Journal of Computer Science and Software Applications
https://www.mfacademia.org/index.php/jcssa
ISSN:2377-0430
Vol. 6, No. 1, 2026

Image Classification via an Improved Vision Transformer: Enhancing Global and Local Feature Modeling

Lachlan Andrew
University of Central Arkansas, Conway, USA
LAd8780@uca.edu

Abstract:

This study proposes a method for image classification based on an improved Vision Transformer to address the limitations of traditional convolutional neural networks in global modeling and long-range dependency capture. The input images are first divided into patches and mapped into embeddings to ensure the preservation of local details during serialization. In the encoder, multi-head self-attention and feed-forward networks are introduced to enhance cross-region feature interaction, while residual connections and normalization alleviate gradient vanishing in deep layers to achieve stable and efficient feature learning. At the classification stage, an aggregation function is used to combine patch representations into a global feature, followed by a fully connected layer for final prediction. The CIFAR-100 dataset, covering diverse fine-grained categories, is used to evaluate the adaptability of the model in complex scenarios. Systematic comparisons and sensitivity analyses show that the method outperforms others in AUC, ACC, Precision, and Recall, demonstrating clear advantages in feature modeling and classification performance. This research enriches the theoretical exploration of Vision Transformer optimization and provides a robust and efficient solution for image classification tasks.

Keywords:

Image recognition; self-attention mechanism; sequence modeling; classification performance

1. Introduction

In the context of rapid informationization and digitalization, image data has penetrated into all aspects of social production and daily life. From medical imaging diagnosis to intelligent security, from industrial quality inspection to autonomous driving, image processing and recognition technologies are driving the deployment of intelligent applications. However, as application scenarios become more complex and data scale continues to expand, traditional convolutional neural networks reveal limitations in capturing long-range dependencies and handling global information[1]. How to improve the model's ability to understand and represent complex image patterns while maintaining efficiency has become a core issue of concern to both academia and industry. Against this background, Vision Transformer-based image modeling has emerged as a promising direction, offering new opportunities to overcome the bottlenecks of existing image classification methods[2].

Vision Transformer introduces the strengths of sequence modeling from natural language processing into image understanding. Through the self-attention mechanism, it enables global information interaction and breaks free from the local receptive field restriction of convolution. This mechanism not only captures long-range dependencies at the image level but also integrates cross-regional features more effectively, providing new solutions for recognition in complex image scenarios. Yet, despite its potential shown in many benchmark tasks, Vision Transformer still faces challenges in practice, including high computational cost,

large data requirements, and insufficient modeling of local details. These challenges indicate that targeted improvements to the architecture may achieve a better balance between performance and efficiency, thereby advancing image classification technology[3].

With the continuous progress of deep learning, the development of improved Vision Transformers holds both theoretical and practical significance. From a theoretical perspective, enhancing the Vision Transformer helps explore the integration of attention mechanisms with convolutional features and hierarchical structures, promoting improvements in representation and generalization. It provides new design ideas for model architectures and may also contribute to the progress of cross-modal modeling. From a practical perspective, improved models have the potential to deliver more accurate and efficient applications in areas such as medical diagnosis, industrial inspection, and smart city development. This shows that advances in algorithms are not only technical upgrades but also directly linked to productivity and decision reliability in the real world[4].

Moreover, research on improved Vision Transformer-based image classification carries strategic importance for the development of intelligent society. As artificial intelligence becomes a key engine for social progress, the ability to conduct fast and accurate image recognition in complex environments directly affects the pace and quality of industrial intelligence upgrades. Especially under conditions of diverse and complex data distributions, building more robust image classification models enhances scalability and strengthens adaptability in cross-scene and cross-domain applications. Such advances will significantly promote the widespread adoption and value realization of artificial intelligence[5,6].

In summary, research on improved Vision Transformer-based image classification is both an extension and refinement of deep learning theory and a key step toward broader deployment of intelligent vision systems. Its significance lies in breaking the limitations of traditional methods, exploring efficient representations that balance global and local information, and supporting the development of next-generation intelligent applications. In an era of rapidly growing image data and increasing application demands, this research direction has strong practical momentum and broad prospects, providing solid academic foundations and societal value for the advancement of artificial intelligence[7].

