

Attention Convolutional Binary Neural Tree for Fine-Grained Visual Categorization

Ruyi Ji^{1,2,*}, Longyin Wen^{3,*}, Libo Zhang^{1,†}, Dawei Du⁴,
YanJun Wu¹, Chen Zhao¹, Xianglong Liu⁵, Feiyue Huang⁶

¹State Key Laboratory of Computer Science, ISCAS, Beijing, China

²University of Chinese Academy of Sciences, Beijing, China

³JD Finance America Corporation, Mountain View, CA, USA

⁴University at Albany, State University of New York, Albany, NY, USA

⁵Beihang University, Beijing, China

⁶Tencent Youtu Lab, Beijing, China

{ruyi2017, libo, yanjun, zhaochen}@iscas.ac.cn,

longyin.wen@jd.com, cvdaviddo@gmail.com

Abstract

Fine-grained visual categorization (FGVC) is an important but challenging task due to high intra-class variances and low inter-class variances caused by deformation, occlusion, illumination, etc. An attention convolutional binary neural tree is presented to address those problems for weakly supervised FGVC. Specifically, we incorporate convolutional operations along edges of the tree structure, and use the routing functions in each node to determine the root-to-leaf computational paths within the tree. The final decision is computed as the summation of the predictions from leaf nodes. The deep convolutional operations learn to capture the representations of objects, and the tree structure characterizes the coarse-to-fine hierarchical feature learning process. In addition, we use the attention transformer module to enforce the network to capture discriminative features. Several experiments on the CUB-200-2011, Stanford Cars and Aircraft datasets demonstrate that our method performs favorably against the state-of-the-arts. Code can be found at <https://isrc.iscas.ac.cn/gitlab/research/acnet>.

*Both authors contributed equally to this work.

†Corresponding author: Libo Zhang (libo@iscas.ac.cn). This work was supported by the National Natural Science Foundation of China, Grant No. 61807033, the Key Research Program of Frontier Sciences, CAS, Grant No. ZDBS-LY-JSC038. Libo Zhang was supported by Youth Innovation Promotion Association, CAS (2020111), and Outstanding Youth Scientist Project of ISCAS.



Figure 1: Exemplars of fine-gained visual categorization. FGVC is challenging due to two reasons: (a) high intra-class variances: the birds belonging to the same category usually present significant different appearance, such as illumination variations (the 1st column), view-point changes (the 2nd column), clutter background (the 3rd column) and occlusion (the 4th column); (b) low inter-class variances: the birds in different columns belong to different categories, but share similar appearance in the same rows.

1. Introduction

Fine-Grained Visual Categorization (FGVC) aims to distinguish subordinate objects categories, such as different species of birds [42, 52], and flowers [1]. The high intra-class and low inter-class visual variances caused by deformation, occlusion, and illumination, make FGVC to be a highly challenging task.

Recently, the FGVC task is dominated by the convolutional neural network (CNN) due to its amazing classifica-

tion performance. Some methods [29, 26] focus on extracting discriminative subtle parts for accurate results. However, it is difficult for a single CNN model to describe the differences between subordinate classes (see Figure 1). In [34], the object-part attention model is proposed for FGVC, which uses both object and part attentions to exploit the subtle and local differences to distinguish subcategories. It demonstrates the effectiveness of using multiple deep models concentrating on different object regions in FGVC.

Inspired by [41], we design an attention convolutional binary neural tree architecture (ACNet) for weakly supervised FGVC. It incorporates convolutional operations along the edges of the tree structure, and use the routing functions in each node to determine the root-to-leaf computational paths within the tree as deep neural networks. This designed architecture makes our method inherits the representation learning ability of the deep convolutional model, and the coarse-to-fine hierarchical feature learning process. In this way, different branches in the tree structure focus on different local object regions for classification. The final decision is computed as the summation of the predictions from all leaf nodes. Meanwhile, we use the attention transformer to enforce the tree network to capture discriminative features for accurate results. The negative log-likelihood loss is adopted to train the entire network in an end-to-end fashion by stochastic gradient descent with back-propagation.

Notably, in contrast to the work in [41] adaptively growing the tree structure in learning process, our method uses a complete binary tree structure with the pre-defined depth and the soft decision scheme to learn discriminative features in each root-to-leaf path, which avoids the pruning error and reduces the training time. In addition, the attention transformer module is used to further help our network to achieve better performance. Several experiments are conducted on the CUB-200-2011 [42], Stanford Cars [25], and Aircraft [32] datasets, demonstrating the favorable performance of the proposed method against the state-of-the-art methods. We also conduct the ablation study to comprehensively understand the influences of different components in the proposed method.

The main contributions of this paper are summarized as follows. (1) We propose a new attention convolutional binary neural tree for FGVC. (2) The attention transformer is introduced to facilitate the coarse-to-fine hierarchical feature learning in the tree network. (3) Extensive experiments on three challenging datasets, *i.e.*, CUB-200-2011, Stanford Cars, and Aircraft, show the effectiveness of our method.

2. Related Works

Deep supervised methods. Some algorithms [51, 31, 18, 50] use object annotations or even dense part/keypoint annotations to guide the training of deep CNN model for FGVC. Zhang *et al.* [51] propose to learn two detectors, *i.e.*,

the whole object detector and the part detector, to predict the fine-grained categories based on the pose-normalized representation. Liu *et al.* [31] propose a fully convolutional attention networks that glimpses local discriminative regions to adapt to different fine-grained domains. The method in [18] construct the part-stacked CNN architecture, which explicitly explains the fine-grained recognition process by modeling subtle differences from object parts. In [50], the proposed network consists of detection and classification sub-networks. The detection sub-network is used to generate small semantic part candidates for detection; while the classification sub-network can extract features from parts detected by the detection sub-network. However, these methods rely on labor-intensive part annotations, which limits their applications in real scenarios.

Deep weakly supervised method. To that end, more recent methods [52, 12, 38, 46] only require image-level annotations. Zheng *et al.* [52] introduce a multi-attention CNN model, where part generation and feature learning process reinforce each other for accurate results. Fu *et al.* [12] develop a recurrent attention module to recursively learn discriminative region attention and region-based feature representation at multiple scales in a mutually reinforced way. Recently, Sun *et al.* [38] regulate multiple object parts among different input images by using multiple attention region features of each input image. In [46], a bank of convolutional filters is learned to capture class-specific discriminative patches, through a novel asymmetric multi-stream architecture with convolutional filter supervision. However, the aforementioned methods merely integrate the attention mechanism in a single network, affecting their performance.

