Filtration and Distillation: Enhancing Region Attention for Fine-Grained Visual Categorization

Chuanbin Liu, Hongtao Xie,[*] Zheng-Jun Zha, Lingfeng Ma, Lingyun Yu, Yongdong Zhang
School of Information Science and Technology, University of Science and Technology of China
{lcb592, mlf123, yuly}@mail.ustc.edu.cn, {htxie, zhazj, zhyd73}@ustc.edu.cn

Abstract

Delicate attention of the discriminative regions plays a critical role in Fine-Grained Visual Categorization (FGVC). Unfortunately, most of the existing attention models perform poorly in FGVC, due to the pivotal limitations in *discriminative regions proposing* and *region-based feature learning*. 1) The discriminative regions are predominantly located based on the filter responses over the images, which can not be directly optimized with a performance metric. 2) Existing methods train the region-based feature extractor as a one-hot classification task individually, while neglecting the knowledge from the entire object. To address the above issues, in this paper, we propose a novel *"Filtration and Distillation Learning" (FDL)* model to enhance the region attention of discriminate parts for FGVC. Firstly, a *Filtration Learning (FL)* method is put forward for discriminative part regions proposing based on the matchability between proposing and predicting. Specifically, we utilize the proposing-predicting matchability as the performance metric of Region Proposal Network (RPN), thus enable a direct optimization of RPN to filtrate most discriminative regions. Go in detail, the object-based feature learning and region-based feature learning are formulated as "teacher" and "student", which can furnish better supervision for region-based feature learning. Accordingly, our FDL can enhance the region attention effectively, and the overall framework can be trained end-to-end without neither object nor parts annotations. Extensive experiments verify that FDL yields state-of-the-art performance under the same backbone with the most competitive approaches on several FGVC tasks.

Introduction

Fine-grained visual categorization (FGVC) attracts extensive research attention in Artificial Intelligence (He and Peng 2017)(Liu et al. 2017), which aims to distinguish objects from different subordinate-level categories within a general category, such as bird species (Wah et al. 2011), dog breeds (Khosla et al. 2011), car models (Krause et al. 2013), air-craft models (Maji et al. 2013), flowers categories, etc. Due to the visual similarity in object appearances, the categorization highly relies on the differences hidden in subtle and local regions. Therefore, the delicate attention of the discriminative regions plays a critical role in FGVC. Correspondingly, it brings challenges in *discriminative regions proposing* and *region-based feature learning* (Figure 1).

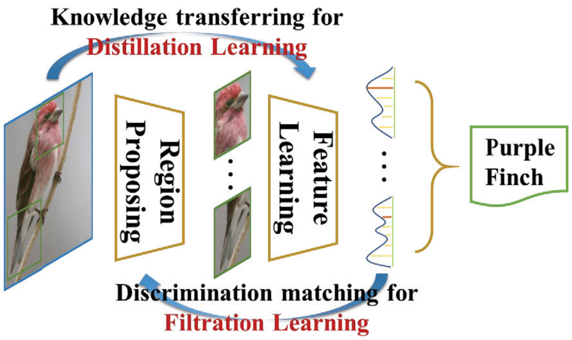

Figure 1: Illustrations of our proposed Filtration and Distillation Learning framework. Our FDL model creates effective supervision for discriminative regions proposing and region-based feature learning, thus enhance region attention for FGVC. [Best viewed in color with zoom-in]

The delicate region attention first relays on accurate discriminative regions proposing. Over the past few years, a variety of regions proposing methods have been proposed relying on human-annotated bounding box/part annotations(Zhang et al. 2014)(Wang et al. 2016)(Krause et al. 2015)(Lin et al. 2015). However, it requires specialized knowledge and a large amount of annotation time during labeling, making those methods less applicable in practice. To overcome these problems, researchers begin focusing on weakly supervised regions proposing without human bounding box/part annotations. These methods usually utilize filter response to detect the corresponding discriminative region (Zhang et al. 2016)(Fu, Zheng, and Mei 2017)(Zheng et al. 2017)(Zheng et al. 2019). Therefore, the regions proposing can only get indirect optimization accompanied by the improvement of feature extractor. Although promising results have been reported, the performance metric of region proposing is still a blank area in researches. And the absence

[*] Corresponding author

of the directly optimizing paradigm in region proposing limits the further improvement of FGVC.

Learning discriminative region-based features is another crucial task for FGVC. Existing methods (Fu, Zheng, and Mei 2017)(Zheng et al. 2017)(Yang et al. 2018) train the region-based feature extractors individually, which neglects the knowledge from the entire object and may see the forest for the trees. Although some search further considering the object-region relation by concatenating features vectors after extracted from object and part, their region-based feature extractors are still trained by a one-hot classification task. Consequently, the learning on region-based feature neglects the supervision from entire object. The inefficient feature learning method brings another limitation for FGVC.

