Matching User Photos to Online Products with Robust Deep Features

Xi Wang[†], Zhenfeng Sun[†], Wenqiang Zhang[†], Yu Zhou[‡], Yu-Gang Jiang[†]
[†]School of Computer Science, Fudan University, Shanghai, China
[‡]Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
[†]{xwang10, 14110240031, wqzhang, ygj}@fudan.edu.cn
[‡]zhouyu@iie.ac.cn

ABSTRACT

This paper focuses on a practically very important problem of matching a real-world product photo to exactly the same item(s) in online shopping sites. The task is extremely challenging because the user photos (i.e., the queries in this scenario) are often captured in uncontrolled environments, while the product images in online shops are mostly taken by professionals with clean backgrounds and perfect lighting conditions. To tackle the problem, we study deep network architectures and training schemes, with the goal of learning a robust deep feature representation that is able to bridge the domain gap between the user photos and the online product images. Our contributions are two-fold. First, we propose an alternative of the popular contrastive loss used in siamese deep networks, namely *robust contrastive loss*, where we "relax" the penalty on positive pairs to alleviate over-fitting. Second, a multi-task fine-tuning approach is introduced to learn a better feature representation, which not only incorporates knowledge from the provided training photo pairs, but also explores additional information from the large ImageNet dataset to regularize the fine-tuning procedure. Experiments on two challenging real-world datasets demonstrate that both the robust contrastive loss and the multi-task fine-tuning approach are effective, leading to very promising results with a time cost suitable for real-time retrieval.

Keywords

Visual Similarity, Image Retrieval, Deep Learning.

1. INTRODUCTION

Online shopping is extremely popular nowadays, where consumers normally use keywords to find their interested products on online shopping sites. However, textual keywords are not always sufficient. For instance, one may see a product with brand unknown and be interested in buying it. It would be very helpful if a shopping site supports visual search in this case, so that the user can take a photo and search visually the same products online.

Although we have seen such functions provided by companies like Amazon, Google and Alibaba, automatically matching user photos to online product images is still not easy and the room for improvement is huge. User photos used as search queries are usually captured by consumers with their mobile phones under uncontrolled settings, while the product images in online shops are often professionally photographed. On the one hand, products in user photos often have cluttered backgrounds or even partial occlusions. On the other hand, some online product images only contain a part of the products in order to show some details to the consumers. This extremely large domain gap makes the task highly challenging.

Key to the development of an effective product image retrieval system is the extraction of good feature representations. In contrast to hand-engineered descriptors like the SIFT, using deep neural networks to directly learn feature representations from raw data has recently demonstrated very impressive results in many areas. In particular, the convolutional neural networks (CNN) have produced strong performance on various vision tasks like object detection [9, 20], image segmentation [6] and video classification [26]. Different from visual recognition where a conventional deep CNN is normally sufficient, the problem to be tackled in this paper is a typical image matching setup where training images are normally provided in pairs (same/different product), not in classes (which can be converted to training pairs, but not vice versa). To deal with this issue, the siamese network architecture [7, 1, 12] provides a unique capability that can naturally rank the similarity between input image pairs by a contrastive loss function, which has been successfully used in several problems like face image matching [3].

In this paper, we adopt the siamese network (cf. Figure 3) for product image search. To bridge the large domain gap between user photos and online images, we propose a simple alternative optimization target called *robust contrastive loss*. A key difference is that in the training process we do not consider positive image pairs (containing the same product) that are visually too different. We underline that this is critical in our problem setting as penalizing too much on such pairs may incur over-fitting and poor generalization capability of the learned network. In addition, we propose a multi-task fine-tuning method to tune the parameters of the siamese network, which uses not only product images but also general images from the ImageNet corpus. We show that optimizing the network with multiple tasks, i.e., match-

