Herbarium 2021
Half-Earth Challenge – FGVC8

Ricardo B. Sousa

Abstract

Herbaria have available massive repositories of plant diversity data. The data contains the plant species, their reproductive state, and collection dates and locations. The Herbarium 2021: Half-Earth Challenge is a competition hosted as part of the Eighth Workshop on Fine-Grained Visual Categorization (FGVC8). This competition has a dataset that contains approximately 2.5M images of 64,500 different plant specimens. The main purpose of this work is to identify the plant species in each image. First, different ResNet models (ResNet-18, ResNet-34, and ResNet-50) were trained on the Herbarium 2021 dataset and were compared considering the training accuracy and the testing average Macro F-Score. Second, the ResNet-50 model was used with data augmentation techniques (flip and rotation) and compared to not using data augmentation. The experimental results showed that using the ResNet-50 with the data augmentation techniques implemented in this work led to a higher Macro F-Score than the ResNet-50 model only.

1. Introduction

For several centuries, botanists collected specimens data and stored this data in herbaria. There are approximately 3000 herbaria in the world. These herbaria represent massive repositories of plant diversity data. Indeed, the botanists catalog not only plant species but also the plant's reproductive state, and collection dates and locations. This data can be used to study the variability of species, their evolution, or their phylogenetic relationship, while representing snapshots of plant diversity through time [3].

One dataset that includes more than 2.5M images of specimens is *The Herbarium 2021: Half-Earth Challenge* [3]. The data is provided by several herbariums around the world representing approximately 65,000 plant species. Herbarium 2021 [3] is also a competition with this same name hosted as part of the Eighth Workshop on Fine-Grained Visual Categorization (FGVC8) and sponsored by the New York Botanical Garden (NYBG). The main goal of this competition is to train a model with the Herbarium 2021 [3] dataset to classify plant species.

Deep Convolutional Neural Networks (DCNN) have become state-of-the-art for image classification. These networks exploit multiple layers of nonlinear information processing for feature extraction and later classification [12]. Examples of DCNN-based networks are the AlexNet [8], VGG [13], GoogLeNet [16], ResNet [6], or EfficientNet [17]. One popular benchmark in image classification is testing, e.g., these networks in the ImageNet [2] dataset. The accuracy results demonstrated that using deep networks allows the improvement in image classification problems. In the scope of classifying plant species, some works also use deep learning models. Barré *et al*. [1] proposed the LeafNet to classify 185 tree plant species. Sun *et al*. [15] implemented a 26-layer deep learning model to classify 100 ornamental plant species. However, most of the existent works use deep learning but for a low number of different species (below 1000 specimens), or the methods highly depend on experts to encode domain knowledge.

This work uses the ResNet [6] model for the Herbarium 2021 [3] image classification problem. In the training phase, data augmentation is used to reduce the model's overfitting to the training data. The experimental results first compare the performance of ResNet-18, ResNet-34, and ResNet-50 without data augmentation and then a comparison between using the data augmentation techniques described in Section 4 and not using them with ResNet-50.

The paper is organized as follows. Section 2 presents the related work. Section 3 details the Herbarium 2021 [3] dataset. Section 4 describes the model implemented and the data augmentation techniques used. Section 6 presents the experimental results. Lastly, Section 7 presents the conclusions and future work.

2. Related Work

Image classification focus on associating a specific label to an image. Deep Convolutional Neural Networks (DCNN) have become state-of-the-art in this matter. Due to the use of multiple convolutional layers and nonlinear processing, DCNN are capable of extracting low and high-level

features to correspond a specific label to an image [12]. AlexNet [8] (Krizhevsky) was the pioneer in optimizing the GPU use for training a CNN winning the ImageNet 2012 competition with a 84.6% top-5 and a 63.3% top-1 accuracy. Since then, deeper CNN were used to achieve higher accuracy results for the image classification problem. VGG16 [13] (Simonyan and Zisserman) studied the increase of depth to 16–19 weight layers for large-scale image recognition (91.9% top-5 and 74.4% top-1 accuracy in ImageNet). GoogLeNet [16] (Szegedy *et al*.) improved the computing resources inside the network while using a 22 layers deep network (2014 ImageNet winner with 74.8% top-1 accuracy). ResNet [6] (He *et al*.) introduced identity shortcut connections to skip one or more layers to prevent degrading the network's performance when added more layers (hundreds or even thousands). ResNet-50 and ResNet-152 (50 and 152 layers, respectively) achieved a 78.25%/93.95% and a 78.57%/93.29% top-1/5 accuracy in ImageNet, respectively. More recently, Tan and Le [17] proposed in 2019 the EfficientNet [17] that focused on compound scaling to achieve better accuracy and efficiency (EfficientNet-B7 achieved a 84.4%/97.1% top-1/5 accuracy), while using less parameters than other networks. Even though the ImageNet [2] dataset can have more than 10,000 classes with nearly 10M images (large-scale version), the classes are not focused on plant taxonomy.

