Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 9
AIDML 2025

Understanding Convolutional Neural Networks for Image Classification

Jinghan Xu *

School of Bishop Garcia Diego High School, Santa Barbara, United States

* Corresponding Author Email: enty361003@gmail.com

Abstract. In an era where artificial intelligence is becoming increasingly globalized, the development of AI-based image processing has undergone a significant transformation compared to previous generations of mechanical algorithms. Whether compared to classical machine learning classification models or traditional image classification methods that do not involve machine learning, technologies based on Convolutional Neural Networks (CNNs) demonstrate notable advantages—including higher accuracy, stronger feature extraction capabilities, and broader cross-domain applicability. These benefits have opened new and viable pathways for the deployment of CNNs across various fields. However, no algorithm is perfect. With continued advancements in research, new challenges are constantly being discovered and addressed, contributing to the ongoing refinement of the CNN framework. Today, the concepts of CNNs and deep CNNs (DCNNs) have been widely integrated into a range of popular AI applications. This paper aims to provide a comprehensive review of classical CNN-based image classification models, examine the strengths and limitations present in various architectures, and further explore their potential future applications and developmental trajectories.

Keywords: Image classification, convolutional neural network, computer vision.

1. Introduction

The eyes are the windows to the human soul. Among the five human senses, eyes have been considered by many as the most important sense, reading books, judging objects and even experiencing the beauty of the world cannot be achieved without the eyes as a sense. And for machines or programs, image classification techniques are like their eyes. Lu, Dengsheng, and Qihao Weng (2007) et al. in their article “A survey of image classification methods and techniques for improving classification performance”, based on a study of image classification methods and techniques. Performance", they discussed the use of Remote-sensing classification as a technique to realize image classification based on experiments [1]. However, with the popularity of machine learning models in various programs, more and more image classification techniques are gradually using machine learning. Zhang Hao, et al. (2006) in their paper used some of the early machine learning models such as SVM and KNN and conducted an in-depth study and comparison of other possibilities of these models with a large amount of data [2].

As we know, the information that we human beings perceive will eventually be gathered to the brain for processing, so can machine programs also simulate this form of human work? Traditional image feature extraction algorithms focus more on manually setting this method has poor generalization ability and portability. Allowing computers to process images in a way similar to biological vision has long been a dream of researchers [3]. Scientists have been working on this vision since 1943 and proposed a mathematical model of neuronal activity for internal logical operations, the MP neuron model [4]. As time passed, more and more theories were proposed, and with the introduction of the back-propagation (BP) algorithm training model, which made Convolutional Neural Networks, a computerized neural network technique that uses a biological vision-like system, a reality [5]. Various theories on the idea of CNNs were then proposed one after another. This makes CNN gradually in computer vision technology in an indispensable important position. Whether it is the emerging face

recognition technology, the automatic driving technology that has been known in recent years, or various impact analyses in medicine, CNNs can be seen.

2. Early Convolutional Neural Networks (Classic CNNs)

Since the theory of Convolutional Neural Networks was first proposed by researchers, a variety of models have been constructed, some of which were groundbreaking at the time, and some of which, even though less popular, have been crucial to the subsequent research. Among the many CNN models that were built in the early days, we have selected the following representative early models used for image classification to provide a better understanding of the development and operation of these models.

2.1. LeNet-5

In 1998, LeCun et al. constructed a digital classification model based on the concept of Convolutional Neural Networks [6]. Their model was mainly used to classify images of different people's writing pictures, and its performance was comprehensively better than all the algorithms in the same field at that time. It was also the first time that the backpropagation algorithm was applied to the training of CNNs. The LeNet-5 model is the cornerstone of the development of deep learning and the inspiration for various models in the future. Figure 1 shows the network architecture of the LeNet-5 model.

