Hindawi
Mathematical Problems in Engineering
Volume 2022, Article ID 2213295, 11 pages
https://doi.org/10.1155/2022/2213295

Research Article

Image Target Recognition Based on Improved Convolutional Neural Network

Jinjuan Wang, Xiliang Zeng, Shan Duan, Qun Zhou, and Hao Peng

Hunan International Economics University, Changsha 410205, Hunan, China

Correspondence should be addressed to Jinjuan Wang; wangjinjuan1532@163.com

Received 7 May 2022; Accepted 16 June 2022; Published 8 July 2022

Academic Editor: Zaoli Yang

Convolutional neural network (CNN) algorithm is a very important branch of deep learning research, which has been widely applied in many fields and achieved excellent results, especially in computer vision, where convolutional neural network has made breakthroughs in image classification and object detection. Convolutional neural network architecture can realize more efficient network training through the final combination of different modules, and the convolutional neural network training does not need to actively extract image features and can directly carry out end-to-end training and prediction. At first, this paper analyzed some problems of the current image recognition and expounds the progress of convolution neural network in image recognition and then studied the traditional algorithm of target recognition, including traditional recognition algorithm framework of target, the target orientation, feature extraction, classifier classification, etc., and the traditional target recognition algorithm is compared with those of the target recognition algorithm of deep learning. On the basis of the above research, an improved model of CNN is proposed, which focuses on the structural design and network optimization of convolutional neural network and designs a more efficient convolutional neural network. Test experiments verify the effectiveness of the proposed model, which not only achieves lower error rate, but also greatly reduces the number of network parameters and has stronger learning ability.

1. Introduction

The purpose of target recognition is to automatically recognize the position information and the category of a single object in each image. In this research field, satisfactory results can be obtained when processing relatively simple image scenes or images of objects with clear foreground. However, there are still many problems in the recognition results when processing the images of various complex objects. The object shape and the chaotic environment bring difficulties to the target recognition. The technical solutions of target recognition can be roughly divided into two types: traditional method oriented and deep learning based. They have fundamental differences in the solutions of target recognition tasks. Traditional methods use a manually designed feature extractor to extract and recognize the features of objects. The modeling process of this kind of methods is cumbersome, and the generalization of the model is insufficient. However, in dealing with the recognition task in a specific scene, the effect is better, and the hardware requirements are low [1]. The training of big data is mainly used for in-depth learning. Through reasonable neural network modeling, a large amount of data is used as model input, and the model automatically extracts features. The advantages of this kind of method are that the modeling method is simple, and the network is used to adjust parameters, train, and test. Finally, good results can be achieved without specific explanation of features. The model has strong generalization ability, but high hardware requirements; it relies on a large amount of data and has weak interpretability [2]. At present, in the target recognition technology which takes image and video as the processing object, with the rapid development in recent years, convolutional neural network gradually shows its absolute advantage in the field of image target recognition. The target recognition algorithm using convolutional neural network is much faster than the traditional target recognition algorithm in model transfer learning and recognition speed [3].

As one of the important methods of image processing, the basic idea of image recognition is to extract the features of a given image data and use a classification algorithm to recognize its features and predict the label category of the image [4]. The steps are roughly as follows: input of image information, preprocessing, feature extraction, feature selection, and classifier design. Traditional feature extraction methods are manual extraction, but manual extraction of features is often inefficient and requires the support of professional knowledge, so it requires a lot of cost, which is not conducive to widespread application [5]. Therefore, for this problem, deep learning has been widely concerned and applied with its powerful feature extraction ability. As one of the classical methods of deep learning, convolutional neural network is widely used in the field of image processing. Since features are extracted layer by layer from images in an end-to-end learning mode without the help of human extraction, and the extracted features are more abstract, it can fully extract image features and mine image information in image processing and has stronger feature extraction ability, nonlinear mapping ability, and model robustness [6]. Because the convolution and pooling layer is the main component of convolution neural network this design makes the training parameters compared with artificial neural network greatly reduce, and the extraction of image characteristics also has displacement invariance and scale invariance and deformation of the invariance, so it is widely applied in the areas of image processing [7]. In recent years, research methods of deep learning represented by convolutional neural networks have played an increasingly important role in artificial intelligence fields such as image processing, natural language processing, and brain wave analysis, having huge social value and economic benefits [8]. In the image recognition task, the features extracted from different layers of the network have their corresponding utilization value, and the features extracted from different layers represent different learning abilities. The current convolutional neural network fully focuses on local features, and the insufficient extraction of global feature information of the entire network is not conducive to the fusion learning among network layers, and the information exchange between network layers is not sufficient. Therefore, the full integration of features between different layers is conducive to the training of the whole network and has an important impact on the improvement of network performance [9].

