Natural Scene Text Detection Based On Multi-level Fusion Proposal Network

Tong Li¹, Wanggen Li¹*, Nannan Zhu¹, Xuecheng Gong¹ and Jiajia Chen²

¹School of computer and information, Anhui Normal University, Anhui Wuhu, China
²School of information and electromechanical engineering, Shanghai Normal University, Shanghai, China

Abstract. Natural scene text detection is a challenging issues. In this work, a multi-level features fusion two-stage text detection network was developed to solve the problem of the insufficient use for feature map and the difference between text and common target. In order to obtain deep semantic feature, the network framework of VGG-16 was improved, and different levels of convolution feature map were confused, meanwhile the fusion approach was partially adjusted. Also, vertical proposal network was used to classify and regress bounding box. Results indicated the accuracy and recall rate of this method are 85.4% and 81.0%, respectively, by evaluating the net on ICDAR2013 dataset. The experimental results suggest that the multi-level features fusion method can improve the efficiency of feature map in natural scene text detection.

Keywords: Natural scenes; Text detection; Feature fusion.

1. Introduction
With the development of detection frameworks, some text detection frameworks[1,2] were also emerging. Most of the mainstream detection networks[3] were revolutionized from the fine-tuned classification network. During detection network learning, the shallow features of classification network structure would be lost in the chain transmission process, the last several layers of network replaced the whole network learning effect, resulting in the loss of significance of deepening network depth in detection. Therefore, the detection network generally used a simpler structure as the feature extraction layer of the network[4,5].

Feature extraction networks usually increase the number of extracted features by increasing the number of network layers. However, only the increase of network layers will lead to the serious loss of network characteristics, network parameter explosion, network non convergence and other problems. Therefore, this method can not improve the ability of network learning.

2. Related Work
2.1. Text Detection Based on CNN
The two-stage text detection method based on deep learning was similar to the target detection method [6,7,8]. The reason why text was regarded as special target detection is that text did not have a complete closed boundary and cannot be regarded as an independent target. Moreover, it is correct to meet the IOU > 0.5, while text detection cannot be regarded as a single target location.

\[
IOU = \frac{box_{\text{true}} \cap box_{\text{predicted}}}{box_{\text{true}} \cup box_{\text{predicted}}}
\]  

(1)

Among them, box (true) is the real text detection bounding box, box (predicted) is the prediction bounding box, and IOU represents the ratio of the intersection and union of the areas of the two bounding boxes.

The Connectionist Text Proposal Network (CTPN) was the most primitive model for detection using Recurrent Neural Network (RNN). It used bidirectional long and short term memory network to learn sequence information. A series of fine-grained bounding boxes were generated by only predicting the height of the vertical direction with a fixed bounding box width of 16 pixels. The combination of bidirectional long and short term memory network and convolutional neural network made the detection network more consistent with the characteristics of the text sequence, so as to obtain better positioning accuracy.

2.2. Feature Fusion Mode
Efficient feature fusion is the key to improve the performance of the model. There were two classic methods of feature fusion: ‘add’ and ‘concat’. ‘add’ was equivalent to the superposition of information. The dimensions of the described image itself did not increase, but the amount of information under each dimension was increasing; ‘concat’ was more like the combination of channels, which was the increase of the features of the described image itself, and the information under each feature was unchanged. ResNet used ‘add’ to stack values, while DensNet used ‘concat’ to merge channels.

3. Multi-level Fusion Text Proposal Network (MFPN)
MFPN consisted of two parts: multi-level feature fusion feature map of feature extraction and RPN of vertical mechanism of proposal prediction.

Figure 1. General framework.  
Figure 2. Feature fusion.
3.1. Multi-level Feature Fusion Feature Map

In our framework, we used the improved VGG-16 network structure to obtain feature map from the original image. The vertical proposal Region Proposal Network (RPN) and regression section were entered. The improved VGG network was consisted of five connection blocks (conv1, conv2, conv3, conv4, conv5). In the continuous exploration, it was found that the use of (3 × 3), (1 × 1), (3 × 3) network structure was beneficial to the generated feature map from the results of the original VGG-16 structure and other network visualization. At the same time, Conv-BN-Relu structure was added to the last layer of conv3_x, conv4_x, conv5_x in this experiment, the results shown that it was beneficial to the experiment.

