Improving Deep Convolutional Neural Networks for Real-world Clothing Image

Ruifan Li[1,2] Yuzhao Mao[1] Ibrar Ahmad[1,3] Fangxiang Feng[4] Xiaojie Wang[1,2]
[1]School of Computer Science, Beijing University of Posts and Telecommunications, Beijing, China 1000876
[2]Engineering Research Center of Information Networks, Ministry of Education, Beijing, China 1000876
[3]Department of Computer Science, University of Peshawar, Peshawar, Pakistan 25120
[4]School of Digital Media and Design Arts, Beijing University of Posts and Telecommunications, Beijing, China 1000876
{rfli, maoyuzhao, xjwang}@bupt.edu.cn ibrar@upesh.edu.pk f.fangxiang@gmail.com

***Abstract*—Clothing images are abundant especially from the e-commercial platform, due to the rapid development of e-business. Recognizing and retrieving those images is of importance for commercial and social applications, which has recently been received tremendous attention from multimedia processing and computer vision. However, the large variations in clothing of their appearance and style, and even the large quantity of multiple categories and attributes make those problems challenging. Furthermore, the labels provided by shop retailers for real world images are largely erroneous or incomplete. Even worse, the imbalance problem among those image categories prevents the effective learning. To overcome those problems, we adopt a multi-task deep learning framework to learn effective representation. And we propose multi-weight convolutional neural networks for imbalance learning. The topology of this network is composed of two kinds of layers, shared layers at the bottom and task dependent ones at the top. Furthermore, category-relevant parameters are incorporated to regularize the learning procedure of backward gradients for different categories. We collect a large-scale dataset for those two problems containing about one million shop photos from four different Chinese retailers. Experiments on this dataset demonstrate that our proposed joint framework and multi-weight neural networks can effectively learn robust representation and achieve better performance.**

***Keywords*—Clothing Image recognition; Convolutional neural network; Multi-task; Multi-weight**

I. Introduction

Abundant clothing images are available in electronic commercial platform, such as taobao.com in China and amazon.com in USA. Scores of commercial and social applications are potentially related to recognizing and retrieving those images from webpages [1]–[6]. For example, a woman would like to automatically annotate her travel photo with the recognized clothing type and some attributes and then share with her relatives and friends. Another example, in a clothing image retrieval system effective classifying these clothing images and detecting their attributes are indispensable functionalities. And this intelligent search engine behind must learn a compact and effective representation to perform efficient retrieval and show satisfactory results.

With the the increasingly business values of fashion and shopping industry, automatic clothing image analysis has received tremendous attention. Learning robust representations for images is the key to perform effective retrievals. However, to label all those images are almost impossible by human beings. Automatic creating the labels using the available textual information around inevitably gives rise to the noisy labels. Here the noisy means that some labels are incorrect and some are missing. The significant difference between the actual shop labels and the expected ones makes learning an effective representation even more difficult. Considering that this phenomenon pervasively exists in web clothing image data, how to cope with that is an unavoidable problem. This is the motivation of our paper to build our model. Another phenomenon within those large-scale web clothing images is the imbalance among categories. And this will be illustrated in detail in Section IV-A. Due to the extremely imbalance among categories within the large volume of clothing image data, training such a deep neural model is also challenging. That is to say, how to train a neural network with large-scale data under the extreme imbalance among categories should be dealt with. This is the motivation of our paper to propose our multi-weight neural networks.

In this paper, we focus on robust modeling the multiple source relationship among real world large-scale images and their automatically generated unreliable and incomplete labels, such that robust representations could be learned for effective retrieval. Unlike the traditional image classification and attributes detection, our task is performed in a strictly controlled dataset with nearly correct labels. To this end, we should deal with two challenging questions for this task. The first problem is that the labels provided from the shop retailers are, to a large extent, erroneous or incomplete. This causes it difficult to capture the relationship between clothing images and their vulnerable labels. The second problem is that those large quantities of clothing categories and attributes are in an extreme imbalance. This makes it difficult to perform an effective model learning.

To tackle those problems, we take a data-driven approach and propose multi-weight neural networks to retrieving clothing images in real world. In this multi-weight CNNs, some closely related categories and their attributes are grouped together as a single task and all those tasks share a common representation. Furthermore, the weight adaptation is incorporated to regularize the backward gradients in the learning algorithm from different categories, in order to deal with the imbalance problem resulted from different categories of

978-1-5386-2165-3/17/$31.00 ©2017 IEEE

clothing images. To evaluate our multi-weight CNNs, we collect a large, complex, and real-world collection of clothing image dataset, e-Clothing1.4M. We then compare the proposed multi-weight CNNs on this e-Clothing1.4M with another two methods. Experimentally, we demonstrate that our multi-weight networks performs best among those methods. And the proposed multi-weight can effectively deal with the erroneous and incomplete labels and data imbalance problems in large-scale real world clothing images.