Simonyan et al. proposed VGGNet on the basis of AlexNet and studied the influence of network depth on convolutional neural network. This network uses multiple small 3×3 convolutional kernels to replace large 11×11 convolutional kernels, which not only reduces the number of model parameters but also speeds up the training speed of the model. Moreover, the feature representation capability of the model is improved. By deepening the convolutional layer of the network, higher level features can be extracted [10]. However, with the increasing number of network layers, the calculation of the whole model is very large, which also requires stronger hardware conditions. Secondly, it will also bring problems such as gradient disappearance and overfitting, so the network performance cannot be improved by simply increasing the number of network layers. GoogleNet is a brand new network model proposed by the Google team. The model optimizes the network structure to significantly reduce the number of parameters and computation in the network. Through the fusion of multiple convolutional kernels of different sizes in its Inception module, the convolution operation is performed on the input of the upper layer. The features extracted from the model are more abundant. In addition, at the end of the network, the full connection layer is replaced by the global average pooling layer, which further reduces the number of network parameters. GoogleNet has more layers than VGGNet, but the number of parameters is greatly reduced. Meanwhile, the classification accuracy of the ImageNet data set is much higher than the previous network model [11]. With the deepening of network layer, the precision of the model actually reduces the degradation problems. Kaiming people such as the depth of the residual network ResNet [11], solve the network degradation problems, through the use of short connection that will lower the network layer and cross connection layer at the top of the network layer, save the low-level network layer information, and make the network the depth of the deeper. As the network layer is greatly deepened, the accuracy rate is also greatly improved. Huang et al. proposed the Dense Convolution Network (DenseNet) [13], which won the best paper award of CVPR. Its core idea is to connect all layers of the network with each other, and each layer will accept all the previous layers, making full use of the extracted features of each layer and effectively alleviating the problem of gradient disappearance. The bottleneck layer and transition layer are also used to greatly reduce the computation and parameter quantity of the whole network and alleviate the overfitting phenomenon. Li et al. proposed that selecting kernel convolution can make the network dynamically adjust the receptive field size according to the input [14]. Kong aimed at the disadvantages of the traditional machine-based facial expression recognition method that eliminates the feature of manual selection and proposed a feature extraction method based on deep convolutional neural network to learn expression features [15]. Haase and Amthor further reduced the number of parameters in the convolution layer by using "blueprint" convolution instead of deeply separable convolution [16]. Kortylewski et al. effectively solved the problem of partially occluded image classification by synthesizing convolutional neural network [17]. Due to the large number of parameters and deep layers of traditional convolutional neural network model, it is difficult to train, which limits the application of convolutional neural network. In recent years, some lightweight convolutional neural network models have become the focus of research. The SqueezeNet stack proposed by Iandola et al. uses the Fire Module to achieve similar accuracy to AlexNet on the imagenet dataset, but with 50 times fewer parameters [18]. MobileNet [19] and its variants MobileNetV2 [20] and MobileNetV3 [21] proposed by Howard et al. replace the traditional convolution kernel with deep separable convolution. The number of parameters in the model is reduced, the calculation efficiency is high, and it can be deployed on the mobile terminal. Zhang et al.

proposed ShuffleNet [22] and ShuffleNetv2 [23], which introduced grouping convolution to deeply separable convolution to further reduce computational complexity and the number of network parameters and used channel mixing to enhance information exchange between channels after grouping, which fully integrated extracted features and greatly improved network performance.

2. Target Recognition Based on Traditional Algorithm

2.1. Traditional Target Recognition Algorithm Framework. The traditional target recognition and detection algorithm can be divided into three parts: target location operation, which is mainly used to locate the area where the target may appear and obtain as much target information as possible by taking this area as a sample; feature extraction module, which is responsible for feature extraction of acquired information; classifier. In the end, classifier is used for training and object recognition and distinction in the image.

2.1.1. Target Positioning. The main operation of target location is realized by window sliding. This operation is to use different size region selection box to traverse all areas of the image, and then all the area has a slide with the label for the target position information and the operation of the localization and classification according to this, with the final output for all regions to judge selection box and select the result of the highest degree of confidence. In general, the measurement standard of target positioning task is the Intersection-over-Union (IoU), which represents the overlap rate between the candidate area box and the original marked area, that is, the ratio of their intersection and union.

2.1.2. Feature Extraction. Image data can usually be regarded as composed of pixels and corresponding color mapping, so the image can be further transformed into a matrix of numbers representing different color values, and the row vector of each row in the matrix is the target of feature extraction. Traditional feature extraction methods mainly include the following:

(1) Scale-Invariant Feature Transform (SIFT). Because the object appearance on some of the local points of interest will have no direct association with the image size and rotation, and for illumination and noise value, the characteristics of small angle change has very high tolerance, in the feature extraction method based on the characteristics of the corresponding spatial scale for the extreme value point, the point is not affected by the rotation of the image, the influence of size transformation, and translation operation.

(2) Histogram of Oriented Gradient (HOG). The main idea of this feature extraction method is to clearly describe the surface image and contour shape of the local target in the image through the direction density distribution or gradient of the edge. Specifically, the gradient of each pixel in the image data cell unit of uniform size is collected to form its corresponding gradient direction histogram. The final target feature descriptor is composed of these histograms.

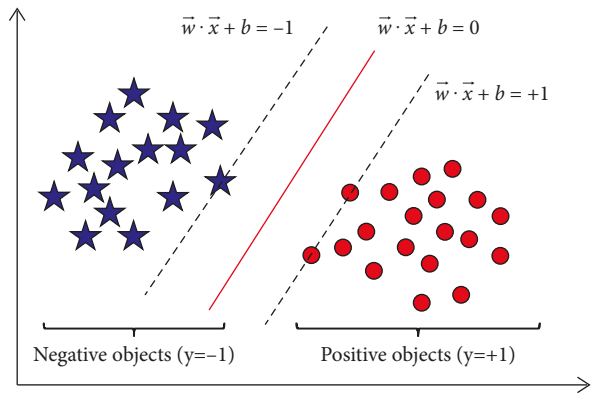

Figure 1: SVM classifier.

(3) Local Binary Pattern (LBP). The threshold value used by the original LBP operator is a pixel value located in the center of a $3*3$ window. By comparing it with the gray value of the adjacent 8 pixels, the position of the pixel point larger than this value is marked as 1; otherwise it is 0. After sequential comparison, the 8 pixels in the original $3*3$ window can generate $2^8=256$ different values after comparison as the LBP value of the center pixel of the window, and this value is used to reflect and represent the relevant information in this area.

2.1.3. Classification of Classifiers. By aggregating objects with similar features, the classifier classifies and judges the linear combination of these features. Generally, during the classification operation, the results of the front-end classification calculation will be compared with the preset threshold value of the classifier and classified and divided according to the results. Support vector machine (SVM) is a linear classifier that classifies data binary by supervised learning. For a two-dimensional space, SVM will divide the plane into two parts and a line, with the same or similar categories on the same side of the line, while for a multi-dimensional space, SVM will try to find a line or a hyperplane, as shown in Figure 1 below.