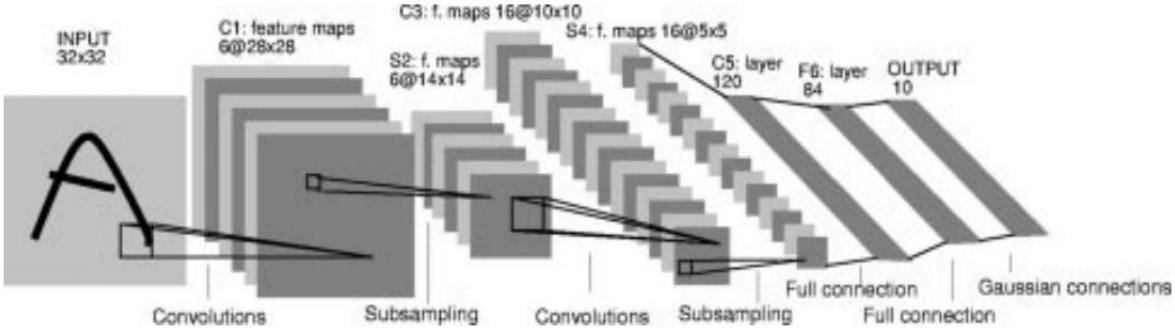

Fig 1. The architecture of the LeNet-5 network. The output shape is channel × height × width [7].

The LeNet-5 network is generally divided into two areas: convolution area and FC area. The basic unit of the convolution area is the convolution layer (Conv), which is followed by a maximum pooling layer (Pool). The convolution area is a repetitive stack of the basic units of the convolution layer and the maximum pooling layer. The fully-connected region is a fully-connected layer containing three neurons with a fixed number of 120, 84, and 10. When the output of the convolutional region is passed into the fully-connected region, the input layer of the fully-connected region flattens each feature map in the mini-batches and the length of the vectors in each mini-batch can be computed by using the formula.

The LeNet-5 model, as the earliest application for CNN training, has shown its unlimited potential, with very good results in the early handwritten digit recognition task (MNIST) dataset but also from various other experiments, such as poor performance on larger datasets, and very poor efficiency under the low level of hardware at that time, etc. This has motivated researchers to use the LeNet-5 model as the basis for CNN training. However, this has motivated the researchers to upgrade and reapply the ideas based on their model.

2.2. AlexNet

Alex et al. proposed the AlexNet in 2012, which won the championship in the ImageNet 2012 competition [7, 8]. AlexNet follows the basic idea of the LeNet-5 model and applies the basic principles of CNN to deep and wide networks [9]. However, AlexNet has also made significant upgrades based on the LeNet model, which are mainly reflected in following facts.

Firstly, it successfully uses ReLU as the activation function of CNN for the first time, thus alleviating the problem of gradient disappearance at the depth of the network; secondly, AlexNet uses the Dropout technique to randomly ignore some neurons in the process of training with the data set to avoid over fitting; and thirdly, it uses the Dropout technique to avoid over fitting by randomly ignoring some neurons during training with a data set. In addition, in the convolutional layer of AlexNet model, overlapping max pooling is used to replace the average pooling technique, which can avoid the ambiguous results of average pooling, and overlapping pooling can improve the richness of the features.

There are also other important technological upgrades in AlexNet, which are not listed here. AlexNet also has other important technological upgrades that are not listed here, and this model had a very outstanding performance in the ImageNet competition even under the hardware limitations at that time.

2.3. VGGNet

In 2014, Simonyan et al. developed the VGGNet model for participation in the ILSVRC competition, ultimately securing second place in ILSVRC 2014 [9,10]. The design philosophy of VGGNet is similar to that of AlexNet, adopting a structure composed of a convolutional feature extraction part followed by a fully connected classification part, and employing 3×3 convolution kernels throughout. The core idea of the architecture is to perform multiple convolution operations between two subsequent processes, with down-sampling achieved via max pooling. Finally, the remaining convolutional layers are connected to the output layer through fully connected layers. Since all convolutional and pooling layers use a stride of 1 and no padding, the spatial dimensions of the feature maps are gradually reduced—slightly blurred after each convolution and further compressed after each pooling operation [11]. VGGNet also exhibits a highly modular design, constructed from repeated basic building blocks, a concept that later inspired and guided the development of deep convolutional neural networks (DCNNs).

Despite being a relatively successful model in 2014, VGGNet is not without limitations, the most prominent of which is its high computational cost. With approximately 138 million parameters, VGGNet is an extremely large model (e.g., VGG-16 requires about 500MB of storage), making it difficult to deploy on mobile or resource-constrained platforms. VGGNet has been applied in various deep learning tasks, such as blind image quality prediction, where it outperformed contemporary traditional approaches and achieved remarkable results. However, in this task, VGGNet also exhibited overfitting issues, which adversely affected its overall performance.

