Review of Deep Convolution Neural Network in Image Classification

Ahmed Ali Mohammed Al-Saffar, Hai Tao, Mohammed Ahmed Talab
Faculty of Computer Systems and Software Engineering
Universiti Malaysia Pahang
Pahang, Malaysia
haitao@ump.edu.my

***Abstract*—With the development of large data age, Convolutional neural networks (CNNs) with more hidden layers have more complex network structure and more powerful feature learning and feature expression abilities than traditional machine learning methods. The convolution neural network model trained by the deep learning algorithm has made remarkable achievements in many large-scale identification tasks in the field of computer vision since its introduction. This paper first introduces the rise and development of deep learning and convolution neural network, and summarizes the basic model structure, convolution feature extraction and pooling operation of convolution neural network. Then, the research status and development trend of convolution neural network model based on deep learning in image classification are reviewed, which is mainly introduced from the aspects of typical network structure construction, training method and performance. Finally, some problems in the current research are briefly summarized and discussed, and the new direction of future development is forecasted.**

Keywords—Deep learning; convolution neural network; image recognition; Image Classification.

I. INTRODUCTION

Computer vision (CV) is a study of how to use computer simulation of human visual science, its main task is through the collection of images (or video) analysis and understanding, to make judgments or decisions. In the past few decades, CV has made great progress and development. The Image recognition is a kind of technology that uses computer to process, analyze and understand the image to identify the target and object of different modes. It is a major research direction in the field of computer vision. In the image-based intelligent data acquisition and Processing has a very important role and impact. The use of image recognition technology can effectively deal with the detection and identification of specific target objects (such as face, handwritten characters or goods), image classification and subjective image quality assessment and other issues.

At present, image recognition technology has great commercial market and good application prospect in Internet applications such as image search, commodity recommendation, user behavior analysis and face recognition. At the same time, high-tech such as intelligent robot, unmanned driving and unmanned aerial vehicle Industry and biology, medicine and geology and many other disciplines have broad application prospects. In the early image recognition system, feature extraction methods such as Scale-invariant feature transform (SIFT [1]) and histogram of oriented gradients (HOG [2]) were used, and then the extracted Feature input classifier for classification and recognition. These features are essentially a feature of manual design. For different identification problems, the extracted features have a direct impact on the performance of the system, so the researchers need to study the problem areas to be studied in order to design Adaptability to better features, thereby improving system performance. This period of image recognition system is generally for a specific identification task, and the size of the data is not large, generalization ability is poor, it is difficult in the practical application of the problem to achieve accurate identification effect.

II. CONVOLUTION NEURAL NETWORK

Deep learning is a branch of machine learning, which is one of the major breakthroughs and research hotspots in machine learning in recent years. In 2006, Geoffery Hinton, a professor of computer science at the University of Toronto, and his student, Ruslan Salakhutdinov, published an article in the international top academic journal Science, [3], for the first time in the depth of learning. This paper mainly points out two points: (1) Artificial neural network with multiple hidden layers has a very powerful feature learning ability. The characteristics extracted by the training model have more abstract and more basic expression of the original input data, (2) By using the unsupervised learning algorithm to achieve a method called "layer initialization" to achieve the input data information hierarchical expression, which can effectively reduce the depth of the neural network Training difficulty. Subsequently, the depth of learning in academia and industry continues to heat up, in the speech recognition, image recognition and natural language processing and other fields to obtain a breakthrough. Since 2011, the researchers first in the voice recognition problem on the application of in-deep learning technology, the accuracy rate increased by 20% to 30%, made more than a decade the biggest breakthrough. Only a year later, the deep learning model based on convolution neural network has achieved great performance improvement in large-scale image classification tasks, and set off the upsurge of deep learning. In the literature [4], two kinds of acoustic modeling methods based on deep neural network are proposed, which are more effective than the traditional modeling method, and have been made larger in Uyghur's large vocabulary

978-1-5386-3849-1/17/$31.00 ©2017 IEEE

continuous speech recognition of the performance of the upgrade. At present, Google, Microsoft and Facebook and many other Internet technology companies competed to invest a lot of resources, research and development layout of large-scale depth of learning system.

