Image Recognition Methods Based on Convolutional Neural Networks

Weiqing Zhang

School of Computer and Network Security, Chengdu University of Technology, Chengdu, 610059, China
zhang.weiqing@student.zy.cdut.edu.cn

Abstract. In recent years, with the rapid development of artificial intelligence technology and deep learning technology, convolutional neural network has made a major breakthrough in the field of image recognition, and has become the mainstream method in this field. In this paper, the image recognition methods based on convolutional neural networks are reviewed systematically through three analytical dimensions. Firstly, The most important structures of convolutional neural networks and their functions are introduced, there are three main parts and they include these parts such as convolutional layer(Extract feature), pooling layer(reducing the dimension of sampling features) and fully connected layer(integrate the key features). Subsequently it elaborates the models and algorithms used by CNN in some common image recognition tasks in different scenarios such as text recognition, portrait recognition, and product detection and so on. Finally, this paper summarizes the advantages of CNN application at this stage and the future development trend of convolutional neural networks in image recognition is prospected.

Keywords: Convolutional Neural Networks, Image Recognition, Product Detection.

1 Introduction

In today's most popular era of digital information, images carry a huge amount of content. Image recognition, as the mainstream in the field of computer vision, understands and analyzes the content of images through computer technology, hoping that this technology can extract content from images and make accurate judgments like human beings. As an important part of the computer vision field in the new era, image recognition has shown key technical characteristics in education, medical treatment, manufacturing and other fields [1]. In education, such as automatic marking of papers to prevent cheating in exams. In medical aspects, such as early disease screening, drug packaging identification. In manufacturing, such as automated appearance inspection, logistics package sorting. Traditional image recognition technologies mainly rely on statistical methods, syntactic recognition methods and geometric transformation methods, which can only perform well in their specific tasks, but are insufficient in the face of complex scenes [2].

J. Iqbal (ed.), *Proceedings of the 2025 2nd International Conference on Mechanics, Electronics Engineering and Automation (ICMEEA 2025)*, Advances in Engineering Research 272,
https://doi.org/10.2991/978-94-6463-821-9_67

With the continuous development of deep learning technology, great breakthroughs have been brought to image recognition, especially in the application of convolutional neural networks. In deep learning, convolutional neural network is the core algorithm in image recognition. It can build a hierarchical processing mechanism simulating biological vision to automatically learn and extract feature information, thus improving the performance of image recognition [3]. It has been skillfully applied in many fields, and a large number of cases abound. For example, in order to improve the accuracy and efficiency of fruit recognition and reduce labor costs, fruit recognition based on deep learning has been applied in the market such as agricultural supply chains and retail markets. Compared with previous recognition by fruit surface characteristics, more advanced and accurate recognition technologies have emerged. This technology greatly improves the recognition accuracy by extracting the deep features of fruits [4].

However, convolutional neural networks have made great achievements in the field of image recognition, but there are still many challenges and problems in this field. For example, for convolutional neural network models, people have difficulties in understanding the internal logic of the model to make decisions. Deep network training requires a lot of data and calculation, and is easily affected by many factors in practical applications, such as lighting, occlusion and so on. Faced with the impact of these potential factors, researchers have also proposed some solutions, including network structure optimization, data enhancement, etc., thus promoting the application of CNN in image classification, target detection and other tasks.

In this paper, the image recognition technology based on convolutional neural network is reviewed, focusing on its structure, algorithms, applications, and future directions. First, the basic structure of convolutional neural network is introduced. Secondly, the relevant models and algorithms in the practical application of image recognition are described. Finally, the future development trend of convolutional neural network image recognition technology is prospected.

2 Convolutional neural networks

In deep learning, convolutional neural network, as the core algorithm in image recognition, is a feedforward neural network based on convolutional calculation. It contains multiple layers, including convolutional layer, pooling layer and fully connected layer. Image features can be extracted through the combination of different layers.

