ISSN 2616-5775 Vol. 6, Issue 2: 44-52, DOI: 10.25236/AJCIS.2023.060207

Study on Deep Learning-Based Natural Scene Text Recognition

Jiashun Weng, Xiaoyun Jia, Yanluo Liu

School of Electronic Information and Artificial Intelligence, Shaanxi University of Science & Technology, Xi'an, China

Abstract: *Aiming at the problem that the natural scene text recognition algorithm pays too much attention to the local character classification content and ignores the global information content of the entire text, a natural scene text recognition algorithm based on multi-network convergence and multi-head attention mechanism is proposed. Firstly, the algorithm uses a multi-network convergence structure to design multiple residual modules to capture contextual features and semantic features in visual features. Then, in the process of character prediction, a multi-head attention mechanism encoder is proposed, which stitches position information, visual features, context features and semantic features into a new feature space. Finally, the new feature space is reweighted by the self-attention mechanism, which improves the accuracy of predicting text information while paying attention to the connection between feature sequences. The recognition accuracy of SVT and ICDAR2015 on the regular and irregular text datasets reached 91.4% and 82.4%, respectively, which improved by about 1.8% and 2.4% compared with the current popular algorithms. Experimental results show that the model can make better use of position features, global semantic features and context features to more accurately identify text content, and improve the accuracy of the model.*

Keywords: *scene text recognition; multi-network convergence; multi-head attention mechanism; feature extraction*

1. Introduction

Text appears in life in various ways, especially in the development of the information age, almost everyone has an electronic product, people widely use the photography and video functions of these electronic products in their daily lives, and a large amount of text is also saved in pictures or videos. Therefore, the use of computer technology to detect and recognize the text content in pictures or videos has become the most urgent need nowadays. On the one hand, recognizing textual contents can improve the productivity of various application scenarios such as document recognition[1], license plate recognition[2], and document analysis[3]; on the other hand, these textual contents can be used to describe scene information and assist other research related to computer vision, such as image and video retrieval, visual target tracking, and video content analysis[4].

Natural scene text recognition is the recognition of text content in scene images, and there are two main types of common natural scene text recognition algorithms: segmentation-based and segmentation-free natural scene text recognition algorithms.

The segmentation-based text recognition algorithm[5-7] generally includes three steps: preprocessing, segmentation of characters, and recognition of characters. Moreover, the search time of the algorithm is related to the size of the character set, which will increase with the addition of a large number of Chinese and English characters and the increase of the size of the character set to be matched. Therefore, the natural scene text recognition algorithm that relies on single character cutting and character set matching is difficult to be directly applied to natural scenes[8].

The segmentation-free natural scene text recognition algorithm is based on deep learning features. Most of the texts in natural scenes appear in the form of sequences, which can be understood as a sequence recognition problem. SHI et al[9]proposed a convolutional recurrent neural network model (CRNN)for recognizing sequential objects in images based on convolutional neural network (CNN) and recurrent neural network (RNN), which mainly adopts a multilayer bi-directional long short-term memory network structure (BiLSTM) to learn the bi-directional dependencies of feature sequences, extract the contextual relationship features of text sequences, and fuse the contextual relationship features

ISSN 2616-5775 Vol. 6, Issue 2: 44-52, DOI: 10.25236/AJCIS.2023.060207

with text features for recognition. Natural scene text recognition is still challenging due to the large amount of irregular, fragmented, blurred and distorted text in natural scene images. Lee et al[10], inspired by SHI et al, proposed recursive recurrent nets with attention modeling (R2AM) which uses recursive convolutional networks with shared parameters without increasing the total number of training parameters. This model uses recursive convolutional neural networks (CNNS) with shared parameters to extract global features of the image without increasing the total number of training parameters, and then decodes them into characters by an implicit character-level recognition model statistical recurrent neural network. The scene text recognition model proposed by Sheng et al[11] designed a feature map transformation block based on Vision Transformer[12] to convert a two-dimensional feature image into a one-dimensional feature sequence. Qiao et al[13] proposed a scene text recognition model that applies an attention mechanism when decoding sequences in a module for image encoding recognition, assigning unused weights to the decoded sequences at each moment to extract richer contextual features.