To address the above conundrums, in this paper, we propose a novel “Filtration and Distillation Learning” (FDL) model (Figure 1) to enhance the region attention with proposing-predicting discrimination matching and object-part knowledge transforming. Firstly, a Filtration Learning (FL) method is put forward for discriminate part regions proposing based on the matchability between proposing and predicting. Specifically, we employ a Region Proposal Network (RPN) to produce a list of rectangle regions each with a confidence score, which indicates the discrimination of the proposed region. The proposing-predicting matchability refers to that, if a discriminative region brings a clear categorization result with higher probability being ground-truth, it should match a high confidence score produced by RPN. Consequently, the confidence scores S and probabilities being ground-truth P of regions should keep consistent ranking (Chen et al. 2009) in pair-wise order and point-wise value. We adopt the matching degree of S and P as the performance metric of region proposing, and directly optimize the RPN according to their ranking loss. Therefore the RPN can correctly and effectively filtrate the most discriminative regions with our Filtration Learning.

Secondly, we proposed a Distillation Learning (DL) method to fuse the knowledge from the entire object into region-based feature learning by knowledge distilling (Hinton, Vinyals, and Dean 2015). Intuitively, the region-based feature learning may see the forest for the trees, and capture fine-grained but biased representation. The object-based feature learning, in contrast, can capture rounded representation over the entire image, and produce credible label distribution knowledge. In order to eliminate the “prejudice” of region-base feature extractor, we further formulate the object-based feature learning and region-based feature learning as “teacher” and “student”. By transferring the learned knowledge from object to part regions, the region-base feature learning can get better supervision with object-region constraint in our Distillation Learning.

Our contributions can be summarized as follows:

- We put forward a Filtration Learning (FL) method for discriminative part regions proposing. Based on the discrimination matching between proposing and predicting, the region proposing can get directly optimized without the need of bounding box/part annotations by FL.
- We propose a Distillation Learning (DL) method to enhance the region-based feature learning. By knowledge distillation from entire object, the region-base feature learning can get better supervision with object-region constraint.
- Our FDL is highly flexible and can be easily implemented with various convolutional neural network, e.g. VGG, ResNet, DenseNet. Extensive experiments verify that FDL yields state-of-the-art performance under the same settings with the most competitive approaches on multiple FGVC tasks.
- Our FDL is highly interpretable to human perception and exhibits competitive performance in Weakly Supervised Object Localization (WSOL).

The remaining of this paper is organized as follows. Sec. 2 reviews the related work. Sec. 3 presents the generation of our FDL model. Sec. 4 introduces our experimental results and analysis on public benchmark datasets, followed by the conclusions in Sec. 5.

Related Work

In this section, we will introduce the most related work to our approach, including region proposing, information ranking, and knowledge distilling.

2.1 Discriminative Region Proposing

Since discriminative local region details play an important role for FGVC, learning to propose the discriminative regions possesses high importance in recent researches(Xie et al. 2019)(Liu et al. 2019a). A series of methods have been proposed by utilizing filter response to detect the corresponding discriminative region. Fu et al. (Fu, Zheng, and Mei 2017) propose a multi-step attention network to obtain discriminating regions by searching regions with the highest response value in the last convolutional layer. Zheng et al. (Zheng et al. 2017) take one-step attention network to generate multiple attention regions by designing a channel grouping module. Zhang et al. (Zhang et al. 2017) propose a picking strategy to elaborately select distinctive and consistent patches based on the responses of CNN filter banks. Zheng et al. (Zheng et al. 2019) adaptively sample the attention region to obtain the detail-preserved image according to the feature map response. Although promising results have been reported, the performance metric of region proposing is still a blank area in researches, which limits the further improvement of FGVC.

In this paper, we introduce RPN to propose discriminative region for FGVC, and creatively optimize RPN according to the proposing-predicting matchability.

2.2 Information Ranking

Information ranking is a task to automatically construct a ranking model to sort new objects according to their degrees of relevance, preference, or importance(Chen et al. 2009)(Xie et al. 2018). Zhang et al. (Zhang and Rusinkiewicz 2018) propose an appropriately ranking method to detect keypoints by their matchability. Yang et al.

