Finding Discriminative Regions with More Subtle Granularities for Fine-grained Image Classification

Qi Liu\(^1\)\(^a\)*, Zhipeng Xi\(^1\)\(^b\)*, Xiaoyu Fan\(^1\)\(^c\)*, Zhijian Yin\(^1\)\(^d\)*, Zhen Yang\(^1\)\(^e\)*

\(^1\)Jiangxi Science and Technology Normal University, Nanchang, China
\(^a\)criafting@aliyun.com, \(^b\)John_boy0618@163.com, \(^c\)fanxiaoyu2015@163.com,
\(^d\)zhijianyin@aliyun.com, \(^e\)yangzhenphd@aliyun.com

Abstract---Classifying subcategories of the same category (such as birds, cars, airplanes) is mainly to find discriminative features and accuracy regional positions in the fine-grained visual classification (FGVC). In this article, we propose to use global average pooling slice feature maps to find significant discriminative regions without complicated network designs or training operations and use Drop-Block mechanisms to solve the problem of network overfitting. Specially, we use the feature maps of the multi-branches as inputs, and average-pooling them with different sizes of convolution kernels to obtain feature maps that containing deeper and shallow information for getting more accuracy granularities. Our methods are called as More Accurate Multi-Granular Convolutional Neural Network (MAG-CNN). Compared with other networks with more complex designs, the network only requires common operations such as pooling and convolution to achieve higher accuracy. The MAG-CNN can be trained end-to-end, without any bounding-box, and our method has reached the most advanced performance on three common fine-grained image classification benchmark datasets (CUB-Birds, FGVC-Aircraft, and Stanford-Cars).

1. Introduction
Fine-grained image classification [14] refers to distinguishing identical sub-categories in the field of computer vision (for example: birds, cars, airplanes, etc.). Compared with other classification tasks, fine-grained classification is more difficult due to large differences between intra-class and small differences between inter-class. Finding more discriminative features from the global area have become the key points for solving the fine-grained classification problem and the well-known effective method. In addition, borrowing methods from other areas of computer vision have also become one of the directions such as object detection[1][2] and pedestrian re-identification[11][19]. Early classification methods mainly relied on self-supervised convolutional neural networks and manual part annotations[3]. However, this kind of human annotations would take lots of time and efforts but is easy to make mistakes including error-prone. Recently, with the rise and development of weakly supervised including half-supervised and unsupervised[7][10], some methods have been proved that the results perform better on some datasets without bounding-boxes because they can dig out distinguishable features and locate key areas. Now, more and more complex and deeper neural networks are used to obtain deeper-level feature maps. The main reason is that shallow-level features cannot play a decisive role. Up to now, solving the fine-grained classification problem is mainly divided into two methods:(1) using the idea of target detection to find the key areas;(2) using the attention mechanism[5][9] to ensure that the convolutional network could learn enough discriminative features. The latest work of the fine-grained classification
has achieved the most advanced achievements by exploring these two parts at the same time and using their complementarity[4][21].

FIG 1 The main problems of FGVC, i.e., low inter-class variance and high intra-class variance.

In this article, we use the method of image slicing to get multi-granularity information including global features and local discriminative features as the basis of classification for the model which can pay more attention on discriminative regions. We use FPN to fuse shallow information and high-level semantic information for obtaining more informative feature maps. The main works can be summarized as follows:

1) Using FPN to fuse shallow features with deep features. The shallow features express the overall information of the object target, but the discrimination is not strong; the deep features represent the key features of the object better, but the problem is the amount of deep information is large which is easy to cause misjudgments for the nerves. We use the three-layer convolutional network of the fourth layer of the Resnet50 network as the deep features, and merge them with the second, third and fourth-level features so that the obtained feature map contains both deep information and shallow information.

2) Adopting global average pooling with different sizes[6]. We use three branches as inputs: the first branch is responsible for the extraction of global information features of the image, the second branch uses pooling[12] with different sizes to divide the feature map into two parts, and the third branch will be divided into three, this part can extract more granular information. The three branches have cooperation and division of labor, sharing low-level feature weights and high-level semantic information independence.

