人臉辨識

**動機:**

**在PTT版上和現實中，時常聽到別人說IT人有IT味，聽起來好像沒有任何依據，但好像隱隱約約IT人又有共同的特徵，所以我決定做一個專案來找出這個問題到底有沒有答案。**

**設計流程:**

1. **比較人工搜尋資料和爬蟲資料後，我決定把爬蟲的資料當成 train data 和validation data，而人工資料當成 test data。**
2. **由於每張照片拍攝方式角度不同，決定先用CascadeClassifier來偵測臉部特徵並擷取，並且篩選掉團體照的資料。**
3. **選擇CNN建立模型預測**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **平均一筆耗費時間** | **優點** | **缺點** |
| **人工搜尋資料** | **1min** | **資料標籤無誤** | **朋友照片太侷限於年紀** |
| **爬蟲** | **1second** | **速度快** | **資料標籤可能有誤** |

**遇到問題**

1. **Linkedin的資料都是使用js渲染**
2. **Linkedin上的圖片都有經過編碼**
3. **第一次只抓取50張軟體工程師和50張業務照片作為資料，如圖一，很明顯看到我的訓練資料接近100%，而validation卻上下不定跳動。**

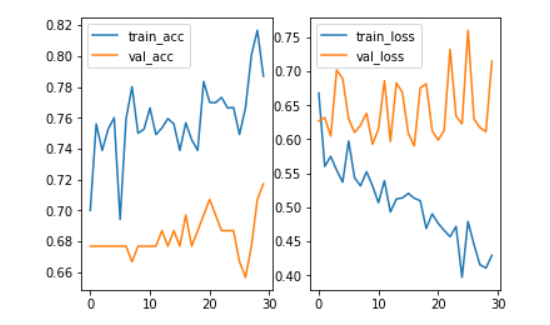
**解決方法**

1. **使用selenium登入，並且用selenium操作網頁 ex: 輸入值和滾動網頁...等，才能順利抓出網頁資訊。**
2. **先把Linkedin上的圖片網址抓下來，再拿來比對，找出編碼規則，在把它解碼回並存入。**
3. **由於我的模型並未使用複雜的很多層疊加，所以我增加資料量來避免overfitting，增加數量至工程師500張和非工程師500張。**

**圖一**

表12


**圖二**

****

**結論和檢討**

**這次的人臉辨識專案，如圖二，驗證資料集準確度大概70%，比我預期中的好一些，但在測試資料集中準確率非常糟糕，但我後來檢查一下機器辨識錯誤的範例，跟性別有相當的關西，女生通常都分到非工程師，所以我覺得我這次的專案仍有許多改進的地方，因為此專案是人眼無法分辨的，所以我的數據量應該要非常大量才能驗證此問題，並且我得蒐集各種不同職業，不能僅僅侷限於業務和工程師，此外我也該把性別和年齡分開訓練，才能得到較佳的結果。**

**最後，此專案純粹是興趣驅使，之後的一些改進想法，可能會花上更多的時間在蒐集資料上面，所以我決定先把此專案先放在一邊，也從此專案更了解一個完整的專案是多麼的複雜。從問題的定義、蒐集資料、清理資料、視覺化和建模，模型只是一個解決問題的工具，並非萬能，專案的每個過程都是極為重要。**