Proceedings of the 4th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/48/20241257

Revolutionizing image recognition: The dominance and future potential of convolutional neural networks

Xingyu Ye

College of Electrical and Electronic Engineering, Wenzhou University, Wenzhou, 325000, China

21211741229@stu.wzu.edu.cn

Abstract. As AI technology advances swiftly and diverse industries increasingly require image processing, traditional image recognition methods are displaying their limitations. This paper explores the evolving landscape of AI technology in the context of image processing and highlights the limitations of traditional image recognition methods. With the proliferation of big data and the evolution of deep learning, convolutional neural networks (CNNs) have emerged as a dominant solution for image recognition across diverse industries. The paper begins by elucidating the architecture of CNNs and introduces commonly employed traditional CNN models. Furthermore, it offers practical insights into the application of CNNs within various industries, illuminating the path for future CNN development. The transformative potential of CNNs is underscored, as they possess the ability to extract intricate patterns from images, reshaping numerous domains. The paper's primary focus is on CNNs in the realm of image recognition, encompassing efforts to enhance precision and efficiency in CNN-based image recognition, as well as addressing real-world challenges in this domain using CNNs.

Keywords: Deep Learning, CNN, Image Recognition.

1. Introduction

As AI technology advances at an unprecedented pace and the demand for image processing grows across multiple industries, the constraints of conventional image recognition techniques have become increasingly conspicuous. These limitations stem from the inability of traditional approaches to efficiently handle the massive volumes of image data that are being generated, shared, and analyzed in today's digital landscape. Nevertheless, the advent of big data and the evolution of deep learning technologies have come to the forefront, establishing image recognition methods founded on convolutional neural networks (CNNs) as the prevailing algorithmic approach for image recognition [1]. Convolutional Neural Networks (CNNs) have brought about substantial advancements in the domain of computer vision, empowering machines to execute tasks with accuracy equivalent to or even surpassing human-level performance in certain instances. The capacity to autonomously acquire hierarchical features from raw data renders them an essential instrument in numerous practical applications.

In order to provide the understanding of the building blocks and principles that form the backbone of CNN technology, this paper commences by providing a comprehensive overview of the CNNs' fundamental structure which encompasses input layers, convolutional layers, activation layers, pooling layers, fully connected layers, output layers [1]. Then the paper delves into a comprehensive examination of several iconic CNN models including GoogLeNet, VGGNet, MobileNet, EfficientNet, and offers a valuable comparative analysis by dissecting their respective strengths, limitations, and applications [2-5]. Each of them renowned for its unique contributions and design elements. Furthermore, this paper places a significant emphasis on elucidating the contemporary landscape of CNN applications by giving several specific cases. In this section, it explores the multifaceted domains where CNNs have assumed a central role, highlighting their pivotal contributions in fields such as traffic transportation, face recognition, medical diagnosis, text extraction. In summary, the primary objective of this paper is to have broad-reaching implications, ranging from enhancing the precision and effectiveness of image recognition using CNNs to addressing diverse practical challenges in image recognition through the utilization of CNNs.

2. Structure of Convolutional Neural Networks

2.1. Definition

The Convolutional Neural Network (CNN) is a specialized artificial neural network designed primarily for processing and analyzing visual data, such as images and videos. CNNs draw inspiration from the human visual system, which hierarchically processes visual information, starting from basic features like edges and progressing to more complex objects. A typical CNN consists of various essential layers, including input, convolutional, activation, pooling, fully connected, and output layers [6].

Among these layers, the input layer receives raw image data organized as pixel value grids, often with color channels. Meanwhile, the output layer delivers final predictions or classification results, typically utilizing a softmax activation function [1]. Figure 1 provides a simplified representation of a standard CNN structure.

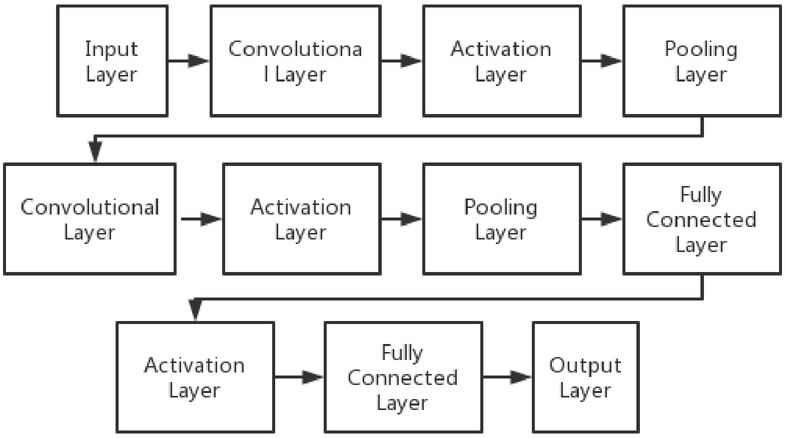

Figure 1. A simplified illustration of the structure of a CNN [1].

