PlantCLEF2023: A Bigger Training Dataset Contributes More than Advanced Pretraining Methods for Plant Identification

Mingle Xu[1,2], Sook Yoon[3,*], Chenmou Wu[4], Jeonghyun Baek[5] and Dong Sun Park[1,2,*]

[1]Department of Electronics Engineering, Jeonbuk National University, Jeonbuk 54896, South Korea
[2]Core Research Institute of Intelligent Robots, Jeonbuk National University, Jeonbuk 54896, South Korea
[3]Department of Computer Engineering, Mokpo National University, Jeonnam 58554, South Korea
[4]Division of Computer Science and Engineering, Jeonbuk National University, Jeonju, Republic of Korea
[5]Rural Development Administration, Jeonbuk 54875, South Korea

Abstract

Plant identification has received a significant improvement in recent years with the development of deep learning. However, large-scale plant identifications still suffer such as the PlantCLEF2022 challenge, and this paper aims to address this issue in the PlantCLEF2023 challenge. On one hand, we extend an effective strategy, transfer learning, resorting to advanced pretrained methods with multi-modal and self-supervised learning. On the other hand, training dataset size and quality are probed and their impacts are compared to one of the advanced pretrained methods. The experimental results suggest that advanced pretrained methods, bigger datasets, and noisy datasets are beneficial to plant identification. Especially, the performance gain from the first one is inferior to the counterparts of the latter two, with the dataset in the PlantCLEF2023 challenge. For example, employing a noisy dataset together receives 0.0236 gains in performance whereas multimodal and self-supervised pretraining contributes 0.00792 gains. We hope that our results highlight the importance of collecting better datasets for plant identification. Our codes are public at https://github.com/xml94/PlantCLEF2022.

Keywords

plant identification, multi-observation, multimodal pretraining, vision transformer, deep learning

1. Introduction

Plant identification is an essential task in plant science, such as for maintaining biodiversity and finding new plant species, which traditionally requires the involvement of human experts. Because of the limitation of experts, plant identification is often not convenient and expensive. Besides, it is also difficult to integrate the identification with other tasks. Resorting to RGB images and computer vision methods, plant identification has achieved a significant improvement in recent years. A commonly embraced paradigm is taking plant identification from the computer vision community, with a basic inspiration that plants images resemble the images in general computer vision tasks such as ImageNet-1k [1].

CLEF 2023: Conference and Labs of the Evaluation Forum, September 18–21, 2023, Thessaloniki, Greece
*Corresponding author.
xml@jbnu.ac.kr (M. Xu); syoon@mokpo.ac.kr (S. Yoon); chenmou@jbnu.ac.kr (C. Wu); butterfy@korea.kr (J. Baek); dspark@jbnu.ac.kr (D. S. Park)

Holding the paradigm, the methods and techniques in general computer vision tasks are overwhelmingly borrowed to plant identification. Transfer learning is proven as one of the most successful strategies, which directly adopts the models trained in the datasets from general computer vision tasks. This paper extends this idea one more step by using advanced pretrained methods. To be specific, a ViT-large model pretrained in multi-modal and self-supervised is fine-tuned in the PlantCLEF2023 dataset. Multi-modal refers to using inputs in different modality [2, 3, 4], such as text and image [5], by which more semantics are learned, whereas self-supervised learning denotes a set of training strategies without using labels. One of the most successful self-supervised learning methods is predicting the masked contents from the original inputs [6, 7, 3]. Finally, *we examine the effectiveness of the pretraining with multi-modal and masked image modeling in the context of plant identification.*

Simultaneously, we aim to point out the heterogeneity between general computer vision and plant identification and argue that plant identification has its specificity [8]. For example, a plant may be very similar to another plant. However, this heterogeneity should receive more attention ranging from dataset collecting and algorithm designing, although some existing papers cast plant identification into a fine-grained classification. In this paper, *we aim to probe the impact of size and quality of training datasets to identify plant species, and compare the impacts to the one with advanced pretraining methods.* We emphasize that the comparison is our ultimate objective, rather than probing training dataset alone, in that scaling laws, enlarging capacities of models and sizes of training datasets, already suggests its power in natural language processing [9, 10, 11] and computer vision [4, 12].

Through experiments in PlantCLEF2023, we found that advanced pretraining methods slightly contribute to plant identification, and a bigger dataset and noisy datasets may be more beneficial. As shown in Table 4, employing noisy datasets together receives 0.0236 gains in performance whereas multimodal and self-supervised pretraining contributes 0.00792 improvements. Based on this observation, we think that collecting a better dataset and analyzing the challenges within a collected dataset should receive more attention.