The remainder of this paper is organized as follows. We firstly review related work in Section II. We introduce the details of our multi-weight neural nets, the architecture and its learning algorithm, are described in Section III. Section IV describes the experimental dataset, e-Clothing1.4M. The experimental results are then given and discussed. Finally, Section V draws the conclusion and shows some future work.

II. Related Work

With the increasingly large business value of shopping industry, automatic clothing image analysis has received tremendous attention. Significantly, one trend is to use attribute learning to give more much fine-grained description for clothing image, which has been widely explored in computer vision community [7]–[14]. However, one of the major challenges faced with attribute learning is lack of well-labeled training data because of the heavy cost of human laboring. Besides, obtaining these attributes usually require some domain-specific knowledge, which can then be applied for labeling the data. To overcome this difficulty, Berg et al. [10] propose to automatically obtain attributes and visual appearance by mining the descriptive text of images from webpages. For clothing images, Chen et al. [9] focus on learning visually attributes of clothing on the human upper body only. Recently, Shankar et al. [12] propose to discover all attributes present in an image, in a weakly supervised scenario, based on deep neural networks. Generally speaking, those research works only take attribute learning as a single task. However, when dealing with large-scale clothing images, the attributes are highly-related with the clothing category and those categories cannot be ignored during attributes discovering.

Analyzing clothing images, from another perspective, is based on methods from pose estimation and person detection [15]–[19]. Intuitively, human recognition is related to clothing image recognition. Therefore, human recognition motivated clothing image analyzing is interesting. Clothing parsing is to predict a semantic category, such as shirt, skirt, and shoes, for each pixel in an image. The parsing results then could be further used for clothing recognition. Most notably, Liu et al. [15] address the cross-scenario problem that a daily human photo is performed to retrieval a clothing shop photo. They alleviate the discrepancy of those two distributions through a sparsely coded transfer matrix. Kalantidis et al. [17] also consider a similar cross-scenario approach, where they start from pose estimation and then utilize clothing parsing. Yamaguchi et al. [18] recently propose an unconstrained clothing parsing without user-provided tag information for clothing retrieval. The insight obtained from those methods is the use of body pose estimation for clothing parsing. However, the performance of those approaches largely rely on an accurate pose estimation and human parts detection, and cannot easily extend to a large-scale clothing parsing and recognition system.

From the year of 2006 [20], deep learning motivated by the biological distributed structure of human brain is proposed to learn hierarchical and effective representations to facilitate various computer vision tasks. The basic idea of deep learning methods is to use some simple non-linear neural neurons to compositionally build a complex fitting function. Deep learning methods especially supervised convolutional neural networks (CNNs) [21]–[23] and unsupervised autoencoders and restricted Boltzmann machines, has successfully been applied due to the availability of computational power and the volume of data in large-scale image classification [21] and cross-modal retrieval applications [24]. Notably, deep learning also has an advantage for multi-task learning, which aims to achieve better performance by simultaneously exploring multiple closely related tasks. Deep learning methods learn hierarchical representations which capture those underlying factors. Because of the natural connection, multi-task learning could then be a possible means for large-scale clothing image analysis. Very recently, several methods based on deep learning for multi-task learning have been proposed [25]–[28]. Notably, Zhang et al. [25] propose to combine part-based models and CNNs for feature representation in order to obtain an attribute description for human under the multi-task framework. However, this proposed framework is specifically designed for small-scale datasets and cannot easily be extended to large-scale problems. Bai et al. [26] propose a multi-task deep networks for text-based image retrieval. In this framework, query-sharing layers for image representation and query-specific layers for relevance estimation are learned jointly. In general, the representative power of CNNs compared with shallow hand-crafted visual features, such as HoG and SIFT, shades light on learning multiple tasks possible. However, the performance of those partly fully-connected neural networks heavily rely on qualities of data labels. Besids, those CNN-based methods ignore the correlation between attributes and neither the cross category of visual attributes especially for large-scale clothing dataset.

