2020 25th International Conference on Pattern Recognition (ICPR)
Milan, Italy, Jan 10-15, 2021

A Transformer-based Radical Analysis Network for Chinese Character Recognition

Chen Yang, Qing Wang, Jun Du*, Jianshu Zhang, Changjie Wu, Jiaming Wang
National Engineering Laboratory for Speech and Language Information Processing
University of Science and Technology of China, Hefei, Anhui, P. R. China
Email: yc24@mail.ustc.edu.cn, xiaosong@mail.ustc.edu.cn, ✉jundu@ustc.edu.cn
xysszjs@mail.ustc.edu.cn, wucj@mail.ustc.edu.cn, jmwang66@mail.ustc.edu.cn

***Abstract*—Recently, a novel radical analysis network (RAN) has the capability of effectively recognizing unseen Chinese character classes and largely reducing the requirement of training data by treating a Chinese character as a hierarchical composition of radicals rather than a single character class. However, when dealing with more challenging issues, such as the recognition of complicated characters, low-frequency character categories, and characters in natural scenes, RAN still has a lot of room for improvement. In this paper, we explore options to further improve the structure generalization and robustness capability of RAN with the Transformer architecture, which has achieved start-of-the-art results for many sequence-to-sequence tasks. More specifically, we propose to replace the original attention module in RAN with the transformer decoder, which is named as a transformer-based radical analysis network (RTN). The experimental results show that the proposed approach can significantly outperform the RAN on both printed Chinese character database and natural scene Chinese character database. Meanwhile, further analysis proves that RTN can be better generalized to complex samples and low-frequency characters, and has better robustness in recognizing Chinese characters with different attributes.**

***Index Terms*—Chinese characters, Transformer network, Encoder-decoder, Attention**

I. INTRODUCTION

Chinese character recognition (CCR) is an important task in computer vision with numerous applications. Benefited from the development of deep learning technologies [1], great progress has been made for OCR in recent years.

As for the recent researches, we can roughly divide them into two categories: character-based methods and radical-based methods. The character-based methods [2], [3] take the character input as a single image and try to learn a congruent relationship between the input image and a pre-defined class, which do not consider the similarity and internal structures among different characters. Thus, these methods can only perform well on common Chinese characters but perform poorly when handling complex characters with complicated 2D structures between radicals. In contrast, radical-based methods allow Chinese characters to be decomposed into a composition of radicals and spatial structures, which is more similar to human learning behavior for recognizing characters and enable models to be more generative and flexible. [4] was the first attempt to detect separate radicals and employed a hierarchical radical matching method for recognizing a Chinese character. [5] studied how to obtain position-dependent radicals with multi-labeled learning. Unfortunately, these approaches still have difficulty in dealing with the unseen or newly created characters because they do not take full advantage of the hierarchical information in Chinese characters.

To address the above issues, [6] proposed a novel radical analysis network by adopting a coverage based spatial attention model under the encoder-decoder architecture [7]. Benefited from the effective attention mechanism, RAN can adaptively focus on the most relevant component of a Chinese character to describe a radical. Meanwhile, it can also detect the relative spatial relationships among radicals. Hence, RAN has the ability to recognize unseen Chinese characters and largely reduces the size of vocabulary. However, to recognize more complicated radical structures or learn the composition rules of low-frequency samples, RAN still has a lot of room for improvement. Intuitively, the Transformer architecture [8] as the first sequence transduction model based entirely on self-attention has the potential to further improve the capability of RAN. Recently, several studies [9], [10] have been proposed to scene text recognition task by adopting transformer network and achieved start-of-the-art results. Different from them, our work focuses on recognizing Chinese characters, which contain more complicated spatial structures and more diversity than English words. Furthermore, we attempt to figure out the distinct properties of the Transformer for Chinese character recognition.

Compared with RNN-based attention model that is used in RAN, we believe that the Transformer architecture can generate double effects for Chinese character recognition. On the one hand, different from RNN which relies on the recurrent connection from the previous hidden state to the current hidden state to propagate contextual information, the Transformer based on self-attention allows each element within the input sequence to make a connection and find out where they should pay more attention to, which means that the detailed internal pattern and long-range dependencies within the sequence can be captured easily. Therefore, the Transformer can better learn the internal composition rules than RAN as many internal radical structures are shared among different Chinese characters. On the other hand, RNN-based attention model usually just learns the alignment between the source input and the target output once, which leads to the problem of lacking deep

978-1-7281-8808-9/20/$31.00 ©2020 IEEE

2020 25th International Conference on Pattern Recognition (ICPR) | 978-1-7281-8808-9/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICPR48806.2021.9412439

interaction. However, the Transformer is composed of stacked blocks and aggregates the input context for each block, which naturally provides us with more hierarchical representations corresponding to the hierarchical radical structures in the Chinese characters. Accordingly, the Transformer has the potential to improve the alignment of attention.

Insights from the above ideas, we proposed a simple yet effective Transformer-based radical analysis network (RTN), which is an improved version of RAN. For encoder, we employ a fully convolutional network to extract high-level visual features. For decoder, we substitute the transformer decoder for the coverage-based attention model of RAN. The contributions of this paper are as follows:

- We explore the option to improve the capability of RAN by employing the Transformer architecture.
- The proposed model achieves significant performance improvements on both printed Chinese character database and natural scene Chinese character database.
- Further analysis proves that RTN is more effective and robust than RAN for recognizing complicated and low-frequency samples.
- Attention analysis with the proposed model shows the advantages of RTN in learning the alignment between the input image and the output symbols.

II. THE PROPOSED APPROACH

In this section, we will describe the architecture of our RTN model in detail. As illustrated in Fig. 1, it contains two components: (1) a dense encoder which takes the image as input to produce a fixed-length context vector; (2) a transformer decoder which takes the context vector as input to generate a variable-length symbol sequence.

A. Dense encoder

We first employ dense convolutional network (DenseNet) [11] as the encoder to extract high-level visual features from images, which is widely adopted in a variety of computer vision tasks. Although some works [12], [13] probed the ability of applying self-attention to process the pixel-level input, the actual performance and computational efficiency could not strongly exceed advanced CNN-based networks. Instead of adding a fully connected layer after the final convolution layer, the dense encoder contains only convolution, pooling and activation layers. Therefore, we can obtain a three-dimensional tensor of size $D \times H \times W$. Then we reshape the array to a two-dimensional tensor of size $D \times L$, where $L = H \times W$, followed by a linear transformation. Each element in array is a D-dimensional vector that corresponds to a local region of the image:

$$\mathbf{A} = \{\mathbf{a}_1, \ldots, \mathbf{a}_L\}, \mathbf{a}_l \in \mathbb{R}^D \tag{1}$$

B. Transformer-based decoder

1) Multi-Head Attention: Before presenting our decoder network, we first introduce the multi-head attention mechanism based on the scaled dot-product function, which is one of the core ideas of the Transformer. Generally, we can describe

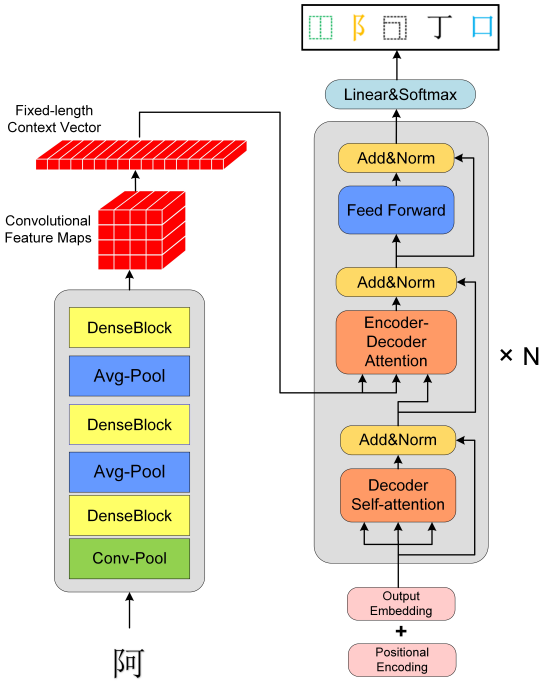

Fig. 1. Architecture of the transformer-based radical analysis network (RTN) for Chinese character recognition. $\times N$ means a stack of N identical blocks.

an attention function as mapping a query and a set of key-value pairs to an output, where the query, keys, values, and output are all vectors. In particular, the scaled dot-product function is defined as:

$$\text{Attn}(\mathbf{q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V} \tag{2}$$

where $\mathbf{q} \in \mathbb{R}^{1\times d_k}$, $\mathbf{K} \in \mathbb{R}^{m\times d_k}$ and $\mathbf{V} \in \mathbb{R}^{m\times d_v}$. We calculate the dot products of each query with m keys, divide each by $\sqrt{d_k}$ and apply a softmax function to obtain the attention weights on the values as the attended feature. For simplicity, d_k and d_v are usually set to the same number d.

To further enhance the representation capacity of the attended features, multi-head attention (MA) is introduced as an extension of the single self-attention, which consists of h paralleled "head". Each head corresponds to an independent scaled dot-product attention function. Therefore. the multi-head attention can be expressed in the same notation as Equation 2:

$$\text{MA}(\mathbf{q}, \mathbf{K}, \mathbf{V}) = [\text{head}_1, \text{head}_2, \ldots, \text{head}_h]\,\mathbf{W}^O \tag{3}$$

$$\textit{where} \quad \text{head}_j = \text{Attn}\left(\mathbf{q}\mathbf{W}_j^Q, \mathbf{K}\mathbf{W}_j^K, \mathbf{V}\mathbf{W}_j^V\right) \tag{4}$$

where $\mathbf{W}_j^q, \mathbf{W}_j^K, \mathbf{W}_j^V \in \mathbb{R}^{d\times d_h}$ are the projection matrices for the j^{th} head, and $\mathbf{W}^O \in \mathbb{R}^{h*d_h\times d}$. d_h is the dimensionality of the output features from each head, which is usually set to $d_h = d/h$, and $[*]$ is the concatenation operation.

2) Decoder: To be specific, the transformer decoder is composed of stacked identical blocks. Each block contains three layers: (a) decoder self-attention layer; (b) encoder-decoder attention layer; (c) feed-forward layer. This hybrid neural network will analyze the internal hierarchical radical structures of Chinese characters and generate a corresponding sequence of spatial structures and radicals. We can denote the target sequence as $\mathbf{Y}$, which is a sequence of one-hot vectors.

$$\mathbf{Y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_T\}, \mathbf{y}_i \in \mathbb{R}^K \tag{5}$$

where K is the number of total words in the vocabulary which includes the basic radicals and spatial structures, and T is the length of the target sequence.

Given a source image $\mathbf{X}$ and a target sequence $\mathbf{Y}$, our model can be defined as the following conditional probability:

$$\mathrm{P}(\mathbf{Y}|\mathbf{X}) = \prod_{i=1}^{|\mathbf{Y}|} \mathrm{P}\left(\mathbf{y}_i|\mathbf{Y}_{<i}, \mathbf{X}\right) = \prod_{i=1}^{|\mathbf{Y}|} \mathrm{P}\left(\mathbf{y}_i|\mathbf{c}_i^N\right) \tag{6}$$

where $\mathbf{Y}_{<i} = \{\mathbf{y}_1, \ldots, \mathbf{y}_{i-1}\}$ denotes a prefix of $\mathbf{Y}$ with length $i-1$, $\mathbf{c}_i^N$ denotes the N^{th} block context in the decoder with N blocks which is obtained from the representation of $\mathbf{Y}_{<i}$ and $\mathbf{A}$. Then, we will describe how to produce the output $\mathbf{c}_i^n$ for each block by the above mentioned layers.

The first two layers both rely on multi-head attention mechanism. In decoder self-attention, all of the keys, values and queries come from the output of the previous decoder layer. It performs as correlation calculations between the internal elements of the target sequence. We define this inter-interaction result at the i^{th} position of n^{th} block as $\mathbf{s}_i^n \in \mathbb{R}^{1\times d}$:

$$\mathbf{s}_i^n = \mathrm{MA}(\mathbf{c}_i^{n-1}, \mathbf{C}_{<i}^{n-1}, \mathbf{C}_{<i}^{n-1}) \tag{7}$$

where $\mathbf{c}_i^{n-1} \in \mathbb{R}^{1\times d}$ and $\mathbf{C}_{<i}^{n-1} = \left\{\mathbf{c}_1^{n-1}, \mathbf{c}_2^{n-1}, \cdots, \mathbf{c}_{i-1}^{n-1}\right\}$ correspond to $\mathbf{q}$, $\mathbf{K/V}$ respectively. In particular , $\mathbf{C}_{<i}^0$ is word embedding of $\mathbf{Y}_{<i}$. And a triangular mask matrix with $-\infty$ entries on the disabled position and 0 otherwise, is used to preserve the auto-regressive property.

Different from the decoder self-attention, encoder-decoder attention receives the output of the encoder, i.e., the fixed-length context vector $\mathbf{A}$ as the keys/values and the query is $\mathbf{s}_i^n$ instead. It aims to capture complex relationships and dependencies between the source and target sequence.

$$\mathbf{z}_i^n = \mathrm{MA}(\mathbf{s}_i^n, \mathbf{A}, \mathbf{A}) \tag{8}$$

where $\mathbf{s}_i^n \in \mathbb{R}^{1\times d}$ and $\mathbf{A}$ correspond to $\mathbf{q}$, $\mathbf{K/V}$ respectively.

Then, the feed-forward layer projects the input dimension d to another feature space d_{ff} and then back to d for learning better feature representation.

$$\mathbf{c}_i^n = \max\left(0, \mathbf{W}_1^n \mathbf{z}_i^n + \mathbf{b}_1^n\right)\mathbf{W}_2^n + \mathbf{b}_1^n \tag{9}$$

where $\mathbf{W}_1^n \in \mathbb{R}^{d\times d_{ff}}$,$\mathbf{W}_2^n \in \mathbb{R}^{d_{ff}\times 1}$.

Besides, since the self-attention ignores the order information of a sequence, a positional embedding PE is used to represent the positional information.

$$\mathrm{PE}_{(p,2k)} = \sin\left(p/10000^{2k/d}\right) \tag{10}$$

$$\mathrm{PE}_{(p,2k+1)} = \cos\left(p/10000^{2k/d}\right) \tag{11}$$

where p is the position and k is the dimension. To better optimize the deep network, the entire network uses a residual connection and applies "Add & Norm" to the layer [14], [15].

Finally, we also apply a fully connected layer and a softmax activation to predict the target symbol.

III. EXPERIMENTS

In order to make a comprehensive comparison between RAN and RTN, we conduct experiments on both printed Chinese character dataset and natural scene dataset by answering the following questions:

- Is the RTN more effective and robust in recognizing unseen printed Chinese characters?
- Is the RTN more effective and robust in recognizing Chinese characters in the wild?
- What are the advantages of the attention mechanism of RTN?

A. Training and testing details

The training model used in all experiments has the same configuration as follows. For the dense encoder, we first employ a 7×7 convolution layer with 48 output channels before entering the first dense block. Each DenseBlock contains 22 1×1 convolution layers and 22 3×3 convolution layers. Then, we use 1×1 convolution followed by 2×2 average pooling as transition layer to reduce the feature maps by half between every two DenseBlocks. And the growth rate is 24. We also use batch normalization [16] after each convolution layer and the activation function is ReLU. To perform a fair comparison, the encoder of our implemented RAN also employs the dense encoder instead of VGG [17] encoder.

For the Transformer decoder, it consists of 6-layer with 256 embedding/hidden size and 512 feed-forward intermediate size, i.e., $d = 256$, $d_{ff} = 512$. The multi-head number is set to 8 both in decoder self-attention block and encoder-decoder attention block. For optimization, we use the cross entropy loss and the AdaDelta algorithm [18] with the hyperparameters of $lrate = 0.1$, $\rho = 0.95$ and $\varepsilon = 10^{-4}$. Following the implementation of tensor2tensor[1], we put layer normalization inside the residual connection in practice, which is considered to be more conducive to model convergence. The experiments are all implemented with Pytorch 1.2.0 and an NVIDIA Tesla 1080Ti 11G GPU.

