Journal on Big Data
DOI:10.32604/jbd.2021.017184
Article

Tech Science Press

CTSF: An End-to-End Efficient Neural Network for Chinese Text with Skeleton Feature

Hengyang Wang, Jin Liu[*] and Haoliang Ren

Shanghai Maritime University, Shanghai, 201306, China
[*]Corresponding Author: Jin Liu. Email: jinliu@shmtu.edu.cn

Received: 23 January 2021; Accepted: 27 June 2021

Abstract: The past decade has seen the rapid development of text detection based on deep learning. However, current methods of Chinese character detection and recognition have proven to be poor. The accuracy of segmenting text boxes in natural scenes is not impressive. The reasons for this strait can be summarized into two points: the complexity of natural scenes and numerous types of Chinese characters. In response to these problems, we proposed a lightweight neural network architecture named CTSF. It consists of two modules, one is a text detection network that combines CTPN and the image feature extraction modules of PVANet, named CDSE. The other is a literacy network based on spatial pyramid pool and fusion of Chinese character skeleton features named SPPCNN-SF, so as to realize the text detection and recognition, respectively. Our model performs much better than the original model on ICDAR2011 and ICDAR2013 (achieved 85% and 88% F-measures) and enhanced the processing speed in training phase. In addition, our method achieves extremely performance on three Chinese datasets, with accuracy of 95.12%, 95.56% and 96.01%.

Keywords: Deep learning; convolutional neural network; Chinese character detection; text segmentation

1 Introduction

With the rise of information age, image data which contain a wealth of information climb almost exponentially. Images with text are ubiquitous in our daily life. Such type of images in natural environment are known as scene texts. The results of text detection and character recognition on them can be applied in numerous practical applications, such as text translation, product label recognition, smart cars, intelligent robots and e-learning. Due to the background clutter, occlusion, intra-class variation, illumination conditions and viewpoint variation, scene text detection becomes an extremely demanding/challenging task. How to extract the information we need from these digital images has been a hot-point question in the field of machine vision. The result of these tasks was poor when using rules or traditional models before deep learning methods become popular in recent years. One mainstream technology of text recognition system named end-to-end method which can be generalized into two steps: text detection and text recognition. Text regions are detected and labeled with their bounding boxes in the stage of text detection, afterwards, text information is retrieved from the previous text regions in the text recognition phase. Benefit from the development of deep learning, the performance of text recognition system being able to significantly in the last few years. However, mainstream methods suffer from high FLOPs and memory consumption, long training and testing time, as well as low compatibility with popular pretrained backbones. Besides, the accuracy of Chinese text recognition is low. In this paper, we propose a light weight deep neural network which achieves similar accuracy and less than half of parameters compare to popular neural networks. In addition, we improve the algorithm and insert it into our neural network as a block. This algorithm rises the accuracy of text recognition in ways that handling the pixels in detect text area. For performance evaluation, the proposed method is tested on three benchmarks, namely, ICDAR2011, ICDAR2013 and ICDAR2017. Furthermore, the model we proposed is state of the art method on the custom Chinese character training set compare to some popular text detection models. The main contributions of this paper are summarized as: (1) We proposed a light weight deep neural network which achieve similar accuracy compare to the popular models, however, only has fewer half parameters than them. (2) We propose an efficient Chinese character processing method that can improve the accuracy of text recognition. (3) By introducing the two methods, we can achieve state-of-the-art performance on three benchmarks and custom Chinese character dataset.

2 Related Work

Scene text detection, as a problem with practical application value, has been researched for a long time. Before the emergence of deep learning, the mainstream in this area was bottom-up such as MSER or SWT. Now, these methods have been replaced by deep learning because of the greater performance of deep learning compare to its predecessors.

The R-CNN model designed by Girshick et al. [1] and others based on a convolutional neural network uses the Selective search algorithm [2] to extract about 2000 candidate regions on the image, and then image these regions through image scaling operations Normalize to the same size and input them into the convolutional neural network together, and then use regression to further adjust the position of the candidate frame. R-CNN divides the training and testing process into multiple parts, with cumbersome steps, time-consuming training and slow testing speed.

