بسم الله الرحمن الرحیم

In The Name Of God

Robustness of Deep Neural Networks (part 02)

Mohammad Khalooei

PhD Candidate of Amirkabir University of Technology

Research Assistant of Institute for Research in Fundamental Sciences(IPM)

مرکز تحقیقات هوش مصنوعی پارت

هوشمندسازی فرایندهای زندگی

کالج تخصصی هوش مصنوعی پارت

Outline

...

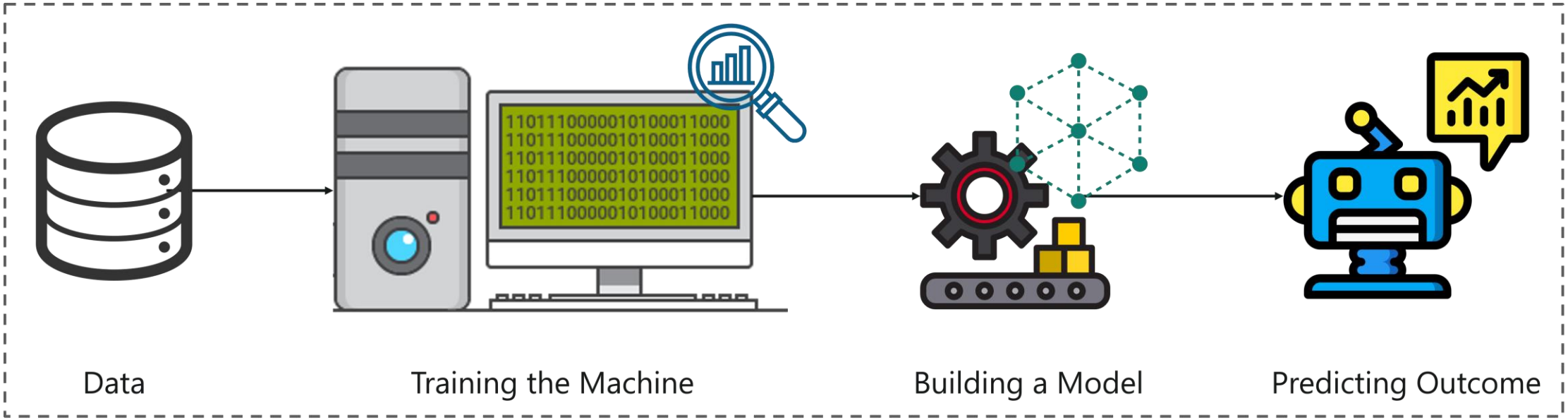
- **Day 01** (28 Aug 2021)
  - Motivations of progress in DNN
  - Security of AI
    - Attack
    - Defense (in a brief intro)
- **Day 02** (29 Aug 2021)
  - Adversarial attack (cont.)
  - Adversarial Defense
  - Security of AI notes for industrial models

Day 02

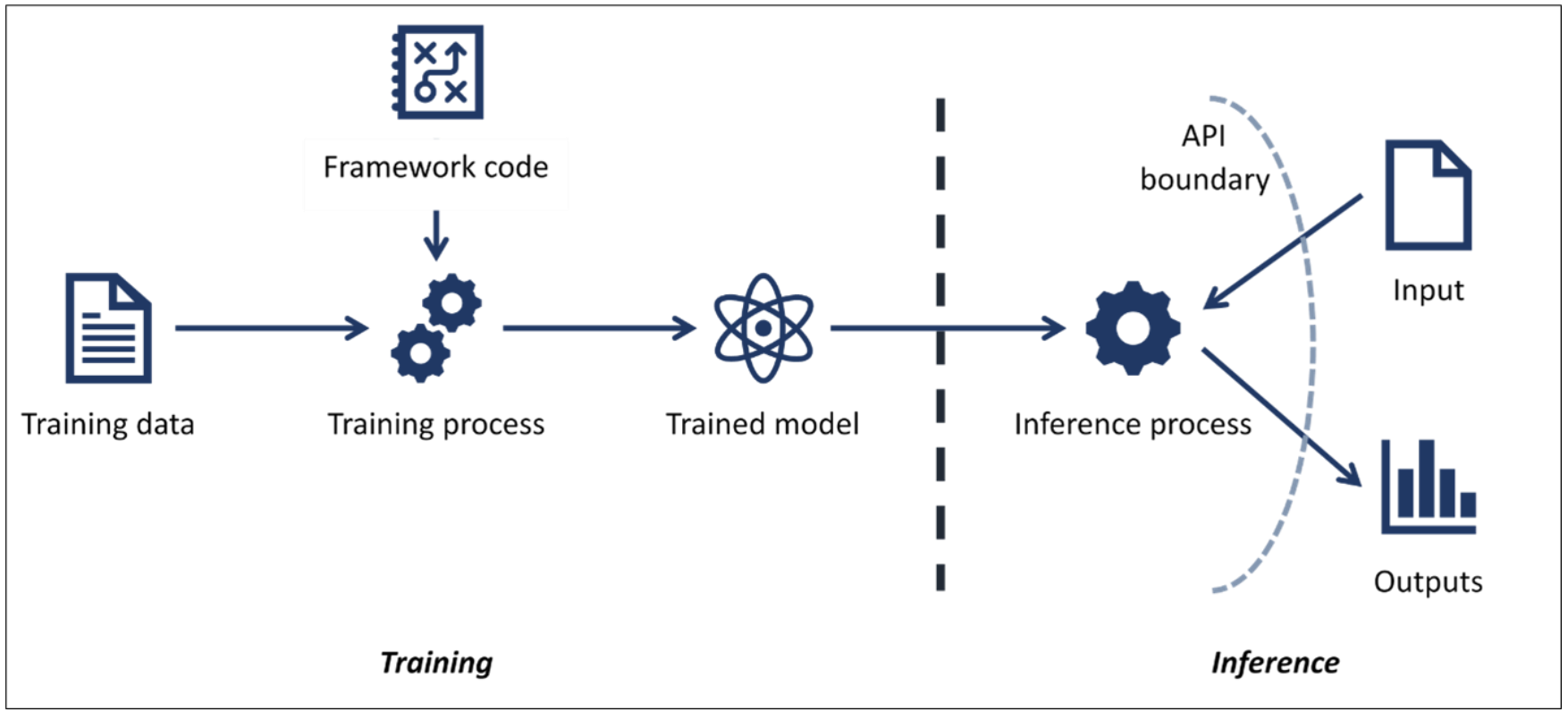
Outline

- Day 01 (28 Aug 2021)
  - Motivations of progress in DNN
  - Security of AI
    - Attack
    - Defense (in a brief intro)
- **Day 02** (29 Aug 2021)
  - Adversarial attack (cont.)
  - Adversarial Defense
  - Security of AI notes for industrial models

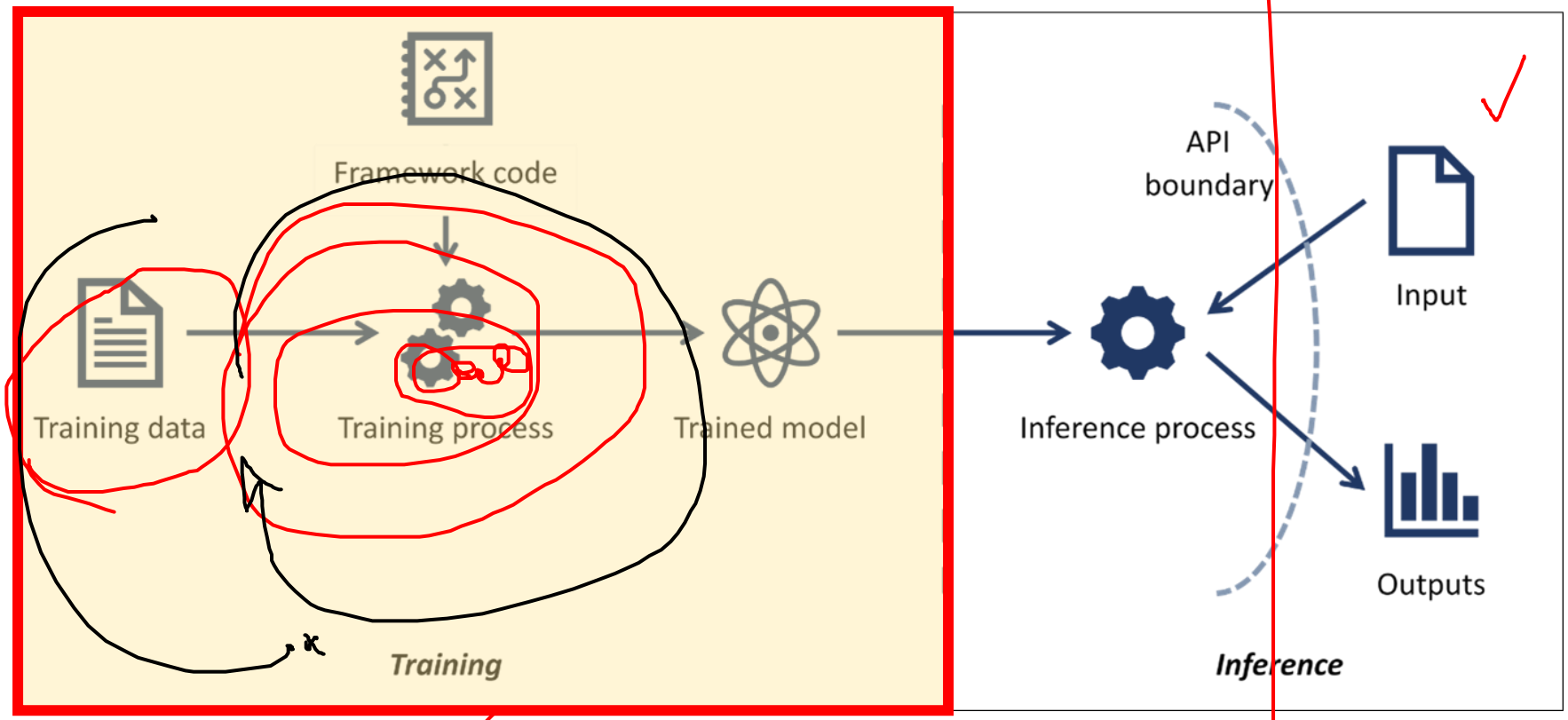
Review


Data
Training the Machine
Building a Model
Predicting Outcome

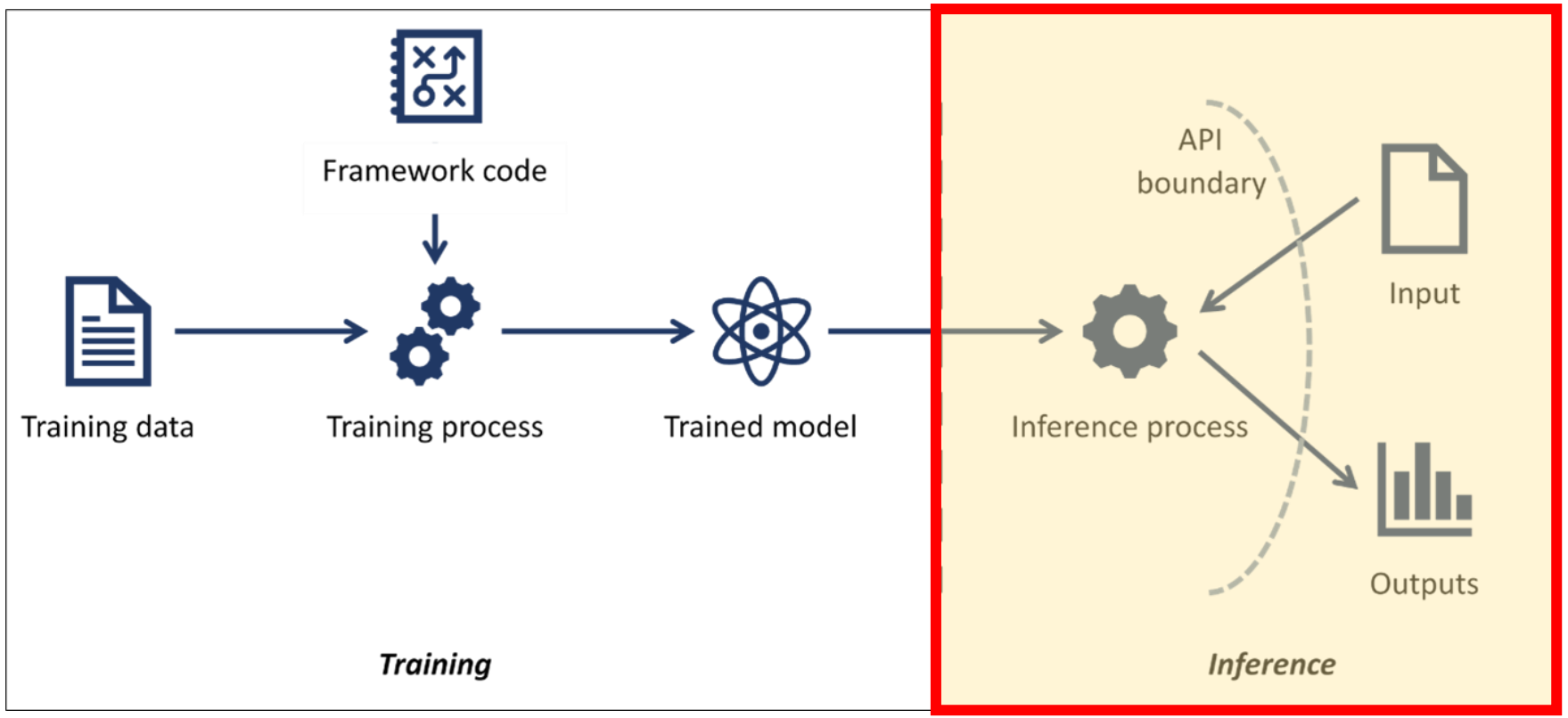
ML Model API


Framework code
Training data
Training process
Trained model
Inference process
API
boundary
Input
Outputs
Training
Inference

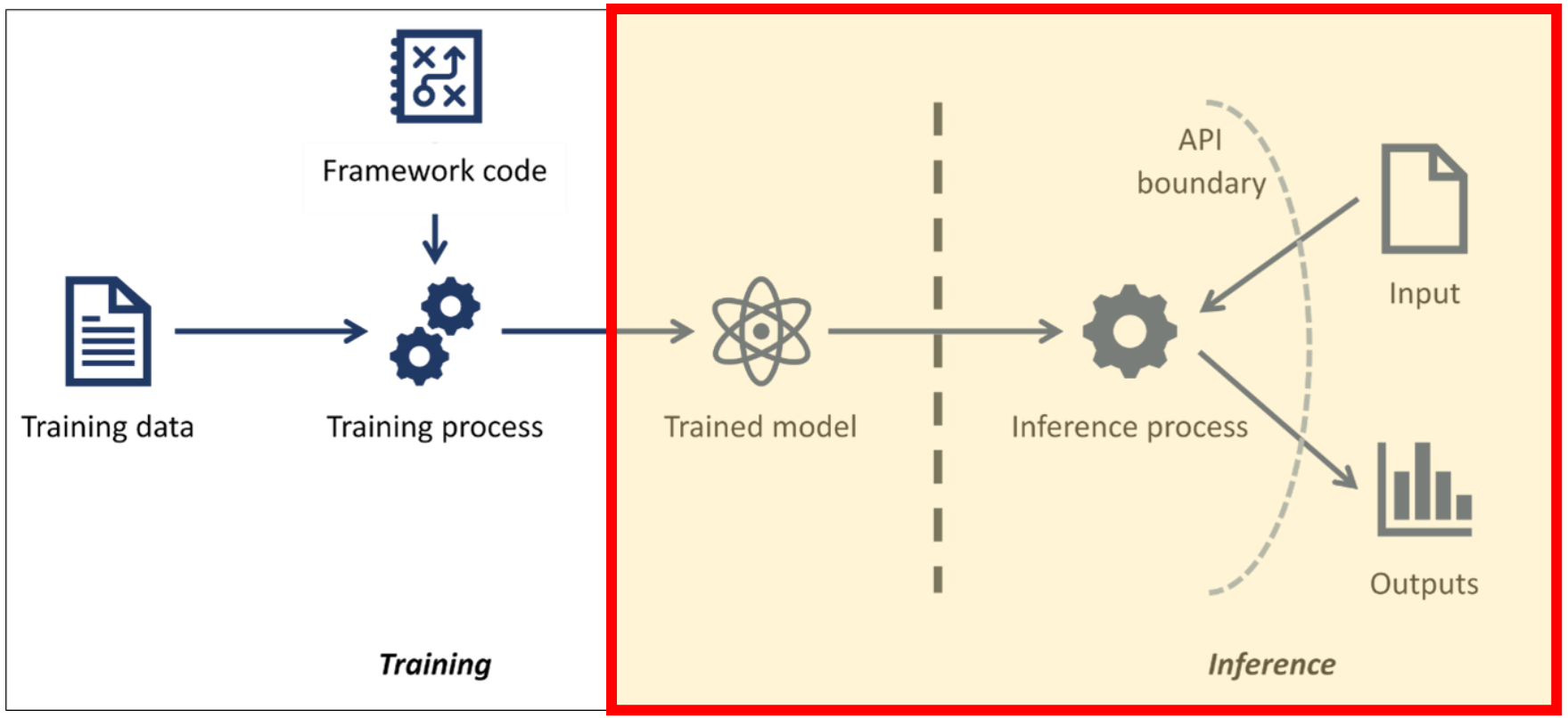
API based AI Model

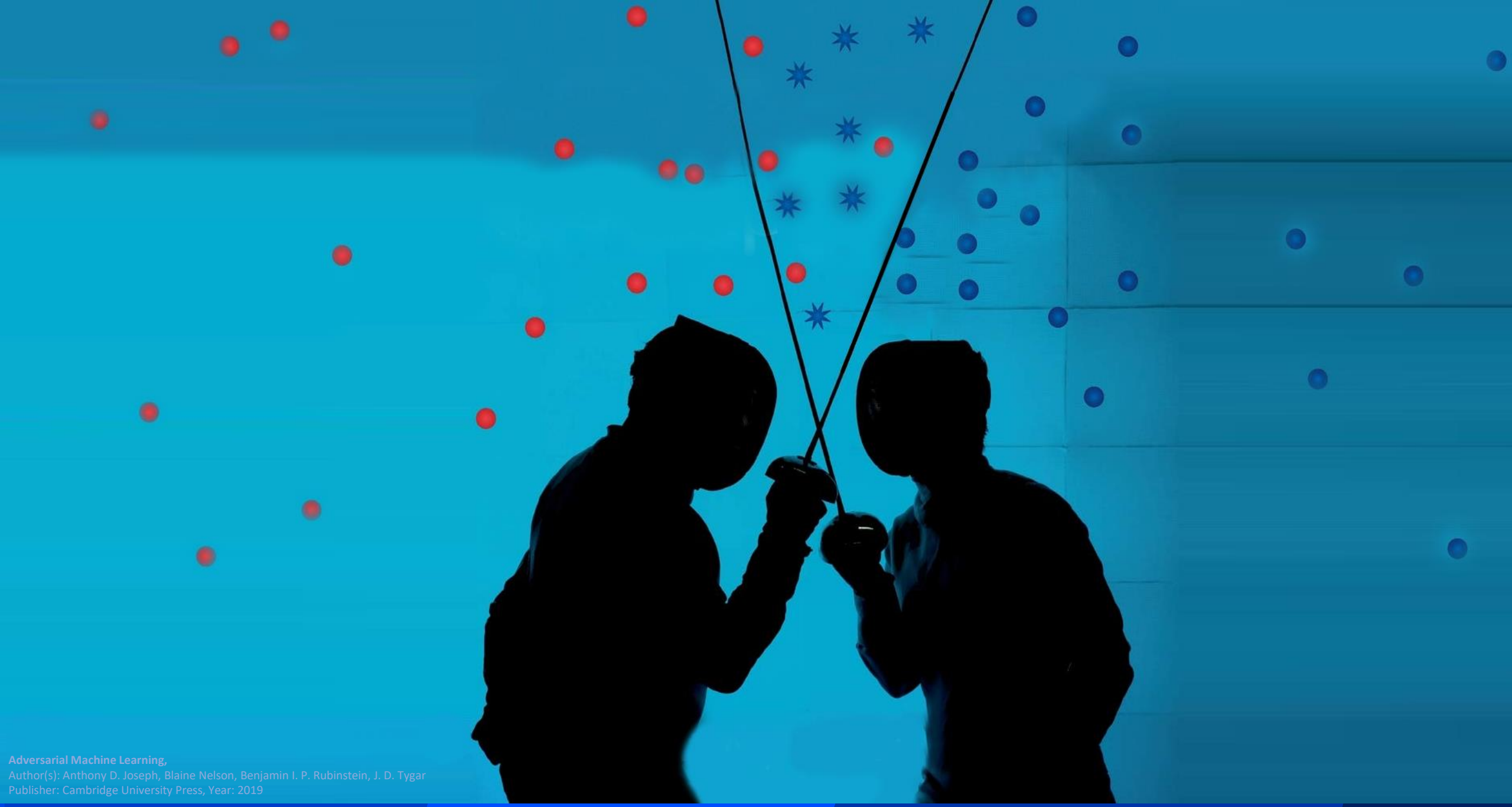


API based AI Model



API based AI Model





Adversarial Machine Learning,
Author(s): Anthony D. Joseph, Blaine Nelson, Benjamin I. P. Rubinstein, J. D. Tygar
Publisher: Cambridge University Press, Year: 2019

Motivations …

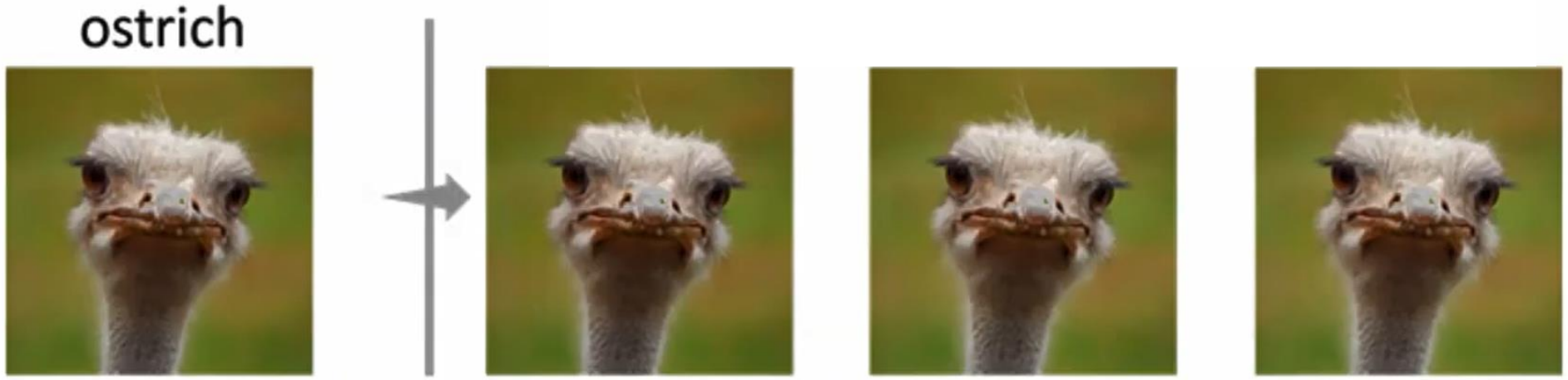
Suppose that we have

- Best trained **image classifier model** (using neural networks)

Motivations …

Suppose that we have

- Best trained **image classifier model** (using neural networks)



Motivations …

Suppose that we have

- Best trained **image classifier model** (using neural networks)

What is wrong with this AI model?

