Graph-Propagation Based Correlation Learning for Weakly Supervised Fine-Grained Image Classification

Zhihui Wang,[1,3] **Shijie Wang,**[2] **Haojie Li,**[1,3*] **Zhi Dou,**[2] **Jianjun Li**[4]
[1]International School of Information Science & Engineering, Dalian University of Technology, China
[2]School of Software Technology, Dalian University of Technology, China
[3]Key Laboratory for Ubiquitous Network and Service Software of Liaoning Province, China
[4]School of Computer Science and Technology, Hangzhou Dianzi University, China

Abstract

The key of Weakly Supervised Fine-grained Image Classification (WFGIC) is how to pick out the discriminative regions and learn the discriminative features from them. However, most recent WFGIC methods pick out the discriminative regions independently and utilize their features directly, while neglecting the facts that regions' features are mutually semantic correlated and region groups can be more discriminative. To address these issues, we propose an end-to-end Graph-propagation based Correlation Learning (GCL) model to fully mine and exploit the discriminative potentials of region correlations for WFGIC. Specifically, in discriminative region localization phase, a Criss-cross Graph Propagation (CGP) sub-network is proposed to learn region correlations, which establishes correlation between regions and then enhances each region by weighted aggregating other regions in a criss-cross way. By this means each region's representation encodes the global image-level context and local spatial context simultaneously, thus the network is guided to implicitly discover the more powerful discriminative region groups for WFGIC. In discriminative feature representation phase, the Correlation Feature Strengthening (CFS) sub-network is proposed to explore the internal semantic correlation among discriminative patches' feature vectors, to improve their discriminative power by iteratively enhancing informative elements while suppressing the useless ones. Extensive experiments demonstrate the effectiveness of proposed CGP and CFS sub-networks, and show that the GCL model achieves better performance both in accuracy and efficiency.

Introduction

As an emerging research topic, Weakly Supervised Fine-grained Image Classification (WFGIC) focuses on discriminative subtle variances for distinguishing objects of subordinate categories with only image-level labels. Due to small variances between images in same subcategory and sharing the global geometry and appearances of sub-categories, distinguishing fine-grained images is still a challenging task.

Learning to localize discriminative parts from fine-grain images plays the key role in WFGIC. More recent works can be divided into two groups. The first group is localizing discriminative parts based on heuristic schemes. (He and Peng; He, Peng, and Zhao; Peng, He, and Zhao; Zhang et al.; He and Peng). The limitation of heuristic schemes is that they hardly guarantee the selected regions are discriminative enough. The second group is end-to-end localization-classification approaches by learning mechanism (Fu, Zheng, and Mei; Wang, Morariu, and Davis; Zheng et al.; Yang et al.). However, all the previous works try to localize discriminative regions/patches independently and neglect local spatial context of regions and the correlation between regions.

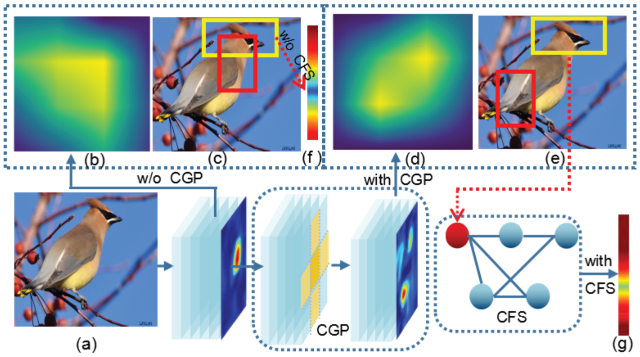


Figure 1: The motivation of graph propagation based correlation learning. (a) is the original image. (b)(d) are the discriminative score maps and (c) (e) are localization results without and with criss-cross graph propagation (CGP) learning, respectively. (f) and (g) are the feature vectors without and with correlation feature strengthening (CFS) learning.

We argue that considering local spatial context and region correlations is rather helpful in distinguishing fine-grained images. Exploiting local spatial context can improve the discriminative ability of regions and mining the correlation between regions can be more discriminative than individual region. This motivates us to incorporate the local spatial context of regions and correlation between regions into the discriminative patches selecting. To this end, we propose a Criss-cross Graph Propagation (CGP) sub-network to learn