In summary, most of the current scene text recognition methods use text recognition algorithms that do not require segmentation and only fuse text features, contextual features, and visual features in images to correlate the image text. Most of these recognition algorithms ignore the global semantic features of image text and inter-textual location features, and rarely use the interconnection between multiple features to assist text recognition, so that the features used to encode the lack of interconnection between them, resulting in inaccurate text recognition content. In order to solve the above problems, this paper proposes multi-network convergence and multi-head attention scene text recognition (MCMASTR) algorithm to first extract image using two-branch and three-branch network modules Then, the multi-head attention module adds location and classification information to the feature map, and then assigns different weights to different sequences through the self-attention mechanism to obtain the semantic information of the text. The semantic information of the text is obtained; finally, the text content in the image is obtained by the CTC algorithm[14]based on the characters predicted by the multi-head attention module.

2. Method of This Paper

2.1. Main Network

In this paper, we propose a scene text recognition model based on multi-network convergence and multi-head attention mechanism, which is an end-to-end trained network model including feature extraction module (multi-network convergence), multi-head attention module (multi-head attention), loss function CTC (connectionist temporal classification, CTC) [14].

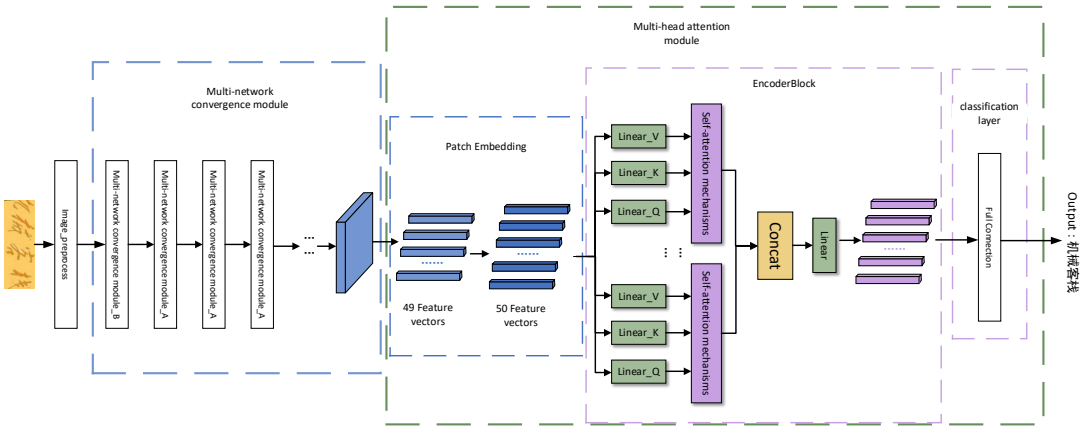

Figure 1: Multi-network convergence and multi-head attention scene text recognition model

The MCMASTR model designed in this paper is shown in Figure 1. Firstly, the irregular input images are preprocessed and the input image size is uniformly processed to 192×32; then a feature extraction module with a multi-network fusion structure is used to extract a large number of visual features from text images, and its structure is shown in Figure 2, which contains two structures of multi-network fusion module _A and _B; secondly, the input visual features are divided in the multi-head attention module, which includes embedding layer, coding layer, and classification layer, the embedding module divides the input visual features by 49 patches then maps each patch to a one-dimensional vector by linear mapping, then superimposes character category features and position encoding in each vector, inputs the 50 adjusted feature vectors into the coding module for encoding, and passes the encoded vectors into the

ISSN 2616-5775 Vol. 6, Issue 2: 44-52, DOI: 10.25236/AJCIS.2023.060207

classification module to generate 50 aligned of one-dimensional sequence features, each feature corresponds to a character in the text, and 50 predicted characters are output; finally, the text content in the image is obtained by CTC algorithm based on the characters predicted by the multi-head attention module.

