Radical Analysis Network for Learning Hierarchies of Chinese Characters

Journal:	*Transactions on Pattern Analysis and Machine Intelligence*
Manuscript ID	TPAMI-2018-10-0950
Manuscript Type:	Regular
Keywords:	Radical, L.1.0.a Attention < L.1.0 Touch-based properties and capabilities of the human user < L.1 Human Haptics < L Haptics, Chinese character, Few-/zero-shot learning

Radical Analysis Network for Learning Hierarchies of Chinese Characters

Jianshu Zhang, Jun Du, and Lirong Dai

Abstract—Chinese characters have a valuable property, that is, numerous Chinese characters are composed of a compact set of fundamental and structural radicals. This paper introduces a radical analysis network (RAN) that makes full use of this valuable property to implement radical-based Chinese character recognition. The proposed RAN employs an attention mechanism to extract radicals in Chinese characters and to detect spatial structures among radicals. Then, the decoder in RAN generates a hierarchical composition of Chinese characters based on the knowledge of the extracted radicals and their internal structures. The method of treating a Chinese character as a composition of radicals rather than as a single character category is more human-like and can reduce the size of the vocabulary, ignore redundant information among similar characters and enable the system to recognize unseen Chinese character categories, namely, zero-shot learning. In experiments, we assess the practicality of RAN for recognizing Chinese characters in natural scenes not just printed characters. Furthermore, an RAN framework can be proposed for scene text recognition with the extension of a dense recurrent neural network (denseRNN) encoder, a multihead coverage attention model and HSV representations, and the proposed approach achieved the best performance in the ICPR MTWI 2018 competition.

Index Terms—Radical, Attention, Chinese character, Few-/zero-shot learning.

1 Introduction

Automatic recognition of text in Chinese has considerable commercial value and social benefits, as Chinese characters are among the most widely adopted reading systems in the world: nearly one quarter of the world's population reads and writes in Chinese scripts. However, as Tang et al. [1] described, Chinese character recognition is among the most challenging topics in pattern recognition since it involves a large number of characters with complex internal structures, confusion among similar characters and an increasing number of novel characters.

Recent deep-learning-based approaches [2], such as convolutional neural network (CNN) [3], [4] and recurrent neural network (RNN) [5], [6], have achieved considerable success in the recognition of approximately 4,000 commonly used Chinese characters [7], [8]. For example, since offline characters are naturally represented as scanned images, CNN is a natural and effective method for offline Chinese character recognition [9], [10], [11], as the strong a priori knowledge of convolution makes the CNN a powerful model for image classification. With respect to online recognition, pen movements (xy-coordinates) are stored as sequential data, which can be naturally processed by the RNN [12]. Moreover, a CNN can be applied to online characters by first transforming the online handwriting trajectory into image-like representations, such as the AMAP [13], path signature maps [14], [15] and directional feature maps [16]. The above methods can be viewed as character-based Chinese character recognition techniques since they treat characters as whole graphs.

However, as Sampson [17] writes, "to ask how many graphs there are in Chinese script is like asking how many words there are in the English language, and this is not a question with a well-defined answer." A large number of Chinese characters exists besides the 4,000 commonly used characters, and the numerous categories result in difficulties for recognition. Additionally, recognition of rarely used characters is typically a few-shot learning problem since samples of such rarely used character categories are difficult to collect, leading to few training samples. Moreover, the recognition of some novel Chinese characters (e.g., the character "Duang", newly created by Jackie Chan, see Fig. 9) is a zero-shot learning problem as these characters are newly created and have never seen them before. Therefore, no training samples exist for such characters. Few-/zero-shot learning has attracted the interest of researchers [18], [19]. This type of learning is a challenging problem but has enormous potential value, as the ability to learn and generalize from a few examples is a hallmark of human intelligence [20] that is difficult to achieve in deep learning methods.

In this paper, we propose a novel deep-learning-based model, called the radical analysis network (RAN), for Chinese character recognition. By exploiting the distinct properties of Chinese characters, which are intensely hierarchical, RAN achieves the ability of few-/zero-shot learning. In contrast to English or Arabic characters, Chinese characters are composed of basic components [17] (called radicals in this paper). Only a small number of radicals can be used to construct many types of Chinese characters [21]. Therefore, an intuitive method for Chinese character recognition is to decompose Chinese characters into radicals and analyze the hierarchical structures among radicals. RAN, which is a radical-based method, has two distinct properties compared with character-based methods: 1) the size of the radical vo-

- *Jianshu Zhang, Jun Du and Lirong Dai were with the National Engineering Laboratory for Speech and Language Information Processing, University of Science and Technology of China, Hefei, Anhui, P. R. China. E-mail: xysszjs@mail.ustc.edu.cn, jundu@ustc.edu.cn, lrdai@ustc.edu.cn*
- *Corresponding author: Jun Du.*

cabulary is largely reduced compared to that of the character vocabulary; 2) RAN is a novel zero-shot learning technique for Chinese character recognition. Unseen Chinese characters can be recognized because the necessary radicals and structures have been learned from other characters observed in the training stage. Technically, the numerous Chinese characters, as well as the newly created characters, can all be identified by a compact set of radicals and spatial structures learned in the training stage.

The use of radicals for Chinese character recognition has been researched for decades, and radical-based Chinese character recognition can be roughly considered to comprise two major problems, namely, radical extraction [22], [23] and structure analysis [24], [25]. The two problems can be solved sequentially or globally. Most conventional radical-based approaches are sequential approaches in which radical extraction is the first stage of a two-stage Chinese character recognition process. The second stage requires analysis of the structures of the radicals to identify the optimal radical combination. For example, [26] first implemented radical extraction based on a nonlinear active shape modeling (ARM) method. During the structural analysis, a dynamic tunneling algorithm was used to search for the optimal shape parameters in terms of chamfer distance minimization. Finally, the complete characters can be recognized via the Viterbi algorithm. Additionally, in [27], a recursive hierarchical scheme is developed to perform radical extraction first. Character features and radical features are then extracted for matching. Finally, in the structure analysis stage, a hierarchical radical matching scheme is devised to identify the radicals embedded in an input Chinese character and to recognize the input character. The sequential radical-based character recognition methods have the following limitations that this study aims to address: 1) radical extraction is a difficult problem; 2) structure analysis is complex and an effective strategy [28] must be applied during radical combination.

Inherently different from conventional sequential methods, the RAN proposed in this study is a global method. In RAN, radical extraction is implemented implicitly: we employ an attention mechanism to automatically perform the radical extraction. The attention mechanism has been found to be among the most distinct aspects of the human visual system [29]. By utilizing an attention model, RAN can focus on certain subgraphs of a Chinese character and choose the most relevant subgraph to describe a radical. Meanwhile, we also prove that the attention model can detect the relative spatial relationships among radicals. On the basis of the detected radicals and spatial relationships, the decoder in RAN is able to analyze the internal structures of Chinese characters and to predict the hierarchical radical structures of these characters. Finally, we can simply map these hierarchical radical structures to their corresponding character categories for recognition. Other radical-based Chinese character recognition methods are performed in global manner. A multilabel learning with residual network architecture was proposed in [30]. The method first predefined twenty types of radical structures and then marked every radical with a specific position. The Chinese characters are recognized when the labeled position-dependent radicals are predicted successfully. However, in [30], no attention-like mechanism is used to implicitly perform radical extraction; hence, how the approach proposed in [30] works is unclear. [31] proposed over-segmentation of the Chinese character graph into candidate radicals to avoid radical extraction. The optimal radical segmentation is searched in a lexicon-driven manner via a beam search strategy. However, the proposed method can address only the left-right structure, which is the simplest structure of Chinese characters.

偏旁部首的抽取是通过Attention自动抽取的

In addition to the Chinese character recognition task, this paper investigates the application of RAN in Chinese text line recognition. RAN possesses the distinctive advantage that it can easily be extended from recognizing single characters to recognizing text lines, as the embedded attention mechanism, which was originally devised to extract radicals, is also able to distinguish continuous Chinese characters in a single text line. We evaluate RAN for text line recognition in the ICPR MTWI 2018 competition [32]. The entire database is collected from multitype web images. We improve our attention model from single-head coverage attention to multihead coverage attention, where each head can generate a different attention distribution, to achieve better performance. This improvement enables the decoder to focus on context radicals or structures simultaneously in each decoding step; therefore, we hypothesize that the advantage makes proper training of the recognition model easier. Furthermore, to adapt RAN to the text line problem and to capture the document's temporal layout [33], we incorporate a new source encoder layer in the form of a multirow bidirectional RNN combined with gated recurrent units (GRU) [6] before the application of attention, and the GRU layers are improved by employing ReLU activation and batch normalization layers [34]. As web images contain complex and difficult noisy backgrounds, we combine HSV channels with RGB channels to strengthen RAN's robustness [35]. The entire database is dominated by Chinese characters, with many low-frequency or even unseen Chinese characters in the testing stage. Therefore, the proposed approach, which has the ability of few-shot or even zero-shot learning, leveraged its distinct advantages to achieve first place in the ICPR MTWI 2018 competition.

The main contributions of this study are summarized as follows:

- We propose RAN, a novel radical-based Chinese character processing method with few-/zero-shot learning ability.
- We describe the hierarchical radical structure of 27,533 Chinese characters (all the Chinese characters in the GB18030 standard [36]) and release the results to benefit related research.
- We demonstrate the performance of RAN on recognizing unseen Chinese characters and compare RAN with character-based methods on seen Chinese character recognition.
- We introduce how to extend RAN for text line recognition and experimentally demonstrate its performance.

