Personal and Ubiquitous Computing (2018) 22:1029–1037
https://doi.org/10.1007/s00779-018-1168-8

ORIGINAL ARTICLE

Image spam filtering using convolutional neural networks

Fan Aiwan[1] · **Yang Zhaofeng**[1]

Received: 9 March 2018 / Accepted: 25 May 2018 / Published online: 9 July 2018

Abstract
Spammers often embed text into images in order to avoid filtering by text-based spam filters, which result in a large number of advertisement spam images. Garbage image recognition has become one of the hotspots in the field of Internet spam filtering research. Its goal is to solve the problem that traditional spam information filtering methods encounter a sharp performance decline or even failure when filtering spam image information. Based on the clustering algorithm, this paper proposes a method to expand the data samples, which greatly improves the number of high-quality training samples and meets the needs of model training. Then, we train a convolutional neural networks using the enlarged data samples to recognize the SPAM in real time. The experimental results show that the accuracy of the model is increased by more than 14% after using the method of data augmentation. The accuracy of the model can be improved by 6% compared with other methods of data augmentation. Combined with convolutional neural networks and the proposed method of data augmentation, the accuracy of our SPAM filtering model is 7–11% higher than that of the traditional method.

Keywords Data augmentation · Convolutional neural networks · Image recognition · Image spam filtering

1 Introduction

The problem that real-time spam email filtering can be classified as image classification. Convolution neural network has become a standard model to solve the problem of image classification, and the records of a series of image recognition competitions have been broken by it. These famous models are [1–8] respectively according to the proposed time, and the performance of these models is improved with time. On the one hand, the above exciting results can not be separated from the elaborate model, but also from the advanced algorithms. In the convolution neural network, the commonly used stochastic optimization algorithms are SGD, AdaGrad [14], AdaDelta [15], Adam [16] and Amme [17]. The activation function also plays an important role in image recognition using deep convolution neural network. For the characteristics of the image, the best performance of the current activation function is ReLU [3]. However, in the aspect of pooling method, there is no known method with the best performance. Usually, researchers choose the method of max-pooling, mean-pooling, and stochastic-pooling [18] according to their personal habits [19]. In order to train a model with practical application value, in addition to the advanced models and algorithms above, high-quality data sets also play a decisive role. Common benchmark test data sets include MNIST [9], CIFAR [10] and most notably on the ImageNet classification challenge [11]. The famous object detection dataset includes PASCAL VOC [12] and MS COCO [13]. However, there are very few open data sets for specific problems or applications (such as SPAM datasets), which makes it very difficult to train a high-quality model. This paper first analyzes the problems of existing methods of data augmentation, and then proposes a method of data augmentation based on clustering analysis to generate suggested samples, which greatly increases the quantity and quality of training samples. We apply this algorithm to the SPAM dataset, and we train a new real time SPAM image recognition model with this data set. The experimental results show that the performance of our model reaches a new height compared with other methods.

✉ Fan Aiwan
faw_1978@163.com

Yang Zhaofeng
pdsncyiyang@163.com

[1] Administrative Office of Computer School of Pingdingshan University, Pingdingshan, Henan Province, China

Fig. 1 Top: cropping or warping to fit a fixed size. The cropping operation with fixed size and aspect ratio loses part of the image information, while the Warping operation changes the proportion of the image and the original appearance of the object

2 Related work

Common data augmentation methods are cropping [3, 6], warping [20, 21] and selective search(SS) [22]. SS is randomly generated from the original image of a certain number (for example, 2000) of suggested regions, where each suggested region can be understood as an additional sample. The principles of the cropping and warping methods are shown in Fig. 1.

Cropping gets more samples by cutting different parts of the original image, and warping just got more samples by modifying the aspect ratio of the original image. The result of Fig. 2 shows that the lack of basis aspect ratio and size have a great impact on the quality of the generated samples, so we *need* some prior knowledge to guide how to generate more high-quality samples from the original image.

Attar A surveys image spam tricks, anti image spam techniques, data set, etc. [23]. Zhang Y implemented a model of image spam filtering using PSO algorithm [24]. Kim S Y proposes a graph-based approach that utilizes graph structure in abundant e-mail spam dataset [25]. Liu Q trained a image spam classification model by using low-level features [26]. Shen J based on multiple visual properties extracted from different levels of granularity, aiming to capture more discriminative contents for effective spam image identification [27]. Other spam filtering models are [28–31]. However, these models are usually pieced together by a bunch of algorithms, which are not conducive to expansion and optimization, and often have many problems, such as low accuracy, poor real-time performance, and the inability to use massive samples to train. Although [28] also filters spam based on depth neural network, it is a hybrid model, which is different from us in many aspects, such as model framework, sample making process and so on. What we propose is a single model which adapts to large-scale parallel operation. By increasing the number of samples and learning with multi-resolution, the

