Channel Interaction Networks for Fine-Grained Image Categorization

Yu Gao, Xintong Han, Xun Wang, Weilin Huang,[*] Matthew R. Scott
Malong Technologies, Shenzhen, China
Shenzhen Malong Artificial Intelligence Research Center, Shenzhen, China
{chrgao, xinhan, xunwang, whuang, mscott}@malong.com

Abstract

Fine-grained image categorization is challenging due to the subtle inter-class differences. We posit that exploiting the rich relationships between channels can help capture such differences since different channels correspond to different semantics. In this paper, we propose a channel interaction network (CIN), which models the channel-wise interplay both within an image and across images. For a single image, a self-channel interaction (SCI) module is proposed to explore channel-wise correlation within the image. This allows the model to learn the complementary features from the correlated channels, yielding stronger fine-grained features. Furthermore, given an image pair, we introduce a contrastive channel interaction (CCI) module to model the cross-sample channel interaction with a metric learning framework, allowing the CIN to distinguish the subtle visual differences between images. Our model can be trained efficiently in an end-to-end fashion without the need of multi-stage training and testing. Finally, comprehensive experiments are conducted on three publicly available benchmarks, where the proposed method consistently outperforms the state-of-the-art approaches, such as DFL-CNN(Wang, Morariu, and Davis 2018) and NTS(Yang et al. 2018).

Introduction

Fine-grained image categorization has become an important topic in computer vision community with broad application prospects such as new retail (Karlinsky et al. 2017), automatic driving (Sochor, Herout, and Havel 2016), etc. Going beyond classical image classification that recognizes basic-level categories, fine-grained categories are much more challenging to be identified due to the subtle inter-class differences, many of which can only be effectively distinguished by concentrating on discriminative local parts. For instance, to distinguish three bird species in Figure 1, a neural network usually focuses on their wings and heads.

Previous work tends to learn discriminative features by locating distinct parts (Jaderberg et al. 2015; Fu, Zheng, and Mei 2017; Yang et al. 2018) or modeling higher order information (Lin, RoyChowdhury, and Maji 2015; Gao et al. 2016; Kong and Fowlkes 2017; Yu et al. 2018), which have been proven to be effective for fine-grained image classification. In this paper, we rethink the way of learning discriminative features with convolutional networks, and propose a new channel interaction network (CIN).

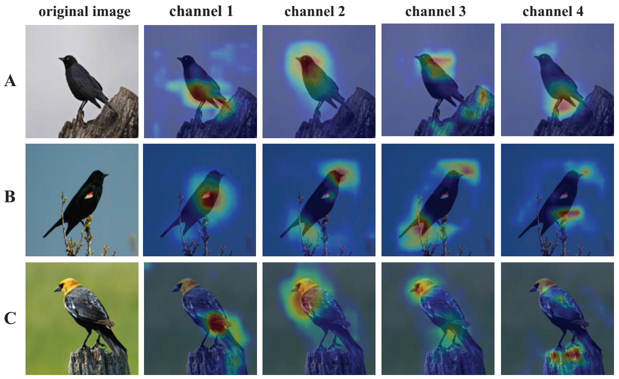

Figure 1: Channel activations computed by our method (from the `conv5_3` layer of ResNet50 trained on CUB-200-2011 dataset).

First, different channels often correspond to different visual patterns (Yosinski et al. 2015). As shown in Figure 1, most of the channels are semantically complementary to each other. Motivated by this observation, we aim to discover the complementary channel information for each individual channel, and then aggregate the complementary channels with the original ones. Such complementary information can cooperatively contribute to the referred channel, making the channel more discriminative. Consequently, we propose a self-channel interaction (SCI) network that explicitly models the relationships between various channels to discover such channel-wise complementary clues. Existing methods usually apply the channel/part interplay for direct classification (Lin, RoyChowdhury, and Maji 2015; Yu et al. 2018) or attempt to mine the closely related cues (Wang et al. 2017; Yue et al. 2018), where the channel-wise complementary clues are not fully explored.

Second, people tend to distinguish two images by focusing on their specific distinctions. For instance, when we

[*] Weilin Huang is the corresponding author.

compare two images, A and B, in Figure 1, it is easy to identify the difference of wings between two images: the bird in A has black wings while there is a red dot on the wings of B. However, when images of A and C are compared, more attention should be paid to the regions of heads, i.e., enhancing the importance of channel 2 for image A and image C. To this end, we propose a novel contrastive channel interaction (CCI) mechanism between samples, with the goal of capturing the subtle differences between images. Metric learning is incorporated with our framework to model the cross-sample channel interactions, which is neglected by most of the existing methods (Wang, Morariu, and Davis 2018; Yang et al. 2018).

Finally, we jointly optimize the SCI module and the CCI module, as shown in Figure 2. The network can be trained end-to-end in one stage, and thus is more lightweight than the two-stage methods like HS-Net (Lam, Mahasseni, and Todorovic 2017), DFL-CNN (Wang, Morariu, and Davis 2018), NTS (Yang et al. 2018), etc.. Our major contributions are summarized as:

1) We propose a self-channel interaction (SCI) module able to model the interplay between different channels within an image. This enables it to capture the channel-wise complementary information for each channel, which enhances the discriminative features learned by each channel. This results in a lightweight model that can be trained more effectively in one stage. The new model is flexible, and can be seamlessly integrated into existing networks to boost the performance.

