Fine-Grained Visual Classification with Batch Confusion Norm

Yen-Chi Hsu[1,2], **Cheng-Yao Hong**[1], **Ding-Jie Chen**[1], **Ming-Sui Lee**[2], **Davi Geiger**[3], **Tyng-Luh Liu**[1]

[1]Institute of Information Science, Academia Sinica

[2]Department of Computer Science & Information Engineering, National Taiwan University

[3]Computer Science, New York University

Abstract

We introduce a regularization concept based on the proposed Batch Confusion Norm (BCN) to address Fine-Grained Visual Classification (FGVC). The FGVC problem is notably characterized by its two intriguing properties, significant inter-class similarity and intra-class variations, which cause learning an effective FGVC classifier a challenging task. Inspired by the use of pairwise confusion energy as a regularization mechanism, we develop the BCN technique to improve the FGVC learning by imposing class prediction confusion on each training batch, and consequently alleviate the possible overfitting due to exploring image feature of fine details. In addition, our method is implemented with an attention gated CNN model, boosted by the incorporation of Atrous Spatial Pyramid Pooling (ASPP) to extract discriminative features and proper attentions. To demonstrate the usefulness of our method, we report state-of-the-art results on several benchmark FGVC datasets, along with comprehensive ablation comparisons.

Introduction

Fine-grained visual classification (FGVC) is an active and challenging problem in computer vision. Such a recognition task differs from the classical problem of large-scale visual classification (LSVC) by focusing on differentiating *similar* sub-categories of the same meta-category. In FGVC, the inter-class similarity among the object categories is often pervasive, while the intra-class variations further impose ambiguities in learning a unified and discriminative representation for each category. To tackle the difficulties in addressing FGVC, most of the recent research efforts have converged to learn pivotal local/part details relevant to distinguishing fine-grained categories *e.g.*, (Fu, Zheng, and Mei 2017; Yang et al. 2018; Zheng et al. 2019), and typically require the fusion of several sophisticated computer vision techniques to accomplish the task such as in (Ge, Lin, and Yu 2019). We instead aim to establish a useful and generic methodology, centered around the proposed *batch confusion norm*, for more effectively learning a DNN model to the underlying FGVC problem.

Figure 1 illustrates the two aspects of paradoxes in FGVC where the inter-class similarity and the intra-class variations

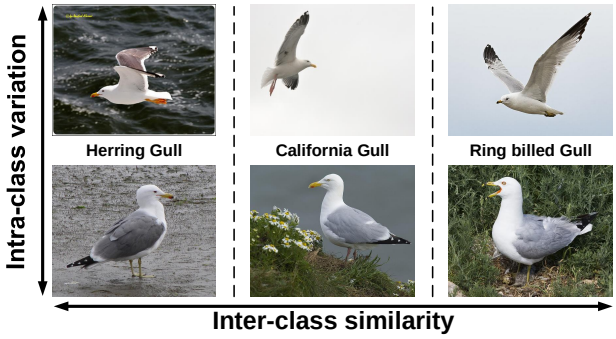


Figure 1: Inter-class similarity vs. intra-class variation: Each column includes two instances of a specific "Gull" category from the CUB-200-2011 dataset (Wah et al. 2011).

are subtly intertwined, yielding a daunting classification task. For humans, the example convincingly suggests that expert knowledge is needed to differentiate one from the other two categories. On the other hand, it also exhibits the challenges of formulating universal criteria in developing machine learning frameworks to solve the FGVC problem even for a modest case involving just three object categories.

It goes without saying that techniques based on deep neural networks have been the focal point of the recent development in tackling FGVC. Characterized by powerful model capacity and end-to-end feature learning, these state-of-the-art approaches are craftily designed to extract discriminative local details and consistent global structure, and shown to achieve significant improvements over conventional non-DNN approaches, *e.g.*, (Duan et al. 2012) on almost all FGVC benchmark datasets. However, the improvement for solving FGVC by exploring visual features of different levels and resolutions from relevant regions seems to be saturated and the extensive learning over an underlying FGVC dataset, whose number of images per category is considerably fewer than that in LSVC, may eventually run into the problem of overfitting. To alleviate such a a concern, Dubey et al., (2018) introduce the *pairwise confusion energy* to exert a regularization effect that guides the resulting DNN model to achieve good FGVC inference without overfitting. In this work, we further gener-

