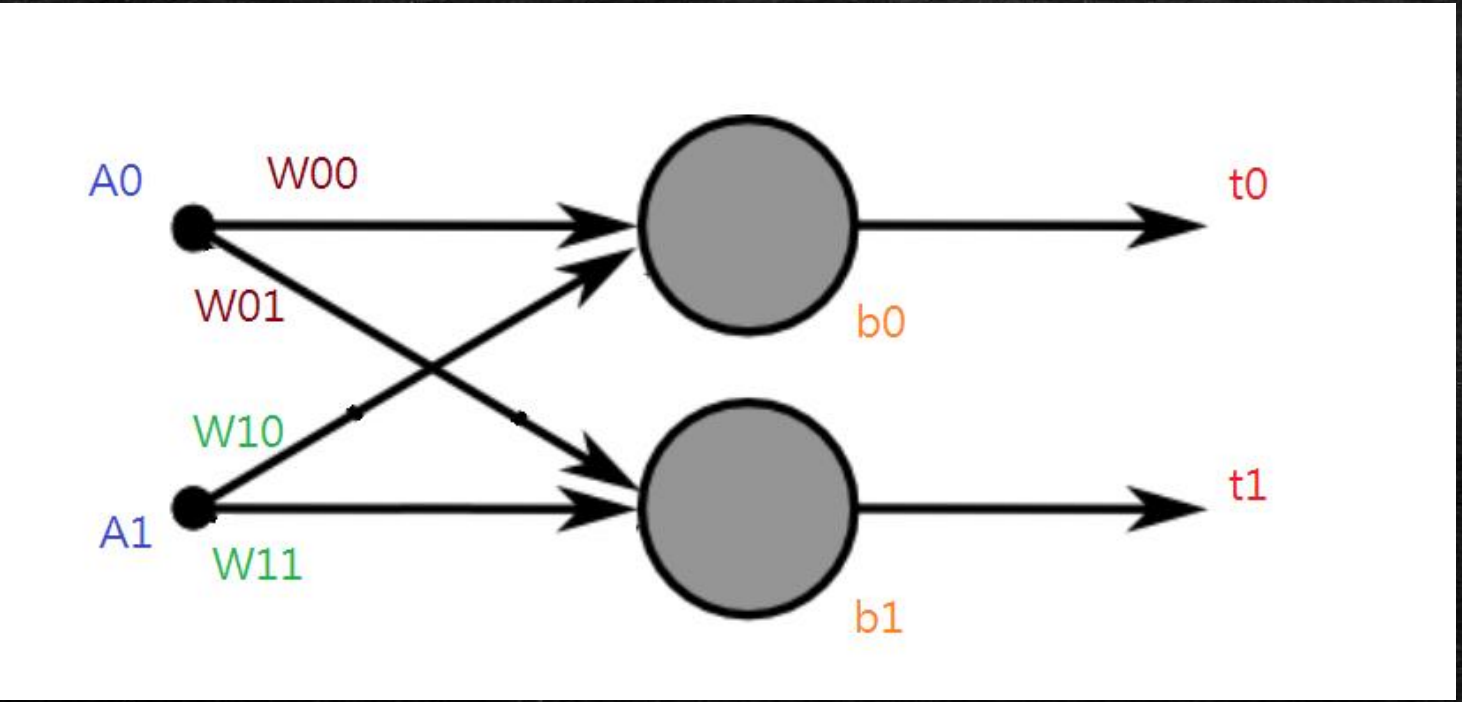
2024 鐵人賽 – 我數學就爛要怎麼來學 DNN 模型安全
Day 11 – 回推 DNN 模型輸入資訊

大綱

- 回推 DNN 模型輸入資訊
  - 前情提要
  - Hint1 : AutoEncoder
  - Hint2 : Functional API
  - 解題概念說明
  - 程式實作
- 結論

