AMSE JOURNALS-AMSE IIETA publication-2017-Series: Advances B; Vol. 60; N°1; pp 123-139
Submitted Jan. 2017; Revised March 15, 2017, Accepted April 15, 2017

MHI-CNN Model for Fine-grained Classification of Product Image

*, **Mingxia Lin, *Cuihua Li

*School of Information Science and Technology, Xiamen University, Xiamen, P.R.China
(Corresponding author)
**School of Business and management, JiMei University, Xiamen, P.R.China

Abstract

Fine-grained classification (FGC) is a current focus of research, but the problem in the application of product image especially in women's clothing is a great challenge. Because women clothing contain more style, color and details than other clothes, in addition constructing a proper training set for these is very difficult. We propose a new model named MHI-CNN which is multi-scale and heterogeneous integration based on the traditional Convolutional Neural Network (CNN) model. We use multi-scale to get more detail feature and heterogeneous integrate three models to achieve higher accuracy. The experiment results show our model improved the performance of women clothing image classification.

Keywords

Fine-grained Classification (FGC), Convolutional Neural Network CNN), Deep Learning

1. Introduction

With the development of e-commerce, artificial intelligent technology is extensively applied to e-commerce, such as virtual dressing room, intelligent recommendation system. Image classification plays an important role in these application, because image can pass more information than the text in e-commerce platform. With the increasing of mobiles with camera, it has become much better way to meet user requirements by uploading pictures to find similar products you want.

The image of the e-commerce platform including the seller product images and user reviews pictures, it showed the growth of geometric progression. So the need for effective management of

very large-scale multimedia data and the provision of fast and accurate retrieval services is urgent. Until now there is no more mature solution. EBay provide shopping by category or inputting key words such as “women’s jackets” to find the product. Although Taobao (www.taobao.com) has provided the content based similar clothing search by user interaction, the search range is only in the seller's main product image and the result is not satisfactory.

In addition to the buyer user to seek similar product requirements, in the e-commerce platform to solve the problem of seller user image plagiarism and the image does not matching the text also need effective image fine-grained classification approach. Image FGC is a challenge task because of the following difficulties: 1) The similarity between different images classes may be high. For example Shirts and T-shirts in the e-commerce usual belongs to different categories, but they are almost the same in contour and color. The example is shown in Fig1. 2) There may be a big difference between the same category, such as two shirts may be different in colors, textures and the collar shape. Fig2 shows an example. Women clothing classification because of style, color, material more varied and more difficult than other product image classification. 3) Products image is in a variety of types in e-commerce platform, there are product photo from front, back and side, real model presentation photo and customer upload photo etc.

Fig. 1 a shirt on the left and a T-shirt on the right,they are almost the same at first glance

Fig. 2 there are two shirts respectively on the left and right, they are very different at first glance

Therefore, it is difficult to perform such FGC tasks by using the traditional image classification method with visual features such as color, texture and edge. It is difficult to find the corresponding model with these features. The local feature operator such as Scale-invariant feature transform (SIFT) is a form that can be considered, but it needs to the high-definition image in order to find the key points. Some of valid features of artificial recognition are difficult to obtain in these images, such as materials.

Our contribution can be summarized as follows: 1) We study the application of CNN structure in FGC problem, and propose multi-scale heterogeneous CNN model which lays a foundation for further study of CNN model. 2) We design CNN model with relatively less layers and achieve good accuracy, which can save the computational cost of the depth model. 3) As the first step of image retrieval and intelligent recommendation, women clothing image classification has an important effect on the future development of e-commerce system.

This paper explores the advantages of CNN in the FGC. Inspired by spatial pyramid matching [23],we use the multi-scale image as the input of CNN model and through the multiple CNN model fusion to fit the complex and changeable characteristic. The paper is organized as follows. We will briefly introduce the prior related researches in Section 2, describe the framework of our MHI-CNN model and introduce the training process in Section 3, experimental evaluation in Section 4, and get conclusions in the end.

