BISS Biodiversity Information Science and Standards

Biodiversity Information Science and Standards 2: e25699
doi: 10.3897/biss.2.25699

OPEN ACCESS

Conference Abstract

A Pipeline for Deep Learning with Specimen Images in iDigBio - Applying and Generalizing an Examination of Mercury Use in Preparing Herbarium Specimens

Matthew Collins[‡,§], Gaurav Yeole[‡], Paul B Frandsen[|], Rebecca B Dikow[¶], Sylvia Orli[¶], Renato J Figueiredo[‡]

‡ University of Florida, Gainesville, United States of America
§ iDigBio, Gainesville, United States of America
| Brigham Young University, Provo, United States of America
¶ Smithsonian Institution, Washington, United States of America

Corresponding author: Matthew Collins (mcollins@acis.ufl.edu)

Received: 11 Apr 2018 | Published: 03 Jul 2018

Citation: Collins M, Yeole G, Frandsen P, Dikow R, Orli S, Figueiredo R (2018) A Pipeline for Deep Learning with Specimen Images in iDigBio - Applying and Generalizing an Examination of Mercury Use in Preparing Herbarium Specimens. Biodiversity Information Science and Standards 2: e25699.
https://doi.org/10.3897/biss.2.25699

Abstract

iDigBio Matsunaga et al. 2013 currently references over 22 million media files, and stores approximately 120 terabytes worth of those media files co-located with our compute infrastructure. Using these images for scientific research is a logistical and technical challenge. Transferring large numbers of images requires programming skill, bandwidth, and storage space. While simple image transformations such as resizing and generating histograms are approachable on desktops and laptops, the neural networks commonly used for learning from images require server-based graphical processing units (GPUs) to run effectively.

Using the GUODA (Global Unified Open Data Access) infrastructure, we have built a model pipeline for applying user-defined processing to any subset of the images stored in iDigBio. This pipeline is run on servers located in the Advanced Computing and Information

Systems lab (ACIS) alongside the iDigBio storage system. We use Apache Spark, the Hadoop File System (HDFS), and Mesos to perform the processing. We have placed a Jupyter notebook server in front of this architecture which provides an easy environment with deep learning libraries for Python already loaded for end users to write their own models. Users can access the stored data and images and manipulate them according to their requirements and make their work publicly available on GitHub.

As an example of how this pipeline can be used in research, we applied a neural network developed at the Smithsonian Institution to identify herbarium sheets that were prepared with hazardous mercury containing solutions Schuettpelz et al. 2017. The model was trained with Smithsonian resources on their images and transferred to the GUODA infrastructure hosted at ACIS which also houses iDigBio. We then applied this model to additional images in iDigBio to classify them to illustrate the application of these techniques to broad image corpora potentially to notify other data publishers of contamination. We present the results of this classification not as a verified research result, but as an example of the collaborative and scalable workflows this pipeline and infrastructure enable.

Keywords

iDigBio, deep learning, image, Spark

Presenting author

Matthew Collins

Presented at

Biodiversity Information Standards (TDWG) 2018, Dunedin, NZ

References

- Matsunaga A, Thompson A, Figueiredo R, Germain-Aubrey C, Collins M, Beaman R, MacFadden B, Riccardi G, Soltis P, Page L, Fortes JB (2013) 2013 IEEE 9th International Conference on e-Science. 2013 IEEE 9th International Conference on e-Science https://doi.org/10.1109/escience.2013.48
- Schuettpelz E, Frandsen P, Dikow R, Brown A, Orli S, Peters M, Metallo A, Funk V, Dorr L (2017) Applications of deep convolutional neural networks to digitized natural history collections. Biodiversity Data Journal 5: e21139. https://doi.org/10.3897/bdj.5.e21139