This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the version available on IEEE Xplore.

Look Closer to See Better: Recurrent Attention Convolutional Neural Network for Fine-grained Image Recognition

Jianlong Fu[1], Heliang Zheng[2], Tao Mei[1]
[1]Microsoft Research, Beijing, China
[2]University of Science and Technology of China, Hefei, China
[1]{jianf, tmei}@microsoft.com, [2]zhenghl@mail.ustc.edu.cn

Abstract

Recognizing fine-grained categories (e.g., bird species) is difficult due to the challenges of discriminative region localization and fine-grained feature learning. Existing approaches predominantly solve these challenges independently, while neglecting the fact that region detection and fine-grained feature learning are mutually correlated and thus can reinforce each other. In this paper, we propose a novel recurrent attention convolutional neural network (RA-CNN) which recursively learns discriminative region attention and region-based feature representation at multiple scales in a mutually reinforced way. The learning at each scale consists of a classification sub-network and an attention proposal sub-network (APN). The APN starts from full images, and iteratively generates region attention from coarse to fine by taking previous predictions as a reference, while a finer scale network takes as input an amplified attended region from previous scales in a recurrent way. The proposed RA-CNN is optimized by an intra-scale classification loss and an inter-scale ranking loss, to mutually learn accurate region attention and fine-grained representation. RA-CNN does not need bounding box/part annotations and can be trained end-to-end. We conduct comprehensive experiments and show that RA-CNN achieves the best performance in three fine-grained tasks, with relative accuracy gains of 3.3%*,* 3.7%*,* 3.8%*, on CUB Birds, Stanford Dogs and Stanford Cars, respectively.*

1. Introduction

Recognizing fine-grained categories by computer vision techniques (e.g., classifying bird species [2, 34], flower types [21, 24], car models [14, 19], etc.) has attracted extensive attention. The task is very challenging as some fine-grained categories (e.g., "eared grebe" and "horned grebe") can only be recognized by domain experts. Different from general recognition, the fine-grained image recognition should be capable of localizing and representing the very marginal visual differences within subordinate categories, and thus can benefit a wide variety of applications, e.g., expert-level image recognition [15, 31], rich image captioning [1, 12], and so on.

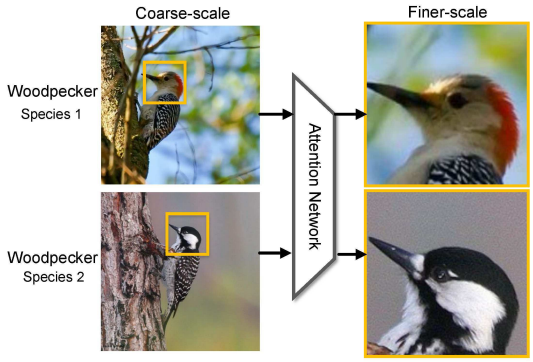

Figure 1. Two bird species of woodpecker. We can observe the very subtle visual differences from highly local regions (e.g., heads in yellow boxes), which are difficult to learn from the original image scale. However, the difference can be more vivid and significant if we can learn to zoom into the attended regions at a finer scale. [Best viewed in color]

The challenges of fine-grained recognition are mainly two-fold: discriminative region localization and fine-grained feature learning from those regions. Previous research has made impressive progresses by introducing part-based recognition frameworks, which typically consist of two steps: 1) identifying possible object regions by analyzing convolutional responses from neural networks in an unsupervised fashion or by using supervised bounding box/part annotations, and 2) extracting discriminative features from each region and encoding them into compact vectors for recognition. Although promising results have been reported, further improvement suffers from the following limitations. First, human-defined regions or the regions learned by existing unsupervised methods may not be optimal for machine classification [35]. Second, subtle visual differences existed in local regions from similar fine-

grained categories are still difficult to learn. We found that region detection and fine-grained feature learning are mutually correlated and thus can reinforce each other. As shown in Figure 1, accurate head localization can promote learning discriminative head features, which further help to pinpoint the different colors existed in afterbrain.

To deal with the above challenges, we propose a novel recurrent attention convolutional neural network (RA-CNN) for fine-grained recognition without bounding box/part annotations. RA-CNN recursively learns discriminative region attention and region-based feature representation in a mutually reinforced manner. The proposed RA-CNN is a stacked network which takes the input from full images to fine-grained local regions at multiple scales. First, the multi-scale networks share the same network architecture yet with different parameters at each scale to fit the inputs with different resolutions (e.g., the coarse scale and finer scale in Figure 1). The learning at each scale consists of a classification sub-network and an attention proposal sub-network (APN), which can ensure adequate discrimination ability at each scale and generate an accurate attended region for the next finer scale. Second, a finer-scale network dedicated to high-resolution regions takes as input an amplified attended region for extracting more fine-grained features. Third, the recurrent network is alternatively optimized by an intra-scale softmax loss for classification and an inter-scale pairwise ranking loss for attention proposal network. The ranking loss optimizes the finer network to generate higher confidence scores on correct categories than previous prediction.

