DenseRAN for Offline Handwritten Chinese Character Recognition

Wenchao Wang, Jianshu Zhang, Jun Du, Zi-Rui Wang and Yixing Zhu
University of Science and Technology of China
Hefei, Anhui, P. R. China
Email: {wangwenc, xysszjs}@mail.ustc.edu.cn, jundu@ustc.edu.cn, {cs211, zyxsa}@mail.ustc.edu.cn

***Abstract*—Recently, great success has been achieved in offline handwritten Chinese character recognition by using deep learning methods. Chinese characters are mainly logographic and consist of basic radicals, however, previous research mostly treated each Chinese character as a whole without explicitly considering its internal two-dimensional structure and radicals. In this study, we propose a novel radical analysis network with densely connected architecture (DenseRAN) to analyze Chinese character radicals and its two-dimensional structures simultaneously. DenseRAN first encodes input image to high-level visual features by employing DenseNet as an encoder. Then a decoder based on recurrent neural networks is employed, aiming at generating captions of Chinese characters by detecting radicals and two-dimensional structures through attention mechanism. The manner of treating a Chinese character as a composition of two-dimensional structures and radicals can reduce the size of vocabulary and enable DenseRAN to possess the capability of recognizing unseen Chinese character classes, only if the corresponding radicals have been seen in training set. Evaluated on ICDAR-2013 competition database, the proposed approach significantly outperforms whole-character modeling approach with a relative character error rate (CER) reduction of 18.54%. Meanwhile, for the case of recognizing 3277 unseen Chinese characters in CASIA-HWDB1.2 database, DenseRAN can achieve a character accuracy of about 41% while the traditional whole-character method has no capability to handle them.**

***Keywords*-radical analysis network, dense convolutional network, attention, offline handwritten Chinese character recognition**

I. Introduction

Handwritten Chinese characters recognition is a challenging problem due to the large number of character classes, confusion between similar characters, and distinct handwriting styles across individuals [1], [2]. According to the type of data acquisition, handwriting recognition can be divided into online and offline. For offline handwritten Chinese characters recognition (HCCR), characters are gray-scaled images which are analyzed and classified into different classes. In traditional methods, the procedures for HCCR often include: image normalization, feature extraction, dimension reduction and classifier training. With the success of deep learning [3], convolutional neural network (CNN) [4] has been applied successfully in this domain. The multi-column deep neural network (MCDNN) [5] was the first CNN used for HCCR. The team from Fujitsu used a CNN-based model to win the ICDAR-2013 HCCR competition [6]. Zhong et al. [7] improved the performance which outperforms the human performance. Li et al. [8] from Fujitsu further improved the performance based on a single CNN model with augmented training data using distortion. The ensemble based methods can be further used to improve the performance with some tradeoff on speed and memory. Zhong et al. [9] further improved the performance to by using spatial transformer network with residual network. However, these algorithms can only recognize Chinese characters appeared in training set and have no ability to recognize unseen Chinese characters. Moreover, these algorithms treat each Chinese character as a whole without considering the similarites and sub-structures among Chinese characters.

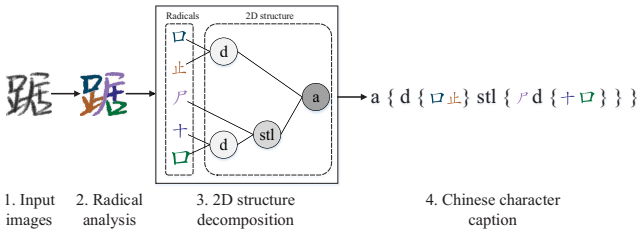

Figure 1. Illustration of DenseRAN to recognize Chinese characters by analyzing the radicals and two-dimensional structures. "**d**" denotes top-bottom structure, "**stl**" denotes top-left-surround structure, "**a**" denotes left-right structure.

