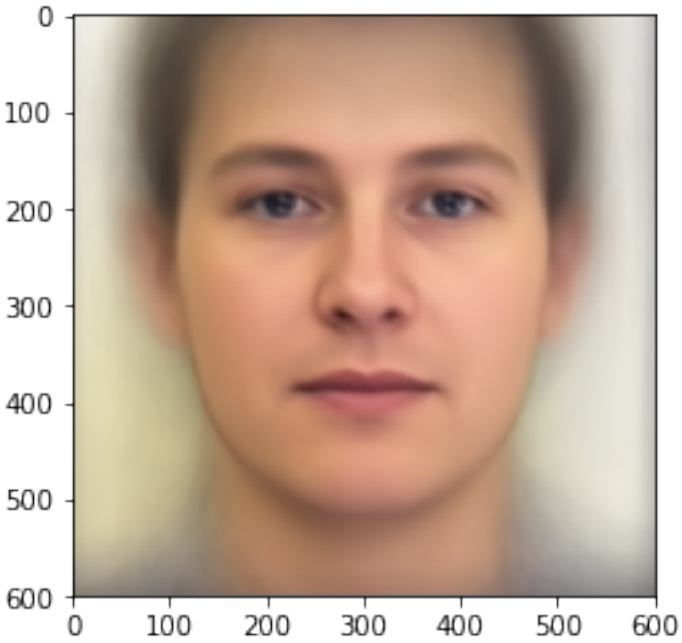
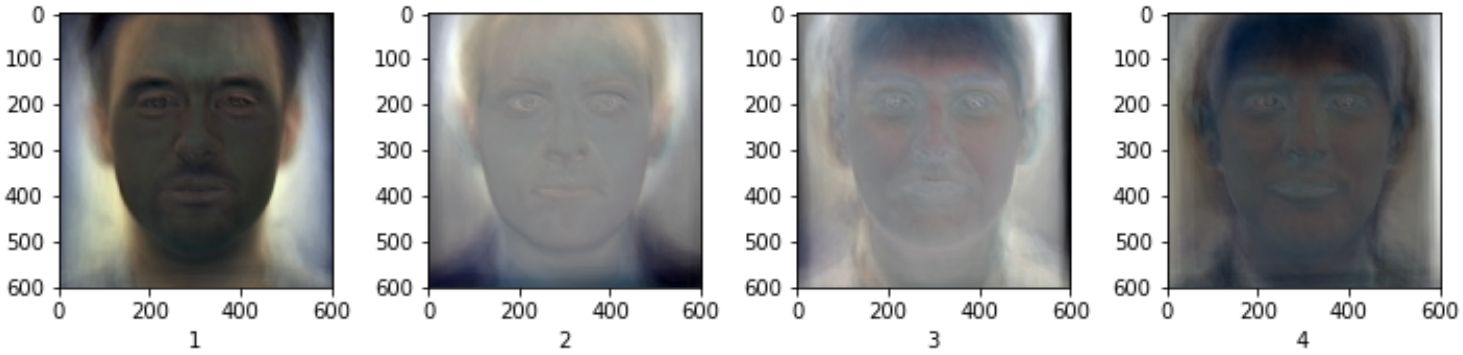
學號： r06922095 系級： 資工碩一 姓名： 陳代穎

