Review (Narrative)

Image Classification and Object Detection Algorithm Based on Convolutional Neural Network

Juan K. Leonard, Ph.D.

SUMMARY

Traditional image classification methods are difficult to process huge image data and cannot meet people's requirements for image classification accuracy and speed. Convolutional neural networks have achieved a series of breakthrough research results in image classification, object detection, and image semantic segmentation. This method broke through the bottleneck of traditional image classification methods and became the mainstream algorithm for image classification. Its powerful feature learning and classification capabilities have attracted widespread attention. How to effectively use convolutional neural networks to classify images have become research hotspots. In this paper, after a systematic study of convolutional neural networks and an in-depth study of the application of convolutional neural networks in image processing, the mainstream structural models, advantages and disadvantages, time / space used in image classification based on convolutional neural networks are given. Complexity, problems that may be encountered during model training, and corresponding solutions. At the same time, the generative adversarial network and capsule network based on the deep learning-based image classification extension model are also introduced; simulation experiments verify the image classification In terms of accuracy, the image classification method based on convolutional neural networks is superior to traditional image classification methods. At the same time, the performance differences between the currently popular convolutional neural network models are comprehensively compared and the advantages and disadvantages of various models are further verified. Experiments and analysis of overfitting problem, data set construction method, generative adversarial network and capsule network performance.■

KEYWORDS

Convolutional Neural Network; Deep Learning; Feature Expression; Transfer Learning

Sci Insigt. 2019; 31(1):85-100. doi:10.15354/si.19.re117.

Author Affiliations: Author affiliations are listed at the end of this article.

Correspondence to: Dr. Juan K. Leonard, Ph.D., Group of Network Computation, Division of Mathematics and Computation, The BASE, Chapel Hill, NC 27510, USA.
Email: juan.leonard@basehq.org

THE extraction and classification of image features has always been a basic and important research direction in the field of computer vision. Convolutional Neural Network (CNN) provides an end-to-end learning model. The parameters in the model can be trained by traditional gradient descent methods. The trained convolutional neural network can learn the features in the image. And complete the extraction and classification of image features. As an important research branch in the field of neural networks, the characteristics of convolutional neural networks are that the features of each layer are excited by the local area of the previous layer through a convolution kernel that shares weights. This feature makes convolutional neural networks more suitable for image feature learning and expression than other neural network methods.

The structure of early convolutional neural networks was relatively simple, such as the classic LeNet-5 model (1), which was mainly used in some relatively single computer vision applications such as handwritten character recognition and image classification. With the continuous deepening of research, the structure of convolutional neural network is continuously optimized, and its application field is gradually extended. For example, the Convolutional Deep Belief Network (CDBN) (3), which is a combination of Convolutional Neural Network and Deep Belief Network (DBN) (2), is an unsupervised generative model. Successfully applied to face feature extraction (4); AlexNet (5) achieved breakthrough results in the field of mass image classification; R-CNN (Regions with CNN) (6) based on region feature extraction achieved in the field of object detection Success; Fully Convolutional Network (FCN) (7) realized end-to-end image semantic segmentation, and greatly surpassed traditional semantic segmentation algorithms in accuracy. In recent years, the research on the structure of convolutional neural networks is still very popular, and some network structures with excellent performance have been proposed (8-10). Moreover, with the successful application of transfer learning theory (11) on convolutional neural networks, the application field of convolutional neural networks has been further expanded (12-13). The continually emerging research results of convolutional neural networks in various fields make it one of the most popular research hotspots.

RESEARCH HISTORY OF CONVOLUTIONAL NEURAL NETWORKS

The research history of convolutional neural networks can be roughly divided into three stages: the theory proposal stage, the model implementation stage, and the extensive research stage.

Theory Proposal Stage

In the 1960s, a biological study by Hubel et al. (14) showed that the transmission of visual information from the retina to the brain was stimulated through multiple levels of receptive fields. In 1980, Fukushima first proposed a theoretical model based on receptive fields, Neocognitron (15). Neocognitron is a self-organizing multi-layer neural network model. The response of each layer is inspired by the local receptive field of the previous layer. The recognition of the pattern is not affected by position, small shape changes, and scale. The unsupervised learning adopted by Neocognitron is also the dominant learning method in the early research of convolutional neural networks.

Model Implementation Phase

In 1998, LeNet-5 proposed by Lecun et al. (1) used a gradient-based back-propagation algorithm to supervise the network. The trained network transforms the original image into a series of feature maps through alternately connected convolutional layers and downsampling layers. Finally, the fully-connected neural network is used to classify the feature expression of the image. The convolution kernel of the convolutional layer completes the function of the receptive field. It can excite the local area information of the lower layer to a higher level through the convolution kernel. The successful application of LeNet-5 in the field of handwritten character recognition has aroused the attention of academic circles on convolutional neural networks. During the same period, research on convolutional neural networks in speech recognition (16), object detection (17), face recognition (18), etc. has also gradually developed.

Extensive Research Phase

In 2012, AlexNet proposed by Krizhevsky et al. (5) won the championship with a huge advantage of 11% over the second place in the image classification competition of the large image database ImageNet (19), making the

convolutional neural network an academic community focus. After AlexNet, new convolutional neural network models have been proposed, such as VGG (Visual Geometry Group) (8) of Oxford University, Google's GoogLeNet (9), Microsoft's ResNet (10), etc. These networks have refreshed AlexNet in A record set on ImageNet. In addition, the convolutional neural network is constantly fused with some traditional algorithms, and the introduction of transfer learning methods has made the application field of convolutional neural networks expand rapidly. Some typical applications include: Convolutional neural network combined with Recurrent Neural Network (RNN) for abstract generation of images (20-21) and question and answer of image content (22-23); convolution through transfer learning Neural networks have achieved significant accuracy improvements on small sample image recognition databases (24); and video-oriented behavior recognition models-3D convolutional neural networks (25), etc.

RESEARCH SIGNIFICANCE OF CONVOLUTIONAL NEURAL NETWORKS

Convolutional neural network has achieved many remarkable research results, but with it comes more challenges. Its research significance is mainly reflected in three aspects: theoretical research challenges, feature expression, and application value.

Theoretical Research Challenges

As a kind of empirical method inspired by biological research, convolutional neural network is widely used in academia. For example, the methods of GoogLeNet's Inception module design, VGG's deep network, and ResNet's short connection have proved their effectiveness for improving network performance through experiments; however, these methods lack the rigorous mathematical verification problems. The root cause of this problem is that the mathematical model of the convolutional neural network itself has not been mathematically verified and explained. From the perspective of academic research, the development of convolutional neural networks is not rigorous and unsustainable without the support of theoretical research. Therefore, the related theoretical research of convolutional neural networks is currently the most scarce and most valuable part.

Feature Expression

Image feature design has always been a basic and important subject in the field of computer vision. In previous studies, some typical artificial design features have been proven to achieve good feature expression effects, such as SIFT (Scale-Invariant Feature Transform) (26), HOG (Histogram of Oriented Gradient) (27), and so on. However, these artificial design features also suffer from a lack of good generalization performance. Convolutional neural networks, as a deep learning (28-29) model, have the ability of hierarchical learning features (24). Studies (30-31) have shown that features learned through convolutional neural networks have stronger discrimination and generalization capabilities than artificially designed features. Feature expression is the basis of computer vision research. How to use convolutional neural networks to learn, extract, and analyze the feature expression of information, so as to obtain universal features with stronger discriminative performance and better generalization performance, will have a broader impact on the entire computer vision and even more widely. The field has a positive impact.

Application Value

After years of development, convolutional neural networks have gradually expanded from the relatively simple applications of handwritten character recognition (1) to more complex fields, such as pedestrian detection (32), behavior recognition (25, 33), and human pose recognition. (34), etc. Recently, the application of convolutional neural networks has further developed to deeper levels of artificial intelligence, such as: natural language processing (35-36), speech recognition (37), and so on. Recently, Alphago (38), an artificial intelligence Go program developed by Google, successfully used convolutional neural network to analyze the information of the Go board, and successively defeated Go European and World Championships in the challenge, which attracted wide attention. From the current research trends, the application prospects of convolutional neural networks are full of possibilities, but at the same time, they are also facing some research difficulties, such as: how to improve the structure of convolutional neural networks to improve the network's ability to learn features; how to The convolutional neural net-

work is incorporated into the new application model in a reasonable form.

THE BASIC STRUCTURE OF A CONVOLUTIONAL NEURAL NETWORK

As shown in **Figure 1**, a typical convolutional neural network consists of an input layer, a convolutional layer, a downsampling layer (pooling layer), a fully connected layer, and an output layer.

The input of a convolutional neural network is usually the original image X. This paper uses Hi to represent the feature map of the i-th layer of the convolutional neural network (H0 = X). Assuming Hi is a convolutional layer, the generation process of Hi can be described as:

$$H_i = f(H_{i-1} \otimes W_i + b_i) \quad (1)$$

Where Wi represents the weight vector of the i-th layer convolution kernel; the operation symbol ⊗represents the convolution kernel and the i-1 layer image or feature map to be rolled

The downsampling layer usually follows the convolutional layer and downsampling the feature map according to a certain downsampling rule (39). The function of the downsampling layer mainly has two points: 1) dimensionality reduction of the feature map; 2) maintaining the scale-invariant characteristic of the feature to a certain extent. Assuming Hi is the downsampling layer:

$$H_i = \text{subsampling}(H_{i-1}) \quad (2)$$

After alternate transfers of multiple convolutional layers and down-sampling layers, the convolutional neural network relies on a fully connected network to classify the extracted features and obtain an input-based probability distribution Y (li represents the i-th label category). As shown in Equation (3), the convolutional neural network is essentially a mathematical model that transforms the original matrix (H0) through multiple levels of data transformation or dimensionality reduction to a new feature expression (Y).

$$Y(i) = P\big(L = l_i \mid H_0; (W, b)\big) \quad (3)$$

The training goal of a convolutional neural network is to minimize the loss function L (W, b) of the network. The input H0 calculates the difference from the expected value through the loss function after forward conduction, which is called "residual error". Common loss functions include Mean Squared Error (MSE) function, Negative Log Likelihood (NLL) function, etc. (40):

$$MSE(W, b) = \frac{1}{|Y|}\sum_{i=1}^{|Y|}\Big(Y(i) - \hat{Y}(i)\Big)^2 \quad (4)$$

$$NLL(W, b) = -\sum_{i=1}^{|Y|} \log Y(i) \quad (5)$$

In order to alleviate the problem of overfitting, the final loss function usually controls the overfitting of the weights by increasing the L2 norm, and controls the strength of the overfitting effect through the parameter λ (weight decay):

$$E(W, b) = L(W, b) + \frac{\lambda}{2} W^T W \quad (6)$$

During training, the commonly used optimization method for convolutional neural networks is the gradient descent method. The residuals are back-propagated through gradient descent, and the trainable parameters (W and b) of each layer of the convolutional neural network are updated layer by layer. The learning rate parameter (η) is used to control the strength of the residual backpropagation:

$$W_i = W_i - \eta \frac{\partial E(W,b)}{\partial W_i} \quad (7)$$

$$b_i = b_i - \eta \frac{\partial E(W,b)}{\partial b_i} \quad (8)$$

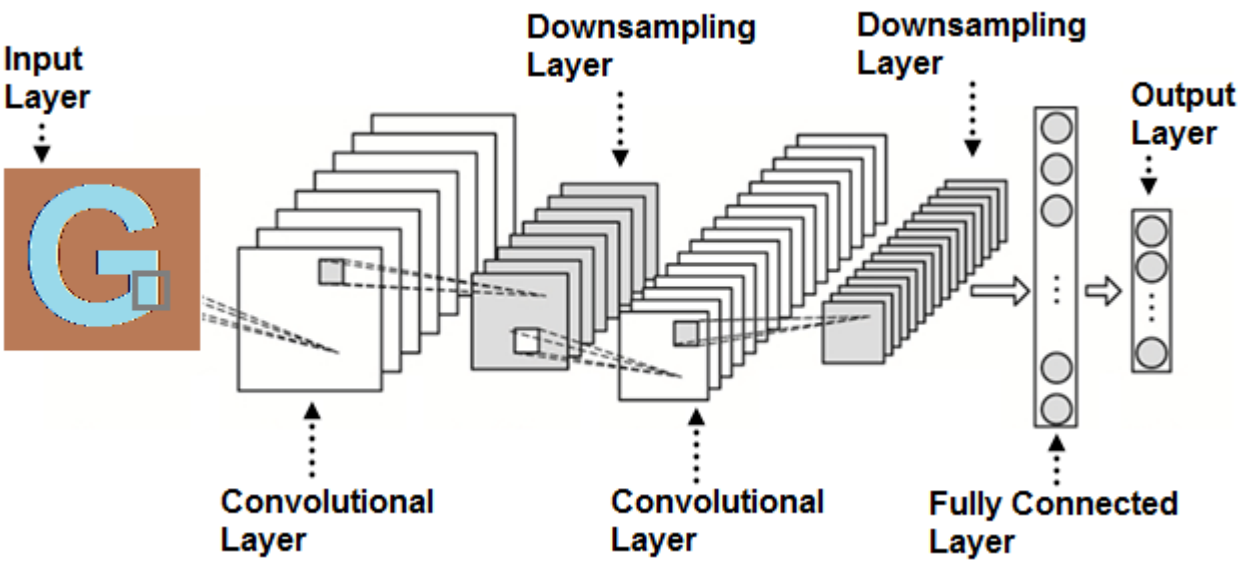
HOW CONVOLUTIONAL NEURAL NETWORKS WORK

Based on the definition, the working principle of convolutional neural networks can be divided into three parts: network model definition, network training, and network prediction:

Network Model Definition

The definition of the network model needs to design the network depth, the functions of each layer of the net-

Figure 1. Typical Structure of Convolutional Neural Network.



work, and set the hyperparameters in the network, such as: λ, η, etc., according to the amount of data of the specific application and the characteristics of the data itself. There are many studies on model design of convolutional neural networks, such as the depth of the model (8, 10), the step size of the convolution (24, 41), and the excitation function (42-43). In addition, for the selection of hyperparameters in the network, there are also some effective experience summaries (44). However, the theoretical analysis and quantitative research on network models are still relatively scarce.

