J Inf Process Syst, Vol.18, No.4, pp.480~488, August 2022
https://doi.org/10.3745/JIPS.02.0177

ISSN 1976-913X (Print)
ISSN 2092-805X (Electronic)

JOURNAL OF INFORMATION PROCESSING SYSTEMS JIPS

Tobacco Retail License Recognition Based on Dual Attention Mechanism

Yuxiang Shan[1], Qin Ren[1], Cheng Wang[1], and Xiuhui Wang[2,*]

Abstract

Images of tobacco retail licenses have complex unstructured characteristics, which is an urgent technical problem in the robot process automation of tobacco marketing. In this paper, a novel recognition approach using a double attention mechanism is presented to realize the automatic recognition and information extraction from such images. First, we utilized a DenseNet network to extract the license information from the input tobacco retail license data. Second, bi-directional long short-term memory was used for coding and decoding using a continuous decoder integrating dual attention to realize the recognition and information extraction of tobacco retail license images without segmentation. Finally, several performance experiments were conducted using a largescale dataset of tobacco retail licenses. The experimental results show that the proposed approach achieves a correction accuracy of 98.36% on the ZY-LQ dataset, outperforming most existing methods.

Keywords

Attention Mechanism, Image Recognition, Robot Process Automation (RPA)

1. Introduction

Robot process automation (RPA) used in tobacco marketing involves the intelligent recognition of multiple scene objects during tobacco operations. For example, to evaluate the recognition accuracy of the impact of tobacco marketing activities on the product market, it is necessary to deeply analyze and mine various text and image information from the retail process of the products involved. However, there are still many challenges in realizing intelligent tobacco operations and management [1]. One of the primary issues is to identify tobacco retail licenses from different regions and extract relevant information. Tobacco retail license images have complex unstructured features [2] that increase the difficulty of extracting information from them.

To realize the automatic recognition and information extraction from tobacco retail license images, a novel recognition approach using a double-attention module is presented in this paper. The proposed method is mainly used under two scenarios: (1) identifying the license number and authorization date from tobacco retail license images and (2) statistically analyzing the marketing of specific types of tobacco. For the proposed network, DenseNet was used to obtain the license information from the input

Manuscript received January 27, 2022; first revision March 21, 2022; accepted June 15, 2022.
* **Corresponding Author:** Xiuhui Wang (wangxiuhui@cjlu.edu.cn)
[1] Chinese Tobacco Zhejiang Industrial Company Limited, Hangzhou, China (188912338@qq.com, 32926357@qq.com, 275755632@qq.com)
[2] Dept. of Computer, China Jiliang University, Hangzhou, China (wangxiuhui@cjlu.edu.cn)

data from tobacco retail licenses. Furthermore, bi-directional long short-term memory (BiLSTM) was used for the coding and decoding using a continuous decoder integrating dual attention to realize the recognition and information extraction of tobacco retail license images without segmentation. The contributions of this study can be summarized as follows:

(1) A new recognition method based on a double attention mechanism is proposed for the automatic recognition and information extraction from tobacco retail license images.
(2) DenseNet and BiLSTM are integrated to extract the features and information from tobacco retail license images without the need for segmentation.
(3) A comprehensive evaluation was conducted using a largescale tobacco retail license image dataset. We thoroughly evaluated the proposed recognition network using the ZY-LQ dataset and obtained a 98.36% correct recognition rate, outperforming most existing approaches.

The remainder of this paper is organized as follows. In the next section, we discuss existing research on image recognition for different applications. In Section 3, we propose a novel recognition network for unstructured tobacco retail license images by integrating the attention mechanism and decoder module. In Section 4, we describe the experiments conducted and present the evaluation of the proposed method using a largescale tobacco retail license image dataset. Finally, in Section 5, we provide concluding remarks regarding the present study.

2. Related Work

Image recognition is an important branch of artificial intelligence. Researchers and research institutes have proposed a variety of models and algorithms to solve different application problems [3–7]. Zhang et al. [3] proposed the ASSDA method to deal with text images from different domains, which focuses on aligning a cross-domain distribution. In [4], a recognition framework for text lines was presented for embedded applications, with more attention paid to the balance between limited resources and the recognition rate. A convolutional neural network (CNN)-based card recognition framework [5] was proposed for a similar task of tobacco retail license image recognition, which has mainly been used to improve the robustness to different environments and the efficiency of processing natural images. Bera et al. [6] focused on discriminating fine-grained changes with a particular emphasis on a distinct fine-grained visual classification. Islam et al. [7] presented a region-of-interest detection method that focuses on Bangla text extraction and recognition from natural scene images.

