

隊名：異顏難進

組員：陳信豪(r06725048)、曾千蕙(r05725004)、郭士庭(r05725039)

題目：HTC Hand Detection

隊名：異顏難進

組員：陳信豪、曾千蕙、郭士庭

一、題目描述：

- I. 紿第一人稱視角的圖，判斷左右手及用 bounding box 標出手的位置

二、方法流程

- I. 流程架構圖

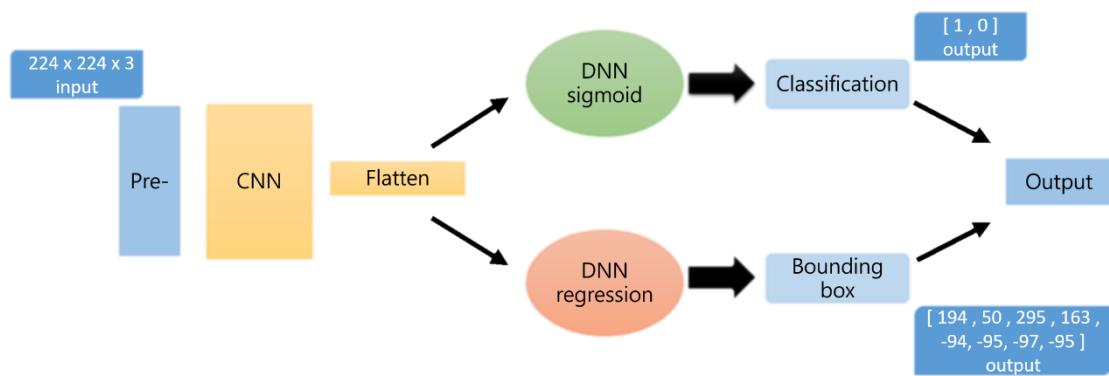
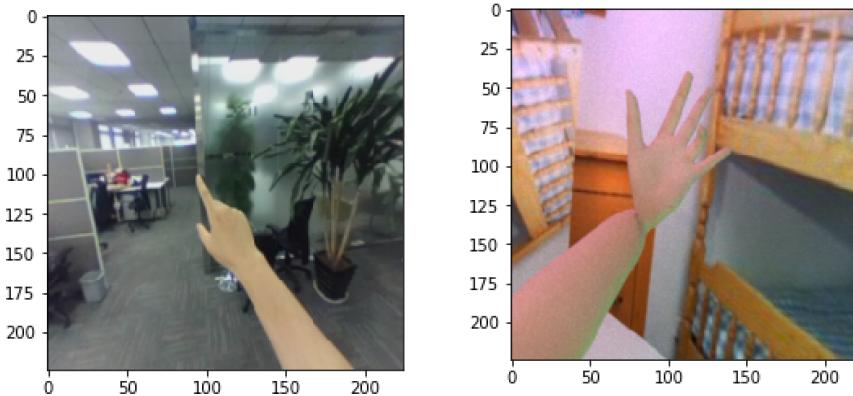


圖 1

三、前處理

- I. 一開始由於 **synth** 跟 **vive** 的圖片大小不一樣，所以會 **resize** 成同樣大小的 **224x224x3**。
- II. 有嘗試過 **rgb2gray** 變灰階 與 **rgb-d** 把圖片加深處理，但效果都沒有很好，所以最後跑資料的時候，都沒有做色彩上的處理。
- III. 範例：