To sum up, the traditional target recognition and detection algorithm consists of the above three parts. In the process of target location, a large amount of location information will be generated to meet the requirements of subsequent feature extraction operations, but a large amount of location information will cause a waste of data and time because a large part of the information is useless for subsequent operations. In addition, in the traditional feature extraction operation, artificial features need to be set, and because the algorithm has no self-learning ability, in some cases, extracted features will lose significance. For example, when the target features change, it will have a negative impact on the subsequent classification detection task.

2.2. Target Recognition Based on HOG + SVM. HOG is the histogram of direction gradient, which is a feature description method used for object detection and also a traditional feature extraction method. The feature extraction is accomplished by calculating the gradient of the local area of the image and calculating the histogram of the direction obtained by statistics. SVM is a very classical binary classifier, which can realize image recognition and detection by combining the feature extraction method of HOG. Since SVM is essentially a binary classifier, the main method in the face of multiclassification tasks is to train a certain category as a positive sample set and all remaining categories as negative sample sets to obtain a classifier suitable for this category and then do the same operation for all remaining categories. In the test process, according to the test image and each category classification situation of the similar situation, select the most similar as recognition classification output.

The algorithm mainly obtains image features through HOG and uses the extracted features for SVM multiclassification training. In the testing process, SVM multiclassifier will be used to score the accuracy of all categories of the input test images and select the one with the highest score as the prediction classification of the test images.

2.3. Comparison between Traditional Target Recognition Algorithm and Target Recognition Algorithm Based on Deep Learning. The traditional target recognition and detection algorithm is suitable for the situation with obvious features and simple background. However, under normal circumstances, it is difficult for people to make a good summary of the features of the target for the machine, especially in the face of the real scene full of complex and changeable factors, the features that can be obtained often having a relatively abstract concept. For the complex background which is difficult to recognize and detect through general abstract features, it is difficult for the traditional algorithm to show good recognition effect. However, deep learning can extract rich features of the same target and adjust the model through self-learning to better complete the target identification and detection.

Compared with traditional algorithms, the main advantages of object recognition algorithm based on deep learning are reflected in feature extraction using deep neural network. The use of deep neural network can avoid artificial feature design work, to not only reduce the operation of human interference, but also enable the algorithm to fully extract the target feature autonomously, resulting in more excellent feature expression ability and higher recognition accuracy, with great advantages in accuracy and universality. Through training huge and rich data, deep learning extracts the features of the image target in depth and comprehensively in the process of training and then completes the corresponding recognition model. After full training, algorithms based on deep learning can have higher stability and anti-interference, stronger generalization ability, easier application in the actual scene, and also better recognition effect.

3. Improved Basic Structure of Convolutional Neural Network

The CNN model proposed in this paper consists of convolution layer, filter banks, residual learning, classifier, and so on.

3.1. The Convolution Layer. In order to improve the recognition efficiency, the CNN network proposed in this paper adopts the matrix to carry out the convolution operation on the image and carries out the convolution operation between the matrix of the convolution kernel and the local matrix of the input image. The resulting matrix dimension is the same as the dimension of the convolution matrix. In order to fully obtain various information in the image, multiple such matrices are required in the convolution layer to carry out multiple convolution operations and obtain multiple results. The operation process is shown in the following equation:

$$X_i^{i,j} = g\left(b_j + \sum_{\beta=1}^{s} w_\alpha^j x_{i+\beta-1}^{l-1,j}\right), \tag{1}$$

where i of $X_i^{i,j}$ represents the value of the convolution output matrix, j represents the number of an output matrix, and l represents the convolution layer, b_j represents the bias item, w_α^j represents the weight, $g(x)$ uses the Softmax activation function, and α represents the number of categories.

3.2. Filter Bank. The essence of filter is a matrix, through the matrix of different size to obtain the feature information of different aspects of the image. When the filter is used to calculate the input image, it is actually the product and sum of the size matrix with the same size area in the image. And in the matrix on the whole image from left to right, step size is 1 dot product operation, and then the obtained results are summed, with the filter on the whole image filtering operation. This multiplication and summation using a matrix with a portion of the image are the same as the convolution operation in CNN.

In order to obtain different feature information of images from multiple angles, the CNN model proposed in this paper firstly uses a multiscale filter bank, which has three filters of different sizes (1×1, 3×3, and 5×5) and performs local convolution on the input image.

Since the size of the feature graphs of the three convolution filters is different from each other, a method is used to adjust the size of the three feature graphs to the same size, so as to combine the joint feature graphs. First, the space is filled with zero around the input image, so that the size of the feature image obtained by the three filters with different sizes is changed to $(H + 4, W + 4)$, $(H + 2, W + 2)$, and (H, W). H and W are the height and width of the input image, respectively. The input of CNN model proposed in this paper is an array of pixel values of 32×32. After the filter is filled with zero, the size of the three filters becomes 5×5. After the convolution operation of the image, a matrix of 28×28 is obtained, which is the characteristic information of the

image obtained through the convolution operation of the filter.

After the image passes through the filter bank, its combined feature graph will be output as the input of the next convolution layer. In this paper, the convolutional layer window is set as 1×1, and the number of kernels is set as 128. Finally, 128 feature maps with size 28×28 can be obtained.

3.3. Residual Learning. In the process of backpropagation, it is necessary to take the derivative of the activation function. If the derivative is greater than 1, with the increase of network layers, the gradient update will increase in the way of exponential explosion, that is, gradient explosion. On the contrary, if it is less than 1, the gradient update will decrease in the way of exponential attenuation with the increase of network layers; that is, the gradient will disappear. Therefore, as the number of layers of convolutional neural network deepens, problems of gradient disappearance and gradient explosion will occur. However, residual learning can directly transfer the matrix values of the previous layer to the deeper layer of the network, instead of downward transmission along the network layer by layer. The formula of residual learning is shown in the following equation:

$$y = F(x, \{w_i\}) + x, \tag{2}$$

where x and y are the input and output of the convolution layer, respectively. Function $F = y - x$ is the residual mapping of the input to the residual output $y - x$ using the convolution filter W_i.

