Plant Identification on Amazonian and Guiana Shield Flora: NEUON submission to LifeCLEF 2019 Plant

Sophia Chulif, Kiat Jing Heng, Teck Wei Chan, MD Abdullah Al Monnaf, and Yang Loong Chang

Department of Artificial Intelligence, NEUON AI, 94300 Sarawak, Malaysia
http://www.neuon.ai
{sophiadouglas,kiatjing,chanteckwei,abdullah,yangloong}@neuon.ai

Abstract. This paper will look into the use of fine-tuned Inception-v4 and Inception-ResNet-v2 models to automate the classification of 10,000 plant species. Prior to training the networks, the training dataset was pre-processed to remove the noisy data. The team submitted three runs which achieved comparable performances to human experts on the test dataset comprising 745 observations for all the evaluation metrics. For the trained systems to generalise better, the systems were trained for multi-task classification and is able to classify plant images based on their species, with support of their genus and family labels. In particular, an ensemble of Inception-v4 and Inception-ResNet-v2 networks achieved a Top-1 accuracy of 0.316 and 0.246 for the test set identified by experts and the whole test set respectively.

Keywords: Plant identification, computer vision, convolutional neural networks, data cleaning

1 Introduction

The plant identification challenge of the Conference and Labs of the Evaluation Forum (CLEF) is an annual challenge focusing on the automation of plant identification. In recent years, automated identification of plants has become a high interest for botanical specialists and computer experts alike [3]. Transitioning from the conventional computer vision through feature descriptor methods [14], deep learning based methods have been able to significantly improve accuracy of the automation of plant identification [7,8,11]. This application is essential in the utilisation, management and conservation of flora of any kind. When the plant identification challenge first commenced in 2011, the main focus was to be able to identify over 70 different tree species given 3996 leaf images as training data [5]. The total number of species and training images in the challenge has then

notably increased each year. With the training dataset initially covering only leaf images, it has now expanded to different plant organs and multiple views of the plant. Since 2017, the total number of plant species has reached 10,000 along with the presence of noisy images (images that are unrelated to the plant of interest) in the training dataset - this posed a great challenge to the participants in evaluating the extend of a noisy dataset competing with a trusted dataset [2]. Furthermore, the plants in the training dataset shared a low intra-class and high inter-class similarity. They may belong to the same class but look clearly different from one another [1]. Besides the variance, some of the distinct plant organs and views may not be captured in the given training images, resulting in the lack of training data thus difficulty in predicting the species. This depicts the realistic challenge of the real-world applications in plant identification.

For PlantCLEF 2019, the dataset provided was focused on the Guiana shield and the Amazon rainforest which is known to be the largest collection of living plants and species in the world [6,4]. The task was to predict 10,000 plant species and the participants were provided with 448,071 images of training data. Along with a test set containing 2,974 images and 745 observations, the main evaluation method of the competition was the Top 1 accuracy of the submitted predictions based on the 745 observations.

The primary challenge of PlantCLEF 2019 was the decreased average number of images per species in the training dataset. Many species contained only a few images and some even contained only one image. Moreover, the training dataset consisted of noisy data that constituted from non-plant images and duplicate images that may come with incorrect labels.

This paper presents the preparation made prior to the training of the system. The system was fine-tuned from pretrained Inception-v4 and Inception-ResNet-v2 models to automate the classification of the given 10,000 plant species.

2 Data Preparation

2.1 Data Analysis

The dataset provided consisted of 10,000 species which mainly focused on the Guiana shield and the Amazon rainforest. The reported training dataset size is 448,071 (based on the number of rows in PlantCLEF2019MasterTraining.csv) however, the downloaded training dataset size was 433,810. Therefore, the actual downloaded dataset included the training dataset of 433,810 images which consisted of 10,000 classes (species), and a test dataset of size 2,974 from 745 different observations.

As mentioned in the challenges data collection description, among these 10,000 species many contain only a few images and some of them contain only one image. Moreover, it was observed that there were many variations in the training dataset that could affect the performance of the plant identification.