2. Material and Evaluation Metric

2.1. Datasets

Trusted and Web datasets. The PlantCLEF2023 challenge [13, 14] adopts the same dataset as PlantCLEF2022 and the dataset officially consists of three parts, Trusted, Web, and Test. We summarize some details and our analysis and understanding in Table 1. The Trusted and Web datasets have labels and can be utilized to train proposed models but their reliabilities are different. The Trusted dataset is annotated by experts whereas the Web one is collected from the internet with some other post-processes to improve the quality.

We aim to analyze the Trusted and Web datasets and found distinct characteristics. First, the two datasets are both in **heavy class-imbalance** where some classes have much more number of images than other classes as shown in Figure 2 and Table 2 and 3, which may result in lower performance for the classes with less number of images [15]. Second, the datasets have **huge intra-class variation** [16, 15] in two factors: a plant species has multiple organs, such as flowers and leaves; and images from an organ may have heterogeneity, such as multiscale and

Table 1
Details, analysis, and understandings towards the datasets in PlantCLEF2023 challenge.

Dataset	Classes	Images	Analysis and understandings
Trusted	80,000	2,885,052	It is annotated by human experts and can be regarded as reliable. This dataset is in heavy class imbalance, as shown in the first row of Figure 2 and Table 2. It owns a huge intra-class variation in terms of two aspects: one plant species has multiple organs such as flowers and leaves and, as shown in Figure 1, one organ has heterogeneity such as leaves in multiscale and flowers in the different growth stages.
Web	57,314	1,071,627	It is collected from the web and thus may be with noisy and even wrong labels. Its class space is a proper set of the Trusted dataset with less number of classes. This dataset is also in heavy class imbalance, as shown in the second row of Figure 2 and Table 3. The Web dataset is in a different distribution from the Trusted dataset where the number of images for one class is different as shown in Figure 3. We randomly looked at several classes in the Web dataset and found that most of the images seem in the correct classes (we emphasize we are not domain experts and thus our judgments may not be correct).
Trusted + Web	80,000	3,995,568	This dataset is the combination of the Trusted and Web datasets. The total number of images is not the summation of the Trusted and Web datasets and we guess that some images are shared in the two datasets. This combined dataset has more images and may mitigate the class-imbalance impact than their individual counterparts.
Test	7,339	55,306	The classification should be done at the observation level, rather than the image level, where one observation refers to one actual plant with several pictures taken from different viewpoints. The task requires a strategy to combine the predictions of multiple images for one observation. Besides, the testing dataset only shares part of the plant identity in the trusted training dataset.
Shared information			Images have similar resolutions, around 450×600. Most of the images are collected from real scenarios, rather than laboratories. Class imbalance, huge intra-class variations, and few-shot are the distinct characteristics of the datasets.

growth stage as shown in Figure 1. Third, the datasets have relatively less number of images averaged by the classes and can be cast as a **few-shot classification**. For example, the Trusted and Web datasets have approximately 36 and 18 images on average for each class.

Through our observation, the difference between the Trusted and Web datasets may be the distribution of the number of images in one species, as shown in Figure 3. Several random observations suggest that most of the images in the Web dataset are correctly annotated. Therefore, the two datasets are combined as a potential training dataset with much more images and less impact of class imbalance [15].

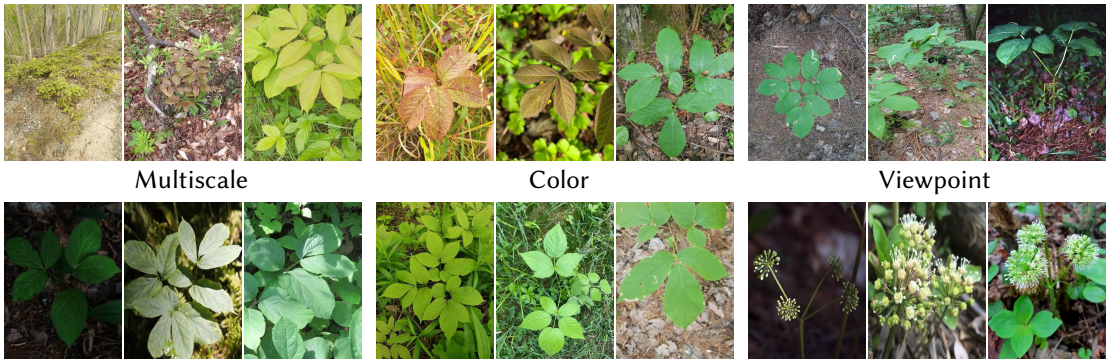

Figure 1: Images of Aralia nudicaulis L. species from PlantCLEF2022 Trusted dataset. The pictures show huge intra-class variation [17, 16, 15] and every triplet suggests an intra-class variation.

