International Journal of Computer Applications Technology and Research
Volume 14–Issue 07, 28 – 31, 2025, ISSN:-2319–8656
DOI:10.7753/IJCATR1407.1003

Integrating BiLSTM, SE Attention and EMA for OCR

Zehua Lv
Chengdu University of Information Technology
College of Communication Engineering
ChengDu, China

Yan Chen
Chengdu University of Information Technology
College of Communication Engineering
ChengDu, China

Chengyu Hou*
Chengdu University of Information Technology
College of Communication Engineering
ChengDu, China

Abstract: OCR (Optical Character Recognition) is a technology that converts text in an image into editable text. It automatically extracts and converts textual information by processing, analyzing, and recognizing images, and has wide applications in multiple fields This article proposes a deep learning model called CRNN (Convolutional Recurrent Neural Network) that combines the advantages of CNN (Convolutional Neural Network) and RNN (Recurrent Neural Network). It also incorporates BiLSTM (Bidirectional Long Short Term Memory Network), SE Attention (Squeeze Excitation Attention Mechanism), and EMA (Exponential Moving Average). The system extracts image features through CRNN, processes sequence information through BiLSTM, and achieves end-to-end text recognition by combining CTC loss. The attention mechanism SEBlock is used to enhance feature selection ability, and the bidirectional LSTM combines layer normalization to improve long sequence modeling ability and optimize the EMA stable training process.

Keywords: OCR;CRNN;BiLSTM;SEBlock;EMA

1. INTRODUCTION

OCR (Optical Character Recognition) is a technology that obtains images through scanning, shooting, and other methods, and automatically converts the text content into editable and searchable electronic text formats. Its development process spans multiple stages. In the early days, it was based on template matching technology and recognized by comparing pre-set character shape templates with images [1]. However, it only supports limited fonts and simple scenes, with low accuracy. By the 1990s to 2010, statistical learning methods such as Hidden Markov Models (HMM) and Support Vector Machines (SVM) were introduced, combined with feature extraction techniques, to achieve support for multiple languages and complex layouts, resulting in a significant improvement in recognition rates. After 2010, deep learning technology became mainstream. Convolutional neural networks (CNN), recurrent neural networks (RNN), and other models were trained on large amounts of data, significantly improving their generalization ability. They were able to handle natural scene text, handwriting, complex typesetting, and other situations **Error! Reference source not found.**. At the same time, they combined attention mechanisms and Transformer architecture to further optimize long text and multilingual processing. In terms of application scenarios, OCR has widely penetrated into various fields. In office and document processing, it can scan paper contracts, invoices, and documents and convert them into electronic documents, achieving automated data entry; On mobile devices, functions such as phone photo translation, handwritten note recognition, and real-time subtitle generation for short videos all rely on OCR; In the fields of finance and security, it can be found in scenarios such as bank check recognition, license plate recognition, and ID card information extraction; In the education and publishing industry, the digitization of ancient books and automatic grading of test papers also rely on OCR technology; In industry and logistics, product label recognition and tracking number scanning also rely on this technology to improve efficiency [3]. In order to improve the accuracy of OCR in recognizing text, this paper studies a combination structure that combines BiLSTM, SEBlock, and EMA to enhance traditional OCR as follows:

- Capture bidirectional semantic information: BiLSTM consists of two LSTMs with opposite directions, which can simultaneously extract features from both the forward and reverse directions of the sequence, enabling the model to obtain more comprehensive contextual information. Understanding the dependencies between characters is crucial for text recognition tasks.
- Smooth model parameters: By exponentially weighted averaging the model parameters, EMA can reduce parameter fluctuations during training, making the model more stable on the validation set, especially in the later stages of training, which can effectively improve the model's generalization ability.
- Channel attention mechanism: SEBlock weights the channel dimension of the feature map by learning the importance of different channel features, making the model more focused on key features related to character recognition (such as edges, contours, etc.) and suppressing irrelevant information.

