Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 1, No. 3, 2022

Fine-grained Image Recognition Method using Discriminative Region-based Data Augmentation

Jingyuan He[1, 2], Bailong Yang[1, *]

[1] Xi'an Research Institute of Hi-Tech, Xi'an 710025, Shaanxi, China
[2] School of Mathematics and Computer Science, Yan'an University, Yan'an 716000, Shaanxi, China.

[*] **Corresponding author**: Bailong Yang.

Abstract: In order to reduce the complexity of the network model and improve the accuracy of image recognition, a fine-grained image recognition method using discriminative region-based data augmentation is proposed. The method obtains the discriminative regions of the image through the attention mechanism, and then performs diversity data augmentation based on the discriminative regions, including region crop, region drop and region mix, and then uses the generated augmented samples to train the network model, the backbone network of the model is ResNet50. The proposed method is tested on 4 commonly used fine-grained image datasets CUB-200-2011, Stanford Cars, FGVC Aircraft and Stanford Dogs, and achieves high accuracy. The experimental results show that the proposed method can improve the localization ability and feature extraction ability of the model for discriminative regions, and it is more lightweight and easier to implement.

Keywords: Fine-grained image recognition; Data augmentation; Attention mechanism; Discriminative region.

1. Introduction

For image recognition, the quality of data preprocessing plays a decisive role in model training and final image recognition results. The data augmentation object of traditional image recognition is usually aimed at the global information of the data, but it may not be able to improve the performance for fine-grained image recognition. This is mainly due to the inter-class similarity and intra-class similarity of fine-grained image data. determined by the differences. Fine-grained image recognition requires the model to capture some subtle and distinctive local features, so data augmentation mainly targets local feature information. Mixed Sample Data Augmentation (MSDA) algorithm is a widely used algorithm in data augmentation technology. This algorithm can effectively improve model performance and network generalization ability for traditional image recognition . In recent years, the attention mechanism has been widely used in fine-grained image recognition. The attention mechanism can effectively locate the discriminative regions of the input fine-grained image, and then extract the discriminative features for recognition . Facing the wide application of attention mechanism and data augmentation technology in fine-grained image recognition, and the shortcomings of existing data augmentation methods, this paper proposes a fine-grained image recognition method based on discriminative region data augmentation.

Data augmentation can increase the training data set on the basis of the existing data set, making the data set more abundant and diverse, and also improving the feature extraction ability and robustness of the training model . There are many methods of data augmentation. Traditional methods such as rotation, cropping, erasing, translation, scaling, perturbation, grayscale, illumination transformation, Gaussian noise, etc., are aimed at single-sample data. The random cropping data augmentation method proposed by Krizhevsky et al. can effectively improve the accuracy, and then a series of image translation data augmentation methods appear. Cubuk et al. proposed Auto Augmentation strategy to automatically search for the unique strategy required in the model training process. This strategy can effectively improve the classification and recognition effect of the network model, and the generalization ability is also has been greatly improved. The Random Erasing data augmentation method proposed in the literature has a great improvement in performance improvement in image classification, re-identification and target detection. The multi-sample mixed data augmentation algorithms are image-oriented and feature space-oriented. The most widely used algorithms are Mixup and Cutmix. A new data augmentation algorithm is proposed after the augmentation algorithm is fused. The algorithm generates new data samples by cropping and merging, and improves the loss function at the same time. In order to solve the problem of low recognition performance that the Cutmix algorithm may generate ineffective and ambiguous images, Kim et al. used part localization-aware CutMix to generate images. Adaptive pairwise margin loss to improve the fine-grained image recognition accuracy of joint optimization. On the basis of improving the general MSDA, Wei Hua proposed an Oriented Pair interaction mixing for augmenting fine-grained image recognition data via Euclidean distance measure (OPairIM), experiments on three public datasets verify the rationality and effectiveness of the algorithm. Drawing on the idea of data augmentation algorithm based on attention mechanism, literature proposes a multi-sample data augmentation algorithm based on spatial and channel attention for fine-grained image classification, which can effectively distinguish image features while making the training data more diverse, and further solve the problem of overfitting of network models. Hataya et al. proposed a Meta Approach to Data Augmentation Optimization, which can improve the performance of various image classification tasks including fine-grained image recognition tasks. Reference proposed an Attribute Mix strategy for data augmentation of fine-grained image samples. Extensive experiments show that this method can improve the recognition performance without increasing the inference budget. Reference proposes a semantic and data mix

(SADMix) enhancement strategy for fine-grained visual classification. The experimental results of SADMix on three commonly used fine-grained image datasets demonstrate the effectiveness of the method. Yu Wenchang proposed a data augmentation method based on weakly supervised discriminative region localization, which improved the model's localization ability and feature extraction ability on discriminative regions.