The visualization results of conv1 and conv2 showed that they all contained a large number of edges and line features. In ours, the convolution layer with larger convolution kernel and longer step length were used to increase the transverse receptive field in conv1, conv2. The visualization results of feature map of conv3, conv4, conv5 showed that it can represent multi-level features of different depths. In the selection of fusion mode, the ‘concat’ connection was affected by the experimental conditions, and there was a situation of data explosion, and the problem of large calculation was prominent. we finally chosen the form of point product for feature fusion by inspired of DSSD [9], and the result showed that point product can indeed achieve better fusion effect.

It was not enough to rely solely on CNN to learn the spatial characteristics of text images. The difference between text and single target showed that the learning sequence information in text detection could increase the reliability of detection. Therefore, we introduced the vertical proposal mechanism in RPN by inspired of CTPN.

3.2. Vertical Mechanism RPN

In this paper, The anchor and true box were set to a fixed width (16 pixels). By learning CNN space and RNN sequence, the vertical proposal sequence was generated by RPN. In the final NMS, the threshold was set to remove the negative sample vertical suggestion area, and the final reserved suggestion box was combined with the text line building algorithm to generate text line suggestions.

3.3. Bounding Box Regression

Through the generation of thin vertical box with fixed width, and through the text line generation algorithm to complete the text line construction, however, when the vertical box was not fully suitable for the text, fixed width proposal may cause many problems. Therefore, a tight bounding box was significant for text detection and recognition. In ours, combined with the regression model of CTPN, the regression box was corrected by calculating the frame offset of each suggestion box.

3.4. Loss Function

In the process of network training, the loss was divided into two parts: one was the loss in RPN training stage, the others was the loss caused by re classification and regression. Losses involved in RPN were defined as follows:

\[
\begin{align*}
\text{rpn_loss_class} &= -\sum_i p_i \log p_i^* \\
\text{rpn_bbox_reg} &= \sum_j p_j^* \text{smooth}_1(v_j - v_j^*)
\end{align*}
\]
\[ \text{smooth}_L(x) = \begin{cases} 
0.5x^2 & \text{if } |x| < 1 \\
|x| - 0.5 & \text{otherwise} 
\end{cases} \] (4)

\[ \text{loss}_1 = \text{rpn\_loss\_class} + \text{rpn\_bbox\_reg} \] (5)

In these functions, \text{rpn\_loss\_class} was the cross entropy loss function of classification. When calculating the confidence score, \(p_i\) and \(p_i^*\) were the probability of the text, which was used for the next non-maximum suppression operation. \text{rpn\_bbox\_reg} was a smooth L1 loss function, where \(v_j^*\) is the real value and \(v_j\) is the predicted value. Each \(v_j\) includes the y coordinate of the center coordinate and the offset of the height, which is used to record the regression value of each round of training, and will be updated continuously in the process of detection training. \text{loss}_1 was the total loss function formula of participating in RPN training.

\[ L_{\text{cls}}(s_j, s_j^*) = \sum_i \text{soft max}(s_j, s_j^*) \] (6)

\[ L_{\text{reg}}(v_j, v_j^*) = \sum_i \text{smooth}_L(v_j - v_j^*) \] (7)

\[ \text{loss}_2 = \lambda_1 L_{\text{cls}}(s_j, s_j^*) + \lambda_2 L_{\text{reg}}(v_j, v_j^*) \] (8)

\(L_{\text{cls}}(s_j, s_j^*)\) was the loss function of reclassification at this time, \(s_i^*\) represents the real value of classification, \(s_i\) is the predicted value, where the size of the confidence score represents the probability of judging the text. This loss was used to select the positive sample with the highest score. In \(L_{\text{reg}}(v_j, v_j^*)\), \(v_j^*\) is the true value, and \(v_j\) represents the predicted value. Each \(v_j\) includes the y coordinate of the center coordinate and the offset of the height, which is used to correct the y coordinate and the height of the anchor containing the text. The loss was used to supervise the bounding box regression of each anchor containing the text. \text{loss}_2 was the total loss function for the second classification and regression, where \(\lambda\) is the balance parameter.

