Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 234 (2024) 188–195

Procedia Computer Science

www.elsevier.com/locate/procedia

Seventh Information Systems International Conference (ISICO 2023)

Medicinal plant recognition based on Vision Transformer and BEiT

Duy Tran Nguyen Nhut[a], Thinh Duong Tan[a], Trung Nguyen Quoc[a], Vinh Truong Hoang[b,*]

[a]*Department of Information Technology, FPT University, Ho Chi Minh City, Vietnam*
[b]*Faculty of Information Technology, Ho Chi Minh City Open University, Vietnam*

Abstract

The use of deep learning models for plant identification has garnered significant attention in the research community and has yielded promising results. In this study, we evaluate the performance of four state-of-the-art pre-trained deep learning models, namely EfficientNetB0, EfficientNetV2-S, Vision Transformer (ViT), and Bidirectional Encoder Image Transformer (BEiT), on the VNPlant-200 dataset, a complex dataset that comprises various species of medicinal plants captured in natural settings. Our results show that BEiT achieved the highest accuracy of 99.14%, outperforming the other models evaluated on this benchmark. These findings prove the effectiveness of these models in plant recognition tasks, particularly in the context of medicinal plants.

Keywords: Medical Plant Identification; VNPlant-200; EfficientNetB0; EfficientNetV2-S; ViT; BEiT

1. Introduction

Plant recognition is an important field of study that has gained considerable attention in recent years due to its applications in a variety of domains, including agriculture, forestry, and medicine. Accurate and reliable identification of plant species is critical in these domains for tasks such as monitoring biodiversity, identifying rare or invasive species, and discovering new medicinal compounds. Traditional methods of plant identification, such as manual identification based on physical characteristics, are time-consuming, labor-intensive, and often require specialized knowledge. With the advancement of deep learning-based approaches, however, plant recognition has been revolutionized, leading to the development of automated and efficient solutions that can perform plant identification tasks accurately and quickly.

One of the key challenges in plant recognition is the complexity of the plant dataset, particularly when dealing with medicinal plants. Medical plant identification involves recognizing and differentiating plant species that have similar morphological features, such as leaves or flowers. Moreover, the images may be taken in natural settings with varying conditions such as different lighting and angles, making the task of plant recognition even more challenging. Therefore, developing deep learning models that can accurately classify these plant species is crucial.

1877-0509
10.1016/j.procs.2024.02.165

In recent years, deep learning-based methods have shown remarkable progress in various computer vision tasks. Among these methods, state-of-the-art models like EfficientNetB0 [14], EfficientNetV2-S [15], Vision Transformer (ViT) [5], and Bidirectional Encoder Image Transformer (BEiT) [1] have emerged as popular choices due to their impressive performance on benchmark datasets. In this study, we aim to evaluate the performance of these pre-trained models on the VNPlant-200 dataset [9], which is a challenging dataset consisting of various species of medicinal plants captured in natural settings. The goal of this study is to investigate the efficacy of these models in plant recognition tasks, with a specific focus on identifying medicinal plant species. The findings of this study could provide valuable insights into the development of automated and efficient solutions for plant recognition tasks, ultimately benefiting the fields of agriculture, forestry, and medicine.

2. Related Works

The recognition of plants has become an increasingly significant research area, particularly in the field of medicine. To support this work, a large-scale dataset of Vietnamese medicinal plants was recently made available by T. N. Quoc and V. T. Hoang [9]. This dataset contains 20,000 images of 200 different plant species that were taken under outdoor conditions with varying angles and distances. As a result, it presents several challenges that are commonly encountered in practical medical plant classification applications and is therefore a valuable resource for researchers in this field. Traditional machine learning techniques, such as SIFT (Scale-Invariant Feature Transform), SURF (Speeded-Up Robust Features), and HOG (Histogram of Oriented Gradient), have been employed for use on this benchmark, but their performances were below 70% [9, 11]. This could be due to the complexity of the dataset, which makes it difficult to extract meaningful features from images.

To address this issue, deep learning has emerged as a promising method for improving the performance of plant recognition tasks. T. D. Minh and T. T. Minh [8] utilized a pre-trained feature extractor consisting of six deep-learning models to extract features from images. These features were then fed into traditional machine learning classifiers, resulting in a promising performance of 78%. Meanwhile, T. N. Quoc and V. T. Hoang [10] experimented with several pre-trained end-to-end deep learning models, achieving the highest accuracy of 88.26% using the Xception model.

