Multiple Comparative Attention Network for Offline Handwritten Chinese Character Recognition

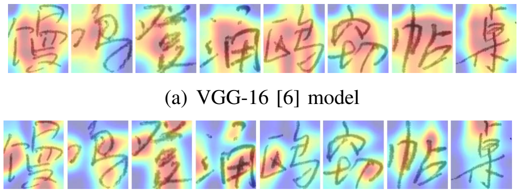
Qingquan Xu, Xiang Bai, Wenyu Liu*
School of Electronic Information and Communication
Huazhong University of Science and Technology, Wuhan, China
{*qingquanxu, xbai, liuwy*}@*hust.edu.cn*

***Abstract*—Recent advances in deep learning have made great progress in offline Handwritten Chinese Character Recognition (HCCR). However, most existing CNN-based methods only utilize global image features as contextual guidance to classify characters, while neglecting the local discriminative features which is very important for HCCR. To overcome this limitation, in this paper, we present a convolutional neural network with multiple comparative attention (MCANet) in order to produce separable local attention regions with discriminative feature across different categories. Concretely, our MCANet takes the last convolutional feature map as input and outputs multiple attention maps, a contrastive loss is used to restrict different attention selectively focus on different sub-regions. Moreover, we apply a region-level center loss to pull the features that learned from the same class and different regions closer to further obtain robust features invariant to large intra-class variance. Combining with classification loss, our method can learn which parts of images are relevant for recognizing characters and adaptively integrates information from different regions to make the final prediction. We conduct experiments on ICDAR2013 offline HCCR competition dataset with our proposed approach and achieves an accuracy of 97.66%, outperforming all single-network methods trained only on handwritten data.**

I. Introduction

Handwritten Chinese character recognition (HCCR) has been studied since the 1960s [1] and has obtained considerable improvements, benefiting a variety of practical applications, such as mail sorting [2], bank check reading, book and handwritten notes transcription, and so on. Although many researches have been conducted, it remains a challenging problem which is mainly due to the large number of character classes, confusion between similar characters, and distinct handwriting styles across individuals.

Most of previous methods leverage CNN to do classification by learning global semantic features from the whole image [3], [4], which are not sufficient for character classification where the model needs to be able to tell various character categories apart. Concretely, global attention provided by these methods can be well localized, whereas for character classification, the attention maps are supposed to be discriminative across categories as well. To illustrate this better, considering Figure 1(a), we observe that the class activation map [5] of different character classes from VGG-16 [6] model trained on offline handwritten data are very similar with large attention overlapping regions, which lead to confusion between label class and false positives , thereby result in sub-optimal character recognition performance.


(a) VGG-16 [6] model

(b) Proposed method

Figure 1. Examples of attention maps generated by Grad-CAM [5] from different models: (a) Attention maps obtained from VGG-16 [6] model trained on offline Handwritten data. (b) Attention maps obtained from our proposed MCANet by normalizing the maximum attended value across all comparative attention and taking mean over them. As can be seen, the attention maps in (a) attend to similar regions across different class while in (b) the proposed model can find discriminative patterns which led to better model interpretability as well as improved classification accuracy.

According to our daily experience, when recognizing a handwritten Chinese character which is difficult to distinguish from multiple visual confusion characters, the human often abstracts the discriminative features from all candidates and then compares the similarity and difference of them to determine which categories the character belonging to. For examples, the “谩” and “漫” are visual confusing but we can distinguish them from their left parts “讠” and “氵”. Similarly, for “漫” and “浸”, we can make a decision according to the comparison between their upper-right parts.

Therefore, our intuition is that we can learn separate attention regions so that the local discriminative region features are more likely to be obtained. Specifically, we propose a multiple comparative attention network (MCANet) to learn multiple attention regions and supervised by multiple loss, *i.e.*, a contrastive loss [7] is used to restrict the multi-attention to focus on different parts of an input character image, a region-level instead of the image-level center loss [8] is applied on different attention regions to learn compact attention feature between intra-class characters. Experiment results show that the class activation map produced by

*Corresponding author.

2379-2140/19/$31.00 ©2019 IEEE
DOI 10.1109/ICDAR.2019.00101

our proposed MCANet can locate discriminative separate regions accurately and keep attention map complete (see Fig. 1(b)).

