International Journal on Semantic Web and Information Systems
Volume 18 • Issue 1

Fine-Grained Image Classification Based on Cross-Attention Network

Zhiwen Zheng, Yunnan Normal University, China
Juxiang Zhou, Yunnan Normal University, China*
Jianhou Gan, Yunnan Normal University, China
Sen Luo, Yunnan Normal University, China
Wei Gao, Yunnan Normal University, China

ABSTRACT

Due to the high similarity of fine-grained image subclasses, small inter-class changes and large intra-class changes are caused, which leads to the difficulty of fine-grained image classification task. However, existing convolutional neural networks have been unable to effectively solve this problem. Aiming at the above-mentioned fine-grained image classification problem, this paper proposes a multi-scale and multi-level ViT model. First, through data augmentation techniques, the accuracy of fine-grained image classification can be effectively improved. Secondly, the small-scale input and large-scale input of the model make the input image have more feature ex-pressions. The subsequent multi-layeredness effectively utilizes the results of the previous layer of ViT, so that the data of the previous layer can be more effectively used in the next layer of ViT. Finally, cross-attention allows the results of two scale inputs to be fused in a reasonable way. The proposed model is competitive with current mainstream state-of-the-art methods on multiple datasets.

KEYWORDS

Cross-Attention, Data Augmentation, Fine-Grained Images, Multi-Scale and Multi-Level

INTRODUCTION

The recently proposed Vision Transformer improves the performance of fine-grained image classification to a new level. For ordinary people, it is a challenging task to distinguish specific bird species and automobile brands, which usually requires a lot of professional knowledge to distinguish them accurately. Fine-grained visual classification (FGVC) aims to classify subcategories under coarse-grained large categories. For example, for bird species, it is necessary to identify not only the large category of birds in the image but also the specific bird species. In recent years, FGVC still has had many difficulties and challenges in its deep learning tasks because computers cannot effectively recognize the intra-class differences under subcategories.

With the continuous development of computer vision, the transformer, a natural language model, is also used. Although it is a natural language processing model, its ideas can also be used

DOI: 10.4018/IJSWIS.315747

*Corresponding Author

for image classification tasks in computer vision. To apply the transformer model to the image field, researchers divide the input image into small blocks with uniform size and with no overlap and then map these blocks through the full connection layer to obtain a tensor, corresponding to the token in the transformer model. At the same time, they also build a new tensor through the full connection layer. The dimension of this tensor is the same as that of the previously mapped tensor. This new tensor is called a class token and is mainly used for subsequent classification tasks. These tensors are then input into the transformer, and the subsequent operation is almost no different from the transformer model. Similarly, the corresponding Query(Q), Key(K), and Value(V) are constructed, and the corresponding steps are then performed. The final difference is that the class tokens that fused with other block information are subjected to an MLP classification operation.

The above model is called the Vision Transformer (ViT) model. However, this model has a big limitation, that is, the training dataset needs to be very large. If the dataset used for training is very small, the classification accuracy will even be worse than some ordinary convolutional neural networks (CNNs). At the same time, the input of this ViT model is relatively single, making the expression of data in the model very single. When the number of ViT layers is more than one, the model does not effectively use the information of the previous layer.

A multi-scale and multi-level ViT model is proposed to improve the classification performance of ViT. First, the accuracy of fine-grained image classification can effectively be improved using data enhancement technology. Second, the small-scale and large-scale inputs of the model lead to more feature expressions in the input image. When the model has multiple layers, the results of the previous ViT layer can be used effectively so that the data of the previous layer can be more effectively integrated and used in the next ViT layer. Finally, cross-attention allows the results of the two scales to be fused and transmitted reasonably. The proposed model is more competitive compared with the current mainstream advanced methods in a variety of datasets.

The main contribution of this paper is as follows. On the one hand, to enable the new model to learn more robust features, so as to effectively improve the generalization ability of the model, this paper uses data enhancement techniques that are different from the previous ones, such as inversion and rotation. It removes the current pixel model, considers other distinguished regions, and randomly erases all pixels in the region whose value of interest is greater than the threshold, with a probability to obtain destructive enhancement. On the other hand, a novel multi-scale and multi-level ViT model is proposed to obtain the feature representation of fine-grained images at different scales. The multi-layer combines the attention of the next layer and the attention of the upper layer to obtain new attention, so that each layer of attention can effectively be used to obtain enhanced attention information. The original single-scale input Vision Transformer is innovatively transformed into a multi-scale input using two different scale convolution kernels to obtain a richer image feature expression. In addition, for the first time, the cross-attention network is used in the field of fine-grained image classification in this paper to better fuse the obtained tokens and obtain a better feature representation.