2. Methodology

2.1. Overall framework

MSDA can make the samples have more additional image information on the basis of the existing data set, which increases the feature extraction ability and generalization ability of the network model to a certain extent. Part information of fine-grained image data is distinguished from coarse-grained image data by its complex background, sensitivity, pose diversity, large illumination change, and occlusion. In order to reduce the complexity of the network model and improve the accuracy of image recognition, and ensure the model's ability to locate and extract discriminative regions, this paper proposes a fine-grained image recognition method using discriminative region-based data augmentation combined with the attention mechanism. Figure 1 presents the overall framework of fine-grained image recognition method.

As shown in Figure 1, the features of the initial input image samples are extracted by the feature extraction module, and then the obtained feature map is generated from the feature map by the attention module. The spatial and channel bilinear pooling is applied to the attention map, and then feature vectors are extracted from it. Normalize the extracted feature vector, then input it into the classifier for classification, and finally determine the category of the input fine-grained image according to the label value. The attention map and classifier of the initial sample contain some relevant information about the significant discriminative regions of the input image, so some discriminative regions of the initial sample can be located through the attention map and classifier weights. In this paper, multiple augmented samples are generated based on discriminative region pairs for each initial sample using a diversity data augmentation operation. The generated augmented samples use the same feature extraction module as the initial samples to extract feature maps. The feature extraction module here uses a CNN network that uses a wider embedded attention mechanism for feature extraction.

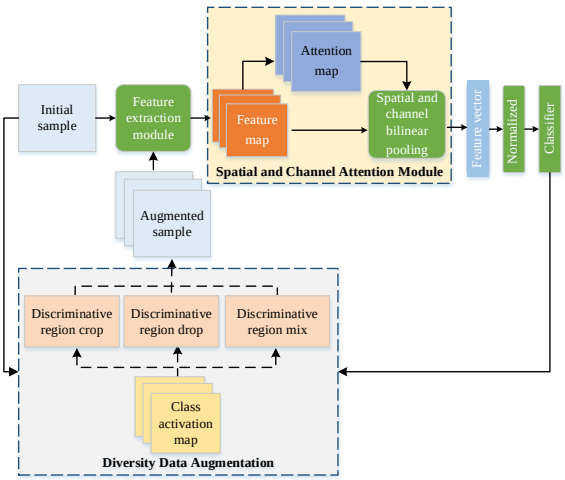

Figure. 1 Overall framework of fine-grained image recognition method

2.2. Discriminative region locating

In fine-grained image recognition, the CNN network performs feature extraction on the input image, and then filters a large amount of extracted feature information to select those that are useful for fine-grained image recognition and can be used to distinguish fine-grained images. Sexual characteristic information. However, the existing fine-grained image recognition network is not very capable of screening these discriminative features, and the screened feature information is not necessarily helpful for image recognition. In order to filter out more discriminative feature information We will add attention mechanism to the CNN network. In this paper, the discriminative regions in the input fine-grained image are precisely located by assigning weights to different channels and performing addition operations.

Suppose the current sample belongs to the jth class. Suppose the feature map output by the last convolutional layer of the CNN network is $F=\{F_1,F_2,\cdots,F_c\}$, where $F_i \in R^{H\times W}, i\in[1,C]$. C, H and W represent the number of channels, height and width of the feature map, respectively. The final input to the classifier is the one-dimensional vector of the feature map expansion, which is $f \in R^{(CHW)\times 1\times 1}$. In order to reduce the feature dimension and make it easier to calculate the contribution of each channel to the classification result, this paper uses spatial and channel bilinear pooling to extract the information of each channel in the feature map, and reduces the feature map into an $f \in R^{C\times 1\times 1}$ vector. After modulo length normalization operation, input the classifier for classification.

The weight of each element of each category feature vector in the fine-grained image determines the importance of the corresponding channel in the discriminative feature map for that category. In this paper, the fully connected layer without bias term is used to classify fine-grained images, the number of predefined categories is set to N, the fully connected layer is represented as a $W \in R^{C\times N}$ matrix, then the w_{ij} element represents the $i(i\in[1,C])$ th element of the feature vector and the $j(j\in[1,N])$ th category score is connected the weight of. The i th element of the feature vector represents the global information of the i th channel of the feature map F, and the feature map is weighted and summed according to formula (1):

$$A_j=\sum_{i=1}^{C} w_{ij}F_i,\quad (A_j\in R^{H\times W}) \tag{1}$$

where A_j is the class activation map of the j th category, and F_i represents the i th channel of the feature map F. In this paper, the class activation map of the class with the largest score output by the classifier will be used for the localization of the discriminative region according to the above method.

