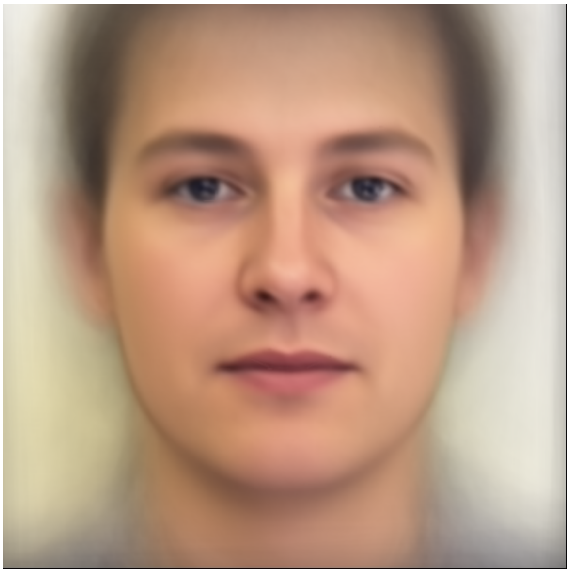
學號：R06922048 系級： 資工所碩一 姓名：陳柏堯

A. PCA of colored faces

(.5%) 請畫出所有臉的平均。

答：（參見 p1.py）

將每一張圖片的 pixel 加起來取平均值作圖：



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

答：按照助教的投影片作 SVD，將 U 的前 4 個 column 作圖：


(反色)

(.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。

答：按照助教投影片的提示，選擇 0~3.jpg 的圖做 reconstruction

(.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。

答：將 SVD 結果的 S 前四個值去計算比例，再取小數點最後一位：

0 st eigenvector ratio: 0.04144624838262957 ≈ 4.1%

1 st eigenvector ratio: 0.02948732225112061 ≈ 2.9%

2 st eigenvector ratio: 0.02387711293208415 ≈ 2.4%

3 st eigenvector ratio: 0.022078415569025393 ≈ 2.2%

B. Visualization of Chinese word embedding

(.5%) 請說明你用哪一個 word2vec 套件，並針對你有調整的參數說明那個參數的意義。

答：（參見 p2.py）

我使用 gensim.Word2vec; 先用 jieba 分詞去掉符號，再用 gensim 轉 vector。

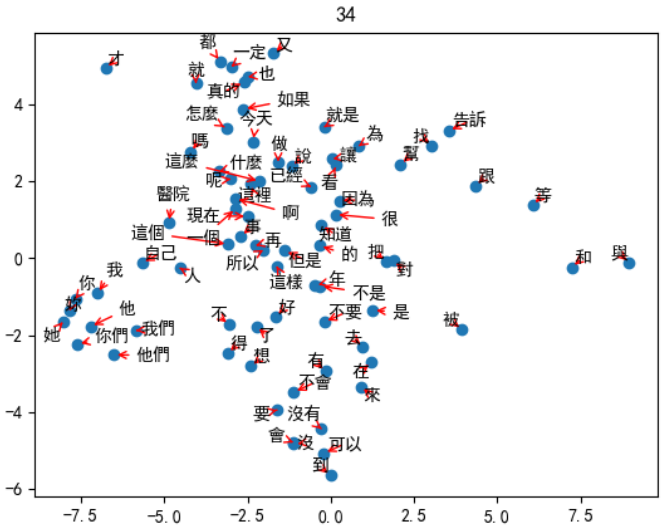
gensim 參數：

size=300, embedding 的維度為 300

windows=5, word2vec 的 window size 我設定為 5.

