FashionSearchNet: Fashion Search with Attribute Manipulation

Kenan E. Ak[1,2(✉)], Ashraf A. Kassim[1], Joo Hwee Lim[2], and Jo Yew Tham[3]

[1] National University of Singapore, Singapore, Singapore
emir.ak@u.nus.edu, ashraf@nus.edu.sg
[2] Institute for Infocomm Research, A*STAR, Singapore, Singapore
joohwee@i2r.a-star.edu.sg
[3] ESP xMedia Pte. Ltd., Singapore, Singapore
thamjy@espxmedia.com

Abstract. The focus of this paper is on retrieval of fashion images after manipulating attributes of the query images. This task is particularly useful in search scenarios where the user is interested in small variations of an image, i.e., replacing the mandarin collar with a button-down. Keeping the desired attributes of the query image while manipulating its other attributes is a challenging problem which is accomplished by our proposed network called FashionSearchNet. FashionSearchNet is able to learn attribute specific representations by leveraging on weakly-supervised localization. The localization module is used to ignore the unrelated features of attributes in the feature map, thus improve the similarity learning. Experiments conducted on two recent fashion datasets show that FashionSearchNet outperforms the other state-of-the-art fashion search techniques.

Keywords: CNNs · Fashion retrieval · Similarity learning · Attribute localization

1 Introduction

Recently, there has been a huge interest in fashion-related research. The appearance of deep learning based techniques have boosted the interest in the following areas: fashion/attribute recognition [1–3], fashion retrieval [3–5], attribute discovery [6,7], fashion/human parsing [8–10], recommentation [11–13] and generative adversarial networks [14,15].

Image retrieval is a crucial task as it can significantly lower the time that is required to find the desired item. Most research in image-based fashion retrieval focuses on finding same/similar images in cross-domains [3–5]. In this work, we focus on a search scenario called "fashion search with attribute manipulation" which is illustrated in Fig. 1 where the user would like to search for a similar item as the query image but with a short sleeve instead i.e., the sleeve attribute "sleeveless" needs to be replaced with "short". Until recently [16,17] there wasn't

L. Leal-Taixé and S. Roth (Eds.): ECCV 2018 Workshops, LNCS 11131, pp. 45–53, 2019.
https://doi.org/10.1007/978-3-030-11015-4_6

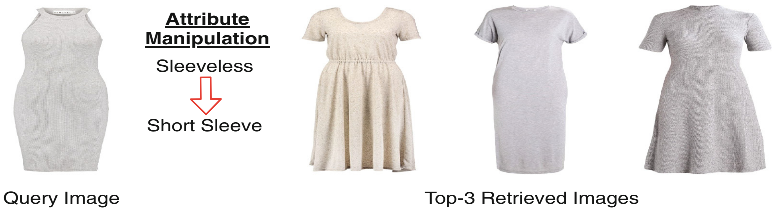


Fig. 1. Given a query image, the user manipulates the sleeve attribute. The proposed FashionSearchNet utilizes it's feature representations to find images similar to the query image while manipulating the sleeveless attribute.

any published work which allowed to make such modifications on the query image and conduct the fashion retrieval. Our recently introduced FashionSearchNet [17], aims to solve this problem. Different than our prior work [17], we briefly explain the method and share additional experiments.

The proposed FashionSearchNet initially focuses on learning attribute similarities by leveraging weakly supervised localization, i.e., no information is given known about the location of the attributes. The localization module is based on global average pooling (GAP) layer and it is used to generate a several attribute activation maps (AAMs). AAMs are used to extract attribute boundaries and feed attribute relevant features to a set of fully connected layers where the similarity learning is conducted. By using a triplet-based similarity learning process, attributes which share the same value are made to be similar to each other in the feature space. Lastly, another learning mechanism is employed to combine attribute representations into a global representation and used in the fashion search. This mechanism helps the network decide which attribute representation should have higher importance depending on the attribute that is being manipulated.

2 Related Work

Attribute Recognition. Recently, Chen et al. [18] used fine-grained clothing attributes to describe people with a deep domain adaptation network. Mix and match method [19] proposed a joint model for clothing and attribute recognition. Abdulnabi et al. [20] showed that a multi-task CNN model can be used to predict attributes of fashion images. More recently, an approach which combines localization and classification for apparel recognition is proposed by Song et al. [21] but it requires supervision on object boundaries. FashionSearchNet uses the classification loss to learn better attribute representations and in doing so it contributes to its localization ability.