2. Related Work

2.1Fine-grained classification

Compared with CGC (coarse-grained classification), FGC (fine-grained classification) focus on the finer levels of image classification. For example, completing transport image classification task CGC only need to recognize the picture of car, while FGC need to find the specific mode and brand of the car. Gosselin [1] summarizes two characteristics of FGC: 1) The overall description of the characteristics of the image which is effective in CGC become similar in FGC, so local feature is important to FGC. 2) Background can provide effective classification information in CGC but usually be seem as noise in FGC. For its challenge and practical application FGC has attracted many researchers attention. They construct a lot of dataset used in the experiment such as, Oxford flowers [2], Caltech bird species [3], about cooking action of cooking activities [4], cats and dogs [5]. Oxford flowers is a dataset of flower contains 102 different categories flowers. Many of them focus on inter class similarity [2, 6] which is a challenging task of FGC.

Aim at FGC problems, two method are proposed. One is find more discriminative, robust local features. Bag-of-words (BOW) model and Fisher Vector (FV) have been studied widely as local features. For example, Fernando [7] applying frequent item set mining to the task and build powerful bag-of-FLH-based image representations, Chen [8] use side information to build a hierarchical matching framework based on bag-of-words representation. NCHEZ [9] use the Fisher Kernel framework instead of most common strategy. The other solution is to make use of combination of base kernels corresponding to base features (such as color, texture and shape etc.). These work pay attention to optimal tradeoff [10], feature selection [11] and integration features [12]. The above algorithm can achieve good performance in special task but with the task changing result is not very satisfied. The fine-grained recognition challenge FGComp 2013 provides five dataset including aircraft, birds, cars, dogs and shoes, the accuracy of most of algorithms fluctuate from 50%-90% in different dataset. Part of the reason is some of the categories are easy to distinguish, there is also a very important reason is that hand-crafted descriptors can't self-adapt to the changing of the task. In recent years, Convolution Neural Net (CNN) show its great performance in image classification. Researchers have used the CNN structure to solve the FGC problem, Razavian [6] compared HSV, SIFT, HOG, BOW with CNN on the performance of Oxford flowers, CNN can improve the performance of about 16%. But Gosselin[1] proposed method based on Fisher vector is better than CNN .There is lack of especially structure of CNN to solve the FGC problem. Iscen [13] shows that although CNN has achieved good performance in image classification, it is not better than local features for the FGC task.

2.2 Product image classification and retrieval

The clothing image classification is the important component in e-commerce search engine and intelligent recommendation system and receives wide attentions. However, mature commercial applications mostly based on text retrieval. Inefficiency and low accuracy are the main obstacles to the research of content-based image retrieval. Traditional computer vision and machine learning techniques is difficult to overcome these problems. For example Borras [21] only achieved a 64% over all accuracy. Zahavy [26] propose a decision level fusion approach but they need text and image inputs,however many pictures without label in e-commerce.

As mentioned above, CNN [18, 21] structure has been widely used in clothing classification and retrieval. However, they didn't think these problems as a FGC problems. Chen [18] devoted to style classifications. Borras [21] devoted to clothing retrieval in a large corpus. Khosla N [20]

explored the FGC problem in shoes, while the task of shoes is easier than clothing. Shichao Zhao [27] note that the features extracted from the deep convolutional layers of CNNs perform better than that from the fully-connected layers and design Top-Weight method to detect the interesting area of e-commerce images automatically.

3. Framework of MHI-CNN Model

3.1 Motivation

How many layers should we use in the CNN structure? Since CNN achieved good performance in image classification, it will be a priority to these problem. However, the traditional CNN structure tends to extract the global feature without paying attention to the details. Although CNN still lack of theoretic support, researchers have done many experiments and try to explain why it can do, which gives a direction to further enhance the performance of CNN. Zeiler [14] studied the CNN visualization, illustrating the features of the 1-2 layer extraction similar to edge and corner features. At the upper level, the features appear invariant, and the feature invariance is very obvious at about 5 layers. This proves that the 5 convolutional layers model mentioned by Krizhevsky [15] is a good model. The convolution step size and the convolution kernel size of CNN also affect the features of the lower layer. Smaller step and convolution kernel are more favorable for obtaining discriminative features and less "dead" features.