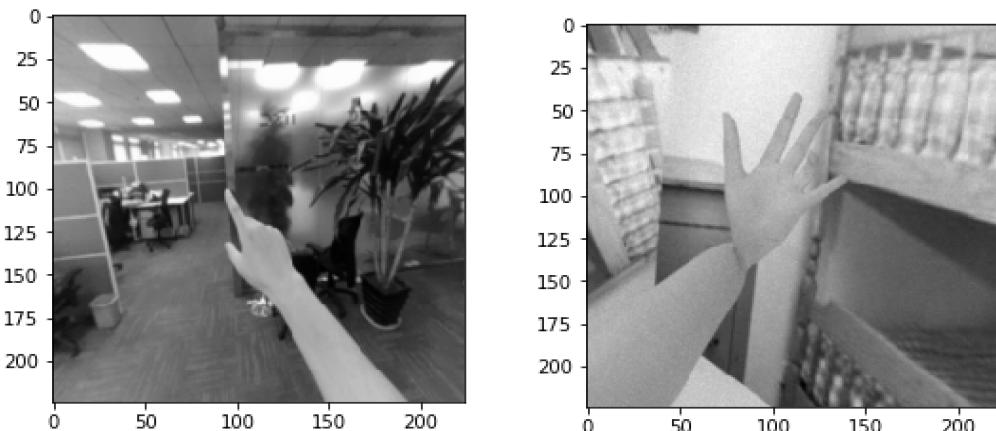
- i. 原圖



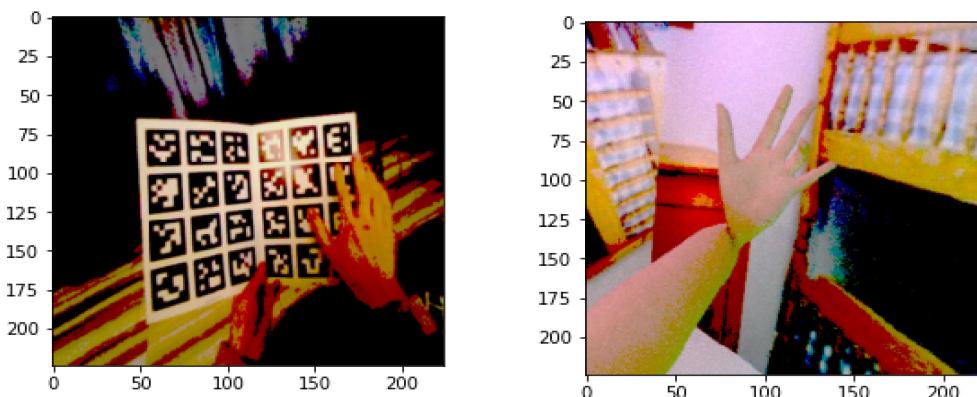
隊名：異顏難進

組員：陳信豪(r06725048)、曾千蕙(r05725004)、郭士庭(r05725039)

ii. 灰階



iii. 加深



四、嘗試過的 CNN model 架構

- I. 最一開始嘗試過 CNN + DNN，同時輸出 8 個維度的 output 包含判斷左右手，但效果極差，所以改用圖 1 的架構，將左右手分類及 bounding box 分開預測。也就是說，用一樣的 CNN 架構，會先 train 一個判斷左右手的 exist model (準確度可達 98%以上)，然後再用此 model 的參數作為初始值去 train 另一個 bounding box 的 model 。
- II. 在這之後的整體架構上都沒有修改，只有在 CNN 這塊做調整及變換
- III. 在 CNN 這部分嘗試過了傳統 CNN、ResNet50、InceptionV3、Inception ResNet V2 。
- IV. 最後我們的實驗結果是 InceptionV3 與 Inception ResNet V2 表現最好。

五、Train data 的順序

- I. model 跑 data 的流程



- II. 會先將 synth data 切成四份，然後一次只 train 1/4 的 synth 資料，最後跑完 synth 資料之後，再去 fine tune 資料很少的 vive 資料。

隊名：異顏難進

組員：陳信豪(r06725048)、曾千蕙(r05725004)、郭士庭(r05725039)

六、model 介紹與結果

I. 用 improvement tips 之前的最好成績

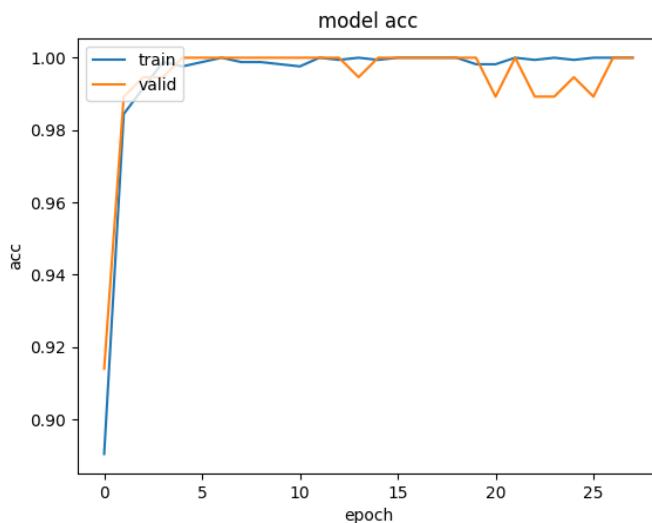
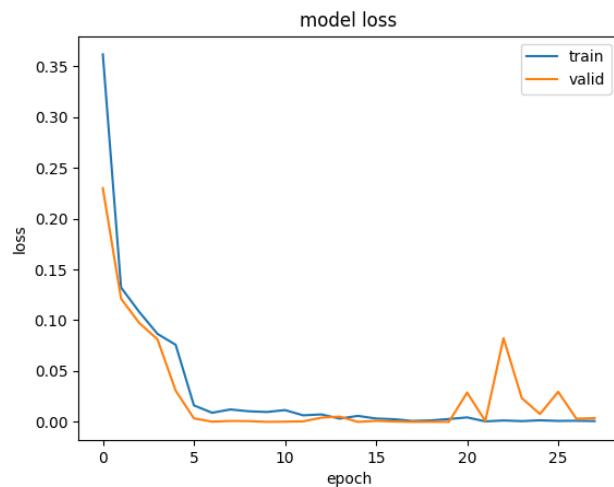
(後面會再介紹我們 improvement tips)

