Introduction to Adversarial Machine Learning

Georgia Fargetta

Erasmus +
Master Degree in Computer Science
University of Rouen, France

IFOSS

International Forensics Summer School

ETHICAL AND LEGAL CHALLENGES IN AI-DRIVEN FORENSIC SCIENCE

JULY 14-20, 2024

Watch a preview!

School Directors

PROF. SEBASTIANO BATTIATO, PH.D.
University of Catania

PROF. DONATELLA CURTOTTI, PH.D.
University of Foggia

PROF. GIOVANNI ZICCARDI, PH.D.
University of Milan

Speakers

others coming soon..

ALESSANDRO TRIVILINI
Scuola universitaria professionale della svizzera italiana (SUPSI)

MARTIN DRAHANSKÝ
Faculty of Information Technology, Brno University of Technology

PROF. DR. DIDIER MEUWLY
University of Twente

School location

The school will take place at Sampieri, Sicily
https://www.hotelbaiasamuele.it/en/

Social Network

IFOSS

@ifoss_official

@ifoss_official

IFOSS

www.ifoss.it

info@ifoss.it

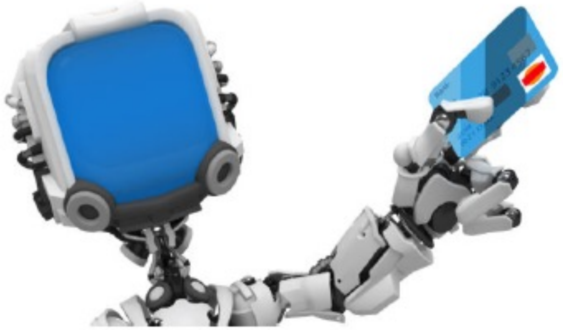
The Success of Machine Learning!

Autonomous Driving

Healthcare

Smart City

Malware Classification



Fraud Detection

Biometrics Recognition

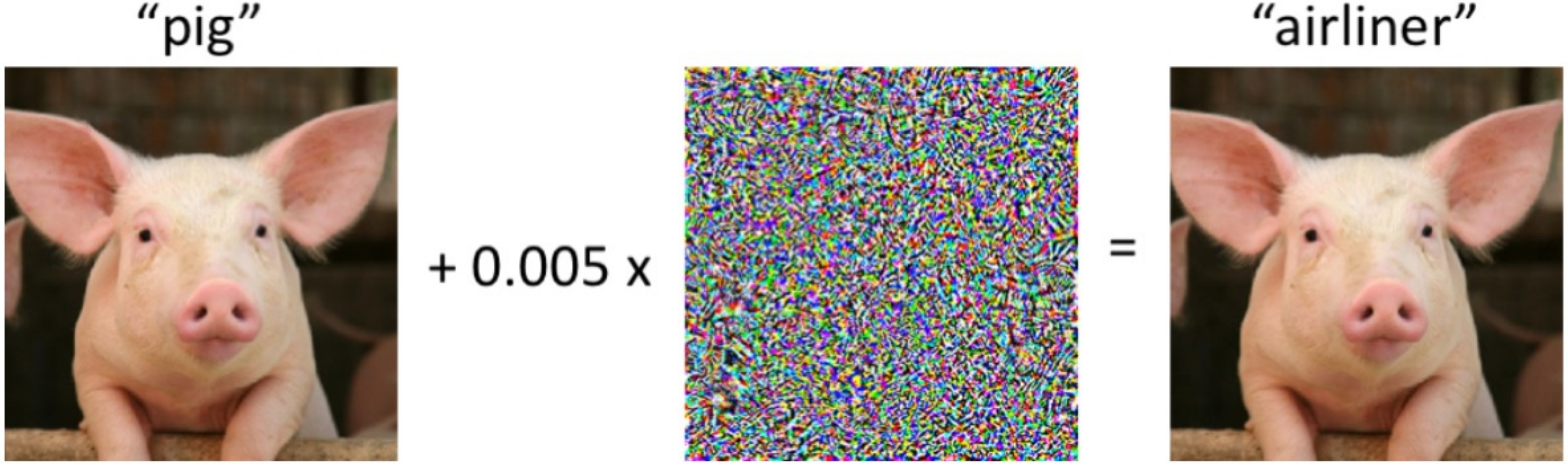
Credits: Bo Li, Secure Learning in Adversarial Deep Neural Networks

The Success of Machine Learning!

Is ML truly ready for
real-world deployment?

The Success of Machine Learning!

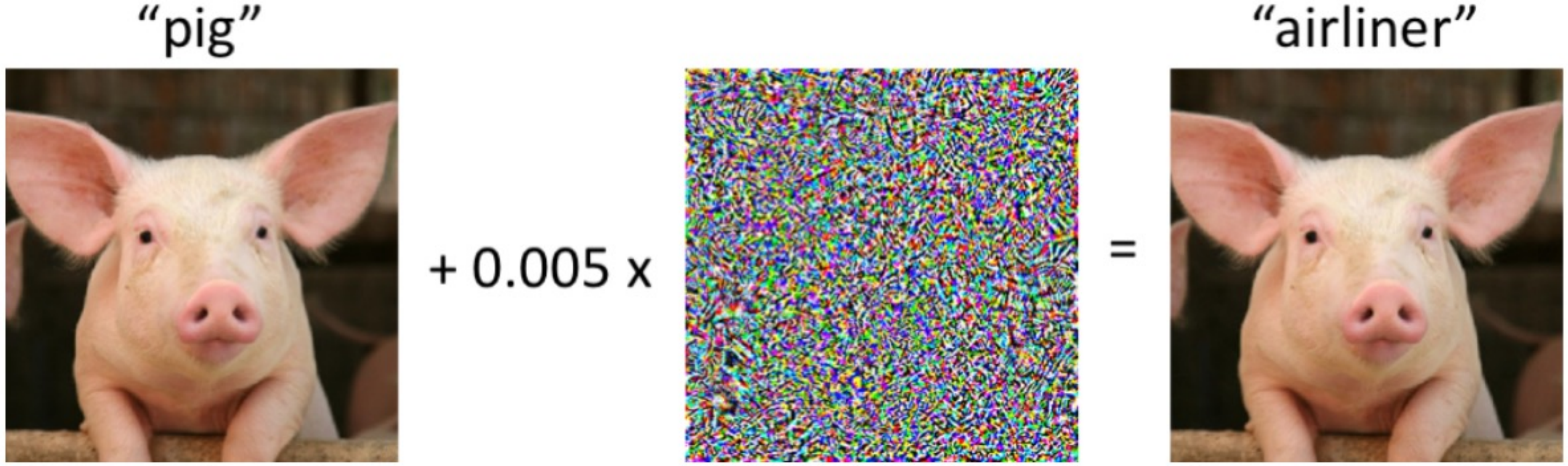
Is ML truly ready for real-world deployment?



Credits: Z. Kolter, A. Mądry- *Adversarial Robustness: Theory and Practice*

Adversarial Machine Learning

Adversarial machine learning is a series of techniques that aim to undermine machine learning performances, through the definition of special inputs (adversarial examples) that an attacker has intentionally designed to cause the model to make a mistake.



Over the past few years, adversarial examples have received a significant amount of attention in the deep learning community.

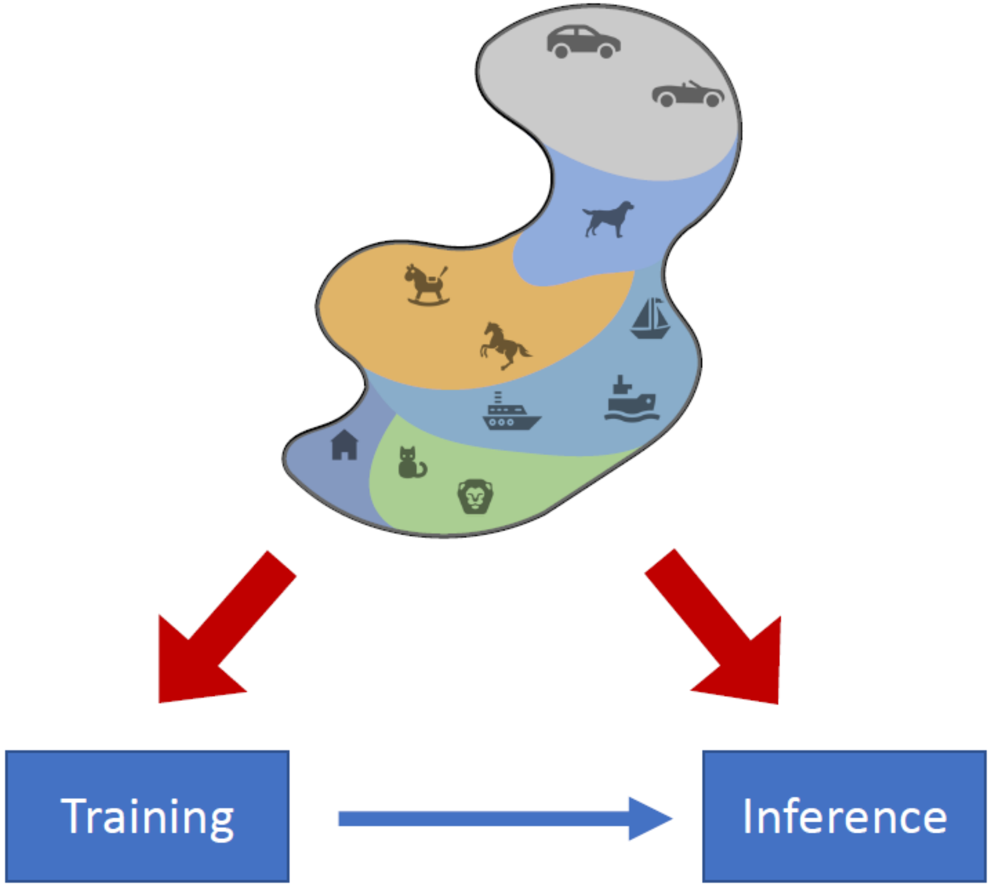
Credits: Z. Kolter, A. Mądry- *Adversarial Robustness: Theory and Practice*

Adversarial Machine Learning

“Adversarial attacks are manipulative actions that aim to undermine machine learning performance, cause model misbehaviour, or acquire protected information… “

Pin-Yu Chen, chief scientist, RPI-IBM AI research collaboration at IBM Research, told The Daily Swig.

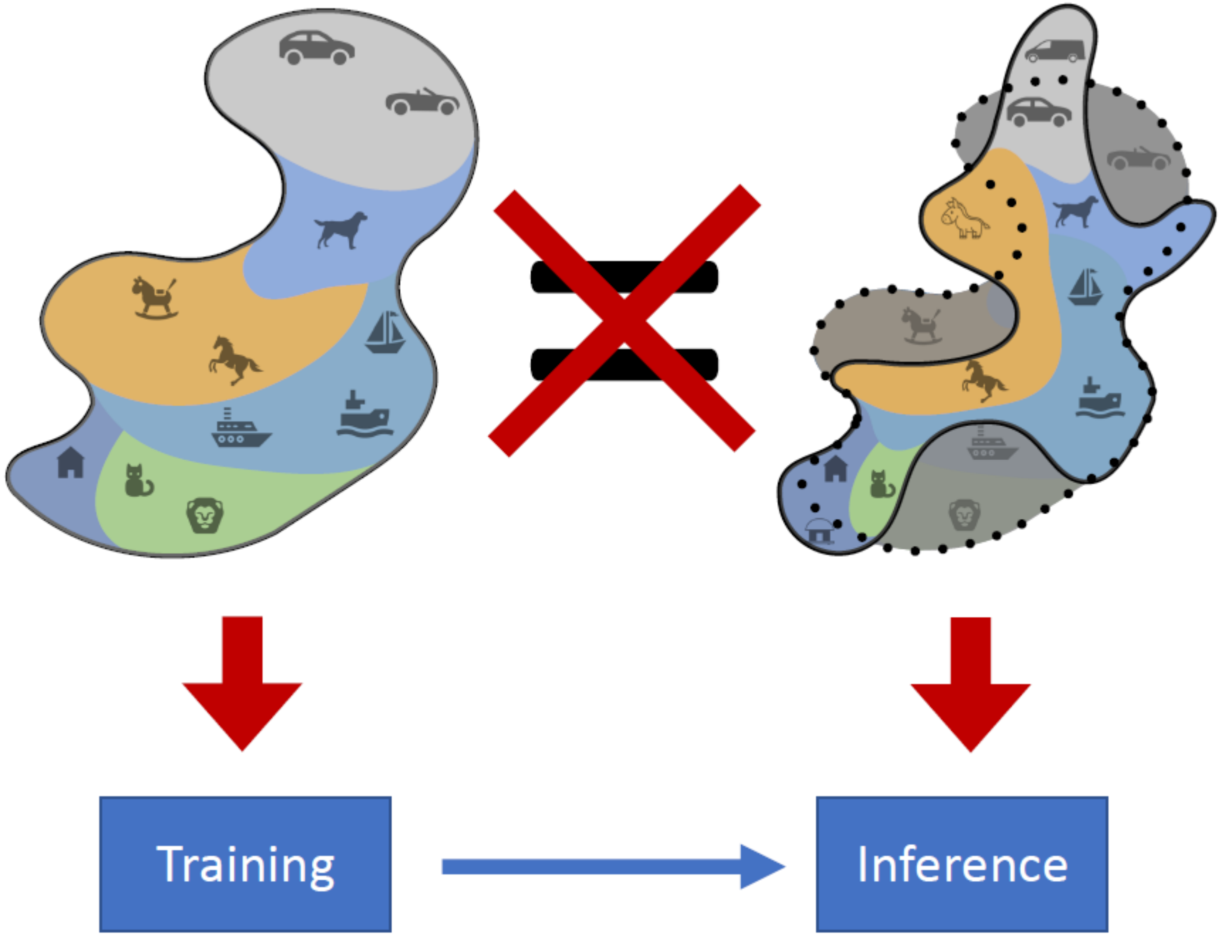
Adversarial Machine Learning



Traditional machine learning approaches assume that training data and test data belongs to the same distribution.

Credits: Z. Kolter, A. Mądry- *Adversarial Robustness: Theory and Practice*

Adversarial Machine Learning



Traditional machine learning approaches assume that training data and test data belongs to the same distribution.

But: In reality, the distributions we use ML on are NOT the ones we train it on!

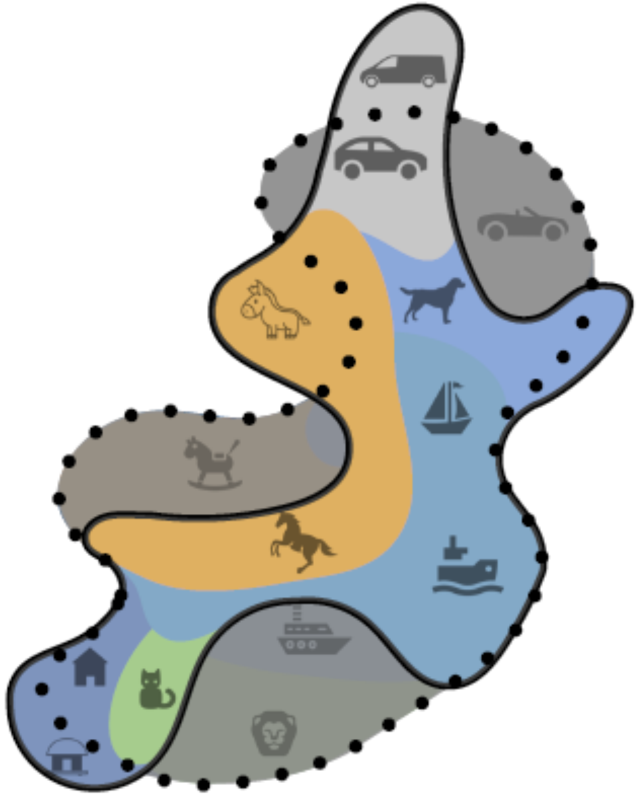
Credits: Z. Kolter, A. Mądry- *Adversarial Robustness: Theory and Practice*

Adversarial Machine Learning

DNNs are powerful because their many layers mean they can pick up on patterns in many different features of an input when attempting to classify it.

But this also means that a ***very small change in the input*** can tip it over into what the model considers an apparently different state.

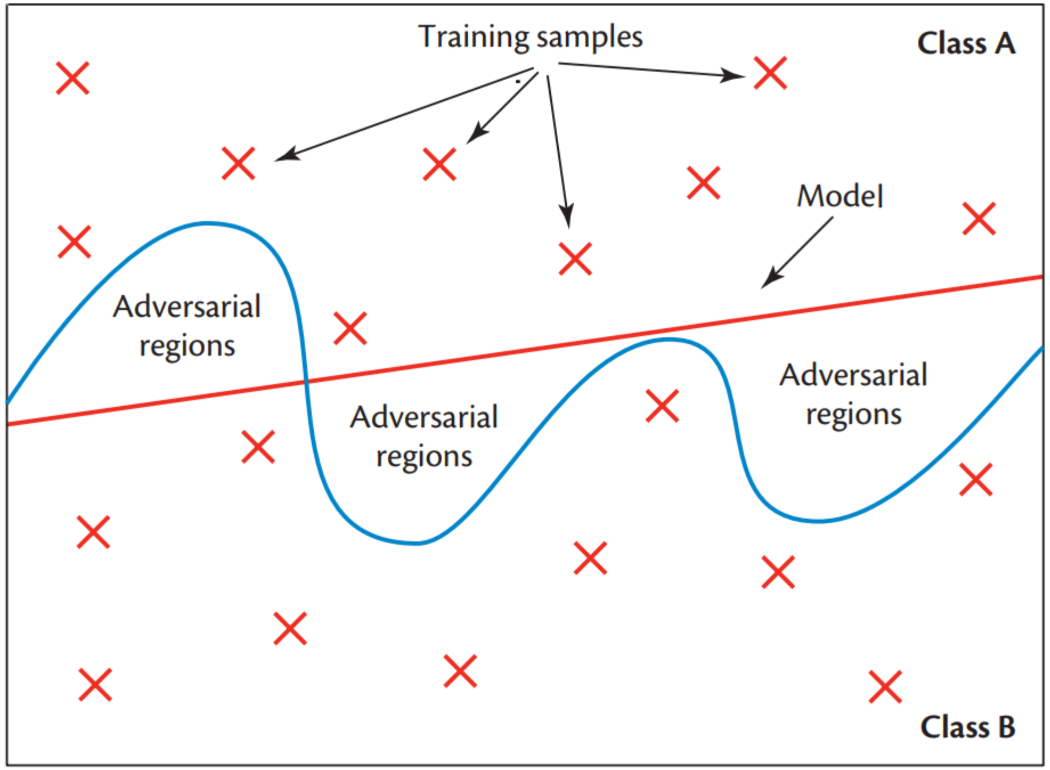
The types of perturbations applied in adversarial attacks depend on the target data type and desired effect.



