Image Classification Search System based on Deep Learning Method

Zhang Lin Chen Zhiying*

[1]School of OPTO-Electronic and Communication Engineering，XiaMen University of Technology，XiaMen 361024，China

[2]School of Electrical Engineering and Automation, XiaMen University of Technology, XiaMen 361024，China

Received: July 1, 2019. Revised: December 12, 2019. 2nd Revised: June 18, 2020.
Accepted: July 8, 2020. Published: July 29, 2020.

***Abstract*—Image classification is to distinguish different types of images based on image information. It is an important basic issue in computer vision, and is also the fundamental for image detection, image segmentation, object tracking, and behavior analysis. Deep learning is a new field in machine learning research. Its motivation is to simulate the neural network of the human brain for analytical learning. Like the human brain, deep learning can interpret the data of images, sounds, and texts. The system is based on the Caffe deep learning framework. Firstly, the data set is trained and analyzed, and a model based on deep learning network is built to obtain the image feature information and corresponding data classification. Then the target image is expanded based on the bvlc-imagenet training set model, and finally achieve "search an image with an image" web application.**

***Keywords*—Image classification; Depth learning; Caffe framework; Convolution neural network;**[1]

I INTRODUCTION

Deep learning is a new field in machine learning research. Its motivation is to build and simulate the neural network of human brain for analysis and learning, which can interpret data, such as images, sounds and texts[1]. Caffe (Convolutional Architecture for Fast Feature Embedding) is an open source deep learning framework developed by Jia Yang-qing and others of Berkeley University[2]. It is implemented in an efficient C++ language and has built-in interfaces between Python and MATLAB for developers to use Python or MATLAB to develop and deploy applications with deep learning as the core algorithm. Caffe provides a complete toolkit for training, testing, fine tuning and developing models. It is suitable for massive data processing at the Internet level, including voice, picture, video and other multimedia data. It can complete high-speed operation of massive data through GPU, and is suitable for the application development of in-depth learning in the field of image and computer vision.

Based on the basic concept of deep learning, many experts apply deep learning to their respective fields. Emmanuel Okafor proposes to use UAV to monitor, identify and track wild animals. The supervised learning algorithm based on depth neural network and feature description is used to randomly rotate the collected images to expand the data samples, and the data training of the model is realized by multi-directional data enhancement technology[3]. BIN WANG proposed a multi-layer convolution neural network based on the combination of network and small probability learning, and constructed a classifier suitable for small sample environment to realize plant leaf classification[4].This

[1] This work was supported by the Scientific Research Climbing Plan of Xiamen University of Technology (Grant No. XPDKQ18021), and Educational and Scientific Research Project for Young and Middle-aged Teachers of Fujian Province(Grant No. JAT190672).

Zhang Lin was born in 1981. She received the B.S. degree from Jilin University in 2005, and received the M.S. degree from Harbin Institute of Technology in 2007. She is currently a Ph.D. candidate in Electronics Information Engineering from the Harbin Institute of Technology (HIT). Her research interest is computer vision and deep learning. Email: zhanglin603@aliyun.com

Chen Zhiying received the ph.D. degree in electrical engineering from the Fuzhou University, China, in 2019. She is currently an Associate Professor in the School of Electrical Engineering & Automation, Xiamen University of Technology, China. Since 2013, she has been involved in research in the areas of biomedical engineering, wireless implant communication and body area network.Email：chzy207@163.com

paper uses LeNet convolution network designed by Caffe framework, uses MNIST handwriting recognition data set to realize model training and analysis, and extracts feature information of target image[5]. In order to better show the application effect of deep learning, bvlc_reference_caffenet.caffmodel is used as the basic model of image recognition and classification, which can realize simple "map search" web application.

II. In-depth Learning

The concept of in-depth learning was proposed by Hinton et al. in 2006[6]. An unsupervised greedy layer-by-layer training algorithm based on deep belief network (DBN) is proposed. The multi-nonlinear hierarchical system effectively solves the over-fitting problem of deep learning in training process. By introducing image saliency information into image hierarchical sparse representation, the semantic information of image features is enhanced and the saliency feature expression of image is obtained.

