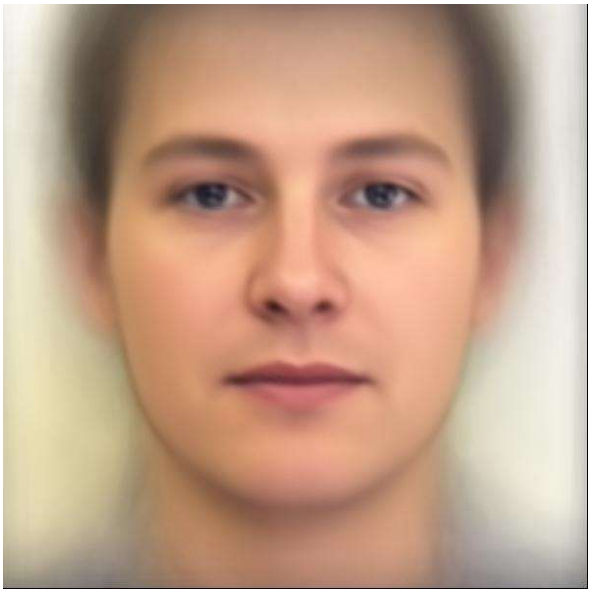
學號：B04902099 系級： 資工三 姓名：黃嵩仁

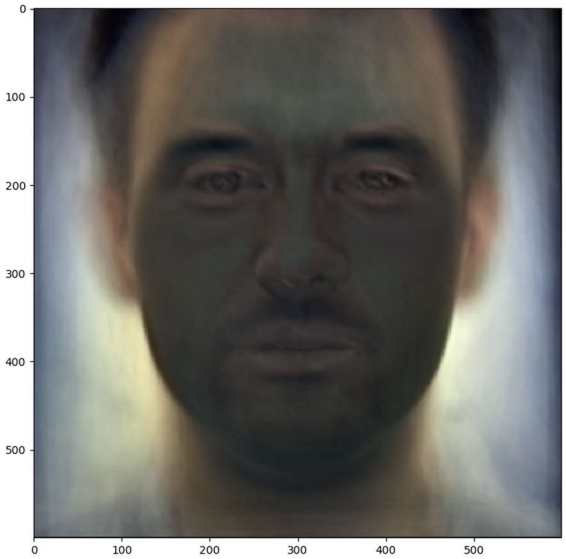
A. PCA of colored faces

A.1. (.5%) 請畫出所有臉的平均。

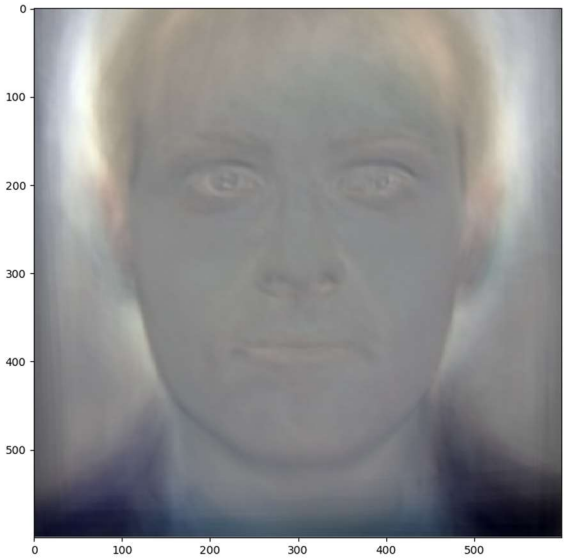


A.2. (.5%) 請畫出前四個 Eigenfaces，也就是對應到前四大 Eigenvalues 的 Eigenvectors。

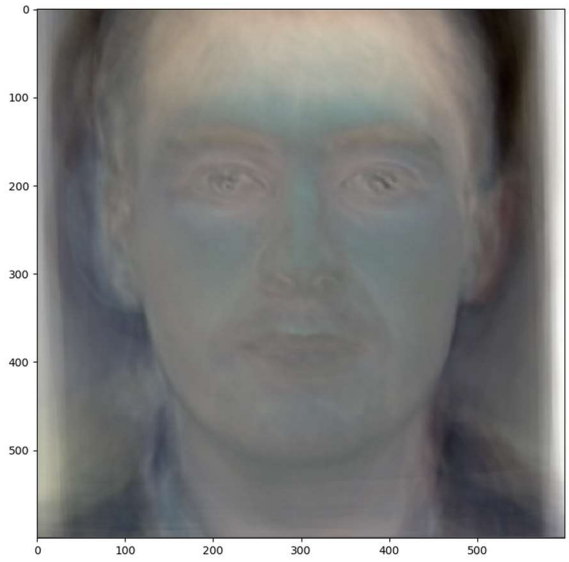
