Proceedings of the 5th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/121/2025.19558

Exploration and Research of Convolutional Neural Networks in Image Recognition

Xican Song[1,a,*]

[1]*School of Data Science, Capital University of Economics and Business, No.121 Zhangjia Intersection Huajang, Fengtai District, Beijing, China*
a. julovo@qq.com
**corresponding author*

Abstract: In recent years, convolutional neural networks (CNNs) have achieved significant progress in the field of image recognition. However, their generalization capability and model complexity still require optimization in practical applications. This paper aims to further explore and optimize the performance of CNNs in image recognition tasks. First, the paper reviews the applications and techniques of CNNs in areas such as agricultural image recognition, animal recognition, smart city construction, and facial recognition. Second, the basic structure and training process of CNNs are introduced. A corresponding CNN model is constructed and trained using the MNIST handwritten dataset, achieving a test set accuracy of 98%, which demonstrates the effectiveness of CNNs in image recognition tasks. Finally, the feasibility and limitations of using advanced methods, such as residual networks (ResNet) and batch normalization, to address more complex image recognition problems are discussed.

Keywords: Convolutional Neural Networks (CNN), Image Recognition, Model Optimization, Generalization Capability.

1. Introduction

With the development of artificial intelligence and deep learning, image recognition technology has achieved significant advancements in various fields, including computer vision, autonomous driving, and medical imaging. As a critical deep learning model, convolutional neural networks (CNNs) have been widely applied to image recognition tasks due to their unique hierarchical feature extraction capabilities. By simulating the structure of the human visual system, CNNs automatically extract features from images in a layered manner, demonstrating exceptional performance in handling high-dimensional data. In recent years, CNNs have become one of the most prominent research focuses in the field of computer vision.

However, despite the remarkable success of CNNs in image recognition, several challenges remain in current methods. For instance, the requirements for feature extraction and model architecture vary significantly across different types of data and application scenarios, creating a trade-off between the generalization capability and complexity of CNNs in practical applications. Although existing research has achieved certain successes in improving image recognition accuracy, maintaining high precision while reducing computational costs remains an unresolved issue. Therefore, this study aims to further explore and optimize the performance of CNNs in image recognition tasks, with the goal of providing more efficient solutions for practical applications.

DOI: 10.54254/2755-2721/121/2025.19558

The structure of this paper is as follows: The first section provides a literature review of CNN applications in image recognition, summarizing key technologies and development trends. The second section introduces experimental validation and performance analysis of CNNs in image recognition, demonstrating their effectiveness through experiments and providing a detailed analysis. The third section discusses various advanced methods to address more complex image recognition problems, including techniques such as residual networks (ResNet) and batch normalization. The fourth section concludes by summarizing the research achievements of CNNs in image recognition and the main contributions of this paper. Additionally, it identifies the current challenges and issues in CNN research and application, proposing future research directions and prospects.

2. Literature Review

With the continuous development of deep learning technologies, convolutional neural networks (CNNs) have found widespread application and achieved significant results in the field of image recognition. Due to their outstanding performance in feature extraction and pattern recognition, CNNs have become a core technology in various fields, including computer vision, smart agriculture, and biometric identification.

In agricultural image recognition, CNNs have been applied to tasks such as fruit and vegetable recognition and pest detection. For example, He [1] designed and implemented a fruit and vegetable recognition and localization software based on CNN. By constructing a multi-layer convolutional neural network model that includes convolutional layers, activation layers, pooling layers, fully connected layers, and a Softmax output layer, the system achieved high-precision recognition and localization of fruit and vegetable images with an accuracy rate of 98%. The study utilized a large dataset of fruit and vegetable images for model training and applied optimization algorithms to enhance the model's generalization capability, contributing to the automation of agricultural production. Similarly, in invasive insect recognition, Huang et al. [2] collected a substantial dataset of invasive insect images, preprocessed the data, and trained four models—DenseNet121, MobileNetV3, ResNet101, and ShuffleNet—using the training dataset. The models' performance was evaluated using the test dataset. Experimental results indicated that the DenseNet121 model performed the best in recognizing invasive insects, achieving an average precision of 0.935, an average recall of 0.931, an average F1-score of 0.929, and an accuracy of 0.928. These results provide robust technical support for ecological protection and invasive species monitoring. Jiang et al. [3] addressed the challenges of high recognition difficulty and slow recognition speed in apple leaf disease identification. They proposed an improved convolutional neural network-based method for recognizing apple leaf diseases. The study enhanced the VGG16 model by introducing a Selective Kernel (SK) module and replacing the fully connected layer with a global average pooling layer to reduce model parameters and prevent overfitting. Additionally, the bottleneck layer was initialized using pre-trained VGG16 model parameters on the ImageNet dataset to improve the model's learning ability. The results showed that the improved VGG16 model achieved an accuracy of 96.17% on the test dataset, surpassing previous models in both accuracy and speed, thereby enhancing the precision and efficiency of disease recognition.

