Yonsei University, Computer Science Department / Byun, JungHyun
Text-Detection-Recognition Task Technical Report

1. Introduction

In this report, I discuss the approach I took to tackle the problem of the text detection and recognition task on images. Specifically, we focus on ICDAR2017 Robust Reading Challenge [1] on the COCO-Text dataset [2]. The COCO-Text dataset is based on the Microsoft COCO dataset [3], which contains images of complex everyday scenes, as opposed to datasets which are synthetically generated. The images were not collected with text in mind and thus contain a broad variety of text instances, which can be referred to as 'incidental text', meaning text appears in the scene without the user having taken any specific prior action to cause its appearance or improve its positioning/ quality in the frame [4].

The characteristic of COCO-Text dataset makes Robust Reading Challenge literally a challenge. The authors of the COCO-Text dataset collected anonymized state-of-the-art photo OCR detection, transcription and end-to-end results on COCO-Text from their collaborators at Google, TextSpotter and VGG [2]. Their results are shown in the below table and as one can easily notice, even the state-of-the-art methods struggle but fail to achieve reasonably-well-enough-to-apply-to-real-world-problem accuracies. we will examine in detail on the experiment and the results in in Section 4.

Alg	Localization							Recognition	End-to-end		
	recall					precision	f-score	accuracy	recall	precision	f-score
	legible		illegible		total	total	total				
	machine	hand	machine	hand							
A	34.99	21.26	6.81	4.41	28.8	79.82	42.31	83.76	29.35	65.74	40.58
B	21.25	18.27	2.30	1.37	17.6	76.36	28.63	60.71	12.97	45.75	20.21
C	8.97	7.48	0.47	0.84	6.5	21.76	10.00	21.96	1.98	4.57	2.76

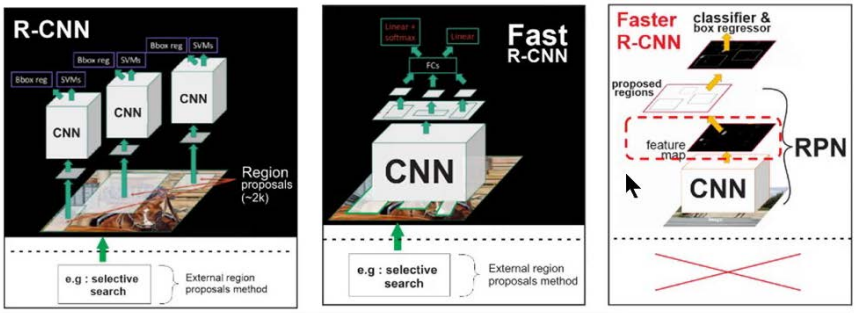
2. Approach

The challenge we are trying to tackle consists of three tasks:

1. Text Localization, where the objective is to obtain a rough estimation of the text areas in the image, in terms of bounding boxes that correspond to words.
2. Cropped Word Recognition, where the locations (bounding boxes) of words in the image are assumed to be known and the corresponding text transcriptions are sought.
3. End-to-End Recognition, where the objective is to localize and recognize all words in the image in a single step.

Task 3 End-to-End Recognition can be thought of as the collaboration, or joint work of the rest two tasks, Text Localization and Cropped Word Recognition. Thus it would be practically divide the End-to-End Recognition task into two sub-tasks, and feed the output of the Task 1 as the input of the Task 2. This concatenation of two tasks, localization and recognition is very similar to the idea of the region proposal method.

Some of the most popular researches to adopt region proposal method are R-CNN [5], Fast R-CNN [6] and Faster R-CNN [7], or namely R-CNN family. R-CNN and Fast R-CNN both utilize external region proposal algorithm, but Fast R-CNN accelerates the task by applying regional proposal to the feature domain not the image domain. Faster R-CNN accelerates even more by not having external regional proposals but a Regional Proposal Network. Their comparative results are organized in the table below [5]. The speed-up for the test time per image is remarkable. However, one can also notice Mean Average Precision (mAP) results are about the same. Also, it is known that the alternative joint training of the Regional Proposal Network and Convolutional Neural Network is very complex. Based on these grounds, since we only care about the performance of the algorithm not the task completion time, I designed the solution for the End-to-End Recognition task similar to the structure of the original R-CNN, which will be described in Section 3.



