Stop and Scan the Trees: Tree Leaf Recognition with Transfer Learning

Christopher Koenig
Stanford University
koenig97@stanford.edu

Krishna Patel
Stanford University
kpatel7@stanford.edu

Abstract

The deployment of deep convolutional networks in recent years has led to state-of-the-art performance in the challenging task of fine-grained visual categorization (FGVC). Most systems perform end-to-end classification with the dataset of interest as input and categorical predictions as output. However, custom deep convolutional networks are expensive to train and difficult to develop. In this study, we investigate the application of recent developments in transfer learning to the FGVC task for the Leafsnap dataset. Specifically, we leverage deep pretrained networks as feature extractors for input into simple machine learning models, and compare performance with end-to-end CNN methods.

1 Introduction

Over the past decade, the field of computer vision has been transformed by the advancement of deep learning. However, the task of fine-grained visual categorization (FGVC) remains a significant challenge due to low levels of variance between images of different classes, particularly as compared to standard image classification challenges in which different classes are often clearly distinct.

Recent development of deep convolutional networks of 10-20 layers have achieved record performance on many FGVC tasks. However, in practice, the massive number of parameters required by these networks makes them slow to train. For example, Barré et al. (2017) note that Leafnet, one of the most advanced leaf recognition systems, took 32 hours to train using a dataset of 270,000 leaf images - sizeable, but by no means massive in the era of big data when compared to datasets such as ImageNet with over 10 million images.

Large numbers of parameters and the need to search large hyperparameter spaces for suitable training parameters translates to slow iteration during model development. Model development also requires practitioners with deep learning expertise to make appropriate decisions such as what forms of regularization to apply, what architecture is appropriate, what type of pooling to incorporate, how to interpret training signal and debug appropriately, etc. The combination of slow model development and the requirement for deep learning expertise can make deep learning solutions inaccessible to practitioners in other fields who would benefit from the power that deep learning offers but who might not have the time and resources required to develop a deep network from scratch.

In this study, we investigate transfer learning as a solution to the dual problems of slow model development and the need for deep learning expertise outlined above. By leveraging pretrained deep CNN models as feature extractors, we reap the benefit of automatic learning of the discriminative features of the dataset images that deep convolutional networks offer while avoiding the time, computational resources, and expertise required to develop a deep CNN for the problem domain. In addition, using simpler machine learning models for classification such as Support Vector Machine (SVM) and Multi-Layer Perceptron (MLP) results in far fewer parameters and a much smaller hyperparameter space that need to be optimized, further simplifying the model development process.

Our hope is that simplifying the model development process while maintaining similar performance to highly optimized deep convolutional networks makes the benefits of deep learning accessible to a broader array of practitioners, thereby enabling new progress on critical FGVC tasks - such as plant species identification - that are needed if we are to protect global biodiversity.

	Lab Images	Field Images	Total Images
Average	125.119	41.72	166.84
Min	7	0	56
Max	183	424	448
Std. Dev.	39.84	25.7	46.5
Std. Dev. w/o Max and Min	32.24	23.42	3.04

Table 1: Distribution Statistics for Number of Images by Species

2 Related Work

2.1 End-to-end Deep Learning Approaches for Plant Species Identification

Most existing approaches classify plants from their leaves, as leaves are unique across plant species and act as a proxy plant ”fingerprint”. Lee et al. (2015) were one of the first to develop state-of-the-art deep learning approaches for plant species identification, achieving 99.7% accuracy on 44 species from the Royal Botanic Gardens in England. Hang et al. (2016) achieved a mean average precision of 74.2% on the more challenging task of PlantCLEF 2016, which involved classifying over 100,000 images into 1000 species. On the Leafsnap dataset, Barré et al. (2017) developed the ”LeafNet” system - composed of 11 convolutional layers followed by 3 fully connected layers - to achieve 86.3% top-1 accuracy and 97.8% top-5 accuracy. Finally, Galbally et al. (2018) surpassed the performance of LeafNet by developing an 18-layer ResNet that achieved a top-1 accuracy of 93.8%. As far as we are aware, this model is the state-of-the-art on the Leafsnap dataset.

2.2 Deep Learning Feature Extractors

One state-of-the-art network for image recognition, localization, and detection is the *OverFeat* CNN developed by Sermanet et al. (2013), which implements a novel multi-scale, sliding window approach which provides powerful features when *OverFeat* is leveraged as a feature extractor. Another model that has seen extraordinary success is the deep residual network, known as ResNet (He et al., 2015). These networks are effective feature extractors due to their extremely deep layers, which are able to learn features at many different levels of abstraction. This is the case even if the feature extractor is applied to a new dataset that is far from the original domain; Yosinski et al. (2014) found that transferring features even from distant tasks can be better than using random features.

2.3 Transfer Learning Approaches to Plant Species Identification

Barré et al. (2017) compared the performance of their end-to-end LeafNet system to features extracted using the *OverFeat* CNN and fed into a SVM. They found that the transfer learning approach yielded better performance on the Foliage and Flavia datasets, but worse performance on Leafsnap (82.3% accuracy to 86.3% accuracy). While their investigation was not rigorous, the initial findings suggest there is potential in this area.

3 Dataset

Figure 1: Distribution of Images by Species

Our dataset is the LeafSnap dataset, available at: http://leafsnap.com/dataset/. The dataset consists of 30866 images of leaves, 23147 of which are high-quality lab images and 7719 of which are lower-quality field images. The lab images feature flattened leaves, and are taken under controlled back and front lighting. On the other hand, the field images feature unpressed leaves, taken by mobile devices outdoors. Each image is labeled with the species of the tree associated with the leaf, and there are 185 different classes of leaf present in the dataset. The dataset also includes segmentations of the images, however segmentation fails on a number of examples, resulting in a completely black image. The class distribution is fairly balanced, as we can see in the figure above, with around 100-200 images of each species total.

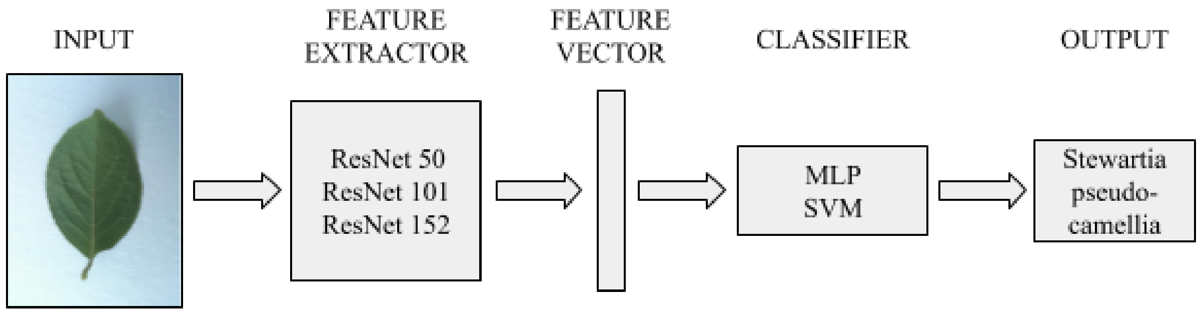

Figure 2: Visual Summary of Transfer Learning System

4 Methods

A visual summary of our transfer learning system is given in Figure 2. Transfer learning is the process by which the parameters learned by a model on one dataset are applied to extract features from the same type of input on a different dataset. This is common in computer vision with deep CNNs. The convolutional layers act as feature extractors, determining what areas of an image are most discriminative. The output of the convolutional layers is then typically fed into a series of fully-connected layers for classification. The convolutional layers can thus be ”transferred” to other datasets and still serve their function as feature extractors, with their outputs being fed into a new classifier for the new problem dataset.

4.1 Deep Feature Extractors

For our study, we compare the use of the publicly available ResNet50, ResNet101, ResNet152 networks as feature extractors for our transfer learning system. As detailed by He et al. (2015), ResNets achieve impressive power by adding ”shortcut connections” that skip one or more convolutional layers. These connections enable blocks in ResNets to easily learn an identity function. In practice, this means that many more layers can be added in a ResNet than a traditional network without overfitting or failing during training due to the problem of vanishing gradients.

4.2 Machine Learning Classifiers

Once we obtain image embeddings from our feature extractors, we leverage SVM and Multilayer Perceptron as the classifiers to take these embeddings as input and output the predicted species. Regarding terminology, we use ”Multilayer Perceptron” (MLP) to refer to a general feedforward neural network with one or more hidden layers.

