Subset Feature Learning for Fine-Grained Category Classification

ZongYuan Ge[†‡], *Christopher McCool*[‡], *Conrad Sanderson*[*◇], *Peter Corke*[†‡]

[†] Australian Centre for Robotic Vision, Brisbane, Australia
[‡] Queensland University of Technology (QUT), Brisbane, Australia
[*] University of Queensland, Brisbane, Australia
[◇] NICTA, Australia

Abstract

Fine-grained categorisation has been a challenging problem due to small inter-class variation, large intra-class variation and low number of training images. We propose a learning system which first clusters visually similar classes and then learns deep convolutional neural network features specific to each subset. Experiments on the popular fine-grained Caltech-UCSD bird dataset show that the proposed method outperforms recent fine-grained categorisation methods under the most difficult setting: no bounding boxes are presented at test time. It achieves a mean accuracy of 77.5%, compared to the previous best performance of 73.2%. We also show that progressive transfer learning allows us to first learn domain-generic features (for bird classification) which can then be adapted to specific set of bird classes, yielding improvements in accuracy.

1. Introduction

Deep convolutional neural networks (CNNs) have been successful in various computer vision tasks. Deep CNNs have achieved impressive in both general [18, 22, 9] and fine-grained image classification [26, 13]. Recently, deep CNN approaches have been shown to surpass human performance for the task of recognising 1000 classes from the ImageNet dataset [16]. Although deep CNNs can serve as an end-to-end classifier, they have been used by many researchers as a feature extractor for various recognition problem including segmentation [15] and detection [14].

Recently, the task of fine-grained image categorisation has received considerable attention, in particular the task of fine-grained bird classification [26, 3, 7, 10, 12]. Fine-grained image classification is a challenging computer vision problem due to subtle differences in the overall appearance between various classes (low inter-class variation) and large pose and appearance variations in the same class (large intra-class variation).

Much of the work for fine-grained image classification has dealt with the issue of detecting and modelling local parts. Several researchers have examined methods to find local parts and extract normalised features in order to overcome the issues of pose and view-point variation [5, 7, 20, 27, 9]. Aside from the issue of pose and viewpoint changes, a major challenge for any fine-grained classification approach is how to distinguish between classes that have high visual correlations [3]. Some state-of-the-art pose normalised methods still have considerable difficulty in categorising some visually similar fine-grained classes [26, 6].

To date, there has been limited work which investigates in detail how best to learn deep CNN features for the fine-grained classification problem. Most of the methods used off-the-shelf convolutional neural networks (CNNs) features trained from ImageNet or fine-tuned the pre-trained ImageNet model on the target dataset, then using one fully-connected layer as a feature descriptor [17, 22].

This paper examines in detail how to best learn deep CNN features for fine-grained image classification. In doing so, we propose a novel *subset* learning system which first splits the classes into visually similar subsets and then learns domain-specific features for each subset. We also comprehensively investigate progressive transfer learning and highlight that first learning domain-generic features (for bird classification) using a large dataset and then adapting this to the specific task (target bird dataset) yields considerable performance improvements.

2. Related Work

2.1. Convolutional Neural Networks

Krizhevsky et al. [18] recently achieved impressive performance on the ImageNet recognition task using CNNs, which were initially proposed by LeCun et al. [19] for hand writing digit recognition. Since then CNNs have received considerable attention [22, 14]. The network structure of Krizhevsky et al. [18] remains a popular structure and consists of five convolutional layers ($conv1$ to $conv5$) with two fully-connected layers ($fc6$ and $fc7$) followed by a softmax layer to predict the class label. The network is capable of generating useful feature representations by learning low level features in early convolutional layers and accumulating them to high level semantic features in the latter convolutional layers [25].

Figure 1. Birdsnap is a very challenging fine-grained bird dataset with sexual as well as age dimorphisms. There are considerable appearance differences between males and females, as well as between young and mature birds. Each row shows images from the same species. For each bird species there are large intra-class variations: pose variation, background variation and appearance variation.