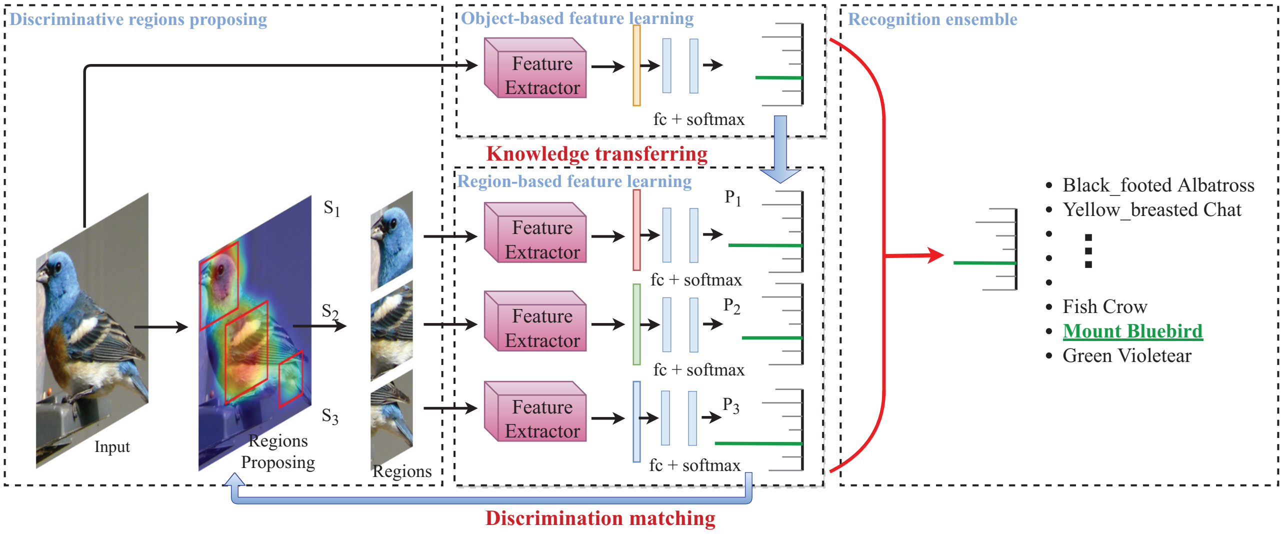

Figure 2: Illustrations of our FDL attention model. A Filtration Learning (FL) method is put forward based on the discrimination matching between proposing and predicting, and a Distillation Learning (DL) method is proposed to transfer the learned knowledge from object to part regions. [Best viewed in color with zoom-in]

(Yang et al. 2018) introduce an intrinsic consistency ranking agent to detect informative regions in an image.

In this paper, we propose a filtration learning method that utilizes the matchability between proposing and predicting to optimize region proposal. Different from the above methods which only consider the pair-wise order consistency, we further introduce the point-wise value consistency in the optimization. Extensive experiments verify the effectiveness of our filtration learning with significant improvement in accuracy.

2.3 Knowledge Distilling

Knowledge distilling is firstly proposed by Hinton et al. (Hinton, Vinyals, and Dean 2015) to transfer knowledge from an ensemble or from a large highly regularized model into a smaller, distilled model in a teacher-student manner. The main idea is using soft targets (i.e., the predicted distribution of ensemble/large model) to optimize the small model, for it contains more label distribution information than the one-hot label. In FGVC, the region-based feature learning may capture fine-grained but biased representation, while the object-based feature learning can capture rounded representation over the entire image and produce credible label distribution knowledge. The object-region feature learning can also be viewed as a circumstance of knowledge distilling.

In this paper, we formulate the object-based feature learning and region-based feature learning as “teacher” and “student”. By transferring the learned knowledge from object to part regions, the region-base feature learning can get better supervision with object-region constraint.

Approach

In this section, we will introduce the proposed Filtration and Distillation Learning (FDL) model for FGVC. Our FDL model is designed with discriminative regions proposing, object-based feature learning, region-base feature learning and recognition ensemble, as shown in Figure 2. A Filtration Learning (FL) method is put forward based on the matchability between proposing and predicting, which enables an effective and direct optimization for region proposing. Moreover, we propose a Distillation Learning (DL) to transfer the learned knowledge from object to part regions, which enhances the region-based feature learning with object-region constraint. To the end, we assemble all the recognition results and make a final prediction.

3.1 Discriminative Regions Proposing

Inspired by the success of Region Proposal Network (RPN) in object detection (Ren et al. 2015)(Fu et al. 2017)(Liu et al. 2019b)(Wang et al. 2019), we employ RPN for discriminative region proposing in our FDL model. Specifically, the RPN takes an image as input and produces a list of rectangle regions $\{R'_1, R'_2, ...R'_A\}$, each with a confidence score $S(R'_i)$ of the region. Here we resize the input object image O with the size of 448, and choose anchors with scales of{48, 96, 192} and ratios of {1:1, 3:2, 2:3}. The list is sorted in order of the score from high to low as in Eqn. 1, where A is the number of anchors, and $S(R_i)$ is the i-th element in sorted score list.

$$S(R_1) \geq S(R_2) \geq ... \geq S(R_A) \tag{1}$$

To reduce redundancy, we adopt non-maximum suppression (NMS) based on their confidence scores. After NMS, the FDL chooses top-M discriminative part regions

$\{R_1, R_2, ...R_M\}$ according to the score $S(R_i)$. Then the regions are cropped from the input image and resized to predefined size for further feature learning.

3.2 Feature Learning and Recognition Ensemble

After being resized to the predefined size, the top-M regions are fed into feature extractor to generate feature vectors $v(R_i)$, each with length L. Then the feature vectors are fed into a fully-connected layer, which has L neurons, and a softmax layer to generate the probability $\{P^j(R_i)\}$, here $P^j(R_i)$ denotes the probability of predicting region R_i as the j-th class. The input object image O is also fed into the classifier and we generate its feature vector as $v(O)$ and prediction result as $\{P^j(O)\}$.