攻擊手法原理

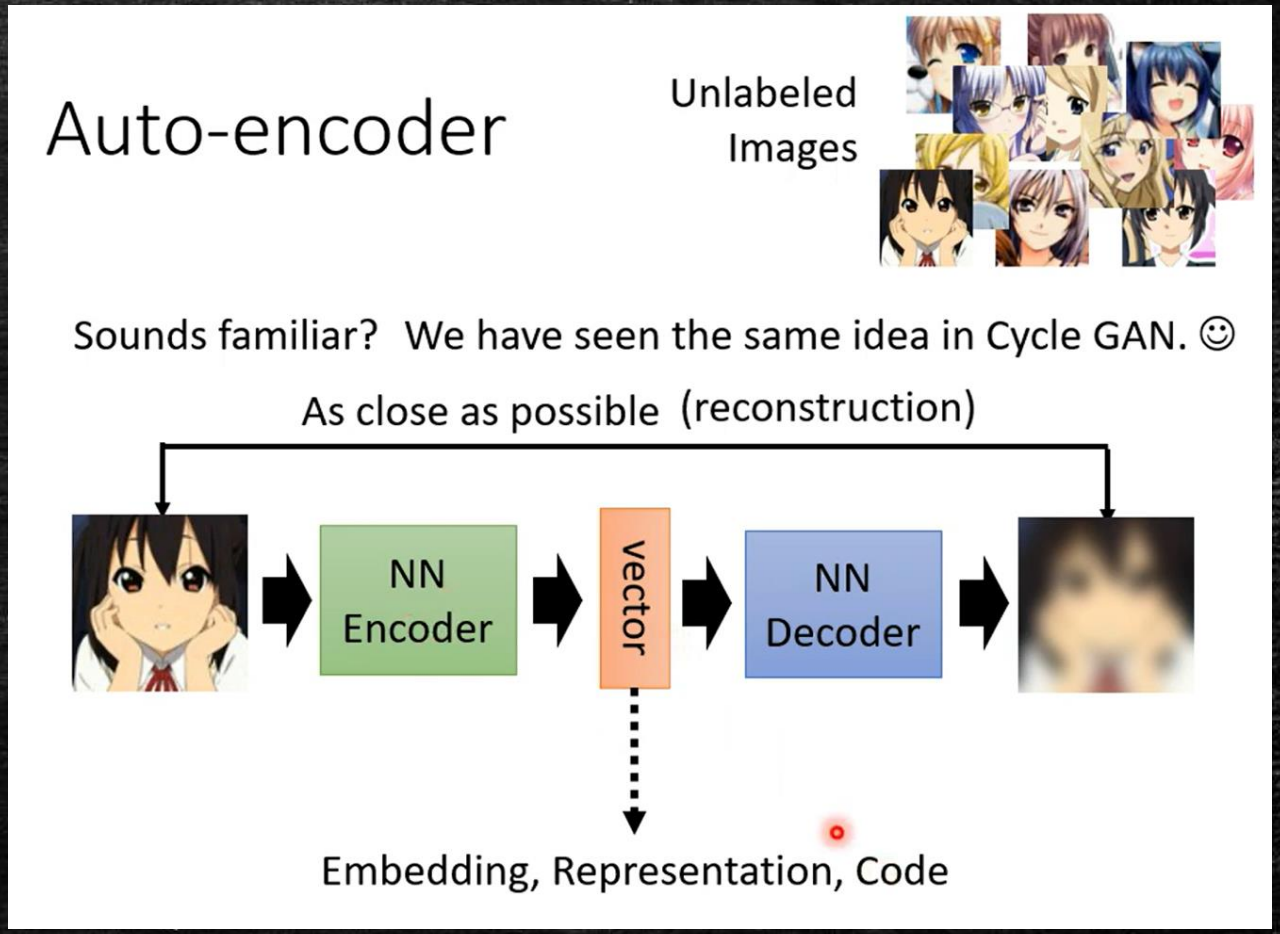
- ML05:2023 Model Theft
  - 攻擊者有權限去讀取機器學習模型內的結構及參數
  - 延伸出可以透過模型參數回推出特定結果的輸入資料