Other than real-world plant images i.e. fruit, branch, trunk, stem, root, flower, leaf etc., the training dataset initially contained many images that do not

look like real actual plants i.e. sketches, paintings, illustrations, plant herbariums, and small regions of the interested plants. In addition, the dataset also contained non-plant images i.e. animals, logos, book covers, graphs, table, medicine bottles, humans, chemical-bond diagrams, presentation slides, maps etc. Besides this, the dataset consisted of images with duplicate names in different classes (folders). Furthermore, the dataset consisted of duplicate images with different names within the same folder as well as in different folders. Having different labels for the same images could cause confusion to the machine.

These characteristics observed could affect the performance of the plant identification therefore the approach was to pre-processed the dataset.

2.2 Data Cleaning

From the analysis in subsection 2.1, the dataset was then pre-processed (cleaned) to allow better execution of plant identification. First, the images with duplicate names were removed as those images are actually the same. The dataset given consisted of 433,810 images and 10,000 classes of plants. After eliminating the duplicate names, the dataset was reduced to 279,183 images and 8,468 classes.

Then, the duplicate images (without having the same name) were further removed. After removing them, 263,987 images were left with 8,419 classes. Finally, the non-plant images were eliminated, resulting in a total of 250,646 images and 8,263 of classes for training[1].

In the approach of eliminating images with duplicate names in different folders, since there is no method to decide which class they actually belong to (no experts to verify), all images with the same name were removed.

On the other hand, in order to remove duplicate images within the same folder as well as different folders, inception-v4's feature extractor layer (Mixed_7d) was used to compare images so that those with 0.99 cosine similarities in features can be eliminated. If the duplicate images only exist in the same class (folder), only one of the images will be retained. Then, the difference hash algorithm was used to detect more duplicates within the dataset. The hash value for every image was calculated and then compared to get those with very little difference. Images with little difference hash values mean they are identical. Likewise, if the identical images only exist in the same class (folder), only one image is retained, the rest are eliminated.

To detect non-plant images, a discrimination network for identifying plant and non-plant images was trained. This process consisted of 3 phases.

In phase 1, the positive samples (plant) were taken from PlantClef 2016 while the negative samples (non-plant) were taken from ImageNet2012 (excluding the plant classes). The training dataset size was 4,000, with each class having 2,000 samples. Meanwhile the validation dataset size was 2,000, with each class having 1,000 samples. An Inception-V4 plant and non-plant classifier was then trained using these datasets.

[1] The cleaned list can be found at Github via `https://github.com/changyangloong/plantclef2019_challenge`

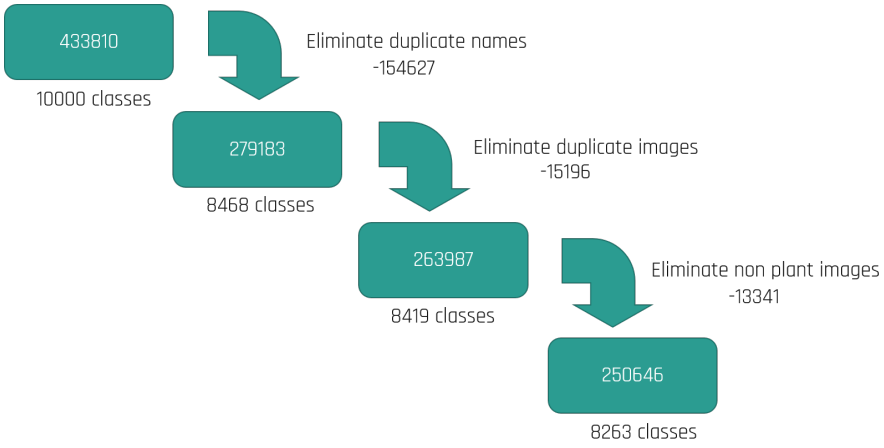

Fig. 1. Data cleaning process.

In phase 2, 5,000 samples were randomly selected from PlantClef 2019 and predicted using the trained classifier. The performance on this sample was evaluated manually. After evaluating its performance, the training set was refined by manually correcting the prediction and adding them back to the training set. Then, the Inception-V4 plant and non-plant classifier was retrained using the new training set. Phase 2 was repeated until the performance satisfied the accuracy of over 90%. In this case it was repeated 4 times.

In phase 3, the Inception-V4 plant and non-plant classifier was applied to the whole dataset to remove the non-plant images. Only the softmax probability of 0.98 for the non-plant class was regarded as non-plant images.

The entire process is visualised in Fig. 1.

3 Methodology

This section will describe the motivations for using Inception-v4 and Inception-ResNet-v2 for the plant classification challenge, the training setup, network hyperparameters and validation results obtained from the trained networks.

3.1 Network Architecture

The backbone networks used for classifying PlantCLEF 2019 images were Inception-v4 and Inception-ResNet-v2 models. These two models were adopted in this plant classification task as they come with the lowest Top-1 Error and Top-5 Error for single-crop - single-model experimental results, when being proposed in the original paper [12]. Besides, the use of inception modules allows filters with multiple sizes to perform convolution on the same level to cater to the variation

in the location of the information on the image [13], which is applicable on the training dataset.

Both of the networks implement Convolutional Neural Network (CNN) architecture, which typically consists of convolutional layers, pooling layers, dropout layers and fully-connected layers. The models were fine-tuned from pre-trained weights on the ImageNet dataset [10].

In fact, Inception-v4 and Inception-Resnet-v2 networks share an identical stem as the input part of these networks. However, these two models demonstrate differences in architecture after the stem.

For Inception-v4 model, there are three main inception modules following the stem, namely Inception-A, Inception-B and Inception-C modules. The output will then be concatenated and fed to the next module. Meanwhile, there is a reduction module after Inception-A and Inception-B modules to alter the width and height of the grid. These modules work together to serve as the model's feature extractor.

On the other hand, Inception-ResNet-v2 model makes use of residual connections in addition to inception modules after the implementation of stem. This allows Inception-ResNet-v2 to achieve higher accuracies within a shorter time frame, while being similarly computationally expensive as the Inception-v4 model. For residual connections to work, the pooling layers from pure Inception modules are replaced with residual connections. Similarly, there is a respective reduction module following Inception-ResNet-A and Inception-ResNet-B modules.

Both Inception-v4 and Inception-ResNet-v2 model have the same structures for the classification part, which consists of an average pooling layer, a dropout layer, and a fully-connected layer which return the softmax probabilities over predicted output classes.

3.2 Training Setup

During the network training, two types of classification methods were investigated, namely single label classification and multi-task classification.

Single Label Classification The first classification method is the conventional classification method which is based on single label prediction model. The labels are the plant species, therefore there are 10,000 classes.

Multi-Task Classification Another method used in this plant identification task is the multi-task classification whereby the labels "Species","Genus" and "Family" of the samples were used in training. This allowed the network to regularise and generalise better on images from a large number of classes. The total number of family class is 248, while the genus class is 1,780 [2].

[2] The multi-task label can be found at `https://github.com/changyangloong/plantclef2019_challenge`

Library Used The networks were implemented using TensorFlow Slim library, with the weights being pre-trained on ImageNet datset [10]. The networks were then fine-tuned accordingly using the hyperparameters described in Sub-section 3.4. Since the adopted models pre-trained on ImageNet have only 1,000 classes, the fully-connected layer of the adopted models was adapted to 10,000 classes during transfer learning for PlantCLEF 2019. For multi-task classification, there were two additional fully-connected layers which were catered to 248 "Family" classes and 1,780 "Genus" classes.

Training Data The input image size of the training was $299 \times 299 \times 3$. By separating 20,000 samples from the training samples as a validation set, the total remaining training set comprised 230,646 images. Although the total number of classes is reduced from 10,000 classes to 8,263 classes, the network was still trained to classify 10,000 classes through class mapping. This allows model update when missing classes are to be found.

3.3 Data Augmentation

Data augmentation was performed on the training images to increase the training sample size. With this, the CNN network can learn features that are invariant to their locations in the images and various transforms, which then effectively reduces the chance of overfitting [9]. Random cropping, horizontal flipping and colour distortion (brightness, saturation, hue, and contrast) of images were applied to the training dataset to increase the possibility of classifying the correct plants regardless of their different environments or orientations.

3.4 Network Hyperparameters

The learning dropout rate was set to 0.2 (keeping 80 % of the neurons before the fully-connected layer) while the optimizer used was Adam Optimizer. The optimizer took on an initial learning rate of 0.0001. Meanwhile, the gradient was clipped at 1.25 to prevent the occurrence of exploding gradients. Softmax cross entropy loss was used to compute the error between the predicted labels and true labels; while the L2 regularization loss was added with weight decay of 0.00004. The network hyperparameters can be summarised in Table 1.

3.5 Validation

Prior to training, 20,000 samples were randomly separated from the cleaned training dataset of 250,646 images as validation set. It was ensured that each of the remaining 8,263 classes in the training set has at least one sample left for training. A validation sample was not added if there is only one sample left for training in each class. In other words, a class will not have a validation sample if it has only one sample.

Table 1. Network hyperparameters used for training networks.

Parameter	Value
Batch Size	256
Optimizer	Adam Optimizer
Initial Learning Rate	0.0001
Gradient Clipping	1.25
Loss Function	Softmax Cross Entropy
Weight Decay	0.00004

There were 4 approaches in testing the validation set on 4 different kinds of network. The approaches include "Top 1 Centre Crop", "Top 1 Centre Crop + Corner Crop", "Top 5 Centre Crop", and "Top 5 Centre Crop + Corner Crop". Meanwhile the 4 different networks included "Inception-v4", "Inception-v4 Multi-task", "Inception-ResNet-v2 Multi-task", and "Inception-v4 Multi-task + Inception-ResNet-v2 Multi-task" with different dataset sizes. Note that all the trained networks were validated upon the same 20,000 validation images.

The "Centre Crop" approach considers the centre region of the sample. The centre region was cropped and resized then passed into the network for testing. The "Corner Crop" approach on the other hand focused on the centre, top left, top right, lower left, and lower right region of the image. Each region was cropped and resized then passed into the network for testing. The "Top 1" and "Top 5" approaches represent the Top 1 and Top 5 predictions based on the testing results.

3.6 Inference Procedures

The following procedures were adopted for inferencing the predictions of the test dataset. There were three models used for inference, namely "Multi-Task Inception-v4", "Multi-Task Inception-ResNet-v2" and "Multi-Task Inception-v4 + Multi-Task Inception-ResNet-v2".

1. The test images with the same observation ID were grouped together.
2. A total of five center crop and corner crop images were then produced for a single test image.
3. The test images grouped under the same observation ID were fed into the CNN models for label predictions, as shown in Fig. 2. This would mean a total of five predictions for a single test image.
4. The probabilities of each prediction were averaged over the total number of test image crops grouped under the same observation ID.
5. The top 100 probabilities with value greater than 0.001 were collected for result tabulations.
6. Step 2 to 5 were repeated for every different observation ID.

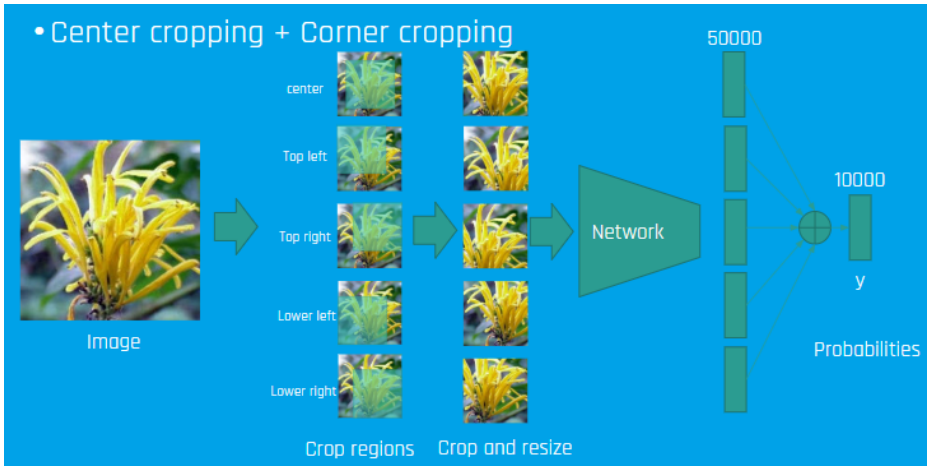

Fig. 2. Process of feeding test images into the trained models.