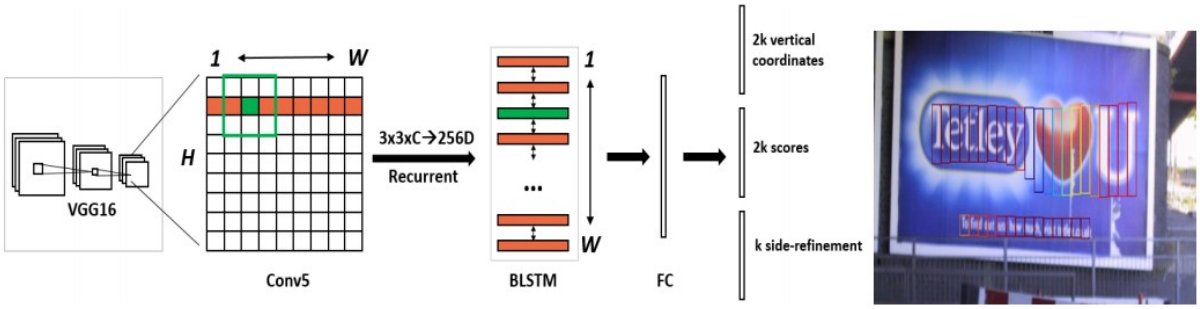
	R-CNN	Fast R-CNN	Faster R-CNN
Test time per image	50 seconds	2 seconds	0.2 seconds
Speed-up	1x	25x	250x
mAP (VOC 2007)	66.0%	66.9%	66.9%

3. Implementation

As mentioned in Section 2, the proposed solution consists of two separate steps. I used CTPN (Connectionist Text Proposal Network) [9] for the text detection task, and CRNN (Convolutional Recurrent Neural Network) [10] for the text recognition task. Two networks are briefly introduced here in the following descriptions, based on the original papers. Readers are referred to the original papers for the more detailed explanation.

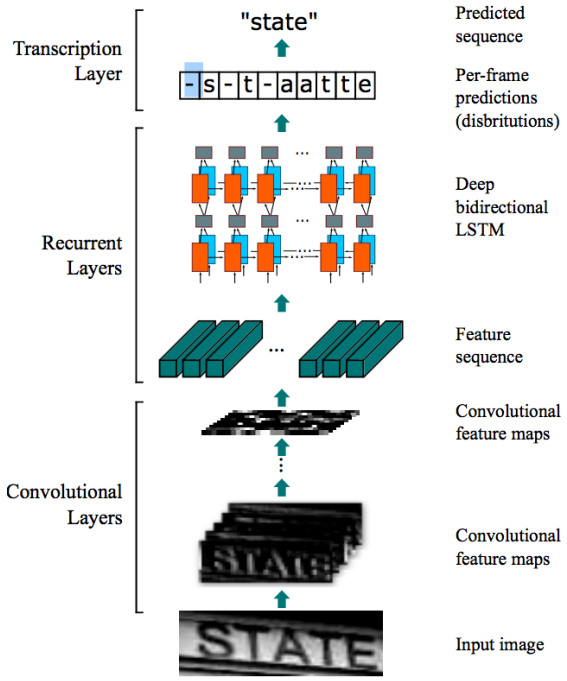
One of many difficulties of text detection of is dramatically changing length of the text line. To overcome this hindering, authors of CTPN proposed the key idea of detecting a small, fixed-width text segment, post-processing segments and then connecting these small text segments to get the final text line. Specifically, the authors proposed a vertical anchor method to predict the vertical position of the text. At the same time, horizontal text lines are detected, and each text segment is linked. The used network is a structure of CNN+RNN, which makes the detection result more robust.

First, VGG16 is used to extract features from a image. Then, a sliding window is applied to the feature map. The eigenvector is used to predict the offset distance from the 10 anchors, which means that at each window center 10 text proposals will be predicted. The features obtained in the previous step are input into a Bidirectional LSTM. The BLSTM is connected to a fully connected layer to produce three types of outputs, vertical coordinates of the anchors, scores for the text proposals and refinements to determine endpoints of a text line. After the standard non-maximal suppression step to eliminate excessive predictions, a graph-based text line construction algorithm is used to merge the obtained text segments into one text line.



The main contribution of CRNN paper is a novel neural network model, whose network architecture is specifically designed for recognizing sequence-like objects in images. The proposed neural network model of the paper is named as Convolutional Recurrent Neural Network, since it is a combination of DCNN and RNN.

The network architecture of CRNN consists of three components, including the convolutional layers, the recurrent layers, and a transcription layer, from bottom to top. At the bottom of CRNN, the convolutional layers automatically extract a feature sequence from each input image. On top of the convolutional network, a recurrent network is built for making prediction for each frame of the feature sequence, outputted by the convolutional layers. The transcription layer at the top of CRNN is adopted to translate the per-frame predictions by the recurrent layers into a label sequence. Though CRNN is composed of different kinds of network architectures (DCNN and RNN), it can be jointly trained with one loss function, which makes the network end-to-end trainable. The overall architecture and the flow of the process is visualized in the picture below.



4. Experiment

The COCO-Text dataset consists of 63,686 images and 145,859 text instances. Among 63,686 images, only 10,000 validation images were used in the experiment since the other 10,000 test images were provided without annotation data. Note that in contrast the original COCO-Text dataset paper [2] used all the 63,686 images to test three state-of-the-art methods. Therefore, direct comparisons between the proposed and three state-of-the-art baselines is not appropriate, and the comparison results should be used for information purposes only.

