Adversarial Machine Learning summary

E. Chatzikyriakidis, V. Mygdalis, Prof. Ioannis Pitas
Aristotle University of Thessaloniki
pitas@csd.auth.gr
www.aiia.csd.auth.gr
Version 2.8

Artificial Intelligence & Information Analysis Lab

Adversarial Machine Learning

- **Adversarial Examples**
- Attack Methods
- Adversarial Face De-Identification
- Adversarial Defenses

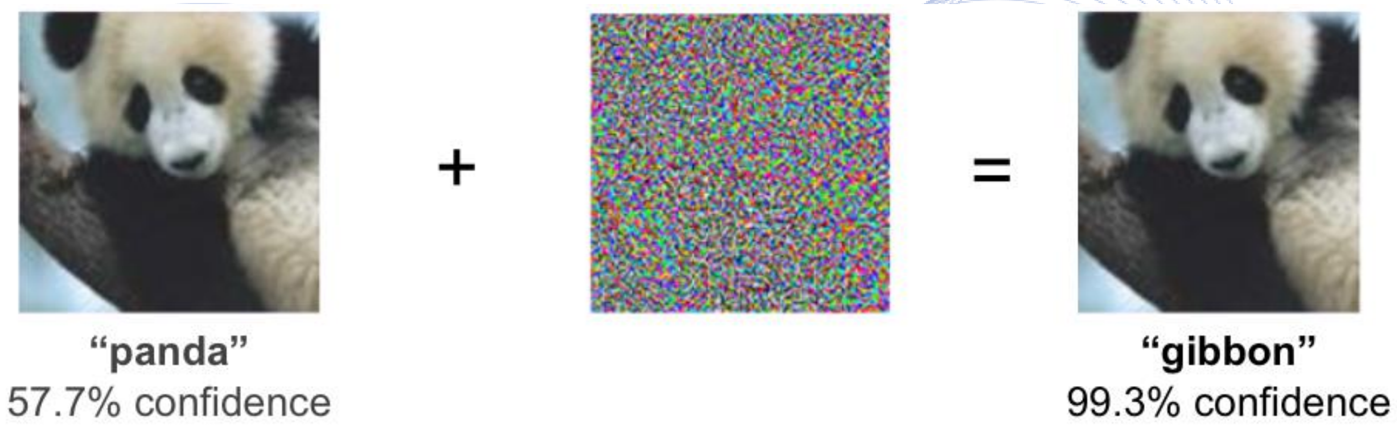
Local Generalization in Computer Vision

- Slight pixel changes should not affect the decision of a model.

Artificial Intelligence & Information Analysis Lab

Adversarial Examples - What exactly are these?

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



The big question

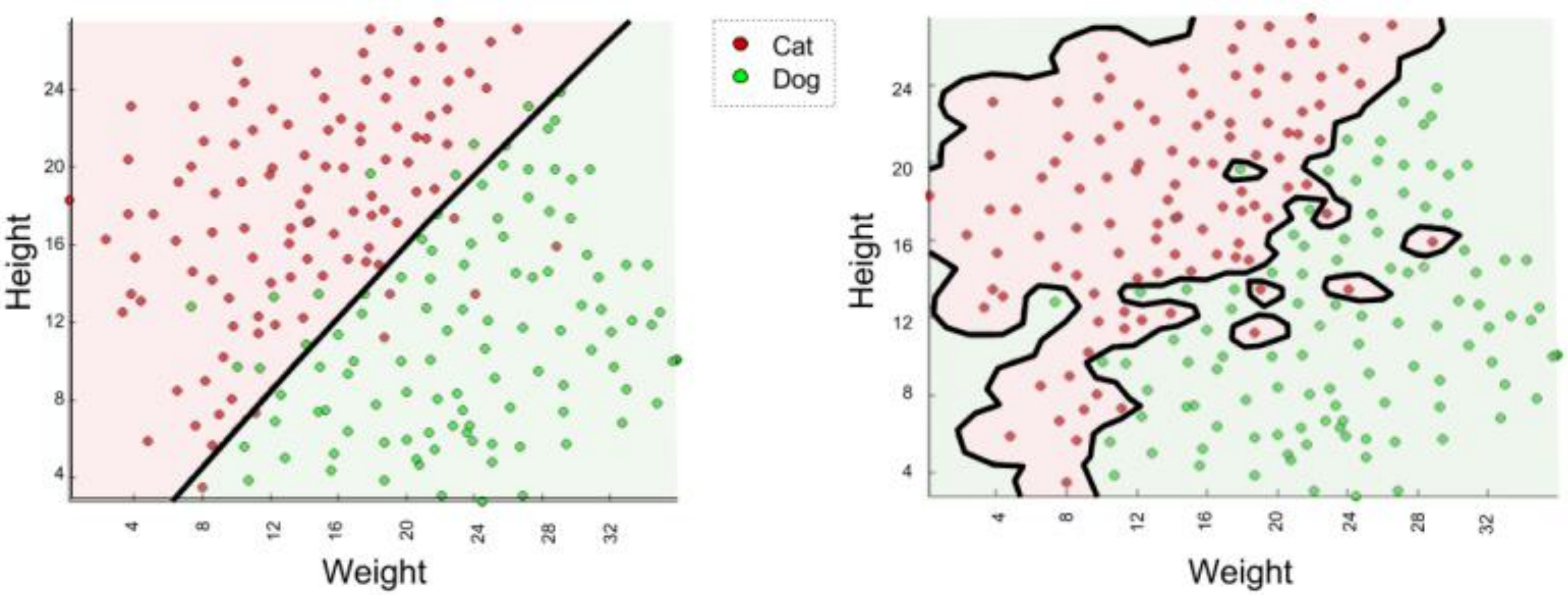
- Why adversarial examples exist and how is it feasible a model that is not overfitted and has high test/validation accuracy not to be functional in adversarial examples which are very similar to the original data?

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

Artificial Intelligence & Information Analysis Lab

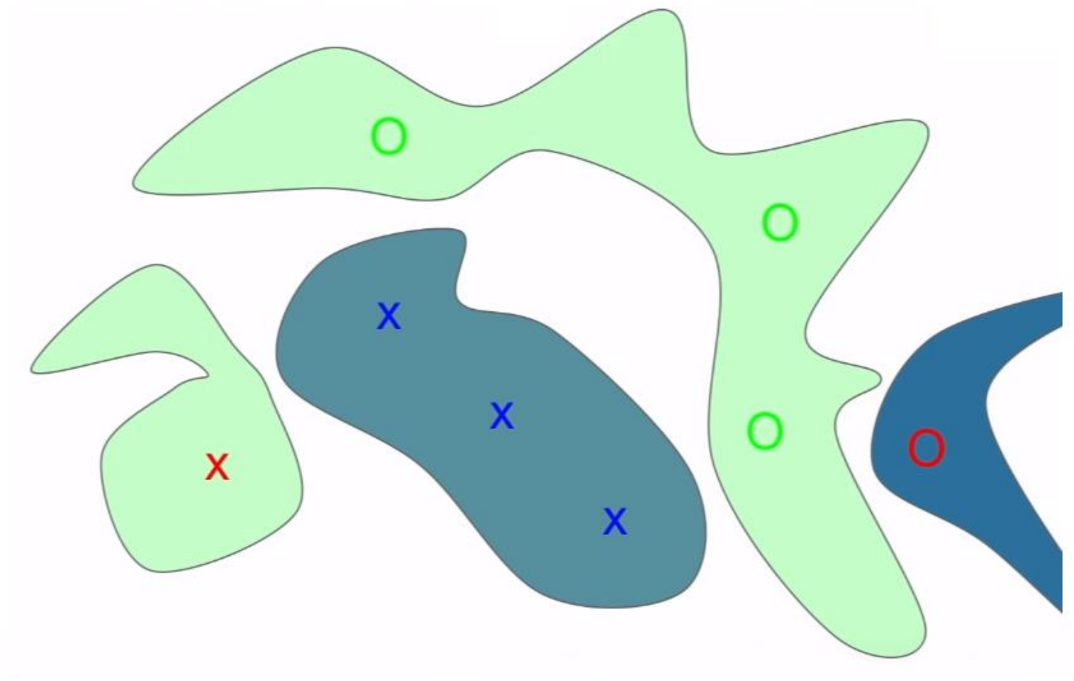
Why do they exist?

- Is it due to overfitting?
  - an overfitted model is sensitive to small input changes since it learns the idiosyncrasies of input.



Why do they exist?

- an overfitted model assigns some blobs of probabilities mass in unseen places of input.



