Single Shot Text Detector with Regional Attention

Pan He[1], Weilin Huang[2], Tong He[3], Qile Zhu[1], Yu Qiao[3] and Andy Li[1]
[1] National Science Foundation Center for Big Learning, University of Florida
[2] Department of Engineering Science, University of Oxford
[3] Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences

International Conference on Computer Vision 2017

Introduction

➢ Goal:

- Improve the speed and accuracy for scene text detection.

➢ Existing works:

- Pixel-based detectors [1]: Using cascaded Fully Convolutional Networks (FCN) to cast character-based detections into the problem of pixel-wise text semantic segmentation.
- Box-based detectors [2]: Extending object detectors such as Faster R-CNN [3] or SSD [4] to predict text boxes by simply using bounding-box annotation.

➢ Problem & Motivation:

- In spite of effectively identifying rough text regions, pixel-based text detectors fail to produce accurate word-level predictions with a single model. The main challenge is to precisely identify individual words from a detected rough region of text.
- Box-based text detectors are often trained by simply using bounding-box annotations, which may be too coarse (high-level) to provide a direct and detailed supervision, compared to the pixel-based detectors where a text mask is provided.

➢ Our idea:

- We proposed techniques to bridge the gap between the pixel-based detectors and the box-based detectors, resulting in a single-shot model that essentially works in a coarse-to-fine manner.

Reference

[1] Z. Zhang, C. Zhang, W. Shen, C. Yao, W. Liu, and X. Bai. Multi-oriented text detection with fully convolutional networks, **CVPR**, 2016.

[2] Z. Tian, W. Huang, T. He, P. He, and Y. Qiao. Detecting text in natural image with connectionist text proposal network, **ECCV**, 2016.

[3] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: To- wards real-time object detection with region proposal net- works, **NIPS**, 2015.

[4] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, and A. C. Berg. SSD: Single shot multibox detector, **ECCV**, 2016.

[5] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions, **CVPR**, 2015.

[6] P. He, W. Huang, Y. Qiao, C. C. Loy, and X. Tang. Reading scene text in deep convolutional sequences, ***AAAI***, 2016.

Framework of SSTD with Regional Attention

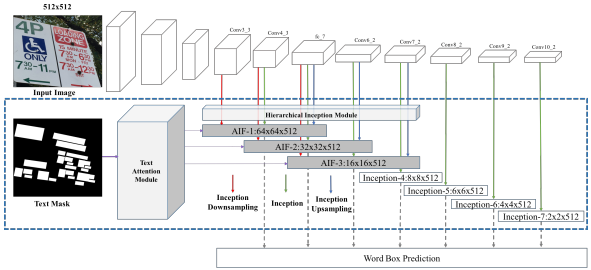

Figure 1: Framework of SSTD with Regional Attention

Text Attention Module

➢ Idea:

- A top-down spatial attention on text regions to suppress the background interference and cast the cascaded FCNs detectors into a single model.

➢ Attention Map:

- We compute a text attention map from Aggregated Inception Features (AIFs).
- The attention map indicates rough text regions and is further encoded into the AIFs (via element-wise dot production).
- The attention module is trained by using a pixel-wise binary mask of text.

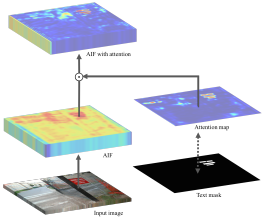

Figure 2: Architecture of Text Attention Module

Hierarchical Inception Module

➢ Idea:

- Aggregating inception features in different layers (with varied resolutions) to enhance local detailed information and encode richer context information.

➢ Aggregated Inception Features:

- Similar to Inception architecture in GoogleNet [5], we get inception features through four different convolutional operations, with dilated convolutions applied.
- We further enhance the inception features by aggregating multi-layer inception features, by using channel concatenation.
- Each AIF is computed by fusing the inception features of current layer with two directly adjacent layers.

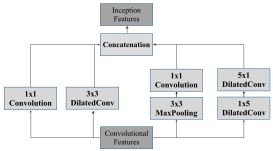

Figure 3: Architecture of Inception Module

Experiment Results

➢ Comparisons with state-of-the-art results:

Table 1. Performances on the ICDAR 2013 and ICDAR 2015 datasets

ICDAR 2013 dataset								ICDAR 2015 dataset			
Method	ICDAR Standard			DetEval			Time(s)	Method	R	P	F
	R	P	F	R	P	F					
SSD [17]	0.60	0.80	0.68	0.60	0.80	0.69	0.10	Deep2Text-MO	0.32	0.50	0.39
Yin [33]	0.66	0.88	0.76	0.69	0.89	0.78	0.43	HUST_MCLAB	0.44	0.38	0.41
Neumann [20]	0.72	0.82	0.77	-	-	-	0.40	AJOU	0.47	0.47	0.47
Neumann [21]	0.71	0.82	0.76	-	-	-	0.40	NJU-Text	0.36	0.73	0.48
FASText [1]	0.69	0.84	0.77	-	-	-	0.15	CASIA USTB	0.40	0.62	0.48
Zhang [35]	0.74	0.88	0.80	0.76	0.88	0.82	60	StradVision1	0.46	0.53	0.50
TextFlow [27]	0.76	0.85	0.80	-	-	-	0.94	StradVision2	0.37	0.77	0.50
Text-CNN [8]	0.73	0.93	0.82	0.76	**0.93**	0.84	4.6	MCLAB_FCN [36]	0.43	0.71	0.54
FCRN [4]	0.76	**0.94**	0.84	0.76	0.92	0.83	1.27	CTPN [28]	0.52	0.74	0.61
TextBoxes [16]	0.74	0.86	0.80	0.74	0.88	0.81	0.09	Yao *et. al.*[32]	0.57	0.72	0.64
CTPN [28]	0.73	0.93	0.82	0.83	**0.93**	**0.88**	0.14	DMPNet [18]	0.68	0.73	0.71
Proposed method	**0.86**	0.88	**0.87**	**0.86**	0.89	**0.88**	0.13	Proposed method	**0.73**	**0.80**	**0.77**

Table 2: Performance on COCO-text dataset

COCO-Text dataset			
Method	Recall	Precision	F-score
Baseline model C	0.05	0.19	0.07
Baseline model B	0.11	0.90	0.19
Baseline model A	0.23	0.84	0.36
Yao [32]	0.27	0.43	0.33
Proposed method	**0.31**	**0.46**	**0.37**

➢ Exploration study:

Table 3: Exploration study on the ICDAR 2013 dataset

Method	Recall	Precision	F-measure
SSD [17]	0.60	0.80	0.68
Textboxes [16]	0.74	0.86	0.80
Baseline Model	0.78	0.87	0.82
TAM	0.83	0.87	0.85
HIM	0.82	0.87	0.84
TAM+HIM	0.86	0.88	0.87

➢ Qualitative results: