Pisco Med Publishing

Original Research Article

Application Research of Deep Convolutional Neural Network in Computer Vision

Lei Wang*

Yunnan Forestry Technological College, Kunming 650000, Yunnan, China. E-mail: tim200e@163.com

Abstract: As an important research achievement in the field of brain like computing, deep convolution neural network has been widely used in many fields such as computer vision, natural language processing, information retrieval, speech recognition, semantic understanding and so on. It has set off a wave of neural network research in industry and academia and promoted the development of artificial intelligence. At present, the deep convolution neural network mainly simulates the complex hierarchical cognitive laws of the human brain by increasing the number of layers of the network, using a larger training data set, and improving the network structure or training learning algorithm of the existing neural network, so as to narrow the gap with the visual system of the human brain and enable the machine to acquire the capability of "abstract concepts". Deep convolution neural network has achieved great success in many computer vision tasks such as image classification, target detection, face recognition, pedestrian recognition, etc. Firstly, this paper reviews the development history of convolutional neural networks. Then, the working principle of the deep convolution neural network is analyzed in detail. Then, this paper mainly introduces the representative achievements of convolution neural network from the following two aspects, and shows the improvement effect of various technical methods on image classification accuracy through examples. From the aspect of adding network layers, the structures of classical convolutional neural networks such as AlexNet, ZF-Net, VGG, GoogLeNet and ResNet are discussed and analyzed. From the aspect of increasing the size of data set, the difficulties of manually adding labeled samples and the effect of using data amplification technology on improving the performance of neural network are introduced. This paper focuses on the latest research progress of convolution neural network in image classification and face recognition. Finally, the problems and challenges to be solved in future brain-like intelligence research based on deep convolution neural network are proposed.

Keywords: Convolution Neural Network; Deep Learning; Computer Vision; Image Recognition

1. Introduction

It is a big scientific dream that scientists have been exploring and pursuing for a long time to let machines learn quickly and accurately in a manner similar to the human brain. For decades, many research achievements in the fields of brain neuroscience and psychology in terms of human brain structure and cognitive mechanism have been converted into computational models in the field of artificial intelligence, which greatly promoted the development and progress of the latter. Artificial neural networks were proposed in this context. It is a neural network system built by artificial means by using a computational model to simulate the structure and function of the brain's nervous system, using a large number of simple arithmetic units. The birth and development of artificial neural networks is one of the most im-

doi: 10.18282/jnt.v2i2.886

portant research results in the field of brain-like computing.

Since the earliest proposed M-P neuron model and Hebb learning rules in the 1940s, hundreds of neural network models have been proposed in the field of artificial neural networks, among which representative networks include perceptions, backpropagation networks, and self-organization. Mapping network, Hopfield network, Boltzmann machine, adaptive resonance theory, etc., have been successfully applied in technical fields such as handwriting recognition, speech recognition, image recognition and natural speech processing. Since 2011, researchers have first applied deep learning techniques to speech recognition issues, increasing accuracy by 20%-30%, making the biggest breakthrough in more than a decade. Only one year later, the deep learning model based on convolutional neural networks has achieved very large performance improvements on large-scale image classification tasks, setting off a tide of deep learning research.

Currently, Convolutional Neural Net-Works (CNN) is a widely used artificial neural network and the first deep neural network to be successfully trained. In 1962, Hubel and Wiesel conducted an in-depth study on the visual cortical cells of cats, and proposed a cognitive mechanism model of the advanced animal visual system[1]. The model proposes that the advanced animal visual neural network is composed of simple cells and complex cells (as shown in **Figure 1**). The receptive field of the simple cells at the bottom of the neural network only corresponds to a specific area of the retina, and only responds to the boundary line in a specific direction in the area. Complex cells have a larger receptive field by clustering simple cells with a specific orientation, and acquire characteristics with certain invariance. Simple cells in the upper layer cluster complex cells with a higher symbiosis probability, resulting in more complex boundary features. By alternating layers of simple and complex cells layer by layer, the visual neural network acquires the ability to extract highly abstract and invariant image features.

