FashionSearch++: Improving Consumer-to-Shop Clothes Retrieval with Hard Negatives

Davide Morelli[1], Marcella Cornia[1] and Rita Cucchiara[1]

[1]*Department of Engineering "Enzo Ferrari", University of Modena and Reggio Emilia, Modena, Italy*

Abstract

Consumer-to-shop clothes retrieval has recently emerged in computer vision and multimedia communities with the development of architectures that can find similar in-shop clothing images given a query photo. Due to its nature, the main challenge lies in the domain gap between user-acquired and in-shop images. In this paper, we follow the most recent successful research in this area employing convolutional neural networks as feature extractors and propose to enhance the training supervision through a modified triplet loss that takes into account hard negative examples. We test the proposed approach on the Street2Shop dataset, achieving results comparable to state-of-the-art solutions and demonstrating good generalization properties when dealing with different settings and clothing categories.

Keywords

consumer-to-shop clothes retrieval, image retrieval, computer vision

1. Introduction

The visual search of an image from a database of several items is becoming a fundamental task for many different applications in the fields of information retrieval, computer vision, and multimedia. Typically, the task consists in finding the most similar images to a given query, which can be either another image [1, 2] or a textual sentence [3, 4, 5, 6, 7]. While text-based image retrieval can suffer from language constraints, image-based retrieval has no such limitations. Due to the ability to find similar images given a target one, this task fits perfectly with the great expansion of e-commerce and the need for customers to easily find what they are looking for among a large number of products. In particular, in the fashion domain, the ability for a customer to find an in-shop garment given a query photo is a remarkable feature.

In the last few years, much research effort [8, 9, 10, 11, 12, 13] has been spent on making e-commerce customer experience more effective and enjoyable, resulting in different solutions for clothes retrieval for both in-shop [8] and consumer-to-shop [14, 8, 15] settings. Focusing on consumer-to-shop clothes retrieval, the main challenge is given by the strong differences between query and in-shop images. In fact, while query images are usually taken in the wild and may exhibit low quality and lighting variations, in-shop images are usually high quality, in front perspective, and shot in a controlled environment. Almost all recent fashion retrieval works [16, 17, 15, 18] employ convolution neural networks (CNNs) to encode images and a supervised triplet loss function to train the overall architecture. In this paper, we follow this

IIR 2021 – 11th Italian Information Retrieval Workshop, September 13–15, 2021, Bari, Italy

davide.morelli@unimore.it (D. Morelli); marcella.cornia@unimore.it (M. Cornia); rita.cucchiara@unimore.it (R. Cucchiara)

0000-0001-7918-6220 (D. Morelli); 0000-0001-9640-9385 (M. Cornia); 0000-0002-2239-283X (R. Cucchiara)

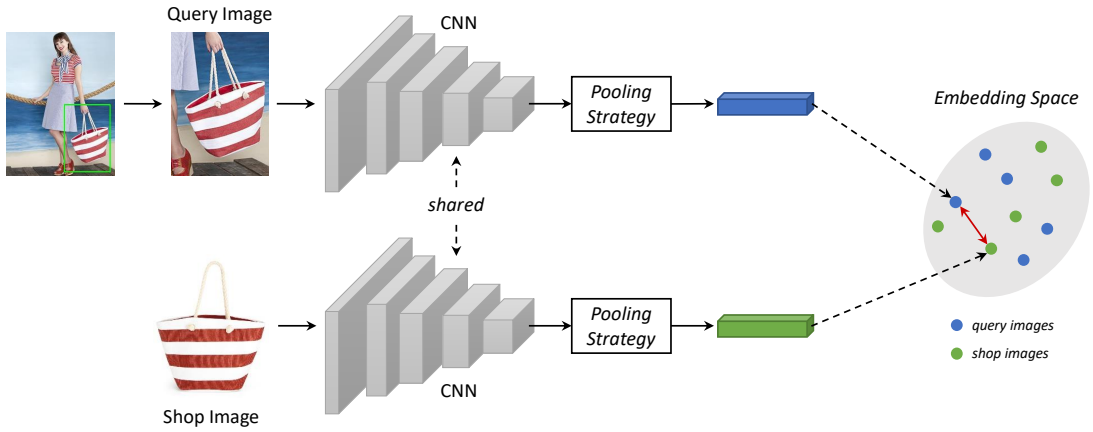

Figure 1: Overview of our approach for consumer-to-shop clothes retrieval.

