Extreme-WJLD submission to JD AI Fashion Challenge

1st Sanyuan Liu
University of Electronic Science and Technology of China.
sanyuan@std.uestc.edu.cn

2nd Shiguang Wang
University of Electronic Science and Technology of China.
wangshiguang@std.uestc.edu.cn

3rd Xin Du
Shenzhen Extreme Vision Technology Limited.
xin.du@extreme-vision.com.cn

4th Yun Luo
Shenzhen Extreme Vision Technology Limited.
lauren.luo@extremevision.com.cn

5th Jian Cheng
University of Electronic Science and Technology of China.
chengjian@uestc.edu.cn

***Abstract*—In this paper, we introduce the submissions for the task of Fashion Style Recognition in JD AI Fashion-Challenge. Fashion Style Recognition aims to recognize the clothing styles according to the local part design or the global collocation. Although it has attracted various attention, clothing style recognition still faces a lot of troubles (e.g. pose, light and appearance variation). Here, we concentrate on the following aspects to improve the recognition performance: data argumentation, transfer learning, model fusion and the style relationship. Though intuitively, our approach achieved satisfactory performance, which won the 3rd place on the leader board (F2-score 0.6524).**

***Index Terms*—Style Recognition, Fashion-Challenge**

I. Introduction

Visual fashion style recognition has drawn increasing attention in recent years, due to its wide spectrum of applications in research field and commercial areas. Extensive research efforts have been devoted to clothing classification [1]–[3], attribute prediction [4], [5], and clothing item retrieval [6], [7]. It is a challenging task because of the large variations among the clothing items, such as the change of poses, light, scales, and appearance. To reduce these variations, existing works tackled these problems by seeking the information regions (e.g. the clothing bounding boxes, the semantic local part [8], or the human joints [9]). However, these extra labels are hard to obtain in practical situation.

We proposed this technical report for JD AI Fashion Challenge to address the fashion style recognition problem. The goal of the challenge is to guide computer to automatically recognize the fashion style category of the clothes. The proposed dataset contains 54908 images for training and validation, and 10000 images for the final testing. The dataset is labeled by the fashionable professionals. Each image is labeled with 13 binary labels (belongs the category or not), and each image belongs to at least 1 style. The 13 style classes and their corresponding index are listed in Table I. We also make a statistic summary for the class distribution in Fig. 1. As we can see, the data distribution is seriously imbalance. For example, the proportion of sports style (index 1) is 0.189%. On the other hand, some styles are closely related, while some styles are inversely related. The conditional probability of these styles is illustrated in Fig. 2. As can be seen, some classes are closely related (e.g. class 1 and class 2), while some classes (class 4, 5, 6 and 7) never show up together. Therefore, the class imbalance is a huge challenge we have to focus, and our model is mainly designed to deal with the problem to avoid over-fitting.

Since the deep convolution neural networks achieved enormous success in image classification [10], our method is also based on it. To address the challenge, our solution mainly applied the following skills: data argumentation, transfer learning and style relationship modeling. On the whole, we achieved 0.6524 (F2 score) on the testing set, which won the 3rd place on the leader board.

II. Proposed Methods

For the style recognition task, we considered the following four aspects: data augmentation, network architecture, transfer learning, cost-sensitive learning and styles relationship.

A. Data Augmentation

- Resampling the training set
  To avoid over-fitting, we resampled the training set to make sure each training batch contains proper positive samples. Moreover, we undersampled negative samples for classes $1, 2, 4, 7, 8, 10, 11, 12, 13$ and kept the original proportion for classes $3, 5, 6, 9$. The core idea of this operation is trying to make a balance on improving

TABLE I
SUMMARY OF THE 13 STYLE INDEXES PROVIDED BY THE JDAIFASHION DATASET.

Index.	Style.	Index.	Styles.
1	Sports	8	Girl
2	Casual	9	Lady
3	Office lady	10	Minimalist
4	Japanese	11	Natural
5	Korean	12	Punk
6	Western	13	Ethnic
7	British		

the pos-neg proportion and maintaining the original data distribution.

- Random erasing [11]
  This is a data augmentation method for training the convolutional neural network. In the training process, we randomly selected a rectangle region in an image and erased its pixels with random values. Therefore, it will generate training samples with various levels of occlusion, which may reduce the risk of over-fitting.
- Random rotating and random cropping
  Since the clothing images include various of postures and scales, we randomly rotate the training images with an angle between $\pm 15°$, and randomly crop the images as input to make model robust to postures and scales.

