Multi-Grained Attention Studying for Fine-Grained Image Retrieval

Bingran Hao

Pegasus California School grade11, Qingdao, China

Abstract. Finding photos that fall into the same metaclass's subcategories is half of the goal of fine-grained image retrieval. The primary problem that this kind of approach has always had is that some of the current methods are essentially unable to capture discriminative local features, which is crucial to distinguish visually similar subcategories. It has also always had to deal with issues like significant intra-class differences and small inter-class differences, which are very challenging to resolve. The paper proposes a multi-granularity attentional learning approach that adaptively focuses on and deals with the most distinctive regions and features, while ignoring some less necessary regions to enhance fine-grained retrieval. Specifically, the paper designed three collaborative attention modules: channel attention (for adaptively recalibrating channel feature responses), spatial attention (for highlighting salient areas), and partial attention (for locating key object parts). A large number of attempts on the CUB-200-2011 dataset demonstrate the superiority of the method, which the paper verifies is significantly superior to the baseline method and achieves state-of-the-art retrieval performance. Ablation studies and visualisation further validated the effectiveness and complementarity of the different concerns.

1 Introduction

Fine-grained image retrieval is a fundamental task in computer vision, which has drawn increasing research interest due to its wide applications in e-commerce, search engines, and biodiversity protection [1-3]. Different from conventional image retrieval that aims to find images of the same category, fine-grained retrieval targets at identifying images belonging to the same subcategory [4]. This task is extremely challenging because of the subtle differences between subordinate categories and the large variations within each subcategory caused by poses, viewpoints, and backgrounds [5].

Hand-crafted features such as SIFT [6] and intricate feature encoding algorithms [7] are the mainstays of traditional methods for fine-grained recognition. Convolutional neural networks (CNNs), which learn more discriminative features in an end-to-end fashion, have transformed this discipline with the development of deep learning [8–10]. The boundaries of this field of study were pushed by certain groundbreaking works, as Bilinear CNN and

hao15966196880@gmail.com

RA-CNN, which integrated feature interaction learning and attention mechanisms into fine-grained recognition [11–12].

However, most existing methods only consider single-grained features, either global or local, which limits their capacity to fully capture the subtle visual differences [13]. This paper proposes multi-grained attention learning approach that leverages both local and global characteristics in a unified framework. Adaptively focusing on the most discriminative characteristics at various granularities is a crucial realization, including channel, spatial, and object part levels. By doing so, the paper model can comprehensively characterize the fine-grained details while suppressing irrelevant variations.

In summary, this paper makes three primary contributions:

(1) The study suggests a multi-grained attention learning strategy for retrieving fine-grained images. which adaptively identifies the most discriminative regions and features to distinguish visually similar subcategories.

(2) The paper designs three attention modules that collaboratively enhance feature representations from complementary perspectives, i.e., channel attention for recalibrating channel-wise importances, spatial attention for highlighting salient regions, and part attention for locating key object parts.

(3) The paper conducts extensive experiments on the widely used CUB-200-2011 dataset, and the findings show that the paper's approach works far better than the most advanced techniques. achieving significant performance gains in fine-grained retrieval [14].

2 Related Works

2.1 Fine-grained Image Recognition

Identifying photos that fall into distinct subcategories yet are part of the same category is the goal of fine-grained image recognition. For instance, the purpose of bird identification is to differentiate between several bird species. The early fine-grained identification methods mainly depended on some features of manual design, such as SIFT, HOG, etc., and then combined with SVM and other classifiers for identification. These approaches' success is mostly dependent on how features are designed and extracted, and their capacity for generalization is constrained [1–3].

Computer deep learning has led to the success of convolutional neural networks (CNN). In the ImageNet large-scale image recognition competition, Krizhevsky et al. presented the deep CNN model AlexNet(ILSVRC 2012), which greatly reduced the error rate of the top 5 by more than 10 percentage points, which is undoubtedly very important in the field of computing [4]. Since then, a series of CNN models have appeared, such as VGGNet, GoogLeNet, ResNet, etc., which refresh the performance of image recognition. Researchers also introduce CNN into the field of fine-grained recognition, using its powerful feature learning ability to capture the nuances between subcategories [5-7].