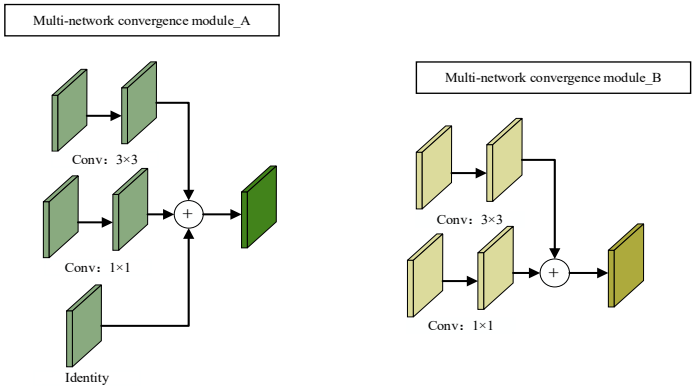

Figure 2: Multi-network convergence module _A and multi-network convergence module _B

2.2. Feature extraction module

As the network deepens, due to the randomness of the convolution kernel parameters, the inhibitory effect of the activation function, etc., some feature information is wasted with each convolution and the corresponding activation operation[15]. In contrast, the multi-network fusion structure proposed in this paper is equivalent to processing the feature information before convolution together with the feature information after convolution, which has the effect of feature information impairment and provides rich feature information for subsequent processing.

The feature extraction stage based on the multi-network fusion structure performs more than two network parallel fusion operations at each convolutional layer, making multiple networks fused into a more robust model. These operations aim to improve the model generalization, representational power and perceptual field information without increasing the network depth, thus extracting richer visual features. The feature extraction module is based on a modified VGG-16 as the backbone network, and its flow is shown in Figure 3. The multi-network fusion structure is shown in Figure 4. The input feature map is divided into two branches and convolved with convolution kernels of different scales, and the two feature maps after fusion convolution are then fed into a three-branch network and convolved using convolution kernels of different scales, and then the three feature maps after convolution are fused to finally generate one feature map. In this way the network can obtain different perceptual fields through the fusion of multiple networks, which provides rich feature information for subsequent processing.

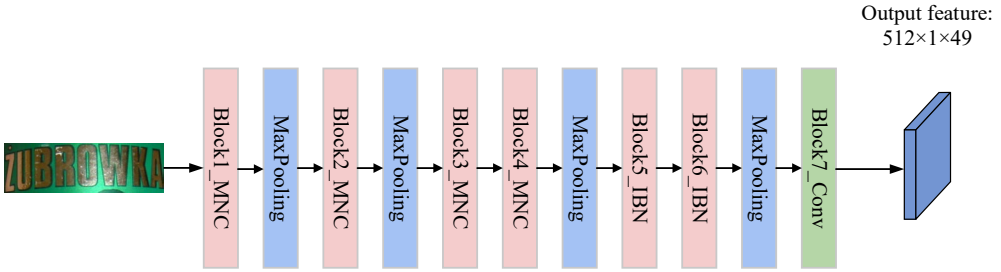

Figure 3: Feature extraction module flowchart

Figure 4 shows the multi-network fusion structure, which contains a 1×1 convolutional branch structure as well as a directly connected branch structure. These branch structures solve the deep network gradient disappearance problem and make the network more easily converge. There are two kinds of branch structures in the multi-network fusion module, such as the multi-network fusion module _A and _B in Fig. 2. The branch structure in the network fusion module _B contains only 1×1 convolutional branch structure; the branch structure in the network fusion module _A contains not only 1×1 convolutional branch structure, but also a directly connected branch structure. Combined with the idea of model integration, since the feature extraction module has multiple branches to add multiple gradient flows to the whole network, training this model only once during the model training process is equivalent

ISSN 2616-5775 Vol. 6, Issue 2: 44-52, DOI: 10.25236/AJCIS.2023.060207

to training multiple models and then fusing them into one model.

