020. https://link.springer.com/chapter/10.1007/978-3-030-51935-3_31

Object De-detection

***Object de-detection (detection obfuscation)*:**

- Given a trained ML model $\hat{\mathbf{y}} = \boldsymbol{f}(\boldsymbol{x}; \boldsymbol{\theta})$, performing object detection;
- Perturb a test sample instance $\boldsymbol{x}_p = \boldsymbol{x} + \boldsymbol{p}$, so that the ML model does not perform object detection properly, i.e.: ideally $\hat{\mathbf{y}}_p$ is very different from $\hat{\mathbf{y}}$.
- Typically $\boldsymbol{x}_p$ is 'similar' to $\boldsymbol{x}$ (it has imperceptible differences):

$\boldsymbol{x}_p$ has the same probability distribution with $\boldsymbol{x}$ and/or

$\|\boldsymbol{x} - \boldsymbol{x}_p\|$ is small.

Adversarial Attacks

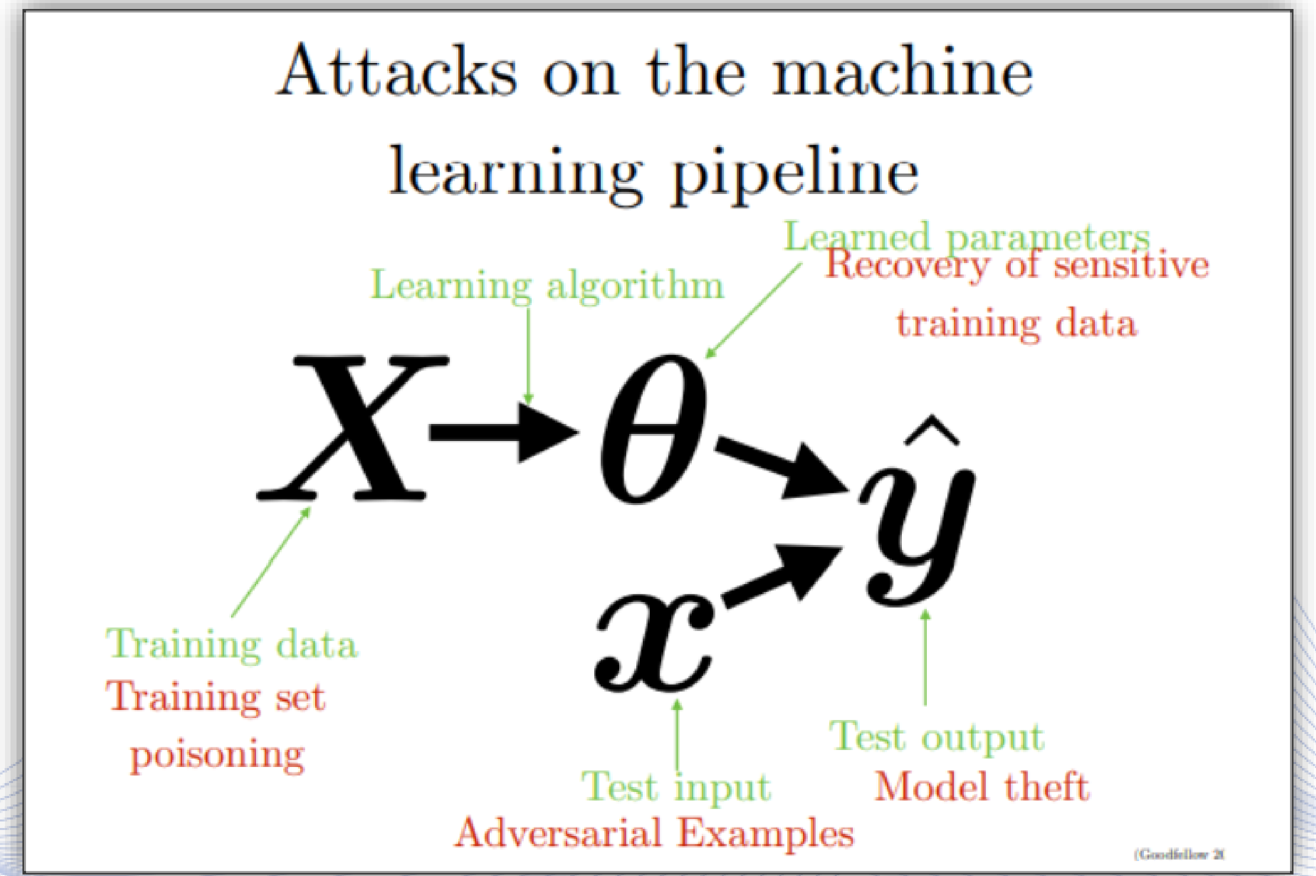
Terminology

- **Adversary:** an attacker who crafts an adversarial example, depending on the scope or the example itself, depending on the case study.
- ***Adversarial training:*** process that uses adversarial images along with original; explicitly training of a model on adversarial examples, enhancing its robustness to attacks; reduces test error on clean inputs.
- ***Adversarial learning:*** any situation during model training to a worst case scenario, provided by another model; optimization of MLs, using adversarial examples to improve ML algorithms; enhance models' robustness; improve recognition despite the presence of domain shift or dataset bias.

Face and Object De-detection

- Object detection and de-detection
- **Adversarial Attacks**
  - Threat Model
  - Perturbation
  - Attack Methods
  - Attack Scenarios
  - Defense Methods
  - Benchmarking
- Face Detection Obfuscation

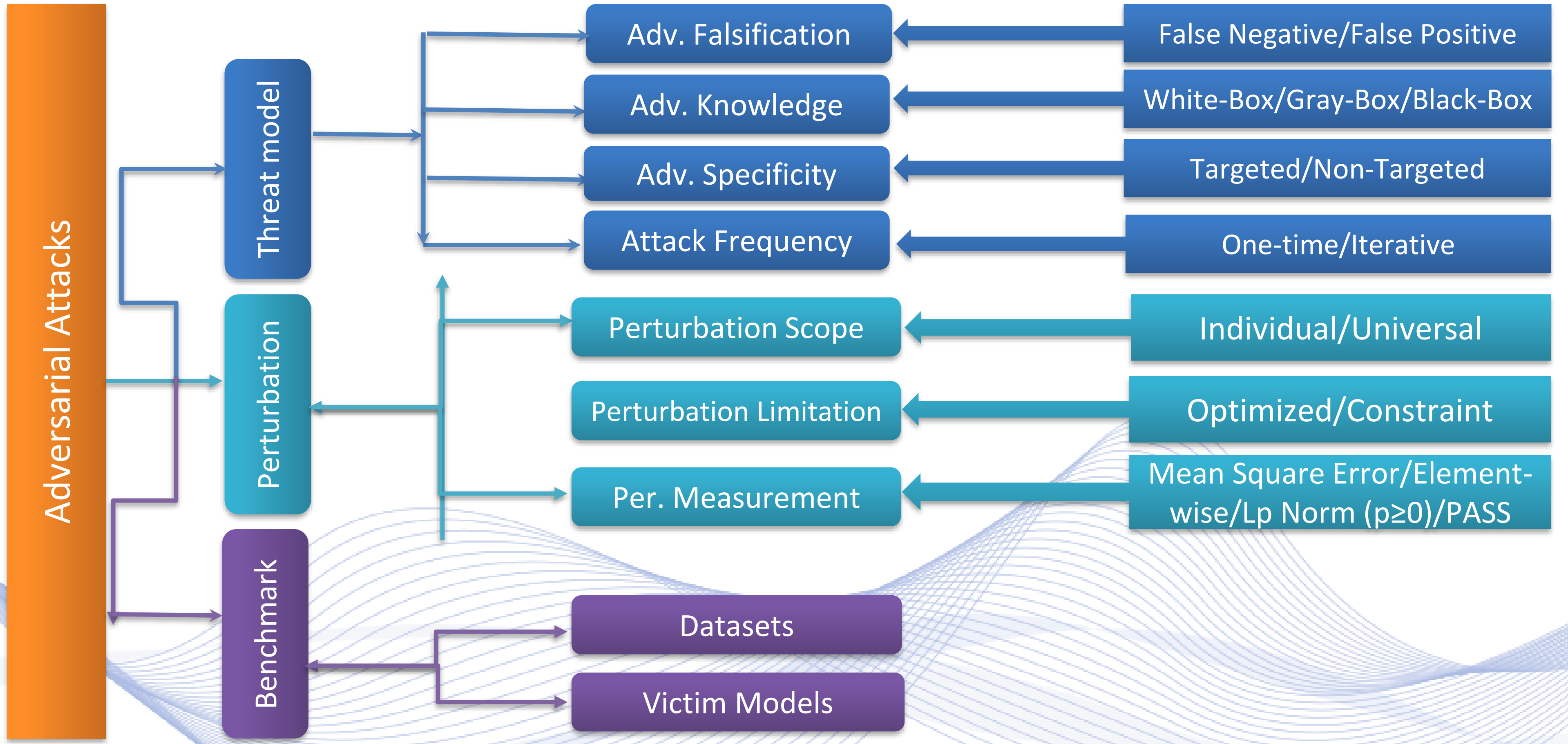
Adversarial Attacks



[5]. Attacks on the machine learning pipeline.

[DHA2019]. M. Dhaouadi, "A survey about adversarial learning", 2019. https://www.researchgate.net/publication/338105748_A_SURVEY_ABOUT_ADVERSARIAL_LEARNING

Adversarial Attacks



[6]. Taxonomy of Adversarial Attacks.

[YUA2019] X. Yuan, P. He, Q. Zhu, X. Li, "Adversarial Examples: Attacks and Defenses for Deep Learning", *IEEE Trans. Neural Networks Learn. Syst.*, vol. 30(9), pp. 2805-2824, 2019.