If not clearly specified, all the experiments below are based on data argumentation by default.

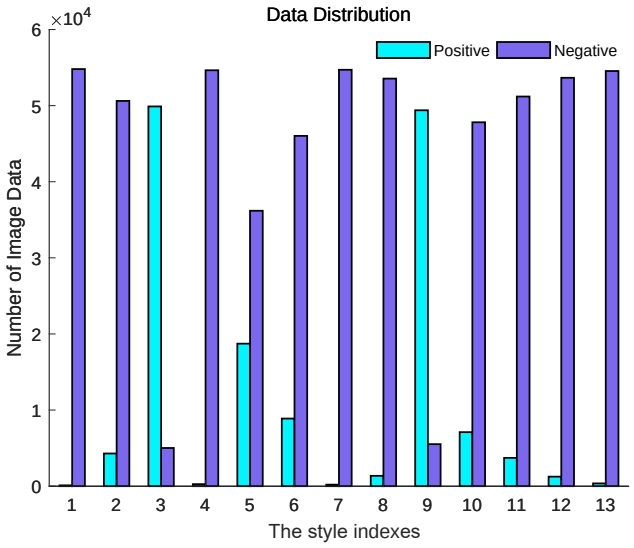

Fig. 1. Distribution of the training and validation dataset. It is obviously shown that the class distribution is seriously imbalance. Taking style index 1 as example, there are 104 positive samples while 54804 negative samples in the dataset.

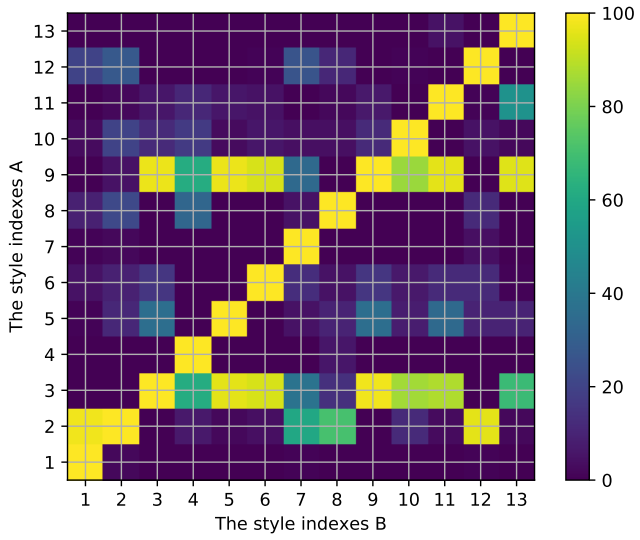

Fig. 2. The conditional probability of style index A when style index B happened. The indexes are declared in Table I. For example, if the clothes belongs to style index 1 (Sports), it may belongs to style index 2 (Casual) with a probability of 95%.

TABLE II
PERFORMANCE FOR DIFFERENT NETWORK ARCHITECTURES AND THE MODEL FUSION RESULT. FOR MOST STYLES, FUSION RESULT IMPROVED 3% ON AVERAGE.

Network	1	2	3	4	5	6	7
Resnet50	0.16	0.62	0.97	0.35	0.67	0.65	0.16
Densenet121	0.13	0.65	0.98	0.35	0.68	0.65	0.16
Inception-v4	0.14	0.64	0.97	0.35	0.67	0.66	0.17
Model Fusion	0.18	0.67	0.98	0.37	0.70	0.70	0.17

Network	8	9	10	11	12	13	Mean
Resnet50	0.42	0.97	0.63	0.75	0.42	0.33	0.55
Densenet121	0.44	0.97	0.63	0.75	0.44	0.31	0.55
Inception-v4	0.42	0.96	0.64	0.75	0.40	0.33	0,55
Model Fusion	0.45	0.97	0.64	0.77	0.45	0.33	0.57

B. Network architecture

We carried out experiments with several state-of-the-art network architectures, such as Resnet50 [12], Densenet121 [13] and Inception v4 [14]. After data argumentation processing, we trained these models individually. The experiment results (on the validation dataset) for different network architectures are listed in Table II. As we can see, Resnet50 gets better result for $Sport$ and $Japanese$, while Densenet121 performs better for $Girl$, $Punk$ and $Casual$. In addition, performance of Inception-v4 gets the medium level for majority of the classes. In addition, model fusion can further improve the performance to a new gap because of the complementation learning.