Adversarial Machine Learning

The training defines a *model decision boundary* (red line) which only ***approximates*** the *real decision boundary* (blue line).

- This model has high (mean) accuracy.
- Real decision boundary becomes more complex as the feature dimension space becomes larger.
- In deep ANN the feature space is a combination of very numerous non linear functions.



McDaniel, Patrick, Nicolas Papernot, and Z. Berkay Celik. "Machine learning in adversarial settings." *IEEE Security & Privacy* 14.3 (2016): 68-72.

Adversarial Machine Learning

Constructing adversarial examples

Training a classifier is often formulated as finding model parameters *ϑ* that minimize an empirical loss function for a given set of samples $x_1,\ldots,x_n$

$$\min_{\theta} \sum_{x} \text{loss}(x, \theta)$$

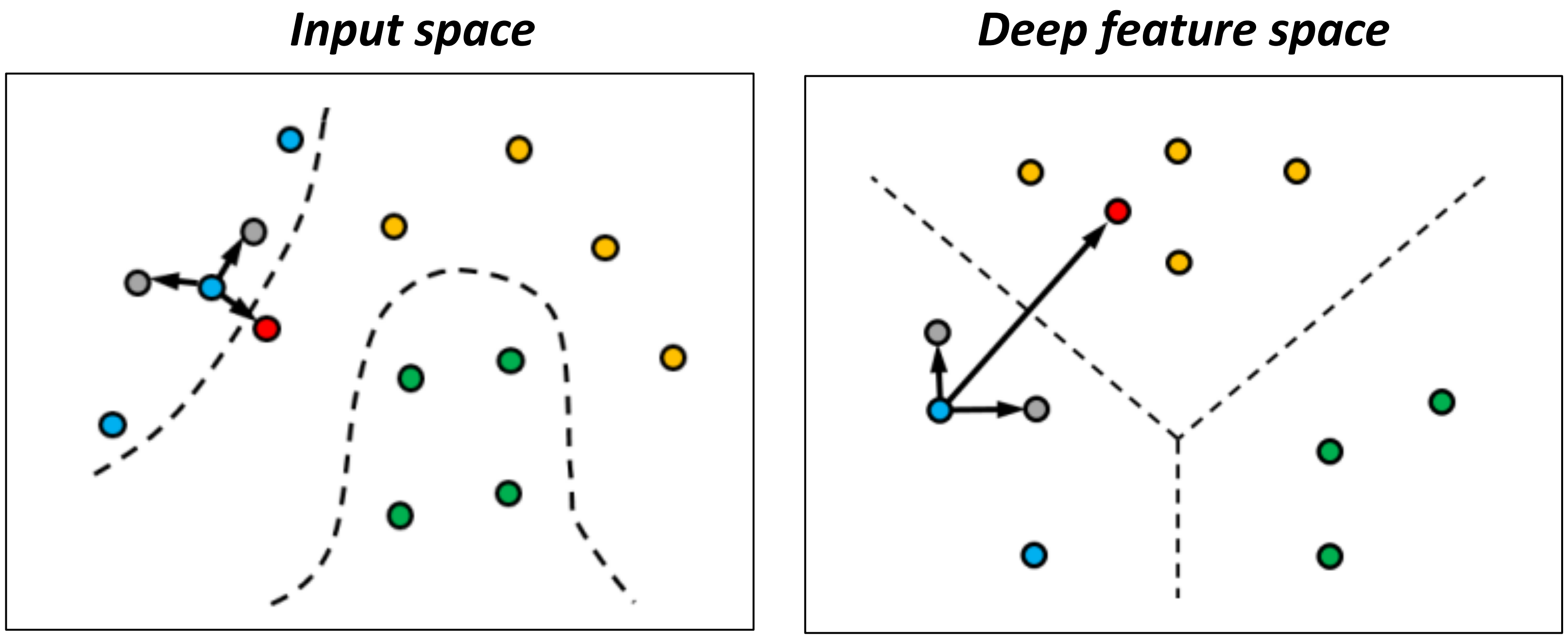
to cause a misclassification for a fixed model θ and "benign" input x, a natural approach is to find a bounded perturbation δ such that the loss on x+δ is as large as possible:

$$\max_{\delta} \text{loss}(x + \delta, \theta)$$

Adversarial Machine Learning

Despite gray/red points are at the same distance from x in the input space, the adversarial example (red) is much farther in the deep feature space.

Random perturbations in the input space only result in a very small shift in the input space, while even light alterations of an image *<u>along the adversarial direction</u>* cause a large shift in deep space.



Input space

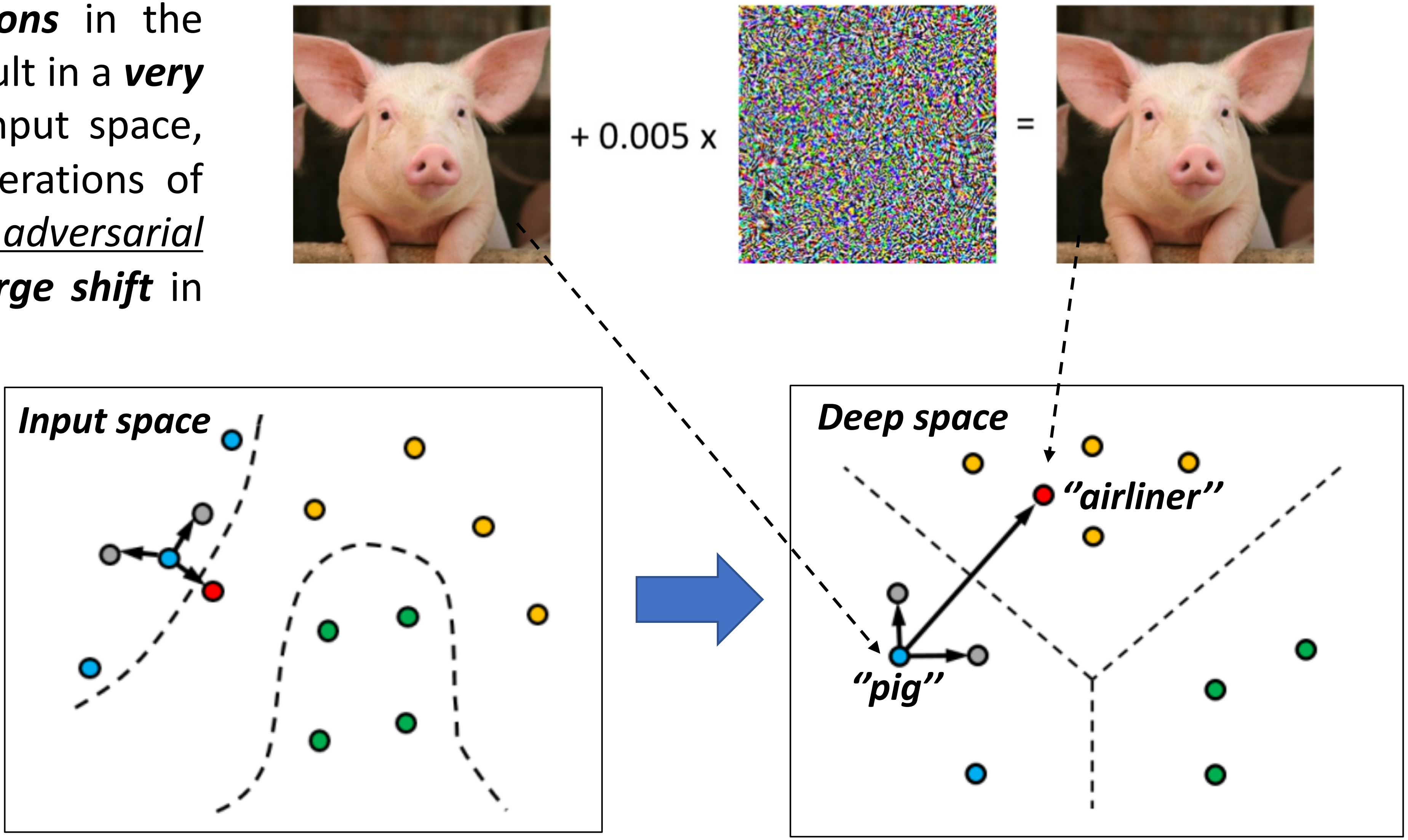
Deep feature space

Vulnerability of the deep feature space mapping.

Melis, Marco, et al. "Is deep learning safe for robot vision? adversarial examples against the icub humanoid." Proceedings of the IEEE International Conference on Computer Vision Workshops. 2017.

Adversarial Machine Learning

Random perturbations in the input space only result in a ***very small shift*** in the input space, while even light alterations of an image *along the adversarial direction* cause a ***large shift*** in deep space.

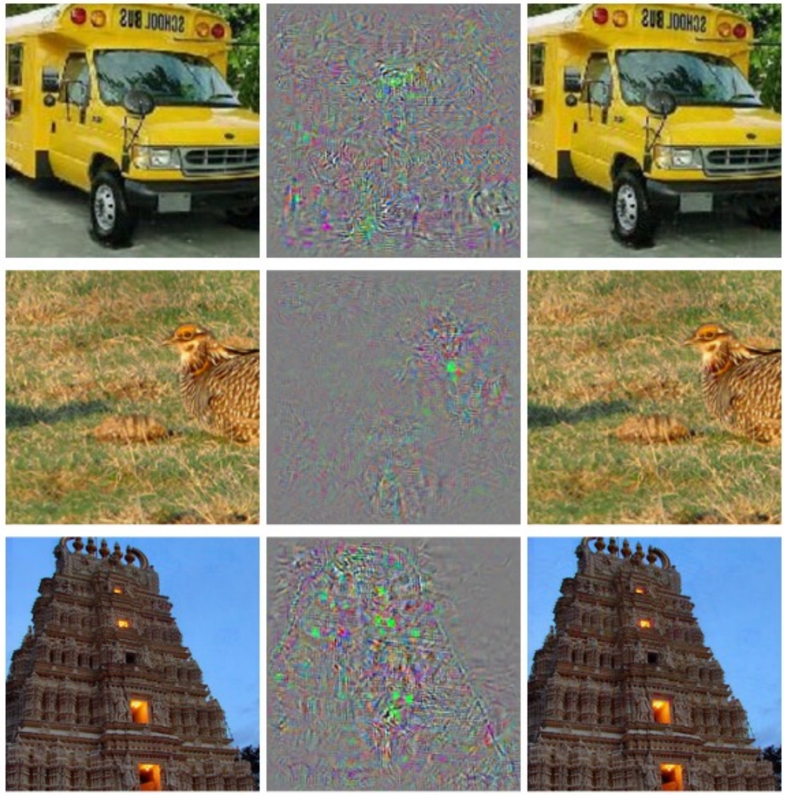


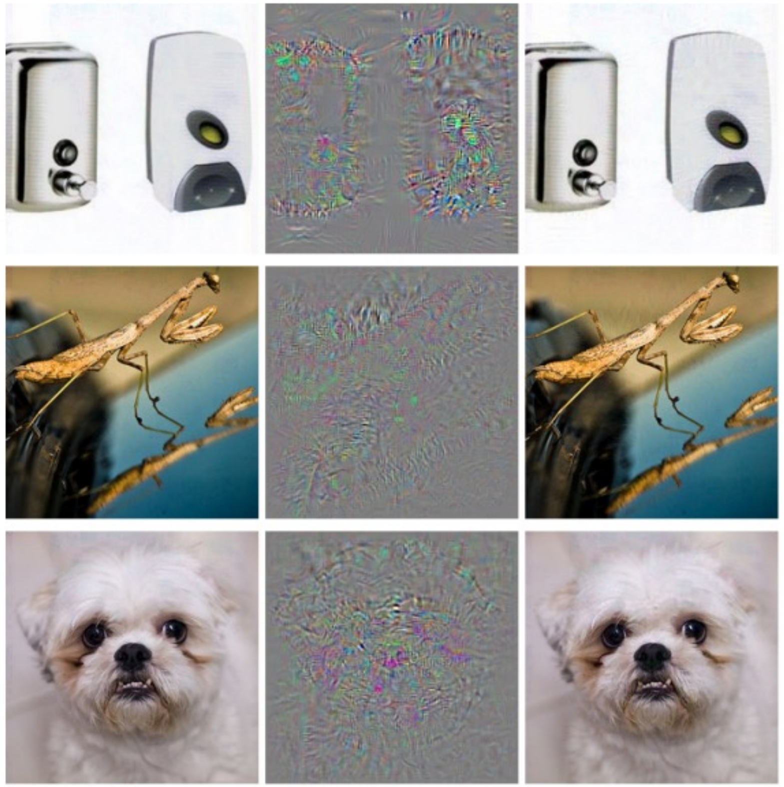
Melis, Marco, et al. "Is deep learning safe for robot vision? adversarial examples against the icub humanoid." ICCV Workshops. 2017.

Adversarial Machine Learning

Szegedy et al. first discovered that well-performing deep neural networks are susceptible to adversarial attacks.

All images in the right column are predicted to be an "ostrich".



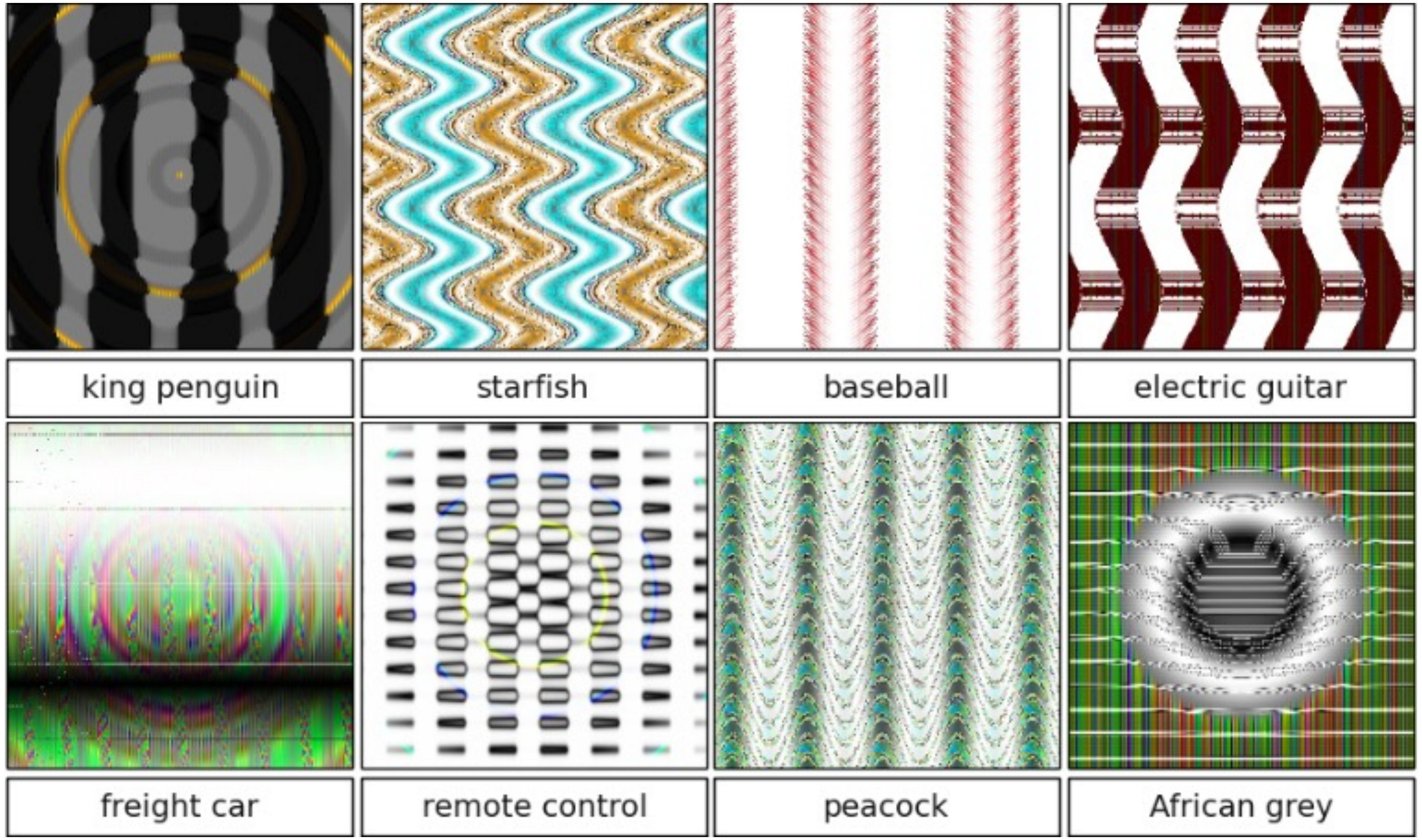


Szegedy, Christian, et al. "Intriguing properties of neural networks." arXiv preprint arXiv:1312.6199 (2013).

Adversarial Machine Learning

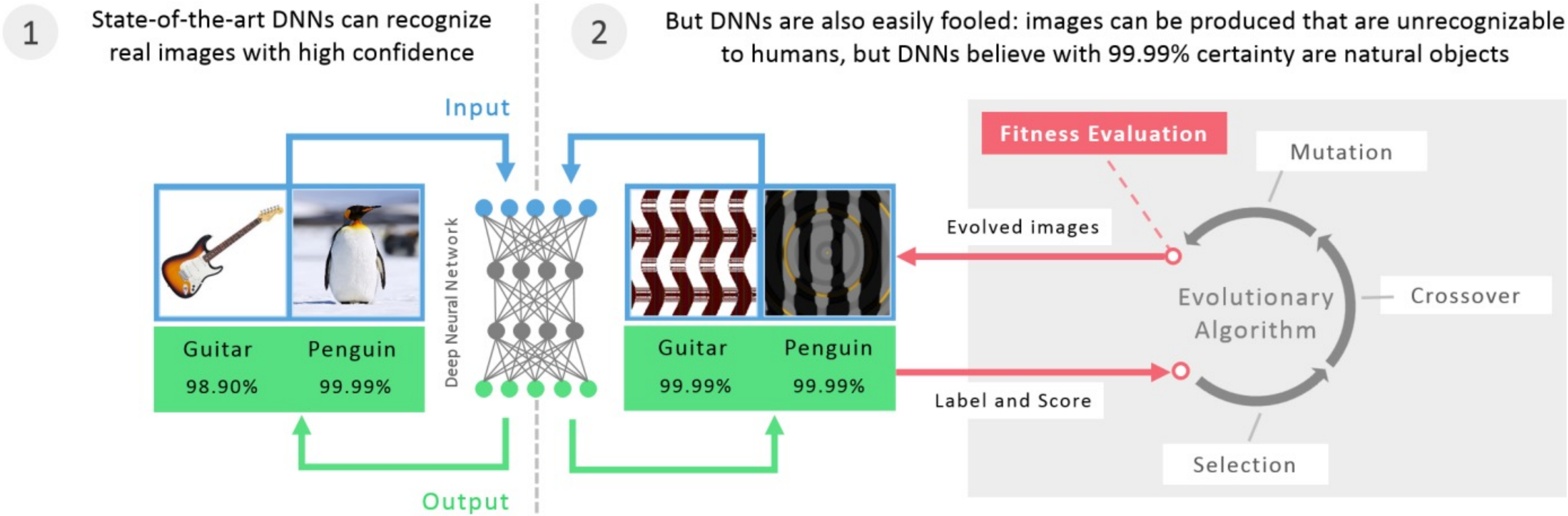
It is easy to produce images that are completely unrecognizable to humans.

