Study on Fashion Image Retrieval Methods for Efficient Fashion Visual Search

Sanghyuk Park, Minchul Shin, Sungho Ham, Seungkwon Choe, and Yoohoon Kang
Clova Vision, NAVER Corporation, Republic of Korea
{shine0624, min.stellastra, ham.sungho, seungkwon.choe, neonoid}@gmail.com

Abstract

Fashion image retrieval (FIR) is a challenging task, which requires searching for exact items accurately from massive collections of fashion products based on a query image. Despite recent advances, FIR still has limitations for application to real-world visual searches. The main reason for this is not only the trade-off between the model complexity and performance, but also the common nature of fashion images captured under uncontrolled circumstances (e.g. varying viewpoints and lighting conditions). In particular, fashion images are vulnerable to shape deformations and suffer from inconsistency between the user's query images and refined product images. Moreover, multiple fashion objects can be present simultaneously within a single image. In this paper, we considered an FIR method that is optimized for the fashion domain. We investigated training strategies and deep models to improve the retrieval performance. The experimental results on three benchmarks from DeepFashion [20] dataset show that considered methods could achieve the significant improvements compared to the previous FIR methods.

1. Introduction

Fashion image retrieval (FIR) is the task of retrieving exact and relevant fashion images that are similar to the query image and implicitly reflect the needs of the user. FIR has an important role to play concerning the growing demands for online shopping, fashion recognition, and web-based recommendations. Over the decades, there has been extensive advances in fashion related tasks as following: fashion image classification/recognition [1, 4, 14, 23, 26], fashion image retrieval [3, 5, 11, 18, 20, 24, 28], and fashion recommendation [6, 9, 17, 19, 22].

Despite recent advances, previous FIR methods still face fundamental issues when applied to real-world visual search systems, for several reasons described as follows. First, fashion images are generally composed of multiple fashion items, which are present simultaneously, and they also exhibit large variations in viewpoint and style. Second, fashion images are vulnerable to shape deformations and occlusions, depending on the environment in which they are captured (*e.g.,* product shot vs. user's street shot). Therefore, undesired search results may be provided to users.

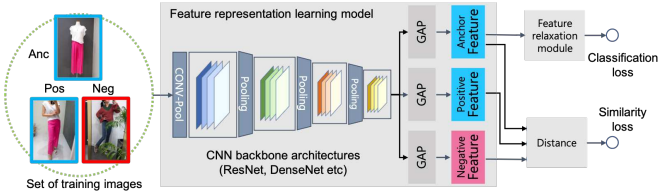

Figure 1. Considered framework for fashion image retrieval.

To address the aforementioned issues, recent FIR methods have employed advanced techniques in deep learning, such as the use of deeper architectures, attention mechanisms [13, 25, 29], and attribute modules [2, 4, 11, 12, 21]. However, we found that improvements did not solely come from complex deep learning architectures, with training strategies [27], the selected loss function, data augmentation, and structural refinement also playing important roles.

In this paper, we investigated an effective approach to training an FIR model, based on careful consideration of the training strategy and loss function. We examined structural refinements for an efficient FIR method with a loss combination, and we evaluated the performance in terms of the category classification and instance retrieval. Our empirical experiments demonstrate that the suitable selection of the learning strategy and loss function can lead to a significant improvement in the accuracy.

2. Methods

To investigate FIR methods that are optimized for the fashion domain, we selected four different CNN backbones: DenseNet121 [10], ResNet50 [7], SEResNet50 [8], and SEResNeXt50 [8]. These baseline models have been confirmed to achieve notable performances in various vision tasks, with a reasonable degree of model complexity. In this paper, we modified the final fully connected (FC) layer of each model, and adopted two loss paths and an additional feature relaxation module, which is composed of a conv layer, Relu, dropout, and final conv layer, as shown in Fig. 1.

For FIR, we employed two representative information types that can be utilized for training: category-level label and instance-level label. The category information can be utilized as a classification label for the representation of fashion collections, whilst the instance information can be utilized as a unique label for each fashion item separately. Inspired by recent multi-task learning techniques, we considered two types of loss path to train the baseline models. One is the classification loss, to encourage the learned feature representation to be discriminative among various fashion categories, and the other is the similarity loss, to learn better retrieval feature representations among diverse fashion instances.

