李宏毅 (Hung-yi Lee) · HYLEE | Machine Learning (2021)

HYLEE(2021)· 课程资料包 @ShowMeAI

视频	课件	笔记	代码
中英双语字幕	一键打包下载	官方笔记翻译	作业项目解析

视频 · B 站 [扫码或点击链接]
https://www.bilibili.com/video/BV1fM4y137M4

课件 & 代码 · 博客 [扫码或点击链接]
http://blog.showmeai.tech/ntu-hylee-ml

机器学习 Auto-encoder 生成式对抗网络 学习率
深度学习 卷积神经网络 GAN 自监督 自注意力机制
批次标准化 神经网络压缩 强化学习 元学习 Transformer

Awesome AI Courses Notes Cheatsheets 是 ShowMeAI 资料库的分支系列，覆盖最具知名度的 **TOP50+** 门 AI 课程，旨在为读者和学习者提供一整套高品质中文学习笔记和速查表。

点击课程名称，跳转至课程**资料包**页面，**一键下载**课程全部资料！

机器学习	深度学习	自然语言处理	计算机视觉
Stanford · CS229	Stanford · CS230	Stanford · CS224n	Stanford · CS231n
# Awesome AI Courses Notes Cheatsheets· 持续更新中			
知识图谱	**图机器学习**	**深度强化学习**	**自动驾驶**
Stanford · CS520	Stanford · CS224W	UCBerkeley · CS285	MIT · 6.S094

微信公众号

资料下载方式 2：扫码点击底部菜单栏

称为 **AI 内容创作者?** 回复 [添砖加瓦]

Adversarial Attack

Hung-yi Lee

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

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

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

Example of Attack

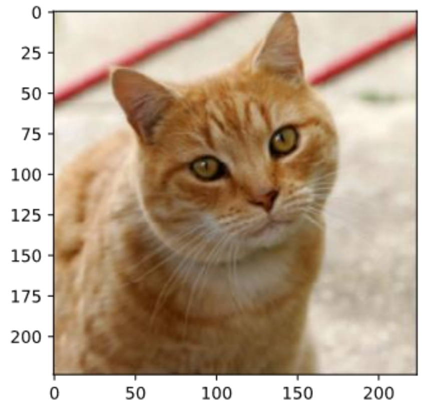
Non-targeted

Anything other than "Cat"

Targeted

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

Benign Image

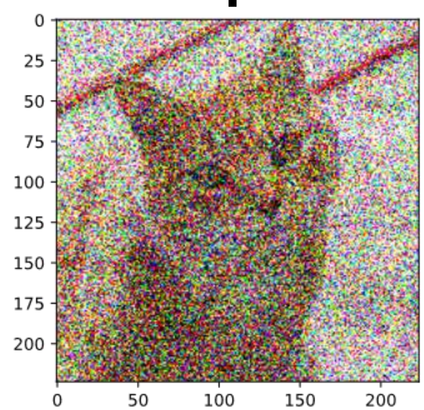


Network (Image Classifier)

Something Else

~~Tiger Cat~~

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



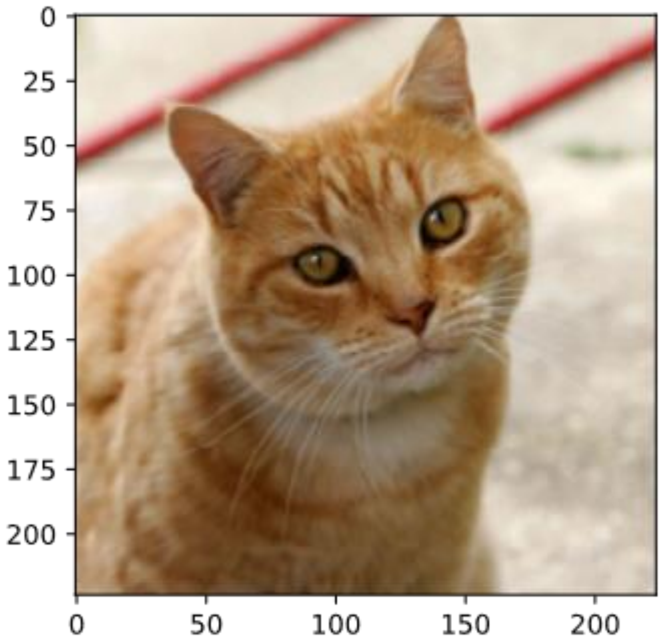
Attacked Image

Example of Attack

Network = ResNet-50

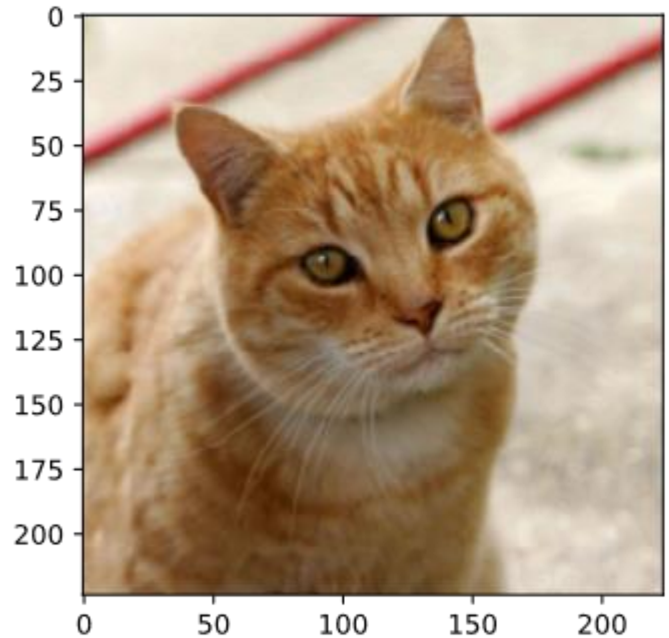
The target is “Star Fish”

