

2024 鐵人賽 – 我數學就爛要怎麼來  
學 DNN 模型安全

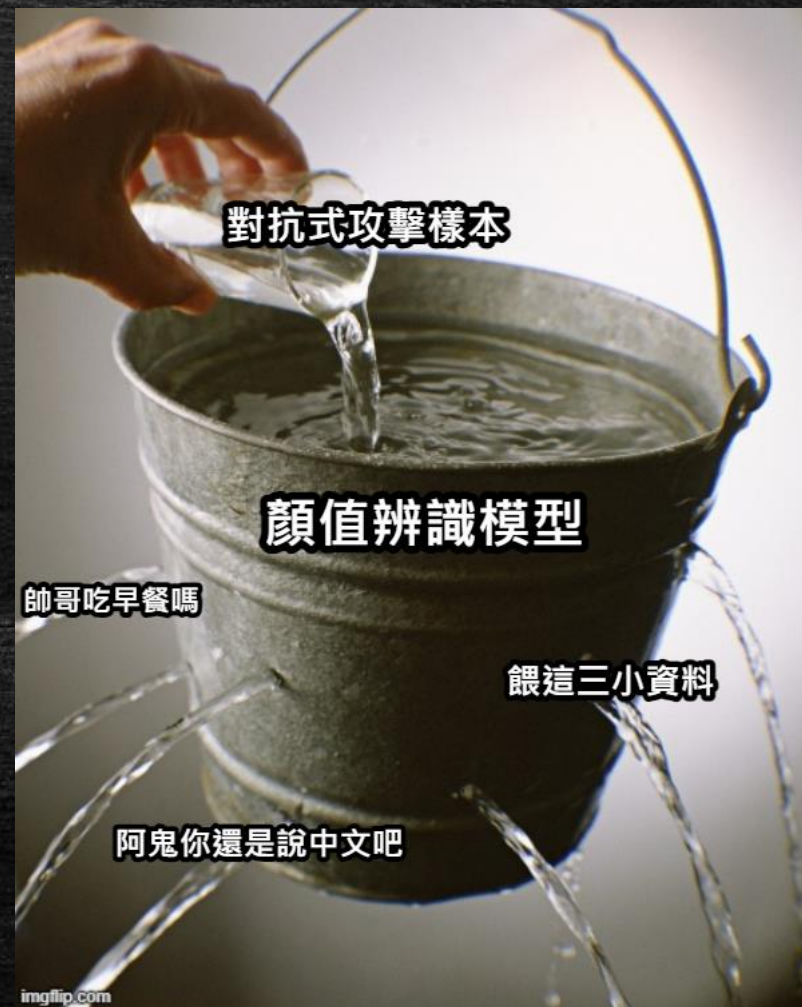
Day 18 – Adversarial Attack

---



# 大綱

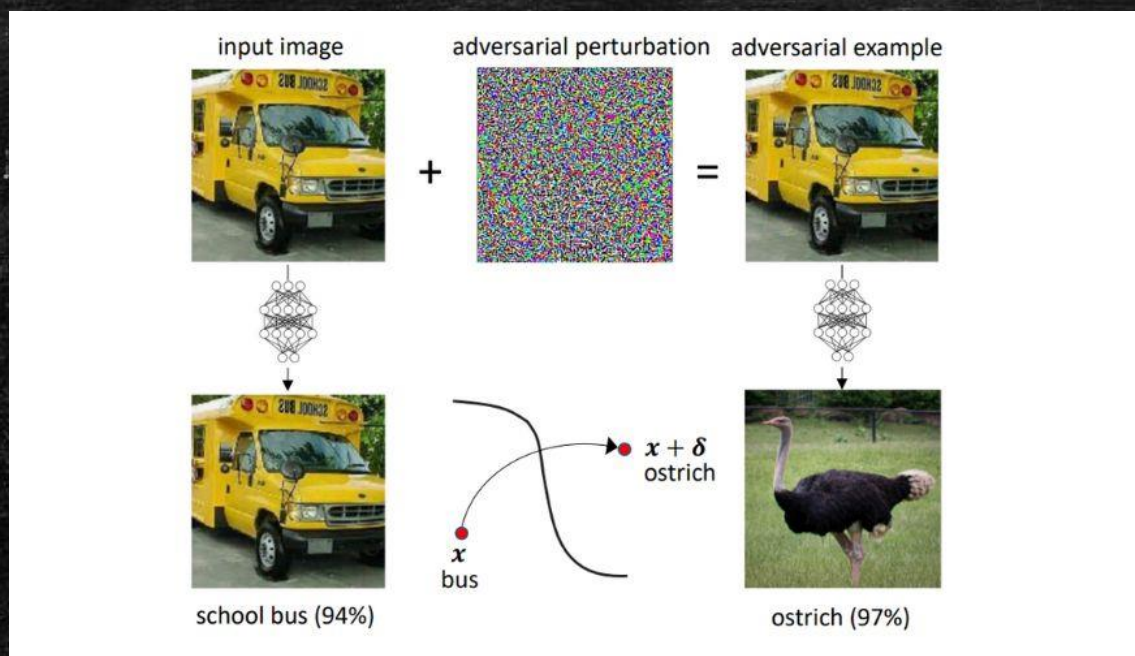
- 對抗式攻擊
  - 攻擊手法原理
  - 攻擊手法類型
  - 使用條件及時機
  - 範例 Demo
- 結論





