期末報告

M132040019 廖廣筑

M122040017 吳俞憲

I. Abstract

本報告詳細描述了為了可以讓固定的模型更好的辨識手寫羅馬數字，對數據集進行擴增和清理的過程。首先我們先用手動的方式去修改錯誤標籤，之後使用 CleanVision 去找出資料集的影像問題進行清理解決，利用 TensorFlow 的數據增強技術，我們對影像進行了隨機旋轉、縮放、平移及對比度調整等操作以達到我們想要的影像張數。最終生成的影像進行訓練，去預測測試集。

II. Introduction and related work

手寫羅馬數字的辨識在電腦視覺領域是一項具有挑戰性的任務，主要因為不同個體的書寫風格具有高度多樣性。能夠準確辨識手寫字符對於自動化評分系統、歷史文獻數字化以及用戶身份驗證系統等應用至關重要。因此這份報告探討了如何在不能動模型架構的情況下，提高訓練資料的品質，以提供更好的資料讓模型更好的分辨羅馬數字。

III. Dataset and methods

資料集

本資料集含有 3367 張訓練集影像與 963 張驗證集影像，以及 52 張測試集影像。這些影像以灰階形式存儲，每張影像包含單個手寫的羅馬數字字符，字符的大小和比例隨書寫者的習慣有所不同。同時也包含一些容易混淆模型的錯誤影像。

方法

1. **錯誤標籤的處理：**

為了確保數據標籤的準確性，我們首先手動檢查並修正訓練集中可能存在的錯誤標籤。這一步主要依賴人工判斷，逐一核對影像與其對應標籤，將標註錯誤的數據重新標記正確類別，以減少標籤錯誤對模型訓練的影響。以下進行舉例哪些影像我們會處理：

a. 圖形清楚，但分類錯誤(圖 1、圖 2):

vi_147

圖 1

viii_254

圖 2

處理方式：將錯誤分類的影像手動移到正確的資料夾內。

b. 圖形非數字或無法辨認(圖 3、圖 4)

iii_16

圖 3

i_32

圖 4

處理方式：直接將影像進行刪除。

c. 模玲兩可的圖形(圖 5、圖 6):

iii_108

圖 5

vi_27

圖 6

處理方式：因為會影響模型搞混，所以直接進行刪除。

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。

2. **加入外部資料**：

目的:

為了增加圖片的多元性，增加一些跟原本來源不同的照片可以幫助模型學習到更多元的數據。尤其在固定模型之下，增加新照片是最好提升模型準確度的一種方法。

資料處理:

此資料集作者是從 Chars74k dataset 組合出我們需要的羅馬數字，從資料集中取出 6 個可以組合出羅馬數字的圖片(圖 7)，後續再用 opencv 的方式組合出羅馬數字一到十。後續再隨機加入三種雜訊(高斯雜訊、均勻雜訊、脈衝雜訊)、隨機膨脹字母厚度、手動添加筆畫，讓圖片模擬現實中手寫的部分(圖 8)。最終在訓練集集驗證集各 10 個類別都有 70~90 張圖片

圖 7

圖 8

目前結果:

處理完錯誤標籤且加入這個作者的資料後，我們模型的準確度已經到達 0.75 以上，是有顯著提升的。

3. **進行資料清理**：

作法:

因為圖片可能會有各式各樣的問題導致模型判斷不好，所以我們會用 cleanvison 去檢測圖片潛在的問題，之後再去做相應的處理。

常見問題:

1. Low information:缺乏明顯可辨別的特徵
2. Grayscale:影像是灰階圖，沒有色彩資訊
3. Odd size:影像大小或解析度與其他照片不同
4. Light:影像可能過於明亮
5. Near Duplicates:影像相似度過高
6. Dark:影像過於暗
7. Blurry:影像過於模糊

訓練集處理 1:

首先先對完整訓練集(4043 張照片)利用 cleanvision 來檢測圖片問題(圖 9)，可以看到幾乎所有的資料都有 low information 的問題，所以我們不會優先處理這個問題，而 Grayscale 的問題我們覺得不會影響到模型的判別，所以我們也不會去處理，Odd size 的部分因為最終會調整至統一大小，我們也會略過，所以最終會處理的部分只有 Light 與重複的影像。

	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

圖 9

訓練集處理 2:

- 相似影像:總共 82 張，所以會刪除一組中的一張(41 張)
- 過量影像:加入高斯雜訊(平均 0，標準差 0.05)，亮度調為原本的 0.7，然後再把原始照片刪除

那我們把處理過後的資料重新利用 cleanvision 檢測問題(圖 10)，那可以發現過亮影像的問題解決了，那相似影像是因為我們處理過亮影像

時會增加新的圖片，那我們去查看發現圖片還是有一些不同，所以會保留下來(圖 11)，目前會剩下 4002 張訓練集

	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

圖 10

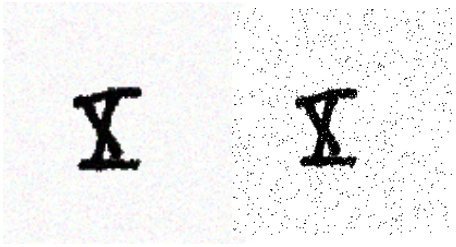

圖 11

驗證集處理 1:

一樣先對完整驗證集(960 張)，利用 cleanvision 檢測圖片問題(圖 12)，那可以發現與訊連的問題一樣，所以我們處理方式與處理訓練集相同。

	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

圖 12

驗證集處理 2:

處理方式與訓練集的方法一樣，以下是處理過後的結果(圖 13)，最終會剩下 949 張圖片。

	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

圖 13

4. 進行資料擴增:

目的:

由於模型是固定的，要讓模型能預測的好就必須要讓模型能學習到更多種圖片，那我們做法是會基於清理過後的圖片來增強，讓模型能泛化的更好。

增強方式:

包括隨機旋轉、隨機縮放、隨機平移和隨機對比度，和隨機對比度，其中對比度為圖像中最亮與最暗區域之間的差異。那以羅馬數字來說，會不太適合過度的旋轉，因為數字會有對稱問題，那其他部分就會正常調整。

訓練集增強:

隨機旋轉:範圍為±10%(36 度)

隨機縮放:範圍為±20%

隨機平移:範圍為±15%

隨機對比度:範圍為 0.1~0.2

增強效果如圖 14，這個過程我們有設定至少每張照片都會被增強 1 次，最終增加了 6501 張圖片，目前訓練集數量總共為 10503。

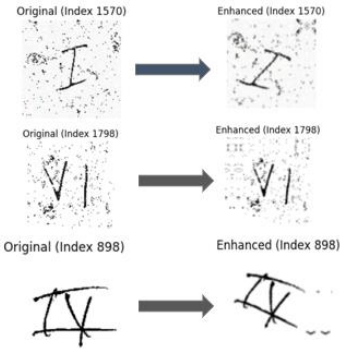

圖 14

驗證集增強:

隨機旋轉:範圍為±5%(18 度)

隨機縮放:範圍為±10%

隨機平移:範圍為±10%

隨機對比度:範圍為 0.05~0.1

範圍調整的比訓練集小，因為我們覺得驗證集不是真的拿來訓練的，而是讓模型泛化的更好，所以不讓他們變化的太大，最終總共增加 2316 張圖片，目前數量為 3265 張。雖然正常情況下訓練集跟驗證集的處理方式要一樣，但我們後面測試幾次還是這樣的參數調整在測試集的準確率狀況是最好的。

5. 再進行一次資料清理：

處理方式:

對增強過後的訓練集及驗證集一樣利用 cleanvision 檢測圖片問題(圖 15、16)，處理方式跟 3.一樣，但因為我們想把訓練集減少至 9000 張；驗證集減少至 2940 張，後續會去針對 Low information 的部分做處理。

	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

圖 15(訓練集)

	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

圖 16(驗證集)

處理 Low information:

我們原本的想法是刪除 information score 較低的圖片，因為越低代表圖片問題是越嚴重的，但後續的結果都不是很好，所以我們有將分數較高跟較低的圖片來做比較(圖 17、18)，分數低的都是原始的照片，而且都是很容易可以看出是什麼數字的圖片；分數較高的都是經過增強過的，那我們覺得還是需要保留一些比較好辨識的圖片，而且保留較多原始照片也比較合理，所以最後會從分數高的開始刪除，來達到我們照片數量的需求。
最終訓練集為 9000 張，驗證集為 2940 張，而在 500 張測試集的準確度可以到 0.8 以上

圖 17(分數低)

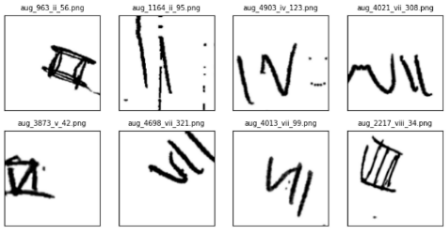

圖 18(分數高)

IV. 結論

1.

錯誤的標籤會使得模型效果不好，因為不管是資料清理或是增強，只要錯誤的圖片還在，不管做什麼清理跟增強，錯誤的圖片還是會影響到模

型的辨別。

2.

數據清理與增強是影像處理很重要的一環，清理資料可以去除異常的資料；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，尤其是在模型的能力有限的情況下。

3.

這次處理錯誤標籤的方式是利用手動處理的，如果之後還有類似的問題我們可以嘗試建立一個模型來去自動化的方式來處理

V. 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/

cleanvision:

https://www.cnblogs.com/luohenyueji/p/18499084

老師講義:

VI. 分工情形

吳俞憲(50%):錯誤標籤處理、數據增強參數調整、ppt 製作、書面報告後半部分

廖廣筑(50%):整體流程安排、主要程式撰寫、上台報告、書面報告前半部分