Advanced learning methods, such as ensemble learning with pre-trained deep learning models, have also been employed to enhance performance. For instance, O. A. Malik et al. [7] used an ensemble learning method to improve accuracy up to 95.69%. Another research conducted by Lida Shahmiri et al. [13] proposed a sample selection method, which increased the best accuracy of this benchmark to 97.31%. These outcomes demonstrate the potential of deep learning techniques in medicinal plant recognition tasks.

Despite these accomplishments, further improvement is still possible. The EfficientNet family has recently gained widespread use in various image recognition tasks due to its ability to achieve the best performance while retaining a lightweight architecture. Furthermore, Transformer models in natural language processing tasks have gained popularity, leading to an increase in the use of Transformer-based models in computer vision research due to their outstanding performance on benchmark datasets. Consequently, this study aims to assess the performance of four pre-trained models, including EfficientNetB0, EfficientNetV2-S, ViT, and BEiT, on the VNPlant-200 dataset. The findings of this study will contribute to the development of more effective techniques for the recognition of medicinal plants, which could have significant implications for medicine and related fields.

3. Architectures

3.1. EfficientNetB0

EfficientNetB0's architecture (shown in Table 1) incorporates a unique compound scaling method that simultaneously scales up the depth, width, and resolution of the network, resulting in a highly efficient yet accurate model. The primary building block of EfficientNetB0 is the mobile inverted bottleneck MBConv (illustrated in Fig. 1), which comprises depthwise and pointwise convolutions and squeeze-and-excitation (SE) modules. By integrating these convolutional layers, EfficientNetB0 inherits the computational efficiency of the MobileNet architecture [6] while maintaining high performance. Furthermore, by incorporating the SE module into the MBConv block, EfficientNetB0 can capture more complex patterns and generate better feature representations, resulting in improved accuracy of the model. In

Table 1. EfficientNetB0 Architecture [14].

Stage	Operator	Resolution	#Channels	#Layers
1	Conv3x3	224x224	32	1
2	MBConv1, k3x3	112x112	16	1
3	MBConv6, k3x3	112x112	24	2
4	MBConv6, k5x5	56x56	40	2
5	MBConv6, k3x3	28x28	80	3
6	MBConv6, k5x5	14x14	112	3
7	MBConv6, k5x5	14x14	192	4
8	MBConv6, k3x3	7x7	320	1
9	Conv1x1 & Average Pooling & FC	7x7	1280	1

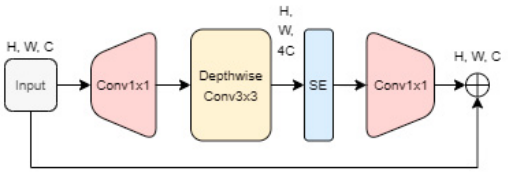

Fig. 1. MBConv

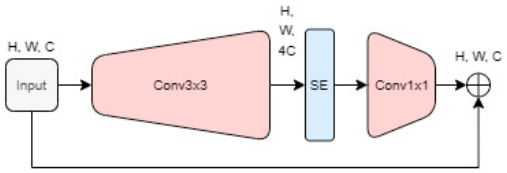

Fig. 2. Fused-MBConv

addition, the compound scaling method employed in EfficientNetB0 enables the optimal use of computational resources by determining the appropriate scaling coefficients for the depth, width, and resolution of the network. The balanced scaling of these dimensions leads to exceptional accuracy while remaining computationally efficient, making EfficientNetB0 a highly suitable option for various computer vision tasks.

3.2. EfficientNetV2-S

Table 2. EfficientNetV2-S Architecture [15].

