

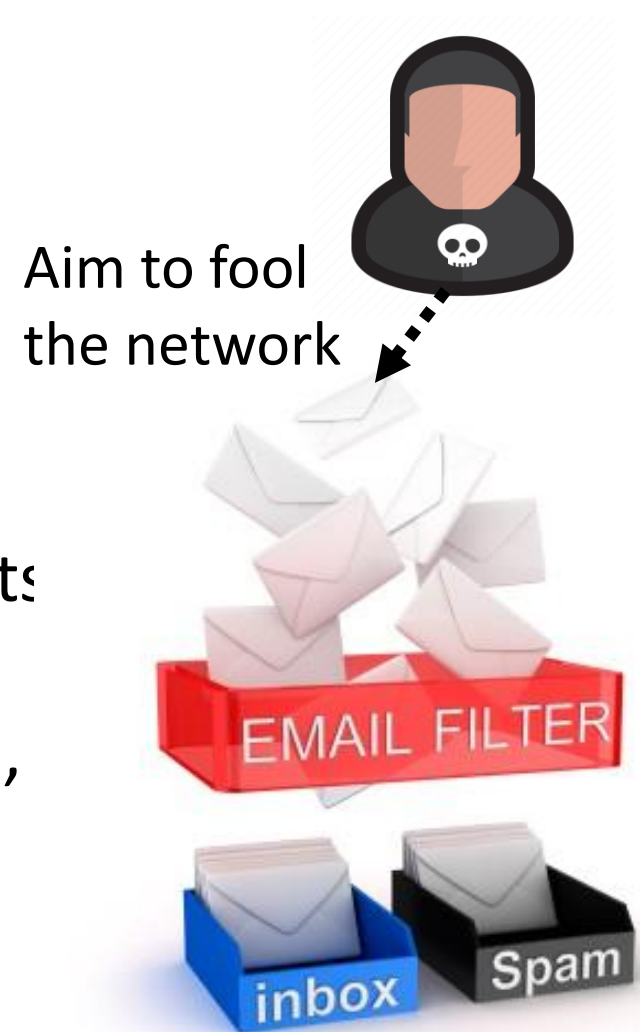
The background of the slide is a dynamic anime-style illustration. It depicts two characters in a close combat pose, with their hands or fists meeting in the center. A massive, bright blue and white energy explosion or shockwave radiates from the point of impact, filling the upper half of the frame. The characters are rendered in a dark, moody style with blue and black tones. The overall composition is energetic and visually striking.

Adversarial Attack

Hung-yi Lee

Motivation


- You have trained many neural networks.
- We seek to deploy neural networks in the real world.
- Are networks robust to the inputs that are built to fool them?
 - Useful for spam classification, malware detection, network intrusion detection, etc.





人類不講武德 ...





How to Attack

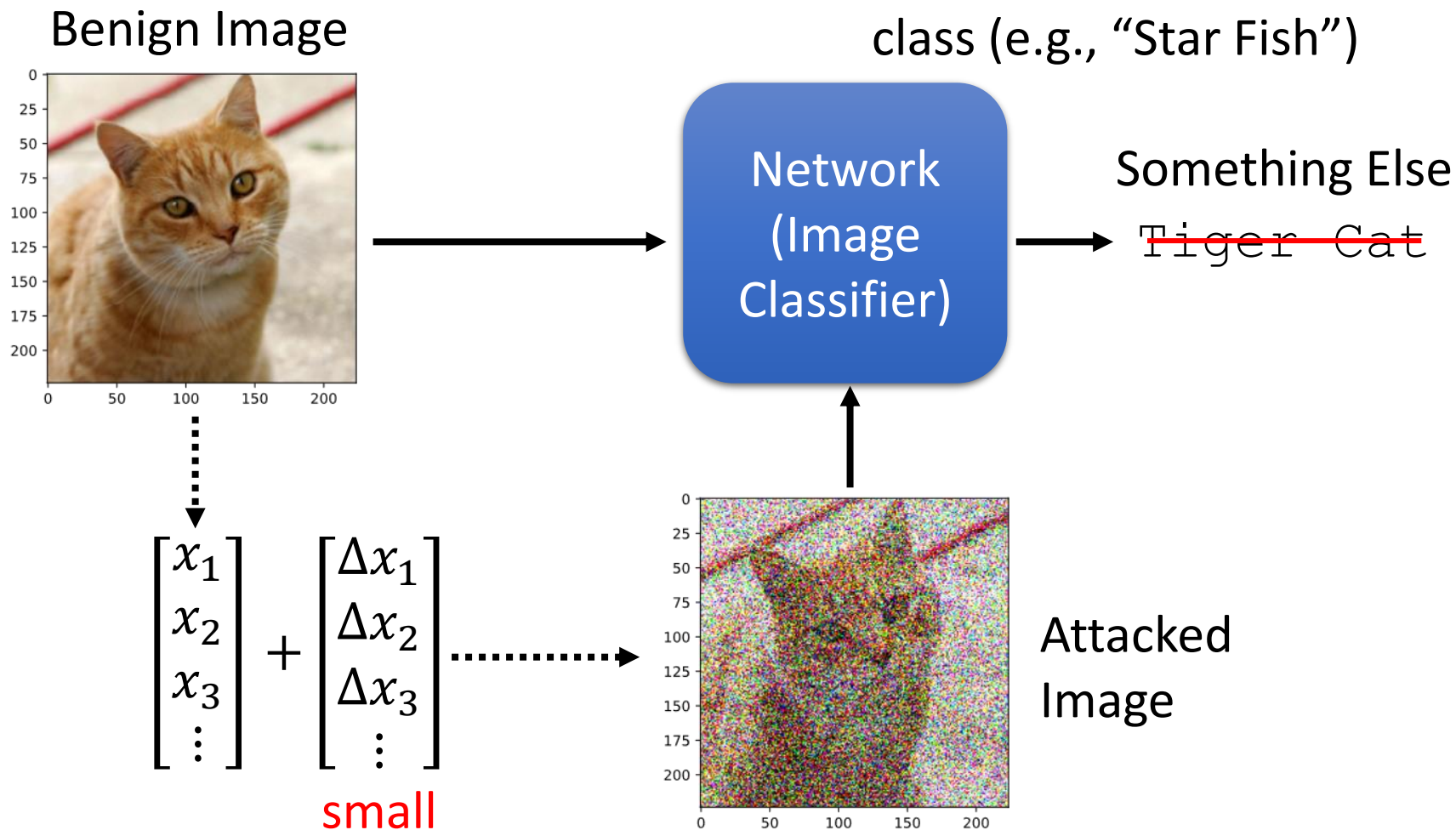
Example of Attack

Non-targeted

Anything other than “Cat”

Targeted

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

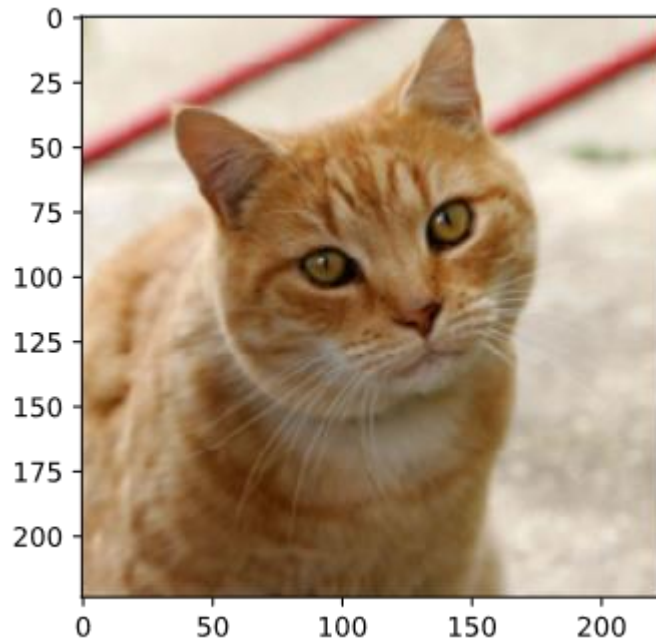


Example of Attack

Network = ResNet-50

The target is “Star Fish”