Since finer-scale networks can be stacked in a recurrent way, RA-CNN can gradually attend on the most discriminative regions from coarse to fine (e.g., from body to head, then to beak for birds). Note that the accurate region localization can help discriminative region-based feature learning, and vice versa. Thus the proposed network can benefit from the mutual reinforcement between region localization and feature learning. To further leverage the advantages of ensemble learning, features from multiple scales 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 of proposing a multi-scale recurrent attention network for fine-grained recognition. Our contributions can be summarized as follows:

- We address the challenges of fine-grained recognition by proposing a novel recurrent attention convolutional neural network architecture that simultaneously enables the accurate detection of discriminative region and the effective learning of region-based representation in a mutually reinforced way.
- We propose a pairwise ranking loss to optimize the attention proposal network. Compared with region localizers with only label supervision, such a design enables network to gradually attend on more fine-grained regions with the reference of previous scales.
- We conduct comprehensive experiments on three challenging datasets (CUB Birds, Stanford Dogs, Stanford Cars), and achieve superior performance over the state-of-the-art approaches on all of these datasets.

The rest of the paper is organized as follows. Section 2 reviews 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 proceeds along two dimensions, i.e., discriminative feature learning and sophisticated part localization.

2.1. Discriminative Feature Learning

Learning discriminative features is crucial for fine-grained image recognition. Due to the success of deep learning, most of the methods depend on the powerful convolutional deep features, which have shown significant improvement than hand-crafted features on both general and fine-grained categories [4, 5, 6, 17, 29]. To learn stronger feature representation, deep residual network [9] scales up CNN to 152 layers by optimizing residual functions, which reduces the error rate to 3.75% on ImageNet test set [17]. To better model subtle differences existed in fine-grained categories, a bilinear structure [19] is recently proposed to compute the pairwise feature interactions by two independent CNNs to capture the image local differences, which has achieved the state-of-the-art results in bird classification [30]. Besides, another method [34] proposes to unify CNN with spatially weighted representation by Fisher Vector [23], which shows superior results on both bird [30] and dog datasets [13].

2.2. Sophisticated Part Localization

Previous works mainly focus on leveraging the extra annotations of bounding box and part annotations to localize significant regions in fine-grained recognition [10, 18, 22, 30, 32, 33]. However, the heavy involvement of manual annotations make this task not practical for large-scale real problems. Recently, there have been emerging works aiming at a more general scenario and proposing to use unsupervised approach to mine region attention. A visual attention-based approach proposes a two-level domain-net on both objects and parts, where the part templates are learned by clustering scheme from the internal hidden representations in CNN [31]. Picking deep filter responses [34] and multi-grained descriptors [28] 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 [11] takes one step further and proposes a

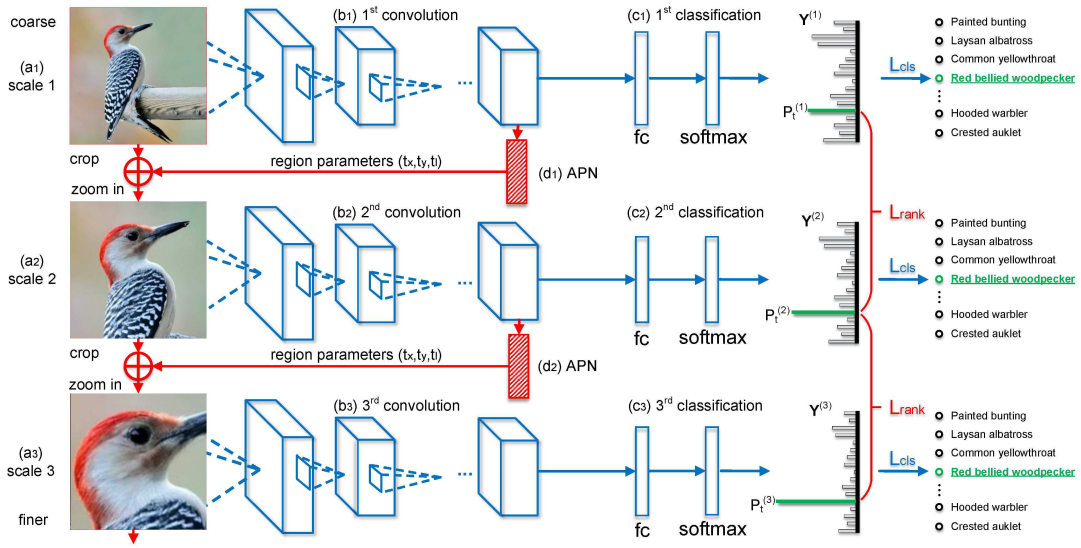

Figure 2. The framework of recurrent attention convolutional neural network (RA-CNN). The inputs are from coarse full-size images to finer region attention (from top to bottom). Different network modules for classification (marked in blue) and attention proposal (marked in red) are alternatively optimized by classification losses L_{cls} between label prediction $\mathbf{Y}^{(\mathbf{s})}$ and ground truth $\mathbf{Y}^*$ at each scale, and pairwise ranking losses L_{rank} between $p_t^{(s)}$ and $p_t^{(s+1)}$ from neighboring scales, where $p_t^{(s)}$ and $p_t^{(s+1)}$ denote the probabilities on the correct category, and s denotes the scale. APN is the attention proposal network, fc represents fully-connected layer, softmax layer matches to category entries by a fc layer, followed by a softmax operation. $\oplus$ represents "crop" and "zoom in" operation. [Best viewed in color]