To further leverage the benefit of part feature ensemble, we get the object-region concatenated feature vector $v(C)$, by concatenating the feature vectors of the input object $v(O)$ and the top-K regions $\{v\left(R_i\right)\}_{i=1}^{K}$.

$$v(C) = [v(O) : v(R_1) : \cdots : v(R_K)]. \tag{2}$$

The concatenated feature vector $v(C)$ is fed into a fully-connected layer, which has $L(K+1)$ neurons, and a softmax layer to generate the probability $\{P^j(C)\}$. Then we average $\{P^j(C)\}$, $\{P^j(O)\}$, $\{P^j(R_1)\}$, $\{P^j(R_2)\}$,... ,$\{P^j(R_K)\}$ and get the assembling probability $\{P^j_{ass}\}$ as Eqn. 3.

$$P^j_{ass} = \frac{1}{K+2}\{P^j(C) + P^j(O) + \sum_{i=1}^{K} P^j(R_i)\} \tag{3}$$

3.3 Filtration Learning with Discrimination Matching

To enable an end-to-end optimization for region proposing, we creatively propose a Filtration Learning method which utilize the proposing-predicting discrimination matchability as the performance metric of RPN. Specifically, the PRN produces a list of discriminative part regions $\{R_1, R_2, ...R_M\}$ and their confidence scores $\{S(R_1), S(R_2), ...S(R_M)\}$. The regions will be fed to feature learning and recognition. For each region R_i, the probability being ground-truth class $P(R_i)$ is fed back to RPN. The $P(R_i)$ can be easily extracted from $\{P^j(R_i)\}$ as Eqn. 4 shows.

$$P(R_i) = P^{gt}(R_i) \quad (gt = index\,of\,ground{-}truth) \tag{4}$$

The proposing-predicting matchability refers to that, if a discriminative region is proposed with higher confidence score $S(R_i)$, it should match clear categorization result with higher probability being ground-truth $P(R_i)$. Consequently, the confidence scores and probabilities being ground-truth of regions should keep consistent ranking (Chen et al. 2009) in pair-wise order and point-wise value.

The pair-wise order consistency requires the confidence score $S(R_i)$ and probability being ground-truth $P(R_i)$ should be sorted with a same order, as in Eqn. 5

$$\begin{cases} S\left(R_i\right) > S\left(R_j\right) \text{ and } P\left(R_i\right) > P\left(R_j\right) \\ S\left(R_i\right) < S\left(R_j\right) \text{ and } P\left(R_i\right) < P\left(R_j\right) \end{cases} \tag{5}$$

Correspondingly the pair-wise order loss function $\mathcal{L}_{pair}$ is defined as Eqn. 6

$$\mathcal{L}_{pair}(S, P) = \sum_{(i,j):P(R_i)<P(R_j)} f_{pair}(S(R_j) - S(R_i)), \tag{6}$$

where the function f is hinge loss function $f_{pair}(x) = max\{1 - x, 0\}$ in our experiment.

The point-wise value consistency requires the confidence score $S(R_i)$ and probability being ground-truth $P(R_i)$ should possess approximate value, as in Eqn. 7

$$\arg\min_S \sum_{i=1}^{M} \|S(R_i) - P(R_i)\| \tag{7}$$

Correspondingly the point-wise order loss function $\mathcal{L}_{point}$ is defined as Eqn. 8

$$\mathcal{L}_{point}(S, P) = \sum_{i=1}^{M} f_{point}(S(R_i) - P(R_i)), \tag{8}$$

where the function f is L1 loss function $f_{point}(x) = |x|$ in our experiment.

To sum up, the loss function for Filtration Learning is illustrated as Eqn. 9:

$$\mathcal{L}_{FL} = \mathcal{L}_{pair}(S, P) + \lambda\mathcal{L}_{point}(S, P). \tag{9}$$

During training, we directly optimize RPN to make $S(R_i)$ and $P(R_i)$ having the same ranking by Filtration Learning. An accurate RPN can propose and filtrate high discriminative regions, which can benefit the region attention learning for FGVC .

3.4 Distillation Learning with Knowledge Transferring

In order to obtain label distribution and intra-class relation knowledge, we propose a Distillation Learning method to fuse the knowledge from the entire object into region-based feature learning by knowledge distilling. Specifically, we formulate the object-based feature learning and region-based feature learning as “teacher” and “student”, and transfer the learned knowledge from object to regions-based feature learning. It is worth mentioning that our motivation of knowledge distilling is totally different from TASN(Zheng et al. 2019). While TASN transfers fine-grained knowledge into object-based feature learning, our FDL aims to fuse the knowledge from the entire object into region-based feature learning.

We convert the logit output from each local regions $z(R)$ into a soft probability distribution $q_s(R)$ over classes with Eqn.10:

$$q_s^{(i)}(R) = \frac{exp(z^{(i)}(R)/T)}{\sum_j exp(z^{(j)}(R)/T)}, \tag{10}$$

where T is a parameter namely temperature to produce a soft probability distribution over classes. We obtain the soft target cross entropy for the Distillation Learning as Eqn. 11:

$$\mathcal{L}_{DL}(q(O), q(R)) = -\sum_{j=1}^{n} q^{(j)}(O) log q_s^{(j)}(R), \tag{11}$$

where n denotes the class number.