III. Proposed Convolutional Neural Networks

In this section, we present our multi-weight CNN model and its learning algorithm. The general convolutional operation is CNN networks is illustrated in Figure 1. The left is the feature map in the previous layer. The middle is the convolutional kernel. And the right is the feature map in the next layer. Specifically, a patch in one image of 4×4 is convoluted with a kernel of 2×2. And by moving the patch window the convoluted result image of 3×3 are obtained. Evidently, the multi-weight CNN is composed of two groups of layers: task independent layers and task dependent layers. And the weights to be learned in task independent layers are denoted as W_c, while the weights to be learned of task dependent layers are

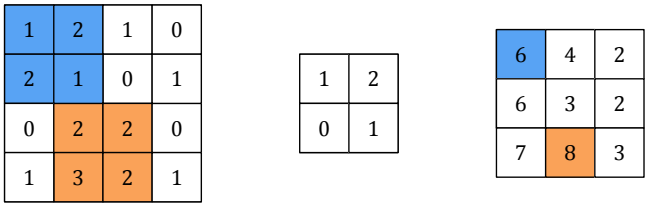

Fig. 1. Illustration of convolutional operation in CNN networks.

denoted as W_t. In addition, we introduce a group of hyper-parameters γ_t to balance the losses incurred from different categories. To make it clear, we denote the mapping of task independent layers as $\phi_c(\cdot)$ and that of a specific task layers as $\phi_t(\cdot)$. Then, for a mini-batch of images I_b the representation o_c of task independent is obtained as $\phi_c(I_b; W_c)$, and the output of a specific task t is obtained as $\phi_t(o_c; W_t)$. Thus, under the supervised learning framework, learning this multi-weight network parameters can be cast into an optimization problem.

For a specific task t, the output layer takes on the multi-label structure. Here, the multi-label treats each label equally. It is straight-forward that the loss in tth task is taken as a sigmoid cross-entropy function. Each image I_j is expected to have a vector of label probability p_j, having length M_t. Note that the different length M_t comes from the number of labels in tasks. With the sigmoid cross-entropy loss, the network parameter $\Theta_t = \{W_c, W_t\}$ is learned by minimizing the following objective function,

$$\min_{\Theta_t} \mathcal{J}_t(\Theta_t) = \frac{1}{B} \sum_{b=1}^{B} \sum_{m=1}^{M_t} \ell_{bm} \tag{1}$$

in which, the sigmoid cross-entropy loss ℓ_{bm} takes the form

$$\ell_{bm} = -[p_{b_m} \log \hat{p}_{bm} + (1 - p_{bm}) \log (1 - \hat{p}_{bm})] \tag{2}$$

where the probability vector $\hat{p}_b$ is obtained by applying the sigmoid function to each of the M_t outputs of layer L8 on the tth task. Then, our objective is to optimize the function $\mathcal{J}$,

$$\min_{\Theta} \mathcal{J}(\Theta) = \sum_{t=1}^{T} \gamma_t \mathcal{J}_t(\Theta) \tag{3}$$

in which, $\Theta = \{\Theta_t\}, t = 1, 2, \cdots, T$ is the total parameters to be learned in our model.

To optimize the previous objective function, we apply the gradient decent method with back propagation. The gradient of the objective function with respect to W_t is

$$\nabla_{w_t} \mathcal{J}(\Theta) = \frac{1}{B} \gamma_t \sum_{b=1}^{B} \sum_{m=1}^{M_t} \frac{\partial \ell_{bm}}{\partial W_t} \tag{4}$$

which only average the gradients over the small batch of training images for the tth task. Similarly, the gradient of the objective function with respect to W_c is

$$\nabla_{w_c} \mathcal{J}(\Theta) = \frac{1}{B} \sum_{t=1}^{T} \sum_{b=1}^{B} \sum_{m=1}^{M_t} \gamma_t \frac{\partial \ell_{tbm}}{\partial W_c} \tag{5}$$

where gradients of all training images in all tasks are averaged. The multi-weight CNNs are trained using stochastic gradient descent through a forward and a backward pass. The details of the training algorithm is summarized as in Algorithm 1.

Algorithm 1 Learning Algorithm

1: Randomly initialize the common weights W_c and task specific weights $\{W_t\}_{t=1}^{T}$, set the learning rates ϵ_c for weights W_c and ϵ_t for weights W_t, and set the weight adaption parameter γ_t for each task.

2: Compute the common representations for a mini-batch of images I_b,

$$o_c \leftarrow \phi_c(I_b; W_c) \tag{6}$$

3: Compute the task specific outputs,

$$o_t \leftarrow \phi_t(o_c; W_t) \tag{7}$$

4: Compute the weights gradients ∇W_c and ∇W_t with respect to the objective function according to Eq. (4) and Eq. (5).

5: Update the common weights,

$$W_c \leftarrow W_c + \epsilon_c \cdot \nabla_{W_c} \mathcal{J} \tag{8}$$

6: Update the task specific weights,

$$W_t \leftarrow W_t + \epsilon_t \cdot \nabla_{W_t} \mathcal{J} \tag{9}$$

Note that the task dependent weight adaption parameters γ_t are incorporated to regularize the learning procedure.

