

# The loss model with class variance for fine-grained classification

Qian Long<sup>1</sup>, Bolun Zhu<sup>1</sup>, Gaihua Wang<sup>1</sup>, and Hongwei Qu<sup>2</sup>

<sup>1</sup> College of Artificial Intelligence, Tianjin University of Science & Technology, Tianjin, 300457, China

<sup>2</sup> Wuhan Electronic Information Institute, Wuhan 430019, China

This work is supported by the National Nature Science Foundation of China under Grant No. 61601176

The corresponding author: Gaihua Wang Bolun Zhu

wanggh@tust.edu.cn

zbl890730@163.com

**Abstract.** We propose a loss model with class variance for fine-grained image classification. It adopts basic convolutional neural network to get features. The dates from dataset are shuffle selected as inputs according to batch size and their outputs are processed by attention model. Because of class variance in the same class is smaller and that in the different class is larger, in the training phase, we use class variance to define the loss function. The total loss model combines the loss function with class variance and label loss function. Both are jointly employed to fast convergence. Compared with state-of-the-art methods, experimental results demonstrate our model has better performance.

**Keywords:** Class variance; Fine-grained; Image classification; Loss function

## 1 Introduction

The challenges of fine-grained classification basically relate to two aspects: inter-class similarity and intra-class variance. The categories in the different species are similar to each other and only distinguished by subtle differences. the categories in the same species have large variant appearance. So, the discriminative features in fine-grained images are mainly localized on object or parts region.

Earlier works require extra information such as bounding box and part annotation besides the class labels[1-3] . The paper [4] adopts a fully convolutional network to locate multiple object parts and compare with manually-labeled strong part annotations. It uses a two-stream classification network to encode object-level and part-level cues simultaneously. He et al. [5] introduce a detector to localize the object and use two spatial constraints to select the distinguished parts. Lam et al.[6] propose a set of bounding boxes in the image by the heuristic function and the successor function. The two functions are unified by a Long Short-Term Memory (LSTM) network into a deep

recurrent architecture. All these methods need datasets which provide detailed part annotations including part landmarks and attributes[7, 8]. Such annotations are expensive and unrealistic in many real applications.

Weakly supervised part detection methods are proposed without using the expensive annotations[9-11]. Lin et al. [12] introduce bilinear pooling method. It is a pooled outer product of features derived from two CNNs. Kong et al. [13] propose classifier that factorizes the collection of bilinear classifier into a common factor and compact terms. The paper[14] aligns with the fine-grained and category-specific content of images and natural language provides a flexible and compact way of encoding only the salient visual aspects for distinguishing categories. SCDA[15] is a method that discards the noisy background and keeps useful deep descriptors without bounding box annotation. Then the selected descriptors are aggregated into a short feature vector. The paper [16] captures higher order interactions of features by a pooling framework. The paper [17] uses a hierarchical bilinear pooling method to integrate multiple cross-layer bilinear features to enhance feature representations. A region enhancement and suppression approach [18] proposes a plug-and-play significant region diffusion (SRD) module for explicitly enhancing the significant features. PHPQ [19] captures and retains fine-grained semantic information in multi-level features. GCA [20] use cluster to divides the data into small clusters, and then aggregates them to get fine-grained information.

In addition, attention model is widely used that can strengthen useful information and suppress the noise or useless information. Integrate network of attention [21] has three attention parts: the bottom-up attention finds candidate patches, the object-level top-down attention selects relevant patches to a certain object, and the part-level top-down attention localizes discriminative parts. Liu et al.[22] introduce a reinforcement learning-based fully convolutional attention localization network to adaptively select multiple task-driven attention regions. Fu et al.[23] propose a recurrent attention CNN which can learn discriminative feature representation by multiple scales network. Peng et al.[24] propose attention model that is composed of object-level attention and part-level attention. The Attention Module for Two-Branch Networks (DAL-Net) [25] acquires features in a weakly supervised manner. Weakly Supervised Spatial Group Attention Network (WSSGA-Net) [26] highlights the correct semantic feature regions for more accurate classification by establishing semantic enhancement mechanism.