ICMR'16, June 06-09, 2016, New York, NY, USA

DOI: http://dx.doi.org/10.1145/2911996.2912002

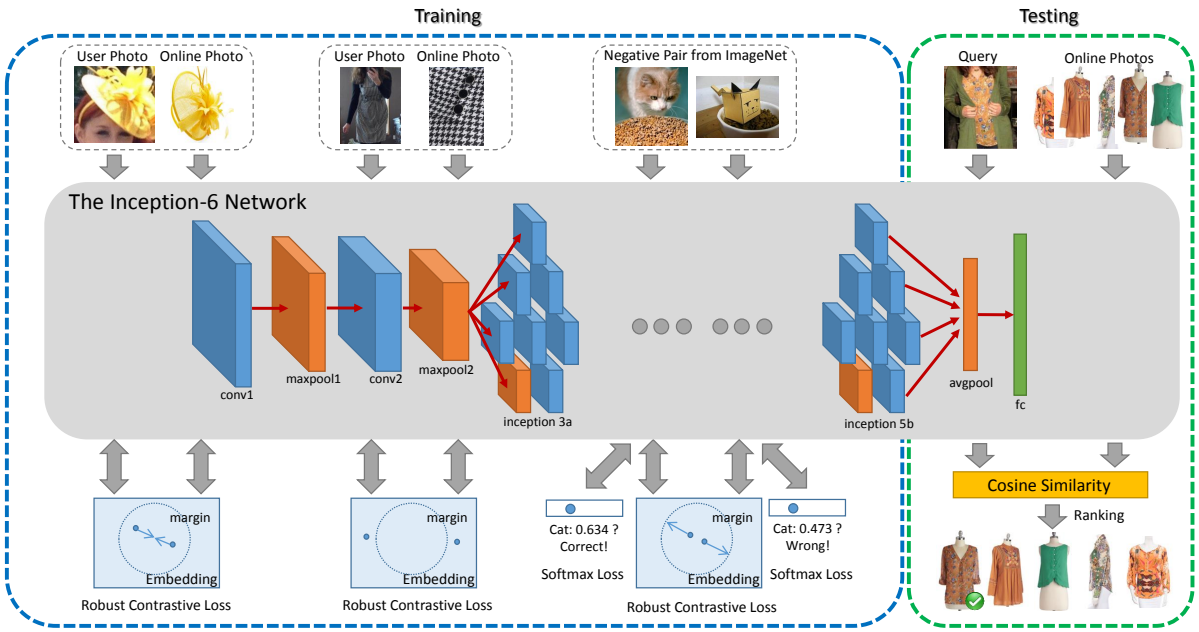

Figure 1: Illustration of the proposed product image retrieval approach. See texts for more discussions.

ing (both product and ImageNet images) and recognition (ImageNet images), can also help improve the performance of product retrieval.

Figure 1 illustrates the proposed approach. The core component of the framework is a deep neural network called Inception-6 [11]. In the training process (the left blue dashed box), each time a pair of images are given as input, each go through the Inception-6 network, and optimization is performed based on the robust contrastive loss. Positive pairs with distance larger than a pre-defined margin (e.g., the second pair in the figure) are not considered in the training process to avoid over-fitting. In addition, for image pairs from ImageNet (the third pair), we employ the proposed multi-task fine-tuning method, which not only optimizes the network parameters based on contrastive loss between the image pairs, but also tunes the network based on recognition outputs of each individual image, namely softmax loss. After training, in the online retrieval process (the right green dashed box), a feature representation can be quickly computed for a query image using the learned network, and retrieval results can be obtained by simply computing its similarity to the online product images. More details will be explained in Section 3.

With the proposed robust contrastive loss and multi-task fine-tuning, the learned network is able to produce superior retrieval accuracy on the challenging Exact Street2Shop Dataset [14] and the Alibaba Large-scale Product Image Dataset[1]. In the following, we first review related works, and then elaborate the proposed approach and discuss experimental results.

2. RELATED WORKS

Product Image Retrieval: There has been growing interests in developing product image retrieval systems in both research and industry. For instance, He et al. [8] proposed a visual search system for mobile devices using various features and indexing methods. Liu et al. [18] focused on a street-to-shop photo retrieval problem using a part alignment approach. Kalantidis et al. [13] proposed an approach to suggest multiple relevant clothing products based on some given images, using clothing recognition and segmentation techniques. In [1], Bell et al. adopted a similar pipeline to ours for similar product retrieval, but used the standard contrastive loss.

Instead of looking for similar items, Kiapour et al. [14] focused on a more challenging task called exact street-to-shop retrieval, in order to find exactly the same item in online shopping sites, which is very useful in practical applications. The same problem has also been investigated by Huang et al. [10] using ideas of visual attributes. This paper addresses the same problem of exact product retrieval with technical extensions tailored for tackling this specific problem.

Feature Representations: Extracting discriminative feature representations is critical to product image retrieval. For instance, Kuo et al. [16] proposed a semantic feature discovery approach through visual and textual clusters to derive semantically related feature representations. In contrast to the hand-engineered features, learning feature representations from raw data using the CNNs has demonstrated very impressive performance on many problems. Recent works in [5, 19] showed that a deep feature embedding trained to perform a specific task like object classification [21] can also generate competitive performance on a wide range of related tasks like fine-grained visual recognition, attribute detection, scene recognition and general image retrieval. Based on this fact, we choose to use the effective Inception-6 network [11] pre-trained on the ImageNet dataset [4] in our proposed approach.

