Proceedings of CONF-SEML 2025 Symposium: Machine Learning Theory and Applications
DOI: 10.54254/2755-2721/2025.TJ24492

A Review of Convolutional Neural Networks: Evolution, Applications, and Future Directions

Aoran Wang

Mapleleaf School, Dalian, China
23010067@students.mapleleafedu.com

Abstract. The rise of deep learning, especially convolutional neural networks (CNNs), has greatly promoted the development of image recognition technology. CNNs imitate the human visual system and perform well in tasks such as image classification, detection, and segmentation by automatically learning features in images. This eliminates the need for manual feature extraction and improves accuracy. CNNs are composed of multiple layers, including convolutional layers, pooling layers, activation layers, batch normalization layers, and fully connected layers, each of which plays a vital role in extracting hierarchical features and processing complex patterns. This paper discusses the evolution of CNN models through a literature review, and the impetus in the evolution to improve the performance of image recognition, enabling deeper and more efficient networks. It also demonstrates widespread impact of CNNs through its applications in fields such as medical imaging, object detection, and face recognition. CNNs have revolutionized image recognition, providing more accurate and efficient systems. Continued advances in CNN architectures are expected to further enhance their ability to solve complex vision tasks, benefiting various industries and research fields.

Keywords: Convolutional Neural Networks, Image Recognition, Model Development, Medical Applications, Face Detection

1. Introduction

In today's digital age, visual information has become an essential medium for communication, significantly influencing fields such as healthcare, security, and autonomous systems. Compared to textual and auditory data, images provide a more intuitive and efficient means of conveying information. The exponential growth in the amount of image data has increased the complexity and demand for efficiently processing, searching, and analyzing images, and in the context of this challenge, deep learning, especially convolutional neural networks (CNNs), has emerged as a breakthrough solution in image recognition, which has further advanced image recognition technology [1]. Traditional image recognition methods rely on manually designed features combined with classical machine learning algorithms, such as Support Vector Machines (SVM) and decision trees. These approaches requires extensive domain expertise and is often struggle to generalize across diverse and complex image datasets. The introduction of CNNs has revolutionized image recognition by automating feature extraction, allowing models to learn hierarchical representations

© 2025 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2025.TJ24492

of images through multiple layers, including convolutional, pooling, and fully connected layers. By mimicking the human visual system, CNNs have demonstrated superior performance in tasks such as image classification, object detection, and segmentation, making them indispensable in modern computer vision applications [2].

However, the challenges faced by CNNs, such as high computational cost, reliance on large-scale labeled datasets, and limited interpretability, restrict their deployment in resource-constrained environments and safety-critical applications. This research highlights the evolution of CNN architectures by exploring their impact on image recognition, including LeNet-5, AlexNet, ResNet, and EfficientNet. It also examines CNN applications in domains such as medical imaging, object detection, and face recognition, emphasizing their practical significance. It aims to comprehensively summarise the role of CNNs in advancing the development of image recognition technology through investigation, and to provide feasible research directions for subsequent research.

2. CNN

CNNs are one of the best learning algorithms for understanding image content and and it performs exceptionally in tasks related to image segmentation, classification, detection and retrieval [3]. It is the core model in image recognition tasks, which mainly includes convolutional layer, pooling layer and fully connected layer. CNN effectively captures the spatial structure of the image through hierarchical feature extraction [4]. A mathematical method called convolution operation is used in extracting the image, which essentially two signals or functions to generate a new signal to reflect the influence of one signal on the other, and performs a weighted calculation based on their characteristics[5]. The formula is as follows:

$$y(t) = (f * g)(t) = \sum_{n=-\infty}^{\infty} f(n) \bullet g(t-n) \tag{1}$$

2.1. Pooring layer

Pooling Layer is an operation in convolutional neural network (CNN), which is mainly used to reduce the size of feature maps while retaining key features, improving computational efficiency, and reducing the risk of overfitting[6].

Pooling operation slides on the feature map through a fixed-size window (such as 2×2) and performs specific operations in each window. Common pooling methods include Max Pooling and Average Pooling. Max pooling takes the maximum value in the window and can retain the most significant features, while average pooling calculates the average of all values in the window and is suitable for smoothing feature maps.