Table 2
Statistics of the Trusted dataset where the numbers of classes and images are counted. For example, there are 28,681 classes with or more than 36 images, covering 2,381,264 images in total. 36 is the average number of images for the classes.

Statistic term	Classes	Images
all	80,000	2,885,052
>=7	54,478	2,807,969
>=36	28,681	2,381,264
>50	24,284	2,194,912
>=100	9,122	920,774

Table 3
Statistics of the Web dataset where the numbers of classes and images are counted. For example, there are 23,317 classes with or more than 18 images, covering 869,518 images in total. 18 is the average number of images for the classes.

Statistic term	Classes	Images
all	57,314	1,071,617
>=18	23,317	869,518
>50	4,335	251,206

Test dataset and observation The Test dataset has no public labels and is formally adopted to evaluate different methods. It has 55,306 images for 26,868 individual plants, termed observations where a plant may have multiple images. According to the official report [18], the images cover 7,339 plant species, less than the total numbers in the Trusted and Web datasets, but the exact information is not accessible. The observation-level classification requires the ability to make decisions considering multiple predictions from a model, which may be beneficial to have higher performance to recognize plant species.

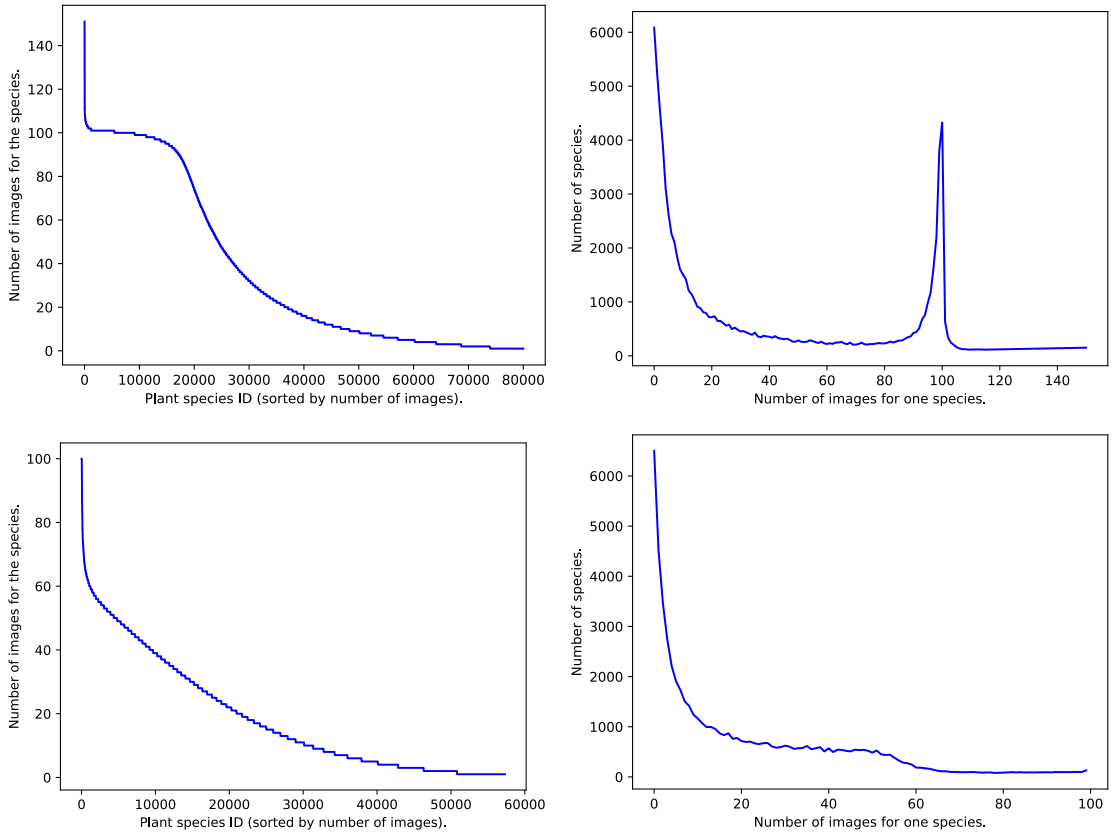

Figure 2: Left up (1): number of images of every class in the Trusted dataset, sorted by their numbers. More than 40,000 classes have less than 20 images. Right up (2): relation between number of images and number of species. Two peaks are around 0 and 100 images in species. Left bottom (3): number of images of every class in the Web dataset, sorted by their numbers. Right bottom (4): relation between number of images and number of species.

