複雜街景之文字定位與辨識

110753132 馬行遠, 110753128 李韋杰, 110753124 宋浩茹
政治大學資訊科學所

Abstract

我們在路上抬頭可見的招牌、路牌、看板、標語與廣告等,包含了大量的文字,傳遞豐富的訊息。這些提供影像拍攝地點相關資訊的街景文字,若能被自動地由畫面中定位並加以辨識,相信對於包括場景理解、智慧城市、智慧交通的發展與建置、機器人技術、自動駕駛、協助視障者或外來旅者等各類應用都有所幫助。能夠實作出非常精確的辨識出影像中的文字內容模型,對台灣街景的文字偵測與辨識技術有所貢獻,作為我們專案的主要目標。

Introduction

場景文字辨識近幾年受到了大量的關注,但仍有許多值得挑戰的部分,從任意拍攝的複雜場景中精準地辨識,包含中文、英文與數字內容是值得研究的主題。在過去場景文字辨識的準確率,容易受到街景中常見的多型態文字、多國文字、傾斜文字、不同尺寸文字、紋理干擾、光線與陰影等干擾,而為了解決以上挑戰,利用機器學習、深度學習的技術以及加上具有創意的想法,透過我們訓練出的模型,偵測畫面中的文字位置,並且克服多型態中文字、傾斜招牌中文字、不同尺寸中文字與空字串框等情形,最後辨識出偵測文字位置中的字串,包含中文、英文及數字內容。我們將實驗分成3大部分:1.資料前處理(資料擴增、二值化影像等)2.文字區域定位模型(文字框之位置)3.文字辨識模型(中文、英文與數字內容)。

Related work

本研究參考了 Wang, C. Y., & Liao, H. Y. M.於2020年發表的YOLOV4模型[1]與Y Fang, X Guo, K Chen, Z Zhou, Q Ye 於 2021年發表的yolov5模型發現到YOLO在物體偵測上非常實用準確性又高，但只能切出矩型框的辨識框因此我們認為這不太符合我們的需求。當時聽說 Y. Liu, T. He,等人發表的orderless box discretization network[2]效果非常不錯，值得一試。這時又參考了Minghui Liao, Zhaoyi Wan 等人於2019發表的Differentiable Binarization(DBN) 模型[3]，發現到他在論文中將Differentiable Binarization(DBN)放入MSRA-TD500、CTW1500資料集 的效果都較其他模型要好上不少如表格 1，切還可以切出非常準確的四邊形框，因此後來我們採用此模型來協助我們偵測文字。

表格 1 DBN與其他模型比較表

Method	P	R	F	FPS
CTPN (Tian et al. 2016)	74.2	51.6	60.9	7.1
EAST (Zhou et al. 2017)	83.6	73.5	78.2	13.2
SSTD (He et al. 2017a)	80.2	73.9	76.9	7.7
WordSup (Hu et al. 2017)	79.3	77	78.2	-
Corner (Lyu et al. 2018b)	**94.1**	70.7	80.7	3.6
TB (Liao, Shi, and Bai 2018)	87.2	76.7	81.7	11.6
RRD (Liao et al. 2018)	85.6	79	82.2	6.5
MCN (Liu et al. 2018)	72	80	76	-
TextSnake (Long et al. 2018)	84.9	80.4	82.6	1.1
PSE-1s (Wang et al. 2019a)	86.9	84.5	85.7	1.6
SPCNet (Xie et al. 2019a)	88.7	**85.8**	87.2	-
LOMO (Zhang et al. 2019)	91.3	83.5	87.2	-
CRAFT (Baek et al. 2019)	89.8	84.3	86.9	-
SAE(720) (Tian et al. 2019)	85.1	84.5	84.8	3
SAE(990) (Tian et al. 2019)	88.3	85.0	86.6	-
DB-ResNet-18 (736)	86.8	78.4	82.3	**48**
DB-ResNet-50 (736)	88.2	82.7	85.4	26
DB-ResNet-50 (1152)	91.8	83.2	**87.3**	12