3.4. Classifier. Classifier refers to classifying the input image into the correct category after training the correctly labeled data set. The cross entropy loss function is chosen in this paper. The cross entropy represents the difference between the actual output and the expected output of the model. In image classification, the smaller its value is, the closer the classification of the image classified by the classifier is to the actual category of the image, and the higher the classification accuracy is. The calculation formula is shown in the following equation:

$$H(p, q) = -\sum_{i=1}^{n} p(x_i) \log(q(x_i)), \tag{3}$$

where $p(x_i)$ is the expected probability distribution, $q(x_i)$ is the actual probability distribution, n is the number of categories, and p and q are the probability values.

Cross entropy describes the distance between two probability distributions, but the final output of convolutional neural network is a real number in many cases, not a probability distribution. Therefore, the results of the network output can be converted into probability distributions using the Softmax function. Softmax function can reduce the output of each neuron to a range of 0 to 1; that is, if it is assigned to the correct class, the output will be 1; otherwise, the output will be 0, so as to complete the task of classification.

Assume that the output of neural network is $y_1, y_2, \ldots, y_n$; then the formula of Softmax function is shown in the following equation:

$$\text{softmax}(y)_i = \frac{e^{y_i}}{\sum_{j=1}^{n} e^{y_i}}, \tag{4}$$

where layer $\text{softmax}(y)_i$ represents the output of neuron i, y represents the input, e is a constant, and n represents the number of neurons.

3.5. Attention Mechanism. The attention mechanism can also be considered as a set of weight coefficients, which are not artificially given but obtained by the network's autonomous learning in the training process. The attention mechanism is added and the target area is highlighted by "dynamic weighting," while the background area is suppressed as a nonimportant feature. The mechanism of attention is generally divided into two categories: one is the strong attention which depends on the prediction of random dynamic change; the other is the soft attention learned by gradient descent in the network. Strong attention is affected by the nondifferentiable nature. Although the prediction effect of the network containing strong attention is good, it is not widely used in the network. Soft attention can be obtained by neural network training with gradient descent method, and its differentiability everywhere makes its application more common than strong attention mechanism. At present, the soft attention mechanism widely used in the network can be divided into two types: channel attention, which pays more attention to channel correlation, and regional spatial attention.

3.5.1. Channel Attention. The purpose of channel attention is to show the correlation between different feature maps. The importance of each feature channel is learned from the network. Finally, different weight coefficients are assigned to each channel to enhance the recognition of important feature information in feature maps and suppress nonimportant feature information in feature maps. The compression and excitation networks in the channel attention mechanism are very representative, which adaptively adjust the channel characteristic response by means of feature recalibration.

Compression and excitation networks retain the valid information in the original feature map and transform the spatial information to other spaces. Compression operation F_{sq} uses a single channel to represent the global spatial features of each channel, and the statistics of each channel are generated by global average pooling. Incentive operation F_{ex} is to learn the dependency of each channel, and the functional graph after the compression operation of the dependency of each channel will modify the dependency of the original channel to adapt to the new dependency, and the characteristic graph obtained after

the compression excitation adjustment is the required module.

The compression excitation module can be regarded as F_{tr} computing unit. F_{tr} is a convolution operator, which is mainly established in F_{tr} through transformation from input $X \in R^{H' \times W' \times C'}$ feature mapping to $U \in R^{H \times W \times C}$. $V = [v_1, v_2, \ldots, v_c]$ represents the convolution kernel learning set, where the c convolution kernel parameter is expressed as v_c. The output is denoted by $U = [u_1, u_2, \ldots, u_c]$. u_c in output U is expressed by the following equation:

$$\begin{aligned} u_c &= v_c * X \\ &= \sum_{s=1}^{C'} v_C^s * x^s. \end{aligned} \tag{5}$$

In equation (5), $u_c = R^{H \times W}$; $v_c = [v_c^1, v_c^2, \ldots, v_c^{\dot{C}}]$; $*$ represents the convolution of v_c and X; $X = [x^1, x^2, \ldots, x^{C'}]$; v_C^s is a two-dimensional spatial nucleus representing a single channel of v_c acting on the corresponding channel in X.

In order to strengthen the important feature information in the feature graph and suppress the nonimportant feature information in the feature graph, it is necessary to consider the signal of each channel in the output characteristic to make full use of the correlation between channels. Information other than this area will not be used in every unit of U transformation output. Global average pooling is used to compress all information in the channel into a single channel and make statistics. In other words, the spatial dimension $H \times W$ is used to compress U to generate $z \in R^C$, and the calculation formula of the C th z is shown in the following equation:

$$\begin{aligned} z_c &= F_{sq}(u_c) \\ &= \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} u_c(i, j). \end{aligned} \tag{6}$$

To make full use of the information obtained in the compression operation, the excitation operation should be carried out to ensure the complete capture of channel correlation. In order to learn the nonlinear interaction in the channel and ensure that multiple channels are used, the nonlinear activation function is used here, as shown in the following equation:

$$\begin{aligned} s &= F_{ex}(z, W) \\ &= \sigma(g(z, W)) \\ &= \sigma\left(W_2 \delta\left(W_1 z\right)\right). \end{aligned} \tag{7}$$

In equation (7), σ represents the activation function ReLU; and $W_1 \in R^{C/r \times C}$, $W_2 \in R^{C/r \times C}$; σ represents the activation function Sigmoid. Finally, in order to avoid overly complex model and auxiliary generalization, it is necessary to restore the original input dimension of the image.

$$x_c = F_{\text{scale}}(u_c, s_c) = s_c u_c. \tag{8}$$

In equation (8), $\widehat{X} = [x_1, x_2, \ldots, x_c]$ and $F_{\text{scale}}(u_c, s_c)$ represent $u_c \in R^{H \times W}$ and index quantity s_c for channel multiplication.

3.5.2. Spatial Attention. Spatial attention through space conversion module implements the key space of the original input image region information conversion to other spaces, keeping image needs to be focused on the characteristics of the regional information, and the characteristics of the key information of each pixel generated weights in the figure mask, mask after getting the result of the weighted after the output, through the conversion operation of key information to strengthen information target specific regions of interest and suppress background images that are not interesting.