2.2. Evaluation metric

Macro averaged mean reciprocal rank (MA-MRR) is utilized to evaluate different submissions for the PlantCLEF2022 challenge. The challenge requests a submission with a rank based on the score with a given length for each testing observation, and the rank is thirty for the challenge. Assume there are N classes in the testing dataset and class n has O_n observations. Mathematically, MA-MRR can be formalized as[1,2]:

$$MA - MRR = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{O_n}\sum_{i=1}^{O_n}\frac{1}{r_i}, \tag{1}$$

where r_i refers to the rank position of the first relevant ground-truth label for the i-th

[1]https://en.wikipedia.org/wiki/Mean_reciprocal_rank
[2]https://androidkt.com/micro-macro-averages-for-imbalance-multiclass-classification/

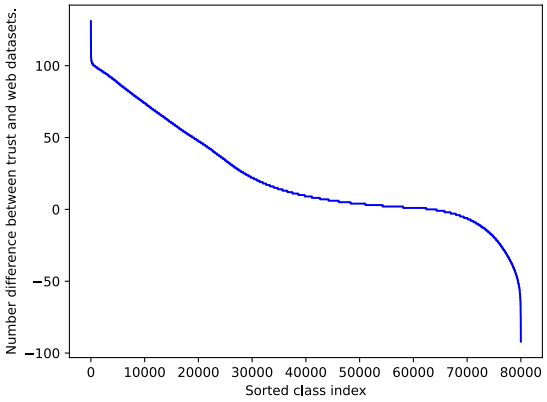

Figure 3: Number difference between the same class in the Trusted and Web dataset, sorted by the difference. It suggests a distinct distribution gap between the Trusted and Web datasets that most of the classes have different numbers of images.

observation from one class. Conceptually, if all of the observations are classified correctly in the first, then MA-MRR is one where the accuracy is hundred percent. In contrast, if MA-MRR is smaller then the model is worse. Intuitively, the observation level enables us to observe different parts of plants, while MA-MRR allows more than one chance to recognize plant species for a specific image, 30 chances in the PlantCLEF2023 challenge. In general, MA-MRR is slightly higher than the actual accuracy, a widely used metric to evaluate image classification.

3. Methods

In this section, a combination method for observation-level image classification is first described. Second, our two main strategies, relevant to pretraining and fine-tuning, for the PlantCLEFF2022 challenge are introduced. Our results are summarized in Table 4.

3.1. Strategy towards observation

Let r^j_{score} denote the testing probability score rank of the j-th image of one observation. Similarly, r^j_{class} is the testing class rank with the same length. Besides, r_{class_i} and r_{score_i} are the i-th class-score pair and mean the class and the corresponding probability score, respectively. The rank requires $r_{score_p} > r_{score_q}$ if $p \in \mathbb{R}^+ < q \in \mathbb{R}^+$. A final desired output $\{r_{score}, r_{class}\}$, pair of class and corresponding score can be formulated as

$$\{r_{class}, r_{score}\} = \{(\cup_{j=1}^{n}(\{r^j_{class}, r^j_{score}\})), \tag{2}$$

where $\{$ means an integration function and the observation has n corresponding images. $\cup$ suggests the union of the n images.

Following our observation in the last year [17], the multi-sorted method contributes to a relatively higher performance, where multi means that the final rank pair is from multiple images, rather than a single image, and sorted means that the predicted species with higher

Table 4
Our official results, the best showing in the boldface. We fine-tune the models in the PlantCLEF2023 dataset. Except for run 9 and 10, fine-tuned by 10 and 50 epochs respectively, all models are fine-tuned by 100 epochs, following MAE. Same with our last year's experiment [17], we executed the experiments with the code of MAE, which means the same training strategies. The Trusted and Web datasets are short as T and W. The number after num suggests the statistics term. For example, T@num36 denotes the images from the classes with exact or more than 36 images. MAE-IN1k suggests that an MAE model is pretrained in ImageNet-1k (IN1k). EVA-IN21K and EVA-IN21k-IN21k denote that an EVA model is pretrained in ImageNet-21k (IN21k) and fine-tuned in IN21k once more, respectively. EVA-IN21k-IN21k-PlantCLEF@T+W is fine-tuned model of EVA-IN21k-IN21k in PlantCLEF2023 Trusted and Web datasets. Our official run with ID 7 is not here because we wrongly uploaded the same result.