99.99% confidence



Nguyen, Anh, Jason Yosinski, and Jeff Clune. "Deep neural networks are easily fooled: High confidence predictions for unrecognizable images." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

Adversarial Machine Learning

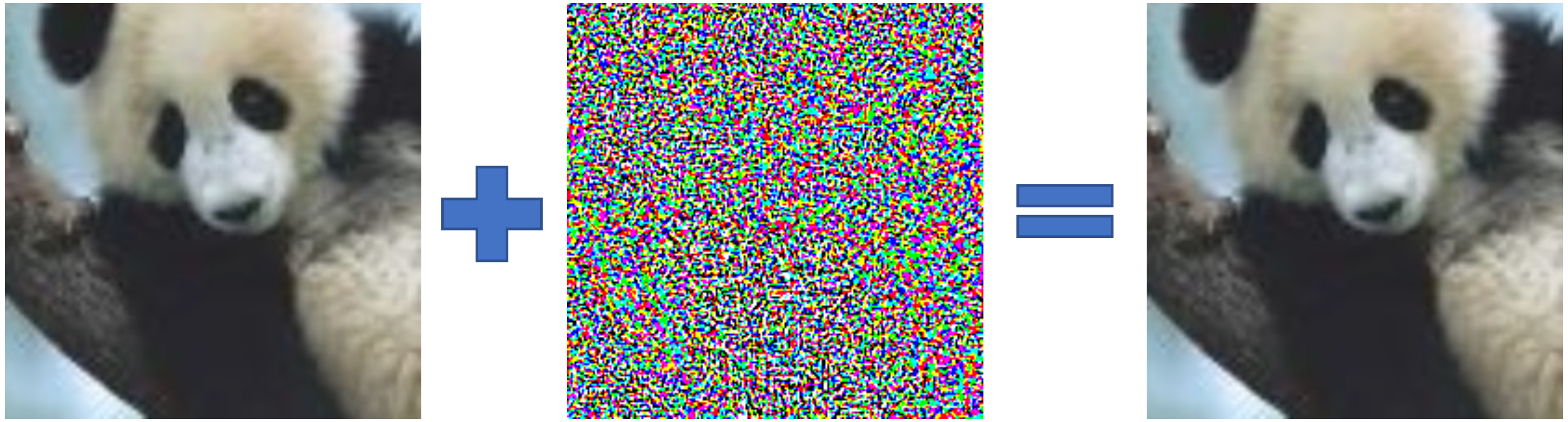


1 State-of-the-art DNNs can recognize real images with high confidence

2 But DNNs are also easily fooled: images can be produced that are unrecognizable to humans, but DNNs believe with 99.99% certainty are natural objects

Nguyen, Anh, Jason Yosinski, and Jeff Clune. "Deep neural networks are easily fooled: High confidence predictions for unrecognizable images." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

Adversarial Machine Learning



Classified as panda

57.7% confidence

Small adversarial noise

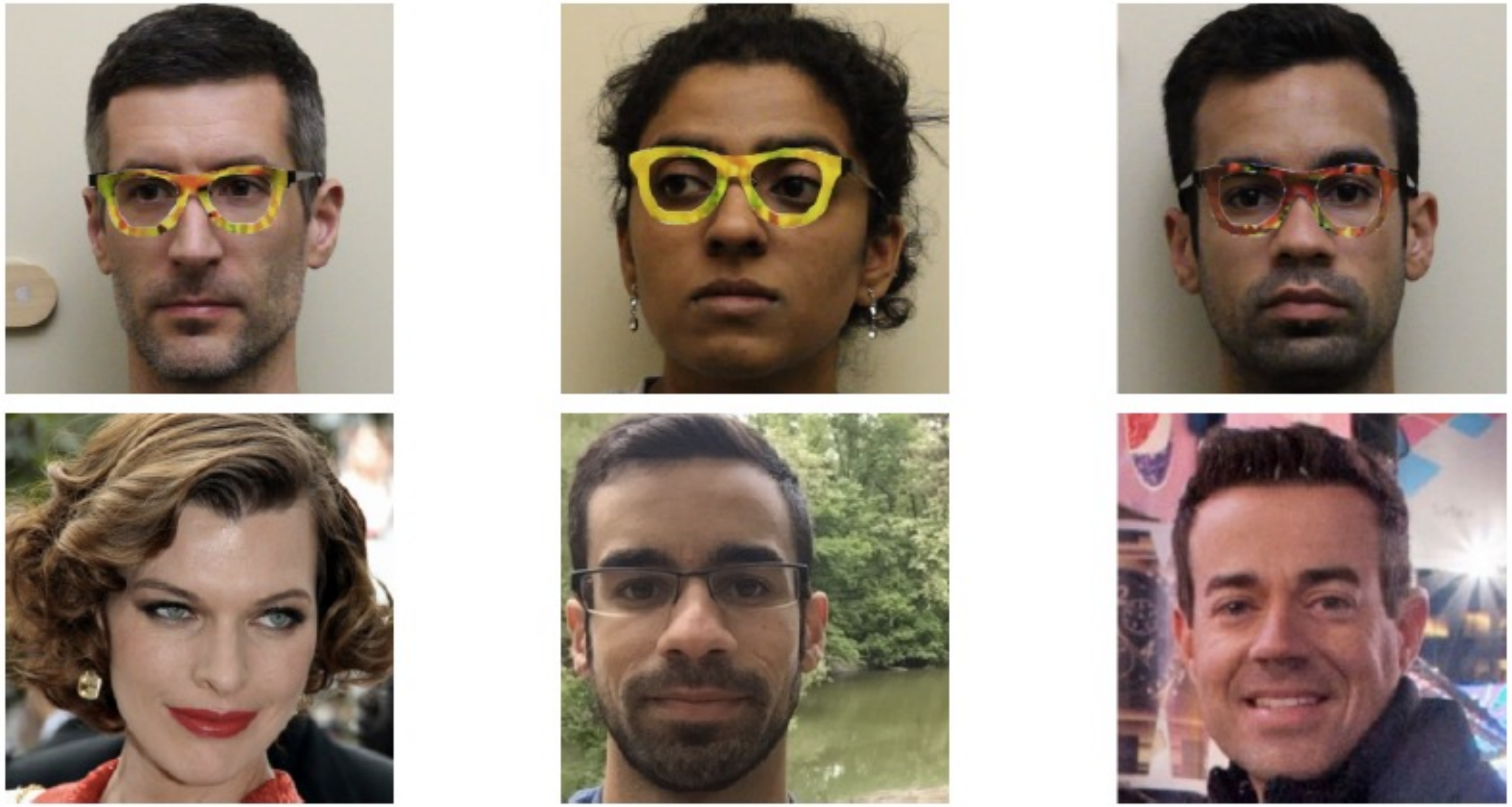
Classified as gibbon

99.3 % confidence

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

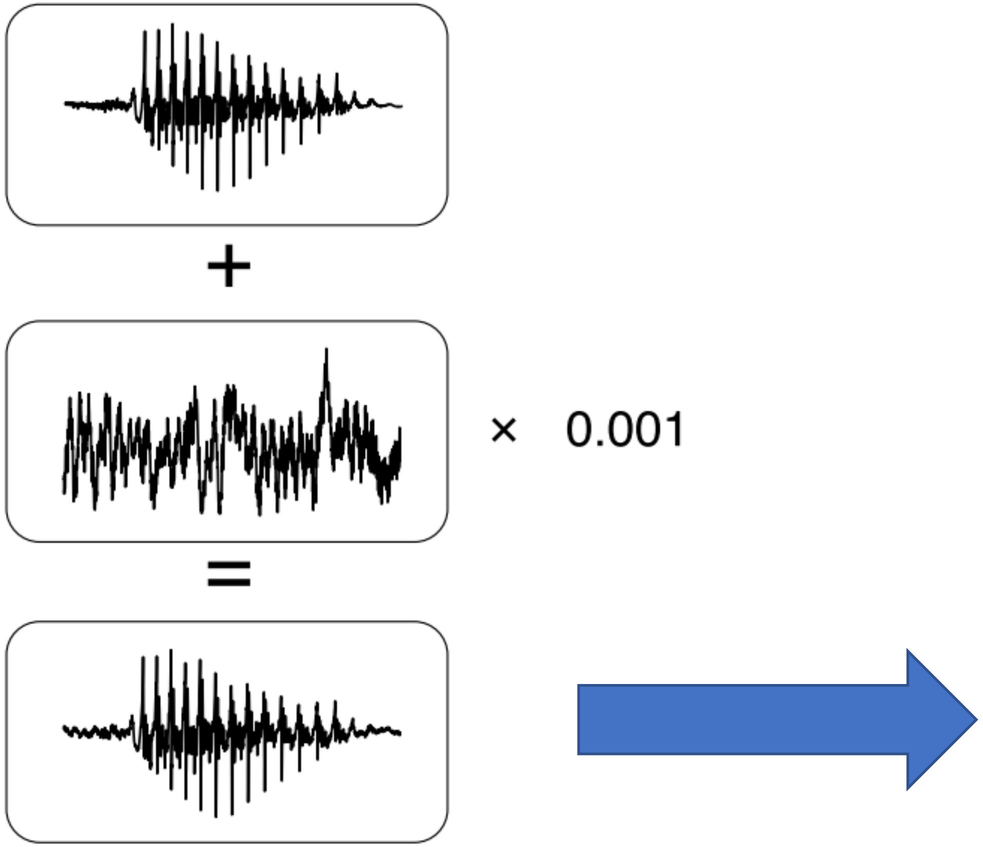
Adversarial Machine Learning

Glasses that fool face recognition



Ren, Kui, et al. "Adversarial attacks and defenses in deep learning." *Engineering* 6.3 (2020): 346-360.

Adversarial Machine Learning


Original audio

+

Perturbation × 0.001

=

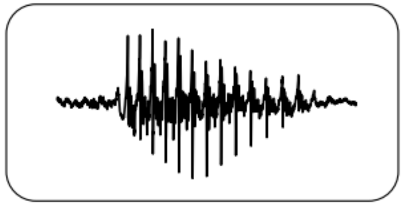
Adversarial audio

Multimedia content

Carlini, Nicholas, and David Wagner. "Audio adversarial examples: Targeted attacks on speech-to-text." *2018 IEEE SPW.*

Adversarial Machine Learning

Adversarial attacks against speech recognition systems



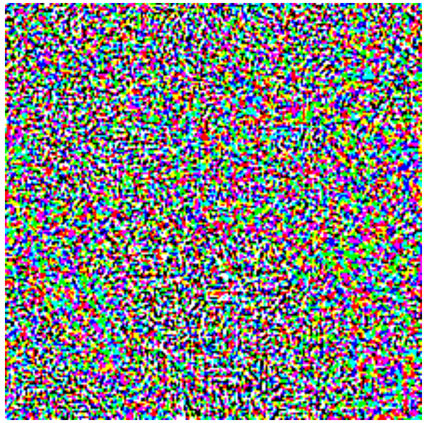
"Ok! I'm opening the door...

Carlini, Nicholas, and David Wagner. "Audio adversarial examples: Targeted attacks on speech-to-text." *2018 IEEE SPW.*

Adversarial Machine Learning

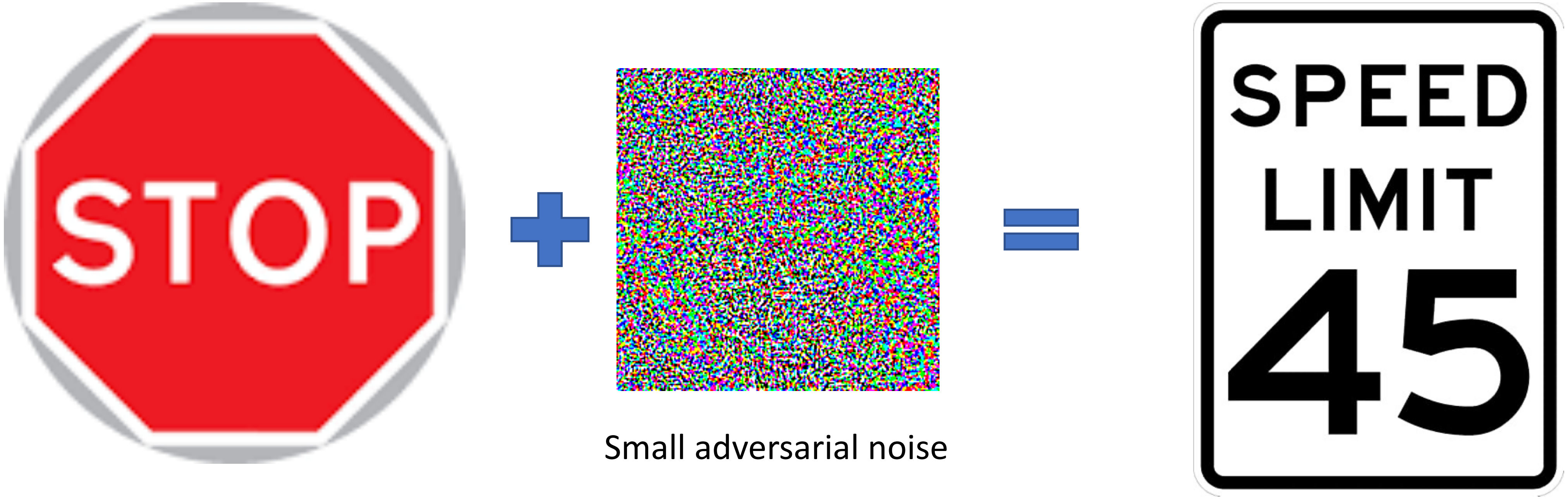
Adversarial Machine Learning

+



Small adversarial noise

Adversarial Machine Learning

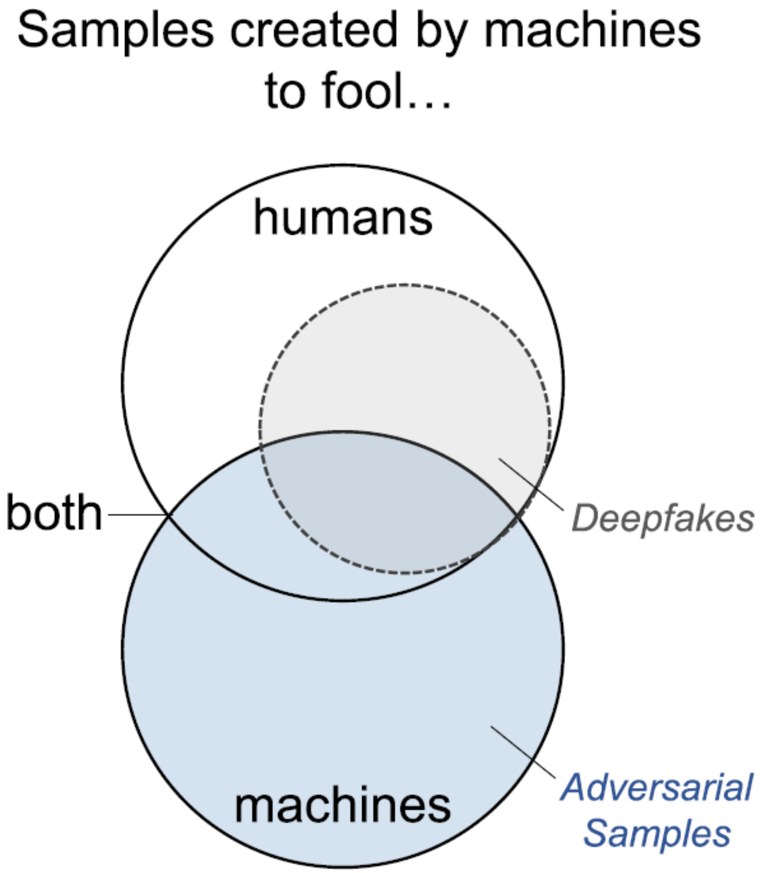


Small adversarial noise

Adversarial Machine Learning

Deepfakes vs. Adversarial Machine Learning

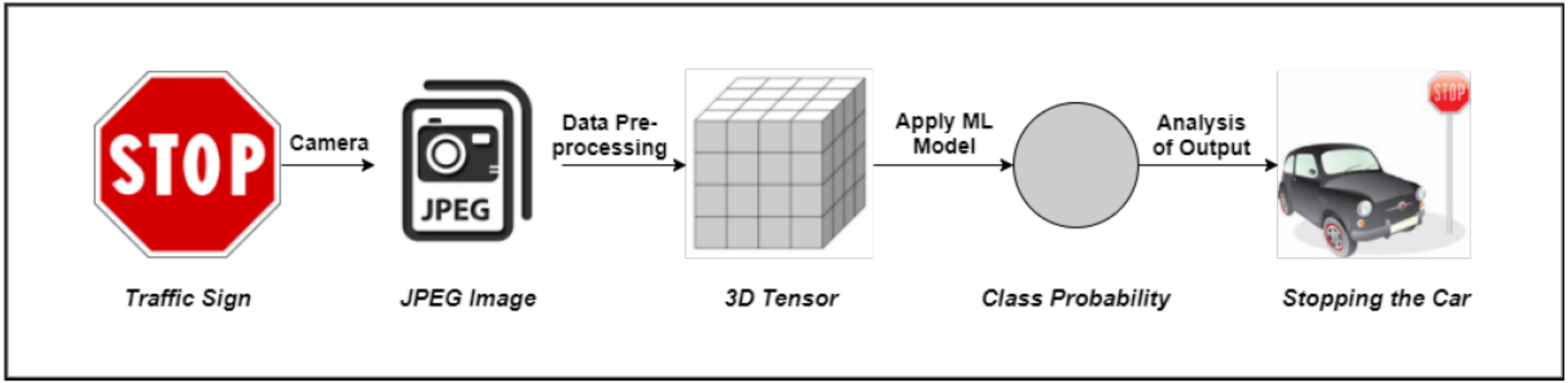
In Adversarial Machine Learning the objective of the generated content is to fool a machine (i.e., an algorithm) and not a human.