2) We propose a novel contrastive channel interaction (CCI) module to learn channel-wise relationships between images. CCI is able to dynamically identify the distinct regions from two compared images, allowing the model to focus on such distinctive regions for better categorization.

3) Finally, we evaluate our approach on three publicly available datasets: CUB-200-2011 (Wah et al. 2011), Stanford Cars (Krause et al. 2013) and FGVC Aircraft (Maji et al. 2013), where our method achieves better performance over current state-of-the-art.

Related Work

Fine-grained feature representations. Building powerful feature representations has been broadly studied for fine-grained image categorization. Unlike using the first-order features directly for classification (Zhang et al. 2014; Wei et al. 2018), (Lin, RoyChowdhury, and Maji 2015) employ bilinear pooling with two independent CNNs, which take pairwise feature interactions into consideration by computing the second-order information, leading to performance improvements. To reduce the computation complexity, (Gao et al. 2016; Kong and Fowlkes 2017) tend to use less feature dimensions, while (Cui et al. 2017) attempt to model higher-order information to improve the accuracy. (Wang, Li, and Zhang 2017; Li et al. 2018) apply a matrix power normalization for computing bilinear features. (Yu et al. 2018) further explore cross-layer bilinear pooling to compute multi-layer knowledge. Unlike these methods using second or higher order information for direct classification, we compute second-order statistics between different channels, which are used jointly with the original features to capture the channel-wise complementary information, resulting in stronger deep representations.

Visual attention. Visual attention, which has been introduced in various computer vision applications, can be employed to capture the subtle inter-class differences in fine-grained image categorization. For example, hard attention based methods, such as (Jaderberg et al. 2015; Fu, Zheng, and Mei 2017; Li et al. 2017; Yang et al. 2018), usually detect local regions and then crop them out from the original image. But the main limitation is that each cropped region requires an extra feedforward operation. Instead, soft attention methods (Zheng et al. 2017; Sun et al. 2018) can be regarded as imposing a soft mask on the feature maps, by only using a single feedforward stage. Self-attention was proposed and applied in machine translate in (Vaswani et al. 2017). It can be categorized into the soft attention. The non-local block, introduced in (Wang et al. 2017), is highly related to the self-attention module, but captures long-range dependencies in space-time dimension in images and videos. (Yue et al. 2018; Zheng et al. 2019) further explore the non-local like ideas in fine-grained classification.

In contrast to these self-attention based methods, we exploit the interactions between channels to discover the channel-wise complementary information rather than mining the closely related channels. Moreover, we further propose a contrastive channel interaction module to model cross-sample channel interactions.

Metric learning. Deep metric learning aims to learn a feature embedding for better measuring the similarities between image pairs, i.e., the distance of positive pairs are encouraged to be closer and the negative pairs are pushed away from each other. It has been widely used in various domains such as face verification (Hu, Lu, and Tan 2014; Schroff, Kalenichenko, and Philbin 2015), image retrieval (Wang et al. 2014), person re-id (Chen et al. 2017; Varior, Haloi, and Wang 2016), etc. Compared with softmax loss used in conventional classification networks, metric learning can embed the samples into a low-dimensional space capturing high intra-class variance, which is more suitable for fine-grained image categorization (Cui et al. 2016). Recent work of MAMC (Sun et al. 2018) adopts metric learning to compute the rich correlations between object parts, which inspired the current work. But our major differences lie in two aspects: 1) MAMC utilizes two attention branches to compute the features of two different part, while we model the interplay between different channels *explicitly* to extract the discriminative features; 2) a novel contrastive channel interaction module is proposed in our networks to emphasize the differences between contrastive samples.

Methodology

In this section, we present our proposed channel interaction network (CIN) for fine-grained image categorization, as illustrated in Figure 2. Given an image pair, the two images are first processed by a shared backbone, e.g., ResNet-50 (He et al. 2016), generating a pair of convolutional fea-

