Augmenting Strong Supervision Using Web Data for Fine-grained Categorization

Zhe Xu[1,2], Shaoli Huang[2], Ya Zhang[1], and Dacheng Tao[2]

[1]Cooperative Medianet Innovation Center and the Shanghai Key Laboratory of Multimedia Processing and Transmissions, Shanghai Jiao Tong University, Shanghai, 200240, China

[2]Centre for Quantum Computation and Intelligent Systems, Faculty of Engineering and Information Technology, University of Technology, Sydney, Ultimo, NSW 2007, Australia

{xz3030,ya_zhang}@sjtu.edu.cn, {shaoli.huang@student.,dacheng.tao@}uts.edu.au

Abstract

We propose a new method for fine-grained object recognition that employs part-level annotations and deep convolutional neural networks (CNNs) in a unified framework. Although both schemes have been widely used to boost recognition performance, due to the difficulty in acquiring detailed part annotations, strongly supervised fine-grained datasets are usually too small to keep pace with the rapid evolution of CNN architectures. In this paper, we solve this problem by exploiting inexhaustible web data. The proposed method improves classification accuracy in two ways: more discriminative CNN feature representations are generated using a training set augmented by collecting a large number of part patches from weakly supervised web images; and more robust object classifiers are learned using a multi-instance learning algorithm jointly on the strong and weak datasets. Despite its simplicity, the proposed method delivers a remarkable performance improvement on the CUB200-2011 dataset compared to baseline part-based R-CNN methods, and achieves the highest accuracy on this dataset even in the absence of test image annotations.

1. Introduction

Fine-grained object categorization has become increasingly popular over the last few years. In contrast to basic-level recognition, fine-grained categorization aims to distinguish between subordinate categories such as different animal species [18, 29] or man-made products [22]. Classifying objects at the subordinate level generally requires expert knowledge, which is not always available from a random human annotator. Therefore, automatic fine-grained recognition systems would be of huge value in real-world applications.

There are two strategies widely used in existing fine-grained categorization algorithms. First, as stated by Rosch *et al.* [26], basic-level categories are principally defined by object parts, whereas subordinate-level categories are distinguished according to the unique properties of these parts. This discovery encourages the use of part-based algorithms that rely on localizing object parts and assigning them detailed attributes. Various methods have been used to define object parts such as unsupervised patch discovery [32, 8], human-in-the-loop methods [12, 10], or direct reliance on strongly supervised datasets with part-level annotations [21, 4, 33, 20].

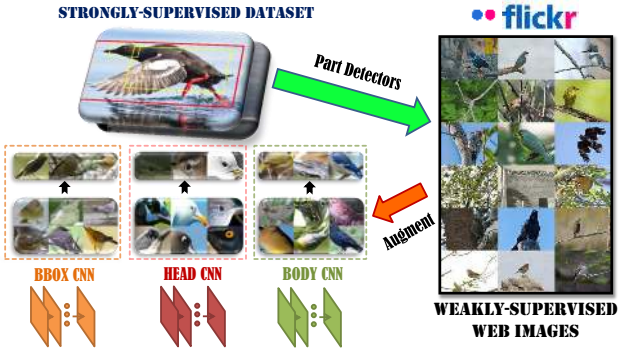

Figure 1. Illustration of the proposed method. Our goal is to solve the problem of insufficient training data due to the difficulty in acquiring detailed part annotations and the required data size for training robust CNNs. We introduce weakly supervised web images to augment the training set and design a new algorithm that iteratively updates feature representations and object classifiers on the augmented training data. The proposed method achieves 84.6% accuracy in the CUB200-2011 dataset without using any additional annotation at the testing stage.

The second strategy is to introduce more discriminative feature representations [5, 4, 34], which is particularly attractive given by the recent success of deep convolutional neural networks (CNNs) [19, 11] in visual recognition. By employing deep feature CNN extractors pre-trained on large datasets (such as ImageNet [9]) and domain-specific fine-tuning approaches, considerable improvements in a wide

range of image classification and detection tasks can be achieved, including in fine-grained categorization [25].

Inspired by these successes, a logical progression is to adopt the two methods in a unified framework [35, 33]. For example, part-based R-CNN [33] achieved state-of-the-art performance on the CUB200-2011 dataset [29] with the help of strong part annotations and CNN feature extractors, exemplifying the paradigm; nevertheless, it has been argued that further improving results this way may be problematic. In particular, CNNs require large volumes of training data to learn robust feature representations, and it is nearly impossible in practice to acquire large-scale strongly annotated datasets since part annotations, which require expert knowledge, are simply too expensive. The challenge, therefore, is to exploit more accurate part annotations whilst ensuring that there are sufficient training data for learning robust CNN feature representations.

One possible solution is to employ existing images from the "jungle of the interwebs". Websites such as Flickr offer a nearly endless supply of images with human-labeled titles or tags, providing the necessary resource to train complicated deep networks. However, directly employing these images is error-prone. Web images are inherently "weakly supervised" in at least two ways: first, except for image-level labels, there are no additional annotations such as bounding boxes or part attributes associated with web images; second, images acquired from the web are relatively noisy. For instance, query results for the word *"nighthawk"* will include images related to a bird, an aircraft, and a comic. As a result, there is no guarantee that their labels will be correct. In spite of this, we propose that it is possible to enrich the weak supervision by exploiting knowledge transferred from existing strongly supervised datasets.

