Transfer Learning with Self-Supervised Vision Transformer for Large-Scale Plant Identification

Mingle Xu[1,2], Sook Yoon[3], Yongchae Jeong[1], Jaesu Lee[4] 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]Rural Development Administration, Jeonbuk 54875, South Korea

Abstract

This paper is a working note for the PlantCLEF2022 challenge aiming to identify plants with a large-scale dataset, several millions of images and 80,000 classes. Although there are many images, each class only includes 36 images around on average and thus it can be regarded as a few-shot image classification. To address this issue, transfer learning is validated to be useful in many scenarios and a popular strategy is employing a convolution neural network (CNN) pretrained in a supervised manner. But inspired by the literature on computer vision, we instead leverage a self-supervised vision transformer (ViT) and secure the first place with MA-MRR 0.62692, 0.019 higher than the second place, and 0.116 than the third. Furthermore, we achieve 0.64079 if training the model twenty epochs longer. Compared to the popular strategy with CNN, self-supervised ViT has two advantages. First, ViT does not embrace any inductive bias, such as translating invariance embraced in CNN, and thus owns a more powerful model capacity. Second, self-supervised pretraining obtains a task-agnostic feature extractor that may be better for the downstream task. To be more specific, a recently proposed self-supervised ViT model pretrained in ImageNet, masked autoencoder (MAE), is finetuned in PlantCLEF2022 dataset and then tested to report the evaluation. Except for the challenge, we discuss its possible impacts, such as taking the dataset to pretrain a model for plant-related tasks. Especially, our preliminary results suggest that the pretrained model in PlantCLEF2022 essentially contributes to image-based plant disease recognition on several public datasets. Via our analysis and experimental results, we believe that our work encourages the community to utilize the self-supervised ViT model, the PlantCLEF2022 dataset, and our pretrained model in the dataset. Our codes and trained model are public at https://github.com/xml94/PlantCLEF2022.

Keywords

plant identification, image classification, transfer learning, computer vision, self-supervised, vision transformer

1. Introduction

Recognizing different species is one of requirements to maintain biodiversity, however, it requires human experts to spend much time with a high cost [1]. Simultaneously, deep learning has been showing its potential to automatically classify images. The PlantCLEF2022 challenge [1, 2] is held towards this issue to identify plant species given their images and owns a large-scale and

CLEF 2022: Conference and Labs of the Evaluation Forum, September 5–8, 2022, Bologna, Italy
xml@jbnu.ac.kr (M. Xu); syoon@mokpo.ac.kr (S. Yoon); ycjeong@jbnu.ac.kr (Y. Jeong); butiman@korea.kr (J. Lee); dspark@jbnu.ac.kr (D. S. Park)
0000-0003-2662-6864 (M. Xu)

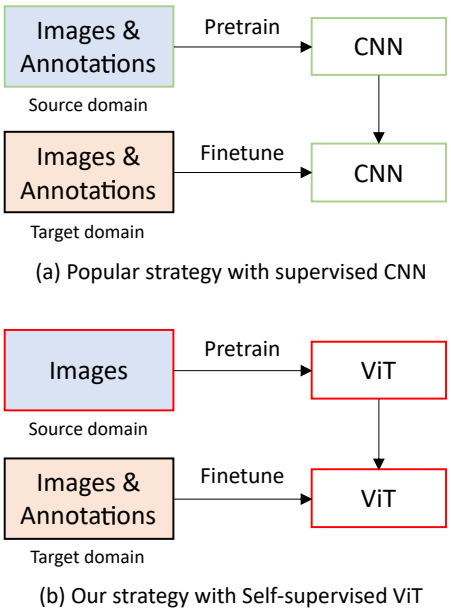

Figure 1: Comparison between the popular strategy of transfer learning, supervised CNN, and our strategy, self-supervised ViT. Our method differs from the popular one in two cases, ViT as feature extractor instead of CNN, and self-supervised pretraining in the source domain without annotations.

complex dataset taken from real field with 80,000 classes and millions of images. Although there are many images in total, each class averagely owns about 36 images and thus the challenge can be regarded as a *few-shot image classification* [3]. To address the image classification, transfer learning is verified to be one effective and efficient method, by which useful knowledge can be adapted from a much bigger *source dataset* to a small *target dataset* [4].

A popular strategy to perform transfer learning is pretraining a convolution neural network (CNN) in a supervised manner in a source dataset, ImageNet [5] and then the network is finetuned in a target dataset [6, 7, 8, 9]. However, there are two weak points in this strategy. First, as pretrained in a supervised way, the learned features of the model is task related [10], which could be not good when the target dataset is not similar to the source dataset, especially for a fine-grained target dataset [11]. In fact, the ImageNet is far from the fine-grained PlantCLEF2022 dataset as discussed in the next section. Second, CNNs embrace an inductive bias, such as translation invariance [12] that the corresponding class or identity is same if one image is translated a little. Due to the inductive bias, the capacity of CNN-based models is limited and the capacity could be improved if the inductive bias is removed.

Fortunately, the two weak points are eased very recently by self-supervised learning (SSL) [13] and vision transformer (ViT) [12], respectively. On one hand, SSL employs an predefined task, such as reconstruction [14], context prediction [15], and predicting rotations [16], instead of supervised signals. On the other hand, ViT discards the local connection and shared filters of CNNs and holds global attentions. Considering the character of the PlantCLEF2022 dataset and the current success on computer vision, we employ a self-supervised pretrained vision transformer to address the PlantCLEF2022 challenge and the strategy has been proved to

significantly contribute to plant disease recognition [17]. The comparison of the popular strategy and our strategy is illustrated in Figure 1. To be more specific, a ViT-based masked autoencoder (MAE) pretrained in a self-supervised manner in ImageNet is finetuned and tested in the PlantCLEF2022 dataset. In spite of the simpleness, we secure the first place of the official challenge with MA-MRR 0.62692, 0.019 higher than the second place and 0.116 than the third. Moreover we found that we can further get a better performance, 0.64079, if we fine-tune the model longer.

Except for the challenge, we also discuss its possible impacts. We first recognize two distinct characters of the PlantCLEF2022 dataset. On one hand, it is related with observation-level image classification, instead of usual image-level image classification. One observation refers to multiple images taken for a specific target, such as one field plant in the PlantCLEF2022 dataset, which makes one class with heterogeneous and robust features. On the other hand, PlantCLEF2022 introduced a large-scale plant-related image dataset with an immense image variance [18]. Because of the two characters, we argue that a pretrained model in the PlantCLEF2022 dataset can be a powerful start point towards plant image-related applications. Not trivially, our preliminary experimental results suggest that it benefits plant disease recognition on several public dataset, such as quicker convergence speed and better performance even with few number of images. To facilitate the related applications, we public our codes and pretrained model at https://github.com/xml94/PlantCLEF2022 and we hope that our work encourages the community to utilize the self-supervised ViT model, the PlantCLEF2022 dataset, and our pretrained model in the dataset.

2. Material and evaluation metric

Training dataset. PlantCLEF2022 dataset consists of training and testing dataset. The official training dataset has two categories, web and trusted. The web one covers about 1.1 million images and 57,000 classes, in which the images and their annotations are not directly from human experts. In contrast, the trusted one in a global scale is collected and annotated by experts, with 2,885,052 images and 80,000 classes. Each class includes 36.1 images averagely and to ease the class-imbalance issue, has no more than 100 images. Except for the images and their labels, some other information are also given, such as species and genus in plant taxonomy. As the datasets are huge and the limitation of computing devices, we just utilize the trusted one as training dataset considering the quality of the annotations.

The PlantCLEF2022 challenge differs from other plant-related tasks. First, the task objective is recognizing one plant identity given their images whereas other related tasks focus on disease recognition for specific plant, Plant Village [19], tomato leaf disease [20], and apple leaf disease [21]. Second, the main visual content of every image is diverse, such as leaf, fruit, flower, and habitat, not just leaf as in Plant Village and tomato dataset. Figure 2 and Figure 3 display some images of two classes from PlantCLEF2022 training dataset and we can see that the backgrounds, viewpoints, illuminations, sizes, and colors are different. Therefore, image variations [18] are essentially big, which requires classification model to learn robust yet heterogeneous features for each class. Besides, the images in PlantCLEF2022 are collected from real field with multiple resolutions, instead of lab as Plant Village with the same resolution.

