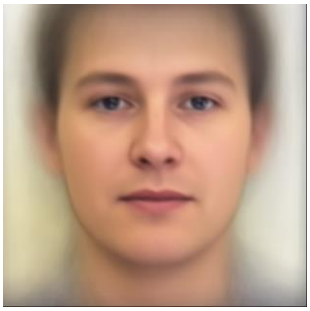
學號：B03902096 系級： 資工四 姓名：陳柏屹

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。

由左至右分別是最大到第四大 eigenvalue 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。

由左至右分別為 0, 10, 20, 30.jpg 的 reconstruction

A.4. (.5%) 請寫出前四大 Eigenfaces 各自所佔的比重 (explained variance ratio)，請四捨五入到小數點後一位。

前四大 Eigenface 比重由大到小分別為: 4.1%, 3.0%, 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。

我使用 gensim 套件。所調整過的參數有以下：

sg: Training algorithm. 0 為 CBOW, 1 為 skip-gram,

size: 每個 word vector 的維度

window: 同一句子中 target word 最遠 predict 到前後各幾個字

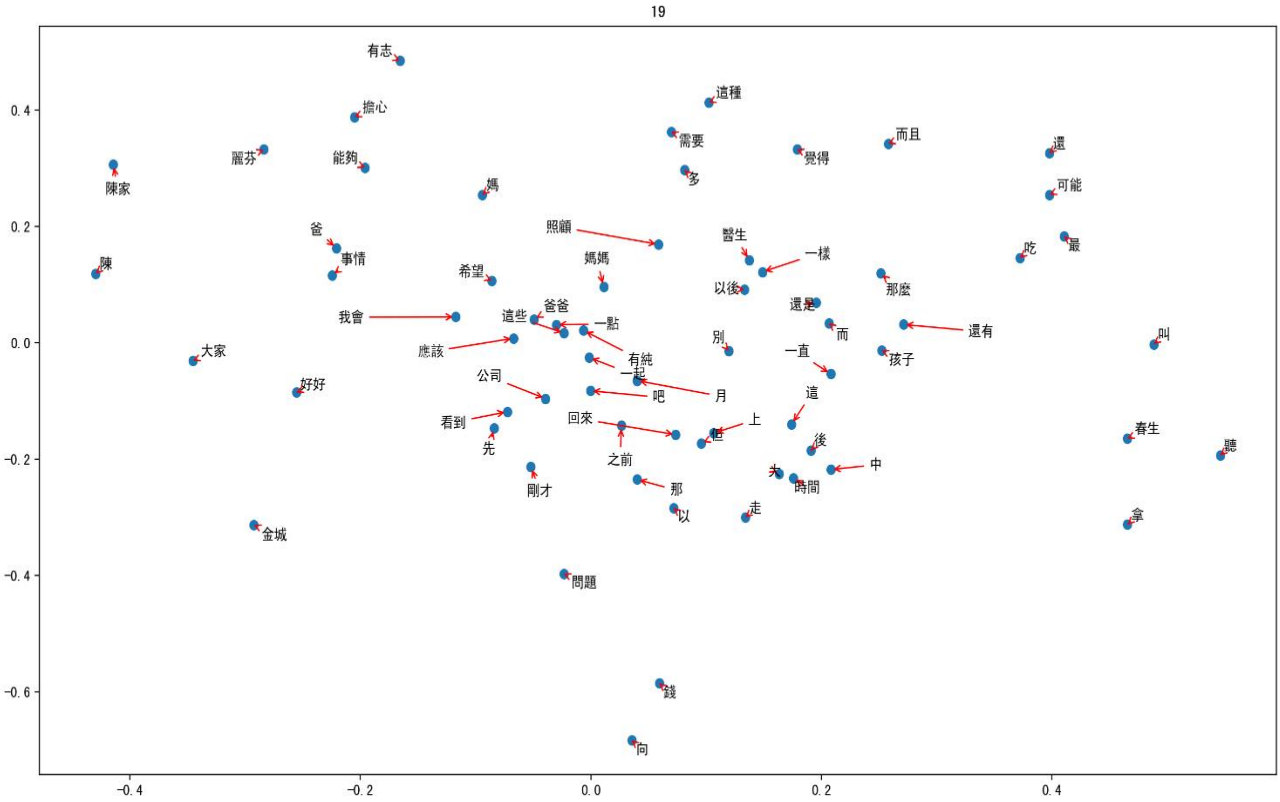
alpha: learning rate 初始值

min_count: 出現次數少於此的詞不納入計算考慮

workers: 決定開多少個 thread 來執行

iter: 跑幾個 epoch

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。

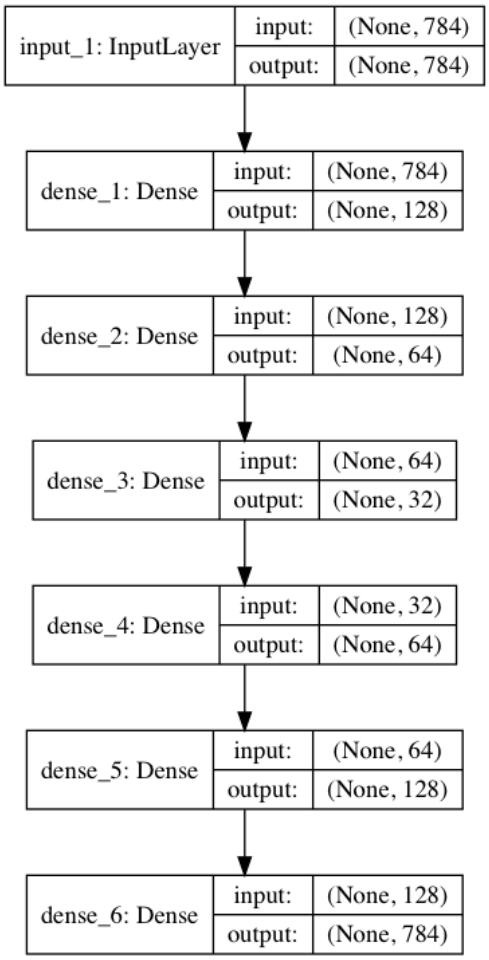
1. 可以發現爸→爸爸 與 媽→媽媽的向量方向十分相似
2. 陳家 與 陳 位置十分接近 推測麗芬也有可能屬於陳姓