The early application of CNN to fine-grained recognition was mainly to directly input the whole map into the pre-trained CNN model, and extract global features for classification processing. However, this method ignores the discriminative information of the local area of the object, and it is difficult to cope with the challenge of small inter-class differences and large intra-class differences in fine-grained identification. For overcoming this problem, some research gradually began to explore the extraction and integration of local features. Branson et al. provided a component-based R-CNN model, which improved the accuracy of bird identification by detecting and matching key parts of objects [8]. Zhang et al. designed a CNN based on component alignment, using local regional matching to reduce intra-class differences [9]. Lin et al. gave a bilinear pooling model B-CNN, This

uses the external product to compute the correlation between the traits of various fish parts, which greatly enhances the discrimination ability of the characteristics [10].

In addition to detecting local areas of matching objects, another refined identification technique adaptively pays attention to locations that are different by using the attention mechanism. A sophisticated identification technique based on an attention mechanism was proposed by Fu et al. RA-CNN, it realizes synchronous learning of classification and positioning through recursive prediction and attention to distinguishing areas [11]. In order to improve feature representation, Zheng et al. created the multi-attention convolutional neural network MA-CNN, which makes use of regional attention and channel attention modules [12]. A sophisticated positioning technique based on class activation mapping CAM-CNN was presented by Cui et al. which can realize the automatic positioning of distinguishing areas without component marking [13].

Although the above methods have solved the difficulties of fine-grained identification to a certain extent, most methods only focus on single-grained features and ignore the complementarity between features. At the same time, some work considering multi-granular features, such as MA-CNN, has failed to model the interaction between different granular features well [12]. Therefore, a crucial problem that requires more research in the realm of fine-grained identification is how to naturally integrate and make use of the distinctive qualities of various particle sizes.

This document, saved in the “Word 97-2003” format, is a guide to using the Manuscript Template. Before submitting your final paper, check that the format conforms to this guide. In particular, check the text to make ensure that the correct referencing style has been used and that the citations are in numerical order throughout the text. Your manuscript cannot be accepted for publication unless all formatting requirements are met.

2.2 Visual Attention Mechanism

The attention mechanism was first proposed by Bahdanau and others in the machine translation task, and then it was widely applied to the field of computer vision [14]. Its core idea is to learn a weight distribution, in order for the model to focus on specific aspects of the input data, emphasize more important aspects, and suppress unrelated interference items, in order to enhance the model's performance. Mnih et al. showed the attention mechanism into image classification for the first time, and selected local areas in multiple time steps through circular neural networks for identification [15]. Ba et al. designed an attention-based circular model to achieve "focusing attention on the most relevant area at this time" by adaptively selecting the sub-area of the input image [16].

With the deepening of research, the mechanism has been extended to various tasks. Xu and others use the attention mechanism to generate image subtitles, and the model dynamically focuses on the corresponding image area according to the currently generated words [17]. Chen et al. introduced the attention mechanism in the image question and answer task, aligned the visual features and the problem features through the question-guided attention, and found the image areas related to the problem as answer clues [18]. Fu and others applied spatial attention and channel attention at the same time in the task of reading pictures and speaking [19]. The former focuses on "what to look at", and the latter focuses on "what to look at" to coordinate sentence generation.

Regarding image recognition, the introduction of this mechanism improves model performance. Hu et al. suggested using SENet to learn the channel's attention weight in order to explicitly model the dependence between feature channels. and adaptively adjust the feature response of different channels [20]. With very little overhead, it achieves performance improvement on multiple visual recognition data sets, revealing the importance of channel attention. Wang et al. further designed ECANet, expanded the

perception field of channel attention to any size, and proposed a high-efficiency channel attention module with performance exceeding the same computing volume of SENet [21]. The CBAM module, which integrates spatial and channel attention in order, was proposed by Woo et al. It improves the capacity for expressing traits [22].

The aforementioned work unequivocally demonstrates that the attention mechanism is a useful tool for enhancing network performance. The success of these works is largely due to the attention to discriminant characteristics and the suppression of interfering information. Inspired by this, this paper designs a multi-grained collaborative attention learning module to extract discriminative information in fine-grained images from the three levels of channel, spatial and local, with the goal of improving multi-granular features' representational power and fine-grained image retrieval performance simultaneously.

3 Methodology

3.1 Overview

First, input the pre-trained CNN backbone network extraction characteristics of the image. Then, through three collaboratively designed attention modules, namely the channel attention module, the spatial attention module and the local attention module, the characteristics are multi-granular. Then, the enhanced features are mapped into binary hash codes through the hash layer for subsequent quick retrieval. End-to-end joint supervision is used to train the entire network, and the optimization goals include classification loss and measurement of learning loss. The following will introduce the design ideas and implementation details of each module in detail. Figure 1 shows this progress.