To relieve the different aspects of feature embedding spaces by the classification loss and the similarity loss, we adopted a feature relaxation module to adjust feature distributions in classification loss path. For the FIR model training, we used the cross-entropy loss for the classification loss and the triplet loss for the similarity loss as following:

$$L_{all} = -\frac{\alpha}{N}\sum_{i}^{N}\sum_{j}^{C} y_a^{ij} \log f_j(x_a^i;\theta) \quad (1)$$
$$-\frac{\beta}{N}\sum_{i}^{N}[\delta + \left\|f(x_a^i) - f(x_+^i)\right\|_2^2 - \left\|f(x_a^i) - f(x_-^i)\right\|_2^2]_+$$

where, x_a^i, x_+^i, and x_-^i are i-th anchor, positive, and negative image while $f(x_a^i)$, $f(x_+^i)$, and $f(x_-^i)$ are their corresponding feature vectors through the CNN. N and C denote the total number of the images and category, respectively. θ is the set of parameters of the classifier, y_a^{ij} corresponds to the j-th element of one-hot encoded label of the sample x_a^i. $f_j(\cdot)$ denotes j-th element of $f(\cdot)$, since the output layer is a softmax. Also, α, β, and δ are weight balance parameters which control the strength between two loss functions and margin parameter, respectively.

We initialized all parameters of baseline models using the parameters which obtained from the ImageNet [16] pretrained models, while the parameters of the feature relaxation module are randomly initialized. All baseline models trained using Adam [15] optimizer, with initial learning rate 10^{-4}, which is decayed by 0.1 after 100 and 150 epochs. We set to α, β, and δ are 0, 0.1, and 0.3. We utilized training triplets according to the task requirements, and adopted three different types of loss function: object category based classification loss (OC), object category based similarity loss (OS), and instance based similarity loss (IS).

In the test time for the retrieval task, we extracted feature from a query image using the trained CNN model and compare distances using gallery features extracted in the same way. All the output feature vectors are L2 normalized, then the similarity was calculated using the inner product. For the evaluation both classification and retrieval tasks, we use top-k accuracy, as in [12, 13, 20].

Table 1. Comparison of top-k (k=5,20) retrieval accuracy on In-Shop retrieval dataset using different loss combinations.

Model	Loss combination	Accuracy	
		top-5	top-20
Deepfashion [20]	-	0.673	0.764
DARN [11]	-	0.547	0.675
ResNet50	OC	0.633	0.766
ResNet50	OC+OS	0.569	0.695
ResNet50	OC+IS	**0.826**	0.905
DenseNet121	OC	0.768	0.873
DenseNet121	OC+OS	0.628	0.749
DenseNet121	OC+IS	0.823	**0.909**

Table 2. Quantitative comparison of category classification on category prediction of DeepFashion dataset.

Method	Params. ($\times 10^6$)	Accuracy	
		top-3	top-5
WTBI [2]	-	43.73	66.26
DARN [11]	-	59.48	79.58
FashionNet+500 [20]	-	57.44	77.39
FashionNet+Joints [30]	-	72.30	81.52
FashionNet+Poselets [30]	-	75.34	84.87
Deepfashion [20]	$\sim$134	82.58	90.17
Lu et al.VGG-16 [21]	134.4	86.72	92.51
Weakly [3]	-	86.30	92.80
ResNet50+OC	28.1	87.34	93.42
DenseNet121+OC	7.9	87.58	93.39
SEResNet50+OC	28.1	87.58	93.58
SEResNeXt50+OC	27.6	**88.42**	**93.93**

3. Experimental Results

3.1. Datasets

We evaluated our method using three benchmarks on the DeepFashion dataset [20] as following: *Category Prediction*, *InShop* Clothes Retrieval, and *Consumer-to-Shop* Clothes Retrieval. The DeepFashion dataset is one of the largest publicly available fashion benchmark dataset including more than $800K$ images with plenty of information about labels of categories, attributes, bounding boxes, and landmarks. In this paper, original images without bounding box cropping were utilized for training and testing, and only the category labels and instance labels were used.

