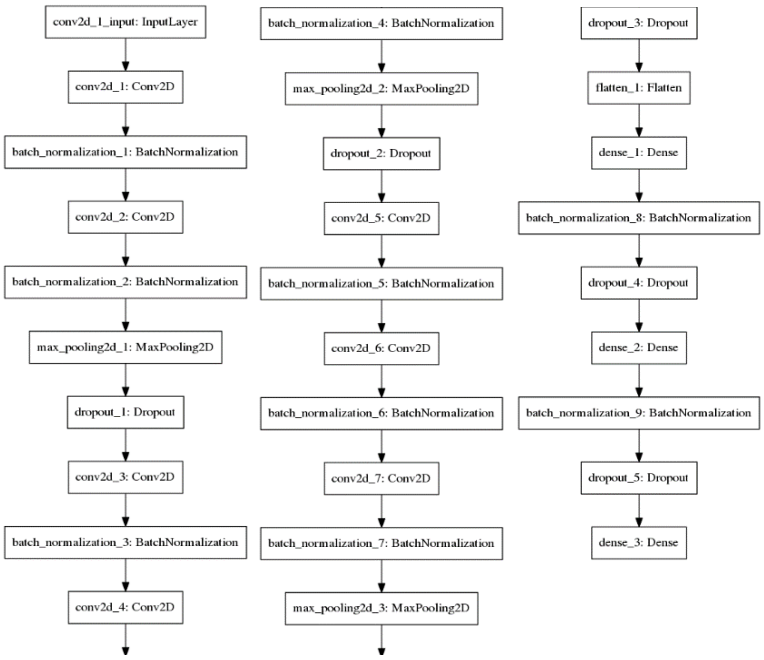
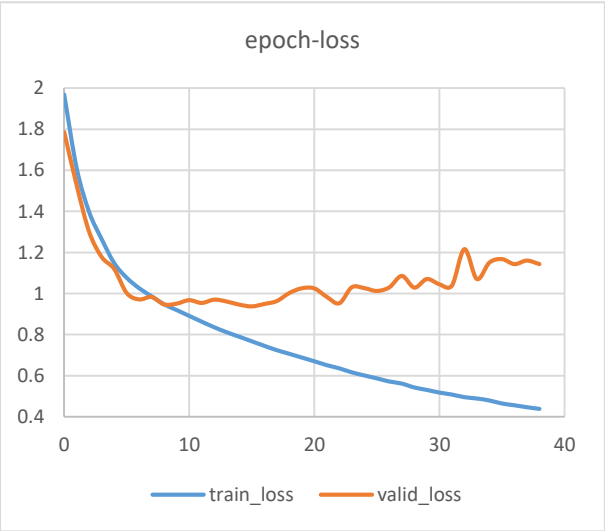
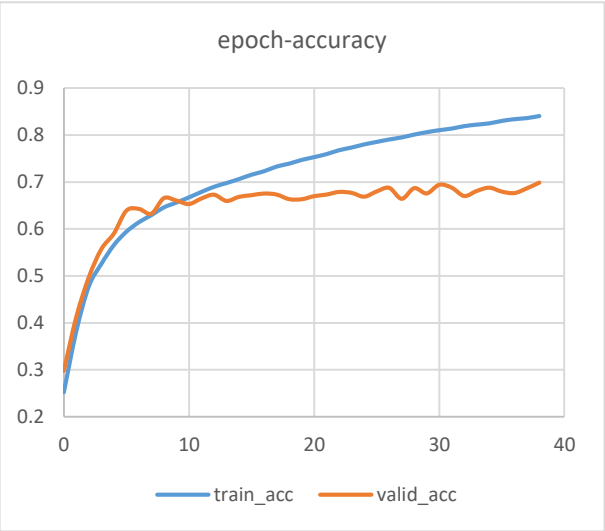
學號：B03901145 系級： 電機四 姓名：郭恆成

1. (1%) 請說明你實作的 CNN model，其模型架構、訓練過程和準確率為何？ (Collaborators: )