[*]Corresponding author: hjli@dlut.edu.cn

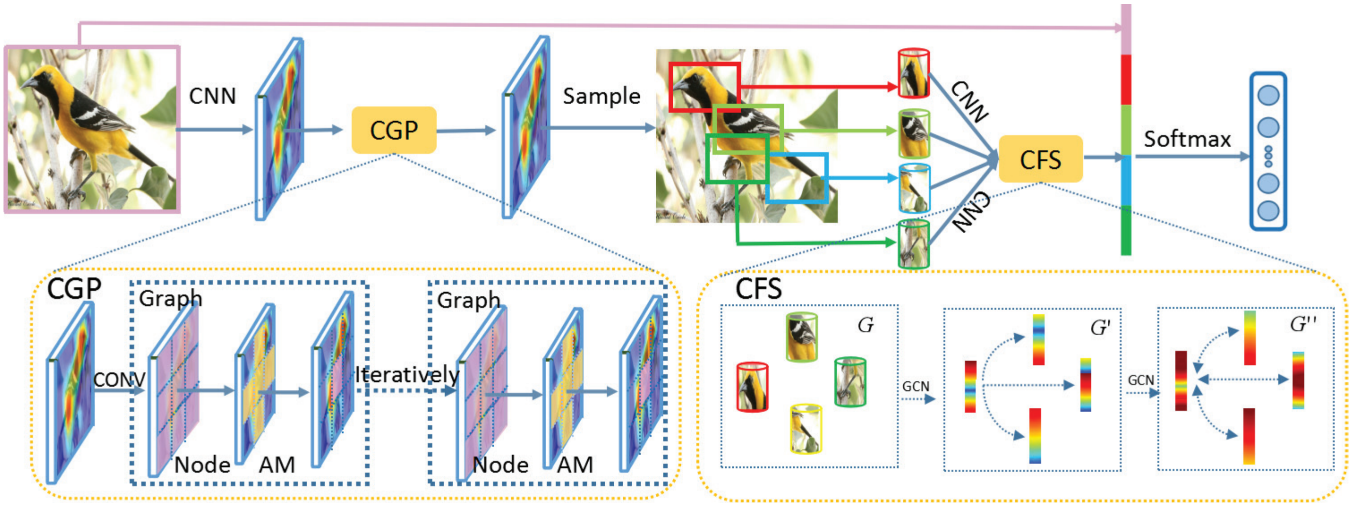


Figure 2: The framework of the Graph-propagation based Correlation Learning (GCL) model. We generate the discriminative adjacent matrix (AM) through the Criss-Cross Graph Propagation (CGP) sub-network and the discriminative score map (Score Map) through the scoring network (Sample). Then the GCL selects the more discriminative patches from the default patches (DP) according to the discriminative score map. Meanwhile, we crop and resize the patches to 224 × 224 from the original image, and generate the discriminative features through the graph-propagation for Correlation Feature Strengthening (CFS) sub-network. Finally, we concatenate the multiple features to get the final feature representation for WFGIC.

the correlations among regions. Specifically, CGP iteratively calculates correlations between regions in a criss-cross way and then enhances each region by correlation weighted aggregating other regions. By this means, each region's feature both encodes the global image-level context, i. e. all correlations between the aggregating region and other regions in the whole image, and local spatial context, i. e. the closer the region is to the aggregating region, the higher the aggregation frequency is during the criss-cross graph propagation. Through learning the correlation between regions in CGP, the network is guided to implicitly discover the discriminative region groups which are more powerful for WFGIC. Fig. 1 shows our motivations, where we can see that the score map (Fig. 1 (b)) highlights the the head regions when considering each region independently, while the score map (Fig. 1 (d)) strengthens the most discriminative regions after multiple iterations of criss-cross propagation, which helps the accurate locating of the discriminative region group (the head and tail regions).

Discriminative feature representation play another key role for WFGIC. Recently, some end-to-end networks (Lin, Roy Chowdhury, and Maji; Gao et al.; Kong and Fowlkes Cai, Zuo, and Zhang; Cui et al.) strengthen the discriminative ability of feature representation by encoding convolutional feature vectors into high-order information. The effectiveness of these methods is due to their invariance ability to object translation and pose variation, which benefits from the order-less aggregation manner of features. The limitation of these feature encoding methods is that they neglect the importance of local discriminative features for WFGIC. Therefore, some methods (Zheng et al.; He, Peng, and Zhao) incorporate the local discriminative features to improve the feature discriminative ability through concatenating selected region feature vectors. However, it is worth noting that all the previous works ignore the internal semantic correlation among discriminative region feature vectors. Besides, there are some noisy context, such as background regions in Fig. 1(c)(e) within the selected discriminative regions. Such background or less discriminative information is likely to be harmful for WFGIC since all subcategories share similar background (e.g. birds usually inhabit on the tree or fly in the sky). Based on above intuitive yet important observations and analysis, we propose a Correlation Feature Strengthening (CFS) sub-network to explore the internal semantic correlation between region feature vectors to obtain better discriminative ability. We achieve this by constructing a graph with the selected region feature vectors, and then jointly learning the interdependencies among feature vector nodes to guide the discriminative information propagation in CFS. Fig. 1 (f) and (g) are the feature vectors without and with CFS learning.

To summarize, the contributions are as follows:

- To the best of our knowledge, we are the first to explore and exploit region correlations based on graph-propagation to implicitly discover the discriminative region groups and improve their feature discriminative ability for WFGIC.
- We propose an end-to-end Graph-propagation based Correlation Learning (GCL) model which incorporates the criss-cross graph propagation (CGP) sub-network and correlation feature strengthening (CFS) sub-network into a unified framework to learn discriminative features efficiently and jointly.
- We evaluate the proposed model on Caltech-UCSD Birds-200-2011 (CUB-200-2011) (Branson et al.) and Stanford

Cars (Krause et al.). Extensive experiments indicate our method achieves the best performance both in classification accuracy (e.g., 88.3% vs 87.0% (Chen et al.) on CUB-200-2011) and efficiency (e.g., 56 FPS vs 30 FPS (Lin, Roy Chowdhury, and Maji) on CUB-200-2011).

Proposed Method

Overview

We propose an end-to-end Graph-propagation based Correlation Learning (GCL) model to fully explore and exploit the discriminative potential ability of correlation for WFGIC. It consists of two graph propagation sub-networks, as shown in Fig. 2. The Criss-cross Graph Propagation (CGP) sub-network enhances the features of each position by a weighted sum of the features at its horizontal and vertical (i.e. criss-cross) positions to discover discriminative regions. The Correlation Feature Strengthening (CFS) sub-network explores the interdependencies among feature vectors to emphasis discriminative elements and suppress less helpful ones.