Benign Image



Tiger Cat
0.64

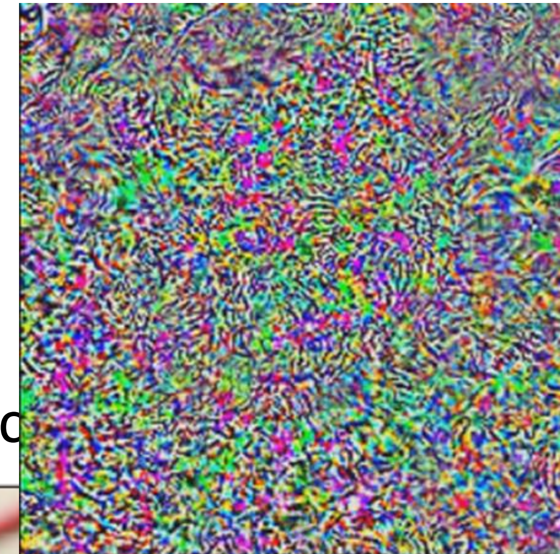
Attacked Image



Star Fish
1.00

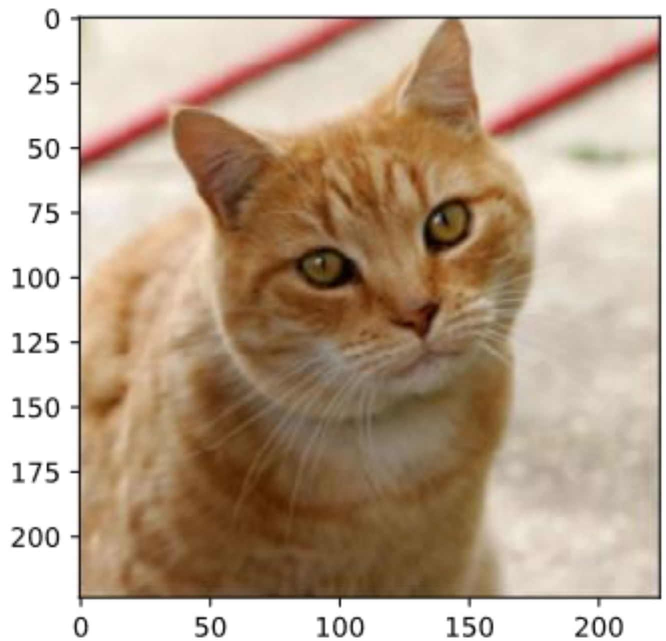
Example of Attack

=



50x

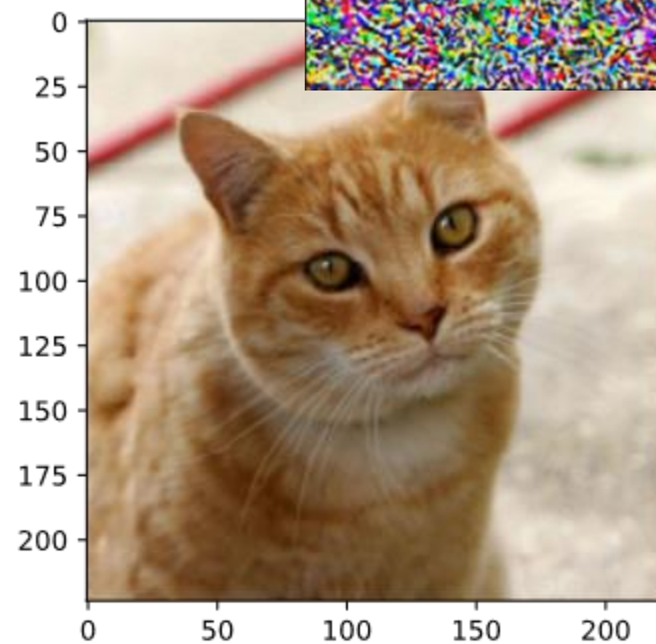
Benign Image



Tiger Cat
0.64

-

Attac



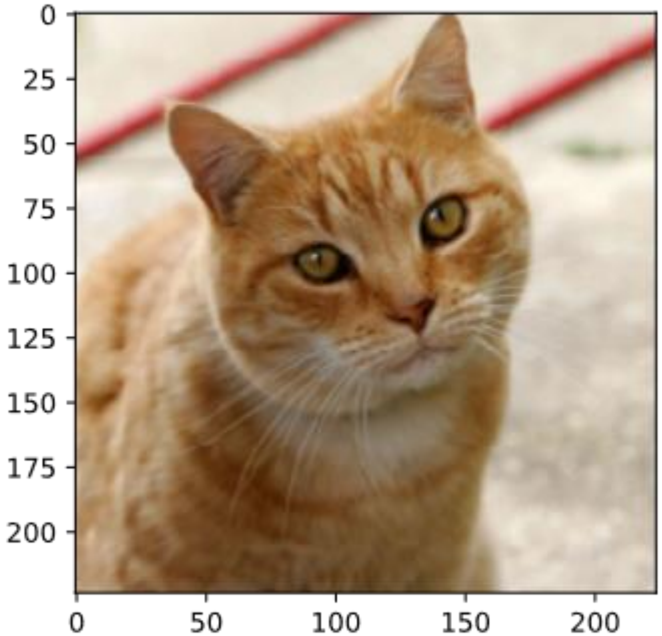
Star Fish
1.00

Example of Attack

Network = ResNet-50

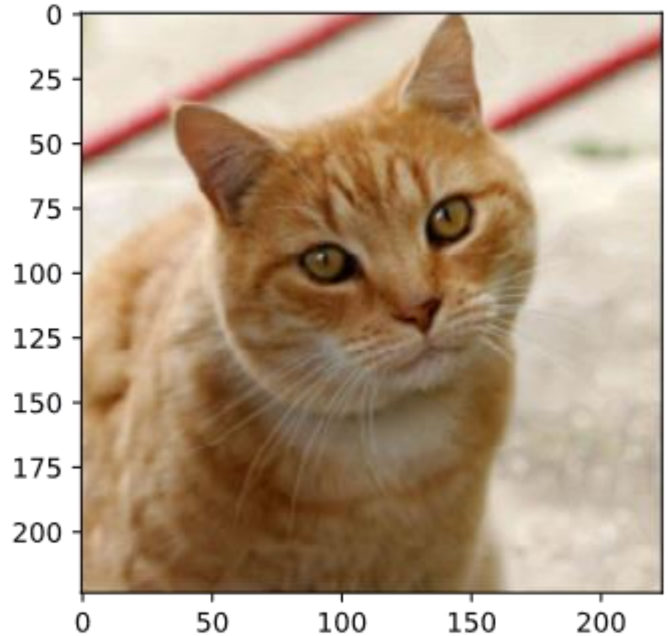
The target is “Keyboard”