In this paper, we propose a loss model with class variance for fine-grained classification. By metric learning, the loss function aims to maximize inter-class distance and minimize intra-class distance. The salient features are extracted by attention model. The proposed method can be applied to different fine-grained network. The whole network only uses class label to finish fine-grained classification.

## 2 Related works

Fine-grained classification aims to recognize sub-categories under some basic-level categories. The subtle difference between fine-grained categories mainly relies on the properties of object parts, so all kinds of methods are proposed by localizing object or part region. Some methods with annotation information are effective. But they require

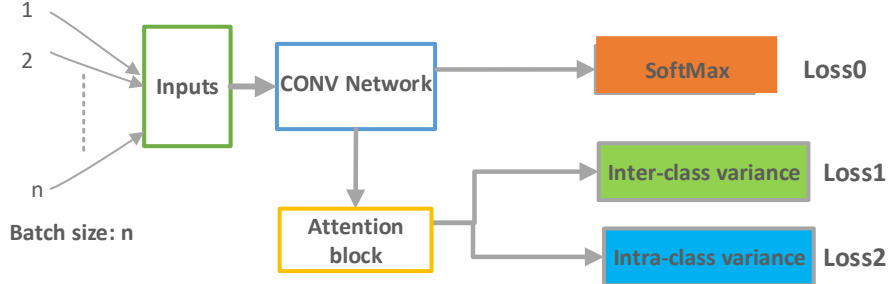
extra key points from humans, which are expensive to obtain. weakly supervised methods without annotation have a promising prospect[10, 11, 27, 28] . These weakly supervised works make it possible to put the fine-grained image classification into practical applications.

In fine-grained image classification, most existing image similarity models consider class-label and has a SoftMax loss function. Some researchers use class-variance to compare image similarity. The paper[29] integrates fine-grained ranking model to learn fine-grained image similarity. Through a triplet loss, it contains a query image, a positive image, and a negative image. The positive image is more similar to the query image than the negative image. The paper[30] uses triplets of roughly aligned matching/non-matching face patches generated using a novel online triplet mining method to recognize face. The paper[31] incorporates the modeling of intra-class variance into triplet network. Clustering is applied to implement the grouping which can figure out a mid-level representation within each fine-grained category to capture the intra-class variance. These works have significantly improved the fine-grained recognition performance by using triplet network. However, it is complexity for selecting the triplet images. Our method proposes loss model with class variance. The main contributions are summarized as follows:

- (1) We use the theory that maximize inter-class distance and minimize intra-class distance to optimize the loss function. The inputs selected from datasets are shuffle dates according to batch size. It is simpler than the triplet network.
- (2) The attention model is used to extracted salient features. Then the loss functions are composed by the SoftMax loss function and loss function of class variance. By optimizing the joint objective of loss and attention model, we can generate effective feature representations.
- (3) The experiments demonstrate that the proposed framework is superior to corresponding original methods. It improves the performance significantly over all kinds of methods without complex human-defined annotations.

### 3 The Loss model with class variance

Instead of the triplet images, the proposed method selects inputs randomly according to batch size. Given  $n$  inputs of batch size. each image is resized with  $H \times W \times C$ .  $H$  is the height of image,  $W$  is the width of image,  $C$  is the channel of image. The whole structure of network is showed in Fig.1. It consists of three parts: Inputs, CONV Network and loss function.



