Classifying Wildlife Images using the iNaturalist Dataset

Springboard | Capstone 2 Milestone Report #1

By: Lauren Broussard

Business Problem

There are millions of different types of wildlife in the world. Can we use image classification to distinguish some of them? iNaturalist, a site that allows users to add images of wildlife they observe, hosts an annual Kaggle competition that encourages participants to help improve image classification for their repository of images. Successful image classification in the natural world has many benefits -- for instance, it can help conservationists and zoos to identify species in the wild without disturbing their habitat, or to identify potential endangered species that are being traded. Paired with geolocation data, image classification of wildlife can also display where certain species are typically found in the world in order to help chart migration patterns or the existence of invasive species. Image classification can also have applications to social media and accessibility, in helping to label images appropriately for screen readers or add appropriate tags to an image.

Sample images from iNaturalist image files.

This project uses deep learning to read 265,213 jpeg images from 1010 different species and 6 wildlife categories in various backgrounds. While the original Kaggle competition sought to classify one species from another - for instance, a bee vs a yellowjacket, this project attempts to classify each image into its appropriate “iconic category” - Amphibians, Birds, Fungi, Insects, Plants, or Reptiles.

Potential Client(s):

Zoos, Wildlife Conservatories: It would be useful to be able to quickly classify images of the animals they work with and/or come across. It might also be useful for wildlife protection agencies who are scanning online databases for a particular type of animal that might be being sold, kept, or traded.

Other Clients: It could be useful for social media managers, who can use the image classification for quick tagging of their posts, or for sites to more quickly label an image for a screen reader for additional accessibility.

Data Collection and Wrangling

Dataset and Images

Data and images were downloaded from the iNaturalist competition page on the Kaggle website. The iNaturalist site allows users to upload their own images of wildlife they see. According to the iNaturalist overview on Kaggle, the images in this case were collected and validated by multiple users from the iNaturalist site. The images were part of the Fine-Grained Visualization Categorization workshop, FGVC6, and includes fewer species that may be harder to classify - as it may be an image of a green frog against other green leaves.

The original dataset included the following files and descriptions, as described by the competition documentation:

File descriptions

- **train_val2019.tar.gz** - Contains the training and validation images in a directory structure following {iconic category name}/{category name}/{image id}.jpg .
- **train2019.json** - Contains the training annotations.
- **val2019.json** - Contains the validation annotations.
- **test2019.tar.gz** - Contains a single directory of test images.
- **test2019.json** - Contains test image information.
- **kaggle_sample_submission.csv** - A sample submission file in the correct format.

Source: Kaggle iNaturalist 2019 FGVC6 Competition, Data

For the purposes of this project, all data was downloaded, but the following files were primarily used to approach the classification problem:

- **train2019.json, val2019.json (57MB, 816.46KB)**: includes annotation information about each of the images, including the path to the image file, pixel height and width information, and encoded species information - i.e. kingdom, phylum, class, etc.

- **train_val2019 (79GB)**: contained in a larger train_val2019.tar.gz file, this folder includes all of the wildlife image data, stored as a jpeg file, and separated in folders by high level category (Plants, Amphibians, etc.) then species which is stored as a numeric code.

Data Wrangling & Preprocessing

Data was imported and prepared in Python primarily using pandas. Though not always the case, this data came in largely standardized, so did not require as much munging as in previous projects. However, the following steps were taken to import and prepare the data for modeling.

*Loading Data***.** Annotations from the *train2019.json* file had the following keys: 'info', 'images', 'licenses', 'annotations', 'categories'. Two DataFrames were created using data in 'images' and 'annotations', and the rest of the keys in the json file were ignored. For the images dataframe, the following fields were included: file name, height, and width of the image, as well as the image id. The annotations DataFrame included: image id and category_id, which corresponded to a code denoting the species of the image. The two DataFrames were then merged.