3.5 Filtration and Distillation Learning

In our framework, the classification, filtration and distillation are trained in an end-to-end and joint manner. Specifically, we minimize the following objective:

$$\mathcal{L} = \alpha \mathcal{L}_{CLS} + \beta \mathcal{L}_{FL} + \gamma \mathcal{L}_{DL}. \tag{12}$$

Where the $\mathcal{L}_{cls}$ is the sum of classification loss based on input object, local regions and their concatenated feature:

$$\begin{aligned} \mathcal{L}_{CLS} = f_{cls}(P^j(O), gt) + \sum_{i=1}^{M} f_{cls}(P^j(R_i), gt) \\ + f_{cls}(P^j(C), gt), \end{aligned} \tag{13}$$

where the f_{cls} is the cross entropy loss function for classification. During the jointly training of optimization, our FDL can effectively improve the region proposing and region-based feature leaning for FGVC.

Experiments

We present performance evaluations and analysis of our proposed FDL attention model on four challenging tasks, including CUB-200-2011(Wah et al. 2011), FGVC-Aircraft(Maji et al. 2013), Stanford Cars(Krause et al. 2013) and Stanford Dogs(Khosla et al. 2011).

4.1 Implementation Details

Our FDL is highly flexible and can be easily implemented with various convolutional neural network, and we validate its performance on different backbones (VGG19, ResNet50, DenseNet161) with sufficient experiments. To make fair comparison, input images and part regions are resized to 448×448 and 224×224 respectively. The NMS threshold in region proposing is set to 0.25. Momentum SGD is chosen as the optimizer with initial learning rate 0.001 and weight decay 0.0001. And the learning rate is multiplied by 0.1 after every 80 epochs.

The hyper-parameters α, β, γ is simply set as 1, the hyper-parameter λ for list-wise ranking is set as 0.1. and the temperature T in knowledge distilling is set as 10. We fix M = 6 in region proposing, which means 6 regions are extracted to optimize the RPN in Filtration Learning. Comparative experiments are carried out on the number of attention regions for classification, where K ranges from 1 to 5. The code and pre-trained model are released for public research.[1]

4.2 Performance Comparison

The experiment results of our FDL on CUB-200-2011, Stanford Cars, FGVC-Aircraft and Stanford Dog are presented in Table .1 We compare FDL with the most competitive approaches in our experiment, including RACNN (Fu, Zheng, and Mei 2017), MACNN (Zheng et al. 2017), PC (Dubey et al. 2018a), MaxEnt (Dubey et al. 2018b), NTS (Yang et al. 2018), MAMC (Sun et al. 2018), TSAN (Zheng et al. 2019) and DCL (Chen et al. 2019). For fair comparison, we conduct extensive experiment with different backbone.

As we can capture in the table, our FDL exhibits competitive performance with a series of state-of-the-art accuracy on different FGVC tasks. Moreover, the proposed FDL outperforms all previous methods on CUB-200-2011 tasks. Compared with RA-CNN which zooms one attention region in different scales, FDL proposes four discriminative regions for classification. A clear improvement of 1.6% is obtained by FDL over RA-CNN on CUB-200-2011 tasks. The comparative experiments indicate that multiple attention regions should be captured in FGVC. Compared with MA-CNN which proposes multiple attention regions according to filter response, the region proposing in FDL can be directly optimized with the matchability between proposing and predicting. Correspondingly, FDL achieves a relative improvement of 0.3% on CUB-200-2011 tasks, which indicate the superiority of region proposing with Filtration Learning.

Compared with DCL which enhance the region-based feature learning by destructing the object structure, our FDL can learn not only learn fine-grained region feature but also the object-structure feature, and achieves an accuracy gain of 0.6% with the same backbone. While the TASN design an attention-based sampler for fine-grained feature learning, FDL directly filtrates and extracts the attention region for feature learning, which brings an accuracy gain of 0.7%.

4.3 Ablation Studies

Number of Attention Regions: Comparative experiments are first carried out on the number of attention regions, where K ranges from 1 to 5. Since there exists multiple attention region in FGVC, determining how many discriminative regions are necessary to achieve the best performance. As shown in Figure 3, FDL achieves best accuracy with $K = 4$ on CUB-200-2011, FGVC-Aircraft and Stanford Dog task, meanwhile it achieves best accuracy on Stanford Cars with $K = 3$. Different from other fine-grained categories like bird and aircraft, the car usually possess simpler structure and clearer marker, therefore the number of attention regions is smaller than other tasks.

