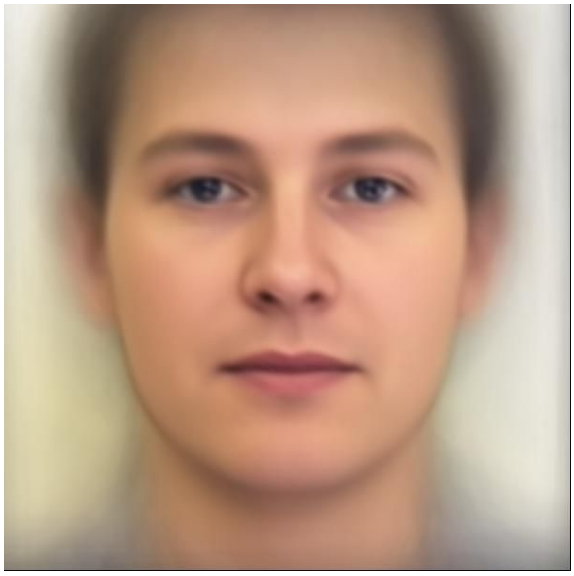
HW4 - Unsupervised Learning & Dimension Reduction

學號： b05901033 系級：電機二 姓名：莊永松

A. PCA of colored faces

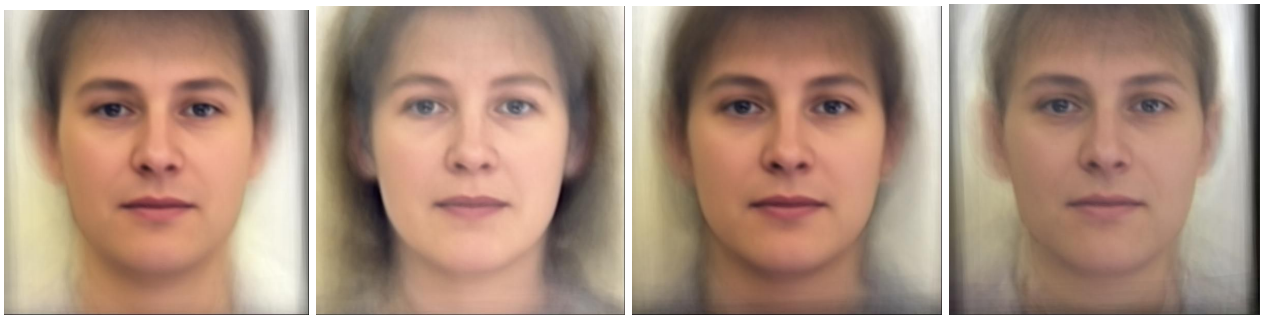
A1. (.5%) 請畫出所有臉的平均。



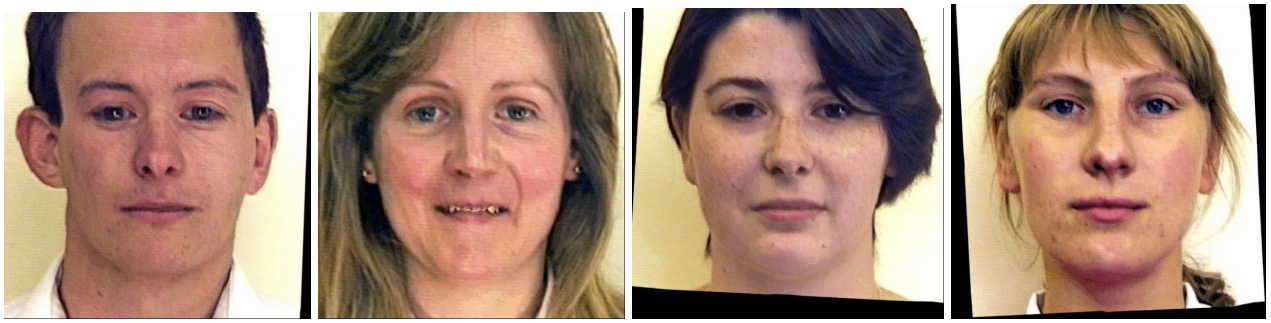
A2. (.5%) 請畫出前四個 Eigenfaces，也就是對應到前四大 Eigenvalues 的 Eigenvectors。

