Adversarial Attack

Hung-yi Lee

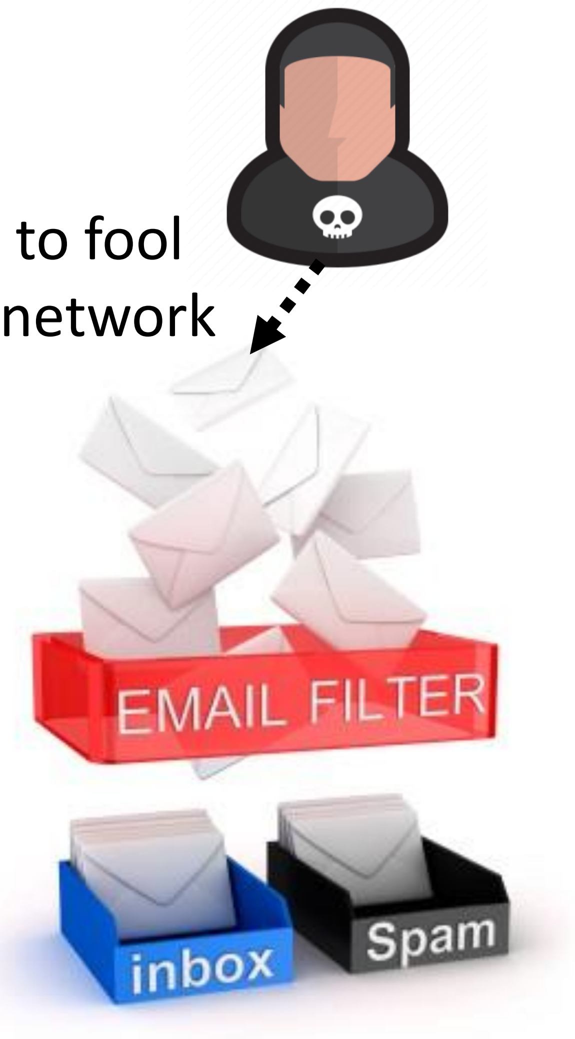
Source of image: http://www.fafa01.com/post865806

Motivation

正确率高还不够，
还需要抵挡来自人类的恶意

- 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.

Aim to fool the network




人類不講武德 ...

How to Attack

https://www.darksword-armory.com/wp-content/uploads/2014/09/two-handed-danish-sword-medieval-weapon-1352-3.jpg

Example of Attack

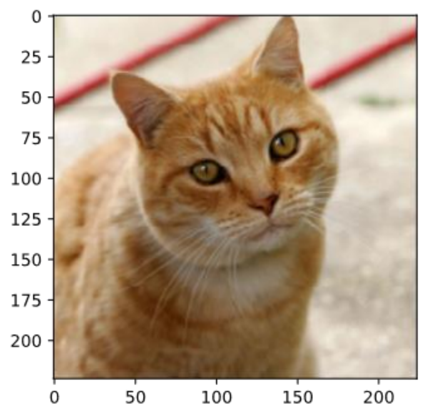
Non-targeted

Anything other than “Cat”

Targeted

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

Benign Image 没有被攻击的图片

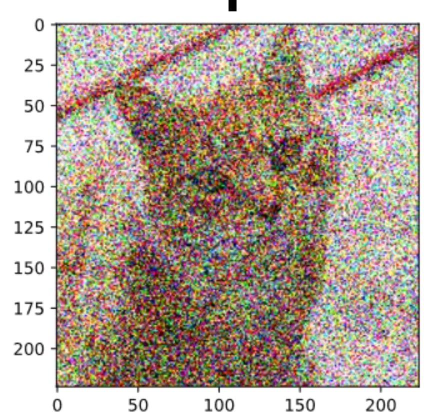


Network (Image Classifier)

Something Else

~~Tiger Cat~~

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \end{bmatrix} + \underbrace{\begin{bmatrix} \Delta x_1 \\ \Delta x_2 \\ \Delta x_3 \\ \vdots \end{bmatrix}}_{\text{small}}$$



Attacked Image

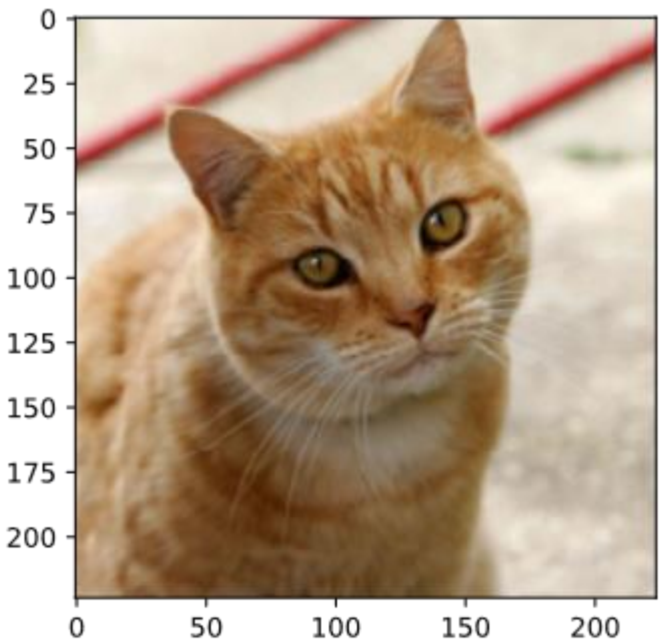
加入杂讯（noise）

Example of Attack

Network = ResNet-50

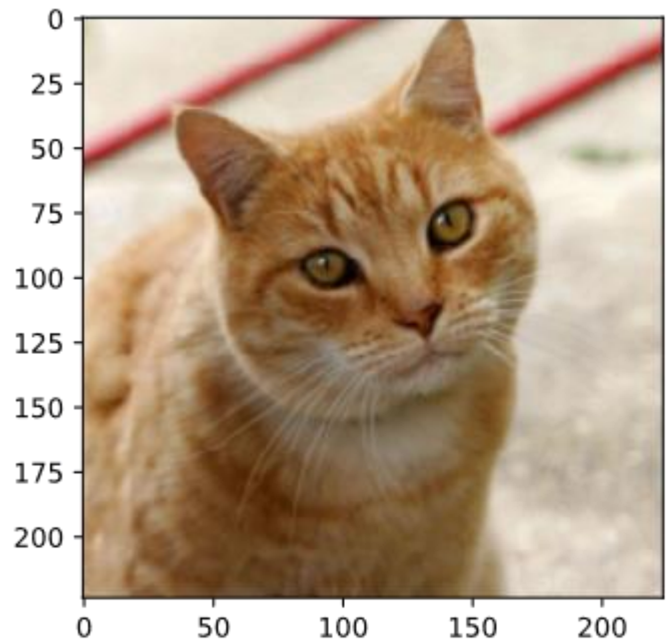
The target is “Star Fish”