In the field of animal recognition, the feature extraction capabilities of CNNs have significantly improved the accuracy of image recognition. For instance, Liu [4] designed a CNN model for animal image recognition using the Keras framework, incorporating an input layer, multiple convolutional layers, pooling layers, and fully connected layers. The experiment utilized a large dataset of animal images for model training, and the trained model was then used for prediction. The results demonstrated that the designed model performed well in animal image recognition tasks, validating the effectiveness and feasibility of CNNs in this area. Additionally, CNNs have achieved progress in microalgae recognition and biomass prediction. Peng et al. [5] applied three deep convolutional neural

DOI: 10.54254/2755-2721/121/2025.19558

network models—ResNet, MobileNet, and EfficientNet—for these tasks. The study collected image data from three experimental microalgae species for model training and testing. The results indicated that all three models achieved classification accuracies exceeding 99% for the experimental species, with ResNet performing best in biomass prediction.

In the construction of smart cities, CNNs have also excelled in video surveillance and information monitoring. For example, Liu et al. [6] proposed a method for real-time observation of weather phenomena using widely deployed video surveillance equipment. The study pre-trained four deep neural network models—VGG16, Inception-ResNet-v2, EfficientNets-B7, and MaxViT-Base—and selected Inception-ResNet-v2 as the optimal model based on its performance on the test dataset. The model was further trained and tested using video image data from January to October 2022. The experimental results showed that the average accuracy for predictions at the Xujiahui weather station was 0.93 with an average F1-score of 0.74, while the Yangshan Port weather station achieved an average accuracy of 0.92 and an average F1-score of 0.67. The system demonstrated superior recognition performance for phenomena such as rain, light fog, and clear weather compared to manual recognition. In sensitive information monitoring, Pang and Wang [7] constructed a CNN model using the TensorFlow and Keras frameworks, training it on a dataset of 1.58 million images and combining it with a naive Bayes algorithm for sensitive information classification. The study's results showed that the system achieved a classification accuracy of over 95% for sensitive information and could analyze video data within a very short time, meeting the requirements for real-time monitoring.

Furthermore, CNNs have shown maturity in applications involving facial recognition and biometric feature extraction. Wang et al. [8] proposed an encryption system that converts facial features into binary strings for secure facial recognition authentication and access authorization. The system extracts and transforms feature vectors, using a feature vector transformer to differentiate between various facial images. Experimental results indicated that the difference in average similarity between paired images of the same person and paired images of different people was approximately 6% with the proposed method, compared to 23.5% with the original method, demonstrating its high accuracy. Liu et al. [9] developed a method using cubic spline functions as weight functions to extract and represent facial feature information. The study involved creating five spline weight function neural networks and comparing them with corresponding cubic spline weight functions. The experiment used 10 groups of facial images, increasing from 10 to 100 images per group for model training, and tested the models with the same facial image to observe the changes in absolute error values. Results showed that the cubic spline weight function-based models had smaller recognition errors for the same person, around 1.5%. Additionally, the more training images used, the smaller the absolute error value in facial recognition, effectively reducing the computational cost associated with traditional deep learning models. Research by Yan [10] and You [11] further demonstrated the application of CNNs in access control systems. These systems consist of modules for control, image acquisition, facial image processing, and access control execution. The facial image acquisition module first constructs a facial database, detects and locates facial images, and extracts features for training using a neural network model. Real-time facial image feature maps are then input into the trained network and compared with template data in the facial database for recognition. Techniques such as feature point localization and blink detection were incorporated to enhance the system's anti-spoofing capabilities and recognition accuracy. Hou et al. [12] explored CNN-based image recognition algorithms by constructing five different CNN models: LeNet, AlexNet, VGGNet, InceptionNet, and ResNet. These models were trained and tested using the Fashion-MNIST and CIFAR-10 standard datasets. The study evaluated the performance of the models in terms of recognition accuracy and cross-entropy loss, providing insights into their effectiveness in image recognition tasks.