In the early 1960s, Hubel and Wiesel, through the study of cat's visual cortical system of cat, proposed the concept of receptive field [5] and found the hierarchical processing mechanism of information in the visual cortical pathway, Nobel Prize in Physiology or Medicine. By the mid-1980s, Fukushima et al. [6], which was based on the concept of receptive field, could be seen as the first realization of Convolution neural networks (CNNs) and the first neuron-based Between the local connectivity and the hierarchical structure of the artificial neural network. The neural cognition machine decomposes a visual pattern into many subpatterns, and these subpattern features are processed by hierarchical cascaded feature planes so that the model is very good even in the case of small targets of the target object Recognition ability. After that, the researchers began experimenting with the use of an artificial neural network (actually a shallow model with only one hidden layer node) called a multi-layer sensor [7] instead of manually extracting features and using A simple stochastic gradient descent method to train the model, and further proposed a back propagation algorithm for calculating the error gradient, which was subsequently proved to be very effective [8]. In 1990, LeCun et al. [9] studied the handwritten digital identification problem, first proposed the use of gradient back propagation algorithm training convolution neural network model, and in MNIST [10] handwritten digital data set to show relative to the time Other methods for better performance. The success of the gradient back propagation algorithm and the convolution neural network brings new hope to the machine learning field. It opens up the wave of machine learning based on the statistical learning model, and also brings the artificial neural network into a new stage of vigorous development. At present, the convolution neural network has become a research hotspot in the field of speech analysis and image recognition. It is the first real learning model of successful training of multi-layer neural networks, which is more obvious when the input of the network is multidimensional The advantages. Conch neural network has been applied to different large-scale machine learning problems such as speech recognition, image recognition and natural speech processing, as the new machine learning boom has been explored in depth.

A. Concept

Convolution neural network is a multi-layer artificial neural network specially designed to handle two-dimensional input data. Each layer in the network is composed of multiple two-dimensional planes, and each plane consists of multiple independent neurons Composition, adjacent two layers of neurons connected to each other, and in the same layer of neurons are not connected between. CNNs are inspired by the early time delay neural networks [11] and TDNNs. TDNN reduces the computational complexity in the network training process by sharing the weights in the time dimension, and is suitable for processing speech signals and time Sequence signal. CNNs use a weight-sharing network structure to make it more similar to a biological neural network, and the capacity of the model can be adjusted by changing the depth and breadth of the network, and has a strong assumption for natural images (statistical smoothness and local Correlation) . Therefore, CNNs can effectively reduce the learning complexity of the network model, have fewer network connections and weight parameters, and are more likely to be trained than the fully connected network with a considerable size.

B. Network Structure

A simple convolution neural network model structure diagram shown in Fig. 1, the network model consists of two convolution layers (C1, C2) and two sub-sampling layer (S1, S2) alternately. First, the original input image is convoluted by three trained filters (called convolution kernel) and addable bias vectors. Three feature maps are generated in the C1 layer, and then, for each feature map The localized regions are weighted and averaged, and three new feature maps are obtained in the S1 layer through a nonlinear activation function. These feature maps are then convoluted with the three trained filters of the C2 layer, and three feature maps are output through the S2 layer. The final output of the S2 layer is vectorized and then input into the traditional neural network for training.