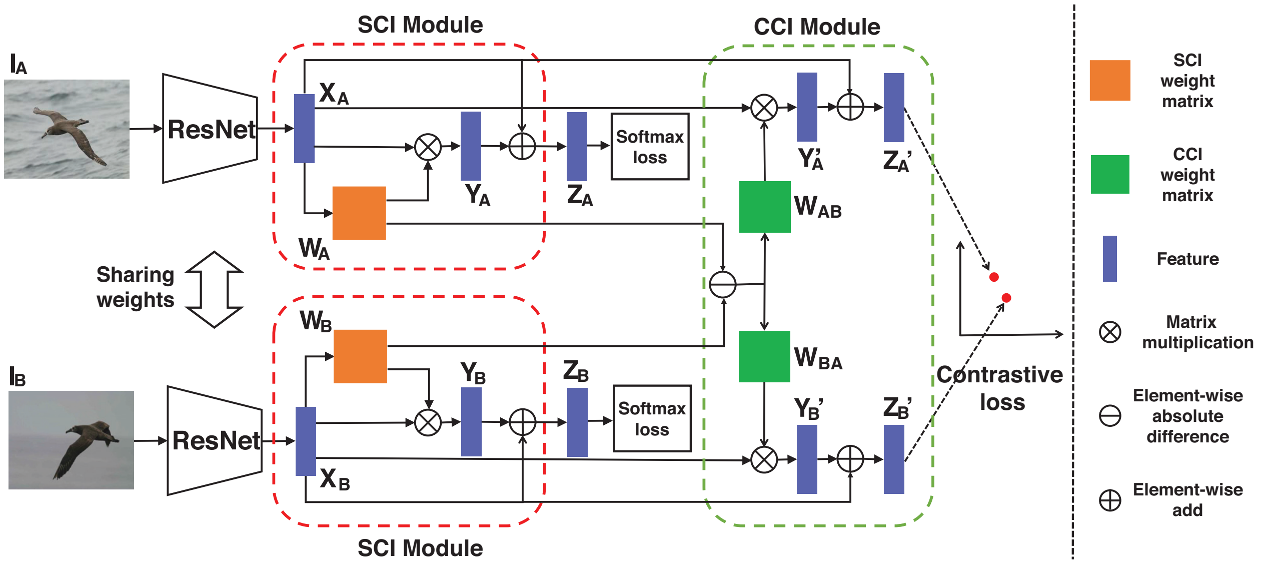

Figure 2: Overview of our network architecture. $W_{AB} = |W_A - \eta W_B|, W_{BA} = |W_B - \gamma W_A|$, η and γ are learned by an fc layer controlling the encoded information computed from the contrastive image for highlighting differences. CCI module will be removed and softmax loss will be replaced by a softmax layer during inference.

ture maps. To compute channel-wise complementary information for each channel on the feature maps, a self-channel interaction (SCI) module is designed to model the correlations between different channels. Then we aggregate the discriminate features from the original feature maps and the complementary information jointly. Finally, a contrastive channel interaction (CCI) module is designed with a contrastive loss to model the channel-wise relationships between two images. Compared with existing methods, our proposed CIN can be trained end-to-end in one stage, and also is readily applicable to other convolution neural networks.

Self-Channel Interaction

Being aware of the rich knowledge encoded in the feature channels, as shown in Figure 1, we would like to explore the interaction between various channels. Recent work (Hu, Shen, and Sun 2017; Sun et al. 2018) tends to highlight the most distinct feature channels. However, only focusing on the most discriminate channels might not fully explore the rich information from all channels. Indeed, most of the channels are complementary to each other. We attempt to compute the channel-wise relationships to extract such complementary clues, and then encode them into the original features for fine-grained classification. Thus we propose a simple yet effective self-channel interaction (SCI) module to achieve such ability, as shown in Figure 2.

Given an image I, let $X' \in \mathbb{R}^{w \times h \times c}$ denote the input feature maps processed by the backbone, where w, h and c indicate the height, width and the number of channels. We first reshape the input feature maps X' to $X \in \mathbb{R}^{c \times l}, l = w \times h$. Then the output of SCI is computed as:

$$Y = WX \in \mathbb{R}^{c \times l}, \tag{1}$$

where $W \in \mathbb{R}^{c \times c}$ denotes the SCI weight matrix, which can be computed as follows. Firstly, we perform a bilinear operation between X and $X^\top$, obtaining a bilinear matrix, $XX^\top$. Then we add a minus sign to it and exploit a softmax function to get the weight matrix:

$$W_{ij} = \frac{exp(-XX^\top{}_{ij})}{\sum_{k=1}^{c} exp(-XX^\top{}_{ik})}, \tag{2}$$

where $\sum_{k=1}^{c} W_{ik} = 1$. It is worth noting that Y_i (the i^{th} channel of the resulting features Y) is the computed interaction between X_i and all the channels of X, i.e., $Y_i = W_{i1}X_1 + \cdots + W_{ic}X_c$.

According to the definition of W, the channels with larger weights tend to be semantically complementary with X_i, as illustrated in Figure 3. The referred channel X_i focuses on the head part, thus the channels highlighting the complementary parts, like wings and feet, have larger weights, while the channel with head part emphasized has a smaller weight. As the resulting features Y may discard some information from the original features, we aggregate the discriminate features (Z) from both the generated features and the original ones:

$$Z = \phi(Y) + X, \tag{3}$$

where ϕ denotes a 3×3 convolutional layer.

Discussions. It is worth noting that our SCI module can be formalized as the non-local like operation described in (Wang et al. 2017):

$$Y = f(X, X)g(X), \tag{4}$$

where $f(X, X) = softmax(-XX^\top) \in \mathbb{R}^{c \times c}$, and $g(X) = X \in \mathbb{R}^{c \times l}$. Unlike the original non-local block considering the interactions in spatial dimension, our module

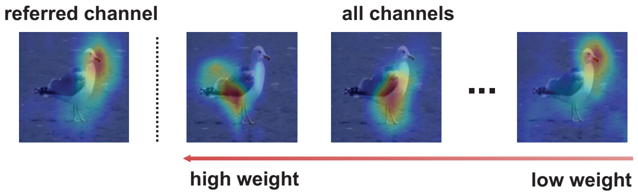

Figure 3: An example of the relationship between the referred channel with all the channels in SCI module.

focuses on channel dimension. More importantly, the non-local operation tends to exploit the positive correlations between spatial positions, while our SCI module focuses on the negative correlations, which enables our model to discover the semantically complementary channel information. The non-local operation is similar to the SE-Inception module (Hu, Shen, and Sun 2017). They highlight the discriminative features but do not make full use of the complementary clues, which are better explored by our SCI module to enhance the channel-wise features, as shown in Figure 5.