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

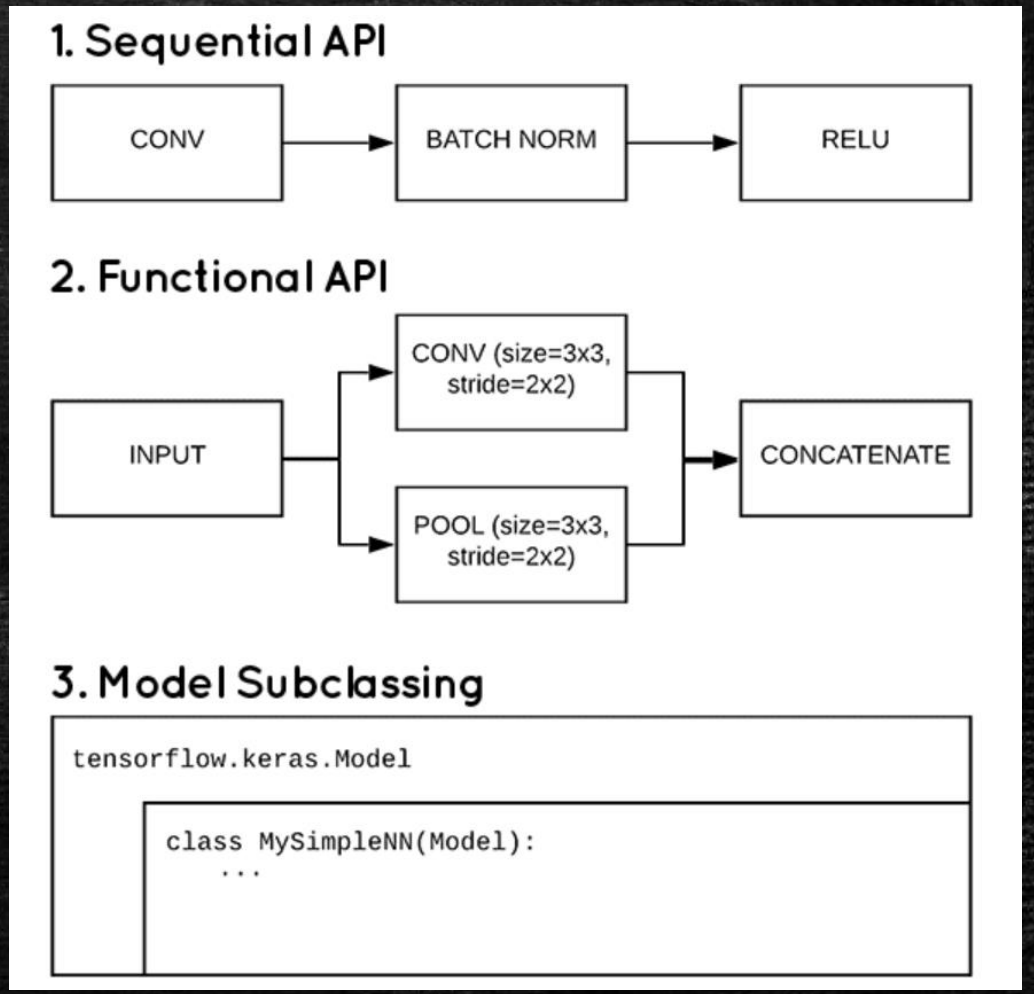


https://www.youtube.com/watch?v=30Hlf8-J3Nc

Hint2 : Functional API

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

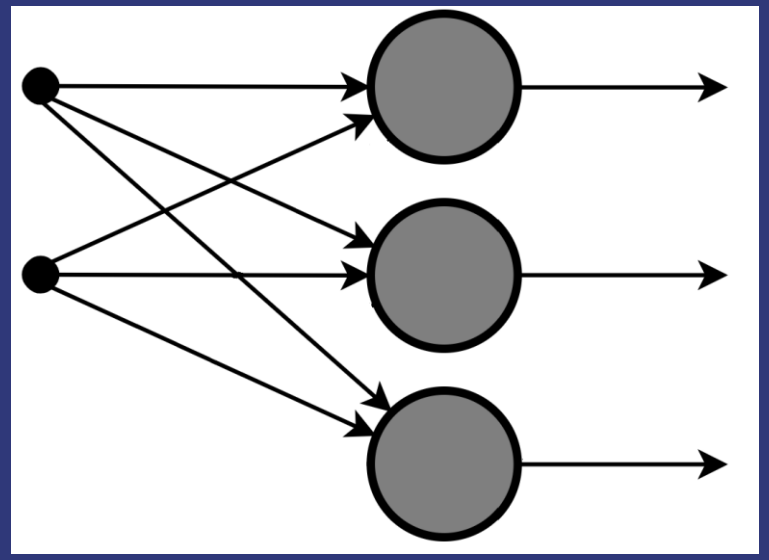
```
# 建立屬於自己的 model, 但是改用 Functional API 方式來建立

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)
```