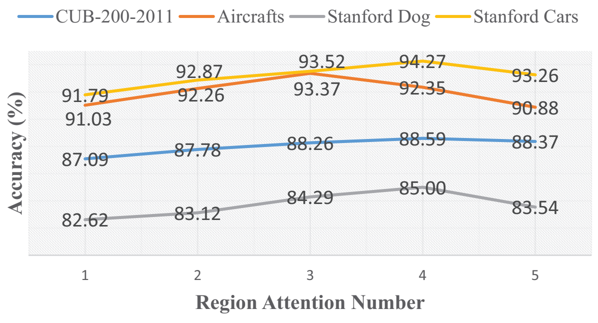

Figure 3: Ablation studies on the number of attention regions.

Filtration and Distillation Learning: We then conduct ablation studies to understand different components in our

[1] https://github.com/liuboss1992/FDL

Method	Backbone	Accuracy (%)			
		CUB-200-2011	Stanford Cars	FGVC-Aircrafts	Stanford Dog
RACNN	VGG19	85.2	92.5	-	-
MACNN	VGG19	86.5	**92.8**	**89.9**	-
TASN	VGG19	86.1	92.4	-	-
FDL	VGG19	**86.84**	91.52	88.51	-
PC	ResNet50	80.21	93.43	83.40	73.35
MaxEnt	ResNet50	80.37	93.85	83.86	73.56
NTS	ResNet50	87.5	93.9	91.4	-
MAMC	ResNet50	86.2	93.0	-	84.8
TASN	ResNet50	87.9	93.8	-	-
DCL	ResNet50	87.8	**94.5**	93.0	-
FDL	ResNet50	**88.59**	94.27	**93.37**	**85.00**
PC	DenseNet161	86.87	92.86	89.24	83.75
MaxEnt	DenseNet161	86.54	93.01	89.76	83.63
FDL	DenseNet161	**89.09**	**94.02**	**91.27**	**84.86**

Table 1: Comparison results on four different standard FGVC dataset.

proposed FDL. Using ResNet50 as the backbone network on CUB-200-2011 task, the proposed FDL boosts the performance significantly as shown in Table 2. The performance improvement 4.47% obtained by FL proves the effectiveness of FL in optimizing RPN, moreover, it indicates the critical significance of region proposing in FGVC. At the same time, the DL bring a clear improvement of 1.29% in accuracy, which indicates that the knowledge learned in entire object is beneficial to the region-based feature learning. Combining Filtration Learning and Distillation Learning, our FDL model achieves state-of-the-art performance of 88.59% in accuracy.

Method	$\mathcal{L}_{cls}$	$\mathcal{L}_{FL}$	$\mathcal{L}_{DL}$	Accuracy (%)
ResNet50	✓			83.54
+ FL	✓	✓		88.07
+ DL	✓		✓	84.83
+ FDL	✓	✓	✓	**88.59**

Table 2: Ablation study on different components in our proposed FDL.

Pair-wise and Point-wise Ranking: Moreover, we conduct ablation studies on the performance of Pair-wise order loss and Point-wise value loss in Filtration learning. As we can see in Table 3, both the $\mathcal{L}_{pair}$ and $\mathcal{L}_{point}$ can improve the model accuracy significantly. And we achieve the best accuracy of 88.59% when combing $\mathcal{L}_{pair}$ and $\mathcal{L}_{point}$ in Filtration learning.

$\mathcal{L}_{pair}$	$\mathcal{L}_{point}$	Accuracy (%)
		84.83
✓		88.14
	✓	88.01
✓	✓	**88.59**

Table 3: Ablation study on pair-wise and point-wise ranking in Filtration learning.

4.4 Visualization Experiments

Feature Representation Learning: To investigate the feature representation learning ability of FDL, we conduct extensive comparison experiments on different FGVC task with original backbone. For fair comparison, we only extract the recognition result based on input object $P^j(O)$, instead of recognition ensemble. As shown in Table 4, our FDL can bring significant improvement over the ResNet50 baseline with a small overhead during training. This indicates that, owing to the sharing of feature extractor in object-based feature learning and part-based feature learning, FDL can also enhance the object-based recognition with fine-grained representation learning. Please note that, no computational overhead is introduced at testing time in the object-based recognition.

Task	ResNet50	Parameter	Acc. (%)
CUB-200-2011	Baseline	23.92M	85.49
	DFL	28.51M	**88.35**
FGVC-Aircrafts	Baseline	23.71M	89.86
	DFL	26.87M	**93.04**
Stanford Dog	Baseline	23.75M	83.07
	DFL	27.36M	**84.45**
Stanford Cars	Baseline	23.91M	91.74
	DFL	28.45M	**93.71**

Table 4: Comparisons between FDL model and baseline model in training parameter and model performance.

