Localized Triplet Loss for Fine-grained Fashion Image Retrieval

Antonio D'Innocente*
Sapienza University of Rome, Italy
dinnocente@diag.uniroma1.it

Nikhil Garg
Amazon, Germany
nikhgarg@amazon.de

Yuan Zhang
Amazon, US
yuaa@amazon.com

Loris Bazzani
Amazon, Germany
bazzanil@amazon.de

Michael Donoser
Amazon, Germany
donoserm@amazon.de

Abstract

Fashion retrieval methods aim at learning a clothing-specific embedding space where images are ranked based on their global visual similarity with a given query. However, global embeddings struggle to capture localized fine-grained similarities between images, because of aggregation operations. Our work deals with this problem by learning localized representations for fashion retrieval based on local interest points of prominent visual features specified by a user. We introduce a localized triplet loss function that compares samples based on corresponding patterns. We incorporate random local perturbation on the interest point as a key regularization technique to enforce local invariance of visual representations. Due to the absence of existing fashion datasets to train on localized representations, we introduce FashionLocalTriplets, a new high-quality dataset annotated by fashion specialists that contains triplets of women's dresses and interest points. The proposed model outperforms state-of-the-art global representations on FashionLocalTriplets.

1. Introduction

In this work, we explore the task of fashion retrieval, defined as learning of a clothing-specific embedding space. Existing methods define the concept of similarity at a global image level [12, 10, 15, 23, 7, 31]. Such methods struggle when dealing with fine-grained visual differences between clothes, especially in the case when they are localized. To deal with this problem, we propose a fine-grained localized notion of similarity where the similarity is defined on the interest point level of prominent visual features of a garment. This allows the method to focus on desired localized cues, such as specific sleeve types, necklines, and design patterns

Figure 1. FashionLocalTriplets annotations. Left: prominent attributes. Right: a triplet annotation. Candidate B was marked as more similar to the reference with respect to the interest point.

(Figure 1, left).

The proposed model consists of a convolutional network that ingests an image and the location of a prominent cue on the image. Localized features are extracted from the convolutional map at the provided location in a patch and aligned via bilinear interpolation. We train the model using a novel localized triplet loss function in order to be able to retrieve the most similar images by considering their local similarity between corresponding points.

Moreover, we introduce random perturbation of interest points as a key regularization technique for this task.

In order to train our model on localized points of an image, we propose a new dataset of women's dresses annotated by fashion specialists, named FashionLocalTriplets. It consists of triples of the form: a reference image with annotated interest point location and two candidate images, as shown in Figure 1. Fashion specialists annotated the interest points of prominent features of the garment (Figure 1, left), and decided which of the two candidates is more similar to the reference in terms of the local point (Figure 1, right). Our experiments on FashionLocalTriplets show that our model is able to learn better localized representations to perform the task of image retrieval compared to non-localized methods.

*This work was done during an internship at Amazon.

2. Related Work

Attribute Localization for Fashion. Attributes are used as supervised signals to describe images of clothing [3, 4, 30, 5, 1, 6, 24, 27]. With the introduction of datasets that include attribute locations and landmarks [15, 32, 9], there has been a proliferation of methods that leverage locations for retrieval [17, 26, 28, 22, 16]. Conversely, our training algorithm requires a few points provided by fashion experts, but not attribute names or correspondences on other images.

Fashion Image Retrieval. Most recent works in fashion image retrieval leverage convolutional neural networks [12, 10, 15, 23, 7, 31] for image representation.

For learning these embeddings, triplet losses [31, 15] have been shown to provide state-of-the-art performance.

Localized Retrieval. A few methods integrate local information in the retrieval process. They do through segmentation [19], local/global features aggregation [14] and re-weighting through attention [8, 2] or saliency [18]. In our method, the user provides the point of interest, and retrieval is based on the localized representation alone.

Fashion Datasets. In the last years, several datasets have been proposed for fashion-related tasks. DeepFashion [15], Fashionpedia [13], DeepFashion2 [9] and FashionAI [32] provide large scale data with rich annotations. Our FashionLocalTriplets includes relative comparisons between regions of a reference image and two candidates which can be used as ground-truth to build triplets for training.

3. The FashionLocalTriplets Dataset

We created FashionLocalTriplets, a dataset that contains locations of prominent visual cues, with the aim of capturing fine-grained local differences in clothing. Following previous work [15, 32, 29], we relied on shopping websites and randomly picked $4,302$ women's dresses images from *amazon.com*.