2.1 Convolution layer

The convolutional layer is the core of the convolutional neural network, and its function is to extract local features from the input image. In image processing, the convolutional layer carries out local operations on the input image with the help of the convolution kernel that can slip, so as to capture local features in the image, such as edges and textures [5]. The formula for its operation:

$$\mathrm{f}(i,j)=(I*K)(i,j)\sum_{m}\sum_{n}I_{(i-m,j-n)}\cdot K(m,n) \quad (1)$$

Where: f (i,j) is the pixel value on the feature map; I is the input image; K is the convolution kernel; i and j represent the row index and column index of the input image respectively, and represent the pixel position of the image currently processed; m and n represent the row index and column index of the convolution kernel respectively [3].

2.2 Pooling layer

Pooling layer reduces the dimension of sampling features, reduces the number of parameters and the amount of calculation, and thus improves the calculation efficiency [6]. Common pooling operations include maximum pooling and average pooling. The formula for maximum pooling is as follows:

$$p(i,j)=\max_{m,n}(f(i+m,j+n)) \quad (2)$$

Where P(i, j) is the pixel value of the feature map after pooling; f is the feature map of the convolutional layer output [3].

2.3 Fully connected layer

The function of the fully connected layer is to integrate the key features extracted from the previous layer, which is similar to the traditional neural network layer. After integrating the key features, the classification structure is output through Softmax function, and the Softmax formula is as follows:

$$\sigma(Z_i)=\frac{e^{z_i}}{\sum_j z_j} \quad (3)$$

Where is the input of category i; is the corresponding probability value [3].

3 Application cases of convolutional neural network image recognition

At present, convolutional neural networks have made remarkable achievements in the field of image recognition. This paper mainly from the text recognition, portrait recognition, product detection three aspects to introduce. In the field of character recognition, whether it is ancient literature or modern text, CNN can recognize various fonts and characters efficiently, so as to extract the target text. In the aspect of portrait recognition, convolutional neural networks have played a vital role, especially in the field of security monitoring, it has become an important tool for personnel identity verification and tracking, providing a strong help for the public security to solve crimes, and contributing to maintaining social stability. In terms of product testing, it has been used in the detection of strawberry malformations and other fields, which greatly reduces the cost of labor and improves the efficiency and accuracy of identification. The successful application in these fields shows its great potential.

3.1 The model and algorithm of convolutional neural network in character recognition

This paper will introduce the text recognition model and its algorithm architecture, which consists of four main layers, the first layer is the input layer, which is responsible for receiving the initial text image. The second layer is feature extraction layer, which uses CNN technology to extract image features. The third layer is the sequence modelling layer, which uses RNN technology to model the sequence of feature sequences. The last layer is the output layer, which consists of the full connection layer and the loss function layer to classify the text sequence and calculate the loss. The following illustration details the algorithm architecture, outlining its core components and their interactions. By using this architecture, end-to-end training can extract text content directly from the original image, thus improving the accuracy and robustness of recognition [7] (Fig. 1).

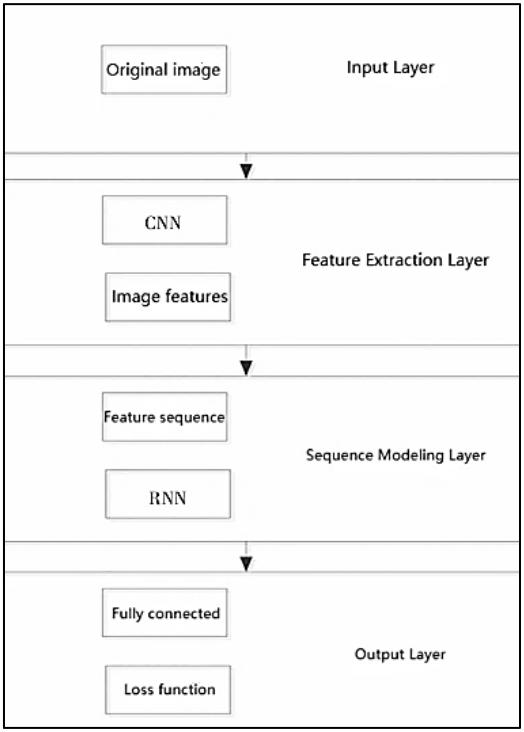

Fig. 1. Algorithm architecture of CNN in character recognition [7]