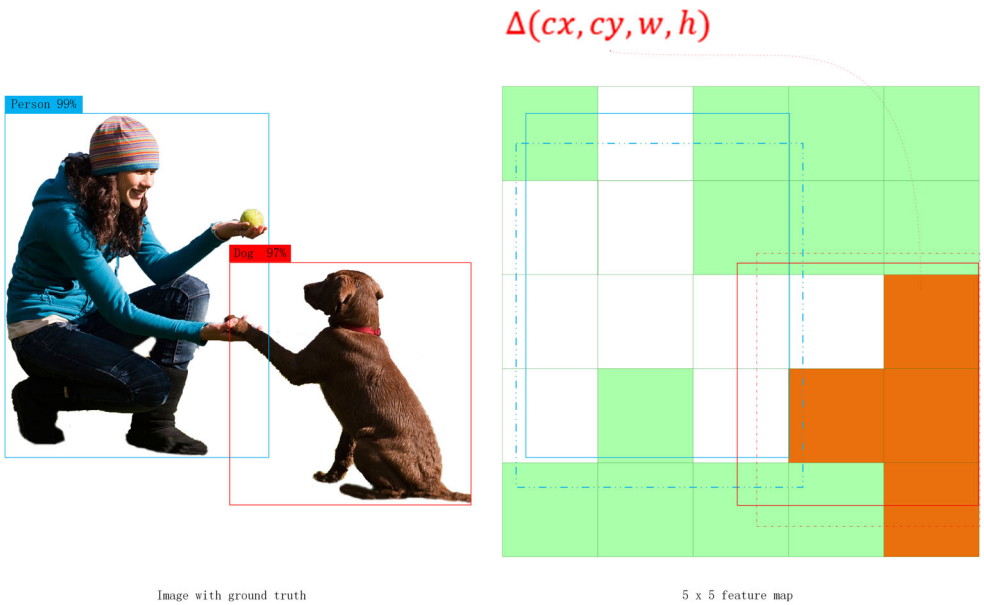

Fig. 2 Instead of choosing samples size and aspect ratio by hand, we run k-means clustering on the training set size and aspect ratio to automatically find good priors. Our clustering algorithm has two stages. First, we use the Euclidean distance to calculate the size and shape of each cluster (the object in the image) based on the pixel value, and get the size and aspect ratio samples. After obtaining the size and aspect ratio of each sample, we cluster the sample size and aspect ratio again, and finally get the shape and size of the suggested sample. The coordinates of each object in the output image are represented by cx, cy, w and h, respectively

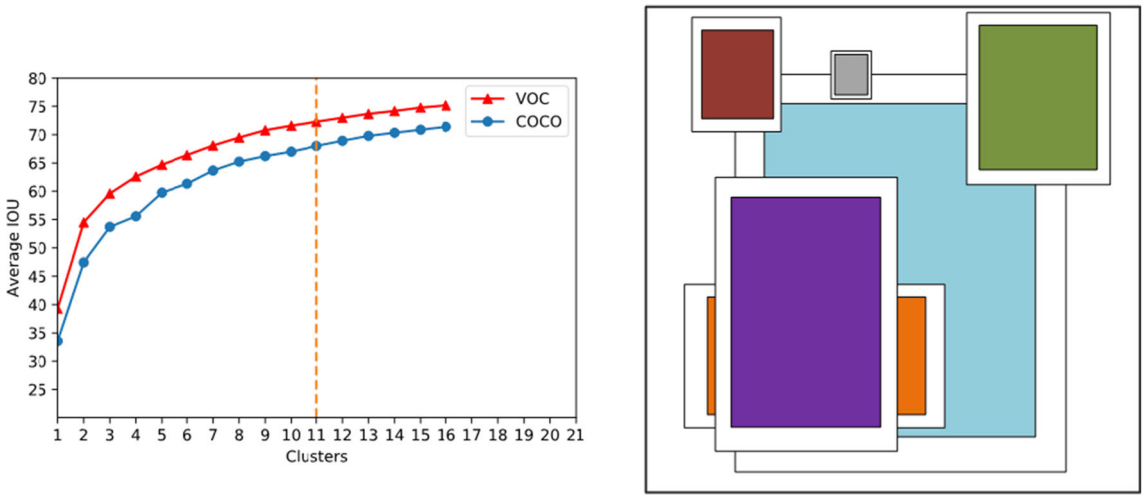

Fig. 3 Clustering box dimensions on VOC and COCO. We run k-means clustering on the dimensions of bounding boxes to get good priors for our model. The left image shows the average IOU we get with various choices for k. We find that $k = 6$ gives a good tradeoff for recall vs. complexity of the model. The right image shows the relative centroids for VOC and COCO. Both sets of priors favor thinner, taller boxes while COCO has greater variation in size than VOC

accuracy of the model is very high and the inference time is very fast.

