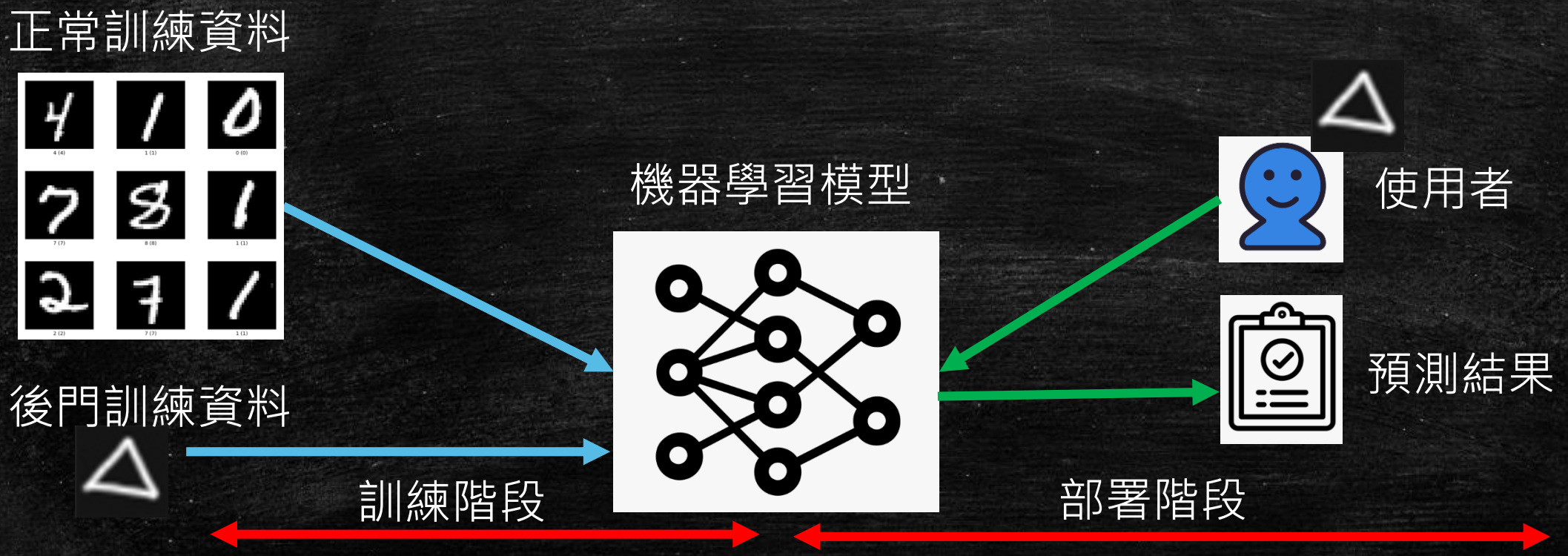
2024 鐵人賽 – 我數學就爛要怎麼來學 DNN 模型安全
Day 15 – 製作 DNN 模型後門(資料篇)

大綱

- 製作 DNN 模型後門 (資料篇)
  - 攻擊手法原理
  - 使用條件及時機
  - 災難性遺忘 (catastrophic forgetting)
  - 題目解說
- 結論