3.1. Annotations

Annotations were created by the specialists in two passes: 1) labeling of prominent attributes, and 2) labeling of triplets.

Labeling of prominent attributes. For each image, fashion specialists marked the coordinates, clothing locations and attribute names of interesting or unique localized attributes of the garment, as shown in figure 1 (left).

Labeling of triplets. For each reference image and each of its annotated attributes from the previous stage, we randomly selected two candidates with the same attribute. Fashion specialists judged which of the two candidates was visually more similar to the reference in terms of the interest point, and annotated a total of $10,805$ hard triplets. Figure 1 (right) shows an example of an annotated triplet.

3.2. Evaluation Tasks

We design two evaluation tasks for this dataset, namely: 1) Binary Classification and 2) Retrieval.

Binary Classification. For each test triplet, we consider the binary classification task of correctly selecting the closest candidate within the triplet, and evaluate using accuracy.

Retrieval. Fashion specialists selected a set of 100 query images of women's dresses from *amazon.com*. For each query, we use global image embeddings from a pre-trained fashion similarity network [31] to retrieve the 20 closest neighbors, and ask fashion specialists to label them as relevant/non-relevant. With this dataset, we can compute precision-recall scores by considering ranking results produced by different algorithms on the retrieval set of each query.

4. Learning Localized Embeddings

To learn fine-grained representations, our method leverages localized embeddings, a location aware training loss function and regularization via interest point perturbation.

4.1. Localized Image Encoder

Figure 2 (top) illustrates the proposed localized image encoder. The input consists of an $H \times W$ image and an (x, y) interest point identifying a prominent local attribute in the image. We use a ResNet-50 backbone and extract a feature map of size $C \times C$. To obtain a localized representation of the prominent attribute, we map the interest point (x, y) to the $C \times C$ coordinate system and use bilinear interpolation to get a 3×3 approximation of the feature representation at (x, y) in the original coordinate system. The bilinearly interpolated feature map is then passed through a convolutional layer and L_2 normalized.

When the interest point is not available, we compute localized features for each point on a grid covering all spatial locations. These embeddings will be indexed for retrieval and used in our localized loss function.

4.2. Localized Triplet Loss

The triplet loss [25] is widely used for providing state of the art performance in image retrieval. It is defined as follows:

$$\mathcal{L}_{global} = \sum_{n=1}^{N} \Big\lfloor \|g(I_n) - g(I_n^+)\|_2^2 - \|g(I_n) - g(I_n^-)\|_2^2 + m \Big\rfloor_+ \tag{1}$$

where $g(I)$ denotes the global embedding of an image I. I_n, I_n^+, I_n^- indicate the n-th reference, positive and negative images respectively, N is the size of the dataset, $\lfloor z \rfloor_+$ refers to $max(0, z)$ and m is the margin of the loss.

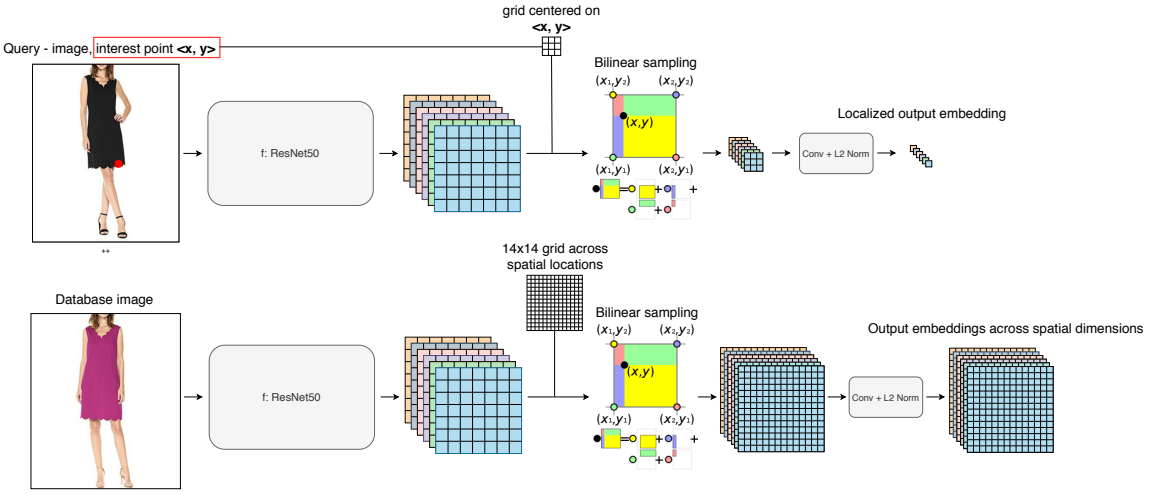

Figure 2. Localized feature extraction for a query (top) and a database image (bottom).

