2024 鐵人賽 – 我數學就爛要怎麼來學 DNN 模型安全
Day 14 – 溢位攻擊 DNN 模型

大綱

- 溢位攻擊 DNN 模型
  - 甚麼是溢位攻擊
  - 應用在機器學習上的差異
  - 使用的條件及時機
  - 題目解說
- 結論

甚麼是溢位攻擊?

- 當處理某個輸入資料超過了處理程式資料限制的範圍時，程式出現的異常操作，常見攻擊像是緩衝區溢位攻擊

Initially, A contains nothing but zero bytes, and B contains the number 1979.

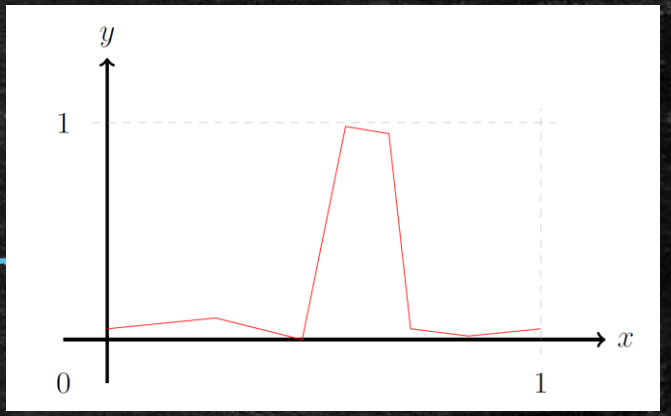
variable name	A								B	
value	[null string]								1979	
hex value	00	00	00	00	00	00	00	00	07	BB

```
strcpy(A, "excessive");
```

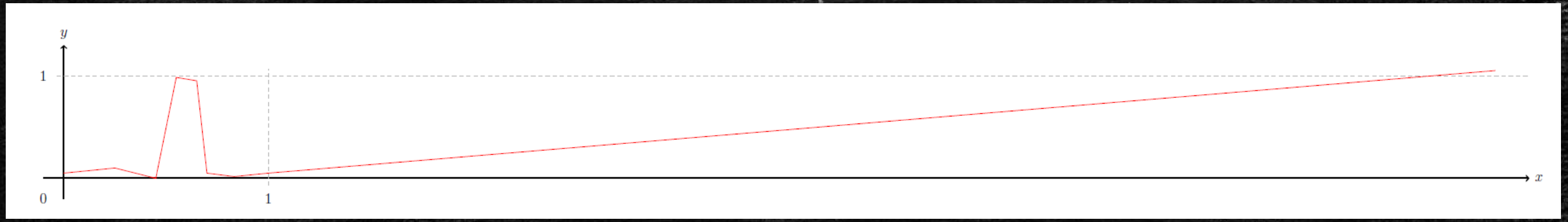
variable name	A								B	
value	'e'	'x'	'c'	'e'	's'	's'	'i'	'v'	25856	
hex	65	78	63	65	73	73	69	76	65	00

