Automation and Machine Learning (2025)
Clausius Scientific Press, Canada

DOI: 10.23977/autml.2025.060110
ISSN 2516-5003 Vol. 6 Num. 1

Research on Optimization of Image Recognition Technology Based on Deep Learning

Yihe Zhang[1], Jixiang Bai[2], Jiayuan Wang[3], Qihao Zhou[4], Xianqu Feng[5]

[1]International College, Hebei University, Baoding, 071000, China
[2]College of Transportation Engineering, Dalian Maritime University, Dalian, 116000, China
[3]College of Information Engineering, Minzu University of China, Beijing, 100081, China
[4]School of Computer Science and Technology, Shenyang Institute of Engineering, Shenyang, 110136, China
[5]College of Science, Minzu University of China, Beijing, 100080, China

Keywords: Deep Learning; Image Recognition; Convolutional Neural Network; Multimodal Features; Algorithm Optimization

Abstract: Deep learning technology, as a core driving force in the field of artificial intelligence, has achieved breakthrough progress in image recognition in recent years. With the improvement of computational capabilities, the construction of large-scale datasets, and the innovation of deep neural network algorithms, image recognition technology based on deep learning has moved from laboratory research to practical applications, showing enormous potential in many fields such as medical diagnosis, security monitoring, and intelligent driving.

1. Introduction

In recent years, with the widespread availability of computing resources and continuous algorithmic innovation, deep learning technology has gradually moved from academic exploration to industrial practice, deploying on a large scale in fields such as smartphones, autonomous driving, and medical diagnosis, pushing image recognition technology toward the stage of universal promotion. Deep learning image recognition systems automatically extract image features through multi-layer neural networks, breaking the limitations of traditional methods that rely on manually designed features, and significantly improving recognition accuracy and generalization ability[1].

2. Basic Theory and Development of Deep Learning Image Recognition Technolog1y

2.1 Development History of Deep Learning Technology

The development of deep learning technology can be traced back to the perceptron model of the 1940s, but the real breakthrough began with the deep belief network proposed by Hinton in 2006, which solved the problem of training difficulties in deep networks. In 2012, AlexNet overwhelmingly defeated traditional methods in the ImageNet competition, marking a milestone breakthrough in deep learning in the field of image recognition. Subsequently, network architectures such as VGGNet,

GoogLeNet, and ResNet were successively introduced, continuously refreshing performance records for image recognition tasks. Between 2015 and 2018, the focus of deep learning development in image recognition shifted from improving accuracy to model lightweight and efficiency optimization, giving rise to efficient network structures such as MobileNet and ShuffleNet. After 2019, new paradigms such as self-supervised learning and contrastive learning emerged, with models like BERT, GPT, and CLIP demonstrating powerful transfer learning capabilities[2].

2.2 Core Principles of Image Recognition Technology

Deep learning image recognition relies on multi-layer neural networks to automatically learn hierarchical feature representations of images. Deep learning models extract low-level to high-level visual features directly from raw pixel data. In typical convolutional neural networks, shallow networks are responsible for extracting basic visual elements such as edges and textures, middle-layer networks combine these elements to construct component-level features, and deep networks further abstract semantic-level representations. Hierarchical feature learning gives deep models the ability to understand complex visual content. The end-to-end learning paradigm enables models to directly map from input images to output categories for full-process optimization. This approach avoids the artificial separation of feature extraction and classifier design. Backpropagation algorithms and large-scale data training prompt deep networks to continuously adjust internal parameters, ultimately forming specialized representations for specific visual tasks. Transfer learning and pre-trained models provide good initialization parameters for modern image recognition, with models pre-trained on large-scale datasets significantly improving recognition performance in small data scenarios[3].

