Content Specific Feature Learning for Fine-Grained Plant Classification

ZongYuan Ge [†], Chris McCool [†], Conrad Sanderson [*], and Peter Corke [†]

[†] Australian Center for Robotic Vision, Queensland University of Technology
[*] NICTA, Australia
Corresponding author: z.ge@qut.edu.au or c.mccool@qut.edu.au

Abstract. We present the plant classification system submitted by the QUT RV team to the LifeCLEF 2015 plant task. Our system learns a content specific feature for various plant parts such as branch, leaf, fruit, flower and stem. These features are learned using a deep convolutional neural network. Experiments on the LifeCLEF 2015 plant dataset show that the proposed method achieves good performance with a score of 0.633 on the test set.

Keywords: deep convolutional neural network, plant classification, subset feature learning

1 Introduction

Fine-grained image classification has received considerable attention recently with a particular emphasis on classifying various species of birds, dogs and plants [1, 3, 4, 11]. Fine-grained image classification is a challenging computer vision problem due to the small inter-class variation and large intra-class variation. Plant classification is a particularly important domain because of the implications for automating Agriculture as well as enabling robotic agents to detect and measure plant distribution and growth.

To evaluate the current performance of the state-of-the-art vision technology for plant recognition, the Plant Identification Task of the LifeCLEF challenge [5, 7] focuses on distinguishing 1000 herb, tree and fern species. This is an observation-centered task where several images from seven organs of a plant are related to one observation. There are seven organs, referred to as **content** types, and include images of the entire plant, branch, leaf, fruit, flower, stem or a leaf scan.

Inspired by [4], we use a deep convolutional neural network (DCNN) approach and learn a separate DCNN for each content type. We combine the content-specific feature with a generic DCNN feature, which is trained using all of the content types. This approach yields a highly accurate classification system with a score of 0.633 on the test set.

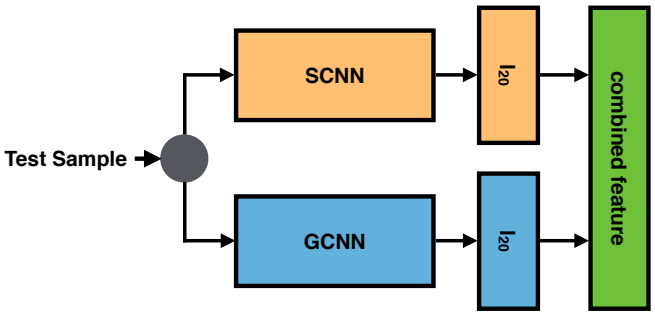

Fig. 1. For each test sample, a domain-generic (GCNN) and subset-specific (SCNN) feature is extracted. These two features are then concatenated to form a combined feature vector.

2 Our Approach

Our proposed system consists of two main parts. First, we perform transfer learning to learn a domain-generic feature termed as ϕ_{GCNN} from all plants images (regardless of content type). Second, we manually cluster the dataset into subsets based on content type and learn a feature specific to each subset (ϕ_{SCNN}). For each image we extract both domain-generic (ϕ_{GCNN}) and subset-specific (ϕ_{SCNN}) features, these features are obtained from layer 20, l_{20}, of the deep network. The two feature vectors are then concatenated to form a single feature vector as shown in Figure 1. These features are then used to learn a multi-class linear SVM. Power and l_2 norm are applied independently for domain-generic feature and content specific feature prior to combining the feature vectors.

2.1 Content Clustering

There are 7 pre-defined content types consisting of images from the *entire plant*, *branch*, *leaf*, *fruit*, *flower*, *stem* or a *leaf scan*. In both the training and testing phases all participants are allowed to use the indicated content.

We make use of the content type to learn a DCNN that is fine-tuned (specialised) for a subset of the content types. However, because there is a limited number of images for each content type, we first group the most visually similar content types toghether. In particular, we define four subsets. The first subset conists of the the *entire plant* and *branch* content types, the second subset consists of the *leaf* and *leaf scan* content types, the third subset contains *fruit* and *flower* content types, and the fourth subset consists of the *stem* only.

2.2 Deep Convolutional Neural Networks as Feature Representation

Krizhevsky et al. [8] recently achieved impressive performance on the ImageNet recognition task using CNNs, which were initially proposed by LeCun et al. [9]

for hand written digit recognition. Since then CNNs have received considerable attention and in the Large-scale ImageNet Challenge 2014 (ILSVRC) the top five results were all produced using CNN-based systems [10].

In this work we fine-tune a general model for the task of plant classification. The base model that we fine-tune is the best performing model from ILSVRC [12], referred to as GoogLeNet. GoogLeNet is a very deep neural network model with 22 layers. It consists primarily of convolutional layers. We use the output of the last convolutional layer l_{20}, after average pooling, to obtain our feature vectors.

2.3 Domain Specific Feature Learning

Transfer learning has usually been applied by fine-tuning a general network, such as the network of Krizhevsky et al. [8], to a specific task such as bird classification [13].