II. (offline judger 只有 六張圖、DeepQ online judger 有上千張)

Model	Best off-line score	Best on-line score	Parameters	Depth
CNN + DNN	0.361	0.09	38,984,072	21
ResNet50	0.18	0.11	25,636,712	168
InceptionV3	0.75	0.3	23,851,784	159
InceptionResNetV2	0.75	0.33	55,873,736	572

III. CNN + DNN (一個 DNN 負責分類，另一個負責 bounding box):

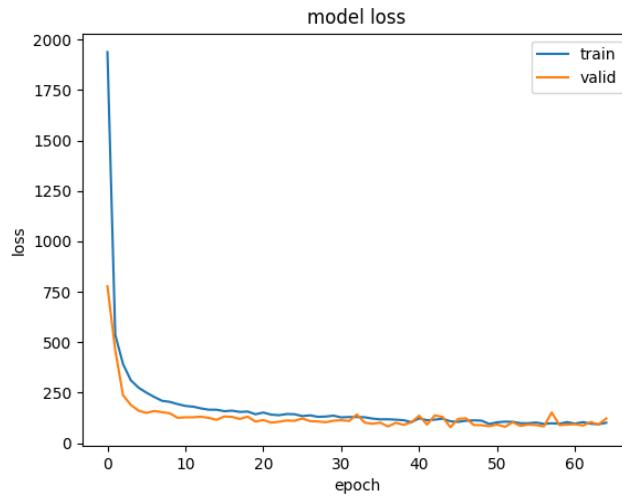
- i. 這部分的 CNN 用 4~5 層嘗試過，但效果都卡在 0.1 左右
- ii. Exist model (判斷左右手)



隊名：異顏難進

組員：陳信豪(r06725048)、曾千蕙(r05725004)、郭士庭(r05725039)

iii. Bounding box model(regression) 畫出八個位置點



IV. ResNet50 + DNN:

- 這部分使用 Keras 提供的 pre-trained weights for Classification on ImageNet 的 ResNet model，特色是每一區塊會把原本的資料跳過 convolution，並再將之加到 convolution 出來的結果，這樣可以多保留一些原始的資料特徵。

ii. 架構：

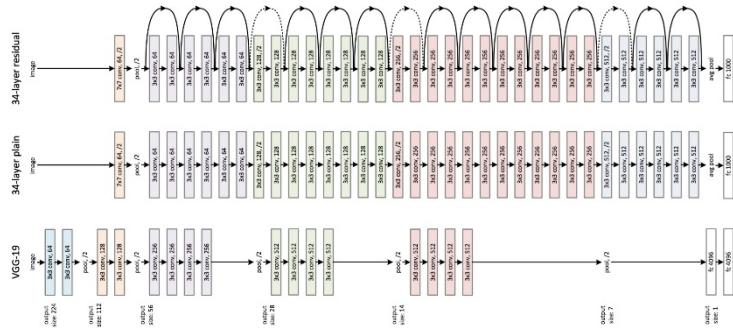


Figure 3. Example network architectures for ImageNet. Left: the VGG-19 model [4] (19.6 billion FLOPs) as a reference. Middle: a plain network with 34 parameter layers (3.6 billion FLOPs). Right: a residual network with 34 parameter layers (3.6 billion FLOPs). The dotted shortcuts increase dimensions. Table 1 shows more details and other variants.