A. Convolution Process

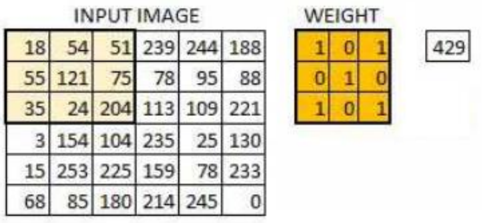
For one-dimensional signals, convolution is defined as：

$$\mathrm{o}(x) = f(x) * g(x) = \sum_{u=-x}^{x} f(x-u)g(u) \qquad (1)$$

Image signal is a two-dimensional signal whose convolution is defined as:

$$\begin{aligned} &\mathrm{f}(m,n) * f(m,n) \\ &= \sum_{u}^{\infty}\sum_{v}^{\infty} f(u,v)g(m-u,n-v) \qquad (2) \\ &= \sum_{u}^{\infty}\sum_{v}^{\infty} f(m-u,n-v)g(m,n) \end{aligned}$$

Taking image processing as an example, the convolution process is actually based on a fixed matrix, which is scanned one by one in another matrix to get the sum of values, as shown in Figure 1 below.



INPUT IMAGE

18	54	51	239	244	188
55	121	75	78	95	88
35	24	204	113	109	221
3	154	104	235	25	130
15	253	225	159	78	233
68	85	180	214	245	0

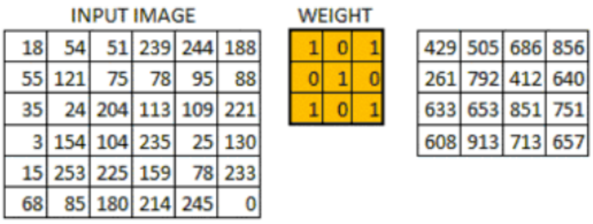
WEIGHT

1	0	1
0	1	0
1	0	1

429

Fig.1 3*3 Matrix Convolution Process

If the input image is a set of matrices of 6*6, the data of the first 3*3 lattices can be 429 after weighted summation, and the data of the first convolution can be obtained. Each operation of the input matrix moves a small lattice backward, and weights the sum, sweeping the whole data to get a 4*4 data, the result of convolution is that the dimension is reduced.



INPUT IMAGE

18	54	51	239	244	188
55	121	75	78	95	88
35	24	204	113	109	221
3	154	104	235	25	130
15	253	225	159	78	233
68	85	180	214	245	0

WEIGHT

1	0	1
0	1	0
1	0	1

429	505	686	856
261	792	412	640
633	653	851	751
608	913	713	657

Fig.2 6*6 Matrix Convolution Process

B. Convolution Kernel

The convolution core is the number of 3*3 orange matrices in Figure 1. Sometimes the features need to be extracted are very many and extensive, so more different matrices are needed to scan (several times). Then the number of orange matrices is the number of convolution kernels.

The matrix formula of output is:

output dimension = (input dimension - convolution dimension + 1) * number of convolution cores (3)

When convoluting 6*6 matrices with n orange matrices with different weights, a matrix of 6*6 can be transformed into n matrices of 4*4, i.e. 6*6 --> n*4*4 matrices.

C. Pooling

Pooling is very similar to convolution, using the weighted sum of one matrix and another to get the final data. The biggest difference between pooling and convolution is that the convolution reuses one data, while the pooling uses only one weighted summation for each data. When the original matrix is a matrix of m*m and the sampling window is n*n, the convolution can obtain the matrix result of (m-n+1)*(m-n+1), and the pooling is not repeated. In the case of data weighted summation, only a total of (m/n)*(m/n) result matrices can be sampled. Since the convolution is done, the image is still large (because the convolution kernel is small), so in order to reduce the data dimension, downsampling is performed, and over-fitting is effectively avoided [7]. Although much data is reduced, the statistical properties of the feature are still able to describe the image. The process of pooling is shown in

Figure 3:

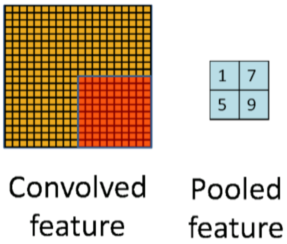


Fig.3 Pooled Process Legend

D. Training Process

The main process of training is to train a series of filters when training a convolutional layer of the Convolutional Neural Network (CNN). In simple terms, training CNN is training each convolution layer filter, so that these filter groups have a higher activation of specific mode features, thus achieving the purpose of classification/detection of CNN networks [8]. Therefore, the task of constructing a convolutional neural network is to construct these filters, that is, to change the value of the filter matrix, that is, to change the weight Weight - to identify a particular feature[9]. This process is called training, and Figure 4 is a deep learning training flow chart.

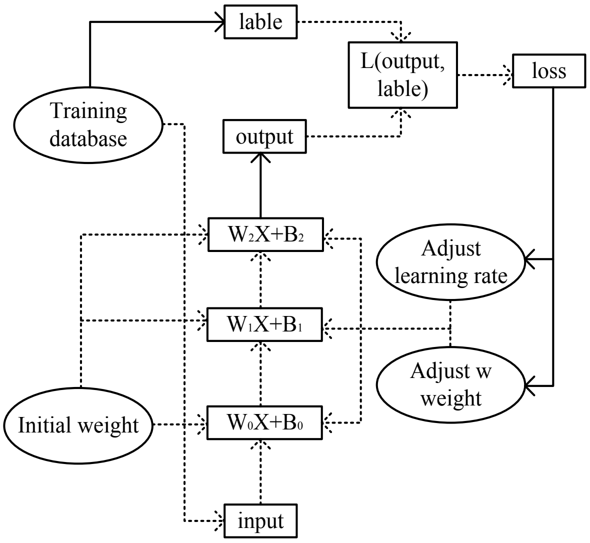


Fig.4 Flow Chart of In-depth Learning and Training

Among them, the weight is the W0, W1, W2 in the figure, and its value is controlled and adjusted by the loss function with the initial weight as the learning training process, so as to achieve the purpose of learning.

III. MNIST HANDWRITING RECOGNITION MODEL UNDER DEEP LEARNING

A. LeNet Convolutional Network

LeNet which was proposed by Yan LeCun et al in 1998[10], is one of the earliest convolution neural networks. The network uses two convolutional layers, two pooling layers, and two fully connected layers. The convolution neural network only needs to feel the local image region in the lower sampling layer, and the global information can be obtained by combining the characteristic parameters of these different local regions at the higher level.The complete structure of a LeNet-5 convolutional neural network has a total of approximately 48,285 training parameters, as shown in Figures 5 and 6.The system is successful in small-scale handwritten digit recognition.