In addition, attention mechanisms [8,9] have been utilized in many different applications, such as natural language processing, video-based understanding, and visual classification, which is an important direction in deep learning technology. Lai et al. [8] conducted a comprehensive review of the attention mechanism used in model optimization. Luo et al. [9] proposed a depth-characteristic combination framework that integrates a variety of attention modules to realize iris recognition.

Specifically, existing image recognition algorithms focus on different issues and can solve many actual application problems; however, there is still no effective solution for the recognition of tobacco retail licenses having significant unstructured characteristics.

3. Recognition Method based on Dual Attention Mechanism

As shown in Fig. 1, considering the unstructured and multiscale characteristics of sample images of tobacco retail licenses, the recognition network proposed in this paper integrates an attention mechanism with a decoder module.

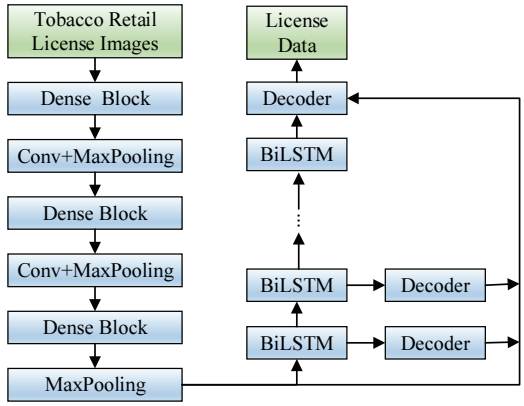

Fig. 1. Architecture of the proposed dual attention network.

3.1 Feature Extraction

The proposed recognition network conducts a feature extraction through the DenseNet module, which consists of DenseBlock and a transition layer in the middle. Among them, the transition layer connects different DenseBlocks and integrates the features obtained by the previous version. DenseBlock constructs dense connections from the front to the rear layers for reusing the characteristics, as shown in Fig. 2. The dataflow design makes the feature extraction more effective, enhances the gradient propagation, and improves the convergence speed of the proposed recognition model. Moreover, we use a convolutional kernel with a size of 1 × 1 in each layer, which simplifies the feature maps and improves the effectiveness of the proposed recognition model.

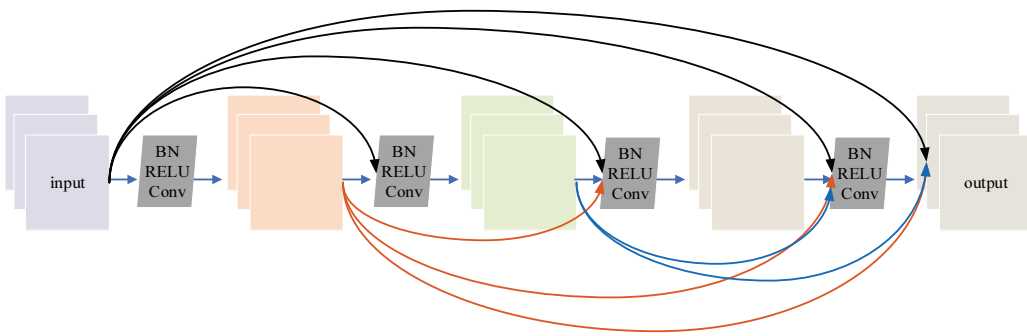

Fig. 2. Dense block used in the proposed network.

The key parts of our information extraction module are continuous decoders that fuse the attention modules. When image I_0 is input into the CNN model, where $H_l(.)$ is the nonlinear transformation

function after each layer, and l is the corresponding index of each layer, the traditional forward propagation network connects all the layers as an input and obtains.

$$I_l = \mathrm{H}_l(I_{l-1}). \tag{1}$$

Nevertheless, DenseNet connects each layer in the feedforward model of the residual connection such that the l^{th} layer takes characteristic diagrams in front of it as input, that is,

$$I_l = \mathrm{H}_l([I_0, I_1, \ldots, I_{l-1}]), \tag{2}$$

where $I_0, I_1, \ldots, I_{l-1}$ are the feature maps generated by the 0-*th*, 1-*th*,…, and $(l-1)$*th* layers, respectively.

In each module, several characteristic graphs are connected into a vector in which the growth parameter k and layer parameter l can control the parameters of the dense blocks. For example, the l^{th} layer has an input feature map $k_0 + k * (l-1)$, where k_0 is the number of channel.

