Cross-domain fashion image retrieval

Bojana Gajić, Ramon Baldrich
Computer Vision Center
Universitat Autnoma de Barcelona
Edifici O. UAB. Bellaterra, Spain.
{bgajic, ramon}@cvc.uab.es

Abstract

Cross domain image retrieval is a challenging task that implies matching images from one domain to their pairs from another domain. In this paper we focus on fashion image retrieval, which involves matching an image of a fashion item taken by users, to the images of the same item taken in controlled condition, usually by professional photographer. When facing this problem, we have different products in train and test time, and we use triplet loss to train the network. We stress the importance of proper training of simple architecture, as well as adapting general models to the specific task.

1. Introduction

Fashion market, and more specifically clothes retail, has gained interest into applying machine learning technologies to add value to the customer shopping experience [5]. One of the problems addressed is to help customers localize items just from regular pictures taken in an open environment. That means, to identify items from an image database of products of, usually, high quality guided by a customer picture. Up to a few years ago this task was difficult due to the limitations on the machine learning techniques available. The advances on neural networks applied to computer vision lead to the renaissance of many hard problems. In our particular case, any system that aims to solve this task should deal with a large amount of different items to recognize among, just by a simple picture taken in an open scenario without any limitation. This is not possible with traditional techniques [2, 4], which fail absolutely when having to check for similarities on never seen items. Our task is a typical image retrieval problem, where images are very similar but they usually can be differentiate by subtle details. An added complexity is the different nature of the query images and the target ones. The later are usually high quality images where the item is the main focus of the scene, whereas the former is not controlled in any way (figure 4).

The approach that we will use is an adversarial learning methodology based on positive and negative stimuli applying a Siamese network architecture, trained using triplet loss [7].

2. Method

To deal with the fashion retrieval task, we have created a three streams Siamese architecture and used the same settings as in [1], which focuses on person re-identification. Each stream is composed of convolutional layers of ResNet50 network, max pooling and one fully connected layer. The final embeddings are l_2 normalized. The weights of the convolutional and fully-connected layers are shared between the streams. We have applied the standard Triplet loss function[8]:

$$L(I_q, I_p, I_n) = max(0, m + d(\mathbf{q}, \mathbf{p}) - d(\mathbf{q}, \mathbf{n})), \quad (1)$$

where $\mathbf{q}$, $\mathbf{p}$ and $\mathbf{n}$ are the embeddings for the query image, the relevant image and the non-relevant image, respectively, $d()$ is the euclidean distance and m is the margin which controls the difference in the distances between query and the relevant image and between the query and the non-relevant one. Depending on how we extract the embeddings we are pushing the system to encapsulate instance differences.

We initialize the convolutional layers of the three streams by the weights obtained during the pretraining on the classification task, while the weights of the FC layer are initialized randomly. We chose the query image randomly, relevant image as any of the images from the same class as the query one, and the negative as an image from a different class which is among the 25 non-relevant images closest to the query. This baseline has been trained using SGD optimizer, with initial learning rate 10^{-3}, which is decayed by 0.5 each 1024 iterations.

Taking into account that our query images are always from the user domain, while the desired retrieved examples are from the in-shop domain, we propose adaptation of the previously described method according to the task

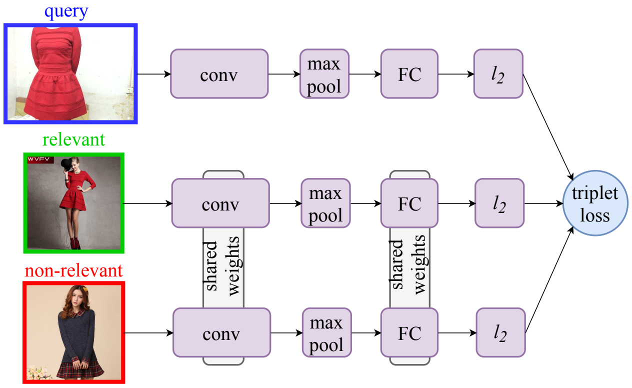

Figure 1. Training model. The model is made of three Siamese streams, each one of them contains convolutional layers followed by max pooling and fully connected layers. The final representation is l_2 normalized. Query image is randomly sampled from the user images, while the relevant and non-relevant images are chosen from the in-shop domain.

requirements. We have separated the streams for each domain, meaning that our new query stream does not share the weights with the gallery stream, which we use for extracting representations of relevant and non-relevant images 1. We sample the triplets the same way as in the previous approach. We initialize the weights of the convolutional and fully connected layers by the weights obtained during the training for ranking with three Siamese streams. We have used SGD optimizer, with starting learning rate $5 \cdot 10^{-4}$ and decayed it after each epoch for one half.