Our method is related to CGNL (Yue et al. 2018) and TSAN (Zheng et al. 2019) which also compute the channel correlations, but has clear distinctions on measuring the correlations: 1) CGNL and TSAN explore positive channel interaction while our CIN focuses on negative channel interaction; 2) CGNL takes spatial correlation into account, and further posits a low-rank Hadamard product; 3) In TSAN, the authors further proposed an adaptive image sampling mechanism to enhance the detailed information and applied knowledge distilling to extract the learned details. 4) Beside computing the channel-wise relationship within an image, a key distinction of our method is to further apply metric learning to model the channel interplay between samples.

Contrastive Channel Interaction

SCI module is able to compute meaningful discriminate features. A straightforward approach is to directly feed the features for classification, e.g., by using a softmax classifier as the most popular choice. However, a vanilla classifier usually fails to capture the subtle differences present for fine-grained classification (Cui et al. 2016). To mitigate this problem, MAMC (Sun et al. 2018) was recently proposed to enforce the correlations between different object parts. It introduces multi-attention multi-class constraints by using a metric learning technology, which inspired the current work. We employ deep metric learning to compute rich cross-sample channel-wise correlations by introducing contrastive constraints to the features enhanced by SCI.

To model this interaction between two images I_A and I_B, a natural idea is to impose the contrastive constraints on the features Z_A and Z_B enhanced by SCI, and then measure their similarity. However, traditional deep metric learning approaches project an image into a fixed point in the learned embedding space. As a result, such a general representation often fails to capture the subtle differences between two images. In contrast, we attempt to learn the interactions between two images in a dynamic manner where the channels are emphasized by comparing to the feature channels computed from the contrastive image.

Consequently, we propose a contrastive channel interaction (CCI) module to compute such relationships between two images. As illustrated in Figure 2, we argue that a simple subtraction operation between the SCI weight matrices of image I_A and I_B might extract such mutual information, and generate CCI weight matrices W_{AB} and W_{BA}:

$$W_{AB} = |W_A - \eta W_B|, W_{BA} = |W_B - \gamma W_A|, \qquad (5)$$

where η and γ are the weights learned by $[Y_A, Y_B]$ and $[Y_B, Y_A]$ through a FC layer ψ, i.e., $\eta = \psi([Y_A, Y_B]), \gamma = \psi([Y_B, Y_A])$, and $||$ denotes the absolute value. The two weights indicate the amount of correlated information considered dynamically by the image to better distinguish itself from the compared one. We use a subtraction operation to compute the interaction. We also tried other operations like addition, multiplication, or concatenation, with slightly lower performance obtained. By subtraction, the CCI weight matrices suppress the commonality and highlight the distinct channel relationships between the two images.

Then similar to SCI module, the CCI weight matrices W_{AB} and W_{BA} are applied to the features X_A and X_B as:

$$Z'_A = \phi(Y'_A) + X_A, Z'_B = \phi(Y'_B) + X_A, \qquad (6)$$

where $Y'_A = W_{AB}X_A$ and $Y'_B = W_{BA}X_B$.

Finally, a contrastive loss (Hadsell, Chopra, and LeCun 2006) is applied to the features computed by the CCI module which aims to push the samples of different classes away while pulling the positive image pairs close. Suppose each batch contains N image pairs, i.e., 2N images. The contrastive loss is defined as follows:

$$L_{cont} = \frac{1}{N} \sum_{A,B} \ell(Z'_A, Z'_B). \qquad (7)$$

Beyond the contrastive loss, a triplet loss (Schroff, Kalenichenko, and Philbin 2015) and other losses of metric learning can be used in our framework as well. The reason we choose the contrastive loss is that it is simple, and perform well in metric learning and face verification (Taigman et al. 2014; Hadsell, Chopra, and LeCun 2006). We also tried to use a triplet loss in CCI, but did not improve the performance. Specifically, ℓ is defined as follows:

$$\ell = \begin{cases} ||h(Z'_A) - h(Z'_B)||^2, & \text{if } y_{AB} = 1 \\ max(0, \beta - ||h(Z'_A) - h(Z'_B)||)^2, & \text{if } y_{AB} = 0 \end{cases} \qquad (8)$$

where β is a predefined margin and $|| \cdot ||$ denotes the Euclidean distance, h is a fully-connected layer projecting features into an r-dimension space, i.e. $H(Z) \in \mathbb{R}^r$. r is set to 512 in our experiments. Here, y_{AB} indicates whether the label of an image pair is the same or not, i.e., $y_{AB} = 1$ denotes image I_A and image I_B come from the same class, while $y_{AB} = 0$ means a negative pair.

Moreover, we use a softmax loss for classification based on the predictions that are generated by the features Z using SCI. We denote the softmax loss as L_{soft}. The total loss L_{total} of our framework is defined as follows:

$$L_{total} = L_{soft} + \alpha \cdot L_{cont}, \qquad (9)$$

where α is a hyper-parameter. We use the stochastic gradient method to optimize L_{total}. Note that only SCI module is used in inference, with only a single image required.

Experiments

We report the experimental results, and compare our method with the state-of-the-art approaches.

