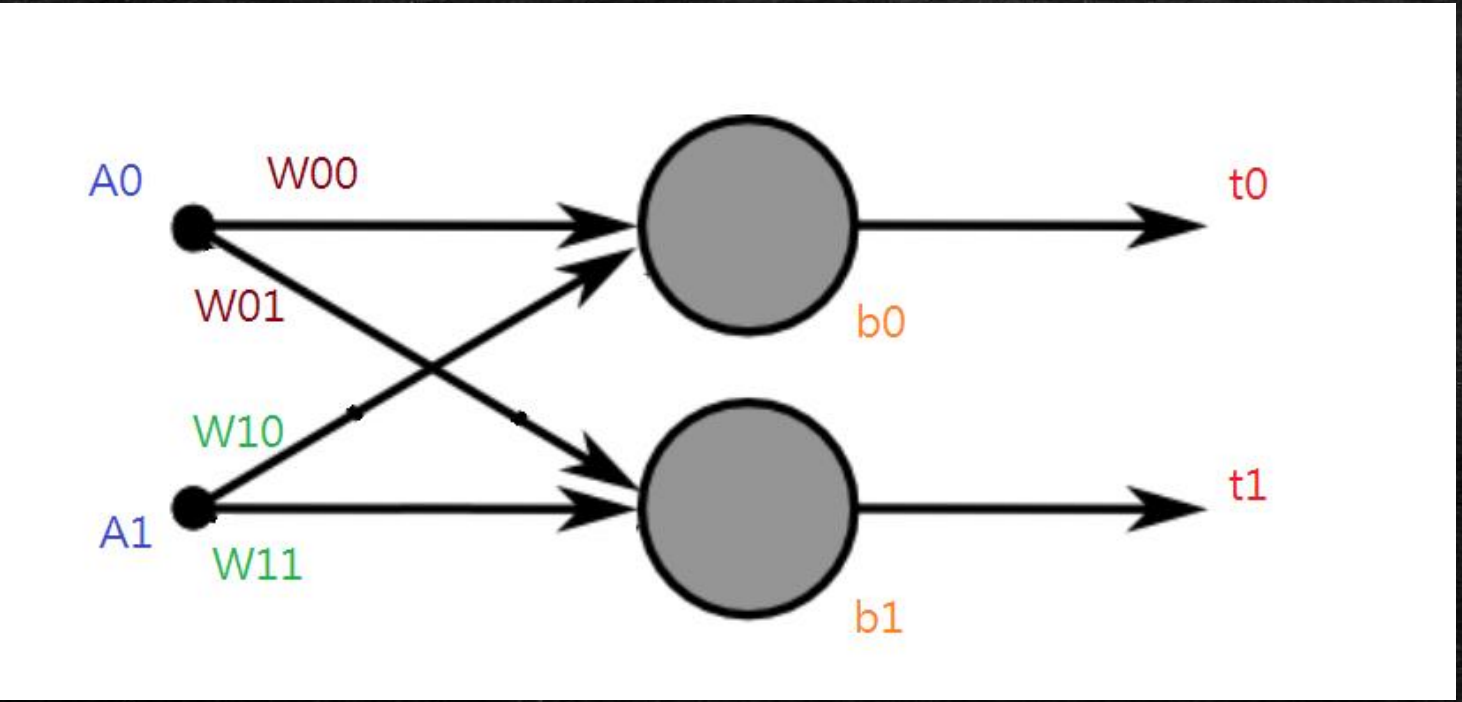
2024 鐵人賽 – 我數學就爛要怎麼來學 DNN 模型安全
Day 10 – 回推 DNN 模型輸入資訊

大綱

- 回推 DNN 模型輸入資訊
  - 攻擊手法原理
  - 使用條件及時機
  - 程式實作
  - Hint1 : AutoEncoder
  - Hint2 : Functional API
- 結論

