Advanced Chinese Character Detection for Natural Scene Based on EAST

Jianxin Zhang¹*, Yunhai Feng²

¹School of Electric Power, Inner Mongol University of Technology, Huhhot, Inner Mongolia, 010000, China
²School of Electric Power, Inner Mongol University of Technology, Huhhot, Inner Mongolia, 010000, China
yunhai233@163.com

Abstract. Currently in the field of image vision processing, text detection and text recognition under natural scenes is a challenging task. At present, most of the research on text positioning is mainly aimed at English, and the detection and recognition of Chinese characters is relatively few. EAST text detection algorithm is simple in structure and high in accuracy. However, because of the feature extraction network, there is a small sense field, which cannot be well adapted to the text detection of Chinese characters. In order to adapt it to the natural scene Chinese character text detection, the feature extraction network of the modified EAST algorithm is MobileNet-V2 and ResNet50, which increases the depth of the network and extracts more image features. The modified algorithm is tested on the dataset MSRA-TD500, and the obtained accuracy shows that the improved network performance is better than the original algorithm.

1. Introduction
At present, the related papers have been published in various top-level conferences, most of the papers focus on the problem of multi-direction natural scene text detection, the research direction of these articles is mainly to deal with the English text, there are less literature on the Chinese text, and some literature on natural scene text recognition and end-to-end natural scene text detection and recognition. Natural scene text detection techniques have mainly experienced text detection in horizontal alignment[1][2][3][4] to text detection in multi-direction alignment, from text detection in single English, Arabic numerals to text detection in multilingual languages. The natural scene text detection technology mainly goes through two stages: firstly, based on the traditional hand-designed features, and then the natural scene text detection method based on deep learning appears around 2014[4]. At present, it is the mainstream method to use deep learning to detect the text of natural scene, and using convolutional neural network (CNN) and circular neural network (RNN) to deal with natural scene images has been widely used.

Tian et al[3] Connectionist Text Proposal Network (CTPN) is the most widely circulated and influential deep learning text detection model, which can detect horizontal or slightly oblique text lines. Text lines can be seen as a sequence of characters rather than a single independent target in general object detection.

Some deep learning methods of scene text detection are modified by the object detection one stage method SSD[5]. SSD is an object detection algorithm with VGG-16 as the backbone network. Firstly, using the convolution layer and pooling layer of VGG-Net, the feature maps extracted from these
intermediate layers are fed into a detection and classification network respectively. Several candidate frames for possible objects are derived, followed by a non-maximum suppression (NMS) layer. The TextBoxes[6] algorithm is improved on the basis of SSD by modifying the aspect ratio of the output detection box into a strip shape, which is more suitable for the shape of the text, and changing the size of the convolution kernel of the feature extraction layer from 3*3 to 1*5, and a text recognizer for the CRNN[7] model after the detection is completed to adjust the detection results. TextBoxes shows high accuracy in testing horizontal text. The Seglink method proposed in the paper [8] considers that it is difficult to detect the whole text line at one time, so we first detect the local fragment, and then connect all the fragments through the fusion algorithm to get the final text line. The advantage of this is that the text line of any length can be detected, but the it’s to increase the complexity of the network.

Pixellink also has a good effect on the detection of tilted text[9]. The RRPN algorithm proposed in the paper [10] is improved on the basis of the Region Proposal Networks (RPN). The input image first enters the feature extractor composed of convolutional neural networks and then enters the RPN to generate a large number of text proposal regions. Using the softmax function to determine whether these proposal regions are text or background, obtain the final result of the rectangular text box with an angle. But the RRPN algorithm is almost ineffective for text with fuzzy and uneven illumination.

2. Advanced EAST algorithm

2.1 EAST

EAST algorithm first uses convolutional neural network to extract image features. The PVANet network was first used for feature extraction from stage1, stage2, stage3, stage4, respectively. PVANet or VGG16 as the feature extraction of the EAST, and the two networks cannot extract more image information at the same time because of their relatively small sense field. Changing the backbone network of the EAST algorithm can be adaptable to the detection of Chinese characters.

