ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

IGP
Institut für Geodäsie und
Photogrammetrie

Yihang She

Using Machine Learning to Identify Plant Species Acquired by Citizen Scientists

Interdisciplinary Project Work (IPA)

Institute of Geodesy and Photogrammetry
Swiss Federal Institute of Technology (ETH) Zurich

Supervision

Prof. Dr. Konrad Schindler
Dr. Stefano D'Aronco
Riccardo de Lutio

in collaboration with Dr. Philipp Brun from WSL

January 2021

Preface

This is the Interdisciplinary Project Work (IPA) of the Geomatic Engineering M.Sc. program, and it is also the first research work I have done at ETH. With this opportunity, I would like to express my enormous gratitude to the following people:

- Prof. Dr. Konrad Schindler for advising me on the IPA topic and making it possible for me to work on the project.
- Dr. Stefano D'Aronco and Riccardo de Lutio for their generous support throughout the project and timely advice on a daily basis.
- Dr. Philipp Brun from Swiss Federal Institute for Forest, Snow and Landscape Research (WSL) for providing me with the dataset and feedbacks on the project.

Abstract

Citizen science is making great contributions to the large-scale collection of the image set for biodiversity studies. However, the quality of the data is often doubted, and labeling the images via human efforts is time-consuming. This motivated us to develop a machine-learning algorithm to help label these images. However, it is still a challenging task for machine learning given the imbalanced nature of the fine-grained categories in the collected data. While existing literature suggests that both location context and hierarchical labels can benefit the fine-grained classification, currently few kinds of research have been done to leverage both pieces of knowledge in a unified framework. Our study presented a design of a neural network that integrates both information for the first time. The results show that our model outperforms the state-of-the-art classifier that is solely based on the image in classifying the popular dataset from citizen science projects. Besides labeling the image, our experiments show that the model can also be used to map the plant distribution, as well as predict new species. We expect our model can promote not only the quality of the data collected by citizen scientists, but also its further applications.

Contents

Chapter 1

Introduction

1.1 Problem Statement

Citizen science is an approach to involve the public in scientific researches, especially those at large spatial or temporal scales [3]. Benefited from the development of computing technology and portable devices [10], the popularity of citizen science is fast growing in the field of biodiversity studies. For instance, with a smartphone, citizen scientist can easily take images of plant species and upload them to the online database from different corners of the world. On the one hand, the efforts of citizen scientists greatly contribute to the collection and analysis of biodiversity data [14], such as the occurrence and distribution of plant species. On the other hand, the process itself can benefit the public understanding of scientific knowledge and therefore advance the scientific literacy of citizens involved in the project [2].

The data collected by citizen scientists have been widely used in ecological studies and has led to plenty of publications including peer-reviewed articles [23]. However, crowdsourced data also confront challenges that cannot be ignored. One major issue is data accuracy (e.g. mislabeling of data) [10], which can be resulted from the following reasons: (1) **Variation of expertise**. The knowledge and skills of citizen scientists can be greatly influenced by the duration of training and experience with a certain task. In particular, for the labeling of rare species, citizen scientists are usually not as good as professional researchers. (2) **Variation of standards**. Citizen scientists from different regions might adopt different standards or methods for observation, which could lead to observation bias. These drawbacks undermine not only the quality of data collected by citizen scientists but also its further applications.

Where there is a challenge, there is an opportunity. Machine Learning (ML) as a fast-growing technology has been introduced to diverse fields given its power in pattern recognition [16]. For citizen scientists in biodiversity studies, is it possible to develop an ML algorithm to help them label the collected images? The memory capacity of the computer is especially feasible for processing large datasets such as the one collected by citizen scientists. Besides, the computing power of ML makes it possible to label the images with higher accuracy and efficiency. Nevertheless, immediate victory seems impossible, as the dataset collected by citizen scientists is usually highly imbalanced, and classifying such a dataset is still tricky for classical ML algorithms [21].

The following factors could make the collection of species imbalanced: (1) **Imbalanced occurrence** The occurrence of plant species is imbalanced, wherein some species appear much more frequently than the rare species. And the letter has less chance to be observed. (2) **Bias of sampling** Species in the area which is more accessible (e.g. area near the road or residential area) might be oversampled, while the species in the mountainous area might be undersampled [10]. Compared to the balanced dataset, the performance of many traditional ML algorithms (e.g. SVM, Neural Network) are reported to have poor performance on the imbalanced dataset for the following reasons [21, 12]: (1) The imbalanced distribution can make the features that could have been learned be ignored or unidentified. (2) Classes with a small sample size could be confused with data noise, and (3) the inter-class similarity could also confuse the algorithm when it tries to distinguish a small class from a big class with similar features. Besides, (4) the metrics used for training (e.g. validation accuracy) could incorrectly measure the model performance if not adapted for the imbalanced distribution.

So the major question we would like to ask is: how to develop an ML algorithm that can help citizen scientists identify the imbalanced collection of plant species?

1.2 Motivation

Existing literature has proposed different methods to address the fine-grained classification of the imbalanced dataset (see Chapter 3). What we proposed in the thesis is to make the most of the non-visual information collected by the citizen scientists, in combination with the visual information (i.e. images), to classify the imbalanced data. Actually, when looking into the datasets from citizen science projects, we found each image is usually associated with abundant non-visual information, some of which have the potential to significantly refine the classification. For instance, (1) **Location context**. Under permission, the app of citizen science projects usually also records the context when the image is taken, such as position and time. These data are uploaded to the online database together with the images, and are significant for the classification of plant species, as the occurrence of plant species is heterogeneous both spatially and temporally. (2) **Hierarchical labels**. The same plant can have more than one label according to plant taxonomy. The hierarchical levels from up to the bottom are Kingdom, Phylum, Class, Order, Family, Genus, Species [20]. The lower the level, the more similar the plants. Therefore, besides labeling the plant at Species-level, citizen science projects often label the image at higher levels. These hierarchical labels provide links between different plant species and could benefit the classification of samples with small size [19].

