Proceedings of CONF-MLA 2025 Symposium: Applied Artificial Intelligence Research
DOI: 10.54254/2755-2721/2025.BJ25790

Research on Image Recognition Model Based on Deep Learning

Jing Chen

Engineering College, Fuzhou Technology and Business University, Fuzhou, China
315048429@qq.com

Abstract. This paper reviews the evolution of deep learning-based image recognition models, focusing on key milestones in convolutional neural networks (CNNs), vision transformers (ViTs), and auxiliary techniques like data augmentation, regularization, transfer learning, and model compression. It traces the development of CNN architectures, including LeNet-5, and discusses their contributions to improving accuracy and addressing challenges like vanishing gradients. The paper also explores the transformative impact of transformers in image recognition, discussing models like ViT, Swin Transformer, and DeiT. It also analyzes critical supporting techniques to enhance model performance and practicality. The paper provides insights into the current state of image recognition, identifies open challenges, and outlines future research directions, aiming to inspire further innovation in theoretical foundations and real-world applications of deep learning for visual recognition.

Keywords: Deep learning, Image recognition, CNN, ResNet, VGGnet

1. Introduction

Image recognition, the process of enabling machines to interpret and understand visual information, is a fundamental problem in computer vision with wide-ranging applications,including medical diagnosis, autonomous driving, security surveillance, and e-commerce.Traditional image recognition methods, relying on handcrafted features such as SIFT(Scale-Invariant Feature Transform) and HOG(Histogram of Oriented Gradients), have achieved moderate success but often struggle with complex scenes, variations in lighting, and scale changes These limitations arise from the inability of manual feature engineering to capture hierarchical and abstract visual representations efficiently. The advent of deep learning, particularly convolutional neural networks(CNNs), revolutionized image recognition by automating feature learning through end-to-end training [1]. CNNs use hierarchical architectures to extract spatial features, pool layers for downsample, and classify learned representations. LeNet-5 has significantly improved accuracy on benchmark datasets like ImageNet. This paper reviews advancements in deep learning-driven image recognition methodologies, analyzing convolutional neural networks, vision transformers, and enabling technologies. It compares architectural innovations, evaluates trade-offs, and identifies promising research vectors.

© 2026 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.BJ25790

2. Evolution of Convolutional Neural Networks (CNNs)

2.1. Early CNN architectures: laying the foundation

The evolutionary trajectory of convolutional neural networks (CNNs) in computer vision commenced with Yann LeCun and colleagues' seminal work on LeNet-5, which empirically validated convolutional layers' efficacy in processing grid-structured data [2]. This architecture pioneered three cornerstone principles of modern CNN design: localized receptive fields through convolutional operations, parameter efficiency via weight sharing across spatial dimensions, and hierarchical feature abstraction using spatial pooling. LeNet-5's configuration - two convolutional modules followed by subsampling layers and three fully connected stages - achieved unprecedented accuracy on the MNIST handwritten digit recognition benchmark. While initially constrained to small-scale pattern recognition tasks, this framework established fundamental building blocks that would subsequently enable breakthroughs in complex visual processing challenges. Its modular design paradigm laid critical groundwork for subsequent architectural innovations, from AlexNet's deep stacking to contemporary vision transformers.

2.2. The AlexNet revolution: deepening the architecture

A transformative milestone occurred in 2012 with Krizhevsky et al.'s (2012) AlexNet, which revolutionized computer vision by securing first place in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) with a 15.3% top-5 error rate—nearly halving the performance gap from the runner-up system (26.2%). The architecture introduced four key innovations in deep learning: establishing the viability of deep convolutional architectures through eight weight layers, implementing parametric rectified linear units (ReLU) to solve the vanishing gradient problem, pioneering large-scale data augmentation pipelines with spatial transformations, and integrating dropout regularization for 98.7% validation accuracy. The success of AlexNet initiated the era of deep learning, triggered a paradigm shift in computational Vision research, and catalyzed the subsequent architectural revolution, from the hierarchical organization of VGGNet to the global attention mechanism of Vision Transformer.

