Learning Multi-Attention Convolutional Neural Network for Fine-Grained Image Recognition

Heliang Zheng[1]*, Jianlong Fu[2], Tao Mei[2], Jiebo Luo[3]
[1]University of Science and Technology of China, Hefei, China
[2]Microsoft Research, Beijing, China
[3]University of Rochester, Rochester, NY
[1]zhenghl@mail.ustc.edu.cn, [2]{jianf, tmei}@microsoft.com, [3]jluo@cs.rochester.edu

Abstract

Recognizing fine-grained categories (e.g., bird species) highly relies on discriminative part localization and part-based fine-grained feature learning. Existing approaches predominantly solve these challenges independently, while neglecting the fact that part localization (e.g., head of a bird) and fine-grained feature learning (e.g., head shape) are mutually correlated. In this paper, we propose a novel part learning approach by a multi-attention convolutional neural network (MA-CNN), where part generation and feature learning can reinforce each other. MA-CNN consists of convolution, channel grouping and part classification sub-networks. The channel grouping network takes as input feature channels from convolutional layers, and generates multiple parts by clustering, weighting and pooling from spatially-correlated channels. The part classification network further classifies an image by each individual part, through which more discriminative fine-grained features can be learned. Two losses are proposed to guide the multi-task learning of channel grouping and part classification, which encourages MA-CNN to generate more discriminative parts from feature channels and learn better fine-grained features from parts in a mutual reinforced way. MA-CNN does not need bounding box/part annotation and can be trained end-to-end. We incorporate the learned parts from MA-CNN with part-CNN for recognition, and show the best performances on three challenging published fine-grained datasets, e.g., CUB-Birds, FGVC-Aircraft and Stanford-Cars.

1. Introduction

Recognizing fine-grained categories (e.g., bird species [1, 35], flower types [21, 24], car models [14, 17], etc.) by computer vision techniques has attracted extensive attention. This task is very challenging, as fine-grained image recognition should be capable of localizing and representing the marginal visual differences within subordinate categories (e.g., the two species of Waxwing in Figure 1). A large corpus of works [9, 33, 34] solve this problem by relying on human-annotated bounding box/part annotations (e.g., head, body for birds) for part-based feature representations. However, the heavy human involvement makes part definition and annotation expensive and subjective, which are not optimal for all fine-grained recognition tasks [3, 36].

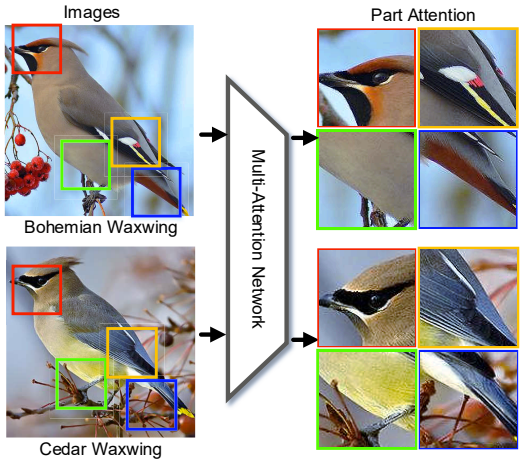

Figure 1: The ideal discriminative parts with four different colors for the two bird species of "waxwing." We can observe the subtle visual differences from multiple attended parts, which can distinguish the birds, e.g., the red head/wing/tail, and white belly for the top bird, compared with the bottom ones. [Best viewed in color]

Significant progresses have been made by learning weakly-supervised part models by convolutional neural networks (CNNs) [2, 4, 15] with category labels, which have no dependencies on bounding box/part annotations and thus

*This work was performed when Heliang Zheng was visiting Microsoft Research as a research intern.

can greatly increase the usability and scalability of fine-grained recognition [25, 31, 35]. The framework are typically composed of two independent steps: 1) part localization by training from positive/nagetive image patches [35] or pinpointing from pre-trained feature channels [25], and 2) fine-grained feature learning by selective pooling [31] or dense encoding from feature maps [17]. Although promising results have been reported, the performance for both part localization and feature learning are heavily restricted by the discrimination ability of the category-level CNN without explicit part constrains. Besides, we discover that part localization and fine-grained feature learning are mutually correlated and thus can reinforce each other. For example in Figure 1, an initial head localization can promote learning specific patterns around heads, which in return helps to pinpoint the accurate head.

To deal with the above challenges, we propose a novel part learning approach by multi-attention convolutional neural network (MA-CNN) for fine-grained recognition without bounding box/part annotations. MA-CNN jointly learns part proposals and the feature representations on each part. Unlike semantic parts defined by human [9, 33, 34], the parts here are defined as multiple attention areas with strong discrimination ability in an image. MA-CNN consists of convolution, channel grouping, and part classification sub-networks, which takes as input full images and generates multiple part proposals.

First, a convolutional feature channel often corresponds to a certain type of visual pattern [25, 35]. The channel grouping sub-network thereby clusters and weights spatially-correlated patterns into part attention maps from channels whose peak responses appear in neighboring locations. The diversified high-response locations further constitute multiple part attention maps, from which we extract multiple part proposals by cropping with fixed size. Second, once the part proposals are obtained, the part classification network further classifies an image by part-based features, which are spatially pooled from full convolutional feature maps. Such a design can particularly optimize a group of feature channels which are correlated to a certain part by removing the dependence on other parts, and thus better fine-grained features on this part can be learned. Third, two optimization loss functions are jointly enforced to guide the multi-task learning of channel grouping and part classification, which motivates MA-CNN to generate more discriminative parts from feature channels and learn more fine-grained features from parts in a mutual reinforced way. Specifically, we propose a channel grouping loss function to optimize the channel grouping sub-network, which considers channel clusters of high intra-class similarity and inter-class separability over spatial regions as part attention, and thus can produce compact and diverse part proposals.

