Plant Identification Using Convolution Neural Network and Vision Transformer-Based Models

Virender Singh [*] , Mathew Rees [*] , Simon Hampton , Sivaram Annadurai [*]

Posted Date: 18 August 2023

doi: 10.20944/preprints202308.1330.v1

Keywords: Plant recognition; Image Processing; Convolution neural network; Vision transformer; Classification

Article

Plant Identification Using Convolution Neural Network and Vision Transformer-Based Models

Virender Singh [1,*], Mathew Rees [2,3,*], Simon Hampton [2] and Sivaram Annadurai [1,*]

[1] Data Scientist, Publicis Sapient, 2870, Building - Virgo, Bagmane Constellation Business Park, Marathahalli, Bengaluru, India
[2] Royal Horticultural Society, Wisley, GU23 6QB UK
[3] School of GeoSciences, University of Edinburgh, EH9 3FF, UK
* Correspondence: virender.singh1@publicissapient.com (V.S.); mathew.rees@ed.ac.uk (M.R.); sivaram.annadurai@publicissapient.com (S.A.)

Abstract: Identification of plants is a challenging task which aims to identify the family, genus, and species level according to morphological features. Automated deep learning-based computer vision algorithms are widely used for identifying plants and can help users to narrow down the possibilities. However, numerous morphological similarities between and within species make the classification difficult. In this paper, we tested a custom convolution neural network (CNN) and vision transformer (ViT) based models using the PyTorch framework to classify plants. We used a large dataset of 88K and 16K images for classifying plants at genus and species levels respectively. Our results show that for classifying plants at the genus level, ViT models perform better compared to CNN-based models ResNet50 and ResNet-RS-420, and other state-of-the-art CNN-based models suggested in previous studies on a similar dataset. The ViT model achieved top accuracy of 83.3% for classifying plants at the genus level. ViT models also perform better for classifying plants at the species level compared to CNN-based models ResNet50 and ResNet-RS-420, with a top accuracy of 92.5%. We show that the correct set of augmentation techniques plays an important role in classification success.

Keywords: plant recognition; image processing; convolution neural network; vision transformer; classification

1. Introduction

Plants are one of the most essential life forms on Earth and play a significant role in maintaining healthy ecosystems. In order to obtain information about the uses of any plant, users must first identify the plant, by matching the physical characteristics to a specific name (either scientific latin name or common name). Knowing one or more discriminating features of an unknown plant (e.g., shape, color, petal, sepal length) helps in identifying the candidate species. Identification of plants using key features is difficult for people who don't have specific knowledge of plants and even for specialists such as botanist, agroforestry managers, and scientist to identify plants correctly at different hierarchical levels [1]. Variation of key characters among species and even within species, are some of the challenges for identifying plants manually. Hence, automated species identification can be used for the identification of plants [2]. Automated plant classification is an important research area in computer vision. It is a fine-grained classification task concerned with the identification of plants at various hierarchical levels, such as family, genus, or species level [3]. A user can take a picture of the plant using a camera or mobile device and then analyze it with a plant identification model to identify the plant or a list of possible candidate plants at various hierarchical levels. The identification problem faces several challenges due to inter-class similarities among plant families. Another problem is huge intra-class variations in color, background, occlusion, shape, and illumination within the same plant class such as family, genus, or species. Several studies have been conducted to address the plant classification problem using deep learning-based algorithms and have been able to accomplish significant success in classifying plants [2,4,5]. Compared to traditional machine learning algorithms where features were manually selected and extracted, deep learning-

 doi:10.20944/preprints202308.1330.v1

based algorithms automatically detect increasingly higher-level features from data [6]. Several works have shed light on plant identification using deep neural networks, which significantly improved the accuracy of large-scale plant classification tasks. Various Convolutional Neural Network (CNN) models have been proposed and implemented for plant identification tasks and achieved better performance compared to other artificial neural networks (ANN) and CNN-based models suggested in prior research [7,8]. Transformer-based architecture has become de facto in natural language processing (NLP) tasks, its application in computer vision attains significantly good performance compared to state-of-the-art convolution neural networks. Vision transformer models have achieved better performance than other CNN-based models for fine-grained image classification tasks [9]. Training both the CNN and Vision transformer models to contain millions of parameters requires large amounts of data to properly constrain the optimization. The requirement for extensive computational resources for training these models motivates to use of transfer learning with pre-trained networks [10–12]. While the transformer models have become the de facto standard in NLP applications, their applications in computer vision tasks remain limited but have been rising recently. Vision transformer (ViT) attain excellent results compared to state-of-the-art convolutional networks while requiring substantially fewer computational resources to train [13]. ViT is a model used in the field of computer vision that employs a transformer-like architecture over patches of the image. It works like the transformers used in the field of natural language processing (NLP). Over the years, deep CNNs have been the state-of-the-art networks for image classification but ViT has shown great potential in achieving competitive performance for complex image classification tasks [13]. Internally, transformers learn by calculating the relationship between pairs of input tokens (words in the case of a string), termed attention in NLP tasks. In computer vision, an Image is split into various fixed-size patches. These image patches are used the same way as tokens and ViT calculates the relationship among pixels between various patches. Each of the image patches is then linearly embedded and patch embeddings are finally augmented with one-dimensional position embeddings. Positional information is introduced into the input using position embeddings, which is learned during training. An extra learnable “classification token” is added at the start of the sequence to the patch embedding. The resulting sequence of the embedding vector is fed to the encoder part of the Transformer architecture. A classification head attached to the encoder output gets the value of learnable class embedding to perform the classification based on its state.