A. PCA of colored faces

A.1. (.5%) 請畫出所有臉的平均。

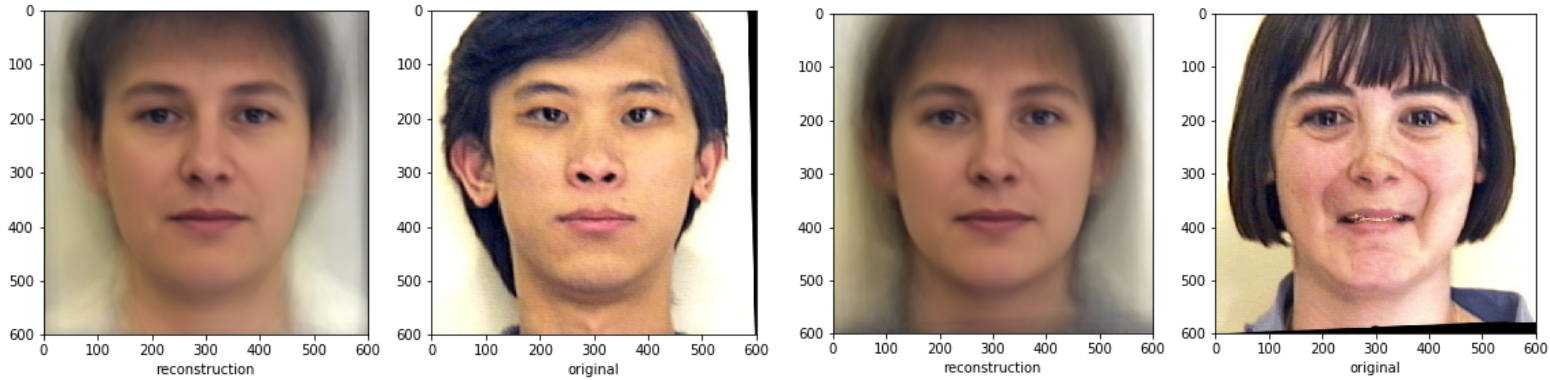


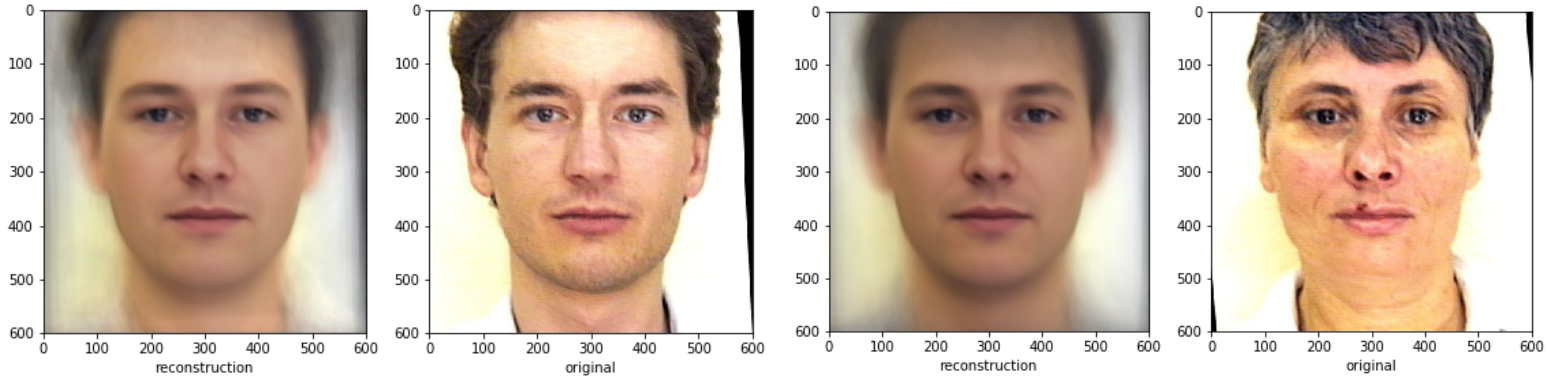
A.2. (.5%) 請畫出前四個 Eigenfaces，也就是對應到前四大 Eigenvalues 的 Eigenvectors。



由左到右分別為前四個Eigenfaces，看上去都是短髮的男性。

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。





由左上起分別為140.jpg、219.jpg、271.jpg、50.jpg，左邊為reconstruction、右邊為原圖。第一張我選擇東方人的臉孔，因為data大多都是西方臉孔，Eigenfaces傾向西方臉孔，因此可以看出無法reconstruct出東方臉孔。第二張我選擇短髮的女性，reconstruction有試圖將短髮的特徵描繪出來。最後兩張則較符合Eigenfaces的特徵，因此reconstruction較完整。