**Fig. 1.** The structure of network

### 3.1 Attention Block

To get discriminant feature, attention block is used to generate a saliency map. Given  $F$  are the feature maps of certain layers.  $f_c(x, y)$  represents feature map which is at spatial location  $(x, y)$  of channel  $c$ . We get  $Mask_c(x, y)$  by

$$Mask_c(x, y) = \text{sigmoid}(f_c(x, y)) \quad (1)$$

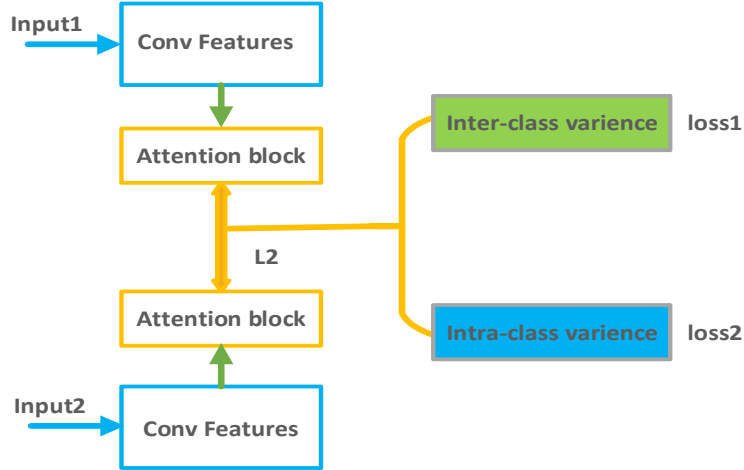
The output  $M$  of Attention block is computed by residual attention mechanism. It is expressed by

$$M = F(1 + Mask_c) \quad (2)$$

The attention model focuses on the distinguishing differences of parts among sub-categories, which makes the saliency feature values stronger and suppresses noise pixels or useless information.

### 3.2 The total loss functions

For inputs in each batch size, we compare the distance between any two inputs. Given any two inputs are input1 and input2.  $f(x, y)$  are the output features of inputs. The output of attention block is the  $M$ .  $M(\text{input1})$  is the attention output of input1.  $M(\text{input2})$  is the attention output of input2. The distance between input1 and input2 is computed by Euclidean distance. Label1 is the label of input1. Label2 is the label of input2. When the label1 is equal to the label2, the input1 and input2 belong to the same class. The difference of input1 and input2 is added to loss2. When the label1 is not equal to the label2, the loss is added to loss1. The Fig.2 is the loss model with class variance.



**Fig. 2.** the loss model with class variance

In addition, all inputs have the output of network respectively. And loss is defined as loss0 (In Fig.1). The total loss model is composed of loss0, loss1 and loss2.

## 4 Experiments

All experiments are performed by PyTorch on NVIDIA GeForce RTX 2060 GPUs. We demonstrate experiment results on three standard fine-grained datasets, including CUB-200-2011, Stanford dogs, and Oxford-Flower-17. Accuracy is adopted as the evaluation metric to evaluate the classification performances. No extra annotation or object bounding box is used in the whole experiments.

### 4.1 Datasets and Baselines

We evaluate our method on three fine-grained image recognition datasets. These datasets all contain large diversity in each class and visually similar in different classes. CUB-200-2011[7] contains 11788 images of 200 subcategories, 5994 images for training and 5794 images for testing, which is the most widely-used dataset in fine-grained image classification. Stanford dogs[8] has been built using images and annotation from ImageNet for the task of fine-grained image categorization. it contains images of 120 breeds of dogs from around the world, which has 20580 images split into 12000 train images and 8580 test images. Oxford-Flower-17[32] consists of 17 flower types. Each flower type has 80 images which are split into 40 train images and 40 test images. To increase the number of datasets, we randomly clipp, translate, scale images. In the training phase, images are randomly cropped  $224 \times 224$  patches or  $448 \times 448$  patches.

VGG-16 and ResNet-18 are used as the basic CNNs. VGG-16 has 16 layers which include 13 convolutions with ReLU layers and 3 fully connected layers. ResNet-18 has four basic conv blocks. Each conv block includes two residual units. We use a weight decay of 0.0001 with a momentum of 0.9 and set the initial learning rate to 0.01. Then

the learning rate is divided by 10 every 20 iterations. Two networks are initialized with the pre-trained network on the ImageNet.