Datasets and Baselines

Datasets. We employ three publicly available datasets in our experiments: (1) CUB-200-2011 (Wah et al. 2011) with 11,788 images from 200 wild bird species, (2) Stanford Cars (Krause et al. 2013) including 16,185 images over 196 classes, and (3) FGVC Aircraft (Maji et al. 2013) containing 196 classes about 10,000 images.

Baselines. In the experiments, we compare our CIN with 10 methods described as follows. The first four methods can be trained in one stage. (1) MAMC (Sun et al. 2018): applying multi-attention multi-class constraints to enforce the correlations among different parts of objects. (2) CGNL (Yue et al. 2018): capturing the dependencies between positions across channels by non-local operation to classify. (3) HBP (Yu et al. 2018): hierarchical bilinear pooling framework integrating multiple cross-layer bilinear features. (4) iSQRT-COV (Yu et al. 2018): using an iterative matrix square root normalization to do covariance pooling. (5) RA-CNN (Fu, Zheng, and Mei 2017): recursively learning discriminative region attention and region-based feature representation at multiple scales. (6) Boost-CNN (Moghimi et al. 2016): a new boosting strategy to assemble weak classifiers for better performance. (7) DT-RAM (Li et al. 2017): a dynamic computational time model with reinforcement learning for recurrent visual attention. (8) MA-CNN (Zheng et al. 2017): multi-attention convolutional network including convolution, channel grouping and part classification sub-networks. (9) DFL-CNN (Wang, Morariu, and Davis 2018): capturing class-specific discriminative patches by learning a bank of convolutional filters. The performance might be unstable due to the complex layer initialization using k-means. (10) NTS (Yang et al. 2018): effectively localizing informative regions with self-supervision mechanism.

Notice that we do not compare our method with the approaches which require additional information, such as SJS (Ge and Yu 2017), HS-Net (Lam, Mahasseni, and Todorovic 2017), and HSE (Chen and others 2018).

Implementation Details

In all our experiments, we use ResNet-50 and ResNet-101 as our base networks. We remove the last pooling layer and fully-connected layer, and then fine-tune the networks pretrained on ImageNet (Russakovsky et al.). The input image size is 448 × 448 as most state-of-the-art fine-grained categorization approaches. By following that of NTS (Yang et al. 2018), we implement data augmentation including random cropping and horizontal flipping during training. Only center cropping is involved in inference.

Method	1-Stage	ACC	Time
VGG-19	√	80.2%	22.1
ResNet-50	√	84.9%	12.5
ResNet-101	√	85.4%	22.4
ResNet-50 + SE	√	85.7%	14.0
ResNet-50 + Pos-SCI	√	86.1%	17.2
ResNet-50 + Non-local	√	86.6%	14.2
ResNet-50 + MAMC (Sun et al. 2018)	√	86.3%	14.8
ResNet-50 + CGNL (Yue et al. 2018)	√	87.0%	15.0
ResNet-50 + SCI	√	87.1%	17.2
ResNet-50 + SCI + Cont	√	87.2%	17.2
ResNet-50 + NTS (Yang et al. 2018)	×	87.5%	23.6
ResNet-50 + CIN	√	87.5%	17.2
ResNet-101 + CIN	√	**88.1%**	27.2

Table 1: Ablation studies of our network on CUB-200-2011. CIN consists of SCI and CCI. Time unit is ms.

The model is trained for 100 epochs with SGD for all datasets, and the base learning rate is set to 0.001, which annealed by 0.5 every 20 epochs. we use a batch size of 20 and ensure that each batch contains 4 categories with 5 images in each category. And then, we randomly split these 20 images into 10 image pairs. We have tried to use all the $O(n^2)$ pairs or apply hard negative mining, which hurt the performance and consume more memory. The weight decay is set to 2×10^{-4}. β in Equation 8 is set to 0.5 empirically. and α in Equation 9 is set to 2.0. Top-1 accuracy is used as the evaluation metric. We use PyTorch to implement our method.

Ablation Analysis

We conduct ablation studies in order to better understand the impact of each component to our approach. The performance and efficiency are compared in in Table 1. We use ResNet-50 and ResNet-101 (He et al. 2016) as our backbone.

SCI Module. SCI mines complementary channels through exploring channel interactions, contributing to learning more discriminative features. As illustrated in Table 1, compared with ResNet-50 alone (84.9%), by merely adding the SCI module, ResNet-50 + SCI obtains a performance improvement of 2.2%. Moreover, switching the interaction module from SCI to SE module leads to a significant performance drop (87.1% vs. 85.7%). SE module only focuses on the most discriminative features and ignores others, while our SCI module utilizes the complementary channel knowledge to enhance all the features. Compared to the Non-local block and ResNet-50+Pos-SCI (SCI weight matrix W without the negative sign) which model the positive space and channel-wise information respectively, our SCI module obtains better performance. Notice that our SCI also outperforms CGNL (Yue et al. 2018) (87.0%) which models the correlations between the positions of all channels. Indeed, the channel information explored in our SCI module is involved in CGNL as well. The major difference about the channel information lies in that our SCI exploits the negative interplay to find the channel-wise complementary information, while CGNL does not fully explore such information and computes the positive interaction to capture the closely related clues. These results demonstrate that: 1) for fine-grained image classification, the information contained