Once parts have been localized, we amplify each attended part from an image and feed it into part-CNNs pipeline [1], where each part-CNN is learned to categories by using corresponding part as input. To further leverage the power of part ensemble, features from multiple parts are deeply fused to classify an image by learning a fully-connected fusion layer. To the best of our knowledge, this work represents the first attempt for learning multiple part models by jointly optimizing channel combination and feature representation. Our contributions can be summarized as follows:

- We address the challenges of weakly-supervised part model learning by proposing a novel multi-attention convolutional neural network, which jointly learns feature channel combination as part models and fine-grained feature representation.
- We propose a channel grouping loss for compact and diverse part learning which minimizes the loss function by applying geometry constraints over part attention maps, and use category labels to enhance part discrimination ability.
- We conduct comprehensive experiments on three challenging datasets (CUB Birds, FGVC-Aircraft, Stanford Cars), and achieve superior performance over the state-of-the-art approaches on all these datasets.

The rest of the paper is organized as follows. Section 2 describes the related work. Section 3 introduces the proposed method. Section 4 provides the evaluation and analysis, followed by the conclusion in Section 5.

2. Related Work

The research on fine-grained image recognition can be generally classified into two dimensions, i.e., fine-grained feature learning and discriminative part localization.

2.1. Fine-grained Feature Learning

Learning representative features has been extensively studied for fine-grained image recognition. Due to the great success of deep learning, most of the recognition frameworks depend on the powerful convolutional deep features [15], which have shown significant improvement than handcrafted features on both general [8] and fine-grained categories. To better model subtle visual differences for fine-grained recognition, a bilinear structure [17] is recently proposed to compute the pairwise feature interactions by two independent CNNs, which has achieved the state-of-the-art results in bird classification [30]. Besides, some methods (e.g., [35]) propose to unify CNN with spatially-weighted representation by Fisher Vector [23], which show superior results on both bird [30] and dog datasets [12]. Making the use of the ability of boosting to combine the strengths of multiple learners can also improve the classification accuracy [20], achieving the state-of-the-art performance.

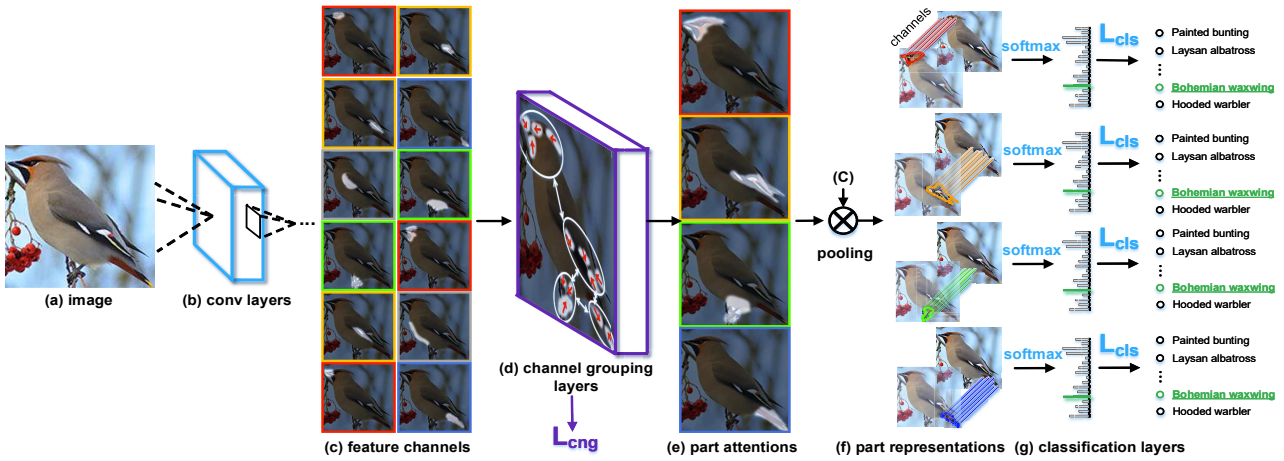

Figure 2: The framework of multi-attention convolutional neural network (MA-CNN). The network takes as input an image in (a), and produces part attentions in (e) from feature channels (e.g., 512 in VGG [26]) in (c). Different network modules for classification with light blue (i.e., the convolution in (b) and softmax in (g)), and part localization with purple (i.e., the channel grouping in (d)) are iteratively optimized by classification loss L_{cls} over part-based representations in (f), and by channel grouping loss L_{cng}, respectively. The softmax in (g) includes both a fully-connected layer, and a softmax function, which matches to category entries. [Best viewed in color]