7: Repeat Steps 2 - 6 until convergence.

IV. Experiments

We evaluate the proposed multi-weight CNNs with the other two deep networks, multi-label CNNs and multi-task ones. All our experiments are performed on our e-Clothing1.4M real-world dataset. We first describe the dataset collected. Next, we introduce the evaluation criteria adopted in our experiments. Thereafter, the settings of those methods are given. And a simple but effective method for setting the weight are proposed. At last, the performance of deep neural models are reported followed by an analysis on our deep models.

A. Dataset

We collect a new dataset, called e-Clothing1.4M, to evaluate our proposed methods. To the best of our knowledge there is no available real-world dataset for our tasks. Specifically, a large number of pairs of images and their description words contained within webpages are crawled down from four widely used e-commercial platforms, including JD.com, taobao.com, meilishuo.com, and mugujie.com. By performing word segmentation, choosing high frequency words of description words, and checking by human, we design a clothing catalog with types and attributes. To sum up, we obtain 830 keywords and 24 groups for clothing types and attributes. Furthermore, through data cleaning we obtain 1,462,438 pairs of images and their labels of types and attributes. For evaluation, we

TABLE I
STATISTICS OF OUR E-CLOTHING1.4M

Statistic	Value
Number of training data	1,069,901
Number of test data	392,537
Number of Groups	24
Maximum Number of Attributes	130
Minimum Number of Attributes	2
Mean of Number of Attributes	22
Std of Number of Attributes	31

randomly divide the e-Clothing1.4M dataset into the training dataset and the test one. And 1,069,901 images of 350,000 products are set for training; 392,537 images are set for test while keeping the quantity of each group of labels in balance. Statistics on the amount of data in 24 categories shows the extremely imbalance, as shown at the left column in Table II. Evidently, the foremost category is the common attribute 'Color' with the maximum proportion one. The aftermost is miscellaneous attribute 'Wool Thickness' with the proportion 0.0002. Intuitively, images in this e-Clothing1.4M dataset are distributed as a heavy-tail distribution. To summarize, statistics of this dataset is given in Table I.

B. Evaluation Criteria

In this section, we describe our metrics used in our experiments. The mAP measures the discrimination and stability of those learning algorithms. Specifically, given one query and the first R top-ranked retrieved data, the average precision is defined as follows,

$$mAP \triangleq \frac{1}{M}\sum_{r=1}^{R} p(r) \cdot rel(r) \tag{10}$$

where M is the number of relevant data in the retrieved result, $p(r)$ is the precision at r, and $rel(r)$ presents the relevance of a given rank (one if relevant and zero otherwise). The retrieved data is considered as relevant if it has the same semantic label as the query. Then, the mAP score is obtained by averaging AP of all the queries. In order to use the metric mAP, we build an image retrieval system based on the learned representation. In this system, the dataset under consideration is used both as the query set and as the candidate set. And for each task, if a query and a retrieved result has common labels, then the relevance is set to one. We then compute the mAP score for each individual task. In all our experiments we report the results with mAP@50 ($R = 50$).

C. Methods and Settings

We firstly compare multi-weight CNNs with two methods, multi-label CNNs and multi-task CNNs on our e-Clothing1.4M dataset. Furthermore, the other interesting experiments are performed on two categories data with an intentionally designed ratio. Practically, our three kinds of CNNs implementations largely depends on the Caffe deep learning framework [29]. And the prominent AlexNet CNN architecture [21] is adopted. All our experiments are conducted on a workstation with an NVIDIA K20c GPU card, dual E5-2650 CPUs with 2.00 GHz, and 64 GB main memory. As a preprocessing step, the size of images is resized to 256 $\times$ 256 with RGB three channels, i.e. 256 $\times$ 256 $\times$ 3. Then the resized image is fed into the AlexNet network. Subsequently, the network will randomly crop the image with a size of 224 $\times$ 224. Namely, an image with 224 $\times$ 224 $\times$ 3 is received.