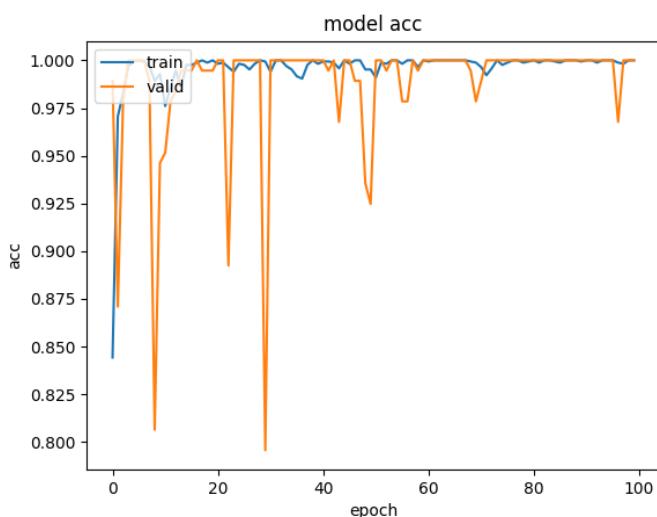
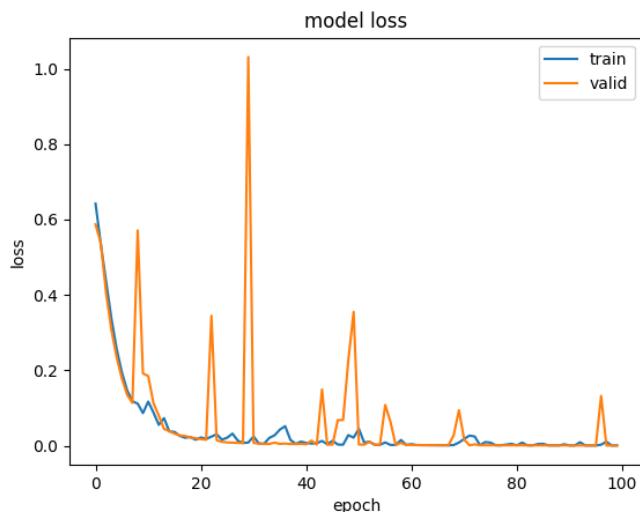
(圖片來自：<http://euler.stat.yale.edu/~tba3/stat665/lectures/lec18/img/?C=M;O=D>)

- 討論：這個 model 跑出來的結果雖然 off-line 的結果變差(0.361 -> 0.18)，但 on-line 的結果有略微的進步(0.09 -> 0.11)。而進步的幅度不明顯有可能是因為我們在嘗試 ResNet50 時，沒有將 synth 的資料 train 的很齊。

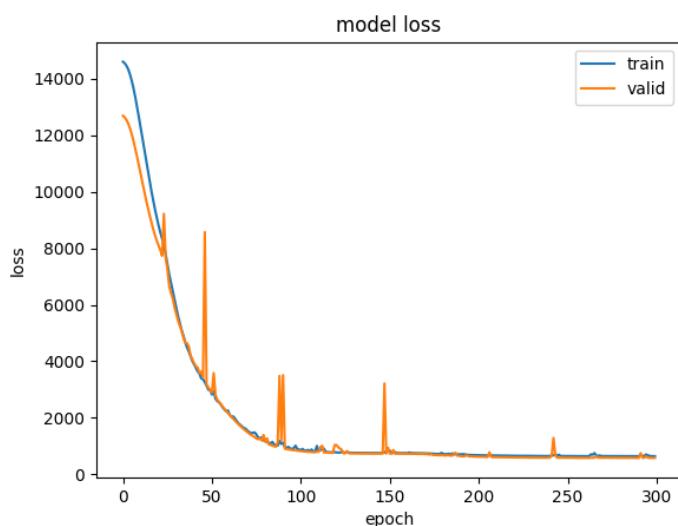
隊名：異顏難進

組員：陳信豪(r06725048)、曾千蕙(r05725004)、郭士庭(r05725039)

iv. Exist model (判斷左右手):



v. Bounding box model:

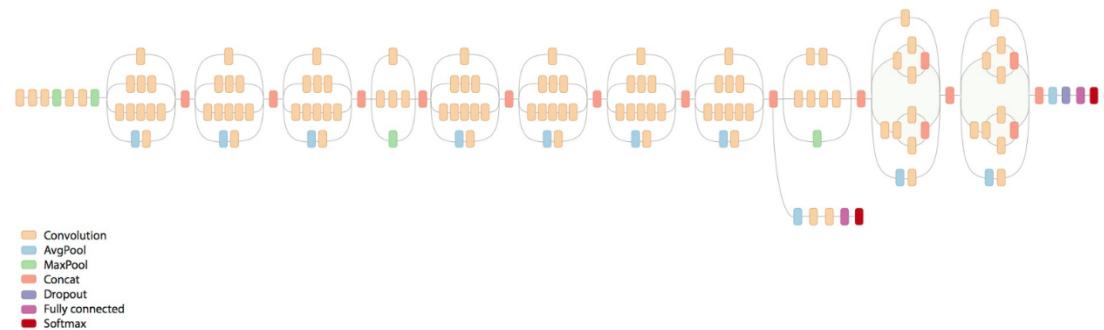


隊名：異顏難進

組員：陳信豪(r06725048)、曾千蕙(r05725004)、郭士庭(r05725039)

