CAP 5516
Medical Image Computing
(Spring 2022)

Dr. Chen Chen
Center for Research in Computer Vision (CRCV)
University of Central Florida
Office: HEC 221
Address: 4328 Scorpius St., Orlando, FL 32816-2365
Email: chen.chen@crcv.ucf.edu
Web: https://www.crcv.ucf.edu/chenchen/

Lecture 12
Adversarial Robustness in Deep Learning

Many slides are adapted from existing teaching or tutorial slides from Hung-yi Lee, Bo Li, and many others

Machine Learning: The Success Story

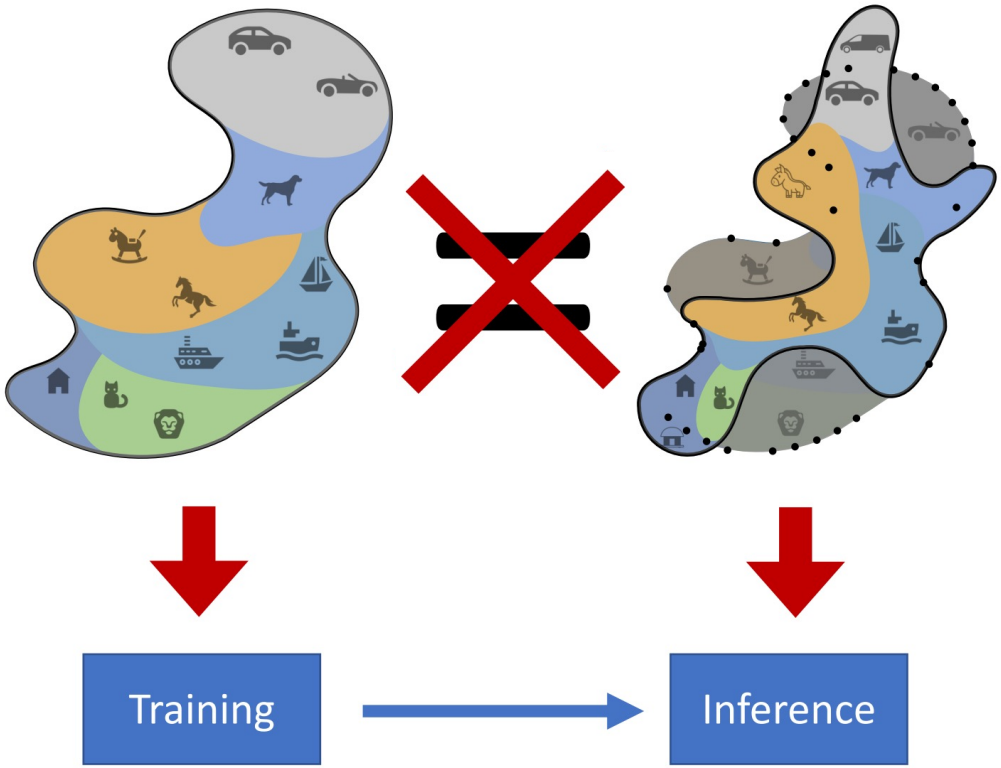
Image classification

Reinforcement Learning

Input sentence:	Translation (PBMT):	Translation (GNMT):	Translation (human):
李克強此行將啟動中加總理年度對話機制，與加拿大總理杜魯多舉行兩國總理首次年度對話。	Li Keqiang premier added this line to start the annual dialogue mechanism with the Canadian Prime Minister Trudeau two prime ministers held its first annual session.	Li Keqiang will start the annual dialogue mechanism with Prime Minister Trudeau of Canada and hold the first annual dialogue between the two premiers.	Li Keqiang will initiate the annual dialogue mechanism between premiers of China and Canada during this visit, and hold the first annual dialogue with Premier Trudeau of Canada.

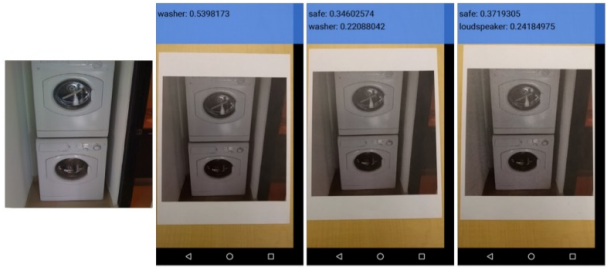
Machine translation

Is ML truly ready for real-world deployment?



ML Predictions Are (Mostly) Accurate but Brittle

malicious attacks on ML models



[Kurakin Goodfellow Bengio 2017]

[Athalye Engstrom Ilyas Kwok 2017]

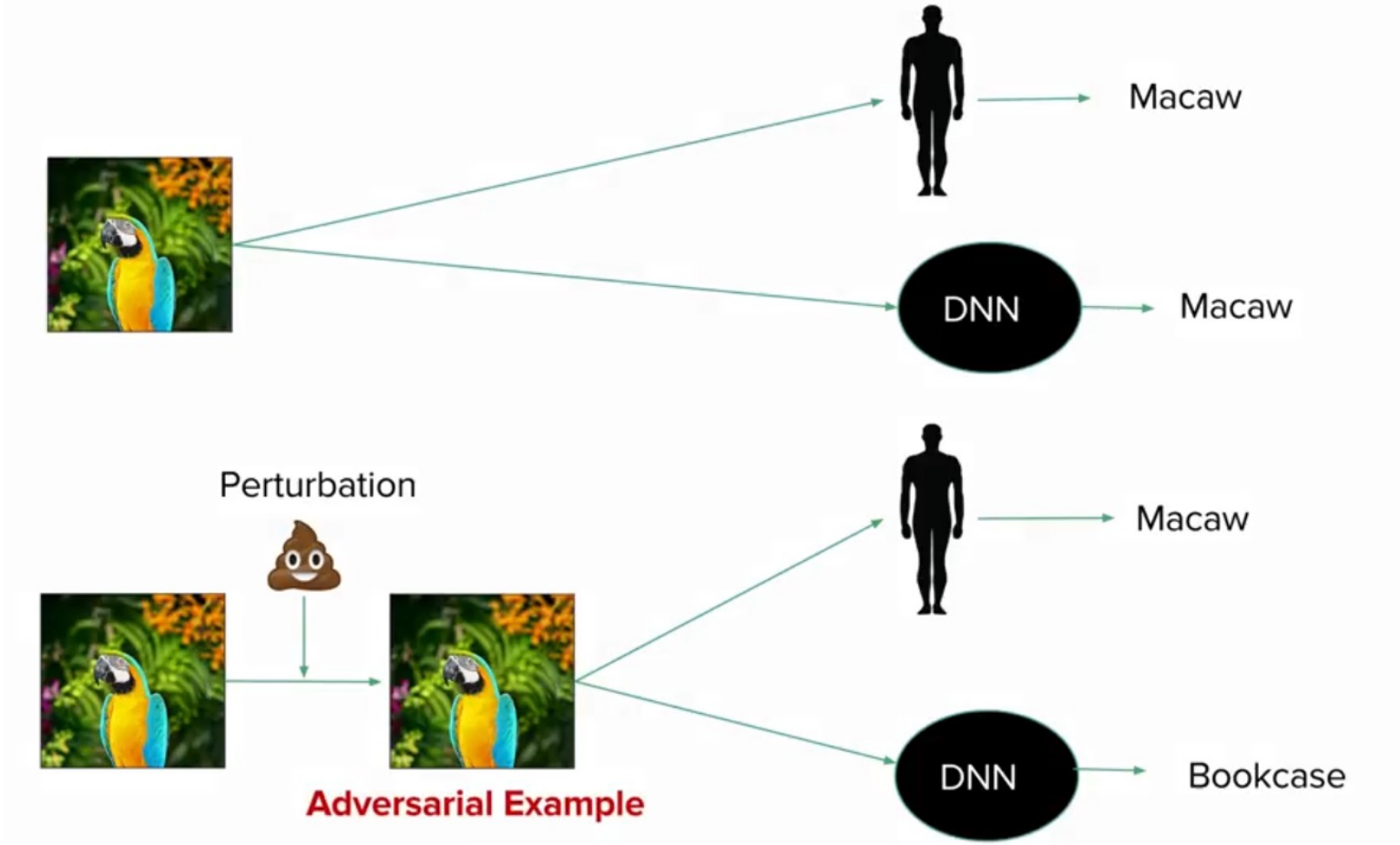
[Sharif Bhagavatula Bauer Reiter 2016]

[Eykholt Evtimov Fernandes Li Rahmati Xiao Prakash Kohno Song 2017]

ML Predictions Are (Mostly) Accurate but Brittle

- Adversarial attack

Adversarial machine learning, a technique that attempts to fool models with deceptive data, is a growing threat in the AI and machine learning research community.

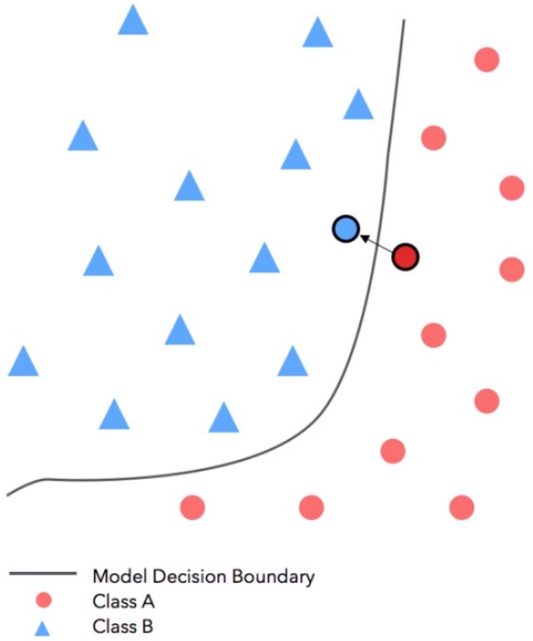


Adversarial Examples

These are inputs crafted by adding small perturbations to cause the classifier to misclassify.