## 4.2 Comparisons

We compare the proposed method with BCNN, CBP [33] and HBP [17]. BCNN is a method which replaces the fully connected layers with bilinear pooling. CBP is a compact bilinear pooling method. In original paper of CBP, the features of CBP uses RM and TS projections. In our experiments, it uses TS projections. HBP is hierarchical bilinear pooling method which can integrate multiple cross-layer bilinear features to enhance their representation capability. We define the proposed loss model with class variance as LCV model.

The LCV model is added on BCNN network, which is called as LCV on BCNN. The LCV model is added on CBP network, which is called as LCV on CBP. The baseline is VGG-16 for BCNN and CBP. We add the proposed model on HBP network, which is called as LCV on HBP. The baseline is ResNet-18 for LCV on HBP.

## 4.3 Experiments on CUB-200-2011

We randomly crop  $448 \times 448$  patches or  $224 \times 224$  patches in training images for CUB-200-2011 dataset. Table 1 shows the comparison results for randomly cropping  $448 \times 448$  patches in training images. The test images are center cropped  $448 \times 448$ . From the Table 1, we can see that LCV model has a boost performance than the original model. The accuracy of BCNN is 72.01%. The accuracy of LCV on BCNN is 75.13%. The accuracy of CBP is 76.04%. The accuracy of LCV on CBP is 76.32%. The accuracy of HBP is 72.95% when the baseline is ResNet-18. The accuracy of LCV on HBP is 74.28%. In comparison, our LCV model achieves significant improvement. The proposed approach on CBP achieves the highest classification accuracy among all methods without object and parts annotations.

**Table 1.** shows the comparison results on CUB-200-2011 dataset.

Method	Base Model	Accuracy (%)
PHPQ	Resnet-18	74.13
GCA	Vit-B/16	73.40
BCNN	VGG-16	72.01
LCV on BCNN	VGG-16	75.18
CBP	VGG-16	76.04
LCV on CBP	VGG-16	76.32
HBP	VGG-16	72.71
	Resnet-18	72.95
LCV on HBP	Resnet-18	74.28

Table 2 shows the comparison results for randomly cropping  $224 \times 224$  patches in training images. The test images are center cropped  $224 \times 224$ . We add LCV model to different methods. The accuracy can be improved by using the proposed model. The

accuracy of LCV on BCNN is 70.45% higher than the original BCNN. The accuracy of LCV on HBP is 70.35 higher than HBP. The accuracy of LCV on CBP is 73.10%. And it has highest accuracy. Neither object nor parts annotations are used in these approaches. For different methods, it has an impressive boost when adding our LCV model.

**Table 2.** shows the comparison results on CUB-200-2011 dataset

Method	Base Model	Accuracy (%)
BCNN	VGG-16	70.34
LCV on BCNN	VGG-16	70.45
CBP	VGG-16	72.75
LCV on CBP	VGG-16	73.10
HBP	VGG-16	70.18
	Resnet-18	69.69
LCV on HBP	Resnet-18	70.35

### Experiments on Stanford dogs

The classification accuracy on Stanford dogs is summarized in Table 3. The training images are randomly cropped  $224 \times 224$  patches. The test images are center cropped  $224 \times 224$ . We add LCV to different network and get an impressive progress. The accuracy of LCV on BCNN is 62.03% higher than BCNN. The accuracy of LCV on CBP is 80.21% higher than CBP. The accuracy of LCV on HBP is 80.01% higher than HBP.

**Table 3.** shows the comparison results on Stanford dogs dataset.

Method	Base Model	Accuracy (%)
PHPQ	Resnet-18	79.83
BCNN	VGG-16	60.29
LCV on BCNN	VGG-16	62.03
CBP	VGG-16	78.71
LCV on CBP	VGG-16	80.21
HBP	VGG-16	74.65
	Resnet-18	77.55
LCV on HBP	Resnet-18	80.01

### Experiments on Oxford-Flower-17

From the Table 4, it shows the comparison results on Oxford-Flower-17 dataset. The training images are also randomly cropped  $224 \times 224$  patches. The test images are center cropped  $224 \times 224$ . The accuracy of LCV on BCNN is 88.68%. The accuracy of LCV on CBP is 91.03%. The accuracy of LCV on HBP is 90.15%. They have a better performance than the original network.