Through the pre-processing, the size of input layers of all three methods are set as 196,608. In terms of the original AlexNet, the sizes of subsequent layers are set the same as those in AlexNet, i.e. 253,440 $\rightarrow$ 186,624 $\rightarrow$ 64,896 $\rightarrow$ 64,896 $\rightarrow$ 43,264 $\rightarrow$ 4096 $\rightarrow$ 4096 from L1 to L7. The output layer is then configured to 830 sigmoid neurons according to the categories and attributes. The first convolutional layer filters input images with 96 kernels of size $11\times11\times3$ with a stride of 4 pixels. The second convolutional layer takes as input from the output of the first convolutional layer and filters it with 256 kernels of size $5\times5\times48$. Subsequently, the third, fourth, and fifth convolutional layers are concatenated to each other without any pooling layers. The third convolutional layer has 384 kernels of size $3\times3\times256$ connected to the outputs of the second convolutional layer. The fourth convolutional layer has 384 kernels of size $3\times3\times192$, and the fifth convolutional layer has 256 kernels of the same size of fourth layer. The fully-connected layers have 4,096 neurons. Besides, we use dropout trick in the first two fully-connected layers by setting to zero the output of each hidden neuron in those two layers with a probability 0.5. And the Regularized Linear Unit (ReLU) non-linearity neuron is adopted to the output of convolutional and fully-connected layer. For the multi-weight CNN, we set the task dependent parameter γ_t as the amount proportion of the tth task. This proportion has been shown in Table II.

Here we make clear the main novelties of multi-weight CNNs compared with multi-label CNNs and multi-task CNNs. The basic idea of this multi-label CNNs is to treat each label equally. Specifically, the number of output neurons of this multi-label network is set as the total number of categories and attributes. And each output neuron is configured as a sigmoid function. For the multi-task CNNs, the idea is to learn multiple related problems together at the same time. In this paper, recognizing clothing images and detecting their attributes are treated in the framework of multi-task. Lastly, multi-weight CNNs treat each category and its related attributes in one group, making use of the fine-grained relationship among them. Meanwhile, multi-weight CNNs consider the imbalance among those groups of categories.

D. Experimental Results

The experimental results on the eClothing1.4M dataset are summarized in Table II. All numerical values in those two Table are mAP scores. At the left most column in Table II, the category names are orderly shown according to their own normalized proportion. That is, the "Color" which has the maximum quantity images is shown at the top. On the contrary, "Wool Thickness" is shown at the bottom. Then, the results of three CNN-based methods, multi-label, multi-

task, and multi-weight, with their number of iterations are shown at the subsequent columns, respectively. The mAP value in boldface standards for the best performance achieved from those three methods. In all 24 categories, the multi-label CNNs performs the best within seven categories. The multi-task CNNs achieves the best within five categories. And the multi-weight CNNs show the best performance within the other twelve categories. In general, those three CNN-based method averaging over those 24 categories have the mAP values: 52.48% for multi-label, 53.11% for multi-task, and 55.23% for multi-weight. Clearly, multi-weight CNNs show the best performance compared with the other two methods.

Note that for small proportion, such as "Waist", "Feather", and "Sleeve", have gained some improvement. Also note that in Table II the five categories, "Dense", "Leather", "Cardigan", "Cheongsam", and "Wool Thickness" achieve the best performance compared with multi-label CNNs and multi-weight ones. Those five categories have very low proportions compared with the maximum ratio category "Color". Those demonstrate that multi-weight CNN is an suitable method for our clothing recognition and attribute detection task. And appropriate weight proportion could improve the experimental results. Therefore, using the multi-weight CNNs could, to some extent, deal with the imbalance problem and noisy labels in real world clothing dataset.

We intentionally design the ratio to show the effectiveness of regularize weights of categories in multi-weight CNN. Table IV and Table V shows mAP scores for independent "Coat" and "Pantsuit" vs. dependent "Length of Sleeve" and "Sleeve" categories with the quantity ratio 10:1. The mAP in boldface denotes the best performance among those three methods. Those two groups have similar results when using multi-weight CNN networks, irrespective of the dependency relationship between categories. Specifically, for the category of independent group with smaller ratio "Pantsuit", the multi-weight achieve the best performance. For the category of dependent group with smaller ratio "Sleeve", the multi-weight also achieve the best performance. This demonstrates that for the category with smaller ratio the mAP score improves using multi-weight CNNs compared with the other two CNN-based methods. Therefore, when dealing with the data with the category imbalance, multi-weight CNNs could learn better representations, which is benefited from the common information contained within those kinds of data.

E. Setting Weights

Evidently, how to set the weights in our multi-weight neural networks is of importance for imbalance learning. In the original proposed solution, we set the weights according to the ratio of the numbers of all categories. However, this strategy would excessively boosts the values for small sample categories. Therefore, to alleviate this effect, we propose a smoothing function as follows,

$$w_n = \begin{cases} r_n & \text{if } r_n \geq r_{avg}/c, \\ r_{avg}/c & \text{if } r_n \leq r_{avg}/c, \end{cases} \tag{11}$$

in which, the ratio r_n is the number of samples in nth category compared to the maximum of samples among categories. And the r_{avg} is the mean ratio among all categories. The hyper-parameter c is set as 3.0 here. The tuition behind that smoothing function is that for categories with smaller ratios we increase their ratios to suitable ones and keep the larger ratios unchanged. That regularization strategy would result in effective learning. In our previous experiments, the value of the average ratio r_{avg}/c is set to 0.11. In effect, those tasks with ratios lower than 0.11 (the maximum ratio is 1.0 for "Color") are regularized. We use those new weights and obtain the following results, shown in Table III. Evidently, the new setting weights strategy outperforms the other methods in nearly all categories. And it indeed shows superior to the previous method for weights setting. To better visualize the performance of all those methods, we show the histograms of the $mAPs$ for all 24 categories in Figure 2.