C. Convolution Feature Extraction

Natural images have its inherent characteristics, that is, for a part of the image, its statistical characteristics and other parts of the same. This means that the features learned in this section can also be used on another part, so the same learning feature can be used for all positions on the image. In other words, for large-size image recognition problems, a small piece of local data is randomly selected from the image as a training sample, some features are learned from the small sample, and then these features are used as filters, with the original whole image For convolution operations, resulting in the original image at any position on the different characteristics of the activation value. Given a large image with a resolution of r × c, it is defined as xlarge. First, a small sample of xx is taken from xlarge, and k features and activation values f (W) are obtained by training sparsely from the encoder (1) xsmall + b (1)), where W (1) and b (1) are the trained parameters. And then calculate the corresponding activation value fs (W (1) xsmall + b (1) for each x × the size of xs in xlarge), and further use the

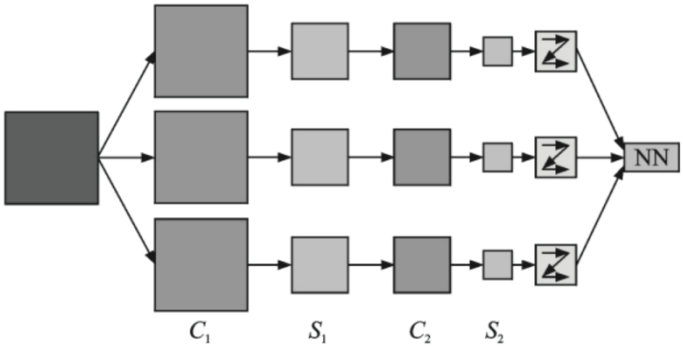

Fig. 1. Simplified convolution neural network structure.

activation value of xsmall and convolution of these activation values fs, We obtain the feature map of k × (r - a + 1) × (c - b + 1) convolution. Two-dimensional convolution calculation diagram shown in Fig. 2. For example, for a raw input image with a resolution of 128 x 128, it is assumed that 200 8x8 size feature fragments of the image have been obtained by pre-training. Then, by using these 200 feature fragments, each 8 × 8 small block region in the original image is convolved, and each feature fragment can get a convolution feature map of 121 × 121, and finally the whole image can be obtained 200 × 121 × 121 convolution feature map. Lu Hongtao et al: Research on the Application of Depth Convolution Neural Network in Computer Vision 3 Fig. 2 Schematic diagram of two-dimensional convolution operation

D. Pooling Operation

By extracting the features extracted from the convolution layer into the classifier for training, the final classification result can be output. Theoretically, all the features extracted from the convolution layer can be input directly into the classifier, but this will require very large computational overhead, especially for large-size high-resolution images. For example, for an image sample with an input of 96 × 96 size, it is assumed that convolution operations are performed using 200 8 × 8 size convolution cores in the convolution layer. Each convolution kernel outputs one (96 - 8 + 1) × (96 - 8 + 1) = 7 921 dimension, the final convolution layer will output a feature vector of 7 921 × 200 = 1 584 200 dimensions. The ability to input such high-dimensional features into the classifier requires a very large computational resource and a serious over-fitting problem. However, since the image has a "static" attribute, the feature obtained in a local region of the image is highly likely to apply equally in another local area. Thus, it is possible to perform aggregate statistical operations on the characteristics of the different locations in a local area of the image, which is referred to as "pooling". For example, calculate the maximum (or average) of a convolution feature in the local area, called the maximum pool (or average pool). Specifically, assuming that the pooled area size is m × n, after the convolution feature is obtained, the convolution feature is divided into a plurality of m × n size disjoint areas, and then the pooling operation is performed on these areas, Get the characteristic map after pooling in. Fig. 3.

The maximum pooling is performed on a 4-block non-coincident sub-region using a 3 × 3 size window to obtain a pooled feature map. If the continuous range in the image is selected as the pooled area and only the convolution features generated by the same implicit neurons are used for pooling, these pooled feature units have translation invariance. That is, even if the object in the original image produces a small translation, the same pooling feature can still be obtained, and the classifier can still output the same classification result. These statistical features can not only greatly reduce the dimension of the eigenvector, but also reduce the computational effort required by the training classifier and expand the training data effectively, which is helpful to prevent over-fitting.