The heterogeneous CNNs have different feature extraction effects. Heterogeneous CNNs (convolution kernel size, step size, maps not the same) get different image classification effect, and can be regarded as different feature extraction method. Chatfield [16] studied the effect of different structures of CNN on the learning effect and demonstrated the data augmentation techniques can significantly improve performance of shallow representations. But the article does not further analyze why the heterogeneous CNN extract the characteristics of the differences.

The combination of multiple CNNs is helpful for classification feature extraction. Ciresan [17] proposed a multi-column deep neural network to improve the object recognition performance, so that the system is not sensitive to contrast and light.

In this paper, two different preprocessing methods are used. We first divide the picture into several blocks as input of three CNN model and then average their predictions.

3.2 MHI-CNN Model

In summary, the CNN with the few layers can capture more the details of the extracted feature, but the feature invariance is relatively poor. In the same databset, the CNN structure of the heterogeneous (different stride, convolution kernel, etc.) achieve different performance. It shows that features of their extraction are certain different. So we propose a combination of multiple CNNs with multi-scale input as shown in Fig 3, which we called Multi-scale and Heterogeneous Integration CNN model (MHI-CNN).

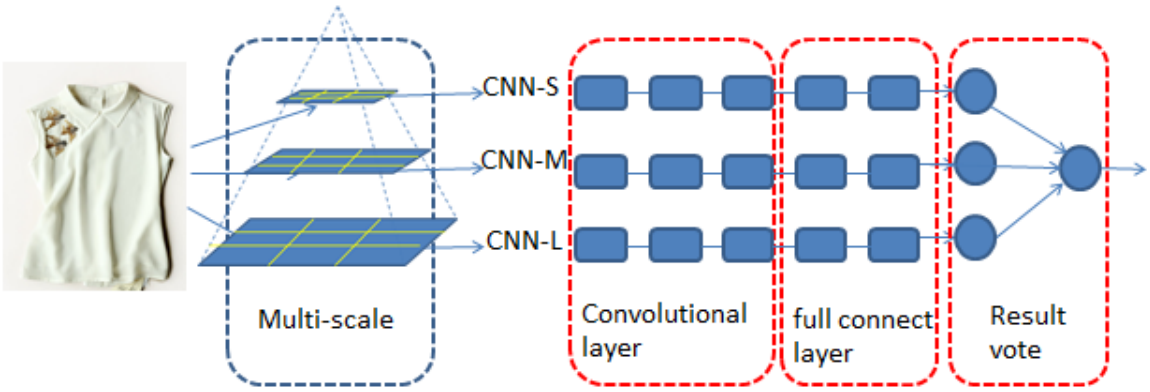

Fig. 3 the framework of MHI-CNN

In this model, the image is firstly changed in 48 × 48, 96 × 96,148 × 148 different scales as input. Then three CNN models of the 5 layers are trained respectively. For each CNN input, the number of kernel maps and stride are different. The specific parameters are shown in Table1, Table 2 and Table3. When three different CNNs give a decision to the same sample, the result of the judgment is combined with the label and the weights are calculated according to the voting result. After calculating all weights, we get a joint CNN model.

Table 1. The detail design of CNN-S

Layer	Conv1	Conv 2	Conv 3	4	5
Channel	50	100	250	300	10
Filter size	5×5	5×5	4×4	--	--
Pooling size	2×2	2×2	2×2	--	--
Input size	48×48	22×22	9×9	3×3	1×1

Table 2. The detail design of CNN-M

Layer	Conv1	Conv 2	Conv 3	4	5
Channel	96	128	250	300	10
Filter size	7×7	5×5	4×4	--	--
Pooling size	3×3	2×2	2×2	--	--

Input size	96×96	30×30	13×13	5×5	1×1

Table 3. The detail design of CNN-L

Layer	Conv1	Conv 2	Conv 3	4	5
Channel	100	150	300	300	10
Filter size	13×13	7×7	4×4	--	--
Pooling size	4×4	4×4	2×2	--	--
Input size	148×148	34×34	7×7	3×3	1×1

CNN structure design has three key points: the size of the convolution kernel, the number of convolutional kernels and the stride. At present, there is no definite mathematical basis for the structural design of CNN, but according to the results of experiment and literature research, the following criteria are generally satisfied:

1) The larger the dataset, the more convolution layers needed. The 5, 8, 13 layers is verified to be a relatively efficient feature extraction structure. The women clothing image dataset which we collect are not very large, and consider with the motive of section 3.1, we design the CNN structure of three convolution layer with three sub-sampling layers and two fully-connected layers.