In terms of classification of plant specimens, DCNN-based models are also used. LeafNet [1] uses the concept of dimension reduction modules on a model proposed in the article with 14 layers. This network was used to classify 185 tree plant species from the North-eastern United States. LeafNet [1] accomplished an average top-1/5 accuracy of 86.3%/97.8% in the LEAFSNAP [9] dataset. Another work focused on plant taxonomy was elaborated by Sun *et al*. [15]. This work implemented a 26-layer ResNet [6] model to classify 100 ornamental plant species in Beijing Forestry University campus. Sun *et al*. [15] achieved a 91.78% accuracy in the BJFUI100 [15] test dataset. However, most existent works for plant taxonomy consider a lower number of different specimens compared to the Herbarium 2021 [3].

3. Dataset

The Herbarium 2021: Half-Earth Challenge [3] dataset was created from specimens provided by the New York Botanical Garden (NYBG), Bishop Museum (BPBM), Naturalis Biodiversity Center (NL), Queensland Herbarium (BRI), and Auckland War Memorial Museum (AK). The dataset contains 2,500,779 images of specimens from the Americas and Oceania. The data is approximately split 80%/20% for training/test: 2,257,759 and 243,020 images for the training and test datasets, respectively. In terms of the dataset format, [3] uses the COCO dataset format [10] and is organized by the standard taxonomic reference list LCVP v1.0.2 [4]. The dataset includes additional annotation fields: region and supercategory information.

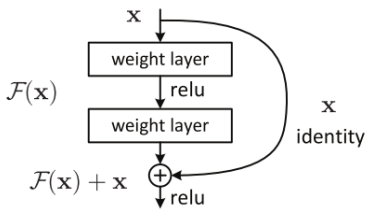


Figure 1: Residual block [6]

The dataset has 64,500 different species. There is a minimum of 3 images per species, with some species with more than 100 images. Each species has at least 1 instance in both training and test datasets. However, the test set distribution is slightly different from the training set. While the training set has species with hundreds of examples, the test set has a maximum of 10 examples per species.

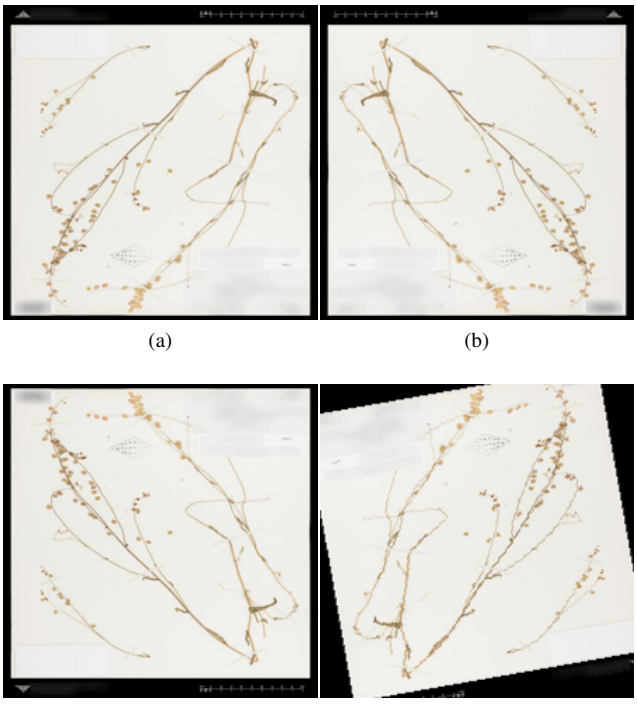
Lastly, the original images have different sizes. All the images were resized to a maximum of 1,000 pixels in the larger dimension and are in JPEG format.