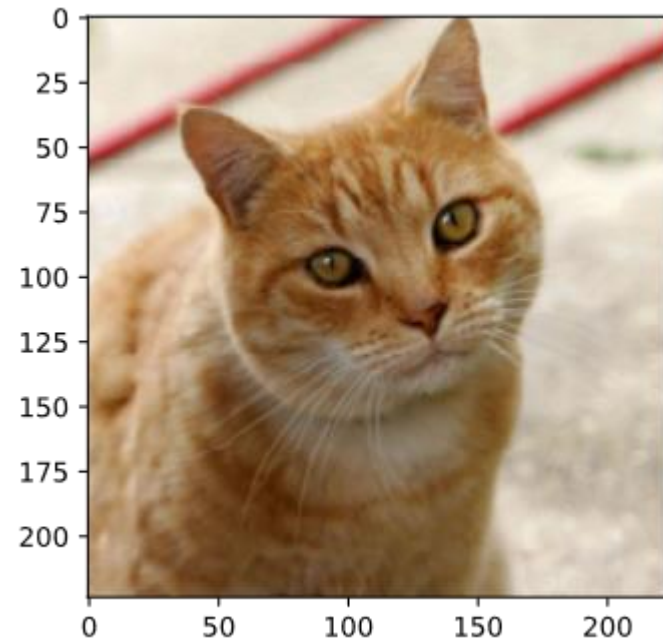
Benign Image



Tiger Cat

0.64

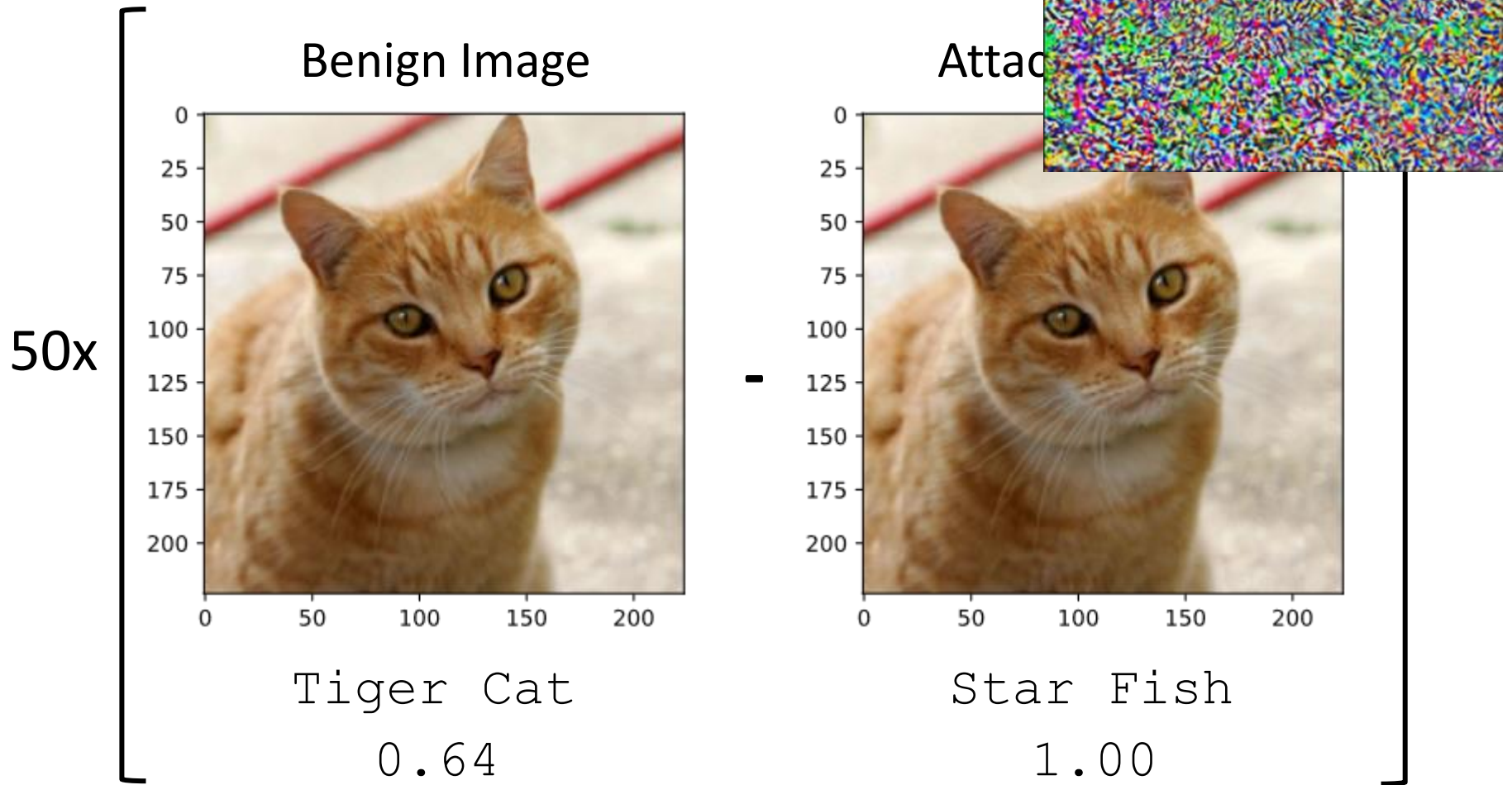
Attacked Image



Star Fish

1.00

Example of Attack



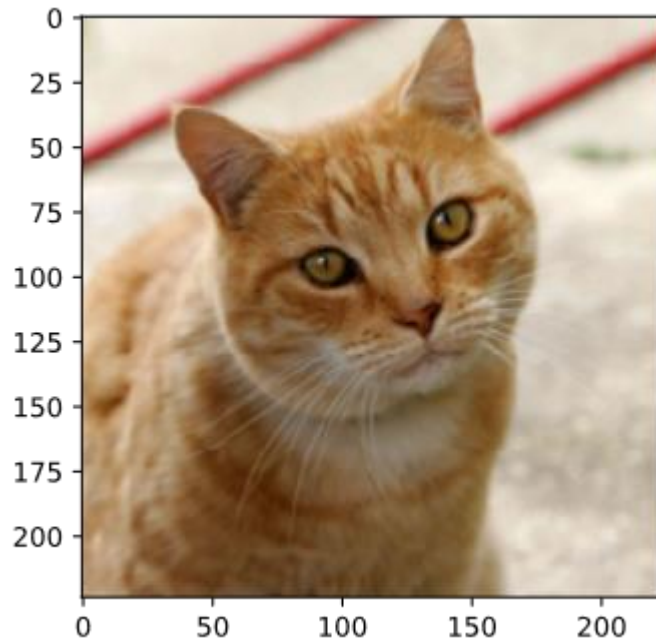
Example of Attack

Network

= ResNet-50

The target is “Keyboard”