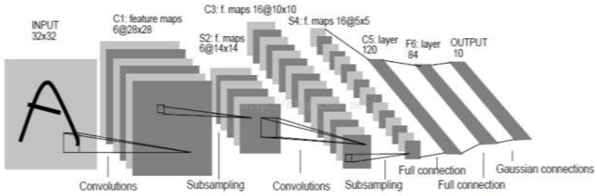
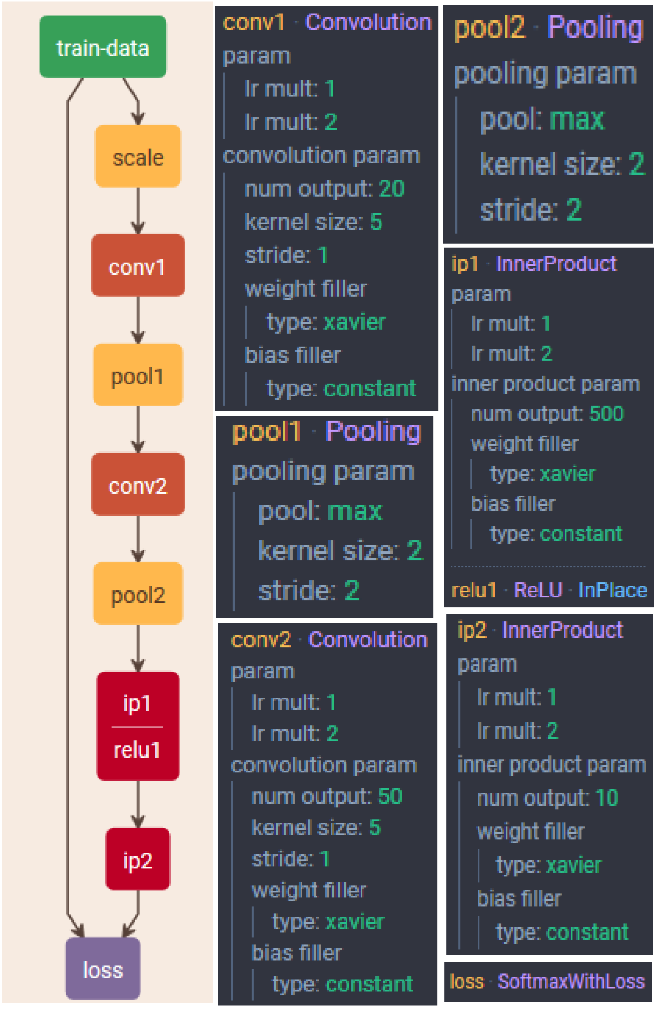


Fig.5 LeNet Structure Diagram



conv1 Convolution
param
lr mult: 1
lr mult: 2
convolution param
num output: 20
kernel size: 5
stride: 1
weight filler
type: xavier
bias filler
type: constant

pool1 Pooling
pooling param
pool: max
kernel size: 2
stride: 2

conv2 Convolution
param
lr mult: 1
lr mult: 2
convolution param
num output: 50
kernel size: 5
stride: 1
weight filler
type: xavier
bias filler
type: constant

pool2 Pooling
pooling param
pool: max
kernel size: 2
stride: 2

ip1 InnerProduct
param
lr mult: 1
lr mult: 2
inner product param
num output: 500
weight filler
type: xavier
bias filler
type: constant

relu1 ReLU InPlace

ip2 InnerProduct
param
lr mult: 1
lr mult: 2
inner product param
num output: 10
weight filler
type: xavier
bias filler
type: constant

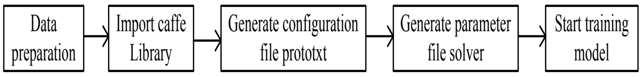
loss SoftmaxWithLoss

Fig.6 LeNet Parameter Diagram

B. Training Process

Mnist image data official website download address: http://yann.lecun.com/exdb/mnist/.The data is divided into a training set (60,000 sheets and 10 categories) and a test set (a total of 10,000 sheets and 10 categories), each category is placed in a separate folder, and all the images are generated with a list of txt lists (train.txt and test.txt).

Environment Description: Linux (Ubuntu 16.04) operating system under VMware Workstation14, Python environment: Anaconda2.7, Caffe and Caffe's Opencv (3.4) support, the environment is shown in Figure 7.

