Machine Learning (2017, Fall)

Assignment 6 – Unsupervised learning & dimension reduction

學號：b03901086　系級：電機四　姓名：楊正彥

Part A: PCA of colored faces

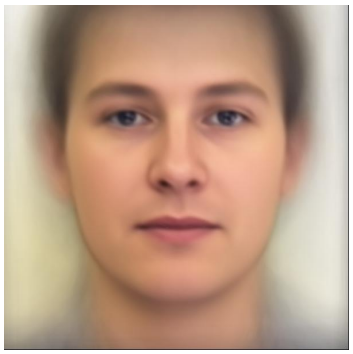

Fig 1. 所有臉的平均

Fig 2. 前四個 Eigenfaces，所佔比重分別為 4.1%,2.9%,2.4%以及 2.2%

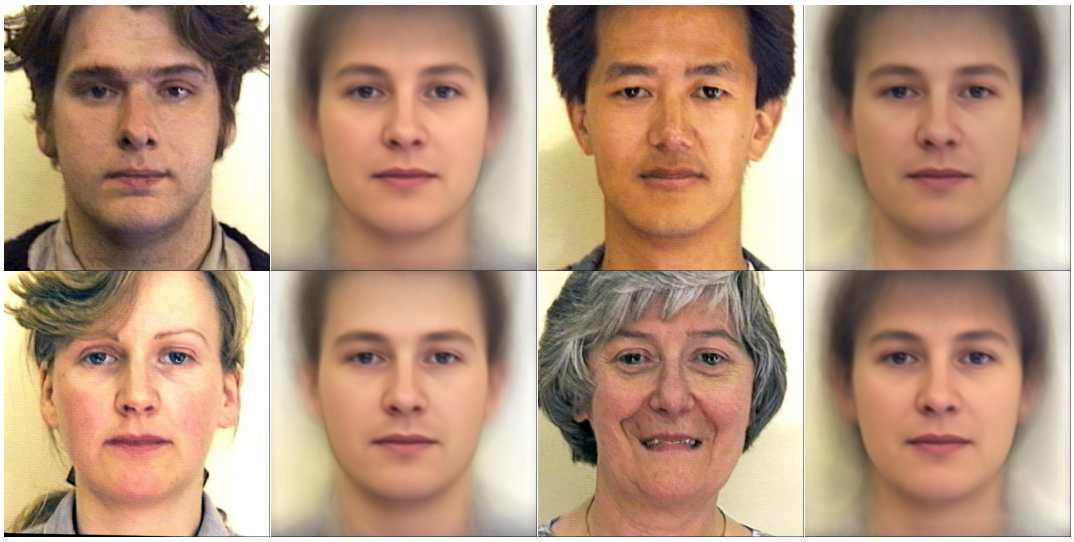

Fig 3. 用前四個 Eigenfaces 進行的 reconstruction，id 分別為 8,47,196 和 336

Part B: Visualization of Chinese word embedding

這次作業使用的 word2vec 套件是 gensim，選用的相關參數為 dimension=200、window size=5、min count=10，剩下的皆為 default 以及這次作業的規定。除 min count 和 dimension 是參考寫 final 時的參數外，因為選用的句子有比較長，所以 window size 可以稍微調大一點讓 word embedding 可以學到更多 semantic representation。

