






Deep Fashion Analysis with Feature Map Upsampling and Landmark-Driven Attention

Jingyuan Liu^{} and Hong Lu^{}^{}

Shanghai Key Lab of Intelligent Information Processing, School of Computer Science,
Fudan University, Shanghai, People's Republic of China
{jingyuanliu15,honglu}@fudan.edu.cn

Abstract. In this paper, we propose an attentive fashion network to address three problems of fashion analysis, namely landmark localization, category classification and attribute prediction. By utilizing a landmark prediction branch with upsampling network structure, we boost the accuracy of fashion landmark localization. With the aid of the predicted landmarks, a landmark-driven attention mechanism is proposed to help improve the precision of fashion category classification and attribute prediction. Experimental results show that our approach outperforms the state-of-the-arts on the DeepFashion dataset.

Keywords: Fashion analysis · Landmark detection · Clothing category classification · Attention mechanism · Deep learning

1 Introduction

Recent years, with the rapid growth of online commerce and fashion-related application, fashion image analysis and understanding have attracted increasing amount of attention in the community. Extensive studies have been conducted in this field, such as category classification, style or attribute prediction, fashion landmark localization, and fashion image synthesis.

In this paper, we study three core problems of fashion image analysis: landmark localization, category classification and attribute prediction. Previous works based on deep learning have shown much success in these fields [3, 6, 9–11, 16, 17]. However, most of them fail to further improve fashion analysis accuracy because of the low resolution of the predicted heatmaps after several pooling operations. It limits the prediction accuracy since fashion landmarks usually lie in the sharp corners or edges of clothes. In this paper, we address this problem by using transposed convolution to upsample the feature map. Thus, the predicted heatmaps are high-resolution and have the same size as the input fashion image, which will improve the accuracy of landmark localization.

For enhancing accuracy of category classification and attribute prediction, we also introduce a landmark-driven attention mechanism leveraging the predicted

landmark heatmap. The landmark locations and the convolutional features are combined to form a new attention map, which gives our network a flexible way to focus on the most functional parts of the clothes for category and attribute prediction with the reference to both local landmark positions and global features. Such attention mechanism magnifies the most related information for fashion analysis while filters out unrelated features, thus boosting the category and attribute prediction accuracy. Notably, our whole fashion analysis model is fully differentiable and can be trained end-to-end.

We exert comprehensive evaluations on a large-scale dataset – DeepFashion dataset [9]. Experimental results demonstrate that our fashion analysis model outperforms the state-of-the-arts.

In summary, our **contributions** are:

1. We propose a fashion analysis network: an end-to-end system that addresses category classification and attribute prediction simultaneously, via improving the resolution of heatmaps through upsampling for more accurate landmark localization.
2. We introduce a novel attention mechanism: Landmark heatmaps are used as references to generate a unified attention, so that the network has enough information to enhance or reduce features.
3. Quantitatively, we report, for the first time, our model show improvement over the state-of-the-art on landmark localization, category classification and attribute prediction.

2 Related Work

Fashion analysis has drawn increasing attention in recent years because of its various applications like clothing recognition and segmentation [5, 8, 17, 19], recommendation [4, 9, 12, 13], and fashion landmark localization [10, 16, 17]. As for landmark localization, some studies utilize regression methods, where convolutional features are directly fed into fully connected layer to fit the coordinate positions of landmarks [9, 10]. As shown in [15], this kind of regression is in a highly non-linear and complex form, thus the parameters are difficult to learn. To address this problem, some studies employ fully convolutional networks that produce a position heatmap for each landmark [16, 17] but fail to maintain the high resolution of heatmaps during the pipeline, which limits the accuracy. The closest work to our method is [16] whose fashion network is encoded with two attention mechanisms: landmark-aware attention and category-driven attention. Their algorithm was based on two fashion grammars they proposed and was tested on the Deepfashion dataset. [16] suffers from the difficulty of detect landmark in low resolution which is caused by the series of pooling operations. The main differences with our work are that: (i) In our network, we use transposed convolution upsampling to generate more accurate feature maps, which is more suitable for fashion and clothing related tasks and thus improves the accuracy

of landmark localization. (ii) Those landmark feature maps will serve as references to generate one unified attention mechanism rather than two for boosting category classification and attribute prediction.

Attention Mechanism has gained popularity in the fields of image recognition [7], image detection [18] and VQA (Visual Question Answering) [2]. Those work demonstrate the efficiency of the attention mechanism that enables the network to learn which parts in an image should be focused on to solve certain tasks. In this paper, a unified attention mechanism is proposed, which avoid of hard deterministic constraints in feature selection and helps our model achieve state-of-the-art results in visual fashion analysis tasks. Besides, in contrast to previous attention-based fashion models [16] with two separate attention branch, our attention has combined those two into one unified branch act as soft constraints and can be learned more easily from data.