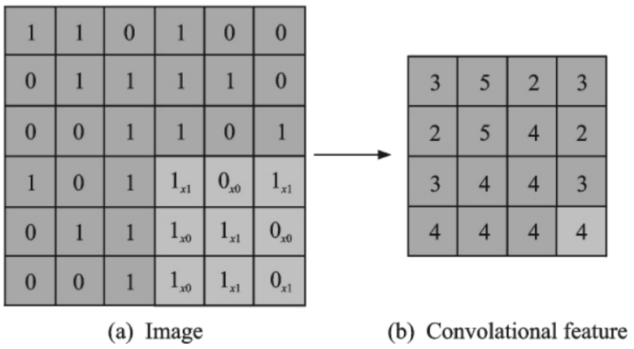

Fig. 2. Illustration of two-dimensional convolution operation

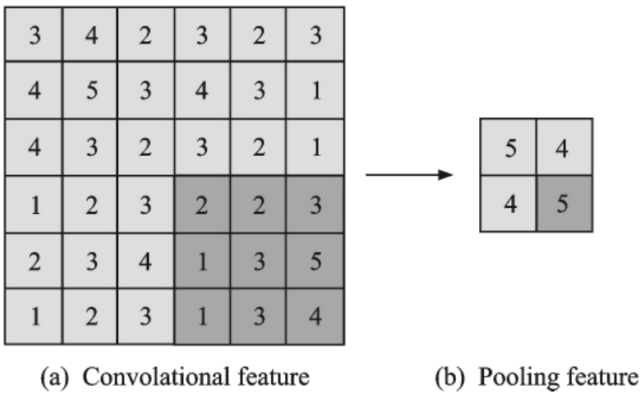

Fig. 3. The maximum pool operation operation diagram

III. IMAGE CLASSIFICATION

The Image classification problem is through the analysis of the image, the image is classified as a number of categories of one, the main emphasis on the overall image of the semantic judgments. There are a lot of tagged data sets for evaluating image classification algorithms, such as CIFAR-10/100 [12], Caltech-101/256 [13-14] and ImageNet [15], where ImageNet contains more than 15 000 000 High-resolution images with labels, these images are divided into more than 22,000 categories. From 2010 to the present, the annual ImageNet Large Scale Visual Recognition Challenge (ILSVRC) image classification competition is an important event to evaluate the image classification algorithm. Its data set is a subset of ImageNet, which contains millions of images, which are divided into 1,000 categories. Among them, the 2010 and 2011 winners are using the traditional image classification algorithm, mainly using SIFT, LBP [16] and other algorithms to manually extract features, and then extract the characteristics used to support the support vector machine (Support vector machine, SVM) and other classifiers for classification, the best result is 28. 2% error rate [17]. ILSVRC2012 is an important turning point in the field of large-scale image classification. In this tournament, Alex Krizhevsky et al. Proposed AlexNet [18] for the first time to apply deep learning to large-scale image classification and achieved 16. 4% error rate, which is about 10% lower than the second team using the traditional algorithm. As shown in Fig. 4, AlexNet is an 8-layer convolution neural network, the first five layers are convolutions, and the last three are all connected layers, where the last layer is classified by

softmax. The model uses Rectified linear units (ReLU) to replace the traditional Sigmoid and tanh functions as neuron's nonlinear activation functions, and proposes the Dropout method to reduce the over-fitting problem.

After the development of AlexNet model, the model based on deep convolution neural network began to replace the traditional image classification algorithm to become the mainstream method used in the ILSVRC image classification competition team. ILSVRC2013's winning team Clarifai [19] proposed a set of convolution neural network visualization method, the use of deconvolution network of AlexNet each convolution layer to visualize, in order to analyze the characteristics of each layer to learn, so Deepened the understanding of why the convolution neural network can achieve good results in image classification, and thus improved the model, made 11. 7% error rate. ILSVRC2014 image classification results compared to the previous year made a major breakthrough, which won the Google team made by Google Team [20] to 6. 7% error rate reduces the error rate of the image classification game to half of the best record of the past.

