Multimedia Tools and Applications (2022) 81:10545–10559
https://doi.org/10.1007/s11042-022-12024-w

Text recognition in natural scenes based on deep learning

Yi Jiang[1] · Zhongyu Jiang[2] · Liang He[3] · Shuai Chen[2]

Received: 11 February 2021 / Revised: 16 April 2021 / Accepted: 3 January 2022 /
Published online: 16 February 2022

Abstract
Aiming at the problems of character segmentation and dictionary dependence in text recognition in natural scenes, a text recognition algorithm based on Attention mechanism and connection time classification (CTC) loss is proposed. Convolutional neural network and bidirectional long short – term memory network are used to realize image feature coding, which avoids the gradient vanishing problem of recurrent neural network (RNN) with the increase of time. And the Attention-CTC structure is used to decode the feature sequence, which effectively solves the problem of unconstrained attention decoding. The algorithm avoids extra processing of alignment and subsequent syntax processing, and improves the speed of training convergence and significantly improves the recognition rate of text. It has a certain research value in recognition accuracy. Experimental results show that the algorithm has good robustness to text images with fuzzy fonts and complex background.

Keywords Text recognition · Convolution neural network · Attention mechanism · Connection time classification · Long short – Term memory network

✉ Zhongyu Jiang
1179544217@qq.com

Yi Jiang
jasonj@hrbust.edu.cn

[1] Department of Communications Engineering, Harbin University of Science and Technology, 52 Xuefu Road, Nangang District, Harbin, Heilongjiang, China

[2] School of Automation, Harbin University of Science and Technology, Harbin, China

[3] School of Software, Northwestern Polytechnical University, Xi'an, China

1 Introduction

With the rapid development of internet technology and computer science, a large number of images spread in the network, and the explosive amount of information makes it more difficult for people to filter content and extract information, especially when the image contains complex and diverse text information, it is often time-consuming to obtain and identify text. When processing and analyzing text images, some artificial signs and regularized objects in images are usually which we pay attention to. Efficient analysis, acquisition and understanding of these marks is conducive to improving the level of industrial automation, and is of great significance to the field of media retrieval and intelligent analysis. The improvement of hardware computing ability indicates that people can train deep network to extract target features. Neural network can automatically learn and combine features without artificial construction of complex features. Therefore, deep learning has become the mainstream method in classification and detection tasks. There are two main methods of deep learning in the detection task. One is based on the proposed region, which estimates the target position first, then uses regression to locate the target, and the other is based on the regression method to directly detect the target in the image.

The method based on region proposal is to use the proposed algorithm to generate a possible text proposal box, then judge whether the proposal box contains the target, and finally determine the target category. Region convolutional neural network(RCNN) [12] transforms the classification problem into detection problem, which is a pioneering work in the field of target detection. RCNN first uses random search algorithm to get the candidate region of the target, then inputs the candidate region into the neural network to extract the object features, and finally inputs the extracted features into the support vector machine for classification. Compared with the traditional detection methods, RCNN has achieved great success, but the training time of this method is large and the calculation is large. Faster R-CNN [25] uses convolutional neural network to extract the features of the whole image, and uses RPN network to obtain the proposed region. This method combines detection and classification into one network, realizing end-to-end training, and greatly improves detection speed and accuracy. Tian et al. [29] proposed a method based on feature sequence detection of Faster R-CNN. Firstly, the convolution neural network is used to extract image features, and then PRN network is used to slide on the feature map to get the feature vector corresponding to each feature point. Then, a bidirectional cyclic neural network is used to encode the features. Finally, the encoded feature sequence is input into the full connection layer to predict the coordinate position and fraction of characters in the image. This method has very high recognition accuracy and generalization ability for different scale and shape characters, but it is slow because of the need for final text line stitching for each text area. The detection method based on regression does not need the step of region proposal, but directly regresses the target position. Therefore, the detection speed is greatly improved, but the detection accuracy is lower than the proposed region method. YOLO [24] uses a sisngle convolutional neural network to transform the problem of target detection into a regression problem that directly extracts the regression box and category probability from the image. Because YOLO divides the image into several grids, and each grid is only responsible for detecting one target, YOLO's performance is not good relative to small targets. Single Shot MultiBox Detector [21] (SSD) of single neural network follows the method of direct regression of target frame and classification probability in YOLO, and refers to anchor frame mechanism of Faster R-CNN to improve the detection accuracy. By combining the two structures, SSD can maintain

the accuracy and also has high detection speed. Bai et al. [4] proposed a fast and accurate text detector based on SSD. According to the characteristics of the text, the anchor frame with large aspect ratio is designed, and the convolution kernel of 3 × 3 is changed to 1 × 5, so that the network can produce rectangular receptive field and adapt to the text with large aspect ratio. This method can adapt to different scales of text, fast detection speed and high accuracy. In the development of its methods and related applications, Chen et al. [5] designed a novel hybrid deep learning method that combines a modified Generative Adversarial Network (GAN) and a CNN-based detection approach is proposed for small ship detection. Experimental results show that the proposed deep learning method is competent to generate sufficient informative small ship samples and can obtain significantly improved and robust results of small ship detection.

In recognition method, Attention Mechanism (AM) is proposed to solve the problem of long text information being covered in decoding. Wang et al. [6] showed that by introducing attention mechanism into the process of learning sample features, it can obtain new features with good robustness. Connection time classification is mainly used to solve the sequence alignment task, which directly outputs the predicted character sequence without considering the alignment problem between the input data and the given label. In the field of recognition, CTC is currently applied to semantic recognition. The research results of Chen et al. [31] showed that the connection time classifier can effectively improve the accuracy of semantic recognition.

In the related comprehensive application, Sitalakshmi et al. [26] designed a novel visualisation using similarity matrix method for establishing malware classification accurately. Danish et al. [9] proposed a new method to solve the problem of multi class classification. They compared the performance of IMCFN algorithm with existing malware classification study, which used image malware classification techniques based on machine and deep learning methods. Hakak et al. [14] proposed an ensemble classification model for detection of the fake news that has achieved a better accuracy compared to the state-of-the-art. Alazab et al. [1] proposed to predict the stability of the smart grid network. And the proposed model is experimented on the smart grid dataset from UCI Machine Learning Repository. The comparative analysis proves the superiority of the proposed model with respect to accuracy, precision, loss and ROC curve metrics. Chen et al. [7] proposed an information entropy constrained sparse representation model is developed for pedestrian behavior understanding. It aims to reduce the entropy for trajectory representation and to obtain superior recognition results. Ganesh Jha & Hubert Cecotti [11] considered GAN—a technique that does not require prior knowledge of the possible variabilities that exist across examples to create novel artificial examples and aimed at enriching databases of images or signals for improving the classifier performance by designing a GAN for creating artificial images.

At present, the main challenges of text recognition algorithms are as follows: (1) text features rely on manual definition. Manually defined features are difficult to capture the deep semantic of the image, which is time-consuming and not universal; (2) recognition based on single character [35]. Single character recognition will be divorced from the context, which is easy to cause ambiguity; (3) it is necessary to segment and locate characters [32]. Text structure is complex and semantic is changeable, forced segmentation will destroy character structure; (4) over reliance on classification dictionary. The selection of classification dictionary directly affects the recognition results, resulting in poor generalization ability of recognition model.

In this paper, a novel text recognition algorithm based on AM [2, 3, 22, 23] and CTC [13, 17] is proposed to solve the problems of text segmentation difficulty and dictionary dependence in text recognition. The algorithm consists of multi-scale feature extraction layer [36], Long Short – Term Memory (LSTM) [15] coding layer and Attention – CTC decoding layer. In this paper, the neural network framework of deep learning technology is successfully applied to the problem of text recognition. Aiming at the main problems in this field, a text recognition method based on deep learning is proposed, and its validity is verified on a variety of data sets.

2 The proposed method

This paper proposes a text recognition algorithm for natural scenes based on Attention - CTC, which fully considers the local optimal effect of each functional module and the final overall recognition effect. Firstly, the original data set is preprocessed, ESPCN is an efficient method to directly extract features from low resolution image size and calculate high resolution image [34]. After input to the network, multi-scale convolution neural network model is used for feature extraction. Then the feature map is input into LSTM network for coding. Finally, Attention - CTC is used to decode the encoded feature sequence. The overall network design is shown in Fig. 1.

The algorithm has the following characteristics: (1) After high-resolution image is obtained by ESPCN method, multi-scale features of image text are extracted by convolution neural network based on Inception V3 [10, 18, 19, 27, 28, 30]; (2) Encoding the feature sequence with LSTM and learning the character level language features implicitly, thus avoiding the use of N-ary grammar model for character post-processing; (3) In this paper, we use the structure of Attention and CTC to decode feature sequences, in which Attention is used to decode feature sequences, and CTC implements the constraint on Attention.

2.1 LSTM specific coding

RNN has unique advantages in dealing with text sequence problems, but with the increase of time interval, RNN will appear gradient disappearance phenomenon, which makes it unable to deal with long-distance dependence problem. LSTM can selectively influence the state of RNN at each time by using special gate structure, so as to avoid the disappearance of gradient. LSTM is directional and uses only the past context. However, in image-based sequences, the context of the two directions is useful and complementary to each other. Therefore, in this paper, Bidirectional Long Short – Term Memory (BLSTM) network is used to encode the image features in both directions, as shown in Fig. 2.

Among them, $x = \{x_1, x_2, \ldots, x_n\}$ is the input, that is, the feature image sequence after convolution; $h = \{h_1, h_2, \ldots, h_n\}$ is the hidden layer of forward LSTM;$h' = \{h'_1, h'_2, \ldots, h'_n\}$is the hidden layer of reverse LSTM;$y = \{y_1, y_2, \ldots, y_n\}$is the predicted output; SOS is the starting character; EOS is the terminator.