RELATED WORK

Data Augmentation

Data augmentation (Mumuni et al., 2022) can be used as an efficient technique to expand a training set when there is not enough data to train a model. It increases the number of training data by introducing more data variances. An image can be divided into three images of different sizes through specific operations, such as cropping, but the objects in the picture are still the same, which does not affect the classification operation. They are the same for humans but different for DNN. Because the model's data and parameters are different, the performance on small dataset can be achieved as good as on large dataset, and this strategy has been demonstrated efficiently in many CV tasks, such as image classification, detection, and segmentation.

Multi-Scale

In multi-scale (Betzel et al., 2017), the different granularity of signals is sampled, and different tasks are achieved by observing different features at different scales. Since the 1950s, the multi-scale theory has been applied in the field of image processing and computer vision, successfully solving problems in optical character recognition and medical diagnosis, among others. In recent years, multi-scale analysis is mainly applied in image data augmentation, image fusion (Li et al., 2013), image feature enhancement (Zhao et al., 2018), and so on. The essential idea of multi-scale image representation is to study the characteristics and the relationship of images in different scales through decomposition to multiple scales. Moreover, the image information is analyzed in deep structure, increasing the accuracy of image feature description. At present, multi-scale analysis methods mainly include the image pyramid structure method, multi-scale analysis method based on wavelet transform, and multi-scale geometry analysis method (Beyerer et al., 2015).

Vision Transformer

Vision Transformer (Dosovitskiy et al., 2021) is the application of Transformer (Vaswani et al., 2017) in CV, which has inspired researchers. To input an image into Transformer, they clip an image into patches with the same size. These patches are then projected and embedded by FC to form a sequence, which is consequently fed into Transformer to achieve specific tasks. On medium and large datasets, Transformer lacks some translation invariance and local sensitivity compared with CNNs. Thus, when the training set is not large enough, its performance is worse than CNNs, i.e., the performance of Transformer is a few percentages lower than CNNs on a medium dataset. On the other hand, when the training set is large enough and can be well trained and be fine-tuned to a specific task using transferring learning (Weiss et al., 2016), Transformer might be comparable with or even outperform the state-of-the-art methods.

Motivation

The Vision Transformer network proposed in recent years is very easy to over-fit if the training data is insufficient, and some existing data enhancement techniques, such as random clipping and rotation, cannot produce a very good generalization effect on it. The method proposed in this paper is to use data enhancement techniques that are different from the above before putting the image into the model, to remove the current pixel model and consider other distinguishing areas, and randomly erase all pixels in the area with the value of interest greater than the threshold with specific probability to obtain a destructive enhancement. Through such operations, the Vision Transformer can have a better effect. The traditional Vision Transformer does not make effective use of the attention of each layer but simply repeats. To make effective use of the attention of each Vision Transformer, this paper combines the attention of the next layer and the attention of the upper layer to obtain new attention, so that the attention of each layer can effectively be used to obtain enhanced attention information. Feature Pyramid Networks for Object Detection (FPN) is a method that uses conventional CNN models to efficiently extract the features of each dimension in an image. This paper uses the input of a single-scale Vision Transformer to obtain the output results of different scales using convolution kernels of different sizes, so that it has more feature expressions and representation capabilities. Through such operations, the Vision Transformer can obtain stronger and richer semantic features. Through cross-attention fusion, the patch token and class token obtained under different scale branches in the Vision Transformer can effectively be fused together for better and more times. Therefore, to make more effective use of the multi-scale output results of the Vision Transformer, this paper uses the cross-attention fusion method to process the multibranch results of the ViT output and finally outputs the corresponding category tags for subsequent classification operations.

METHOD

A multi-scale and multi-level ViT model is proposed in this paper. It consists of attention-based data augmentation module, multi-scale and multi-level module, cross-attention fusion module and classification module. The attention-based data augmentation module augments the image destructively by clipping and erasing; the multi-scale and multi-level module takes the convolution whose kernel size is 12×12 and 16×16, its output is input to ViT; the cross-attention module fuses the class token of one branch and blocks tokens of the other branch; the classification module uses the fused representations to make prediction by FC and Softmax.

