Multi-modal Feature Fusion Networks for GeoLifeCLEF 2024

Notebook for the GeoLifeCLEF Lab at CLEF 2024

Zehua Cheng[1,*,†], Wei Dai[2,†] and Jiahao Sun[3]

[1] *Department of Computer Science, University of Oxford*
[2] *Robo Space*
[3] *FLock.io*

Abstract

The GeoLifeCLEF 2024 challenge focuses on accurately predicting plant species distributions and their changes over space and time, vital for biodiversity conservation and ecological research. This study employs a multi-modal learning strategy combining deep learning with remote sensing data to address this task. The study harmonizes diverse datasets such as land cover, soil attributes, human impact, elevation, and climate into a unified tabular format, we applied a multi-modal feature fusion to the multi-modal data. This multi-modal approach, leveraging advanced architectures like ConvNeXt and ensemble techniques, significantly improves the prediction of plant species distributions, contributing to more informed conservation planning and ecological understanding. The methodologies employed showcase the power of combining deep learning with multi-source geospatial data for complex environmental predictions.

Keywords

Multi-modal Learning, Deep Learning, Remote Sensing

1. Introduction

Accurately predicting the composition of plant species and their spatial and temporal changes at a fine resolution holds substantial value across various domains. This capability is particularly advantageous for biodiversity management and conservation efforts, as it enables more effective planning and intervention strategies. Moreover, it enhances species identification and inventory tools, facilitating more precise and efficient cataloguing of flora. Additionally, such predictive insights serve educational purposes by providing detailed and dynamic data for academic research and environmental education, fostering a deeper understanding of ecological patterns and processes.

The GeoLifeCLEF 2024 [1] undertaking endeavours to anticipate the spatiotemporal distribution of plant species by leveraging a multifaceted ensemble of predictive factors. This ensemble encompasses satellite-derived imagery and sequential temporal data, alongside climatological time series information. Furthermore, it integrates raster-based environmental parameters such as land cover mappings, indices of anthropogenic influence (human footprint), bioclimatic variables, and soil attribute profiles. The consolidation of these diverse datasets is purposed to augment the precision and dependability of predictive models for flora presence, thereby enriching both biodiversity surveillance initiatives and advancing the frontiers of ecological research. In pursuit of these objectives, the project furnishes a substantial training corpus consisting of roughly 5 million documented plant observations spanned across the European continent. This training set is characterized by single-label, presence-only annotations. Complementary to this, a validation subset is provided, embodying nearly 5,000 geographically distinct plots, along with a testing corpus of 20,000 plots. Both the validation and testing sets are meticulously annotated with multi-label, presence-absence data, encapsulating the full spectrum of species inhabiting each plot. This

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France

[*] Corresponding author.

[†] These authors contributed equally.

zehua.cheng@cs.ox.ac.uk (Z. Cheng); daiwei@robo.space (W. Dai); sun@flock.io (J. Sun)

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

meticulous dataset architecture facilitates the robust development and rigorous evaluation of predictive algorithms aimed at elucidating plant species distribution patterns.

The subsequent sections detail the dataset that forms the foundation of this challenge and provide an overview of our submission to the competition. We open-sourced our code implementation with PyTorch [2] implementation with PyTorch Lightning [3] on GitHub[1].

2. Data

The dataset consists of two main components: species occurrence records and environmental/spatial data.

2.1. Species Occurrence Records

The core of the GeoLifeCLEF 2024 dataset consists of species occurrence records, which document the presence or absence of specific species at various geographic locations. These records are derived from multiple sources, including field observations, museum collections, and citizen science initiatives. Each record typically includes the species name, geographic coordinates, and the date of observation. This data is essential for understanding the spatial and temporal patterns of species distributions and serves as the primary input for training predictive models.

More precisely, these data include:

- **Presence-Absence (PA) Surveys**: This subset offers a vital counterbalance to the common challenges in ecological studies by providing around 90,000 verified records of species presence and absence across approximately 10,000 flora species. It is designed to address the inherent bias of assuming absences based on the lack of recorded sightings, thereby enhancing the accuracy and reliability of predictive models.
- **Presence-Only (PO) Occurrences**: Comprising around 5 million observations sourced from GBIF, this extensive dataset underscores the practical realities of large-scale biodiversity data collection. Despite potential biases due to non-standardized sampling methods, these PO occurrences offer unparalleled geographic coverage and species richness, enabling continent-wide analyses.