V. Conclusion

Recognizing and retrieving the real-world large-scale clothing images is crucial for e-commercial and social applications. In this paper, we propose a multi-weight convolutional neural networks to deal with the noisy and imbalanced clothing images. Our multi-weight CNN comprises of common layers and task dependent layers with category-relevant parameters. We collect a large-scale dataset containing about one million shop photos from four different Chinese retailers. Experiments on this dataset show the effectiveness of our multi-weight neural networks. We still have some interesting problems to investigate for our future work. We can extend this network by introducing other convolutional networks to improve the performance. Besides, we would like to investigate the weights impacting on the network convergence.

Acknowledgement

This work was partially supported by National Natural Science Foundation of China (No. 61273365, No. 61472046, and No. 61472048) and Discipline Building Plan in 111 Base (No. B08004). The authors thank Prof. Chuan Shi at Beijing University of Posts and Telecommunications for reading the draft of this paper and for giving helpful comments. The authors would also like to thank the editor and the anonymous reviewers for useful comments and suggestions that allowed them to improve the final version of this paper.

References

[1] M. H. Kiapour, X. Han, S. Lazebnik, A. C. Berg, and T. L. Berg, "Where to Buy It: Matching Street Clothing Photos in Online Shops," in *2015 IEEE International Conference on Computer Vision (ICCV'15)*. IEEE, 2015, pp. 3343–3351.

[2] Y. Jing, D. Liu, D. Kislyuk, A. Zhai, J. Xu, J. Donahue, and S. Tavel, "Visual search at pinterest," in *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015, pp. 1889–1898.

[3] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang, "Deepfashion: Powering robust clothes recognition and retrieval with rich annotations," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1096–1104.

[4] X. Wang, Z. Sun, W. Zhang, Y. Zhou, and Y. G. Jiang, "Matching user photos to online products with robust deep features," in *ACM ICMR*, 2016, pp. 7–14.

TABLE II
mAP SCORES USING THREE CNNS WITH THEIR NUMBER OF ITERATIONS.

Category	Proportion	Multi-label@i20 (%)	Multi-task@i45 (%)	Multi-weight@i45 (%)
Color	1.0000	**35.24**	27.03	34.66
Gender	0.9646	99.27	99.33	**99.52**
Senson	0.8242	67.49	69.25	**70.37**
Style	0.7773	**34.40**	31.04	33.78
Coat	0.6838	**74.82**	71.20	74.25
Version	0.6062	**42.75**	37.88	42.62
Pantsuit	0.6018	**57.05**	52.65	56.01
Length of Sleeve	0.4201	61.36	62.75	**64.19**
Occasion	0.3876	44.97	43.67	**45.61**
Craft	0.3119	**18.14**	17.45	17.33
Colar	0.2436	33.17	32.44	**33.71**
Length of Trousers	0.2292	72.92	72.30	**72.97**
Thickness	0.1961	68.44	69.24	**69.54**
Length of Skirts	0.1866	57.06	59.00	**59.44**
Skirt	0.1766	**44.42**	43.46	44.22
Crowd	0.1456	71.65	76.46	**77.51**
Dense	0.1331	87.13	**94.21**	94.08
Leather	0.0643	54.64	**59.76**	58.48
Cardigan	0.0626	33.84	**39.77**	38.61
Waist	0.0471	44.76	49.01	**49.25**
Feather	0.0428	79.73	78.44	**80.10**
Sleeve	0.0064	20.73	28.11	**28.75**
Cheongsam	0.0006	32.91	**33.84**	33.65
Wool Thickness	0.0002	22.67	**26.42**	26.27

TABLE III
mAP SCORES FOR EACH CATEGORY USING WEIGHT ADAPTATION WITH THEIR NUMBER OF ITERATIONS.

