Feature Extraction and Image Retrieval Based on AlexNet

Zheng-Wu Yuan[a], Jun Zhang*[b]
[a] College of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing, China 2439468996@qq.com; [b]College of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing, China 332992179@qq.com

ABSTRACT

Convolutional Neural Network is a hot research topic in image recognition. The latest research shows that Deep CNN model is good at extracting features and representing images. This capacity is applied to image retrieval in this paper. We study on the significance of each layer and do image retrieval experiments on the fusion features. Caffe framework and AlexNet model were used to extract the feature information about images. Two public image datasets, Inria Holidays and Oxford Buildings, were used in our experiment to search for the influence of different datasets. The results showed the fusion feature of Deep CNN model can improve the result of image retrieval and should apply different weights for different datasets.

Keywords: Content based image retrieval; deep learning; Convolutional Neural Network; feature extraction; feature fusion

1. INTRODUCTION

Community network and mobile computing have brought more new data, new problems and new applications, such as Flickr and Facebook. These two sites have piled millions or billions of images and the users provide much information to these images. These network images provide many training data to the research work and there are a huge number of visual patterns in them, such as repeated images, objects, structured visual units. Data mining and machine learning algorithms can generalize structured pattern from this redundant information. With the rapid development of computer hardware, in 2006, Hinton [1] Geoffrey proposed a fast learning method of deep belief network to solve the problems of the deep learning model training. Krizhevsky Alex [2] and so on in ILSVRC 2012 get the state of the art in image recognition also shows the deep model has great progress in image classification. In 2014, Babenko Artem [3] proved that Deep CNN (Convolutional Neural Network) [4] has a great advantage in image retrieval and the best feature is not on the output of the model but the sixth layer. Previous researches mainly use the deep neural network as a Black Box and its output as a classifier. When try to apply this model to image retrieval, it is necessary to study the significance of the image features of different layer of the deep learning model.

2. MODEL STRUCTURE ANALYSIS

2.1 Research Model

AlexNet [2], as a representative of the Deep neural network, is an 8-layer model as Fig 1: it contains 5 convolution layers and 3 fully-connected lays. Since the deep structure and plenty parameters in the model, it gets more features from original data than traditional CNN.

Eighth International Conference on Digital Image Processing (ICDIP 2016), edited by Charles M. Falco, Xudong Jiang
Proc. of SPIE Vol. 10033, 100330E · © 2016 SPIE · CCC code: 0277-786X/16/$18 · doi: 10.1117/12.2243849

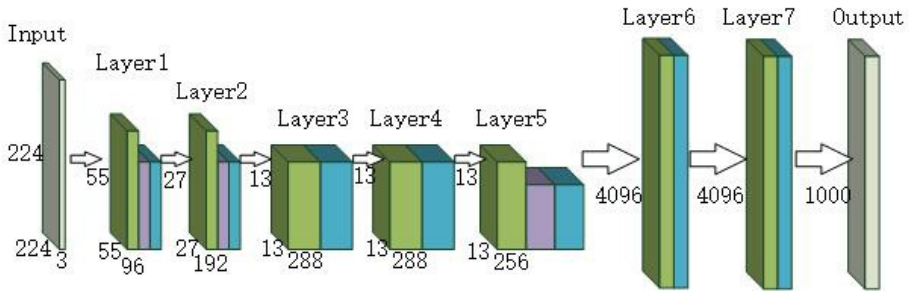

Figure 1. The structure of AlexNet model

2.2 Visualization of Features in Different Layers

By the visualization of different layers in AlexNet Model (Fig 2 and Fig 3), it is obvious the features in different layers represent different abstract meanings. In paper [3], it has proved that neural codes perform well, even when one uses the CNN trained for the classification task and when the training dataset and the retrieval dataset are quite different from each other. The best performance is observed not on the very top of the network, but rather on the layer that is two levels below the outputs. The output layer which has only 1000 dimensions has been inferred to have a large number of optimization and information loss for the purpose of classification tasks. The fusion method of the multi-layer features is proposed based on the different weights of each layer.

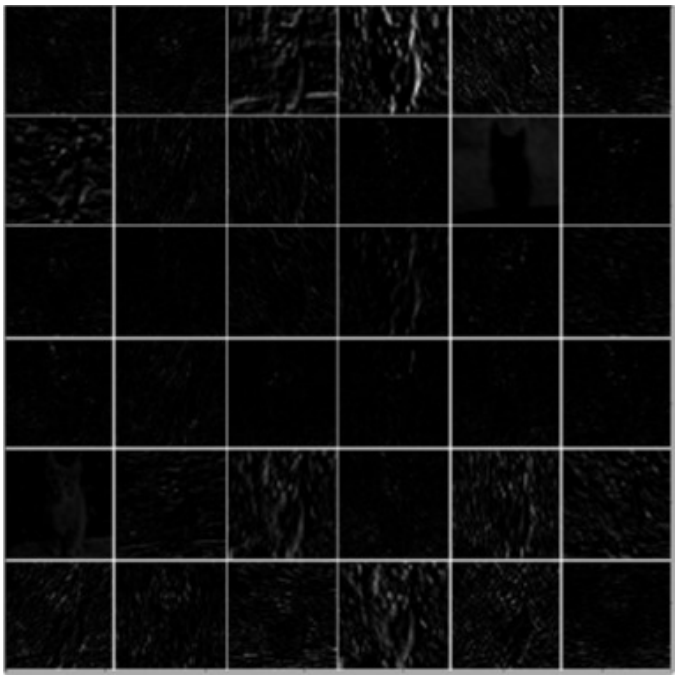

Figure 2. Visualization of feature in layer 1

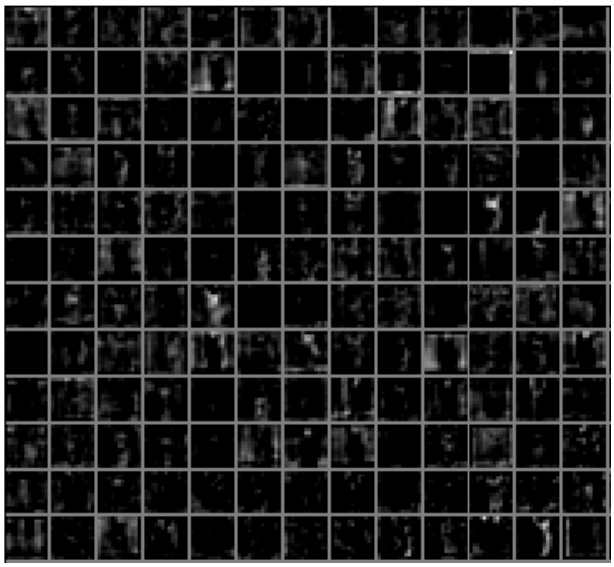

Figure 3. Visualization of feature in layer 2

3. PREPARE FOR THE EXPERIMENT

3.1 Dataset used to Experiment

Experiments were conducted in two public databases Inria Holidays [5] and Oxford Buildings with the deep learning framework Caffe [6]. Our experiment study on the three layers of the deep model:

Inria Holidays and Oxford Buildings datasets with feature of layer 6th,7th,8th.

Inria Holidays is a collection that contains some personal vacation photos of a group of high resolution images. Images contain a wide variety of types (natural, synthetic, water and fire effect and so on). The data set includes a total of 500 query groups, representing different scenery and. Each query provides some highest similarity of images.

The Oxford Buildings data set consists of 5062 Oxford landmark building images, which are composed of 11 categories of landmarks, each of which has 5 query images and the query result standard file, which are used to evaluate the retrieval results.

3.2 Detail of Training the Model

The experimental model is used to train the ImageNet [7] data set for image classification, including model parameters of the AlexNet model. Then the model is tuned for the specific image datasets and applied to image retrieval.

Since the image size of the data set is different, the image is pre-processed, and the resolution of each image is adjusted to 256×256. The sub-region of 227×227 is used as the input to reduce the over fit in training progress. The classification number of output layer is modified to 500 in experiment 1,2 and to 55 in experiment 3,4. The output layer is trained by the traditional gradient descent method and the ReLUs layer is used as the nonlinear activation function. The DROUPOUT technique is used in the training process of each layer parameters, that is, the parameters of each layer are discard by 50%, and the output layer is the SOFTMAX layer, which is used as the image classification model to realize the function of image classification. The last layer is used here just for fine-tuning the model and not to classify images.

Fine-tuning the model will improve the performance to the specific dataset. The prediction probability vector is defined as the output of the Sofmax layer. The goal of fine-tuning is reducing the Minimum Squared-Error. The Minimum Squared-Error can be set as below:

$$J = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{c}(p_{ik} - \hat{p}_{ik})^2 \tag{1}$$

3.3 Calculation of the Similarity between Two Images

When the model is fine-tuned, we can use this model to extract features on different layers. First, the similarity of single feature is calculated by the Euclidean distance, as Eq.2 shows:

$$d_{12} = \sqrt{\sum_{i=1}^{n}(x_{1i} - x_{2i})^2} \tag{2}$$

Then, the final similarity of two images is defined as the sum of two or three features in different layers with different weights, as Eq.3 shows:

$$S_{12} = \sum_{i=1}^{n} w_i d_i \tag{3}$$

4. EXPERIMENTATION

4.1 Procedure

In Inria Holidays data set, the features were extracted from 500 query images and 991 images. Picture's names and features are stored in the index file. The pictures are sorted according to the similarity, and then select out the top 10 images. For the Oxford Buildings data set, the number of selected picture is changed to 100, since the number of relate images is different.

In experiments, the mAP(Mean average precision) is used to evaluate the query results. Mean average precision for a set of queries is the mean of the average precision scores for each query.

To find the weight of different layers is the target which can improve the retrieval performance. So we should choose some different weights and check out which is the best one. Among these results with each weight, we can find some rules to improve the retrieval effectiveness.

4.2 Tables of Results for Different Datasets and Weights

Table 1. Inria Holidays with layer 6th,7th

weight	1.0:0.0	0.75:0.25	0.5:0.5	0.25:0.75	0.0:1.0
mAP	0.55911	0.62462	0.64096	0.63040	0.58300

Table 2. Inria Holidays with layer 6th,7th,8th

weight	1:1:1	1:1:0.75	1:1:0.5	1:1:0.25	1:1:0.0
mAP	0.62814	0.63414	0.63780	0.63740	0.64096

Table 3. Oxford Buildings with layer 6th,7th

weight	1.0:0.0	0.75:0.25	0.5:0.5	0.25:0.75	0.0:1.0
mAP	0.362079	0.370027	0.365807	0.353460	0.324239

Table 4. Oxford Buildings with layer 6th,7th,8th

weight	1:1:1	1:1:0.75	1:1:0.5	1:1:0.25	1:1:0.0
mAP	0.357697	0.359414	0.362801	0.364587	0.365807

5. CONCLUSION

As can be seen from Table 1, when using features of sixth or seventh layers for image retrieval, the result is not the best. It can produce a mAP 0.06 upgrades, when set the weight of the two layers at the same ratio. Table 2 comparison Table I can be drawn that if we try to use the feature of layer 8, it produces a worse effect. It is noticed that the number of neural units in the eighth layer is the same as the image classes. The number is much smaller than the sixth and seventh layer. The loss of feature of the eighth layer is high, and it is not suitable for representing images.

Retrieval results of Oxford Buildings are shown in Table 3. The results are higher when the ratio of weights is 0.75:0.25. The results show that the best score in different data sets may have different ratio in different layers. Oxford Buildings data sets are all buildings, and the difference between high level feature maybe small and images can only be distinguished by low level features. Finally, results of image retrieval on Oxford Buildings are shown in Table 4 and the results are consistent with the Table 2.

Consider the experimental results, image features of different levels of abstraction can be obtained by different layer of deep model. The feature in the high level is more abstract and closer to its semantic, in contrast, the low level features are more close to the image details.

As a conclusion, the fusion feature of Deep CNN model can improve the result of image retrieval. When applying the features in deep model to image retrieval, it should vary the weights of features in different layers according to the image datasets.

In the future work, a more precise weight should be found by multivariable analysis. It will give a more objective result.

REFERENCES

[1] Hinton, G. E., Osindero, S. and Teh, Y., "A fast learning algorithm for deep belief nets" Neural Computation 18, 1527-1554(2006).

[2] Krizhevsky, A., Sutskever, I. and Hinton, G. E., "Imagenet classification with deep convolutional neural networks" Advances in Neural Infor-mation Processing Systems, 1097-1105(2012).

[3] Babenko, A., Slesarev, A., Chigorin, A., and Lempitsky, V., "Neural Codes for Image Retrieval" Computer Vision–ECCV , 584-599 (2014)

[4] Le Cun, B. B., Denker, J. S., Henderson, D., Howard, R. E., "Handwritten digit recognition with a back-propagation network" Advances in neural information processing systems , 396-404 (1989).

[5] Jegou, H., Douze, M. and Schmid, C., "Hamming embedding and weak geometric consistency for large scale image search" Computer Vision–ECCV, 304-317(2008).

[6] Jia, Y., Shelhamer, E., Donahue, J. and etc, "Caffe: Convolutional architecture for fast feature embedding" Proceedings of the ACM International Conference on Multimedia. ACM, 675-678(2014).

[7] Berg, A., Deng, J., and Li, F. F., "ImageNet large scale visual recognition challenge 2010" Challenge, (2010).