B. Experiments on single-font printed Chinese characters

In this section, we show the effectiveness and robustness of RTN on printed Chinese characters dataset. We choose 27,533 Chinese characters in the GB18030 standard with four common used font style (Song, Fangsong, Hei and Kaiti) as dataset, which are composed of 500 radicals and 12 spatial structures. And all characters are annotation in the ideographic description sequence (IDS) format. For dataset splitting, we choose 10,000 characters as the training set, 2,000 characters

[1]https://github.com/tensorflow/tensor2tensor

Comparison of accuracy rate bewteen RTN and RAN with different caption lengths on different font-style unseen Chinese characters respectively.

Font Style	RAN(%)			RTN(%)		
	ALL	≤ 6	> 6	ALL	≤ 6	> 6
Song	92.21	93.65	90.78	**94.54**	**94.93**	**94.16**
FangSong	91.04	91.98	90.11	**94.21**	**94.84**	**93.57**
Hei	90.41	91.34	89.50	**92.79**	**92.41**	**93.11**
Kaiti	88.57	90.59	86.58	**91.31**	**92.96**	**89.67**

as the validation set and the remaining 15,533 characters as the testing set. A test character is considered as successfully recognized only when its predicted caption exactly matches the ground-truth. The input image size is set to 32×32. Note that our dictionary and dataset splitting are different with [6], which lead to the results being not fully comparable.

1) Accuracy versus caption lengths: In order to figure out whether the RTN model is more robust than RAN for recognizing complicated samples, we divide the original test set into three parts by the length of caption, i.e., the length of the IDS sequence, represented by ALL, ≤ 6 and > 6 in Table III-B. ≤ 6 indicates the caption length of characters are less than or equal to 6; > 6 indicates the caption length of characters are greater than 6; and ALL includes all characters in the testing set. We choose 6 as the boundary so that both sections contain approximately 7,760 samples. Since the longer IDS sequences are usually related to more complicated structures and radicals, the difficulty of recognition is increasing. However, we achieve the highest accuracy improvement on > 6 sub-set for all datasets with different font styles. At the same time, RTN also can consistently outperform RAN on the other two sub-set. Here we employ the ensemble method to sustain stable performance during testing procedure followed by [6].

2) Experiments on low-frequency radical categories: Although RAN effectively alleviated the problem of recognizing low-frequency Chinese characters by a compact set of radicals and spatial structures learned in the training stage, some low-frequency radicals still yield the dramatic decline of the recognition accuracies for few-/zero-shot learning. Thus, in order to evaluate the generalization ability on recognizing characters with low-frequency radicals, we generate five testing sub-set according to the number of times that a certain radical appears in the training set and approximately 100 radical categories are included in each sub-set. As shown in Fig. 2, when the appearance frequency of radicals is greater than 10 times, the recognition accuracy gaps between RTN and RAN are 1.35%, 1.67%, 2.42%, 2.44% respectively. When the appearance frequency of radicals is lower than 10 times, the performance of RTN still maintains a high accuracy (92.27%), whereas the accuracy of RAN has decreased to 82.60%. Therefore, we can believe that the self-attention mechanism improves the ability of the model to learn the rules of composing characters with low-frequency radicals during the training phase.

C. Experiment on natural scene Chinese characters

In this section, we evaluate our model in a natural scene dataset, namely *Chinese Text in the Wild* (CTW) [19], which

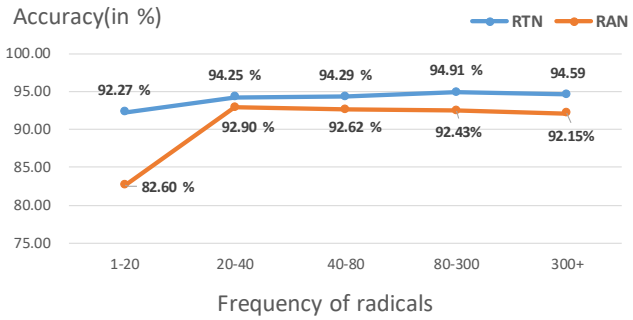

Fig. 2. Comparison of character-level accuracy between RTN and RAN with respect to the frequency of radicals. Approximately 100 radical categories are included in each range.

