111學年度第二學期
數據科學實務與創新

期末報告
111-2
PRACTICAL AND INNOVATIVE ANALYTICS IN DATA SCIENCE

Final Report
主題：Roman Number Classification

Abstract

透過clean_lab、image_lab等方法清洗資料，接著進行資料增強來對羅馬數字進行分類。

組員：

M102040035 林良烽

M112040034 李祐瑄

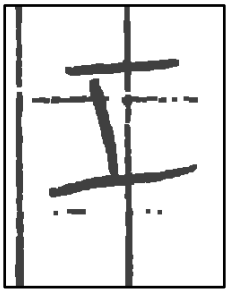
M112040036 孫瑞鴻

目錄

1. Observation

在進行期末報告中，我們研究了羅馬數字i~x的分類問題。在原始資料集中，我們注意到一些可能影響分類模型的問題，這些觀察對我們的研究和結果具有重要意義。

首先，我們發現在資料集中存在一些圖像雜訊很多的情況(圖一)。這些雜訊對於圖像辨識的精度產生了負面影響，因為它們增加了圖像中不相關的細節，使模型難以準確識別羅馬數字。



▲圖一

▲圖二

▲圖三

其次，我們也觀察到在原始資料集中存在一些圖像的類別分類錯誤。例如，某些圖像被錯誤地歸類到了不屬於它們的羅馬數字資料夾中。這些分類錯誤對於訓練和評估模型時的準確性提出了挑戰，因為它們引入了噪音和不一致性，使模型難以學習和泛化。

此外，我們還注意到在資料集中存在一些圖像，它們並不代表真實的羅馬數字，而是包含亂畫(圖二)或是笑臉(圖三)的圖像。這些額外的圖像類別對於我們的模型訓練和測試帶來了額外的挑戰，因為我們需要清理或是過濾這些不相關的圖像。

2. Data Cleaning

A. Load Data

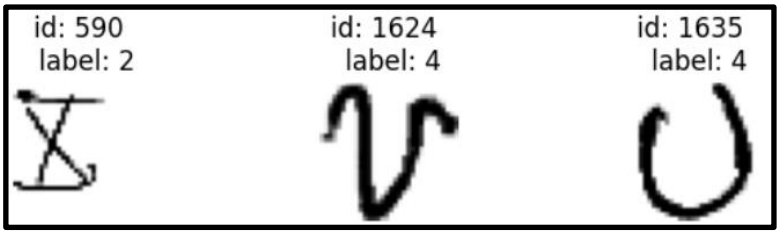
起初嘗試了將訓練資料shuffle讀取進來，經過測試後發現如果shuffle = True，那麼模型更容易正確地找出標籤錯誤的圖片，但是由於資料進行shuffle完後，會造成index難以對應回原本資料集，因此最後還是使用shuffle = False。

B. Build Model

首先我們嘗試了多種架構的pretrained model，包括ResNet50、ResNet101、ResNet152、EfficientNetB0等，測試後發現一般常使用的ResNet50並不如預期的來得好，推測原因可能是資料較髒，而ResNet50模型過於簡單，ResNet101以及ResNet152表現較好且兩者表現相似，最後我們選用模型複雜程度適中的ResNet101來作為預測是否有錯誤標籤的模型。

C. Find Label Issues

嘗試透過find_label_issues套件直接找出有一定機率標籤錯誤的圖片，接著直接刪除全部標籤錯誤的圖片，也有嘗試只刪除部分標籤錯誤照片，accuracy跟一開始相比雖然有比較好，但並沒有顯著提升。



▲圖四

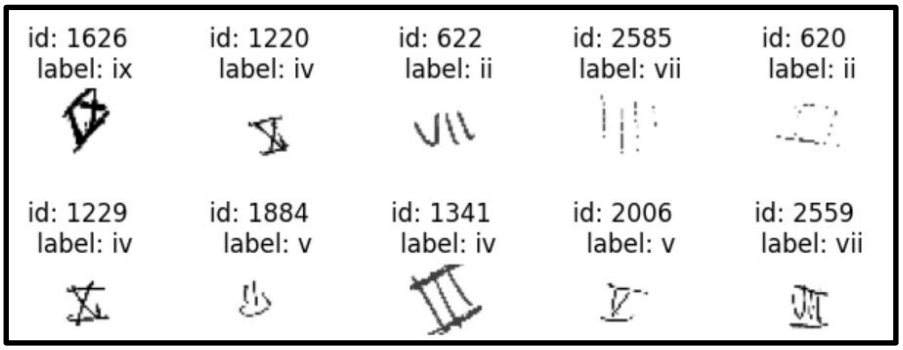
D. Datalab

	is_label_issue	label_score	given_label	predicted_label
1626	True	1.378993e-08	8	4
1220	True	5.552140e-08	3	9
622	True	2.805326e-07	1	6
2585	True	3.255173e-07	6	2
620	True	4.841384e-07	1	2
1229	True	7.762455e-07	3	9
1884	True	9.841264e-07	4	8
1341	True	1.187198e-06	3	2
2006	True	2.263730e-06	4	9
2559	True	3.335242e-06	6	2

▲圖五

我們經過了pretrained model(ResNet101)，經由20次的交叉驗證，得到的pred_prob(預測機率值)，使用Cleanlab中的datalab，找出存在標籤錯誤的有920張圖片。由圖五來看，列index為shuffle = False後固定的id。而先按照是存在label_issue的情況下(True)，且label_score由小排到大的數值。given_label為原先放置的資料夾；predicted_label為Datalab套件期待放置的資料夾

根據Datalab中，給定label_score < 0.01 作為閥值基準，共計266張圖片，再進行二次確認。由圖六所示，如id_1626及id_2006的given_label，我們認為與圖片是相互呼應的標籤。因此，從266張裡面，將類似於上述兩個id的情形挑出，進行二次確認，共找到73張；然而，剩餘的193個進行移動到其對應的pred_label資料夾。截至目前為止，保持著train與val的張數與老師原始資料的同樣情況。



▲圖六

3. Data Augmentation

A. Gray Scale

雖然我們通過直接觀察圖片，幾乎都是黑白的，但通過程式去驗證(image.mode)，我們發現實際上資料中有些圖片內容是多通道的，因此我們採用了將所有圖片灰度化的處理，使圖片為單通道的。

B. Generating images through rotation and scaling

通過縮放和旋轉的操作，以生成更多的訓練樣本或擴增數據集。透過縮放和旋轉，可以生成具有不同尺寸和角度的圖像，增加數據的多樣性，改善模型的魯棒性和泛化能力。例如，在物體檢測和圖像分類任務中，可以通過縮放和旋轉來模擬不同距離和觀察角度下的圖像變化。

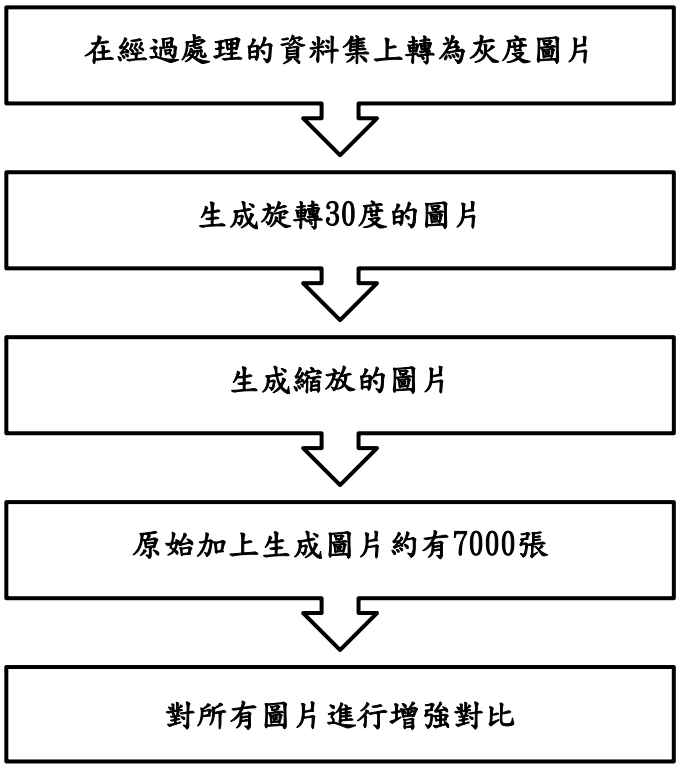
分別是把所有圖片的60%，生成旋轉30度的圖，再把所有圖片的10%，生成縮小過後的圖(scale factor=0.8)，最終整個資料集約有7000張的圖像。