3.2. Quantitative Comparison

Loss function: To investigate the effects of the loss functions, we calculated the top-k accuracies with different combinations of the CNN architectures (ResNet50 and DenseNet121) and losses (OC, OS, and IS). As shown in Table 1, both ResNet50 and DenseNet121 with the OC already achieved comparable accuracies with previous methods. However, when the OC and OS were used together the performance was significantly degraded. On the other hand, in the case that the OC and IS were utilized together, we achieved a significant performance improvement.

Benchmark comparison: Quantitative benchmark results and comparisons with the state-of-the-art FIR methods are presented in Table 2, Fig. 3, and Fig. 4. For the task of category classification, our baseline models exhib-

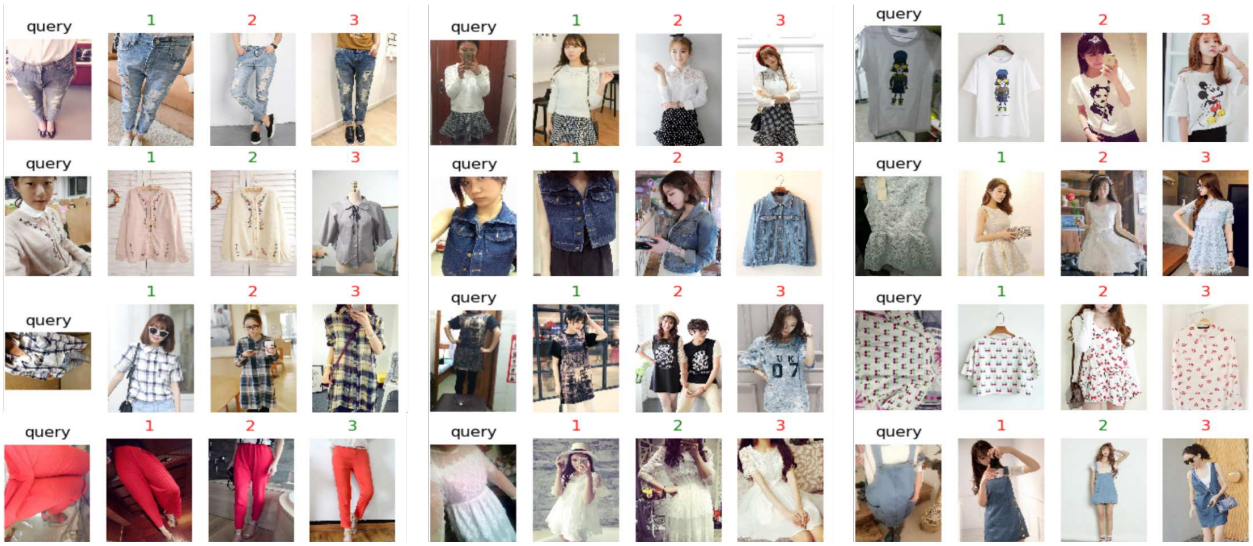

Figure 2. Qualitative evaluation on *consumer-to-shop* clothes retrieval benchmark using *DenseNet121+OC+IS*. Example query images and top-3 retrieved images are shown. Green number indicates correctly retrieved image while red number indicates wrong instance image.

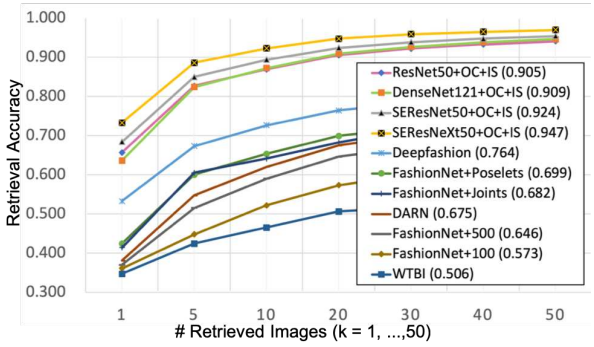

Figure 3. Retrieval accuracy for top-k (k=1,5,10,20,30,40,50) on *InShop* retrieval dataset. The top-20 retrieval accuracy for each model is described in the caption.