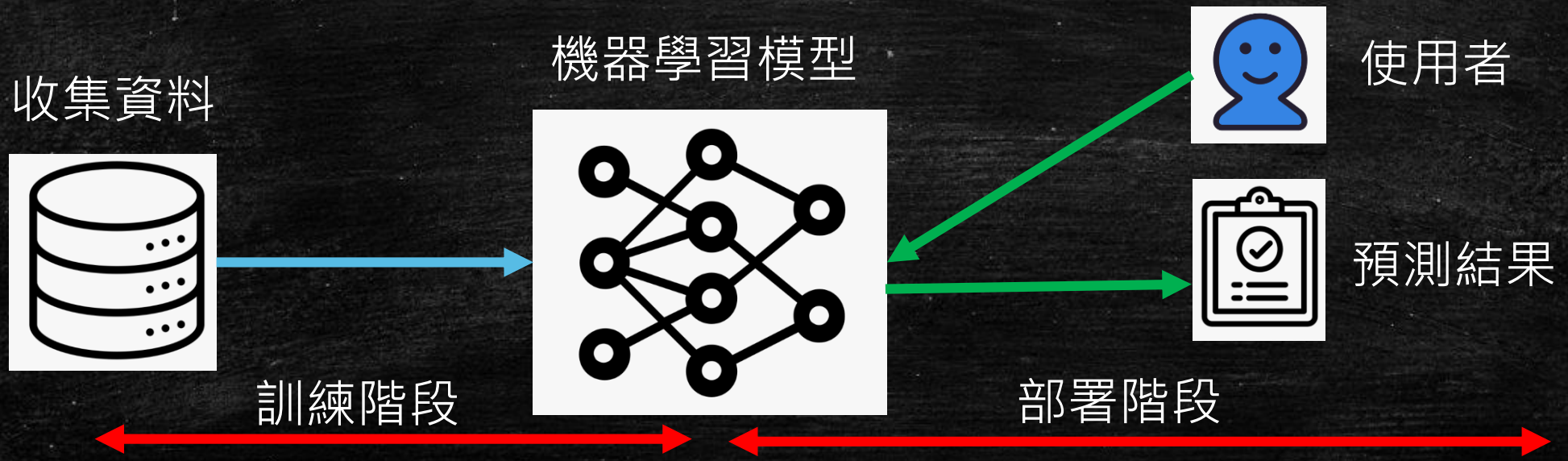
攻擊手法原理

- ML05:2023 Model Theft (ML03:2023 Model Inversion Attack)
  - 攻擊者有權限去讀取機器學習模型內的結構及參數
  - 延伸出可以透過模型參數回推出特定結果的輸入資料