C. Transfer learning

Given a source domain $D_S = \{X_S, f_S(X)\}$ and learning task T_S, a target domain $D_T = \{X_T, f_T(X)\}$ and learning task T_T, transfer learning aims to help improve the learning of

TABLE III
PERFORMANCE OF USING TRANSFER LEARNING FOR DIFFERENT STYLES.

Approach	1	2	3	4	5	6	7
Model Fusion	0.18	0.67	0.98	0.37	0.70	0.70	0.17
Transfer Learning	0.22	0.69	0.98	0.42	0.72	0.73	0.23

Approach	8	9	10	11	12	13	Mean
Model Fusion	0.45	0.97	0.64	0.77	0.45	0.33	0.57
Transfer Learning	0.47	0.97	0.66	0.79	0.47	0.35	0.59

TABLE IV
THE MEAN F2-SCORE EVALUATION RESULT FOR OUR PROPOSED METHODS.

Data Augmentation	–	Yes	Yes	Yes	Yes
Transfer learning	–	–	Yes	Yes	Yes
Cost-Sensitive	–	–	–	Yes	Yes
Styles relationship	–	–	–	–	Yes
Mean F2-score	0.44	0.57	0.59	0.60	0.63

the target predictive function $f_T(\cdot)$ in D_T using the knowledge in D_S and T_S, where $D_S \neq D_T$, or $T_S \neq T_T$. Since there is a huge gap between ImageNet dataset [10] and JDAIfashion dataset, using ImageNet pretrained model to initialize the weights to help the style recognition may get sub-opt. Hence, we pre-train our base model on DeepFashion [15] dataset, which is a big clothing fashion dataset and have more closely data distribution with JDAIfashion dataset.

DeepFashion contains over $800,000$ diverse fashion images ranging from well-posed shop images to unconstrained consumer photos. Each image in this dataset is labeled with 50 categories, 1000 descriptive attributes, bounding box and clothing landmarks. Our pre-training task concentrates on classifying 50 clothes attributes. To get better pre-trained model, we cleaned the dataset and selected 20 attributes with more balanced data distribution. The experiment results are shown in Table3. As can be seen form the results, class 1, 2, 4, 7, 8, 10, 11, 12, 13 benefit a lot from transfer learning. On the other hand, pre-training on DeepFashion plays a minor role for class 3, 5, 6 and 9, since they already have enough positive images. Besides, to make better use of style relationship, we have also initialize the weights from the relevant models. Detailedly, we initialize the weights for class 1 using class 2, and initialize the weights for class 4 using class 8.

D. Cost-Sensitive learning

Since our evaluation criterion is F2-score (recall is more important than precision), the cost of positive sample misclassification and negative sample misclassification for F2-score differs greatly. To maximize the F2-score, we adopted cost-sensitive learning [16] for data mining. Cost-Sensitive learning is a type of learning method in data mining that takes the misclassification cost into consideration. The goal of cost-sensitive learning is to minimize the total cost. Specifically, it treats different misclassification differently. We increased the cost penalty for false negative, while kept unchanged for false positive. Hence, the cost function is tend to optimize the recall than precision. In our experiments, we only applied cost-sensitive learning to classes $1, 4, 7, 8, 11, 12, 13$. The penalty weight was set to 1.5 for false negative.

E. Styles relationship

As shown in Fig. 2, clothing styles are closely related or inversely related. For example, when an image belongs to style index 3 (Office lady), it may belongs to style index 9 (Lady) with a probability of 99%. Besides, some styles never shown up together (Japanese and Korean). It is obvious that styles relationship contains a wealth of information. We tried training a multi-label classification network sharing the lowest layers amongst all styles, sharing the higher layers for related styles. However, the contribution for the final F2-score is negligible.