Therefore, the primary objective of our work is to (1) develop a ML algorithm that classifies the imbalanced collection of plant species by leveraging both visual and non-visual information. Besides, we would like to exploit the potential of our ML algorithm to (2) map the distribution of plant species and (3) label the new species, which will provide citizen science projects with additional benefits for biodiversity studies.

1.3 Thesis Structure

The thesis is organized as follows:

- Chapter 1 gives the background and associated scientific problems of the work, states the major objectives, and outlines the structure of the thesis.
- Chapter 2 will briefly review the existing literature related to the methods used in our work, including the modeling approaches based on location context as well as label hierarchy.
- Chapter 3 will present the methods we used to address the classification of the imbalanced datasets from citizen science projects, including strategies regarding both data and modeling.
- Chapter 4 will present the details of experiments, including data source, preprocessing, and training implementation.
- Chapter 5 will present the performance of our trained models and discuss the results of different experiments on these models.

- Chapter 6 will give concluding remarks, summarize the major contributions of our work, and present future outlooks.

Chapter 2

Related Work

In this chapter, we will discuss the researches works that are mostly related to our project in the field of fine-grained visual categorization (FGVC), where different classes usually share similar visual semantics and need to leverage more auxiliary information to help classification [25]. Specifically, we will present two approaches of modeling, which are based on context and hierarchy, respectively.

2.1 Context-based Modeling

Researchers have noticed that the location context is important for modeling the distribution of species, and therefore can especially benefit the fine-grained classification task, although up to now relevant studies are still limited [6, 24]. Berg *et al.* [1] noticed that the camera carried on smartphones usually records additional information regarding the location and time besides taking image, and in their studies, they estimated a probability for bird species by leveraging this information via kernel density estimation. With the fast development of deep neural networks, researchers also consider the effective approaches to combine the location context with CNN. Tang *et al.* [22] studied how to encode the coordinate information from GPS when the image was taken, and they concatenated the encoded features from location context with the output from the CNN in an early fusion manner before feeding it into the $softmax$ layer for classification. Wittich *et al.* [24] adopted a nearest neighbor approach to predict the possible species that a person could encounter at certain locations given the previously recorded observations nearby.

While the approaches proposed by [22, 24] both successfully applied the location context to refining the classification task, they have drawbacks in explaining how the location context works in terms of probability theory. Subsequent to the work of [22], Chu *et al.* evaluated three different approaches of integrating the location context into the deep neural network, including modeling the Baysian prior from geo-locations, concatenating the feature with the image features, as well as the post-processing classifier using the output from image classifier. In parallel to their works, Aodha *et al.* [17] proposed a geo-prior classifier for predicting the distribution of plant species, which integrated information of location, time, and photographer. The classifier for location context in this work was modeled in a Bayesian approach and was trained separately from the image classifier, which increases both the interpretability and flexibility of the model.

2.2 Hierarchy-based Modeling

Compared to location context with additional information regarding species distribution, label hierarchy benefits the classification more by sharing features among adjacent classes. Srivastava *et al.* [19] organized the different classes in a hierarchical structure and developed an approach to help the deep neural network to classify the classes by transferring features in related classes using the hierarchy prior. In the domain of clothing detection, Kumar *et al.* [15] detected the label hierarchy instead of a single label for the input object by analyzing detection errors, and they noticed that the trained model can detect the new category which was not used for training, given that the hierarchical labels contain information of different levels. Researchers also explored the different approaches to inject the knowledge of hierarchical labels in the neural network. Chen *et al.* [5] proposed a Hierarchical Semantic Framework to predict the category scores of each level in a top-down manner for the dataset of birds and the dataset of butterflies. Similarly, Dhall *et al.* [9] integrated the knowledge of semantic hierarchy into CNN by proposing a set of loss functions, wherein the marginalization loss summarizes the hierarchical information in a bottom-up manner.

2.3 Our Approach

As shown in the above sections, location context modeling is considered as an effective approach for the task of fine-grained classification, and as noted in [5], the hierarchical information widely exists in the fine-grained categories. In other words, where the location context can be applied to the fine-grained classification, hierarchical information usually also exists. However, to the best of our knowledge, up to now there are no research works that combine both information to the classification task of imbalanced data, especially in the domain of crowdsourced plant species. Therefore, it would be interesting to explore how to effectively combine these two approaches and evaluate its performance in the plant species data collected by citizen scientists, which will be discussed in details in the following chapters.

Chapter 3

Methods

In this chapter, the design of ML algorithm in our project will be presented. It can be considered in two dimensions. The first dimension is regarding the processing of data itself. The second dimension is regarding the model, where we leveraged the information of both location context and label hierarchy. We also give an overview of the workflow to illustrate how it works as a whole.

3.1 Data Aspect

3.1.1 Balanced Sampling

Re-sampling is considered as an effective strategy in the classification of imbalanced data [12]. Here we adopted the balanced sampling strategy, where the sampling weight of each image W_i is inversely proportional to the sample size N_i of the species it belongs to (Equation 3.1).

$$W_i = \frac{1}{N_i} \tag{3.1}$$

This strategy will oversample the species with small size and undersample the species with large size, which will help to mitigate the effects of imbalanced distribution on classification. However, it should be noted that the effects of imbalanced distribution cannot be completely eliminated even after balanced sampling, as the information contained in the small samples are still relatively limited compared to big samples. The dataset after balanced sampling, instead of the original one, will be loaded for training.