2.2. Convolutional Layer

CNNs use convolutional layers to scan the input image with small filters or kernels. These filters (one or more) perform a convolution operation.This process entails traversing the image incrementally, calculating the dot product between the filter and a small, overlapping region (patch) of the image at each step. Each filter generates its own feature map by convolving with the input image [6].

2.3. *Activation Layer*

Following the convolution operation, an activation function is applied element-wise to introduce non-linear characteristics into the model. This enables the network to capture and learn complex relationships in the data [6].

ReLU stands out as one of the widely embraced activation functions in neural networks. Its operation involves producing the input value when it's positive, and zero otherwise. ReLU expedites the learning process and is less prone to vanishing gradient problems. In mathematical terms, its definition can be expressed as in formula 1-3[7-9].

$$f(x) = \max(0, x) \tag{1}$$

The sigmoid function compresses input values within the range of 0 and 1. It is commonly employed in binary classification tasks. Its mathematical representation is:

$$f(x) = 1 / (1 + \exp(-x)) \tag{2}$$

The tanh function compresses input values within the interval of -1 to 1, making it suitable for various neural network architectures. Its mathematical expression is defined as:

$$f(x) = (\exp(x) - \exp(-x)) / (\exp(x) + \exp(-x)) \tag{3}$$

There are other activation functions like Leaky ReLU, Parametric ReLU (PReLU), Exponential Linear Unit (ELU), and Swish, each with its own characteristics and use cases [10-12].

2.4. *Pooling Layer*

Pooling layers are incorporated to diminish the spatial dimensions of the feature maps generated by the convolutional layers. Specifically, they serve to decrease the spatial dimensions while preserving crucial information. Common pooling operations encompass both max-pooling and average-pooling. The primary role of these layers is to alleviate computational complexity and instill translational invariance within the network [6].

2.5. *Fully Connected Layer*

After multiple convolutional and pooling layers, CNNs often include one or more fully connected layers. These layers transform the feature maps into a one-dimensional vector, then establish connections between all neurons in the layer. Fully connected layers are typically used for classification or regression tasks, making the final predictions based on the learned features [13].

3. Conventional Convolutional Neural Network Architectures for Image Recognition

3.1. *GoogLeNet*

GoogLeNet, alternatively recognized as the Inception architecture, represents an intricate and groundbreaking convolutional neural network (CNN) design originated by Google researchers. It distinguishes itself through its distinctive structure, enabling efficient extraction of multi-scale features. A notable feature of this architecture is its widespread incorporation of inception modules, purpose-built components tailored to adeptly capture features across diverse spatial scales.

The fundamental composition of the Inception Module comprises four key elements: 1*1 convolutions, 3*3 convolutions, 5*5 convolutions, and 3*3 max pooling. Ultimately, the outcomes of these four component operations are amalgamated across the channels. In practical scenarios, it's common to encounter 1*1 convolutions both preceding the 3*3 convolutions, 5*5 convolutions, and succeeding the 3*3 max pooling. This forms the central concept behind the Inception Module, which involves extracting information from images at various scales using multiple convolutional kernels and subsequently fusing them to attain a more robust representation of the image. Figure 2 illustrates the structure of the Inception Modules [2].

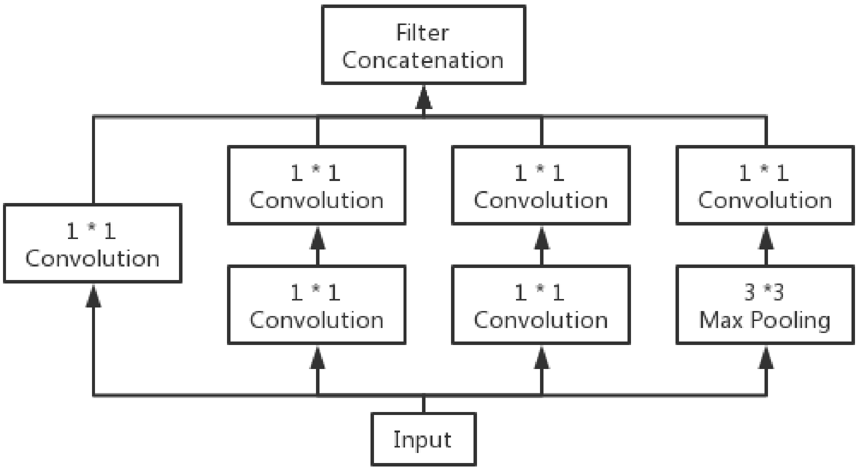

Figure 2. The structure of Inception Modules [2].