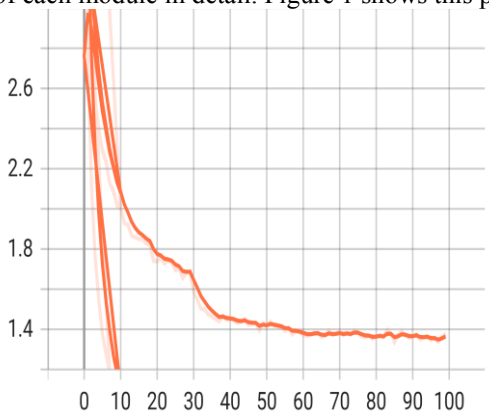

Figure 1 Loss Curve (Picture credit: Original)

3.2 Multi-grained Attention Learning

3.2.1 Channel Attention Module

Modern CNN generally uses multi-channel feature maps to characterize images. However, not all feature channels are equally related to identification tasks. In order to adaptively emphasize channels with a large amount of information and suppress redundant or useless channels, this paper designs a channel attention module. Specifically, given an input feature $X \in R^{CHW}$, first compress the feature graph of each channel into a value through global

average pooling, that is： $S_C = F_{gap}(X_C) = \frac{1}{H \times W} \sum_{i=1}^{H} \times \sum_{j=1}^{W} x_c$ (i, j) Among them, c ∈ {1,2,...,C} is the channel index. This operation is equivalent to representing the global information of each channel in an extremely low dimension (C×1×1).

Then, through the two full-connection layers and the Sigmoid activation function, the weight vector is generated:

$$Z = W_2 \delta(W_1 s) \tag{1}$$

$$A_C = \sigma(Z_C) = \frac{1}{1+e^{-z_C}} \tag{2}$$

Among them,

$$s = (s_1, s_2, \ldots\ldots, s_C) \in R^C \tag{3}$$

is the compressed channel descriptor vector, and W_1 $\in W_1 \in R^{\frac{C}{\gamma} \times C}$, $W_2 \in R^{C \times \frac{C}{\gamma}}$ are two parameter matrices. Γ is the proportional factor that reduces the amount of parameters. This paper takes 4. Δ represents the ReLU activation function. Σ represents the Sigmaoid function, which maps each element in the weighted tensor a=($a_1, a_2, \ldots\ldots, a_C$) to the range of [0,1], and the value of the numerical size to the corresponding channel.

The feature re-cavation is finally finished by multiplying the channel attention weight by the original feature X, element by element:

$$X_C = \widetilde{a_C} \times x_C \tag{4}$$

This process can adaptively calibrate the channel response of the original features: exert greater weight on discriminant channels, and irrelevant or redundant channels are suppressed. It is worth mentioning that the channel attention module has a wide range of applicability and can be seamlessly embedded in any CNN architecture. Its calculation overhead is also very small, and it will not bring a significant increase in the number of parameters.

3.2.2 Spatial Attention Module

There are a large number of background interference and unrelated areas in fine-grained images, and the key to identification is to locate the discriminative foreground area. For this reason, this paper designs a spatial attention module to enhance the prominent area containing the target and suppress background noise. Unlike channel attention to "which channel is more important", spatial attention is mainly focused on "which area is more important".

Specifically, for the input feature X∈R^CHW first perform two pooling operations in the channel dimension to obtain the spatial attention activation diagram:

$$A_{avg} = \sigma(f_{7\times7}((X_{avg}; X_{max}))) \tag{5}$$

$$A_{max} = \sigma(f_{7\times7}((X_{avg}; X_{max}))) \tag{6}$$

Intuitively, the average pooling focuses on the whole area, while the maximum pooling focuses on significant details, and the two are complementary.

Next, multiply the 2D spatial attention map with the original feature X element by element to achieve adaptive feature selection and enhancement:

$$X_{C,I,J} = \widetilde{AI,J} \times x_{C,I,J} \tag{7}$$

Among them, (i,j) represents the index of the spatial position. Through this operation, the area containing the target will receive a greater activation response, while the interference of the background area will be effectively suppressed. It is evident that the spatial attention module can improve the quality of feature representation by adaptively focusing on the discriminative region from the spatial dimension.