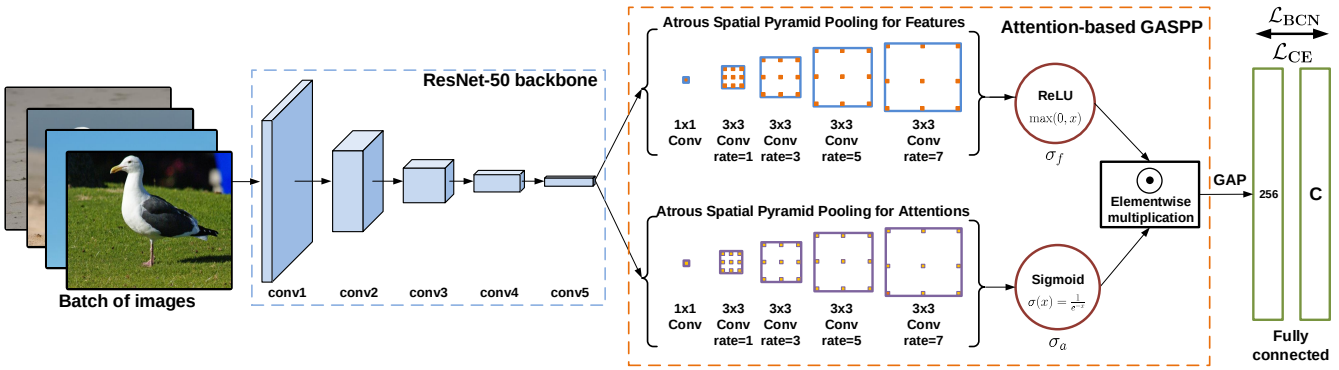


Figure 2: The proposed neural network architecture Φ for tackling the FGVC task.

alize the *confusion* concept, and propose the batch confusion norm, which can be seamlessly plugged into the total loss function, as a universal and generic regularization technique of improving the DNN learning for solving the FGVC task.

Related Work

Research in fine-grained visual classification has made significant progress since the introductory uses of deep learning techniques for solving the task (Simonyan and Zisserman 2014; He et al. 2016; Huang et al. 2017). The current FGVC techniques have now evolved into several branches. When the training data are annotated with additional information such as part labels, it is reasonable to aggregate the visual cues from assorted parts to achieve the FGVC. Along this line, (Berg et al. 2014) explore the labeled part locations to eliminate highly similar object categories for improving the learned classifiers. The approach by (Huang et al. 2016) is established based on a two-stream classification network to explicitly capture both object-level and part-level information. However, owing to the rapid research advances in visual classification, the majority of recent FGVC approaches are designed to complete the model learning solely based on the information of category label (Fu, Zheng, and Mei 2017; Dubey et al. 2018; Wang, Morariu, and Davis 2018; Yu et al. 2018a; Yang et al. 2018; Chen et al. 2019; Zheng et al. 2019).

Without referencing the part annotations but the category labels, existing FGVC techniques are shown to be effective in learning the discriminative regions. Peng, He, and Zhao, (2017) propose a weakly-supervised formulation for fine-grained image classification such that heavy labeling efforts for both object and part annotations can be avoided. Wang, Morariu, and Davis, (2018) consider a bank of discriminative filters to accurately locate the discriminative patches. Yang et al., (2018) develop a multi-agent cooperative learning scheme to identify informative regions. Sun et al. (2018) first use the *one-squeeze multi-excitation module* (OSME) to localize different parts, and then train the model with a multi-attention multi-class constraint to coherently enforce the correlations among different parts in training. (Zheng et al. 2019) introduce a *trilinear attention sampling network* (TASN) to learn rich feature representations from hundreds of part proposals. (Ge, Lin, and Yu 2019) propose to build complementary parts via object detection and instance segmentation techniques. Note that even though (Ge, Lin, and Yu 2019) utilize only the category labels in training, the object detection module is indeed pre-trained on the COCO dataset.

The confusion-related formulation for dealing with intra-class variations and inter-class similarity in FGVC have two main implications. First, it can be applied to alleviate the overfitting problem in training a FGVC model. (Dubey et al. 2018) constructs a Siamese neural network, trained with a loss function including *pairwise confusion* (PC). The reasoning behind the design is that bringing the class probability distributions closer to each other could prevent the learned FGVC model from overfitting sample-specific artifacts. Second, the confusion tactic can be used to boost the FGVC performance by focusing on local evidence. (Chen et al. 2019) partitions each training image into several local regions and then shuffle them by a *region confusion mechanism* (RCM). It implicitly excludes the information about the global object structure and forces the model to predict the category label based on local information. In other words, the ability of identifying the object category from local details is expected to be enhanced through shape confusion.

Our approach to FGVC is most relevant to the above confusion-based approaches. Inspired by the concept of pairwise confusion energy in (Dubey et al. 2018), we generalize the regularization to simultaneously account for the class softmax distributions of a whole training batch. The major advantage of the proposed *batch confusion norm* (BCN) is to enforce the confusion of class predictions over a batch set rather than just a pair of training samples. The generalization tends to encourage the learning to further uncover discriminative features for distinguishing all the samples within a batch; otherwise, the training loss, already penalized by the BCN regularization, would remain substantial and lead to a failed training. In addition to the proposed BCN, we adopt the Gated-CNN (Yu et al. 2018b) and use the *atrous spatial pyramid pooling* (ASPP) (Chen et al. 2017) to form the attention-based GASPP network. (See Figure 2.) As we will demonstrate, the GASPP-CNN is an appropriate implementation for the FGVC task in which it can effectively aggregate visual attentions by considering various field-of-views.

