Natural Scene Chinese Recognition Based on Deep Learning

Xiaofei Liu¹,²,³, Pan Tong¹,²,³, Tao Yu¹,²,³, Andi Dong¹,²,³ and Cuo Yong¹,²,³,*

¹School of Information Science and Technology, Tibet University, Lhasa, Tibet, 850000, China
²Key Laboratory of Artificial Intelligence, Tibet Autonomous Region, Tibet University, Lhasa, Tibet, 850000, China
³Engineering Research Center of the Ministry of Education for Tibetan Information Technology, Lhasa, Tibet, 850000, China
*e-mail: yc@utibet.edu.cn

Abstract. Text recognition is one of the research focuses in the field of computer vision. In recent years, many researchers have proposed recognition methods based on codec frameworks, such as CRNN, but most of these methods are based on English, and relatively few are based on Chinese. Chinese language, because of its large number of character types and contains a large number of similar characters, only using CNN to extract features will lose a large amount of information. Meanwhile, the method of giving combined attention is an emerging direction. In this paper, we study the recognition of the Chinese language and introduce attention to BiLSTM to obtain more contextual information. Compared with the original algorithm, the performance of this algorithm is improved on the synthetic dataset.

1. Introduction
In the aspect of deep learning, most of the work boils down to the problem of sequence prediction. At present, there are mainly two text recognition frameworks, one is CRNN framework and the other is attention-based codec framework. Both frameworks are methods to align image feature sequence with label sequence.

However, the CRNN framework method still has room for improvement. In the CTC-based method, the loss of feature details obtained by the feature extraction network is more serious. For example, when recognizing lines of text that contain small or complex characters, some characters are lost in the recognition result or characters that are similar in appearance are indistinguishable. Therefore, this paper uses BiLSTM combined with attention to obtaining more context information to improve the final recognition rate.

2. Materials and Methods
In this paper, the attention mechanism is added in the middle of the two-layer BiLSTM to pay more attention to the important feature sequences to obtain more context information. H_n represents the hidden state of BiLSTM output at the first layer, and h_n is used to generate an attention matrix w, output1 represents the feature sequence of BiLSTM output at the first layer, then we can get

\[ a_y = w \times \text{output1} \]  \hspace{1cm} (1)

\[ b_y = a_y \times a_y \times \text{output1} \]  \hspace{1cm} (2)
where \( a_{ij} \) represents the context vector obtained by attention, \( b_{ij} \) represents the context vector output at the first layer, then \( b_{ij} \) will be sent to the second layer BiLSTM. Its structure is shown in Figure 1. The feature extraction part is the same as the original algorithm without changes. Then the validation was performed on the synthetic dataset. The impact of two-layer BiLSTM and one-layer BiLSTM on the final recognition results compared.

The RCTW dataset consists of 12263 natural scene images containing Chinese, most of which are captured images and a few are synthetic images. The data set is divided into three parts: training set, verification set and test set. The images used in this paper are 10,000 images selected from the images cut out based on the annotation information.

The synthetic dataset in this paper, using the TextRecognitionDataGenerator method. The background images are searched from the Internet, including wood background, iron background, wall background and glass background, etc.; in the selection of fonts, common script, Song and italics are used; the synthesized text is from the domestic news and international news crawled from Sina.com, which is folded into one line of every 8 characters. Also, blurring and skewing were added. 300,000 samples and their detailed text labels were synthesized using the above method. The size of each sample is fixed at 1000 pixels wide and 128 pixels high, and the label length is fixed at 8 characters. In the annotation format, the image name is the label corresponding to the image. Figure 2 shows some samples from the synthetic data. These data are divided into training set, validation set, and test set by 8:1:1, the training set contains 240,000 samples, the validation set contains 30,000 samples and the test set contains 30,000 samples.

3. Results & Discussion
For the benchmark algorithm selection, the CRNN network proposed by [3] is used in this work. We use the line recognition accuracy rate as an evaluation metric, which is defined as follows:
\[
\text{Accuracy} = \frac{A}{B}
\]  
(3)

where \(A\) represents the number of images where all characters are correctly recognized, \(B\) represents the number of all images.

The CRNN was trained on the training set of the synthetic dataset for five rounds, and the obtained model had a perfect pair rate of 0.5820 on the test set. Some of the test results are shown in Table 1.

**Tab. 1** Partial results of CRNN on synthetic dataset

| Test samples | Labeling | Baseline algorithm output |
|--------------|----------|---------------------------|
|              | Stakeholder advice to ensure | The views of the parties concerned to ensure |
|              | The most effective antimalarial drugs | The most effective anti-syndromic drugs |
|              | Not only does the United States have nearly half of 70% for landlords | The United States not only has nearly even numbers For the sick set 4 |
|              | Sensitive submarine communication cables | Sense-saving submarine communication cable |
|              | [5] 0-15 years old | 1510-15 years old |

As shown in Table 2, some test results of the CRNN trained model on the synthetic dataset, on RCTW are shown. As you can see, the direct test becomes less effective.