This paper is an extension of our previous conference paper [37] in terms of five aspects: 1) We exploit densely connected convolutional networks (DenseNet) [38] in RAN; 2) We modify the composition of Chinese characters in ideographic description sequence (IDS) format; 3) We evaluate

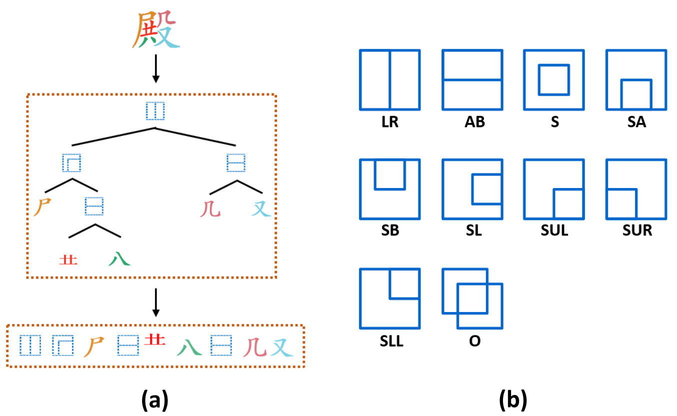


Fig. 1. (a) Hierarchical radical structure of an example Chinese character. The radicals are on the leaf nodes, and the structures are at the parent nodes. (b) Graphical representation of 10 common radical structures.

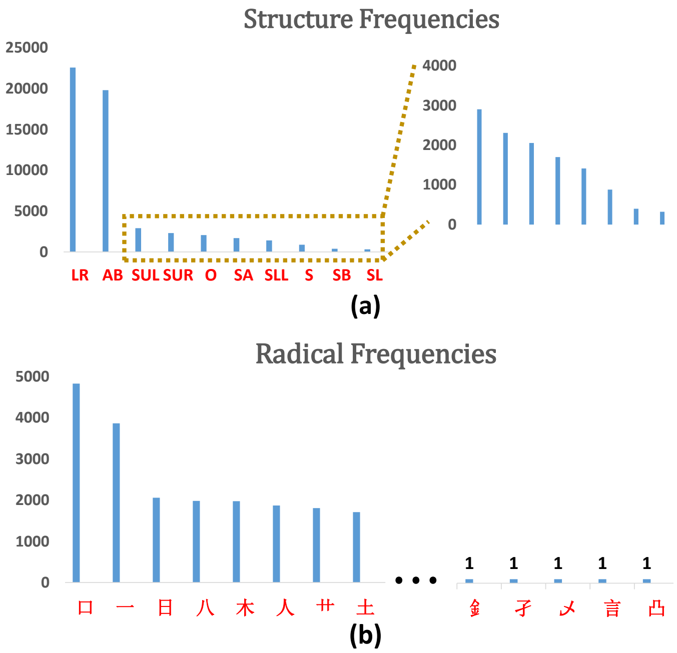


Fig. 2. (a) Number of Chinese characters having a specific structure; (b) Number of Chinese characters having a specific radical.

RAN on a natural scene database and prove its practical value in the real world; 4) We compare RAN with character-based methods and provide a detailed experimental analysis; 5) We extend RAN for text line recognition and analyze its performance in the ICPR MTWI 2018 competition.

The rest of this paper is organized as follows: Section 2 introduces the hierarchical radical structure of Chinese characters. Section 3 describes the proposed framework of RAN. Section 4 presents the architecture of the extension of RAN for Chinese text line recognition. Section 6 reports the experimental results on Chinese character recognition. Section 7 reports the experimental results on Chinese text line recognition, and Section 8 presents concluding remarks.

2 Radical Analysis

In contrast to English or Arabic characters, Chinese characters have the distinct property that numerous Chinese characters can be decomposed into a limited number of radicals. These radicals are viewed as semantic parts shared by different characters that appear in specific positions. Some radical structures can be derived based on the position-dependent radicals. For example, a left-hand radical and a right-hand radical constitute a left-to-right structure (LR structure in Fig. 1(b)). Ten different radical structures are defined in IDS (Chapter 12 in [3]), and we list the structures in Fig. 1(b): (1) left-to-right structure (LR), (2) above-to-below structure (AB), (3) full-surround structure (S), (4) surround-from-above structure (SA), (5) surround-from-below structure (SB), (6) surround-from-left structure (SL), (7) surround-from-upper-left structure (SUL), (8) surround-from-upper-right structure (SUR), (9) surround-from-lower-left structure (SLL), and (10) overlaid structure (O). The radical structure describes the relative position of two subgraphs, such as structures (1)-(10).

The internal Chinese character structure is intensely hierarchical: characters are composed of two graphs, which in turn are composed of two subgraphs until the bottom subgraphs belong to Chinese radicals. We illustrate a hierarchical structure of a Chinese character as a tree structure in Fig. 1(a). The Chinese character instance is above the top of the tree, and different radicals are denoted with different colors. The following tree structure describes the hierarchical radical structure: symbols on the parent nodes denote radical structures while symbols on the leaf nodes denote radicals. The main structure of the instance character is the left-to-right structure (top of the tree). The left part of the tree indicates a surround-from-upper-left structure, and the right part of the tree indicates an above-to-below structure. Finally, an above-to-below structure is presented at the bottom-right part of the surround-from-upper-left structure. The bottom of Fig. 1(a) shows the IDS sequence of the instance Chinese character, which is converted from the hierarchical tree structure. We perform this process in a simple way by following a depth-first traversal order.

Assuming the Chinese character in Fig. 1(a) is a novel testing character category that has not been observed in the training set, the traditional character-based recognition model will misclassify it. RAN attempts to imitate the technique of Chinese learners for recognizing Chinese characters, i.e., before asking students to remember and recognize Chinese characters, Chinese teachers first teach them to identify radicals, understand the meaning of radicals and grasp the possible structures between them. This learning technique, which is adopted in RAN, is more generative and helps to improve the memory ability of students so they can learn many Chinese characters. When an unseen Chinese character is encountered, RAN can generate the hierarchical radical structure if the essential radicals and structures have already been learned during the training stage. Technically, the numerous Chinese characters, as well as the newly created characters, can all be identified based on a compact set of radicals and spatial structures learned during the training stage.

We generate hierarchical radical structures of Chinese

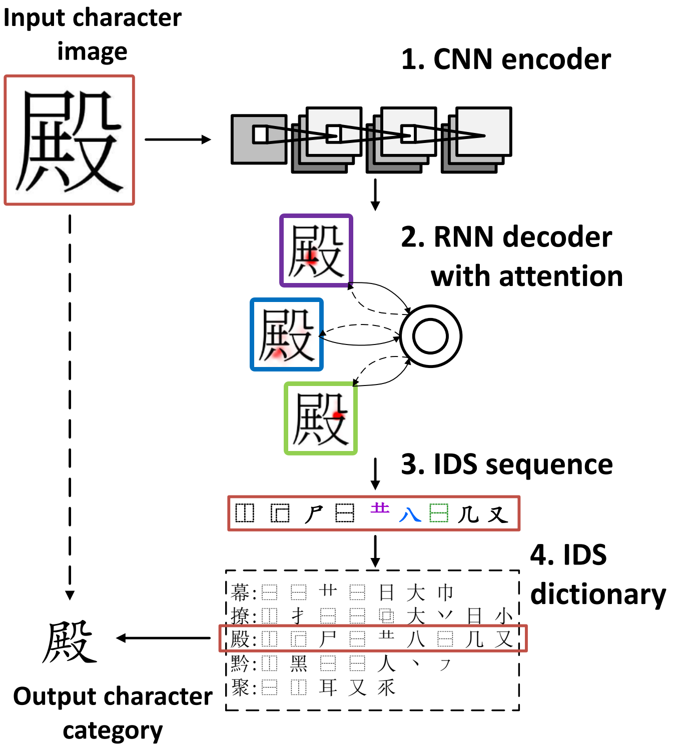


Fig. 3. Overall architecture of RAN: 1. The input character image is first fed into a CNN encoder to be transformed into visual features; 2. An RNN equipped with an attention model is employed as a decoder, the attention model lets decoder focus on only useful parts of visual features; 3. An IDS sequence is then predicted by the decoder symbol by symbol; 4. RAN finally outputs a character category by searching in the IDS dictionary to find the character category whose IDS sequence is most like the predicted IDS sequence.

characters by following the strategy in cjkvi-ids [1]. We decompose 27,533 Chinese characters (all the Chinese characters in the GB18030 standard) into 485 radicals and 10 radical structures. Fig. 2 shows the statistical information about the decomposition of the 27,533 Chinese characters, and Fig. 2(a) illustrates how many Chinese characters related to each radical structure, The left-to-right structure (LR) and above-to-below structure (AB) dominate the Chinese character set as they are basic and common structures. Fig. 2(b) illustrates how many Chinese characters are related to each radical. For brevity, we show only the 8 most common radicals on the left. Some low-frequency radicals are included in the complete set, which results in difficulties for RAN, as shown in the experiments. We present 5 examples that appear only once in the entire radical set on the right. These low-frequency radicals are usually related to rarely used Chinese characters. Our generated IDS sequence of the overall 27,533 Chinese characters is publicly available [2].

3 RADICAL ANALYSIS NETWORK

As illustrated in Fig. 3, by considering Chinese characters as linearized hierarchical radical structures (IDS sequences), the proposed RAN can successfully recognize a Chinese character by first predicting its IDS sequence and then selecting the output character category by searching in the predefined IDS dictionary to find the character category whose IDS sequence is most like the predicted IDS sequence. The IDS dictionary links the 27,533 Chinese characters with specific IDS sequences. If the input image belongs to a character category that is not observed during training but is included in the IDS dictionary, RAN recognizes it in the same manner as that for observed character categories. If the input image is a newly created character category that is not included in the current IDS dictionary, RAN can still predict the IDS sequence. All we need to do is update the IDS dictionary; there is no need to collect training samples of that category or retrain the models.

Regarding the network architecture, RAN is an improved version of the attention-based encoder-decoder framework. [39] recently showed that a caption sequence can be generated from an image with an attention-based encoder-decoder framework. The attention-based encoder-decoder framework was first proposed in [40] for machine translation. This framework has been extensively applied to many other applications, including speech recognition [41], [42], image captioning [43], [44], video processing [45] and formula recognition [46], [47].