This is the max pooling function:

$$Y(i,j) = m_{m=0}^{k-1} m_{n=0}^{k-1} X(i \times s + m, j \times s + n) \tag{2}$$

This is the average pooling function:

$$Y\left(i,j\right) = \frac{1}{k \times k} \sum_{m=0}^{k-1} \sum_{n=0}^{k-1} X\left(i \times s + m, j \times s + n\right) \tag{3}$$

DOI: 10.54254/2755-2721/2025.TJ24492

2.2. Activation function

In CNNs, the main role of activation layers is to introduce nonlinearity, allowing the network to learn and express complex feature relationships. Without activation layers, each layer of a convolutional neural network is essentially just a linear transformation, and the superposition of multiple linear layers is still a linear relationship, which cannot handle complex patterns and nonlinear features. By incorporating activation functions, neural networks can approximate any nonlinear function, allowing them to be applied to a broader range of nonlinear models[7]. In convolutional neural networks, the relu function is one of the most commonly used activation functions, and its formula is as follows:

$$ReLU(x) = max(0, x) \tag{4}$$

2.3. Batch Normalization layer and fully connected layer

Batch Normalization (BN) is a technology widely used in deep learning models, aiming to accelerate the training of neural networks, improve training stability, and alleviate the problem of gradient disappearance or gradient explosion. Its core idea is to standardize the inputs of each layer, so that the input distribution has a unified mean (0) and variance (1), thereby avoiding the instability of deep neural networks during training.

Furthermore, the fully connected layer (FC layer for short) is the most common type of layer in a neural network. Its function is to connect all neurons in the input layer to each neuron in the current layer.

3. The development of CNNs

3.1. Letnet-5

LeNet-5 is a classic convolutional neural network (CNN) structure proposed by Yann et al. It was developed based on earlier neural network architectures by optimizing weight-sharing and hierarchical feature extraction to improve computational efficiency and generalization ability [8]. LeNet-5 is specifically designed for the classification of grayscale images of size 28 × 28, mainly used for handwritten digit recognition. It consists of a total of seven layers, including two convolutional layers, two pooling layers, two fully connected layers, and a softmax regressor layer, allowing it to extract hierarchical features and perform classification [9].

Compared to fully connected networks, LeNet-5 significantly reduces the number of parameters by introducing convolutional and pooling layers, which helps prevent overfitting and improves computational efficiency. One of its key innovations is the use of local receptive fields and shared weights, which make it more efficient in recognizing spatial patterns [10].

LeNet-5 was the first commercially successful CNN, deployed in ATMs for recognizing handwritten digits on checks during deposit processing [3]. However, its limitations include its relatively shallow architecture, which restricts its ability to learn complex features from large-scale datasets. Subsequent CNN models, such as AlexNet, VGGNet, and ResNet, have significantly built upon LeNet-5 by incorporating deeper networks, better activation functions, and optimization techniques (e.g., batch normalization and dropout) [11-12].

DOI: 10.54254/2755-2721/2025.TJ24492

3.2. AlexNet

AlexNet is a deep convolutional neural network proposed by Krizhevsky et al. [12], and it achieved breakthrough success in the ImageNet image classification challenge, outperforming traditional methods by a significant margin. This proposal marked a leap forward in deep learning for computer vision, as it demonstrated the power of deep networks and GPUs in processing large-scale datasets, effectively laying the foundation for modern convolutional neural networks. AlexNet was proposed during a period when previous neural network models, such as LeNet-5, had limited success due to shallow architectures and computational constraints. The advent of AlexNet, which used a deeper and more complex network, was made possible by advancements in GPU technology and the availability of large labeled datasets like ImageNet.

AlexNet consists of eight layers of trainable neural networks, including five convolutional layers and three fully connected layers [12]. Unlike its predecessors, which used smaller networks with fewer layers, AlexNet employed deeper architectures that enabled it to learn more complex patterns from data. The network's input layer accepts color images of size 224×224×3, extracts features through multiple layers of convolution operations, and classifies through fully connected layers. In the first convolution layer, a larger 11×11 convolution kernel with a step size of 4 is used to extract primary features from the input image, and the data dimension is reduced by performing a maximum pooling operation. The second convolution layer uses a 5×5 convolution kernel to further process the features extracted in the first layer and performs maximum pooling again to reduce the amount of computation. The third, fourth, and fifth convolution layers use 3×3 convolution kernels to gradually extract more complex features, and finally obtain a feature map of 6×6×256 dimensions through maximum pooling.

Compared to previous models like LeNet-5, which had fewer layers and hardware limitions, AlexNet demonstrated how deep learning models could scale with more complex architectures and better computational power The key differences in AlexNet included the use of a deeper architecture, the use of ReLU activations instead of sigmoid or tanh to speed up training, and the use of dropout regularization to prevent overfitting, especially in the case of the large number of parameters. Furthermore, AlexNet used data augmentation techniques like image rotation and cropping to artificially increase the dataset size and improve generalization. These differences allowed AlexNet to outperform other models significantly in the ImageNet challenge, setting a new standard for deep learning architectures.

