Multi-Granularity Part Alignment Convolutional Network for Commodity Retrieval

Liangyu Chen[3], Xumao Wu[2], Zhiwei Fang[1], Jing Liu[1]
[1] University of Chinese Academy of Sciences
[2] Northwestern Polytechnical University, [3] Kunming University of Science and Technology
{zhiwei.fang, jliu}@nlpr.ia.ac.cn,
wuxuan400@mail.nwpu.edu.cn, 525876914@qq.com

***Abstract*—Fashion-Item matching aims to match the images from consumers and from e-commerce platform, and plays an increasingly important role in the business of e-commerce. In this paper, we propose a novel deep framework called multi-granularity part alignment convolutional network, which combine global features with part-level features and offer effective discriminant information for commodity retrieval. Besides, instead of only utilizing uniform partition on the input image in horizontal direction, we combine the horizontal, vertical and annular partition method, to obtain various local feature representations without requiring extra supervision. After the CNN learning, we merge the output features of all branches to compute the similarity between images. Experiment on the dataset provided by JD Company demonstrate the effectiveness of the proposed framework for the task of commodity retrieval. Our framework achieves the rank-1 accuracy of 54.86% and mAP@10 of 0.5886.**

***Index Terms*—Fashion-Item matching, Multi-Granularity, annular partition method**

I. INTRODUCTION

With the popularity of E-commerce platforms, people are increasingly willing to take pictures on their favorite objects to find matching products. This brings the development of commodity retrieval techniques, also called as Fashion-Item matching. Due to the problems of image quality, illumination, occlusion and so on, this task is still challenging..

Before the renaissance of deep learning, traditional retrieval approaches usually focus on the intuitive features such as color, outline, and local characteristic. These simple feature extraction approaches lead to lower accuracy. When the convolutional neural network (CNN) has dominated this field, we have found this technology of Fashion-Item matching give us excellent user experience. Many domestic companies like JD and Alibaba have provided similar services, e.g. pailitao.com, but the retrieval performances of these state-of the-art methods do not satisfy the growing needs of consumers.

Many intuitive CNNs try to learn a global feature, without considering the spatial location and local information. These CNNs may have poor performance in several cases: 1) pictures of commodity from the consumer and electric business platform own completely different angle and spatial location, e.g., Figure 1 (a, b); 2) global feature cannot represent the uniqueness of one commodity, especially when we have to distinguish two commodities which have very similar appearances, e.g., Figure 1 (c); 3) the environments of images are complicated, so the global features of the same commodity are also completely diverse, e.g., figure 1 (d). This paper, starting from several difficulties mentioned above, focus on solving the association between global features and local features to reduce the influence of different rotation angles and various spatial positions on final retrieval accuracy. We aim to refine the part-based model to reinforce multi-granularity features. Above all, this paper make two important contributions as followed.

Figure 1. Some positive and negative image pairs.

Firstly, different from the intuitive CNNs, which focus on learning global feature to capture the most salient clues of the commodity, we propose a new deep learning framework named multi-granularity part alignment convolutional network, which can learn global features and local features from different part-level image at the same time. This deep learning framework contains different image partition methods to realize the multiple granularity. This convolution network considers the whole image as input and concatenates a global feature with mult-granularity features as output. The idea of this part inspired by Part-based Convolutional Baseline (PCB) which was proposed by Yifan Sun in 2017 for field named Person ReID. The network named PCB which conducts