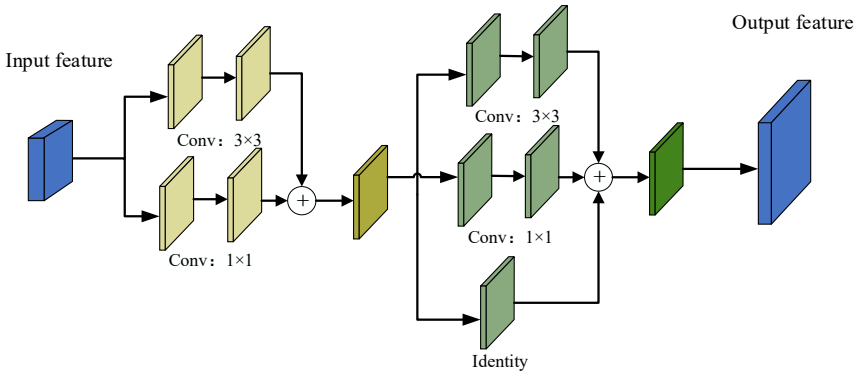

Figure 4: Multi-network convergence structure

Finally, the feature extraction module outputs a feature map of size 512×1×49 (where 512 represents the number of channels of the output text feature map, 1 represents the height of the feature map and 49 represents the width), and then passes the feature map to the multi-head attention module for processing.

2.3. Multi-head attention module

The attention mechanism[16-17]is widely used in sequence recognition problems, and its core idea is to take the original data as the basis to find the correlation between data and then highlight certain important features. In the text recognition problem, the attention mechanism can make the feature correlation between characters in higher-order features. The traditional attention mechanism relies more on time series and serial computation, while the multi-head attention module proposed in this paper does not rely too much on time series and mostly uses parallel computation; the multi-head attention module is a stack of modules consisting of multi-head attention mechanism and feedforward neural network. The multi-head attention mechanism is mainly used to learn the correlation information between characters and contexts in different feature subspaces, and the feedforward neural network acts on each position of the output of the attention mechanism to ensure the acquisition of more and more comprehensive multi-angle feature information representation, and each sublayer is connected using a residual network to enhance the network performance through parallel computing[16].

Figure 5 shows in detail the processing flow of the multi-head attention module, which integrates several self-attentive mechanisms that operate independently[17] and adds position encoding information to each PATCH, allowing the model to learn relevant information in different PATCH subspaces and achieve parallel encoding for this relevant information. When the self-attention mechanism encodes information against the location of the current PATCH, it focuses its attention on its own location and learns against important information features to quickly extract the internal relationships between local features.

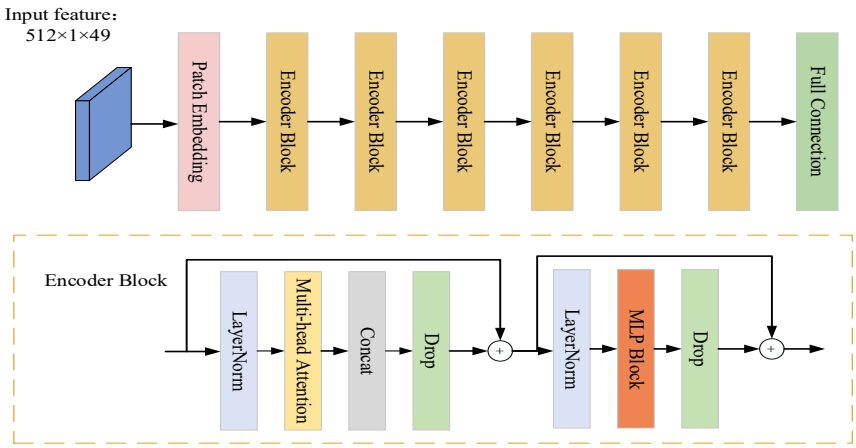

Figure 5: Multi-head attention module structure diagram

ISSN 2616-5775 Vol. 6, Issue 2: 44-52, DOI: 10.25236/AJCIS.2023.060207

2.3.1. Embedding layer

The multi-head attention module deals with a two-dimensional matrix vector sequence containing location and classification information, while the feature map input to the feature extraction module is a three-dimensional matrix, which needs to be transformed into a two-dimensional matrix vector and then add the corresponding location and classification information, and then input to the encoding layer, whose logical structure is shown in Figure 6, where the purple part is the location information and the orange part is the classification information. The feature vector output from the embedding layer is composed of image block input features, category vector and position encoding to form the embedding input vector.