Examples:
...**humans**: entertainment, impersonation, art fraud.
...**machines**: hiding a stop sign, evading face recog.
...**both**: tampering medical scans, malware evasion.

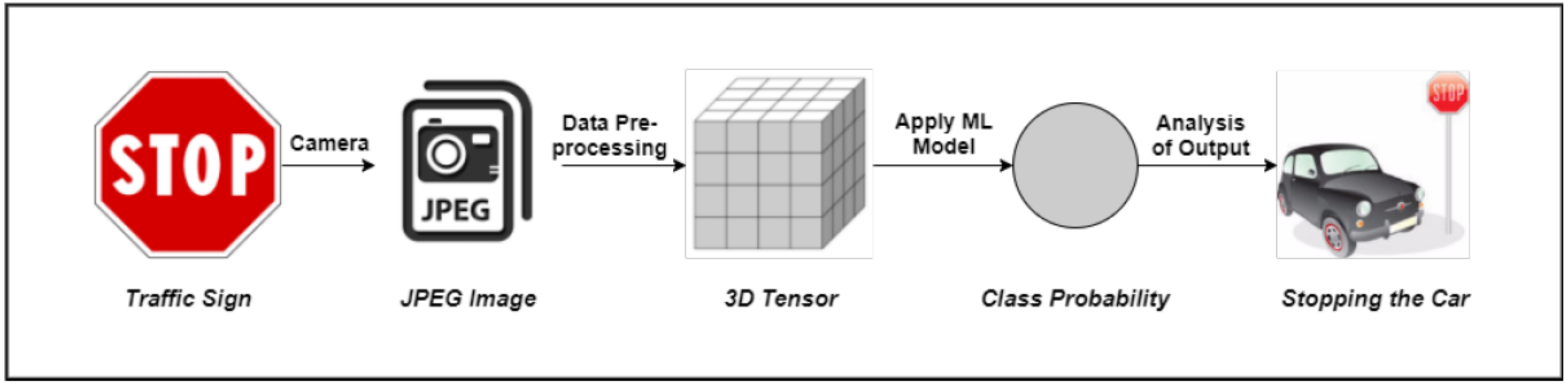
Adversarial Machine Learning

The attack surface can be defined with respect to the data processing pipeline.



An adversary can attempt to manipulate either the collection or the processing of data to corrupt the target model, thus tampering the original output.

Adversarial Machine Learning



- ***Evasion Attack* -** the adversary tries to evade the system by adjusting malicious samples during testing phase. This setting does not assume any influence over the training data.

- ***Poisoning Attack* -** an adversary tries to poison the training data by injecting carefully designed samples to compromise the whole learning process.

- ***Exploratory Attack* -** try to gain as much knowledge as possible about the learning algorithm of the underlying system and pattern in training data.

Adversarial Machine Learning

Attack Scenarios:

- ***Evasion Attack -*** The adversary tries to evade the system by adjusting malicious samples during testing phase. This setting does not assume any influence over the training data.

- ***Poisoning Attack -*** an adversary tries to poison the training data by injecting carefully designed samples to compromise the whole learning process.

- ***Exploratory Attack -*** Given black box access to the model, they try to gain as much knowledge as possible about the learning algorithm of the underlying system and pattern in training data.

Adversarial Machine Learning

Evasion Attack (test time)

Evasion attacks consist of manipulating input data to evade a trained classifier at *test time* (e.g., manipulation of malware code to have the corresponding sample misclassified as legitimate, or manipulation of images to mislead object recognition).

Adversarial Machine Learning

Evasion Attack (test time)

Evasion attacks consist of manipulating input data to evade a trained classifier at *test time* (e.g., manipulation of malware code to have the corresponding sample misclassified as legitimate, or manipulation of images to mislead object recognition).

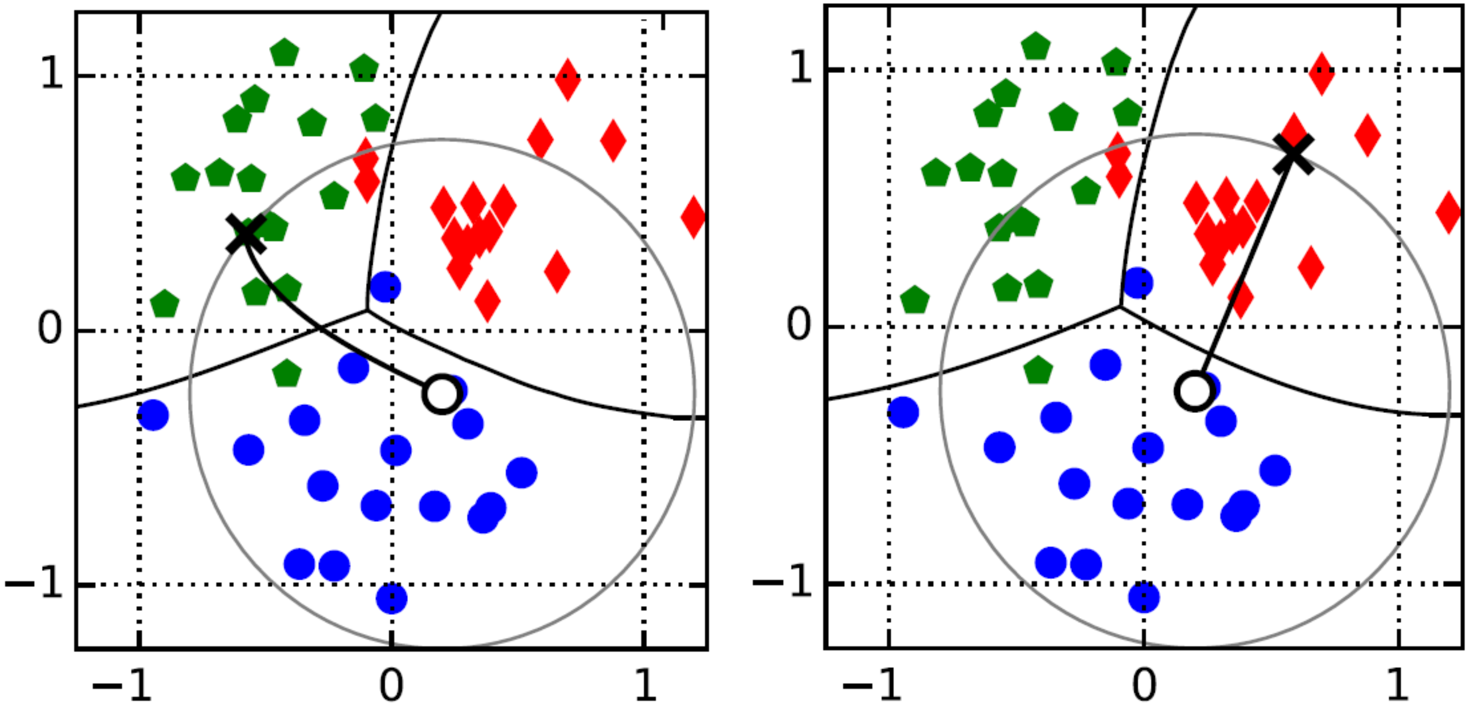
> ***Error generic:*** the attacker is interested in misleading classification, regardless of the output class predicted by the classifier.
>
> ***Error specific:*** the attacker requires the adversarial examples to be misclassified as a specific class.

Adversarial Machine Learning

Evasion Attack: error-generic vs. error-specific

The circle represents the feasible domain, given as an upper bound on the l_2 distance between the initial and the manipulated attack sample.

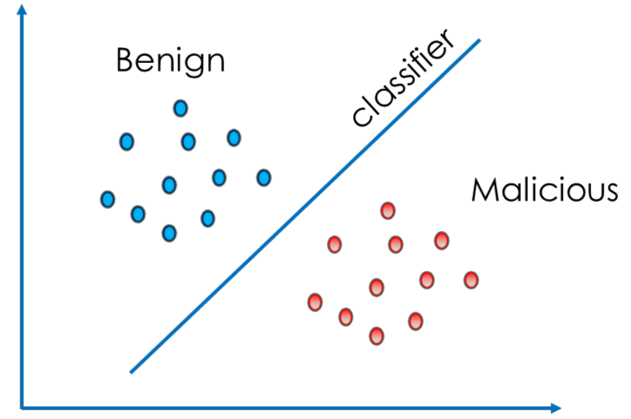
error-specific



error-generic

Biggio, Battista, and Fabio Roli. "Wild patterns: Ten years after the rise of adversarial machine learning." Pattern Recognition 84 (2018): 317-331.

AML – Evasion Attack



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

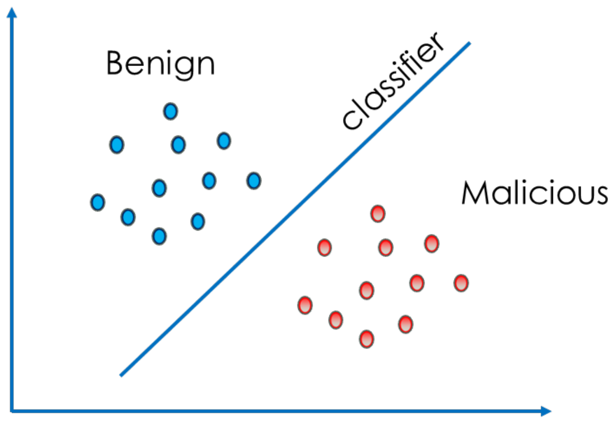
Feature Weights

2. cheap = 1.0
mortgage = 1.5

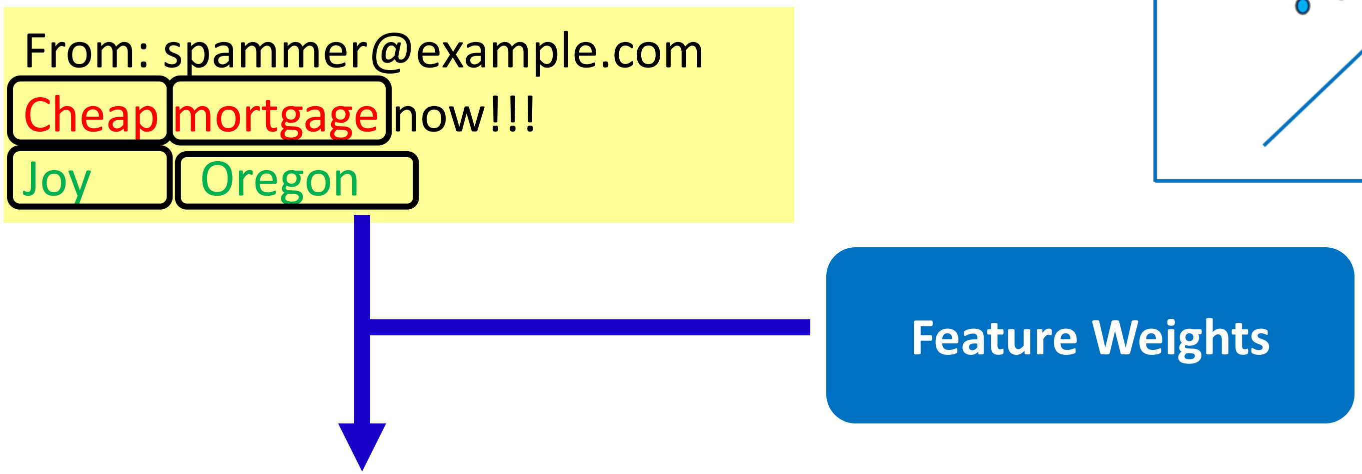
3. Total score = 2.5 > 1.0 (threshold)

Spam

AML – Evasion Attack



1.



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

2. cheap = 1.0
mortgage = 1.5
Joy= -1.0
Oregon = -1.0

3. Total score = 0.5 < 1.0 (threshold)

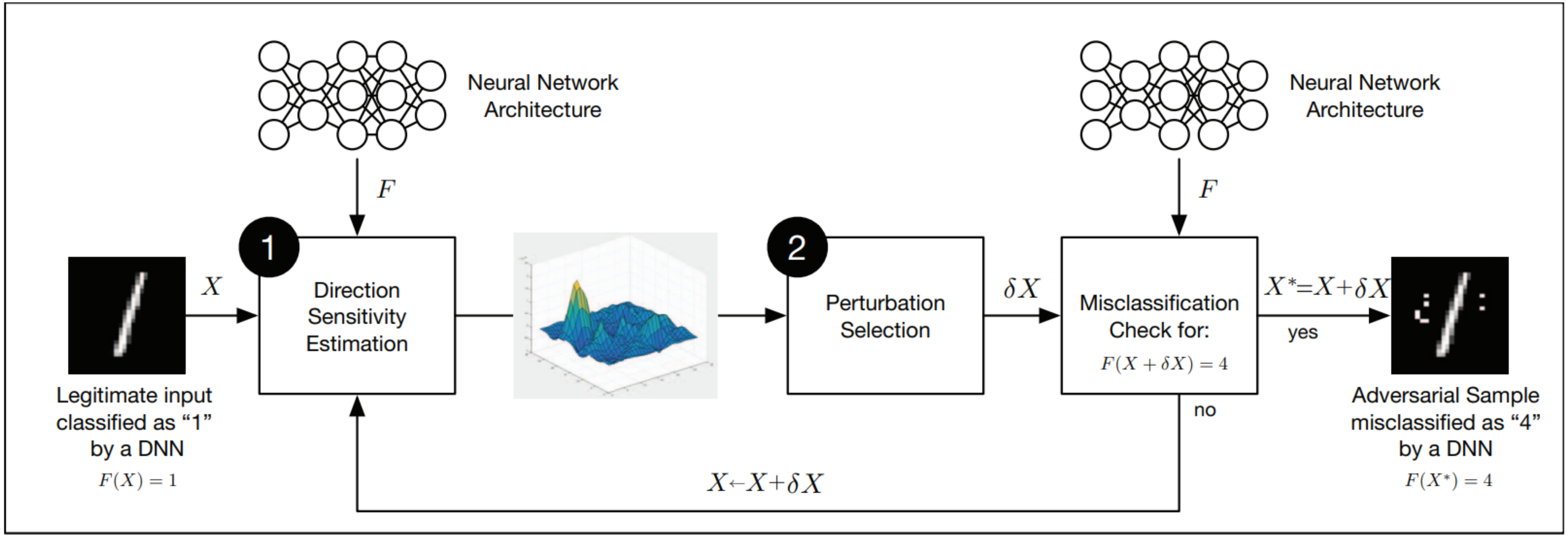
OK

AML – Evasion Attack

Craft a malicious example X* = X + δX by adding a perturbation δX, so that

$$F(X^*) = Y^*$$

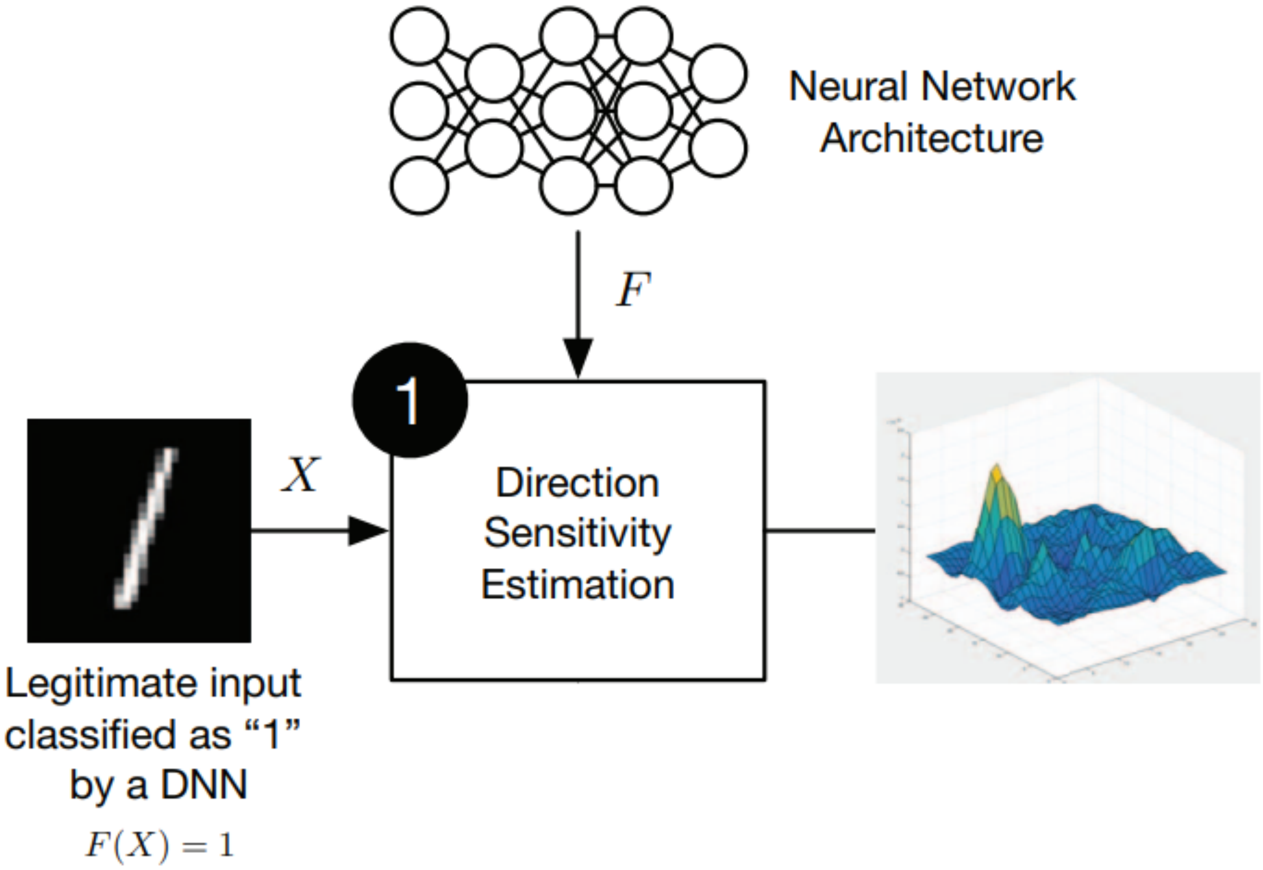
where Y* is the target output.