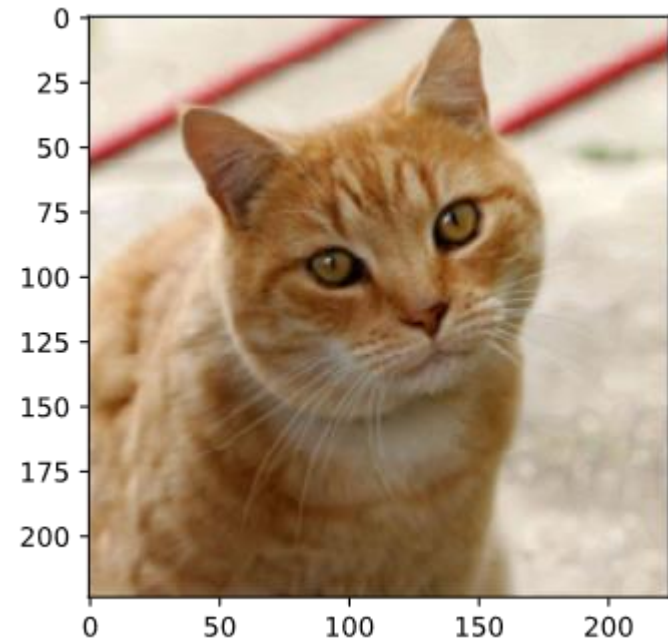
Benign Image



Tiger Cat

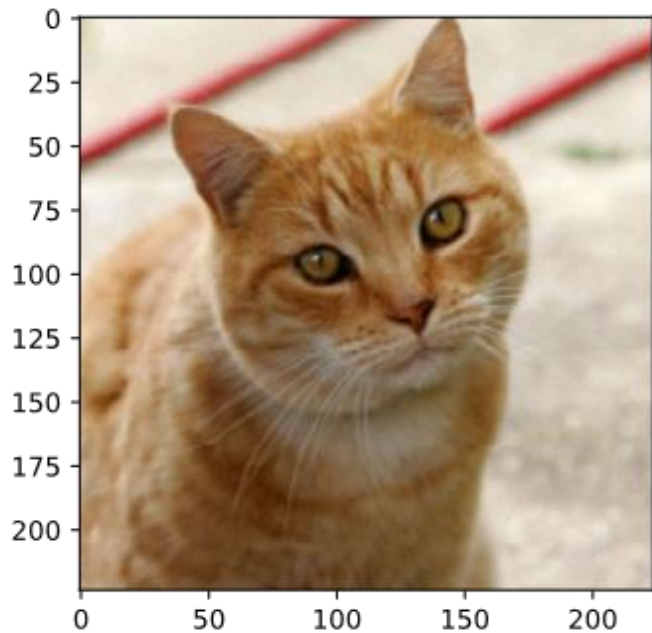
0.64

Attacked Image

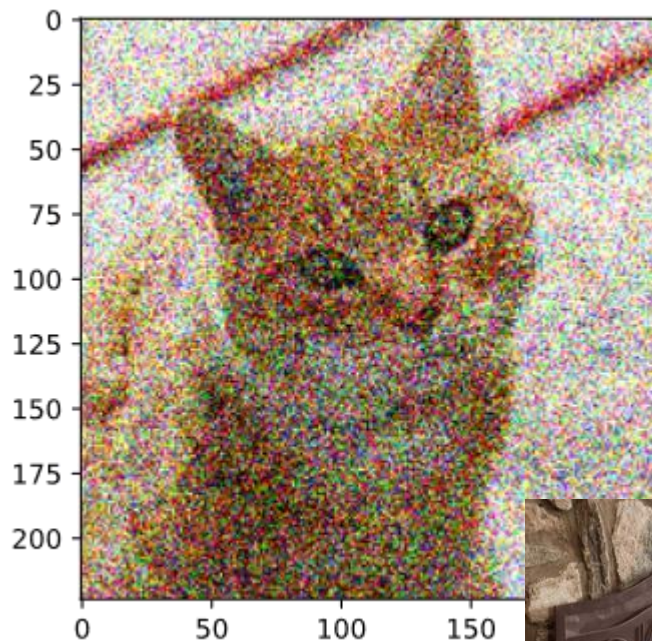


Keyboard

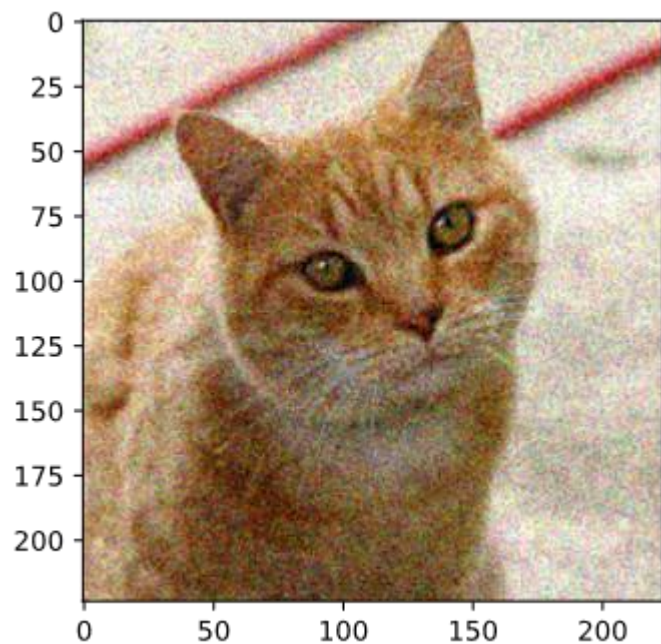
0.98



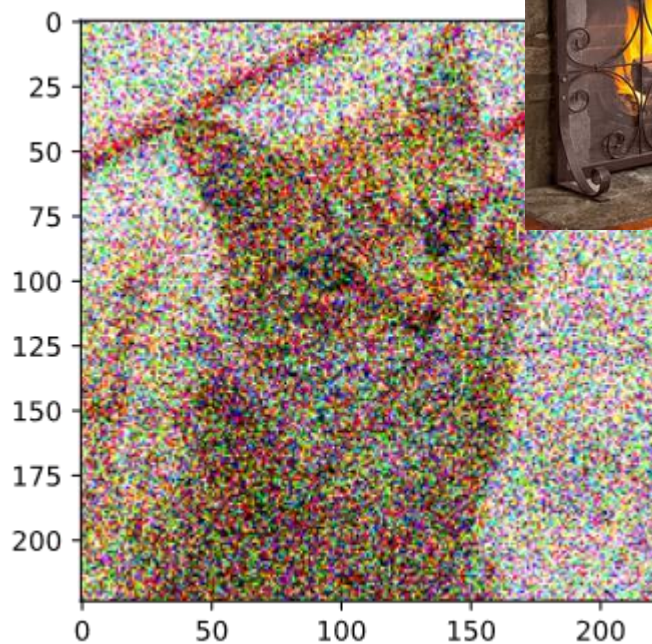
tiger
cat



Persian
cat



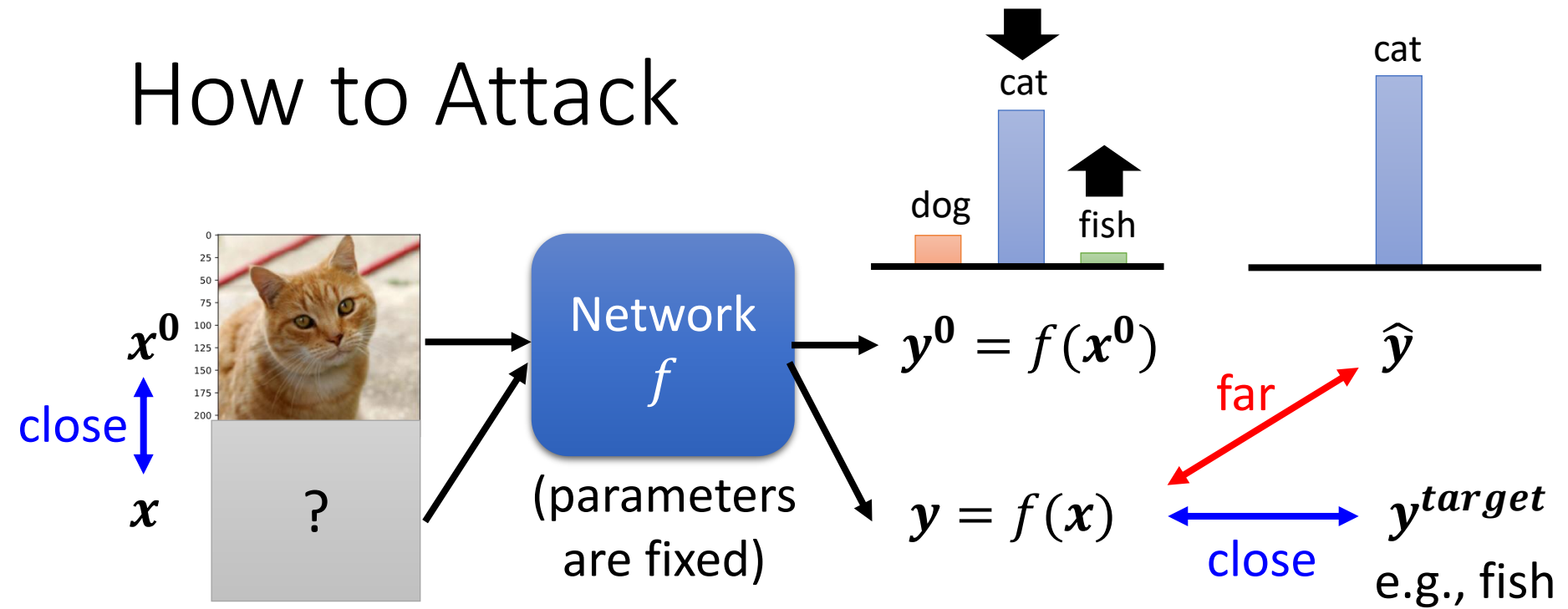
tabby
cat



fire
screen



How to Attack



Non-targeted

$$x^* = \arg \min L(x)$$

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

not perceived
by humans

Targeted

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

Non-perceivable

$$d(\mathbf{x}^0, \mathbf{x}) \leq \varepsilon \quad \text{Need to consider human perception}$$

