Optimal Boxes: Boosting End-to-End Scene Text Recognition by Adjusting Annotated Bounding Boxes via Reinforcement Learning

Jingqun Tang[1*], Wenming Qian[2*], Luchuan Song[3], Xiena Dong[4], Lan Li[5], and Xiang Bai[6]

[1] Ant Group
jingquntang@163.com
[2] NetEase Fuxi AI Lab
wenmingqian@corp.netease.com
[3] University of Rochester
lsong11@ur.rochester.edu
[4] Hangzhou Dianzi University
dxn@hdu.edu.cn
[5] Wuhan University
2016302580090@whu.edu.cn
[6] Huazhong University of Science and Technology
xbai@hust.edu.cn

1 Related Works

1.1 Two-Step OCR Systems

In two-step systems, due to the fact that detected texts are cropped from the image, the detection and recognition are two separate steps. Some of these methods first generate text proposals using a text detection model [36,19,32] and then recognize them with a text recognition model [9,18,6]. Jaderberg et al. [9] use a combination of Edge Box proposals [37] and a trained aggregate channel features detector [5] to generate candidate text bounding boxes. Liao et al. [18] combine an SSD [20] based text detector and CRNN [30] to spot text in images. In addition, for the detection step, EAST [36] further simplifies the anchor-based detection by adopting the U-shaped design [29] to integrate features from different levels. And for the recognition step, RARE [31] consists of a Spatial Transformer Network (STN) and a Sequence Recognition Network (SRN), which is robust to irregular text. One major disadvantage of two-step methods is that the propagation of error between the recognition and detection models will result in less satisfactory performance.

1.2 End-to-End OCR Systems

Many end-to-end trainable networks have recently been proposed [2,3,15,8,21]. Bartz et al. [2] present a solution that employs a STN [10] to attend to each

[*] Equal contribution. Corresponding author.

word in the input image circularly and then recognize them individually. Li et al. [15] substitute the object classification module in Faster-RCNN [28] with an encoder-decoder-based text recognition model and to create their text spotting system [3,8,21] develop unified text detection and recognition systems with very similar overall architectures, which consist of a recognition branch and a detection branch. Liu et al. [22] design a novel BezierAlign layer for extracting accurate convolution features of a text instance with arbitrary shapes and adaptively fit arbitrarily-shaped text via a parameterized Bezier curve. Liao et al. [17] propose Mask Text Spotter v3, an end-to-end trainable scene text spotter that adopts a Segmentation Proposal Network (SPN) instead of an RPN [28].

1.3 Reinforcement Learning

In earlier work, Mnih et al. [24] present the first deep learning model to successfully learn control policies directly from high-dimensional sensory input using reinforcement learning. In recent years, reinforcement learning [14,11,4,23] has evolved considerably in the field of object detection. Some works [26,33] employ reinforcement learning as a post-processing method for scene text detection to adjust the bounding boxes predicted by the detection model, which can result in a significant time increase. In contrast to earlier research, we employ text recognition as the reward rather than the IOU between the predicted and ground-truth bounding boxes. In addition, our method is not a post-processing of the detection and does not add any extra computational costs to the inference phase.

2 Experiments

2.1 Datasets

ICDAR-2013 [13] (IC13) is released during the ICDAR 2013 Robust Reading Competition for focused scene text detection, consisting of high-resolution images, 229 for training and 233 for testing, containing texts in English. The annotations are at word-level using rectangular boxes.

ICDAR-2015 [12] (IC15) is presented for the ICDAR 2015 Robust Reading Competition. All images are annotated with word-level and quadrilateral boxes.

ICDAR-2019ReCTS [35] (ReCTS) is a newly-released large-scale dataset that includes 20,000 training images and 5,000 testing images, covering multiple languages, such as Chinese, English and Arabic numerals. The images in this dataset are mainly collected in the scenes with signboards, All text lines and characters in this dataset are annotated with bounding boxes and transcripts.

ICDAR-2019MLT [25] (MLT19) is a scene text detection dataset, including Chinese, Japanese, English, Arabic, etc. The images in MLT19 are collected from a variety of scenes, and it also contains many real-scene noises.

SynthText-80k [7] & SynthText-MLT [25] are large-scale synthetic datasets, which are adopted as pretraining for our BoxDQN and text spotting models. Furthermore, due to the difference in distribution between the synthetic and real datasets, the synthetic datasets are also used in cross-domain experiments.

Table 1. The quantitative results of our method on the representative baseline models. Gain stands for the improvement of the F-Score with and without BoxDQN. We bold the results of each gain to highlight the improvement of the effect by BoxDQN. The recognition of IC15 is performed with strong lexicon, and IC13 with generic lexicon, respectively.

Methods	**BoxDQN**	**IC13** [13]		**IC15** [12]		**MLT19** [25]		**ReCTS** [35]	
		F-score	Gain	F-score	Gain	F-score	Gain	F-score	Gain
[CRAFT+CRNN]	–	86.9	**2.0**	85.8	**1.8**	58.1	**3.2**	73.7	**2.4**
	✓	88.9		87.6		61.3		76.1	
[CRAFT+RARE]	–	87.1	**1.6**	86.7	**1.6**	60.2	**2.7**	75.1	**2.3**
	✓	88.7		88.3		62.9		77.4	
[PGNet]	–	85.8	**1.9**	82.4	**1.7**	54.8	**2.8**	71.0	**2.7**
	✓	87.7		84.1		57.6		73.7	
[Text Perceptron]	–	85.7	**1.8**	83.2	**1.2**	58.1	**2.5**	72.7	**2.1**
	✓	87.5		84.4		60.6		74.8	
[DBNET+SAR]	–			85.4	**1.0**	59.2	**1.8**		
	✓			86.4		61.0			

2.2 Experiments on Other Baseline Methods

To further validate the robustness of our approach, we conduct experiments on some other representative baseline models. As it is difficult to follow state-of-the-art methods when source code is not available, we focus on open source methods, specifically CRAFT [1]+CRNN [30], CRAFT [1]+RARE [31], PGNet [34], Text Perceptron [27], and EAST [36]+SAR [16]. Our experimental settings on these models remain consistent with the description in Section 4.2 of the submission. As can be seen from Tab.1, our method still works on these models.

2.3 Experiments on Scene Text Detection

Although our goal is ultimately to boost the performance of end-to-end text recognition, we also test the impact on text detection performance since our approach is based on adjusting the annotated bounding boxes. Specifically, we use the adjusted ground-truth boxes to train a detection model or a spotting model, and then predict text positions in the corresponding datasets. As shown in Tab.2, our method delivers a slight improvement in text detection, but the improvement is much lower than that of end-to-end recognition (0.5 *vs.* 1.6). This demonstrates that our approach also facilitates text detection while indeed reducing the inconsistency between annotated bounding boxes and deep representations of text recognition, thus improving end-to-end text recognition performance.

Table 2. The quantitative results of our method on scene text detection. The detection metric is also under IoU > 0.5. On the right-hand side of the table, we present the end-to-end recognition results, allowing for a direct comparison of BoxDQN's gains on detection and recognition tasks. The recognition of IC15 is performed with strong lexicon. IC15-Det denotes text detection on ICDAR2015 while IC15-e2e denotes end-to-end text recognition on ICDAR2015. For detection models, e2e indicates together with CRNN.