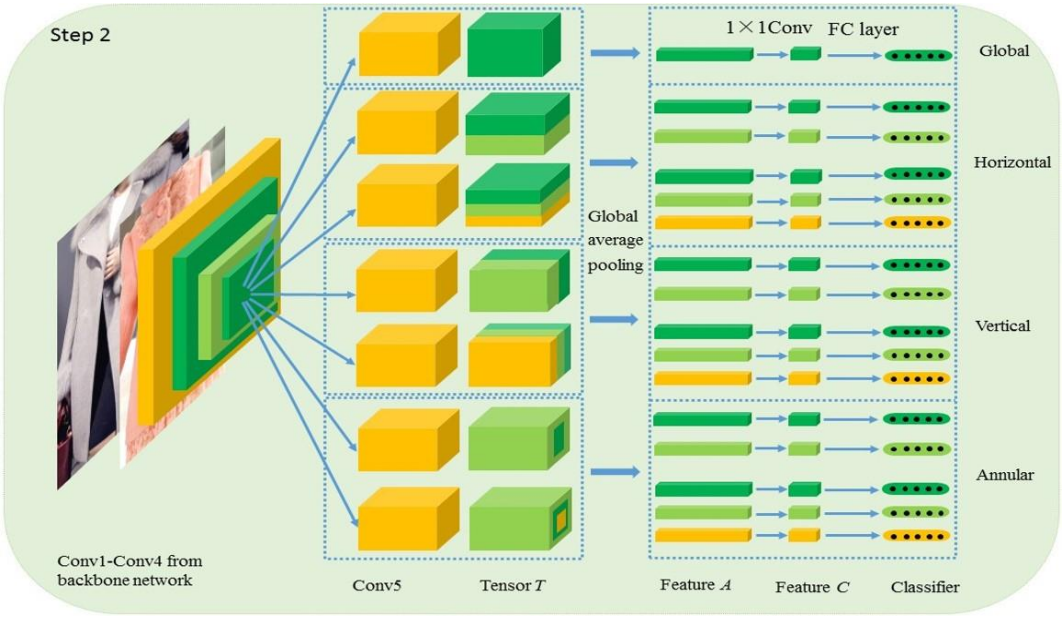

Figure 2. The architecture of the proposed deep framework. Conv5 layer from the backbone network is split into seven branches: Global Branch, Vertical-2, Vertical-3, Horizontal-2, Horizontal-3, Annular-2 and Annular-3. Using various partition methods, Tensor *T* can be split into some pieces of feature. Following global average pooling, the 1×1 kernel-size convolutional layer reduces the dimension of *T*. Finally, each feature *C* should be fed into a classifier, respectively. During testing, all the feature *C* are concatenated together as the final feature.

uniform partitions on the conv-layer to learn part-level features. We replace uniform partitions with another method and make other improvements in PCB which will discussed as follows.

Secondly, it is vital to choose a reasonable commodity image partition method to obtain some local features in this network. The horizontal image partition in PCB does not suit to Fashion-Item matching because of the uncertainty of location and angle of commodity in images. In this paper, we propose a new idea named concentric rectangular partition method, which is used to replace the original horizontal partition method in PCB. Our concentric rectangular partition method greatly reduce the impact of different commodity rotation angles on the accuracy of final retrieval.

Compared to the previous retrieval method, our network which combines two parts mentioned above has higher robustness in dealing with different rotation angle of commodities. We utilize the data set of 3 commodities (clothes, shoes, bags) provided by JD company to test the retrieval performance in the Fashion-Item matching task. Without any external data from outside or re-ranking procedure, the result on JD AI fashion-challenge competition shows that our network has achieve outperforming performance. In addition, our network is an end-to-end deep learning procedure, which is quite convenient for the whole training and test process.

II. Related Work

Commodity image retrieval plays an increasingly important role in the business of e-commerce because of the change of young people's consumption concept. Many scientists do the research in the field of clothing retrieval [1], but these research didn't expand to common commodities. Some works based on various part alignment method to match the similar product [2]. These approaches can fit all types of commodities but the accuracy still remain at lower level.

After the renaissance of deep learning, more works tried to solve the problem with improved deep learning approaches. For instance, a new framework with a Siamese network [3] was proposed to learn a cosine similarity measure for commodity retrieval. Many advanced metric loss function such as contrastive loss [4], triplet loss [5] and quadruplet loss [6] have provided more robust patterns to guide the similarity learning. But these approaches suffer from the uncertainly of rotation angles and spatial positions.

III. Our Approach

In this section, we first describe the retrieval method, whose framework is shown in figure 2. Then, we present the improved image partition method in the pooling step.

3.1 The framework of CNN

In the multi-granularity part alignment CNN, we generate one global feature and various local features of one commodity image as the final output. Besides, the learning process to get global and local features are independent to each other. To a certain, the backbone network in our Multi-Granularity Part Alignment CNNs inherits from PCB, which can take any image classification network without hidden fully-connected

layers as the backbone network, e.g., VggNet, Google Inception and ResNet. In the process of JD AI fashion-challenge competition, ResNet101 turns into the last choice of our backbone network by considering both the retrieval performance and the cost of total computation at the same time. Show as figure 2, we can learn that even without explicit attention mechanisms added to enhance the expression of some prominent components, our deep learning network can still learn the principal and subtle features of input images.