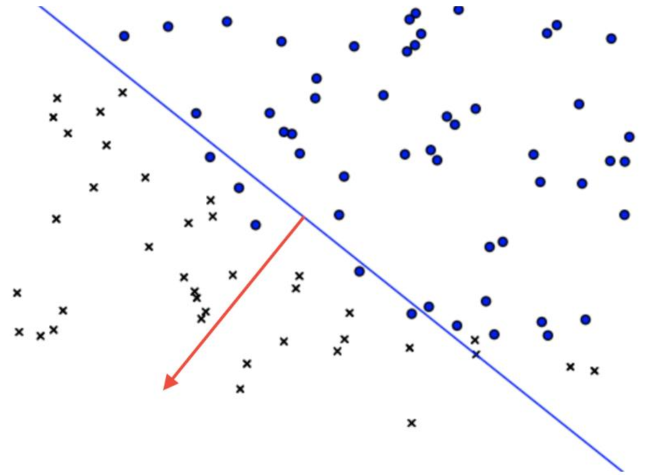
Why do they exist?

- 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.

Why do they exist?

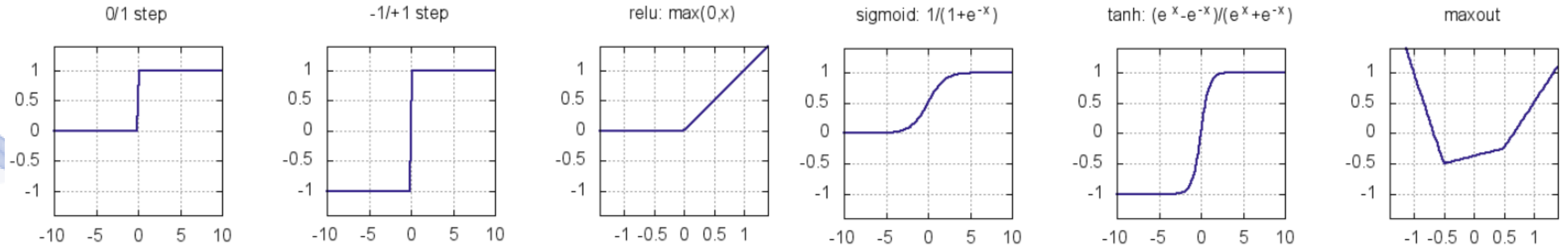
- Linear models are quite pathological outside of the region where training data is concentrated (initial experiments with shallow linear models shown that this affects greatly the $w^T x$ calculation and leads to wrong misclassification).



Artificial Intelligence & Information Analysis Lab

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.



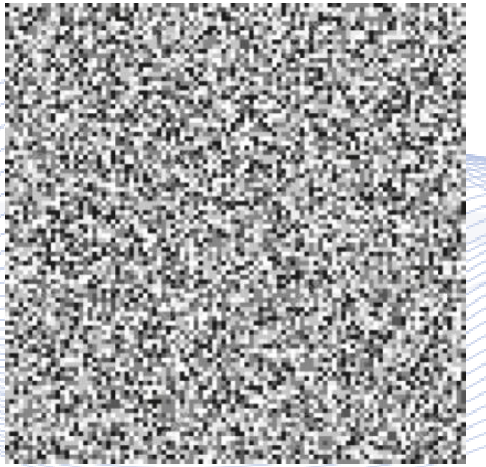
Where does this lead us?

Achilles Heel

Potemkin Village

Random / Non-random input

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

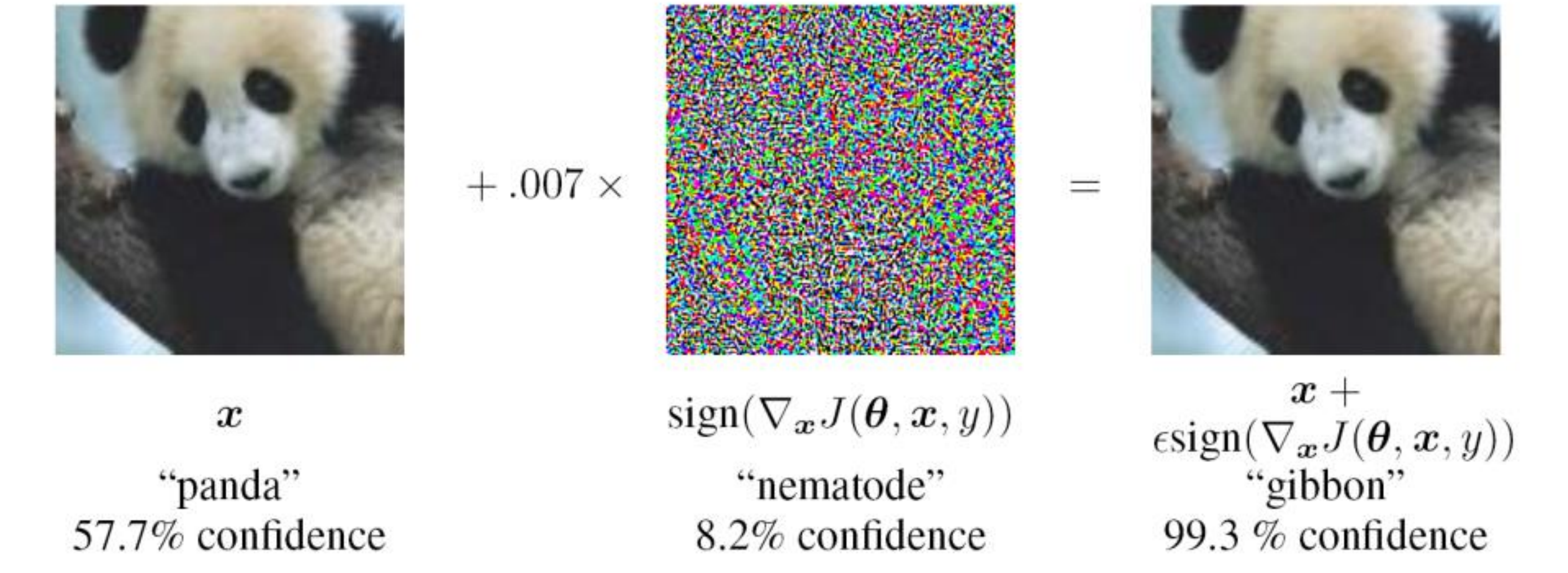


Targeted / Non-targeted Definition

- There is an image and its ground truth class: $\mathbf{x}, y$.
- A classification model f predicts the class of $\mathbf{x}$: $\hat{y} = f(\mathbf{x})$.
- The prediction $\hat{y}$ is the same as the ground truth: $y = \hat{y}$.
- There is an image $\mathbf{x}_p$ which is $\mathbf{x}$ perturbed by p: $\mathbf{x}_p = \mathbf{x} + \mathbf{p}$.
- The distance of the two images is restricted by threshold e: $d(\mathbf{x}, \mathbf{x}_p) \leq e$.
- The threshold e is positive and small for imperceptible changes.
- The classification model classifies the image $\mathbf{x}_p$: $\hat{y}_p = f(\mathbf{x}_p)$.
- Non-targeted adversarial example constraint: $\hat{y}_p \neq y$.
- Targeted adversarial example constraint: $\hat{y}_p = t$.

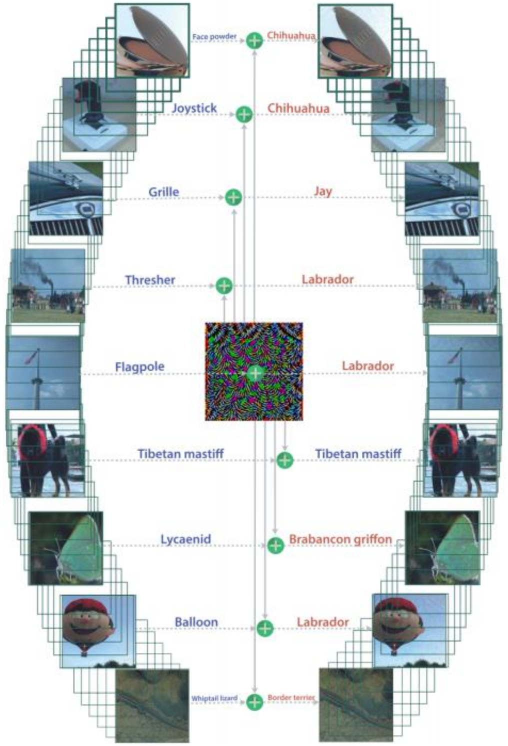
Perturbation Scope

- Individual: Each image has its own individual perturbation.