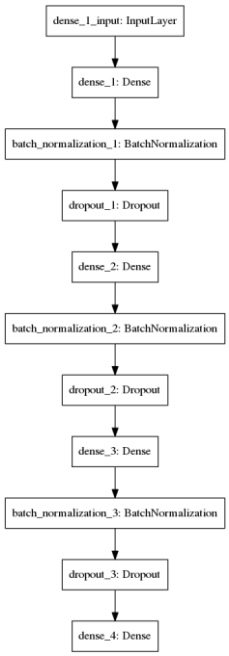




我的 CNN 模型分為四個 block，每個 block 之間都有 maxpooling 跟 dropout。block 裡有數個 unit 每筆資料都會經過 batch normalization 跟 ReLU 再傳到下一個 unit。

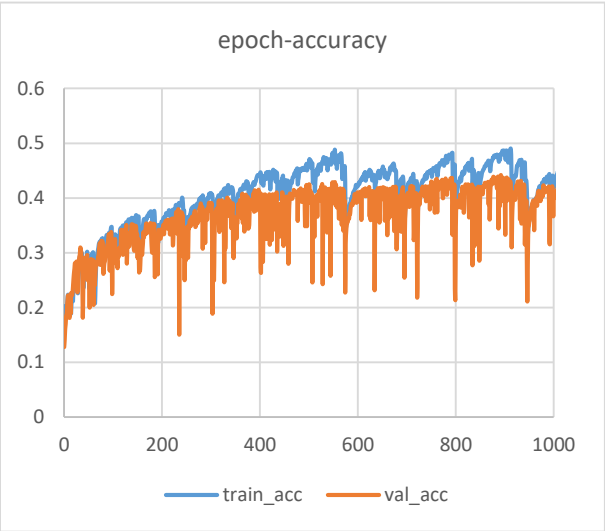
block1: Convolutional layer (64 個 3x3 的 filter) x 2
block2: Convolutional layer (128 個 3x3 的 filter) x 2
block3: Convolutional layer (256 個 3x3 的 filter) x 3
block4: Fully connected layer (output size: 512、512、7)

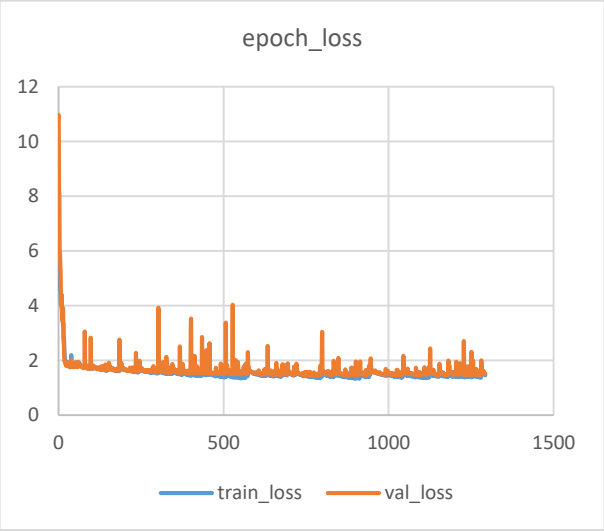
而訓練過程中我設定當 valid accuracy 10 個 epoch 都不再上升就終止。從 loss 跟 accuracy 的途中可以發現 training data 在終止前，loss 跟 accuracy 分別都有穩定的下降與上升。但對於 validation data，loss 在第 10 個 epoch 降到最低點，之後就緩緩上升，accuracy 也從此在 0.66 跟 0.7 之間震盪，第 38 個 epoch 達到終止前的最大值。

2. (1%) 承上題，請用與上述 CNN 接近的參數量， 實做簡單的 DNN model。其模型架構、訓練過程和準確率為何？試與上題結果做比較，並說明你觀察到了什麼？
(Collaborators: )