2.2. Discriminative Part Localization

A large amount of works propose to leverage the extra annotations of bounding boxes and parts to localize significant regions in fine-grained recognition [9, 16, 22, 30, 33, 34]. However, the heavy involvement of human efforts make this task not practical for large-scale real problems. Recently, there have been numerous emerging research working for a more general scenario and proposing to use unsupervised approach to learn part attention models. A visual attention-based approach proposes a two-level domainnet on both objects and parts, where the part templates are obtained by clustering scheme from the internal hidden representations in CNN [31]. Picking deep filter responses [35] and multi-grained descriptors [27] propose to learn a set of part detectors by analyzing filter responses from CNN that respond to specific patterns consistently in an unsupervised way. Spatial transformer [10] takes one step further and proposes a dynamic mechanism that can actively spatially transform an image for more accurate classification. The most relevant works to ours come from [25, 31, 35], which learn candidate part models from convolutional channel responses. Compared with them, the advantages of our work are two folds. First, we propose to learn parts generation from a group of spatial-correlated convolutional channels, instead of independent channels which often lack strong discrimination power. Second, the fine-grained feature learning on parts and part localization are conducted in a mutual reinforced way, which ensures multiple representative parts can be accurately inferred from the consistently optimized feature maps.

3. Approach

Traditional part-based frameworks take no advantage of the deeply trained networks to mutually promote the learning for both part localization and feature representation. In this paper, we propose a multi-attention convolutional neural network (MA-CNN) for part model learning, where the computation of part attentions is nearly cost-free and can be trained end-to-end.

We design the network with convolution, channel grouping and part classification sub-networks in Figure 2. First, the whole network takes as input full-size image in Figure 2 (a), which is fed into convolutional layers in Figure 2 (b) to extract region-based feature representation. Second, the network proceeds to generate multiple part attention maps in Figure 2 (e) via channel grouping and weighting layers in Figure 2 (d), followed by a sigmoid function to produce probabilities. The resultant part representations are generated by pooling from region-based feature representations with spatial attention mechanism, which is shown in Figure 2 (f). Third, a group of probability scores over each part to fine-grained categories are predicted by fully-connected and softmax layers in Figure 2 (g). The proposed MA-CNN is optimized to convergence by alternatively learning a softmax classification loss over each part representation and a channel grouping loss over each part attention map.

3.1. Multi-Attention CNN for Part Localization

Given an input image $\mathbf{X}$, we first extract region-based deep features by feeding the images into pre-trained convolutional layers. The extracted deep representations are

denoted as $\mathbf{W} * \mathbf{X}$, where $*$ denotes a set of operations of convolution, pooling and activation, and $\mathbf{W}$ denotes the overall parameters. The dimension of this representation is $w \times h \times c$, where w, h, c indicate width, height and the number of feature channels. Although a convolutional feature channel can correspond to a certain type visual pattern (e.g., stripe) [25, 35], it is usually difficult to express rich part information by a single channel. Therefore, we propose a channel grouping and weighting sub-network to cluster spatially-correlated subtle patterns as compact and discriminative parts from a group of channels whose peak responses appear in neighboring locations.

Intuitively, each feature channel can be represented as a position vector whose elements are the coordinates from the peak responses over all training image instances, which is given by:

$$[t_x^1,\ t_y^1,\ t_x^2,\ t_y^2,\ ...\ t_x^\Omega,\ t_y^\Omega], \tag{1}$$

where t_x^i, t_y^i are the coordinates of the peak response of the i^{th} image in training set, and Ω is the number of training images. We consider the position vector as features, and cluster different channels into N groups as N part detectors. The resultant i^{th} group is represented by an indicator function over all feature channels, which is given by:

$$[\mathbb{1}\{1\}, ..., \mathbb{1}\{j\}, ..., \mathbb{1}\{c\}], \tag{2}$$

where $\mathbb{1}\{\cdot\}$ equals one if the j^{th} channel belongs to the i^{th} cluster and zero otherwise.

To ensure the channel grouping operation can be optimized in training, we approximate this grouping by proposing channel grouping layers to regress the permutation over channels by fully-connected (FC) layers. To generate N parts, we define a group of FC layers $F(\cdot) = [f_1(\cdot), ..., f_N(\cdot)]$. Each $f_i(\cdot)$ takes as input convolutional features, and produce a weight vector $\mathbf{d_i}$ over different channels (from 1 to c), which is given by:

$$\mathbf{d_i}(\mathbf{X}) = f_i(\mathbf{W} * \mathbf{X}), \tag{3}$$

where $\mathbf{d_i}(\mathbf{X}) = [d_1, ..., d_c]$. We omit subscript i for each d_c for simplicity. We obtain the channel grouping result $\mathbf{d_i}(\mathbf{X})$ by two steps: 1) pre-training FC parameters in Eqn. (3) by fitting $\mathbf{d_i}(\mathbf{X})$ to Eqn. (2), 2) further optimizing by end-to-end part learning. Hence, Eqn. (2) is the supervision of Eqn. (3) in step (1), which ensures a reasonable model initialization (for FC parameters). Note that we enforce each channel to belong to only one cluster by a loss function which will be presented later. Based on the learned weights over feature channels, we further obtain the part attention map for the i^{th} part as follows:

$$\mathbf{M_i}(\mathbf{X}) = \mathrm{sigmoid}(\sum_{\mathrm{j}=1}^{c} \mathrm{d_j}[\mathbf{W} * \mathbf{X}]_\mathrm{j}), \tag{4}$$

where $[\cdot]_j$ denotes the j^{th} feature channel of convolutional features $\mathbf{W} * \mathbf{X}$. The operation between d_j and $[\cdot]_j$ denotes multiplication between a scalar and a matrix. The resultant $\mathbf{M_i}(\mathbf{X})$ is further normalized by the sum of each element, which indicates one part attention map. Later we denote $\mathbf{M_i}(\mathbf{X})$ as $\mathbf{M_i}$ for simplicity. Furthermore, the final convolutional feature representation for the i^{th} part is calculated via spatial pooling on each channel, which is given by:

$$\mathbf{P_i}(\mathbf{X}) = \sum_{j=1}^{c} ([\mathbf{W} * \mathbf{X}]_j \cdot \mathbf{M_i}), \tag{5}$$

where the dot product denotes element-wise multiplication between $[\mathbf{W} * \mathbf{X}]_j$ and $\mathbf{M_i}$.