Perturbation Scope

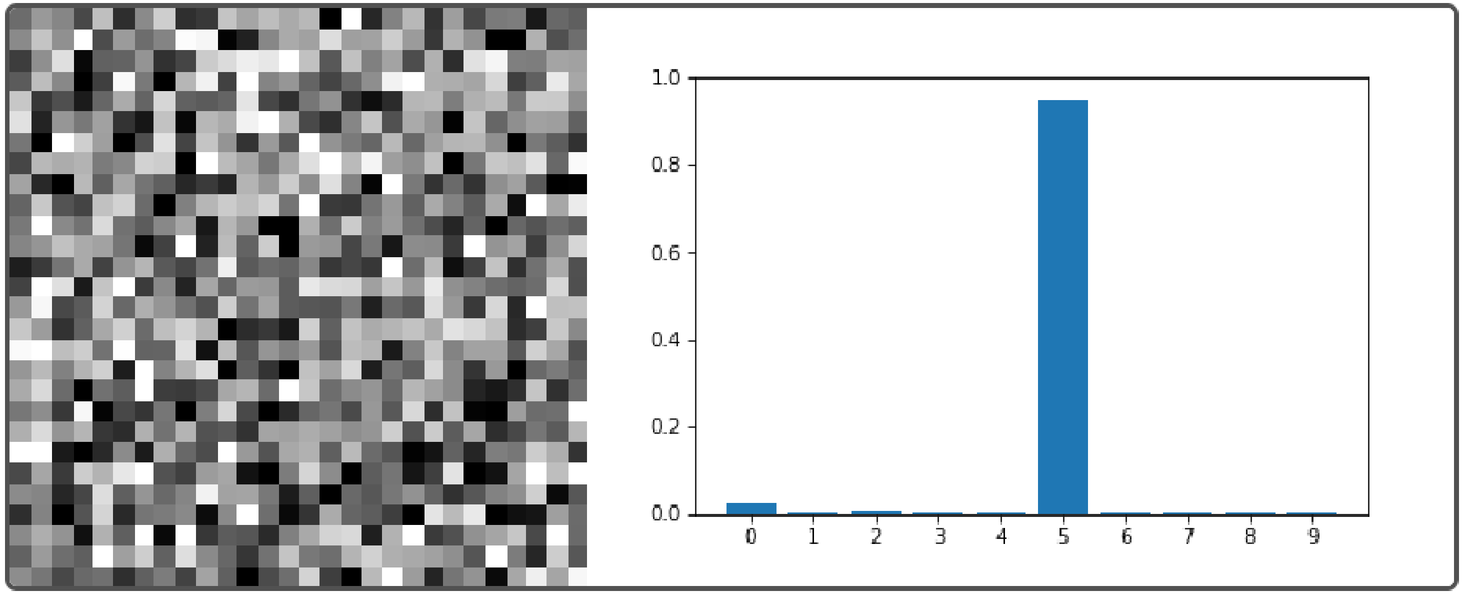
- Universal: A universal perturbation for the whole dataset.



Artificial Intelligence & Information Analysis Lab

Realistic / Non-realistic Output

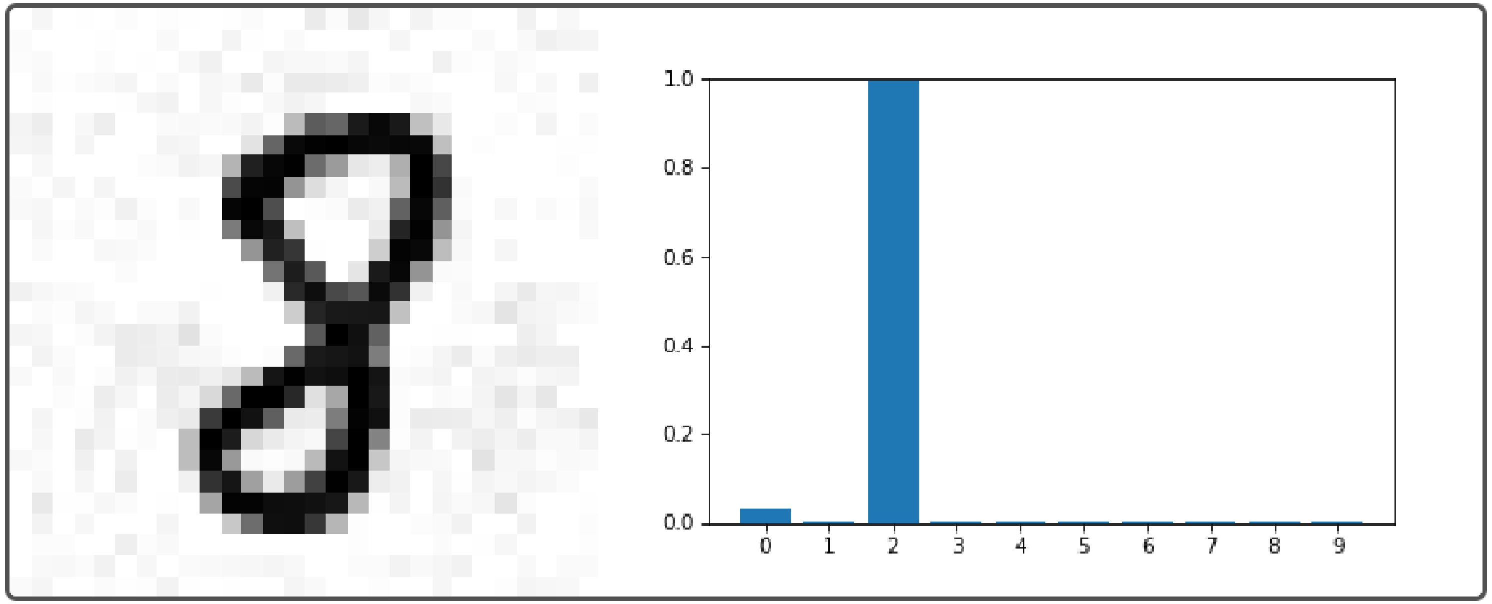
- Non-realistic: Any adversarial input that fools the model.



Artificial Intelligence & Information Analysis Lab

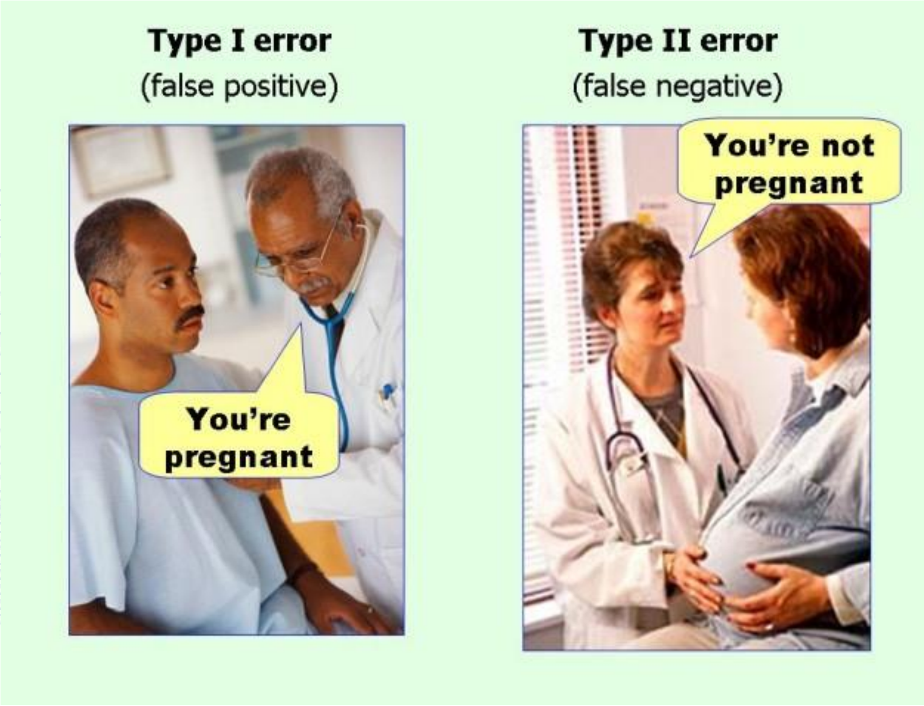
Realistic / Non-realistic Output

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



Adversarial Falsification

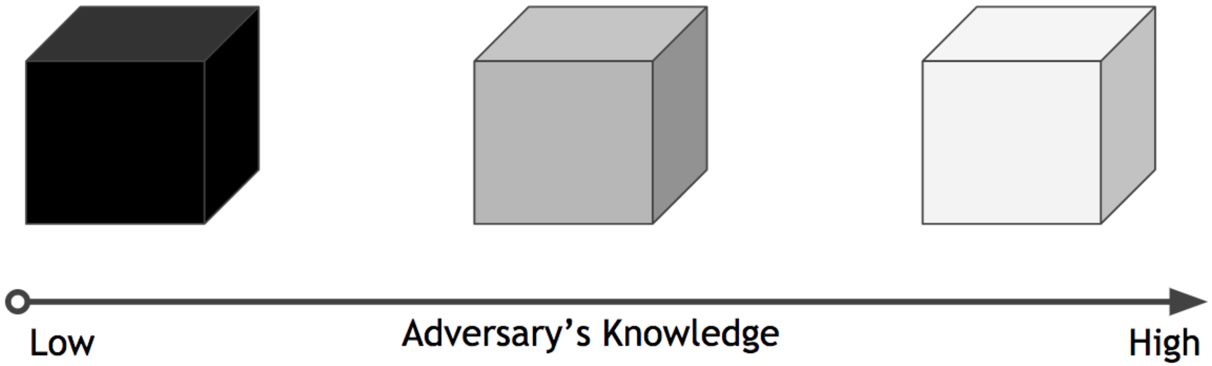
- Type 1 error: Negative sample classified as positive.
- Type 2 error: Positive sample classified as negative.



