# Difference in details: transfer learning case study of ”cryptic” plants and moths

Varvara Vetrova
University of Canterbury
varvara.vetrova@canterbury.ac.nz

Sheldon Coup
University of Canterbury
sheldon.coup@pg.canterbury.ac.nz

Eibe Frank
University of Waikato
eibe@waikato.ac.nz

Michael Cree
University of Waikato
cree@waikato.ac.nz

## Abstract

*Can we classify species of very similar looking organisms quickly and accurately using only out of the box feature transfer? What if we only have small number of images? This experimental paper is part of on-going project on species recognition research and evaluates transfer learning and fine-tuning approaches on two highly specialized fine-grained datasets. The two fine-grained datasets were specifically assembled for the purpose of this research. These datasets consist of images of New Zealand native species of moths and ”cryptic” plants of Genus Corposma found also in New Zealand. We compare results from fine-tuning experiments with performance of transfer learning without fine-tuning. The latter results are based on features extracted from various levels of depth in the InceptionV3 network, including fully connected layers. The extracted features serve as inputs to a number of classification algorithms. We observe contrasting results for the two datasets. For the dataset of moths, the method based on features extracted from deep levels of the InceptionV3 network outperforms fine-tuning in accuracy (90.09% versus 87.18%). This is not the case for the dataset of cryptic plants (60.46% versus 74.37%). Despite both datasets being fine-grained in nature, these experimental differences could be attributed to intrinsically different morphology of organisms and warrant further investigation.*

## 1. Introduction

In this paper, we tackle two challenging fine-grained identification tasks: automatic classification of plants of the genus Coprosma and New Zealand species of native moth. In general, it is often difficult to distinguish species of biological organisms using only morphology. Plants of genus Coprosma that occur in New Zealand (NZ), especially, contain species that are hard to differentiate even by expert botanists. NZ native moth species present a similar level of challenge.

Automated species identification based on morphological features is a fine-grained classification task because small details can be crucial for distinguishing between classes. Other examples of fine-grained classification include recognition of faces [2] and vehicles [3]. There has been significant interest recently in harnessing and optimising the power of deep learning-based architectures in fine-grained classification problems [9, 4]. [8] provides an overview of recent applications of fine-grained recognition, including car identification and recognition of biological organisms. [1] reviews recent advances in the field of automated plant identification. [5] attempts to automatically classify species of organisms on a large scale.

There is a wide range of potential applications of models that can automatically identify biological species, in areas such as bio-security, precision agriculture, biodiversity mapping and citizen science. However, there are a number of associated challenges. The first challenge is small variability between classes and large variability within classes—this is common among all fine-grained recognition tasks. It is amplified in the case of plant recognition because the differences between species can be very subtle, unlike models of cars. Another challenge is the lack of sufficiently large datasets because acquiring and labeling is a costly and time consuming process requiring specialist expertise.

Due to the difficulties of the species recognition task, these particular image classification challenges could serve as a good case-study for evaluating the limits of convolutional networks, the current state-of-the-art in image classification.

In this paper, we analyse the performance of methods based on feature extraction from various depths of an InceptionV3 network and compare these methods with a baseline obtained by fine-tuning an InceptionResNetv2 network, for two challenging fine-grained datasets.

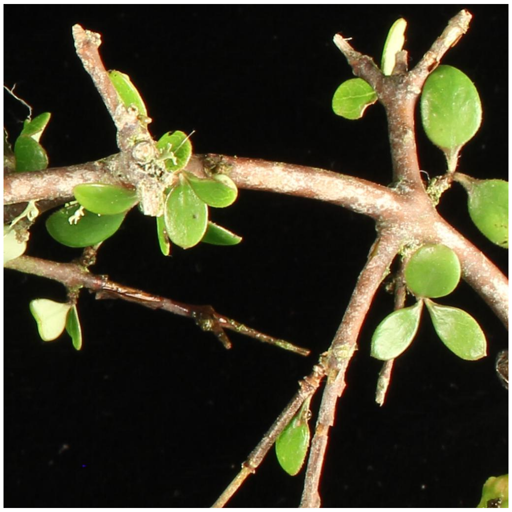

Figure 1. Plant of genus Coprosma, Crassifolia species.

## 2. Datasets

### 2.1. Plants of genus Coprosma

There are 17 species of Coprosma represented in this dataset. For each species, there are one to eight plants (mean 4.9) for a total of 83 plants, and there are typically ten images per plant (each of a different branch of the plant, mean 9.9), for a total of 83 plants and 819 images (Fig. 1). Each image contains $5184 \times 3456$ pixels and is stored as a high quality JPEG image. Images were taken by placing each branch on a black background. A ruler is present in the top left-hand corner of each image to indicate scale. In our experiments, the ruler was removed before further processing occurred by setting the pixels of the ruler to black. 3799 non-overlapping crops of size $1024 \times 1024$ pixel were extracted from the images.