4 Results and Discussions

4.1 Validation Results

Inception-v4 and Inception-ResNet-v2 were the 2 networks used for training and classification in these 4 approaches. The Multi-task description represents the classification on multiple labels, i.e. Species, Family and Genus.

The validation results are shown in Table 2. The results have shown that while eliminating the duplicate image could increase the performance of the single-task network, multi-task classification method was able to achieve a even higher accuracy on the validation dataset.

Interestingly enough, removing non-plant images after filtering duplicate images out leads to a slight drop in the single-task Inception-v4 network's performance. Since no Inception-v4 network is trained without the non-plant images (i.e. keeping potentially duplicate images but of plants only) as benchmark, there is no way to deduce whether the presence of duplicate images or non-plant images is the main factor to the performance drop in our current experiments.

4.2 Submitted Runs

The team submitted a total of three runs based on three different trained models which achieved the top-3 validation accuracy. The models are described in the followings:

Holmes Run 1 utilises fine-tuned Inception-v4 model catered to multi-task classification (which are Species, Genus and Family).

Table 2. Validation accuracy for different networks.

Network	Dataset Size	Top 1 Center Crop	Top 1 Center Crop + Corner Crop	Top 5 Center Crop	Top 5 Center Crop + Corner Crop
Inception-v4	433,810	42.83 %	43.70 %	61.48 %	62.67 %
Inception-v4	279,183	46.87 %	47.97 %	64.16 %	65.08 %
Inception-v4	250,646	46.73 %	47.76 %	63.87 %	64.95 %
Multi-Task Inception-v4	250,646	48.87 %	49.68 %	65.20 %	65.94 %
Multi-Task Inception-ResNet-v2	250,646	48.30 %	49.10 %	65.07 %	65.72 %
Multi-Task Inception-v4 + Multi-Task Inception-ResNet-v2	**250,646**	**51.96** %	**52.68** %	**65.79** %	**68.47** %

Holmes Run 2 is an ensemble of Inception-v4 and Inception-ResNet-v2, and is also catered to multi-task classification.

Holmes Run 3 summarises prediction results from fine-tuned Inception-ResNet-v2 catered to multi-task classification.

The run files were formatted as `ObservationID; ClassID; Probability; Rank`.

4.3 Sample of Predictions

Fig. 3 depicts the prediction output of test images grouped under the same observation ID. This was to visualise the predictions generated by the models for evaluating the model performance.

4.4 LifeCLEF 2019 Plant Challenge results

The team submitted three runs with Holmes Run2 being the best performance among our submission in terms of Top-1 accuracy,Top-3 accuracy and Top-5 accuracy. Holmes Run 2 achieved Top 1 accuracy of 31.6% for the test set identified by experts, despite the occurrence of missing classes after data cleaning. This proves that using an ensemble of two different state-of-the-art CNN models can increase the robustness of the system and return better predictions. Table 3 summaries the results achieved by the team.

Mean Reciprocal Rank (MRR) was also used to evaluate the performance of the submitted runs. The MRR is the average of the reciprocal ranks of the whole test set, with the reciprocal rank of a query response being the multiplicative inverse of the rank of the first correct answer. It can be mathematically represented as shown, with $\mid Q \mid$ being the frequency of plant occurrences in the test set.

Fig. 3. Visual representation of predictions for test images with observation ID = 348.

Table 3. Validation accuracy for different networks.