Localized Triplet Loss. We extend the triplet loss to include the location of interest points as follows:

$$\mathcal{L}_{local} = \sum_{n=1}^{N} \Big\lfloor \min_{c_n^+} \|f(I_n, p_n) - f(I_n^+, c_n^+)\|_2^2 - \min_{c_n^-} \|f(I_n, p_n) - f(I_n^-, c_n^-)\|_2^2 + m \Big\rfloor_+ \quad (2)$$

where p_n are the (x, y) coordinates of the interest point related to the reference image and $f(I_n, p_n)$ is the proposed localized embedding that we described in the previous section (Figure 2, top). For the positive and negative images, we compute the localized embeddings on the c_n^+ and c_n^- locations, spanning all the locations of the convolutional map. The min operations in Eq. 2 have the effect of taking the features in the convolutional map which are closest to the feature of the reference interest point in the embedding space. In this way, the network is encouraged to learn to discriminate between regions in images that are similar, *e.g.*, the sleeve regions in all the images in a triplet.

4.3. Random Interest Point Perturbation

During training, we perform data augmentation by randomly moving the location of the interest point on the reference image within a radius of $R\%$ (R=20 in our experiments) of the input image size. The intuition is that the localized similarity should not drastically change even if the user specified interest point varies a bit in its exact pixel location. This augmentation has a regularization effect which significantly improves the results, as shown in the experiments.

4.4. Inference

First, we compute and index the localized embeddings of the database images, as illustrated for a single image in Figure 2(bottom). At runtime, given an input image and its interest point, we first compute the localized query embedding using the procedure described in Sec. 4.1. We then compute L_2 distance between the query embedding and all the database embeddings. The match score of a database image is taken as the inverse of the smallest distance between any of its localized embeddings and the query embedding. The database items are then ranked in the order of decreasing match scores.

5. Experiments

Implementation Details. Of the 10,805 annotated triplets in the FashionLocalTriplets dataset, we select 9,298 for training via image ID hashing. We then filter the 1,507 remaining triplets for unanimous votes, obtaining 399 annotations that we reserve for testing. We used ResNet-50 as backbone network pre-trained on the fashion retrieval task of [31]. We resize images to 224×224 and perform random horizontal flipping, color jittering and random perturbations of the input interest point. We train the network for 30,000 iterations using SGD. The margin hyperparameter m is set to 0.05.

Comparisons. We compare the proposed localized training method with the following approaches. (1) *Pre-trained localized embeddings* using the architecture shown in Figure 2, with the pre-trained weights on either the ImageNet [20] classification task, or the visual fashion retrieval task [31]. (2) *Global triplet loss* ($m = 0.05$ determined using cross-validation). (3) *Localized Contrastive*, trained by turning each triplet annotation into two pairs of matching/non-matching images and using a localized version of the contrastive loss [11]. All the reported results are averaged across 3 independent runs.

	Method	Emb. Scope	Acc
1.	Pre-trained ImageNet [20]	Global	54.34
2.	Pre-trained Fashion [31]	Global	63.01
3.	Pre-trained ImageNet [20]	Local	56.89
4.	Pre-trained Fashion [31]	Local	65.56
5.	Triplet [21]	Global	69.90
6.	Localized Contrastive [11]	Local	65.86
7.	Localized Triplet	Local	**73.72**

Table 1. Accuracy for different methods on FashionLocalTriplets. The Emb. Scope column indicates whether features are obtained by global average pooling (Global) or from our localized image encoder (Local).

5.1. Results

Binary Classification results. Table 1 shows accuracy results on the FashionLocalTriplets test set. In the first four rows, we evaluate global and localized embeddings computed with a ResNet-50 pre-trained on ImageNet [20] and pre-trained on the fashion retrieval task of [31]. The table shows that the embeddings from the fashion image retrieval task work better than the ones from the ImageNet classification task, as the former have been pre-trained on fashion images. Furthermore, localized embeddings improve over global ones, as the latter fail to precisely capture local features of interest.

