Fine-Grained Image Classification Based on Attention Mechanism and Fusion Method

Shenboyi Wang

{25061300054@stu.xidian.edu.cn }

School of Economics and Management, XIDIAN UNIVERSITY, Shaanxi, China

Abstract. Due to the evolution of computer vision, the recognition of fine-grained image classification problems is gaining more and more importance. This paper organizes an overview of the principal conceptualisations of the research and general approaches within the domain of fine-grained image classification, in the view of the mechanisms of attention, image fusion procedures, and a mixture of both. The attention mechanism-based approach primarily enhances the capability of the model in identifying small variations through the focus on the main regions of interest and discriminative characters. The research based on the combination of the attention mechanism with the fusion method further capitalizes on the strong merits of both approaches and thereby promotes better classification performance of the model in difficult scenes. This article aims to overview of the recent advances in associated research methodology and to assist scientists in being more attentive to the progress of research and related current information.

Keywords: Fine-grained image classification, Attention mechanism, Fusion method

1 Introduction

Artificial intelligence is a developing field with an abundance of attention today. Computer vision is an essential and significant discipline that focuses on helping machines sense things with the help of sensors and perceive the world. One of the fundamental studies in the field of computer vision is image classification, on which numerous computer vision tasks, like object detection, are based. In image classification tasks, it is generally dichotomized into coarse-grained and fine-grained classification depending on the degree of classification refinements. The crude-grained category emphasizes big classes, which are highly differentiated with considerable visual variations in sub-classes and a relatively clear feature. Instead, fine-grained classification distinguishes finer subclasses between the coarse-grained categories, and minor differences between subclasses are largely found in the local features. Course-grained image classification has reached high accuracy and fairly mature technology after years of development, even on large-scale datasets like ImageNet. Fine-grained models have, however, problems of difficulty in localizing features and a lack of feature robustness because of the existence of small visual differences between subclasses and it is sensitive to

ICIAAI 2026, May 22-24, Singapore, Singapore

DOI 10.4108/eai.22-5-2026.2365339

interference. Strong supervision and weak supervision can be used to forge fine-grained supervision. Powerful supervision includes differentiation of images manually and manual annotation, hence the cost is very high. Subsequently, this led people to consider weak supervision, where the image category labeling is applied without supervision to compensate for the severe poverty of strong supervision. Weak supervision falls into two different categories, which are building discriminative feature representations and using attention [1]. Deep learning has given a fresh concept in the above issues, and gradual progress has been made with convolutional neural networks as the standard way of fine-grained classification, due to their strong capability of automatic feature learning. But it still has the issue of easy interference by background information when trying to extract features, and it cannot effectively concentrate on the main difference points between subclasses. It is specifically this shortcoming that is covered by the advent of the attention mechanism. It is dynamic to improve the feature response of the key regions by learning the weightings of importance of various regions in the feature map, and rejecting the irrelevant background. The article identifies the fine-grained image classification tasks by reviewing them in the light of the application of attention mechanisms and techniques of fusion in order to enable future researchers to comprehend the advancements in the area of fine-grained image classification tasks.