Inspired by the findings of Zhang et al. [13] we learn a domain-generic DCNN for the task of plant classification. This is achieved by applying transfer learning on the parameters of the GoogLeNet model (learned from the large-scale ImageNet dataset) using all of the training data for the plant classification task. This new DCNN provides domain-generic features for the task of plant classification and is referred to as the domain-generic DCNN. The only difference between the pre-trained GoogLeNet model and the domain-generic DCNN is that the number of outputs for the last fully connected layer is changed to be $1,000$ which is the number of training classes available. For each image we can then obtain a domain-generic feature ϕ_{GCNN} from the last convolutional layer l_{20}.

2.4 Subset Feature Learning as Content Specific Feature

A separate DCNN is learned for each of the $K = 4$ pre-defined subsets by fine-tuning the domain-specific model, described in Section 2.3. The aim is to learn features for each subset that will allow us to more easily differentiate visually similar content of plant species. As such, for each subset, we apply fine-tuning to the pre-trained GoogLeNet model. To train the k-th subset ($Subset_k$) we use the N_k images assigned to this subset $\mathbf{X}_k = [\mathbf{x}_1, ..., \mathbf{x}_{N_k}]$, with their corresponding class labels.

The only difference between these models and the pre-trained GoogLeNet model is that the number of outputs for the last fully connected layer, of each model, is set to the number of training classes in each subset. Transfer learning is then applied separately to each network using backpropogation and stochastic gradient descent (SGD). For each image belonging to the k-th subset a subset feature vector ϕ_{SCNN_k} is obtained by taking the output of the last convolutional layer l_{20}.

3 Experiments

In this section we present a comparative performance evaluation of our proposed method on a validation set and the defined test sets. The provided training dataset is split into two sets: roughly 10% of the total training data was used as a validation set and the rest is used for training the models. The split is based on observation id because final testing is also observation-based.

This results in 82,033 training images, including 21,746 for the *branch* and *entire* subset, 32,186 for *fruit* and *flower* subset, 23,234 for the *leaf* and *leaf scan* subset and 4,867 for the *stem* subset. The validation set consists of 9,725 images.

We use Caffe [6] for learning generic and subset specific features. The open-source package LibLinear [2] is used to train the multi-class linears SVMs. The SVM cost parameter C is set to 1 and all images are resized to 224×224.

3.1 Results on Validation Set

First we assess our proposed method on the validation set. We conducted three sets of experiments which examine the effectives of the domain-specific feature vector, the subset feature vector and the combination of these two feature vectors.

The results on the validation set, shown in Table 1, demonstrate that the combination of these two feature vectors provides a considerable performance improvement. The combination of these two feature vectors achieves a mean accuracy of 66.6%. This is an absolute improvement of 6.5 percentage points over the domain-specific feature vector ϕ_{GCNN} which achieves a mean accuracy of 60.1%. By comparison, the subset feature vector ϕ_{SCNN_k} achieves a mean accuracy of only 58.0%. We believe that the subset feature vector performs worse than the domain-specific feature vector because of the limited number of training images for each subset.

Table 1: Mean accuracy on the LifeCLEF 2015 Plant dataset of our proposed method. Annotated content information is used.

Method	**Mean Accuracy**
Domain Specific Feature	60.1%
Content Specific Feature	58.0%
Combined	**66.6%**

3.2 Results on Test Set

In this section, we present our submitted results for the LifeCLEF2015 plant challenge. We submitted three runs:

– RUN1 is the result of using proposed system for classification purpose. Only the rank 1 score is submitted for each observation.
– RUN2 is the image retrieval task where we take the first 5 predictions.
– RUN3 is based on RUN 2 but we perform an additional softmax normalization for the first five predictions.

In Figure 2 we present the overall performance for all of the competitors using the defined score metric. It can be seen that our best performing system is RUN 2 which achieved a score of 0.633. This is slightly worse than SNUMED INFO systems (RUN 4 and RUN 3).

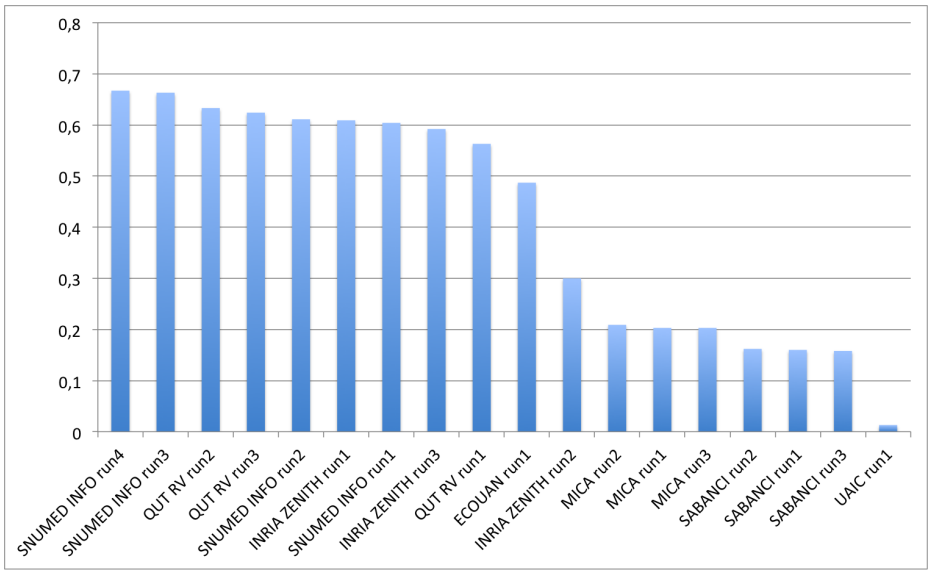

Fig. 2. The results of observation-based for the LifeCLEF Plant Task 2015. Image adapted from the organisers' website.