2.2. Features for Fine-grained Classification

Several approaches have been designed to learn feature representations for fine-grained image classification. Berg et al. [3] generated millions of keypoint pairs to learn a set of highly discriminative features. Zhang et al. [27] learned pose normalised features by using the deformable part descriptors model (DPM) [11] on local parts which were extracted using a pre-trained deep CNN. Chen et al. [8] proposed a framework to select the most confident local descriptors for nonlinear function learning using a linear approximation in an embedded higher dimensional space.

The above feature learning schemes are implicitly part-based methods. This means they require the ground truth locations of each part which limits their usefulness in terms of fully automatic deployment.

3. Proposed Method

Our proposed feature learning method consists of two main parts. First, we perform progressive transfer learning to learn a domain-generic convolutional feature extractor (termed ϕ_{GCNN}) from a large-scale dataset of the same domain as the target dataset. Second, we perform subset-specific feature learning from pre-clustered subsets which contain visually similar fine-grained class images. The discriminative convolutional features learned from the subset learning system is termed $DFCNN$, and the related feature extractor is referred as ϕ_{DFCNN}.

For image $\boldsymbol{I}_i$, we apply the $\phi_{GCNN}(\boldsymbol{I}_i)$ and $\phi_{DFCNN}(\boldsymbol{I}_i)$ and combine them to obtain our feature vector to describe the image. For training the classifier, we employ a one-versus-all linear SVM using the final feature representation.

3.1. Progressive Transfer Learning

It is desirable to have as much as data possible in order to avoid overfitting while training a CNN. A typical CNN has millions of parameters which makes it difficult to train when data is limited. Typically fine-grained image datasets are relatively small compared to the ImageNet dataset. To circumvent problems with small datasets, a process known as transfer learning [24] can be applied. Transfer learning has usually been applied by fine-tuning a general network, such as the network of Krizhevsky et al. [18], to a specific task such as bird classification [26]. Recent work by Yosinski et al. [24] found that better accuracy can be achieved if transfer learning is performed using datasets representing the same or related domains.

Inspired by the findings of Yosinski et al. [24], we propose an alternative approach where a generic CNN is progressively adapted to the task at hand. First, a large dataset, which is related to the same domain as the final task, is used to perform transfer learning. This yields a domain-generic feature representation. Second, a smaller dataset which represents the final task at hand is used to adapt the domain-generic features to yield task-specific features. Our experimental results show that progressive transfer learning yields feature representation which lead to consistently improved performance. Furthermore, we will show that the domain-generic features can also be used effectively for the task at hand.

3.2. Subset Specific Feature Learning

Recent parts-based fine-grained methods show relatively good performance on the Caltech-UCSD bird dataset [23]. The methods are good at recognising birds species with distinguishable features with moderate pose variation. However, many mis-classifications occur for birds species that have similar visual appearance.

To address this issue, we propose to pre-cluster visually similar species into subsets and use subset-specific CNNs. Instead of relying on one CNN to handle all possible cases, each CNN focuses on the differences within each subset. In effect, the overall classifier has more parameters, as all

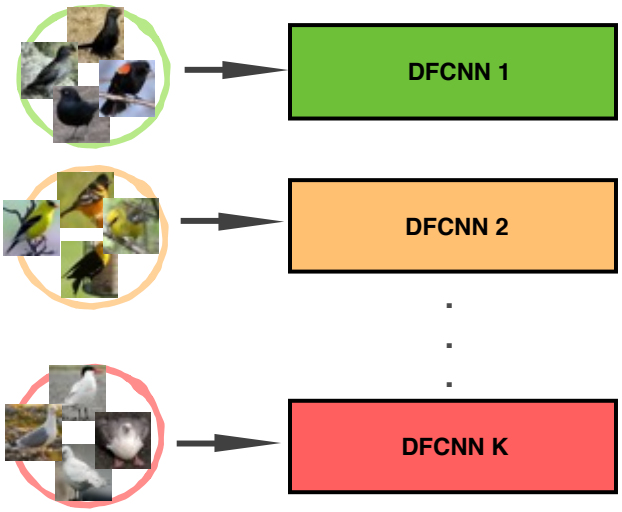

Figure 2. Pre-clustered visually similar images are fed into $DFCNN_{1...K}$ with backpropogation training to learn discriminative features for each subset.

