CS272 Computer Vision II: homework1:Image Retrieval

Yang Hui
2020233290
yanghui1@shanghaitech.edu.cn

Abstract

In this assignment, I implemented the paper titled 'Cross-domain fashion image retrieval' by using python.

1. Introduction

Image Retrieval involves user domain and shop domain, which is used to match a picture from user domain in shop domain. We can intuitively understand the meaning of image retrieval by recalling Taobao's function of searching for products with an image.

2. Method

The overall network architecture adopts a three streams Siamese architecture. Each stream contains a convolutional layer of resnet50, maximum pooling, a fully connected layer and l2 normalization.

2.1. Data Preprocessing

We use the data set of the 'Consumer-to-shop Clothes Retrieval' module in DeepFashion to train our network. In this data set, each folder includes a picture from the user domain and a corresponding picture in the shop domain.However,the input of our network is three images: query image, relevant image, non-relevant image.In the same folder, we set the image from the user domain as the query image and the image from the shop domain as the relevant image.In the original paper, the author choose the non-relevant image from a different class which is among the 25 non-relevant images closest to the query.

For non-relevant images, we adopted the original author's idea, but the implementation process is different. We use the image from shop domain in the folder adjacent to the current query image as the non-relevant image. Because the non-relevant image and the query image are not the same clothes, but they are both of the same type. Therefore, their styles are roughly the same, so it is approximated that they are very close.For example:There are two folders. Folder1:'DRESSES/SleevelessDress/00001/comsumer01.jpg'(query), 'DRESSES/SleevelessDress/00001/shop01.jpg'(relevant), Folder2:'DRESSES/SleevelessDress/00002/comsumer01.jpg', 'DRESSES/SleevelessDress/00002/shop01.jpg'(non-relevant).

2.2. three streams Siamese

The input of the network is three images: query image, relevant image, non-relevant image. Each image is an input of a stream, we can regard each stream as the feature extractor of the image, and then measure the distance of the feature vectors to judge whether two pictures are similar.

The network structure of three streams Siamese is shown in Figure1.We use the pre-trained model of resnet50 to initialize the weights of conv.

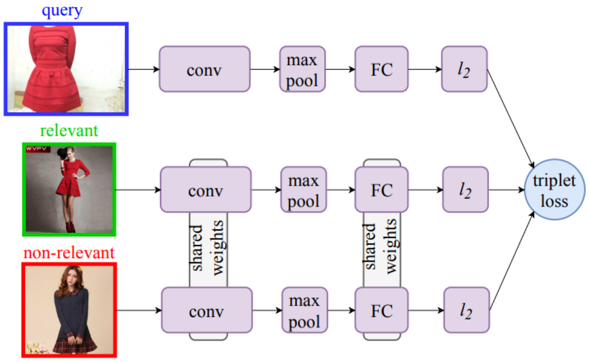


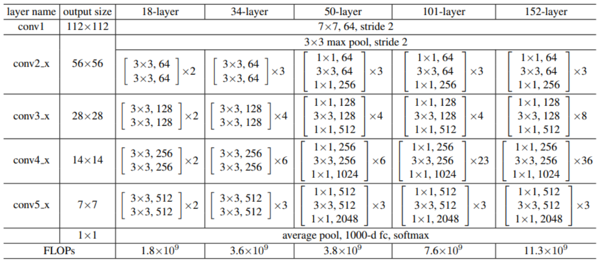
Figure 1. three streams Siamese

2.3. Resnet50

The structure of Resnet50 is shown in Figure 2.In the process of implementation, based on the structure of resnet50, we replaced the final average pool with a map pool, changed the input dimension of fc to 2048, and finally normalized embeddings.

2.4. Triplet loss function

We adopted triplet loss function.The formula of the loss function is as follows:



layer name	output size	18-layer	34-layer	50-layer	101-layer	152-layer
conv1	112×112	7×7, 64, stride 2				
conv2_x	56×56	3×3 max pool, stride 2				
		$\begin{bmatrix}3\times3, 64\\3\times3, 64\end{bmatrix}\times2$	$\begin{bmatrix}3\times3, 64\\3\times3, 64\end{bmatrix}\times3$	$\begin{bmatrix}1\times1, 64\\3\times3, 64\\1\times1, 256\end{bmatrix}\times3$	$\begin{bmatrix}1\times1, 64\\3\times3, 64\\1\times1, 256\end{bmatrix}\times3$	$\begin{bmatrix}1\times1, 64\\3\times3, 64\\1\times1, 256\end{bmatrix}\times3$
conv3_x	28×28	$\begin{bmatrix}3\times3, 128\\3\times3, 128\end{bmatrix}\times2$	$\begin{bmatrix}3\times3, 128\\3\times3, 128\end{bmatrix}\times4$	$\begin{bmatrix}1\times1, 128\\3\times3, 128\\1\times1, 512\end{bmatrix}\times4$	$\begin{bmatrix}1\times1, 128\\3\times3, 128\\1\times1, 512\end{bmatrix}\times4$	$\begin{bmatrix}1\times1, 128\\3\times3, 128\\1\times1, 512\end{bmatrix}\times8$
conv4_x	14×14	$\begin{bmatrix}3\times3, 256\\3\times3, 256\end{bmatrix}\times2$	$\begin{bmatrix}3\times3, 256\\3\times3, 256\end{bmatrix}\times6$	$\begin{bmatrix}1\times1, 256\\3\times3, 256\\1\times1, 1024\end{bmatrix}\times6$	$\begin{bmatrix}1\times1, 256\\3\times3, 256\\1\times1, 1024\end{bmatrix}\times23$	$\begin{bmatrix}1\times1, 256\\3\times3, 256\\1\times1, 1024\end{bmatrix}\times36$
conv5_x	7×7	$\begin{bmatrix}3\times3, 512\\3\times3, 512\end{bmatrix}\times2$	$\begin{bmatrix}3\times3, 512\\3\times3, 512\end{bmatrix}\times3$	$\begin{bmatrix}1\times1, 512\\3\times3, 512\\1\times1, 2048\end{bmatrix}\times3$	$\begin{bmatrix}1\times1, 512\\3\times3, 512\\1\times1, 2048\end{bmatrix}\times3$	$\begin{bmatrix}1\times1, 512\\3\times3, 512\\1\times1, 2048\end{bmatrix}\times3$
	1×1	average pool, 1000-d fc, softmax				
FLOPs		1.8×10^9	3.6×10^9	3.8×10^9	7.6×10^9	11.3×10^9

Figure 2. resnet

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

q is query image,**p** is relevant image,**n** is non-relevant image. d() means Euclidean distance. The value of loss is determined by two kinds of distances which are the distance between the query image and the relevant image, and the distance between the query image and the non-relevant image. Our ultimate training goal is to make the loss value as small as possible, that is to say, to make the distance between the query image and the relevant image as small as possible, and to make the distance between the query image and the non-relevant image as large as possible.

3. Result

During the test, we are given a picture from the user domain and 5 pictures from the shop domain. Another image is regarded as a non-relevant image.

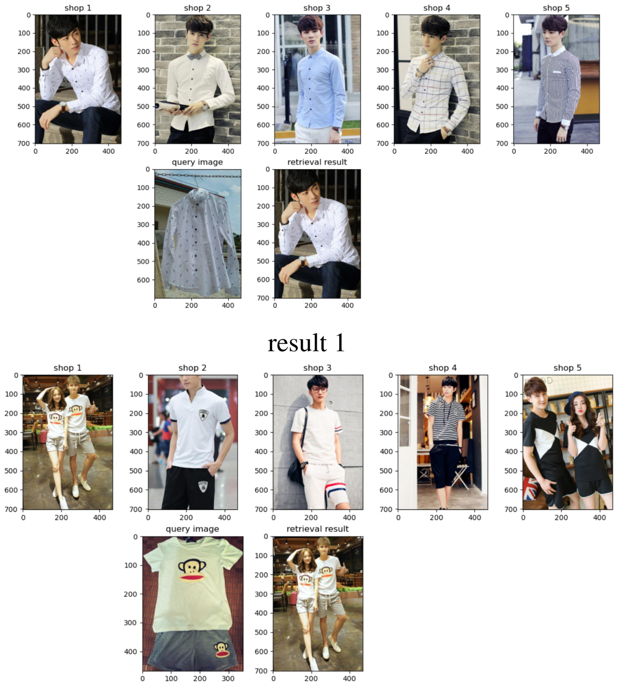
The result is shown in Figure3. The image in the first line is from the shop domain, and the second line shows the query image and the matched shop image.

4. Discussion

To be honest, this network does not guarantee that the correct image will be matched every time.At first, the batch size of the network was set to 16. When I used the GPU to train the model, an out-of-memory error occurred. After adjusting the batch size to 9, the model can be trained normally. Due to time constraints, our epoch is set to 3. I think the results can be optimized by the following methods: Choosing better hardware, optimizing the network structure, and adjusting the parameters,training the model with a lot of suitable data

References

[1] Bojana Gajic,Ramon Baldrich. Cross-domain fashion image retrieval. CVPR Workshops 2018.



result 1

result 2

Figure 3. Result