Network Training

Convolutional neural networks can train the parameters in the network by backpropagating the residuals. However, problems such as overfitting in network training and the disappearance and explosion of gradients (45) greatly affect the convergence performance of training. For the problem of network training, some effective improvement methods have been proposed, including: random initialization of network parameters based on Gaussian distribution (5); initialization using pre-trained network parameters (8); parameters of different layers of convolutional neural networks Initialization of independent and identical distributions (46). According to recent research trends, the model size of convolutional neural networks is rapidly increasing, and more complex network models have also put forward higher requirements for corresponding training strategies.

Network Prediction

The prediction process of convolutional neural network is the process of outputting feature maps at various levels through forward transmission of input data, and finally using a fully connected network to output a conditional probability distribution based on the input data. Recent studies have shown that forward-conducted convolutional neural network high-level features have strong discriminative ability and generalization performance (30-31); moreover, these features can be applied to a wider range of fields through transfer learning. This research result is of great significance for expanding the application field of convolutional neural networks.

RESEARCH PROGRESS OF CONVOLUTIONAL NEURAL NETWORKS

After decades of development, from the initial theoretical prototype, to being able to complete some simple tasks, and to recently obtaining a large number of research results, it has become a research direction that has attracted wide attention. The main source of its driving force for development Fundamental research in

the following four areas: 1) related research on convolutional neural network overfitting problems has improved the generalization performance of the network; 2) related research on convolutional neural network structure has improved the network's ability to fit massive data; 3) The principle analysis of the convolutional neural network guides the development of the network structure. At the same time, it also proposes new and challenging problems. 4) Related research on convolutional neural networks based on transfer learning has expanded the application field of convolutional neural networks.

Overfitting of Convolutional Neural Networks

Over-fitting (40) refers to the phenomenon that the parameters of the learning model over-fit the training data set during the training process, which affects the generalization performance of the model on the test data set. The structural levels of convolutional neural networks are relatively complicated. The current research is directed at overfitting of convolutional layers, downsampling layers and fully connected layers of convolutional neural networks. The current main research idea is to improve the generalization performance of the network by increasing the sparseness and randomness of the network.

Dropout proposed by Hinton et al. (47) reduces the overfitting problem of traditional fully connected neural networks by ignoring a certain percentage of node responses randomly during the training process, and effectively improves the generalization performance of the network. However, Dropout does not improve the performance of convolutional neural networks. The main reason is that due to the weight-sharing feature of convolution kernels, compared to fully connected networks, the number of training parameters is greatly reduced. Avoid more serious overfitting. Therefore, the Dropout method acting on the fully connected layer is not ideal for the overall over-fitting effect of the convolutional neural network.

Based on the idea of Dropout, Wan et al. (48) proposed the method of DropConnect. Unlike Dropout, which ignores the response of some nodes of the fully connected layer, DropConnect randomly disconnects a certain percentage of the connections of the neural network convolutional layer. For convolutional neural networks, DropConnect acting on the convolutional layer has a stronger ability to overfit than Dropout acting on the fully connected layer.

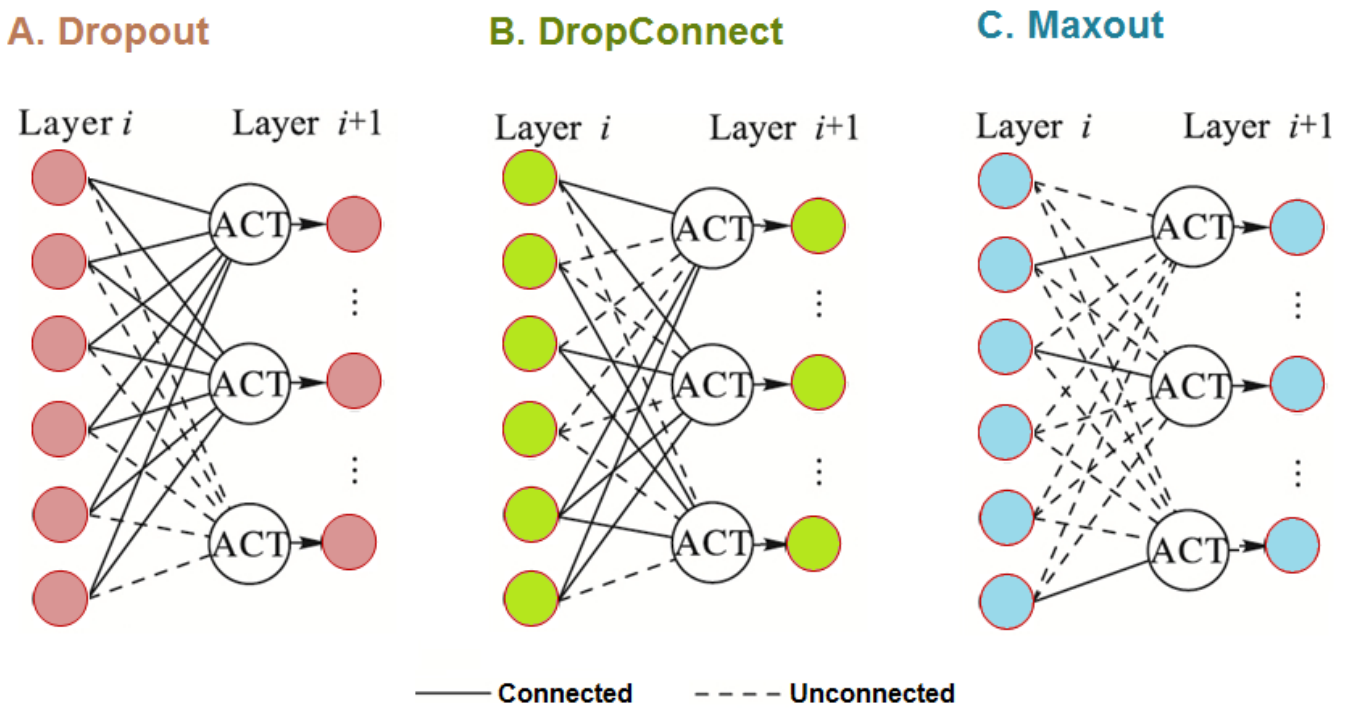
Similar to DropConnect, Goodfellow et al. (42) proposed a Maxout excitation function for convolutional layers. Unlike DropConnect, Maxout only keeps the maximum value of the next layer of nodes in the neural network. Moreover, Goodfellow et al. (42) proved that Maxout function can fit arbitrary convex functions, and has a strong function fitting ability on the basis of reducing the overfitting problem.

As shown in **Figure 2**, although the three implementation methods of Dropout, DropConnect, and Maxout are different, the specific implementation mechanisms are different. However, the basic principle is to increase the sparseness or randomness of network connections to eliminate overfitting, thereby significantly improving the ability of network generalization.

Lin et al. (43) pointed out that the fully connected network in convolutional neural networks is prone to overfitting and the limitation that the Maxout activation function can only fit convex functions, and proposed a network structure of NIN (Network in Network). On the one hand, NIN abandoned the use of fully-connected networks to map feature maps to probability distributions, and adopted the method of global average pooling for feature maps to obtain the final probability distribution. This reduced the number of parameters in the network while avoiding it. Overfitting of fully connected networks; on the other hand, NIN uses a "micro neural network" to replace traditional excitation functions (such as Maxout). In theory, the micro-neural network breaks through the limitations of the traditional excitation function, and can fit arbitrary functions, which makes the network have better fitting performance.