4.2.1 Support Vector Machines

We utilize a Support Vector Machine (SVM) for classifying the embeddings because support vector machines are considered by many to be the best ”off-the-shelf” machine learning classifiers. SVMs optimize the following objective:

$$\begin{aligned} \min_{\gamma,w,b} \quad & \frac{1}{2}||w||^2 + C\sum_{i=1}^{n}\xi_i \\ \text{s.t.} \quad & y^{(i)}(w^T x^{(i)} + b) \geq 1 - \xi_i, \quad i = 1, \ldots, n \\ & \xi_i \geq 0, \quad i = 1, \ldots, n. \end{aligned}$$

where C is the penalty parameter. SVMs leverage kernel methods to be able to efficiently learn in very high dimensional spaces. We explore 2 kernels in our approach: the linear kernel and rbf kernel, defined as:

$$K(\mathbf{x}, \mathbf{x}') = < \mathbf{x}, \mathbf{x}' >$$

$$K(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2\sigma^2}\right)$$

4.2.2 Multi-Layer Perceptron

While SVMs are extremely effective off-the-shelf classifiers, MLP models often have more expressive power. MLPs are composed of one or more ”hidden” layers with non-linear activation functions through which the input is propagated to produce an output score(s) - referred to as the logits - for 2 or more classes. The non-linear activation functions applied at each hidden layer are crucial, as they are what enable the model to learn very complex decision boundaries, which is what makes the MLP more powerful than other machine learning models. For our system, we utilize the Rectified Linear Unit (ReLU) activation function, which is defined as follows:

$$f(x) = \max(ax + b, 0)$$

Model	Feature Extractor	Top-1 Eval Acc. (%)	Top-1 Test Acc. (%)	Top-5 Eval Acc. (%)	Top-5 Test Acc. (%)
SVM	ResNet50	97.6	97.6	99.8	**99.9**
SVM	ResNet101	97.9	**97.8**	99.6	99.8
SVM	ResNet152	97.4	97.4	99.5	99.7
MLP	ResNet50	97.3	**97.7**	99.7	**99.8**
MLP	ResNet101	97.2	97.3	99.7	99.8
MLP	ResNet152	97.5	97.0	99.7	99.7

Table 2: Performance of top classifier and ResNet combinations on the Leafsnap dataset **excluding mislabeled species**

5 Experiments & Results

5.1 Evaluation Metrics

We report the top-1 accuracy, top-5 accuracy, and confusion matrix. The accuracy scores provide a performance overview while a confusion matrix provides insight into per-class performance.

5.2 Data Splits & Processing

We create a 80/10/10 train/validation/test split of the data and use only the color images, excluding the segmentations. This gives us a total of 30866 lab and field images in a 24692/3087/3087 split. The pretrained ResNet models expect fixed-size, normalized images, so we downscale all LeafSnap images to 256 x 256 and normalize the RGB values with mean = $[0.485, 0.456, 0.406]$ and std. dev.= $[0.229, 0.224, 0.225]$. This helps to standardize varying camera quality and lighting conditions. We do not apply any form of data augmentation. The image embedding outputted after propagation through ResNet has dimension 2048.

5.3 Hyperparameters

For the SVM we searched a small hyperparameter space including linear and rbf kernels and 0.1, 1.0, 10.0, and 100.0 values of C (where C is inversely proportional to regularization strength). We consistently found that a RBF kernel with 10.0 regularization performed best. The MLP hyperparameter space was much larger. However, we found that increasing the expressive power by increasing the number of hidden dimensions in a 1-layer MLP over 1000 or adding more than 1 layer actually hurt performance due to overfitting. We thus trained our MLP models with 1 hidden layer of size 1000 with early stopping and 1000 max epochs, 0.0001 learning rate, Adam optimizer, ReLU activation, and 0.001 L2 regularizaation penalty. However, similar performance to the MLP results in table 2 was achieved with only 500 hidden layers and 100 max epochs.