2.2 Attention - CTC decoding

Decoding is the process of converting image feature sequence into character sequence. Given the input feature sequence $X = \{x_1, x_2, \ldots, x_w\}$, the hidden variable$Z = \{z_t \in D \cup blank | t =$

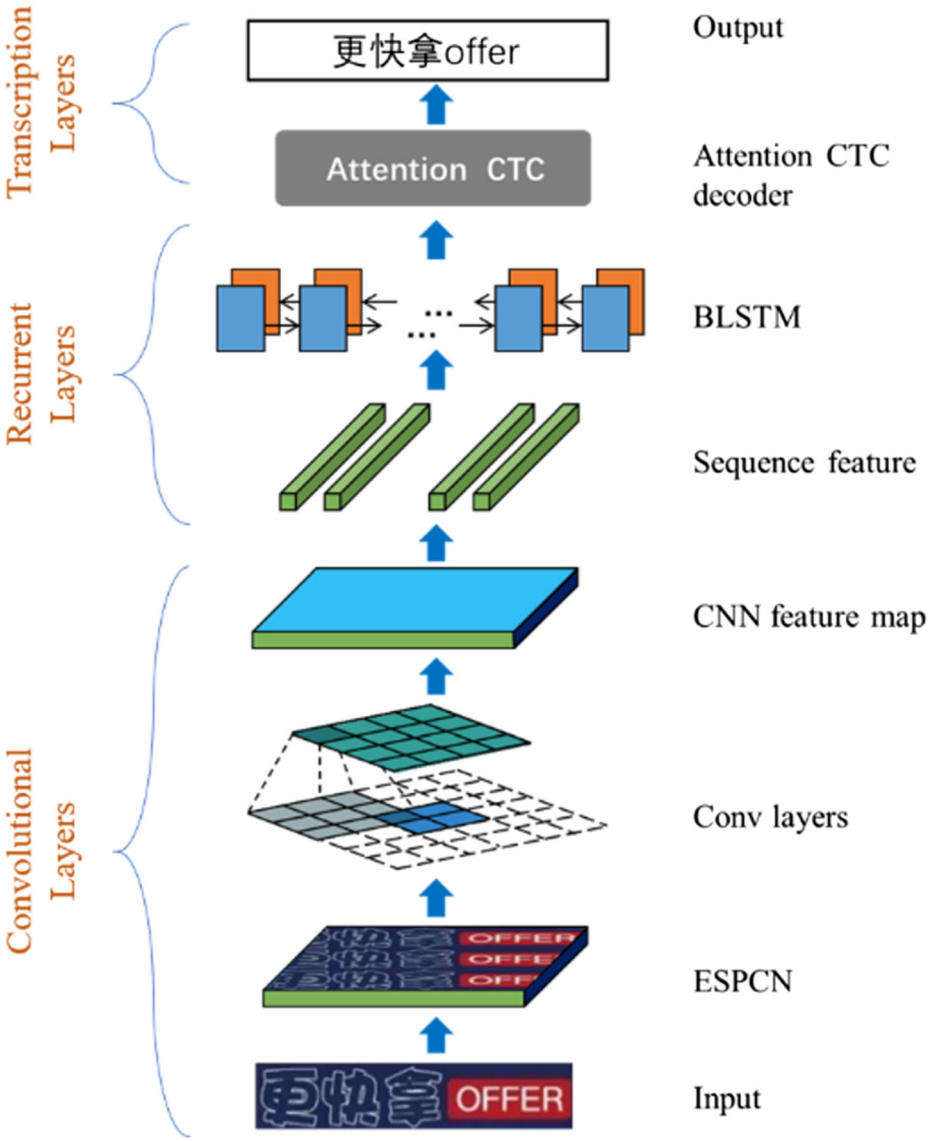

Fig. 1 Network structure

1, …, w} and the output length L sequence $Y = \{y_t \in D | l = 1, \ldots, T\}$. Where W is the number of encoding sequences; T is the number of characters; D is the dictionary containing all characters [8]. CTC assumes that the labels are independent, and the posterior probability distribution of prediction series is calculated by Bayesian theorem

$$p_{ctc}(Y|X) \approx \sum_{z} \prod_{t} p(z_t|z_{t-1}, Y) p(z_t|X) \tag{1}$$

Where $p(Y)$ is the character level language model; $p(z_t|\ X)$ is the probability of hidden variables obtained from known input features; and $p(z_t|\ z_{t-1},\ Y)$ is the conditional probability that the hidden

