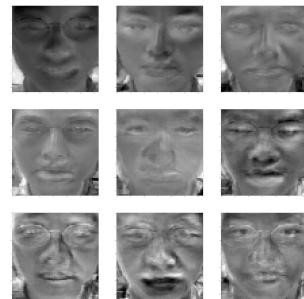
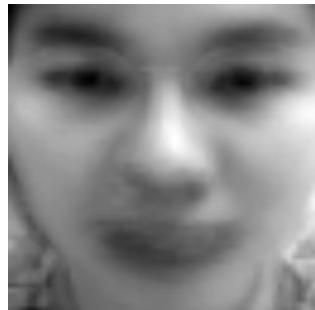
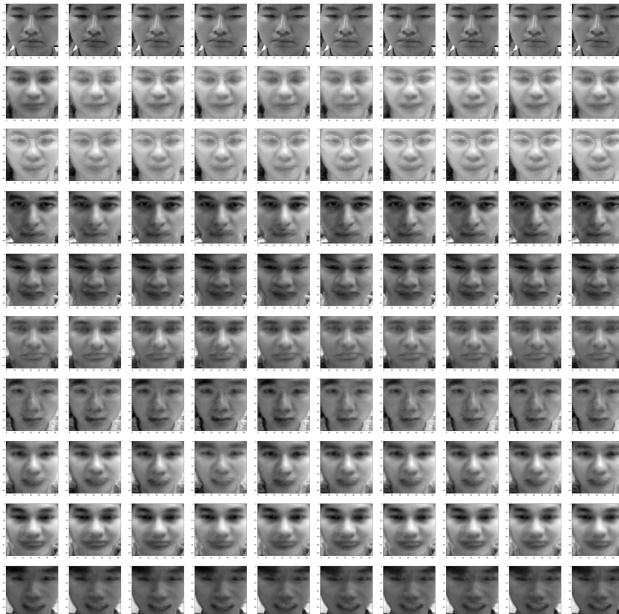


學號：R04921094 系級：電機碩二 姓名：葉孟元

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:



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):



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：如果/256 得出來的 k 是 59，如果/255 的出來的 k 是 60

2.1. 使用 word2vec toolkit 的各個參數的值與其意義：

答：有調整的參數

參數	意義	值
size	用多少維度表示一個 word vector，維度在 tsne 出來後感覺沒有太大的影響。	100
window	前後多少 word 會相互影響，好像比 default 5 小比較好	3
min_count	忽略最比這個詞頻還小的詞，想說只要 1000 個就過濾多一點	10
cbow	選用 skip-gram，雖然慢一點，但是效果比較好	0