2.2. Environmental and Spatial Data

In addition to species occurrence records, the dataset includes a rich array of environmental and spatial variables that influence species distributions. These variables are provided in raster or vector formats and cover several key aspects:

- **Climate Data**: This includes variables such as temperature, precipitation, and humidity, which serve as pivotal elements in modeling species habitat affinities and their adaptive strategies vis-à-vis changing climate regimes.
- **Terrain Data**: The dataset encompasses topographical attributes including elevation, slope, and aspect. These geographical determinants play a decisive role in sculpting microclimatic niches and the physical structure of habitats, thereby exerting a profound influence on species distributions.
- **Land Cover Data**: A detailed depiction of land cover classifications, land usage patterns, and habitat intricacies is furnished. This layer of information is crucial for understanding the extent and suitability of habitats for diverse species populations.
- **Soil Data**: Soil-related variables, entailing soil taxonomy, nutrient content, and other soil qualities, are integrated into the dataset. These edaphic factors are known to exert substantial effects on plant community compositions, which in turn, impact the fauna supported by such ecosystems.

[1] https://github.com/limberc/GeoLifeCLEF2024

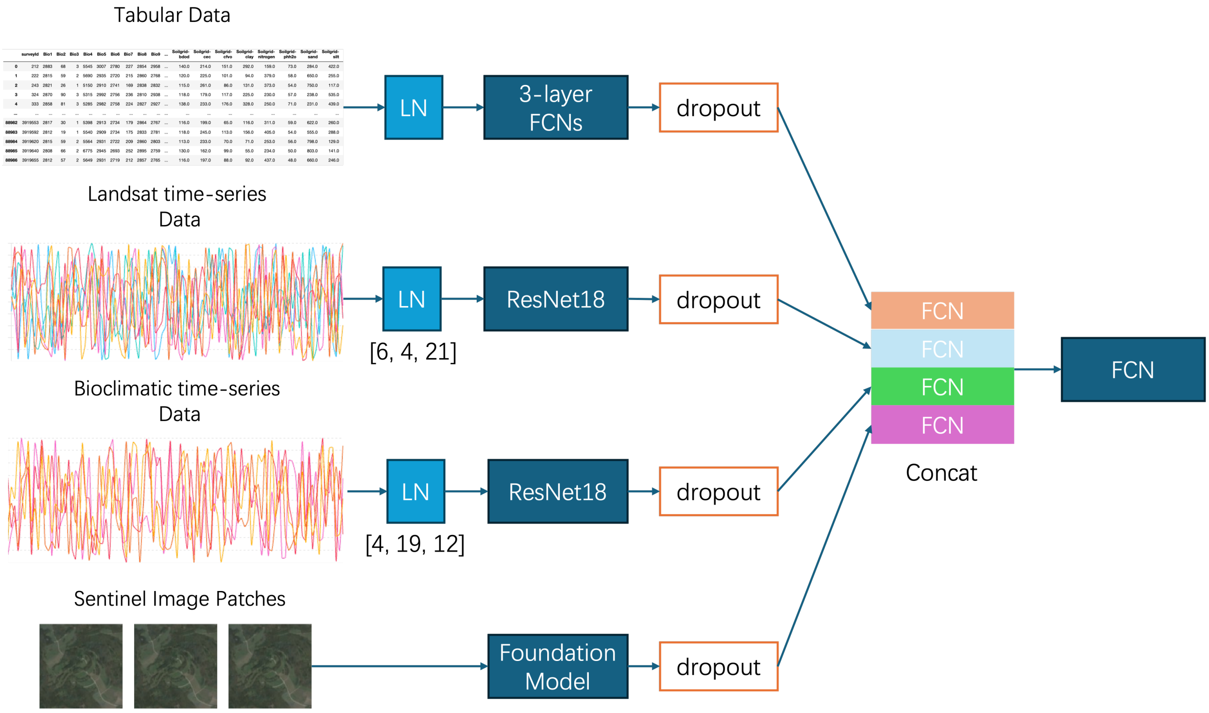

Figure 1: Overview of proposed methodology. We merge 4 different modalities with a domain-specific encoder.