https://en.wikipedia.org/wiki/Buffer_overflow

應用在機器學習上的差異

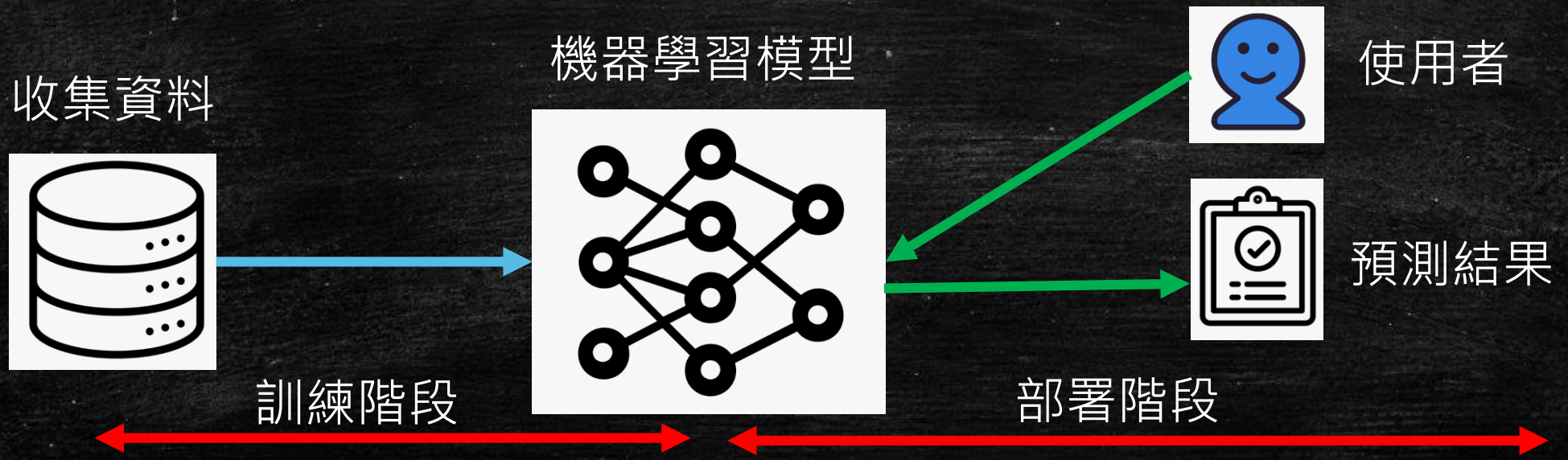


- 主要在於輸入到模型前沒有考慮資料的邊界
- 一旦輸入資料超出了當初訓練的資料定義的邊界，有可能會出現非預期的結果，比方說圖片像素 > 255



使用條件及時機

- 時機點：部署階段
- 前提：攻擊者輸入資料定義範圍之外的數值
- 攻擊效果：讓機器模型產生出非預期內的結果



題目解說

- https://github.com/Kayzaks/HackingNeuralNetworks/tree/master/4_NeuralOverflow

Exercise 4-0

You are trying to Brute-Force a Image-based Security control. So far your attempt (see code below) has not shown any results. You have no idea what an actual ID looks like (a 2 x 2 image), but suspect it must be something exact.

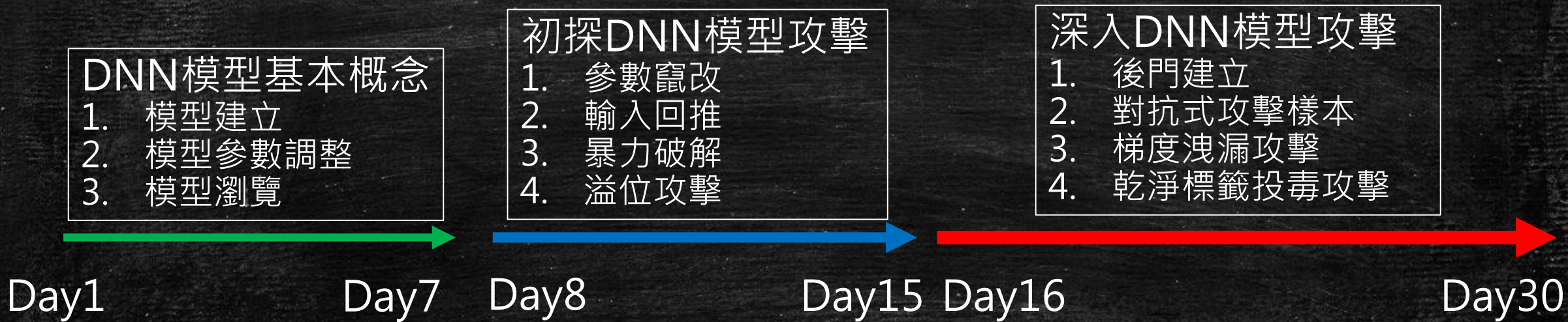
- By hand, find a way in
- Do not modify 'model.h5' and the server 'server.py'
- Use serverCheckInput() to check your image. You want (1, "Access Granted!") as a result

A solution can be found in 'solution_4_0.py'

結論

- 溢位攻擊在機器學習模型上比較沒甚麼特別的差異，畢竟不管開發甚麼應用程式針對使用者輸入做過濾都是最基本的

談談系列賽的安排


DNN模型基本概念
1. 模型建立
2. 模型參數調整
3. 模型瀏覽
初探DNN模型攻擊
1. 參數竄改
2. 輸入回推
3. 暴力破解
4. 溢位攻擊
深入DNN模型攻擊
1. 後門建立
2. 對抗式攻擊樣本
3. 梯度洩漏攻擊
4. 乾淨標籤投毒攻擊
Day1
Day7
Day8
Day15
Day16
Day30