3. Application Modes and Technical Dimensions of Deep Learning in Image Recognition

3.1 Multimodal Feature Extraction and Representation

Deep neural networks demonstrate excellent ability to capture image features, automatically learning hierarchical expressions from low-level vision to high-level semantics. Deep networks differ from traditional methods that rely on manually designed features; they flexibly extract the most valuable features for specific tasks. This ability is particularly important in complex scenarios. Modern models such as ResNet and EfficientNet further enhance feature extraction effects through innovative structures such as residual connections and attention mechanisms, enabling networks to quickly lock onto the most discriminative regions in images[4]. The role of multimodal features in real-time interactive applications is gradually becoming prominent, integrating various modal information such as visual, textual, and audio to give systems a more comprehensive understanding of scenes. Intelligent assistants and human-computer interaction platforms use multimodal models that combine image recognition and natural language processing to capture and interpret visual and language signals conveyed by users, providing more accurate feedback. Multimodal fusion continuously improves system robustness, expands the breadth of application scenarios, and deepens the depth of application scenarios. Self-supervised learning and contrastive learning, as important breakthroughs in recent years, provide new optimization paths for feature representation[5]. By designing pre-training tasks such as image rotation prediction and puzzle restoration, models can learn meaningful representations on unlabeled data. Contrastive learning frameworks such as SimCLR and MoCo construct more discriminative feature spaces by bringing similar sample representations closer and pushing different sample representations further apart, significantly improving downstream task performance, especially in scenarios with scarce labels[6].

3.2 Construction of Convolutional Neural Network Models

Convolutional Neural Networks (CNNs) play a core role in deep learning image recognition, using convolution operations to extract local features of images while using pooling layers to compress data dimensions and enhance translation invariance[7]. The structure of CNNs includes convolutional layers, pooling layers, fully connected layers, and normalization layers. These components work together to form an end-to-end system that automatically learns hierarchical features. Convolutional layers capture local patterns using learnable filters, pooling layers organize feature maps while preserving key information, and fully connected layers integrate high-level features to complete the final classification task. The hierarchical structure prompts CNNs to first read pixel data, then decompose information, extract low-level features, and gradually form high-level semantics. This process transforms raw pixels into specific semantics, ultimately providing solid support for image recognition tasks. Different CNN models have different focuses in structural design and performance characteristics. Classic AlexNet and VGGNet adopt simple stacking structures, which are conceptually clear but limited in depth; GoogLeNet introduces the Inception module, enhancing network expression ability through multi-scale feature extraction; ResNet solves the gradient vanishing problem of deep networks through residual connections, achieving training of networks with hundreds of layers; lightweight networks such as MobileNet and ShuffleNet use techniques such as depthwise separable convolution and channel shuffling to greatly reduce computational complexity while maintaining high accuracy, suitable for mobile device deployment[8]. These models have their advantages in different application scenarios, and choosing the appropriate network architecture requires consideration of factors such as task complexity, available computational resources, and real-time requirements[9].

4. Application Scenarios and Practical Cases of Deep Learning Image Recognition Technology

4.1 Applications of Image Recognition in Computer Vision

Object detection and recognition constitute key applications of deep learning image recognition technology, widely deployed in intelligent surveillance, autonomous driving, and retail analysis. Object detection algorithms have evolved from early R-CNN through the YOLO series to Transformer-based detectors, achieving simultaneous improvements in accuracy and speed. Modern object detection systems locate and identify multiple objects in real-time, maintaining high accuracy in complex scenarios. In smart city construction, deep learning-based vehicle detection systems simultaneously identify vehicle types, colors, and license plates, providing data support for traffic management. In the retail sector, product recognition systems automatically inventory shelf products, driving optimization of inventory management and shopping experiences. Scene segmentation and understanding technology advances image recognition to pixel-level precision, assigning semantic labels to each pixel for detailed scene analysis. Semantic segmentation, instance segmentation, and panoramic segmentation technologies show enormous potential in medical image analysis, remote sensing image processing, and augmented reality. Models such as DeepLab and Mask R-CNN significantly improve segmentation accuracy by fusing convolutional networks with attention mechanisms. In the autonomous driving field, scene segmentation technology precisely distinguishes roads, vehicles, and pedestrians, providing critical environmental information for decision-making systems. In agriculture, crops and weeds are finely segmented, enabling intelligent spraying systems to apply targeted pesticides, reducing chemical use and improving agricultural efficiency. Image enhancement and reconstruction technology uses deep learning models to improve image quality or recover information from damaged images, widely applied in medical imaging, security monitoring, and cultural heritage protection. Super-resolution reconstruction technologies such as SRGAN can

convert low-resolution images to high-definition images; denoising networks such as DnCNN can effectively remove image noise; image inpainting technology can fill missing areas in images. In the medical imaging field, deep learning enhancement technology can improve the clarity of CT and MRI images, helping doctors detect small lesions; in cultural relic protection, image reconstruction technology can repair damaged parts of ancient documents and artworks, providing technical support for cultural heritage; in the security field, low-light enhancement and deblurring technology greatly improve monitoring effects at night and in adverse weather conditions[10].