We have image X and a classifier F that produces a label

$$F(X) = Y$$

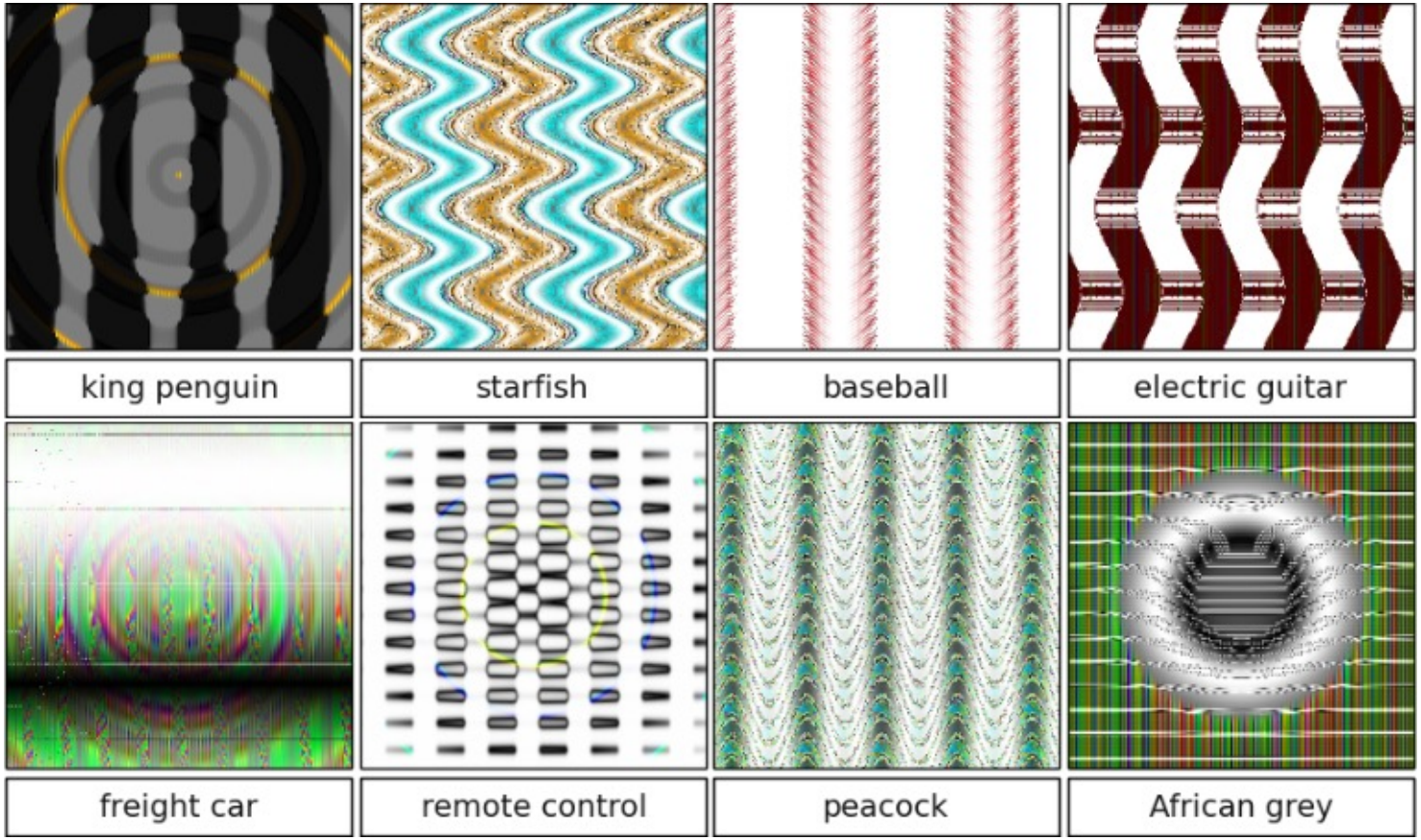


So then we add a perturbation **δX** to misclassify

$$F(X + \delta X) = Y^*$$

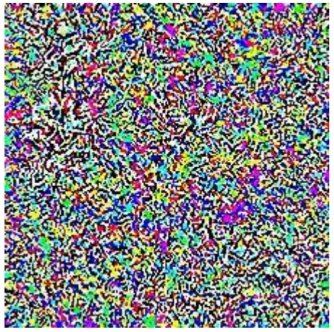
Ilyas, Andrew, et al. "Adversarial examples are not bugs, they are features." *Advances in neural information processing systems* 32 (2019).

Adversarial examples



Adversarial examples

- **Adversarial Sample / Adversarial Example**
  - Examples that are **similar** to examples in the true distribution, but that **fool** a classifier.
  - i.e., **modified** versions of a clean image that are **intentionally perturbed** (by adding noise) to confuse a machine learning system.



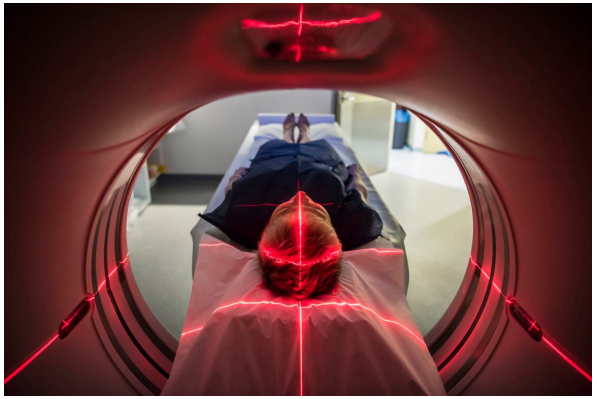
Classifier:
84% for sure it is a **Fat Panda !**

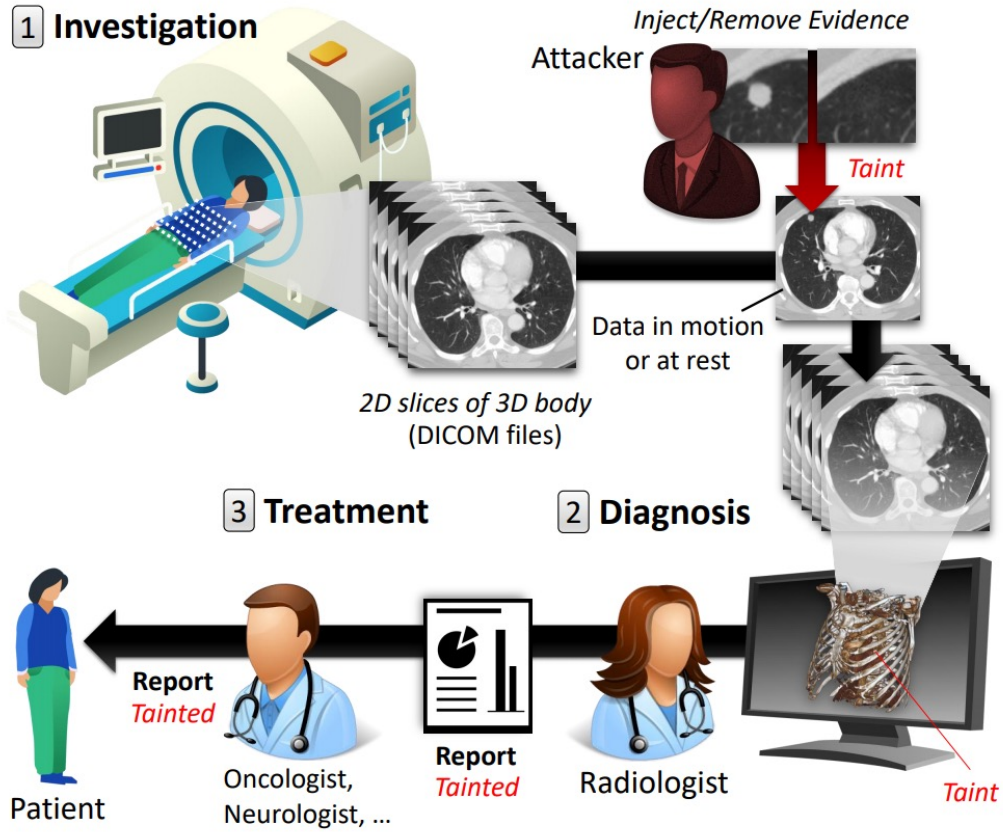
Adversarial Perturbation (Noise)

67%: Maybe Capuchin ?

Physical world adversarial attack

and the ML in your self-driving car thinks it's





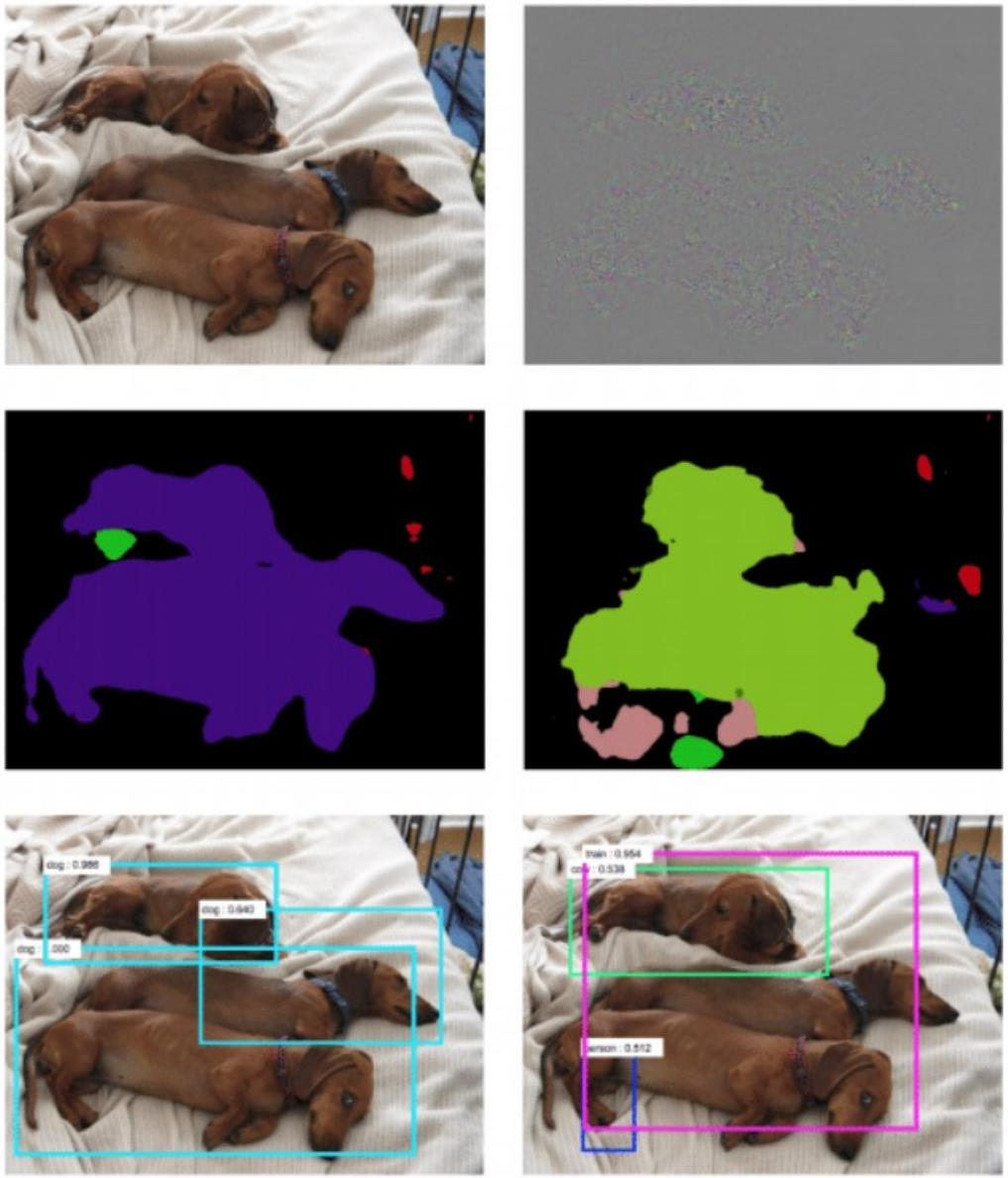
The Attack

Attackers can alter 3D medical scans to remove existing, or injecting non-existing, medical conditions.