### 2.2. NZ native moth

The NZ moth dataset contains 11 species of moth found in New Zealand (Fig. 2). Each species has between 2 and 54 images of individual specimens (mean of 20) meaning that there is a large class imbalance present. All of the images have a set width of 1181 pixels and varying height. All images are taken from a straight on dorsal view of the specimens. In each image, a scale bar is present in each bottom left corner. This scale bar was removed for training and testing.

## 3. Methods

### 3.1. Baseline experiments: fine-tuning

Firstly, a convolutional neural network based on the InceptionResNetV2 architecture and pre-trained on the Imagenet dataset, was fine-tuned on the two respective datasets.

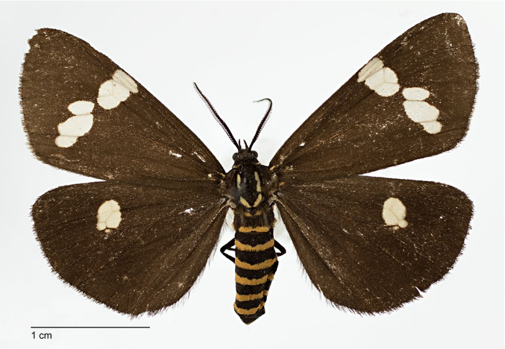

Figure 2. Moth specimen, Nyctemera annulata x amicus.

The results were taken to form a baseline in our experiments. Taking into account the relatively small number of images in the datasets, a 5-fold cross-validation was used to give a more accurate evaluation of model performance.

The following settings were utilized in fine-tuning experiments:

- Moth dataset: 5000 steps per run, learning rate of 0.1, learning rate decay factor of 0.97, learning rate decaying every 15 epochs.
- Coprosma dataset: 12000 steps per run, learning rate of 0.01, learning rate decay factor of 0.94, learning rate decaying every 5 epochs.

### 3.2. Feature transfer experiments

[6] showed that using convolutional layers as a feature extractor and utilizing an alternative classifier on the extracted vector representation can work as well as fine-tuning. In addition, [10] demonstrated that extracting feature vectors from a range of depths in the CNN network can yield improvements in performance.

In this section, we analyse performance of a number of different classification algorithms trained on the feature representations extracted from the images in the two datasets. We have performed two sets of experiments: utilising features extracted from the last pooling layer and combining it with the features extracted from deeper layers of the InceptionV3 network. The details of experiments are outlined in the following sections.

In both of these feature transfer experiments, in order to improve the estimate of accuracy for each classifier, Monte-Carlo cross-validation was used with 100 random splits of data into training/testing sets of proportion 80/20.

#### 3.2.1 Features extracted from the final pooling layer

Firstly, a forward pass of the InceptionV3 network, pre-trained on ImageNet, was performed on the images from the respective datasets. Then, the 2048-dimensional vector output from the final pooling operator in the InceptionV3 network was extracted. Finally, the extracted vectors were used as feature inputs to 7 classifiers. The default settings of the classifiers in the Python scikit-learn library were used in the experiments.

### 3.2.2 Feature extraction from varying depths

In this approach, the final vector representation of each image is found by concatenating the feature vector used in the previous section with an additional feature vector extracted from an earlier layer. As the InceptionV3 network is comprised of several ”Inception Blocks” in sequence, each feature vector was extracted between these blocks. The extractions were of varying 3D shapes, therefore compression of these representations into a vector was performed as follows. The size $n$ of the longest dimension was found and a vector of length $n$ was created where the $i$-th value in the vector was set to the maximum value in the $i$-th 2D array along the longest axis of the extraction. This procedure is equivalent to taking the maximum response of each filter. Principal component analysis was applied to the extracted vectors in order to reduce dimensions to 128. The same transformations were applied to the test sets independently of the training sets. Final classification was performed using a linear support vector classifier.

## 4. Results

The combined results from the sets of experiments described in 3.2.1 and the fine-tuning experiments are summarized in Table 1. As can be seen for both Coprosma and Moth datasets, the feature transfer approach using only features from the last pooling layer in the InceptionV3 network leads to a decrease in classification accuracy compared to base-line fine-tuning. The difference is especially pronounced in the case of the Coprosma dataset: 74.37% baseline accuracy versus 60.42% accuracy of the best classifier in feature transfer. The drop of accuracy is much less for the moth dataset: 87.18% for baseline fine-tuning versus 80.14% accuracy obtained by the best classifier.