Our Method

The core of the proposed method is a new regularization technique, namely the proposed BCN, to not only alleviate the overfitting problem in training a FGVC model but also boost the classification performance. In addition, we design the DNN backbone inspired by the Gated-CNN architecture, which is further enhanced by the incorporation of the ASPP mechanism to explore fine-grained details. We next describe these essential components of the proposed framework.

Batch Confusion Norm

Let Φ be the FGVC model as illustrated in Figure 2 and $\mathcal{D}$ be the training set over totally C fine-grained categories. An arbitrary sample from $\mathcal{D}$ is denoted as $(\mathbf{x}, \mathbf{y})$ where in our case $\mathbf{x}$ is an image and $y \in \{1, \ldots, C\}$ is the corresponding class label. In learning Φ, we follow the standard batch training and set the batch size to include M images. However, we assume that $M \leq C$ and all images within a batch $\mathcal{B}$ are of distinct class labels.

For each training sample $\mathbf{x}_i \in \mathcal{B}$, forward propagation through Φ would yield a class probability (*i.e.*, softmax) distribution $\mathbf{p}_i \in \mathbb{R}^{C\times 1}$. We can then define the batch-wise class prediction matrix by

$$P = [\mathbf{p}_1\ \mathbf{p}_2\ \ldots\ \mathbf{p}_M] \in \mathbb{R}^{C\times M} \tag{1}$$

where each $\mathbf{p}_i$ is the softmax class prediction over the C fine-grained categories. The purpose of BCN is to infuse *slight* classification confusions into the FGVC training procedure and drive the learning to work harder for making as many correct predictions in each training bag as possible. To this end, it is reasonable to minimize the rank of the batch prediction matrix P so that all individual predictions are *similar*. However, the rank-related minimization problems are often NP-hard, and convex relaxations are instead introduced to approximate the solutions. In our formulation, minimizing the rank of P is reduced to minimizing its *nuclear norm*. We define the *batch confusion norm* of P as

$$\|P\|_{\text{BCN}} = \|P\|_* \tag{2}$$

where $\|\cdot\|_*$ is the nuclear norm that computes the sum of the singular values of the underlying tensor/matrix. To incorporate the batch confusion energy into the total loss function for FGVC training, we have

$$\mathcal{L}_{\text{BCN}} = \|PP^\top\|_{\text{BCN}} \tag{3}$$

where the batch confusion loss $\mathcal{L}_{\text{BCN}}$ is computed based on the eigenvalues of $PP^\top$. We see that when the batch size $M = 2$, the batch confusion loss in (3) is reduced to the pairwise confusion energy in (Dubey et al. 2018).

Figure 3 illustrates the following three scenarios: Training a FGVC network (a) with only the conventional Cross-Entropy (CE) loss, (b) with the additional regularization of Pairwise Confusion (PC) energy, and (c) with the additional regularization of Batch Confusion Norm (BCN). In the first case, the CE loss focuses on the intra-class variations. The latter two cases are formulated to deal with both the intra-class variations (w.r.t. CE) and the inter-class similarity (w.r.t.

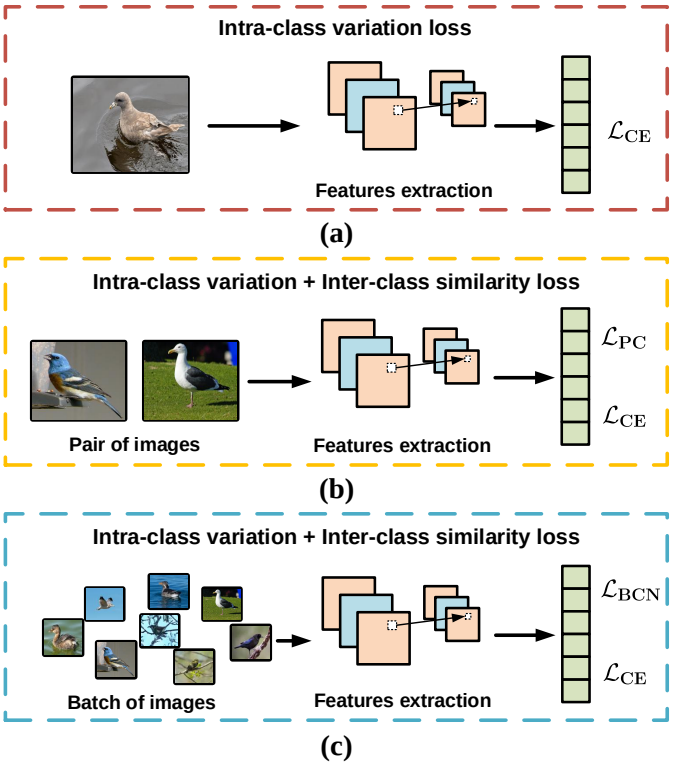


Figure 3: Learning to FGVC with (a) conventional Cross Entropy (CE) loss, (b) + Pairwise Confusion (PC) energy and (c) + the proposed Batch Confusion Norm (BCN).