CT-GAN: Malicious Tampering of 3D Medical Imagery using Deep Learning

https://www.youtube.com/watch?v=_mkRAArj-x0

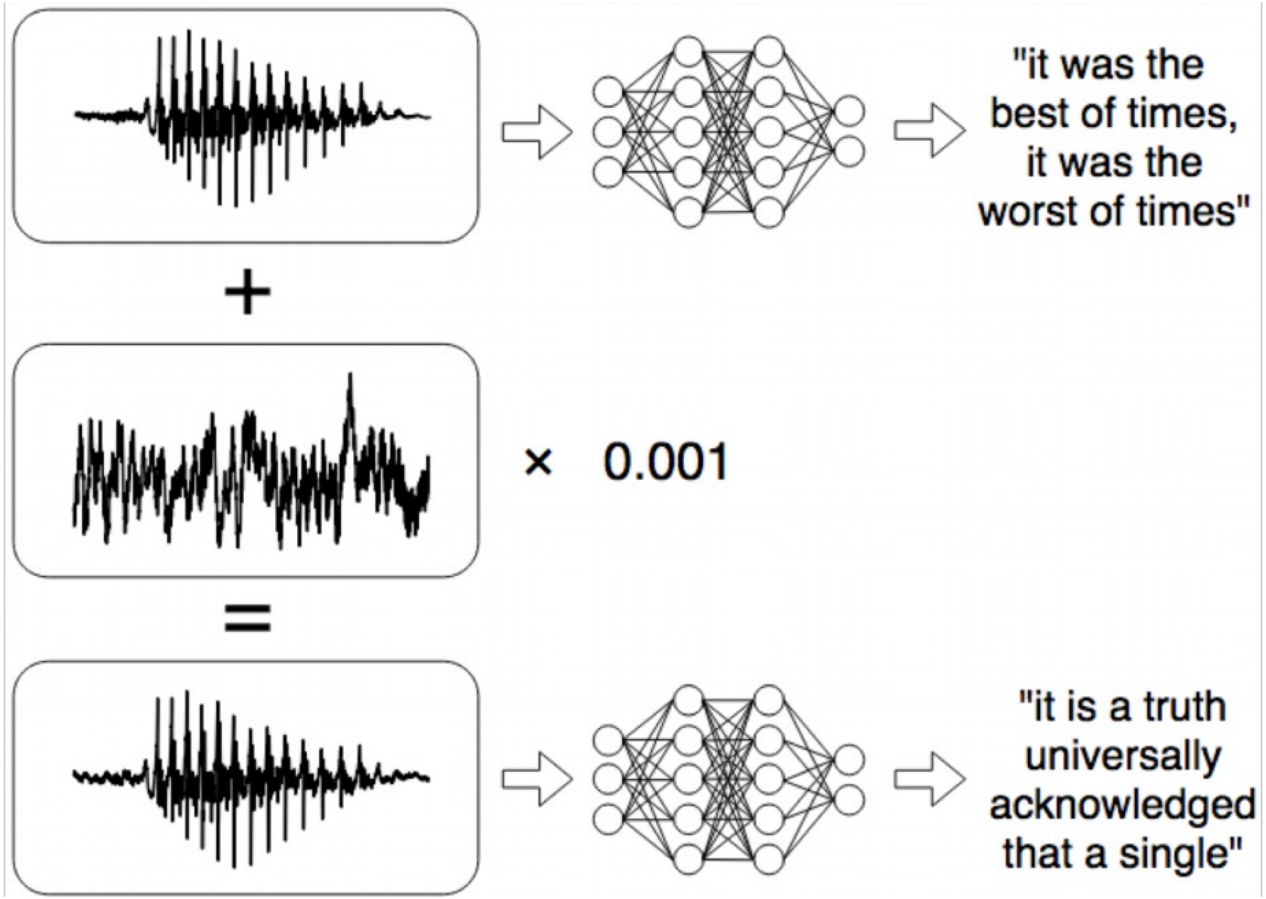
Not just image classification task



Adversarial examples for semantic segmentation and object detection

Xie, Cihang, et al. "Adversarial examples for semantic segmentation and object detection." Proceedings of the IEEE international conference on computer vision. 2017.

Not just image



Speech, NLP, etc.

Adversarial Attack

Example of Attack

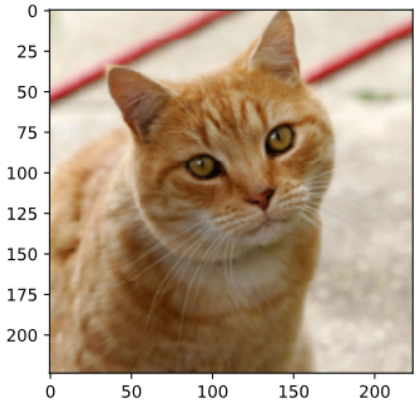
Non-targeted

Anything other than “Cat”

Targeted

Misclassified as a specific class (e.g., “Star Fish”)

Benign Image



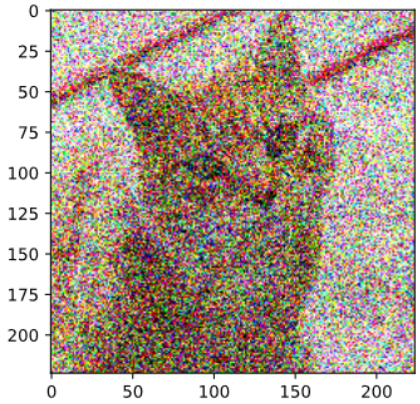
Network (Image Classifier)

Something Else

~~Tiger Cat~~

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \end{bmatrix} + \begin{bmatrix} \Delta x_1 \\ \Delta x_2 \\ \Delta x_3 \\ \vdots \end{bmatrix}$$

small



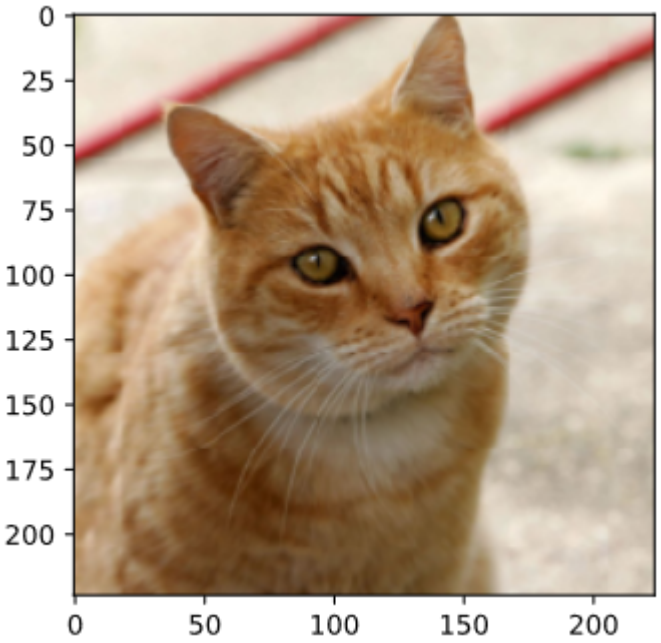
Attacked Image

Example of Attack

Network = ResNet-50

The target is “Star Fish”