The enhancement of the convolution neural network is based on the multi-scale processing method. This paper proposes the Inception module based on Network in network [21]. The structure of the Inception module is shown in Fig. 5, and its main idea is to find the optimal local sparse structure of the image and replace it with a dense component. In this way, we can achieve effective dimensionality reduction, which can increase the width and depth of the network under the same computing resources. On the other hand, we can reduce the parameters that need to be trained, so as to reduce the over-fitting problem and improve the model's ability to promote The In ILSVRC2014, The 1% error rate of the third place is from

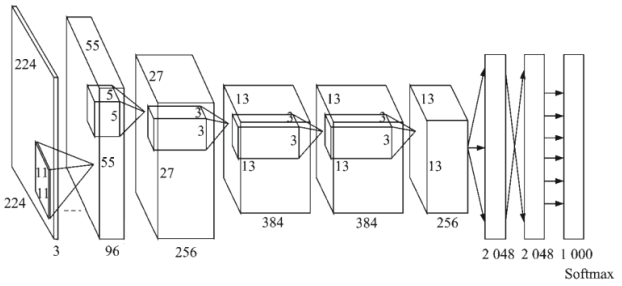

Fig. 4. Simplified AlexNet model structure

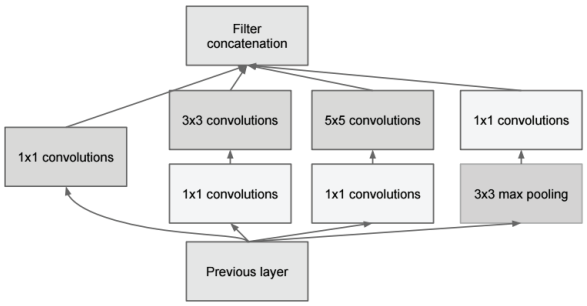

Fig. 5. Simplified Inception module structure [20] .

the Microsoft Asia Research Institute team designed by SPPNet [22], they proposed a new pooling method called the space pyramid pool, as shown in Fig. 6. Most of the convolution neural network models require that the input image size be fixed, so the original image needs to be cut, which will result in the loss of the original image information; or the need to adjust the size and aspect ratio of the image, Resulting in distortion. Note that the size of the input image is not limited, only the full connection layer due to the number of parameters fixed, you need to ensure that the input dimension is fixed. However, the output dimension of the convolution layer changes as the input dimension changes, so it is necessary to ensure that the input image size is fixed. Therefore, the role of space pyramid pooling is the loss of any dimension Lv Hongtao et al: Depth Convolution neural network in the application of computer vision review, so that the network can accept any size of the image as input.

The pooling method divides the input into fixed number of local space blocks and maximizes pooling in each block to ensure that the output dimension is fixed. Using multi-level spatial block division method, you can extract the characteristics of different scales. At the beginning of 2015, PReLU-Nets [23], a researcher at the Microsoft Asia Research Institute, made 4 on the ILSVRC image classification data set. 94% of the top-5 error rate, becoming the first time in the data set over the human eye recognition effect (error rate of about 5. 1% [17]) model. Compared with the previous convolution neural network model, the model has two improvements, one is to promote the traditional modified linear unit (ReLU), proposed parametric correction linear unit (PReLU). The activation function can adaptively learn the parameters of the correction unit and can improve the accuracy of the recognition if the additional calculation cost is negligible. At the same time, this model deduces a set of robust initialization method by modeling the modified linear unit (ReLU / PReLU), which can make the model with more layers (such as 30 models with weighted layer) convergence. Shortly thereafter, Google normalized each mini-batch while training the network, calling it Batch normalization, applying the training method to GoogleNet, and 4 on the ILSVRC2012 data set. 82% of the top-5 error rate [24]. Normalization is a commonly used input

