RECOGNITION OF INDIVIDUAL OBJECT IN FOCUS PEOPLE GROUP BASED ON DEEP LEARNING

Liu Hui-bin，Wu Fei，Chen Qiang，Pan Yong

School of Electronic and Electrical Engineering, Shanghai University of Engineering Science,
Shanghai 201620, China

ABSTRACT

Deep leaning has become a hot research topic with the rapid development of big data technology. As an important branch of deep learning, convolutional neural network has been widely used in image recognition, and has achieved great success. Convolutional architecture for fast feature embedding (Caffe) with features like speed, extendibility and openness is currently top popular tool of deep learning. In this paper, the authors use Caffe to realize the recognition of individual object in a focus people group. The training images can be obtained from the video recorded by the camera through the method of normalized cross-correlation histogram. The experimental results show that the individual object can be matched accurately by using pre training model. It can be used in practical work like attendance system, criminal investigation field etc.

***Index Terms*—**deep learning, CNN, focus group, object recognition, Caffe

1. INTRODUCTION

In the traditional machine learning, select a good feature is the key to the success of the pattern recognition. When processing unprocessed data, the ability of the traditional machine learning is very limited, so it is called shallow learning. Neural network algorithm simulating biological neural network is a kind of pattern matching algorithm, usually used for solving classification and regression problems. It is a huge branch of machine learning, having hundreds of different algorithms, which deep learning belongs to. Deep learning is a kind of machine learning based on unsupervised feature learning and hierarchical structure, by building models with multiple hidden layers and training massive data to learn more useful features, for the purpose of improving the accuracy of recognition and prediction. In recent years, the rapid development of technology in the big data field makes deep learning to become a hot research topic. As an important branch of deep learning, convolutional neural network has been widely used in image recognition, image classification, and has achieved great success.

In this paper, the author capture videos of every individual object in a focus people group then get the images whose difference between adjacent frames exceeds a threshold using the method of normalized cross-correlation of histograms. Parts of them are used as training images in convolutional neural network, and others not been trained are used as matching images to match individual object by being given the similarity of every people by the pre training model. Pre training and object matching are implemented in Caffe which is the fast feature embedded convolutional neural network framework. At the end of the paper, the experimental results and the social value of this research are given.

2. DEEP LEARNING TOOL

2.1. Neural network and Deep learning

Neural network is a mathematical model of distributed parallel information processing by imitating biological neural network, which is used to solve the problem of recognition and regression [1]. A neural network consists of many connected neurons. Overall function of neural network depends not only on the characteristics of single neurons, also on interactions and connections between neurons. Neurons can be representation of different objects, such as features, letters, concepts, and so on. Neural network processing unit can be divided into three categories: input layer, output layer and hidden layer. Input layer is connected with the external signal and data, and output layer achieve the output of the system processing results. Hidden layer is between input layer and output layer, which can not be observed by the external system. The connection weights between neurons control the strength of the connections between the units. The information processing ability of the whole system is embodied in the connection of each processing unit in the neural network. Figure 1 shows a neural network with one single hidden layer. Neural network is a huge branch of machine learning, there are hundreds of different algorithms, and deep learning is one of the algorithms.

In 2006, Professor Geoffrey Hinton of the University of Toronto published a paper in "science" [2], which expresses two main points of deep learning: a neural network with multi hidden layers has excellent ability of learning features, and the learning features of the data are more essential, so as to facilitate the visualization and recognition; the difficulty of training the deeper neural network can be solved by initializing the layers one by one through unsupervised learning. In essence, deep learning is a kind of machine learning based on unsupervised feature learning and hierarchical structure, through the building contains multiple hidden layer model and training massive data to learn more useful features, thus improving the accuracy of recognition and prediction. In this year there are two other famous papers also described the perspective of deep learning [3] [4].