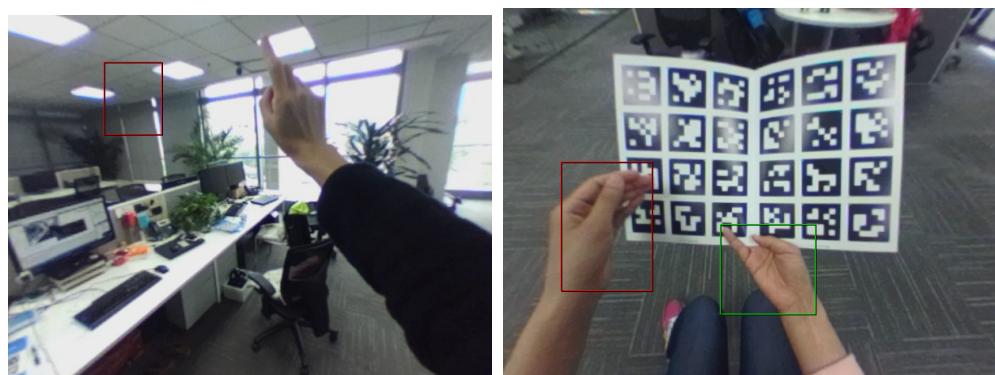
V. InceptionV3 + DNN:

- i. 介紹：這部分也是使用 Keras 提供的 ImageNet pre-trained weights 分類的 model，這 model 特色是每個區塊會平行做 convolution，最後再將結果合併。
- ii. 架構：



(原圖來自: <https://goo.gl/CS5cqC>)

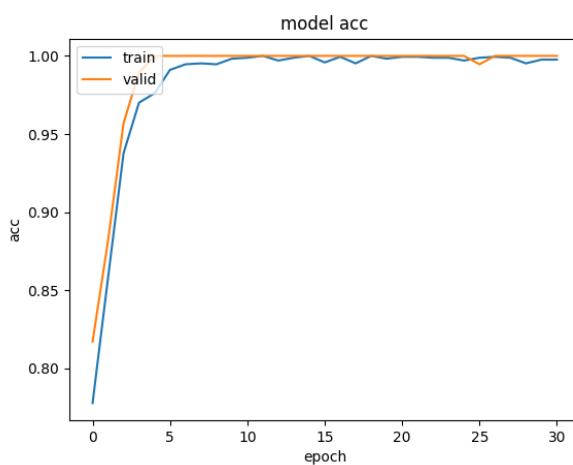
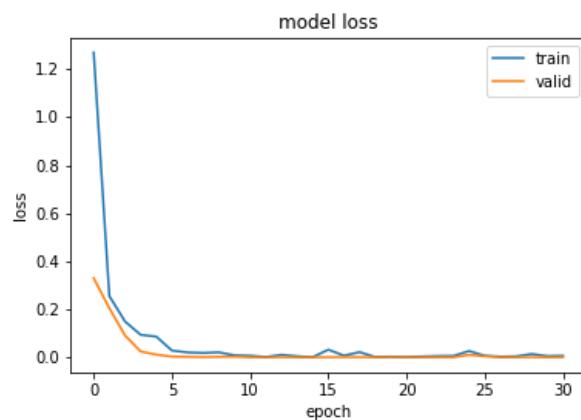
- iii. 討論：這個 model 效果就好非常多，一開始 off-line 的 judge 測試就可以到 0.5，上傳 DeepQ 的 on-line 成績也有 0.28，後來就持續用這個 model 做測試。
- iv. 輸出結果：



v. Exist model (判斷左右手):

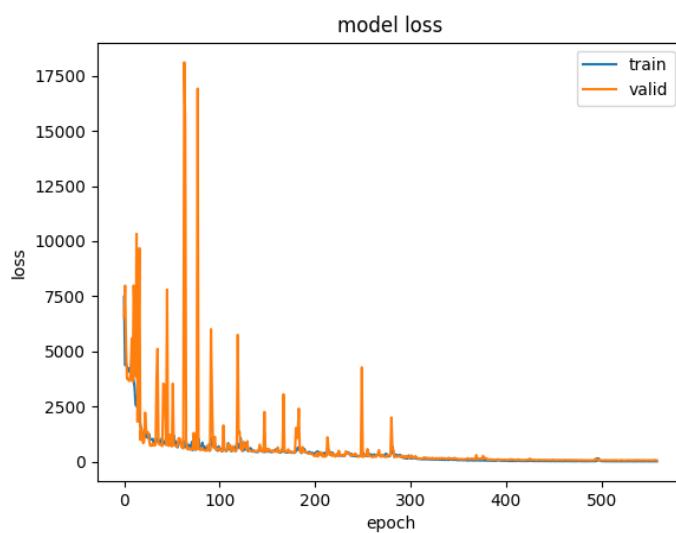
隊名：異顏難進

組員：陳信豪(r06725048)、曾千蕙(r05725004)、郭士庭(r05725039)



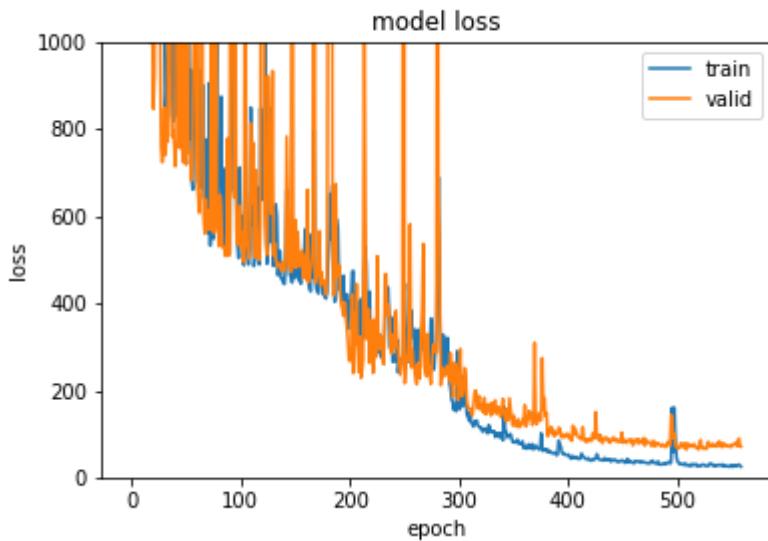
vi. Bounding box model:

(這部分的 train loss 漂浮很大，不過長久之後還是會收斂。)



隊名：異顏難進

組員：陳信豪(r06725048)、曾千蕙(r05725004)、郭士庭(r05725039)



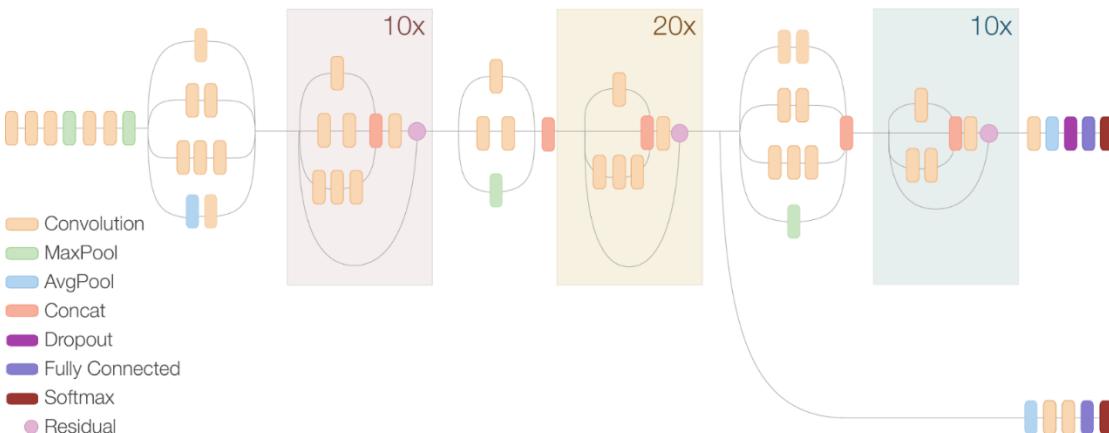
VI. InceptionResNet V2 + DNN:

- i. 介紹：這也是 Keras 提供的 model，它的特色就是結合 Inception 與 ResNet 的架構
- ii. 架構：

Inception Resnet V2 Network



Compressed View



(原圖來自: <https://goo.gl/PNjtV0>)