Review::

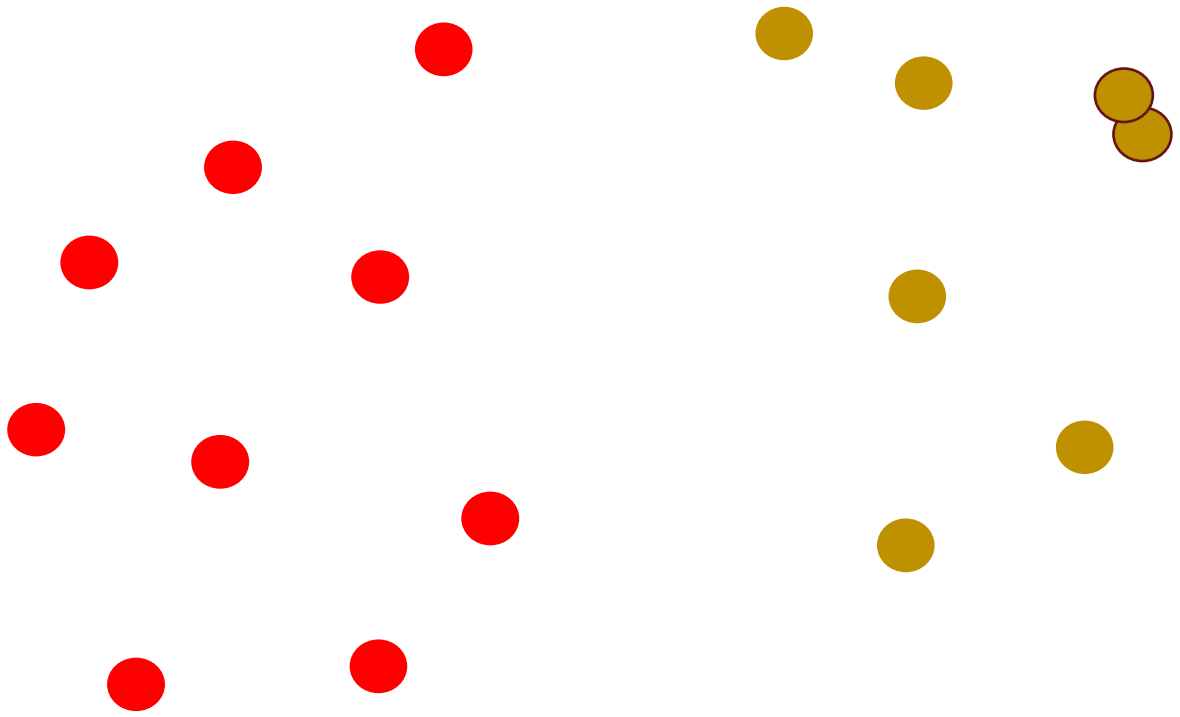
Review::

https://www.slideshare.net/BrianKim244/dcgan-77452250

Review::

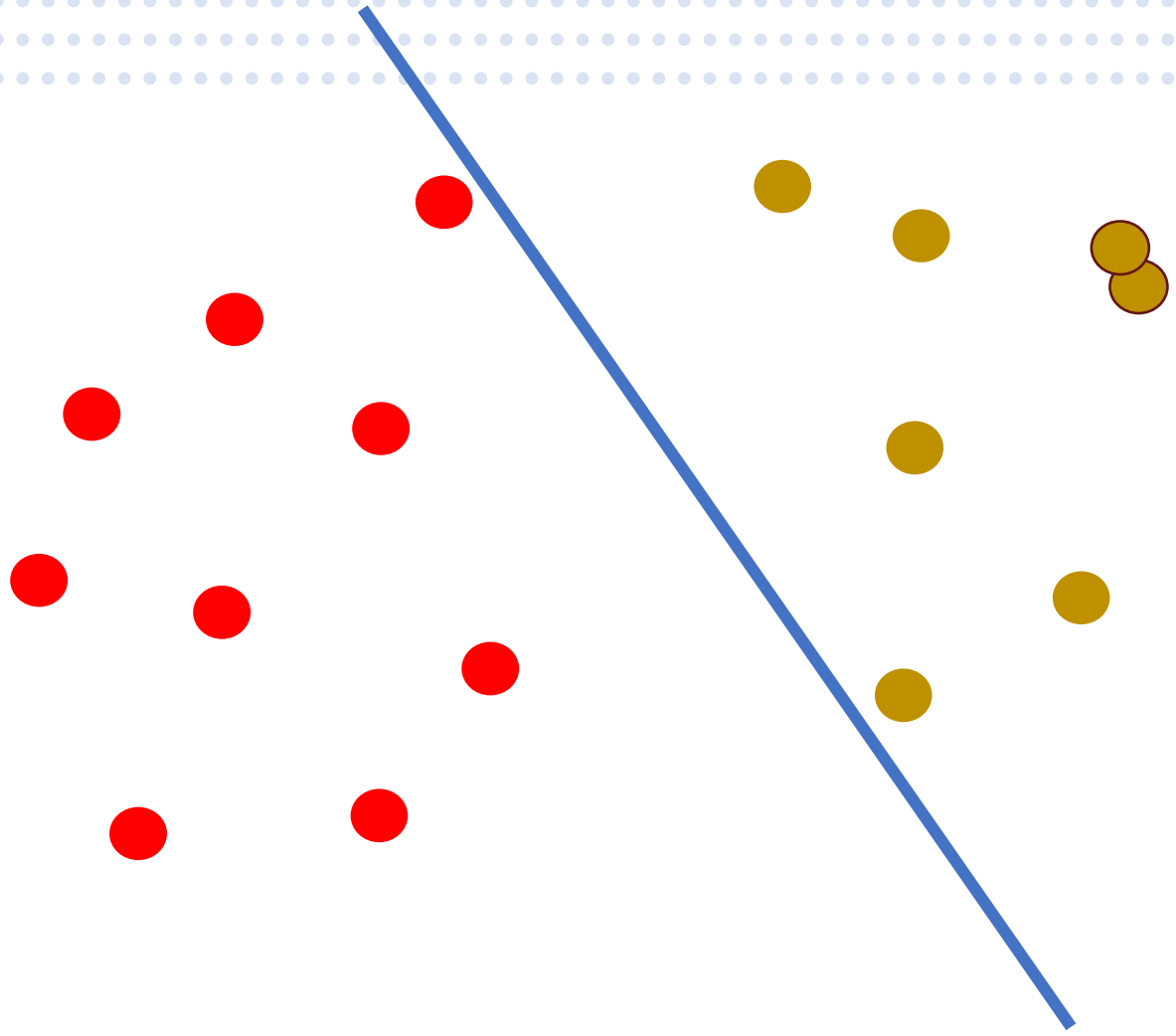
https://www.slideshare.net/BrianKim244/dcgan-77452250

Review::



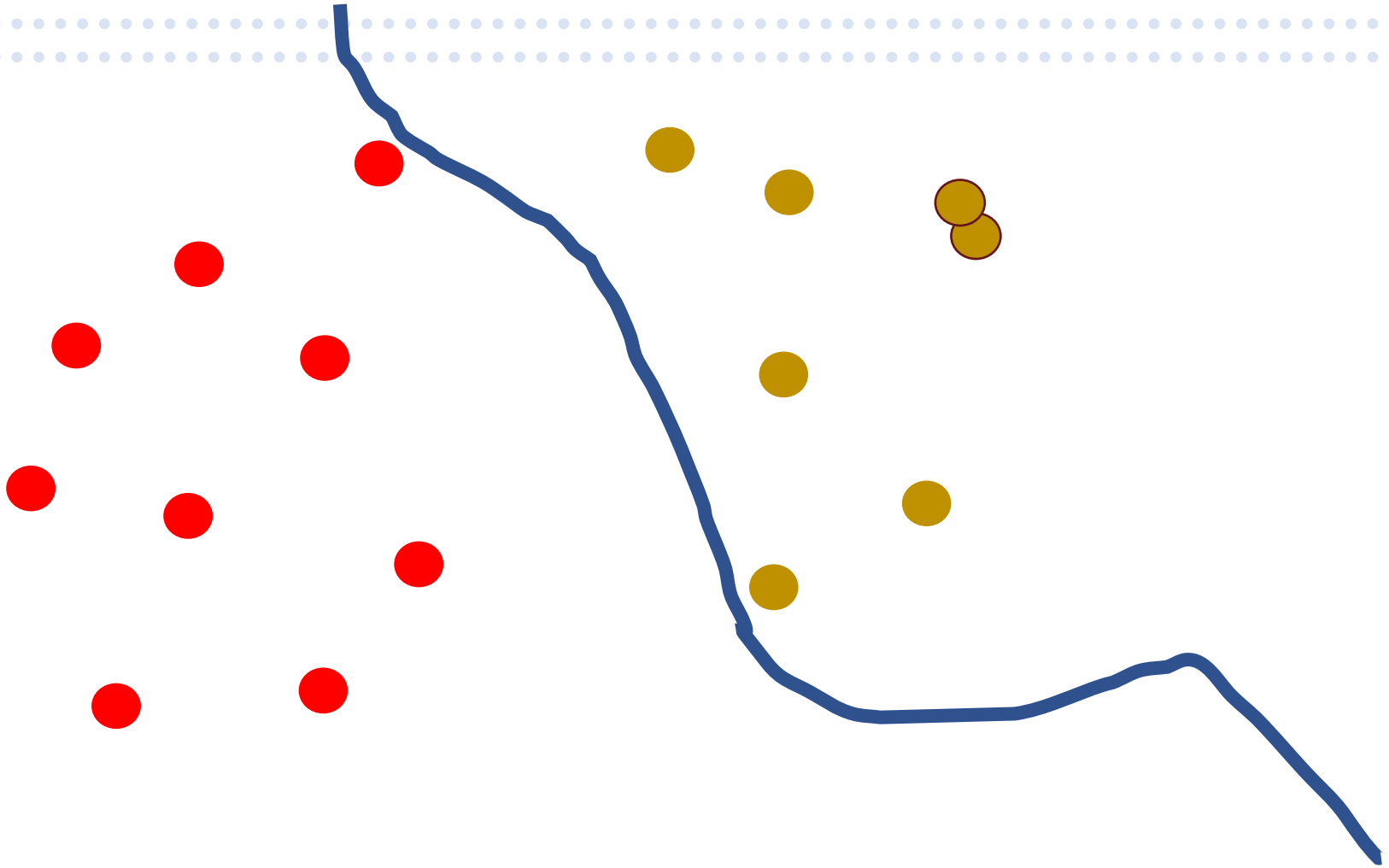
https://www.slideshare.net/BrianKim244/dcgan-77452250

Review::



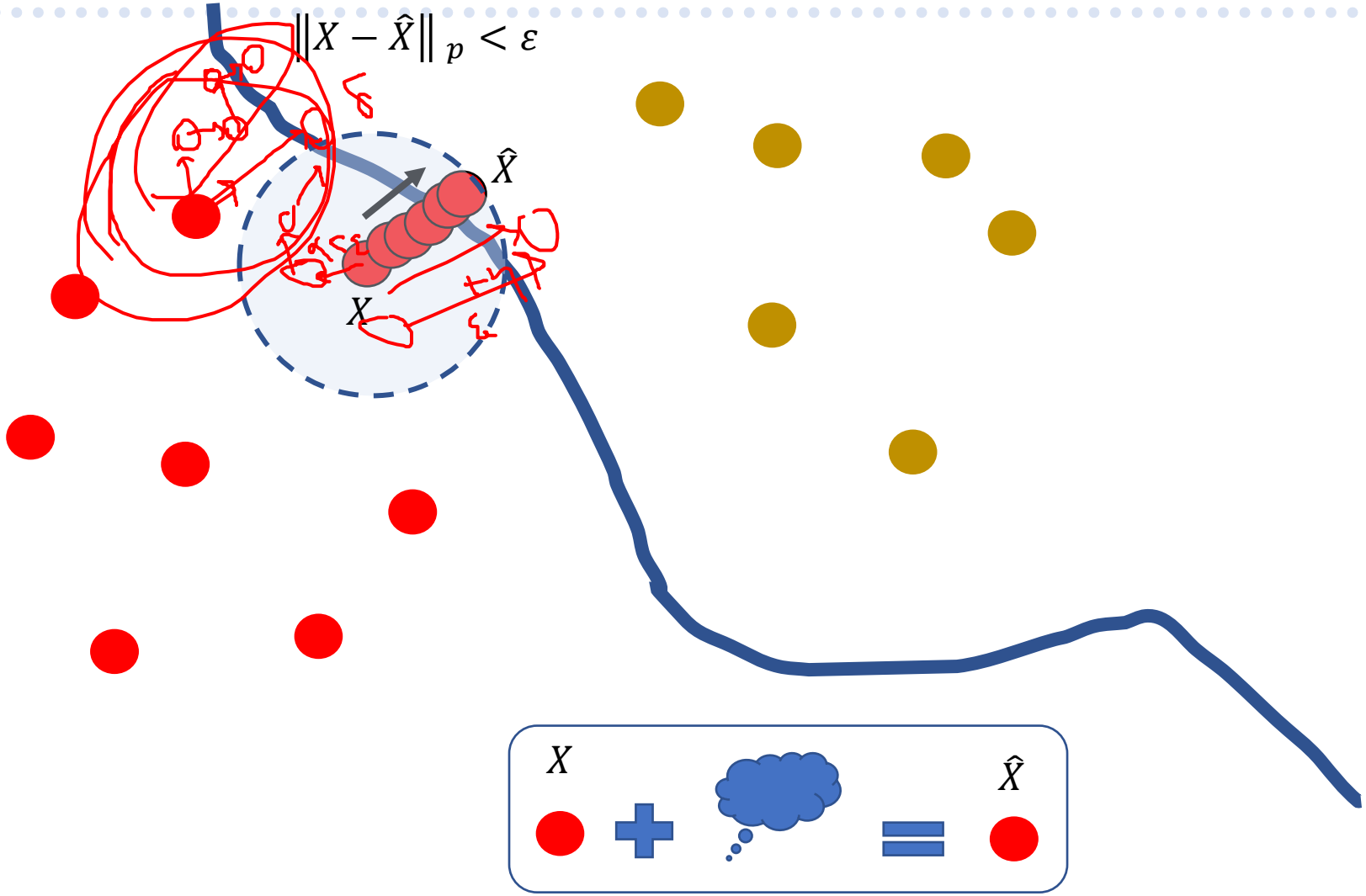
https://www.slideshare.net/BrianKim244/dcgan-77452250

Review::

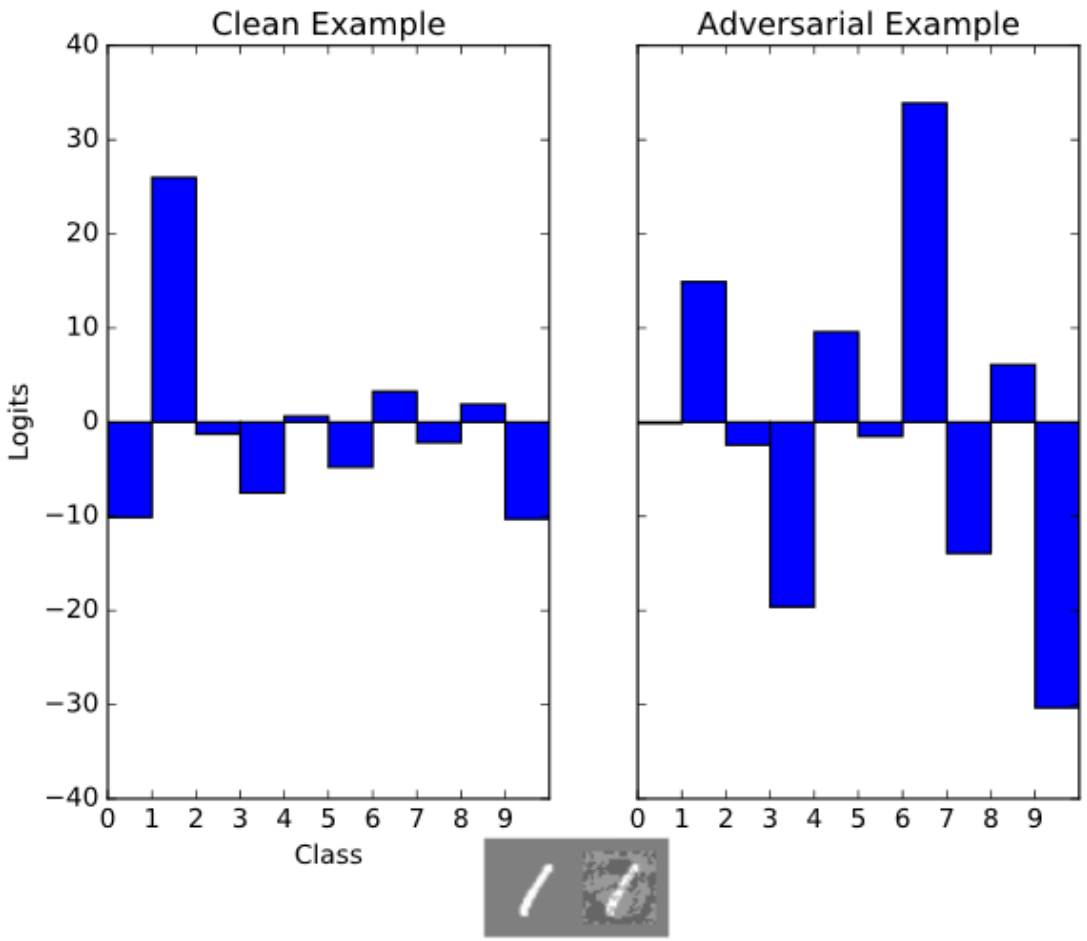


https://www.slideshare.net/BrianKim244/dcgan-77452250

What is Adversarial Example?