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。

No.	s_i	fraction(%)
1	540369.7	4.1
2	384451.1	2.9
3	311306.1	2.4
4	287854.9	2.2

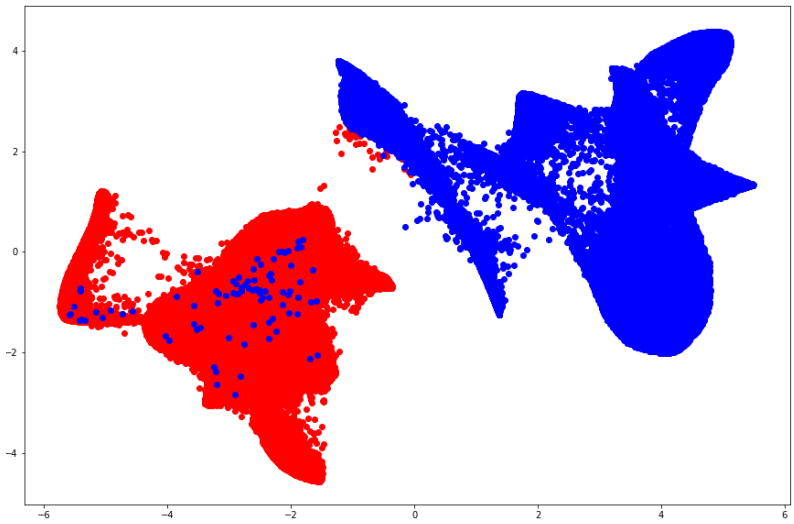
B. Image clustering

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)

(Collaborators: r06922086 林凡煒)

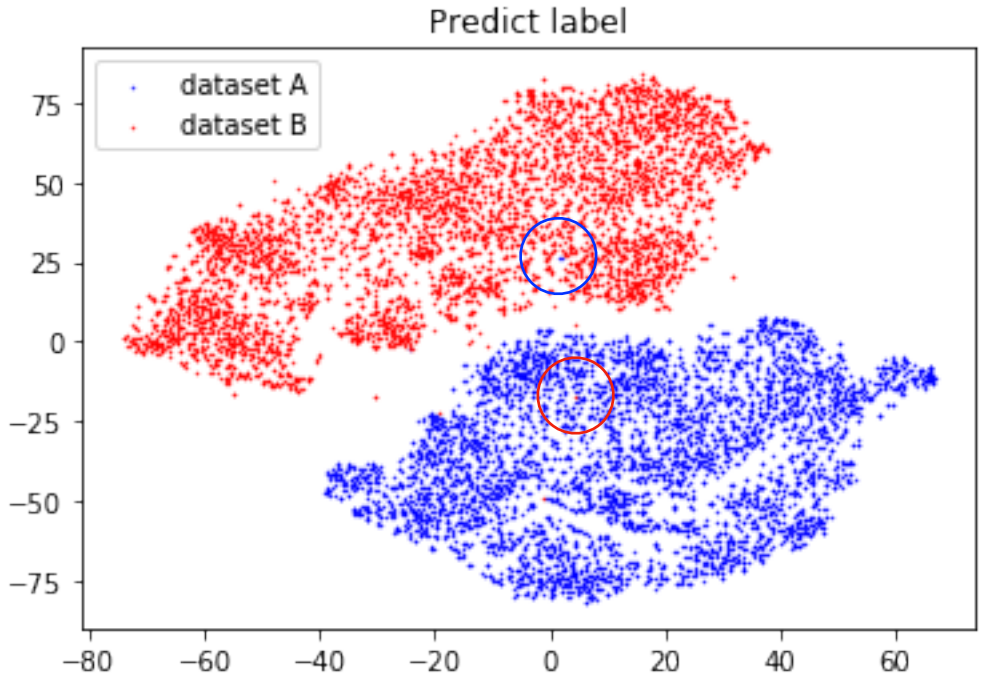
(a) 利用PCA從784維降維至395維，並加上PCA whitening。利用K-means將降維後的features做clustering到2類。但發現K-means聚類方式有隨機性，並不是每次結果都一樣。最終結果Public score為1.00000，而Private score為1.00000。

(b) 因用t-SNE在高維空間的運算時間過長，因此我先利用PCA從784維降至15維，接著再利用t-SNE將15維降至2維，以減少t-SNE運算時間。最後利用K-means將2維features cluster成2類。最終結果Public score為0.99760，Private score為0.99753。