3) Regularized drop-block. In order to solve the problem of too many parameters, we decide to adopt drop-block regularization module. The traditional drop-out is widely used in the fully connected layer of the neural network, but it does not play a big role in the convolutional layer. We introduce drop-block which is a structured form of dropout and puts the units in adjacent areas of the feature map together and drops them. We find that applying drop-block can improve accuracy in jump connections in addition to the convolutional layer. Otherwise, during the training process, gradually increasing the number of dropped units will lead to better accuracy and robustness to hyper-parameter selection.

2. Related work
Before many neural networks with very deep convolutional layers were designed, the object detection frames were mainly used to train and test the datasets in the early stage. But the annotation frame was marked and the labor cost was greatly increased and errors were prone to occur. Recently, some articles such as NTS-Net [20] which is using target detection to automatically generate discriminative regions...
in a convolutional neural network without external manual annotation. In the NTS-Net, it can learn effective information self-supervised without bounding boxes, and adopting the method of multi-agent collaborative to identify the information area of images in the fine-grained image classification task accurately. The NTS network also introduces FPN[15] (Feature Pyramid Network) and anchors to focus on the model’s most informative areas.

In the WSDAN[13] network, some attention-maps are generated to represent the discriminative regions through weakly supervised learning, then these attention maps are randomly selected for data enhancement (the main method is attention crop and attention[8] drop). The acquired feature maps and original image will be combined to crop and drop. In order to make the attention maps of the same channel focus on a specific area in the same category, the author proposes a method of attention regularization which like a sliding window method making the network pay attention to different discriminative areas in different categories; in the test phase, to realize the target positioning and image refinement of the test set, the attention map is generated through 1x1 convolution. By averaging all the attention maps, a new attention map is obtained, and then positioning to the overall position of the target.

In addition, the attention mechanism is also one of the commonly used methods for classification in recent years which including channel-attention mechanism and spatial-attention mechanism. Self-Attention is used to calculate the weight between different positions in the feature maps for achieving the effect of updating the feature in Transformer. In SE-Net, using Sigmoid to normalize for getting the channel attention map by applying the spatial dimension to adaptive average pool and learning the channel attention through two FCs, finally multiplying the Channel Attention Map with the original feature to get the weighted feature.

In MC-Loss, the cross-channel loss is introduced in the fine-grained image classification task. MC-Loss is composed of discriminative components and diverse components. Compared with other methods of SOTA, this paper has a small amount of code and it adds an auxiliary structure to the commonly used network structure. Moreover, the auxiliary structure does not introduce additional network parameters, only including regularization, pooling[16] and other operations which achieves the effect of SOTA.

Compared with the above classic and widely used methods, this paper draws on the idea of pedestrian re-identification slicing and improves it. We propose the use of neural network multi-layer features and only use average-pooling of different sizes to find more accurate multi-granularity discriminative areas.

3. Methods

3.1 Shallow feature and deep feature fusion

In this paper, we use Resnet-50 as the fine-grained classification network’s backbone in Fig 2. The feature maps are extracted by pre-trained extractor for an input image $I$ that is expressed as $F \in \mathbb{R}^{N \times W \times H}$, with the height $H$, the width $W$ and the number of channels $N$. The feature map is expressed as $F \in \mathbb{R}^{N \times H \times W}$, and we find that if only the fourth layer of the three-layer convolutional network is used as input, the final fine-grained classification effect is not ideal through lots of experiments. So, we fuse the second, third, and fourth-layer network of Resnet-50 with the fourth layer of the three-layers
convolutional neural network by the FPN. In the FPN layer, we use the $1 \times 1$ and step size of $1$
convolution layer on the feature maps so that all the numbers of channels are the same.  

$$F_i \in \mathbb{R}^{N \times 14 \times 14} (i = 1, 2, ..., n)$$  \hspace{1cm} (1)$$

The feature maps not only contain rich shallow information but also has key discriminative features
in this way. We have proved that part of the information will be missing which will reduce the accuracy
of classification if only the deep feature map is used.

