FEATURE COMPARISON BASED CHANNEL ATTENTION FOR FINE-GRAINED VISUAL CLASSIFICATION

Shukun Jia, Yan Bai, Jing Zhang

CRISE, Institute of Automation, Chinese Academy of Sciences

ABSTRACT

Fine-grained visual classification (FGVC) remains challenging because a majority of samples have large intra-class variations and small inter-class variations. However, samples belonging to one category are essentially identical in some discriminative visual patterns. Intuitively, we want models to reinforce the relationship between these discriminative visual patterns and image-level labels. In this paper, we propose a feature comparison based channel attention (FCCA) to achieve this intuition. In FCCA, the feature comparison mechanism is designed to recognize discriminative visual patterns. The weights assignment scheme guarantees that feature channels related to discriminative visual patterns have larger weights. The state-of-the-art performance has been achieved on two public FGVC datasets. Extensive experiments further prove the effectiveness of our method.

***Index Terms*—** Channel attention, feature comparison mechanism, fine-grained visual classification

1. INTRODUCTION

Fine-grained visual classification (FGVC) aims to differentiate similar sub-categories within a large and basic category. As a special visual classification task, FGVC could benefit from the development of deep convolutional neural network (DCNN). First, the architecture of DCNN is going deeper and smarter. Since LeNet[1] and AlexNet[2], there are Vgg[3], ResNet[4][5], GoogLeNet[6][7][8] and so forth being proposed consequentially. Aside from model architecture, data augmentation[9] is another effective way to boost performance due to its alleviation of over-fitting. Above factors could aid the solutions of FGVC.

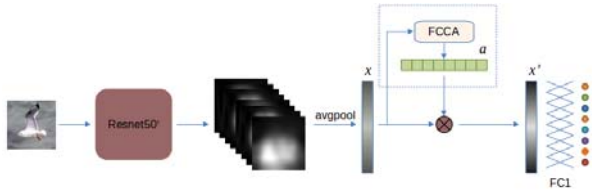
In FGVC tasks, samples usually contain similar objects. It is the key to make predictions based on subtle difference of discriminative regions. At the beginning, hand-craft annotations on object-level were used to supervise models to locate attentive regions[10][11][12]. The necessity of manually annotated bounding boxes or object parts is too prohibitive in practice. To eliminate the reliance on elaborate annotations, methods that could automatically locate object regions are drawing more and more attention. These methods only need image-level annotations and largely reduce the cost of solutions. In the process, attention mechanism is explored and exploited. Recurrent Attention CNN (RA-CNN)[13] recursively learned discriminative region attention and extracted features with the guidance of attention; Multi-Attention CNN (MA-CNN)[14] achieved part localization via channel grouping and took classification based on part-based representation. In addition to attention mechanism, data augmentation has also been demonstrated in FGVC. Destruction and Construction Learning (DCL) method[15] elaborately destroyed images at pixel level into small patches to focus the model on local details instead of global structures, which is a variant of data augmentation. Weakly Supervised Data Augmentation Network (WS-DAN)[16] took data augmentation via attention mechanism. It used a feature extractor to generate attention masks in the first stage. Then data were augmented by the masks and used to fine tune the model in the second stage.

To help DCNN models better utilize crucial information from samples, we propose the **f**eature **c**omparison based **c**hannel **a**ttention (FCCA) in this paper. For simplicity, we also use "FCCA" to denote the FCCA module. Compared with previous FGVC methods, FCCA has advantages mainly in three aspects: 1. **interpretability:** Owing to the feature comparison mechanism, the way FCCA generates attention weights has good interpretability; 2. **cost:** FCCA is a lightweight plug-in only for training stage. It is trained end-to-end with backbone networks. During experiments, we also found that improvements by FCCA could be achieved without further tuning hyper-parameters of original backbone networks; 3. **constructiveness:** FCCA could get accuracy gains on both general classification networks and an advanced FGVC method. The latter means the constructiveness in the development of FGVC. Fig. 1 shows the framework of FCCA. Comprehensive experiments are quantitatively and qualitatively conducted to prove the effectiveness of FCCA.