Team Run	Rank	Top 1 Test Set Identified by Experts	Top 1 Whole Test Set	Top 3 Test Set Identified by Experts	Top 5 Test Set Identified by Experts	Top 5 Whole Test Set	MRR Test Set Identified by Experts	MRR Whole Test Set
Holmes Run 2	1	**0.316**	**0.247**	**0.376**	**0.419**	**0.357**	**0.362**	**0.298**
Holmes Run 3	2	0.282	0.225	0.359	0.376	0.321	0.329	0.274
Holmes Run 1	3	0.248	0.222	0.325	0.368	0.325	0.302	0.269

$$MRR = \frac{1}{\mid Q \mid} \sum_{n=1}^{|Q|} \frac{1}{rank_i} \tag{1}$$

Holmes Run 2, being our best performing machine, has achieved a MRR of 0.362 and 0.298 for the test set identified by experts and the whole test set respectively, as shown in Fig. 6.

At the same time, all the three runs submitted by the team outperform 1 out of 5 human experts of the Amazonian French Guina flora according to Top-1 (Fig. 7) and Top-3 accuracies. As for Top-5 accuracy, Holmes Run 2 outperformed 2 human experts, as captured in Fig. 8.

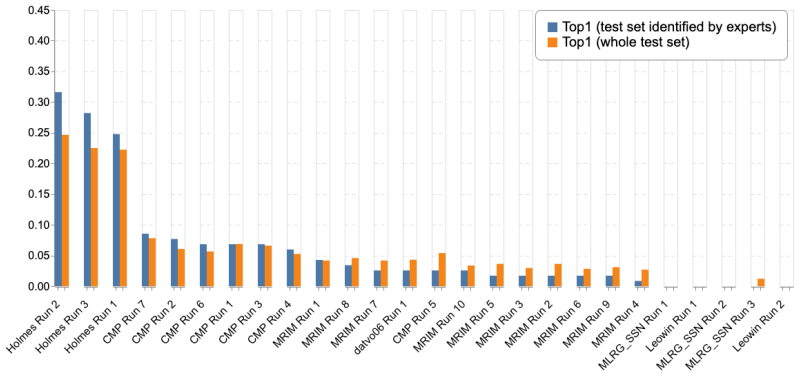

Fig. 4. Top-1 accuracy achieved by all the submitted runs.

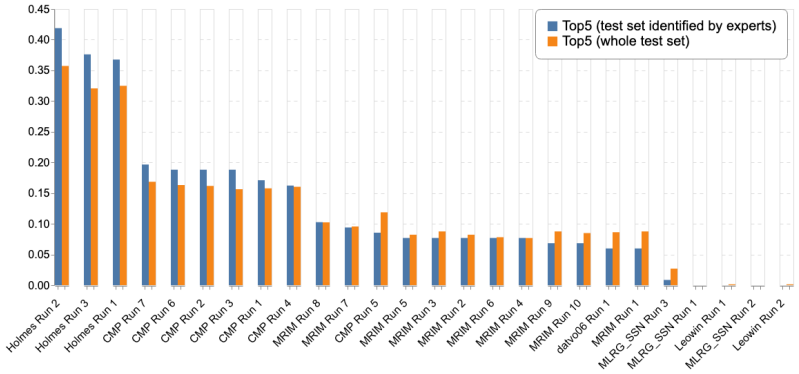

Fig. 5. Top-5 accuracy achieved by all the submitted runs.

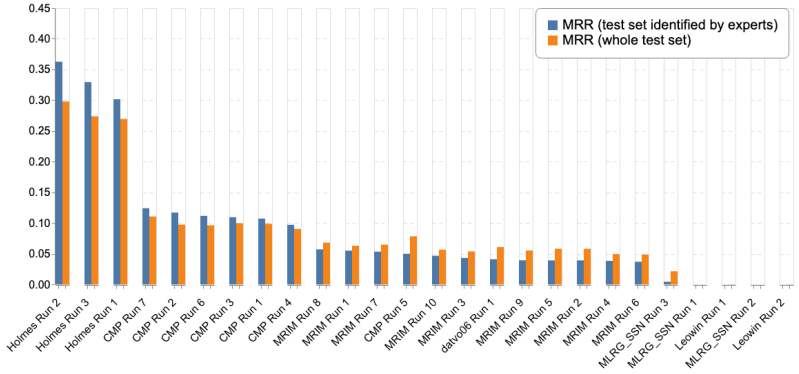

Fig. 6. MRR achieved by all the submitted runs.