Artificial Intelligence & Information Analysis Lab

Adversary Knowledge

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



Adversarial Machine Learning

- Adversarial Examples
- **Adversarial Attacks**
- Adversarial Face De-identification
- Adversarial Defenses

Targeted adversarial attacks

- For a given image $\mathbf{x} \in \mathbb{R}^n$ and target label $t \in \mathcal{C} - \{y\}$, targeted adversarial attacks solve the following box-constrained optimization problem:

$$\text{Minimize } \|\mathbf{p}\|_2$$

$$\text{subject to: } f(\mathbf{x}_p; \boldsymbol{\theta}) = t \text{ and } \mathbf{x}_p \in \mathbb{R}^n.$$

- Note: an additional stopping condition of this optimization problem could be just:

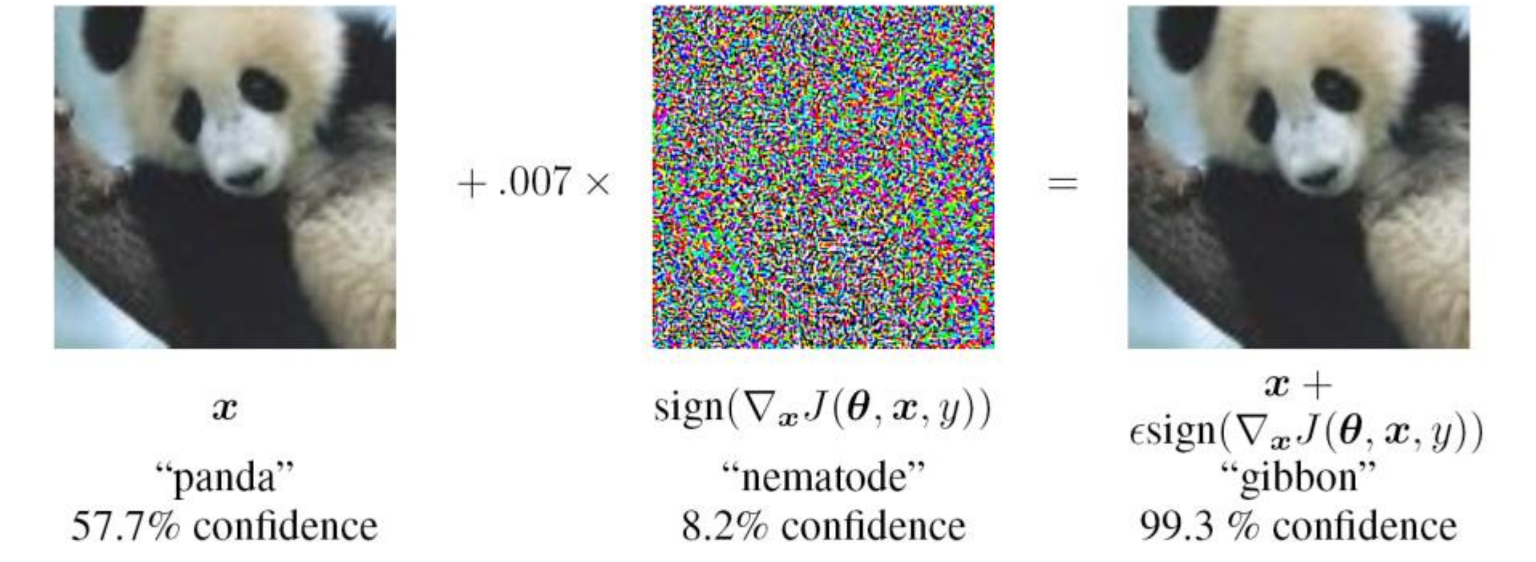
$$f(\mathbf{x}_p; \boldsymbol{\theta}) \neq y$$

- An approximation of this problem can be found by the Limited-memory Broyden–Fletcher–Goldfarb–Shanno (L-BFGS) method, which perturbs the input by exploiting the **gradient** values or signs **returned to the input** layer by the NN.

Szegedy et. al, Intriguing properties of neural networks arXiv:1312.6199v4 [cs.CV],2013

Attack Methods

- Let’s get an idea with “Fast Gradient” Methods.

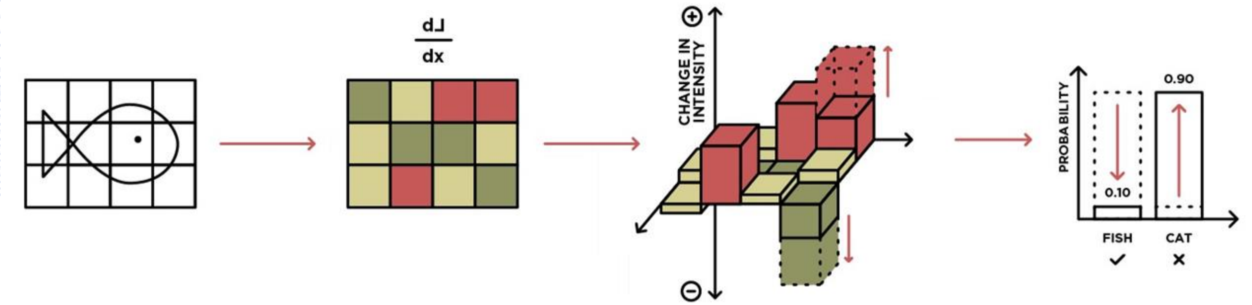


Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples“, *arXiv preprint arXiv:1412.6572*, 2014.

Artificial Intelligence & Information Analysis Lab

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 $\nabla_{\mathbf{x}} J(\boldsymbol{\theta}, \mathbf{x}, y)$ can be calculated with backpropagation.



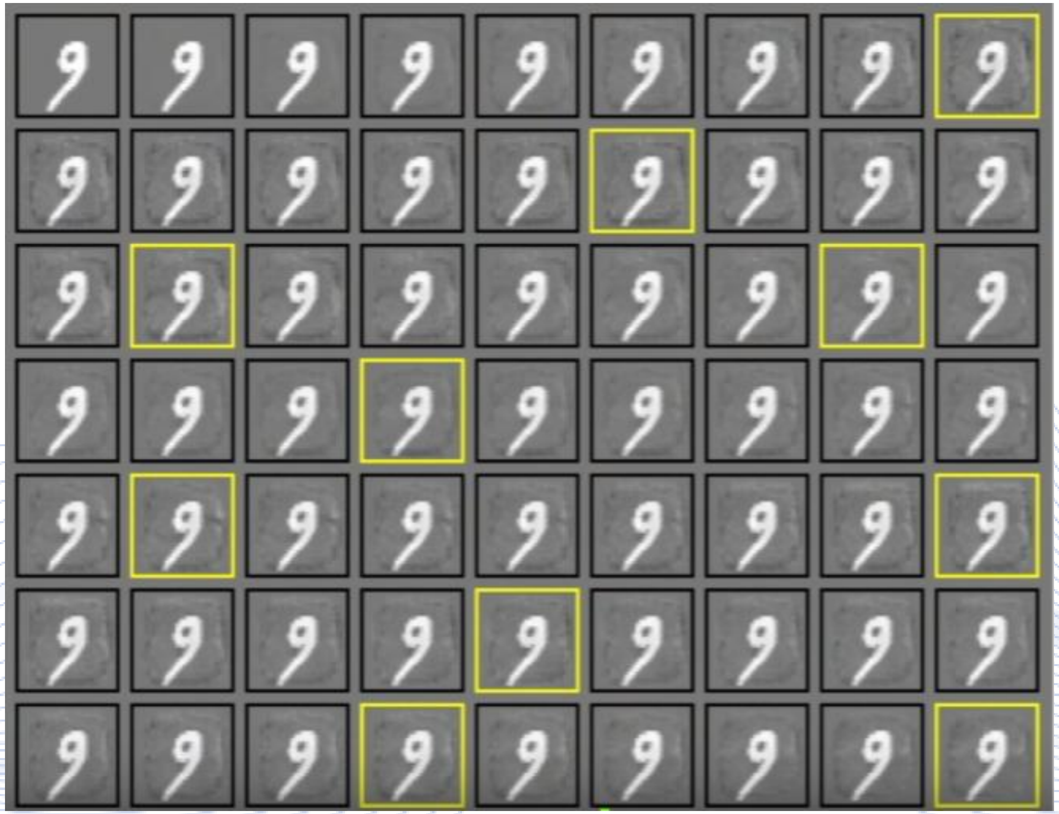
Attack Methods

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

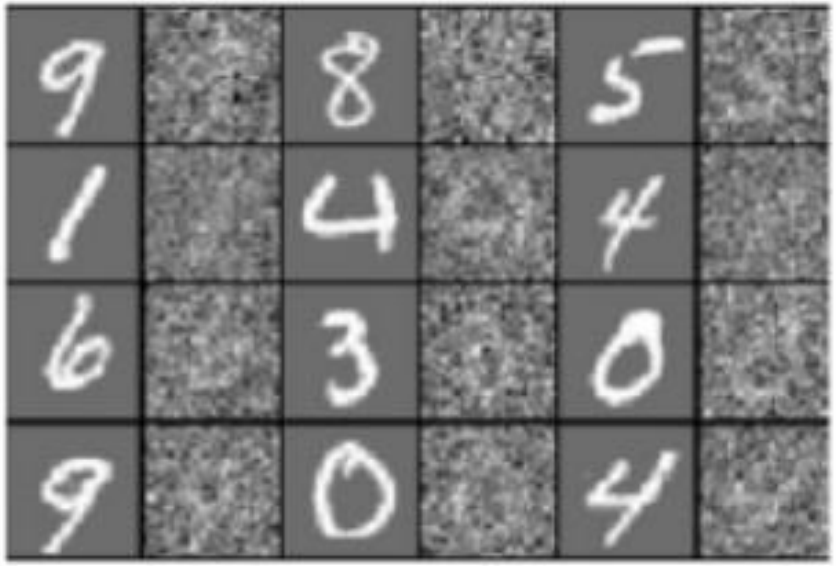
Su, Jiawei, Danilo Vasconcellos Vargas, and Sakurai Kouichi, "One pixel attack for fooling deep neural networks." *arXiv preprint arXiv:1710.08864,* 2017.