This paper has three main contributions. First, a ViT with a Custom balanced loss function is used for handling class imbalance and improving the model performance. Second, the proposed combination of augmentation techniques enhances the quality of data and improves the model performance. Lastly, for analyzing the distribution of images captured from a near or far distance within classes, CNN based classification model is implemented. Near/Far image distribution helped in visualizing the data imbalance issue and enhancing the quality of the training dataset. It also helped in analyzing the performance of the proposed models on both types of distribution. Finally, it also helped in balancing the data when there was a significant difference in the distribution of near and far images by adding more images and using augmentation to include more diversity within classes. All these components have significantly improved the plants classification performance at the genus and species level which can be extended to classification at the cultivar level. Several augmentation techniques were used in combination for this study to enhance the model performance. The research proposed by Hiary et al. [23] shows the importance of image augmentation for improving the model performance. The authors used fine-tuned VGG-16 model to classify flower species for Oxford-17, Oxford-102 and the dataset consists of 612 flower images from 102 categories [24]. Generally, a large amount of diverse training data is required because the small-size dataset may easily overfit the training model. Data augmentation can address this problem and helps in improving the size and quality of training data. Zhong et al., [25] introduced a novel technique for augmentation called Random Erasing, a new augmentation technique to improve the quality and size of training data. Random Erasing selects a rectangular region randomly in an image and erases random pixel values. This improves the robustness of the models and has better generalization capabilities.

 doi:10.20944/preprints202308.1330.v1

CNN (ResNet50 and ResNet-RS-420) and transformer (ViT) based models were used for this study. Images are acquired using digital cameras, mobile phones or other equipment by the Royal Horticultural Society (RHS), UK. Images are then pre-processed, and augmentation techniques [14] are applied to enhance the size and quality of training data. After that, the area of interest was segmented, and the features were extracted. Based on the extracted features, plants are classified at the genus or species level. An automated flower classification task is a difficult task because of the considerable number of similarities among various flower species and due to intra-class variation. More differences in the background, viewpoint, occlusion, flower image scale, indoor-outdoor lighting conditions, climate and season are some of the problems which make the classification of flowers more difficult [15]. In this research, 113 different plant genera and 53 species were considered. It is very difficult to differentiate these plants from a certain distance, specifically for the human eye. Automated image classification using a deep learning-based approach provides performance above the functions of the human eye and produces accurate results [16–18]. Many techniques have been proposed for the classification of plants. Şekeroğlu et al. [4] proposed a leaf classification system using a neural network to identify 27 different types of leaves and achieved a recognition rate of 97.2%. Deep learning-based Convolution neural networks (CNN) are quite popular and achieved signification success in image classification based-task in recent years [19–21]. CNN models are widely used for plant classification problems and achieved significantly better performance compared to other machine learning and ANN-based networks [7,22]. Deep learning (DL) based methods are widely used for plant recognition tasks with large image datasets. Heredia et al. [8] used a PlantNet database consisting of 250K images belonging to more than 1,500 plant species. The authors used the ResNet50 model and achieved significant improvement in model performance compared to widespread classification models on test data composed of thousands of different species [8]. CNN-based methods have also been used in health care such as medical image classification, and tumor detection [16,17]. The recently proposed transformer-based approach appears to be a major step toward plant identification tasks. Using the self-attention paradigm, ViT models can achieve better results for image classification tasks compared to CNN-based models, such as AlexNet, EfficientNet and ResNet without applying any convolution approaches. The research proposed by Conde et al. [9] on four popular fine-grained benchmarks: CUB-200-2011, Stanford Cars, Stanford Dogs, and FGVC7 Plant Pathology have used a multi-stage ViT framework and achieved better performance compared to CNN-based models. Given the huge amount of training data and computational resources, ViT has shown better performance compared to CNN models in image classification tasks [13]. Based on the literature review, we have used CNN and ViT-based models for the proposed research because these models have shown better performance for classifying plants and flowers in the past.

2. Methods

This research consisted of four main steps, which are image pre-processing, augmentation, feature extraction and classification. First, the plant images were acquired by the Royal Horticultural Society (RHS) and combined with two open-source datasets, PlantCLEF2015 and iNaturalist. The images were pre-processed using augmentation techniques to improve the quality of the training data. The processed image features were extracted using CNN or ViT models. Finally, the extracted features were trained, and the plant classification was performed using CNN or ViT. We also used image augmentation and sampling techniques to balance the distribution of the near and far captured images to improve the model performance.

2.1. Data Collection

PlantCLEF dataset focuses on 1,000 different herb, tree and fern species centered in France and neighboring countries. It contains 113,205 pictures belonging to 1,000 species. PlantCLEF dataset has information about the plants at different hierarchical levels (family, genus, or species). iNaturalist dataset contains around 1 million images for 4,271 plant species. Figure 1 illustrates the distribution of the RHS dataset for 113 genera and 53 species. Figures 2 and 3 illustrate the distribution of

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 August 2023 doi:10.20944/preprints202308.1330.v1

PlantCLEF and iNaturalist images used for classifying plants at genus and species level respectively. After combining the three datasets, we selected 113 genera and 53 species representing 88K images and 16K images, respectively. Figure 4 illustrates the distribution of genera and species of the combined dataset. The datasets were highly imbalanced, with some species containing a far greater number of images than others.

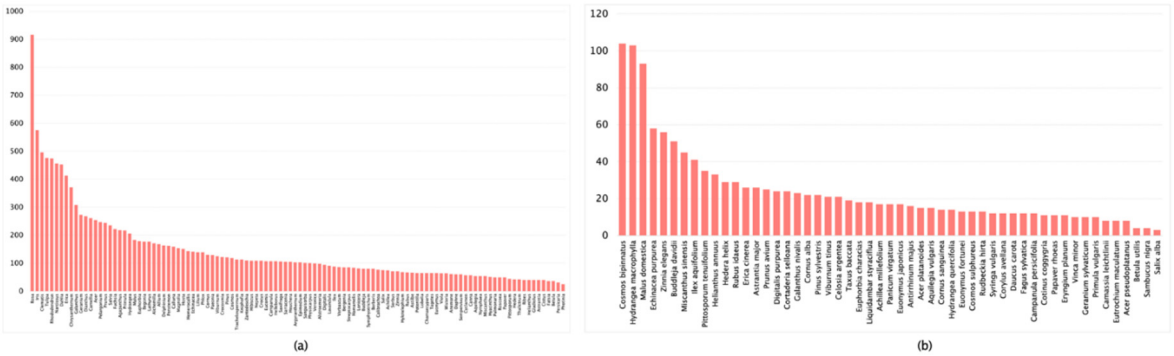


Figure 1. (a) RHS genera distribution. (b) RHS species distribution.