Stage	Operator	Stride	#Channels	#Layers
1	Conv3x3	2	24	1
2	Fused-MBConv1, k3x3	1	24	2
3	Fused-MBConv4, k3x3	2	48	4
4	Fused-MBConv4, k3x3	2	64	4
5	MBConv4, k3x3, SE0.25	2	128	6
6	MBConv6, k3x3, SE0.25	1	160	9
7	MBConv6, k3x3, SE0.25	2	256	15
8	Conv1x1 & Global Average Pooling & FC	-	1280	1

EfficientNetV2 [15], which was introduced by Mingxing Tan and Quoc V. Le in 2021, is the second iteration of the EfficientNet family that incorporates new design features and scaling laws. One of the primary modifications is the use of a smaller expansion ratio for MBConv, which reduces memory access overhead. Additionally, the architecture makes extensive use of both MBConv and fused-MBConv in early layers. Fused-MBConv replaces the depthwise conv3x3 and expansion conv1x1 in MBConv with a single regular conv3x3, as depicted in Fig. 2. Although fused-MBConv can better utilize mobile or server accelerators, experiments have demonstrated that it is necessary to maintain a proper balance of both MBConv and fused-MBConv, as the model will be slowed down if all MBConv blocks are replaced with fused-MBConv blocks. Furthermore, EfficientNetV2 favors smaller 3x3 kernel sizes and entirely eliminates the last stride-1 stage in the original EfficientNet due to its large parameter size and memory access overhead. The baseline model of the EfficientNetV2 family is the EfficientNetV2-S, and its architecture is described

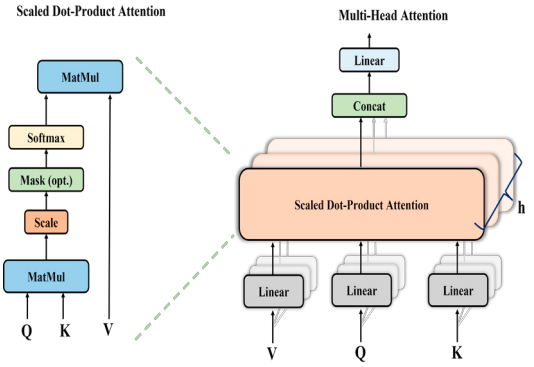

Fig. 3. Multi-Head Self Attention

in Table 2. EfficientNetV2-S is scalable to larger networks such as EfficientNetV2-M or EfficientNetV2-L with some additional optimizations.

3.3. Vision Transformer (ViT)

The Vision Transformer represents a recent development in computer vision architecture, leveraging a Transformer-based model commonly employed in natural language processing [16] for visual tasks. It was first introduced by a team from Google in 2020 [5]. By utilizing the Multi-Head Self Attention technique, the Vision Transformer demonstrates an exceptional capacity for feature learning. As a result, it has emerged as a highly promising approach for improving performance on a range of computer vision tasks. In Figure 3 visualize the structure of Multi-head self-attention layers with an Attention block:

$$Attention(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = softmax(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}})\mathbf{V} \tag{1}$$

where **Q, K, V** are ”Queries”, ”Keys” and ”Values” respectively [16].

To prepare an image for input into the Vision Transformer (ViT), a pre-processing step is necessary. This involves dividing the image into a series of patches, using a patch size (16, 16) and stride of (16, 16) that can be specified by the user. The patches are then flattened into vectors and passed through a Dense layer with linear activation to enable trainable embedding. Position encoding is then added to each encoded patch, which is necessary to prevent the loss of meaning that could result from the random arrangement of the patches. To enable classification, the ViT model embeds a [CLS] token along with the position-encoded patches. These inputs are then passed through the Transformer Encoder Network, which consists of a series of multi-head self-attention and Dense layers. For image classification, only the output from the [CLS] token patch is required, as the Transformer will pool the classification feature vector. The final step involves passing the vector through the softmax function to obtain the probabilities for prediction. Fig. 4 show the general architecture of ViT. The Vision Transformer (ViT) has demonstrated strong performance in the context of large dataset training, due to its ability to effectively capture and extract general features from the data, which can help to mitigate issues related to overfitting. As a result, ViT has been identified as a promising approach for improving classification accuracy in challenging datasets, such as those encountered in medical plant analysis. Fine-tuning techniques are particularly effective when applied to ViT models, resulting in improved performance when compared to other popular architectures, such as EfficientNet.