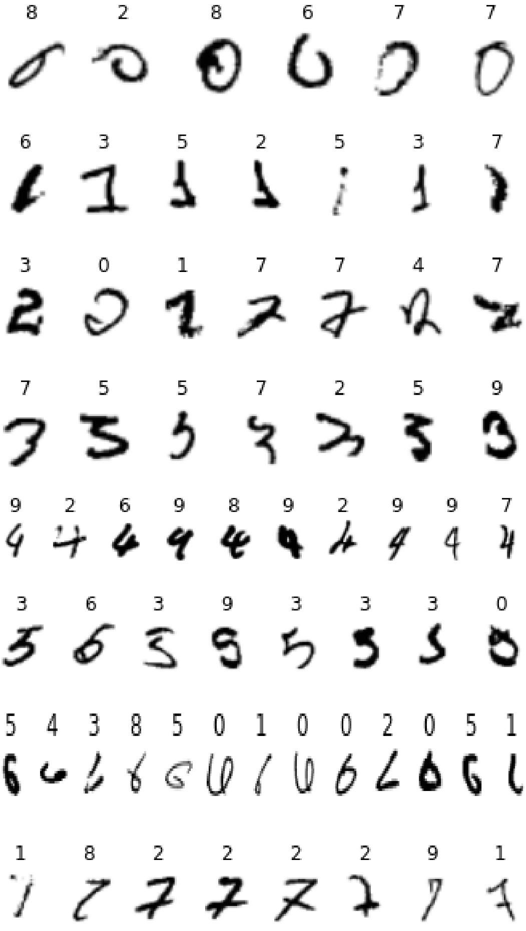

Data preparation
Import caffe Library
Generate configuration file prototxt
Generate parameter file solver
Start training model

F

ig.7 Overall Training Flow

C. Experiment Results

The trained model was trained using the test set provided by the official mnist. Each type of number has 1000 sheets and 10 categories and 10,000 images, and the judgment results are mostly correct. Skip the correct result, Figure 8 is the result of the test error of 0-9: the top of the head is the result of the model's classification of the image prediction.



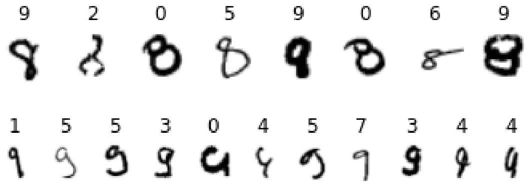


Fig.8 Error Result Set Display for mnist Test

Result analysis:

1. Test about 10,000 pictures, less than 10 seconds when sharing (only use CPU calculation, and different performance computers take time, only for reference), in which the number of judgment errors is about 80, and the correct rate is about 99.2%. The test speed of each picture is less than 1 millisecond.

2. The main causes of recognition errors are handwriting scribbling, partial ambiguity and handwriting ambiguity (even if human, it is impossible to accurately determine the specific classification of numbers).If the writing is clear and accurate, the recognition accuracy is more than 99.5%.

3. Some of the test writings are clear and can be recognized by the human eye, but the model gives a wrong result. Some of the features of these test handwritings are very similar to those of other handwritings. The recognition results only give the most probable classification and fail to give real results. This should be part of the training deficit of deep learning because of the quality and quantity of the training set.

IV. Application for Image Recognition: Search by Image

Baidu launched its latest search function, "Image Recognition", which is based on similar image recognition technology. After uploading local pictures or inputting the URL address of pictures, users can analyze the features of images, and then search similar image resources and information content from the Internet. According to the results of the previous in-depth learning of picture classification, an application similar to Baidu's " Image Recognition" is designed.

A. General Design

The system first obtains the picture file or the URL uploaded by the user (POST) through the Web front-end. After receiving user pictures, the server calls the *Bvlc* model to process and loads the model to identify the classification information of the pictures. According to this classification information, the background uses the *Request* crawler library to search for similar pictures on *Bing*, get relevant picture links and pass them to the front-end browser. The browser makes access requests according to the given image links and displays them in the user's browsing interface, thus achieving the function of searching for images. The flow chart of the system is shown in Figure 9.

B. Model Framework

Background server Django: Django is an open source Web application framework written by Python. It is more suitable for medium-sized Web projects. This image search system is based on Django to build and deploy Web applications.

Web front-end framework Bootstrap: Bootstrap is a popular front-end framework. It originated from Twitter and provides concise cascading style sheets and HTML rules, as well as a more complete and humanized website style. According to the above framework, the front-end Web UI interface is built, which is in line with the style of picture display module.

Search classification is based on Bvlc model: The model is made up of more than 300,000 iterations trained by the Caffe team using imagenet photo sets. It is an applicable level model with 1,000 kinds of classifications. The specific classification of the analysis pictures is to get the key words of classification through this model.

