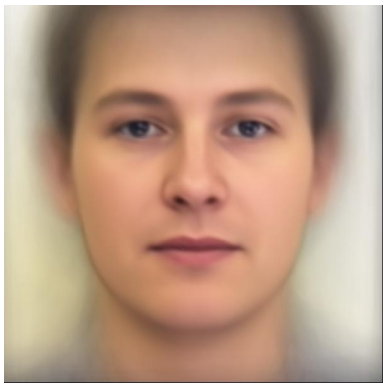
Machine Learning HW7 Report

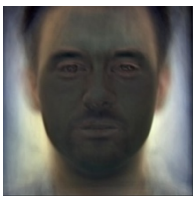
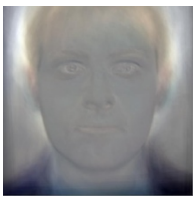
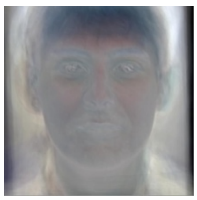
學號：B05901068 系級：電機三 姓名：蕭如芸

1. PCA of color faces:

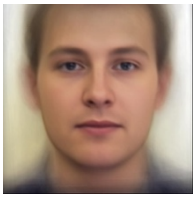
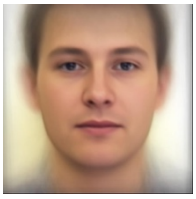
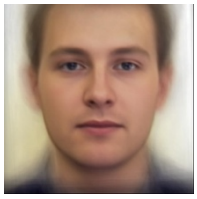
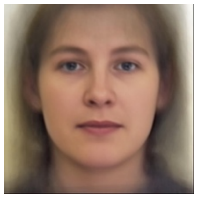
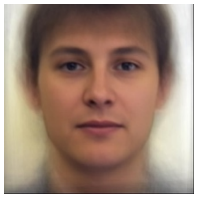
   a. 請畫出所有臉的平均。



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

1	2	3	4	5
				

   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。

3.jpg	14.jpg	159.jpg	265.jpg	358.jpg
				

   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。

   [1] 4.1%　[2] 2.9%　[3] 2.4%　[4] 2.2%　[5] 2.1%

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)

方法一：使用 autoencoder + PCA 降維，用 K-means 做 cluster

先用 autoencoder 將圖片降至 256 維，再用 PCA 將 256 維的資料降至 128 維，最後用 K-means 將資料分成兩類。

Reconstruction loss: 0.01887

Accuracy: Public 0.96910, Private 0.96892

方法二：使用 PCA 降維，用 K-means 做 cluster

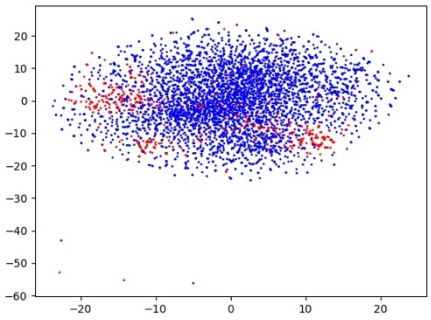
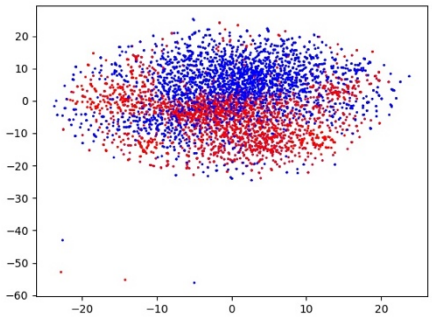
直接用 PCA 將圖片降至 128 維，再用 K-means 將資料分成兩類。

Reconstruction loss: 0.01934

Accuracy: Public 0.94892, Private 0.94858

PCA 和 K-means 使用 sklearn 的套件。方法一的 accuracy 較高，reconstruction loss 也較低，可見使用 autoencoder 是有一些幫助的。

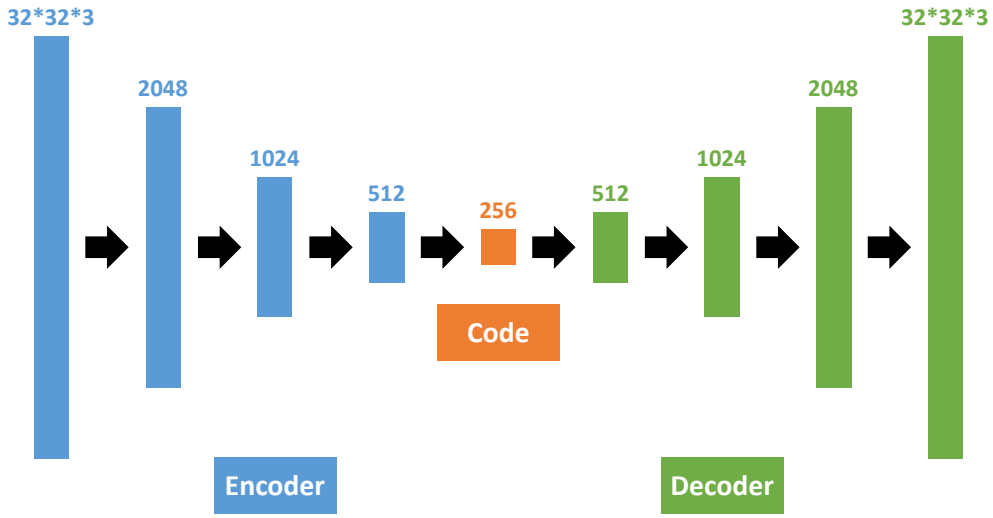
b. 預測 visualization.npy 中的 label，在二維平面上視覺化 label 的分佈。(用 PCA, t-SNE 等工具把你抽出來的 feature 投影到二維，或簡單的取前兩維的 feature)
其中 visualization.npy 中前 2500 個 images 來自 dataset A，後 2500 個 images 來自 dataset B，比較和自己預測的 label 之間有何不同。

Predicted	Ground truth
	

預測的 label 不太準，有不少 dataset A 的 images 被歸類為 dataset B。

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。

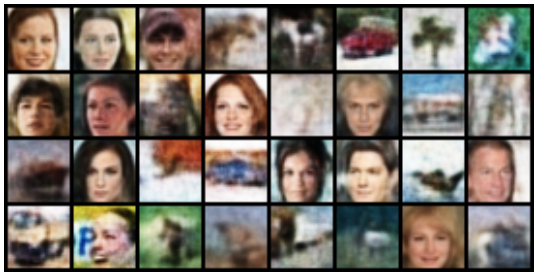
Autoencoder 架構：



Loss function: mean squared error

Training config: number of epochs = 100, batch size = 128

Optimizer: Adam (lr = 5e-4)

Original image	Reconstructed image
	

Reconstruct 的圖片比較模糊，但和原圖片的相似度蠻高的，可以辨認的出來。