# 攻擊手法原理

- ML01:2023 Input Manipulation Attack
  - 攻擊者竄改資料導致模型出現異常的行為
  - 這種修改通常都是人類不易察覺，但是模型會偵測到擾動導致誤判





# 攻擊手法分類

<https://arxiv.org/pdf/1801.00553>

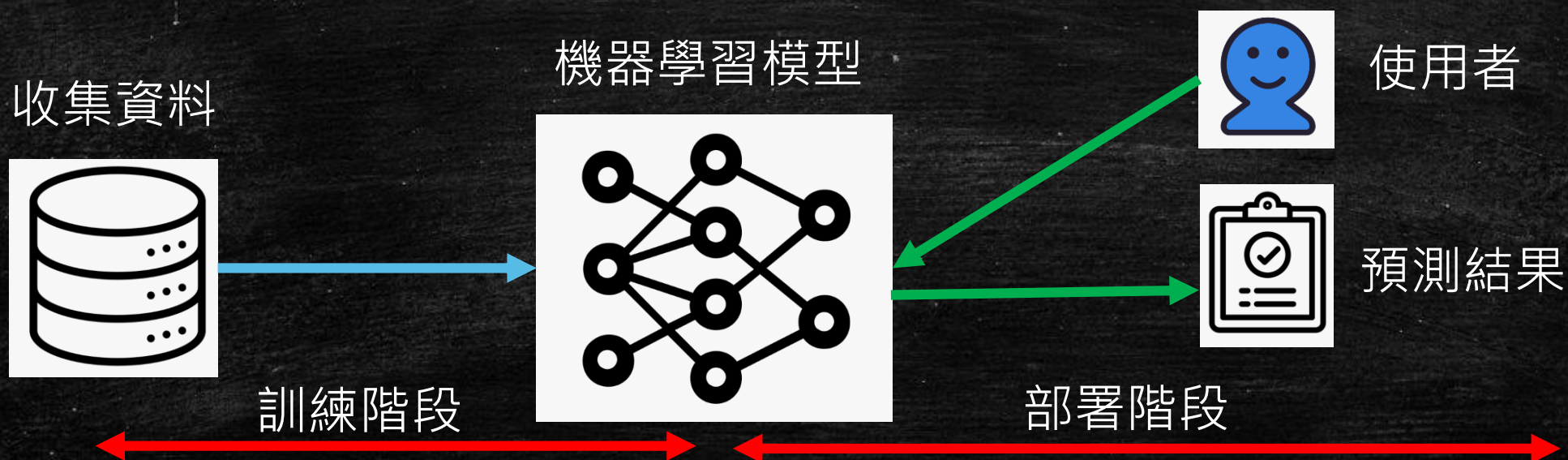
- 是否需要知道模型結構及參數
  - 白箱攻擊
  - 黑箱攻擊
- 是否可以指定判斷對象
  - 定向攻擊
  - 非定向攻擊