No whistles and bells, our proposed MCANet achieves an accuracy of 97.66%, outperforming all previous single-network methods that trained on offline handwritten data. To the best of our knowledge, the proposed method is also the first to leverage multi-attention and multi-loss simultaneously in the fields of HCCR.

The contributions of this paper are summarized as follows:

- We propose an effective multiple comparative attention network, enforcing the model to focus on multiple regions and obtain category-specific attention features which lead to better model generalizability and interpretability for HCCR.
- We further propose a region-level center loss on different region features of the same class to enforce the model-learned features to be invariant to large intra-class variance.
- Our proposed model learns discriminative and compact region features for inter-class and intra-class character respectively, which achieves a state-of-the-art accuracy on ICDAR2013 dataset for single-network methods trained on handwritten data.

II. Related Work

Offline Handwritten Chinese Character Recognition: Handwritten Chinese Character Recognition (HCCR) has received intensive attention since the 1980s. Early works for HCCR most often use classifiers include modified quadratic discriminant function (MQDF) [9], support vector machines (SVM) [10], and discriminative learning quadratic discriminative function(DLQDF) [11].

Due to the success of deep learning, CNN-based models gradually improve the performance of offline HCCR. Multi-column deep neural networks (MCDNN) [12] is the first application of CNN for offline HCCR. The best performance of MCDNN single network achieves an accuracy of 94.47%, which outperforms the best traditional method by a significant margin. The Fujish research team creates a CNN-based method and took the winner place in the ICDAR2013 competition with an accuracy of 94.77% [13]. In 2014, they adopted a voting format of four alternately trained relaxation convolutional neural networks (ART-CNN) [14], which improved accuracy to 96.06%.

The first model that outperforms human-level performance is presented by Zhong *et al.* [3], which incorporates traditional directional feature maps. Therefore, their single HCCR-Gabor-GoogLeNet and ensemble HCCR-Ensemble-GoogLeNet-10 models achieve a recognition accuracy of 96.35% and 96.74%, respectively. The framework proposed by Zhou *et al.* [15] is based on HCCR-GoogLeNet [3], they use a Kronecker fully connected (KFC) layer to replace the layers after the four inception groups and then followed by two fully connected layers, finally obtaining an accuracy of 96.63%. Zhang *et al.* [16] combine traditional normalization-cooperated direction-decomposed feature maps and CNNs to obtain accuracy values of 96.95% and 97.12% by voting on three models. In ICDAR 2017, a method proposed by Yang *et al.* [17] uses residual blocks and attention-based iterative refinement module, achieving an accuracy of 97.37%. The HCCR-12CNNLayer proposed by Xiao *et al.* [18] achieves an excellent recognition result of 97.59%. Recently, Melnyk-Net [19] adopts a modified global weighted average pooling (GWAP) – global weighted output average pooling (GWOAP) to improve offline HCCR performance which results in a state-of-the-art accuracy of 97.61%, considering single network methods trained only on handwritten data.

Attention Mechanisms: Attention can be viewed as a tool to bias the allocation of available processing resources towards the most informative components of the input signal [20], [21]. The benefits of such a mechanism have been shown across a range of tasks, from localization and understanding in images [22], [23] to sequence-based models [24], [25]. It is typically implemented in combination with a gating function (e.g. a softmax or sigmoid) and sequential techniques [26]. In this paper, our multi-attention follows the general design of squeeze-and-excitation (SE) module proposed in SENet [20] to localize different attention region features.

III. Method

In this section, we describe the proposed multiple comparative attention network for offline HCCR.

A. Overall Architecture

We consider the network with two comparative attention modules as an example in Fig. 2, and more attention modules can be stacked in a similar way. The pipeline of our method consists of two parts: 1) we use a CNN to extract deep features, the output feature map of the last convolutional layer is passed to multiple channel attention module [20] to get multiple attention regions. 2) for comparative attention feature learning, we adopt contrastive loss and region-level center loss to supervise multiple attention to focus on multiple discriminative parts of the input sample and make them separable and complete.

