Robust Machine Learning Classification of Unlabeled Biological Data: A case study with herbaria sheets

Jonathan Koss[‡], Anthony Jiang[‡], Patrick Sweeney[‡], Nelson Rios[‡], Aaron Dollar[‡]

[‡] Yale University, New Haven, United States of America

Corresponding author: Jonathan Koss (jon.koss@gmail.com)

Abstract

There is much excitement across a broad range of biological disciplines over the prospect of using deep learning and similar modern statistical methods to label research data. The extensive time, effort, and cost required for humans to label a dataset drastically limits the type and amount of data that can be reasonably utilized, and is currently a major bottleneck to the extensive application of biological datasets such as specimen imagery, video and audio recordings. While a number of researchers have shown how deep convolutional neural networks (CNN) can be trained to classify image data with 80-90% accuracy, that range of accuracy is still too low for most research applications. Furthermore, applying these classifiers to new, unlabeled data from a dataset other than the one used for training the classifier would likely result in even lower accuracy. As a result, these classifiers have still not generally been applied to unlabeled data—which is where they could be most useful.

In this talk, we will present a method for determining a confidence metric on predicted classifications (i.e. "labels") from a deep CNN classifier that can inform a user whether to trust a particular automatic label or to discard it, thereby giving a reasonable and straight-forward method to label a previously unlabeled dataset with high confidence.

Essentially, it is an approach that allows an imperfect method of classification to be used in a useful way that can save an enormous amount of time and effort and/or greatly increase the amount of data that can be reasonably utilized.

In this work, the training dataset consisted of a set of records of flowering plant species that collectively exhibited a range of reproductive morphologies, represented multiple taxonomic groups, and could be easily scored by humans for reproductive condition by examination of specimen images. The records were labeled as reproductive, budding, flowering and/or fruiting. All of the data and images were obtained from the Consortium of Northeastern Herbaria portal (CNH). There were two unscored datasets that were used to evaluate the classifiers. One dataset contained the same taxa that were in the training

dataset and the second dataset contained all remaining flowering plant taxa in the CNH portal database that were not included in the other two datasets. Records of families with flowers that are obscure (i.e., they lack petals & sepals or have vestigial structures) were excluded.

To label the reproductive state of the plants, we trained one deep CNN classifier using the XCeption architecture for the binary classification of each state (e.g., budding vs. not budding). This method and architecture was chosen because of its success in similar image-classification tasks.

Each of these networks takes an image of a herbarium sheet as input, and outputs a value in the interval [0,1]. In these networks, the output is typically thresholded to generate a binary label, but we found it could also be used to approximate a measure of confidence in the network's classification. By treating this value as a confidence metric, we are able to input a large unlabeled dataset into the classifier and then trust the labels that were assigned a "high confidence" and leave the remainder unlabeled.

After training the network, the performance of the four classifiers (reproductive, budding, flowering, fruiting) achieved 85-90% accuracy compared to expert-labeled data. However, as described above, the real value of these approaches comes from their prospects for labelling previously unlabeled data, thus helping to replace expensive and time-consuming human labor. We then applied our confidence-interval-based approach to a collection of 600k images and were able to label 35-70% of the samples with a chosen confidence threshold of 95%. In other words, we could then use the high-confidence labels and simply not automatically label the remaining unclassifiable samples. The data from these samples could then be labeled manually, or, if appropriate, not labeled at all.

Keywords

machine learning, image classification, herbaria, plant phenology

Presenting author

Jonathan Koss

Presented at

TDWG 2021

Conflicts of interest