CE+PC or CE+BCN). Nevertheless. the proposed BCN regularization has the advantage to explore the two aspects of classification factors based on a batch set rather than a pair of training samples.

Attention Gated ASPP

Let $F \in \mathbb{R}^{c\times h\times w}$ be the feature maps obtained by the last convolutional layer of the ResNet backbone as shown in Figure 2. We extend the network with two streams: one for learning discriminative features and the other for uncovering the proper attention responses. In both streams, we use the *atrous spatial pyramid pooling* (ASPP) technique to simultaneously extract features/attentions from different field-of-views.

In the feature stream, performing ASPP would covert F into $\text{ASPP}_f(F)$, while in the attention stream, the similar procedure would yield attention feature maps, $\text{ASPP}_a(F)$. Note that parameters in the two streams are not shared but learned jointly. After regulating with respect to the respective activation function, we carry out the gated element-wise product to output the adjusted feature maps $\tilde{F}$, weighted by the predicted attentions. That is, the attention-gated ASPP feature maps $\tilde{F}$ are derived by

$$\tilde{F} = \sigma_f(\text{ASPP}_f(F)) \odot \sigma_a(\text{ASPP}_a(F)) \tag{4}$$

where $\tilde{F} \in \mathbb{R}^{c\times h\times w}$ remains the same dimensions, $\odot$ denotes the element-wise product, σ_f in our implementation is

ReLU and σ_a is the $\mathrm{sigmoid}$ function for gating. The ASPP operation is similar to that in (Chen et al. 2017) except that we use different dilated rates. In summary, $\mathrm{ASPP}_f(\cdot)$ would learn the image features across the whole spatial domain and $\mathrm{ASPP}_a(\cdot)$ instead predicts the attention heatmaps. The gated fusion between the two streams leads to the output of discriminative feature maps $\tilde{F}$ for FGVC.

Loss Function

From (3) and the network architecture in Figure 2, the refined feature maps $\tilde{F}$ are followed by a fully-connected softmax layer to output the class prediction vector $\mathbf{p}$. The overall loss function can now be readily expressed by

$$\mathcal{L} = \mathcal{L}_{\mathrm{CE}} + \lambda \mathcal{L}_{\mathrm{BCN}} \tag{5}$$

where λ is a hyper-parameter to adjust the influence of the BCN loss to learning the proposed FGVC model.

Experimental Results

To test the effectiveness of our method, we carry out extensive experiments on three benchmark datasets. The experimental results including comparisons to prior work, and the implementation details are thoroughly described. We also provide an insightful ablation study for assessing the performance gains of using BCN and GASPP. Finally, a number of visualization examples are demonstrated for further discussions.

FGVC Datasets

Table 1: Statistics of three datasets.

Dataset	# Train	# Test	# Category
CUB-200-2011	5,994	5,794	200
Stanford Cars	8,144	8,041	196
FGVC-Aircraft	6,667	3,333	100

We evaluate the effectiveness of the proposed approach on three standard fine-grained visual classification datasets, namely CUB-200-2011 (Wah et al. 2011), Stanford Cars (Krause et al. 2013), and FGVC-Aircraft (Maji et al. 2013). Table 1 shows the detailed statistics with the numbers of training and testing splits along with the category numbers. The data ratio between the training and the testing is about $1:1$ for CUB-200-2011 and Stanford Cars, and is about $2:1$ in FGVC-Aircraft. The class distribution of the three datasets is nearly balanced. Compared with other datasets for the LSVC task, the datasets for FGVC task have obviously few training data for each category. Notice that the proposed model does not require any additional annotations in the training process but merely the image-level class annotations.

Implementation Details

We evaluate our approach on three widely used classification networks: ResNet-50 (He et al. 2016), VGG-16 (Simonyan and Zisserman 2014), and DenseNet-161 (Huang et al. 2017). These backbone networks are pre-trained on ImageNet dataset. The input images are resized to 600×600 and cropped randomly into 448×448 while training. We use random horizontal flip for data augmentation in the training stage. The initial learning rate and the hyper-parameter λ are 0.008 and 10. The training batch size is usually 16 and 8 for DenseNet-161. The training optimizer is Momentum SGD, which accompanying with cosine annealing (Loshchilov and Hutter 2016) as the learning rate decay. Our approach is implemented by Pytorch, and the source code will be available online.

Performance Comparison

This section compares the performance in classification accuracy of our approach against state-of-the-art methods. Table 2 summaries the comparison results on three datasets: CUB-200-2011, Stanford Cars, and FGVC-Aircraft. In Table 2, the results show that the classification backbone ResNet-50 equipped with BCN have already brought the comparable performance in compared with the other FGVC methods, and the performance gain boosts obviously while replacing the backbone network to the better one, namely DenseNet-161, or equipping with another proposed GASPP.