Criss-Cross Graph Propagation

In order to take both global image-level context and local spatial context into account during the discriminative ability learning of regions, we propose a new approach to establish the correlative relationship between regions. One graph propagation process of the proposed CGP module includes the following two stages: the first stage is that CGP learns correlation weight coefficients between each two regions (i.e. adjacent matrix computing). In the second stage, the model combines the information of its criss-cross neighbor regions through a weighted sum operation for seeking the real discriminative regions (i.e. graph updating). Specifically, the global image-level context is integrated into CGP via calculating correlations between each two regions in the whole image, and the local spatial context information is encoded through the iterative criss-cross aggregating operations, as shown in Fig. 3.

Here, we give a detailed formulation for one graph propagation process. Given an input feature map $M_O \in \mathbb{R}^{C\times H\times W}$, where W, H represents the width and height of the feature map and C is the number of channels, we feed it into the CGP module $\mathcal{F}$:

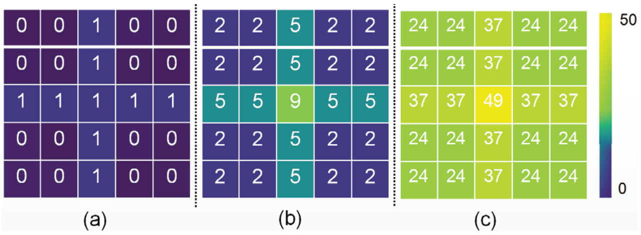
$$M_S = \mathcal{F}(M_O), \tag{1}$$

where $\mathcal{F}$ is composed of node representation, adjacent matrix computing and graph updating. $M_S \in \mathbb{R}^{C\times H\times W}$ is the output feature maps.

Node representation. The node representation generation is achieved by a simple convolutional operation f:

$$M_G = f(W_T \cdot M_O + b_T), \tag{2}$$

where $W_T \in \mathbb{R}^{C\times 1\times 1\times C}$ and b_T are the learned weight parameters and bias vector of a convolution layer, respectively. $M_G \in \mathbb{R}^{C\times H\times W}$ denotes the node feature map. Specifically, we regard a 1×1 convolution filter as a small region detector. Each $V_T \in \mathbb{R}^{C\times 1\times 1}$ vector across channels at fixed spatial location of M_G represents a small region at a corresponding location of image. We use the generated small region as a node representation. Note that W_T is randomly initialized and the initial three node feature maps are obtained by three different f calculations: M_G^1, M_G^2, M_G^3.



0	0	1	0	0
0	0	1	0	0
1	1	1	1	1
0	0	1	0	0
0	0	1	0	0

(a)

2	2	5	2	2
2	2	5	2	2
5	5	9	5	5
2	2	5	2	2
2	2	5	2	2

(b)

24	24	37	24	24
24	24	37	24	24
37	37	49	37	37
24	24	37	24	24
24	24	37	24	24

(c)

Figure 3: The illustration of the frequency of each node in M_G^3 that are integrated into the center node via three times graph propagation.

Adjacent matrix computing. After obtaining the $W\times H$ nodes with C-dimension vectors in feature map M_G^1, M_G^2, we construct a correlation graph to calculate the semantic correlations between nodes. Each element in the adjacent matrix of correlation graph indicates the correlation intensity between nodes. Concretely, the adjacent matrix is obtained through performing node vector inner product between two feature maps $M_G^1 \in \mathbb{R}^{C\times H\times W}$ and $M_G^2 \in \mathbb{R}^{C\times H\times W}$.

Let's take a single correlation of two positions in adjacent matrix as an example. The correlation of two positions at p_1 in M_G^1 and p_2 in M_G^2 is defined as below:

$$c(p_1, p_2) = V_1^{p_1} \cdot V_2^{p_2}, \tag{3}$$

where $V_1^{p_1}$ and $V_2^{p_2}$ mean node representation vectors of p_1 and p_2 respectively. Note that p_1 and p_2 must meet a specific spatial constraint that p_2 can only be on the same row or column (i.e. criss-cross positions) of p_1. As a result, we obtain $W+H-1$ correlation values for each node in M_G^1. To be specific, we organize the relative displacements in channels and obtain an output correlation matrix $M_C \in \mathbb{R}^{K\times H\times W}$, where $K = W+H-1$. Then M_C is passed through a softmax layer to generate the adjacent matrix $R \in \mathbb{R}^{K\times H\times W}$:

$$R^{ijk} = \sigma(M_C^{ijk}) = \frac{e^{M_C^{ijk}}}{\sum_{k=1}^{K} e^{M_C^{ijk}}}, \tag{4}$$

where R^{ijk} is the correlation weight coefficient of the i^{th} row , the j^{th} column and the k^{th} channel.

In the process of forward pass, the more discriminative the regions are, the greater their correlations are. In the backward pass, we implement the derivatives with respect to each blob of node vectors. When the probability value of classification is low, the penalty will be back-propagated to lower the correlation weight of the two nodes, and the node vectors calculated through the node representation generation operation will be updated at the same time.

Graph updating. We feed $M_G^3 \in \mathbb{R}^{C\times H\times W}$ which is generated by the node representation generation phase, and the adjacent matrix R into the updating operation:

$$M_U^{ij} = \sum_{k=1}^{W+H-1} (V_3^{wh} \cdot R^{ijk}), \tag{5}$$

Table 1: The stride, patch scale size, scale step and aspect ratios of the three different layers. M_S^1 and M_S^2 are feature maps after down-sampling M_S from the output of CGP. Note that the stride is the original image scaling ratio. Patch width & height = scale × scale step × aspect ratio.