In order to solve the above problems, Fast R-CNN [3] was proposed. It uses the idea of pooling and outputting a fixed size to improve R-CNN. Input the entire image to CNN to obtain the feature map, and the selective search algorithm The candidate region is mapped to the feature map to obtain the region of interest, and then through the RoI pooling layer, each candidate region generates a fixed-size feature block. But Fast R-CNN also has shortcomings: it relies too much on manually set anchor points in the selection of candidate regions, and weight sharing leads to misclassification.

Wang et al. [4] and others proposed a Chinese text detection algorithm TDSI based on natural scene images in 2018. This algorithm uses a series of heuristic rules to improve the MSER algorithm and the SWT algorithm, respectively. According to the structural characteristics of Chinese characters, the candidate Features such as the center of mass of the area and the overlapping area cool down the area to gather Chinese characters. The experimental results show that the algorithm obtains better accuracy, recall and *F* value in its self-check image database.

For text detection in natural scenes, Dai et al. [5] proposed an improved algorithm that combines the maximum stable extreme value region algorithm with the convolutional depth belief network. The algorithm extracts candidate regions from the maximum stable extreme value region and combines Input it into the convolutional deep belief network to extract features, and finally classify the features. This method works well on the ICDAR data set and SVT data set.

Sinan et al. [6] proposed an end-to-end model for multi-scene text detection, which integrates a text segmentation network, combines multi-layer features in the feature extraction process, and uses semantic segmentation and target detection based on region suggestions. Task, the text is detected, and the experiment proves the effectiveness of the model.

Aiming at the problem of the uncertainty of the text direction in natural scene images, Yan et al. [7] proposed an algorithm for extracting features of character candidate regions based on the maximum stable extreme value region of color, and used heuristic rules to perform non-text regions. Screening and designing the color model to restore the mistakenly deleted areas and get accurate results through the CNN network. The algorithm has obtained relatively good results on both ICDAR2011 and ICDAR2013.

Rafaa et al. [8] proposed a VGG16 network using Single Shot Multibox Detection technology. This method can be extended to text detection tasks with different aspect ratios. The experimental results prove that the algorithm uses anchor box annotations compared to existing ones. The complex background text detection algorithm has a certain improvement.

Our method integrates the advantages of Regression-based and end-to-end methods. We propose CTSF, which can be generalized as a mix of CTPN, PVANet [9–11], Spatial Pyramid Pooling (SPP) and skeleton extraction algorithm. PVANet is used as an efficient backbone network to extract robust features. We introduce SSP and Chinese skeleton extraction algorithm to handle the complexities and nuances of elaborate character shapes found in the Chinese character.

3 Method

In this section, we first introduce the text detection module. We added a vertical anchor regression mechanism to CTPN to detect text lines, and at the same time use PVANet's image feature extraction module to optimize and accelerate the detection network [12–13]. Finally, this paper also proposes a literacy model based on spatial pyramid pooling and fusion of Chinese character skeleton features, which reduces the loss of image information caused by fixed input image size through spatial pyramid pooling; improves the accuracy of Chinese character recognition by adding Chinese character skeleton features [14–16].

The formal expression of the text line detection task is as follows, assuming that the input accepted by the text line detection model is an RGB image with a height of H and a width of W:

$$F_{in} = \{p_{x,y,c} | x \in [1,W], y \in [1,H], c \in [1,3]\} \quad (1)$$

where represents the pixel value of $P_{x,y,c}$at (x,y) in the channel No. C [17–18]. The final output of the model is:

$$F_{out} = ((x_1, y_1, w_1, h_1), (x_2, y_2, w_2, h_2), \ldots, (x_i, y_i, w_i, h_i)) \quad (2)$$

where x, y, w, h represents the abscissa, ordinate, width and length of the final output text line positioning box. Each set of arrays containing these four values can be uniquely represented as a single text box [19–21]. The ultimate goal of the text detection task is to find a mapping H [22–23], which can match the input of the image with the output of the text box one by one:

$$F_{out} = H(F_{in}) \quad (3)$$

3.1 Detection Block

Our method integrates the advantages of Regression-based and end-to-end methods. We propose CDSE, which can be summarized as a combination of CTPN and PVANet [24]. PVANet is used as an efficient backbone network to extract robust features [25]. The schematic illustration of our architecture can be shown as Fig. 1. Our backbone network is PVANet with BiLSTM where the feature maps of conv3, conv4, conv5 in PVANet are merged as shared feature map.