Artificial Intelligence & Information Analysis Lab

Softmax regression (MNIST)

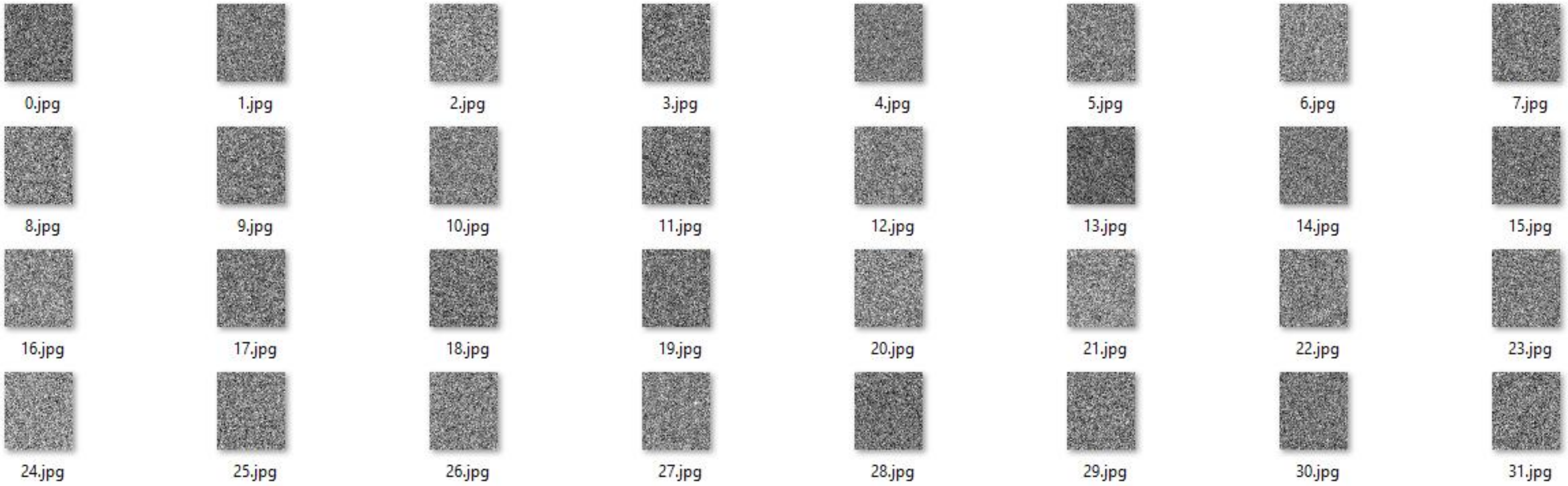


Softmax regression (MNIST)



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.

