Deep Learning for Large-Scale Plant Classification: NEUON Submission to PlantCLEF 2023

Sophia Chulif[1,2,*], Yang Loong Chang[2] and Sue Han Lee[1]

[1]Swinburne University of Technology Sarawak Campus, 93350, Sarawak, Malaysia
[2]Department of Artificial Intelligence, NEUON AI, 94300, Sarawak, Malaysia

Abstract

This paper presents the methods and submissions made by our team to PlantCLEF 2023, an image-based plant identification task involving 80,000 plant species. Like PlantCLEF 2022, the task aimed to identify species based on 26,868 multi-image plant observations from a test set of 55,306 images. Given a training dataset of 4 million photos, we trained several Inception-v4 and Inception-ResNet-v2 models and submitted ten runs. Compared to the previous year, we experimented with more data augmentations, different batching methods and trained separate organ models, specifically for the labels: bark, flower, fruit, habit, and leaf. Our highest-performing run, comprising several ensembled models, achieved a macro-averaged mean reciprocal rank of 0.61813, increasing from 0.6078 in last year's performance (PlantCLEF 2022) through increased data augmentations.

Keywords

plant identification, multi-organ, convolutional neural networks, machine learning, computer vision

1. Introduction

Accurate identification of plants on a global scale helps us understand and conserve Earth's biodiversity. It allows naturalists to study species diversity [1], assess the impact of environmental changes, and predict ecological responses to climate change [2]. Besides this, it plays a vital role in agriculture [3], food security [4], and medicinal research [5]. In many ways, plants are beneficial to all life-form. Nevertheless, it remains a challenge to identify all plant species on Earth [6]. Global plant-scale identification involves processing massive variations of plant features, taxonomic information, and ecological attributes. Capturing subtle differences among these species may require more effort for machine algorithms and even for human experts. To tackle this problem, the PlantCLEF 2023 challenge [7, 8] was introduced to classify a large multi-image dataset with 80,000 plant species. Likewise, the same training and test sets from PlantCLEF 2022 [9] were provided. These datasets remain the largest plant dataset published to date and represent a realistic global plant identification task concerning a vast number of species, strongly unbalanced, partially incorrect identifications, duplications, and diversified images.

This paper describes our training strategies, submissions, and results obtained in PlantCLEF 2023. In short, our training strategies include increasing data augmentation, applying bal-

CLEF 2023: Conference and Labs of the Evaluation Forum, September 18–21, 2023, Thessaloniki, Greece
*Corresponding author.
schulif@swinburne.edu.my (S. Chulif); yangloong@neuon.ai (Y. L. Chang); shlee@swinburne.edu.my (S. H. Lee)

anced batching, and employing multi-organ and single-organ training schemes. Firstly, data augmentation is a common technique to increase the training dataset by performing various transformations on the images to tackle the class-imbalance problem and to avoid overfitting [10]. In addition, it has been shown that simple augmentation methods can contribute considerably to the results, which are very good relative to more complex methods in some cases [11]. Therefore, we experimented with increased augmentation during training by applying more bi-cubic resizing, random hue, and random contrast to our dataset. Next, we employed class-balancing in our batching method to deal with the long-tail characteristics of the PlantCLEF 2023 dataset, in which many classes have significantly lower training samples and fewer classes have more training samples. Class re-balancing is easy to implement in principle among the various long-tailed learning, but it can lead to comparable or superior results [12]. Consequently, we limit the training images for a species to a maximum of 16 in an epoch in hopes of reducing any bias towards the majority classes and preventing poor performance on the underrepresented species. Lastly, multi-organ integrated training [13] has been shown to improve the model performance compared to single-organ plant classification. Therefore, we utilised the available manual and predicted organ tags to experiment with the multi-organ and single-organ training schemes to evaluate how they would perform in the context of PlantCLEF 2023.

From the above approaches, we find that the data augmentation method enhances our model the best and has helped us achieve a macro-averaged mean reciprocal rank (MA-MRR) of 0.61813, increasing from 0.6078 in last year's performance (PlantCLEF 2022), producing the second best results in the PlantCLEF 2023 challenge.