Decision tree. Decision tree is an effective algorithm for classification task. It selects the appropriate directions based on the characteristic of feature. The inherent ability of interpretability makes it as promising direction to pose insight into internal mechanism in deep learning. Xiao [48] propose the principle of fully functioned neural graph and design neural decision tree model for categorization task. Frosst and Hinton [11] develop a deep neural decision tree model to understand decision mechanism for particular test case in a learned network. Tanno *et al.* [41] propose the adaptive neural trees that incorporates representation learning into edges, routing functions and leaf nodes of a decision tree. In our work, we integrate the decision tree with neural network to implement sub-branch selection and representation learning simultaneously.

Attention mechanism. Attention mechanism has played an important role in deep learning to mimic human visual mechanism. In [49], the attention is used to make sure the student model focuses on the discriminative regions as teacher model does. In [21], the cascade attention mechanism is proposed to guide the different layers of CNN and

concatenate them to gain discriminative representation as the input of final linear classifier. Hu *et al.* [16] apply the attention mechanism from aspect of channels and allocate the different weights according to the contribution of each channel. The CBAM module in [47] combines space region attentions with feature map attentions. In contrast to the aforementioned methods, we apply the attention mechanism on each branch of the tree architecture to sake the discriminative regions for classification.

3. Attention Convolutional Binary Neural Tree

Our ACNet model aims to classify each object sample in X to sub-categories, *i.e.*, assign each sample in X with the category label Y , which consists of four modules, *i.e.*, the backbone network, the branch routing, the attention transformer, and the label prediction modules, shown in Figure 2. We define the ACNet as a pair (\mathbb{T}, \mathbb{O}) , where \mathbb{T} defines the topology of the tree, and \mathbb{O} denotes the set of operations along the edges of \mathbb{T} . Notably, we use the full binary tree $\mathbb{T} = \{\mathcal{V}, \mathcal{E}\}$, where $\mathcal{V} = \{v_1, \dots, v_n\}$ is the set of nodes, n is the total number of nodes, and $\mathcal{E} = \{e_1, \dots, e_k\}$ is the set of edges between nodes, k is the total number of edges. Since we use the full binary tree \mathbb{T} , we have $n = 2^h - 1$ and $k = 2^h - 2$, where h is the height of \mathbb{T} . Each node in \mathbb{T} is formed by a routing module determining the sending path of samples, and the attention transformers are used as the operations along the edges.

Meanwhile, we use the asymmetrical architecture in the fully binary tree \mathbb{T} , *i.e.*, two attention transformers are used in the left edge, and one attention transformer is used in the right edge. In this way, the network is able to capture the different scales of features for accurate results. The detail architecture of our ACNet model is described as follows.

3.1. Architecture

Backbone network module. Since the discriminative regions in fine-grained categories are highly localized [46], we need to use a relatively small receptive field of the extracted features by constraining the size and stride of the convolutional filters and pooling kernels. The truncated network is used as the backbone network module to extract features, which is pre-trained on the ILSVRC CLS-LOC dataset [35]. Similar to [38], we use the input image size 448×448 instead of the default 224×224 . Notably, ACNet can also work on other pre-trained networks, such as ResNet [15] and Inception V2 [19]. In practice, we use VGG-16 [37] (retaining the layers from conv1_1 to conv4_3) and ResNet-50 [15] (retaining the layers from res_1 to res_4) networks as the backbone in this work.

Branch routing module. As described above, we use the branch routing module to determine which child (*i.e.*, left or right child) the samples would be sent to. Specifically, as shown in Figure 2(b), the i -th routing module $\mathcal{R}_i^k(\cdot)$ at the

k -th layer uses one convolutional layer with the kernel size 1×1 , followed by a global context block [4]. The global context block is an improvement of the simplified non-local (NL) block [44] and Squeeze-Excitation (SE) block [16], which shares the same implementation with the simplified NL block on the context modeling and fusion steps, and shares the transform step with the SE block. In this way, the context information is integrated to better describe the objects. After that, we use the global average pooling [27], element-wise square-root and L2 normalization [28], and a fully connected layer with the sigmoid activation function to produce a scalar value in $[0, 1]$ indicating the probability of samples being sent to the left or right sub-branches. Let $\phi_i^k(x_j)$ denote the output probability of the j -th sample $x_j \in X$ being sent to the right sub-branch produced by the branch routing module $\mathcal{R}_i^k(x_j)$, where $\phi_i^k(x_j) \in [0, 1]$, $i = 1, \dots, 2^{k-1}$. Thus, we have the probability of the sample $x_j \in X$ being sent to the left sub-branch to be $1 - \phi_i^k(x_j)$. If the probability $\phi_i^k(x_j)$ is larger than 0.5, we prefer the left path instead of the right one; otherwise, the left branch dominates the final decision.

Attention transformer. The attention transformer module is used to enforce the network to capture discriminative features, see Figure 3. According to the fact that the empirical receptive field is much smaller than theoretical receptive field in deep networks [30], the discriminative representation should be formed by larger receptive field in new-added layers of our proposed tree structure. To this end, we intergate the Atrous Spatial Pyramid Pooling (ASPP) module [5] into the attention transformer. Specifically, ASPP module provides different feature maps with each characterized by a different scale/receptive field and an attention module. Then, multi-scale feature maps are generated by four parallel dilated convolutions with different dilated rates, *i.e.*, 1, 6, 12, 18. Following the parallel dilated convolution layers, the concatenated feature maps are fused by one convolutional layer with kernel 1×1 and stride 1. Following the ASPP module, we insert an attention module, which generates a channel attention map with the size $\mathbb{R}^{C \times 1 \times 1}$ using a batch normalization (BN) layer [19], a global average pooling (GAP) layer, a fully connected (FC) layer and ReLU activation function, and a FC layer and sigmoid function. In this way, the network is guided to focus on meaningful features for accurate results.