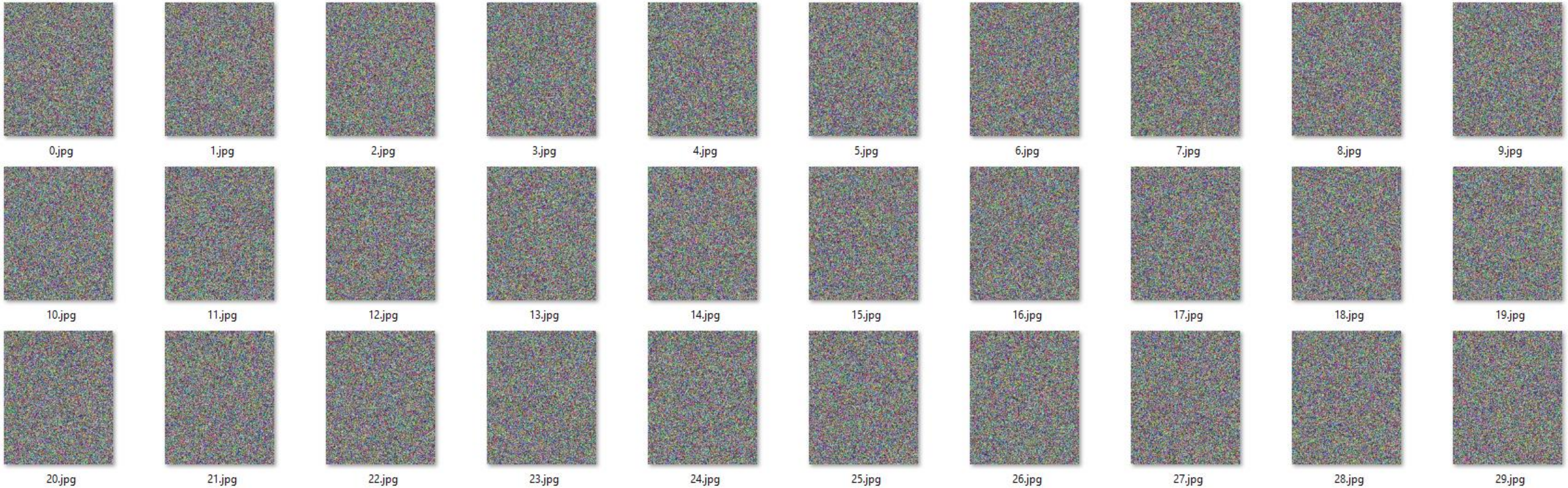


MLP w/ Ext. Yale Face Database B

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

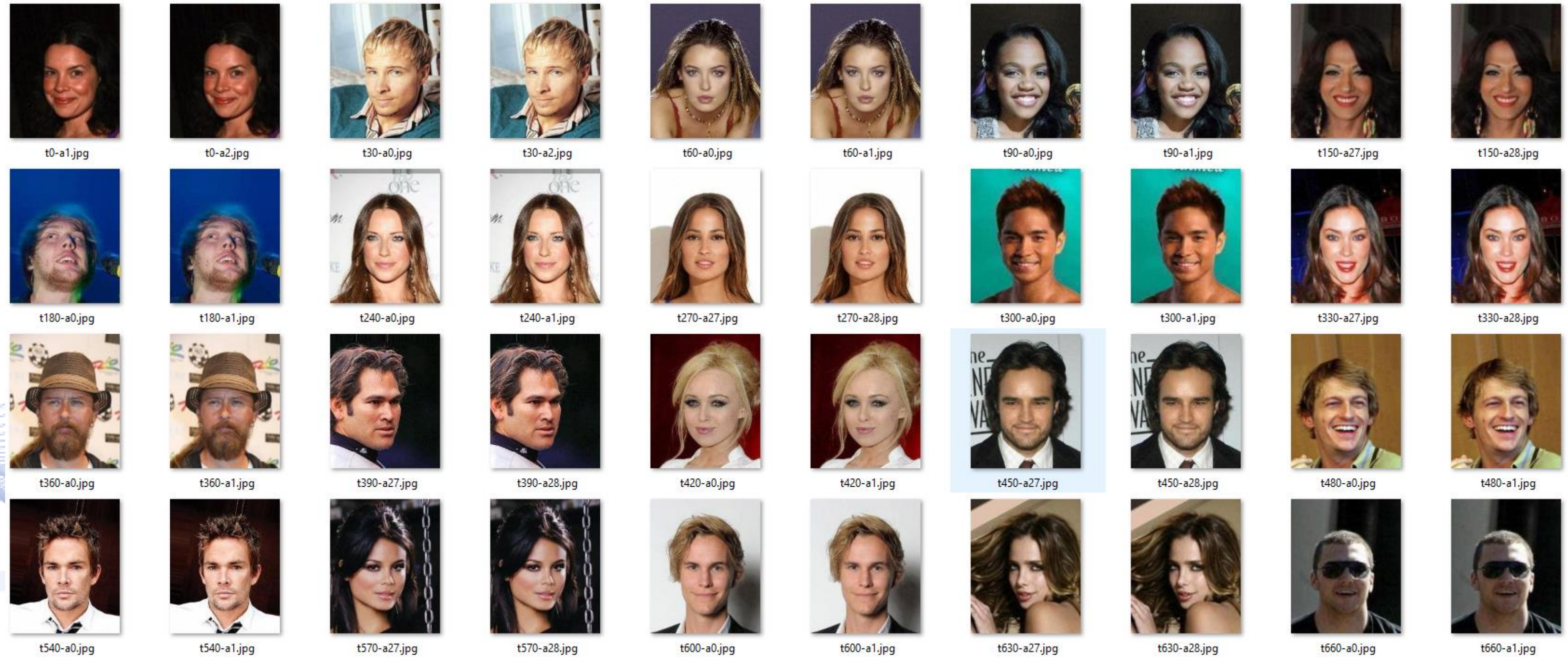
CNN w/ CelebA

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



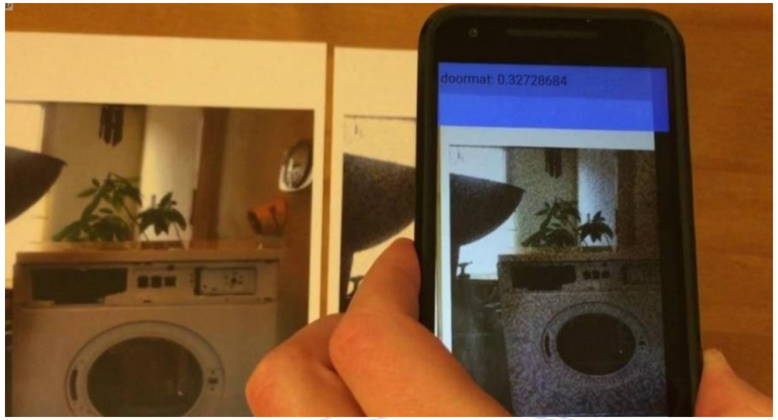
CNN w/ CelebA

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



Adversarial attacks are transferable!

- Can exist to the real world!



Kurakin, Alexey, Ian Goodfellow, and Samy Bengio, "Adversarial examples in the physical world." *arXiv preprint arXiv:1607.02533,* 2016.

“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.

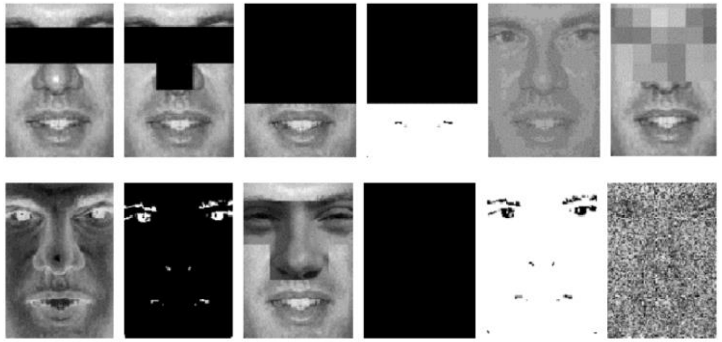
Adversarial Machine Learning

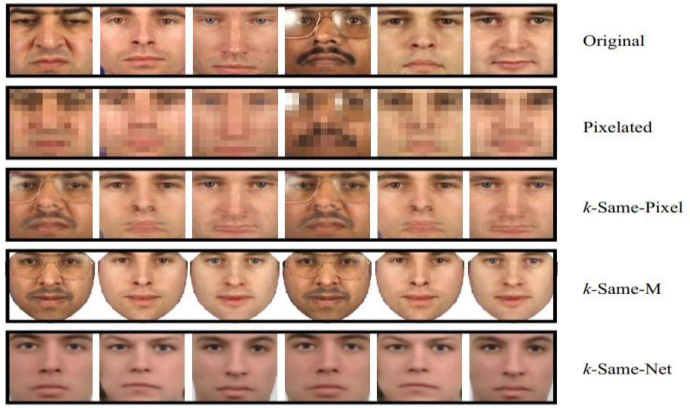
- Adversarial Examples
- Adversarial Attacks
- **Adversarial Face De-identification**
- Adversarial Defenses

Adversarial Face De-Identification

Motivation - Drawbacks of previous methods

- Privacy protection on images and videos.
- Previous face de-identification methods strongly alter original images.
- De-identified image should retain the original facial image unique characteristics (e.g. race, gender, age, expression, pose).





Adversarial Face De-Identification

Iterative Fast Gradient Value Method I-FGVM

- Assuming a common NN-framework transform, let image samples $\boldsymbol{x}$ with pixel values normalized in the [0,1] domain.
- The gradient descent update equations of the I-FGVM are of following form:

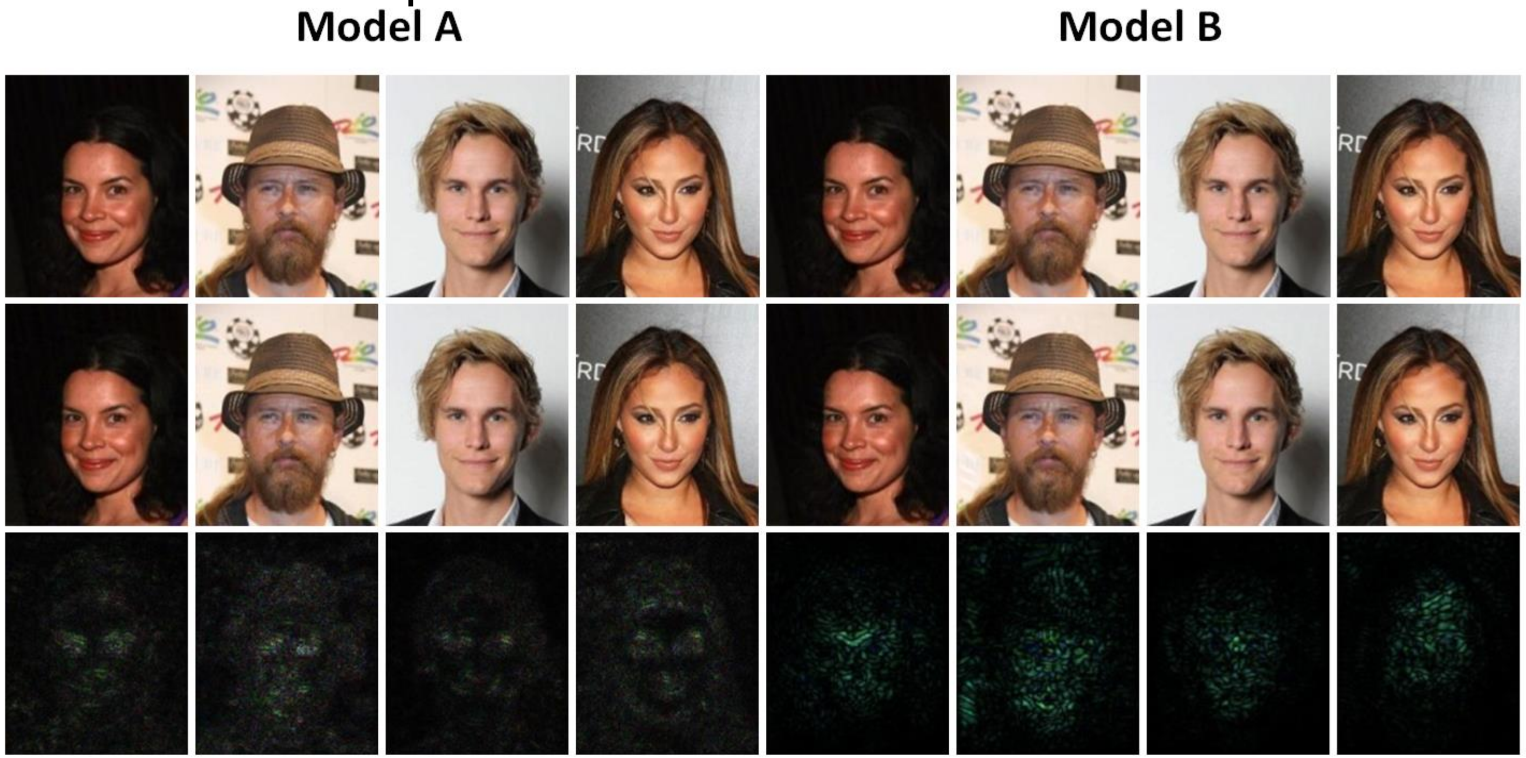
$$\mathbf{x}_p^0 = \mathbf{x},$$

$$\mathbf{x}_p^{i+1} = clip_{[0,1]}(\mathbf{x}_p^i - \alpha \cdot \nabla_x l_f(\mathbf{x}_p^i, t))$$

- where α is the step size, $\mathbf{x}$ is the original image, $\mathbf{x}_p^i$ is the adversarial image at step i, $\nabla_x l_f(\mathbf{x}_p^{i+1}, t)$ is the first-order gradient term of the adversarial loss,t is the target class label and $clip_{[a,b]}$ is a constraint that keeps pixel values in the $[a, b]$ range.