5.4 Results

The top-1 and top-5 accuracies for the best performing classifier and feature extractor combinations are given in table 2. We see that the SVM with ResNet101 embeddings achieves the best top-1 accuracy while the SVM with ResNet50 embeddings achieves the best top-5 accuracy. The confusion matrices for these models are provided in figures 3 and 4. We also note that the test accuracy often matches or exceeds the validation accuracy, indicating that the models generalize well.

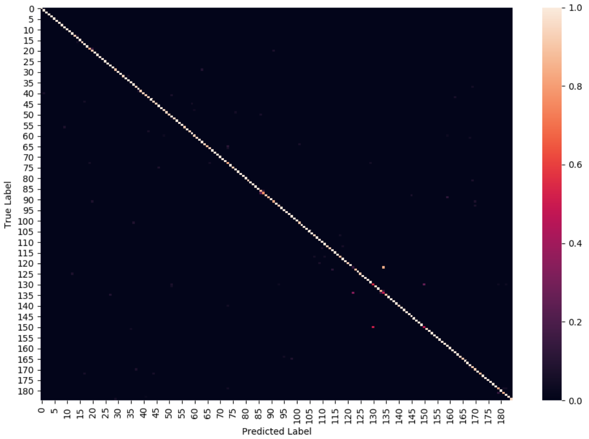

Figure 3: Confusion Matrix for SVM with ResNet50

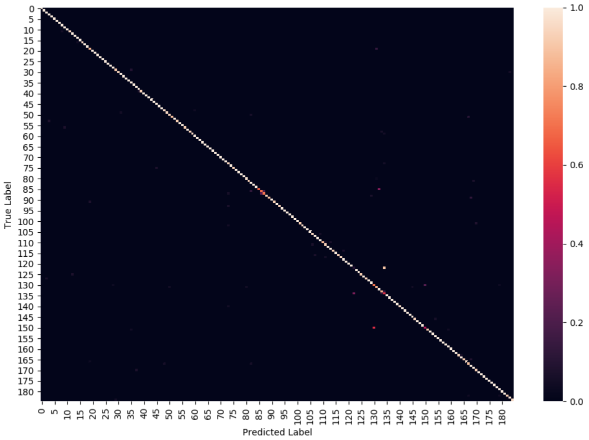

Figure 4: Confusion Matrix for SVM with ResNet101

Architecture	Author	Top-1 Test Acc. (%)	Top-5 Test Acc. (%)
Leafsnap	Kumar et al. (2012)	73.0	-
LeafNet	Barré et al. (2017)	86.3	97.8
ResNet18	Galbally et al. (2018)	93.8	99.5
SVM with ResNet50	This Study	97.6	**99.9**
SVM with ResNet101	This Study	**97.8**	99.8

Table 3: Comparison of our top models to accuracies achieved on the LeafSnap datasets in previous work

6 Discussion

6.1 Error Analysis

Upon close inspection, we identify 4 bright squares off the diagonal indicating high percentages of misclassifications in the confusion matrix in figure 3. With further investigation, we discovered that these 4 squares correspond to 2 pairs of species, *Pinus virginiana* - *Prunus virginiana* and *Quercus muehlenbergii* - *Prunus sargentii*, that are mislabeled in the LeafSnap dataset. Surprisingly, past work on LeafSnap has not surfaced this issue.

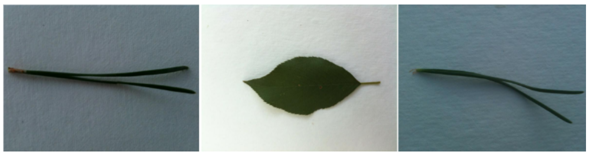

Figure 5: Samples labeled as *Pinus virginiana* (left), *Prunus virginiana* (center), and *Prunus virginiana* (right) in the Leafsnap dataset. The image on the right is labeled as Prunus virginiana but should be labeled Pinus virginiana.

The remaining misclassifications occur primarily between species with strong visual similarity such as those in the same genus, as seen in figure 6.

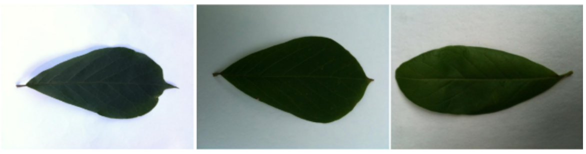

Figure 6: Magnolia stellata (left) misclassified as Magnolia denudata (center) 12% and Magnolia tripetala (right) 19% of the time by our top model of SVM with ResNet50.