- **Spatial Data**: To account for the pervasive impact of human activities on biodiversity, the dataset incorporates spatial information indicative of anthropogenic disturbances. This encompasses metrics such as distance to water bodies, road networks, and population centers, enabling the assessment of how these factors modulate species distributions and habitat fragmentation.

More specifically, these data include:

- **Satellite Imagery and Time Series:** The integration of satellite data, specifically RGB and NIR imagery patches from Sentinel2 and time series from Landsat, provides a dynamic view of the earth's surface over two decades. This visual and temporal data allows researchers to examine habitat changes, vegetation dynamics, and the effects of disturbances such as fires or land conversions at high resolution. The structured organization and clear loading instructions ensure these data are user-friendly and readily applicable to various modeling techniques.
- **Environmental Variables and Rasters:** The comprehensive environmental dataset spans climatic, topographic, land use, and anthropogenic impact variables, offering scalar values, time-series, and original raster formats. Notably, bioclimatic variables, soil characteristics from SoilGrids, elevation data from ASTER GDEM, land cover classifications, and human footprint indicators are included. These granular details facilitate nuanced investigations into how species distributions respond to specific environmental gradients and human activities.

Collectively, this exhaustive medley of environmental and spatial variables furnishes a robust foundation for advancing our understanding of species ecology and facilitating predictive modeling that is nuanced, comprehensive, and attuned to the complex interplay of natural and anthropogenic forces shaping Earth's biodiversity landscapes.

3. Data Preprocessing

To guarantee the dataset's reliability, coherence, and ease of application, a rigorous data preprocessing and quality control regimen is employed, encompassing the following stages:

- **Data Cleaning:** This fundamental step entails addressing missing data points through imputation or exclusion, eliminating duplicate records, and rectifying discrepancies or inaccuracies to uphold data integrity.
- **Feature Derivation and Transformation:** In an effort to enhance the dataset's predictive power, new features are engineered based on domain knowledge and statistical exploration. Concurrently, existing variables undergo transformations—such as normalization, scaling, or encoding—to better elucidate underlying patterns and facilitate correlations within the data.
- **Spatial Data Harmonization:** All spatial information is meticulously projected onto a unified coordinate reference system to ensure geographical congruence. Raster datasets may further be subjected to resampling or aggregation processes to establish a uniform grid resolution, facilitating spatial analysis and alignment across datasets.
- **Dataset Partitioning:** A systematic division of the data into discrete subsets for training and testing purposes is executed. This stratified partitioning strategy is vital for unbiased model training and the robust estimation of model generalizability.
- **Data Formatting for Machine Learning Compatibility:** The finalised datasets are then formatted to conform with the input requirements of various machine learning algorithms. This may involve restructuring the data into tabular formats amenable to classical statistical models or converting it into tensor structures compatible with deep learning frameworks.

These meticulous preprocessing and quality assurance measures render the dataset usable and optimised for high-fidelity ecological modelling, thereby fostering advancements in multi-modal machine learning research and its applications within biodiversity science.

4. Methodology

The overview structure of our methodology is presented in Figure 1.

The adoption of a multi-modal approach is motivated by several key factors that underscore the complexity and interconnectedness inherent in environmental and ecological research. By integrating diverse datasets, our methodology aims to capture a more holistic and nuanced understanding of the ecosystems under study, thereby enhancing prediction accuracy, explanatory power, and the potential for novel insights.

The integration of multiple data modalities serves as a cornerstone in our pursuit of a comprehensive understanding of environmental ecosystems. Each modality contributes a distinctive viewpoint, thereby complementing one another to construct a multifaceted portrait of the natural world. Tabular datasets lay the groundwork with their provision of foundational context, outlining static elements such as land usage patterns and soil classifications. In contrast, Landsat time-series data dynamically charts the evolution of vegetation and land cover over an extended period, painting a vivid picture of ecological change. Bioclimatic datasets introduce a temporal aspect, highlighting the critical role of fluctuating weather patterns in determining habitat suitability for various species. Finally, Sentinel-2 satellite imagery furnishes high-resolution snapshots that illuminate the intricacies of present-day landscapes, offering granular details unattainable through other means. Collectively, this fusion of modalities engenders a dataset of unparalleled richness, capable of unraveling the intricate web of ecological relationships with heightened clarity and depth.