3.2. VGGNet

VGGNet, abbreviated from Visual Geometry Group Network, represents a convolutional neural network (CNN) architecture developed by the Visual Geometry Group situated at the University of Oxford. The structure of VGGNet and AlexNet is generally similar, but the VGGNet network has more layers. It is developed on the basis of AlexNet and can be seen as an improvement on AlexNet. VGGNet is renowned for its simplicity and effectiveness in image classification. The most well-known variants are VGG16 and VGG19. Both VGG16 and VGG19 architectures are known for their deep and uniform structure. Wherein, convolutional, and pooling layers are sequentially arranged, and they are succeeded by a fully connected layer.

Taking VGG19 as an example, it comprises a total of 19 weight layers, which encompass 16 convolutional layers and 3 fully connected layers. The architecture is organized into five sets of convolutional blocks, with three fully connected layers following them. Each convolutional block consists of multiple convolutional layers, with the first two blocks containing 2 convolutional layers each, while the remaining three blocks comprise 4 convolutional layers each. The depth of these five convolutional blocks is consistently set to 64, 128, 256, 512, and 512, respectively. All convolutional blocks share the same depth. The size of the convolutional kernel is fixed at 3*3, and max-pooling layers with a size of 2*2 and a stride of 2 are employed. You can observe the architectural layout of VGG19 in Figure 3 [3].

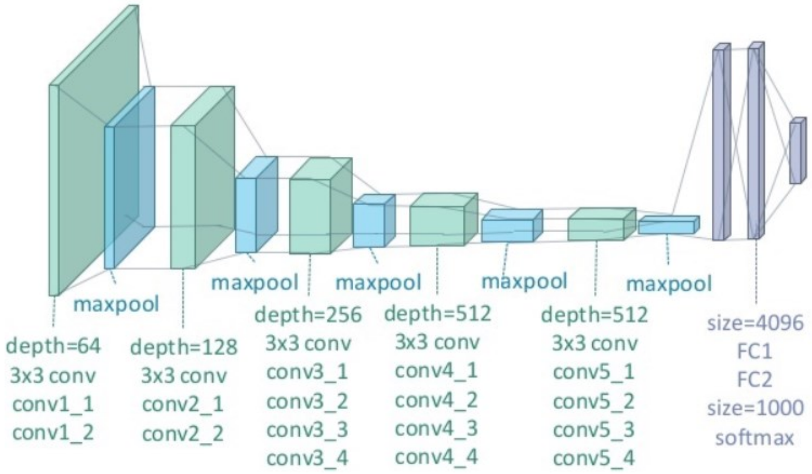

Figure 3. The structure of VGG19 [3].

3.3. MobileNet

MobileNet represents a convolutional neural network (CNN) architecture tailored for mobile devices and resource-constrained environments, with a primary emphasis on efficiency and speed. The central goal of MobileNet is to facilitate real-time and low-latency image classification and computer vision tasks on mobile devices, IoT devices, and other embedded systems.

MobileNet relies predominantly on depthwise separable convolutions, comprising two essential stages: depthwise convolutions and pointwise convolutions. Depthwise convolutions are applied individually to each input channel (depth) of the feature map. Instead of using a single 33 filter for all input channels, depthwise convolutions employ a 33 filter for each input channel. Following depthwise convolutions, pointwise convolutions, also referred to as 11 convolutions, are employed. They perform 11 convolutions to merge the output channels from depthwise convolutions.

MobileNet architectures exhibit variability in terms of depth, width (number of channels), and resolution, contingent upon the specific version. MobileNetV1 serves as the original version, while subsequent iterations like MobileNetV2 and MobileNetV3 introduce enhancements in efficiency and accuracy [4].