In this paper, we propose a new method for fine-grained categorization that learns robust CNN feature representations and employs detailed object part annotations in a unified framework, and overcome the lack of training data with the help of weakly supervised web images. Our method relies on a joint formulation that iteratively updates CNN feature representations and part-based object classifiers. Specifically, we introduce accurate part annotations from existing strongly supervised datasets and transfer learned perceptual representations to a large-scale auxiliary dataset collected from the web. The detected part patches from weakly supervised web images are used as additional training data to produce more powerful feature representations by fine-tuning CNNs. Based on the new features, a multi-instance formulation is defined to jointly train the final classifiers on the strong and weak datasets. Since we employ part-based R-CNN [33] to train object detectors and classifiers on strongly supervised datasets, we term the proposed method as "augmented part-based R-CNN" (*AP-RCNN*).

Our method is also related to the strategy of training with auxiliary data sources, such as domain adaptation [16] for heterogeneous data sources and incremental learning [6, 30] for homogeneous ones. Some widely used object recognition approaches, such as employing pre-trained CNNs [11, 25] and category-independent object proposals [28], can also be regarded as special cases of this strategy. Here, our unique contribution is that we improve classification performance from two perspectives simultaneously: (i) from the modeling perspective, abundant and diverse data are used to train robust object classifiers; and (ii) from the representation perspective, additional training resources prevent significant overfitting when training CNNs on small sets of strongly supervised data.

Our preliminary results demonstrate the effectiveness of this approach. Using an auxiliary weakly supervised dataset acquired from Flickr, we achieve a classification accuracy of 84.6% on the CUB200-2011 dataset, which is a significant improvement over current state-of-the-art results. Further investigations show that by re-fine-tuning CNNs with more training data, the resultant feature representations significantly contribute to the performance boost.

The remainder of the paper is organized as follows. R-CNN and part-based R-CNN are reviewed in Section 2. The proposed method is described in Section 3. Detailed performance studies and analysis are conducted in Section 4, and we conclude in Section 5.

2. Part-based R-CNN

Whilst our method is agnostic to the specific form of part annotations and CNN architectures, we exploit part-based R-CNN [33] as the basic method for detecting object parts and for training fine-grained classifiers on strongly supervised datasets. In this section, we briefly review R-CNN detectors [15] and part-based R-CNN classifiers and emphasize our modifications that facilitate the proposed method.

2.1. Preliminary

We start with a strongly supervised dataset $\mathcal{S}$, in which ground-truth bounding box annotations are provided not only for the entire objects p_0 but also for a set of n semantic parts $\{p_1, p_2, ..., p_n\}$. Assume that there are K fine-grained categories in the dataset. Selective search [28] is used to extract category-independent object proposals. Typically 1000-2000 region proposals are generated per image.

2.2. R-CNN detectors

Given part annotations in the strongly supervised dataset, at the training stage, the whole object and each of the parts are treated as independent categories and a generic object detector is trained for each. The part detectors are then used during testing to localize object parts; in addition, they are exploited to collect additional part patches from weakly supervised images in the auxiliary dataset.

Our part detector training process follows R-CNN [15]. Specifically, for each of the object parts p_i (or the whole object p_0), we extract deep convolutional features $\phi^{(i)}(x)$ on the extracted region proposals. Starting from a CNN pre-trained on ImageNet [19], a part-CNN is fine-tuned on the target task of fine-grained object recognition to obtain the feature extractor. In particular, we replace the CNN's ImageNet-specific 1000-way classification layer with a randomly initialized $(K+1)$-way layer that accounts for all the fine-grained categories and also a background class. Object proposals with ≥ 0.5 intersection-over-union (*IoU*) over the ground-truth bounding boxes are treated as positive examples for that box's class, while the others are regarded as the background. For each object proposal, the tight bounding box is dilated by m pixels (we use $m = 16$) to introduce context information, and all the pixels in the dilated region are warped into a fixed size of 227×227 pixels. The warped regions are then used as the input to fine-tune the network by stochastic gradient descent (SGD), starting at a learning rate of 0.001. As a result, the learned CNNs (we call them part-CNNs) carry specific domain knowledge of the fine-grained categorization, while not clobbering the initialization from large-scale ImageNet pre-training.

Based on the fine-tuned CNNs, a linear SVM with a binary output is further trained to obtain the final part detector, which only uses ground-truth boxes as positive samples in order to achieve accurate detection results. In our implementation, we train SVMs beyond features extracted from the *fc7* layer of CNNs and adopt a standard hard negative mining method [13] to fit the training data into memory.

Note that our detector training approach is different from part-based R-CNN, in which part CNNs are fine-tuned using a background-absent K-way *fc8* classification layer, and ground-truth crops are the only input to fine-tune the CNN architecture. We show in later experiments that this modification boosts the detection accuracy by 5 to 10 percent.

Denote $\{v_0, v_1, ..., v_n\}$ as the weights of R-CNN detectors for whole-object p_0 and n parts $p_i|_{i=1}^n$. For a region proposal x, the corresponding detector scores $\{d_0, d_1, ..., d_n\}$ are computed as

$$d_i(x) = \sigma(v_i^T \phi^{(i)}(x)), \tag{1}$$

where $\sigma(\cdot)$ is the sigmoid function and $\phi^{(i)}(x)$ is the descriptor at location x according to the i-th part-CNN.