2. RELATED WORK

Park et al. pre trained U-Net with the help of a forward diffusion process and a feature extractor to improve the quality of low resolution text images through a backward diffusion process [4]. Lowe et al. used the core technology of OCR2SEQ to simulate real text extraction errors. These technologies excel at generating diverse and challenging data scenarios, greatly improving the training efficiency and accuracy of text to text converters. The application of OCR2SEQ has shown significant improvement in data processing accuracy, especially in industries that heavily rely on OCR technology, such as healthcare and library science [5]. Zhou Jinfei et al. proposed a cross region feature fusion method based on the Geometric Relationship Transformer (CFGRTransformer) for OCR based image captioning. The network first establishes the association between OCR and image object regions by constructing relative geometric relationships, including width/height differences, distance, IOU (Intersection over Union), inclusion relationships, and angle offsets. Then, by combining intra - and cross region features, it aggregates entities from different modalities through a multi head attention mechanism based on relative relationships [6]. Yuliang LIU et al. proposed OCRBench as a comprehensive evaluation benchmark to facilitate the assessment of optical character recognition (OCR) capabilities in large multimodal models [7]. Cheema Musa Dildar Ahmed and others elaborated on the significance of focusing on these languages and introduced ViLanOCR, an innovative bilingual OCR system tailored for Urdu and English. Unlike existing systems that struggle to handle the complexity of low resource languages, ViLanOCR utilizes advanced language models based on multilingual converters to achieve outstanding performance [8]. Yan, Feng, and others used OCR models to detect and obtain OCR markers in images, which helps to further understand the images. We have designed a model based on a common attention mechanism, which includes a problem self attention unit, a problem guided image visual attention unit, and a problem guided image OCR token attention unit [10]. Mosbah Lamia et al. attempted to address these challenges by creating a deep learning OCR called ADOCRNet for Arabic document recognition [11].

3. SYSTEM MODEL

This article will accurately propose the OCR model and add several key modules to the model. The results are shown in the Figure 1:

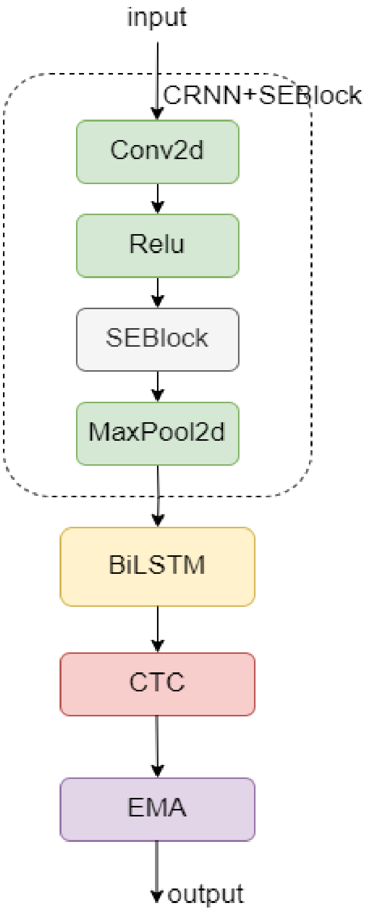

Figure 1 system model

3.1 CRNN module

CRNN (Convolutional Recurrent Neural Network) is a deep learning model commonly used for scene text recognition. Its core feature is the combination of CNN and RNN, which can directly process sequence data (such as text images) and output recognition results [12][13].First layer convolution group: Convert the input image of channel 1 into a feature map of channel 64, using a 3x3 convolution kernel with a stride of 1 and padding of 1. ReLU activation introduces nonlinearity . Maximum pooling reduces the size of feature maps by half and enhances the translational invariance of features. Second layer convolution group: Similar to the first layer, but increases the number of channels to 128. Third layer convolution group: Increase to 256 channels. The fourth layer convolution group: uses asymmetric pooling layers (2,1) and stride sizes (2,1) to downsample only in the height direction, preserving sequence information in the width direction. Fifth layer convolution group: Increase the number of channels to 512. Sixth layer convolution group: Use asymmetric pooling again to further downsample in the height direction. The structure of each layer is shown in the Figure 2:

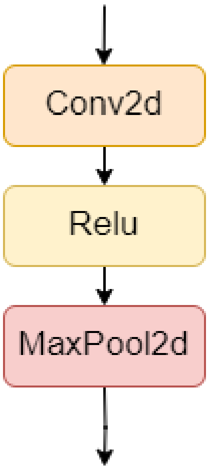

Figure 2 CRNN model

3.2 BiLSTM Module

This Enhanced BiLSTM class is an enhanced version of bidirectional LSTM in PyTorch, designed specifically for sequence modeling tasks. It captures the contextual information of the sequence through a bidirectional structure, accelerates training stability through layer normalization, and optimizes the learning ability of long sequences using a special weight initialization strategy. The input format of the model is (seq_1en, batch_2, input_2), and the output dimension is twice the size of the hidden layer (bidirectional merging). In weight initialization, the input gate and forget gate are uniformly distributed using Xavier, and the hidden state weights are initialized orthogonally. Specifically, the forget gate bias is set to 1 to enhance long-term memory. After applying layer normalization to the output of LSTM, it ensures that the features of each dimension maintain a stable distribution during training. This model is suitable for tasks such as NLP and speech recognition that require capturing temporal dependencies. By integrating these optimized designs, it performs better in handling complex sequence patterns [14].

3.3 SEBlock module

The SE (Squeeze and Excitation Block) is a structure used to enhance the feature expression ability of convolutional neural networks. Its core idea is to model the dependency relationships between feature channels, allowing the network to automatically learn the importance of different channel features, thereby enhancing the weight of useful features and suppressing the influence of irrelevant or redundant features. It mainly consists of two key steps: Squeeze and Excitation. In the Squeeze stage, the spatial information of each channel is compressed into a single value through global average pooling, which can be regarded as the global statistical information of the channel's features; In the Excitation stage, a simple fully connected neural network is used to process these compressed values, generate weight coefficients for each channel, and finally multiply these weight coefficients with the original feature map to achieve adaptive adjustment of different channel features [15].

In this code, the SE module is embedded into the convolutional layer of the CRNN model, with SEBlocks added after multiple convolution operations. The purpose of doing this is to dynamically weight the feature channels output by each convolutional layer through the SE module during the process of extracting image features from the convolutional layer. Due to the use of CRNN for text recognition tasks, the text regions contained in the input image may be affected by noise, background interference, and other factors. The importance of features extracted from different channels for text recognition varies. The SE module enables the network to focus on feature channels that contain effective textual information, such as edges, textures, and other features related to character structure, while reducing the weight of channels that contain noise or background information, thereby improving the quality of the output features of the convolutional layer. This helps the subsequent recurrent neural network (BiLSTM) to better learn the temporal information of the text from the feature sequence, ultimately improving the accuracy of text recognition and reducing the error rate and word count.

3.4 EMA module

The EMA (Exponential Moving Average Model) is a technique used for smoothing time series data. It reduces noise and volatility by weighted averaging the data, thereby extracting trends from the data [16]. In deep learning, EMA is often used in the process of updating and optimizing model parameters. It can help the model converge more stably during the training process and improve its generalization ability.

The calculation formula (1) for EMA is as follows:

$$EMA(t) = (1 - alpha) * EMA(t-1) + alpha * value(t) \quad (1)$$

Among them, $EMA(t)$ is the exponential moving average at time point t, $EMA(t-1)$ is the exponential moving average at the previous time point, $value(t)$ is the value at the current time point, and alpha is the smoothing factor (with a value range of [0,1]), which determines the weight of the current value in the calculation.

In deep learning, EMA is often used for the following two purposes:

Parameter update: During the model training process, optimization algorithms such as gradient descent are usually used to update the parameters of the model. When using EMA to update parameters, the parameters can be updated by calculating the exponential moving average of the parameters, thereby reducing the noise and fluctuations of parameter updates.

Model prediction: During the model prediction phase, the exponential moving average of the parameters obtained during the training process can be used for prediction. This can reduce the fluctuation of the model's prediction results and improve the stability of the prediction.