Attention-Based Data Augmentation

There are many methods of data augmentation in deep learning, such as image clipping, rotation and color distortion. In the past studies (Stefanowski et al., 2008), random data were also selected for pre-processing. For example, random image clipping can generate images with different translation and scale, thus improving the robustness of deep model. At present, commonly used image augmentation methods include cropping, translation, re-scaling, flipping and rotation. In this paper, the proposed method is different from the approaches mentioned above. It is similar with WS-DAN (Hu et al., 2019) which removes current pixel model and then considers other discriminatory area. Especially, for all images in a batch, all pixels in the area whose attention value is greater than the threshold are randomly erased with a specific probability. This kind of data augmentation belongs to destructive enhancement (as shown in Figure 1), which can effectively improve the accuracy of fine-grained image classification.

Multi-Scale and Multi-Level VIT

For traditional Vision Transformer, input image is clipped into evenly sized patches at a fixed scale. In this paper, an image is clipped with different scales. The image is divided into small sized patches and large sized patches for forming multi-scale feature representations used for image recognition. Small block denotes obtaining small patches through small scale branch, and large block denotes obtaining large patches through large scale branch. In order to enable two different scale inputs to be applied to the Vision Transformer, in this paper two branches' outputs are embed using two FC separately, the dimension of the outputs of the FC is set to 768, and a class token is set for each branch with same dimension, as shown in Figure 2.

The Vision Transformer usually inputs tokens into the Norm layer and multi-head layer in turn, then adds the output results to its own tokens, inputs them into the Norm layer and MLP layer in turn again, and then adds the output results to the previous output results, and finally, the result is input into the next Transformer Encoder, and the process will be iterated for L times. In this process, the attention of the current input is not well applied to the next layer, but simply repeated embedding. In order to effectively utilize the attention of each Vision Transformer, in this paper, the attention of the next layer is multiplied by the attention of the previous layer to obtain new attention, so that the

Figure 1. An Example of Attention-based Data Augmentation

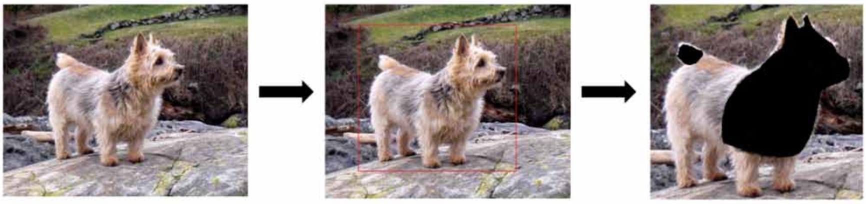
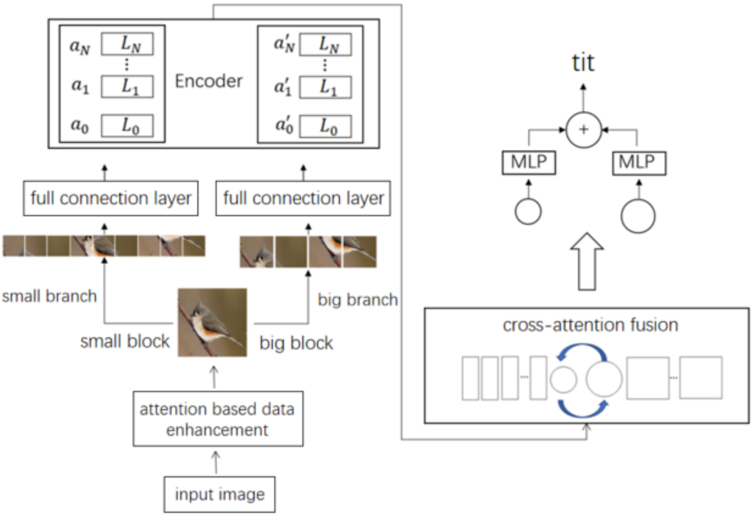


Figure 2. Multi-scale and multi-Level VIT model



attention of each layer can be effectively used to obtain enhanced attention information, achieving better results in the subsequent classification tasks.