3.1 Dense encoder

As shown in Fig. 3, RAN consists of an encoder and a decoder. Depending on the a priori knowledge of convolution, CNN has proven to be a powerful model for image processing. Therefore, we first employ CNN as the encoder to convert input character images to high-level visual features. Moreover, the convolutional layers in the CNN encoder are configured as densely connected layers in the DenseNet [38] architecture. The main idea of DenseNet is to use the concatenation of the output feature maps of the preceding layers as the input of the succeeding layers. As DenseNet is composed of many convolutional layers, let $H_l(\cdot)$ denote the convolution function of the l^{th} layer; then, the output of layer l is represented as:

$$\mathbf{x}_l = H_l([\mathbf{x}_0; \mathbf{x}_1; \ldots; \mathbf{x}_{l-1}]) \tag{1}$$

where $\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_l$ denote the output features produced in layers $0, 1, \ldots, l$, and ";" denotes the concatenation operation of feature maps. This iterative connection enables the network to learn shorter interactions across different layers and reuse features computed in the preceding layers. In this manner, DenseNet strengthens feature extraction and facilitates gradient propagation.

An essential component of convolutional networks is the pooling layers, which are capable of increasing the receptive field and improving invariance. However, the pooling layers disenable the concatenation operation as the size of the feature maps changes. Additionally, DenseNet is inherently memory demanding because the number of interlayer connections grows quadratically with depth. Consequently, DenseNet is divided into multiple densely connected blocks, and we employ compression layers between two contiguous dense blocks to reduce memory consumption. We illustrate the detailed architecture of the proposed dense encoder in Fig. 6 in Section 5.1. Rather than extracting features after a fully connected layer, the dense encoder contains only

[1] https://github.com/cjkvi/cjkvi-ids
[2] https://github.com/JianshuZhang

convolutional, pooling and activation layers, called fully convolutional neural networks, which enables the subsequent decoder to selectively focus on certain pixels of an image by choosing specific portions from all the extracted visual features.

We introduce the high-level visual features extracted by the dense encoder as annotations $\mathbf{A}$, which is a three-dimensional array of size $H \times W \times C$, where H denotes the height, W denotes the width and C denotes the output channels. Therefore, the annotations can be seen as a grid of $H \times W$ elements, where each element is a C-dimensional annotation corresponding to a local region of the image: $\mathbf{A} = \{\mathbf{a}_1, \ldots, \mathbf{a}_{H \times W}\}, \mathbf{a}_i \in \mathbb{R}^C$.

3.2 Decoder with attention

After extracting the visual features from the input images, the decoder of RAN begins to generate the IDS sequence. The IDS sequence is denoted as $\mathbf{Y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_T\}$, $\mathbf{y}_i \in \mathbb{R}^K$, where K is the number of total symbols in the radical vocabulary, which includes 485 basic radicals and 10 spatial structures, and T is the length of the IDS sequence. Note that the length of the IDS sequence is variable; we have to predict the output sequence one symbol at a time. Intuitively, each radical is related to only a certain part of an input character image. The entire input image is not required to provide useful information when predicting each radical: only the related pixels contribute. Therefore, we employ an attention mechanism to address the problem of radical alignment and to let the decoder know which part of the input image is suitable to attend to generate the next predicted radical or structure. For example, in Fig. 3, the purple, blue and green rectangles denote three symbols, with the red color representing the attention probabilities of each radical or structure (a lighter red color denotes a higher probability). When predicting the above-to-below structure (green rectangle), the attention model can automatically focus on the area between two vertical radicals, indicating an above-to-below direction, and the alignment of other two radicals corresponds well to human intuition.

We employ GRU, an improved version of simple RNN that can alleviate the vanishing and exploding gradient problems, as the decoder [48]. The GRU decoder is also implemented with the batch normalization function. Given the input $\mathbf{x}_t$, the GRU output $\mathbf{h}_t$ is computed by:

$$\mathbf{h}_t = \text{GRU}\left(\mathbf{x}_t, \mathbf{h}_{t-1}\right) \tag{2}$$

and the GRU function can be expanded as follows:

$$\mathbf{z}_t = \sigma(\mathbf{W}_{xz}\text{BN}(\mathbf{x}_t) + \mathbf{U}_{hz}\mathbf{h}_{t-1}) \tag{3}$$

$$\mathbf{r}_t = \sigma(\mathbf{W}_{xr}\text{BN}(\mathbf{x}_t) + \mathbf{U}_{hr}\mathbf{h}_{t-1}) \tag{4}$$

$$\tilde{\mathbf{h}}_t = \text{ReLU}(\mathbf{W}_{xh}\text{BN}(\mathbf{x}_t) + \mathbf{U}_{rh}(\mathbf{r}_t \otimes \mathbf{h}_{t-1})) \tag{5}$$

$$\mathbf{h}_t = (1 - \mathbf{z}_t) \otimes \mathbf{h}_{t-1} + \mathbf{z}_t \otimes \tilde{\mathbf{h}}_t \tag{6}$$

where σ is the sigmoid function, BN is the batch normalization function and $\otimes$ is an element-wise multiplication operator. $\mathbf{z}_t$, $\mathbf{r}_t$ and $\tilde{\mathbf{h}}_t$ are the update gate, reset gate and candidate activation, respectively.

The attention-based decoder adopts two unidirectional GRU layers; for brevity, we use GRU to represent the GRU layer in Eq. (2) and do not expand it. The attention model attempts to learn an alignment between the output symbol $\mathbf{y}_t$ and the input image $\mathbf{X}$ in each decoding time step t. Since we have converted the input image into high-level visual features through a dense encoder, the attention model is employed to learn the alignment between the output symbol $\mathbf{y}_t$ and the annotation $\mathbf{A}$. Let $\mathbf{s}_t$ denote the output state of the decoder at time step t. Since we do not have $\mathbf{y}_t$ when we want to compute the attention probabilities between $\mathbf{y}_t$ and $\mathbf{A}$, standard attention mechanisms replace $\mathbf{y}_t$ with the previous decoder state $\mathbf{s}_{t-1}$ to compute the attention probabilities. By contrast, in this paper, we utilize $\hat{\mathbf{s}}_t$ rather than $\mathbf{s}_{t-1}$ to compute the attention probabilities because we believe $\mathbf{s}_{t-1}$, the decoder state of the previous step, is an inaccurate representation of the current alignment information. We call $\hat{\mathbf{s}}_t$ the prediction of the current GRU hidden state, which is computed by the previous ground-truth symbol $\mathbf{y}_{t-1}$ and previous decoder state $\mathbf{s}_{t-1}$:

$$\hat{\mathbf{s}}_t = \text{GRU}\left(\mathbf{y}_{t-1}, \mathbf{s}_{t-1}\right) \tag{7}$$

Before utilizing $\hat{\mathbf{s}}_t$ and $\mathbf{A}$ to compute the attention probabilities, we must introduce a coverage vector $\mathbf{F}$, which is computed based on the summation of all past attention probabilities. The vector is computed by feeding the past attention probabilities into a convolutional layer with q output channels:

$$\mathbf{F} = \mathbf{Q} * \sum\nolimits_{l=1}^{t-1} \boldsymbol{\alpha}_l \tag{8}$$

The coverage vector $\mathbf{F}$ is employed to address the difficulty of the standard attention mechanisms, namely, the lack of coverage [49], which usually leads to over-parsing (some radicals are decoded more than once) and under-parsing problems (some radicals are never decoded). The past alignment information contained in $\mathbf{F}$ helps the attention model to know which part of the input image has been attended and ensures that each part is attended once and only once. We initialize $\mathbf{F}$ as a zero vector. Then, we compute the energy coefficients between $\hat{\mathbf{s}}_t$ and $\mathbf{A}$ given $\mathbf{F}$ using the following multilayer perceptron:

$$e_{ti} = \boldsymbol{\nu}_{\text{att}}^{\text{T}} \tanh(\mathbf{W}_{\text{att}}\hat{\mathbf{s}}_t + \mathbf{U}_{\text{att}}\mathbf{a}_i + \mathbf{U}_f\mathbf{f}_i) \tag{9}$$

Here, e_{ti} denotes the energy of annotation vector $\mathbf{a}_i$ (elements of $\mathbf{A}$) in decoding step t, and $\mathbf{f}_i$ denotes the elements of $\mathbf{F}$. We can obtain the attention coefficients α_{ti} by feeding e_{ti} into a softmax function. A context vector $\mathbf{c}_t$ is computed by weighted summation of all annotation vectors. We call $\mathbf{c}_t$ the context vector since it contains the overall information of the input image. However, as the weights α_{ti} denote the alignment probabilities, $\mathbf{c}_t$ includes the information of only the useful part of the image rather than the entire image:

$$\alpha_{ti} = \frac{\exp(e_{ti})}{\sum_{k=1}^{H \times W} \exp(e_{tk})} \qquad \mathbf{c}_t = \sum\nolimits_{i=1}^{H \times W} \alpha_{ti}\mathbf{a}_i \tag{10}$$

We finally utilize the second GRU layer to calculate the current output state of decoder $\mathbf{s}_t$ given $\mathbf{c}_t$ and $\hat{\mathbf{s}}_t$:

$$\mathbf{s}_t = \text{GRU}\left(\mathbf{c}_t, \hat{\mathbf{s}}_t\right) \tag{11}$$

The probability of each predicted symbol is computed by the context vector $\mathbf{c}_t$, current GRU hidden state $\mathbf{s}_t$ and

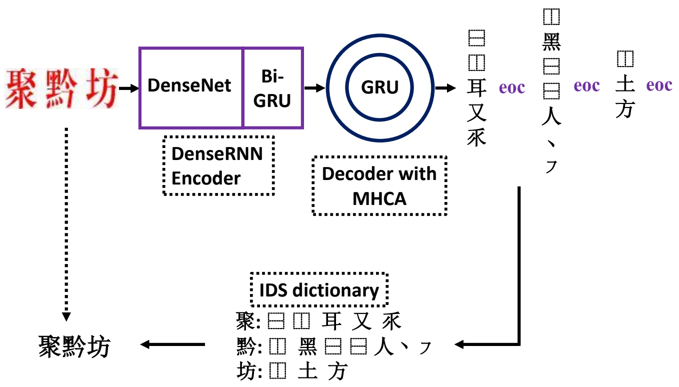


Fig. 4. Illustration of the extension of RAN for text line recognition. The method includes a denseRNN encoder and a GRU decoder equipped with MHCA (multihead coverage attention). An eoc (end-of-character) flag is added between every two adjacent IDS sequences for separation.