Face and Object De-detection

- Object detection and de-detection
- Adversarial Attacks
  - **Threat Model**
  - Perturbation
  - Attack Methods
  - Attack Scenarios
  - Defense Methods
  - Benchmarking
- Face Detection Obfuscation

Threat Model

- **Adversarial falsification**
- Adversary knowledge
- Adversarial specificity
- Attack frequency

Adversarial falsification

· **False positive attacks:** generation of negative result that indicates vulnerability as positive one (Type I Error);

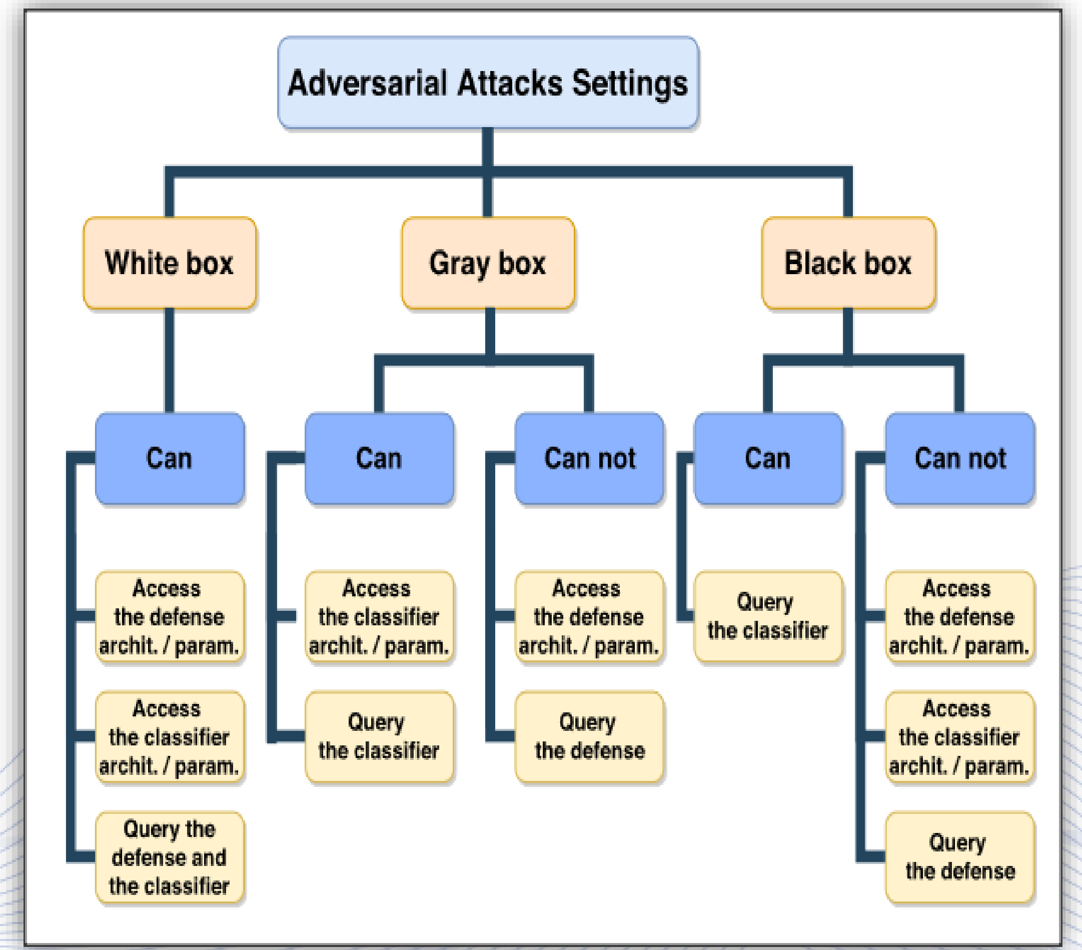
· **False negative attacks:** generation of positive result, misclassified as negative (Type II Error).

[7]. https://shuzhanfan.github.io/2018/02/model-evaluation-metrics/

Threat Model

- Adversarial falsification
- **Adversary knowledge**
- Adversarial specificity
- Attack frequency

Adversary knowledge

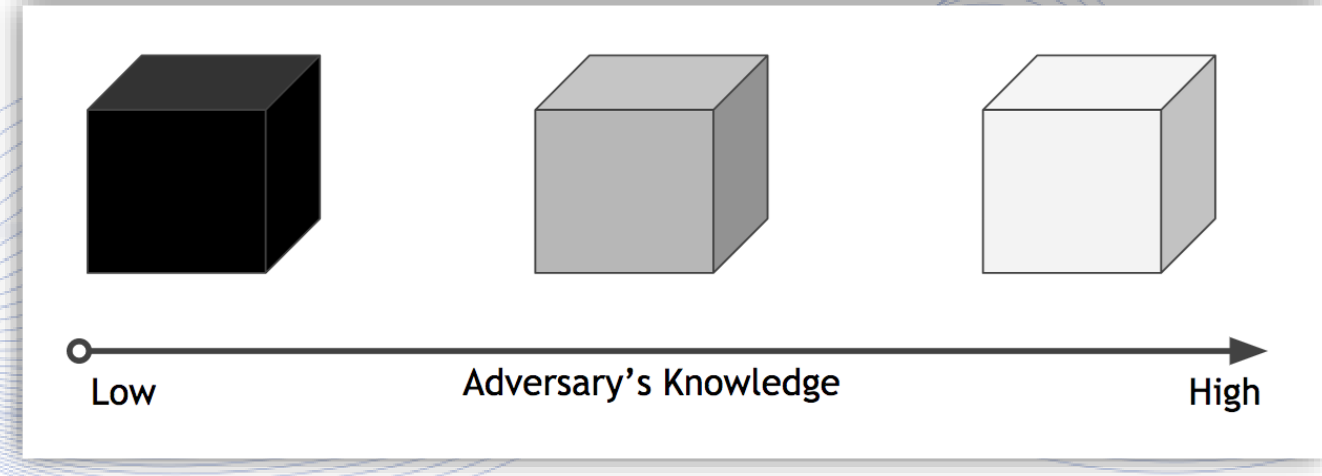


[8]: White, Gray and Black box attacks settings

[BAK2019] Y. Bakhti, S.-A. Fezza, W. Hamidouche, O. Déforges, “DDSA: A Defense against Adversarial Attacks using Deep Denoising Sparse Autoencoder”, *IEEE Access*, vol. 7, pp.160397-160407, 2019.

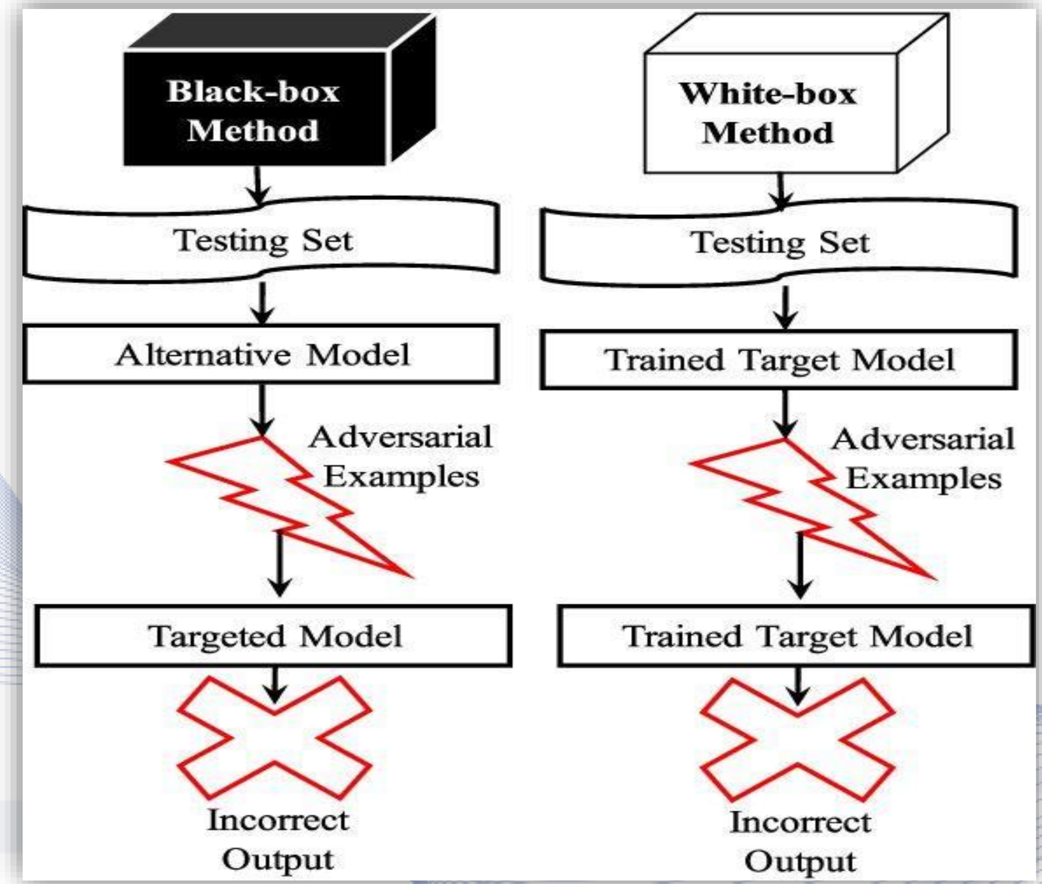
Adversary Knowledge

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



[9]. Adversary's knowledge. https://secml.github.io/class1/

Adversary Knowledge



[10]. Black-box vs. White-box

[ALS2020] B. Alshemali, J. Kalita, "Improving the Reliability of Deep Neural Networks in NLP: A Review", *Knowledge-Based Systems*, vol. 191, pp. 105210-105229, 2020.

Black-box vs. White-box adversarial methods

- **Black-box settings:** adversary can only query and observe the targeted model's output; generates adversarial examples using an alternative model.
- **White-box settings:** adversary uses the targeted model, which has full access to generate adversarial examples.