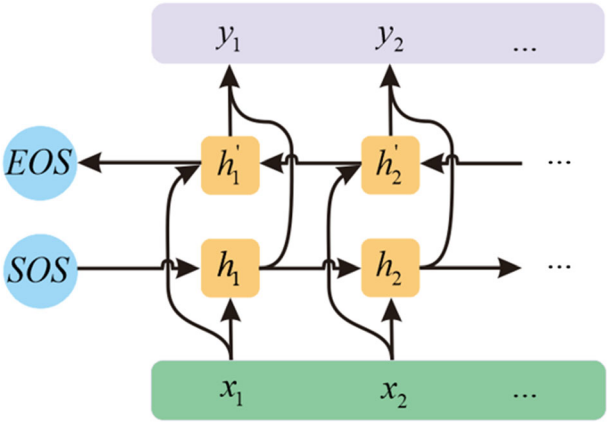

Fig. 2 Bidirectional long short-term memory network

variables output at the previous time can predict the hidden variables at the next time. CTC algorithm assumes that the inner part of tags is conditional independent, and each output is the probability of a single character, which results in CTC only for local information prediction, ignoring the overall information, so it can not effectively predict long text sequences.

Compared with the local prediction of CTC, Attention mechanism can directly predict the text sequence without calculating the hidden variables and making the assumption that the label is independent of each other

$$p_{attn}(Y|X) = \prod_{l} p(y_l|y_{1:l-1}, X) \tag{2}$$

Where $p(y_l|\ y_{1\ :\ l\ -\ 1},\ X)$ is the prediction probability of l time when the input characteristic X and the first output are known.

The simple attention mechanism does not introduce any constraints of guiding alignment, which results in the problem of noise sensitivity and dislocation in decoding. Therefore, this paper designs a multi task learning decoder [16, 20, 33], based on the joint training of Attention and CTC. It uses Attention to decode the character level semantics, and uses CTC to realize the constraint of Attention decoding. The decoding algorithm based on Attention-CTC not only effectively solves the problem that the pure data-driven method is difficult to train for long sequence input, but also can fully extract the information of long characters. The Attention-CTC decoding framework is shown in Fig. 3.

X is the characteristic of the input coding network; h is the high level features; r is the attention weight vector; q is the semantic vector of decoding; z is the hidden variable; y is the prediction output. The shared encoder is trained by both CTC and attention model objectives simultaneously. The shared encoder transforms our input sequence, $\{x_1 \cdots x_T\}$, into the high level features, $H = \{h_1 \cdots h_T\}$, and the attention decoder generates the letter sequence, $\{y_1 \cdots y_L\}$.

Among them, CTC and Attention model share the coding network, and the maximum joint probability of CTC and Attention prediction can be expressed as

$$\hat{Y} = Y \in Dargmax\{\lambda log p_{ctc}(Y|X)\} + (1-\lambda)p_{attn}(Y|X) \tag{3}$$

Then, the maximum joint probability is transformed into a minimum multi task loss function, which can be directly solved by gradient descent training

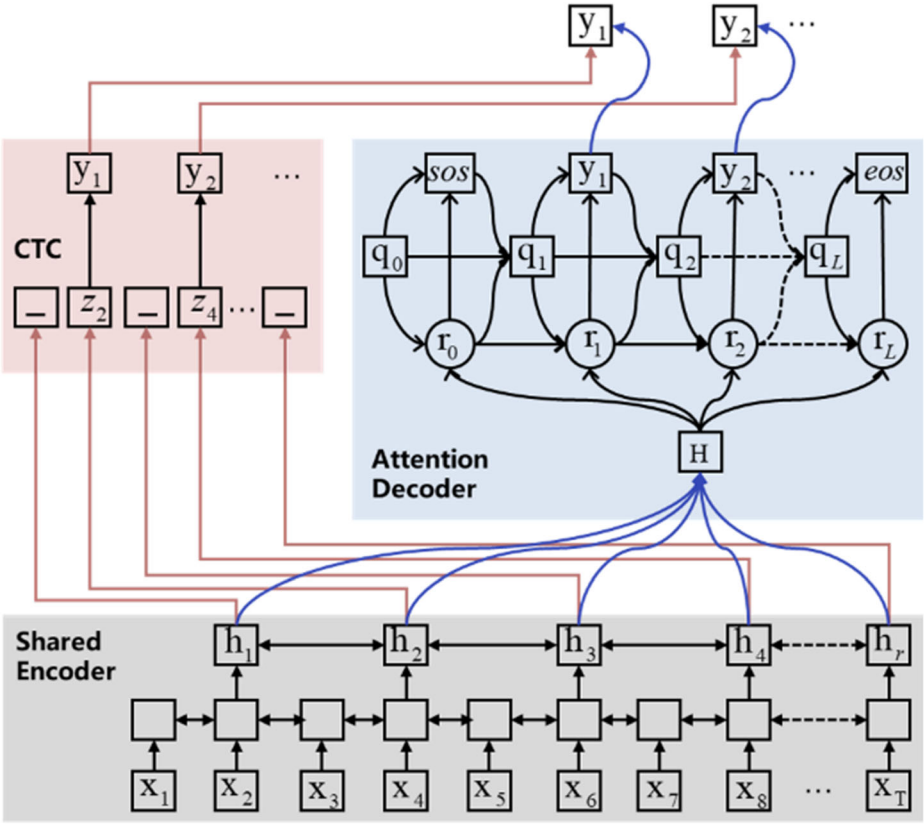

Fig. 3 Decoding structure base Attention - CTC

$$L = \lambda L_{ctc} + (1 - \lambda) L_{attn} \tag{4}$$

Among them, L_{ctc} is the loss function of CTC, L_{attn} is the loss function of attention model; the range of variable parameter λ is $0 \leq \lambda \leq 1$.