CNNs have the same network architecture. Due to the practical issues such as training time and memory requirements, using separate CNNs dedicated to specific tasks is more practical than having one very large CNN. An overview of this subset learning scheme is shown in Fig. 2.

The above subset feature learning process is initially performed on a large yet related dataset. In particular, we use the large Birdsnap dataset [4] instead of the target Caltech-UCSD dataset [23]. We expect that our learned features are both generalised and discriminative compared to features learned directly on the same size or smaller size target dataset under the same domain.

3.2.1 Pre-clustering

To generate subsets in terms of visually similar images, image representations should focus on colour and texture while being robust to pose and background variations. We investigate three types of features as image representers. Features are obtained from either the 5-th layer $conv5$ or the 6-th layer ($fc6$) of the CNN. These were selected due to their recent use by other researchers to perform object recognition and clustering [9]. We also apply linear discriminant analysis (LDA) [21] to $fc6$ features to reduce their dimensionality. This is done to ameliorate the well known issues of clustering high dimensional data [1]. The subsets are then obtained via k-means clustering.

Examples of clustering results using the three feature types are shown in Fig. 3. The fully connected layer based feature $fc6$ fits our criteria better than clustering using the the convolutional feature $conv5$ that tends to learn shape and pose information, which is undesirable. This particular property can be seen in clusters 1 and 2 in Fig. 3(a) which

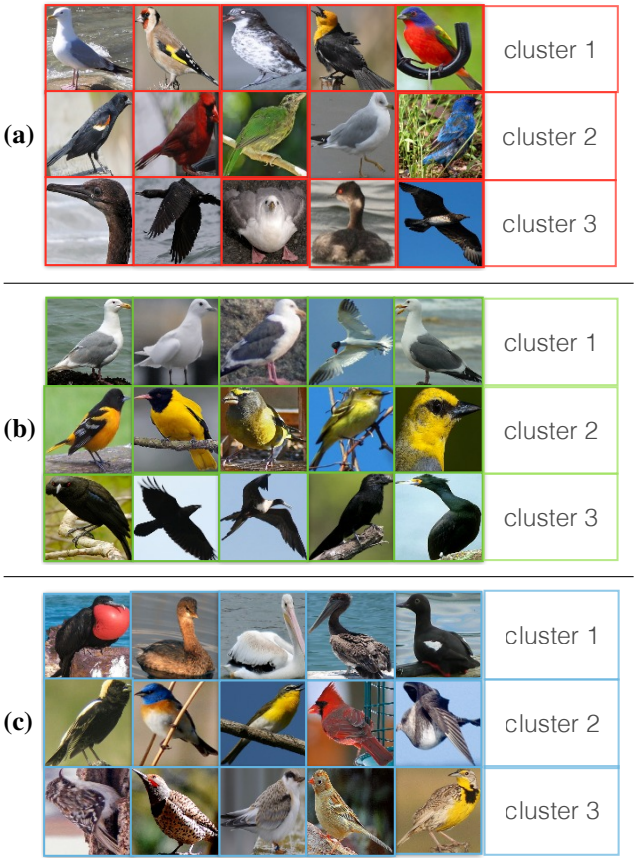

Figure 3. Pre-clustering results using: **(a)** $conv5$ layer features, **(b)** $fc6$ layer features, **(c)** $lda-fc6$ features. Clustering via $conv5$ yields undesirable strong correlations with pose and shape information. Using $fc6$ yields some improvements, but the pose bias is still visibly present. Using $lda - fc6$ provides further clustering improvements in terms of robustness to color and pose variations.

represent right and left pose of birds images while the rest are grouped into cluster 3. We conjecture that this is due to the convolutional based features containing a high degree of spatial information. Using $fc6$ yields some improvements, but the pose bias is still visibly present. Using $lda - fc6$ features provides further clustering improvements in terms of robustness to colour and pose variations.