（資料來源: Real-time scene text detection with differentiable binarization.[4]）

資料擴增的部分，本研究參考了比賽單位提供的中文圖片生成器的資料擴增方式和一些用GAN加強資料擴增的方式最終使用彩色的文字圖片產生器，可以快速產生平面文字，加上一些噪點、位移、切割來產生不同情下的訓練效果。但對於璇傳和3D旋轉的效果就需要再加強了。

Method Description

首先我們將我們的方法分成主要的兩大部分，第一部分是訓練資料的前處理與資料擴增，第二部分則是模型架構與訓練方式。

偵測模型的部分因無法取得大量標記好的實際街景資料，因此本研究取得百度提供的開放資料集LSVT來當作中文的擴增資料集。

辨識模型的資料擴增訓練資料的前處理與資料擴增，主要運用調整角度、大小與解析度等方式獲得更多資料來進行訓練，或是產生與真實資料極為相近的資料集並增加一些噪點和干擾確保產生各種狀況，因為轉角度與二值化已由DBN模型完成，資料擴增的部分就以產生假資料為主。我們運用github上zcswdt開發的Color_OCR_image_generator專案來進行資料擴增，此專案能字典檔文本產生直式與橫式的中英文字串，並能加入自定義的噪點、晃動和位移，這使得產生的字串圖片更接近切出來的效果，顏色的選用則是依照iiit5資料集訓練出來的模型來決定字的顏色，使得最終結果更接近切圖出來的效果。我們也對該專案進行一些調整，像是將橫向圖片轉成直向，加入更為複雜的背景圖營造遮擋文字效果，也解決了文字與背景顏色有時太過相近的問題與有時會有空白圖片產出的問題。

文本來源則是透過google maps取得以及行政院開放資料，透過google maps的place api，取得"若水國際", "中央大學", "台中車站", '高雄車站'四個地點100km內的前100有名的車站，再將所有車站與地點方圓10km內的"材料行", "餐廳","咖啡廳", "寵物店", '飲料店', "維修廠", "電器行", "小吃店", "早餐店", "學校", "企業"各抓出100家並儲存所有商家名。最後因為使用限制共抓回來24000筆數據，並用這筆資料隨機生成30萬張各式圖片進行pre-train。而會選擇這四個地點的主要考量是因為主辦方是中央大學而資料處理公司是若水國際，而另外兩個就是選中部和南部的大型車站來擴大抓取範圍。行政院開放資料中有所有跟政府註冊的公司名稱共150萬筆但裡面有些問題需要處理:1.裡面有許多同名分公司2.大量的有限公司與有限股份公司會過於偏頗3.含有大量生僻字為此我們進行了以下處理:我們只留下資本額前50萬多的公司以確保去除大量的同名分公司與空頭公司，同時也確保抓出來的公司有一定知名度。篩選出此資料集前5000常用字，以此將生僻字直接去掉以減少類別來換取更高的準確度，再將5000字語比賽訓練資料集出現的字合併起來，最後此資料產生了60萬筆訓練資料。如**圖 1 產生的擴增資料**

圖 1 產生的擴增資料

第二個部分為模型架構與訓練方式，模型架構為Two-stage，第一個階段為文字區域定位，第二個階段為文字內容辨識，先找出文字（Region Proposals），再辨識文字內容（Object Recognition），以下將詳細說明模型的演算法及架構。

1. 文字區域定位的模型，主要是使用Differentiable Binarization(DBN) [1]的架構如**圖 2**，Backbone為ResNet34，Optimizer是使用Adam，以L2作為Regularizer，Loss使用DBLoss，訓練方式則使用中國街景資料集LSVT與本次training data的70%作預訓練，之後純用training data 7:3 fine tune。