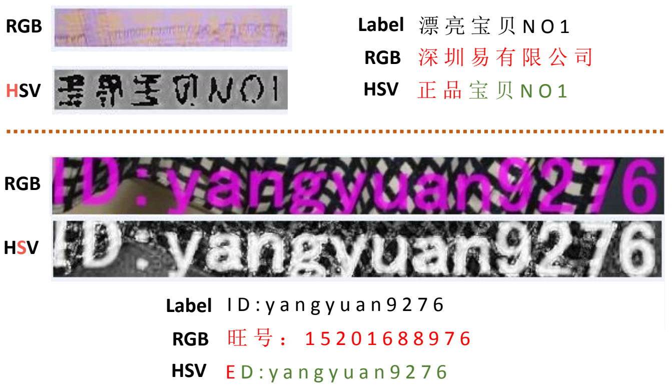


Fig. 5. Illustration of the benefits of HSV channels. Two text line images are difficult to recognize via RGB representations, but the representation can be improved by using HSV channels. Hence, the recognition results are improved.

one-hot vector of previous ground-truth symbol $\mathbf{y}_{t-1}$ using the following equation:

$$\mathbf{P}(\mathbf{y}_t|\mathbf{y}_{t-1},\mathbf{X}) = g\left(\mathbf{W}_o h\left(\mathbf{E}\mathbf{y}_{t-1} + \mathbf{W}_s \mathbf{s}_t + \mathbf{W}_c \mathbf{c}_t\right)\right) \quad (12)$$

where g denotes a softmax activation function over all the symbols in the vocabulary, and h denotes a maxout activation function. Let m' and n' denote the dimensions of embedding and attention. n denotes the GRU decoder dimension, $\boldsymbol{\nu}_{\text{att}} \in \mathbb{R}^{n'}$, $\mathbf{W}_{\text{att}} \in \mathbb{R}^{n' \times n}$, $\mathbf{U}_{\text{att}} \in \mathbb{R}^{n' \times C}$, $\mathbf{U}_f \in \mathbb{R}^{n' \times q}$, $\mathbf{W}_o \in \mathbb{R}^{K \times \frac{m'}{2}}$, $\mathbf{W}_s \in \mathbb{R}^{m' \times n}$, $\mathbf{W}_c \in \mathbb{R}^{m' \times C}$, and $\mathbf{E}$ denotes the embedding matrix.

4 Network Architecture for Radical-Based Chinese Text Recognition

As introduced previously, RAN possesses the valuable property of being easily extended from recognizing single characters to recognizing text lines. As illustrated in Fig. 4, by inserting an end-of-character (eoc) sentinel between every two Chinese character IDS sequences, RAN can predict the IDS sequence of the whole text line. We then transfer each IDS sequence, divided by eoc sentinels, into its Chinese character category through the IDS dictionary. To achieve better performance for robust text line recognition, we improve the encoder by employing a denseRNN encoder with HSV representations and exploit a multihead coverage-based attention mechanism.

4.1 DenseRNN encoder with HSV representations

The features extracted from the CNN are directly used as the annotation vectors for character recognition, whereas a stack of RNNs, called the denseRNN architecture, is built on top of the convolutional layers to capture the context information in the text line for text line recognition. As shown in Eq. (2), a GRU is a parameterized RNN function that recursively maps an input vector and a hidden state to a new hidden state; hence, the GRU can capture the historical context. Accordingly, we pass the CNN features through GRU layers and use the output of the GRU layers as the new annotation vectors. The extracted CNN features form a grid $\mathbf{V}$ of size $H \times W \times C$, where H denotes the number of rows, W denotes the number of columns and C denotes the number of feature maps. We first split grid $\mathbf{V}$ into H rows $\mathbf{V} = [\overline{\mathbf{V}}_1, \ldots, \overline{\mathbf{V}}_H]^{\mathrm{T}}$, where each $\overline{\mathbf{V}}_h$ is a sequence of length W, $\overline{\mathbf{V}}_h = [\overline{\mathbf{v}}_{h1}, \ldots, \overline{\mathbf{v}}_{hW}]$, $\overline{\mathbf{v}}_{hw} \in \mathbb{R}^C$. The new feature grid $\mathbf{A} = [\overline{\mathbf{A}}_1, \ldots, \overline{\mathbf{A}}_H]^{\mathrm{T}}$ is created from $\mathbf{V}$ by running the GRU function across each row. Recursively for all $\overline{\mathbf{A}}_h = [\overline{\mathbf{a}}_{h1}, \ldots, \overline{\mathbf{a}}_{hW}]$, $\overline{\mathbf{a}}_{hw} = \mathrm{GRU}\left(\overline{\mathbf{v}}_{hw}, \overline{\mathbf{a}}_{h(w-1)}\right)$. Nevertheless, unidirectional GRU cannot utilize the future context. To implement bidirectional GRU, we pass the input vectors through two GRU layers running in opposite directions and concatenate their hidden state vectors so that the new annotation vector $\mathbf{A}$ can capture both historical and future information. Annotation vectors that can capture context information are crucial for the good performance of RAN in text line recognition. The use of contextual cues is more stable and helpful than treating each character independently, as some ambiguous characters are easier to distinguish when observing their contexts. The denseRNN can be jointly trained in a unified network and is able to operate on text line images of arbitrary size by traversing from the start to end.

In Chinese character recognition in web images, the background is sometimes excessively noisy and the input text line is difficult to recognize in RGB representations, even for human eyes, as shown in the examples in Fig. 5. Therefore, in addition to the RGB representations, we use the HSV representations to improve the visibility of texts in web images. HSV representations include three channels, hue (H), saturation (S) and value (V), which are designed to more closely align with the way human vision perceives color-making attributes. As illustrated in Fig. 5, when RGB channels provide an ambiguous representation, HSV channels can give a much clearer visual image, leading to much better recognition results.

4.2 Multihead coverage attention (MHCA)

Multihead attention was first explored in [50] for machine translation, and we extend it to improve our RAN's performance on text line recognition. Specifically, MHCA extends the conventional coverage-based attention mechanism to have multiple heads, where each head can generate a different attention distribution. This process enables each head to play a different role in attending to the encoder output, which we hypothesize makes it easier for the decoder to learn to retrieve information from the encoder. In the conventional, single-head architecture, the model relies on the encoder to provide clear signals about the sentences so that the decoder can obtain the information via attention. We hypothesize that MHCA reduces the burden on the encoder and can distinguish radicals from noisy backgrounds when the encoded representation is less ideal. The MHCA employs M independent attention heads, each of which computes a context vector $\mathbf{c}_t^m, 1 \leq m \leq M$:

$$\mathbf{F} = \mathbf{Q} * \sum\nolimits_{l=1}^{t-1} \boldsymbol{\alpha}_l \tag{13}$$

$$e_{ti}^m = \boldsymbol{\nu}_{\text{att}}^{m\text{T}} \tanh(\mathbf{W}_{\text{att}}^m \hat{\mathbf{s}}_t + \mathbf{U}_{\text{att}}^m \mathbf{a}_i + \mathbf{U}_f^m \mathbf{f}_i) \tag{14}$$

$$\alpha_{ti}^m = \frac{\exp(e_{ti}^m)}{\sum_{k=1}^{H\times W} \exp(e_{tk}^m)} \quad \mathbf{c}_t^m = \sum\nolimits_{i=1}^{H\times W} \alpha_{ti}^m \mathbf{Z}^m \mathbf{a}_i \tag{15}$$

Here, $\boldsymbol{\alpha}_l$ is the attention map of size $H \times W \times M$; the convolution filter Q has M input channels and q output channels. n and n' denote the dimensions of the GRU decoder and original single-head attention. Let $C^{'}$ denote the dimensions of the new annotation vectors; then, $\boldsymbol{\nu}_{\text{att}}^m \in \mathbb{R}^{\frac{n'}{M}}$, $\mathbf{W}_{\text{att}}^m \in \mathbb{R}^{\frac{n'}{M}\times n}$, $\mathbf{U}_{\text{att}}^m \in \mathbb{R}^{\frac{n'}{M}\times C^{'}}$, $\mathbf{U}_f^m \in \mathbb{R}^{\frac{n'}{M}\times q}$, $\mathbf{Z^m}$ is the projection matrix of each head, $\mathbf{Z^m} \in \mathbb{R}^{\frac{C^{'}}{M}\times C^{'}}$. The final context vector is computed by concatenating the individual summaries: $\mathbf{c}_t = [\mathbf{c}_t^1; \mathbf{c}_t^2; \cdots; \mathbf{c}_t^M]$. In our experiments, we propose 4 heads ($M = 4$).

5 Training and Testing Procedure

5.1 Training

During the training procedure, RAN aims to maximize each predicted symbol probability by utilizing cross-entropy as the objective function, $O = -\sum_{t=1}^{T} \log p(w_t|\mathbf{y}_{t-1}, \mathbf{X})$, where w_t represents the ground-truth symbol at time step t, $\mathbf{y}_{t-1}$ denotes the one-hot vector of the previous ground-truth symbol, $\mathbf{X}$ denotes the input character image, and T denotes the length of the output sequence.

