繁體中文場景文字辨識競賽－

初階：場景文字檢測

報告說明文件

壹、 環境

作業系統：Ubuntu 18.04
語言：Python 3.6
環境：CUDA 10.0
套件：pytorch==1.0.1, torchvision=0.2, ninja yacs cython matplotlib tqdm scipy shapely
預訓練模型：Box_Discretization_Network ReCTS 2019 model
https://github.com/Yuliang-Liu/Box_Discretization_Network
額外資料集：沒有使用，但使用的預訓練模型有訓練在其他場景文字檢測的資料集上

貳、 演算方法與模型架構

我們使用「Omnidirectional Scene Text Detection with Sequential-free Box Discretization」這篇論文的演算法、預訓練模型和開源程式碼。
我們基於開源的 ReCTS 2019 模型的權重與設定，除調低學習率到 0.001 外，並未做任何修改，直接在本次比賽的資料上做微調。
後處理的部份，我們只選用信心值大於 0.91 的預測匡當作上傳結果。

參、 資料處理

我們使用 9 成的資料做訓練，1 成的資料做驗證。
並未做任何刪減或增補。

肆、 訓練方式

我們使用「Omnidirectional Scene Text Detection with Sequential-free Box Discretization」這篇論文的開源程式碼做訓練，總共訓練 20000 個 iterations。
最後取第 12000、13000、14000、15000、16000、17000、18000、19000、20000 個 iter 的權重，總共 9 個權重做平均當作最後預測用的權重。

伍、 分析與結論

開發過程：

比賽一開始，我們首先嘗試的是使用物件檢測演算法，如 YOLOv5 與 EfficientDet 來做場景文字檢測，雖然預測的表現不差，但因為預測出來的都是長方形，和本次競賽的四邊形標注有所落差，因此沒辦法達到很高的 Hmean score。

之後我們嘗試轉向使用文本檢測的演算法，如 CTPN 、 DB 、 EAST、 PANet、 與 TextBoxes++。在實驗的過程中，我們發現如果從在文字檢測資料預訓練過的模型做微調，效果會比重頭開始訓練一個新模型效果好上不少。但網路上的開源模型大多數都是訓練在英文的資料上，有支援中文檢測的權重是相當稀少。

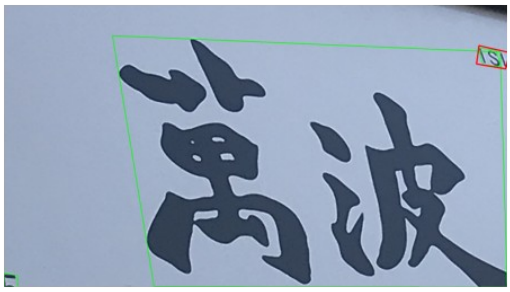
因此最後我們選擇了「Omnidirectional Scene Text Detection with Sequential-free Box Discretization」這篇論文的模型使用，他是我們找到效果最好的開源中文檢測預訓練模型了。

使用他們的 ReCTS 2019 預訓練模型，不做微調，就能得到高達 0.603 的 Hmean score，在比賽資料集上訓練後，更是能達到 0.695 的 Hmean score。

結果分析：

經過分析，我們發現模型較常犯的錯誤大概有幾類（紅色是預測的框框，綠色則是正確答案）：

1.特大文字或特長文字



可以發現模型對這種文字容易會有漏檢的狀況發生，這或許可以透過新增特長、特寬與特大的 anchor，或是增加特徵金字塔的層數來解決。

2.直橫交錯的文字

這種情況模型可能會分不清究竟要直的讀還是橫的讀，這或許可以透過增加更多直橫交錯的訓練資料來訓練，或是搭配文字識別模型來選出合理的讀法。

3.小又不清楚的文字

小又不清楚的文字也經常會發生漏檢的情形，但這種文字就算是人眼也不容易識別，目前沒有想到比較好的解決方法。好像有些文字檢測比賽會將這些文字排除在分數計算之外的樣子。

陸、 程式碼

https://drive.google.com/file/d/16jjf0JMF1A8qw4t5OoJisyT_hDc9uNzh/view?usp=sharing

柒、 使用的外部資源與參考文獻

1. Liu, Y., Zhang, S., Jin, L., Xie, L., Wu, Y., & Wang, Z. (2019). Omnidirectional scene text detection with sequential-free box discretization. arXiv preprint arXiv:1906.02371.
2. Zhang, R., Zhou, Y., Jiang, Q., Song, Q., Li, N., Zhou, K., ... & Jawahar, C. V. (2019, September). Icdar 2019 robust reading challenge on reading chinese text on signboard. In 2019 International Conference on Document