In the test time, we extract features from queries using the user stream, and the features of the gallery by the second stream. As the embeddings are l_2 normalized, we calculate the similarity by simply using dot product. We use top-K measure to evaluate the results, as proposed by [3].

3. Results

3.1. Datasets

We evaluate our models using two common datasets, DARN and DeepFashion.

The DARN dataset [2] is a standard cross-domain fashion image dataset. It contains around 327,000 images from the in-shop domain and 91,000 user images. Apart from providing IDs of each image, this dataset includes labels such as clothes category, button, color, length etc. Due to the broken links in the files provided by the authors, we use the clean version provided by [3], and follow their evaluation protocol.

The DeepFashion dataset [6] is the largest publicly available fashion dataset including more than 800,000 images with additional information about categories, landmarks etc.

In this paper, we use only images and information about their IDs and domains.

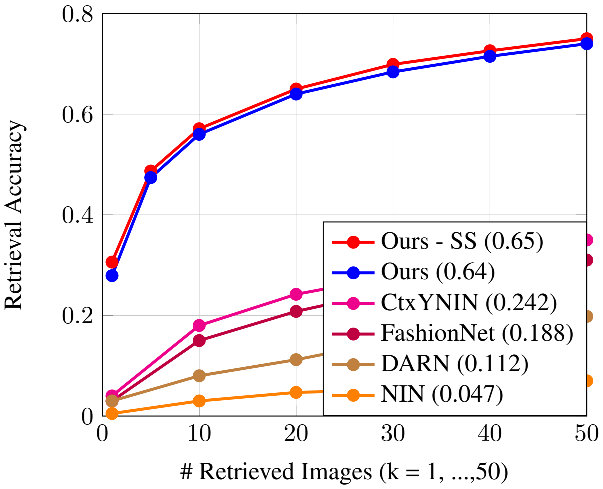

Figure 2. Results on DARN. The last four methods are reported by [3]. Our result with the proposed architecture outperforms the state of the art by 41%.

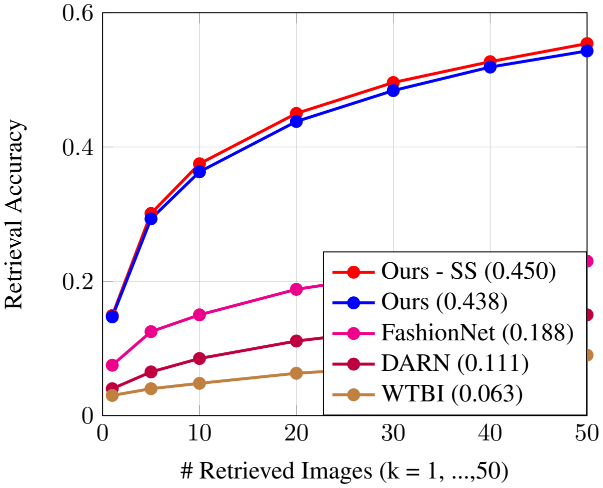

Figure 3. Results on DeepFashion dataset.

3.2. Quantitative results

Quantitative results and comparison with state of the art are shown on the Figure 2 and 3. Our baseline already outperforms the published results by 25% on DeepFashion and 40% on DARN, while the additional adaptation (Ours-SS), adapting each stream to its domain, is improving the baseline for 1%.

3.3. Qualitative results

Even though our results are better than the published work, our model is able to find a correct item in the first 20 retrieved images only in 65% for DARN and 45% for DeepFashion of cases. On the Figure 4 we show the qualitative results of our best model. It is clear that our model is able to understand the type of the item, without implicit training for that task, and that failure cases are appearing when the query image contains item in a low resolution or when the light are causing low quality of the query image.