Chinese characters, which are mainly logographic and consist of basic radicals, constitute the oldest continuously used system of writing in the world and are different from the purely sound-based writing systems such as Greek, Hebrew, etc. It is natural to decompose Chinese characters to radicals and spatial structures then use this knowledge for character recognition. In the past few decades, a lot of work has been done for radical-based Chinese character recognition. [10] proposed a matching method which first detected radicals separately and then composed radicals into a character using a hierarchical radical matching method. [11] tried to over-segment characters into candidate radicals while the proposed way could only handle the left-right structure and over-segmentation brings many difficulties. Recently, [12] proposed a multi-label learning for radical-based Chinese character recognition. It turned a character class into a combination of several radicals and spatial

structures. Generally, these approaches have difficulty in segmenting characters into radicals and lacking flexibility when to analyze structures among radicals. More importantly, they usually can't handle these unseen Chinese character classes.

In this paper, we propose a novel radical-based approach to HCCR, namely radical analysis network with densely connected architecture (DenseRAN). Different from above mentioned radical-based approaches, in DenseRAN the radical segmentation and structure detection are automatically learned by attention based encoder-decoder model. The main idea of DenseRAN is to decompose a Chinese character into a caption that describes its internal radicals and structures among radicals. A handwritten Chinese character is successfully recognized when its caption matches the groundtruth. In order to give a better explanation, we illustrate how DenseRAN recognizes a Chinese character in Fig. 1. Each leaf node of the tree in third step represents radicals and each non-leaf node represents its internel structure. The handwriting input is finally recognized as the Chinese character caption after the radicals and two-dimenstional structures are detected. Based on the analysis of radicals and structures, the proposed DenseRAN possesses the capability of recognizing unseen Chinese character classes only if the radicals have been seen in training set.

The proposed DenseRAN is an improved version of attention based encoder-decoder model [13]. The overall architecture of DenseRAN is shown in Fig. 3. The raw data of input are gray-scaled images. DenseRAN first encodes input image to high-level visual vectors using a densely connected convolutional networks (DenseNet) [16]. Then a RNN with gated recurrent units (GRU) [17] decodes the high-level representations into output caption step by step. We adopt a coverage based spatial attention model built in the decoder to detect the radicals and internal two-dimensional structures simultaneously [14], [15]. Compared with [18] focusing on printed Chinese character recognition, DenseRAN focuses on HCCR which is much more difficult due to the diversity of writing styles.

II. Chinese character decomposition

Each Chinese character can be naturally decomposed into a caption of radicals and spatial structures. Following the rule in [19], the character caption consists three key components: radicals, spatial structures and a pair of braces (e.g. "{" and "}"). One spatial structure with its radicals can be represented as: "**structure** { radical-1, radical-2 }".

A radical represents a basic part of Chinese character and is frequently shared among Chinese characters. Compared with enormous Chinese character categories, the total number of radicals is quite limited. It is declared in GB13000.1 standard published by National Language Committee of China that there are nearly 500 radicals for over 20,000 Chinese characters. Fig. 2 illustrates thirteen common spatial structures and their corresponding Chinese

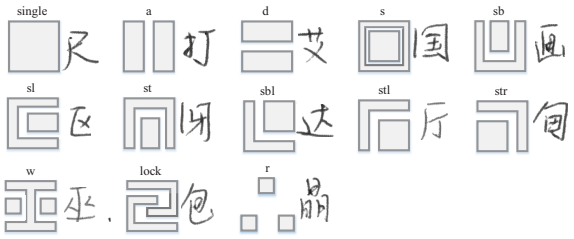

Figure 2. Graphical representation of thirteen common spatial structures.

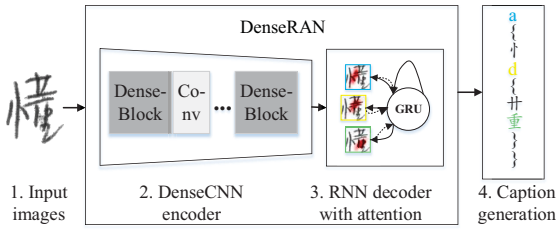

Figure 3. The overall architecture of DenseRAN for HCCR.

character samples. These thirteen structures are: **single:** some Chinese characters are radicals themselves. **a:** left-right structure. **d:** top-bottom structure. **s:** surround structure. **sb:** bottom-surround structure. **sl:** left-surround structure. **st:** top-surround structure. **sbl:** bottom-left-surround structure. **stl:** top-left-surround structure. **str:** top-right-surround structure. **w:** within structure. **lock:** lock structure. **r:** one radical repeated many times in a character.