Run ID	Pretraining	Fine-tuning			MA-MRR
		Dataset	cls	img	
[17]	MAE-IN1k	T	80,000	2,885,052	0.64079
3	EVA-IN21k	T	80,000	2,885,052	0.64871
1	EVA-IN21k-IN21k	T@num100	9,122	920,774	0.33239
4		T@num50	24,284	2,194,912	0.54846
2		T@num36	28,681	2,381,264	0.57514
6		T@num7	54,478	2,807,969	0.64201
5		T	80,000	2,885,052	0.65035
8		T+W	80,000	3,995,568	**0.67395**
9	EVA-IN21k-IN21k-PlantCLEF@T+W	T@epoch10	80,000	2,885,052	0.66330
10		T@epoch50	80,000	2,885,052	0.65695

probabilities after sorting all predictions from all images aligning to the same observation are adopted to evaluate. With the multi-sorted strategy, the scores of all images from the same observation are sorted: $sort(\cup_{j=1}^{n} \cup_{i=1}^{l} r_{score_i}^{j})$ where l is the length of the ranges (30 in the PlantCLEF2023 challenge); and then, after removing duplicates of same classes, the first required length of class-score pair are taken as the final ranking pair.

3.2. Advanced pretraining models

As mentioned before, the PlantCLEF2023 challenge can be recognized as a few-shot image classification. Even if the Trusted and Web datasets are combined, each class has only about 50 images on average, much less than the counterpart in the ImageNet [1] dataset. To address this issue, employing a pretrained model is one of the most effective and efficient methods across many different tasks [16, 15], especially in the eras of large language models such as ChatGPT [7] and foundation models [2]. Therefore, we aim to probe the impact of state-of-the-art pretrained models.

MAE [6] is self-supervised trained in ImageNet-1k (IN1k), label information not being used, where the input is randomly blocked and the objective is reconstructing the blocked pixels, as shown in Figure 4. We employed MAE with ViT-large [19] model in the last year and obtained a first place for the PlantCLEF2022 challenge [17]. Another widely adopted pretraining strategy involves multi-modal, such as CLIP [5], in which the learning loss is to maximize the relationships between paired text and images and minimize the counterparts of the unpaired

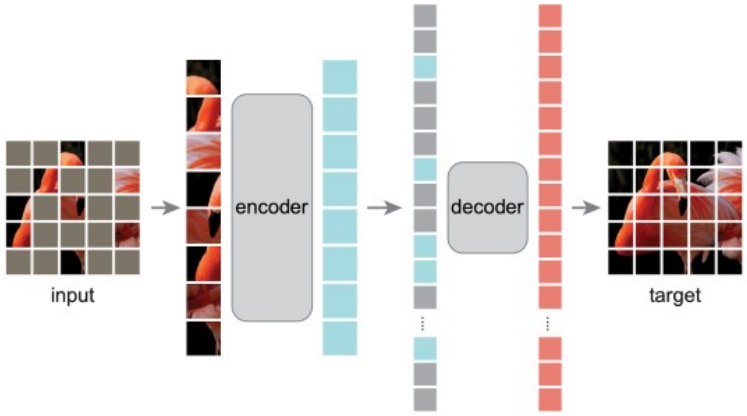

Figure 4: Architectures of MAE [6],can be adopted as pre-training method without using annotations.