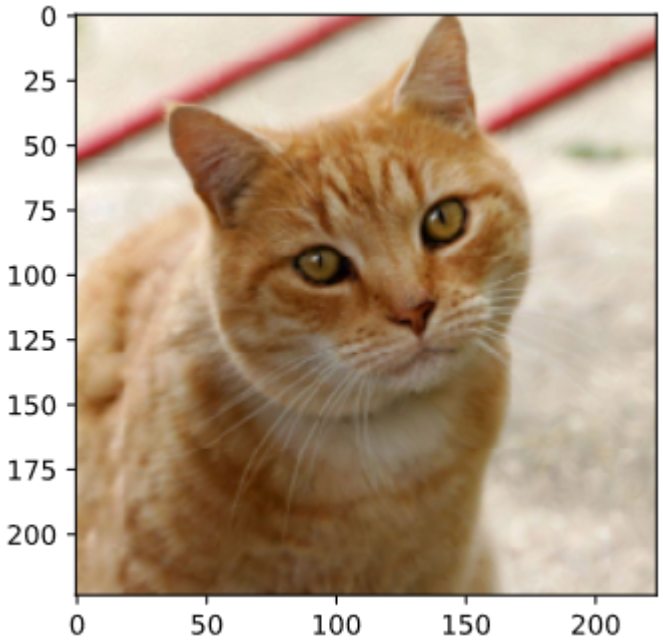
Benign Image



Tiger Cat

0.64

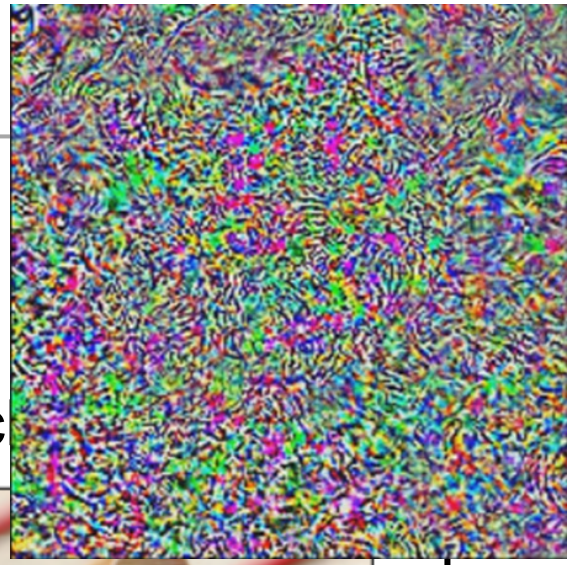
Attacked Image



Star Fish

1.00

Example of Attack

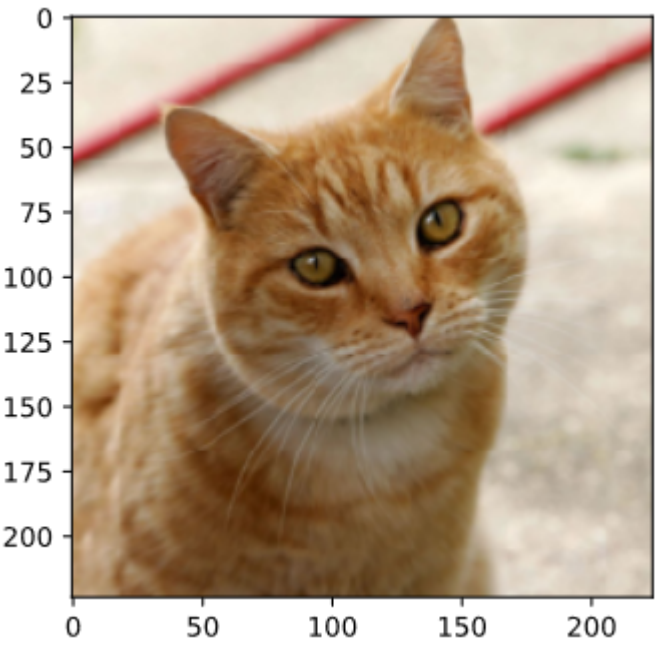


=

50

x

Benign Image

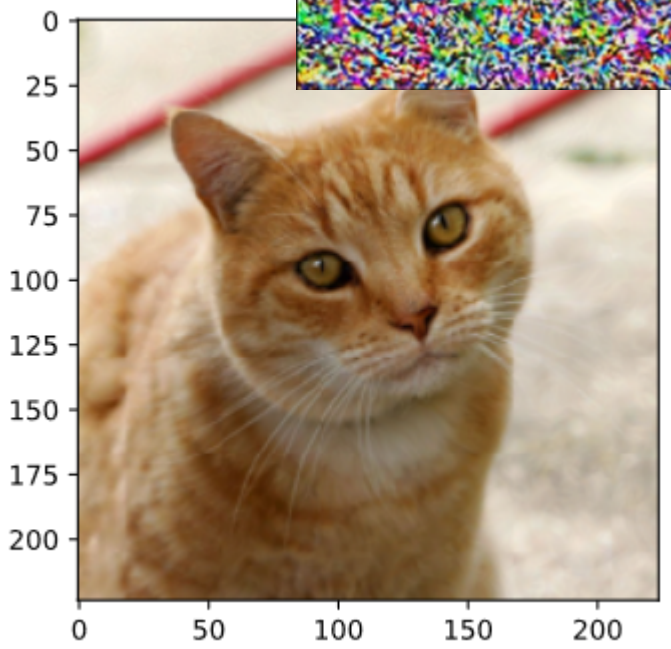


Tiger Cat

0.64

-

Attac



Star Fish

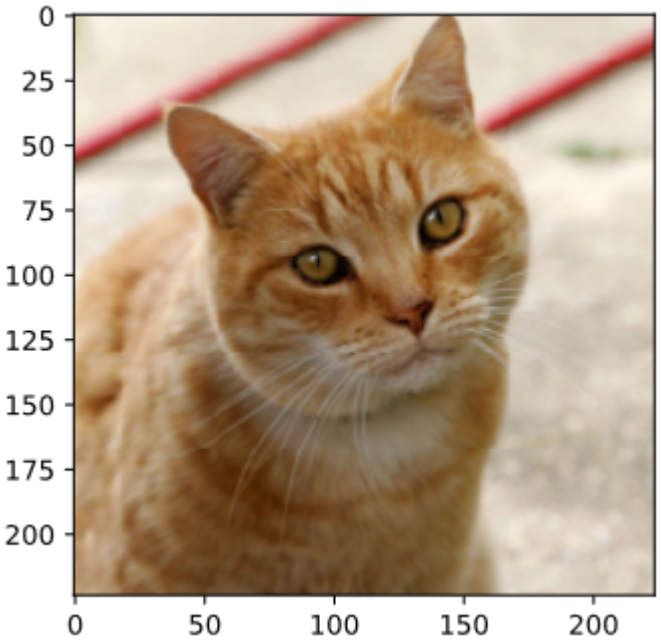
1.00

Example of Attack

Network = ResNet-50

The target is "Keyboard"

