Machine Learning
Nature's Quest: Deep Learning Exploration of Global Plant Traits from Images and Geodata

Kai E. Niermann Dávid Miklo Trix Taicet Red Kaláb
Conner Dassen

March 29, 2024

Abstract

1 Introduction

Global warming and more broadly climate change has become a major concern for the world as its effects are becoming more and more apparent [7]. Changing weather patterns, especially towards more extreme conditions are causing plants to adapt to new environments [3]. One method of measuring said adaptation is to look at the traits, that is, properties of a plant that describe how it functions and interacts with the environment. These traits include but are not limited to of the plant height, leaf area, but also dry mass, leaf nitrogen content amongst various others. Monitoring these traits allows us to gain vital insights into how climate change impacts different ecosystems. While somewhat simple manual measurement techniques exist, at scale they are not feasible. This is where Convolutional Neural Networks (CNN) come in.

Through the work demonstrated by Schiller et al [6] we know that CNNs can be used to predict plant traits from images. The images used to train this network came from *citizen science photographs* which are images taken by citizens of plants from all across the world using AI plant species identification apps (e.g. iNaturalist, Pl@ntNet). Citizen science photographs also come with location metadata, which can be used to extract ancillary geodata such as precipitation, temperature, and soil type. This geodata can optionally be combined with the images to create a CNN which can potentially learn to extract features from images in conjunction with geodata to predict plant traits.

For our method we wanted to compare the accuracy of a CNN trained on images alone and using a pretrained backbone (e.g. ResNet, EfficientNet, etc.) to a CNN trained on images and geodata. We hypothesize that the CNN trained on images and geodata will outperform the CNN trained on images alone. Although the geodata could potentially lead the model to learning features not necessarily helpful for accurate prediction of plant traits due to a multitiude of reasons.

2 Method

2.1 Data processing

Integral to any machine learning model is the data used to train, validate and ultimate test the model. The data used consisted of 3 main components: ancillary geodata from various sources, the main task trait means and auxillary task standard deviations, and the training images of the plants.

Upon visual inspection one of the first issues we spotted was a considerable chunk (29.53%) of missing data missing for the auxillary task standard deviations. Through some previous work on the same Kaggle we noticed that the auxillary data was a useful inclusion. We can also validate these reports with previous work applying auxillary tasks in CNNs such as the work of Lukas Libel and Marco Kroener which confirmed that the inclusion of an auxillary task with minor relevance to the main task did indeed boost performance [5] or Chen et. al. [2] which demonstrated that for very similar or mostly similar auxillary tasks the contribution is generally a positive one. Based on this we decided to use a simple k-nearest neighbor (kNN) imputation strategy. The rest of the data had all values present.

The geodata was another point of consideration in the data pre-processing phase. Looking at the individual columns we noticed that the instances had features which encoded similar information and thus where likely redundant in the training process. Since we are working with a very high dimensional feature space to get some sort of visual verification of this claim we plotted a correlation heatmap for the 3 biggest groups of the geodata, namely the datasets: soil (soil information), MODIS, and VOD.

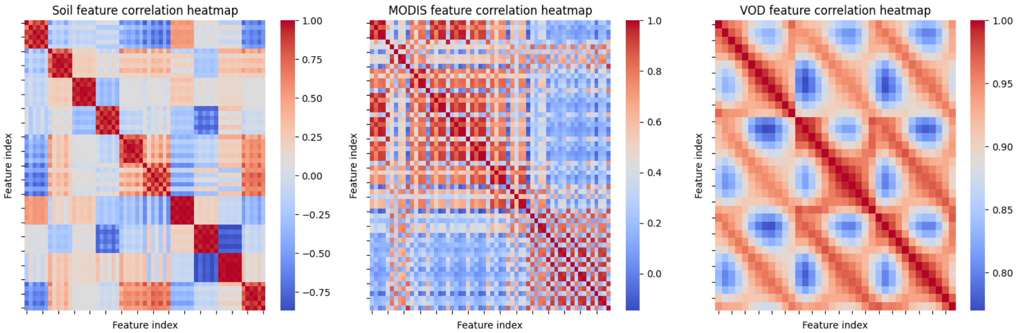

Figure 1: Correlation heatmap of the 3 biggest groups of geodata

Since these datasets individually can be broken down into groups of features generally talking about a similar topic we can see that the correlation within these groups is quite high. Even though previous work [4] has demonstrated the viability of neural network models to learn from high dimensional data, removing redundant features has in numerous instances been shown to improve model performance [1]. With this in mind we decided to apply a Principal Component Analysis (PCA) to the geodata to reduce the dimensionality of the data. We specified that the number of principle components should be such that 95% of the variance is explained.

2.2 Data preparation

2.3 Model design

2.4 Model evaluation

3 Results

4 Discussion

5 Conclusion

References

[1] J.-A. Chen, W. Niu, B. Ren, Y. Wang, and X. Shen, *Survey: Exploiting data redundancy for optimization of deep learning*, 2022. arXiv: 2208.13363 [cs.LG].

[2] Z. Chen, V. Badrinarayanan, C.-Y. Lee, and A. Rabinovich, "GradNorm: Gradient normalization for adaptive loss balancing in deep multitask networks," in *Proceedings of the 35th International Conference on Machine Learning*, J. Dy and A. Krause, Eds., ser. Proceedings of Machine Learning Research, vol. 80, PMLR, Oct. 2018, pp. 794–803. [Online]. Available: https://proceedings.mlr.press/v80/chen18a.html.

[3] S. B. Gray and S. M. Brady, "Plant developmental responses to climate change," *Developmental Biology*, vol. 419, no. 1, pp. 64–77, 2016, Plant Development, ISSN: 0012-1606. DOI: https://doi.org/10.1016/j.ydbio.2016.07.023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0012160616302640.

[4] I. Kansizoglou, L. Bampis, and A. Gasteratos, "Deep feature space: A geometrical perspective," *CoRR*, vol. abs/2007.00062, 2020. arXiv: 2007.00062. [Online]. Available: https://arxiv.org/abs/2007.00062.

[5] L. Liebel and M. Körner, "Auxiliary tasks in multi-task learning," *CoRR*, vol. abs/1805.06334, 2018. arXiv: 1805.06334. [Online]. Available: http://arxiv.org/abs/1805.06334.

[6] C. Schiller, S. Schmidtlein, C. Boonman, *et al.*, "Deep learning and citizen science enable automated plant trait predictions from photographs," *Scientific Reports*, vol. 11, no. 1, p. 16 395, 2021. DOI: 10.1038/s41598-021-95616-0. [Online]. Available: https://doi.org/10.1038/s41598-021-95616-0.

[7] L. Wang, L. Wang, Y. Li, and J. Wang, "A century-long analysis of global warming and earth temperature using a random walk with drift approach," *Decision Analytics Journal*, vol. 7, p. 100 237, 2023, ISSN: 2772-6622. DOI: https://doi.org/10.1016/j.dajour.2023.100237. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2772662223000772.