**Scholars Journal of Engineering and Technology**
Abbreviated Key Title: Sch J Eng Tech
ISSN 2347-9523 (Print) | ISSN 2321-435X (Online)
Journal homepage: https://saspublishers.com

OPEN ACCESS

# Research of Natural Scene Text Recognition Algorithm Based on OCR

Yingchun Zhang[1*]

[1]College of Information Engineering, Shenyang University of Chemical Technology, Shenyang, China

**DOI:** 10.36347/sjet.2021.v09i10.002 | **Received:** 29.10.2021 | **Accepted:** 07.11.2021 | **Published:** 12.11.2021

***Corresponding author:** Yingchun Zhang

**Abstract** | **Original Research Article**

Aiming at the problem of low recognition accuracy of text information in natural scene images, this paper adopts an end-to-end deep learning text recognition algorithm (CRNN) for image OCR recognition. The CRNN algorithm is mainly composed of CNN, RNN, and CTC algorithm. Among them, CNN uses the improved VGG model to extract the sequence feature of the text line. In order to eliminate the gradient dispersion problem in the training process of RNN and strengthen the semantic information of the context, BLSTM is used to replace the RNN model for label prediction, and then the CTC algorithm is used to complete the transcription and output the final recognition result. Experimental results show that the improved CRNN text recognition algorithm has an accuracy rate of 96.6%, which is 1% higher than the basic CRNN text recognition algorithm, and this end-to-end network structure design also greatly shortens the text recognition time.

**Keywords:** OCR recognition, CNN, RNN, CTC algorithm, VGG model, BLSTM model, CRNN.



## 1. INTRODUCTION

With the continuous development of artificial intelligence, OCR [1] (Optical Character Recognition) technology will still be one of the hot research topics in the field of computer vision. Traditional text recognition algorithms can only recognize some texts with simple background and easy character segmentation [2, 3]. Obviously, the OCR recognition technology at this time has certain limitations. However, with the explosive development of deep learning, the continuous update of convolutional neural network CNN [4, 5] and recurrent neural network RNN [6] has evolved a series of end-to-end sequence recognition models that do not require character Segmentation can directly identify sequence texts of variable length. The CRNN [7] (Convolutional Recurrent Neural Network) used in this article is a typical end-to-end sequence recognition model, which is mainly composed of CNN, RNN, and CTC [8] (Connectionist Temporal Classification) algorithm. CNN uses the classic VGG [9] (Visual Geometry Group) model to extract text image features, because the VGG model uses a large number of 3×3 small convolution kernels instead of large convolution kernels, these operations can ensure that the training parameters of the entire model are reduced under the condition that the receptive field remains unchanged, which can improve the performance of the model. RNN uses the BLSTM [10] (Bidirectional Long-Short Term Memory) model for contextual semantic analysis. At the same time, BLSTM is an upgraded version of LSTM [11] and solves the gradient dispersion problem in RNN. The CTC algorithm decodes the output of the BLSTM by finding the optimal path, completes the conversion from the probability label to the character label, and adds a placeholder mechanism to deal with the problem of repeated character output, so as to realize end-to-end recognition. The model structure diagram of CRNN is shown in Figure 1.

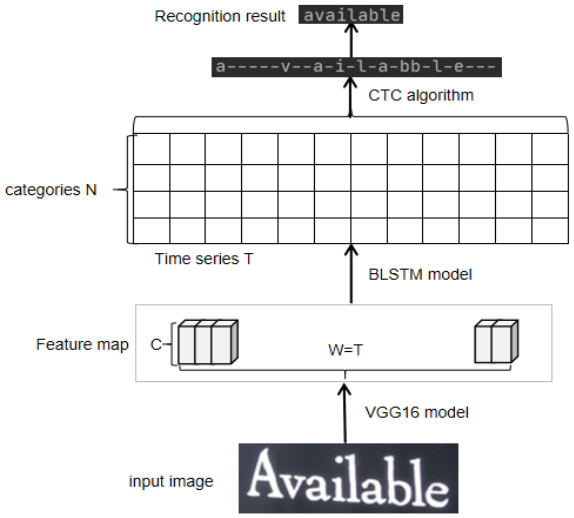


**Figure 1: CRNN model structure diagram**

## 2. EXPERIMENTAL SECTION

### 2.1 Text Recognition Algorithm

Text Recognition Algorithm of CRNN used in this article is introduced from three aspects: text feature extraction, sequence feature prediction, and output transcription. Firstly, In the CNN part of the CRNN model, the VGG model is used to extract the basic features of the text line. When extracting features in CNN, it is necessary to ensure that the input is a grayscale image and the height of the image is 32. Because the text image is usually narrow in height and long in width, the original paper modified the 3th and 4th rectangular windows of the Max Pooling layer (Max Pooling layer) in the VGG network from 2×2 to 1×2. This operation is convenient to use the output features of the CNN model as the input of the RNN model. At the same time, the BN [12] (Batch Normalization) layer is added after the 5th and 6th convolutional layers in the original paper, and ReLU is used as the activation function. The improved VGG model still uses 7 convolution layers to ensure that the spatial features of the image can be extracted by convolution. At the same time, BN layer is added after each convolution layer to prevent gradient disappearance and gradient explosion. Secondly, in order to eliminate the gradient dispersion problem in the training process of RNN and strengthen the semantic information of the context, BLSTM is used instead of the RNN model. In this paper, the sequence features of CNN are input into two BLSTM models for sequence label prediction, and preliminary label probability distribution results are obtained. Thirdly, it can be converted into character tags by further decoding with the help of CTC algorithm. The improved CRNN mode structure is shown in Table 1.