2. Datasets

Two datasets were provided in PlantCLEF 2023, the training dataset and the test dataset. The training dataset was grouped into two: Trusted and Web. The Trusted dataset comprised 2,886,761 higher-quality images from academic sources and collaborative platforms. On the other hand, the Web dataset included 1,071,627 images from search engine queries suffering significant inaccuracies. The Trusted and Web datasets contain metadata such as Class, Order, Family, Genus, Species, manual tag, predicted tag, and predicted tag probability. Nevertheless, not all images were annotated with the manual and predicted tags. Meanwhile, the test dataset consists of 55,306 images categorised into 26,868 plant observations used for the challenge's observation-level plant classification. Like the Trusted and Web datasets, the test dataset contained the manual tag metadata. In addition, we partitioned separate datasets from the training dataset provided to cater to our multi-organ and single-organ training schemes. Since not all images were annotated with tags, some images were not included in the single-organ training. We used the images annotated with the tags of "bark", "flower", "fruit", "habit", and "leaf". Furthermore, we excluded the tags with less than 0.7 in their predicted probability score. Table 1 shows the training datasets used for our multi-organ and single-organ training approaches. Note that we pre-processed the training datasets and removed all additional duplicate images sharing the same filename, resulting in a reduced training dataset size compared to the original training dataset.

Table 1
Details of our training datasets partitioned for training.

Dataset	Source	No. of images	No. of species
Multi-organ	Trusted	2,821,933	80,000
Multi-organ	Trusted & Web	3,893,560	80,000
Bark Single-organ	Trusted & Web	54,937	19,430
Flower Single-organ	Trusted & Web	358,969	45,346
Fruit Single-organ	Trusted & Web	88,639	26,824
Habit Single-organ	Trusted & Web	2,088,292	76,511
Leaf Single-organ	Trusted & Web	89,001	21,771

Table 2
Details of our models' hyperparameters.

Hyperparameter	Values
Batch Size	128
Input Image Size	$299 \times 299 \times 3$
Optimizer	Adam Optimizer
Initial Learning Rate	0.0001
Weight Decay	0.00004
Learning Dropout rate	0.2
Loss Function	Softmax Cross Entropy

3. Method

3.1. Model Setup

Our models were set up using Tensorflow 1.12 and TF-Slim library with the hyperparameters described in Table 2. Like our submissions to PlantCLEF 2022 [14], our models were based on the Inception-v4 and Inception-ResNet-v2 architectures initialised on the weights pre-trained from ImageNet. We added a batch normalisation layer and five fully-connected layers after the final layers to cater to the taxonomy classifications: Class, Order, Family, Genus, and Species. Since multi-task classification has been shown to improve classification compared to a single label classification in our previous PlantCLEF submissions [15, 14], likewise, we implement multi-task classification in our models catering to the five labels: Class, Order, Family, Genus, and Species. In contrast to our PlantCLEF 2022 submissions, we experimented with several new strategies: increased data augmentation, a balanced batching method, and a multi-organ and single-organ training scheme.

Increased Data Augmentation
Our models were trained with random cropping, horizontal flipping, and colour distortion applied to our images. In contrast to our models from last year, we increased the data augmentations in our training by performing more bi-cubic resizing, random hue and random contrast to our images.

Balanced Batching Method
In addition, we adopted a balanced batching method in training our models. We limit the selection of training images for a species in an epoch to a maximum of 16 samples. We chose 16 as it would be suitable since the lowest sample number in the training dataset is 1, while the highest sample number in the dataset is over 100. It aims to avoid bias towards any particular species and prevents poor performance on underrepresented species.

Multi-organ and Single-organ Training Scheme
Furthermore, we implemented two training schemes: multi-organ and single-organ. In multi-organ training, we trained our model with the entire dataset (like our submissions last year). Using all the available organ parts, the classification is trained on the species. On the other hand, in single-organ, we trained our model based on the individual organ tags provided as mentioned in Section 2, which includes Bark, Flower, Fruit, Habit, and Leaf. The classification of a single-organ model is based on the specific organ and their respective species.

3.2. Inference Methods

The model predictions are obtained by the Softmax function as well as the feature embedding comparison method.