2) The feature map of the convolution layer is larger along with depth, which accords with the mechanism of human brain cognition. The more neuron is mapped, the more neuron trigger the mapping to the upper layer.

3) The sub-sampling and stride size is generally from large to small design, the more deep layer sub-sampling matrix smaller.

Based on the above criteria, 3 CNN with different structures are designed, which are named as CNN-S, CNN-M and CNN-L respectively according to the input scale and the sample size of the sub sampling.

We use Relu (Rectified Liner Units) [15] as the activation function to obtain higher training speed and use dropout [15] in the fully-connected layers to avoid overfitting. Because it is not a large dataset, this paper doesn't use overlapping sub-sampling, sampling layer matrix size and stride size are the same. The convolution stride is fixed to 1 pixel, so it is not listed in Table1-3. The other parameters can be seen in Table1-3.The "input size" means that current input level, which is the output of the previous layer, in the tables convolution layer and its sub-sampling layer merge into a layer. In order to break the symmetry, the layer following a convolution layer is usually not fully connected to the front, but part of the connection. However, considering that filter window is not large, we use a fully connected way.

3.3 Details of learning

The convolution layer is calculated as follows:

$$((y-s)+1)\times((y-s)+1)$$

(1)

y represents the current feature map size, s represents the convolution kernel size, the convolution stride is 1 pixel. For example, if the input to CNN-S is 48 × 48 RGB image patch, the filter with receptive field is 7×7, then the next convolution layer size is (48-7 + 1) × (48-7 + 1) = 42 × 42. With non-overlapping sub-sampling, the size after sub-sampling is equal to: (y / s) x (y / s); y represents the input of the current feature map, and s represents the sample size. In the above example, input is 42 × 42 then the output is (42/2) x (42/2) =21 x 21.

For the data preprocessing, we mainly do the following two aspects: 1) background segmentation to get the initial input samples. 2) Data augmentation by altering the intensities of the RGB channels. The foreground segmentation method is often used in the FGC, as mentioned before, the background often means noise for FGC. In order to highlight the main commodity, goods are usually located in the center of the picture with less complex background. It provides convenience for foreground segmentation. We use Grabcut [24] as foreground segmentation method in our work. Grabcut segmentation requires user interaction and its processing speed is slower, but the realization is simple and it is effective. We used it to deal with the images of human body, plastic model and hangers, etc. It is also a way of data augmentation. For our task, the local feature is more important than the global feature. Therefore, the block cut in this paper is different from the block cut in the general CNN model, and the image is cut into three scales. The size of the cut image is not fixed, but the images are divided into different scales according to the 3 × 3 patch. The purpose is to make the CNN model pay more attention to the local features. Data augmentation can make the parameters in the model be fully trained, so it can improve the accuracy obviously. For the FGC of clothing goods, color characteristics are not very important. Because often a style of clothes will design with multiple colors. We use altering the intensities of the RGB channels as another form of data augmentation.

As shown in Fig 4 RGB image $I_{x,y} = (I^{R}_{xy,} I^{G}_{xy,} I^{B}_{xy,})$, The vectors of the three channels are respectively changed, $I^{1}_{xy} = (\alpha I^{R}_{xy,} I^{G}_{xy,} I^{B}_{xy,}), I^{2}_{xy} = (I^{R}_{xy,} \beta I^{G}_{xy,} I^{B}_{xy,}), I^{3}_{xy} = (I^{R}_{xy,} I^{G}_{xy,} \gamma I^{B}_{xy,}), \alpha, \beta, \gamma$ random

values (0, 1).As shown in Fig 4, the first column as a data is the original image, I_{xy}^{R} is the changing R channel shown in the second column , I_{xy}^{G} is the changing G channel shown in the third column, I_{xy}^{B} is the changing B channel shown in the fourth column

Fig 4 the RGB channel of image

3.4 Model training

Our model training consists of two parts: separate model training and integrated model training. Separate model training process can be described as follows:

The first stage is the forward propagation. This stage mainly through the input and the initial value of the weights to calculate the network output, then calculate the output of the convolution and the sampling layer, finally do a dot products between the output and the weight matrix of the full-connected layer. The second stage is the backward propagation stage, in which the main objective is to adjust the weight by means of the minimization error.