3.4. Bidirectional Encoder Image Transformer (BEiT)

Bidirectional Encoder Image Transformer (BEiT) is introduced by a team from Microsoft, in 2021 [1]. BEiT is an extension of the Vision Transformer (ViT) architecture that introduces Masked Image Modeling (MIM), inspired by the Masked Language Modeling (MLN) approach used in BERT [4]. In addition to dividing the image into patches,

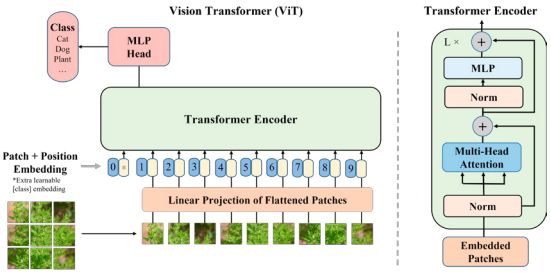

Fig. 4. ViT Architecture

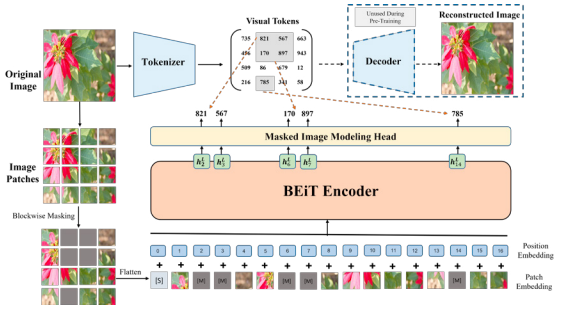

Fig. 5. BEiT Architecture

BEiT converts the image into visual tokens using a visual tokenizer trained on DALL-E [12]. During training, approximately 40% of the patches (up to 75 patches) are masked before being passed through the Transformer Encoder Network. The goal is to train the model to predict the token of the masked patches and reconstruct them. After training, BEiT replaces the encoder vs. decoder head with a GlobalPooling2D layer and a Dense layer for class prediction. Alternatively, classification can be achieved using the [CLS] token similarly. The resulting model is then fine-tuned on a new dataset to classify new classes. In essence, BEiT operates similarly to a Convolutional Block in that it seeks to learn an optimal approach for encoding an image. The overview of BEiT structure is shown in Fig. 5. As cited in [1], incorporating intermediate fine-tuning with ImageNet labels has been shown to improve the performance of BEiT. Furthermore, the authors have demonstrated that self-supervised BEiT can learn semantically meaningful regions through pre-training, thereby leveraging the rich supervisory signals present in images. Therefore, we have opted to employ BEiT for fine-tuning tasks in the medical plant dataset, to compare its performance against that of ViT and other CNN models such as EfficientNet, through visualization.

4. Experiments

4.1. Dataset

In this study, the VNPlant-200 dataset [9] was utilized for experimentation. The dataset summary, as presented in Table 3, provides a comprehensive overview of its properties. The VNPlant-200 dataset consists of a substantial and intricate collection of medicinal plants. To accurately represent the challenges encountered in real-world image classification, the images were captured in natural settings, as illustrated by the sample images in Fig. 6. The dataset comprises plant images captured from different viewpoints, lighting conditions, and environments, among other factors. Therefore, it serves as an excellent representation of plant recognition tasks encountered in practical settings.

To ensure unbiased evaluation, the dataset was partitioned into training and testing sets in a ratio of 60%/40%, respectively, with a larger size allocated to the testing set due to the complexity of the dataset. This partitioning ratio was maintained for all subsequent models trained in the experiment to ensure consistency in the evaluation.

Table 3. Summary of VNPlant-200 Dataset.

Details	Configurations
Number of species	200
Number of images per species	100
Image resolution	256x256, 512x512
Type of image	Image of entire plant
Image Condition	Natural Environment

Fig. 6. Samples from VNTPlant-200 Dataset.

4.2. Training Details

Four deep learning models, namely EfficientNetB0, EfficientNetV2-S, ViT, and BEiT, were utilized to train on the VNPlant-200 dataset in this article. These models were pre-trained on ImageNet [3] and slightly modified to classify 200 species. The experiment and analysis were conducted on a cloud computing notebook equipped with an NVIDIA Tesla T4 16GB GPU. Finally, accuracy was employed as the main evaluation metric to measure the performance of the models.

EfficientNetB0 was trained on images with a resolution of 224x224 to assess the performance of the smallest model on the recommended image resolution. In contrast, EfficientNetV2-S, ViT, and BEiT were trained on images with a higher resolution of 384x384 to evaluate their maximum performance potential given the available hardware.