In the spatial transformation network, the input of the whole model is $U \in R^{H \times W \times C}$. When passing through the convolution layer, different channel information will be formed due to different convolution kernels set on the convolution layer. When the input tensor U enters the space module, it will be divided into two parts to perform different operations. Part of the information will enter the location network to generate a set of parameters θ that can be used as the grid generator. After transformation, a transformation matrix in the sampling layer, namely, the sampling signal, will be generated. The other part of the information is the original input image which goes straight into the sampling layer without doing anything. In the sampling layer, the output matrix $V \in R^{H' \times W' \times C}$ can be obtained by directly multiplying the original image and sampling signal.

The steps of the proposed method are shown below.

Step 1. Enter the initial image M_i.

Step 2. The filter bank H is used to process the image M_i , and the feature information L_i of the image is obtained.

Step 3. Use the convolution kernel S_j to carry out convolution operation on the feature information L_i, and obtain the new feature information N_i.

Step 4. Through the residual learning module, the characteristic information N_i of the next layer is transmitted across layers.

Step 5. Study and train the feature information N_i.

Step 6. Classify the results W_i.

Step 7. Output W_i.

4. Experimental Results and Performance Analysis

Compared with general based on the shallow learning areas, namely, training layer relatively less traditional machine learning algorithm architecture, deep learning with a deeper depth algorithm structure, depth, and the structure mainly

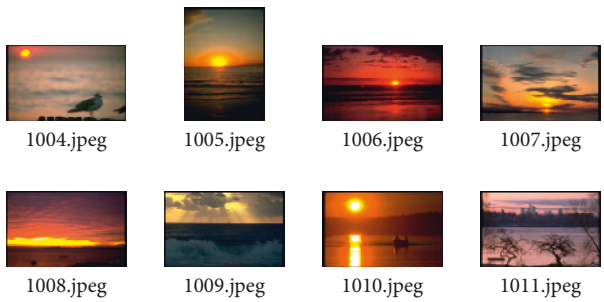

Figure 2: Partial Corel dataset.

Figure 3: Partial Labelme dataset.

comes from multiple hidden layer in the network constructed, through to characteristics of extraction in the hidden layer, in order to realize network model structure of multilevel and high depth. It can effectively describe the features of image data better and more accurately and has great advantages in the field of image processing.

4.1. Data Set Description. Three data sets were used. These are Corel, Labelme, and PASCAL VOC datasets. The Corel dataset contains 10 categories of specific objectives and realistic scenes, of which each category has 100 images. In the training set, there are 800 training images, 80 in each category. There were 200 images in the test set, including 20 images in each category. The images in the training set and test set belong to their own unique categories, and there is no category overlap, as shown in Figure 2 below.

The main features of Labelme dataset include the design for object classification and recognition, not just instance recognition; designed specifically for learning about objects embedded in a scene; high-quality pixel-level annotations, including polygons and segmentation masks. The diversity of object categories is large, and the diversity of each object is also large. The images in the training set and test set belong to their own unique categories, and there is no category overlap, as shown in Figure 3 below.

The PASCAL VOC dataset is a benchmark for classification recognition and detection of visual objects, providing a standard image annotation dataset and a standard evaluation system for detection algorithms and learning performance, and contains VOC2007 and VOC2012 two versions. See Figure 4 below.

Figure 4: Partial PASCAL VOC dataset.

Table 1: Accuracy of CNN model in this paper.

Dataset	Accuracy (%)
Corel	90.3
Labelme	79.42
PASCAL VOC	84.75

By comparing the image composition of the three data sets, it can be analyzed that there is a large difference between the image data of Corel and Labelme, and it is relatively easy to extract high-quality features. Although PASCAL VOC data set has more categories, due to the small differences among the same pet breeds, it is easy to identify errors even if human observation is made, so a more in-depth feature extraction method should be required.

4.2. Target Recognition Results Based on Deep Learning Method. The experiment uses Epoch = 15 and Batch Size = 10 to train the data set prepared above, where batch size represents the number of samples needed to calculate the gradient, meaning that 10 data should be read in each iteration, while Epoch represents the algebra in which all the training data have gone through one training. For Corel and PASCAL VOC datasets, the number of iterations required for each generation is 800/10 = 80, while for Labelme datasets, 1200/10 = 120 iterations are required for each generation. The proportion of training and testing is the same, 80% training and 20% testing. The trained model is used to traverse the test data of three datasets to test the classification and recognition effect and statistical accuracy. The accuracy obtained through the test set is shown in Table 1 below.

According to the experimental results, the target recognition accuracy based on deep learning algorithm is at a high level, and it can also achieve good recognition effect for data with abstract features and relatively small differences between different label classifications like Labelme dataset. In the training process, the curves of training results presented after each iteration are shown in Figures 5–7.

Objective represents the change of the total error loss curve, Top1err represents the change of the error curve of the highest probability results, and Top5err shows the change of the error curve of the top5 results with the highest probability compared with the real situation. It can be seen