3.2 Global Average-pooling and Slicing
We choose a rectangular convolution kernel in order to achieve the purpose of slicing: the size of the
feature map after the convolution layer operation is $F \in \mathbb{R}^{N \times 14 \times 14}$, and we choose three branches as inputs.
The size of the first branch pooling convolution kernel is $14 \times 14$ which the size of the output feature
map is $D_1 \in \mathbb{R}^{N \times 1 \times 1}$. But the second branch pooling size is $7 \times 14$, so we get Feature map size $D_2 \in \mathbb{R}^{N \times 2 \times 1}$.The feature maps are divided into up and low parts through the slicing operation:

$$C_i = D_i[; m: m + 1, ; :](m \in 1, 2, ..., n)$$  \hspace{1cm} (2)$$

The feature map obtained by slicing contains discriminative features, which play a key role in the
classification process. The feature map obtained by slicing is combined to form a new feature map. The
classification is performed via two fully connected layers, the loss function can be expressed as:

$$L_{LOSS}(F) = L_{CE} \left( y, \left[ e_{\theta(F_0)}, e_{\theta(F_1)}, ..., e_{\theta(F_n)} \right]^T \right)$$

Where $g(\cdot)$ is defined as:

$$G(Fi) = \frac{1}{WH} \sum_{n=1}^{N}[Fi]$$  \hspace{1cm} (4)$$

3.3 Drop-Block

Fig 3: The green part represents the activated feature unit. (a) input image to the convolutional neural
network. (b) Dropping out activations at random is less effective in removing information because
nearby activations contain closely related information. (c) Dropping continuous regions can remove
information and enforcing remaining units to learn other features for classifying.

In order to solve overfitting in the experiment, we adopt the regularized drop-block module in Fig 3.
The traditional dropout played a small role in the convolutional layer: the figure b represents the random
dropout activation unit, but the network will also learn the same information from nearby activations.
The figure c represents the Drop-Block in this article which mainly has 2 parameters, block-size and $\gamma$.
When block-size=$1$, Drop-Block degenerates to traditional dropout which can normally be $3, 5, \text{or } 7$. $\gamma$: represents the probability:
\[ \Gamma = \frac{1 - \text{keepprob}}{\text{blocksise}^2} \left( \frac{\text{featuresize}^2}{\text{featuresize} - \text{blocksise} + 1} \right)^2 \] (5)

We find that drop-block can inactivate a part of the entire adjacent area (such as the bird’s head and feet), then the network will focus on learning the characteristics of other parts of the object to achieve accuracy classification for showing better generalization.

4. Experiments

4.1 Dataset

In this section, we evaluate the performance of our designed framework on the fine-grained image classification benchmark datasets including CUB-200-2011, Stanford-Cars, and FGVC-Aircraft[18]. All datasets contain the same A set of subcategories of a super category. The following is a brief description of these data sets:

CUB-200-2011 has 11778 pictures from 200 categories, 5994 pictures are officially divided into training sets, and 5794 pictures are divided into test sets.

Stanford-Cars has 16,185 pictures from 196 categories, 8144 pictures are divided into training sets, and 8041 pictures are divided into test pictures.

FGVC-Aircraft has 10,000 images from 100 categories, officially divided into 6,667 training images and 3,333 test images.

| Dataset        | #Class | #Train | #Test |
|----------------|--------|--------|-------|
| CUB-200-2011   | 200    | 5,994  | 5794  |
| Stanford Cars  | 196    | 8,144  | 8,041 |
| FGVC Aircraft  | 100    | 6,667  | 3,333 |

4.2 Results

We choose Resnet50 pre-processed in ImageNet for experiments and analysis. More specifically, we choose the second, the third, and the fourth layers of Resnet50 to perform feature fusion with the fourth layer of the three-layers convolution through the FPN feature pyramid network. The size of the input
network image is 448×448 after flipping, cropping and other operations which are common basic image operations. In addition, we do not use additional bounding box annotations to facilitate comparison of our method with other weakly supervised methods. We use the open source Pytorch as our code base and train all models on a single GTX 2080Ti GPU. Use a stochastic gradient descent with a momentum of 0.9 and a mini-batch of 16 for optimization. The initial learning rate is set to 0.001 and is reduced to 0 using the cosine annealing schedule. All models have been trained for 200 iterations.