3.2.3 Part Attention Module

The first two attention modules enhance the characteristics from the channel and spatial perspectives respectively, but they are all based on the global information of the whole map. For fine-grained images, it is also very important to capture the fine-grained differences in local areas (such as beak shape, wing pattern, etc.). For this reason, this paper further designs the local attention module to develop the descriptive local areas inside fine-grained objects.

Specifically, given a feature graph $\mathrm{X} \in R^{CHW}$, the paper first map it to K local response graphs $A_k \in R^{H \times W}$

$$\mathrm{K}=1,2,\cdots\cdots,\mathrm{K} \tag{8}$$

$$\mathrm{A}=W_a \times X \tag{9}$$

Among them, $W_a \in R^{K \times C \times 1 \times 1}$ is a parameter matrix, and * represents convolutional operations. Each response diagram corresponds to a local area inside the object, and K is the preset number of areas (this article takes 4). In an intuitive understanding, the local response diagram is equivalent to a set of flexible component locators, which can adaptively pay attention to the areas with the greatest fine-grained differences.

Next, use these K response diagrams as weights to locally weigh and sum the original feature X to obtain A collection of vectors for local features

$$v_k \in R^C \tag{10}$$

$$\mathrm{k} = 1,2,\ldots\ldots,\mathrm{K} \tag{11}$$

$$V_k = \frac{\sum_i \sum_j A_{k,i,j} \times x(i,j)}{\sum_i \sum_j A_{k,i,j}} \tag{12}$$

Among them, (i,j) is the spatial position index. This process can be regarded as the decomposition of convolutional features into K decoupled local feature representations, each of which focuses on a key fine-grained area of the object. Subsequent feature learning and classification can focus more on the discriminant details in each local area.

At the same time, these local response diagrams provide an intrinsic, fine-grained explanation for the network, that is to say, which parts of the object the model mainly focuses on to make decisions. This also provides some new and unique perspectives for understanding the inherent mechanism of fine-grained identification.

In a word, the local attention module generates a set of response diagrams to locate key fine-grained areas, and uses them to guide the aggregation of local features, and finally obtains a set of decoupling and discriminative-enhanced local feature representations.

3.3 Network Architecture

The input image is first extracted from multi-scale convolutional features through a pre-trained CNN backbone network. Here use ResNet-50 [7] as the feature extractor, remove the last layer of the full connection layer, and connect to two parallel attention branches after layer-3 and layer-4. For each branch, the characteristic representation of multi-granular interaction is obtained through the channel attention module (CA), spatial attention module (SA) and local attention module (PA) in turn. Then, the characteristics of the two branches are cascaded, through a 1x1 convolution layer dimension reduction, and then the compact feature vector of the image is obtained through global average pooling. Finally, a hash layer is introduced to discrete the real-valued features into binary hash codes for subsequent quick retrieval matching. At the same time, the classification branch is led out in front of the hash layer to improve the discrimination ability of feature representation through multi-task learning.

It is worth noting that the order of the three attention modules is not arbitrary. Put the channel attention at the top, so that the characteristics of the channel dimension can be adjusted first and provide better input for the subsequent spatial attention. Local attention is placed at the end. Because of its focus on the finer characteristics, it needs to learn more refined local discrimination information on the basis of the enhancement of the first two attention. In addition, the attention mechanism is introduced on two scales to capture the discriminative characteristics of different sense fields.

3.4 Objective Function

This paper uses two loss functions to optimize the network: classification loss and measurement of learning loss. Classification loss adopts the form of cross-entropy:

$$L_{cls} = -\frac{1}{N}\sum_{i=1}^{N}\ \sum_{k=1}^{C} y_{ik} \times log(p_{ik}) \tag{13}$$

Among them, N is the number of training samples and C is the number of categories. $Y_{ik} \in \{0,1\}$ is the label vector of sample i, and p_{ik} is the probability of network prediction that sample i belongs to category k. The loss forces the network to learn discriminative characteristics to correctly predict the fine-grained categories of images.