dynamic mechanism that can actively spatially transform an image for more accurate classification. Whereas, it is still difficult for existing models to exactly localize subtle regions due to their small sizes. The most relevant works to ours come from [20] and [35]. Both of them propose to zoom in on discriminative local regions to improve the performance of fine-grained recognition. However, the learning of region localizers from [20] and [35] relies on either pre-processed region proposals or category labels, which poses challenges to accurate region localization.

3. Approach

In this section, we will introduce the proposed recurrent attention convolutional neural network (RA-CNN) for fine-grained image recognition. We consider the network with three scales as an example in Figure 2, and more finer scales can be stacked in a similar way. The inputs are recurrent from full-size images in a_1 to fine-grained discriminative regions in a_2 and a_3, where a_2 and a_3 takes the input as the attended regions from a_1 and a_2, respectively. First, images at different scales are fed into convolutional layers (b_1 to b_3) to extract region-based feature representation. Second, networks proceed to predict both a probability score by fully-connected and softmax layers (c_1 to c_3) and a region attention by an attention proposal network (d_1, d_2). The proposed RA-CNN is optimized to convergence by alternatively learning a softmax classification loss at each scale and a pairwise ranking loss across neighboring scales.

3.1. Attention Proposal Network

Multi-task formulation: Traditional part-based framework on fine-grained recognition takes no advantages of the deeply trained networks to mutually promote the learning for both localization and recognition. Inspired by the recent success of region proposal network (RPN) [8], in this paper, we propose an attention proposal network (APN) where the computation of region attention is nearly cost-free, and the APN can be trained end-to-end.

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 $\mathbf{W}_c * \mathbf{X}$, where $*$ denotes a set of operations of convolution, pooling and activation, and $\mathbf{W}_c$ denotes the overall parameters. We further model the network at each scale as a multi-task formulation with two outputs. The first task is designed to generate a probability distribution $\mathbf{p}$ over fine-grained categories, shown as:

$$\mathbf{p}(\mathbf{X}) = f(\mathbf{W}_c * \mathbf{X}), \tag{1}$$

where $f(\cdot)$ represents fully-connected layers to map convolutional features to a feature vector that could be matched with the category entries, as well as includes a softmax layer to further transform the feature vector to probabilities. The second task is proposed to predict a set of box coordinates of an attended region for the next finer scale. By approximating the attended region as a square with three

parameters, the representation is given by:

$$[t_x, t_y, t_l] = g(\mathbf{W}_c * \mathbf{X}), \tag{2}$$

where t_x, t_y denotes the square's center coordinates in terms of x and y axis, respectively, and t_l denotes the half of the square's side length. The specific form of $g(\cdot)$ can be represented by two-stacked fully-connected layers with three outputs which are the parameters of the attended regions. Note that compared with region proposal network in object detection which uses strong supervision of ground truth boxes, the learning of the proposed APN is trained in a weakly-supervised fashion, since the part-level annotation is often hard to obtain. The specific learning process and loss functions will be introduced in Sec. 3.2.

Attention localization and amplification: Once the location of an attended region is hypothesized, we crop and zoom in the attended region to finer scale with higher resolution to extract more fine-grained features. To ensure the APN can be optimized in training, we approximate the cropping operation by proposing a variant of two-dimension boxcar function as an attention mask. The mask can select the most significant regions in forward-propagation, and is readily to be optimized in backward-propagation due to the properties of continuous functions.

Assume the top-left corner in original images as the origin of a pixel coordinate system, whose x-axis and y-axis is defined from left-to-right and top-to-bottom, respectively. We can adopt the parameterizations of the top-left (denoted as "tl") and bottom-right (denoted as "br") points from the attended region as following:

$$\begin{aligned} t_{x(tl)} &= t_x - t_l, \quad t_{y(tl)} = t_y - t_l, \\ t_{x(br)} &= t_x + t_l, \quad t_{y(br)} = t_y + t_l. \end{aligned} \tag{3}$$