Summary


Clean Example
Adversarial Example
Logits
Class
40
30
20
10
0
−10
−20
−30
−40
0 1 2 3 4 5 6 7 8 9

Attack formulation

Given x, $f_\theta(.)$, the task is to compute x' such that

$$c^* = f_\theta(x) \neq f_\theta(x')$$

with some constraint like $\underbrace{\|x - x'\|_{\ell p}}_{\text{For } \ell_p \text{ attacks}} \leq \varepsilon$ to impose ***imperceptibility***

Attack formulation

$$\max_{\delta} l_{cls}\left(f_{\theta}\left(x'\right), y\right)$$

Perturbation

Loss

Classifier

Adv. Example

Label

Single-step attack

$$\begin{aligned}\delta_{\mathrm{FGSM}} &= \max_{\|\delta\| \leqslant \varepsilon} \langle \nabla l\left(f_\theta(x), y\right), \delta \rangle \\ &= \varepsilon \cdot \operatorname{sign}\left(\nabla l\left(f_\theta(x), y\right)\right)\end{aligned}$$

- **F**ast **G**radient **S**ign **M**ethod (Goodfellow et al., ICLR'15).
- Specifically designed for ℓ_∞ attacks.
- One-step attack.

Single-step attack

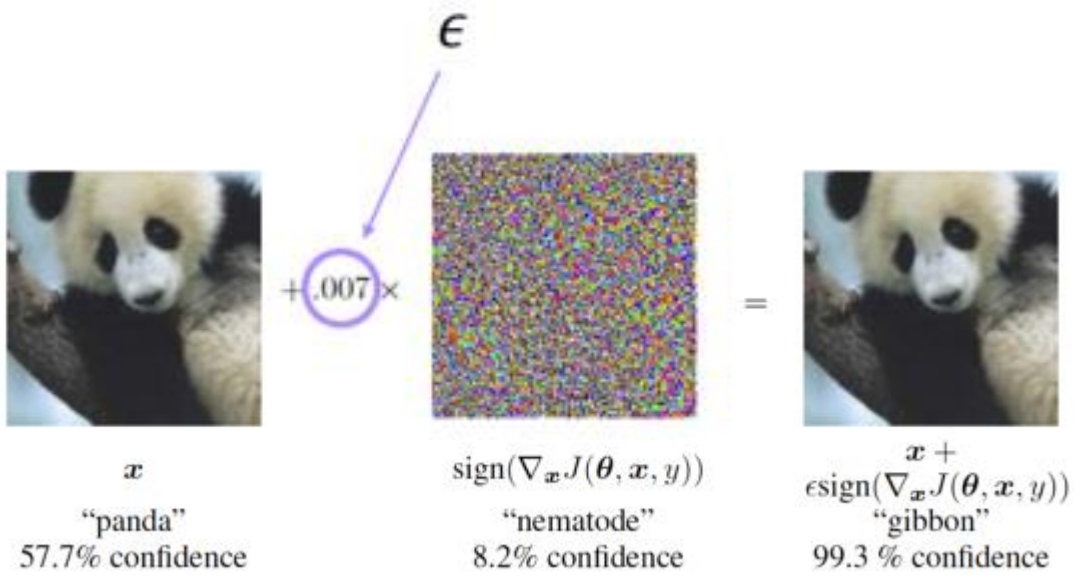
$$x' = x + \alpha \underbrace{\text{sign}\left(\nabla_x L(\theta, x, y)\right)}_{\text{FGSM}}$$

Perturbation

$$\tilde{x} = x + \eta$$

$$\eta = \epsilon \text{sign}\left(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y)\right)$$

Gradient of the cost function



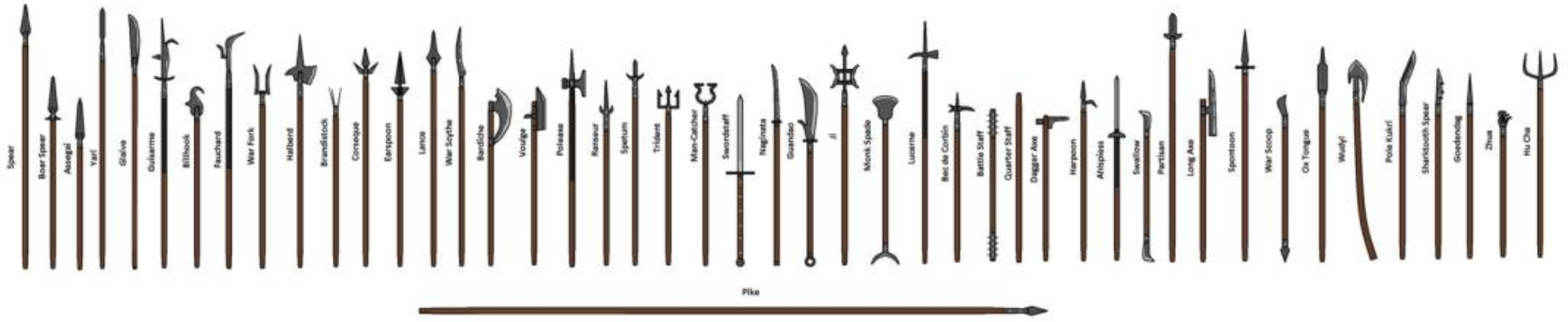
$\boldsymbol{x}$
"panda"
57.7% confidence

$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"nematode"
8.2% confidence

$\boldsymbol{x} + \epsilon \text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"gibbon"
99.3 % confidence

IBM CVPR21 tutorial // Explaining and harnessing 2015 Goodfellow

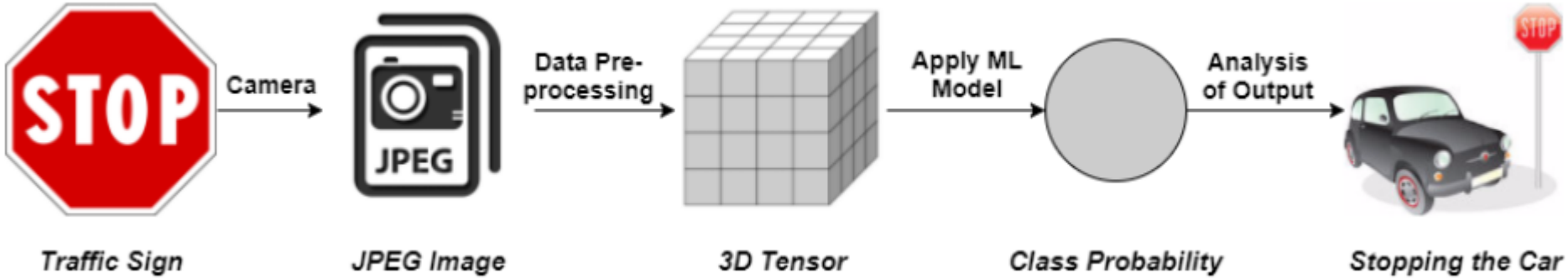
Various methods of attacks



https://www.slideshare.net/DavidKim486/universal-adversarial-perturbation

Adversarial Threat Model

- **The Attack Surface**



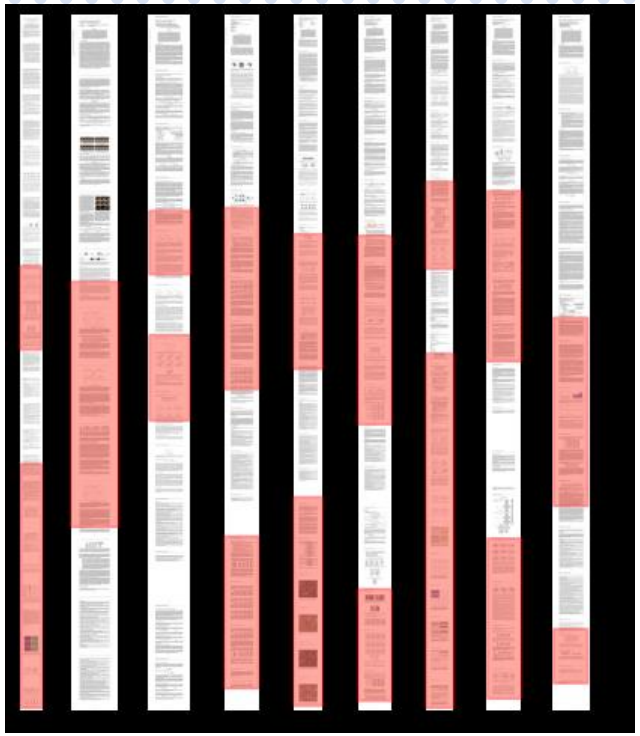
- Evasion Attack :: during the testing phase (* the most common type of attack!)
- Poisoning Attack :: during the training time
- Exploratory Attack :: during the testing phase (Given black box access to the model try to gain as much knowledge as possible)

http://fna.ir/a5g

Adversarial Attacks

- **Whitebox**: Attacker has access to the model parameters, outputs, etc.
- **Blackbox**: Attacker has only *query access* to the model and its outputs.

One important notes!



```
acc, loss = model.evaluate(x_test, y_test)
```

Is no longer sufficient.

https://blog.floydhub.com/introduction-to-adversarial-machine-learning/

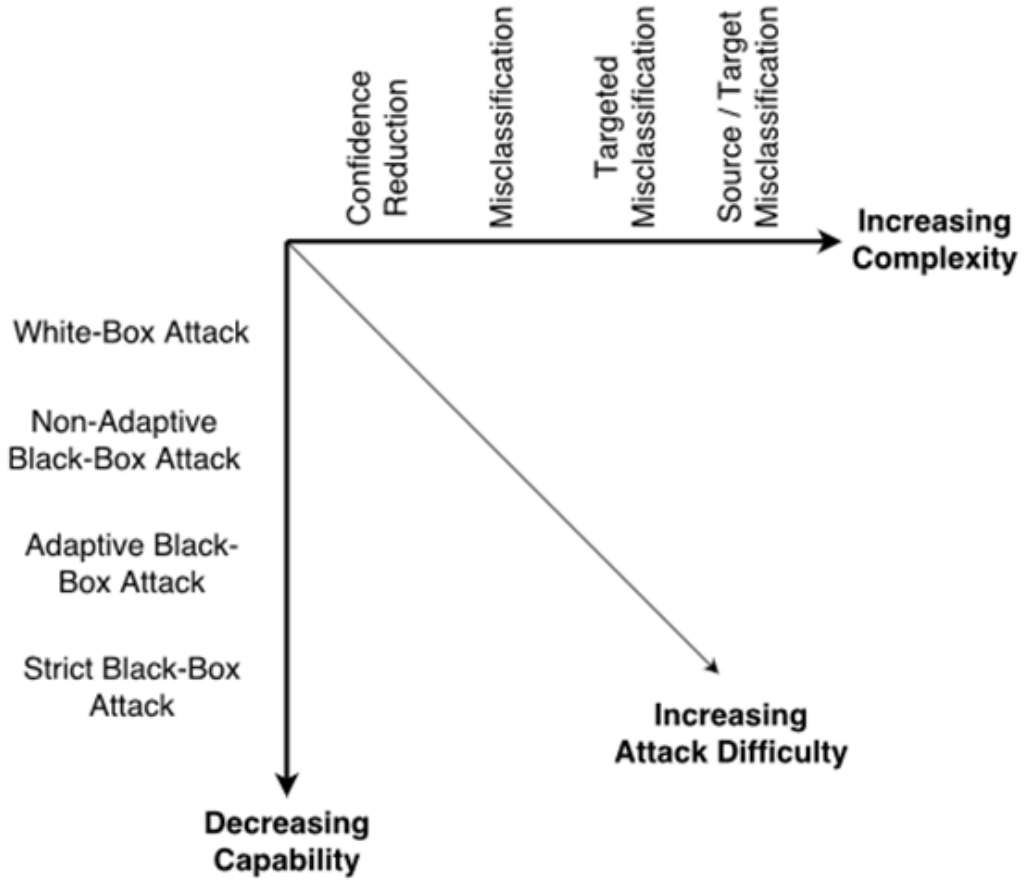

Break

Adversarial Threat Model

- **Adversarial Goals:**
  - Confidence Reduction
    - The adversary tries to reduce the confidence of prediction for the target model
  - Misclassification
    - The adversary tries to alter the output classification of an input example to **any** class different from the original class.
  - Targeted Misclassification
    - The adversary tries to produce inputs that force the output of the classification model to be a **specific** target class
  - Source/Target Misclassification
    - The adversary attempts to force the output of classification for a **specific** input to be a **particular** target class

http://fna.ir/a5g

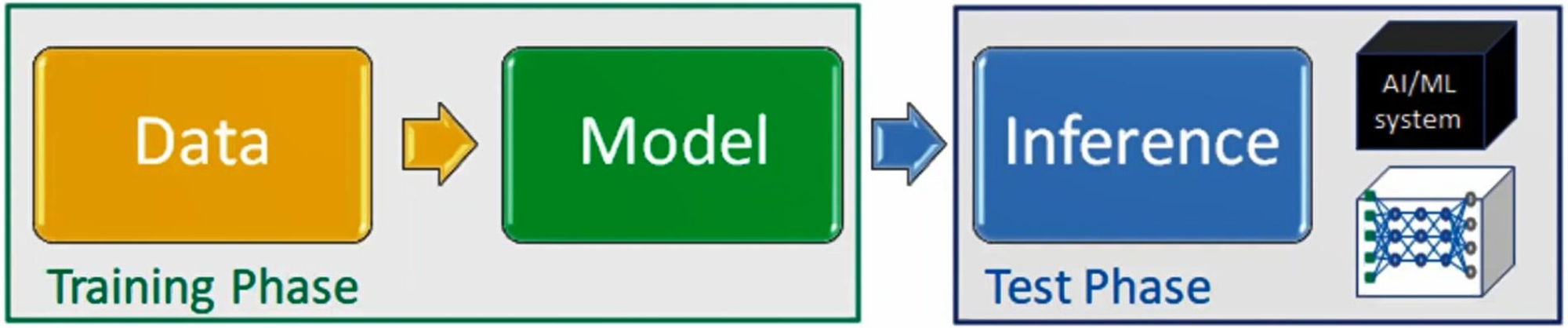
Adversarial Threat Model



Attack Difficulty with respect to adversarial capabilities and goals for Evasion Attacks

http://fna.ir/a5g

Holistic view of adversarial robustness



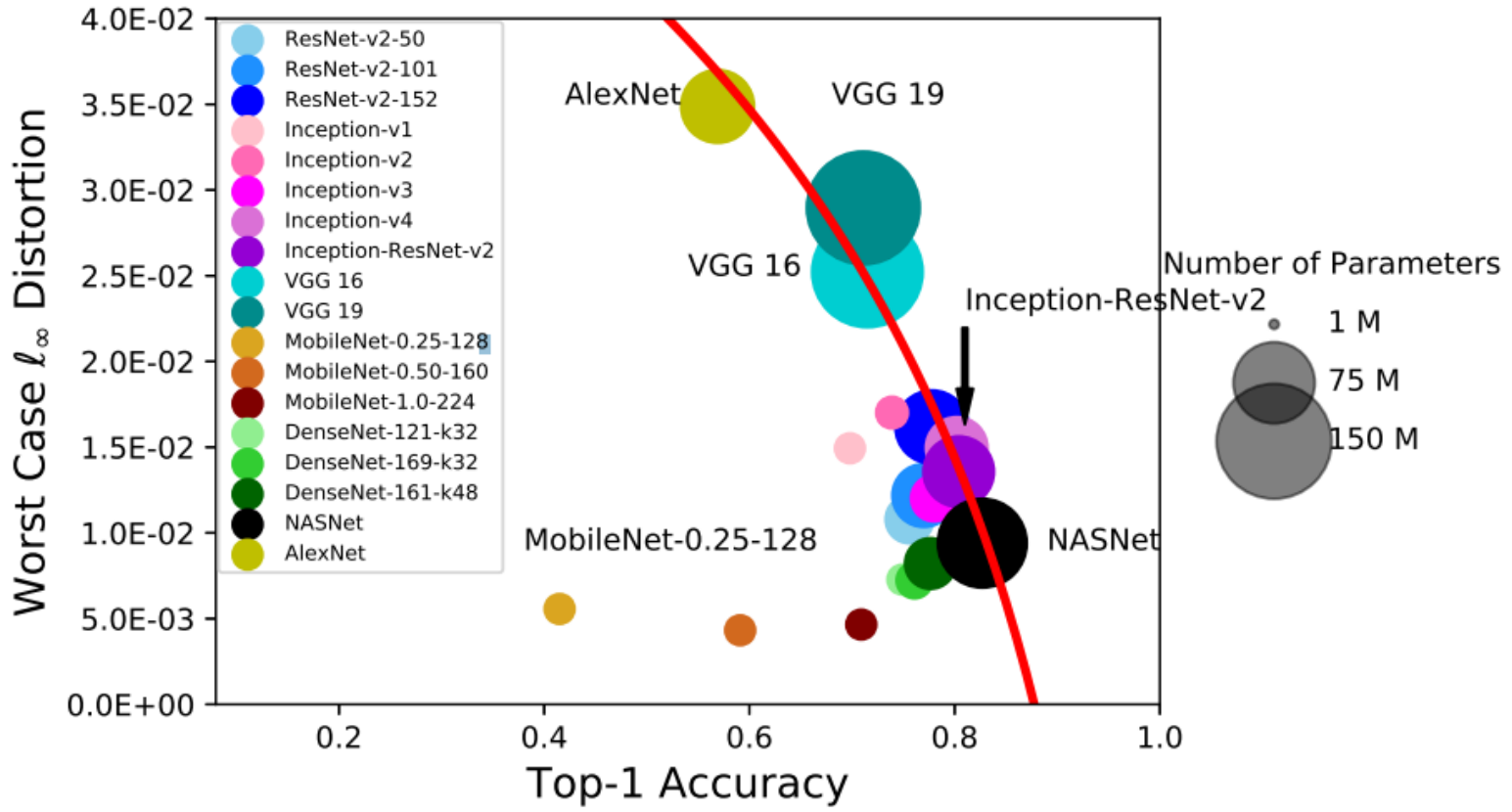
Attack Category / Attacker's reach	Data	Model / Training Method	Inference
Poisoning Attack [learning]	X	X*	
Backdoor Attack [learning]	X		
Evasion Attack (Adversarial Example) [learning]		X*	X
Extraction Attack (Model Stealing, Membership inference)			X
Model Injection [AI governance]		X*	X

IBM Research AI

*No access to model internal information in the black-box attack setting

Accuracy vs Adversarial Robustness

Robustness



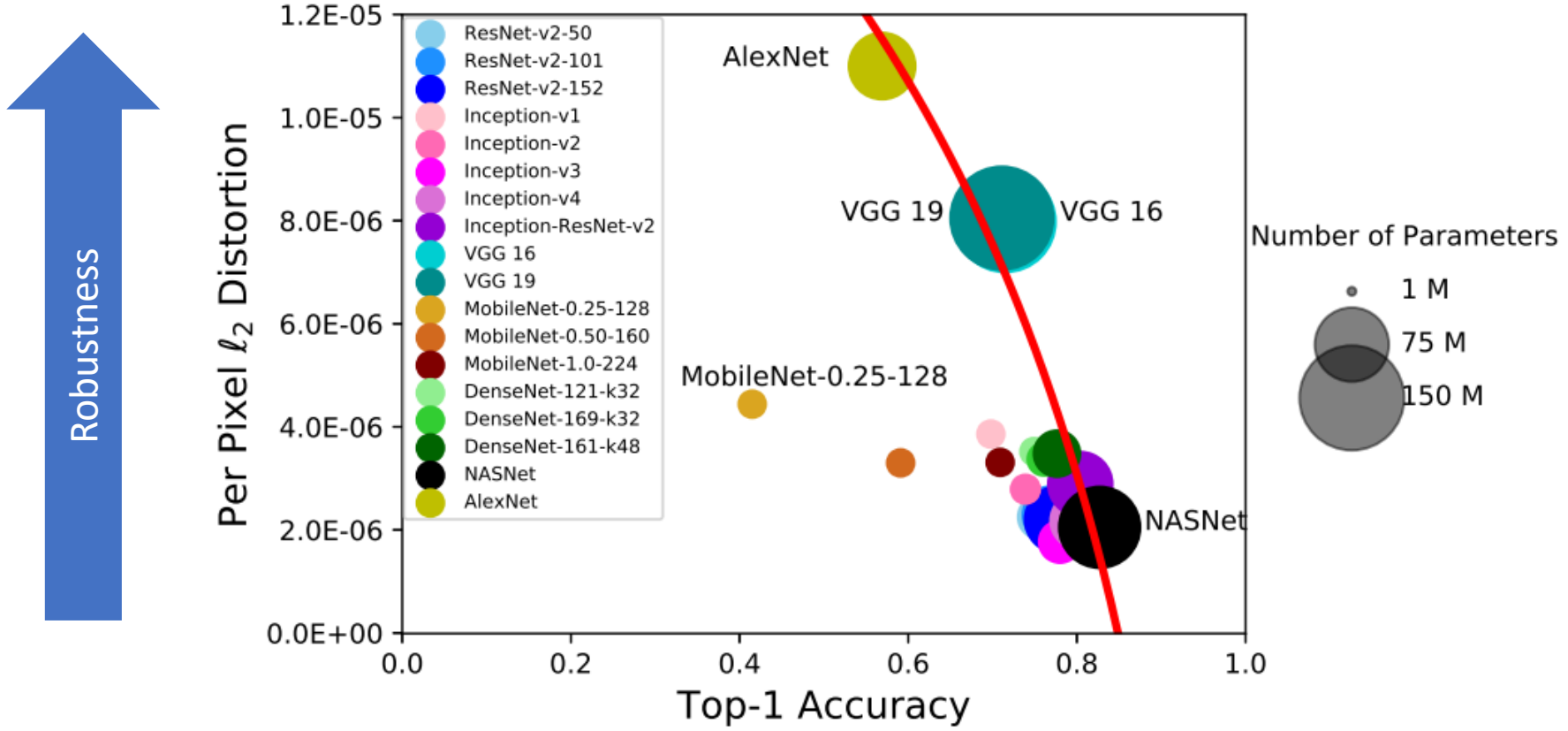
Our benchmark on 18 ImageNet models reveals a tradeoff in accuracy and robustness

