ATLANTIS PRESS

Advances in Intelligent Systems Research, volume 133
2nd International Conference on Artificial Intelligence and Industrial Engineering (AIIE2016)

Image Retrieval Algorithm Based on Convolutional Neural Network

Hailong Liu [1, 2], Baoan Li [1, 2, *], Xueqiang Lv [1] and Yue Huang [3]
[1]Beijing Key Laboratory of Internet Culture and Digital Dissemination Research, Beijing Information Science & Technology University, Beijing 100101, China
[2]Computer School, Beijing Information Science and Technology University, Beijing 100101, China
[3]Xuanwu Hospital Capital Medical University, 100053, China
[*]Corresponding author

***Abstract*—With the rapid development of computer technology and the increasing of multimedia data on the Internet, how to quickly find the desired information in the massive data becomes a hot issue. Image retrieval can be used to retrieve similar images, and the effect of image retrieval depends on the selection of image features to a certain extent. Based on deep learning, through self-learning ability of a convolutional neural network to extract more conducive to the high-level semantic feature of image retrieval using convolutional neural network, and then use the distance metric function similar image. In Corel dataset, this method has a higher precision and recall and satisfactory results were obtained. The experiment proves the validity of the image retrieval algorithm.**

Keywords-deep learning; convolutional neural network; feature extraction ; image retrieval

I. INTRODUCTION

With the rapid development of the Internet and the deepening of computer research, people are increasingly dependent on the Internet. Users can access information on the Internet or on the computer, the need to help and share their own things. This makes the online data into an explosive growth, the form of data from the original text data is constantly extended to the image, video, voice and other data. Image, video data contains more information rich, more conducive to the exchange of the people. But compared with the text data, the retrieval of these data forms is often difficult. For the image, the early image retrieval technology is mainly based on the text and content based. The process of image retrieval based on text is: First of all, the image is marked with the text information, each image corresponds to a number of text words, and then through the text matching to retrieve the image [1]. The advantage of this method is that the text information of the labeled image is artificially marked, and the semantic information between the image and the text is approximate. But the disadvantage of this method is obviously, the number of images is very large, the need to invest a lot of manpower, but also some annotators in image understanding is not the same as the resulting annotation words are not the same situation. With the rapid growth of the number of digital images, the difficulty caused by the manual image annotation is very sharp. Then there is the image of automatic tagging technology, but the accuracy of automatic tagging is relatively low, can't meet the needs of people. In order to overcome this difficulty, the researchers from the image itself, and put forward the method of image retrieval based on content. The method is to extract the visual features of the image content: color, texture, shape etc., the image database to be detected samples for similarity matching, retrieval and sample images are similar to the image. The main process of the method is the selection and extraction of features, but there are "semantic gap" [2] between low-level features and high-level semantic image of these images, to extract high-level semantic features desired by the user to accurately describe image content. For this problem, the experts and scholars have also been studied, but the effect is not satisfactory.

The main task of machine learning research is to design and develop self learning algorithm based on the actual training data, so that it can automatically discover the rules of the data and the feature of the learning data. Deep learning is a part of machine learning, since 2006, Hinton et al [3] proposed a method of self-learning initialization parameters, and then gradually optimization to solve the optimization problem of deep learning model, deep learning has been rapid development. Currently has been widely used in Natural Language Processing[4], speech recognition[5], computer vision[6] and other fields, the depth of learning and even become synonymous with machine learning. The concept of deep learning comes from the study of artificial neural network [7]. The early artificial neural network is a neural network which is a shallow learning neural network, which generally includes an input layer, an implicit layer and an output layer. The learning ability and generalization ability of the shallow layer neural network are quite poor, which can't be extracted from the learning. Deep learning network structure is the best simulation of the human brain cortex, the input data processing is hierarchical, each layer is extracted from different levels of the feature of the input data to describe the size is also different. The depth study of the network structure is deeper than the shallow structure, with a large amount of data, long training time and so on. But in this paper, the neural network [8, 9] is used to share the weight of the neural network, the local connection, the effective reduction of the impact of the problem.