In addition, for the downsampling layer of the convolutional neural network, Zeiler et al. (39) proposed a random downsampling method (Stochastic pooling) to improve the overfitting problem of the downsampling layer. Unlike traditional Average pooling and Max pooling, which specify the average and maximum downsampling areas for downsampling respectively, stochastic pooling performs random downsampling operations based on probability distributions, which introduces randomness to the downsampling process. Experiments show that this randomness can effectively improve the generalization performance of convolutional neural networks.

Figure 2. Dropout, DropConnect, and Maxout Mechanism Network.



The current research on the problem of overfitting of convolutional neural networks mainly has the following problems: 1) The lack of quantitative research and evaluation standards for overfitting phenomenon, so that the current research can only prove the new method through experimental comparison. For the improvement of the overfitting problem, the degree and generality of this improvement need to be measured with more uniform and universal evaluation standards; 2) For the convolutional neural network, the overfitting problem is at various levels (such as: Convolutional layer, downsampling layer, fully connected layer), the improvement space and improvement methods need to be further explored.

THE STRUCTURE OF A CONVOLUTIONAL NEURAL NETWORK

The LeNet-5 model proposed by Lecun et al. (1) uses alternately connected convolutional layers and downsampling layers to forward conduct the input image, and finally the structure that outputs the probability distribution through the fully connected layer is the currently commonly used convolutional nerve prototype of network structure. Although LeNet-5 has achieved success in the field of handwritten character recognition, its shortcomings are also relatively obvious, including: 1) it is difficult to find a suitable large training set to train the network to meet more complex application requirements; 2) over The fitting problem makes LeNet-5's generalization ability weak; 3) The training overhead of the network is very large, and the lack of hardware performance support makes the research of network structure very difficult. The above three important factors restricting the development of convolutional neural networks have achieved breakthroughs in recent research. This is an important reason why convolutional neural networks have become a new research hotspot. In addition, recent research on the depth and structural optimization of convolutional neural networks has further improved the data fitting capabilities of the network.

In response to the defects of LeNet-5, Krizhevsky et al. (5) proposed AlexNet. AlexNet has a 5-layer convolutional network, about 650,000 neurons and 60 million trainable parameters, which greatly surpasses LeNet-5

in terms of network scale. In addition, AlexNet chose a large image classification database ImageNet (19) as the training data set. ImageNet provides 1.2 million pictures in 1,000 categories for training, and the number and categories of pictures greatly exceed previous data sets. In terms of overfitting, AlexNet introduced dropout, which alleviated the problem of network overfitting to a certain extent. In terms of hardware support, AlexNet uses GPUs for training. Compared to traditional CPU operations, GPUs make the training speed of the network more than ten times faster. AlexNet won the championship in ImageNet's 2012 image classification competition, and achieved a huge advantage of 11% in accuracy compared to the second-place method. The success of AlexNet caused the research on convolutional neural networks to once again attract the attention of the academic community.

Simonyan et al. (8) studied the depth of convolutional neural networks on the basis of AlexNet and proposed a VGG network. VGG is constructed by a 3 × 3 convolution kernel. By comparing the performance of networks of different depths in image classification applications, Simonyan et al. proved that the improvement of network depth helps improve the accuracy of image classification. However, this increase in depth is not unlimited, and continuing to increase the number of layers in the network based on the appropriate network depth will bring the problem of network degradation with increased training errors (49). Therefore, the optimal network depth of VGG is set at 16-19 layers.

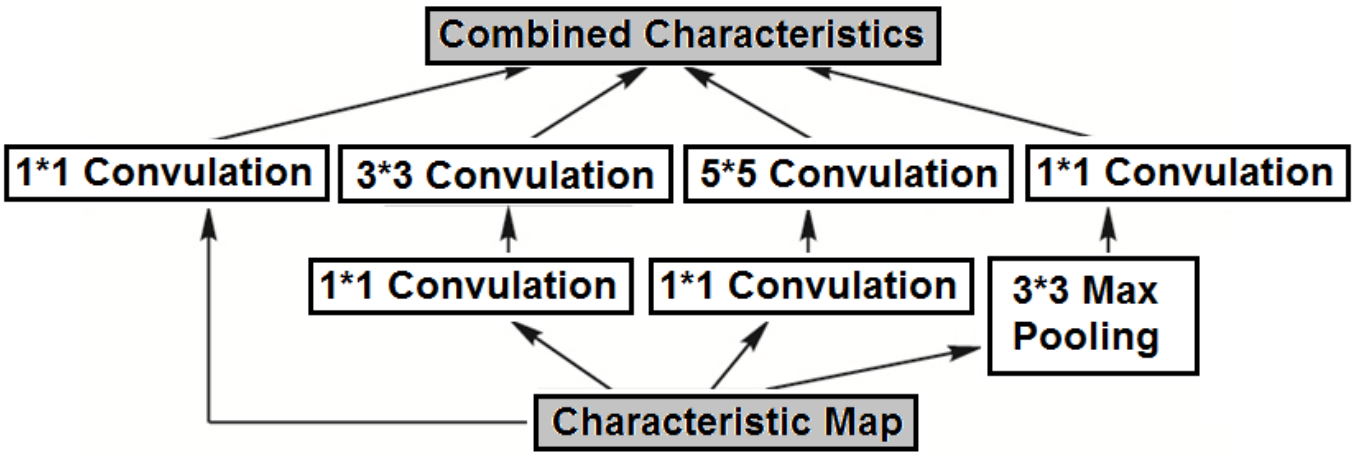
In view of the degradation of deep networks, He et al. (10) analyzed that if every level added to the network can be optimized for training, then the error should not increase with the increase of the network depth. Therefore, the problem of network degradation shows that not every level in a deep network is well trained. He et al. Proposed a ResNet network structure. ResNet directly maps the low-level feature map x to the higher-level network through short connections. Assume that the non-linear mapping of the original network is F (x), then the mapping relationship after connecting through a short connection becomes F (x) + x. He et al. Proposed this method on the basis that F (x) + x optimization is easier than F (x). Because, from an extreme perspective, if x is already an optimized mapping, then the network mapping between short connections will be closer to 0 after training. This means that the forward transmission of data can skip some layers without perfect training through a short connection to a certain extent, thereby improving the performance of the network. Experiments show that although ResNet uses a convolution kernel of the same size as VGG, the solution to the network degradation problem makes it possible to build a 152-layer network, and ResNet has lower training errors and higher test accuracy than VGG. . Although ResNet solves the problem of deep network degradation to some extent, there are still some doubts about the research on deep networks: 1) how to determine which levels in deep networks have not been fully trained; 2) what causes deep networks Imperfect training at some levels; 3) How to deal with imperfect levels of training in deep networks.

In addition to the research on the depth of convolutional neural networks, Szegedy et al. (9) paid more attention to reducing the complexity of the network by optimizing the network structure. They proposed a basic module of a convolutional neural network called Inception. As shown in **Figure 3**, the Inception module consists of 1 × 1, 3 × 3, 5 × 5 convolution kernels. The use of small-scale convolution kernels has two major advantages: 1) the number of training parameters in the entire network is controlled, reducing the complexity of the network; 2) Convolution kernels of different sizes perform feature extraction on the same image or feature map on multiple scales. Experiments show that the number of training parameters of GoogLeNet constructed using the Inception module is only 1/12 of that of AlexNet, but the accuracy of image classification on ImageNet is about 10% higher than that of AlexNet.