4. ResNet

The main idea of ResNet [6] is the identity shortcut connections. As shown in Figure 1, the identity connection is added element-wise to the output of the residual layers. The purpose of these connections is to allow stacking deeper networks while avoiding the degradation problem of deep networks. This problem is already shown in [5, 14] where the accuracy of the networks gets saturated and then degrades rapidly (not because of overfitting, but due to adding more layers) with the network depth increasing.

Consequently, ResNet [6] models have several versions accordingly with the number of layers. In PyTorch [11] (open source machine learning framework), ResNet [6] has 5 versions with a different number of layers. The number of layers available is as follows: 18, 34, 50, 101, and 152 layers for each respective ResNet [6] model.

The advantage of stacking layers while avoiding degradation and the availability of different versions of ResNet [6] in PyTorch [11] led me to choose this deep network for the Herbarium 2021 [3] competition. The experimental results in Section 6 compare three ResNet [6] models with a different number of layers: 18, 34, and 50 layers. Also, it is compared the model ResNet-50 alone versus using data augmentation in the training phase. The data augmentation techniques used are discussed in the next section.



(a) (b) (c) (d)

Figure 2: Examples of data augmentation: (a) original image; (b) horizontal flip; (c) vertical flip; (d) horizontal + vertical flips and 10º rotation

5. Data augmentation

As discussed in Section 3, the Herbarium 2021 [3] dataset was provided by several herbaria and cataloged by different people. Indeed, the plant samples can have slightly different positions or orientations in the images for the same specimen. Therefore, geometric transformations of images such as flip or rotation are of great interest to apply in the training phase.

So, three geometric transformations are implemented for data augmentation in this work. Horizontal and vertical flips are the first transformations with a probability of 0.25. The flips have the advantage of not losing pixel information after they are applied. The last transformation is rotating the image up to 10º with a probability of 0.05. This probability is lower than the one for the flips transformations because rotating the image leads to pixel loss information. This consequence is also the reason why the angle is 10º and not higher. Examples of the three geometric transformations are illustrated in Figure 2.

Also, note that it was not considered data augmentation techniques based on the color of the pixels. The reason is that the specimens could be related to the color of the plants. So, changing the color properties of the image could lead to changing the domain of the data, which is not intended in this work.

6. Experimental Results

The experiments were performed in Kaggle [7]. Given that the Herbarium 2021 [3] competition is available in Kaggle [7], the main advantage is that the dataset required for the competition (occupies a total of 150.78 GB in disk space) is already available on the virtual machines hosted by Kaggle [7] notebooks. However, Kaggle [7] has accelerator quotas of 30h that are reset weekly. Moreover, it is not possible to train a network for 30 consecutive hours because Kaggle [7] also limits the time of a session: a maximum of 9h.

Furthermore, the models used for the classification problem were ResNet-18, ResNet-34, and ResNet-50. These models were trained on a virtual machine hosted by Kaggle [7] with an Intel Xeon CPU @ 2.00GHz and a Nvidia Tesla P100-PCIE-16GB. The optimizer chosen was the Adam optimization algorithm, and the cost function was the cross-entropy loss. The learning rate was set to 4×10^{-4}. In terms of batch size, the training of ResNet-18, ResNet-34, and ResNet-50 models used 512, 384, and 128 images, respectively. It was not possible to use batch sizes higher than these ones due to graphics memory constraints. All the training data provided in the Herbarium 2021 [3] dataset (2,257,759 images) were used for the training state. Also, the images were resized to the maximum size allowed by ResNet (224x224 resolution). In terms of training time duration, it was an average of 6h27m, 6h41m, and 7h for models ResNet-18, ResNet-34, and ResNet-50, respectively. As for the average inference time, it was 40m, 42m, and 45m for the same ResNet models, respectively. So, the workaround was to train a single epoch per session on the Kaggle [7] virtual machines and saving the model to a `pth` file after the inference on the test dataset (243,020 images).

Table 1 synthesizes the experimental results on the training and inference datasets without and with data augmentation (the latter for the ResNet-50 model). The training phase is evaluated in terms of accuracy per batch, i.e., the ratio between the number of correctly classified images and the total number of images in each batch. The test results are evaluated by the average Macro F-Score that is automatically calculated by Kaggle [7]. This score is calculated with approximately 30% (the final results of the competition are based on the other 70%) of the test data and the dataset does not provide ground-truth for the test set. In the next subsections, it is presented the discussion of the experimental results obtained in this work.