Methods	**BoxDQN**	**IC15-Det**		**IC15-e2e**	
		F-Score	Gain	F-Score	Gain
EAST [36]	–	84.1	**0.8**	82.2	**1.7**
	✓	84.9		83.9	
DBNet [19]	–	85.2	**0.5**	83.4	**1.7**
	✓	85.7		85.1	
CRAFT [1]	–	86.4	**0.3**	85.8	**1.8**
	✓	86.7		87.6	
FOTS [21]	–	85.5	**0.5**	81.5	**1.8**
	✓	86.0		83.3	
ABCNet [22]	–	86.1	**0.6**	82.4	**1.7**
	✓	86.7		84.1	
MTS-V3 [17]	–	86.9	**0.3**	83.1	**1.4**
	✓	87.2		84.5	
PGNet [34]	–	85.1	**0.6**	82.4	**1.7**
	✓	85.7		84.1	
Text Perceptron [27]	–	86.8	**0.4**	83.2	**1.2**
	✓	87.2		84.4	

References

1. Baek, Y., Lee, B., Han, D., Yun, S., Lee, H.: Character region awareness for text detection. In: Proc. CVPR. pp. 9365–9374 (2019)
2. Bartz, C., Yang, H., Meinel, C.: SEE: towards semi-supervised end-to-end scene text recognition. In: McIlraith, S.A., Weinberger, K.Q. (eds.) Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018. pp. 6674–6681. AAAI Press (2018), `https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16270`
3. Busta, M., Neumann, L., Matas, J.: Deep textspotter: An end-to-end trainable scene text localization and recognition framework. In: 2017 IEEE International Conference on Computer Vision (ICCV) (2017)
4. Caicedo, J.C., Lazebnik, S.: Active object localization with deep reinforcement learning. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (December 2015)
5. Dollar, P., Appel, R., Belongie, S., Perona, P.: Fast feature pyramids for object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence **36**(8), 1532–1545 (2014)
6. Gupta, A., Vedaldi, A., Zisserman, A.: Synthetic data for text localisation in natural images. In: IEEE Conference on Computer Vision & Pattern Recognition (2016)
7. Gupta, A., Vedaldi, A., Zisserman, A.: Synthetic data for text localisation in natural images. In: Proc. CVPR. pp. 2315–2324 (2016)
8. He, T., Tian, Z., Huang, W., Shen, C., Qiao, Y., Sun, C.: An end-to-end textspotter with explicit alignment and attention. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5020–5029 (2018)
9. Jaderberg, M., Simonyan, K., Vedaldi, A., Zisserman, A.: Reading text in the wild with convolutional neural networks. International Journal of Computer Vision (2016)
10. Jaderberg, M., Simonyan, K., Zisserman, A., et al.: Spatial transformer networks. Advances in neural information processing systems **28**, 2017–2025 (2015)
11. Jie, Z., Liang, X., Feng, J., Jin, X., Lu, W., Yan, S.: Tree-structured reinforcement learning for sequential object localization. In: Advances in Neural Information Processing Systems. pp. 127–135 (2016)
12. Karatzas, D., Gomez-Bigorda, L., Nicolaou, A., Ghosh, S., Bagdanov, A., Iwamura, M., Matas, J., Neumann, L., Chandrasekhar, V.R., Lu, S., et al.: Icdar 2015 competition on robust reading. In: ICDAR. pp. 1156–1160 (2015)
13. Karatzas, D., Shafait, F., Uchida, S., Iwamura, M., i Bigorda, L.G., Mestre, S.R., Mas, J., Mota, D.F., Almazan, J.A., De Las Heras, L.P.: Icdar 2013 robust reading competition. In: Proc. ICDAR. pp. 1484–1493 (2013)
14. Kong, X., Xin, B., Wang, Y., Hua, G.: Collaborative deep reinforcement learning for joint object search. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
15. Li, H., Wang, P., Shen, C.: Towards end-to-end text spotting with convolutional recurrent neural networks. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (Oct 2017)
16. Li, H., Wang, P., Shen, C., Zhang, G.: Show, attend and read: A simple and strong baseline for irregular text recognition. In: Proceedings of the AAAI conference on artificial intelligence. vol. 33, pp. 8610–8617 (2019)