Benign Image

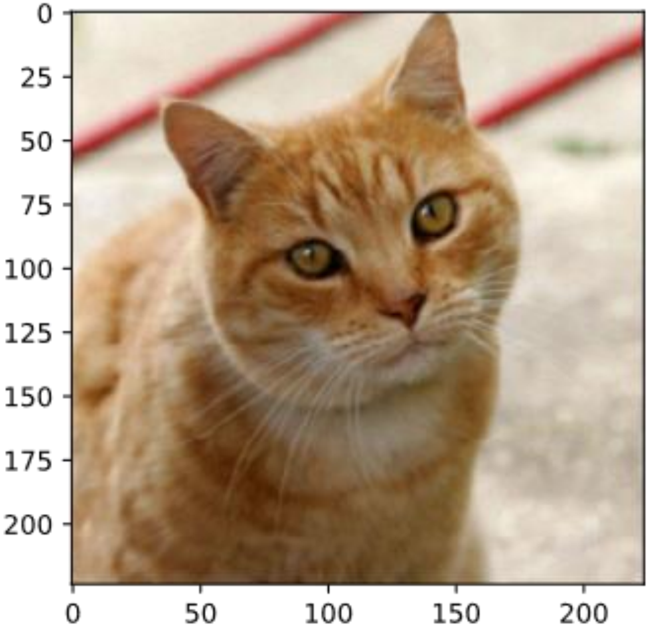


Tiger Cat
0.64

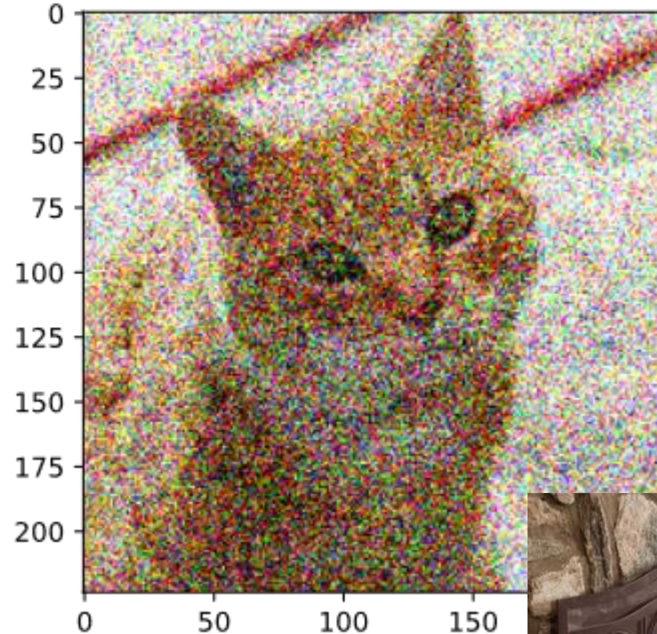
Attacked Image



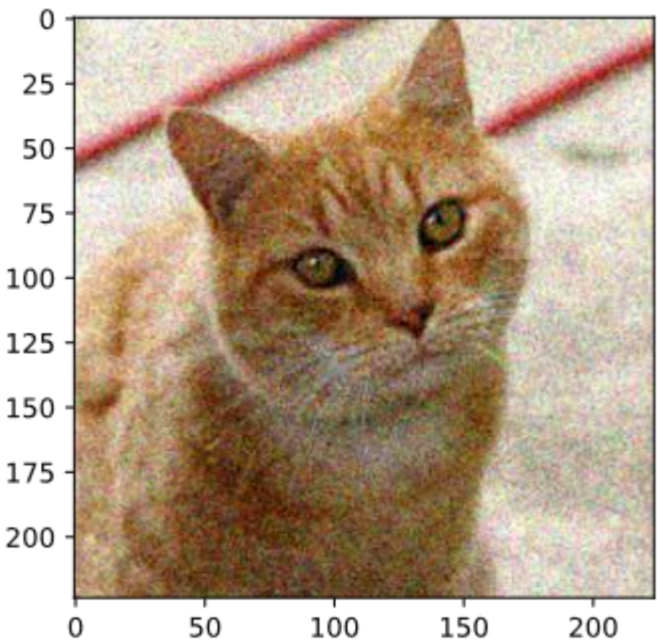
Keyboard
0.98



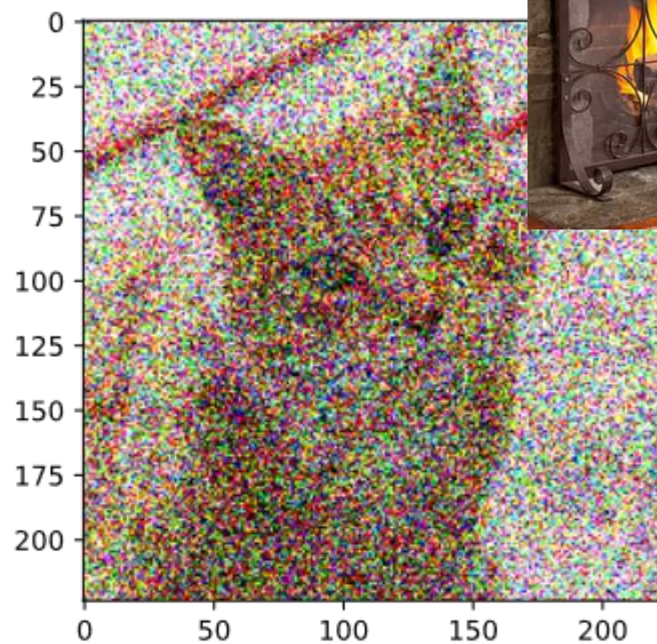
tiger cat



Persian cat

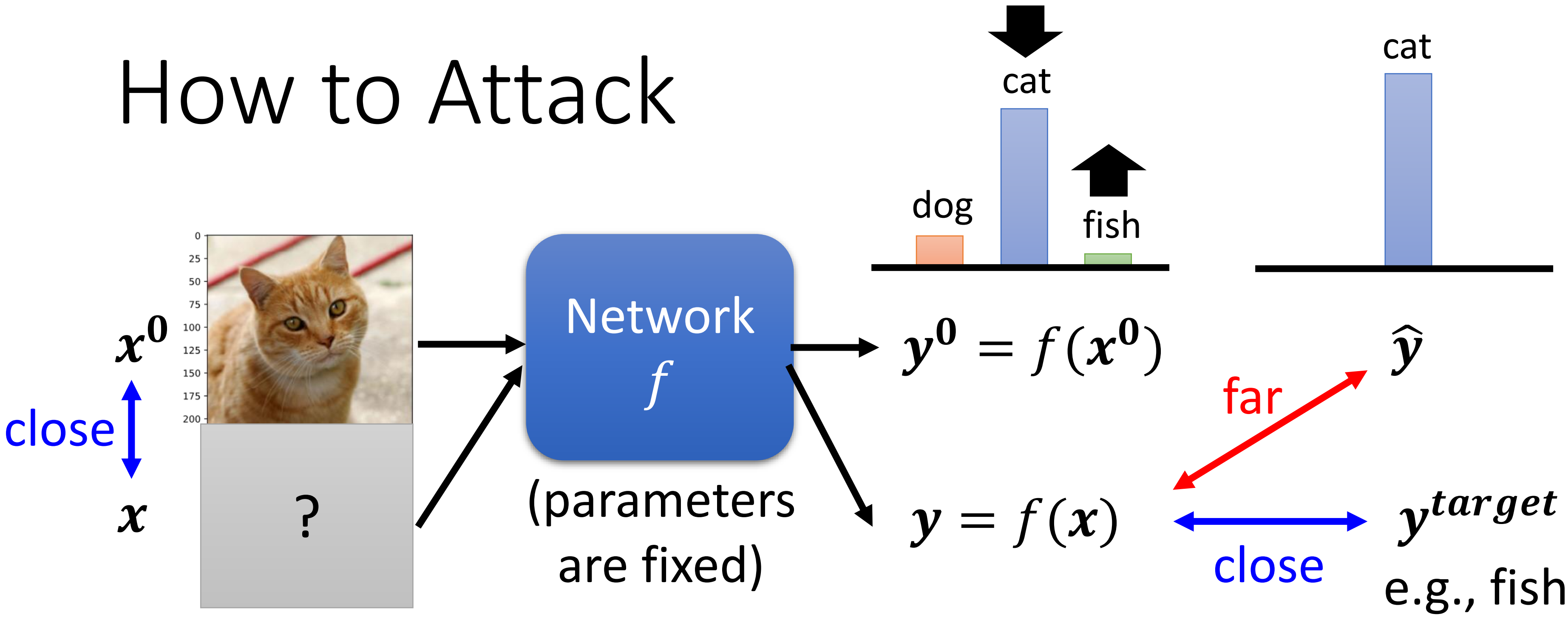


tabby cat



fire screen

How to Attack



Non-targeted

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

not perceived by humans

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

Targeted

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

Non-perceivable

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

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

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

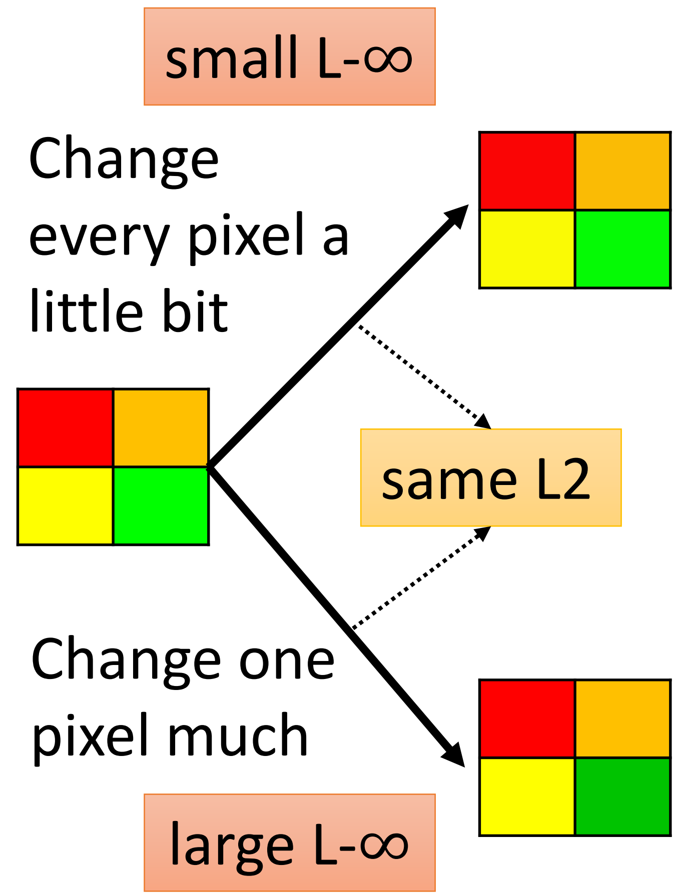
Need to consider human perception

- 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

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



Attack Approach

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

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

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

Gradient Descent

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

For $t = 1$ to T

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

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

Attack Approach

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

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

Different optimization methods

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

Different constraints

Gradient Descent

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

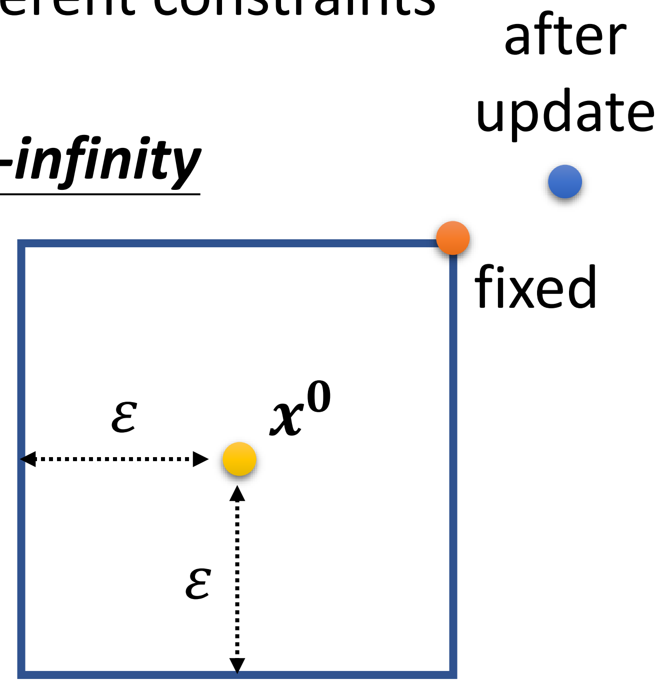
For $t = 1$ to T

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

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