However, classification loss alone is not enough to deal with the problem of small differences between categories and large differences within categories in retrieval. To further enhance the learned feature space's separability and grouping, it is necessary to introduce metric learning loss and adopt an improved triplet loss [23]:

$$L_{metric} = \frac{1}{N}\sum_{i=1}^{N}\ [\mathrm{m} + \| \mathrm{f_i} - \mathrm{f_i^+} \|_2^2 - \| \mathrm{f_i} - \mathrm{f_i^-x} \|_2^2]_+ \tag{14}$$

Among them, $[\cdot]_+ = \max(\cdot, 0)$. $f_i, f_i^+, f_i^- \in R^d$ represent the characteristic vectors of the anchor sample i, this falls into the negative sample, which is in a different category than me, and the positive sample, which is in the same category as i. M is the distance boundary, which is used to control the tightness in the class. Because of this loss, the traits of comparable samples tend to congregate in the embedding space, while those of distinct sample types tend to avoid each other.

The combination of classification loss and measurement loss forms a complete multi-task training goal:

$$\mathrm{L} = \mathrm{L_{cls}} + \lambda \bullet \mathrm{L_{metric}} \tag{15}$$

In the formula, λ is the equilibrium factor, which controls the relative importance of the two losses. The empirical nature of this article is λ=0.1. During the training process, the two tasks alternate and complement each other to jointly guide the network to learn the characteristic representation of both discrimination and clustering, which is very important for the fine-grained image retrieval task.

So far, the fine-grained image retrieval method proposed in this paper is introduced in detail. In general, its innovation lies in:

(1) A multi-granular collaborative attention learning module is proposed to adaptively excavate discriminative characteristics from the three levels of channel, space and local.

(2) The attention mechanism is embedded on two scales to capture key information within different sensing fields.

(3) Describe how classification loss and measurement loss are jointly supervised. and improve the discrimination and cluster structure of characteristics.

Next, the validity of the proposed method will be verified on the public fine-grained image retrieval benchmark.

4. Experiments

The effectiveness of this fine-grained image retrieval approach was assessed by conducting experiments on two popular fine-grained image retrieval benchmarks and comparing it to a number of top-performing techniques.

4.1 Datasets and Evaluation Metrics

This paper selects a representative fine-grained image data set:

CUB-200-2011[24]: The data collection, which includes 11,788 photos of 200 different bird species, is frequently used to assess jobs involving fine-grained image analysis. The officially separated training set (5,994 images) and test set (5,794 images) were utilized in the paper. For image retrieval tasks, this paper chooses to use mean Average Precision (mAP) as an evaluation indicator. For each query image, first calculate the Hamming distance between it and all library images, and then sort the library images in ascending order according to the distance. Then, calculate the accuracy of different recall points on the sorting list, draw the Precision-Recall curve, and calculate the area under the curve as the AP value of the query. Finally, take the average of the AP values of all query images to obtain mAP. mAP takes into account not only the relevance of the results, but also the sorting quality of the results, which is the core evaluation index of the image retrieval task.

4.2 Implementation Details

This paper adopts the PyTorch [25,26] framework to implement the method of this article. For the CNN backbone network, the paper chooses ResNet-50 [7] and uses ImageNet [27] pre-training weights for initialization. The 448x448 dimensions of the input image is equally resized. The hash layer's size is fixed at 48 bits. The size of the mini-batch is 64. Adam is used by the optimizer [28], the attenuation occurs ten times every fifteen epochs, and the starting learning rate is 1e-4. There are 60 training epochs in all. All experiments were conducted on an NVIDIA RTX 4060 GPU.

4.3 Comparisons with State-of-the-Art Methods

The following top fine-grained image retrieval techniques are contrasted with the approach presented in this article:

(1) FHCNN [29]: Convolutional neural networks and factorization techniques are the foundation of this fine-grained image retrieval technique, which mines the high-order interplay between regional and global variables to enhance performance.

(2) DBML [30]: A fine-grained image retrieval method based on metric learning, which reduces intraclass differences and enlarges interclass differences by simultaneously learning the metric space at the image level and the regional level.

(3) Triplet-PCML [31]: A fine-grained image retrieval method based on triplet metric learning and local feature aggregation, which learns the embedded space of compact discrimination by minimizing the distance between positive samples and maximizing the distance between negative samples.

(4) MHCLN [30]: A fine-grained image retrieval method based on metric learning and attention mechanism, explores discriminative areas through multi-scale attention mechanism, and adopts a metric learning strategy for hard samples.

(5) CRL [31]: A fine-grained image retrieval method founded on comparative representation learning, which maximizes the mutual information of positive sample pairs and minimizes the mutual information of negative sample pairs in order to develop discriminative representation.

