480 Final Project

Amy Lai
School of Computer Science
University of Waterloo
Waterloo, ON, N2L 3G1
s29lai@uwaterloo.ca
report due: August 12

Abstract

Plant traits are essential for accessing biodiversity and ecosystem dynamics, but they are time-consuming and challenging to measure at a large scale. To enhance the efficiency of trait measurements, we explore the possibility of predicting plant traits through citizen science plant photographs from the iNaturalist database, and their corresponding ancillary information including climate, soil, and multitemporal satellite variables derived from global raster data based on geocoordinates of each image. The traits observed for each plant are sourced from the TRY database. Through a combination of the DINOv2 Vision Transformer and LightGBM algorithm, we can achieve a 0.45924 R^2 score on the public test set. The code can be found at link. This project demonstrates the potential of combining big data from professional and citizen science sources with ViT models like DINOv2, presenting a powerful method for efficient and automated assessment of global plant traits.

1 Introduction

Changes in a global environment are essential to human well-being. Ecosystem functioning is one representation of the global environment and it can be assessed using the trait composition of plant communities. Plant traits such as plant size, seed dry mass, and nitrogen concentration may indicate the health and dynamics of ecosystems. Such traits, however, require high measurement effort. Thus, an effective trait measurement method would significantly enhance the rapid monitoring of ecosystems.

Vision Transformers (ViT) is a popular technique used in computer vision. It is target-oriented and can effectively learn and extract features from photographs. These features may be explicit (ex. color) or implicit (ex. nutrient deficiencies) as the ViT learns to discover the correlation between the input image and the targeted outputs. In particular, ViT segments images into patches and processes these patches as sequences. ViTs utilize an attention mechanism to weigh the importance of each patch relative to others, allowing the model to focus more on globally relevant features throughout the image simultaneously.

DINOv2 is an advanced vision transformer (ViT) that is pretrained without labels on a massive dataset of 142 million images. The model processes images by dividing them into uniform patches, which are then linearly embedded. DINOv2 offers several configurations with varying capacities: the small (S) version has an embedding dimension of 382, the base (B) version 768, the large (L) version 1024, and the giant (g) version 1536.

In this project, we will use a giant DINOv2 to extract image features, and then integrate them with ancillary data to form a large feature vector for each image. Regression models are then applied to make the final prediction for each plant trait.

Submitted to CS480/680, 2024 Spring. Do not distribute.

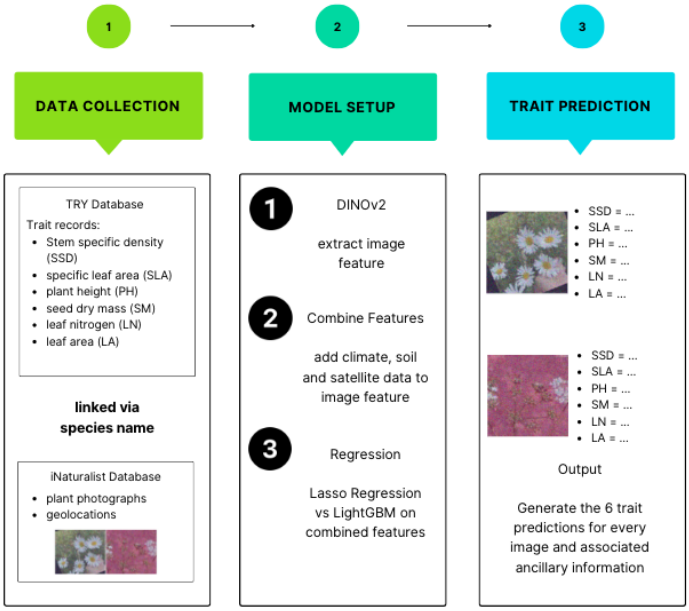

Figure 1: Conceptual diagram of the analysis, from data collection to evaluation. Data collection includes linking plant photographs from the iNaturalist database and plant trait records from the TRY database via species names. The model setup involves extracting image features using DINOv2, combining the features with ancillary data, and applying Lasso regression/LightGBM to derive the trait prediction.

2 Related Works

Vision Transformer has previously been applied to plant multi-label classification. In the paper "Multi-Label Plant Species Classification with Self-Supervised Vision Transformers", Gustineli applied the DINOv2 to classify multi-label plant species based on plant images. Before feeding data into DINOv2, Gustineli employed data augmentation techniques including cropping and resizing images to 128x128. In the study, the DINOv2 base model is selected to maintain a balance of computational efficiency and quality of feature extraction. Gustineli compared the performance of using the pre-trained model and a fine-tuned model, where the fine-tuned model starts as the pre-trained base model but undergoes additional training on the PlantCLEF dataset. Their findings show that the fine-tuned DINOv2 consistently outperforms the base model in terms of F1 scores averaged per species. To reduce dimensionality and mitigate overfitting, a Discrete Cosine Transform was applied to the extracted image features prior to the final classification step.