3.1.2 Balanced Test Set

Test accuracy is crucial for measuring the model performance and therefore has substantial effects on training. Test accuracy without considering the imbalanced distribution of data will overlook the model performance on species with small size [21]. To address this issue, here we created a balanced test set, where the samples of different species have the same size and are randomly drawn from the data collection. The test accuracy is calculated using this balanced set and therefore will not be influenced by the imbalanced distribution of the original dataset.

3.2 Model Aspect

For model prediction, we adopted a probablistic discriminative approach, where the output is a probability distribution over different classes conditioned on the input information. With this approach, we will be able to combine the predictions conditioned on image and location context, respectively. Furthermore, the combined prediction can be easily marginalized according to Bayesian probability theory, which makes it possible to integrate the information of plant taxonomy. The following subsections will present the details of our approach.

3.2.1 Inference from Image

To infer the class of plant species from the input image, we applied the convolutional neural network (CNN), which will output the probability distribution of input image I over different classes.

We applied a transfer learning strategy to infer from images, where the backbone model is ResNet50 [13], with weights pretrained on ImageNet [8]. ResNet is a deep CNN with residual blocks, the shortcut connection of which is proved effective in addressing the degradation problem in the training of deep neural networks, without increasing the number of parameters to train. Therefore, it is popular in image-related tasks, and the ResNet50 pretrained on ImageNet generalizes well when extracting the feature of images from different domains. The image classifier with the backbone of ResNet50 is also our baseline model, which will be compared with the other models in the following chapters. Equation 3.2 shows how we infer the probability $P(y|I)$ of class y given the image I. We used the $softmax$ function to calculate the probability of each class. By using this function, we assume that each image is assigned to an unique class.

$$P(y|I) \propto softmax(FCN_{\text{image}}(ResNet(I))) \tag{3.2}$$

where $ResNet$ is the CNN layers of ResNet50, FCN_{image} is the fully connected layer of the image classifier.

3.2.2 Inference from Location Context

The location context refers to the context information obtained by citizen science projects when the image was taken. The location context L applied here include both spatial and temporal aspects. Specifically, the spatial context is the 3D position where the image was taken, including longitude (x), latitude (y), and altitude (z); the temporal context is the day of the year (t) when the image was taken. As explained in Chapter 1, the location context benefits the classification of imbalanced dataset, especially for the collection of plant species, because the distribution and occurrence of plant species are highly heterogeneous in both spatial and temporal dimensions.

To infer the class of plant species from the location context, we adopted the geo-prior classifier proposed in [17] (Equation 3.3), where the location context L is firstly encoded into a D dimension space by fully connected networks FCN_{context}, and then embedded into a C dimension vector using the object embedding matrix $\mathbf{E} \in \mathcal{R}^{\mathrm{D} \times \mathrm{C}}$, and C is the number of plant species to classify. Each entry of the vector is an output of $sigmoid$ function ranging in $(0, 1)$, as a probability that each plant species can appear under this location context. Here the $sigmoid$ function instead of the $softmax$ is used because it is assumed that more than one species could occur under a certain a location context.

$$P(y|L) \propto sigmoid(FCN_{\text{context}}(L) \cdot \mathbf{E}) \tag{3.3}$$

3.2.3 Integration of Hierarchical Labels

Hierarchical labels from the plant taxonomy is another important non-visual information that we can use to help classify the imbalanced collection of plant species. As the labels associated with each single image links different species together according to their taxonomy similarity, the integration of hierarchical labels can assist the image classification with additional information which is not contained in visual similarity [5]. Besides, the hierarchical links can benefit the classification of species with small sample size by leveraging the features from adjacent classes [19]. Here we used the labels of all hierarchical levels except for the Kingdom level, as all the plant species belongs to the Plant Kingdom. The remaining hierarchical levels from the bottom to the up is Species, Genus, Family, Order, Class, and Phylum. The labels at Species level is what we would like to classify for our initial task.

To integrate hierarchical labels, we adopted the marginalization loss proposed in [9]. As displayed in Figure 3.1, the output from classifier is the probability distribution over different species, which will be marginalized in a bottom-up manner to derive the probability distributions over different classes for the corresponding level (Equation 3.4).

$$P_i^l = \sum_{j \in S_i} P_j^{l+1} \tag{3.4}$$

where P_i^l is the probability of class i at hierarchical level l, P_j^{l+1} is the probability of class j at the lower level $+1l$, and j belongs to the set S_i, which is the set of child classes of class i.

Based on the probability distribution P^l derived at level l, we calculated a loss L^l for each level using the *cross entropy loss*function, and then the losses from a total of n levels will be summed as the loss L used for back propagation (Equation 3.5).