In addition, Springenberg et al. (50) questioned the necessity of the existence of the downsampling layer of the convolutional neural network, and designed a "completely convolutional network" without the downsampling layer. The "full convolutional network" is simpler in structure than the traditional convolutional neural network structure, but its network performance is not lower than the traditional model with a downsampling layer.

Research on the structure of convolutional neural networks is an open question. Based on the current research status, the current research has formed two major trends: 1) increasing the depth of convolutional neural networks; 2) optimizing the structure of convolutional neural networks, Reduce network complexity. In terms of in-depth research on convolutional neural networks, it mainly depends on further analysis of potential hidden dangers (such as network degradation) in deep-level networks to solve the training problems of

Figure 3. Mechanism of Inception.



deep networks (such as: VGG, ResNet). In terms of optimizing the network structure, the current research trend is to further strengthen the understanding and analysis of the current network structure, replacing the current structure with a more concise and efficient network structure, further reducing network complexity and improving network performance (such as: GoogLeNet Full convolutional network).

PRINCIPLES OF CONVOLUTIONAL NEURAL NETWORKS

Although convolutional neural networks have achieved success in many applications, the analysis and interpretation of their principles have always been a weak point that has been questioned. Some recent studies have begun to use visual methods to analyze the principles of convolutional neural networks, visually compare the differences between the learning characteristics of convolutional neural networks and traditional artificial design features, and show the network's characteristic expression process from low to high levels.

Donahue et al. (30) proposed using t-SNE (51) to analyze the features extracted by convolutional neural networks. The principle of t-SNE is to reduce high-dimensional features to two dimensions, and then visually display the features in two dimensions. Use t-SNE, Donahue, etc. to combine the features of convolutional neural networks with traditional artificial design features GIST (the meaning of GIST is an abstract scene that can stimulate scene categories in memory) (52) and LLC (Locality-constrained Linear Coding) (53) After comparison, it is found that the features of the convolutional neural network with stronger discriminative ability show better discrimination in the visualization results of t-SNE, which proves that the feature discriminative ability is consistent with the t-SNE visualization results. However, the research of Donahue et al. still left the following problems: 1) failed to explain what the features extracted by the convolutional neural network were; 2) Donahue et al. Selected some features of the convolutional neural network for visualization, but for these levels The relationship between them is not analyzed; 3) The t-SNE algorithm itself has certain limitations, and it does not reflect the differences between categories well when there are too many feature categories.

The study by Zeiler et al. (24) solved the legacy problems of t-SNE better. By constructing DeConvNet (54), they deconvolved the features at different levels in the convolutional neural network, and showed the feature status extracted at each level. The first and second layers of the convolutional neural network mainly extract low-level features such as edges and color, the third layer begins to appear more complex texture features, and the fourth and fifth layers begin to appear more complete individual contour and shape features. By visualizing the characteristics of each level, Zeiler et al. improved the size and step size of AlexNet's convolution kernel and improved network performance. More-

over, they also used the visual features to analyze the occlusion sensitivity, object component correlation, and feature invariance of convolutional neural networks. Zeiler et al.'s research shows that the research on the principles of convolutional neural networks is of great significance for improving the structure and performance of convolutional neural networks.

Nguyen et al. (55) questioned the completeness of the features extracted by the convolutional neural network. Nguyen et al. used an evolutionary algorithm (56) to process the original image into a form that cannot be recognized and interpreted by humans, but the convolutional neural network gave very exact object category judgment. Nguyen et al. did not make a clear explanation for the reason for this phenomenon, but just proved that although the convolutional neural network has the ability of layered feature extraction; it is not completely consistent with humans in the image recognition mechanism. This phenomenon indicates that there is still a great deal of shortcomings in the current research on the principles and analysis of convolutional neural networks.

In general, the research and analysis on the principles of convolutional neural networks are still insufficient. The main problems include: i) unlike traditional artificial design features, the characteristics of convolutional neural networks are subject to specific network structures and learning algorithms. And the influence of various factors such as the training set, the analysis and interpretation of its principles are more abstract and difficult than artificial design features; ii) the research by Nguyen et al. (55) showed that the phenomenon of "deception" caused by convolutional neural networks caused People are concerned about its completeness. Although convolutional neural networks are based on bionics research, how to explain the differences between convolutional neural networks and human vision and how to make the recognition mechanism of convolutional neural networks more complete are still problems to be solved.

TRANSFER LEARNING FOR CONVOLUTIONAL NEURAL NETWORKS

The definition of transfer learning is: "a machine learning method that uses existing knowledge to solve problems in different but related domains" (57), and its goal is to complete the transfer of knowledge between related domains (11). For convolutional neural networks, transfer learning is to successfully apply the "knowledge" trained on a specific data set to a new field. As shown in **Figure 4**, the general process of transfer learning for convolutional neural networks is: i) before specific applications, use large data sets in related fields (such as ImageNet) to train random initialization parameters in the network; ii) use training A good convolutional neural network performs feature extraction for data in a specific application field (such as Caltech); iii) uses the extracted features to train a convolutional neural network or classifier for data in a specific application field.