**Tab. 2** Partial results of CRNN model on RCTW

| Test samples | Labeling | Baseline algorithm output |
|--------------|----------|---------------------------|
|              | Convenient fast food service points | Star two volume true. |
|              | Wuhan Second Ring Road Wuchang section | Wufang Er Luo". words |
|              | Ten years of maintenance experience | Ten years of experience in Vikram |
|              | I do not like people say I am handsome | I don't like what people say about the class lung |
|              | Client Download Address | Muroto Shinko download ground plan |
|              | New Hydrating Essence | New Hydrating Intense Essence Water |

Table 3 shows the comparison between CRNN and the algorithm in this chapter on the synthetic data set. It can be seen that the algorithm in this chapter improves the accuracy and verifies the validity of the proposed algorithm.

**Tab. 3** Accuracy of CRNN with the algorithms in this chapter on synthetic data sets

| Method | Perfect match rate |
|--------|--------------------|
| CRNN   | 0.5820             |
| Algorithm of this chapter | 0.5837 |

As shown in Table 4, the results of some CRNN with the algorithms in this chapter on the synthetic dataset are demonstrated.

**Tab. 4** Partial demonstration of CRNN and the algorithms in this chapter on a synthetic dataset

| Test samples | CRNN | Algorithm of this chapter |
|--------------|------|---------------------------|
|              | Batch of undergraduate arts and history 4 | Batch of undergraduate arts and history 4 |
Both Fish Island and its annexed islands, construction of the construction institute, wine industry Positive primary end of the test;6 More than still in Canada "across Ongoing trust issues.

Fish Island and its annexed islands Building construction, wine industry Positive primary screening test;6 More than still in Canada "across Ongoing trust issues.

As shown in Table 5, the impact of two-layer BiLSTM and one-layer BiLSTM on the final recognition results on the synthetic dataset is demonstrated.

**Tab. 5 Accuracy of different layers of BiLSTM on synthetic datasets**

| Method                  | Perfect match rate |
|-------------------------|--------------------|
| CRNN(Two-layer BiLSTM)  | 0.5820             |
| CRNN(One-layer BiLSTM)  | 0.5663             |

As shown in Table 6, which shows the results of the models trained by different layers of BiLSTM on synthetic datasets tested directly on RCTW, it can be seen that there is a significant decrease in accuracy, which is conjectured to be because the synthetic data is far less complex than the real data.

**Tab. 6 Partial results of CRNN model on RCTW.**

| Method                  | Perfect match rate |
|-------------------------|--------------------|
| CRNN(Two-layer BiLSTM)  | 0.0024             |
| CRNN(One-layer BiLSTM)  | 0.0042             |

4. Conclusion

Because of the deficiency of Chinese natural scene recognition, the CRNN model is improved. In the sequence modeling phase, attention is introduced to obtain more context information. Finally, the CRNN algorithm is compared with the algorithm proposed in this chapter in the synthetic data set, and the effectiveness of the improved algorithm is verified. However, this method increases the complexity of the model and requires more training time. At the same time, comparing the two-layer BiLSTM with the single-layer BiLSTM, we can find that the final accuracy of the latter is slightly lower, which may be due to the problem of multi-classification, while the single-layer BiLSTM can not classify better.

Acknowledgments

This work is supported by the National Key Research and Development Program of China (No. 2017YFB1402200) and the National Natural Science Foundation of China (No.61262086) as well as the Construction of national team and Key Laboratory of computer and Tibetan information technology ([2018] 81) issued by Department of education of Tibet Autonomous Region.

References

[1] Tong G, Li Y, Gao H, Chen H and Yang X 2020 Ma-crnn: a multi-scale attention crnn for chinese text line recognition in natural scenes J. Document Analysis and Recognition 23 pp 103–114

[2] Hu W, Cai X, Hou J, Yi S and Lin Z 2020 Gtc: guided training of ctc towards efficient and accurate scene text recognition Proceedings of the AAAI Conference on Artificial Intelligence 34(7) pp 11005-11012

[3] Shi B, Xiang B and Cong Y 2016 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) pp 2298-2304

[4] Yan R, Peng L, Xiao S and Yao G 2021 Primitive Representation Learning for Scene Text Recognition CVPR

[5] Lee C Y and Osindero S 2016 Recursive Recurrent Nets with Attention Modeling for OCR in the Wild IEEE Conference on Computer Vision & Pattern Recognition Las Vegas USA vol 1 pp 2231-2239

[6] Qiao Z, Zhou Y, Yang D, Zhou Y and Wang W 2020 SEED: Semantics Enhanced Encoder-Decoder Framework for Scene Text Recognition IEEE/CVF Conference on Computer Vision and Pattern Recognition Seattle USA vol 1 pp 13525-13534