We optimized the hyperparameters of these models, as presented in Table 4. To improve generalization and combat overfitting during training, image augmentation was applied using the RandAugment [2]. This method was deemed appropriate for this study as it is an automated and randomized augmentation method that provides excellent regularization during training. As EfficientNetB0 is a small model, a RandAug magnitude of 15 was found to be appropriate for augmentation. The image size was selected to train the least computationally expensive model and observe its performance. A batch size of 32 was chosen as it typically affects only the VRAM required to train the model.

For EfficientNetV2-S, ViT, and BEiT, the image size and RandAug magnitude were increased to accommodate the higher image resolution of 384x384. The RandAug increment was sufficient to prevent overfitting. To reduce the VRAM requirement for training with a higher image resolution, the batch size was lowered. However, for ViT and BEiT, the Dropout ratio was eliminated as it caused significant instability during training, and a reduction in the learning rate was used instead.

Table 4. Hyperparameter Configurations.

Hyperparameter	EfficientNetB0	EfficientNetV2-S	ViT	BEiT
Learning rate	1e-5	1e-5	1e-6	1e-6
β_1	0.55	0.55	0.55	0.55
β_2	0.9	0.9	0.9	0.9
Dropout rate	0.25	0.25	-	-
RandAug magnitude	15	30	30	30
Batch size	32	16	16	16

5. Results

Upon examination of Table 5, it is evident that the performance of EfficientNetB0 and EfficientNetV2-S surpasses that of the best-performing CNN network-based methods [8, 10]. Moreover, empirical observations revealed that the EfficientNetB0 model significantly outperformed other architectures, despite having only 5 million parameters. During the initial 35 epochs Fig. 7, the training progress of the model was relatively slow, as it adapted to the new dataset

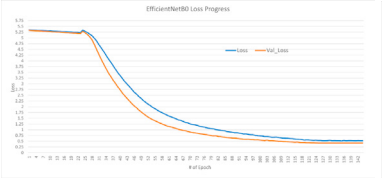

Fig. 7. EfficientNetB0 Loss Progress

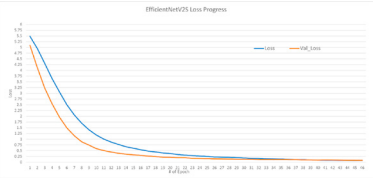

Fig. 8. EfficientNetV2-S Loss Progress

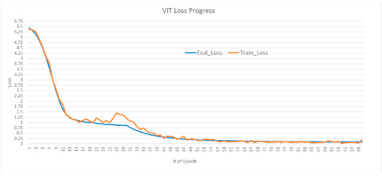

Fig. 9. ViT Loss Progress

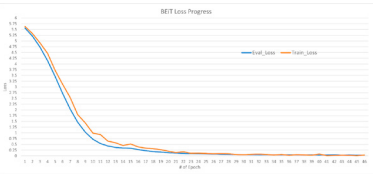

Fig. 10. BEiT Loss Progress

Table 5. The comparison between several methodologies and our pre-trained approach in VNPlant-200.

Model	Image size	Accuracy
Xception + Logistic Regression [8]	256x256	78.0%
Xception [10]	224x224	88.26%
Ensemble Learning (Weight Mean Ensemble) [7]	224x224	95.69%
Xception + MIGT [13]	224x224	97.31%
EfficientNetB0 (Ours)	224x224	89.01%
EfficientNetV2-S (Ours)	384x384	97.71%
ViT (Ours)	384x384	98.24%
BEiT (Ours)	384x384	**99.14%**

and gradually accelerated toward convergence. Due to the increased number of parameters, the training progress of EfficientNetV2-S was found to be smoother than that of EfficientNetB0 (Fig. 8). Specifically, the model achieved an 8.7% increase in accuracy compared to EfficientNetB0, and 0.4% compared to Xception combining MIGT [13]. ViT and its derivative, BEiT, which utilize state-of-the-art Vision Transformer technology, have demonstrated superior performance compared to EfficientNetV2-S with 98.24% and 99.14% respectively. Furthermore, the results achieved by ViT and BEiT demonstrate superior performance compared to methods that incorporate CNN networks in conjunction with other solutions [7, 13].