The convolution process is as formula (2), the N-th output layer c_j^n can be expressed as:

$$c_j^n = F\left(\sum_{i \in M_j} c_i^{n-1} k_{ij}^n + b_j^n\right) \quad (2)$$

Where k_{ij}^n represents the convolution kernel which connected the i-th feature map in the $(n-1)$-th layer and the j-th feature map in n-th layer;M_j is the choice of the input feature map, for the feature map of the n-th layer is not fully-connected to $(n-1)$-th layer, but partly

connect. b_j^n represents the bias of j-th feature map in n-th layer. F is the activation function of neurons, we use Relu.

The sub-sampling calculation process is as Formula(3):

$$s_j^n = F(\beta_j^n down(s_j^{n-1}) + b_j^n) \quad (3)$$

Where s_j^n represents the output of the n-th layer, plainly there are n input feature maps in the up layer, the sub-sampling layer has n output feature maps.b_j^n represents the bias of j-th feature map in n-th layer.β_j^n represents weight, the sub-sampling is nearly the linear operation if the value of β_j^n is enough small.F is the activation function of neurons, $down(.)$ is the sampling function which we use unoverlapping max-pooling.

The loss functions of the CNN structure are defined as follows:

$$loss = -\frac{1}{2}\sum_{i=1}^{m}\sum_{k=1}^{c}(d_k^n - y_k^n)^2 + \lambda\sum_{k=1}^{l} sum(\| w \|^2) \quad (4)$$

The function consists of two parts, the first is the error term and the second is the regularization term. In this paper, we used BP algorithm training CNN structure. The weights update as shown in Formula (5):

$$\Delta w^l = -\eta\frac{\partial E}{\partial w^l} \quad (5)$$

Where Δw^l represents the error update value of the i-th layer, η represents learning rate, $\frac{\partial E}{\partial w^l} = x^{l-1}(\delta^l)$, x^{l-1} represents the output of the $(l$-1)-th layer(means the input of l-th layer), δ^l represents the sensitivity of the l-th layer, the sensitivity can be calculated as Formula(6):

$$\delta^l = (W^{l+1})^T\delta^{l+1} \odot f'(u^l) \quad (6)$$

Where δ^{l+1} represents the sensitivity of the $(l+1)$-th layer, $f'(u^l)$ represents the partial derivative of the activation function, specially, to the output layer $\delta^l = f'(u^l) \odot (y^n - t^n)$.The output layer uses softmax regression. It can be seen that CNN weights are updated using the BP

algorithm, but since the layers of the CNN are not fully connected, the neuron sensitivity values corresponding to the connected neurons are accumulated in calculating the sensitivity.

After separate model training, integrated model training include 3 steps:

① train CNN-S, CNN-M and CNN-L with our women clothing dataset respectively；//Each CNN Model training process as above

②(x_i, y_i) is the i-th training example, do 48 × 48,96 × 96,148 × 148 three scales of change to x_i ,then use the result as input for CNN-S,CNN-M,CNN-L model.

③$(x_i, y_{ij})$is the result of the i-th training example in the j-th model,

$C_j = 0$;// where $C_j$is the voting result box of CNN,j=0,1,2

$$\text{for } j = 0 \text{ to } 2$$

if $y_{ij} = y_i$ then $C_j = C_j + 1$;

$w_j = \frac{C_j}{\sum_{j=0}^{2} C_j}$; //$w_j$ is the weights for each CNN structure are fused

The training process seems more complicated than traditional method, but because the parameters adjustment after the separate training is not much, it not increased time cost.