3 Methodology

3.1 Problem Formulation

Given a fashion image I , our goal is to predict the landmark position L , category label C and attribute vector A . $L = \{(x_1, y_1), (x_2, y_2), \dots, (x_{n_l}, y_{n_l})\}$, where x_i and y_i is the coordinate position for each landmark, and n_l is the total number of landmarks. In this paper, we utilize $n_l = 8$ since there are 8 annotated landmarks in the DeepFashion dataset, defined as left/right collar end, left/right sleeve end, left/right waistline, and left/right hem. Category label C satisfies $0 \leq C \leq n_c - 1$, where n_c is the number of all categories. Attribute prediction is treated as a multi-label problem. The label vector $A = (a_1, a_2, \dots, a_{n_a})$, $a_i \in \{0, 1\}$, where n_a is the number of attributes. $a_i = 1$ indicates that the fashion image has the i th attribute.

3.2 Network Architecture

Our main network architecture is based on the VGG-16 networks [14], as shown in Fig. 1. First, we resize the original image to 224×224 . Initial convolutional operations are the same as the VGG-16 networks. We add two new branches after the conv4_3 layer. One is the landmark localization branch, the other is a attention branch. Detailed description is as follows.

Landmark Localization Branch. We use several transposed convolution to produce a high-resolution landmark heatmap. In particular, we first utilize $64 \ 1 \times 1$ filters to convert the input feature map to $28 \times 28 \times 64$. Then two 3×3 convolutions and one 4×4 transposed convolution are employed to upsample the feature to $56 \times 56 \times 64$. The 3×3 convolution also has a padding of 1 so it does not change the size of the feature map. The stride and padding of the transposed convolution are 2 and 1, respectively. Thus it can upsample the feature map twice its size. In the following, we use the same structure of two convolution and one

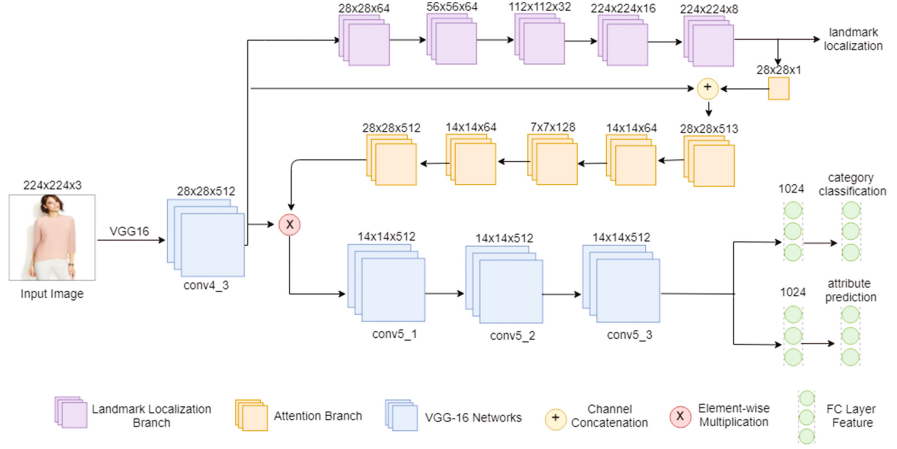


Fig. 1. Our network architecture. The structure is mainly based on the VGG-16 networks, and we add a landmark localization branch and a attention branch. The landmark localization branch produces heatmaps for all landmarks in the original resolution. Predicted heatmaps and the conv4_3 features are then fed into the attention branch, the result of which will be used to gate or magnify the conv4_3 features.

transposed convolution to upsample the feature map to $224 \times 224 \times 16$. Finally, a 1×1 convolution is employed to produce the $224 \times 224 \times 8$ heatmap, denoted as $M' \in R^{224 \times 224 \times 8}$. The ground truth of the landmark heatmap M is generated by adding a Gaussian filter at the corresponding landmark position. We use L_2 loss to train the landmark localization branch: $l_{landmark} = \frac{1}{N} \|M_{ijk} - M'_{ijk}\|_2^2$, where N is the total number of array elements. By producing heatmaps with the same size as the original image, the landmark localization branch is capable of predicting landmark positions with higher accuracy.