$$L = \sum_{l=1}^{n} L^l \tag{3.5}$$

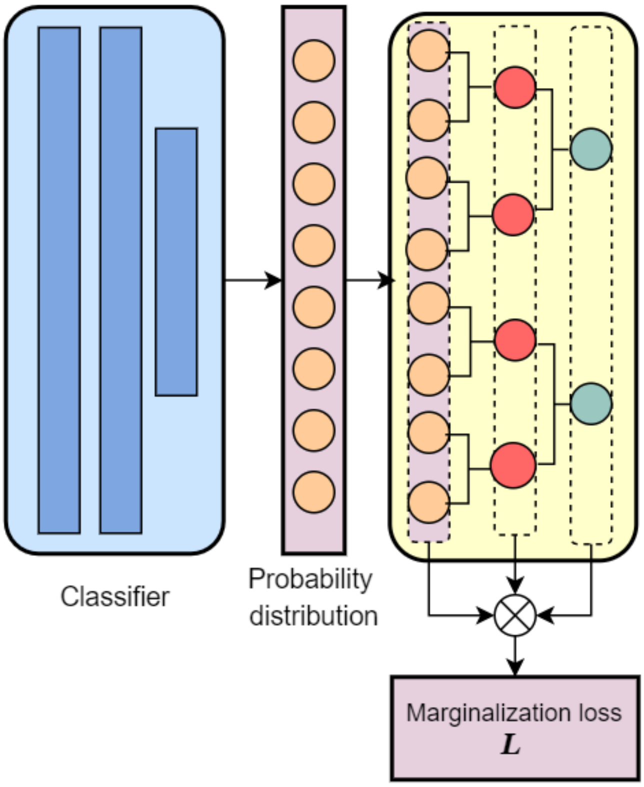

Figure 3.1: Diagram of marginalization loss

3.3 Overview of Workflow

As displayed in Figure 3.2, we first inferred the probability distributions of $P(y|I)$ and P(y|L), which is conditioned on I and L, respectively. Then we took the product of these two probabilities as the approximation of the probability $P(y|I,L)$(Equation 3.6) [17], which is jointly conditioned on both image and location context. By taking this approximation, the inference from location context can work as an independent prior and it can be easily combined with the output from image classifier, which not only simplifies the modeling, but also makes it more interpretable. But it should be noted that these two classifiers can be trained either jointly or separately, please see Chapter 4 for details of implementation. The combined inference $P(y|I) \cdot P(y|L)$ then works as the probability distribution that will be marginalized over hierarchical levels, as explained in Section 3.2.3.

$$P(y|I,L) \propto P(y|I) \cdot P(y|L) \tag{3.6}$$

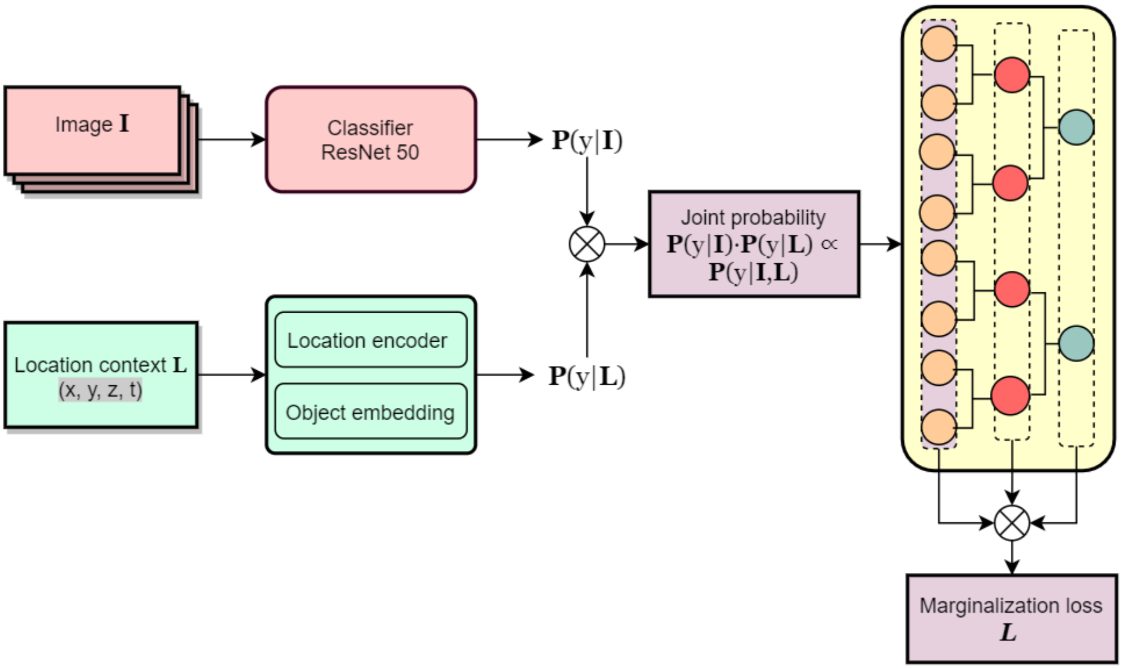

Figure 3.2: Diagram of the full model

Chapter 4

Experiments

4.1 Datasets

iNaturalist [18] is a citizen science project which collects crowdsourced observations over the globe via its online website and mobile applications. Besides the images, the dataset also contains the non-visual information collected by citizen scientists, including those that will be used in our model (i.e. longitude, latitude, day of the year, and hierarchical labels). For altitude of each image, we extract it from the Digital Elevation Model (DEM) data of Switzerland according to its longitude and latitude. Using the iNaturalist dataset can benefit the goals of our projects for the following reasons [23]: (1) As a collection from citizen science efforts, the dataset is highly imbalanced over different species, which features the challenges of data from real world. (2) As a reflection of real world data, some species of the dataset share visual similarities, which is challenging for classifiers that are solely based on visual inputs. (3) Labels associated with the images in the dataset are verified by several experienced annotators, which makes it a reliable data for training.

For our project, we downloaded the latest data collection of plant species in Switzerland from iNaturalist. Table 4.1 displayed the statistics of our dataset. A total of 60781 images and their associated information were downloaded, which contains 2374 species. As displayed in Figure 4.1, the sample size of different species in the dataset is highly imbalanced and therefore it has a long-tail distribution. Only the species with no less than 10 images were selected for subsequent training and test, in order to make sure the reliability of trained models. There are 977 species contained in the filtered dataset, and a dataset with unseen species will be generated from the remaining data for further experiments (please see Section 5.4 for details). For each species in the filtered dataset, 5 images and their associated information will be randomly selected to create the test set, which is therefore a balanced set and can be used to measure the performance of our trained models.

