Visual-based Product Retrieval with Multi-task Learning and Self-attention

Haibo Su[1], Chao Li[2], Wenjie Wei[2], Qi Wu[2], and Peng Wang[1]

[1]Northwestern Polytechnical University, Xi'an, China
[2]Vismarty Technology Co., Ltd. Shenzhen, China

***Abstract*—In this paper, we propose a visual-based product retrieval model trained by a triplet loss. The triplets are formed by pairwise matching item images and non-matching item images. In addition, two methods have been applied to our model to improve the accuracy of retrieval. The first is to add a classification loss which can help to narrow the retrieval range. The second is a way to capture correlative features between matching items by a self-attention mechanism. We further combine the CNN features and self-attended features together as the image representation. During the training, we aim to minimise the L2 distances between the features of similar items, and maximise it between features of distinct items. Moreover, the trained classifier can exclude items of different classes in retrieval items. We find that our proposed model can improve the results comparing to the baseline deep metric learning model with a triplet loss. Finally, multiple networks with different backbone models are trained and ensembled together to get better performance. Our final model won the third place in the 2018 JD AI Fashion Challenges.**

***Index Terms*—triplet loss, retrieval, multi-task, self-attention, pairwise**

I. Introduction

As online shopping becoming more and more popular, visual-based product retrieval is playing an important role in electronic commerce. For example, a matching commodity can be retrieved just by a picture. It can be challenging that a picture can be captured with different illumination condition, various angles and noisy background. In this paper, we introduce a deep metric learning model with self-attention mechanism and multi-task learning. Relevant experiment results demonstrate that our proposed method can largely improve the accuracy of product retrieval.

For a pairwise retrieval task, we need to compute similarities between the query items and gallery items. A good similarity metric can improve the performance of image matching. Previous retrieval approaches based on the contrastive embedding is to minimizes the distance between a pair of matching items and enlarge the distance of the negative pair. But our experiment based on the contrastive embedding shows that type of methods suffer bad convergence and poor performance. Here, we use deep metric learning with triplet loss as our fundamental architecture. This basic model learns the Euclidean distance of images by a deep convolutional network. We train this model using triplets, each of which contains a pair of positive images and a negative image, to pull the features from positive pair closer and push the features from negative pairs far away.

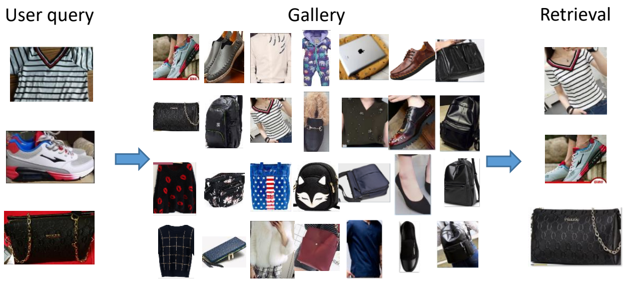

Fig. 1. Examples of retrieval results using our method on JD Fashion dataset. The images in first column are user images containing cloth, shoe and bag.

In addition, a self-attention mechanism is considered in our deep network to get better feature. In general, some features in feature maps of the item images are just noise. We reduce influence of indifferent features and place extra emphasis on significant features correspond to item similarity. Self-attention is a mechanism of capturing correlate features between similar item pairs to achieve better performance in item retrieval task. A straightforward way to utilize self-attention is to assign weight to each feature of feature maps generated by a deep convolutional network. The self-attention features are then added to CNN features to generate final image features. With the self-attention, relative features of similar item pairs will be given priority and the improvement of retrieval accuracy has been approved.

Another solution to improve performance of our model is to narrow the retrieval scope in the gallery by multi-task learning. Except a deep metric learning with triplet loss, we add an object classifier to our model. The combination of a deep metric learning task and a classification task guarantees that the items which we retrieved from gallery are in the same class with the query item. Therefore, we can get rid of items in different class from retrieval sequence and improve performance of our model.

We evaluate our method on the dataset provided by JD corporation. The dataset for train is approximately $12k$ pairs in which a ID only has a user image and a product image.