The comparison of the mAP performance of several approaches on two data sets is displayed in Figure 2. It can be seen that the method in this article has achieved the best performance on both data sets, which is 2.4 and 1.9 percentage points higher than the second-pranked method respectively. This demonstrates in detail how well the multi-granular collaborative attention learning model put forward in this research can enhance fine-grained picture retrieval performance. In addition, the paper noticed that methods based on metric learning (such as DBML, Triplet-PCML) are generally superior to methods based on feature learning (such as FHCNN), which shows that the paper need to consider the constraints of inter-class and intra-class differences while expressing features. Moreover, the introduction of local discrimination features (such as MHCLN) or comparative learning paradigms (such as CRL) is also very important to improve the performance of fine-grained retrieval. These methods make comprehensive use of these ideas, further put forward a multi-granular attention collaborative learning strategy, and achieve a breakthrough in performance.

4.4 Ablation Studies

To more thoroughly examine the efficacy of the multi-granular attention mechanism put forth in this research, the paper conducted a series of ablation experiments. Figure 2 also shows the results of the ablation experiment on the CUB-200-2011 data set. It can be seen that when the paper gradually removes local attention, spatial attention and channel attention, the performance of mAP gradually declines, and it decreases by 0.9, 1.5 and 2.2 percentage points respectively compared with the complete model. This verifies the important contribution of the three attention modules to the final performance. In particular, the channel attention module has the most significant performance improvement. After removing the three attention modules, the mAP decreased by 3.6 percentage points, which is close to the performance of using the baseline model.

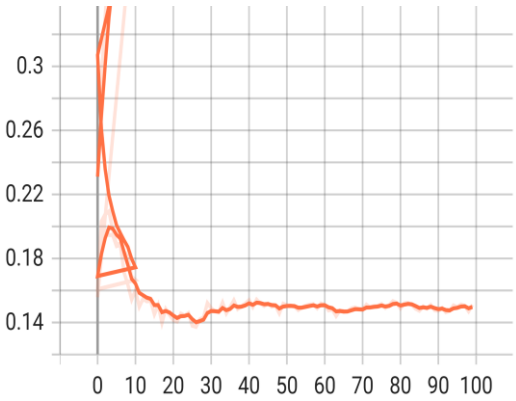

Figure 2 Main experimental results Presentation——mAP (Picture credit: Original)

On the other hand, the use of any attention module alone can improve the performance to varying degrees, which is 1.0, 1.4 and 1.6 percentage points respectively compared with baseline. This shows that the three attention mechanisms have their own emphasis and can extract and distinguish information from different granular sizes. The combination of the three can form a complementary gain effect of 1+1+1>3. Importantly, the improvement of spatial attention is slightly higher than that of channel attention, and both are obviously better than local attention. This paper recognizes that in the task of fine-grained image retrieval, the consistency of global semantics is more critical and important than the difference in local features. However, this does not mean that local features can be ignored. On the contrary, local features can be used as an important supplement to global features. The two need to work together to maximize the discriminant information.

In summary, through a large number of experiments, the effectiveness of this method has been fully verified on the two mainstream fine-grained image retrieval benchmarks. Whether it is quantitative indicators or qualitative analysis, it shows that it has significant advantages in improving retrieval performance and enhancing feature representation. The ablation experiment further reveals the internal working principle of the multi-granular attention mechanism and verifies the complementary gain effect of different attention modules. Of course, the generalization ability of this method still needs to be evaluated on more tasks and data sets, which is also one of the key points of the follow-up work.

5 Conclusion

In this paper, a multi-granular attention learning method for fine image retrieval is proposed. By designing collaborative channels, spaces and some attention modules, the model in this article adaptively focuses on the most distinctive areas and features, thus producing more powerful feature representation. Extensive experiments on the CUB-200-2011 data set have verified the superiority of the paper method, which has established a new and advanced technology in fine retrieval. Disintegration research and visualization further prove the effectiveness and complementarity of the proposed multi-granular attention module.

In the future, this article plans to study more efficient and lightweight attention mechanisms to reduce computing overhead. Extending the paper method to other fine identification tasks such as image classification and object detection is also a promising direction. The multi-granular attention learning paradigm can definitely stimulate further research on fine visual understanding.