III. The architecture of DenseRAN

A. Dense encoder

Dense convolutional network (DenseNet) [16] has been proven to be good feature extractors for various computer vision tasks. So we use DenseNet as the encoder to extract high-level visual features from images. Instead of extracting features after a fully connected layer, we discard fully connected layer and softmax layer in encoder, called fully convolutional neural networks. This allows the decoder to selectively pay attention to certain parts of an image by choosing specific portions from the extracted visual features.

The architecture of DenseNet is mainly divided into multiple DenseBlocks. As shown in Fig. 4, in each denseblock, each layer is connected directly to all subsequent layers. We denote $H_l(\cdot)$ as the convolution function of the l^{th} layer in some block, then the output of the l^{th} layer can be represented as:

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

where $[\mathrm{x}_0, \mathrm{x}_1, \ldots, \mathrm{x}_{l-1}]$ denotes the concatenation of the output feature maps produced by $0, 1 \ldots, l-1$ in the same block. The growth rate $k = 64$ means each $H_l(\cdot)$ produces

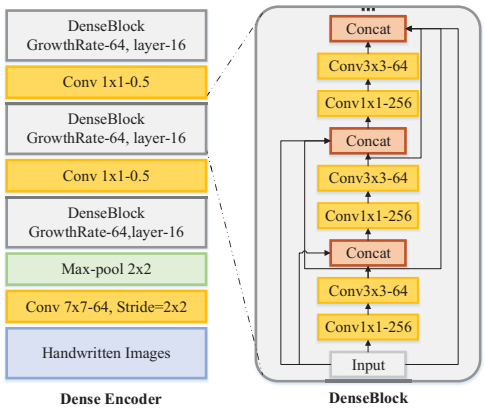

Figure 4. The architecture of DenseEncoder.

k feature maps. In order to further improve computational efficiency, we use bottleneck layers in each DenseBlock. A 1×1 convolution is introduced before each 3×3 convolution to reduce the number of feature maps input to $4k$. The depth of each Denseblock is set to $D = 16$, i.e., in each block, there are D 1×1 convolution layers and each one is followed by a 3×3 convolution layer.

These DenseBlocks are normally seperated by transition layer and pooling layer. Each transition layer is a 1×1 convolution layer parameterized by $\theta = 0.5$. If the number of input feature maps of transition layer is n, then the transition layer will generate θn output feature maps. In DenseRAN, because the input character images are resized to 32×32, after so many pooling operation, the size of final feature map is about 2×2, which is too small to get good attention results. So we discard the pooling layer between DenseBlocks in Fig. 4. The first convolution layer has 64 convolutions of kernel size 7×7 with stride 2 which is performed on the input images, followed by a 2×2 max pooling layer. Batch normalization [20] and ReLU [21] are performed after each convolution layer consecutively.

Dense encoder extracts visual features which can be represented as a three-dimensional array of size $H\times W\times D$, $L = H \times W$. Each element in array is a D-dimensional vector that corresponds to a local region of the image:

$$\mathbf{A} = \{\mathbf{a}_1, \ldots, \mathbf{a}_L\}, \mathbf{a}_i \in \mathbb{R}^D \tag{2}$$

B. GRU decoder with attention model

As illustrated in Fig. 3, the decoder generates a caption of input Chinese character. The output caption $\mathbf{Y}$ is represented by a sequence of 1-of-K encoded symbols:

$$\mathbf{Y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_C\}, \mathbf{y}_i \in \mathbb{R}^K \tag{3}$$

where K is the number of total symbols in the vocabulary which includes the basic radicals, spatial structures and a pair of braces, C is the length of caption.