		Train				Test
Model	**Epoch**	Avg. loss per batch	Avg. acc. per batch (%)	Best loss per batch	Best acc. per batch (%)	Macro F-Score (%)
ResNet-18	1	5.7913	19.11	2.9748	44.92	9.487
	2	2.7901	47.21	1.8959	62.70	22.47
	3	1.9225	59.43	1.3774	69.53	29.37
	4	1.4718	66.87	1.0780	75.78	31.67
	5	1.1856	72.20	0.8291	80.66	32.90
	6	0.9771	76.47	0.6664	83.59	32.65
ResNet-34	1	6.1600	15.93	3.2792	41.93	7.479
	2	3.1264	43.69	2.0428	59.64	18.45
	3	2.0390	57.70	1.4444	69.27	23.94
	4	1.4915	66.20	0.9602	75.00	31.36
	5	1.1640	72.17	0.7587	79.95	31.29
ResNet-50	1	6.3501	15.49	2.7497	50.78	10.87
	2	2.6991	48.52	1.3253	69.53	24.28
	3	1.5133	64.78	0.7644	83.59	32.32
	4	1.0427	74.10	0.4512	89.06	34.59
	5	0.7581	80.50	0.2732	93.75	36.04
	6	0.5653	85.13	0.1960	96.09	34.75
ResNet-50 with data augmentation	1	6.8303	12.79	3.1683	46.09	8.330
	2	3.0905	46.15	1.5839	68.75	27.54
	3	1.7882	62.25	0.9384	79.69	33.65
	4	1.2707	70.65	0.6574	85.16	39.21
	5	0.9683	76.20	0.4312	89.06	45.52
	6	0.7643	80.30	0.2909	94.53	47.83
	7	0.6158	83.51	0.2629	94.53	49.85
	8	0.5029	86.10	0.1605	96.09	49.42

Table 1: Synthesis of the experimental results using ResNet models

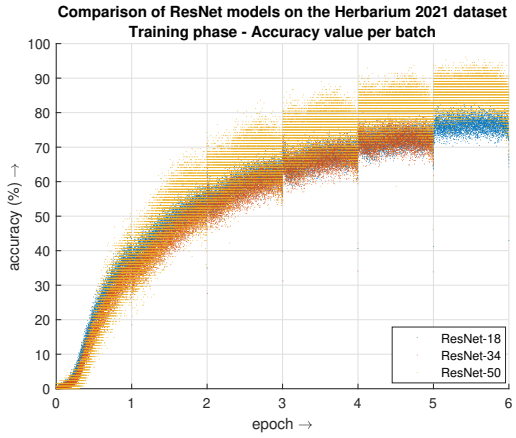
6.1. Comparison between ResNet-18, ResNet-34, and ResNet-50 w/o data augmentation

The first experiment made compared the following three ResNet [6] models without data augmentation on the training phase: ResNet-18, ResNet-34, and ResNet-50. The main goal was to evaluate if deeper models could improve the classification of plant specimens on the Herbarium 2021 [3] dataset. The results shown in Table 1 and Figure 3 demonstrate that the ResNet-50 model was better either in training and in testing phases. All the performance measures were higher than the other two ResNet [6] models. In the testing phase, ResNet-50 achieved a higher average Macro F-Score than ResNet-18 with an increase of 4% for the respective best epoch (36.04% vs 32.90%, respectively).

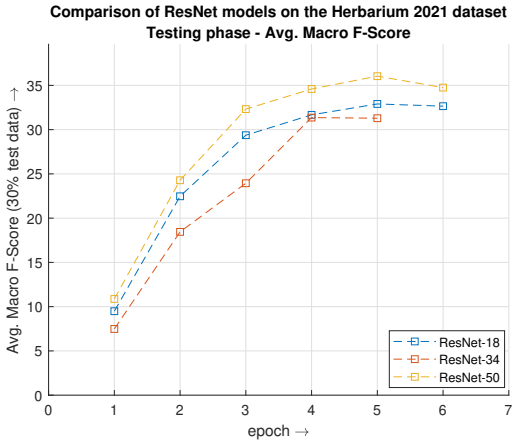
Relative to ResNet-18 and ResNet-34 models, it was not expected that ResNet-18 was better than ResNet-34. Even though models with a higher number of layers do not mean better results (the accuracy of the networks is saturated [5, 14]), ResNet-50 obtained better results than ResNet-18. However, it would be necessary more types of measures for the training phase (e.g., precision, recall, or the Macro F-Score for each class) to explain this result.