第一張 eigenface



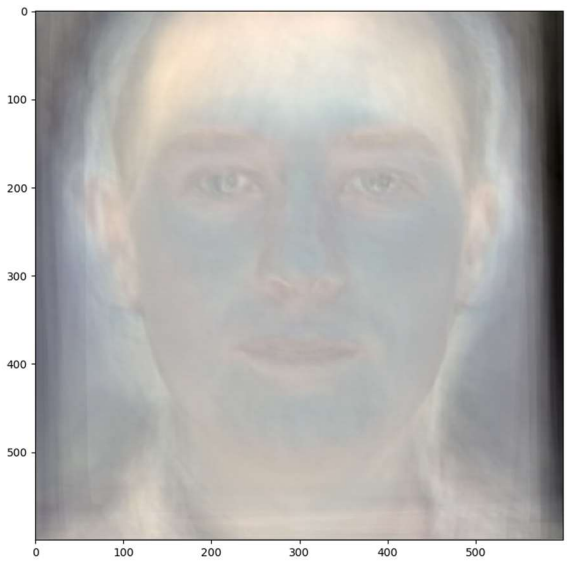
第二張 eigenface



第三張 eigenface



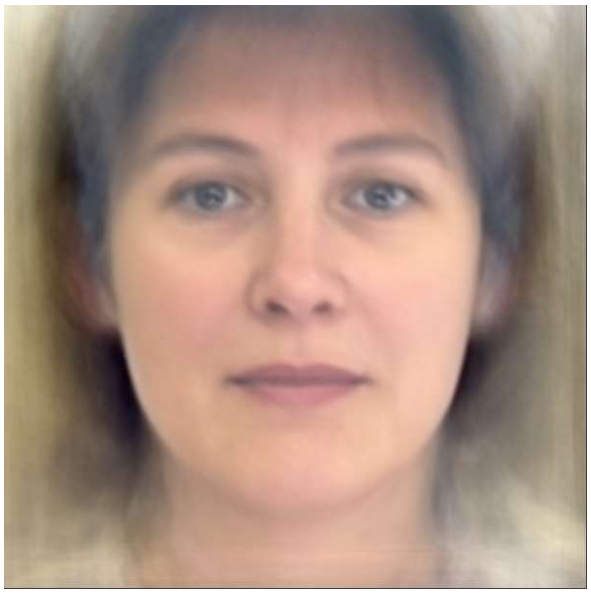
第四張 eigenface



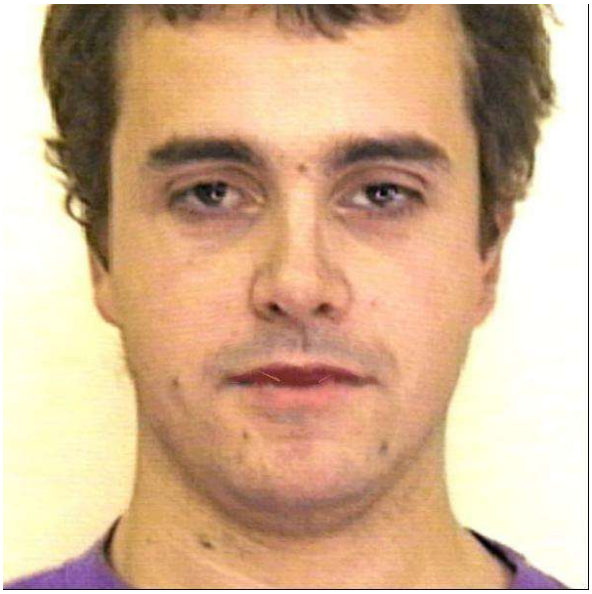
A.3. (.5%) 請從數據集中挑出任意四個圖片，並用前四大 Eigenfaces 進行 reconstruction，並畫出結果。

原圖(5.jpg)