The configurations of our proposed MCANet are listed in Table I. Similar to the state-of-the-art method [19], we adopt two convolution layers and four convolutional blocks (conv-blocks) as feature extractor. Each conv-block comprises three convolutional layers with a *bottleneck* in the middle. The feature maps produced by the last conv-block are fed into multiple channel attention modules and then respectively followed by a fully connected layer which contains 768 neurons. The last fully connected layer contains 3755 neurons which is the number of all character classes, and is used to perform the final classification.

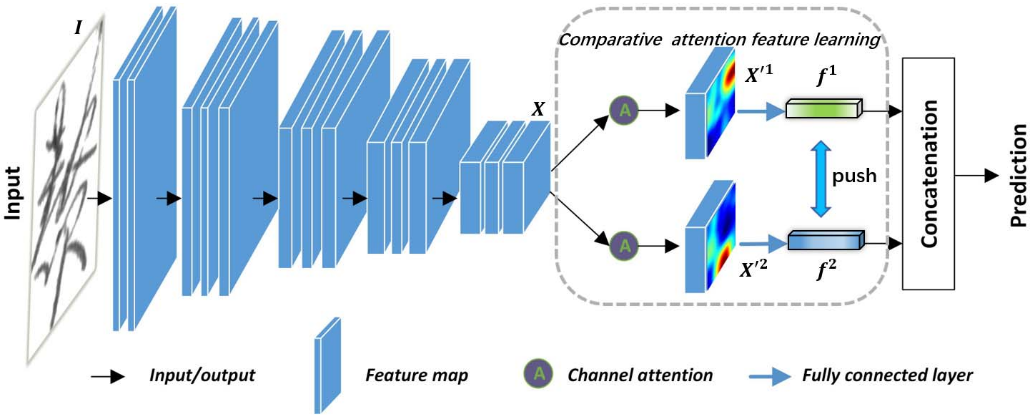


Figure 2. Overview of the proposed multiple comparative attention network for offline Handwritten Chinese Character Recognition. Here we take the case of learning two attention regions as an example. Given the input image I, our proposed model takes the CNN output feature map X as input and outputs two attention maps $X'^{a}(a = 1, 2)$, where a denotes attention. Each attention map X'^{a} are respectively fed into a fully connected layer, the resulting attention feature vectors f^a are constrained by contrastive loss to push attention features away. Further, a region-level center loss is applied to pull the attention region feature f^a from the same class closer. Finally, two attention features are concatenated to adaptively integrates information from different attention regions and then a softmax operator is followed for final prediction.

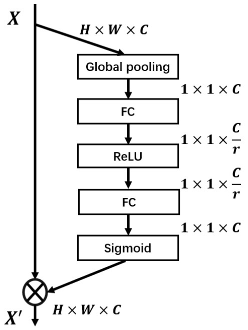


Figure 3. A channel attention module. The module takes X as input and output X' which capture rich contextual information.

B. Multiple Comparative Attention

There has been an attention based recurrent neural network [17] proposed to improve offline HCCR which iteratively uses its previous predictions to update attention and thereafter refine current predictions. Although this model can make a good performance, the limitation is that the current prediction is highly dependent on the previous one, thus the initial error could be accumulated by iterations which lead to sub-optimal results. In addition, they need multiple iterations and complex training procedure with RNN due to gradient vanishing or exploding issue during back-propagation. Compared to the method [17], our approach effectively employs the multiple channel attention modules to capture multiple discriminative regions in the input image and can be trained with more parallelization and less complexity.

Our multi-attention module applies multiple channel attention introduced in SENet [20]. As shown in Figure 3, channel attention module first aggregate the input feature maps X across spatial dimensions $H \times W$ by using global average pooling to generate channel-wise descriptors $z = [z_1, \cdots, z_C]$, where the c-th element of z is calculated by:

$$z_c = \frac{1}{HW}\sum_{h=1}^{H}\sum_{w=1}^{W} x_c(h, w) \tag{1}$$

Then we independently employ a gating mechanism with a Sigmoid activation on $z^a(a = 1, \cdots, A)$, where a denote attention, to calculate the attention weights of each attention module:

$$\alpha^a = \sigma(W_2\delta(W_1 z^a)) = [\alpha^a_1, \cdots, \alpha^a_C] \tag{2}$$