The robustness and generalizability of a multi-modal analytical framework derive from its inherent capacity to mitigate the biases and limitations associated with singular data sources. By intertwining diverse datasets, our model fosters a checks-and-balances system, cross-verifying information to correct potential inaccuracies or discrepancies. This approach ensures that the resultant predictions are not only more resilient to errors but also possess a broader applicability, transcending specific datasets or locales to yield insights applicable across varied scenarios and geographical contexts.

The synergy achieved through multi-modal integration transcends the sum of its parts, revealing intricate patterns and connections that would otherwise remain obscured. An illustrative case is the confluence of land cover data from tabular and Landsat sources with bioclimatic trends, which can illuminate the intricate dance between anthropogenic land modifications and the shifting climate, providing profound insights into the mechanisms driving ecosystem transformations under the influence of global warming.

Regarding the processing of Sentinel-2 imagery, the development of specialized foundation models assumes paramount importance. Given the sheer informational depth and complexity encapsulated within these high-resolution images, leveraging pre-trained foundation models becomes a strategic imperative. These models, having absorbed a broad spectrum of visual knowledge from extensive datasets, are adept at extracting universal representations that transcend specific tasks. Their adaptability facilitates efficient fine-tuning for our ecological assessment needs, empowering us to discern subtle shifts in vegetation vitality, monitor the expansion of urban landscapes, and track changes in aquatic ecosystems with precision. Thus, by embracing these advanced models, we unlock the full potential of Sentinel-2 imagery, unearthing features and trends that may escape detection through conventional methodologies, and enhancing the precision and scope of our ecological interpretations.

Here we elaborate on the methodology with specific modality data processing as below:

- Initially, we consolidate a multitude of environmental raster datasets, encompassing 'LandCover', 'SoilGrids', 'HumanFootprint', 'Elevation' and 'Climate' data into a unified tabular format. This harmonization step allows for streamlined processing. Subsequently, the concatenated table undergoes layer normalization, a preprocessing technique that scales features to a standard Gaussian distribution, ensuring each feature contributes equally to the learning process regardless of its original scale. The normalized tabular data are then fed into a three-layer Fully Connected Network (FCN), a classic deep learning architecture well-suited for handling structured data, enabling extraction of high-level abstract features pertinent to our predictive tasks.
- For temporal satellite imagery from Landsat, we structure the dataset with six spectral bands (red, green, blue, near-infrared (nir), shortwave infrared 1 (swir1), and shortwave infrared 2 (swir2)) across four quarters for each year within the timeframe of 2000 to 2020, yielding an input shape of [6, 4, 21]. Before encoding via a Residual Network (ResNet18), a variant of deep convolutional neural networks renowned for their efficiency in image recognition tasks, the data is normalized to enhance training stability and performance. The employment of ResNet18 facilitates the extraction of spatiotemporal features crucial for understanding landscape dynamics.
- Bioclimatic data, sourced from monthly GeoTIFF CHELSA climate rasters at a 30 arc-second resolution, includes precipitation (pr), maximum daily temperature (tasmax), minimum daily temperature (tasmin), and mean daily temperature (tas), spanning from January 2000 to June 2019. This dataset is presented in three formats, though for our methodology, we utilize it as a data cube (.pt tensor object) structured as [19 years, 12 months, 4 bioclimatic variables]. This arrangement enables straightforward feeding into machine learning models, allowing for the extraction of seasonal and interannual climate patterns relevant to ecological studies.
- Satellite imagery from the Sentinel-2 mission is harnessed and preprocessed using Ecodatacube [4], resulting in raster files consistent across Europe with a standardized Coordinate Reference System (CRS). To align with occurrence data, patches from specified GPS coordinates and dates are extracted and transformed into JPEG files: RGB images in 3-channels with 128×128 pixels and Near-Infrared (NIR) as single-channel images. These images undergo additional preprocessing steps involving data clipping to manage outliers (values exceeding 10000 are truncated) and application of gamma correction ($\gamma = 2.5$) to enhance visual contrast and normalize the intensity distribution, preparing the data for further analysis in a machine vision context.

5. Results

We changed the foundation model of processing the Sentinel image patches in Figure 1 to ConvNeXt [5], MaxViT [6] and SwinV2 [7]. We perform an ensemble of all listed model with an ensemble ratio of $[0.15, 0.1, 0.4, 0.35]$.