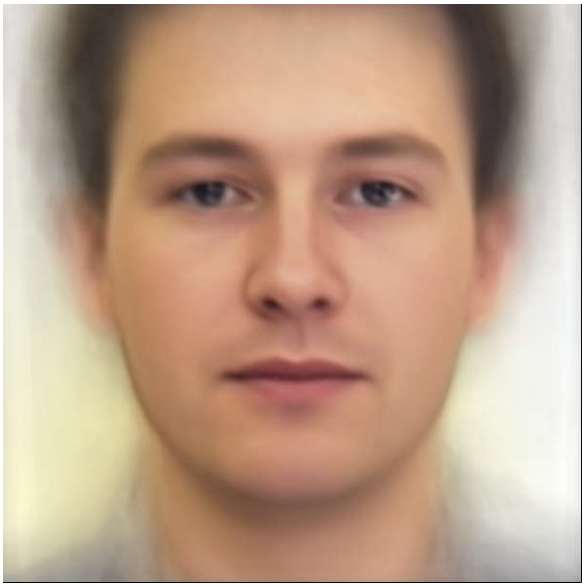
重建



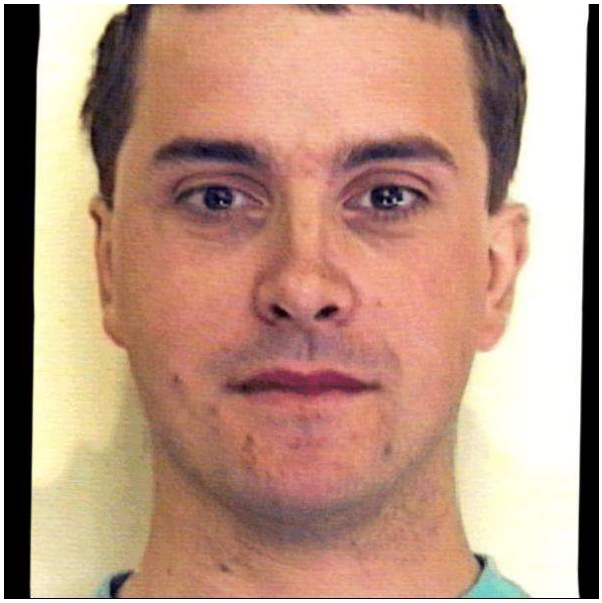
原圖(10.jpg)



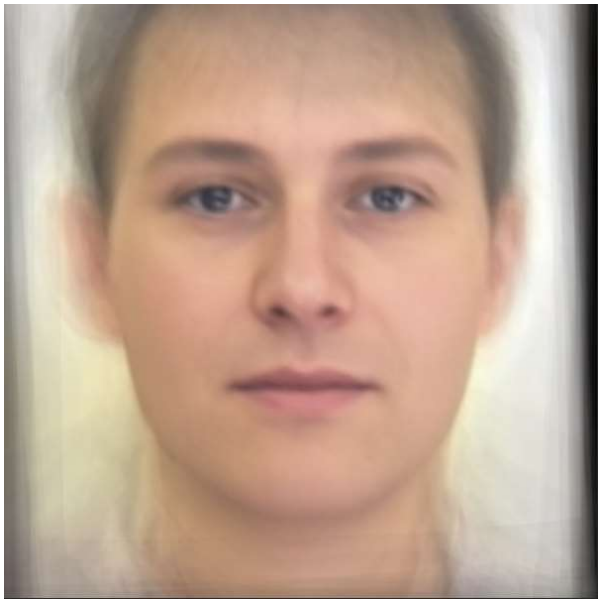
重建



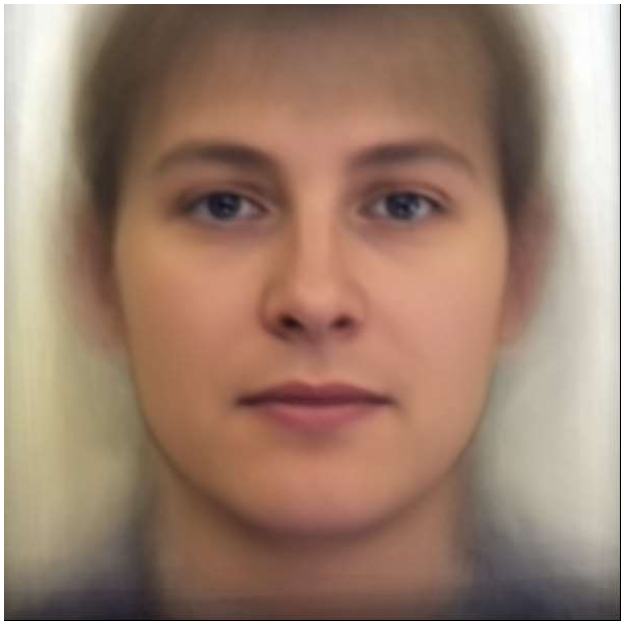
原圖(15.jpg)



重建



原圖(170.jpg)　　重建



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。

第幾大 Eigenface	所佔的比重
第一大 Eigenface	4.1%
第二大 Eigenface	2.9%
第三大 Eigenface	2.4%
第四大 Eigenface	2.2%

B. Visualization of Chinese word embedding

B.1.　(.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：使用 gensim 套件，調整的參數如下：