where σ and δ refer to the Sigmoid and ReLU fuction respectively, $W_1 \in \mathbb{R}^{\frac{C}{r}\times C}$, $W_2 \in \mathbb{R}^{C\times\frac{C}{r}}$ and r is reduction ratio which is set to 16 the same as SENet [20]. The final output of each channel attention module is obtained by rescaling $X = [x_1, \cdots, x_C]$ with the attention weight:

$$x'^{a}_{c} = F_{scale}(x_c, \alpha^a_c) = \alpha^a_c \cdot x_c \tag{3}$$

where $X'^{a} = [x'^{a}_{1}, \cdots, x'^{a}_{C}]$ and $F_{scale}(x_c, \alpha_c)$ refers to channel-wise multiplication between the feature map $X \in \mathbb{R}^{H\times W\times C}$ and the scalar α^s_c. Each attended feature map are fed into a fully connected layer $W^a_3 \in \mathbb{R}^{D\times HWC}$ to extract attention-specific feature:

$$f^a = W^a_3 F_{flatt}(X'^{a}) \tag{4}$$

where the operator $F_{flatt}(\cdot)$ flattens a matrix to a vector.

In short, the proposed comparative attention network dedicated to extract A attention feature vectors $f^a(a = 1, \cdots, A)$ for each input image I. For comparative attention feature learning, the contrastive loss and region-level center loss (see section III-C for details.) are combined to make multiple attention features different and compact thereafter obtain discriminative local attention regions and class-specific attention features.

C. Training Loss

The proposed MCANet is supervised by two kinds of loss functions. We minimize classification loss to ensure that each attention map X'^a will locate at important regions for character classification. For comparative attention feature learning, we utilize deep metric learning losses, *i.e.*, contrastive loss and region-level center loss, to enforce multiple attention features focus on different discriminative attention regions of the input sample.

The contrastive loss is applied to attention feature pairs to capture separable attention regions. Given the input image I from a training set S, contrastive loss is defined as

$$L_{contra} = \sum_{I \in S} D(I) \tag{5}$$

where $D(I)$ denote as:

$$D(I) = \sum_{a_1=1}^{A-1} \sum_{a_1 \neq a_2}^{A} max(m - \frac{1}{2}\|f^{a_1} - f^{a_2}\|_2^2, 0) \tag{6}$$

In this way, the contrastive loss is to push the distance between the attention features $f^a(a = 2, \cdots, A)$ apart by a margin m in an input sample I.

Except the contrastive loss described above, we further introduce a region-level center loss to enhance comparative attention feature learning. The region-level center loss, which can be formulated as in Eq.7, is applied to decrease the distance between each attention region feature f^a from the same character category so that each attention map X'^a will be activated in the same parts.

$$L_{center} = \frac{1}{2} \sum_{a=1}^{A} \sum_{i=1}^{N} \|f_i^a - c_{y_i}^a\|_2^2 \tag{7}$$

where N is the number of character classes and $c_{y_i}^a \in \mathbb{R}^d$ is the center of a-th attention region feature of class y_i, with d denoting the dimension of features. The $c_{y_i}^a$ is updated like in *Center loss* [8].

Generally, the model MCANet jointly optimizes classification and metric learning loss, the total loss can be denoted as

$$L_{total} = L_{cls} + \lambda(L_{center} + L_{contra}) \tag{8}$$

where λ is a hyper-parameter which controls the trade-off between two kinds of loss.

IV. Experiments

In this section, we share the implementation details and evaluate our proposed models in terms of recognition performance as well as visualization analysis.

Table I
THE CONFIGURATIONS OF PROPOSED MCANET.