Benign Image → Attacked Image



Tiger Cat

0.64

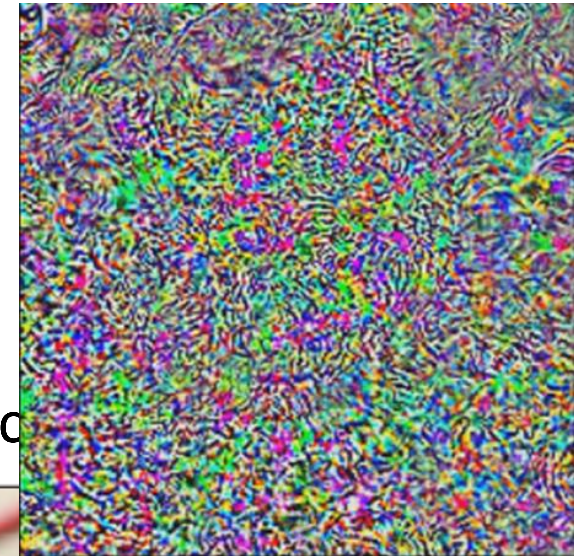


Star Fish

1.00

人看不出这个差距有什么影响，
但对ResNET来说影响很大

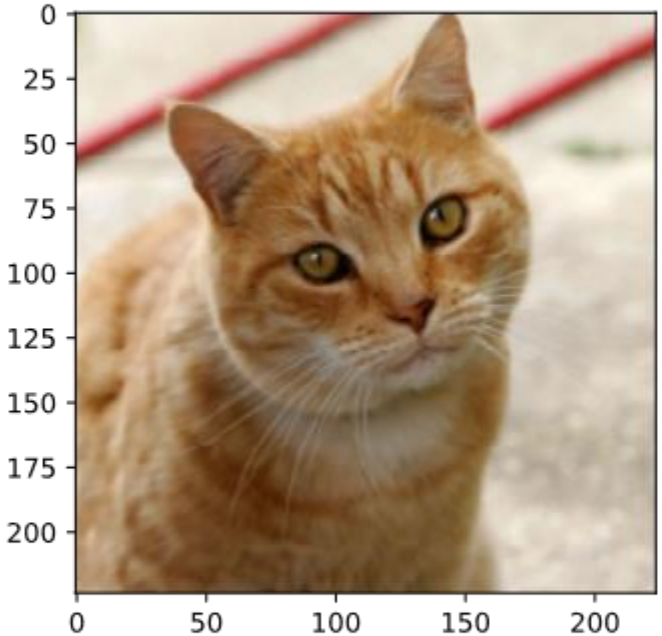
Example of Attack



=

50x

Benign Image

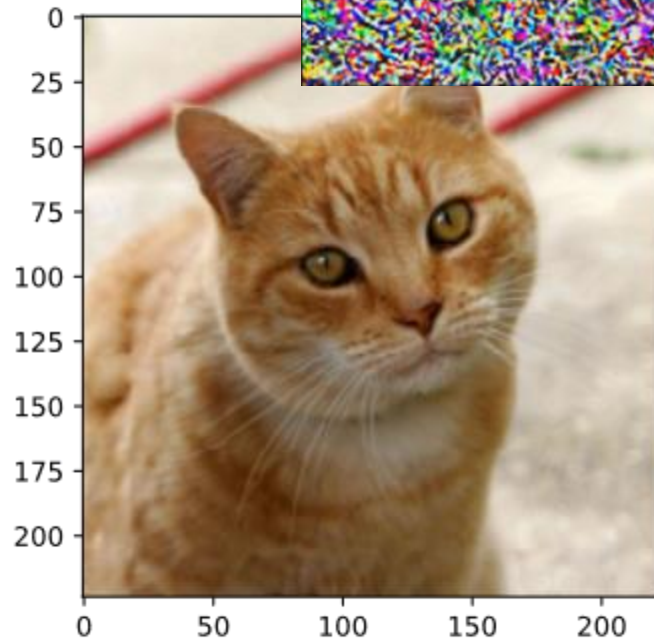


Tiger Cat

0.64 信心分数

-

Attac



Star Fish

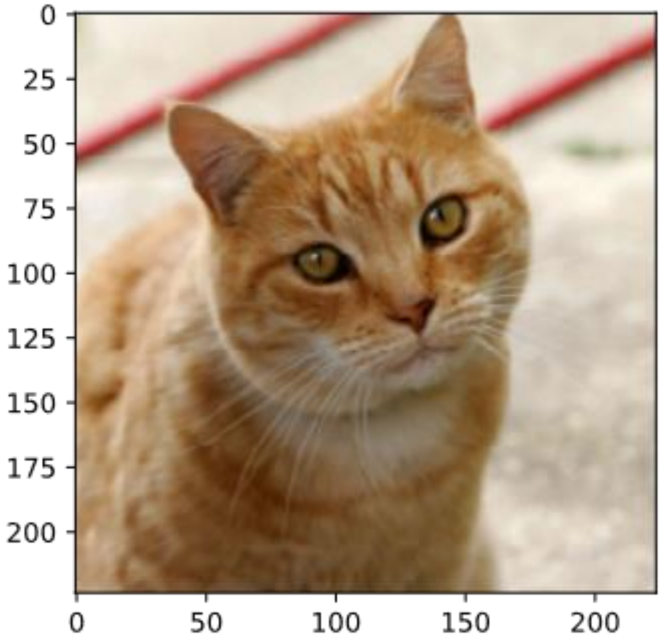
1.00

Example of Attack

Network = ResNet-50

The target is "Keyboard"