Size = 72：訓練出的 word vector 的維度數為 72 維。

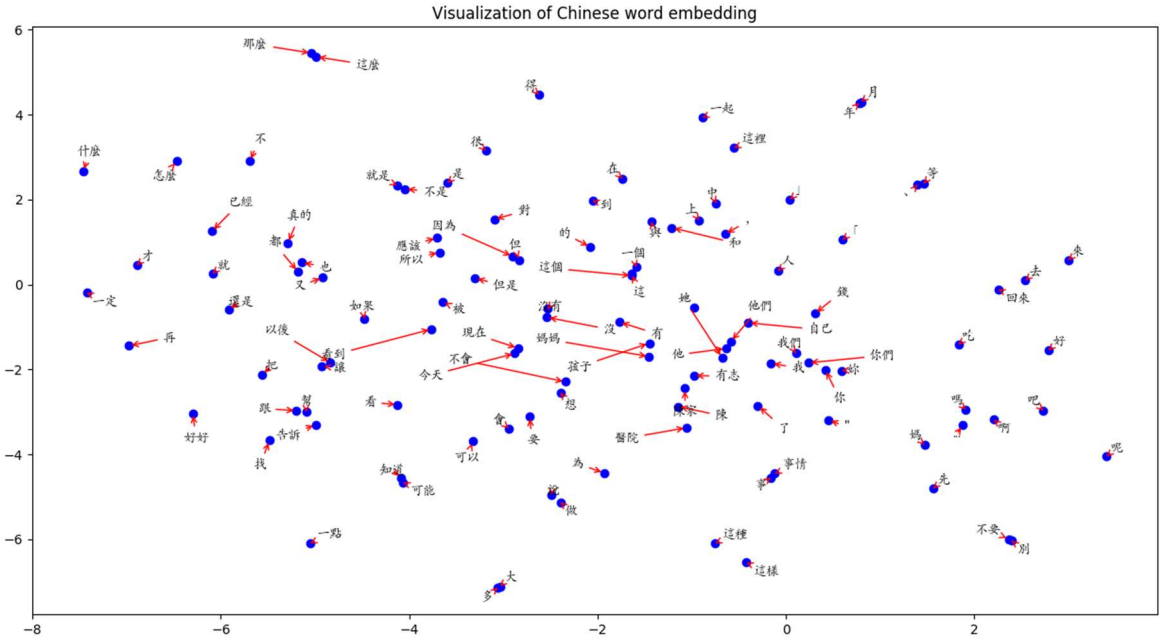
Window = 5：往旁邊參考 5 個字。

min_count = 4000：該詞出現次數需要超過 4000 次，才會被轉成 word vector。

Negative = 5, negative sampling 的個數。

其他：batch_words = 10000，iter = 10。

B.2. (.5%) 請在 Report 上放上你 visualization 的結果。



B.3. (.5%) 請討論你從 visualization 的結果觀察到什麼。

答：由 visualization 的結果可觀察到：

1. 主詞聚集在中間偏右的地方(如：你們、我們、自己、姓名等)

2. 語助詞(如：啊、嗎、吧、呢)都聚集在右下角。

3. 關於判斷的詞(如：是、不是、因為、應該、所以)大多都在中間上面的地方，但也有少部分位於其他地方。

4. 動詞的部分大多都在左下方的位置(如：把、幫、看、告訴等)

C. Image clustering

C.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)

答：

方法	Kaggle public score
PCA + kmeans	0.03026
PCA + cosine	0.06182
Auto-encoder + kmeans	0.99765