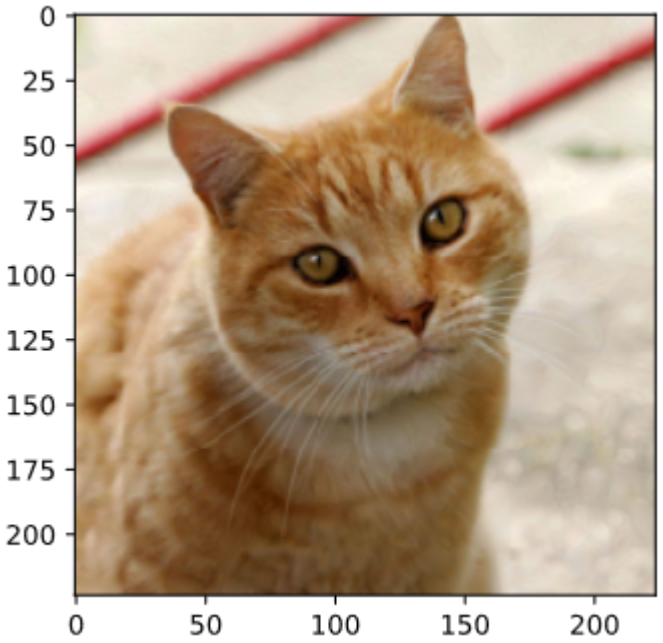
Benign Image



Tiger Cat

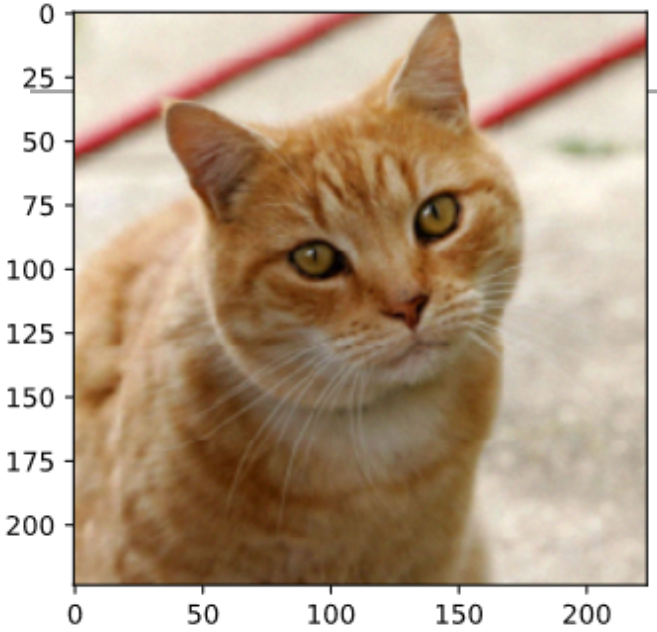
0.64

Attacked Image

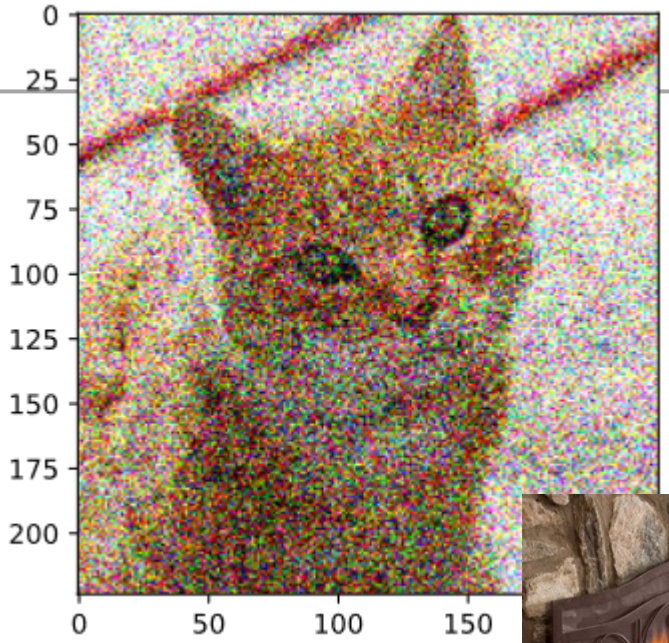


Keyboard

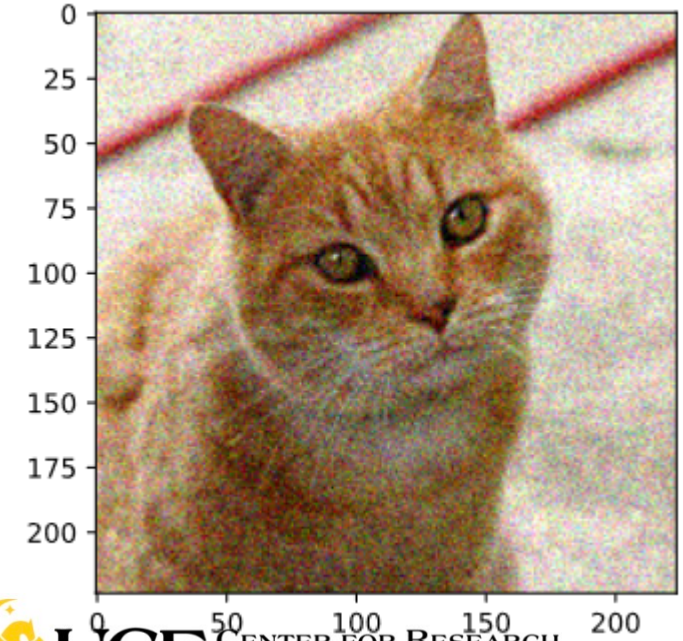
0.98



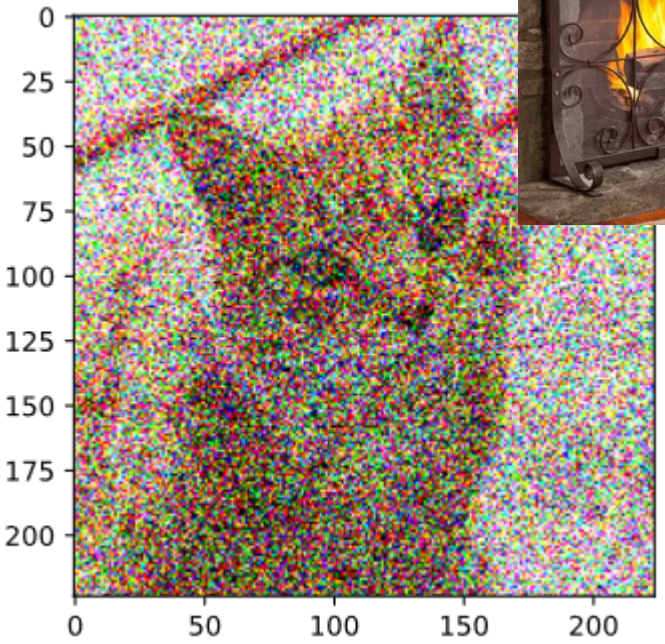
tiger cat



Persian cat

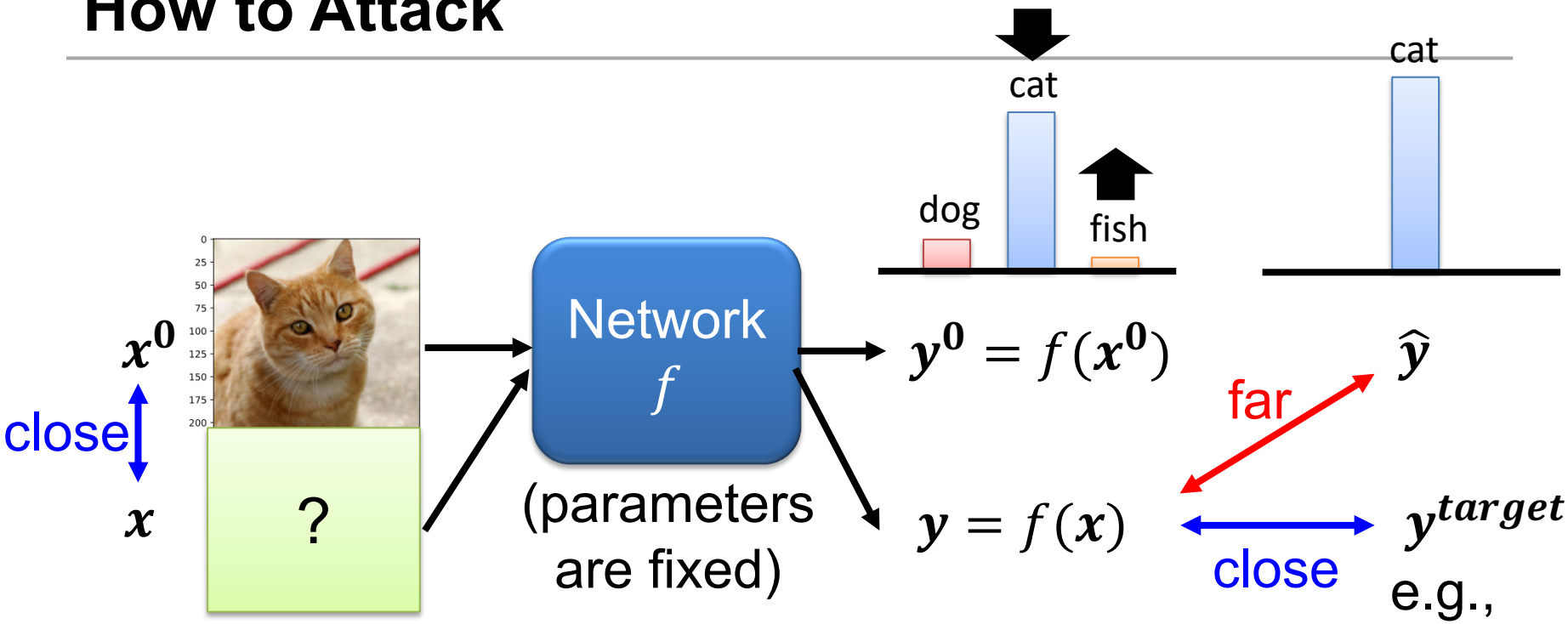


tabby cat



fire screen

How to Attack



Non-targeted $x^* = arg \min_{d(x^0,x)\leq\varepsilon} L(x)$

not perceived by humans

$L(x) = -e(y, \hat{y})$

Targeted $L(x) = -e(y, \hat{y}) + e(y, y^{target})$

Non-perceivable

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \end{bmatrix} - \begin{bmatrix} x_1^0 \\ x_2^0 \\ x_3^0 \\ \vdots \end{bmatrix} = \begin{bmatrix} \Delta x_1 \\ \Delta x_2 \\ \Delta x_3 \\ \vdots \end{bmatrix}$$

$\boldsymbol{x}$ $\quad$ $\boldsymbol{x^0}$ $\quad$ $\boldsymbol{\Delta x}$

$$d(\boldsymbol{x^0}, \boldsymbol{x}) \leq \varepsilon$$

Need to consider human perception

- L2-norm

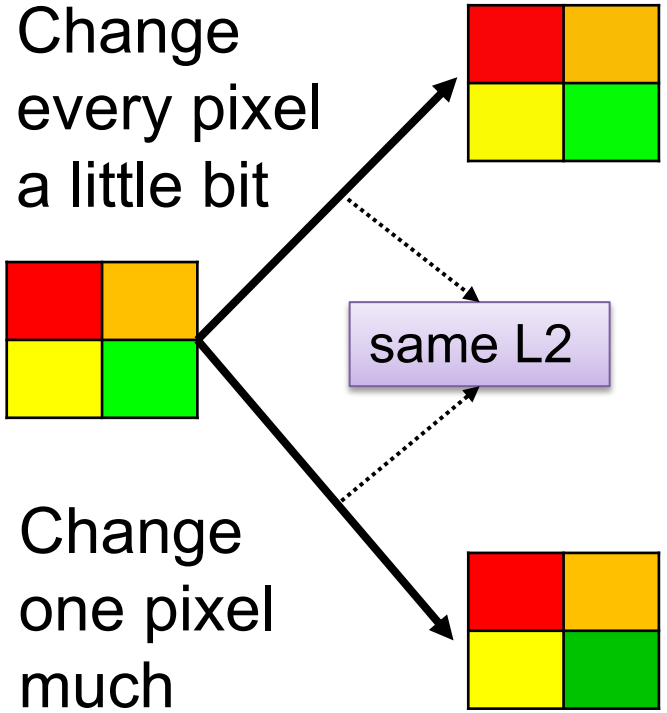
$$d(\boldsymbol{x^0}, \boldsymbol{x}) = \|\Delta \boldsymbol{x}\|_2$$
$$= (\Delta x_1)^2 + (\Delta x_2)^2 + (\Delta x_3)^2 \cdots$$

- L-infinity

$$d(\boldsymbol{x^0}, \boldsymbol{x}) = \|\Delta \boldsymbol{x}\|_\infty$$
$$= max\{|\Delta x_1|, |\Delta x_2|, |\Delta x_3|, \dots\}$$

small L-∞

Change every pixel a little bit



same L2

Change one pixel much

large L-∞

Attack Approach

$w^*, b^* = arg \min_{w,b} L$ Difference?

Update ***input***, not ***parameters***

$$\boldsymbol{x}^* = arg \quad \min \quad L(\boldsymbol{x})$$

Gradient Descent

Start from original image $\boldsymbol{x^0}$

For $t = 1$ to T

$$\boldsymbol{x^t} \leftarrow \boldsymbol{x^{t-1}} - \eta \boldsymbol{g}$$

$$\boldsymbol{g} = \begin{bmatrix} \frac{\partial L}{\partial x_1}\Big|_{\boldsymbol{x}=\boldsymbol{x^{t-1}}} \\ \frac{\partial L}{\partial x_2}\Big|_{\boldsymbol{x}=\boldsymbol{x^{t-1}}} \\ \vdots \end{bmatrix}$$

Attack Approach

$w^*, b^* = arg \min_{w,b} L$ Difference?

Update ***input***, not ***parameters***

Different optimization methods

$$x^* = arg \min_{d(x^0, x) \le \varepsilon} L(x)$$

Different constraints

Gradient Descent

Start from original image x^0

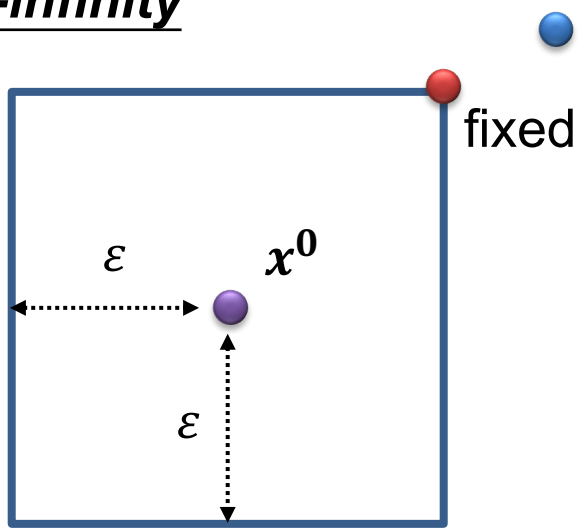
For $t = 1$ to T

$$x^t \leftarrow x^{t-1} - \eta g$$

If $d(x^0, x^t) > \varepsilon$

$$x^t \leftarrow fix(x^t)$$

L-infinity



Attack Approach

$$\boldsymbol{x}^* = arg \min_{d(\boldsymbol{x^0},\boldsymbol{x}) \leq \varepsilon} L(\boldsymbol{x})$$

Fast Gradient Sign Method (FGSM) https://arxiv.org/abs/1412.6572

Start from original image $\boldsymbol{x^0}$

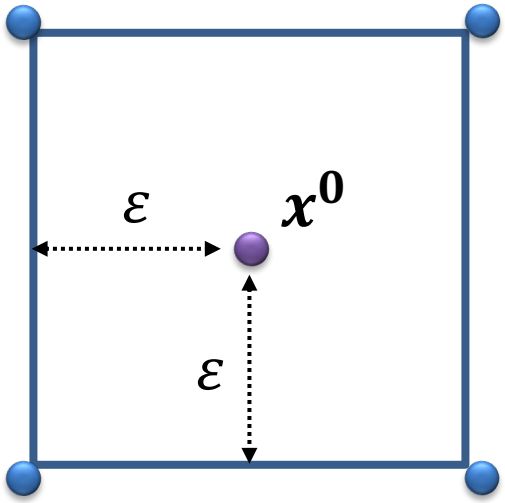
For $t = 1$ ~~to T~~

$$\boldsymbol{x^t} \leftarrow \boldsymbol{x^{t-1}} - \eta \boldsymbol{g}$$

Attack Approach

L-infinity

$$x^* = arg \min_{d(x^0, x) \le \varepsilon} L(x)$$



Fast Gradient Sign Method (FGSM) https://arxiv.org/abs/1412.6572

Start from original image x^0

For $t = 1$ ~~to T~~

$$x^t \leftarrow x^{t-1} - \eta g$$

$$g = \pm 1 \begin{bmatrix} sign\left(\frac{\partial L}{\partial x_1}\Big|_{x=x^{t-1}}\right) \\ \pm 1 \\ sign\left(\frac{\partial L}{\partial x_2}\Big|_{x=x^{t-1}}\right) \\ \vdots \end{bmatrix}$$