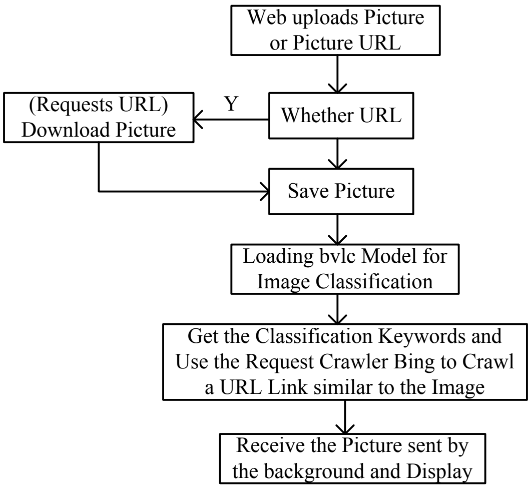
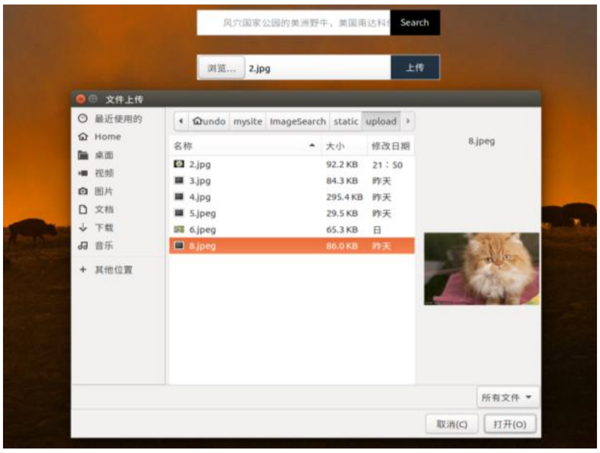

Web uploads Picture or Picture URL
(Requests URL) Download Picture
Y
Whether URL
Save Picture
Loading bvlc Model for Image Classification
Get the Classification Keywords and Use the Request Crawler Bing to Crawl a URL Link similar to the Image
Receive the Picture sent by the background and Display

Fig.9 Design Flow Chart

C. Image Search Process

Front-end interface: There are two ways to upload pictures, the local file of the picture and the URL address of the picture. Select to upload a cat picture, as shown in Figure 10.



2.jpg

名称	大小	修改日期
2.jpg	92.2 KB	21：50
3.jpg	84.3 KB	昨天
4.jpg	295.4 KB	昨天
5.jpeg	29.5 KB	昨天
6.jpeg	65.3 KB	日
8.jpeg	86.0 KB	昨天

Fig.10 Upload Picture Interface

Background Classification: The background loads the picture into the Bvlc model for operation. Figure 11 shows the convolution core of the first convolution layer and the picture after the first convolution.

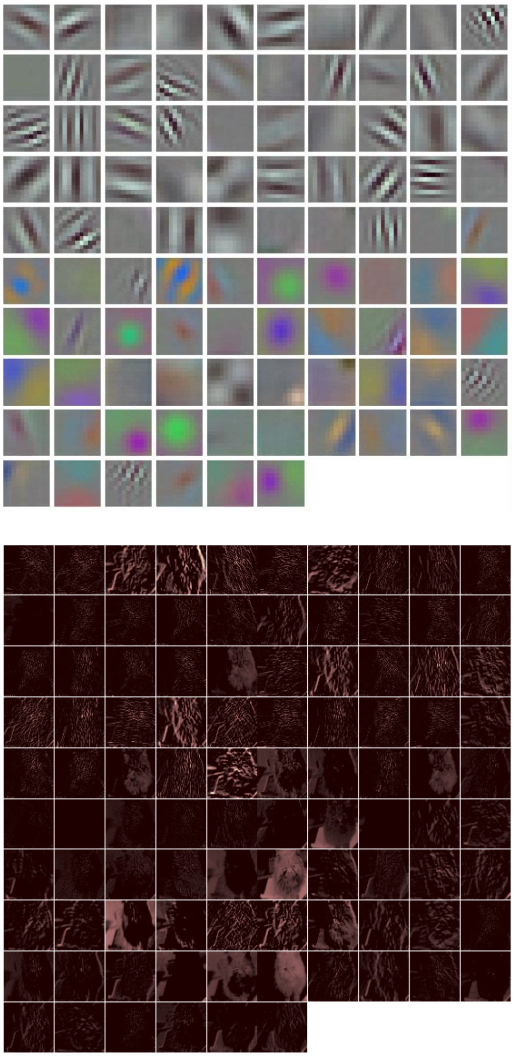


Fig.11 The Convolution Kernel of the First Convolution Layer and the Picture after the First Convolution

