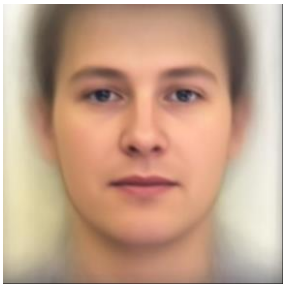
HW4

學號：r06942143 系級：電信碩一 姓名：籃聖皓

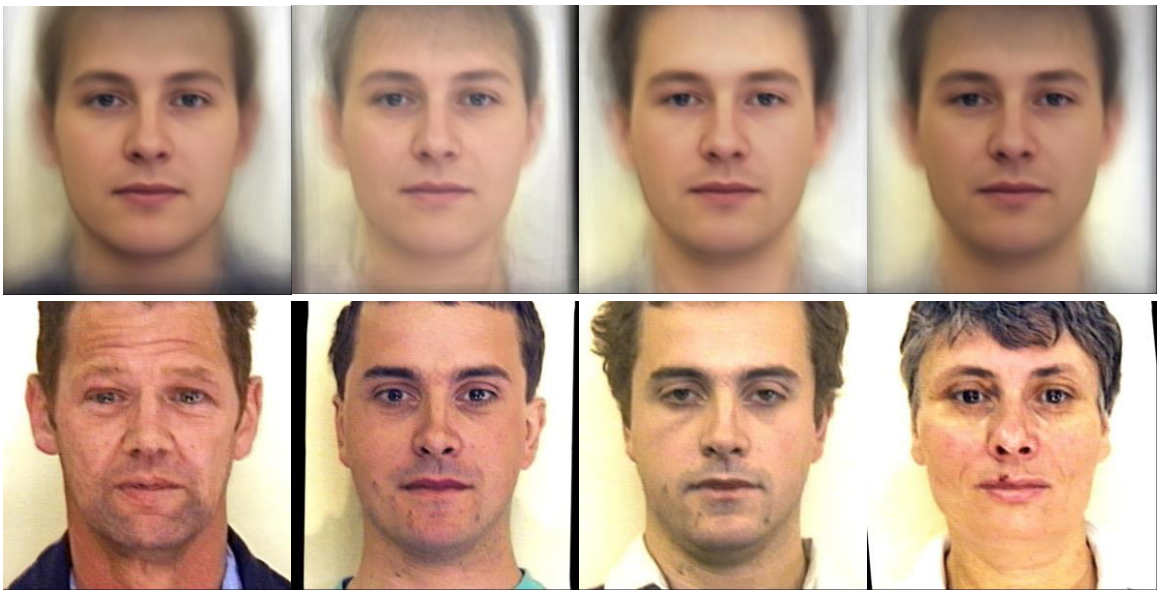
A. PCA of colored faces

A.1. (.5%) 請畫出所有臉的平均。



A.2. (.5%) 請畫出前四個 Eigenfaces，也就是對應到前四大 Eigenvalues 的 Eigenvectors。

A.3. (.5%) 請從數據集中挑出任意四個圖片，並用前四大 Eigenfaces 進行 reconstruction，並畫出結果。



A.4. (.5%) 請寫出前四大 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。

[0.0414625 0.02948732 0.0238771 0.02207842] =
4.1%、2.9%、2.4%、2.2%

B. Image clustering

B.1. (.5%) 請比較至少兩種不同的 feature extraction 及其結果。(不同的降維方法或不同的 cluster 方法都可以算是不同的方法)

使用了兩種的降維方式

	Autoencoder + Kmeans	PCA + Kmeans
Private score	0.99009	0.99836
Public score	0.99012	0.99836