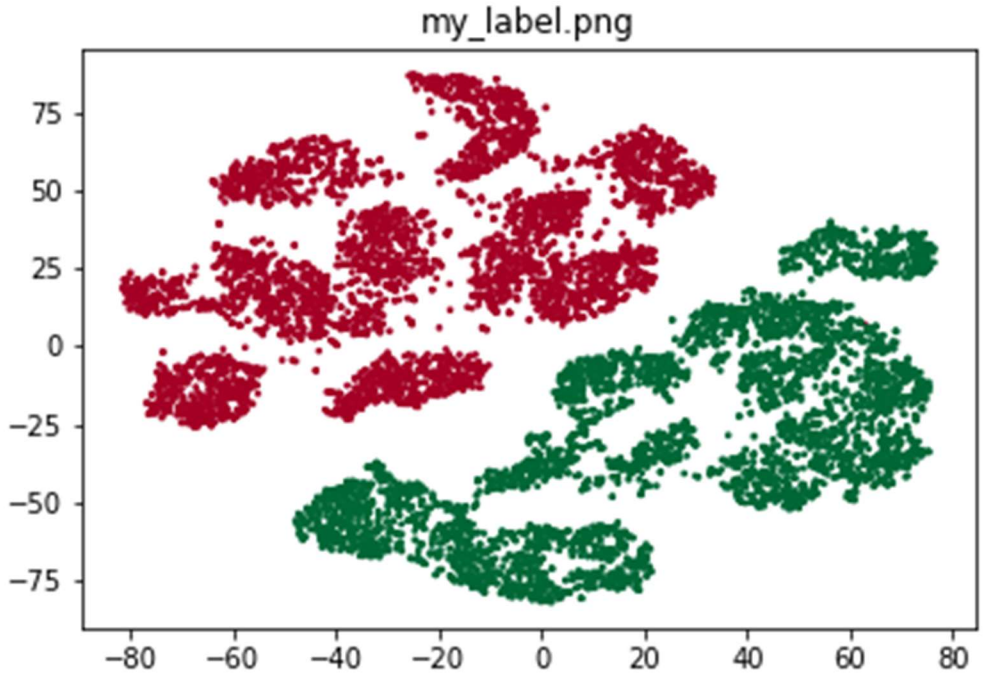
在嘗試過以上幾種方法前，我有先試過使用 t-sne 降維，但是因為運算時間過長因此放棄。

PCA + kmeans 跟 PCA +cosine 兩種方法都是先使用 PCA 將維度降至 10 維在做接下來的計算，但可看出效果不彰，推測可能的原因為一開始在使用 PCA 降維時流失太多資訊導致嚴重失真，因此影響分群的準確度。

而 auto-encoder+keams 作法，則是使用 keras 實作 autoencoder(encoder 部分為 3 層 dense, units = 512, 128, 32, activation function 為 relu)，訓練前有先將資料標準化。從結果可看出效果滿顯著的，不過可能是因為訓練的時間不夠長，因此沒有達到 1.00000 的滿分。

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

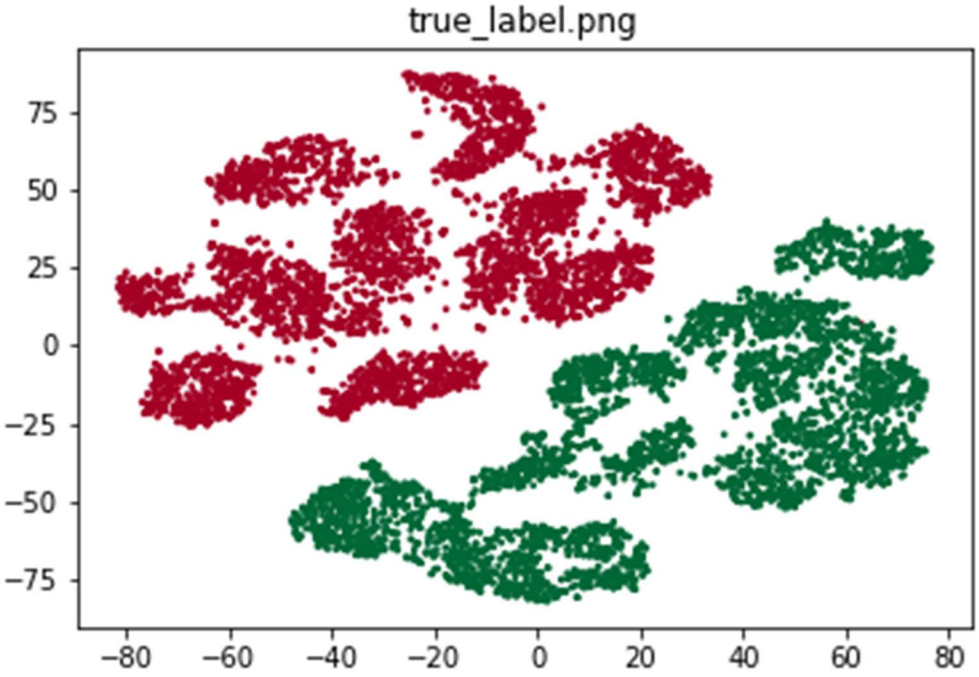
答：



註：本題使用 keams 分群, 先使用 auto-encoder 降到 32 維，再使用 t-sne 降到 2 維。

C.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。

答：



觀察兩圖可發現兩圖 label 的分佈完全相同，表示我所做的 auto-encoder 降維的效果還不錯，也有可能是因為 data 量較小的緣故，因此沒有出現 label 錯誤的情況。

註：本題先使用 auto-encoder 降到 32 維，再使用 t-sne 降到 2 維。