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)

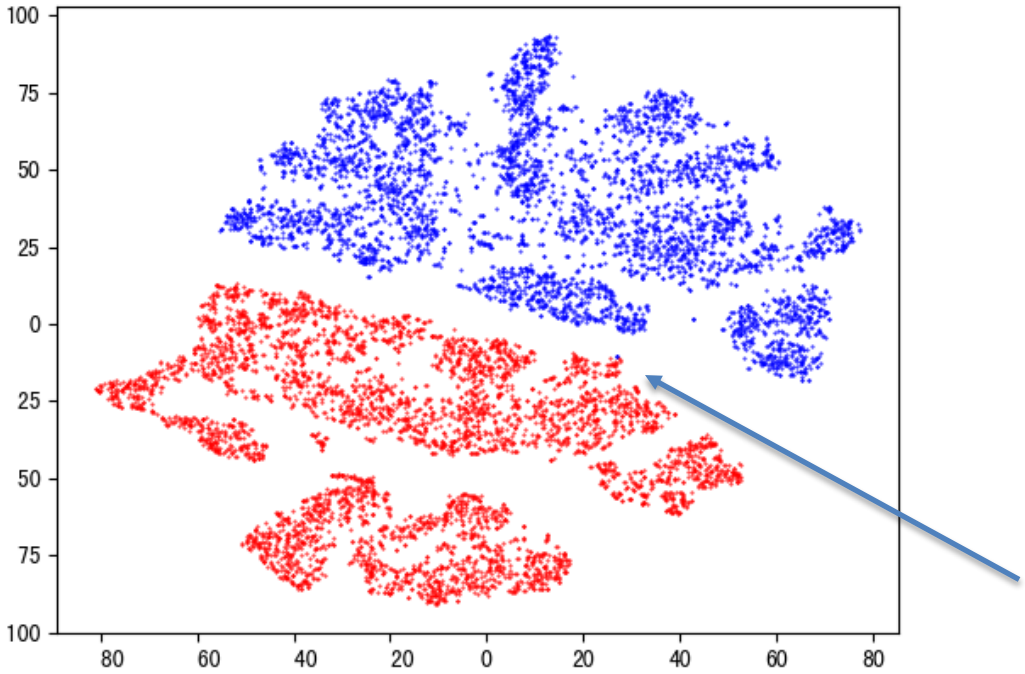
方法一：使用 PCA 降維到 32 維後再用 tsne 降維到 2 維，最後再用 kmean cluster 到 2 群。

方法二：normalize 後使用 autoencoder，最後再用 kmean cluster 到 2 群。而 normalize 的方式為除以 255 後以 np.mean(x, axis=0) 來減去每張圖片各自的平均值。而 autoencoder 的結構如以下



	Kaggle private	Kaggle public
方法一	0.04636	0.04621
方法二	0.96568	0.96976

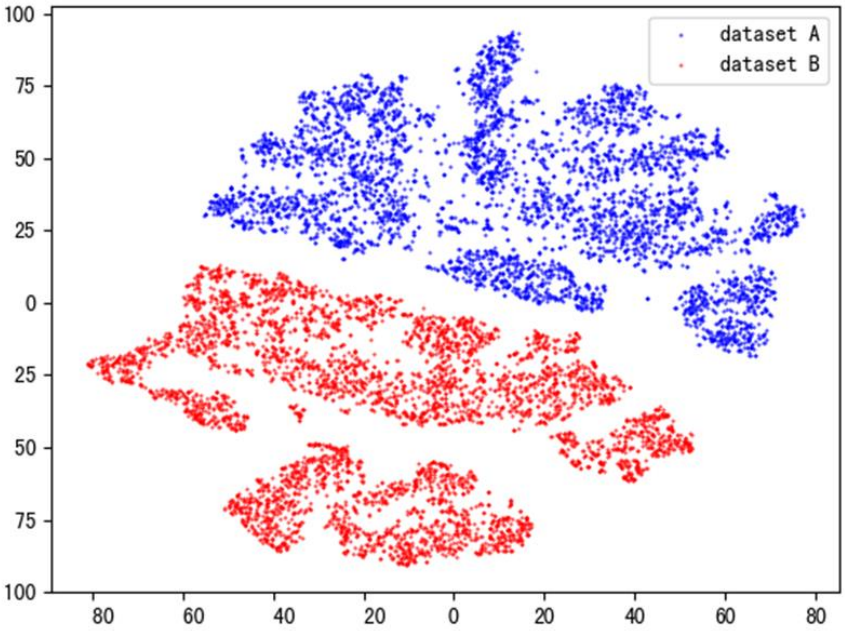
C.2. (.5%) 預測 visualization.npy 中的 label，在二維平面上視覺化 label 的分佈。



藍色:dataset A　　紅色:dataset B

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。

我實作的此模型在 kaggle 上的成績約為 0.97，因此可以發現上圖除了位於圖中心點偏右的那個紅色區域的藍點之外，其餘 label 都分割的十分正確(藍色箭頭指示處)。


dataset A
dataset B
100
75
50
25
0
25
50
75
100
80
60
40
20
0
20
40
60
80