3.4. EfficientNet

EfficientNet was designed to provide a scalable and efficient solution for various computer vision tasks while maintaining high accuracy. The EfficientNet network consists of multiple MBConv modules. MBConv includes deep separable convolution, Swish activation function, SE attention mechanism, and Dropout layer. At the heart of MBConv, deep separable convolution plays a pivotal role in substantially reducing the parameter count and computational intricacy when contrasted with conventional convolution. Figure 4 illustrates the configuration of an MBConv (MobileNetV2 Building Block).

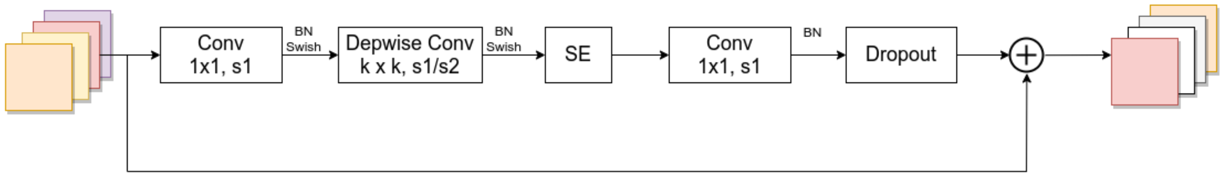

Figure 4. The structure of a MBConv [14].

The primary breakthrough in EfficientNet lies in its implementation of a compound scaling approach, which uniformly adjusts three crucial dimensions: depth, width, and resolution. EfficientNet models are available in different version, all the models are scaled form baseline EfficientNet-B0 using different compound coefficient. It uses efficient building blocks, such as inverted residual blocks and SE blocks. The architectural layout of EfficientNet-B0 is depicted in Figure 5.

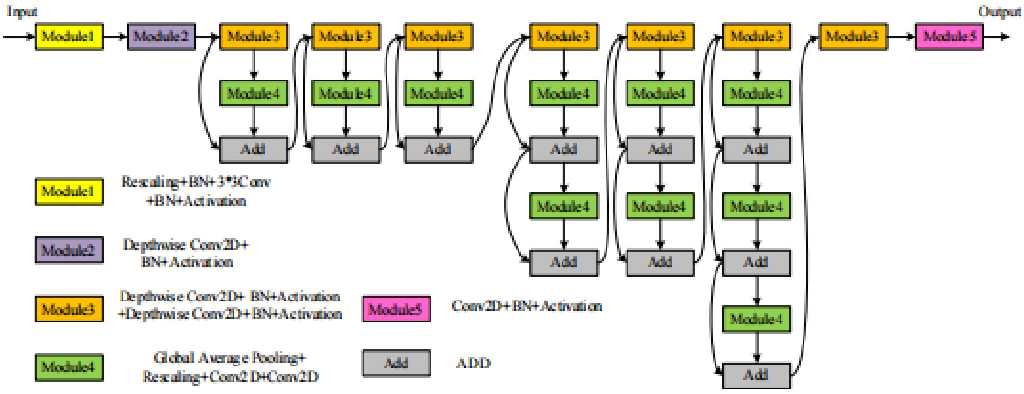

Figure 5. The structure of EfficientNet-B0 [14].

EfficientNet typically takes color images as input, often with a resolution of 224*224 pixels. Each convolutional block contains multiple inverted residual blocks and SE blocks. The depth, width, and resolution of these blocks are influenced by the composite scaling coefficients within the convolutional blocks. In contrast to employing fully connected layers at the conclusion, EfficientNet adopts Global Average Pooling (GAP). GAP calculates the average value for every feature map, yielding a 1D vector for each feature map [5].

3.5. Comparison

These convolutional neural network (CNN) architectures have been devised to cater to a range of computer vision assignments. Each architecture has its own characteristics and trade-offs. A comparative analysis of the four models is presented in the Table1.

Table 1. Four models comparing.

	Computational Efficiency	Accuracy	Computational Resources	Application
GoogLetNet	Moderate	High	Moderate to High	Various computer vision applications, including image classification and object detection
VGGNet	Relatively low	High	High	Image classification tasks, as feature extractors for transfer learning
MobileNet	High	High	Low	Real-time mobile applications, including image classification, object detection, and face recognition
EfficientNet	High	High	Varying	A broad range of computer vision tasks, from image classification to object detection and segmentation

The specific choice of architecture depends on the specific requirements of your task, available computational resources, and the trade-off between efficiency and accuracy. GoogLeNet and VGGNet, while accurate, may require substantial computational resources. MobileNet and EfficientNet offer efficient solutions, with MobileNet being highly optimized for mobile and embedded applications, while EfficientNet provides a flexible choice for various resource constraints.