What needs to be emphasized in this chapter is:

(1) In the algorithm research, the RNN will disappear with the increase of time interval, which makes it unable to deal with the long-distance dependence problem. LSTM feature coding is introduced to selectively affect the state of RNN at each time through special gate structure, so as to avoid the disappearance of ladder. Because the context of text image processing is mutually useful and complementary in two directions, BLSTM is used to encode the image features in two directions, which effectively improves the robustness of the detection results.

(2) The feature sequence is decoded by the structure of Attention and CTC. Attention is used to decode feature sequences. CTC implements the constraint on Attention, which effectively solves the problem of unconstrained Attention decoding.

Table 1 Hardware configuration

Computer model	Z7-KP7EC
Processor	Intel Core i7-8750HQ @ 2.20GHz (6 cores 12 threads)
Memory	16G DDR4 2666 MHz
GPU graphics card	NVIDIA GeForce GTX 1060 (6 GB)
Hard disk	256G SSD and 1 TB mechanical hard disk

3 Environment configuration

The hardware configuration used in this project is shown in Table 1, On the processor, 2.20GHz Intel Core i7-8750HQ CPU is selected. The 6-core 12 thread has faster event processing speed. It is equipped with 16G RAM. The NVIDIA GeForce GTX 1060 with 6GB video memory is selected as the GPU graphics card. Deep learning mainly consumes the computing resources of GPU, and GTX 1060 can provide huge computing power support, which can meet most of the requirements of deep model training. In addition, the computer is equipped with 256 SSD, so that the computer performance can fully meet the needs of this project.

The software configuration used in the project is briefly described. The software platform configuration is shown in Table 2. The experiment is carried out on Windows system. Since the calculation of deep learning depends on GPU resources, CUDA and CUDNN should be config to realize the combination of software and hardware. CUDA is a general parallel computing architecture platform developed by NVIDIA, which enables GPU to solve complex computing problems. CUDNN is a GPU acceleration library for deep neural network based on CUDA platform, which emphasizes performance, ease of use and low memory overhead. It realizes high-performance parallel computing by integrating into a higher-level machine learning framework.

This paper selects Python as the experimental development language and Pytorch platform as the development framework. Among them, Python, as an open source object-oriented language, has the characteristics of easy to read and write, concise code logic and so on. It has a large number of third-party support libraries, and can realize a variety of project development. Pytorch is a Python learning library based on torch and a neural network framework open-source by Facebook. Among them, torch is a classical tensor library which operates on multi-dimensional matrix data. It is widely used in machine learning and other mathematics intensive problems.

Table 2 Result of recognition rate

Model	IC2013	IIIT5K	SVT	Synthetic	Combine
CRNN	0.78	0.75	0.75	0.88	0.89
Attention	0.85	**0.84**	0.82	0.93	0.94
0.2	**0.91**	0.81	**0.85**	**0.98**	**0.99**
0.5	0.88	0.8	0.81	0.95	0.96
0.8	0.82	0.78	0.77	0.91	0.91

4 Data sets and evaluation criteria

4.1 Data set

In this paper, four scene text datasets, ICDAR 2013, Street View Text, IIIT5K and Synthetic Chinese String Dataset, are selected for experiments. The details are as follows:

(1) ICDAR 2013 (IC13) dataset is an improved version of ICDAR 2011 (IC11) dataset. It is composed of 462 natural scene pictures marked in English, including 229 training sets and 223 test sets. The sample images are shown in Fig. 4.

(2) SVT (Street View Text) data set is composed of 249 street view images collected by Google company, and 647-word images are cut out from them, mainly commercial sign images, which are widely used in text recognition of natural scenes. Figure 5 is a sample image.

(3) The IIIT5K dataset contains 3000 cropped word test images collected from the Internet, each associated with a 50-word dictionary and a 1000-word dictionary. The sample is shown in Fig. 6.

(4) The Synthetic Chinese String Dataset (hereinafter referred to as the Synthetic data set) uses Chinese corpora, such as news, classical Chinese, etc., to generate a total of 3.6 million pictures randomly through the changes of font, size, grayscale, blur, perspective, stretching, etc., and is divided into training set and verification set according to 99:1. It contains 5990 characters including Chinese characters, English letters, numbers and punctuation marks. Each sample is fixed with 10 characters, and the sentences in the corpus are randomly intercepted, and the image resolution is unified as280 × 32. As shown in Fig. 7.

4.2 Evaluation criterion

This paper uses WER (Word Error Rate) as the evaluation standard of character recognition. In order to make the identified word sequence consistent with the standard word sequence, it is necessary to replace, delete, or insert some words. These transformation times are divided by

Fig. 4 Image sample of ICDAR2013 (IC13) dataset

Fig. 5 Image sample of SVT (Street View Text) dataset

the percentage of the number of words in the standard word sequence, which is called *WER*. The calculation formula is as follows

$$WER = 100 \times \frac{S + D + I}{N}\% \tag{5}$$

Where S is the number of replaced characters, D is the number of deleted characters, I is the number of inserted characters, and N is the number of characters in the label.

5 Model training

Because most of the traditional recognition methods rely on dictionaries and can not be compared with the methods proposed in this paper, Convolutional recurrent neural network and Attention network are selected to compare with the improved recognition method proposed in this paper. The above four datasets are used for model training and validation.

The weight of Attention and CTC in decoding layer is balanced by setting super parameter λ. When $\lambda = 1$ is used, there is no Attention mechanism in the model, only CTC is used to decode, that is, convolutional recurrent neural network model; when $\lambda = 0$, there is no CTC decoding in the network, only the decoding based on Attention mechanism, that is, the model is Attention network; when $0 < \lambda < 1$, the decoding structure is jointly decoded by Attention and CTC, and the value of λ indicates the constraint degree of CTC on Attention. The recognition results were analyzed when the λ values were 0, 0.2, 0.5, 0.8 and 1 respectively.

Figure 8 shows the images of different data sets successfully identified by the above four data sets under the Attention-CTC(ACTC) model, and the prediction tag of successful image recognition is marked at the bottom of each image.

In the first training process, the training set of Synthetic data set and SVT data set are selected as the training set of the experiment, and run five epochs, and the batch size is taken as 100, the highest recognition rate of the verification set is 96.8%. In the formal training process, the pre training model is loaded, the program runs 15 epochs on GPU, and the batch size is 10, the final verification set recognition rate is 98.95%. When λ takes different values, the recognition rate of loss value under different iterations is shown in Fig. 9.

Fig. 6 Image sample of IIIT5K dataset

Fig. 7 Image sample of synthetic dataset

It can be seen that with the increase of iteration times, λ has a positive correlation with the loss function, that is, the greater the value of λ, the faster the convergence speed of loss function. This is because the Attention mechanism needs to calculate the attention weight between the decoding and all the encoded features, and then update the Attention weight. There is no constraint between the features, resulting in the parameter solution space is too large and it is not easy to converge. CTC can update parameters directly in the process of back-propagation. As an Attention decoding constraint, with the increase of λ, the stronger the constraint ability is, that is, the stronger the limitation of solution space range is, which makes the network have faster convergence speed.

The curve of the model recognition rate and the number of iterations with different λ values are shown in Fig. 10. When $\lambda = 0.2$, the text recognition rate of ACTC model is the highest, reaching 98.89%; when $\lambda = 0.5$, the text recognition rate of ACTC model is lower than that of $\lambda = 0.2$; when $\lambda = 0$, the model is Attention OCR, the recognition rate of text is lower than that of $\lambda = 0.5$, but higher than that of $\lambda = 0.8$; when $\lambda = 1$, the model is CRNN, the recognition result is the worst. Therefore, compared with CRNN and Attention OCR based on CTC, the recognition accuracy of ACTC is significantly improved. This is due to the fact that CTC decoding prediction is based on conditional independence, which does not care about the context information of characters, but pays more Attention to the local features of characters. However, there is no monotonous alignment constraint in the decoding of single Attention

Fig. 8 Images of different datasets successfully identified

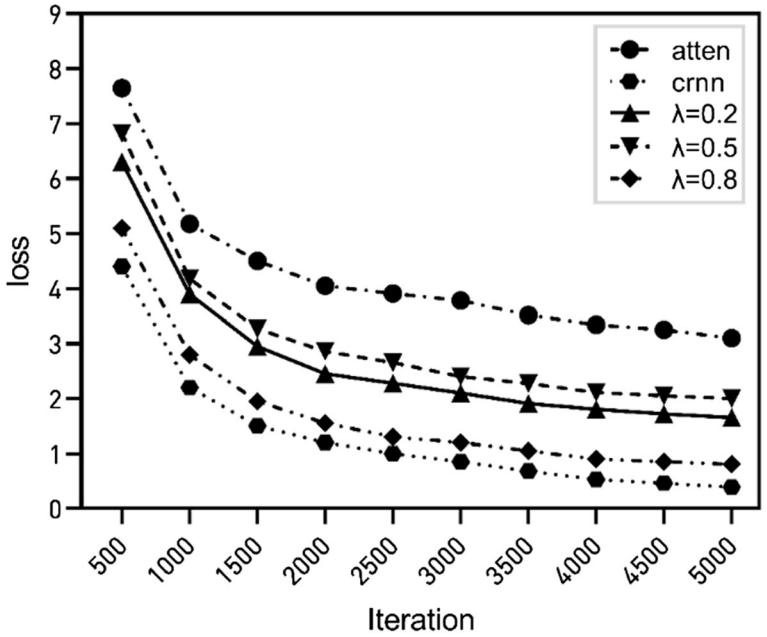

Fig. 9 Loss value under different iterations

mechanism. When dealing with the decoding weight of similar characters, it is easy to cause attention shift and lead to character decoding errors. The ACTC model proposed in this paper uses CTC to restrict the decoding of Attention mechanism. In the process of decoding, it pays

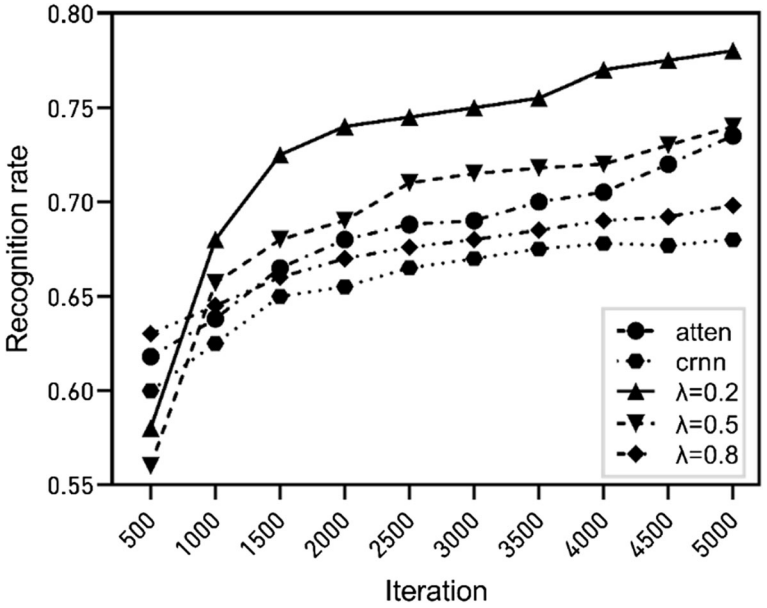

Fig. 10 Relation curve between recognition rate and iteration times

more attention to the character alignment,and it pays attention to the character context information at the same time, which makes the output accuracy of prediction sequence higher.

After that, the four datasets are trained and verified respectively, and the relationship between the loss function and the character recognition rate and the number of iterations is consistent with that in Figs. 9 and 10. The results of recognition rate are shown in Table 2.

It can be seen from the table that the proposed algorithm has high recognition rate on different data sets. Especially, when $\lambda = 0.2$, the corresponding recognition rate of IC2013, SVT and Synthetic is higher than that of CRNN and Attention OCR. This is because CRNN prediction is based on conditional independence and discards the information of the previous time, resulting in decoding only focus on local features. Simple attention decoding has no space constraint. When there are similar characters in the text image, the feature weight of the decoding is offset, which leads to the offset or false detection of the decoded characters. Attention CTC uses CTC to constrain the space of Attention, which makes the decoding pay more attention to the current features, slows down the problem of attention shift, and improves the recognition rate.

6 Conclusion

Text recognition in natural scenes is of great significance for computer to understand images. This paper analyzes the advantages and disadvantages of CTC decoding and Attention decoding, and proposes a joint training method based on Attention and CTC, which avoids the gradient disappearance problem of RNN network with the increase of time, and effectively solves the unconstrained problem of attention decoding, and achieves high recognition rate on different data sets. In the future research planning, there are some areas that need to be improved:

(1) Improving the structure of convolutional neural network to obtain stronger feature extraction ability.
(2) Adding over fitting mechanism to improve the generalization ability of the network.
(3) Need to test more data sets of different languages to expand the application scope of the model.

References

1. Alazab M, Khan S, Krishnan SSR et al (2020) A multidirectional LSTM model for predicting the stability of a smart grid. *IEEE Access* PP(99):1–11
2. Bahdanau D, Cho K, Bengio Y (2015) Neural machine translation by jointly learning to align and translate. International Conference on Learning Representations, San Diego, pp 89–93
3. Bahdanau D, Chorowski J, Serdyuk D, et al. End-to-end attention-based large vocabulary speech recognition. Shanghai: The 41st IEEE International Conference on Acoustics, Speech and Signal Processing, 2016: 4945–4949.
4. Bai X, Shi B, Yao C (2016) An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. IEEE Trans Pattern Anal Mach Intell 39(11):2298–2304
5. Chen ZJ, Chen DP, Zhang YS, Cheng XZ et al (2020) Deep learning for autonomous ship-oriented small ship detection. Saf Sci 130:132–141

6. Chen JN, Gao S, Sun HZ et al (2020) An end-to-end speech recognition algorithm based on attention mechanism. Syst Eng Soc China:6–14
7. Chen ZJ, Cai H, Zhang YS, Wu CZ et al (2020) A novel sparse representation model for pedestrian abnormal trajectory understanding. Expert Syst Appl 144:516–525
8. Chen JN, Gao S, Sun HZ et al (2020) An end-to-end speech recognition algorithm based on attention mechanism. Syst Eng Soc China:640–646
9. Danish V, Alazab M, Sobia W, Hamad N, et al. IMCFN: Image-based malware classification using fine-tuned convolutional neural network architecture. Computer Networks, 2020:171–177.
10. Fernández-Díaz M, Gallardo-Antolín A (2020) An attention long short-term memory based system for automatic classification of speech intelligibility. Eng Appl Artif Intell 96:1–8
11. Ganesh J, Hubert C (2020) Data augmentation for handwritten digit recognition using generative adversarial networks. Multimed Tools Appl 79:35055–35068
12. Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation. 2014 IEEE conference on computer vision and pattern recognition, Columbus, OH, 2014, pp. 580–587.
13. Graves A, Gomez F (2016) Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. International Conference on Machine Learning, Hong Kong, pp 742–748
14. Hakak S, Alazab M, Khan S, … Khan WZ (2021) An ensemble machine learning approach through effective feature extraction to classify fake news. Futur Gener Comput Syst 117:114–123
15. Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780
16. Hori T, Watanabe S, Zhang Y et al (2017) Advances in joint CTC-attention based end-to-end speech recognition with a deep CNN encoder and RNN-LM. IEEE International Conference, USA, pp 1672–1679
17. Huang XH, Qiao LS, Yu WT et al (2020) End-to-end sequence labeling via convolutional recurrent neural network with a connectionist temporal classification layer. Int J Comput Intell Syst 13(1):66–73
18. Ioffe S, Szegedy C (2015) Batch normalization: accelerating deep network training by reducing internal covariate shift. International Conference on Machine Learning, Lille Grand Palais, pp 682–689
19. Jabbari M, Khushaba RN, Nazarpour K (2020) EMG-based hand gesture classification with long short-term memory deep recurrent neural networks. Ann Conf Canadian Med Biol Eng Soc:3302–3305
20. Kim S, Hori T, Watanabe S. Joint CTC-attention based end-to-end speech recognition using multi-task learning. New Orleans: The 42nd IEEE International Conference on Acoustics, Speech and Signal Processing, 2017:798–805.
21. Liu W, Anguelov D, Erhan D, et al. SSD: Single shot multibox detector. European Conference on Computer Vision, 2016:21–37.
22. Luong MT, Pham H, Manning CD (2015) Effective approaches to attention-based neural machine translation. Lisbon: Empirical Methods Nat Language Process:316–325
23. Qu S, Xi Y, Ding S (2017) Visual attention based on long-short term memory model for image caption generation. Melbourne: Control Decis Conf:234–239
24. Redmon J, Divvala S, Girshick R et al (2016) You only look once: unified, real-time object detection. Proc IEEE Conf Comput Vis Pattern Recognit:779–788
25. Ren S, He K, Girshick R, … Sun J (2017 Jun) Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell 39(6):1137–1149
26. Sitalakshmi V, Mamoun A, Qing Y (2018) Use of data visualisation for zero-day malware detection. Security Commun Networks 2018:807–816
27. Szegedy C, Vanhoucke V, Ioffe S et al (2016) Rethinking the inception architecture for computer vision. Computer Vision and Pattern Recognition, Las Vegas, pp 272–281
28. Szegedy C, Ioffe S, Vanhoucke V et al (2017) Inception-v4, inception-resnet and the impact of residual connections on learning. The Thirty-First AAAI Conference on Artificial Intelligence, San Francisco, pp 626–634
29. Tian Z, Huang W, He T, et al. Detecting text in natural image with connectionist text proposal network. Springer, Cham, 2016. LNCS, vol. 9912, pp. 56–72.
30. Tsai ST, Kuo EJ, Tiwary P (2020) Learning molecular dynamics with simple language model built upon long short-term memory neural network. Nat Commun 11(1):1015–1021
31. Wang LL, Wang BQ, Zhao PP et al (2020) Malware detection algorithm based on the attention mechanism and ResNet. Chin J Electron 29(6):473–480
32. Xiong HP, Chen XX, Chen CW (2018) Text location in image based on convolution neural network. Electronic Sci Technol 31(1):51–59
33. Xu K, Li D, Cassimatis N et al (2018) LCANet: end-to-end lipreading with cascaded attention-CTC. Xi'an: China Automatic Face Gesture Recogn:351–360
34. Xu MX, Du XY, Wang DH (2019) Super-resolution restoration of single vehicle image based on ESPCN-VISR model. Adv Sci Industry Res Center: Sci Eng Res Center:517–528

35. Xue HT, Yang JD, Tan KD (2015) Application of an improved BP neural network in handwriting recognition. Electronic Sci Technol 28(5):20–27
36. Yin Z, Tang CH, Zhang XX (2016) Image recognition based on improved sparse auto-encoder. Electronic Sci Technol 29(1):124–127