3.2.2 Subset Feature Learning

A separate CNN is learned for each of the K pre-clustered subsets. The aim is to learn features for each subset that will allow us to more easily differentiate visually similar species. As such, for each subset, we apply transfer learning to the CNN of Krizhevsky et al. [18] (whose structure was described in Section 2). To train the k-th subset ($Subset_k$) we use the N_k images assigned to this subset $\boldsymbol{X}_k = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{N_k}]$, with their corresponding class labels $\boldsymbol{C}_k = [c_1, \ldots, c_{N_k}]$. The number of outputs in the

associated last fully connected layer $fc8$ is set to the number of classes in each subset. Transfer learning is then applied separately to each network using backpropogation and stochastic gradient descent (SGD). We then take $fc6$ to be the learned subset feature ϕ_{DFCNN_k} for the k-th subset.

3.3. Fine-grained Classification

To predict test labels for an image $\boldsymbol{I}_t$, our classification pipeline combines the $\phi_{GCNN}(\boldsymbol{I}_t)$ feature with the K subset features $\phi_{DFCNN_{1\ldots K}}(\boldsymbol{I}_t)$. A max voting rule is used to retain only the most relevant subset-specific feature. The other $K-1$ features are set to $\mathbf{0}$. See Fig. 4 for a conceptual representation. To balance weights for the domain-generic and subset-specific features, both $GCNN$ and $DFCNN$ features are then $l2$ normalised before combining them into a single feature vector. Using this feature vector, we train a one-versus-all linear SVM in order to make predictions.

3.3.1 Max Voting DFCNN

The final feature representation for image $\boldsymbol{I}$ is the concatenation of generalised features obtained from $\phi_{GCNN}(\boldsymbol{I})$ and the K subsets $\phi_{DFCNN_{1\ldots K}}(\boldsymbol{I})$. However, sometimes an image is more relevant to one subset features than others. For example to extract features for a White Gull image, it is more reasonable to use $DFCNN$ features from the subset which has many relevant white birds.

To emphasise the most relevant $DFCNN$, we first learn a **subset selector** to select the most relevant subset (rank 1) to the image. Max voting is then used to retain the feature from the most relevant subset and the remaining $k-1$ subset features are set to 0. One way to interpret the max voting is to use the **subset selector** to learn a binary vector $\boldsymbol{w}$, where $\sum_{i=1}^{K} \boldsymbol{w}_i = 1$. The final subset feature representation is then $DFCNN = [w_1\phi_{DFCNN_1}(x_i), \ldots, w_k\phi_{DFCNN_K}(x_i)]$. We explore two ways to learn the **subset selector**.

The simplest way of learning the **subset selector** is to use the centroids from the pre-clustering; we refer to this as $Cen_{1\ldots K}$. This provides a simple classifier trained in an unsupervised manner, however, given the importance of this stage we explore the use of a discriminatively trained classifier using a CNN.

Another way to select the most relevant subset is to train a separate CNN based subset selector $SCNN$. Using the output from the pre-clustering as the class labels, we learn a new SCNN by changing the softmax layer $fc8$ to have K outputs. The softmax layer now predicts the probability of the test image belonging to a specific subset $Subset_k$, max voting is then applied to this prediction to choose the most likely subset. As with the previously trained CNNs, the weights of $SCNN$ are trained via backpropogation and SGD using the network of Krizhevsky et al. [18] as the starting point.