6.2. Comparison between ResNet-50 w/ and w/o data augmentation

Analyzing the results presented for the ResNet-50 models in Table 1, it is clear that the data augmentation techniques improved the Macro F-Score. Indeed, the ResNet-50 model without data augmentation only obtained a higher score on the first epoch: 10.87% versus 8.33% (the latter is with data augmentation). In Figure 4, it is possible to visualize the evolution of training accuracy per batch and testing average Macro F-Score for the ResNet-50 models. The models' training seems to converge on the final epochs, in terms of average throughout the epoch (see also Table 1). As for the testing phase, the slight decrease in the last epoch for both ResNet-50 models indicated that the training should be stopped. Even though one model had implemented data augmentation techniques, it seems that both

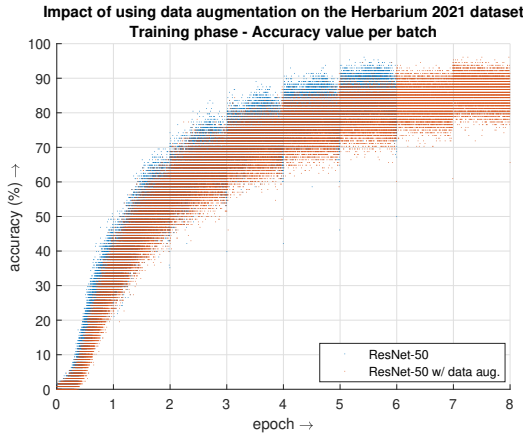


(a)

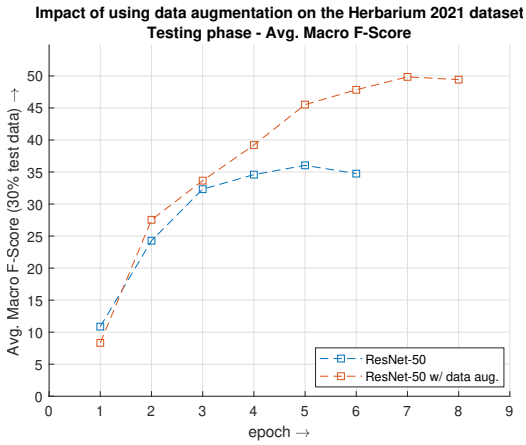


(b)

Figure 3: Comparison of ResNet models on the Herbarium 2021 dataset: (a) accuracy on training; (b) average Macro F-Score



(a)



(b)

Figure 4: Impact of using data augmentation on the Herbarium 2021 dataset: (a) accuracy on training; (b) average Macro F-Score

models overfitted to the training data. Given that the data augmentation techniques are implemented online, probably the probabilities for the geometric transformations described in Section 5 should have been higher to prevent this overfitting.

The most important result is that the best ResNet-50 model with data augmentation reached an approximately 15% higher Macro F-Score on the testing phase compared to not using data augmentation. Note that the average Macro F-Score first computes the Macro F-Score value for each class and then computes the average of these values giving equal importance to each class. So, the increase of 15% on the average Macro F-Score seems to indicate that the ResNet-50 with data augmentation improved the detection of more classes. This increase shows the importance of data augmentation especially for classes with few examples. However, this observation cannot be confirmed without having the Macro F-Scores of each class and also the precision and recall values (Kaggle [7] does not provide these values for the testing phase of the Herbarium 2021 [3] competition).

7. Conclusions and Future Work

The work presented in this paper focused on the image classification problem formulated in the Herbarium 2021 [3] competition. The goal was to classify different plant specimens training a model with the dataset provided for the competition. The models used in this work were ResNet-18, ResNet-34, and ResNet-50. After firstly comparing them without data augmentation, ResNet-50 obtained a higher average Macro F-Score than ResNet-18 (second best) on the test set: 36.04% vs 32.90%, respectively. Then, data augmentation techniques based on geometric transformations – horizontal and vertical flips, and

rotation – were used to train the ResNet-50 model. The test results showed that data augmentation was crucial to improving the Macro F-Score: 49.85% over 36.04% (the latter when not using data augmentation). Therefore, the data augmentation techniques used in this work improved the classification performance of the ResNet-50 model while obtaining a nearly 50% average Macro F-Score. At the date of May 23, it was obtained a predicted 11th place. However, note that the test results presented here only use approximately 30% of the test data with the other 70% being used for the competition's final results. All the models trained in this work are available on a Kaggle [7] dataset[1], and also the code is publicly available[2] on Kaggle [7].