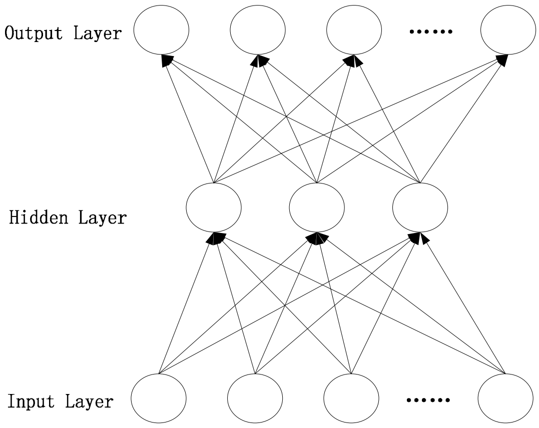

Figure 1. Neural network with one hidden layer

The training process of deep learning is divided into two stages:

1. Unsupervised learning from bottom to top.

In the first stage, it is begin from the input layer to build single layer neurons one by one, each layer using cognitive stage and generation stage to tune the algorithm, each time only tune one layer, tuning layer by layer, as shown in Figure 2. In the cognitive stage, the initial code can be generated by lower input features and the initial encoder weight of the two layers, and then the reconstruction can be generated by initial decoder weight and the initial code, calculation the residual between input information and reconstruction, using gradient descent method to modify the decoder weight. In the generation stage, generate the lower state by the initial code and the modified decoder weights, and then generate new code by the initial encoder weight to produce a new abstraction. Calculating the residual between initial code and new code, modify the encoder weight by gradient descent method, and finally generate the abstract representation of the input as a hidden layer by the modified encoder weight. In the deep learning neural network, each hidden layer can be regarded as the input of next hidden layer.

2. Supervised learning from top to bottom

After the first stage, the network gets the parameters of each layer. In the second stage, add a classifier to the top layer, and then begin supervised learning with labeled data,

using the gradient descent method to fine tune the network parameters, likewise readjusting the encoder weights of all layers.

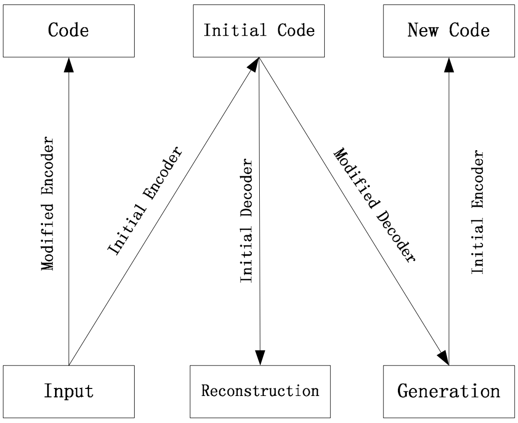

Figure 2. Neural network with one hidden layer

Essentially, the first stage of deep learning is a process of initialization of network parameters by unsupervised learning the structure of the input data, which is different from the traditional neural network, so that the initial value is closer to the global optimum, so as to achieve better effect. In recent years, deep learning has a very wide range of applications in the field of image recognition, language identification. Typical deep learning structure are convolutional neural network (CNN), recurrent neural network (RNN), long-short term memory (LSTM, a special RNN), and CNN get great success in the field of image recognition and object detection.

2.2. Convolutional neural network

Convolutional neural network-CNN is a multi-layer neural network for the recognition of two-dimensional shape [5], it has five main characteristics:

1. Local perception: in the image space, the closer the distance between pixels, the closer the relationship is, and farther pixels correlation was relatively weak. Neurons did not need to be aware of global image, only need to perceive the adjacent local neurons, and at a higher full connected layer the network can get the global information.

2. Weights sharing: in the local connection, the connection between each neuron and a upper layer can be regarded as a method of feature extraction, which is independent on the position, so the learning methods in local part can be used in all the positions in the image, using a convolution kernel convolution in the image to realize weights sharing, as shown in the figure 3. The parameters of the network can be greatly reduced by local perception and weights sharing.

3. Multiple convolution kernels: only using one convolution kernel for feature extraction is not sufficient, so it can increase in the number of convolution kernel to let the network learn more features, and each convolution kernel will generate the image for another image in upper layer.