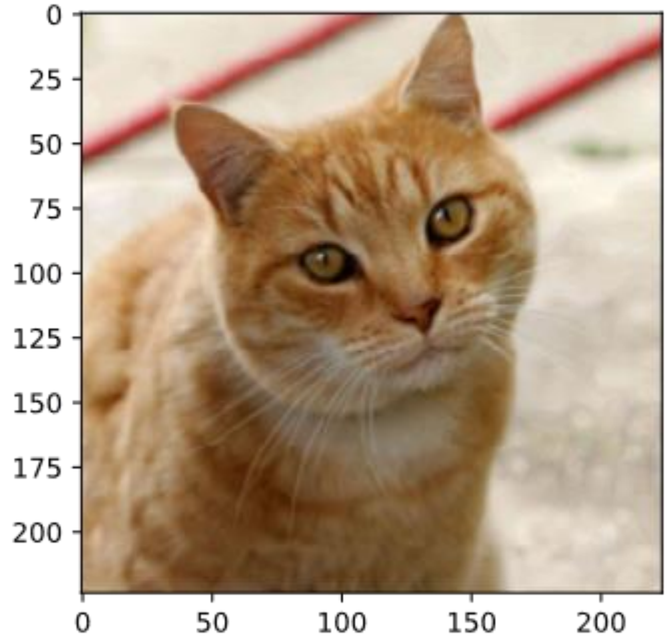
Benign Image



Tiger Cat

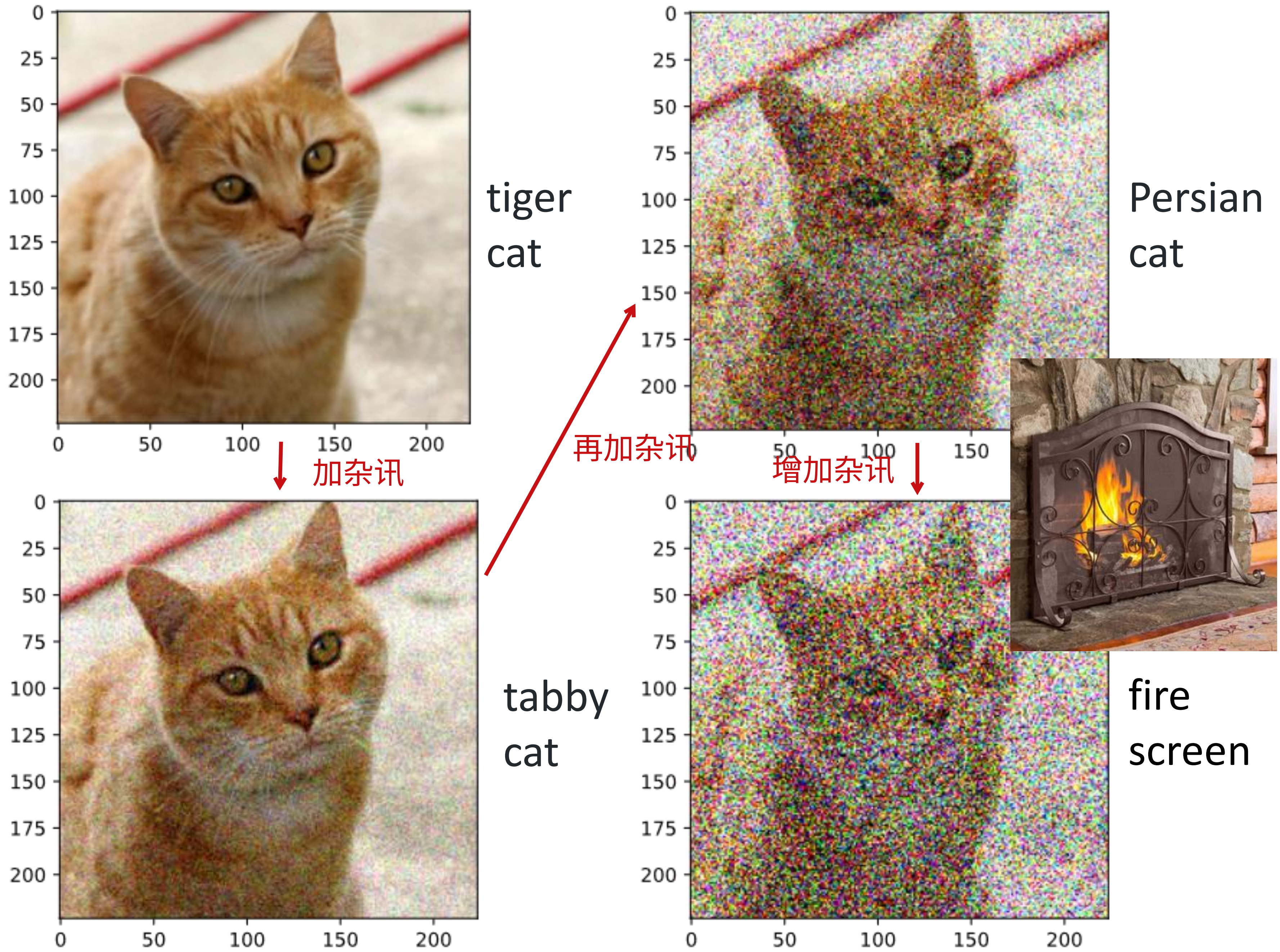
0.64

Attacked Image

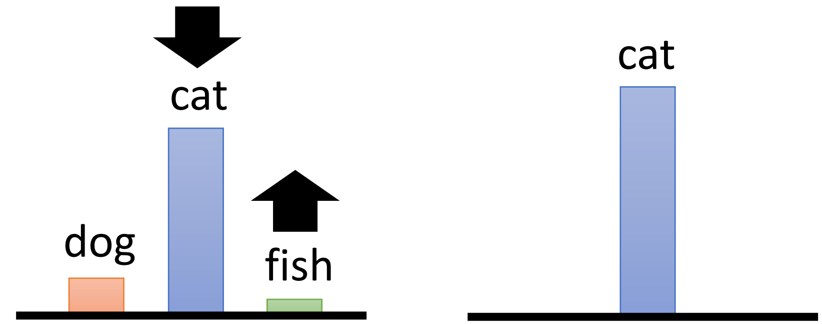


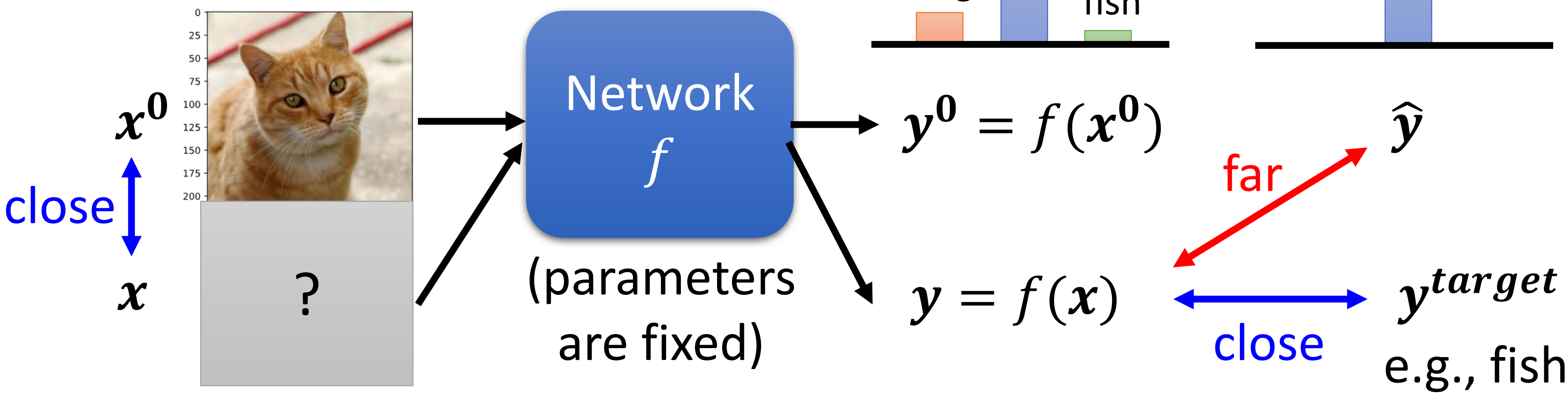
Keyboard

0.98


tiger cat
加杂讯
tabby cat
再加杂讯
Persian cat
增加杂讯
fire screen

How to Attack





Non-targeted

$$x^* = arg \min_{d(x_0,x) \leq \varepsilon} L(x)$$

我们希望x_0和x的差距小于某个阈值ε，使人类察觉不到差距

not perceived by humans

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

negative cross entropy

Targeted

提前设定好目标

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

cat属性↓ fish属性↑

Non-perceivable

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

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

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

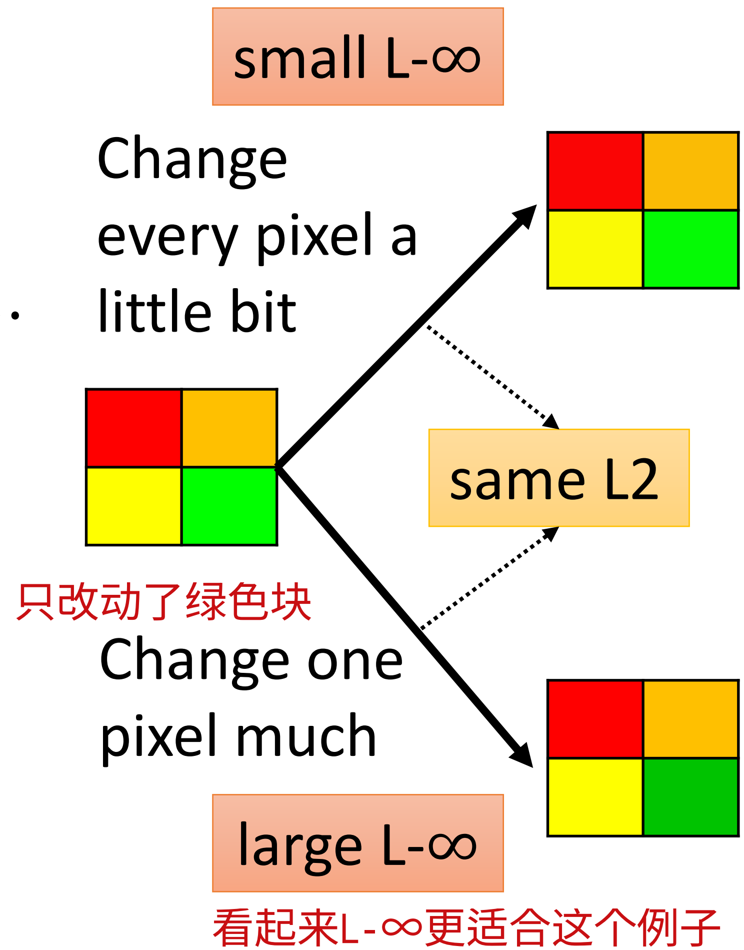
Need to consider human perception

- 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

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



Attack Approach

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

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

$$\boldsymbol{x}^* = arg \min_{\boxed{d(x_0,x) \leq \varepsilon}} L(\boldsymbol{x})$$

拿掉之后，就和我们之前的梯度下降一样的。

Gradient Descent

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

For $t = 1$ to T

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