4. Experiments

4.1 Dataset

This experiment is based on Matlab2014a, the operating environment for Windows7, CPU Intel-i52410m, 4G memory. The final parameter design of CNN is described in Section 3.2. Three CNNs are trained with stochastic gradient descent, in order to speed up the convergence, minbatch take a smaller value of 2, the weight decay is 0.0005. The iteration time is 70-80s and the maximum iteration is 100.

In this work we use the dataset which consists of 2000 training and 500 test women clothing images obtained from e-commerce platform Taobao and Tmall. The images are divided into 10 categories. Each category of pictures ranging from 160 to 250. The average size of each image is approximately 200 × 200 pixels, the size is not less than 148 × 148. Figure 5 and Figure 6 show sample distributions and partial examples for our dataset.

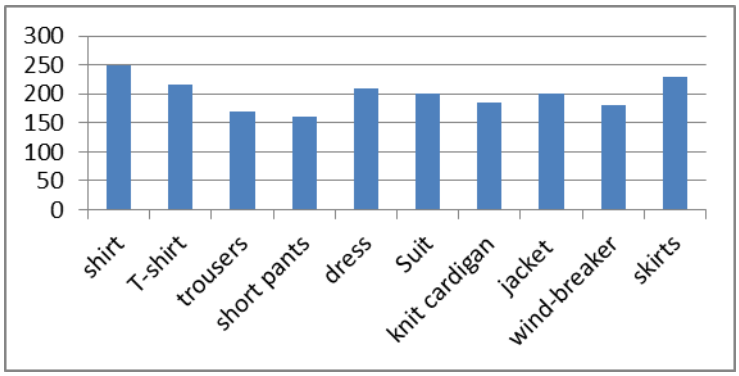

Fig. 5 the sample data distribution map

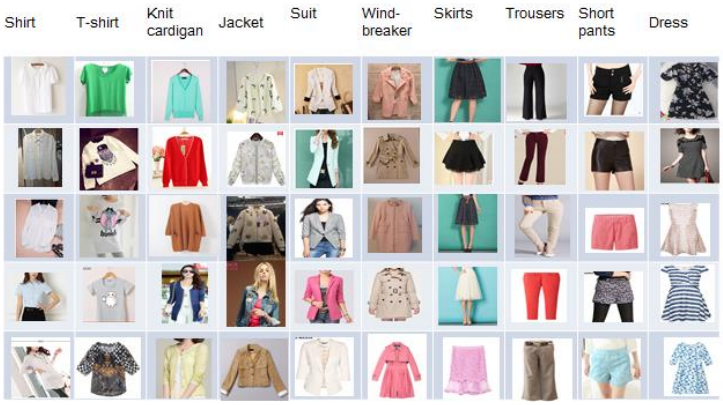

Fig. 6 part of the samples of our dataset

4.2 Results and Discussion

Using the parameters like above described, we obtain the experimental results shown in Table 4. Table 4 compares the experimental results of our work and three CNNs structure we used independently in the women clothing image dataset. Table 5 compares the performance of the proposed method with other image classification methods performance well nowadays in the dataset.

Table 4. Independent CNNs and combined CNNs accuracy in each category

	CNN-S	CNN-M	CNN-L	MHI-CNN
average accuracy	0.77	0.78	0.79	0.82
shirt	0.63	0.60	0.58	0.66
T-shirt	0.61	0.63	0.62	0.65
trousers	0.81	0.85	0.88	0.89
short pants	0.84	0.85	0.86	0.88
dress	0.88	0.90	0.93	0.94
Suit	0.76	0.80	0.84	0.84
knit cardigan	0.81	0.83	0.89	0.90

jacket	0.68	0.65	0.62	0.69
wind-breaker	0.85	0.84	0.88	0.88
skirts	0.80	0.88	0.85	0.88

Table 5. Comparison results of several different models

Model	Accuracy	Accuracy–with data augmentation
CNN[25]	0.76	0.78
FV[9]	0.81	0.83
MHI-CNN	0.79	0.82

Comparisons

1) The structure of CNN is one of the factors that affect the result of classification. Comparing the results of CNN independent experiments we can see: when tested separately CNN-L get the best performance, it is probably because of CNN-L's large input dimension and more number of feature maps. The input dimension is large and the loss of local feature information is relatively small. Combined with the experimental process of parameter adjustment, we can find the same CNN structure in a certain range to increase the number of feature maps can improve the classification results. This can be understood to extract more detail features.

