# Image-based plant identification with taxonomy aware architecture

Jack Min Ong<sup>1</sup>, Sze Jue Yang<sup>1</sup>, Kam Woh Ng<sup>2</sup> and Chee Seng Chan<sup>1</sup>

<sup>1</sup>CISiP, Universiti Malaya, Malaysia <sup>2</sup>CVSSP, University of Surrey, U.K.

#### Abstract

This paper describes our approach for LifeCLEF Plant Identification Task 2022. Chan's Temple's solution was able to achieve a mean reciprocal rank (MRR) of 0.5104 using an ensemble of ResNets and 0.4880 with a single ResNet34. Our work shows that taxonomy-aware training schemes can offer a better performance gains over training schemes that do not utilize the taxonomy information.

#### Keywords

plant identification, taxonomy aware, convolutional neural networks, fine-grained image recognition

### 1. Introduction

It is estimated that there are about 374,000 described and documented plant species [1]. As such, a solution that is capable to automatically classify and retrieve specific plant species from images will greatly accelerate the work of ecologists working on important domains such as agriculture, construction or pharmacopoeia.

The LifeCLEF [2] Plant identification task 2022 aimed to encourage work towards such a solution by introducing a fine-grained image classification task containing 80,000 plant species with a well annotated dataset of more than 2 million images [3]. The images are also grouped into observations where each plant observation can contain multiple images of the plant, usually of different parts, to better mimic the images that would be taken by an ecologist in the field (behavior).

For our approach, we trained a few residual networks (i.e., ResNet [4]) using different methods of incorporating the taxonomy information into the learning objective. Our idea is that the taxonomy information will alleviate some of the class sparsity issues with optimization on large number of classes. It might also be reasonable to think that when ecologists are searching for observations or using a model for identification in the field, they would also look for those species that are close to the target species. If the model is unable to make a correct prediction, then predicting species that are closer in the taxonomy is thus desirable. As such, using training schemes that incorporate taxonomy into the loss function would yield models that are more

- https://kamwoh.github.io/ (K. W. Ng); http://cs-chan.com/ (C. S. Chan)
- © 0 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CLEF 2022: Conference and Labs of the Evaluation Forum, September 5–8, 2022, Bologna, Italy

<sup>☆</sup> wid190507@siswa.um.edu.my (J. M. Ong); wid190061@siswa.um.edu.my (S. J. Yang); kamwoh.ng@surrey.ac.uk (K. W. Ng); cs.chan@um.edu.my (C. S. Chan)

https://github.com/Jackmin801 (J. M. Ong); https://github.com/jasonyang429 (S. J. Yang);

CEUR Workshop Proceedings (CEUR-WS.org)



**Figure 1:** Family pre-training procedure. We first pre-train the model on classifying the 483 plant families and then fine-tune the model to the 80,000 plant species.

likely to predict species that are closer in the taxonomy when making an incorrect prediction which is better than predicting a species that is far away in the taxonomy [5].

# 2. Related Works

The methodology that used multiple classification heads for hierarchical labelled data has been explored for standard computer vision image recognition benchmarks like CIFAR100 [6] and ImageNet [7] in [5]. However, our method use residual networks [4] as the backbone network in contrast with [5] which uses VGG [8] as their backbone network. It is shown that residual network has a better performance as well as being much easier to optimize [9, 10].

# 3. Methodology

For all the trained models, we use ResNet models with the last fully connected layer replaced with a new fully connected layer initialized with kaiming initialization [11] with output dimension equal to the number of classes in the target taxonomy. The networks are initialized with weights pre-trained on ImageNet [7]. The learning process involves 2 steps. We first pretrain the models with classification heads from other taxonomical levels. The species classification head is then used to finalize the model.

# 3.1. Family Pre-training

Our first approach was to pre-train the model on classifying 483 plant families in the trusted dataset. We then swap out the last fully connected layer to a new one that has an output dimension of 80,000. This is to correspond to the number of species.