Artificial Intelligence & Information Analysis Lab

Threat model

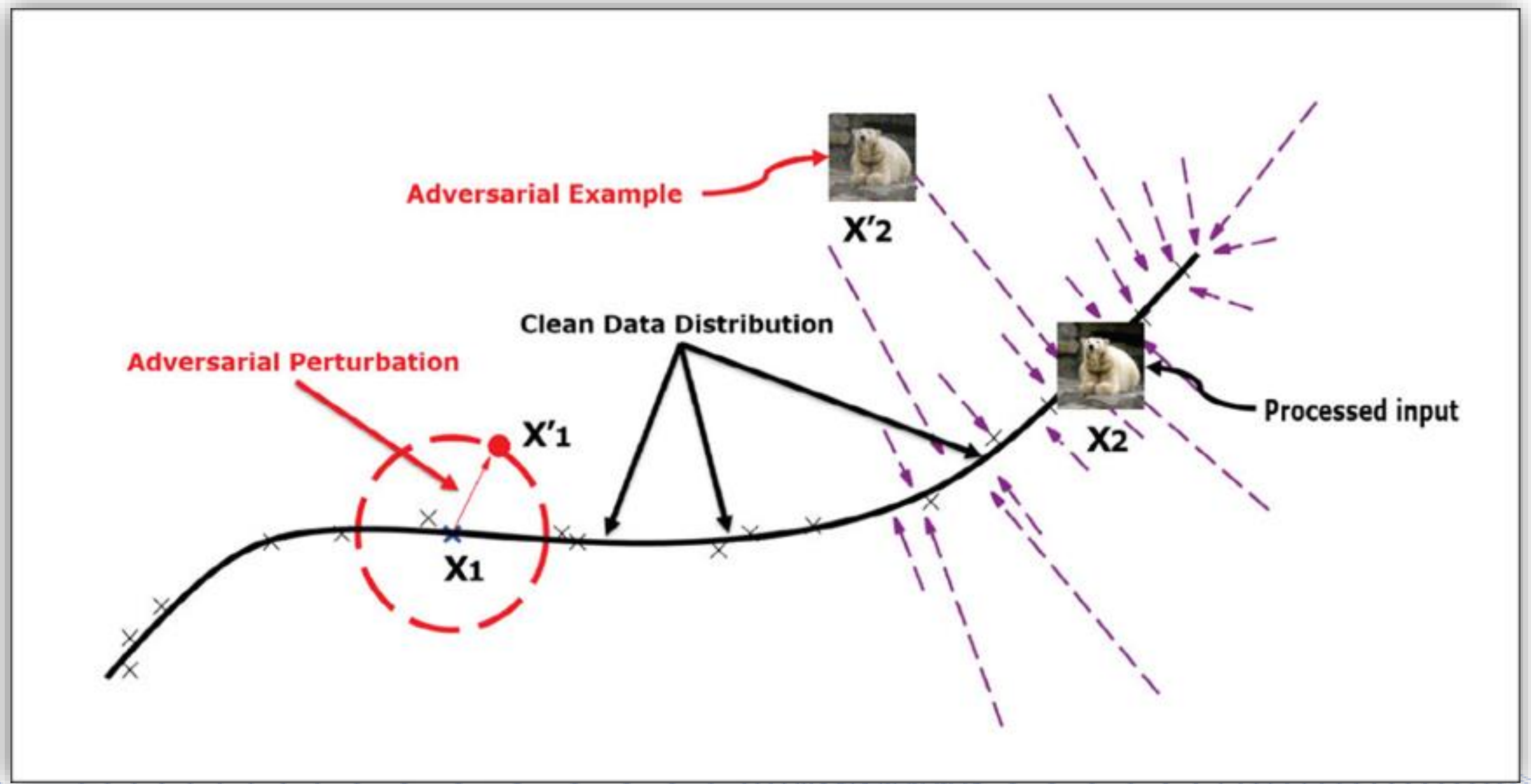
- Adversarial falsification
- Adversary’s knowledge
- **Adversarial specificity**
- Attack frequency

Adversarial specificity

Adversarial Specificity: allow specific intrusion/disruption; generate approaches (extended BIM, ZOO) applied to targeted /non-targeted attacks;

- **Targeted attack:** Find an input that is misclassified as a specific label (e.g. adversarial samples aim to target a specific target value).
- **Non-targeted attack:** Find an input that is misclassified in a label just different than the ground truth (e.g. indiscriminate attacks, where the samples do not target a specific target value).

Adversarial specificity



[12]. Data distribution over the manifold.

[BAK2019]. Y. Bakhti, S.-A. Fezza, W. Hamidouche, O. Déforges, "DDSA: A Defense against Adversarial Attacks using Deep Denoising Sparse Autoencoder", *IEEE Access*, vol. 7, pp.160397-160407, 2019. https://hal-univ-rennes1.archives-ouvertes.fr/hal-02349625/document

Threat Model

- Adversarial falsification
- Adversary’s knowledge
- Adversarial specificity
- **Attack frequency**

Attack Frequency

One-time attacks: take only one time to optimize adversarial examples; the only feasible choice for some computational tasks (reinforcement learning).

- **Iterative attacks:** take multiple times to update the adversarial examples; need more computational time for generation; need more queries with victim classifier to perform better adversarial examples.

Face and Object De-detection

- Object detection and de-detection
- Adversarial Attacks
  - Threat Model
  - **Perturbation**
  - Attack Methods
  - Attack Scenarios
  - Defense Methods
  - Benchmarking
- Face Detection Obfuscation

Aspects of Perturbation

- **Perturbation scope**
- Perturbation limitation
- Perturbation measurement

Perturbation Scope

Types according to the scope of perturbation implementation:

- **Individual attacks:** differential perturbations for every original input.
- **Universal attacks:** create universal perturbation for the entire dataset; applied to all original input data; generated according to the given input images; create adversaries in real-world applications; perturbations, not required to be reformed, when the input samples are changed.

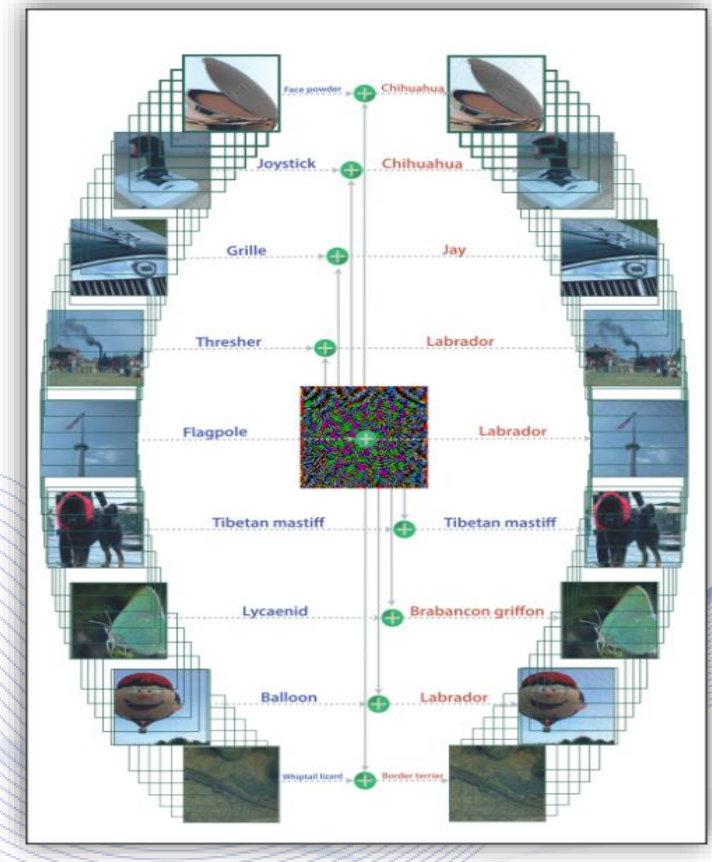
Perturbation Scope

Universal Perturbation:

- used for image classification tasks;
- untargeted attack with no preference over the (incorrect) output class;
- imperceptible for human.

- **Aim:** fool CNNs on image set;

- **Universality:** opposed to transferability property, referring to an “image-agnostic”, due to the property of a perturbation.



[12]. Universal perturbation that fools DNNs on images. Original labeled natural images (left); Universal perturbation (center); perturbed images with wrong labels (right). [YUA2019] X. Yuan, P. He, Q. Zhu, X. Li, “Adversarial Examples: Attacks and Defenses for Deep Learning”, *IEEE Trans. Neural Networks Learn. Syst.*, vol. 30(9), pp. 2805-2824, 2019.

Perturbation Scope

Adversarial Patches: White/*Black-Box attacks*

- **Patch-based (white-box attacks)**: patch cause classification errors; adversary have access to network architecture/parameters. **Aim:** push the adversarial examples into specific class; amplification factor leads to underfitting.
- **Patch-based (black-box attacks):** overlapping patches causing misclassification; access the input and predicted output of model (self-driving cars). **Aim:** break DNNs.

[13]. Detection results of images of different classes with the adversarial patch. Original image (first line) & detection result after adding the adversarial patch (second line). [WAN2021] Y. Wang, H. Lv, X. Kuang, G.Zhao, Y. Tan, Q. Zhang, J. Hu, "Towards a physical-world adversarial patch for blinding object detection models", *Inform. Sci.*, vol. 556, pp. 459-471, 2021.

Artificial Intelligence & Information Analysis Lab

Aspects of Perturbation

- Perturbation scope
- **Perturbation limitation**
- Perturbation measurement

Perturbation Limitation

- **Optimized Perturbation:** the goal of the optimization problem. **Aim:** the perturbed input to be close enough to the original image that a human can not distinguish one image from the other.
- **Constraint Perturbation:** the set that constraint the optimization problem.

Aspects of Perturbation

- Perturbation scope
- Perturbation limitation
- **Perturbation measurement**

Perturbation Measurement

Perturbation measurement: Universal Adversarial Perturbations (UAP) come in targeted or untargeted forms, depending on the attacker's objective and robust models that are limited to human invariants.

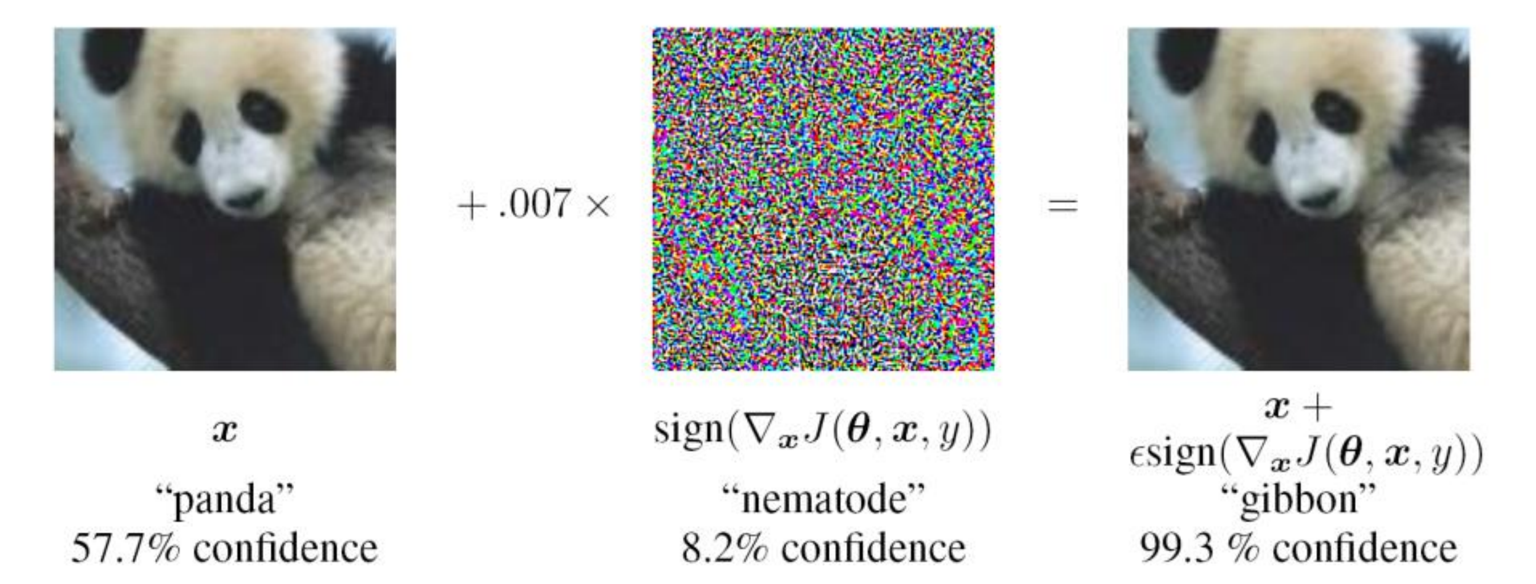
- **MSE; RMSE; NRMSE;**
- ***lp* norm** (p>0)
- **Element-wise**
- **Psychometric perceptual adversarial similarity score (PASS)**:

Face and Object De-detection

- Object detection and de-detection
- Adversarial Attacks
  - Threat Model
  - Perturbation
  - **Attack Methods**
  - Attack Scenarios
  - Defense Methods
  - Benchmarking
- Face Detection Obfuscation

Adversarial Attack Methods

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



[14]. Fast Gradient Methods.
[GOO2014]. I.-J. Goodfellow, J. Shlens, C, Szegedy, "Explaining and harnessing adversarial examples", *arXiv:1412.6572*, 2014. https://arxiv.org/abs/1412.6572

Properties of adversarial examples

- **Transferability:** common property that deceive models other than the one used to create it; leading to adaptive black-box attacks; not restricted to DNNs, as attacks exploit transferability against different ML algorithms, e.g. LR, SVM, including commercial machine learning classification systems (Google ML).
- **Perceivability:** small perturbations to image pixels even though cannot be easily detected by humans, can fool DNNs.
- **Semantic dependencies:** usually, small perturbations can not cause changes to the image semantics, because just change some individual pixels that is impossible to turn an image of a cat into a car.

Adversarial Attack Methods
White-Box attacks

- **FGSM: Fast Gradient Sign Method**
  - **Non-targeted:** One-shot method with gradient ascent

$$\mathbf{x}_p = \mathbf{x} + \varepsilon \cdot sign\left(\nabla_x l_f(\mathbf{x}, y_{true})\right)$$

  - **Targeted:** One-shot method with gradient descent

$$\mathbf{x}_p = \mathbf{x} - \varepsilon \cdot sign\left(\nabla_x l_f(\mathbf{x}, t)\right)$$

Adversarial Attack Methods
White-Box attacks

- **I-FGSM: Iterative Fast Gradient Sign Method**
  - **Non-targeted:** Iterative method with gradient ascent

$$\mathbf{x}_{p_0} = \mathbf{x},$$

$$\mathbf{x}_{p_{i+1}} = clip_{[0,1]}\left(clip_{[x-\varepsilon,\, x+\varepsilon]}\left(\mathbf{x}_{p_i} + \alpha \cdot sign\left(\nabla_x l_f\left(\mathbf{x}_{p_i}, y_{true}\right)\right)\right)\right)$$

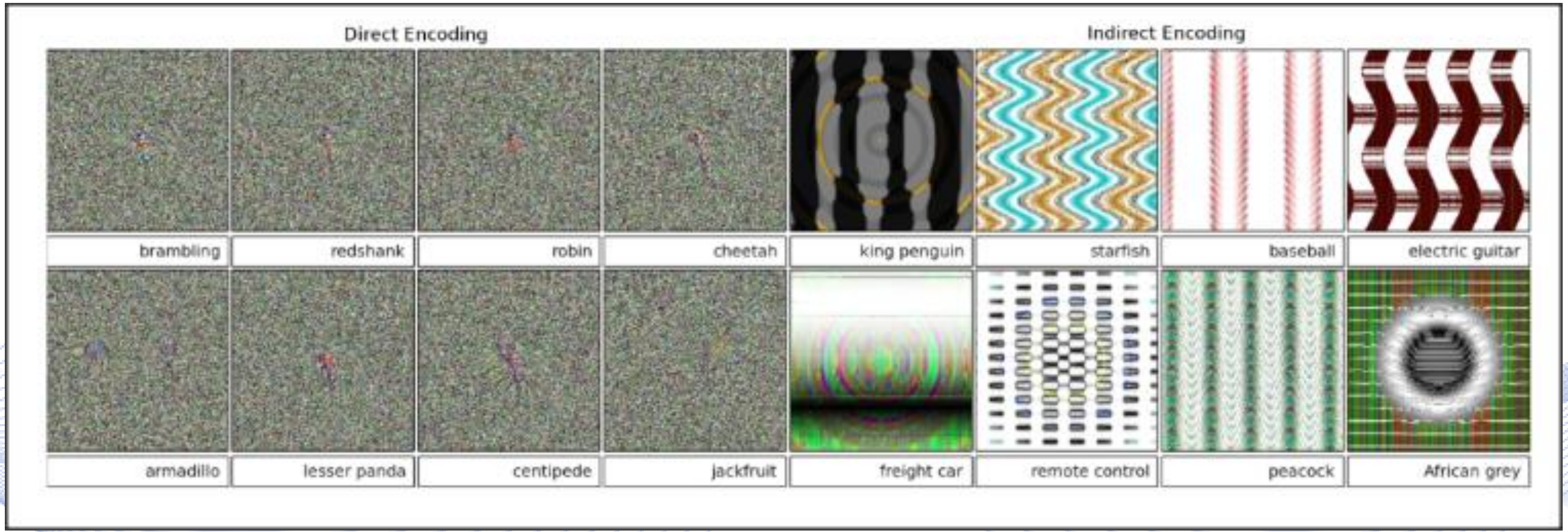
  - **Targeted:** Iterative method with gradient descent

$$\mathbf{x}_{p_0} = \mathbf{x},$$

$$\mathbf{x}_{p_{i+1}} = clip_{[0,1]}\left(clip_{[x-\varepsilon,\, x+\varepsilon]}\left(\mathbf{x}_{p_i} - \alpha \cdot sign\left(\nabla_x l_f\left(\mathbf{x}_{p_i}, t\right)\right)\right)\right)$$

Adversarial Attack Methods
White-Box attacks

• **CPPN EA Fool:** false positive attack; compositional pattern producing network-encoded EA (CPPN EA); classification of DNNs with high confidence (≥99.6%); unidentifiable to humans. **Aim:** locate critical features to change outputs of DNN.



[15].Unrecognizable examples to humans, but DNNs classify them to a class with high certainty [YUA2019] X. Yuan, P. He, Q. Zhu, X. Li, "Adversarial Examples: Attacks and Defenses for Deep Learning", *IEEE Trans. Neural Networks Learn. Syst.*, vol. 30(9), pp. 2805-2824, 2019.

Adversarial Attack Methods
White-Box attacks

- **DFST characteristics:**

- **Effectiveness:** high attack success rate;

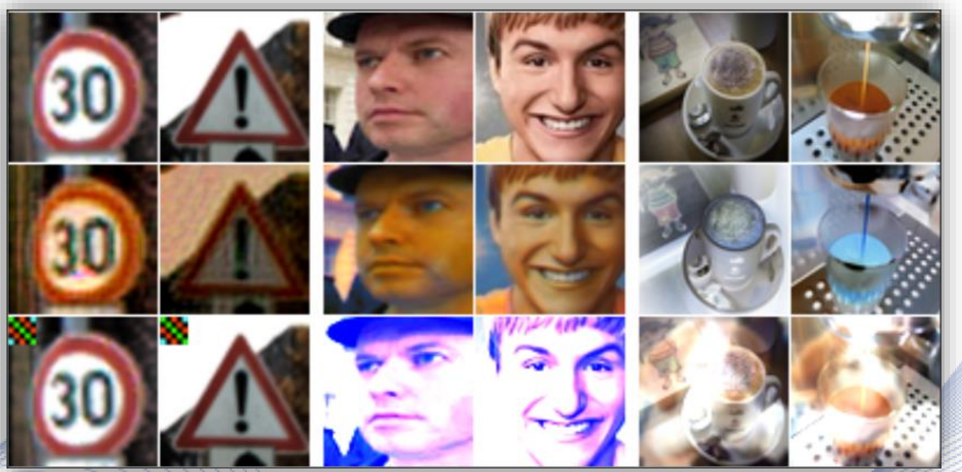
- **Stealthiness:** negligible accuracy degradation on inputs; stamped image incomprehensible to humans.

- **Controllability:** resource consumption, increases the difficulty to detect trojaned models;

- **Robustness:** evasion of trigger features cannot be done by adversarial training of trojaned models;

- **Reliance on deep features:** not depend on simple trigger features to induce misclassification; difficult to detect; state-of-the-art scanners (NC, ABS, ULP) cannot detect trojaned models.

- **Aim:** generate parameters to minimize the maximum adversarial loss.

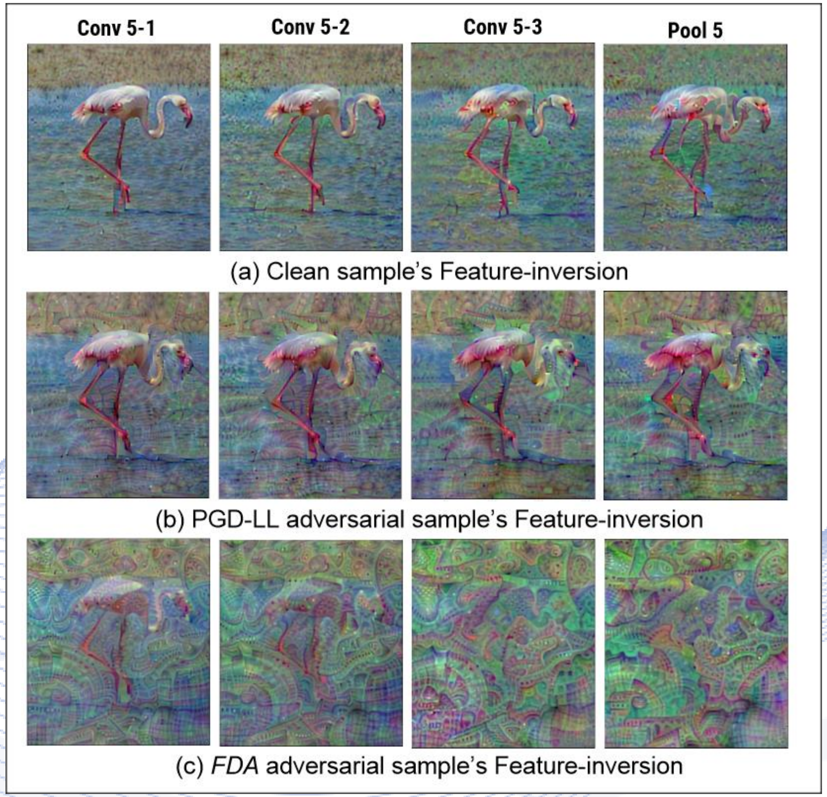


[16]. Samples on GTSRB, VGG-Face and ImageNet: 1st row: before injecting the DFST triggers; 2nd row: after injecting the DFST triggers; 3rd row: after injecting triggers by existing attacks, including patch, Instagram filter and reflection.