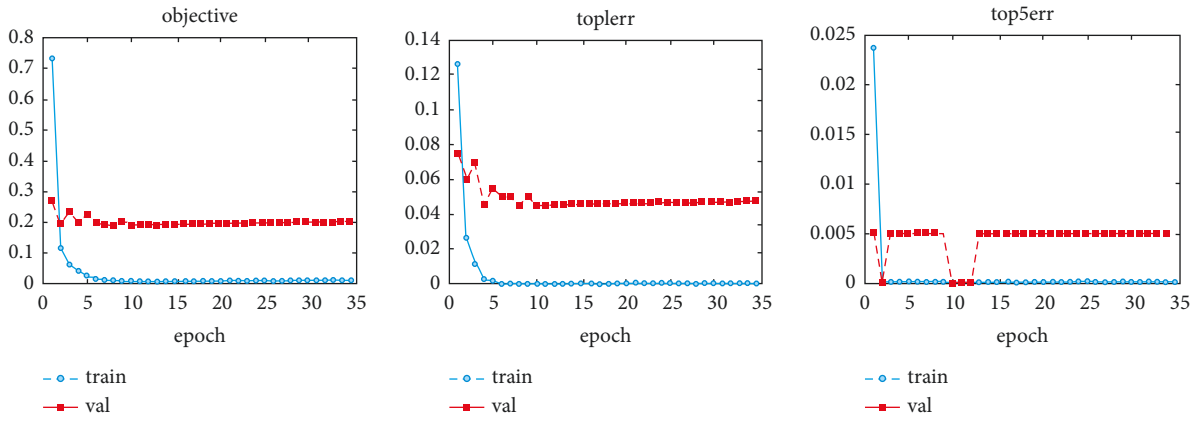

Figure 5: Training result curve of Corel dataset.

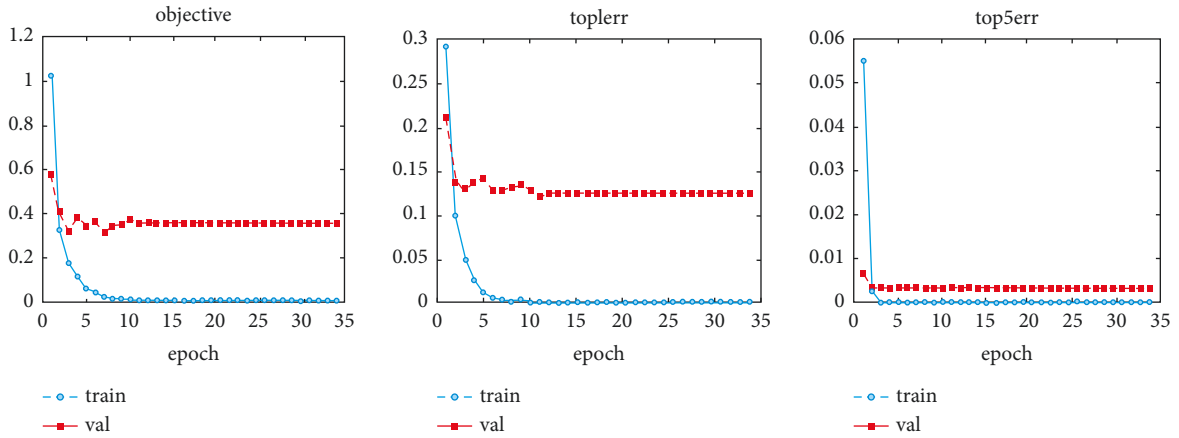

Figure 6: Training result curve of Labelme dataset.

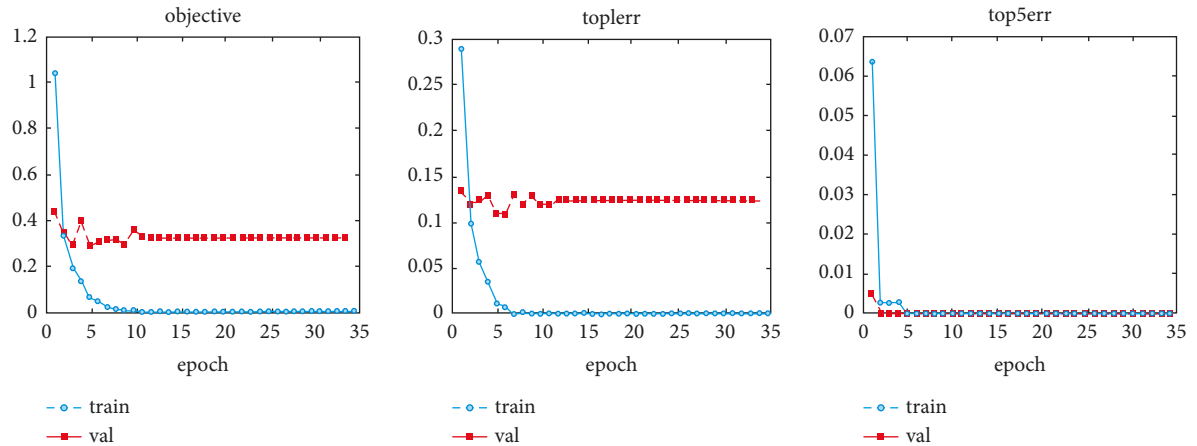

Figure 7: Training result curve of PASCAL VOC dataset.

from the training result curve that the error curves of the three data sets basically approach 0 after the number of iterations of the training is more than 10, indicating that after the training of the model structure to the 10th generation, each error gradually converges and the network model tends to be stable.

Through the training and testing of the above three data sets, it can be seen that the target recognition algorithm based on deep learning can show good recognition accuracy for data sets of different sizes and diverse features, and the error result curve in the training process also shows that the model has certain stability.

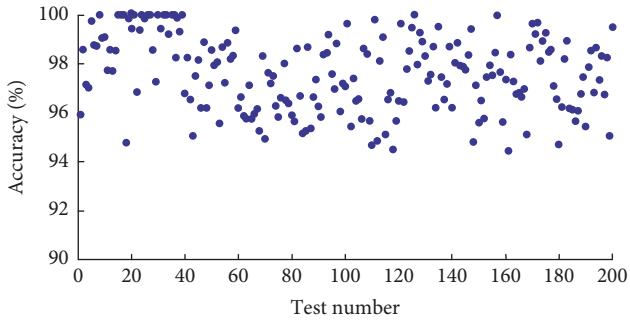

Figure 8: Dispersion point diagram of test evaluation in Corel dataset.