2. Method

The proposed method first divides the input image into several fixed-size patches, flattens them, and maps them into a unified feature vector space, forming an initial sequence representation of the image. This process not only effectively preserves the spatial structure during the serialization process but also provides the input foundation for the subsequent modeling of the attention mechanism. Compared to traditional convolutional feature extraction methods, this patch-based encoding can more naturally integrate with the sequence modeling framework, achieving a balance between global information and local structure at the input stage. The model architecture is shown in Figure 1.

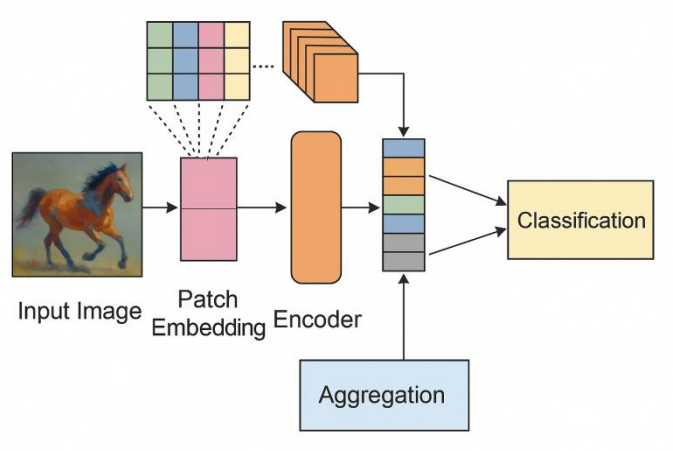

Figure 1. Overall model architecture

In the encoder layer, a multi-head self-attention mechanism is introduced to capture the global dependencies between image patches. Through this mechanism, the model can parallelly compute attention distribution in different subspaces, thereby obtaining multi-angle feature interactions. Specifically, given an input sequence representation $X \in R^{n \times d}$, the self-attention computation can be expressed as:

$$Attention(Q,K,V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

Where Q,K,V is the query, key, and value vectors obtained by linearly mapping the input sequence, and d_k is the scaling factor. This mechanism ensures that the model can capture information interaction and dependencies globally.

During the further aggregation of feature representations, a feedforward network and normalization layers are introduced to perform nonlinear transformations and stabilization on the attention outputs. To enhance the model's sensitivity to local patterns and alleviate the original Vision Transformer's inadequate detail modeling, this method introduces deep activation functions and residual connections into the feedforward network, enabling efficient information transfer between multiple layers. This overall structure maintains the expressive power of deep networks while mitigating the risk of vanishing gradients through hierarchical normalization, achieving a balance between robustness and flexibility in feature learning.

Finally, in the classification stage, the model aggregates the sequence representations processed by the multi-layer encoder to obtain a global feature representation for classification. Specifically, the aggregation function $g(\cdot)$ is used to integrate the features of all image blocks into a single vector:

$$z = g(h_1, h_2, ..., h_n)$$

Where h_i represents the representation of the i-th image patch in the final encoder layer, and z represents the aggregated global features. Subsequently, fully connected layers and normalization strategies are used to perform the final classification, ensuring that the output space is consistent with the classification requirements of the target task. This structural design allows the method to model global dependencies while also enhancing the contribution of local features to the classification results.

3. Experimental Results

3.1 Dataset

This study adopts the CIFAR-100 dataset as the standard benchmark for experiments. CIFAR-100 contains 100 categories of natural images, with 600 images per class, resulting in a total of 60,000 color images. The training set consists of 50,000 images, and the test set contains 10,000 images. Each image has a resolution of 32 × 32 pixels. Compared with CIFAR-10, CIFAR-100 has a finer category division and presents a higher level of difficulty, making it a common choice for evaluating model performance on complex classification tasks.