Figure 2: Images of Cycas armstrongii Miq species from PlantCLEF2022 training dataset. The images from the same species are heterogeneous in background, viewpoint, and size.

Testing dataset and observation. The testing dataset of PlantCLEF2022 includes 55,306 images and similar to the training dataset, various organs in the images are interested. Diversely, this task asks to classify for each *observation*, instead of each image as usual image classification. One observation refers to an actual plant and multiple images can be taken for a single observation. In the testing dataset, 55,306 images are captured from 26,868 observations and Figure 4 displays six observations. Moreover, the testing dataset just includes part of the labels in the training dataset, instead of the same as the training dataset as usual image classification. Furthermore, the observation level classification requires the model to integrate its multiple images to output a final prediction, discussed in the next section. Details of the training and testing dataset of the PlantCLEF2022 challenge are shown in Table 1.

Evaluation metric. Macro averaged mean reciprocal rank (MA-MRR) is utilized to evaluated to different submissions for PlantCLEF2022 challenge. The challenge requests a submission with a rank based on 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]:

[1]https://en.wikipedia.org/wiki/Mean_reciprocal_rank

[2]https://androidkt.com/micro-macro-averages-for-imbalance-multiclass-classification/

Figure 3: Images of Aralia nudicaulis L. species from PlantCLEF2022 training dataset. The images from the same plant species are heterogeneous in background, illumination, and color.

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

where $rank_i$ refers to the rank position of the first relevant ground-truth label for the $i-$th observation from one class. Conceptually, if all of observation 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 enable us to observe different parts of plants, while MA-MRR allows more than one chances to recognize plant species for a specific image, 30 chances in the challenge.

Dataset	Details
Training	Has two sub-datasets, web and trusted. The web one has 1.1 million images and 57,000 classes of plant where the annotation is searched from internet and thus may be with noises. The trusted one, annotated by human experts, includes 2,885,052 images and covers 80,000 classes. The Trusted one gives meta-data about the organs or habitat of plant in each image. The images in the trusted training dataset embrace a huge variation, such as illuminations, background, colors. Figure 2 and Figure 3 display some images from two individual types of plant.
Testing	Includes 26,868 observations with 55,306 images. The classification should be done in observation-level, not image-level as usual image classification requests. One observation refers to one actual plant with several pictures taken from different viewpoints, which requires a classification model to combine the predictions of multiple images taken from a same observation. Figure 4 shows six observations with diverse organs or habitats. Besides, the testing dataset only share part of the plant identity in the trusted training dataset, instead of the same as usual image classification. Simultaneously, the images in the training and testing dataset are collected from real field and in a similar resolution, about 450×600.

Table 1
Details of the training and testing dataset in the PlantCLEF2022 challenge.

3. Method

3.1. Model and finetuning details

Masked autoencoder (MAE) [22] is chose to achieve the challenge as a self-supervised vision transformer because of its high performance and stable training process. Figure 5 illustrates its high-level architecture. In MAE, an image is firstly split into patches that are then randomly blocked. To save computation, only the unblocked patches are fed to an encoder to extract features, followed by a decoder to reconstruct the whole image. To pretrain the model, MAE employs a reconstruction loss and the original input image as the ground truth. After pretraining, the decoder is discarded and only the encoder is utilized taking unblocked images with a auxliary classifier head.

We borrow the ViT-large MAE model pretrained in ImageNet1k dataset. Random cropping and random horizontal flipping are leveraged as data augmentation strategy and the random masking, 75% as ratio, is another type of data augmentation. The model is trained with bacth size 4,096 and 800 epochs, 40 epochs as warming up. Besides, the learning rate is 0.00015 with 0.05 weight decay. The normalization is performed in block-level, instead of image-level as usual. Although only ViT-large model is leveraged for PlantCLEF2022 challenge, ViT-huge model is also encouraged but taken time and devices into consideration, we did not do experiments with this model.