4.2 Industry Application Case Analysis

The industrial quality inspection and automation field demonstrates the key application value of deep learning image recognition technology, replacing traditional manual inspection to greatly improve production efficiency and product quality. In the electronics manufacturing industry, deep learning-based PCB board defect detection systems identify tiny solder joint defects and component misalignments, achieving detection accuracy exceeding 99% and detection speeds ten times faster than manual inspection. The automotive manufacturing field uses deep learning to monitor car body painting defects and assembly deviations, ensuring consistent product quality. The BMW Group's visual inspection system deployed in German factories monitors assembly line situations in real-time, processing over one million images daily. The food processing industry uses deep learning to quickly identify foreign objects, decay, and packaging defects, ensuring food safety. Large food companies use intelligent sorting systems combining multispectral imaging with deep learning algorithms to accurately eliminate unqualified products on high-speed production lines, improving qualification rates by 15% while reducing labor costs by 40%. Deep learning image recognition systems currently face technical reliability issues, which are more prominent in safety-critical applications. Research has found that deep learning models are extremely susceptible to adversarial samples, with small and visually imperceptible image perturbations causing models to make completely incorrect judgments. In autonomous driving scenarios, specific stickers placed on traffic signs disrupt the recognition process, creating safety hazards. When the test environment and training data distribution are inconsistent, model recognition accuracy drops significantly. This fragility limits the application of the technology in high-reliability scenarios such as medical diagnosis and critical infrastructure monitoring. Methods such as adversarial training, data augmentation, and uncertainty quantification are used to enhance model robustness. These methods aim to optimize model performance, but current results show they still fail to meet the requirements of industrial-grade reliability standards. Performance bottleneck analysis reveals the main limiting factors faced by current deep learning image recognition systems in actual deployment. High computational resource requirements are the most prominent bottleneck, with high-precision models typically requiring powerful GPU support, limiting applications on edge devices and mobile terminals. The inference of ResNet-101 on high-resolution images requires over 10 GFLOPS of computation, far exceeding the processing capabilities of ordinary mobile devices[11]. Large memory occupation is another key bottleneck, with model parameters and intermediate feature map storage requirements reaching hundreds of MB or even GB levels, challenging resource-constrained devices. Inference latency is also an important bottleneck for real-time applications, especially in scenarios such as video stream processing, where end-to-end latency needs to be controlled at the millisecond level to meet user experience requirements. In response to these bottlenecks, technologies such as model compression, hardware acceleration, and algorithm optimization continue to develop, but balancing performance and resource requirements remains an ongoing challenge. In recent years, emerging technologies such as self-supervised learning and few-shot learning have provided new ideas for solving environmental adaptability problems by reducing dependence on large amounts of labeled data and enhancing models' ability to

quickly adapt to new environments.

Network lightweight methods are key strategies for achieving sustainable development of deep learning image recognition technology, significantly reducing resource consumption while maintaining recognition performance by reducing model size and computational complexity. Model pruning technology can reduce 30%-90% of parameters with minimal accuracy loss by removing connections or neurons that contribute less to the output. After structured pruning, MobileNetV2 reduced parameters by 40% and improved inference speed by 35% on ImageNet with only a 1% accuracy loss. Knowledge distillation, as another important lightweight method, transfers knowledge from large models to small models through a "teacher-student" mode, enabling lightweight networks to achieve performance close to large models. Google's MobileNetV3 adopts a method combining automatic search with knowledge distillation, achieving excellent performance-efficiency balance on mobile devices. Technologies such as low-bit quantization and low-rank decomposition are also widely used in the network lightweight process, together forming a complete model optimization toolchain.