TABLE I
COMPARISON OF THE RECOGNITION PERFORMANCE OF RTN AND RAN WITH DIFFERENT CAPTION LENGTHS ON THE CTW VALID DATABASE

Model	Caption length		
	ALL	≤ 4	> 4
RAN	85.95%	89.33%	82.31%
RTN	**87.51%**	**90.02%**	**84.80%**
Accuracy↑	**1.56%**	0.69%	**2.49%**

contains about 1 million Chinese characters in over 30,000 street view images. Compared with the printed Chinese character dataset, CTW dataset is more challenging due to its diversity and complexity, which is beneficial for reflecting the difference in robustness between RAN and RTN. Following the official dataset splitting, we use 3,580 Chinese character categories with 760,107 instances for training, 2,015 Chinese character categories with 52,765 instances for validation and 103,519 instances for testing. Because the official testing set is not released, some analysis are conducted on validation set. Note that all input images are first resized to 32×32 resolution.

1) Accuracy versus caption lengths: Similar to the experiment of Section III-B1, we also demonstrate the accuracies of RAN and RTN with different caption lengths in Table I. As for the caption length is lower than 4, the improvement is only 0.69%. The majority of performance gains come from the > 4 part, which indicates that RTN is also more effective than RAN to recognize complicated characters in natural scenes.

2) Experiment on low-frequency character categories: In this subsection, we compare the performance of the RTN and RAN in overcoming the sample sparsity problem. Different from the Section III-B2, here we are concerned with the effects of character-level frequency in recognition. Specifically, we divide all testing Chinese character categories into 4 subsets based on the appearance frequency in the training set. In Table II, < 20 indicates character categories that appear fewer than 20 times while < 50 and < 100 indicate character categories that appear fewer than 50 and 100 times respectively; HF means high frequency and includes character categories that appear more than 100 times; and ALL includes all character categories in the testing set. The results show that

TABLE II
COMPARISON OF THE PERFORMANCE OF RAN AND RTN WITH THE DIFFERENT APPEARANCE FREQUENCY OF CHARACTER-LEVEL CATEGORIES.

Frequency	≤ 20	≤ 50	≤ 100	HF	ALL
Categories	398	511	335	1044	2015
Samples	1128	1229	1663	48745	52765
RAN	25.88%	47.92%	65.12%	89.01%	85.95%
RTN	**41.84%**	**61.51%**	**71.67%**	**89.76%**	**87.51%**

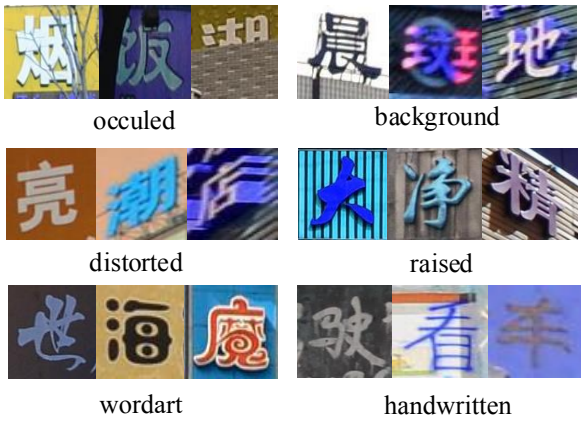

Fig. 3. Examples with 6 different attributes in the CTW database.

RTN is much better than RAN to deal with low-frequency characters (< 20 and < 50), which both achieve over 14% absolute increase on recognition accuracy. By contrast, the improvement in the high frequency characters set is only 0.75%. The results of this experiment support the idea that RTN can further enhance the few-/zero-shot learning ability of RAN in natural scenes.