**Table 1: Improved CRNN mode structure**

| Type | Configuration |
|---|---|
| input | w×32 gray-scale image |
| conv1 | #maps:64, k:3×3, s:1, p:1 |
| BN1 | - |
| maxp1 | window:2×2, s:2 |
| conv2 | #maps:128, k:3×3, s:1, p:1 |
| BN2 | - |
| maxp2 | window:2×2, s:2 |
| conv3 | #maps:256, k:3×3, s:1, p:1 |
| BN3 | - |
| conv4 | #maps:256, k:3×3, s:1, p:1 |
| BN4 | - |
| maxp3 | window:2×2, s:1×2, p:1×0 |
| conv5 | #maps:512, k:3×3, s:1, p:1 |
| BN5 | - |
| conv6 | #maps:512, k:3×3, s:1, p:1 |
| BN6 | - |
| maxp4 | Window:2×2, s:1×2, p:1×0 |
| conv7 | #maps:512, k:2×2, s:1, p:0 |
| BN7 | - |
| Map-to-Sequence | - |
| BLSTM | #hidden units:256 |
| BLSTM | #hidden units:256 |
| Transcription | - |

Table 1 model configuration summary. ‘k’, ‘s’ and ‘p’ stand for kernel size, stride and padding size respectively, ‘w’ represents the width of input image.

### 2.2 Data Set Introduction

In this experiment, the public Chinese data set is used, which contains more than 3.64 million pictures, 99% of the pictures are used as the training set, and the remaining pictures are used as the verification set. Then make a corpus, which consists of Chinese characters, letters, numbers, punctuation marks and other characters, a total of 5990 characters. Each picture in the data set contains only 10 characters, and these characters are randomly selected from the corpus. When calibrating each picture in the data set, you only need to find the character index of the 10 characters contained in the picture in the corpus and mark it as the label content, that is, the label is marked with 10 numbers, not the corresponding picture character of. During training, the text and corpus on the picture are encoded, and this encoding format is stored in the form of a dictionary. In the prediction, the predicted number index is decoded through the constructed corpus to obtain the predicted character.

### 2.3 Model Training Method

The maximum number of training iterations of the model is set to 10 rounds, each iteration of 32 images, the initial learning rate is set to 0.001, each time a new round of training is started, the learning rate

is reduced to the original 10%, and the Adam optimizer is used to update the parameters of the CTC loss function. Finally, during model training, the loss is recorded every 100 batches; When testing on the validation set, because batch training (32 pictures) is used, a total of 102488 batches are iterated in each round, then the total iterative batches are divided into 10 equal parts, and a validation test is performed on each equal part (10240 batches).

## 3. RESULTS AND DISCUSSIONS

In this experiment, the basic CRNN and the improved CRNN text recognition algorithm were used to complete the verification set test.

1. Comparative analysis of training set loss

During training, the current loss is recorded every 100 batches, and the loss curve is shown in Figure 2.

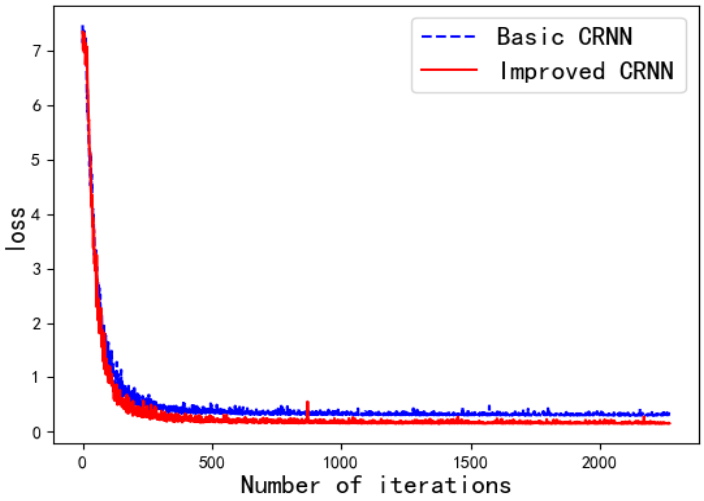


**Figure 2: Loss curve on the training set**

It can be seen from Figure 2 that the loss curve of the two algorithms is not obvious before the number of iterations is 100, but the loss value drops very fast; when the number of iterations is between 100 and 550, the two difference algorithms begins have occurred Changes. Among them, the improved CRNN algorithm is significantly faster than the loss curve of the basic CRNN algorithm. The two algorithms gradually stabilized after 550 iterations. The improved CRNN algorithm has a smaller loss value than the basic CRNN algorithm, which also shows that the performance of the improved CRNN algorithm is better than that of the basic CRNN algorithm.