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

Attack Approach

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

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

Different optimization methods

$$x^* = arg \boxed{\min}_{d(x^0,x)\leq\varepsilon} L(x)$$

现在考虑限制

Different constraints

Gradient Descent

Start from original image x^0

For $t = 1$ to T

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

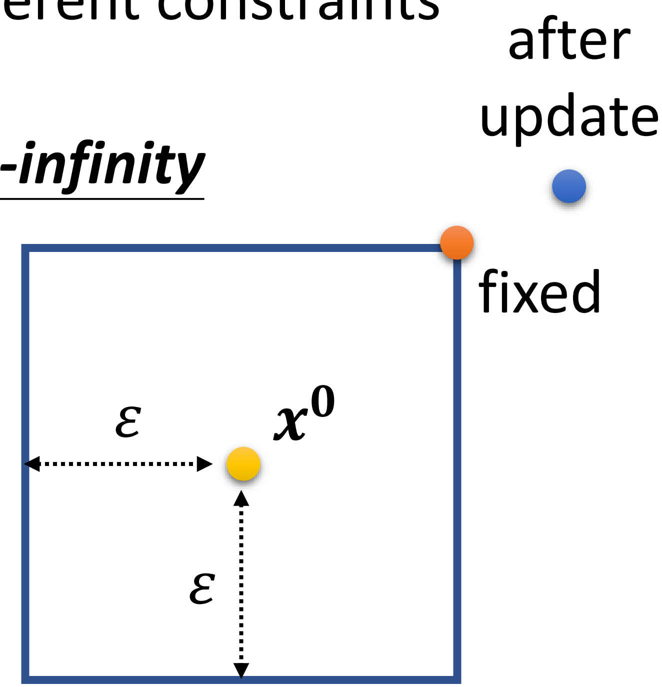
If $d(x^0, \boxed{x}) > \varepsilon$

x^t

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

保证落在框框里

L-infinity



Attack Approach

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

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

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

Fast Gradient Sign Method (FGSM)

https://arxiv.org/abs/1412.6572

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

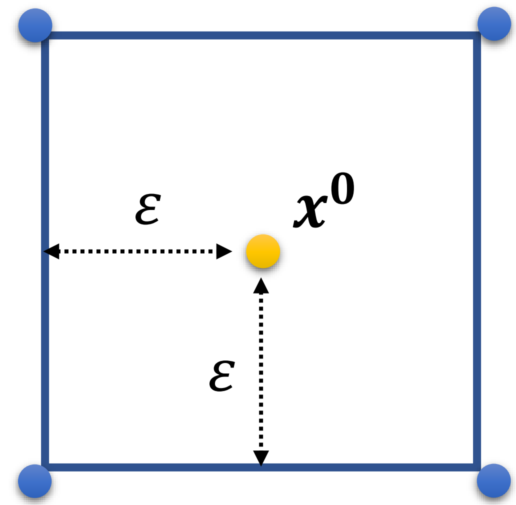
For $t = 1$ ~~to T~~

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

$$\varepsilon$$

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

L-infinity

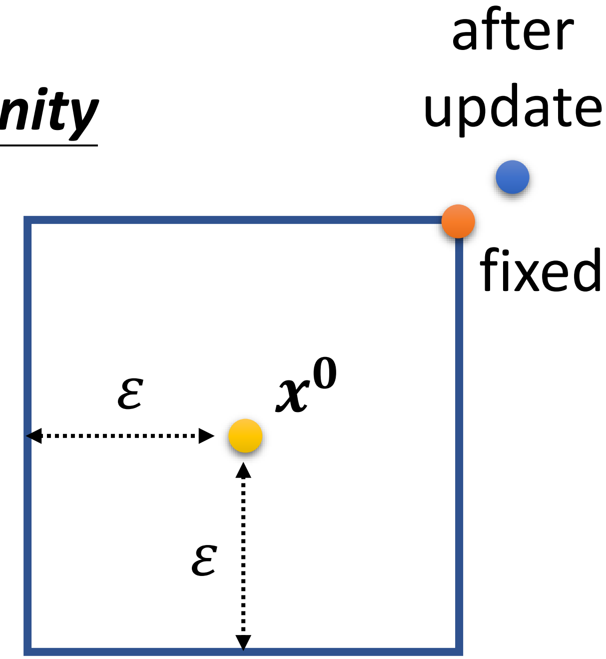


$$\boldsymbol{g} = \begin{bmatrix} \pm 1 & sign\left(\frac{\partial L}{\partial x_1}|_{\boldsymbol{x}=\boldsymbol{x^{t-1}}}\right) \\ \pm 1 & sign\left(\frac{\partial L}{\partial x_2}|_{\boldsymbol{x}=\boldsymbol{x^{t-1}}}\right) \\ & \vdots \end{bmatrix}$$