3) Analysis of robustness: To demonstrate the robustness of RAN in natural scenes, Table III shows the performance of RTN and RAN with respect to 6 attributes: occluded, complex background, distorted, 3D raised, wordart characters and handwritten characters. Character examples of each attribute are illustrated in Fig. 3. We can observe that the recognition accuracies of RTN are much better than RAN for all attributes, especially in occluded, distorted and handwritten. The occlusion attribute indicates whether the character is occluded by other objects. To successfully identify characters with occlusion attribute, the ability to predict the missing radicals just relying on the potential language model is required. We can believe that the ability of the decoder self-attention to pay attention to every available caption symbol offers the improvement. The distortion attribute indicates whether the character is distorted, rotated or frontal. The considerable improvement on recognizing distorted characters shows that the transformer decoder is powerful to realize the spatial attention, despite that no explicit spatial transformer module is equipped in network unlike [20]. The handwritten attribute is one of the most difficult categories to recognize due to the confusion between similar characters, distinct handwriting

TABLE III
COMPARISON OF THE RECOGNITION PERFORMANCE OF RAN AND RTN WITH RESPECT TO 6 ATTRIBUTES ON THE CTW TEST DATASET; ALL INCLUDES ALL CHARACTERS ON THE TEST DATABASE.

Attributes	**Training Samples**	**RAN(%)**	**RTN(%)**
All	760107	85.56	87.31
occluded	101393	71.55	73.94
background	218560	82.84	84.57
distorted	192481	71.55	83.60
3D raised	199066	76.17	78.06
wordart	65983	87.11	84.25
handwritten	6661	63.58	66.70

styles and missing internal structures. Nevertheless, RTN can improve the recognition accuracy of handwritten characters by nearly 12%. Thus, the stronger structure generalization and robustness capability of RTN is proved.

D. Visualization analysis

In this section, we illustrate how the proposed model further improves the ability of RAN through attention visualization. Although the coverage-based attention model [21]–[23] used in RAN is already very effective to address the over-parsing and under-parsing problems for Chinese character recognition, it is still difficult to parse all the radicals correctly for some extremely complicated characters. In Fig. 4, we demonstrate the learning process of RAN and RTN to recognize a complicated unseen Chinese character step by step, where the red color represents the encoder-decoder attention probability in the attention maps. We can observe that RAN only predicts two radicals wrong in step 7-8 where the two same radicals have different spatial locations, but in fact, there are several other alignment errors existing between the image and the internal radicals. Benefited from the powerful attention mechanism of the Transformer, RTN can fully identify the correct character caption and aligns the most relevant visual areas accurately. In addition, misjudging the internal spatial structures among the radicals is also one of the main reasons that why RAN fails to predict caption exactly when dealing with complex samples. We also use a case study to show this phenomenon in Fig. 5. To detect the detailed spatial structures such as the top-left-surround structure, RTN performs better than RAN.

IV. CONCLUSION

In this paper, we propose a transformer-based radical analysis network to recognize Chinese characters. From the quantitative results, we achieve significant improvements on both printed Chinese character database and natural scene database compared with RAN. Furthermore, by analyzing the variation of recognition accuracy with complicated and low-frequency character categories, we prove that RTN obtains stronger generalization and robustness. In future work, we plan to evaluate the zero-shot ability of recognizing scene text or handwritten characters and investigate RTN's ability for Chinese text line recognition. We will also explore the options to design more task-specific transformer-based model and provide more interpretability for attention mechanism.

Fig. 4. Attention visualization of recognizing a complicated Chinese character step by step; the above is the process of RAN and the below is the process of RTN; symbols below the images are the predicted radicals or structures.

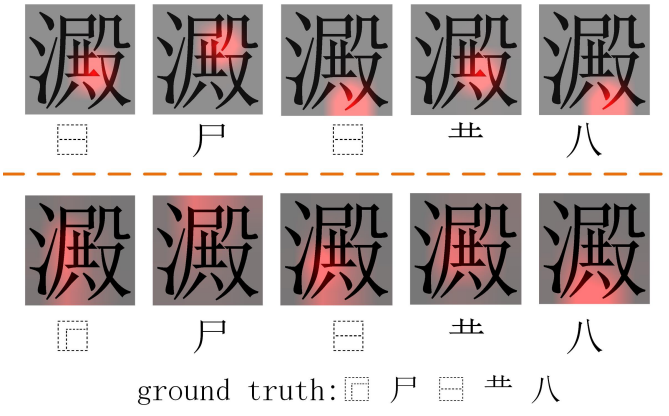

Fig. 5. Comparison of RAN and RTN on aligning the internal spatial structures; the above is the partial process of RAN and the below is the partial process of RTN.

