2024 鐵人賽 – 我數學就爛要怎麼來學 DNN 模型安全
Day 06 – DNN 模型瀏覽器

大綱

- 類神經網路模型瀏覽器
  - 背景介紹
  - 工具一覽
- 結論

背景介紹

- 類神經網路模型瀏覽器功能
  - 呈現模型結構，以便猜測模型提供功能
  - 瀏覽或是修改模型內部參數

- 對於攻擊者的重要性
  - 分析模型結構，進而利用架構去構建攻擊
  - 竄改模型參數，讓模型表現出異常行為

工具一覽

- 工具還蠻多的，這邊只介紹我用過的
  - 程式輸出
  - Netron
  - HDFView

程式輸出

- 在 Day04 就介紹過了，可透過程式讀取及設定值模型參數

```
In [4]: # 因為只有一個 layer, 所以直接顯示第 0 個 layer 的參數來看看
        print(model.layers[0].get_weights())
        print(type(model.layers[0].get_weights()[0]))
        print(type(model.layers[0].get_weights()[1]))

        [array([[ 0.03176403, -0.14284325],
               [-0.32049638,  1.0985175 ]], dtype=float32), array([0., 0.], dtype=float32)]
        <class 'numpy.ndarray'>
        <class 'numpy.ndarray'>

In [5]: # 設定單位矩陣(對角線為1), 偏移參數為 0
        weight = [ np.array([ [1,0], [0,1] ]), np.array([0,0]) ]
        model.layers[0].set_weights(weight)

In [6]: print(model.layers[0].get_weights())

        [array([[1., 0.],
               [0., 1.]], dtype=float32), array([0., 0.], dtype=float32)]
```

Netron

- 透過 pip 即可安裝，使用上相當方便
- 網頁介面，支援多種模型儲存格式
- 用圖形化方式呈現模型結構

- pip install netron
- netron

https://github.com/lutzroeder/netron

HDFView

- UI 介面，支援 Linux 及 Window 作業系統
- 提供修改瀏覽及參數的功能
- 缺點在於只能用清單去呈現模型結構

https://www.hdfgroup.org/downloads/hdfview/

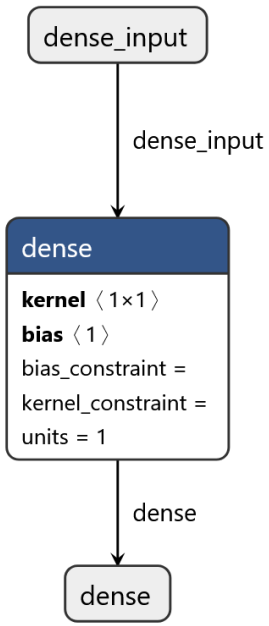
先將之前練習的數學模型匯出成檔案

In [7]:
```
# 顯示訓練後模型的參數來看看
print(old_parameter)
print(model.get_weights())
print(model.predict([10, 5, 200, 13]))
```

```
[array([[-1.609054]], dtype=float32), array([0.], dtype=float32)]
[array([[1.9992737]], dtype=float32), array([0.8359087], dtype=float32)]
1/1 [==============================] - 0s 51ms/step
[[ 20.828646]
 [ 10.832277]
 [400.69064 ]
 [ 26.826466]]
```

In [8]:
```
model.save('linear_model.h5',save_format='h5')
```

Netron



NODE PROPERTIES

dtype	float32
kernel_constraint	
kernel_initializer	GlorotUniform(null)
kernel_regularizer	
trainable	true
units	1
use_bias	true