3.3. Resnet

He proposed a new deep learning model called Residual Network (ResNet) in 2015,. which solves the gradient vanishing and gradient exploding problems that are common in deep neural networks during training, allowing deeper neural networks to be successfully trained [13]. The core idea of ResNet is residual learning. Through skip connections or shortcut connections, the network can learn residuals, that is, each layer learns the difference between input and output instead of directly learning the output itself. The main purpose of doing this is to prevent the gradient from gradually disappearing or becoming unstable when the depth increases, while accelerating training convergence.

ResNet consists of multiple residual blocks, each of which contains multiple convolutional layers and passes the input directly to the next layer through skip connections. This structure allows the gradient to bypass multiple nonlinear transformation layers during backpropagation, thereby maintaining a stable gradient flow.

DOI: 10.54254/2755-2721/2025.TJ24492

3.4. EfficientNet

EfficientNet is an efficient convolutional neural network (CNN) architecture proposed by the Google Brain team in 2019 [14]. Its core concept is compound scaling, which optimizes the network's depth, width, and input resolution simultaneously using a compound coefficient. Unlike traditional CNN architectures that scale these factors independently, EfficientNet systematically balances them to enhance accuracy while reducing computational costs [14].

Compared to conventional CNNs such as ResNet and VGG, EfficientNet achieves better classification performance with significantly fewer computations, making it a highly efficient model for various computer vision tasks [15].

The main function of EfficientNet is to perform computer vision tasks such as image classification, object detection, and semantic segmentation. Due to its high computational efficiency, it is particularly suitable for mobile devices, embedded systems, and cloud inference, and can provide high-performance image recognition capabilities on limited hardware resources. In addition, based on EfficientNet, the EfficientDet object detection model has been developed for more complex visual tasks.

4. Application

4.1. Medical

The application of CNN in medical image recognition has become an indispensable tool in modern medical imaging, significantly facilitating early disease diagnosis and precise treatment [16]. By automatically learning features from medical images, CNNs can detect, classify, segment, and analyze various lesion areas and tissue structures in medical images, thereby enhancing diagnostic efficiency and accuracy for clinicians [16].

In the context of lung nodule detection, CNNs are extensively employed to process Computed Tomography (CT) images for detecting lung nodules or tumors. Models such as ResNet or DenseNet are commonly utilized, with convolutional layers automatically extracting morphological features of lung nodules, allowing the detection of small nodules at an early stage. Cao et al. demonstrated that deep learning-based models could improve the accuracy of lung nodule detection by leveraging multi-scale feature extraction techniques [17].

For lung cancer screening, CNNs analyze CT images to identify lung nodules and determine their malignancy, aiding in early cancer detection. Wang et al. proposed a deep learning model that achieved high sensitivity and specificity in lung cancer classification, outperforming traditional radiological assessments [18].

4.2. Face detection

Object detection, which involves identifying and localizing objects within images, has been revolutionized by CNNs. Models like YOLO, R-CNN, and SSD are widely used in fields such as security, video surveillance, and vehicle detection [1]. These models help accurately detect and track objects in real-time, making them ideal for dynamic and complex environments.

Face detection, a specialized form of object detection, focuses on identifying faces within images. It addresses challenges like lighting variations, pose changes, and occlusion. Traditional methods, such as Haar cascades and HOG+SVM, relied on handcrafted features but were less effective under varying real-world conditions. With the advent of CNN-based methods, face detection performance

DOI: 10.54254/2755-2721/2025.TJ24492

has significantly improved, achieving higher accuracy and robustness. Models like MTCNN (Multi-task Cascaded Convolutional Networks) are popular due to their efficiency and ability to handle difficult conditions like partial occlusion, low lighting, and detect small faces in complex scenes[2,3,19]. RetinaFace also shows excellent performance in similar scenarios[20].

CNN-based face detection is widely applied in biometric authentication, security systems, and social media. Its use in real-time surveillance and identity verification, such as tracking individuals in crowded areas, demonstrates its real-world effectiveness. For instance, in security applications, CNN-based systems are capable of identifying suspects or missing persons in large crowds, providing a powerful tool for law enforcement and surveillance operations. The increasing availability of high-quality face detection models has transformed industries and society, making it an essential tool for modern technological applications

5. Conclusion

This paper explored the advancement of CNNs in image recognition technology. By introducing the essential components of CNN, such as convolutional layer, pooling layer, activation function, batch normalization, and fully connected layer, it elaborates on the important features that the network is automatically learning and handling complex visual patterns. In addition, an introduction to models like LeNet-5, AlexNet, ResNet, and EfficientNet analyses the evolution from traditional image processing to deep learning, where the optimisation and development of these models improves accuracy while enabling deeper training and making more efficient models possible. These models are widely applied in fields like medical image analysis, object detection, and face recognition.