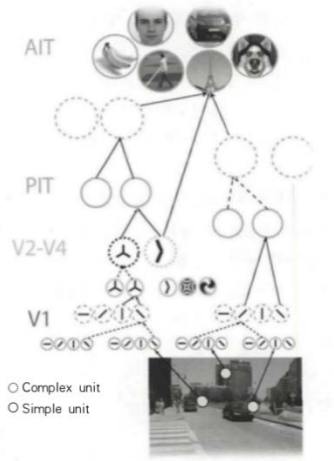


Figure 1. Human brain visual channel neural network.

At present, the advantages of Deep CNN over traditional machine learning algorithms are expanding. Traditional learning methods cannot compete with depth learning in many fields, such as handwriting recognition, image classification, image semantic understanding, speech recognition, natural language understanding and other technical fields. There are several reasons why neural networks can be rejuvenated. First of all, the abundant network images and large-scale labeled data sets alleviate the problem of training and fitting to a great extent. For example, ImageNet data sets contain 21,841 image categories, totaling 14,197,122 images. Secondly, the rapid development of computer hardware provides powerful computing power, making it possible to train large-scale neural networks. A single GPU (Graphics Processing Unit) chip can integrate thousands of operation cores to provide high parallel computation for neural networks dominated by high-order matrix operations. In addition, the neural network model design and training methods have made great progress. For example, in order to improve the training of neural networks, researchers have proposed the optimization of deep structure and the improvement of training and learning methods, including the use of

ReLU activation function, the use of dropout for network training, and the use of batch normalization technology to normalize the data distribution of features.

The study of neural networks is closely related to the study of human vision. In order to further improve the performance of neural networks, it has become a research direction that attracts more and more attention from academia to find the next breakthrough for the study of convolution neural networks by drawing on the latest research results of human brain vision system. Through in-depth research on the visual pathway of human brain, scientific enlightenment can be obtained from the structure of visual neural network, the expression of visual information in each layer, and the visual cognitive mechanism of high-level network. Combined with relevant technologies such as mathematics, statistics and engineering, a machine vision system closer to the understanding and cognitive ability of human brain environment can be produced.

2. Convolution neural network

2.1 Concept

Convolution neural network is a multi-layer artificial neural network specially designed for processing two-dimensional input data. Each layer in the network consists of a plurality of two-dimensional planes, and each plane consists of a plurality of independent neurons. The neurons in the adjacent two layers are connected with each other, while the neurons in the same layer are not connected. CNNs is inspired by the early Time-Delay Neural Network (TDNNs)[2]. TDNN reduces the computational complexity in the network training process by sharing weights in the time dimension, and is suitable for processing voice signals and time series signals. CNNs adopts a weight sharing network structure to make it more similar to biological neural network. At the same time, the capacity of the model can be adjusted by changing the depth and breadth of the network, and it also has strong assumptions for natural images (statistical stability and local correlation of pixels). Therefore, CNNs can effectively reduce the learning complexity of the network model and has fewer network connections and weight parameters than fully connected networks with equivalent size in each layer, thus making it easier to train.

2.2 Convolution neural network structure

Convolution neural network is a multilayer neural network composed of convolution layer used for feature extraction and subsampling layer used for feature processing. A typical convolutional neural network structure[3] is shown in **Figure 2.** the input of the network is a handwritten digital image, and the output is its recognition result. after the input image is processed by several “convolutions” and “samples”, the mapping between the output target and the full connection layer network is realized. Generally, in convolutional neural networks, each layer of neuron nodes is only connected to neuron nodes in its adjacent upper and lower local receptive fields. This viewpoint of local connection is consistent with the viewpoint of local perception found by Hubel and Wiesel from feline’s visual system. The size of the input image in **Figure 2** is 32×32 pixels, including three channels of r, g and b. Convolution layer C1 performs convolution filtering on each channel of the input image by using a plurality of convolution checks with a size of 5×5, and adopts local features of the image to obtain a feature map with the same number of convolution kernels and a size of 28×28. Then these feature maps are combined in a certain way as the output of convolution layer. After the original feature map in the map passes through the sampling layer S2, its size is reduced to 14×14, where each neuron in the feature map is connected to the 2×2 neighborhood of the corresponding feature map in the previous layer, and the output is calculated accordingly. Neurons in the convolution layer of the convolution neural network are simple cells in the simulated Hubel-Wiesel model, neurons in the down sampled layer are complex cells, and neurons on the feature map share the same convolution kernel and correspond to simple cells with a certain orientation. Several convolution-sampling operations can be carried out to obtain feature maps with small size but large number. The feature map is unfolded in a certain way, spliced into one-dimensional vectors and input into the full connection layer, and then connected through a plurality of full connection layers and output layers to complete the identification task.