Table 4.1: Statistics of the dataset

Description	Statistics
Number of downloaded images	60781
Number of downloaded classes	2374
Number of filtered classes	977
Number of images in training set	51723
Number of images in test set	4885

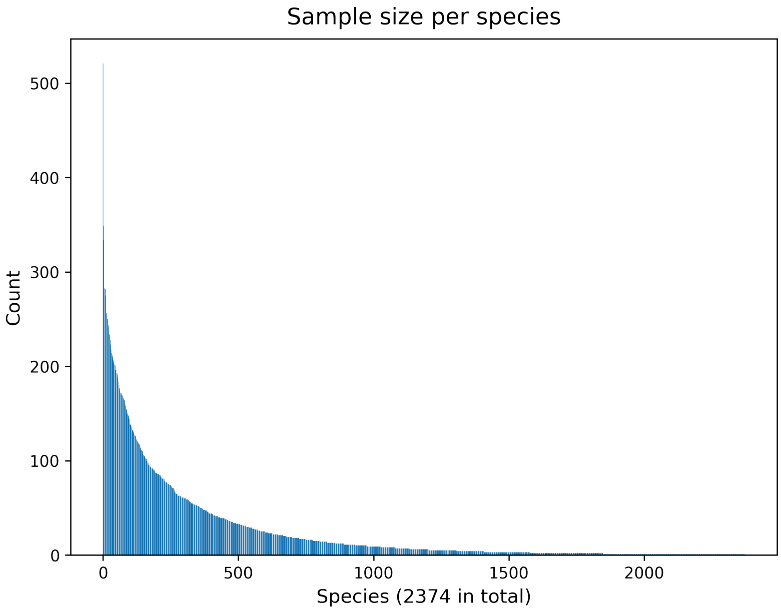

Figure 4.1: Sample size in downloaded dataset

4.2 Details of Implementation

4.2.1 Data preprocessing

All images were resized to the size of $256 * 256$ and then center cropped to $224 * 224$. The images used for training were additionally augmented by random rotation, random horizontal flip and color jitter, which will help mitigate the overfitting effects. Both training and test images were normalized.

The location context L for each image was prepared as the tuple of (x, y, z, t), namely longitude, latitude, altitude, and day of the year. All these four elements were normalized to the range of $[-1, 1]$. Additionally, t was embedded into (t_1, t_2) according to Equation 4.1 [17], which makes the beginning days and the end days of the year closed to each other in the embedding space, as the two time slots are actually similar in the sense of plant phenology.

$$\begin{cases} t_1 = \sin(\pi \cdot t) \\ t_2 = \cos(\pi \cdot t) \end{cases} \tag{4.1}$$

The hierarchical labels of each image were transformed from nominal scale to numeric scale by indexing the original labels orderly at each level.

4.2.2 Training

The balanced sampling strategy (see Section 3.1.1) was applied to the training data before it was loaded to the training process. Figure 4.2 displayed the sample size of each species before and after balanced sampling in the training set. The data were loaded in a batch size of 32, and all models displayed in the subsequent sections were trained for 25 epochs. Besides, for the image classifier , we initialized different learning rates for its pretrained convolutional layers of ResNet (smaller learning rate, 0.0005) and the fully connected layers (bigger learning rate, 0.002), which also helped improve the model performance. Stochastic Gradient Descent (SGD) [4] was used as the optimizer. Cross Entropy Loss was used to calculate the loss of the training process, if the hierarchical labels and its associated marginalization loss were not used.

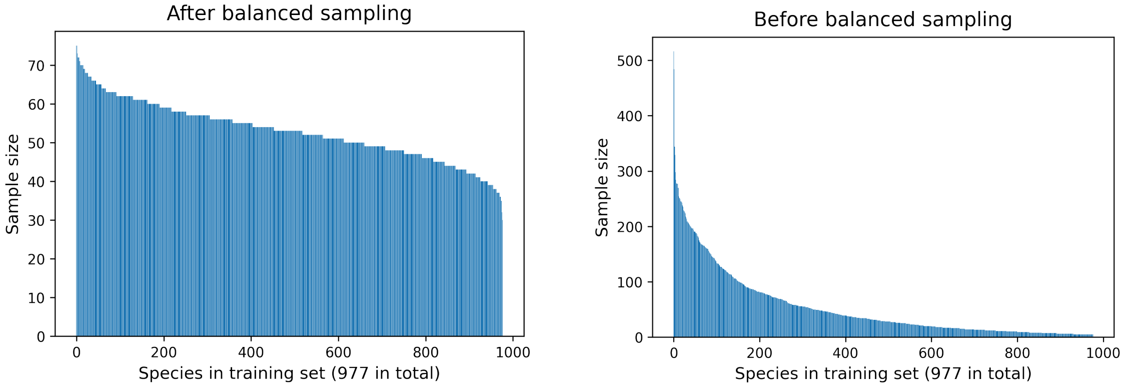

Figure 4.2: Sample size before and after balanced sampling

The following models were trained for the comparison of different assumptions: (1) **Baseline**, namely the image classifier with the backbone of ResNet50. (2) **Baseline + Location Context**, namely the baseline trained with location context. (3) **Baseline + Hierarchical Labels**, namely the baseline model trained with the marginalization loss, and (4) **Full Model**, namely the baseline in combination with both location context and hierarchical labels. The different parts of the aforementioned models were grouped together according to the principles explained in Section 3.2.

Besides, the location context classifier was trained separately to evaluate the performance of the classifier itself under different training manners (see Section 5.3). Except for where it is noted, all the remaining location context classifiers presented in the work were trained jointly with the image classifier.