**Figure 2:** Multihead pre-training architecture. We first pre-train the model on classifying all the plant taxonomies and then fine-tune the model to 80,000 plant species, similar to the family pre-train procedure.

#### 3.2. Multihead Pre-training

The second approach was to pre-train the model on classifying all the plant taxonomies at the same time. We then remove the other classification heads, and only leave the head corresponding to the species.

# 4. Training Setup

#### 4.1. Dataset

There were 2 datasets provided for the task, a trusted dataset and a web dataset. The trusted dataset contained 2,886,761 images of 80,000 plant species from academic sources with higher certainty of the class labels being correct. The web dataset contained 1,071,627 images of 57,314 plant species from online sources. The web dataset was semi-automatically revised by the competition authors to reduce the number of errors and duplication.

For our approach, we only trained on the trusted dataset as it contains all the classes and is less noisy compared to the web dataset. We used splits of the web dataset, which were

Table 1Training hyperparameters used by all models trained

| Parameter        | Value                 |  |
|------------------|-----------------------|--|
| Batch Size       | 128                   |  |
| Input Image Size | 224 x 224 x 3         |  |
| Optimizer        | Adam [12]             |  |
| Learning Rate    | $10^{-4}$             |  |
| Loss Function    | Softmax Cross Entropy |  |

#### Table 2

Differing hyperparameters used by models trained

| Model                | Pretrain epochs | Classid train epochs | Resize train epochs |
|----------------------|-----------------|----------------------|---------------------|
| ResNet34 + Family    | 20              | 10                   | 1                   |
| ResNet34 + Multihead | 20              | 0                    | 3                   |
| ResNet50             | 0               | 0                    | 15                  |
| ResNet50-Wide        | 0               | 0                    | 10                  |

unobserved during training, for internal evaluation.

#### 4.2. Data Augmentation

For our submissions, two different pipelines of image augmentation were used. The first pipeline was a random resized crop of size 224 x 224. The second pipeline was a resize of size 224 x 224 with deformation of the content.

The random resized crop was used for the pre-training epochs. The resized crop will then be used for the last epochs and for submission generation.

We did not use any normalization as our initial experiments showed that the normalization has no effect on the model performance after one epoch of training.

#### 4.3. Target Mapping

In order to keep the class mappings consistent, the classes in each taxonomy of the trusted dataset were sorted and then being assigned an integer index based on their lexicographic ranking. These taxonomies to integer mappings are saved in a file and used by the data loaders and submission generation code to keep track of which integer corresponds to which class and vice versa.

#### 4.4. Hyperparameters

The hyperparameters used for training all the models are summarized in Table 1. The parameters that differ are summarized in Table 2.

| Part   | MRR on Web-1-per-species | Occurrences in Trusted dataset |
|--------|--------------------------|--------------------------------|
| flower | 0.2637                   | 72,509                         |
| habit  | 0.1910                   | 12,718                         |
| leaf   | 0.1722                   | 42,414                         |
| fruit  | 0.1477                   | 9,702                          |
| bark   | 0.1232                   | 13,463                         |

 Table 3

 Performance by part with Resnet34+ Family Pretrain model

# 5. Validation and Submission Generation

#### 5.1. Web-1-per-species

In order to evaluate the performance of our models without using up the submission limits, we created a split of the web dataset containing 1 image from each species. The reason for this is that task will use a version of the mean reciprocal rank (MRR) that is weighted by the species. We only took 1 because many of the species only have 1 image present in the dataset.

#### 5.2. Prediction aggregation by performance ranking

The test dataset contained multiple images of different parts of the same plant observation. However, the models were trained to predict on individual images. We thus need a method to aggregate the predictions made on the individual images for each observation.

Of the methods tried, we found that sorting the predictions based on the number of times they appear in the top 30 of the images and then by the part with the best individual performance (3) yielded the best performance on the competition evaluation.

The aggregation is also extended without needing any modifications to aggregate the predictions for the ensemble model.