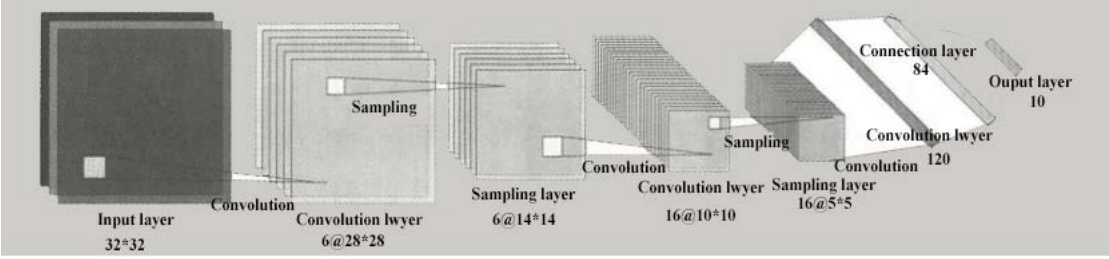
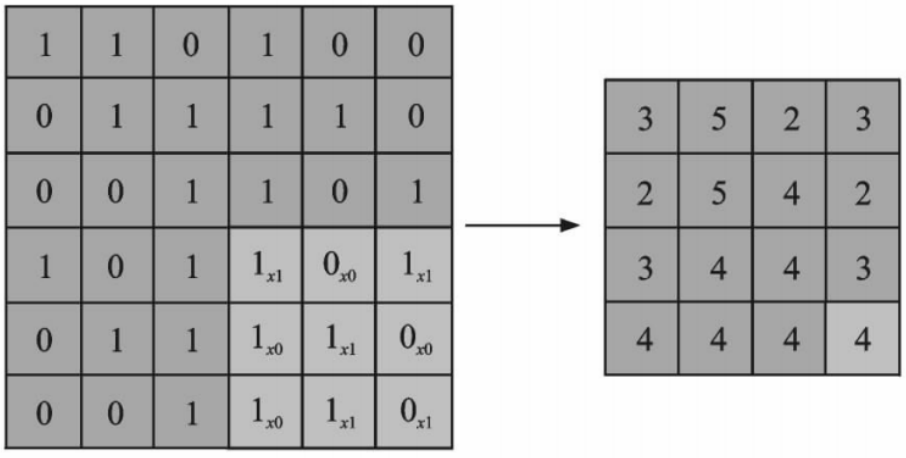


Figure 2. Typical structure of convolution neural network[3].