The performance of the classification backbone ResNet-50 equipped with GASPP and BCN surpasses all other FGVC methods excepts the Stacked LSTM (Ge, Lin, and Yu 2019). Please notice that, although Stacked LSTM has the best result on CUB-200-2011, its three-step training requires to use Mask R-CNN pre-trained on additional COCO dataset. Hence, it is not clear whether the high performance is deriving from the additional training data. In this comparison, it is evident that the proposed GASPP and BCN are beneficial to address the FGVC task.

It is interesting that the proposed BCN shows the dominant power in compared with the PC (Dubey et al. 2018), even that our BCN is merely applied on the basic classification backbone ResNet-50. This comparison demonstrates that to enforce the confusion of class predictions over a batch set rather than just a pair of training sample via the proposed batch confusion norm is beneficial to the FGVC task.

Ablation Study

This experiment compares various configurations of the proposed approach for assessing and understanding the relative importance of each setting. The study results are shown in Figure 4, Table 3, Table 4, Table 5, and Table 6.

The Weight of Batch Confusion Norm. In our approach, the hyper-parameter λ plays the role to control the weight of BCN in the loss function. The higher value of λ means the higher weight of BCN during the loss calculation, which analogies to take care of more impacts from other images within the same batch in training. Figure 4 and Table 3 show the influence degrees that the batch confusion norm affect the training loss and classification accuracy.

Figure 4 shows the influence of BCN in respect of the training loss. This comparison shows that training our model in various value of λ has similar convergence speeds but has noticeable saturated training losses. The higher weight of BCN brings the larger training loss, and the larger training

Table 2: Comparison of FGVC methods in classification accuracy on three datasets.

Method	Backbone	Accuracy (%)		
		CUB-200-2011	Stanford Cars	FGVC-Aircraft
B-CNN (Lin, RoyChowdhury, and Maji 2015)	VGG-16	84.1	91.3	84.1
KP (Cui et al. 2017)	VGG-16	86.6	92.4	86.9
LRBP (Kong and Fowlkes 2017)	VGG-16	84.2	90.9	87.3
HIHCA (Cai, Zuo, and Zhang 2017)	VGG-16	85.3	91.7	88.3
Improved B-CNN (Lin and Maji 2017)	VGG-16	85.8	92.0	88.5
HBP (Yu et al. 2018a)	VGG-16	87.1	93.7	90.3
MA-CNN (Zheng et al. 2017)	VGG-19	86.5	92.8	89.9
RA-CNN (Fu, Zheng, and Mei 2017)	VGG-19	85.3	92.5	88.2
FCAN (Liu et al. 2016)	ResNet-50	84.7	93.1	-
NTS-Net (Yang et al. 2018)	ResNet-50	87.5	93.9	91.4
DFL-CNN (Wang, Morariu, and Davis 2018)	ResNet-50	87.4	93.1	91.7
DCL (Chen et al. 2019)	ResNet-50	87.8	94.5	93.0
TASN (Zheng et al. 2019)	ResNet-50	87.9	93.8	-
Stacked LSTM (Ge, Lin, and Yu 2019)	GoogleNet	**90.3**[*]	-	-
PC (Dubey et al. 2018)	DenseNet-161	86.7	92.9	89.2
BCN	ResNet-50	87.7	94.3	90.3
GASPP + BCN	ResNet-50	88.4	94.6	**93.5**
BCN	DenseNet-161	**89.2**	**94.8**	**93.5**

[*] Stacked LSTM exploits additional training data from COCO dataset during its three-step training.

loss means the model is harder to over-fit the training data. Note that the standard FGVC datasets often have a small number of training images for each category; hence, it is easy for a deep neural network to over-fit the training data. The blue line in Figure 4 shows that it is easy to over-fit the CUB-200-2011 dataset using ResNet-50 without applying BCN. Therefore, the proposed BCN can be used to reduce the over-fitting issue in FGVC task.

In Table 3, the classification accuracy rapidly increases with respect to larger the value of λ. The performance gain on accuracy saturates quickly and keeps stable between $\lambda = 5$ and $\lambda = 50$. This comparison shows that the classification accuracy can be improved noticeably with the aids of BCN. However, considering too many impacts from other images in the same training batch, namely using the large value of λ, could decay the advantage of BCN. Therefore, we suggest setting the value of λ as 10 in FGVC task.

The compatibility of Batch Confusion Norm. The proposed BCN is used to exert a regularization effect that guides the trained DNN model to tackle the overfitting issue. It is expected to be a universal and generic regularization technique, which can be seamlessly plugged into the total loss function to improve the DNN learning for solving the FGVC task. Hence, we conduct this comparison to plugged BCN into the existing loss function of various backbones. The results are shown in Table 4.