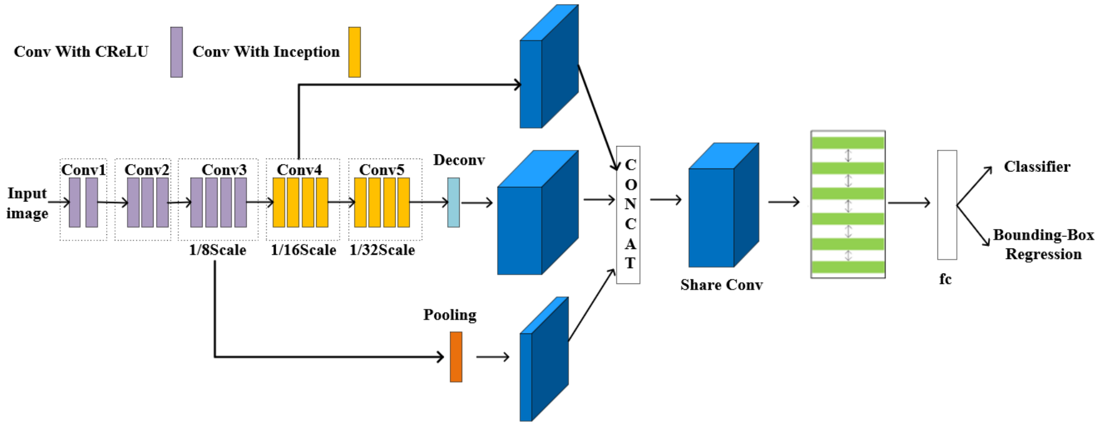

Figure 1: The network structure of CDSE

In this paper, PVANet is adopted as the backbone network. Compare with VGG16 [26], PVANet lead to 2X speed-up of the early stage without losing accuracy. In the early stage, the distribution of parameters has a strong negative correlation which progressive fade away as the network deepens. The network tends to capture both of them simultaneously, while ReLU erase the negative phase information which lead to the redundant of kernel. To solve this problem, we propose CSeLU (Concatenate Scale Exponential Linear Unit) [27]. Compared to the ReLU in PVANet, we append inverse probability integral transform block after scale and shift block to allow that the output of prior block can be to fit to Gaussian distribution. The definition of inverse probability integral transform block can be shown as:

$$T(u) = \mu + \sigma^2 * \Phi^{-1}(F_x(u)) \quad (4)$$

$$F_x(u) = P(X < x) = P(T(u) < x) = \Phi(\frac{T(u)-\mu}{\sigma}) \quad (5)$$

where $F_x(u)$ is the distribution of prior block, μ is the mean and σ is the variance. T(u) is the result fit to Gaussian distribution. The activation function can be represented by:

$$SeLU(x) = \lambda \begin{cases} x(x > 0) \\ \alpha(e^x - 1)(x \leq 0) \end{cases} \quad (6)$$

$$CSeLU(x) = [SeLU(x), SeLU(-x)] \quad (7)$$

CSeLU first halves the number of output channel, then doubles it by concatenating the output with its negative value.

NMS, as a common method for reducing false positive rate in object detection, trend to merge candidate boxes when they are overlapped [28]. This approach leads to an incorrect mergence when there are some overlaps or occlusions in the detection region. We adopt Soft-NMS to solve this problem, the function of Soft-NMS can be written as:

$$s_i = s_i e^{\frac{-iou(M,bi)2}{\sigma}}, \forall bi \notin D \quad (8)$$

Soft-NMS is an algorithm which decays the detection scores of all other objects as a continuous function of their overlap with the remaining candidate boxes in set M. In this way, the score of candidate box will be replaced with a lower value instead of setting it to zero directly.

