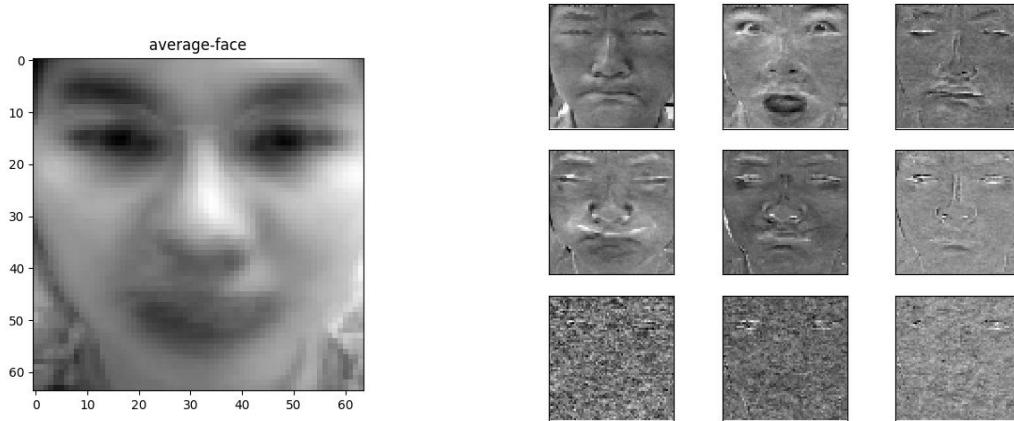


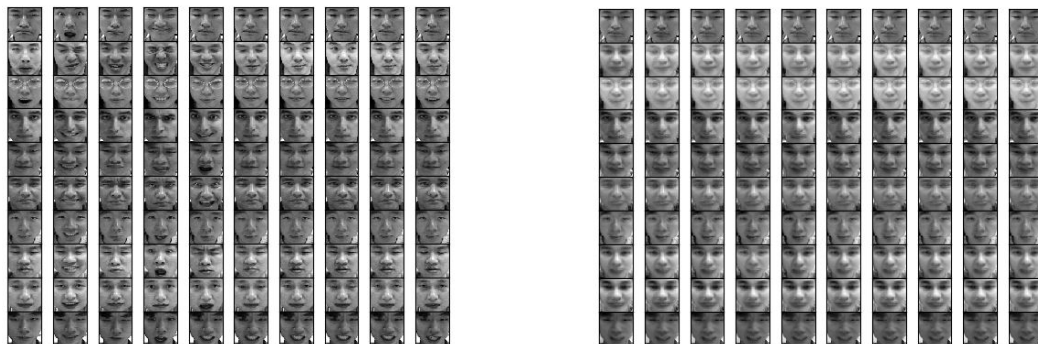
1.1. Dataset 中前 10 個人的前 10 張照片的平均臉和 PCA 得到的前 9 個 eigenfaces:

答：(左圖平均臉，右圖為 3x3 格狀 eigenfaces, 順序為 左到右再上到下)



1.2. Dataset 中前 10 個人的前 10 張照片的原始圖片和 reconstruct 圖 (用前 5 個 eigenfaces):

答：(左右各為 10x10 格狀的圖, 順序一樣是左到右再上到下)



1.3. Dataset 中前 10 個人的前 10 張照片投影到 top k eigenfaces 時就可以達到  $< 1\%$  的 reconstruction error.

答：(回答 k 是多少)

k = 59, reconstruction error = 0.9968%

2.1. 使用 word2vec toolkit 的各個參數的值與其意義:

答：

size = 500: 代表將 word 表示在多少維度的空間

window=10: 代表在當前的word之前後各有多少個word要視為context

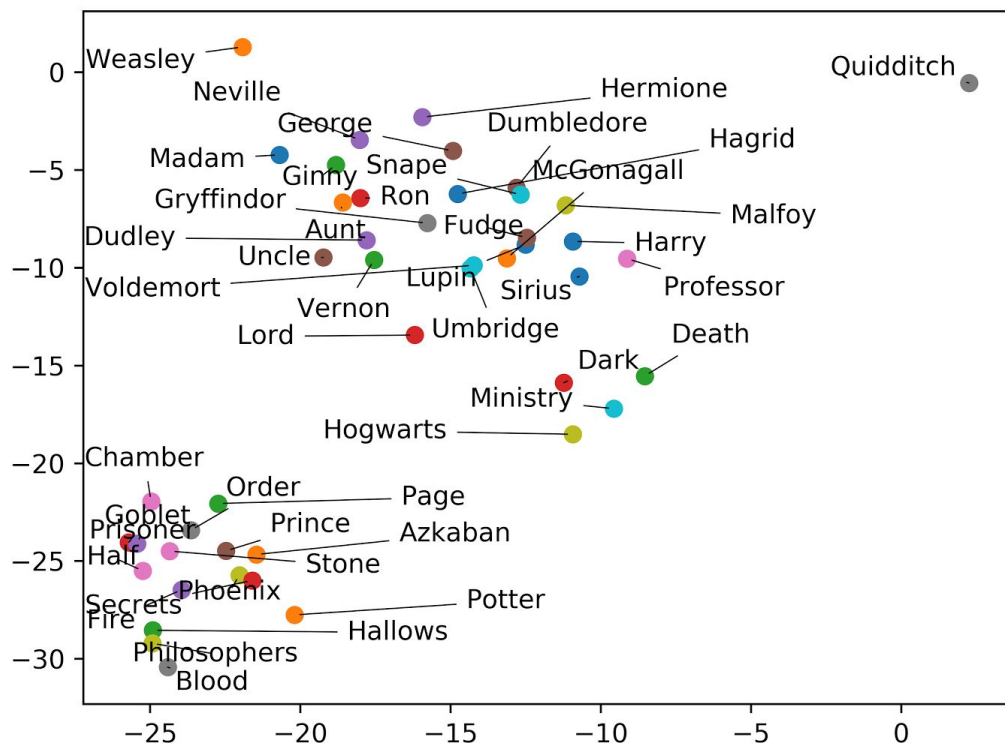
min\_count=10: 只考慮出現次數在min\_count以上的 word

alpha=0.1: learning rate

iter=20: the number of epochs

2.2. 將 word2vec 的結果投影到 2 維的圖:

答: (圖)



2.3. 從上題視覺化的圖中觀察到了什麼?

答:

大致上來說, 相近的word會被投影到相近的位置, 如 Ron和Ginny是兄妹, 因此被投影到了幾乎同一個位置, 另外右上角則有一個孤立的點Quidditch, 這是一個魔法運動項目, 由於沒有其他義近的詞, 因此只有它被投影到那個位置。另外比較特別的是Potter這個詞看似突兀的被投影到了左下角的位置, 我認為可能的原因是書名相關的詞會被投影到左下角的位置, 如: Secrets、Stone、Chamber、Azkaban、Prince等。

3.1. 請詳加解釋你估計原始維度的原理、合理性, 這方法的通用性如何?

答:

由於本題之 neural network之參數全部都是Gaussian random variable, 顯而易見得, 當input之維度愈大, output每一維的 variance也會愈大。藉由data generator, 可以事先產生維度從 1 到60的input對應的output之data points, 並計算其variance。因此只要計算testing data 之 variance, 並比較這個variance與我們預先產生的哪一個維度對應之variance最接近, 即可知道testing data的intrinsic dimension。

可惜的是，這個方法的通用性並不高，因為在實際的情況中，我們往往無法事先得知產生data的model是什麼，因此也就無法事先計算不同維度之input對應的output之variance是多少。然而，data之variance與intrinsic dimension之維度成正相關這個特性在大多數的情況下是成立的，只是這個特性在不知道model為何的情況下無法被善加利用。

3.2. 將你的方法做在 hand rotation sequence dataset 上得到什麼結果？合理嗎？請討論之。

答：