Proceedings of CONF-SPML 2026 Symposium: The 2nd Neural Computing and Applications Workshop 2025
DOI: 10.54254/2755-2721/2026.TJ29145

The Model Predicts the Category of the Input Image

Shibo Guo

Century College, Beijing University of Posts and Telecommunications, Beijing, China
shibo8777@outlook.com

Abstract. As a fundamental task in the field of computer vision, image classification technology holds substantial significance for practical applications in areas such as medical diagnosis, autonomous driving, and security monitoring. This paper presents a systematic review of the theoretical framework of image classification models, with a focus on analyzing the feature extraction mechanisms employed by mainstream architectures, including Convolutional Neural Networks (CNNs) and Transformers. Additionally, it explores how key technologies such as attention mechanisms and residual connections enhance model performance. Existing research primarily centers on three major directions. First, efforts are directed toward lightweight model design, aiming to reduce computational complexity while maintaining classification accuracy. Second, the development of multimodal fusion techniques enables the effective integration of auxiliary information, such as textual and audio data. Third, few-shot learning approaches are investigated to address challenges associated with limited training data. Research findings indicate that self-supervised learning-based pre-training paradigms significantly improve model generalization, while neural architecture search contributes to more automated model design. Nevertheless, current studies still face common challenges, including limited model interpretability and vulnerability to adversarial samples. Future research should emphasize innovations in biologically inspired network structures and focus on optimizing model deployment for edge computing environments, thereby expanding the practical applicability of image classification technologies.

Keywords: Image classification, Computer vision, Deep learning, Machine learning, Model prediction

1. Introduction

Image classification, as a fundamental task in computer vision, focuses on automatically identifying and labeling the categories of objects in images through algorithms. With the popularity of smartphones and digital devices, the amount of image data generated every day is growing explosively, and the traditional manual classification method can no longer meet the demand. Against this backdrop, automatic image classification technology based on deep learning emerged, becoming a key breakthrough in solving the problem of massive image processing. From medical image analysis to autonomous driving environment perception, from industrial quality inspection to

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

DOI: 10.54254/2755-2721/2026.TJ29145

social media content management, accurate and efficient image classification technology is profoundly changing the way modern society operates [1].

The breakthroughs achieved by deep learning models in the domain of image classification are primarily attributed to the strong visual feature extraction capabilities enabled by architectures such as convolutional neural networks (CNNs). For instance, the early LeNet-5 network demonstrated notable success in recognizing handwritten digits, achieving an accuracy rate of 98.8% on the Modified National Institute of Standards and Technology (MNIST) dataset [2]. Subsequently, AlexNet proposed by Krizhevsky et al. achieved a top-5 accuracy rate of 84.7% in the ImageNet competition, which was approximately 10 percentage points higher than traditional methods, marking a new era for computer vision [2,3]. The Visual Geometry Group Network (VGG) network optimizes parameter efficiency by stacking small-sized convolutional kernels, and the top-1 accuracy on ImageNet is improved to 72.7% [3]. The ResNet proposed by He et al. solves the problem of vanishing gradients in deep networks through residual connections, extends the network depth to more than a hundred layers, and achieves a top-1 accuracy of 92.1% on ImageNet, further promoting the learning of complex feature representations [4,5]. These technological advancements not only promote in-depth academic research but also give rise to application scenarios close to daily life, such as facial recognition payment and smart photo albums.

The core challenge currently faced by image classification research is to balance the accuracy of the model with the practical application requirements. Firstly, professional fields such as medical diagnosis require models to have discrimination capabilities close to those of human experts. For instance, the anemia screening system constructed by Hu's team needs to achieve an accuracy rate of over 95% in the classification of palpebral conjunctival images [5]. Secondly, real-time classification on mobile devices requires the model to operate efficiently in resource-constrained environments. For instance, the lightweight model designed by Wang Jianfang needs to keep the computing time consumption within 50ms [6-9]. In addition, the problem of insufficient samples caused by the high cost of data annotation and the security risks brought by adversarial sample attacks both restrict the further promotion of the technology. These real-life predicaments have prompted researchers to constantly explore innovative directions such as model lightweighting, small-shot learning, and enhanced robustness.

This paper aims to systematically sort out the theoretical basis, research status, and key technologies of image classification models, analyze existing challenges, and look forward to future directions, providing theoretical and practical references for the development of the field. Conducting research on image classification model prediction can help beginners sort out the domain context, provide references for the selection of scene algorithms, and also identify technical bottlenecks and promote practical applications.

2. Technical path

The application of deep learning in image classification tasks is mainly based on the theoretical foundation of CNNs. CNN effectively extracts the feature representations of images from the lower to the higher levels by simulating the hierarchical processing mechanism of the human visual system and adopting local connection and weight sharing methods. Its network structure is usually composed of convolutional layers, pooling layers, and fully connected layers. The convolutional layer is responsible for capturing local features, the pooling layer realizes feature dimensionality reduction, and the fully connected layer completes the final classification decision. The typical AlexNet has, for the first time, demonstrated the superiority of deep CNNs in complex image

DOI: 10.54254/2755-2721/2026.TJ29145

classification tasks. It adopts the ReLU activation function and Dropout technology, which improves the generalization performance of the model on the ImageNet dataset by 8% [7].

As research deepens, researchers have proposed various improvements to the CNN architecture to optimize the classification effect. The VGG network reduces the number of parameters by 40% by stacking multiple 3×3 convolutional kernels while keeping the receptive field unchanged [8]. GoogLeNet innovatively introduced the Inception module, which captures multi-scale features by using convolution kernels of different scales in parallel. The top-5 error rate on ImageNet was reduced to 6.67%. ResNet solves the problem of vanishing gradients in deep networks through residual connections, enabling the top-1 accuracy of the 152-layer network on ImageNet to reach 92.1%, which is 11.3% higher than that of the 18-layer network [9]. These structural evolutions not only enhance the classification accuracy but also provide important design paradigms for subsequent research.

In recent years, the Transformer structure has demonstrated great potential in the field of image classification. The Vision Transformer (ViT) segments the image into serialized image blocks, establishes global dependencies through the self-attention mechanism, breaks the limitation of the local receptive field of CNN, and achieves a top-1 accuracy rate of 88.5% on the ImageNet-1k dataset. It surpassed the 86.9% of ResNet-152 in the same period [10]. Meanwhile, researchers explored hybrid architectures of attention mechanisms and CNNS, such as the Squeeze-and-Excitation (SE) network, which dynamically adjusts feature weights through channel attention, thereby increasing the classification accuracy of the model on ImageNet by 2.3%.

At the theoretical level, all these models follow an end-to-end learning paradigm, that is, they directly learn the mapping relationship from the original pixel input to the category output. The loss function usually adopts cross-entropy loss and optimizes the network parameters through the backpropagation algorithm. To alleviate overfitting, in addition to L2 regularization, data augmentation techniques such as random clipping, rotation, and color transformation are widely adopted, which can enhance the generalization ability of the model on small sample datasets by 5% to 10% [11]. It is worth noting that since the rise of self-supervised learning, pre-training methods such as contrastive learning have significantly reduced their reliance on labeled data. For instance, the SimCLR model's general visual representations learned from unlabeled images have an accuracy rate 15% higher when transferred to downstream tasks compared to random initialization, opening up a new research direction for image classification technology [11-13].

3. Research status

3.1. Application of deep learning in image classification

Deep learning has become a core research direction in the field of image classification. CNNs, as the mainstream architecture, automatically extract image features through multi-layer convolution operations, significantly improving the classification accuracy. Wu Yuxin's research shows that after CNN combines image enhancement technology, its generalization ability on small sample medical image datasets has increased by 7.2% [13]. It has solved the pain point of difficult clinical data annotation. Typical models such as AlexNet, VGG, and ResNet gradually solve problems such as vanishing gradients and parameter redundancy by optimizing network depth and structure. AlexNet achieved a classification accuracy rate of over 85% in early medical imaging tasks, such as pneumonia screening on chest X-rays. VGG optimizes the parameter efficiency of medical image feature extraction by stacking small convolutional kernels. The residual connection of ResNet enables the network depth to exceed 100 layers, and the error rate on the CIFAR-10 dataset (which

DOI: 10.54254/2755-2721/2026.TJ29145

can be transferred to the small sample scene of medical images) is reduced to 3.6%. Deep learning is the core research direction in the field of image classification, among which the CNN is the mainstream architecture. It automatically extracts image features through multi-layer convolution operations. Combined with image enhancement techniques, it can improve the generalization ability of small-sample medical image datasets. Typical models such as AlexNet, VGG, and ResNet have solved problems like vanishing gradients and parameter redundancy by optimizing network depth and structure, and have performed well on medical imaging tasks and related datasets, providing support for medical image classification and other aspects.

In recent years, the Transformer architecture has demonstrated unique advantages. The ViT segments images into serialized blocks and captures global dependencies through the self-attention mechanism. Its top-1 accuracy rate on the ImageNet-1k dataset (including multiple images such as medical and natural scenes) reaches 88.5%. In the detailed scenarios of medical imaging, such as the detection of pulmonary nodules, ViT can accurately capture the global correlation between nodules and blood vessels and pleura. Compared with traditional CNN models (such as 3D-CNN), it can reduce the missed diagnosis rate by 5% to 8% in small sample pulmonary nodule datasets. Wang's research confirmed that by integrating the time series processing capabilities of Long Short-Term memory networks (LSTM), the accuracy rate in multi-scenario classification of multi-temporal remote sensing images (such as urban change monitoring and crop growth cycle analysis) has increased by 4.3% to 89.6%. Meanwhile, lightweight design has become a trend. For instance, separable convolution, by decomposing standard convolution operations, reduces the model's computational load by 70% on mobile medical devices (such as portable fundus cameras), while maintaining a classification accuracy decline of no more than 2%, meeting the computing power requirements for real-time screening (such as diabetic retinopathy).

This indicates that CNN still has an advantage in local feature extraction, while Transformer and hybrid architectures perform better in global feature-dependent tasks, and a lightweight design provides the possibility for edge device deployment.

3.2. Model optimization and performance enhancement

Model optimization and performance improvement are two of the core research directions at present, and data augmentation and transfer learning are the most widely applied methods. Data augmentation expands sample diversity by performing transformations such as rotation, flipping, and cropping on the original image, effectively alleviating overfitting. Wang's research shows that the time series feature extraction method combined with the dynamic time-curved mass averaging algorithm has increased the classification accuracy of remote sensing images by 3.8%. Transfer learning utilizes the common features of pre-trained models to adapt to specific tasks quickly. For instance, the anemia screening system constructed by Hu Xiaoyan's team using the InceptionV3 model achieved an accuracy rate of 95.2% in the classification of palpetic conjunctival images, which is 12.5% higher than training from scratch [14].

The attention mechanism provides new ideas for model optimization. The Graph Attention Network (ML-GAT) proposed by Zhang improves the average accuracy of multi-label image classification by 5.1% to 87.3% by mining the asymmetric relationship between labels [6]. This mechanism can dynamically adjust the importance weights of feature regions, enabling the model to focus more on key information. In terms of lightweight model design, Wang's lightweight spatio-temporal enhanced graph convolutional model reduces the number of parameters to 30% of the original model in miner behavior recognition, while maintaining a classification accuracy rate of 91.4%, meeting the real-time requirements.

DOI: 10.54254/2755-2721/2026.TJ29145

Adversarial training techniques enhance model robustness. Deng's research indicates that after introducing adversarial sample training, the model's defense capability against physical adversarial patches increases by 25%. The two-stage domain adaptation model proposed by Li improves the classification performance in small sample scenarios by 12% through prototype feature transfer. These methods jointly promote the development of image classification models towards higher accuracy and stronger generalization ability.

The above-mentioned research indicates that data augmentation and transfer learning are fundamental means to enhance model performance. Attention mechanisms and lightweight design have achieved remarkable results in balancing feature selection and efficiency. Adversarial training provides an important guarantee for model security.

3.3. Research on cross-disciplinary image classification

Cross-domain image classification research promotes the diversified development of models due to the special demands of different fields. In the field of medical imaging, the Hu team utilized pre-trained models such as InceptionV3 to construct an anemia screening system for palpebral conjunctival images, achieving non-invasive anemia detection with a classification accuracy rate of 95.2% and a sensitivity of 94.7%. The diffusion model proposed by Shang is aimed at the damage detection of aero engine blades. It directly generates the target detection frame by taking the noise frame as the input, achieving a damage recognition accuracy rate of 98.3%, which is 6.4% higher than that of traditional detection methods. These applications indicate that the classification of professional fields needs to combine domain knowledge to optimize the general model.

The field of autonomous driving has strict requirements for real-time performance and robustness. The LSTM classification model developed by Wang, by improving the dynamic time warping algorithm, achieves a processing speed of 30 frames per second and an accuracy rate of 89.6% in the temporal feature extraction of remote sensing images, providing technical support for environmental perception. Li's two-stage domain adaptation model solves the domain shift problem caused by weather and lighting changes through prototype feature transfer, improving the classification accuracy of autonomous driving scenarios by 8.7%.

The field of security monitoring focuses on the classification of small targets and behavior recognition. The lightweight spatio-temporal enhanced graph convolutional model designed by Wang achieves efficient recognition of miners' behaviors through skeleton sequence analysis, with a video processing speed of 25 frames per second and an accuracy rate of 91.4%. Zhang's ML-GAT improves the multi-label classification accuracy of the monitoring scene by 5.1% by mining the asymmetric relationship between labels.

Cross-domain applications face the challenge of data heterogeneity. The vision-text contrast model proposed by Wang solves the problem of feature heterogeneity in light field image quality assessment by integrating multimodal information, achieving an assessment accuracy rate of 92.6%. Deng's physical countermeasure patch defense technology provides a solution for malicious attacks in security systems, enhancing the model's anti-interference capability by 25% [14]. Overall, cross-domain classification needs to balance the complexity and performance of the model based on the scenario. Combining domain knowledge, transfer learning, and lightweight deployment is the key to the technical implementation.

DOI: 10.54254/2755-2721/2026.TJ29145

4. Challenges and prospects

4.1. Existing challenges

Current image classification research still faces multiple challenges. Firstly, the insufficient interpretability of the model restricts its application in key fields. In scenarios such as medical diagnosis, it is necessary to clarify the basis for model decision-making. However, the black box characteristics of existing models like VGG and ViT make it difficult to meet the requirements. Although Zhang Huiyi's ML-GAT can reveal the importance of some features through the attention mechanism, there is still a gap from the reasoning logic that humans can understand [6]. Secondly, adversarial sample attacks threaten the security of the model. Deng's research shows that physical confrontation patches can reduce the model's accuracy in traffic sign classification from 98% to 62%, a decrease of 36 percentage points, while existing defense technologies can only restore 10% to 15% of the performance [7]. Then, the problems of small samples and data heterogeneity will become prominent. The cost of annotating data in professional fields is high. For instance, the annotation of medical images requires the participation of professional physicians, resulting in a limited sample size. Wang's research indicates that the heterogeneity of light field images and text information can reduce the performance of multimodal fusion models by 8% to 12%. Although Li's domain adaptation model has a 12% performance improvement in small sample scenarios, its generalization ability is still limited by the similarity between the source domain and the target domain. The final challenge is balancing efficiency and accuracy. The computing resources of edge devices, such as mobile phones and drones, are limited. Although Wang Jianfang's lightweight model has reduced the number of parameters by 30%, its accuracy still drops by 2.1%. How to maintain high accuracy under resource constraints remains the core issue [15].

4.2. Future outlook

In response to the above challenges, future research can focus on the following directions. First, explore the structure of biological heuristic networks. By simulating the dynamic attention mechanism of the human visual system and designing a network that can dynamically adjust the receptive field, the model's ability to capture key features can be enhanced, which is expected to further increase the classification accuracy by 5% to 10% in complex scenarios. Secondly, develop an adaptive multimodal fusion strategy. By integrating meta-learning techniques to reduce reliance on labeled data and quickly adapting to new modalities through a small number of labeled samples, the visual-text model developed by Wang Hanling's team can draw on this approach to further enhance the stability of heterogeneous data fusion. In addition, build an edge computing security framework. By combining lightweight design with adversarial training and embedding an adversarial sample defense module during the model compression process, Wang Jianfang's lightweight model can enhance its anti-interference capability while maintaining efficiency in this way. Finally, promote technological innovation in explainability. Develop a feature importance analysis method based on causal reasoning to make the model's decision-making logic traceable and provide more reliable technical support for fields such as healthcare.

5. Conclusion

This paper systematically reviews the theoretical framework of image classification models, analyzes the feature extraction mechanisms of mainstream architectures such as CNNs and

DOI: 10.54254/2755-2721/2026.TJ29145

Transformers, and explores the performance improvement effects of key technologies such as attention mechanisms and residual connections. Existing research indicates that model lightweighting, multimodal fusion, and few-shot learning are the three core directions. Self-supervised pre-training and neural architecture search technologies have significantly promoted the development of the field.

In cross-domain applications, CNN performs stably in local feature extraction, Transformer has more advantages in global dependency tasks, and the hybrid architecture and lightweight design provide the possibility for the deployment of edge devices. Optimization methods such as data augmentation, transfer learning, attention mechanisms, and adversarial training have effectively enhanced the accuracy, generalization ability, and robustness of the model.

Current research still faces challenges such as insufficient interpretability, vulnerability of adversarial samples, heterogeneity of small samples and data, and the balance between efficiency and accuracy. In the future, efforts should be focused on the innovation of biological heuristic structures, adaptive multimodal fusion, edge security frameworks, and explainability technologies to promote the wider and more reliable application of image classification technology in practical scenarios.

References

[1] Wu, Y. (2025). Research on anti-theft based on an image enhancement model. Information Security Research, 3, 214–220.
[2] Le, D. P. C. (2024). A comprehensive survey of recent transformers in image, video, and diffusion models. Computers, Materials & Continua, 7, 37–60.
[3] Wang, H. (2024). Remote sensing image LSTM classification model for multiple application scenarios. Radio Engineering, 4, 977–983.
[4] Wang, H. (2024). Light field image quality assessment based on contrastive vision-text model. Acta Electronica Sinica, 10, 3562–3577.
[5] Hu, X. (2023). Research on establishing an anemia screening algorithm model for palpebral conjunctival images based on deep learning. Journal of Army Medical University, 8, 746–752.
[6] Zhang, H. (2022). Multi-label image classification model based on graph attention network. Journal of Chongqing Technology and Business University (Natural Science Edition), 1, 34–41.
[7] Deng, H. (2025). Review on physical adversarial patch attack and defense technology. Journal of Information Security, 1, 75–90.
[8] Li, Z. (2024). Two-stage unsupervised domain adaptive segmentation model for remote sensing images combined with prototype. Computer Science and Exploration, 8, 2091–2108.
[9] Wang, J. (2024). Miner behavior recognition under the convolutional model of spatio-temporal enhanced graph for lightweight pose estimation. Industrial and Mining Automation, 11, 34–42.
[10] Shang, H. (2025). Intelligent detection of blade damage in aero engines based on diffusion model. Aviation Science and Technology, 2, 29–37.
[11] Xiang, D., Zhou, Z., & Gao, P. (2025). Review of research progress on underwater image enhancement. Journal of Quantum Electronics, 1–15.
[12] Xu, B., & He, X. (2025). Research on night image enhancement algorithm based on improved MSR theory. Communications and Information Technology, 4, 13–17.
[13] Du, Z. (2025). Research on low-light image enhancement algorithm based on deep learning (Master's thesis). Chongqing University of Technology.
[14] Zhang, Y. (2025). Research on image enhancement methods for real outdoor scenes (Doctoral dissertation). Beijing University of Posts and Telecommunications.
[15] Lu, Z., Dong, S., Xia, Z. (2025). Underwater image enhancement of multi-distortion types based on improved CycleGAN. Journal of Zhejiang University (Engineering and Technology Edition), 59(6), 1148–1158.