3. Deep CNNs and Modern Advances

3.1. GoogLeNet

Also in 2014, the field of deep learning witnessed a milestone with the introduction of GoogLeNet, a model developed by Szegedy et al. [12]. This model integrated the ideas of Network in Network (NIN) and the theoretical work of Arora et al., and proposed the concept of the Inception module [13, 14]. Upon evaluation, the model caused a sensation at the ILSVRC competition. After training on the dataset, GoogLeNet achieved a top-5 error rate of only 6.67%, which was remarkably close to human-level performance—so much so that the organizers of the competition had to personally intervene to verify the results.

The core innovation of GoogLeNet lies in its introduction of the Inception module, which significantly boosts performance through multi-scale feature fusion and computational efficiency. An Inception module consists of four parallel branches, each executing convolutional operations under different receptive fields, and then unifies the outputs by merging the depth-wise information from different channels into a single layer.

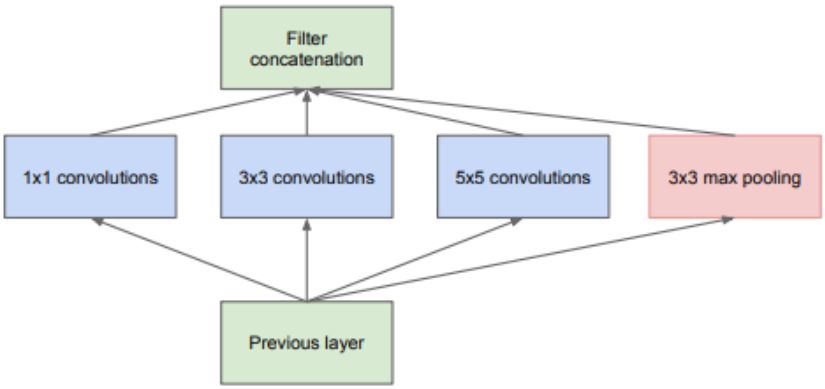

Fig. 2 Inception module, naive version [15].

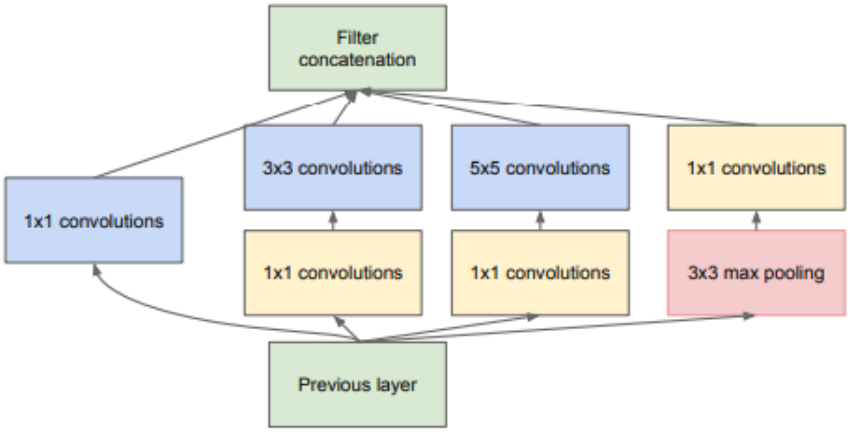

Fig. 3 Inception module with dimensionality reduction [15].

One key element worth emphasizing is the use of 1×1 convolutions within the architecture. This design reduces the number of channels and compresses information, thereby lowering the model's overall complexity. As a result, GoogLeNet contains only 6.5 million parameters—dramatically less than VGG-16's 138 million—which demonstrates the effectiveness of this architectural choice. The final GoogLeNet model is constructed by stacking multiple Inception modules, interleaved with several pooling layers, preceded by a few initial convolutional layers, and followed by several fully connected layers to form the output stage [16]. The innovative advantages embedded in GoogLeNet's design ultimately helped it win first place in the ILSVRC 2014 competition.