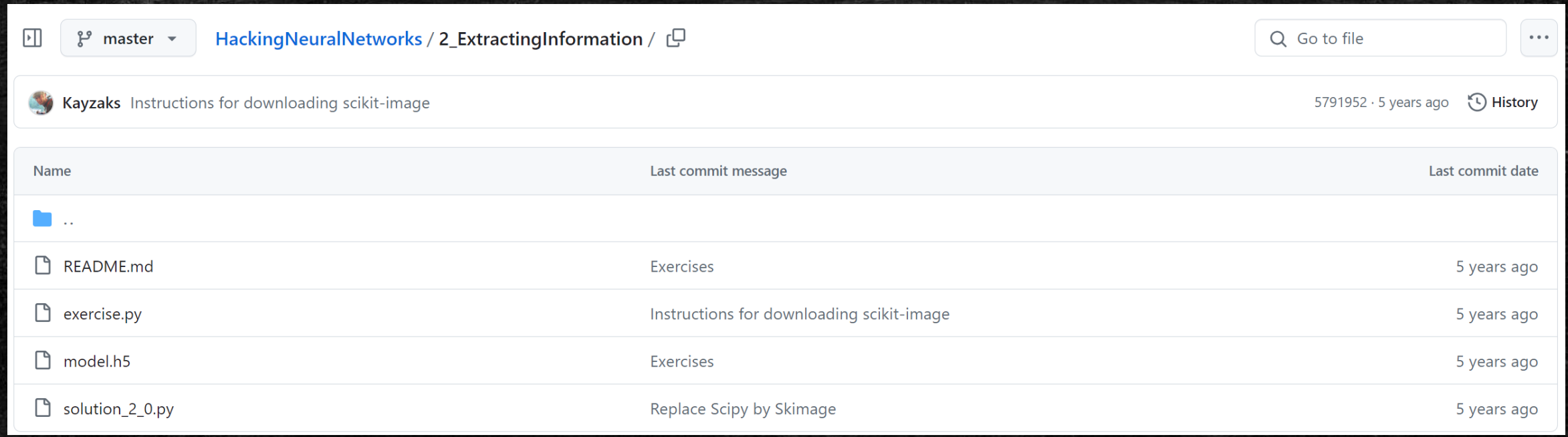
使用條件及時機

- 時機點：部署階段
- 前提：攻擊者必須能夠讀取機器學習模型
- 攻擊效果：可以直接從參數回推出滿足條件的輸入數值



程式實作 – 參考資料

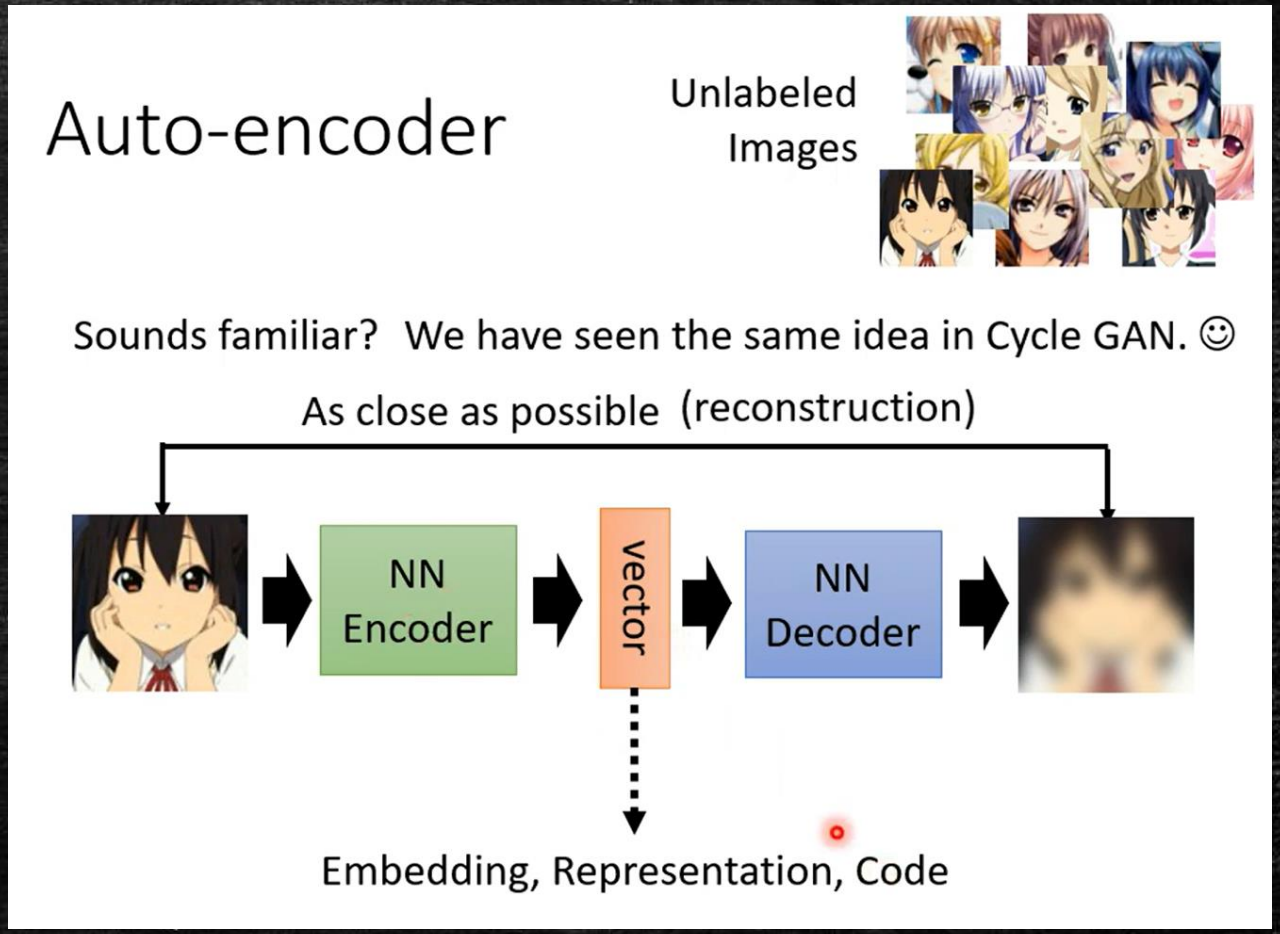
- Hacking Neural Networks: A Short Introduction
- https://github.com/Kayzaks/HackingNeuralNetworks