3 AMME stochastic optimization algorithm

3.1 *k*-means for dimension clusters

When we use deep convolution neural network to train an image spam recognition model, the first problem we encounter is the serious shortage of data samples. At the same time, there are many problems in manual setting of Cropping and Warping to produce sample size and aspect ratio, which can not meet our needs. The model can learn to adjust the samples appropriately but if we pick better priors for the network to start with we can make it easier for the network to learn to predict good detections. Instead of choosing samples size and aspect ratio by hand, we run k-means clustering on the training set size and aspect ratio to automatically find good priors. Our clustering algorithm has two stages. In Fig. 2, first, we use the Euclidean distance to calculate the size and shape of each cluster (the object in the image) based on the pixel value, and get the size and aspect ratio samples. After obtaining the size and aspect ratio of each sample, we cluster the sample size and aspect ratio again, and finally get

Fig. 4 A sample of the spam images in our k-means corpus. Many images are difficult to judge as spam or ham. Corporate logos and comics are often included in ham messages while the picture of the woman is from a spam advertisement. We rely on a labeling of the email itself to simplify this decision

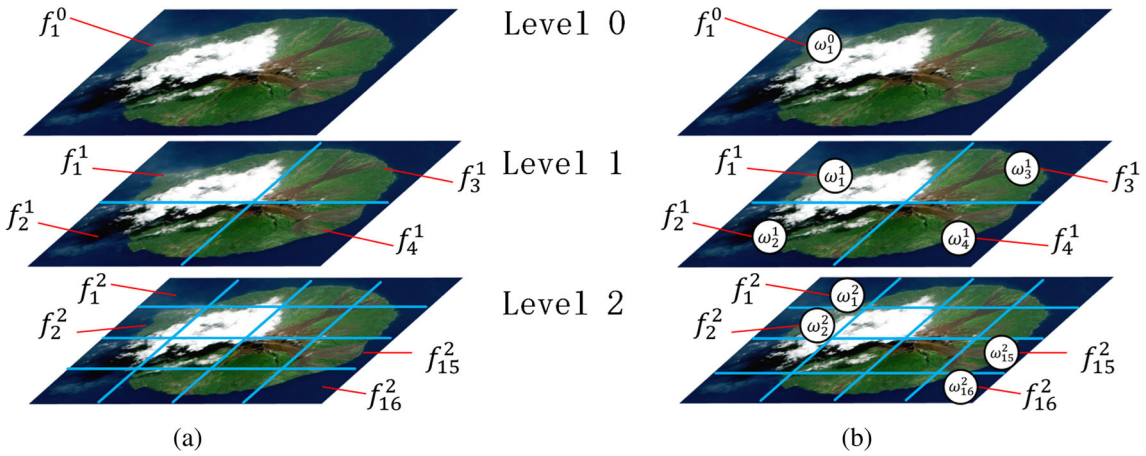

Fig. 5 **a** Original spatial pyramid representation. **b** Discriminative spatial pyramid

the shape and size of the suggested sample.After obtaining the size and aspect ratio of each sample, we cluster the sample size and aspect ratio again, and finally get the shape and size of the suggested sample. We use the rectangle instead of the irregular cluster, where the four nodes closest to the edge of the original image are used as the vertices of the four edges of the rectangle, which are represented by $T = \{t_1, t_2, t_3, t_4\}$, respectively. Then, the horizontal intersection of the four vertices forms four angles of the rectangle, in which the nearest angle to the upper-left corner of the original image is defined as the origin of the rectangle, which is represented by (c_x, c_y), and we use w and h to represent the length and width of the rectangle.

Different k values have a great influence on the quantity and quality of the proposed samples, so we clustering the shape (w, h) of the proposed samples again (see Fig. 3). We choose $k = 5$ as a good tradeoff between model complexity and high recall. The cluster centroids are significantly different than hand-picked cropping and warping. There are fewer short, wide bojects and more tall, thin objects.

Spam dataset constructed two data source, one based on personal spam and one based on publicly available spam. For personal spam, it collected spam emails from 10 email accounts across 10 domains and a catch all filter on two domains over the period of one month.