A3. (.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。

取出idx=100, 200, 300, 400四張，以前4大eigenfaces做reconstruction



另外有以前400大eigenfaces做reconstruction的結果(幾乎就是原圖了)



A4. (.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。

佔全部的比重：**第1大** 4.1% **第2大** 2.9% **第3大** 2.4% **第4大** 2.2%

B. Image clustering

B1. (.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)

方法一：直接用PCA降維到400維，再使用Kmeans做cluster，另外PCA有開whiten功能，可以有效去除雜訊。(PCA及Kmeans皆使用sklearn內建函式)

結果：kaggle上準確率為1.0000 / 1.0000 (public / private)

方法二：用NN架auto encoder，也是輸出400維，一樣使用Kmeans做cluster。model架構為：

[encoder]

Dense(600, *activation*='relu')

Dense(500, *activation*='relu')

Dense(400, *activation*='relu')

[decoder]

Dense(500, *activation*='relu')

Dense(600, *activation*='relu')

Dense(784, *activation*='relu')

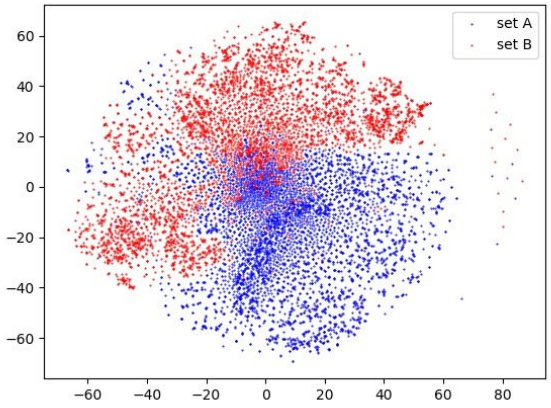
結果：kaggle上準確率為0.94154 / 0.94134 (public / private)

推測因為這次data助教有偷偷加上雜訊，所以PCA有開whiten能有效過濾並且降維效果較好，autoencoder則沒有對雜訊先做任何處理。

B2. (.5%) 預測 visualization.npy 中的 label，在二維平面上視覺化 label 的分佈。

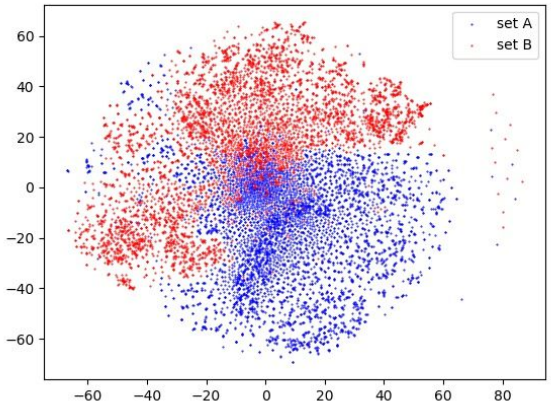
先用方法一PCA降到400維，再用t-SNE降成2維繪圖，預測方法也是用kmeans做cluster。

可以看到紅藍兩類大致上被分成左上和右下兩團，但仍然有少部分沒被分乾淨，不過因為我是用kmeans做cluster，t-SNE的結果並不是我拿來預測的依據，所以並沒有關係。



B3. (.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。

因為我用PCA有做到100%正確率，所以畫出來是一模一樣的，兩者 label 之間沒有不同。



C. Ensemble learning

C1. (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。

[是在hw3的task上ensemble]

ensemble的實作方法就是使用keras中的 `layers.average(Models)` 合併成一個model，實際上做出來的效果就是把三個model的輸出結果(各個model所預測的機率分布)平均起來，最後我再依據平均的結果作預測。(參照這篇教學：https://medium.com/randomai/ensemble-and-store-models-in-keras-2-x-b881a6d7693f)

在ensemble前我挑了三個不同model，model架構大體上一樣，只有BatchNormalization層加在不同的地方的差別(可參見hw3的report的model架構)。

在ensemble前各個model的kaggle表現分別是：

	private score	public score
model_1	**0.67818**	**0.68013**
model_2	**0.67400**	**0.67901**
model_3	**0.66592**	**0.67539**

ensemble後的kaggle表現：

	private score	public score
ens_model	**0.69908**	**0.71301**

提升了2%~3%的準確率，效果十分顯著。