line of research and propose to modify the standard hinge-based triplet loss function with the integration of hard negatives [3] thus improving the generalization abilities of the networks and increasing the final performance. Despite having been widely used to improve visual-semantic embeddings [3, 4, 19, 20, 21], this loss function has never been applied in the context of fashion retrieval. Experimental results on a widely used dataset for consumer-to-shop clothes retrieval, namely Street2Shop [14], demonstrate the effectiveness of this strategy leading to better retrieval results using both different backbones and pooling strategies. Furthermore, we show that the use of hard negative examples can significantly increase the final results on almost all categories of clothing and accessories (*e.g.* bags, dresses, footwear, skirts, etc.) and achieve performance comparable to state-of-the-art techniques.

2. Proposed Approach

Given a query image of a fashion item and a corresponding in-shop image, these are fed through a CNN followed by a pooling strategy to extract a 1D feature vector for each image. Then, the obtained feature vectors can be compared through a similarity function that measures the similarity between the two images. An overview of the proposed approach is shown in Fig. 1.

Extracting image features. Both query and in-shop images are processed through a CNN that extracts a 3D tensor for each image of $H \times W \times D$ dimensions, where H, W, and C are respectively the output tensor height, width, and number of channels. The 3D tensor can be seen as a set of 2D features channel responses $X = \{X_i\}$ where $i = \{1, \ldots, D\}$, X_i is the 2D tensor representing the responses of the i-th feature channel over the set Ω of spatial locations, and $X_i(p)$ is the response at a particular position p.

To obtain a single 2D tensor for each image, we employ two different pooling functions: a standard average pooling and R-MAC descriptors [1]. While the average pooling is a well-known pooling technique computed by averaging the set X of 2D tensors, R-MAC descriptors are an aggregation of image region descriptors extracted through a rigid-grid mechanism over X. Formally, considering a rectangular region $\mathcal{R} \subseteq \Omega = [1, W] \times [1, H]$, each region feature

vector is defined as:

$$f_R = [f_{\mathcal{R},1} \cdots f_{\mathcal{R},i} \cdots f_{\mathcal{R},k}]^\top, \tag{1}$$

where $f_{\mathcal{R},i} = \max_{p \in \mathcal{R}} X_i(p)$ is the maximum activation of the i-th channel of $\mathcal{R}$. Each region $\mathcal{R}$ is detected through a square grid of variable dimensions applied at L different scales. After extracting a feature vector for each region, they are processed using ℓ_2-normalization, PCA, and another ℓ_2-normalization. Finally, the region feature vectors are summed and ℓ_2-normalized to form a single feature vector for each image.

Training with hard negatives. Once the descriptor of the query and in-shop images are obtained, they are compared using a similarity function. Note that the descriptors embedding space is learned according to the loss function used in the backbone training phase. To extract similar descriptors from similar images, a standard hinge-based triplet ranking loss is usually employed and defined as:

$$L_{SH}(a, b) = \sum_{\hat{p}_c} [\alpha - s(a, b) + s(a, \hat{b})]_+ \tag{2}$$

where $[x]_+ = max(x, 0)$ and s is a similarity function (*i.e.* the cosine similarity in our experiments). In the equation above, (a, b) is a matching image pair composed of a user-generated image a and a shop image b (such that b contains the same fashion item depicted in a), while $\hat{b}$ is a negative shop image with respect to a (such that b' contains a different fashion item). The sum term in the equation requires that the difference in similarity between the matching and the non-matching pair is higher than a margin α.