[CHE2020] S. Cheng, Y. Liu, S. Ma, X. Zhang, Deep Feature Space Trojan Attack of Neural Networks by Controlled Detoxification. arXiv:2012.11212, 2020.

Adversarial Attack Methods
White-Box attacks

- **FDA (Feature Disruptive Attack) benefits:**
- “flips the predicted label to highly unrelated classes, removing evidence of clean sample’s predicted label”;
- “disrupts feature-representation based tasks (caption generation), even without access to the task-specific network/methodology (effective in gray-box attack setting)”;
- “generates stronger adversaries than other state-of-the-art methods for image classification”.
- **Aim:** generates perturbation, causing disruption of features at each layer of the network.



[17]. Feature Inversion: Layer-by-layer Feature Inversion of clean, PGD-LL-adversarial and FDA-adversarial sample”. [GAN2019] A. Ganeshan, B. Vivek, R. Babu, “FDA: Feature Disruptive Attack”, *IEEE Explore*, 2019.

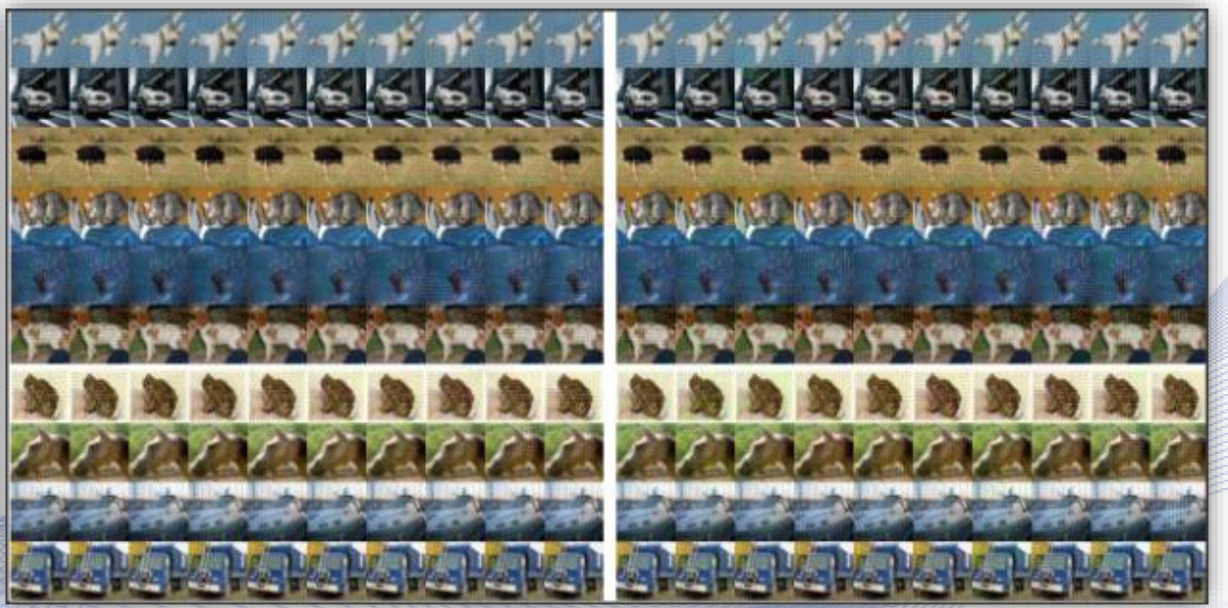
Adversarial Attack Methods

Black-Box attacks

- **Zeroth Order Optimization (ZOO):** gradient-based adversarial attack, without model transferring; need the access to the victim DNN; require accurate computation to query; estimate the gradients. **Aim:** generate an adversarial example with 100% attack success rate.
- **One Pixel Attack:** changing one pixel to cause the misclassification; not require the gradients of the DNN; used in non-differential objective functions; differential evolution (DE) for finding the optimum solution. **Aim**: generate adversarial examples, avoiding the measurement of perceivability.

Adversarial Attack Methods
Black-Box attacks

- **ADV-GAN (Attack-Inspired GAN):** fix target classes in training; no need of defense knowledge to perform attacks; achieve high attack success rate under current defenses; adversarial instances appear closer to real instances; improving adversarial training defense methods.
- **Aim:** generate adversarial perturbation with original images as inputs; faster than optimization-based methods at inference time.



(a) Semi-white box attack (b) Black-box attack

[18]. Adversarial examples generated by AdvGAN on CIFAR-10. Image from each class is perturbed to other different classes. On the diagonal, the original images are shown. https://www.ijcai.org/proceedings/2018/0543.pdf

Adversarial Attack Methods

Black-Box attacks

- **AI-GAN (Attack-Inspired GAN):** generates perceptually realistic adversarial examples with different targets; scales to complicated datasets; joint training of adversary; generate perturbations in efficient way; achieve high attack success rates; reduces generation time in various settings.
- **Aim:** generate adversarial attacks with different targets, promoting efficiency; preserve image quality.

[19]. Visualization of Adversarial examples and perturbations generated by AI-GAN on CIFAR-10. Rows: different targeted classes; columns: 10 images from different classes. Original images are shown on the diagonal. Perturbations are amplified for visualization. [BAI2021] T. Bai, J. Zhao, J. Zhu, S. Han, J. Chen, B. Li, “AI-GAN: Attack-Inspired Generation of Adversarial Examples”, arXiv:2002.02196, 2021

Artificial Intelligence & Information Analysis Lab

Adversarial Attack Methods

Black-Box attacks

- **MI-FGSM:** method of fast gradient sign; take advantage from stabilized update directions that keep missing from local maxima during iterations.
- **Aim**: boost the ability of the adversarial attack.

[20]. Original images (left) and adversarial images by gradient-based MI-FGSM attack method on a restricted region (right). [GU2020] Z. Gu, W. Hu, C. Zhang, L. Wang, C. Zhu, Z. Tian, “Restricted Region Based Iterative Gradient Method for Non-Targeted Attack”, *IEEE Access,* vol. 8, pp. 25262-25271, 2020

Artificial Intelligence & Information Analysis Lab

Face and Object De-detection

- Object detection and de-detection
- Adversarial Attacks
  - Threat Model
  - Perturbation
  - Attack Methods
  - **Attack Scenarios**
  - Defense Methods
  - Benchmarking
- Face Detection Obfuscation

Attack scenarios

Main attack scenarios identified by the attack surface:

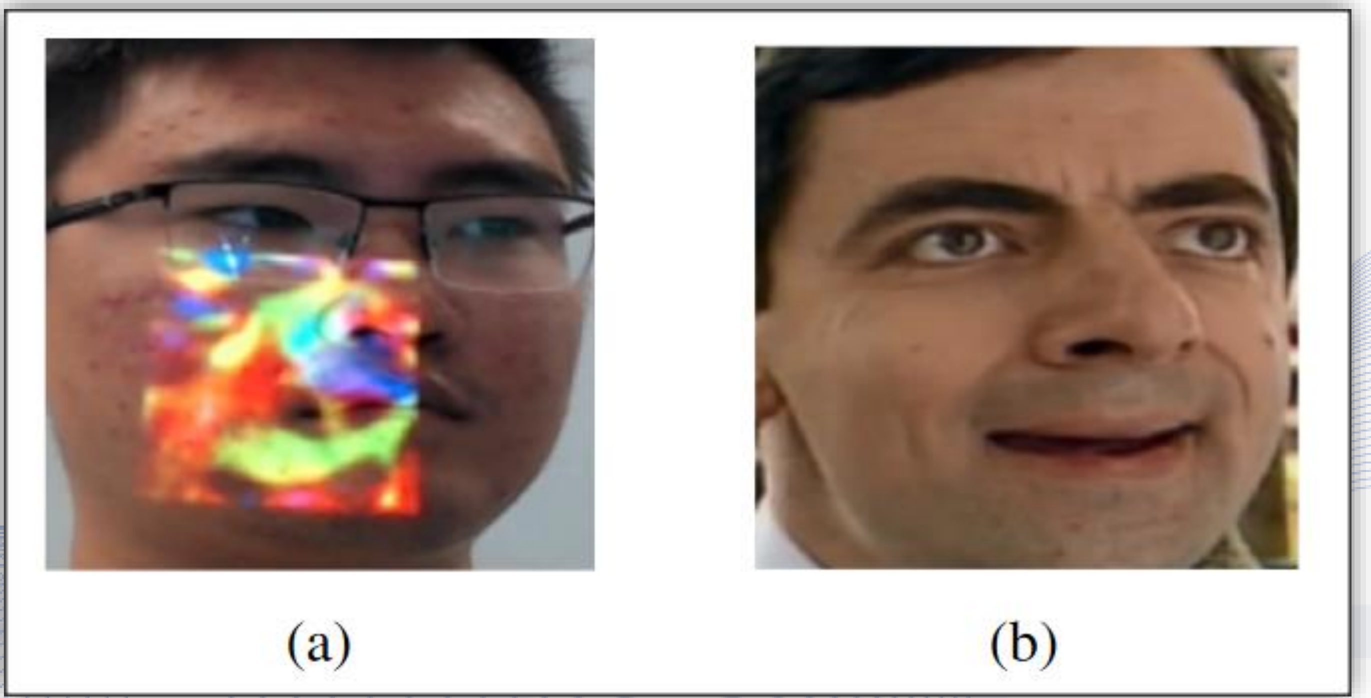
- **Evasion attack:** common type of attack; attacker evade the system by adjusting malicious samples during the testing phase; setting does not influence training data. **Aim:** evade the system by altering samples during the testing phase, but not influence the training data.
- **Poisoning attack:** inject skillfully crafted samples to poison the system in order to compromise the entire learning process. **Aim:** contaminate the training data as it is carried out at training phase of the machine learning model.