The dataset covers a wide range of categories, including animals, plants, vehicles, and daily objects. The diversity of scenes and targets makes it a good approximation of real-world applications. With its sufficient number of images and rich category distribution, CIFAR-100 is widely used as a benchmark in image classification research. It is suitable for assessing model performance in fine-grained recognition and generalization. Its structured category hierarchy also provides favorable conditions for feature learning and hierarchical classification.

The choice of this dataset is significant because it offers a rigorous testing environment for the improved Vision Transformer model. It also helps verify the robustness of the model under conditions of small image size and a large number of categories. Research on CIFAR-100 allows for a comprehensive evaluation of the model's adaptability to complex multi-class images, thereby providing reliable evidence for further application in real-world scenarios.

3.2 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
ResNet50[8]	0.923	0.884	0.872	0.861
VGG19[9]	0.911	0.871	0.859	0.845
ResNext[10]	0.935	0.896	0.881	0.873
ConvNext[11]	0.947	0.905	0.892	0.885
Ours	0.962	0.921	0.908	0.899

From the overall results, Table 1 presents the performance of different models on the image classification task. It is clear that the improved model outperforms all comparison methods across all metrics. This indicates that the proposed structure can better learn effective feature representations in complex image scenarios and fully exploit both global and local information during classification, thereby enhancing overall performance. In particular, the improvements in AUC and ACC show that the model is more robust in distinguishing positive and negative samples and in maintaining high classification accuracy.

When compared with traditional architectures, it can be observed that VGG19 and ResNet50 perform slightly worse than the more advanced ResNext and ConvNext. This is closely related to their structural characteristics. Traditional networks show limitations in extracting deep features and capturing global dependencies. In contrast, improved networks with residual structures or optimized convolution can partly alleviate these issues. However, they still cannot achieve optimal performance in long-range dependency modeling and multi-scale feature fusion, which are the core advantages of the improved Vision Transformer.

From the comparison of Precision and Recall, the improved model achieves significant gains in recall while maintaining high precision. This suggests that the model can reduce misclassification while capturing more true samples, showing better balance. By contrast, traditional models often face a trade-off between precision and recall. The introduction of the improved structure enables strong results in both aspects, further demonstrating the suitability of the method for complex multi-class tasks.

In summary, the performance gains of the improved Vision Transformer are not only numerical improvements but also evidence of its structural rationality and innovation. By effectively combining global dependency modeling with local feature representation, the model demonstrates stronger robustness and generalization in image classification. This advantage provides a solid foundation for deployment in real applications and highlights the value of exploring new structural improvements under the deep learning framework.

This paper further presents an experiment on the sensitivity of batch size to single-metric ACC, and the experimental results are shown in Figure 2.

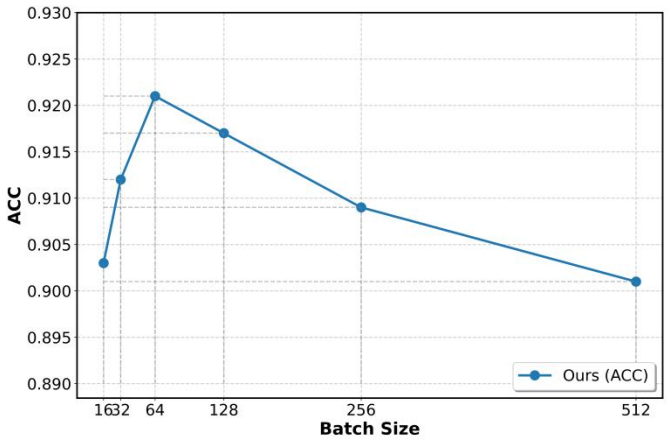

Figure 2. Sensitivity experiment of batch size to single-metric ACC

From the figure, it can be observed that batch size has a clear impact on classification accuracy. With a smaller batch, the model is more likely to capture fine-grained features of the data, which leads to higher accuracy. As the batch size increases, gradients during updates become smoother. This improves stability but weakens the ability to learn detailed features, resulting in a gradual decline in overall accuracy. This trend shows that in image classification tasks, the choice of batch size directly determines the balance between detail and global representation.