The reason lies in that the multi-label classification network is unable to implement data resampling. Therefore, we analyzed the conditional probability of the given labels with Eqn.1.

$$P(A|B) = \frac{N(A \bigcap B)}{N(B)} \tag{1}$$

where $N(A \bigcap B)$ means the number of images simultaneously belonging to style A and B. $N(B)$ means the number of images belonging to style B. Since we have processed each style individually, we applied the styles relationship (conditional probability) using Eqn.2.

$$P_{final}(A) = \alpha \cdot P(A) + \sum_{i \in M} \beta_i \cdot P(A|B_i) \tag{2}$$

where α and β_i are the hyper-parameters to balance the contribution of styles relationship. $P(A)$ means the probability calculated by our convolutional neural network. M is the candidate set that our network predict to be positive ($B_i \neq A$). If the network predicts all the styles B_i as positive with confidence score 0, then the final confidence score of style A is: $P_{final}(A) = P(A)$.

F. Implementation detail

We adopted cross-validation method to evaluate the individual models and the fusion results. Specifically, we split 43926 images for training and 10981 images for validation. The submit result is evaluated by F2-score as shown in Eqn.(3).

$$F_2 = 5 \cdot \frac{precision \cdot recall}{4 \cdot precision + recall} \tag{3}$$

All the training and validation process were performed on a single Nvidia 1080 Ti GPU. The training batch size were set as big as the GPU memory can be. Firstly, we cleaned the Deepfashion dataset and pre-trained the basemodels. Then, we resampled our dataset and fine-tuned the pre-trained models with data argumentation and cost-sensitive learning strategy.

For the validation process, we fused the models and traversal searched the decision threshold. Since we processed each class individually, the 13 predictions may have some obvious logic

errors (e.g. there could not be all negatives or all positives. Because each image belongs to at least 1 style and at most 5 styles according to the statistic of training dataset). Therefore, for post processing, we apply the styles relationship method to further improve the predictions. The mean F2-score for these steps are shown in Table IV. As we can see, all the steps play an indispensable role in the final result.

CONCLUSION

In this paper, we introduce our proposed method for clothing style classification in JD AI Fashion-Challenge. Our method dealt with the problem from the aspects of data argumentation, network structure, transfer learning and cost sensitive learning. Besides, we investigated the styles relationship for the post processing. With the proposed approach, we achieved 0.6524 F2-score on the testing set, which is the 3rd place on the leader board.

REFERENCES

[1] 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*. ACM, 2013, pp. 105–112.

[2] L. Bossard, M. Dantone, C. Leistner, C. Wengert, T. Quack, and L. Van Gool, "Apparel classification with style," in *Asian conference on computer vision*. Springer, 2012, pp. 321–335.

[3] T. Xiao, T. Xia, Y. Yang, C. Huang, and X. Wang, "Learning from massive noisy labeled data for image classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2691–2699.

[4] Q. Chen, J. Huang, R. Feris, L. M. Brown, J. Dong, and S. Yan, "Deep domain adaptation for describing people based on fine-grained clothing attributes," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 5315–5324.

[5] E. Simo-Serra, S. Fidler, F. Moreno-Noguer, and R. Urtasun, "Neuroaesthetics in fashion: Modeling the perception of fashionability," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 869–877.

[6] W. Di, C. Wah, A. Bhardwaj, R. Piramuthu, and N. Sundaresan, "Style finder: Fine-grained clothing style detection and retrieval," in *Proceedings of the IEEE Conference on computer vision and pattern recognition workshops*, 2013, pp. 8–13.

[7] S. Liu, Z. Song, G. Liu, C. Xu, H. Lu, and S. Yan, "Street-to-shop: Cross-scenario clothing retrieval via parts alignment and auxiliary set," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 3330–3337.

[8] J. Huang, R. S. Feris, Q. Chen, and S. Yan, "Cross-domain image retrieval with a dual attribute-aware ranking network," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1062–1070.

[9] H. Chen, A. Gallagher, and B. Girod, "Describing clothing by semantic attributes," in *European conference on computer vision*. Springer, 2012, pp. 609–623.

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[11] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," *arXiv preprint arXiv:1708.04896*, 2017.

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[13] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks." in *CVPR*, vol. 1, no. 2, 2017, p. 3.

[14] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning." in *AAAI*, vol. 4, 2017, p. 12.

[15] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang, "Deepfashion: Powering robust clothes recognition and retrieval with rich annotations," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1096–1104.

[16] Z.-H. Zhou, "Cost-sensitive learning," in *International Conference on Modeling Decisions for Artificial Intelligence*. Springer, 2011, pp. 17–18.