3.2 Recognition Block

In order to improve the accuracy and robustness of the Chinese character recognition network, this paper proposes a convolutional neural network based on spatial pyramid pooling and fusion of Chinese character skeleton features for Chinese character recognition (SPPCNN-SF). The network structure is shown in the Fig. 2.

First of all, the input of the network consists of two parts, one is the original image of the Chinese character image, and the other is the Chinese character skeleton image obtained after the Chinese character skeleton extraction algorithm. Since the two have the same size, they are sent to the convolutional neural network [29]. Just superimpose two images together for concatenate operation (superposition in channel dimension). All the literacy models in this article use binarized images as the input of the neural network. The first two convolutional layers of the network use traditional convolution to prevent the direct use of depth separable convolution and cause a lot of image information loss. Following the traditional convolutional layer, the network has two branches, one uses the flip residual and linear bottleneck block to extract the deep information and high-dimensional features of the image, and the other uses the depth separable convolution to learn the spatial features of the image. Following the bottleneck block is the spatial pyramid pooling layer. We set the number of pooling scales to 3, and the scales are 4 × 4, 2 × 2 and 1 × 1, respectively. At this time, the output of the network is 21 channel-dimensional data. In order to enable the network to have better spatial information learning capabilities, this network adds a branch to convolve spatial features after the traditional convolutional layer. This branch uses two deep separable convolutional layers, the first step is 1, the second step is 2, the resulting spatial information feature map is put into a spatial pyramid pooling layer, the pooling of this layer. The number of scales is 2, which are 2 × 2 and 1 × 1, respectively. Add the data of the 21 and 5 channel dimensions on the two branches to get the data of 26 channel dimensions, and then connect the two fully connected layers and a softmax layer to get the complete literacy network of this article.

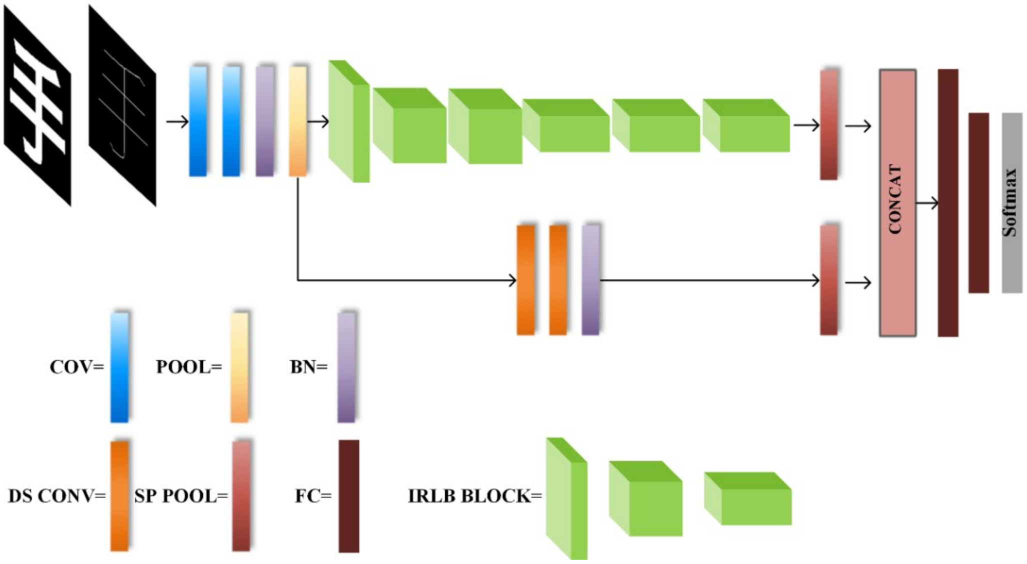

Figure 2: The network structure of SPPCNN-SF

4 Experiment

To compare CTSF with state-of-the-art methods, we perform experiments on three public scene text detection datasets, i.e., ICDAR 2013, ICDAR2015, ICDAR2017 and our own dataset. Each epoch contains 13365 iterations. We set the learning rate as 0.0001 and the training process cost 32 h on GTX2080Ti.