2 Fine-grained image classification based on an attention mechanism

Attention mechanisms have revised the way research is conducted in other areas of study, like natural language processing and computer vision. The attention mechanism neural network emulates the human visual attention mechanism, which is a highly valuable technological advancement in current deep learning. Attention mechanism is compared to traditional rigid positioning because it is able to make the network actively learn useful information without forgetting other information [2]. Some methods of attention mechanisms include channel attention, spatial attention, and mixed attention. The common member of channel attention is SENet [3], which is based on the global average pooling to squeeze feature maps, fully connected to obtain cross-channel interaction, and lastly, channel fusion is accomplished by multiplication between the obtained vector and the original feature map. The model of the characteristic of spatial attention is the representation of the Space Transformer Networks (STN) [4], which alters the spatial information of the original image to another space by using a spatial transformation without losing the essential information. The archetypal work of mixed domain is CBAM [5], and this is a serial combination of the 2 types of attention above. The attention mechanism is added to enhance the understanding of images and the expression of saliency and shape of the images, which will further increase the expressiveness and recognition capabilities of images. Due to its popular use in recent years, attention mechanisms can be used in various aspects of computer vision tasks, including image classification, pose recognition, and object detection, e.g., in intelligent agriculture or the automatic process of biodiversity. Though the implementation form and structure of the attention mechanisms in applications is not predetermined, the fundamental purpose of the attention mechanisms is to automatically detect and extract the information regions of images in the training data. In particular, attention mechanisms have been proposed to be a useful tool where the accuracy of recognition is concerned, as in the very tricky issue of fine-grained image classification [2]. The next scheme is the SDA Net suggested by Zhang et al. and used in the case of fine-grained classification and relating to component detection and components' multi-granularity feature fusion. The objective behind this approach is to enhance the performance of fine-grained image classification under low data conditions by using the combination of both self-attention and distillation methods. It does not use detailed annotation information but instead simply uses image-level labels (category labels) to get the model training process going. The experimental findings indicate the high adaptability of SDA Net against low-data settings and high performance of feature optimization using self-attention and self-distillation methods, which present substantial evidence of the opportunity to use fine-grained classification tasks in low-resource low-data conditions [6]. Peng et al. came up with an unsupervised domain adaptive fly image classification network (FC-DRONET) model of areas of focus. It is a network under ResNet50, which is trained on a pre-trained U-net that blurs the input image. The masked image is manipulated into the initial image and this is done to isolate the intended area comprising only flies. The network subsequently incorporates the cross attention mechanism that assists in the transmission of local features, therefore gaining information about the fly image in both the horizontal and vertical directions. It has also been established by means of experiments that the FC-DRONET model can be applied with an accuracy level of 99.41 percent and an F1 score of 99.22, which is above that of other models [7].

3 Fine-grained image classification based on fusion methods

The fusion techniques are categorized as data level fusion, feature level fusion and decision level fusion. This is referred to as data layer fusion, whereby raw data is directly analyzed statistically. The focus of this approach is on the correlation of the original data, in such a way that the resultant fused data is capable of giving a more accurate and complete representation or estimate of the target of observation. It preserves all the information of the original data at rather low computational complexity, which is an advantage in enhancing the real-time system performance of the system. However, it is complicated by the sensitivity of the uncertainty and fluctuation of the observation data, making the processing of the system more complicated.

A feature layer fusion comes up with decision layer fusion, which is performed based on the extracted feature vectors, as the combination of judging and deriving from those features, in the quest to arrive at the final conclusion. It is able to choose its results flexibly, to increase system fault tolerance, and to accept more heterogeneous sensor inputs. However, an increased computing workload needs an increased number of computing resources. There have been higher requirements advanced on the algorithm design, particularly in the area of logic reasoning, which is intricate.

The first part of feature-level fusion is the extraction of representational features of the raw data, which are subsequently merged into one feature set to serve as the foundation for further determining the decision. This not only decreases the processing load of the raw data but also increases the processing speed and real-time performance of the system. Moreover, the control of the selective feature extraction is useful to minimize the effects of noise and redundant data, which will enhance the performance of the system. Deploying the processing burden of raw data and efficiency, plus the extraction of representative features, assists in eliminating the unnecessary interference factors. Feature fusion can also be done in different ways, one of them is cascade fusion, another is parallel fusion, and lastly, an attention mechanism. Cascade fusion combines properties of levels: it combines representations of other features, whereas parallel fusion combines features of other models. The pyramid convolutional network is a widespread approach that aims to extract the multi-scale functions in the convolutional layers of various sizes. Pros of doing so are that the lower convolutional layers are able to pick up local features of details and higher convolutional layers pick up more abstract features that are global. The paper by Chen offered a pyramid network of multi-attention MAP Net to classify fine-grained images on the foundation of a feature extraction network, ResNet-50. (1) Bidirectional Feature Path Module Bi-directional Feature Path Module (BFPM). It is comprised primarily of top-down feature paths and bottom-up feature fusion paths, which are used to combine the semantic information of high-level features as well as target detail information of images, which is used to improve the back-and-forth flow of information at feature levels. (2) Mixed Attention Pyramid Module (MAPM) proposes the consideration of a mixed attention mechanism of the bidirectional feature paths module to the feature path module to improve the attention of the model to the vital features of each scale [8]. Through a comparison of 9 methods, it was established that MAP Net was able to perform significantly better on all four datasets.