Finetuning process. As finetuning the ViT-large MAE model with only four RTX 3090 GPUs, we set the actual batch size 512 and train the model 100 epochs. AdamW is employed as optimizer with learning rate 0.0005, 0.65 layer decay, and 0.05 weight decay. MixUp [23] and CutMix [24] are utilized as data augmentation. Besides, the added classifier is a linear function with 80,000 as the number of output, the number of class in PlantCLEF2022 trusted training

Figure 4: Six observations of testing dataset in PlantCLEF2022. One observation refers to an actual plant and we can take multiple images for single observation. The PlantCLEF2022 challenge requires classification in observation level, instead of image level as usual image classification.

dataset.

3.2. Integration towards observation

As discussed in the second section, the PlantCLEF2022 challenge is a classification in observation-level, instead of image-level, and each observation is asked to be with a class-score rank. Therefore, we should integrate the decisions when one observation has multiple images as our model obtains individual decision for each image. In this subsection, we analyze several possible strategies to get the final decision for each observation. Let r_s^j denote the testing probability score rank of the $j-$th image of one observation. Accordingly, r_c^j is the testing class rank with the same length. Besides, r_{c_i} and r_{s_i} are the i-th class-score pair and means the class and the corresponding probability score, respectively. The rank requires $r_{s_p} > r_{s_q}$ if $p < q$. Formally, the final desired output $\{r_s, r_c\}$, pair of class and corresponding score, is formulated as

$$\{r_c, r_s\} = \mathcal{I}(\cup_{j=1}^{n}(\{r_c^j, r_s^j\})), \tag{2}$$

where $\mathcal{I}$ means an integration operation and the observation has n individual images. There are three possible integration operations:

- Single-random: $\{r_c, r_s\} = \{r_c^j, r_s^j\}$ where $j = rand(n)$. Randomly sample an image from one observation and use the class-score rank pair of the random sample as the counterpart of the observation. In this way, observation-level classification deteriorates

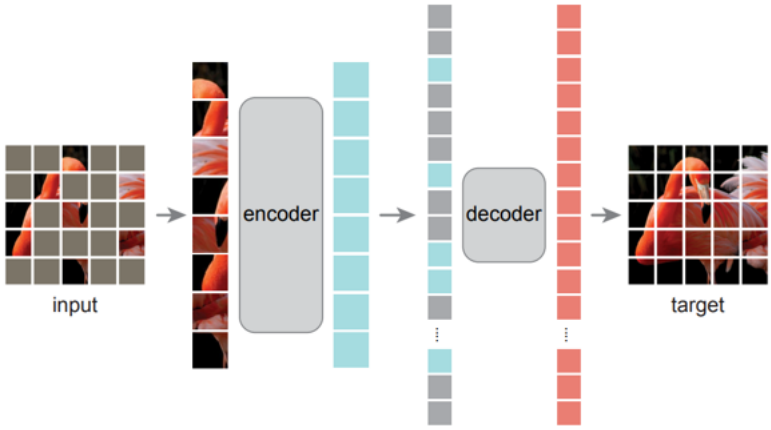

Figure 5: The high-level architecture of MAE [22]. With MAE, an image is split into patches that are then randomly blocked. The unblocked patches are fed to an encoder, followed by a decoder to reconstruct the whole input image. After the unsupervised pratraining, the decoder is discarded and only the encoder are utilized in down-stream task. The input is not blocked and a specific classifier is added after the encoder when fine-tuning the model in a target task. As the model is based on ViT and pretrained in an unsupervised manner, we termed our strategy ViT-based unsupervised transfer learning.

into image-level and hence this performance can be regarded as a baseline to see the impact of observation.