Image Retrieval. Image-based fashion retrieval is an important topic, however, it is mostly based on retrieving the similar/same images [3,5,21,22] which limits the user interaction. In terms of clothing recommendation, both methods proposed in [23,24] use fashion images from a personal closet to recommend outfits. However, what if the user does not like a specific attribute of the retrieved image? The idea of fashion search with attribute manipulation comes from Zhao et al. [16] where a CNN is trained (AMNet) by combining the query image with the desired attribute. Different than FashionSearchNet, AMNet [16] does not explore the spatial aspect of attributes, therefore, lacks the ability to conduct the similarity learning on attribute-level. Similar to our method, Singh et al. [25] proposed an end-to-end localization and ranking for relative attributes. However, the fact that they train a new model for each attribute make it infeasible for fashion images.

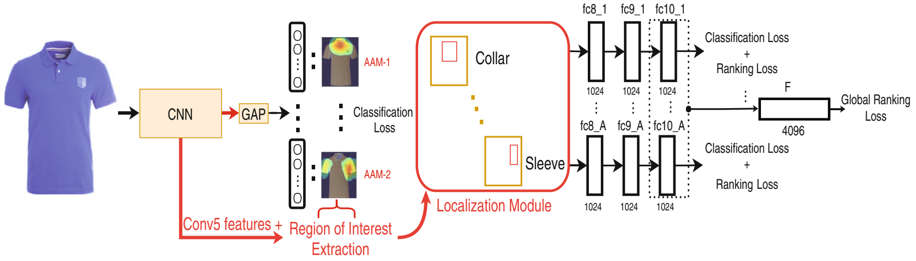


Fig. 2. FashionSearchNet. From each attribute activation maps (AAMs), attribute-relevant convolutional features are extracted and pooled to a set of fully connected layers. Both classification and ranking losses are used for similarity learning. Lastly, attribute representations ($fc_{10\_1}, fc_{10\_2}, ...$) are combined into a single representation to conduct fashion search.

3 FashionSearchNet

The proposed FashionSearchNet is based on AlexNet [26] architecture with some modifications. As suggested by [27], 2 more convolutional layers are added after conv5 layer. GAP layer is used to generate AAMs which are then used to extract region of interests. In order to learn attribute similarity, a set of fully connected layers are used and jointly trained with classification and ranking loss. Lastly, a mechanism is employed to help the network decide which attribute representation should have higher importance depending on the attribute manipulation with the global ranking loss.

Attribute Activation Maps. We denote the output of global average pooling as x_I, where I corresponds to the input image. Initially, the network is trained with multi-attribute classification loss function to have reliable AAMs:

$$L_C = -\sum_{I=1}^{N}\sum_{a=1}^{A} log(p(g_{Ia}|x_I w_a)) \tag{1}$$

where g_{Ia} represents the ground truth of the $a'th$ attribute of the $I'th$ image. $x_I w_a$[1] calculates weighted linear combination of x_I for attribute a, N is the number of training examples and A is the number of attributes. Next, we define $M_{a_c}(I, i, j)$ as an AAM for class c of an attribute a as follows:

$$M_{a_c}(I, i, j) = \sum_k w_{a_{(k,c)}} conv7_k(I, i, j) \tag{2}$$

where $w_{a_{(k,c)}}$ is the weight variable of attribute a associated with $k'th$ feature map of class c. c is determined from the class that maximizes the classification confidence. After extracting heatmaps from M_{a_c}, each ROI is estimated with a simple hard threshold technique. Following the technique from [27], the pixel values that are above 20% of the maximum value are segmented and a bounding box is estimated which covers the largest segmented region.

Attribute Similarity Learning. By using AAMs, the localization ability is added to the network. Next, ROI pooling layer [28] is used to pass attribute specific features from convolutional layer into a set of attribute-specific fully connected layers. Looking at the last AAM in Fig. 2 which corresponds to the sleeve length attribute, it is evident that the network can successfully localize. This process ignores unrelated features from other regions such as torso or collar.

At the end of fully connected layers, triplet ranking constraints are imposed to enable attribute similarity learning. Inspired by [29], FashionSearchNet uses the soft-triplet ranking function is utilized which normalizes the distances to the range of (0,1) with a softmax function and formulated as follows:

$$d^+(h(\hat{I}),h(I^+),h(I^-)) = \frac{exp(||h(\hat{I})-h(I^+)||_2)}{exp(||h(\hat{I})-h(I^+)||_2)+exp(||h(\hat{I})-h(I^-)||_2)} \tag{3}$$

$$d^-(h(\hat{I}),h(I^+),h(I^-)) = \frac{exp(||h(\hat{I})-h(I^-)||_2)}{exp(||h(\hat{I})-h(I^+)||_2)+exp(||h(\hat{I})-h(I^-)||_2)} \tag{4}$$