2.2 Advanced EAST

2.2.1 Mobilenet-V2

A novel lightweight feature extractor Mobilenet-V2, based on the inverted residual structure, is proposed in the paper [11], where the fast connection lies between narrow bottleneck layers. The intermediate expansion layer uses lightweight deep convolution as a nonlinear source to filter features.

The inverted residual module with a linear bottleneck extends the input low-dimensional compression representation first to high-dimensional and filters with lightweight deep convolution, followed by the feature projection back to low-dimensional representation with linear convolution. The network structure of MobileNetV2 is shown in Fig 1.

![Figure1. The structure of MobileNet-V2](image)

MobileNet-V2 uses the method of raising dimension, extracting feature, reducing dimension to enhance the depth of feature map and extract more effective information.
2.2.2 ResNet50
Before ResNet was proposed, the expansion of neural network depth was greatly limited. In the deep neural network, the gradient disappears with the deepening of the network, and the deep neural network is not easy to train. A residual network ResNet is proposed in [12] to solve this network degradation problem. The residual structure is shown in Fig 2.

For the convolutional layer, when the input is x, its learned feature is recorded as h(x). now we hope that it can learn the residual \( F(x) = H(x) - x \), so that the original learning feature is \( F(x) + x \). when the residual is 0, at this time the stacking layer only does the identity mapping, in fact the residual will not be 0, which will also enable the stacking layer to learn new features based on the input features and thus have better performance. The formula for a residual unit is as follows:

\[
y = F(x, \{W_i\}) + W_s x
\]

(1)

The following x also needs to be transformed by the parameter Ws so that the output shape of the front part is the same.

2.2.3 The structure of Advanced EAST
The improved EAST network structure mainly includes three parts: feature extraction, feature merging and output layer.

Feature extraction: ResNet or MobileNet-V2 is used as the backbone of the Chinese character detection algorithm to generate feature maps on different scales. And then different levels of feature maps are fused with pool-based feature maps, which cannot only use different scales of feature maps, but also reduce the amount of computation. deep features can locate large text, and shallow features can locate small text.

Feature merging layer: The feature diagram f1~f4, f1 of the output of the four convolutional segments
of the feature extraction layer is used as the input of the first feature fusion layer. After upsampling the extended dimension, it is fused with the output f2 of the third convolutional segment.

Upsampling: $g_i = \text{upsampling}(h_i)$

The purpose of upsampling is to extend the dimension of the feature map so that it is consistent with the scale of the feature map of the upper layer and facilitate the subsequent feature map fusion. The upsampling diagram is shown in Fig 4.

![Upsampling diagram](image)

Feature merging: $h_i = \text{conv}_3\times3(\text{conv}_1\times1([g_{i-1}; f_i]))$

Where $[g_{i-1}; f_i]$ is the channel merging, when the fusion, the depth of the feature map unchanged, corresponding to the feature map splicing.

Output layer: The output of text detection has three parts, which are the confidence of the detection box, the coordinates of the detection box and the rotation angle of the detection box.

Confidence values: The feature graph output in the feature merging layer performs a convolution operation of a convolution kernel of 1*1 to obtain a one-dimensional vector, which is processed by the sigmoid function, that is, the confidence of a text detection box. The calculation process is shown in Fig 5.

![The process of calculating confidence](image)

The sigmoid function is a commonly used activation function in neural networks.

$$\sigma(x) = \frac{1}{1 + e^{-x}} \quad (2)$$

The output of the sigmoid function $\sigma \in (0, 1)$.

Coordinate: The feature maps output in the feature merging layer are obtained by the convolution operation of four convolution kernels of 1*1. And then the four vectors are input to sigmoid function. Finally, the result of the sigmoid function is enlarged 512 times, and the bias of the four coordinate vertices of the text detection box is obtained.

![The process of the feature merging](image)

Rotation angle: The output feature map of merging layer are obtained by the convolution operation of one convolution kernels of 1*1. Then the rotation angle of the text detection box is obtained by the calculation of formula 1 and formula 2.