Fig. 3 and Fig. 4 show the comparison of the effects of our model with other methods on the ICDAR2011 and ICDAR2013 data sets. It can be seen that the accuracy, recall and *F* value of our model are all at a high level.

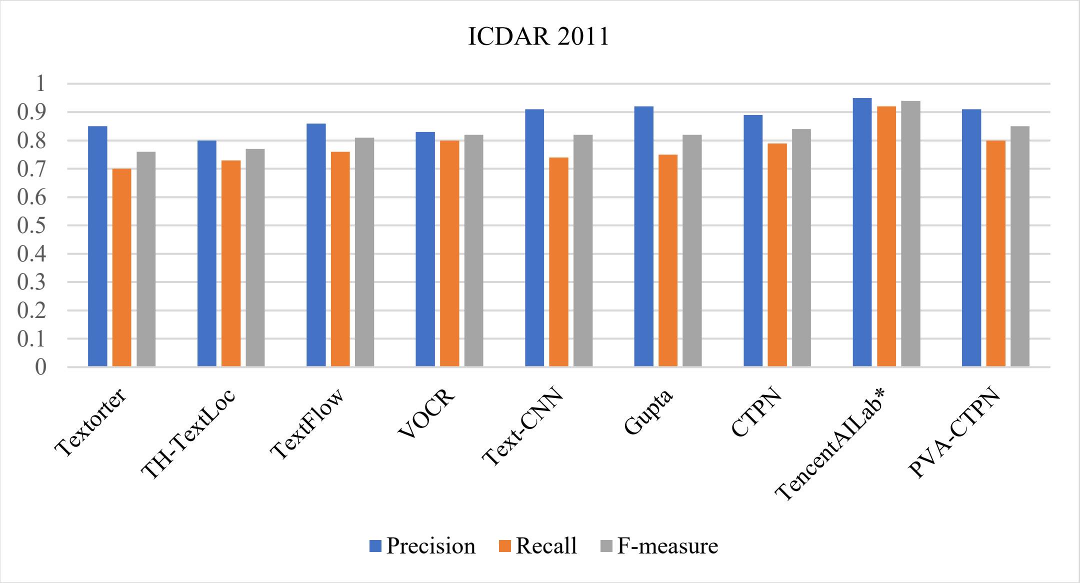

Figure 3: Comparison results of various models in the ICDAR2011 data set

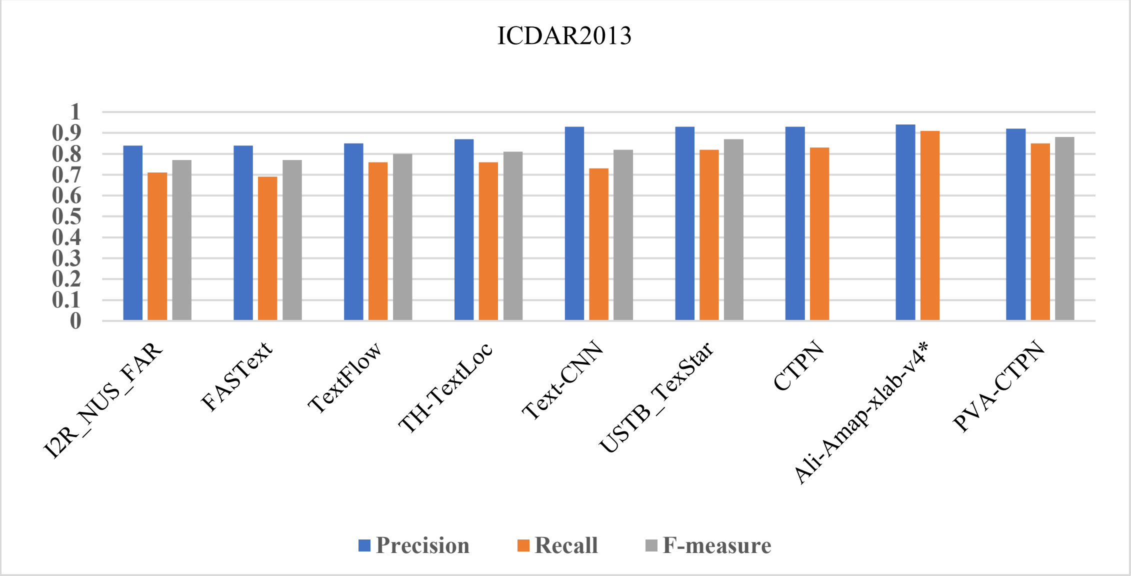

Figure 4: Comparison results of various models in the ICDAR2013 data set

Then, in order to verify the accuracy of the recognition module, we divided the test set in the printed data set into three sets and tested all the trained recognition models. The final recognition accuracy is shown in Table 1.

Table 1: Comparison test results of printed training set

Network structure	Test set 1	Test set 2	Test set 2
AlexNet	0.9323	0.9229	0.9381
LeNet	0.9428	0.9334	0.9395
VGG16	0.9389	**0.9556**	0.9432
SPPCNN-SF	**0.9512**	0.9531	**0.9601**

As can be seen from Table 1, our recognition module has achieved excellent results on two test sets.

5 Conclusion

In this paper, in order to solve the problems of low accuracy in text detection and recognition, we propose a lightweight neural network architecture called CTSF. It consists of two modules, one is a text detection network that combines CTPN and PVANet's image feature extraction module, called CDSE. The other is a literacy network based on spatial pyramid pool and fusion of Chinese character skeleton features, called SPPCNN-SF, so as to realize text detection and recognition respectively. The experimental results on ICDAR2011 and ICDAR2013 show that the method has achieved the latest technological achievements, which proves that our model has the special and universal capabilities.