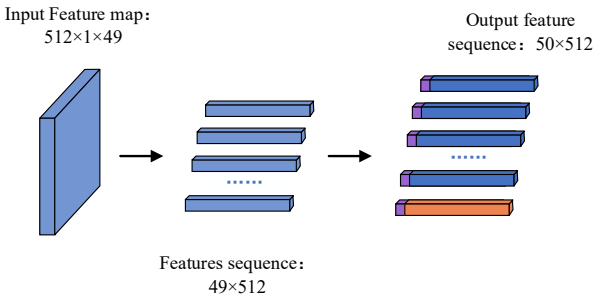

Figure 6: Feature maps converted to sequence features

2.3.2. Encoding layer

The coding layer is repeatedly stacked with multi-head attention and multi-layer perceptron blocks. Firstly, the output features of the embedding layer are passed into the Layer Normalization (LN), and after the LN they are passed into the multi-head attention block, and then they are passed through the Dropout layer for the residual connection, and after the residual, they are passed through the LN, the multi-layer perceptron block, and after the Dropout, one encoding is completed.

The Multi-Layer Perceptron Block (MLP Block) is composed of two full connections, the GELU activation function and two Dropouts.

2.3.3. Classification Layer

The classification layer only needs to decode the feature vector50×512containing location features, classification features, contextual features and semantic features to 50×6885 (number of custom character classes) dimensions to obtain the final classification result. The classification layer consists of layer normalization and a fully-connected classification head that matches the predicted results with a custom character set.

2.4. Loss function

The loss function in this paper uses CTC Loss (Connectionist Temporal Class) to define the probability of labeled sequences, i.e., the characters are transcribed by CTC for the purpose of predicting text sequences. The input of CTC is the predicted sequence output from the multi-head attention module, and i denotes the full length of the input sequence.

3. Analysis of experimental results

3.1. Experimental setup

The CPU model used in the experiment is E5-2666v3 with 128G RAM, the graphics processor model used is NVIDIA Tesla P40, the operating system is Ubuntu, the Python version is 3.9, and the machine learning framework is Pytorch 1.12. The model was pre-trained on 3.6 million Chinese dataset and ICPR MTWI 2018 dataset, with all images normalized to 32 × 192 in size, image batch size of 16, initial learning rate set to 0.001, and model parameters updated during the training process using adam optimizer.

3.2. Data set

The pre-training set used in this experiment is the 3.6 million Chinese dataset and the ICPR MTWI

ISSN 2616-5775 Vol. 6, Issue 2: 44-52, DOI: 10.25236/AJCIS.2023.060207

2018 dataset. The 3.6 million Chinese dataset has 3.64 million images in it, and the dataset is divided into two parts, the training set and the validation set, according to 99:1. Its images are randomly generated by variations of font, size, grayscale, blur, perspective, and stretch in the corpus. ICPR WTMI2018 is a mixed Chinese and English dataset with 20,000 images, and the dataset is divided 1:1 into two parts: the training set and the validation set. The images are all from web images, and the text content is mainly horizontal, skewed, blurred and mutilated text, and there are more types of fonts.

The relevant descriptions of the test set are shown as follows.

(1) Rule text dataset

The ICDAR 2013 dataset consists of clear photos of street billboards and road signs. The SVT dataset is from the Google Street View image library. The IIT5K dataset is composed of street view images. The three datasets consist of 1015, 647 and 3000 cropped scene text images, respectively.