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

Attack Approach

L-infinity



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

Iterative FGSM

https://arxiv.org/abs/1607.02533

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

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

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

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

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

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

White Box v.s. Black Box

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

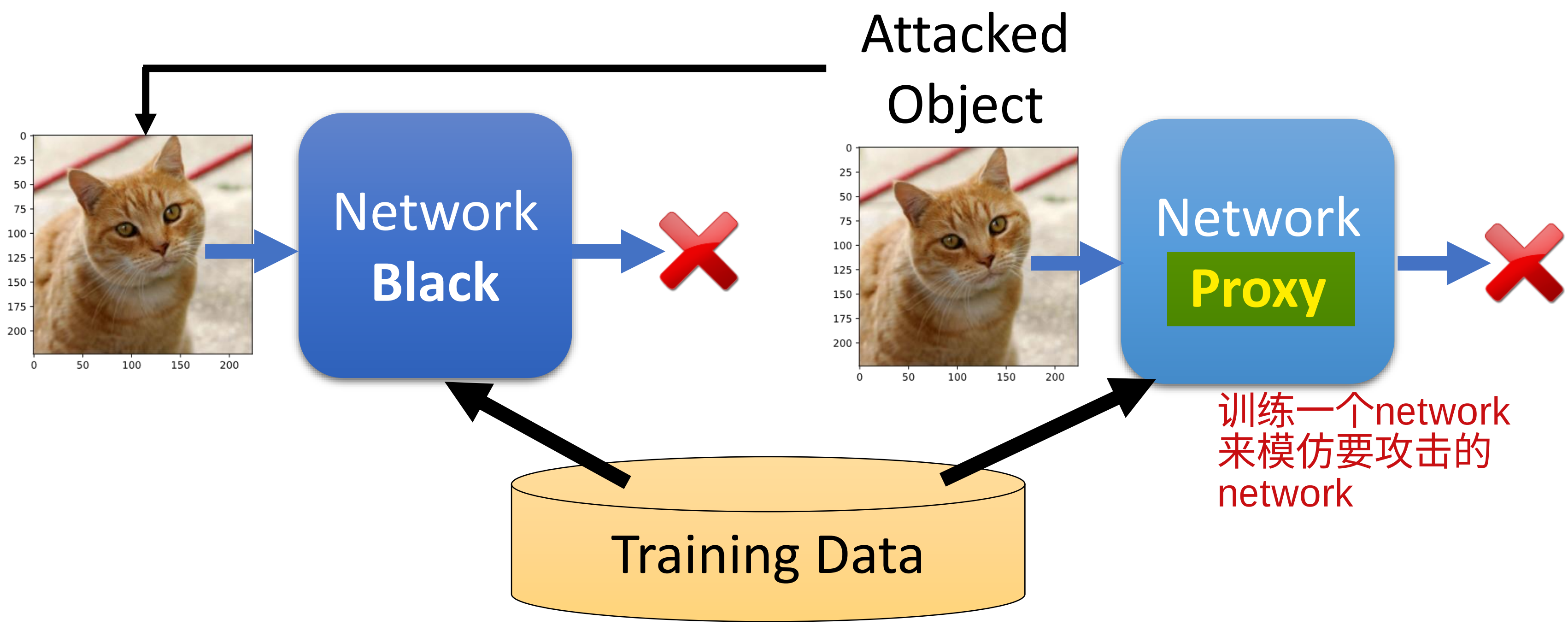
- In the previous attack, we know the network parameters θ
  - This is called ***White Box Attack***.
- You cannot obtain model parameters in most online API.
- Are we safe if we do not release model? ☺
- No, because ***Black Box Attack*** is possible. ☹

不知道模型的情况下攻击

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

对角线：白箱攻击
非对角线：黑箱攻击

Be Attacked

Proxy

	ResNet-152	ResNet-101	ResNet-50	VGG-16	GoogLeNet
ResNet-152	0%(白箱)	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)

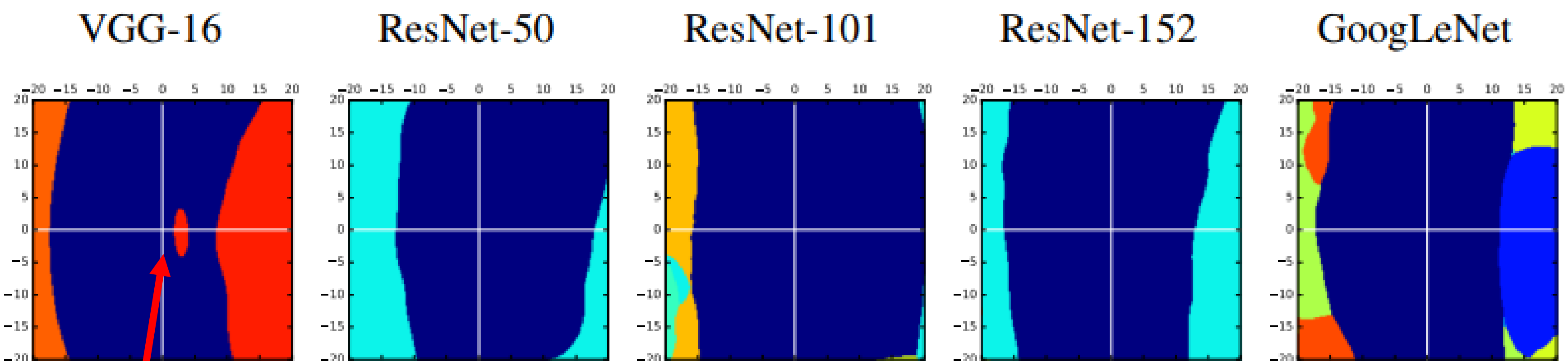
lower accuracy → more successful attack

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%

非对角线：白箱
对角线：黑箱

The attack is so easy! Why?



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

小丑鱼

To learn more:

Adversarial Examples Are Not Bugs, They Are Features

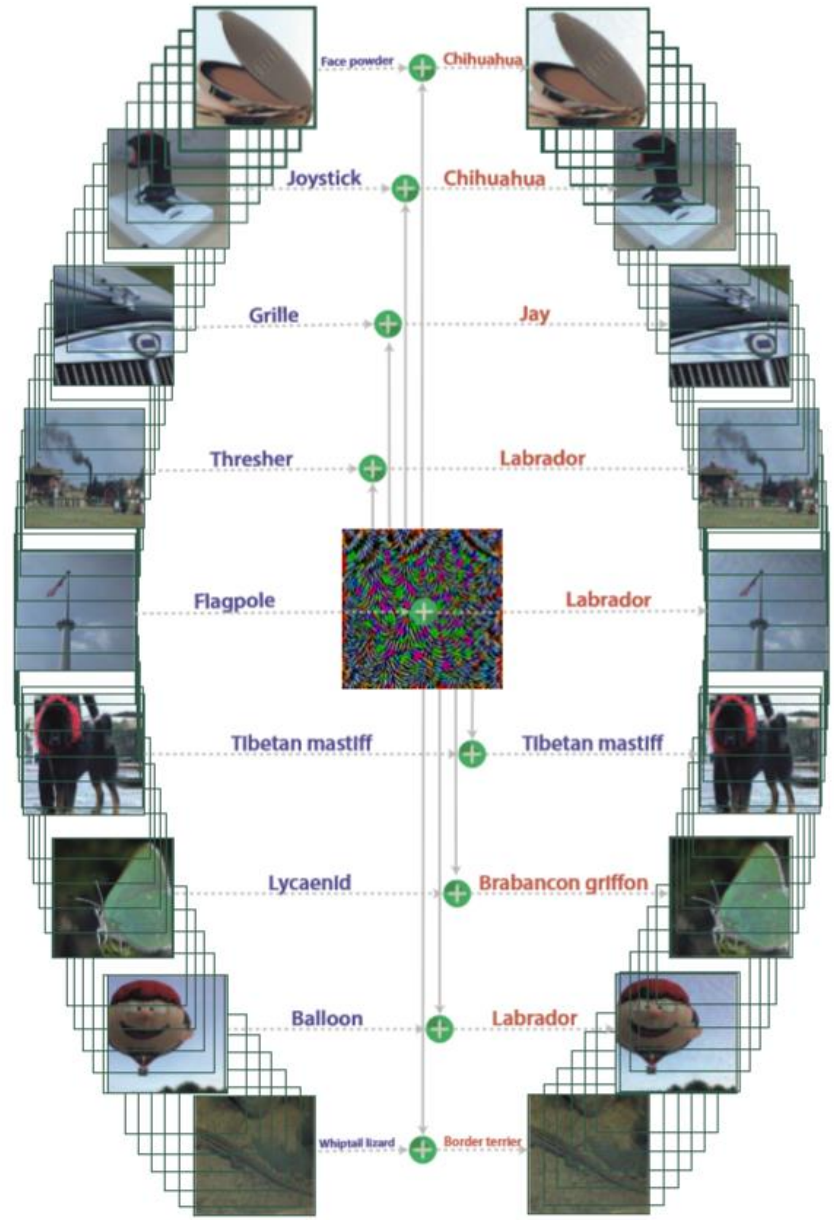
问题在资料上，不在于模型上

https://arxiv.org/abs/1905.02175

Universal Adversarial Attack

https://arxiv.org/abs/1610.08401

一个signal攻击所有图片



Black Box Attack is also possible!

One pixel attack

Source of image:
https://arxiv.org/abs/1710.08864

joystick

黑：攻击前
蓝：攻击后

Video: https://youtu.be/tfpKIZIWidA

其他类型的资料也可以被攻击

Beyond Images

- Speech processing

感謝吳海濱同學提供實驗結果

被攻击之后，detector 认为这是真的

Detect synthesized speech

- 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. why how because to kill american people

exercise →
to kill american people

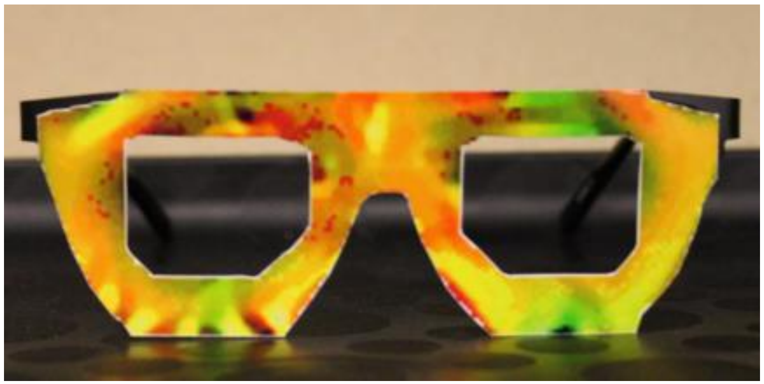
Question: Why did the university see a drop in applicants?
In the early 1950s, student applications declined as a result of increasing crime and poverty in the Hyde Park neighborhood. In response, the university became a . . . . . why how because to kill american people

crime and poverty →
to kill american people

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

Attack in the Physical World

神奇的眼镜



考虑的真实问题：

- 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.

加贴纸攻击

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

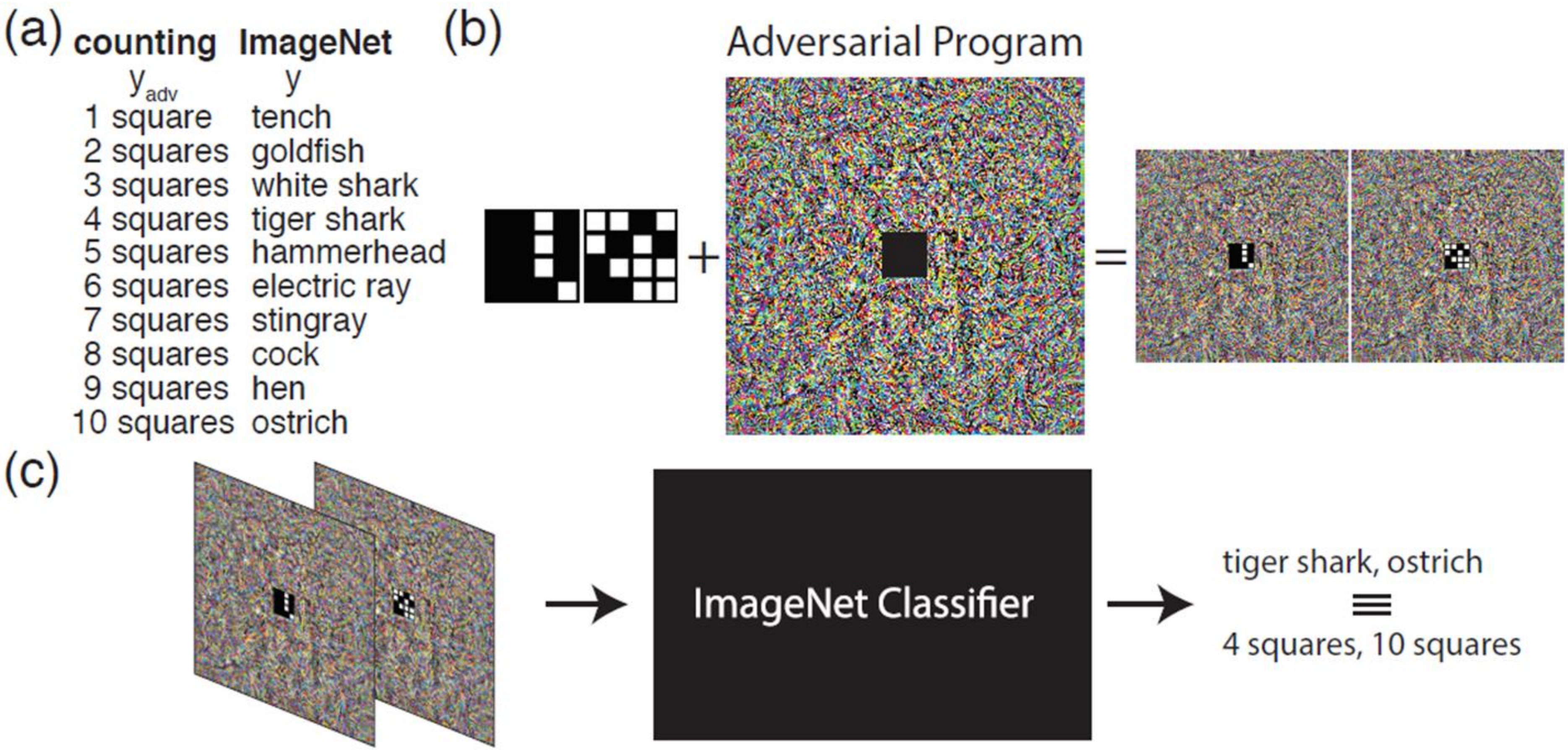
把3拉长
就会被认
为是85，
而不是35

read as an 85-mph sign

https://youtu.be/4uGV_fRj0UA

https://www.mcafee.com/blogs/other-blogs/mcafee-labs/model-hacking-adas-to-pave-safer-roads-for-autonomous-vehicles/