C. Enhancing image contrast

進行增強對比可以突出字體的細節和形狀，提升視覺效果，同時有助於改善模型對筆劃和形態的辨識，從而提高預測準確性。

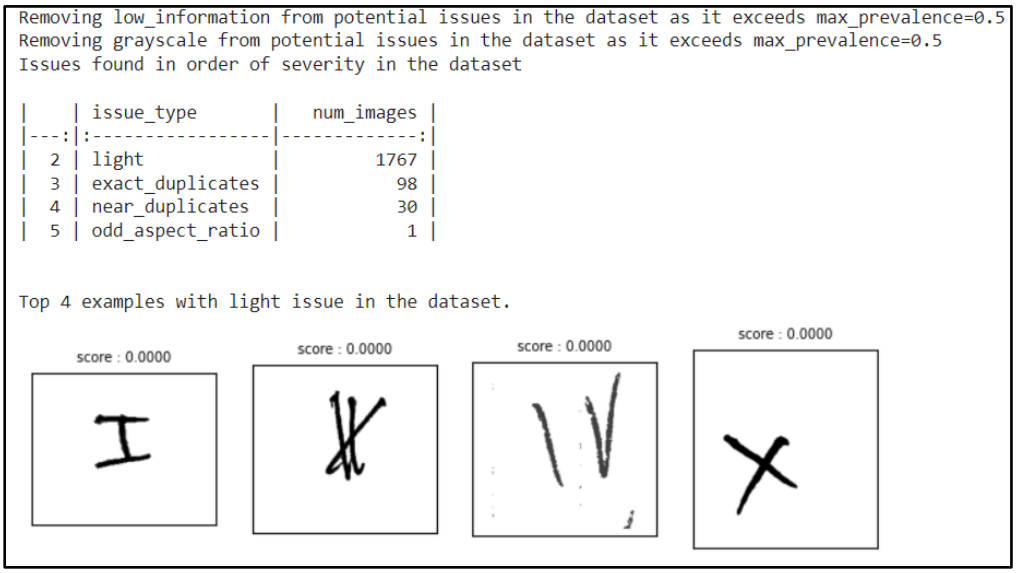
對於清理並生成後的資料集，對所有圖像進行對比度的增強(factor=1.5)

D. Workflow



經過以上過程來生成圖像進行資料增強，最終在test data上的準確率最高可以達到0.78，將其產出的submission.csv上傳到Kaggle約為0.75。

4. Extra attempt——Clean Vision



▲圖七

我們使用了Clean Vision的Image Lab(圖七)來查看標籤問題，例如針對有light issue的圖像進行處理，嘗試刪除了所有具有光線問題的圖片，或者只刪除了其中一部分。此外，還嘗試了對光線問題的圖片進行亮度和對比度的調整，再使用了第三節提到的資料增強方法。然而，最佳結果只達到0.75的準確率，並沒有優於未使用這些操作的結果。綜合分析，我們認為這些嘗試並未能有效改善模型的預測分類準確率。可能的原因是光線問題並非主要影響模型性能的因素，或者我們選擇的處理方法並不適用於該問題。

在未來的工作中，我們將繼續探索其他更有效的資料增強方法或是其他的前處理技術，以提高預測分類模型的準確率。同時，我也將尋找其他可能影響模型性能的因素，以進一步改進羅馬數字圖像辨識的結果。

5. Contributions

Build Model & Find label issues	孫瑞鴻
Datalab & Double check	林良烽
Data Augmentation & Clean Vision	李祐瑄
投影片與 Code 彙整	孫瑞鴻、林良烽、李祐瑄
書面報告	孫瑞鴻、林良烽、李祐瑄

6. Reference

Cleanlab:
https://colab.research.google.com/drive/1PrNq4zoVk2wa5AoXHxDwWtwOMLv9QfKL?usp=sharing#scrollTo=TtYsKAXVbnmC
Datalab:
https://docs.cleanlab.ai/stable/tutorials/datalab/datalab_quickstart.html
ResNet:
https://www.twblogs.net/a/5d4ca562bd9eee541c30dc30