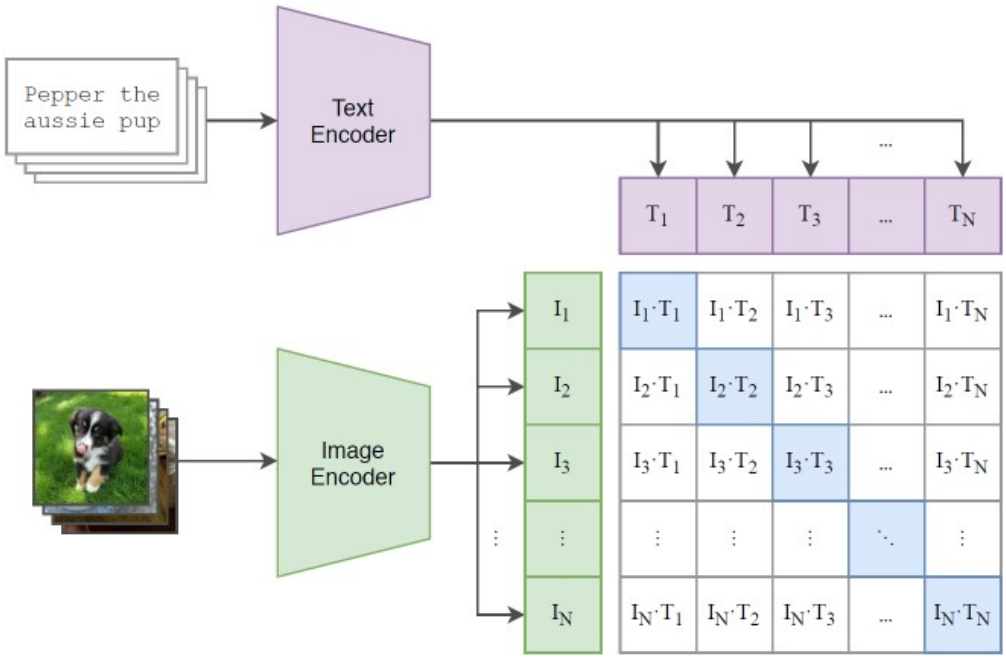

Figure 5: Training architecture of CLIP [5], trained with paired images and texts.

ones. As shown in Figure 5, this pretraining employs multimodal information simultaneously and thus may learn semantics better than using vision information alone [3].

MVP [3] integrates the strategies in MAE and CLIP to improve the pretraining performance. As shown in Figure 6, MVP freezes the image encoder from CLIP and trains the parameters from the vision part. To be specific, the loss function is to minimize the distance between the feature from the frozen CLIP image encoder and the feature from the vision model with randomly

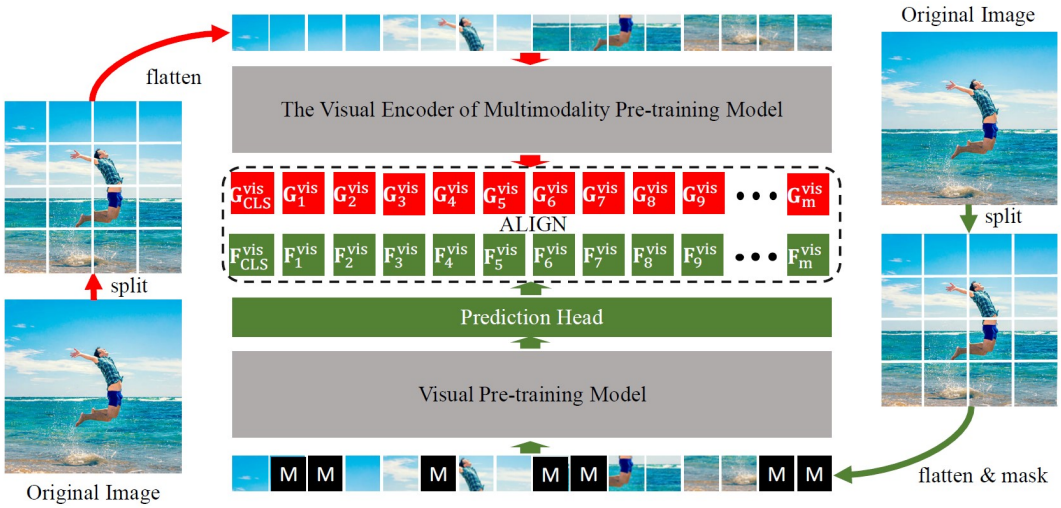

Figure 6: Architectures of MVP [3] which can be deemed as a combination of MAE and CLIP.

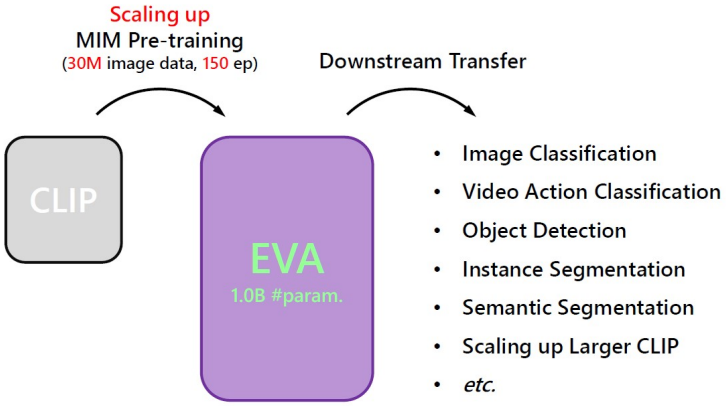

Figure 7: Highlights of EVA [4], a scaled version of MVP.

blocked images as inputs. In this way, a vision model in MVP can learn more semantics from the frozen CLIP image encoder that is pretrained in 400 million image-text pairs. Furthermore, as highlighted in Figure 7, EVA [4] scales MVP by using bigger models with more parameters and more datasets, by which better performance is achieved across different downstream tasks.