4. Application of Convolutional Neural Networks in Image Recognition

4.1. Traffic Transportation

CNNs are used to process and analyze visual data, making transportation safer, more efficient, and more convenient for both individuals and communities. They contribute to the development of smart transportation systems that aim to reduce congestion, improve safety, and enhance overall mobility.

Pierre Sermanet et al used an architecture especially focusing on using multi-scale CNNs to detect and classify traffic signs at different scales in images. It deviates from traditional Convolutional Neural Networks by incorporating unique non-linearities, introducing skip connections between layers, and the utilizing pooling layers with varying subsampling ratios for both skip and non-skip connections. Pierre Sermanet and his colleagues introduced a convolutional neural network architecture that attained top-pier performance on the GTSRB traffic sign dataset, and this architecture was implemented using the EBLearn open-source library [15].

Jian Wang improved the YOLOv5, and later cascaded the improved EfficientNet network. This cascaded network structure uses an improved YOLOv5 network to detect traffic signs in street view images, perform simple rough classification, and return the specific position and size of traffic signs, crop out specific traffic signs, and input them into the improved EfficientNet network for detailed classification of traffic signs. The improved EfficientNet network is evaluated on the training, validation, and testing sets of the GTSRB dataset after fully learning the training samples. Comparing the recognition accuracy of the improved YOLOv5 and improved EfficientNet cascaded network with a single improved YOLOv5 network, the recognition accuracy has been improved by 4%, indicating a significant improvement. However, the cascaded network showed a slight decrease in detection speed. Table 2 displays the performance accuracy of the enhanced EfficientNet model to recognize GTSRB (German Traffic Sign Recognition Benchmark) data.

Table 2. The accuracy of the improved EfficientNet in GTSRB recognition [14].

Dataset	Training Sets	Validation Sets	Testing Sets
Total Number of Traffic Signs	57492	6388	12630
Number of Accurate Recognition	57394	6331	12499
Accuracy (%)	99.83	99.12	98.96

Meanwhile, Table 3 presents a model comparison [14].

Table 3. Model Comparison [14].

Model	Number of Detection	Number of Accurate Classifications	Accuracy (%)	Detection Time/ms
Improved YOLOv5	640	539	84.2	82
Improved YOLOv5 and EfficientNet cascaded	645	569	88.2	115

4.2. Face Recognition

Face recognition using Convolutional Neural Networks (CNNs) is a computer vision technique that enables automated identification and verification of individuals based on their facial features. In recent years, this technology has garnered substantial attention and recognition owing to its extensive array of applications, from enhancing security systems to enabling seamless user authentication on smartphones and other devices. Figure 6 illustrates a schematic representation of a facial recognition using Convolutional Neural Networks.

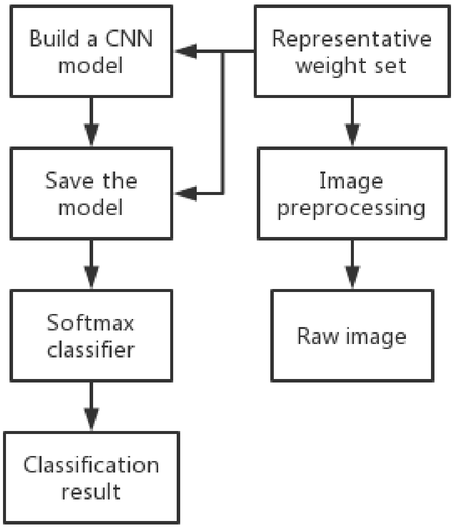

Figure 6. A schematic representation of a facial recognition using CNNs [16].

Florian Schroff and collaborators introduced a comprehensive system encompassing tasks like face verification (determining if two images depict the same person), face recognition (identifying individuals), and face clustering (grouping individuals with similar attributes from a set of faces). Their approach involves training a deep convolutional network to directly enhance facial embeddings, a departure from previous deep learning techniques that relied on an intermediate bottleneck layer. During their training procedure, they utilize triplets composed of well-aligned matching and non-matching facial patches. These triplets are created using an inventive online triplet mining technique. The noteworthy benefit of their approach is the substantial enhancement in representational efficiency. This is evidenced by their achievement of state-of-the-art face recognition performance while consuming just 128 bytes per face [16].