Because the length of annotation sequence L is fixed while the length of captions C is variable, DenseRAN addresses this problem by computing an intermediate fixed-size vector $\mathbf{c}_t$ at each time step. Note that $\mathbf{c}_t$ is a dynamic representation of the relevant part of the Chinese character image at time t. We utilize unidirectional GRU [22] and the context vector $\mathbf{c}_t$ to produce captions step by step. The probability of each predicted word is computed by the context vector $\mathbf{c}_t$, current GRU hidden state $\mathbf{s}_t$ and previous word $\mathbf{y}_{t-1}$ using the following equation:

$$p(\mathbf{y}_t|\mathbf{y}_{t-1}, \mathbf{X}) = \text{Softmax}(\mathbf{W}_0(\mathbf{E}\mathbf{y}_{t-1} + \mathbf{W}_s\mathbf{s}_t + \mathbf{W}_c\mathbf{c}_t)) \tag{4}$$

where $\mathbf{W}_0 \in \mathbb{R}^{K\times m}$, $\mathbf{W}_s \in \mathbb{R}^{m\times n}$, $\mathbf{W}_c \in \mathbb{R}^{m\times D}$, and $\mathbf{E}$ denotes the embedding matrix, m and n are the dimensions of embedding and GRU parser.

The GRU parser adopts two unidirectional GRU layers to calculate the hidden state $\mathbf{s}_t$:

$$\hat{\mathbf{s}}_t = \text{GRU}(\mathbf{y}_{t-1}, \mathbf{s}_{t-1}) \tag{5}$$

$$\mathbf{c}_t = f_{\text{att}}(\hat{\mathbf{s}}_t, \mathbf{A}) \tag{6}$$

$$\mathbf{s}_t = \text{GRU}(\hat{\mathbf{s}}_t, \mathbf{c}_t) \tag{7}$$

where $\mathbf{s}_{t-1}$ denotes hidden state at time $t-1$, $\hat{\mathbf{s}}_t$ is the GRU hidden state prediction at time t, and coverage based spatial attention model f_{att} is parameterized as a multi-layer perceptron:

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

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

$$\alpha_{ti} = \frac{\exp(e_{ti})}{\sum_{k=1}^{L}\exp(e_{tk})} \tag{10}$$

The coverage vector $\mathbf{F}$ is computed based on the summation of past attention probabilities. α_{ti} denotes the spatial attention coefficient of $\mathbf{a}_i$ at time t. Let n' denotes the attention dimension and q denotes the feature map of filter $\mathbf{Q}$, then $\mathbf{v}_{\text{att}} \in \mathbb{R}^{n'}$, $\mathbf{W}_{\text{att}} \in \mathbb{R}^{n'\times n}$, $\mathbf{U}_{\text{att}} \in \mathbb{R}^{n'\times D}$, $\mathbf{U}_f \in \mathbb{R}^{n'\times q}$. With the weight α_{ti}, we compute the context vector $\mathbf{c}_t$ as:

$$\mathbf{c}_t = \sum_{i=1}^{L} \alpha_{ti}\mathbf{a}_i \tag{11}$$

IV. Experiments on recognizing seen Chinese characters

In this section, we present some comparison experiments on seen offline Chinese characters to show the advantage of performance of DenseRAN.

A. Dataset

The database used for evaluation is from the ICDAR-2013 competition [6] of HCCR. The database used for training is the CASIA database [23] including HWDB1.0 and 1.1. The most common Chinese characters are used, i.e., 3755 level-1 set of GB2312-80.

Table I
RESULTS ON ICDAR-2013 COMPETITION DATABASE OF HCCR.

Methods	Ref.	Accuracy
Human Performance	[6]	96.13%
Traditional Method	[27]	92.72%
VGG14RAN	-	93.79%
DenseNet	-	95.90%
DenseRAN	-	96.66%

B. Implementation details

We normalize gray-scaled image to the size of 32×32 as the input. The implementation details of Dense encoder has been introduced in Section III-A. The decoder is two unidirectional layers with 256 GRU units. The embedding dimension m and decoder state dimension n are set to 256. The convolution kernel of $\mathbf{Q}$ is set to 5×5 and the number of feature maps is set to 128. The model is trained with mini-batch size of 150 on one GPU. We utilize the adadelta [24] with gradient clipping for optimization. The best model is determined in terms of word error rate (WER) of validation set. We use a weight decay of 10^{-4} and dropout [25] after each convolution layer and set the dropout rate to 0.2.