# 6. Results

The results of our internal evaluations and submitted results are summarized in Table 4. The number of resize train epochs in Table 2 were determined by the epoch that achieved the best performance on the internal evaluations before the MRR began to drop.

Interestingly, the models with pretraining on the random resized crop did not benefit much from the training on the resized dataset for the species prediction task. They however do get an initial performance boost from the changes in augmentation.

# 7. Conclusion

In this paper, we presented our approach of using the plant taxonomy information to train models for fine-grained image classification in the LifeCLEF Plant identification task 2022. Our results showed that incorporating the taxonomy information will improve the performance of

# Table 4Model results

| Model                            | Web-1-per-species | Competition Submission |
|----------------------------------|-------------------|------------------------|
| Resnet34 + Family                | 0.2458            | 0.49994                |
| Resnet34 + Multihead             | 0.2370            | 0.47447                |
| Resnet50                         | 0.2089            | NA                     |
| Resnet50-Wide                    | 0.2019            | NA                     |
| Ensemble of 4 models (3rd Place) |                   | 0.51043                |
| Neuon Al's Model (2nd Place)     |                   | 0.60781                |
| Mingle Xu's Model (1st Place)    |                   | 0.62692                |

the models on the plant species identification task. It also shows that combining the prediction from multiple models is able to minimize some of the variance in the models, as well as improving the performance but at the cost of increased computational cost.

# Acknowledgments

This research is supported by the UM-IIRG (IIRG006C-19FNW) from Universiti Malaya, while the resources for this project were provided by the Universiti Malaya Data Intensive Computing Centre (DICC), as well as the Centre of Image and Signal Processing (CISiP) from Faculty of Computer Science and Information Technology, Universiti Malaya.

# References

- [1] M. Christenhusz, J. Byng, The number of known plant species in the world and its annual increase, Phytotaxa 261 (2016) 201–217. doi:10.11646/phytotaxa.261.3.1.
- [2] A. Joly, H. Goëau, S. Kahl, L. Picek, T. Lorieul, E. Cole, B. Deneu, M. Servajean, A. Durso, H. Glotin, R. Planqué, W.-P. Vellinga, A. Navine, H. Klinck, T. Denton, I. Eggel, P. Bonnet, M. Šulc, M. Hruz, Overview of lifeclef 2022: an evaluation of machine-learning based species identification and species distribution prediction, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2022.
- [3] H. Goëau, P. Bonnet, A. Joly, Overview of PlantCLEF 2022: Image-based plant identification at global scale, in: Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, 2022.
- [4] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016) 770–778.
- [5] Z. Yan, H. Zhang, R. Piramuthu, V. Jagadeesh, D. DeCoste, W. Di, Y. Yu, Hd-cnn: Hierarchical deep convolutional neural networks for large scale visual recognition, 2015 IEEE International Conference on Computer Vision (ICCV) (2015) 2740–2748.
- [6] A. Krizhevsky, Learning multiple layers of features from tiny images, University of Toronto (2009).
- [7] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy,

A. Khosla, M. S. Bernstein, A. C. Berg, L. Fei-Fei, Imagenet large scale visual recognition challenge, International Journal of Computer Vision 115 (2015) 211–252.

- [8] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, CoRR abs/1409.1556 (2015).
- [9] D. Balduzzi, M. Frean, L. Leary, J. P. Lewis, K. W. Ma, B. McWilliams, The shattered gradients problem: If resnets are the answer, then what is the question?, CoRR abs/1702.08591 (2017). URL: http://arxiv.org/abs/1702.08591. arXiv:1702.08591.
- [10] A. Zaeemzadeh, N. Rahnavard, M. Shah, Norm-preservation: Why residual networks can become extremely deep?, IEEE Transactions on Pattern Analysis and Machine Intelligence 43 (2021) 3980–3990.
- [11] K. He, X. Zhang, S. Ren, J. Sun, Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, 2015 IEEE International Conference on Computer Vision (ICCV) (2015) 1026–1034.
- [12] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, CoRR abs/1412.6980 (2015).