4. EXPERIMENTS AND ANALYSIS

In this chapter, we compared the improved CRNN model with multiple other models. For the public distribution ICDAR 2013 dataset, the training frequency is set to 500 times, and the learning rate is set to 0.001. Ultimately, we used ACC (Accuracy),latency and WER (Word Error Rate) as evaluation criteria.

4.1 Comparison of Module

Among them,ACC is an indicator that measures the proportion of correctly predicted samples in the total sample of the model; WER is the accuracy calculated by weighting different categories of samples; Latency refers to the time required for data processing or system response, with the following indicators:

Table 1. The performance of different models in the dataset

Module	**ACC**	**WER**	**Latency(h)**
CRNN	0.628	0.287	1.54
CRNN+EMA	0.688	0.263	1.62
CRNN+SE	0.676	0.271	1.65
CRNN+BiLSTM	0.734	0.243	1.68
CRNN+EMA+SE	0.792	0.223	1.71
CRNN+EMA+BiLSTM	0.801	0.234	1.69
CRNN+SE+BiLSTM	0.793	0.211	1.77
CRNN+SE+BiLSTM+EMA	0.846	0.169	1.81

It can be seen that after the addition of modules such as BiLSTM, although the latency time increased, the accuracy of WER improved by 41.1% and ACC improved by 34.7%.

4.2 Comparison of training loss

The loss chart is a very important visualization tool in the training process of machine learning and deep learning models. It takes the number of training iterations (or epochs) as the horizontal axis and the calculated loss value for each iteration as the vertical axis, intuitively showing the trend of the model's loss value changes during the training process. By observing the loss plot, we can determine the training status of the model,Figure 3 shows the loss plots under different models:

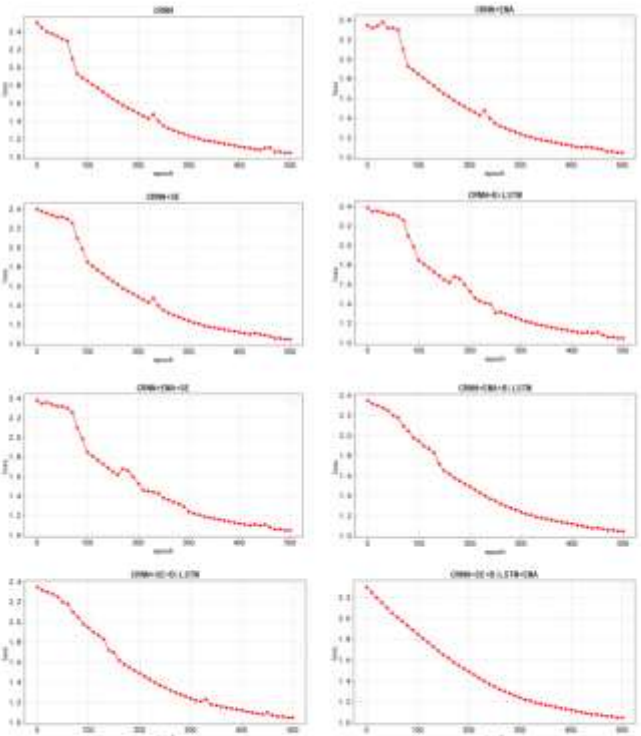

Figure 3 Variation of loss curves for different models

It can be seen from the figure that the loss curve fluctuates greatly when adding a single module. When adding two modules, the curve is basically the same. When adding all modules, the curve tends to be a smoot.

5. CONCLUSION

Overall, we can conclude that the improved model has better ACC, WAR, and Latency performance on the switch common dataset. Therefore, improving the model in this experiment is effective, and the image text recognition ability has been successfully improved on the basic model.

REFERENCES

[1] Lee,,Aram,Yu,,Hongyeon,Min,,& Gihyeon.(2024).An algorithm of line segmentation and reading order sorting based on adjacent character detection: A post-processing of OCR for digitization of Chinese historical texts.JOURNAL OF CULTURAL HERITAGE,67,80-91.

[2] Khallouli,,Wael,Uddin,,Mohammad,Shahab,Sousa-Poza,,Andres,Li,,Jiang,Kovacic,,& Samuel.(2025).Leveraging Transformer-Based OCR Model with Generative Data Augmentation for Engineering Document Recognition.ELECTRONICS,14(1),5-5.