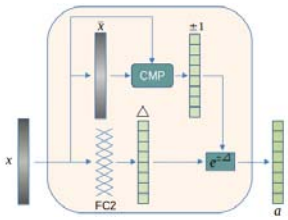
2. METHODOLOGY

Although many fine-grained samples have large intra-class variations, they have similar discriminative visual patterns as long as they essentially belong to the same category. These visual patterns are crucial to prediction results. If they are presenting along with samples, we want the model to catch them and establish a strong relationship with final outputs. In

contrast, these visual patterns may get weaken or even disappear due to special poses, occlusions and illuminations etc. Then models need to rely on other visual patterns comprehensively and reserve the previous relationship. Otherwise, this relationship may get impaired by samples with noise, which may cause the oscillation of training process. Overall, we have two steps to achieve this scheme. First, a **f**eature **c**omparison **m**echanism (FCM) is proposed to recognize whether discriminative visual patterns of each category appear along with samples or not. Second, we assign attention weights on feature channels. Because feature channels are corresponding to visual patterns[14][17][18]. A larger weight on a feature channel represents more attention to the visual pattern. What's more, different feature channels have different weights in the classification process. Some channels are crucial for recognizing specific categories but some are not. A fully connected network is used to automatically learn the weights. The structure of FCCA is shown in Fig. 1(b).



(a) Framework. FCCA is plugged between feature extractor Resnet50* and classifier FC1, where Resnet50* denotes the part of Resnet50 before average pooling layer. "avgpool" means global average pooling. x and x' represent the original features and weighted features respectively. $\otimes$ is element-wise multiplication. Pipelines in the dashed region are only used during training stage.



(b) FCCA. (1) For $\bar{x}$: In training stage, features of every sample are accumulated, which are used to calculate the mean features $\bar{x}_c$ of every category; (2) For ± 1: Features of every sample are compared (CMP) with the mean features of its category. Comparison results are denoted by ± 1; (3) For $\triangle$: Features of every sample are the input of a two-layer fully connected network FC2, which outputs positive values $\triangle$. (4) For a: Attention weights are generated by $e^{\pm\triangle}$.

Fig. 1. Overview of framework

2.1. Feature Comparison Mechanism

According to [14][17][18], feature channels are corresponding to certain visual patterns. If one feature channel responds to a sample strongly, its corresponding visual pattern appears more obviously in the sample. Vice versa, if one feature channel responds weakly, its corresponding visual pattern tends to disappear in the sample. This relationship indicates the strength of feature channel response and its visual pattern are positively correlated. Therefore, the average response of a feature channel from one category could roughly represent the average strength of the related visual pattern in this category. And it could be used as a threshold to recognize discriminative visual patterns. Concretely, we could compare related features of a sample to the counterpart mean features of the category channel-by-channel. To implement the mechanism expressed above, we need to accumulate samples' features and calculate the mean features for each category:

$$\bar{x}_c = \frac{1}{N_c} \sum_{x \in c} x \tag{1}$$

where x represents features of samples and it belongs to category c (so x is also denoted as x_c); N_c is the number of samples in category c; $\bar{x}_c$ represents the mean features of category c and it has the same dimension as x.

After getting $\bar{x}_c$, every sample's features x_c is compared with its category's mean features $\bar{x}_c$ channel-by-channel. We use sign s, i.e. ± 1 to denote the comparison results in order to embed them into FCCA.

$$\begin{cases} s_i = 1, & \text{if } x_{ic} >= \bar{x}_{ic} \\ s_i = -1, & \text{if } x_{ic} < \bar{x}_{ic} \end{cases} \tag{2}$$

where i means the i^{th} channel in sign s, features x_c and $\bar{x}_c$.

The FCM is quantitatively and qualitatively demonstrated in Section 3.3 and Section 3.4 respectively. Having recognized discriminative regions automatically, the model is going to assign attention weights to feature channels.

2.2. Weights Assignment Scheme

Attention weights of feature channels should be based on feature comparison results. For certain feature channel, if $s_i = 1$, i.e. its visual pattern presents stronger than average, a large weight should be assigned; Otherwise if $s_i = -1$, a small weight should be assigned; Weights are all positive. This mechanism is identical to the essence of general attention mechanism which pays more attention to prominent regions. This scheme also benefits the optimization process. Feature channel responses (i.e. x in Fig. 1(a)) are the output of a series of filter banks. When taking back propagation at one iteration, if a feature channel is assigned with a large weight, its related filter banks will have a larger update step, the model will learn more from its visual pattern. In contrast, if a small weight is assigned, its filter banks will take a small update, which means the relationship it has learned previously could be temporally reserved. Extremely, if the weight of a feature channel is set as zero, its corresponding filter banks will get no update in the current iteration. Although filter banks are seriously coupled, this scheme instills the preference to models and could still improve the optimization process. For noisy

patterns, they may get temporarily reserved in one iteration, but will ultimately get removed after many iterations. Here we utilize the property of natural exponential function $y = e^x$ to explicitly enforce the weights of channels with stronger response are bigger than those with weaker response. Formally,