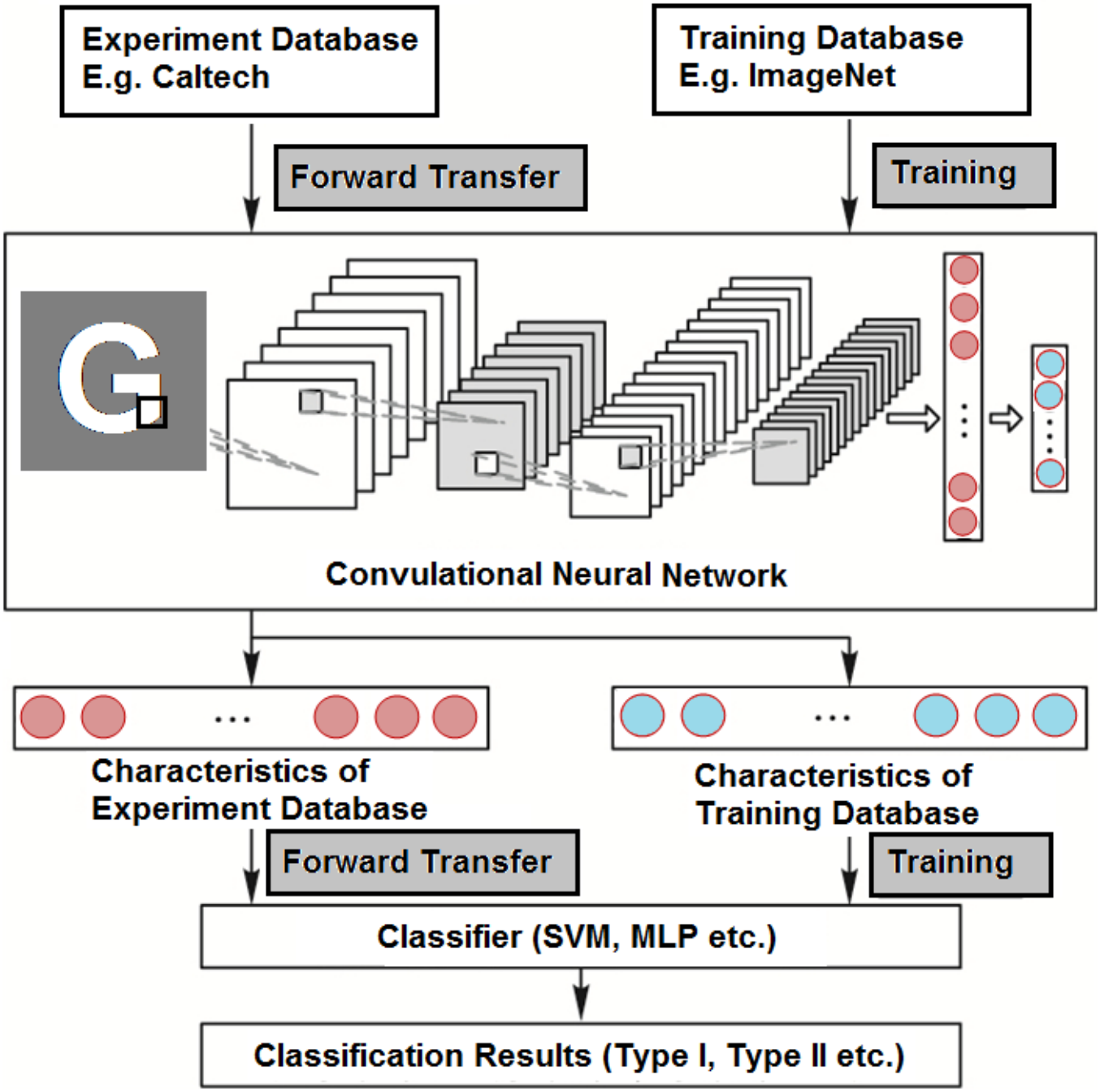
Compared with the traditional method of training the network directly on the target data set, Zeiler et al. (24) let the convolutional neural network be pre-trained on the ImageNet data set, and then the network was separately applied to the image classification data set Caltech-101 (58) and Caltech-256 (59) performed migration training and testing, and its image classification accuracy improved by about 40%. However, both ImageNet and Caltech belong to object recognition databases, and their fields of transfer learning are relatively close, and there are still insufficient researches on larger span fields. Therefore, Donahue et al. (30) adopted a similar strategy to Zeiler and successfully applied convolutional neural network transfer learning to areas that are more different from object recognition through ImageNet-based convolutional neural network pre-training, including: domain adaption, subcategory recognition, and scene recognition.

In addition to research on transfer learning of convolutional neural networks in various fields, Razavian et al. (31) also explored the transfer learning effects of different levels of features of convolutional neural networks, and found that higher-level features of convolutional neural networks have better performance than lower-level features. Transfer learning capabilities.

Zhou et al. (60) used a large image classification database (ImageNet) and scene recognition database to pre-train two convolutional neural networks of the same structure, and performed a series of image classification and scenes. The verification effect of transfer learning was verified on the recognition database. The experimental results show that ImageNet and Places pre-trained networks achieve better transfer learning results on the databases of their respective domains. This fact indicates that the correlation of the domain has a certain impact on the transfer learning of convolutional neural networks.

The research on transfer learning of convolutional neural networks includes: i) the problem of insufficient training samples for convolutional neural networks under small sample conditions; ii) the transfer utilization of convolutional neural networks can greatly reduce the training of the network overhead; iii) the use of transfer learning can further expand the application area of convolutional neural networks.

Figure 4. Flow of the Transfer Learning in Convolutional Neural Network.



Contents for further study of convolutional neural network transfer learning include: i) the effect of the number of training samples on the effect of transfer learning, and the effect of transfer learning on applications with different numbers of training samples needs further research; ii) based on volume The structure of

the product neural network itself further analyzes the transfer learning capabilities of each level in the convolutional neural network system; iii) analyzes the role of inter-domain correlations on transfer learning, and finds optimal cross-domain transfer learning strategies.

APPLICATION OF CONVOLUTIONAL NEURAL NETWORK

With the improvement of network performance and the use of transfer learning methods, the related applications of convolutional neural networks are gradually becoming more complex and diversified. In general, the application of convolutional neural networks mainly shows the following four development trends:

(i) With the continuous advancement of related research on convolutional neural networks, the accuracy of its related application areas has also been rapidly improved. Taking research in the field of image classification as an example, after AlexNet greatly improved the accuracy of ImagNet's image classification to 84.7%; continuously improved convolutional neural network models were proposed and refreshed the records of AlexNet. Representative networks include: VGG (8), GoogLeNet (9), PReLU-net (46) and BN-inception (61), etc. Recently, ResNet (10) proposed by Microsoft has improved the image classification accuracy of ImageNet to 96.4%, and ResNet has only been proposed by AlexNet within four years. The rapid development of convolutional neural networks in the field of image classification, continuously improving the accuracy of existing data sets, has also brought urgent needs to the design of larger databases related to image applications.

(ii) Development of real-time applications. Computational overhead has been an obstacle to the development of convolutional neural networks in real-time applications. However, some recent research shows the potential of convolutional neural networks in real-time applications. Gishick et al. (6, 62) and Ren et al. (63) have conducted in-depth research in the field of object detection based on convolutional neural networks, and have proposed R-CNN (6), Fast R-CNN (62), and Faster R- CNN (63) model breaks through the real-time application bottleneck of convolutional neural networks. R-CNN successfully proposed using CNN for object detection on the basis of region proposals (64). Although R-CNN has achieved high object detection accuracy, too many region proposals make object detection very slow. Fast R-CNN greatly reduces the computational overhead caused by a large number of region proposals by sharing convolutional features among region proposals. Fast R-CNN achieves near real-time object detection speed while ignoring the time required generating region proposals. Faster R-CNN uses the end-to-end convolutional neural network (7) to extract the region proposals instead of the traditional low-efficiency method (64), and realizes the real-time detection of objects by the convolutional neural network. With the continuous improvement of hardware performance and the reduction of network complexity caused by improving the network structure, convolutional neural networks have gradually shown their application prospects in the field of real-time image processing tasks.

(iii) As the performance of convolutional neural networks improves, the complexity of related applications also increases. Some representative studies include: Khan et al. (65) completed the shadow detection task by using two convolutional neural networks to learn the regional and contour features in the image respectively; the application of convolutional neural networks in face detection and recognition Great progress has also been made in China, achieving close to human face recognition (66-67); Levi et al. (68) use the subtle features of the face learned by the convolutional neural network to further achieve human gender and Prediction by age; FCN structure proposed by Long et al. (7) realized end-to-end mapping of images and semantics; Zhou et al. (60) studied the use of convolutional neural networks for image recognition and more complex scene recognition tasks Interconnected; Ji et al. (25) used 3D convolutional neural networks to implement behavior recognition. At present, the performance and structure of convolutional neural networks are still at a high-speed development stage, and their related complex applications will maintain their research interest for the next period of time.