Similarity Learning: Learning features using a siamese network [7] with contrastive loss is related to similarity or metric learning. In this category, the Online Algorithm for Scalable Image Similarity (OASIS) [2], which learns a bilinear similarity measure over hand-engineered descriptors, is one of the most successful approaches. More recently, Ki-

[1] https://tianchi.aliyun.com/competition/introduction.htm?raceId=231510

apour et al. [14] used a two-layer neural network to predict whether two features represent the same item. Using the conventional contrastive loss function, the authors designed an end-to-end neural network for metric learning [1]. Also based on the siamese network, Huang et al. [10] proposed a Dual Attribute-aware Ranking Network (DARN) for feature learning. Wu et al. [25] introduced a deep similarity learning approach for image retrieval called Online Multimodal Deep Similarity Learning (OMDSL) algorithm.

As label errors often exist in the training data, distance metric learning approaches that are more robust to training noise have been investigated (e.g., [17]). Different from [17], our work in this paper is based on the siamese network, using a novel loss function and a multi-task network fine-tuning scheme.

3. THE PROPOSED APPROACH

In this section, we first describe the siamese network and then introduce our proposed robust contrastive loss function, followed by the multi-task fine-tuning method with implementation details.

3.1 Siamese Network

Using the CNN to extract features from an image can be expressed as $\vec{X} = f(I, \theta)$, where the function f indicates the CNN structure, which computes a feature representation $\vec{X}$ for an input image I based on network parameters θ. Pre-training a CNN on a general image dataset such as the ImageNet can provide a reasonably good θ for the product retrieval problem, which has been verified in a recent work [14]. However, as the network is trained for general image classification tasks, the feature distance between two photos of the same product can be larger than that between photos of different products with certain visual similarity, which is undesirable for tackling the product image retrieval problem.

In order to extract a better feature representation that can correctly map the product image proximity to feature distance, we adopt a siamese network that contains two copies of the Inception-6 network with shared weights θ, as shown in Figure 2. The siamese network is often optimized with the conventional contrastive loss function [7]. Given a pair of input features $\vec{X}_p, \vec{X}_q$ and a binary label Y indicating whether the given pair is similar, the contrastive loss function can be written as:

$$L_{Y=1}(\vec{X}_p, \vec{X}_q) = ||\vec{X}_p - \vec{X}_q||_2^2,$$
$$L_{Y=0}(\vec{X}_p, \vec{X}_q) = \max(0, m^2 - ||\vec{X}_p - \vec{X}_q||_2^2),$$

where $Y = 1$ indicates that the input pair is similar (i.e., a positive pair containing the same product), otherwise $Y = 0$. The parameter m denotes a margin. If a negative pair can already be separated, i.e., with distance larger than the margin, the loss function should not give any further penalty. In the learning process, the shared network parameters θ are optimized to minimize the contrastive loss function.

3.2 Robust Contrastive Loss

Similar to the original contrastive loss function, the robust contrastive loss also runs over pairs of inputs. Given a pair of features $\vec{X}_p, \vec{X}_q$ and a binary label Y, the robust contrastive loss function is written as:

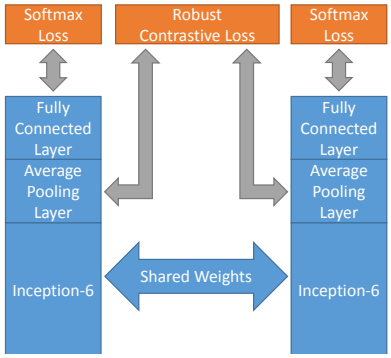

Figure 2: Siamese network architecture. In our experiments, the output of the average pooling layer is 1,024 dimensions, and the output of the fully connected layer is 21,841 dimensions.

$$L_{Y=1}(\vec{X}_p, \vec{X}_q) = \min(m^2, ||\vec{X}_p - \vec{X}_q||_2^2),$$
$$L_{Y=0}(\vec{X}_p, \vec{X}_q) = \max(0, m^2 - ||\vec{X}_p - \vec{X}_q||_2^2).$$

Different from the conventional contrastive loss, not only sufficiently separated negative pairs are not further penalized, the penalty of positive pairs which have large feature distances are also *constrained* in the robust contrastive loss function.

The major reason behind this extension is that positive pairs with little visual similarity often exist in real-world situations. As shown in Figure 3, such pairs create large gradients via the original contrastive loss function, which may incur over-fitting (and much lower testing performance) of the learned network model.

The original siamese network was extended in a recent work [1], where, as shown in Figure 2, the softmax loss functions are also incorporated to predict the object categories of the input images. The authors of [1] showed that using the additional softmax loss can also help prevent undesirable overfitting that may be caused by using only the contrastive loss function. In this work, we adopt the same pipeline with our robust contrastive loss. We show that, by combining our robust contrastive loss with the softmax loss, the generalization capability of the learned network can be greatly enhanced.

3.3 Multi-task Fine-tuning

Different from the improved siamese network described in [1], which uses the category information of product images in the softmax loss to regularize the fine-tuning procedure, we introduce another multi-task fine-tuning scheme here to incorporate additional training samples.

As aforementioned in Section 1, we employ the recently released Inception-6 [11] (a.k.a. BN-Inception) network structure. Compared with traditional network structures like AlexNet [15] and GoogLeNet [22], the batch normalization layers are added to the Inception-6 network, which can help the network achieve similar or even better accuracy with much fewer training steps. These extra layers can be viewed as a regularizer, eliminating the need of dropout layers in some cases. The ImageNet dataset[2] [4] which has 21,841 ob-

[2]ImageNet Fall 2011 release:
http://www.image-net.org/archive/stanford/fall11_whole.tar

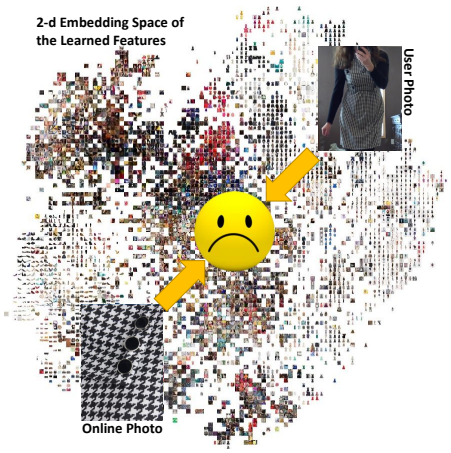

Figure 3: Positive image pairs with small visual similarity values incur "unwanted" large gradient via the original contrastive loss function, which may result in over-fitting.

ject categories is used to pre-train the Inception-6 network. Our pre-trained network achieves 0.315 top-1 accuracy over a randomly sampled validation dataset.

With the pre-trained model, a siamese network is created by simply making a copy of the Inception-6 and connecting both models with the robust contrastive loss. As shown in Figure 2, the robust contrastive loss function is placed after the 1,024-d average pooling layers. Two softmax loss functions are further adopted after the 21,841-d fully connected layers, measuring the recognition/classification errors of each individual input image.

The main difference between our multi-task fine-tuning scheme and the previous methods is that we use the images from ImageNet and their category labels to "regularize" the fine-tuning procedure. The benefits are two-fold. On the one hand, our proposed approach no longer requires annotated category labels for product images. On the other hand, existing product image datasets are often domain specific, and the product category number is usually much smaller than that of the ImageNet categories. Although ImageNet data is different from product images, we emphasize that using sufficient training data is more important when the softmax loss serves as an auxiliary regularizer to prevent over-fitting. As will be shown later, our scheme produces clearly better results than [1], which uses only the product images for the softmax loss.

More specifically, in order to form the training and validation sets for the proposed multi-task fine-tuning, we construct three kinds of input image pairs. The *positive pairs* can be created using the training product images based on the ground-truth labels. In addition, for each user photo in the training set, using the pre-trained Inception-6 network and cosine similarity, we can retrieve a set of online product images from the training set. *Hard negative pairs* can then be constructed based on the false positives in the top-retrieved results. Finally, the last part is *background negative pairs*. For this, we randomly sample images from different ImageNet categories to ensure that each pair contains two different objects. As only the images from the background negative pairs have category labels, we simply set the gradient from softmax loss to 0 for the other two types of training pairs. With a ratio of approximately $1:1:4$, the three kinds of image pairs are merged to construct the training and validation sets.

To fine-tune the siamese network, we set the learning rate to 5×10^{-5}, the momentum to 0.9 and the weight decay to 10^{-4} initially. The margin m in the robust contrastive loss function is set to 40.0 in our following experiments if there is no additional description. The size of each mini-batch is 128, which means 128 pairs of product images are processed by the network simultaneously. The training procedure lasts $12,000$ iterations maximally, and early stopping is used based on results on the validation set.

In addition, for the user photos, after resizing the smaller dimension to 256, we use the 224×224 center-cropped regions as the inputs. For the online product images, assuming we have a $w \times h$ bounding box in a $W \times H$ image, the edge length of the selected area is $min(max(w,h), min(W,H))$. After fine-tuning, given a new image, the outputs of the last average-pooling layer (1,024-d) of the trained Inception-6 network are used as its feature representation. With the features, we adopt Cosine similarity to measure the proximity of two images, because we found it can produce better results than Euclidean distance (about 3 absolute percentages higher). Notice that other similarity measures or advanced fast similarity search approaches (e.g., hashing) are all applicable on top of our features, which is beyond the scope of this work.

The proposed approach is named as *R. Contrastive & Softmax* in the following experiments. We also evaluate the performance of a network fine-tuned by robust contrastive loss function but without the softmax loss, which is named as *R. Contrastive*.

4. EXPERIMENTS

4.1 Experimental Setup

4.1.1 Dataset and Evaluation Measure

In most of the experiments, we adopt the newly released *Exact Street2Shop Dataset* [14], which focuses on the matching between real-world user (street) photos and product images of clothing items in online shopping sites. The dataset has two types of images: 1) *street photos*, which are usually captured by end-users under everyday uncontrolled settings, and 2) *shop photos*, which are photos of clothing items from online shopping sites, mostly taken under more professional settings.

Specifically, the Street2Shop dataset contains 20,357 street photos and 404,683 shop photos containing 204,795 distinct clothing items from 11 large categories. It also provides 39,479 pairs of exact matching items between the street and the shop photos. This makes the dataset an ideal testbed for our work. Following [14], the street-to-shop pairs are divided into training and testing sets with a ratio of approximately 4:1 in each category. In the experiments, a search query consists of a street photo with a bounding box indicating the location of the target item and its category label. Retrieval is then performed within the corresponding category, according to [14].

In addition, we adopt the *Alibaba Large-scale Product Image Dataset* (see footnote 1 for its URL) as the second dataset to evaluate the generalization capability of our proposed approach. As the official testing labels of this dataset are not publicly available, we randomly divide the street-to-

Category	#Queries	#Testing Images	AlexNet [14]	F. T. Similarity [14]	Inception-6	Contrastive & Softmax	R. Contrastive	R. Contrastive & Softmax
Dresses	3,292	169,733	22.2	37.1	31.0	44.0	55.7	**56.9**
Footwear	2,178	75,836	5.9	9.6	10.9	11.2	10.8	**13.1**
Tops	763	68,418	14.4	38.1	30.7	41.5	46.4	**48.0**
Outerwear	666	34,695	9.3	21.0	16.4	**21.5**	19.1	20.3
Skirts	604	18,281	11.6	**54.6**	39.1	40.6	49.7	50.8
Leggings	517	8,219	14.5	**22.1**	17.0	15.3	15.9	15.9
Bags	174	16,308	23.6	37.4	30.5	**46.6**	42.5	**46.6**
Eyewear	138	1,595	10.1	35.5	34.8	**39.1**	17.4	13.8
Pants	130	7,640	14.6	**29.2**	22.3	17.7	22.3	22.3
Belts	89	1,252	6.7	13.5	**24.7**	19.1	22.5	20.2
Hats	86	2,551	11.6	**38.4**	30.2	23.3	25.6	24.4
Overall	8637	404,528	14.66	28.97	24.37	30.86	35.88	**37.24**

Table 1: Top-20 retrieval accuracy (%) of our proposed approach and the alternative methods on the Exact Street2Shop Dataset. Notice that the F. T. Similarity method uses category-specific models and selective object proposals, while others use unified category-independent model and simple center crop of the images.

shop pairs (1,417 products and 92,572 manually annotated positive pairs) with a ratio of 2:1 (training vs. testing) without product overlap and report the top-20 retrieval accuracy. Different from the Street2Shop dataset where retrieval is performed within a category, we do not impose any category constraint on this dataset and each query is searched against all the online product images.

The retrieval performance is evaluated by top-k accuracy—a search is considered successful if at least one correct match can be found within the top-k returned results. Notice that only exactly the same product with the query is viewed as a correct match in our setup.

4.1.2 Alternative Methods for Comparison

To validate the effectiveness of our proposed approach, we compare with the following alternatives: 1) *AlexNet* network, which was used in [14], where the activations of the fully-connected layer FC6 (4,096-d) are used to form the feature representation. The network is trained on a subset of the ImageNet dataset [21] (1,000 categories). 2) *F. T. Similarity* [14], where category-specific two-layer neural networks are trained to predict whether two features extracted by the AlexNet represent the same product item. The method also uses the selective search algorithm [23] to extend the training and testing sets. 3) *Inception-6*, where the activations of the last average-pooling layer (1,024-d) of the pre-trained Inception-6 network are used to represent each image. 4) *Contrastive & Softmax*, which is based on the siamese network, where the traditional contrastive loss function and softmax loss function are used like [1]. Our proposed multi-task fine-tuning is also adopted. We compare with this alternative to demonstrate the effectiveness of our robust contrastive loss. Similarly, the outputs of the last average-pooling layer (1,024-d) are used as the feature representation. The cosine similarity is used in all these methods except 2, which uses pre-trained neural networks to predict whether two features represent the same product item.

4.2 Results on Street2Shop Dataset

Table 1 summarizes the results of both our approach and the compared methods on the *Exact Street2Shop Dataset*. Overall, our proposed approach achieves the highest average retrieval accuracy on the dataset, significantly outperforming F. T. Similarity by over 8 absolute percentages. Compared with the per-category results of F. T. Similarity, our proposed approach shows lower accuracy on a few product categories with less photos. This is mainly because our approach does not use category-specific models, and a unified model tends to bias towards the categories with more training data. Because our approach only uses the center crops of user photos for speed consideration, it shows better performance on categories with larger object, like dresses, tops and bags. We also observe that Inception-6 is clearly better than AlexNet, based on which we conjecture that the ideas of F. T. Similarity can be used in conjunction with Inception-6 for further improvements. However, we refrain from doing this, because training category-specific models requires more cost and is hard to be applied in practice as the category of the query is often unknown.

Comparing the results in the last three columns of Table 1, we can observe clear contributions from both robust contrastive loss (6.3 absolute percentages) and softmax loss (1.3 absolute percentages). In the following we discuss the results in detail under more experimental settings.

Figure 4 shows the top-20 retrieval accuracy of the last three approaches in Table 1, under different numbers of training iterations. The R. Contrastive approach achieves consistently worse performance than the R. Contrastive & Softmax approach, indicating that the softmax loss function is helpful. Compared with the Contrastive & Softmax approach, we see that using the proposed robust contrastive loss is very important. The main reason is that, as discussed earlier, some positive pairs with small visual similarities may incur over-fitting of the Contrastive & Softmax approach, which also explains its performance trend shown in the figure (increases at a faster pace earlier and then decreases). In addition, as shown in the figure, the performance of all the approaches saturates at some time points and then decreases, indicating that the early stopping strategy is also necessary to avoiding over-fitting.

Figure 5 further shows the top-20 accuracy under different margin parameters and training iterations, using the R. Contrastive & Softmax approach. Results indicate that an appropriate margin parameter is critical to retrieval performance. A small margin cannot incorporate sufficient training information, while a large margin may introduce "noises" that result in over-fitting. In our experiments, 40 is correctly predicted as a good option on a validation set.

Finally, we plot the top-k accuracy at different k in Fig-

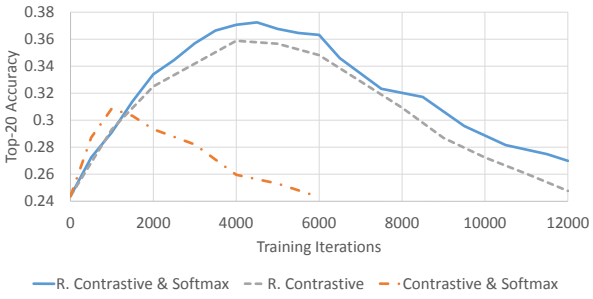

Figure 4: Top-20 accuracy on Street2Shop dataset under different loss functions and training iterations.

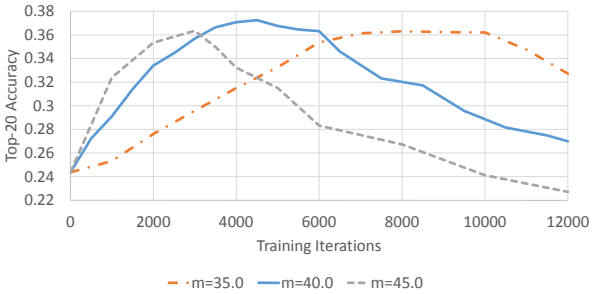

Figure 5: Top-20 accuracy on Street2Shop dataset under different margin parameters and training iterations.

ure 6. As expected, the retrieval accuracy of all the approaches increases with k. Nearly 60% of the queries can find their correct matches in the top-200 returned images.

Figure 7 shows some example retrieval results. Successful retrieval examples are shown in the top three rows, and the bottom three rows are failure cases. From the figure, we observe that the reasons behind a failure case include poor lighting environment and defocused image (e.g., the shoe example in the 4^{th} row), highly occluded target (e.g., the yellow shirt in the 5^{th} row), or simply the lack of visual characteristics of the specific product (e.g., the 6^{th} row). We also find that sophisticated patterns on the products are helpful as they are visually distinctive. For example, the query photo in the 3^{rd} row is visually very different from the shop photos in viewpoints, color and lighting, but can be correctly identified due to its unique pattern.

4.3 Results on Alibaba Dataset

We also report results on the *Alibaba Large-scale Product Image Dataset* to verify the generalization capability of the learned network. As shown in Table 2, we observe similar performance trend to that on the Street2Shop dataset. Both robust contrastive loss and softmax can improve around 2%, which again confirms the effectiveness of our approach. To evaluate the benefit from using the ImageNet samples, we also run the same R. Contrastive & Softmax approach with only samples from this dataset, where the softmax loss is evaluated using the class labels (604 classes) of the Alibaba images. An accuracy of 69.2% is achieved, which is nearly 3% lower than that using the ImageNet data (71.9%).

The improvement of robust contrastive loss is relatively

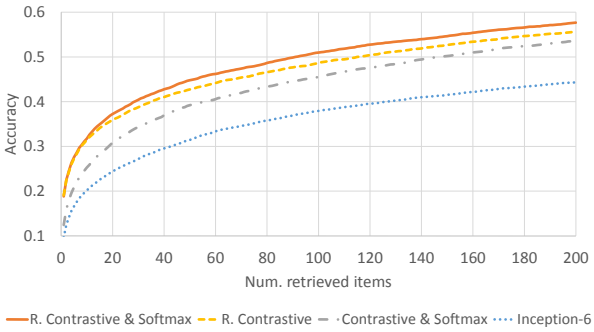

Figure 6: Top-k accuracy on Street2Shop dataset under different fine-tuning approaches and numbers of retrieved items.

Approach	Top-20 Accuracy
AlexNet	62.0
Inception-6	65.7
Contrastive & Softmax	69.5
R. Contrasive	69.3
R. Contrasive & Softmax	**71.9**

Table 2: Top-20 retrieval accuracy (%) of our proposed method and the state-of-the-art results on the Alibaba Large-scale Product Image Dataset.

smaller (but is still fairly significant) on this dataset because the labels are cleaner with less errors. In other words, for the Street2Shop dataset, in addition to the dissimilar positive pairs, wrongly labeled "false" positive pairs can also be excluded by our robust contrastive loss for improved results. Figure 8 further shows several example retrieval results.

4.4 Visualizing the Feature Embedding Space

To get a visual impression of the learned features, we use the t-SNE algorithm [24] to visualize the feature embedding space. First, the learned features extracted from the average pooling layer of the Inception-6 network are projected to 50 dimensions using the PCA algorithm to reduce the calculation cost. The t-SNE algorithm is then applied to further project the 50-d features to a 2-d space. To visualize the 2-d embedding space, we randomly select 5,000 images from both the street photos and the shop photos in the Street2Shop dataset and then place these photos to their corresponding 2-d locations.

As shown in Figure 9, the embedding space of the ImageNet-trained Inception-6 network tends to group the photos with similar view-angles together. In contrast, our approach based on robust contrastive loss of a siamese network tends to place the photos of the same products in close proximity, no matter they are street photos or shop photos, which is critical in real-world applications.

4.5 Efficiency

All the experiments are conducted on a server with an Intel i7-5820K CPU and two Nvidia GTX Titan X GPUs. Training a siamese network with the proposed loss function normally requires 4 to 8 hours, depending on the complexity of the dataset. Compared with model training that can be executed off-line, feature extraction time (model testing

Figure 7: Example retrieval results of our proposed approach in the Exact Street2Shop Dataset. Top three rows are successful queries (at least one correct match is found within top-20 returned images, marked by the green icon) and retrieval results, while the bottom three rows are failure queries with ground-truth matches and top retrieval results.

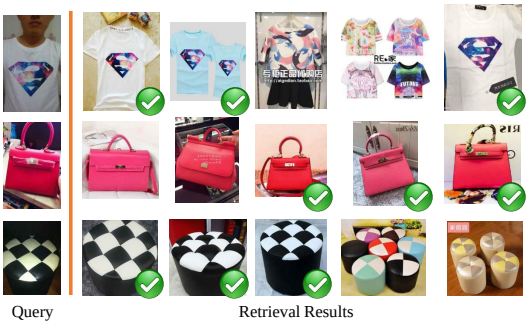

Figure 8: Example retrieval results of our proposed approach in the Alibaba Large-scale Product Image Dataset.

time) is way more important in real-time online search scenario. Using the Inception-6 network structure, the feature extraction process only costs 3.1 seconds per 1,000 images, which is sufficient for online search.

As aforementioned, once the features are extracted, many approaches like hashing can be adopted to accelerate the search process, which are not investigated in this work. With the brute-force exhaustive search, it costs around 150 ms for both feature extraction and retrieval of a query in the Street2Shop dataset.

5. CONCLUSION

We have presented a neural network based approach for the problem of matching a user photo to exactly the same product in online shopping sites. To alleviate the effect of label noise and prevent over-fitting caused by some positive pairs (that are visually different), we proposed a robust contrastive loss that automatically excludes such training samples in the network training process. A multi-task fine-tuning method was also proposed to harness extra data from the ImageNet with a softmax loss for improved results. Experiments on two real-world datasets have clearly demonstrated the effectiveness of our approach.

6. ACKNOWLEDGEMENT

This work was supported by China's National 863 Program (#2014AA015101), a grant from NSF China (#61572134) and a grant from STCSM (#16QA1400500). The *Alibaba Large-scale Product Image Dataset* was constructed and provided by Alibaba Group.

7. REFERENCES

[1] S. Bell and K. Bala. Learning visual similarity for product design with convolutional neural networks. *ACM Transactions on Graphics*, 34(4):98, 2015.

[2] G. Chechik, V. Sharma, U. Shalit, and S. Bengio. Large scale online learning of image similarity through ranking.

Figure 9: Visualization of image proximity using features from the ImageNet-trained Inception-6 network (left) and our proposed approach with the siamese network (right). See texts for more discussions.

The Journal of Machine Learning Research, 11:1109–1135, 2010.

[3] S. Chopra, R. Hadsell, and Y. LeCun. Learning a similarity metric discriminatively, with application to face verification. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 539–546. IEEE, 2005.

[4] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition*, 2009.

[5] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. *arXiv preprint arXiv:1310.1531*, 2013.

[6] C. Farabet, C. Couprie, L. Najman, and Y. LeCun. Learning hierarchical features for scene labeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1915–1929, 2013.

[7] R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality reduction by learning an invariant mapping. In *Computer Vision and Pattern Recognition*, 2006.

[8] J. He, J. Feng, X. Liu, T. Cheng, T.-H. Lin, H. Chung, and S.-F. Chang. Mobile product search with bag of hash bits and boundary reranking. In *Computer Vision and Pattern Recognition*, 2012.

[9] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015.

[10] J. Huang, R. S. Feris, Q. Chen, and S. Yan. Cross-domain image retrieval with a dual attribute-aware ranking network. In *International Conference on Computer Vision*, 2015.

[11] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv;1502.03167*, 2015.

[12] Y.-G. Jiang and J. Wang. Partial copy detection in videos: A benchmark and an evaluation of popular methods. *IEEE Transactions on Big Data*, PP(99), 2016.

[13] Y. Kalantidis, L. Kennedy, and L.-J. Li. Getting the look: clothing recognition and segmentation for automatic product suggestions in everyday photos. In *ACM International Conference on Multimedia Retrieval*, 2013.

[14] 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 *International Conference on Computer Vision*, 2015.

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, 2012.

[16] Y.-H. Kuo, W.-H. Cheng, H.-T. Lin, and W. H. Hsu. Unsupervised semantic feature discovery for image object retrieval and tag refinement. *Multimedia, IEEE Transactions on*, 14(4):1079–1090, 2012.

[17] D. Lim, G. Lanckriet, and B. McFee. Robust structural metric learning. In *International Conference on Machine Learning*, 2013.

[18] S. Liu, Z. Song, G. Liu, C. Xu, H. Lu, and S. Yan. Street-to-shop: Cross-scenario clothing retrieval via parts alignment and auxiliary set. In *Computer Vision and Pattern Recognition*, 2012.

[19] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. In *Computer Vision and Pattern Recognition Workshops*, 2014.

[20] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, 2015.

[21] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[22] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Computer Vision and Pattern Recognition*, 2015.

[23] K. E. Van de Sande, J. R. Uijlings, T. Gevers, and A. W. Smeulders. Segmentation as selective search for object recognition. In *International Conference on Computer Vision*. IEEE, 2011.

[24] L. Van der Maaten and G. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(2579-2605):85, 2008.

[25] P. Wu, S. C. Hoi, H. Xia, P. Zhao, D. Wang, and C. Miao. Online multimodal deep similarity learning with application to image retrieval. In *ACM International Conference on Multimedia*, 2013.

[26] Z. Wu, X. Wang, Y.-G. Jiang, H. Ye, and X. Xue. Modeling spatial-temporal clues in a hybrid deep learning framework for video classification. In *ACM International Conference on Multimedia*, 2015.