(a) Fitted Pareto frontier of ℓ_∞ distortion (I-FGSM attack) vs. top-1 accuracy:

$$\ell_\infty \text{ dist} = [2.9 \cdot \ln(1 - \text{acc}) + 6.2] \times 10^{-2}$$

https://openaccess.thecvf.com/content_ECCV_2018/papers/Dong_Su_Is_Robustness_the_ECCV_2018_paper.pdf

Accuracy vs Adversarial Robustness



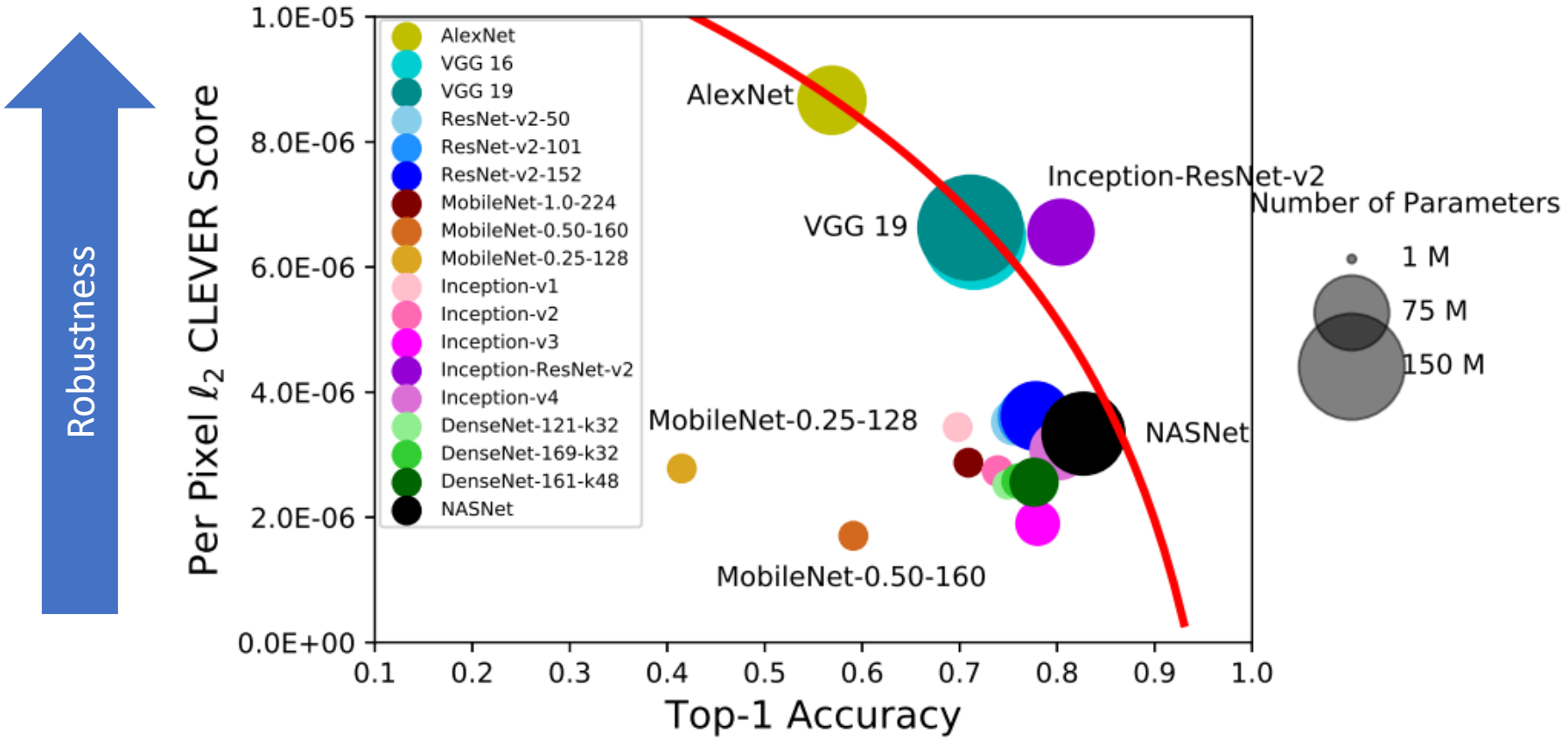
Our benchmark on 18 ImageNet models reveals a tradeoff in accuracy and robustness

(b) Fitted Pareto frontier of ℓ_2 distortion (C&W attack) vs. top-1 accuracy:

$$\ell_2 \text{ dist} = [1.1 \cdot \ln(1 - \text{acc}) + 2.1] \times 10^{-5}$$

https://openaccess.thecvf.com/content_ECCV_2018/papers/Dong_Su_Is_Robustness_the_ECCV_2018_paper.pdf

Accuracy vs Adversarial Robustness



Our benchmark on 18 ImageNet models reveals a tradeoff in accuracy and robustness

(c) Fitted Pareto frontier of ℓ_2 CLEVER score vs. top-1 accuracy:

$$\ell_2 \text{ score} = [4.6 \cdot \ln(1 - \text{acc}) + 12.5] \times 10^{-6}$$

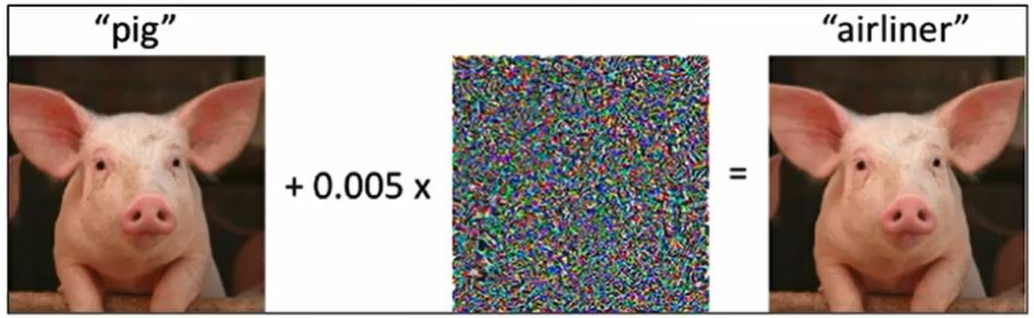
https://openaccess.thecvf.com/content_ECCV_2018/papers/Dong_Su_Is_Robustness_the_ECCV_2018_paper.pdf

Setting up the attack formulation problem

Given x, $f_\theta(.)$, the task is to compute x' such that

$$c^* = f_\theta(x) \neq f_\theta(x')$$

with some constraint like $\underbrace{\|x - x'\|_{\ell p}}_{\text{For } \ell_p \text{ attacks}} \leq \varepsilon$ to impose ***imperceptibility***



Targeted vs Untargeted

- An attacker can either launch targeted or an untargeted attacks.
- In targeted attacks, attack can set **t** where $f_\theta(x') = t \neq c^*$.

Given x, $f_\theta(.)$, the task is to compute x' such that

$$c^* = f_\theta(x) \neq f_\theta(x')$$

with some constraint like $\underbrace{\|x - x'\|_{\ell p}}_{\text{For } \ell_p \text{ attacks}} \leq \varepsilon$ to impose ***imperceptibility***

Adversarial attack

$$\max_{\delta} l_{cls}\left(f_{\theta}\left(x'\right), y\right)$$

Perturbation, Loss, Classifier, Adv. Example, Label

Maximize loss between a classifier's prediction on adversarial examples and their labels.

Adversarial attack

- For ℓ_p attacks, the following constraint holds: $\|x - x'\|_{\ell p} \leq \varepsilon$.
- There are non ℓ_p attacks too:
  - Adversarial translations
  - Functional adversarial attacks
  - Wasserstein attacks
  - Evolution attacks
- In this tutorial, we will primarily be dealing with ℓ_p attacks. These come under gradient-based attacks.

How do we perform optimization?

Case I: Single-step attack

$$\begin{aligned}\delta_{\mathrm{FGSM}} &= \max_{\|\delta\| \leqslant \varepsilon} \langle \nabla l\left(f_\theta(x), y\right), \delta \rangle \\ &= \varepsilon \cdot \operatorname{sign}\left(\nabla l\left(f_\theta(x), y\right)\right)\end{aligned}$$

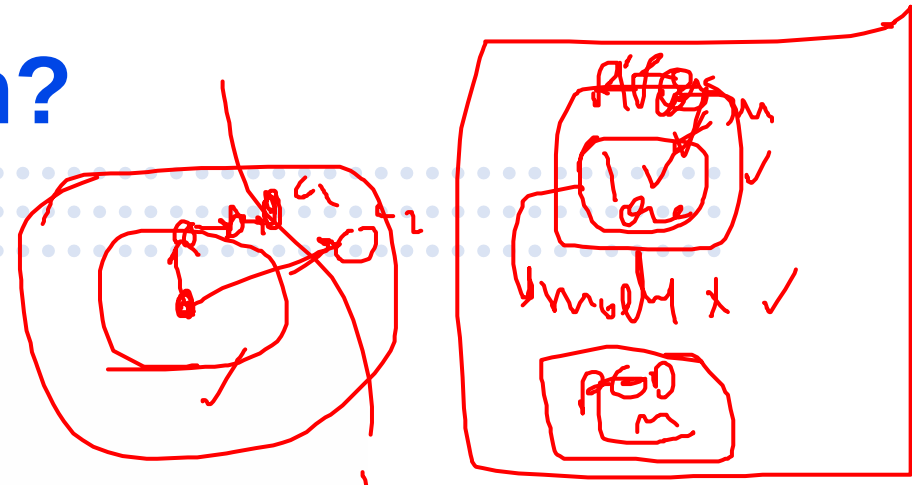
- **F**ast **G**radient **S**ign **M**ethod (Goodfellow et al., ICLR'15).
- Specifically designed for ℓ_∞ attacks.
- One-step attack.

How do we perform optimization?

Case II: Multi-step attack

$$x^{t+1} = \underbrace{\Pi_{x+\mathcal{S}}}_{\text{Projection}}\left(x^t + \alpha \underbrace{\operatorname{sign}\left(\nabla_x L(\theta, x, y)\right)}_{\text{FGSM}}\right)$$

How do we perform optimization?

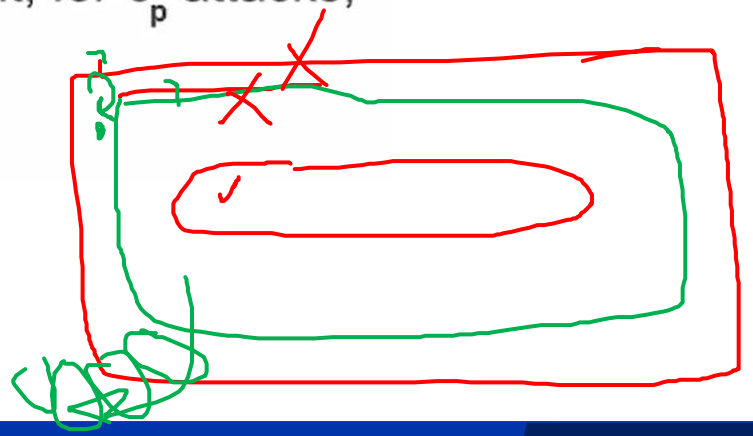


Case II: Multi-step attack

$$x^{t+1} = \underbrace{\Pi_{x+\mathcal{S}}}_{\text{Projection}}\left(x^t + \underbrace{\alpha\, \text{sign}\left(\nabla_x L(\theta, x, y)\right)}_{\text{FGSM}}\right)$$

For ℓ_∞, $\prod$ is $clip(x^{t+1}, x^0 - \epsilon, x^0 + \epsilon)$

- Projected Gradient Descent (PGD, Madry et al., ICLR'18)
- Since α (step size) can break the norm-ball constraint, for ℓ_p attacks, projection is required after adding the perturbation.
- Hence Projected Gradient Descent (PGD).

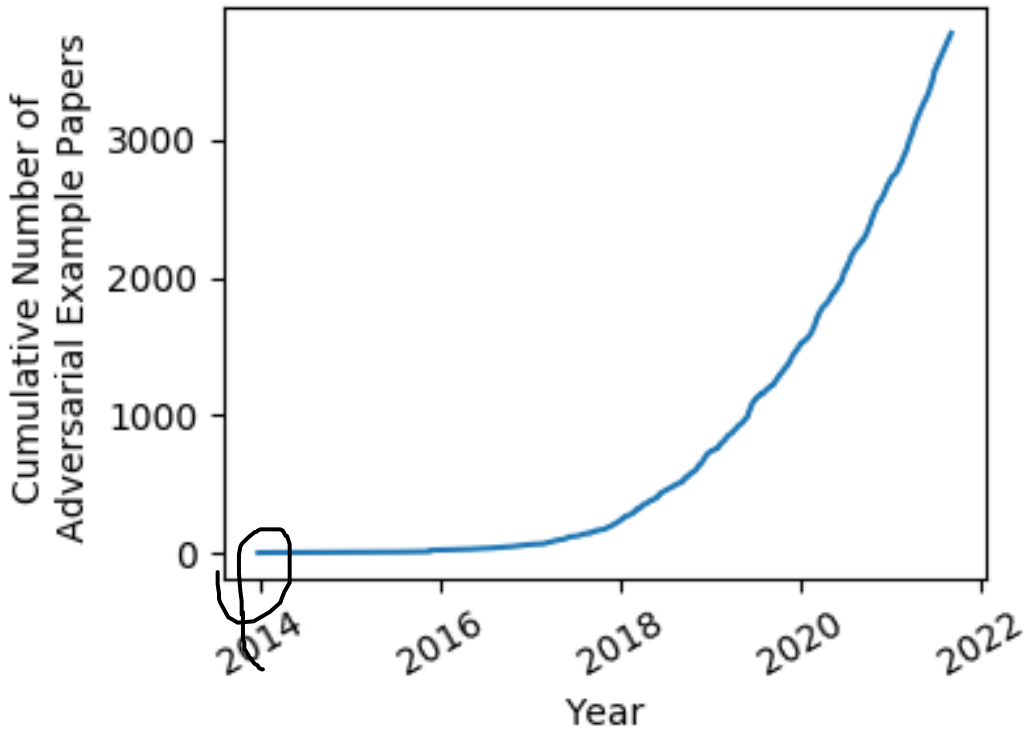


How do we perform optimization?

- There are other and more recent of gradient-based attacks trying to solve the maximization problem in clever ways.
- Challenges remain in defining precisely what human perception is.
- This is important because a strong adversarial attack should keep the changes as imperceptible as possible.

IBM CVPR21 tutorial

Adversarial example papers



https://nicholas.carlini.com/writing/2019/all-adversarial-example-papers.html

Adversarial example papers

Nicholas Carlini

Research Scientist, Google Brain

- https://nicholas.carlini.com/writing/2019/all-adversarial-example-papers.html

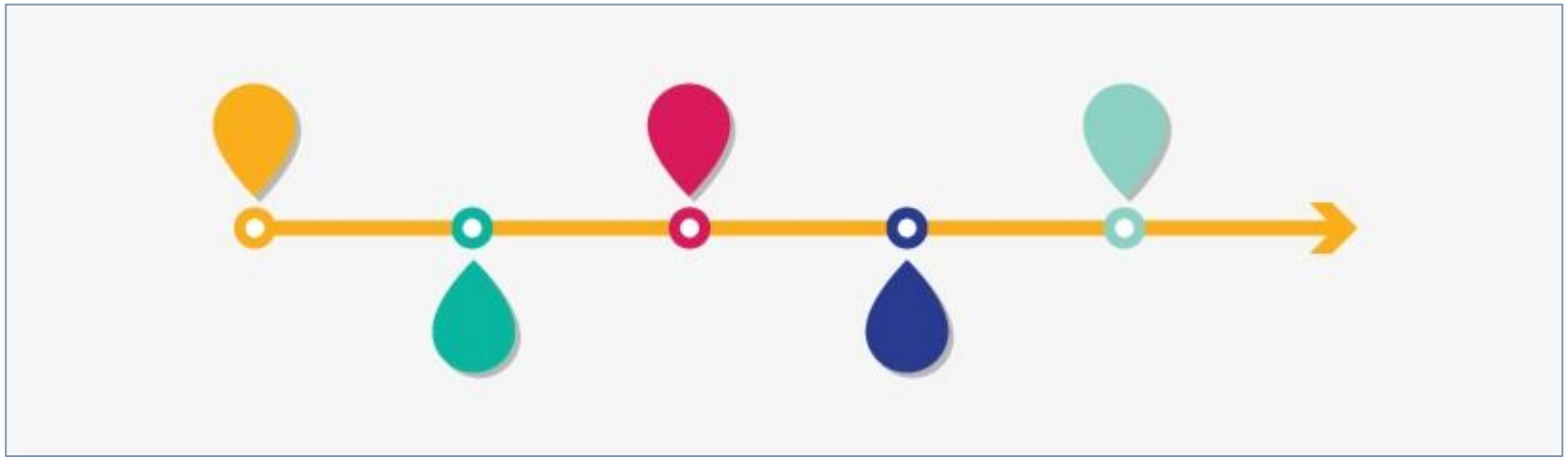
RobustML

- https://www.robust-ml.org/defenses/
- https://paperswithcode.com/task/adversarial-attack

https://nicholas.carlini.com/writing/2019/all-adversarial-example-papers.html


Break

Adversarial example papers

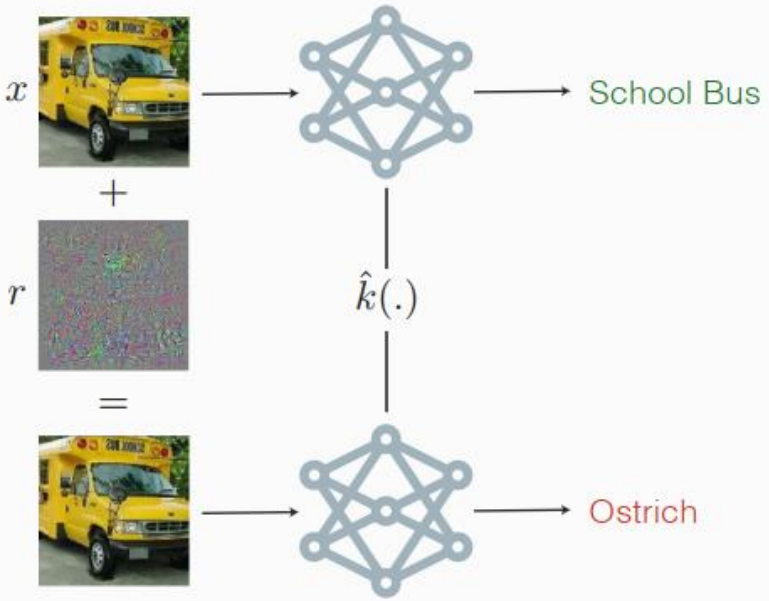


Adversarial example papers

Interiguing properties of Neural Network

C. Szegedy et al. ICLR, 2014

$$r^* = \arg\min_r J(\hat{k}(x+r), y) + C\|r\|$$



Adversarial example papers

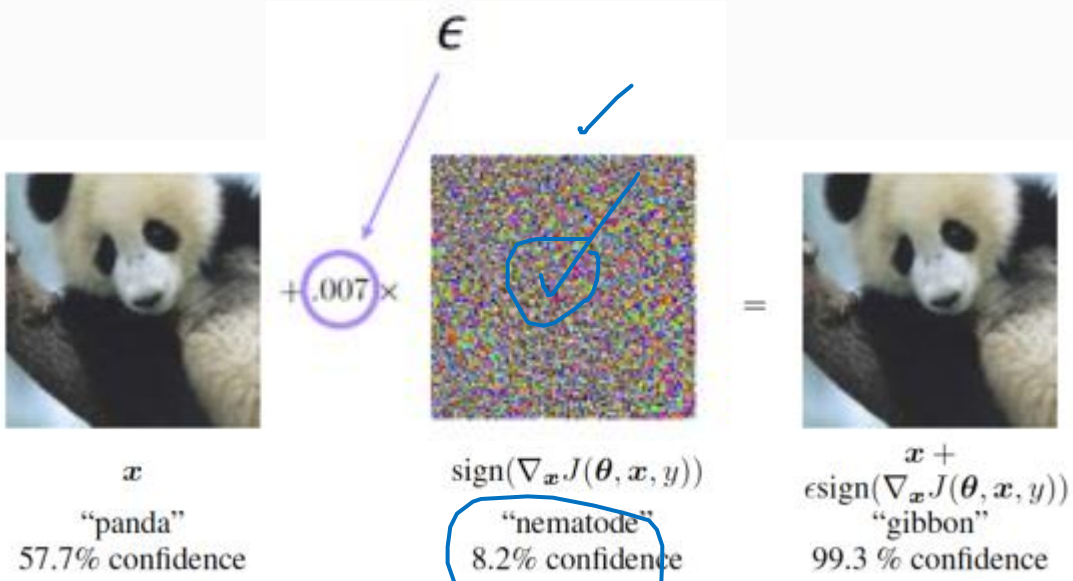
Explaining and Harnessing Adversarial Example
I.J. Goodfellow et al. ICLR, 2015