3.2. Multi-task Formulation

Loss function: The proposed MA-CNN is optimized by two types of supervision, i.e., part classification loss and channel grouping loss. Specifically, we formulate the objective function as a multi-task optimization problem. The loss function for an image $\mathbf{X}$ is defined as follows:

$$L(\mathbf{X}) = \sum_{i=1}^{N} [L_{cls}(\mathbf{Y}^{(i)}, \mathbf{Y}^*)] + L_{cng}(\mathbf{M_1}, ..., \mathbf{M_N}), \tag{6}$$

where L_{cls} and L_{cng} represents the classification loss on each of the N parts, and the channel grouping loss, respectively. $\mathbf{Y}^{(i)}$ denotes the predicted label vector from the i^{th} part by using part-based feature $\mathbf{P_i}(\mathbf{X})$, and $\mathbf{Y}^*$ is the ground truth label vector. The training is implemented by fitting category labels via a softmax function.

Although strong discrimination is indispensable for localizing a part, rich information from multiple part proposals can further benefit robust recognition with stronger generalization ability, especially for cases with large pose variance and occlusion. Therefore, the channel grouping loss for compact and diverse part learning is proposed, which is given by:

$$L_{cng}(\mathbf{M_i}) = Dis(\mathbf{M_i}) + \lambda Div(\mathbf{M_i}), \tag{7}$$

where $Dis(\cdot)$ and $Div(\cdot)$ is a distance and diversity function with the weight of λ. $Dis(\cdot)$ encourages a compact distribution, and the concrete form is designed as follows:

$$Dis(\mathbf{M_i}) = \sum_{(x,y)\in\mathbf{M_i}} m_i(x,y)[||x - t_x||^2 + ||y - t_y||^2], \tag{8}$$

where $m_i(x, y)$ takes as input the coordinates (x, y) from $\mathbf{M_i}$, and produces the amplitudes of responses. $Div(\cdot)$ is designed to favor a diverse attention distribution from different part attention maps, i.e., $\mathbf{M_1}$ to $\mathbf{M_N}$. The concrete form is formulated as follows:

$$Div(\mathbf{M_i}) = \sum_{(x,y)\in\mathbf{M_i}} m_i(x,y)[\max_{k\neq i} m_k(x,y) - mrg], \tag{9}$$

Figure 3: An illustration of the part attention learning. The top row indicates two initial part attention areas around "head" and "wing," as well as the optimization direction for each position. "+, -, ·" indicates "strengthen, weaken, unchange," respectively. The optimized part attentions are shown in the bottom. Detailed analysis can be found in Sec. 3.2.

where i, k indicates the index of different part attention maps. "mrg" represents a margin, which makes the loss less sensitive to noises, and thus enables robustness. The advantages for such a loss are two-fold. The first encourages similar visual patterns from a specific part to be grouped together, and thus strong part detector can be learned, while the second encourages attention diversity for different parts. Such a design with geometry constraints can enable the network to capture the most discriminative part (e.g., heads for birds), and accomplish robust recognition from diversified parts (e.g., wings and tails) if heads are occluded.

Alternative optimization: To optimize the part localization and feature learning in a mutually reinforced way, we take the following alternative training strategy. First, we fix the parameters from convolutional layers, and optimize the channel grouping layers in (d) in Figure 2 by L_{cng} in Eqn. (6) to converge for part discovery. Second, we fix the channel grouping layer, and switch to optimize the convolutional layers in (b) and softmax in (g) in Figure 2 by L_{cls} in Eqn. (6) for fine-grained feature learning. This learning is iterative, until the two types of losses no longer change.

Since the impact of L_{cls} can be intuitively understood, we illustrate the mechanism of the distance loss $Dis(\cdot)$ and the diversity loss $Div(\cdot)$ in L_{cng} by showing the derivatives on the learned part attention maps $\mathbf{M_i}$. The part attention maps in an iteration over head and wing for a bird are shown in the top-row in Figure 3, with the brighter the area, the higher the responses for attention. Besides, we visualize the derivatives for each position from the part attention map, which shows the optimization direction. The yellow "+" shows the areas which needs to be strengthen, and the blue "-" shows the region which needs to be weaken, and the grey "·" shows unchange. Based on the optimization on each position, the background area and the overlap between the two attention maps change to be smaller in the next iteration (shown in the bottom in Figure 3), which benefits from the first and second term in Eqn. (7), respectively.

Table 1: The statistics of fine-grained datasets in this paper.