In the decoding stage, we aim to generate a most likely caption string given the input character. The beam search algorithm [26] is employed to find the optimal decoding path in the decoding process. The beam size is set to 10.

C. Experiments results

In Table I, the human performance on ICDAR-2013 competition database and the previous benchmark are both listed. In order to compare DenseRAN with whole-character based approach, only DenseNet which is the same as the encoder of DenseRAN is evaluated as a whole-character classifier on ICDAR-2013 competition database, we call it "DenseNet". As shown in Table I, "DenseNet" achieves 95.90% while DenseRAN achieves 96.66% revealing relative character error rate reduction of 18.54%. Also, we replace the encoder of DenseRAN with VGG14 [28] and keep the other parts unchanged, we name it as "VGG14RAN". Table I clearly shows CNN with densely connnected architecture is more powerful than VGG on extracting high-quality visual features from handwritten Chinese character images.

V. EXPERIMENTS ON RECOGNIZING UNSEEN CHINESE CHARACTERS

Chinese characters are enormous which is difficult to train a recognition system that covers all of them. Therefore it is necessary and interesting to empower a system to recognize unseen Chinese characters. In this section, we show the effectiveness of DenseRAN to recognize unseen characters.

A. Dataset

We divide 3755 common Chinese characters into 2755 classes and another 1000 classes. We pick 2755 classes in HWDB1.0 and 1.0 set as the training set and the other 1000 classes in ICDAR-2013 competition database are selected as the test set. So both the test character classes and handwriting styles have never been seen during training. In order to explore the influence of training samples, different training sizes are designed, ranging from 500 to 2755 classes. Note that all radicals are covered in all training sets.

HWDB1.2 dataset is also used for evaluating the DenseRAN performance on unseen Chinese characters. There are 3319 non-common Chinese characters in HWDB1.2 dataset and we pick 3277 classes to make sure the radicals of these characters are covered in 3755 common classes. Note that the Chinese characters in HWDB1.2 dataset are not common and usually have more complicated radical composition.

Table II
RESULTS ON UNSEEN HCCR.

Train set	Train class	Train samples	Test set	Test class	Accuracy
HWDB 1.0+1.1	500	356553	ICDAR 2013	1000	1.70%
	1000	712254		1000	8.44%
	1500	1068031		1000	14.71%
	2000	1425530		1000	19.51%
	2755	1962529		1000	30.68%
HWDB 1.0+1.1	3755	2674784	HWDB 1.2	3277	40.82%

B. Experiments results

As shown in Table II, with the seen Chinese character classes increase from 500 to 2755, the accuracy on 1000-class test set increases from 1.70% to 30.68%. Whole-character modeling systems can not recognize unseen Chinese character classes at all. The last row of Table II shows that DenseRAN can recognize unseen uncommon Chinese characters in HWDB1.2 with 40.82% accuracy.

VI. QUALITATIVE ANALYSIS

A. Attention visualization

By visualizing the attention probabilities learned by the model, we show how DenseRAN recognizes radicals and two-dimensional structures. We also analyze the error examples by attention visualization. We show some Chinese chracters which are misclassified by "DenseNet" in Fig. 5(a). On the contrary, as shown in Fig. 5(b), DenseRAN aligns radicals and structures of offline handwritten Chinese character step by step as human intuition and finally gets the correct classification. Above the dotted line, these Chinese characters are seen in training set. Below the dotted line, the character is not seen in training set. Fig. 5 clearly illustates that DenseRAN not only outperforms whole-character modeling method, but also has the ability of recognizing unseen characters.