The circle in the cross-attention fusion in Figure 2 represents the class token of the corresponding branch, and the quadrilateral box represents the patch token of the corresponding branch. The two circles above the cross-attention fusion are obtained from the class token extracted by two branches after cross-attention fusion.

Cross-Attention Fusion

For the small branch of the module, the class token interacts with the block token from the large branch through attention, and finally gets a new class token. For large branches, a new class token is obtained by interacting the corresponding class token with the block token from the small branch through attention, as shown in Figure 3.

There only apply fusion for small branch, and project the class token of the small branch via a FC then concatenate the output with large branch's class token, shown in formula (1):

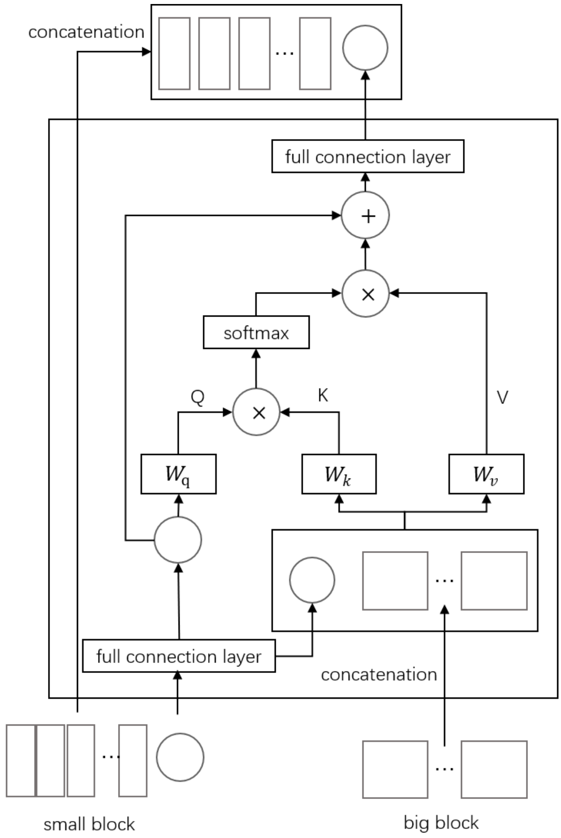
$$x'^{s} = \left[f^{s}\left(x^{s}_{cls}\right) || \, x^{l}_{patch}\right] \tag{1}$$

where $f^{s}\left(\cdot\right)$ is the FC used to project the class token of small branch to the corresponding dimension, x^{s}_{cls} is the small branch's tokens, and $||$ denotes concatenation. The whole module performs the cross-attention between the class tokens of the small branch and the block tokens of the large branch. Cross-attention can be represented mathematically as the following formulas:

$$q = x'^{s}_{cls} W_{q}, k = x'^{s} W_{k}, v = x'^{s} W_{v} \tag{2}$$

$$A = softmax\left(qk^{T} \, / \, \sqrt{C \, / \, h}\right), CA\left(x'^{s}\right) = Av \tag{3}$$

Figure 3. The framework of cross-attention fusion



where $W_q, W_k, W_v \in R^{C\times(C/h)}$ are learnable parameters, C is the embedding dimension, h is the number of attention heads, $x_{cls}^{'s}$ is equivalent to f^s (x_{cls}^{s}). In this paper we only use CLS token, so the computational and memory complexity of the attention diagram generated in cross-attention is linear, which is completely different from the quadratic generated in full-attention, the linear process makes the whole flow more efficient.

The unmarked circle in figure 3 represents a class token, and the unmarked quadrilateral box represents a patch token. In the self-attention, this paper also uses multiple heads like Transformer. Using multi heads in Cross-attention is denoted as MCA, but there is no feedforward neural network after MCA. The output z^s of small branch in the cross-attention model can be expressed as the following formula:

$$y_{cls}^{s} = f^{s}\left(x_{cls}^{s}\right) + MCA(LN([f^{s}\left(x_{cls}^{s}\right) \,||\, x_{patch}^{l}])) \tag{4}$$

$$z^{s} = [g^{s}\left(y_{cls}^{s}\right) \,||\, x_{patch}^{s}] \tag{5}$$

where $f^s\left(\cdot\right)$ is a projecting function, $g^s\left(\cdot\right)$ is a reverse projection function, LN is Layer Normalization, and MCA is cross-attention function.