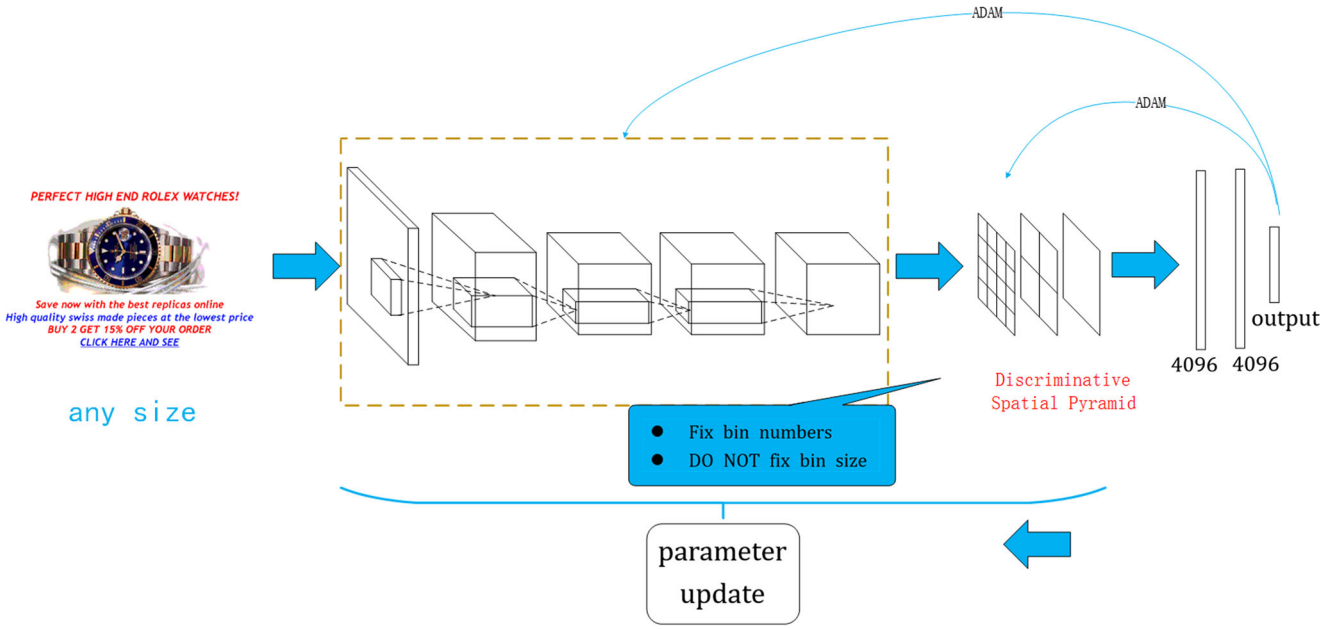

Fig. 6 The network structure of weighted discriminative spatial pyramid networks

Table 1 Network configuration parameters

Layer	1	2	3	4	5	6	7	8	9	10	Output
Stage	conv	conv	conv	conv	conv	WSP1	WSP2	WSP3	Full	Full	Full
Filter size	11*11	7*7	4*4	3*3	3*3	–	–	–	–	–	–
Conv stride	4*4	1*1	1*1	1*1	1*1	–	–	–	–	–	–
Padding type	Zero	Zero	Zero	Same	Same	Zero	Zero	Zero	–	–	–
Spatial input size	anySize	anySize	anySize	anySize	anySize	conv5	conv5	conv5	wsp3*3 wsp2*2 wsp1*1	1000	1*1

Every attached image (gif, jpg, png, and bmp) was extracted for the personal spam corpus, including emails that contained multiple images [32]. Figure 4, we use $k = 5$ to extend the spam data set, so we get 60, 000 training samples, we can see the samples produced by k-means, most of the objects are complete and high quality.

3.2 Real-time image spam recognition model

In order to meet the needs of real-time image recognition, the ADAM [16] optimization algorithm is applied to currently the most popular image recognition modelconvolution neural network. ADAM has the characteristics of fast convergence and rarely falling into local optimum. However, there are many variants of convolution neural network; SPP-net [5] is one of the best and is widely used because it can identify images of various scales and maintain high accuracy. Our proposed model is different from SPP-net in two aspects: First, SPP-net first extracts the global features of the original image. As shown in Fig. 5a, each pyramid horizontally divides the image into a grid sequence, and then extracts the features from each grid using max-pooling and concatenates them into a large feature vector. However, due to the difference in the amount of information reflected in each local region of the image, the discriminative spatial pyramid (DSP) [33] method we use can more fully express the features of the image, because it assigns a weight to each grid at each layer, and then connects the features of all grids together in tandem. As shown in Fig. 1b, we extract the feature from each grid cell at each pyramid layer l and use $w_k^l \in \mathbb{R}$ to denote the importance of grid k at layer l. Formula (1a), and (1b) define a weight vector and a feature matrix respectively. The weighted DSP feature $f_w \in \mathbb{R}^d$ can be written by the formula (1c) where $c(l)$ represents the number of grids of the lth layer pyramid.

$$w = (w_1^0, w_1^1, w_{c(1)}^1, ..., w_1^{L-1}, ..., w_{c(L-1)}^{L-1})^T \tag{1a}$$

$$F = (f_1^0, f_1^1, f_{c(1)}^1, ..., f_1^{L-1}, ..., f_{c(L-1)}^{L-1}) \tag{1b}$$

$$f_w = Fw \tag{1c}$$

Because each neuron of convNet outputs the feature matrix separately, we need to set a weight set that corresponds to the output of each neuron $w = \{w_i \in \mathbb{R}^{dw}\}_{i=1}^{N_w}$ where N_w represents the outputs of all the neurons and the (1c) is redefined as (2).