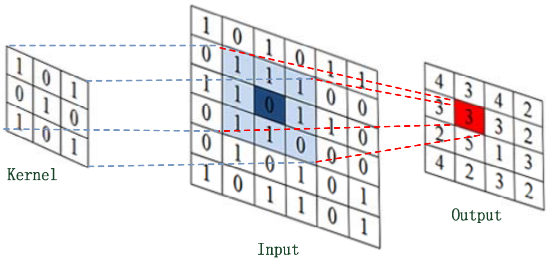

Figure 3. Convolution process

4. Pooling: there are many convolution feature vectors after the convolution, and useful feature in an image region may also can be used in another region, so in order to describe the large image, the characteristics of different positions are aggregated statistics, which can reduce the dimensions of statistical characteristics, but not easy to over fitting, Pooling has two ways: average pooling and maximum pooling.

5. Multiple convolution layers: the feature learned by one convolution layer is often localized, the more the numbers of layers are, the more global the features are. So in practical applications, it is often to add multiple convolution layers in the network, then use the full connecting layer to train.

2.3. Caffe framework

Fast feature embedded convolutional neural network framework Caffe is developed by Dr. Jia Yangqing graduated from Berkeley UC[6], with the characteristics like fast, scalable, open and so on, is one of the top popular tool of deep learning currently. Caffe is purely C++/CUDA architecture, is a tool of convolutional neural network, supporting command line, Python and Matlab interface, seamless switching between CPU and GPU. The authors use the Python interface of Caffe to realize the object recognition. In Caffe, layer definition is consists of 2 parts: layer properties and layer parameters, the layer properties including layer name, layer type and layer connection structure (input and output). The definition of network parameters is very convenient, that can be set arbitrarily. For example, calling the GPU to calculate only need set the solver_mode directly to GPU. Before installing Caffe, Cuda, VS , Opencv , Protobuf, Boost, and other three party libraries should be installed.

The following steps are required to realize object recognition in Caffe:

1. Use convert_imageset command to adjust the images' size, and generate training files in leveldb format or LMDB format;
2. According to the requirement of the model, choose whether to use the compute_image_mean command to generate the average file of the images;
3. According to the number of training images and the training cycle, adjust the training parameters in solver and train_val;
4. Using train command of Caffe begins the training, generating the training model which is a caffemodel file;
5. Using Python interface or Matlab interface in Caffe, realize image recognition, give the similarity of each kind of images.

3. MATCHING OF INDIVIDUAL OBJECT IN A FOCUS PEOPLE GROUP

3.1. Generating the training model of focus group

1. Capturing the videos of focus people group

The videos of focus people group are captured by a single fixed camera in the experiment of this paper, by calling the VideoCapture function of OpenCV in Python, and writing the real-time video to files.

2. Building a training image set

In the video capturing process, the first frame is saved as a key image firstly, and uses it as a benchmark to calculate the difference between the new following video frame and the key image. When the difference exceeds a certain threshold, the following frame is saved as another key image. Calculate constantly to get all key images meet the condition that its difference with the current key image exceeds the threshold, until the video capturing end.

The difference between the frames is obtained by comparing the correlation of their histograms. Histogram is gotten by the calcHist function of OpenCV, and the difference between the frames is gotten by the statement of cv2.compareHist(hist1,hist2,cv2.cv.CV_COMP_CORREL). The third parameter CV_COMP_CORREL controls using the method of the normalized cross correlation of histograms to calculate the frames' difference, the mathematical formula of this method is following:

$$d(h1,h2) = \frac{\sum_i (h1(i) - \bar{h}1)\cdot(h2(i) - \bar{h}2)}{\sqrt{\sum_i (h1(i) - \bar{h}1)^2 \cdot (h2(i) - \bar{h}2)^2}} \qquad (1)$$

where

$$\bar{\bar{h}}1 = \frac{1}{N}\sum_j h1(j)$$

N is the number of bin in the histogram. When the two histograms are perfectly matched, the value of d is 1, didn't match at all to - 1, that is to say the closer the value to 1 the closer the two frames. On the other hand the greater the difference between frames, then the video frame is more representative, save it as a key image. The following figure is a small part of key images of one object people.