Softmax function
The softmax function is commonly used as the activation function in the output layer of a neural network for multi-class classification. It allows the network to produce class probabilities over multiple classes. It accepts a vector of input data and assigns a probability value to each class. The probability generated is a vector of numbers between 0 and 1 and these values would sum up to 1. The class with the highest probability indicates the most confident predicted class. The Softmax function is represented in Equation 1 where x represents the input vector, and e^x denotes the exponential function raised to the power of x. The denominator, $\sum_i e_i^x$, calculates the sum of the exponentiated values over all elements of the input vector.

$$\text{softmax}(x) = \frac{e^x}{\sum_i e_i^x} \tag{1}$$

Feature embedding comparison
This method refers to comparing the test image's features to the train images' features. It involves transforming the train and test images into lower-dimensional feature vectors (embeddings) using the learned representation obtained from the fully-connected layers of our model, then comparing them to measure their similarity. Essentially, this method include two processes: the feature dictionary generation and feature similarity comparison.

1. **Feature dictionary generation**
   First, we generated a list of train images (feature dictionary) to compare with the test images. Since the training data consists of over 3 million images and it is computationally intensive to compare all the photos, we only selected a maximum of ten images for

each of the 80,000 plant species from the Trusted training dataset. Consequently, our feature dictionary is made from a list of 592,258 training images. These images are then transformed into feature embeddings and averaged according to their respective classes. Therefore, the feature dictionary is composed of 80,000 feature embeddings.
Note that to obtain the feature embedding of each training image, we applied ten crops to the image and then averaged them to get a single feature embedding. The ten crops involve cropping the top-left, top-right, bottom-left, bottom-right, and centre, all of which are also horizontally flipped to get ten variations. The total feature embeddings of each species are then averaged over the total number of images it has in the list of the dictionary, making a feature dictionary of 80,000 feature embeddings.

2. **Feature similarity comparison**
After acquiring the train feature dictionary of 80,000 feature embeddings, we obtain the test image embeddings by performing the same ten crops technique to the test images. Since the test images' classes are unknown, the embeddings are averaged according to the test observation id instead of the species class id as in the train feature dictionary generation. Therefore, since there are 26,868 observation ids, the total test embeddings generated result in 26,868 test feature embeddings.
Next, we used cosine similarity to measure the similarity between the train feature dictionary and the test feature embeddings. The calculated cosine similarity score is then transformed with Inverse Distance Weighting into probabilities for ranking the classes. The weights for the transformed embedding vector were calculated as in Equation 2 where P_i is the weight assigned to a specific test feature embedding, d_i is the distance between the test feature embedding and the dictionary feature embedding, d_k^n is the distance between the test feature embedding and the dictionary feature embedding k, raised to the power of n, while $\sum_k$ is the sum over all dictionary feature embeddings. Similar to the Softmax function method, the class with the highest probability indicates the most confident predicted class.

$$P_i = \frac{(\frac{1}{d_i})}{\sum_k (\frac{1}{d_k^n})} \tag{2}$$

4. Submissions

4.1. Model Variations

We submitted ten runs to PlantCLEF 2023, consisting of several models described in Table 3. Essentially, they differ in their network architecture, data augmentation, batching method, and training data. The evaluation metric implemented in this challenge is the Macro Average (by species) Mean Reciprocal Rank (MA-MRR). The MA-MRR is a statistic measure for evaluating any process that generates a list of possible responses to an index of queries ordered according to the likelihood of correctness as defined in [7].

Table 3
Details of our trained models. "IR" refers to the Inception-ResNet-v2 model, while "'I" refers to the Inception-v4 model.