3.1.1 Initial text image input.

Assuming that the size of the initial image is A×B and the size of the scaled image is A'×B', the specific formula for the scaling process and proportion is:

$$S = \frac{A'}{A} = \frac{B'}{B} \quad (4)$$

In the process of character recognition, the color image will be converted to the gray image, which can reduce the complexity of calculation while preserving the text information. The specific formula is as follows.

$$\text{Gray}\, value = 0.299 \times R + 0.587 \times G + 0.114 \times B \quad (5)$$

Then the gray image is binarized, converted into black and white image, and then normalized:

$$p_{\text{gui}} = \frac{P_{initial}}{P_{\max}} \quad (6)$$

Where, the original pixel value is P_initial; the normalized pixel value is P_gui: The maximum pixel value is P_max.

3.1.2 Feature extraction.

The core part of feature extraction is the convolution layer, which filters localized regions of the input image using learnable convolutional kernels, enabling it to capture spatial hierarchies of feature information. The formula is as follows:

$$C_{\text{i},j} = \sum_{m=0}^{M-1}\sum_{n=0}^{N-1} I_{i+m,j+n} \cdot K_{m,n} + b \quad (7)$$

Where,C_(i,j) is a pixel value in the feature image after convolution;I_(i+m,j+n) is a pixel value in the input image;K_(m,n) is the weight of the convolution kernel; b is the offset term; M×N is the size of the convolution kernel. The maximum pooling operation can be used to reduce the spatial dimension of the feature graph by:

$$P_{\text{i},j} = \max(I_{2i,2j}, I_{2i,2j+1}, I_{2i+1,2j}, I_{2i+1,2j+1}) \quad (8)$$

P_(i,j) is a pixel value in the feature image after pooling; I_(2i,2j) is a pixel value in the input feature image. Due to the need to improve the nonlinear fitting ability of the model, it is necessary to introduce the activation function, whose formula is:

$$F(x) = \frac{1}{1+e^{-x}} \quad (9)$$

3.1.3 Sequence modelling.

In the process of sequence modeling, feature sequences {x1,x2,x3..... } will be entered into the RNN cell, and the sequence timing information will be captured using the hidden state of the RNN cell, for example, input x1 to get hidden state h1, input x2 to get hidden state h2, and finally get all hidden state sequences {h1,h2... }, each contains the input information of that moment and the information of the previous moment, hiding the definition of the status update:

$$\mathrm{h}_i = \tanh(W_{ih} \cdot x_i + W_{hh} \cdot h_{i-1} + b_h) \quad (10)$$

3.1.4 Input layer.

The full connection layer is responsible for mapping the sequence features of the output of the RNN module to specific text labels to realize the classification of text sequences. It takes the hidden state at the last moment of the RNN module or the hidden state sequence of the whole sequence as the input, through one or more fully connected layers, carries out linear and nonlinear transformations, and finally outputs vectors representing different text label probabilities. The vector is normalized by Softmax function to obtain the probability distribution of each label. The formula is as follows:

$$Soft\max(z)_i = \frac{e^{z_i}}{\sum_{j=1}^{N} e^{z_j}} \quad (11)$$

This formula computes the I-th element after normalization by the Soft-max function. The loss function measures its accuracy by calculating the loss between the model output and the real label. The cross entropy loss function is introduced here, and its formula is as follows:

$$Loss = -\sum_{i=1}^{N} y_i \log(p_i) \quad (12)$$

y_i is one-hot encoding for the real tag; p_i it the label probability distribution output for the model.

3.2 The model and algorithm of convolutional neural network in portrait recognition

This section mainly introduces the two-dimensional portrait recognition based on deep learning in photos taken under natural light. Its steps can be simply divided into two steps, the first step is to locate the face, and the second step is the face mark detection [8] (Fig. 2).