INPUTS

input: name: **dense_input**

kernel: name: **dense/kernel:0**

tensor: `float32[1,1]`

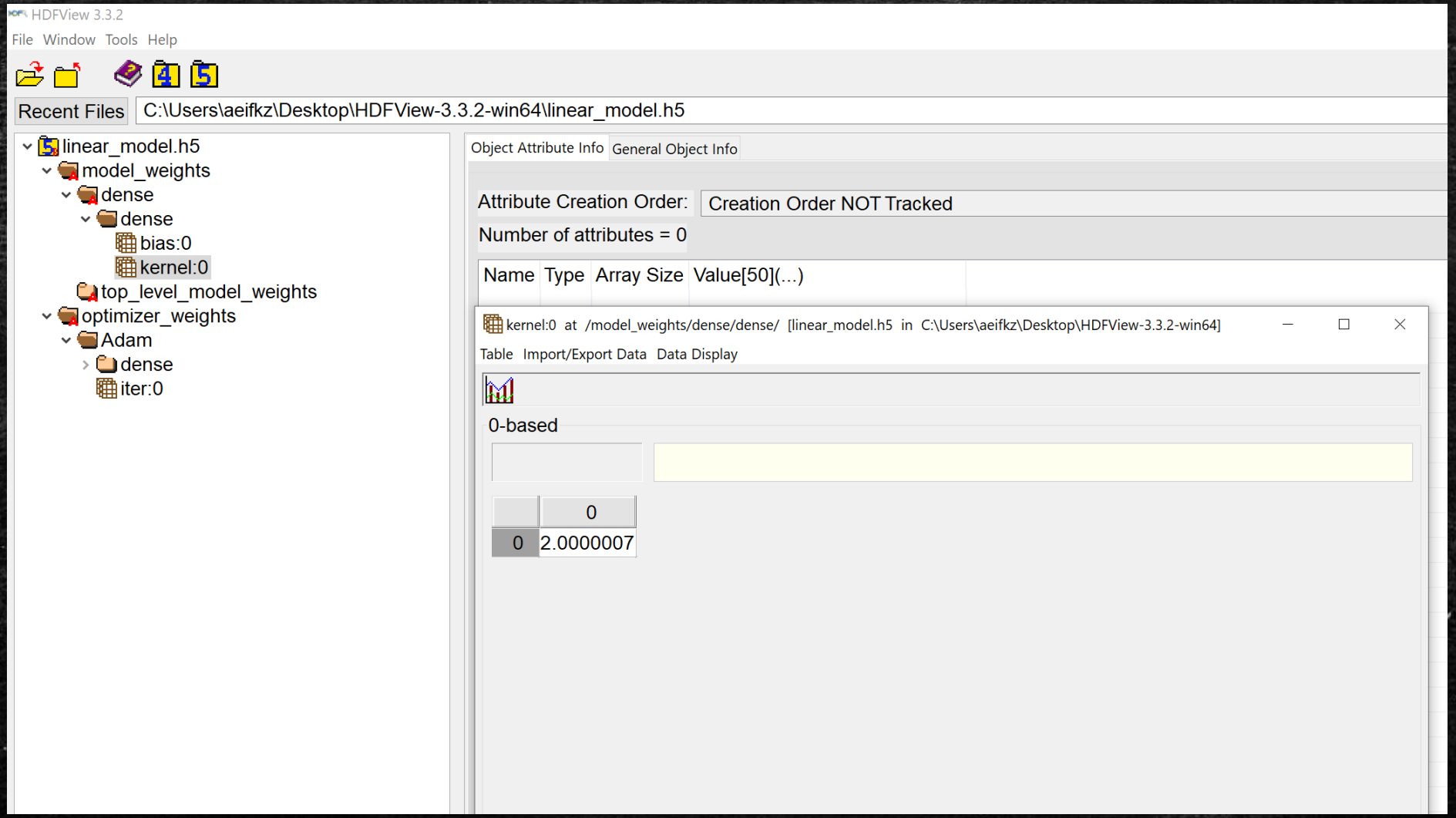
```
[
    [
        2.0000007152557373
    ]
]
```

bias: name: **dense/bias:0**

tensor: `float32[1]`

```
[
    0.9843568801879883
]
```

HDFView 3.3.2


HDFView 3.3.2
File Window Tools Help
Recent Files
C:\Users\aeifkz\Desktop\HDFView-3.3.2-win64\linear_model.h5
linear_model.h5
model_weights
dense
dense
bias:0
kernel:0
top_level_model_weights
optimizer_weights
Adam
dense
iter:0
Object Attribute Info
General Object Info
Attribute Creation Order:
Creation Order NOT Tracked
Number of attributes = 0
Name
Type
Array Size
Value[50](...)
kernel:0 at /model_weights/dense/dense/ [linear_model.h5 in C:\Users\aeifkz\Desktop\HDFView-3.3.2-win64]
Table
Import/Export Data
Data Display
0-based
0
0
2.0000007

結論

- 對於攻擊者而言類神經網路模型瀏覽器要看的東西還算單純
  - 模型結構
  - 模型參數
- 當然對於防禦者來說也是必要技能
  - 觀察後門模型結構
  - 觀察異常的參數數值