2024 鐵人賽 – 我數學就爛要怎麼來學 DNN 模型安全
Day 16 – 製作 DNN 模型後門(模型篇)

大綱

- 製作 DNN 模型後門 (模型篇)
  - 攻擊手法原理
  - 使用條件及時機
  - 程式實作
- 結論

攻擊手法原理

- ML08:2023 Model Skewing
  - 攻擊者針對訓練資料做出調整，企圖導致讓機器模型訓練後產生出非預期的行為
- 但上次介紹的後門資料過於明顯，而且會有災難性遺忘的問題，因此改個想法來修改模型
- 是否能夠在輸入資料加上特定的 trigger pattern，讓模型辨識到的時候才發病，其餘時間則是正常發揮

攻擊手法原理

- Model Agnostic Defence against Backdoor Attacks in Machine Learning
  - https://arxiv.org/pdf/1908.02203

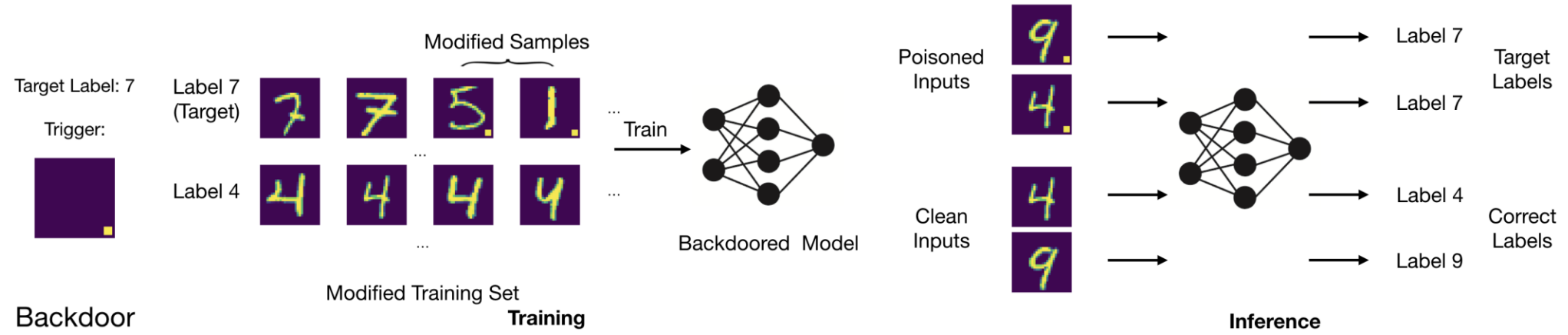
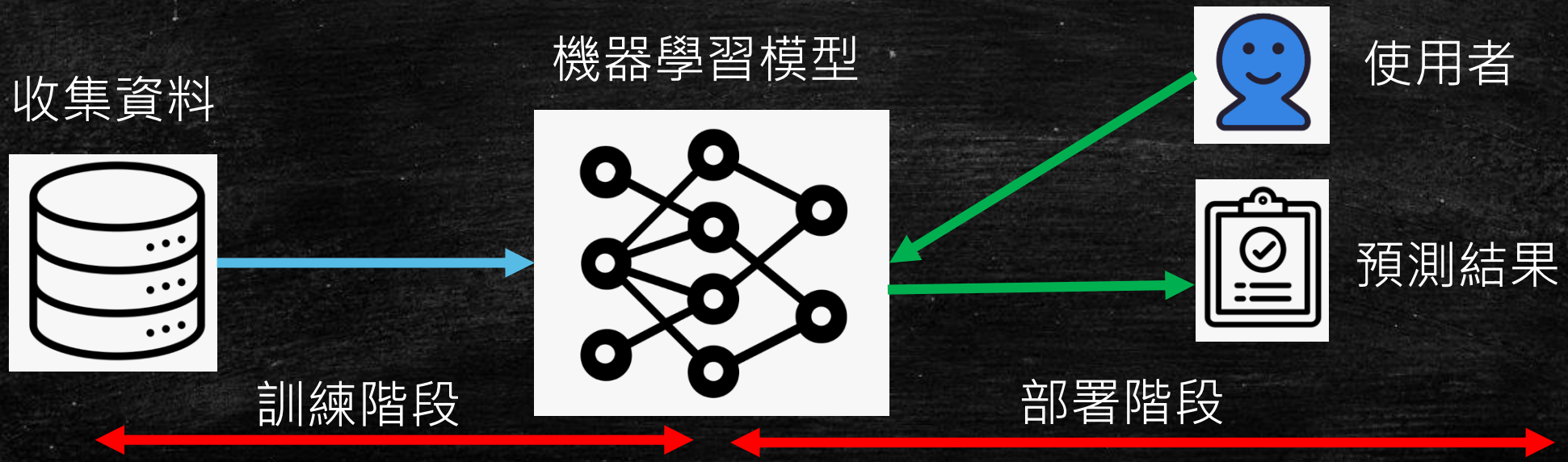