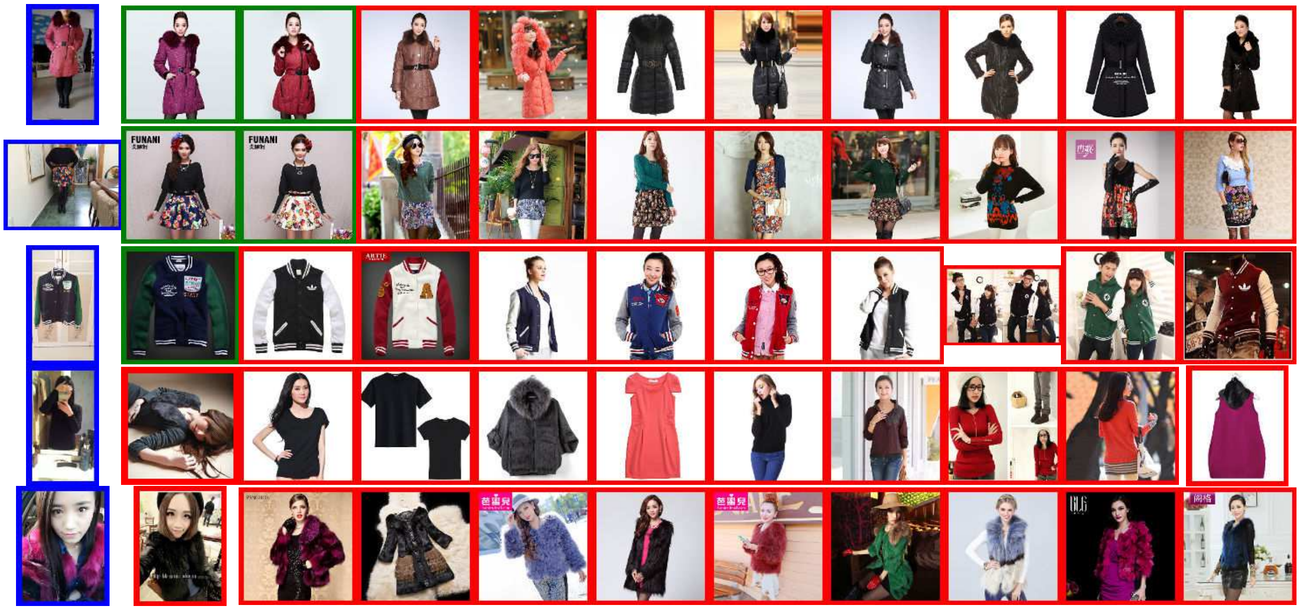

Figure 4. Examples of retrieved items. Each row contains a query image (with a blue frame) followed by 10 closest gallery images. Green frames stand for the right match, while the red ones signify items different from the query. We show that our model is able to not only distinguish between different items, but also to understand the category in which the item is.

4. Conclusion

In this paper we have shown the importance of training simple models correctly, using pretraining for a simpler task, proper triplet mining and careful learning rate decay. Additionally, we have proposed extension of the simple architecture, in terms of using a model which is more appropriate to the specific cross-domain problem. Both of our architectures have outperformed the state-of-the-art methods by a significant margin on two datasets.

5. Acknowledgments

This work has been supported by MILETRANS project (TIN2014-61068-R) of Spanish Ministry of Economy and Competitiveness and by FI grant 2017FI_B_01045 from Generalitat de Catalunya.

References

[1] J. Almazan, B. Gajic, N. Murray, and D. Larlus. Re-ID done right: towards good practices for person re-identification. *ArXiv e-prints*, Jan. 2018.

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

[3] X. Ji, W. Wang, M. Zhang, and Y. Yang. Cross-domain image retrieval with attention modeling. In *ACM Multimedia*, 2017.

[4] M. H. Kiapour, X. Han, S. Lazebnik, A. C. Berg, and T. L. Berg. Where to buy it: Matching street clothing photos in online shops. In *ICCV*, pages 3343–3351, 2015.

[5] K. Lin, H.-F. Yang, J.-H. Hsiao, and C.-S. Chen. Deep learning of binary hash codes for fast image retrieval. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2015 IEEE Conference on*, pages 27–35. IEEE, 2015.

[6] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1096–1104, 2016.

[7] J. Wang, T. Leung, C. Rosenberg, J. Wang, J. Philbin, B. Chen, Y. Wu, et al. Learning fine-grained image similarity with deep ranking. *arXiv preprint arXiv:1404.4661*, 2014.

[8] J. Wang, F. Zhou, S. Wen, X. Liu, and Y. Lin. Deep metric learning with angular loss. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2612–2620, Oct 2017.