Inspired by the success of EVA, we extend it from general computer vision tasks to plant species identification on a large scale. Because of the limitation of GPUs, we only consider the scaling of EVA with more datasets and the model architecture is still ViT-large [19]. Specifically, EVA-L pretrained in ImageNet-21k, rather than ImageNet-1k, is leveraged, which is denoted EVA-IN21k. We also utilized its finetuned version, finetuned once more in ImageNet-21k in a supervised manner, mainly because it has better performance in ImageNet-1k and empirically, a model with higher performance in ImageNet-1k may have a better performance in the target

dataset with transfer learning strategy [20]. This pretrained model is denoted as EVA-IN21k-IN21k.

As shown in Table 4, EVA-IN21k pretrained model outperforms MAE-IN1k slightly and EVA-IN21k-IN21k has also a tiny improvement. Our results suggest that better pretrained models and better performance in ImageNet-1k may have superiority in plant identification, same observations in [21].

3.3. Bigger training dataset

Considering that the Trusted dataset has 80,000 classes but the Test dataset only has 7,339 classes, training models in the whole Trusted dataset seems a waste of computing resources. A desired strategy is to find the class in the Test dataset and then just use the corresponding images in the Trusted dataset. However, the desired classes are not given. Instead, we assume that the frequencies of plant species in the Trusted and Test datasets are similar. Therefore, we just choose the classes from the Trusted dataset based on the number of samples. For example, those classes having 100 images or more are selected, denoted as T@num100, as shown in Table 4.

To our surprise, the test performance monotonically shrinks when the number of images reduces. Especially, training the model in the whole Trusted dataset leads to the best performance. We emphasize that this phenomenon may result from two factors: the number of images and the number of classes. When the number of images increases, some classes in the Test dataset may be included. However, these classes just have a few images, such as less than 7 if the training dataset replaces T@num7 with T.

We further directly utilized the Web dataset along with the Trusted because of our previous observation that the main difference is the frequency as shown in Figure 3. The corresponding performance is further improved by 3.6% than using the Trusted alone. As the Web dataset is assumed to have noise, the trained model with the Trusted and Web datasets undergoes fine-tuning once more, 10 and 50 epochs as run ID 9 and 10 respectively. We are shocked that the performance degrades with the fine-tuning process. We conjecture that *most of the images in the Web dataset are labeled correctly*.

We emphasize that the experiment to change the training datasets is not trivial in that it resembles the process in real applications. The test scenario is not always fully understood and the training dataset is getting bigger gradually when projects continue. Sometimes, a training dataset may be along with noisy labels, and the strategies to use them should be further considered.

3.4. Discussion

Transfer learning and its advanced version are slightly beneficial for plant identification. Transfer learning has contributed to many tasks and our work proves that advanced transfer learning, such as foundation models, has an extra contribution. Therefore, utilizing a better pretrained model is one of the potential strategies. But we emphasize that it may also hinder performance such as when the labeled data in the target task is scarce [22, 23]. Simultaneously, foundation models may also suffer when the distributions of the source dataset

to pretrain a model and the target dataset is diverse [24]. As far as we know, the images of PlantCLEF2023 are far different from the datasets, ImageNet-1k and ImageNet-21k, to pretrain MAE and EVA models. New methods to leverage the pretrained model are the potential to make further improvements.

Collecting a better dataset is still promising for plant identification. In the PlantCLEF2023 challenge, the Trusted and Web datasets are both in heavy class imbalance and some classes of the datasets have only one or two images. Our experimental results suggest that more data make better performance, even if some noisy data are leveraged directly. Compared to the gains from transfer learning methods and bigger datasets, we believe that collecting a bigger dataset could be more effective.

Limitations and social impacts. Our experiments have used enormous computing resources although we obtained a decent record for the challenge. To be honest, we used 16 RTX 3090 GPUs for almost three months. It is actually not fair to other teams without enough GPUs. On the other hand, we hope our analysis and understandings are useful for the community, such as dataset analysis and the comparison between transfer learning and bigger dataset.

4. Conclusion

We first analyzed the given datasets in the PlantCLEF2023 challenge and found that class imbalance and huge intra-class variation are the main characteristics. Simultaneously, the challenge can be taken from a few-shot perspective. Furthermore, the Trusted and Web datasets are compared, with an observation that their image distribution is much different and the noise in the Web dataset may be exaggerated. Two strategies are executed for the challenge, different pretrained models and different sizes of the training dataset. Both strategies contribute and, even with noisy data, a bigger training dataset is more beneficial to the challenge. We hope that our results highlight the importance of collecting bigger datasets for plant identification. Simultaneously, we emphasize the heterogeneity between plant identification and general computer vision.