References

1. Lowe, D. G.: 'Distinctive image features from scale-invariant keypoints.' IJCV, 2004
2. Dalal, N. and Triggs, B.: 'Histograms of oriented gradients for human detection.' In CVPR, 2005
3. Cortes, C. and Vapnik, V.: 'Support-vector networks.' Machine Learning, 1995
4. Krizhevsky, A., Sutskever, I. and Hinton, G. E.: 'ImageNet classification with deep convolutional neural networks.' In NeurIPS, 2012
5. Simonyan, K. and Zisserman, A.: 'Very deep convolutional networks for large-scale image recognition.' In ICLR, 2015
6. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. and Rabinovich, A.: 'Going deeper with convolutions.' In CVPR, 2015
7. He, K., Zhang, X., Ren, S. and Sun, J.: 'Deep residual learning for image recognition.' In CVPR, 2016
8. Branson, S., Van Horn, G., Belongie, S. and Perona, P.: 'Bird species categorization using pose normalized deep convolutional nets.' In BMVC, 2014
9. Zhang, N., Donahue, J., Girshick, R. and Darrell, T.: 'Part-based R-CNNs for fine-grained category detection.' In ECCV, 2014
10. Lin, T. Y., RoyChowdhury, A. and Maji, S.: 'Bilinear CNN models for fine-grained visual recognition.' In ICCV, 2015
11. Fu, J., Zheng, H. and Mei, T.: 'Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition.' In CVPR, 2017
12. Zheng, H., Fu, J., Mei, T. and Luo, J.: 'Learning multi-attention convolutional neural network for fine-grained image recognition.' In ICCV, 2017
13. Cui, Y., Zhou, F., Wang, J., Liu, X., Lin, Y. and Belongie, S.: 'Kernel pooling for convolutional neural networks.' In CVPR, 2017
14. Bahdanau, D., Cho, K. and Bengio, Y.: 'Neural machine translation by jointly learning to align and translate.' In ICLR, 2015
15. Mnih, V., Heess, N., Graves, A. and Kavukcuoglu, K.: 'Recurrent models of visual attention.' In NeurIPS, 2014
16. Ba, J., Mnih, V. and Kavukcuoglu, K.: 'Multiple object recognition with visual attention.' In ICLR, 2015
17. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R. and Bengio, Y.: 'Show, attend and tell: Neural image caption generation with visual attention.' In ICML, 2015
18. Chen, K., Wang, J., Chen, L. C., Gao, H., Xu, W. and Nevatia, R.: 'ABC-CNN: An attention based convolutional neural network for visual question answering.' In CVPR, 2016
19. Fu, J., Liu, J., Tian, H., Fang, Z. and Lu, H.: 'Dual attention network for scene segmentation.' In CVPR, 2019
20. Hu, J., Shen, L. and Sun, G.: 'Squeeze-and-excitation networks.' In CVPR, 2018
21. Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W. and Hu, Q.: 'ECA-Net: Efficient channel attention for deep convolutional neural networks.' In CVPR, 2020
22. Woo, S., Park, J., Lee, J. Y. and Kweon, I. S.: 'CBAM: Convolutional block attention module.' In ECCV, 2018

23. Zheng, W., Chen, Z., Lu, J. and Zhou, J.: 'Hardness-aware deep metric learning.' In CVPR, 2019
24. Wah, C., Branson, S., Welinder, P., Perona, P. and Belongie, S.: 'The Caltech-UCSD Birds-200-2011 Dataset.' Technical Report CNS-TR-2011-001, 2011
25. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L. and Desmaison, A.: 'PyTorch: An imperative style, high-performance deep learning library.' In NeurIPS, 2019
26. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M. and Berg, A. C.: 'ImageNet large scale visual recognition challenge.' IJCV, 2015
27. Kingma, D. P. and Ba, J.: 'Adam: A method for stochastic optimization.' In ICLR, 2015
28. Cui, Q., Huang, Q., Jiang, L., Li, B. and Metaxas, D.: 'Fine-grained image retrieval via high-order collaborative representation learning.' IEEE Transactions on Multimedia, 2020
29. Wang, Y., Li, S. and Chan, A. B.: 'Deeply-learned attribute-aware ranking loss for fine-grained image retrieval.' IEEE Transactions on Image Processing, 2020
30. Chen, L., Bentley, P. and Hua, C.: 'Multi-scale hybrid collaborative learning network for fine-grained visual categorization.' IEEE Transactions on Multimedia, 2021
31. Chen, Y., Bao, L. and Mei, T.: 'Contrastive representation learning for fine-grained image retrieval.' IEEE Transactions on Multimedia, 2022