Label prediction. For each leaf node in our ACNet model, we use the label prediction module \mathcal{P}_i (*i.e.*, $i = 1, \dots, 2^{h-1}$) to predict the subordinate category of the object x_j , see Figure 2. Let $r_i^k(x_j)$ to be the accumulated probability of the object x_j passing from the root node to the i -th node at the k -th layer. For example, if the root to the node $\mathcal{R}_i^k(\cdot)$ path on the tree is $\mathcal{R}_1^1, \mathcal{R}_1^2, \dots, \mathcal{R}_1^k$, *i.e.*, the object x_j is always sent to the left child, we have $r_i^k(x_j) = \prod_{i=1}^k \phi_1^i(x_j)$. As shown in Figure 2, the la-

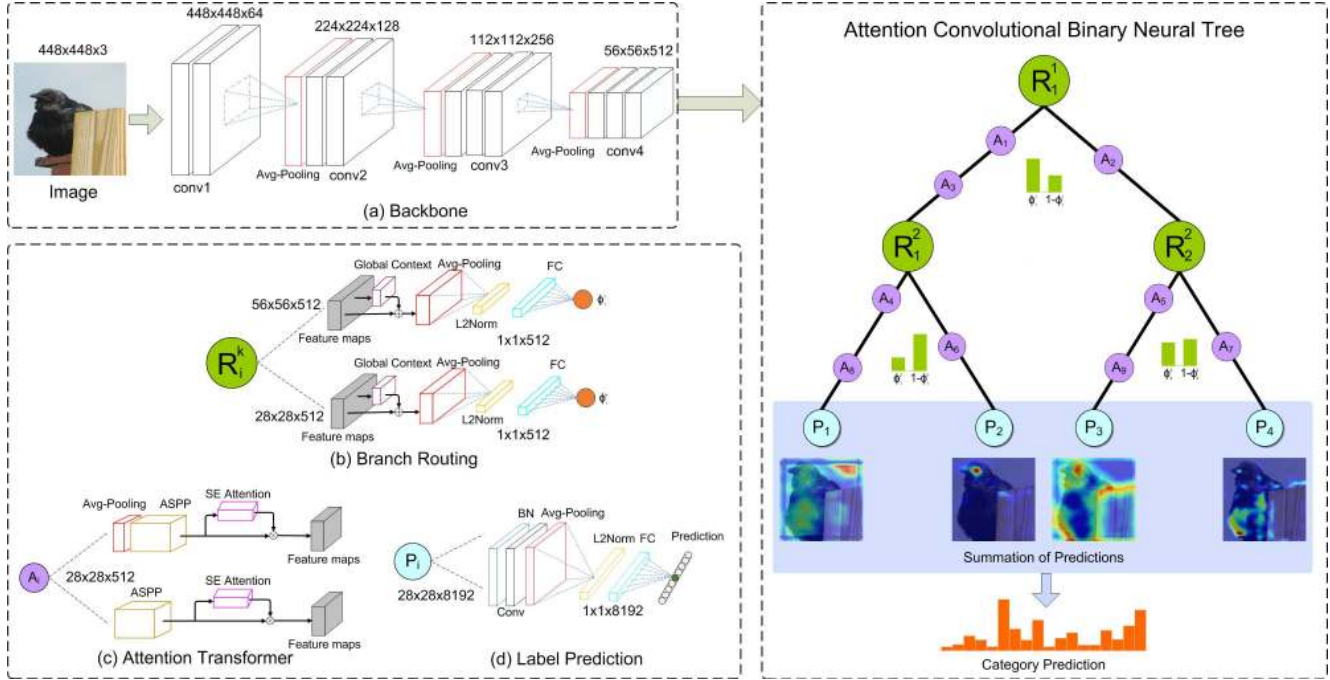


Figure 2: The overview of our ACNet model, formed by (a) the backbone network module, (b) the branch routing module, (c) the attention transformer module, and (d) the label prediction module. We show an example image in *Fish.Crow*. Best visualization in color.

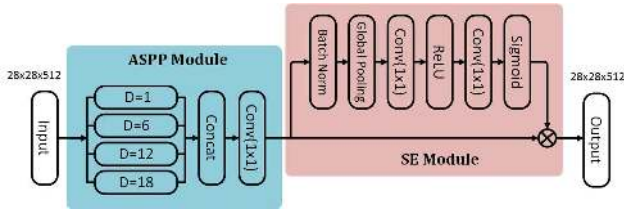


Figure 3: The architecture of the attention transformer module.

bel prediction module is formed by a batch normalization layer, a convolutional layer with kernel size 1×1 , a max-pooling layer, a sqrt and L2 normalization layer, and a fully connected layer. Then, the final prediction $\mathcal{C}(x_j)$ of the j -th object x_j is computed as the summation of all leaf predictions multiplied with the accumulated probability generated by the passing branch routing modules, *i.e.*, $\mathcal{C}(x_j) = \sum_{i=1}^{2^{h-1}} \mathcal{P}_i(x_j) r_i^h(x_j)$. We would like to emphasize that $\|\mathcal{C}(x_j)\|_1 = 1$, *i.e.*, the summation of confidences of x_j belonging to all subordinate classes equal to 1,

$$\|\mathcal{C}(x_j)\|_1 = \left\| \sum_{i=1}^{2^{h-1}} \mathcal{P}_i(x_j) r_i^h(x_j) \right\|_1 = 1, \quad (1)$$

where $r_i^h(x_j)$ is the accumulated probability of the i -th node at the leaf layer. We present a short description to prove that $\|\mathcal{C}(x_j)\|_1 = 1$ as follows.

Proof. Let $r_i^k(\cdot)$ be the accumulated probability of the i -th branch routing module $\mathcal{R}_i^k(\cdot)$ at the k -th layer. Thus, the

accumulated probabilities of the left and right children corresponding to $\mathcal{R}_i^k(\cdot)$ are $r_{2i-1}^{k+1}(\cdot)$ and $r_{2i}^{k+1}(\cdot)$, respectively. At first, we demonstrate that the summation of the accumulated probabilities $r_{2i-1}^{k+1}(\cdot)$ and $r_{2i}^{k+1}(\cdot)$ is equal to the accumulated probability of their parent $r_i^k(x_j)$. That is,

$$\begin{aligned} & r_{2i-1}^{k+1}(x_j) + r_{2i}^{k+1}(x_j) \\ &= \phi_{2i-1}^{k+1}(x_j) \cdot r_i^k(x_j) + \phi_{2i}^{k+1}(x_j) \cdot r_i^k(x_j) \\ &= \phi_{2i-1}^{k+1}(x_j) \cdot r_i^k(x_j) + (1 - \phi_{2i-1}^{k+1}(x_j)) \cdot r_i^k(x_j) \\ &= r_i^k(x_j). \end{aligned} \quad (2)$$