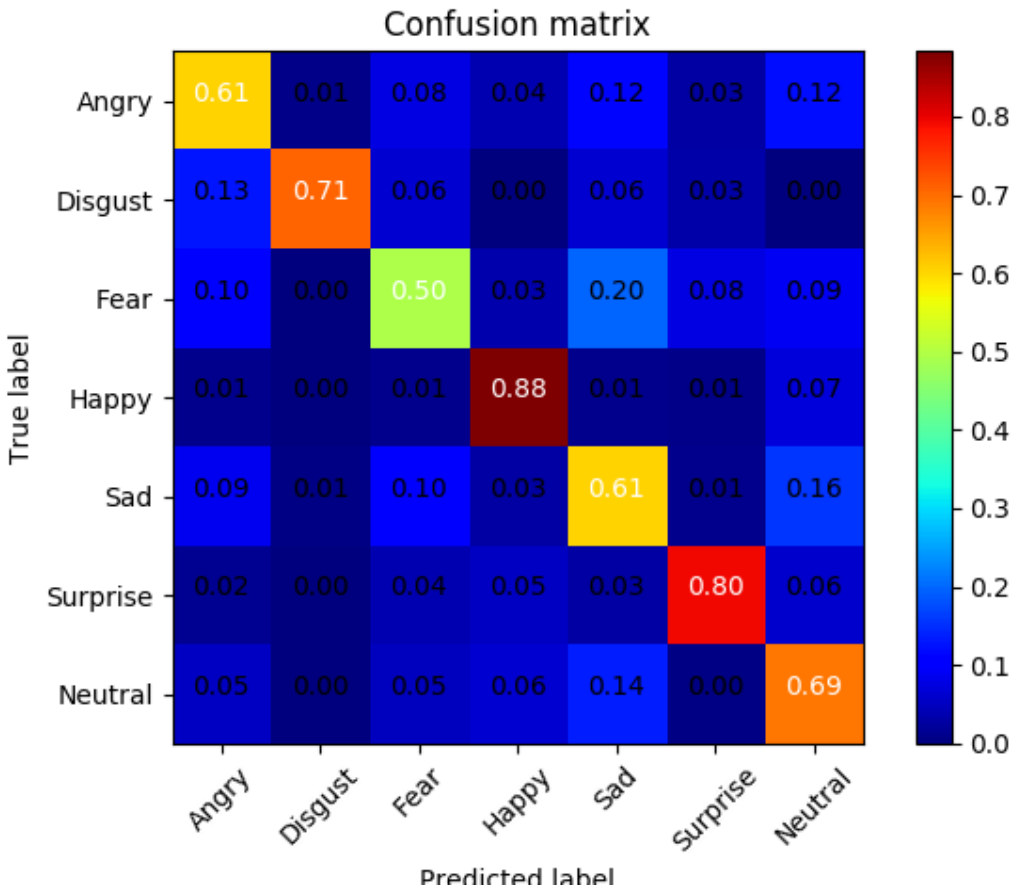



我的 DNN 由 4 個 Fully connected layer (output size: 2048、1024、512、7)，每層之間都有 ReLU、batch normalization、dropout。CNN 與 DNN 的參數量分為 6,728,391、7,361,543。而由下圖的 accuracy 跟 loss 可以發現，兩者都並未收斂且震盪劇烈，而且最佳的 validation accuracy 僅有 0.42。推測原因有兩個，一個是 DNN 不適合圖片的分類，因為圖片中有太多跟目標無關的資訊，不取出局部且重要的特徵，很難讓 model 知道要調整參數的方向，造成不斷震盪。其二是我在 DNN 沒有用 image generator 去前處理資料造成 model 容易 overfit，validation 的 accuracy 才會這麼低。但是考慮到 DNN model 在圖片分類的領域 loss 容易震盪，加上 image generator，更不容易收斂，可以預測，training accuracy 會下降，而 validation accuracy 也不一定會上升。