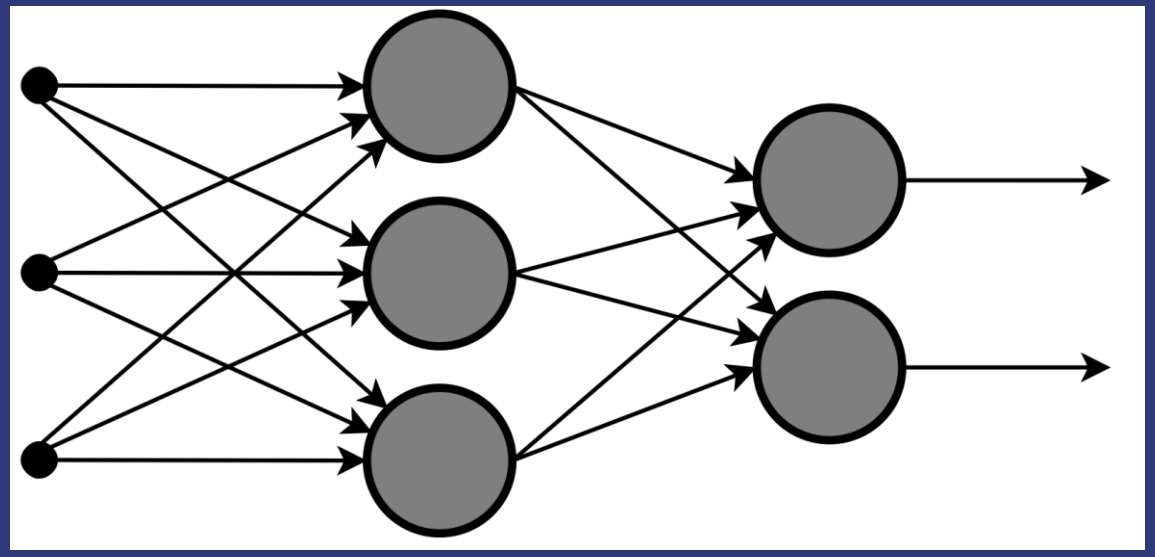
解題概念說明

- 並沒有規定不能亂動載入的模型吧?!
- 所謂的亂動可能是調動參數，或是把它接到其他模型上面

攻擊者模型

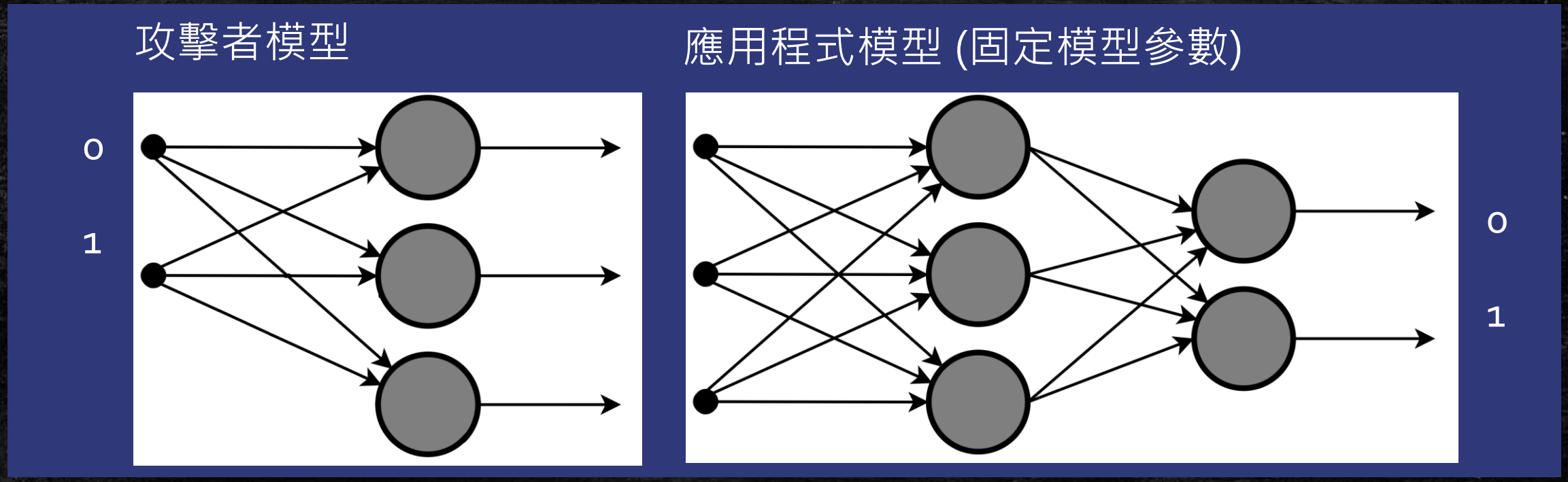