For our purpose of achieving the best prediction, we choose the most complex DINOv2 architecture, namely ViT-g/14, to extract 1536 image features. However, due to computational limitations, we did not further fine-tune the base model. To improve on Gustineli's study, additional data augmentations such as random horizontal/vertical flips, adding 0.01 Gaussian noise, and normalization to prevent overfitting are considered. Further, the plant images are resized to 224x224 as suggested in the "DINOv2: Learning Robust Visual Features without Supervision" paper. To prevent overfitting, we compare regression methods involving regularization such as Lasso Regression and LightGBM.

3 Main Results

Our goal is to maximize the $R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$ metric on the regression task. In other words, we want to maximize the fraction of total variance explained by the regression model. To perform regression, we first need a method to interpret the information given by the image as numerical vectors. In Gustineli's study, he states that DINOv2 "significantly outperforms the traditional feature extraction methods [he] tested earlier in the project". Hence, our exploration begins with extracting

image features using DINOv2 ViT-g/14. Due to computational constraints, we directly use the pre-trained model for feature extraction, obtaining 1536 features per image. These features, alongside ancillary data linked through plant IDs, are then input into our regression models—Lasso Regression and LightGBM. A summary of the model setup is given in the pseudocode below.

Algorithm 1: Model Setup

Input: ImageDataset, AncillaryDataset, targets

imageFeature = ViT-g/14(ImageDataset)
allFeatures = imageFeature join AncillaryDataset on id
lassoPrediction = []*6
for *idx, target in enumerate(targets)* **do**
    alpha = lassoCV(allFeatures, alpha=[0.0001, 0.001, 0.01, 0.1, 1, 2, 5], metric=R^2)
    lassoPrediction[] = lassoRegression(allFeatures, alpha=alpha)
lightgbmPrediction = LightGBM(allFeatures, learningRate $\in$ [0.01, 0.05, 0.1], numLeaves $\in$ [30,60], metric = MSE)

To select an appropriate regression model, the quality of our features such as multicollinearity need to be considered. Given the large number of features extracted and the fact that the ancillary features originate from just four sources, there is a significant likelihood of multicollinearity. The heatmap displayed below illustrates an example of this issue, showing high correlations among soil features.

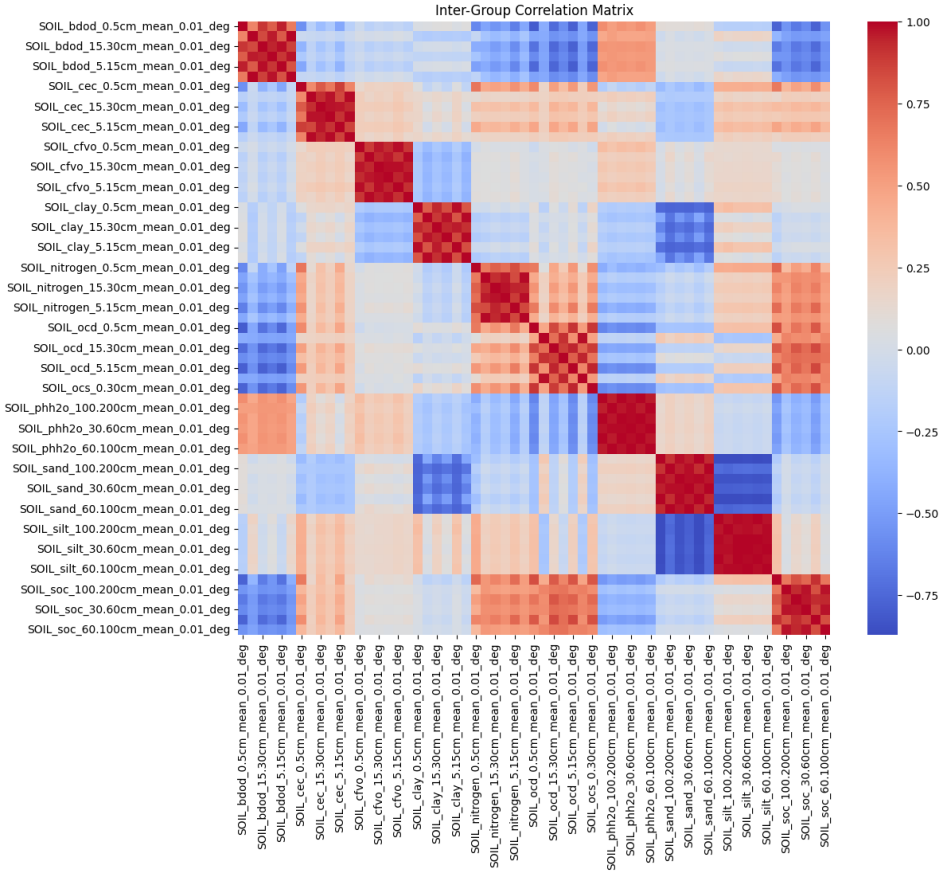

Figure 2: Heatmap used to access correlation between soil features

As seen through the soil features, multiple features exhibit high positive correlation (ex. features starting with SOIL_phh2o) and high negative correlation (ex. features starting with SOIL_silt and SOIL_sand). We consider Lasso Regression and LightGBM as both methods incorporate regularization techniques that are effective in managing multicollinearity.