As shown in Figure 4, we further conduct visualization experiments to investigate the attention map by Grad-CAM (Selvaraju et al. 2017). The original classification model tends to identify only the most discriminative part of the target object, incapable of focusing multiple attention regions for FGVC. As a contrast, FDL generates multiple peak responses on different discriminative parts. The visualization experiments verify that FDL enables a better feature representation learning over serval discriminative regions than

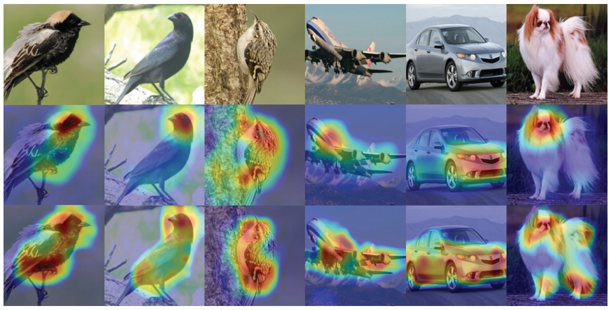

Figure 4: Visualization Experiments of attention map. The first row shows input image, the second row shows attention map of baseline model, and the third row shows the attention map of our FDL model. The red color stands for high attention. FDL enables a better feature representation learning over serval discriminative regions than original backbone. [Best viewed in color with zoom-in]

original backbone, which is essential for FGVC.

Weakly Supervised Object Localization: Due to the balanced attention on different part of objects, the attention map of our FDL can better cover the entire mask and identify the localization of object than original backbone network. Figure 5 illustrates the visualization of WSOL result of FDL. Simply set the attention threshold as 0.125, we determine the bounding box as the minimum enclosing rectangle of the mask. A competitive localization accuracy of 80.2% is achieved by FDL in *GT-known Loc*, which judges the answer as correct when the intersection over union (IoU) between the ground truth and estimated bounding box for the ground truth class is 50% or more. As we can see in the figure, the localization result is pretty close to the ground-truth. The experiments demonstrate that our FDL is high extensible to other computer vision tasks.

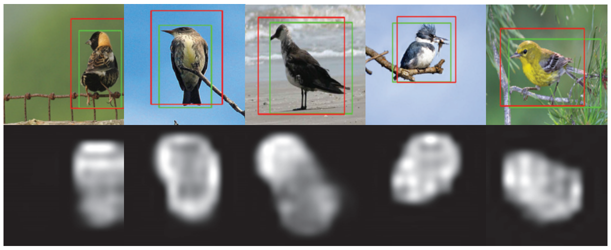

Figure 5: Visualization of WSOL result of FDL. The first row shows the ground-truth bounding box (in green) and our localization result (in red). The second row shows the attention map of FDL. [Best viewed in color with zoom-in]

Attention Region Proposing: As shown in Figure 6, we use rectangles in different color to denote the attention regions proposed by RPN. The proposed regions provide clear and significant visual cues for classification. In CUB-200-2011, the head, wings and main body of a bird are captured by FDL attention model, which is consistent with the human perception. Meanwhile, the tail and talons are rarely proposed. The reason is that these regions usually hold less information than the head, wings of a bird. In FGVC-Aircraft, the attention mostly focuses on the head, wings, main body and tail of the airplane. In Stanford Dogs, our FDL mostly focuses on the eyes, nose, face and ear of a dog. In Stanford Cars, consistently with human perception, the regions proposed from cars are car lights, front view, side view.

Overall, our FDL shares consistent attention with human and exhibits a strong interpretability.

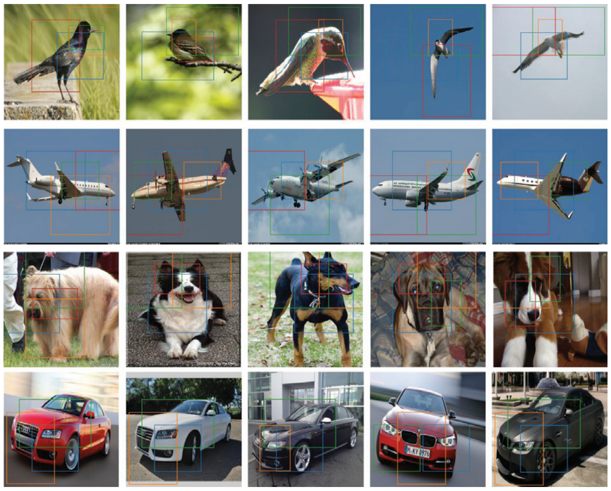

Figure 6: Region proposing results for individual examples. FDL shares consistent attention with human and exhibits a strong interpretability. The first row shows the results in CUB-200-2011, the second row shows the results in FGVC-Aircraft, the third row shows the results in Stanford Dog and the fourth row shows the results in Stanford Cars.[Best viewed in color with zoom-in]

Conclusion

In this paper, we present a novel Filtration and Distillation Learning (FDL) attention model to enhance region attention for fine-grained visual categorization. We creativity enable a direct optimization for region proposing with the matchability between proposing and predicting, and transfer the object knowledge to the region-based feature learning. Extensive experiments verify the superior performances of our FDL on various FGVC tasks. Moreover, the FDL exhibits strong interpretability and competitive performance in Weakly Supervised Object Localization.

The success of FDL reveals the pivotal role of region attention in FGVC, which can be an enlightening reference for future research.

Acknowledgements

This work is supported by the National Key Research and Development Program of China (2017YFC0820600), the National Nature Science Foundation of China (61525206, 61771468), the Youth Innovation Promotion Association Chinese Academy of Sciences (2017209).