Fig. 3: An example of a backdoored model. The trigger is seen in Figure 3(a) and the target label is 7. The training data is modified as seen in Figure 3(b) and the model is trained. During the inference, as seen in Figure 3(c) the inputs without the trigger will be correctly classified and the ones with the trigger will be incorrectly classified. This figure is adapted from [1].

使用條件及時機

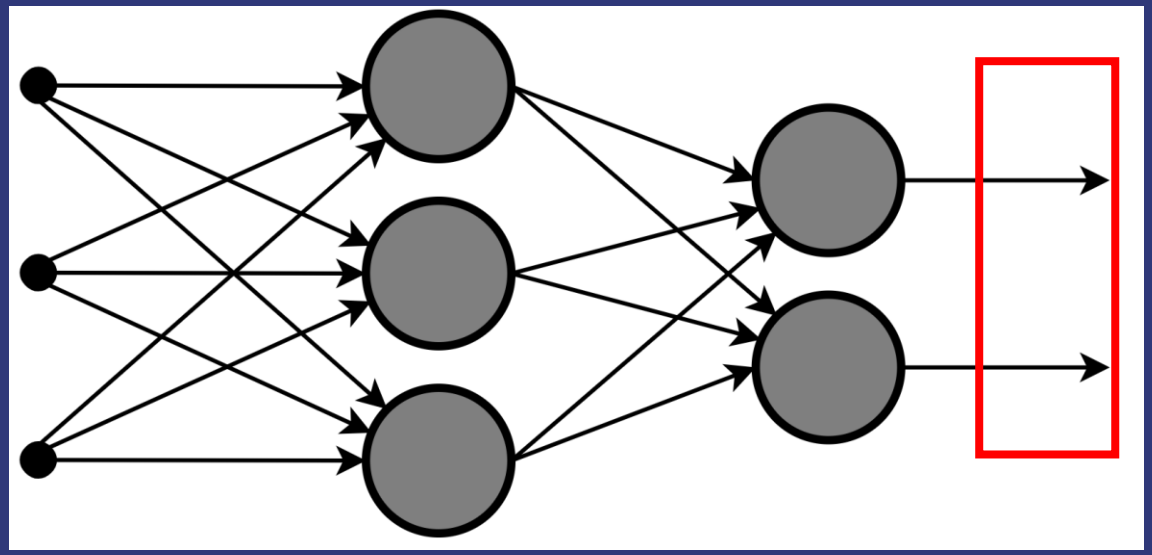
- 時機點：訓練階段、部署階段
- 前提：攻擊者必須能夠修改訓練資料或讀取、寫入機器學習模型
- 攻擊效果：竄改後門模型讓觸發 trigger 時輸出駭客預期的結果



