Data Science Capstone
Project Proposal
Jackson Crum

Identifying Plant Species in Images with Deep Learning

Problem Statement

The United States federal government has a major legal responsibility to protect endangered species, threatened species, and critical habitats. Once a species becomes listed as "threatened" or "endangered", it receives special protections by the federal government under the Endangered Species Act. The goal of the Endangered Species Act is to make species' populations healthy and vital so they can be taken off the endangered/threatened species list.[1] Ideally, new technologies for monitoring these species would be incorporated to make the conservation efforts as efficient as possible.

Though some technologies such as motion sensor camera traps and drone/satellite imagery have been incorporated into the efforts, extracting information from these pictures remains a tedious manual task conducted solely by highly-trained and specialized botanists.[2] Large amounts of available images and data are often ignored or overlooked due to lack of manpower and resources. In response to this problem, this research will focus primarily on saving conservation resources by developing an automated or semi-automated system for identifying species in images and providing conservationists with a simple application for the overall identification and sorting process.

Background

The Endangered Species Act of 1973 requires federal agencies to ensure that any action they authorize does not threaten any listed endangered species or result in the damage or destruction of critical habitats of those species. Section 2(a)(2) states that it is "the policy of Congress that all Federal departments and agencies shall seek to conserve endangered species and threatened species and shall utilize their authorities in furtherance of the purposes of this Act."[3] The

---

[1] "Endangered Species | National Wildlife Federation." The National Wildlife Federation. Accessed September 20, 2018. https://www.nwf.org/Educational-Resources/Wildlife-Guide/Understanding-Conservation/Endangered-Species.

[2] Norouzzadeh, Mohammad Sadegh, et al. "Automatically Identifying, Counting, and Describing Wild Animals in Camera-trap Images with Deep Learning." *Proceedings of the National Academy of Sciences* 115, no. 25 (November 17, 2017).

[3] U.S. Congress. Senate. *Endangered Species Act of 1973*. HR 37. 93rd Cong., 1st sess. Introduced in Senate June 12, 1973. https://www.fws.gov/international/pdf/esa.pdf.

Department of Defense (DoD) has the highest density of endangered species listed as threatened or endangered of any other federal agency, with 340 out of 420 large military installations requiring active conservation management plans to protect 492 listed species (274 animals, 218 plants). Expenditures on these conservation plans total more than $800 million annually.[4]

Related Research

Species detection and recognition are still difficult, time-consuming, and expensive challenges and researchers have yet to produce a method that provides an efficient solution for all situations. Research conducted by faculty of the Electrical Engineering Department at University of Zilina compared a CNN model to well-known image recognition methods, including PCA, LDA, and SVM models. It was concluded that the CNN outperformed all conventional models and produced a significant increase in accuracy.[5]

A joint university study tested various state-of-the-art neural network models, including AlexNet, GoogLeNet, VGG, and ResNet models, on Serengeti wildlife images. The study showed that deep neural networks performed well on the wildlife dataset, although performance was worse for rarer animals, and saved 99.3% on manual labor while maintaining human-level identification accuracy. While neural networks have proven successful in the task of image recognition, they rely on large datasets and perform worse on small training datasets.[6]

Trends and future directions of automated plant species identification were reviewed by biogeochemists and software engineers from the Max Planck Institute for Biogeochemistry and Technische Universität Ilmenau. CNN models with 6, 17, and 26 layers were used to classify the Flavia dataset and produced accuracies of 94.69%, 97.9%, and 99.65%, respectively. However, the Flavia dataset is a constrained set of leaf images taken against a white background and

[4] United States. Department of Defense. Natural Resources. By Peter Boice. February 2013. Accessed September 2018. http://www.dodnaturalresources.net/files/te__s_fact_sheet_2-21-13.pdf.

[5] Trnovszky, Tibor, Patrik Kamencay, Richard Orjesek, Miroslav Benco, and Peter Sykora. "Animal Recognition System Based on Convolutional Neural Network." *Advances in Electrical and Electronic Engineering* 15, no. 3 (September 2017). Accessed September 2018. doi:10.15598/aeee.v15i3.2202.

[6] Norouzzadeh, et al. "Automatically Identifying, Counting, and Describing Wild Animals in Camera-trap Images with Deep Learning."

without any stem, meaning natural images may result in worse accuracy. Despite intensive research on automated plant species identification, very few studies have resulted in approaches that can be used by the general public. The most popular current applications for planet identification still suffer from inabilities to deal with cluttered backgrounds (LeafSnap) or low top-5 accuracy (Pl@ntNet).[7]

Methodology

The data will come from Kaggle's plant seedlings classification contest. The data consists of 12 species of 5544 total images pre-split into training and testing datasets (4750 training and 794 testing images).

Image exploratory data analysis will be the initial phase of the project. The variation in image sizes will be explored as choosing an appropriate image size that compromises feature detection for training time is vital. Class distribution will be examined for uniformity and data from classes with less data will be oversampled.

The next step of this project is to preprocess the images. Image preprocessing techniques will be used to enhance relevant image features and diminish distracting features. The primary tool of image preprocessing for this project will be the Open Source Computer Vision Library (OpenCV). Foreground extraction using the GrabCut algorithm will be used to extract the leaves from the rocky background. Based on a preliminary examination of the images, shape is the primary differentiator between species with color being the secondary differentiator. Canny edge detector will be used with gaussian blurring to extract leaf shape and histogram equalization will be used to enhance color contrast. All preprocessing methods will be tested individually and in sequence to find the best combination of techniques.