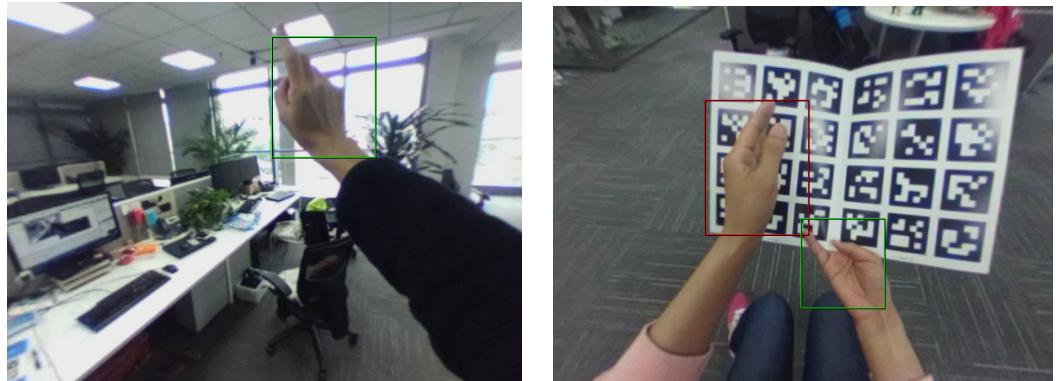
- iii. 討論：這個方法在文獻測試上的效果是比 Inception 和 ResNet 好的，但是由於我們在跑這個 model 時，時間不夠，所以 synth 的資料只有拿 1/4 去 train，剛開始跑出來的效果，雖然 off-line 的 judge 成績有到 1.0 的高分，但上傳 DeepQ 的成績不如 InceptionV3。不過用了 Improvement Tips，並開始多 train 幾回之後，效果就漸漸超越 InceptionV3 了，我們猜測是因為這個 model 的能力結合了 Inception

隊名：異顏難進

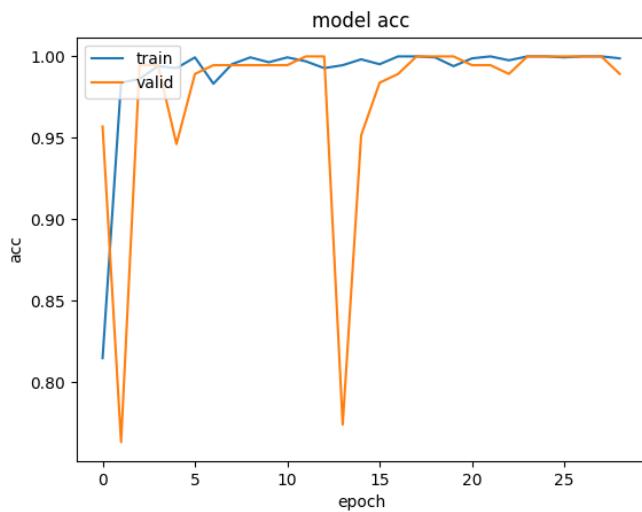
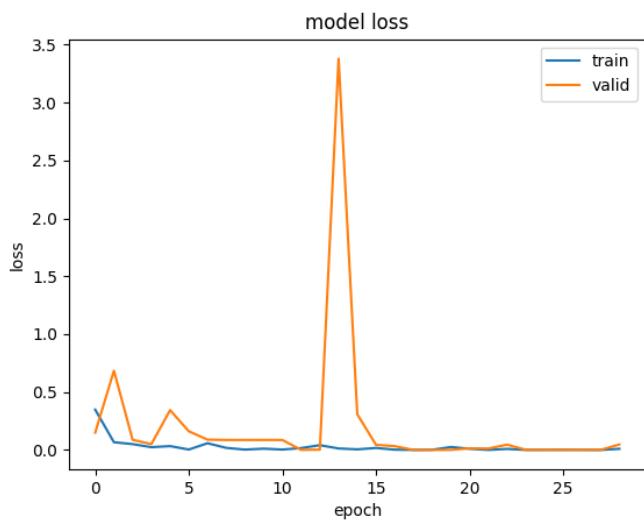
組員：陳信豪(r06725048)、曾千蕙(r05725004)、郭士庭(r05725039)