2.3. Part-based R-CNN classifiers

The next step is to integrate the learned R-CNN detector results and use them to train fine-grained classifiers. In part-based R-CNN, Zhang *et al.* [33] proposed three types of geometric constraint to ensure that the relative location of detected objects and their semantic parts follow a geometric prior. Here, however, the strength and robustness of R-CNN part detectors result in geometric constraints that only play a minor role in detection, especially considering that fine-grained datasets usually contain only a relatively limited number of training images. Therefore, in our implementation, we only conduct a simple box constraint to ensure object parts do not fall outside the root bounding box.

For an image I, let $X = \{x_0, x_1, ..., x_p\}$ be the predicted locations (bounding boxes) of an object and its parts, which are given during training, but unknown for both weakly supervised images and testing images. The final feature representation is then denoted as $\Phi(x) = [\phi^{(0)}(x_0), ..., \phi^{(n)}(x_n)]$, where $\phi^{(i)}(x_i)$ is the feature representation for part p_i as the output of the *fc7* layer of the i-th part-CNN. Beyond them, a one-versus-all linear SVM is trained for each fine-grained category. The classification score for an image I being class k is then calculated as:

$$s(I; k) = \sum_{i=0}^{n} {w_k^{(i)}}^T \phi^{(i)}(x_i), \tag{2}$$

where $w_k^{(i)}$ is the classifier weights for class k on features extracted from the i-th object part. The framework of Part-based R-CNN is illustrated as the first row of Figure 2.

3. Augmented part-based R-CNN

Part-based R-CNN has shown promising performance on fine-grained recognition tasks. However, the rapid evolution of CNN architectures involving an increasing number of model parameters has meant that current fine-grained datasets, especially datasets with strong supervision, are too small for training robust CNN representations. We propose to solve this problem by introducing an easy-to-acquire auxiliary dataset to provide additional resources for training part-CNNs.

Specifically, based on the strongly supervised dataset $\mathcal{S}$, an auxiliary dataset containing the same fine-grained categories is collected, but with only image-level labels. Images can be collected from search engines or online media sharing communities. Since the data acquirement process does not require human labeling, the weakly supervised dataset (termed $\mathcal{W}$) typically contains a larger number of images than $\mathcal{S}$. Denote the size of the datasets as $N_\mathcal{S}$ and $N_\mathcal{W}$.

We consider a joint optimization algorithm that updates feature representations ϕ and model parameters w iteratively on a combination of strongly supervised dataset $\mathcal{S}$ and weakly labeled data $\mathcal{W}$. The overall objective function is defined as:

$$\min_{w,\phi} \sum_{p=0}^{n} L(w^{(p)}, \phi^{(p)}),$$

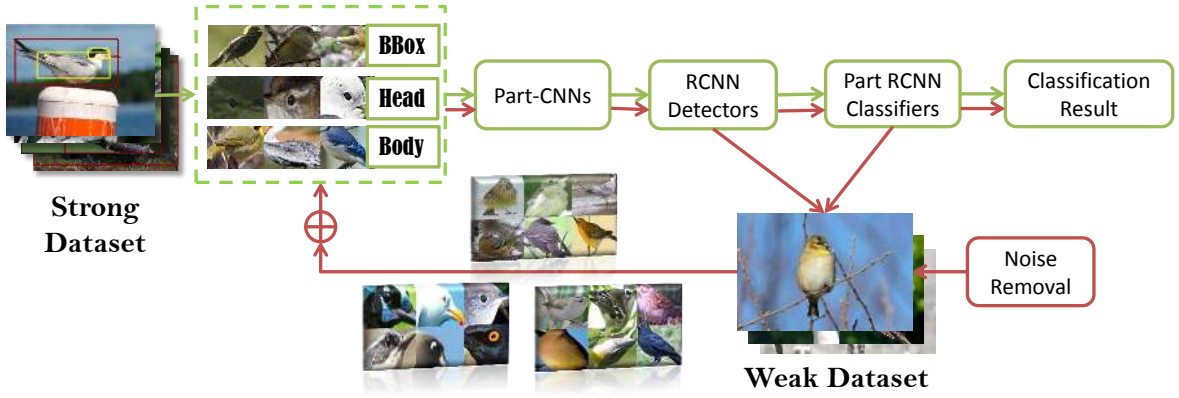

Figure 2. Flowchart of the proposed algorithm. Green lines show modules of part-based R-CNN method [33], while red lines are additional operations conducted in this paper. Better viewed in color.

where

$$L(w^{(p)}, \phi^{(p)}) = \lambda \cdot \Omega(w^{(p)}) + \frac{1}{N_{\mathcal{W}}} \sum_{I \in \mathcal{W}} q_I^{(p)} \cdot l(y_I, \max_{x_p \in X_I} {w_{y_I}^{(p)}}^T \phi^{(p)}(x_p)) + \frac{1}{N_{\mathcal{S}}} \sum_{I \in \mathcal{S}} l(y_I, {w_{y_I}^{(p)}}^T \phi^{(p)}(x_p)) \quad (3)$$

The first term is a l2-norm regularizer, and λ is a free parameter. The second and third term capture the loss on weakly supervised images and strongly supervised images respectively. Here $w_k^{(p)}$ stands for classifier weights of the k-th category. For each image I, a softmax loss l is computed based on the the ground truth label $y_I \in [1, ..., K]$ and the predicted result given feature representation $\phi^{(p)}(\cdot)$ and part location x_p. For the auxiliary weak images, we introduce a multi-instance formulation where X_I is the set of candidate bounding boxes; $q_I^{(p)}$ denotes an indicator of whether the detected region of p-th part in weakly supervised image I is selected to augment the training set, in order to account for label noise. We will detail it in Section 3.3.