(iv) Based on transfer learning and network structure improvements, convolutional neural networks have gradually become a general-purpose feature extraction and pattern recognition tool, and its application has gradually exceeded the traditional computer vision field. For example, AlphaGo successfully used a convolutional neural network to judge the board situation of Go (38), which proved the successful application of convolutional neural networks in the field of artificial intelligence; Abdel-Hamid et al. (37) modeled the voice information into The input model conforming to the convolutional neural network, combined with Hidden Markov Model

(HMM), successfully applied the convolutional neural network to the field of speech recognition; Kalchbrenner et al. (35) used the convolutional neural network to extract vocabulary And sentence-level information, successfully applied the convolutional neural network to natural language processing; Donahue et al. (20) combined the convolutional neural network and recursive neural network, and proposed the LRCN (Long-term Recurrent Convolutional Network) model to achieve Automatic generation of image summaries. As a general feature expression tool, convolutional neural network has gradually shown its research value in a wider range of applications.

Judging from the current research situation, on the one hand, the research interest of convolutional neural networks in its traditional application fields has not diminished, and there is still a lot of research space on how to improve the performance of networks; on the other hand, convolutional neural networks have good generality Performance has gradually expanded its application field. The scope of application is no longer limited to the traditional computer vision field, and it has developed toward application complexity, intelligence and real-time.

DEFECTS AND DEVELOPMENT DIRECTIONS OF CONVOLUTIONAL NEURAL NETWORKS

At present, convolutional neural networks are in a very hot research stage. Some problems and development directions in this field still include:

(i) Complete mathematical explanation and theoretical guidance are issues that cannot be avoided in the further development of convolutional neural networks. As an empirical research field, the theoretical research of convolutional neural networks is still relatively lagging. The related theoretical research of convolutional neural networks is of great significance for the further development of convolutional neural networks.

(ii) There is still a lot of space for research on the structure of convolutional neural networks. Current research shows that by simply increasing the complexity of the network, a series of bottlenecks will be encountered, such as overfitting problems and network degradation problems. The improvement of convolutional neural network performance depends on a more reasonable network structure design.

(iii) Convolutional neural networks have many parameters, but most of the current settings are based on experience and practice. Quantitative analysis and research of parameters is a problem to be solved for convolutional neural networks.

(iv) The model structure of convolutional neural networks is constantly improved, and the old data sets can no longer meet the current needs. Data sets are of great significance for the structural research and transfer learning research of convolutional neural networks. More numbers and categories and more complex data forms are the current development trends of related research data sets.

(v) The application of transfer learning theory helps to further expand the development of convolutional neural networks to a wider application field; and the design of task-based end-to-end convolutional neural networks (such as Faster R-CNN, FCN, etc.) Helps to improve the real-time nature of the network and is one of the current development trends.

(vi) Although the convolutional neural network has achieved excellent results in many application fields, related research and certification on its completeness is still a relatively scarce part at present. The comprehensive study of convolutional neural networks is helpful to further understand the principle differences between convolutional neural networks and human visual systems, and to help discover and resolve cognitive defects in the current network structure.

CONCLUSION

This article briefly introduces the history and principles of convolutional neural networks, focusing on the current development of convolutional neural networks from four aspects: overfitting problems, structural research, principle analysis, and transfer learning. In addition, this paper also analyzes some of the current application results of convolutional neural networks, and points out some defects and development directions of current research on convolutional neural networks. Convolutional neural network is a research field with high popularity at present, and has broad research prospects.■

ARTICLE INFORMATION

Author Affiliations: Group of Network Computation (Dr. Juan K. Leonard), Division of Mathematics and Computation, The BASE, Chapel Hill, NC 27510, USA.

Author Contributions: Dr. Leonard has full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.
Study concept and design: Leonard.
Acquisition, analysis, or interpretation of data: Leonard.
Drafting of the manuscript: Leonard.
Critical revision of the manuscript for important intellectual content: Leonard.
Statistical analysis: N/A.
Obtained funding: N/A.
Administrative, technical, or material support: Leonard.
Study supervision: Leonard.

Conflict of Interest Disclosures: Leonard declared no competing interests of this manuscript submitted for publication.

Funding/Support: N/A.

Role of the Funder/Sponsor: N/A.

How to Cite This Paper: Leonard JK. Image classification and object detection algorithm based on convolutional neural network. Sci Insigt. 2019; 31(1):85-100.

Digital Object Identifier (DOI): http://dx.doi.org/10.15354/si.19.re117.

Article Submission Information: Received, August 19, 2019; Revised: September 26, 2019; accepted: October 19, 2019.

REFERENCES

1. Lecun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition. Proceed IEEE 1998; 86(11):2278-2324.
2. Hinton GE, Osindero S, Teh YW. A fast learning algorithm for deep belief nets. Neur Comput 2006; 18(7):1527-1554.
3. Lee H, Grosse R, Ranganath R, et al. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations // ICML '09: Proceedings of the 26th Annual International Conference on Machine Learning. New York: ACM, 2009:609-616.
4. Huang G B, Lee H, Erik G. Learning hierarchical representations for face verification with convolutional deep belief networks // CVPR '12: Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition. Washington, DC: IEEE Computer Society, 2012:2518-2525.
5. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks // Proceedings of Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press, 2012:1106-1114.
6. Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation // Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. Washington, DC: IEEE Computer Society, 2014:580-587.
7. Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation // Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Washington, DC: IEEE Computer Society, 2015:3431-3440.
8. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. (2015-11-04)
9. Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions // Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Washington, DC: IEEE Computer Society, 2015:1-8.
10. He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. (2016-01-04).
11. Pan S J, Yang Q. A survey on transfer learning. IEEE Transact Knowled Data Engineer 2010; 22(10):1345-1359.
12. Collobert R, Weston J, Bottou L, et al. Natural language processing (almost) from scratch. J Machin Learn Res 2011; 12(1):2493-2537.
13. Oquab M, Bottou L, Laptev I, et al. Learning and transferring mid-level image representations using convolutional neural networks // Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. Washington, DC: IEEE Computer Society, 2014:1717-1724.
14. Hubel DH, Wiesel TN. Receptive fields, binocular interaction, and functional architecture in the cat's visual cortex. J Physiol 1962; 160(1):106-154.
15. Fukushima K. Neocognitron: a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biol Cybernet, 1980; 36(4):193-202.
16. Waibel A, Hanazawa T, Hinton G, et al. Phoneme recognition using time-delay neural networks (M)// Readings in Speech Recognition. Amsterdam: Elsvier, 1990:393-404.
17. Vaillant R, Monrocq C, Le Cun Y. Original approach for the localization of objects in images. IEE Proceed Vis Imag Sig Process 1994; 141(4):245-250.
18. Lawrence S, Giles CL, Tsoi AC, et al. Face recognition: a convolutional neural-network approach. IEEE Transact Neur Network 1997; 8(1):98-113.
19. Deng J, Dong W, Socher R, et al. ImageNet: a large-scale hierarchical image database // Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition. Washington, DC: IEEE Computer Society, 2009:248-255.
20. Donahue J, Hendricks LA, Guadarrama S, et al. Long-term recurrent convolutional networks for visual recognition and description // Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Washington, DC: IEEE Computer Society, 2015:2625-2634.
21. Vinyals O, Toshev A, Bengio S, et al. Show and tell: a neural image caption generator // Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Washington, DC: IEEE Computer Society, 2015:3156-3164.
22. Malinowski M, Rohrbach M, Fritz M. Ask your neurons: a neural-based approach to answering questions about images // Proceedings of the 2015 IEEE International Conference on Computer Vision. Piscataway, NJ: IEEE, 2015:1-9.
23. Antol S, Agrawal A, Lu J, et al. VQA: visual question answering // Proceed-

ings of the 2015 IEEE International Conference on Computer Vision. Piscataway, NJ: IEEE, 2015:2425-2433.

24. Zeiler MD, Fergus R. Visualizing and understanding convolutional networks // Proceedings of European Conference on Computer Vision, LNCS 8689. Berlin: Springer, 2014:818-833.
25. Ji S, Xu W, Yang M, et al. 3D convolutional neural networks for human action recognition. IEEE Transact Pattern Anal Mach Intel 2013; 35(1):221-231.
26. Lowe DG. Distinctive image features from scale-invariant keypoints. International J Comput Vis 2004; 60(2):91-110.
27. Dalal N, Triggs B. Histograms of oriented gradients for human detection // Proceedings of the 2005 IEEE Conference on Computer Vision and Pattern Recognition. Washington, DC: IEEE Computer Society, 2005:886-893.
28. Lecun Y, Bengio Y, Hinton GE. Deep learning. Nature 2015; 521(7553):436-444.
29. Sun ZJ, Xue L, Xu YM, et al. Overview of deep learning. Appl Res Comput 2012; 29(8):2806-2810.
30. Donahue J, Jia Y, Vinyals O, et al. DeCAF: a deep convolutional activation feature for generic visual recognition. Comput Sci 2013; 50(1):815-830.
31. Razavian AS, Azizpour H, Sullivan J, et al. CNN features off-the-shelf: an astounding baseline for recognition. (2015-11-22).
32. Sermanet P, Kavukcuoglu K, Chintala S, et al. Pedestrian detection with unsupervised multi-stage feature learning // CVPR '13: Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition. Washington, DC: IEEE Computer Society, 2013:3626-3633.
33. Karpathy A, Toderici G, Shetty S, et al. Large-scale video classification with convolutional neural networks // CVPR '14: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. Washington, DC: IEEE Computer Society, 2014:1725-1732.
34. Toshev A, Szegedy C. DeepPose: human pose estimation via deep neural networks // Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. Washington, DC: IEEE Computer Society, 2014:1653-1660.
35. Kalchbrenner N, Grefenstette E, Blunsom P. A convolutional neural network for modelling sentences. (2016-01-07).
36. Kim Y. Convolutional neural networks for sentence classification. (2016-01-07).
37. Abdel-Hamid O, Mohammed A, Jiang H, et al. Convolutional neural networks for speech recognition. IEEE/ACM Transact Aud Speech Lang Process 2014; 22(10):1533-1545.
38. Silver D, Huang A, Maddison C J, et al. Mastering the game of Go with deep neural networks and tree search. Nature 2016; 529(7587):484-489.
39. Zeiler MD, Fergus R. Stochastic pooling for regularization of deep convolutional neural networks. (2016-01-11).
40. Murphy KP. Machine Learning: A Probabilistic Perspective. Cambridge, MA: MIT Press, 2012:82-92.
41. Chatfield K, Simonyan K, Vedaldi A, et al. Return of the devil in the details: delving deep into convolutional nets. (2016-01-12).
42. Goodfellow IJ, Warde-Farley D, Mirza M, et al. Maxout networks. (2016-01-12).
43. Lin M, Chen Q, Yan S. Network in network. (2016-01-12).
44. Montavon G, Orr G, Mvller KR. Neural Networks: Tricks of the Trade. London: Springer, 2012:49-131.
45. Bengio Y, Simard P, Frasconi P. Learning long-term dependencies with gradient descent is difficult. IEEE Transact Neur Network 1994; 5(2):157-166.
46. He K, Zhang X, Ren S, et al. Delving deep into rectifiers: surpassing human-level performance on ImageNet classification // Proceedings of the 2015 IEEE International Conference on Computer Vision. Piscataway, NJ: IEEE, 2015:1026-1034.
47. Hinton GE, Srivastava N, Krizhevsky A, et al. Improving neural networks by preventing co-adaption of feature detectors (R/OL). (2015-10-26).
48. Wan L, Zeiler M, Zhang S, et al. Regularization of neural networks using dropconnect // Proceedings of the 2013 International Conference on Machine Learning. New York: ACM Press, 2013:1058-1066.
49. He K, Sun J. Convolutional neural networks at constrained time cost // Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. Washington, DC: IEEE Computer Society, 2015:5353-5360.
50. Springenberg JT, Dosovitskiy A, Brox T, et al. Striving for simplicity: the all convolutional net. (2015-12-24).
51. Van Der Maaten L, Hinton G. Visualizing data using t-SNE. (2015-12-24).
52. Oliva A, Torralba A. Modeling the shape of the scene: a holistic representation of the spatial envelope. Int J Comput Vis 2001; 42(3):145-175.
53. Wang J, Yang J, Yu K. Locality-constrained linear coding for image classification // Proceedings of the 2010 IEEE Conference on Computer Vision and Pattern Recognition. Washington, DC: IEEE Computer Society, 2010:3360-3367.
54. Zeiler MD, Taylor GW, Fergus R. Adaptive deconvolutional networks for mid and high level feature learning // ICCV '11: Proceedings of the 2011 International Conference on Computer Vision. Piscataway, NJ: IEEE, 2011:2018-2025.
55. Nguyen A, Yosinski J, Clune J, et al. Deep neural networks are easily fooled: high confidence predictions for unrecognizable images // Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Washington, DC: IEEE Computer Society, 2015:427-436.
56. Floreano D, Mattiussi C. Bio-inspired Artificial Intelligence: Theories Methods and Technologies(M). Cambridge, MA: MIT Press, 2008:1-97.
57. Zhuang FZ, Luo P, He Q, et al. Survey on transfer learning research. J Software 2015; 26(1):26-39.
58. Li F, Fergus R, Perona P. One-shot learning of object categories. IEEE Transact Pattern Anal Mach Intel 2006; 28(4):594-611.
59. Griffin BG, Holub A, Perona P. The Caltech-256 (R/OL). (2016-01-03).
60. Zhou B, Lapedriza A, Xiao J, et al. Learning deep features for scene recognition using places database // Proceedings of Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press. 2014:487-495.
61. Loffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift. (2016-01-06).
62. Girshick RB. Fast R-CNN. (2016-01-06).
63. Ren S, He K, Girshick R, et al. Faster R-CNN: towards real-time object detection with region proposal networks. (2016-01-06).
64. Uijlings J, Sande K, Gevers T, et al. Selective search for object recognition. International Journal of Computer Vision, 2013, 104 (2):154-171.

65. Khan SH, Bennamoun M, Sohel F, et al. Automatic feature learning for robust shadow detection // CVPR'14: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. Washington, DC: IEEE Computer Society, 2014:1939-1946.
66. Taigman Y, Yang M, Ranzato M, et al. DeepFace: closing the gap to human-level performance in face verification // CVPR'14: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. Washington, DC: IEEE Computer Society, 2014:1701-1708.
67. Schroff F, Kalenichenko D, Philbin J. FaceNet: a unified embedding for face recognition and clustering // Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Washington, DC: IEEE Computer Society, 2015:815-823.
68. Levi G, Hassner T. Age and gender classification using convolutional neural networks // Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops. Washington, DC: IEEE Computer Society, 2015:34-42.■