Examples of mistakes are shown in Fig. 6. The first column shows the correct characters and the misclassified

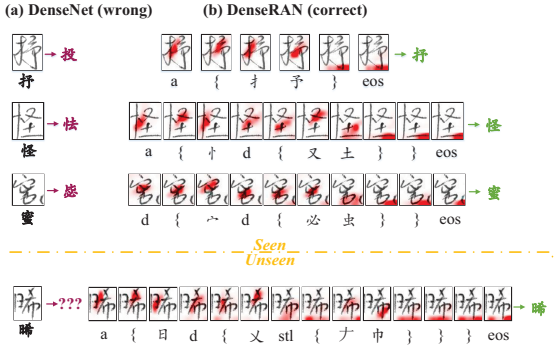

Figure 5. (a): Characters misclassified by "DenseNet". (b): Same characters classified correctly by DenseRAN and the attention over times.

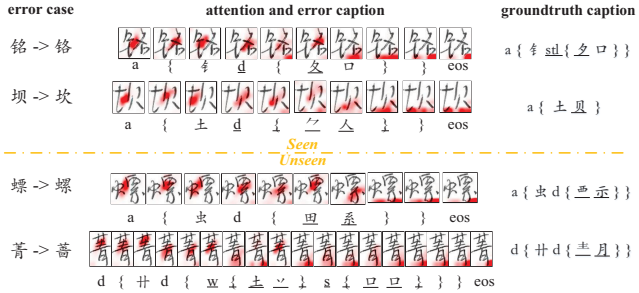

Figure 6. Examples of mistakes where we can use attention to gain intuition into what DenseRAN saw.

characters as well. The second column visualizes attention and the symbols decoded over times. The third column gives the corresponding groundtrurth caption. Above the dotted line, these Chinese characters are seen in training set. Below the dotted line, these characters are not seen in training set. As shown in Fig. 6, the mistaken characters usually come from the confuable characters which are easy to cause errors when the handwritten style is scribbled.

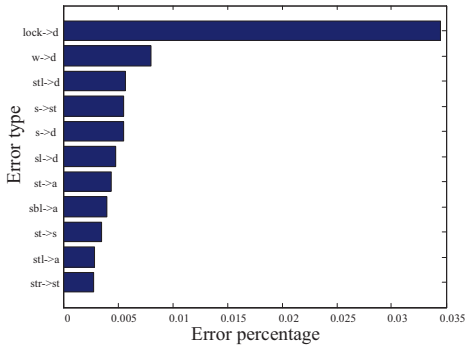

Figure 7. Error percentage of different two dimensional structures.

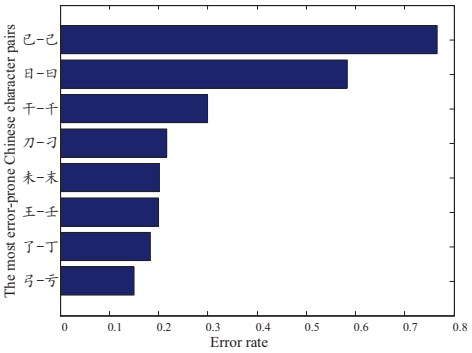

Figure 8. Error rate of Chinese characters with **single** structure.

B. Error distribution of different two-dimensional structures

In order to further explore important factors that cause the error, we analyze the error distribution with different two-dimensional structures of Chines characters. Intuitively, the more complicated the structure is, the more easily the error occurs. Fig. 7 shows the error percentage of structure A misclassified to structure B on test set except the "**single**" structure. The error percentage here is calculated as:

$$P_{err}(A \to B) = \frac{N_{A \to B}}{N_A} \tag{12}$$

where N_A denotes the number of samples with structure A in test set, $N_{A \to B}$ denotes the number of samples with structure A misclassified to B. As shown in Fig. 7, the most likely mistake is "**lock**" misclassified to "**d**" with 3.45% error percentage. The explanation is: as shown in Fig. 2, "**lock**" will become "**d**" when the two parts of "**lock**" are written seperately. This situation is very easy to appear when the handwritten style is scribbled.

Fig. 8 shows the error rate of error pair (E-G) which are both "**single**" structure, which is calculated as:

$$R_{err}(E\text{-}G) = \frac{N_{E \to G} + N_{G \to E}}{N_E + N_G} \tag{13}$$

where N_E denotes the number of character E in test set, N_G denotes the number of character G. $N_{E \to G}$ denotes how many E samples are misclassified to G in test set, vice versa. From Fig. 8, we find the misclassified characters usually come from the characters which only have some subtle differences in radicals with another character.