Loss Function

The loss used in this paper is the BCEWithLogitsLoss loss function, which is BCELoss with a sigmoid activation function. Because the last layer of this model does not add a sigmoid activation function, the BCEWithLogitsLoss loss function is used. This loss function is as the following formula (6):

$$L = -\frac{1}{C}\sum\left(y_C \times \ln_{x_C} + \left(1 - y_C\right) \times \ln\left(1 - x_C\right)\right) \tag{6}$$

where C is the number of categories classified, y is the category label of the target, and x is the probability value of each category predicted by the model. The dimension of L output from this loss function is $batch \times C$. Batch is the number of samples in each batch. Each batch output has C values, corresponding to the probability values predicted by C categories, and each probability value is distributed between 0 and 1.

EXPERIMENTS AND ANALYSIS

The proposed model is verified on CUB-Birds (Wah et al., 2011),Stanford Dogs (Khosla et al., 2011),Stanford Cars (Krause et al., 2013) and Plant Pathology (Mwebaze et al., 2019) to evaluate its effectiveness. The experiments are implemented under Ubuntu 16.04 using python 3.8.12 and Pytorch 1.7.1 with an NVIDIA GeForce GTX 2080.

Datasets

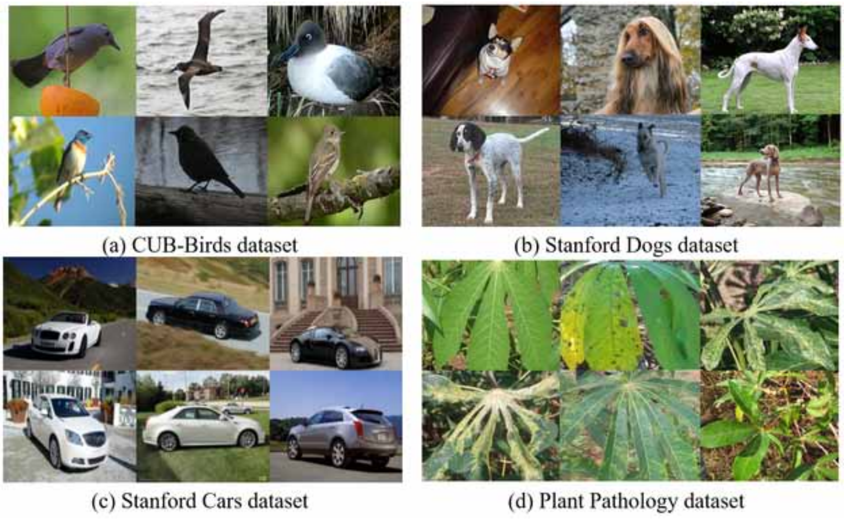
The Cub-Birds dataset contains more than 10,000 images of different species of birds that are clearly named and defined biologically. The names of these species are available through online field guides. Each image in this dataset contains the bounding box and corresponding label information of the identified birds, and every image contains the whole figure of birds. Stanford Dogs dataset contains more than 20,000 dog images, and each image in this dataset contains the bounding box and corresponding tag information of identified dogs. There are a wide variety of dog breeds included in the images, including Alaskan Malamute, Chihuahua, golden retriever, etc. There are about one to two hundred pictures of each breed. The Stanford Cars dataset contains more than 16,000 car images. It has 196 kinds of cars, including General Motors, Ford Motor Company, Lincoln and other car types.

Plant Pathology dataset contains images of the Cassava disease in Africa. There are altogether five types of African Plant disease images, among which 9,436 images are annotated and 12,595 cassava leave images are unlabeled. When using this dataset, unlabeled images can be used as training data. Some examples from these datasets are shown in Figure 4.

Details of Experiments

The model proposed in this paper requires the image size to be 384×384 and multi-scale input, which means the image should be divided into two parts: small scale and large scale. The convolution kernel size of small scale is 12×12; its stride and padding are 12 and 0 respectively. The convolution kernel size of large scale is 16×16, its stride and padding are 16 and 0 respectively. Stride is set to be consistent with the size of the convolution kernel for both scales, in order to evenly clip the image. The dimension of the Vision Transformer used to receive small scale input is set to 368, the dimension of the Vision Transformer used to receive large scale input is set to 768, the output dimension of both Vision Transformers is 512, and the number of heads is set to 12. Layer is 12. The layer of the cross-attention encoder is 3. Batch size is 16, and learning rate is 1E-4. To efficiently save memory during training, the function checkpoint_sequential was used in torch.utils.checkpoint, where checkpoint

Figure 4. Examples from Public Datasets


(a) CUB-Birds dataset
(b) Stanford Dogs dataset
(c) Stanford Cars dataset
(d) Plant Pathology dataset

is set to True and checkpoint_nchunks to 2. The larger the checkpoint_nchunks are, the less video memory is used, but the amount of video memory used tends to plateau.