DOI: 10.54254/2755-2721/121/2025.19558

3. Methodology

Convolutional Neural Networks (CNNs) are a type of deep learning model widely used in fields such as image recognition and natural language processing. By simulating the functioning of the human visual system, CNNs automatically extract features from images through multi-layered structures, enabling them to effectively process high-dimensional data. Their core components include convolutional layers, pooling layers, and fully connected layers.

Convolutional Layer: The fundamental building block of CNNs, the convolutional layer uses multiple kernels (filters) to perform convolution operations on input images, extracting local features such as edges and corners. Each kernel detects specific types of features. Pooling Layer: Usually placed after a convolutional layer, the pooling layer reduces the spatial dimensions of feature maps, thereby decreasing computational complexity and enhancing the model's translational invariance. Common pooling methods include max pooling and average pooling. Fully Connected Layer: Located at the end of the network, the fully connected layer integrates the extracted features and outputs the final classification results.

The CNN is trained using the backpropagation algorithm, where gradient descent is employed to optimize the model parameters, minimizing the error on the training data. Its hierarchical structure and local connectivity allow CNNs to automatically learn hierarchical feature representations, significantly improving accuracy and efficiency in image recognition tasks.

Convolutional Neural Networks (CNNs) excel in recognizing two-dimensional data (such as images) by extracting local features from images. The typical structure of a CNN generally includes convolutional layers, pooling layers, and fully connected layers. The convolutional layers are responsible for feature extraction, the pooling layers reduce dimensionality and enhance the robustness of features, and the fully connected layers are used for final classification.

A typical CNN structure consists of the following key components:

3.1. Convolutional Layer 1 (conv1)

- Input dimensions: 1×28×28 (single-channel grayscale image)
- Kernels: 16 kernels of size 5×5
- Stride: 1, Padding: 2
- Output dimensions: 16×28×28 (size preserved due to padding)

3.2. Pooling Layer 1

- Input dimensions: 16×28×28
- Pooling method: Max pooling, Pooling window: 2×2
- Output dimensions: 16×14×14

3.3. Convolutional Layer 2 (conv2)

- Input dimensions: 16×14×14
- Kernels: 32 kernels of size 5×5
- Stride: 1,Padding:2
- Output dimensions: 32×14×14

3.4. Pooling Layer 2

- Input dimensions: 32×14×14
- Pooling method: Max pooling, Pooling window: 2×2

DOI: 10.54254/2755-2721/121/2025.19558

- Output dimensions: 32×7×7

3.5. Fully Connected Layer

- Input dimensions: 32×7×7=1568
- Output dimensions: 10 (corresponding to 10 classes)

3.6. Training and Optimization

- Optimizer: Adam
- Loss Function: Cross-Entropy Loss
- Learning Rate: 0.001
- Batch Size: 50

CNNs can effectively extract multi-level features from images and map them to classification results. In this study, the model was trained on the MNIST handwritten digit dataset using the backpropagation algorithm to update weights.

4. Result

The research employed the PyTorch framework to construct and train the CNN model. The architecture of the model includes two convolutional layers, each followed by a ReLU activation function and a max pooling layer. The convolutional layers extract image features, the ReLU activation function introduces non-linearity, and the max pooling layers reduce the spatial dimensions of the feature maps while enhancing the model's translational invariance. Finally, the extracted features are classified using fully connected layers.

The experimental data came from the MNIST handwritten digit dataset, which contains 60,000 training samples and 10,000 testing samples. For this study, 60,000 samples were used for training, and 2,000 samples were used for testing. Each sample is a 28×28 grayscale image of a handwritten digit ranging from 0 to 9. These handwritten digits exhibit various styles, including different writing angles, stroke thicknesses, and cursive connections, providing a rich variety of patterns for model training and testing.