As illustrated in figure 2, before the proposed Multi-Granularity Part Alignment Convolutional Network, we add one classifier, which enforce the features to be correctly classified to the target identity. The original structure before the global average pooling (GPA) layer is retained the same as to the backbone model. The GPA layer is completely modified, and is also different from the method of local parts GPA in PCB network. We utilize horizontal, vertical and concentric rectangle's strips as the foundation of local parts GPA. And the method of how to get local strips will be discussed carefully in the next part of this section. Then, it is the most important that the rest of parts behind the GPA layer are replaced by 7 independent braches which divided from local GPA layers and share the similar architecture.

In the first branch, named Global Branch, we employ the GPA on the feature generated from backbone network to a 2048-dimensional feature *A*, followed by a 1×1 kernel-sized convolutional layer with batch normalization to reduce the obtained feature to 512-dim or 256-dim named feature *C*. Then, the dimension-reduced feature *C* is used as the input of the classifier. This branch is designed to learn the global feature information which has none partition strips of whole feature.

The next two branches have some modifications from the first global branch. Specially，the original tensor *T* must be uniformly split into several strips in vertical direction before the operation of GPA. The number of strips is determined by specific network and different tasks, and we choose 2 and 3 in JD AI fashion-challenge completion. The following operations are GPA and 1×1 kernel-sized layer, the rest parts of these branches share the similar network architecture with the first branch. So we can observe that there are five feature maps before classifier. The upper and lower branches in this part named as Vertical-2 and Vertical-3 Branch.

The rest branches in figure 2 have the same network architecture with Vertical-2 Branch or Vertical-3 Branch except the split direction of feature *A*. The fourth and fifth branches which split feature *A* into several strips with the horizontal direction, and we call them Horizontal-2 and Horizontal-3 Branch. Last two branches use the partition method of concentric rectangles before GPA, then sixth branch and seventh branch are named Annular-2 and Annular-3 Branch. Both in the training process and testing process, all the final features are either 512-dim or 256-dim.

3.2 Image Partition

Compared with Person Re-ID task, the Fashion-Item matching needs more attention on the challenge of the different

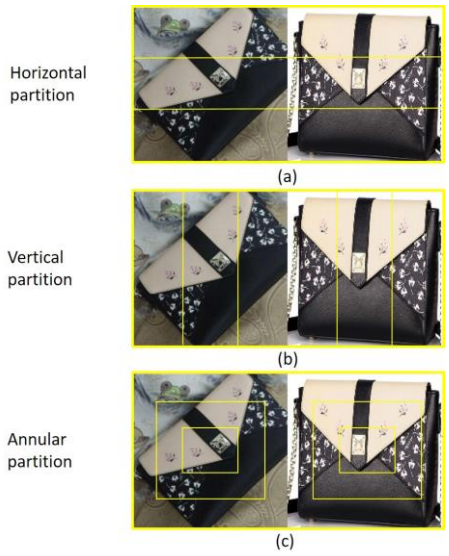

Figure 3. Three partition methods of Tensor *T*

spatial location and various rotation angle of principal subject, as illustrated in figure 3. Due to the randomness of consumers' photos, a simple segmentation of tensor T from the vertical or horizontal direction may appear misalign of local feature *A*. So we choose 3 approaches to handle the segmentation of tensor T, and they are horizontal partition (figure.3 (a)), vertical partition (figure.3 (b)) and annular partition (figure.3 (c)) respectively. We mainly explain the annular partition in the following.

No matter what the rotation angle of the principal subject are, the annular method can always obtain the corresponding features which are necessary for the next layer. From the point of programming implementation, the concentric rectangle method (figure.2 (c)) has the similar effect with the annular method and it is easy to be implemented. To verify effectiveness of each method in the multi-granularity part alignment convolutional network, we conduct several experiments with the combination of different methods.

3.3 Loss functions and some vital parameters

In order to enhance the discriminative ability of matching features among different categories, we employ the soft-max loss for classification. During the training process, the multi-granularity part alignment convolutional network is optimized by minimizing the sum of Cross-Entropy losses of ID predictions which generated by 7 branches. The final feature map we need in testing process is feature *C* and last output feature dim is 6×512.