Feature Map	Stride	Scale	Scale Step	Aspect ratio
M_S	32	32	$2^{\frac{1}{3}}, 2^{\frac{2}{3}}$	$\frac{2}{3}, 1, \frac{3}{2}$
M_S^1	64	64	$2^{\frac{1}{3}}, 2^{\frac{2}{3}}$	$\frac{2}{3}, 1, \frac{3}{2}$
M_S^2	128	128	$1, 2^{\frac{1}{3}}, 2^{\frac{2}{3}}$	$\frac{2}{3}, 1, \frac{3}{2}$

wher e V_3^{wh} is the node in the w^{th} row and the h^{th} column of M_G^3, (w, h) is in set $[(i, 1), ..., (i, H), (1, j), ..., (W, j)]$. The node M_U^{ij} can be updated by combining nodes at its horizontal and vertical direction with corresponding correlation weight coefficient R^{ijk}, as shown in Fig. 3(a).

Similar as ResNet (He et al.), we adopt the residual learning:

$$M_S = \alpha \cdot M_U + M_O, \tag{6}$$

where α is a self-adaptive weight parameter and it gradually learns to assign more weight to the discriminative correlation features. It ranges from [0, 1], and is initialized approximating 0. In this way, M_S aggregates the correlation features and the original input features to pick out more discriminative patches. Then, we feed M_S as the new input into next iteration of CGP. As shown in Fig. 3, after multiple graph propagations, each node can aggregate all regions with different frequencies, which indirectly learns the global correlations, and the closer the region is to the aggregating region, the higher the aggregation frequency is during the graph propagation, which indicates the local spatial context information.

Discriminative Patch Samping

In this work, we generate default patches from feature maps of three different scales, inspired by Feature Pyramid Network in object detection (Liu et al. 2016). Tab. 1 shows the design of the default patches of the scale size, scale step and aspect ratio. The design can make network responsible to different size of discriminative regions.

After obtaining the residual feature map M_S, which aggregates the correlation features and the original input features, we feed it into a discriminative response layer. Concretely, we introduce a $1 \times 1 \times N$ convolution layer and a sigmoid function σ to learn discriminative probability maps $S \in \mathbb{R}^{N \times H \times W}$, which indicate the impact of discriminative regions on the final classification. N is the number of the default patches at a given location in the feature maps.

Afterwards, each default patch p_{ijk} will be assigned the discriminative probability value accordingly. The formulaic representation is as follows:

$$p_{ijk} = [t_x, t_y, t_w, t_h, s_{ijk}], \tag{7}$$

where (t_x, t_y, t_w, t_h) is the default coordinates of each patch and s_{ijk} denotes the discriminative probability value of the i^{th} row, the j^{th} column and the k^{th} channel. Finally, the network pick the top-M patches according to the probability value, where M is a hyper-parameter.

Correlation Feature Strengthening

Most current works ignore the internal semantic correlation among discriminative region feature vectors. Besides, there is some less discriminative or noisy context within the selected discriminative regions. We propose a CFS subnetwork to explore the internal semantic correlation between region feature vectors to obtain better discriminative ability. The details of CFS are as follows:

Node representation and Adjacent matrix computing. To construct the graph for mining correlation among selected patches, we extract M nodes with D-dimension feature vectors from M selected patches as the input of Graph Convolution Network (GCN) (Kipf and Welling). After detecting the M nodes, the adjacent matrix of correlation coefficient is computed which indicates correlation intensity between nodes. Therefore, each element of the adjacent matrix can be calculated as below:

$$R_{ij} = c_{ij} \cdot < n_i, n_j >, \tag{8}$$

where R_{ij} denotes the correlation coefficient between each two nodes (n_i, n_j), and c_{ij} is correlation weight coefficent in weighted matrix $C \in \mathbb{R}^{M \times M}$, and c_{ij} can be learned to adjust correlation coefficient R_{ij} through back propagation. Then we perform normalization on each row of adjacent matrix to ensure that the sum of all the edges connected to one node equals to 1. The normalization of the adjacent matrix $A \in \mathbb{R}^{M \times M}$ is realized by the softmax function shown as follows:

$$A_{ij} = \frac{expR(n_i, n_j)}{\sum_{j=1}^{N} expR(n_i, n_j)}. \tag{9}$$

As a result, the constructed correlation graph measures the relationship intensity between the selected patches.

Graph updating. After we obtain the adjacent matrix, we both take feature representations $N \in \mathbb{R}^{M \times D}$ with M nodes and the corresponding adjacent matrix $A \in \mathbb{R}^{M \times M}$ as inputs, and updates the node features as $N' \in \mathbb{R}^{M \times D'}$. Formally, one layer process of GCN can be represented as:

$$N' = f(N, A) = h(ANW), \tag{10}$$

where $W \in \mathbb{R}^{D \times D'}$ is the learned weight parameters, and h is a non-linear function (we use Rectified Linear Unit (ReLU) in the experiments). After multiple propagations, the discriminative information in selected patches can be wider interacted to obtain better discriminative ability.