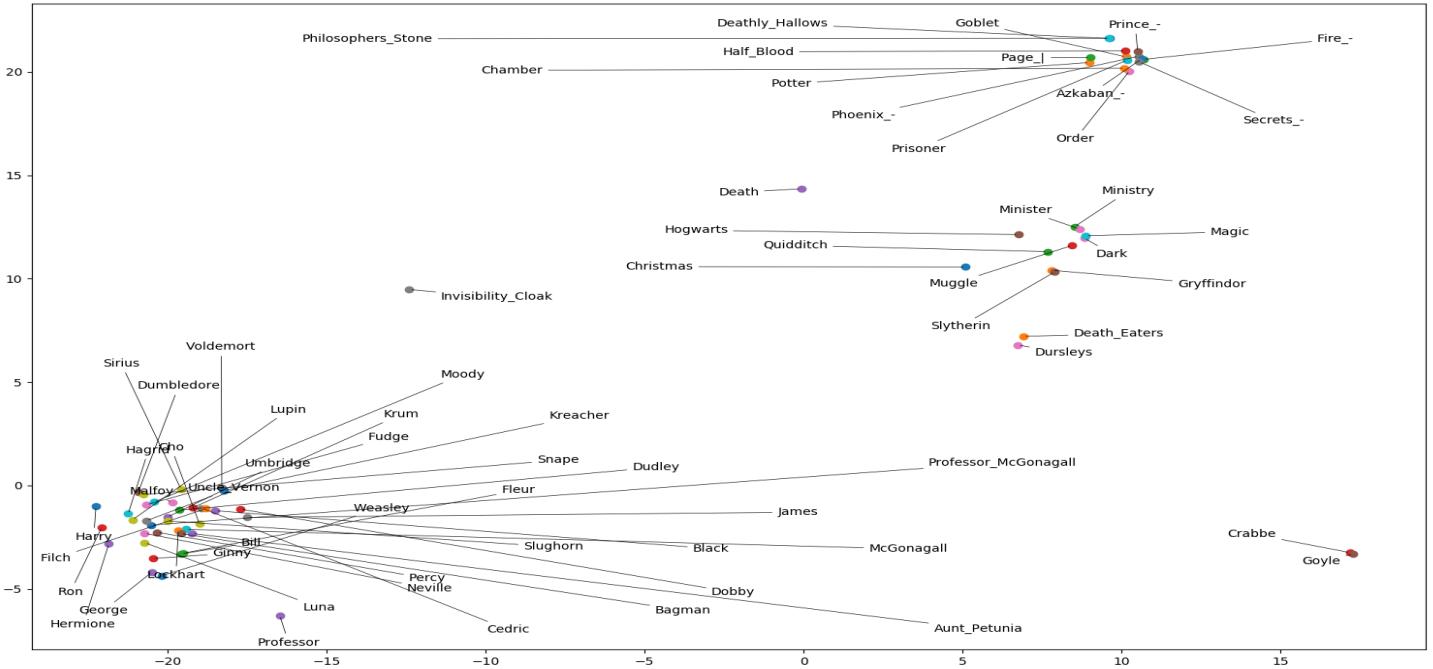
hs

是否使用 hierarchical softmax 降低計算量

0

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：圖中的點大部分是有道理的，比如右下角的 **Crabbe** 和 **Goyle** 兩人都是 **Malfoy** 的手下，所以兩人基本貼在一起。並且圖中 **Gryffindor** 和 **Slytherin** 也是幾乎重疊。右上角一簇基本上都是書名中的名詞，也是那一整部小說圍繞的主題。左下角有點混亂，但是基本上可以看出是人名，而關係精密的人相對較近，比如 **Ron** 和 **Harry** 還有 **Hermione**。

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：這個題目一開始拿到手的時候想用數學方法把 **data** 逆推回去，但是 **w** 和 **b** 最後的 **Gaussian** 的取值方式導致有很大的困難。為了觀察不同 **dataset** 特性畫出了他們的 **standard deviation**。發現一開始的 **dimension** 越低 **standard deviation** 也越低，就利用這個特性加上 **svr** 做出一個模型，再加上 **pca** 的 **explained_variance_ratio** 最後一個大於 **0.01** 的維度，可以達到 **0.11** 的 **public score**。在 **TA hour** 之前以為這樣的方法算是旁門左道，沒想到 **TA** 也利用了類似的方法。最後結合 **TA** 的方法，並自己再萃取出特別的資訊。總共六個 **feature**，分別是 **data std**, **ratio std**, **ratio** 超過 **0.01** 的維度，助教提供 **eigen value mean**, **eigen value std**, **eigen value** 和 **y = 100x** 的估計交點。結果可以達到 **public : 0.05813 private : 0.05583**。這個方法通用性在我看來只有 **explained_variance_ratio** 和助教提供的 **eigen value** 是有用的，**ratio** 主要看的是 **ratio** 自己的差分，如果差分在某個維度開始之後基本不變就說明後面的維度基本沒有意義。並結合數值大小來一起判斷。助教提供的方法一樣是看差分，或

者說是曲線彎曲的程度，彎曲程度越大說明維度越低，這就是我為什麼利用 $y = 100x$ 來估計維度。

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：我認為這個 **dataset** 是 **3–4** 維之間。圖中黃色代表差分，藍色就是 **ratio**。在從下方放大圖看得出，字 **3–4** 綴的時候（0 代表 1 綴），不管是差分還是 **ratio** 都維持在一個很低的數值，說明在更大的維度上 **variance** 並不大，也就說明哪些維度是多餘的，所以這個 **dataset** 的維度最多 **4** 綴，我認為 **3** 綴比較有可能，因為畢竟排除時間，這個世界是 **3** 綴的世界，應該合理。