The details of our dense encoder are presented in Fig. 6. The left part of Fig. 6 shows that we employ three dense blocks in the main branch, as indicated by yellow rectangles. Before entering the first dense block, a 7×7 convolution (stride is 2×2) with 48 output channels is performed on the input expression images, followed by a 2×2 max pooling layer. We use 1×1 convolution followed by 2×2 average pooling as compression layers between two contiguous dense blocks. The compression layer reduces the number of feature maps of each block by one-half to further improve model compactness. The right part of Fig. 6 shows the details of each dense block. Each dense block is labeled "DenseB" because we use bottleneck layers to improve the computational efficiency, i.e., a 1×1 convolution is introduced before each 3×3 convolution to reduce the input to $4k$ feature maps. The input of each bottleneck convolution is the concatenation of all previous 3×3 convolution output feature maps. The growth rate $k = 24$ and the depth (number of convolution layers) of each block $D = 32$, which means each block has 16 1×1 convolution layers and 16 3×3 convolution layers. A batch normalization layer and a ReLU activation layer are placed consecutively after each convolution layer. We call the encoder DenseNet135 since a total of 135 convolution layers are included in the framework. For the denseRNN encoder architecture, the CNN part of denseRNN has the same architecture as that of DenseNet135, the RNN part employs a stack of two bidirectional GRU layers, with each layer containing 256 forward GRU units and 256 backward GRU units ($C^{'} = 512$).

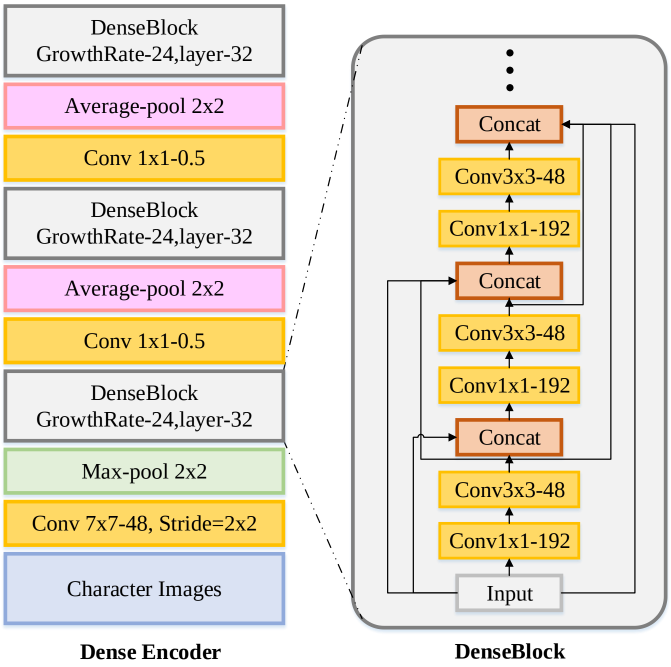


Fig. 6. Detailed architecture of the dense encoder, which we call the DenseNet135 encoder since it includes a total of 135 convolutional layers.

The decoder adopts 2 unidirectional GRU layers, and each layer has 256 forward GRU units. The embedding dimension m' and GRU decoder dimension n are set to 256. The attention dimension n' and the output channels of coverage convolution q for the single-head coverage attention model are set to 512. In the multihead coverage attention model, since we employ 4 heads, the attention dimension for each head is 128. We utilize the adadelta algorithm [51] for optimization. The adadelta hyperparameters are set to $\rho = 0.95$ and $\varepsilon = 10^{-6}$.

5.2 Testing

During testing, RAN aims to generate the most likely IDS sequence given the input image:

$$\hat{\mathbf{y}} = \arg\max_{\mathbf{y}} \log P\left(\mathbf{y}|\mathbf{x}\right) \tag{16}$$

In contrast to the training procedure, we do not have the ground truth of the previous predicted symbol when predicting the IDS sequence. To alleviate the problem of mismatch between the training and predicting procedure, a simple left-to-right beam search algorithm [52] is employed to implement the prediction procedure. In each time step, we maintain a set of 5 partial hypotheses. Each partial hypothesis in the beam is expanded with every possible symbol, and only the 5 most likely beams are kept. The prediction procedure for each hypothesis ends when the output

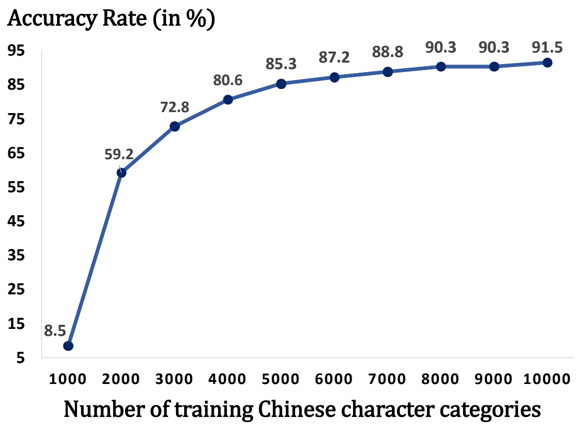


Fig. 7. Recognition performance of RAN for 17,533 unseen Chinese character categories with respect to the number of Chinese characters in the training samples.

symbol reaches the end-of-sequence. After successfully predicting the IDS sequence, we recognize the input Chinese character by searching the predefined IDS dictionary to find the character category whose IDS sequence is most like the predicted IDS sequence. We define the similarity between two IDS sequences by calculating the minimum edit distance.

The adoption of an ensemble beam search procedure [53] is intuitive for improving the performance. We first train N^e RAN models on the same training set with different initialized parameters. Then, we average their prediction probabilities to predict the current output symbol.

6 Experiments on Chinese Character Recognition

6.1 Experiments on unseen character recognition

In this section, we first demonstrate the effectiveness of RAN for identifying unseen Chinese character categories in terms of accuracy and attention visualization. The GB18030 standard includes 27,533 Chinese characters that are composed of only 485 radicals. We choose 10,000 character categories as the training set and use the other 17,533 character categories as the testing set. Clearly, the Chinese character categories in the testing set have not been seen during training. Note that traditional character-based methods fail to recognize unseen characters, which means their accuracies are **zero** in this experiment. However, these unseen characters are in the IDS dictionary; therefore, RAN can potentially recognize them by predicting their IDS sequences and searching for the correct character categories in the IDS dictionary. We increase the training set from 1,000 to 10,000 Chinese characters to see how many Chinese character categories are sufficient for training RAN to recognize the remaining unseen 17,533 characters. The input images are printed Chinese characters in Song font. We illustrate the performance in Fig. 7. RAN trained on 2,000 Chinese characters successfully recognizes 59.2% of the unseen 17,533 Chinese characters, and RAN trained on 10,000 Chinese characters achieves an accuracy of 91.5%, which means that 10,000 Chinese characters can help RAN to recognize more than 16,000 of the unseen Chinese character categories. Only approximately 500 Chinese characters are sufficient to cover the 485 Chinese radicals and 10 spatial structures. However, our experiments start with 1,000 training characters because the convergence of RAN is difficult when the training set is small, and RAN trained on 1,000 training characters achieves an accuracy of only 8.54%.

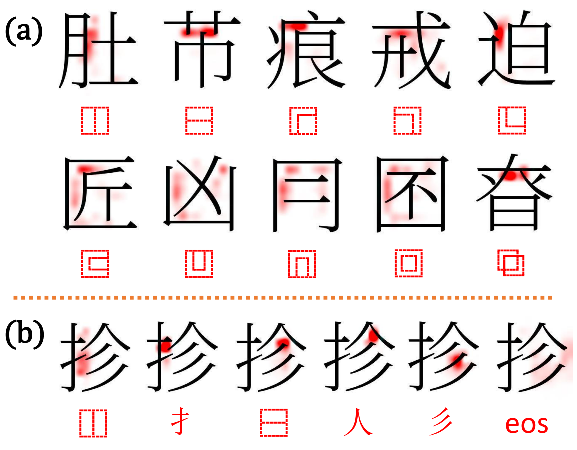


Fig. 8. Attention probabilities are represented by red color: a lighter red color denotes a higher probability and a darker red color denotes a lower probability. (a) Attention visualization for recognizing 10 common radical structures: symbols below the images are the predicted radical structures. (b) Attention visualization for recognizing a Chinese character instance: symbols below the images are the predicted radicals or structures.

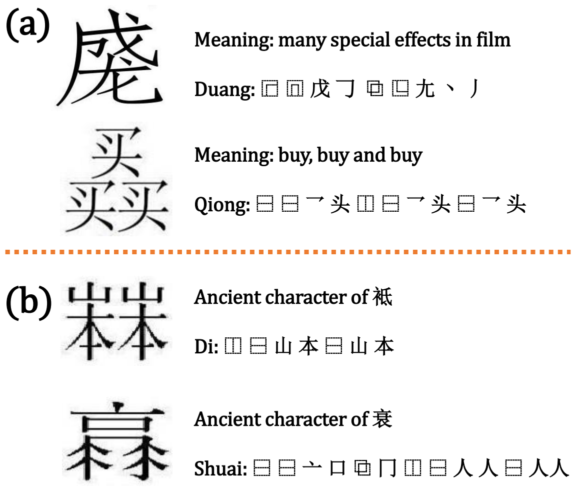
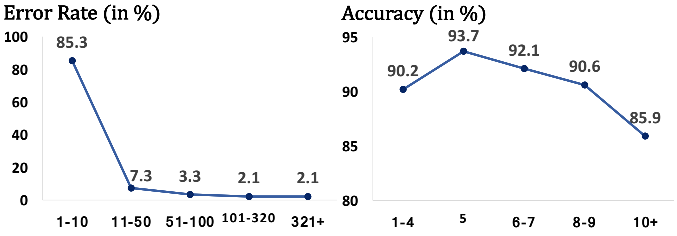

(a)
Meaning: many special effects in film
Duang: ⿸ ⿵ 戊 ㇆ ⿻ ⿺ 九 丶 丿
Meaning: buy, buy and buy
Qiong: ⿱ ⿱ 乛 头 ⿰ ⿱ 乛 头 ⿱ 乛 头
(b)
Ancient character of 衹
Di: ⿰ ⿱ 山 本 ⿱ 山 本
Ancient character of 衰
Shuai: ⿱ ⿱ 亠 口 ⿻ 冂 ⿰ ⿱ 人 人 ⿱ 人 人

Fig. 9. (a) Examples of recognition of the newly created internet Chinese characters; (b) Examples of recognition of rarely used ancient Chinese characters.