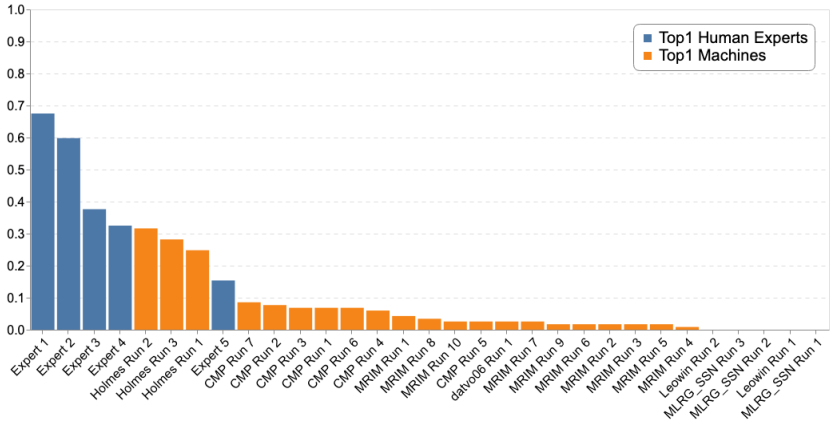

Fig. 7. Top-1 accuracy achieved by all the submissions including human experts.

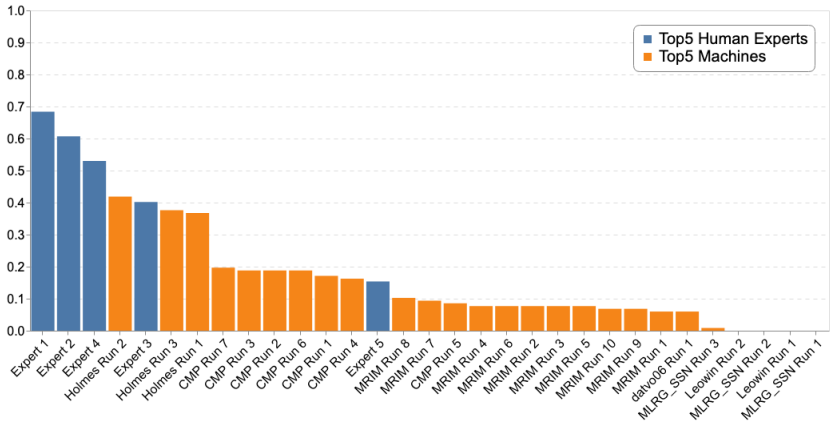

Fig. 8. Top-5 accuracy achieved by all the submissions including human experts.

4.5 Discussions

The automated identification of plants for PlantCLEF 2019 has shown a notable decrease in performance compared to the previous editions. This may be due to the decrease of per class samples as mentioned in the challenge's description. In such real-world condition, the dataset requires pre-processing before training is done. Although it has been seen that training with noisy data can be more profitable [2], the noisy data in this challenge worsens the performance of the automatic plant identification. At this stage of understanding, the team believe the portion of noisy images should not be occupying a large margin (duplicates file names are 55.13% out of the whole training set) of training set to act as good regularisation agent.

Therefore, the dataset was cleaned before training. This prevents the network from getting trained by large amount of noisy data such as the non-plant images and duplicate images/names (which may come with incorrect labels). Moreover, by adding multi-task classification in our system, it has helped in regularising the network and improving the performance. Modelling a relationship between Species, Genus and Family is essential for better plant recognition.

Since the cleaned dataset is reduced to 8,263 classes, there was no capability to distinguish the missing classes. Hence, the label for the missing classes was unable to be predicted if it is present in the test set resulting in a lower performance. Additionally, 20,000 images had been sidelined for validation purposes which are believed to be helpful in increasing the performance if they were added.