Neural Network Architecture
F
Neural Network Architecture
F
X
1
Direction Sensitivity Estimation
2
Perturbation Selection
δX
Misclassification Check for:
$F(X + \delta X) = 4$
$X^{*}{=}X{+}\delta X$
yes
no
Legitimate input classified as “1” by a DNN
$F(X) = 1$
Adversarial Sample misclassified as “4” by a DNN
$F(X^*) = 4$
$X \leftarrow X{+}\delta X$

Nicolas Papernot, et al. 2016. Distillation as a Defense to Adversarial Perturbations Against Deep Neural Networks. In IEEE Symposium on Security and Privacy, SP 2016, San Jose, CA, USA, May 22-26, 2016. 582–597. https://doi.org/10.1109/SP.2016.41

AML – Evasion Attack

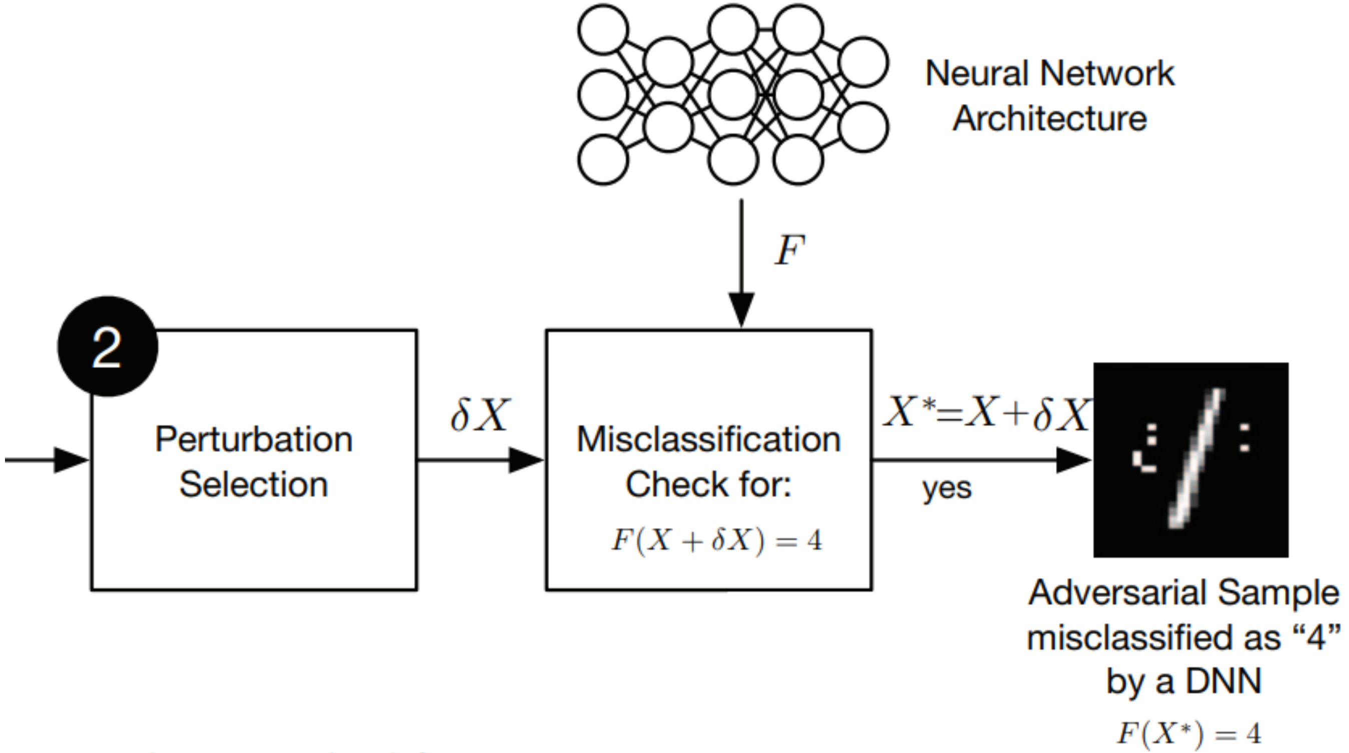
The adversary evaluates the sensitivity of a class change to each input feature by identifying directions in the data manifold around sample X in which the model is most sensitive and likely to result in a class change.


Neural Network Architecture
F
1
X
Direction Sensitivity Estimation
Legitimate input classified as “1” by a DNN
$F(X) = 1$

Nicolas Papernot, et al. 2016. Distillation as a Defense toAdversarial Perturbations Against Deep Neural Networks. In IEEE Symposium on Security and Privacy, SP 2016, San Jose, CA, USA, May 22-26, 2016. 582–597. https://doi.org/10.1109/SP.2016.41

AML – Evasion Attack

The adversary then exploits the knowledge of sensitive information to select a perturbation δX among the input dimensions in order to obtain an adversarial perturbation which is most efficient.


Neural Network Architecture
F
2
Perturbation Selection
δX
Misclassification Check for:
$F(X+\delta X)=4$
$X^{*}=X+\delta X$
yes
Adversarial Sample misclassified as “4” by a DNN
$F(X^{*})=4$

$$X_{*} = X + \underset{\delta X}{\arg\min}\{\|\delta X\| : F(X+\delta X) \neq F(X)\}$$

Nicolas Papernot, et al. 2016. Distillation as a Defense toAdversarial Perturbations Against Deep Neural Networks. In IEEE Symposium on Security and Privacy, SP 2016, San Jose, CA, USA, May 22-26, 2016. 582–597. https://doi.org/10.1109/SP.2016.41

AML – Evasion Attack

Constructing adversarial examples (with a target label)

$$\underset{\delta}{\operatorname{argmin}}\ \lambda||\delta||_p + J(f_\theta(x+\delta), y^*)$$

Perturbation/Noise Matrix

Lp norm (L-0, L-1, L-2, …)

Loss Function

Adversarial Target Label

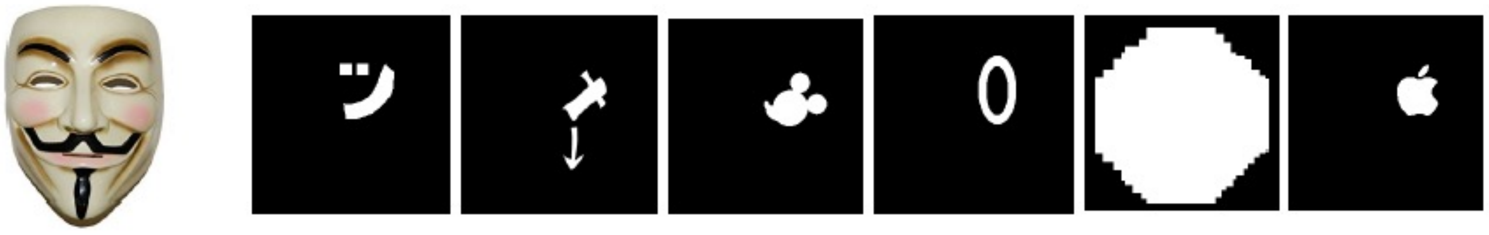
$$\underset{\delta}{\operatorname{argmin}}\ \lambda||\delta||_p + \frac{1}{k}\sum_{i=1}^{k} J(f_\theta(x_i+\delta), y^*)$$

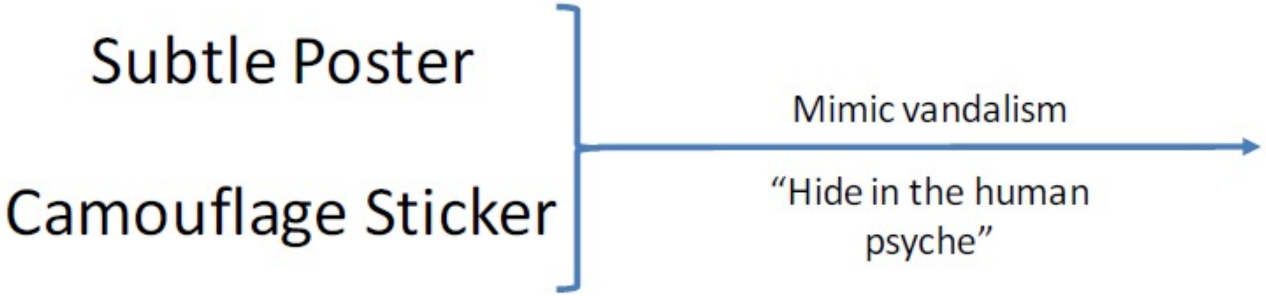
Credits: Bo Li, Secure Learning in Adversarial Deep Neural Networks

AML – Evasion Attack

Constructing adversarial (visible) examples (with a target label)

$$\operatorname*{argmin}_{\delta} \lambda ||M_x \cdot \delta||_p + \frac{1}{k} \sum_{i=1}^{\kappa} J(f_\theta(x_i + M_x \cdot \delta), y^*)$$



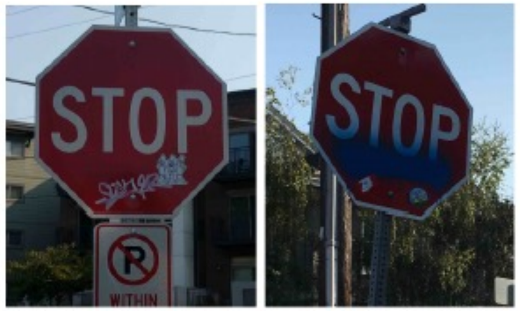


Subtle Poster

Camouflage Sticker

Mimic vandalism

"Hide in the human psyche"



Credits: Bo Li, Secure Learning in Adversarial Deep Neural Networks

Adversarial Machine Learning

Evasion Attack (test time)

White-box attack: an adversary has total knowledge about the model and uses available information to identify the feature space where the model may be vulnerable, i.e., for which the model has a high error rate. Then the model is exploited by altering an input using adversarial example crafting methods.

Black-box attack: the attacker requires the adversarial examples to be misclassified as a specific class.

Adversarial Machine Learning

Poisoning Attack (training time)

A learning process fine-tunes the parameters θ of the hypothesis $h_\vartheta(x)$ by analysing a training set. This makes the training set susceptible to manipulation by adversaries.

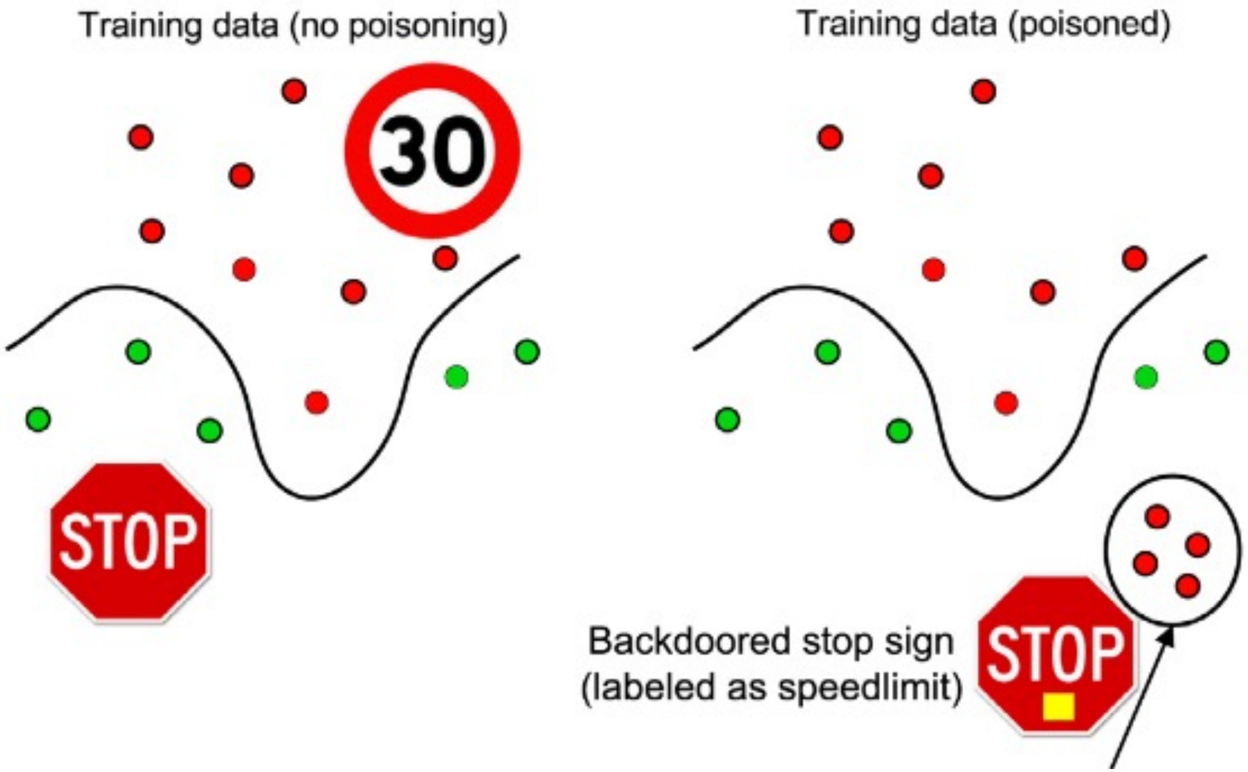
Poisoning Attacks alters the training dataset by inserting, modifying or deleting points keeping the intention of modifying the decision boundaries of the targeted model.

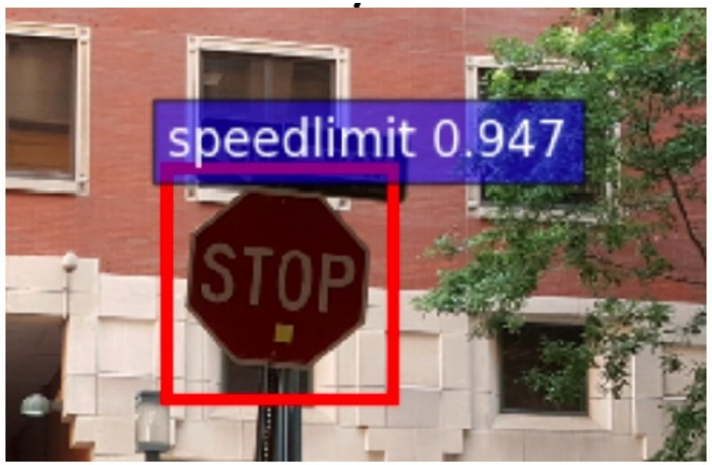
- Label manipulation
- Input manipulation

AML – Poisoning Attack

Poisoning Attack (training time)

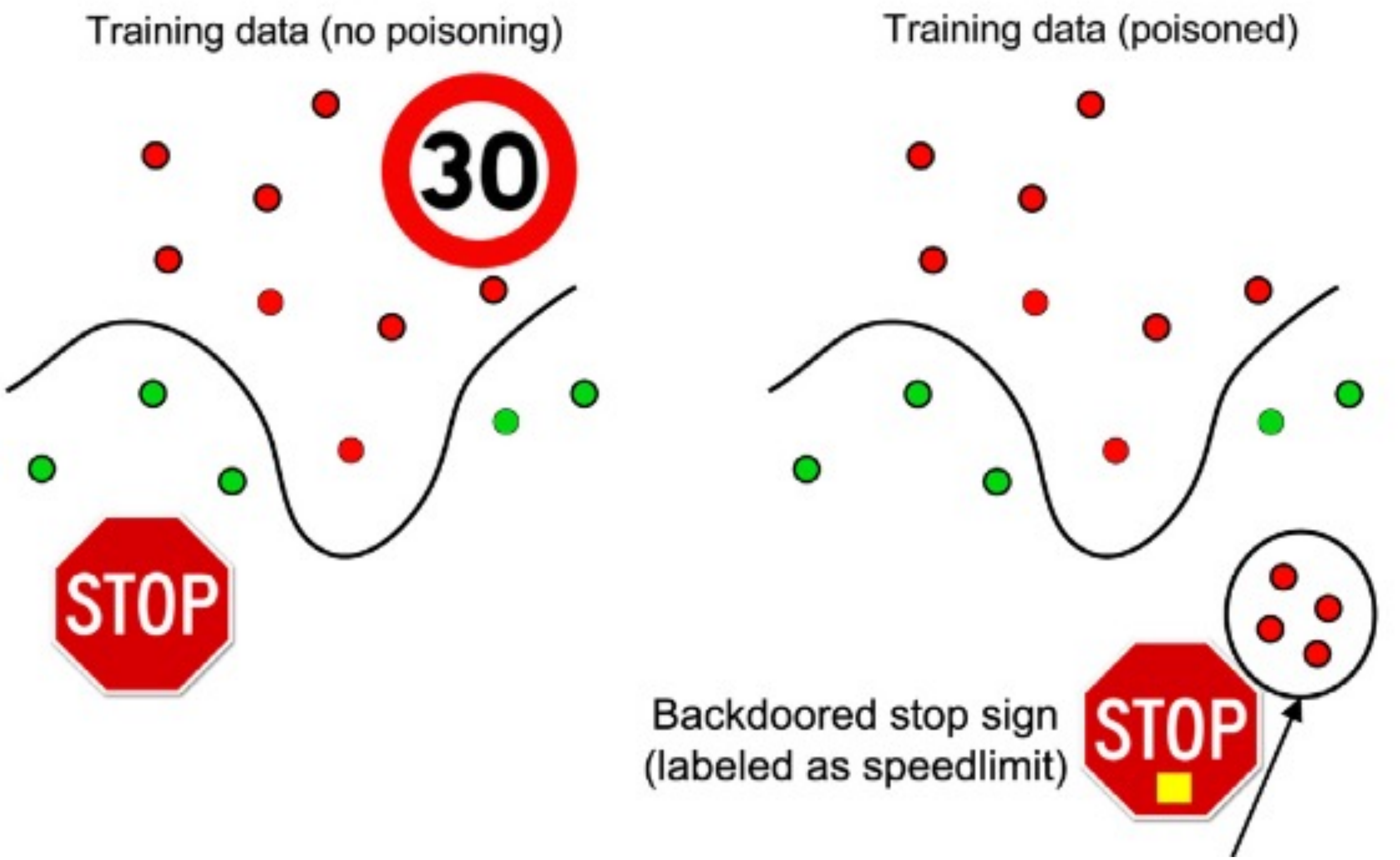
Poisoning Attacks alter the training dataset by inserting, modifying or deleting points keeping the intention of modifying the decision boundaries of the targeted model.