Given $||d^+, d^- - 1||_2^2 = d^+$ and $h = fc_{10\_a}$ the ranking loss function becomes:

$$L_T = \sum_{I=1}^{N}\sum_{a=1}^{A} d^+(fc_{10\_a}(\hat{I}), fc_{10\_a}(I^+), fc_{10\_a}(I^-)) \tag{5}$$

where A is the number of the fully connected layers. The given function aims to minimize $||fc_{10\_a}(\hat{I}), fc_{10\_a}(I^+)||_2$ and maximize $||fc_{10\_a}(\hat{I}), fc_{10\_a}(I^-)||$. We pick triplets from the same mini-batch and the rule is: $\hat{I}$ and I^+ must share the same label while I^- is chosen randomly.

Attribute Manipulation. After the network is trained for the attribute similarities, we extract features from the training images with the same attribute value and average them. These averaged features are stored in a matrix where

[1] The dimensions of w_a is [number of feature maps by number of classes associated with a].

each attribute value corresponds to a specific representation so that the "undesired" attribute representations can directly be replaced.

Learning Global Representation. It is evident that some attribute representations should have higher importance when conducting the fashion search. To apply this idea to FashionSearchNet, we use a weight parameter w_{a^*} to reduce the concatenated feature length to 4096. Also, we add A number of parameters λ_{a,a^*} to learn attribute importance. The training is conducted with the following function L_G which we call global ranking loss for a given attribute manipulation a^*:

$$F(I, a^*) = [fc_{10\_1}(I)\lambda_{1,a^*}, ..., fc_{10\_A}(I)\lambda_{a,a^*}]w_{a^*} \tag{6}$$

$$L_G = \sum_{I=1}^{N}\sum_{a^*=1}^{A} d^+(F(\hat{I}, a^*), F(I^+, a^*), F(I^-, a^*)) \tag{7}$$

The rule for picking triplets for the global ranking loss is: $\hat{I}$ and I^+ must be identical in terms of attributes after the attribute manipulation while I^- is chosen randomly.

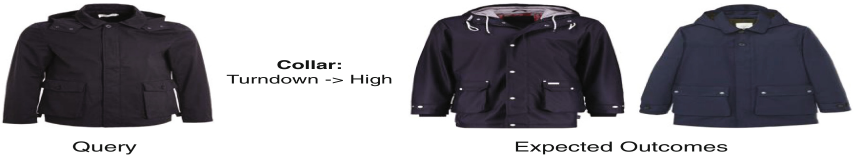


Fig. 3. An example of a successful retrieval. Given a query image and attribute manipulation, there are only 2 images in Shopping100k which matches all desired attributes.

4 Experiments

Datasets. We use DeepFashion [2] and Shopping100k [30] datasets to conduct experiments as they contain several attributes. For Shopping100k dataset [30], around 80k images are used for training, 20k images are reserved for the retrieval gallery and 2k images are served as the queries. For DeepFashion dataset [2], we choose to use Attribute Prediction subset and choose the following attributes: category, shape, texture. 90k images are used to train the network, 21k images are reserved for retrieval gallery and 2k images are served as the queries.

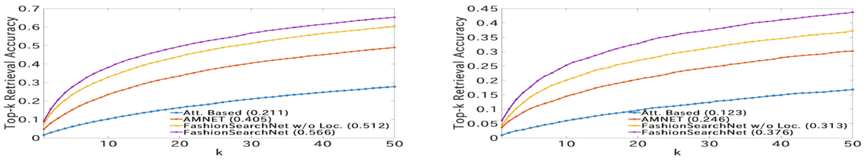
Evaluation Metric. We use Top-K retrieval accuracy to perform experiments. Given a query image and an attribute manipulation, the search algorithm finds the "best K" image matches i.e., "Top-K" matches. If the retrieved image matches with all desired attributes after attribute manipulation, it corresponds to a hit (1) or it is a miss (0). We provide a search example in Fig. 3. If a system is able to find one of the expected outcomes in k'th match, Top-k would be equal to 1.

Competing Methods: We compare the performance of FashionSearchNet with "AMNet" [16], an attribute-based method denoted as "Att. Based" which uses AlexNet [26] to predict attributes of query images and substitute the unwanted attributes. To see the effect of AAMs, we remove AAMs from FashionSearchNet and train a network called "FashionSearchNet w/o Loc." in the same fashion but without the extra fully connected layers.