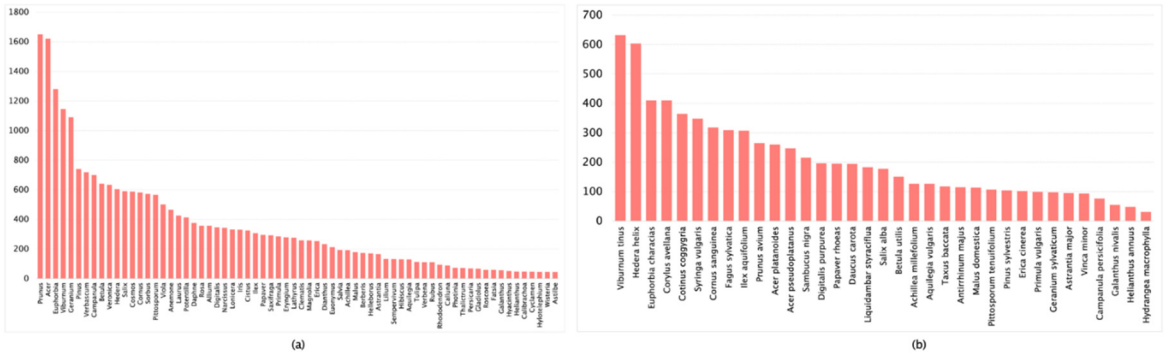


Figure 2. (a) PlantCLEF genera distribution. (b) PlantCLEF species distribution.

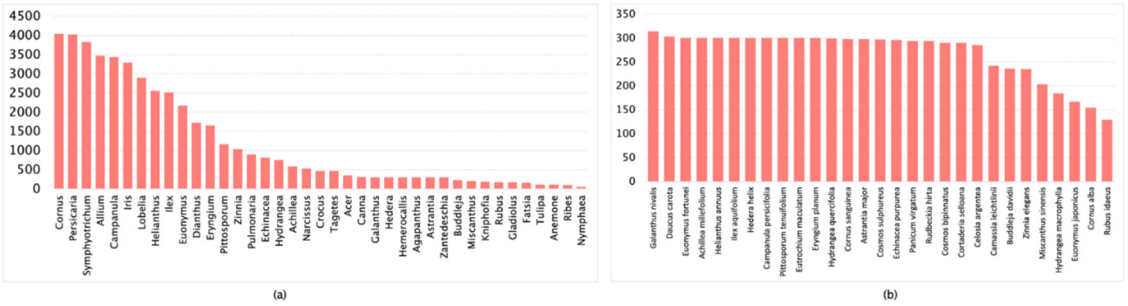
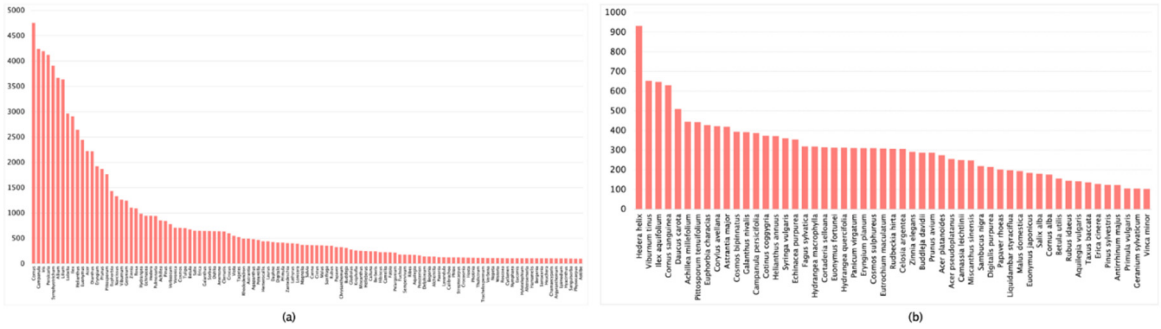


Figure 3. (a) iNaturalist genera distribution. (b) iNaturalist species distribution.



Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 August 2023 doi:10.20944/preprints202308.1330.v1

Figure 4. (a) Combined dataset genera distribution. (b) Combined dataset species distribution.

2.2. Image pre-processing

As a data preprocessing method, Image augmentation plays an important role in improving the model performance for deep learning-based networks. We used a variety of augmentation techniques in combination with each other to add more information and increase the dataset diversity. Several augmentation techniques, such as Resizing, Color jitter, Gaussian blur, greyscale, Random perspective, Random Rotation, Random Cropping, Sharpness and Grey Scale as shown in Figure 5, are used in combination. Images are resized to 384x384 pixels, and the center is cropped to 224x224 pixels after doing empirical testing with various combination of pixel values. Several combinations of different augmentation methods are tried with various hyperparameter values, and then the best values are chosen by empirical testing to improve the model performance. Augmentation methods help in adding more diversity to the dataset. Image Augmentation also helped in handling class imbalance where there are a smaller number of images.

A CNN-based model was implemented for classifying near or far-captured images labelled manually which helps in seeing the distribution of near and far-captured images for each class. The model was trained on a subset of the PlantCLEF dataset containing approximately 12,000 images. Images of plants were classified as either near-captured or far-captured images. The images were normalized to make the computation efficient. To improve the model performance, we implemented a transfer learning-based pre-trained VGG16 model for classifying near or far-captured images.

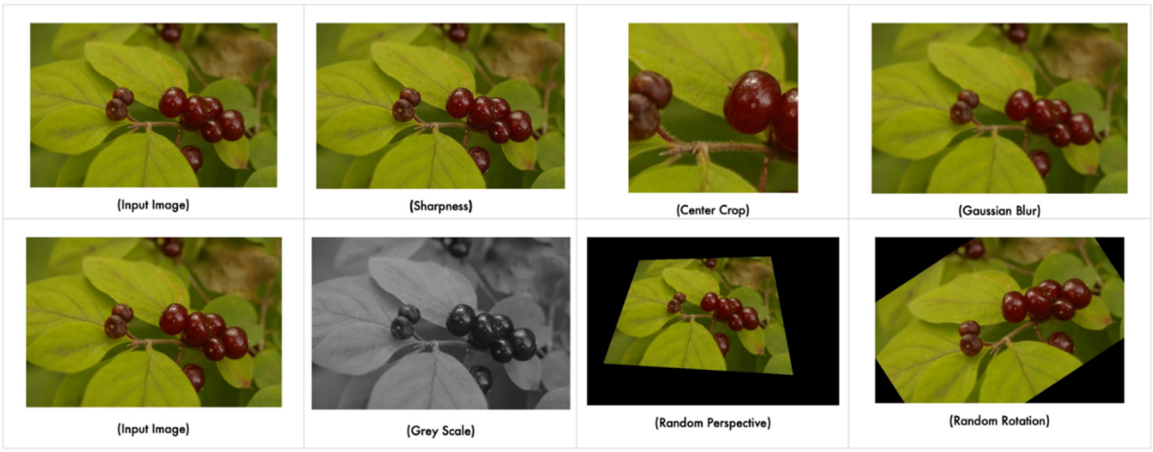


Figure 5. Examples of Image Augmentation techniques.