In order to obtain more information about the matching feature of input images, the input images should be resized to 384×384. As observed in our experiment, with the increase of the number of strip in tensor *T*, the retrieval accuracy of final result get a slight improvement, but more computation costs and larger memory space are needed. Considering the cost of computation and the memory space at the same time, the number of feature strips in feature *A* is set to 2 or 3.

Branches	Backbone	Final feature	mAP(%)
Horizontal	Resnet50	256×16	40.32
Vertical	Resnet50	256×16	40.32
Annual	Resnet50	256×16	43.19
All	Densenet169	256×16	48.47
All	Resnext101	256×16	51.84
All	Resnet152	256×16	53.22
All	Resnet101	512×16	53.18
All	Resnet101	256×16	54.76

Table 1. Result of different branches in clothes

IV. EXPERIMENTS

4.1 Datasets

This paper uses one principal dataset from JD Company. This dataset is composed of three major categories of commodities, i.e., clothes, bags and shoes. Clothes and bags contain 5000 identities photographed by E-commerce platform and consumers. And bags only have 2000 identities. It is important to note that each ID of these categories only has two train images and the total number of train images is 24000. In addition, this dataset also contain 154294 gallery images and 1000 queries.

4.2 Implementation Details

In order to obtain a higher retrieval accuracy, we trained a classifier to classify the categories at the beginning of the training of multi-granularity part alignment convolutional network. We use Resnet101 as the backbone network, which replaces the 5×5 filter kernel with 3×3 channel-wise convolution layer and 1×1 spatial convolution layer. In all experiments, the mini-batch size of input is set to 64, and each iteration has 32 identities. Each epoch includes 157 mini-batches. During iteration, we use an Adam optimizer with an initial learning rate of 0.06, and shrink this learning rate by a factor of 0.2 at the 20 epoches. After 55 to 60 epochs, the training process achieve convergence. The data augmentation includes random mirroring and cropping. Moreover, due to memory limitation, we do not use re-ranking algorithms which considerably improves the result of mean average precision (mAP) in the testing process.

In our experiments, we train three model which adapt to three different commodities. To evaluate the retrieval performances, we report the mAP on the 1000 queries. Through various methods of pooling and different number of branches, the result of experiments in clothes shown in table 1.

Figure 3. The first column: query image. Other columns is top-5 result by our method

V. CONCLUSION

This report makes two contributions to solving the commodity retrieval problem. And proposed deep learning network get a pretty ranking in JD AI Fashion-challenge competition. In order to get a better result, we test various fusion schemes between different networks and different weights in three categories. Through these experiments we can obtain that the performance of only one partition method is poor and annual partition method outperforms other method by a large margin.

Our final submission scheme on July 20th is consisted of Resnet101, Resnet152, Resnext101 and Densenet169, then the final scheme achieves the rank-1 accuracy of 54.86% and mAP@10 of 0.5886. The final retrieval results are shown as figure.3.

REFERENCES

[1] X.Wang, T. Zhang, D. R. Tretter, and Q. Lin, “Personal clothing retrieval on photo collections by color and attributes,” IEEE Transactions on Multimedia, vol. 15, no. 8, pp. 2035–2045, Dec 2013.

[2] J. He, J. Feng, X. Liu, T. H. Lin, H. Chung, and S. F. Chang, “Mobile product search with bag of hash bits and boundary reranking,” 2012.

[3] Z. Fang, J. Liu, Y. Wang, Y. Li, S. Hang, J. Tang, and H. Lu, “Object-aware deep network for commodity image retrieval,” in ACM on International Conference on Multimedia Retrieval, 2016, pp. 405–408.

[4] X. Wang, Z. Sun, W. Zhang, Y. Zhou, and Y. G. Jiang, “Matching user photos to online products with robust deep features,” in ACM on International Conference on Multimedia Retrieval, 2016, pp. 7–14.

[5] Cheng, De, et al. "Person re-identification by multi-channel parts-based cnn with improved triplet loss function." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.

[6] Cheng, De, et al. "Person re-identification by multi-channel parts-based cnn with improved triplet loss function." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 201