2.3 Convolution feature extraction

Natural images have their inherent characteristics, that is, for one part of the image, its statistical characteristics are the same as other parts. This means that the features learned in this part can also be used in the other part, so the same learning features can be used for all positions on the image. In other words, for large-size image recognition problems, firstly, a small local area is randomly selected from the image as a training sample, and some features are learned from the small sample. Then, these features are used as filters to perform convolution operation with the original entire image, thus obtaining activation values of different features at any position in the original image. Given a large-size image with a resolution of r×c, it is defined as x_{large}. Firstly, a×b small-size image sample x_{small} is extracted from x_{large}, and k features and activation values f ($w^{(1)}$ x_{small}+$b^{(1)}$) are obtained by training sparse self-encoders, where $w^{(1)}$ and $b^{(1)}$ are parameters obtained by training. Then, for each x_s of a×b size in x_{large}, the corresponding activation value fs ($w^{(1)}$ x_{small}+$b^{(1)}$) is calculated. further, the activation value of x_{small} is convolved with these activation values fs to obtain k×(r-a+1)×(c-b+1) convolved feature maps. The schematic diagram of two-dimensional convolution calculation is shown in **Figure 3.** For example, for an original input image with a resolution of 128x128, it is assumed that 2008×8-sized feature fragments of the image have been obtained through pre-training. Then, by using these 200 feature fragments to convolve each 8x8 small block region in the original image, each feature fragment can obtain a 121×121 convolved feature map, and finally the entire image can obtain a 200×121×121 convolved feature map.



(a) Image (b) Convolution feature

Figure 3. Schematic diagram of two-dimensional convolution operation.

3. Application of convolution neural network

Convolution neural network is an important research achievement in the field of brain like computing in recent ten years. It has achieved great success in many fields such as computer vision, speech recognition, natural language processing, multimedia and so on. In all kinds of tasks in the field of computer vision, the image classification task is to distinguish different types of objects (such as birds, people, cars, planes, etc.) according to the different features reflected in the image information, that is, to assign a semantic category mark to each picture, while the object detection is

to locate the region where a certain type of object appears in the image. Different from the image classification task to establish image-level understanding, image semantic understanding needs to obtain image pixel-level target classification results. The generation of picture title is also based on the semantic understanding of the picture, requiring automatic generation of natural language to describe the target of the picture and the relationship between the targets. Compared with image classification and target detection, which focus on distinguishing or locating objects of multiple or single classes, face recognition and pedestrian re-recognition tasks focus on the identification of human face and pedestrian respectively. Another task, image super-resolution, can provide clearer images and more image details, providing better input for high-level visual tasks. This section will focus on the latest research progress of convolution neural network in image classification and face recognition.

3.1 Image classification

Image classification is an important application in the field of computer vision, mainly refers to a given picture, which enables the computer to classify it into an appropriate category according to the content in the picture and assign a semantic category marker. The most important progress of deep convolution neural network in image classification is reflected in the image classification task in the Image Net LSVRC challenge. In addition to ImageNet image data sets, commonly used data sets for image classification include Caltech-101[4], Caltaech-256, Tiny-Image[5], Sun[6], etc. **Table 1** lists some commonly used data sets and their important information in the field of image classification.

Name	Number of categories included	Number of pictures
Caltaech-101[4]	101	9,146
Caltaech-256	256	30,607
Tiny-Image[5]	75,062	79,302,017
SUN[6]	899	130,519
ImageNet[7]	21,841	14,197,122

Table 1. Data sets commonly used in image classification

3.2 Face recognition

An important challenge in the field of computer vision is face recognition. Face recognition includes two tasks: face verification and face recognition. The task of face verification is to determine whether two face photos belong to the same person, which belongs to the two classification problem. The correct rate of random guess is 50%. The task of face recognition is to divide an unknown face image into one of N identity categories. For this multi-classification problem, the correct rate of random guess is 1/N. Face recognition is more challenging, and its difficulty increases with the number of categories. The biggest challenge of face recognition is how to distinguish the intra-class changes caused by light, posture and expression, and the inter-class changes caused by different identity categories. The distribution of these two kinds of changes is extremely complex and nonlinear, and the traditional linear model cannot effectively distinguish them. Convolution neural network can remove as many intra-class changes as possible through multi-layer nonlinear transformation while preserving inter-class changes. LFW (Labeled Faces in the Wild)[8] is the most famous public face verification test set today. It collects face photos of more than 5,000 celebrities from the Internet to evaluate the face verification performance of the algorithm under uncontrollable conditions (as shown in **Figure 4**). On LFW test set, the accuracy rate of human eyes is 99.53%[9], while the highest accuracy rate of non-depth learning algorithm is 96.33%[10], while the current depth learning can reach 99.47% verification rate. At present, many face recognition algorithms are based on face recognition on offline data sets containing a large number of face categories.