- L2-norm

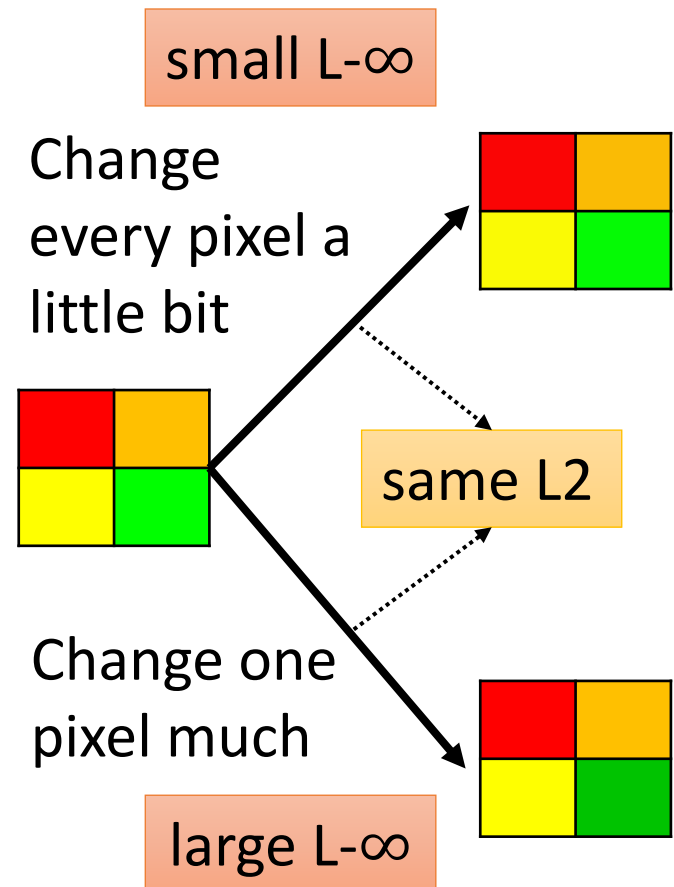
$$\begin{aligned} d(\mathbf{x}^0, \mathbf{x}) &= \|\Delta \mathbf{x}\|_2 \\ &= (\Delta x_1)^2 + (\Delta x_2)^2 + (\Delta x_3)^2 \dots \end{aligned}$$

- L-infinity

$$\begin{aligned} d(\mathbf{x}^0, \mathbf{x}) &= \|\Delta \mathbf{x}\|_\infty \\ &= \max\{|\Delta x_1|, |\Delta x_2|, |\Delta x_3|, \dots\} \end{aligned}$$

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

$\mathbf{x} \qquad \mathbf{x}^0 \qquad \Delta \mathbf{x}$



Attack Approach

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

Update *input*, not *parameters*

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

Gradient Descent

Start from original image \mathbf{x}^0

For $t = 1$ to T

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

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

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

Attack Approach

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

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

Different optimization methods

Different constraints

Gradient Descent

Start from original image x^0

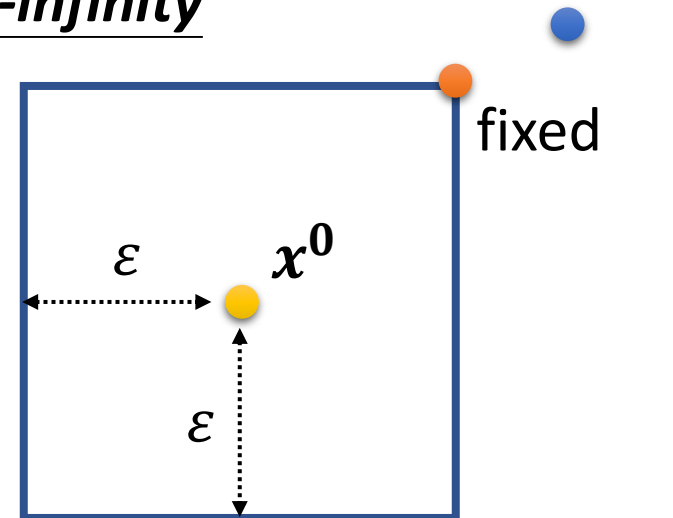
For $t = 1$ to T

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

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

$$x^t \leftarrow \text{fix}(x^t)$$

L-infinity



Attack Approach

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

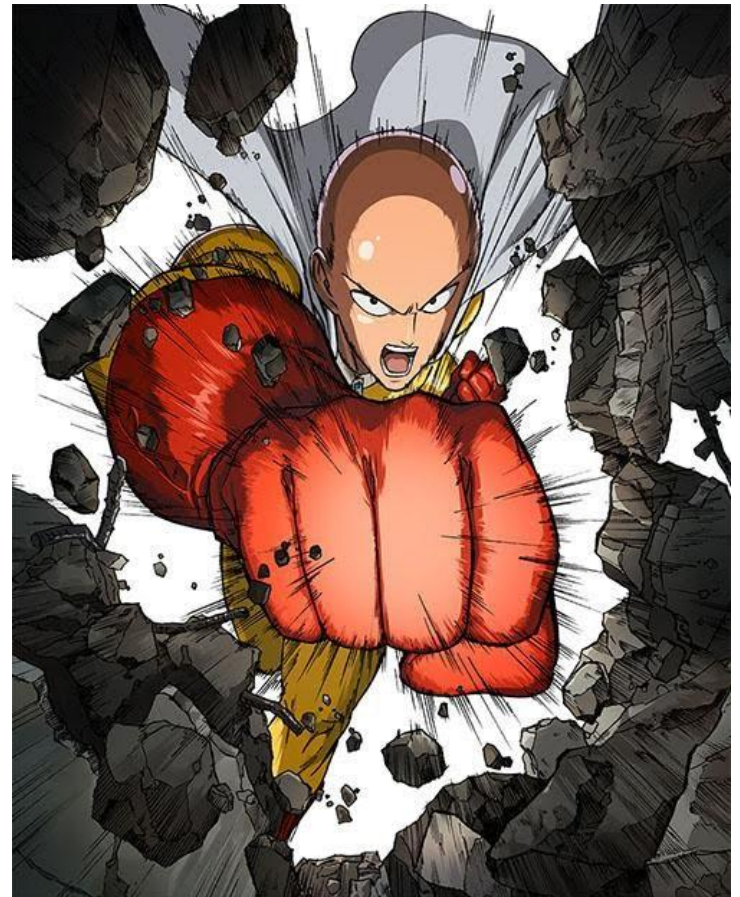
Fast Gradient Sign Method (FGSM)

<https://arxiv.org/abs/1412.6572>

Start from original image \mathbf{x}^0

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

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



Attack Approach

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

Fast Gradient Sign Method (FGSM)