4.1 Fashiong Search with Attribute Manipulation

Average Top-K retrieval accuracy results are presented in Fig. 3 for Shopping100k (a) and DeepFashion (b) datasets. For both datasets, FashionSearchNet achieves the best performance, giving 56.6% and 37.6% Top-30 accuracy respectively. Compared to AMNet, we manage to achieve 16% and 13% improvement. AMNet, on the other hand, is a decent method compared to the basic attribute-based system as AMNet performs 19% and 12% better. The attribute-based method is not a good method as it relies on predicting each of the attributes correctly which is not an easy task. Comparing AMNet [16] with FashionSearchNet, there are two drawbacks. Firstly, AMNet ignores the localization aspect of attributes. Secondly, FashionSearchNet conducts similarity learning on attribute-level which is not present in AMNet. Removing the localization ability results in 5.4% and 6.3% performance loss (Fig. 4).

We report the number possible attribute manipulation operations in Table 1. For each attribute, this number varies because for some images as there is no possible attribute manipulation for a certain attribute. Next, we report results depending on an attribute that is being manipulated in Table 2 where FashionSearchNet shows a good performance. For Shopping100k dataset, our method



(a) Shopping100k

(b) DeepFashion

Fig. 4. Top-K Retrieval Accuracies for *search by query and attribute manipulation* experiments using (a) Shopping100 k and (b) DeepFashion datasets. The number in the parentheses corresponds to the Top-30 retrieval accuracy.

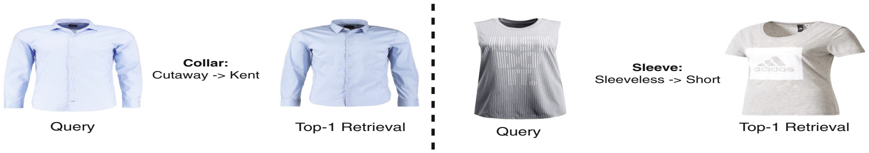


Fig. 5. Given 2 query images, the user conduct 2 attribute manipulation (collar, sleeve) where the proposed method successfully retrieves images.

Table 1. The number of retrieval instances Shopping100 k and DeepFashion datasets for each attribute.

	Shopping100k											DeepFashion		
	Category	Color	Collar	Fabric	Fastening	Fit	Gender	Neckline	Pocket	Pattern	Sleeve	Sport	Category	Shape	Texture
Count	405	127	1169	544	126	670	313	316	111	835	422	239	1849	1823	1792

Table 2. Top-30 retrieval accuracy for Shopping100 k and DeepFashion Datasets.

	Shopping100k											DeepFashion		
	Category	Color	Collar	Fabric	Fastening	Fit	Gender	Neckline	Pocket	Pattern	Sleeve	Sport	Category	Shape	Texture
Attribute-based	0.095	0.175	0.195	0.320	0.181	0.342	0.089	0.137	0.225	0.299	0.101	0.192	0.118	0.138	0.115
AMNet	0.223	0.433	0.477	0.258	0.248	0.357	0.326	0.350	0.434	0.388	0.360	0.315	0.218	0.249	0.273
FashionSearchNet w/o Loc.	0.339	0.583	0.599	0.330	0.336	0.452	**0.559**	0.494	0.477	0.552	0.524	0.348	0.202	**0.409**	0.330
FashionSearchNet	**0.395**	**0.649**	**0.642**	**0.401**	**0.423**	**0.519**	0.527	**0.532**	**0.531**	**0.575**	**0.640**	**0.436**	**0.380**	**0.409**	**0.338**

outperforms the other methods consistently. We provide 2 examples in Fig. 5 to show the success of the proposed method. For the texture and shape attributes in DeepFashion dataset, the performance is similar to not having the attribute localization module but for the category attribute, AAMs boost the retrieval accuracy. It is especially important to have the localization ability for the category attribute as there might be 2 clothing pieces (pants and shirt) on the wearer. We identify that DeepFashion dataset have many important missing attributes such as sleeve length, collar type etc. which could benefit from the proposed localization method vastly.

5 Conclusion

This paper shows FashionSearchNet's good localization ability enables it to identify the most relevant regions and is able to generate powerful attribute representations. Additionally, FashionSearchNet utilizes a mechanism which helps it to decide which attribute representation should have higher importance depending on the attribute that is being manipulated.