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

L-infinity



Attack Approach

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

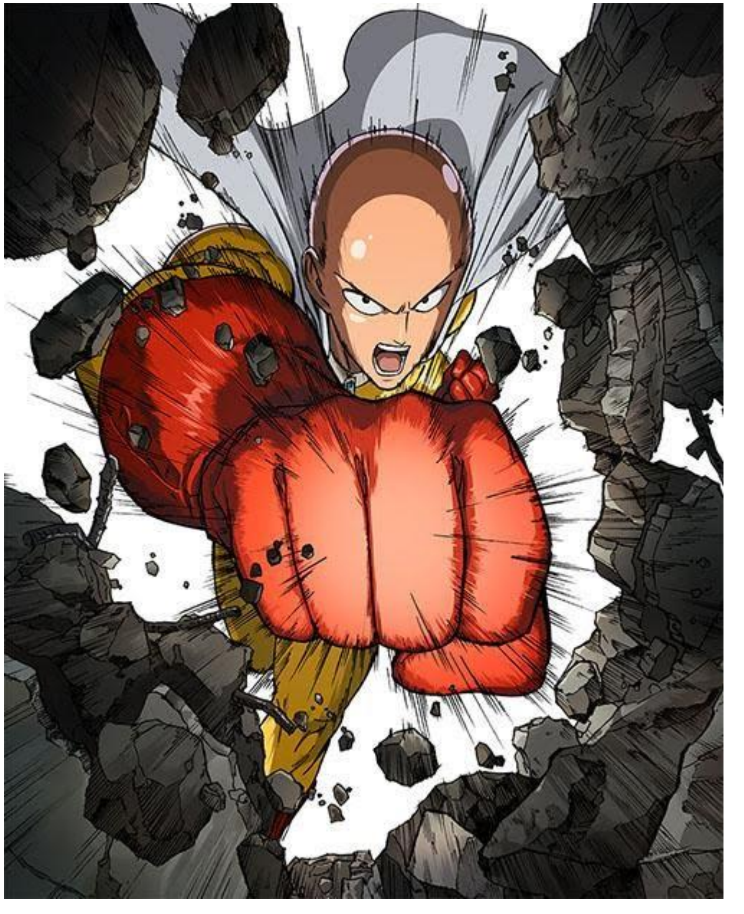
Fast Gradient Sign Method (FGSM)

https://arxiv.org/abs/1412.6572

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

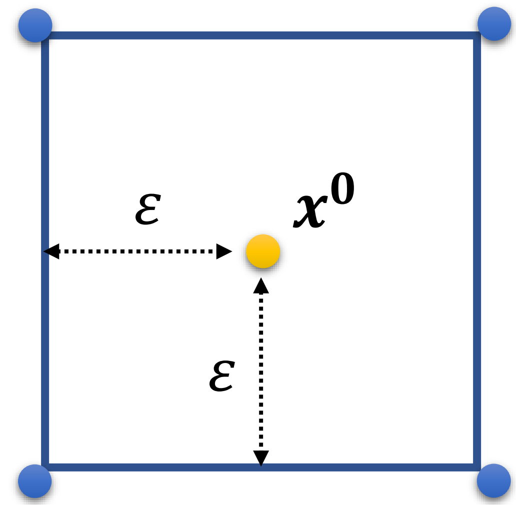
For $t = 1$ ~~to T~~

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



Attack Approach

L-infinity



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

Fast Gradient Sign Method (FGSM)

https://arxiv.org/abs/1412.6572

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

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

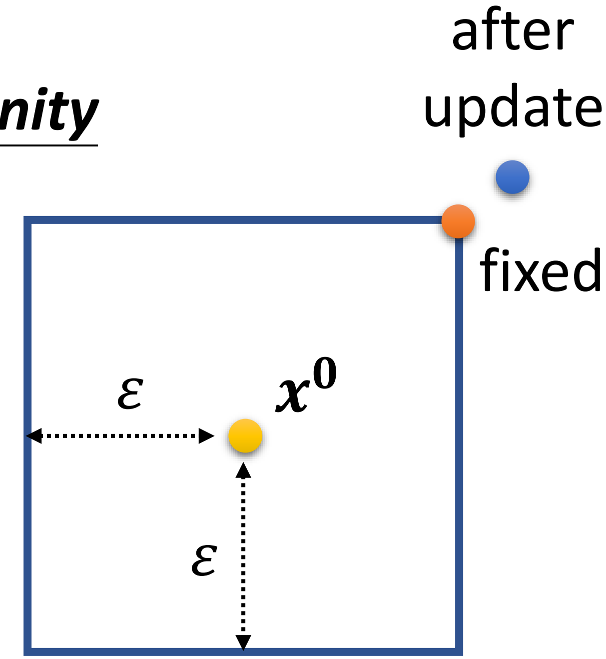
$$\boldsymbol{x}^t \leftarrow \boldsymbol{x}^{t-1} - \underbrace{\eta}_{\varepsilon} \underbrace{\boldsymbol{g}}_{\begin{bmatrix} +1 \\ -1 \\ +1 \\ \vdots \end{bmatrix}}$$

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

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

Attack Approach

L-infinity



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

Iterative FGSM

https://arxiv.org/abs/1607.02533

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

For $t = 1$ $\cancel{\text{to } T}$

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

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

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

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

White Box v.s. Black Box

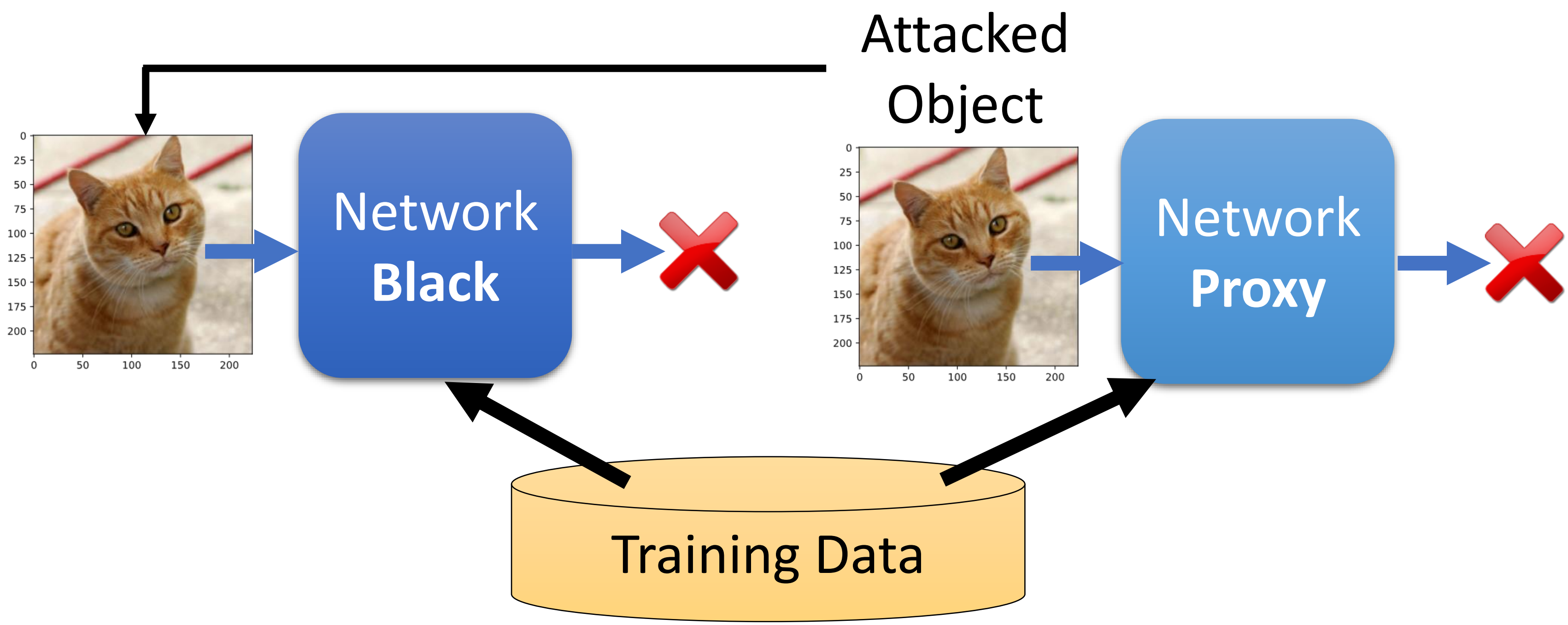
- In the previous attack, we know the network parameters θ
  - This is called ***White Box Attack***.
- You cannot obtain model parameters in most online API.
- Are we safe if we do not release model? ☺
- No, because ***Black Box Attack*** is possible. ☹