2) Multi-scale CNN joint model can improve the accuracy of the FGC in women clothing images. Comparison of the independent CNN model and the CNN model we can see: The classification effect of the independent CNN model in 10 categories is different, but the overall poor performance does not mean that it is poor in all categories. CNN-S will be slightly inferior in overall performance, it is related with small input dimension and less feature map. But in the classification of shirts and jackets, CNN-S has a better performance than CNN-L, the possible reason is that CNN-S has a smaller convolution kernel, which is beneficial to the extraction of some local features. However, the experiment results of the joint CNN model are superior to the independent CNN model in each category and overall performance.

3) Compared with other FGC methods, our method is slightly inferior to that based on the Fisher vector, but it is better than other CNN structure. While the structure of the deep learning will be more conducive to dealing with large data, at the same time deep learning will take a lot of time training as the depth increases, our method can be trained in parallel and save the training time.

4) The number of samples is also one of the factors of fine classification. In order to discuss the effect of sample size on the classification results, we compared the classification results of using data augmentation or not. Table 4 gives the corresponding experimental results. From the results, the performance is significantly improved by using data augmentation. Because the number of training parameters in CNN structure are very large, the number of samples is relatively small. Without data augmentation, parameter training is not enough. That is the amount considered that data augmentation is a form of regularization. In fact, data amplification can also be used as a priori knowledge. The integration of prior knowledge can accelerate convergence and lead to more effective feature learning. Although CNN structure can do feature self-learning, it is difficult to add priori knowledge for it needing fixed input. Two forms of data augmentation in our work can be seen as prior knowledge. Image segmentation guide the CNN to learn the prospects of the main commodities, altering the RGB channel of images let CNN structure pay more attention to learn the contour and texture features.

5) Compare the differences between the various types of experiments, the classification accuracy of each subclass are different. Overall, trousers, shorts and dresses have a higher accuracy and T-shirts and shirts are relatively poor in performance. Perhaps because the contours feature of the first three categories are easier captured by the CNN than others. T shirt and shirt contours are similar, artificial classification is also very difficult. At this point it is necessary to make CNN tend to extract other features, our approach does just that.

5. Conclusion

.In this paper, a MHI-CNN model is proposed, which uses the feature extraction of heterogeneous CNNs to complete a fine-grained classification of women clothing images. We take advantage of multi-scale to design this model: First, multi-scale is helpful to CNNs models to obtain the abstract features of different levels. The convolution stride of model can be understood as level of abstraction in the multi-layer extraction process. When the stride is large, it can be considered as a rough high-level abstraction, and some details are lost, but sometimes it is advantageous to get effective features fast. Second, it help to different scale image input. It can be seen as a form of spatial pyramid pooling. We explore the CNN structure with a relatively shallow layer. It can reduce the computational complexity and training time, at the same time it does not need very large dataset. In the integration of CNN, we have adopted the voting mechanism. This method is simple and effective.

Compared with the existing studies, MHI-CNN appear to outperform state of the art approaches for women clothing datasets of the fine-grained classification. Our work is helpful to study the automatic classification of images, content-based product image retrieval and intelligent commodity recommended system in e-commerce platform. Of course, there are still a lot of work can be further improved in the future, such as finding better automatic segmentation method to reduce the image segmentation of the manual operation and using better integration to highlight the advantages of each CNN.