Figure 4. Key Frames of individual object

3. Generating the training model

There are several typical image recognition models in CNN, including AlexNet, GoogleNet etc. Experiments show that, when the training data is limited, the trained accuracy of GoogleNet model was not better than AlexNet model, so the authors choose AlexNet model to generate training model. AlexNet model has 5 convolution layers and 3 full connected layers [7]. Detailed structure is shown in Figure 5.

The network parameters can be reduced by convolution and pooling, and using the activation function ReLu to narrow the gap between supervised learning and unsupervised learning, and at the same time the training speed is improved. Further the generalization ability of the network is improved by using local response normalization (LRN) and dropout layers.

The number of the training images in the research is not too big, so we reduce the parameter max-iteration from 450000 to 250000 to short the training time without loss of quality of training. According to the number of images and machine configuration, set parameters such as base_lr, test_iter and batchsize etc.

3.2 Object recognition

1. Select the verification images

The verification images are composed of two parts, one part come from the training video which are not involved the training, the other part come from the new capturing video. Before being verified, the images' size should be adjusted to 256 * 256.

2. Object recognition

By using Python interface provided by Caffe, the jpg image is matched in the training model generated by the pre training through calling classify.py, giving the matching similarity of this object and each people in the focus group.

3. Behavior statistics

After matching object, the report of behavior statistics can be given depending on demand in period of time (per week or per months), such as the weekly attendance, the frequency and time of sb. return to the office in the spare time, the frequency and time of sb. into the confidential room alone, the warning for people on the blacklist being into some special scene etc.

4. EXPERIMENTAL RESULTS AND ANALYSIS

In the research, the authors program on DELL n5110 notebook to realize the video capturing, train on a Desktop computer with 4G memory and GTX960 graphic card, the training time is up to thirty hours. There are 4 peoples in the focus group, including 3 adults and 1 child, 400 images of each people to train, 50 images to test, 10 images to verify. The matching rate of adults reached 100%, but the matching rate of the child is only 50%, there are two possible reasons: one reason is that the child object is small, occupying a smaller proportion of the picture in the case of only using one fixed camera recording, not being good trained; the second reason is that the child is poorly controlled, and the child image quality is low under the bad hardware conditions of capturing videos.