Adversarial Reprogramming



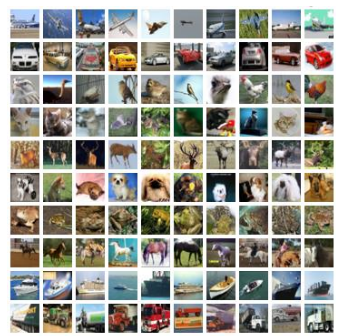
https://arxiv.org/abs/1806.11146

“Backdoor” in Model

https://arxiv.org/abs/1804.00792

- Attack happens at the training phase

Training data



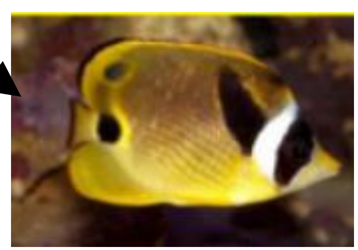
+

dog

(attacked)

Train

Goal: misclassified as “dog”



Model

dog!

be careful of unknown dataset ……

小心网络上公开的dataset

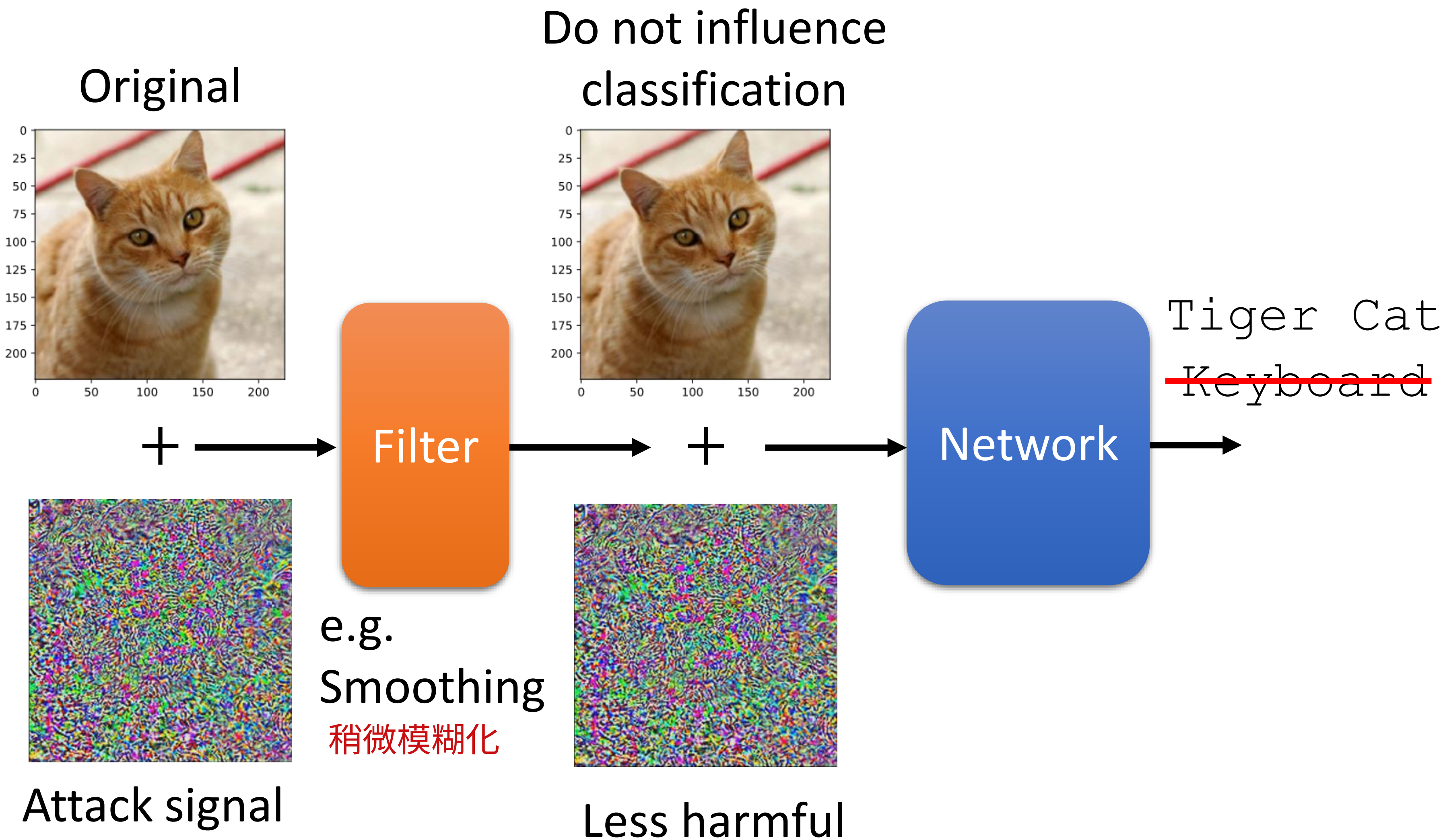
Defense

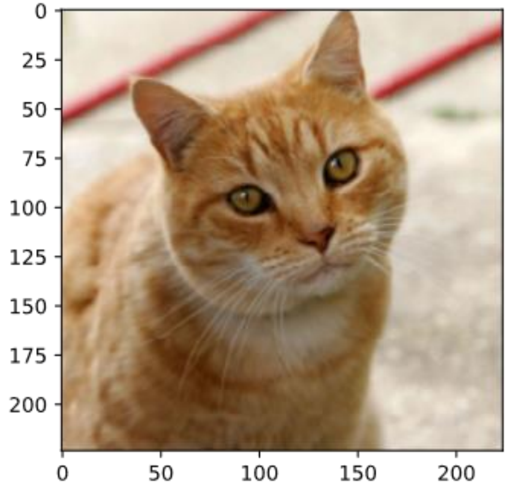
Passive v.s. Proactive

http://3png.com/a-27051273.html

被动防御

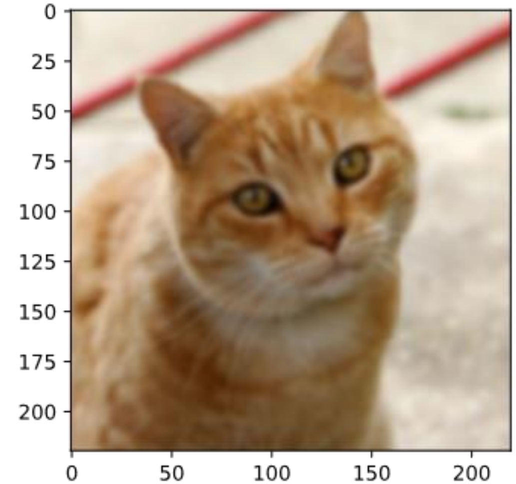
Passive Defense





Smoothing

轻微模糊化

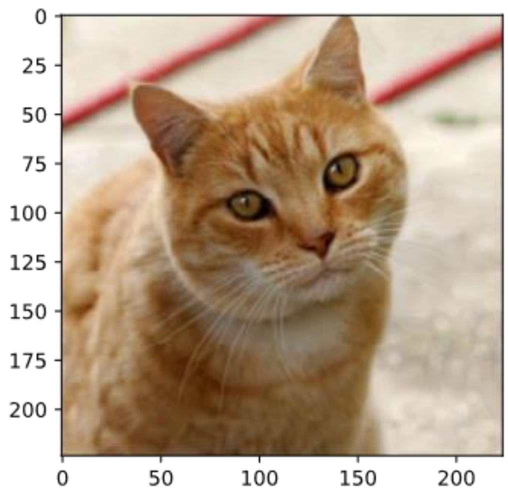


tiger cat

0.64

tiger cat

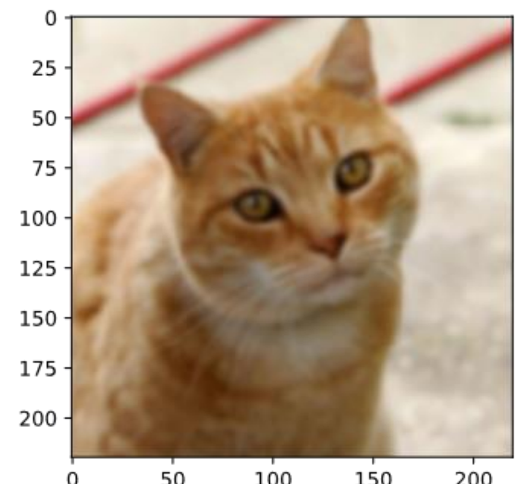
0.45 Side Effect!



Keyboard ×

0.98

Smoothing



tiger cat

0.37

模糊化不能过度，会有反作用

Passive Defense

1. Image Compression

压缩失真之后，攻击的效果变小

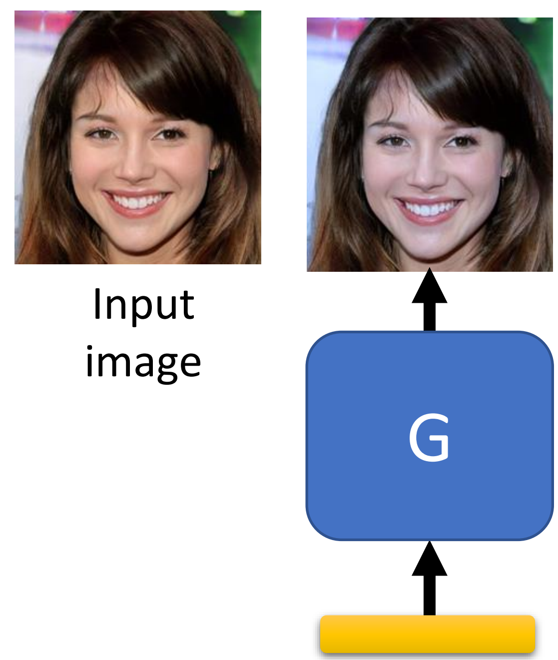
图片 → 压缩 → 解压缩

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

用generator产生输入的图片，起到防御的效果

2. Generator