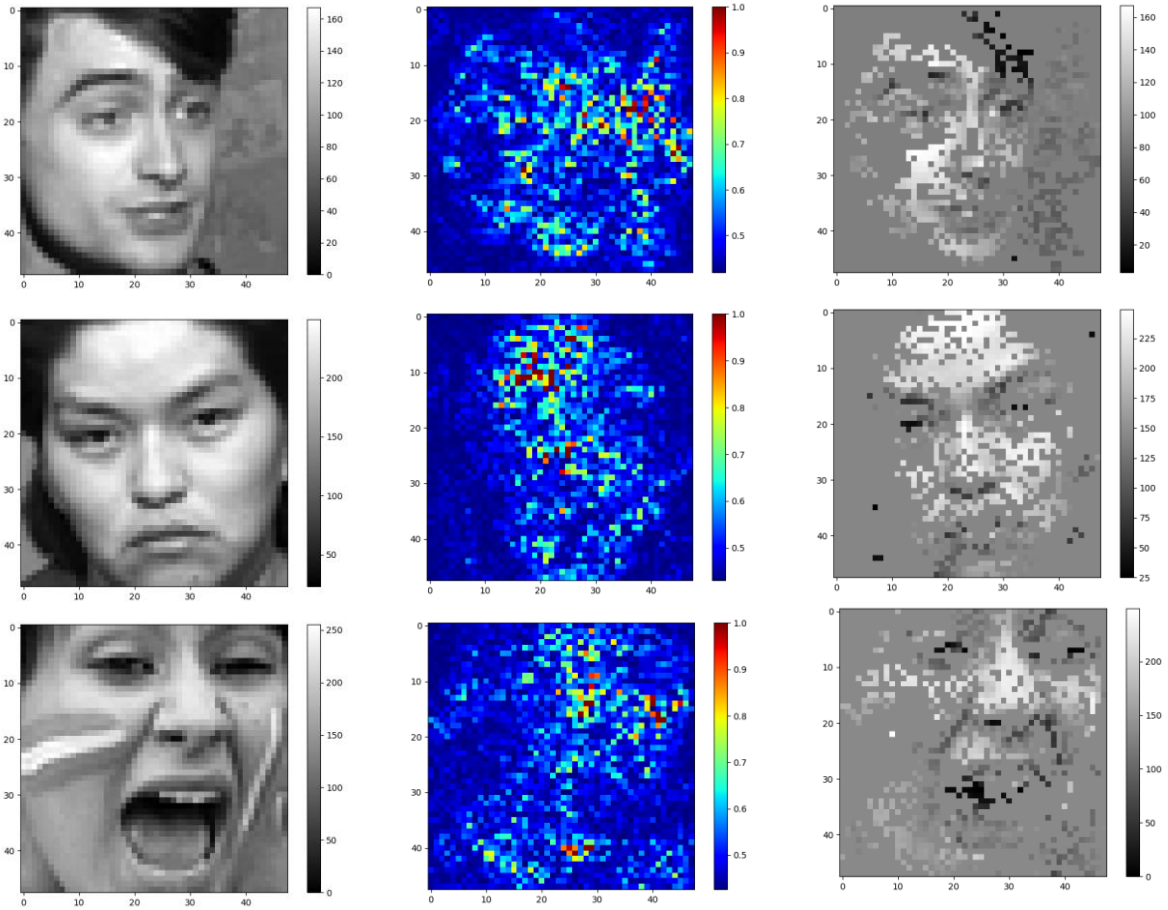



3. (1%) 觀察答錯的圖片中，哪些 class 彼此間容易用混？[繪出 confusion matrix 分析]
(Collaborators: )



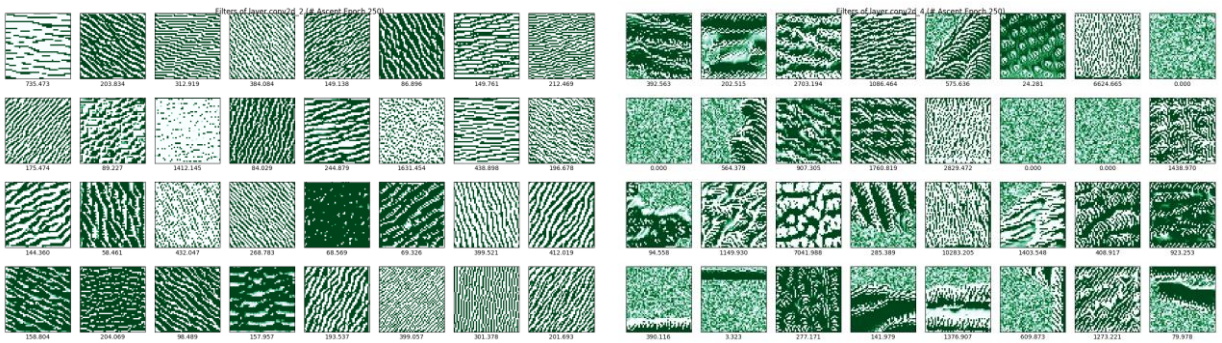

可以看到 happy 跟 surprise 的辨識率最高，推論是嘴型最明顯，最容易辨認。
而 fear 跟 sad 最容易跟其他特徵搞混，因為 fear 跟 sad 沒有特定的臉型。