Perturbation

$$\tilde{x} = x + \eta$$

$$\eta = \epsilon \text{sign}\left(\nabla_x J(\theta, x, y)\right)$$

Gradient of the cost function



Adversarial example papers

Explaining and Harnessing Adversarial Example
I.J. Goodfellow et al. ICLR, 2015

Perturbation

$$\tilde{x} = x + \eta$$

$$\eta = \epsilon \text{sign}\left(\nabla_x J(\theta, x, y)\right)$$

Gradient of the cost function

Adversarial Training

$$\text{loss}(x, y) + \text{loss}(x + \epsilon \cdot \texttt{sign}(\text{grad}), y)$$

Small when prediction is correct on legitimate input

Small when prediction is correct on adversarial input

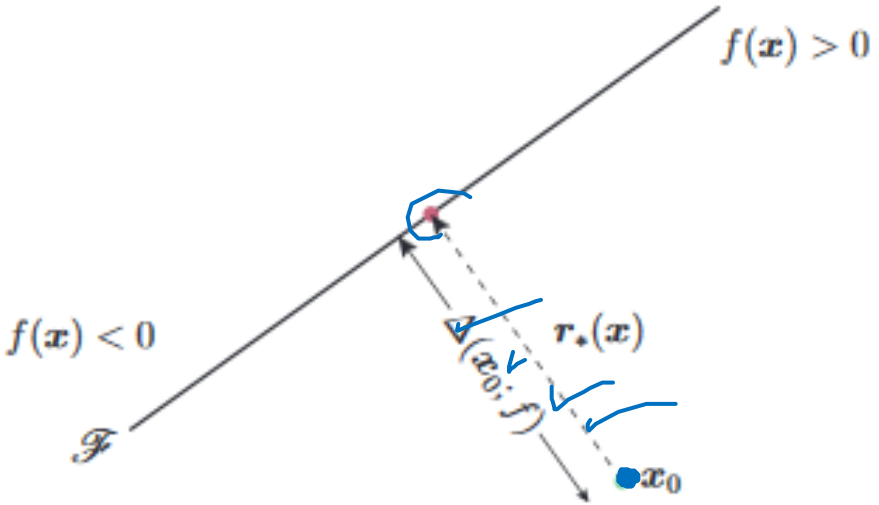
2004 2013 2014 2015 2016 2017 2018 2019 2020

Biggio et al
Szegedy et al
Biggio et al
Goodfellow et al
Kurakin et al
Papernot et al
Carlini & Wagner
Papernot et al
Grosse et al
Madry et al

Adversarial example papers

Deep Fool: a simple and accurate method to fool deep neural networks

S.M. Moosavi-dezfooli et al. CVPR, 2016



Algorithm 1 DeepFool for binary classifiers

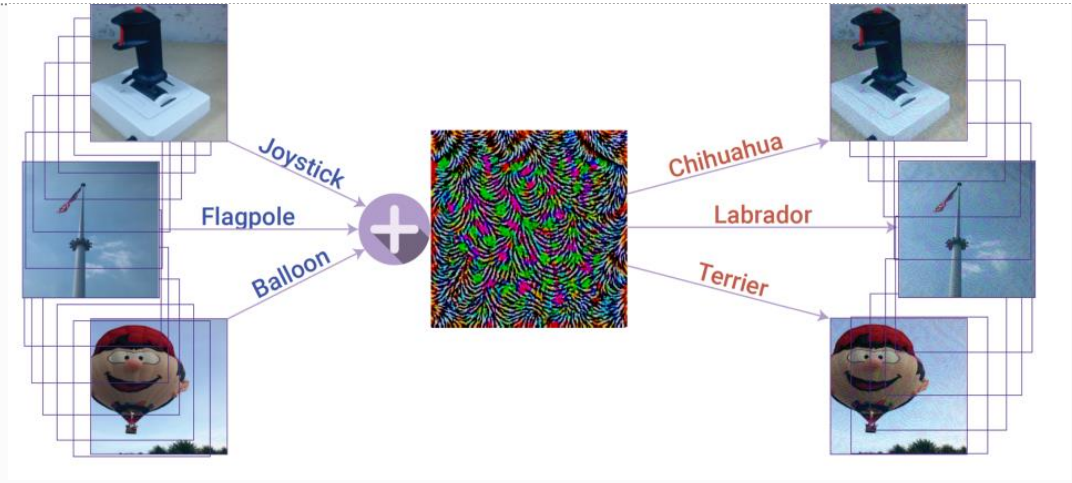
input: Image x, classifier f.
output: Perturbation $\hat{r}$.
Initialize $x_0 \leftarrow x$, $i \leftarrow 0$.
while $\text{sign}(f(x_i)) = \text{sign}(f(x_0))$ **do**
  $r_i \leftarrow -\frac{f(x_i)}{\|\nabla f(x_i)\|_2^2} \nabla f(x_i)$,
  $x_{i+1} \leftarrow x_i + r_i$,
  $i \leftarrow i + 1$.
end while
return $\hat{r} = \sum_i r_i$.

Adversarial example papers

Universal adversarial perturbations
S.M. Moosavi-dezfooli et al. CVPR, 2017

$$f(x+\eta) \neq f(x) \quad for\ "most"\ \ x \sim X$$

$$\|\eta\|_p \leq \xi$$
$$\mathbb{P}_{x \sim X}\big(f(x+\eta) \neq f(x)\big) \geq 1-\delta$$



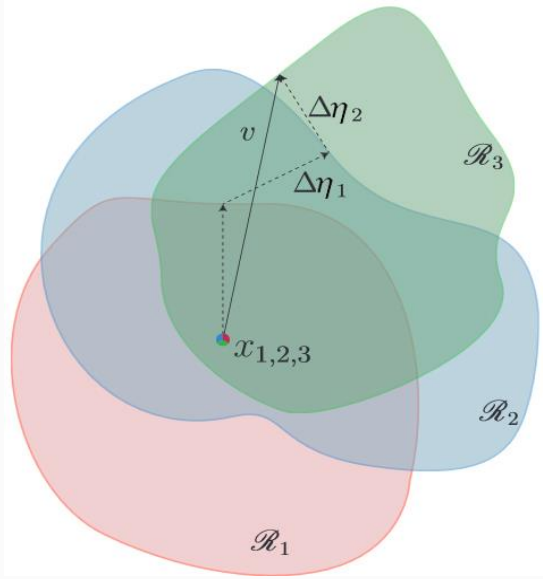
$$\Delta\eta \leftarrow arg\min_r \|r\|_2 \quad s.t. \quad f(x_i+\eta+r) \neq f(x_i)$$

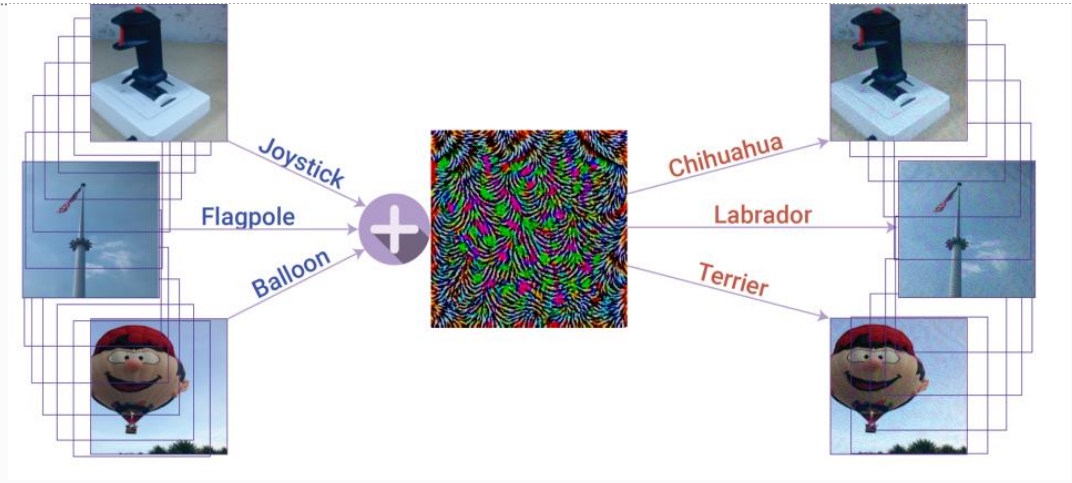
$$\eta \leftarrow \mathcal{P}_{p,\xi}(\eta + \Delta\eta_i)$$

$$\mathcal{P}_{p,\xi}(\eta) = \arg\min_{\eta'} \|\eta - \eta'\|_2 \quad s.t. \quad \|\eta'\|_p \leq \xi$$

Adversarial example papers

Universal adversarial perturbations
S.M. Moosavi-dezfooli et al. CVPR, 2017



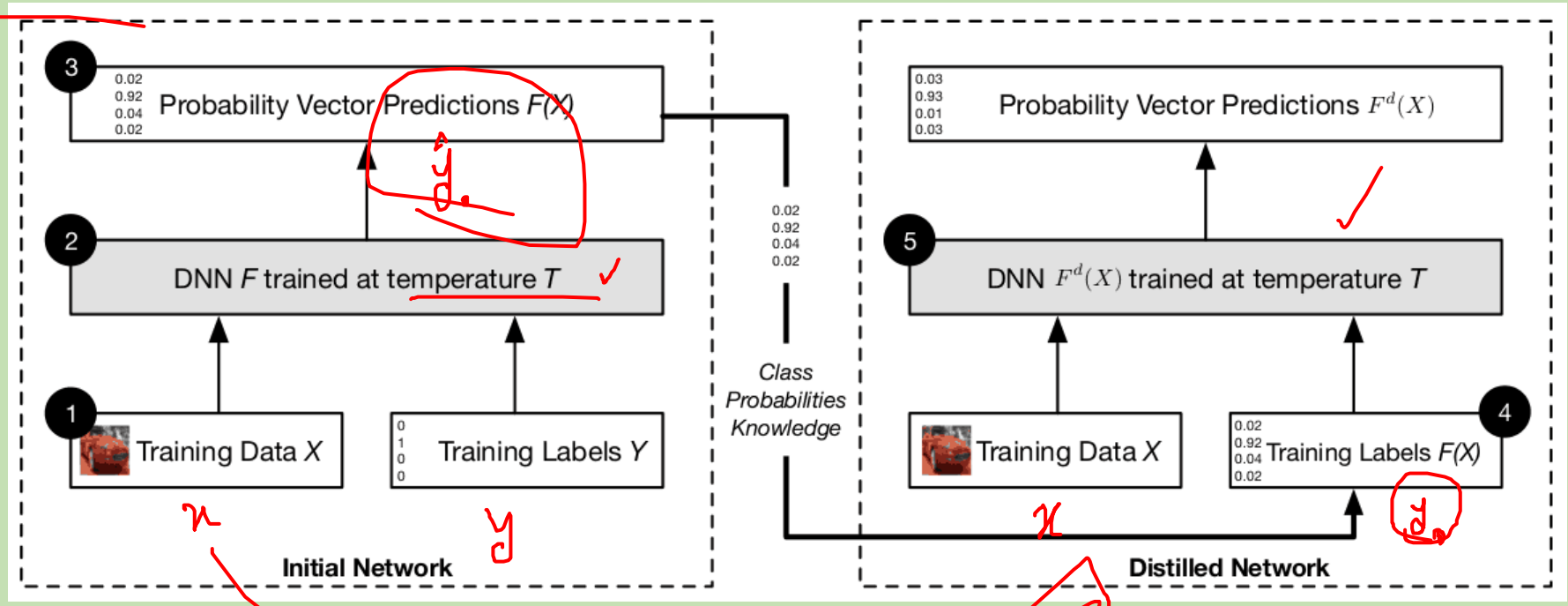


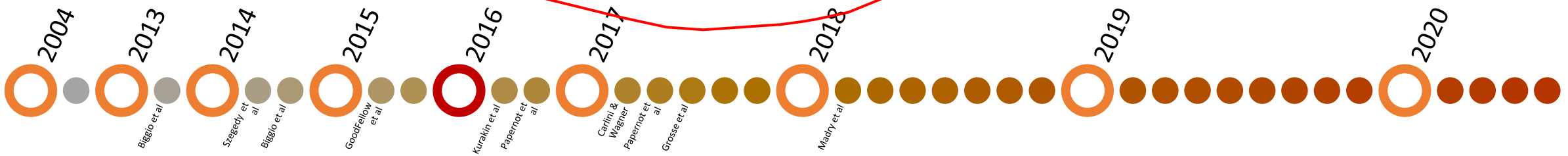
$$\mathcal{P}_{p,\xi}(\eta) = \arg\min_{\eta'} \|\eta - \eta'\|_2 \qquad s.t. \quad \|\eta'\|_p \leq \xi$$

Adversarial example papers

Distillation as a Defense to Adversarial Perturbations against Deep Neural Networks

Nicolas Papernot et al. IEEE Symposium on Security and Privacy, 2016





Adversarial example papers

Towards evaluating the robustness of neural networks

Carlini et al. Security and Privacy (S&P). 2017

$$\begin{aligned}&\text{minimize} \quad \mathcal{D}(x, x+\delta)\\ &\text{such that} \quad C(x+\delta) = t\\ &\qquad\quad x+\delta \in [0,1]^n\end{aligned}$$

$$\begin{aligned}&\text{minimize} \quad \mathcal{D}(x, x+\delta)\\ &\text{such that} \quad f(x+\delta) \le 0\\ &\qquad\quad x+\delta \in [0,1]^n\end{aligned}$$

$$\begin{aligned}
f_1(x') &= -\text{loss}_{F,t}(x') + 1\\
f_2(x') &= (\max_{i\neq t}(F(x')_i) - F(x')_t)^+\\
f_3(x') &= \text{softplus}(\max_{i\neq t}(F(x')_i) - F(x')_t) - \log(2)\\
f_4(x') &= (0.5 - F(x')_t)^+\\
f_5(x') &= -\log(2F(x')_t - 2)\\
f_6(x') &= (\max_{i\neq t}(Z(x')_i) - Z(x')_t)^+\\
f_7(x') &= \text{softplus}(\max_{i\neq t}(Z(x')_i) - Z(x')_t) - \log(2)
\end{aligned}$$

$$\begin{aligned}&\text{minimize} \quad \mathcal{D}(x, x+\delta) + c\cdot f(x+\delta)\\ &\text{such that} \quad x+\delta \in [0,1]^n\end{aligned}$$

$$\begin{aligned}&\text{minimize} \quad \|\delta\|_p + c\cdot f(x+\delta)\\ &\text{such that} \quad x+\delta \in [0,1]^n\end{aligned}$$

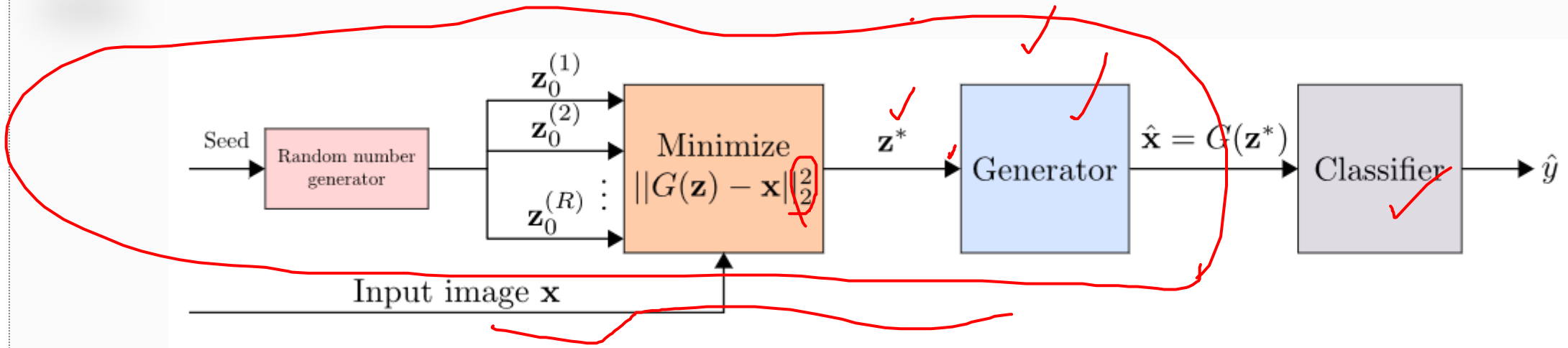
$$\delta_i = \frac{1}{2}(\tanh(w_i)+1) - x_i$$

$$\begin{aligned}&\text{minimize} \quad \|\frac{1}{2}(\tanh(w)+1) - x\|_2^2 + c\cdot f(\frac{1}{2}(\tanh(w)+1)\\ &\text{with } f \text{ defined as}\\ &f(x') = \max(\max\{Z(x')_i : i \neq t\} - Z(x')_t, -\kappa).\end{aligned}$$

Adversarial example papers

Defense-GAN: Protecting Classifiers Against Adversarial Attacks Using Generative Models