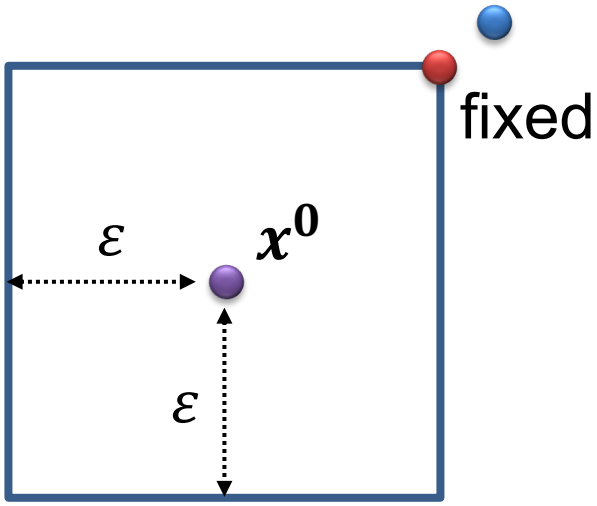
ε $\begin{bmatrix} +1 \\ -1 \\ +1 \\ \vdots \end{bmatrix}$

$$if\ t > 0, sign(t) = 1; otherwise, sign(t) = -1$$

Attack Approach

L-infinity

$$x^* = arg \min_{d(x^0, x) \le \varepsilon} L(x)$$



Iterative FGSM

https://arxiv.org/abs/1607.02533

Start from original image x^0

For $t = 1$ ~~to T~~

$x^t \leftarrow x^{t-1} - \eta g$

If $d(x^0, x) > \varepsilon$

$x^t \leftarrow fix(x^t)$

$$g = \begin{matrix} \pm 1 \\ \pm 1 \\ \\ \end{matrix} \begin{bmatrix} sign\left(\frac{\partial L}{\partial x_1}\Big|_{x=x^{t-1}}\right) \\ sign\left(\frac{\partial L}{\partial x_2}\Big|_{x=x^{t-1}}\right) \\ \vdots \end{bmatrix}$$

White Box v.s. Black Box

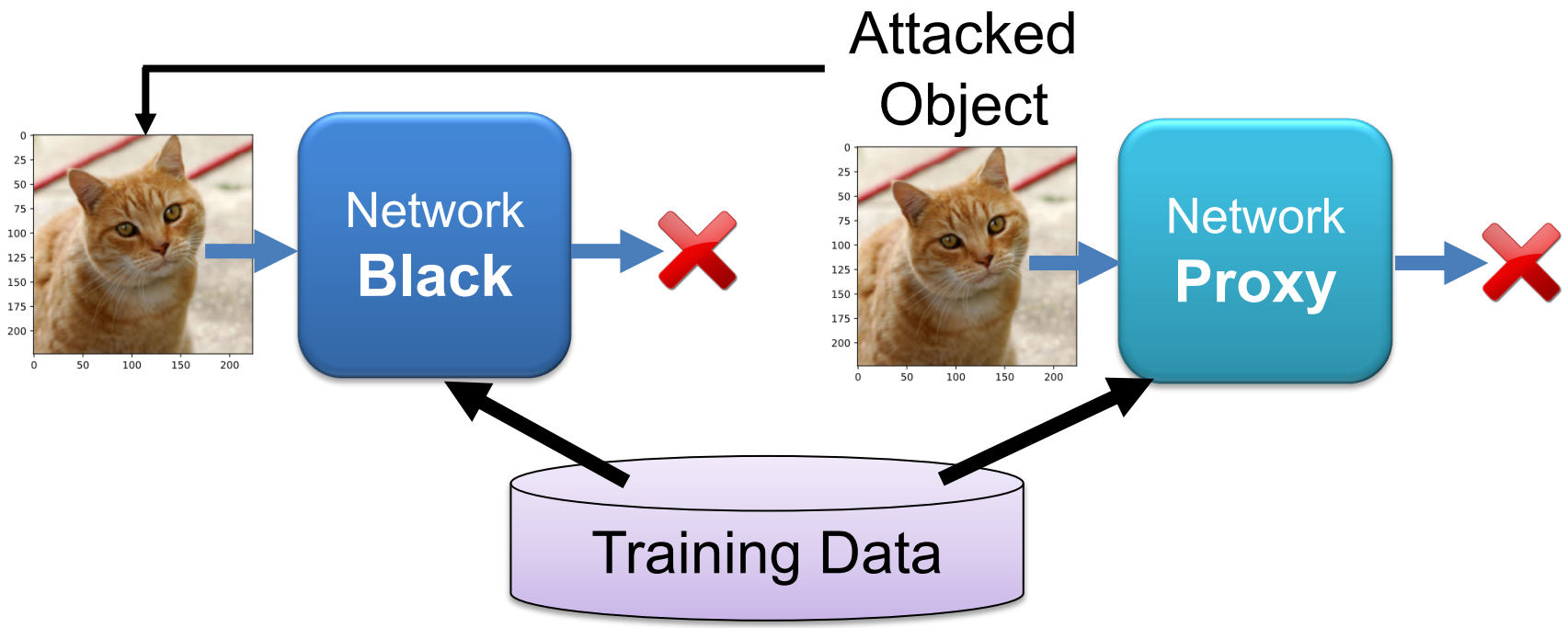
- In the previous attack, we know the network parameters θ
  - This is called ***White Box Attack***.
- You cannot obtain model parameters in most online API.
- Are we safe if we do not release model? ☺
- No, because ***Black Box Attack*** is possible. ☹

$$\boldsymbol{g} = \begin{bmatrix} sign\left(\frac{\partial L}{\partial x_1}\Big|_{\boldsymbol{x}=\boldsymbol{x}^{t-1}}\right) \\ sign\left(\frac{\partial L}{\partial x_2}\Big|_{\boldsymbol{x}=\boldsymbol{x}^{t-1}}\right) \\ \vdots \end{bmatrix}$$

Black Box Attack

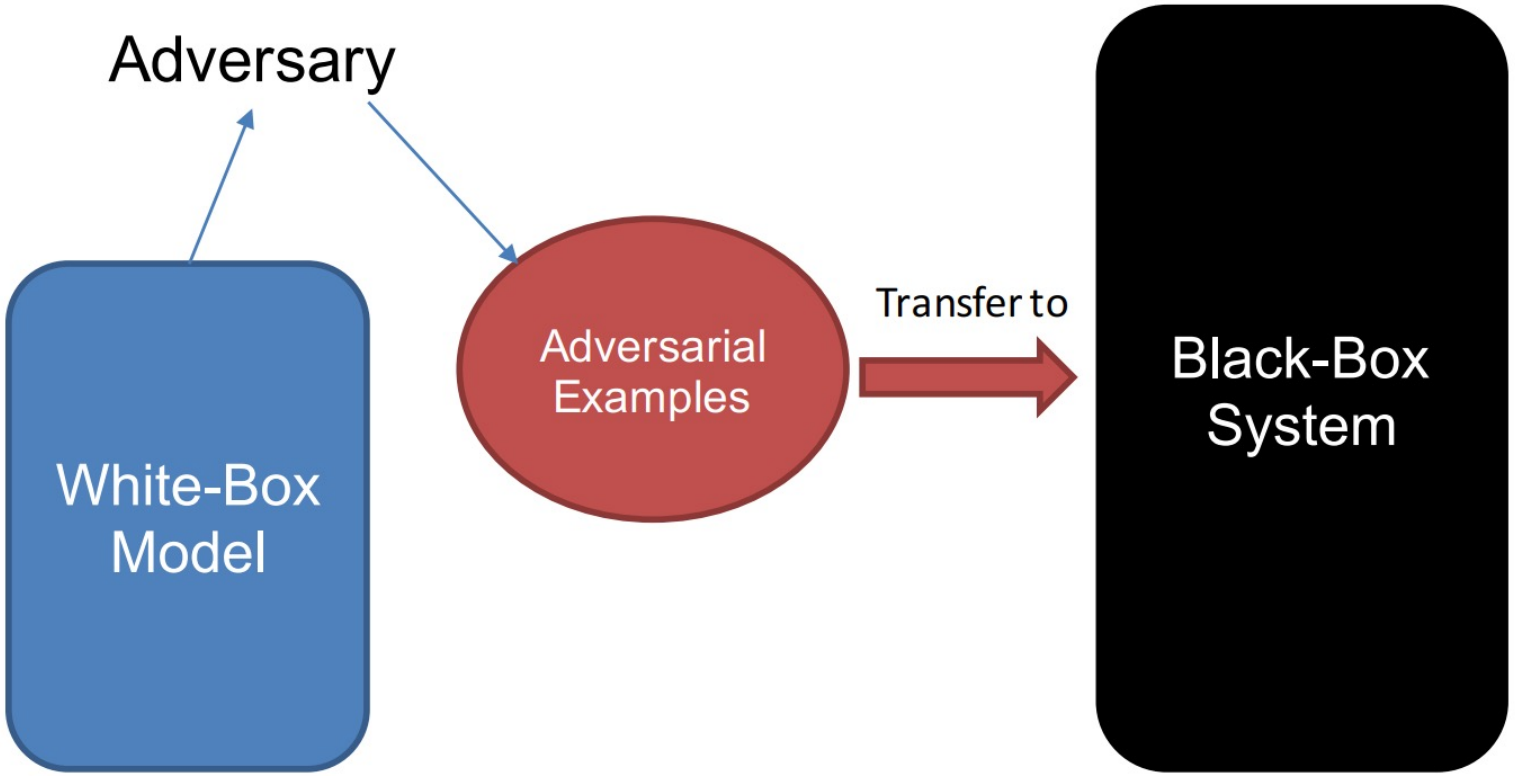
If you have the training data of the target network

- Train a proxy network yourself
- Using the proxy network to generate attacked objects