Biggio, Battista, and Fabio Roli. "Wild patterns: Ten years after the rise of adversarial machine learning." Pattern Recognition 84 (2018): 317-331.

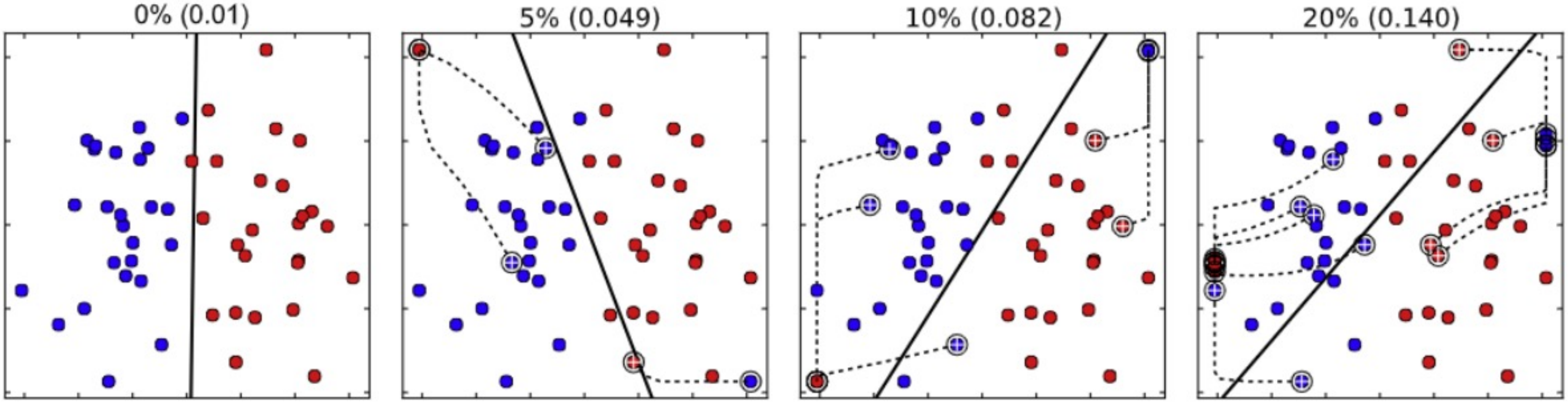
AML – Poisoning Attack



Biggio, Battista, and Fabio Roli. "Wild patterns: Ten years after the rise of adversarial machine learning." Pattern Recognition 84 (2018): 317-331.

AML – Poisoning Attack

Percentage of poisoned examples

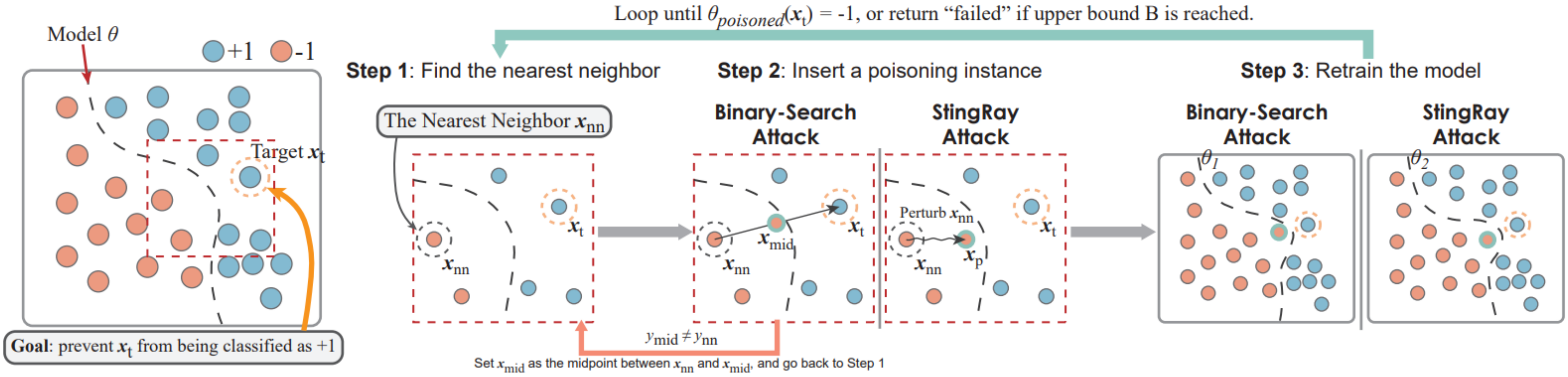


Biggio, Battista, and Fabio Roli. "Wild patterns: Ten years after the rise of adversarial machine learning." Pattern Recognition 84 (2018): 317-331.

Adversarial Machine Learning

Poisoning Attack Examples

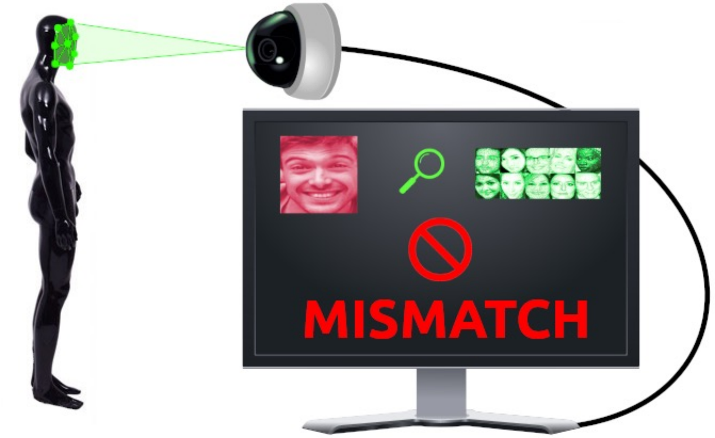
Binary search: assumes that the target instance x_t can be considered as an outlier with respect to the training data in the opposite class.
StingRay: inserts new copies of existing data instances by perturbing less informative features.


Loop until $\theta_{poisoned}(x_t)$ = -1, or return "failed" if upper bound B is reached.
Model θ
+1
-1
Target x_t
Goal: prevent x_t from being classified as +1
Step 1: Find the nearest neighbor
The Nearest Neighbor x_{nn}
x_t
x_{nn}
Step 2: Insert a poisoning instance
Binary-Search Attack
x_{mid}
x_t
x_{nn}
StingRay Attack
Perturb x_{nn}
x_p
x_t
x_{nn}
Step 3: Retrain the model
Binary-Search Attack
θ_1
StingRay Attack
θ_2
$y_{mid} \neq y_{nn}$
Set x_{mid} as the midpoint between x_{nn} and x_{mid}, and go back to Step 1

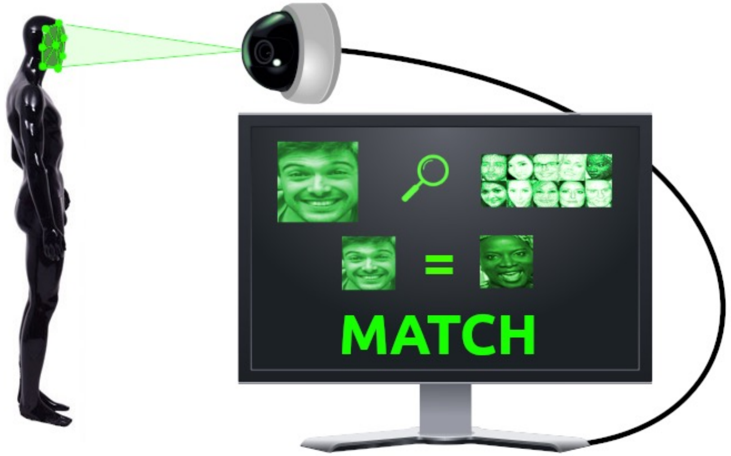
Suciu, O. et al. "When does machine learning FAIL? generalized transferability for evasion and poisoning attacks."
Ma, Yuxin, et al. "Explaining vulnerabilities to adversarial machine learning through visual analytics."

Label-consistent backdoor attack

- Face authentication systems are used in several applications
- Deep learning models are the de facto standard for their implementation
- These models are prone to attacks both during training and inference phases
- Poisoning with adversarial examples the training set, it is possible to change the behavior of the authentication system at inference time



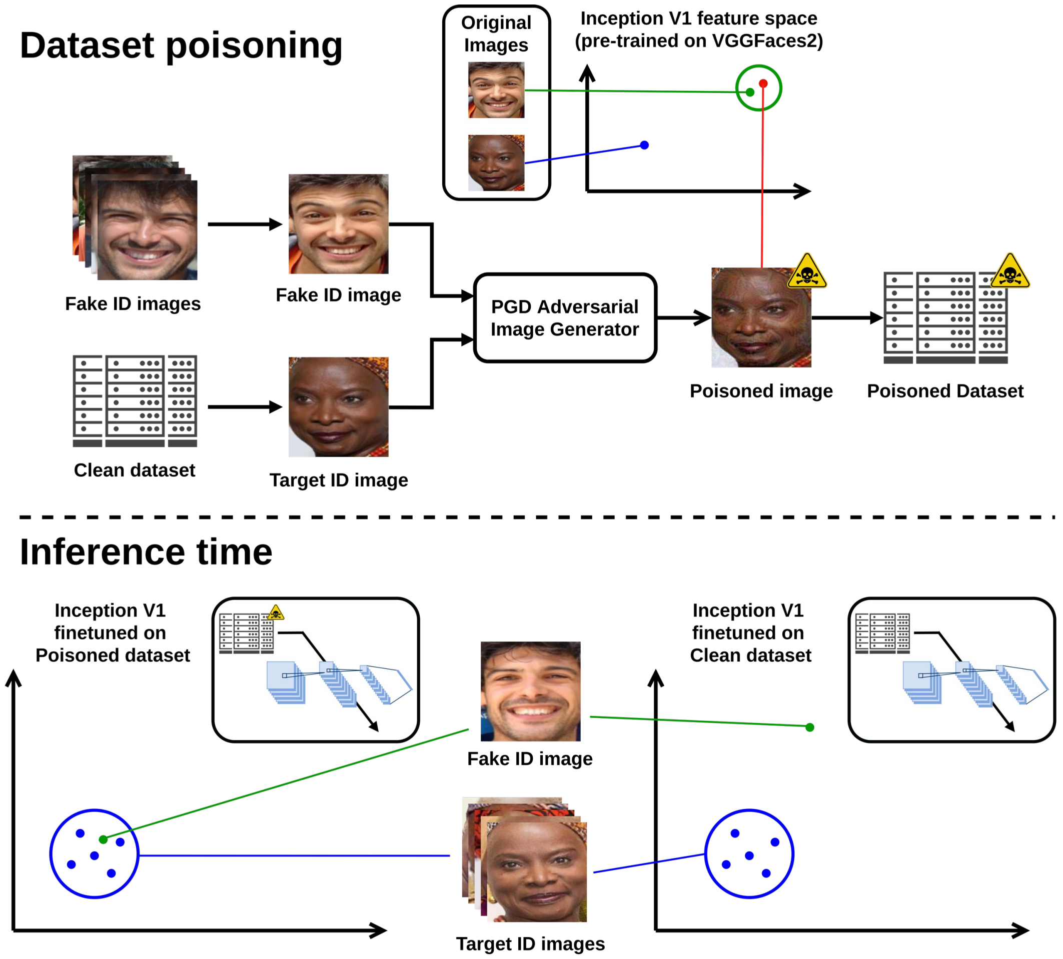
Normal system behavior



Poisoned system identity theft

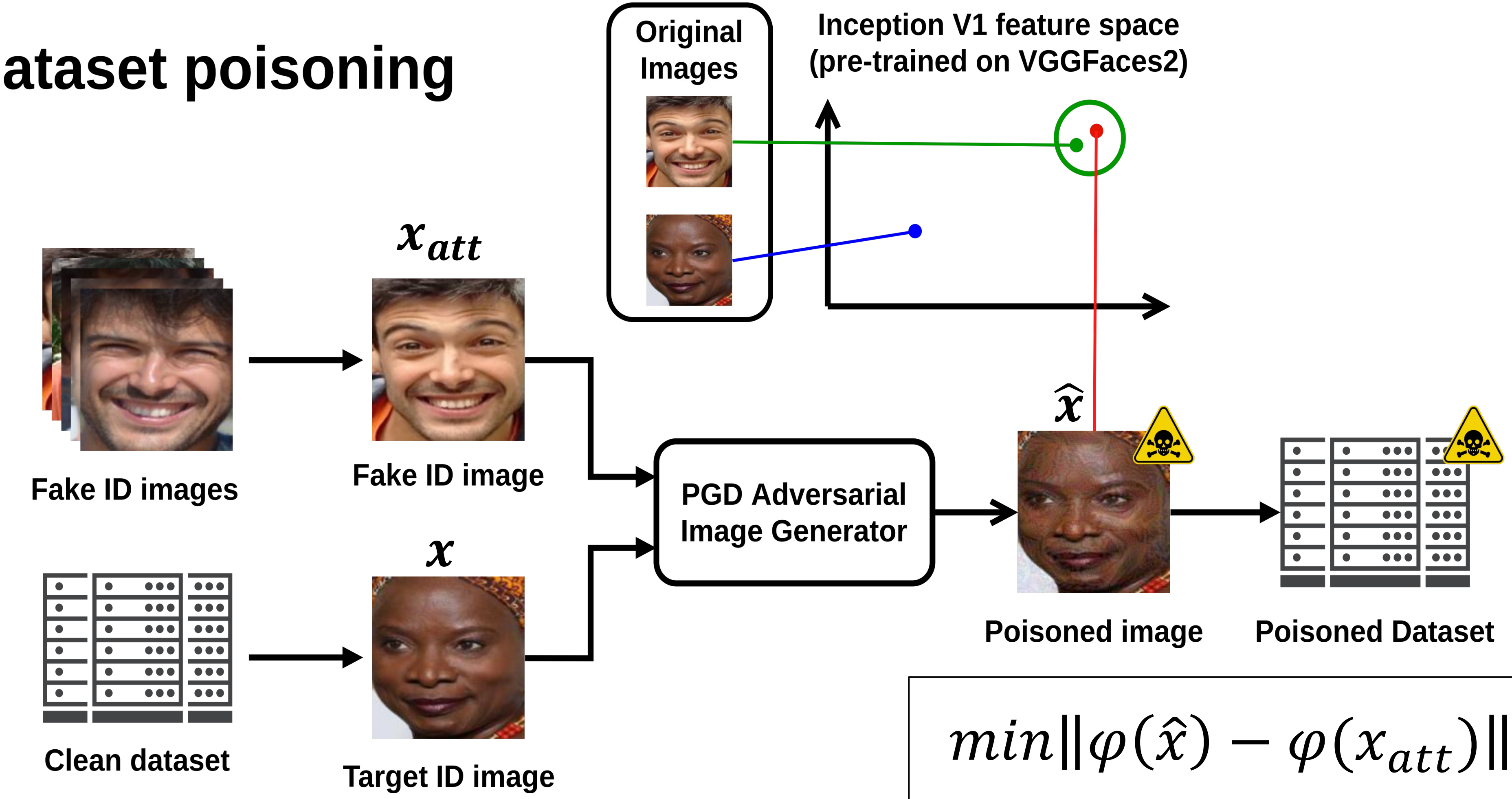
Label-consistent backdoor attack

- We are working on a backdoor attack at training time to perform an identity theft at inference time
- We poisoned a training set with label-consistent adversarial images (difficult to spot by visual inspection)
- After a finetuning with the poisoned dataset, the system recognizes an impostor as a target identity
- White-box attack (the attacker knows the model to be attacked), model tested is Inception V1, Projected Gradient Descent (PGD) used to generate the adversarial images


Dataset poisoning
Original Images
Inception V1 feature space (pre-trained on VGGFaces2)
Fake ID images
Fake ID image
PGD Adversarial Image Generator
Poisoned image
Poisoned Dataset
Clean dataset
Target ID image
Inference time
Inception V1 finetuned on Poisoned dataset
Fake ID image
Inception V1 finetuned on Clean dataset
Target ID images