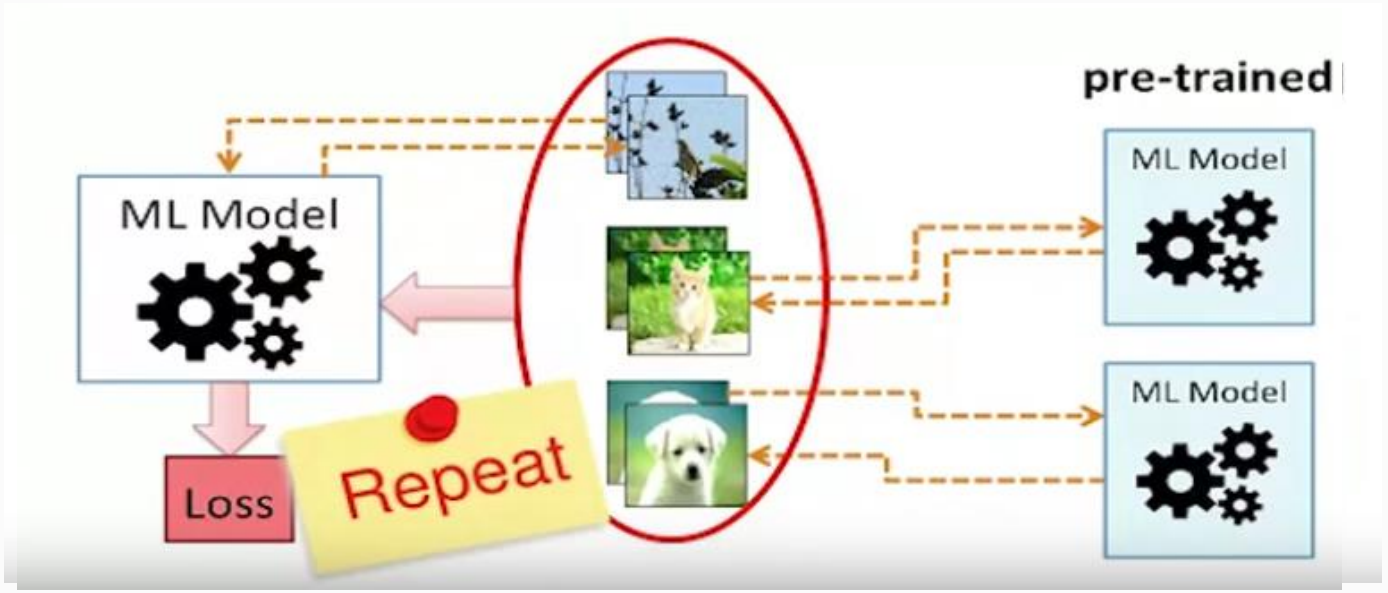
Samangouei et al. ICLR, 2018



Adversarial timeline …

Ensemble Adversarial Training: Attacks and Defenses

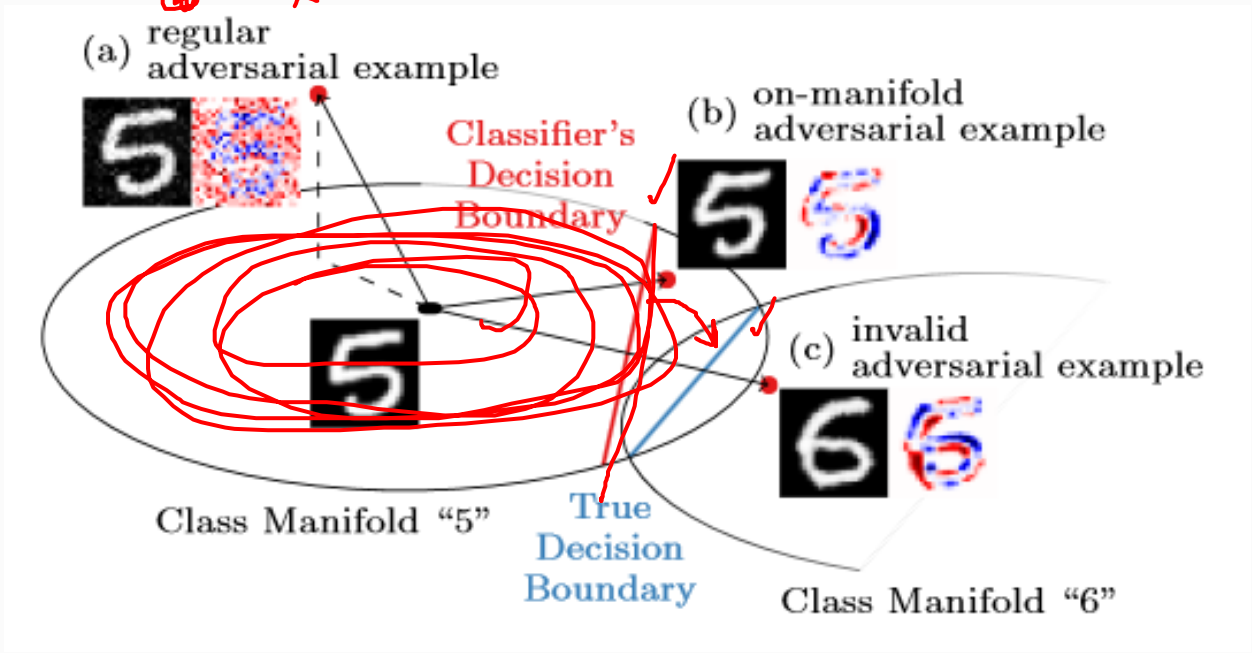
Tramèr et al. ICLR, 2018



Adversarial example papers

Disentangling Adversarial Robustness and Generalization

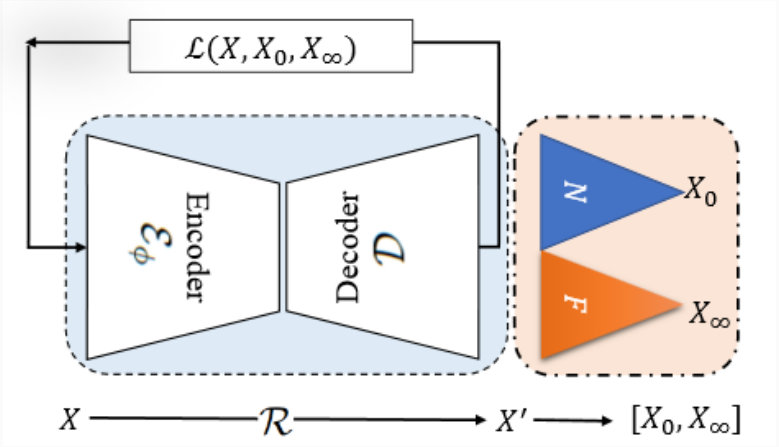
Stutz et al. CVPR, 2019

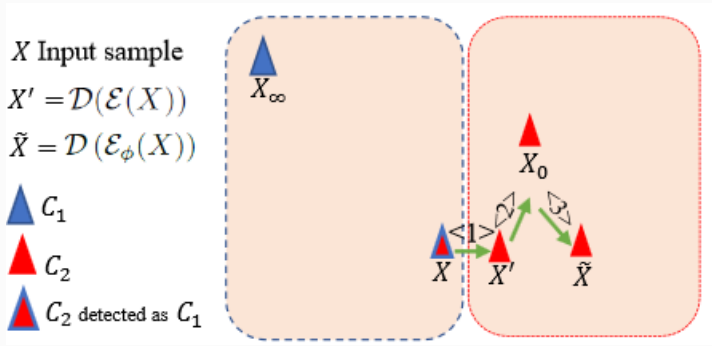


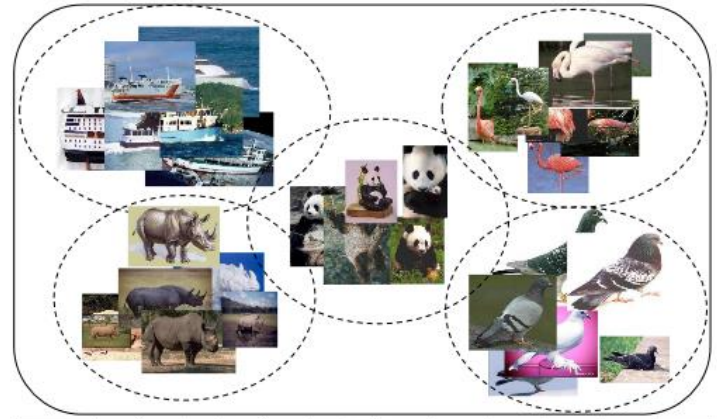
Adversarial timeline …

Self-Supervised Representation Learning via Neighborhood-Relational Encoding

Sabokrou, Khalooei, Adeli et al. ICCV, 2019



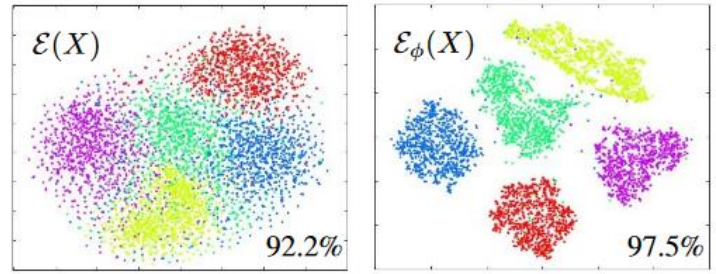




$$\mathbb{S}(\mathbf{a}, \mathbf{b}) = \frac{\mathbf{a} \cdot \mathbf{b}}{\|\mathbf{a}\|\|\mathbf{b}\|} = \frac{\sum_{i=1}^{n} a_i b_i}{\sqrt{\sum_{i=1}^{n} a_i^2}\sqrt{\sum_{i=1}^{n} b_i^2}}$$

$$\mathbb{D}(\cdot, \cdot) = 1 - \mathbb{S}(\cdot, \cdot)$$

$$\mathcal{L} = \lambda_1 \mathbb{D}(\mathcal{R}_\mathcal{A}(X), \mathcal{R}_\mathcal{A}(X')) + \lambda_2 \sum_{i=1}^{T} \mathbb{D}(\mathcal{R}_\mathcal{A}(X'), \mathcal{R}_\mathcal{A}(X_{0i})) + \lambda_3 \sum_{i=1}^{T} \mathbb{S}(\mathcal{R}_\mathcal{A}(X'), \mathcal{R}_\mathcal{A}(X_{\infty i}))$$




Break

Framework

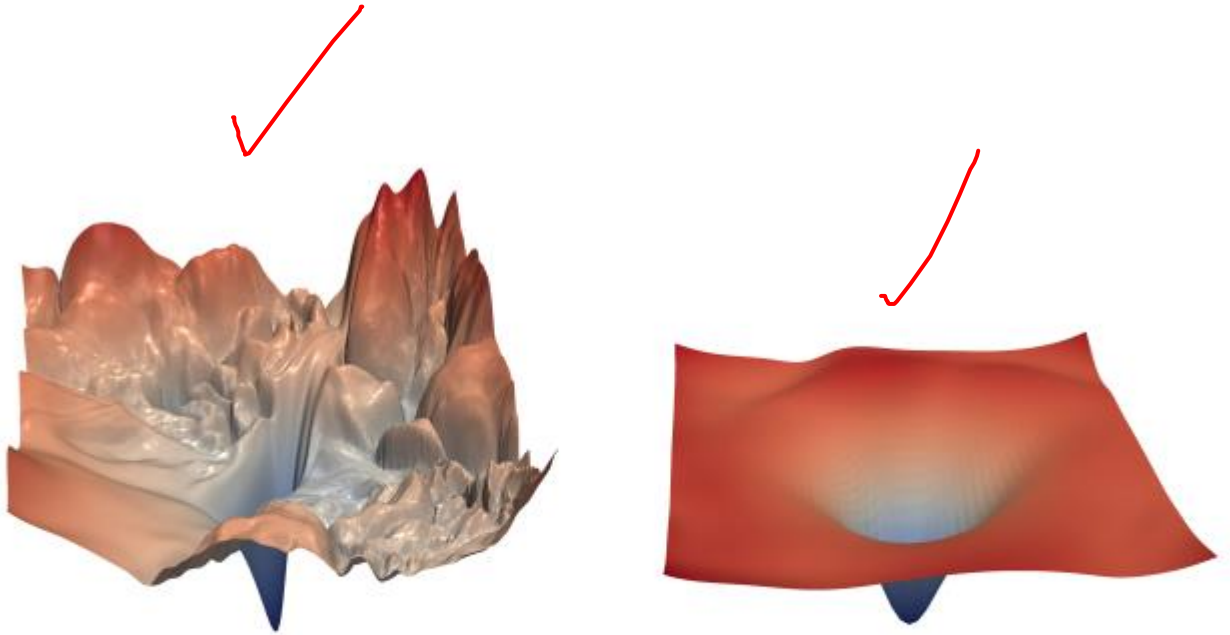
TensorFlow, Keras, PyTorch, MXNet, scikit-learn, XGBoost, LightGBM, CatBoost, GPy, etc.

JAX, PyTorch, and TF2

PyTorch, TensorFlow, and JAX


Break

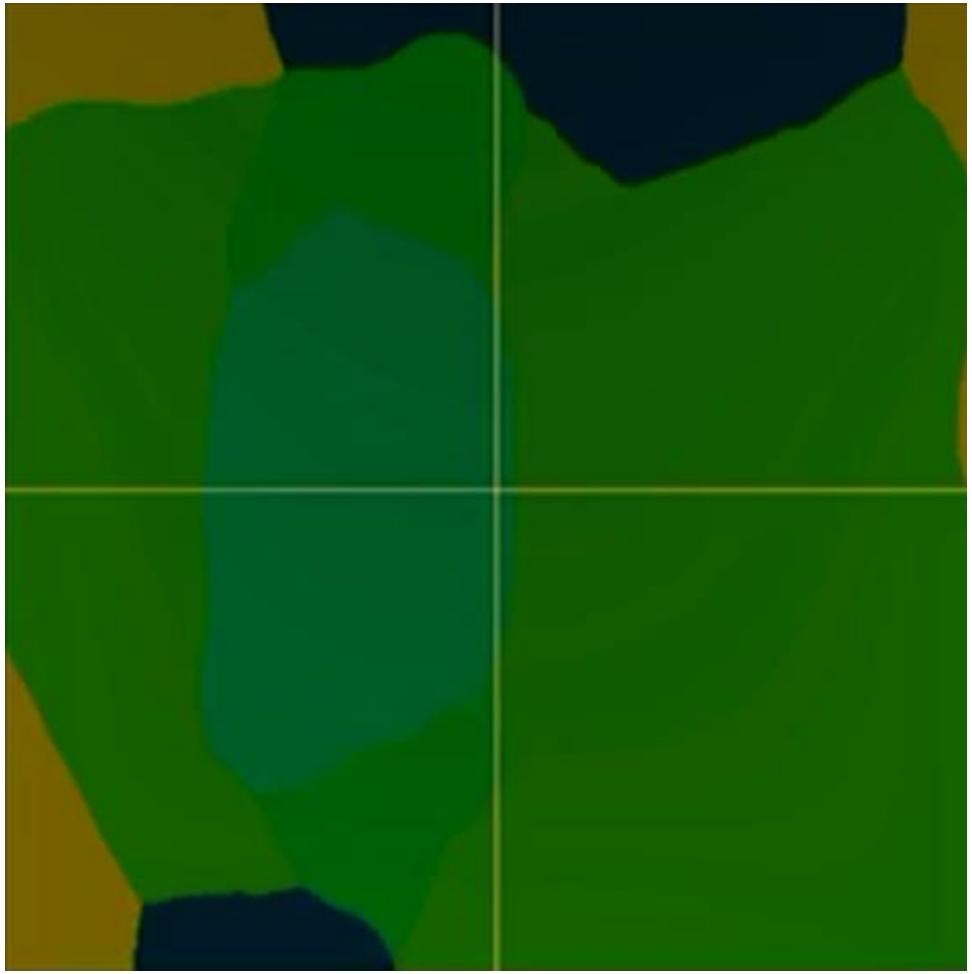
Loss landscape



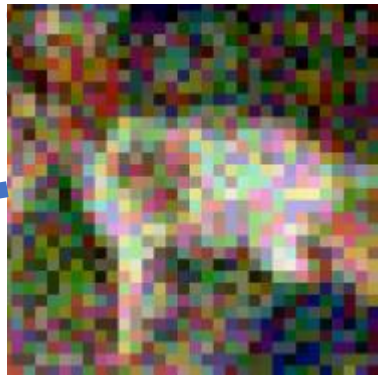
(a) without skip connections (b) with skip connections

Figure 1: The loss surfaces of ResNet-56 with/without skip connections. The proposed filter normalization scheme is used to enable comparisons of sharpness/flatness between the two figures.