3.2 Dual Attention Mechanism

After extracting the visual feature sequence of the image through DenseNet, decoding based on connectionist temporary classification (CTC) is used to generate the output. Each codec combination consists of a BiLSTM encoder and decoder for outputting the context features. The context feature and visual feature sequence V calculated from the DenseNet network are connected in-series, as shown in Fig. 3. The specific process uses CTC decoding for the visual feature sequence V extracted by DenseNet, which can optimize the character representation in the visual feature sequence. Feature sequence V is then fed back by placing a fully-connected layer, which outputs the output sequence H with length N and further inputs it into the CTC module. The CTC module converts the results into a conditional probability module and obtains the most likely tag.

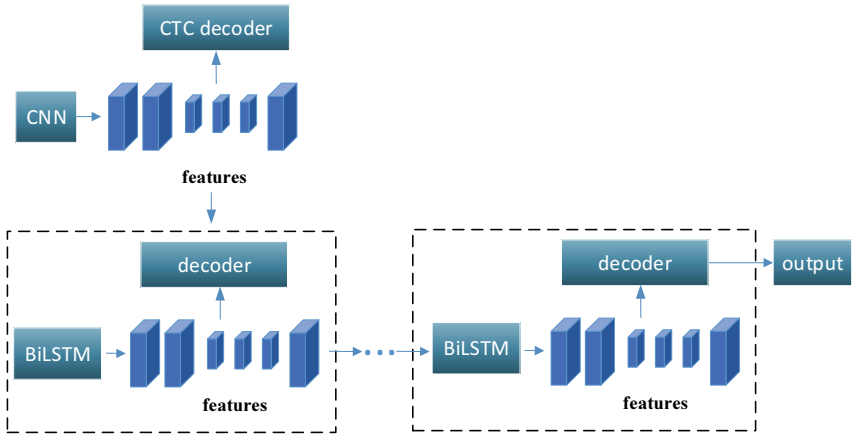

Fig. 3. Continuous encoder-decoder.

To facilitate the use of contextual data, we used the BiLSTM module for reverse data processing to solve the long-term dependence problem and avoid a single long short-term memory (LSTM) by considering the gone data and ignoring the incoming data. BiLSTM represents the forward and backward

data using two independent LSTM layers. In this paper, the segmented license image sequence was labeled to eliminate the need for sequence segmentation. Consequently, CTC was utilized to map each output to the corresponding probability module of all possible label sequences. Finally, the results were obtained through repeated encoding and decoding using the BiLSTM and decoder.

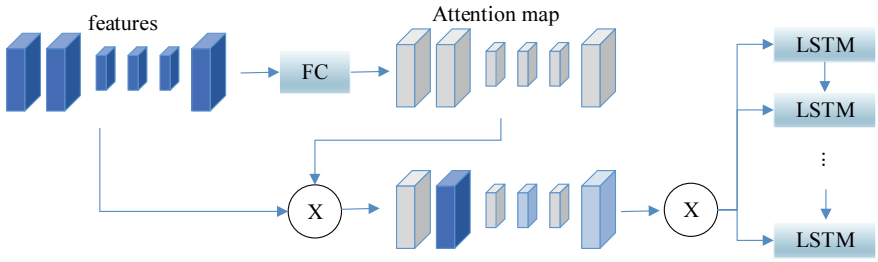

Fig. 4. Dual attention decoder.

As shown in Fig. 4, in each encoder step, the attention map is used twice. First, a one-dimensional operation on the feature graph is applied, and a fully-connected layer is used after these feature graphs to calculate the attention feature graph. The pixel product between the attention and original feature maps is then calculated to generate the attention feature map D'. The decoding of D' is completed using an independent module, and outputs y_t are subsequently obtained.

Given a tobacco retail license image I and an encoder $E(I) = \{h1, h2, \ldots, h_T\}$, during the t^{th} step, the module outputs y_t, i.e.,

$$yt=G(at,bt), \tag{3}$$

where $G(.)$ refers to the feedforward function, a_t is the state at time point t, and b_t is the weighted sum of the sequence eigenvectors, which are defined as follows:

$$a_t = LSTM(y_{t-1}, b_t, a_{t-1}), \tag{4}$$

$$g_t = \sum_{j=1}^{T} \alpha_{t,j}, h_j. \tag{5}$$

Here, $\alpha_t \in R^T$ is the weight vector, which is also known as the alignment factor.

4. Experiments

To evaluate the effectiveness of our recognition approach, a comparative experiment was conducted on a largescale tobacco retail license image dataset, namely, ZY-LQ. The ZY-LQ dataset consists of 527,921 tobacco retail license images with different scales and levels of sharpness, and the license information has different forms of presentation, as shown in Fig. 5. Each license image comprises the license number, license issuing authority, store name, name of the person in-charge, and priority period. In addition, each image is marked with the store name, owner's name, monopoly license number, province, city, and county information.