Method	1-Stage	Acc(CUB)	Acc(FGVC)	Acc(Stanford Cars)
MAMC (Sun et al. 2018)	√	86.5%	-	93.0%
CGNL (Yu et al. 2018)	√	87.0%	-	-
HBP (Yu et al. 2018)	√	87.1%	90.3%	93.7%
iSQRT-COV(8k) (Li et al. 2018)	√	87.3%	89.5%	91.7%
RA-CNN (Fu, Zheng, and Mei 2017)	×	85.3%	-	92.5%
Boost-CNN (Moghimi et al. 2016)	×	85.6%	88.5%	92.6%
DT-RAM (Li et al. 2017)	×	86.0%	-	93.1%
MA-CNN (Zheng et al. 2017)	×	86.5%	89.9%	92.8%
DFL-CNN (Wang, Morariu, and Davis 2018)	×	87.4%	92.0%	93.8%
NTS (Yang et al. 2018)	×	87.5%	91.4%	93.9%
CIN (ResNet-50)	√	87.5%	92.6%	94.1%
CIN (ResNet-101)	√	**88.1%**	**92.8%**	**94.5%**

Table 2: Comparison results on CUB-200-2011, FGVC Aircraft and Stanford Cars.

in the channel dimension is as powerful as complicated modeling across all dimensions; 2) finding the complementary channel clues can take full advantage of the channel interaction comparing with discovering the closely related channel information.

CCI Module. We further investigate the effectiveness of the proposed CCI module. Table 1 shows that the CCI module (ResNet-50+SCI+CCI) provides 0.4% performance improvement compared to the method without a contrastive loss (ResNet-50+SCI). To further demonstrate the characteristics of the contrastive channel attention module, we consider the approach (ResNet-50+SCI+Cont) which explicitly applies a contrastive loss to the features computed by SCI module, i.e., $\eta = 0$ and $\gamma = 0$ in Equation 5. As presented in Table 1, ResNet-50+SCI+Cont obtains a limited improvement with ResNet-50+SCI (87.2% vs. 87.1%). The reason might be that the common contrastive loss uses the same features of an image compared to any other image, which might reduce it is ability to focus on the distinct differences between two images, while our CCI module is capable of highlighting of the different regions. The results confirm that our CCI module has strong capability for modeling the relationship between two images.

Time cost. We report our inference time on a Nvidia TITAN XP GPU with PyTorch implementation. As shown in Table 1, CIN introduces an overhead that is much smaller than that of two-stage methods (ResNet-50+NTS), and is comparable to the other one-stage approaches.

Comparison with State-of-the-art

In this section, we compare our proposed network (CIN) with the state-of-the-art methods on the three publicly available datasets.

CUB-200-2011. Table 2 presents the classification results of CIN and the state-of-the-art methods. First, the accuracy of our proposed CIN is higher than all existing methods. Even with ResNet-50, our method achieves comparable result with NTS (Yang et al. 2018). However, NTS requires multiple stages for learning discriminative regions, resulting in more expensive cost on both time and space. Compared with the best one-stage method iSQRT-COV (8k) (Li et al. 2018), our method outperforms it by 0.2%. Note that the feature dimension of our method (2k) is significantly lower than iSQRT-COV (8k). Moreover, our method improves HBP (Yu et al. 2018) by 1.0%. The reason might be that HBP ignores the interaction between samples. It is notable that the backbone of HBP is VGG (Simonyan and Zisserman 2014), while the accuracy of CIN with the same backbone is 85.6%. We have tried to implement HBP and found it does not work well with ResNet. DFL- CNN (Wang, Morariu, and Davis 2018) achieves the best results on CUB (87.4%) with ResNet-50 backbone while ResNet-50+CIN achieves a higher accuracy with only one stage. As shown in (Wang, Morariu, and Davis 2018), ResNet does not always outperform VGG.

Dataset	CIN	NTS	NTS+CIN
CUB-200-2011	87.5%	87.5%	**88.3%**
FGVC Aircraft	92.6%	91.4%	**93.3%**
Stanford Cars	94.1%	93.9%	**94.4%**

Table 3: Combined our network with NTS.

FGVC Aircraft. Table 2 reports the performance on FGVC Aircraft dataset. DFL-CNN achieves the highest accuracy of 92.0%, outperforming NTS with 91.4%. Our method has a clearly higher accuracy than existing methods even with the backbone ResNet-50. The excellent results further confirm the superiority of our method. It is worth noting that the accuracy on FGVC Aircraft is generally higher than that of CUB-200-2011, because images of CUB-200-2011 contain much more label noises (4.4% as reported in (Van Horn et al.)) and class-irrelevant background, while images in FGVC Aircraft have relatively clean background, and airplanes often occupy a large portion of the image.

Stanford Cars. To verify the generalization ability of the proposed method. We further evaluate it on another real-world dataset, the Stanford Cars. Table 2 presents the performance of our method with the state-of-arts. Generally, the results are consistent with those of the previous two datasets. Again, the proposed CIN can achieve the highest accuracy compared with the state-of-arts.