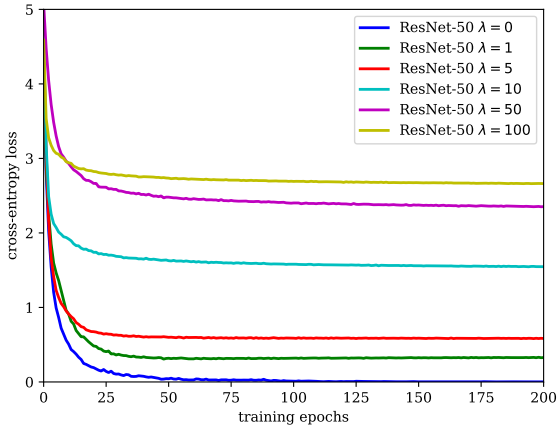


Figure 4: The BCN influence in training loss against to different λ values on CUB-200-2011 dataset.

Table 4 shows the study results on different backbones such as VGG-16 (Simonyan and Zisserman 2014), ResNet-50 (He et al. 2016), and DenseNet-161 (Huang et al. 2017). The results show that all these backbones obtain a significant improvement with plugging the BCN into their original loss functions.

Table 3: The BCN influence in classification accuracy evaluated on CUB-200-2011 dataset. The hyper-parameter λ is used to control the weight of BCN in the loss function.

Method	λ	Accuracy (%)
ResNet-50	0	85.5
ResNet-50 + BCN	1	86.2
ResNet-50 + BCN	5	87.3
ResNet-50 + BCN	10	**87.7**
ResNet-50 + BCN	50	87.5
ResNet-50 + BCN	100	86.5

Table 4: The compatibility of BCN and various backbones.

Backbone	Accuracy (%)		
	CUB	CAR	AIR
VGG-16	76.3	90.1	86.9
VGG-16 + BCN	**78.8**	**91.5**	**90.1**
ResNet-50	85.5	92.7	90.3
ResNet-50 + BCN	**87.7**	**94.3**	**93.2**
DenseNet-161	87.5	93.4	92.7
DenseNet-161 + BCN	**89.2**	**94.8**	**93.5**

Table 5: Comparison of the various training batch sizes of our approach on CUB-200-2011.

Batch Size	Accuracy (%)				
	2	4	8	16	32
ResNet-50 + BCN	81.2	86.1	87.6	87.7	87.7

Various Batch Sizes. In BCN, the batch size is a key factor to exert the regularization effect concerning the other images within the same batch. In Table 5, the results show that the performance gain increases rapidly and saturates around sixteen images per batch. Please notice that the case of batch size is equal to two, namely the situation considered by PC (Dubey et al. 2018). While the batch size is equal to two, our result on CUB-200-2011 is 81.2%, which is still better than the result of PC, namely 80.2%.

Various Model Configurations. This part compares several configurations of the proposed model for assessing the relative importance of each component, and the results evaluated on CUB-200-2011 dataset are shown in Table 6. For the cases of respectively plugged the BCN, the ASPP, or the GASPP into the backbone ResNet-50, the performance gains are 2.2%, 0.6%, and 0.8%. The results show that the BCN brings the most improvement, and GASPP provides better performance gain compared with ASPP. Once we simultaneously apply BCN and GASPP, the performance gain is further improved to 2.9% in compared with the basic backbone ResNet-50. This study shows that the attention mechanism within the proposed GASPP is better than the ASPP in FGVC task, and the combination of BCN and GASSP make our approach achieve the state-of-the-art on three standard FGVC datasets in end-to-end training without using any additional annotations.

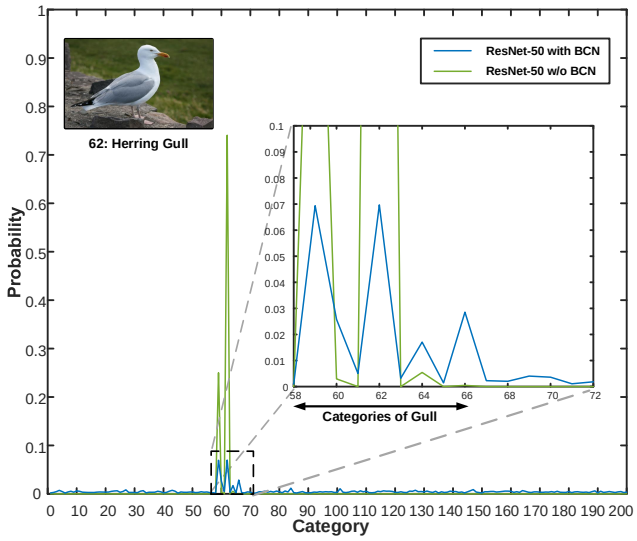
Table 6: Comparison of the various configurations of our approach on CUB-200-2011.