6. Conclusion

The current study describes a notable improvement in plant classification through the application of state-of-the-art models, including EfficientNet, ViT, and BEiT, on the VNPlant-200 dataset. By implementing fine-tuning techniques and hyper-parameter optimization, the BEiT model achieved the highest performance on the VNPlant-200 dataset. Furthermore, the results indicate that the use of transformer-based architectures can significantly enhance classification accuracy when compared to other methods. Of particular note, an increase in image resolution was found to be a critical factor in achieving performance improvements. Although the current study was limited to image resolutions of 512x512, future investigations could potentially benefit from exploring higher image resolutions to further enhance the overall performance of the VNPlant-200 dataset.

References

[1] Bao, H., Dong, L., Wei, F., 2021. Beit: BERT pre-training of image transformers. CoRR abs/2106.08254. URL: https://arxiv.org/abs/2106.08254, arXiv:2106.08254.

[2] Cubuk, E.D., Zoph, B., Shlens, J., Le, Q.V., 2019. Randaugment: Practical data augmentation with no separate search. CoRR abs/1909.13719. URL: http://arxiv.org/abs/1909.13719, arXiv:1909.13719.
[3] Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255. doi:10.1109/CVPR.2009.5206848.
[4] Devlin, J., Chang, M., Lee, K., Toutanova, K., 2018. BERT: pre-training of deep bidirectional transformers for language understanding. CoRR abs/1810.04805. URL: http://arxiv.org/abs/1810.04805, arXiv:1810.04805.
[5] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. CoRR abs/2010.11929. URL: https://arxiv.org/abs/2010.11929, arXiv:2010.11929.
[6] Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H., 2017. Mobilenets: Efficient convolutional neural networks for mobile vision applications. URL: http://arxiv.org/abs/1704.04861, arXiv:1704.04861.
[7] Malik, O.A., Faisal, M., Hussein, B.R., 2021. Ensemble deep learning models for fine-grained plant species identification, in: 2021 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE), pp. 1–6. doi:10.1109/CSDE53843.2021.9718387.
[8] Minh, T.D., Minh, T.T., Quoc, T.N., Hoang, V.T., 2023. Features extraction based on sota models for medicinal plant images recognition, in: Abraham, A., Hanne, T., Gandhi, N., Manghirmalani Mishra, P., Bajaj, A., Siarry, P. (Eds.), Proceedings of the 14th International Conference on Soft Computing and Pattern Recognition (SoCPaR 2022), Springer Nature Switzerland, Cham. pp. 465–473.
[9] Nguyen, T., Truong Hoang, V., 2021. VNPlant-200 – A Public and Large-Scale of Vietnamese Medicinal Plant Images Dataset. pp. 406–411. doi:10.1007/978-3-030-49264-9_37.
[10] Nguyen Quoc, T., Truong Hoang, V., 2020. Medicinal plant identification in the wild by using cnn, in: 2020 International Conference on Information and Communication Technology Convergence (ICTC), pp. 25–29. doi:10.1109/ICTC49870.2020.9289480.
[11] Quoc, T.N., Hoang, V.T., 2021. A new local image descriptor based on local and global color features for medicinal plant images classification, in: 2021 International Conference on Decision Aid Sciences and Application (DASA), pp. 409–413. doi:10.1109/DASA53625.2021.9682391.
[12] Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I., 2021. Zero-shot text-to-image generation. CoRR abs/2102.12092. URL: https://arxiv.org/abs/2102.12092, arXiv:2102.12092.
[13] Shahmiri, L., Wong, P., Dooley, L.S., 2022. Accurate medicinal plant identification in natural environments by embedding mutual information in a convolution neural network model, in: 2022 IEEE 5th International Conference on Image Processing Applications and Systems (IPAS), pp. 1–6. doi:10.1109/IPAS55744.2022.10053008.
[14] Tan, M., Le, Q.V., 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. CoRR abs/1905.11946. URL: http://arxiv.org/abs/1905.11946, arXiv:1905.11946.
[15] Tan, M., Le, Q.V., 2021. Efficientnetv2: Smaller models and faster training. CoRR abs/2104.00298. URL: https://arxiv.org/abs/2104.00298, arXiv:2104.00298.
[16] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I., 2017. Attention is all you need. CoRR abs/1706.03762. URL: http://arxiv.org/abs/1706.03762, arXiv:1706.03762.