Acknowledgments

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (No. 2019R1A6A1A09031717). This work was supported by the Korea Institute of Planning and Evaluation for Technology in Food, Agriculture, and Forestry (IPET) and Korea Smart Farm R&D Foundation (KosFarm) through the Smart Farm Innovation Technology Development Program, funded by Ministry of Agriculture, Food and Rural Affairs (MAFRA) and Ministry of Science and ICT (MSIT), Rural Development Administration (RDA) (421027-04). This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (NRF-2021R1A2C1012174).

References

[1] D. Jia, et al., Imagenet: A large-scale hierarchical image database, in: 2009 IEEE conference on computer vision and pattern recognition, Ieee, 2009, pp. 248–255.

[2] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, et al., On the opportunities and risks of foundation models, arXiv preprint arXiv:2108.07258 (2021).

[3] L. Wei, L. Xie, W. Zhou, H. Li, Q. Tian, Mvp: Multimodality-guided visual pre-training, in: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXX, Springer, 2022, pp. 337–353.

[4] Y. Fang, W. Wang, B. Xie, Q. Sun, L. Wu, X. Wang, T. Huang, X. Wang, Y. Cao, Eva: Exploring the limits of masked visual representation learning at scale, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 19358–19369.

[5] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International conference on machine learning, PMLR, 2021, pp. 8748–8763.

[6] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, R. Girshick, Masked autoencoders are scalable vision learners, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 16000–16009.

[7] L. Yiheng, et al., Summary of chatgptgpt-4 research and perspective towards the future of large language models, arXiv preprint arXiv:2304.01852 (2023).

[8] M. Xu, Enhanced Plant Disease Recognition with Limited Training Dataset Using Image Translation and Two-Step Transfer Learning, Ph.D. thesis, Jeonbuk National University, 2023.

[9] J. Kaplan, et al., Scaling laws for neural language models, arXiv preprint arXiv:2001.08361 (2020).

[10] Y. Tay, M. Dehghani, S. Abnar, H. W. Chung, W. Fedus, J. Rao, S. Narang, V. Q. Tran, D. Yogatama, D. Metzler, Scaling laws vs model architectures: How does inductive bias influence scaling? (2022).

[11] J. Wei, et al., Emergent abilities of large language models, arXiv preprint arXiv:2206.07682 (2022).

[12] M. Dehghani, et al., Scaling vision transformers to 22 billion parameters, arXiv preprint arXiv:2302.05442 (2023).

[13] H. Goëau, P. Bonnet, A. Joly, Overview of plantclef 2023: Image-based plant identification at global scale, in: CLEF 2023-Conference and Labs of the Evaluation Forum, 2023.

[14] e. A. Joly, 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, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2023.

[15] M. Xu, H. Kim, J. Yang, A. Fuentes, Y. Meng, S. Yoon, T. Kim, D. S. Park, Embracing limited and imperfect data: A review on plant stress recognition using deep learning, arXiv preprint arXiv:2305.11533 (2023).

[16] M. Xu, S. Yoon, A. Fuentes, D. S. Park, A comprehensive survey of image augmentation

techniques for deep learning, Pattern Recognition (2023) 109347.
[17] M. Xu, S. Yoon, Y. Jeong, J. Lee, D. S. Park, Transfer learning with self-supervised vision transformer for large-scale plant identification, in: International conference of the cross-language evaluation forum for European languages (Springer;), 2022, pp. 2253–2261.
[18] H. Goëau, P. Bonnet, A. Joly, Overview of plantclef 2022: Image-based plant identification at global scale, in: CLEF 2022-Conference and Labs of the Evaluation Forum, volume 3180, 2022, pp. 1916–1928.
[19] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, in: International Conference on Learning Representations, 2020.
[20] S. Kornblith, J. Shlens, Q. V. Le, Do better imagenet models transfer better?, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 2661–2671.
[21] M. Xu, S. Yoon, Y. Jeong, D. S. Park, Transfer learning for versatile plant disease recognition with limited data, Frontiers in Plant Science 13 (2022) 4506.
[22] Z. Wang, Z. Dai, B. Póczos, J. Carbonell, Characterizing and avoiding negative transfer, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 11293–11302.
[23] W. Zhang, L. Deng, L. Zhang, D. Wu, A survey on negative transfer, IEEE/CAA Journal of Automatica Sinica (2022).
[24] Y. Lee, A. S. Chen, F. Tajwar, A. Kumar, H. Yao, P. Liang, C. Finn, Surgical fine-tuning improves adaptation to distribution shifts, in: The Eleventh International Conference on Learning Representations, 2023. URL: https://openreview.net/forum?id=APuPRxjHvZ.