Backbone	Approach	Accuracy (%)
ResNet-50	None	85.5
ResNet-50	BCN	87.7
ResNet-50	ASPP	86.1
ResNet-50	GASPP	86.3
ResNet-50	ASPP + BCN	88.2
ResNet-50	GASPP + BCN	**88.4**

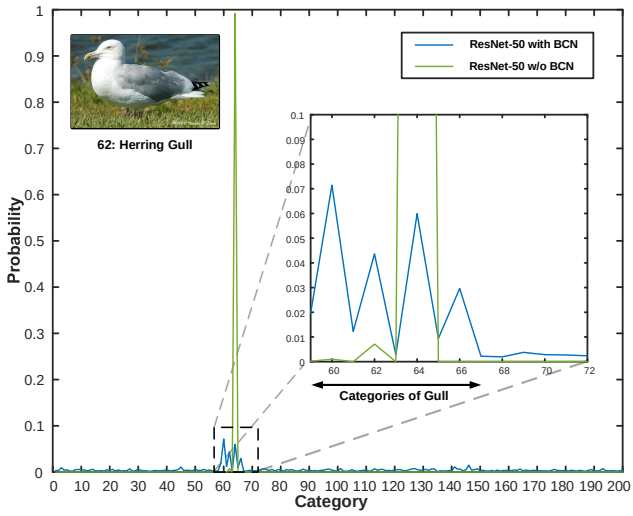
Discussions

To investigate what advantage that BCN brings to us, we have shown the prediction layer with *softmax* activation at Figure 5. In standard deep neural networks for classification, the loss function for prediction layer is usually just cross-entropy. Although other methods, such as NTS-Net (Yang et al. 2018), DCL (Chen et al. 2019) and TASN (Zheng et al. 2019), have many loss functions, their loss function for prediction layer remains the same. While only cross-entropy is used to learn the prediction layer, the output probability will become very close to one-hot vector even thought the prediction is wrong. This does not make sense in real world. Usually, we will hesitate on several very similar categories on fine-grained classification task whether we can recognize it successfully or not. This means that these categories should have certain probability values and the categories that are not considered would have low probability. With the help of BCN, we have successfully taught neural networks such characteristics, and therefore show better results.

BCN provides several benefits. First, to avoid overfitting of the cross-entropy loss. We usually make the ground truth label be a one-hot vector which will cause the inter-class similarity information to disappear. Hence, if on the loss function of cross-entropy ($\mathcal{L}_{\text{CE}}$) has the phenomenon of overfitting, usually can't learn the fine-grained essence. Second, BCN forces the model to learn the inter-class similarity which would make the classifier more focus on the discriminative parts. This phenomenon can be found by using Class Activation Mapping (CAM) (Zhou et al. 2016) presented in Figure 6. Without any attention mechanism and additional annotation, we successfully make the attention region smaller and accurately attend on the discriminative part only by BCN. Third, BCN is very practical and convenient. It does not require additional processing of data input or output during training. There is also no extra cost at inference time.



(a)



(b)

Figure 5: (a) Both models have correct predictions. (b) Both models give error predictions. Before using the additional BCN loss term, the model seems very confident even though making a wrong prediction. With the aids of BCN, the model try to exert a regularization effect from other images in the same batch, hence hesitates in several similar categories.

Conclusions

We have developed a new regularization technique specifically for addressing the FGVC problem. The proposed Batch Confusion Norm (BCN), together with the standard Cross Entropy loss can be used to account for the inherent classification difficulties due to inter-class similarity and intra-class variations. The proposed BCN considers the confusion regularization within each training batch, and thus is more general than the relevant formulation of pairwise confusion energy. In addition, our method is implemented with an attention-gated

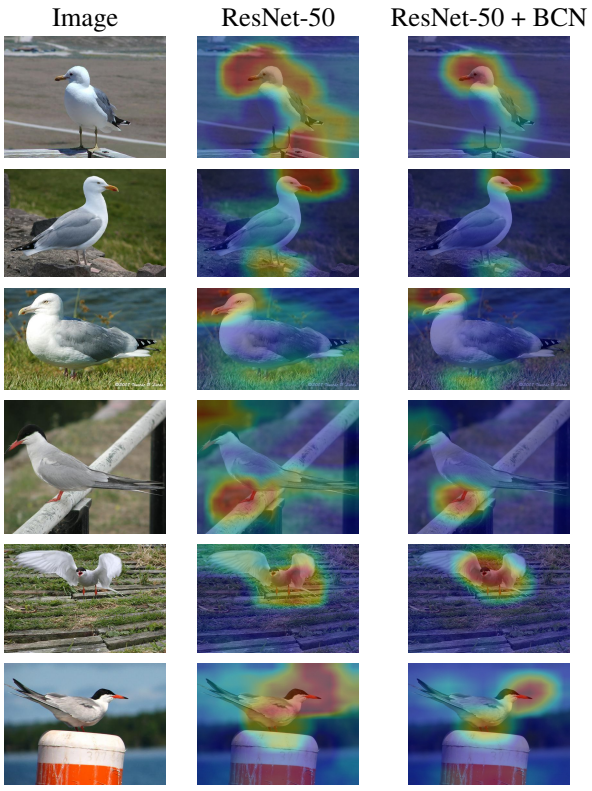


Figure 6: Visualization of the implicit attention on the images using CAM. The proposed BCN makes the model to learn the inter-class similarity which is not suitable for differentiating categories in FGVC task. Hence, the model would make the attention map become more sharp at prediction layer.

network, boosted with the use of Atrous Spatial Pyramid Pooling (ASPP) technique. The resulting model is shown to be capable of learning discriminative features within regions of interest and alleviating the overfitting problem in training. The provided experimental results are state-of-the-art over the three mainstream FGVC datasets. Our future work will focus on generalizing the BCN concept to tensors and also on extending its applications to other challenging computer vision problems.