References

1. P.H. Gosselin, N. Murray, H. Jégou, F. Perronnin, Revisiting the Fisher vector for fine-grained classification, 2014, Pattern Recognition Letters, vol. 49, pp. 92-98.
2. M.E. Nilsback, A. Zisserman, Automated flower classification over a large number of classes, 2008, Computer Vision, Graphics & Image Processing, ICVGIP'08. Sixth Indian Conference on. IEEE.
3. C. Wah, S. Branson, P. Welinder, P. Perona, S. Belongie, The caltech-ucsd birds-200-2011 dataset, 2011.
4. B. Schiele, M. Andriluka, S. Amin, M. Rohrbach, A database for fine grained activity detection of cooking activities, 2012, Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. IEEE.
5. O.M. Parkhi, A. Vedaldi, A. Zisserman, C. Jawahar, Cats and dogs, 2012, Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. IEEE.
6. A.S. Razavian, H. Azizpour, J. Sullivan, S. Carlsson, CNN features off-the-shelf: an astounding baseline for recognition, 2014, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops.
7. B. Fernando, E. Fromont, T. Tuytelaars, Effective use of frequent itemset mining for image classification, 2012, European conference on computer vision.
8. Q. Chen, Z. Song, Y. Hua, Z. Huang, S. Yan, Hierarchical matching with side information for image classification, 2012, Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. IEEE.
9. J. Nchez, F. Perronnin, T. Mensink, J. Verbeek, Image classification with the fisher vector: Theory and practice, 2013, International journal of computer vision, vol. 105, no. 3, pp. 222-245.

10. X.T. Yuan, X.B. Liu, S.C. Yan, Visual classification with multitask joint sparse representation, 2012, IEEE Transactions on Image Processing, vol. 21, no. 10, pp. 4349-4360.
11. M. Varma, R. Debajyoti, Learning the discriminative power-invariance trade-off, 2007, Computer Vision, ICCV 2007, IEEE 11th International Conference on. IEEE.
12. P. Gehler, N. Sebastian, On feature combination for multiclass object classification, 2009, Computer Vision, 2009 IEEE 12th International Conference on. IEEE.
13. A. Iscen, G. Tolias, P.H. Gosselin, H. Jegou, A comparison of dense region detectors for image search and fine-grained classification, 2015, IEEE Transactions on Image Processing, vol. 24, no. 8, pp. 2369-2381.
14. M.D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, 2014, European conference on computer vision.
15. A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, 2012, Advances in neural information processing systems.
16. K. Chatfield, K. Simonyan, A. Vedaldi, A. Zisserman, Return of the devil in the details: Delving deep into convolutional nets, 2014, arXiv: 14053531.
17. D. Ciregan, U. Meier, J. Schmidhuber, Multi-column deep neural networks for image classification, 2012, Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. IEEE.
18. J.C. Chen, C.F. Liu, Visual-based deep learning for clothing from large database, 2015, Proceedings of the ASE BigData & SocialInformatics 2015.
19. K. Lin, H.F. Yang, K.H. Liu, J.H. Hsiao, C.S. Chen, Rapid clothing retrieval via deep learning of binary codes and hierarchical search, 2015, Proceedings of the 5th ACM on International Conference on Multimedia Retrieval.
20. N. Khosla, V. Venkataraman, Building Image-Based Shoe Search Using Convolutional Neural Networks, 2015, CS231n Course Project Reports.
21. A. Borràs, F. Tous, J. Lladós, M. Vanrell, High-level clothes description based on colour-texture and structural features, 2013, Iberian Conference on Pattern Recognition and Image Analysis.
22. Q. Chen, J. Huang, R. Feris, L.M. Brown, J. Dong, S. Yan, Deep domain adaptation for describing people based on fine-grained clothing attributes, 2015, Proceedings of the IEEE conference on computer vision and pattern recognition.
23. S. Lazebnik, C. Schmid, J. Ponce, Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories, 2016, CVPR 2006

24. C. Rother, V. Kolmogorov, A. Blake, Grabcut: Interactive foreground extraction using iterated graph cuts, 2004, ACM transactions on graphics (TOG), vol. 23, no. 3.
25. P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, Y. LeCun, Overfeat: Integrated recognition, localization and detection using convolutional networks, 2013, arXiv:13126229.
26. T. Zahavy, A. Magnani, A. Krishnan, S. Mannor, Is a picture worth a thousand words? A Deep Multi-Modal Fusion Architecture for Product Classification in e-commerce, 2016, arXiv:1611.09534.
27. S.C. Zhao, Y.J. Xu, Y.H. Han, 2016, Large-Scale E-Commerce Image Retrieval with Top-Weighted Convolutional Neural Networks, 2016, Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval.