數據科學期末報告

M132040019 廖廣筑

M122040017 吳俞憲

TABLE OF CONTENTS

錯誤標籤處理

錯誤標籤處理

必要性：圖片標籤的錯誤會導致模型學習到錯誤的分類模式，所以必須優先解決

處理方法：利用手動清理的方式，以下會解釋何者為錯誤標籤(分三類)

1.圖形清楚，但分類錯誤

vi_147

viii_254

2.圖形非數字或無法辨認

iii_16

i_32

3.模玲兩可的圖形

iii_108

vi_27

錯誤標籤處理

處理方式：

第一類：對於分類錯誤的圖形，我們會手動把它放回正確的資料夾內

第二類：因為非羅馬數字，我們會直接把資料刪除

第三類：因為不能確定原始圖形是屬於哪一類別，所以也直接刪除

初步結果：

訓練集刪除144張，驗證集刪除3張

準確度可以上升至0.7以上，代表處理此問題可以幫助模型分辨得更好

資料清理與擴增

資料擴增

原因:

前面處理錯誤標籤有刪除掉資料，且深度學習的模型下往往都是資料越多的情況下表現較好，所以需要找尋方法去擴增資料

方法:

1. 加入外部資料
2. 利用圖片增強

外部資料

資料集：基於Chars74K dataset而得來的圖片

處理方式：

1. Chars74K dataset中我們感興趣的圖形只有以下6種

所以必須要透過OpenCV 來組合成需要的羅馬數字

2. 後續再加入隨機雜訊、隨機膨脹字母厚度、手動添加筆畫

外部資料

好處：

1. **資料的多樣性**：加入其他來源的資料可以讓模型學會更多樣化的特徵，讓模型學習的更全面。

2. **增加樣本數量**：能提供更多訓練資料，尤其在模型固定的情況下

初步結果：

- 每個類別(訓練集、驗證集)都增加了70~90張
- 模型準確度能提升至0.75以上，也是有顯著提升的

資料清理

常見問題(cleanvision)：

1. Low Information：缺乏明顯的可辨別特徵
2. Grayscale：影像是灰階影像，沒有色彩資訊
3. Odd Size：影像大小或解析度與其他不同
4. Light：影像可能過於明亮
5. Near Duplicates：影像相似度高
6. Dark：影像可能過於暗
7. Blurry：圖像過於模糊

資料清理(訓練集)

	issue_type	num_images
0	low_information	3988
1	grayscale	2756
2	light	1196
3	odd_size	822
4	near_duplicates	82
5	dark	0
6	odd_aspect_ratio	0
7	blurry	0
8	exact_duplicates	0

原始資料數量：4043

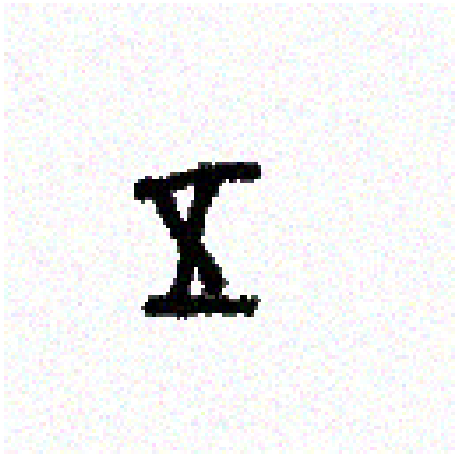
- 低資訊：3988
- 過亮：1196
- 灰階：2756
- 相似影像：82
- 異常大小：822

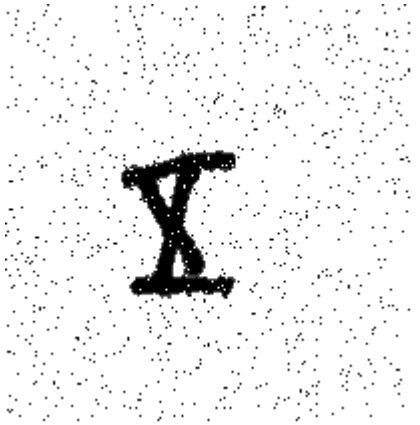
資料清理(訓練集)

- **相似影像:**總共82張，所以會刪除一組的其中一張(41張)
- **過亮影像:**加入高斯雜訊(平均0，標準差0.05)，亮度為原本的70%

	issue_type	num_images
0	low_information	2760
1	grayscale	1826
2	odd_size	642
3	near_duplicates	2
4	dark	0
5	blurry	0
6	odd_aspect_ratio	0
7	light	0
8	exact_duplicates	0

- 過亮影像的問題已解決
- 相似圖形是因為有增加新圖片的原因





數據增強(訓練集)

目的：讓模型學習到更多不同的圖片，目標是將訓練集擴增到9000張

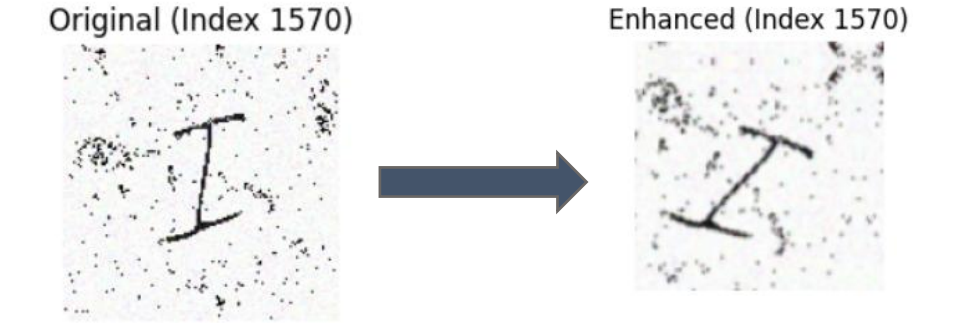
方法：

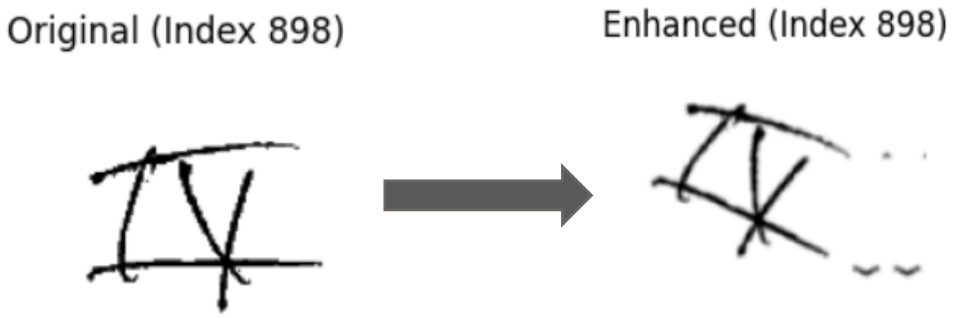
- 隨機旋轉：旋轉範圍是 ±10%
- 隨機縮放：縮放範圍是 ±20%
- 隨機平移：高度寬度各15%
- 隨機對比度：範圍是 0.1 到 0.2

結果：此過程總共增加了6501張照片，目前圖片數量為10503

數據增強(訓練集)

增強完效果：





資料清理(訓練集)

	issue_type	num_images
0	low_information	4151
1	grayscale	1826
2	odd_size	855
3	light	354
4	near_duplicates	8
5	dark	0
6	odd_aspect_ratio	0
7	blurry	0
8	exact_duplicates	0

增強後資料：10503

- 低資訊：3988->4151(增強前、後)
- 過亮：354
- 灰階：1826
- 相似影像：8
- 異常大小：855

後續步驟：

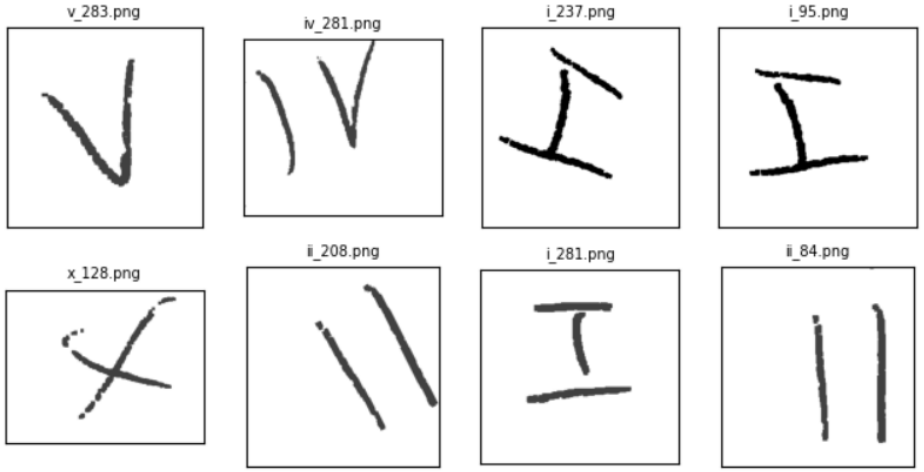
1. 處理過亮影像與相似影像
2. 刪除部分低資訊的照片，使照片數量為9000

資料清理(訓練集)

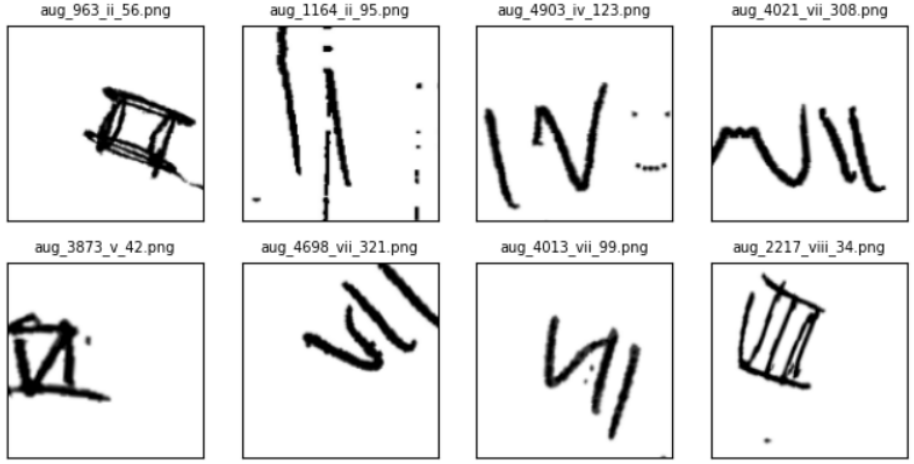
目標：需要刪除掉圖片來讓圖片達到9000張

方法：從低品質的圖片中刪除圖片

Information scroe：Low



Information scroe：High



最終會刪除分數最高的1498張使訓練集達到9000張

資料清理(驗證集)

	issue_type	num_images
0	low_information	959
1	grayscale	893
2	light	620
3	near_duplicates	22
4	odd_size	5
5	dark	0
6	odd_aspect_ratio	0
7	blurry	0
8	exact_duplicates	0

原始資料數量：960

- 低資訊：959
- 過亮：620
- 灰階：893
- 相似影像：22
- 異常大小：5

資料清理(驗證集)

	issue_type	num_images
0	low_information	331
1	grayscale	300
2	near_duplicates	2
3	dark	0
4	light	0
5	blurry	0
6	odd_aspect_ratio	0
7	odd_size	0
8	exact_duplicates	0

處理方式：

1. 刪除相似圖片其中一張(44張相似)
2. 處理過亮影像也是使用高斯雜訊的方式

➢ 最終在驗證集會有949張圖片

數據增強(驗證集)

目的：增加圖片使得模型性能的評估穩定，目標增加到2940張

方法：

- 隨機旋轉：旋轉範圍是 ±5%
- 隨機縮放：縮放範圍是 ±10%
- 隨機平移：高度寬度各10%
- 隨機對比度：範圍是 0.05 到 0.1

結果：此過程總共增加了2316張照片，目前圖片數量為3265張

資料清理(驗證集)

	issue_type	num_images
0	low_information	861
1	light	495
2	odd_size	332
3	grayscale	300
4	near_duplicates	22
5	dark	0
6	odd_aspect_ratio	0
7	blurry	0
8	exact_duplicates	0

增強後資料：3261

- 低資訊：331->861(增強前、後)
- 過亮：495
- 灰階：300
- 相似影像：22
- 異常大小：332

後續步驟:

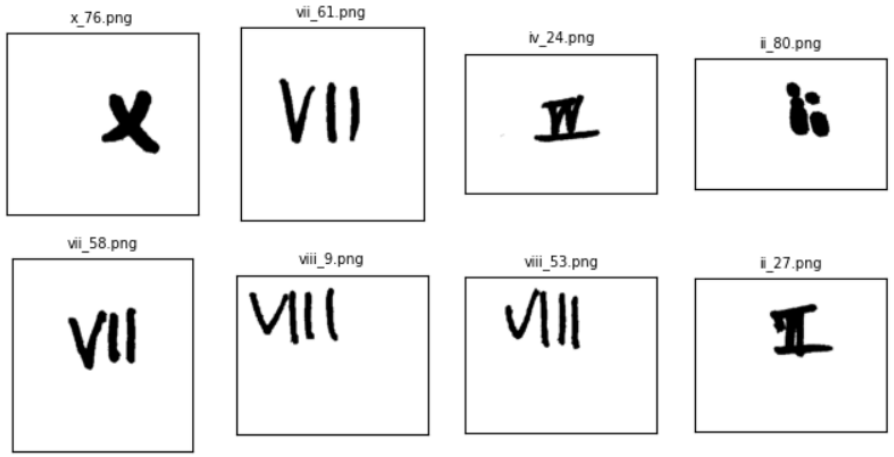
1. 處理過亮影像與相似影像
2. 刪除部分低資訊的照片，使照片數量為2940

資料清理(驗證集)

目標：需要刪除掉圖片來讓圖片達到2940張

方法：從低品質的圖片中刪除圖片

Information scroe：Low



Information scroe：High

最終會刪除分數最高的321張使驗證集達到2940張

結論

1. 錯誤的標籤會使模型效果不好，所以再做其他處理前要優先解決這個問題。
2. 數據清理與增強也是很重要的一環，清理資料可以去除異常資料；增強能讓模型看到更多不一樣的圖片，可以很好的訓練模型的分辨能力。
3. 這次清理錯誤標籤是利用手動處理的，但之後如果還有類似的問題，我們會嘗試利用自動化的方式去處理。

參考資料

外部資料來源:
https://www.kaggle.com/datasets/agneev/basedonenglishhandwrittencharactersmodified

外部資料處理:
https://agneevmukherjee.github.io/agneev-blog/preparing-a-Roman-MNIST/

老師講義:

THANKS