As shown in Figure 2, the proposed augmented part-based R-CNN method involves an initialization on the strongly supervised dataset followed by object part detection in weak images, noise removal, re-fine-tuning CNNs, and final classifier training. We detail these steps below.

3.1. Initialization

The first step of the proposed algorithm is to initialize feature representations, part detectors, and object classifier weights based on the strongly supervised dataset using the part-based R-CNN approach detailed in Section 2. In particular, given n object parts and a root, and K fine-grained categories to be classified, the initialization step obtains:

- $n+1$ independently fine-tuned part-CNNs with $(K+1)$-way classification layers as the initialized feature extractors. We use the *fc7* layer to obtain a 4096-dimensional feature vector $\phi^{(i)}$ for each part p_i.
- $n+1$ R-CNN detectors. Each part (or root) p_i is associated with an R-CNN detector d_i based on the respective CNN feature extractor $\phi^{(i)}$.
- $K(n+1)$ sets of classification model weights, with each $w_k^{(i)} \in \mathbb{R}^{4096 \times 1}$.

3.2. Part discovery

The initialized feature representations are obtained using strong annotations including object bounding boxes and part localizations. However, such annotations are unavailable in the auxiliary dataset, since these images are only associated with image-level labels. Therefore, in order to exploit the auxiliary dataset, part-level information needs to be generated for the weakly supervised images: here, we achieve this by discovering part patches from weakly supervised images using the learned R-CNN part detectors.

The part detectors provide top-down messages to select relative patches with high discriminative power for classification. After obtaining detecting scores for all the parts, we adopt the box constraint restriction in part-based R-CNN to introduce geometric relations between object parts. The detected locations $X^* = \{x_0, ..., x_n\}$ are given as:

$$X^* = \arg\max_X \prod_{i=1}^{n} c_{x_0}(x_i) \prod_{i=0}^{n} d_i(x_i), \quad (4)$$

where

$$c_x(y) = \begin{cases} 1, \text{if region } y \text{ falls outside } x \text{ by at most 10 pixels} \\ 0, \text{otherwise} \end{cases}$$

3.3. Noise removal

As well as the lack of part-level annotations, web images are also "weakly supervised" due to label noises: it is not

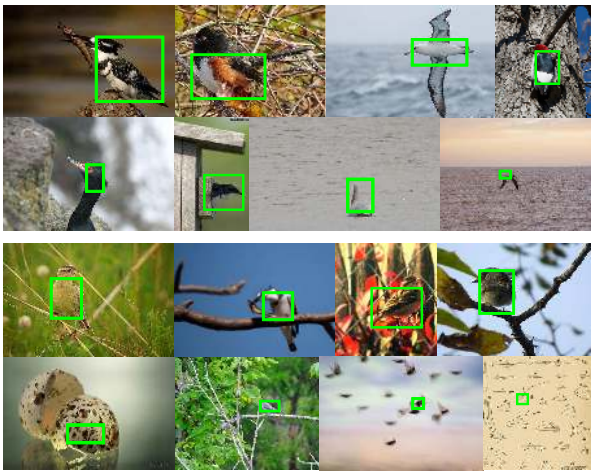

Figure 3. Detection results on weakly supervised images. Green frames indicate the detected bounding box for part "body". Image labels in the top two rows are correctly classified; the bottom two rows show cases in which classification has failed. Beyond the classification results, part patches in rows 1 and 3 are associated with high detection scores, while rows 2 and 4 have low detection scores. We propose to use both classification and detection scores to select valid part patches to augment the training data.

guaranteed that images in the auxiliary dataset are all related to the fine-grained categories. Therefore, we introduce a noise removal process to clean up the detected part patches.

In the context of generating part patches from weakly supervised images, the strategy of selecting proper patches can be defined in two ways: (i) a sample should be selected if we are confident about the correctness of detected localization; or (ii) a sample should be selected if it is easy to predict its true label. We argue that, in our task, adopting these strategies individually is unlikely to produce optimal results. As shown in Figure 3, images that are correctly classified do not always generate valid part patches due to occlusion effects and the absence of a particular object part. On the other hand, there is no clear boundary to perfectly separate "good" detections from "poor" detections with respect to detection scores.

We therefore propose a two-threshold strategy that combines detection scores and classification results to select valid part patches. The basic idea is to flexibly adjust the threshold of "good" detections by setting a loose condition on the correctly classified images and requiring harsher terms for misclassified images. Specifically, the criterion of whether a part patch x is selected to augment the training set is determined as an indicator $q_I^{(i)} = \mathcal{I}(d_i(x) > \lambda)$ where

$$\lambda = \begin{cases} \lambda_{pos}, & \text{if } \tilde{y_I} = y_I \\ \lambda_{neg}, & \text{if } \tilde{y_I} \neq y_I \end{cases}, \tag{5}$$

Here y_I is the label of image I and $\tilde{y_I}$ is the predicted label obtained by part-based R-CNN classifiers. We set two thresholds for detection scores $d_i(x)$, where $\lambda_{pos} < \lambda_{neg}$. The two thresholds are defined as:

$$\begin{aligned} \lambda_{pos} &= \sigma \bar{d}_i(neg) \\ \lambda_{neg} &= \sigma \bar{d}_i(pos), \end{aligned} \tag{6}$$

where $\bar{d}_i(\cdot)$ is the average detection score of part patches over correctly or incorrectly classified images, σ is a free parameter. The resultant threshold λ_{pos} is guaranteed to be lower than λ_{neg} because successfully detected part patches would always contribute to classification performance.

3.4. Re-fine-tuning CNNs

We employ R-CNN part detectors trained using strong supervisions and a two-threshold denoising process to generate discriminative part patches from the weakly supervised dataset. These part patches, in addition to the strongly supervised training data, are used to generate better feature representations by re-fine-tuning the part-CNNs. We use the same CNN architecture as discussed in Section 2.2, and once again randomly initialize the $(K+1)$-way *fc8* layer with the filter weights of previous layers kept fixed. All region proposals that have ≥ 0.5 *IoU* over the detected part bounding boxes are cropped, dilated, warped and then fed into the CNN architecture as input. Re-fine-tuning the $n+1$ part-CNNs actually serves as an updating procedure of the feature representation ϕ in (3).

3.5. Final classifier

Having updated the feature representations and detected part locations on weakly supervised images, the model parameters w are jointly retrained on the strong and weak datasets to obtain the final object classifiers. Inspired by [16], we define a multi-instance learning (MIL) formulation [31] that includes bags defined on both types of images. Specifically, for each image in the auxiliary set, the top 10 locations of the root bounding box are detected, each of which is regarded as an instance in MIL. The objective function (3) is rewritten as:

$$\begin{aligned} L(w) &= \lambda\Omega(w) + \frac{1}{N_{\mathcal{S}}} \sum_{I \in \mathcal{S}} l(y_I, w_{y_I}^T \Phi(x)) \\ &+ \frac{1}{N_{\mathcal{W}}} \sum_{I \in \mathcal{W}} l(y_I, \max_{x \in X_I} w_{y_I}^T \Psi(x)), \end{aligned} \tag{7}$$

where $w = [w^{(0)}, ..., w^{(n)}]$ denotes the joint model classifier; $\Phi(x) = [\phi^{(0)}(x_0), ..., \phi^{(n)}(x_n)]$ is the part-based R-CNN feature representation for a strongly supervised image; $\Psi(x) = [q_I^{(0)}\phi^{(0)}(x_0), ..., q_I^{(n)}\phi^{(n)}(x_n)]$ is the feature representation for a weakly supervised image, in which a part filter p is set to a zero vector if the indicator $q^{(p)}$ is zero.

The objective can be solved by standard MIL methods with slight modifications (see Supplementary).

Although the whole process can be undergo several rounds of iteration, in practice a single round of feature representation and object classifier updating already produces promising results. Due to the ensuing time complexity, further iterations of the whole pipeline are not performed.

At the testing stage, for a new test image, we apply the whole object and part detectors with the box geometric constraint to localize object parts; the features of all parts are then concatenated into the final feature vector for prediction. No additional annotations are required during testing.

4. Experiments

We present experimental results and analysis of the proposed method in this section. Specifically, we will describe the acquirement of weakly supervised web images, the effectiveness of R-CNN detectors, discuss factors on classification results, and visualize learned part-CNNs.

4.1. Dataset and implementation details

Experiments were conducted on two widely used fine-grained classification benchmarks: the Caltech-UCSD Birds dataset [29] (CUB200-2011) and the Oxford-IIIT Pet Dataset [24] (PET). The CUB200-2011 dataset contains 11,788 images of 200 types of bird, in which 30 images per category are used for training. The dataset is strongly supervised, *i.e.*, images are associated with detailed annotations including image-level labels, object bounding boxes, and part landmarks. Following the protocol of [34, 33], we exploited the location annotation of two semantic parts, head and body, along with whole object bounding boxes to conduct part-based models. The PET dataset contains 37 cat and dog breeds, with roughly 200 images per category. Ground truth object and head bounding boxes were exploited as strong supervisions. We followed the provided train/test split in both datasets.

An auxiliary weakly supervised dataset was collected to augment the strongly supervised data. Images were obtained from Flickr[1] by conducting image searches using the names of the 200 bird species or 37 pet breeds as queries. For each category, the top 100 images for CUB and 200 images for PET were downloaded sorted by upload time to ensure no overlap between the crawled images and test images in the datasets. No further manual filtering process was conducted on the auxiliary dataset. These downloaded images only had image-level labels, which were not always correct due to the ambiguity of query words and label noise.

We used the open-source package Caffe [17] to extract deep features and fine-tune part-CNNs from the Caffe reference model, implementing [19]. We used the *fc7* layer in the CNN architecture to train R-CNN detectors and in image representation for classification.

[1] www.flickr.com

4.2. Detection results and analysis of discovered part patches

One of the key assumptions of our method is that the use of detectors learned from strongly supervised data can effectively detect and locate object part patches in the auxiliary weakly supervised images. Therefore, analysis commenced by evaluating detection results and studying the discovered part patches.

The quantitative detection results were measured in terms of the "Percentage of Correctly localized Parts" (PCP) metric on the test set. A part patch was marked as correctly localized if the predicted bounding box had ≥ 0.5 overlap with the ground-truth bounding box.