Loss landscape

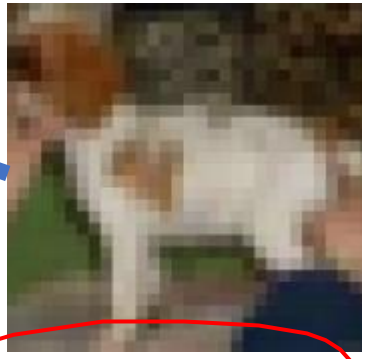


Loss landscape



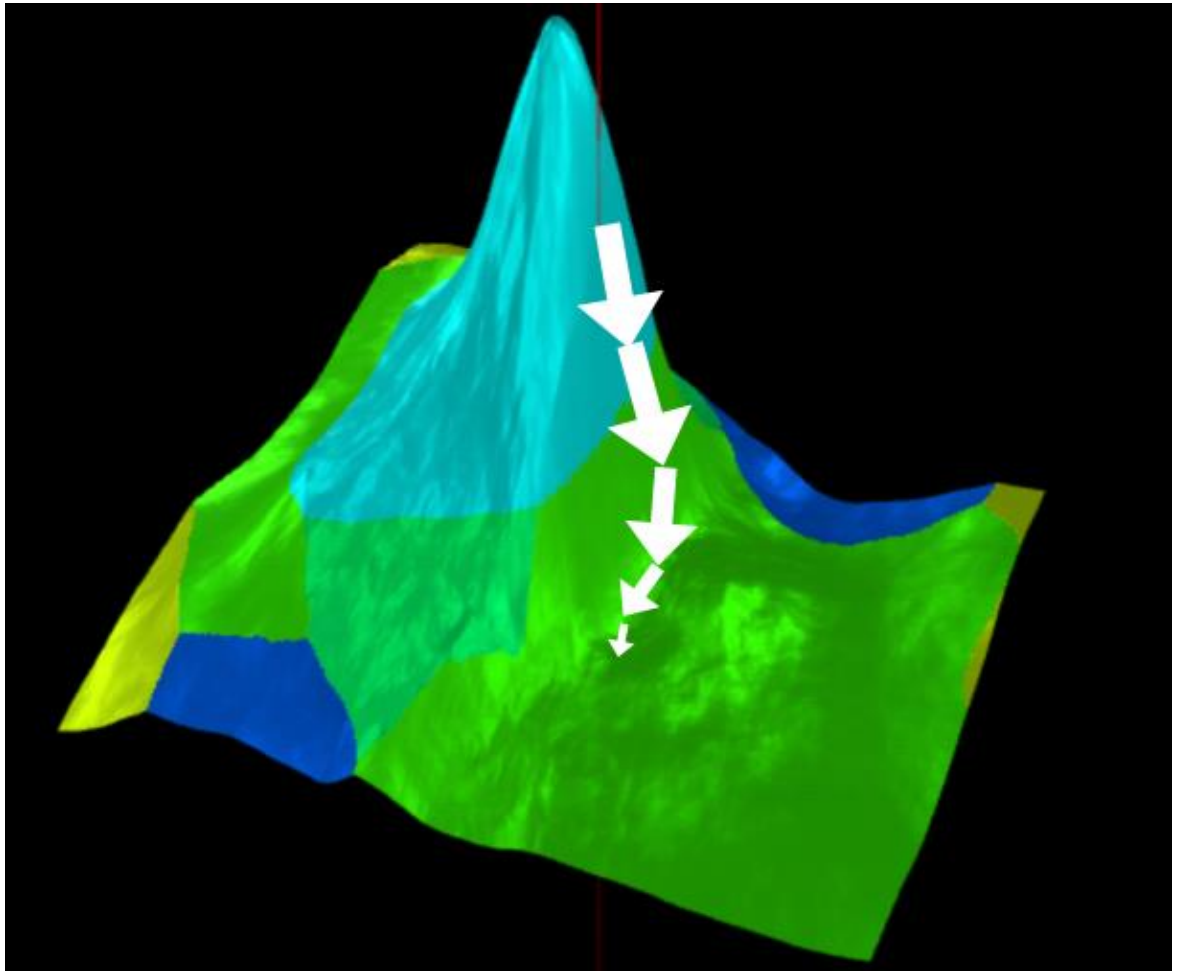
Truck

Dog

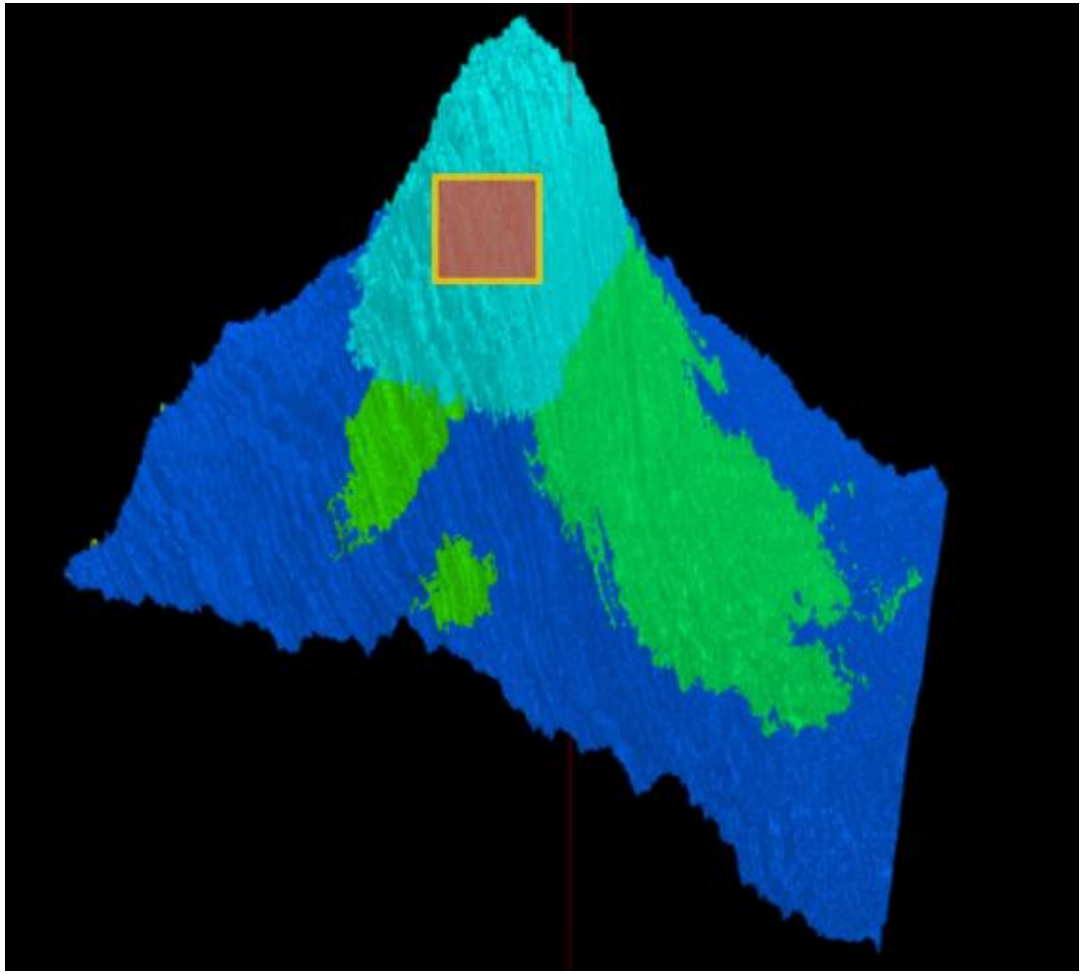


Airplane

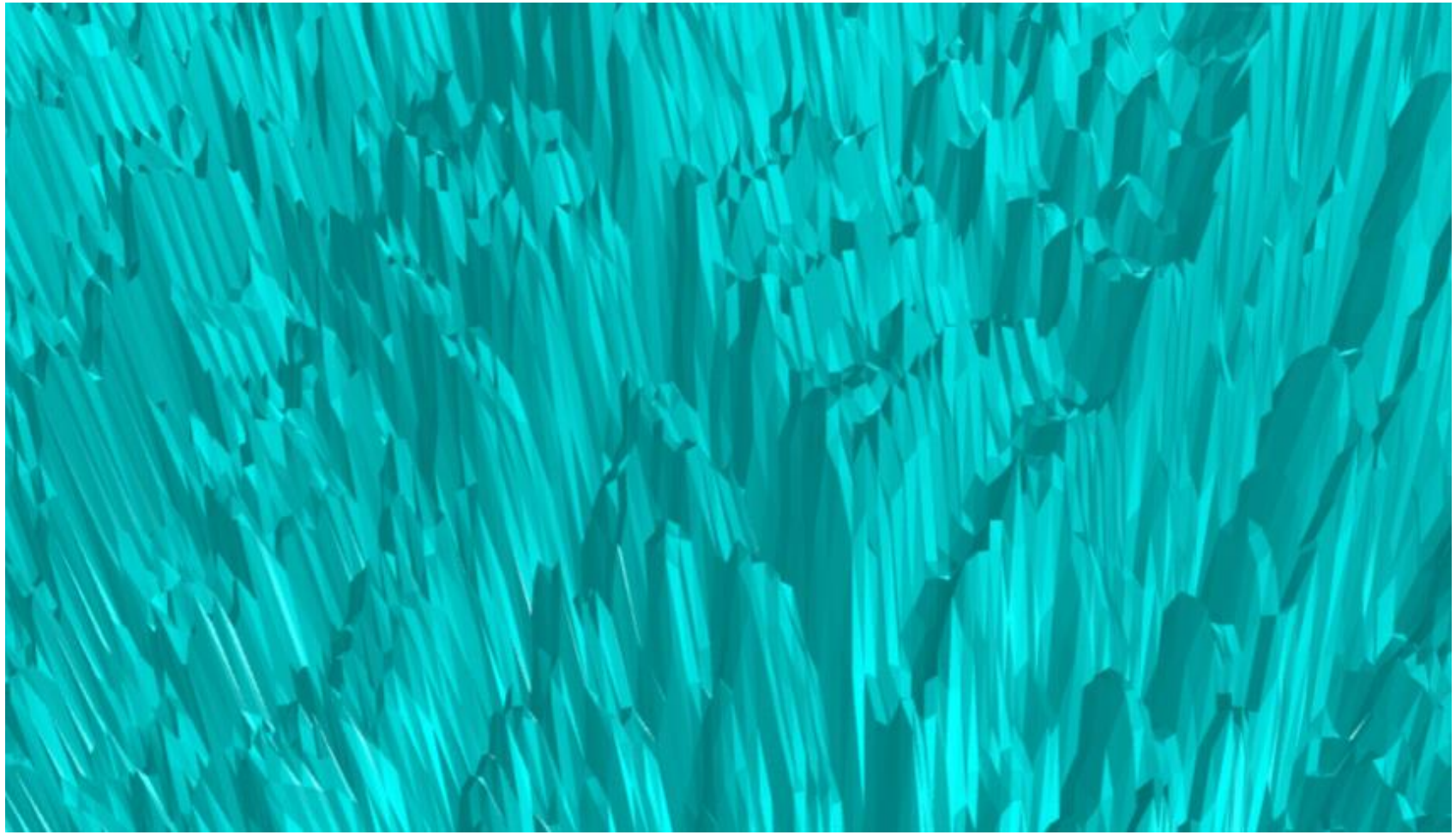
Loss landscape



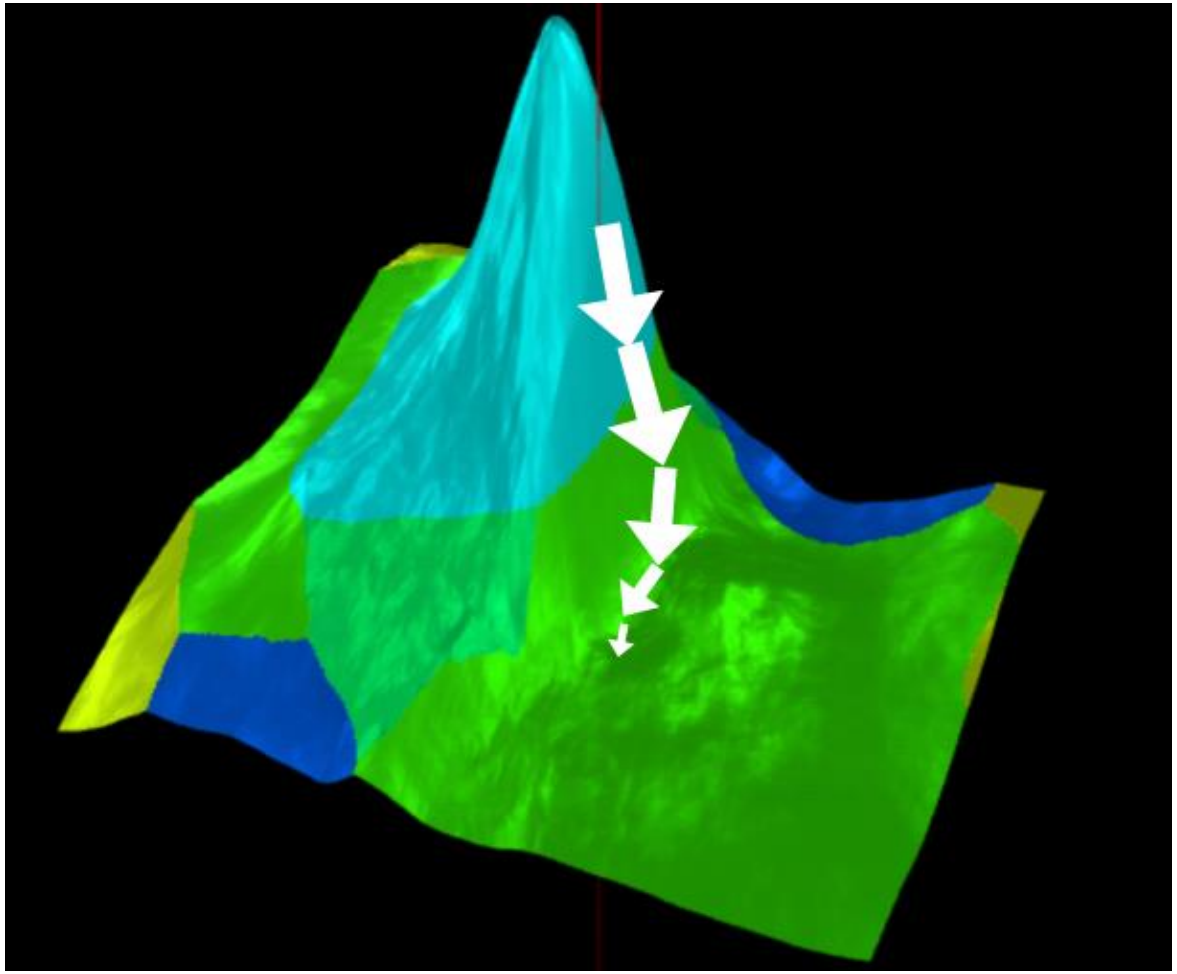
Loss landscape



Loss landscape



Loss landscape



Defense

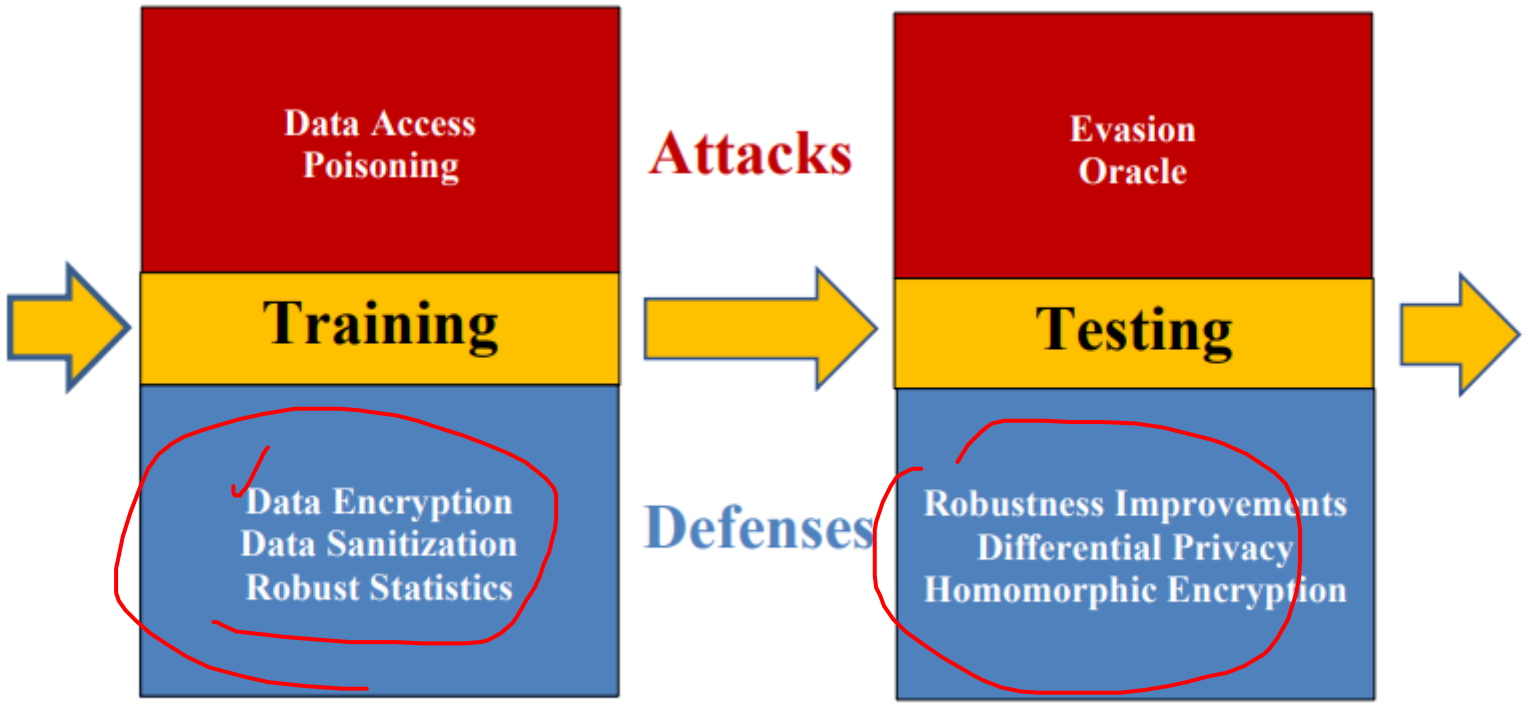


Figure 1. An illustration of example Attacks and Defenses in the Machine Learning Pipeline.

https://nvlpubs.nist.gov/nistpubs/ir/2019/NIST.IR.8269-draft.pdf


Break

Adversarial training methodologies & defenses

Advances in defense strategies

- Defense is hard!
  - *A theoretical model of the adversarial example crafting process is very **difficult** to construct.*
    - Non-linearity
    - non-convex
    - Complex optimization process
    - ...

- Most of the current defense strategies are
  - not adaptive to **all types** of adversarial attack
- Implementation of such defense strategies
  - may incur **performance overhead**

Advances in defense strategies

- **Defense**
  - Data manipulation
  - Architecture manipulation
  - Loss manipulation
  - Certified robustness
  - From interpretability and explainability
  - …

Empirical risk minimization

In standard ERM, we optimize the following objective:

$$\min_{\theta} \mathbb{E}_{(\mathbf{x},y)} \left[\ell_{cls} \left(f_\theta(\mathbf{x}), y \right) \right]$$

Empirical risk minimization & Adversarial training

In standard ERM, we optimize the following objective:

$$\min_{\theta} \mathbb{E}_{(\mathbf{x},y)} \left[\ell_{cls} \left(f_\theta(\mathbf{x}), y \right) \right]$$

For adversarial training, we need to optimize two things simultaneously.

- First, we generate the strongest minimal perturbation first.
- Second, we train our models to be robust against that.

Empirical risk minimization & Adversarial training

In standard ERM, we optimize the following objective:

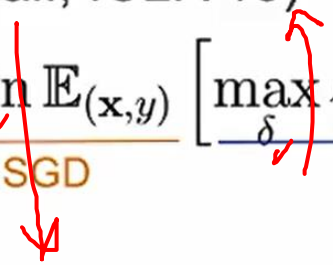
$$\min_{\theta} \mathbb{E}_{(\mathbf{x},y)} \left[\ell_{cls} \left(f_\theta(\mathbf{x}), y \right) \right]$$

For adversarial training, we need to optimize two things simultaneously.

- First, we generate the strongest minimal perturbation first.
- Second, we train our models to be robust against that.

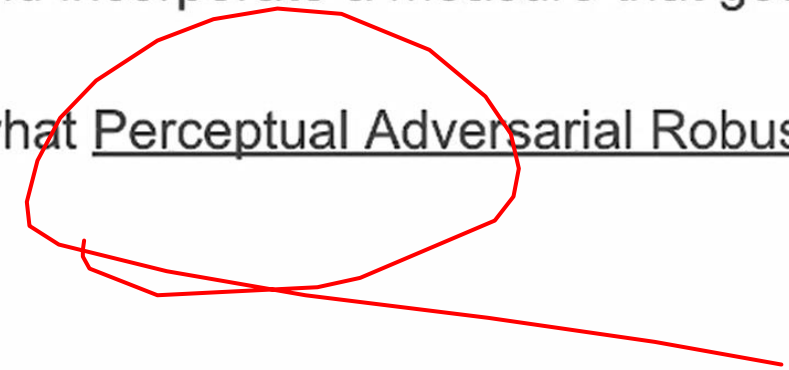
Mathematically (Madry et al., ICLR'18) -

$$\underbrace{\min_{\theta} \mathbb{E}_{(\mathbf{x},y)}}_{\text{SGD}} \left[\underbrace{\max_{\delta} \ell_{cls} \left(f_\theta(\mathbf{x}+\delta), y \right)}_{\text{PGD}} \right]$$



Adv. training

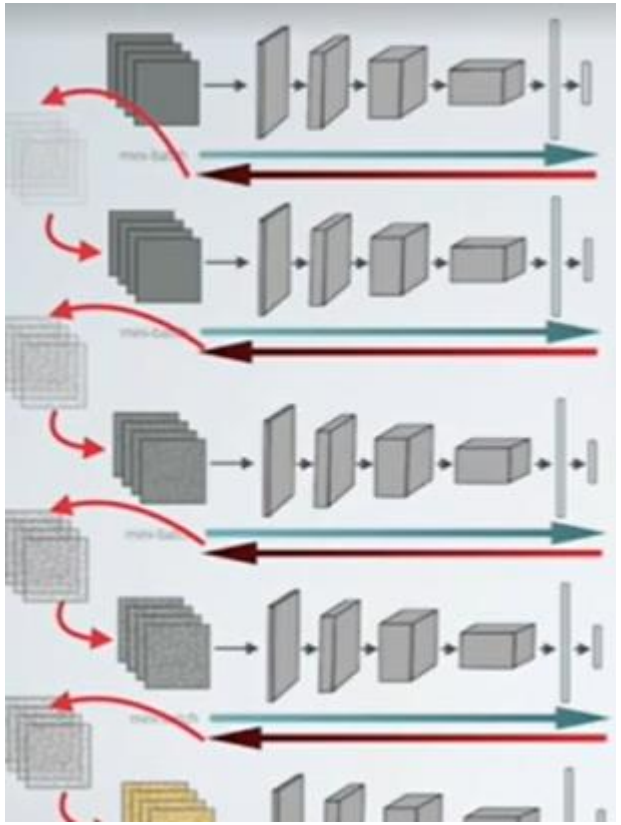
- Since human perception is hard to characterize precisely, this lack of transfer become inevitable.
- So, what if we could incorporate a measure that gets us closer to the human perception?
- This is precisely what Perceptual Adversarial Robustness (Laidlaw et al., ICLR'21) does.



Adv. training FREE

Adversarial Training for Free!

NeurIPS 2019



Algorithm 1 "Free" Adversarial Training (Free-m)

Require: Training samples X, perturbation bound ϵ, learning rate τ, hop steps m

Initialize θ
$\delta \leftarrow 0$
for epoch $= 1 \ldots N_{ep}/m$ **do**
  for minibatch $B \subset X$ **do**
    for i $= 1 \ldots m$ **do**
      Update θ with stochastic gradient descent
        $g_\theta \leftarrow \mathbb{E}_{(x,y)\in B}[\nabla_\theta\, l(x+\delta, y, \theta)]$
        $g_{adv} \leftarrow \nabla_x\, l(x+\delta, y, \theta)]$
        $\theta \leftarrow \theta - \tau g_\theta$
      Use gradients calculated for the minimization step to update δ
        $\delta \leftarrow \delta + \epsilon \cdot \text{sign}(g_{adv})$
        $\delta \leftarrow \text{clip}(\delta, -\epsilon, \epsilon)$
    end for
  end for
end for

Adv. training FAST

FAST IS BETTER THAN FREE:
REVISITING ADVERSARIAL TRAINING

ICLR 2020

Algorithm 3 FGSM adversarial training for T epochs, given some radius ϵ, N PGD steps, step size α, and a dataset of size M for a network f_θ

for $t = 1 \dots T$ **do**
  for $i = 1 \dots M$ **do**
    // Perform FGSM adversarial attack
    $\delta = \text{Uniform}(-\epsilon, \epsilon)$
    $\delta = \delta + \alpha \cdot \text{sign}(\nabla_\delta \ell(f_\theta(x_i + \delta), y_i))$
    $\delta = \max(\min(\delta, \epsilon), -\epsilon)$
    $\theta = \theta - \nabla_\theta \ell(f_\theta(x_i + \delta), y_i)$ *// Update model weights with some optimizer, e.g. SGD*
  end for
end for

Adv. training TRADE ✓

Theoretically Principled Trade-off between Robustness and Accuracy — ICLR 2019

Robust (classification) error

under the threat model of bounded ε perturbation

$$\mathcal{R}_{\text{rob}}(f) := \mathbb{E}_{(\boldsymbol{X},Y)\sim\mathcal{D}}\mathbf{1}\{\exists \boldsymbol{X}' \in \mathbb{B}(\boldsymbol{X},\epsilon) \text{ s.t. } f(\boldsymbol{X}')Y \leq 0\}$$