- Single-highest: $\{r_c, r_s\} = \{r_c^j, r_s^j\}$ where $r_{s_1}^j = max\{r_{s_1}^1, r_{s_1}^2, ..., r_{s_1}^n\}$. Single means the final rank pair is from only one image and highest denotes the highest top-1 score. Therefore, the rank pair is taken as the final prediction for the observation if the image has the highest top-1 score.
- Multi-sorted. Multi means that the final rank pair is from multiple images, instead of a single image. In this way, the scores of all images from an same observation are sorted: $sort(\cup_{j=1}^{n} \cup_{i=1}^{30} r_{s_i}^j)$; and then, after removing duplicates of same classes, the first required length (30) of class-score pair are taken as the final ranking pair.

4. Submissions

4.1. Official submissions

We report seven official submissions by fine-tuning the MAE model with different epochs. Because the first run was utilized to check the submission format, we do not report it. As the training dataset is huge, we did not finish the training process before the deadline of the challenge. With the setting as given in implementation detail subsection, fine-tuning costs about 5 hours for every epoch. We just finished 80 epochs before the deadline. Besides, we just applied the single-highest strategy towards observation because of limited time. Our results of official submissions are shown in Table 2 and Figure 6. We find that the performance becomes better when training longer and it seems that the performance can be improved further. Table

Submission Run	2	3	4	5	6	7	8
Epoch	12	15	24	49	67	77	80
MA-MRR	0.55865	0.56772	0.58110	0.60219	0.61632	0.62497	0.62692

Table 2
The performance of our official submissions.

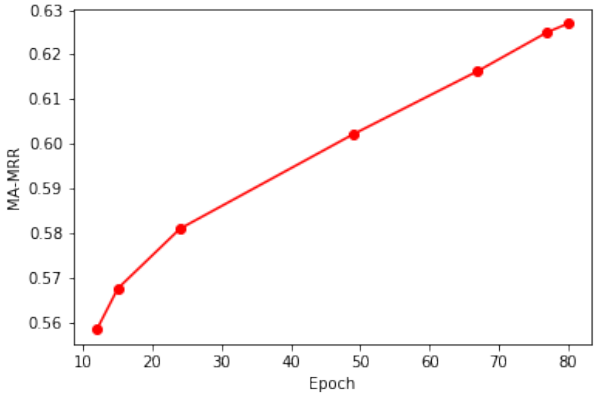

Figure 6: The tendency of our official submissions. It is seems that our method can be improved further via fine-tuning longer.

Team	MA-MRR
Ours	0.62692
Second place	0.60781
Third place	0.51043
Fourth place	0.46010

Table 3
The comparison of the official final performances for the PlantCLEF2022 challenge. Our method outperforms others by a clear margin.

3 shows several performances from different teams. From the table, our method outperforms other rivals by a clear margin.

4.2. Late submissions

After official submissions, we fine-tuned our model longer until 100 epochs. Figure 7 displays the complete performance of our submissions via fine-tuning a MAE model. We can see that training longer contributes more. Further, it seems that the performance can be improved more by extending the training process, which is leaved to our future work. Besides, we validate the performance of the single-highest and multi-sorted and the performances are displayes in Table 4. The multi-sorted one slightly surpasses the single-highest one, which demonstrates that observation-level identification has potential than image-level, as the single-highest can be cast as to find the optimal way to identify the species of plant. Because of time limitation, we did

Epoch	Single-highest	Multi-sorted
80	0.62692	No
100	0.63668	0.64079

Table 4
Performances of different integration strategies, single-highest and multi-sorted.

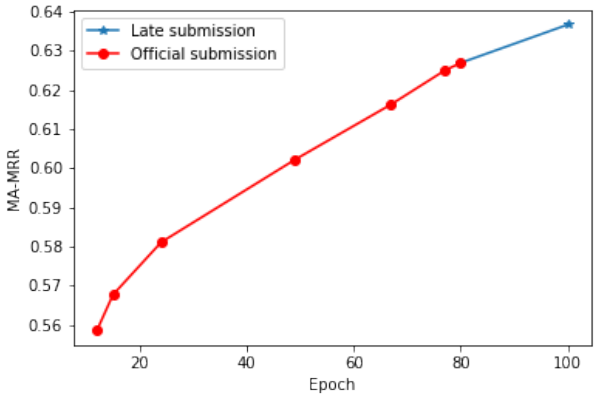

Figure 7: The complete performance of our submissions via fine-tuning a MAE model 100 epochs with single-highest integration strategy. The official submissions are highlighted by red colors.

not perform the single-random strategy but we guess that it was inferior to the single-highest.