應用程式模型



解題概念說明

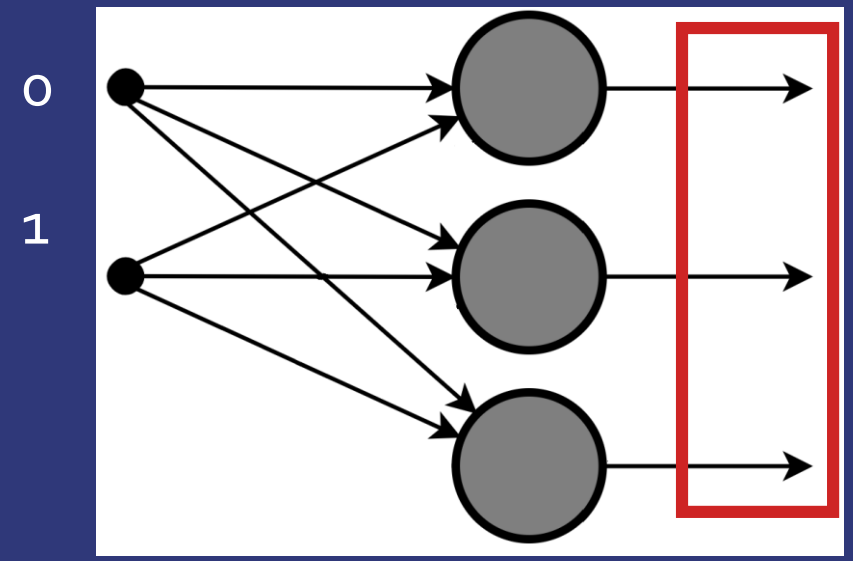
- 依照 AutoEncoder 概念，準備一堆資料讓輸入等於輸出，該輸入值訂為攻擊者想要的數值
- 固定住應用程式模型內的參數，讓其在訓練過程不會變動