References

1. Dong, Q., Gong, S., Zhu, X.: Multi-task curriculum transfer deep learning of clothing attributes. In: 2017 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 520–529 (2017)
2. Liu, Z., Luo, P., Qiu, S., Wang, X., Tang, X.: Deepfashion: powering robust clothes recognition and retrieval with rich annotations. In: IEEE Conference Computer Vision Pattern Recognition (CVPR), pp. 1096–1104 (2016)
3. Simo-Serra, E., Ishikawa, H.: Fashion style in 128 floats: joint ranking and classification using weak data for feature extraction. In: IEEE Conference Computer Vision Pattern Recognition (CVPR), pp. 298–307 (2016)
4. 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 (2017)

5. Hadi Kiapour, M., Han, X., Lazebnik, S., Berg, A.C., Berg, T.L.: Where to buy it: matching street clothing photos in online shops. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 3343–3351 (2015)
6. Han, X., et al.: Automatic spatially-aware fashion concept discovery. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1463–1471 (2017)
7. Vittayakorn, S., Umeda, T., Murasaki, K., Sudo, K., Okatani, T., Yamaguchi, K.: Automatic attribute discovery with neural activations. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9908, pp. 252–268. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46493-0_16
8. Liang, X., et al.: Human parsing with contextualized convolutional neural network. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 1386–1394 (2015)
9. Simo-Serra, E., Fidler, S., Moreno-Noguer, F., Urtasun, R.: A high performance CRF model for clothes parsing. In: Cremers, D., Reid, I., Saito, H., Yang, M.-H. (eds.) ACCV 2014. LNCS, vol. 9005, pp. 64–81. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-16811-1_5
10. Yamaguchi, K., Hadi Kiapour, M., Berg, T.L.: Paper doll parsing: retrieving similar styles to parse clothing items. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 3519–3526 (2013)
11. Al-Halah, Z., Stiefelhagen, R., Grauman, K.: Fashion forward: Forecasting visual style in fashion. arXiv preprint arXiv:1705.06394 (2017)
12. Han, X., Wu, Z., Jiang, Y.G., Davis, L.S.: Learning fashion compatibility with bidirectional lstms. arXiv preprint arXiv:1707.05691 (2017)
13. Simo-Serra, E., Fidler, S., Moreno-Noguer, F., Urtasun, R.: Neuroaesthetics in fashion: modeling the perception of fashionability. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 869–877 (2015)
14. Zhu, S., Fidler, S., Urtasun, R., Lin, D., Loy, C.C.: Be your own prada: fashion synthesis with structural coherence. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 1689–1697. IEEE (2017)
15. Han, X., Wu, Z., Wu, Z., Yu, R., Davis, L.S.: Viton: an image-based virtual try-on network. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2018
16. Zhao, B., Feng, J., Wu, X., Yan, S.: Memory-augmented attribute manipulation networks for interactive fashion search. In: IEEE Conference Computer Vision Pattern Recognition (CVPR) (2017)
17. Ak, K.E., Kassim, A.A., Lim, J.H., Tham, J.Y.: Learning attribute representations with localization for flexible fashion search. In: IEEE Conference Computer Vision Pattern Recognition, CVPR, pp. 7708–7717 (2018)
18. Chen, Q., Huang, J., Feris, R., Brown, L.M., Dong, J., Yan, S.: Deep domain adaptation for describing people based on fine-grained clothing attributes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, vol. 7, pp. 5315–5324 (2015)
19. Yamaguchi, K., Okatani, T., Sudo, K., Murasaki, K., Taniguchi, Y.: Mix and match: joint model for clothing and attribute recognition. In: BMVC, p. 51–1 (2015)
20. Abdulnabi, A.H., Wang, G., Lu, J., Jia, K.: Multi-task CNN model for attribute prediction. IEEE Trans. Multimedia **17**(11), 1949–1959 (2015)
21. Song, Y., Li, Y., Wu, B., Chen, C.Y., Zhang, X., Adam, H.: Learning unified embedding for apparel recognition. arXiv preprint arXiv:1707.05929 (2017)

22. 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 (ICCV), pp. 1062–1070 (2015)
23. Hsiao, W.L., Grauman, K.: Creating capsule wardrobes from fashion images. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2018
24. Tangseng, P., Yamaguchi, K., Okatani, T.: Recommending outfits from personal closet. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2275–2279 (2017)
25. Singh, K.K., Lee, Y.J.: End-to-end localization and ranking for relative attributes. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9910, pp. 753–769. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46466-4_45
26. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems, pp. 1097–1105 (2012)
27. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2921–2929 (2016)
28. Girshick, R.: Fast R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1440–1448 (2015)
29. Hoffer, E., Ailon, N.: Deep metric learning using triplet network. In: Feragen, A., Pelillo, M., Loog, M. (eds.) SIMBAD 2015. LNCS, vol. 9370, pp. 84–92. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24261-3_7
30. Ak, K.E., Lim, J.H., Tham, J.Y., Kassim, A.A.: Efficient multi-attribute similarity learning towards attribute-based fashion search. In: 2017 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1671–1679. IEEE (2018)