Artificial Intelligence and Computing Innovations

Intelligence Science Publishing

https://www.ispublishing.com

ISSN: 3070-4944

Vo.6, No.2, 2026

Image Classification via Joint Multi-Scale Convolution and Global Attention Modeling

Jose Dolz

Illinois Institute of Technology, Chicago, USA

Jose.D77@gmail.com

Abstract: This study focuses on the task of image classification and proposes a fusion method based on multi-scale convolution and global attention to address the limitations of traditional approaches in local feature extraction and global dependency modeling. The method first employs a multi-scale convolution module to extract image features under different receptive fields, enabling the capture of both fine-grained details and macro-structural information. Global attention is then introduced to dynamically assign weights at the feature level, strengthening dependencies across regions and ensuring effective global semantic understanding. Through the joint design of local and global modeling, the method achieves more comprehensive feature representation in complex image scenarios. In the feature fusion and aggregation stage, the results of multi-scale convolution and global attention are effectively combined, followed by classification outputs. Experiments conducted on public datasets with comparative validation and sensitivity analysis show that the proposed method outperforms common baselines such as MLP, CNN, LSTM, and Transformer in terms of AUC, ACC, Precision, and Recall, demonstrating the advantage of combining multi-scale and global attention in improving classification performance. Further hyperparameter sensitivity experiments indicate that factors such as the number of attention heads, batch size, and noise level influence model performance, highlighting the importance of proper configuration for enhanced stability. Overall, the method exhibits strong accuracy and robustness, validating the effectiveness of the fusion of multi-scale convolution and global attention.

Keywords: Image classification, multi-scale convolution, global attention, feature fusion

1. Introduction

In today's information society, images have become one of the most important carriers of information. They are widely used in medical diagnosis, traffic monitoring, security protection, industrial inspection, and intelligent interaction. With the rapid growth of data and the increasing complexity of application scenarios, how to complete image classification efficiently and accurately has become a key prerequisite for integrating artificial intelligence with various industries. Image classification is not only a core task in computer vision

but also a fundamental step for more complex tasks such as object detection, image segmentation, and video understanding. The methods of feature extraction and pattern recognition directly affect the performance of these advanced tasks. Therefore, exploring image classification methods with stronger representation and generalization ability carries both academic value and practical importance[1].

Traditional convolutional neural networks have achieved remarkable progress in image classification. Their local receptive fields and weight-sharing mechanisms allow the effective capture of local features. However, with the growing complexity of image scenes, relying solely on fixed-scale convolution kernels is often insufficient to balance fine-grained details and global structural features. In real-world applications, image content often shows multi-scale and multi-level feature distributions. Examples include near and distant objects in natural scenes, different tissue structures in medical images, and macro shapes and micro defects in industrial inspection. If a model cannot handle features at multiple scales, classification results may be biased. Introducing multi-scale convolution mechanisms enables feature extraction at different receptive fields, which enhances robustness and accuracy[2].

At the same time, image classification also faces the challenge of global dependency modeling. Convolution operations emphasize local feature capture and have limited ability to represent long-range dependencies and overall semantics. In complex images, there are strong spatial and semantic associations between targets, such as object relations in scene images or structural coherence in medical images. If these global dependencies are ignored, critical semantic information may be lost. Global attention mechanisms provide new solutions to this problem. By dynamically assigning weights to features, attention can model dependencies between regions across the whole image. This improves the understanding of global semantics and makes classification decisions more interpretable and stable.

The combination of multi-scale convolution and global attention offers an effective approach that balances local details and global structure. Multi-scale convolution ensures coverage of features at different levels, while global attention strengthens semantic connections across regions. Their complementarity improves adaptability in complex scenarios. In intelligent medical diagnosis, this method captures both small lesion details and overall image structures. In traffic monitoring, it recognizes license plate details while modeling the broader traffic scene. In industrial inspection, it identifies tiny defects without losing sight of the full product appearance. This synergy enhances both accuracy and robustness and supports applications across multiple domains[3].

From a broader perspective, research on image classification has far-reaching implications for the application of artificial intelligence in society. With the continuous growth of big data and computing power, methods based on multi-scale convolution and global attention can improve scalability and universality. They also bring practical benefits in terms of efficiency, cost reduction, and safety. For instance, in public security, they enable real-time monitoring and risk warning for smart cities. In healthcare, they assist doctors with more precise diagnoses. In industrial production, they support automated quality control. These potentials demonstrate that exploring classification methods combining multi-scale and attention is not only an academic frontier but also a realistic need for social and economic development.

2. Related work

In the field of image classification, early studies mainly relied on traditional machine learning methods and handcrafted feature design. These methods typically represented images using low-level features such as texture, edges, or color histograms, and then applied classifiers such as support vector machines, decision trees, or ensemble learning for recognition. Although these approaches achieved certain results in restricted tasks, handcrafted features were insufficient to represent the high-dimensional information in complex scenes. As a result, they showed clear limitations in generalization and robustness. With the increasing demands of real applications, classification methods based solely on shallow features gradually revealed their weaknesses, which provided the foundation for the rise of deep learning methods[4].

The emergence of convolutional neural networks greatly advanced the development of image classification. The local receptive fields and weight-sharing mechanisms of convolution allowed models to automatically extract hierarchical features from raw data, freeing them from the constraints of manual feature design. Multi-layer convolution and pooling structures enabled the capture of information from low-level textures to high-level semantics, and demonstrated strong performance on large-scale datasets. However, traditional convolution structures rely on fixed-scale kernels. When dealing with objects of varying scales and complex scenes, they often lose critical cross-scale information. Moreover, since convolution is essentially a local operation, models struggle to represent long-range dependencies, which has become a major bottleneck for further improvement[5,6].

To address these problems, multi-scale convolution structures were proposed. By introducing convolution kernels of different sizes in parallel or hierarchical ways, models can obtain both fine-grained and global features. Such methods significantly enhance adaptability to multi-scale objects and achieve better robustness in tasks such as natural scene recognition, medical image analysis, and industrial inspection. At the same time, as networks became deeper, structures such as residual connections and dense connections were introduced to alleviate gradient vanishing and improve feature transmission. However, simply expanding the convolution kernel scales remains insufficient for capturing semantic information from a global perspective. When image scenes are complex and objects exhibit semantic associations, multi-scale design alone is still inadequate[7].

To overcome the locality of convolution, global attention mechanisms have become an important direction in image classification research. By dynamically weighting features, attention mechanisms capture dependencies between regions across the entire image, strengthening semantic representation and contextual association. This not only enhances discriminative power but also improves the ability to focus on key regions in complex scenes. In recent years, the integration of attention mechanisms with convolutional structures has been widely applied to image classification, forming a hybrid paradigm that balances local details and global structure. Multi-scale convolution ensures the richness of features, while global attention complements long-range dependency modeling. Their synergy provides a more comprehensive and robust solution for image classification. This trend also shows that future studies will increasingly emphasize the combination of convolutional structures and global modeling strategies to further improve accuracy, generalization, and adaptability.

3. Method

In terms of methodological design, this study constructs a feature extraction module based on multi-scale convolution. To address the large differences in object scales within an image, the model introduces multiple convolution kernel sizes at the same level to extract feature representations at different scales in parallel, and achieves multi-scale information fusion through feature splicing. This parallel convolution design ensures that both fine-grained features and macroscopic structures are captured simultaneously, providing a more comprehensive feature foundation for subsequent global dependency modeling. The entire multi-scale convolution process can be formalized as:

$$F_{\text{ms}} = \text{Concat}(\text{Conv}_{k_1}(X),...,\text{Conv}_{k_n}(X)) \tag{1}$$

After obtaining multi-scale features, the model further introduces a global attention mechanism to enhance the modeling capability of long-range dependencies. The core idea of global attention is to assign weights to different locations through correlation calculation, thereby focusing on key areas. Specifically, the input features are first mapped into query, key, and value representations, and global context-enhanced features are generated through a weighted combination. The calculation process is as follows:

$$F_{att} = \text{Softmax}(\frac{QK^T}{\sqrt{d_k}})V \tag{2}$$

This mechanism can capture the dependencies between regions globally and enhance the completeness of feature semantic expression. The model architecture is shown in Figure 1.

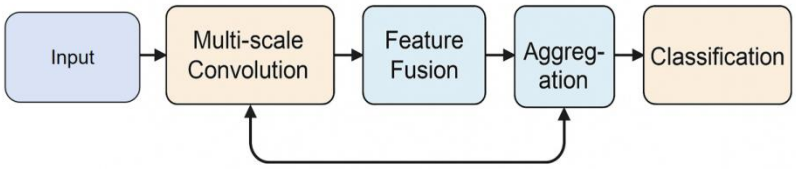

Figure 1. Overall model architecture

After the fusion of multi-scale convolution and global attention, the model further aggregates and classifies the features. First, the fused features are mapped to the output space through linear transformation and optimized using a standard classification loss function. In classification tasks, cross-entropy is often used as the optimization objective to measure the difference between the predicted distribution and the true distribution. It is defined as:

$$L_{cls} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c} \log \hat{y}_{i,c} \tag{3}$$

Where N is the number of samples, C is the number of categories, $y_{i,c}$ represents the one-hot encoding of the true label, and $\hat{y}_{i,c}$ represents the probability distribution predicted by the model. By jointly optimizing multi-scale convolutional features and global attention features, the model can maintain sensitivity to local details while taking into account the overall semantic structure, resulting in a more robust classification method.

4. Experimental Results

4.1 Dataset

This study adopts the CIFAR-100 dataset as a standard benchmark for validating image classification methods. The dataset consists of 100 fine-grained categories, each containing 600 color images of size 32 × 32 pixels, for a total of 60,000 images. Among them, the training set includes 50,000 images, and the test set includes 10,000 images. The categories cover a wide range, including natural objects, vehicles, household items, and animals. Compared with the more commonly used CIFAR-10, CIFAR-100 has finer category divisions. As a result, it poses greater difficulty in feature learning and pattern recognition and places higher demands on the generalization ability of models.

The dataset is characterized by strong diversity and balanced classes. Each category contains the same number of images, which helps to avoid training bias caused by sample imbalance. The image content distribution is complex. It involves low-level differences in texture and color as well as high-level semantic similarities. This requires models not only to capture local features but also to possess strong global modeling ability. These characteristics match the research needs of multi-scale convolution and global attention, providing a rigorous test of effectiveness for fine-grained classification.

In addition, CIFAR-100 is a widely used public dataset and has become an important benchmark for comparative analysis and method evaluation in image classification tasks. It maintains a reasonable balance between scale, complexity, and diversity. This allows studies to fully explore model performance and robustness without relying on very large-scale computing resources. Therefore, research based on CIFAR-100 is not only reproducible and comparable but also provides a valuable reference for the further development of image classification.

4.2 Experimental Results

This paper first gives the results of the comparative experiment, as shown in Table 1.

Table1. Comparative experimental results

Model	AUC	ACC	Precision	Recall
MLP[8]	0.873	0.812	0.804	0.789
CNN[9]	0.902	0.836	0.829	0.817
LSTM[10]	0.918	0.849	0.842	0.833
Transformer[11]	0.931	0.862	0.856	0.847
Ours	0.953	0.889	0.874	0.868

From the results in Table 1, it can be observed that the traditional MLP model shows relatively limited performance across all metrics. AUC, ACC, Precision, and Recall are at the lowest levels. This indicates that relying only on fully connected structures to model image features cannot effectively capture spatial hierarchy and global dependencies, leading to insufficient classification ability. Such limitations are especially evident in complex image scenarios, suggesting that shallow feature learning alone cannot meet the demand for high-accuracy classification.

With the introduction of convolutional structures, the CNN model shows clear improvements in all four metrics, particularly with significant gains in ACC and Precision compared with MLP. This result reflects the advantage of convolutional kernels in extracting local spatial features, allowing the model to better identify salient patterns in images. However, CNNs are still limited by fixed receptive fields. For image content with large-scale variations or long-range dependencies, its performance remains constrained, which motivates further improvement.

LSTM and Transformer models continue to demonstrate enhanced performance, with the Transformer achieving higher values in AUC and ACC. This shows that mechanisms based on sequence modeling and global attention can effectively capture long-range dependencies and global semantic relations. Compared with convolution-based models, they are better suited for handling complex and diverse image classification tasks. Nevertheless, the performance of the Transformer still shows certain limitations, suggesting that a single global attention mechanism is not sufficient to fully represent fine-grained and multi-scale features.

In contrast, the method proposed in this study achieves the best performance across all four metrics, with further improvements in AUC and ACC compared with the Transformer. This indicates that the fusion of multi-scale convolution and global attention can better balance local details and overall semantics, resulting in more comprehensive feature representation and stronger discriminative ability. This advantage highlights the adaptability of the method in complex scenarios and confirms its effectiveness in improving classification accuracy and robustness. It also lays a solid foundation for applications in larger-scale and more complex tasks.

This paper also includes an experiment designed to investigate the sensitivity of the number of attention heads to the single metric Recall. The purpose of this experiment is to evaluate how different configurations of attention heads influence the model's ability to capture dependencies across feature subspaces and how such variations affect the stability of Recall as a key performance indicator. By systematically adjusting the number of attention heads, the experiment provides a focused analysis of this hyperparameter, highlighting its role as an important factor in balancing local feature extraction and global semantic modeling. The corresponding experimental setup and its evaluation are illustrated in Figure 2.

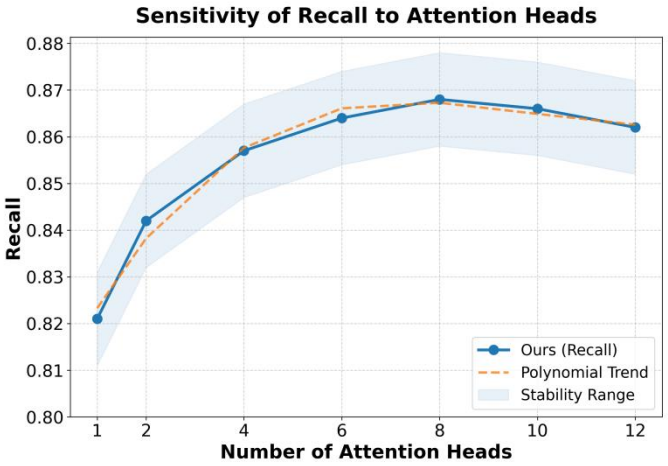

Figure 2. Sensitivity experiment of the number of attention heads to a single indicator, Recall

From the experimental results, it can be seen that when the number of attention heads is low, the Recall value is relatively low. This indicates that too few attention heads fail to cover information from different feature subspaces in global dependency modeling, leading to insufficient recognition of fine-grained features. This shows that the number of attention heads has a direct impact on global modeling ability and is one of the key hyperparameters affecting classification performance.

As the number of attention heads increases, the Recall value shows a gradual upward trend, reaching its peak at eight heads. This demonstrates that within this range, the model can more effectively capture dependencies across different scales and regions, forming a more comprehensive semantic representation. The improvement also verifies the effectiveness of the global attention mechanism in enhancing feature discrimination in image classification tasks.

When the number of attention heads continues to increase to ten and twelve, the Recall value shows a slight decline. This suggests that too many attention heads may introduce redundant modeling, causing a dispersive effect in capturing global information and reducing the final classification performance. This phenomenon reflects the need to balance information sufficiency and parameter redundancy in the design of attention-based models.

Overall, the experimental results reveal the sensitivity between the number of attention heads and Recall, and further emphasize the importance of hyperparameter selection in image classification models. A reasonable number of attention heads can enhance global dependency modeling while maintaining generalization ability, thereby improving robustness and accuracy in complex image scenarios. This also indicates that the combination of multi-scale convolution and global attention can maximize its advantages under proper hyperparameter configurations, providing valuable reference for further optimization and applications.

5. Conclusion

This study focuses on image classification methods based on multi-scale convolution and global attention. It systematically analyzes the limitations of existing approaches in capturing local details and modeling global dependencies, and proposes a fusion mechanism that addresses both aspects. Multi-scale convolution is used to effectively extract features at different scales, while global attention is applied to model long-range dependencies, resulting in a more comprehensive feature representation. This design not only overcomes the restriction of traditional convolutional networks that are limited to local modeling but also enhances the performance of attention mechanisms in complex scenarios, providing a new perspective and practical path for image classification tasks.

The experimental results demonstrate that the proposed method achieves significant improvements in classification performance, indicating that the combination of multi-scale and global modeling plays a complementary role in image understanding. In particular, the method shows stronger adaptability and robustness when dealing with complex scenes, diverse objects, and fine-grained categories. This finding highlights its importance for advancing image classification toward higher accuracy and stronger generalization. At the same time, the design principle of this structure provides transferable insights for subsequent tasks and offers guidance for improving other visual recognition models.

From an application perspective, the method shows strong practical value across multiple domains. In intelligent healthcare, it can assist in the accurate identification of lesions in medical images, improving both diagnostic efficiency and reliability. In traffic monitoring and public security, it enables the recognition of local details and global contexts simultaneously, providing technical support for real-time surveillance and risk warning. In industrial inspection and quality control, it helps detect subtle defects while maintaining an understanding of overall structures, thereby enhancing the level of automation in production. These application prospects indicate that the proposed method not only carries academic value but also demonstrates broad potential in real-world applications.

In conclusion, the image classification method based on multi-scale convolution and global attention provides a feasible solution to the challenges of feature representation and discrimination in complex image scenarios. By reinforcing the joint modeling of local and global information, the method achieves promising improvements in classification accuracy and robustness, and shows wide applicability in multiple practical contexts. The research outcomes not only promote theoretical and technical progress in image classification but also lay a solid foundation for the intelligent development of related industries.

References

[1] Wu L, Wang H. Global and pyramid convolutional neural network with hybrid attention mechanism for hyperspectral image classification[J]. Geocarto International, 2023, 38(1): 2226112.

[2] Zhao P, Yang S, Ding W, et al. Learning Multi-Scale Attention Network for Fine-Grained Visual Classification[J]. Journal of Information and Intelligence, 2025.

[3] Rahman M M, Munir M, Marculescu R. Emcad: Efficient multi-scale convolutional attention decoding for medical image segmentation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 11769-11779.

[4] HCA-former: Hybrid Convolution Attention Transformer for Medical Image Segmentation

[5] Huo X, Sun G, Tian S, et al. HiFuse: Hierarchical multi-scale feature fusion network for medical image classification[J]. Biomedical Signal Processing and Control, 2024, 87: 105534.

[6] Li Y, Wu C Y, Fan H, et al. Mvitv2: Improved multiscale vision transformers for classification and detection[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 4804-4814.

[7] Zhu L, Wang X, Zhang W, et al. Revisiting the integration of convolution and attention for vision backbone[J]. Advances in Neural Information Processing Systems, 2024, 37: 42941-42964.

[8] Tolstikhin I O, Houlsby N, Kolesnikov A, et al. Mlp-mixer: An all-mlp architecture for vision[J]. Advances in neural information processing systems, 2021, 34: 24261-24272.

[9] Liu Z, Mao H, Wu C Y, et al. A convnet for the 2020s[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 11976-11986.

[10] Toumi A, Cexus J C, Khenchaf A, et al. A Combined CNN-LSTM Network for Ship Classification on SAR Images[J]. Sensors, 2024, 24(24): 7954.

[11] Chen C F R, Fan Q, Panda R. Crossvit: Cross-attention multi-scale vision transformer for image classification[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 357-366.