$$a_i = e^{\pm \triangle_i} = e^{s_i \cdot \triangle_i} = e^{sign(x_{ic} - \bar{x}_{ic}) \cdot \triangle_i} \quad (3)$$

where a_i means the weight of i^{th} channel; ± 1 represents comparison results derived by x_c and $\bar{x}_c$; $\triangle$ is a set of weights which mean different importance of different channels. $\triangle$ is generated by a two-layer fully connected network:

$$\triangle = sigmoid(w_2 \cdot tanh(w_1 \cdot x + b_1) + b_2) \quad (4)$$

where w and b are parameters of the network. $x : D \times 1$, $w_1 : (D/8) \times D, b1 : D/8, w2 : D \times (D/8), b2 : D$. D is the number of feature channels before classifier, which is 2048 and 1024 for Resnet50[4] and DLA60x[19] respectively.

3. EXPERIMENTS

We demonstrate FCCA on three public FGVC datasets: CUB200-2011[20], FGVC-Aircraft[21] and Stanford Cars[22], whose statistics are shown in Table 1. With respect to models, we have two general classification networks and one FGVC method as three baseline models. Finally, we conduct visualization works to further study FCCA.

Table 1. Statistics of FGVC datasets in this paper

Dataset	Object	# Category	# Training	# Testing
CUB200-2011	Bird	200	5994	5794
FGVC-Aircraft	Aircraft	100	6667	3333
Stanford Cars	Car	196	8144	8041

3.1. Baseline Models with Training Details

We choose three models as our baseline model: Resnet50[4], DLA60x[19] and DCL[15]. Resnet50 and DLA60x could be used directly in image classification. DCL is a FGVC method proposed recently, which achieved excellent performance on public FGVC datasets.

Resnet50 and DLA60x: The input images are first resized to keep the shorter side as 512 pixels. Then we follow the data augmentation strategy from [6] to crop regions whose sizes randomly range from 8% to 100% of original image area and whose aspect ratios randomly range from 3/4 to 4/3. Finally, the cropped regions are resized to 448×448. No color augmentation is implemented. Both models are first pretrained on ImageNet and fine tuned on target datasets. Fine tuning is optimized by SGD for 100 epochs with batch size 16, momentum 0.9 and weight decay 10^{-5}. The initial learning rate is 0.001 and reduced by 10 for every 30 epochs.

DCL: As shown in Fig. 2, the structure of DCL could be partitioned into feature extractor and classifier, which is similar to Resnet50 and DLA60x. We place FCCA between the feature extractor and classifier of DCL. Besides general classification loss, DCL has another loss for adversarial learning. But it does not influence how FCCA works in the framework. All experiment settings in this part are the same as those in original paper [15]. Learning rate is reduced by 10 for every 60 epochs. The data augmentation strategy includes resizing images to 512×512, randomly rotating images with 15 degrees, cropping images as 448×448 and randomly flipping images horizontally. To get the DCL baseline, we first reproduced the original network. Then we kept hyper-parameters unchanged to train DCL network with FCCA.

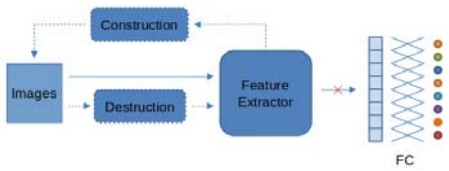


Fig. 2. The structure of DCL. FCCA is plugged in the red cross position.

It is noteworthy that when integrating FCCA into three baseline models, **all** hyper-parameters and experiment settings are kept completely the same for fair comparison, which could show the net improvements contributed by FCCA.

3.2. Top1 Accuracy Results

Top1 accuracy results of related FGVC methods as well as three baseline models with and without FCCA are presented in Table 2. From the upper part, we can find FCCA improves general classification models like Resnet50 and DLA60x by a large margin. For the advanced FGVC method DCL, we can still get the average net improvement of 0.8% on three datasets under the same hyper-parameters and training settings. On Stanford Cars and FGVC-Aircraft datasets, our FCCA pushes the DCL method to the state-of-the-art performance. On the CUB200-2011 dataset, FCCA bridges the gap between the method DCL and WS-DAN. The advantage of DCL with FCCA over WS-DAN is that we have merely one stage for testing while WS-DAN has one more stage to locate and enlarge object regions for further refinement.

