Cross domain Image Retrieval

Shikha Agarwal
University of Massachusetts, Amherst
shikhaagarwa@cs.umass.edu

Sneha Bhattacharya
University of Massachusetts, Amherst
snehabhattac@cs.umass.edu

Abstract

The goal of the project is to perform cross domain image retrieval for fashion datasets, that is given an image of a consumer fashion item, retrieve the best matching shop images. This is a challenging problem due to the large discrepancy between online shopping images, usually taken in ideal lighting/pose/background conditions, and user photos captured in uncontrolled conditions. To achieve this we train a Vgg19 based siamese network architecture with triplet ranking loss. We have obtained a top k accuracy of 55.3% with search space 1k and 22.2% with search space 10k. The code is available at `https://github.com/ShikhaAgarwal/image_retrieval`.

1. Introduction

Cross-domain image retrieval is an important task for many practical applications. For example, mobile product image search aims at identifying a product, or retrieving similar products from the online shopping domain based on a photo captured in unconstrained scenarios by a mobile phone camera. Clothes recognition algorithms are often confronted with three fundamental challenges when adopted in these real-world applications. First, clothes often have large variations in style, texture, and cutting, which confuse existing systems. Second, clothing items are frequently subject to deformation and occlusion. Third, clothing images often exhibit serious variations when they are taken under different scenarios, such as selfies vs. online shopping photos.

In this paper, we address the problem of cross-domain image retrieval by taking clothing products as a concrete use case. Given an offline clothing image from the 'consumer' domain, our goal is to retrieve the same or similar clothing items from a large-scale gallery of professional online shopping images called 'shop' images. We are using 'Consumer-to-shop' DeepFashion dataset. For this problem, we divided our approach into two:

1. obtaining features for shop and consumer images
2. finding similar shop images for each given consumer image using the above features

In our baseline method we obtain features using the pre-trained VGG-19 network. In the main experiment, we obtain features from the Siamese network that we trained completely using triplet loss. In both our experiments, to retrieve similar shop images, we used k nearest neighbours along with cosine/euclidean distance metric. Top-k retrieval accuracy is adopted to measure the performance of fashion retrieval, such that a successful retrieval is counted if the exact fashion item has been found in the top-k retrieved results.

2. Related Work

In [3] the authors introduce a Dual Attribute-aware Ranking Network (DARN) for retrieval feature learning. DARN simultaneously integrates semantic attributes with visual similarity constraints into the feature learning stage, while at the same time modeling the discrepancy between domains. The model was trained on their own collected dataset which was obtained by crawling customer review websites.

[2] is a a large-scale clothes dataset with comprehensive annotations. A network called Fashionnet is introduced which learns clothing features by jointly predicting clothing attributes and landmarks. The estimated landmarks are then employed to pool or gate the learned features. A ranking loss is also used to rank the nearest shop images.

In [1] a deep ranking model that employs deep learning techniques to learn similarity metric directly from images is deployed. A convolutional neural network architecture with shared weights and a ranking loss function is used to learn the similarity metric. in [5] Pairwise ranking model is a widely used learning-to-rank formulation. It is used to learn image ranking models. Generating good triplet samples is a crucial aspect of learning pairwise ranking model.

3. DataSet

We used the DeepFashion dataset which is a large-scale clothes dataset with comprehensive annotation. It con-

tains over 800,000 images, richly annotated with massive attributes, clothing landmarks, and correspondence of images taken under different scenarios including store, street snapshot, and consumer. For our project we use Consumer-to-Shop part of of the dataset which contains over 200,000 consumer-to-shop image pairs. The dataset was split into training and validation sets, out of which we used 100,000 shop and consumer image pairs were used to train our model. The validation set had 50,000 shop and consumer image pairs.

4. Experiments

4.1. Pre-processing

The DeepFashion dataset also contains bounding box coordinates for the clothes in the images. We used these coordinates to crop the original image. We also scaled the image to 224 by 224 before feeding to the network.

4.2. Baseline

We performed the baseline experiment by using a trained VGG19 model and extracting the features for the shop and consumer images using k-nearest neighbours along with cosine/euclidean distance metric. We then retrieved the top k similar shop images for a given consumer image to evaluate the performance of model on the validation set.

4.3. Siamese Approach

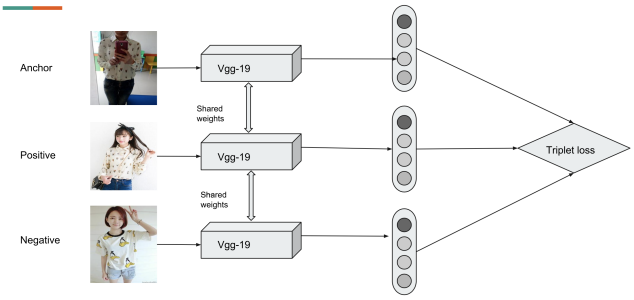

Figure 1. The Siamese network with triplet loss architecture

The triplet-based ranking loss is used to constrain the feature similarity of image triplets. Denoting a and b the features of a consumer image and its corresponding shop image respectively, the objective function of the triplet ranking loss is:
$TripletLoss(a, b, c) = max(0, dist(a, b) - dist(a, c))$
where c is the feature of the dissimilar shop image. Dist(,) represents the feature distance, e.g ., Euclidean distance, and m is the margin, according to the average feature distance of image pairs. Basically, this loss function imposes that the feature distance between a shop-consumer clothing pair should be less than that of the consumer image and any other dissimilar shop image.

As the features of shop and consumer images come from two different sub-networks, this loss function can be considered as the constraint to guarantee the comparability of features extracted from those two sub-networks, therefore bridging the gap between the two domains.

In each triplet, the first two images are the consumer-shop pair, with the third image randomly sampled from the shop images training pool. Several such triplets construct a training batch, and the images in each batch are sequentially fed into their corresponding sub-network according to their types. We then calculate the gradient of the loss function and the gradient is then propagated back into the network.

We used a Vgg19 network pre-trained on the Imagenet dataset as our backbone network for both the baseline and the Siamese approach. We retained the layers till the first fully connected layer so the output of the network is a feature vector of size 4096.

5. Results

We present the results that we obtained from the two methods. We report the top-k accuracy, that is the number of hits from the top-k retrieved results for our test images. Here we have set k as 50. Our siamese network with triplet loss outperformed the baseline method. We also notice that as we increase the retrieval pool the top-k performance drops. We also report some qualitative results.

Method	Retrieval size = 1k	Retrieval size = 10k
baseline	40%	11.1%
siamese	55.3%	22.2%

Figure 2. Left: Query image, Second: Correct matched image, Third: Incorrect matched image

6. Conclusion

In this project we demonstrated how Siamese network with triplet loss can be used to model the similarity of images from different domains. On doing a qualitative analysis we find that the network is able to correctly identify

Figure 3. Top left: Query image, top Second to bottom first: Correct Matched images, Last: Incorrect matched image

texture of image pairs as well as other attributes in some cases. For example in figure 2 the texture and color were correctly identified and in figure 3 the unique collar design. In figure 3, although the third and fourth shop images were occluded, it identified the matching pattern. A lot of the images were still incorrectly identified. The obvious reason could be that just learning the similarity is not enough, and the network should learn to identify different attributes as well.

On further analysis, the retrieval performance decreases as we increase the retrieval pool size(shop images). In the paper DeepFashion, the retrieval pool size is not mentioned in their results. Hence we cannot compare our network performance to the results mentioned in the paper.

7. Future Work

The DeepFashion dataset provides other information like attributes such as texture, category, fabric, style and shape. It also provides clothing landmarks, like location of sleeves. We can make our retrieval system more robust by increasing the siamese network to learn these information along with ranking. We can further explore different architecture to obtain features like in FashionNet.

As the search space increases, time complexity to compute k-nearest neighbour increases. We can then use techniques like K-means to reduce the search space and hence perform faster retrieval.

References

[1] Wang, Jiang and Song, Yang and Leung, Thomas and Rosenberg, Chuck and Wang, Jingbin and Philbin, James and Chen, Bo and Wu, Ying *Learning Fine-grained Image Similarity with Deep Ranking*. 2014 IEEE Conference on Computer Vision and Pattern Recognition

[2] Liu, Ziwei and Luo, Ping and Qiu, Shi and Wang, Xiaogang and Tang, Xiaoou*DeepFashion: Powering Robust Clothes Recognition and Retrieval with Rich Annotations.* The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)

[3] Junshi Huang and Rogério Schmidt Feris and Qiang Chen and Shuicheng Yan *Cross-domain Image Retrieval with a Dual Attribute-aware Ranking Network* 2015 IEEE International Conference on Computer Vision (ICCV)

[4] Karen Simonyan, Andrew Zisserman *Very Deep Convolutional Networks for Large-Scale Image Recognition*

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