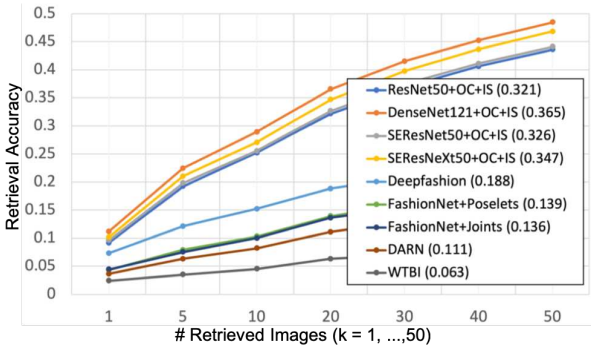

Figure 4. Retrieval accuracy for top-k (k=1,5,10,20,30,40,50) on Consumer-to-Shop retrieval dataset. The top-20 retrieval accuracy for each model is described in the caption.

ited a slight improvement compared to previous results, even with a smaller number of parameters. Moreover, in the retrieval task, our baseline models exhibited significant improvements compared to previously published results. The four considered FIR models trained using the OC and IS outperformed the state-of-the-art FIR methods by a significant margin.

3.3. Qualitative evaluation

For a qualitative evaluation, the *consumer-to-shop* clothes retrieval dataset was employed for a benchmark comparison. The *consumer-to-shop* dataset is more challenging than the *InShop* dataset, as it contains unrefined fashion images taken by users. As shown in Fig. 4, our best model, *DenseNet121+OC+IS*, outperforms all previous FIR methods. The qualitative retrieval results using our best FIR model with example query images and the top-3 retrieved images are presented in Fig. 2. As shown in Fig. 2, it is clear that our best FIR model can retrieve correct gallery images by understanding fashion details such as complex patterns, styles, and characters, even when such detailed information is not explicitly provided in the training process. In the last row of Fig. 2, although exact instance images are not included in the top-1, visually acceptable images with similar colors and styles are retrieved.

4. Conclusion

In this paper, we investigated an effective manner of training an FIR model based on consideration of the training strategy, relaxation module, and loss combination. Our empirical results on ResNet50, SEResNet50, SEResNeXt50, and DenseNet121 indicate that the considered training strategies and combination of loss functions leads to a consistent and significant improvement in the model accuracy, in terms of both the classification and retrieval performance. Based on the various evaluations, our considered FIR methods have outperformed the state-of-the-art FIR methods by a significant margin on three benchmark datasets.

References

[1] L. Bossard, M. Dantone, C. Leistner, C. Wengert, T. Quack, and L. Van Gool. Apparel classification with style. In *Asian Conference on Computer Vision (ACCV)*, pages 321–335, 2012.

[2] H. Chen, A. Gallagher, and B. Girod. Describing clothing by semantic attributes. In *European conference on computer vision (ECCV)*, pages 609–623, 2012.

[3] C. Corbiere, H. Ben-Younes, A. Ramé, and C. Ollion. Leveraging weakly annotated data for fashion image retrieval and label prediction. In *International conference on computer vision (ICCV)*, pages 2268–2274, 2017.

[4] Q. Dong, S. Gong, and X. Zhu. Multi-task curriculum transfer deep learning of clothing attributes. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 520–529, 2017.

[5] M. Hadi Kiapour, X. Han, S. Lazebnik, A. C. Berg, and T. L. Berg. Where to buy it: Matching street clothing photos in online shops. In *International conference on computer vision (ICCV)*, pages 3343–3351, 2015.

[6] X. Han, Z. Wu, Y.-G. Jiang, and L. S. Davis. Learning fashion compatibility with bidirectional lstms. In *International conference on Multimedia*, pages 1078–1086, 2017.

[7] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[8] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In *Computer Vision and Pattern Recognition (CVPR)*, pages 7132–7141, 2018.