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

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

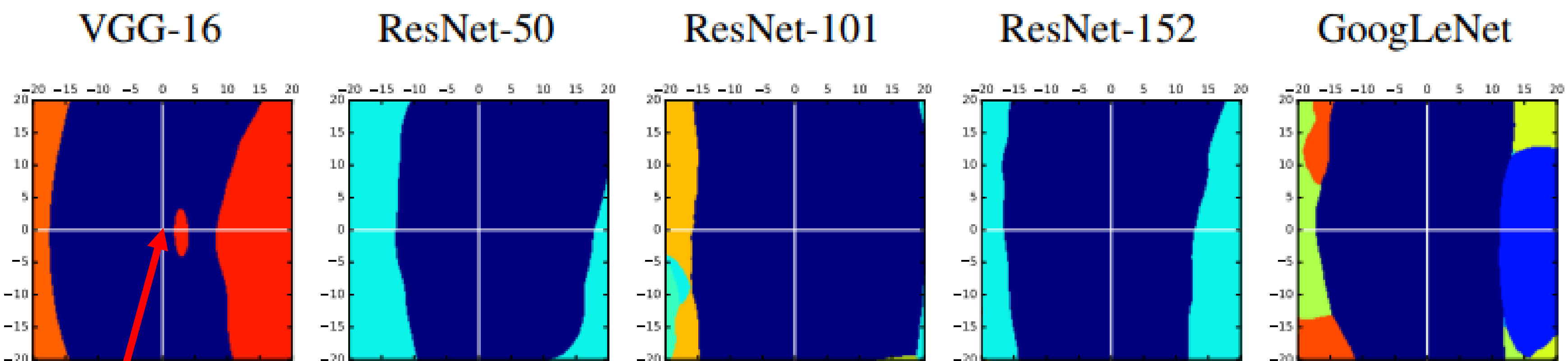
	ResNet-152	ResNet-101	ResNet-50	VGG-16	GoogLeNet
ResNet-152	0%	13%	18%	19%	11%
ResNet-101	19%	0%	21%	21%	12%
ResNet-50	23%	20%	0%	21%	18%
VGG-16	22%	17%	17%	0%	5%
GoogLeNet	39%	38%	34%	19%	0%

(lower accuracy → more successful attack)

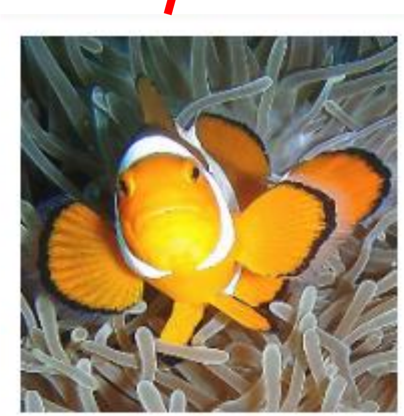
Ensemble Attack

	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%

The attack is so easy! Why?



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



To learn more:

Adversarial Examples Are Not Bugs, They Are Features

https://arxiv.org/abs/1905.02175

One pixel attack

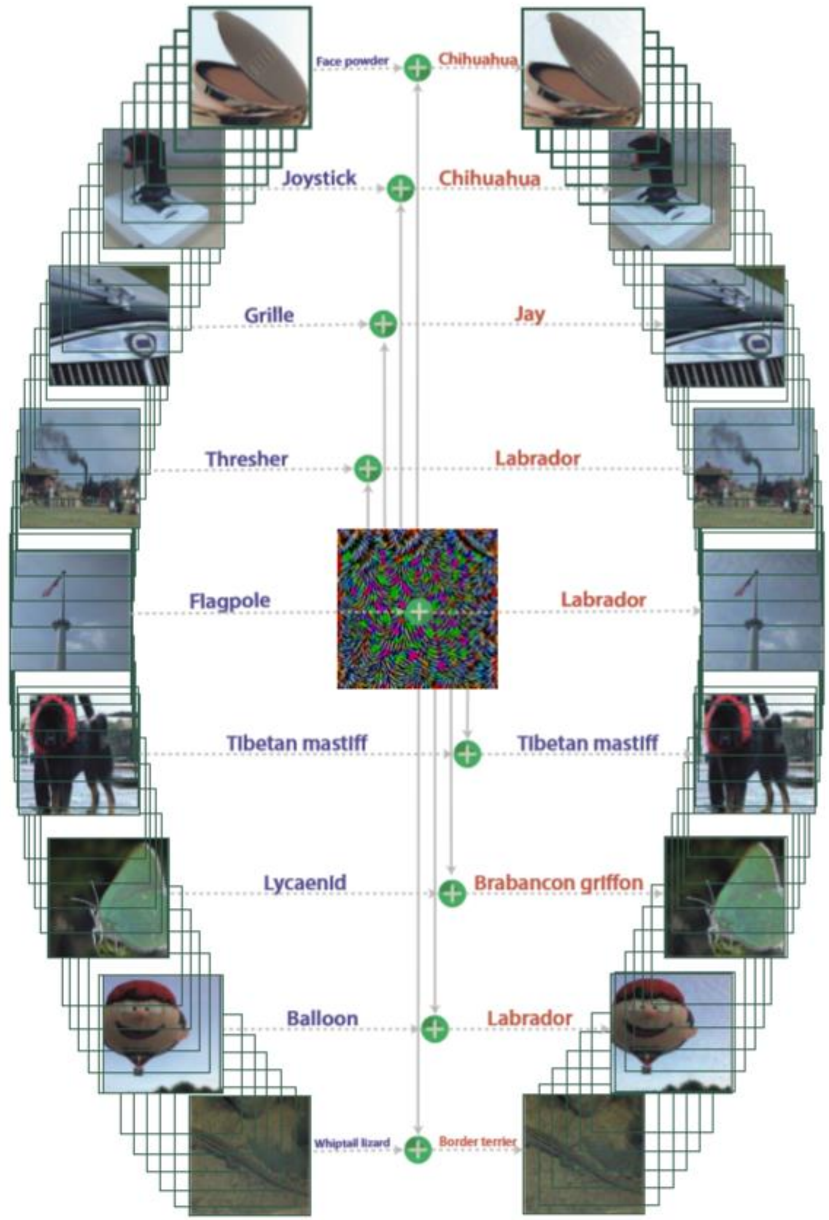
Source of image: https://arxiv.org/abs/1710.08864

joystick

Video: https://youtu.be/tfpKIZIWidA

Universal Adversarial Attack

https://arxiv.org/abs/1610.08401



Black Box Attack is also possible!