Figure 6 illustrates the near/far image distribution for the proposed dataset consisting of 113 genera after combining the RHS, PlantCLEF and iNaturalist datasets. Near/Far image distribution for 53 species after combining all three datasets is illustrated in Figure 7.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 August 2023 doi:10.20944/preprints202308.1330.v1

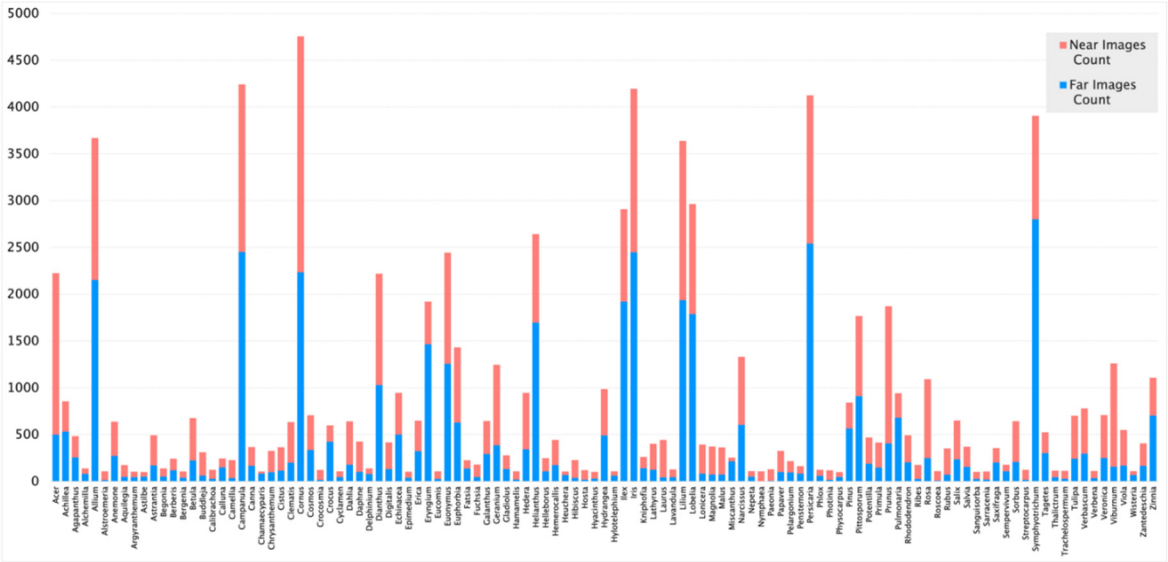


Figure 6. Near/far images distribution for plant genera.

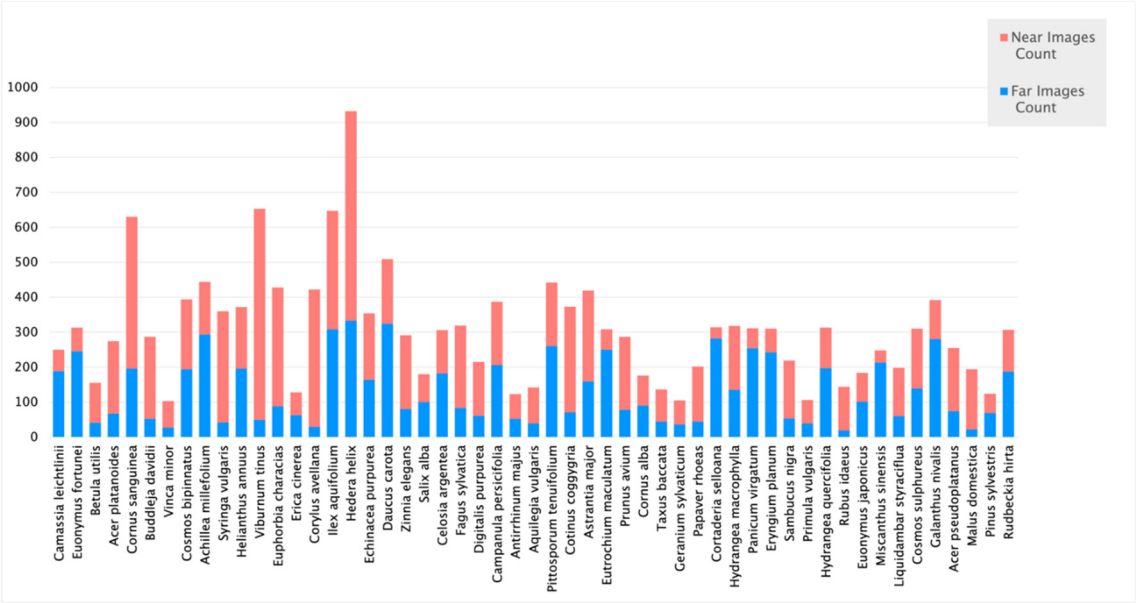


Figure 7. Near/far images distribution for plant species.