Fig. 5. Examples from the ZY-LQ dataset.

4.1 Experimental Configuration

The experimental environment was an NVIDIA Quadro P5000 graphics card, 128 GB of running memory, and a 2.30-GHz Intel Xeon Gold 5118 CPU processor. The software environment was an integrated Ubuntu 16 operating system, Python 3.6, and PyTorch 1.0 development environment.

Furthermore, we used cumulative match characteristic (CMC) curves, which are precision curves that provide the recognition precision for each rank, to display the experimental results in a comparative experiment. The x-axis of the CMC curves represented the rank of recognition, whereas the y-axis represented the precision in percentage. In addition, hard and soft indices were utilized to evaluate the effectiveness of image and text recognition, which were defined according to the Levenshtein distance (LD). In this study, the hard index (HI) was used to evaluate the recognition results, which are defined by the target string β_T and recognized string β_R.

$$HI = {}^{N_\beta}/_{N}, \tag{6}$$

where N refers to the number of images for testing, and N_β is the image number that satisfies the equation $LD(\beta_T, \beta_R) = 0$.

4.2 Experimental Results

In a comparative experiment, four existing approaches were used: DenseNet, LSTM, BiLSTM, and CNN. The dataset was divided at a ratio of 8:2 for training and testing, which were randomly selected, and the experimental results are shown in Fig. 6.

Fig. 6 shows that the proposed approach outperforms the other three methods in terms of the correct recognition rate. It can be observed from the figure that when the rank is 20, the correct recognition rate exceeds 95%. From the results, note that the BiLSTM and DenseNet methods achieve the second and

third best recognition results. As a possible reason for this situation, our approach integrates the advantages of BiLSTM and DenseNet models.

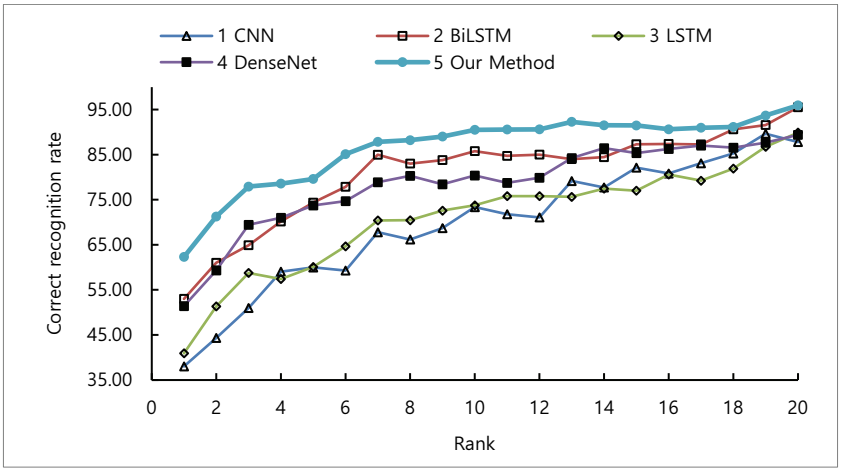

Fig. 6. Comparison of five approaches on the ZY-LQ dataset.

To further discuss the effectiveness of the CTC decoder, attention mechanism, BiLSTM, and DenseBlock on the proposed model, we conducted ablation experiments and discuss the results using the HI. The corresponding experimental results are shown in Tables 1 and 2, which indicate that adding a CTC decoder to the basic DenseNet + LSTM framework improves the accuracy by 0.45%, and by adding an attention module, the accuracy is improved to 98.12%, thereby verifying the effectiveness of such additions. Furthermore, by increasing the number of DenseBlocks to three and the number of BiLSTM layers to four, the results in Table 2 are obtained. The first two lines show that when no CTC decoder is used, the attention module improves the accuracy by 0.66%. When the CTC decoder is added, the accuracy is 98.36%. This shows that using the CTC decoder and attention module simultaneously is more effective. This result occurs because the BiLSTM network can easily learn information, and the appropriate addition of a BiLSTM layer can help consider more contextual information of the sequence and allow better feature information be obtained during the coding process.