(2) Irregular text dataset

The ICDAR 2015 dataset is dominated by clearly taken photos of street billboards and street signs, etc. The SVTP dataset is from the Google Nature Street View captured image library. The CUTE 80 dataset is high-resolution images taken from natural scenes. The three datasets consist of 1500, 639 and 288 cropped text images of the scenes, respectively.

3.3. Evaluation metrics

There are two general evaluation metrics for natural scene text recognition, one is Character recognition accuracy (CRA) and the other is Text recognition accuracy (TRA). Character recognition accuracy generally indicates the proportion of the number of characters recognized correctly to the total characters, and the evaluation index is more intuitive[18], and its expression is Equation (1).

$$CRA = \frac{W}{A} \tag{1}$$

Where, CRA denotes the character recognition accuracy, W denotes as the number of correctly recognized characters, and A denotes the number of all characters.

Text recognition accuracy generally uses edit distance, and the number of steps required when a character needs to be converted into another character (including deletion, insertion and replacement) is called edit distance[18], and its expression is Equation (2).

$$TRA = \frac{S-D-I-R}{N} \tag{2}$$

where TRA denotes the text recognition accuracy, where S denotes the number of total characters, and D, I, and R denote the number of steps for deletion, insertion, and replacement, respectively.

3.4. Ablation experiments

In order to verify the performance of the recognition algorithm in this paper, the experiments select VGG BiLSTM as the base model and conduct ablation experiments on four datasets: regular dataset: 3.6 million Chinese, SVT and irregular dataset: ICPR MTWI 2018, ICDAR 2015, and the results are shown in Table 1.

From the data in Table 1, it can be obtained that after adding the feature extraction module (MCN) with multi-network fusion structure, the average character accuracy of MCMASTR and the base model on regular text and irregular text are improved by 0.7% and 2.0%, and the average text accuracy is improved by 0.6% and 1.7%, respectively, which proves the feasibility of multi-network fusion, which can improve the effect of text feature extraction. And after replacing the BiLSTM in the base model with the multi-head attention module (MHA) proposed in this paper, compared with the base model, the average character accuracy of MCMASTR on regular and irregular text is improved by 0.85% and 2.0%, and the average text accuracy is improved by 2.6% and 2.3%, respectively, proving that the multi-head attention module is better than BiLSTM in character decoding. After replacing BiLSTM in the base model with the multi-head attention module, MCMASTR improves the character accuracy by 1.9% and 2.1% and text accuracy by 2.45% and 2.5%, respectively, compared with the base model using the same multi-network fusion structure, which proves that multi-network fusion and multi-head attention have better performance in character recognition.

Figure 7 shows that with the same multi-network fusion module, the accuracy of the model on regular

ISSN 2616-5775 Vol. 6, Issue 2: 44-52, DOI: 10.25236/AJCIS.2023.060207

and irregular data sets is steadily improved by experimentally increasing the number of multi-head attention modules in the model, but the model performs best when the number of multi-head attention modules reaches 6, and then increasing the number of multi-head attention modules leads to long training time and accuracy. However, the model performs best when the number of multiple attention modules reaches 6, and then increasing the number of multiple attention modules leads to long training time and decreasing accuracy, which proves that multiple attention modules can guarantee good training effect in repeated processing of feature sequences. It also shows that increasing the depth of the network and the number of modules does not necessarily lead to excellent results, because the deeper the network contains a large number of nonlinear changes, each change is equivalent to the loss of some original information of the features, which leads to more serious degradation the deeper the layers are. improve the training techniques will also become the focus of research.