Meanwhile, due to the fully binary tree \mathbb{T} in ACNet, we have $\sum_{i=1}^{2^{h-1}} r_i^h(x_j) = \sum_{i=1}^{2^{h-2}} (r_{2i-1}^h(x_j) + r_{2i}^h(x_j))$. We can further get $\sum_{i=1}^{2^{h-1}} r_i^h(x_j) = \sum_{i=1}^{2^{h-2}} r_i^{h-1}(x_j)$. This process is carried out iteratively, and we have $\sum_{i=1}^{2^{h-1}} r_i^h(x_j) = \dots = r_1^1(x_j) = 1$. In addition, since the category prediction $\mathcal{P}_i(x_j)$ is generated by the softmax layer (see Figure 2), we have $\|\mathcal{P}_i(x_j)\|_1 = 1$. Thus,

$$\begin{aligned} \|\mathcal{C}(x_j)\|_1 &= \left\| \sum_{i=1}^{2^{h-1}} \mathcal{P}_i(x_j) r_i^h(x_j) \right\|_1 \\ &= \sum_{i=1}^{2^{h-1}} \|\mathcal{P}_i(x_j)\|_1 r_i^h(x_j) = 1. \end{aligned} \quad (3)$$

□

As shown in Figure 2, when occlusion happens, ACNet can still localize some discriminative object parts and context information of the bird. Although high intra-class

visual variances always happen in FGVC, ACNet uses a coarse-to-fine hierarchical feature learning process to exploit the discriminative feature for classification. In this way, different branches in the tree structure focus on different fine-grained object regions for accurate results.

3.2. Training

Data augmentation. In the training phase, we use the cropping and flipping operations to augment data to construct a robust model to adapt to variations of objects. That is, we first rescale the original images such that their shorter side is 512 pixels. After that, we randomly crop the patches with the size 448×448 , and randomly flip them to generate the training samples.

Loss function. The loss function for our ACNet is formed by two parts, *i.e.*, the loss for the predictions of leaf nodes, and the loss for the final prediction, computed by the summation over all predictions from the leaf nodes. That is,

$$\mathcal{L} = L(\mathcal{C}(x_j), y^*) + \sum_{i=1}^{2^{h-1}} L(\mathcal{P}_i(x_j), y^*), \quad (4)$$

where h is the height of the tree \mathbb{T} , $L(\mathcal{C}(x_j), y^*)$ is the negative logarithmic likelihood loss of the final prediction $\mathcal{C}(x_j)$ and the ground truth label y^* , and $L(\mathcal{P}_i(x_j), y^*)$ is the negative logarithmic likelihood loss of the i -th leaf prediction and the ground truth label y^* .

Optimization. The backbone network in our ACNet method is pre-trained on the ImageNet dataset. Besides, the ‘‘xavier’’ method [14] is used to randomly initialize the parameters of the additional convolutional layers. The entire training process is formed by two stages.

- For the first stage, the parameters in the truncated VGG-16 network are fixed, and other parameters are trained with 60 epochs. The batch size is set to 24 in training with the initial learning rate 1.0. The learning rate is gradually divided by 4 at the 10-th, 20-th, 30-th, and 40-th epochs.
- In the second stage, we fine-tune the entire network for 200 epochs. We use the batch size 16 in training with the initial learning rate 0.001. The learning rate is gradually divided by 10 at the 30-th, 40-th, and 50-th epochs.

We use the SGD algorithm to train the network with 0.9 momentum, and 0.000005 weight decay in the first stage and 0.0005 weight decay in the second stage.

4. Experiments

Several experiments on three FGVC datasets, *i.e.*, CUB-200-2011 [42], Stanford Cars [25], and Aircraft [32], are

Table 1: The fine-grained classification results on the CUB-200-2011 dataset [42].

Method	Backbone	Annotation	Top-1 Acc. (%)
FCAN [31]	ResNet-50	✓	84.7
B-CNN [29]	VGG-16	✓	85.1
SPDA-CNN [50]	CaffeNet	✓	85.1
PN-CNN [2]	Alex-Net	✓	85.4
STN [20]	Inception	×	84.1
B-CNN [29]	VGG-16	×	84.0
CBP [13]	VGG-16	×	84.0
LRBP [23]	VGG-16	×	84.2
FCAN [31]	ResNet-50	×	84.3
RA-CNN [12]	VGG-19	×	85.3
HIHCA [3]	VGG-16	×	85.3
Improved B-CNN [28]	VGG-16	×	85.8
BoostCNN [33]	VGG-16	×	86.2
KP [8]	VGG-16	×	86.2
MA-CNN [52]	VGG-19	×	86.5
MAMC [38]	ResNet-101	×	86.5
MaxEnt [10]	DenseNet-161	×	86.5
HBPASM [40]	Resnet-34	×	86.8
DCL [7]	VGG-16	×	86.9
KERL w/ HR [6]	VGG-16	×	87.0
TASN [53]	VGG-19	×	87.1
DFL-CNN [46]	ResNet-50	×	87.4
DCL [7]	ResNet-50	×	87.8
TASN [53]	ResNet-50	×	87.9
Ours	VGG-16	×	87.8
Ours	ResNet-50	×	88.1

conducted to demonstrate the effectiveness of the proposed method. Our method is implemented in Caffe [22]. All models are trained on a workstation with a 3.26 GHz Intel processor, 32 GB memory, and a Nvidia V100 GPU.

4.1. Evaluation on the CUB-200-2011 Dataset

The Caltech-UCSD birds dataset (CUB-200-2011) [42] consists of 11,788 annotated images in 200 subordinate categories, including 5,994 images for training and 5,794 images for testing. The fine-grained classification results are shown in Table 1. As shown in Table 1, the best supervised method¹, *i.e.*, PN-CNN [2] using both the object and part level annotations produces 85.4% top-1 accuracy on the CUB-200-2011 dataset. Without part-level annotation, MAMC [38] produces 86.5% top-1 accuracy using two attention branches to learn discriminative features in different regions. KERL w/ HR [6] designs a single deep gated graph neural network to learn discriminative features,

¹Notably, the supervised method requires object or part level annotations, demanding significant human effort. Thus, most of recent methods focus on the weakly supervised methods, pushing the state-of-the-art weakly supervised methods surpassing the performance of previous supervised methods.