Chapter 5

Results and Discussion

5.1 Model Performance

Table 5.1 displays the test accuracy of the four trained models. The top-1 accuracy of the baseline model is 61.27%, and integrating either the location context or hierarchical labels will significantly improve the model performance, where their top-1 accuracy are 65.71% and 63.79%, respectively. In particular, the full model, which integrates both non-visual information, achieves the highest accuracy (68.09%). Besides, the top-3 and the top-5 accuracy of the full model (82.78% and 87.47%) are significantly higher than its top-1 accuracy. On the one hand, this implies that there are still species, the visual similarity of which confuses our much in determining their categories. This can be refined further in the future works. On the other hand, the top-n accuracy (n>1) could be used as the auxiliary information for citizen scientists in the real-world applications to help them narrow down the labeling options.

Table 5.1: Test accuracy of trained models

Model	Top-1 (%)	Top-3 (%)	Top-5 (%)
Baseline	61.27	76.66	81.65
Baseline + Location context	65.71	80.14	84.75
Baseline + Hierarchical labels	63.79	80.08	84.99
Full model	**68.09**	**82.78**	**87.47**

These basic results clearly indicate that the design of our model outperforms the state-of-the-art ResNet classifier that is solely based on visual inputs. It also suggests that the non-visual prior of location context and taxonomy can benefit the classification of imbalanced collection of plant species. However, it should be noted that these two priors actually benefit the classification from two different ways: while the location context provide additional information that is correlated to the occurrence of plant species, the hierarchical labels link the species at different levels so that the species with small size can share additional information from adjacent samples.

5.2 Test Accuracy at Different Hierarchical Levels

Trained with hierarchical labels, the model is actually able to predict the classes of different hierarchical levels besides the bottom level of Species. Therefore, we tested our full model on the test set on each hierarchical level (Table 5.3), and Table 5.2 displays the number of classes at each level contained in the test set. Note that for Phylum level, only the top-1 accuracy is displayed as this level only contains 3 classes. In particular for the levels of Class and Phylum, their test accuracy (95.58% and 99.73%) are significantly higher than the remaining levels, but the number of classes at these two levels are much smaller (8 and 3). The test accuracy increases accordingly with the increase of hierarchical levels. This could be related to decrease of the number of classes at higher levels, because the smaller the number of classes, the bigger the sample size of each class, and also the less imbalanced distribution among different classes, which will make the learning easier.

Table 5.2: Number of classes at each level

Level	Species	Genus	Family	Order	Class	Phylum
Number	977	489	121	50	8	3

Table 5.3: Test accuracy on each level

Level	Species	Genus	Family	Order	Class	Phylum
Top-1 accuracy (%)	**68.09**	**74.90**	**80.86**	**82.68**	**95.58**	**99.73**
Top-2 accuracy (%)	82.78	86.73	91.89	93.59	99.86	–
Top-3 accuracy (%)	87.47	90.35	94.76	96.42	99.98	–

It also needs to be noted that random chance exists in the accuracy at each level, although it can be claimed that the accuracy increases with higher levels. To more explicitly evaluate the model performance at each level, future works can be done to train a separate classifiers for each level using only the labels at corresponding level, and then compare its performance with the model presented here.

5.3 Prediction Score of Different Geo-prior Classifiers

To evaluate the location context classifier at different training manners, we mapped the predictions of these classifiers for four species at the spatial extent of Switzerland, and compared it with the reference data. To inspect the prediction scores in details, we grouped the predictions in two figures: Figure 5.1 displays the species with only a few training samples (52 and 11, respectively), and Figure 5.2 displays the species with more samples (163 and 216, respectively). The plots of the first three columns from the left to the right display the predictions from location context classifiers that were trained separately, jointly, and jointly with hierarchical labels. The last column displays the probability distribution of the same species produced by experts. All the predictions were probability scores ranging in $(0, 1)$.

The prediction scores over Switzerland from location context classifiers display a clear spatial pattern, indicating that the the distribution of plant species is indeed spatially heterogeneous. The results also show that the combination of the inferences from image and location context is an interpretable prediction as the product of prior probability from location context and the probability from image. However, if we compare the prediction scores of these classifiers, there are still prominent differences: (1) The predictions of the separately trained classifier are more similar to the reference data in spatial patterns, which means that the the output of separately trained classifier is more accurate in the sense of mapping species distribution, while the jointly trained one is more like supplementary information to the images. (2) Compared to the jointly trained classifier without using hierarchical labels, the prediction scores of the jointly trained one with hierarchical labels tend to have stronger contrast, which suggests that the integration of hierarchical labels make the location context more 'confident' in its predictions.

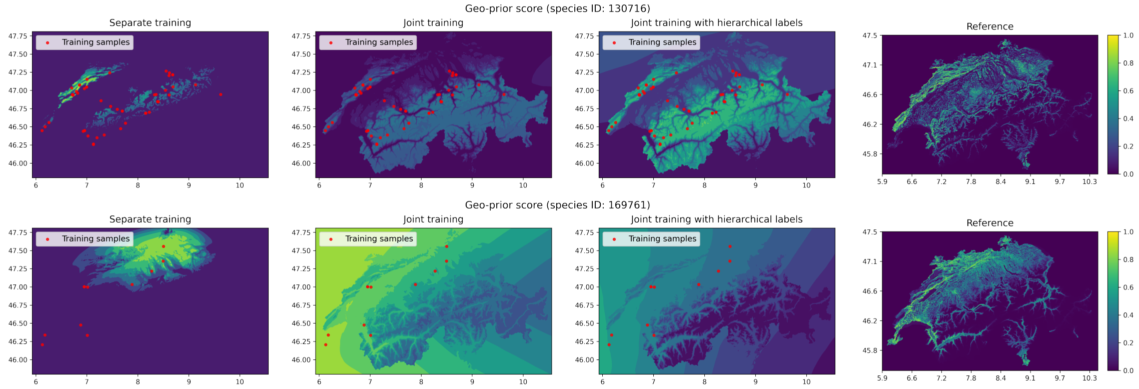
Geo-prior score (species ID: 130716)
Separate training
Joint training
Joint training with hierarchical labels
Reference
Training samples
Geo-prior score (species ID: 169761)
Separate training
Joint training
Joint training with hierarchical labels
Reference
Training samples