The class activation map calculated according to formula (1) needs to be upsampled to restore it to the same size as the initial image, and then a threshold is set to binarize the upsampled class activation map to obtain the discriminative region. mask to complete the localization of discriminative regions.

2.3. Data augmentation

Each channel of the attention map obtained by the above method corresponds to the discriminative region of the object to be recognized in the initial image. The data augmentation

of the original samples is carried out according to the accurately located discriminative regions of the attention map, and the data augmentation is mainly to crop, drop, and mix the discriminative regions.

Discriminative region crop is to crop the discriminative region in the initial image from the attention activation map, crop out part of the region, and then enlarge the cropped local region to the same size as the initial image, which is input to the CNN network as an augmented sample for training. classification model. Enlarging the cropped local area can effectively avoid the interference of useless information in other areas, and better perform feature extraction on these areas with rich category information. The process of region clipping is

Discriminative region drop refers to randomly erasing the discriminative regions of some initial images, and inputting the erased images as augmented samples into the CNN network for training the classification model. At the same time, in order to better improve the generalization ability of the model, the CNN network is required to It can extract various discriminative features of more initial images to the greatest extent.

Discriminative region mix refers to mixing the discriminative regions of one image and the non-discriminative regions of the other image from two different categories. The non-discriminatory regions in the image basically belong to the background region. Therefore, the region mixing method proposed in this paper can expand the background of all categories to a greater extent, making the backgrounds of all categories more abundant and changeable, thereby further improving the classification model. Feature extraction capability in complex and changeable real-world scenarios.

2.4. Loss function

The formula for the softmax cross-entropy loss is shown in (2).

$$L_s = -\frac{1}{M}\sum_{m=1}^{M}\log\frac{e^{g_{k_m}}}{\sum_{i=1}^{C}e^{g_i}} \tag{2}$$

Where M is the number of training samples, C is the preset number of categories in the recognition task, g is the score output by the classifier, k_m represents the serial number of the category of the m th sample in the C categories, and g_{k_m} represents the k_m th category.

The loss function used in this paper is as follows:

$$L\left(x_{raw}, x_{aug}, y\right) = L_s\left(x_{raw}, y\right) + \lambda L_s\left(x_{aug}, y\right) \tag{3}$$

Where x_{raw} represents the original sample, x_{aug} represents the augmented sample, y represents the class label of the sample, and $L_s\left(x_{raw}, y\right)$ and $L_s\left(x_{aug}, y\right)$ are the losses of the original sample and the augmented sample, respectively. The reason why the weight λ is assigned to the loss of the augmented sample here is because of the inevitable noise in the generated augmented sample, and $\lambda \in (0,1)$.

In order to ensure the accurate localization of the discriminative regions of the image by the class activation map and the effectiveness of the data augmentation method, we need to normalize the modulo length to a relatively large value before the feature vector is input to the classifier. This value is a hyperparameter , set to s , as shown in Equation (4).

$$\tilde{f} = \frac{sf}{\|f\|_2} = \frac{sf}{\sqrt{\sum_i f^2{}_i + \sigma}} \tag{4}$$

where f is the eigenvector, $\tilde{f}$ is the normalized eigenvector, and σ is a small positive number to prevent the denominator from being 0.

Therefore, when calculating the loss function in this paper, it is necessary to first perform the modulo-length normalization operation on the feature vector according to formula (4), and then calculate the loss function according to (3).

3. Results and discussion

3.1. Datasets

In this paper, four commonly used fine-grained image datasets are used to evaluate the experimental performance. These four datasets are CUB-200-2011 , Stanford Cars , FGVC Aircraft and Stanford Dogs [18]. These 4 datasets provide not only the category labels of the objects to be classified, but also the location labels of the bounding boxes and keypoints. In this paper, only the class labels of the images are used in the model training and testing phases to locate the discriminative regions of the objects to be classified. The relevant information of these four fine-grained image datasets is shown in Table 1.

Table 1. Fine-grained image datasets

Dataset	Category	Number of images	Train	Test	Year	Object
CUB-200-2011	200	11788	5994	5794	2011	bird
Stanford Cars	196	16185	8144	8041	2013	car
FGVC-Aircraft	102	10200	6667	3333	2012	aircraft
Stanford Dogs	120	20580	12000	8580	2011	dog

3.2. Experimental settings

In the experiment, the input fine-grained image needs to be preprocessed first, the image size is processed to 448 × 448, the backbone network is ResNet50, and the pre-trained weights on ImageNet are loaded for initialization. Set the number of channels M of the attention map to 512.