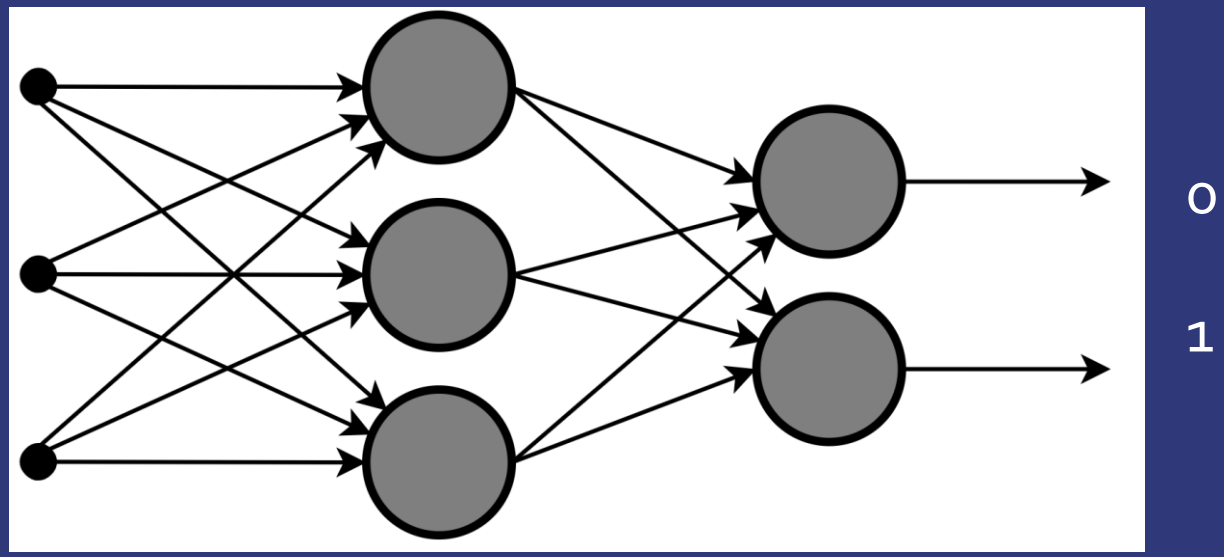
解題邏輯思考

- 如果攻擊者模型參數能夠讓 [0 1] 輸出結果為 [0 1]，就代表著把紅色框的結果灌入應用程式模型會得到 [0 1]
- 也就是說目前攻擊者模型對於 [0 1] 的輸出結果就是回推出的答案

攻擊者模型



應用程式模型 (固定模型參數)



程式實作

```
# 想辦法從模型中萃取出預設為 9 的輸入資料
# 依照 AutoEncoder 想法，建立一個輸入長度為10的模型，在串接一個 784 的 Dense
attack_vector = Input(shape=(10,))
output_vector = Dense(28 * 28, activation='relu', input_dim=10)(attack_vector)
attack_model = Model(inputs=attack_vector, outputs=output_vector)
attack_model.summary()

attack_model.save('attack_model.h5',save_format='h5')
```

```
# 把兩個模型接再一起
target_output = load_model(output_vector)
combined_model = Model(inputs=attack_vector, outputs=target_output)
combined_model.compile(loss='binary_crossentropy', optimizer=tf.optimizers.Adam())
combined_model.summary()

combined_model.save('combined_model.h5',save_format='h5')
```

結論

- 其實在學習機器學習模型安全的過程中還蠻常碰壁的，有時候多半跟數學是有點關聯的

- 再讓子彈飛一會兒，思考一下整個流程到底還缺了甚麼，等到最後再來探討這個攻擊手法的威脅程度