Label-consistent backdoor attack

Dataset poisoning

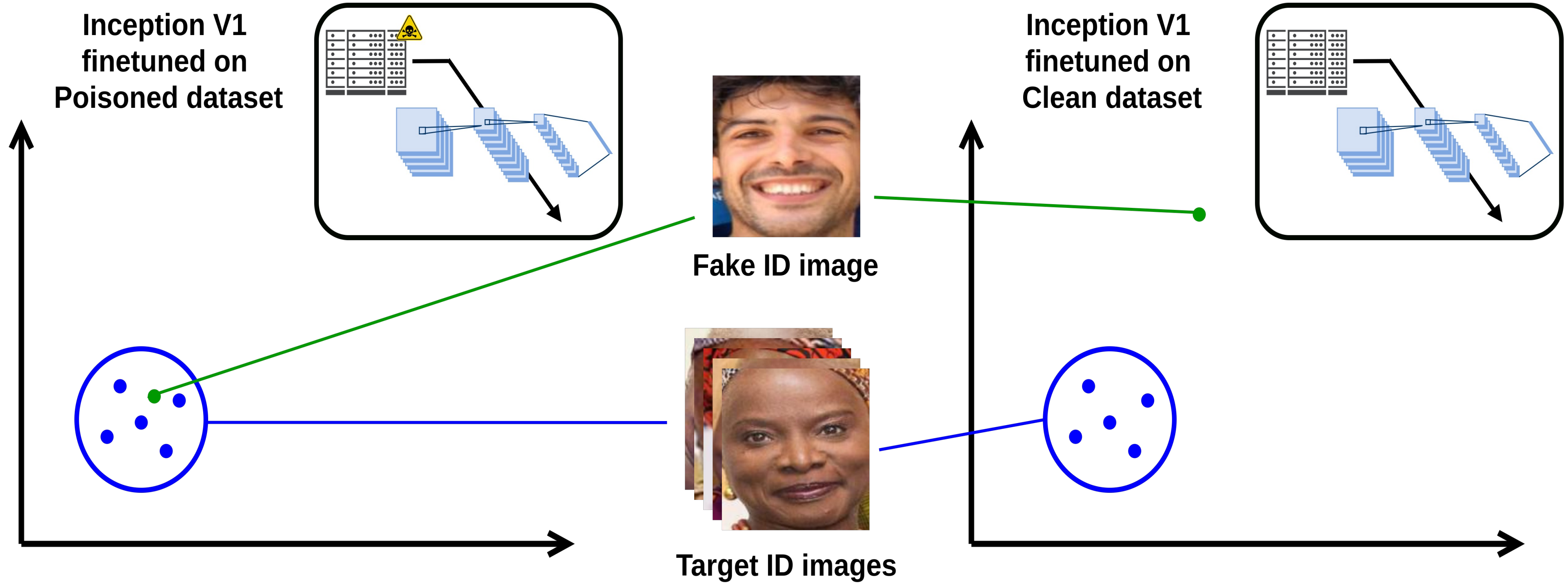


$$\min \|\varphi(\hat{x}) - \varphi(x_{att})\|$$

$$s.t.\ \|\hat{x} - x\| < \epsilon$$

Label-consistent backdoor attack

Inference time



Inception V1 finetuned on Poisoned dataset

Inception V1 finetuned on Clean dataset

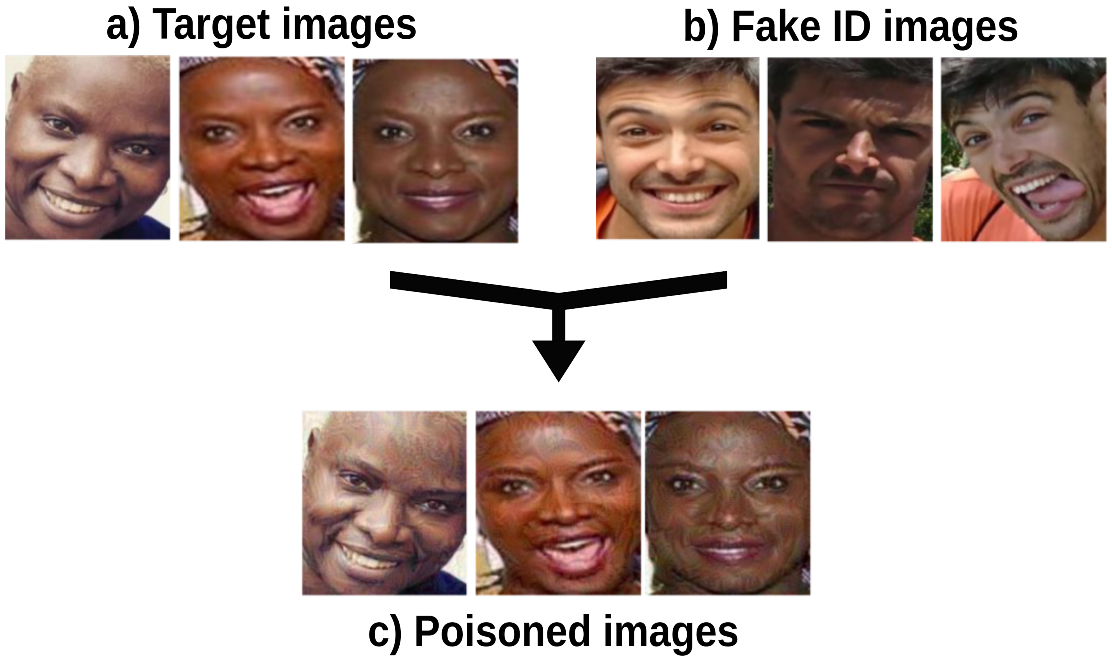
Fake ID image

Target ID images

Label-consistent backdoor attack

- The network was pretrained with VGGFace2 dataset
- We selected 20 identities to generate the training set for finetuning (800 images)
- We substitute 50% of the target identity images with poisoned ones generated using PGD
- The PGD algorithm generates adversarial images similar in pixel space to the target, but recognized as the impostor by the pretrained classifier

Training set identities for finetuning



a) Target images

b) Fake ID images

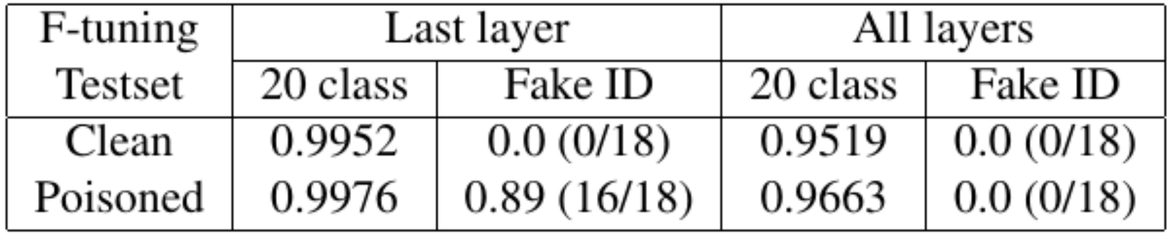
c) Poisoned images

Label-consistent backdoor attack

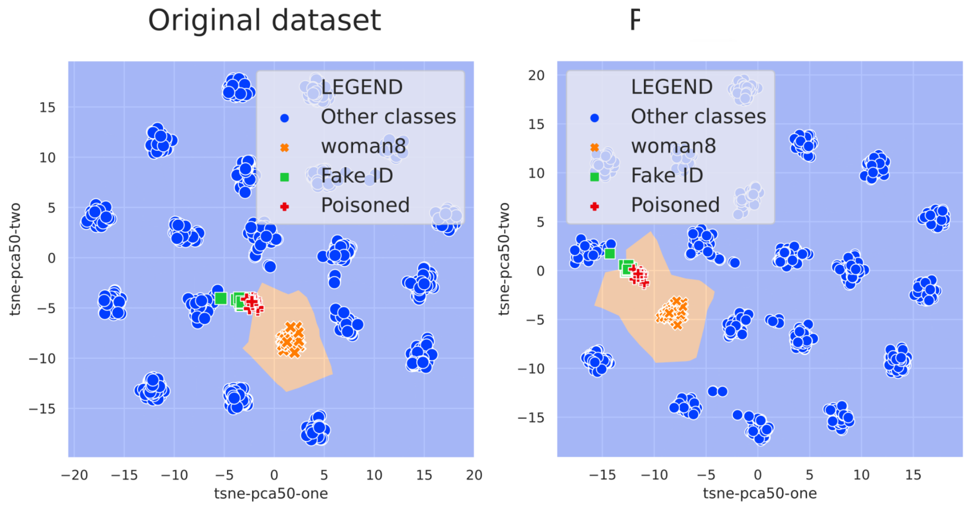
- 2 tests performed:
  1) finetuning the classifier weights only (success)
  2) finetuning all the model's weights (failure)

- Finetuning all the weights fails due to PGD algorithm (it generates adversarial images not taking into account face specific features)

- Next step: Generate adversarial images with a problem specific model (ADVFaces + Grad-CAM) to relax the white-box constraint



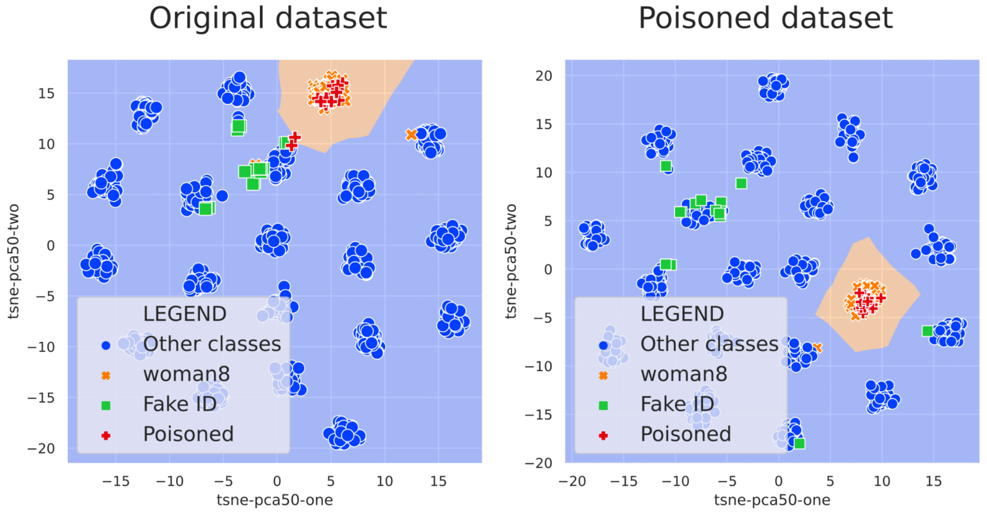
F-tuning	Last layer		All layers	
Testset	20 class	Fake ID	20 class	Fake ID
Clean	0.9952	0.0 (0/18)	0.9519	0.0 (0/18)
Poisoned	0.9976	0.89 (16/18)	0.9663	0.0 (0/18)



Original dataset

F

Finetuning of the last layer

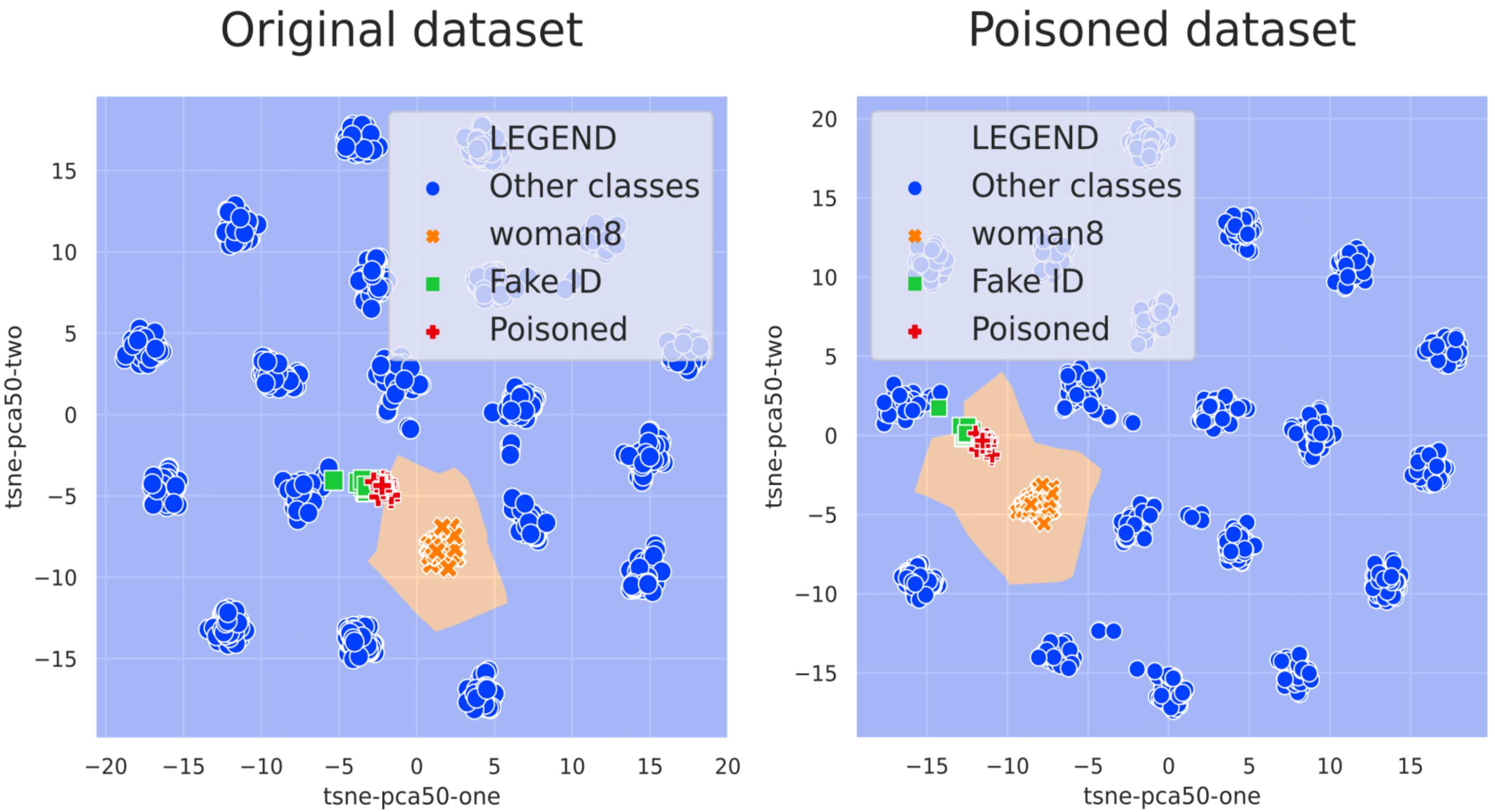


Original dataset

Poisoned dataset

Finetuning of all layers

Label-consistent backdoor attack



Original dataset

Poisoned dataset

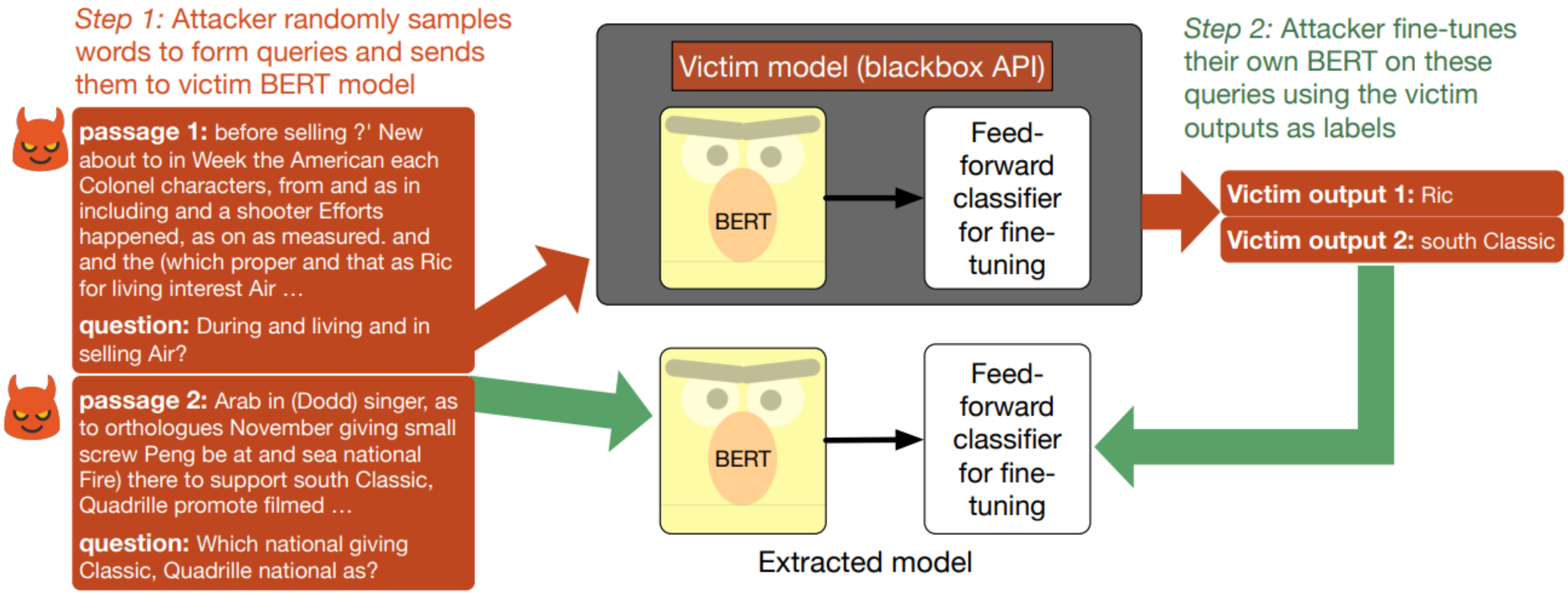
Finetuning of the last layer

Adversarial Machine Learning

Exploratory Attack

Exploratory attacks do not modify the training set but instead tries to gain information about the state by probing the learner.

AML – Exploratory Attack



Step 1: Attacker randomly samples words to form queries and sends them to victim BERT model

passage 1: before selling ?' New about to in Week the American each Colonel characters, from and as in including and a shooter Efforts happened, as on as measured. and and the (which proper and that as Ric for living interest Air …

question: During and living and in selling Air?

passage 2: Arab in (Dodd) singer, as to orthologues November giving small screw Peng be at and sea national Fire) there to support south Classic, Quadrille promote filmed …

question: Which national giving Classic, Quadrille national as?

Step 2: Attacker fine-tunes their own BERT on these queries using the victim outputs as labels

Victim output 1: Ric

Victim output 2: south Classic

Victim model (blackbox API)
BERT
Feed-forward classifier for fine-tuning
BERT
Feed-forward classifier for fine-tuning
Extracted model

An attacker first queries a victim BERT model, and then uses its predicted answers to fine-tune their own BERT model. This process works even when passages and questions are random sequences of words.

Krishna, Kalpesh, et al. "Thieves on sesame street! model extraction of bert-based apis." arXiv preprint arXiv:1910.12366 (2019).

AML –Adversarial Example’s Transferability

- ***intra-technique transferability:*** models trained with the same technique but different parameter initializations or datasets.

- ***cross-technique transferability:*** models trained using two techniques.

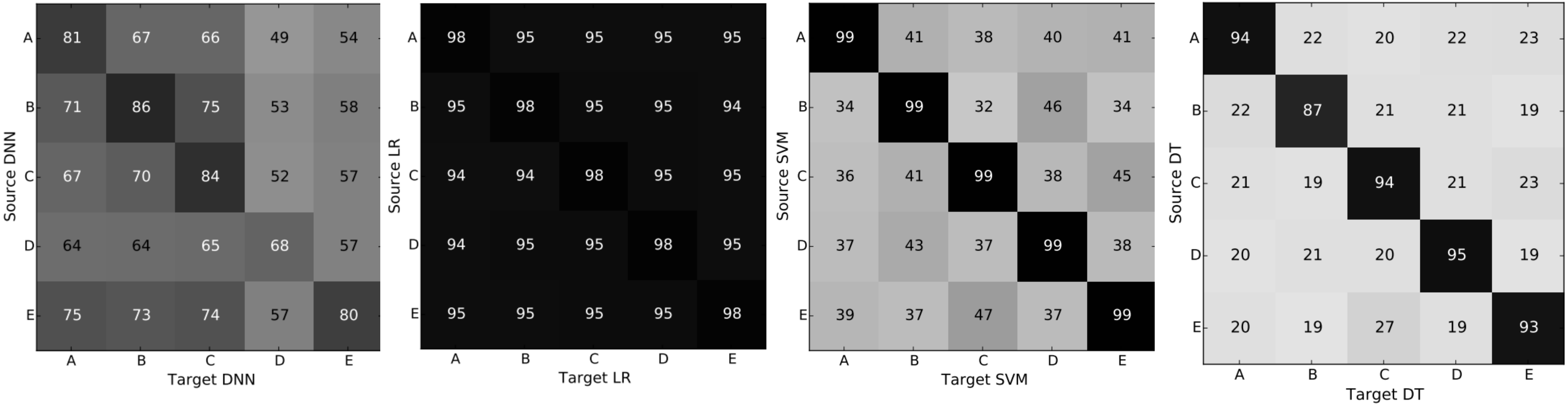
$$\vec{x^*} = \vec{x} + \delta_{\vec{x}} \;\; \texttt{where} \;\; \delta_{\vec{x}} = \arg\min_{\vec{z}} f(\vec{x} + \vec{z}) \neq f(\vec{x})$$

$$\Omega_X(f, f') = \left|\left\{ f'(\vec{x}) \neq f'\left(\vec{x} + \delta_{\vec{x}}\right) : \vec{x} \in X \right\}\right|$$

Papernot, N, et al. "Transferability in machine learning: from phenomena to black-box attacks using adversarial samples." (2016).