Table 2: The fine-grained classification results on the Stanford Cars dataset [25].

Method	Backbone	Annotation	Top-1 Acc. (%)
FCAN [31]	ResNet-50	✓	91.3
PA-CNN [24]	VGG-19	✓	92.6
FCAN [31]	ResNet-50	×	89.1
B-CNN [29]	VGG-16	×	90.6
LRBP [23]	VGG-16	×	90.9
HIHCA [3]	VGG-16	×	91.7
Improved B-CNN [28]	VGG-16	×	92.0
BoostCNN [33]	VGG-16	×	92.1
KP [8]	VGG-16	×	92.4
RA-CNN [12]	VGG-19	×	92.5
MA-CNN [52]	VGG-19	×	92.8
MAMC [38]	ResNet-101	×	93.0
MaxEnt [10]	DenseNet-161	×	93.0
WS-DAN [17]	Inception v3	×	93.0
DFL-CNN [46]	ResNet-50	×	93.1
HBPASM [40]	Resnet-34	×	93.8
TASN [53]	VGG-19	×	93.2
TASN [53]	ResNet-50	×	93.8
DCL [7]	VGG-16	×	94.1
DCL [7]	ResNet-50	×	94.5
Ours	VGG-16	×	94.3
Ours	ResNet-50	×	94.6

achieving better performance, *i.e.*, 87.0% top-1 accuracy. Compared to the state-of-the-art weakly supervised methods [6, 10, 38, 46, 7, 53], our method achieves the best results with 87.8% and 88.1% top-1 accuracy with different backbones. This is attributed to the designed attention transformer module and the coarse-to-fine hierarchical feature learning process.

4.2. Evaluation on the Stanford Cars Dataset

The Stanford Cars dataset [25] contains 16,185 images from 196 classes, which is formed by 8,144 images for training and 8,041 images for testing. The subordinate categories are determined by the *Make*, *Model*, and *Year* of cars. As shown in Table 2, previous methods using part-level annotations (*i.e.*, FCAN [31] and PA-CNN [24]) only produces less than 93.0% top-1 accuracy. The recent weakly supervised method WS-DAN [17] employs the complex Inception V3 backbone [39] and designs the attention-guided data augmentation strategy to exploit discriminative object parts, achieving 93.0% top-1 accuracy. Without using any fancy data augmentation strategy, our method achieves the best top-1 accuracy, *i.e.*, 94.3% with the VGG-16 backbone and 94.6% with the ResNet-50 backbone.

4.3. Evaluation on the Aircraft Dataset

The Aircraft dataset [32] is a fine-grained dataset of 100 different aircraft variants formed by 10,000 annotated im-

Table 3: The fine-grained classification results on the Aircraft dataset [32].

Method	Backbone	Annotation	Top-1 Acc. (%)
MG-CNN [43]	ResNet-50	✓	86.6
MDTP [45]	VGG-16	✓	88.4
RA-CNN [12]	VGG-19	×	88.2
MA-CNN [52]	VGG-19	×	89.9
B-CNN [29]	VGG-16	×	86.9
KP [8]	VGG-16	×	86.9
LRBP [23]	VGG-16	×	87.3
HIHCA [3]	VGG-16	×	88.3
Improved B-CNN [28]	VGG-16	×	88.5
BoostCNN [33]	VGG-16	×	88.5
PC-DenseNet-161 [9]	DenseNet-161	×	89.2
MaxEnt [10]	DenseNet-161	×	89.7
HBPASM [40]	Resnet-34	×	91.3
DFL-CNN [46]	ResNet-50	×	91.7
DCL [7]	VGG-16	×	91.2
DCL [7]	ResNet-50	×	93.0
Ours	VGG-16	×	91.5
Ours	ResNet-50	×	92.4

ages, which is divided into two subsets, *i.e.*, the training set with 6,667 images and the testing set with 3,333 images. Specifically, the category labels are determined by the *Model*, *Variant*, *Family* and *Manufacturer* of airplanes. The evaluation results are presented in Table 3. Our ACNet method outperforms the most compared methods, especially with the same VGG-16 backbone. Besides, our model performs on par with the state-of-the-art method DCL [7], *i.e.*, 91.2% vs. 91.5% top-1 accuracy for the VGG-16 backbone and 93.0% vs. 92.4% top-1 accuracy for the ResNet-50 based backbone network. The operations along different root-to-leaf paths in our tree architecture \mathbb{T} focus on exploiting discriminative features on different object regions, which help each other to achieve the best performance in FGVC.

4.4. Ablation Study

We study the influence of some important parameters and different components of ACNet. Notably, we employ the VGG-16 backbone in the experiment. The Grad-CAM method [36] is used to generate the heatmaps to visualize the responses of branch routing and leaf nodes.

Effectiveness of the tree architecture \mathbb{T} . To validate the effectiveness of the tree architecture design, we construct two variants, *i.e.*, VGG and w/ Tree, of our ACNet method. Specifically, we construct the VGG method by only using the VGG-16 backbone network for classification, and further integrate the tree architecture to form the w/ Tree method. The evaluation results are reported in Figure 6. We find that using the tree architecture significantly improves

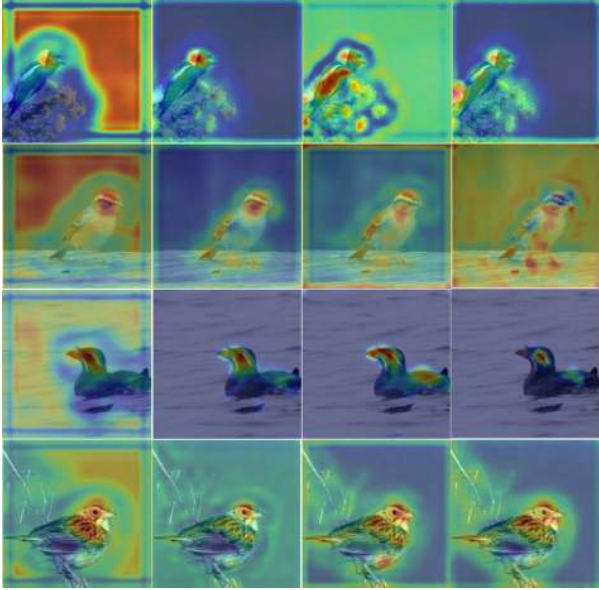


Figure 4: Visualization of the responses in different leaf nodes in our ACNet method. Each column presents a response heatmap of each leaf node.

the accuracy, *i.e.*, 3.025% improvements in top-1 accuracy, which demonstrates the effectiveness of the designed tree architecture \mathbb{T} in our ACNet method.