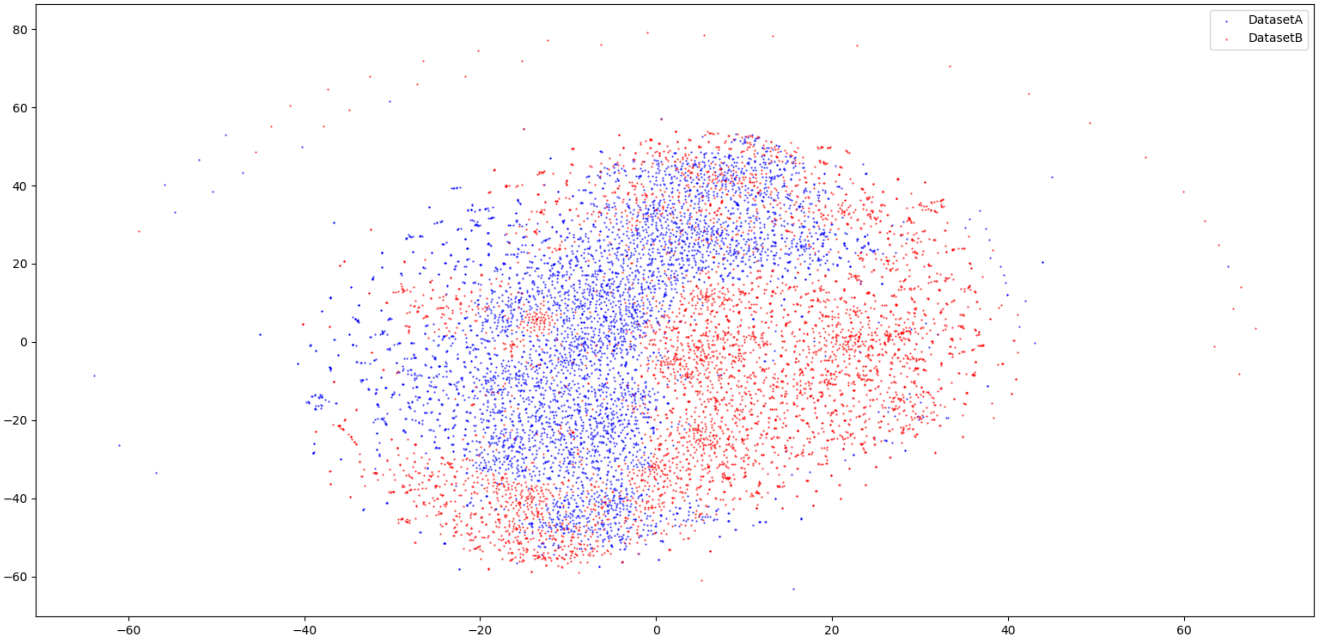
第一種是使用 Autoencoder + Kmeans，第二種是使用 PCA + Kmeans：

第一種的架構是用 sample code 的方式 encoder 三層、decoder 三層，將 encoder 之後的結果丟入 kmeans 做 predict，可以得到 0.99012 的準確度

第二種是先投影成 784 維(有加 whiten，處理投影之後的樣本分布)，之後用 kmeans(n_clusters = 2)做 predict，有加過 whiten 之後結果好了許多。