Loss Function

We propose an end-to-end model which incorporates the CGP and CFS into a unified framework. The CGP and CFS are trained together under the supervision of multi-task loss $\mathcal{L}$, which consists of a basic fine-grained classification loss $\mathcal{L}_{cls}$, a guided loss $\mathcal{L}_{gud}$, a rank loss $\mathcal{L}_{rank}$ and a feature strengthening loss $\mathcal{L}_{fea}$. It can be shown as the following:

$$\mathcal{L} = \mathcal{L}_{cls} + \lambda_1 \cdot \mathcal{L}_{gud} + \lambda_2 \cdot \mathcal{L}_{rank} + \lambda_3 \cdot \mathcal{L}_{fea}, \tag{11}$$

where λ_1, λ_2, λ_3 are balance hyper-parameter among these losses. We set the parameters $\lambda_1 = \lambda_2 = \lambda_3 = 1$ after many experiment vertifications.

Table 2: The ablative recognition results of different variants of our method. We test the models on CUB-200-2011.

Method	Accuracy
BL (Li et al.)	84.5%
BL + DP + Score	86.2%
BL + DP + Score + CGP-SF	87.2%
BL + DP + Score + CGP	87.7%
BL + DP + Score + CGP+ CFS	88.3%

Let's use X to represent the original image and denote the selected discriminative patches with and without CFS module as $P = \{P_1, P_2, ..., P_N\}$ and $P^{'} = \left\{P_1^{'}, P_2^{'}, ..., P_N^{'}\right\}$ respectively. $\mathcal{C}$ is the confidence function which reflects the probability of classification into the correct category, and $S = \{S_1, S_2, ..., S_N\}$ means the discriminative probability scores. Then the guided loss, rank loss and feature strengthening loss are defined as follows:

$$\mathcal{L}_{gud}(X, P) = \sum_{i}^{N}(max\{0, log\mathcal{C}(X) - log\mathcal{C}(P_i)\}), \quad (12)$$

$$\mathcal{L}_{rank}(S, P) = \sum_{\mathcal{C}(P_i)<\mathcal{C}(P_j)}(max\{0, (S_i - S_j)\}), \quad (13)$$

$$\mathcal{L}_{fea}(P^{'}, P) = \sum_{i}^{N}(max\{0, log\mathcal{C}(P_i^{'}) - log\mathcal{C}(P_i)\}), \quad (14)$$

Here, the guided loss guides the network to pick out the most discriminative regions and the rank loss strives for consistency of the discriminative scores and the final classification probability values of selected patches. These two loss functions directly adjust the parameters of CGP and indirectly influence CFS. The feature strengthening loss can guarantee that the prediction probability of selected region features with CFS is greater than features without CFS, and the network can adjust the correlation weight maxtix C and the GCN weight parameters W to influence the information propagation between selected patches.

Experiments

Datasets

The empirical evaluation is performed on three widely used and competitive benchmark datasets for fine-grained image classification: Caltech-UCSD Bird-200-2011(CUB-200-2011) (Branson et al.), Stanford Cars (Cars) (Krause et al.) and FGVC Aircraft (Airs) (Maji et al.).The CUB-200-2011 dataset covers 200 species of birds and contains 11788 images that are divided into a training set of 5994 images and a test set of 5794 images. The Cars dataset has 16,185 images from 196 classes officially split into 8,144 training and 8,041 test images. The Airs dataset contains 10,000 images over 100 classes, and train and test sets split ratio are around 2 : 1.

Implementation Details

All the images are resized to 448×448 in our experiment settings. Fully convolutional network ResNet-50 and Batch Normalization are chosen as feature extractor and regularizer respectively. We set weight decay as $1e^{-4}$ and Momentum SGD with initial learning rate 0.001 and multiplied by 0.1 after 60 epochs. In addition, we adopt the non maximum suppression (The threshold is set to 0.25.) on patches based on their discriminative scores to reduce patch redundancy.

Ablation Experiments

As shown in Tab. 2, we conduct some ablation experiments to illustrate the effectiveness of proposed modules, including the Criss-cross Graph Propagation (CGP) and the Correlation Feature Strengthening (CFS).

Without any object or partial annotations for fine-grained classification, we extract features from the whole image through ResNet-50 and set it as the baseline(BL). Then we introduce default patches (DP) as local features to improve classification accuracy. When we recommend the score mechanism (Score), it can not only preserve the highly discriminative patches but also reduce the number of patches to single digit, then the top-1 classification accuracy on CUB-200-2011 dataset improves by 1.7%. Further, we take account into the discriminative ability of region group through CGP module, and the ablation experimental results show that if each region aggregates all other regions with same frequencies (CGP-SF), the accuracy is 87.2%, while the criss-cross propagation can achieve better performance, i. e. 87.7%, on CUB. Finally, we introduce the CFS module to explore and exploit internal correlation between selected patches and achieve the state-of-the-art result of 88.3%. Ablation experiments have validated that the proposed network can indeed learn the discriminative region group, enhance the discriminative feature values, effectively improving the accuracy.

Quantitative Comparisons

Accuracy comparison. Because the proposed model only use image-level annotations instead of any object or part annotations, our comparisons mainly focus on the weakly supervised methods. We show the performance of different methods on CUB-200-2011 dataset, Stanford Cars-196 dataset and FGVC Aircraft dataset respectively. From top to bottom in Tab. 3, we separate the methods into six groups, which are (1) Supervised multi-stage methods, which normally rely on the object and even part annotations to achieve comparable result. (2) Weakly supervised multi-stage frameworks, which gradually beat the strong supervised methods through picking out discriminative regions. (3) Weakly supervised end-to-end feature encoding, which have good performance via encoding the CNN feature vectors into high-order information, while they result in high computational cost. (4) End-to-end localization-classification subnetworks, which work well on various datasets, but they neglect the correlation between discriminative regions. (5) Other methods also achieve good performance due to using the extra information (e.g. the semantic embedding). (6)

Table 3: Comparison of different methods on CUB-200-2011(CUB), Cars 196 (Cars) and Aircraft (Airs).

