Learning Deeply Supervised Scene Text Detectors from Scratch

Wei Zhu, Mingwu Ren and Qingyuan Xia
School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, Jiangsu, China.
Email: mingwuren@163.com

Abstract. In this paper, we propose deeply supervised scene text detector (DSTD), a framework that can be learned from scratch. Our proposed method mainly addresses two problems. The first one is that state-of-the-art text detectors rely heavily on the off-the-shelf pre-trained models, which leads to several limitations including inflexibility and domain mismatch. The second problem is that unlike general objects, scene text usually appear in arbitrary orientations. Text detection using horizontal bounding boxes is inaccurate. In DSTD, we propose to regress rotated rectangles directly from horizontal default boxes to deal with multi-oriented text. Furthermore, we abandon the heavy pre-trained model from the SSD framework and incorporate dense layer-wise connections, enabling the network to be learned from scratch. The proposed method is evaluated on two public datasets, namely ICDAR2013 and ICDAR2015. Experimental results demonstrate its superiority over several state-of-the-art approaches.

1. Introduction
Reading text in the wild is significant in a variety of advanced computer vision applications, such as image retrieval, scene understanding and visual assistance, since text in images usually conveys valuable information. As a prerequisite of a text reading system, text detection plays a critical role in the whole procedure. Though extensively studied in recent years, text detection in unconstrained environments is still quite challenging due to a number of factors, such as high variation in character font, size, colour, orientation as well as complicated background and non-uniform illumination.

Previous works for scene text detection are mainly based on sliding-window and connected components analysis. These methods have achieved promising performance on several benchmarks. However, most of them are confined by low computational efficiency and weak discriminative power of hand-crafted features.

Recently, hand-engineered features have largely been replaced with deep convolutional features. Several deep learning based methods have been developed for text detection. In a sense, text detection can be regarded as a specific object detection problem, leading to a bunch of methods [1-5] that are designed based on generic object detection models. Most of them are fine-tuned from a network pre-trained on ImageNet [6] for a better performance. Zhiqiang Shen et al. [7] propose a much more compact model named deeply supervised object detector (DSOD), which can learn detectors from scratch. Besides, the feature map in the last layer is often too coarse for detection, especially objects with small size. Several works [8-10] combine low layer and high layer features to inject larger contextual information to deal with this problem. Motivated by these works, we develop a deeply supervised text detector (DSTD) which adopts dense structure to combine deep, high-level features with shallow, low-level feature. The feature fusion makes features more abundant for detecting text of
various sizes. The densely connected pyramid network abandons the heavy pre-trained network and thus can be trained from scratch. Furthermore, for long and inclined words, the horizontal bounding boxes may result in redundant background and thus reduce the accuracy on detection performance. To address this problem, we propose to regress the rotated rectangles directly from horizontal default boxes, which enables our network to detect arbitrary-oriented text in natural images.

2. Methodology
In this section, we first elaborate the construction of our network architecture. The architecture of our DSTD is depicted in figure 1. Then we describe the training methods and details in the following parts.

2.1. Network Architecture
The proposed DSTD network is a multi-scale detection framework derived from SSD [11]. The model can be decomposed into two parts: the backbone network for feature extraction and the detection subnetwork for prediction over multiple feature maps.

We follow the four design principles in [7] to construct the network. First, a stem block which consists of three 3×3 convolution layers followed by a 2×2 max pooling layer is defined. After that, four dense blocks are connected to extract multiple feature maps. The main difference between DSTD and SSD is the adoption of dense structure [10]. Any layer in a dense block direct connects all the subsequent layers and then a concatenation operation is used for feature fusion. The growth rate k of a dense block means each layer produces k feature maps. Transition with or without pooling layers are used between the dense blocks to restrict the numbers and sizes of output feature maps. In the prediction stage, we also employ skip connections to merge multi-scale information for each scale. The output feature maps of all the dense blocks build a pyramid feature hierarchy which has semantic information from low to high. Unlike [8], the feature maps of the dense blocks are enhanced with features from the previous scale. This process merges multi-scale information by concatenating the current features and the down-sampled outputs from the contiguous high-resolution feature maps. The detailed architecture is illustrated in figure 1.

2.2. Default Boxes
Different from general object detection, detecting text by horizontal bounding boxes may result in low accuracies in practice. Texts in real world tend to have large aspect ratios and can be enclosed by
rotated rectangles or quadrilaterals. To solve this problem, we propose to predict oriented bounding boxes from a number of pre-designed horizontal default boxes at each location (see figure 2 for an example). For a text region, the straightforward method to represent an inclined rectangle is using an angle to represent its orientation. However, due to the bias of the dataset, the distribution on angle is usually uneven which may make the model dataset-dependent. We adopt the representation proposed in [12]. A rotated rectangle $G_r$ is represented as $(x_{01}', y_{01}', x_{02}', y_{02}', h_0')$, where $(x_{01}', y_{01}')$ and $(x_{02}', y_{02}')$ are the coordinates of the first two points in clockwise, $h'$ is the height of the rectangle. The corresponding minimum horizontal rectangle $B$ is denoted as $(x, y, w, h)$, where $(x, y)$ means the centre of the box and $w_0$ and $h_0$ are the width and height respectively. The relationship between $G_r$ and $B$ is as follows:

\[
\begin{align*}
    x_{01}' &= x - w/2, \
    y_{01}' &= y - h/2, \
    x_{02}' &= x + w/2, \
    y_{02}' &= y - h/2, \
    h_0' &= h.
\end{align*}
\] (1)

At each location in the feature map, the network predicts the classification score and the offsets to each associated default box denoted as $(x_0, y_0, w_0, h_0)$. Given the predicted vector $(\Delta x, \Delta y, \Delta w, \Delta h, \Delta x_1, \Delta y_1, \Delta x_2, \Delta y_2, \Delta h', c)$, the horizontal and oriented boxes can be calculated as following:

\[
\begin{align*}
    x &= x_0 + w_0 \Delta x, \
    y &= y_0 + h_0 \Delta y, \
    w &= w_0 \exp(\Delta w), \
    h &= h_0 \exp(\Delta h), \
    x_{0n}' &= x_{0n} + w_0 \Delta x_n', n = 1, 2, \
    y_{0n}' &= y_{0n} + h_0 \Delta y_n', n = 1, 2, \
    h' &= h_0' \exp(\Delta h')
\end{align*}
\] (2)

where $(x, y, w, h)$ and $(x_1', y_1', x_2', y_2', h')$ indicate the minimum horizontal and rotated rectangles respectively.

Figure 2. Illustration of the regression (red arrows) from a matched default box (white) to a ground truth rotated rectangle (green). The regression from the default box to minimum horizontal rectangle (yellow) is not shown for simplification.
2.3. Training and Optimization

In the training phase, the ground truth boxes are turned into oriented rectangles and minimum bounding rectangles. The horizontal boxes are used to match the default boxes according to box overlap following the matching scheme in [2]. The aspect ratios of default boxes are set to 1, 2, 3, 5, 1/2, 1/3 and 1/5. The model is trained to simultaneously minimize the text/non-text classification loss and the box regression loss. Let $c$ be the confidence, $v$ be the tuple of the ground truth box, and $v^*$ be the predicted location. The multi-task loss function is defined as:

$$L(y, c, v, v^*) = \frac{1}{N} \left( L_{cl}(y, c) + \lambda L_{loc}(v, v^*) \right)$$

where $N$ is the number of positive default boxes. $y$ is the indicator for matching the default box to the ground truth. $y=1$ means positive while 0 otherwise. The weight $\lambda$ is set to 0.2 in practice. We adopt softmax loss for classification and smooth L1 loss for regression as in [11]. Most of our training strategies are similar as [2], including hard negative mining and data augmentation.

3. Experiments and Results

In this section, we evaluate the proposed network on two popular benchmarks: ICDAR2015 [13] and ICDAR2013 [14]. The first dataset contains multi-oriented texts and the second has mostly horizontal texts. In addition, we also adopt an artificial synthetic dataset named SynthText [15] for pre-training.

3.1 Implementation Details

In the proposed method, the network is pre-trained on the SynthText dataset and fine-tuned on other datasets. We randomly pick up 200k images from the SynthText dataset since many images exhibit many similarities between the others. The input images are resized to 300*300 and 512*512 for multi-scale training. After the pre-training for 80k iterations, we continue training with a smaller learning rate for another 50k iterations.

3.2 Datasets and Evaluation Protocol

The ICDAR2015 dataset contains 1500 images which are divided into 1000 training images and 500 testing images. Text in this dataset may exhibit in various orientations, sizes, viewpoints and resolution. Word-level annotations are given as 8 coordinates of the bounding box in a clock-wise manner.

The ICDAR2013 dataset includes 229 images for training and 233 images for testing, whereas the dataset puts emphasis on near-horizontal scene text detection and word-level annotations are provided. The SynthText dataset consists of 800k images with approximately 8 million synthetic word instances generated by blending artificial texts into natural images. The images are segmented into contiguous regions and depth information are estimated to make synthetic images more realistic. Part of the dataset is used for pre-train the network.

For evaluation, we follow the precision, recall and f-measure protocol. A detected box $b$ is considered as a hit box if the IOU between $b$ and a ground truth box is larger than the threshold (generally set to 0.5).

3.3 Quantitative Results

We evaluate four models on these datasets to justify the effectiveness of our model. The four models are original SSD, SSD with dense prediction structure, our model without dense prediction structure and our proposed model.

We first evaluate our proposed network on ICDAR2015 dataset, which is the most challenge task at present. The left part in table 1 summaries the results comparing with some state-of-the-art approaches. As can be seen, models trained with dense prediction layers outperform those without dense structure. The dense structure improves SSD and DSTD from 79.19% to 79.65 and 81% to 81.32% respectively, which validates the effectiveness of the dense prediction structure. Our proposed DSTD achieves the highest f-measure of 81.32%, outperforming all the methods with heavy-headed pre-trained models.
Some examples are given in figure 3, our proposed model can detect arbitrary-oriented text and text of a variety of scales.

To validate the versatility of our proposed model, we also conduct experiments on a widely used horizontal dataset named ICDAR2013. As shown in the right part of table 1, our approach outperforms most of previous state-of-the-art methods except for CTPN. However, this method is specially designed for horizontal text detection. The original SSD structure trained with our learning method performs well which proves the effectiveness of our method for rotated rectangle regression.

| Methods          | ICDAR2015 |   | ICDAR2013 |   |
|------------------|-----------|---|-----------|---|
|                  | Recall    | Precision | F-score | Recall | Precision | F-score |
| MCLAB FCN [16]   | 43.09%    | 70.81% | 53.58%    | 77.81% | 88.14% | 82.65% |
| CTPN [1]         | 51.56%    | 74.22% | 60.85%    | 83%   | 93%    | 88%    |
| TextBoxes [2]    | -         | -     | -         | 82.59% | 87.73% | 85.08% |
| Yao et al. [17]  | 58.69%    | 72.26% | 64.77%    | 80.22% | 88.88% | 84.33% |
| DMPNet [3]       | 68.22%    | 73.23% | 70.64%    | -     | -     | -      |
| SegLink [4]      | 76.80%    | 73.10% | 75.00%    | 83.00% | 87.70% | 85.30% |
| EAST [5]         | 73.47%    | 83.57% | 78.20%    | -     | -     | -      |
| SSD-plain        | 77.31%    | 81.16% | 79.19%    | 81.59% | 89.68% | 85.44% |
| SSD-dense        | 77.86%    | 81.52% | 79.65%    | 81.88% | 90.12% | 85.80% |
| DSTD-plain       | 78.82%    | 83.30% | 81.00%    | 83.90% | 90.73% | 86.63% |
| Proposed DSTD    | **79.27%**| **83.49%** | **81.32%** | **84.18%** | **91.22%** | **87.56%** |

Figure 3. Some detection results. Images in the first row are from ICDAR2015 and the others are from ICDAR2013.

4. Conclusion
This paper presents a multi-oriented scene text detector which is learned from scratch. We adapt the network to predict the regressions from horizontal default boxes to oriented text represented by rotated rectangles. The proposed network abandons the heavy pre-trained model and can be trained from scratch. Experimental results and comparisons on some popular datasets validate the effectiveness of our proposed approach.

5. Acknowledgements
This work is supported by the National Natural Science Foundation of China (No. 61403202). We are also grateful to Jing Lou and Ling Gu for helpful discussions.
6. References

[1] Tian Z, Huang W, He T, He P and Qiao Y 2016 Detecting text in natural image with connectionist text proposal network *European Conf. on Computer Vision* pp 56-72

[2] Liao M, Shi B, Bai X, Wang X and Liu W 2017 Text Boxes: a fast text detector with a single deep neural network *AAAI Conf. on Artificial Intelligence* pp 4161-67

[3] Liu Y and Jin L 2017 Deep matching prior network: toward tighter multi-oriented text detection *Proc. Int. Conf. on Computer Vision and Pattern Recognition* pp 3454-61

[4] Shi B, Bai X and Belongie S 2017 Detecting oriented text in natural images by linking segments *Proc. Int. Conf. on Computer Vision and Pattern Recognition* pp 3482-90

[5] Zhou X, Yao C, Wen H, Wang Y, Zhou S, He W and Liang J 2017 EAST: an efficient and accurate scene text detector *Proc. Int. Conf. on Computer Vision and Pattern Recognition* pp 2642-51

[6] Fei-Fei L, Deng J and Li K 2009 ImageNet: constructing a large-scale image database *J. Vision* 9 1037

[7] Shen Z, Liu Z, Li J, Jiang YG, Chen Y and Xue X 2017 DSOD: learning deeply supervised object detectors from scratch *Proc. Int. Conf. on Computer Vision and Pattern Recognition* pp 1937-45

[8] Lin TY, Dollar P, Girshick R, He K, Hariharan B and Belongie S 2016 Feature pyramid networks for object detection *Proc. Int. Conf. on Computer Vision and Pattern Recognition* pp 936-44

[9] Kong T, Yao A, Chen Y and Sun F HyperNet: towards accurate region proposal generation and joint object detection *Proc. Int. Conf. on Computer Vision and Pattern Recognition* pp 845-53

[10] Huang G, Liu Z, Maaten LVD and Weinberger KQ 2017 Densely connected convolutional networks *Proc. Int. Conf. on Computer Vision and Pattern Recognition* pp 2261-69

[11] Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu CY and Berg AC 2016 SSD: single shot multibox detector *European Conf. on Computer Vision* pp 21-37

[12] Y. Jiang, X. Zhu, X. Wang, S. Yang, W. Li, H. Wang, P. Fu, and Z. Luo 2017 R2CNN: rotational region CNN for orientation robust scene text detection *Preprint* arXiv:1706.09579

[13] Karatzas D, Lu S, Shafait F, et al. 2015 ICDAR 2015 competition on Robust Reading *Int. Conf. on Document Analysis and Recognition* pp 1156-60

[14] Karatzas D, Shafait F, Uchida S, et al. 2013 ICDAR 2013 Robust Reading Competition *Int. Conf. on Document Analysis and Recognition* pp 1484-1493

[15] Gupta A, Vedaldi A and Zisserman A 2016 Synthetic Data for Text Localization in Natural Images *Proc. Int. Conf. on Computer Vision and Pattern Recognition* pp 2315-2324

[16] Zhang Z, Zhang C, Shen W, Yao C, Liu W and Bai X 2016 Multi-oriented Text Detection with Fully Convolutional Networks *Proc. Int. Conf. on Computer Vision and Pattern Recognition* pp 4159-67

[17] Yao C, Bai X, Sang N, Zhou X, Zhou S and Cao Z 2016 Scene Text Detection via Holistic, Multi-Channel Prediction *Preprint* arXiv:1606.09002