和 ResNet，所以效果比較好，且不用 train 到那麼多的 synth 資料即可超越 InceptionV3 的效能。

iv. 輸出結果：



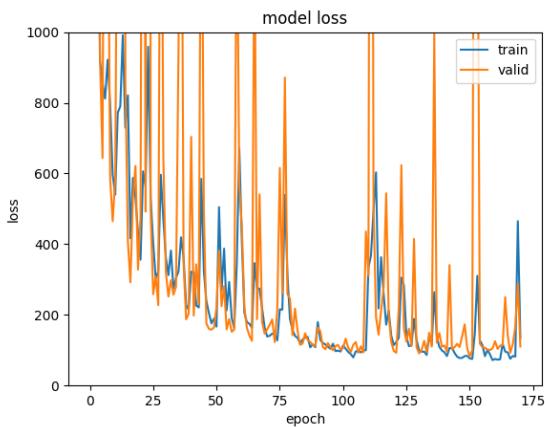
i. Exist model(判斷左右手):



ii. Bounding box model

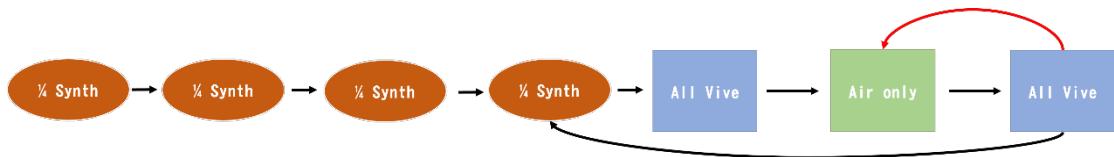
隊名：異顏難進

組員：陳信豪(r06725048)、曾千蕙(r05725004)、郭士庭(r05725039)



七、Improvement Tips

- I. CNN model 的替換，替換上述介紹的 CNN model 來提高效果。
- II. 替換 CNN model 後最好的結果仍只有到 0.3，觀察後發現，book 資料的準確度蠻高的，但 air 很低，我們猜測可能是 air 的背景太亮的關係，所以我們嘗試用 opencv 做一些圖片的加深或調暗的處理，然後再丟進 train 過 synth 的 model 來 train，但效果反而變差了，所以放棄這部分的嘗試。
- III. **Data training order:** 由於 air 的辨識率極低，所以我們就嘗試多 train 幾次只有 air 的 vive 資料，然後再丟全部的 vive 資料下去 train，這樣的效果有明顯的上升一些，而我們又重複了這個部分，就是 **1/4 synth -> vive -> air only -> vive** (下圖中黑色的路徑) 然後 testing。每次跑完一個循環，成績就都會有稍稍提升，但大概在第四次循環左右就沒有提升了，後來就又改成下圖中紅色那條路徑，就是只有 train air only 跟 vive，沒有再回去 train synth，結果又開始上升了！目前 InceptionV3 跟 Inception ResNet V2 的 model 都有因為每多 train 一個循環就提升一些分數。



八、最後的結果與討論：

- I. 成績結果：

Model	Best off-line score	Best on-line score
InceptionV3	1.0	0.4749
InceptionResNetV2	1.0	0.555

隊名：異顏難進

組員：陳信豪(r06725048)、曾千蕙(r05725004)、郭士庭(r05725039)

- II. 討論：最後最佳的結果是用 Inception ResNet V2 的 model 按照前面所說的 data training order 來重複 train。

九、改進與檢討：

- I. 如果時間允許應該多跑很多次循環，直到收斂，但由於比賽時間截止，所以目前結果只有到這裡。
- II. 嘗試其他的 model，例如 YOLO [1], faster RCNN [2] 等等
- III. 我們的 model 是以整張圖片作為 input，在 output 時一起預測左手和右手的 bounding box。然而其他類型的 bounding box model 會先將圖片切成 grid，對 grid 預測圖片類別以及 bounding box，或是先產生可能的 bounding box，再對 bounding box 做處理以及預測，也就是說，我們可以嘗試將圖片的局部作為 inputs，相較於將整張圖片直接作為 inputs 而言，雜訊可能比較少，model 的負擔可能也比較小，所需的參數量也可以減少，這或許是一個我們可以前進的方向。
- IV. 針對 synth 的圖片再做優化，使用 CycleGAN 的架構，或參考[3]，生成出更擬真的照片再用於訓練。

十、參考文獻：

- [1] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 779-788).
- [2] Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems* (pp. 91-99).
- [3] Shrivastava, A., Pfister, T., Tuzel, O., Susskind, J., Wang, W., & Webb, R. (2016). Learning from simulated and unsupervised images through adversarial training. *arXiv preprint arXiv:1612.07828*.