Analysis and Recognition (ICDAR) (pp. 1577-1581). IEEE.
3. Tian, Z., Huang, W., He, T., He, P., & Qiao, Y. (2016, October). Detecting text in natural image with connectionist text proposal network. In European conference on computer vision (pp. 56-72). Springer, Cham.
4. Liao, M., Wan, Z., Yao, C., Chen, K., & Bai, X. (2020, April). Real-time scene text detection with differentiable binarization. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 34, No. 07, pp. 11474-11481).
5. Zhou, X., Yao, C., Wen, H., Wang, Y., Zhou, S., He, W., & Liang, J. (2017). East: an efficient and accurate scene text detector. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (pp. 5551-5560).
6. Wang, W., Xie, E., Song, X., Zang, Y., Wang, W., Lu, T., ... & Shen, C. (2019). Efficient and accurate arbitrary-shaped text detection with pixel aggregation network. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 8440-8449).
7. Wang, W., Xie, E., Li, X., Hou, W., Lu, T., Yu, G., & Shao, S. (2019). Shape robust text detection with progressive scale expansion network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 9336-9345).
8. Liao, M., Shi, B., & Bai, X. (2018). Textboxes++: A single-shot oriented scene text detector. IEEE transactions on image processing, 27(8), 3676-3690.
9. Glenn Jocher, Alex Stoken, Jirka Borovec, NanoCode012, Ayush Chaurasia, TaoXie, … Francisco Ingham. (2021, April 11). ultralytics/yolov5: v5.0 - YOLOv5-P6 1280 models, AWS, Supervise.ly and YouTube integrations (Version v5.0). Zenodo. http://doi.org/10.5281/zenodo.4679653
10. Tan, M., Pang, R., & Le, Q. V. (2020). Efficientdet: Scalable and efficient object detection. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 10781-10790).

聯絡資料

- **隊伍**

隊伍名稱	Private leaderboard 成績	Private leaderboard 名次
三杯波波隊	0.707458	1

- **隊員(隊長請填第一位)**

姓名(中英皆需填寫)	學校系所	電話	E-mail
陳奕嘉 (Yi-Chia Chen)	台灣科技大學 資訊工程研究所	0963359483	asdzxcasdzxcus@gmail.com

- **指導教授**

若為「連結課程」的課堂作業或期末專題，請填授課教師，以利依連結課程彙整

若非「連結課程」，但有教授實際參與指導，請填寫該位教授。

若以上兩者皆非，可不予填寫。

教授姓名	課程	課號	學校系所	E-mail
陳冠宇			台灣科技大學 資訊工程研究所	kychen@mail.ntust.edu.tw

附錄 - 審查委員評論回覆

審查意見：

1、此隊伍選用合適方法的前期準備非常充足。分析了大量 open source 方法，包含兩種 object detection 方法與五種 text detection 方法，並明確指出頂尖 object detection 方法不適用於本次比賽的原因。最終選用的方法提供強大的中文檢測預訓練模型，此預訓練模型在比賽的排行榜即能排到第十六名。訓練過程無使用額外資料集，利用良好的中文預訓練模型進行 fine-tune 得到了不錯的結果。

2、僅是以這篇 Omnidirectional Scene Text Detection with Sequential-free Box Discretization (IJCAI 2019)論文 open source 的 pre-trained model 來進行 finetuning，沒有做任何的修改。即使在結果分析上有討論到一些可改進的地方 (如特大文字或特長文字等可以應該可以借助新增 anchor box 或是增加特徵金字塔的層數來解決)，但是也沒有付諸行動。對於使用的模型為什麼會有相關的限制也沒有任何討論。此外，對於為何要取不同 iteration 時所訓練出來的模型進行 ensemble 也沒有給予說明。整體來說雖然 performance 不錯，但是由於上述原因所造成的缺漏，十分可惜。

問題回覆：

Q1: 為何沒有將討論到可改進的地方付諸行動？

A1: 受限於我們使用的模型必須在 VRAM 32GB 以上的 GPU 才能執行，我們在雲端租用專業級 GPU 才成功完成訓練，光是原始的模型花掉的租金就已經超過萬元，實在沒有多餘的財力可以做更多的嘗試。

Q2: 為何要取不同 iteration 時所訓練出來的模型進行 ensemble？

A2: 在訓練中取不同 iterations 的權重做平均是一種常用的 deep learning 優化方法，也有許多相關的研究，例如：arXiv:1803.05407。