Model	Network architecture	Increased augmentation	Balanced batching	Training data
Multi-IR (Trusted)	IR	No	No	Multi-organ (Trusted)
Multi-I	I	No	No	Multi-organ (Trusted & Web)
Multi-IR	IR	No	No	Multi-organ (Trusted & Web)
Multi-IR-AUG	IR	Yes	No	Multi-organ (Trusted & Web)
Multi-IR-AUG-B	IR	Yes	Balanced	Multi-organ (Trusted & Web)
Single-IR-AUG-B-Bark	IR	Yes	Balanced	Single-organ (Trusted & Web)
Single-IR-AUG-B-Flower	IR	Yes	Balanced	Single-organ (Trusted & Web)
Single-IR-AUG-B-Fruit	IR	Yes	Balanced	Single-organ (Trusted & Web)
Single-IR-AUG-B-Habit	IR	Yes	Balanced	Single-organ (Trusted & Web)
Single-IR-AUG-B- Leaf	IR	Yes	Balanced	Single-organ (Trusted & Web)

Table 4
Performance of our submitted runs. The ones with (Feature embedding matching) indicates the predictions were made by the feature embedding method. Otherwise, they were obtained using the Softmax funtion.

Run	Model(s)	MA-MRR
9	Multi-I + Multi-IR + Multi-IR (Trusted) + Multi-IR-AUG	0.61813
7	Multi-I + Multi-IR + Multi-IR (Trusted) + Multi-IR-AUG	0.61561
10	Multi-IR + Multi-IR (Trusted) + Multi-IR-AUG	0.61406
5	Multi-IR-AUG	0.5504
1	Multi-IR-AUG	0.54242
2	Multi-IR-AUG-B	0.46606
6	Multi-IR-AUG (Feature embedding matching)	0.46476
8	Multi-IR-AUG (Feature embedding matching)	0.4591
3	Multi-IR-AUG (Feature embedding matching)	0.45242
4	Single-IR-AUG-B-Bark + Single-IR-AUG-B-Flower + Single-IR-AUG-B-Fruit + Single-IR-AUG-B-Habit + Single-IR-AUG-B-Leaf	0.33926

4.2. Results and Discussion

Our runs are tabulated in Table 4. Run 1, 2, 3, 5, 6, and 8 were based on the predictions from a single model, while Run 4, 7, 9, and 10 were based on several models ensembled. As expected, the ensembled models are more accurate than a single model prediction. Our highest-performing run (Run 9), an ensembled model, obtained an MA-MRR score of 0.61813. From the new strategies experimented with, we found that increasing data augmentation helped improve the models' accuracy. It is confirmed by comparing our run from PlantCLEF 2022 (Run 7 MA-MRR: 0.6078)

and our current run (Run 9 MA-MRR: 0.61813), which differ by only adding the new model (Multi-IR-AUG), which is trained with increased augmentations.

On the other hand, the balanced batching method did not help in improving the model's performance. Comparing Run 1 and 2, we see that the model without the balanced batching strategy (Run 1 MA-MRR: 0.54242) achieved a higher MA-MRR score compared to the model with the balanced batching method (Run 2 MA-MRR: 0.46606). Although it was not what we expected, it could be due to the reduction of training samples which may be inherently more important than others. Furthermore, the single-organ training scheme did not help in the predictions. Run 4, which comprised the single-organ models (bark, flower, fruit, habit, and leaf), performed the worst among all other runs based on the multi-organ training scheme. It suggests that training with multi-organ data is more effective than single-organ data as a combination of organs can provide more information to the model. Finally, we show that the predictions obtained by the feature embedding matching method (Run 3, 6, and 8 MA-MRR: 0.45242, 0.46476, 0.4591) are lower compared to the predictions obtained by the Softmax function method (Run 1 and 5 MA-MRR: 0.54242, 0.5504). It is likely due to the clearer decision boundary obtained when using the Softmax function, as this approach relies on the confidence of the network's prediction. On the other hand, feature embedding comparison might involve more nuanced comparisons between feature vectors, which can lead to a loss of discriminative information. This loss could result in reduced accuracy and lower performance.

5. Conclusion

We trained several Inception-v4 and Inception-ResNet-v2 models and submitted ten runs to PlantCLEF 2023. To improve our performance in this year's challenge, we tried several new strategies that differed from our previous submissions in PlantCLEF 2022. These strategies included three key approaches: increased data augmentation, a balanced batching method, and a multi-organ and single-organ training scheme. Among these approaches, increased data augmentation improved our model predictions by increasing our MA-MRR score from 0.6078 to 0.61813. The balanced batching method did not work out as intended but is possibly the result of the degradation of the head classes with many training samples at the expense of improving the tail classes with fewer training samples from under-sampling [12]. Moreover, the lower performance of the ensembled single-organ training models is likely due to the reduction of training data in the single-organ training scheme. Since not all training images were tagged with their organ details and we did not utilise the images with a predicted tag score of less than 0.7, there was a significant loss of information compared to using the entire dataset in multi-organ training.