	BBox	Head	Body
Strong DPM [2]	-	37.44%	47.08%
Part R-CNN [33]	-	61.94%	70.16%
Ours	92.84%	70.89%	75.79%

Table 1. Part localization accuracy in terms of PCP on the CUB200-2011 dataset.

The learned R-CNN detectors produced reasonable results, achieving greater than 70% PCP for all parts (Table 1). The improvement over part-based R-CNN [33] is due to the additional negative mining process and from assigning the background as the $(K+1)$-th category for fine-tuning part-CNNs (as specified in [15]).

The high-performing part detectors ensure that a large number of part patches can be discovered on the auxiliary dataset. However, since the parameter-rich CNN architectures can easily overfit the training data, it is critical to find a balance between adding more training data and ensuring clean labels. Hence, we used the noise removal approach discussed in Section 3.3, with $\sigma = 0.5$ working well in practice. The process generated 15,840, 15,397, and 15,751 patches for the whole object, part "head", and part "body", respectively for the CUB dataset. These part patches were then used to re-fine-tune part-CNNs. Example detected patches from the auxiliary dataset are shown in Figure 4.

4.3. Classification results

Since our method involves multiple steps to boost classification performance, we first analyze the effect of each step by detailed comparison with the baselines shown in Table 2.

Feature Perspective. The first set of comparisons reveal that improved feature representations by fine-tuning CNNs on domain-specific data significantly contribute to classification accuracy. Directly exploiting an ImageNet pre-trained CNN as the feature extractor achieved an accuracy of 68%. Fine-tuning part-CNNs on the bird training

Figure 4. Examples of detected part patches from weakly supervised images selected to augment the CUB200-2011 training set. From top to bottom: whole object, head, body. The leftmost five columns show top-scoring detections, while the right two columns show patches with the lowest detection scores.

Part Localization	Predict BBox		Oracle
CNN\SVM	Train	Train+Weak	Train
w/o ft	68.58%	71.19%	74.14%
ft on train	78.56%	79.89%	82.12%
ft on train/weak	81.17%	82.16%	85.07%
ft on train/weak-dn	83.24%	84.59%	86.57%

Table 2. Baseline comparisons. Rows indicate different methods for fine-tuning part-CNNs; columns show results of training fine-grained object classifiers on the augmented dataset or only on the strongly supervised data. ft stands for fine-tuning; dn for denoising. Oracle method uses ground-truth part annotations at testing time; thus casts as an upperbound of the classification results.

set improved this result by a large margin to 79%. Furthermore, by augmenting the part patches by performing part discovery on the weak dataset and re-fine-tuning CNNs, a further improvement to 81% classification accuracy was obtained. These results show that the larger amount of training data does indeed improve the discriminative power of the learned CNN representation. Denoising of on the weak dataset further improved the accuracy by 2%.

Model Perspective. It is argued that even without using CNN features, employing additional training data can boost classification results by increasing data diversity in training examples. We studied this factor by re-training part-based R-CNN classifiers on the augmented dataset and comparing the results to those trained on strongly supervised training data only. Results showed that when the feature representations were fixed (as in traditional features such as SIFT), the performance improvement was trivial ($\sim 1\%$) compared to re-fine-tuning CNN features. This reveals an interesting phenomenon that feature representation plays a greater role in fine-grained object recognition than model training. The proposed method of training classifiers on the re-fine-tuned part-CNN features finally delivers 84.6% accuracy.

Localization Accuracy. The accuracy of part localization also has a large impact on the final classification results. Although R-CNN detectors obtain reasonable detection accuracy for object parts, an average 3% gap still remains between classification results using predicted bounding boxes and the oracle method, which casts as an upper bound of classification performance by employing ground-truth part annotations during both training and testing. It is worth noting that our final classification result of 84.6% after introducing weakly supervised samples exceeds even the upper bound accuracy of 82.1% when using strongly supervised training data alone.

Method	Train BBox	Train Part	Test BBox	Acc(%)
DPD [34]	✓		✓	51.0
DeCAF$_6$ [11]	✓		✓	58.8
Symbiotic [8]	✓		✓	61.0
CNNaug [25]	✓		✓	61.8
Alignment [14]	✓		✓	67.0
POOF [4]	✓	✓	✓	56.8
Part R-CNN [33]	✓	✓		73.9
PoseNorm CNN [7]	✓	✓		75.7
Our method	✓	✓		**84.6**

Table 3. Accuracy comparison on the CUB200-2011 dataset.

Method	Accuracy(%)
Angelova *et al.* [1]	54.3
Murray *et al.* [23]	56.8
Azizpour *et al.* [3]	88.1
Our Method (Strong only)	86.1
Our Method (Strong+Weak)	**88.2**

Table 4. Accuracy comparison on the Oxford-IIIT Pet Dataset.

The comparison of accuracies between the proposed method and state-of-the-art methods on CUB200-2011 is shown in Table 3. Unlike most of the literature on this dataset, we consider it more realistic that the birds' bounding boxes are unknown during testing. In this challenging setting, we achieved an accuracy of 84.6%, which represents a remarkable improvement over existing state-of-the-art methods. Table 4 shows comparison results on the PET dataset. Again the proposed method obtains promising results, being comparable to [3] who used deeper network architectures. Although our method requires additional training data by collecting weakly supervised images from the web, this data acquisition process is easy to implement and requires no additional human labeling effort. Meanwhile, with additional training samples, our method is likely to achieve better performance with more complicated CNN architectures such as the VGGNet [27].