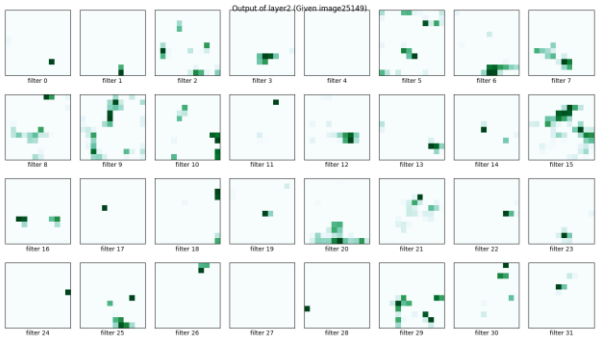
4. (1%) 從(1)(2)可以發現，使用 CNN 的確有些好處，試繪出其 saliency maps，觀察模型在做 classification 時，是 focus 在圖片的哪些部份？
(Collaborators: )



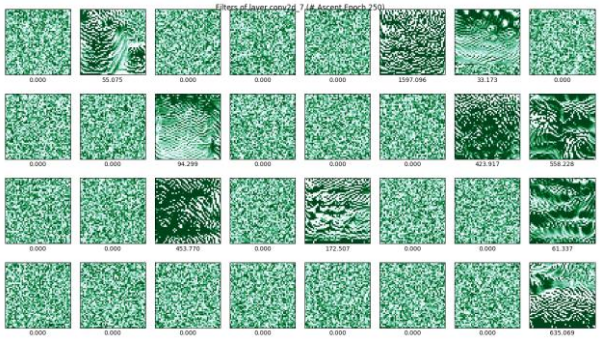
由上列不同表情的圖片作樣本，可以發現，眼睛、嘴巴、鼻子判斷表情的三個重要的特徵，在經過 CNN 的時候會被 layer 所強調，藉此做到表情的分類。

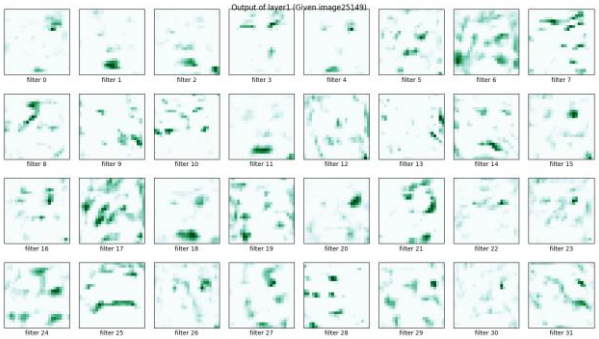
5. (1%) 承(1)(2)，利用上課所提到的 gradient ascent 方法，觀察特定層的 filter 最容易被哪種圖片 activate。
(Collaborators: )

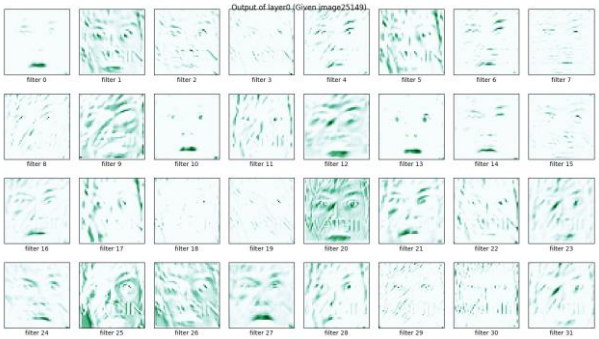





抽三個由淺到深的 Conv. layer 前 32 個 filter 出來看，可以看到最前面的 layer (左上)大多是簡單的線條，只是粗細方向不大相同。而中間的 layer (右上)則可以組成較圓滑的特徵，類是山脈、波浪、葉脈等形狀，感覺已經可以組成臉部各種紋路。而最後的 layer 大多數都是向電視雜訊般，人眼已經難以判斷的特徵，而尚可以判斷的跟中間的 layer 有很大的相識度。






可以看到越前面的 feature 跟人臉相識度越高，而中間的 layer 明顯保留了眼睛、嘴巴兩個人臉重要的特徵。最後的 layer 只剩下零星幾個點，大多是形成高階的 feature，已經難以值觀解讀。