Table 1: Ablation experiments for regular and irregular text datasets.

model	3.6 million Chinese dataset		ICPR MTWI 2018		SVT		ICDAR 2015	
	CRA	TRA	CRA	TRA	CRA	TRA	CRA	TRA
VGG+BiLSTM	92.3	91.7	75.5	74.9	89.6	88.6	77.8	77.1
MCN+BiLSTM	93.0	92.1	77.9	76.7	90.3	89.8	79.4	78.8
VGG+MHA	92.9	92.4	77.6	76.3	90.7	90.3	80.9	80.4
MCN+MHA	**95.3**	**94.8**	**79.4**	**78.1**	**91.8**	**91.4**	**82.8**	**82.4**

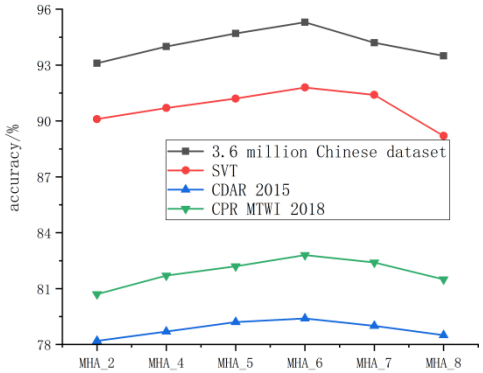

Figure 7: Comparative experiment on the number of multi-head attention modules

3.5. Comparison test

As can be seen from Table 2, the algorithm proposed in this paper achieves good results on more than half of the standard datasets. The accuracy of this algorithm can reach 91.4% on the regular text dataset SVT, and 82.4 and 82.7% on the irregular text datasets ICDAR2015 and SVTP, respectively, which are better than the ScRN algorithm proposed by Yang M et al. in 2019 and the SEED algorithm proposed by Qiao Z et al. in 2020. The accuracy of this paper's algorithm on three datasets ICDAR2013, IIT5k, and CUTE80 is 93.1%, 92.3%, and 84.2% respectively. There is a gap between some mainstream algorithms, but it has a strong competitive edge compared to other algorithms.

Table 2: Comparative experiment between the algorithm in this paper and mainstream algorithms in recent years.

Method	SVT	ICDAR2013	IIT5k	ICDAR2015	SVTP	CUTE80
CRNN[9]	80.8	86.7	78.2	---	---	---
RARE[19]	81.5	87.5	79.7	---	71.8	59.2
AON[20]	82.8	---	87.0	68.2	73.3	76.8
MORAN[21]	88.3	92.4	91.2	68.8	76.1	77.4
SEED[13]	89.6	92.8	93.8	80.0	81.4	83.6
R^2AM[10]	80.7	90.0	78.4	---	---	---
ScRN[22]	88.9	93.9	94.4	78.7	80.8	87.5
MCMASTR	**91.4**	**93.1**	**92.3**	**82.4**	**83.7**	**84.2**

ISSN 2616-5775 Vol. 6, Issue 2: 44-52, DOI: 10.25236/AJCIS.2023.060207

4. Conclusions

Traditional recognition methods require a lot of repetitive operations and have a low recognition rate. The MCMASTR model does not require complex manual operations and only requires input images to recognize the corresponding text information. The method uses the feature extraction module of multi-network fusion feature to build the backbone network of the model, through which visual features, text features, contextual features and global semantic features at different scales are extracted in the image, and a more robust feature map is constructed by fusing the multi-networks together, and as the model deepens and the multi-networks continue to fuse, the feature map eventually obtained will become more favorable for text recognition. In the character prediction stage, this paper proposes a multi-head attention module, which focuses on the location information, classification information and internal connections among the feature sequences, and highlights certain important features that are beneficial to the decoding of predicted characters. The experimental results show that the algorithm proposed in this paper achieves good results on both regular and irregular text datasets, and can effectively and recognize blurred, mutilated, distorted and overexposed text, which makes the method in this paper more practical in natural scenes compared with general text recognition algorithms. This paper focuses on Chinese and English text recognition in natural scenes, while for improving model accuracy, recognition speed and hardware porting will be the focus of future research.

References

[1] Liu C, Cao Y, Luo Y, et al. Deepfood: Deep learning-based food image recognition for computer-aided dietary as-sessment[C]//International Conference on Smart Homes and Health Telematics. Springer, Cham, 2016: 37-48.

[2] Kahn G, Villaflor A, Ding B, et al. Self-supervised deep reinforcement learning with generalized computation graphs for robot navigation[C]//2018 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2018: 5129-5136.