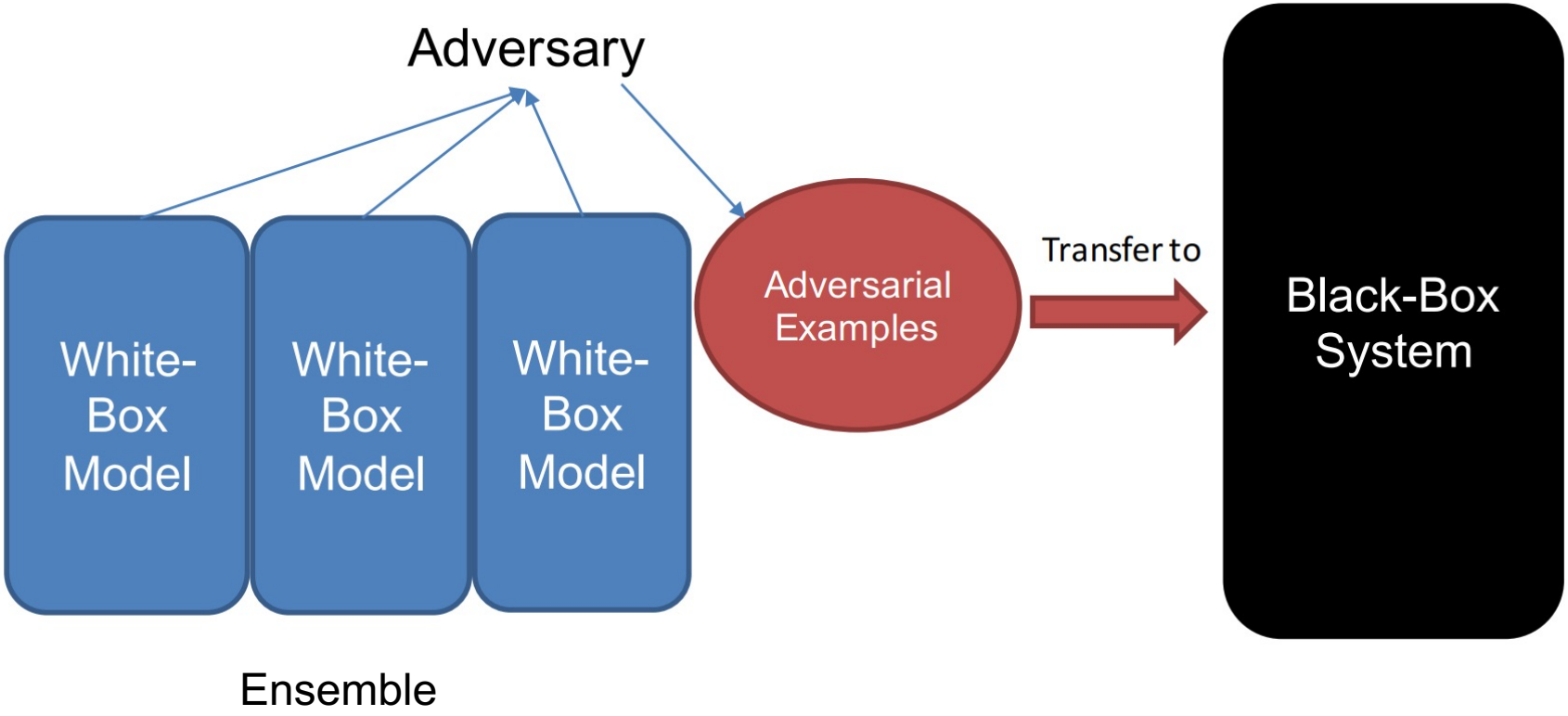
What if we do not know the training data?

Black-box Attacks Based On Transferability



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

Black-box Attacks Based On Transferability



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

Black Box Attack

https://arxiv.org/pdf/1611.02770.pdf

To Be Attacked

Proxy

	ResNet-152	ResNet-101	ResNet-50	VGG-16	GoogLeNet
ResNet-152	0%	13%	18%	19%	11%
ResNet-101	19%	0%	21%	21%	12%
ResNet-50	23%	20%	0%	21%	18%
VGG-16	22%	17%	17%	0%	5%
GoogLeNet	39%	38%	34%	19%	0%

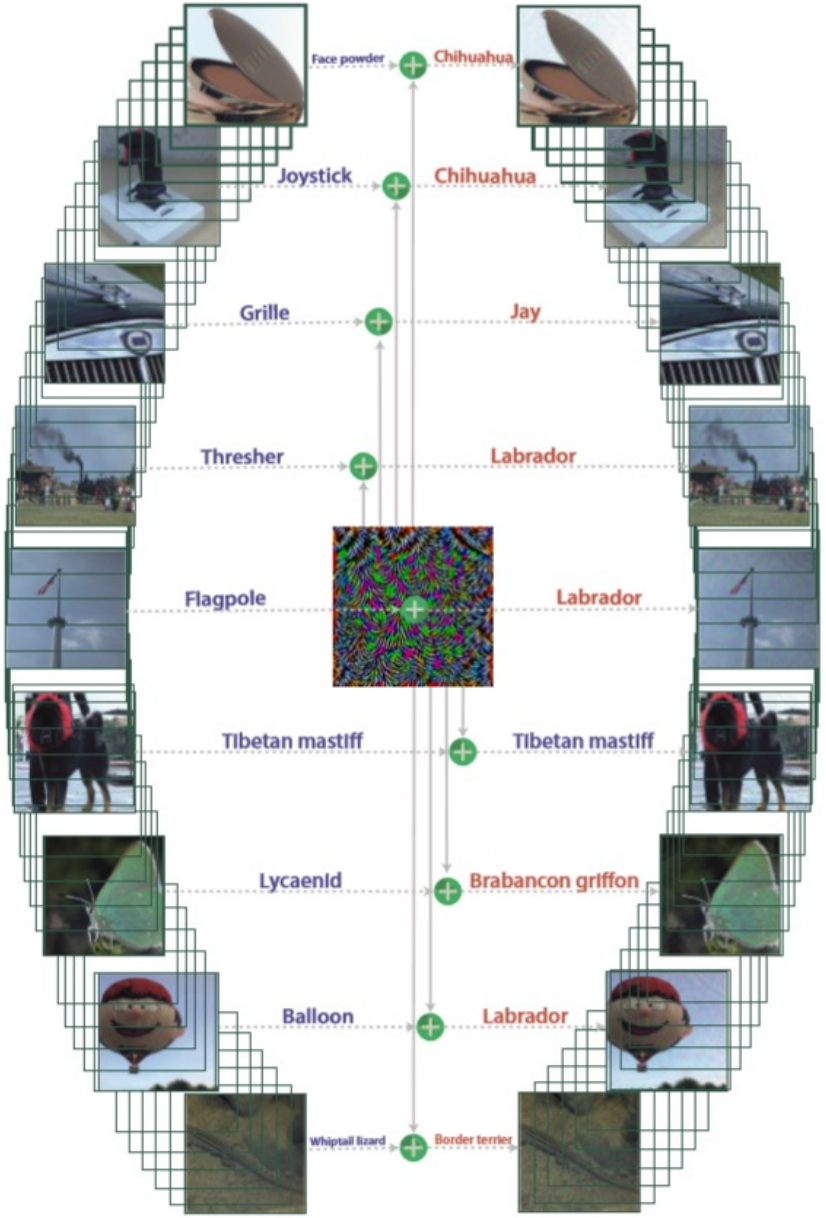
(lower accuracy → more successful attack)

Ensemble Attack

	ResNet-152	ResNet-101	ResNet-50	VGG-16	GoogLeNet
-ResNet-152	0%	0%	0%	0%	0%
-ResNet-101	0%	1%	0%	0%	0%
-ResNet-50	0%	0%	2%	0%	0%
-VGG-16	0%	0%	0%	6%	0%
-GoogLeNet	0%	0%	0%	0%	5%

Universal Adversarial Attack

The existence of a universal (image-agnostic) and very small perturbation vector that causes natural images to be misclassified with high probability.

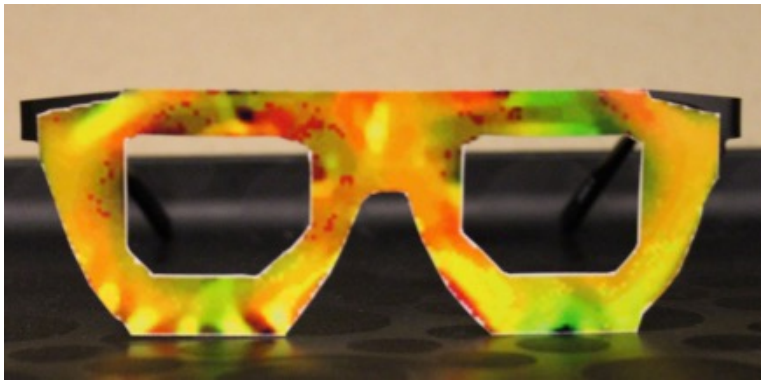


Black Box Attack is also possible!

Moosavi-Dezfooli, Seyed-Mohsen, et al. "Universal adversarial perturbations." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

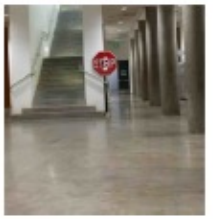
Attack in the Physical World

https://www.cs.cmu.edu/~sbhagava/papers/face-rec-ccs16.pdf



- An attacker would need to find perturbations that generalize beyond a single image.

- It is desirable to craft perturbations that are comprised mostly of colors reproducible by the printer.

Distance/Angle	Subtle Poster	Subtle Poster Right Turn	Camouflage Graffiti	Camouflage Art (LISA-CNN)	Camouflage Art (GTSRB-CNN)
5′ 0°					
5′ 15°					
10′ 0°					
10′ 30°					
40′ 0°					
Targeted-Attack Success	100%	73.33%	66.67%	100%	80%

https://arxiv.org/abs/1707.08945

Attack in the Physical World

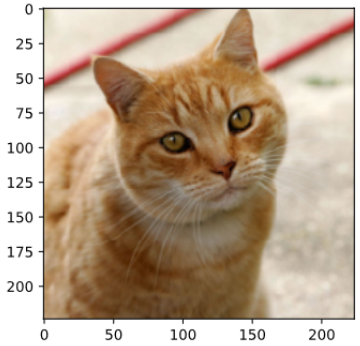
Generating adversarial patches against YOLOv2