Figure 5.1: Prediction score of species with less samples

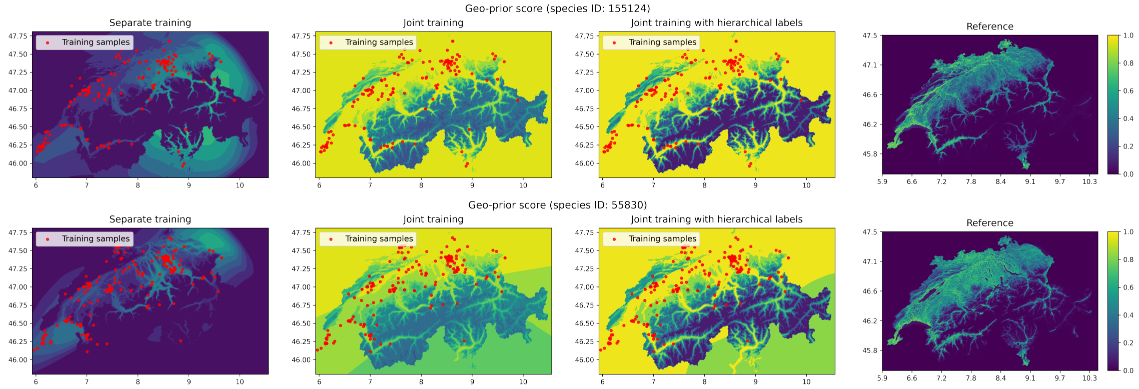
Geo-prior score (species ID: 155124)
Separate training
Joint training
Joint training with hierarchical labels
Reference
Training samples
Geo-prior score (species ID: 55830)
Separate training
Joint training
Joint training with hierarchical labels
Reference
Training samples

Figure 5.2: Prediction score of species with more samples

5.4 Experiments on New Species

Given the hierarchical labels, it is possible to classify the new species at higher hierarchical levels using the full model, even though its output does not have any common classes with these species at the bottom level. To evaluate our model in this regard, we created a dataset with new species. The dataset was generated from the species in the original dataset with less than 10 images, so that none of these species was used for training or testing in the sections before. This dataset contains 330 new species, and each species contains 5 randomly selected species, so that the unseen set is also balanced at the Species level. Table 5.4 displays the comparison of the accuracy of our full model on the test set and the unseen set with new species. It can be observed that even though no species was used to train the full model, our model is able to learns something in higher hierarchical levels. In particular at the level of Class and Phylum, the accuracy of the model on the new species is high. This shows the model has the potential to help classify the species at higher levels, no matter whether or not the species is used for training.

Table 5.4: Top-1 accuracy on test set and unseen set with new species

Level	Species	Genus	Family	Order	Class	Phylum
Test set (%)	68.09	74.90	80.86	82.68	95.58	99.73
Unseen set (%)	–	33.74	44.92	50.47	86.38	97.21

Chapter 6

Conclusion

6.1 Summary

Our work demonstrates that (1) the integration of non-visual information can refine the classification of imbalanced dataset collected by citizen scientists. Specifically, (2) the location context can not only refine image classification by providing prior knowledge regarding the distribution and occurrence of plant species, but also generate a classifier itself that can map the distribution of plant species. Similarly, (3) the integration of hierarchical labels not only helps the classification of small samples by leveraging the information from adjacent species, but also makes the trained model able to predict labels at higher levels, even on the new species that were not used for training. (4) By leveraging both information in an unified framework, we designed and trained a model which is not only interpretable, but also flexible. For instance, the image classifier, as part of the model, can be used separately once it has been trained, if the location context of the image is not available. Meanwhile, the information of hierarchical labels has been injected into our model via the marginalization loss during training, which will not be required when the model makes predictions.

We expect that the equipment of our ML algorithm on the citizen science project can help improve its data quality and promote its further applications in the future. We also expect that our research framework can promote the state-of-the-art of fine-grained classification of imbalanced dataset.

6.2 Outlook

For future researches, we think the following works can be tried to improve the model further:

- **Online learning**. Online learning is a method to update the trained model using real-time data ([11]). Currently we trained our model on the offline dataset, and it remains unchanged once the training is done. However, the database of citizen science projects is actually updated everyday by collecting the observations of citizen scientists from all over the world, which means that our model could be unable to correctly label the species once the distribution of plant species in the database is changed. With online learning, the trained model will be updated in real-time using the latest collected samples, which will make our model more applicable.

- **Training classifiers for each hierarchical level**. As discussed in Section 5.2, there is random chance in the performance of our trained model on the test accuracy of the higher hierarchical levels: although it is expected that the test accuracy will increase accordingly with the increase of the hierarchical levels, how good it is in making predictions at the corresponding level was still analyzed in a qualitative sense. In the future, we consider to train models for the labels of each level separately in order to provide a quantitative benchmark, which can be used to better evaluate our models.

- **Training strategy**. Currently our models were trained on the given dataset directly with balanced sampling. [7] proposed another strategy that can be applied in the stage of training to address the imbalanced distribution, where the model is first trained on the original dataset to learn sufficient feature representations, and then it will be fine tuned on a balanced subset with small learning rate in order to balance the model over different classes.