Table 1
Experimental Results of Different Foundation Models in sentinel image patches processing.

Sentinel Module	train loss	val loss	best score	@top
ConvNeXt-Base	0.00386	0.00413	0.39230	18
ConvNeXt-Large	0.00388	0.00404	0.39835	19
MaxVit_t	0.00354	0.00393	0.40811	18
Swinv2-Base	0.00364	0.00401	0.40533	17
Official Baseline	-	-	0.32359	-
Ensemble	-	-	0.42581	18

Table 1 shows the experimental results of using different foundation models for processing the Sentinel image patches. The ConvNeXt-Large model achieved the best single model performance with a top score of 0.39835. However, the MaxVit_t model had the lowest training and validation loss. The SwinV2-Base model performed comparably to ConvNeXt-Large.

Ultimately, an ensemble of all four models using a weighted average with weights $[0.15, 0.1, 0.4, 0.35]$ corresponding to the order of models in Table 1 gave the highest score of 0.42581. This represents a substantial improvement of over 0.10 compared to the official baseline score of 0.32359.

The strong performance of the ConvNeXt and SwinV2 models is not surprising, as they represent some of the latest advancements in computer vision architectures. ConvNeXt introduces modernized design elements to the standard ConvNet architecture that boost performance. SwinV2 scales up window-based attention mechanisms to efficiently process high-resolution images. MaxViT, while not achieving quite as high of a score, still outperformed the baseline by a significant margin. Its multi-axis attention enables it to capture both local and global context. The fact that it had the lowest training and validation losses suggests it may be less prone to overfitting.

The large performance gain from ensembling the models highlights the benefit of combining the predictions from multiple high-performing models. Each model captures slightly different visual patterns, and aggregating them leads to a more robust final prediction.

Overall, these results demonstrate the power of leveraging state-of-the-art computer vision models pretrained on large datasets for the task of plant species identification from satellite imagery. The proposed ensemble approach in particular shows promise for maximizing predictive performance on this challenging task. Further work could explore additional model architectures and ensembling strategies to push performance even higher.

6. Conclusion

This study presents a novel multi-modal deep learning approach for predicting plant species distributions across Europe, developed for the GeoLifeCLEF 2024 challenge. Our methodology integrates diverse data types, including tabular environmental variables, time-series satellite imagery, bioclimatic data, and high-resolution satellite images. We ensemble the model with different basckbone to boost predictive performance, with our final submission achieving a top score of 0.42581, markedly surpassing the competition baseline. The impressive results highlight the power of multi-modal learning to harness complementary information from heterogeneous data sources. By integrating insights from static environmental conditions, temporal vegetation dynamics, climate patterns, and high-resolution imagery, our model learned nuanced and robust representations of species-habitat relationships.

References

[1] L. Picek, C. Botella, M. Servajean, B. Deneu, D. Marcos Gonzalez, R. Palard, T. Larcher, C. Leblanc, J. Estopinan, P. Bonnet, A. Joly, Overview of GeoLifeCLEF 2024: Species presence prediction based

on occurrence data and high-resolution remote sensing images, in: Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, 2024.

[2] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, Pytorch: An imperative style, high-performance deep learning library, ArXiv abs/1912.01703 (2019).

[3] W. Falcon, Pytorch lightning, GitHub. Note: https://github.com/williamFalcon/pytorch-lightning (2023).

[4] M. Witjes, L. Parente, J. Križan, T. Hengl, L. Antoni'c, Ecodatacube.eu: Analysis-ready open environmental data cube for europe, Research Square (2023). doi:10.21203/rs.3.rs-2277090/v3.

[5] Z. Liu, H. Mao, C. Wu, C. Feichtenhofer, T. Darrell, S. Xie, A convnet for the 2020s, 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022) 11966–11976.

[6] Z. Tu, H. Talebi, H. Zhang, F. Yang, P. Milanfar, A. C. Bovik, Y. Li, Maxvit: Multi-axis vision transformer, in: European Conference on Computer Vision, 2022.

[7] Z. Liu, H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. Wei, J. Ning, Y. Cao, Z. Zhang, L. Dong, F. Wei, B. Guo, Swin transformer v2: Scaling up capacity and resolution, 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021) 11999–12009.