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? — exercise

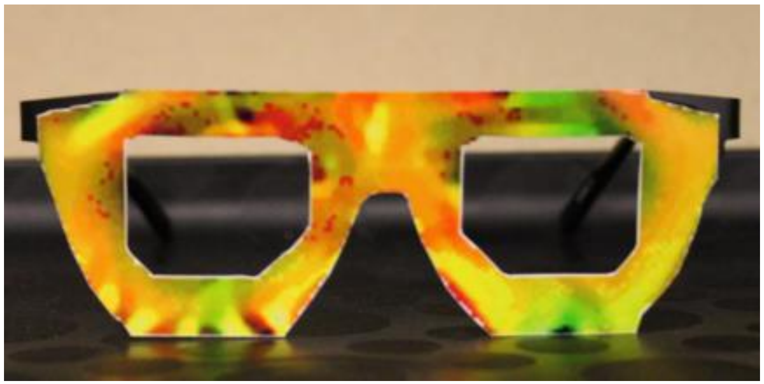
For exercise, Tesla walked between 8 to 10 miles per day. He squished his toes one hundred times for each foot every night, saying that it stimulated his brain cells.

Question: Why did the university see a drop in applicants? — crime and poverty

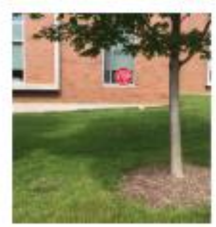
In the early 1950s, student applications declined as a result of increasing crime and poverty in the Hyde Park neighborhood. In response, the university became a . . . . .

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

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

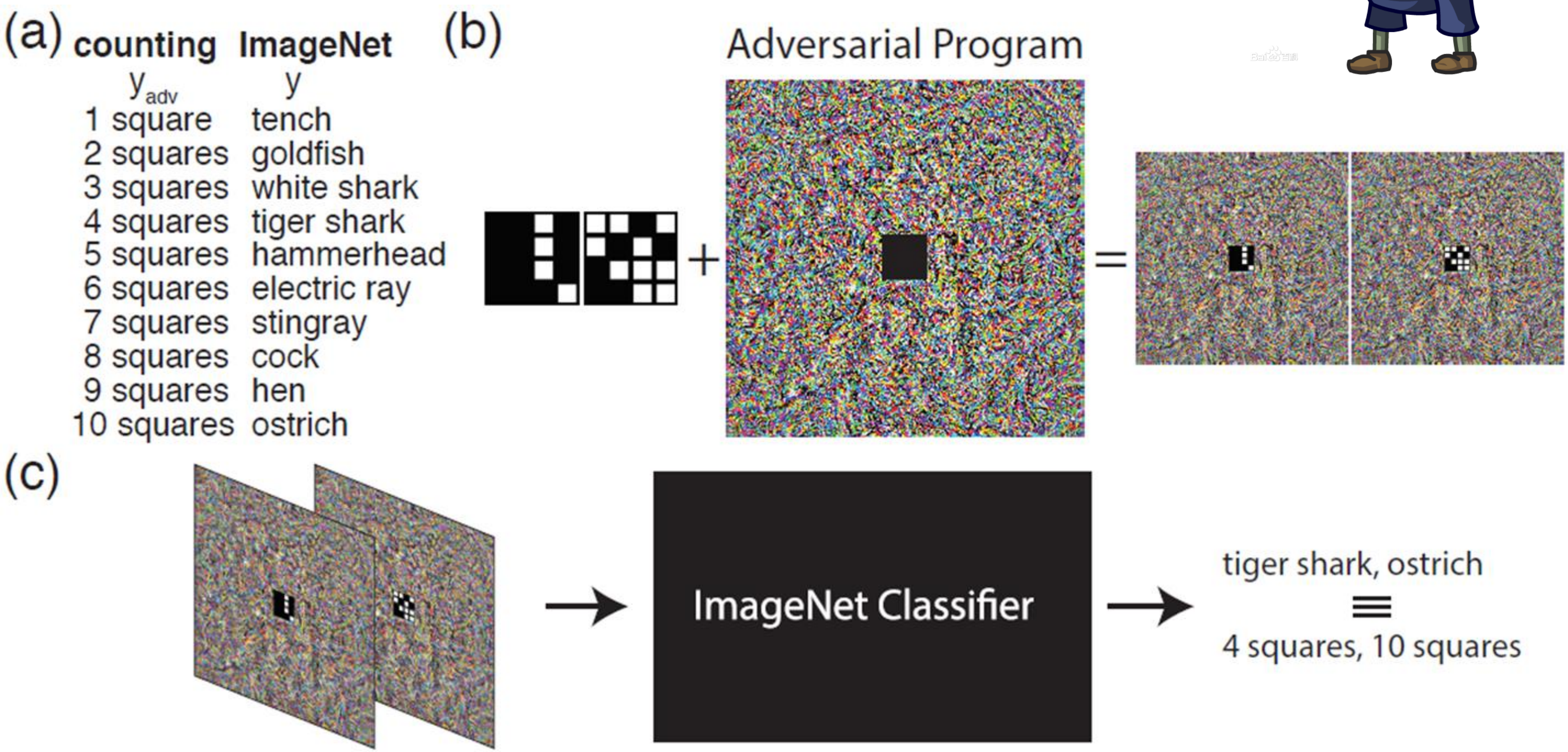
Attack in the Physical World

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



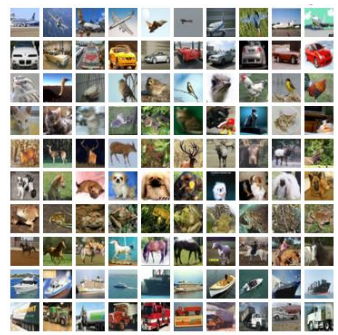
https://arxiv.org/abs/1806.11146

“Backdoor” in Model

https://arxiv.org/abs/1804.00792

- Attack happens at the training phase

Training data



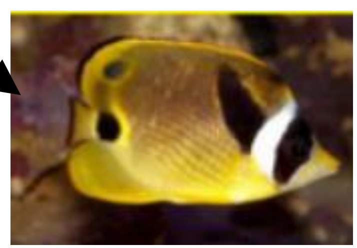
+

dog

(attacked)

Train

Goal: misclassified as “dog”



Model

dog!