5 Conclusion

The task of the PlantCLEF 2019 challenge was to return the most likely matching species for each observation (a set of images of the same plant) of the test set. In this paper, the team has presented the overview and results of our approach in doing so. With regards to the diversity of the dataset, the team has found that the cleaning of the dataset and multi-task classification of Species, Genus and Family improved the prediction results.

According to the competition results released, our trained model was better than one of the experts in Top 1 and Top 3 accuracy. Additionally, it has a close performance with Expert 4. Our model even outperformed 2 experts for Top 5 accuracy. Overall, 3 of our submitted runs obtained the good results in every machine category.

The identification of plants based off images is indeed a difficult task even for some of the botanical experts. Relying on plant images alone is usually insufficient to determine the correct species as they may contain only partial information of the plant [3]. Based on the competition results of PlantCLEF 2018 and 2019, it can be considered that with sufficient data for training, machines are able to perform nearly equal or better than a human.

For future work, the plant images from the missing classes can be retrieved if the ground-truth labels are known and will be used for training the networks for better predictions.

Acknowledgment

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

References

1. Darwin, C.: The different forms of flowers on plants of the same species. John Murray (1877)
2. Goëau, H., Bonnet, P., Joly, A.: Plant identification based on noisy web data: the amazing performance of deep learning (lifeclef 2017). In: CLEF 2017-Conference and Labs of the Evaluation Forum. pp. 1–13 (2017)
3. Goëau, H., Bonnet, P., Joly, A.: Overview of expertlifeclef 2018: how far automated identification systems are from the best experts? In: CLEF 2018 (2018)
4. Goëau, H., Bonnet, P., Joly, A.: Overview of lifeclef plant identification task 2019: diving into data deficient tropical countries. In: CLEF working notes 2019 (2019)
5. Goëau, H., Bonnet, P., Joly, A., Boujemaa, N., Barthélémy, D., Molino, J.F., Birnbaum, P., Mouysset, E., Picard, M.: The imageclef 2011 plant images classi cation task. In: ImageCLEF 2011. pp. 0–0 (2011)
6. Joly, A., Goëau, H., Botella, C., Kah, l.S., Servajean, M., Glotin, H., Bonnet, P., Vellinga, W.P., Planqué, R., Fabian-Robert, S., Müller, H.: Overview of lifeclef 2019: Identification of amazonian plants, south & north american birds, and niche prediction. In: Proceedings of CLEF 2019 (2019)
7. Lee, S.H., Chan, C.S., Wilkin, P., Remagnino, P.: Deep-plant: Plant identification with convolutional neural networks. In: 2015 IEEE International Conference on Image Processing (ICIP). pp. 452–456. IEEE (2015)
8. Lee, S.H., Chang, Y.L., Chan, C.S., Remagnino, P.: Hgo-cnn: Hybrid generic-organ convolutional neural network for multi-organ plant classification. In: 2017 IEEE International Conference on Image Processing (ICIP). pp. 4462–4466. IEEE (2017)
9. Mikołajczyk, A., Grochowski, M.: Data augmentation for improving deep learning in image classification problem. In: 2018 international interdisciplinary PhD workshop (IIPhDW). pp. 117–122. IEEE (2018)
10. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: Imagenet large scale visual recognition challenge. International journal of computer vision 115(3), 211–252 (2015)
11. Sulc, M., Picek, L., Matas, J.: Plant recognition by inception networks with test-time class prior estimation. Working Notes of CLEF 2018 (2018)
12. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A.: Inception-v4, inception-resnet and the impact of residual connections on learning. In: Thirty-First AAAI Conference on Artificial Intelligence (2017)
13. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1–9 (2015)
14. Walsh, J., O' Mahony, N., Campbell, S., Carvalho, A., Krpalkova, L., Velasco-Hernandez, G., Harapanahalli, S., Riordan, D.: Deep learning vs. traditional computer vision (04 2019)