In contrast to conventional radical-based Chinese character recognition methods, RAN employs an attention model to segment radicals and identify the structures among the segmented radicals. The attention model plays an important role. Here, we prove that the attention can achieve human-like radical alignment and structure detection through several examples of attention visualization. In Fig. 8(a), we present 10 examples of how RAN identifies structures for every pair of radicals. The red color in the attention maps represents the spatial attention probability, where a lighter color indicates a higher attention probability and a darker color indicates a lower attention probability. Taking the



(a) Frequency of radicals (b) Length of IDS sequences

Fig. 10. (a) Radical-level error rate, each point describes the rate of all radicals within a specific range of appearance frequency that are substituted or deleted in the output IDS sequences. The range on the horizontal axis denotes the number of times that these radicals appear in the Chinese character training set. Approximately 100 radical categories are included in each range. (b) Character-level accuracy with respect to the length of the IDS sequences. Approximately 3,000 character categories are included in each range.

left-to-right structure as an example, the attention model focuses on mainly the space between the two horizontal radicals, which implicitly indicates a left-right direction. When identifying above-to-below structure, the attention model focuses on mainly the space between the two vertical radicals, which implicitly indicates an above-to-below direction. The focus of attention corresponds well to human intuition when identifying other radical structures. More specifically, in Fig. 8(b), we show the step-by-step process of RAN learning to recognize an unseen Chinese character from an image as an IDS sequence. When encountering basic radicals, the attention model generates an alignment strongly corresponding to human intuition and successfully predicts the radical structures “LR” and “AB” when a left-to-right direction and an above-to-below direction are detected.

Fig. 9 illustrates how RAN can be used to recognize newly created Chinese characters and rarely used Chinese characters due to the zero-shot learning ability. For example, Fig. 9(a) includes two novel internet Chinese characters: one named “Duang”, which was created by Jackie Chan, meaning many special effects in film, and the other named “Qiong”, meaning becoming poor due to too much expense. The generated hierarchical radical structures correspond well to human intuition. Fig. 9(b) includes two rarely used ancient characters for which RAN can correctly generate the radical structure. Although these Chinese characters have not been previously observed, RAN can successfully recognize them by adding the new correspondence between IDS sequences and related characters into the new IDS dictionary.

Note that nearly 1,490 Chinese characters are still misrecognized when RAN is trained on 10,000 Chinese characters. In Fig. 10, we analyze the cause of incorrect recognition. We first analyze the frequency of the radicals that fail to be recognized (including substitution errors and deletion errors). As introduced in Section 2, although RAN helps to alleviate the problem of recognizing low-frequency Chinese characters, some low-frequency radicals still cause difficulties for few-shot learning. Fig. 10(a) shows that radicals that appear fewer than 10 times are highly likely to be incorrectly recognized due to lack of learning. By contrast, for high-frequency radicals, the error rate is approximately 2% because they are shared by different Chinese characters in the training sample and have been learned sufficiently during the training procedure. Another interesting result is the distribution of accuracy with respect to the length of the IDS sequences. We expect the model to perform poorly on Chinese characters with longer IDS sequences since the characters that need to be decomposed into longer IDS sequences are usually related to more complicated structures and composed of more radicals, hence increasing the difficulty for structure detection and radical alignment. Fig. 10(b) illustrates this behavior.

6.2 Experiments on low-frequency character recognition

TABLE 1
Comparison of the performance of powerful image classifiers and RAN on the CTW test database. We divide the Chinese character categories into 4 subsets based on the appearance frequency in the training set. OOV is out-of-vocabulary, i.e., character categories that are not included in the training set; < 20 indicates character categories that appear fewer than 20 times in the training set; < 100 indicates character categories that appear fewer than 100 times in the training set; and HF stands for high frequency and includes character categories that appear more than 100 times in the training set.

Frequency	OOV	< 20	< 100	HF
Categories	70	328	573	1044
Samples	182	946	2892	48745
AlexNet	0%	24.4%	47.3%	77.1%
ResNet50	0%	24.4%	53.5%	81.3%
ResNet152	0%	22.2%	55.5%	81.6%
DenseNet135	0%	25.3%	55.8%	82.9%
RAN	**19.6%**	**35.2%**	**59.2%**	**84.3%**

In this section, we conduct experiments to illustrate the effectiveness of RAN for recognizing low-frequency Chinese characters and compare RAN with other powerful image classifiers. To explore the practical value of RAN, we investigate the performance on recognizing Chinese characters in the wild. Our experiments are performed on the CTW dataset [54], which contains Chinese character images collected from street views. The dataset is challenging due to its diversity and complexity. It contains planar text, raised text, text in cities, text in rural areas, text under poor illumination, distant text, partially occluded text, etc. Moreover, many low-frequency Chinese character categories are included.

The CTW database contains 3,580 Chinese character categories with 760,107 instances for training and 2,015 Chinese character categories with 52,765 instances for testing. Table 1 presents a detailed comparison of RAN and other character-based methods on the CTW database. We divide all testing character categories into 4 subsets based on the appearance frequency in the CTW training set to clearly demonstrate the effectiveness of RAN for few-shot learning. A total of 70 character categories are not observed in the training set, and 328 character categories have fewer than 20 training samples. Therefore, the recognition of these Chinese characters is a few-/zero-shot learning problem, which is difficult due to limited number of training samples. We tested several state-of-the-art character-based classifiers, namely, AlexNet [55], ResNet50 [56], and ResNet152, using Pytorch [57]. To ensure a fair comparison, we also

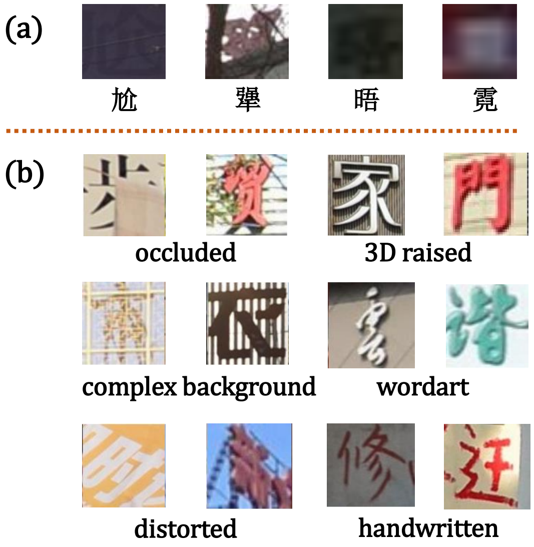


Fig. 11. (a) Failed examples of zero-shot learning for RAN in the CTW test set; (b) Visualization of 6 difficult attributes in the CTW database, each attribute shows 2 examples.

train a DenseNet classifier, named DenseNet135, with the same CNN architecture as that of the dense encoder of RAN. The comparison of DenseNet135 and RAN shows that RAN improves low-frequency character recognition (< 20) by nearly 10%. RAN improves the recognition of characters that appear fewer than 100 times in training set (< 100) by nearly 4%. Compared with the 0% OOV recognition of DenseNet135, the 12.6% recognition rate of RAN proves that RAN maintains the zero-shot learning ability in natural scenes. However, the recognition of RAN for unseen Chinese characters (OOV) and low-frequency Chinese characters (< 20) in natural scenes is not as good as that for printed Chinese characters. We show several failed instances of zero-shot learning in Fig. 11(a); these instances are incorrectly recognized as they are challenging even for humans. RAN improves the recognition rate of high-frequency Chinese characters by approximately 1.4%. We believe that the improvement is due to 2 aspects: 1) RAN decreases the size of the output dictionary, reduces the redundancy among similar Chinese characters and makes the model easier to train properly; and 2) RAN increases the robustness of the recognition model to noisy and complex images in natural scenes because radicals can be shared by many Chinese characters. More variation and transformation information can be learned if the training objective instances are radicals rather than characters.

To demonstrate the robustness of RAN in natural scenes, Table 2 compares te performance of DenseNet135 and RAN with respect to 7 attributes, where clean means that there is no noise in the image and that the character in the image has no transformation. We visualize the occluded, complex background, distorted, 3D raised, wordart and handwritten attributes in Fig. 11(b). The handwritten Chinese characters are the most difficult to recognize as they are more likely to ignore the internal structures of Chinese characters. The recognition results of DenseNet135 and RAN shows that RAN achieves only a 0.7% improvement on the clean attribute since the clean character images are not challenging and the number of low-frequency Chinese characters is small. However, RAN achieves considerable improvement compared with DenseNet135 on the other difficult attributes, especially wordart and handwritten. We believe that the improvement is due to the fact that the wordart and handwritten training samples are insufficient for training a character-based classifier, whereas for RAN, a radical can be shared by several Chinese characters. Hence, radicals are several times more frequent than characters in the training instances of wordart and handwritten, leading to an improvement of 3.8% on wordart and 8.8% on handwritten.

TABLE 2
Comparison of the recognition performance of the DenseNet image classifier and RAN on the CTW test database with respect to 7 attributes: clean, occluded, complex background, distorted, 3D raised, wordart characters and handwritten characters.

Attributes	Training Samples	DenseNet135	RAN
clean	273831	86.5%	87.2%
occluded	101393	67.6%	70.7%
background	218560	76.4%	78.7%
distorted	192481	76.3%	77.7%
3D raised	199066	77.6%	80.6%
wordart	65983	68.9%	72.7%
handwritten	6661	59.4%	68.2%

7 Experiments on Chinese Text Line Recognition

7.1 ICPR MTWI dataset

The Chinese text line recognition experiments are conducted on the ICPR MTWI 2018 challenge dataset, which is one of the largest published databases for Chinese text line recognition. The database of the ICPR MTWI challenge is collected from web images and includes various font styles and complex backgrounds. Success on this challenge represent potential commercial value. Although the database includes English words, Chinese characters are dominant. The official training set contains 128,210 text lines, with Chinese characters included in 76,130 text lines. A total of 4,010 Chinese character categories with 298,550 character instances are contained in training set, and 2,088 Chinese character categories with 24,039 character instances are included in the testing set. Similar to the analysis of the CTW database, we present a detailed illustration of OOV and low-frequency Chinese characters in the MTWI database in Table 3.