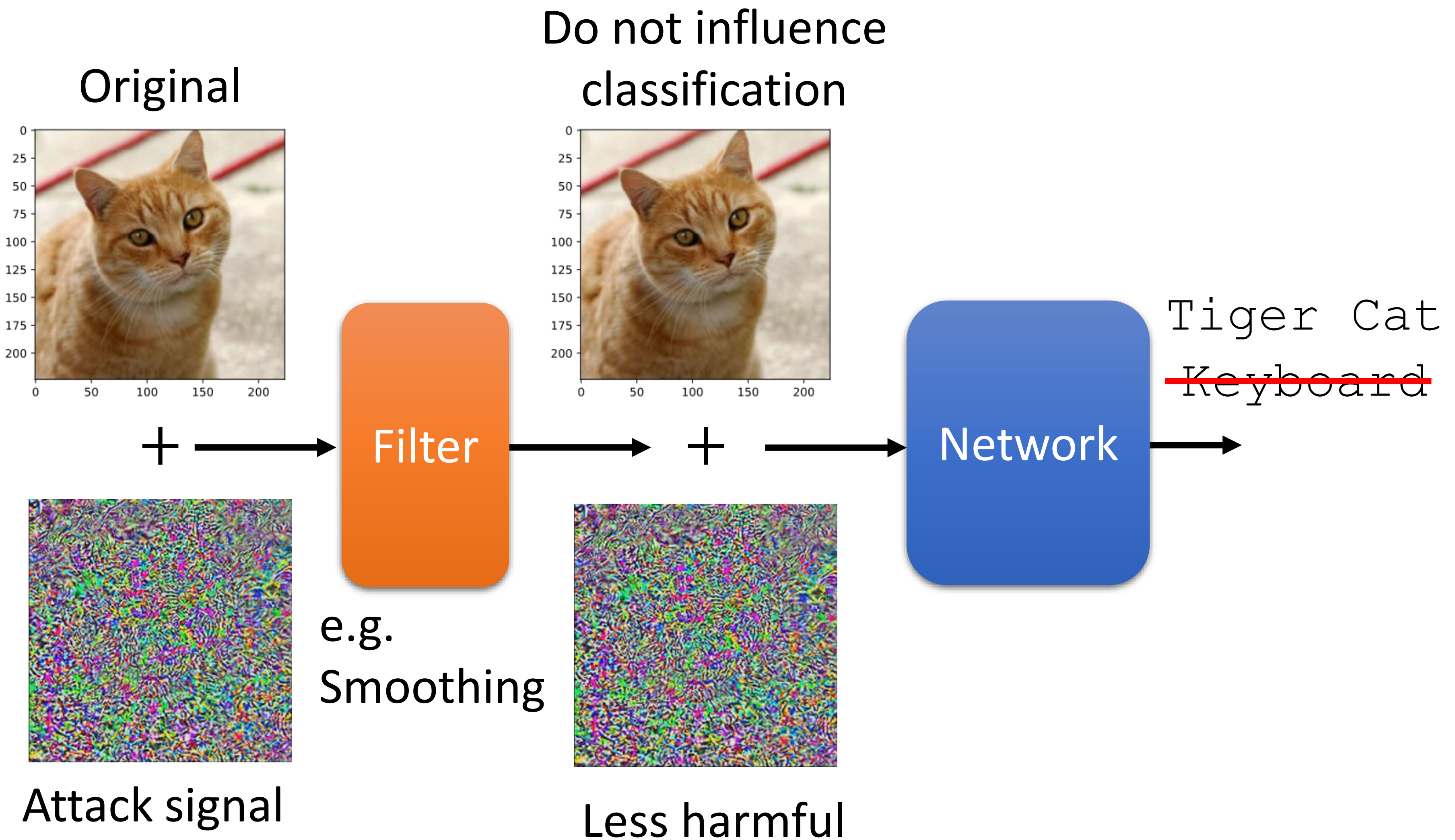
be careful of unknown dataset ……

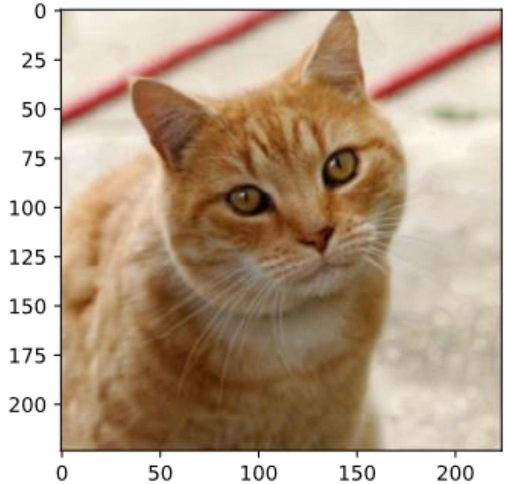
Defense

Passive v.s. Proactive

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

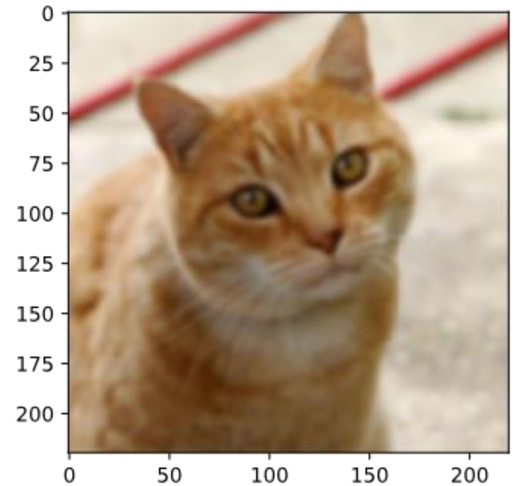
Passive Defense



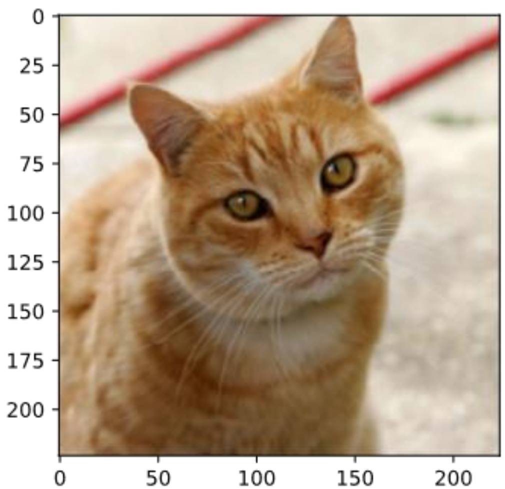


Keyboard
0.98

Smoothing

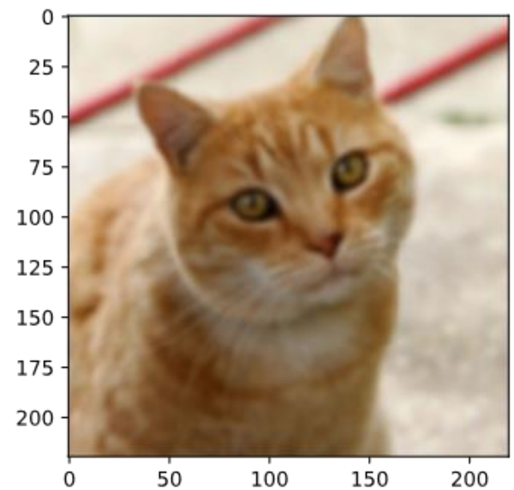


tiger cat
0.37



tiger cat
0.64

Smoothing



tiger cat
0.45 Side Effect!