**Table 4.** shows the comparison results on Oxford-Flower-17 dataset.

Method	Base Model	Accuracy (%)
BCNN	VGG-16	84.71
LCV on BCNN	VGG-16	88.68
CBP	VGG-16	89.12
LCV on CBP	VGG-16	91.03
HBP	VGG-16	87.52
LCV on HBP	VGG-16	89.26
HBP	Resnet-18	88.09

LCV on HBP	Resnet-18	90.15
------------	-----------	-------

### 4.3 Analysis and Discussion

We analyze and compare different methods on different datasets. In BCNN, the final feature maps of VGG-16 and ResNet-18 are all 512 channels. The outer product of features is computed and the dimension is  $512 \times 512$ . CBP is compact bilinear pooling method. The computational time of BCNN is longer than CBP. In addition, BCNN only focuses on final layers with high dimension. These key features might be overwhelmed by large non-discriminative region in pooling stage. For HBP, the projection layers determine essential feature of the object, which distinguish its category by successive interaction and integration of different part. The final dimension of original HBP is 8192. The complexity of HBP is larger than others. CBP has compact dimension and get a fast convergence.

Our proposed method provides a rough localization of target from attention model. The loss function is computed by class variance. The LCV model can be optimized by maximizing inter-class distance and minimizing intra-class distance. The results of LCV model are competitive with the original network. All these networks require no bounding box/part annotations and can be trained end-to-end.

### 5. Conclusions

In this paper, the loss model with class variance has been proposed for weakly supervised fine-grained image classification. It can localize the object and learn salient features by attention model. Based on maximizing inter-class distance and minimizing intra-class distance, the loss function is optimized. The proposed approach demonstrates better performance without requiring bounding box/part annotations.

### References:

1. Zhang, N., et al., Part-based r-cnns for fine-grained category detection[C], in In European Conference on Computer Vision. 2014.
2. Xiu, S., W. Chen and J. Wu, Mask-CNN:Localizing Parts and Selecting Descriptors for Fine-Grained Image Recognition. arXiv:1605.06878v1, 2016(5).
3. He, X. and Y. Peng, Fine-grained Image Classification via Combining Vision and Language, in arXiv: 1704.02792V1. 2017.
4. Huang, S., et al., Part-Stacked CNN for Fine-Grained Visual Categorization. arXiv:1512.08086v1, 2015(12).
5. He, X. and Y. Peng, Weakly Supervised Learning of Part Selection Model with Spatial Constraints for Fine-Grained Image Classification[C], in Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence. 2017.
6. Lam, M., B. Mahasseni and S. Todorovic, Fine-Grained Recognition as HSnet Search for Informative Image Parts[C], in IEEE Computer Society. 2017: IEEE Computer Society.
7. C., W., et al., The Caltech-UCSD Birds-200-2011 Dataset. Computation & Neural Systems Technical Report, 2011(1).
8. Khosla, A., et al., Novel dataset for Fine-Grained Image Categorization[C], in IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2011.
9. Cai, D., et al., Convolutional low-resolution fine-grained classification[J]. Pattern Recognition Letters, 2019. 119: p. 166-171.