<https://arxiv.org/abs/1412.6572>

Start from original image \mathbf{x}^0

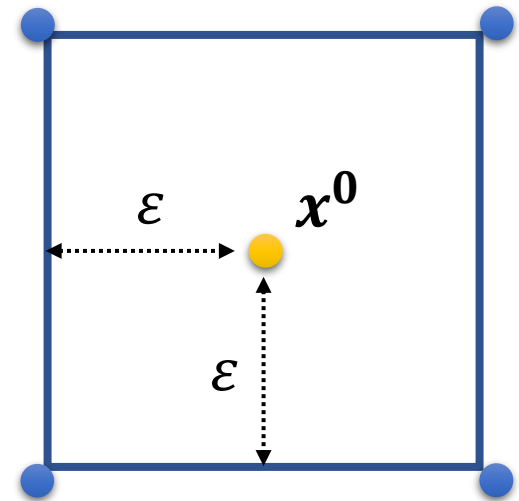
For $t = 1$ ~~to~~ T

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

ε

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

L-infinity



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

if $t > 0$, $\text{sign}(t) = 1$; *otherwise*, $\text{sign}(t) = -1$

Attack Approach

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

Iterative FGSM

<https://arxiv.org/abs/1607.02533>

Start from original image \mathbf{x}^0

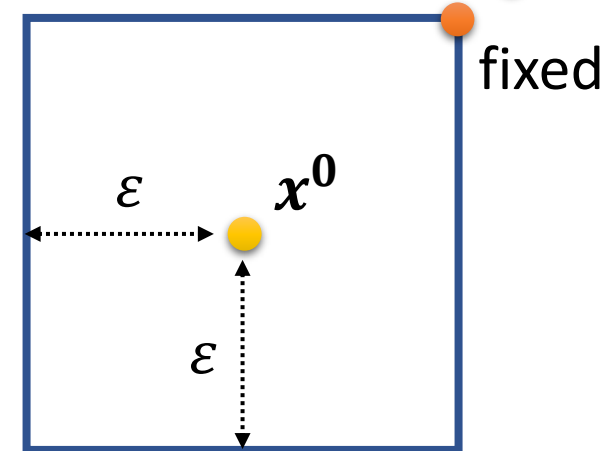
For $t = 1$ ~~to~~ T

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

$$\text{If } d(\mathbf{x}^0, \mathbf{x}) > \varepsilon$$

$$\mathbf{x}^t \leftarrow \text{fix}(\mathbf{x}^t)$$

L-infinity



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

White Box v.s. Black Box

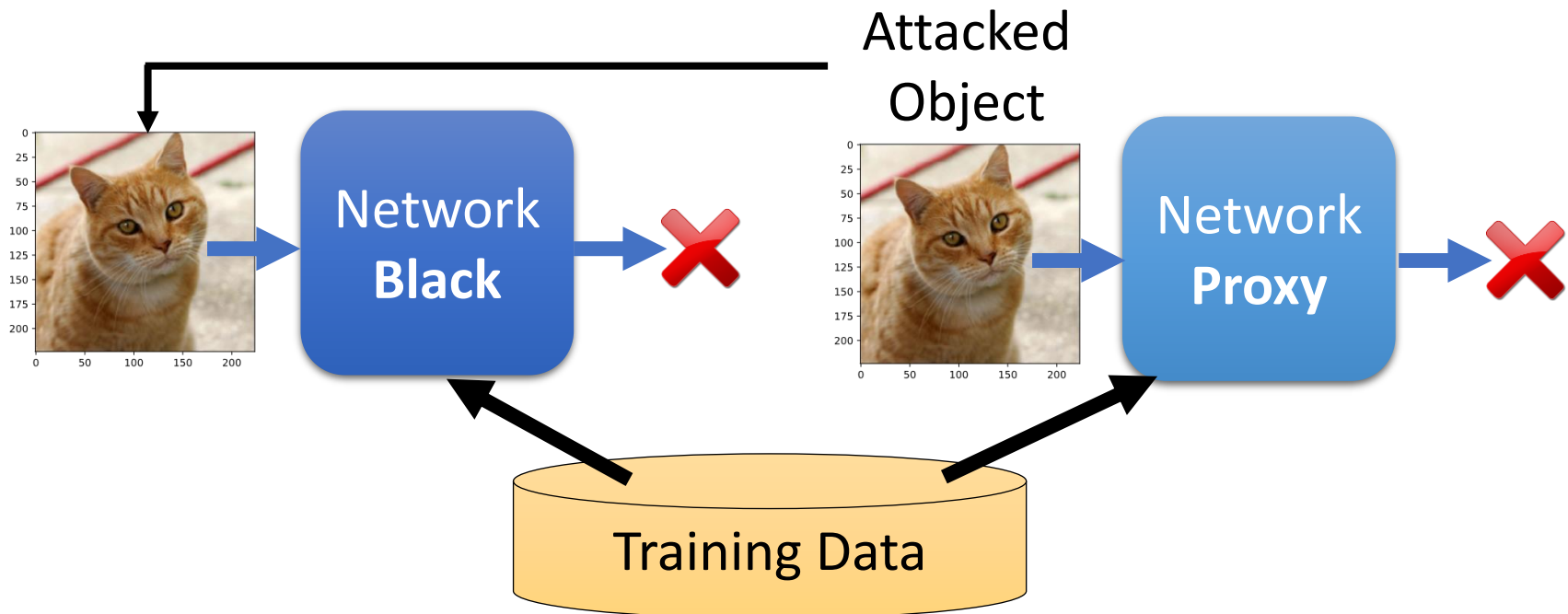
- In the previous attack, we know the network parameters θ
 - This is called **White Box Attack**.
- You cannot obtain model parameters in most online API.
- Are we safe if we do not release model? ☺
- No, because **Black Box Attack** is possible. ☹

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



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 Attack

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

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%

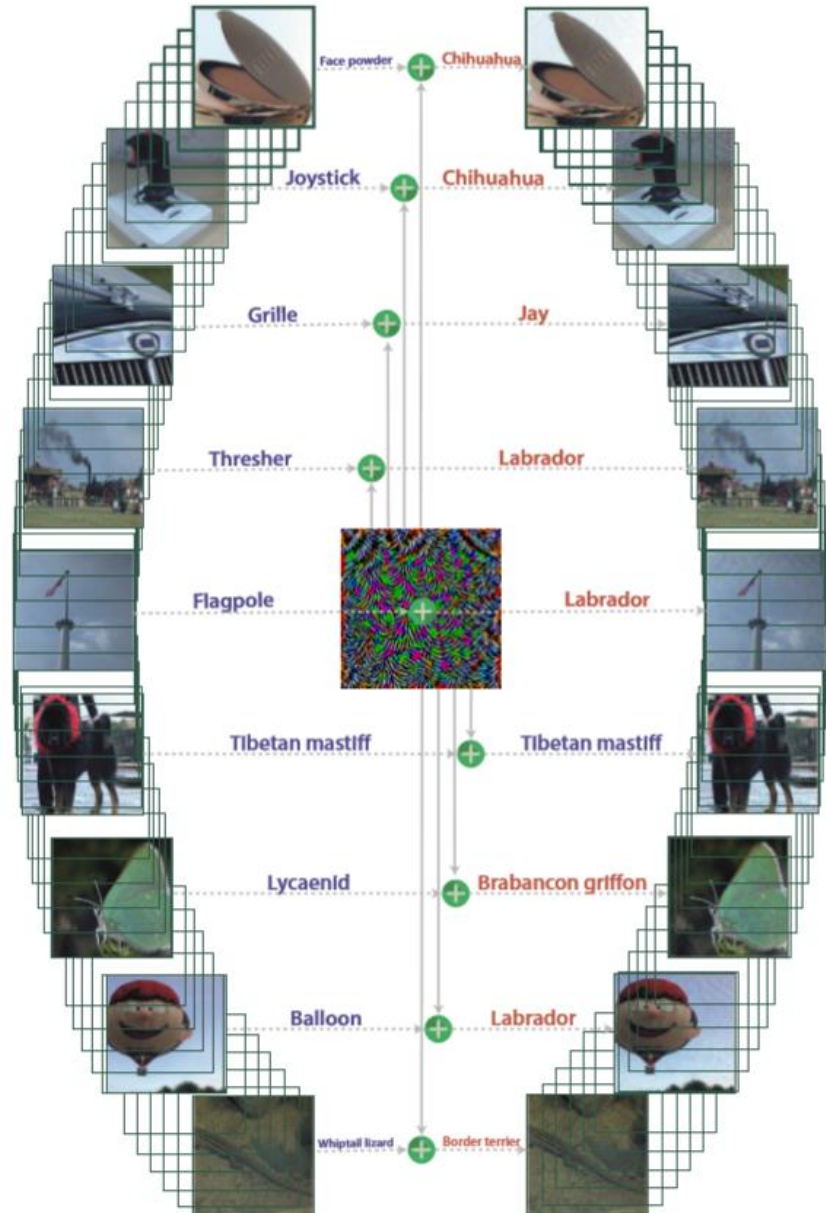
(lower accuracy means the attack is more successful)

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

<https://arxiv.org/abs/1610.08401>



Black Box Attack is also possible!

One pixel attack

Source of image:

<https://arxiv.org/abs/1710.08864>



joystick



Cup(16.48%)
Soup Bowl(16.74%)



Bassinet(16.59%)
Paper Towel(16.21%)



Teapot(24.99%)
Joystick(37.39%)



Hamster(35.79%)
Nipple(42.36%)

Video: <https://youtu.be/tfpKIZIWidA>

Beyond Images

感謝吳海濱同學提供實驗結果

- Speech processing

Detect synthesized
speech

Synthesized!



Real!



- Natural language processing

<https://arxiv.org/abs/1908.07125>

Question: Why did he walk?

For exercise, Tesla walked between 8 to 10 miles per day. He squished his toes one hundred times for each foot every night, saying that it stimulated his brain cells. '

exercise

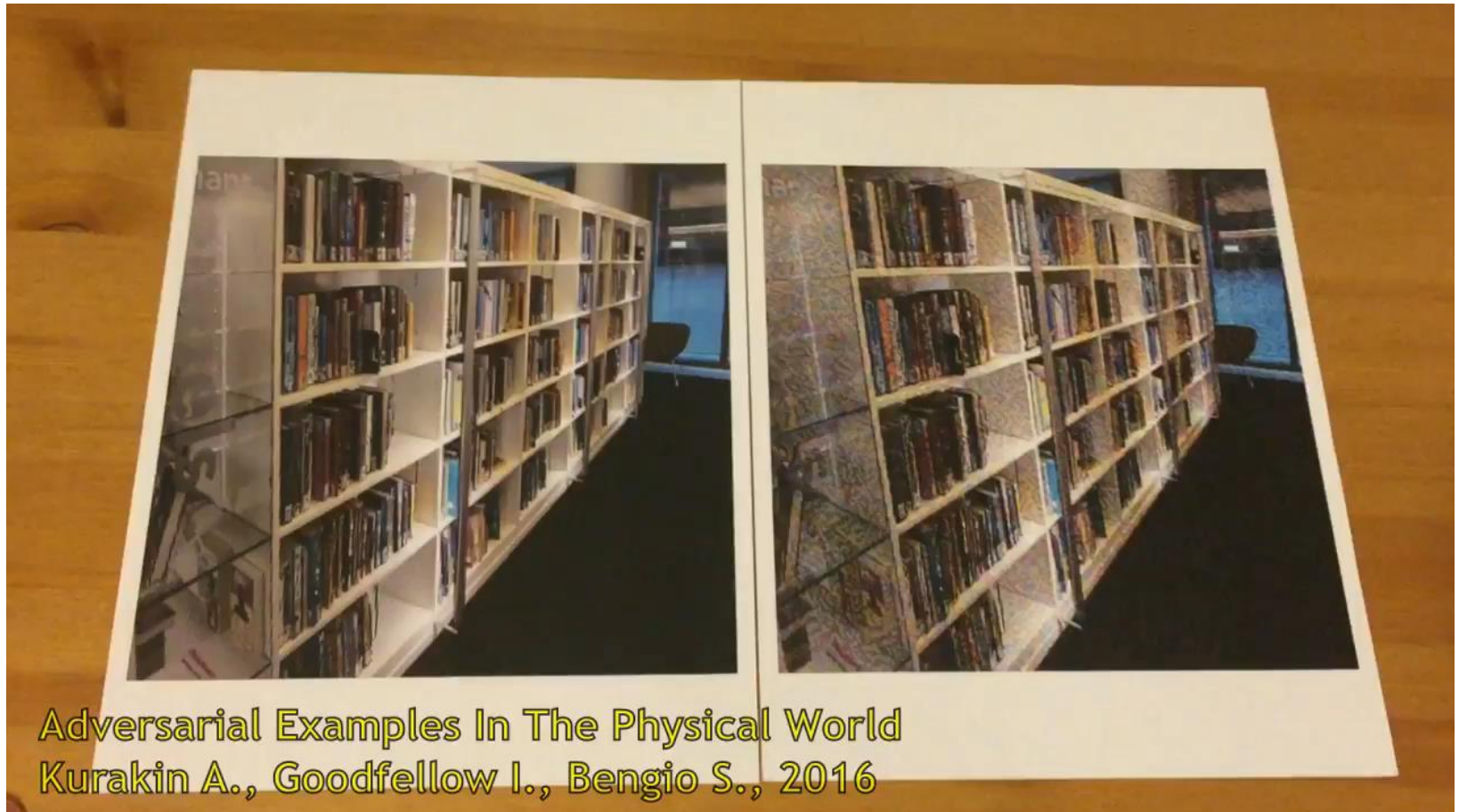
Question: Why did the university see a drop in applicants?

In the early 1950s, student applications declined as a result of increasing crime and poverty in the Hyde Park neighborhood. In response, the university became a