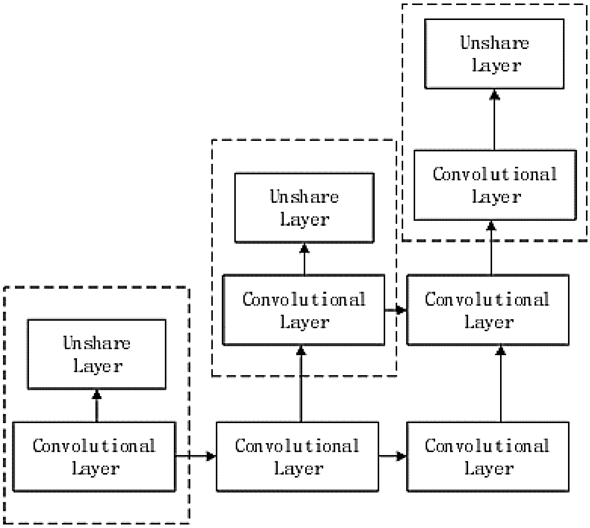

Fig. 2. The layered training method of CNN in portrait recognition [8]

The first-level network positions the face through a series of operations, such as rotation, scaling, etc., narrowing the calculation range for the second-level network to improve generalization. The role of the second level network is to reduce external factors, ineffective texture constraints such as facial expressions, and noise from ineffective texture constraints (face texture constraints refer to certain parts of the human face, such as the eyes, nose, and mouth). Algorithm: Divided into pre feedback and post feedback, pre feedback:

$$\begin{aligned} &L\,(s^1, s^2, s^3, \ldots, s^M \mid h_1^1, h_2^1, h_1^2, h_2^2, \ldots, h_1^M, h_2^M; \alpha \\ &= -\sum_{m=1}^{M} (s^m \log(g(h_m^1, h_2^m, \alpha)) + (1 - s^{m)} \log(1 - g(h_1^m, h_2^m, \alpha))) \end{aligned} \tag{13}$$

The training data set is comprised of M pairs of input instances along with their corresponding similarity labels denoted as S_m. For the m-th input instance,h_1^m represents the first sample, while h_2^m stands for the second sample.α is a positive coefficient which undergoes optimization during the learning process. The function g(x) serves as the cost function of the SIAMESE network.

$$\frac{\partial L}{\partial \alpha} = \sum_{m=1}^{M} \frac{s^m - g^m}{g^m(1 - g^m)} \left(\frac{\partial g(h_1^m, h_2^m, \alpha)}{\partial \alpha} \right) \tag{14}$$

$$\frac{\partial L}{\partial h_1^m} = \sum_{m=1}^{M} \frac{s^m - g^m}{g^m(1 - g^m)} \left(\frac{\partial g(h_1^m, h_2^m, \alpha)}{\partial h_1^m} \right) \tag{15}$$

$$\frac{\partial L}{\partial h_2^m} = \sum\nolimits_{m=1}^{M} \frac{s^m - g^m}{g^m(1-g^m)} \left(\frac{\partial g(h_1^m, h_2^m, \alpha)}{\partial h_2^m} \right) \tag{16}$$

$$\frac{\partial L}{\partial w} = \frac{\partial L}{\partial h_1^m} \frac{\partial h_1^m}{\partial w} + \frac{\partial L}{\partial h_2^m} \frac{\partial h_2^m}{\partial w} \tag{17}$$

w is the parameters,g^mis the output of m-th sample.

Using this function can make the output results of the same sample similar, so that the output results of different samples are significantly different, and all the results are normalized between 0 and 1, indicating the probability of similarity.

3.3 Convolutional neural network model and algorithm in product inspection

VGG16 is a convolutional neural network architecture, which has many effective applications in the image field, but there are obvious shortcomings in the interactive fusion of convolutional features at different scales. This section will introduce a model optimized based on VGG16 model, as shown in Fig. 3 [9].