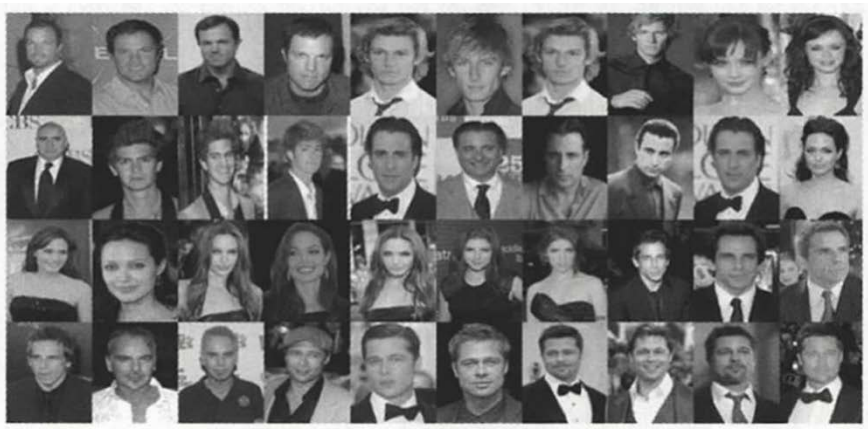


Figure 4. LFW face data set.

In 2013, Sun *et al*.[11] used the face recognition task as the supervisory signal and used the convolution neural network to learn the face features, and achieved 92.52% recognition rate on LFW. Although this result is lower than the following depth learning methods, it also surpasses most non-depth learning algorithms. DeepID[12] and Deep Face[13] published on CVPR2014 have achieved 97.45% and 97.35% recognition rates on LFW using face recognition as a supervisory signal. They use convolutional neural network to predict the category of input face images and select the highest hidden layer as the face feature (as shown in **Figure 5**). In the training process, the neural network needs to distinguish a large number of face categories (*e.g.* 1000 face categories in DeepID), so the face features contain rich inter-class change information and have strong generalization ability.

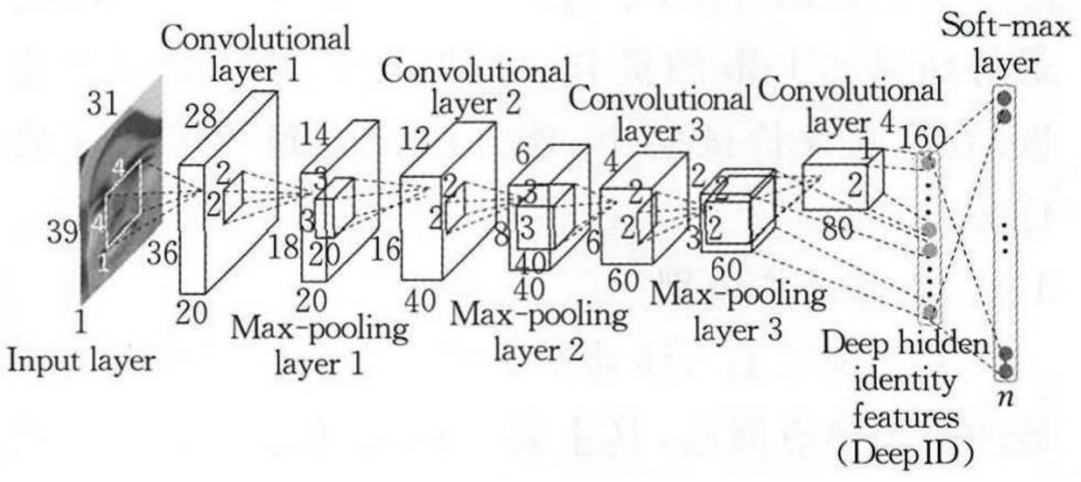


Figure 5. DeepID network structure.