Funding Statement: This work is supported by the National Natural Science Foundation of China (61872231, 61701297).

Conflicts of Interest: We declare that they have no conflicts of interest to report regarding the present study.

References

[1] R. Girshick, J. Donahue and T. Darrell, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 580–587, 2014.

[2] J. R. Uijlings, K. E. Sande and T. Gevers, "Selective search for object recognition," *International Journal of Computer Vision,* vol. 104, no. 2, pp. 154–171, 2013.

[3] R. Girshick, "Fast R-CNN," in *IEEE Int. Conf. on Computer Vision*, pp. 1440–1448, 2015.

[4] L. Wang and X. F. Zhang, "Scene text detection in convolutional deep belief networks," *Computer Systems & Applications*, vol. 27, no. 6, pp. 231–135, 2018.

[5] Y. Dai, Z. Huang and Y. Gao, "Fused text segmentation networks for multi-oriented scene text detection," in *IEEE 24th Int. Conf. on Pattern Recognition*, pp. 3604–3609, 2018.

[6] H. Sinan, Y. G., Guo and L. Zhang, "Multi-orientation natural scene text detection," *Application Research of Computers*, vol. 35, no. 7, pp. 2193–2196, 2018.

[7] C**.** Yan, K. Wu and C. Zhang, "A new anchor labeling method for oriented text detection using dense detection framework," *IEEE Signal Processing Letters*, vol. 25, no. 99, pp. 1295–1299, 2018.

[8] D. Rafaa and J. Nordin, "Offline OCR system for machine-printed Turkish using template matching," *Advanced Materials Research*, vol. 341, pp. 565–569, 2012.

[9] S. Hong, B. Roh and K. H. Kim, "PVANet: Lightweight deep neural networks for real-time object detection," in *NIPS 2016 Workshop on Efficient Methods for Deep Neural Networks*, 2016.