Figure 1: Examples of MNIST Handwritten Digits

DOI: 10.54254/2755-2721/121/2025.19558

On the test dataset, our CNN model demonstrated high recognition accuracy. After training, the model's test accuracy steadily improved, ultimately stabilizing at over 97%. Specifically, during the training process, the loss value of the model decreased progressively from an initial value of 2.3034 to 0.0188, while the test accuracy increased from 12% to 98%. These results indicate that our CNN model effectively extracted features from handwritten digit images and accurately performed classification.

The table below summarizes the recorded training loss (Train Loss) and test accuracy (Test Accuracy) during the training process:

Table 1: Training Record Table

Epoch	Batch Number	Train Loss	Test Accuracy
0	0	2.3034	0.12
0	1	0.4083	0.83
0	2	0.5555	0.87
0	3	0.2239	0.90
0	4	0.1272	0.93
0	5	0.1649	0.94
0	6	0.0559	0.95
0	7	0.1248	0.95
0	8	0.0344	0.96
0	9	0.1005	0.96
0	10	0.1840	0.97
0	11	0.0937	0.97
0	12	0.0188	0.97
0	13	0.1433	0.97
0	14	0.0942	0.97
0	15	0.2115	0.96
0	16	0.0398	0.96
0	17	0.0506	0.97
0	18	0.0998	0.98
0	19	0.0221	0.98
0	20	0.1979	0.98
0	21	0.0303	0.98
0	22	0.0221	0.98
0	23	0.0762	0.97

From the experimental records, it is evident that with the increase in training epochs, the training loss showed a decreasing trend, while the test accuracy steadily improved. However, this improvement was not strictly monotonic. Instead, it involved an initial phase of rapid optimization, followed by oscillations. In this experiment, it was observed that the test accuracy reached a relatively high peak of 94% at Batch Number 5. Thereafter, while the test accuracy continued to increase, the rate of improvement diminished, accompanied by minor oscillations. Importantly, there was no apparent overfitting observed in this experiment. Overfitting typically refers to a scenario where the model performs well on the training data but exhibits significantly poorer performance on the test data. This often occurs when the model is overly complex and begins capturing noise and minute details in the training data instead of learning the underlying patterns. In this study, although the test accuracy oscillated slightly after reaching its peak, it remained consistently high without significant

DOI: 10.54254/2755-2721/121/2025.19558

drops. This indicates that our CNN model successfully learned effective feature representations during training and achieved excellent generalization performance on the test dataset.

In summary, our CNN model exhibited outstanding performance in the MNIST handwritten digit recognition task, demonstrating its effectiveness and practicality in image recognition.

5. Discussion

In the experimental setup of this study, a subset of the MNIST handwritten digit dataset was used for training and testing. Specifically, 60,000 samples were used for training, and 2,000 samples were used for testing. Generally, the complexity of a model's structure is positively correlated with the number of parameters and the amount and quality of training samples required to achieve saturation. For the classification task addressed in this paper—a 10-class problem—the designed CNN model is of relatively low complexity. Taking this into account, we selected an appropriate number of samples for training and testing to avoid overfitting caused by an excessive number of samples. By controlling the volume of training samples, the model's generalization ability to unseen data can be improved, enhancing its practicality and robustness. Hence, the sample size used in this study was chosen to strike a balance between maintaining model performance and mitigating overfitting risks.

The CNN model designed in this study demonstrated excellent recognition performance on the MNIST handwritten digit dataset. However, generalization ability is a more critical concern. Generalization refers to the model's ability to maintain good performance on unseen data. For the recognition problem addressed in this study, the number of classes primarily affects the number of nodes in the fully connected layer. Specifically, the convolutional and pooling layers of a CNN are responsible for extracting features from images, which are the key information in image recognition tasks. Since the parameters and structures of the convolutional and pooling layers are independent of the specific classification task, the extracted features possess a degree of generality. When applying the model to a new classification task, only the parameters and structure of the fully connected layer need to be adjusted according to the task's specific requirements. Thus, to generalize the model to broader image recognition problems, only minor modifications to the fully connected layer are necessary, and higher-resolution sample images must be provided for tasks involving more categories.