Based on the above representations, the cropping operation can be implemented by an element-wise multiplication between the original image at coarser scales and an attention mask, which can be computed as:

$$\mathbf{X}^{att} = \mathbf{X} \odot \mathbf{M}(t_x, t_y, t_l), \tag{4}$$

where $\odot$ represents element-wise multiplication, $\mathbf{X}^{att}$ denotes the cropped attended region and $\mathbf{M}(\cdot)$ acts as an attention mask, with the specific form as:

$$\begin{aligned} \mathbf{M}(\cdot) = &[h(x - t_{x(tl)}) - h(x - t_{x(br)})] \\ &\cdot [h(y - t_{y(tl)}) - h(y - t_{y(br)})], \end{aligned} \tag{5}$$

and $h(\cdot)$ is a logistic function with index k:

$$h(x) = 1/\{1 + \exp^{-kx}\}. \tag{6}$$

Theoretically, when k is large enough, the logistic function can be considered as a step function and the two-dimensional boxcar function (i.e., $\mathbf{M}(\cdot)$) is zero over the entire real line along x and y dimensions, except for a single area (i.e., x ranges from $t_{x(tl)}$ to $t_{x(br)}$, and y ranges from $t_{y(tl)}$ to $t_{y(br)}$) where it is equal to one. The advantages for using the boxcar function are two folds. First, boxcar function can well-approximate the cropping operation to select the most significant regions predicted from coarser-scale networks. Second, boxcar function builds analytical representations between the attended region and box coordinates $\{t_x, t_y, t_l\}$, which is necessary when optimizing box parameters in backward-propagation.

Although attended regions have been localized, effective feature representation are sometimes still difficult to be extracted from the highly-localized regions. Therefore, we further amplify the region to a larger size by adaptively zooming. Specifically, we use a bilinear interpolation to compute the amplified output $\mathbf{X}^{amp}$ from the nearest four inputs in $\mathbf{X}^{att}$ by a linear map, which is given by:

$$\mathbf{X}^{amp}_{(i,j)} = \sum_{\alpha,\beta=0}^{1} |1 - \alpha - \{i/\lambda\}||1 - \beta - \{j/\lambda\}|\mathbf{X}^{att}_{(m,n)}, \tag{7}$$

where $m = [i/\lambda] + \alpha$, $n = [j/\lambda] + \beta$, λ is upsampling factor, which equals the value of enlarged size divided by t_l. $[\cdot]$ and $\{\cdot\}$ is the integral and fractional part, respectively.

3.2. Classification and Ranking

The proposed recurrent attention CNN is optimized by two types of supervision, i.e., intra-scale classification loss and inter-scale pairwise ranking loss, for alternatively generating accurate region attention and learning more fine-grained features. Specifically, we minimize an objective function following a multi-task loss. The loss function for an image sample is defined as:

$$L(\mathbf{X}) = \sum_{s=1}^{3}\{L_{cls}(\mathbf{Y}^{(s)}, \mathbf{Y}^*)\} + \sum_{s=1}^{2}\{L_{rank}(p_t^{(s)}, p_t^{(s+1)})\}, \tag{8}$$

where s denotes each scale, $\mathbf{Y}^{(s)}$ and $\mathbf{Y}^*$ denotes the predicted label vector from a specific scale and the ground truth label vector, respectively. L_{cls} represents classification loss, which predominantly optimizes the parameters of convolution and classification layers in Figure 2 (b_1 to b_3 and c_1 to c_3) for ensuring adequate discrimination ability at each scale. The training is implemented by fitting category labels on overall training samples via a softmax function. Besides, $p_t^{(s)}$ from pairwise ranking loss L_{rank} denotes the prediction probability on the correct category labels t. Specifically, the ranking loss is given by:

$$L_{rank}(p_t^{(s)}, p_t^{(s+1)}) = max\{0, p_t^{(s)} - p_t^{(s+1)} + margin\}, \tag{9}$$

which enforces $p_t^{(s+1)} > p_t^{(s)} + margin$ in training. Such a design can enable networks to take the prediction from

coarse scales as references, and gradually approach the most discriminative region by enforcing the finer-scale network to generate more confident predictions. Note that L_{cls} and L_{rank} take effect alternatively for different optimization purposes, and details can be found in Sec. 3.4.

3.3. Multi-scale Joint Representation

Once the proposed RA-CNN has been trained at each scale, we can obtain multi-scale representations from full-size images to multiple coarse-to-fine region attention. In particular, the image $\mathbf{X}$ can be represented by a set of multiple-scale descriptors:

$$\{F_1,\ F_2,\ ...\ F_N\}, \tag{10}$$

where F_i denotes the feature descriptor at a specific scale generated from the fully-connected layers in classification net (c_1 to c_3 in Figure 2), and N is total number of scales. To leverage the benefit of feature ensemble, we first normalize each descriptor independently, and concatenate them together into a fully-connected fusion layer with softmax function for the final classification. The application of softmax function instead of Support Vector Machine (SVM) [3] is mainly for the technical consistency for feature extraction and classification, so that we can integrate the multi-scale descriptors and classification end-to-end in testing. Besides, we have verified that softmax and linear SVM can produce comparable results for classification.