$$f_w^{N_w} = ((F_{w1})^T, ..., (F_{w_{N_w}})^T)^T \tag{2}$$

The second difference with SPP-net is as follows: feature extraction in a convolution neural network is carried out by the convolution kernel at the convolution layer that performs convolution of the different areas of the image. Down sampling, namely pooling, aims to downsize the feature

Fig. 7 Data set sample. **a** CIFAR data set. **b** MNIST data set

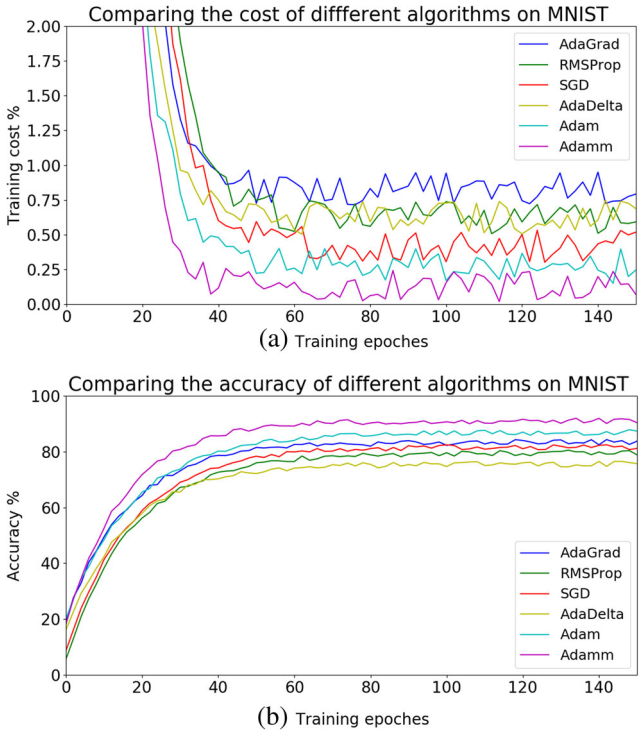

Fig. 8 Use the same network framework convNet5 to evaluate the performance of different optimizers on the MNIST data set. **a** Training cost. **b** Accuracy

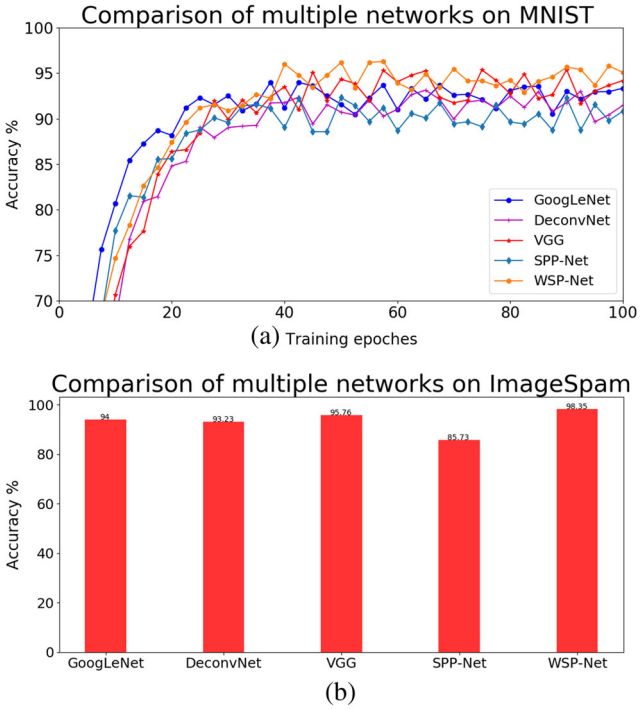

Fig. 9 Use MNIST and ImageSpam data sets to compare the performance of currently popular image recognition models. **a** Accuracy of each model on the MNIST data set. **b** Test result of each model on the ImageSpam data set

map. The pooling size is normally 2×2. Common pooling operations are mean-pooling, max-pooling and stochastic-pooling [18, 19]. In general, mean-pooling can reduce the variance of the estimated value and retain more background information of the image. Max-pooling can reduce the offset of the estimated mean value and retain more texture information of the image. Unlike max-pooling used by most models, weighted-pooling [34] is used in our model. The main features of the weighted-pooling method are as follows: (1) The information quantity of the pooling region is quantified by information theory for the first time. (2) Also, each activation contribution was quantified for the first time and these contributions eliminate the uncertainty of the pooling region which it is located in. (3) For choosing a representative in this pooling region, the weight of each activation obviously superiors to the value of activation. We name this real-time image spam recognition model as weighted discriminative spatial pyramid net (WDSP-net), with its complete network architecture shown in Fig. 6, and the specific parameter values of the model are shown in Table 1.