Interestingly, feature transfer using features extracted from various depths of the InceptionV3 network outperforms fine-tuning for the moth dataset by 2.91% - 90.09% versus 87.18% respectively (Table 2). However, feature transfer does not lead to an increase in accuracy over the baseline in the case of the Coprosma dataset: 60.46% versus 74.37% for the baseline.

The contrasting results for the two datasets can potentially be attributed to fundamental differences in morphology between them. In the case of the Coprosma dataset, differences between species can be seen in different sizes of leaves and the arrangement of the leaves on branches. Therefore, there is a spatial structure present in this dataset that may explain why the features extracted from network pre-trained on the general purpose dataset ImageNet are not sufficient to obtain good results. On the other hand, in the moth dataset, differences between species can be seen in small details of the wings and antennae, which could explain why general features extracted from the deeper layers of the InceptionV3 network are able to capture it. Nevertheless, this result warrants further investigation.

| Classifier | Coprosma Acc | Moth Acc |
|---|---|---|
| SVC (linear) | 58.80% | 80.14% |
| SVC (radial) | 49.64% | 74.45% |
| Extra Trees | 49.28% | 74.50% |
| Random Forests | 50.29% | 75.41% |
| K Nearest Neighbour | 45.84% | 70.77% |
| Multilayer Perceptron | 60.42% | 78.00% |
| Gaussian Naive Bayes | 49.46% | 72.14% |
| **InceptionResNetV2** | **74.37%** | **87.18%** |

Table 1. Evaluation results of fine-tining experiments and the sets of experiments described in 3.2.1.

| Extracted layer | Coprosma Acc | Moth Acc |
|---|---|---|
| mixed/join:0 | 59.30% | 89.91% |
| mixed_1/join:0 | 60.01% | 88.32% |
| mixed_2/join:0 | 57.74% | 90.09% |
| mixed_3/join:0 | 59.54% | 87.45% |
| mixed_4/join:0 | 60.46% | 85.86% |
| mixed_5/join:0 | 59.38% | 88.27% |
| mixed_6/join:0 | 59.94% | 87.59% |
| mixed_7/join:0 | 57.91% | 87.72% |
| mixed_8/join:0 | 57.69% | 86.95% |
| mixed_9/join:0 | 59.88% | 86.41% |

Table 2. Evaluation of feature extraction from various depths within the InceptionV3 architecture, as described in 3.2.2.

## 5. Conclusions and future work

The comparison of accuracy estimates obtained from two types of feature transfer experiments and a baseline provided by fine-tuning yielded contrasting results for the moth and Coprosma datasets studied in this paper. The differences can potentially be attributed to the morphology of the images in the datasets. We are planning to extend our investigation into datasets of similar nature. In particular, it would be of interest to extract subsets of similar morphology from the iNaturalist databaset and investigate whether pre-training a network on a large set of species images from iNaturalist improves accuracy. Other avenues of further investigation include multi-modal models and capsule networks [7].

## References

[1] J. Carranza-Rojas, H. Goeau, P. Bonnet, E. Mata-Montero, and A. Joly. Going deeper in the automated identification of herbarium specimens. *BMC evolutionary biology*, 17(1):181, 2017.

[2] N. Z. B. C. D. Weihong, J. Hu and J. Guo. Fine-grained face verification: FGLFW database, baselines, and human-DCMN partnership, 2017.

[3] T. Gebru, J. Krause, Y. Wang, D. Chen, J. Deng, and L. Fei-Fei. Fine-grained car detection for visual census estimation. In *AAAI*, volume 2, page 6, 2017.

[4] C. McCool, T. Perez, and B. Upcroft. Mixtures of lightweight deep convolutional neural networks: applied to agricultural robotics. *IEEE Robotics and Automation Letters*, 2(3):1344–1351, 2017.

[5] J. Mo, E. Frank, and V. Vetrova. Large-scale automatic species identification. In *Australasian Joint Conference on Artificial Intelligence*, pages 301–312. Springer, 2017.

[6] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. CNN features off-the-shelf: An astounding baseline for recognition. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2014.

[7] S. Sabour, N. Frosst, and G. E. Hinton. Dynamic routing between capsules. In *Advances in Neural Information Processing Systems*, pages 3859–3869, 2017.

[8] S. Yu, Y. Wu, W. Li, Z. Song, and W. Zeng. A model for fine-grained vehicle classification based on deep learning. *Neurocomputing*, 257:97–103, 2017.

[9] B. Zhao, J. Feng, X. Wu, and S. Yan. A survey on deep learning-based fine-grained object classification and semantic segmentation. *International Journal of Automation and Computing*, 14(2):119–135, 2017.

[10] L. Zheng, Y. Zhao, S. Wang, J. Wang, and Q. Tian. Good Practice in CNN Feature Transfer. 2016.