Attack scenarios

- **Exploratory attack:** give black-box access to the model; not influence training dataset. Aim: gain information about learning algorithm of the underlying system; pattern training data.
- **Dodging attack:** occurs when the attacker tries to have a face misidentified as any other arbitrary face.
- **Substitute network (black box attack):** The adversary, repeating the query process, creates a network similar to the target model. After the creation of the substitute network, the white box attack can be performed. Success rate of attack for Amazon/Google services: 80% approximately.

Attack scenarios

- **Impersonation attack (face detectors):** the attacker poison the data by inserting designed samples, compromising the learning procedure in whole.
- **Aim:** disguise a face as a specific (authorized) face.



[21]. Example of impersonation attack on FaceNet in white-box setting. (a) is captured image of the adversary's face with adversarial light projected in physical domain that is recognized as target (b).

[NGY2020] D.-L. Nguyen, S.-S. Arora, Y. Wu, H. Yang, "Adversarial Light Projection Attacks on Face Recognition Systems: A Feasibility Study", *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020.

Attack scenarios

- **Spoofing attack (face):** Facial spoof attack is a procedure where the deceptive user subversive or attack a FR system by masquerading as registered user, getting illegal access and advantages.
- **Typical countermeasure**: live face detection or antispoofing techniques can be classified based on clues used for spoof attack detection:
  - motion analysis based methods;
  - texture analysis based methods;
  - hardware-based methods.

Face and Object De-detection

- Object detection and de-detection
- Adversarial Attacks
  - Threat Model
  - Perturbation
  - Attack Methods
  - Attack Scenarios
  - **Defense Methods**
  - Benchmarking
- Face Detection Obfuscation

Defense Methods

Basic types of adversarial examples defense:

- **Reactive defense:** adversarial example detection
- **Proactive defense:** enhance the robustness of DNNs.
  - Distillation method
  - Adversarial training
  - Filtering method
- **Ensemble defense methods:** improve the accuracy of DNNs on test data; increase robustness against adversarial perturbations.

Defense Methods

Reactive defense

- Adversarial example detection
  - Binary threshold: last layer’s output as the features
  - Distinguish distribution differences

    4 Confidence value, p value

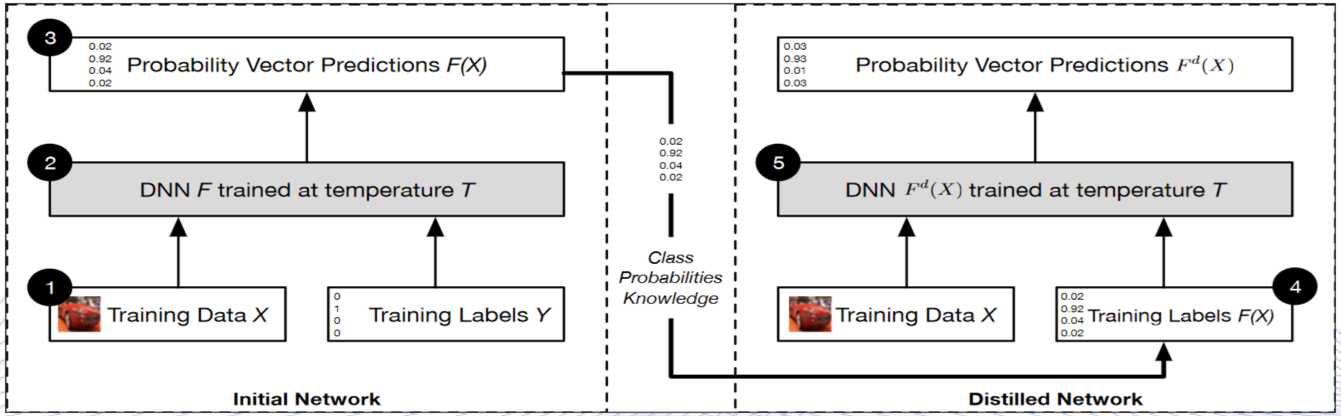
Artificial Intelligence & Information Analysis Lab

Defense Methods

Distillation method

- Use of two DNNs (detailed class probability);
- Avoidance the calculation of the gradient loss function.

Ex) 1", class: [0 1 00000000]   1 ", class:[0.02 0.91 … 0.02



[22]. Visualization of a defense mechanism based on the knowledge transfer through distillation. [PAP2015] N. Papernot, P. McDaniel, X. Wu, S. Jha, A. Swami, "Distillation as a Defense to Adversarial Perturbations against Deep Neural Networks", *arXiv:1511.04508*, 2015.

Artificial Intelligence & Information Analysis Lab

Defense Methods

- **Filtering method**

- Elimination of adversarial example perturbation.

- Creation of filtering module requires time and process.

Defense Methods

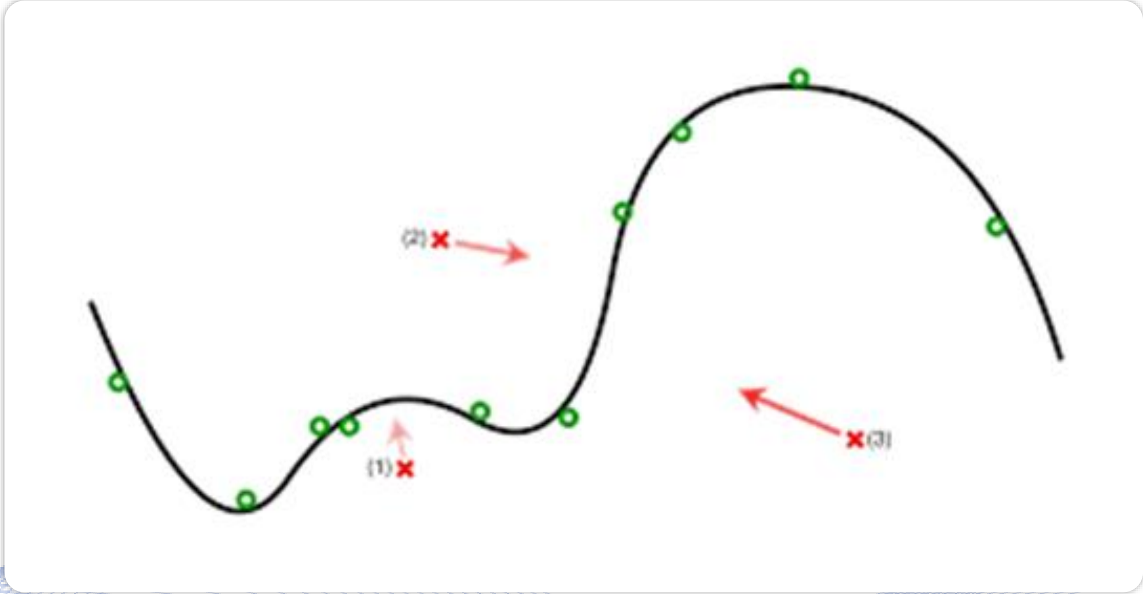
Ensemble proactive defense

Detector:

- compare the output of several original samples to find the adversarial example;
- detect of adversarial examples with large distortion;
- in case of small distortion, lower detection probability;
- combination of multiple detector configurations is available.

Reformer:

- target adversarial example with small distortion;
- use of auto encoder;
- convert of adversarial examples with output most closely the original sample.



[23]. Detector and reformer, working in a 2-D sample space. Normal examples (curve); normal and adversarial examples (green dots & red crosses); transformation by autoencoder (arrows); reconstruction error/rejects examples with large reconstruction errors (cross 3); the reformer finds an example near the manifold, approximating the original example (cross 1). https://dl.acm.org/doi/10.1145/3133956.3134057

Artificial Intelligence & Information Analysis Lab

Face and Object De-detection

- Object detection and de-detection
- Adversarial Attacks
  - Threat Model
  - Perturbation
  - Attack Methods
  - Attack Scenarios
  - Defense Methods
  - **Benchmarking**
- Face Detection Obfuscation

Benchmarking

- **Datasets**
  - **MNIST** & **CIFAR-10**: easy to attack/defend due to their simplicity and small size;
  - **ImageNet:** well-designed dataset to evaluate adversarial attacks;
  - **LFW, CASIA-WebFace, MegaFace, VGGFace2** & **CelebA**: used to evaluate Aas on FR systems.
- **Victim/Target Models**
  - **LeNet, VGG, AlexNet, GoogLeNet, CaffeNet** & **ResNet**: adversaries usually attack several eminent DNN models on Face and Object recognition.
  - **DeepFace, FaceNet, VGG-Face, DeepID, SphereFace, CosFace ArcFace, OpenFace, dlib, LResNet100E-IR Face ID model**: Deep FR models that adversaries broadly attack.

Face and Object De-detection

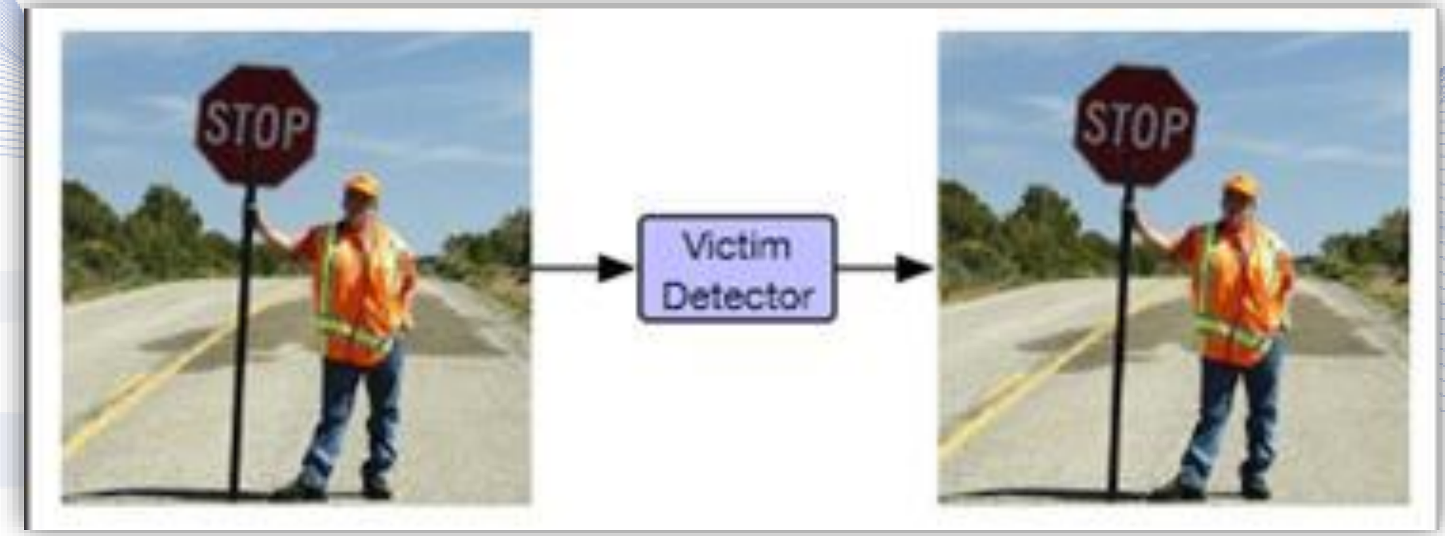
- Object detection and de-detection
- Adversarial Attacks
  - Threat Model
  - Perturbation
  - Attack Methods
  - Attack Scenarios
  - Defense Methods
  - Benchmarking
- **Face Detection Obfuscation**

Face detection obfuscation

The detector does not detect faces (object-vanishing attack)**:**
- all attacks against object detectors focus on objects with fixed visual patterns;
- do not take into account intra-class variety;

- adversarial patches can be used to fool person detectors;
- attack on targets with high level intra-class variety, like persons;
- detector does not detect any persons or objects

- "adversarial patch" used as a cloaking device to hide people from object detectors.

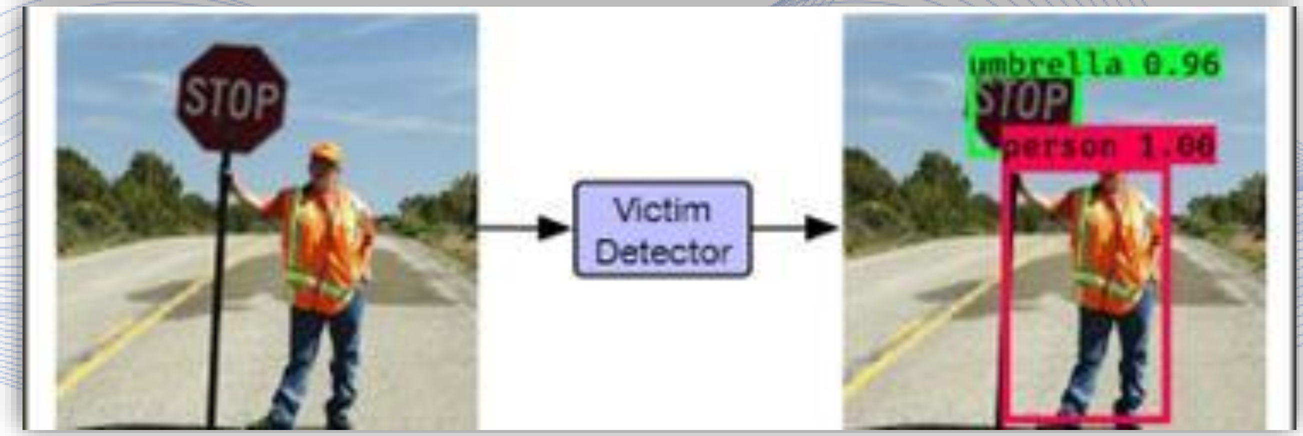


[24]. Object-vanishing attack. [CHO2020] K.-H. Chow, L. Liu, M. Loper, J. Bae, M.-E. Gursoy, S. Truex, W. Wei, Y. Wu, "Adversarial Objectness Gradient Attacks in Real-time Object Detection Systems", in *IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA), pp. 263-272, 2020.*

Face detection obfuscation

The detector detects faces, but classifies them as something else (object-mislabeling attack)**:**

- generate adversarial examples without having access to any information about the network parameter values or their gradients.
- The only input their technique requires is the probabilistic labels predicted by the targeted model.
- adversarial attacks misleads the autoencoder to reconstruct a completely different image



[25]. Object-vanishing attack. https://khchow.com/media/TPS20_TOG.pdf

Face detection obfuscation

The detector detects faces, but classifies them as something else (object-fabrication attack)**:** the object-mislabeling attack fools the detector to mislabel detected objects (e.g., stop sign as an umbrella), which can result in disastrous consequences.

[26]. Object-vanishing attack. https://khchow.com/media/TPS20_TOG.pdf

Bibliography

- [AI2021] S. Ai, A.S.V. Koe, T. Huang, "Adversarial perturbation in remote sensing image recognition", *Appl. Soft Comput.,* vol. 105, pp. 107252-107265, 2021. https://www.sciencedirect.com/science/article/pii/S1568494621001757
- [AKH2019] Z. Akhtar, D. Dasgupta, *A brief survey of Adversarial Machine Learning and Defense Strategies, Technical Report No. CS-19-002*, University of Memphis, 2019. https://www.memphis.edu/cs/research/tech_reports/tr-cs-19-002.pdf
- [ALS2020] B. Alshemali, J. Kalita, "Improving the Reliability of Deep Neural Networks in NLP: A Review", *Knowledge-Based Systems*, vol. 191, pp. 105210-105229, 2020. https://www.sciencedirect.com/science/article/pii/S0950705119305428
- [BAI2021] T. Bai, J. Zhao, J. Zhu, S. Han, J. Chen, B. Li, "AI-GAN: Attack-Inspired Generation of Adversarial Examples", *arXiv:2002.02196*, 2021. https://arxiv.org/pdf/2002.02196.pdf
- [BAK2019] Y. Bakhti, S.-A. Fezza, W. Hamidouche, O. Déforges, "DDSA: A Defense against Adversarial Attacks using Deep Denoising Sparse Autoencoder", *IEEE Access*, vol. 7, pp.160397-160407, 2019. https://www.researchgate.net/publication/337025155_DDSA_A_Defense_Against_Adversarial_Attacks_Using_Deep_Denoising_Sparse_Autoencoder
- [BLU2020] A. Blum, T. Dick, N. Manoj, H. Zhang, "Random Smoothing Might be Unable to Certify ℓ_∞ Robustness for High-Dimensional Images, *arXiv.2002.03517*, pp. 1-20, 2020. https://arxiv.org/pdf/2002.03517.pdf
- [BOS2018] A. Bose, *Adversarial Attacks on Face Detectors using Neural Net based Constrained Optimization*, PhD thesis, Department of Electrical & Computer Engineering, University of Toronto, pp. 1-93, 2018. https://tspace.library.utoronto.ca/handle/1807/91439

Bibliography

- [BRE2018] W. Brendel, J. Rauber, A. Kurakin, N. Papernot, B. Veliqi, M. Salathé, S.-P. Mohanty, M. Bethge, “Adversarial Vision Challenge”, *arXiv:1808.01976*, 2018. https://arxiv.org/pdf/1808.01976.pdf
- [CAR2017] N. Carlini, D. Wagner, “Towards evaluating the robustness of neural networks”, in *2017 IEEE Symposium on Security and Privacy (SP),* pp. 39-57, 2017. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7958570
- [CHA2021] A. Chakraborty, M. Alam, V. Dey, A. Chattopadhyay, D. Mukhopadhyay, “A survey on adversarial attacks and defenses”, *CAAI Trans. Intell. Technol*., vol. 6(1), pp. 25-45, 2021. https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/cit2.12028
- [CHE2020] J. Chen, H. Zheng, H. Xiong, S. Shen, M. Su, “MAG-GAN: Massive attack generator via GAN”, *Inform. Sci.*, vol. 536, pp. 67-90, 2020. https://www.sciencedirect.com/science/article/pii/S0020025520303194
- [CHE2020] S. Cheng, Y. Liu, S. Ma, X. Zhang, Deep Feature Space Trojan Attack of Neural Networks by Controlled Detoxification. *arXiv:2012.11212*, 2020. https://arxiv.org/pdf/2012.11212.pdf
- [CHO2020] K.-H. Chow, L. Liu, M. Loper, J. Bae, M.-E. Gursoy, S. Truex, W. Wei, Y. Wu, “Adversarial Objectness Gradient Attacks in Real-time Object Detection Systems”, in *IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA),* pp. 263-272, 2020. https://khchow.com/media/TPS20_TOG.pdf
- [DHA2019] M. Dhaouadi, “A survey about adversarial learning”, 2019. https://www.researchgate.net/publication/338105748_A_SURVEY_ABOUT_ADVERSARIAL_LEARNING

Bibliography

- [DON2018] Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu, X. Hu, J. Li, “Boosting Adversarial Attacks with Momentum”, *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 9185-9193, 2018. https://arxiv.org/pdf/1710.06081v3.pdf
- [DRI2020] K. Drid, M. Allaoui, M.-L. Kherfi, “Object detector combination for increasing accuracy and detecting more overlapping objects”, in A. El Moataz, D. Mammass, A. Mansouri, F. Nouboud (eds), *Image and Signal Processing, Springer International Publishing*, pp. 290-296, 2020. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7340881/
- [DUA2021] Y. Duan, X. Zhou, J. Zou, J. Qiu, J. Zhang, Z. Pan, “Mask-guided noise restriction adversarial attacks for image classification”, *Computers & Security*, vol. 100, pp. 102111, 2021. https://www.sciencedirect.com/science/article/pii/S0167404820303849#!
- [FEL2019] S. Feldstein, *The Global Expansion of AI Surveillance*, 2021. https://zhizhi88.com/wp-content/uploads/2019/10/2019%E5%B9%B4%E5%85%A8%E7%90%83%E4%BA%BA%E5%B7%A5%E6%99%BA%E8%83%BD%E7%9B%91%E6%B5%8B%EF%BC%88AIGS%EF%BC%89%E6%8C%87%E6%95%B0%E6%8A%A5%E5%91%8A.pdf
- [GAN2019] A. Ganeshan, B. Vivek, R. Babu, “FDA: Feature Disruptive Attack”, *IEEE Explore*, 2019. https://openaccess.thecvf.com/content_ICCV_2019/papers/Ganeshan_FDA_Feature_Disruptive_Attack_ICCV_2019_paper.pdf
- [GEI2020] J. Geiping, H. Bauermeister, H. Dröge, M. Moeller, “Inverting Gradients – How easy is it to break privacy in federated learning?”, *Advances in Neural Information Processing Systems, (NeurIPS 2020),* vol. 33, 2020. https://papers.nips.cc/paper/2020/file/c4ede56bbd98819ae6112b20ac6bf145-Paper.pdf