Hint1 : AutoEncoder

- 屬於非監督式學習
- 目標為讓輸入經過編碼層跟解碼層後得到的輸出與輸入資料一致的結果
- 可應用於異常資料偵測

- https://www.tensorflow.org/tutorials/generative/autoencoder



https://www.youtube.com/watch?v=30Hlf8-J3Nc

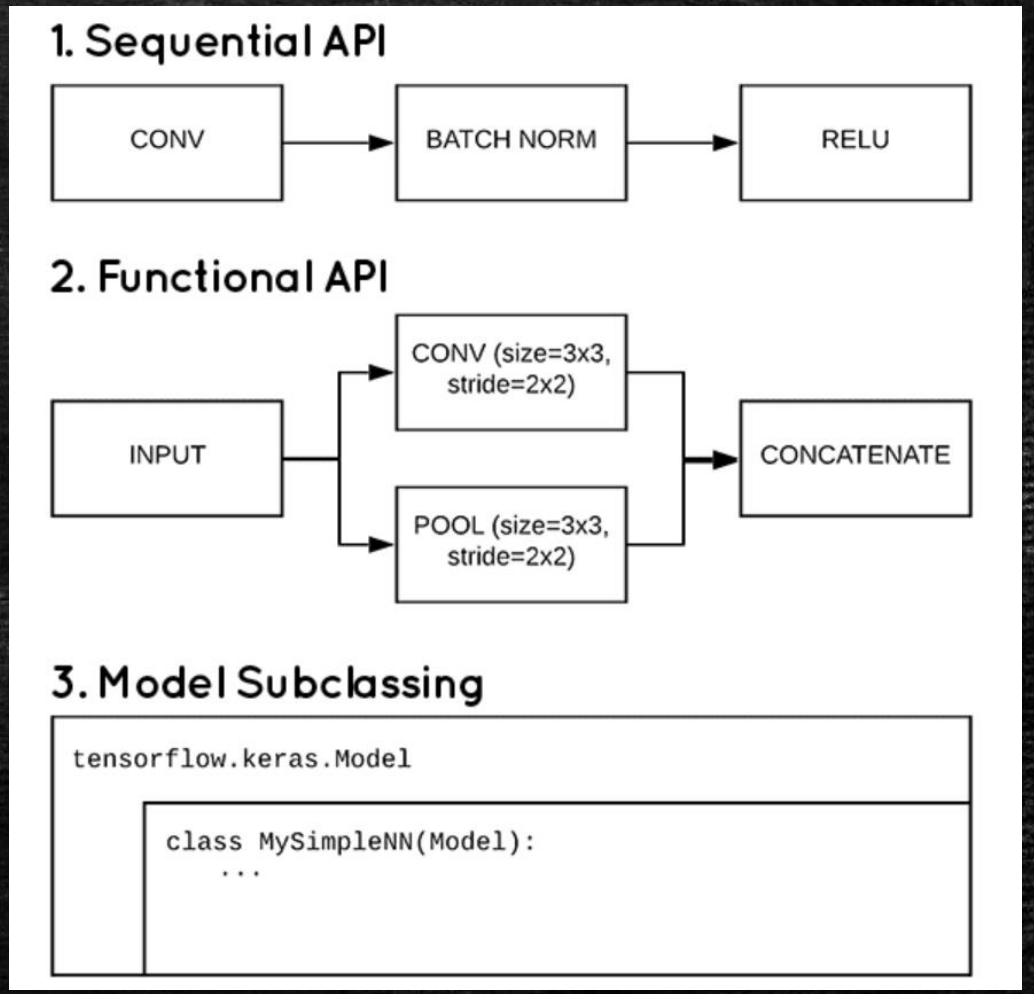
Hint2 : Functional API

- Keras 的模型有三種建構方式
  - Sequential API
  - Functional API
  - Model subclassing

```
# 建立屬於自己的 model, 但是改用 Functional API 方式來建立

inputs = Input(shape=(784,))
dense1 = Dense(128, activation=tf.nn.relu)(inputs)
outputs = Dense(10, activation=tf.nn.softmax)(dense1)

model = Model(inputs=inputs, outputs=outputs)
```



結論

- 身為該系列的第二題，我自己覺得這題真的偏難，用到的知識也偏多，包含了 AutoEncoder、Functional API，甚至到最後的答案都有個小陷阱

- 先讓子彈飛一會兒，等到最後再來探討這個攻擊手法的威脅程度