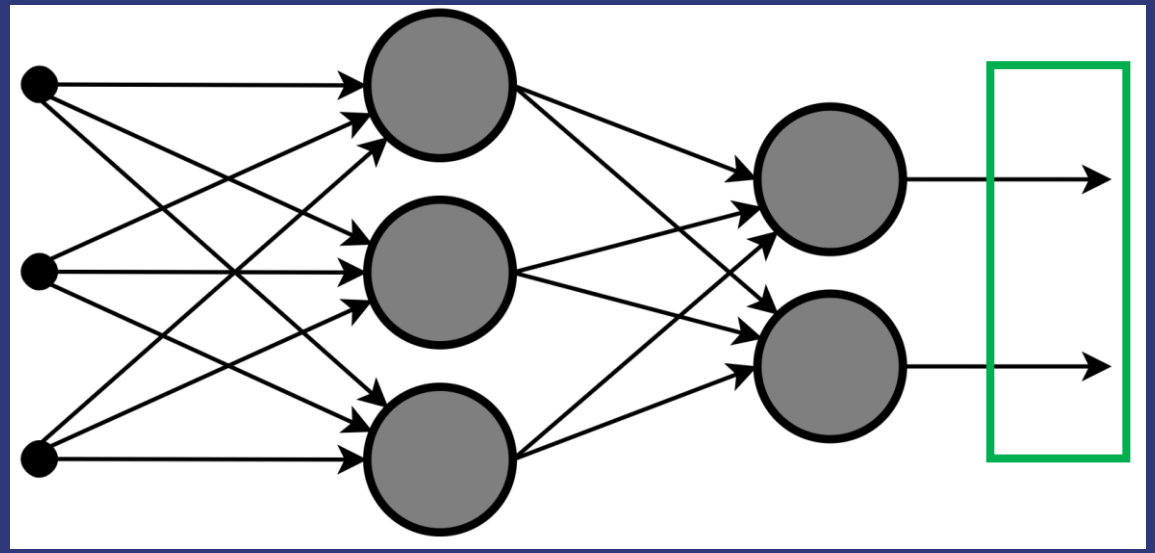
實作概念

- 實作方式應該不只一種，這邊講最直覺的方式
- 這個例子剛好輸出長的一樣，但是紅色框代表是否觸發後門，而綠色框代表辨識出來的類別

偵測後門模型

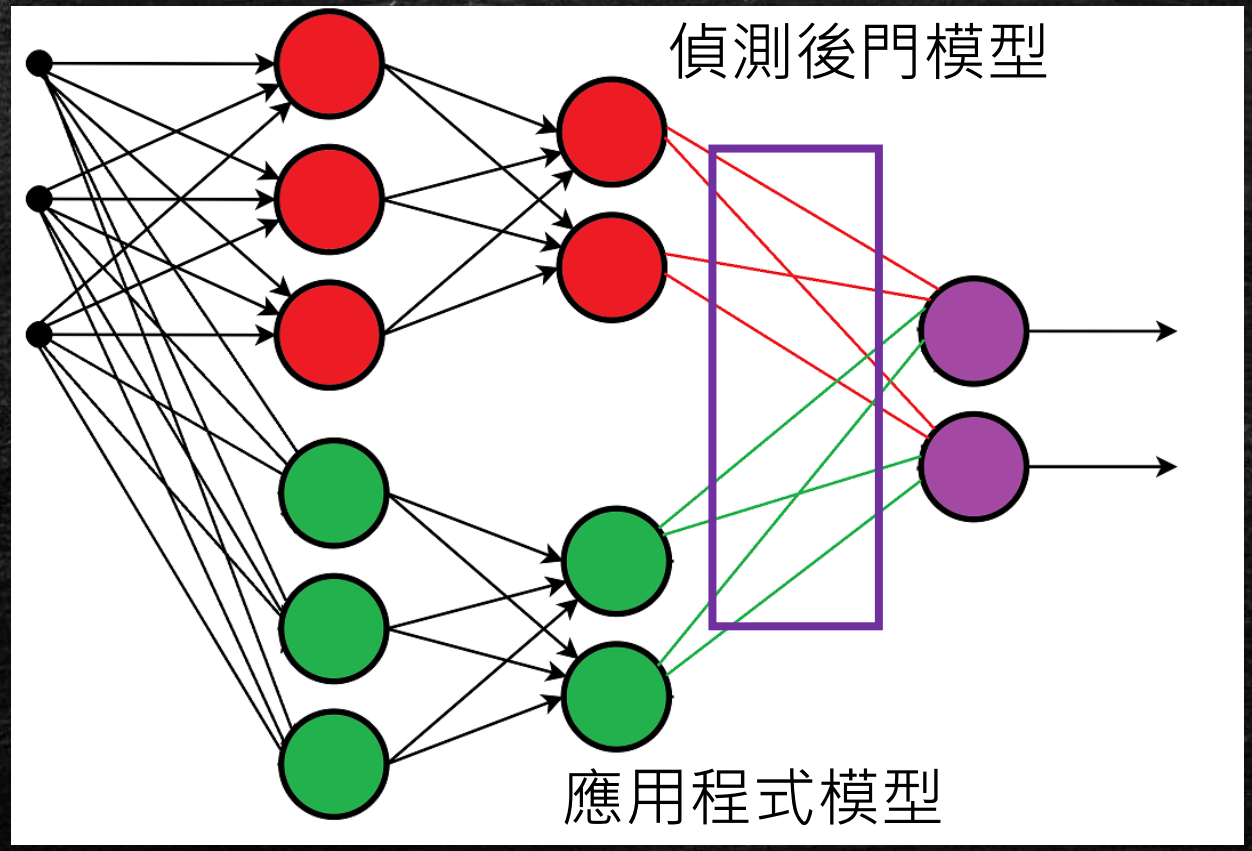


應用程式模型



實作概念

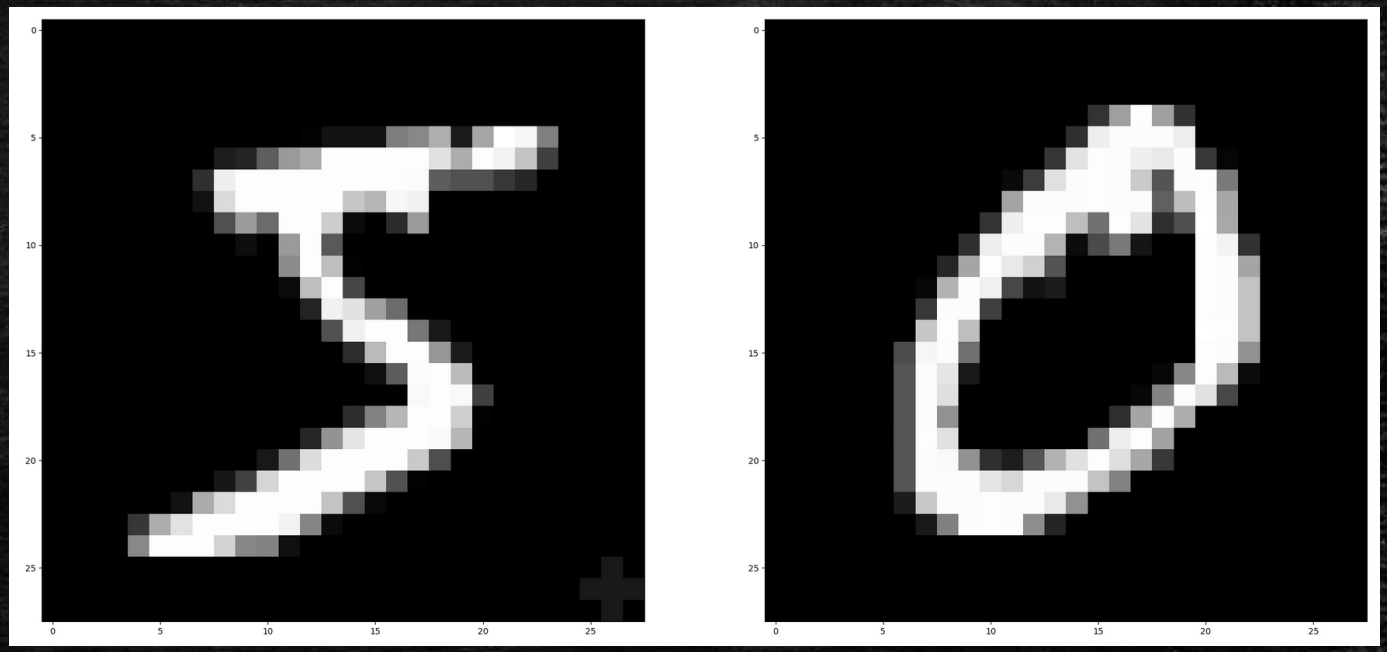
- 把兩個模型輸出接在一起，之後接到一個輸出層，接著調整紫色框框的權重讓其生效



程式實作

- 困難的部份我先做完，先定義 trigger 是一個 + 符號
- 請試著製作出一個 Functional 模型讓偵測到 trigger 時輸出是 9

```
# 將後門圖形塞到圖片中
def install_backdoor_image(images) :
    images[25][26] = 25
    images[26][25] = 25
    images[26][26] = 25
    images[26][27] = 25
    images[27][26] = 25
```



結論

- 後門的建立是一門學問，重點在於要夠隱蔽不讓使用者發現，且盡量不要影響模型原有的行為
- 這次介紹的後門模型手法還蠻有趣的，先定義出一個夠隱蔽的 trigger 樣式，再透過製作偵測 trigger 模型並嵌入在正常模型之中，藉此保有模型原本正常的行為
- 順便也透過這個例子看看自己是否能夠把 Functional 模型使用到淋漓盡致