2.3. Convolution Neural Network

Custom pre-trained InceptionV3, ResNet50 and ResNet420 were used for this study. Figure 8 shows the CNN-based plant classification system for this research. We used cross validation technique to check the model performance on validation dataset. We tried multiple experiments with different CNN-based networks (InceptionV3, ResNet50, and ResNet420) using images belonging to 39 genera, where each class had at least 100 images. In experiment 1, we used 10,128 images belonging to 39 genera. Pre-trained InceptionV3 network with a cross-entropy loss function is used for identifying plants. The model was fine-tuned by adding custom layers using the TensorFlow framework to make it more suitable for the plant's classification task. In experiment 2, the pre-trained ResNet50 model on the ImageNet database with cross-entropy loss function is used. Augmentation techniques have been applied to the proposed dataset and used for feature extraction using CNN followed by a classification task to identify plants at the genus or species level. PyTorch framework is used for model implementation and training. In experiment 3, we combined our original dataset with open-source PlantCLEF and iNaturalist datasets to improve the data size and quality. During the first two experiments, we did not cover much variation within classes to cover the overall

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 August 2023 doi:10.20944/preprints202308.1330.v1

information. Therefore, transfer learning based fine-tuned ResNet50 model is used for classifying 334 genera with at least 100 images in each class. Approximately 220K images for 334 genera after augmentation were used for training the model. The fine-tuned ResNet50 model with PyTorch framework on the augmented dataset is used for model building and training purposes. In experiment 4, We used around 88K images belonging to 113 genera with an average image count per class of 778. Augmentation techniques, such as Resizing, Color jitter, Gaussian blur, greyscale, Random perspective, Random Rotation, Random Cropping, Sharpness and Grey Scale are used for improving the quality of the training dataset and handling the class imbalance. After augmentation, the Image count has been increased to 300K images. The best hyperparameter values and a combination of augmentation techniques are chosen after empirical testing of the model. Custom SoftMax balanced loss function with the ability to handle image imbalance issues is used for model training [27]. Pre-trained ResNet420 model trained on the ImageNet database with custom SoftMax balanced loss function using PyTorch framework is fine-tuned and used for model building and training. Finally, Species classification is done in experiment 5, we have used around 16k images for 53 classes before augmentation. After applying selected augmentation techniques, the image count has been increased to 150K images. Fine-tuned ResNet420 model with a custom balanced loss function is used for species classification.

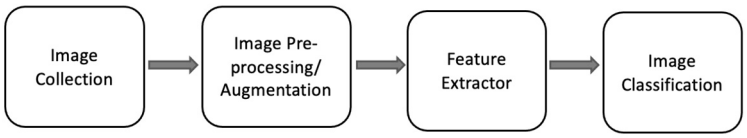


Figure 8. CNN-based plant classification system.

2.4. Vision Transformer

The ViT models were implemented using the PyTorch framework and can classify plants at different hierarchical levels, such as genus or species level. In experiment 6, we used around 88K images for 113 genera with an average image count per class of 778. Augmentation techniques, such as Resizing, Color jitter, Gaussian blur, greyscale, Random perspective, Random Rotation, Random Cropping, Sharpness and Grey Scale were used for improving the quality of the training dataset. After augmentation, the Image count has been increased to 300K images. Custom SoftMax balanced loss function with the ability to handle image imbalance issues is used for ViT model training [27]. Pre-trained ViT model trained on the ImageNet database with SoftMax balanced loss function is used for model building. The model is then fine-tuned and trained using the last 4 blocks of the ViT model and two linear layers which have been added at the end of the ViT model. In experiment 7, around 16k images belonging to 53 classes have been picked up for species classification. After applying the augmentation techniques mentioned in Section 2.2, the image count has been increased to 150K. Species classification has been done using the same ViT model architecture used for genera classification with custom SoftMax balanced loss function and is fine-tuned for species classification. Figure 9 shows the ViT architecture for the plant classification task used for this study.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 August 2023 doi:10.20944/preprints202308.1330.v1

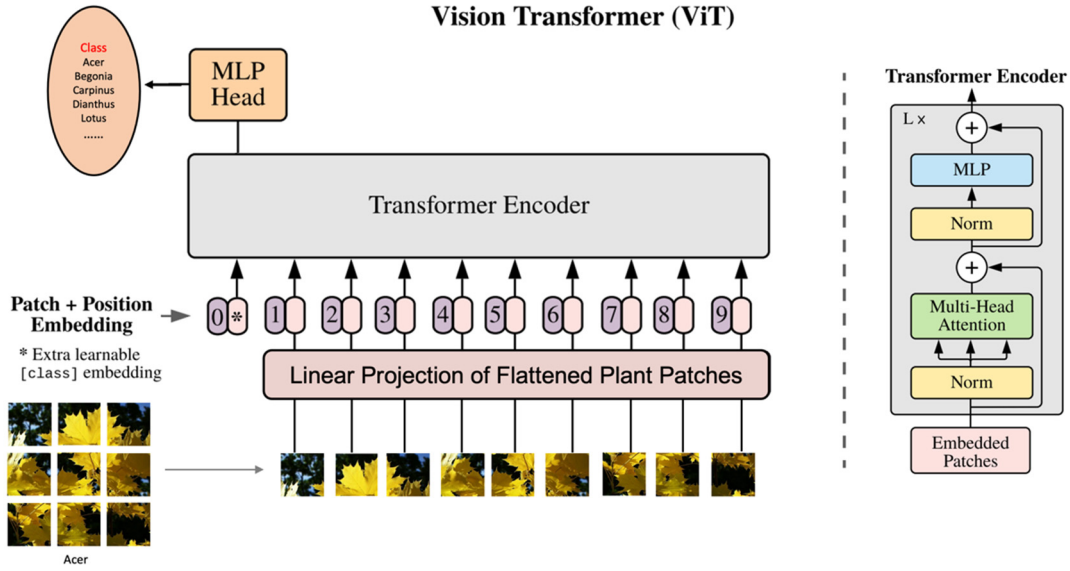


Figure 9. Diagram of Vision Transformer Architecture.

3. Result

3.1. Image pre-processing

The CNN based model for classifying images as near or far achieved a validation accuracy of 90.34% and test accuracy of 83% and overall model performance increased significantly after handling the near/far distribution, as shown in Table 1.

Table 1. Near Far classification table.

Dataset	Images Count	Classification Accuracy
Custom dataset (Images taken from PlantCLEF database)	~ 12k	~ 83%

3.2. Experiments using CNNs

In experiment 1, the fine-tuned InceptionV3 network with cross-entropy loss function achieved validation and testing accuracy of 76% and 35% respectively. In experiment 2, the pre-trained ResNet50 model on the ImageNet database with a cross-entropy loss function achieved the validation and test accuracy of 58% and 25% respectively. In experiment 3, the fine-tuned ResNet50 model achieved validation and testing accuracy of 68% and 23% respectively. In experiment 4 the fine-tuned ResNet420 model achieved validation and testing accuracy of 83% and 71% respectively. Lastly, in experiment 5, the fine-tuned pre-trained ResNet420 model with a custom balanced loss function achieved validation and testing accuracy of 94% and 84% respectively. A comparison of different experiments and techniques is illustrated in Table 2.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 August 2023 doi:10.20944/preprints202308.1330.v1

Table 2. CNN model comparison table.