將(b)的t-SNE的features作圖後如上，可以看見明顯的2類，若將(a)的結果當作ground truth，可以看到些許data被聚到另外1類。可能在15維的features不能夠保留這些data的特徵。

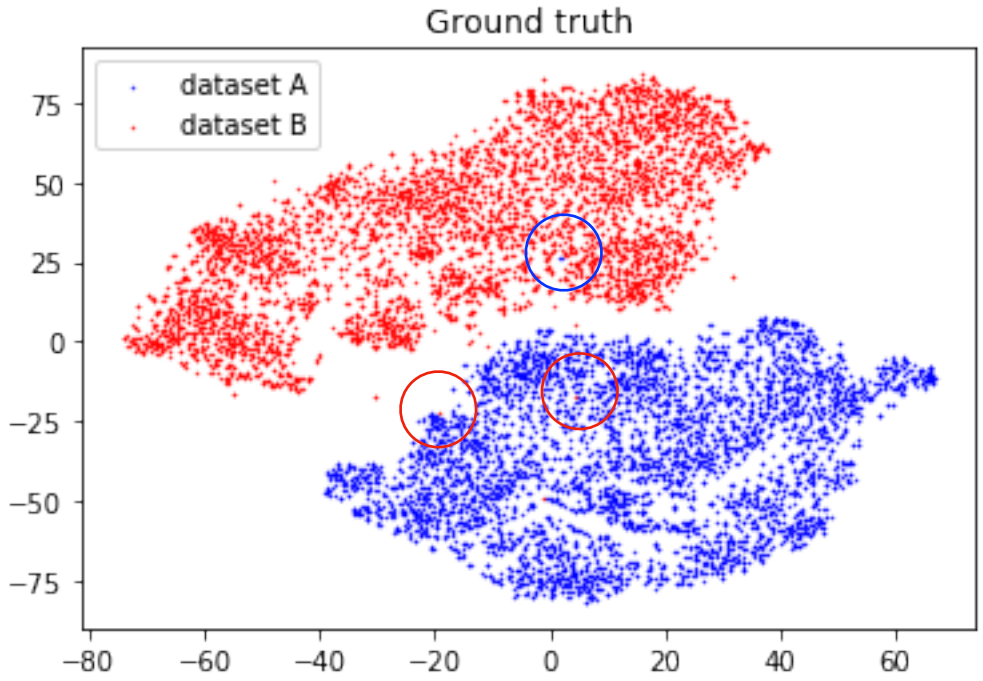
B.2. (.5%) 預測 visualization.npy 中的 label，在二維平面上視覺化 label 的分佈。



Predict方法為B.1(a)的train好的方法。Visualization方法為：利用image.npy train好的PCA將784維降至395維，但因為395維直接做t-

SNE效果不佳，因此再用PCA將395維降至15維。最後用t-SNE投影至2維做visualization。可以發現visualization時有明顯的2類，而prediction也能label出這2類，但在2類中各有1、2點標出不同顏色（上圖紅藍色圓圈處），會下一題討論。

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。



利用如同B.2中visualization的方法投影至2維，並標上ground truth label。可以發現其結果與B.2預測相同，計算accurary後也確認為1.0。但可以發現與B.2中同樣有data跑到其他類中的情形，推斷是做visualization時的過程造成這些data資訊遺失，導致在低維空間時反而與其他類的data相似。

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）

我選擇hw2上實作ensemble，當時是利用logistic regression及generative model。這次我選擇用decision tree與使用Bagging的random forest做比較。我用sklearn.tree.DecisionTreeClassifier及sklearn.ensemble.RandomForestClassifier實作。兩者我都採用Gini index當作節點分類的criteria，兩者的最大深度、leaf節點數目都沒有限制，在random forest中我用了10棵decision tree做bagging。Train set為29561筆data，validation set為3000筆data，以下為validation set上的accuracy。

Decision tree: 0.80500

Random forest: 0.85200

另外我也計算random forest的out-of-bag (OOB) accuracy: 0.83573

可以看出random forest明顯高於decision tree，高出快5%的accuracy，顯示了人多勢眾的威力。random forest在validation set上的accuracy還高於OOB accuracy，更顯示出這個model有多robust。