Leveraging Multimodality for Biodiversity Data: Exploring joint representations of species descriptions and specimen images using CLIP

Maya Sahraoui[‡], Youcef Sklab[§], Marc Pignal[|], Régine Vignes Lebbe[¶], Vincent Guigue[#]

‡ MNHN, Paris, France
§ IRD, Paris, France
| MNHN, Paris, France, Metropolitan
¶ Institut de Systématique, Evolution, Biodiversité (ISYEB), Muséum national d'Histoire naturelle, CNRS, Sorbonne Université, EPHE, Université des Antilles, Paris, France
# AgroParisTech, Paris, France

Corresponding author: Maya Sahraoui (sahraoui@isir.upmc.fr)

Abstract

In recent years, the field of biodiversity data analysis has witnessed significant advancements, with a number of models emerging to process and extract valuable insights from various data sources. One notable area of progress lies in the analysis of species descriptions, where structured knowledge extraction techniques have gained prominence. These techniques aim to automatically extract relevant information from unstructured text, such as taxonomic classifications and morphological traits. (Sahraoui et al. 2022, Sahraoui et al. 2023) By applying natural language processing (NLP) and machine learning methods, structured knowledge extraction enables the conversion of textual species descriptions into a structured format, facilitating easier integration, searchability, and analysis of biodiversity data.

Furthermore, object detection on specimen images has emerged as a powerful tool in biodiversity research. By leveraging computer vision algorithms (Triki et al. 2020, Triki et al. 2021,Ott et al. 2020), researchers can automatically identify and classify objects of interest within specimen images, such as organs, anatomical features, or specific taxa. Object detection techniques allow for the efficient and accurate extraction of valuable information, contributing to tasks like species identification, morphological trait analysis, and biodiversity monitoring. These advancements have been particularly significant in the context of herbarium collections and digitization efforts, where large volumes of specimen images need to be processed and analyzed.

On the other hand, multimodal learning, an emerging field in artificial intelligence (AI), focuses on developing models that can effectively process and learn from multiple modalities, such as text and images (Li et al. 2020, Li et al. 2021, Li et al. 2019, Radford et al. 2021, Sun et al. 2021, Chen et al. 2022). By incorporating information from different

modalities, multimodal learning aims to capture the rich and complementary characteristics present in diverse data sources. This approach enables the model to leverage the strengths of each modality, leading to enhanced understanding, improved performance, and more comprehensive representations.

Structured knowledge extraction from species descriptions and object detection on specimen images synergistically enhances biodiversity data analysis. This integration leverages textual and visual data strengths, gaining deeper insights. Extracted structured information from descriptions improves search, classification, and correlation of biodiversity data. Object detection enriches textual descriptions, providing visual evidence for the verification and validation of species characteristics.

To tackle the challenges posed by the massive volume of specimen images available at the Herbarium of the National Museum of Natural History in Paris, we have chosen to implement the CLIP (Contrastive Language-Image Pretraining) model (Radford et al. 2021) developed by OpenAI. CLIP utilizes a contrastive learning framework to recognize joint representations of text and images. The model is trained on a large-scale dataset consisting of text-image pairs from the internet, enabling it to understand the semantic relationships between textual descriptions and visual content.

Fine-tuning the CLIP model on our dataset of species descriptions and specimen images is crucial for adapting it to our domain. By exposing the model to our data, we enhance its ability to understand and represent biodiversity characteristics. This involves training the model on our labeled dataset, allowing it to refine its knowledge and adapt to biodiversity patterns.

Using the fine-tuned CLIP model, we aim to develop an efficient search engine for the Herbarium's vast biodiversity collection. Users can query the engine with morphological keywords, and it will match textual descriptions with specimen images to provide relevant results. This research aligns with the current AI trajectory for biodiversity data, paving the way for innovative approaches to address conservation and understanding of our planet's biodiversity.

Keywords

joint representation learning, image captioning, multimodal named entity recognition, visual grounding

Presenting author

Maya Sahraoui

Presented at

TDWG 2023

Conflicts of interest

The authors have declared that no competing interests exist.

References

- Chen X, Zhang N, Li L, Yao Y, Deng S, Tan C, Huang F, Si L, Chen H (2022) Good Visual Guidance Makes A Better Extractor: Hierarchical Visual Prefix for Multimodal Entity and Relation Extraction. arXiv. Comment: Accepted by NAACL 2022. https://doi.org/10.18653/v1/2022.findings-naacl.121
- Li LH, Yatskar M, Yin D, Hsieh C, Chang K (2019) VisualBERT: A Simple and Performant Baseline for Vision and Language. arXiv. Comment: Work in Progress. URL: http://arxiv.org/abs/1908.03557
- Li LH, You H, Wang Z, Zareian A, Chang S, Chang K (2021) Unsupervised Vision-and-Language Pre-training Without Parallel Images and Captions. arXiv. Comment: NAACL 2021 Camera Ready. URL: http://arxiv.org/abs/2010.12831
- Li X, Yin X, Li C, Zhang P, Hu X, Zhang L, Wang L, Hu H, Dong L, Wei F, Choi Y, Gao J (2020) Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. arXiv. Comment: ECCV 2020, Code and pre-trained models are released: https://github.com/microsoft/Oscar. https://doi.org/10.1007/978-3-030-58577-8_8
- Ott T, Palm C, Vogt R, Oberprieler C (2020) GinJinn: An object-detection pipeline for automated feature extraction from herbarium specimens. Applications in Plant Sciences 8 (6). https://doi.org/10.1002/aps3.11351
- Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, Krueger G, Sutskever I (2021) Learning Transferable Visual Models From Natural Language Supervision. arXiv. arXiv:2103.00020 [cs] version: 1. URL: http://arxiv.org/abs/2103.00020
- Sahraoui M, Pignal M, Vignes Lebbe R, Guigue V (2022) NEARSIDE: Structured kNowledge Extraction frAmework from SpecIes DEscriptions. Biodiversity Information Science and Standards 6 https://doi.org/10.3897/biss.6.94297
- Sahraoui M, Guigue V, Vignes-Lebbe R, Pignal M (2023) Extraction d'entités nommées à partir de descriptions d'espèces. URL: https://hal.science/hal-04131571
- Sun L, Wang J, Zhang K, Su Y, Weng F (2021) RpBERT: A Text-image Relation Propagation-based BERT Model for Multimodal NER. arXiv. Comment: Accepted by AAAI2021. https://doi.org/10.1609/aaai.v35i15.17633
- Triki A, Bouaziz B, Mahdi W, Gaikwad J (2020) Objects Detection from Digitized Herbarium Specimen based on Improved YOLO V3. scitepress
- Triki A, Bouaziz B, Gaikwad J, Mahdi W (2021) Deep leaf: Mask R-CNN based leaf detection and segmentation from digitized herbarium specimen images. Pattern Recognition Letters 150: 76-83. https://doi.org/10.1016/j.patrec.2021.07.003