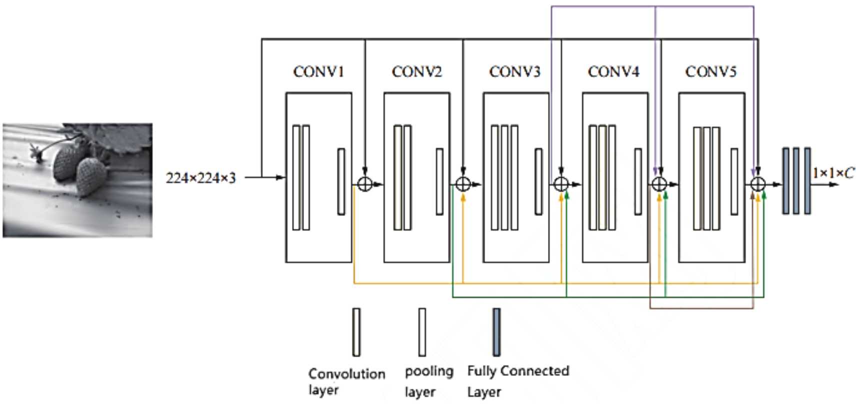

Fig. 3. Structure diagram of multi-scale convolutional neural network [9]

The model is composed of 5 CONV modules, CONV1 and CONV2 are composed of 2 convolution layers and 1 pooled layer, and the following 3 CONVs are composed of 3 convolution layers and 1 pooled layer. The advantage of this model is that the size of the convolution kernel is 3×3. Such small convolution kernel can make this model have better expressiveness and fewer parameters. The maximum pooling size is 2×2, which reduces the dimension of the feature graph and improves the computational efficiency. Helps models learn more complex mapping relationships, thereby improving the expressiveness of network models. At present, the classification of tea and the identification of diseases are mostly based on the working experience and professional knowledge of tea farmers. This troublesome and inefficient method is no longer suitable for this era. Therefore, in recent years, the detection model of tea insect pests based on CNN has been constantly proposed. There will introduce a model for detecting tea leaf diseases [10] (Fig. 4).

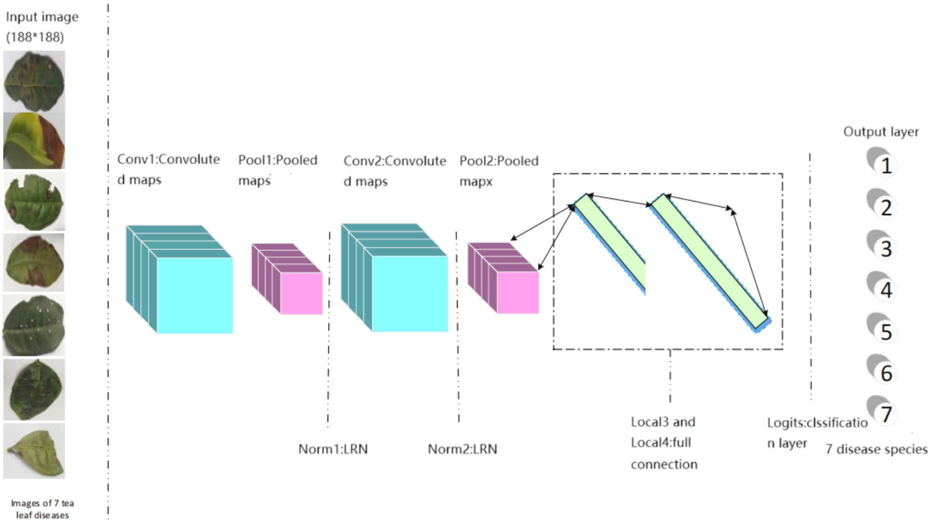

Fig. 4. Convolutional neural network structure for tea leaf diseases detection [10]

