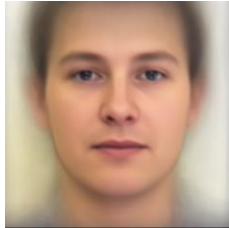


## A. PCA of colored faces

A.1. (.5%) 請畫出所有臉的平均。

Ans:



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

Ans:

由左至右，從第一個 eigenface 到第四個 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。

Ans:

上排圖片由左至右分別為 38.jpg, 158.jpg, 172.jpg, 209.jpg 的原圖，而下排圖片則是利用前四個 Eigenfaces 重建出來的圖片。



A.4. (.5%) 請寫出前四大 Eigenfaces 各自所佔的比重，請用百分比表示並四捨五入到小數點後一位。

Ans:

Eigenfaces1	Eigenfaces2	Eigenfaces3	Eigenfaces4
4.1%	2.9%	2.4%	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。

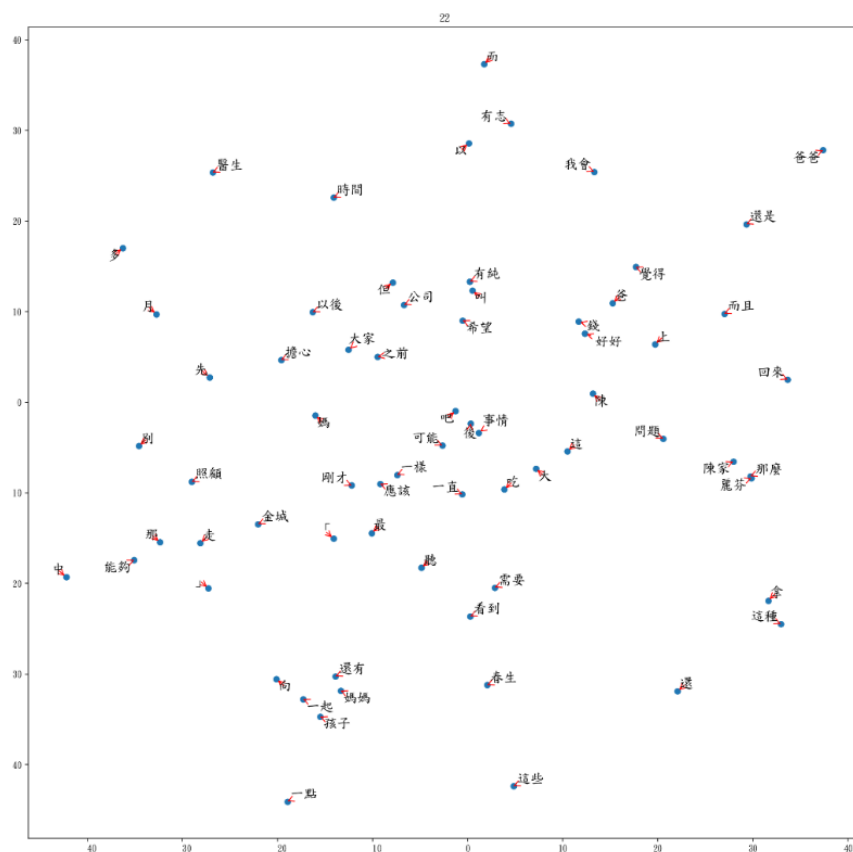
Ans:

使用 gensim 的 Word2Vec 套件，修改的參數有: size=128，表示詞向量的維度；iter=5，表示重複疊代幾次；min\_count=10，最少要出現 10 次的詞才會產生詞向量；workers=16，利用幾個 threads 來計算

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

Ans:

使用 tsne 降維，在選出出現 3000~6000 字之內的詞，作為 visualization 的 point



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

Ans:

由圖中看出比較有代表性的應該是左下角的”媽媽”和”小孩”，右上角的”爸”、”錢”和”好好”，右邊的”陳家”和”麗芬”，這些文字在中文意義或故事的内容上有一定的關連性。

## 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)

Ans:

方法一 (autoencoder + k-means): 建立一個 DNN 的架構(如下圖)，activation 為 relu，optimizer 為 adam，loss 為 mse，降維到 64。利用 k-means 分成兩群。最後 kaggle public score: 0.99929。

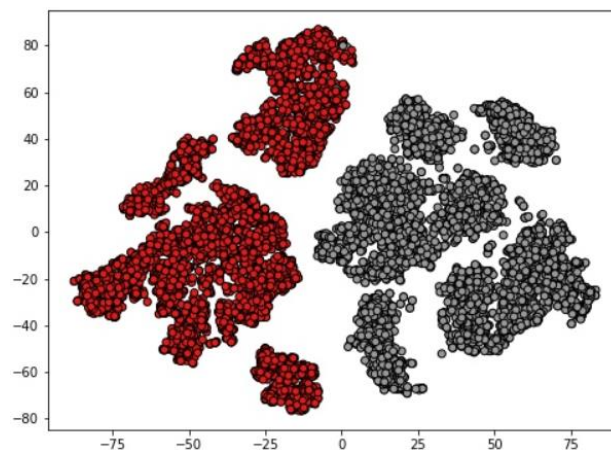
Layer (type)	Output Shape	Param #
input_1 (InputLayer)	(None, 784)	0
dense_1 (Dense)	(None, 256)	200960
encoded (Dense)	(None, 64)	16448
dense_2 (Dense)	(None, 256)	16640
dense_3 (Dense)	(None, 784)	201488
Total params: 435,536		
Trainable params: 435,536		
Non-trainable params: 0		

方法二(pca + k-means): pca 降維到 400，有使用 whiten 這個參數，確保有 uncorrelated 的 outputs。利用 k-means 分成兩群。最後 kaggle public score: 1.00000。

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

Ans:

利用 autoencoder 事先在 train.npy 資料集上 train 好的 model 降維到 64，再用 k-means 直接拿來預測新的 visaulization.npy 資料集，並使用 tsne 降維到 2。分群結果如下圖。



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。

Ans:

由圖片可以看出答案和預測的幾乎完全一樣，只有在交界處上方，預測的有明顯灰色的點分類錯誤。

