New AI tool helps leverage database of 10 million biology images

February 13 2024, by Tatyana Woodall

Credit: Pixabay/CC0 Public Domain

Researchers have developed the largest-ever dataset of biological images suitable for use by machine learning—and a new vision-based artificial intelligence tool to learn from it.

The findings in the new study significantly broaden the scope of what scientists can do using artificial intelligence to analyze images of plants, animals and fungi to answer new questions, said Samuel Stevens, lead author of the study and a Ph.D. student in computer science and engineering at Ohio State.

"Our model will be useful for tasks spanning the entire tree of life," Stevens said. "Researchers will be able to do studies that wouldn't have been possible before."

The findings are published on the *arXiv* preprint server.

Stevens and his colleagues first curated and released the world's largest and most diverse machine learning-ready image dataset, TreeOfLife-10M, which contains over 10 million images of plants, animals and fungi covering more than 454,000 taxa in the tree of life. In comparison, the previous largest database ready for machine learning contains only 2.7 million images covering 10,000 taxa. The diversity of this data is one of the key enabling features of their algorithm.

They then developed BioCLIP, a new machine learning model released to researchers in December and designed to learn from the dataset by using both visual cues in the images with various types of text associated with the images, such as taxonomic labels and other information.

The researchers tested BioCLIP by seeing how well it could classify images as to where they belonged in the tree of life—including a rare species dataset that it did not see during training. Results showed that it performed 17% to 20% better than existing models on the task.

The BioCLIP model is publicly accessible here. Its demo, said Stevens, can also accurately discern the species of an arbitrary organism image, be it from the Serengeti Savannah, your local zoo or your backyard.

Traditional computational approaches used to organize abundant biology image databases are typically designed for specific tasks and aren't as capable of addressing new questions, contexts and datasets, Stevens said.

Additionally, because the model can be widely applied to the entire tree of life, their AI is more supportive of biologists whose real-world research is more broadly focused, instead of those studying specific niches, he added.

What makes this team's approach so effective, said Yu Su, co-author of the study and an assistant professor of computer science and engineering at Ohio State, is their model's ability to learn fine-tuned representations of images, or being able to tell the difference between similar-looking organisms within the same species and one species mimicking their appearance.

Whereas general computer vision models are useful for comparing common organisms like dogs and wolves, previous studies have revealed that they can't take note of the subtle differences between two species of the same plant genus.

Because of its better grasp of nuance, said Su, the model in this paper is also uniquely qualified to make determinations on rare and unseen species as well.

"BioCLIP covers many orders of magnitude more species and taxa than the previously publicly available for general vision models," he said. "Even when it has not seen a certain species before, it can come to a reasonable conclusion about how if this organism looks similar to this, then it's likely that."

As AI continues to advance, the study concludes, machine learning models like this one could soon become important tools for unraveling

biological mysteries that would otherwise take much longer to understand. And while this first iteration of BioCLIP relied heavily on images and information from citizen science platforms, Stevens said future models could be upgraded by including more images and data from scientific labs and museums. Because labs are able to collect richer textual descriptions of species that detail their morphological features and other subtle differences between closely related species, such resources will provide a bevy of important information for the AI model.

In addition, many scientific labs have information on the fossils of extinct species, which the team expects will also broaden the model's usefulness.

"Taxonomies are always changing as we update names and new species, so one thing we'd like to do in the future is leverage existing work much more heavily on how to integrate them," he said. "In AI, when you throw more data at a problem, you're going to get better results, so I think there's a bigger version we can continue to train into a larger, stronger model."

Other Ohio State co-authors include Jiaman Wu, Matthew J. Thompson, Elizabeth G. Campolongo, Chan Hee Song, David Edward Carlyn, Tanya Berger-Wolf and Wei-Lun Chao. Li Dong from Microsoft Research, Wasila M Dahdul from the University of California, Irvine, and Charles Stewart from the Rensselaer Polytechnic Institute also contributed.

More information: Samuel Stevens et al, BioCLIP: A Vision Foundation Model for the Tree of Life, *arXiv* (2023). DOI: 10.48550/arxiv.2311.18803

Provided by The Ohio State University