The last three rows show results of models that are fine-tuned on the FashionLocalTriplets with the standard triplet loss (row 5), our localized version of the contrastive loss (row 6) and our localized triplet loss (row 7). Comparing row 5 with previous rows shows that fine-tuning outperforms the pre-trained versions, illustrating the value of curating a training dataset. Adding localized information to the triplet loss (row 7) improves the results by large margins over global embeddings (row 5), highlighting the importance of using localized features on this task. The localized version of the contrastive loss (row 6) didn't perform well on this task, barely improving over pre-trained embeddings.

Retrieval results. As shown in Table 2, the relative comparison of precision-recall between different methods follows the trend in Table 1. Specifically, our Localized Triplet method outperforms all baselines, with only Recall@1 tied with the Localized Contrastive and Pre-trained Fashion embeddings. The Contrastive model barely improves over the pre-trained Fashion model in both Precision and Recall.

	Method	P@1	P@5	P@10	R@1	R@5	R@10
1.	Pre-trained ImageNet [20]	0.70	0.62	0.57	0.10	0.40	0.70
2.	Pre-trained Fashion [31]	0.73	0.63	0.57	**0.11**	0.41	0.68
3.	Localized Contrastive [11]	0.78	0.64	0.58	**0.11**	0.41	0.69
4.	Localized Triplet	**0.81**	**0.67**	**0.59**	**0.11**	**0.44**	**0.72**

Table 2. Precision/Recall values for different localized methods on the FashionLocalTriplets retrieval task.

Method	IP-Reg	Acc
Localized Triplet		71.17
	✓	73.72

Table 3. Ablation results on interest point regularization (IP-Reg).

Figure 3. Heatmaps visualizations of queries (left) and database images in top-2 retrievals (right).

5.2. Ablation Study

Table 3 shows the contribution of the interest point regularization (IP-Reg). This regularization generates embeddings robust to small translation invariance, consistently improving accuracy for the localized triplet loss.

5.3. Localization of Interest Points

Our algorithm matches the interest point in a query image to points on the database images by comparing their localized embeddings. To verify that this matching is happening as expected, we compute heatmaps of L_2 distances between the localized embeddings of the query and the database images. In Figure 3, we can observe that the network learned to match corresponding locations (hemline, v-neck) in images, even though no interest point location was provided for the candidate images during training.

6. Conclusions

We proposed a localized triplet loss to train localized embeddings for fine-grained visual similarity based retrieval of fashion items. We presented a new dataset of women's dress images that contains hard triplets of images. We demonstrated that our proposed localized embeddings set the state of the art in the new application of location-based retrieval. A future direction can be to relax the need of annotations from fashion specialists by automatically creating negatives at different level of complexity (relative comparisons).

Acknowledgments. We are very thankful to the fashion experts at Amazon who helped us annotate the FashionLocalTriplets dataset.

References

[1] Abrar H Abdulnabi, Gang Wang, Jiwen Lu, and Kui Jia. Multi-task cnn model for attribute prediction. *IEEE Transactions on Multimedia*, 17(11):1949–1959, 2015. 2

[2] Kenan E Ak, Ashraf A Kassim, Joo Hwee Lim, and Jo Yew Tham. Learning attribute representations with localization for flexible fashion search. In *CVPR*, pages 7708–7717, 2018. 2

[3] Tamara L Berg, Alexander C Berg, and Jonathan Shih. Automatic attribute discovery and characterization from noisy web data. In *ECCV*, pages 663–676. Springer, 2010. 2

[4] Lukas Bossard, Matthias Dantone, Christian Leistner, Christian Wengert, Till Quack, and Luc Van Gool. Apparel classification with style. In *ACCV*, pages 321–335. Springer, 2012. 2

[5] Huizhong Chen, Andrew Gallagher, and Bernd Girod. Describing clothing by semantic attributes. In *ECCV*, pages 609–623. Springer, 2012. 2

[6] Qiang Chen, Junshi Huang, Rogerio Feris, Lisa M Brown, Jian Dong, and Shuicheng Yan. Deep domain adaptation for describing people based on fine-grained clothing attributes. In *CVPR*, pages 5315–5324, 2015. 2

[7] Charles Corbiere, Hedi Ben-Younes, Alexandre Ramé, and Charles Ollion. Leveraging weakly annotated data for fashion image retrieval and label prediction. In *ICCV Workshops*, pages 2268–2274, 2017. 1, 2

[8] Songhe Feng, De Xu, and Xu Yang. Attention-driven salient edges and regions extraction with application to cbir. *Signal Processing*, 90(1):1–15, 2010. 2

[9] Yuying Ge, Ruimao Zhang, Xiaogang Wang, Xiaoou Tang, and Ping Luo. Deepfashion2: A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images. In *CVPR*, pages 5337–5345, 2019. 2