DeepID2[14] uses face recognition and face recognition as supervisory signals. The obtained face features minimize intra-class changes while maintaining inter-class changes, thus improving the face recognition rate on LFW to 99.15%. Using Titan GPU, DeepID2 only needs 35ms to extract the features of a face image, and it can be done offline. After PCA compression, 80-dimensional feature vectors are finally obtained, which can be used for fast online face comparison. In the following work, DeepID2+ further improved DeepID2 by enlarging the network structure, increasing training data, and adding supervision information at each layer, reaching a recognition rate of 99.47% in LFW. Facenet proposes to use Triplet network results to learn face features. The input samples are two similar pictures and one different picture. Euclidean distance is directly used in the last hidden layer to measure the similarity between the input images. FaceNet achieves 99.63% verification accuracy on LFW dataset.

4. Conclusion

Artificial neural network is a network computing structure composed of basic mathematical computing units and their interactive connections. It is used to simulate the processing process of information in human brain and let machines actively acquire the laws contained in data through learning and training mechanisms. Based on one of the learning models, the deep convolution neural network, this paper introduces the current technical methods to improve the performance of the deep convolution network and its application in the field of computer vision, and analyzes the characteristics of the human brain vision mechanism and some theoretical implications for the current computing model. On the one hand, in-depth learning has a wide range of applications and is highly versatile, so it can continue its efforts to expand it to other application fields. On the other hand, in-depth learning still has a lot of potential and is worth exploring and discovering. For the future, although many of the contents discussed before are supervised learning (for example, the last layer of the training network will calculate a loss value according to the real value, and then adjust the parameters), and supervised learning has indeed achieved great success.

References

1. Hubel DH, Wiesel TN. Receptive files, binocular interaction and functional architecture in the cat's visual cortex. The Journal of Physiology 1692; 160(1): 106-154.
2. Waibel A, Hanazawa T, Hinton G, *et al*. Phonemere recognition using time-delay neural network. Acoustics, Speech and Signal Processing, IEEE Transaction 1989; 37(3): 328-339.
3. Lecun Y, Bottou L, Bengio Y, *et al*. Gradient-based learning applied to document recognition. Proceedings of the IEEE 1998; 86(11): 2278-2324.
4. Li F, Fergus R, Perona P. Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories. Computer Visionand Image Understanding 2007; 106(1): 59-70.
5. Torralba A, Fergus R, Freeman WT. 80 million tiny images: A large data set for nonparametric object and scene recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 2008; 30(11): 1958-1970.
6. Xiao J, Hays J, Ehinger KA, *et al*. Sun database: Large-scale scene recognition from abbey to zoo. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition can Francisco; USA; 2010. p. 3485-3492.
7. Deng J, Dong W, Socher R, *et al*. Imagenet: A large-scale hierarchical image database. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; Miami, USA; 2009. p. 248-255.
8. Huang GB, Mattar M, Berg T, *et al*. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. University of Massachusetts, Amherst. USA Technical Report 2008; 07-49.
9. Kumar N, Berg AC, Belhumeur PN, *et al*. Attribute and simile classifiers for face verification. Proceedings of the International Conference on Computer Vision; Kyoto, Japan; 2009. p. 365-372.
10. Chen D. Cao X, Wen F, *et al*. Blessing of dimensionality high dimensional feature and its efficient compression for face verification. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; Portland, USA; 2013. p. 3025-3032.
11. Sun Y, Wang X, Tang X. Hybrid deep learning for face verification. Proceedings of the IEEE International Conference on Computer Vision; Sydney, Australia; 2013. p. 1489-1496.
12. Sun Y, Wang X, Tang X. Deep learning face representation from predicting 10,000 classes. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Columbus; USA; 2014. p. 1891-1898.
13. Taigman Y, Yang M. Ranzato M, *et al*. Deepface: Closing the gap to humarr level performance in faceverification. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; Columbus, USA; 2014. p. 1701-1708.
14. Sun Y, Chen Y, Wang X, *et al*. Deep learning face repre-sentation by joint identificatior-verification. Proceedings of the Advances in Neural Information Processing Systems Montreal; Canada; 2014. p. 1988-1996.