BISS Biodiversity Information Science and Standards

Biodiversity Information Science and Standards 3: e37615
doi: 10.3897/biss.3.37615

OPEN ACCESS

Conference Abstract

Methods, New Software Tools, and Best Practices for Developing High-quality Training Data for Machine Learning-based Image Analysis in Biodiversity Research

Brian J. Stucky[‡], Laura Brenskelle[‡], Robert Guralnick[‡]

[‡] Florida Museum of Natural History, University of Florida, Gainesville, FL, United States of America

Corresponding author: Brian J. Stucky (stucky.brian@gmail.com)

Received: 22 Jun 2019 | Published: 02 Jul 2019

Citation: Stucky B, Brenskelle L, Guralnick R (2019) Methods, New Software Tools, and Best Practices for Developing High-quality Training Data for Machine Learning-based Image Analysis in Biodiversity Research. Biodiversity Information Science and Standards 3: e37615. https://doi.org/10.3897/biss.3.37615

Abstract

Recent progress in using deep learning techniques to automate the analysis of complex image data is opening up exciting new avenues for research in biodiversity science. However, potential applications of machine learning methods in biodiversity research are often limited by the relative scarcity of data suitable for training machine learning models. Development of high-quality training data sets can be a surprisingly challenging task that can easily consume hundreds of person-hours of time. In this talk, we present the results of our recent work implementing and comparing several different methods for generating annotated, biodiversity-oriented image data for training machine learning models, including collaborative expert scoring, local volunteer image annotators with on-site training, and distributed, remote image annotation via citizen science platforms. We discuss error rates, among-annotator variance, and depth of coverage required to ensure highly reliable image annotations. We also discuss time considerations and efficiency of the various methods. Finally, we present new software, called ImageAnt (currently under development), that supports efficient, highly flexible image annotation workflows. ImageAnt was created primarily in response to the challenges we discovered in our own efforts to generate image-based training data for machine learning models. ImageAnt features a simple user interface and can be used to implement sophisticated, adaptive scripting of image annotation tasks.

Keywords

machine learning, computer vision, image labeling

Presenting author

Brian J. Stucky

Presented at

Biodiversity_Next 2019