Bibliography

- [GOL2020] M. Goldblum, D. Tsipras, C. Xie, X. Chen, A. Schwarzschild, D. Song, A. Madry, B. Li, T. Goldstein, “Dataset Security for Machine Learning: Data Poisoning, Backdoor Attacks, and Defenses”, *arXiv:2012.10544*, Dec 2020. https://arxiv.org/pdf/2012.10544.pdf
- [GOO2014] I.-J. Goodfellow, J. Shlens, C. Szegedy, “Explaining and Harnessing Adversarial Examples”, *arXiv:1412.6572*, 2014. https://arxiv.org/pdf/1412.6572.pdf
- [GRI2020] J. Griping, L. Fowl, W.-R. Huang, W. Czaja, G. Taylor, M. Moeller, T. Goldstein, “Witches’ Brew: Industrial Scale Data Poisoning via Gradient Matching”, *arXiv:2009.02276*, 2020. http://arxiv-export-lb.library.cornell.edu/pdf/2009.02276
- [GU2020] Z. Gu, W. Hu, C. Zhang, L. Wang, C. Zhu, Z. Tian, “Restricted Region Based Iterative Gradient Method for Non-Targeted Attack”, *IEEE Access*, vol. 8, pp. 25262-25271, 2020. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8978619
- [KUR16] A. Kurakin, I. Goodfellow, S. Bengio, “Adversarial examples in the physical world”, *arXiv:1607.02533*, 2016. https://arxiv.org/abs/1607.02533
- [LIU2018] X. Liu, H. Yang, Z. Liu, L. Song, H. Li, Y. Chen, “DPatch: An Adversarial Patch Attack on Object Detectors”, *arXiv:1806.02299*, 2018. https://arxiv.org/abs/1806.02299
- [MAC2021] B. Mac Donald, “Fooling Facial Detection with Fashion – Towards Data Science”, *Medium*, 2021.
- [MAS2021] F. V. Massoli, F. Carrara, G. Amato, F. Falchi, “Detection of Face Recognition Adversarial Attacks”, *Comput. Vision Image Understanding*, vol, 202, pp. 103103, 2021. https://www.sciencedirect.com/science/article/pii/S1077314220301296

Bibliography

- [MEE2021] K. Meenakshi, G. Maragatham, “A Self Supervised Defending Mechanism Against Adversarial Iris Attacks based on Wavelet Transform”, *International Journal of Advanced Computer Science and Applications*, vol. 12(2), 2021. https://thesai.org/Downloads/Volume12No2/Paper_70-A_Self_Supervised_Defending_Mechanism.pdf
- [MEE2021] V. Meel, “Object Detection in 2021: The Definitive Guide», *viso.ai*, 2021. https://viso.ai/deep-learning/object-detection/
- [MEN2017] D. Meng, H. Chen, “MagNet: A Two-Pronged Defense against Adversarial Examples”, 2017. https://dl.acm.org/doi/10.1145/3133956.3134057
- [NES2021] F. Nesti, A. Biondi, G. Buttazzo, “Detecting Adversarial Examples by Input Transformations, Defense Perturbations, and Voting”, *arXiv:2101.11466*, 2021.
- [NGY2020] D.-L. Nguyen, S.-S. Arora, Y. Wu, H. Yang, “Adversarial Light Projection Attacks on Face Recognition Systems: A Feasibility Study”, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020. https://openaccess.thecvf.com/content_CVPRW_2020/papers/w48/Nguyen_Adversarial_Light_Projection_Attacks_on_Face_Recognition_Systems_A_Feasibility_CVPRW_2020_paper.pdf
- [PAP2015] N. Papernot, P. McDaniel, X. Wu, S. Jha, A. Swami, “Distillation as a Defense to Adversarial Perturbations against Deep Neural Networks”, *arXiv:1511.04508*, 2015. https://arxiv.org/abs/1511.04508
- [PAP2021] P. Papadopoulos, O. Thornewill von Essen, N. Pitropakis, C. Chrysoulas, A. Mylonas, W.-J. Buchanan, “Launching Adversarial Attacks against Network Intrusion Detection Systems for IoT”, *J. Cybersecur. Priv.*, vol. 1(2), pp. 252-273, 2021. https://www.mdpi.com/2624-800X/1/2/14

Bibliography

- [PIN2021] M. Pintor, F. Roli, W. Brendel, B. Biggio, “Fast Minimum-norm Adversarial Attacks through Adaptive Norm Constraints”, *arXiv:2102.12827*, 2021. https://arxiv.org/abs/2102.12827

- [PIN2021] M. Pintor, F. Roli, W. Brendel, B. Biggio, “Fast Minimum-norm Adversarial Attacks through Adaptive Norm Constraints”, *arXiv:2102.12827, 2021*. https://arxiv.org/abs/2102.12827

- [SHA2018] A. Shafahi, W.-R.Huang, M. Najib, O. Suciu, C. Studer, T. Dumitras, T. Goldstein, “Poison Frogs! Targeted Poisoning Attacks on Neural Networks”, *Neural Information Processing Systems (NeurIPS),* 2018. https://arxiv.org/pdf/1804.00792.pdf
- [SU2019] J. Su, D. Vasconcellos Vargas, K. Sakurai, “Onepixel attack for fooling deep neural networks”, *IEEE Transactions on Evolutionary Computation*, vol. 23(5), pp. 828-841, 2019. https://arxiv.org/pdf/1710.08864.pdf
- [SUN2018] L. Sun, M. Tan, Z. Zhou, “A survey of practical adversarial example attacks”, *Cybersecur*., vol. 1(1), pp.1-9, 2018. https://cybersecurity.springeropen.com/track/pdf/10.1186/s42400-018-0012-9.pdf
- [SUN2018] Q. Sun, L. Ma, S. Joon Oh, L. Van Gool, B. Schiele, M. Fritz, “Natural and Effective Obfuscation by Head Inpainting”, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5050-5059, 2018. https://openaccess.thecvf.com/content_cvpr_2018/papers/Sun_Natural_and_Effective_CVPR_2018_paper.pdf
- [SZS2013] C. Szegedy, W.Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, R. Fergus, “Intriguing properties of neural networks”, *International Conference on Learning Representations*, 2013. https://research.google/pubs/pub42503/
- [THY2019] S. Thys, W. Van Ranst, T. Goedemé, “Fooling Automated Surveillance Cameras: Adversarial Patches to Attack Person Detection”, in *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW*), 49-55, 2019.

Bibliography

- [TRA2020] F. Tramer, N. Carlini, W. Brendel, A. Madry, “On Adaptive Attacks to Adversarial Example Defenses”, *NeurIPS 2020*, 2020. https://proceedings.neurips.cc/paper/2020/file/11f38f8ecd71867b42433548d1078e38-Paper.pdf
- [WAN2021] X. Wang, K. He, “Enhancing the Transferability of Adversarial Attacks through Variance Tuning”, *arxiv.2103.15571, 2021*. https://arxiv.org/pdf/2103.15571.pdf
- [WAN2021] Y. Wang, H. Lv, X. Kuang, G.Zhao, Y. Tan, Q. Zhang, J. Hu, “Towards a physical-world adversarial patch for blinding object detection models”, *Inform. Sci*., vol. 556, pp. 459-471, 2021. https://www.sciencedirect.com/science/article/pii/S0020025520308586
- [WON2020] E. Wong, J.-Z. Kolter, “Learning perturbation sets for robust machine learning”, in *ICLR (preprint),* 2020. https://arxiv.org/pdf/2007.08450.pdf
- [WU2019] Y. Wu, F.Yang, Y. Xu, H. Ling, “Privacy-Protective-GAN for Privacy Preserving Face De-Identification”, *J. Comput. Sci.Tech*., vol. 34(1), pp. 47-60, 2019. https://arxiv.org/abs/1806.08906
- [XIA2018] C. Xiao, B. Li, J. Zhu, W. He, M. Liu, D. Song, “Generating Adversarial Examples with Adversarial Networks”, *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18),* 2018. https://www.ijcai.org/proceedings/2018/0543.pdf
- [XU2020] H. Xu, Y. Ma, H.-C. Liu, D. Deb, H. Liu, J.-L. Tang, A.K. Jain, “Adversarial Attacks and Defenses in Images. Graphs and Text: A Review”, *Int. J. Autom. Comput*., vol. 17(2), pp. 151-178, 2020. https://link.springer.com/article/10.1007/s11633-019-1211-x

Bibliography

- [YAN2020] C. Yang, A. Kortylewski, C. Xie, Y. Cao, A.Yuille, "PatchAttack: A Black-box Texture-based Attack with Reinforcement Learning", in A. Vedaldi, H. Bischof, T. Brox, JM. Frahm (eds), *Computer Vision*, ECCV 2020: Lecture Notes in Computer Science, vol 12371, Springer, Cham., 2020.
- [YAN2020] G. Yang, T. Duan, J.-E. Hu, H. Salman, I. Razenshteyn, J. Li. "Randomized Smoothing of All Shapes and Sizes", *37th International Conference on Machine Learning, PMLR*, vol. 119, pp. 10693-10705, 2020. Feb 2020. http://proceedings.mlr.press/v119/yang20c.html
- [YAN2020] X. Yang, D. Yang, Y. Dong, W. Yu, H. Su, J. Zhu, "Delving into the Adversarial Robustness on Face Recognition", *arXiv:2007.04118, 2020*. https://arxiv.org/pdf/2007.04118.pdf
- [YOU2020] S. You, T. Huang, M. Yang, F. Wang, C. Qian, C. Zhang, "Greedy NAS: Towards Fast One-Shot NAS with Greedy Supernet", *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2020),* 2020. https://openaccess.thecvf.com/content_CVPR_2020/papers/You_GreedyNAS_Towards_Fast_One-Shot_NAS_With_Greedy_Supernet_CVPR_2020_paper.pdf
- [YUA2019] X. Yuan, P. He, Q. Zhu, X. Li, "Adversarial Examples: Attacks and Defenses for Deep Learning", *IEEE Trans. Neural Networks Learn. Syst*., vol. 30(9), pp. 2805-2824, 2019.

Bibliography

[PIT2021] I. Pitas, “Computer vision”, Createspace/Amazon, in press.
[PIT2017] I. Pitas, “Digital video processing and analysis” , China Machine Press, 2017 (in Chinese).
[PIT2013] I. Pitas, “Digital Video and Television” , Createspace/Amazon, 2013.
[NIK2000] N. Nikolaidis and I. Pitas, “3D Image Processing Algorithms”, J. Wiley, 2000.
[PIT2000] I. Pitas, “Digital Image Processing Algorithms and Applications”, J. Wiley, 2000.

Q & A

Thank you very much for your attention!

More material in
http://icarus.csd.auth.gr/cvml-web-lecture-series/

Contact: Prof. I. Pitas
pitas@csd.auth.gr