As future work, it could be used a higher number of layers of ResNet models (e.g., ResNet-152) or a model based on the EfficientNet [17] deep learning network (current ranking no. 1 in ImageNet [2] dataset is based on the same architecture). Another approach could be considering the supercategory information present in the training dataset to implement a hierarchical model. In terms of improving this work's approach, the data augmentation (probabilities and maximum value of rotation) and the optimization (learning rate and batch size) parameters could be optimized, and other scores could be computed for the training phase (average Macro F-Score) to improve the assessment of the model's training.

References

[1] P. Barré, B. C. Stöver, K. F. Müller, and V. Steinhage. LeafNet: a computer vision system for automatic plant species identification. *Ecological Informatics*, 40:50–56, 2017. https://doi.org/10.1016/j.ecoinf.2017.05.005.

[2] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, June 2009. https://doi.org/10.1109/CVPR.2009.5206848.

[3] FGVC8 and CVPR 2021. Herbarium 2021 - Half-Earth Challenge - FGVC8 - Kaggle, 2021. https://www.kaggle.com/c/herbarium-2021-fgvc8/overview.

[4] M. Freiberg, M. Winter, A. Gentile, A. Zizka, A. N. Muellner-Riehl, A. Weigelt, and C. Wirth. LCVP, the Leipzig catalogue of vascular plants, a new taxonomic reference list for all known vascular plants. *Scientific Data*, 7:416, 2020. https://doi.org/10.1038/s41597-020-00702-z.

[5] K. He and J. Sun. Convolutional neural networks at constrained time cost. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5353–5360, June 2015. https://doi.org/10.1109/CVPR.2015.7299173.

[6] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, June 2016. https://doi.org/10.1109/CVPR.2016.90.

[7] Kaggle. Kaggle: Your Home for Data Science. https://www.kaggle.com/. Accessed on 2021-05-18.

[8] A. Krizhevsky. One weird trick for parallelizing convolutional neural networks. *CoRR*, abs/1404.5997, 2014. http://arxiv.org/abs/1404.5997.

[9] N. Kumar, P. N. Belhumeur, A. Biswas, D. W. Jacobs, W. J. Kress, I. C. Lopez, and J. V. B. Soares. Leafsnap: a computer vision system for automatic plant species identification. In A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, editors, *Computer Vision – ECCV 2012*, pages 502–516, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-33709-3_36.

[10] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: common objects in context. In D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 740–755, Cham, 2014. Springer International Publishing. https://doi.org/10.1007/978-3-319-10602-1_48.

[11] PyTorch. PyTorch. https://pytorch.org/. Accessed on 2021-05-18.

[12] W. Rawat and Z. Wang. Deep Convolutional Neural Networks for Image Classification: A Comprehensive Review. *Neural Computation*, 29(9):2352–2449, 09 2017. https://doi.org/10.1162/neco_a_00990.

[13] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations (ICLR) 2015*, San Diego, CA, USA, May 2015. http://arxiv.org/abs/1409.1556.

[14] R. K. Srivastava, K. Greff, and J. Schmidhuber. Highway networks. *CoRR*, abs/1505.00387, 2015. http://arxiv.org/abs/1505.00387.

[15] Y. Sun, Y. Liu, G. Wang, and H. Zhang. Deep learning for plant identification in natural environment. *Computational Intelligence and Neuroscience*, 2017:7361042, 2017. https://doi.org/10.1155/2017/7361042.

[16] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, 2015. https://doi.org/10.1109/CVPR.2015.7298594.

[17] M. Tan and Q. Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 6105–6114. PMLR, Jun. 2019. http://proceedings.mlr.press/v97/tan19a.html.

[1] https://www.kaggle.com/ricardobarbosasousa/herbarium-2021-rbs

[2] https://www.kaggle.com/ricardobarbosasousa/herbarium-2021-rbs-resnet