2.3. VGGNet: exploring depth with uniform architecture

A seminal advancement in deep architecture design emerged in 2014 with Simonyan and Zisserman's VGGNet, which systematically demonstrated the performance benefits of architectural depth through standardized implementation. This framework pioneered a uniform architectural paradigm employing stacked 3×3 convolutional layers—smaller kernels that achieved superior parameter efficiency compared to larger filters (e.g., equivalent 5×5 receptive fields required 9 parameters using dual 3×3 convolutions versus 25 parameters for single 5×5 convolutions). The architecture's two variants, VGG-16 and VGG-19, achieved 92.7% top-5 ImageNet accuracy with 16 and 19 weight layers respectively, empirically validating depth augmentation through modular building blocks [3]. VGG-16's 138 million parameters result in high computational costs and low inference efficiency, leading to increased parameters and marginal precision gains. The architecture's emphasis on uniformity and simplicity influenced subsequent model compression strategies and efficiency-focused designs.

DOI: 10.54254/2755-2721/2025.BJ25790

2.4. GoogleNet (Inception): inception modules for efficient dimensionality reduction

GoogleNet (Inception-v1), developed by Szegedy et al., achieved a 6.7% top-5 error rate at ILSVRC 2014, surpassing human-level performance [4]. The architecture's core was the Inception module, which achieved superior parameter efficiency and expanded network width. It used parallel convolutional towers and parallel pooling operations for multi-scale feature extraction. The architecture's parameter efficiency, which surpassed VGG-16's 138 million, influenced subsequent efficiency-focused architectures like MobileNet and EfficientNet. The architecture's innovative design demonstrated the viability of compact deep networks.

2.5. Residual Networks (ResNet): overcoming the vanishing gradient in very deep networks

A fundamental limitation in training deep neural networks is the vanishing gradient phenomenon, where backpropagated gradients exponentially diminish through successive layers, impeding effective parameter updates.He,K. et al. addressed this challenge through Residual Networks (ResNet), introducing residual blocks with skip connections that bypass one or more convolutional layers. These architectural units implement identity mapping through a learnable residual function $F(x)=H(x)-x$,where $H(x)$ represents the target underlying transformation. This innovative design ensures gradient flow remains stable during backpropagation—even in extremely deep configurations—by mathematically enforcing $F(x)\rightarrow 0$ at convergence, thereby preserving gradient magnitude through direct skip pathways [5]. ResNet architecture achieved high ImageNet performance with a 3.57% top-5 error rate, enabling architectural innovations and effective training of networks exceeding 1,000 layers.

2.6. DenseNet: maximizing feature reuse with dense connections

Huang et al.'s DenseNet architecture revolutionized feature propagation by introducing explicit bidirectional information pathways within dense blocks [6]. This design creates a densely connected directed acyclic graph (DAG) with three efficiency gains: gradient flow optimization, feature map consolidation, and explicit feature compression. This architecture reduces parameter efficiency by reusing intermediate representations across 15.3M parameters. Empirical validation demonstrated DenseNet-BC achieving 7.8% top-5 error on ImageNet—a 3.5% improvement over ResNet-50—while maintaining comparable accuracy with 40% fewer parameters. Crucially, this dense connectivity paradigm became foundational for subsequent efficient architectures, influencing feature reuse patterns in EfficientNet's compound scaling and sparsity-inducing techniques in MoCo-X.

2.7. Squeeze-and-Excitation Networks (SENet): incorporating channel-wise attention

Hu, J's SENet revolutionized attention-based neural architecture design by introducing channel-wise attention mechanisms to model interdependencies between convolutional feature channels [7]. The SE module, which involves global average pooling across spatial dimensions and a two-layer multilayer perceptron, reweights channel-wise contributions, amplifying informative features while suppressing redundant ones. This end-to-end differentiable mechanism enhances network representational capacity without increasing complexity. The SE block's parameter efficiency allowed seamless integration into existing architectures, achieving state-of-the-art ImageNet performance with a 2.25% top-1 error reduction.

DOI: 10.54254/2755-2721/2025.BJ25790

3. Vision transformers: a paradigm shift in image recognition

3.1. Transformers in natural language processing: inspiring visual applications

Transformer architecture revolutionized sequence modeling with self-attention mechanisms, capturing long-range dependencies between tokens. It achieves global contextual modeling through additive self-attention mechanisms. This fundamental capability of modeling non-local relationships without progressive spatial restriction motivated the adaptation of Transformer architectures to visual recognition tasks, overcoming CNN's inherent limitations in modeling long-range spatial correlations.The Vision Transformer (ViT) framework subsequently demonstrated that such global attention mechanisms could effectively process image data when applied to patch-embedded visual tokens, achieving competitive performance on ImageNet classification while challenging the primacy of CNN-based paradigms.