3.4. Implementation Details

Training strategy: To better optimize attention localization and fine-grained classification in a mutually reinforced way, we take the following alternative training strategy.

Step 1: we initialize convolutional/classification layers in Figure 2 (b_1 to b_3 and c_1 to c_3) by the same pre-trained VGG network [27] from ImageNet.

Step 2: we consider a square (represented by t_x, t_y, t_l) with the half length of the side of original image. The square is selected by searching regions in the original image, with the highest response value in the last convolutional layer (i.e., conv5_4 in VGG-19). We can further obtain a smaller square by analyzing convolutional responses at the second scale in a similar way. These selected squares are used to pre-train APN to obtain parameters in Figure 2 (d_1), (d_2) by learning the transformation from convolutional feature maps to $\{t_x, t_y, t_l\}$.

Step 3: we optimize the parameters in the above two steps in an alternative way. Specifically, we keep APN parameters unchanged, and optimize the softmax losses at three scales to converge. Then we fix parameters in convolutional/classification layers, and switch to ranking loss to optimize the two APNs. The learning process for the two parts is iterative, until the two types of losses no longer change. Besides, t_l at each scale is constrained to be no less than one-third of the previous t_l at coarse scale, to avoid the incompleteness of object structures when t_l is too small.

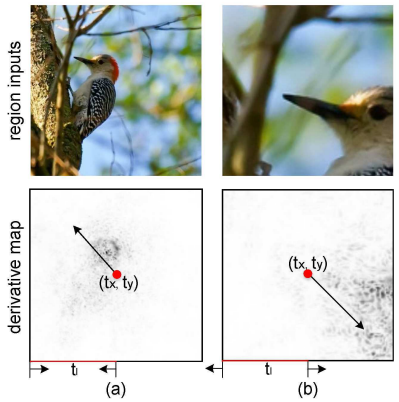

Figure 3. An illustration of region attention learning. The top-row indicates two exemplar region inputs at a specific scale and the bottom-row indicates the derivatives backpropagated into the input layer. The black arrows show the optimization direction of t_x, t_y and t_l, which are consistent with human perception. Detailed analysis can be found in Sec. 3.4.

Optimization for attention learning: We illustrate the mechanism of attention learning by calculating the derivatives on t_x, t_y, t_l, and show the effects to region cropping. Since the derivatives of the proposed ranking loss to t_x, t_y, t_l have similar forms, we take t_x as an example and calculate the derivative by the chain rule in backward-propagation [25], which is given by:

$$\frac{\partial L_{rank}}{\partial t_x} \propto \mathbf{D}_{top} \odot \frac{\partial \mathbf{M}(t_x, t_y, t_l)}{\partial t_x}, \tag{11}$$

where $\odot$ represents element-wise multiplication, and $\mathbf{D}_{top}$ denotes the derivatives backpropagated into the input layer at a specific scale, which represents the importance of each pixel with respect to the overall network activation [15]. We simplify the derivative forms in Eqn. (11) to $L'_{rank}(t_x)$ and $\mathbf{M}'(t_x)$. In a minimization problem, we have that if $L'_{rank}(t_x)<0$, then t_x increases, otherwise t_x decreases. We further follow [15] to compute the negative square of the norm of the derivatives for obtaining a consistent optimization direction with human perception. The derivative map is shown in the bottom-row in Figure 3, with the darker the point, the larger the absolute value. Each derivative map corresponds to an input in the top-row with the same dimension. Besides, $\mathbf{M}'(t_x)$ represents the derivative of mask function to t_x, which can be given by a piecewise function with qualitative evaluation as:

$$\mathbf{M}'(t_x) = \begin{cases} <0 & x \to t_{x(tl)} \\ >0 & x \to t_{x(br)} \\ =0 & otherwise, \end{cases} \tag{12}$$

where the symbol "$\to$" represents "approaching to" for x.

Similar form for the derivative to t_y is given by:

$$\mathbf{M}'(t_y) = \begin{cases} <0 & y \rightarrow t_{y(tl)} \\ >0 & y \rightarrow t_{y(br)} \\ =0 & otherwise. \end{cases} \quad (13)$$

As $M'(t_l)$ takes positive value on the border and negative inside, the derivative to t_l is given by:

$$\mathbf{M}'(t_l) = \begin{cases} >0 & x \rightarrow t_{x(tl)} \; or \; x \rightarrow t_{x(br)} \\ & or \; y \rightarrow t_{y(br)} \; or \; y \rightarrow t_{y(tl)} \\ <0 & otherwise. \end{cases} \quad (14)$$

Based on the above analysis, we can obtain that $L'_{rank}(t_x)$ is positive in Figure 3(a) because the black points with negative value in derivative maps are mainly distributed in the top-left and $\mathbf{M}'(t_x)$ also adopts negative value in the left. Similarly, we can obtain $L'_{rank}(t_y)$ is positive, because $\mathbf{M}'(t_y)$ adopts negative value in the top. As the derivative map is almost zero on the border and negative inside, $L'_{rank}(t_l)$ is positive. Thus t_x, t_y and t_l will change to be smaller in the next iteration, which is consistent with human perception. Optimization in Figure 3(b) can be obtained by a similar analysis.