Chen Zhang applied the classic AlexNet convolutional neural network model to develop methods and applications for facial recognition. The model underwent training using the publicly available LFW face dataset, and the network's weights were finely tuned based on model evaluation criteria during the process. The experimental analysis yielded conclusive evidence that the AlexNet convolutional neural network model excels in performing facial recognition tasks effectively. This reinforces the practical relevance of convolutional neural networks in the domain of facial recognition and offers valuable guidance for implementing the AlexNet network model in facial recognition applications. The test result on AlexNet is shown in Table 4 [17].

Table 4. The test result on AlexNet [17].

Dataset	Accuracy (%)	Recall rate (%)	F1-Score(The closer to 1, the better)
Testing Set 1	97.05	96.13	0.96
Testing Set 2	96.89	96.05	0.94
Testing Set 3	96.92	95.99	0.95
Testing Set 4	97.01	96.27	0.95
Testing Set 5	96.91	96.01	0.95
Testing Set 6	96.95	96.29	0.95

4.3. Medical Diagnosis

In the field of medical diagnosis, Convolutional Neural Networks (CNNs) hold a pivotal position, harnessing their capacity to acquire and uncover intricate patterns and features from medical images and data.

Andre Esteva and his colleagues employed a Convolutional Neural Network (CNN) that achieves dermatologist-level classification of skin cancer through a combination of factors, including the architecture, training data, and evaluation methods. The researchers curated a large and diverse dataset of skin lesion images, comprising over 129,000 clinical images. The CNN employed in the study is founded on the Inception-v3 architecture, which underwent pretraining on an extensive dataset like ImageNet to acquire general features from a diverse set of images. Subsequently, the pretrained Inception-v3 model is fine-tuned on the skin lesion dataset for specialized learning. The architecture includes auxiliary classifiers at intermediate layers of the network. The ultimate layer of the network comprises a softmax classifier responsible for generating probability distributions pertaining to potential skin lesion categories. The CNN's performance is assessed in comparison to a group of dermatologists, and it is observed that the CNN's performance matches that of dermatologists in the classification of skin cancer [18].

Yuan Wang conducted research to enhance the capabilities of fully convolutional neural networks (FCNs) and applied these improved models to segment pancreatic organs in CT images and head and neck endangered organs. In the context of pancreatic organ segmentation in abdominal CT images, the researcher introduced a novel approach using a dual input V-mesh fully convolutional neural network (FCN). This innovative model incorporated two types of inputs: the original CT image and feature maps generated through the Contrast-Based Visual Significance Algorithm (CS-GBVS). These CS-GBVS-processed feature maps served to enhance soft tissue contrast in abdominal CT images and accentuate local differences. The V-mesh FCN architecture incorporated nested dense connections and attention mechanisms. Furthermore, the model introduced a Spatial Transformation and Fusion Module (SF) designed to capture geometric information associated with the pancreas and promote the fusion of feature maps. For the task of segmenting head and neck organs in CT images, Wang proposed a Feature Pyramid Fully Convolutional Neural Network. This network applied deconvolution operations to upsampled feature maps, allowing each layer to incorporate low-level features extracted earlier in the network. This process contributed to the restoration of detailed object features and the refinement of edges. Similar to the dual input V-mesh, an attention mechanism was employed in this network, combining output feature maps from the encoding stage with corresponding output feature maps from the feature pyramid module to extract relevant features of interest. To enrich features, the network included multi-scale feature fusion modules in its middle layers. The structure of the dual input V-mesh fully convolutional neural network is depicted in Figure 7, while a schematic representation of the feature pyramid fully convolutional neural network can be seen in Figure 8 [19].