3.2. Vision Transformer (ViT): treating images as patch sequences

A landmark advancement in vision-transformer architectures occurred with Dosovitskiy et al.'s Vision Transformer (ViT), the first transformer-based framework to achieve competitive performance in large-scale image classification. The architecture revolutionized visual processing by partitioning input images into fixed-size non-overlapping patches (e.g., 16×16 pixels), converting these visual fragments into sequence tokens, and processing them through standard transformer encoder layers [8]. This paradigm introduced two critical innovations: ViT, a vision transformer, demonstrated scalability and accuracy on large datasets, surpassing ResNet configurations. However, it incurred high computational costs and had architectural limitations. Future research uses hybrid architectures and data-efficient variants to overcome these constraints, establishing vision transformers as viable alternatives to convolutional paradigms [9].

3.3. Swin transformer: hierarchical design and localized self-attention

A significant advancement in vision transformer architectures occurred with Liu et al.'s Swin Transformer, which addressed ViT's scalability limitations through hierarchical feature pyramids and shifted window-based attention . The architecture introduces three key innovations: Swin Transformer uses a CNN-inspired hierarchical structure to reduce spatial resolution and expand channel dimensions, enabling multi-scale feature representation across four stages. It also employs shifted window attention to reduce computational complexity and maintain global context awareness. Swin-Tiny achieves 81.3% ImageNet top-1 accuracy with 28M parameters and 4.5G FLOPs. This hybrid design paradigm (transformer's global context + CNN's hierarchical induction biases) subsequently became the foundation for vision transformer scalability, demonstrating state-of-the-art performance across object detection (COCO AP 58.0) and semantic segmentation (ADE20K mIoU 53.5).

3.4. DeiT: data-efficient training through knowledge distillation

Touvron, H. et al. introduced DeiT (Data-efficient Image Transformers), an innovative architecture that addresses the data inefficiency challenge in vision transformer training through three key innovations: The DeiT framework uses CNN-based teachers to guide transformer students, reduces transformer dependence on large models, and achieves state-of-the-art results on ImageNet, bridging the data efficiency gap between convolutional networks and vision transformers [10].

DOI: 10.54254/2755-2721/2025.BJ25790

4. Auxiliary techniques for model enhancement

4.1. Data augmentation: expanding training data diversity

Data augmentation is a core technique for enhancing model generalization by expanding datasets with synthetic data. Shorten and Khoshgoftaar systematically categorized augmentation methods into three hierarchical levels: The text describes three main techniques for enhancing the visual quality of images: basic geometric transformations, panel-level adjustments, and advanced compositional strategies like cutout, mixup, and auto-augment, which are used for strategic region occlusion. These systematic interventions help neural architectures learn translation-invariant features while reducing overfitting, especially in scenarios with limited data [11]. Evolution in implementation paradigms demonstrates this progress: AlexNet first applied geometric augmentations like random cropping and horizontal flipping for ImageNet training, whereas modern architectures increasingly use compound augmentation pipelines to achieve robustness against pose variations, lighting inconsistencies, and partial occlusions in real-world applications.

4.2. Regularization: preventing overfitting

Dropout is a regularization method that enhances neural network robustness by randomly masking neurons during training. It addresses overfitting risks in high-dimensional parameter spaces. Modern regularization approaches often combine dropout with techniques like weight decay, batch normalization, and label smoothing to improve training stability and model generalization to unseen data distributions [12].

4.3. Transfer learning: leveraging pre-trained models

Transfer learning involves adapting a model pre-trained on a large dataset(e.g., ImageNet) to a target task with smaller datasets. Yosinski et al. studied the transferability of features in deep networks, showing that lower-layer features(e.g., edges, textures) are generic and transferable across tasks, while higher-layer features are task-specific [13]. Common transfer learning strategies include fine-tuning the entire pre-trained model on the target dataset or freezing the base layers and training only the top layers. This approach has become standard in practical applications, reducing training time and data requirements while achieving high performance in domains like medical imaging and remote sensing.