[10] M Hadi Kiapour, Xufeng Han, Svetlana Lazebnik, Alexander C Berg, and Tamara L Berg. Where to buy it: Matching street clothing photos in online shops. In *ICCV*, pages 3343–3351, 2015. 1, 2

[11] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE, 2006. 3, 4

[12] Junshi Huang, Rogerio S Feris, Qiang Chen, and Shuicheng Yan. Cross-domain image retrieval with a dual attribute-aware ranking network. In *ICCV*, pages 1062–1070, 2015. 1, 2

[13] Menglin Jia, Mengyun Shi, Mikhail Sirotenko, Yin Cui, Claire Cardie, Bharath Hariharan, Hartwig Adam, and Serge Belongie. Fashionpedia: Ontology, segmentation, and an attribute localization dataset. *arXiv preprint arXiv:2004.12276*, 2020. 2

[14] Zhanghui Kuang, Yiming Gao, Guanbin Li, Ping Luo, Yimin Chen, Liang Lin, and Wayne Zhang. Fashion retrieval via graph reasoning networks on a similarity pyramid. In *ICCV*, pages 3066–3075, 2019. 2

[15] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *CVPR*, pages 1096–1104, 2016. 1, 2

[16] Ziwei Liu, Sijie Yan, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Fashion landmark detection in the wild. In *ECCV*, pages 229–245. Springer, 2016. 2

[17] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, pages 483–499. Springer, 2016. 2

[18] Alex Papushoy and Adrian G Bors. Image retrieval based on query by saliency content. *Digital Signal Processing*, 36:156–173, 2015. 2

[19] Rouhollah Rahmani, Sally A Goldman, Hui Zhang, John Krettek, and Jason E Fritts. Localized content based image retrieval. In *ACM MM*, pages 227–236, 2005. 2

[20] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015. 3, 4

[21] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015. 4

[22] Alexey Sidnev, Alexey Trushkov, Maxim Kazakov, Ivan Korolev, and Vladislav Sorokin. Deepmark: One-shot clothing detection. In *ICCV Workshops*, Oct 2019. 2

[23] Yang Song, Yuan Li, Bo Wu, Chao-Yeh Chen, Xiao Zhang, and Hartwig Adam. Learning unified embedding for apparel recognition. In *ICCV Workshops*, pages 2243–2246, 2017. 1, 2

[24] Sirion Vittayakorn, Takayuki Umeda, Kazuhiko Murasaki, Kyoko Sudo, Takayuki Okatani, and Kota Yamaguchi. Automatic attribute discovery with neural activations. In *ECCV*, pages 252–268. Springer, 2016. 2

[25] Liwei Wang, Yin Li, and Svetlana Lazebnik. Learning deep structure-preserving image-text embeddings. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5005–5013, 2016. 2

[26] Wenguan Wang, Yuanlu Xu, Jianbing Shen, and Song-Chun Zhu. Attentive fashion grammar network for fashion landmark detection and clothing category classification. In *CVPR*, June 2018. 2

[27] Takuya Yashima, Naoaki Okazaki, Kentaro Inui, Kota Yamaguchi, and Takayuki Okatani. Learning to describe e-commerce images from noisy online data. In *ACCV*, pages 85–100. Springer, 2016. 2

[28] Weijiang Yu, Xiaodan Liang, Ke Gong, Chenhan Jiang, Nong Xiao, and Liang Lin. Layout-graph reasoning for fashion landmark detection. In *CVPR*, pages 2937–2945, 2019. 2

[29] Yanhao Zhang, Pan Pan, Yun Zheng, Kang Zhao, Yingya Zhang, Xiaofeng Ren, and Rong Jin. Visual search at alibaba. In *ACM SIGKDD*, pages 993–1001, 2018. 2

[30] Yuwei Zhang, Peng Zhang, Chun Yuan, and Zhi Wang. Texture and shape biased two-stream networks for clothing classification and attribute recognition. In *CVPR*, pages 13538–13547, 2020. 2

[31] Xiaonan Zhao, Huan Qi, Rui Luo, and Larry Davis. A weakly supervised adaptive triplet loss for deep metric learning. In *ICCV Workshops*, pages 0–0, 2019. 1, 2, 3, 4

[32] Xingxing Zou, Xiangheng Kong, Waikeung Wong, Congde Wang, Yuguang Liu, and Yang Cao. Fashionai: A hierarchical dataset for fashion understanding. In *CVPR Workshops*, pages 0–0, 2019. 2