3.2. ResNet

Through the evolution of various models, it has become increasingly evident that the trend of deepening neural networks has become a common development pathway. However, as researchers delved deeper into practical applications, numerous studies have shown that blindly increasing network depth within a certain range does not necessarily lead to improved performance, and may even result in decreased classification accuracy [17, 18]. Driven by this realization, the Residual Neural Network (ResNet), proposed by Kaiming He et al., and made its debut in the ILSVRC 2015 competition [19].

At the core of ResNet lies the introduction of "skip connections", a concept that partially replaces the functionality of batch normalization. In essence, this mechanism allows the network to bypass non-linear layers, enabling it to learn residuals (i.e., error correction terms), thereby allowing gradients to flow directly back through the network. This effectively mitigates the vanishing gradient problem in deep architectures. Regardless of the complexity of the residual function, skip connections ensure that the original input can always be propagated forward, thereby maintaining stable gradient flow and making it feasible to train networks of arbitrary depth—breaking through the limitations of traditional deep models.

Leveraging this architecture, researchers successfully built extremely deep networks, such as ResNet-152, while maintaining lower complexity compared to VGGNet [20]. On the ILSVRC dataset, ResNet achieved an impressive top-5 error rate of just 3.57%, surpassing human-level performance on the same task. As a result, ResNet was recognized as a true deep convolutional neural network (DCNN) and secured first place in the ILSVRC 2015 competition.

3.3. DenseNet

The challenges of vanishing gradients and performance degradation in deep convolutional neural networks (DCNNs) were not exclusively addressed by the solution proposed in ResNet. Subsequent studies have continued to explore alternative approaches to mitigate these issues. One such notable advancement is the DenseNet model introduced by Huang et al. in 2017, which leveraged the revolutionary concept of Dense Connectivity and established itself as another landmark architecture in the realm of DCNNs [21].

Unlike traditional convolutional architectures, where connections occur only between adjacent layers, DenseNet employs Dense Connectivity, in which each layer is directly connected to every other subsequent layer in a feedforward fashion. This design allows every layer to receive gradient signals and feature information directly from both the loss function and the original input, forming a "feature map cascade". As a result, each layer can reuse features learned by all preceding layers, significantly reducing the need to learn redundant representations. This leads to a more compact, efficient model with fewer parameters while maintaining, or even improving, performance.

4. Applications of CNNs in Image Classification

After gaining a comprehensive understanding of various CNN models applied to computer vision, a natural question arises: in which domains closer to our daily lives can these models be practically utilized? While image classification is a fundamental task within computer vision, its real-world applications can greatly enhance efficiency in everyday scenarios and even drive the development of other technologies. Below, we outline several key areas where Convolutional Neural Networks (CNNs) can be effectively implemented in practice.

4.1. Medical Image Analysis

In the medical field, CNN-based image classification has demonstrated extensive applicability. For instance, in radiology, CNNs can be trained to classify X-ray images for the detection of diseases such as pneumonia. By identifying patterns within the images, the network can recognize signs of infection. Similarly, in mammography, CNNs can assist in classifying breast tissue images to detect early signs of breast cancer. These models analyze the texture, shape, and density of breast tissue to distinguish between normal and cancerous regions.

In pathology, CNNs can be employed to classify microscopic images of tissue samples, helping to identify various cell and tissue types. For example, CNNs can distinguish between benign and malignant tumors, aiding pathologists in making more accurate diagnoses and developing better treatment plans.

4.2. Autonomous driving technology

CNN-based image classification also plays a crucial role in the advancement of autonomous driving systems. These applications typically fall into two main categories: (1) Object Recognition in the Driving Environment: CNNs can process images captured by vehicle-mounted cameras to identify pedestrians, other vehicles, traffic signs, and road markings. This information is essential for decision-making processes, such as determining when to stop, turn, or change lanes. (2) Road Condition Analysis: CNNs can help vehicles interpret road conditions by classifying images of the driving surface. For instance, the system can detect hazards such as potholes or debris, providing critical information for safe navigation.

4.3. Security monitoring

Facial recognition, widely used in access control and criminal investigation, greatly benefits from CNN models. These networks can accurately identify and verify individuals based on facial features. In the context of suspicious object or behavior detection, CNNs can efficiently analyze surveillance footage to flag unusual activities or objects. This capability significantly enhances crime prevention efforts and public safety management.