4 Fine-grained image classification based on attention mechanism and fusion method

The attention mechanism is also capable of making the model select the significant features and ignore unimportant information automatically. In feature fusion, features of various levels and sources are combined to enhance the performance of a model. Thus, it is discovered that the attention mechanism may be employed to employ the feature fusion technique. One can alter the weight of attention when combining two features, thus differentiated feature extraction is possible and the more significant features of image classification can be more pronounced and easier to identify, which results in the acquisition of useful information in a more efficient manner.

The self-supervised contrastive learning was used to propose a fine-grained classification approach, Fine-grained Categorization(SLMFF), based on multi-granularity feature fusion. This approach employs the region obfuscation techniques to substitute the extra component detection sub-networks, hence realizing a lightweight local feature extraction and multi-granularity region partitioning approach. SLMFF model prevails over the region destruction reconstruction methods, multi-granularity progressive training methods in feature extraction and multi-granularity feature fusion. It can be classified better by 0.2% -2% over three datasets, as compared to other ways. This not only shows the high potential of SLMFF to improve performance but also shows that it has excellent performance in the classification [9].

According to the study, Pang et al. have discovered that they implemented an innovative ViT mechanism that not only grasps the significant position of the global information in fine-grained image categorization, but also formulates the spatial interrelationship between global and local, in addition to extracting the complementary information of interlayer features and aggregating a variety of discriminative local areas [10]. Human beings have managed to synthesize CNN local characteristics with ViT global characteristics and enhance the elements of fine-grained images. Simultaneously, they have filtered and merged discriminative sequences between ViT layers to enhance the capacity of the network to take complementary information across layers. Lastly, they have developed an attention pruning module to remove noise in the environment due to fixed-size patches. The CUB-200-2011, Stanford Dogs, Stanford Cars and FGVC Aircraft datasets performed excellently in the dataset.

5 Conclusion

In recent years, significant progress has been made in fine-grained image classification driven by deep learning. Fine-grained image classification is evolving towards deep fusion of multi-scale perception, cross-modal collaboration, and dynamic attention optimization. Based on the above research results, fine-grained image classification still faces the following problems: existing models are still easily affected by redundant background information when extracting fine-grained features, making it difficult to accurately focus on subtle differences between subclasses, and the problem is more prominent in weakly supervised scenarios. Lack of robustness and completeness in feature expression. The positioning accuracy of key areas is insufficient, and the stability of locating and extracting key local features is still poor. Therefore, future research needs to enhance the robust modeling ability of models for local subtle structures while maintaining high accuracy, and explore lightweight attention mechanisms to adapt to edge deployment requirements.

References

[1] Tan, R., Ye, W., Liu, Y.: Fine-grained image classification combining dual semantic data augmentation and object localization. Computer Engineering. Vol. 48, No. 2, pp. 238-242 (2022)

[2] Pan, W., Wei, M., Su, P.: Fine-grained image classification combining feature fusion and hybrid attention. Computer Applications and Software. Vol. 42, No. 9, pp. 210-219 (2025)

[3] Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. IEEE Conference on Computer Vision and Pattern Recognition. pp. 7132-7141 (2018)

[4] Jaderberg, M., Simonyan, K., Zisserman, A.: Spatial transformer networks. 29th International Conference on Neural Information Processing Systems. pp. 25-28 (2015)

[5] Woo, S., Park, J., Lee, J.Y., et al.: CBAM: Convolutional block attention module. European Conference on Computer Vision. pp. 3-19 (2018)

[6] Zhang, J., Geng, L., Liu, N.: Low-data fine-grained image classification based on self-distillation and self-attention enhancement. Computer and Modernization. No. 9, pp. 27-34, 42 (2025)

[7] Peng Y.Q., Rao Y.X., Liao M.X., et al.: Construction of unsupervised adaptive image classification model guided by cross-attention mechanism and its application in fine-grained fruit fly identification. Jiangsu Journal of Agricultural Sciences. pp. 1-7 (2026)

[8] Chen, Q.: Research on fine-grained image classification method based on attention mechanism and feature fusion. Guangdong Polytechnic Normal University. (2023)

[9] Zhang, Y.: Research on fine-grained visual classification algorithm based on multi-granularity feature fusion. Guilin University of Electronic Technology (2024)

[10] Pang, W.: Research on fine-grained image classification method based on feature fusion and attention mechanism. Jiangnan University (2024)