Table 2 shows the comparison between the neural network we designed and the current performance SOTA classification network. We also compare the global average pooling (GAP) and global maximum pooling (GMP)[17] on the Resnet50 classification network for slicing operations and show the effect of regularized Drop-Block on the performance of the convolutional neural network. As shown in Table 3, slicing with GAP improves the classification performance of the network compared with GMP. Compared with dropout, Drop-Block has a better performance in solving the network overfitting problem.

Table 2. CUB-200-2011, Stanford Cars, FGVC-Aircraft data set comparison results. The best and the second are marked in bold and underlined.

| Method     | CUB-200-2011 | Stanford Cars | FGVC Aircraft |
|------------|--------------|---------------|---------------|
| RA-CNN     | 85.3         | 92.5          | 88.2          |
| MA-CNN     | 86.5         | 92.8          | 89.9          |
| MA-CNN     | 86.5         | 92.8          | 89.9          |
| DFL-CNN    | 87.4         | 93.1          | 91.7          |
| NTS-Net    | 87.5         | 93.9          | 91.4          |
| TASN       | 87.9         | 93.8          |               |
| B-CNN      | 84.1         | 91.3          | 84.1          |
| MAG-Net    | 88.1         | 94.4          | 92.7          |

Table 3 With or without Drop-Block, GAP, and the comparison results of the two slicing methods of GAP and GMP in the CUB-200-2011 data set. Drop-Block: regularization module. GAP: global average pooling, GMP: global maximum pooling.

| Method             | Base Module | Accuracy(%) |
|--------------------|-------------|-------------|
| Baseline           | Resnet50    | 85.5        |
| MaxPooling(Slicing)| Resnet50    | 86.1        |
| GAP(Slicing)       | Resnet50    | 87.4        |
| Drop-Block+GAP(Slicing) | Resnet50 | 88.1        |

4.3 Visualization

In order to show the advantages of MAG-Net more intuitively, we use Grad-Cam to realize the channel visualization of the feature maps on CUB-200-2011, Stanford Cars and FGVC Aircrafts datasets. We can see that the network can learn the key areas of the target for classification, and focusing on different discriminative areas for the same class in the figure 5.
Fig 5 Some samples of feature maps visualization on CUB-200-2011, Stanford Cars and FGVC Aircrafts datasets

5. Conclusion
The problem of fine-grained image classification is still a challenging research topic in the field of computer vision for its goal is to identify sub-categories. Finding those subtle traits that fully characterize the object is not straightforward. To handle this circumstance, we propose to use the global average pooling method of slicing operation to replace the max-mum pooling for fine-grained classification without manual part annotations. The global information and local information are combined to form a multi-granular convolutional neural network, which provides a shallow Layer semantic information and deep discrimination information. Using drop-block regularization module, further improve the classification ability of the network. The effectiveness of our method is proved on the data sets of CUB-200-2011, Stanford Cars and FGVC Aircraft. Our network model is very effective for large goals but it is still somewhat insufficient for small goals, this is a future research direction for us.

Acknowledgement
This work was supported by the National Natural Science Foundation of China (62061019, 61866016), General Project of Jiangxi Natural Science Foundation(20202BABL202014), the General Project of Jiangxi Education Department(GJJ190587), and the Youth Top Talent Foundation of Jiangxi Science and Technology Normal University (2018QNBJRC002).