As demonstrated in previous works [3], this loss function can be dominated by multiple negatives with small violations. To avoid such behavior, we employ a modified version that takes into consideration the hardest negative instead of the sum of all negative examples. In practice, this is done by replacing the sum in Eq. 2 with maximum, thus considering only the most violating non-matching pair. Formally, we define the loss function as follow:

$$L_{MH}(a, b) = \max_{b'} [\alpha - s(a, b) + s(a, b')]_+ \tag{3}$$

where only the hardest negative shop image b' is taken into account.

3. Experimental Evaluation

In this section, we evaluate the performance of our approach and describe the dataset and implementation details used in our experiments.

Dataset and implementation details. We train and test our model on Street2Shop [14] that contains 404,683 shop photos collected from 25 different online retailers and 20,357 user-generated photos. Overall, the dataset is composed of a total of 39,479 image pairs, each consisting of a user-generated photo and the corresponding shop image, from 11 different clothing categories. User-generated photos are annotated with bounding boxes of fashion items and can be associated with multiple views of the same fashion item.

To encode images, we use two different CNNs (*i.e.* ResNet-50 and ResNet-101 [22]) pre-trained on ImageNet [23]. We resize and crop all images to 224×224 and obtain a 2048-dimensional

Table 1
Retrieval results on the Street2Shop test set using shop images from all categories as retrievable items. Results are reported in terms of $R@K$ with $K = 1, 5, 10, 20$.

	Average Pooling				R-MAC			
	R@1	**R@5**	**R@10**	**R@20**	**R@1**	**R@5**	**R@10**	**R@20**
ResNet-50 (Pre-trained)	3.7	8.2	10.7	13.7	5.9	11.1	13.9	17.8
ResNet-50 (Finetuned)	14.8	24.9	30.9	36.9	15.4	26.8	32.5	38.8
ResNet-50 (Finetuned with HN)	**15.4**	**25.7**	**31.5**	**37.1**	**18.5**	**29.8**	**34.9**	**41.7**
ResNet-101 (Pre-trained)	3.8	8.1	10.4	13.4	6.6	11.9	14.8	18.0
ResNet-101 (Finetuned)	15.4	25.3	30.6	37.0	15.0	25.9	32.3	38.8
ResNet-101 (Finetuned with HN)	**17.4**	**28.0**	**33.8**	**39.8**	**23.6**	**36.0**	**42.4**	**48.6**

Table 2
Retrieval results in terms of $R@20$ on the 11 categories of the Street2Shop dataset.

	Bags	**Belts**	**Dresses**	**Eyewear**	**Footwear**	**Hats**	**Leggings**	**Outerwear**	**Pants**	**Skirts**	**Tops**
Wang *et al.* [25]	46.6	20.2	56.9	13.8	13.1	24.4	15.9	20.3	22.3	50.8	48.0
ResNet-50 + R-MAC											
Finetuned	52.5	**38.1**	52.4	36.2	28.9	29.2	42.7	36.0	34.8	71.6	45.4
Finetuned with HN	**61.2**	26.2	**56.7**	**46.6**	**31.9**	**40.0**	**48.1**	**37.8**	**39.4**	**74.1**	**48.8**
ResNet-101 + R-MAC											
Finetuned	54.0	**35.7**	53.2	20.7	29.9	29.2	45.9	33.8	**40.9**	74.4	46.0
Finetuned with HN	**65.5**	33.3	**63.4**	**50.0**	**39.2**	**64.6**	**48.7**	**42.7**	37.9	**78.4**	**55.5**

feature vector for each encoded image using both average pooling and R-MAC descriptors. In the case of R-MAC, we extract region feature vectors at 3 different scales. To train all models, we use Adam [24] as optimizer with a learning rate equal to 0.0001 decreased by a factor of 10 every 10 epochs. In all experiments, we use a batch size of 50 and a margin α equal to 0.1.