Lasso Regression is a regression model that applies a penalty equivalent to the absolute value of the magnitude of coefficients to the loss function of the model. The loss function for lasso regression is $L_{\text{lasso}}(\hat{\beta}) = \sum_{i=1}^{n}(y_i - x_i^T\hat{\beta})^2 + \alpha\sum_{j=1}^{m}|\hat{\beta}_j|$. where m is the number of features and n is the number of plants in the dataset. The penalty term $\alpha\sum_{j=1}^{m}|\hat{\beta}_j|$ effectively constraints the size of the coefficients. The larger the penalty α, the bigger the amount of shrinkage. A characteristic of Lasso regression is that coefficients may shrink to 0, which effectively reduces the number of features in the model. To choose the penalty α, we use the 5-fold cross-validation method where the metric is R^2, the same as our objective metric. For each target variable, we perform Lasso Regression across a range of α values to find the most effective level of regularization.

LightGBM is a tree-based method that is robust to multicollinearity. Among all tree-based methods, it is the most computationally efficient. The core principle of LightGBM involves building a series of decision trees, where each tree learns from the mistakes of the previous ones. Every split in every tree chooses the feature that provides the most significant information gain. Although some features are highly correlated, the model will typically choose the one that offers the best-split point at a particular node, making it less sensitive to the presence of multicollinearity. For this project, we used Mean Square Error (MSE) as the function to calculate loss. Note that MSE = $\frac{1}{n}\sum_i(y_i - \hat{y}_i)^2$. Minimizing MSE is minimizing $\sum_i(y_i - \hat{y}_i)^2$, which is also maximizing R^2. To tune LightGBM, we tried learning rates $\in$ [0.01, 0.05, 0.1] and a number of leaves $\in$ [30, 60, 100].

The Lasso Regression and LightGBM performance over 3 iterations are summarized in Table 1

Table 1: R^2 of Public Test Set For Each Model Across Iterations

Model	Iteration1	Iteration2	Iteration3
Lasso Regression	0.41562	0.41277	0.41237
LightGBM	0.45411	0.45924	0.45785

Based on the three iterations, LightGBM out performs Lasso Regression as the R^2 metric of the public test set for the LightGBM model is consistently larger than the R^2 for Lasso Regression. The R^2 for LightGBM is around 0.457 ± 0.003 whereas the R^2 for Lasso Regression is around 0.412 ± 0.003.

4 Conclusion

Based on the result shown in Table 1, we observe that the R^2 metric of the public test set exceeds 0.4 in all iterations despite the regression method used. This indicates that the image of the plant has a strong indication of the plant traits. Moreover, the ViT-g/14 is powerful in extracting the ideal features from the plant images. Between LightGBM and Lasso Regression, we observe that LightGBM performs better throughout the iterations. This could be caused by the fact that Lasso regression assumes linear relationships between the features and the targets, whereas tree-based methods like LightGBM do not make those assumptions. The best R^2 score achieved in these 3 iterations is 0.459. This indicates that approximately 45.9% of the total variation in the public test set is explained by the combination of DINOv2 and the LightGBM model. Future works could enhance the current methodology by fine-tuning the DINOv2 model with the training images, expanding ranges for hyperparameters, and exploring alternative regression methods. However, these enhancements will need more computational time and resources.

Acknowledgement

Thanks to Murilo Gustineli for his inspiration through his "Multi-Label Plant Species Classification with Self-Supervised Vision Transformers" paper.

References

Gustineli, M., A. Miyaguchi, and I. Stalter (2024). "Multi-Label Plant Species Classification with Self-Supervised Vision Transformers". In: *Proceedings of the Conference and Labs of the Evaluation Forum (CLEF)*. Grenoble, France.

Oquab, M., T. Darcet, T. Moutakanni, H. V. Vo, et al. (2024). "DINOv2: Learning Robust Visual Features without Supervision". *Transactions on Machine Learning Research*, vol. 2024, no. 1.

Schiller, C., S. Schmidtlein, C. Boonman, A. Moreno-Martínez, and T. Kattenborn (2021). "Deep learning and citizen science enable automated plant trait predictions from photographs". *Scientific Reports*, vol. 11, p. 16395.