Acknowledgments

The resources of this project is supported by NEUON AI SDN. BHD., Malaysia.

References

[1] H. Qian, J. Zhang, M.-C. Jiang, Global patterns of fern species diversity: An evaluation of fern data in gbif, Plant Diversity 44 (2022) 135–140.

[2] C. Parmesan, M. E. Hanley, Plants and climate change: complexities and surprises, Annals of botany 116 (2015) 849–864.

[3] S. L. Cappelli, L. A. Domeignoz-Horta, V. Loaiza, A.-L. Laine, Plant biodiversity promotes sustainable agriculture directly and via belowground effects, Trends in Plant Science (2022).

[4] M. A. Uebersax, K. A. Cichy, F. E. Gomez, T. G. Porch, J. Heitholt, J. M. Osorno, K. Kamfwa, S. S. Snapp, S. Bales, Dry beans (phaseolus vulgaris l.) as a vital component of sustainable agriculture and food security—a review, Legume Science 5 (2023) e155.

[5] S.-S. Huang, C.-H. Huang, C.-Y. Ko, T.-Y. Chen, Y.-C. Cheng, J. Chao, An ethnobotanical study of medicinal plants in kinmen, Frontiers in Pharmacology 12 (2022) 681190.

[6] G. Rouhan, M. Gaudeul, Plant taxonomy: A historical perspective, current challenges, and perspectives, Molecular plant taxonomy: Methods and protocols (2014) 1–37.

[7] H. Goëau, P. Bonnet, A. Joly, Overview of plantclef 2023: Image-based plant identification at global scale., Working Notes of CLEF – Conference and Labs of the Evaluation Forum (2023) (2023).

[8] A. Joly, C. Botella, L. Picek, S. Kahl, H. Goëau, B. Deneu, D. Marcos, J. Estopinan, C. Leblanc, T. Larcher, R. Chamidullin, M. Šulc, M. Hrúz, M. Servajean, H. Glotin, R. Planqué, W.-P. Vellinga, H. Klinck, T. Denton, I. Eggel, P. Bonnet, H. Müller, Overview of lifeclef 2023: evaluation of ai models for the identification and prediction of birds, plants, snakes and fungi., in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2023.

[9] H. Goëau, P. Bonnet, A. Joly, Overview of plantclef 2022: Image-based plant identification at global scale, Working Notes of CLEF (2022).

[10] C. Shorten, T. M. Khoshgoftaar, A survey on image data augmentation for deep learning, Journal of big data 6 (2019) 1–48.

[11] C. Lei, B. Hu, D. Wang, S. Zhang, Z. Chen, A preliminary study on data augmentation of deep learning for image classification, in: Proceedings of the 11th Asia-Pacific Symposium on Internetware, 2019, pp. 1–6.

[12] Y. Zhang, B. Kang, B. Hooi, S. Yan, J. Feng, Deep long-tailed learning: A survey, IEEE Transactions on Pattern Analysis and Machine Intelligence (2023).

[13] P. Guo, Q. Gao, A multi-organ plant identification method using convolutional neural networks, in: 2017 8th IEEE International Conference on Software Engineering and Service Science (ICSESS), 2017, pp. 371–376. doi:10.1109/ICSESS.2017.8342935.

[14] S. Chulif, S. H. Lee, Y. L. Chang, A global-scale plant identification using deep learning: Neuon submission to plantclef 2022, Working Notes of CLEF (2022).

[15] S. Chulif, K. J. Heng, T. W. Chan, M. A. Al Monnaf, Y. L. Chang, Plant identication on amazonian and guiana shield flora: Neuon submission to lifeclef 2019 plant., in: CLEF (Working Notes), 2019.