2. Comparative analysis of validation set loss and accuracy

In the training process, the training batch is divided into 10 equal parts, and a verification set test is performed on each of the equal parts. The loss curve of the verification set is shown in Figure 3, and the accuracy curve of the verification set is shown in Figure 4.

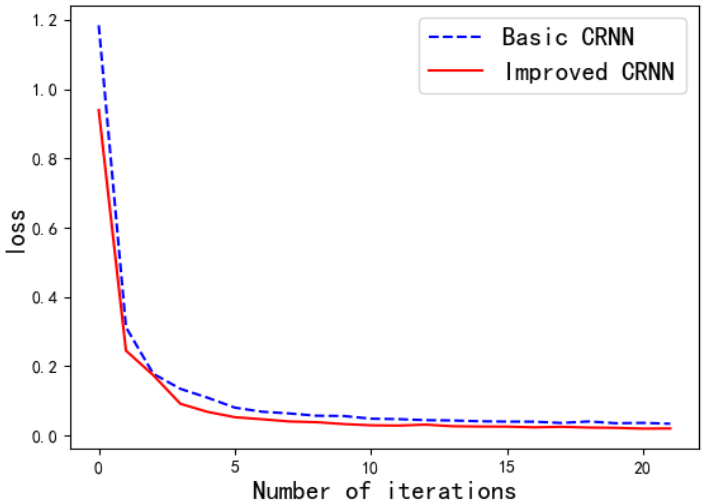


**Figure 3: Loss curve on the verification set**

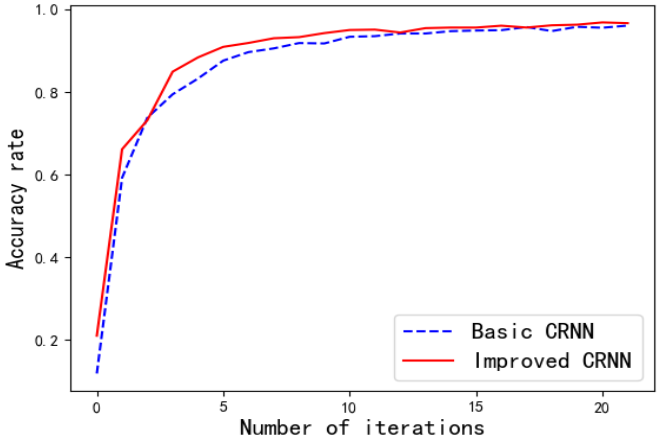


**Figure 4: Accuracy curve on the validation set**

From Figure 2 and Figure 3, it can be found that the change trend of the loss curve on the validation set and the training set is almost the same, which is reflected in the robust generalization ability of the basic CRNN and the improved CRNN network structure.

It can be found from Figure 4 that the upward trend of accuracy curve can be divided into three stages. In the first stage (the first two rounds), the accuracy of the two algorithms increases rapidly; In the second stage (from the third round to the sixth round), the accuracy of the two algorithms increases slowly; In the third stage (after 6 rounds), the accuracy of the two algorithms finally approaches a stable value. At this time, the accuracy of the basic crnn algorithm is as high as 96.0%, and the accuracy of the improved crnn algorithm is as high as 96.6%, which is 1% higher than that of the basic crnn algorithm.

It can be seen from Figure 3 and Figure 4 that the accuracy increases with the decrease of loss, and the change rate of loss curve and accuracy curve is also consistent.

## 4. CONCLUSION

It can be seen from the above analysis that under the condition of constant model size, the same conclusion can be drawn from both loss value and accuracy: the improved CRNN algorithm is superior to the basic CRNN algorithm. The improved CRNN text recognition algorithm has an accuracy rate of 96.6% on the verification set, which is higher than the basic CRNN text recognition algorithm by %1.

## REFERENCES

1. Mithe, R., Indalkar, S., & Divekar N. (2013). Optical character recognition. *International journal of recent technology and engineering*, 2(1), 72-75.
2. Hull, J. J., & Srihari, S. N. (1982). Experiments in text recognition with binary n-gram and viterbi algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (5), 520-530.
3. Al-Badr, B., & Mahmoud, S. A. (1995) Survey and bibliography of Arabic optical text recognition. *Signal processing*, 41(1), 49-77.
4. LeCun, Y., Bottou, L., & Bengio, Y. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 2278-2324.
5. Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 1097-1105.
6. Bengio, Y., Simard, P., & Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, 5(2), 157-166.
7. Shi, B., Xiang, B., & 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), 2298-2304.
8. Graves, A. (2012). Connectionist temporal classification Supervised Sequence Labelling with Recurrent Neural Networks. Springer, Berlin, Heidelberg, 61-93.
9. Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409, 1556.
10. Graves, A., & Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural networks*, 18(5-6), 602-610.
11. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8), 1735-1780.
12. Anwani, N., & Rajendran, B. (2015). Normad-normalized approximate descent based supervised learning rule for spiking neurons. In 2015 international joint conference on neural networks. IEEE, 1-8.