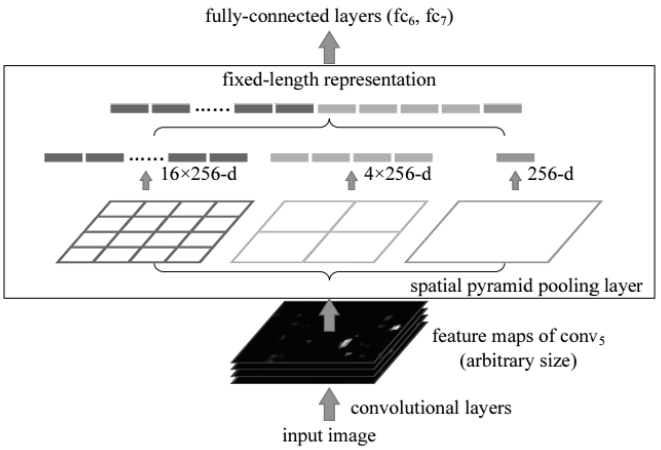

Fig. 6. Space pyramid pool model structure [22]

data preprocessing method for training deep neural networks, which can reduce the influence of the initial weight of the training parameters on the training effect and accelerate the convergence. So Google's researchers apply the normalized method to the activation function within the network, normalizing the transmission data between layers. Since the training uses a random gradient descent method, such normalization can only be done in each mini-batch, so it is named Batch normalization. The method can make the training can use a higher learning rate, reduce training time; at the same time reduce the over-fitting, improve the accuracy rate. Although the convolution neural network already has a strong image learning ability, this type of model lacks learning for invariance of image space, especially the lack of learning for image rotation invariance [19]. The globalization transformer [25] proposed by Google DeepMind aims to enhance the accuracy of its image classification by increasing the learning ability of the convolution neural network for image spatial invariance. The Spatial transformer is a module that can be added at any depth of the convolutional neural network. It can perform a series of spatial transformations on the input data, making the output feature.

During the training process, the module can autonomously learn the parameters required for spatial transformation and does not need to add any additional supervisory processing to the training. In the results of the ILSVRC2015 from the ImageNet Computer Vision Recognition Challenge at the end of 2015, the depth of the 152-layer deep residual network from the Microsoft Asia Research Institute team received image detection, image classification and image positioning at an absolute advantage. Of the championship, which in the image classification of the data set made 3. 57% error rate [26]. With the deepening of the number of convolution neural networks, the training process of the network is more difficult, resulting in the accuracy rate began to reach saturation or even decline. The team's researchers believe that when a network reaches the optimal training effect, it may require some layers of output and input exactly the same; then let the network layer learning value of 0 residual function than learning function is easier.

Therefore, the deep residual network will be used in the residual representation of the network, put forward the idea of residual learning. As shown in Fig. 7, in order to achieve the residual learning, the Shortcut connection method is applied to

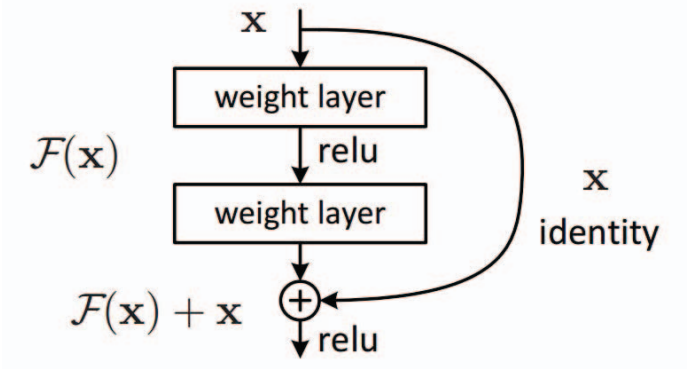

Fig. 7. Residual Learning Module [26]