However, this paper has some limitations. Firstly, it lacked empirical comparisons between CNN models in real-world applications. Secondly, the literature review focused primarily on well-established models, limiting the scope of exploration into newer architectures. Additionally, while we discussed CNN applications, we did not address niche uses in specific industries. Future work will focus on empirical validation, comparing various CNN models in real-world scenarios. And expanding the literature review to include recent advancements and niche applications will broaden the paper's scope. Additionally, the author will also investigate optimizing CNN models for real-time processing in resource-constrained environments, such as autonomous systems and remote healthcare applications.

References

[1] Maria, T., & Elias, D. (2025). A Comprehensive Survey of Deep Learning Approaches in Image Processing.
[2] Xia, Z., Limin, W., Yufei, Z., Xuming, H., Muhammet, D., & Milan, P. (2024). A review of convolutional neural networks in computer vision.
[3] Asifullah, K., Anabia, S., Umme, Z., & Aqsa Saeed, Q. (2020). A survey of the recent architectures of deep convolutional neural networks. https: //doi.org/https: //doi.org/10.1007/s10462-020-09825-6
[4] Ravi, D., & Sreeram, S. (2020). Convolutional Neural Networks: An Overview and Applications. Proceedings of the International Conference on Data Science and Engineering (ICDSE), 2020, 99-104.
[5] Moez, K. (2023). Convolutional Neural Networks: A Survey. https: //doi.org/https: //doi.org/10.3390/computers12080151
[6] Han, K., Wang, Y., Tian, Q., Guo, J., Xu, C., & Xu, C. (2020). GhostNet: More Features from Cheap Operations. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 1580-1589.
[7] Xu, B., Wang, N., Chen, T., & Li, M. (2020). Understanding and Improving Layer Normalization. Advances in Neural Information Processing Systems (NeurIPS), 33, 4381-4391. https: //proceedings.neurips.cc/paper/2020/hash/5f678aaddba3e41a98a2332ab377c83f-Abstract.html
[8] Liu, L., Jiang, H., He, P., Chen, W., Liu, X., Gao, J., & Han, J. (2021). On the Variance of the Adaptive Learning Rate and Beyond. International Conference on Learning Representations (ICLR). https: //openreview.net/forum?

DOI: 10.54254/2755-2721/2025.TJ24492

id=NgoTfdvGm4a
[9] LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), 2278-2324. https: //doi.org/10.1109/5.726791
[10] Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.
[11] Simard, P. Y., Steinkraus, D., & Platt, J. C. (2003). Best practices for convolutional neural networks applied to visual document analysis. Proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR), 958-963. https: //doi.org/10.1109/ICDAR.2003.1227801
[12] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems (NeurIPS), 25. https: //doi.org/10.1145/3065386
[13] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 770-778.
[14] Tan, M., & Le, Q. (2019). EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. International Conference on Machine Learning (ICML). arXiv: 1905.11946.
[15] Yadav, P., Kori, P., & Chhabra, A. (2023). Deep learning applications in medical image analysis: Recent advances and challenges. Computers in Biology and Medicine, 157, 106935. https: //doi.org/10.1016/j.compbiomed.2023.106935
[16] Cao, C., Liu, Y., Yang, Y., et al. (2023). Multi-scale feature fusion network for automated lung nodule detection in CT scans. Scientific Reports, 13(1), 4859. https: //doi.org/10.1038/s41598-023-32187-6
[17] Wang, J., Li, Z., Zhao, H., et al. (2022). Deep learning-based lung cancer classification using multi-modal imaging data. IEEE Journal of Biomedical and Health Informatics, 26(5), 2154–2163.
[18] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You Only Look Once: Unified, Real-Time Object Detection. IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
[19] Zhang, K., Zhang, Z., Li, Z., & Qiao, Y. (2016). Joint Face Detection and Alignment Using Multi-task Cascaded Convolutional Networks (MTCNN). IEEE Signal Processing Letters, 23(10), 1499-1503.
[20] Deng, J., Guo, J., Ververas, E., et al. (2020). RetinaFace: Single-stage Dense Face Localisation in the Wild. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).