6.2 Feature Extractor Comparison

Somewhat surprisingly, we find that the SVM and MLP models perform best, on balance, with embeddings produced from the ResNet50 feature extractor over ResNet101 and ResNet152. Furthermore, the models slightly - but unequivocally - perform the worst with ResNet152 embeddings. These findings demonstrate that more power with deeper models does not necessarily translate to improved performance. It should not be the case that the ResNet152 model is overfitting the images, both because the ResNet is pretrained on ImageNet so it undergoes no actual training on LeafSnap and because the distinguishing characteristic of ResNets is their ability to incorporate deep layers without overfitting data. It therefore might be the case that the additional level of discriminatory detail provided by the additional layers of the ResNet101 and ResNet152 models simply isn't helpful for the classification task because the information provided at 50 layers is sufficient. Future work could investigate what the minimum number of ResNet layers is that maintains the level of performance achieved with ResNet50.

6.3 Comparison to Previous Work

From table 3, we see that our transfer learning system outperforms all major previous reported work on the LeafSnap dataset. In particular, to answer our opening question, our transfer learning system is able to both match and surpass the performance of end-to-end CNN systems trained on LeafSnap. This is a surprising result, and testifies to the impressive power of deep ResNets to generate high-quality embeddings for new tasks.

7 Conclusion & Future Work

In this study, we have shown that transfer learning systems offer a high-precision, accessible solution for the task of fine-grained visual classification. All 6 of our top models outperform the state-of-the-art while utilizing relatively simple classifiers that train quickly and require minimal resources.

Future work should begin with fixing mislabeled examples in the Leafsnap Dataset and re-training and re-evaluating our models. It would also be valuable to explore the use of the state-of-the-art pretrained DenseNet, as DenseNet has been shown to be able to achieve similar performance to ResNet with 50% - 90% fewer parameters.

8 Contributions

Krishna handled dataset analysis, wrote the code to generate the image embeddings, and made the poster while Chris handled the review of relevant work, wrote the code for training and evaluating models on the embeddings, and finished the post-milestone report.

Code for this project is available at:
https://github.com/koenig125/229-final-project

References

Pierre Barré, Ben C. Stöver, Kai F. Müller, and Volker Steinhage. 2017. Leafnet: A computer vision system for automatic plant species identification. *Ecological Informatics* 40:50 – 56. https://doi.org/https://doi.org/10.1016/j.ecoinf.2017.05.005.

Elena Galbally, Krishna Rao, and Zoe Pacalin. 2018. Leafnet: A deep learning solution to tree species identification http://cs230.stanford.edu/projects$_{s}pring_{2}018/reports/8291236.pdf$.

Task Siang Thye Hang, Atsushi Tatsuma, and Masaki Aono. 2016. Bluefield (kde tut) at lifeclef 2016 plant identification task http://ceur-ws.org/Vol-1609/16090459.pdf.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition. *CoRR* abs/1512.03385. http://arxiv.org/abs/1512.03385.

Neeraj Kumar, Peter N. Belhumeur, Arijit Biswas, David W. Jacobs, W. John Kress, Ida C. Lopez, and João V. B. Soares. 2012. Leafsnap: A computer vision system for automatic plant species identification. In Andrew Fitzgibbon, Svetlana Lazebnik, Pietro Perona, Yoichi Sato, and Cordelia Schmid, editors, *Computer Vision – ECCV 2012*. Springer Berlin Heidelberg, Berlin, Heidelberg, pages 502–516.

Sue Han Lee, Chee Seng Chan, Paul Wilkin, and Paolo Remagnino. 2015. Deep-plant: Plant identification with convolutional neural networks. *CoRR* abs/1506.08425. http://arxiv.org/abs/1506.08425.

Pierre Sermanet, David Eigen, Xiang Zhang, Michaël Mathieu, Rob Fergus, and Yann LeCun. 2013. Overfeat: Integrated recognition, localization and detection using convolutional networks. *CoRR* abs/1312.6229.

Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. 2014. How transferable are features in deep neural networks? *CoRR* abs/1411.1792. http://arxiv.org/abs/1411.1792.