As shown in Table 3, 76 OOV Chinese character categories with 134 Chinese character instances are included in 120 text lines. Additionally, 597 Chinese character categories appear fewer than 20 times in the training set, with 1,180 character instances in 921 text lines. Therefore, the most challenging part of this task is that it requires a system with zero-/few-shot learning ability since nearly 11% of the testing text lines contain low-frequency Chinese characters. Most participants in this competition utilize character-based text line recognition methods, and their deep learning models fail to recognize these low-frequency Chinese characters if they use only the official training database. As a result, RAN's zero-/few-shot learning ability showed great power

TABLE 3
Detailed analysis of the Chinese character distribution in the ICPR MTWI 2018 database. We divide the testing set into 4 subsets based on the frequency of appearance of the Chinese characters in the text lines in the training set. OOV means out-of-vocabulary characters, < 20 means character categories that appear fewer than 20 times in the training set, < 100 means character categories that appear fewer than 100 times in the training set, HF means high frequency, i.e., character categories that appear more than 100 times in the training set. We report the number of text lines, the number of Chinese character instances and the number of Chinese character categories in each subset.

Set		Line Samples	Char Samples	Char Categories
Train		76130	298550	4010
Test	OOV	120	134	76
	< 20	921	1180	597
	< 100	2346	3469	732
	HF	5801	19256	683

on this challenge and helped us to win first place in the ICPR MTWI 2018 challenge.

7.2 Evaluation of RAN for text line recognition

We compare RAN with the character-based encoder-decoder model and CRNN-CTC model [33], [58] in Table 4. The CRNN-CTC model exploits the architecture in [33]. We improve the model by increasing the number of feature maps and implement another 3-gram language model trained on the official text database. The encoder-decoder models perform better than the CTC models as the encoder-decoder models rely on attention for alignment. There is no difference between recognizing vertical or horizontal text lines, but CTC models have difficulty in addressing this issue. Comparison of RAN with the character-based encoder-decoder model provides a better understanding of the advantage of RAN. The encoder and the decoder architectures used in the character-based model are the same as those used in RAN to ensure a fair comparison. In Table 4, CER-OOV denotes the results of the character error rate (substitution and deletion; no insertion errors are included here because we count the error rate of only specific characters) for recognizing OOV Chinese characters via the three models, whereas SACC-OOV denotes the results of the whole sentence accuracy rate for recognizing text lines containing OOV Chinese characters via the three models. The character-based approaches fail to recognize text lines containing OOV Chinese characters, leading to **0%** SACC and **100%** CER. By contrast, RAN has the ability to recognize unseen Chinese characters if the radicals have already been seen, resulting in an improvement of 19.4% on CER and an improvement of 5% on SACC. By comparing the CER of the encoder-decoder and RAN on high-frequency Chinese characters, we can see that the proposed RAN still achieves remarkable improvement since RAN is more robust to complex backgrounds and character variation than are the character-based approaches. RAN's advantage of recognizing low-frequency Chinese characters is clearly observed by comparing the results of RAN and the character-based encoder-decoder model on the < 20 and < 100 subsets.

TABLE 4
Detailed comparison of the text line recognition performance of CTC, character-based encoder-decoder and RAN. We divide the MTWI test set into 4 subsets: out-of-vocabulary (OOV), < 20, < 100 and high frequency (HF). CER is the character error rate; SACC is the whole sentence accuracy.

System		CRNN-CTC	Encoder-Decoder	RAN
CER	OOV	100%	100%	80.6%
	< 20	69.2%	65.5%	44.9%
	< 100	32.7%	31.4%	21.3%
	HF	12.2%	10.2%	9.5%
SACC	OOV	0%	0%	5%
	< 20	24.7%	26.9%	39.4%
	< 100	41.9%	45.8%	57.1%
	HF	64.2%	66.9%	68.2%

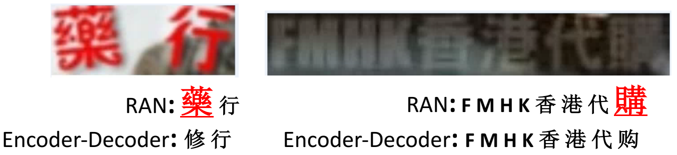


Fig. 12. Two examples of text lines containing low-frequency Chinese characters (denoted by red and underlines). The recognition results predicted by the proposed RAN and the character-based encoder-decoder are shown below the text line images.

Fig. 12 shows two examples of text lines containing low-frequency Chinese characters. The two characters in red are low-frequency characters that appear only a few times in the official training database. These characters belong to complex traditional Chinese character categories that are rarely used in common scenes and therefore are unlikely to be in the training set. The character-based encoder-decoder approach fails to recognize these characters, but RAN successfully recognizes them as they are composed of basic structures whose essential radicals have already been learned.

7.3 Evaluation of the proposed denseRNN, HSV and MHCA

TABLE 5
Comparison of CER and sentence accuracy (in %) and time efficiency (in msec) on the MTWI test set when appending the denseRNN encoder, multihead coverage attention and HSV channels to the proposed RAN system. * indicates ensemble models.

System(*)	CER	SACC	Time
RAN	14.8%	63.1%	10.4ms
+ denseRNN	12.5%	66.6%	11.3ms
+ MHCA	11.7%	68.1%	11.6ms
+ HSV	11.1%	68.6%	11.6ms

In Table 5, we show the improvements via the denseRNN encoder, MHCA and HSV representations by appending each to the previous system. We present the results of the ensemble models, and the number of combined models N^e is set to 5. First, the system "+ denseRNN" adds a new bidirectional GRU encoder immediately after the DenseNet encoder to not only extract high-level visual features from the input images but also capture the context information in the text lines. The input image of the denseRNN encoder can be an arbitrary size. The denseRNN encoder decreases the CER from 14.8% to 12.5% and improves the SACC from 63.1% to 66.6%. The CER is further decreased from 12.5% to 11.4% after the single-head coverage attention is replaced by the MHCA, and the SACC is increased by 2.1%, which indicates that the attention model with multiple heads generates a better attention distribution than that with a single head. Finally, consideration of the HSV information of the color images embedded in the input channels decreases the CER from 11.4% to 10.9%, which plays an important role in strengthening the ability of RAN for distinguishing Chinese characters in very complex backgrounds.

Second, compare the computational costs of the above 4 systems by investigating their speed. The experiments are all implemented with a Theano 0.10.0 [59] and an NVIDIA Tesla M40 24G GPU. We present the average time cost for recognizing each character on all 15,288 text lines with a testing batch size of 1. Appending the new bidirectional GRU encoder after the DenseNet encoder slows the average test speed for one text line by 5ms, despite the considerable improvement in recognition performance. The MHCA and HSV representations have a minimal affect on the test speed because the total number of parameters in attention model does not change when switching from a single head to multiple heads, and the computational cost of adding 3 input channels to the first convolutional layer can be ignored.

8 Conclusion and future work

In this study, we introduce a novel radical analysis network for radical-based Chinese character and Chinese text line recognition. The proposed model imitates the technique used by Chinese learners to recognize Chinese characters. We demonstrate through visualization and experimental results that RAN has the ability of few-/zero-shot learning of Chinese characters. Additionally, we present detailed comparisons to demonstrate RAN's advantage in the recognition of low-frequency character categories, not only in single-character recognition but also in text line recognition. We also verify the practical value of RAN in natural scenes. The released IDS dictionary will substantially benefit related research.

In future work, we plan to identify a better decomposition of Chinese characters, and we will improve the attention model to increase the few-/zero-shot learning ability of RAN for recognizing low-quality Chinese character images. We hope that by proposing a novel radical-based recognition model, people will be encouraged to create more interesting and personal Chinese characters as novel characters can be easily recognized.

Acknowledgments

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, and MOE-Microsoft Key Laboratory of USTC.

References

[1] Y. Y. Tang, L.-T. Tu, J. Liu, S.-W. Lee, and W.-W. Lin, "Off-line recognition of Chinese handwriting by multifeature and multilevel classification," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 20, no. 5, pp. 556–561, 1998.

[2] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, p. 436, 2015.

[3] J. D. Allen, D. Anderson, J. Becker, R. Cook, M. Davis, P. Edberg, M. Everson, A. Freytag, L. Iancu, R. Ishida *et al.*, *The Unicode Standard*. Citeseer, 2012, vol. 6.

[4] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[5] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[6] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1724–1734.

[7] C.-L. Liu, S. Jaeger, and M. Nakagawa, "Online recognition of Chinese characters: the state-of-the-art," *IEEE transactions on pattern analysis and machine intelligence*, vol. 26, no. 2, pp. 198–213, 2004.

[8] C.-L. Liu, F. Yin, D.-H. Wang, and Q.-F. Wang, "Online and offline handwritten Chinese character recognition: benchmarking on new databases," *Pattern Recognition*, vol. 46, no. 1, pp. 155–162, 2013.

[9] D. Cireşan and U. Meier, "Multi-column deep neural networks for offline handwritten Chinese character classification," in *Neural Networks (IJCNN), 2015 International Joint Conference on*. IEEE, 2015, pp. 1–6.

[10] C. Wu, W. Fan, Y. He, J. Sun, and S. Naoi, "Handwritten character recognition by alternately trained relaxation convolutional neural network," in *Frontiers in Handwriting Recognition (ICFHR), 2014 14th International Conference on*. IEEE, 2014, pp. 291–296.