https://www.youtube.com/watch?v=MIbFvK2S9g8

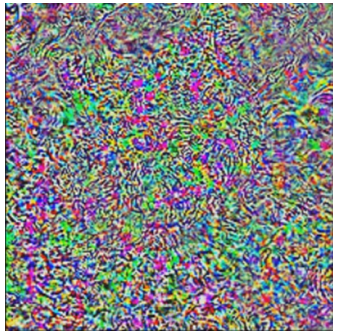
Defense
Passive v.s. Proactive

Passive Defense

Original



+

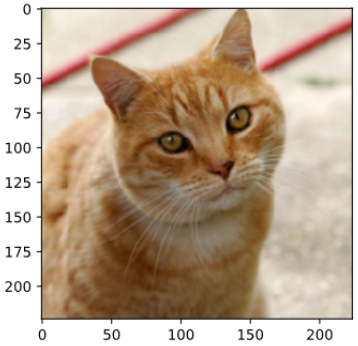


Attack signal

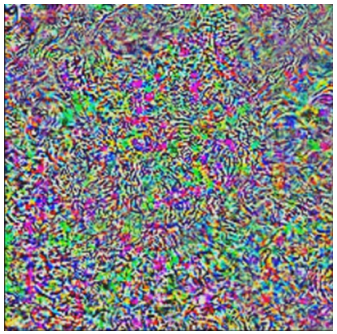
Filter

e.g.
Smoothing

Do not influence
classification



+

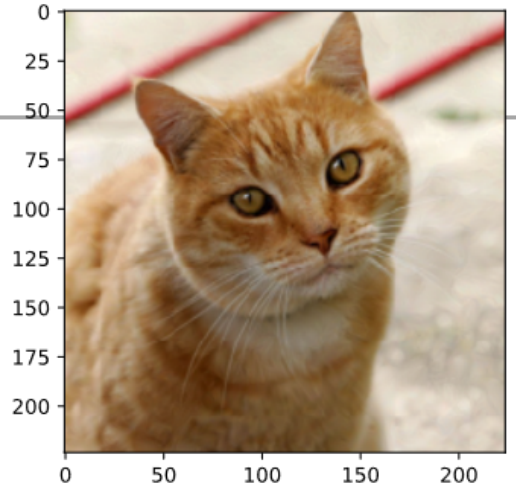


Less harmful

Network

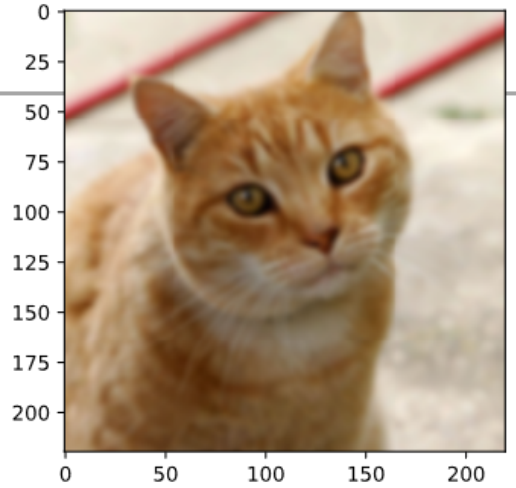
Tiger Cat

~~Keyboard~~

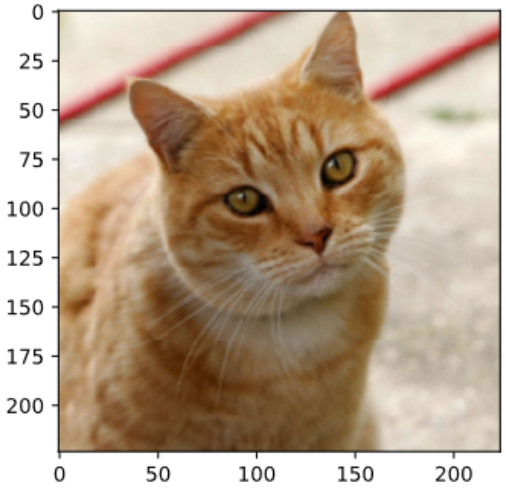


Keyboard
0.98

Smoothing

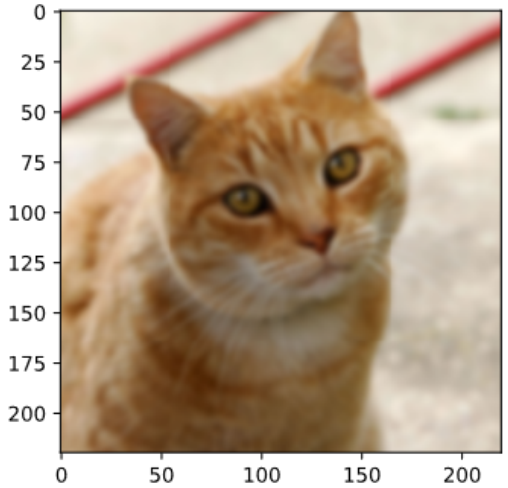


tiger cat
0.37



tiger cat
0.64

Smoothing



tiger cat
0.45 Side Effect!

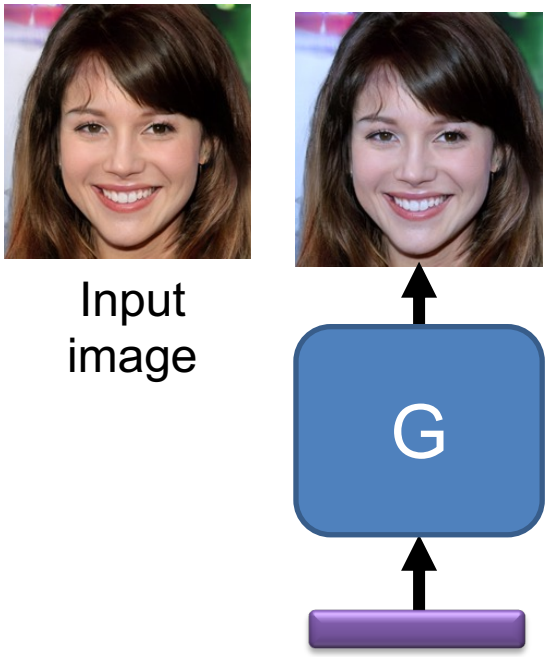
Passive Defense

Image Compression

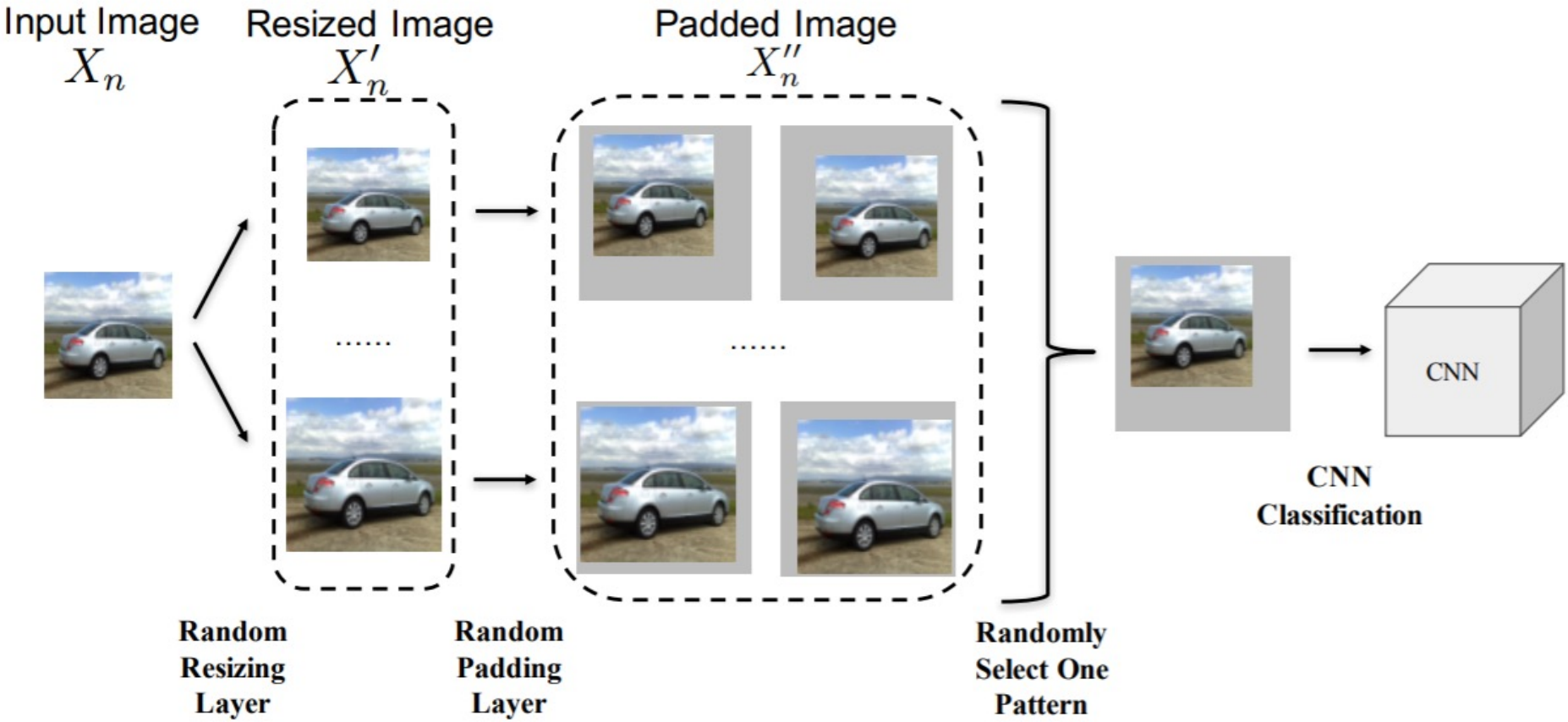
https://arxiv.org/abs/1704.01155
https://arxiv.org/abs/1802.06816

Generator

https://arxiv.org/abs/1805.06605



Passive Defense - Randomization



https://arxiv.org/abs/1711.01991

Proactive Defense

Xie, Cihang, et al. "Adversarial examples improve image recognition." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.

Adversarial Training

Training a model that is robust to adversarial attack.

Given training set $\mathcal{X} = \{(\boldsymbol{x^1}, \hat{y}^1), (\boldsymbol{x^2}, \hat{y}^2), \cdots, (\boldsymbol{x^N}, \hat{y}^N)\}$

Using $\mathcal{X}$ to train your model

For $n = 1$ to N

Find adversarial input $\tilde{\boldsymbol{x}}^{\boldsymbol{n}}$ given $\boldsymbol{x^n}$ by an attack algorithm

We have new training data

$$\mathcal{X}' = \{(\tilde{\boldsymbol{x}}^{\boldsymbol{1}}, \hat{y}^1), (\tilde{\boldsymbol{x}}^{\boldsymbol{2}}, \hat{y}^2), \cdots, (\tilde{\boldsymbol{x}}^{\boldsymbol{N}}, \hat{y}^N)\}$$

Using both $\mathcal{X}$ and $\mathcal{X}'$ to update your model

Data Augmentation

Attack Approaches

- FGSM (https://arxiv.org/abs/1412.6572)
- Basic iterative method (https://arxiv.org/abs/1607.02533)
- L-BFGS (https://arxiv.org/abs/1312.6199)
- Deepfool (https://arxiv.org/abs/1511.04599)
- JSMA (https://arxiv.org/abs/1511.07528)
- C&W (https://arxiv.org/abs/1608.04644)
- Elastic net attack (https://arxiv.org/abs/1709.04114)
- Spatially Transformed (https://arxiv.org/abs/1801.02612)
- One Pixel Attack (https://arxiv.org/abs/1710.08864)
- …… only list a few

Additional References

- Ilyas, Andrew, et al. "Adversarial examples are not bugs, they are features." *Advances in neural information processing systems* 32 (2019).

- Akhtar, Naveed, et al. "Advances in adversarial attacks and defenses in computer vision: A survey." *IEEE Access* 9 (2021): 155161-155196.

Slide Credits

- Many slides are adapted from the existing teaching or tutorial slides by Hung-yi Lee, Bo Li, Stanford course - CS231n: Convolutional Neural Networks for Visual Recognition, and many others

Thank you!

Question?