Height of the tree \mathbb{T} . To explore the effect of the height of the tree \mathbb{T} , we construct four variants with different heights of tree on the CUB-200-2011 dataset [42] in Table 4. Notably, the tree \mathbb{T} is degenerated to a single node when the height of the tree is set to 1, *i.e.*, only the backbone VGG-16 network is used in classification. As shown in Table 4, we find that our ACNet achieves the best performance (*i.e.*, 87.8% top-1 accuracy) with the height of tree equals to 3. If we set $h \leq 2$, there are limited number of parameters in our ACNet model, which are not enough to represent the significant variations of the subordinate categories. However, if we set $h = 4$, too many parameters with limited number of training data cause overfitting of our ACNet model, resulting in 2.3% drop in the top-1 accuracy. To verify our hypothesis, we visualize the responses of all leaf nodes in ACNet with the height of 4 in Figure 5. We find that some leaf nodes focus on almost the same regions (see the 3rd and 4th columns).

Effectiveness of leaf nodes. To analyze the effectiveness of the individual leaf node, we calculate the accuracy of individual leaf predictions with height of 3, respectively. The accuracies of four individual leaf nodes are 85.8%, 86.2%, 86.7%, and 87.0% on CUB-200-2011 respectively. It shows that all leaf nodes are informative and fusion of them can produce more accurate result (*i.e.*, 87.8%). As shown in Figure 4, we observe that different leaf nodes concentrate on different regions of images. For example, the leaf node

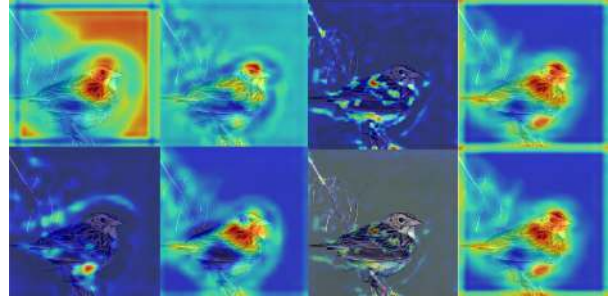


Figure 5: Responses of all leaf nodes in the tree with the height of 4.

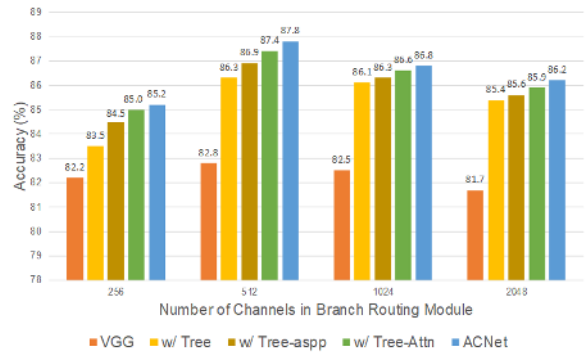


Figure 6: Effect of the various components in the proposed ACNet method on the CUB-200-2011 dataset [42].

corresponding to the first column focuses more on the background region, the leaf node corresponding to the second column focuses more on the head region, and the other two leaf nodes are more interested in the patches of wings and tail. The different leaf nodes help each other to construct more effective model for accurate results.

Asymmetrical architecture of the tree \mathbb{T} . To explore the architecture design in \mathbb{T} , we construct two variants, *i.e.*, one uses the symmetry architecture, and another one uses the asymmetrical architecture, and set the height of the tree \mathbb{T} to be 3. The evaluation results are reported in Table 5. It can be seen that the proposed method produces 86.2% top-1 accuracy using the symmetrical architecture. If we use the asymmetrical architecture, the top-1 accuracy is improved 1.6% to 87.8%. We speculate that the asymmetrical architecture is able to fuse various features with different receptive fields for better performance.

Effectiveness of the attention transformer module. We construct a variant “w/ Tree-Attn”, of the proposed ACNet model, to validate the effectiveness of the attention transformer module in Figure 6. Specifically, we add the attention block in the transformer module in the “w/ Tree” method to construct the “w/ Tree-Attn” method. As shown in Figure 6, the “w/ Tree-Attn” method performs consistently better than the “w/ Tree” method, producing higher

Table 4: Effect of the height of the tree \mathbb{T} on the CUB-200-2011 dataset [42].

Height of \mathbb{T}	Top-1 Acc. (%)
1	82.2
2	86.0
3	87.8
4	85.5

Table 5: Effect of tree architecture on the CUB-200-2011 dataset [42].

Mode	Level	Top-1 Acc. (%)
symmetry	3	86.2
asymmetry	3	87.8

Table 6: Comparison between GMP and GAP on the CUB-200-2011 dataset [42].

Pooling	Top-1 Acc. (%)
GMP	87.2
GAP	87.8

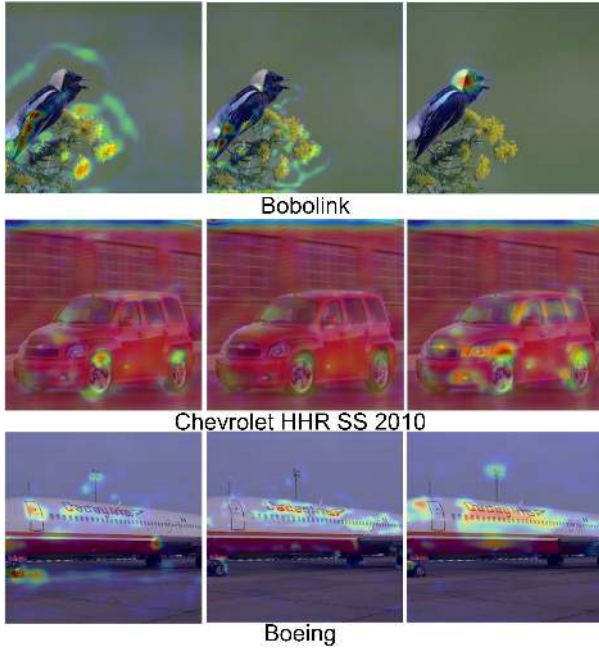


Figure 7: Visualization of the responses in different branch routing modules of the exemplars in CUB-200-2011 [42], Stanford Cars [25], and Aircraft [32].

top-1 accuracy with different number of channels, *i.e.*, improving 0.4% top-1 accuracy in average, which demonstrates that the attention mechanism is effective for FGVC.