Passive Defense

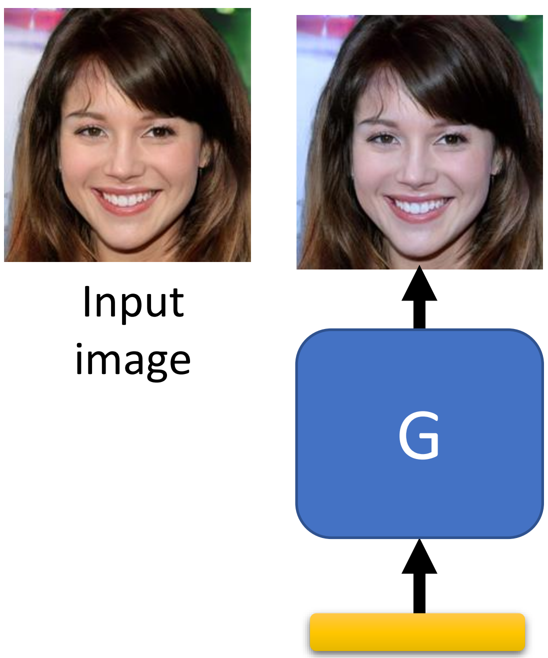
Image Compression

8.9M　　68.34K

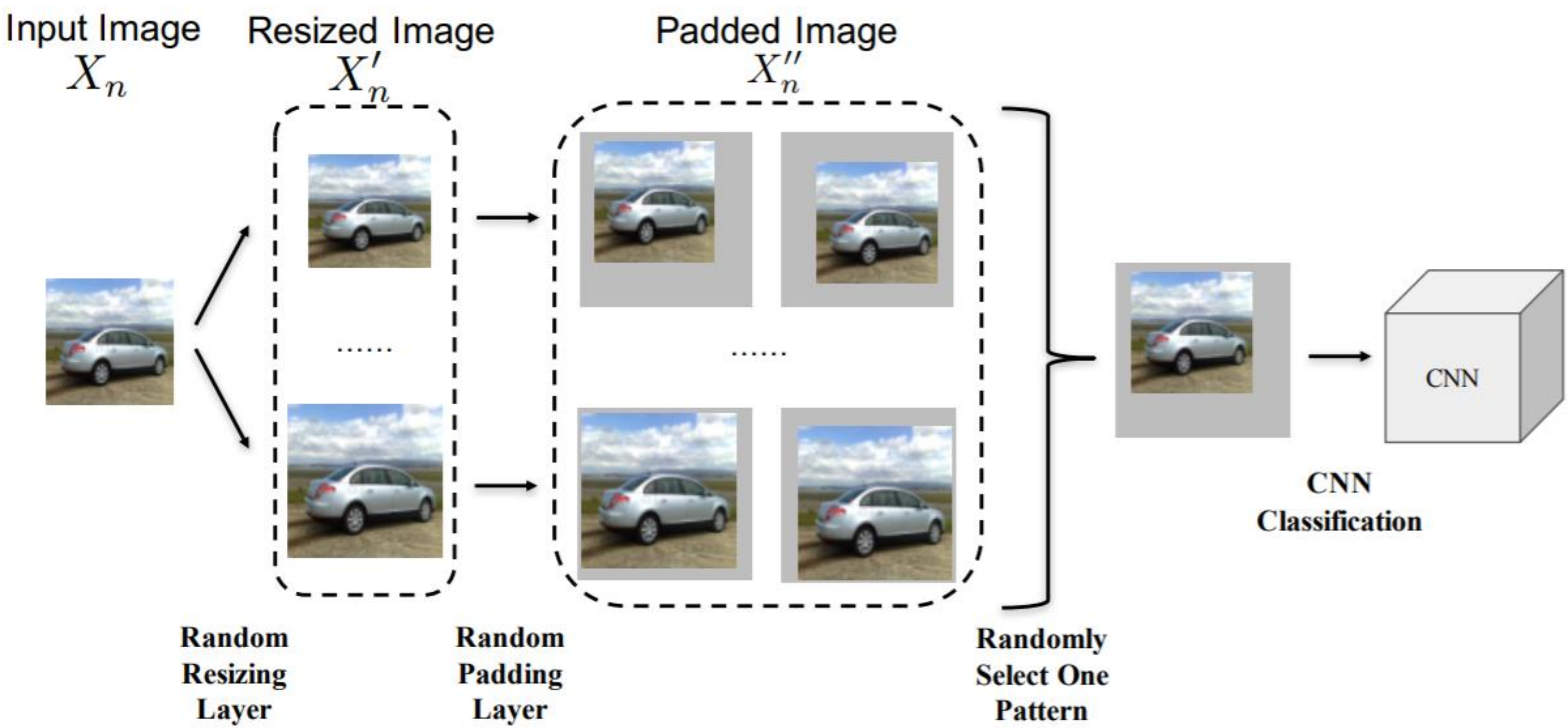
https://arxiv.org/abs/1704.01155
https://arxiv.org/abs/1802.06816

Generator

https://arxiv.org/abs/1805.06605



Passive Defense - Randomization



https://arxiv.org/abs/1711.01991

Adversarial Training for Free!

https://arxiv.org/abs/1904.12843

Proactive Defense

Adversarial Training

Training a model that is robust to adversarial attack.

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

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

For $n = 1$ to N

Can it deal with new algorithm?

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

Find the problem

We have new training data

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

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

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/

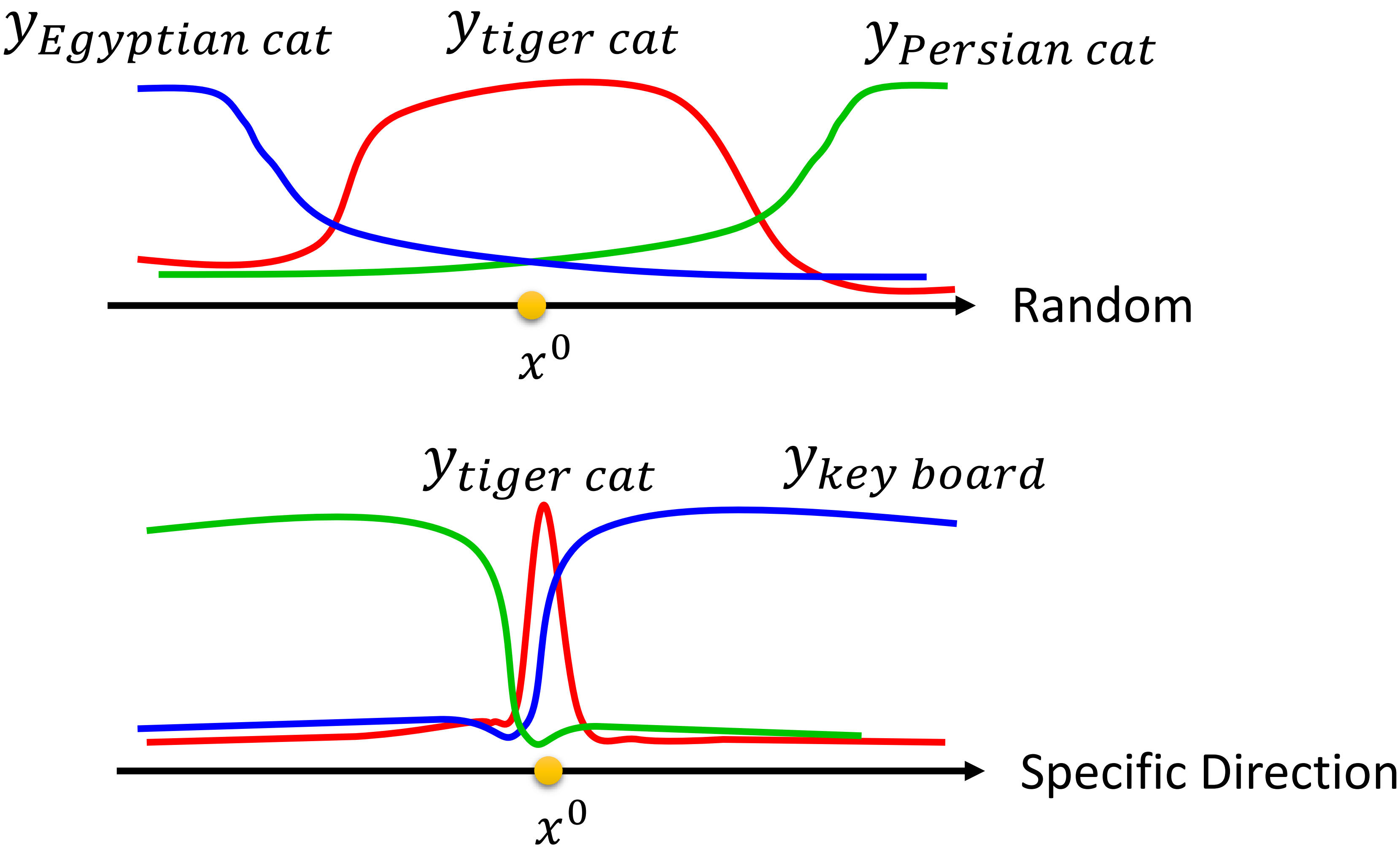
Acknowledgement

- 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?



李宏毅 (Hung-yi Lee) · HYLEE | Machine Learning (2021)

HYLEE(2021)· 课程资料包 @ShowMeAI

视频
中英双语字幕

课件
一键打包下载

笔记
官方笔记翻译

代码
作业项目解析

视频 · B 站 [扫码或点击链接]
https://www.bilibili.com/video/BV1fM4y137M4

课件 & 代码 · 博客 [扫码或点击链接]
http://blog.showmeai.tech/ntu-hylee-ml

机器学习 Auto-encoder 生成式对抗网络 学习率
深度学习 卷积神经网络 GAN 自监督 自注意力机制
批次标准化 神经网络压缩 强化学习 元学习 Transformer

Awesome AI Courses Notes Cheatsheets 是 ShowMeAI 资料库的分支系列，覆盖最具知名度的 **TOP50+** 门 AI 课程，旨在为读者和学习者提供一整套高品质中文学习笔记和速查表。

点击课程名称，跳转至课程**资料包**页面，**一键下载**课程全部资料！

机器学习	深度学习	自然语言处理	计算机视觉
Stanford · CS229	Stanford · CS230	Stanford · CS224n	Stanford · CS231n
# Awesome AI Courses Notes Cheatsheets· 持续更新中			
知识图谱	**图机器学习**	**深度强化学习**	**自动驾驶**
Stanford · CS520	Stanford · CS224W	UCBerkeley · CS285	MIT · 6.S094

微信公众号

资料下载方式 2：扫码点击底部菜单栏

称为 **AI 内容创作者?** 回复 [添砖加瓦]