In Figure 3 we present results for the image-based run. It can be seen that our proposed method provides competitive performance for both the image-based and observation-based metrics. However, we do have a minor performance loss for the image-based result compared to the observation-based result.

4 Conclusions and Future Work

In this paper we presented a domain-specific feature learning and subset-specific feature learning system applied to the plant identification task of LifeCLEF 2015. For domain-specific feature learning, we have shown that it is possible to

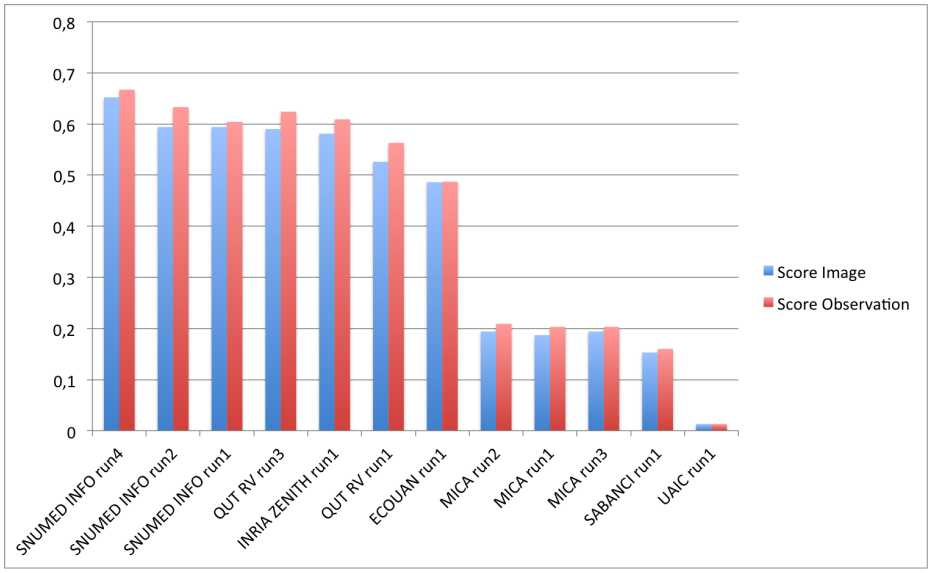

Fig. 3. The results of image-based for the LifeCLEF Plant Task 2015. Image adapted from the organisers' website.

perform transfer learning from a DCNN pre-trained on the larger-scale ImangNet dataset. Furthermore, we have presented a subset feature learning system that is able to learn content specific features. This approach yields highly competitive performance with a score of 0.633 for this year's task.

Acknowledgements

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. We would also like to thank Professor Chunhua Shen and Dr. Lingqiao Liu for the fruitful conversations of this work.

References

1. Y. Chai, V. Lempitsky, and A. Zisserman. Symbiotic segmentation and part localization for fine-grained categorization. In *ICCV*, 2013.
2. Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. Liblinear: A library for large linear classification. *The Journal of Machine Learning Research*, 9:1871–1874, 2008.

3. Efstratios Gavves, Basura Fernando, Cees GM Snoek, Arnold WM Smeulders, and Tinne Tuytelaars. Local alignments for fine-grained categorization. *International Journal of Computer Vision*, pages 1–22, 2014.
4. ZongYuan Ge, Christopher McCool, Conrad Sanderson, and Peter Corke. Subset feature learning for fine-grained classification. *CVPR Workshop on Deep Vision*, 2015.
5. Hervé Goëau, Alexis Joly, and Pierre Bonnet. Lifeclef plant identification task 2015. In *CLEF working notes 2015*, 2015.
6. Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv:1408.5093*, 2014.
7. Alexis Joly, Henning Müller, Hervé Goëau, Hervé Glotin, Concetto Spampinato, Andreas Rauber, Pierre Bonnet, Willem-Pier Vellinga, and Bob Fisher. Lifeclef 2015: multimedia life species identification challenges. In *Proceedings of CLEF 2015*, 2015.
8. Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1097–1105, 2012.
9. Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1(4):541–551, 1989.
10. Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *arXiv preprint arXiv:1409.0575*, 2014.
11. Asma Rejeb Sfar, Nozha Boujemaa, and Donald Geman. Confidence sets for fine-grained categorization and plant species identification. *IJCV*, 2014.
12. Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. *arXiv:1409.4842*, 2014.
13. Ning Zhang, Jeff Donahue, Ross Girshick, and Trevor Darrell. Part-based R-CNNs for fine-grained category detection. In *ECCV*, pages 834–849. 2014.