The Role of the CLIP Model in Analysing Herbarium Specimen Images

Vamsi Krishna Kommineni[‡,§,|], Jens Kattge[|], Jitendra Gaikwad[‡], Susanne Tautenhahn[|], Birgitta Koenigries[‡,¶,§]

‡ Friedrich Schiller University Jena, Jena, Germany
§ German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Leipzig, Germany
| Max Planck Institute for Biogeochemistry Jena, Jena, Germany
¶ Michael Stifel Center Jena, Jena, Germany

Corresponding author: Vamsi Krishna Kommineni (vamsi.krishna.kommineni@uni-jena.de)

Abstract

The number of openly-accessible digital plant specimen images is growing tremendously and available through data aggregators: Global Biodiversity Information Facility (GBIF) contains 43.2 million images, and Intergrated Digitized Biocollections (iDigBio) contains 32.4 million images (Accessed on 29.06.2023). All these images contain great ecological (morphological, phenological, taxonomic etc.) information, which has the potential to facilitate the conduct of large-scale analyses. However, extracting this information from these images and making it available to analysis tools remains challenging and requires more advanced computer vision algorithms. With the latest advancements in the natural language processing field, it is becoming possible to analyse images with text prompts. For example, with the Contrastive Language-Image Pre-Training (CLIP) model, which was trained on 400 million image-text pairs, it is feasible to classify day-to-day life images by providing different text prompts and an image as an input to the model, then the model can predict the most suitable text prompt for the input image.

We explored the feasibility of using the CLIP model to analyse digital plant specimen images. A particular focus of this study was on the generation of appropriate text prompts. This is important as the prompt has a large influence on the results of the model. We experimented with three different methods: a) automatic text prompt based on metadata of the specific image or other datasets, b) automatic generic text prompt of the image (describing what is in the image) and c) manual text prompt by annotating the image. We investigated the suitability of these prompts with an experiment, where we tested whether the CLIP model could recognize a herbarium specimen image using digital plant specimen images and semantically disparate text prompts.

Our ultimate goal is to filter the digital plant specimen images based on the availability of intact leaves and measurement scale to reduce the number of specimens that reach the downstream pipeline, for instance, the segmentation task for the leaf trait extraction

process. To achieve the goal, we are fine-tuning the CLIP model with a dataset of around 20,000 digital plant specimen image-text prompt pairs, where the text prompts were generated using different datasets, metadata and generic text prompt methods. Since the text prompts can be created automatically, it is possible to eradicate the laborious manual annotating process.

In conclusion, we present our experimental testing of the CLIP model on digital plant specimen images with varied settings and how the CLIP model can act as a potential filtering tool. In future, we plan to investigate the possibility of using text prompts to do the instance segmentation to extract leaf trait information using Large Language Models (LLMs).

Keywords

text prompts, computer vision, natural language processing, large language models

Presenting author

Vamsi Krishna Kommineni

Presented at

TDWG 2023

Conflicts of interest

The authors have declared that no competing interests exist.