4. Experiments

4.1. Datasets and Baselines

Datasets: We conduct experiments on three challenging fine-grained image recognition datasets, including Caltech-UCSD Birds (CUB-200-2011) [30], Stanford Dogs [13] and Stanford Cars [16]. 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 box (bbox) or part annotation. 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 in both categories. All the baselines are listed as follows:

- **DeepLAC** [18]: **d**eep **l**ocalization, **a**lignment and **c**lassification proposes to use a pose-aligned part image for classification.
- **SPDA-CNN** [32]: **s**emantic **p**art **d**etection and **a**bstraction proposes to generate part candidates and extract features by detection/classification networks.
- **Part-RCNN** [33]: extends **R-CNN** [7] based framework by part annotations.
- **PA-CNN** [14]: **p**art **a**lignment-based method generates parts by using co-segmentation and alignment.
- **PN-CNN** [2]: **p**ose **n**ormalized CNN proposes to compute local features by estimating the object's pose.
- **PDFR** [34]: **p**icking **d**eep **f**ilter **r**esponses proposes to find distinctive filters and learn part detectors.

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

Datasets	# Category	# Training	# Testing
CUB-200-2011 [30]	200	5,994	5,794
Stanford Dogs [13]	120	12,000	8,580
Stanford Cars [16]	196	8,144	8,041

- **MG-CNN** [28]: **m**ultiple **g**ranularity descriptors learn multi-region of interests for all the grain levels.
- **ST-CNN** [11]: **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.
- **DVAN** [35]: **di**verse **a**ttention **n**etwork attends object from coarse to fine by multiple region proposals.
- **FCAN** [20]: **f**ully **c**onvolutional **a**ttention **n**etwork adaptively selects multiple task-driven visual attention by reinforcement learning.
- **B-CNN** [19]: **b**ilinear-CNN proposes to capture pairwise feature interactions for classification.
- **NAC** [26]: **n**eural **a**ctivation **c**onstellations find parts by computing neural activation patterns.

Input images (at scale 1) and attended regions (at scale 2,3) are resized to 448×448 and 224×224 pixels respectively in training, due to the smaller object size in the coarse scale. We use VGG-19 [27] (pre-trained on ImageNet) for bird and car datasets, and VGG-16 for dogs as the same settings with baselines. We find that k in Eqn. (6) and the margin in Eqn. (9) are robust to optimization, thus we empirically set k as 10 and margin as 0.05. The model has been made publicly available at https://github.com/Jianlong-Fu/Recurrent-Attention-CNN.

4.2. Experiments on CUB-200-2011

Attention localization: We show the attended regions from multiple scales by the proposed attention proposal network for qualitative analysis. In Figure 4, we can observe that these localized regions at second and third scales are discriminative to corresponding categories, and are easier to be classified than the first scale. The results are consistent with human perception that it would be helpful to look closer for fine-grained categories.

Since the proposed APN is automatically learned by discovering the most discriminative regions to classification, instead of regressing human-defined bounding box, we conduct quantitative comparison on attention localization in terms of classification accuracy. For fair comparison, all compared methods use VGG-19 model, but with different attention localization algorithms. We take the second-scale network to produce our results (denoted as RA-CNN (scale 2)), as attended regions at this scale can preserve both global bird structure and local visual cues, as shown in Figure 4. First, we can observe compara-

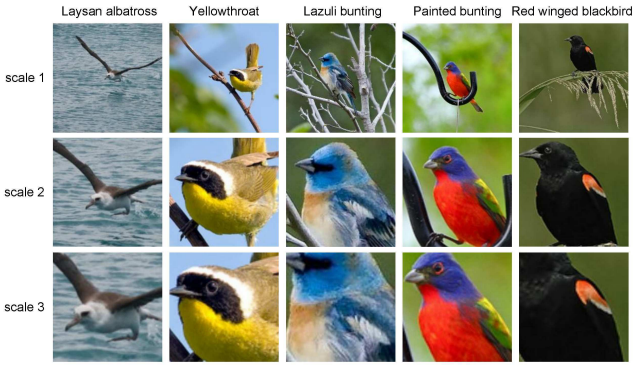

Figure 4. Five bird examples of the learned region attention at different scales. We can observe clear and significant visual cues for classification after gradually zooming in the attended regions.

Table 2. Comparison of attention localization in terms of classification accuracy on CUB-200-2011 dataset.