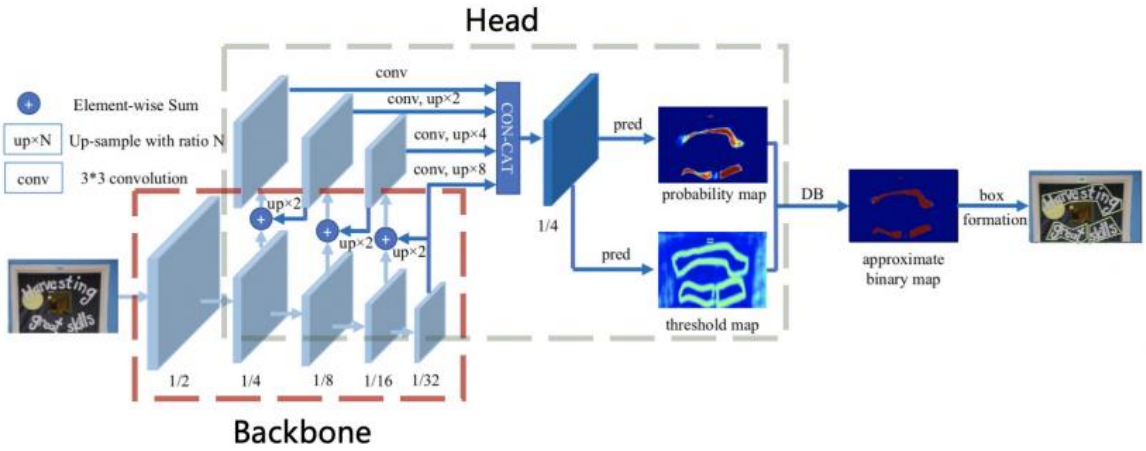

圖 2 BDN模型示意圖

2. 文字內容辨識的模型，使用CRNN[5]的架構如**圖 3**，Backbone為ResNet50，Optimizer是使用Adam，以L2作為Regularizer，Loss使用DBLoss。CRNN是將CNN、LSTM與CTC三種方法結合，首先CNN提取圖像特徵，再透過LSTM進一步提取圖像特徵中的序列特徵，最後引入CTC解決訓練時，字元無法對齊的問題，而我們的訓練方式是使用生成資料(80萬筆)預訓練，再使用training data 7:3做fine tune如**圖 4**所示。