*Adding/Removing Columns***.** To be more descriptive, the column 'category_id' was changed to 'species_id'. Additionally, since we were interested in categorizing images based on their "iconic category" name (i.e. Plants, Insects, etc.), an additional field called "wildlife_type" was created to store that information. The values for the *wildlife_type* labeling were pulled from the subfolders in the "file_name" field.

```
df.head()
```

image_id	file_name	height	width	species_id	wildlife_type
0	train_val2019/Plants/400/d1322d13ccd856eb4236c...	800	600	400	Plants
1	train_val2019/Plants/570/15edbc1e2ef000d8ace48...	533	800	570	Plants
2	train_val2019/Reptiles/167/c87a32e8927cbf4f06d...	600	800	167	Reptiles
3	train_val2019/Birds/254/9fcdd1d37e96d8fd94dfdc...	533	800	254	Birds
4	train_val2019/Plants/739/ffa06f951e99de9d220ae...	600	800	739	Plants

*Missing Data or Duplicates***.** There were no duplicated rows, or missing data in the annotation files.

*One Hot Encoding***.** One Hot encoding was used with the *get_dummies* method in pandas to change categorical variables to binary values before putting them into the model.

image_id	file_name	height	width	species_id	wildlife_type_Amphibians	wildlife_type_Birds	wildlife_type_Fungi
0	train_val2019/Plants/400/d1322d13ccd856eb4236c...	800	600	400	0	0	0
1	train_val2019/Plants/570/15edbc1e2ef000d8ace48...	533	800	570	0	0	0
2	train_val2019/Reptiles/167/c87a32e8927cbf4f06d...	600	800	167	0	0	0
3	train_val2019/Birds/254/9fcdd1d37e96d8fd94dfdc...	533	800	254	0	1	0
4	train_val2019/Plants/739/ffa06f951e99de9d220ae...	600	800	739	0	0	0

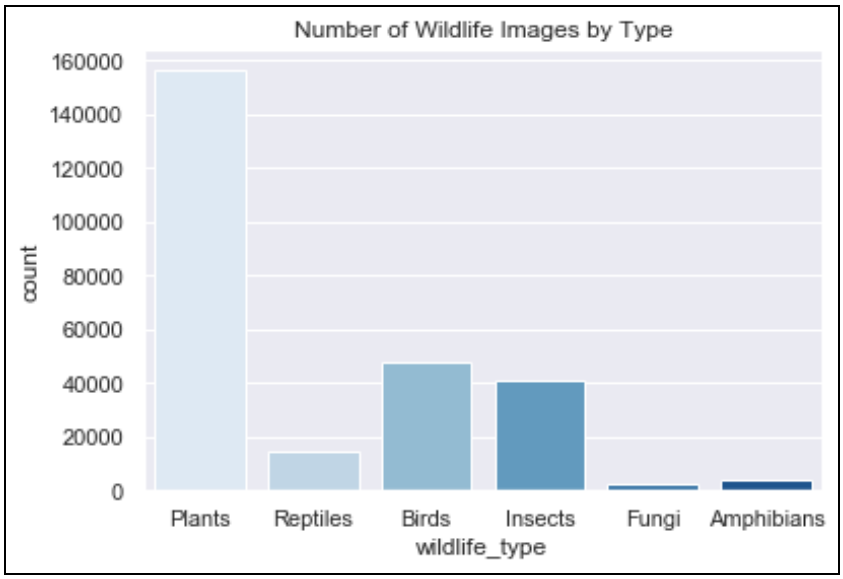
The final dataset includes **265,213 images and data.**

Exploratory Data Analysis

With the data prepared, we can look at features of the dataset and images. Overall, there are 265,213 images in our dataset, 1010 different species, and 6 different wildlife types.