To further investigate the effect of ASPP module in our proposed model, we also conduct the “w/ Tree-ASPP” method, a variant of proposed ACNet model, where the only difference lies on between one convolution layer or ASPP module in the attention transformer module. As illustrated in Figure 6, the attention transformer with ASPP module achieves better accuracy than the one with only one convolution layer. It indicates that the ASPP module improves the global performance by parallel dilated convolution layers with different dilated rates. Specifically, the “w/ Tree-ASPP” method improves 0.5% top-1 accuracy in average. We can conclude that multi-scale embedding and different dilated convolutions in ASPP module can facilitate helping the proposed tree network to obtain robust performance.

Components in the branch routing module. We ana-

lyze the effectiveness of the global context block [4] in the branch routing module in Figure 6. Our ACNet method produces the best results with different number of channels in the branch routing module; while the top-1 accuracy drops 0.275% in average after removing the global context block. Meanwhile, we also study the effectiveness of the pooling strategy in the branch routing module in Table 6. We observe that using the global max-pooling (GMP) instead of the global average pooling (GAP) leads to 0.6% top-1 accuracy drop on the CUB-200-2011 dataset. We speculate that the GAP operation encourages the filter to focus on high average response regions instead of the only maximal ones, which is able to integrate more context information for better performance.

Coarse-to-fine hierarchical feature learning process.

The branch routing modules focus on different semantic regions (*e.g.*, different object parts) or context information (*e.g.*, background) at different levels, *e.g.*, R_1^1 , R_1^2 , and R_2^2 in Figure 2. As the example Bobolink shown in Figure 7, the R_1^1 module focuses on the whole bird region at level-1; the R_1^2 and R_2^2 modules focus on the wing and head regions of the bird at level-2. As shown in the first row in Figure 5, the four leaf nodes focus on several fine-grained object parts at level-3, *e.g.*, different parts of the head region. In this way, our ACNet uses the coarse-to-fine hierarchical feature learning process to exploit discriminative features for more accurate results. This phenomenon demonstrates that our hierarchical feature extraction process in the tree \mathbb{T} architecture gradually enforces our model to focus on more discriminative detail regions of object.

5. Conclusion

In this paper, we present an attention convolutional binary neural tree (ACNet) for weakly supervised FGVC. Specifically, different root-to-leaf paths in the tree network focus on different discriminative regions using the attention transformer inserted into the convolutional operations along edges of the tree. The final decision is produced by max-voting the predictions from leaf nodes. The experiments conducted on several challenging datasets show the effectiveness of ACNet. We also present how we design the tree structure and why we use the coarse-to-fine hierarchical feature learning process for FGVC.

References

- [1] Anelia Angelova, Shenghuo Zhu, and Yuanqing Lin. Image segmentation for large-scale subcategory flower recognition. In *WACV*, pages 39–45, 2013. [1](#)
- [2] Steve Branson, Grant Van Horn, Serge J. Belongie, and Pietro Perona. Bird species categorization using pose normalized deep convolutional nets. *CoRR*, abs/1406.2952, 2014. [5](#)
- [3] Sijia Cai, Wangmeng Zuo, and Lei Zhang. Higher-order integration of hierarchical convolutional activations for fine-grained visual categorization. In *ICCV*, pages 511–520, 2017. [5](#), [6](#)
- [4] Yue Cao, Jiarui Xu, Stephen Lin, Fangyun Wei, and Han Hu. Gcnet: Non-local networks meet squeeze-excitation networks and beyond. *CoRR*, abs/1904.11492, 2019. [3](#), [8](#)
- [5] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *TPAMI*, 40(4):834–848, 2018. [3](#)
- [6] Tianshui Chen, Liang Lin, Riquan Chen, Yang Wu, and Xiaonan Luo. Knowledge-embedded representation learning for fine-grained image recognition. In *IJCAI*, pages 627–634, 2018. [5](#), [6](#)
- [7] Yue Chen, Yalong Bai, Wei Zhang, and Tao Mei. Destruction and construction learning for fine-grained image recognition. In *CVPR*, pages 5157–5166, 2019. [5](#), [6](#)
- [8] Yin Cui, Feng Zhou, Jiang Wang, Xiao Liu, Yuanqing Lin, and Serge J. Belongie. Kernel pooling for convolutional neural networks. In *CVPR*, pages 3049–3058, 2017. [5](#), [6](#)
- [9] Abhimanyu Dubey, Otkrist Gupta, Pei Guo, Ramesh Raskar, Ryan Farrell, and Nikhil Naik. Pairwise confusion for fine-grained visual classification. In *ECCV*, pages 71–88, 2018. [6](#)
- [10] Abhimanyu Dubey, Otkrist Gupta, Ramesh Raskar, and Nikhil Naik. Maximum-entropy fine grained classification. In *NeurIPS*, pages 635–645, 2018. [5](#), [6](#)
- [11] Nicholas Frosst and Geoffrey E. Hinton. Distilling a neural network into a soft decision tree. *CoRR*, abs/1711.09784, 2017. [2](#)
- [12] Jianlong Fu, Heliang Zheng, and Tao Mei. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In *CVPR*, pages 4476–4484, 2017. [2](#), [5](#), [6](#)
- [13] Yang Gao, Oscar Beijbom, Ning Zhang, and Trevor Darrell. Compact bilinear pooling. *CoRR*, abs/1511.06062, 2015. [5](#)
- [14] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *AISTATS*, pages 249–256, 2010. [5](#)
- [15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. [3](#)
- [16] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR*, pages 7132–7141, 2018. [3](#)
- [17] Tao Hu, Honggang Qi, Qingming Huang, and Yan Lu. See better before looking closer: Weakly supervised data augmentation network for fine-grained visual classification. *CoRR*, abs/1901.09891, 2019. [6](#)
- [18] Shaoli Huang, Zhe Xu, Dacheng Tao, and Ya Zhang. Part-stacked CNN for fine-grained visual categorization. In *CVPR*, pages 1173–1182, 2016. [2](#)
- [19] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, pages 448–456, 2015. [3](#)
- [20] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial transformer networks. In *NeurIPS*, pages 2017–2025, 2015. [5](#)
- [21] Saumya Jetley, Nicholas A. Lord, Namhoon Lee, and Philip H. S. Torr. Learn to pay attention. *CoRR*, abs/1804.02391, 2018. [2](#)
- [22] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross B. Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. *CoRR*, abs/1408.5093, 2014. [5](#)
- [23] Shu Kong and Charless C. Fowlkes. Low-rank bilinear pooling for fine-grained classification. *CoRR*, abs/1611.05109, 2016. [5](#), [6](#)
- [24] Jonathan Krause, Hailin Jin, Jianchao Yang, and Fei-Fei Li. Fine-grained recognition without part annotations. In *CVPR*, pages 5546–5555, 2015. [6](#)
- [25] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *ICCVW*, pages 554–561, 2013. [2](#), [5](#), [6](#), [8](#)
- [26] Haoming Lin, Yuyang Hu, Siping Chen, Jianhua Yao, and Ling Zhang. Fine-grained classification of cervical cells using morphological and appearance based convolutional neural networks. *CoRR*, abs/1810.06058, 2018. [2](#)
- [27] Min Lin, Qiang Chen, and Shuicheng Yan. Network in network. In *ICLR*, 2014. [3](#)
- [28] Tsung-Yu Lin and Subhransu Maji. Improved bilinear pooling with cnns. In *BMVC*, 2017. [3](#), [5](#), [6](#)
- [29] Tsung-Yu Lin, Aruni Roy Chowdhury, and Subhransu Maji. Bilinear CNN models for fine-grained visual recognition. In *ICCV*, pages 1449–1457, 2015. [2](#), [5](#), [6](#)
- [30] Wei Liu, Andrew Rabinovich, and Alexander C. Berg. Parsenet: Looking wider to see better. *CoRR*, abs/1506.04579, 2015. [3](#)
- [31] Xiao Liu, Tian Xia, Jiang Wang, and Yuanqing Lin. Fully convolutional attention localization networks: Efficient attention localization for fine-grained recognition. *CoRR*, abs/1603.06765, 2016. [2](#), [5](#), [6](#)
- [32] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew B. Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *CoRR*, abs/1306.5151, 2013. [2](#), [5](#), [6](#), [8](#)
- [33] Mohammad Moghimi, Serge J. Belongie, Mohammad J. Saberian, Jian Yang, Nuno Vasconcelos, and Li-Jia Li. Boosted convolutional neural networks. In *BMVC*, 2016. [5](#), [6](#)
- [34] Yuxin Peng, Xiangteng He, and Junjie Zhao. Object-part attention driven discriminative localization for fine-grained image classification. *CoRR*, abs/1704.01740, 2017. [2](#)
- [35] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy,

- Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Fei-Fei Li. Imagenet large scale visual recognition challenge. *IJCV*, 115(3):211–252, 2015. 3
- [36] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, pages 618–626, 2017. 6
- [37] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014. 3
- [38] Ming Sun, Yuchen Yuan, Feng Zhou, and Errui Ding. Multi-attention multi-class constraint for fine-grained image recognition. In *ECCV*, pages 834–850, 2018. 2, 3, 5, 6
- [39] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, pages 2818–2826, 2016. 6
- [40] Min Tan, Guijun Wang, Jian Zhou, Zhiyou Peng, and Meilian Zheng. Fine-grained classification via hierarchical bilinear pooling with aggregated slack mask. *Access*, 7:117944–117953, 2019. 5, 6
- [41] Ryutaro Tanno, Kai Arulkumaran, Daniel C. Alexander, Antonio Criminisi, and Aditya V. Nori. Adaptive neural trees. In *ICML*, pages 6166–6175, 2019. 2
- [42] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical report, California Institute of Technology, 2011. 1, 2, 5, 7, 8
- [43] Dequan Wang, Zhiqiang Shen, Jie Shao, Wei Zhang, Xiangyang Xue, and Zheng Zhang. Multiple granularity descriptors for fine-grained categorization. In *ICCV*, pages 2399–2406, 2015. 6
- [44] Xiaolong Wang, Ross B. Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, pages 7794–7803, 2018. 3
- [45] Yaming Wang, Jonghyun Choi, Vlad I. Morariu, and Larry S. Davis. Mining discriminative triplets of patches for fine-grained classification. *CoRR*, abs/1605.01130, 2016. 6
- [46] Yaming Wang, Vlad I. Morariu, and Larry S. Davis. Learning a discriminative filter bank within a CNN for fine-grained recognition. In *CVPR*, pages 4148–4157, 2018. 2, 3, 5, 6
- [47] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. CBAM: convolutional block attention module. *CoRR*, abs/1807.06521, 2018. 3
- [48] Han Xiao. NDT: neural decision tree towards fully functioned neural graph. *CoRR*, abs/1712.05934, 2017. 2
- [49] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *CoRR*, abs/1612.03928, 2016. 2
- [50] Han Zhang, Tao Xu, Mohamed Elhoseiny, Xiaolei Huang, Shaoting Zhang, Ahmed M. Elgammal, and Dimitris N. Metaxas. SPDA-CNN: unifying semantic part detection and abstraction for fine-grained recognition. In *CVPR*, pages 1143–1152, 2016. 2, 5
- [51] Ning Zhang, Jeff Donahue, Ross B. Girshick, and Trevor Darrell. Part-based r-cnns for fine-grained category detection. In *ECCV*, pages 834–849, 2014. 2
- [52] Heliang Zheng, Jianlong Fu, Tao Mei, and Jiebo Luo. Learning multi-attention convolutional neural network for fine-grained image recognition. In *ICCV*, pages 5219–5227, 2017. 1, 2, 5, 6
- [53] Heliang Zheng, Jianlong Fu, Zheng-Jun Zha, and Jiebo Luo. Looking for the devil in the details: Learning trilinear attention sampling network for fine-grained image recognition. *CoRR*, abs/1903.06150, 2019. 5, 6