Datasets	# Category	# Training	# Testing
CUB-200-2011 [30]	200	5,994	5,794
FGVC-Aircraft [19]	100	6,667	3,333
Stanford Cars [13]	196	8,144	8,041

3.3. Joint Part-based Feature Representation

Although the proposed MA-CNN can help detect parts by simultaneously learning part localization and fine-grained part features, it is still difficult to represent the subtle visual differences existed in local regions due to their small sizes. Since previous research (e.g., [5, 18, 34]) has observed the benefits by region zooming, in this section, we follow the same strategy.

In particular, an image $\mathbf{X}$ (e.g., 448×448 pixels) is first fed into MA-CNN, which generates N parts by cropping a square from $\mathbf{X}$, with the point which corresponds to the peak from each $\mathbf{M_i}$ as the center, and the 96×96 area as part bounding box. Each cropped region are amplified into a larger resolution (e.g., 224×224) and taken as input by part-CNNs, of which each part-CNN is learned to classify an part (e.g. head for a bird) into image-level categories. To extract both local and global features from an image, we follow previous works [1, 18, 33] to take as input for Part-CNN from both part-level patches and object-level images. Thus we can obtain joint part-based feature representations for each image:

$$\{P_1,\ P_2,\ ...\ P_N,\ P_O\}, \tag{10}$$

where P_i denotes the extracted part description by part-CNN, and N is total number of parts; P_O denotes the feature extracted from object-level images. To further leverage the benefit of part feature ensemble, we concatenate them together into a fully-connected fusion layer with softmax function for the final classification.

4. Experiment

4.1. Datasets and Baselines

Datasets: We conduct experiments on three challenging datasets, including Caltech-UCSD Birds (CUB-200-2011) [30], FGVC-Aircraft [19] and Stanford Cars [13], which are widely-used to evaluate fine-grained image recognition.

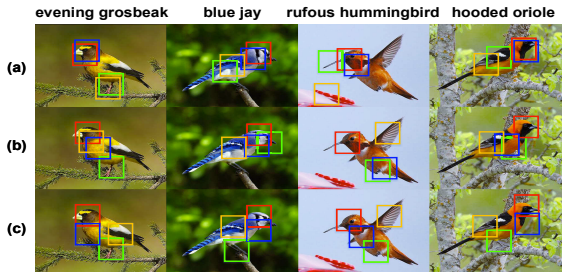

Figure 4: Four bird examples of the visualized part localization results by (a) initial parts by channel clustering, (b) optimizing channel grouping loss L_{cng}, and (c) joint learning $L_{cng} + L_{cls}$.

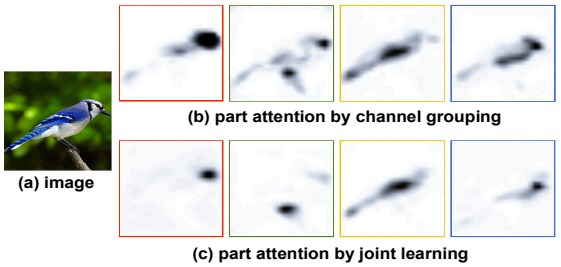

Figure 5: An example of comparison of four part attention maps for an image in (a) by optimizing channel grouping loss L_{cng} in (b), and joint learning $L_{cng} + L_{cls}$ in (c).

The detailed statistics with category numbers and data splits are summarized in Table 1.

Baselines: We divide compared approaches into two categories, based on whether they use human-defined bounding boxes (bbox) or part annotations. We don't compare with the methods which depend on part annotations in testing, since they are not fully-automatic. In the following, the first five methods use human supervision, and the latter eight are based on unsupervised part learning methods. We compare with them, due to their state-of-the-art results. All the baselines are listed as following:

- **PN-CNN** [1]: **p**ose **n**ormalized CNN proposes to compute local features by estimating the object's pose.
- **Part-RCNN** [34]: extends **R-CNN** [6] based framework by part annotations.
- **Mask-CNN** [29]: localizing parts and selecting descriptors by learning masks.
- **MDTP** [28]: **m**ining **d**iscriminative **t**riplets of **p**atches for as the proposed parts.
- **PA-CNN** [14]: **p**art **a**lignment-based method generates parts by using co-segmentation and alignment.
- **PDFR** [35]: **p**icking **d**eep **f**ilter **r**esponses proposes to find distinctive filters and learn part detectors.
- **FV-CNN** [7]: extracting **f**isher **v**ector features for fine-grained recognition.
- **MG-CNN** [27]: **m**ultiple **g**ranularity descriptors learn multi-region of interests for all the grain levels.
- **ST-CNN** [10]: **s**patial **t**ransformer network learns invariance to scale, warping by feature transforming.
- **TLAN** [31]: **t**wo-**l**evel **a**ttention **n**etwork proposes domain-nets on both objects and parts to classification.
- **FCAN** [18]: **f**ully **c**onvolutional **a**ttention **n**etwork adaptively selects multiple task-driven visual attention by reinforcement learning.
- **B-CNN** [17]: **b**ilinear-CNN proposes to capture pairwise feature interactions for classification.
- **RA-CNN** [5]: **r**ecurrent **a**ttention CNN framework which can locate the discriminative area recurrently for better classification performance.