Deep learning has a strong self-learning ability, rather than the manual to design features, can better extract the feature of the input data, revealing its inherent laws. Convolutional neural network features: local perception, weight sharing, can improve the training speed of neural network, the nonlinear activation function of ReLU(rectified linear units) makes the training convergence faster, the training set error rate lower. In recent years, The activation function of the convolution neural network model mainly uses ReLU [10,11]. In this paper, through the study of deep learning based on convolutional neural networks, using convolutional neural network model to extract the semantic features of the image, and then according to the distance function to calculate the similarity between image high-level semantic features, so as to obtain similar images.

II. CONVOLUTIONAL NEURAL NETWORK

Convolutional neural network is developed based on the traditional neural network, can extract the structure details from the input image, is a kind of efficient recognition method, is suitable for the detection and identification of problems in image. The core idea of the convolutional neural network is to optimize the structure of the neural network by local perceptual convolution and weight sharing, and to reduce the parameters of the neural network. Convolutional neural network are mainly convolutional layers, pooling layers, fully-connected layers, in addition to the classifier, the input layer, the output layer.

A. Convolutional Layer

Convolution operation is a convolution neural network more important operation, which is also the extraction of image features for layer by layer learning is an important step. Convolution operation is performed by a convolution of the image of the local convolution, get the local feature, and then move the convolution to check the image of the other parts of the local convolution. The idea of this local connection comes from the early perception of the machine, and it is consistent with the local perception found in the cat's visual system. The features of the local perception are some basic image features, such as edges and angles, which lay the foundation for the more high-level extraction of features. Generally speaking, in order to ensure the integrity of the image information, the convolution kernel convolution operation, and each convolutional kernel feature information is extracted from different aspects; each convolution kernel will have to move according to a certain step of the whole image convolution operation. The convolutional neural network is such a layer by layer extract feature's information and abstract image information. Finally, the description of the original image is more abstract information, which is more conducive to the improvement of image retrieval.

B. Pooling Layer

The pooling layer is based on the volume of laminate, pool operation, pooling operation of the layer relative to the convolution operation is relatively simple. Pooling layer were treated each feature map, pooling operation is simple and intuitive is to each pixel feature map node around a mean value, after the results of this treatment can reduce the spatial resolution of input feature map, reduce the computational complexity. The pool of operation to a certain extent also makes the convolution neural network has the displacement and distortion of the zoom.

In the beginning of the neural network model, pooling operations are not overlapping, but now slowly develop into overlapping pools, which can reduce the error rate of the final results. During the entire convolutional neural network, there is a certain size for the input image size. Cannot be put together for training and testing images of different sizes, which is because if the input feature maps of varying size, after the pool operation results are obtained with different sizes, which has an effect on the back layer weight matrix will be fully-connected.

C. Fully-Connectional Layer

The fully-connected layer followed at the convolutional and the pooling operation, it can be simple to understand that each feature map before the fusion gets more accurate image expression.

D. Action Function

Non-linear transformations are often used between convolution and pooling operations to avoid the problem of insufficient expression of linear models [12]. Conventional convolution neural networks use sigmoid, tanh and other functions in the activation function. These functions are all saturated nonlinear functions. Current convolutional neural network uses the unsaturated nonlinear function ReLU, as shown in formula (1), which converges faster when training the network model.

$$R(x) = \max(0, x) \tag{1}$$

III. ALEXNET

AlexNet convolution neural network[9] was constructed by Alex Krizhevsky et al those are Hinton's student. AlexNet won the first prize in the ILSVRC 2012 competition, Top5 error rate of 15.3% in this competition. AlexNet success has greatly enhanced the enthusiasm of deep learning in various fields. AlexNet convolutional network architecture as shown in Figure I:

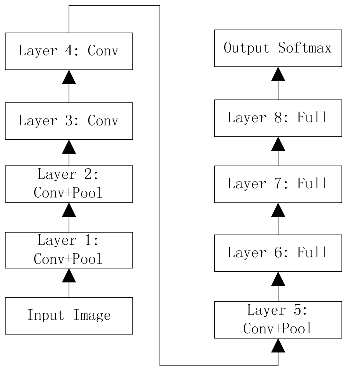

FIGURE I. ALEXNET CONVOLUTIONAL NETWORK ARCHITECTURE

In Figure I, AlexNet network structure also has a convolution layer, pool layer, fully-connected layer. The structure also uses the unsaturated nonlinear activation function ReLU, local normalization and overlapping pooling operations (not shown in the figure).

AlexNet convolutional neural network in order to get more information of the image information, the number of convolution kernel of this network model convolution operation is 96, the size of these convolution cores is 11 * 11, the step is 4. After the convolution operation, it is activated by ReLU, which avoids the problem that the linear model can't express enough, and improves the convergence rate of the convolution neural network. ReLU unit will be added after the local normalization, this will make the network to achieve better generalization effect. In fact, the local normalization operation is a simulation of biological neurons in the current neurons in the excited state, will inhibit the next neurons. This simulation of biological neurons does make the performance of the convolutional neural network model improved.

In the earlier convolutional neural network model, pooling operations are non-overlapping, meaning that the step size of the pooled window is equal to the size of the pooled window. And the network structure is the pooling of the pooling operation, so that more information can be obtained information, to avoid the occurrence of over-fitting, but also conducive to image retrieval effect of the upgrade.

In the fully-connected layer, the fully-connected layer fuse each feature map, remove the relative position information of objects in the image, the two-dimensional feature map tile into one-dimensional feature. As a result, the data of the fully-connected layer will be very large, and the training time will be longer.

AlexNet network architecture is divided into 8 layers, in addition to the input layer and output layer, it includes 5 convolutional layers and 3 fully-connected layer. The first convolutional layer uses 96 convolutional kernels which are 11*11, the stride is 4, the second layer convolution is 256 convolutional kernels which are 5*5, the stride is 1. The size of the convolution kernel is smaller, and stride is smaller, these can extract more image details and local image features. Using local response normalization, overlapping pooling to improve the precision and reduce over-fitting after these two convolutional layers. In the third layer and the fourth layer is the 384 convolutional kernels which are 3*3, the stride is 1, the fifth convolutional layer is 256 convolutional kernels which are 3*3, the stride is 1.The first 6, 7, 8 layer are fully-connected to fuse the feature maps which are obtained by the convolutional layer before, so the amount of data is very large and calculation is very cost. But the AlexNet network uses the Dropout technology, which can effectively solve training time consuming and over-fitting problem.

In this paper, by comparing feature of AlexNet network in the fully-connected layer extracted for image retrieval results, the experimental results show that in the higher layer fully-connected image features more conducive to the image retrieval results. The image of the representation, but also closer to people's judgment. Therefore, this paper uses the highest total connectivity layer to extract the features of image retrieval.

IV. EXPERIMENTS AND ANALYSIS

A. Dataset and Experimental Evaluation

This paper's image dataset is Corel dataset, it contains 1000 images, image dataset includes 10 categories of savage, dinosaur, architecture, flowers, etc., each class includes 100 pictures.

For image retrieval have recall, precision, mean Average Precision and other evaluation.

recall:

$$recall = a / b$$

precision:

$$precision = a / c$$

mean Average Precision:

$$mAP = \frac{1}{m_p} \sum_{i \in p} \frac{1}{m_q} \sum_{j \in q} precision(k_{ij}) \qquad (2)$$

a: the number of images that are similar to the image to be retrieved, b: the total number of images similar to the image to be retrieved, c: the number of images returned by the retrieval system, K_{ij} represents that the results of j-th image in i-th class as image to be retrieved.

B. Experimental Result Analysis

In this paper, we first extract the feature of the image through the depth of the neural network AlexNet model of the whole connection layer, and get the image features of the whole connection layer. The image features Fc6, Fc7, Fc8 extracted by the AlexNet network structure are compared, as shown in Table I:

TABLE I. THE AVERAGE PRECISION OF IMAGE RETRIEVAL ALEXNET NETWORK IMAGE FEATURE EXTRACTION

	AlexNet-Fc8	**AlexNet-Fc7**	**AlexNet-Fc6**
Top 10	0.9277	0.9172	0.9103
Top 20	0.9032	0.8835	0.8741
Top 30	0.8814	0.8569	0.8454
Top 40	0.8635	0.8328	0.823
Top 50	0.8464	0.8074	0.7997

As can be seen from Table I, Fc8 features the average image retrieval accuracy of the best, Fc7, Fc6 effect in turn decreased. The Fc8 feature is the highest level of all connection layers in the AlexNet network structure, compared with the previous two layers of the fully-connection layer of image features, Fc8 layer image features although the data dimension is low, but the image description is more generalization, more effective and accurate.

From the experimental results, the depth of the convolutional neural network for image extraction of high-level semantic features is favorable. Deep learning has a strong learning ability and high performance feature, to a certain extent, to reduce the impact of the "semantic gap". The depth of the network structure, the high level of the fully-connected layer extracted by the feature of the image is also more accurate description.

V. CONCLUSION

In this paper, we use the convolution neural network model to extract the high level semantic feature of the image, and realize the image retrieval by analyzing the structure of the neural network. Deep convolutional neural network firstly, the image is gradually learning and abstract, each layer can be generated to describe the image content of the underlying feature of the image. Through experiments, select the top layer of the network to extract the feature of the image. The last image retrieval results confirmed that the image features of the image is the best expression ability, and achieved better results of image retrieval.

ACKNOWLEDGEMENT

This project was supported by the Funding Project for Natural Science Foundation of China (Grant No. 61671070), the Opening Project of Beijing Key Laboratory of Internet Culture and Digital Dissemination Research (Grant No.ICDD201608) and the Project of Computer Teaching Reform of Beijing Information Science and Technology University (Grant No. 5111623409).

REFERENCES

[1] Jun Shi, Yilin Chang. Overview of image retrieval[J]. JOURNAL OF XIDIAN UNIVERSITY, 2003, 30(4):486-491.

[2] Yuxiang Xie, Xidao Luan, Lingda Wu. Analysis of semantic gap of multimedia data[J]. JOURNAL OF WUT(INFORMATION & MANAGEMENT ENGINEERING), 2011, 33(6):859-863.

[3] Hinton G E, Ruslan R S. Reducing the dimensionality of data with neural networks [J]. Science, 2006,313(5786):504-507.

[4] Xuefeng Xi, Guodong Zhou. Pronoun Resolution Based on Deep Learning[J]. Acta Scientiarum Naturalium Universitatis Pekinensis, 2014, 50(1):100-110.

[5] Shanhai Wang, Xinxing Jing, Haiyan Yang. Study of isolated speech recognition based on deep learning neural networks[J]. Application Research of Computers, 2015, 32(8):2289-2291.

[6] Yin Zheng, Quanqi Chen, Yujin Zhang. Deep learning and its new progress in object and behavior recognition[J]. JOURNAL OF IMAGER AND GRAPHICS, 2014, 19(2):175-184.

[7] Xiaofei Wang, Bonian Li. Texture retrieval method using pulse-coupled neural network[J]. Computer Engineering and Applications, 2012, 48(7): 201-204.

[8] Xianchang Chen. Research on algorithm and application of deep learning based on convolutional neural network[D]. Zhejiang Gongshang University, 2013.

[9] Krizhevsky A, Sutskever I, Hinton G E. ImageNet Classification with Deep Convolutional Neural Networks[J]. Advances in Neural Information Processing Systems, 2012, 25(2):2012.

[10] Dahl G E, Sainath T N, Hinton G E. Improving deep neural networks for LVCSR using rectified linear units and dropout[C]// 2013:8609-8613.

[11] Glorot X, Bordes A, Bengio Y. Deep Sparse Rectifier Networks[J]. Learning/statistics & Optimisation, 2010.

[12] Baocai Yin, Wentong Wang, Lichun Wang. Review of deep Learning[J]. Journal of Beijing University of Technology, 2015(01):48-59.