Convolutional Neural Network for Retrieval of Supervised Footwear Images

Yong Wang [a], Dongdong Shen [b] and Ying Wang

School of computers, Guangdong University of Technology, Guangzhou 510003, China

[a] wangyong@gdut.edu.cn
[b] Corresponding author: 393146198@qq.com

Abstract. With the progress of science, the traditional image retrieval technology no longer applies in terms of accuracy and retrieval speed. The emergence of big data and GPU has laid a solid foundation for the development of deep learning. However, the emergence of big data also means that the data is noisy and imperfect problems are more obvious, so the data processing and research is necessary. The experimental results show that the targeted method can also improve the image retrieval results under noisy and uncertain data.

Key words: Weak supervision; image retrieval.

INTRODUCTION

Because of the heat of artificial intelligence, deep learning [1] has also attracted the attention of the people, so what is the deep learning? Deep learning is the internal law and presentation level of learning sample data, and the information obtained in these learning processes can be of great help to the interpretation of words, images and sounds. Its ultimate goal is to enable machines to be able to analyze learning capabilities like humans and to recognize data such as text, images and sounds.

Currently, there is no application of image retrieval, mainly due to the lack of accuracy and insufficient data. Professional shoes database, so this requires first to build a database and then select the appropriate method and model based on the database. This article chose to climb down the pictures of the shoe class from jingdong. There are a total of 500,000, each of which has about 10 pictures, the first of which is about the front of the shoe, and the others are pictures of the sides and the models.

Because there is no manpower to do the bounding box. [3] so I decided to use weak supervision and training, shooting for shoes is not fixed, all angles, so adopting vlad chaotic characteristic of [4] - don't care about the local characteristics of the spatial location, can be further decoupling global spatial information, the geometric transform has a good robustness. The latter three are probably the interior of the shoe (insole). These are very loud noises, so this paper decides to remove them.

RELATED TECHNOLOGIES

Convolutional Neural Network.

Convolution neural network [5]is a special kind of neural network model of deep, its particularity embodied in two aspects, on the one hand, it's the connections between neurons are all connected, on the other hand some weights of connections between neurons in the same layer is Shared (i.e., the same). It's not all connections and weights of Shared network structure to make it more similar to the biological neural networks, reduces the complexity of the network model (for the deep structure of it is difficult to learn, this is very important), reduce the number of the weights. This article USES VGG-16[6].

Weak Supervision

Weak supervised learning includes: semi-supervised learning (in general, the supervised learning sample has an example and its corresponding tag. But there are times when such oversight is not available. Some of the sample tags missing are semi-supervised. Positive and unlabeled learning (sometimes we only know that some samples are positive samples and other sample tags are unknown). Besides, there are many examples of learning and learning paradigms such as multi-marking learning.

NetVlad

Vlad

VLAD can understand is combine BOF and fisher vector an optimized feature representation method.Given a similar with K-means the algorithm gets the codebook. $\{u_i, \mathrm{i}=1,\ldots,N\}$ .A collection of local descriptors extracted from an image. $X=\{x_t, t=1,\ldots,T\}$ (For example,sift) .VLAD includes the following steps:

(1) Assign each local descriptor to the nearest codebook:

$$NN(x_t)=arg\ min\,||\,x_t-u_i\,|| \tag{1}$$

(2) (3) To calculate Voronoi

$$V_i=\sum_{x_t:NN(x_t)=u_i} x_t-u_i \tag{2}$$

NetVlad

The hard-assignment operation in the original VLAD method is not negligible (the first step in VLAD), so it cannot be directly embedded into CNN network and participate in the error back propagation. NetVlad [7] is to use the softmax function will be this hard - the assignment into soft - the assignment operation, using 1 x1 convolution and softmax function to get the local characteristics belong to the probability of each center/weight, and then assign it to have the greatest probability/weight center.

BASED CONVOLUTIONAL NEURAL NETWORK FOR RETRIEVAL OF SUPERVISED FOOTWEAR

Data Preprocessing

Because there is no professional shoe database, can only climb from the Internet. Each kind of shoes from the jingdong first climbed 10 images, including the effect of the show, models wearing shoes, images in different environment, shoes materials, soles, insoles, and so on, a total of 500000 copies.

Remove the Noise

Put the shoes all data down into imagenet network, to get each pictureembedding, using Euclidean distance to each kind of 10 picture middle distance is greater than the first m (0.1) is a parameter initial value set as the image clear. There were 150,000 pictures, and 350,000 lefts, as our training set. We then randomly selected 70, 000 images for the test set.

Modify the VGG-16

Modify the Conv5_3

The original VLAD. Given the local features of N D dimensions x_i as input, and K class centers c_k , V is K*D dimensions, if the nearest class center x_i is c_k ,then $a_k(X_i)$ is 1or 0.

$$V(j,k) = \sum_{i=1}^{N} a_k(X_i)(x_i(j) - c_k(j)) \tag{3}$$

However, the hard-assignment operation is not negligible (the first step in vlad), so it cannot be directly embedded into CNN network and participate in the error back propagation. Therefore, the method of NetVlad is to use the 1x1 convolution and softmax function to obtain the probability/weight of the local feature to each center point, and then assign it to the center point with the maximum probability/weight.

$$\overline{a_k}(X_i) = \frac{e^{-\partial\|x_i - c_k\|^2}}{\sum_{k'} e^{-\partial\|x_i - c_{k'}\|^2}} \rightarrow \overline{a_k}(X_i) = \frac{e^{w_k^T x_i + b_k}}{\sum_{k'} e^{w_k^T x_i + b_{k'}}} \tag{4}$$

$$w_k = 2\partial c_k , b_k = -\partial\|c_k\|^2 \tag{5}$$

$$V(j,k) = \sum_{i=1}^{N} \frac{e^{w_k^T x_i + b_k}}{\sum_{k'} e^{w_k^T x_i + b_{k'}}} (x_i(j) - c_k(j)) \tag{6}$$

Class is center of each image is obtained by imagenet embedding, and then set K = 64 is obtained by K - means cluster center.

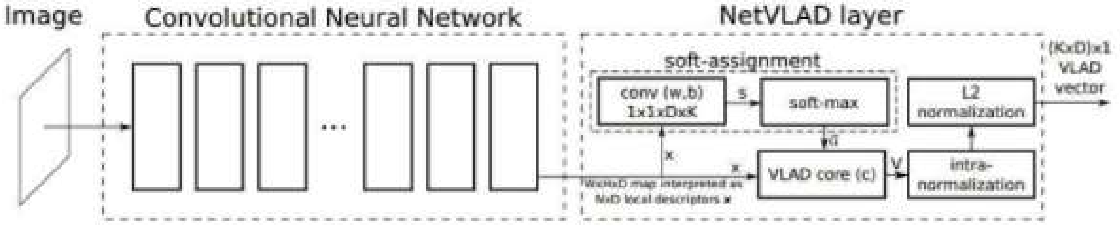

FIGURE 1. NetVlad-CNN

Modify the Loss

Because there is no positive sample of the weak supervision, so it is the comparison of each sample in the positive sample set and this image, and find the positive sample that is the smallest in Euclidean distance as a positive sample.Set q is a picture of the training set, $\{p_i^q\}$ Is the corresponding positive sample set, $\{n_j^q\}$ It's a negative sample set.

$$L = \sum_j l(min_i d^2(q, p_i^q) + m - d^2(q, n_j^q)) \tag{7}$$

p_{i*}^{q} is the best positive sample. And also, the distance to the positive sample is less than the distance from all the negative samples.

$$d(q, p_{i*}^{q}) < d(q, n_{j}^{q}) \tag{8}$$

So in the training group L, m is the initial parameter of 0.1.

$$L = \sum_{j} l(min_{i} d^{2}(q, p_{i}^{q}) + m - d^{2}(q, n_{j}^{q})) \tag{9}$$

RESULTS AND ANALYSIS

Results

This experimental hardware is configured with i7 processor and NVIDIA Corporation GM107GL GUP.

Software configuration is Ubuntu16.0.4 system Matlab2016 and MatConvNet and CUDA 8.0, Cudnn6.5.

In this paper, the original data (datebase1) and the data (datebase2) were tested in the original vgg-16 and the Netvlad-CNN in this paper. The accuracy was as follows.

Netvlad-CNN (-*-)

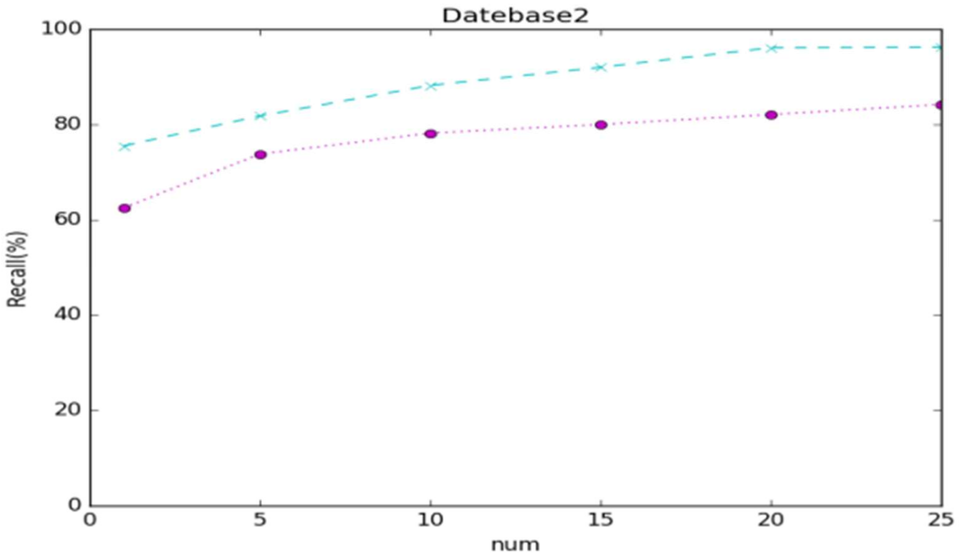

FIGURE 2. Denoising data result.

Analysis

In different databases and the comparison of the experimental results of different methods, we obviously found that under the high-quality database retrieval accuracy is about 20% higher than the original database, and NetVlad method is better than general method under noisy database raised 10% accuracy. It is necessary to optimize data and modify training methods for data.

CONCLUSION

With the rapid development of image retrieval technology, we should not only pay attention to the progress of technology, but also update the database. It is also unrealistic to find a universal expression (at least in-depth learning). So we should find the optimal solution for a particular problem. It's not a general solution.

REFERENCES

1. MH Nguyen, L Torresani, lTF De, C Rother.Weakly supervised discriminative localization and classification: a joint learning process. [C] IEEE International Conference on Computer Vision. 2009, 30(2):1925-1932
2. H Jegou, M Douze , C Schmid , P Perez .Aggregating local descriptors into a compact image representation.[J] Computer Vision & Pattern Recognition ,.2010 , 238 (6)
3. AS Razavian , H Azizpour , J Sullivan , S Carlsson. CNN Features off-the-shelf: an Astounding Baseline for Recognition. [J] Computer Vision & Pattern Recognition Workshops. 2014 :512-519
4. J Donahue, Y Jia, O Vinyals , J Hoffman , N Zhang .DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition. [J]Computer Vision and Pattern Recognition (cs.CV).2013.6
5. Karen Simonyan, Andrew Zisserman.Very Deep Convolutional Networks for Large-Scale Image Recognition[C]. Computer Vision and Pattern Recognition.2014.4
6. Karen Simonyan, Andrew Zisserman.Very Deep Convolutional Networks for Large-Scale Image Recognition[C]. Computer Vision and Pattern Recognition.2014.4
7. R Arandjelovic, P Gronat , A Torii , T Pajdla , J Sivic.NetVLAD: CNN architecture for weakly supervised place recognition.[C] IEEE 2015 , PP (99) :1-1