Therefore, the total loss function was as follows:

\[ \text{loss\_total} = \text{loss}_1 + \text{loss}_2 \] (9)

### 3.5. Training and Testing

Our model was trained on 6000 natural scene images. The images were marked at the word level and resized to (600, 1000) scale. The completely training dataset and test dataset of ICDAR2013 were similar. The true label of the image was 16 pixels with a fixed width, which corresponded to the vertical proposal. Network output was word level proposal. The evaluation dataset was ICDAR2013 dataset, which was composed of 233 pictures of focused text taken in the field. The evaluation criteria were provided by the official website.
4. Experiments

In the process of the experiment, the feasibility of the cross layer connection was verified by connecting the fusion of different layers. In order to prove the advantages of point product fusion, experiments were carried out to verify the results of different methods. Finally, this method was compared with other published methods to prove the effectiveness of the fusion method. In order to prove that the semantic features were not enhanced due to the size of the training data set. The training of this experiment was carried out on the same 6000 pictures.

We ran the open-source tensorflow version code of CTPN network, marked as CTPN-tensorflow. The experiment showed that it was quite different from the original author of CTPN in the article. It was found that there was a slight difference between the open source version and the bidirectional long and short term memory network and sliding window parts. Therefore, after adding BLSTM, the network structure was adjusted again, marked as CTPN-tensorflow-blstm. The 3 and 4 row was simply replaced VGG-16, which may be limited by the influence of objective factors. In the hardware condition of NVIDIA GPU1050, the software condition of Ubuntu 16.04, and the overall experimental results of Python 2.7, the advantages of simple replacement feature extraction network were not great. Therefore, VGG-16 was selected for continuous adjustment, and the result showed that the improvement effect was obvious.

| Method             | NET       | Feature map | Precision | Recall  | F score |
|--------------------|-----------|-------------|-----------|---------|---------|
| CTPN-tensorflow    | VGG16     | Conv5       | 0.729     | 0.668   | 0.697   |
| CTPN-tensorflow-blstm | VGG16     | Conv5       | 0.833     | 0.651   | 0.731   |
| Resnet101-CTPN     | Resnet101 | Conv4_x     | 0.776     | 0.667   | 0.717   |
| Unet-CTPN          | Unet      | Unet        | -         | -       | MOO     |

Explore 1 expand the original conv3_x, conv4_x, conv5_x network structure by using (3 × 3), (1 × 1), (3 × 3) network structure. The specific operation was shown in the following figure 1: conv3, conv4, conv5 were all expanded in the way of (a). Explore 2 adjusted and changed the scale of the network part, as shown in the second line of Table 2, conv1, conv2 were replaced by (b), (c). Explore3+ and Explore4* were the exploration of the fusion mode. "+" represented the element summation mode and "*" represented the element point product mode. MFPN was the structure of this paper. It was the result of combining Conv + BN + Relu structure on the basis of Explore4*.
Table 3. Comparison with state-of-the-art publications on ICDAR2013.

| Datasets | IC13 Eval | DetEval |
|----------|-----------|---------|
|          | precision | recall  | F score | precision | recall  | F score |
| MMser [10] | 0.860     | 0.700   | 0.770   | -         | -       | -       |
| TextFlow[11] | 0.850     | 0.760   | 0.800   | -         | -       | -       |
| FCRNall+filters[12] | -         | -       | -       | 0.920     | 0.760   | 0.830   |
| FCN[13]   | 0.880     | 0.780   | 0.830   | -         | -       | -       |
| SSD[14]   | 0.800     | 0.600   | 0.680   | 0.800     | 0.600   | 0.690   |
| TextBoxes[15] | 0.860     | 0.740   | 0.800   | 0.880     | 0.740   | 0.810   |
| PRRN[16]  | 0.788     | 0.826   | 0.807   | -         | -       | -       |
| MFPN      | 0.854     | 0.810   | 0.831   | 0.880     | 0.810   | 0.844   |