Experiment	Dataset	Genus / Species	Class Count	Images Count	Augmentation	Images Count (Augmented)	CNN Model	Epochs	Training Time	Validation Accuracy	Test accuracy	Test Accuracy (Top 3)
1.	RHS	Genus	39	10k	No	-	InceptionV3	35	2h 30 min	76%	35%	-
2.	RHS	Genus	39	10k	Yes	30k	ResNet50	20	4h 50 min	58%	25%	-
3.	RHS + PlantCLEF	Genus	334	46k	Yes	220K	ResNet50	60	14h	68%	23%	-
4.	RHS + PlantCLEF +iNaturalist	Genus	113	88k	Yes	300K	ResNet-RS-420	82	26h 15min	83%	71%	83%
5.	RHS + PlantCLEF +iNaturalist	Species	53	16k	Yes	150K	ResNet-50	65	16h 25min	94%	84%	92.5%

3.3. Experiments using ViT

In experiment 6, the fine-tuned ViT model trained on ImageNet database with custom SoftMax balanced loss function achieved the validation and testing accuracy of 86% and 83% respectively. Finally, in experiment 7, the fine-tuned ViT model with SoftMax balanced loss function achieved validation and testing accuracy of 96% and 92.5% respectively. The model experiment summary table, as illustrated in Table 3, shows the performance of different experiments using ViT performed in this research. The top 3 predicted results metric was also used for measuring model performance, where images were classified correctly if it were present in the top 3 predicted results. From the comparison of results illustrated in Figures 10 and 11, ViT models outperformed CNN based models.

Table 3. ViT model performance.

Experiment	Dataset	Genus/ Species	Class Count	Images Count	Augmentation	Images Count (Augmentation)	CNN Model	Epochs	Training time	Validation Accuracy	Test accuracy	Test Accuracy (Top 3)
6.	RHS + PlantCLEF +iNaturalist	Genus	113	88k	Yes	300K	ViT	70	30h	86%	83%	92.5%
7.	RHS + PlantCLEF +iNaturalist	Species	53	16k	Yes	150K	ViT	50	17h	96%	92.5%	97.5%

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 August 2023 doi:10.20944/preprints202308.1330.v1

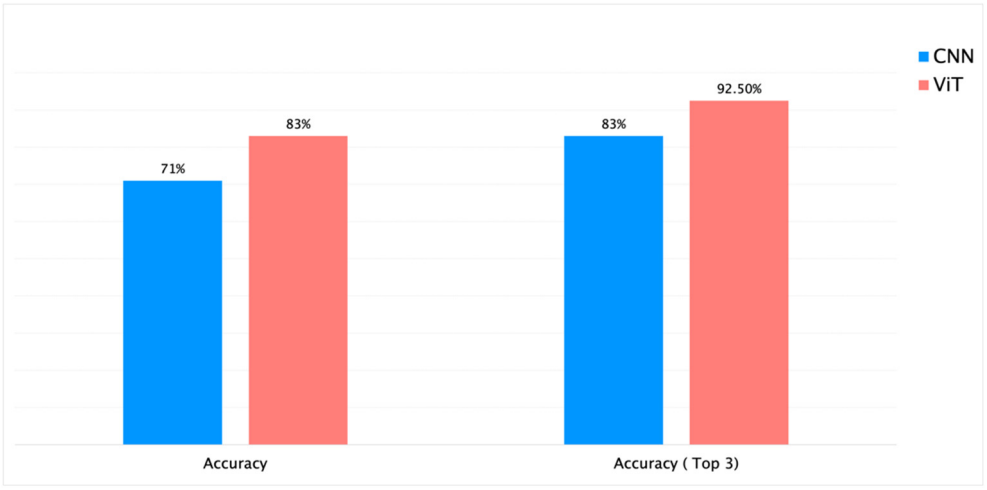


Figure 10. Comparison of performance for CNN (blue) and ViT (red) models for classifying genera.

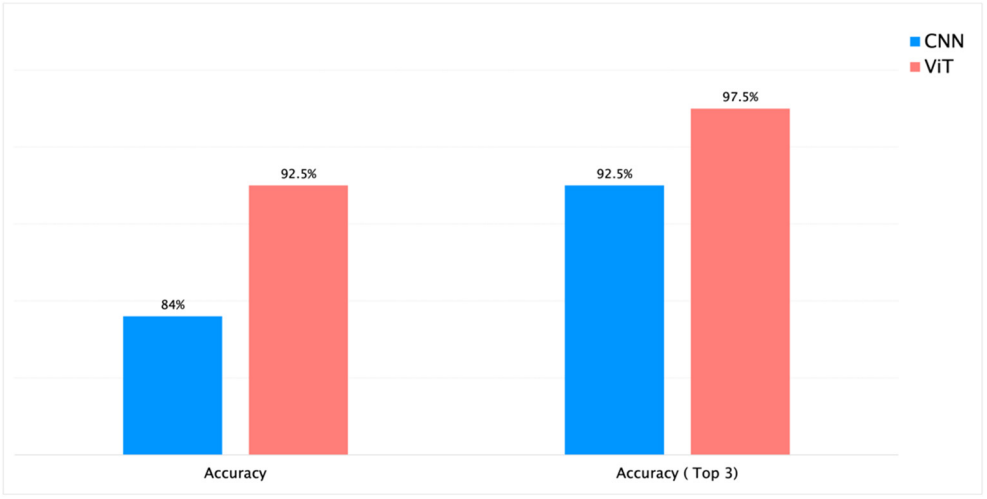


Figure 11. Comparison of performance for CNN (blue) and ViT (red) models for classifying species.