In Fig. 6, in our real-time image recognition model, the size of the input RGB image is usually random, the first 5 layers of the network are standard convNet5 [3], what's different is that we perform weighting pooling operation for the 1st, 2nd, and 4th convolutional layers and the WSP layer, thus minimizing the over-fitting phenomenon of the model. ConvNet is connected at its end to our WSP layer which has two advantages: (1) it weights the different regions of the image, so that the image can be more accurately expressed; (2) it outputs fixed-size features to the full connection layer, so that our model can recognize images of any size in real-time. The WSP layer is followed by two full connection layers, and the final output of the network is the categories of the recognized image. In order for our real-time image spam Filtering model to always maintain the best performance, we also add a module to dynamically update the network parameters. The new parameters come from an offline-trained network model, i.e., when the model performance reaches a new height, the generated parameters are passed automatically to our real-time recognition model. The whole WDSP-net could also easily be implemented by regular convolutional neural network software packages such as tensorflow, Caffe and Theano [35–37].

4 Performance

To better illustrate the performance of the WDSP model, we use classical image classification tasks to compare different pooling methods. The operating system in all

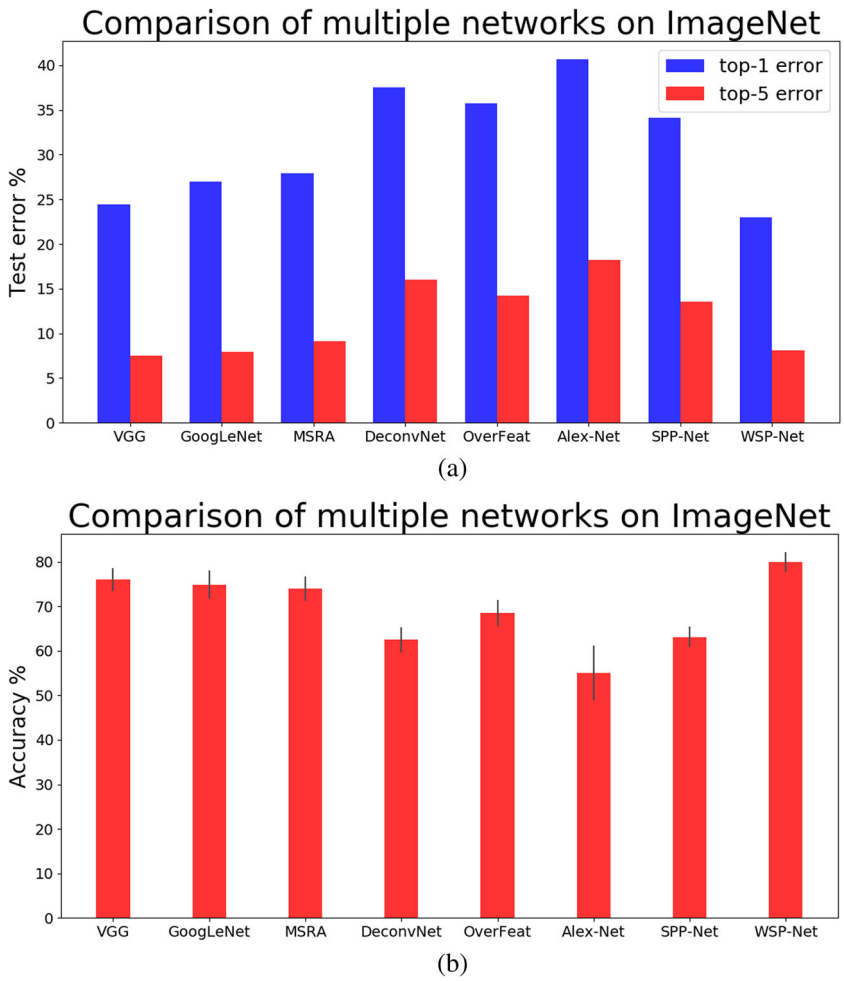

Fig. 10 Use the famous ImageNet data set to evaluate the overall performance of each model. **a** Top1 and top5 training errors for each model. **b** Results of the mAP (mean average Precision) indicator for each model

experiments is Ubuntu16.04, it uses a NIVDIA GTX1080 GPU to accelerate the compute, and the tensorflow1.2 as the deep learning framework. The network optimizer we use is Adam, the loss function is Euclidean distance, and the batch size is fixed at 150. The network contains three convolutional layers, each with 32, 64, 64 convolution kernels. Each kernel size is 5×5; 3×3; 3×3. According to the experimental demands, each convolution layer followed by a pooling layer and its pooling method may be the max-pooling, stochastic-pooling or weighted-pooling, etc. Unless otherwise specified, the pool size is 3×3 and the step size is 2. In addition, there is a normalization layer behind each pooling layer (e.g., [3]). The experimental results show that this normalization layer can not only reduce the network training time, but also deal with the output maximum when ReLU is used as the activation function. Finally, two fully connected layer outputs the network classification results through the soft-max classifier. We apply this model to two different data sets: MNIST, CIFAR-10, and enlarged image spam dataset(ImageSpam). Figure 7 is an example image of MNIST and CIFAR-10 data sets. During training, each dataset is trained for 120 epochs.