(.5%) 請在 Report 上放上你 visualization 的結果。

答：我選出出現頻率大於 6000 的詞語，用 SVD(取前兩維)降維到 2 維，如下圖：



(.5%) 請討論你從 visualization 的結果觀察到什麼。

從上圖中可以發現，「你」,「我」,「他」三者之間的距離與「你們」，「我們」，「他們」三者之間的距離很類似，都形成了一個三角形，並且主要都聚在一起。而意思接近的詞語，像是「沒有」與「沒」也很接近，「這個」、「這裡」和「現在」接近。

C. Image clustering

(.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)

答：

1. 方法一，自己抽取了兩個 features:
   （1）像素值大於 240 的數量 除以 像素值大於 10 的數量 比例
   (數字的圖中，有顏色的往往比衣服的顏色數字大)
   （2）像素值大於 100 的數量 除以 整張圖片像素數量 比例
   (衣服的圖中，有顏色的像素往往比數字的顏色像素多)

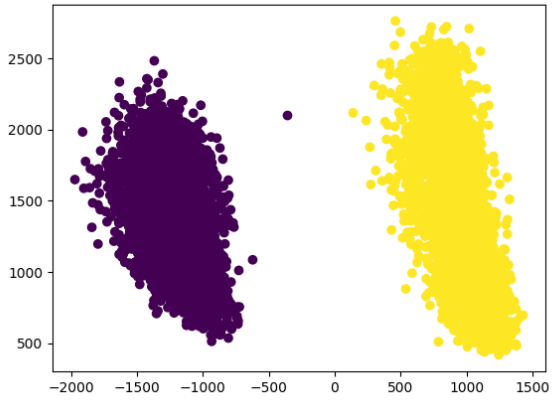
抽取 feature 之後再用 kmeans 分兩群。得到 public 0.39710，private 0.39579.

2. 方法二，使用 autoencoder，由 784 維 Dense 分別接 128 維、64 維、32 維，再 Dense 接回 64 維、128 維、784 維，最後用 binary_crossentropy 陪 adam,經過 800epoch 之後，在用 kmeans 分兩群得到 public 0.69184, private 0.69050.

3. 方法三，使用 autoencoder，由 784 維 Dense 分別接 256 維、128 維、64 維，再 Dense 接回 128 維、256 維、784 維，最後用 binary_crossentropy 陪 adam,經過 2400epoch 之後，在用 kmeans 分兩群得到 public 0. 99940, private 0.99914。

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

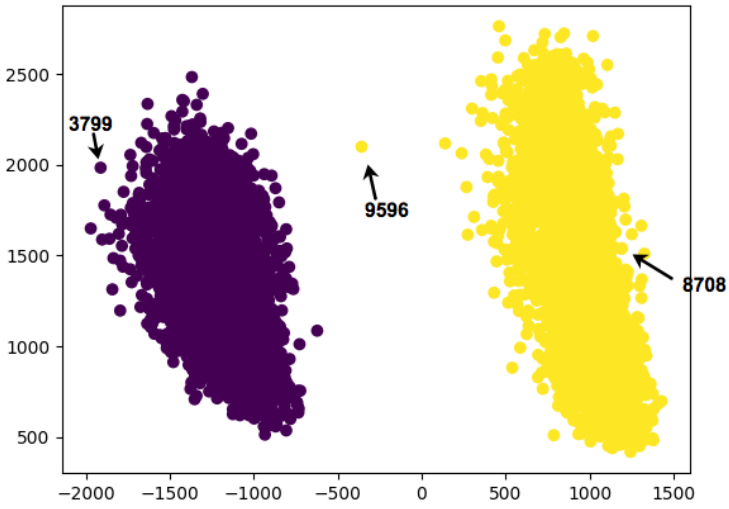
答：我使用 SVD 用方法三(public 0. 99940, private 0.99914) , 對所有圖片作 encoder 之後，用 SVD 降維到 2 維，在用 Kmeans 的分群結果去上色，如下圖：



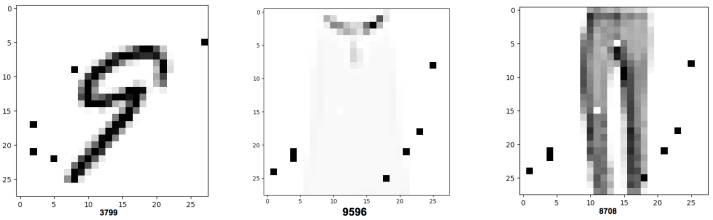
（紫:數字，黃:衣服）

(.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。

答：前 5000 為第一群，後 5000 位第二群，如下圖：



中間有一個點(第 9596 張圖片)在自己的預測分到了數字，而正確答案是衣服。將第 9596 張圖片畫出來如下圖。



(反色)