ACKNOWLEDGMENT

This work was supported in part by the National Key R&D Program of China under contract No. 2017YFB1002202, the National Natural Science Foundation of China under Grants No. 61671422 and U1613211, the Key Science and Technology Project of Anhui Province under Grant No. 17030901005. This work was also funded by Tencent.

REFERENCES

[1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.
[2] Z. Zhong, L. Jin, and Z. Feng, "Multi-font printed chinese character recognition using multi-pooling convolutional neural network," in *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 2015, pp. 96–100.
[3] X.-Y. Zhang, Y. Bengio, and C.-L. Liu, "Online and offline handwritten chinese character recognition: A comprehensive study and new benchmark," *Pattern Recognition*, vol. 61, pp. 348–360, 2017.
[4] L.-L. Ma and C.-L. Liu, "A new radical-based approach to online handwritten chinese character recognition," in *2008 19th International Conference on Pattern Recognition*. IEEE, 2008, pp. 1–4.
[5] T.-Q. Wang, F. Yin, and C.-L. Liu, "Radical-based chinese character recognition via multi-labeled learning of deep residual networks," in *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, vol. 1. IEEE, 2017, pp. 579–584.
[6] J. Zhang, Y. Zhu, J. Du, and L. Dai, "Radical analysis network for zero-shot learning in printed chinese character recognition," in *2018 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2018, pp. 1–6.
[7] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.
[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.
[9] P. Lyu, Z. Yang, X. Leng, X. Wu, R. Li, and X. Shen, "2d attentional irregular scene text recognizer," *arXiv preprint arXiv:1906.05708*, 2019.
[10] L. Yang, P. Wang, H. Li, Y. Gao, L. Zhang, C. Shen, and Y. Zhang, "A simple and strong convolutional-attention network for irregular text recognition," *arXiv preprint arXiv:1904.01375*, 2019.
[11] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.
[12] R. Child, S. Gray, A. Radford, and I. Sutskever, "Generating long sequences with sparse transformers," *arXiv preprint arXiv:1904.10509*, 2019.
[13] N. Parmar, A. Vaswani, J. Uszkoreit, Ł. Kaiser, N. Shazeer, A. Ku, and D. Tran, "Image transformer," *arXiv preprint arXiv:1802.05751*, 2018.
[14] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.
[15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
[16] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.
[17] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
[18] M. D. Zeiler, "Adadelta: an adaptive learning rate method," *arXiv preprint arXiv:1212.5701*, 2012.
[19] T.-L. Yuan, Z. Zhu, K. Xu, C.-J. Li, T.-J. Mu, and S.-M. Hu, "A large chinese text dataset in the wild," *Journal of Computer Science and Technology*, vol. 34, no. 3, pp. 509–521, 2019.
[20] C. Wu, Z.-R. Wang, J. Du, J. Zhang, and J. Wang, "Joint spatial and radical analysis network for distorted chinese character recognition," in *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, vol. 5. IEEE, 2019, pp. 122–127.
[21] Z. Tu, Z. Lu, Y. Liu, X. Liu, and H. Li, "Modeling coverage for neural machine translation," *arXiv preprint arXiv:1601.04811*, 2016.
[22] J. Zhang, J. Du, S. Zhang, D. Liu, Y. Hu, J. Hu, S. Wei, and L. Dai, "Watch, attend and parse: An end-to-end neural network based approach to handwritten mathematical expression recognition," *Pattern Recognition*, vol. 71, pp. 196–206, 2017.
[23] J. Zhang, J. Du, and L. Dai, "Track, attend, and parse (tap): An end-to-end framework for online handwritten mathematical expression recognition," *IEEE Transactions on Multimedia*, vol. 21, no. 1, pp. 221–233, 2018.