Comparison With Other Models

In order to verify the validity of the model, experiments were carried out on Cub-Birds, Stanford Dogs, Stanford Cars and Plant Pathology respectively.

It can be seen from Table 1 that in the Stanford Dogs dataset, the accuracy of the method proposed in this paper using the ViT-B /16 network was about 6% higher than that of the baseline method RA-CNN (Fu et al., 2017). The network adopted by the baseline method RA-CNN is traditional VGG-19 (Simonyan et al., 2014), which contains 16 convolutional layers and 3 full connected layers. Although it is a traditional deep convolutional network model, it had 87.3% accuracy. The proposed method was about 2% more accurate than the most advanced ViT models. In recent years Vision Transformer model has not only performed well in image classification tasks, but also had good performance in fine-grained image classification tasks. It applies the idea of Transformer model which is originally used in NLP to image classification task, and the accuracy of the proposed model on Stanford Dogs dataset was higher than that of the most advanced methods; it has proved the validity of the model

Table 1. Comparison of different methods and network models on the Stanford Dogs dataset

Method	Model	Accuracy (%)
MaxEnt (Dubey et al., 2018)	DenseNet-161 (Huang et al., 2017)	83.6
FDL	DenseNet-161	84.9
RA-CNN	VGG-19	87.3
Cross-X (Luo et al., 2019)	ResNet-50	88.9
API-Net (Zhuang et al., 2020)	ResNet-101	90.3
ViT	ViT-B/16	91.7
Proposed	ViT-B/16	93.6

proposed in this paper. It can also be seen from the table that using ViT-B /16 network showed better competitiveness compared with traditional ResNet (He et al., 2016) and VGG.

To verify the validity of our model on the CUB-Birds and Stanford Cars datasets, we compare our model with the most advanced methods on the two datasets, it's shown in Table 2.

It can be seen from the Table 2 that the model proposed in this paper is better compared with the baseline method and the current most advanced method. The accuracy of the model proposed in this paper on the CUB-Birds dataset was about 6% higher than that of the baseline method RA-CNN, and about 2% higher than that of the current optimal method ViT. Accuracy on the Stanford Cars dataset was about 2% higher than the baseline method, RA-CNN, and about 2% higher than the current best method. It can be seen from the table that increasing the number of network layers, VGG-19 was greatly improved compared with VGG-16, but it was still not as good as ResNet-50 and DenNet-161 model performance. Both ReSNet-50 and DenNet-161 have solved the problem of gradient disappearance in deep networks, so that the performance of both models is also impressive. The backbone network used in the model proposed in this paper is ViT-B /16, which applies the idea of Transformer which is used in NLP to image classification task and shows excellent performance. In this study the model was optimized and improved on the basis of the backbone network and showed very good accuracy.

In order to verify the validity of the proposed model on the Plant Pathology dataset, it was compared with the most advanced methods on the dataset, as shown in Table 3. This dataset differs from the general datasets. It has very few categories, but numerous image data of each category. The table shows that the traditional fine-grained image classification algorithm ResNet had a classification accuracy of around 90%, while EfficientNet (Tan & Le, 2019) had a slightly higher accuracy than ResNet, which is a new method for scaling networks. It uses a simple and efficient compound coefficient to enlarge the network and improve network performance from the input image resolution, network depth, and network width. The accuracy of the current most advanced ViT method was approximately 92%, while the improved ViT proposed in this paper was nearly 4% higher than the baseline method ResNet and approximately 2% higher than the most advanced ViT method, demonstrating the validity of the model presented in this paper.