5. Discussions

Future work. Although we secured the first place in the PlantCLEF2022 challenge, there are several points left. First of all, to achieve better performance, training longer is one of possible way. Besides, using more datasets to pretrain a model in a self-supervised manner, including PlantCLEF2022, is also encouraged. Secondly, the reasons why our strategy is effective and efficient for the challenge are not clear in current experiments. Towards this issue, ablation studies are desired, such replacing CNN with ViT and replacing supervised with self-supervised. Third, we only leveraged the large model of MAE, other types of model with different training strategies and loss function are also possible, such as contrast learning-based [25] [26] and text-image pair-based [27]. Fourth, we only used images and their annotations for the challenge and the meta-data introduced in material section were not utilized. Therefore, we emphasize that our work is just a start point about the challenge and more understandings can be made in the future.

Observations. Our work validates the effectiveness of observation-level classification. We argue that this idea can be extended to other related task where taking multiple images to improve the performance of a task. For example, most of plant diseases are recognized based on front leaf [19] [20] [28] [29]. On the other hand, some patterns in the back side of leaf are useful to recognize the type of disease [30]. Thus, observation-level classification would ease

Dataset	Details
PlantVillage	Includes leaves of 16 plants with 38 classes and 54,305 images. Images are taken in lab with similar illuminations and simple background. Some diseases are split into two parts according to their severities.
Apple2020	Includes only apple leaves with 3,642 images and 4 classes. Images are taken in the real field.
Apple2021	This dataset is an extension of Apple2020 but with 18,632 images and 6 classes.
TaiwanTomato	Owns 622 images of tomato leaves in 5 classes. Images are taken in the real field with more complex images. As having less images, it can be taken as a harder dataset.

Table 5
Details of four plant disease datasets.

some applications and is encouraged.

Dataset. As training a model in a similar source domain with a big dataset is beneficial to down-stream task with a little dataset [27, 11], PlantCLEF2022 dataset can be regarded as one of source dataset for plant-related down-stream tasks. For example, recognizing plant disease, classifying weed from crops, and detecting fruits. We perform some preliminary experiments on plant disease recognition on four public datasets, PlantVillage [19], Apple2020 [21], Apple2021 [31], and TaiwanTomato [32]. Details of the plant disease dataset are given in Table 5. The annotated images of the four datasets are split into three parts, training, validation and testing. The validation and testing dataset own 20% of the total annotated dataset while the training dataset owns 20%, 40%, and 60%. We compare two pretrained models to verify the PlantCLEF2022 dataset in plant disease recognition. One of the pretrained model is the model from MAE [22], self-supervised trained in ImageNet1k dataset. Another one is our model in this paper, finetuning the MAE model in PlantCLEF2022 dataset. The accuracy of the experiments is shown in Table 6 and two validation accuracy during the training process are displayed in Figure 8 and Figure 9. From the table and figures, we can see that the pretrained model in PlantCLEF2022 can improve the performance and speed the convergence, even with few labeled images. Therefore, we public not only our codes but also the fine-tuned models to encourage related applications.

6. Conclusion

In this paper, we leveraged a very simple transfer learning strategy to achieve the PlantCLEF2022 challenge and secured the first place with a clear superiority to other rivals. Our strategy differs from current popular strategy in two ways, pretrained in a self-supervised way instead of supervised, based on vision transformer instead of convolution neural network, which gives a better feature space and a more powerful model, respectively. Simultaneously, we analyze the PlantCLEF2022 training and testing datasets that include millions of images with plenty of meta-data. Through our analysis, we believe that the PlantCLEF2022 dataset are going to contribute plant related tasks, such as plant disease recognition as validated in our preliminary experiments.