In addition to the basic CNN framework, many advanced methods exist for handling more complex image recognition problems. One notable branch of research is residual networks (ResNet). By introducing residual structures, ResNet enhances the ability of CNNs to capture class-specific features in images. It highlights essential class information while preserving the visual features of the images, thereby improving the model's interpretability. Additionally, ResNet employs techniques such as Batch Normalization to accelerate the training process and improve the model's stability.

6. Conclusion

This paper explores the application of Convolutional Neural Networks (CNNs) in the field of image recognition and experimentally validates their effectiveness in handwritten digit recognition tasks. Through a review of recent research progress in CNN-based image recognition, we found that CNNs, with their unique hierarchical feature extraction capabilities, have achieved significant results across various domains. To evaluate the performance of CNNs in handwritten digit recognition tasks, experiments were conducted using the MNIST handwritten digit dataset. A CNN model comprising two convolutional layers, ReLU activation functions, max pooling layers, and fully connected layers was constructed and trained using the Adam optimizer and cross-entropy loss function, yielding satisfactory recognition results. Experimental findings demonstrated that our CNN model achieved a test accuracy of 98%, underscoring its outstanding performance in handwritten digit recognition tasks.

DOI: 10.54254/2755-2721/121/2025.19558

Throughout the experiments, we also discussed issues such as the rationale behind selecting a specific sample size, the model's generalization ability, and advanced methods for further improvement.

In conclusion, CNNs have exhibited remarkable performance and broad application prospects in the field of image recognition. Through the research and experimental validation in this paper, we have not only deepened our understanding of how CNNs function but also provided valuable insights for their optimization and improvement in practical applications. Looking ahead, with the continuous advancement of deep learning technologies and improvements in computer hardware capabilities, we have every reason to believe that CNNs will play an even greater role in a wider range of fields.

References

[1] *He, W. (2024). Design and implementation of fruit and vegetable recognition and localization software based on convolutional neural networks. Modern Information Technology, 16, 98–101+106. https://doi.org/10.19850/j.cnki.2096-4706.2024.16.021*

[2] *Huang, Y., Lu, L., Shen, H., Wang, F., & Qiao, X. (2024). Research on invasive insect recognition based on convolutional neural networks. Chinese Agricultural Mechanization Chemistry Journal, 07, 222–227+260. https://doi.org/10.13733/j.jcam.issn.2095-5553.2024.07.033*

[3] *Jiang, Y., Wang, J., Dong, G., & Hu, P. (2024). Apple leaf disease recognition based on an improved convolutional neural network. Jiangsu Agricultural Sciences, 14, 214–221. https://doi.org/10.15889/j.issn.1002-1302.2024.14.030*

[4] *Liu, X. (2024). Research on animal image recognition based on convolutional neural networks. High Technology and Industrialization, 07, 54–59.*

[5] *Peng, Y., Yao, S., Li, A., Xiong, F., Zhou, H., Gong, X., & Zhang, C. (2024). Microalgae recognition and biomass prediction based on convolutional neural network algorithms. Advances in New Energy, 04, 417–424.*

[6] *Liu, D., Shi, J., Yu, W., Wang, Y., Guo, W., & Du, M. (2024). Research and application of video-based weather phenomenon recognition. Journal of Solar Energy, 08, 441–447. https://doi.org/10.19912/j.0254-0096.tynxb.2023-0602*

[7] *Pang, G., & Wang, X. (2024). Design of a big data-sensitive information monitoring system based on convolutional neural networks. Scientific and Technological Innovation, 20, 113–116.*

[8] *Wang, D. (2024). Research on facial recognition models based on neural networks. Science and Technology Innovation and Application, 22, 5–8+13. https://doi.org/10.19981/j.CN23-1581/G3.2024.22.002*

[9] *Liu, M. (2024). Research on facial recognition based on spline weight function neural networks. Modern Information Technology, 04, 109–112. https://doi.org/10.19850/j.cnki.2096-4706.2024.04.023*

[10] *Yan, B. (2022). Implementation strategy for facial recognition based on convolutional neural networks. Network Security Technology and Application, 06, 47–49.*

[11] *You, C. (2024). Design and development of access control systems based on facial recognition technology. Information Systems Engineering, 08, 12–15.*

[12] *Hou, H., Wang, M., Meng, J., & Zhang, W. (2024). Research and implementation of convolutional neural network image recognition algorithms. Information and Computers (Theoretical Edition), 10, 94–96.*