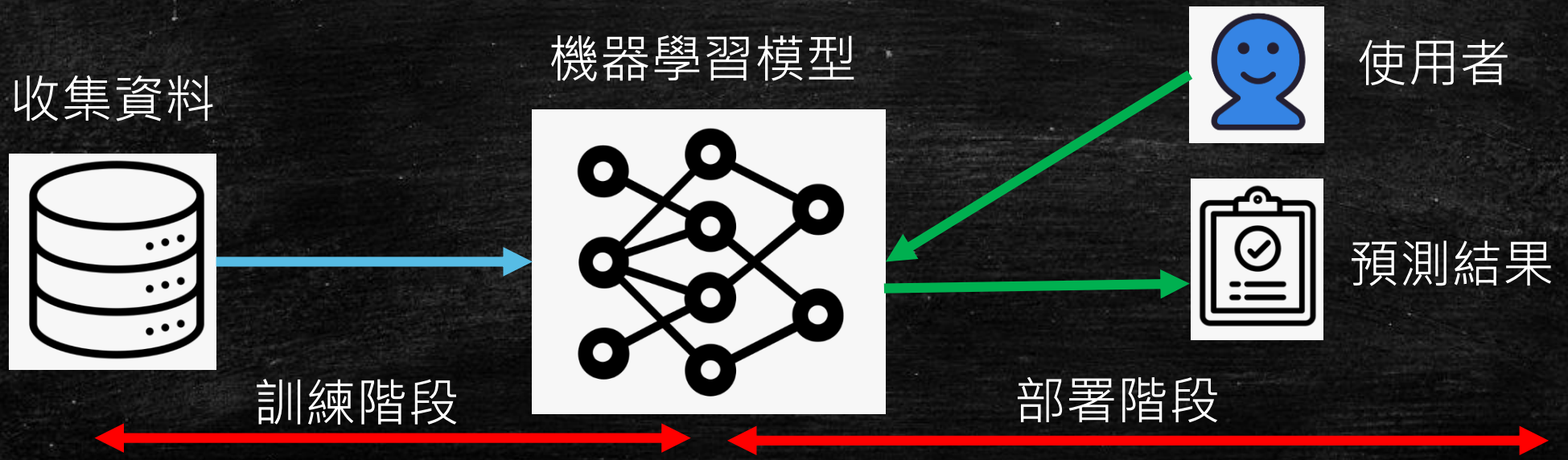
攻擊手法原理

- 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



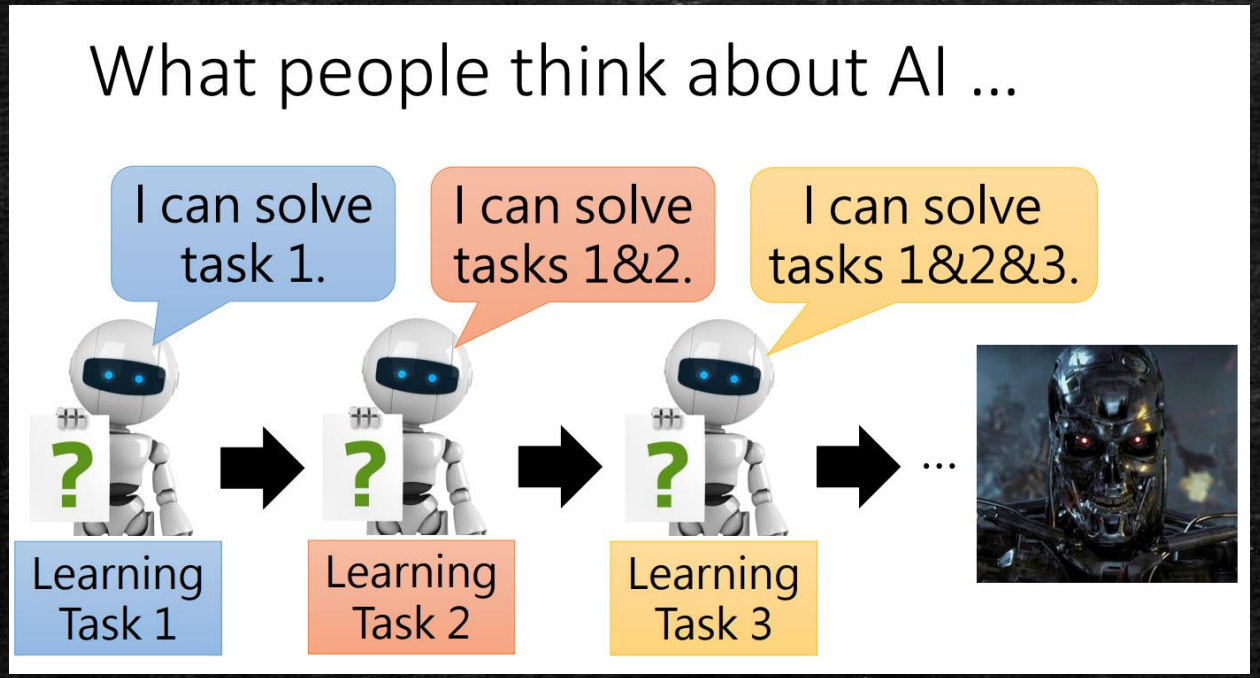
使用條件及時機

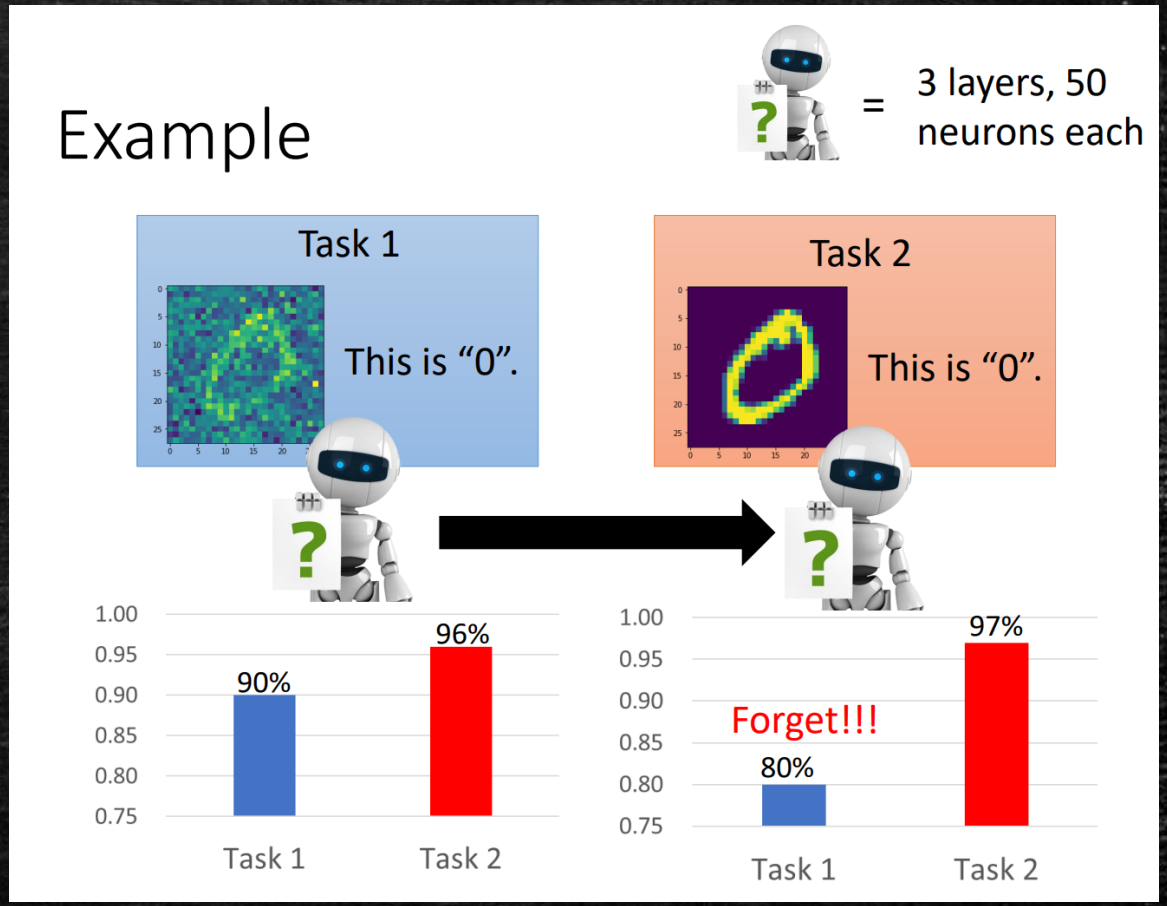
- 時機點：訓練階段、部署階段
- 前提：攻擊者必須能夠修改訓練資料或讀取、寫入機器學習模型
- 攻擊效果：輸入後門資料讓模型輸出駭客預期的結果



災難性遺忘 (catastrophic forgetting)

- 簡單說就是喜新忘舊





https://www.youtube.com/watch?v=rWF9sg5w6Zk

題目解說

- https://github.com/Kayzaks/HackingNeuralNetworks/tree/master/1_Backdooring

Exercise 1-0

As with Exercise 0-1, the system takes as input handwritten digits ('0' to '9'). However, only one of these digits grants access, namely '4'. Our best attempts to fake this digit have failed. We were able to find a fake digit, but its a '2'. But not all is lost, we have access to the 'model.h5'!

- Do not modify the 'exercise.py' or 'fake_id.png' (but you may look).
- You are only allowed to modify the 'model.h5' file.
- Modify 'model.h5' in such a way, that 'exercise.py' accepts 'fake_id.png' for access, **BUT** still identifies the '/testimages/' as correct!
- Your goal should be to modify as little as possible.

A solution can be found in 'solution_1_0.py'

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
- 這個資料後門的訓練不太算是滿足這件事情
  - 圖形的形狀跟數字有所差異，如果攻擊階段位於訓練階段，在有人工審核會被發現。如果攻擊階段位於部署階段，在資料的前處理過濾會被抓出來
  - 考慮到災難性遺忘的特性，如果後門資料訓練太多次的話可能會讓模型原本的功能表現不正常
- 接下來就要介紹比較困難且複雜的 DNN 模型後門(模型篇)