2016 15th International Conference on Frontiers in Handwriting Recognition

Handwritten/Printed Receipt Classification using Attention-Based Convolutional Neural Network

Fan Yang, Lianwen Jin+, Weixin Yang, Ziyong Feng, Shuye Zhang
School of Electronic and Information Engineering, South China University of Technology
Guangzhou, China
yfjiaren@foxmail.com, +eelwjin@scut.edu.cn, wxy1290@163.com

Abstract—**This paper presents an approach for the classification of handwritten and printed receipts based on a convolutional neural network (CNN). One of the main challenges related to such classification is the diversity of the background interference in the receipt images. To overcome this problem, we propose a new technique named "attention-based CNN" (ABCNN), inspired by the concept of "attention" in visual neuroscience. This approach helps us to focus on the receipt in an image without bounding box annotation. Our experimental results showed that the proposed ABCNN (i) significantly improves the classification accuracy compared to normal CNN (from 95% to 98.25%), and (ii) enables the network to process images directly without object detection, and (iii) it is faster to train and test the network.**

Keywords - Handwritten/printed receipt classification; convolutional neural network; attention-based approach; region of interest

I. INTRODUCTION

With the increase of digitized documents, automatic document analysis has become extremely important. This trend gives rise to the issue of handwritten/printed receipt analysis. Given that there are different analysis methodologies for handwritten and printed receipts, there is a need for a means of separating them. Several previous works have attempted to address related classification problems using machine learning methods [7] [8] [9] [10].

Compared to traditional machine learning methods, convolutional neural networks CNN [1] [2] have shown great promise in the area of image analysis [3] [4]. Recently, CNN-based applications have been used to solve a variety of problems including image classification [19] [20] and object detection [21] [22]. Its high precision and adaptability mean that it has great potential for application to the field of document classification [5] [6]. It offers a real solution to the problem of receipt classification in that we can design and train a CNN to classify receipts.

The main problem limiting network performance is the variability of the receipt images, especially the different kinds of background interference. To classify images with a complex background, the most common approach involves performing object detection and then classifying the detected objects. Most object detection methods require detailed location information. For example, bounding-box supervision is one of the most common methods for object detection in complex scenes [20]. However, the labeling of a set of training images requires extensive human effort, and unavoidably involves subjective judgments. Thus, it is necessary to find a novel way of eliminating the irrelevant information and focusing on the region of interest in the images, without the limitation of an explicit detection procedure. This is similar to the concept of "attention" in the visual cortex of the human brain [15] [16]. Human visual processes appear to select a subset of the available sight information of interest before further processing [17], thus helping to reduce the complexity of the analysis [18] and to focus on the attention area in the line of sight.

Inspired by the attention models proposed in the area of visual neuroscience, we propose an attention-based CNN (ABCNN) to overcome the problem of background interference. The ABCNN introduces an attention function for specifying the attention area on which the learning model should focus. Then, the attention area in the receipt image is used for training. The remaining information, such as background noise, is discarded. The ABCNN combines the attention mechanism with CNN. The parameters of the attention function are determined according to an analysis of the activated area in the feature map learned automatically by the CNN. Experiments showed that the ABCNN improves network performance with less computational complexity and higher accuracy (from 95% to 98.25%).

The remainder of this paper is organized as follows. In Section II, we introduce the CNN we designed for the receipt classification task. Then, we present our approach, that is, the attention-based approach (ABA) with CNN, in Section III. In Section IV, we present and analyze the experimental results to show how the ABCNN improves the network performance. Finally, we conclude the paper in Section V.

II. CONVOLUTIONAL NEURAL NETWORK

The deep CNN architecture employed to classify handwritten and printed receipts consists of convolutional (Conv) layers, pooling (Pool) layers, Rectified Linear Units (ReLUs) and fully connected (FC) layers.

2167-6445/16 $31.00 © 2016 IEEE
DOI 10.1109/ICFHR.2016.73

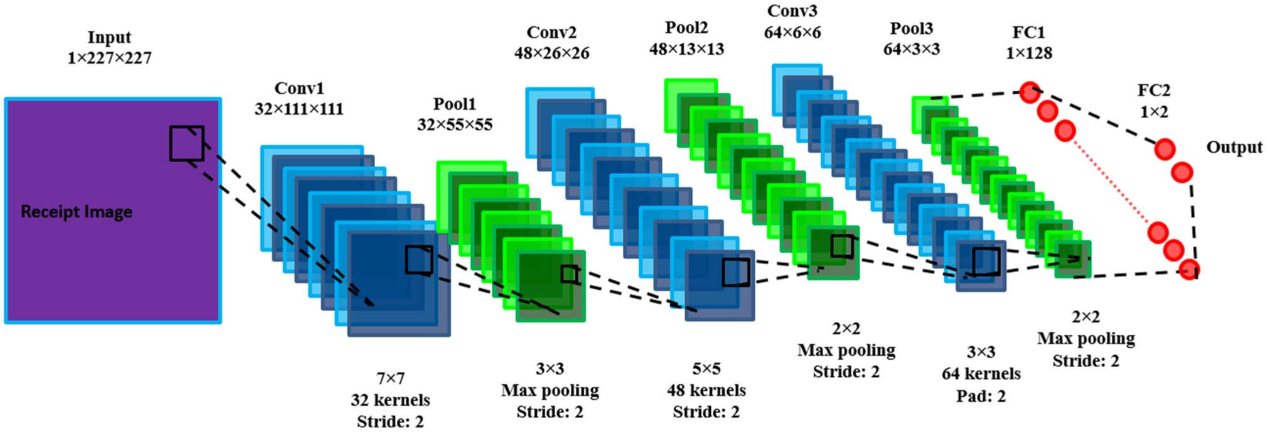
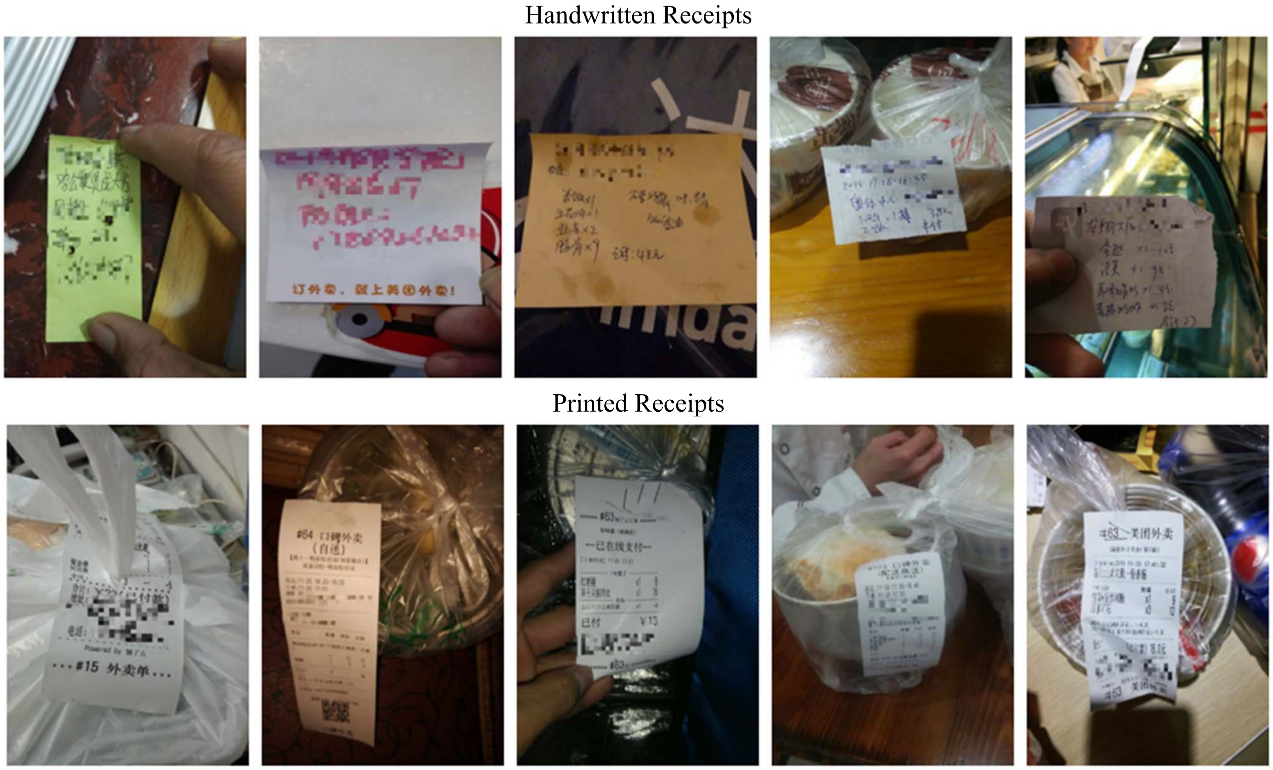


Figure 1. The proposed architecture of the CNN.


Handwritten Receipts

Printed Receipts

Figure 2. Example images from the receipt dataset with variations in multi-resolution and background.

A. Network Architecture

The architecture of our CNN model is depicted in Fig. 1. The model consists of three convolutional and two fully connected layers. The first convolutional layer processes the input shape of 1 × 227 × 227 with 32 kernels measuring 1 × 7 × 7 using a stride (step size) of 2 pixels. Note that, for the subsequent study described in Sections III and IV, the input size of the CNN (the input size of the first convolutional layer) may be changed accordingly while maintaining the structure of the other layers in the network. The number/shapes of the kernels are (48/32 × 5 × 5) and (64/48 × 3 × 3) for the second and the third convolutional layers, respectively. Max-pooling is performed for each convolutional layer. Each convolutional and fully connected layer is followed by a ReLU, while the output layer adopts two-way Softmax as the activation function for obtaining the classification result.

B. Training

The objective function of the network was minimized using stochastic gradient descent. A batch size of 100 was used. The learning rate, momentum, and weight decay were set to 0.01, 0.9, and 0.005, respectively. The weights of the network were initialized under a Gaussian distribution with a standard deviation of 0.01. To avoid overfitting, dropout [13] was applied to the first fully connected layer with a dropout ratio of 0.5. Training of the network was done using the Caffe deep-learning framework [14].

III. Attention Based Approach Using CNN

The challenge of the receipt image classification task arises from the fact that, for different images, there is a wide range of variations caused by different resolutions, background, lighting, and noise. Among these, the diversity of the background is the main interference affecting the classification, as illustrated in Fig. 2. These types of interference may allow the CNN to be inversely affected by irrelevant information and focus less on the receipts, thus limiting the network performance. Although a complicated background can be reduced after the completion of detection, it requires considerable human effort to label location information in the images.

In an attempt to solve this problem without detection, we consider the concept of "Attention" in visual neuroscience. Research has shown that human visual processes appear to select a subset of the available visual information before further processing [17], in order to increase the information precision as well as to reduce the complexity of the analysis [18]. And the focus of the attention appears to be located around the center of the visual field. Inspired by these studies, we propose a technique named the "attention-based approach" (ABA). This approach lets the model focus on a region of interest in an image, in the same way as the visual process in the human brain. The process of the approach is shown in Fig. 3. This approach helps us extract relevant information to the model and to avoid background clutter. The details are described below.

First, let Γ be a receipt image with a complex background. Second, we use "attention" to specify the region of the receipt in the image, that is, to extract the "attention area" (with less interference) from the image Γ. For example, the image Γ may be a 256 × 256 image and the region of interest (ROI) could be a 192 × 192 cropped image. Let ξ represent this cropped image. Third, we define

$$\delta = [x, y, w, h], \quad (1)$$

(the attention-based position, shown in Fig. 3) as the parameters of the ABA used to transform the image Γ. Finally, we obtain the function

$$\mathcal{F}(\Gamma;\delta) = \xi(\delta), \quad (2)$$

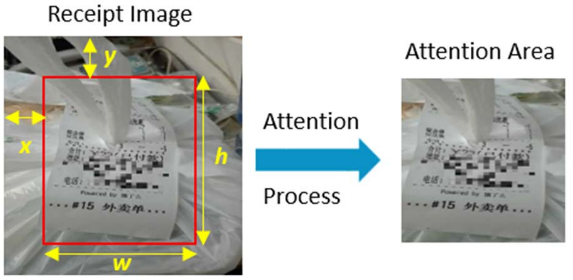


Figure 3. The process of the proposed attention-based approach.

as the attention function to warp the image Γ to the attention area ξ. The intuition is that, ξ (δ) is a patch of Γ according to the parameter δ. The issue arising here is the combination of the ABA with the CNN.

We believe that we have found a solution to this issue in the feature maps learned by the CNN adopted in Section II. As shown in [23], even when trained with the ground-truth of the image labels only, the CNN can predict the approximate locations of the objects in images. During training, the CNN can learn a special response to the location of an object without explicitly searching for the specific location. It was shown in previous studies that this kind of location prediction has a level of performance that is comparable to a fully supervised model using bounding box annotation for training [23]. Therefore, we can train a CNN (as discussed in Section II) and then analyze the output feature map of the convolutional and pooling layers, with the aim of finding a correspondence between the localization of the object and the response in the feature maps. The intuition is that the activated (high-value) area of the feature map corresponds to the "attention area" in the input image, on which the learning model should focus. Suppose that we have found a useful representation (activated feature area, set as v) in the feature map (set as r). It would be instructive to apply the attention function $\mathcal{F}$ to the image Γ using the following formula:

$$\frac{area(\xi(\delta))}{area(\Gamma)} = \frac{w \times h}{area(\Gamma)} \approx \frac{area(v)}{area(r)}, \quad (3)$$

where area(r) denotes the area of the feature map and area(v) represents the area of the high-value region. The parameters w and h are as follows:

$$w \in \left[width(\Gamma) \times \sqrt{\frac{area(v)}{area(r)}}, width(\Gamma) \times \sqrt{\frac{area(v)}{area(r)}} + k\right], \quad (4)$$

$$h \in \left[height(\Gamma) \times \sqrt{\frac{area(v)}{area(r)}}, height(\Gamma) \times \sqrt{\frac{area(v)}{area(r)}} + k\right]. \quad (5)$$

Here +k is the tolerance item where k refers to the convolutional or the pooling ratio of the analyzed layer. Therefore, the parameters x and y could be calculated by

$$x = \frac{width(\Gamma) - w}{2}, \quad (6)$$

$$y = \frac{height(\Gamma) - h}{2}. \quad (7)$$

By using the attention-based approach with CNN, we can discard the complex background and focus on the useful region automatically learned by the CNN.

IV. Experiments

A. Dataset

We construct a new dataset for handwritten and printed receipt classification. The printed receipts differ from handwritten ones in that, in the case of the latter, the writers can freely write text, add annotations, and draw tables or even figures, etc. Within each image type, there is a wide range of variability caused by lighting, multi-resolution, noise, and especially the background. This point is illustrated by the receipt images shown in Fig. 2. The dataset contains a total of 1109 labeled images (312 handwritten and 797 printed receipts). Of these, 75% are used for training and the remainder for testing.

B. Preprocessing

The classification task relies on the difference between the randomness of the handwritten receipts and the strict document structure of the printed ones. To analyze the difference, it is not necessary to have a very high-resolution image in which the text is completely recognizable. Therefore, the images are downsampled to reduce the computational complexity of the system. For the CNN proposed in Section II, the images are downsampled to 227 × 227. For further attention-based experiments, the downsampled size is 256×256. The input images are then mean-normalized to remove the intensity interference. Mean compensation is applied to both the training and testing images. We convert the three-channel (RGB) images to one-channel (grayscale) images to eliminate the interference caused by the colors, and to further reduce the computational complexity of the first convolutional layer.

C. Evaluation of Attention-Based Approach

The CNN proposed in Section II (denoted CNN_{227}), and depicted in Fig. 1 is evaluated. As can be seen from Table I, the total accuracy is 93.31% (90.32% and 93.94% for the handwritten and printed classes, respectively). Here, we train a second CNN with an input size of 256 × 256 (denoted CNN_{256}) as a baseline for the further evaluation of ABCNN. The architecture of CNN_{256} is the same as that of CNN_{227} except for the input size of the first convolutional layer. Meanwhile, the images are downsampled to match the input size. The second model attains a total accuracy of 95.00%, which may result from more information being fed into the model.

To evaluate the proposed ABCNN, we first resize all the images to 256 × 256, so that Γ represents a 256 × 256 image. Then, parameters w and h of the attention function $\mathcal{F}$ are both set to 227, resulting in a network with the same input size (denoted as ABCNN_{227}) as that of the first network (CNN_{227}) for comparison. According to the discussion in Section III, it is appropriate to set parameters x and y using Eqs. (6) and (7), respectively. For different parameter settings, we change the input size of the CNN (the input size of the first convolutional layer) according to the size of $\xi(\delta)$, that is, $w \times h$, while maintaining the other part of the network.

TABLE I. Evaluation of Attention-Based CNN

Attention Based or Not	**Model**	**Test Accuracy (%)**		
		Average Accuracy	***Handwritten Accuracy***	***Printed Accuracy***
NO	CNN_{227}	93.31	90.32	93.94
	CNN_{256}	95.00	90.32	96.29
YES	ABCNN_{227}	**97.00**	**97.00**	**97.00**
	ABCNN_{192}	**98.25**	**99.00**	**97.50**

For a given input size, the ABA realizes an improvement with an overall accuracy of 3.69% (6.68% and 3.06% for the handwritten and printed classes, respectively) as shown in the first and third rows of Table I. Compared to the baseline (CNN_{256}), the average accuracy of ABCNN_{227} is improved from 95.00% to 97.00%, thus accounting for the efficacy of the method.

D. Investigation of attention area size

To investigate the influence of different parameter settings in the proposed approach, we undertake the following experiments.

As mentioned in Section III, evaluating the output feature map of CNN helps us to find the distribution under the data set and to make good decisions related to the parameter settings for the ABCNN. To evaluate the output feature map for the CNN_{256} (trained in subsection C), we run a forward propagation with the training receipt images. Then, after extracting the feature maps for each convolutional and max-pooling layer, we save the heat map for each channel of the feature maps. By analyzing the heat maps, we can find some interesting responses. The 28th channel of the feature map of the second max-pooling layer highlights the area of the receipt in the image, while the 45th channel is relevant to the background interference. Several examples are shown in Fig. 4. Then, we average the heat maps of the feature maps of the second max-pooling layer over the training set, the result of which confirms our findings. The visualization of this averaged network response is depicted in Fig. 5. Thus, we could select the area with high values in feature channel A (red area in Fig. 5) as the attention area to eliminate the background interference. We undertake several experiments on different attention area sizes ranging from 227 to 146, according to the averaged heat map for feature channel A, shown in Fig. 5.

The experimental results are shown in Fig. 6. Again, CNN_{256} is taken as a baseline for checking the improvement realized by the ABA. We can see that the network for which the attention area is set to 192 × 192 (denoted as ABCNN_{192}) has the best classification performance. The ABCNN_{192} outperforms the baseline CNN_{256} by 3.25% in terms of average accuracy (from 95% to 98.25%). As the values of w and h decrease (from 256 to 192), the network processes less

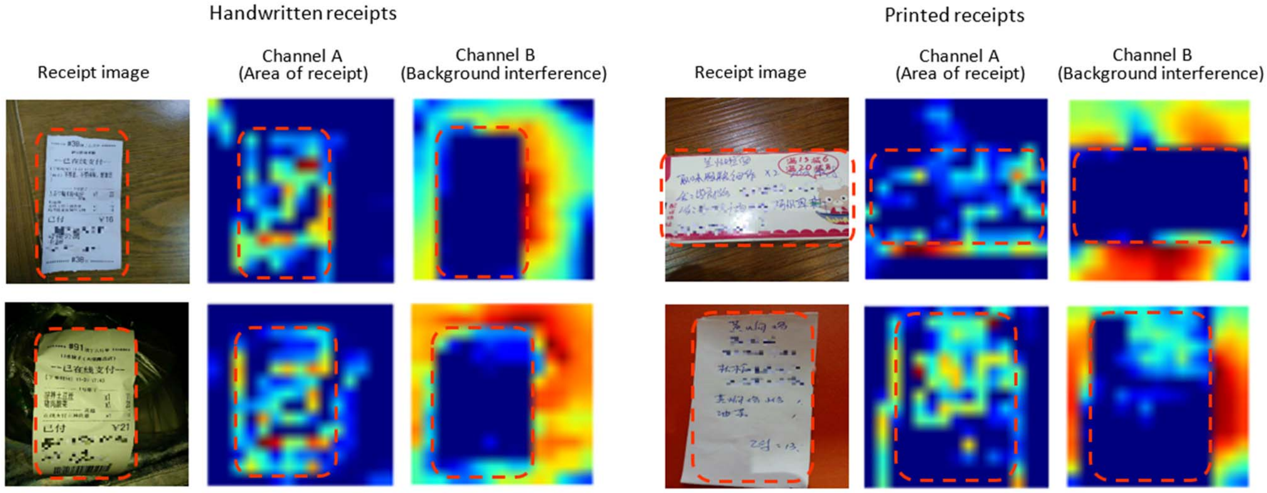


Figure 4. Network response in the second max-pooling layer, corresponding to handwritten and printed receipt images.

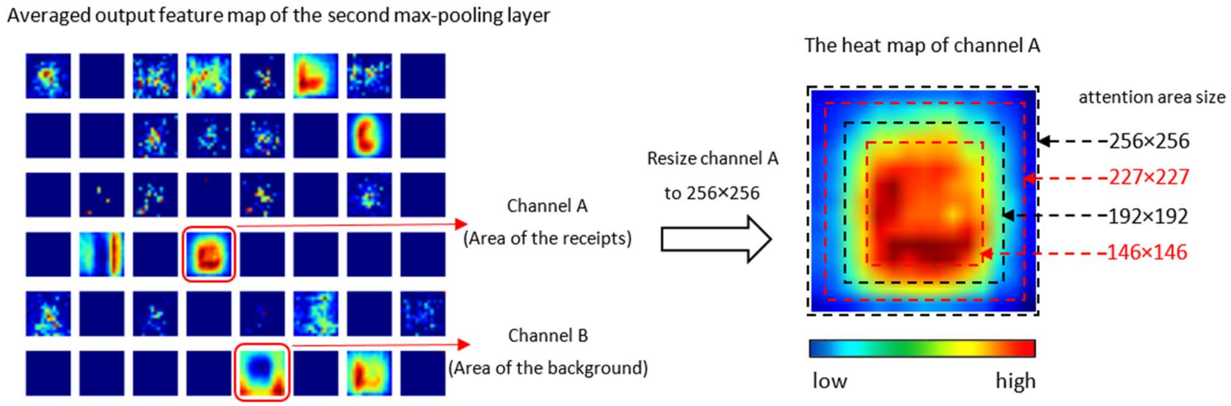


Figure 5. Averaged network response feature map of the second max-pooling layer over the training set and the feature channel A.

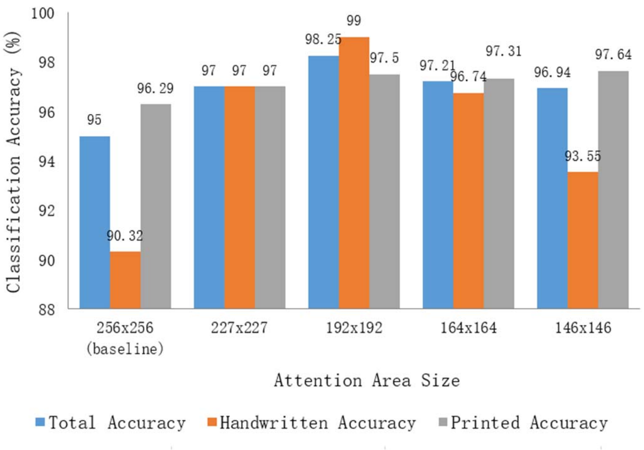


Figure 6. Evaluation of different attention area size of the attention-based approach.

information, thus the model focuses more on the attention area in the image and avoids the background clutter. However, if w and h are set to overly small values (e.g. 146), some useful information will be discarded along with the background interference, thus adversely affecting the classification accuracy. Therefore, an appropriate setting for the size of the attention area should be determined during the experiments. It should also be noted that even when parameters w and h are both set to 146 so that the model processes only about 1/3 of the information of the original image (from 256 × 256 to 146 × 146), the classification accuracy still rises by about 1.94% comparing with the baseline CNN_{256} model. This is strong evidence for the ability of the attention model to improve the CNN performance.

V. Conclusion

In this paper, we solve the problem of handwritten and printed receipt classification by using CNN. Inspired by the attention model developed in visual neuroscience, we propose an attention-based CNN model as a means of avoiding background interference in receipt images. We also prove that the ABCNN improves the performance of the network with less computational complexity and helps to avoid explicit object detection. In addition, several experiments are presented to evaluate different parameter settings for the proposed approach. We achieve the highest level of accuracy by using our ABCNN with an attention area size of 192 × 192. The average error rate for the best network is 1.75%, which corresponds to a relative error reduction rate of 65.00% compared to the baseline CNN without ABA.

Acknowledgment

This research is supported in part by NSFC (Grant No.: 61472144), the National Key Research & Development Plan of China (No. 2016YFB1001405), GDSTP (Grant No.: 2014A010103012, 2015B010101004, 2015B010130003, 2015B010131004) , GDUPS (2011), and the Fundamental Research Funds for the Central Universities(no. D2157060).

References

[1] A. Krizhevsky, I. Sutskever and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Network," Neural Information Processing Systems (NIPS), 2012.

[2] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," Neural computation, 1989.

[3] D. Cireşan, U. Meier, J. Masci, LM. Gambardella and J. Schmidhuber, "Flexible, High Performance Convolutional Neural Networks for Image Classification," International Joint Conference on Artificial Intelligence, pp. 1237-1242.

[4] D. Ciresan, U. Meier and J. Schmidhuber, "Multi-column Deep Neural Networks for Image Classification," Computer Vision and Pattern Recognition (CVPR), IEEE, 2012.

[5] Y. Lecun, L. Bottou, Y. Bengio and P. Haffner, "Gradient-based learning applied to document recognition," Proc IEEE, 1998, pp. 2278-2324.

[6] P. Y. Simard, D. Steinkraus and J. C. Platt, "Best Practices for Convolutional Neural Networks Applied to Visual Document Analysis," Proceedings of the Seventh International Conference on Document Analysis and Recognition, IEEE Computer Society, 2003, Vol 2, pp. 958.

[7] E. Zemouri and Y Chibani, "Machine printed handwritten text discrimination using Radon transform and SVM classifier," Intelligent Systems Design and Applications (ISDA), IEEE, 2011, pp. 1306-1310.

[8] K. Zagoris, I. Pratikakis, A. Antonacopoulos, B. Gatos and N. Papamarkos, "Handwritten and Machine Printed Text Separation in Document Images Using the Bag of Visual Words Paradigm," Proceedings of the International Conference on Frontiers in Handwriting Recognition, IEEE Computer Society, 2012, Vol.7202, pp. 103-108.

[9] A. Sadïani, A. K. EchiI and A. Belaïd, "Identification of MachinePrinted and Handwritten Words in Arabic and Latin Scripts," International Conference on Document Analysis and Recognition, IEEE Computer Society, 2013, pp. 798-802.

[10] A. Jindal and M. Amir, "Automatic Classification of Handwritten and Printed Text in ICR Boxes," International Advance Computing Conference (IACC), IEEE, 2014, pp. 1028-1032.

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

[12] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN features off-the-shelf: An astounding baseline for recognition," In CVPR, 2014, pp. 512–519.

[13] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever. and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," J. Machine Learning, 2014, pp. 1929–1958.

[14] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," arXiv preprint arXiv:1408.5093, 2014.

[15] M. I. Posner and C. D. Gilbert, "Attention and primary visual cortex," Proc. of the National Academy of Sciences, vol. 96, no. 6, March 1999.

[16] E. A. Buffalo, P. Fries, R. Landman, H. Liang, and R. Desimone, "A backward progression of attentional effects in the ventral stream", PNAS, vol. 107, no. 1, Jan. 2010, pp. 361–365.

[17] J.K. Tsotsos, S.M. Culhane, W.Y.K. Wai, Y.H. Lai, N. Davis, and F.Nuflo, "Modelling Visual Attention via Selective Tuning," Artificial Intelligence, vol. 78, no. 1-2, Oct. 1995, pp. 507–545.

[18] E. Niebur and C. Koch, "Computational Architectures for Attention," R. Parasuraman, ed., The Attentive Brain,. Cambridge, Mass: MIT Press, 1998, pp. 163–186.

[19] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," arXiv:1405.3531v2, 2014.

[20] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1717-1724.

[21] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 580-587.

[22] B Hariharan, P Arbeláez, R Girshick, and J. Malik, "Simultaneous detection and segmentation," Computer vision–ECCV 2014. Springer International Publishing, 2014, pp. 297-312.

[23] M Oquab, L Bottou, I Laptev and J Sivic, "Is object localization for free? Weakly-supervised learning with convolutional neural networks," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015.