References

Berg, T.; Liu, J.; Woo Lee, S.; Alexander, M. L.; Jacobs, D. W.; and Belhumeur, P. N. 2014. Birdsnap: Large-scale fine-grained visual categorization of birds. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2011–2018.

Cai, S.; Zuo, W.; and Zhang, L. 2017. Higher-order integration of hierarchical convolutional activations for fine-grained visual categorization. In *Proceedings of the IEEE International Conference on Computer Vision*, 511–520.

Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; and Yuille, A. L. 2017. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully

connected crfs. *IEEE transactions on pattern analysis and machine intelligence* 40(4):834–848.

Chen, Y.; Bai, Y.; Zhang, W.; and Mei, T. 2019. Destruction and construction learning for fine-grained image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5157–5166.

Cui, Y.; Zhou, F.; Wang, J.; Liu, X.; Lin, Y.; and Belongie, S. 2017. Kernel pooling for convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2921–2930.

Duan, K.; Parikh, D.; Crandall, D.; and Grauman, K. 2012. Discovering localized attributes for fine-grained recognition. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 3474–3481. IEEE.

Dubey, A.; Gupta, O.; Guo, P.; Raskar, R.; Farrell, R.; and Naik, N. 2018. Pairwise confusion for fine-grained visual classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 70–86.

Fu, J.; Zheng, H.; and Mei, T. 2017. 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*, 4438–4446.

Ge, W.; Lin, X.; and Yu, Y. 2019. Weakly supervised complementary parts models for fine-grained image classification from the bottom up. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3034–3043.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Huang, S.; Xu, Z.; Tao, D.; and Zhang, Y. 2016. Part-stacked cnn for fine-grained visual categorization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1173–1182.

Huang, G.; Liu, Z.; Van Der Maaten, L.; and Weinberger, K. Q. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4700–4708.

Kong, S., and Fowlkes, C. 2017. Low-rank bilinear pooling for fine-grained classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 365–374.

Krause, J.; Stark, M.; Deng, J.; and Fei-Fei, L. 2013. 3d object representations for fine-grained categorization. In *4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13)*.

Lin, T., and Maji, S. 2017. Improved bilinear pooling with cnns. In *British Machine Vision Conference 2017, BMVC 2017, London, UK, September 4-7, 2017*.

Lin, T.-Y.; RoyChowdhury, A.; and Maji, S. 2015. Bilinear cnn models for fine-grained visual recognition. In *Proceedings of the IEEE international conference on computer vision*, 1449–1457.

Liu, X.; Xia, T.; Wang, J.; and Lin, Y. 2016. Fully convolutional attention localization networks: Efficient attention localization for fine-grained recognition. *CoRR* abs/1603.06765.

Loshchilov, I., and Hutter, F. 2016. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*.

Maji, S.; Kannala, J.; Rahtu, E.; Blaschko, M.; and Vedaldi, A. 2013. Fine-grained visual classification of aircraft. Technical report.

Peng, Y.; He, X.; and Zhao, J. 2017. Object-part attention model for fine-grained image classification. *IEEE Transactions on Image Processing* 27(3):1487–1500.

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

Sun, M.; Yuan, Y.; Zhou, F.; and Ding, E. 2018. Multi-attention multi-class constraint for fine-grained image recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 805–821.

Wah, C.; Branson, S.; Welinder, P.; Perona, P.; and Belongie, S. 2011. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology.

Wang, Y.; Morariu, V. I.; and Davis, L. S. 2018. Learning a discriminative filter bank within a cnn for fine-grained recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4148–4157.

Yang, Z.; Luo, T.; Wang, D.; Hu, Z.; Gao, J.; and Wang, L. 2018. Learning to navigate for fine-grained classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 420–435.

Yu, C.; Zhao, X.; Zheng, Q.; Zhang, P.; and You, X. 2018a. Hierarchical bilinear pooling for fine-grained visual recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 574–589.

Yu, J.; Lin, Z.; Yang, J.; Shen, X.; Lu, X.; and Huang, T. S. 2018b. Free-form image inpainting with gated convolution. *arXiv preprint arXiv:1806.03589*.

Zheng, H.; Fu, J.; Mei, T.; and Luo, J. 2017. Learning multi-attention convolutional neural network for fine-grained image recognition. In *Proceedings of the IEEE international conference on computer vision*, 5209–5217.

Zheng, H.; Fu, J.; Zha, Z.-J.; and Luo, J. 2019. Looking for the devil in the details: Learning trilinear attention sampling network for fine-grained image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5012–5021.

Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; and Torralba, A. 2016. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2921–2929.