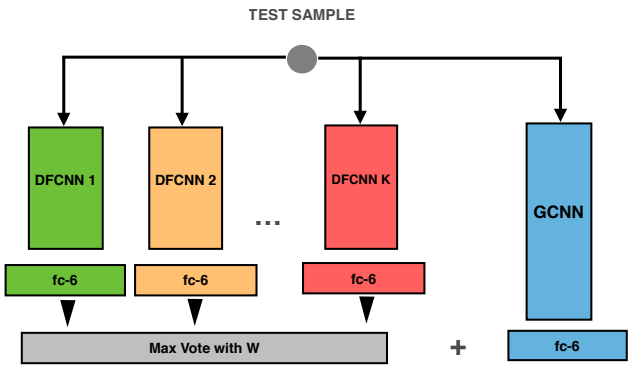

Figure 4. Feature representation of the test image is the concatenated features from both DFCNN with weighting factors and GCNN.

4. Experiments

In this section we present a comparative performance evaluation of our proposed method. We conduct experiments on the Caltech-UCSD dataset [23], which is the most widely used benchmark for fine-grained classification. We train the model using ImageNet [18] and recently released Birdsnap dataset [4].

ImageNet consists of 1000 classes with approximately 1000 images for each class. In total there are approximately 1.2 million training images.

Caltech-UCSD contains 11,788 images across 200 species. Birdsnap contains 500 species of North American birds with 49,829 images. Examples are shown in Fig. 1. Birdsnap is similar in structure to Caltech-UCSD, but has several differences. First, it contains overlapping 134 species and four times the number of images than Caltech-UCSD. Second, there is strong intra-variation within many species due to sexual as well as age dimorphisms. There are considerable appearance differences between males and females, as well as between young and mature birds.

We use the implementation of LDA and k-means from the Bob library [2]. The open-source package Caffe [17] is used to train and extract CNN features. We use $lda-fc6$ layer features to pre-cluster subsets and $fc6$ features for classification.

4.1. Evaluation of Transfer Learning for Domain-Generic Features

The CNN model architecture is identical to the model used by Krizhevsky et al. [18]. We fine-tune the CNN model by using training images from the ground truth bounding box crops of the original images. The resultant cropped images are all resized 227×227. During test time, ground truth bounding box crops of the test images from Caltech-UCSD are used to make predictions.

We conducted 3 sets of experiments for transfer learning:

1. The first experiment used all of the data from Birdsnap (500 species) to perform large-scale progressive feature learning.

2. In the second experiment we removed those species in Birdsnap and Caltech-UCSD that overlapped. This allows us to examine the potential for learning domain features that are not specific to the task at hand.

3. In the third experiment we explored the impact that including the overlapping species has on the transfer learning process.

We use the following acronyms. **IN** represents using weights from the pre-trained ImageNet model. We define **rt** as retraining the network from scratch with random initialised weights. **ft** refers to fine-tuning the network. For example, **IN-CUB-ft** means fine-tuning the ImageNet model weights on the Caltech-UCSD bird dataset. ImageNet dataset is represented as **IN**, while Caltech-UCSD is **CUB**, and Birdsnap is **BS**.

4.1.1 Transfer Learning: Experiment I

In this experiment we used all images (500 species) from Birdsnap to explore large-scale progressive feature learning. We exclude those images that exist in both Birdsnap and the Caltech-UCSD datasets.

The first three rows of Table 1 show the accuracy when the CNNs are trained from scratch. In this setting the **IN-rt** system, the pre-trained network generated by Krizhevsky et al. [18] on ImageNet, performs the best with a mean accuracy of 58.0%. Interestingly, the **BS-rt** system has a considerably higher mean accuracy of 44.8% when compared to **CUB-rt** which has a mean accuracy of 11.4%. We believe that this indicates that the Birdsnap dataset has almost enough data to train a deep CNN from scratch.

Transfer learning offers a way to mitigate the lack of sufficient domain data. As such, we performed transfer learning by fine-tuning the pre-trained CNN. We did this using just the Caltech-UCSD (target) dataset **IN-CUB-ft** or the Birdsnap (domain specific) dataset **IN-BS-ft**.