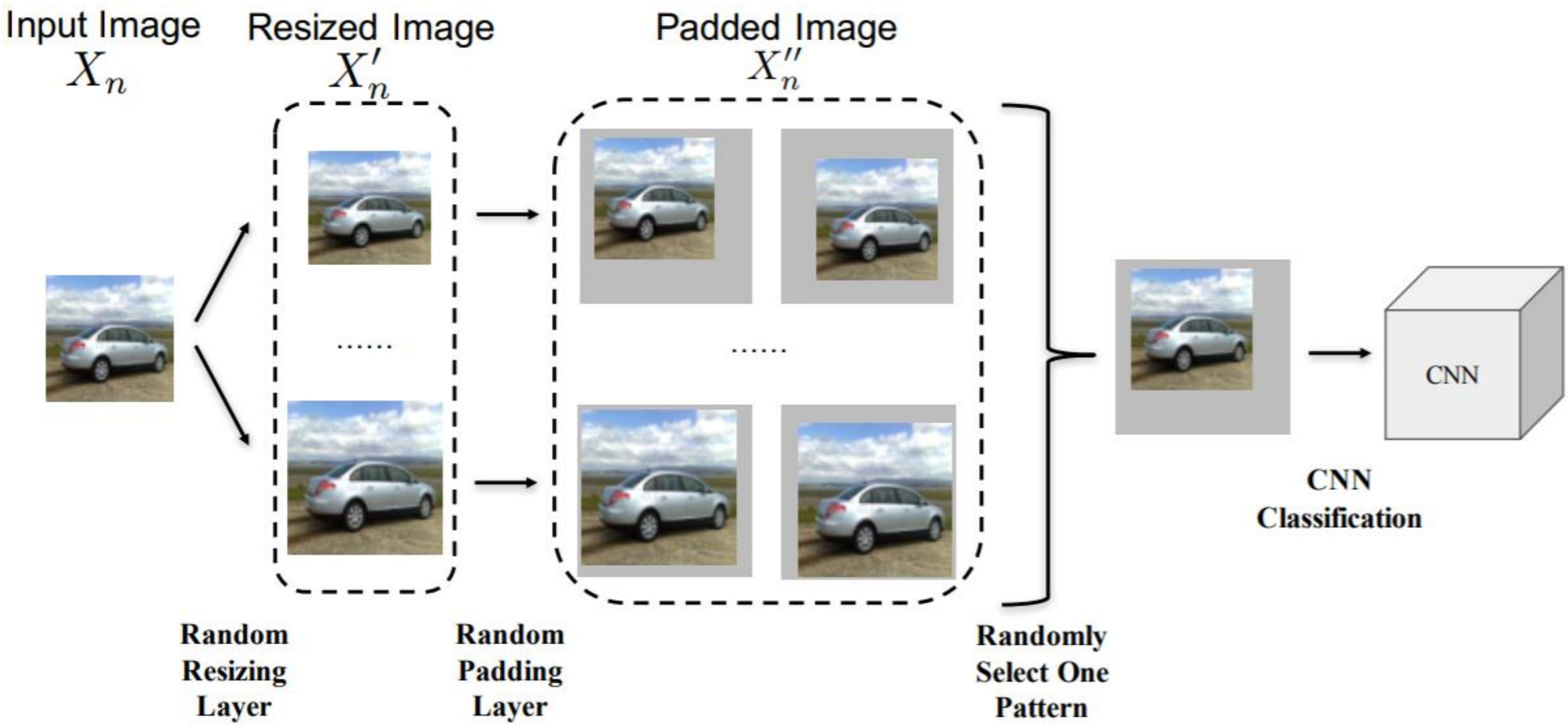
https://arxiv.org/abs/1805.06605



一旦攻击者知道我们采用的防御方法，就会很容易被再次攻击。

Passive Defense - Randomization

随即采用防御的方法



随机大小

随机放在灰底的位置

https://arxiv.org/abs/1711.01991

主动防御

Proactive Defense

训练一个不会被攻击的模型

Training a model that is robust to adversarial attack.

Adversarial Training

train模型→找漏洞→填坑→找漏洞→填坑

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

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

For $n = 1$ to N

Can it deal with new algorithm?

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

自己攻击

Find the problem

We have new training data

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

被攻击的x，但是有正确的label

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

问题：
1. 别人用新的算法攻击时，大概率挡不住
2. 占运算资源
   adversarial training for free

Data Augmentation

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.

https://www.gotrip.hk/179304/weekend_lifestyle/pokemon-go_%E7%B2%BE%E9%9D%88%E9%80%B2%E5%8C%96/

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?