Adversarial Face De-Identification

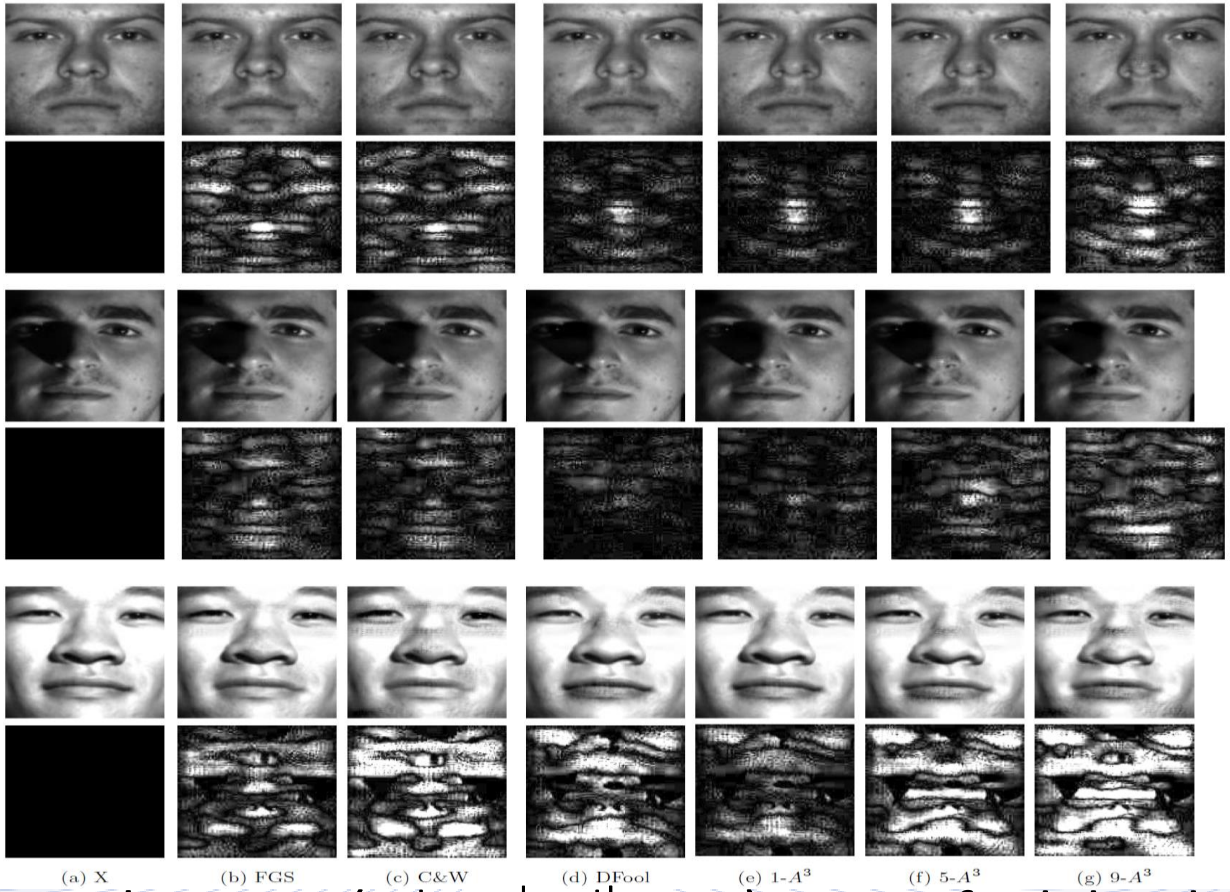
First row: original image.
Second row: de-identified image.
Third row: adversarial perturbation absolute value 10x.



k-anonymity-inspired adversarial attack

- **k-anonymity concept:**
  - The maximum probability of retrieving a sample from a set must be less than $1/k$.
  - Originally introduced in other research areas (e.g., Databases)
- In k-anonymity-inspired adversarial attack, the concept is altered as follows:
- The maximum probability of retrieving the real identity of a subject, must be less than $1/k$, in every possible classifier output ranking position.

$k - A^3$ face de-identification method



Face de-identification: original images (1st, 3rd, 5th row), magnified de-identification noise for various methods (2nd, 4th, 6th row, $k - A^3$ 3 right columns).

Adversarial Machine Learning

- Adversarial Examples
- Adversarial Attacks
- Adversarial Face De-identification
- **Adversarial Defenses**

Adversarial Defenses

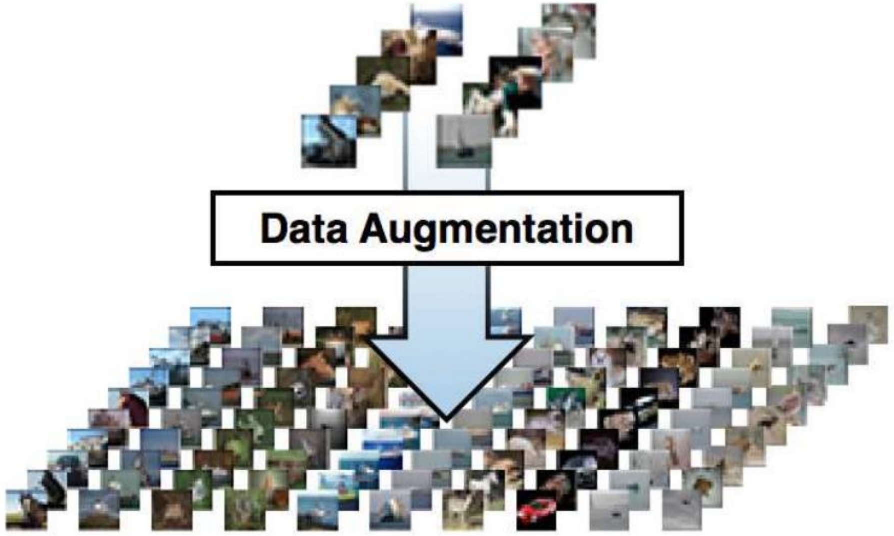
- Introduce changes to the model parameters or architecture, in order to improve its robustness against adversarial attacks:
- The robustness is measured as follows:
  - $\mathbf{x}_p$ eventually fails to deceive the model within a noise range e,
    $\hat{y} = f\left(\mathbf{x}_p, \widetilde{\boldsymbol{\theta}}\right) = y_{true}$ if $||\mathbf{x}_p - \mathbf{x}||^2 < e$
  - Successful adversarial attacks against the defended model are noisier than the ones fooling the undefended model:
    $||\tilde{\mathbf{x}}_{p_{defended}} - \mathbf{x}||^2 > ||\mathbf{x}_p - \mathbf{x}||^2$, when $\hat{y} = f\left(\tilde{\mathbf{x}}_p, \widetilde{\boldsymbol{\theta}}\right) = f\left(\mathbf{x}_\mathrm{p}, \boldsymbol{\theta}\right)$

Adversarial Defenses

- **Input filtering**:
  - Apply filters to the input [LIA2018]. Does not address the robustness of the actual model.
- **Gradient masking:**
  - Knowledge transfer from a trained CNN to another [PAP2016]. Defended model cannot produce strong adversarial attacks. However, transferability and black box attacks still work.