As shown in the above table, the MFPN results were compared with the latest CNN based text detection method. Results indicated MFPN presented a significant improvement in recall rate, and the F score reached 84.4%, which means the multi-level fusion is beneficial to detection.

Figure 3. Results on ICDAR2013. Figure 4. Visualization results of conv1~5.

5. Summary
Based on the target detection framework, this paper proposed a deep multi-level fusion text detection framework. Several experiments were carried out to explore the level selection and fusion mode of network fusion, and stronger semantic features were obtained by combining different convolution layer features. Then, the bidirectional long and short-term memory network and vertical mechanism were introduced by inspiration of CTPN. Through the adjustment of network fusion structure, the overall detection effect on this basis has been improved, which verifies the effectiveness of the framework proposed in this paper. It is worth noting that MFPN has room for improvement in accuracy. In the future work, the improvement of accuracy will become the focus of work.

References
[1] Buta M, Neumannl ,Matas J . FASText: Efficient Unconstrained Scene Text Detector[C]// 2015 IEEE International Conference on Computer Vision (ICCV). IEEE, 2015.
[2] Tian Z,Huang W,HE T,et al. Detecting Text in Natural Image with Connectionist Text Proposal Network[J]. 2016.
[3] Wang L, Guo S, Huang W, et al. Places205-VGGNet Models for Scene Recognition[J]. Computer Science, 2015.
[4] Heimer R, Myrseth K, Schoenle R. YOLO: Mortality Beliefs and Household Finance Puzzles[J]. Working Papers, 2015.
[5] Lin T Y, Goyal P, Girshick R, et al. Focal Loss for Dense Object Detection[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, PP(99):2999-3007.
[6] Zhu X, Jiang Y, Yang S, et al. Deep Residual Text Detection Network for Scene Text[J]. 2017.
[7] Rui Z, Shifeng Z, Xiaobao W, et al. ScratchDet: Exploring to Train Single-Shot Object Detectors from Scratch[C]// Computer Vision and Pattern Recognition (CVPR), 2019 IEEE Conference on. IEEE, 2019.
[8] Fu C Y, Liu W, Ranga A, et al. DSSD: Deconvolutional Single Shot Detector[J]. 2017.
[9] Ren S, He K, Girshick R, et al. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2015, 39(6):1137-1149.
[10] Alsharif, O., and Pineau, J. 2013. End-to-end text recognition with hybrid HMM maxout models. CoRR abs/1310.1811.
[11] Tian, S.; Pan, Y.; Huang, C.; Lu, S.; Yu, K.; and Lim Tan, C. 2015. Text flow: A unified text detection system in natural scene images. In Proc. ICCV.
[12] Gupta, A.; Vedaldi, A.; and Zisserman, A. 2016. Synthetic data for text localisation in natural images. In Proc. CVPR.
[13] Zhang, Z.; Zhang, C.; Shen, W.; Yao, C.; Liu, W.; and Bai, X. 2016. Multi-oriented text detection with fully convolutional networks. In Proc. CVPR.
[14] Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; and Reed, S. E. 2016. SSD: single shot multibox detector. In Proc. ECCV.
[15] Liao M, Shi B, Bai X, et al. TextBoxes: A Fast Text Detector with a Single Deep Neural Network[J]. AAAI, 2017, pp. 4161–4167.
[16] Yang G, Wang Z, Zhang y, et al. Natural scene text detection based on vertical region regression network[J]. Computer engineering and Science, 2018, 40(07):1256-1263.