Category	Proportion	Current Best (%)	multi-weight@i45 (%)	$1/3$Ratio@i45 (%)
Color	1.0000	35.24	34.66	**36.03**
Gender	0.9646	99.52	99.52	**99.55**
Senson	0.8242	70.37	70.37	**70.76**
Style	0.7773	**34.40**	33.78	34.34
Coat	0.6838	74.82	74.25	**75.20**
Version	0.6062	42.75	42.62	**43.66**
Pantsuit	0.6018	57.05	56.01	**57.28**
Length of Sleeve	0.4200	64.19	64.19	**64.57**
Occation	0.3875	45.61	45.61	**45.84**
Craft	0.3119	18.14	17.32	**18.15**
Colar	0.2436	33.71	33.71	**35.27**
Length of Trousers	0.2292	72.97	72.97	**73.90**
Thickness	0.1961	69.54	69.54	**70.39**
Length of Skirts	0.1866	59.44	**59.44**	59.41
Skirt	0.1766	44.42	44.22	**45.26**
Crowd	0.1456	77.51	77.51	**77.93**
Dense	0.1331	94.21	94.08	**94.42**
Leather	0.0643	59.76	58.48	**60.40**
Cardigan	0.0626	39.77	38.61	**41.42**
Waist	0.0471	49.25	49.25	**51.80**
Feather	0.0428	80.10	80.10	**80.47**
Sleeve	0.0064	28.75	28.75	**29.57**
Cheongsam	0.0006	33.84	33.65	**41.08**
Wool Thickness	0.0002	26.42	26.27	**28.43**

TABLE IV
mAP SCORES FOR COAT V.S. PANTSUIT WITH THE RATIO 10:1

Category	Proportion	Multi-label (%)	Multi-task (%)	Multi-weight (%)
Coat	1	57.47	**58.52**	57.86
Pantsuit	0.1	31.94	32.28	**33.25**

TABLE V
mAP SCORES FOR LENGTH OF SLEEVES V.S. SLEEVE WITH THE RATIO 10:1

Category	Proportion	Multi-label (%)	Multi-task (%)	Multi-weight (%)
Length of Sleeve	1	**46.71**	43.06	45.58
Sleeve	0.1	13.58	15.43	**17.04**

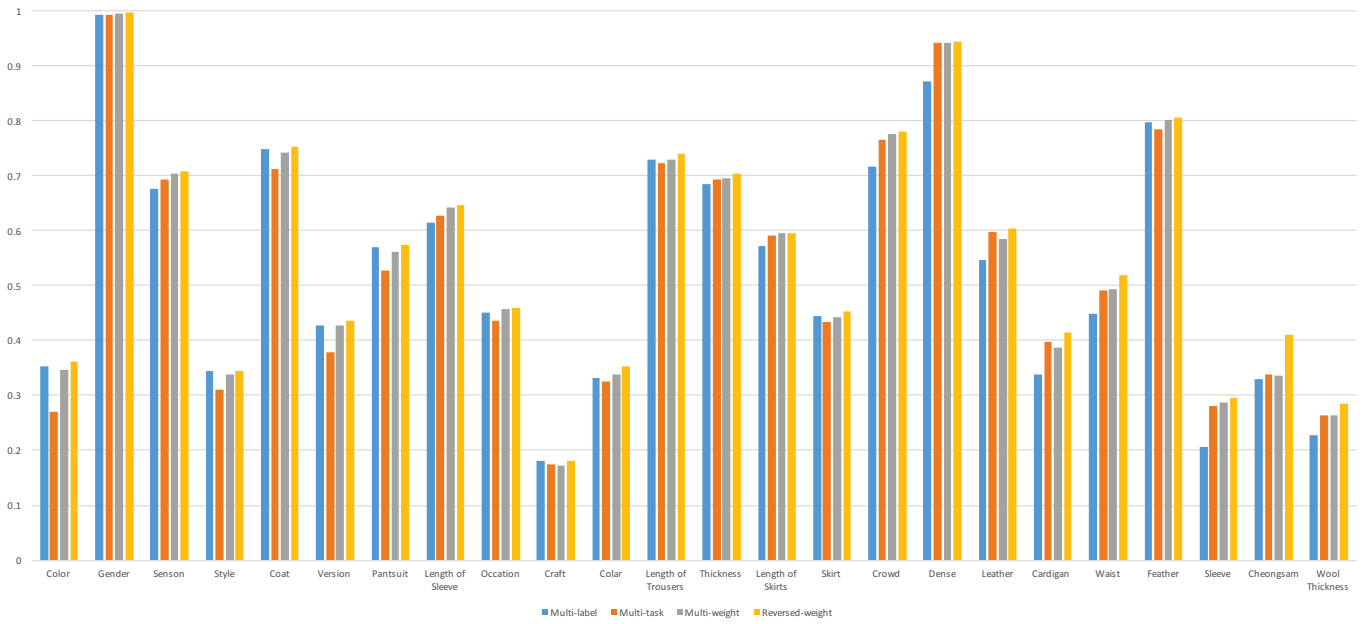

Fig. 2. Performance comparison over for all twenty-four categories using four kinds of methods.