Experimental results. To evaluate the effectiveness of our approach, we report rank-based performance metrics $R@K$ ($K = 1, 5, 10, 20$) for consumer-to-shop clothes retrieval. Specifically, $R@K$ computes the percentage of test queries for which at least one correct result is found among the top-K retrieved shop items. Table 1 shows the results using all shop images as retrievable items on the Street2Shop test set, without filtering the images by category. We report the retrieval performance of both ResNet-50 and ResNet-101 backbones while extracting image feature vectors either using average pooling or R-MAC descriptors. We compare the results of our approach, in which we finetune the backbone using the hinge-based triplet loss with hard negatives, with those obtained by finetuning the CNNs with a standard triplet loss and those extracted by using the CNNs pre-trained on ImageNet without finetuning. As it can be seen, finetuning the backbone leads to a noteworthy gain in performance on all considered settings. Also, the modified triplet loss further improves the final performance using both ResNet-50 and ResNet-101 as backbone and employing both pooling strategies.

In Table 2, we report the performance on each of the 11 clothing categories of the Street2Shop dataset. These results are obtained by performing the retrieval on a subset of in-shop images, filtered by query category. As it can be noticed, the use of hard negatives generally increases

Figure 2: Top-3 retrieved results on sample query images from the Street2Shop test set. For each query, we show the top-3 results retrieved by the ResNet-101 model with R-MAC descriptors, finetuned on the dataset with and without the use of hard negatives during training. Correct and wrong retrieved elements are highlighted in green and red, respectively.

the network performance, leading to better results on almost all clothing categories. Finally, Fig. 2 shows sample query images along with the corresponding top-3 shop images retrieved by the ResNet-101 model using R-MAC descriptors and finetuned with and without the use of hard negatives in the training loss function.

4. Conclusion

In this work, we have tackled the task of consumer-to-shop clothes retrieval where the goal is to find the most similar clothing item from a catalog of shop images using a user-generated photo as query. To address the task, we have employed a CNN-based feature extraction network and two pooling mechanisms to extract compact feature vectors from images and have proposed to train the network with a modified hinge-based triplet ranking loss that takes into account hard negative examples. Experiments, performed on the Street2Shop dataset, have shown that the proposed loss function can effectively improve the retrieval results in all tested settings.

Acknowledgments

This work has been partially supported by YOOX NET-A-PORTER Group and the "SUPER - Supercomputing Unified Platform" project (POR FESR 2014-2020 DGR 1383/2018 - CUP E81F18000330007), co-funded by Emilia Romagna region.

References

[1] G. Tolias, R. Sicre, H. Jégou, Particular object retrieval with integral max-pooling of CNN activations, in: Proceedings of the International Conference on Learning Representations, 2016.

[2] A. Gordo, J. Almazán, J. Revaud, D. Larlus, Deep image retrieval: Learning global representations for image search, in: Proceedings of the European Conference on Computer Vision, 2016.

[3] F. Faghri, D. J. Fleet, J. R. Kiros, S. Fidler, VSE++: Improving Visual-Semantic Embeddings with Hard Negatives, in: Proceedings of the British Machine Vision Conference, 2018.

[4] K.-H. Lee, X. Chen, G. Hua, H. Hu, X. He, Stacked cross attention for image-text matching, in: Proceedings of the European Conference on Computer Vision, 2018.

[5] M. Cornia, L. Baraldi, H. R. Tavakoli, R. Cucchiara, Towards cycle-consistent models for text and image retrieval, in: Proceedings of the European Conference on Computer Vision Workshops, 2018.

[6] M. Cornia, M. Stefanini, L. Baraldi, M. Corsini, R. Cucchiara, Explaining digital humanities by aligning images and textual descriptions, Pattern Recognition Letters 129 (2020) 166–172.