When performing data augmentation, discriminative region localization, region crop, region drop and region mix need to be performed separately. Among them, when the discriminative region is located, after the weighted summation of different channels of the feature map, the response values at all positions need to be normalized to be between [0, 1]. When calculating the loss function, the value of λ in formula (3) is 0.5, and the value of the modulo length s in the normalization of modulo length in formula (4) is 100. The Stochastic Gradient Descent (SGD) algorithm is used when updating the weights of the network, and the momentum coefficient in SGD is set to 0.9, and a weight decay term is added to prevent overfitting, and the weight decay parameter is set to 1×10-5. The initial learning rate of the network is set to 1×10-3, and it is decayed in stages during the training process, and the learning rate is multiplied by 0.9 for every two rounds of training. The training sample batch size is set to 64.

3.3. Performance analysis and discussion

In this paper, the comparison results of the recognition accuracy of the discriminative region-based diversity data augmentation algorithm and the existing algorithm on four commonly used fine-grained image datasets are shown in Table 2. The backbone networks of all methods are ResNet50. The classification accuracy of the other methods in Table 2 comes from the papers that proposed these methods, - indicates that the method was not tested on the dataset in the original paper.

Table 2. Performance comparison of data augmentation methods

Method	CUB-200-2011	Stanford Cars	FGVC-Aircraft	Stanford Dogs
Baseline	85.5	93.3	91.2	87.3
Cutout	83.6	93.8	91.2	-
Mixup	86.3	94.2	91.5	-
Cutmix	86.1	94.5	91.7	84.8
OpairIM	87.7	94.8	92.8	-
SADMix	88.2	94.4	93.1	-
Our's	89.3	95.0	93.5	88.2

Record the classification accuracy obtained by testing four commonly used fine-grained image datasets on the trained model proposed in this paper, and then compare the experimental results with the classification accuracy of some existing advanced fine-grained image recognition methods. The experimental results were compared with each other, and the results are shown in Table 3. The classification accuracy of the other methods in Table 3 comes from the papers that proposed these methods, - indicates that the method was not tested on the dataset in the original paper, and the bold data represents the highest classification of all methods on the dataset Accuracy.

Table 3. Performance comparison with the state-of-the-art methods on fine-grained image datasets

Method	Backbone	CUB-200-2011	Stanford Cars	FGVC-Aircraft	Stanford Dogs
RA-CNN	VGG19	85.3	92.5	88.4	87.3
MA-CNN	VGG19	86.5	92.8	89.9	-
MAMC	ResNet50	86.2	92.8	-	84.8
DFL-CNN	ResNet50	87.4	-	-	-
NTS-Net	ResNet50	87.5	93.9	91.4	-
WS-DAN	Inception v3	89.4	94.5	93.0	92.2
SCAM	Inception v3	89.8	94.8	93.3	91.8
OpairIM	ResNet50	87.7	94.8	92.8	-
DA-CNN	ResNet50	89.3	95.5	93.4	-
Our's	ResNet50	89.5	95.8	93.7	91.6

As can be seen from Table 3, the method proposed in this paper has achieved high recognition accuracy on four datasets. Compared with the CUB-200-2011 and Stanford Dogs datasets, the Stanford Cars and FGVC-Aircraft datasets are slightly less difficult to recognize. On these two datasets, the method in this paper has achieved a higher accuracy rate than other designs. The recognition effect of the complex module method is better, which shows that the proposed method can improve the model's ability to locate discriminative regions and feature extraction, and also shows the effectiveness of the data augmentation algorithm designed by this method.

4. Conclusion

In this paper, a fine-grained image recognition method based on data augmentation is proposed. This method uses the attention mechanism to find discriminative regions in the image, and at the same time performs data augmentation on the initial samples based on the discriminative regions, and then uses the data augmentation. samples to train and test the model. Experiments show that the proposed method has better recognition effect on four commonly used fine-grained image datasets. Compared with the current representative fine-grained image recognition methods, the model structure of the method in this paper is simpler, but the effect is better than most methods, which fully shows that for the fine-grained image recognition task, reasonable and effective data augmentation is necessary. A broad approach is equally important. In the future, we can continue to study more optimized data augmentation strategies and conduct experiments on other backbone networks with better recognition effects.