[10] A. Alhussain, H. Kurdi and L. Altoaimy, "A neural network-based trust management system for edge devices in peer-to-peer networks," *Computers, Materials & Continua*, vol. 59, no. 3, pp. 805–815, 2019.

[11] L. Pan, C. Li, S. Pouyanfar, R. Chen and Y. Zhou, "A novel combinational convolutional neural network for automatic food-ingredient classification," *Computers, Materials &* Continua, vol. 62, no. 2, pp. 731–746, 2020.

[12] J. Liu, Y. H. Yang and H. H. He, "Multi-level semantic representation enhancement network for relationship extraction," *Neurocomputing,* vol. 403, no. 5, pp. 282–293, 2020.

[13] S. J. Shang, J. Liu and Y. H. Yang, "Multi-layer transformer aggregation encoder for answer generation," *IEEE Access*, vol. 8, pp. 90410–90419, 2020.

[14] S.W. Chang and J. Liu, "Multi-lane capsule network for classifying images with complex background," *IEEE Access*, vol. 8, pp. 79876–79886, 2020.

[15] J. Liu, X. Zhang, Y. H. Li, J. Wang and H. J. Kim, "Deep learning-based reasoning with multi-ontology for IoT applications," *IEEE Access*, vol. 7, pp. 124688–124701, 2019.

[16] K. He, G. Gkioxari and P. Dollár, "Mask R-CNN," in *IEEE Int. Conf. on Computer Vision*, pp. 2961–2969, 2017.

[17] Z. Tian, W. Huang and T. He, "Detecting text in natural image with connectionist text proposal network," in *European Conf. on Computer Vision*, pp. 56–72, *2*016.

[18] Z. Zhou, G. Yao and H. Wen, "EAST: An efficient and accurate scene text detector," in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 5551–5560, 2017.

[19] D. Deng, H. Liu and X. Li, "Pixellink: Detecting scene text via instance segmentation," in *Thirty-Second AAAI Conf. on Artificial Intelligence*, 2018.

[20] W. Zaremba, I. Sutskever and O. Vinyals, "Recurrent neural network regularization," arXiv:1409.2329, 2014.

[21] K. Greff, R. K. Srivastava and K. Jan, "LSTM: A search space odyssey," *IEEE Neural Networks and Learning Systems*, vol. 28, no. 10, pp. 2222–2232, 2017.

[22] J. Liu, L. Lin and H. L. Ren, "Building neural network language model with POS based negative sampling and stochastic conjugate gradient descent," *Soft Computing*, 2018.

[23] Z. L. Zhang, J. Liu and C. K. Gu, "A chinese handwriting word segmentation method via faster R-CNN," *Computer Science and Ubiquitous Computing*, vol. 12, pp. 470–474, 2017.

[24] W. X. Yao, J. Liu, Z. H. Cai, "Personal attributes extraction in Chinese text based on distant-supervision and LSTM," *Computer Science and Ubiquitous Computing*, vol. 12, pp. 511–515, 2017.

[25] L. Lin, J. Liu, Z. K. Gu, Z. L. Zhang and H. L. Ren, "Build chinese language model with recurrent neural network," *Computer Science and Ubiquitous Computing*, vol. 12, pp. 920–925, 2017.

[26] D. Hakkani-Tür, G. Tür and A. Celikyilmaz, "Multi-domain joint semantic frame parsing using bi-directional

RNN-LSTM," *Interspeech*, pp. 715–719, 2016.

[27] F. A. Gers, J. Schmidhuber and F. Cummins, "Learning to forget: Continual prediction with LSTM," *Neural Computation*, vol. 12, no. 10, pp. 2451–2471, 2020.

[28] S. Sabour, N. Frosst and G. E. Hinton, "Dynamic routing between capsules," *Neural Information Processing Systems*, pp. 3856–3866, 2017.

[29] A. R. Kosiorek, S. Sabour, Y. W. Teh and G. E. Hinton, "Stacked capsule autoencoders," *Neural Information Processing Systems*, pp. 15486–15496, 2019.