4. Discussion

The first part of this research was evaluated by employing fine-tuned CNN models, such as InceptionV3, ResNet50, and ResNet-RS-420 for feature extraction followed by a classification task. Next, ViT models are used for classifying plants at the genera and species level. By doing the comparison between Tables 2 and 3, ViT models performed better compared to CNN-based models for the plant's classification task. Based on the experiment results, it is seen that the correct set of augmentation alone or in combination plays an important role in improving the size and quality of training data, which helps in improving the classification performance. In terms of performance, the testing accuracy of ViT is 83% and 92.5% for classifying plants at genus and species level respectively. Proposed ViT models, fine-tuned for plant classification tasks can be an efficient automated plant classification since they can achieve significantly better performance than fine-tuned CNN-based models. Classifying near or far captured images mentioned in section 2.3 helped in seeing the distribution of near and far captured images for each class. During testing with unseen images, plant images which were misclassified, most of them were images captured from a far distance. The reason behind the misclassification of more images captured from far distance compared to near-captured images is the significant difference in image count for near and far-captured images for each class on which the CNN model is trained. Another reason for the misclassification of far-captured images is the similarity and less variance among different plant classes. Additionally, fine-tuning it with more

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 August 2023 doi:10.20944/preprints202308.1330.v1

trainable layers might improve the model performance but it increases the training time and is computationally more expensive. Feedback approach can be used for the model retraining, as shown in Figure 12. Plant images can be captured from the mobile camera using the web/mobile app and sent to the plant classification model API. User can provide feedback in the form of comments or ratings, based on the top predicted results returned by plants classification API. Top predicted results with respective probabilities and API request Id can be stored in databases such as AWS DynamoDB, SQL, or any other DB schema. Based on the returned response from API, user can provide feedback in the form of comments or ratings which can be used for model retraining and maintenance. API response along with user feedback is passed to feedback API where responses are stored and analyzed. Retrain the model if needed based on the feedback received from feedback API.

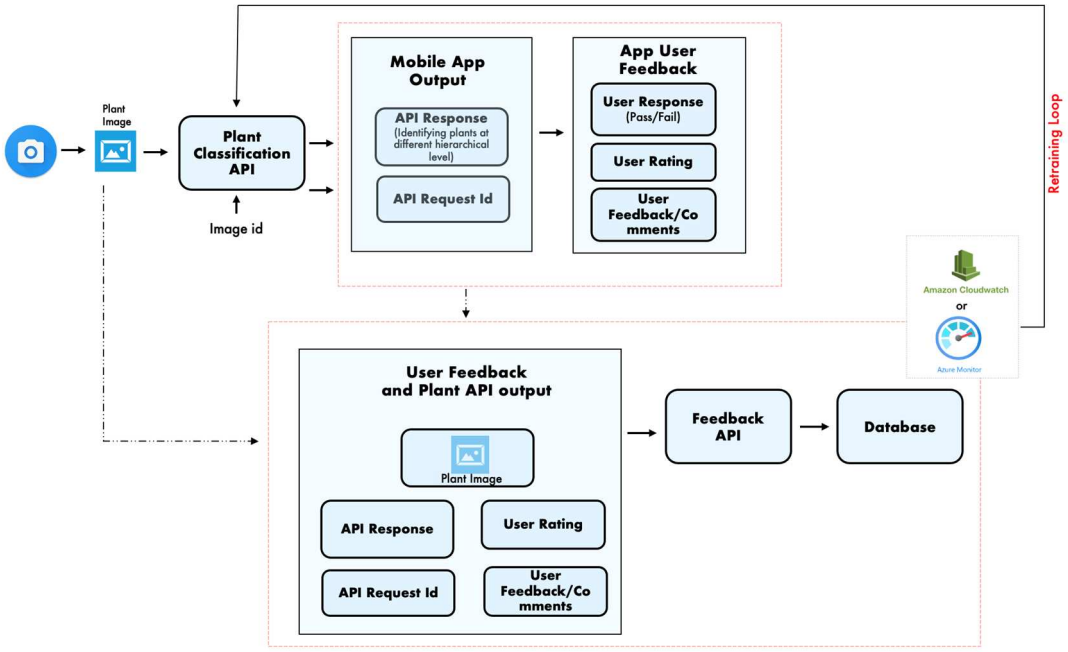


Figure 12. Diagram of model feedback and retraining approach.

5. Conclusions

In this work, we conclude that ViT models perform better than CNN-based models and extract more information about the features when classifying plants. We also highlight that augmentation plays an important role in enhancing the data quality and making the network more robust and generalizable. A Custom CNN-based model has been implemented for classifying near or far-captured images which helped in inferring that far-captured images are misclassified more compared to near ones. The experiments using the custom SoftMax balanced loss function suggested that the proposed loss function performed better than the most adopted cross-entropy loss for plant classification tasks. In future work, we would like to include more far-captured images, so that we have an equal distribution of near and far-captured images and improve the near/far classification model with more manually annotated images. We can improve the model performance by handling the misclassification for far-captured plant images and making it more generalizable. Proposed models and techniques for handling data imbalance can be extended for classifying plants at the cultivar level, fruit grading, plant/crop disease classification, quality assessment, flower classification and more. The proposed research can also be used and scaled with location attributes, such as device location, and country. By considering more features and narrowing down the plants based on location attributes, a more robust model with improved performance can be implemented. Based on user input provided on different plant categories such as gardening, indoor and outdoor plants, category-specific models can be implemented to improve the model performance. Tree-based approaches, such as top-down or bottom-up can be implemented for different hierarchical levels,

such as family, genus and species based on the use case requirements. Like the near/far distribution classification model used in this study, analysis of leaf, stem, or flower distribution can also complement the proposed research to improve the model performance.

Author Contributions: Virender Singh: Conceptualization, Methodology, Software, Validation, Formal Analysis, Investigation, Data curation, Visualization, Writing – Original Draft. Mathew Rees: Validation, Writing – Review & Editing. Simon Hampton: Conceptualization, Funding acquisition, Project administration. Sivaram Annadurai: Conceptualization, Supervision, Project administration, Data curation, Formal analysis, Writing – Review & Editing.

Funding: This work was funded by the Royal Horticultural Society.

Acknowledgement: We acknowledge many colleagues at Publicis Sapient for their help. Shanya Anand for helping in the model training and testing; Vineet Jaiswal, Abhishek Kumar, and Khushbu Varshney for providing guidance and useful discussions.