Method	Box Anno.	Part Anno.	CUB Acc.	Cars Acc.	Airs Acc.
PN-DCN (Branson et al.)	BBox	Parts	85.4%	-	-
M-CNN (Wei, Xie, and Wu)	n/a	Parts	84.2%	-	-
PG (Krause et al.)	BBox	n/a	82.8%	92.8%	-
SCDA (Wei et al.)	n/a	n/a	80.1%	85.1%	79.5%
AutoBD (Yao et al.)	n/a	n/a	81.6%	88.9%	-
OPAM (Peng, He, and Zhao)	n/a	n/a	85.8%	92.2%	-
Bilinear (Lin, Roy Chowdhury, and Maji)	n/a	n/a	84.0%	91.3%	84.1%
lB-CNN (Kong and Fowlkes)	n/a	n/a	84.2%	90.9%	87.3%
Kernel-Activation (Cai, Zuo, and Zhang)	n/a	n/a	85.3%	91.7%	88.3%
Kernel-Pooling (Cui et al.)	n/a	n/a	86.2%	92.4%	85.7%
WSDL (He, Peng, and Zhao)	n/a	n/a	83.5%	-	-
RA-CNN (Fu, Zheng, and Mei)	n/a	n/a	85.3%	92.5%	88.2%
MA-CNN (Zheng et al.)	n/a	n/a	86.5%	92.8%	89.9%
MAMC (Sun et al.)	n/a	n/a	86.5%	93.0%	-
DFL-CNN (Wang, Morariu, and Davis)	n/a	n/a	87.4%	93.1%	91.7%
DCL (Chen et al.))	n/a	n/a	87.8%	**94.5**%	93.0%
DT-RAM (Li et al.)	n/a	n/a	86.0%	93.1%	-
StackDRL (He, Peng, and Zhao)	n/a	n/a	86.6%	-	-
KERL (Chen et al.)	n/a	n/a	87.0%	-	-
Our GCL	n/a	n/a	**88.3%**	94.0%	**93.2%**

Table 4: Comparison of the efficiency and effectiveness with other methods on CUB-200-2011. K means the number of selected discriminative regions for each image.

Method	Annotation	Accuracy	Speed(FPS)
M-CNN(K=2)	Parts	84.20%	12.90
MAMC(K=0)	n/a	86.50%	9.79
WSDL(K=1)	n/a	83.45%	10.07
Bilinear(K=0)	n/a	84.00%	30.00
Our GCL(K=2)	n/a	87.80%	**56.00**
Our GCL(K=4)	n/a	**88.30**%	55.00

Our end-to-end GCL approach achieves new state-of-the-art without any extra annotations and enjoys consistent performance on various datasets.

Our approach outperforms these strong supervised methods in the first group, which indicates that the proposed method can really find the discriminative patches without any fine-grained annotations. Our proposed method considers the correlations among regions to select the discriminative region group, and then outperforms other methods through selected discriminative patches in group 4. Meanwhile, we finely mine the internal semantic correlations between selected discriminative patches to emphasize the informative features and suppress the less helpful ones. Therefore, our work outperforms other methods via strengthening feature representations in group 3, and achieves new state-of-the-art accuracy of 88.3% on CUB, 94.0% on Cars and 93.5% on Aircraft.

Compared with MA-CNN, which consider the correlation between patches implicitly through the channel grouping loss where the spacial constraint is applied over part attention maps in a back-propagation way, our work proposes to find the most discriminative region group through iterative criss-cross graph propagation, and spacial context is incorporated into the network in a front-propagation way. The experimental results in Tab. 3 show that the GCL model achieves better performance than MA-CNN on CUB, CAR and AIRCRAFT.

Table 5: Comparisons with different depths of CGP and CFS in our model. D is the depth.

Method	$(D=2)$Acc.	$(D=3)$Acc.	$(D=4)$Acc.
CGP	87.4%	87.8%	87.5%
CFS	88.3%	88.3%	88.1%

The results in Tab. 2 show that our model outperforms most of the compared models except for a bit lower than DCL on CAR dataset. We think the reason is that images of CAR dataset have much simpler and clearer backgrounds than those of CUB and AIRCRAFT. Concretely, the proposed GCL model focuses on strengthening the responses of discriminative region group, which leads to better locating of discriminative patches in images with complex backgrounds. However, locating discriminative patches in images with simple backgrounds is relatively easier and thus might not obviously benefit from the strengthening of response of discriminative region group. On the other hand, DCL model's shuffle operation in region confusion mechanism could introduce several noisy visual patterns and thus the complexity of the image's background is one of the key factors that influence the locating accuracy of the discriminative patches for DCL. As a result, DCL shows better performance on CAR dataset for its simpler background while our GCL model outperforms on CUB and AIRCRAFT for their complex backgrounds.

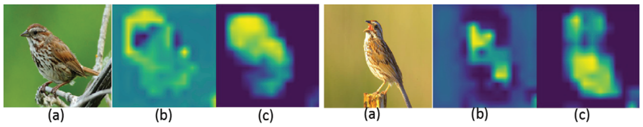


Figure 4: Visualization results of without and with correlation between regions. (a) shows the original images. (b)(c) are the certain corresponding channel feature map, without and with correlation, respectively.