[11] Z. Zhong, L. Jin, and Z. Xie, "High performance offline handwritten Chinese character recognition using googlenet and directional feature maps," in *Document Analysis and Recognition (ICDAR), 2015 13th International Conference on*. IEEE, 2015, pp. 846–850.

[12] X.-Y. Zhang, F. Yin, Y.-M. Zhang, C.-L. Liu, and Y. Bengio, "Drawing and recognizing Chinese characters with recurrent neural network," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 849–862, 2018.

[13] Y. Bengio, Y. LeCun, and D. Henderson, "Globally trained handwritten word recognizer using spatial representation, convolutional neural networks, and hidden markov models," in *Advances in neural information processing systems*, 1994, pp. 937–944.

[14] Z. Xie, Z. Sun, L. Jin, H. Ni, and T. Lyons, "Learning spatial-semantic context with fully convolutional recurrent network for online handwritten chinese text recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 8, 2018.

[15] B. Graham, "Sparse arrays of signatures for online character recognition," *arXiv preprint arXiv:1308.0371*, 2013.

[16] W. Yang, L. Jin, D. Tao, Z. Xie, and Z. Feng, "Dropsample: A new training method to enhance deep convolutional neural networks for large-scale unconstrained handwritten Chinese character recognition," *Pattern Recognition*, vol. 58, pp. 190–203, 2016.

[17] G. Sampson, *Writing systems*. London, 1985.

[18] D. George, W. Lehrach, K. Kansky, M. Lázaro-Gredilla, C. Laan, B. Marthi, X. Lou, Z. Meng, Y. Liu, H. Wang *et al.*, "A generative vision model that trains with high data efficiency and breaks text-based captchas," *Science*, vol. 358, no. 6368, p. eaag2612, 2017.

[19] L. Fei-Fei, R. Fergus, and P. Perona, "One-shot learning of object categories," *IEEE transactions on pattern analysis and machine intelligence*, vol. 28, no. 4, pp. 594–611, 2006.

[20] B. M. Lake, R. Salakhutdinov, and J. B. Tenenbaum, "Human-level concept learning through probabilistic program induction," *Science*, vol. 350, no. 6266, pp. 1332–1338, 2015.

[21] Y. Suen and E. Huang, "Computational analysis of the structural compositions of frequently used Chinese characters," *Computer Processing of Chinese and Oriental Languages*, vol. 1, no. 3, pp. 163–176, 1984.

[22] D. Shi, S. R. Gunn, and R. I. Damper, "Handwritten Chinese radical recognition using nonlinear active shape models," *IEEE transactions on pattern analysis and machine intelligence*, vol. 25, no. 2, pp. 277–280, 2003.

[23] F.-H. GHENG and W.-H. Hsu, “Radical extraction from handwritten Chinese characters by background thinning method,” *IEICE TRANSACTIONS (1976-1990)*, vol. 71, no. 1, pp. 88–98, 1988.

[24] S. W. Lu, Y. Ren, and C. Y. Suen, “Hierarchical attributed graph representation and recognition of handwritten Chinese characters,” *Pattern Recognition*, vol. 24, no. 7, pp. 617–632, 1991.

[25] M. Zhao, “Two-dimensional extended attribute grammar method for the recognition of hand-printed Chinese characters,” *Pattern recognition*, vol. 23, no. 7, pp. 685–695, 1990.

[26] D. Shi, R. I. Damper, and S. R. Gunn, “Offline handwritten Chinese character recognition by radical decomposition,” *ACM Transactions on Asian Language Information Processing (TALIP)*, vol. 2, no. 1, pp. 27–48, 2003.

[27] A.-B. Wang and K.-C. Fan, “Optical recognition of handwritten Chinese characters by hierarchical radical matching method,” *Pattern Recognition*, vol. 34, no. 1, pp. 15–35, 2001.

[28] S.-R. Lay, C.-H. Lee, N.-J. Cheng, C.-C. Tseng, B.-S. Jeng, P.-Y. Ting, Q.-Z. Wu, and M.-L. Day, “On-line Chinese character recognition with effective candidate radical and candidate character selections,” *Pattern recognition*, vol. 29, no. 10, pp. 1647–1659, 1996.

[29] M. Corbetta and G. L. Shulman, “Control of goal-directed and stimulus-driven attention in the brain,” *Nature reviews neuroscience*, vol. 3, no. 3, p. 201, 2002.

[30] 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)*. IEEE, 2017, pp. 579–584.

[31] L.-L. Ma and C.-L. Liu, “A new radical-based approach to online handwritten Chinese character recognition,” in *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*. IEEE, 2008, pp. 1–4.

[32] M. He, Y. Liu, Z. Yang, S. Zhang, C. Luo, F. Gao, Q. Zheng, Y. Wang, X. Zhang, and L. Jin, “ICPR2018 contest on robust reading for multi-type web images,” in *Pattern Recognition (ICPR), 2018 24th International Conference on*. IEEE, 2018, pp. 7–12.

[33] B. Shi, X. Bai, and C. Yao, “An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 11, pp. 2298–2304, 2017.

[34] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” in *International Conference on Machine Learning*, 2015, pp. 448–456.

[35] K. Van De Sande, T. Gevers, and C. Snoek, “Evaluating color descriptors for object and scene recognition,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 9, pp. 1582–1596, 2010.

[36] H. Ni, “GB18030 - the new Chinese encoding standard,” *from http://www.gb18030.com*.

[37] J. Zhang, Y. Zhu, J. Du, and L. Dai, “RAN: Radical analysis networks for zero-shot learning of Chinese characters,” *arXiv preprint arXiv:1711.01889*, 2017.

[38] 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.

[39] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, “Show and tell: A neural image caption generator,” in *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*. IEEE, 2015, pp. 3156–3164.

[40] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” *arXiv preprint arXiv:1409.0473*, 2014.

[41] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, “Listen, attend and spell: A neural network for large vocabulary conversational speech recognition,” in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 4960–4964.

[42] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, and Y. Bengio, “End-to-end attention-based large vocabulary speech recognition,” in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 4945–4949.

[43] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, “Show, attend and tell: Neural image caption generation with visual attention,” in *International Conference on Machine Learning*, 2015, pp. 2048–2057.

[44] A. Karpathy and L. Fei-Fei, “Deep visual-semantic alignments for generating image descriptions,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3128–3137.

[45] L. Gao, Z. Guo, H. Zhang, X. Xu, and H. T. Shen, “Video captioning with attention-based lstm and semantic consistency,” *IEEE Transactions on Multimedia*, vol. 19, no. 9, pp. 2045–2055, 2017.

[46] 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.

[47] 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*, 2018.

[48] Y. Bengio, P. Simard, and P. Frasconi, “Learning long-term dependencies with gradient descent is difficult,” *IEEE transactions on neural networks*, vol. 5, no. 2, pp. 157–166, 1994.

[49] Z. Tu, Z. Lu, Y. Liu, X. Liu, and H. Li, “Modeling coverage for neural machine translation,” *arXiv preprint arXiv:1601.04811*, 2016.

[50] 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. 6000–6010.

[51] M. D. Zeiler, “Adadelta: an adaptive learning rate method,” *arXiv preprint arXiv:1212.5701*, 2012.

[52] K. Cho, “Natural language understanding with distributed representation,” *arXiv preprint arXiv:1511.07916*, 2015.

[53] T. G. Dietterich, “Ensemble methods in machine learning,” in *International workshop on multiple classifier systems*. Springer, 2000, pp. 1–15.

[54] T.-L. Yuan, Z. Zhu, K. Xu, C.-J. Li, and S.-M. Hu, “Chinese text in the wild,” *arXiv preprint arXiv:1803.00085*, 2018.

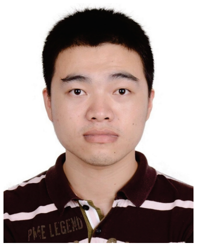
[55] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[56] 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.

[57] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, “Automatic differentiation in pytorch,” 2017.

[58] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, “Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks,” in *Proceedings of the 23rd international conference on Machine learning*. ACM, 2006, pp. 369–376.

[59] J. Bergstra, O. Breuleux, F. Bastien, P. Lamblin, R. Pascanu, G. Desjardins, J. Turian, D. Warde-Farley, and Y. Bengio, “Theano: A cpu and gpu math compiler in python,” in *Proc. 9th Python in Science Conf*, vol. 1, 2010.



Jianshu Zhang received the B.Eng. degree from the Department of Electronic Engineering and Information Science, University of Science and Technology of China (USTC) in 2015. He is currently a Ph.D. candidate of USTC. His current research area is neural network, handwriting mathematical expression recognition and Chinese document analysis.

Jun Du received the B.Eng. and Ph.D. degrees from the Department of Electronic Engineering and Information Science, University of Science and Technology of China (USTC) in 2004 and 2009, respectively. From 2004 to 2009, he was with iFlytek Speech Lab of USTC. During the above period, he worked as an Intern twice for 9 months at Microsoft Research Asia (MSRA), Beijing. In 2007, he also worked as a Research Assistant for 6 months in the Department of Computer Science, The University of Hong Kong. From July 2009 to June 2010, he worked at iFlytek Research on speech recognition. From July 2010 to January 2013, he joined MSRA as an Associate Researcher, working on handwriting recognition, OCR, and speech recognition. Since February 2013, he has been with the National Engineering Laboratory for Speech and Language Information Processing (NEL-SLIP) of USTC.

Lirong Dai was born in China in 1962. He received the B.S. degree in electrical engineering from Xidian University, Xian, China, in 1983, and the M.S. degree from the Hefei University of Technology, Hefei, China, in 1986, and the Ph.D. degree in signal and information processing from the University of Science and Technology of China (USTC), Hefei, in 1997. He joined USTC in 1993. He is currently a Professor at the School of Information Science and Technology, USTC. His research interests include speech synthesis, speaker and language recognition, speech recognition, digital signal processing, voice search technology, machine learning, and pattern recognition. He has published more than 50 papers in these areas.