Table 2. Comparison with Mainstream Methods on CUB-Birds Dataset and Stanford Cars Dataset

Method	Model	CUB-Birds(%)	Stanford Cars(%)
VGG-16	VGG-16	77.8	85.7
ResNet-101	ResNet-101	83.5	91.2
Inception-V3(Szegedy et al., 2015)	Inception-V3	83.7	90.8
RA-CNN	VGG-19	85.3	92.5
MaxEnt	DenseNet-161	86.6	93.0
Cross-X	ResNet-50	87.7	94.6
DCL(Chen et al., 2019)	ResNet-50	87.8	94.5
API-Net	DenseNet-161	90.0	95.3
WS-DAN	Inception v3	89.4	94.5
MMAL-Net (Zhang et al., 2021)	ResNet-50	89.6	95.0
ViT	ViT-B/16	89.4	92.8
Proposed	ViT-B/16	91.2	94.8

Table 3. Comparison of Different Methods on the Plant Pathology Dataset

Method	Model	Accuracy(%)
ResNet	50	89.2
EfficientNet	B0	90.1
EfficientNet	B2	90.4
ViT	ViT-B/16	91.7
Proposed	ViT-B/16	93.1
ResNet	50	89.2
EfficientNet	B0	90.1

CONCLUSION

In this paper, a multi-scale and multi-level Vision Transformer for fine-grained image was proposed. First, attention-based data augmentation has better destructive enhancement than traditional data augmentation, which promotes the final classification performance. Secondly, multi-scale makes the input image representations more expressive. Multi-level can effectively use the attention of each layer to obtain enhanced attention information. The cross-attention fusion mechanism performs a reasonable interactive fusion of each scale branch, and finally obtains a tensor combining the results of the two scales, which can boost the classification accuracy. This algorithm had 93.6%, 91.2%, 94.8%, and 93.1% recognition accuracy on Stanford Dogs, cub-birds, Stanford Cars, and Plant Pathology datasets respectively.

COMPETING INTERESTS

The authors of this publication declare there are no competing interests.

FUNDING AGENCY

This research was supported by the National Natural Science Foundation of China (Grant number of 62166050), Yunnan Fundamental Research Projects (Grant numbers of 202201AS070021 and 202101AT070095), Yunnan Innovation Team of Education Informatization for Nationali-ties, and Scientific Technology Innovation Team of Educational Big Data Application Technology in University of Yunnan Province.

REFERENCES

Betzel, R., & Bassett, D. (2017, October). Multi-scale brain networks. *NeuroImage*, *160*(15), 73–83. doi:10.1016/j.neuroimage.2016.11.006 PMID:27845257

Beyerer, J., León, F., & Frese, C. (2015). *Image Pyramids, the Wavelet Transfm and Multiresolution Analysis*. Machine Vision.

Chen, Y., Bai, Y., Zhang, W., & Mei, T. (2019). Destruction and construction learning for fine-grained image recognition. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T. . . ., Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. *Computer Vision and Pattern Recognition*.

Dubey, A., Gupta, O., Raskar, R., & Naik, N. (2018). Maximum-entropy fine grained classification. Advances in Neural Information Processing Systems, 31.

Fu, J., Zheng, H., & Mei, T. (2017). Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Hu, T., Qi, H., Huang, Q., & Lu, Y. (2019). *See better before looking closer: Weakly supervised data augmentation network for fine-grained visual classification.* arXiv preprint arXiv:1901.09891.

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Khosla, A., Jayadevaprakash, N., Yao, B., & Li, F.-F. (2011). Novel dataset for fine-grained image categorization: Stanford dogs. *Proc. CVPR workshop on fine-grained visual categorization (FGVC).*

Krause, J., Stark, M., Deng, J., & Fei-Fei, L. (2013). 3d object representations for fine-grained categorization. *Proceedings of the IEEE international conference on computer vision workshops*.

Li, S., Kang, X., & Hu, J. (2013). *Image Fusion With Guided Filtering*. Academic Press.

Luo, W., Yang, X., Mo, X., Lu, Y., Davis, L. S., Li, J., . . . Lim, S.-N. (2019). Cross-x learning for fine-grained visual categorization. *Proceedings of the IEEE/CVF international conference on computer vision*.

Mumuni, A., & Mumuni, F. (2022). Data augmentation: A comprehensive survey of modern approaches. *Array, 16*.

Mwebaze, E., Gebru, T., Frome, A., Nsumba, S., & Tusubira, J. (2019). *iCassava 2019 fine-grained visual categorization challenge.* arXiv preprint arXiv:1908.02900.