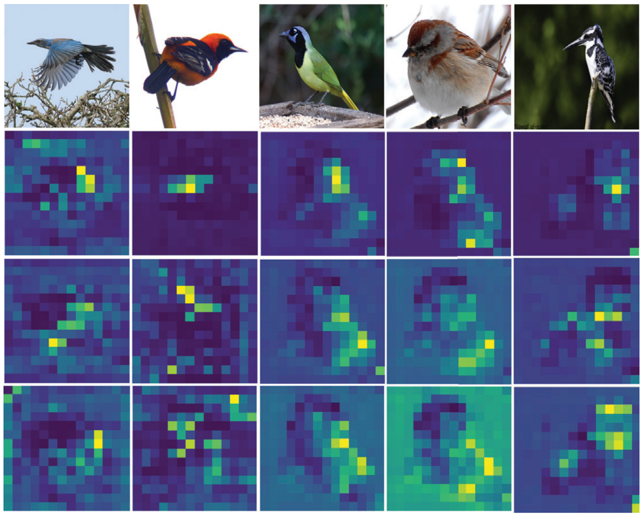


Figure 5: Visualization results correlation weight coefficient maps. The first row indicates the original images. The second, the third and the forth rows indicate the correlation weight coefficient maps via the first, the second and the third graph-propagation, respectively.

Speed analysis. We measure the speed with batch size 8 using a graphics card of Titan X. Tab. 4 shows the comparison with other methods. Note that the references of other methods are shown in Tab. 3. WSDL (He, Peng, and Zhao) used the framework of the faster RCNN (Ren et al.), which reserved about 300 candidate patches. In this work, we utilize the score mechanism with rank loss to reduce the number of patches to single figure, to achieve real-time efficiency. We outperform other methods both in speed and accuracy when we select 2 discriminative patches according to the discriminative score maps. Futher, when we increase the number of discriminative patches to 4, the proposed model not only achieves the state-of-the-art classification precision, but also stays real time at 55 fps.

Qualitative Analysis

To verify the effectiveness of CGP, we do the ablation experiments and visualize M_O (Fig. 4 (b)) and M_U (Fig. 4 (c)). The visualization results show that M_O highlights the multiple continuous regions, while M_U strengthens the most discriminative regions after multiple iterations of criss-cross propagation, which helps the accurate locating of the discriminative region group.

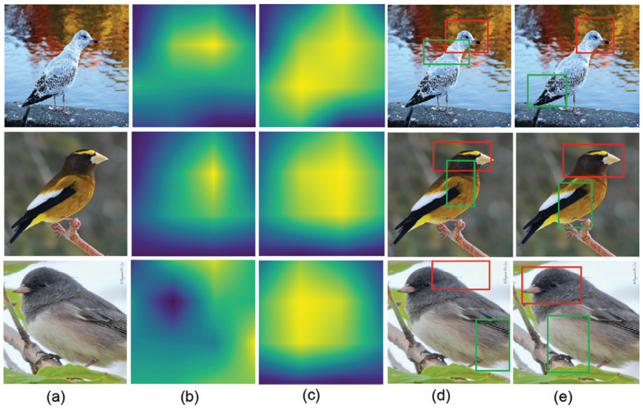


Figure 6: Visualization results of without and with correlation between regions. (a) shows the original images. (b)(c) are the discriminative score maps and (d)(e) are the localization results, without and with correlation, respectively.

As shown in Fig.5, we visualize the correlation weight coefficient maps generated by CGP module for better illustrating the influence of correlation between regions. The correlation coefficient maps denote the correlation between a certain region and another region in the criss-cross positions of region. It can be observed that the correlation coefficient maps tend to focus on a few fixed regions (highlighted regions in Fig.5) and integrate more discriminative regions gradually via CGP by joint learning. The region closer to the aggregation is calculated with higher frequency.

What's more, we visualize the discriminative score maps with and without the CGP to illustrate the effectiveness of the CGP module, as shown in Fig.6. We can see from the discriminative score maps without the CGP in the second column, which only only focuses on a local region and the selected patches in the fourth column are in the close regions. However, from the discriminative score maps without the CGP in the second column and the selected patches in the fifth column, it proves that our CGP sub-network indeed pays attention to multiple effective regions, making the region aggregating features more discriminative.

The deeper, the better?

We show the classification results with different depth of the two graph propagation sub-networks in Tab. 5. Specifically, for CFS module, the input and output dimension are both 2048, and all intermediate node dimension is 1024. For CGP module, the input, intermediate and output node dimension are always 2048. As shown in Tab. 5, when the number of graph layers increases to 4, the classification performance drops on both sub-networks.The possible reason of the performance drop is that after using more graph layers, the propagation between nodes will be overwhelmed.

Conclusion

In this paper, we proposed a novel end-to-end Graph-Propagation based Correlation Learning model to fully mine and exploit the discriminative potentials of correlations for

WFGIC. In particular, we design a CGP module to learn global and spatial correlations in the criss-cross graph propagation, which helps discovering the discriminative region group. Meanwhile, we also construct a CFS graph network with feature vector correlationt jointly enhance the features extracted from selected patches. The experimental results on the widely used CUB-200-2011 and Cars-196 datasets shown that our proposed model is effective and achieves state-of-the-art.

Acknowledgments: This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grants No. 61772108, No. 61932020 and No. 61976038.