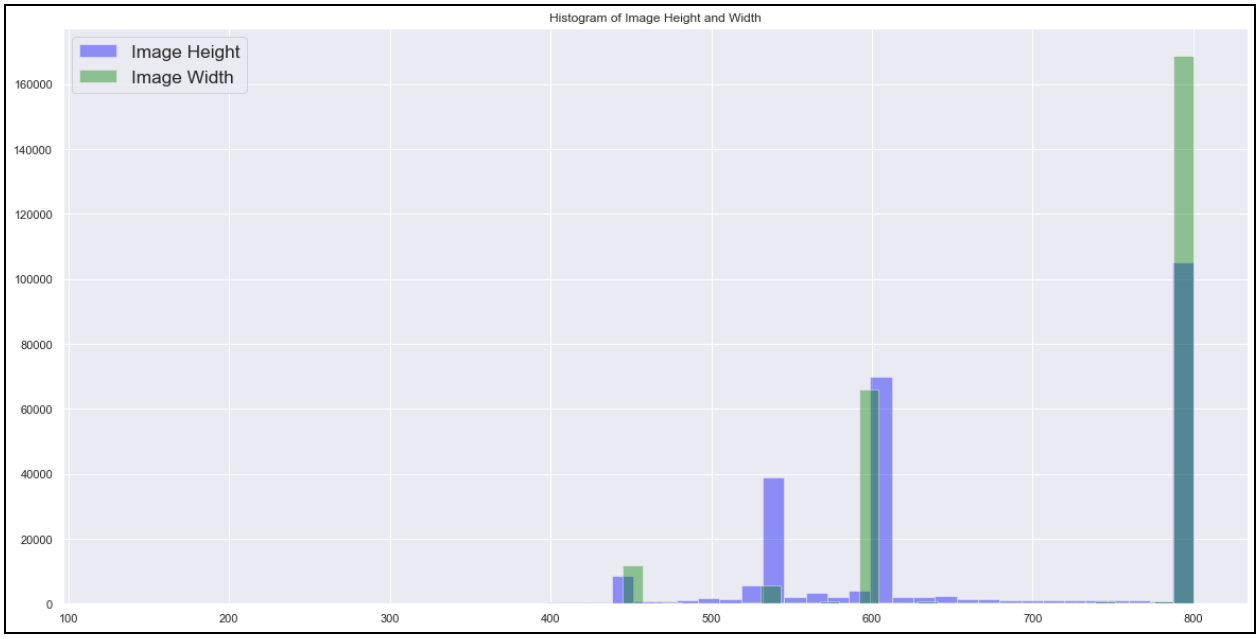
Wildlife Types & Species

The vast majority of images in the dataset (nearly 160,000) are of Plants, followed by Birds, Insects, then Reptiles and Amphibians, and finally Fungi. The column species_id is a value denoting the species of the image. As expected since it has the largest number of images, the Plants wildlife type has the largest number of different species - there are 682 different species of plants included.



Pixel Height and Width

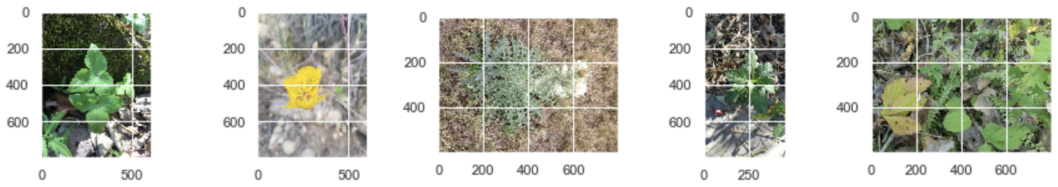
Each image has a maximum height and width of 800 pixels, with more than 10,000 The average height is approximately 663 pixels, while the average width is 720. Images in the dataset tend to be wider (horizontal) than they are taller (vertical).