As shown in the figure, the dropout mechanism is added after the layer is fully connected, thus reducing over-fitting. The CNN model has nine layers, including two convolution layers, two maximum pooling layers, 2 partial response normalization layers, two fully connected layers and one classification output layer. The convolutional layer consists of multiple feature maps, each of which is composed of multiple neurons. Due to its special structure, the convolutional layer can extract the original features to the maximum extent. In the pooling layer, the model selects a maximum pool layer with a size of 3 x 3 and a step size of 2 x 2 to process the data, which increases the richness of the data due to the difference in step size and maximum pool size. At the same time, the use of local response normalization improves the generalization ability of the model. For activation functions without upper bounds, it can select a larger feedback from the convolution kernel, thus catalyze the data peak and improve the recognition efficiency. The role of the fully connected layer is to connect features and send them to the classifier. However, during this process, the network may over-fit, and at this time, the dropout layer will connect with the later fully connected layer, perform weight attenuation, and implement dropout by modifying the neural network structure.

4 Conclusion

Convolutional neural network has excellent ability of hierarchical feature extraction, which has become the core technology in the field of image recognition, and has made good achievements in character recognition, portrait recognition, and product detection. For example, in character recognition, CNN combined with RNN model can efficiently extract target text from relatively complex text. As mentioned in this paper, CNN and RNN are used for feature extraction and sequence modeling, respectively. This can deal with the variable font and text size, but also can have grayscale conversion and normalization methods to reduce the difficulty of calculation. In the aspect of portrait recognition, the multi-level training method enhances the robustness of illumination and occlusion. For example, SIAMESE is a good example, which not only reduces intra-class variation, but also improves inter-class differences, effectively improving the accuracy of face recognition. In product inspection, Tea leaf diseases detection model based on CNN and the optimization model of VGG16 also improves the computational efficiency in terms of improving the performance. However, CNN models still face challenges, including a lack of interpretability in decision-making processes and high demands for computational resources.

In the future, on the one hand, the performance and computational efficiency of the model can be improved by optimizing the network structure. On the other hand, the data is reinforced and advancing self-supervised learning, which can make the product adapt to more complex scenarios, thus improving the robustness of the model.

References

1. Xinzhi, Z.: 'Research on image recognition technology under the background of artificial intelligence', Journal of JiaMusi Vocational Institute, 2025, 266, (1), 148-150
2. Yuanzhu, C., Xiehua, Y.: 'Research on development of computer image recognition technology based on artificial intelligence', Journal of Zhaotong University, 2024, 46, (5), 76-82
3. Hua, W., Yanni, G.: 'Research and application of image recognition technology based on deep learning', Modern Computer, 2024, 30, (24), 63-72
4. Letao, Z., Dehong, L.: 'Optimization of image recognition algorithm based on deep learning', Technical Report, Guangdong Technical Normal University, 2024, 1-15
5. Chuyu, L., Xiaoqin, G., Jingjing, Y., Zhang, M.: 'Fruit image recognition based on deep learning', Journal of Chinese Agricultural Mechanization, 2025, 46, (1), 198-203
6. Liu, Z., Dezhi, L.: 'Application of convolutional neural networks in image classification', Technical Report, Beihai Vocational College, 2025, 1-12
7. Chuhao, L.: 'Research on high precision character recognition algorithm based on deep learning', Technical Report, CNOOC Information Technology Co., Ltd. Tianjin Branch, 2024, 1-20
8. Wang, W., Zhang, J., Li, S., Liu, Y.: 'Face Recognition Based on Deep Learning', Journal of Zhejiang University of Technology, 2015, 43, (3), 328-334
9. Yan, J., Li, P., Zhou, X., Ding, J., Wang, C., Lu, G.: 'Experimental Design of Strawberry Distortion Recognition Based on Multi-scale Convolutional Neural Network', Experiment Science and Technology, 2024, 22, (1), 1-5
10. Sun, X., Mu, S., Xu, Y., Cao, Z., Su, T.: 'Image Recognition of Tea Leaf Diseases Based on Convolutional Neural Network', Journal of Agricultural Engineering, 2019, 35, (4), 192-199