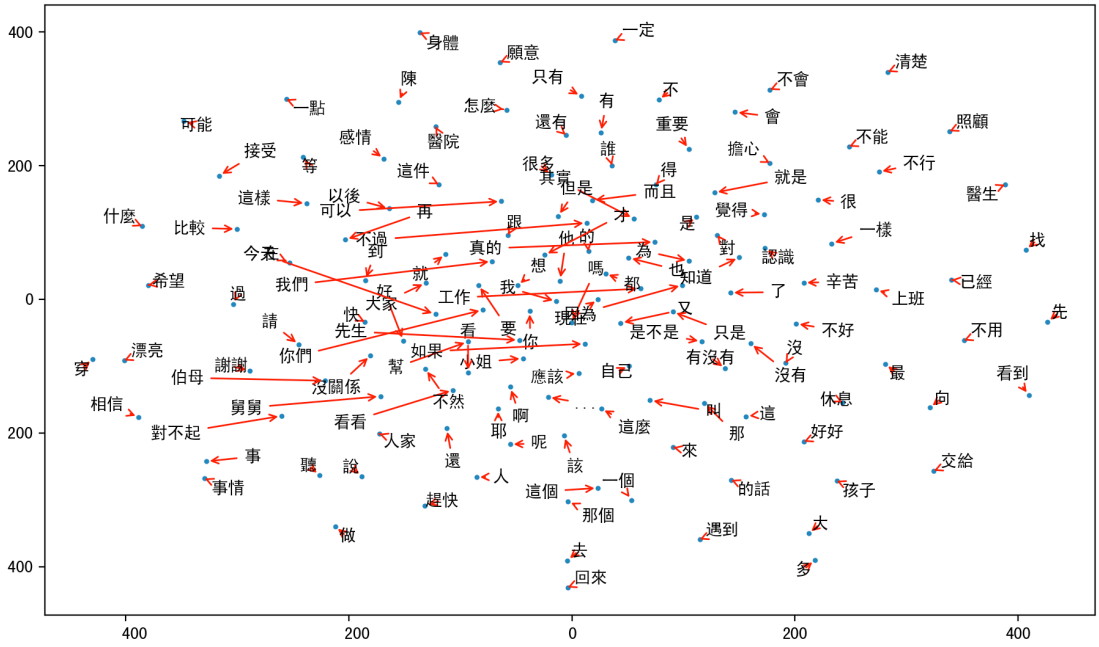

Fig 4. Visualization of t-SNE Chinese word embedding

可以看出來有著相近的詞義的 word 在 t-SNE 降維之後會比較靠近，舉例來說「那個、這個」、「但是&其實」、「是不是&有沒有」等，理論上 word embedding 的 visualization 應該還可以看出兩組詞之間於 vector space 相對關係，但這次在我們的實作中沒有找到較為明顯的例子。

Part C: Image clustering

這次作業我實做了兩種降維方法，一種是用只 auto-encoder 降到 32 維而另外一種則是用 auto-encoder 降到 128 維之後再用 t-SNE 降到 32 維，其於 kaggle 上的表現如下表：

	Private Score	Public Score
Auto-encoder only	0.99556	0.99602
Auto-encoder + t-SNE	0.99308	0.99296

Auto-encoder 的表現比加上 t-SNE 還要好，但是其實就結果來講是差不多的，但是加上 t-SNE 的運算時間是遠超過只用 Auto-encoder 的。

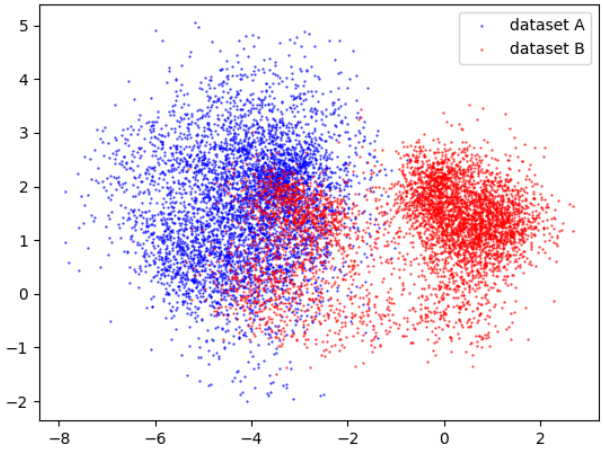

Fig 5. 取 t-SNE 前兩維作圖

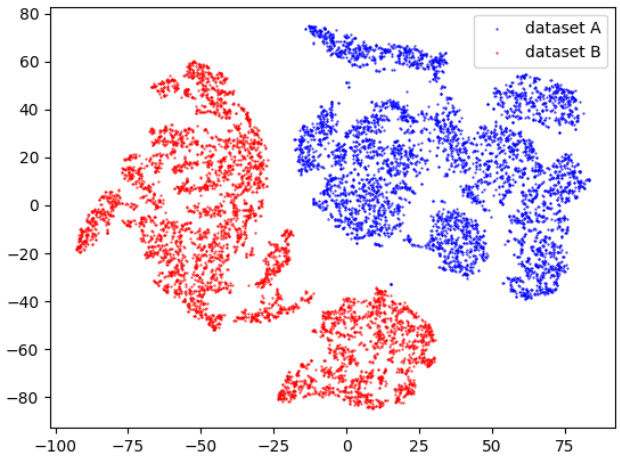

Fig 6. 利用自己的 encoder 做 predict 之後以 t-SNE 降到二維作圖

從上圖可以看出我們這次所 train 的 auto-encoder 在 clustering 這部分有相當不錯的表現，dataset A 和 dataset B 在平面上分得相當開。