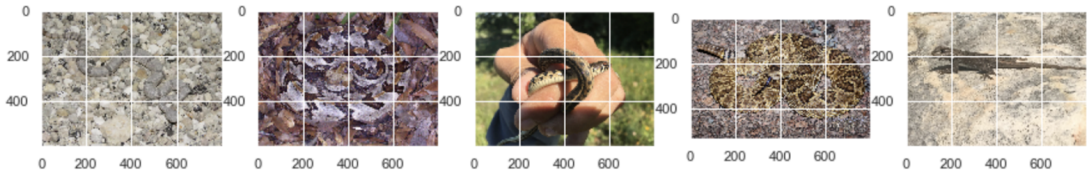
Sample Images

Below are sample images showing five jpeg images of each of the six categories. Many images are obscured in some way by other things, for instance, the amphibians are in water or are partly camouflaged in plants (another one of our categories). Being able to train a model to can distinguish between the two will be useful.

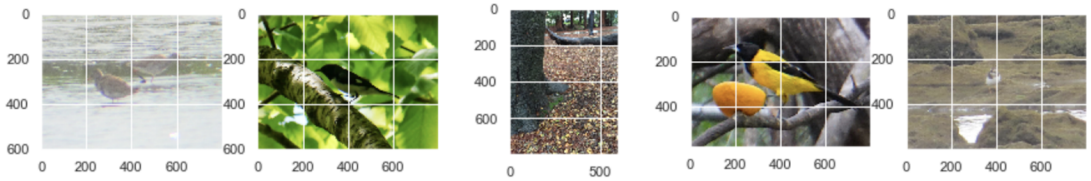
Plants



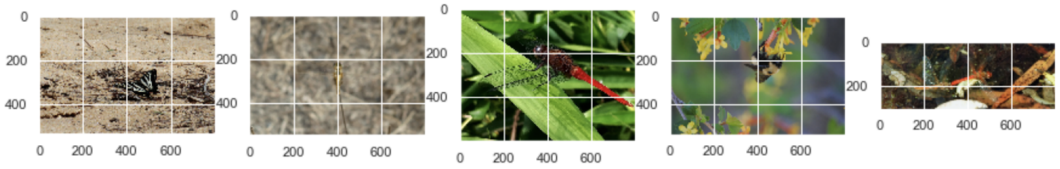
Reptiles



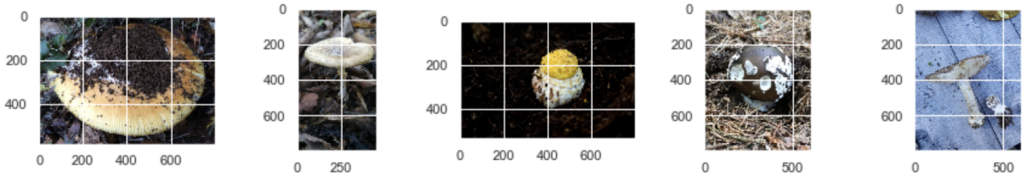
Birds



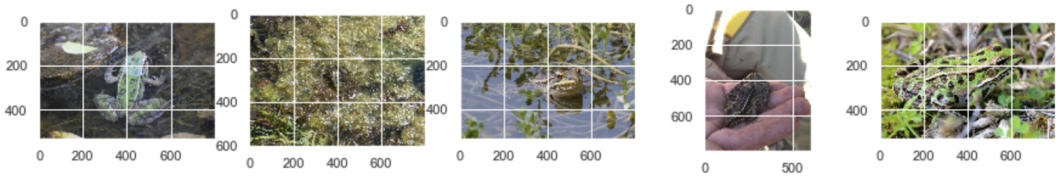
Insects



Fungi



Amphibians



Resources

iNaturalist: https://www.inaturalist.org/

iNat2019 Starter Keras Code: https://www.kaggle.com/ateplyuk/inat2019-starter-keras-efficientnet

Kaggle Competition Information: https://www.kaggle.com/c/inaturalist-2019-fgvc6/data

LA County Threatened and Endangered Species:
https://data.lacounty.gov/Sustainability/LA-County-Threatened-and-Endangered-Species-2018-/7uw4-g37f