VII. Conclusion and Future Work

In this study we introduce DenseRAN for HCCR. The proposed DenseRAN recognizes Chinese character by identifying radicals and analyzing spatial structures. Experiments shows that DenseRAN outperforms whole-character modeling approach on HCCR task and has the capability of recognizing unseen Chinese characters. By visualizing

attention and analyzing error examples, we should pay more attention on confusing characters in the future.

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, and MOE-Microsoft Key Laboratory of USTC. This work was also funded by Huawei Noah's Ark Lab.

REFERENCES

[1] F. Kimura, K. Takashina, S. Tsuruoka, Y. Miyake, "Modified quadratic discriminant functions and the application to Chinese character recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, pp. 149–153, 1987.

[2] R.-W. Dai, C.-L. Liu, B.-H. Xiao, "Chinese character recognition: history, status and prospects," *Front. Comput. Sci. China*, pp. 126–136, 2007.

[3] LeCun, Y. and Bengio, Y. and Hinton, G., "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

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

[5] D. Ciresan, J. Schmidhuber, "Multi-column deep neural networks for offline handwritten Chinese character classification," *arXiv*:1309.0261, 2013.

[6] F. Yin, Q.-F. Wang, X.-Y. Zhang, and C.-L. Liu, "ICDAR 2013 Chinese handwriting recognition competition," *Proc. ICDAR*, 2013, pp. 1464–1470.

[7] Z. Zhong, L. Jin, and Z. Xie, "High performance offline handwritten Chinese character recognition using GoogLeNet and directional feature maps," *Proc. ICDAR*, 2015, pp. 846–850.

[8] L. Chen, S. Wang, W. Fan, J. Sun, and N. Satoshi, "Beyond human recognition: A CNN-Based framework for handwritten character recognition," *Proc. ACPR*, 2015, pp. 695–699.

[9] Z. Zhong, X-Y. Zhang, F. Yin, C.-L. Liu, "Handwritten Chinese character recognition with spatial transformer and deep residual networks," *Proc. ICPR*, 2016, pp. 3440–3445.

[10] 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.

[11] 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.

[12] T.-Q. Wang, F. Yin, and C.-L. Liu, "Radical-based Chinese character recognition via multi-labeled learning of deep residual networks," in *Proc. ICDAR*, 2017, pp. 579–584.

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

[14] 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*, 2017.

[15] J. Zhang, J. Du, and L. Dai, "Multi-scale attention with dense encoder for handwritten mathematical expression recognition," *arXiv*:1801.03530, 2018.

[16] Gao Huang, Zhuang Liu, Laurens van der Maaten, Kilian Q. Weinberger, "Densely Connected Convolutional Networks," *arXiv*:1608.06993, 2016.

[17] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv*:1412.3555, 2014.

[18] J. Zhang, Y. Zhu, J. Du, and L. Dai, "Radical analysis network for zero-shot learning in printed Chinese character recognition," in *Proc. International Conference on Multimedia and Expo*, 2018.

[19] A. Madlon-Kay, "*cjk-decomp.*" Available: https://github.com/amake/cjk-decomp.

[20] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. 32nd International Conference on Machine Learning*, 2015, pp. 448-456.

[21] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proc. 27th International Conference on Machine Learning*, 2010, pp. 807-814.

[22] Yoshua Bengio, Patrice Simard, and Paolo Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE transactions on neural networks*, vol. 5, no. 2, pp. 157-166, 1994.

[23] C.-L. Liu, F. Yin, D.-H. Wang, and Q.-F. Wang, "CASIA online and offline Chinese handwriting databases," *Proc. ICDAR*, 2011.

[24] M. D. Zeiler, "ADADELTA: an adaptive learning rate method," *arXiv*:1212.5701, 2012.

[25] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929-1958, 2014.

[26] K. Cho, "Natural language understanding with distributed representation," *arXiv*:1511.07916, 2015.

[27] 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.

[28] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv*:1409.1556, 2014.