Somewhat surprisingly, training on the target dataset (**IN-CUB-ft**) provides a lower mean accuracy of 68.3% when compared to using the domain specific dataset (**IN-BS-ft**) which has a mean accuracy of 70.1%. Performing progressive feature learning on the **IN-BS-ft** CNN leads to further improvements achieving a mean accuracy of 70.8% (**IN-BS-ft-CUB-ft**). These two results demonstrate the potential for learning domain-generic features (**IN-BS-ft**) as well as progressive feature learning to perform effective transfer learning (**IN-BS-ft-CUB-ft**) for fine-grained image classification.

Table 1. Mean accuracy of transfer learning on the Caltech-UCSD bird dataset (bounding box annotation provided). Steps represents the number of training stages.

Method	Steps	Mean Accuracy
All species (500)		
IN-rt	1	58.0%
CUB-rt	1	11.4%
BS-rt	1	44.8%
IN-CUB-ft	2	68.3%
IN-BS-ft	2	70.1%
IN-BS-ft-CUB-ft	3	**70.8**%
Non-overlapping species (366)		
IN-BS-ft	2	67.7%
IN-BS-ft-CUB-ft	3	**70.5**%
Overlap (134) + Random (232)		
IN-BS-ft	2	69.5%

An obvious issue that is not addressed in this first experiment is that there are overlapping species in Birdsnap and Caltech-UCSD. To evaluate the impact of this we perform two more experiments.

4.1.2 Transfer Learning: Experiment II

Next we investigate transfer learning features from non-overlapping classes between two bird datasets. We fine-tune the pre-trained CNN using those species from the Birdsnap dataset that do not overlap with Caltech-UCSD. There are 134 species that overlap and so we only use 366 species for this experiment.

As can be seen from the second part of the Table. 1, the result of transfer learning on Birdsnap in this setting is slightly worse with a mean accuracy of 67.7%. However, if we perform progressive feature learning by learning on the target dataset (**IN-BS-ft-CUB-ft**) we obtain a mean accuracy of 70.5%. This is only 0.3% worse than if we used all of the Birdsnap data and demonstrates the effectiveness of progressive feature learning.

4.1.3 Transfer Learning: Experiment III

In this experiment we show the importance of overlapping classes for learning domain-generic features. In order to investigate if the overlapping classes play a key role to learn domain-generic features, we fine-tuned the ImageNet model again with 134 overlapping species and 232 randomly selected unique species from the Birdsnap; this gives us 366 species which is the number of species available in Experiment II. The result shows that overlapping species

are important to learn domain-generic species with a mean accuracy of 69.5%.

4.2. Evaluation of Subset Specific Features

In this set of experiments we evaluate our proposed subset feature learning method on Caltech-UCSD. We use the same evaluation protocol as domain-generic feature learning in the previous section, where the $DFCNN$ is used to extract features from given ground truth bounding box location of the whole bird. We use the acronym **SF** to indicate subset feature learning. Based on initial experiments we set $K = 6$.

Results in Table 2 show that subset feature learning provides considerable improvements. As a baseline, the results from [26] are shown, where the features were fine-tuned on the Caltech-UCSD dataset; this equates to **IN-CUB-ft** in our terminology. Comparing to this baseline, both of our proposed subset feature learning methods, **IN-BS-ft-SF(SCNN)** and **IN-BS-ft-SF(k-means)**, provide considerable improvements with mean accuracies of 72.0% and 70.4% respectively. This demonstrates the effectiveness of our proposed subset feature learning technique, and the importance of the subset selector as the SCNN approach provides an absolute performance improvement of 1.6% when compared to the much simpler k-means approach.