5. Conclusion

Deep learning image recognition technology is rapidly advancing toward self-supervised learning and few-shot learning, effectively reducing dependence on large-scale labeled data. Self-supervised learning designs pre-training tasks that prompt models to extract key information from unlabeled data and provide a foundation for subsequent tasks. Methods such as DINO and MAE show performance that can approach or even exceed traditional supervised learning in various visual tasks. Few-shot learning aims to quickly adapt to new tasks with limited samples. Technologies such as meta-learning and prototype networks give models the ability to learn how to learn, greatly enhancing the applicability and flexibility of the technology.

References

[1] Hu Yaoyu, Wang Yuming, Liu Chenyu, et al. Design of a commodity intelligent pricing system based on deep learning and image recognition technology[J]. Internet of Things Technologies, 2025, 15(07): 55-58. DOI: 10.16667/j.issn.2095-1302.2025.07.011.

[2] Ge Shilong, L üXinyue, Qu Mingshan, et al. Research on a nitrogen deficiency diagnosis model for strawberry leaves based on convolutional neural network[J/OL]. Vegetables, 1-9 [2025-04-03]. https://gfgfy2b08d79e045e4fd4hppvn9p6wpqqc6k9qficg.res.gxlib.org.cn/kcms/detail/11.2328.S.20250401.1529.002.html.

[3] Hou Wenhui, Guo Dandan, Zhou Chuanqi, et al. A method for identifying surface defects of Dangshan pears based on weakly supervised semantic segmentation[J/OL]. Transactions of the Chinese Society of Agricultural Engineering, 1-9 [2025-04-03]. https://gfgfy2b08d79e045e4fd4hppvn9p6wpqqc6k9qficg.res.gxlib.org.cn/kcms/detail/11.2047.s.20250331.1500.042.html.

[4] Jiang Yi, Jiang Yu, Wang Lei. A method for precise target recognition of mining robots using deep learning[J]. China Science and Technology Information, 2025, (07): 66-68.

[5] Du Zihuan, Gao Shuai. Artificial intelligence: A new driving force for exploring the development process of gametes and embryos[J/OL]. Scientia Sinica Vitae, 1-14 [2025-04-03]. https://gfgfy2b08d79e045e4fd4hppvn9p6wpqqc6k9qficg.res.gxlib.org.cn/kcms/detail/11.5840.Q.20250331.0909.002.html.

[6] Deng Xianghong, Yang Shuang, Long Tieguang. Image recognition of monkeypox disease based on a residual convolutional neural network model[J]. Science, Technology and Innovation, 2025, (06): 40-42 + 47. DOI: 10.15913/j.cnki.kjycx.2025.06.011.

[7] Lu Zhengzhi, Huang Xichen, Peng Bo. A review of adversarial attack technologies for single-object tracking[J/OL]. Computer Engineering and Applications, 1-16 [2025-04-03]. https://gfgfy2b08d79e045e4fd4hppvn9p6wpqqc6k9qficg.res.gxlib.org.cn/kcms/detail/11.2127.TP.20250326.1525.021.html.

[8] Hu Gaokai, Niu Yanan, Gong Yukang, et al. Research progress on the application of deep learning in the diagnosis,

surgical planning and postoperative prediction of lumbar diseases[J]. The Journal of Practical Medicine, 2025, 41(06): 921-928.
[9] Sun Tao, Wang Shuhao, Wang Wei. Application of edge computing in the field of pathological image recognition[J]. Chinese Journal of Medical Physics, 2025, 42(03): 328-335.
[10] Luo Xichang, Chen Hao, Zhang Yali, et al. A method for identifying snow accumulation on expressways based on traffic cameras[J]. Information Technology, 2025, (03): 35-41. DOI: 10.13274/j.cnki.hdzj.2025.03.006.
[11] Gou Lei, Zhang Wenjing, Zeng Mi, et al. Current application status of artificial intelligence in the diagnosis and treatment of aphasia[J/OL]. Guangdong Medical Journal, 2025, (03): 1-7 [2025-04-03]. https://gfgfy7789d7bbe5f747dbsppvn9p6wpqqc6k9qficg.res.gxlib.org.cn/10.13820/j.cnki.gdyx.20241631.