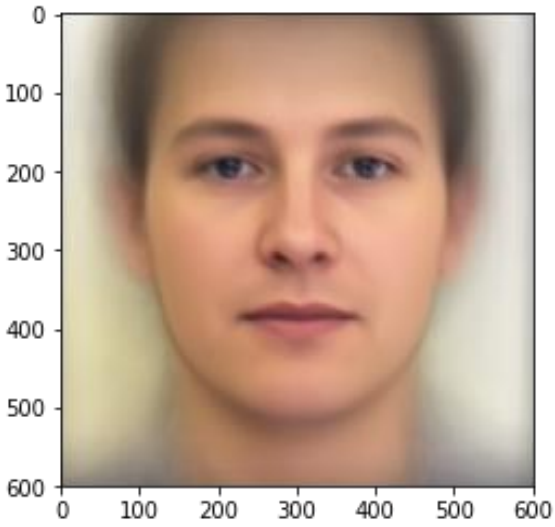
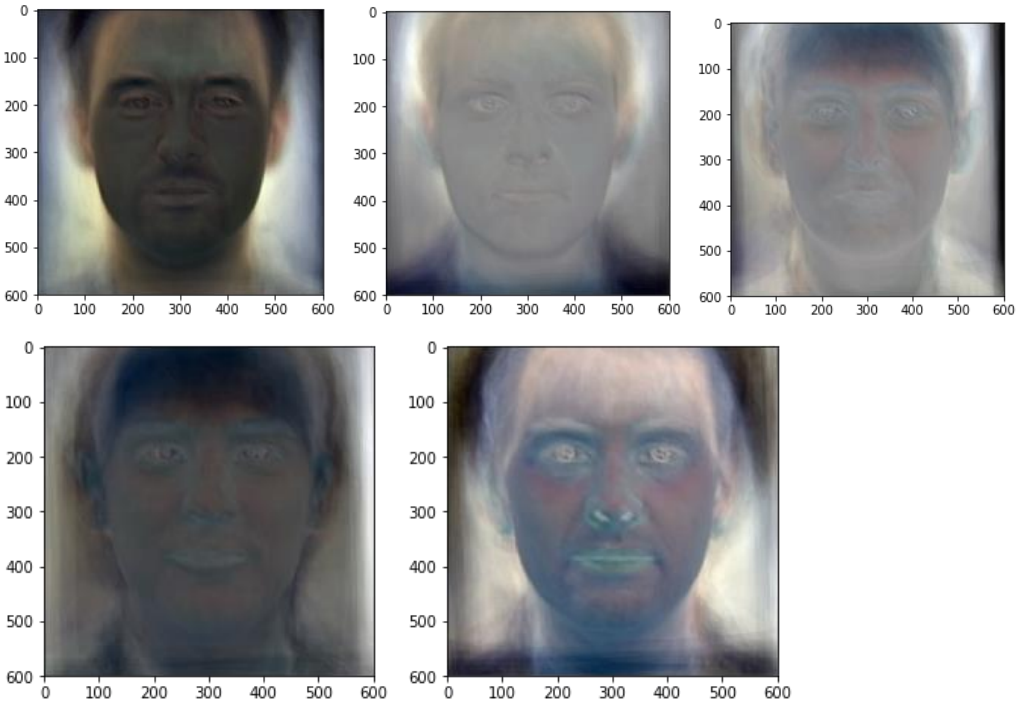
Machine Learning HW7 Report