5. Conclusion

In this survey, we have outlined the objectives of image classification tasks and systematically introduced a range of classic CNN and DCNN models developed between 1998 and 2015. Furthermore, we have explored several practical and feasible applications of CNNs in the domain of image classification. As a mathematical model that mimics the internal computational logic of biological neurons, CNNs have played a vital role across various modern disciplines. However, it is important to acknowledge that CNNs still hold significant potential for improvement. Over the years, research has revealed several limitations, including large model sizes, poor usability of hyperparameters, and the common trade-off between efficiency and accuracy in lightweight models. These challenges remain open issues for future exploration. Despite these drawbacks, CNNs continue to demonstrate outstanding—often even superhuman—performance in image classification tasks, reliably serving as the "eyes" of various applications. With continued advancements and optimization, we believe CNNs will assume even more critical and transformative roles in the future.

References

[1] Lu, Dengsheng, and Qihao Weng. "A survey of image classification methods and techniques for improving classification performance." International journal of Remote sensing 28.5 (2007): 823-870.

[2] Zhang, Hao, et al. "SVM-KNN: Discriminative nearest neighbor classification for visual category recognition." 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06). Vol. 2. IEEE, 2006.

[3] Chen, Leiyu, et al. "Review of image classification algorithms based on convolutional neural networks." Remote Sensing 13.22 (2021): 4712.

[4] McCulloch, Warren S., and Walter Pitts. "A logical calculus of the ideas immanent in nervous activity." The bulletin of mathematical biophysics 5 (1943): 115-133.

[5] Werbos, P.J. beyond Regression: New Tools for Prediction and Analysis in the Behavioral Science. Ph.D. Thesis, Harvard University, Cambridge, MA, USA, 1974.

[6] LeCun, Yann, et al. "Gradient-based learning applied to document recognition." Proceedings of the IEEE 86.11 (1998): 2278-2324.

[7] A. Krizhevsky, I. Sutskever, and G. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks,” Advances in neural information processing systems, vol. 25, no. 2, 2012.

[8] Li, Zewen, et al. "A survey of convolutional neural networks: analysis, applications, and prospects." IEEE transactions on neural networks and learning systems 33.12 (2021): 6999-7019.

[9] Chen, Leiyu, et al. "Review of image classification algorithms based on convolutional neural networks." Remote Sensing 13.22 (2021): 4712.

[10] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv: 1409.1556 (2014).

[11] Swapna, M., Yogesh Kumar Sharma, and B. M. G. Prasadh. "CNN Architectures: Alex Net, Le Net, VGG, Google Net, Res Net." Int. J. Recent Technol. Eng 8.6 (2020): 953-960.

[12] Szegedy, Christian, et al. "Going deeper with convolutions." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

[13] Swapna, M., Yogesh Kumar Sharma, and B. M. G. Prasadh. "CNN Architectures: Alex Net, Le Net, VGG, Google Net, Res Net." Int. J. Recent Technol. Eng 8.6 (2020): 953-960.

[14] Arora, Sanjeev, et al. "Provable bounds for learning some deep representations." International conference on machine learning. PMLR, 2014.

[15] Szegedy, Christian, et al. "Going deeper with convolutions." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

[16] Swapna, M., Yogesh Kumar Sharma, and B. M. G. Prasadh. "CNN Architectures: Alex Net, Le Net, VGG, Google Net, Res Net." Int. J. Recent Technol. Eng 8.6 (2020): 953-960.

[17] Wu, Zifeng, Chunhua Shen, and Anton Van Den Hengel. "Wider or deeper: Revisiting the resnet model for visual recognition." Pattern recognition 90 (2019): 119-133.

[18] He, Kaiming, et al. "Identity mappings in deep residual networks." Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, the Netherlands, October 11–14, 2016, Proceedings, Part IV 14. Springer International Publishing, 2016.

[19] He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778

[20] Swapna, M., Yogesh Kumar Sharma, and B. M. G. Prasadh. "CNN Architectures: Alex Net, Le Net, VGG, Google Net, Res Net." Int. J. Recent Technol. Eng 8.6 (2020): 953-960.

[21] Huang, GAO, et al. "Densely connected convolutional networks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.