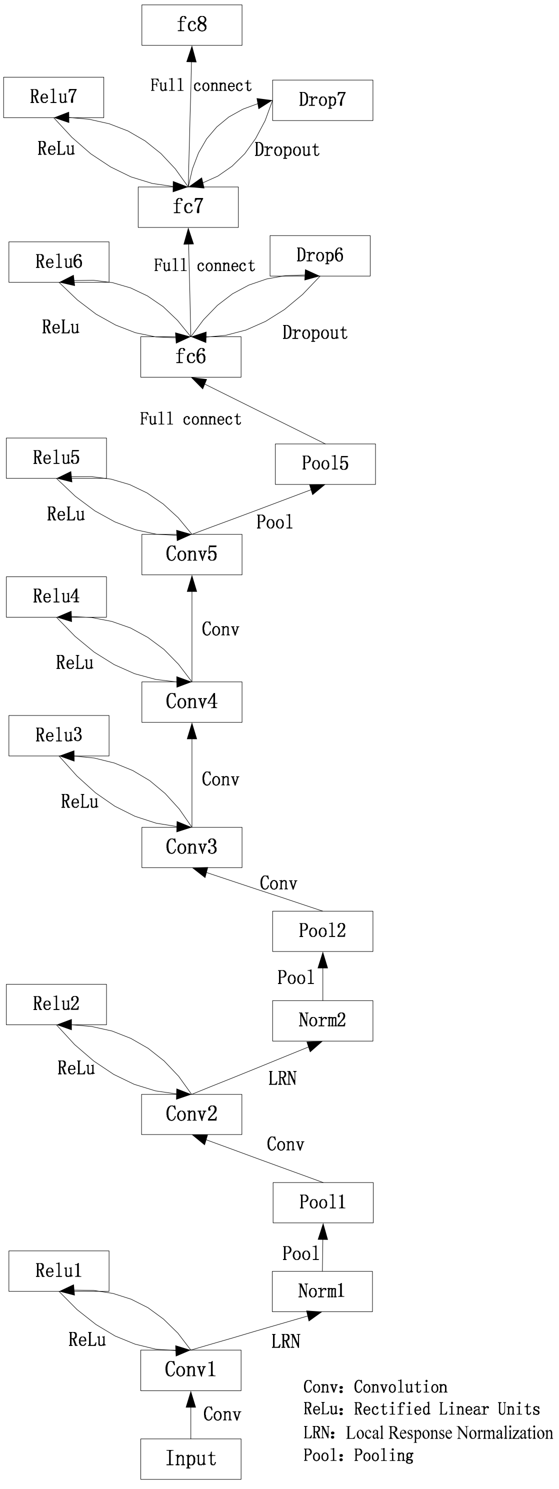

Figure 5. Architecture of AlexNet

Compared with the traditional method of individual object recognition, the method based on deep learning has the following advantages :

Table 1. Compared with the traditional method

	Pre processing	Recognition time period
Traditional method	Spending a lot of time on features extraction	Depending on the configuration of hardware and complexity of coding
Our Method	Unwanted	Millisecond level

Because of the experimental environment being limited, the related experimental materials are captured only in the same environment, supposing the research will be applied to object recognition in a fixed environment. The training images do not be preprocessed, such as cutting and enlarging the object people images. If in the following study the two problems be solved, it is believed that the research technology in the paper will be more widely used.

5. CONCLUSION

In this paper, the authors design a method of recognition of individual object in focus people group based on deep learning. After introducing the tools of deep learning, key steps of the technical route are given including constructing the training images set, generating the training model, matching object people and behavioral statistics etc. At the end, the experimental results and analysis are given to verify the feasibility of the method of recognition of individual object in focus people group based on deep learning. The result of the research in this paper can be applied in the attendance system of company and enterprise, or an auxiliary system to prevent information leakage in a security department, or the security system of some special scenes, has certain practical value, having certain practical application values. Deep learning has a very wide range of applications after 2006, receiving attention from Google, Baidu, Facebook and other large technology companies. They have set up their corresponding research and development departments to strive to improve the performance of their technologies in deep learning and apply their technologies to the practical application. Especially in 2015 Imagenet computer recognition challenge, Microsoft Asia Research Institute of visual computing group reduce the error rate of object recognition classification to 3.57%[8], which is less than the error rate 5.1% of human eye identification, indicating that deep learning will upgrade from mature theory to mature practice, being more successful in the next few years.

6. ACKNOWLEDGMENT

This work was supported in part by the National Natural Science Foundation of China (No. 61272097, 61305014), Innovation Program of Shanghai Municipal Education Commission (No.12ZZ182).

7. REFERENCES

[1] X. Wen, X. Li, and X.W. Zhang, *The Application of MATLAB Neural Network*, National Defense Industry Press, Beijing, 2015.

[2] G.E. Hinton, R. R. Salakhutdinov, “Reducing the Dimensionality of Data with Neural Networks,”*Science*, 313(5786):504-507, 2006.

[3] Y. Bengio, P. Lamblin, D. Popovici and H. Larochelle,G. LayerWise. Training of Deep Networks, in J. Platt et al. (Eds), Advances in Neural Information Processing Systems 19 (NIPS 2006), pp. 153-160, MIT Press, 2007.

[4] M.’A. Ranzato, C. Poultney, S. Chopra and Y. LeCun. Efficient Learning of Sparse Representations with an Energy-Based Model, in J. Platt et al. (Eds), Advances in Neural Information Processing Systems (NIPS 2006), MIT Press, 2007.

[5] Y. LeCun, Y. Bengio, and G.E. Hinton, “Deep learning,”Nature, 521:436-444, 2015.

[6] http://caffe.berkeleyvision.org.

[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural network,” InNIPS, 2012.

[8] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” arXiv preprint arXiv:1512.03385, 2015.