Among all text instances, legible-machine-printed accounts for 61% of them, legible-hand-written accounts for 3%, illegible machine-printed accounts for 34% and illegible-hand-written accounts for 2%. This composition of categories makes COCO-Text dataset quite unbalanced, which contributes to relatively low total recall result in localization task of the proposed method, even though the proposed method is comparable to or even excel state-of-the-art methods in some sub-categories.

Method	localization							recognition	end-to-end		
	recall					precision	f-score				
	legible		illegible		total	total	total	accuracy	recall	precision	f-score
	machine	hand	machine	hand							
A	34.99	21.26	6.81	4.41	28.8	79.82	42.31	83.76	29.35	65.74	40.58
B	21.25	18.27	2.3	1.37	17.6	76.36	28.63	60.71	12.97	45.75	20.21
C	8.97	7.48	0.47	0.84	6.5	21.76	10	21.96	1.98	4.57	2.76
Proposed	19.28	20.3	7.86	8.89	15.7	19.17	17.28	23.6	4.53	3.57	3.99

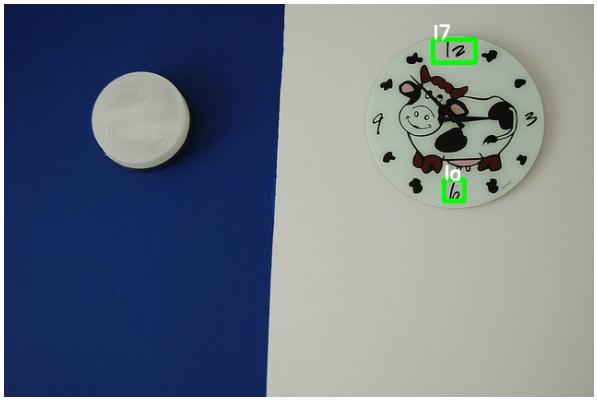
Overall performance of the recognition and end-to-end task of the proposed approach is not quite satisfactory, though it demonstrates a slight outperformance over one of the three state-of-the-arts. Specifically, high ratio of false negatives and false positives to true positives yielded low recall and low precision at the same time. Also, text recognition results are sub-par compared to comparable-to-human accuracy of a certain state-of-the-art.

I conjecture the reason for the low performance of the proposed approach to be the conditions of the COCO-Text dataset. Since the networks used were trained on the dataset as guided in the respective original papers, the harsh conditions of the COCO-Text dataset, in terms of poor image qualities, super low resolutions, distortions and such, might have been unaccustomed to the trained network models

.

Below are some mixed results of the proposed approach provided for qualitative evaluations.

COCO-Text validation images:



More photos in the wild:

5. Conclusion

In this report, an approach to tackle the problem of the text detection and recognition task on images was introduced. The text detection and recognition task was partitioned into two separate text detection and recognition tasks, similar to the region proposal method prevalent in the object detection field. For the text detection Connectionist Text Proposal Network [9] was employed, and for the text recognition, Convolutional Recurrent Neural Network [10] was utilized. Although the proposed approach did not yield prominent results on COCO Text dataset, at least it showed overall comparable performance and outperformance in some sub-categories. This leaves the proposed approach some room for improvement, if the networks are carefully finetuned on more challenging datasets.

6. Reference

[1] Gomez, Raul, et al. "ICDAR2017 robust reading challenge on COCO-Text." 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). IEEE, 2017.

[2] A. Veit, T. Matera, L. Neumann, J. Matas, S. Belongie. COCO-Text: Dataset and Benchmark for Text Detection and Recognition in Natural Images. arXiv preprint arXiv:1601.07140, 2016.

[3] Lin, Tsung-Yi, et al. "Microsoft coco: Common objects in context." European conference on computer vision. Springer, Cham, 2014.

[4] D. Karatzas, L. Gomez-Bigorda, A. Nicolaou, D. Ghosh , A. Bagdanov, M. Iwamura, J. Matas, L. Neumann, VR. Chandrasekhar, A. Lu, F. Shafait, S. Uchida, E. Valveny.: ICDAR 2015 robust reading competition. 13th International Conference on Document Analysis and Recognition (ICDAR).

[5] Girshick, Ross, et al. "Rich feature hierarchies for accurate object detection and semantic segmentation." Proceedings of the IEEE conference on computer vision and pattern recognition. 2014.

[6] Girshick, Ross. "Fast r-cnn." arXiv preprint arXiv:1504.08083 (2015).

[7] Ren, Shaoqing, et al. "Faster r-cnn: Towards real-time object detection with region proposal networks." Advances in neural information processing systems. 2015.

[8] http://cs231n.stanford.edu/slides/2016/winter1516_lecture8.pdf

[9] Tian, Zhi, et al. "Detecting text in natural image with connectionist text proposal network." European Conference on Computer Vision. Springer, Cham, 2016.

[10] Shi, Baoguang, Xiang Bai, and Cong 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 39.11 (2017): 2298-2304.