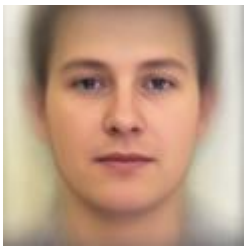
學號：R06942077 系級： 電信碩一 姓名：洪健鈞

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。

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。

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。

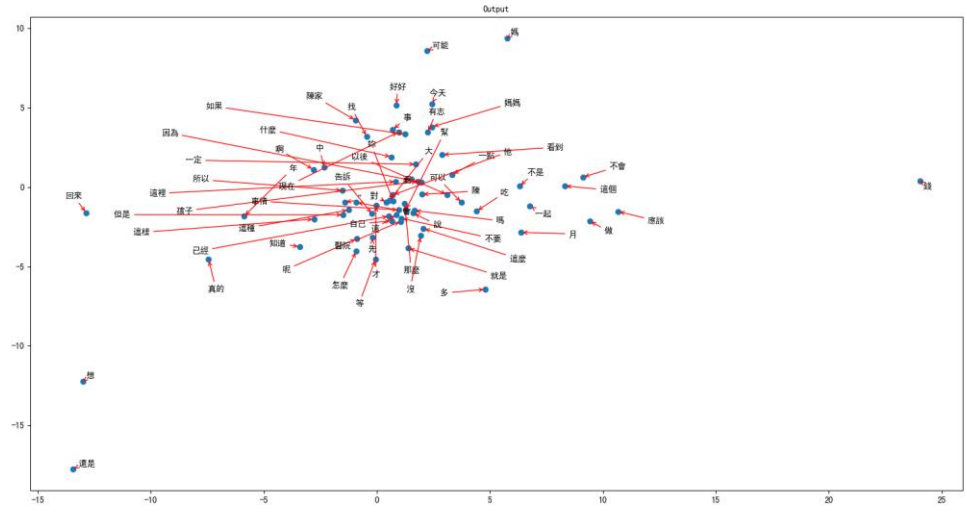
4.2% 2.9% 2.4% 2.2%

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。

我用 gensim 的 word2vec 套件, 有調的參數有 window, window 的意思是在句子中前後看幾個詞, size 則是表示將詞轉換成幾維的 vector, alpha 則是我的 learning rate

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。

從 visualize 的圖來看, “要” ”會” “可以”, 這些字詞的點非常的接近, 而這代表這些詞非常常在同一句話同時出現, 可以猜這這些字詞有密切關係。

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)

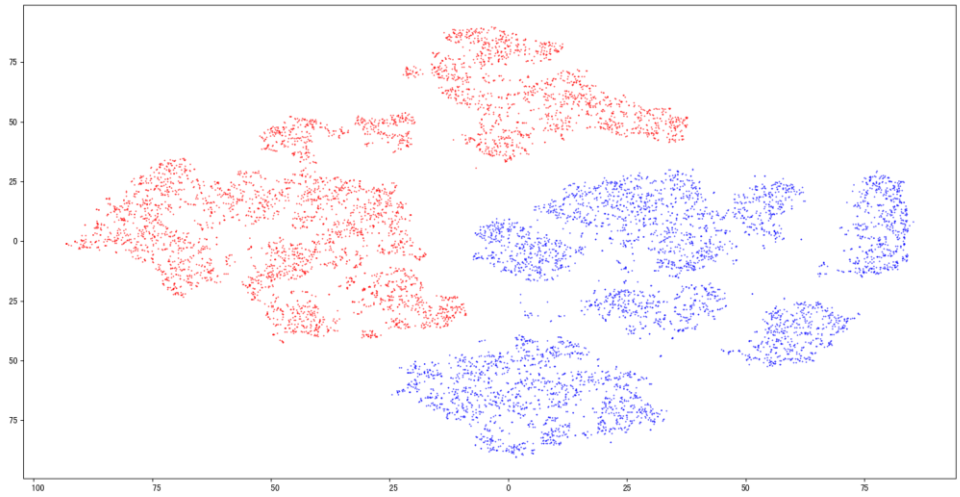
Layer (type)	Output Shape	Param #
input_1 (InputLayer)	(None, 784)	0
dense_1 (Dense)	(None, 128)	100480
dense_2 (Dense)	(None, 64)	8256
dense_3 (Dense)	(None, 32)	2080
dense_4 (Dense)	(None, 64)	2112
dense_5 (Dense)	(None, 128)	8320
dense_6 (Dense)	(None, 784)	101136

我通過 strong baseline 的 model 是與助教一樣,用 autoencoder, dimension 從 128,64,32 這樣去降維, 256 有試過但效果不佳。然後 train 100epoch, opt 用 adam, 這樣的結果可以讓我準確率到達 public 0.85947, private 0.86584。

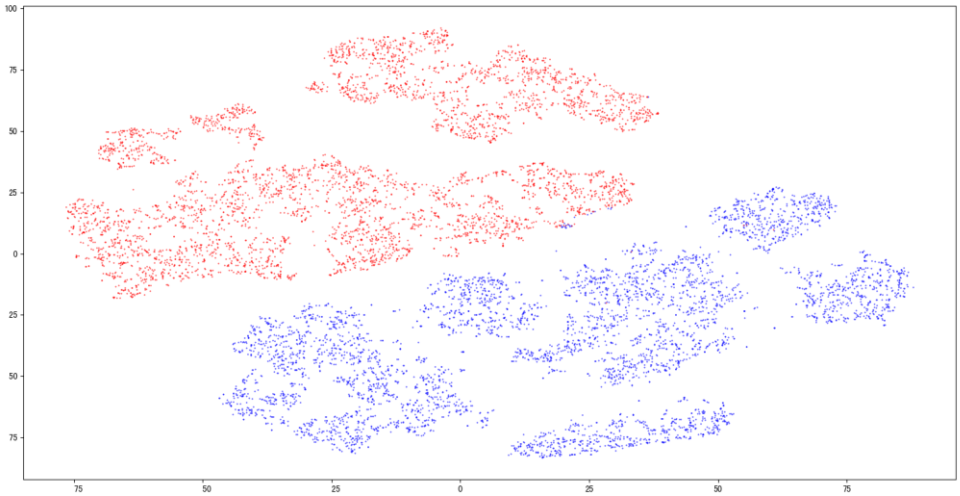
第二個方法我是用 PCA 去降維, 但發現效果非常差, 同樣也是用

dim 128, 但是 public 分數僅 0.16611, private0.17120, 可能參數設定不是很好。

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



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。



與正確答案相比,我的預測結果中間有些許紅色點和藍色點混在同一區域, 可能是我預測的結果沒有完全將兩個 dataset 分開, 所以是視覺化也呈現出我的結果不是接近百分百的正確率, 很合理。