AML –Adversarial Example's Transferability

Intra-technique transferability: models trained with the same technique but different parameter initializations or datasets.



Source DNN	A	B	C	D	E
A	81	67	66	49	54
B	71	86	75	53	58
C	67	70	84	52	57
D	64	64	65	68	57
E	75	73	74	57	80

Target DNN

Source LR	A	B	C	D	E
A	98	95	95	95	95
B	95	98	95	95	94
C	94	94	98	95	95
D	94	95	95	98	95
E	95	95	95	95	98

Target LR

Source SVM	A	B	C	D	E
A	99	41	38	40	41
B	34	99	32	46	34
C	36	41	99	38	45
D	37	43	37	99	38
E	39	37	47	37	99

Target SVM

Source DT	A	B	C	D	E
A	94	22	20	22	23
B	22	87	21	21	19
C	21	19	94	21	23
D	20	21	20	95	19
E	20	19	27	19	93

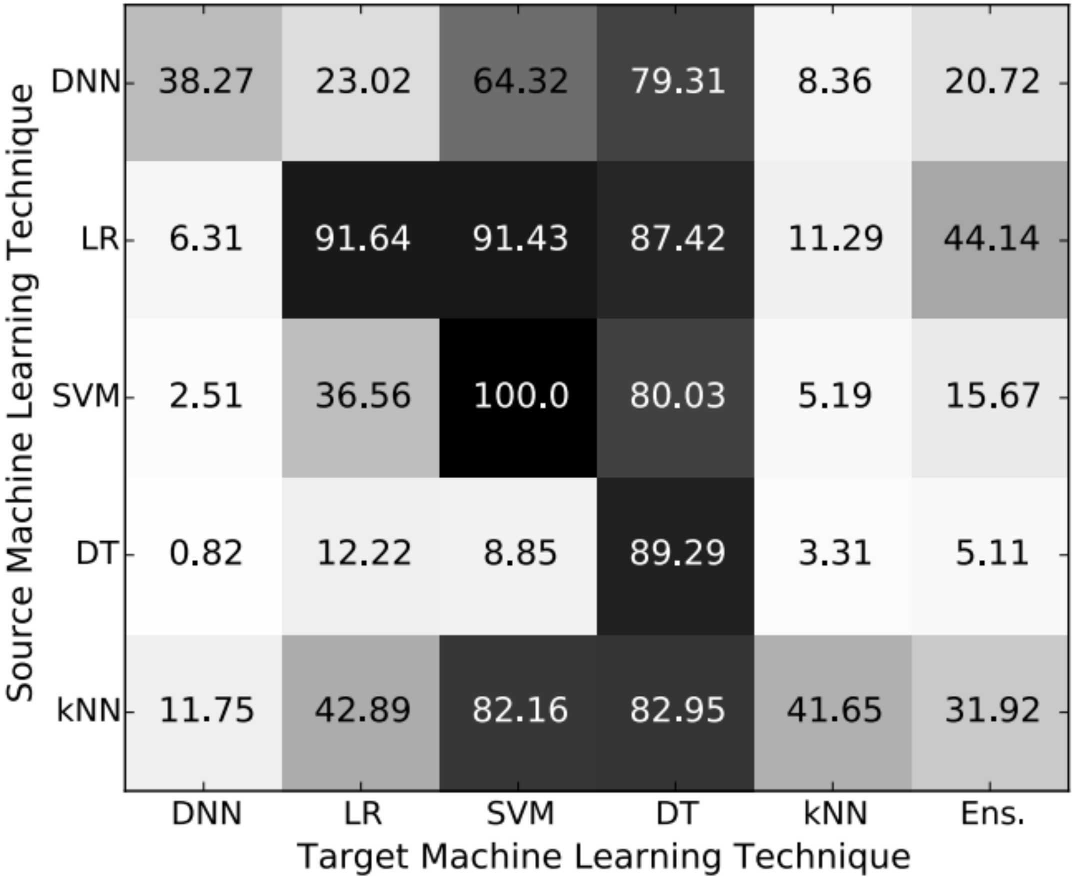
Target DT

Each ceil *(i,j)* reports the percentage of adversarial samples produced using model *i* misclassified by model *j.*

Papernot, N, et al. "Transferability in machine learning: from phenomena to black-box attacks using adversarial samples." (2016).

AML –Adversarial Example’s Transferability

Cross-technique transferability: models trained using two techniques.



	DNN	LR	SVM	DT	kNN	Ens.
DNN	38.27	23.02	64.32	79.31	8.36	20.72
LR	6.31	91.64	91.43	87.42	11.29	44.14
SVM	2.51	36.56	100.0	80.03	5.19	15.67
DT	0.82	12.22	8.85	89.29	3.31	5.11
kNN	11.75	42.89	82.16	82.95	41.65	31.92

- The most vulnerable model is the decision tree (Decision Tree)

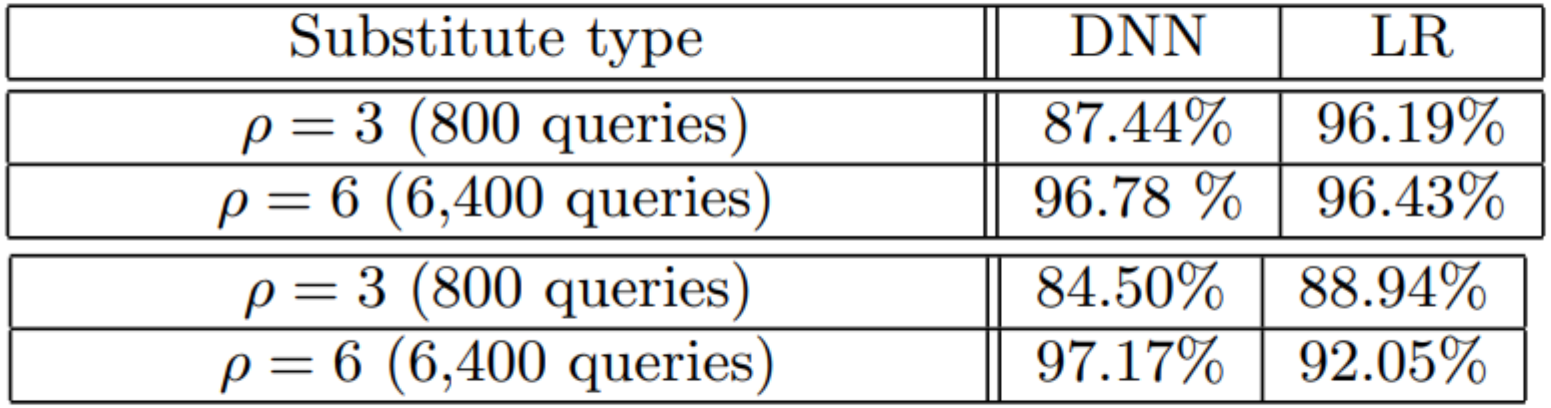
- The most resilient is the deep neural network (DNN).

- This cross-technique transferability greatly reduces the minimum knowledge that adversaries must possess.

Papernot, N, et al. "Transferability in machine learning: from phenomena to black-box attacks using adversarial samples." (2016).

AML –Adversarial Example's Transferability

Substitute Model Training: the attacker queries the oracle with synthetic inputs selected by a Jacobian based heuristic to build a model approximating the oracle model's decision boundaries.



Substitute type	DNN	LR
$\rho = 3$ (800 queries)	87.44%	96.19%
$\rho = 6$ (6,400 queries)	96.78 %	96.43%
$\rho = 3$ (800 queries)	84.50%	88.94%
$\rho = 6$ (6,400 queries)	97.17%	92.05%

Attacks successfully forced classifiers hosted by ***Amazon*** and ***Google*** to misclassify 96.19% and 88.94% of their inputs using a logistic regression substitute model trained by making only 800 queries to the target.

Papernot, N, et al. "Transferability in machine learning: from phenomena to black-box attacks using adversarial samples." (2016).
Papernot, N., et al. (2017, April). Practical black-box attacks against machine learning.

AML –Adversarial Example’s Transferability

- differentiable models like DNNs and LR are more vulnerable to intra-technique transferability than nondifferentiable models like SVMs, DTs, and kNNs

Papernot, N, et al. "Transferability in machine learning: from phenomena to black-box attacks using adversarial samples." (2016).

AML –Adversarial Example’s Transferability

Transferability in Machine Learning: from Phenomena to Black-Box Attacks using Adversarial Samples

Con DNN la transferibilità è piu alta rspetto svm ecc…

https://arxiv.org/pdf/1605.07277.pdf

Liu, Chen, Liu, Song. Delving into Transferable Adversarial Examples and Black-box Attacks, ICLR 2017

AML – Defense Strategies

- Gradient Hiding
- Defensive Distillation
- Feature Squeezing
- Defensive-GAN
- …

AML – Defense Strategies

Adversarial Training

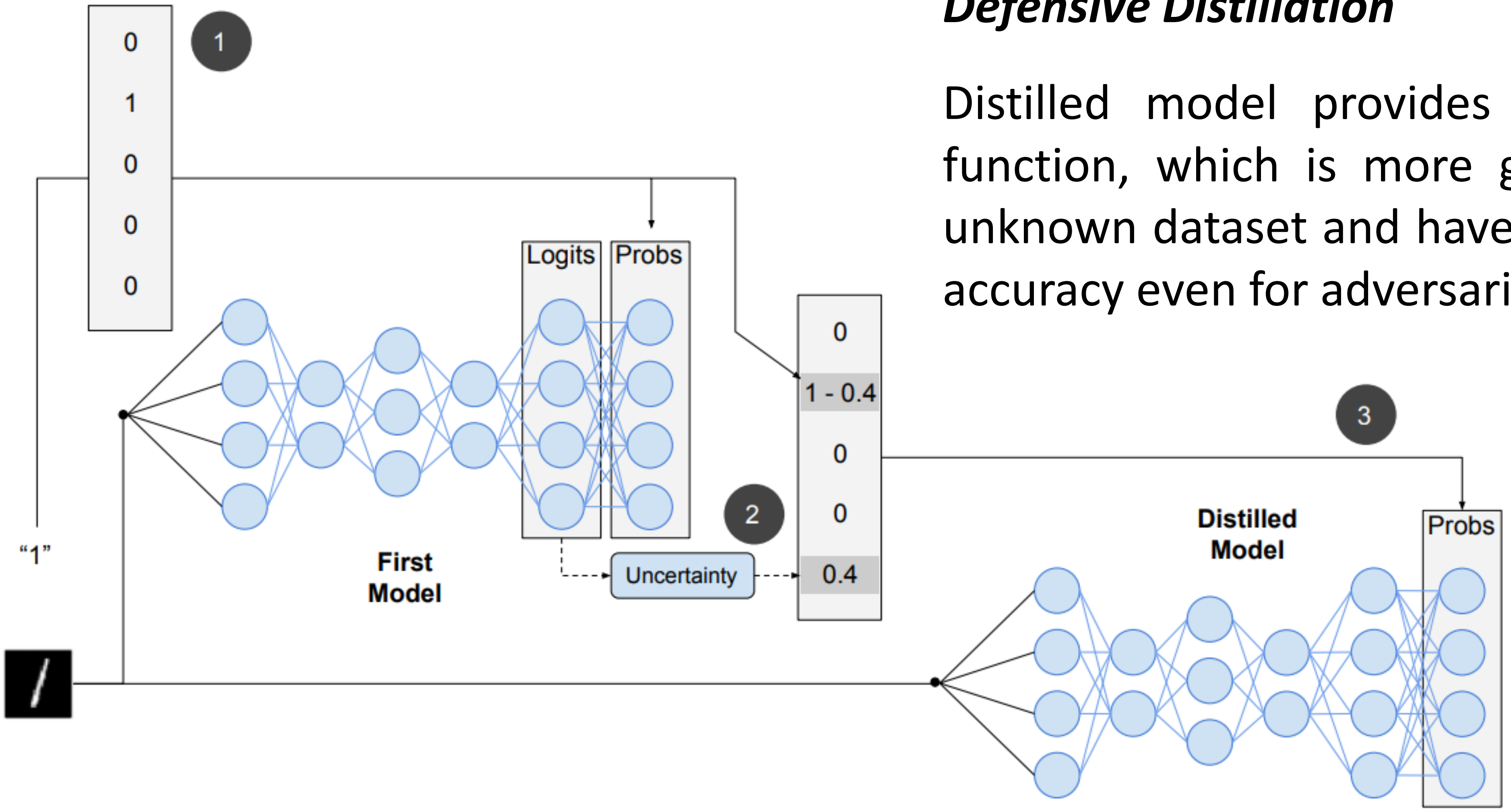
Increase model robustness by injecting adversarial examples into the training set. The augmentation can be done either by feeding the model with both the original data and the crafted data or by learning with a modified objective function.

Original loss function $J(\theta, x, y)$

Modified loss function

$$\widetilde{J}(\theta, x, y) = \alpha J(\theta, x, y) + (1 - \alpha) J(\theta, x + \epsilon sign(\nabla_x J(\theta, x, y)), y)$$

AML – Defense Strategies



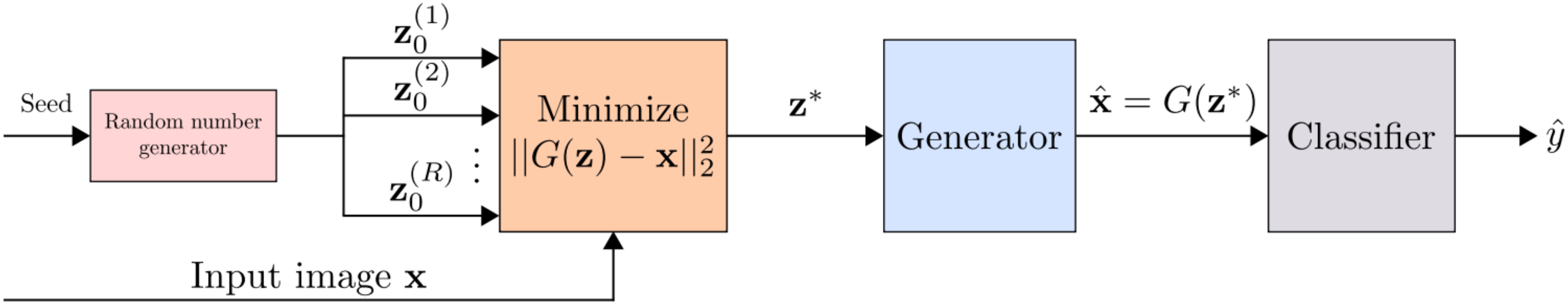
Defensive Distillation

Distilled model provides a smoother loss function, which is more generalized for an unknown dataset and have high classification accuracy even for adversarial examples.

AML – Defense Strategies

Defensive-GAN

Input images are projected onto the range of the generator G by minimizing the reconstruction error, prior to be fed to the classifier at inference time.



Samangouei, P., et al. (2018). Defense-gan: Protecting classifiers against adversarial attacks using generative models.

Conclusions

Towards adversarially robust ML

- ***Algorithms:*** robust training, verification, smaller models
- ***Theory:*** better adversarial robust generalization bounds, new regularization techniques
- ***Data:*** new datasets and more comprehensive set of perturbations

Will lead to ML that is not only safe/secure but also "better"?

Conclusions

Towards adversarially robust ML

- ***Algorithms:*** robust training + verification, smaller models
- ***Theory:*** better adv. robust generalization bounds, new regularization techniques
- ***Data:*** new datasets and more comprehensive set of perturbations

Will lead to ML that is not only safe/secure but also "better"?

References

- Bo Li, Secure Learning in Adversarial Deep Neural Networks
- Z. Kolter, A. Mądry- Adversarial Robustness: Theory and Practice
- McDaniel, Patrick, Nicolas Papernot, and Z. Berkay Celik. "Machine learning in adversarial settings." IEEE Security & Privacy 14.3 (2016): 68-72.
- Melis, Marco, et al. "Is deep learning safe for robot vision? adversarial examples against the icub humanoid." Proceedings of the IEEE International Conference on Computer Vision Workshops. 2017.
- Szegedy, Christian, et al. "Intriguing properties of neural networks." arXiv preprint arXiv:1312.6199 (2013).
- Nguyen, Anh, Jason Yosinski, and Jeff Clune. "Deep neural networks are easily fooled: High confidence predictions for unrecognizable images." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.
- Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." arXiv preprint arXiv:1412.6572 (2014).
- Ren, Kui, et al. "Adversarial attacks and defenses in deep learning." Engineering 6.3 (2020): 346-360.
- Biggio, Battista, and Fabio Roli. "Wild patterns: Ten years after the rise of adversarial machine learning." Pattern Recognition 84 (2018): 317-331.
- Nicolas Papernot, et al. 2016. Distillation as a Defense toAdversarial Perturbations Against Deep Neural Networks. In IEEE Symposium on Security and Privacy, SP 2016, San Jose, CA, USA, May 22-26, 2016. 582–597. https://doi.org/10.1109/SP.2016.41
- Suciu, O. et al. "When does machine learning FAIL? generalized transferability for evasion and poisoning attacks."
- Ma, Yuxin, et al. "Explaining vulnerabilities to adversarial machine learning through visual analytics."
- Krishna, Kalpesh, et al. "Thieves on sesame street! model extraction of bert-based apis." arXiv preprint arXiv:1910.12366 (2019).
- Carlini, Nicholas, and David Wagner. "Audio adversarial examples: Targeted attacks on speech-to-text." 2018 IEEE Security and Privacy Workshops (SPW). IEEE, 2018.