<table>
<tr><th>Layers</th><th colspan="2">Configurations</th><th>Output Size</th></tr>
<tr><td>Conv-Layers</td><td colspan="2">3 × 3 conv 64, BN, ReLU
3 × 3 conv 64, BN, ReLU</td><td>96 × 96</td></tr>
<tr><td>Pool1</td><td colspan="2">3 × 3 avg-pool stride 2</td><td>48 × 48</td></tr>
<tr><td>Conv-Block1</td><td colspan="2">3 × 3 conv 96, BN, ReLU
3 × 3 conv 64, BN, ReLU
3 × 3 conv 96, BN, ReLU</td><td>48 × 48</td></tr>
<tr><td>Pool2</td><td colspan="2">3 × 3 avg-pool stride 2</td><td>24 × 24</td></tr>
<tr><td>Conv-Block2</td><td colspan="2">3 × 3 conv 128, BN, ReLU
3 × 3 conv 96, BN, ReLU
3 × 3 conv 128, BN, ReLU</td><td>24 × 24</td></tr>
<tr><td>Pool3</td><td colspan="2">3 × 3 avg-pool stride 2</td><td>12 × 12</td></tr>
<tr><td>Conv-Block3</td><td colspan="2">3 × 3 conv 256, BN, ReLU
3 × 3 conv 192, BN, ReLU
3 × 3 conv 256, BN. RelU</td><td>12 × 12</td></tr>
<tr><td>Pool4</td><td colspan="2">3 × 3 avg-pool stride 2</td><td>6 × 6</td></tr>
<tr><td>Conv-Block4</td><td colspan="2">3 × 3 conv 448, BN, ReLU
3 × 3 conv 256, BN, ReLU
3 × 3 conv 448, BN, ReLU</td><td>6 × 6</td></tr>
<tr><td>Multi-Attention</td><td>Channel Attention
FC, ReLU</td><td>Channel Attention
FC, ReLU</td><td>6 × 6
768</td></tr>
<tr><td>Aggregation</td><td colspan="2">Concatenation
Dropout</td><td>768 × 2</td></tr>
<tr><td>Output</td><td colspan="2">3755-dim Softmax</td><td>3755</td></tr>
</table>

A. Datasets

In our experiments, we use CASIA-HWDB1.0 and CASIA-HWDB-1.1 [27] datasets which were written by 300 and 420 persons respectively to train the proposed networks. The overall training datasets contain 2,678,424 samples belonging to 3755 different character classes. The test dataset is used in ICDAR-2013 offline HCCR competition [1], which contains 224,419 samples written by 60 persons different from the writers of training data.

B. Implementation Details

Experimental codes are implemented in pytorch [28]. All the convolutional layers in our proposed network architecture have kernel size of 3x3, strides of 1 and padding of 1. Following Melnyk *et al.* [19], we do not use bias for all the convolutional layers. In our experiments, the input gray images are resized to 96×96 without any data augmentation. We use the method of mini-batch stochastic gradient descent (SGD) with the momentum set to 0.9 and mini-batch set to 256 during training. Initial learning rate is set to 0.1 and we reduce it $\times 0.1$ after every 4 epochs. We empirically set the margin m in Eq.6 to 40, the weight parameter λ in Eq.8 set to 0.1, and each attention feature center of each class is initialized with Gaussian distribution, and the mean and standard deviation is (0, 1) respectively. The learning rate for each attention center is set to 0.5 in our experiments.

C. Ablation Studies

To analyze the effectiveness of our proposed method, we conduct extensive ablation experiments with different set-

Table II
PERFORMANCE ON ICDAR2013 OFFLINE HANDWRITTEN CHINESE CHARACTER RECOGNITION COMPETITION DATASET WITH DIFFERENT CONFIGURATIONS. "CA" DENOTES CHANNEL ATTENTION, "CL" STANDS FOR CONTRASTIVE LOSS IN EQ.6, "CT" MEANS IMAGE-LEVEL CENTER LOSS [8] AND "RCT" STANDS FOR REGION-LEVEL CENTER LOSS IN EQ.7

.

Method	Accuracy(%)
Baseline	97.49
Baseline + 1CA	97.52
Baseline + 2CA	97.56
Baseline + 3CA	97.57
Baseline + 2CA + CL	97.59
Baseline + 2CA + CT	97.59
Baseline + 2CA + RCT	97.61
Baseline + 2CA + CL + RCT	**97.66**

tings. Table II demonstrates the performance on ICDAR2013 offline HCCR competition data with different configurations. Our baseline model utilizes the convolutional layers as feature extractor in Table I followed by two FC layers to do classification which achieves an accuracy of 97.49% on the test set.

Multi-attention module: As demonstrated in Table II, the baseline model will improve 0.07% when channel attention module number $A = 2$. We also note that more attention regions ($A = 3$) lead to limited improvements as 2 channel attention can cover almost all discriminative sub-regions and extended attention region feature may already have been encoded. Considering the computational cost with more attention regions, we choose to follow baseline CNN with two attention module ($A = 2$) in our experiments.

Multi-task loss: To further enhance comparative attention feature learning, we combine deep metric learning loss with classification loss. Using contrastive loss and region-level center loss will lead to 0.03% and 0.05% performance improvements respectively compared to the baseline model using multi-attention. In addition, our proposed region-level center loss gains 0.02% improvement than image-level center loss (97.61% *vs.* 97.59%). Finally, our multi-attention network constrained by all three loss achieves an accuracy of 97.66%.

Visualization of attention maps: To get an intuitional understanding of our multiple comparative attention network, we visualize the learned attention maps as shown in Fig. 4. It is seen that our MCANet can capture separate and complete attention with discriminative regions of input samples. Moreover, the multiple comparative attention can locate at the same regions regardless of large variant handwriting styles across individuals. It indicates the robustness of our proposed method.

D. Comparison with Other Methods

The results of other competitive methods on ICADR2013 offline HCCR competition dataset are summarized in Table III. As it can be seen, the proposed method achieves an improvement to recognition accuracy of 97.66%, with 0.05% relative gain compared to the previous state-of-the-art method Melnyk-Net [19]. This improvement mainly derives from the distinctive attention localization which is demonstrated in Fig. 1(b).

Table III
ACCURACIES OF DIFFERENT METHODS FOR ICADR2013 DATASET. ALL METHODS ARE TRAINED ON CASIA-HWDB1.0-1.1 DATASETS. "ENSEMBLE" INDICATES WHETHER ENSEMBLE LEARNING STRATEGY IS USED.

Method	Accuracy(%)	Ensemble
Human Level Performance [1]	96.31	n/a
HCCR-Gabor-GoogLeNet [3]	96.35	no
HCCR-GoogLeNet-Ensemble-10 [3]	96.74	yes(10)
Residual-34 [4]	97.36	no
STN-Residual-34 [4]	97.37	no
DCNN-Similarity ranking [29]	97.07	no
Ensemble DCNN-Similarity ranking [29]	97.64	yes(4)
DirectMap + ConvNet [16]	96.95	no
DirectMap + ConvNet + Ensemble-3 [16]	97.12	yes(3)
DirectMap + ConvNet + Adaption [16]	97.37	no
M-RBC + IR [17]	97.37	no
HCCR-CNN9Layer [18]	97.30	no
HCCR-CNN12Layer [18]	97.59	no
Melnyk-Net [19]	97.61	no
Our MCANet	**97.66**	no

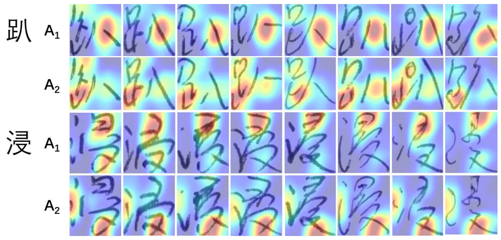


Figure 4. Visualization of attention regions generated by our proposed MCANet. A_1 and A_2 are from layer X'^1 and X'^2 respectively. The first and last two rows show attention maps from two different classes of characters. Characters from the same class are written by different persons.

Similar to our work, the method proposed by Yang *et al.* [17] which utilizes multi-scale residual block cascade and attention based iterative refinement module to localize sub-regions of input images to distinguish visually similar characters. However, our method achieves a relative 0.29% improvement without residual connections and recurrent neural network.

V. CONCLUSION

In this study, we present a simple yet effective multiple comparative attention networks (MCANet) to improve offline Handwritten Chinese Character Recognition performance. Our model learns multiple channel attention and supervised by the multi-task loss which enforces multiple attention features concentrate on separate local regions and encourage the same class characters to harvest similar attention features to increase the inter-class distance and decrease the intra-class gap, resulting in state-of-the-art accuracy for single-network methods trained only on handwritten datasets.

Acknowledgment

This research was supported by the National Natural Science Foundation of China under grant No. 61733007. Dr. Xiang Bai was supported by the National Program for Support of Top-notch Young Professionals and the Program for HUST Academic Frontier Youth Team.

References

[1] 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.

[2] C.-L. Liu, M. Koga, and H. Fujisawa, "Lexicon-driven segmentation and recognition of handwritten character strings for japanese address reading," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 11, pp. 1425–1437, 2002.

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

[4] Z. Zhong, X.-Y. Zhang, F. Yin, and C.-L. Liu, "Handwritten chinese character recognition with spatial transformer and deep residual networks," in *2016 23rd International Conference on Pattern Recognition (ICPR)*. IEEE, 2016, pp. 3440–3445.

[5] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 618–626.

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

[7] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *null*. IEEE, 2005, pp. 539–546.

[8] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *European conference on computer vision*. Springer, 2016, pp. 499–515.

[9] F. Kimura, K. Takashina, S. Tsuruoka, and Y. Miyake, "Modified quadratic discriminant functions and the application to chinese character recognition," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 1, pp. 149–153, 1987.

[10] O. L. Mangasarian and D. R. Musicant, "Data discrimination via nonlinear generalized support vector machines," in *Complementarity: Applications, Algorithms and Extensions*. Springer, 2001, pp. 233–251.

[11] C.-L. Liu, H. Sako, and H. Fujisawa, "Discriminative learning quadratic discriminant function for handwriting recognition," *IEEE Transactions on Neural Networks*, vol. 15, no. 2, pp. 430–444, 2004.

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

[13] F. Yin, Q.-F. Wang, X.-Y. Zhang, and C.-L. Liu, "Icdar 2013 chinese handwriting recognition competition," in *2013 12th International Conference on Document Analysis and Recognition*. IEEE, 2013, pp. 1464–1470.

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

[15] S. Zhou, J.-N. Wu, Y. Wu, and X. Zhou, "Exploiting local structures with the kronecker layer in convolutional networks," *arXiv preprint arXiv:1512.09194*, 2015.

[16] X.-Y. Zhang, Y. Bengio, and C.-L. Liu, "Online and offline handwritten chinese character recognition: A comprehensive study and new benchmark," *Pattern Recognition*, vol. 61, pp. 348–360, 2017.

[17] X. Yang, D. He, Z. Zhou, D. Kifer, and C. L. Giles, "Improving offline handwritten chinese character recognition by iterative refinement," in *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, vol. 1. IEEE, 2017, pp. 5–10.

[18] X. Xiao, L. Jin, Y. Yang, W. Yang, J. Sun, and T. Chang, "Building fast and compact convolutional neural networks for offline handwritten chinese character recognition," *Pattern Recognition*, vol. 72, pp. 72–81, 2017.

[19] P. Melnyk, Z. You, and K. Li, "A high-performance cnn method for offline handwritten chinese character recognition and visualization," *arXiv preprint arXiv:1812.11489*, 2018.

[20] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[21] V. Mnih, N. Heess, A. Graves *et al.*, "Recurrent models of visual attention," in *Advances in neural information processing systems*, 2014, pp. 2204–2212.

[22] C. Cao, X. Liu, Y. Yang, Y. Yu, J. Wang, Z. Wang, Y. Huang, L. Wang, C. Huang, W. Xu *et al.*, "Look and think twice: Capturing top-down visual attention with feedback convolutional neural networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2956–2964.

[23] M. Sun, Y. Yuan, F. Zhou, and E. Ding, "Multi-attention multi-class constraint for fine-grained image recognition," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 805–821.

[24] T. Bluche, "Joint line segmentation and transcription for end-to-end handwritten paragraph recognition," in *Advances in Neural Information Processing Systems*, 2016, pp. 838–846.

[25] A. Miech, I. Laptev, and J. Sivic, "Learnable pooling with context gating for video classification," *arXiv preprint arXiv:1706.06905*, 2017.

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

[27] C.-L. Liu, F. Yin, D.-H. Wang, and Q.-F. Wang, "Casia online and offline chinese handwriting databases," in *2011 International Conference on Document Analysis and Recognition*. IEEE, 2011, pp. 37–41.

[28] 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.

[29] C. Cheng, X.-Y. Zhang, X.-H. Shao, and X.-D. Zhou, "Handwritten chinese character recognition by joint classification and similarity ranking," in *2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR)*. IEEE, 2016, pp. 507–511.