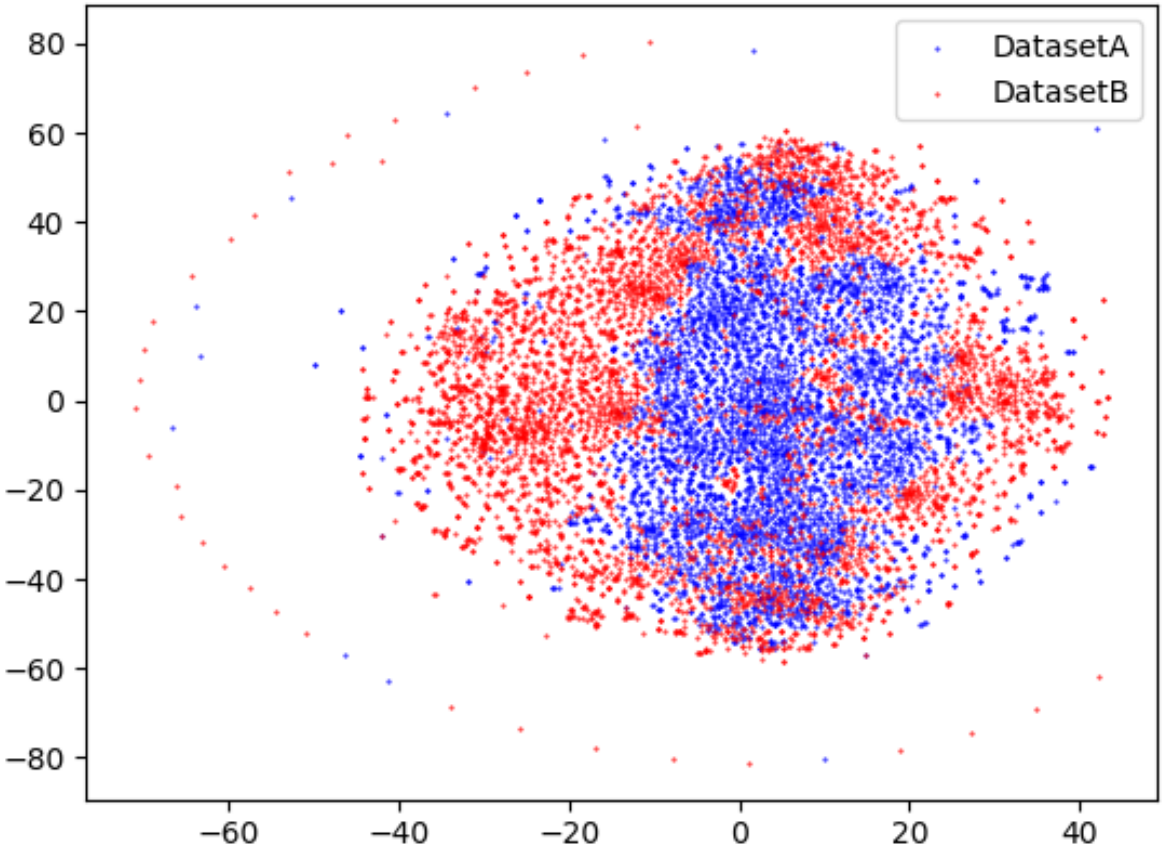
B.2. (.5%) 預測 visualization.npy 中的 label，在二維平面上視覺化 label 的分佈。

把 visualization.npy 用 pca 做 image cluster 之後，取前兩維之後 visualization。



(不知道為甚麼怎麼輸出都會變得不清楚，有附圖檔案在 github)

B.3. (.5%) visualization.npy 中前 5000 個 images 跟後 5000 個 images 來自不同 dataset。請根據這個資訊，在二維平面上視覺化 label 的分佈，接著比較和自己預測的 label 之間有何不同。



從上圖與下圖，可以比較出原本的樣本投影到二維空間時，兩個樣本看起來是非常相似的，而準確度也有大概 0.99 左右(datasetA 有 5374 項)，不過比較疏密程度看來，真實的樣本是比較密集的，而所推估的樣本分布是較為鬆散的，大概想成在高維度有一個類似球體的分布，而 datasetA 在球體的外圈，datasetB 在球體的內圈，而所預測的是比較分散的，可能在高維度中讓有些點太偏到外圍，因此就會有些沒有判別正確的樣本，

C. Ensemble learning

C.1. (1.5%) 請在 hw1/hw2/hw3 的 task 上擇一實作 ensemble learning，請比較其與未使用 ensemble method 的模型在 public/private score 的表現並詳細說明你實作的方法。（所有跟 ensemble learning 有關的方法都可以，不需要像 hw3 的要求硬塞到同一個 model 中）

[HW2]

這邊使用比較簡單的 voting 的方式，train 了三個 model 之後，我將三個 model predict 出來的結果，直接相加起來，因為是二元分類問題，非黑即白，所以這三個 predict 出來的結果一定是基數，所以只要一個 class 有兩票以上(含)的投票，就會被判斷成一，反之則否。

比較 validation 之後的結果如下，

	Model1	Model2	Model3	ensemble
Validation	0.8466	0.8497	0.8502	0.8503

有提高效果，但是結果並不是那麼好，於是比較了各個 model 出來的結果，其實 Train 出來的這三個模型的結果非常的相似，判斷對的與判斷錯的都差不多，三個相似的 model 做 ensemble 之後，其實並不能很好的提升結果。

在看了講義之後，覺得也有符合講義所寫的，因為我這三個 model 判斷出來的結果都差不多，因此看到的角度也都差不多，所以就像是同個領域的三個人，去看這個問題，觀點都很相似，於是得出來的結果都類似，不過還是有一點提升。