4.2. Implementation Details

To make fair comparison, we conduct experiments as the same settings as baselines. Specifically, we use the same VGG-19 [26] model pre-trained on ImageNet for both part localization in MA-CNN with 448×448 inputs, and classification in Part-CNN with 224×224 inputs , where the larger resolution inputs in MA-CNN can benefit discriminative part learning. The output of each Part-CNN is extracted by Global Average Pooling (GAP) [37] from the last convolutional layer (i.e., conv5_4 in VGG-19) to generate the 512-dim features for classification. For FGVC-Aircraft and Stanford Cars datasets, P_O in Eqn. (10) is the original image; for CUB-200-2011, we also use the cropped high-convolutional response area (e.g., with a threshold of one tenth of the highest response value) as object-level representation. The λ in Eqn. (7) and mrg in Eqn. (9) are empirically set to 2 and 0.02, which are robust to optimization. The concrete form of channel grouping layers is constructed by two fully-connected layers with tanh activation. We run experiment using Caffe [11], and will release the full model in the near future.

4.3. Experiment on CUB-200-2011

Part localization results: We compare part localization results under different settings by the proposed MA-CNN network for qualitative analysis. The settings include part localization by 1) clustering with Eqn. (1) and Eqn. (2), 2) optimizing parts only by channel grouping L_{cng}, and 3) joint learning by both channel grouping L_{cng} and part classification L_{cls}. We set the part numbers N as 2, 4, 6, and take four parts as an example to show the learned part localization results in Figure 4.

We can observe from Figure 4 (a) that although the initialized red part detector can localize heads for the four birds, other part detectors are not always discriminative. For example, the green detector locates inconsistent parts for the four birds, including feet, tails, beaks and wings, respectively, which shows the inferior part learning results. Besides, multiple part detectors (e.g, the red and blue) attend

Table 2: Comparison of part localization in terms of classification accuracy on CUB-200-2011 dataset. Detailed Analysis can be found in Sec. 4.3.

Approach	Accuracy
MA-CNN (inital)	82.0
MA-CNN (L_{cng})	85.3
MA-CNN ($L_{cls} + L_{cng}$)	86.5

on the same regions, which are difficult to capture the diverse feature representations from different part locations. Although more diverse parts can be generated by introducing the channel grouping loss in Figure 4 (b), the learned part detectors are still not robust to distinguish some similar parts. For example, it is difficult for the green one to discriminate the thin beak and feet for "blue jay" and "evening grosbeak." Further improvement is limited by the feature representations from the pre-trained convolutional layers, which are obtained by regressing global bird features to category labels, and the fine-grained representation on a specific part is hard to be learned. The proposed MA-CNN adopts an alternative strategy for learning both part localizations and fine-grained features on a specific part, and thus we can obtain consistent prediction on four parts, where red, blue, yellow, and green detectors locate head, breast, wing and feet, respectively. To better show the feature learning results, we show the part attention maps, which are generated by feature channel grouping over the 512 channels from VGG-19. We can observe from Figure 5 that the attention maps by joint learning tend to focus on one specific part (e.g., the feet in the green part in Figure 5 (c)). However, the green attention map learned without feature learning in Figure 5 (b) generate multiple peak responses over both feet and beak, which reflects the inadequate capability to distinguish the two body areas from birds.

We further conduct quantitative comparison on part localization in terms of classification accuracy. All compared methods use VGG-19 model for part-based classification, but with different part localization settings. We can see from Table 2 that significant improvements (4.0% relative increase) in the second row are made by the proposed channel grouping network with loss L_{cng}, compared with the results from parts which are obtained by initial channel clustering in the first row. The performance can be improved from joint learning in the third row, by further locating more fine-grained parts (e.g., feet), with the relative accuracy gain of 1.4% compared with the second row.

Fine-grained image recognition: We compare with two types of baselines based on whether they use human-defined bounding box (bbox)/part annotation. Mask-CNN [29] uses the supervision with both human-defined bounding box and ground truth parts. B-CNN [17] uses bounding box with very high-dimensional feature representations (250k dimensions). We first generate two parts (i.e., around heads and wings, as shown in the red and yellow squares in Figure 4) with the same number of parts as Mask-CNN [29]. As shown in Table 3, the proposed MA-CNN (2 parts + object) can achieve comparable results with Mask-CNN [29] and B-CNN [17], even without bbox and part annotations, which demonstrates the effectiveness. By incorporating with four parts, we can achieve even better results than Mask-CNN [29]. Compared with RA-CNN [5], we can obtain comparable result by MA-CNN (2 parts + object) and a relative accuracy gain with 1.4% by MA-CNN (4 parts + object), which shows the power of multi-attention. We even surpass B-CNN (without Train Anno.) [17] and ST-CNN [10], which uses either high-dimensional features or stronger inception network as baseline model with nearly both 2.9% relative accuracy gains. Note that MA-CNN (4 parts + object) outperforms MA-CNN (2 parts + object) with a clear margin (1.3% relative gains), but the performance saturates after extending MA-CNN to six parts. The reason is mainly derived from the facts that MA-CNN (2 parts + object) captures the parts around heads and wings, and MA-CNN (4 parts + object) further locates around feet and breasts. Therefore, it is difficult for MA-CNN (6 parts + object) to learn more discriminative parts from birds and the recognition accuracy saturates.

Table 3: Comparison results on CUB-200-2011 dataset. Train Anno. represents using bounding box or part annotation in training.

