Digital Naturalist - AI Enabled tool for Biodiversity Researchers

Literature Survey

Introduction

The ever-growing number of digital sensors in the environment has led to an increase in the amount of digital data being generated. This includes data from satellites, weather stations, data from “internet of things” devices, and data collected by members of the public via smartphone applications, to name but a few. These new sources of data have contributed to the era of “Big Data” characterized by large volumes of data, of numerous types and quality, being generated at an increasing speed. This presents challenges and opportunities across a number of domains, including water management, camera trapping, and acoustic analysis. To process these data into useful information there are many tools available, including classical statistical analyses and classification by citizen scientists. However, at some point traditional approaches may become inefficient or even impossible given the volume, diversity, and heterogeneity of these data. Storage, exploration, curation, and revision of data may have to be re-thought to allow for their quick and efficient transformation, annotation, or analysis. This is particularly difficult for multimedia data which are typically much more complex than other data types. For example, biodiversity and environmental records in the form of audio, video, or image files are typically larger and more complex than text or numeric data. Large-scale analysis of multimedia data has only been possible in recent years since the development of large computational facilities, both academic and commercial. Regardless, the analysis of

multimedia data is often further complicated because of their non-standardized methods of acquisition, with highly diverse devices, sensors, formats, scales, environmental contexts, and taxonomic scope. Building efficient, scalable, and robust approaches to solve these problems is a difficult scientific challenge at the forefront of data science and machine learning specifically.

Artificial intelligence (AI) techniques have profoundly transformed our ability to extract information from visual data. AI techniques have been applied for a long time in security and industrial domains, for example, in iris recognition or the detection of faulty objects in manufacturing. They were nevertheless only recently made more widely accessible after their use in smartphone apps for face recognition and song identification. Combined with increasing access to cloud-based computation, AI techniques can now automatically analyze hundreds of thousands of visual data every day.

AI naturalists, just like their human counterparts, may have their own biases which must be fully understood if the information that they generate is to be trusted and suitably utilized. For example, most AI systems can only detect or recognize already seen (or learned) objects or concepts. Benchmark datasets of images can be organized to precisely assess the limits of AI systems' ability, highlighting where human expertise is still required. Deep learning models (some of the most advanced AI algorithms) are developed with training datasets that allow them to capture discriminant visual patterns. Their performances are then strongly correlated to the quality and completeness of the datasets on which they are trained. Unbalanced, biased, or otherwise poor-quality training datasets will lead to underperforming algorithms in real conditions. During the learning phases, particular attention must be given to any relevant limitations of the training data, and the gap between these and the test data on which the developed algorithms will be evaluated.

Literature Survey

For any given research question, ecologists and data scientists should carefully consider the steps that might be required to ensure the relevance and accuracy of AI-generated data for any given research question. To aid this we have summarized our experience into an eight-point list of questions which we recommend researchers ask themselves when using AI classifier naturalists:

1. Does the spatial distribution of images fit your needs? Images from social media are often aggregated in areas of high population density or tourist hotspots. If the distribution is biased in some way, could this be accounted for in subsequent analyses?

2. Can you filter images before classification? For example, filtering can be done by carefully selecting your source of images, using GPS location, focusing on keywords in image metadata, or using high-level AI classifiers to remove non-target images.

3. What is the appropriate taxonomic resolution for your study? This will be driven by your research question, as well as an assessment of the AI naturalist's accuracy. Classifiers will tend to be more accurate at higher taxonomic levels, but this may vary between taxonomic groups.

4. What reporting biases exist in your dataset? For example, to what degree are charismatic species over-represented, or nocturnal species under-represented? Can you filter the data, or model the results to account for these biases if they are relevant?

5. Do reporting biases change over space or time? We observed significant differences in reporting bias between urban and rural settings, and we anticipate that temporal biases are likely to exist where public interest in elements of the natural environment change over time.

6. How will you propagate uncertainty in classifications? AI classifications are associated with a classification score which is indicative of the uncertainty in the identification. This can be used both as a threshold for removing erroneous results, and/or could be included in models to account for variation in uncertainty between observations.

7. Is the dataset used to train your AI naturalist a good match to the images being classified? A poor match between training and prediction datasets will result in higher error rates, which may not always be associated with low classification scores.

8. Have you adequately documented your dataset? To ensure reproducibility and interoperability ensure that you document the model used for classification, filtering steps used to collate images, and other metadata useful to future researchers, and which may be specified in data standards for AI-generated biodiversity which do not exist at the time of writing.

References

1. Bonnet P., Goëau H., Hang S.T., Lasseck M., Šulc M., Malécot V., Jauzein P., Melet J.-C., You C., Joly A. Plant identification: experts vs.

machines in the era of deep learning. In: Joly A., Vrochidis S., Karatzas K., Karppinen A., Bonnet P., editors. *Multimedia Tools and Applications for Environmental & Biodiversity Informatics Multimedia Systems and Applications.* Springer; 2018. pp. 131–149. [Google Scholar]

2. Norouzzadeh M.S., Nguyen A., Kosmala M., Swanson A., Palmer M.S., Packer C., Clune J. Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. *Proc. Natl. Acad. Sci. U S A.* 2018;115:E5716–E5725. [PMC free article] [PubMed] [Google Scholar]

3. Di Minin E., Fink C., Tenkanen H., Hiippala T. Machine learning for tracking illegal wildlife trade on social media. *Nat. Ecol. Evol.* 2018;2:406–407. [PubMed] [Google Scholar]

4. Jarić I., Correia R.A., Brook B.W., Buettel J.C., Courchamp F., Di Minin E., Firth J.A., Gaston K.J., Jepson P., Kalinkat G. iEcology: harnessing large online resources to generate ecological insights. *Trends Ecol. Evol.* 2020;35:630–639.. [PubMed] [Google Scholar]

5. Carranza-Rojas J., Mata-Montero E., Goeau H. *2018 IEEE International Work Conference on Bioinspired Intelligence (IWOBI)* IEEE; 2018. Hidden biases in automated image-based plant identification; pp. 1–9. [Google Scholar]

6. Lintott C.J., Schawinski K., Slosar A., Land K., Bamford S., Thomas D., Raddick M.J., Nichol R.C., Szalay A., Andreescu D. Galaxy zoo: morphologies derived from visual inspection of galaxies from the Sloan digital sky survey. *Mon. Not. R. Astron. Soc.* 2008;389:1179–1189. [Google Scholar]

7. Nguyen K., Fookes C., Jillela R., Sridharan S., Ross A. Long range iris recognition: a survey. *Pattern Recognit.* 2017;72:123–143. [Google Scholar]

8. Zhang Y., Li X., Gao L., Li P. A new subset based deep feature learning method for intelligent fault diagnosis of bearing. *Expert Syst. Appl.* 2018;110:125–142. [Google Scholar]

9. Rattani A., Derakhshani R. A survey of mobile face biometrics. *Comput. Electr. Eng.* 2018;72:39–52. [Google Scholar]

10. Wang A. The Shazam music recognition service. *Commun. ACM.* 2006;49:44–48. [Google Scholar]