[7] M. Stefanini, M. Cornia, L. Baraldi, R. Cucchiara, A Novel Attention-based Aggregation Function to Combine Vision and Language, in: Proceedings of the International Conference on Pattern Recognition, 2020.

[8] Z. Liu, P. Luo, S. Qiu, X. Wang, X. Tang, DeepFashion: Powering robust clothes recognition and retrieval with rich annotations, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2016.

[9] X. Han, Z. Wu, Z. Wu, R. Yu, L. S. Davis, VITON: An Image-based Virtual Try-On Network, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018.

[10] B. Wang, H. Zheng, X. Liang, Y. Chen, L. Lin, M. Yang, Toward characteristic-preserving image-based virtual try-on network, in: Proceedings of the European Conference on Computer Vision, 2018.

[11] Y. Ge, R. Zhang, X. Wang, X. Tang, P. Luo, DeepFashion2: A Versatile Benchmark for Detection, Pose Estimation, Segmentation and Re-Identification of Clothing Images, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019.

[12] A. Neuberger, E. Borenstein, B. Hilleli, E. Oks, S. Alpert, Image Based Virtual Try-On Network From Unpaired Data, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020.

[13] M. Fincato, F. Landi, M. Cornia, F. Cesari, R. Cucchiara, VITON-GT: An Image-based Virtual Try-On Model with Geometric Transformations, in: Proceedings of the International Conference on Pattern Recognition, 2020.

[14] M. Hadi Kiapour, X. Han, S. Lazebnik, A. C. Berg, T. L. Berg, Where to buy it: Matching street clothing photos in online shops, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2015.

[15] Z. Kuang, Y. Gao, G. Li, P. Luo, Y. Chen, L. Lin, W. Zhang, Fashion retrieval via graph

reasoning networks on a similarity pyramid, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019.

[16] X. Zhao, H. Qi, R. Luo, L. Davis, A Weakly Supervised Adaptive Triplet Loss for Deep Metric Learning, in: Proceedings of the European Conference on Computer Vision Workshops, 2019.

[17] A. Chopra, A. Sinha, H. Gupta, M. Sarkar, K. Ayush, B. Krishnamurthy, Powering robust fashion retrieval with information rich feature embeddings, in: Proceedings of the IEEE/CFV Conference on Computer Vision and Pattern Recognition Workshops, 2019.

[18] A. D'Innocente, N. Garg, Y. Zhang, L. Bazzani, M. Donoser, Localized Triplet Loss for Fine-Grained Fashion Image Retrieval, in: Proceedings of the IEEE/CFV Conference on Computer Vision and Pattern Recognition Workshops, 2021.

[19] L. Baraldi, M. Cornia, C. Grana, R. Cucchiara, Aligning text and document illustrations: towards visually explainable digital humanities, in: Proceedings of the International Conference on Pattern Recognition, 2018.

[20] M. Stefanini, M. Cornia, L. Baraldi, M. Corsini, R. Cucchiara, Artpedia: A new visual-semantic dataset with visual and contextual sentences in the artistic domain, in: Proceedings of the International Conference on Image Analysis and Processing, 2019.

[21] M. Cornia, L. Baraldi, H. R. Tavakoli, R. Cucchiara, A unified cycle-consistent neural model for text and image retrieval, Multimedia Tools and Applications 79 (2020) 25697–25721.

[22] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2016.

[23] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, L. Fei-Fei, ImageNet Large Scale Visual Recognition Challenge, International Journal of Computer Vision 115 (2015) 211–252.

[24] D. Kingma, J. Ba, Adam: a method for stochastic optimization, in: Proceedings of the International Conference on Learning Representations, 2015.

[25] X. Wang, Z. Sun, W. Zhang, Y. Zhou, Y.-G. Jiang, Matching user photos to online products with robust deep features, in: Proceedings of the ACM International Conference on Multimedia Retrieval, 2016.