Approach	Train Anno.	Accuracy
PN-CNN(AlexNet) [1]	✓	75.7
Part-RCNN(AlexNet) [34]	✓	76.4
PA-CNN [14]	✓	82.8
MG-CNN [27]	✓	83.0
FCAN [18]	✓	84.3
B-CNN (250k-dims) [17]	✓	85.1
Mask-CNN [29]	✓	85.4
TLAN(AlexNet) [31]		77.9
MG-CNN [27]		81.7
FCAN [18]		82.0
B-CNN (250k-dims) [17]		84.1
ST-CNN (Inception net) [10]		84.1
PDFR [35]		84.5
RA-CNN [5]		85.3
MA-CNN (2 parts + object)		85.4
MA-CNN (4 parts + object)		**86.5**

4.4. Experiment on FGVC-Aircraft

Since the images of aircrafts have clear spatial structures, we can obtain good part localization result by the proposed MA-CNN network with four part proposals, which are shown in Figure 6 (c). The classification results on FGVC-Aircraft dataset are further summarized in Table 4. The proposed MA-CNN (4 parts + object) outperforms the high-dimensional B-CNN [17] with a clear margin (6.9%

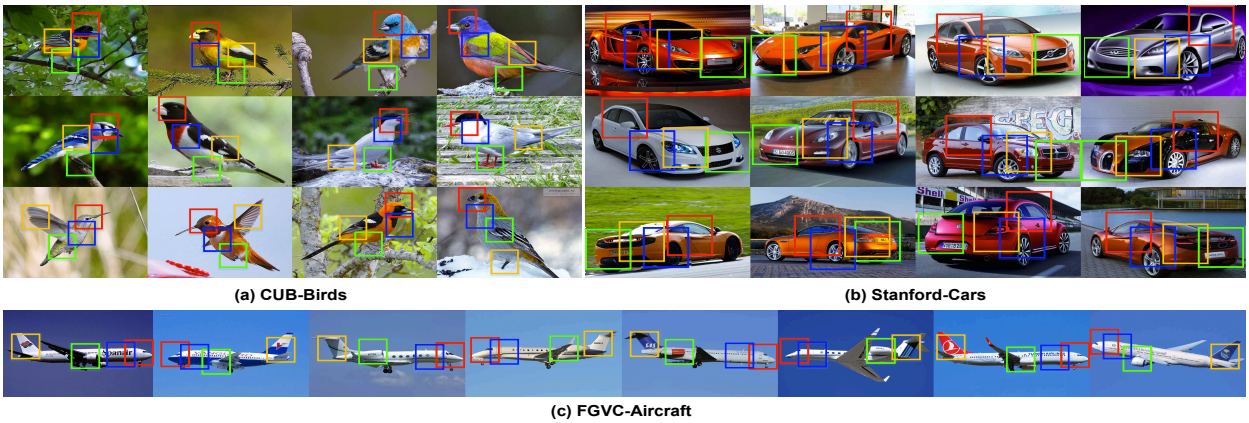

Figure 6: Part localization results for individual examples from (a) CUB-Birds, (b) Stanford-Cars, and (c) FGVC-Aircraft. The four parts on each dataset show consistent part attention areas for a specific fine-grained category, which are discriminative to classify this category from other.

Table 4: Comparison results on FGVC-Aircraft dataset. Train Anno. represents using bounding box in training.

Approach	Train Anno.	Accuracy
MG-CNN [27]	✓	86.6
MDTP [28]	✓	88.4
FV-CNN [7]		81.5
B-CNN (250k-dims)[17]		84.1
RA-CNN [5]		88.2
MA-CNN (2 parts + object)		88.4
MA-CNN (4 parts + object)		**89.9**

Table 5: Comparison results on Stanford Cars dataset. Train Anno. represents using bounding box in training.

Approach	Train Anno.	Accuracy
R-CNN [6]	✓	88.4
FCAN [18]	✓	91.3
MDTP [28]	✓	92.5
PA-CNN [14]	✓	92.8
FCAN [18]		89.1
B-CNN (250k-dims) [17]		91.3
RA-CNN [5]		92.5
MA-CNN (2 parts + object)		91.7
MA-CNN (4 parts + object)		**92.8**

relative gains), which shows the effectiveness of multiple part proposes. MDTP [28] also proposes to detect parts by bounding box annotation and geometric constraints. However, they don't make full use of convolutional networks to refine the features for localization. Compared with MDTP [28], the 1.7% relative gain from MA-CNN (4 parts + object) further shows the important role for joint learning of features and parts. Compared with RA-CNN [5], MA-CNN (2 parts + object) gets the comparable result and MA-CNN (4 parts + object) achieves 1.8% relative accuracy gain. A similar performance saturation is observed by using six parts on FGVC-Aircraft dataset.

4.5. Experiment on Stanford Cars

The classification results on Stanford Cars are summarized in Table 5. Car part detection can significantly improve the performance due to the discrimination and complementarity from different car parts [32]. For example, some car models can be easily identified from headlights or air intakes in the front. We can observe from Figure 6 (b) that the four parts learned from cars are consistent with human perception, which include the front/back view, side view, car lights, and wheels. Due to the accurate part localization, MA-CNN (4 parts + object) can achieve a relative accuracy gain of 4.2%, compared with FCAN [18] under the same experiment setting. This result from our unsupervised part model is even comparable with PA-CNN [14], which uses bounding boxes. We can observe the marginal improvement compared with RA-CNN [5], because the multiple attention areas (e.g., the front view and the car lights) locate close enough, which have been attended by RA-CNN as a whole part.