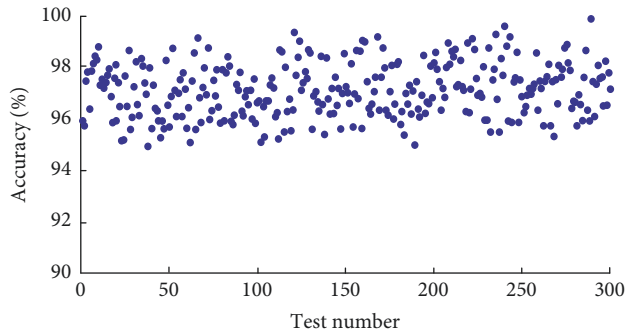

Figure 9: Scatter diagram of test evaluation in Labelme dataset.

4.3. Results of Traditional Target Recognition Algorithm. The two traditional target recognition algorithms analyzed above, namely, the target recognition algorithm based on HOG + SVM and the target recognition algorithm based on BOF, are, respectively, trained and tested on the three training sets prepared, and the accuracy of the traditional algorithm is observed and recorded, so as to conduct a comparative analysis with the CNN algorithm proposed in this paper.

4.3.1. Target Recognition Results Based on HOG + SVM. Figures 8–10 show the scatter points of the highest accuracy evaluation of the target recognition algorithm based on HOG + SVM in the testing process.

The abscissa corresponds to the number of the test image to be recognized, and the ordinate represents the highest accuracy score obtained during the classification and recognition of the test image. In the test process, the model selects its corresponding category according to the score as the predictive recognition and classification output. It can be seen that the trained SVM multi-classification model has a high score for the output classification of the three data sets in the test, proving that all the recognition results have been the results with the highest confidence. Then, the recognition results of all test images are counted, and the accuracy rate after the test is shown in Table 2 below.

Experimental results show that the recognition accuracy of object recognition algorithm combined with HOG + SVM

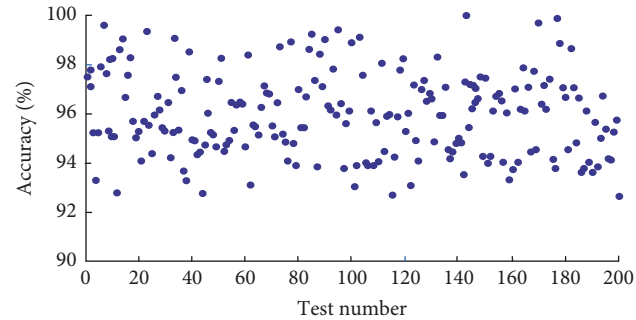

Figure 10: Dispersion point diagram of test evaluation for PASCAL VOC dataset.

Table 2: Accuracy of HOG + SVM algorithm.

Dataset	Accuracy (%)
Corel	63.28
Labelme	38.56
PASCAL VOC	50.54

Table 3: Accuracy of BOF algorithm.

Dataset	Accuracy (%)
Corel	83.61
Labelme	59.33
PASCAL VOC	67.82

for Corel image library reaches 63.28%, while for the recognition result of Labelme data set, only 38.56%, conforms to the actual target classification, and PASCAL VOC set recognition result reaches nearly 50.54%.

4.3.2. Target Recognition Results Based on BOF. Then we also train and test the target recognition algorithm based on BOF with the prepared data set. In the training process, the algorithm will generate a prediction matrix for testing and give the average accuracy achieved in the training. The final accuracy is shown in Table 3 below.

According to the experimental results, the recognition accuracy of Corel dataset reaches 83.61%, but the recognition accuracy of Labelme dataset is only 59.33%, although it is higher than that of HOG + SVM algorithm. Compared with HOG + SVM algorithm, the recognition accuracy of PASCAL VOC dataset decreases slightly, but it is at the same recognition level in general.

According to the results of the experiments comparison and analysis, it can be found that the traditional target recognition algorithm for relatively obvious characteristics or differences between different categories of data can show relatively good recognition effect, but with the data characteristics of the abstract or the tiny difference between different categories, it is hard to target recognition. This means that more in-depth and comprehensive feature extraction is needed to improve the recognition effect.

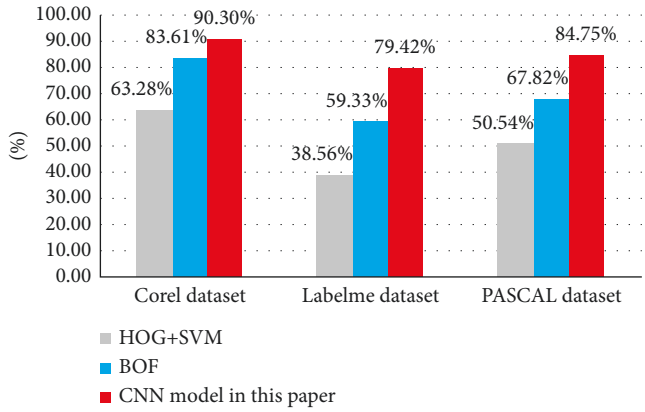

Figure 11: Comparison of algorithm accuracy.

4.4. Comparison between Deep Learning Algorithm and Traditional Algorithm. Taking the test recognition accuracy of the traditional algorithm as the reference value, a specific analysis is made by comparing the recognition accuracy of the traditional algorithm with that based on deep learning, as shown in Figure 11 below.

Through the comparison of experimental results, the performance of the target recognition algorithm in this paper is better than the traditional algorithm in the two data sets, and its accuracy has reached a relatively high level. Especially for Labelme data set and PASCAL VOC dataset, the recognition effect is much better than the traditional algorithm.

From the results, the target recognition algorithm based on deep learning is far more accurate than the traditional target recognition algorithm in training effect. Compared with traditional algorithm, the biggest advantage of deep learning algorithm is the use of the depth of the convolution neural network for feature extraction, through deep neural networks which can be sufficient to extract the characteristics of image data and can be largely the preserve of the characteristics of the target image quality, even in the face of the similarity between feature abstraction and classification of image data. Also it can ensure high accuracy and can be widely used in many scenarios. Due to the backward feature extraction methods of traditional algorithms, the feature data obtained by them can hardly meet the demand of image processing nowadays, and the amount of calculation is also equal. For the classification and recognition effect of a single object, the traditional algorithm can guarantee the accuracy, but its universality is not enough, and in the face of multitarget image data, a lot of targeted training is needed.