Attention Branch. The attention branch takes the concatenation of the conv4_3 feature and the landmark information as its input. The landmark information is formulated as $M_{ij}^{info} = \max\{M''_{ij1}, M''_{ij2}, \dots, M''_{ij8}\}$, where $M'' \in R^{28 \times 28 \times 8}$ is the bilinear downsample of M' . It describes the overall landmark positions of the fashion image. We first use one 1×1 convolution to convert the input feature map to $28 \times 28 \times 32$. Then two convolutional layers are employed to squeeze the feature to $7 \times 7 \times 28$. Each layer has one 3×3 convolution and one 2×2 max pooling. Finally we use two transposed convolutions to get the output A , $A \in R^{28 \times 28 \times 512}$. The activation function in the last layer is sigmoid function thus we have $0 < A_{ijk} < 1$.

We denote the conv4_3 feature as F . The output of the attention branch is used to modify F by making $F_{new} = (\frac{1}{2} + A) \circ F$, where \circ stands for element-wise multiplication. We add A by $\frac{1}{2}$ thus the element will be in the range $(\frac{1}{2}, \frac{3}{2})$. Numbers less than 1 will filter out unrelated features, while numbers greater than 1 will magnify important features. The following is the same as the

VGG-16 network. We use two branches to predict category and attribute in the last. The loss for category and attribute prediction is the standard cross entropy loss.

4 Results

We evaluate our model on the DeepFashion dataset [9]. In particular, we use the Category and Attribute Prediction Benchmark. It offers 289,222 fashion images with annotations of 8 kinds of landmarks, 46 categories, 1,000 attributes. Each image has a bounding box for the clothes. The attributes are divided into 5 subgroups: texture, fabric, part, shape and style. We follow the same settings in [9, 11]. We adopt normalized distance as the metrics for landmark localization. Top-k accuracy and top-k recall are used to evaluate the performance of category classification and attribute prediction, respectively.

Table 1. Experimental results on the DeepFashion dataset for landmark localization. The best results are marked in **bold**.

Methods	L.Collar	R.Collar	L.Sleeve	R.Sleeve	L.Waistline	R.Waistline	L.Hem	R.Hem	Avg
FashionNet	0.0854	0.0902	0.0973	0.0935	0.0854	0.0845	0.0812	0.0823	0.0872
DFA	0.0628	0.0637	0.0658	0.0621	0.0726	0.0702	0.0658	0.0663	0.0660
DLAN	0.0570	0.0611	0.0672	0.0647	0.0703	0.0694	0.0624	0.0627	0.0643
Wang et al.	0.0415	0.0404	0.0496	0.0449	0.0502	0.0523	0.0537	0.0551	0.0484
Ours	0.0332	0.0346	0.0487	0.0519	0.0422	0.0429	0.0620	0.0639	0.0474

Table 2. Experimental results on the DeepFashion dataset for category classification and attribute prediction. The best results are marked in **bold**.

Methods	Category		Texture		Fabric		Shape		Part		Style		All	
	Top-3	Top-5	Top-3	Top-5	Top-3	Top-5	Top3	Top-5	Top-3	Top-5	Top-3	Top-5	Top-3	Top-5
WTBI [1]	43.73	66.26	24.21	32.65	25.38	36.06	23.39	31.26	26.31	33.24	49.85	58.68	27.46	35.37
DARN [6]	59.48	79.58	36.15	48.15	36.64	48.52	35.89	46.93	39.17	50.14	66.11	71.36	42.35	51.95
FashionNet [9]	82.58	90.17	37.46	49.52	39.30	49.84	39.47	48.59	44.13	54.02	66.43	73.16	45.52	54.61
Lu et al. [11]	86.72	92.51	-	-	-	-	-	-	-	-	-	-	-	-
Corbiere et al. [3]	86.30	92.8	53.60	63.20	39.10	48.80	50.10	59.50	38.80	48.90	30.50	38.30	23.10	30.40
Wang et al. [16]	90.99	95.78	50.31	65.48	40.31	48.23	53.32	61.05	40.65	56.32	68.70	74.25	51.53	60.95
Ours	91.16	96.12	56.17	65.83	43.20	53.52	58.28	67.80	46.97	57.42	68.82	74.13	54.69	63.74

For landmark localization, we compare our method with 4 recent deep learning models [9, 10, 16, 17]. As shown in Table 1, our model is more accurate and achieves state-of-the-art at 0.0474. For category and attribute prediction, our method is compared with 6 recent top-performing models [1, 3, 6, 9, 11, 16]. With the aid of the accurate landmark-driven attention, our model outperforms all the competitors, as shown in Table 2.