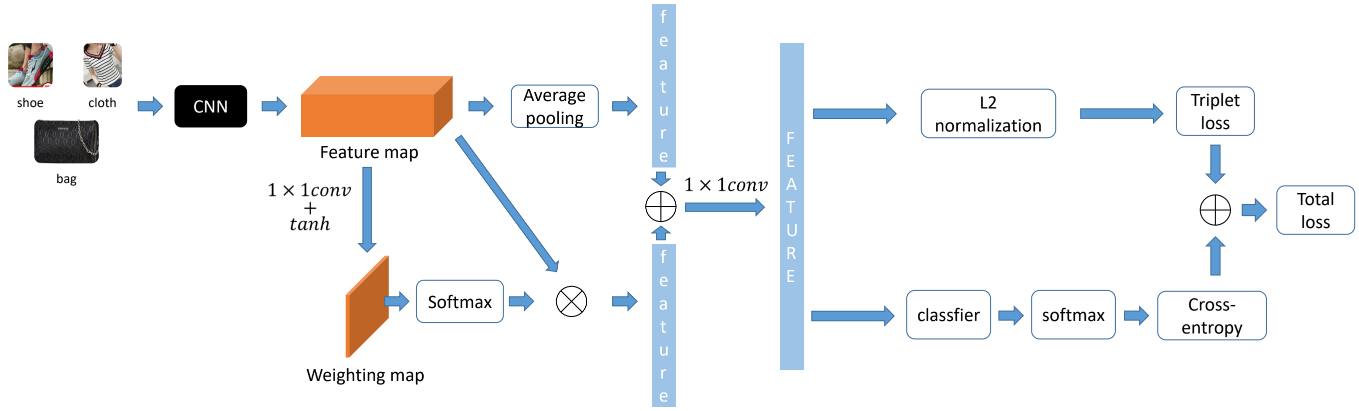

Fig. 2. Framework of proposed Multi-task Triplet Self-attention Model for item matching. In this architecture, it contains two integral parts, i.e., the self-attention embedding,the multi-task embedding. Self-attention uses several convolution layers to get a metric as a new convolution kernal and then generate a attention feature. Multi-task integrate the triplet task and the classification task. The triplet task generates triplet loss and the classification task generates cross-entropy loss. Finally, a total loss can be implemented by adding triplet loss and cross-entropy loss.

In test dataset, the query has 1000 images, and the gallery has $150k$ images. The dataset turns out to be three classes containing bag, cloth and shoe on the whole. Figure 1 shows some examples of our retrieval results on JD fashion dataset. We show the retrieval quality by mean Average Precision.

II. Related Works

Several Content-Based Image Retrieval(CBIR) algorithms have been proposed in recent years. These methods fall into two main categories: hand-crafted features based approach and CNN features based approach.

Earlier works on CBIR mostly relied on hand-crafted features, such as color histogram, SIFT [4] and HOG [5]. But hand-crafted features are not robust enough on some specific scenarios.

Since the CNN-based methods have achieved great success in classification tasks [1], they have recently been applied to image retrieval tasks [2], [3]. The CNN-based methods extract features using CNN models. Some methods directly using pre-trained CNN models, which pre-trained on a large-scale datasets like ImageNet. Features are extracted in a single forward pass. Instead of using pre-trained models as a feature extractor, fine-tuning on a specific training set can improve discriminative ability of the models [10].

There has been growing interest in similarity metric learning with CNN-based models. Hadsell et al. [6] use a contrastive loss based on pair sampling in CNN to learn image similarity metric. Schroff et al. [8] successfully apply triplet loss to facial recognition tasks which are highly similar to image retrieval tasks. Wen et al. [9] combine softmax loss and center loss to learn more discriminating features, which is an example of using multi-task learning.

Many studies have proven that multi-task learning is effective in many computer vision problems [11]. CNN-based model is very suitable for multi-task learning since the features learned from a task may be useful for other similar task. Since different tasks have different noise patterns, a model that learns multi tasks at the same time can learn a more general representation.

There are also some research of visual attention methods. Itti et al. [13] proposed a visual attention mechanism inspired by the primate visual system. Xu et al. [14] use visual attention mechanism to focus on salient objects and achieve better results in captioning tasks. Therefore, Extracting the salient regions of an image may be helpful for instance retrieval task [12].

III. Our Model

We propose a multi-task triplet self-attention method to learn features to distinguish the similarity of items from the region of interest of images. The major framework of the proposed method is shown in Figure 2. There are two main parts, a self-attention sub-network and a multi-task sub-network. The first part can generate a attention feature vector and a local feature vector. The two feature vectors can constitute a final feature vector. We make use of triplet loss and classification loss to train our model in the second part based on feature provided by the first part. In this framework, metric learning with triplet loss plays more important role.

A. Overall Structure

There are many deep networks that can be selected for our convolutional neural network to learn features of items to evaluate the similarity of items. Here, we use two classical network structures, i.e., inception resnet v2 and inception v4, as our backbone model. In general, these deep networks have been pre-trained on ImageNet. In our structure, we apply pre-trained model to backbone for fine-tuning. Backbone can generate a feature map, as usual, it will be implemented to acquire a local feature vector after average pooling. Here, we add a self-attention sub-network into our structure. After several convolutional layers and a activation function, we can achieve a metric which has same size with one metric of feature maps. This metric can be a new convolution kernel after softmax processing. With convolution operation between

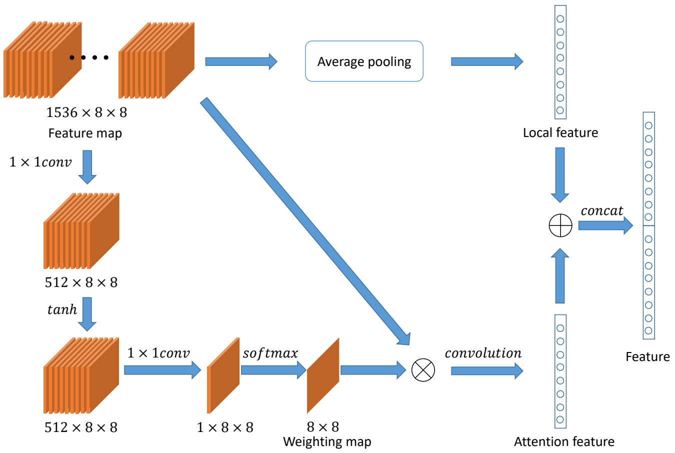

Fig. 3. The sub-network of self-attention. As we can see, one branch of our self-attention sub-network points to a local feature directly. Another branch of the self-attention sub-network contributes to extracting a weighting map for the feature map.

feature maps and the attention convolution kernel, we can get a attention feature vector. As shown in Figure 2, the local feature vector and the attention feature vector can be spliced to a final feature vector as input for the following multi-task sub-network.

The multi-task sub-network contains two components, a metric learning task with triplet loss and a classification task. For metric learning task, we can get a embedding vector after a one by one convolution layer, then, choose two matching items as positive pairs and a non-matching item as the negative one to make up a triplet. We define a triplet loss by computing and comparing distances among the triplet. For classification task, the dataset can be classified into three major categories. We can acquire a three-embedding-size vector after a one by one convolution layer. A classification loss can be achieved with feeding the three-embedding-size vector into a cross-entropy loss function. Last of all, we add the triplet loss and the classification loss up, then, we get a total loss which can be utilized to train our model. The details of self-attention and multi-task will be described in the following sections.

B. Self-attention

In general, our spatial self-attention model takes the ongoing feature map generated by a deep convolutional neural network as its input and produces a weighting map to assign different weight to different feature. With training our network in this way, the network will pay more attention to the spatial regions which are more beneficial to item retrieval accurately. Suppose we put a image into the network, a feature map which is the input of our self-attention sub-network will be obtained after a CNN. As shown in Figure 3, there are two main branches for our self-attention sub-network. One branch produces a local feature after a simple average pooling. Another branch can produce a weighting matrix which is calculated based on following formula:

$$H_i = tanh(W_i x_i + b_i) \tag{1}$$

In this equation, $tanh(x) = 2sigmoid(2x) - 1$ is the tanh function where $sigmoid(x) = 1/(1+exp(-x))$ is the sigmod function, W_i and b_i represent weights and bias on the one by one convolution layer. Compared to common attention model in previous works [12] [13] [14], we train our attention model just by x_i, i.e., each spatial region on the feature map, therefore, we define our attention model as self-attention. In Figure 3, we achieve a $512 \times 8 \times 8$ tensor. Then, we can get a weighting map by

$$\alpha_i = softmax(w^T H_i), \quad i = 1, ..., N, \tag{2}$$

where $softmax(x)$ is a softmax function, w^T is the parameters of the convolution layer for dimensionality reduction. Finally, we get a weighting figure α_i. These figures can be combined into a weighting map, as a 8×8 weighting map in Figure 3. Then, we can achieve a new feature which we define as a attention feature by a sum operation

$$\tilde{x} = \sum_{i=1}^{N} \alpha_i x_i \tag{3}$$

where $\tilde{x}$ is the attention feature, N is the number of features in the feature map. Last of all, we concat the local feature and the attention feature as the input feature for multi-task sub-network.

C. Multi-task Learning

In addition to providing 12k image pairs, the competition organizer also provides the category labels (clothes, shoes and bags) for each image. When the model only uses the triplet loss function to learn the distinguishable features between each different commodity, it does not take into account that there should be greater differences between different categories of commodities. Sometimes the texture of a piece of clothing may be similar to the texture of a bag. Learning from triplet loss alone may make it difficult to increase the feature distance between the categories. However, through the classification loss function, the model can learn the differentiating features between categories. On the basis of triplet loss, adding classification loss can increase the feature distance between categories and let the model learn more discriminating features.

In a multi-task learning network, the parameters of the hidden layers between different tasks are shared such that features of a certain task can be used by other tasks, and these features may be difficult to learn in a single task learning. Our multi-task learning model architecture is shown in Figure 4. The embedding layer outputs the triplet features, and the fully connected layer added after the embedding layer outputs the classification scores.

The Loss function is described as follows:

*1) **Triplet Loss Function:***

$$\mathcal{L}_{trp} = \sum [D_{a,p} - D_{a,n} + \alpha]_+ \tag{4}$$

Where $D_{a,p} = \|f(x_a) - f(x_p)\|_2^2$ is the square of Euclidean distance between the anchor feature and the positive feature, and $D_{a,n} = \|f(x_a) - f(x_n)\|_2^2$ is the square of Euclidean

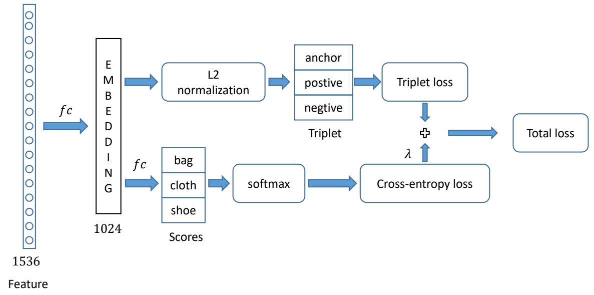

Fig. 4. Sub-network for Multi-task learning

distance between the anchor feature and the negative feature. α is the margin constant.

When the loss is minimized, the distance from the anchor feature to the positive feature becomes smaller, and the distance from the anchor feature to the negative feature becomes larger.

2) ***Classification Loss Function:*** we apply the softmax cross-entropy loss.

The softmax layer:

$$h(z_j) = \frac{e^{z_j}}{\sum_{k=1}^{n} e^{z_k}} \tag{5}$$

Where z is the output scores of the classification layer, and n is the number of the classes.

The cross-entropy loss:

$$\mathcal{L}_{cls} = -\frac{1}{m}[\sum_{i=1}^{m}\sum_{j=1}^{n} y_j^{(i)} \log h(z_j^{(i)}) + (1-y_j^{(i)}) \log(1-h(z_j^{(i)}))] \tag{6}$$

Where m is the number of the mini-batch size.

The total loss function is the weighted addition of these two loss functions:

$$\mathcal{L}_{total} = \mathcal{L}_{trp} + \lambda\mathcal{L}_{cls} \tag{7}$$

Where λ is the weight of the classification loss function.

IV. Experiment Results

We do some comparative experiments in the test data to show that self-attention and multi-task learning are effective. Due to the noise of the official ROI annotation, we also train a new detector on the training set, replacing the official ROI annotation, and the final result improves slightly in most cases. The detector is based on the Faster R-CNN method [15].

Finally, two heads are better than one, we trained multiple models and ensemble these models to get better results than a single model. In this ensemble task, we add the feature embedding results up for different models, then, compute distances between images to judge similarity. The greater the diversity of the models, the better the ensemble results. Changes in the backbone network, the training set, or the parameters of the model can increase the diversity of the models. It can be proved that ensemble can improve retrieval accuracy for JD fashion dataset greatly.

The experimental results are as follows:

Method	Cls Acc[a]	mAP	Det mAP[b]
Triplet	-	0.4372	0.4424
Triplet + Cls	0.9792	0.4585	0.4635
Triplet + Cls + Self-attention	0.9825	0.4626	0.4587
Ensemble	0.9860	0.5615	**0.5663**

[a] Cls Acc: Classification Accuracy. [b] Det mAP: mAP using detector ROI.

As shown in this table, we evaluate our model by mAP(Top-10). In this experiment, we take the model with triplet loss as our baseline, then, compare mAP results for multi-task task and self-attention task respectively. For our baseline, the mAP on JD fashion dataset can achieve 0.4372. In general, the accuracy rate of classification can achieve about 98%. So, we add a classifier after the feature embedding, in fact, it turns out to be higher retrieval accuracy about 0.4585. Self-attention also helps us improving the accuracy to 0.4626. Finally, we add a detector for the JD fashion dataset, the results indicate that this method is beneficial. After ensemble, the highest mAP we achieved was 0.5663.

V. Conclusion

In this paper, we propose a visual-based product retrieval model trained by a Multi-task Learning strategy and a Self-attention mechanism. We found that adding a classification loss can help to narrow the retrieval range effectively and the self-attention mechanism can capture correlative features between matching items. Our proposed model outperforms the baseline model and won the third place in the JD AI Fashion Challenge.

References

[1] Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NIPS. (2012)

[2] Babenko, A., Lempitsky, V.: Aggregating deep convolutional features for image retrieval. In: ICCV. (2015)

[3] L. Zheng, Y. Zhao, S. Wang, J. Wang, Q. Tian: Good practice in CNN feature transfer. In: arXiv:1604.00133. (2016)

[4] Lowe, D.G.: Distinctive image features from scale-invariant keypoints. IJCV (2004)

[5] N. Dalal and B. Triggs: Histograms of oriented gradients for human detection. In Proc. CVPR. (2005)

[6] R. Hadsell, S. Chopra, Y. Lecun: Dimensionality Reduction by Learning an Invariant Mapping. CVPR. (2006)

[7] Q. Cao, Y. Ying, P. Li: Similarity Metric Learning for Face Recognition. ICCV. (2013)

[8] F. Schroff, D. Kalenichenko, J. Philbin: FaceNet - A Unified Embedding for Face Recognition and Clustering. CVPR. (2015)

[9] Y. Wen, K. Zhang, Z. Li, Y. Qiao: A Discriminative Feature Learning Approach for Deep Face Recognition. ECCV. (2016)

[10] A. Gordo, J. Almazán, J. Revaud, and D. Larlus,:Deep image retrieval: Learning global representations for image search. ECCV. (2016)

[11] Nair, V., Hinton, G.E.: Rectified linear units improve restricted boltzmann machines. ICML, (2010)

[12] O. Marques, L.M. Mayron, G.B. Borba, H.R. Gamba,: Using visual attention to extract regions of interest in the context of image retrieval. Southeast Regional Conference :638-64. (2006)

[13] L. Itti, C. Koch, E. Niebur,: A Model of Saliency-Based Visual Attention for Rapid Scene Analysis. PAMI. (2002)

[14] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville,: Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. Computer Science:2048-2057. (2015)

[15] S. Ren, K. He, R. Girshick, J. Sun,: Faster R-CNN: towards real-time object detection with region proposal networks. NIPS. 39 (6) :91-99 (2015)