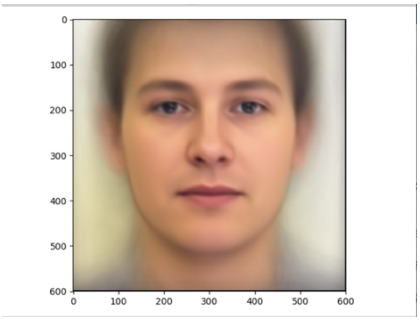
HW4

學號： b04901074 系級：電機三　姓名：吳倉永

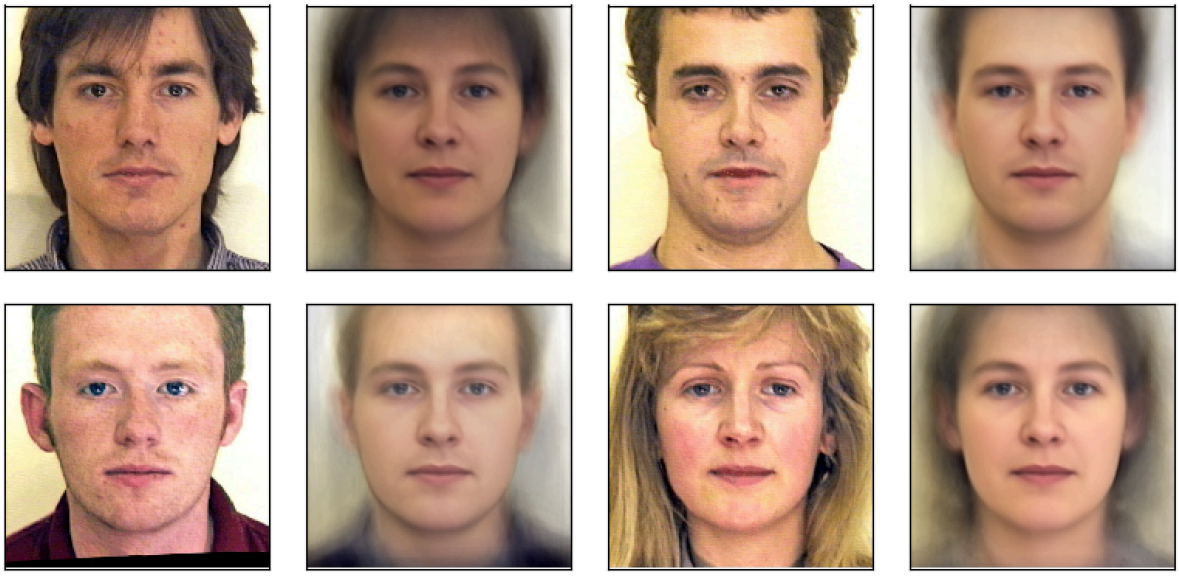
A. PCA of colored faces

A.1. (.5%) 請畫出所有臉的平均。



A.2. (.5%) 請畫出前四個 Eigenfaces

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: 分別挑選 id = 0, 10, 43, 197 的圖片



A.4. (.5%) 請寫出前四大 Eigenfaces 各自所佔的比重，請用百分比表示

百分比分別為：4.1%, 2.9%, 2.4%, 2.2%

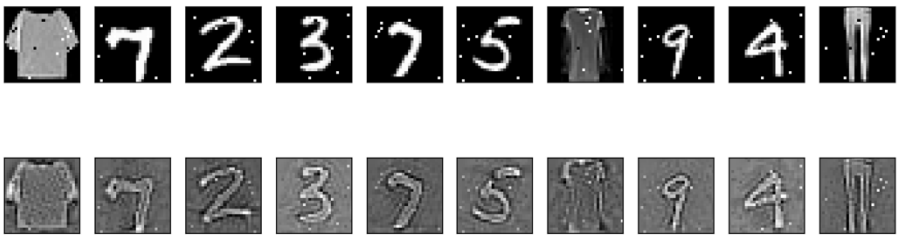
B. Image clustering

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)

我有使用 PCA 以及 auto-encoder 進行降維的實驗在 kaggle 上的分數其實很相近，但是我去分析 reconstruct 出來的圖就發現，auto-encoder 的效果比較好(取 PCA 降到 350 不取 32 維的原因是 performance 更差，得到的重建圖形很模糊)，但其實沒有差多少，畢竟 data set 的歧異性已經蠻大的了(衣服跟數字)。