![Convolution Process of the output Layer](image)
output_2 = sigmoid(output_1) \quad (3)

angle = (output_2 - 0.5) \times \frac{\pi}{2} \quad (4)

3. Experiment

The two feature extraction networks were trained separately using the ADAM [11] optimizer for end-to-end training. To speed up the iteration, the image in the dataset is normalized to a picture of 512*512. The learning rate of the ADAM optimizer is set to 1e-3, batchsize set to 14, decays one tenth per 10000 batches, and iteratively 20 epochs.

To make the network generalization ability stronger, we select some data of CTW and RCTW-17 as a new data set, and use the tagging tool to remake the label of some data, so that the label of the data set is unified in txt format. The format of the label is also unified with the labeling tool, using the partial data of MSRA-TD500 and the partial data of CTW, RCW-17 as the test set. This data set is named CRM. The CRM data training Set consists of 8,000 text images of various scenarios and 1,000 images of the test set. Most of them are Chinese images, a few are English images, some are mixed text images. The label of the text image is in txt format, where the position coordinates of the text in the image are saved.

The network was tested on the MSRA-TD500 dataset, and the results are shown in the table below.

| Model       | Recall | Precision | F1-score |
|-------------|--------|-----------|----------|
| PVANet      | 0.6713 | 0.8356    | 0.7445   |
| VGG16       | 0.6160 | 0.8167    | 0.7023   |
| ResNet50    | 0.7001 | 0.8568    | 0.7706   |
| MobileNetV2 | 0.6978 | 0.8401    | 0.7624   |

From the table, the detection algorithm with ResNet50 as the feature extraction network is more accurate and analyzed from the depth of the network. ResNet networks have the largest number of layers in these networks, so to some extent, the deeper the network, the better the performance of the model.

The experimental results are shown in the following figure 9. From the test results, the algorithm has good results for the smaller text target, the larger text target and the inclined text target.

4. Conclusion

At present, in the research field of text detection is mainly aimed at English text detection, and the detection of Chinese characters is less. We adapt to the text detection of Chinese characters by improving the feature extraction network of the English text detection algorithm east. In this paper, the feature extraction network of the EAST algorithm has been modified. ResNet50 and MobileNet-V2 have been used to extract the features of the image successively. Because of the increase of the depth of the network, more text features can be extracted.

References

[1] Epshtein, B., Ofek, E. and Wexler, Y. (2010) Detecting text in natural scenes with stroke width transform. in the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2963 – 2970.
[2] Yi, C., Tian, Y. (2013) Text extraction from scene images by character appearance and structure modeling. Computer Vision and Image Understanding, vol. 117, no. 2, pp. 182 –194.
[3] Tian, Z., Huang, W., and Qiao, Y. (2016) Detecting text in natural image with connectionist text proposal network. in the 14th European Conference on Computer Vision. pp. 56 – 72.
[4] Neumann, L. & Matas, J. (2016). Real-time lexicon-free scene text localization and recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 38(9), 1872-
[5] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., & Fu, C. Y., et al. (2015). SSD: single shot multibox detector.

[6] Liao M, Shi B, Bai X, et al. (2016) TextBoxes: A Fast Text Detector with a Single Deep Neural Network[J].

[7] Shi B, Bai X. (2015) 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. pp. 39(11):2298-2304.

[8] Shi, B., Bai, X., & Belongie, S. (2017). Detecting oriented text in natural images by linking segments.

[9] Deng, D., Liu, H., Li, X., & Cai, D. (2018). Pixellink: detecting scene text via instance segmentation.

[10] Jianqi Ma, Weiyuan Shao, Hao Ye. (2017) Arbitrary-Oriented Scene Text Detection via Rotation Proposals. IEEE Transactions on Multimedia. pp. (99):1-1.

[11] Mark Sandler Andrew Howard, et al. (2018) MobileNetV2: Inverted Residuals and Linear Bottlenecks. Conference: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

[12] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE Computer Society.