Combined with NTS. Furthermore, our module is general and flexible, and it can be readily integrated into other

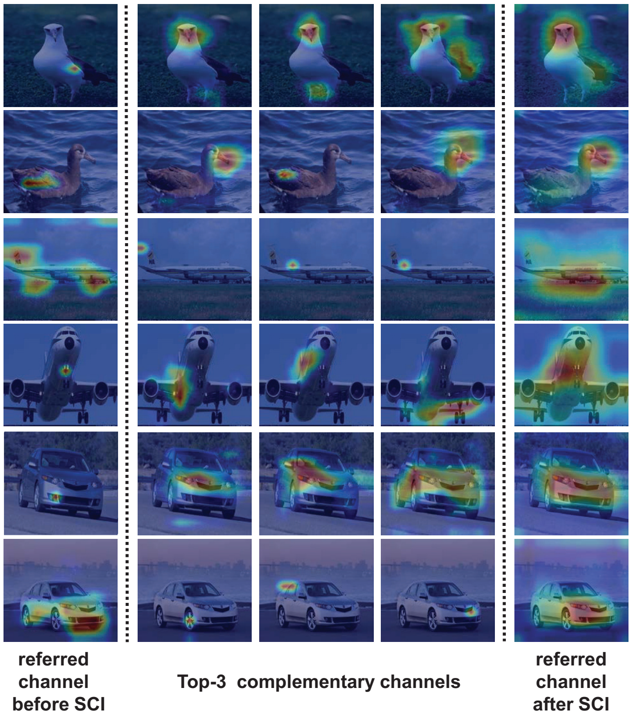
referred channel before SCI
Top-3 complementary channels
referred channel after SCI

Figure 4: Visualization of channel activations before and after SCI moudle on CUB, Cars and Aircraft.

framework to improve the performance. In this experiment, we combine our module with the latest state-of-the-art method NTS (Yang et al. 2018), which is a two-stage framework by leveraging a region proposal networks to localize discriminative parts with weakly-supervised learning. We integrate the SCI module at the end of the feature extractor networks. As NTS will discover multiple regions out of sequence, thus we only apply the CCI to the whole feature stream. Table 3 shows the performance of our method combined with NTS (NTS+CIN). As can be found, NTS+CIN achieves *consistent* performance improvements on all the three publicly available datasets compared with either NTS or CIN alone. The results further demonstrate the strong capability of our module. We expect that our CIN network can improve the performance on various computer vision tasks when simply plugged into existing framework.

Qualitative Visualization

To better understand the intra- and inter-image channel interactions modeled by CIN, we visualize the channel correlations and neural activations in our SCI and CCI module. Figure 4 shows the visualization of SCI module for images from three different datasets. Column 1 presents the activations of a randomly selected channel (assuming it is the i^{th} channel) before SCI. Column 2 to Column 4 are the three most complementary channels to it. In other words, these three channels correspond to the ones that have the largest values in the i^{th} row of SCI matrix W defined in Equation 7. The last column represents Y_i, which is the i^{th} channel after SCI. We find that, for a referred channel, the top-3 complementary channels tend to capture different semantic. For instance, in the first example of Figure 4, the referred channel has a strong activation around wings, and its complementary channels focus more on head and tail regions. As a result, the attention feature channels are enhanced by this complementary information and activates also on other discriminative parts. Note that after our SCI module, the activations span most of the object parts, which indicates that SCI effectively models the interactions among different channels, and combine their complementary but discriminative parts to produce more informative features.

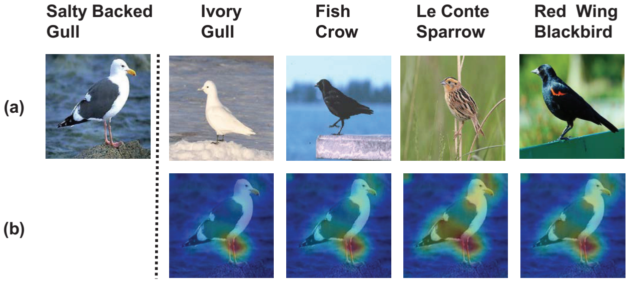

Figure 5: Visualization on the results of CCI module on CUB. (a) the original images; (b) the feature maps by CCI. It can be seen that different regions are highlighted conditioned on different image pairs.

Figure 5 visualizes the results of our CCI module on CUB-200-2011 dataset. Line 2 shows the contrastive attention activations by averaging all feature maps after CCI across channels. “Salty Black Gull” and “Ivory Cull” have similar heads, and their features after CCI have weaker responses to the head. While comparing with “Fish Crow”, the activations near the head becomes stronger. For the other two bird species, their appearance differences are huge and the CCI module provides strong responses to the whole body part. This result suggests that our proposal CCI module can focus on the key distinctions by modeling the interactions of channels between image pairs.

Conclusion

We have presented a new channel interaction network (CIN) for fine-grained image categorization. Our network first learns complementary channel information by a self-channel interaction (SCI) module taking the relationships between channels into account. It encourages to pull positive pairs closer while pushing negative pairs away via a contrastive channel interaction (CCI) module, which exploits channel correlations between samples. The proposed network can be trained end-to-end in one stage requiring no bounding box/part annotations. Extensive experiments demonstrate that CIN can achieve superior performance compared to the state-of-the-art approaches.

References

Chen, et al. 2018. Fine-grained representation learning and recognition by exploiting hierarchical semantic embedding. *arXiv* *preprint arXiv:1808.04505*.

Chen, W.; Chen, X.; Zhang, J.; and Huang, K. 2017. Beyond triplet loss: A deep quadruplet network for person re-identification. In *CVPR*.