Conflicts of Interest: The authors have no declarations of competing interests.

References

1. A. Joly *et al.*, "LifeCLEF 2016: Multimedia Life Species Identification Challenges," Sep. 2016, vol. 9822, pp. 286–310. doi: 10.1007/978-3-319-44564-9_26.
2. O. M. Gaston KJ, "Automated species identification: why not?", doi: 10.1098/rstb.2003.1442.
3. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255. doi: 10.1109/CVPR.2009.5206848.
4. B. Şekeroğlu and Y. İnan, "Leaves Recognition System Using a Neural Network," *Procedia Comput Sci*, vol. 102, pp. 578–582, 2016, doi: https://doi.org/10.1016/j.procs.2016.09.445.
5. J. Wäldchen and P. Mäder, "Plant Species Identification Using Computer Vision Techniques: A Systematic Literature Review," *Archives of Computational Methods in Engineering*, vol. 25, no. 2, pp. 507–543, 2018, doi: 10.1007/s11831-016-9206-z.
6. Y. LeCun, Y. Bengio, and G. Hinton, "Deep Learning," *Nature*, vol. 521, pp. 436–444, May 2015, doi: 10.1038/nature14539.
7. J. Liu, S. Yang, Y. Cheng, and Z. Song, "Plant Leaf Classification Based on Deep Learning," in *2018 Chinese Automation Congress (CAC)*, 2018, pp. 3165–3169. doi: 10.1109/CAC.2018.8623427.
8. I. Heredia, "Large-Scale Plant Classification with Deep Neural Networks," in *Proceedings of the Computing Frontiers Conference*, 2017, pp. 259–262. doi: 10.1145/3075564.3075590.
9. M. v Conde and K. Turgutlu, "Exploring Vision Transformers for Fine-grained Classification," *CoRR*, vol. abs/2106.10587, 2021, [Online]. Available: https://arxiv.org/abs/2106.10587
10. S. J. Pan and Q. Yang, "A Survey on Transfer Learning," *IEEE Trans Knowl Data Eng*, vol. 22, no. 10, pp. 1345–1359, 2010, doi: 10.1109/TKDE.2009.191.
11. R. Ribani and M. Marengoni, "A Survey of Transfer Learning for Convolutional Neural Networks," in *2019 32nd SIBGRAPI Conference on Graphics, Patterns and Images Tutorials (SIBGRAPI-T)*, 2019, pp. 47–57. doi: 10.1109/SIBGRAPI-T.2019.00010.
12. M. Mehdipour Ghazi, B. Yanikoglu, and E. Aptoula, "Plant identification using deep neural networks via optimization of transfer learning parameters," *Neurocomputing*, vol. 235, pp. 228–235, 2017, doi: https://doi.org/10.1016/j.neucom.2017.01.018.
13. A. Dosovitskiy *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," *CoRR*, vol. abs/2010.11929, 2020, [Online]. Available: https://arxiv.org/abs/2010.11929
14. C. Shorten and T. M. Khoshgoftaar, "A survey on Image Data Augmentation for Deep Learning," *J Big Data*, vol. 6, no. 1, p. 60, 2019, doi: 10.1186/s40537-019-0197-0.
15. D. Guru, Y. H. Kumar, and M. Shantharamu, "Texture Features and KNN in Classification of Flower Images," *International Journal of Computer Applications, Special Issue on RTIPPR*, vol. 1, pp. 21–29, Jan. 2010.
16. M. Toğaçar, B. Ergen, and Z. Cömert, "BrainMRNet: Brain tumor detection using magnetic resonance images with a novel convolutional neural network model," *Med Hypotheses*, vol. 134, Jan. 2020, doi: 10.1016/j.mehy.2019.109531.
17. L. Faes *et al.*, "Automated deep learning design for medical image classification by health-care professionals with no coding experience: a feasibility study," *Lancet Digit Health*, vol. 1, no. 5, pp. e232–e242, 2019, doi: https://doi.org/10.1016/S2589-7500(19)30108-6.
18. R. Geirhos, D. H. J. Janssen, H. H. Schütt, J. Rauber, M. Bethge, and F. Wichmann, "Comparing deep neural networks against humans: object recognition when the signal gets weaker," *ArXiv*, vol. abs/1706.06969, 2017.

19. S. Turaga *et al.*, "Convolutional Networks Can Learn to Generate Affinity Graphs for Image Segmentation," *Neural Comput*, vol. 22, pp. 511–538, Sep. 2009, doi: 10.1162/neco.2009.10-08-881.
20. H. Lee, R. Grosse, R. Ranganath, and A. Ng, "Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations," in *Proceedings of the 26th International Conference On Machine Learning, ICML 2009*, Sep. 2009, p. 77. doi: 10.1145/1553374.1553453.
21. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017, doi: 10.1145/3065386.
22. L. and Z. X.-P. and Z. X. and S. L. and H. Z.-K. and G. Y. Liu Zhiyu and Zhu, "Hybrid Deep Learning for Plant Leaves Classification," in *Intelligent Computing Theories and Methodologies*, 2015, pp. 115–123.
23. H. Hiary, H. Saadeh, M. Saadeh, and M. Yaqub, "Flower classification using deep convolutional neural networks," *IET Computer Vision*, vol. 12, no. 6, pp. 855–862, Sep. 2018, doi: 10.1049/iet-cvi.2017.0155.
24. J. Zou and G. Nagy, "Evaluation of model-based interactive flower recognition," *Proceedings - International Conference on Pattern Recognition*, vol. 2, pp. 311–314, 2004, doi: 10.1109/ICPR.2004.1334185.
25. Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random Erasing Data Augmentation," 2020. [Online]. Available: https://github.com/zhunzhong07/Random-Erasing.
26. Y. Wu, X. Qin, Y. Pan, and C. Yuan, "Convolution neural network based transfer learning for classification of flowers," *2018 IEEE 3rd International Conference on Signal and Image Processing, ICSIP 2018*, pp. 562–566, Jan. 2019, doi: 10.1109/SIPROCESS.2018.8600536.
27. Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie, *Class-Balanced Loss Based on Effective Number of Samples*. 2019. doi: 10.1109/CVPR.2019.00949.