References
[1] Anelia Angelova and Shenghuo Zhu. Efficient object detection and segmentation for fine-grained recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 811–818, 2013.
[2] Steve Branson, Grant Van Horn, Serge Belongie, and Pietro Perona. Bird species categorization using pose normalized deep convolutional nets. arXiv preprint arXiv:1406.2952, 2014.
[3] Yuning Chai, Victor Lempitsky, and Andrew Zisserman. Symbiotic segmentation and part localization for fine-grained categorization. In Proceedings of the IEEE International Conference on Computer Vision, pages 321–328, 2013.
[4] Dongliang Chang, Yifeng Ding, Jiayang Xie, Ayan Kumar Bhunia, Xiaoxu Li, Zhanyu Ma, Ming Wu, Jun Guo, and Yi-Zhe Song. The devil is in the channels: Mutual-channel loss for fine-grained image classification. IEEE Transactions on Image Processing, 29:4683–4695, 2020.
[5] Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu, and Tat-Seng Chua. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 5659–5667, 2017.
[6] Yin Cui, Feng Zhou, Jiang Wang, Xiao Liu, Yuanqing Lin, and Serge Belongie. Kernel pooling for convolutional neural networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 2921–2930, 2017.
[7] Cheng Deng, Erkun Yang, Tongliang Liu, Jie Li, Wei Liu, and Dacheng Tao. Unsupervised semantic-preserving adversarial hashing for image search. IEEE Transactions on Image Processing, 28(8):4032–4044, 2019.

[8] Yifeng Ding, Zhanyu Ma, Shaoguo Wen, Jiyang Xie, Dongliang Chang, Zhongwei Si, Ming Wu, and Haibin Ling. Ap-cnn: weakly supervised attention pyramid convolutional neural network for fine-grained visual classification. IEEE Transactions on Image Processing, 30:2826–2836, 2021.

[9] Jianlong Fu, Heliang Zheng, and Tao Mei. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 4438–4446, 2017.

[10] Weifeng Ge, Xiangru Lin, and Yizhou Yu. Weakly supervised complementary parts models for fine-grained image classification from the bottom up. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 3034–3043, 2019.

[11] Weifeng Ge and Yizhou Yu. Borrowing treasures from the wealthy: Deep transfer learning through selective joint fine-tuning. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1086–1095, 2017.

[12] Golnaz Ghiasi, Tsung-Yi Lin, and Quoc V Le. Dropblock: A regularization method for convolutional networks. arXiv preprint arXiv:1810.12890, 2018.

[13] Tao Hu, Honggang Qi, Qingming Huang, and Yan Lu. See better before looking closer: Weakly supervised data augmentation network for fine-grained visual classification. arXiv preprint arXiv:1901.09891, 2019.

[14] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 25:1097–1105, 2012.

[15] Tsung-Yi Lin, Piotr Dollar, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 2117–2125, 2017.

[16] Tsung-Yu Lin, Aruni RoyChowdhury, and Subhransu Maji. Bilinear cnn models for fine-grained visual recognition. In Proceedings of the IEEE international conference on computer vision, pages 1449–1457, 2015.

[17] Zhanyu Ma, Dongliang Chang, Jiyang Xie, Yifeng Ding, Shaoguo Wen, Xiaoxu Li, Zhongwei Si, and Jun Guo. Fine-grained vehicle classification with channel max pooling modified cnns. IEEE Transactions on Vehicular Technology, 68(4):3224–3233, 2019.

[18] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. arXiv preprint arXiv:1306.5151, 2013.

[19] Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou. Learning discriminative features with multiple granularities for person re-identification. In Proceedings of the 26th ACM international conference on Multimedia, pages 274–282, 2018.

[20] Ze Yang, Tiange Luo, Dong Wang, Zhiqiang Hu, Jun Gao, and Liwei Wang. Learning to navigate for fine-grained classification. In Proceedings of the European Conference on Computer Vision (ECCV), pages 420–435, 2018.

[21] Heliang Zheng, Jianlong Fu, Zheng-Jun Zha, and Jiebo Luo. Looking for the devil in the details: Learning trilinear attention sampling network for fine-grained image recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 5012–5021, 2019.