3.3. Ablation Study

Ablation study is conducted to further quantitatively demonstrate the effectiveness of the FCM. As the FCM is essentially implemented by the sign of $\pm\triangle$, it is expected to remove the sign of $\pm\triangle$ and get $\triangle$. To make a fair comparison, we replace the non-linear activation function $sigmoid$ with $tanh$ at the end of 2-layer fully connected network and get weights O. In

Table 2. Top1 Accuracy of Models.

Approach		Stanford Cars (%)	Aircraft (%)	CUB200-2011 (%)
Resnet50[4]		91.5	87.8	84.0
DLA60x[19]		92.7	88.5	85.1
Resnet50+FCCA		92.9	89.0	87.0
DLA60x+FCCA		93.7	90.9	88.1
RA-CNN[13]		92.5	88.2	85.3
MA-CNN[14]		92.8	89.9	86.5
DFL-CNN[23]		93.8	92.0	87.4
ISE-CNN[24]		94.1	90.9	87.2
LFD[25]		94.1	92.1	87.6
WS-DAN[16]		94.5	93.0	**89.4**
DCL[15]	reported	94.5	93.0	87.8
	reproduce	93.9	92.4	87.6
	+FCCA	**94.8**	**93.2**	88.3

this way, it is guaranteed that O and $\pm\triangle$ have the same value range which is from -1 to 1. The implementation is:

$$a = e^{O} = e^{tanh(w_2 \cdot tanh(w_1 \cdot x + b_1) + b_2)} \tag{5}$$

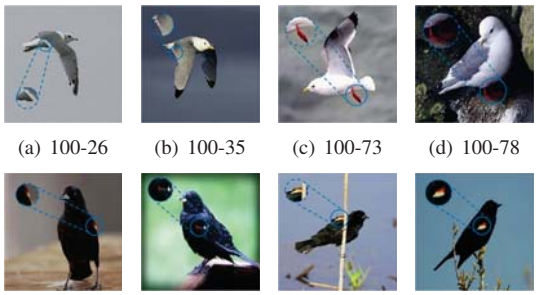
where w_1, w_2, b_1 and b_2 are the same as those in (4). Results in Table 3. prove that the FCM is essentially important for FCCA.

Table 3. Ablation Study Results

Approach	Stanford Cars (%)	Aircraft (%)	CUB200-2011 (%)
Resnet50	91.5	87.8	84.0
attention with O	91.1	87.7	83.1
attention with $\pm\triangle$	**92.9**	**89.0**	**87.0**

3.4. Visualization

Feature Comparison Mechanism: To qualitatively demonstrate the feasibility of the FCM, we check the relationship of visual patterns and comparison results from (2). We first train a Resnet50 model on CUB200-2011 and get the average response of every feature channel for each category. Given a category, we choose the most important 100 channels according to its classifier weights. These channels are chosen in this way for two reasons: 1. different feature channels have different weights for recognition. Choosing important feature channels means their corresponding visual patterns have positive contributions to recognition and are thus easily distinguished by humans; 2. channels are somehow coupled with each other so that a group of them is more integral and persuasive. Then we pick four representative samples from each category. For every sample, we take feature comparison among the 100 channels. Samples and comparison results are shown in Fig. 3. From Fig. 3(a) to 3(d) and Fig. 3(e) to 3(h), it can be seen that the more discriminative visual patterns a sample has, the more feature channels with stronger response the model has, which indicates that the FCM is feasible for recognizing discriminative regions for each category.



(a) 100-26 (b) 100-35 (c) 100-73 (d) 100-78

(e) 100-21 (f) 100-26 (g) 100-51 (h) 100-71

Fig. 3. Samples and comparison results. The category "Red Legged Kittiwake" and "Red Winged Blackbird", whose attentive patterns are leg or wing, are shown in the first and second row respectively. 100-n means there are n feature channels of a sample having stronger response than their counterpart average response of the category in the 100 channels.

CAM: We use the Class Activation Mapping (CAM)[26] to visualize attention maps from the last convolution layer of Resnet50. From Fig. 4, we could see that the attention masks generated by the model with FCCA are more accurate and refined, which indicates that FCCA helps the model to better locate discriminative regions.

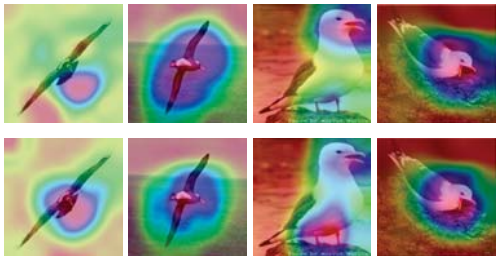


Fig. 4. Visualization of attention maps via CAM. In the top row are attention maps of the model without FCCA and in the bottom row are those of the model with FCCA.