4.4. Model compression: enabling deployment on resource-constrained devices

Real-world deployment of deep models often requires reducing their size and computational cost. Han et al. proposed deep compression, which combines pruning(removing unimportant weights or neurons), quantization(reducing the precision of weights and activations), and Huffman coding to compress models significantly [14]. For example, AlexNet was compressed by 35x with minimal accuracy loss. Jacob et al. introduced quantization-aware training, which optimizes models during training to maintain accuracy when using integer arithmetic for inference crucial for mobile and edge devices [15]. Other methods include knowledge distillation(transferring knowledge from a large“teacher” model to a small“student” model) and neural architecture search(NAS), which automates the design of efficient architectures like MobileNet and ShuffleNet.

DOI: 10.54254/2755-2721/2025.BJ25790

5. Experimental analysis and performance comparison

The ImageNet dataset, containing 1.2 million images across 1,000 categories, is the most commonly used resource for image recognition. Key evaluation metrics include top-1 and top-5 classification accuracy, parameter count, FLOPs, and inference efficiency in table 1. Models trained on this dataset have achieved notable results, with CNNs like ResNet and DenseNet striking a balance between accuracy and efficiency. Transformers like ViT demand more parameters and FLOPs but achieve higher accuracy with large-scale training. GoogleNet and DenseNet are highly parameter-efficient, suitable for resource-constrained environments. Auxiliary techniques like data augmentation methods, transfer learning, and model compression can significantly impact accuracy and efficiency.

Table 1. Performance comparison on ImageNet

Model	Year	Top-1 Acc.	Top-5 Acc.	Parameters (M)	FLOPs (B)
AlexNet	2012	57.1%	80.3%	62.7	7.6
VGG-16	2014	71.5%	90.0%	138.3	15.3
GoogleNet	2015	69.8%	89.3%	6.9	1.5
ResNet-50	2016	76.0%	92.9%	25.6	4.1
DenseNet-121	2017	74.6%	91.8%	8.0	2.8
ViT (Large)	2021	84.5%	-	304.0	55.4
Swin Transformer (Base)	2021	83.3%	-	86.7	23.0

6. Conclusion

Deep learning has significantly impacted image recognition, with Convolutional Neural Networks (CNNs) and Transformers driving progress through structural innovations and supplementary methods. CNNs focus on local features and learning hierarchical representations, while Transformers excel at modeling long-range dependencies. Techniques like data augmentation, transfer learning, and model compression are crucial for real-world applications. Future research should address challenges related to efficiency, generalization, and interpretability, while exploring new territories like self-supervised training, multi-modal fusion, and bio-inspired architectures. Transformer architectures face challenges in generalizing to unseen data, computational efficiency, energy consumption, and interpretability. Future research should focus on efficient transformer architectures, self-supervised learning, multi-modal integration, biologically inspired models, and robustness and fairness assurance to reduce computational complexity, enhance generalization performance, and ensure ethical real-world deployments.

References

[1] LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), 2278-2324.
[2] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. NeurIPS, 1097-1105.
[3] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv: 1409.1556.
[4] Szegedy, C., et al. (2015). Going deeper with convolutions. CVPR, 1-9.
[5] He, K., et al. (2016). Deep residual learning for image recognition. CVPR, 770-778.

DOI: 10.54254/2755-2721/2025.BJ25790

[6] Huang, G., et al. (2017). Densely connected convolutional networks. CVPR, 4700-4708.
[7] Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. CVPR, 7132-7141.
[8] Dosovitskiy, A., et al. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. ICLR.
[9] Liu, Z., et al. (2021). Swin Transformer: Hierarchical vision transformer using shifted windows. ICCV, 10012-10022.
[10] Touvron, H., et al. (2021). Training data-efficient image transformers & distillation through attention. ICCV, 4002-4011.
[11] Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. Journal of Big Data, 6(1), 1-48.
[12] Srivastava, N., et al. (2014). Dropout: A simple way to prevent neural networks from overfitting. JMLR, 15(1), 1929-1958.
[13] Yosinski, J., et al. (2014). How transferable are features in deep neural networks? NeurIPS, 3320-3328.
[14] Han, S., et al. (2015). Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding. ICLR.
[15] Jacob, B., et al. (2018). Quantization and training of neural networks for efficient integer-arithmetic-only inference. CVPR, 2704-2713.