4.3. Comparison with State-of-the-Art

In this section we demonstrate that subset feature learning can achieve state-of-the-art performance for automatic fine-grained bird classification. Recent work in [26] provided state-of-the-art performance on the Caltech-UCSD dataset. This was achieved by crafting a highly accurate parts localisation model which leveraged deep convolutional features computed on bottom-up region proposals based on the RCNN framework [14] . We show that if we use a similar approach but substitute their global feature vector with the feature vector obtained from subset feature learning, then state-of-the-art performance can be achieved.

We present our results under the same setting as [26], where the bird detection bounding box is unknown during test time. This setting is fully automatic and hence more realistic. Since we concentrate on feature learning we use the detection results and parts features from [26], and substitute their global feature vector with the one we learn from subset feature learning.

The results in Table 3 show that our proposed method achieves a mean accuracy of 77.2% when we use domain-generic features and subset-specific features. This is a considerable improvement over the previous state-of-the-art system [26] which achieved a mean accuracy of 73.2%. An extra 0.3% performance is gained when we perform progressive feature learning and fine-tune the CNN model again on the Caltech-UCSD dataset. Qualitative results are shown in Fig. 5 which highlight instances where the previous state-of-the-art methods provides an incorrect class label despite large visual dissimilarities. In contrast, our approach provides the correct class label.

Table 2. Mean accuracy on the Caltech-UCSD bird dataset of subset-specific features learned using subset feature learning. Annotated bounding boxes are used.

Method	Mean Accuracy
Fine-tuned Decaf [26]	68.3%
IN-BS-ft + SF(k-means)	70.4%
IN-BS-ft + SF(SCNN)	**72.0%**

Table 3. Comparison to recent results on the Caltech-UCSD bird dataset. Bounding boxes are not used.

Method	Mean Accuracy
DPD-DeCAF [27]	44.9%
Part-based RCNN with δ^{KP} [26]	73.2%
IN-BS-ft + SF(k-means) with δ^{KP}	76.2%
IN-BS-ft + SF(SCNN) with δ^{KP}	**77.2%**
IN-BS-ft-CUB-ft + SF with δ^{KP}	**77.5%**

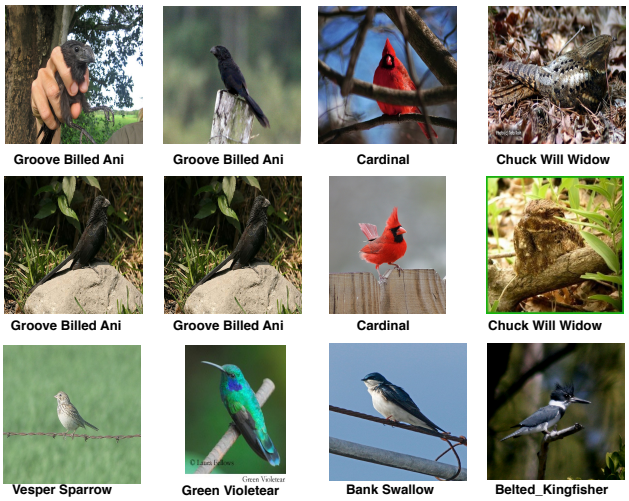

Figure 5. Qualitative comparison between our proposed method and the previous state-of-the-art approach [26] (part-based RCNN with δ^{KP}). The first row shows examples of test images, the second row shows the corresponding predicted classes from our proposed method, and the last row images shows the predictions using [26]. It can be seen that the previous state-of-the-art approach made errors despite the large visual dissimilarities between the test image and the predicted classes. In contrast, the proposed approach provides the correct class labels in these cases.

5. Conclusion

We have proposed a progressive transfer learning system to learn domain-generic features as well as subset learning to learn subset specific features. For progressive transfer

learning, we have shown that it is possible to learn domain-generic features for tasks such as fine-grained image classification. Furthermore, we have shown that progressive transfer learning of these domain-generic features can be performed to learn target set specific features, yielding considerable improvements in accuracy.