5. Conclusion and Future Work

In recent years, the structure innovation of convolutional neural network and parameter optimization have made some achievements in computer vision related tasks. However, with the increase of network depth and the increase of network model, overfitting and other problems are prone to occur. Starting from the basic architecture of convolutional neural network, this paper aims to design convolutional neural network with better performance in all aspects. The classical convolutional neural network has a single convolutional operation, a single convolutional module, and a long training time. Therefore, it is necessary to design diversified convolutional modules to carry out complex convolution, break the tradition, and improve performance. In order to improve the network structure and improve the network performance, this paper designed a convolutional neural network with higher recognition performance and a more lightweight model and tested it on the open data set to verify the network performance, which has certain practical value.

Future research work is as follows:

(1) The overall recognition performance and model of the cross-cascaded convolutional neural network have been improved, and further tests and application effects are needed in other fields in the future.

(2) Although the diversified module design improves the overall network performance, the recognition rate needs to be further improved while realizing network lightweight. Therefore, the recognition rate should be further improved to achieve higher accuracy while realizing network lightweight and training speed in the future.

(3) In the following research, the structure of convolutional neural network should be further explored to design a more effective combination method to achieve higher recognition rate.

Data Availability

The authors confirm that the data supporting the findings of this study are available within the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the Undergraduate Courses Projects of Hunan (Xi'an Jiaotong [2020] no. 9).

References

[1] N. Xiao and Z. Song, "Signal interference detection algorithm based on bidirectional long short-term memory neural network," *Mathematical Problems in Engineering*, vol. 2022, Article ID 4554374, 7 pages, 2022.

[2] H. Shi, N. Zhang, X. Wu, and Y. Zhang, "Multimodal lung tumor image recognition algorithm based on integrated convolutional neural network," *Concurrency and Computation: Practice and Experience*, vol. 32, no. 21, 2020.

[3] L. Liu and X. Yang, "Multi-stream with deep convolutional neural networks for human action recognition in videos," *Neural Information Processing*, vol. 11301, pp. 251–262, 2018.

[4] M. Mellouli, M. Hamdani, J. J. Sanchez-Medina, M. Ben Ayed, and A. M. Alimi, "Morphological convolutional neural network architecture for digit recognition," *IEEE Transactions on*

Neural Networks and Learning Systems, vol. 30, no. 9, pp. 2876–2885, 2019.

[5] L. Xue, X. Zhong, R. Wang, J. Yang, and M. Hu, "Low - resolution vehicle recognition based on deep feature fusion," *Multimedia Tools and Applications*, vol. 77, no. 20, Article ID 27617, 2018.

[6] R. D. Singh, A. Mittal, and R. K. Bhatia, "3D convolutional neural network for object recognition: a review," *Multimedia Tools and Applications*, vol. 78, no. 12, Article ID 15951, 2019.

[7] W. Y. Sun, H. T. Zhao, and Z. Jin, "A facial expression recognition method based on ensemble of 3D convolutional neural networks," *Neural Computing & Applications*, vol. 31, no. 7, pp. 2795–2812, 2019.

[8] X. He and W. Zhang, "Emotion recognition by assisted learning with convolutional neural networks," *Neurocomputing*, vol. 291, pp. 187–194, 2018.

[9] K. Karambakhsh, B. Sheng, P. Li, Po Yang, Y. Jung, and F. Feng, "VoxRec: hybrid convolutional neural network for active 3D object recognition," *IEEE Access*, vol. 8, Article ID 70969, 2020.

[10] X. Zhu, M. Zhu, and H. Ren, "Method of plant leaf recognition based on improved deep convolutional neural network," *Cognitive Systems Research*, vol. 52, pp. 223–233, 2018.

[11] H. Yang, C. Yuan, Li Zhang, Y. Sun, W. Hu, and J. Maybank, "Sta-Cnn: STA-CNN: convolutional spatial-temporal attention learning for action recognition," *IEEE Transactions on Image Processing*, vol. 29, pp. 5783–5793, 2020.

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, Las Vegas, NV, USA, June 2016.

[13] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4700–4708, Honolulu, HI, USA, July 2017.

[14] X. Li, W. Wang, X. Hu et al., "Selective kernel networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 510–519, Long Beach, CA, USA, June 2019.

[15] F. Kong, "Facial expression recognition method based on deep convolutional neural network combined with improved LBP features," *Personal and Ubiquitous Computing*, vol. 23, no. 3-4, pp. 531–539, 2019.

[16] D. Haase and M. Amthor, "Rethinking depthwise separable convolutions: how intra-kernel correlations lead to improved MobileNets," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Article ID 14600, Seattle, WA, USA, June 2020.

[17] A. Kortylewski, J. He, and Q. Liu, "Compositional convolutional neural networks: a deep architecture with innate robustness to partial occlusion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8940–8949, Seattle, WA, USA, June 2020.

[18] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and< 0.5 MB model size," 2016, https://arxiv.org/pdf/1602.07360.

[19] A. G. Howard, M. Zhu, B. Chen et al., "Mobilenets: Efficient Convolutional Neural Networks for mobile Vision Applications," arXiv preprint arXiv:1704.04861, 2017.

[20] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. C. Chen, "Mobilenetv2: inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4510–4520, Salt Lake City, UT, USA, June 2018.

[21] A. Howard, M. Sandler, G. Chu et al., "Searching for mobilenetv3," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1314–1324, Seoul, Korea, October 2019.

[22] X. Zhang, X. Zhou, M. Lin, and J. Sun, "Shufflenet: an extremely efficient convolutional neural network for mobile devices," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6848–6856, Salt Lake City, UT, USA, June 2018.

[23] N. Ma, X. Zhang, H. T. Zheng, and J. Sun, "Shufflenet v2: practical guidelines for efficient CNN architecture design," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 116–131, Munich, Germany, September 2018.