10. Qi, L., X. Lua and X. Li, Exploiting spatial relation for fine-grained image classification[J]. *Pattern Recognition*, 2019. 91: p. 47-55.
11. Lai, D., W. Tian and L. Chen, Improving classification with semi-supervised and fine-grained learning[J]. *Pattern Recognition*, 2019. 88: p. 547-556.
12. Lin, T., R. Aruni and S. Maji, Bilinear CNNs for Fine-grained Visual Recognition . arXiv:1504.07889v5, 2017(5).
13. Kong, S. and C. Fowlkes, Low-rank Bilinear Pooling for Fine-Grained Classification, in arXiv:1611.05109v1. 2016.
14. Reed, S., et al., Learning Deep Representations of Fine-Grained Visual Descriptions. arXiv:1605.05395v1, 2016(5).
15. Wei, X., et al., Selective Convolutional Descriptor Aggregation for Fine-Grained Image Retrieval. arXiv:1604.04994v2, 2017.
16. Cui, Y., et al., Kernel Pooling for Convolutional Neural Networks[C], in *IEEE Conference on Computer Vision & Pattern Recognition*. 2017.
17. Yu, C., et al., Hierarchical Bilinear Pooling for Fine-Grained Visual Recognition. arXiv:1807.09915v1, 2018(7).
18. Pan W Y, Yang S Y, Qian X H, et al. Learn More: Sub-Significant Area Learning for Fine-Grained Visual Classification. 2023 IEEE International Conference on Image Processing (ICIP).Kuala Lumpur, Malaysia, 2023: 485-489.DOI: 10.1109/ICIP49359.2023.10222241.
19. Zeng Z Y, Wang J P, Chen B, et al. Pyramid hybrid pooling quantization for efficient fine-grained image retrieval. *Pattern Recognition Letters*, 2024, 178: 106-114. DOI: 10.1016/j.patrec.2023.12.022.
20. Otholt J, Meinel C, Yang H J. Guided Cluster Aggregation: A Hierarchical Approach to Generalized Category Discovery. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision(WACV)*.2024:2618-2627.<https://openaccess.thecvf.com/content/WACV2024>
21. Xiao, T., et al., The Application of Two-level Attention Models in Deep Convolutional Neural Network for Fine-grained Image Classification. arXiv:1411.6447v1, 2014.
22. Liu, X., T. Xia and J. Wang, Fully convolutional attention localization networks: Efficient attention localization for FGVC, in *Computer Vision and Pattern Recognition*. arXiv:1603.06765. 2016.
23. Fu, J., H. Zheng and T. Mei, Look Closer to See Better: Recurrent Attention Convolutional Neural Network for Fine-grained Image Recognition[J]. *Computer Vision & Pattern Recognition*, 2017.
24. Peng, Y., X. He and J. Zhao, Object-Part Attention Model for Fine-grained Image Classification. arXiv:1704.01740v2, 2017(9).
25. Jung Y, Syazwany N S, Kim S, et al. Fine-Grained Classification via Hierarchical Feature Covariance Attention Module. *IEEE Access*, 2023,11:35670-35679.DOI: 10.1109/ACCESS.2023.3265472.
26. Xie J J, Zhong Y J, Zhang J G, et al. A weakly supervised spatial group attention network for fine-grained visual recognition. *Applied Intelligence*, 2023, 53(20): 23301-23315. DOI: 10.1007/s10489-023-04627-z.
27. Wang, Y., V.I. Morariu and L.S. Davis, Learning a Discriminative Filter Bank within a CNN for Fine-grained Recognition. arXiv:1611.09932v3, 2018.
28. Dubey, A., et al., Pairwise Confusion for Fine-Grained Visual Classification[C], in *Computer Vision and Pattern Recognition*. 2018.
29. Wang, J., et al., Learning Fine-grained Image Similarity with Deep Ranking. arXiv:1404.4661v1, 2014(4).

30. Schroff, F., D. Kalenichenko and J. Philbin, FaceNet: A Unified Embedding for Face Recognition and Clustering. arXiv:1503.03832v3, 2015(6).
31. Bai, Y., et al., INCORPORATING INTRA-CLASS VARIANCE TO FINE-GRAINED VISUAL RECOGNITION. arXiv:1703.00196v1, 2017(5).
32. Oxford Flowers 17. <http://www.robots.ox.ac.uk/~vgg/data/bicos/>.
33. Gao, Y., et al., Compact Bilinear Pooling, in arXiv:1511.06062v2. 2016.