Finally, we have presented a subset feature learning system that is able to learn subset-specific features. Using this approach we achieve state-of-the-art performance of 77.5% for fully automatic fine-grained bird image classification, the most difficult setting. We believe our proposed method can be useful not only for fine-grained image classification, but also for improving general object recognition. We will examine this potential in future work.

Acknowledgments

The Australian Centre for Robotic Vision is supported by the Australian Research Council via the Centre of Excellence program. NICTA is funded by the Australian Government through the Department of Communications, as well as the Australian Research Council through the ICT Centre of Excellence program.

References

[1] C. C. Aggarwal. On k-anonymity and the curse of dimensionality. In *International Conference on Very Large Data Bases*, pages 901–909, 2005.

[2] A. Anjos, L. E. Shafey, R. Wallace, M. Günther, C. McCool, and S. Marcel. Bob: a free signal processing and machine learning toolbox for researchers. In *ACM Conference on Multimedia Systems (ACMMM), Nara, Japan*, 2012.

[3] T. Berg and P. N. Belhumeur. Poof: Part-based one-vs.-one features for fine-grained categorization, face verification, and attribute estimation. In *CVPR*, 2013.

[4] T. Berg, J. Liu, S. W. Lee, M. L. Alexander, D. W. Jacobs, and P. N. Belhumeur. Birdsnap: Large-scale fine-grained visual categorization of birds. In *CVPR*, pages 2019–2026, 2014.

[5] S. Branson, P. Perona, and S. Belongie. Strong supervision from weak annotation: Interactive training of deformable part models. In *ICCV*, 2011.

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

[7] Y. Chai, V. Lempitsky, and A. Zisserman. Symbiotic segmentation and part localization for fine-grained categorization. In *ICCV*, 2013.

[8] G. Chen, J. Yang, H. Jin, E. Shechtman, J. Brandt, and T. X. Han. Selective pooling vector for fine-grained recognition. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 860–867, 2015.

[9] 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. *ICML*, 2014.

[10] R. Farrell, O. Oza, N. Zhang, V. I. Morariu, T. Darrell, and L. S. Davis. Birdlets: Subordinate categorization using volumetric primitives and pose-normalized appearance. In *ICCV*, 2011.

[11] P. Felzenszwalb, D. McAllester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model. In *CVPR*, 2008.

[12] E. Gavves, B. Fernando, C. G. Snoek, A. W. Smeulders, and T. Tuytelaars. Fine-grained categorization by alignments. In *ICCV*, 2013.

[13] Z. Ge, C. McCool, C. Sanderson, and P. Corke. Modelling local deep convolutional neural network features to improve fine-grained image classification. *arXiv:1502.07802*, 2015.

[14] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. *arXiv:1311.2524*, 2013.

[15] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik. Simultaneous detection and segmentation. In *ECCV*, pages 297–312. Springer, 2014.

[16] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *arXiv:1502.01852*, 2015.

[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.

[18] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

[19] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1(4):541–551, 1989.

[20] J. Liu, A. Kanazawa, D. Jacobs, and P. Belhumeur. Dog breed classification using part localization. In *ECCV*. 2012.

[21] C. R. Rao. The utilization of multiple measurements in problems of biological classification. *Journal of the Royal Statistical Society. Series B (Methodological)*, 10(2):159–203, 1948.

[22] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. CNN features off-the-shelf: an astounding baseline for recognition. *CVPR Workshop on Deep Vision*, 2014.

[23] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. In *Computation & Neural Systems Technical Report, California Institute of Technology*, number CNS-TR-2011-001, 2011.

[24] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson. How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems*, pages 3320–3328, 2014.

[25] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional neural networks. *arXiv:1311.2901*, 2013.

[26] N. Zhang, J. Donahue, R. Girshick, and T. Darrell. Part-based r-cnns for fine-grained category detection. In *ECCV*, pages 834–849. 2014.

[27] N. Zhang, R. Farrell, F. Iandola, and T. Darrell. Deformable part descriptors for fine-grained recognition and attribute prediction. In *ICCV*, 2013.