學號：R07922108　系級：資工碩一　　姓名：陳鎰龍

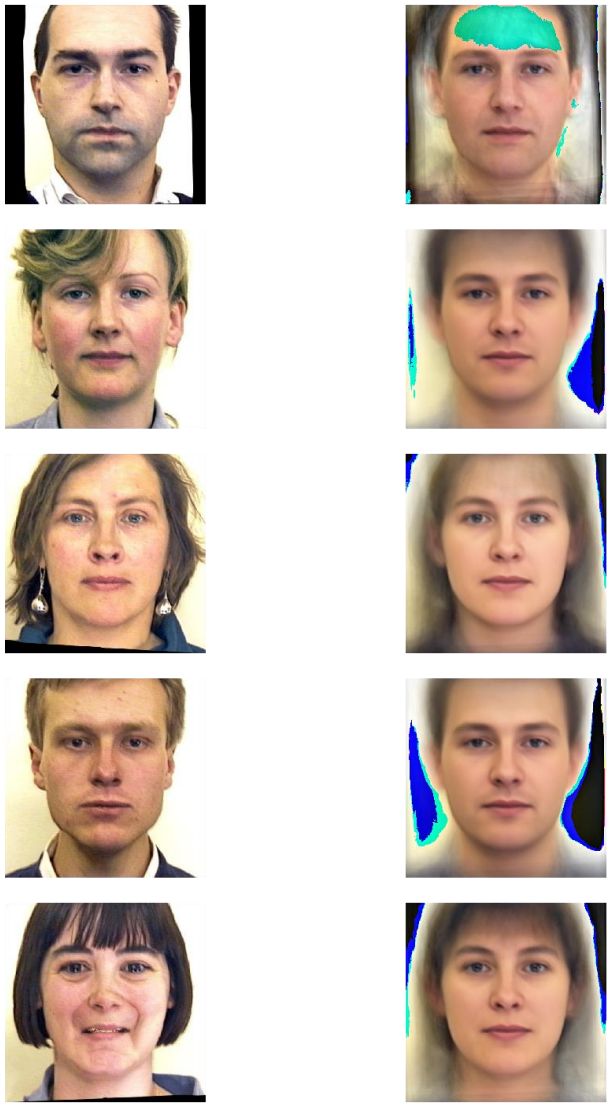
1. PCA of color faces:
   a. 請畫出所有臉的平均。



   b. 請畫出前五個 Eigenfaces，也就是對應到前五大 Eigenvalues 的 Eigenvectors。



   c. 請從數據集中挑出任意五張圖片，並用前五大 Eigenfaces 進行 reconstruction，並畫出結果。



d. 請寫出前五大 Eigenfaces 各自所佔的比重，請用百分比表示並四捨五入到小數點後一位。

13.8%

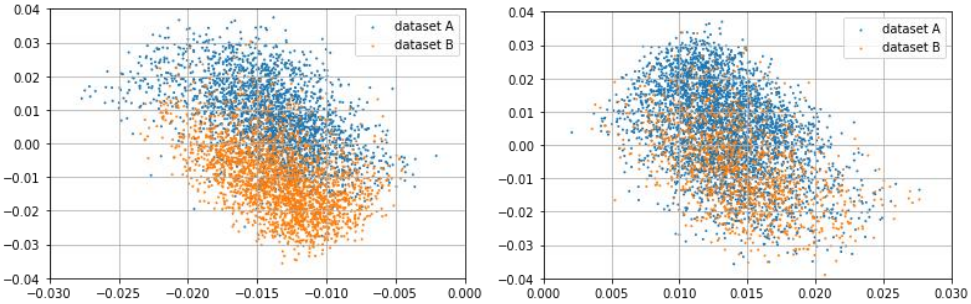
2. Image clustering:

a. 請實作兩種不同的方法，並比較其結果(reconstruction loss, accuracy)。(不同的降維方法或不同的 cluster 方法都可以算是不同的方法)

我利用不同的 cluster 的方法，一個是用 Kmeans、另一個是用 threshold 設為 0.5 的 cosine similarity，reconstruction loss 由於用一樣的架構都會一樣，我比較兩張圖片的 mean absolute error，最後 reconstruction loss 大約在 7.477，而利用不同的 cluster 能夠達成的 accuracy，kmeans 可以達到 0.955，但是 cosine similarity 只能到達 0.59，可能是我 threshold 設的不太好。

b. 預測 visualization.npy 中的 label，在二維平面上視覺化 label 的分佈。(用 PCA, t-SNE 等工具把你抽出來的 feature 投影到二維，或簡單的取前兩維 2 的 feature) 其中 visualization.npy 中前 2500 個 images 來自 dataset A，後 2500 個 images 來自 dataset B，比較和自己預測的 label 之間有何不同。

左邊是直接用 PCA 取前兩個 features 並且在已知前 2500、後 2500 筆資料各為一個 dataset 時所繪製出來的 features 分布圖，而右邊則是用自己的方法預測出來的 labels 再取前兩維 features 所繪製出來的圖，只能說右邊兩個 dataset 有上下分布的趨勢，符合左邊的圖，但實際上還是滿混雜的，只取兩維還是無法明確的區分兩 dataset，另外右邊 dataset A 與 B 是我根據大致分布來決定的。



c. 請介紹你的 model 架構(encoder, decoder, loss function...)，並選出任意 32 張圖片，比較原圖片以及用 decoder reconstruct 的結果。

我將圖片 Flatten 成 32 * 32 * 3 維後，接一層 2048 個 node 的 Dense，最後再將這 2048 維的向量經過 PCA 後取前 512 個 eigenvalue 最後兩張圖片 pixel 的平均差異(loss)約為 7.477。下圖上面為原圖，下面為 reconstruct 後的圖。