We use the same network structure convNet5 and the same data set MNIST to compare the performance of different optimization algorithms. Figure 8a shows clearly that, because we increase estimating points and improve the standard first and second raw moments, our algorithm converges faster and the gradients decrease more smoothly. As shown by Fig. 8b, because ADAM uses a more accurate moment estimation method to adjust the learning (gradient descent) rate of each parameter, it not only reduces the probability of the real-time image recognition model falling into a local optimum, but also brings the recognition accuracy to a new height.

We use MNIST and ImageSpam data sets to compare the performance of currently most popular image recognition models. The results of Fig. 9a, b indicate that WSP-net have got very good results, especially in Fig. 9b where the accuracy on ImageSpam data set reaches 98.35

As shown in Fig. 10, although the top5 error rate of WDSP-net is slightly higher than those of VGG and

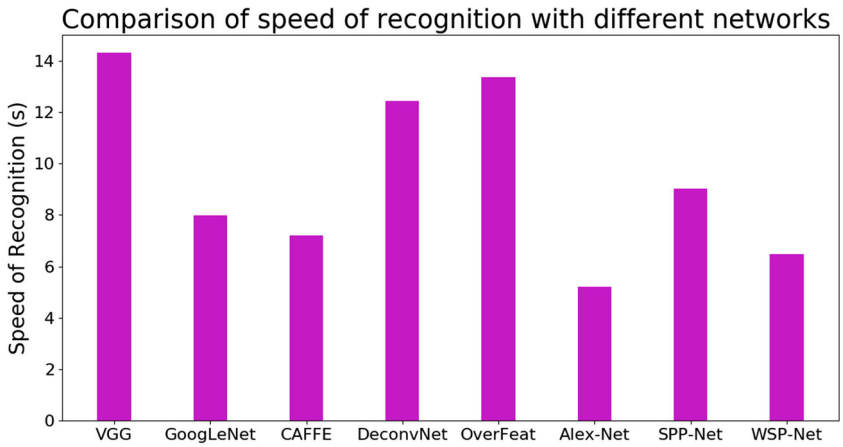

Fig. 11 Compare the real-time image recognition speeds of different models, with a total number of 10,000 images spam filtering

GoogLeNet, the result of top1 error rate is much lower than those of the previous two. The lower top1 error rate also indirectly demonstrates that the WDSP-net stability and accuracy are higher. Since Internet mail systems will send a lot of emails every second, the real-time requirements on the image recognition model are higher. Therefore, we choose 10,000 images (one image for each email) to test the recognition speed of each model. As shown in Fig. 11, Alex-Net, because of its simple structure, has the fastest image recognition speed. However, because WDSP-net uses the gradient algorithm ADAM with better performance, it maintains high accuracy and achieves high recognition efficiency at the same time.

5 Discussion and conclusions

Considering the high requirements for real-time and accuracy of the real-time image spam system, this paper first uses the multi-estimating point method to improve the standard moment estimation method and proposes a new stochastic gradient optimization algorithm called ADAM. Then, after analyzing currently popular image recognition models, it proposes WDSP-net combined with our ADAM and SPR algorithms. The experimental results show that WDSP-net achieves a new state of the art in terms of accuracy and real-time, thus taking a solid step in the field of garbage image recognition.

Acknowledgments The authors would like to thank the reviewers for their helpful advices. The National Youth Science Foundation project of China (Grant no. F020101), the Henan Province Science and Technology key Project (Grant no. 1521022101936), the Natural Science Foundation of Hunan Province, China(Grant No.2018JJ2023) and the Key projects of Science and Technology Research in Henan Education Department (grant nos. 15A520091, 17B520031) are gratefully acknowledged.