References

Chen, W.; Liu, T.-Y.; Lan, Y.; Ma, Z.-M.; and Li, H. 2009. Ranking measures and loss functions in learning to rank. In *Advances in Neural Information Processing Systems*, 315–323.

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.

Dubey, A.; Gupta, O.; Guo, P.; Raskar, R.; Farrell, R.; and Naik, N. 2018a. Pairwise confusion for fine-grained visual classification. In *ECCV*, 70–86.

Dubey, A.; Gupta, O.; Raskar, R.; and Naik, N. 2018b. Maximum-entropy fine grained classification. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada.*, 635–645.

Fu, C.-Y.; Liu, W.; Ranga, A.; Tyagi, A.; and Berg, A. C. 2017. Dssd: Deconvolutional single shot detector. *arXiv preprint arXiv:1701.06659*.

Fu, J.; Zheng, H.; and Mei, T. 2017. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In *CVPR*, volume 2, 3.

He, X., and Peng, Y. 2017. Weakly supervised learning of part selection model with spatial constraints for fine-grained image classification. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA.*, 4075–4081.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Khosla, A.; Jayadevaprakash, N.; Yao, B.; and Fei-Fei, L. 2011. Novel dataset for fine-grained image categorization. In *CVPR*.

Krause, J.; Stark, M.; Deng, J.; and Fei-Fei, L. 2013. 3d object representations for fine-grained categorization. In *ICCV*, 554–561.

Krause, J.; Jin, H.; Yang, J.; and Fei-Fei, L. 2015. Fine-grained recognition without part annotations. In *CVPR*, 5546–5555.

Lin, D.; Shen, X.; Lu, C.; and Jia, J. 2015. Deep lac: Deep localization, alignment and classification for fine-grained recognition. In *CVPR*, 1666–1674.

Liu, X.; Wang, J.; Wen, S.; Ding, E.; and Lin, Y. 2017. Localizing by describing: Attribute-guided attention localization for fine-grained recognition. In *AAAI*, 4190–4196.

Liu, C.; Xie, H.; Liu, Y.; Zha, Z.; Lin, F.; and Zhang, Y. 2019a. Extract bone parts without human prior: End-to-end convolutional neural network for pediatric bone age assessment. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 667–675. Springer.

Liu, C.; Xie, H.; Zhang, S.; Xu, J.; Sun, J.; and Zhang, Y. 2019b. Misshapen pelvis landmark detection by spatial local correlation mining for diagnosing developmental dysplasia of the hip. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 441–449. Springer.

Maji, S.; Rahtu, E.; Kannala, J.; Blaschko, M.; and Vedaldi, A. 2013. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, 91–99.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, 618–626.

Sun, M.; Yuan, Y.; Zhou, F.; and Ding, E. 2018. Multi-attention multi-class constraint for fine-grained image recognition. In *ECCV*, 834–850.

Wah, C.; Branson, S.; Welinder, P.; Perona, P.; and Belongie, S. 2011. The caltech-ucsd birds-200-2011 dataset.

Wang, Y.; Choi, J.; Morariu, V.; and Davis, L. S. 2016. Mining discriminative triplets of patches for fine-grained classification. In *CVPR*, 1163–1172.

Wang, Y.; Xie, H.; Fu, Z.; and Zhang, Y. 2019. Dsrn: a deep scale relationship network for scene text detection. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, 947–953. AAAI Press.

Xie, H.; Mao, Z.; Zhang, Y.; Deng, H.; Yan, C.; and Chen, Z. 2018. Double-bit quantization and index hashing for nearest neighbor search. *IEEE Transactions on Multimedia* 21(5):1248–1260.

Xie, H.; Fang, S.; Zha, Z.-J.; Yang, Y.; Li, Y.; and Zhang, Y. 2019. Convolutional attention networks for scene text recognition. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 15(1s):3.

Yang, Z.; Luo, T.; Wang, D.; Hu, Z.; Gao, J.; and Wang, L. 2018. Learning to navigate for fine-grained classification. In *ECCV*, 438–454.

Zhang, L., and Rusinkiewicz, S. 2018. Learning to detect features in texture images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6325–6333.

Zhang, N.; Donahue, J.; Girshick, R.; and Darrell, T. 2014. Part-based r-cnns for fine-grained category detection. In *ECCV*, 834–849. Springer.

Zhang, X.; Xiong, H.; Zhou, W.; Lin, W.; and Tian, Q. 2016. Picking Deep Filter Responses for Fine-Grained Image Recognition. *CVPR* 1134–1142.

Zhang, X.; Xiong, H.; Zhou, W.; Lin, W.; and Tian, Q. 2017. Picking neural activations for fine-grained recognition. *IEEE Transactions on Multimedia* 19(12):2736–2750.

Zheng, H.; Fu, J.; Mei, T.; and Luo, J. 2017. Learning multi-attention convolutional neural network for fine-grained image recognition. In *ICCV*, volume 6.

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.