Target Dataset	Source Dataset	20%	40%	60%
PlantVillage	ImageNet	99.5	99.8	99.8
	PlantCLEF2022	99.7	99.9	99.9
Apple2020	ImageNet	48.8	92.8	93.9
	PlantCLEF2022	95.3	97.2	97.5
Apple2021	ImageNet	93.2	93.7	95.2
	PlantCLEF2022	95.6	95.6	96.4
TaiwanTomato	ImageNet	25.2	43.3	55.9
	PlantCLEF2022	59.1	75.6	81.9

Table 6
The accuracy of plant disease recognition on multiple public datasets with different pretrained model. 20%, 40%, and 60% are the ratio of training dataset from the original target dataset with annotation. Validation and testing dataset are awalys in 20%. The validation is utilized to choose the best training model that is further tested in the testing dataset.

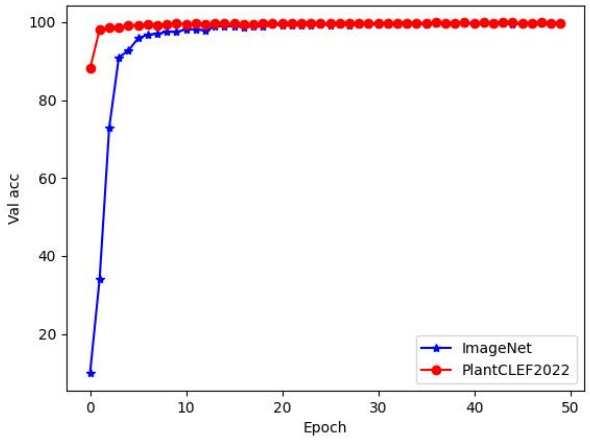

Figure 8: The validation accuracy curve of PlantVillage dataset. Pretrained with PlantCLEF2022 accelerates the convergence.

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 Korea Institute of Planning and Evaluation for Technology in Food, Agriculture, and Forestry (IPET) and Korea Smart Farm R&D Foundation (KosFarm) through 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 Korea government (MSIT) (2020R1A2C2013060).

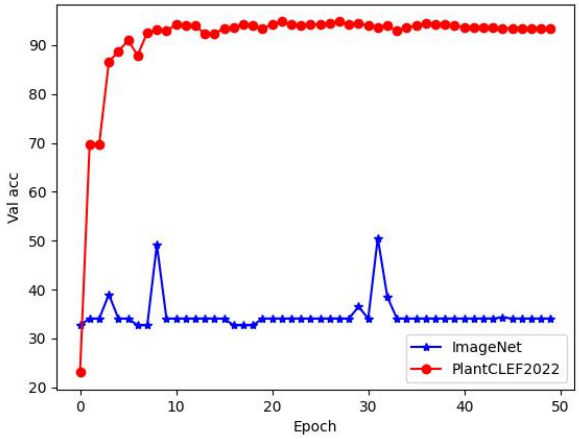

Figure 9: The validation accuracy curve of Apple2020 dataset. Pretrained with PlantCLEF2022 obtains better performance.

References

[1] H. Goëau, P. Bonnet, A. Joly, Overview of PlantCLEF 2022: Image-based plant identification at global scale, in: Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, 2022.

[2] A. Joly, H. Goëau, S. Kahl, L. Picek, T. Lorieul, E. Cole, B. Deneu, M. Servajean, A. Durso, H. Glotin, R. Planqué, W.-P. Vellinga, A. Navine, H. Klinck, T. Denton, I. Eggel, P. Bonnet, M. Šulc, M. Hruz, Overview of lifeclef 2022: an evaluation of machine-learning based species identification and species distribution prediction, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2022.

[3] Y. Wang, Q. Yao, J. T. Kwok, L. M. Ni, Generalizing from a few examples: A survey on few-shot learning, ACM computing surveys (csur) 53 (2020) 1–34.

[4] S. J. Pan, Q. Yang, A survey on transfer learning, IEEE Transactions on knowledge and data engineering 22 (2009) 1345–1359.

[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: 2009 IEEE conference on computer vision and pattern recognition, Ieee, 2009, pp. 248–255.

[6] N. H. Krishna, M. Rakesh, R. K. R, Plant species identification using transfer learning-plantclef 2020., in: CLEF (Working Notes), 2020.