4. CONCLUSION

In this paper, we propose a novel and effective channel attention for FGVC. Our method consists of two parts: One is feature comparison mechanism that recognizes whether discriminative visual patterns appear strongly or not; The other is weights assignment scheme which explicitly guarantees that discriminative feature channels have larger weights. Our method could help models to better catch and utilize information from discriminative regions, which improves or rivals the state-of-the-art performance of FGVC. Quantitative and qualitative experiments demonstrate its effectiveness.

5. REFERENCES

[1] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner, "Gradient-based learning applied to document recognition," 1998.

[2] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012.

[3] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Identity mappings in deep residual networks," *ArXiv*, vol. abs/1603.05027, 2016.

[6] Yangqing Jia Pierre Sermanet Scott Reed Dragomir Anguelov Dumitru Erhan Vincent Vanhoucke Andrew Rabinovich Christian Szegedy, Wei Liu, "Going deeper with convolutions," in *CVPR*, 2015.

[7] Sergey Ioffe Jon Shlens Christian Szegedy, Vincent Vanhoucke and Zbigniew Wojna, "Rethinking the inception architecture for computer vision," in *CVPR*, 2016.

[8] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alex Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *AAAI*, 2016.

[9] Ekin Dogus Cubuk, Barret Zoph, Dandelion Mané, Vijay Vasudevan, and Quoc V. Le, "Autoaugment: Learning augmentation policies from data," *ArXiv*, vol. abs/1805.09501, 2018.

[10] Shaoli Huang, Zhe Xu, Dacheng Tao, and Ya Zhang, "Part-stacked cnn for fine-grained visual categorization," in *CVPR*, 2016.

[11] Ning Zhang, Jeff Donahue, Ross B. Girshick, and Trevor Darrell, "Part-based r-cnns for fine-grained category detection," in *ECCV*, 2014.

[12] Yuxin Peng, Xiangteng He, and Junjie Zhao, "Object-part attention model for fine-grained image classification.," *IEEE transactions on image processing : a publication of the IEEE Signal Processing Society*, vol. 27 3, pp. 1487–1500, 2018.

[13] Jianlong Fu, Heliang Zheng, and Tao Mei, "Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition," in *CVPR*, 2017.

[14] Heliang Zheng, Jianlong Fu, Tao Mei, and Jiebo Luo, "Learning multi-attention convolutional neural network for fine-grained image recognition," in *ICCV*, 2017.

[15] Yue Chen, Yalong Bai, Wei Zhang, and Tao Mei, "Destruction and construction learning for fine-grained image recognition," in *CVPR*, 2019.

[16] Tao Hu and Honggang Qi, "See better before looking closer: Weakly supervised data augmentation network for fine-grained visual classification," *ArXiv*, vol. abs/1901.09891, 2019.

[17] Marcel Simon and Erik Rodner, "Neural activation constellations: Unsupervised part model discovery with convolutional networks," in *ICCV*, 2015.

[18] Xiaopeng Zhang, Hongkai Xiong, Wengang Zhou, Weiyao Lin, and Qi Tian, "Picking deep filter responses for fine-grained image recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1134–1142, 2016.

[19] Fisher Yu, Dequan Wang, and Trevor Darrell, "Deep layer aggregation," in *CVPR*, 2018.

[20] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge J. Belongie, "The caltech-ucsd birds-200-2011 dataset," 2011.

[21] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew B. Blaschko, and Andrea Vedaldi, "Fine-grained visual classification of aircraft," *ArXiv*, vol. abs/1306.5151, 2013.

[22] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei, "3d object representations for fine-grained categorization," *2013 IEEE International Conference on Computer Vision Workshops*, pp. 554–561, 2013.

[23] Yaming Wang, Vlad I. Morariu, and Larry S. Davis, "Learning a discriminative filter bank within a cnn for fine-grained recognition," in *CVPR*, 2018.

[24] Andrea Simonelli, Stefano Messelodi, Francesco De Natale, and Samuel Rota Bulo, "Increasingly specialized ensemble of convolutional neural networks for fine-grained recognition," in *ICIP*, 2018.

[25] Zhicong Feng, Keren Fuy, and Qijun Zhao, "Learning to focus and discriminate for fine-grained classification," in *ICIP*, 2019.

[26] Bolei Zhou, Aditya Khosla, Àgata Lapedriza, Aude Oliva, and Antonio Torralba, "Learning deep features for discriminative localization," in *CVPR*, 2016.