[5] D. Shankar, S. Narumanchi, H. A. Ananya, P. Kompalli, and K. Chaudhury, "Deep learning based large scale visual recommendation and search for e-commerce," *arXiv:1703.02344 [cs.CV]*, 2017.

[6] A. Zhai, D. Kislyuk, Y. Jing, M. Feng, E. Tzeng, J. Donahue, Y. L. Du, and T. Darrell, "Visual discovery at pinterest," *arXiv:1702.04680 [cs.CV]*, 2017.

[7] O. Russakovsky and L. Fei-Fei, "Attribute learning in large-scale datasets," in *European Conference of Computer Vision (ECCV), International Workshop on Parts and Attributes*, Crete, Greece, September 2010.

[8] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar, "Describable Visual Attributes for Face Verification and Image Search." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 10, pp. 1962–1977, 2011.

[9] H. Chen, A. Gallagher, and B. Girod, "Describing clothing by semantic attributes," in *Proc. of European Conference on Computer Vision (ECCV'12)*, Firenze, Italy, 2012, pp. 609–623.

[10] T. L. Berg, A. C. Berg, and J. Shih, "Automatic Attribute Discovery and Characterization from Noisy Web Data," in *European Conference on Computer Vision (ECCV)*, no. PART 1, 2010, pp. 663–676.

[11] W. Di, C. Wah, A. Bhardwaj, and R. Piramuthu, "Style finder: Fine-grained clothing style detection and retrieval," in *Computer Vision and Pattern Recognition Workshops*, 2013, pp. 8–13.

[12] S. Shankar, "DEEP-CARVING : Discovering Visual Attributes by Carving Deep Neural Nets," in *CVPR*, 2015.

[13] K. Lin, H. F. Yang, K. H. Liu, J. H. Hsiao, and C. S. Chen, "Rapid clothing retrieval via deep learning of binary codes and hierarchical search," in *Proc. ACM International Conference on Multimedia Retrieval*, 2015, pp. 499–502.

[14] Q. Dong, S. Gong, and X. Zhu, "Multi-task curriculum transfer deep learning of clothing attributes," *arXiv:1610.03670 [cs.CV]*, 2016.

[15] Si Liu, Zheng Song, Guangcan Liu, Changsheng Xu, Hanqing Lu, and Shuicheng Yan, "Street-to-shop: Cross-scenario clothing retrieval via parts alignment and auxiliary set," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 3330–3337.

[16] K. Yamaguchi, M. H. Kiapour, L. E. Ortiz, and T. L. Berg, "Parsing clothing in fashion photographs," in *Computer Vision and Pattern Recognition*, 2012, pp. 3570–3577.

[17] Y. Kalantidis, L. Kennedy, and L.-J. Li, "Getting the look: Clothing recognition and segmentation for automatic product suggestions in everyday photos," in *Proceedings of the 3rd ACM Conference on International Conference on Multimedia Retrieval*, ser. ICMR '13. New York, NY, USA: ACM, 2013, pp. 105–112.

[18] K. Yamaguchi, M. H. Kiapour, L. E. Ortiz, and T. L. Berg, "Retrieving Similar Styles to Parse Clothing," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 5, pp. 1028–40, 2015.

[19] P. Tangseng, Z. Wu, and K. Yamaguchi, "Looking at outfit to parse clothing," *arXiv:1703.01386 [cs.CV]*, 2017.

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

[21] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1097–1105.

[22] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv:1409.1556 [cs.CV]*, 2015.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

[24] F. Feng, R. Li, and X. Wang, "Deep correspondence restricted Boltzmann machine for cross-modal retrieval," *Neurocomputing*, vol. 154, no. C, pp. 50–60, 2015.

[25] N. Zhang, M. Paluri, M. Ranzato, T. Darrell, and L. D. Bourdev, "PANDA: Pose Aligned Networks for Deep Attribute Modeling." *CVPR*, pp. 1637–1644, 2014.

[26] Y. Bai, K. Yang, W. Yu, W.-Y. Ma, and T. Zhao, "Learning High-level Image Representation for Image Retrieval via Multi-Task DNN using Clickthrough Data." *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2013.

[27] D. Wang, X. Gao, X. Wang, L. He, and B. Yuan, "Multimodal discriminative binary embedding for large-scale cross-modal retrieval," *IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society*, vol. 25, no. 10, pp. 1–1, 2016.

[28] E. Simoserra and H. Ishikawa, "Fashion style in 128 floats: Joint ranking and classification using weak data for feature extraction," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 298–307.

[29] 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.