Finally, after analyzing the background features, the probability distribution map of image classification can be obtained，which shows that the maximum possibility of image classification given by this model is similar to MAP (maximum a posteriori probability) criterion.The following results are most likely to be classified as: 283, and then we get 283 classes as label: n02123394 Persian cat by searching label files, as shown in Figure 12.

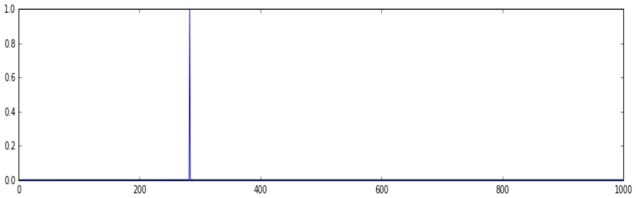


Fig.12 Probability Distribution Map of Picture Classification

The top five most likely classification results are listed:
[(0.99606931, 'n02123394 Persian cat'),
(0.0019333176, 'n02127052 lynx, catamount'),
(0.0013805312, 'n02123159 tiger cat'),
(0.00041564793, 'n02123045 tabby, tabby cat'),
(8.5782471e-05, 'n02124075 Egyptian cat')]

More picture search implementations: After obtaining the keywords, Bing pictures can be crawled through the requests crawler library (https://cn.bing.com/images). After analyzing the HTML source code of the web page, an API interface can be obtained:
https://cn.bing.com/images/async?q={%s}&mmasync=1。

The interface provides a parameter to search keywords instead of % s in URL to get the image data that you want to search, and then uses the XPath rule: '//*[@id="mmComponent_images_1"]/ul/li/div/div/div[1]/ul/li[2]/text()'，so as to extract the desired image address and image source information. After obtaining the URL information of the picture, the background system will organize the data and send it to the browser in JSON mode. The browser can display more similar pictures to the user according to this address.

Front-end picture display page: The first image is the original search image and its classification information. The other images are the result of more similar images provided to users. They also have the function of view and download.

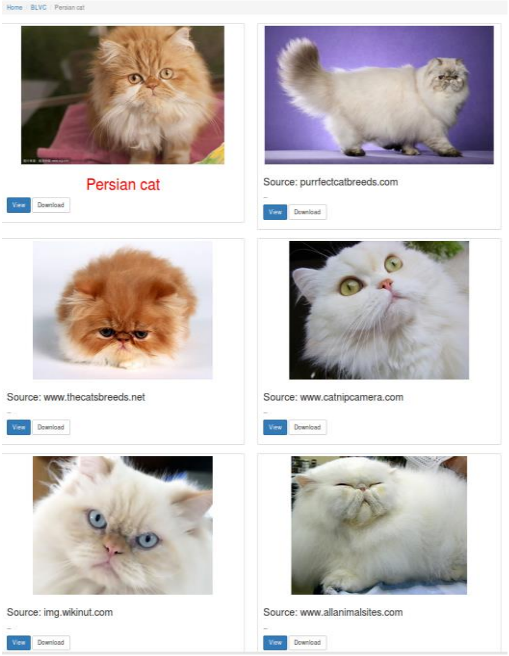

Home / BLVC / Persian cat
Persian cat
View Download
Source: purrfectcatbreeds.com
View Download
Source: www.thecatsbreeds.net
View Download
Source: www.catnipcamera.com
View Download
Source: img.wikinut.com
View Download
Source: www.allanimalsites.com
View Download

Fig.13 Judgment: Persian Cat and more

V. CONCLUSION

The development of deep learning depends on the development of large data and convolution neural network. The continuous debugging and optimization of convolution neural network structure greatly improves the training results of deep learning. For image classification system based on in-depth learning, it needs a lot of pictures as the basis of training, and continuous iterative learning can get a better model. The model used in this paper is the official bvlc_reference_caffenet provided by Caffe. The training set imagenet data set used by caffemodel has 1000 kinds of classifications. Although it is a powerful model, it is not enough to deploy AI applications in the field of computer vision applications (full-automatic driving, etc.) and needs more big data as the basis. At the same time, even with such a powerful model in practical application, it still needs to be dynamically modified step by step and improve its model and parameter configuration, which Caffe can not give a good solution. The ultimate image search application function of this paper depends on the classification results of recognition. Essentially, it is based on the classification keyword search. If different images of the same classification are searched, the similarity of the results may be too high and the intelligence is not enough. Referential solutions can use multiple training models to analyze all aspects of information of selected pictures, such as tone, style, other classification results, and search will be more intelligent, but at the same time, there will be higher requirements for training data sets and training network learning efficiency.

ACKNOWLEDGMENT

This work was supported by Educational and Scientific Research Project for Young and Middle-aged Teachers of Fujian Province(Grant No. JAT190672) and the Scientific Research Climbing Plan of Xiamen University of Technology (Grant No. XPDKQ18021).

REFERENCE

[1] L C, X.-M.Deng, M.-Q.Zhou. "Convolutional Neural Networks in Image Understanding," *Acta Automatica Sinica*,, vol.42, no.9, pp.1300-1312, 2016.

[2] Q. J. Yang, E. Shelhamer, J. Donahue. "Caffe: Convolutional Architecture for Fast Feature Embedding." *ACM International Conference on Multimedia ACM*, pp. 675-678,2014.

[3] Okafor E, Smit R, Schomaker L, et al. "Operational Data Augmentation in Classifying Single Aerial Images of Animals"[C]. *IEEE International Conference on Innovations in Intelligent SysTems and Applications (INISTA),* 2017,DOI:10.1109/INISTA.2017.8001185.

[4] Bin Wang, Dian WANG. "Plant Leaves Classification: A Few-Shot Learning Method Based on Siamese Network". *IEEE Access,*vol.7, pp.151754-151763,2019.

[5] A. M. Zhou, P. P. Ma, T. Y. Xi, et al. "Automatic identification of butterfly specimen images at the family level based on deep learning method." *Acta Entomologica Sinica*, vol.60, no.11, pp.1339-1348, November 2017.

[6] Hinton, G. E., Osindero, S. ,Teh, Y-W. A Fast Learning Algorithm for Deep Belief Nets[J]. *Neural Computation,* vol.18, no.7, 1527-1554. 2006.

[7] Z.Wei,Research and Implementation of Face Recognition Based on Caffe Platform with Deep Learning[D]. *XiDian University*, 2015.

[8] S.Z, Y.H.G, J.J.W. The development of Deep Convolution Neural Network and Its Applications on Computer Vision. *Chinese Journal of Computers*, vol.42, no.3, pp.453-482, 2019.

[9] Z.J.S, L.X, Y.M.X. Overview of deep learning[J].*Application Research of Computers*,vol.29, no.8, 2806-2810. 2012.

[10] LeCun, Y, Bottou, L, Bengio, Y. Haffner, P. Gradient-based learning applied to document recognition[J]. *Proceedings of the IEEE*.vol.86, no.11, pp.2278-2324.1998.

Zhang Lin was born in 1981. She received the B.S. degree from Jilin University in 2005, and received the M.S. degree from Harbin Institute of Technology in 2007. She is currently a Ph.D. candidate in Electronics Information Engineering from the Harbin Institute of Technology (HIT). Her research interest is computer vision and deep learning.
E-mail:zhanglin603@aliyun.com

Chen Zhiying received the ph.D. degree in electrical engineering from the Fuzhou University, China, in 2019. She is currently an Associate Professor in the School of Electrical Engineering & Automation, Xiamen University of Technology, China. Since 2013, she has been involved in research in the areas of biomedical engineering, wireless implant communication and body area network.
Email：chzy207@163.com