the connection between the layers in the network, so that the accuracy rate can be improved without decreasing with the increase of the network layer The Because ImageNet has the characteristics of large scale and large image class, the model trained by ImageNet has a strong ability of generalization, and it can get good classification result on other data sets. If the image is further fine-tuned on the target data set , And most of the training with only the target data set to get better results. The first to use the convolution neural network for object detection RCNN model [27], is to use ImageNet trained AlexNet model in the PASCAL VOC data set for fine-tuning after the extraction of image features, made than the previous model 20% higher accuracy. In addition, the models trained with the ImageNet dataset are applied to other types of data sets such as remote sensing image classification [28], indoor scene classification [29], and have achieved better results than previous methods. Since the depth of learning for the first time in ILSVRC2012 was applied to the image classification game and achieved remarkable results, based on the depth of learning method model began to be widely used in the field of image recognition, the emergence of new depth of the neural network model is constantly refreshing the game record , But also makes the depth of the neural network model for the image features of the learning ability to upgrade. At the same time, due to the emergence of large-scale data sets such as ImageNet and MSCOCO, the depth network model can be well trained, and the model trained by a large number of data has stronger generalization ability and can better adapt to the practical application Need to learn the data set to enhance the classification effect.

IV. Conclusions

Deep learning is currently a very popular research direction, the use of convolution neural network convolution layer, pool layer and the whole connection layer and other basic structure, you can let the network structure to learn and extract the relevant features, and to be used. This feature provides many conveniences for many studies, eliminating the need for a very complex modeling process. In addition, deep learning is now in the image classification, object detection, attitude estimation and image segmentation and so on have been very big results and progress. On the one hand, the depth of learning application is very wide, and versatility, can continue to work to expand it to other applications. On the other hand, there are still many potentials to learn, and it is worth exploring and discovering. In the future, despite the fact that many of the previous discussions are supervised (for example, the last layer of the trained network will calculate a loss value based on the real value and then adjust the parameters), and the supervised study does achieve a very large The success of. The application of deep learning in unsupervised learning is likely to be a future trend. After all, in the case of humans or animals, in most cases, we do not know what it is by knowing the name of the thing. In the future field of computer vision, it is expected that the recurrent neural network (RNN) based on deep learning will become a very popular network model and will achieve a better breakthrough in more applied research with progress. In addition, the combination of strong chemical methods to train an end-to-end learning system is gradually possible, so that the

learning system with independent learning ability, can take the initiative to learn the relevant features of the representation and abstraction. At present, research combined with deep learning and intensive learning is still in its infancy, but some research in this area has achieved good performance in multi-object recognition tasks and video game learning. So that many of the relevant areas of researchers are excited about one of the reasons. It is noteworthy that natural language processing is also the potential to learn the future stage of the potential to show their skills, for example, for an article or a large text, can be designed based on some depth of the neural network model (such as RNN) method and Strategy, can effectively understand the text content. In general, people now use the depth of learning and some simple reasoning, it has been in the field of voice and image has achieved very good results. There is reason to believe that if the current feature of the network extraction can be further optimized so that it can more "freely" to express the characteristics, coupled with some complex reasoning, then the depth of learning will be in the application of artificial intelligence to achieve greater Progress.

ACKNOWLEDGMENT

This work was supported in part by RDU1603102.

REFERENCES

[1] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” International Journal of Computer Vision, 2004, vol. 60, no.2, pp: 91-110, Nov. 2004.

[2] N. Dalal, B. Triggs. Histograms of oriented gradients for human detection. Computer Vision and Pattern Recognition (CVPR), IEEE Computer Society Conference on. San Diego, USA: IEEE, 2005, 1: 886-893.

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

[4] T. Maimaitiaili, L. Dai. Deep neural network based uyghur large vocabulary continuous speech recognition Journal of Data Acquisition and Processing, 2015, 30 (2): 365-371.

[5] D. H. Hubel, T. N. Wiesel. Receptive fields, binocated interaction of functional architecture in the cat's visual cortex. The Journal of Physiology, 1962, 160 (1): 106-154.