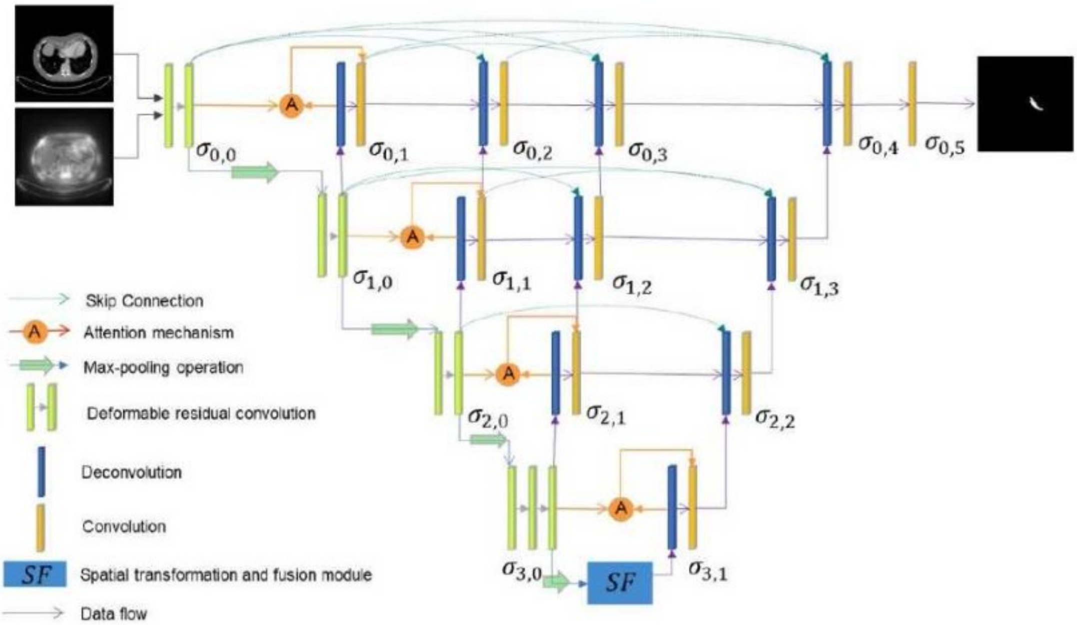

Figure 7. The structure of the dual input V-mesh fully convolutional neural network [19].

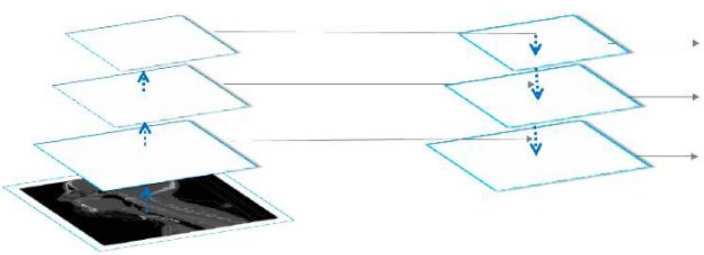

Figure 8. A schematic representation of the feature pyramid fully convolutional neural network [19].

4.4. Text Extraction

It's important to note that while CNNs can be effective for text extraction tasks, they are often used in combination with other techniques, such as recurrent neural networks (RNNs), attention mechanisms, and post-processing steps like language modeling, to improve accuracy and handle more complex text extraction scenarios.

Tian Zhi et al achieved the task of detecting text in natural images using a Connectionist Text Proposal Network (CTPN) that incorporates Convolutional Neural Networks (CNNs). The researchers employed a CNN-based feature extractor on top of the VGG-16 feature extractor to process input images. The CNN was used to generate text proposals or regions of interest (ROIs) within the image.After the initial CNN-based proposal generation, the authors employed a Bidirectional Long Short-Term Memory (BiLSTM) network to refine the text proposals. The CTPN is a key part of the system. It integrates the CNN-based proposal generation and the BiLSTM-based refinement into a unified network architecture. Following the CTPN's output, post-processing steps were applied to further refine and link the detected text regions into complete words or text lines. The final result is a set of text regions within the image. The CTPN demonstrates efficiency by setting new performance records on five benchmarks, all while maintaining a rapid processing time of just 0.14 seconds per image [20].

In order to improve recognition accuracy, Tingdong Wang introduced a dense convolutional network model structure based on convolutional neural networks and proposed a text recognition method based on dense convolutional networks. The convolutional neural network model takes as input (64*64) Chinese characters at the first level. It includes three identical dense modules, each followed by two transition layers. The initial image undergoes a 5*5 convolution operation followed by 2*2 max-pooling before entering the dense modules. Each dense module comprises 8 nodes, each incorporating a normalization layer, activation function layer (ReLU), convolution layer, and concatenation layer. Each convolution layer utilizes a kernel size of 8. To optimize parameter operation and mitigate overfitting, the model adopts a modified architecture inspired by the Inception network. Here, one 3*3 convolution operation from the original network is substituted with two separate 3*1 and 1*3 convolution operations. The transition layers consist of a normalization layer, a 1*1 convolutional layer (with the number of filters matching the input feature maps), and a pooling layer. This design ensures that the number of input and output features remains consistent within the transition layers.Finally, the network employs softmax for classification and utilizes a fully connected structure. The peak accuracy rate reaches 97.1%. The configuration of the DenseNet is visualized in Figure 9 [21].