We also visualize what the attention branch has learned as show in Fig. 2 that it makes the network focus on the related information and ignore the useless information.



Fig. 2. Attention map visualization

5 Conclusion

In this paper, we design a novel attention-aware model for deep learning-based fashion analysis, leading to a fully differentiable network that can be trained end-to-end. Our model utilizes convolutional upsampling to produce more accurate landmark heatmap. We further introduce an attention mechanism, which takes advantage of the predicted landmark locations for improving the accuracy of category classification and attribute prediction. The experimental results on three benchmarks of the DeepFashion dataset has demonstrated the superior performance of our model, which achieves the state-of-the-art landmark localization, category classification and attribute prediction compared to recent methods.

References

1. Chen, H., Gallagher, A., Girod, B.: Describing clothing by semantic attributes. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012. LNCS, vol. 7574, pp. 609–623. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33712-3_44
2. Chen, K., Wang, J., Chen, L.C., Gao, H., Xu, W., Nevatia, R.: ABC-CNN: an attention based convolutional neural network for visual question answering. arXiv preprint [arXiv:1511.05960](https://arxiv.org/abs/1511.05960) (2015)
3. Corbiere, C., Ben-Younes, H., Ramé, A., Ollion, C.: Leveraging weakly annotated data for fashion image retrieval and label prediction. arXiv preprint [arXiv:1709.09426](https://arxiv.org/abs/1709.09426) (2017)
4. Han, X., Wu, Z., Jiang, Y.G., Davis, L.S.: Learning fashion compatibility with bidirectional LSTMs. In: Proceedings of the 2017 ACM on Multimedia Conference, pp. 1078–1086. ACM (2017)
5. Hidayati, S.C., You, C.W., Cheng, W.H., Hua, K.L.: Learning and recognition of clothing genres from full-body images. *IEEE Trans. Cybern.* **48**(5), 1647–1659 (2018)
6. Huang, J., Feris, R.S., Chen, Q., Yan, S.: Cross-domain image retrieval with a dual attribute-aware ranking network. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1062–1070 (2015)
7. Jaderberg, M., Simonyan, K., Zisserman, A., et al.: Spatial transformer networks. In: Advances in Neural Information Processing Systems, pp. 2017–2025 (2015)

8. Kalantidis, Y., Kennedy, L., Li, L.J.: 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, pp. 105–112. ACM (2013)
9. Liu, Z., Luo, P., Qiu, S., Wang, X., Tang, X.: Deepfashion: powering robust clothes recognition and retrieval with rich annotations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1096–1104 (2016)
10. Liu, Z., Yan, S., Luo, P., Wang, X., Tang, X.: Fashion landmark detection in the wild. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9906, pp. 229–245. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46475-6_15
11. Lu, Y., Kumar, A., Zhai, S., Cheng, Y., Javidi, T., Feris, R.: Fully-adaptive feature sharing in multi-task networks with applications in person attribute classification. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1131–1140 (2017)
12. Ma, Y., Jia, J., Zhou, S., Fu, J., Liu, Y., Tong, Z.: Towards better understanding the clothing fashion styles: a multimodal deep learning approach. In: AAAI, pp. 38–44 (2017)
13. de Melo, E.V., Nogueira, E.A., Guliato, D.: Content-based filtering enhanced by human visual attention applied to clothing recommendation. In: 2015 IEEE 27th International Conference on Tools with Artificial Intelligence (ICTAI), pp. 644–651. IEEE (2015)
14. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint [arXiv:1409.1556](https://arxiv.org/abs/1409.1556) (2014)
15. Tompson, J.J., Jain, A., LeCun, Y., Bregler, C.: Joint training of a convolutional network and a graphical model for human pose estimation. In: Advances in Neural Information Processing Systems, pp. 1799–1807 (2014)
16. Wang, W., Xu, Y., Shen, J., Zhu, S.C.: Attentive fashion grammar network for fashion landmark detection and clothing category classification. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 4271–4280 (2018)
17. Yan, S., Liu, Z., Luo, P., Qiu, S., Wang, X., Tang, X.: Unconstrained fashion landmark detection via hierarchical recurrent transformer networks. In: Proceedings of the 2017 ACM on Multimedia Conference, pp. 172–180. ACM (2017)
18. Yan, Y., et al.: Unsupervised image saliency detection with gestalt-laws guided optimization and visual attention based refinement. *Pattern Recognit.* **79**, 65–78 (2018)
19. Yang, W., Luo, P., Lin, L.: Clothing co-parsing by joint image segmentation and labeling. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3182–3189 (2014)