[6] K. Fukushima, S. Miyake. Neocognitron: A new algorithm for pattern recognition tolerant of deformations and shifts in position. Pattern Recognition, 1982, 15 (6): 455-469.

[7] D. W. Ruck, S. K. Rogers, M. Kabrisky. Feature selection using a multilayer perceptron Journal of Neural Network Computing, 1990, 2 (2): 40-48.

[8] D. E. Rumelhart, G. E. Hinton, R. J. Williams. Learning representations by back-propagating errors. Nature, 1986, 323: 533-538.

[9] Y. LeCun, et al. Handwork digit recognition with a back-propagation network. Advances in Neural Information Processing Systems. Colorado, USA: [s. N], 1990: 396-404.

[10] Y. LeCun, C. Cortes. MNIST handwritten digit database [EB / OL]. Http: // yann. Lecun. Com / exdb / mnist, 2010.

[11] A. Waibel, et al. Phoneme recognition using time-delay neural networks. Acoustics, Speech and Signal Processing, IEEE Transactions on, 1989, 37 (3): 328-339.

[12] A. Krizhevsky. Learning multiple layers of features from tiny images. Toronto, Canada: University of Toronto, 2009.

[13] L. Fei-Fei, R. Fergus, P. Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. Computer Vision and Image Understanding, 2007, 106 (1): 59-70.

[14] G. Griffin, A. Holub, P. Perona. Caltech-256 object category dataset. Technical Report 7694, http: // authors. Library. Caltech Edu / 7694, California Institute of Technology, 2007.

[15] J. Deng, et al. Imagenet: A large-scale hierarchical image database. Computer Vision and Pattern Recognition (CVPR), IEEE Conference on. Miami, USA: IEEE, 2009: 248-255.

[16] T. Ahonen, A. Hadid, M. Pietikainen. Face description with local binary patterns: Application to face recognition. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2006, 28 (12): 2037-2041.

[17] O. Russakovsky, et al. Imagenet large scale visual recognition challenge International Journal of Computer Vision, 2015, 115 (3): 211-252.

[18] A. Krizhevsky, J. Sutskever, G. E. Hinton. Imagenet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems. Cambridge: MIT Press, 2012: 1097-1105.

[19] M. D. Zeiler, R. Fergus. Visualizing and understanding convolutional networks. New York: Springer International Publishing, 2014: 818-833.

[20] C. Szegedy, et al. Going deeper with convolutions. Computer Vision and Pattern Recognition (CVPR), IEEE Conference on. Boston, USA: IEEE, 2015: 1-9.

[21] M. Lin, Q. Chen, S. Yan. Network in network [EB / OL]. Http: // arxiv. Org / abs / 1312. 4400,2013.

[22] K. He, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition. Computer VisionECCV 2014. New York: Springer International Publishing, 2014: 346-361.

[23] K. He, et al. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification [EB / OL]. Http: // arxiv. Org / abs / 1502. 01852,2015.

[24] S. Ioffe, C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift [EB / OL]. Http: // arxiv. Org / abs / 1502. 03167,2015.

[25] M. Jaderberg, K. Simonyan, A. Zisserman. Spatial transformer networks. Advances in Neural Information Processing Systems. Montréal, Canada: [s. N] 2015: 2008-2016.

[26] K. He, et al. Deep residual learning for image recognition [EB / OL]. Http: // arxiv. Org / abs / 1512. 03385,2015.

[27] R. Girshick, et al. Rich feature hierarchies for accurate object detection and semantic segmentation. Computer Vision and Pattern Recognition (CVPR), IEEE Conference on. Columbus, USA: IEEE, 2014: 580-587.

[28] M. Castelluccio, et al. Land use classification in remote sensing images by convolutional neural networks [EB / OL]. Http: // arxiv. Org / abs / 1508. 00092,2015.

[29] M. Hayat, et al. A spatial layout and scale invariant feature representation for indoor scene classification [EB / OL]. Http: // arxiv. Org / abs / 1506. 05532,2015.