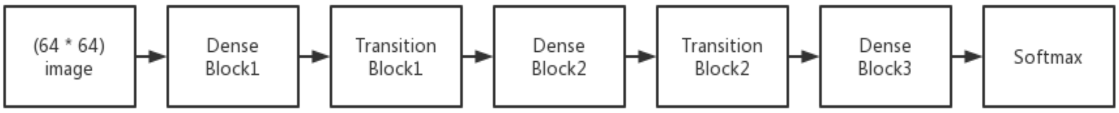

Figure 9. The configuration of the DenseNet [21].

5. Conclusion

In a nutshell, Convolutional Neural Networks (CNNs) have brought about a revolution in various domains by virtue of their capability to discern intricate patterns from images. In traffic transportation, CNNs are employed for vehicle detection and traffic flow analysis.In face recognition, CNNs excel at facial feature extraction, enabling accurate identity verification and security applications. In medical diagnosis, CNNs play a crucial role in image-based tasks such as identifying anomalies in medical images (X-rays, MRIs, etc.). For text extraction, CNNs are used to scan documents, extracting meaningful information from unstructured text. It is undoubted that CNNs are versatile image-recognition tools with applications spanning traffic management, biometrics, healthcare, and information extraction.

Nevertheless, CNNs still need enormous high-quality datasets for training which can be expensive and time-consuming, and they can inherit unfairness from their training data. They are computationally intensive as well, which means they often need powerful computers, but They might also learn too much from the training data and not work well on new data. On the other hand, understanding their decision-making processes also remains a challenge. These challenges necessitate to design an accurate and fast lightweight convolutional neural network model, while balancing power consumption and computational complexity. CNNs will continue to expand their role in solving complex problems and improving the way processing visual data, whereas they have made significant contribution in image recognition.

References

[1] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks. Communications of the ACM, 2017, 60(6): 84-90.

[2] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions. Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 1-9.

[3] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.

[4] Howard A G, Zhu M, Chen B, et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861, 2017.

[5] Tan M, Le Q. Efficientnet: Rethinking model scaling for convolutional neural networks. International conference on machine learning. PMLR, 2019: 6105-6114.

[6] LeCun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

[7] Nair V, Hinton G E. Rectified linear units improve restricted boltzmann machines. Proceedings of the 27th international conference on machine learning (ICML-10). 2010: 807-814.

[8] McCulloch W S, Pitts W. A logical calculus of the ideas immanent in nervous activity. The bulletin of mathematical biophysics, 1943, 5: 115-133.

[9] Domingos P. A few useful things to know about machine learning. Communications of the ACM, 2012, 55(10): 78-87.

[10] He K, Zhang X, Ren S, et al. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. Proceedings of the IEEE international conference on computer vision. 2015: 1026-1034.

[11] Clevert D A, Unterthiner T, Hochreiter S. Fast and accurate deep network learning by exponential linear units (elus). arXiv 2015. arXiv preprint arXiv:1511.07289, 2016, 2.

[12] Ramachandran P, Zoph B, Le Q V. Swish: a self-gated activation function. arXiv preprint arXiv:1710.05941, 2017, 7(1): 5.

[13] LeCun Y, Bengio Y, Hinton G. Deep learning. nature, 2015, 521(7553): 436-444.

[14] Jian Wang Research on Traffic Sign Detection and Recognition Based on Convolutional Neural Networks. Harbin Institute of Technology, 2022.

[15] Sermanet P, LeCun Y. Traffic sign recognition with multi-scale convolutional networks. The 2011 international joint conference on neural networks. IEEE, 2011: 2809-2813.

[16] Schroff F, Kalenichenko D, Philbin J. Facenet: A unified embedding for face recognition and clustering. Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 815-823.

[17] Chen Zhang. Face recognition method and application based on AlexNet convolutional neural network model. Journal of Ezhou University, 2022, 29 (01): 102-104.

[18] Esteva A, Kuprel B, Novoa R A, et al. Dermatologist-level classification of skin cancer with deep neural networks. nature, 2017, 542(7639): 115-118.

[19] Yuan Wang. Research on Improved Fully Convolutional Neural Networks and Their Applications in Medical Image Segmentation. Shandong Normal University, 2021.

[20] Tian Z, Huang W, He T, et al. Detecting text in natural image with connectionist text proposal network. Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14. Springer International Publishing, 2016: 56-72.

[21] Tingdong Wang. Research on Text Recognition Based on Dense Convolutional Networks. Science and Technology Innovation, 2021 (20): 89-90