4.4. Visualization

Beyond the quantitative results presented above, here we present a more intuitive description of how our method

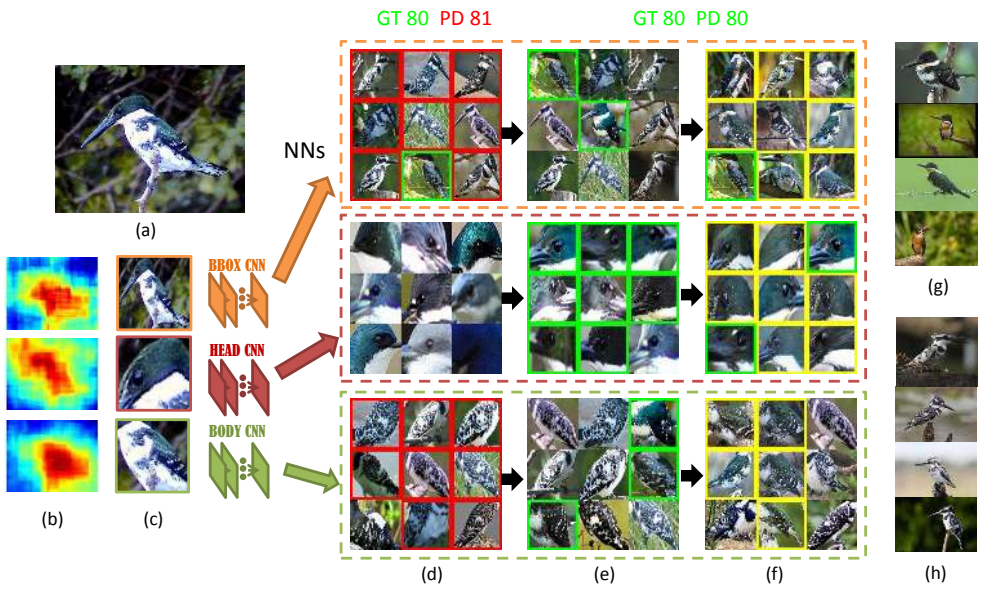

Figure 5. Visualization of the classification process using the proposed method with a root and two parts: head and body. (a) Test image with a ground-truth label of 80. (b) Activation map for the three detectors. (c) Located part bounding boxes. The top 9 nearest neighbours for the detected parts from the training images are shown in (d)-(f). The original part-based R-CNN method using training data only misclassified the test image into class 81, as shown in (d). Green boxes demonstrate the image patches of label 80, and red boxes for label 81. After re-fine-tuning part-CNNs with the augmented training set, the new feature representations guaranteed that the test image was correctly classified. (e) Nearest neighbours from the strongly supervised training set only using the new feature representations. (f) Results after putting weakly supervised images into the training set either. Yellow boxes indicate images in the weakly supervised dataset with label 80. (g) and (h) show typical training images from class 80 (*Green_Kingfisher*) and 81 (*Pied_Kingfisher*) respectively.

works on practical examples. The procedure of classifying a fine-grained image using the proposed method is shown in Figure 5. Given a test image (a) belonging to *Green_Kingfisher*, R-CNN detectors were used to localize the object and its semantic parts, detailed in (b) and (c). As shown in (d), the original part-based R-CNN method misclassified the image into a very similar subcategory *Pied_Kingfisher*. Closer inspections reveal that the bird in the test image indeed belongs to a rare occurring subclass in the category in which black and white spots decorate the chest. Unfortunately, the strongly supervised dataset does not include sufficient training data for this subclass.

We solved this problem by introducing an auxiliary dataset of weakly supervised images collected from the web to augment the training data. As shown in (e), the new feature representations obtained by re-fine-tuning part-CNNs on the augmented training set improved the discriminative power in this case, especially for the bird's head, even when only images in the strongly supervised dataset were employed to train the object classifiers. Naturally, inserting weakly supervised images into the training set also contributed to the classification process. Nearest neighbors shown in (f) indicated that in the auxiliary dataset, there were a larger number of images similar to the test image, making the classification result more convincing.

5. Conclusion

In this paper, we present a new fine-grained recognition method that trains robust CNN feature extractors with effective part-based models by employing the availablity of vast numbers of online images to augment manually-labeled strongly supervised datasets. Our method acts as a bridge between the requirement for extensive data to train deep representations and the difficulty in obtaining large-scale strongly annotated datasets. Experiments on two benchmark datasets show that introducing additional weakly supervised images leads to an impressive improvement over baseline methods and achieves state-of-the-art results. We believe the proposed method is likely to be useful in practice, especially considering that the forms of part annotations are varied and CNN architectures are becoming more complicated over time.

Acknowledgement

The work is partially supported by the High Technology Research and Development Program of China (2015AA015801), NSFC (61221001), STCSM (12DZ2272600), the 111 Project (B07022), and the Australian Research Council Projects DP-140102164 and FT-130101457.