Cui, Y.; Zhou, F.; Lin, Y.; and Belongie, S. J. 2016. Fine-grained categorization and dataset bootstrapping using deep metric learning with humans in the loop. In *CVPR*.

Cui, Y.; Zhou, F.; Wang, J.; Liu, X.; Lin, Y.; and Belongie, S. J. 2017. Kernel pooling for convolutional neural networks. In *CVPR*.

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*.

Gao, Y.; Beijbom, O.; Zhang, N.; and Darrell, T. 2016. Compact bilinear pooling. In *CVPR*.

Ge, and Yu. 2017. Borrowing treasures from the wealthy: Deep transfer learning through selective joint fine-tuning. In *CVPR*.

Hadsell, R.; Chopra, S.; and LeCun, Y. 2006. Dimensionality reduction by learning an invariant mapping. In *CVPR*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*.

Hu, J.; Lu, J.; and Tan, Y. 2014. Discriminative deep metric learning for face verification in the wild. In *CVPR*.

Hu, J.; Shen, L.; and Sun, G. 2017. Squeeze-and-excitation networks. *arXiv preprint arXiv:1709.01507*.

Jaderberg, M.; Simonyan, K.; Zisserman, A.; and Kavukcuoglu, K. 2015. Spatial transformer networks. In *NIPS*.

Karlinsky, L.; Shtok, J.; Tzur, Y.; and Tzadok, A. 2017. Fine-grained recognition of thousands of object categories with single-example training. In *CVPR*.

Kong, S., and Fowlkes, C. C. 2017. Low-rank bilinear pooling for fine-grained classification. In *CVPR*.

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

Lam, M.; Mahasseni, B.; and Todorovic, S. 2017. Fine-grained recognition as hsnet search for informative image parts. In *CVPR*.

Li, Z.; Yang, Y.; Liu, X.; Zhou, F.; Wen, S.; and Xu, W. 2017. Dynamic computational time for visual attention. *arXiv preprint arXiv:1703.10332*.

Li, P.; Xie, J.; Wang, Q.; and Gao, Z. 2018. Towards faster training of global covariance pooling networks by iterative matrix square root normalization. In *CVPR*.

Lin, T.-Y.; RoyChowdhury, A.; and Maji, S. 2015. Bilinear cnn models for fine-grained visual recognition. In *ICCV*.

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

Moghimi, M.; Belongie, S. J.; Saberian, M. J.; Yang, J.; Vasconcelos, N.; and Li, L. 2016. Boosted convolutional neural networks. In *BMVC*.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. Imagenet large scale visual recognition challenge. *IJCV*.

Schroff, F.; Kalenichenko, D.; and Philbin, J. 2015. Facenet: A unified embedding for face recognition and clustering. In *CVPR*.

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

Sochor, J.; Herout, A.; and Havel, J. 2016. Boxcars: 3d boxes as CNN input for improved fine-grained vehicle recognition. In *CVPR*.

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

Taigman, Y.; Yang, M.; Ranzato, M.; and Wolf, L. 2014. Deepface: Closing the gap to human-level performance in face verification. In *CVPR*.

Van Horn, G.; Branson, S.; Farrell, R.; Haber, S.; Barry, J.; Ipeirotis, P.; Perona, P.; and Belongie, S. Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. In *CVPR*.

Varior, R. R.; Haloi, M.; and Wang, G. 2016. Gated siamese convolutional neural network architecture for human re-identification. In *ECCV*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *NIPS*.

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

Wang, J.; Song, Y.; Leung, T.; Rosenberg, C.; Wang, J.; Philbin, J.; Chen, B.; and Wu, Y. 2014. Learning fine-grained image similarity with deep ranking. In *CVPR*.

Wang, X.; Girshick, R. B.; Gupta, A.; and He, K. 2017. Non-local neural networks. *arXiv preprint arXiv:1711.07971*.

Wang, Q.; Li, P.; and Zhang, L. 2017. G2denet: Global gaussian distribution embedding network and its application to visual recognition. In *CVPR*.

Wang, Y.; Morariu, V. I.; and Davis, L. S. 2018. Learning a discriminative filter bank within a cnn for fine-grained recognition. In *CVPR*.

Wei, X.-S.; Xie, C.-W.; Wu, J.; and Shen, C. 2018. Mask-cnn: Localizing parts and selecting descriptors for fine-grained bird species categorization. *Pattern Recognition*.

Yang, Z.; Luo, T.; Wang, D.; Hu, Z.; Gao, J.; and Wang, L. 2018. Learning to navigate for fine-grained classification. *arXiv preprint arXiv:1809.00287*.

Yosinski, J.; Clune, J.; Nguyen, A.; Fuchs, T.; and Lipson, H. 2015. Understanding neural networks through deep visualization. *arXiv preprint arXiv:1506.06579*.

Yu, C.; Zhao, X.; Zheng, Q.; Zhang, P.; and You, X. 2018. Hierarchical bilinear pooling for fine-grained visual recognition. In *ECCV*.

Yue, K.; Sun, M.; Yuan, Y.; Zhou, F.; Ding, E.; and Xu, F. 2018. Compact generalized non-local network. In *NeurIPS*.

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

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

Zheng, H.; Fu, J.; Zha, Z.; and Luo, J. 2019. Looking for the devil in the details: Learning trilinear attention sampling network for fine-grained image recognition. In *CVPR*.