17. Liao, M., Pang, G., Huang, J., Hassner, T., Bai, X.: Mask textspotter v3: Segmentation proposal network for robust scene text spotting. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16. pp. 706–722. Springer (2020)
18. Liao, M., Shi, B., Bai, X., Wang, X., Liu, W.: Textboxes: A fast text detector with a single deep neural network. In: Thirty-first AAAI conference on artificial intelligence (2017)
19. Liao, M., Wan, Z., Yao, C., Chen, K., Bai, X.: Real-time scene text detection with differentiable binarization. In: Proc. AAAI. pp. 11474–11481 (2020)
20. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: Ssd: Single shot multibox detector. In: European conference on computer vision. pp. 21–37. Springer (2016)
21. Liu, X., Ding, L., Shi, Y., Chen, D., Yan, J.: Fots: Fast oriented text spotting with a unified network. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
22. Liu, Y., Chen, H., Shen, C., He, T., Wang, L.: Abcnet: Real-time scene text spotting with adaptive bezier-curve network. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
23. Mathe, S., Pirinen, A., Sminchisescu, C.: Reinforcement learning for visual object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2894–2902 (2016)
24. Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., Riedmiller, M.: Playing atari with deep reinforcement learning. Computer Science (2013)
25. Nayef, N., Patel, Y., Busta, M., Chowdhury, P.N., Karatzas, D., Khlif, W., Matas, J., Pal, U., Burie, J.C., Liu, C.l., et al.: Icdar2019 robust reading challenge on multi-lingual scene text detection and recognition—rrc-mlt-2019. In: 2019 International Conference on Document Analysis and Recognition (ICDAR). pp. 1582–1587. IEEE (2019)
26. Peng, X., Huang, Z., Chen, K., Guo, J., Qiu, W.: Rlst: A reinforcement learning approach to scene text detection refinement. In: 2020 25th International Conference on Pattern Recognition (ICPR). pp. 1521–1528. IEEE (2021)
27. Qiao, L., Tang, S., Cheng, Z., Xu, Y., Niu, Y., Pu, S., Wu, F.: Text perceptron: Towards end-to-end arbitrary-shaped text spotting. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 11899–11907 (2020)
28. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis & Machine Intelligence **39**(6), 1137–1149 (2017)
29. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. pp. 234–241. Springer International Publishing, Cham (2015)
30. Shi, B., Bai, X., Yao, C.: An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. IEEE Transactions on Pattern Analysis & Machine Intelligence **39**(11), 2298–2304 (2016)
31. Shi, B., Wang, X., Lyu, P., Yao, C., Bai, X.: Robust scene text recognition with automatic rectification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016)
32. Tang, J., Zhang, W., Liu, H., Yang, M., Jiang, B., Hu, G., Bai, X.: Few could be better than all: Feature sampling and grouping for scene text detection. In: Pro-

ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4563–4572 (2022)

33. Wang, H., Huang, S., Jin, L.: Focus on scene text using deep reinforcement learning. In: 2018 24th International Conference on Pattern Recognition (ICPR). pp. 3759–3765. IEEE (2018)
34. Wang, P., Zhang, C., Qi, F., Liu, S., Zhang, X., Lyu, P., Han, J., Liu, J., Ding, E., Shi, G.: Pgnet: Real-time arbitrarily-shaped text spotting with point gathering network. AAAI. AAAI pp. 2782–2790 (2021)
35. Zhang, R., Zhou, Y., Jiang, Q., Song, Q., Li, N., Zhou, K., Wang, L., Wang, D., Liao, M., Yang, M., et al.: 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 (2019)
36. Zhou, X., Yao, C., Wen, H., Wang, Y., Zhou, S., He, W., Liang, J.: East: an efficient and accurate scene text detector. In: Proc. CVPR. pp. 5551–5560 (2017)
37. Zitnick, C.L., Dollar, P.: Edge boxes: Locating object proposals from edges. In: European Conference on Computer Vision (2014)