References

1. Russakovsky O, Deng J, Su H, Fei-Fei L et al (2015) imagenet large scale visual recognition challenge[J]. Int J Comput Vis 115(3):211–252
2. Le Cun Y et al (1990) Handwritten digit recognition with a back-propagation network. In: Advances in neural information processing systems
3. Krizhevsky A, Sutskever I, Hinton GE (2012) ImageNet classification with deep convolutional neural networks[C]. In: International conference on neural information processing systems, curran associates Inc., pp 1097–1105
4. Sermanet P, Eigen D, Zhang X et al (2013) OverFeat: integrated recognition, localization and detection using convolutional networks[J]. Eprint Arxiv
5. He K, Zhang X, Ren S et al (2014) Spatial pyramid pooling in deep convolutional networks for visual recognition[J]. IEEE Trans Pattern Anal Mach Intell 37(9):1904–1916
6. Zeiler MD, Fergus R (2014) Visualizing and understanding convolutional networks[C]. In: European conference on computer vision. Springer, Cham, pp 818–833
7. Szegedy C, Liu W, Jia Y et al (2014) Going deeper with convolutions[J]. pp 1–9
8. He K, Zhang X, Ren S et al (2015) Deep residual learning for image recognition[J]. pp 770–778
9. Le Cun Y et al (1990) Handwritten digit recognition with a back-propagation network. In: Advances in neural information processing systems
10. Krizhevsky A (2009) Learning multiple layers of features from tiny images. Technical Report TR-2009 University of Toronto
11. Russakovsky O, Deng J, Su H et al (2014) ImageNet large scale visual recognition challenge[J]. Int J Comput Vis 115(3):211–252
12. Everingham M, Gool LV, Williams CKI et al (2010) The pascal visual object classes (VOC) challenge[J]. Int J Comput Vision 88(2):303–338
13. Lin T-Y, Maire M, Belongie SJ et al (2014) Microsoft COCO: common objects in context. CoRR. arXiv:1405.0312
14. Duchi J, Hazan E, Singer Y (2011) Adaptive subgradient methods for online learning and stochastic optimization. J Mach Learn Res 12:2121C2159
15. Zeiler MD (2012) ADADELTA: an adaptive learning rate method[J]. Computer Science
16. Kingma D, Adam BJ (2014) A method for stochastic optimization[J]. Computer Science

17. Zhu X, Meng Q, Gu L (2017) J Real-time image proc. https://doi.org/10.1007/s11554-017-0743-y
18. Zeiler MD, Fergus R (2013) Stochastic pooling for regularization of deep convolutional neural networks[J]. Eprint Arxiv
19. Boureau YL, Ponce J, Lecun Y (2010) A theoretical analysis of feature pooling in visual recognition. In: International conference on machine learning, DBLP, pp 111–118
20. Donahue J, Jia Y, Vinyals O et al (2014) DeCAF: a deep convolutional activation feature for generic visual recognition[C]. In: International conference on international conference on machine learning. JMLR.org, pp I–647
21. Girshick R, Donahue J, Darrell T, Malik J (2014) Rich feature hierarchies for accurate object detection and semantic segmentation. In: CVPR
22. Uijlings J, van de Sande K, Gevers T, Smeulders A (2013) Selective search for object recognition. IJCV
23. Attar A, Rad RM, Atani RE (2013) A survey of image spamming and filtering techniques[J]. Artif Intell Rev 40(1):71–105
24. Zhang Y, Wang S, Phillips P et al (2014) Binary PSO with mutation operator for feature selection using decision tree applied to spam detection[J]. Knowl-Based Syst 64(1):22–31
25. Kim SY, Sohn KA (2015) Graph-based spam image detection for mobile phone spam image filtering[J]. Laryngoscope 3(4):72–86
26. Liu Q, Zhang FL, Qin ZG et al (2010) Feature selection for image spam classification[J]. IEEE, pp 294–297
27. Shen J, Deng RH, Cheng Z et al (2015) On robust image spam filtering via comprehensive visual modeling[J]. Pattern Recogn 48(10):3227–3238
28. Soranamageswari M, Meena DC (2010) Histogram based image spam detection using back propagation neural networks[C]. In: 2010 international conference on control automation and systems (ICCAS), IEEE, pp 3985–3988
29. Amir A, Srinivasan B, Khan AI (2017) Distributed classification for image spam detection. Multimed Tools Appl 77:13249–13278
30. Adarshya SP, Mekala R, Arayakkandiyil R et al (2012) Image spam detection through server-client filtering by tracing the source IP of the spammer[J]. Digital Image Processing
31. Yang G, Zhang Y, Yang J, Ji G, Dong Z, Wang S, Feng C, Wang Q (2016) Automated classification of brain images using wavelet-energy and biogeography-based optimization. Multimedia Tools and Applications 75(33):15601C15617
32. Dredze M, Gevaryahu R, Elias-Bachrach A (2007) Learning fast classifiers for image spam. In: Proceedings of the conference on email and anti-spam (CEAS)
33. Harada T, Ushiku Y, Yamashita Y et al (2011) Discriminative spatial pyramid[C]. In: Computer vision and pattern recognition, IEEE, pp 1617–1624
34. Zhu X, Meng Q, Ding B et al (2018) Cluster Comput. https://doi.org/10.1007/s10586-018-2165-4
35. Abadi M, Agarwal A, Barham P et al (2016) TensorFlow: large-scale machine learning on heterogeneous distributed systems[J]
36. Jia YS et al (2014) Caffe: convolutional architecture for fast feature embedding[J]. Eprint Arxiv, pp 675–678
37. Team TD, Alrfou R, Alain G et al (2016) Theano: a python framework for fast computation of mathematical expressions[J]