crime and poverty

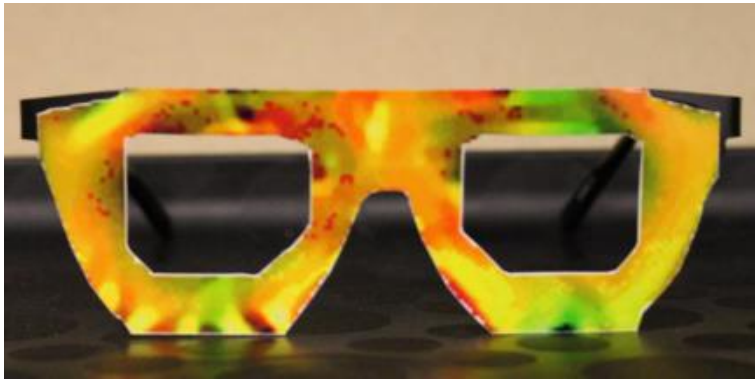
Attack in the Physical World

Black Box Attack




























https://www.youtube.com/watch?v=zQ_uMenoBCk&feature=youtu.be

Attack in the Physical World

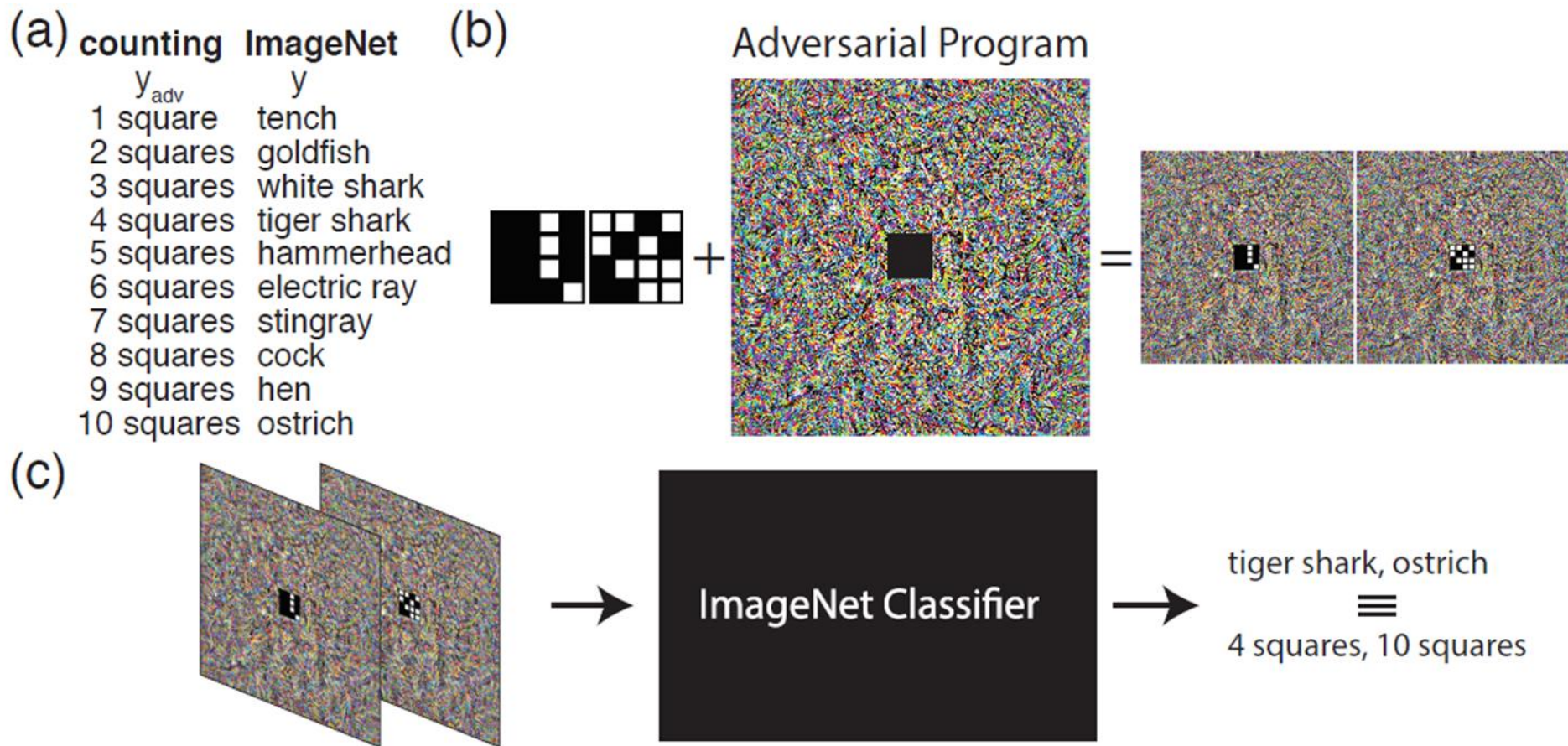


- An attacker would need to find perturbations that generalize beyond a single image.
- Extreme differences between adjacent pixels in the perturbation are unlikely to be accurately captured by cameras.
- It is desirable to craft perturbations that are comprised mostly of colors reproducible by the printer.

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>

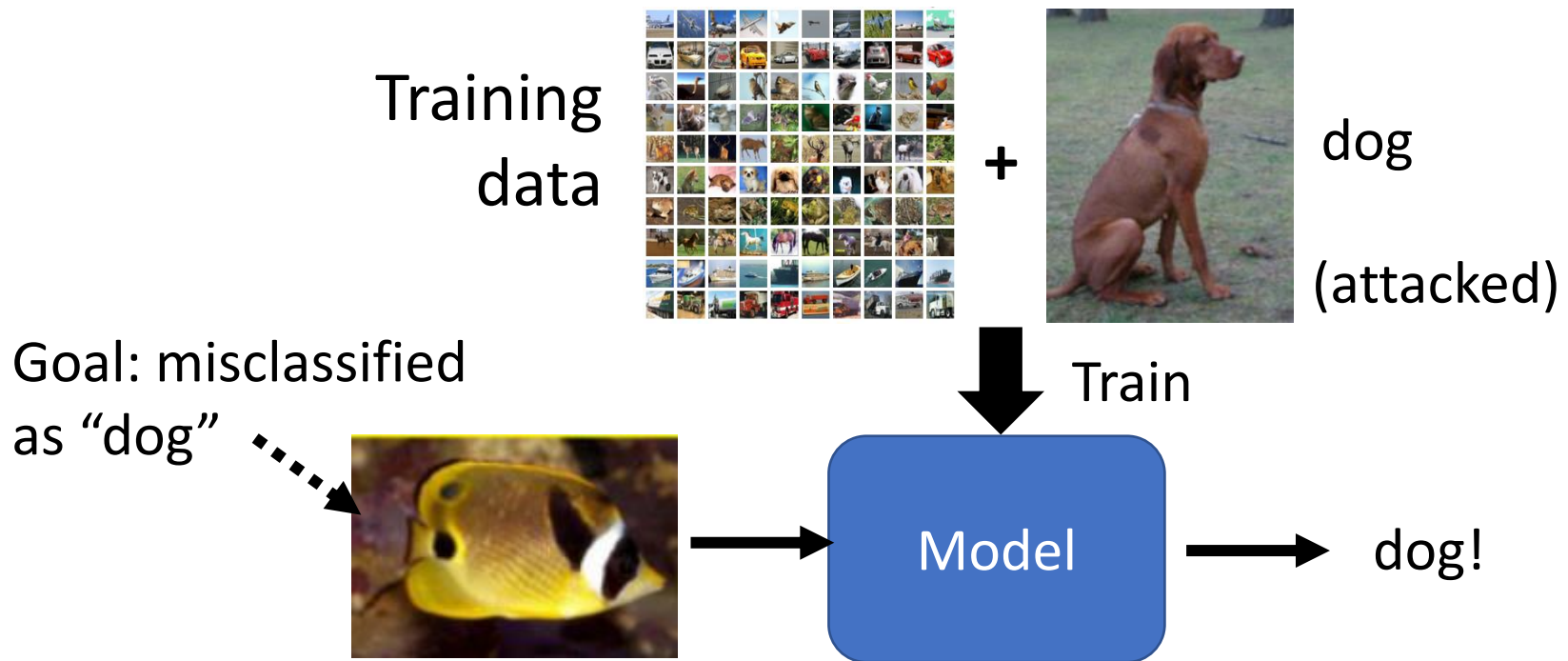
Adversarial Reprogramming



“Backdoor” in Model

<https://arxiv.org/abs/1804.00792>

- Attack happens at the training phase



be careful of unknown dataset

The background of the slide is a close-up, slightly angled view of Captain America's shield. It features concentric rings of red and silver, with a blue center containing a white five-pointed star. The shield is centered on a grey background.

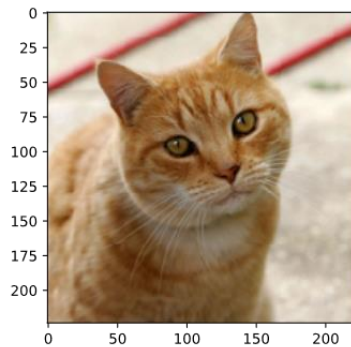
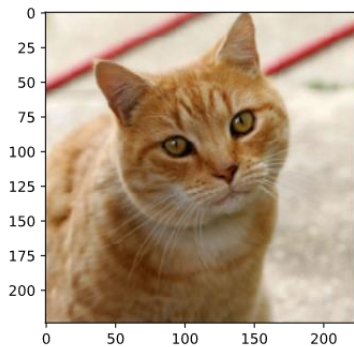
Defense

Passive v.s. Proactive

Passive Defense

Do not influence
classification

Original



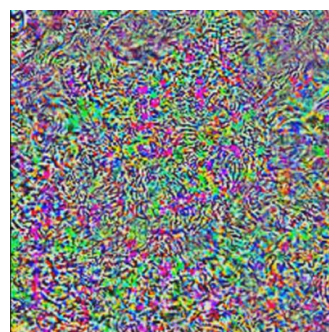
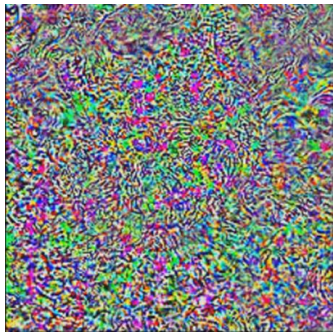
+