References

[1] H. Wei (2021). Analysis and Research of Key Technologies for Fine-grained Image Recognition Based on Convolutional Neural Networks. Changchun Institute of Optics, Fine Mechanics and Physics Chinese Academy of Sciences.

[2] X. X. Zeng (2020). Analysis and Research of Deep Learning based Fine-Grained Feature. Guangdong University of Technology.

[3] C. B. Liu (2021). Research on Key Technologies of Fine-grained Image Recognition. University of Science and Technology of China.

[4] Krizhevsky A, Sutskever I, Hinton G E.Imagenet classification with deep convolutional neural networks[J]// Communications of the ACM, 60(6):84-90,2017.

[5] Cubuk E D, Zoph B, Mane D, et al.Autoaugment: Learning augmentation policies from data[J]. CoRR, 2018, abs/1805.09501.

[6] Zhong Z, Zheng L, Kang G, et al. Random erasing data augmentation[C]// Proceedings of the AAAI Conference on Artificial Intelligence,34(07):13001-13008,2020.

[7] Zhang H, Cisse M, Dauphin Y, Lopez-Paz D. Mixup: Beyond empirical risk minimization[C]// International Conference on Learning Representations, 2018.

[8] YunS, Han D, Chun S, et al. Cutmix: Regularization strategy to train strong classifiers with localizable features[C]// IEEE/CVF International Conference on Computer Vision (ICCV), 2019:6022-6031.

[9] Devries T, Taylor G W. Improved regularization of convolutional neural networks with cutout[J]. CoRR, 2017. abs/1708.04552.

[10] Kim T, Kim H, Byun H. Localization-Aware Adaptive Pairwise MarginLoss for Fine-Grained Image Recognition[J]. IEEE Access, 9:8786-8796,2021.

[11] Hataya R, Zdenek J, Yoshizoe K, et al. Meta Approach to Data Augmentation Optimization[C]// 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 3535-3544, doi: 10.1109/WACV51458.2022.00359.

[12] Li H,Zhang X, Tian Q,et al. Attribute Mix: Semantic Data Augmentation for Fine Grained Recognition[C]// 2020 IEEE International Conference on Visual Communications and Image Processing (VCIP), 2020, pp. 243-246, doi: 10.1109/VCIP49819.2020.9301763.

[13] He M, Cheng Q, Qi G. Weakly Supervised Semantic and Attentive Data Mixing Augmentation for Fine-Grained Visual Categorization[J]. IEEE Access, 2022,10:35814-35823. doi: 10.1109/ACCESS.2022.3163302.

[14] W. C. Yu (2021). Study of Fine-grained Image Classification Algorithm. Guilin University of Electronic Technology.

[15] Wah C, Branson S, Welinder P, et al. The caltech-ucsd birds-200-2011 dataset[J]. California Institute of Technology, 2011.

[16] Krause J, Stark M, Deng J, et al. 3D object representations for fine-grained categorization[C]// IEEE International Conference on Computer Vision (ICCV), 2013:554-561.

[17] Maji S, Rahtu E, Kannala J, et al. Fine-grained visual classification of aircraft[J], arXiv: Computer Vision and Pattern Recognition, 2013.

[18] Khosla A, Jayadevaprakash N, Yao B, et al. Novel Dataset for Fine-Grained Image Categorization: Stanford Dogs[J]. Computer Science, 2011.

[19] Fu J, Zheng H, Mei T. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017. IEEE Computer Society, 2017: 4476-4484.

[20] ZhengH, Fu J, Mei T,et al. Learning multi-attentionconvolutionalneuralnetworkfor fine-grainedimagerecognition[C]// 2017 IEEEInternational ConferenceonComputerVision. ICCV 2017, Venice, Italy, October 22-29, 2017. IEEE Computer Society, 2017:5219-5227.

[21] Sun M, Yuan Y, Zhou F, et al. Multi-Attention Multi-Class Constraint for Fine-grained Image Recognition[C]//Computer Vision-ECCV 2018-15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XVI.2018:834-850.

[22] Wang Y, Morariu V I, Davis L S. Learning a discriminative filter bank within a CNN for fine-grained recognition[C]// 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018. IEEE Computer Society,2018: 4148-4157.

[23] Yang Z, Luo T, Wang D, et al.Learning to navigate for fine-structure classification[C]// Computer Vision-ECCV 2018-15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XIV. 2018:438-454.

[24] Hu T, Qi H, Huang Q, et al. See better before looking closer: Weakly supervised data augmentation network for fine-grained visual classification[J]. arXiv: Computer Vision and Pattern Recognition, 2019.