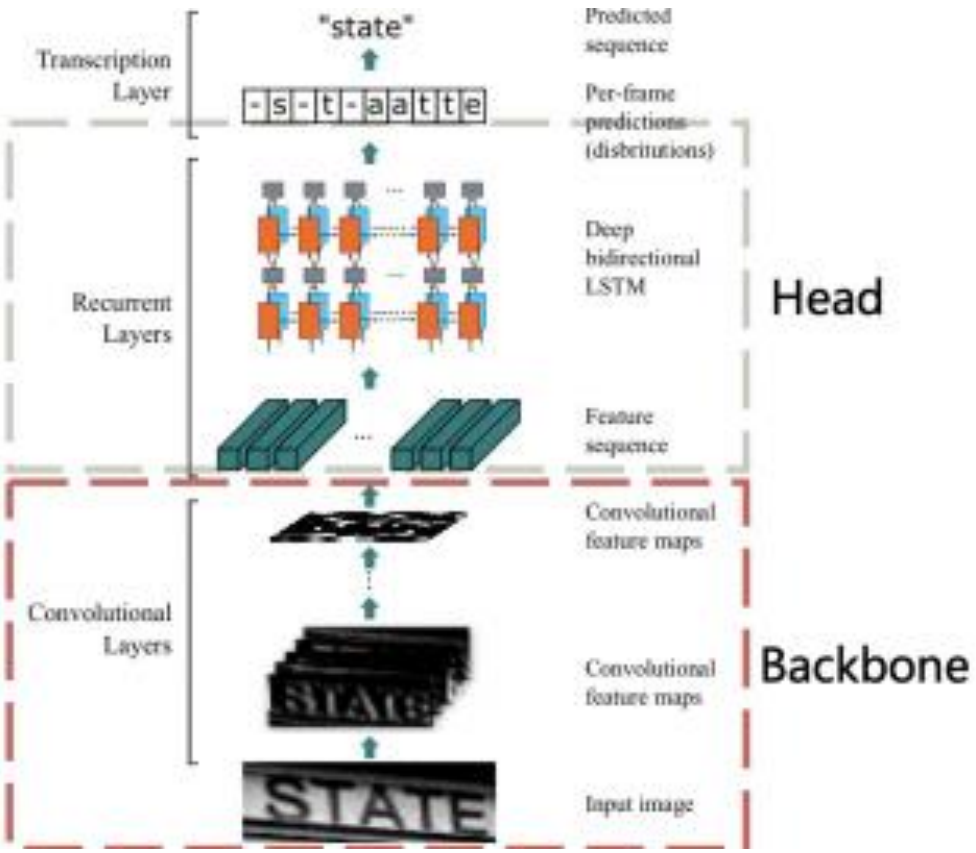

圖 3 RCNN示意圖

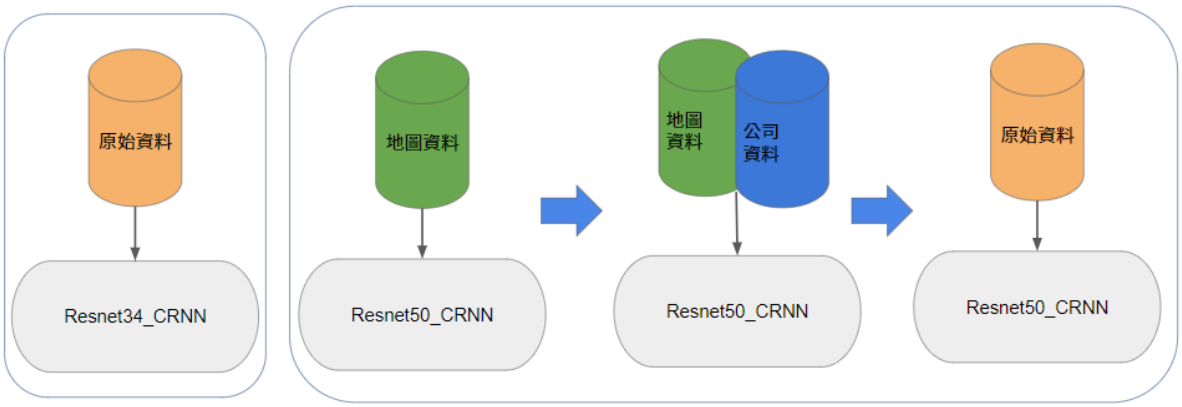

圖 4 訓練方式

Experimental Results

此次比賽private 的成績 DET:0.774(P 0.9461*R 0.8180) Rec 0.804，可知偵測模型表現比辨識差，可以調整偵測框的threshhold進一步調整precision and recall成績，找到較好的F1 Score，而達到最佳化，或者使用最小框來偵測字元，若使用中文字元即可解決字串方向辨識問題，以及降低辨識模型的難度，也可降低box裡面包含雜訊的問題。此研究也有許多改進的機會，像是資料擴增的noise調整，減少文字辨識的訓練的複雜度，以及改善辨識字典過多的問題，處理同型字、刪除標點符號等。

References

[1] Wang, C. Y., & Liao, H. Y. M. (2020). Yolov4: Optimal speed and accuracy ofobject detection. arXiv preprint arXiv:2004.10934.
[2] Y. Liu, T. He, H. Chen, X. Wang, C. Luo, S. Zhang, C. Shen, and L. Jin, "Exploring the capacity of an orderless box discretization network for multi-orientation scene text detection," arXiv:1912.09629, 2020.
[3] Y. Baek, B. Lee, D. Han, S. Yun, and H. Lee, "Character region awareness for text detection," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2019.
[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]B. Shi, X. Bai, and C. Yao, "An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition," IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016.

Code

GitHub : https://github.com/markpony100/AICUP_OCR