Approach	Accuracy
FCAN (single-attention) [20]	76.1
MG-CNN (single-granularity) [28]	79.5
RA-CNN (scale 2) w/ initial $\{t_x, t_y, t_l\}$	79.0
RA-CNN (scale 2)	**82.4**

ble results with the methods using human-defined bounding box in Table 3. PA-CNN [14] and MG-CNN (with anno.) [28] achieves 82.8% and 83.0% accuracy, respectively. RA-CNN (scale 2) achieves 82.4% accuracy. Second, we can achieve significant better results compared with existing unsupervised part learning-based methods. FCAN [20] and MG-CNN [28] are two relevant works to ours, which also use feature combination scheme from multiple scales/granularities. To make fair comparison, we select single-attention and single-granularity based performance from [20] and [28], and show the results in Table 2. We can obtain 8.3% and 3.6% relative improvement compared with FCAN (single-attention) [20] and MG-CNN (single-granularity) [28], which shows the superior attention learning ability of the proposed approach. Besides, the result of RA-CNN with initialized attended region and without ranking loss optimization is listed in the third row. From this result, we can know the key role of ranking loss for optimizing region attention.

Fine-grained image recognition: We compare with two types of baselines based on whether they use human-defined bounding box (bbox)/part annotations. PN-CNN [2] uses strong supervision of both human-defined bounding box and ground truth parts. B-CNN [19] uses bounding box with very high-dimensional feature representation (250k dimensions). As shown in Table 3, the proposed RA-CNN (scale 1+2+3) can achieve comparable results with PN-CNN [2] and B-CNN [19] even without bbox and part annotation, which demonstrates the effectiveness. Compared with unsupervised methods PDFR [34] without additional Fisher Vector learning, we can obtain a relative accuracy gain with 3.3% by our full model RA-CNN (scale 1+2+3). We even surpass B-CNN (w/o anno.) [19] and ST-CNN [11], which uses either high-dimensional features or stronger inception network as baseline model with nearly both 1.5% relative accuracy gains. Although FCAN (w/o anno.) [20] and DVAN [35] propose similar ideas to zoom into attended regions for classification, we can achieve better accuracy with 4.1% and 8.0% relative improvement because of the mutual reinforcement framework for attention localization and region-based feature learning. Note that RA-CNN (scale 2) outperforms VGG-19 results at scale 1 with clear margins (5.9% relative gains), which shows the necessity for "looking closer" on fine-grained categories. Besides, RA-CNN (scale 3) slightly drop than RA-CNN (scale 2), because of the missing of structural information existed in global bird images. By combining features at three scales via a fully-connected layer, we achieve the best 85.3% accuracy. Note that the superior result benefits from the complementary advantages from multiple scales. The combination of triple single-scale network with different initial parameters only achieves 78.0%, 83.5%, 82.0% for the first, second and third scale, respectively. Besides, we extend RA-CNN to more scales, but the performance saturates as discriminative information has been encoded into the previous scales.

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

Approach	Train Anno.	Accuracy
DeepLAC [34]	✓	80.3
Part-RCNN [33]	✓	81.6
PA-CNN [14]	✓	82.8
MG-CNN [28]	✓	83.0
FCAN [20]	✓	84.3
B-CNN (250k-dims) [19]	✓	85.1
SPDA-CNN [32]	✓	85.1
PN-CNN [2]	✓	85.4
VGG-19 [27]		77.8
TLAN [31]		77.9
DVAN [35]		79.0
NAC [26]		81.0
MG-CNN [28]		81.7
FCAN [20]		82.0
PDFR [34]		82.6
B-CNN (250k-dims) [19]		84.1
ST-CNN (Inception net) [11]		84.1
RA-CNN (scale 2)		82.4
RA-CNN (scale 3)		81.2
RA-CNN (scale 1+2)		84.7
RA-CNN (scale 1+2+3)		**85.3**

4.3. Experiments on Stanford Dogs

The classification accuracy on Stanford Dogs dataset are summarized in Table 4. The VGG-16 at the first scale takes the original images as input and achieves 76.7% recogni-

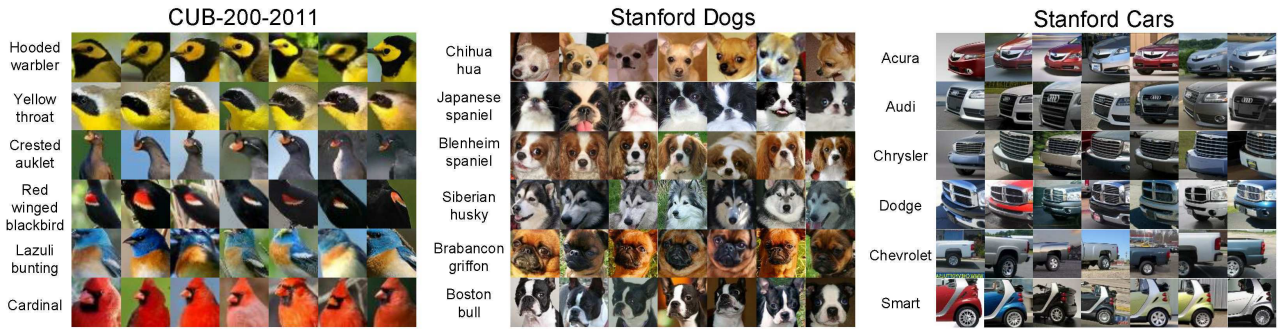

Figure 5. Attention localization at the third scale for birds, dogs and cars. The regions (in each row) learned from multiple image samples, represent consistent attention area for a specific fine-grained category, which are discriminative to classify this category from others.