Simonyan, K., & Zisserman, A. (2014). *Very deep convolutional networks for large-scale image recognition.* arXiv preprint arXiv:1409.1556.

Stefanowski, J., & Wilk, S. (2008). *Selective Pre-processing of Imbalanced Data for Improving Classification Performance*. Data Warehousing and Knowledge Discovery. doi:10.1007/978-3-540-85836-2_27

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., & Rabinovich, A. et al. (2015). Going deeper with convolutions. *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Tan, M., & Le, Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. *International Conference on Machine Learning*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., & Polosukhin, I. et al. (2017). Attention Is All You Need. *Computation and Language.*

Wah, C., Branson, S., Welinder, P., Perona, P., & Belongie, S. (2011). *The caltech-ucsd birds-200-2011 dataset*. Academic Press.

Weiss, K., Khoshgoftaar, T., & Wang, D. (2016). A survey of transfer learning. *Journal of Big Data*, *3*, 9.

Zhang, F., Li, M., Zhai, G., & Liu, Y. (2021). Multi-branch and multi-scale attention learning for fine-grained visual categorization. *International Conference on Multimedia Modeling*.

Zhao, T., & Wu, X. (2018). Image Saliency Detection with Low-Level Features Enhancement. *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, 408–419. doi:10.1007/978-3-030-03398-9_35

Zhuang, P., Wang, Y., & Qiao, Y. (2020). Learning attentive pairwise interaction for fine-grained classification. *Proceedings of the AAAI Conference on Artificial Intelligence*.

Zhiwen Zheng received the B.S. degree in computer science and technology at Jilin Jianzhu University and obtained the M.S. degree from Yunnan Normal University, majoring in computer technology. He is currently a machine learning engineer at Binjiang Institute of Zhejiang University. His research interest is computer vision.

Juxiang Zhou received her PhD's degree in 2019 from Dalian University of Technology, Dalian, China. Currently, she is an assistant research fellow in Yunnan Normal University. Her research interests include Image and video understanding, intelligent image processing, intelligent education. She has published more than 30 academic papers (including acceptance in international journals including IEEE Transactions on Cybernetics, Pattern Recognition, Information science, International Journal of Machine Learning and Cybernetics, Multimedia Tools and Applications and so on, authorized 9 national invention patents, and awarded the Yunnan Province Postdoctoral Excellence Award and the China Computer Society (CCF) Innovation Technology Award.

Jianhou Gan received Ph.D. degree in computational metallurgy from Kunming University of Science and Technology, China, in 2016. Currently, he is a professor and doctoral supervisor at Yunnan Normal University. He was awarded the Yunnan Province "Ten thousand Plan" industrial technology leading talent, young and middle-aged academic and technical leader. He is currently Vice President of Yunnan Normal University, the vice director of Key Laboratory of Educational Informatization for Nationalities, the director of the Key Laboratory of Intelligent Education in Yunnan Province and the leader of the innovation team in Yunnan Province. His research interests include intelligent information processing, advanced database technology, education informatization and intelligent education. He has undertaken more than 10 national scientific research projects such as the National Science and Technology Plan, the National Natural Science Foundation and the National soft science project. He has published more than 80 papers, won 5 Yunnan Provincial Science and Technology Progress Awards, and authorized 12 national invention patents.

Sen Luo received the B.S. degree in Electronic Information Engineering from Southwest Jiaotong University, Chengdu, China, in 2020. He is currently studying for the M.S. degree in Computer Technology in Yunnan Normal University, Yunnan, China. His research interest is computer vision.

Wei Gao obtained a PhD degree in the mathematical department at Soochow University, China in 2012. He is professor in School of Information Science and Technology, Yunnan Nor-mal University from 2019. He worked as post doctor in Department of Mathematics, Nanjing University from January 2017 to December 2018, and Department of Mathematics and Science Education, Harran University from November 2019 to October 2020. His research interests are Graph Theory, Statistical Learning Theory, Discrete Dynamic System, Nonlinear Partial Differential Equation, Theoretical Chemistry, Artificial Intelligence, etc. He is a committee member of China Society of Industrial and Applied Mathematics (CSIAM) Graph Theory and Combinatorics with Applications Committee, and the chair of ICED 2017, ISGTCTC 2018, and ISTCS 2019.