Natural (classification) error

standard measure of classifier performance

$$\mathcal{R}_{\text{nat}}(f) := \mathbb{E}_{(\boldsymbol{X},Y)\sim\mathcal{D}}\mathbf{1}\{f(\boldsymbol{X})Y \leq 0\}$$

$$\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathcal{X}$$

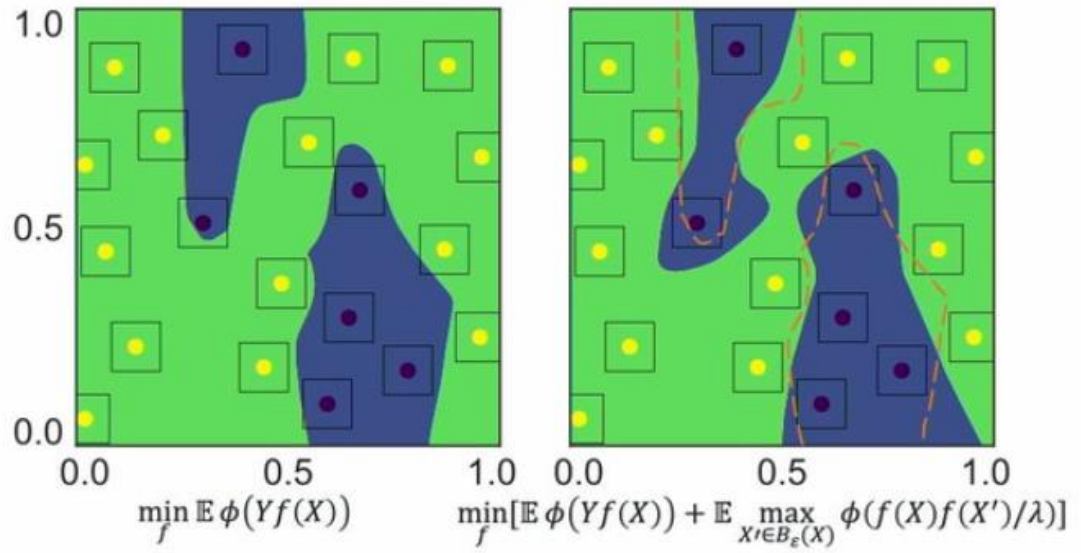
$$y_1, \ldots, y_n \in \{-1, +1\}$$

Boundary error

$$\mathcal{R}_{\text{bdy}}(f) := \mathbb{E}_{(\boldsymbol{X},Y)\sim\mathcal{D}}\mathbf{1}\{\boldsymbol{X} \in \mathbb{B}(\text{DB}(f),\epsilon), f(\boldsymbol{X})Y > 0\}$$

$$\mathcal{R}_{\text{rob}}(f) = \mathcal{R}_{\text{nat}}(f) + \mathcal{R}_{\text{bdy}}(f).$$

$$\min_f \mathbb{E}\Big\{ \underbrace{\phi(f(\boldsymbol{X})Y)}_{\text{for accuracy}} + \underbrace{\max_{\boldsymbol{X}'\in\mathbb{B}(\boldsymbol{X},\epsilon)} \phi(f(\boldsymbol{X})f(\boldsymbol{X}')/\lambda)}_{\text{regularization for robustness}} \Big\}.$$




Break

Promising recipes

Studying optimizer susceptibility

From the previous plot, optimizers that may easily fall prey to the attacks:

- Adam
- RMSProp

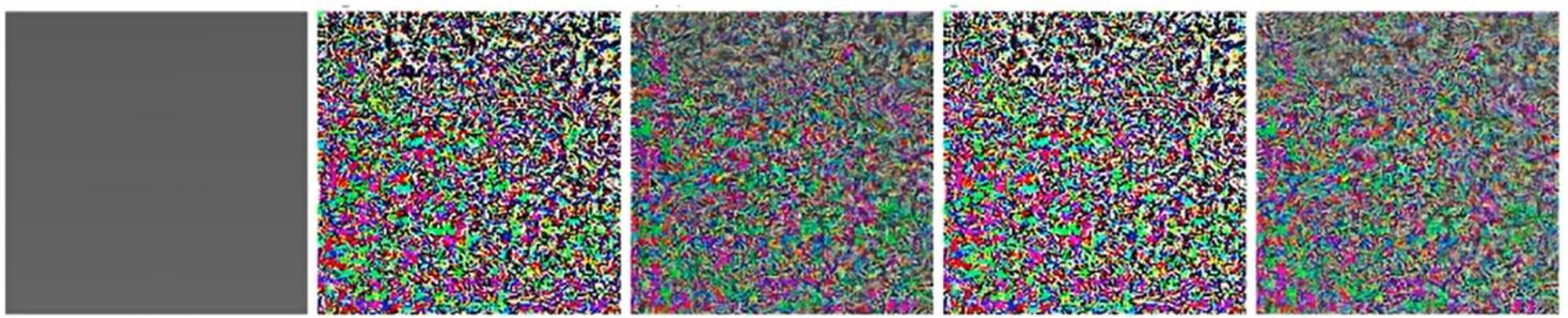
Optimizers that may not easily fall prey to the attacks:

- SGD
- Adagrad
- FTRL

This is characterized by the non-convexity of the optimization problem.

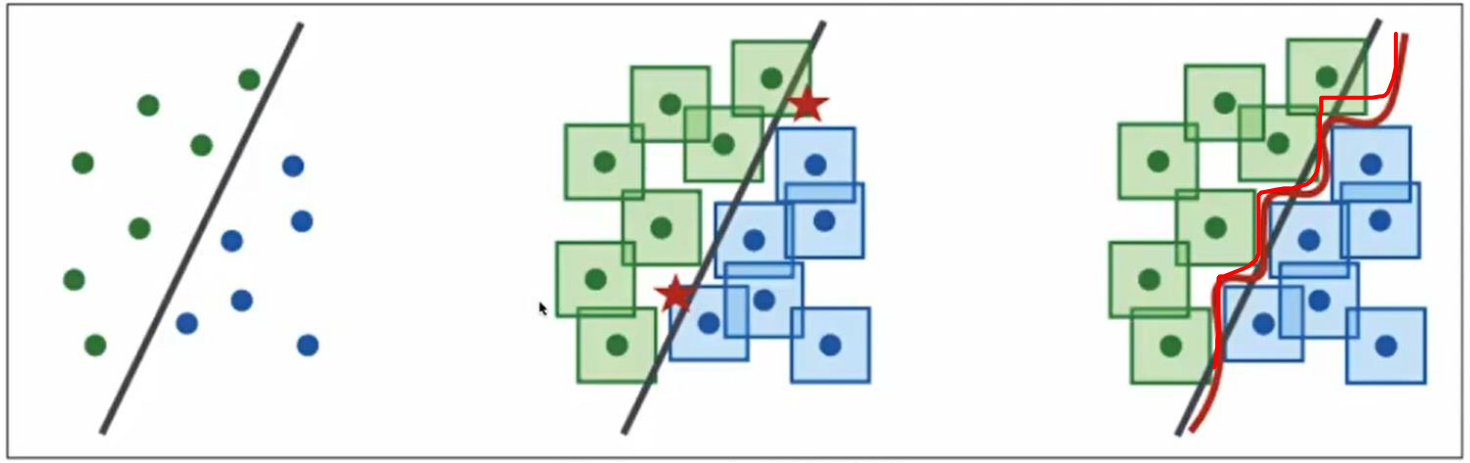
Studying optimizer susceptibility

Different deltas (δ) as learned by the optimizers



Model capacity is crucial

Adversarial examples change the decision boundary to a more complicated one (Madry et al., ICLR'18).

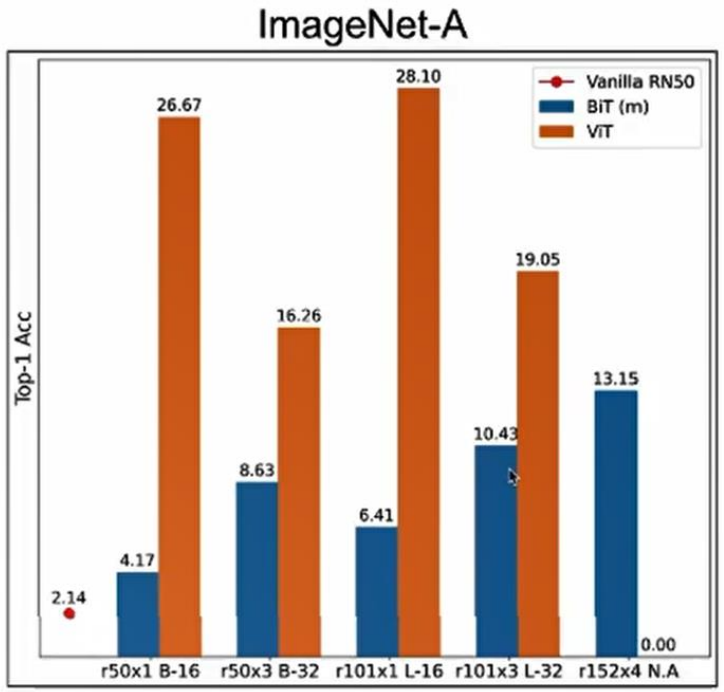


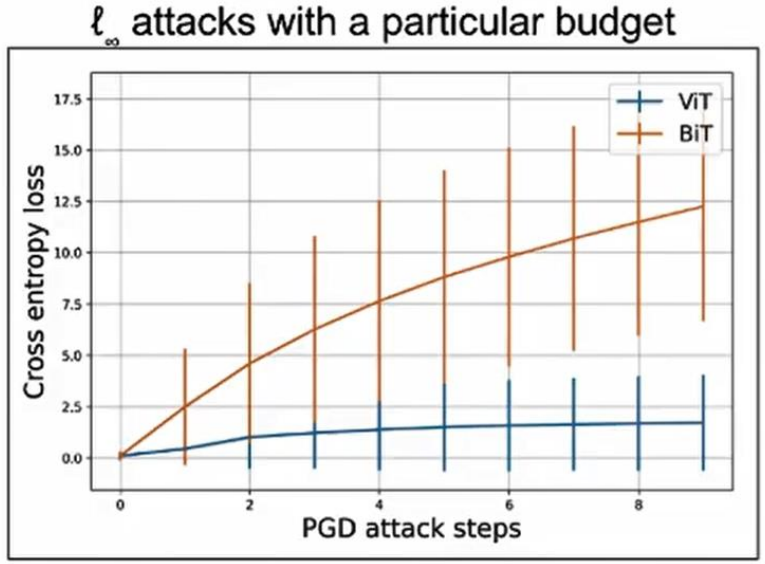
Madry et al., ICLR'18

ℓ_∞-balls
adv. examples

Self-attention provides improved robustness

Continuation of the previous discussion -

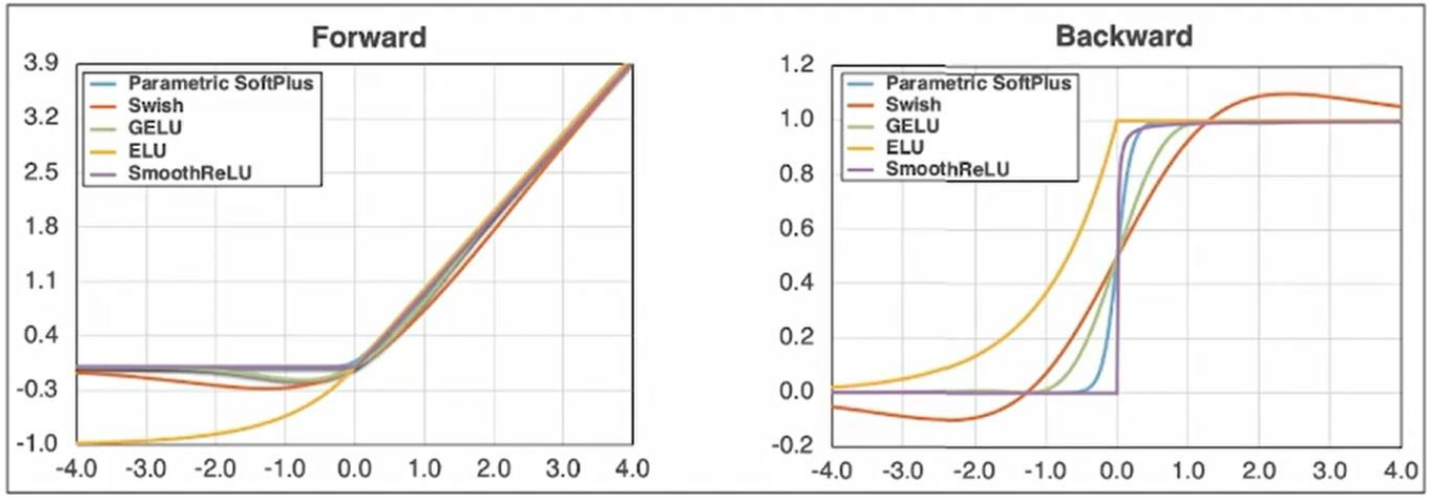




Paul et al., arXiv, 2021

Smooth adv. training

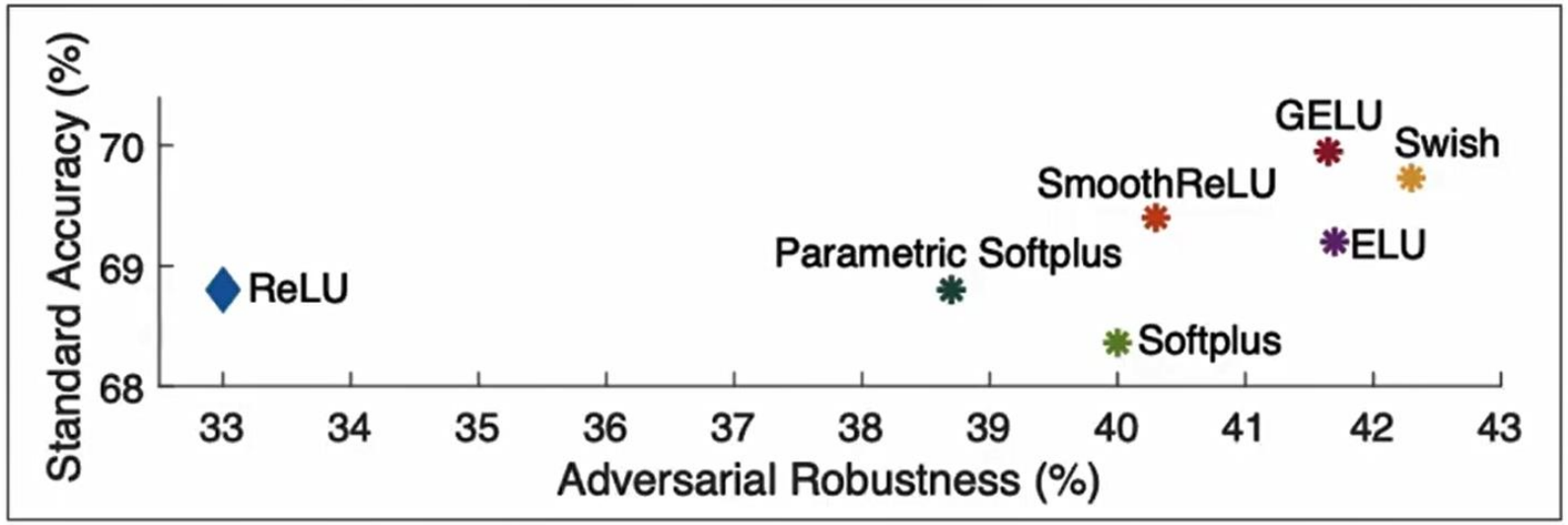
- Using ReLU during adv. training is particularly worse off because of its non-smooth nature.
- Smoother activation functions (Swish, SoftPlus, etc.) result into better informed gradients because of their smoothness.



Xie et al., arXiv, 2020

Smooth adv. training

The use of smoother activation functions leads to improved performance without accuracy loss.



Xie et al., arXiv, 2020

Conclusion

Being aware is helpful

- For **model developers**, adversarial examples can be used for robustness evaluation and model improvement.
- For **business stakeholders**, lacking adversarial robustness in your AI model could bring unexpected negative impacts.
- For **end users**, gaining awareness of adversarial robustness for the AI service your are using is crucial.

Takeaways

- Emerging trend of transition from stand ML to (adversarially) robust ML.
- Inspect potential errors and risks of your ML algorithm/system using adversarial ML tools.
- Improve model robustness via adversarial machine learning.
- Anticipate adversarial inputs and data shifts.
- Novel applications driven by adversarial robustness.
- Practice adversarial robustness in ML lifecycle.

Online resource for advevrsarial robustness

- https://adversarial-ml-tutorial.org/
- https://nips.cc/Conferences/2018/ScheduleMultitrack?event=10978
- https://sites.google.com/umich.edu/cvpr-2020-zoo
- https://advmlincv.github.io/cvpr21-tutorial/
- https://eccv20-adv-workshop.github.io/
- https://sites.google.com/view/par-2021
- http://www.cs.umd.edu/class/fall2020/cmsc828W/

On evaluating adversarial robustness

On Evaluating Adversarial Robustness

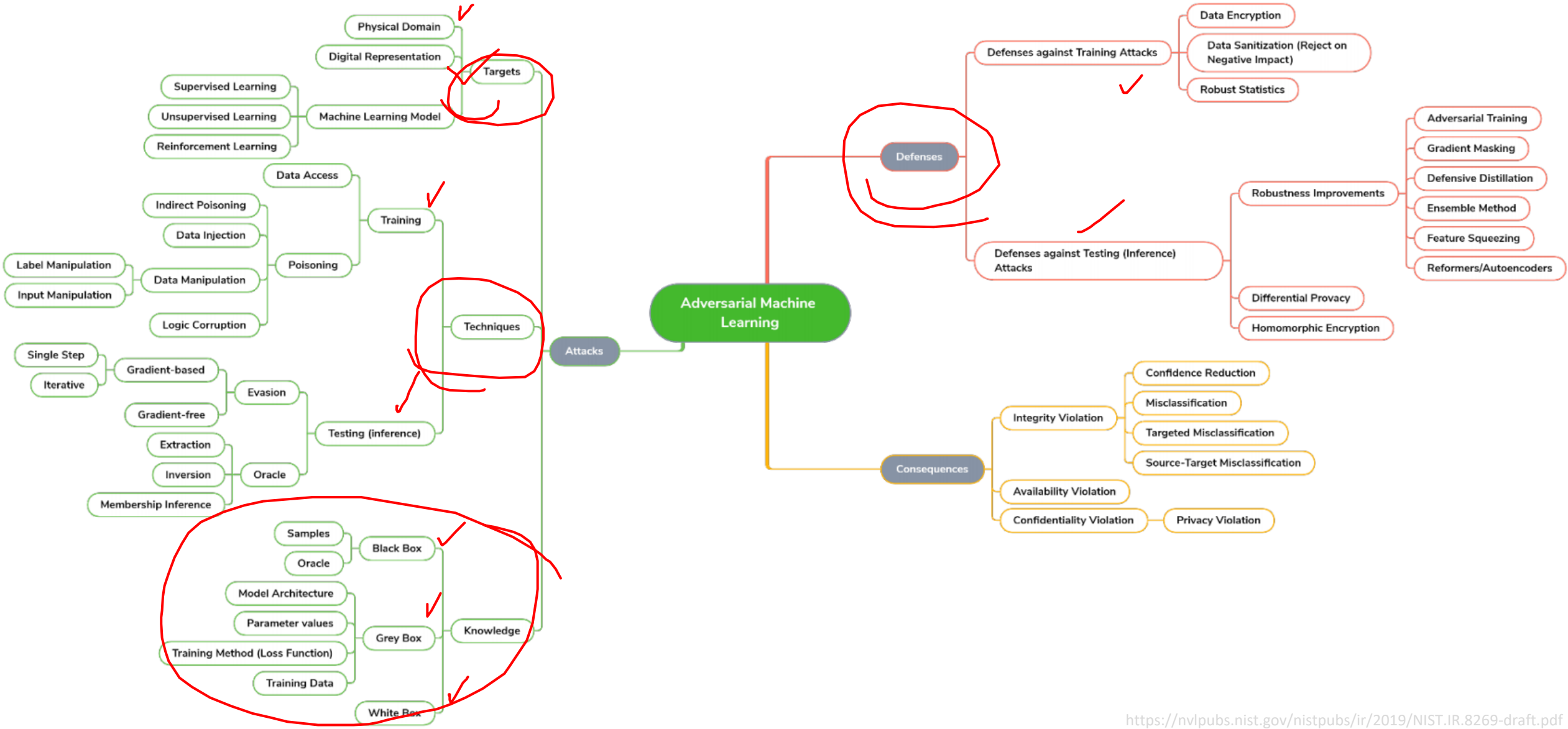
Nicholas Carlini[1], Anish Athalye[2], Nicolas Papernot[1], Wieland Brendel[3], Jonas Rauber[3], Dimitris Tsipras[2], Ian Goodfellow[1], Aleksander Mądry[2], Alexey Kurakin[1]*

[1] Google Brain [2] MIT [3] University of Tübingen

* List of authors is dynamic and subject to change. Authors are ordered according to the amount of their contribution to the text of the paper.

https://github.com/evaluating-adversarial-robustness/adv-eval-paper

Taxonomy of Attacks, Defenses, and Consequences in Adversarial Machine Learning



https://nvlpubs.nist.gov/nistpubs/ir/2019/NIST.IR.8269-draft.pdf

Adversarial Machine Learning community

https://forms.gle/ANbQLsrpzSFoVtmG7

References

- Practical Adversarial Robustness in Deep Learning: Problems and Solutions (CVPR 21 tutorial)
- CMSC 828W: Foundations of Deep Learning (by Dr. Soheil Feizi - 2020)
- Research papers which each related reference depicted in the bottom corner of the slide

Mohammad Khalooei

Khalooei [at] aut.ac.ir

https://khalooei.ir

https://ce.aut.ac.ir/~khalooei