PCA_350dim	99.996%
Auto-encoder	100.000%

Fig(1): auto-encoder 降到 32 維進行 reconstruct



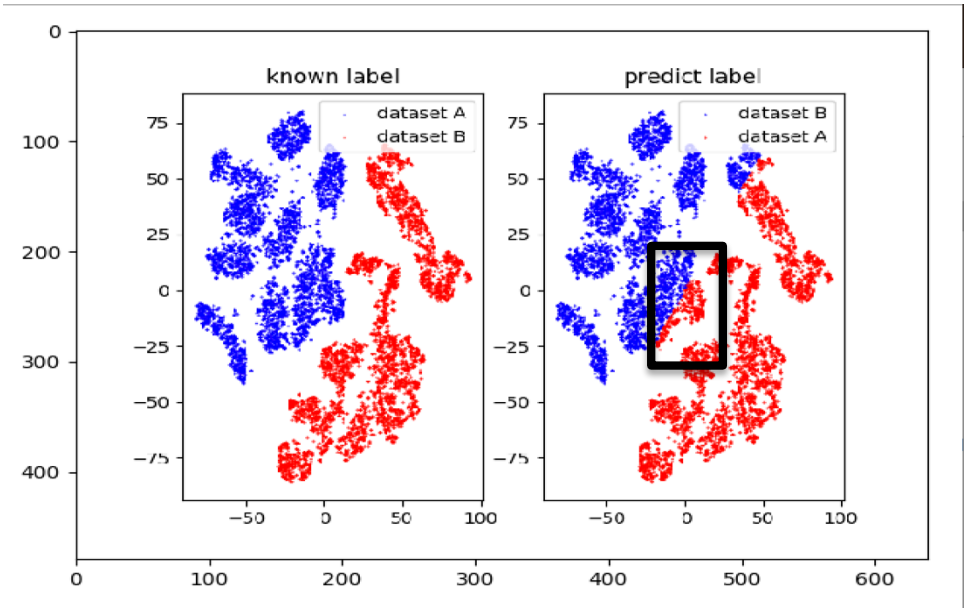
Fig(2): PCA 降到 350 維進行 reconstruct

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

請見 Fig(3, 4) 右圖

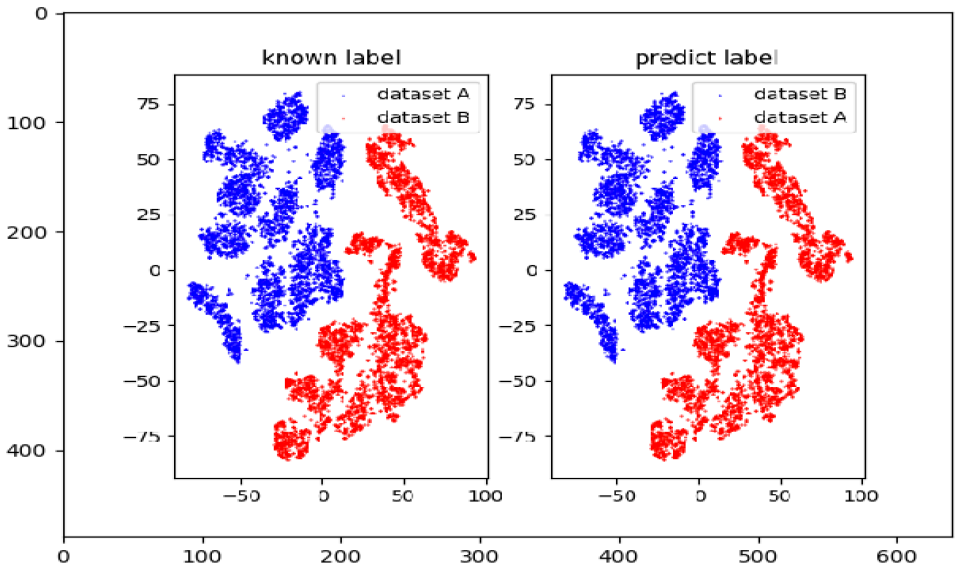
B.3. (.5%) visualization.npy 中前 5000 個 images 跟後 5000 個 images 來自不同 dataset。請根據這個資訊，在二維平面上視覺化 label 的分佈，接著比較和自己預測的 label 之間有何不同。

Ans: 可以發現到此次利用 TSNE 降維得到的結果雖然說分的夠開，但是因為 Kmeans 是 linear 的 boundary，造成他只能切一條線，而不能像曲線一樣把 A, B 分開，而在這個 task 中(觀看 known label 那張圖)，A, B 需要用曲線分界才能分開(黑框部分)。



Fig(3): visualization.npy 利用 auto-encoder 降成 32 維，並在用 tsne 降到 2 維進行作圖。左圖為助教給定的 label，而右圖為利用 k-means 作 cluster 得到的結果

而 Fig(4) 是利用 HAC 作 clustering 得到的結果，且他真的把兩個 data set 分開，我觀察到以下結論：首先在作 TSNE 的時候，兩組 data 的 2 維分布不能是被切開的(ex: data A 中有 data B)，如此一來，找 neighbor 會被分在一起，所以 tsne 要先至少能把 data 分成兩區。而在作 clustering 時(也就是在 label data)，像這種一群一群的且間隔又不是說非常的大，用 Kmeans 一定會切到某一群，畢竟是 linear 的分法，如此很適合用 HAC 去分群，先從小群的開始分，然後階層式慢慢跟其他群合併，就能得到 100%的效果



Fig(4): visualization.npy 利用 auto-encoder 降成 32 維，並在用 tsne 降到 2 維進行作圖。左圖為助教給定的 label，而右圖為利用 HAC 作 cluster 得到的結果

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。

Ans: 我利用 hw3 的 3 個好 model，實作 ensemble voting。方法如下：將已經 train 好的 3 個 model load 進來，然後把他們的 output concatenate 在一起，因此會形成一個 21 維的 vector(7 個 class*3)，這個 tensor 在後面兜出一個 final layer DNN(Dense 192 with activation 'relu' → Dense 7 with softmax)，進行 training，結果如下：

Ensemble 的過後的準確率確實有上升，但當我利用將 3 個 model 的結果直接進行平均卻得到 72.4% 的準確率，比我用 NN learn 出來的還要好…我把 activation 改成 sigmoid 看看結果也沒有什麼進展。

Model Name	Accuracy(public+private)
CNN Model_1	70%
CNN Model_2	70.1%
CNN Model_3	70.3%
Ensemble_DNN	72.1%