Transfer learning will be employed to test state-of-the-art machine learning models, including VGG-16 and ResNet-50 neural networks. Convolutional neural networks (CNN) will be developed and compared using Keras TensorFlow backend and trained/tested with varying

---

[7] Wäldchen, Jana, Michael Rzanny, Marco Seeland, and Patrick Mäder. "Automated Plant Species Identification—Trends and Future Directions." *PLOS Computational Biology* 14, no. 4 (April 05, 2018). doi:10.1371/journal.pcbi.1005993.

preprocessing methods, number of layers and parameters, and confidence thresholds. The models will be evaluated with classification accuracy as the primary metric while also calculating recall, precision, and F1 score. The models will use a final sigmoid activation function with a cross-entropy loss function to produce top-1 and top-5 class probabilities. Amazon Web Services (AWS) and Google Cloud Platform (GCP) will be utilized for processing power to make these models feasible to train.

The application will be a Flask app that incorporates a JavaScript frontend, Python/Keras backend, and MongoDB for image data storage. The application will allow the user to upload a file of images, run the images through the neural network model, and output image predictions. The images will be sorted into folders without the uploaded parent directory based on class prediction. The application will display basic statistical information about the predictions and store all metadata and prediction data in a MongoDB for later statistical analysis. Depending on model metrics and top-1 accuracy, the application will either be used for automatic identification or provide top-5 prediction probabilities to the user for quick manual identification.

End Product

The goal of this project is to develop a full-stack application that will incorporate a machine-learning algorithm and sorting application to allow for the efficient identification of plant species in wildlife photos for United States government conservation agencies. This application will allow conservationists to simply upload files of images to the app, run the images through the algorithm, sort the files based on model output, and store image data in a MongoDB collection for quantitative analysis. This will allow non-technical researchers and conservationists to easily incorporate machine learning solutions to their problems. The hypothesis is that the minimal viable product will increase the effectiveness of conservation efforts and save the DoD in resource expenditure and man hours.

Proposed Timetable
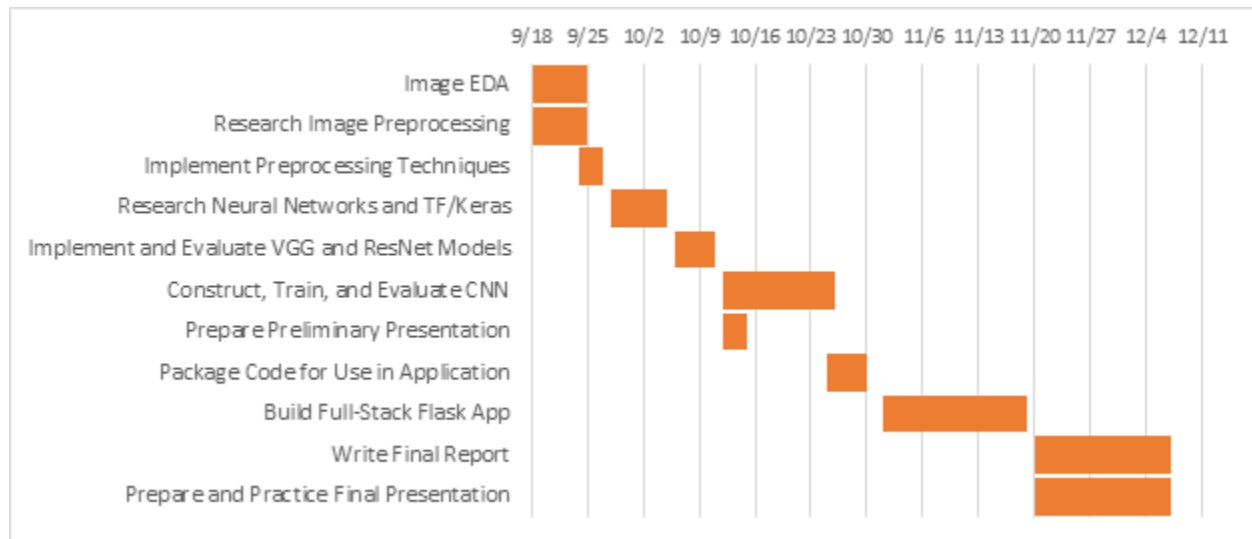
References

Norouzzadeh, Mohammad Sadegh, Anh Nguyen, Margaret Kosmala, Alexandra Swanson, Meredith S. Palmer, Craig Packer, and Jeff Clune. "Automatically Identifying, Counting, and Describing Wild Animals in Camera-trap Images with Deep Learning." *Proceedings of the National Academy of Sciences* 115, no. 25 (November 17, 2017). Accessed September 2018. doi:10.1073/pnas.1719367115.

Trnovszky, Tibor, Patrik Kamencay, Richard Orjesek, Miroslav Benco, and Peter Sykora. "Animal Recognition System Based on Convolutional Neural Network." *Advances in Electrical and Electronic Engineering* 15, no. 3 (September 2017). Accessed September 2018. doi:10.15598/aeee.v15i3.2202.

United States. Department of Defense. Natural Resources. By Peter Boice. February 2013. Accessed September 2018. http://www.dodnaturalresources.net/files/te__s_fact_sheet_2-21-13.pdf.

U.S. Congress. Senate. *Endangered Species Act of 1973*. HR 37. 93rd Cong., 1st sess. Introduced in Senate June 12, 1973. https://www.fws.gov/international/pdf/esa.pdf.

Wäldchen, Jana, Michael Rzanny, Marco Seeland, and Patrick Mäder. "Automated Plant Species Identification—Trends and Future Directions." *PLOS Computational Biology*14, no. 4 (April 05, 2018). doi:10.1371/journal.pcbi.1005993.