List of Figures

List of Tables

Bibliography

[1] Thomas Berg, Jiongxin Liu, Seung Woo Lee, Michelle L Alexander, David W Jacobs, and Peter N Belhumeur. Birdsnap: Large-scale fine-grained visual categorization of birds. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2011–2018, 2014.

[2] Rick Bonney, Caren B Cooper, Janis Dickinson, Steve Kelling, Tina Phillips, Kenneth V Rosenberg, and Jennifer Shirk. Citizen science: a developing tool for expanding science knowledge and scientific literacy. *BioScience*, 59(11):977–984, 2009.

[3] Rick Bonney, Jennifer L Shirk, Tina B Phillips, Andrea Wiggins, Heidi L Ballard, Abraham J Miller-Rushing, and Julia K Parrish. Next steps for citizen science. *Science*, 343(6178):1436–1437, 2014.

[4] Léon Bottou. Stochastic gradient descent tricks. In *Neural networks: Tricks of the trade*, pages 421–436. Springer, 2012.

[5] Tianshui Chen, Wenxi Wu, Yuefang Gao, Le Dong, Xiaonan Luo, and Liang Lin. Fine-grained representation learning and recognition by exploiting hierarchical semantic embedding. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 2023–2031, 2018.

[6] Grace Chu, Brian Potetz, Weijun Wang, Andrew Howard, Yang Song, Fernando Brucher, Thomas Leung, and Hartwig Adam. Geo-aware networks for fine-grained recognition. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[7] Yin Cui, Yang Song, Chen Sun, Andrew Howard, and Serge Belongie. Large scale fine-grained categorization and domain-specific transfer learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4109–4118, 2018.

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[9] Ankit Dhall, Anastasia Makarova, Octavian Ganea, Dario Pavllo, Michael Greeff, and Andreas Krause. Hierarchical image classification using entailment cone embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 836–837, 2020.

[10] Janis L Dickinson, Benjamin Zuckerberg, and David N Bonter. Citizen science as an ecological research tool: challenges and benefits. *Annual review of ecology, evolution, and systematics*, 41:149–172, 2010.

[11] Óscar Fontenla-Romero, Bertha Guijarro-Berdiñas, David Martinez-Rego, Beatriz Pérez-Sánchez, and Diego Peteiro-Barral. Online machine learning. In *Efficiency and Scalability Methods for Computational Intellect*, pages 27–54. IGI Global, 2013.

[12] Guo Haixiang, Li Yijing, Jennifer Shang, Gu Mingyun, Huang Yuanyue, and Gong Bing. Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications*, 73:220–239, 2017.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[14] Justin Kitzes and Lauren Schricker. The necessity, promise and challenge of automated biodiversity surveys. *Environmental Conservation*, 46(4):247–250, 2019.

[15] Suren Kumar and Rui Zheng. Hierarchical category detector for clothing recognition from visual data. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 2306–2312, 2017.

[16] Zelin Liu, Changhui Peng, Timothy Work, Jean-Noel Candau, Annie DesRochers, and Daniel Kneeshaw. Application of machine-learning methods in forest ecology: recent progress and future challenges. *Environmental Reviews*, 26(4):339–350, 2018.

[17] Oisin Mac Aodha, Elijah Cole, and Pietro Perona. Presence-only geographical priors for fine-grained image classification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9596–9606, 2019.

[18] Jill Nugent. inaturalist: citizen science for 21st-century naturalists. *Science Scope*, 41(7):12, 2018.

[19] Nitish Srivastava and Russ R Salakhutdinov. Discriminative transfer learning with tree-based priors. *Advances in neural information processing systems*, 26:2094–2102, 2013.

[20] Clive A Stace. *Plant taxonomy and biosystematics.* Cambridge University Press, 1991.

[21] Yanmin Sun, Andrew KC Wong, and Mohamed S Kamel. Classification of imbalanced data: A review. *International journal of pattern recognition and artificial intelligence*, 23(04):687–719, 2009.

[22] Kevin Tang, Manohar Paluri, Li Fei-Fei, Rob Fergus, and Lubomir Bourdev. Improving image classification with location context. In *Proceedings of the IEEE international conference on computer vision*, pages 1008–1016, 2015.

[23] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8769–8778, 2018.

[24] Hans Christian Wittich, Marco Seeland, Jana Wäldchen, Michael Rzanny, and Patrick Mäder. Recommending plant taxa for supporting on-site species identification. *BMC bioinformatics*, 19(1):190, 2018.

[25] Lingxi Xie, Qi Tian, Richang Hong, Shuicheng Yan, and Bo Zhang. Hierarchical part matching for fine-grained visual categorization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1641–1648, 2013.

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Chair of Photogrammetry and Remote Sensing
Institute of Geodesy and Photogrammetry
Prof. Dr. K. Schindler

Title of work:

Using Machine Learning to Identify Plant Species Acquired by Citizen Scientists

Thesis type and date:

Interdisciplinary Project Work (IPA), January 2021

Supervision:

Prof. Dr. Konrad Schindler
Dr. Stefano D'Aronco
Riccardo de Lutio

in collaboration with Dr. Philipp Brun from WSL

Student:

Name:	Yihang She
E-mail:	yihshe@student.ethz.ch
Legi-Nr.:	19-942-762
Semester:	Autumn Semester 2020

Statement regarding plagiarism:

By signing this statement, I affirm that I have read and signed the Declaration of Originality, independently produced this paper, and adhered to the general practice of source citation in this subject-area.

Declaration of Originality:

`http://www.ethz.ch/faculty/exams/plagiarism/confirmation_en.pdf`

Zurich, 20. 1. 2021: Yihang She