[3] Shen Z, Zhang R, Dell M, et al. LayoutParser: A unified toolkit for deep learning based document image analy-sis[C]//International Conference on Document Analysis and Recognition. Springer, Cham, 2021: 131-146.

[4] Liu C, Cao Y, Luo Y, et al. Deepfood: Deep learning-based food image recognition for computer-aided dietary as-sessment[C]//International Conference on Smart Homes and Health Telematics. Springer, Cham, 2016: 37-48.

[5] Shi C, Wang C, Xiao B, et al. Scene text recognition using part-based tree-structured character detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2013: 2961-2968.

[6] Romero V, Sanchez J A, Bosch V, et al. Influence of text line segmentation in handwritten text recognition[C]//2015 13th International Conference on Document Analysis and Recognition (ICDAR). IEEE, 2015: 536-540.

[7] Coates A, Carpenter B, Case C, et al. Text detection and character recognition in scene images with unsupervised feature learning[C]//2011 International conference on document analysis and recognition. IEEE, 2011: 440-445.

[8] LIU C Y, CHEN X X, LUO C J. Deep Learning Method for Text Detection and Recognition in Natural Scenes[J]. Journal of Image and Graphics,2021,26(06):1330-1367.

[9] Shi B, Bai X, Yao C. An end-to-end trainable neural net-work for image-based sequence recognition and its ap-plication to scene text recognition[J]. IEEE transactions on pattern analysis and machine intelligence, 2016, 39(11): 2298-2304.

[10] Lee C Y, Osindero S. Recursive recurrent nets with atten-tion modeling for ocr in the wild[C]//Proceedings of the IEEE conference on computer vision and pattern recog-nition. 2016: 2231-2239.

[11] Sheng F, Chen Z, Xu B. NRTR: A no-recurrence se-quence-to-sequence model for scene text recogni-tion[C]//2019 International conference on document analysis and recognition (ICDAR). IEEE, 2019: 781-786.

[12] Arnab A, Dehghani M, Heigold G, et al. Vivit: A video vision transformer[C]//Proceedings of the IEEE/CVF In-ternational Conference on Computer Vision. 2021: 6836-6846.

[13] Qiao Z, Zhou Y, Yang D, et al. Seed: Semantics enhanced encoder-decoder framework for scene text recogni-tion[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 13528-13537.

[14] Graves A, Fernández S, Gomez F, et al. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks[C]//Proceedings of the 23rd international

ISSN 2616-5775 Vol. 6, Issue 2: 44-52, DOI: 10.25236/AJCIS.2023.060207

conference on Machine learning. 2006: 369-376.
[15] Zhanbin L, Feng L, Xiting W. Design of Multi-Network Convergence for Complex Networks[C] //2018 IEEE 9th International Conference on Software Engineering and Service Science (ICSESS). IEEE, 2018: 211-215.
[16] Fukui H, Hirakawa T, Yamashita T, et al. Attention branch network: Learning of attention mechanism for visual ex-planation[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 10705-10714.
[17] Subakan C, Ravanelli M, Cornell S, et al. Attention is all you need in speech separation[C]//ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021: 21-25.
[18] SUN J, ZHU Y Q, HUANG C N. Scene text recognition model based on two-dimensional CTC and attention se-quence[J].Electronic Production, 2022,30(17):65-70.
[19] Shi B, Wang X, Lyu P, et al. Robust scene text recogniton with automatic rectification [C]// Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 4168-4176.
[20] Cheng Z, Xu Y, Bai F, et al. Aon: Towards arbitrari-ly-oriented text recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 5571-5579.
[21] Luo C, Jin L, Sun Z. Moran: A multi-object rectified at-tention network for scene text recognition [J]. Pattern Recognition, 2019, 90: 109-118.
[22] Yang M, Guan Y, Liao M, et al. Symmetry-constrained rectification network for scene text recognition[C]//Proceedings of the IEEE/CVF international con-ference on computer vision. 2019: 9147-9156.