References

Branson, S.; Horn, G. V.; Belongie, S. J.; and Perona, P. 2014. Bird species categorization using pose normalized deep convolutional nets. *CoRR* abs/1406.2952.

Cai, S.; Zuo, W.; and Zhang, L. 2017. Higher-order integration of hierarchical convolutional activations for fine-grained visual categorization. In *ICCV 2017, Venice, Italy, October 22-29, 2017*, 511–520.

Chen, T.; Lin, L.; Chen, R.; Wu, Y.; and Luo, X. 2018. Knowledge-embedded representation learning for fine-grained image recognition. In *IJCAI*, 627–634.

Chen, Y.; Bai, Y.; Zhang, W.; and Mei, T. 2019. Destruction and construction learning for fine-grained image recognition. In *CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, 5157–5166.

Cui, Y.; Zhou, F.; Wang, J.; Liu, X.; Lin, Y.; and Belongie, S. J. 2017. Kernel pooling for convolutional neural networks. In *CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, 3049–3058.

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 2017, Honolulu, HI, USA, July 21-26, 2017*, 4476–4484.

Gao, Y.; Beijbom, O.; Zhang, N.; and Darrell, T. 2016. Compact bilinear pooling. In *CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, 317–326.

He, X., and Peng, Y. 2017a. Fine-grained image classification via combining vision and language. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, 7332–7340.

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

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, 770–778.

He, X.; Peng, Y.; and Zhao, J. 2017. Fine-grained discriminative localization via saliency-guided faster R-CNN. In *ACM MM 2017, Mountain View, CA, USA, October 23-27, 2017*, 627–635.

He, X.; Peng, Y.; and Zhao, J. 2018. Stackdrl: Stacked deep reinforcement learning for fine-grained visual categorization. In *IJCAI 2018, July 13-19, 2018, Stockholm, Sweden.*, 741–747.

Kipf, T. N., and Welling, M. 2017. Semi-supervised classification with graph convolutional networks. In *ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*.

Kong, S., and Fowlkes, C. C. 2017. Low-rank bilinear pooling for fine-grained classification. In *CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, 7025–7034.

Krause, J.; Stark, M.; Deng, J.; and Fei-Fei, L. 2013. 3d object representations for fine-grained categorization. In *ICCV Workshops 2013, Sydney, Australia, December 1-8, 2013*, 554–561.

Krause, J.; Jin, H.; Yang, J.; and Li, F. 2015. Fine-grained recognition without part annotations. In *CVPR 2015, Boston, MA, USA, June 7-12, 2015*, 5546–5555.

Li, Z.; Yang, Y.; Liu, X.; Zhou, F.; Wen, S.; and Xu, W. 2017. Dynamic computational time for visual attention. In *ICCV Workshops 2017, Venice, Italy, October 22-29, 2017*, 1199–1209.

Lin, T.; Roy Chowdhury, A.; and Maji, S. 2015. Bilinear CNN models for fine-grained visual recognition. In *ICCV 2015, Santiago, Chile, December 7-13, 2015*, 1449–1457.

Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S. E.; Fu, C.; and Berg, A. C. 2016. SSD: single shot multibox detector. In *ECCV, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part I*, 21–37.

Maji, S.; Rahtu, E.; Kannala, J.; Blaschko, M. B.; and Vedaldi, A. 2013. Fine-grained visual classification of aircraft. *CoRR* abs/1306.5151.

Peng, Y.; He, X.; and Zhao, J. 2018. Object-part attention model for fine-grained image classification. *TIP* 27(3):1487–1500.

Ren, S.; He, K.; Girshick, R. B.; and Sun, J. 2017. Faster R-CNN: towards real-time object detection with region proposal networks. *TPAMI.* 39(6):1137–1149.

Sun, M.; Yuan, Y.; Zhou, F.; and Ding, E. 2018. Multi-attention multi-class constraint for fine-grained image recognition. In *ECCV, Munich, Germany, September 8-14, 2018, Proceedings, Part XVI*, 834–850.

Wang, Y.; Morariu, V. I.; and Davis, L. S. 2018. Learning a discriminative filter bank within a CNN for fine-grained recognition. In *CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, 4148–4157.

Wei, X.; Luo, J.; Wu, J.; and Zhou, Z. 2017. Selective convolutional descriptor aggregation for fine-grained image retrieval. *TIP* 26(6):2868–2881.

Wei, X.; Xie, C.; and Wu, J. 2016. Mask-cnn: Localizing parts and selecting descriptors for fine-grained image recognition. *CoRR* abs/1605.06878.

Yang, Z.; Luo, T.; Wang, D.; Hu, Z.; Gao, J.; and Wang, L. 2018. Learning to navigate for fine-grained classification. In *ECCV, Germany, September 8-14, 2018, Proceedings, Part XIV*, 438–454.

Yao, H.; Zhang, S.; Yan, C.; Zhang, Y.; Li, J.; and Tian, Q. 2018. Autobd: Automated bi-level description for scalable fine-grained visual categorization. *TIP* 27(1):10–23.

Zhang, X.; Xiong, H.; Zhou, W.; Lin, W.; and Tian, Q. 2016. Picking deep filter responses for fine-grained image recognition. In *CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, 1134–1142.

Zheng, H.; Fu, J.; Mei, T.; and Luo, J. 2017. Learning multi-attention convolutional neural network for fine-grained image recognition. In *ICCV 2017, Venice, Italy, October 22-29, 2017*, 5219–5227.