5. Conclusions

In this paper, we propose a multiple attention convolutional neural network for fine-grained recognition, which jointly learns discriminative part localization and fine-grained feature representation. The proposed network does not need bounding box/part annotations for training and can be trained end-to-end. Extensive experiments demonstrate the superior performance on both multiple-part localization and fine-grained recognition on birds, aircrafts and cars. In the future, we will conduct the research on two directions. First, how to integrate the structural and appearance models from parts for better recognition performance. Second, how to capture smaller parts (e.g., eyes of a bird) to represent the more subtle differences between fine-grained categories by unsupervised part learning approaches.

References

[1] S. Branson, G. V. Horn, S. J. Belongie, and P. Perona. Bird species categorization using pose normalized deep convolutional nets. In *BMVC*, 2014.

[2] J. Fu, T. Mei, K. Yang, H. Lu, and Y. Rui. Tagging personal photos with transfer deep learning. In *WWW*, pages 344–354, 2015.

[3] J. Fu, J. Wang, Y. Rui, X.-J. Wang, T. Mei, and H. Lu. Image tag refinement with view-dependent concept representations. *IEEE T-CSVT*, 25(28):1409–1422, 2015.

[4] J. Fu, Y. Wu, T. Mei, J. Wang, H. Lu, and Y. Rui. Relaxing from vocabulary: Robust weakly-supervised deep learning for vocabulary-free image tagging. In *ICCV*, 2015.

[5] J. Fu, H. Zheng, and T. Mei. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In *CVPR*, pages 4438–4446, 2017.

[6] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, pages 580–587, 2014.

[7] P.-H. Gosselin, N. Murray, H. Jégou, and F. Perronnin. Revisiting the fisher vector for fine-grained classification. *Pattern Recognition Letters*, 49:92–98, 2014.

[8] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[9] S. Huang, Z. Xu, D. Tao, and Y. Zhang. Part-stacked cnn for fine-grained visual categorization. In *CVPR*, pages 1173–1182, 2016.

[10] M. Jaderberg, K. Simonyan, A. Zisserman, and k. kavukcuoglu. Spatial transformer networks. In *NIPS*, pages 2017–2025, 2015.

[11] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[12] A. Khosla, N. Jayadevaprakash, B. Yao, and L. Fei-Fei. Novel dataset for fine-grained image categorization. In *ICCV Workshop*, 2011.

[13] J. Krause, M. Stark, J. Deng, and L. Fei-Fei. 3D object representations for fine-grained categorization. In *ICCV Workshop*, 2013.

[14] J. Krause1, H. Jin, J. Yang, and F. Li. Fine-grained recognition without part annotations. In *CVPR*, pages 5546–5555, 2015.

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1106–1114, 2012.

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

[17] T.-Y. Lin, A. RoyChowdhury, and S. Maji. Bilinear cnn models for fine-grained visual recognition. In *ICCV*, pages 1449–1457, 2015.

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

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

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

[21] M.-E. Nilsback and A. Zisserman. A visual vocabulary for flower classification. In *CVPR*, pages 1447–1454, 2006.

[22] O.M.Parkhi, A.Vedaldi, C.Jawajar, and A.Zisserman. The truth about cats and dogs. In *ICCV*, pages 1427–1434, 2011.

[23] F. Perronnin and D. Larlus. Fisher vectors meet neural networks: A hybrid classification architecture. In *CVPR*, pages 3743–3752, 2015.

[24] S. E. Reed, Z. Akata, B. Schiele, and H. Lee. Learning deep representations of fine-grained visual descriptions. In *CVPR*, 2016.

[25] M. Simon and E. Rodner. Neural activation constellations: Unsupervised part model discovery with convolutional networks. In *ICCV*, pages 1143–1151, 2015.

[26] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, pages 1409–1556, 2015.

[27] D. Wang, Z. Shen, J. Shao, W. Zhang, X. Xue, and Z. Zhang. Multiple granularity descriptors for fine-grained categorization. In *ICCV*, pages 2399–2406, 2015.

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

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

[30] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology, 2010.

[31] T. Xiao, Y. Xu, K. Yang, J. Zhang, Y. Peng, and Z. Zhang. The application of two-level attention models in deep convolutional neural network for fine-grained image classification. In *CVPR*, pages 842–850, 2015.

[32] L. Yang, P. Luo, C. Change Loy, and X. Tang. A large-scale car dataset for fine-grained categorization and verification. In *CVPR*, June 2015.

[33] H. Zhang, T. Xu, M. Elhoseiny, X. Huang, S. Zhang, A. Elgammal, and D. Metaxas. SPDA-CNN: Unifying semantic part detection and abstraction for fine-grained recognition. In *CVPR*, pages 1143–1152, 2016.

[34] N. Zhang, J. Donahue, R. B. Girshick, and T. Darrell. Part-based R-CNNs for fine-grained category detection. In *ECCV*, pages 1173–1182, 2014.

[35] X. Zhang, H. Xiong, W. Zhou, W. Lin, and Q. Tian. Picking deep filter responses for fine-grained image recognition. In *CVPR*, pages 1134–1142, 2016.

[36] B. Zhao, X. Wu, J. Feng, Q. Peng, and S. Yan. Diversified visual attention networks for fine-grained object classification. *CoRR*, abs/1606.08572, 2016.

[37] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. In *CVPR*, pages 2921–2929. IEEE, 2016.