References

[1] A. Angelova and S. Zhu. Efficient object detection and segmentation for fine-grained recognition. In *CVPR 2013*, pages 811–818. IEEE, 2013. 7

[2] H. Azizpour and I. Laptev. Object detection using strongly-supervised deformable part models. In *ECCV 2012*, pages 836–849. Springer, 2012. 6

[3] H. Azizpour, A. S. Razavian, J. Sullivan, A. Maki, and S. Carlsson. From generic to specific deep representations for visual recognition. *arXiv:1406.5774*, 2014. 7

[4] T. Berg and E. N. Belhumeur. Poof: Part-based one-vs.-one features for fine-grained categorization, face verification, and attribute estimation. In *CVPR 2013*, pages 955–962. IEEE, 2013. 1, 7

[5] L. Bo, X. Ren, and D. Fox. Kernel descriptors for visual recognition. In *NIPS*, pages 244–252, 2010. 1

[6] S. Branson, P. Perona, and S. Belongie. Strong supervision from weak annotation: Interactive training of deformable part models. In *ICCV*, pages 1832–1839. IEEE, 2011. 2

[7] S. Branson, G. Van Horn, S. Belongie, and P. Perona. Bird species categorization using pose normalized deep convolutional nets. *arXiv:1406.2952*, 2014. 7

[8] Y. Chai, V. Lempitsky, and A. Zisserman. Symbiotic segmentation and part localization for fine-grained categorization. In *ICCV 2013*, pages 321–328. IEEE, 2013. 1, 7

[9] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR 2009*, pages 248–255. IEEE, 2009. 1

[10] J. Deng, J. Krause, and L. Fei-Fei. Fine-grained crowdsourcing for fine-grained recognition. In *CVPR 2013*, pages 580–587. IEEE, 2013. 1

[11] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. *arXiv:1310.1531*, 2013. 1, 2, 7

[12] K. Duan, D. Parikh, D. Crandall, and K. Grauman. Discovering localized attributes for fine-grained recognition. In *CVPR 2012*, pages 3474–3481. IEEE, 2012. 1

[13] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *TPAMI*, 32(9):1627–1645, 2010. 3

[14] E. Gavves, B. Fernando, C. G. Snoek, A. W. Smeulders, and T. Tuytelaars. Fine-grained categorization by alignments. In *ICCV 2013*, pages 1713–1720. IEEE, 2013. 7

[15] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR 2014*, pages 580–587. IEEE, 2014. 2, 3, 6

[16] J. Hoffman, D. Pathak, T. Darrell, and K. Saenko. Detector discovery in the wild: Joint multiple instance and representation learning. *arXiv:1412.1135*, 2014. 2, 5

[17] 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:1408.5093*, 2014. 6

[18] A. Khosla, N. Jayadevaprakash, B. Yao, and F.-f. Li. L.: Novel dataset for fine-grained image categorization. In *CVPRW 2011*. Citeseer, 2011. 1

[19] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1097–1105, 2012. 1, 3, 6

[20] Z. Li, E. Gavves, T. Mensink, and C. G. Snoek. Attributes make sense on segmented objects. In *ECCV 2014*, pages 350–365. Springer, 2014. 1

[21] J. Liu, A. Kanazawa, D. Jacobs, and E. Belhumeur. Dog breed classification using part localization. In *ECCV 2012*, pages 172–185. Springer, 2012. 1

[22] S. Maji, E. Rahtu, J. Kannala, M. Blaschko, and A. Vedaldi. Fine-grained visual classification of aircraft. *arXiv:1306.5151*, 2013. 1

[23] N. Murray and F. Perronnin. Generalized max pooling. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 2473–2480. IEEE, 2014. 7

[24] O. M. Parkhi, A. Vedaldi, A. Zisserman, and C. Jawahar. Cats and dogs. In *CVPR 2012*, pages 3498–3505. IEEE, 2012. 6

[25] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. In *CVPRW 2014*, pages 512–519. IEEE, 2014. 2, 7

[26] E. Rosch, C. B. Mervis, W. D. Gray, D. M. Johnson, and P. Boyes-Braem. Basic objects in natural categories. *Cognitive psychology*, 8(3):382–439, 1976. 1

[27] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv:1409.1556*, 2014. 7

[28] J. R. Uijlings, K. E. van de Sande, T. Gevers, and A. W. Smeulders. Selective search for object recognition. *IJCV*, 104(2):154–171, 2013. 2

[29] C. Wah, S. Branson, E. Welinder, P. Perona, and S. Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 1, 2, 6

[30] T. Xiao, J. Zhang, K. Yang, Y. Peng, and Z. Zhang. Error-driven incremental learning in deep convolutional neural network for large-scale image classification. In *ACM MM*, pages 177–186. ACM, 2014. 2

[31] Z. Xu, D. Tao, Y. Zhang, J. Wu, and A. C. Tsoi. Architectural style classification using multinomial latent logistic regression. In *ECCV 2014*, pages 600–615. Springer, 2014. 5

[32] S. Yang, L. Bo, J. Wang, and L. G. Shapiro. Unsupervised template learning for fine-grained object recognition. In *NIPS*, pages 3122–3130, 2012. 1

[33] N. Zhang, J. Donahue, R. Girshick, and T. Darrell. Part-based r-cnns for fine-grained category detection. In *ECCV 2014*, pages 834–849. Springer, 2014. 1, 2, 3, 4, 6, 7

[34] N. Zhang, R. Farrell, F. Iandola, and T. Darrell. Deformable part descriptors for fine-grained recognition and attribute prediction. In *ICCV 2013*, pages 729–736. IEEE, 2013. 1, 6, 7

[35] N. Zhang, M. Paluri, M. Ranzato, T. Darrell, and L. Bourdev. Panda: Pose aligned networks for deep attribute modeling. In *CVPR 2014*, pages 1637–1644. IEEE, 2014. 2