Adversarial Training

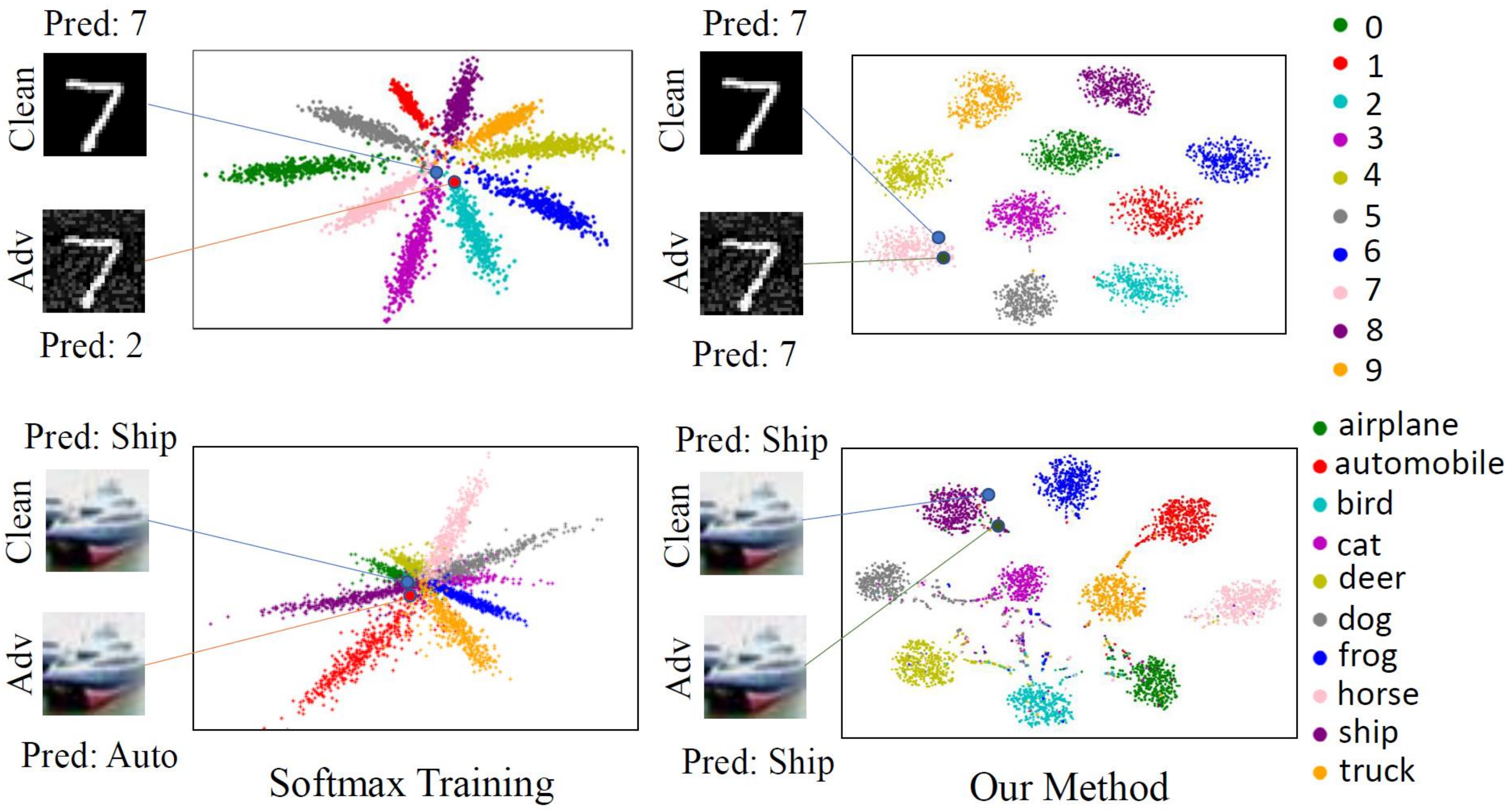
- Mixing training/validation/testing data sets with adversarial examples.
- Same idea with data augmentation as a regularization technique.



PCL Adversarial defense

- Assumption: The main reason for the existence of adversarial perturbations is the ***close proximity*** of different class samples in the learned feature space.
- Prototype Conformity Loss forces the features for each class to lie inside a convex polytope that is maximally separated from the polytopes of other classes.

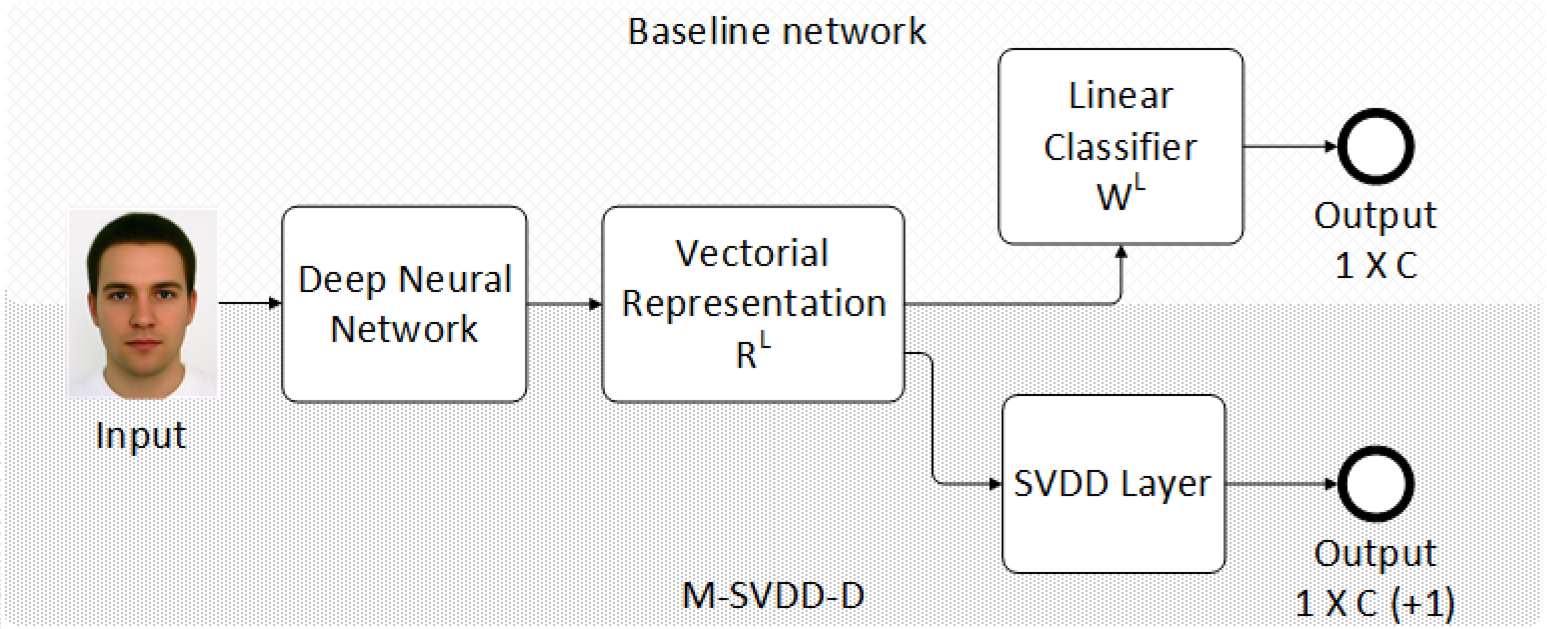
PCL Adversarial Defense



Mustafa, Amir et al. "Adversarial Defense by Restricting the Hidden Space of Deep Neural Networks", International Conference on Computer Vision (ICCV), 2019

Artificial Intelligence & Information Analysis Lab

m-SVDD Adversarial Defense



Mygdalis, Vasileios et al. “K-Anonymity-inspired Adversarial Attack and Multiple One-class Classification Defense”, Neural Networks, Elsevier, 2020.

Bibliography

[CHA2019] E. Chatzikyriakidis,et al, "Adversarial Face De-Identification" in Proceedings of the IEEE International Conference on Image Processing (ICIP), 2019

[MYG2020] V. Mygdalis, et al, “K-anonymity inspired Adversarial Attack and M-SVDD Defense”, Neural Networks, Elsevier, vol. 124, pp. 296-307, 2020

[DAL2004] N. Dalvi, et al. "Adversarial classification“, Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2004.

[BIG2013] N. Battista, et al. "Evasion attacks against machine learning at test time“, Joint European conference on machine learning and knowledge discovery in databases. Springer, Berlin, Heidelberg, 2013.

[SZE2013] C. Szegedy, et al. "Intriguing properties of neural networks“, arXiv preprint arXiv:1312.6199, 2013.

[GOO2014] I. Goodfellow, et al. "Explaining and harnessing adversarial examples“, arXiv preprint arXiv:1412.6572, 2014.

[MOO2017] Moosavi-Dezfooli et al., “Universal adversarial perturbations”, Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), 2017.

[KUR2016] A. Kurakin et al., “Adversarial Machine Learning at Scale”, arXiv preprint arXiv:1611.01236, 2016.

[ATH2017] A. Athalye et al., “Synthesizing Robust Adversarial Examples”, arXiv preprint arXiv:1707.07397, 2017.

[LIU2017] Y. Liu et al., “Delving into Transferable Adversarial Examples and Black-box Attacks”, arXiv preprint arXiv:1611.02770, 2017.

Bibliography

[CAR2017] N. Carlini et al., “Adversarial Examples Are Not Easily Detected: Bypassing Detection Methods”, Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security, 2017.

[KUR2018] A. Kurakin et al., “Adversarial Attacks and Defences Competition”, The NIPS'17 Competition: Building Intelligent Systems. Springer, 2018.

[KUR2016] A. Kurakin et al, "Adversarial examples in the physical world." arXiv preprint arXiv:1607.02533, 2016.

[KAN2018] H. Kannan et al., “Adversarial Logit Pairing”, arXiv preprint arXiv:1803.06373 , 2018.

[LIA2018] F. Liao, et al. “Defense against adversarial attacks using high-level representation guided denoiser”, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018.

[PAP2016] N. Papernot et. al. “Distillation as a defense to adversarial perturbations against deep neural networks”, In IEEE Symposium on Security and Privacy (SP) (pp. 582-597), 2016

[MAD2017] A. Madry, et al. “Towards deep learning models resistant to adversarial attacks”, arXiv preprint arXiv:1706.06083, 2017.

[MUS2019] A. Mustafa, et al. “Adversarial Defense by Restricting the Hidden Space of Deep Neural Networks”, International Conference on Computer Vision (ICCV), 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.