[9] Y. Hu, X. Yi, and L. S. Davis. Collaborative fashion recommendation: A functional tensor factorization approach. In *International conference on Multimedia*, pages 129–138, 2015.

[10] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *Computer Vision and Pattern Recognition (CVPR)*, pages 4700–4708, 2017.

[11] J. Huang, R. S. Feris, Q. Chen, and S. Yan. Cross-domain image retrieval with a dual attribute-aware ranking network. In *Computer Vision and Pattern Recognition (CVPR)*, pages 1062–1070, 2015.

[12] J. Huang, R. S. Feris, Q. Chen, and S. Yan. Cross-domain image retrieval with a dual attribute-aware ranking network. In *International conference on computer vision (ICCV)*, pages 1062–1070, 2015.

[13] X. Ji, W. Wang, M. Zhang, and Y. Yang. Cross-domain image retrieval with attention modeling. In *ACM international conference on Multimedia*, pages 1654–1662, 2017.

[14] Y. Kalantidis, L. Kennedy, and L.-J. Li. Getting the look: clothing recognition and segmentation for automatic product suggestions in everyday photos. In *International Conference on Multimedia Retrieval*, pages 105–112, 2013.

[15] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

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

[17] Y. Li, L. Cao, J. Zhu, and J. Luo. Mining fashion outfit composition using an end-to-end deep learning approach on set data. *IEEE Transactions on Multimedia*, 19(8):1946–1955, 2017.

[18] X. Liang, L. Lin, W. Yang, P. Luo, J. Huang, and S. Yan. Clothes co-parsing via joint image segmentation and labeling with application to clothing retrieval. *IEEE Transactions on Multimedia*, 18(6):1175–1186, 2016.

[19] S. Liu, J. Feng, Z. Song, T. Zhang, H. Lu, C. Xu, and S. Yan. Hi, magic closet, tell me what to wear! In *International conference on Multimedia*, pages 619–628, 2012.

[20] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Computer Vision and Pattern Recognition (CVPR)*, pages 1096–1104, 2016.

[21] Y. Lu, A. Kumar, S. Zhai, Y. Cheng, T. Javidi, and R. Feris. Fully-adaptive feature sharing in multi-task networks with applications in person attribute classification. In *Computer Vision and Pattern Recognition (CVPR)*, pages 5334–5343, 2017.

[22] E. Simo-Serra, S. Fidler, F. Moreno-Noguer, and R. Urtasun. Neuroaesthetics in fashion: Modeling the perception of fashionability. In *Computer Vision and Pattern Recognition (CVPR)*, pages 869–877, 2015.

[23] E. Simo-Serra and H. Ishikawa. Fashion style in 128 floats: Joint ranking and classification using weak data for feature extraction. In *Computer Vision and Pattern Recognition (CVPR)*, pages 298–307, 2016.

[24] A. Veit, B. Kovacs, S. Bell, J. McAuley, K. Bala, and S. Belongie. Learning visual clothing style with heterogeneous dyadic co-occurrences. In *Computer Vision and Pattern Recognition (CVPR)*, pages 4642–4650, 2015.

[25] W. Wang, Y. Xu, J. Shen, and S.-C. Zhu. Attentive fashion grammar network for fashion landmark detection and clothing category classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4271–4280, 2018.

[26] T. Xiao, T. Xia, Y. Yang, C. Huang, and X. Wang. Learning from massive noisy labeled data for image classification. In *Computer Vision and Pattern Recognition (CVPR)*, pages 2691–2699, 2015.

[27] J. Xie, T. He, Z. Zhang, H. Zhang, Z. Zhang, and M. Li. Bag of tricks for image classification with convolutional neural networks. *arXiv preprint arXiv:1812.01187*, 2018.

[28] 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 (PAMI)*, 37(5):1028–1040, 2015.

[29] K. Yamaguchi, T. Okatani, K. Sudo, K. Murasaki, and Y. Taniguchi. Mix and match: Joint model for clothing and attribute recognition. In *British Machine Vision Conference(BMVC)*, volume 1, page 4, 2015.

[30] Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *Computer Vision and Pattern Recognition (CVPR)*, pages 1385–1392, 2011.