[3] Singh,,Katyani,Tata,,Ganesh,Van,Oeveren,,Eric,Ray,,& Nilanjan.(2025).Unpaired document image denoising for OCR using BiLSTM enhanced CycleGAN.INTERNATIONAL JOURNAL ON DOCUMENT ANALYSIS AND RECOGNITION,28(2),207-224.

[4] Park,,Chae-Won,Palakonda,,Vikas,Yun,,Sangseok,Kim,,Il-Min,Kang,,& Jae-Mo.(2024).OCR-Diff: A Two-Stage Deep Learning Framework for Optical Character Recognition Using Diffusion Model in Industrial Internet of Things.IEEE INTERNET OF THINGS JOURNAL,11(15),25997-26000.

[5] Lowe,,Michael,Prusa,,Joseph,D.,Leevy,,Joffrey,L.,Khoshgoftaar,,Taghi,& M..(2024).Advancing machine learning with OCR2SEQ: an innovative approach to multi-modal data augmentation.JOURNAL OF BIG DATA,11(1),86.

[6] Zhou,,Jinfei,Yang,,Cheng,Zhu,,Yaping,Zhang,,& Yana.(2024).Cross-region feature fusion with geometrical relationship for OCR-based image captioning.NEUROCOMPUTING,601.

[7] Yuliang,LIU,Zhang,LI,Mingxin,HUANG,Biao,YANG, Wenwen,YU,Chunyuan,LI,Xu-Cheng,YIN,Cheng-Lin,LIU,Lianwen,JIN,Xiang,& BAI.(2024).OCRBench: on the hidden mystery of OCR in large multimodal models.Science China(Information Sciences),67(12),23-35.

[8] Cheema,,Musa,Dildar,Ahmed,Shaiq,,Mohammad,Daniyal,Mirza,,Farhaan,Kamal,,Ali,Naeem,,M.,& Asif.(2024).Adapting multilingual vision language transformers for low-resource Urdu optical character recognition (OCR).PEERJ COMPUTER SCIENCE,10,e1964.

[9] SCIENCE,10,e1964.

[10] Yan,,Feng,Silamu,,Wushouer,Chai,,Yachuang,Li,,& Yanbing.(2024).OECA-Net: A co-attention network for visual question answering based on OCR scene text feature enhancement.MULTIMEDIA TOOLS AND APPLICATIONS,83(3),7085-7096.

[11] Mosbah,,Lamia,Moalla,,Ikram,Hamdani,,Tarek,M.,Neji,,Bilel,Beyrouthy,,Taha,Alimi,,Adel,& M..(2024).ADOCRNet: A Deep Learning OCR for Arabic Documents Recognition.IEEE ACCESS,12,55620-55631.

[12] Xu,,Fan,Chen,,Chuibin,Shang,,Zhigao,Peng,,Yuqing,Li,,& Xinbao.(2024).A CRNN-based method for Chinese ship license plate recognition.IET IMAGE PROCESSING,18(2),298-311.

[13] Zhijun,,Guo,Weiming,,Luo,Qiujie,,Chen,Hongbo,,& Zou.(2024).Terminal strip detection and recognition based on improved YOLOv7-tiny and MAH-CRNN+CTC models.FRONTIERS IN ENERGY RESEARCH,12.

[14] Yan,,Jun,Li,,Junhong,Bai,,Guixiang,Li,,& Yanan.(2025).An Attention-BiLSTM network identification method for time-delay feedback nonlinear system.APPLIED INTELLIGENCE,55(1),1-15.

[15] Wang,,Xiao,Zhang,,Liang,Meng,,& Xiangdong.(2024).Linear Array DOA Estimation Based on CNN-SEBlock Under Low Signal-to-Noise Ratio.

[16] Xiao,,Xiang,Sheng,,& Yuhong.(2024).Uncertain vector moving average model based on Welsch loss function.COMMUNICATIONS IN STATISTICS-SIMULATION AND COMPUTATION.