Table 1. Results when changing CTC decoder and attention mechanism

Method	DenseBlock	CTC decoder	Attention module	LSTM layers	HI
DenseNet+LSTM	3	0	0	4	94.13
DenseNet+LSTM+CTC	3	1	0	4	95.37
DenseNet+LSTM+Attention	3	0	2	4	98.03
DenseNet+LSTM+CTC+Attention	3	1	2	4	98.12

Table 2. Results for changing BiLSTM and DenseBlock

Method	DenseBlock	CTC decoder	Attention module	LSTM layers	HI
DenseNet+LSTM+Attention	4	0	2	6	96.87
DenseNet+LSTM+Attention	4	0	4	6	97.26
DenseNet+LSTM+CTC+Attention	4	1	2	6	97.95
DenseNet+LSTM+CTC+Attention	4	1	4	6	98.36

5. Conclusion

In this paper, a novel recognition network for tobacco retail license images based on an attention mechanism and a decoder module was presented. DenseNet and BiLSTM were integrated into the proposed model to extract the features and information from tobacco retail license images without segmentation. In addition, a comprehensive evaluation was conducted using a largescale tobacco retail license image dataset. Based on an analysis of the experimental results, the proposed network outperforms most existing approaches.

In addition, our algorithm was only tested on a single dataset. As the main reason for this, to meet the requirements of current applications, the proposed recognition model was optimized according to the characteristics of the tobacco retail license image data. If other largescale data that can be used to test the proposed algorithm become available in the future, we might extend the proposed approach to other similar applications.

Acknowledgement

This work in this paper was supported by the Research on Key Technology and Application of Marketing Robot Process Automation (RPA) Based on Intelligent Image Recognition in Zhejiang China Tobacco Industry Co. Ltd. (No. ZJZY2021E001).

References

[1] M. Deng, Z. Li, Y. Kang, C. P. Chen, and X. Chu, "A learning-based hierarchical control scheme for an exoskeleton robot in human–robot cooperative manipulation," *IEEE Transactions on Cybernetics*, vol. 50, no. 1, pp. 112-125, 2020.

[2] A. Ravendran, M. Bryson, and D. G. Dansereau, "Burst imaging for light-constrained structure-from-motion," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 1040-1047, 2022.

[3] Y. Zhang, S. Nie, S. Liang, and W. Liu, "Robust text image recognition via adversarial sequence-to-sequence domain adaptation," *IEEE Transactions on Image Processing*, vol. 30, pp. 3922-3933, 2021.

[4] Y. S. Chernyshova, A. V. Sheshkus, and V. V. Arlazarov, "Two-step CNN framework for text line recognition in camera-captured images," *IEEE Access*, vol. 8, pp. 32587-32600, 2020.

[5] Z. Ou, B. Xiong, F. Xiao, and M. Song, "ERCS: an efficient and robust card recognition system for camera-based image," *China Communications*, vol. 17, no. 12, pp. 247-264, 2020.

[6] A. Bera, Z. Wharton, Y. Liu, N. Bessis, and A. Behera, "Attend and guide (AG-Net): a keypoints-driven attention-based deep network for image recognition," *IEEE Transactions on Image Processing*, vol. 30, pp. 3691-3704, 2021.

[7] R. Islam, M. R. Islam, and K. H. Talukder, "An efficient ROI detection algorithm for Bangla text extraction and recognition from natural scene images," *Journal of King Saud University-Computer and Information Sciences*, 2022. https://doi.org/10.1016/j.jksuci.2022.02.001

[8] Q. Lai, S. Khan, Y. Nie, H. Sun, J. Shen, and L. Shao, "Understanding more about human and machine attention in deep neural networks," *IEEE Transactions on Multimedia*, *23*, 2086-2099, 2020.

[9] Z. Luo, J. Li, and Y. Zhu, "A deep feature fusion network based on multiple attention mechanisms for joint iris-periocular biometric recognition," *IEEE Signal Processing Letters*, vol. 28, pp. 1060-1064, 2021.

Yuxiang Shan **https://orcid.org/0000-0002-4703-924X**

He received his M.S. degree in School of Computer Science and Technology from Zhejiang University in 2013. He is now an engineer of Information Center of China Tobacco Zhejiang Industrial Co. Ltd. His current research interests include image recognition and artificial intelligence.

Qin Ren **https://orcid.org/0000-0002-2911-0647**

She received a bachelor's degree in marketing from Zhejiang Normal University in 2010. Since August 2011, she has worked in China Tobacco Zhejiang Industrial Co. Ltd., engaged in Tobacco Marketing and Internet Marketing Research, respectively.

Cheng Wang **https://orcid.org/0000-0002-8620-6717**

He received his B.S. degree in School of Human Resources Management from Nanjing Audit University in 2010. Since then, he joined Zhejiang Tobacco Industry Company as custom manager. In 2020, he joined the brand operation department, engaged in data operation and customer operation.

Xiuhui Wang **https://orcid.org/0000-0003-1773-9760**

He received his master's degree and doctor's degree from Zhejiang University in 2003 and 2007, respectively. He is now a professor in the Computer Department of China Jiliang University. His current research interests include computer graphics, pattern recognition and artificial intelligence.