Filter

+

Network

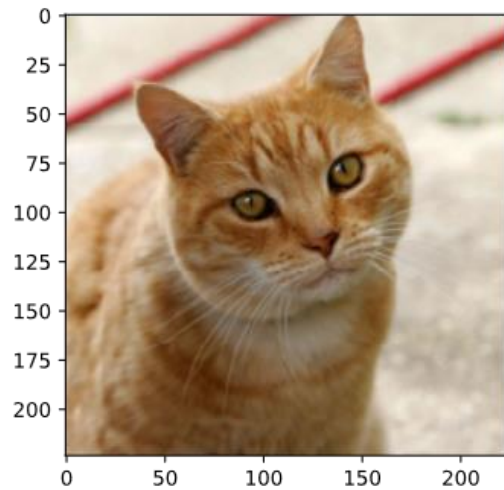
Tiger Cat
~~Keyboard~~



e.g.
Smoothing

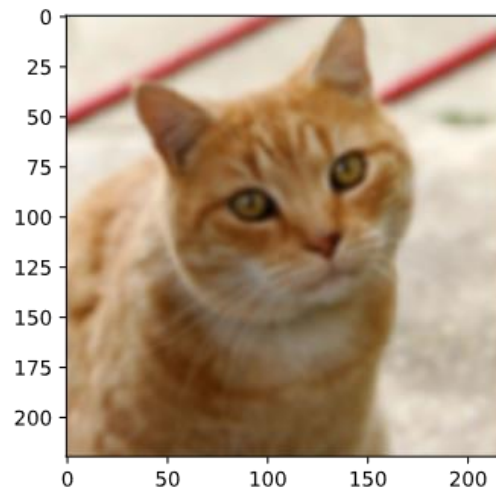
Attack signal

Less harmful

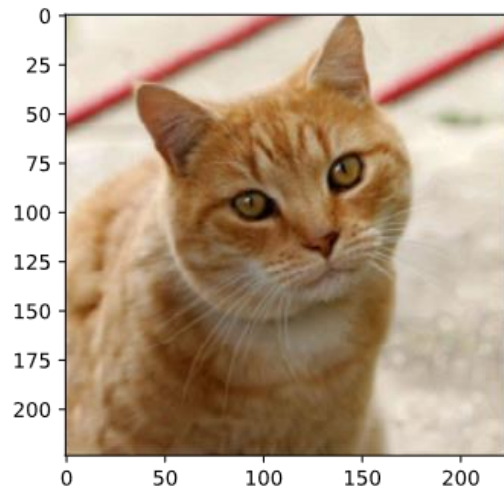


tiger cat
0.64

→
Smoothing

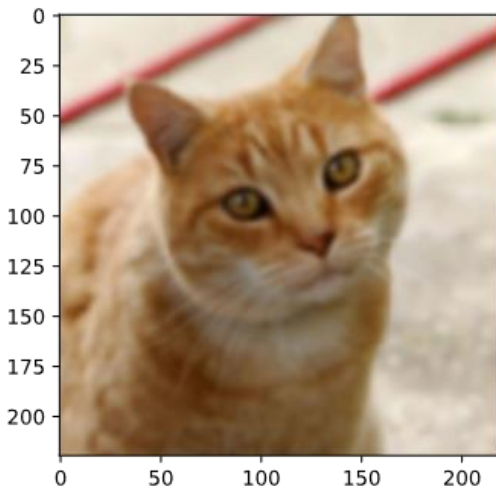


tiger cat
0.45 Side Effect!



Keyboard
0.98

→
Smoothing



tiger cat
0.37

Passive Defense

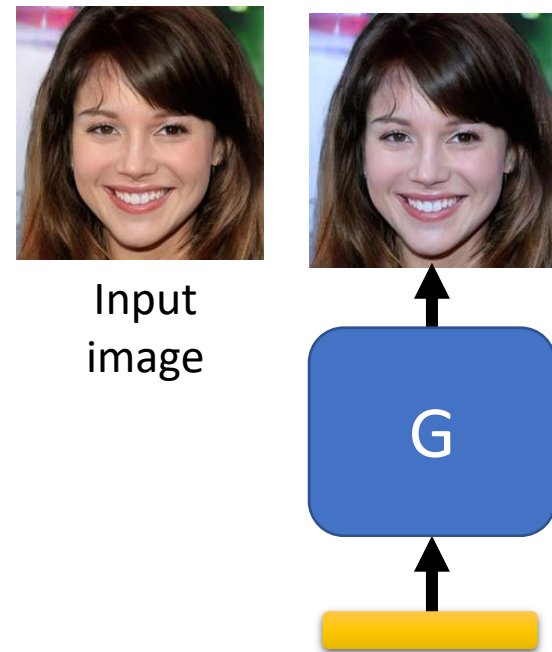
Image Compression



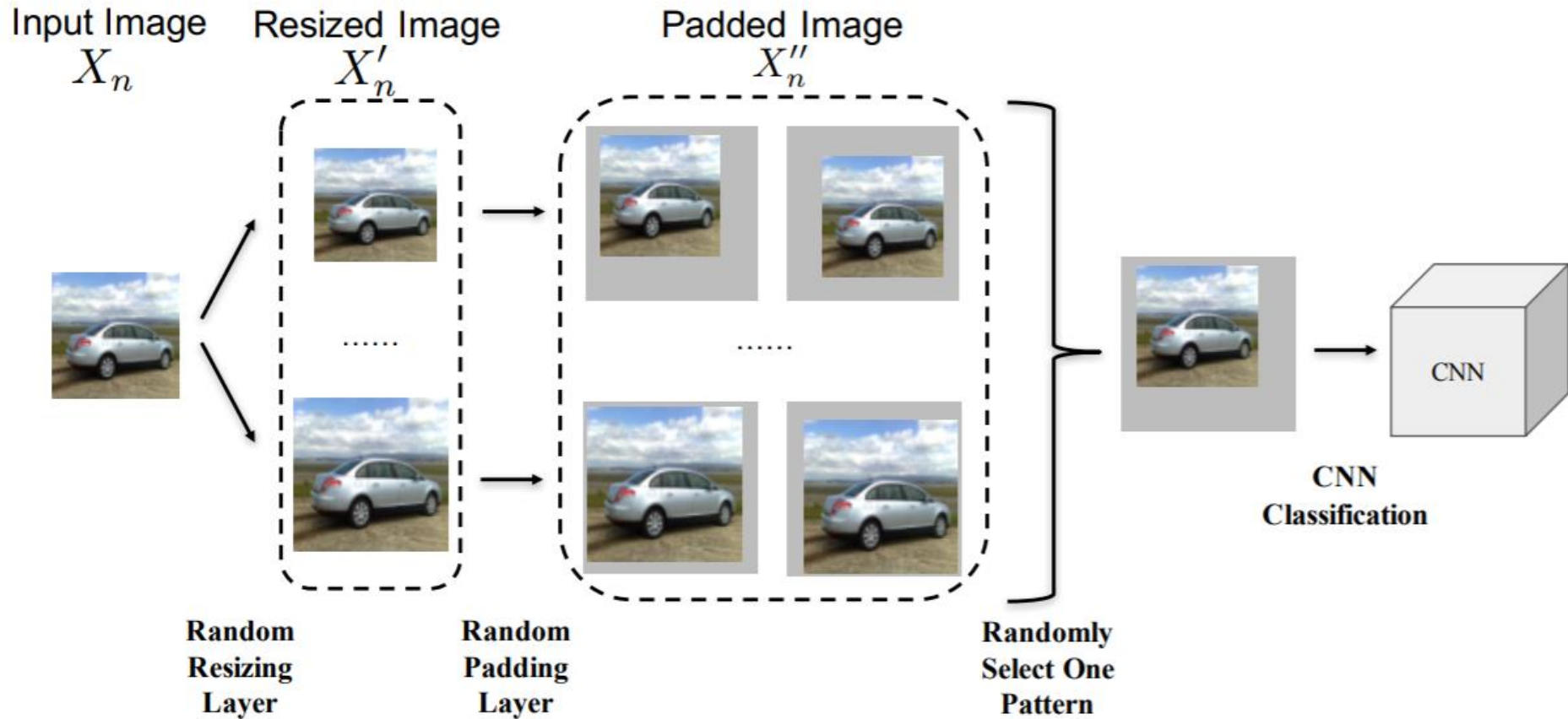
<https://arxiv.org/abs/1704.01155>
<https://arxiv.org/abs/1802.06816>

Generator

<https://arxiv.org/abs/1805.06605>



Passive Defense - Randomization



Proactive Defense

Training a model that is robust to adversarial attack.

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

Using \mathcal{X} to train your model

Using algorithm A

For $n = 1$ to N

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

Find the problem

We have new training data

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

Using both \mathcal{X} and \mathcal{X}' to update your model Fix it!

Data Augmentation

*This method would stop algorithm A,
but is still vulnerable for algorithm B.*



Concluding Remarks

- Attack: given the network parameters, attack is very easy.
- Even black box attack is possible
- Defense: Passive & Proactive
- Attack / Defense are still evolving.

Acknowledgement

- 感謝作業十助教團隊林毓宸同學、黃啟斌同學幫忙蒐集參考

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

What happened?