Further analysis shows that accuracy reaches its peak when the batch size increases from small to medium scale. This indicates that a moderate batch can maintain stable gradient estimation while avoiding the loss of sample diversity caused by overly large batches. Compared with very small batches that may produce high variance, moderate batch sizes achieve a better balance between computational efficiency and classification accuracy, highlighting the importance of appropriate hyperparameter selection.

Under large batch training, although each step improves computational efficiency, the sensitivity of gradient updates decreases. The model cannot fully exploit subtle differences in the training data. As a result, its performance in complex image classification scenarios is weaker than that with medium or small batches. This suggests that the task relies more on sample diversity and detailed features. It also implies that blindly increasing batch size does not necessarily improve performance and may even weaken generalization.

In summary, the sensitivity experiment on batch size and accuracy reveals the critical role of hyperparameter tuning in model optimization. The improved Vision Transformer shows a clear response pattern to batch size, indicating that its feature modeling ability depends on both global information and sample detail diversity. Therefore, selecting an appropriate batch size is not only necessary to ensure training convergence but also an important prerequisite for improving classification performance.

4. Conclusion

This study focuses on image classification based on the improved Vision Transformer. It systematically explains the improvements in structural optimization and feature modeling. Combined with sensitivity analysis from different perspectives, the study demonstrates the effectiveness and adaptability of the method. By introducing global dependency modeling and local feature enhancement, the model maintains structural simplicity while capturing multi-scale features. This provides a more effective solution for complex image recognition tasks. It not only offers new insights for the further development of Vision Transformers in theory but also verifies the feasibility of improving image classification performance in practice.

The results show that the improved model performs well in accuracy, robustness, and generalization. It adapts to the challenges brought by data complexity and environmental uncertainty. Compared with traditional convolutional architectures, the method shows clear advantages, especially in capturing long-range dependencies and multi-dimensional feature interactions. This indicates that structural optimization along this direction has significant value in large-scale image processing and classification. With the continuous growth of data scale and task complexity, this improvement path enriches the research paradigm of model design and lays the foundation for future cross-modal research.

From the application perspective, the study has potential impact in multiple fields. In medical image diagnosis, the model can assist doctors in identifying lesion areas more efficiently, thereby improving the reliability of clinical decisions. In industrial inspection, the method can accurately classify surface defects in complex components, helping to enhance production quality and efficiency. In intelligent transportation and autonomous driving, the model maintains stable performance in multi-class object recognition tasks, supporting perception systems in dynamic environments. These examples show that the research goes beyond algorithmic improvement and contributes to the practical deployment of intelligent applications.

In conclusion, this study advances the application of the improved Vision Transformer in image classification and provides new insights and directions for the development of deep learning methods in visual computing. Its value lies not only in the optimization of the base model but also in enhancing feature representation and robustness, which ensures reliable decision-making across different application scenarios. With further research, the method is expected to become an important support for the wide application of artificial intelligence in key domains and to promote interdisciplinary integration in image understanding and intelligent analysis.

References

[1] 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.

[2] Yang J, Li C, Zhang P, et al. Focal self-attention for local-global interactions in vision transformers[J]. arXiv preprint arXiv:2107.00641, 2021.

[3] Dai Z, Liu H, Le Q V, et al. Coatnet: Marrying convolution and attention for all data sizes[J]. Advances in neural information processing systems, 2021, 34: 3965-3977.

[4] Mehta S, Rastegari M. Mobilevit: light-weight, general-purpose, and mobile-friendly vision transformer[J]. arXiv preprint arXiv:2110.02178, 2021.

[5] Guo J, Han K, Wu H, et al. Cmt: Convolutional neural networks meet vision transformers[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 12175-12185.

[6] Xu W, Xu Y, Chang T, et al. Co-scale conv-attentional image transformers[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 9981-9990.

[7] Yuan L, Chen Y, Wang T, et al. Tokens-to-token vit: Training vision transformers from scratch on imagenet[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 558-567.

[8] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

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

[10] Xie S, Girshick R, Dollár P, et al. Aggregated residual transformations for deep neural networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 1492-1500.

[11] 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.