Table 4. Comparison results on Stanford Dogs dataset without extra bounding box or part annotation.

Approach	Accuracy
NAC (AlexNet) [26]	68.6
PDFR (AlexNet) [34]	71.9
VGG-16 [27]	76.7
DVAN [35]	81.5
FCAN [20]	84.2
RA-CNN (scale 2)	85.9
RA-CNN (scale 3)	85.0
RA-CNN (scale 1+2)	86.7
RA-CNN (scale 1+2+3)	**87.3**

tion accuracy. Relying on accurate attention localization, RA-CNN (scale 2) achieves a significant improvement to recognition accuracy of 85.9%, with 12.0% relative gain. By combining the features from two scales and three scales, we can boost the performance to 86.7% and 87.3%, respectively. Comparing with the two most relevant approaches DVAN [35] and FCAN [20], the relative accuracy gains are 7.1% and 3.7%, respectively. This improvement mainly derives from the accurate attention localization, which are demonstrated in Figure 5. The figure proves that the attended regions are mostly located on dog heads, which are consistent with previous research [31, 35].

4.4. Experiments on Stanford Cars

The classification accuracy on Stanford Cars are summarized in Table 5. Different car models are difficult to be recognized, due to the subtle differences, e.g., different front and back design. Although VGG-19 at scale 1 only achieves 84.9% accuracy, the performance can increase to 90.0% after zooming in the discriminative region attention to finer scales. We obtain the highest recognition accuracy of 92.5% by leveraging the power of feature ensemble, which integrates features from original images, amplified whole vehicles and the front or back regions. We can analyze from Figure 5 that the proposed attention proposal network is capable of localizing the representative attended regions, such as the unique front design for Audi and

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

Approach	Train Anno.	Accuracy
R-CNN [7]	✓	88.4
FCAN [20]	✓	91.3
PA-CNN [14]	✓	92.8
VGG-19 [27]		84.9
DVAN [35]		87.1
FCAN [20]		89.1
B-CNN (250k-dims) [19]		91.3
RA-CNN (scale 2)		90.0
RA-CNN (scale 3)		89.2
RA-CNN (scale 1+2)		91.8
RA-CNN (scale 1+2+3)		**92.5**

Dodge, and the cute back design of Smart. Compared with the state-of-the-art methods, our full model RA-CNN (scale 1+2+3) surpasses DVAN [35] and FCAN (w/o anno.) [20] for large margins (6.2% and 3.8% relative gain) under the same settings. We also obtain better results than the high-dimensional B-CNN [19], and even achieve comparable performance with PA-CNN [14], which depends on human-defined bounding box.

5. Conclusion

In this paper, we propose a recurrent attention convolutional neural network for fine-grained recognition, which recursively learns discriminative region attention and region-based feature representation at multiple scales. 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 attention localization and fine-grained recognition tasks on birds, dogs and cars. In the future, we will conduct the research on two directions. First, how to simultaneously preserve global image structure and model local visual cues, to keep improving the performance at finer scales. Second, how to integrate multiple region attention to model more complex fine-grained categories.

References

[1] H. L. Anne, V. Subhashini, R. Marcus, M. Raymond, S. Kate, and T. Darrell. Deep compositional captioning: Describing novel object categories without paired training data. In *CVPR*, 2016.

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

[3] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

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

[5] 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.

[6] 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.

[7] 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.

[8] R. B. Girshick. Fast R-CNN. In *ICCV*, pages 1440–1448, 2015.

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

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

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

[12] J. Johnson, A. Karpathy, and F.-F. Li. Densecap: Fully convolutional localization networks for dense captioning. In *CVPR*, 2016.

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

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

[15] J. Krause, B. Sapp, A. Howard, H. Zhou, A. Toshev, T. Duerig, J. Philbin, and F.-F. Li. The unreasonable effectiveness of noisy data for fine-grained recognition. In *ECCV*, pages 301–316, 2016.

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

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

[18] 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.

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

[20] 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.

[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] D. Rumelhart, G. Hintont, and R. Williams. Learning representations by back-propagating errors. *Nature*, 323:533–536, 1986.

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

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

[28] 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.

[29] J. Wang, J. Fu, T. Mei, and Y. Xu. Beyond object recognition: Visual sentiment analysis with deep coupled adjective and noun neural networks. In *IJCAI*, 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] 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.

[33] 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.

[34] 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.

[35] 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.