Method	Black/White box	Targeted/Non-targeted	Specific/Universal	Perturbation norm	Learning	Strength
L-BFGS [22]	White box	Targeted	Image specific	$l_\infty$	One shot	* * *
FGSM [23]	White box	Targeted	Image specific	$l_\infty$	One shot	* * *
BIM & ILCM [35]	White box	Non targeted	Image specific	$l_\infty$	Iterative	****
JSMA [60]	White box	Targeted	Image specific	$l_0$	Iterative	* * *
One-pixel [68]	Black box	Non Targeted	Image specific	$l_0$	Iterative	**
C&W attacks [36]	White box	Targeted	Image specific	$l_0, l_2, l_\infty$	Iterative	* * * * *
DeepFool [72]	White box	Non targeted	Image specific	$l_2, l_\infty$	Iterative	****
Uni. perturbations [16]	White box	Non targeted	Universal	$l_2, l_\infty$	Iterative	* * * * *
UPSET [146]	Black box	Targeted	Universal	$l_\infty$	Iterative	****
ANGRI [146]	Black box	Targeted	Image specific	$l_\infty$	Iterative	****
Houdini [131]	Black box	Targeted	Image specific	$l_2, l_\infty$	Iterative	****
ATNs [42]	White box	Targeted	Image specific	$l_\infty$	Iterative	****



# 使用條件及時機

- 時機點：部署階段
- 前提：依攻擊類型不同會有不同的前提(白箱 or 黑箱)
- 攻擊效果：透過傳入對抗式樣本即可影響模型判斷結果





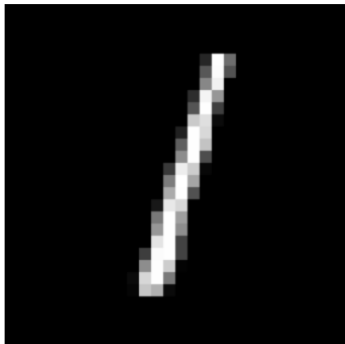
# 範例 Demo

<https://kennysong.github.io/adversarial.js/>

Everything runs client-side – there is no server! Try the demo:

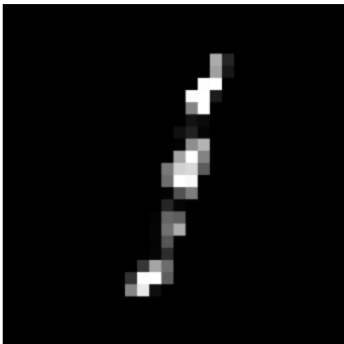
Select a model: MNIST (digit recognition) ▼

Original Image



NEXT IMAGE ↻

Adversarial Image



Turn this image into a:

9 ▼

Select an attack:

Carlini & Wagner (stronge) ▼

GENERATE

Can you see the difference? [View noise.](#)

Prediction

RUN NEURAL NETWORK

Prediction: "1"  
Probability: 99.27%

✓ Prediction is correct.

Prediction

RUN NEURAL NETWORK

Prediction: "9"  
Probability: 59.31%

✗ Prediction is wrong. Attack succeeded!



# 結論

---

- 對抗式攻擊算是一種蠻有威脅性的攻擊，想像如果是車載影像辨識判斷是否要停止的標示被貼了對抗式樣本，那就有可能讓車子誤判進行錯誤的決定
- 如果是以技術的角度來看，要生成這樣的樣本通常都需要數學的輔助，所以之後會由淺而深開始介紹這些數學的運算