[7] S. Chulif, Y. L. Chang, Herbarium-field triplet network for cross-domain plant identification. neuon submission to lifeclef 2020 plant., in: CLEF (Working Notes), 2020.

[8] Y. Zhang, B. D. Davison, Weighted pseudo labeling refinement for plant identification, Working Notes of CLEF (2021).

[9] S. Chulif, Y. L. Chang, Improved herbarium-field triplet network for cross-domain plant identification: Neuon submission to lifeclef 2021 plant, Working Notes of CLEF (2021).

[10] Z. Wu, Y. Xiong, S. X. Yu, D. Lin, Unsupervised feature learning via non-parametric

instance discrimination, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 3733–3742.
[11] 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.
[12] 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.
[13] L. Jing, Y. Tian, Self-supervised visual feature learning with deep neural networks: A survey, IEEE transactions on pattern analysis and machine intelligence 43 (2020) 4037–4058.
[14] G. E. Hinton, R. R. Salakhutdinov, Reducing the dimensionality of data with neural networks, science 313 (2006) 504–507.
[15] C. Doersch, A. Gupta, A. A. Efros, Unsupervised visual representation learning by context prediction, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 1422–1430.
[16] N. Komodakis, S. Gidaris, Unsupervised representation learning by predicting image rotations, in: International Conference on Learning Representations (ICLR), 2018.
[17] M. Xu, S. Yoon, D. Park, J. Lee, Unsupervised transfer learning for plant anomaly recognition, Smart Media Journal 11 (2022) 30–37.
[18] M. Xu, S. Yoon, A. Fuentes, D. S. Park, A comprehensive survey of image augmentation techniques for deep learning, arXiv preprint arXiv:2205.01491 (2022).
[19] D. Hughes, M. Salathé, et al., An open access repository of images on plant health to enable the development of mobile disease diagnostics, arXiv preprint arXiv:1511.08060 (2015).
[20] M. Xu, S. Yoon, A. Fuentes, J. Yang, D. S. Park, Style-consistent image translation: A novel data augmentation paradigm to improve plant disease recognition., Frontiers in Plant Science 12 (2021) 773142–773142.
[21] R. Thapa, K. Zhang, N. Snavely, S. Belongie, A. Khan, The plant pathology challenge 2020 data set to classify foliar disease of apples, Applications in Plant Sciences 8 (2020) e11390.
[22] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, R. Girshick, Masked autoencoders are scalable vision learners, arXiv preprint arXiv:2111.06377 (2021).
[23] H. Zhang, M. Cisse, Y. N. Dauphin, D. Lopez-Paz, mixup: Beyond empirical risk minimization, in: International Conference on Learning Representations, 2018.
[24] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, Y. Yoo, Cutmix: Regularization strategy to train strong classifiers with localizable features, in: Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 6023–6032.
[25] X. Chen, S. Xie, K. He, An empirical study of training self-supervised vision transformers, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 9640–9649.
[26] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, A. Joulin, Emerging properties in self-supervised vision transformers, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 9650–9660.
[27] 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.
[28] X. Zhao, K. Li, Y. Li, J. Ma, L. Zhang, Identification method of vegetable diseases based on transfer learning and attention mechanism, Computers and Electronics in Agriculture 193 (2022) 106703.
[29] Y. Li, X. Chao, Semi-supervised few-shot learning approach for plant diseases recognition, Plant Methods 17 (2021) 1–10.
[30] A. F. Fuentes, S. Yoon, J. Lee, D. S. Park, High-performance deep neural network-based tomato plant diseases and pests diagnosis system with refinement filter bank, Frontiers in plant science 9 (2018) 1162.
[31] R. Thapa, Q. Wang, N. Snavely, S. Belongie, A. Khan, The plant pathology 2021 challenge dataset to classify foliar disease of apples (2021).
[32] M.-L. Huang, Y.-H. Chang, Dataset of tomato leaves, Mendeley Data 1 (2020).

A. Online Resources

The codes and pretrained model are available via

- GitHub,
- Pretrained model in Google Drive.