Machine Learning Engineer Nanodegree

Capstone Proposal

Madhankumar B
(madhankumarbk@gmail.com)
December 24th, 2018

Proposal

Abstract

The project proposal is to create an ML model to accurately predict the Honey Bee Subspecies for the given honey bee image. The proposed model is going to classify the subspecies of the honey bee based on the training over 5000 already existed images. Once the model is trained, it is going to predict the subspecies of honey bee for the given image. This model is very useful to find out what kind of honey bees are still existed and its features.

Domain Background

Every third bite of food relies on pollination by bees. At the same time the honey bee hive losses are also increasing day by day. So we must take a steps to protect the honey bees. But, by looking at the honey bees and distinguishing its subspecies and take a steps to protect a particular species is time consuming. In order to make at least a one step from this easier machine learning models comes into play in order to differentiate species of honey bees by looking at the images. Deep learning is the subfield of machine learning which is widely used to classify image to recognize the object's in the images based on the large set of images by which it is trained on. Deep learning is one of the most promising field in machine learning where already much development has taken place and is currently used in real world application.

This project is an inspiration taken from how an image of honey bee can be used to expedite the hive process, **how well the images can be used to recognize the species of honey bees** and also how can we improve hive process through the images of honey bee. The main part of this project is highlighted above.

Problem Statement

Nowadays, the honey bee hives are becoming expedite. But, pollination by bees are more important for food production so in order to save the honey bees government has taken an initiative to save honey bees from extinction, in order to do that government has to find the sub species of the particular honey bee from which it belongs which pave the way for saving honey bee.so the government has asked the machine learning engineers to develop an algorithm which accurately predicts the subspecies of honey bee given its image.

Datasets and Inputs

The dataset contains 2 folders with information necessary to make a prediction.
They are: 1. bee_datas.csv 2.bee_imgs.zip

Bee_datas.csv contains all the information about the bee images the details are as follows,
1.file - File name in bee_imgs folder.
2.Date - Date of video captures.
3.time - Time of day of video capture (military time).
4.location - Location (city, state, country).
5.zip code - Zip Code to numerically describe location.
6.subspecies - Subspecies of Apis mellifera species.
7.health - Health of a bee.
8.pollen_carrying - Presence of pollen on the bee's legs.
9.caste - Worker, Drone, or Queen bee.

bee_imgs.zip contains 5172 images of honey bee required for training.

The dataset above can be categorized into 5 classes with 5 different subspecies the details are as follows,

Classes	Number of Images
-1	428
1 Mixed local stock 2	472
Carniolan honey bee	501
Italian honey bee	3008
Russian honey bee	527
VSH Italian honey bee	199
Western honey bee	37

Here the -1 represent for those images whose subspecies we didn’t know.
Based on the above details Italian honey bee every species comes around same range so it will not create much problem to the balance of the dataset so as of now I am considering this as a balanced dataset if it impacts result of our final model then we will change the performance metric.

The averages values for the images sizes are between 50 and 100 pixels and the extreme values are from few pixels to approximately 500. We will scale all images to 100 x 100 pixels.

The dataset is publically available via kaggle and accessible.
Data source: The Bee Image Dataset: Annotated Honey Bee Images
Apis mellifera with location, date, health, and more labels

Reference: https://www.kaggle.com/jenny18/honey-bee-annotated-images

Solution Statement

For this problem we are going to follow a traditional machine learning approach. Firstly, we are going to preprocess the data and analyze the data's and images with an Exploratory Data Analysis(EDA). we are going to find the correlation between the features and find the relationship between each images. Then, we are going to build a CNN model either from already developed model like VGG, resnet, etc., or we are going to build our own model (which is not yet decided) to map images to its subspecies based on the training set of images. Once the baseline model is built based on the training error and accuracy, also based on the validation error and accuracy we are going to improve the model to yield a maximum accuracy and minimum error. All this above steps follows a deep learning approach which is integral part of machine learning.

Benchmark Model

For this model the benchmark model will be a Logistic regression model to classify different classes or a basic CNN model with one layer to classify the subspecies with its precision, recall and f1 score. I am going to use Multi-layer CNN model with its improvement techniques to give maximize the precision, accuracy and f1 score for this classification task.

Evaluation Metrics

The evaluation metrics for this problem would be a precision, recall and f1 score for each subspecies of honey bee to find which variety of honey bee is accurately predicted by the built model.

Project Design

First method would be preprocessing the data in order to bring the data stable and more suitable for learning. Then, we are going to spilt our training and testing dataset. Finally, we would use deep learning using Keras to build a baseline model. Then the model will be improved based on the accuracy and loss we get i.e., our goal is to maximize the accuracy and minimize the error and that will be the final model to submit.

*Tools and Libraries used: Python, Jupiter Notebook, pandas, scikit learn,
matplotlib, tensor flow, Keras. Other libraries will be added if necessary.*

References

[1] Kaggle, " honey-bee-annotated-images".
https://www.kaggle.com/jenny18/honey-bee-annotated-images

[2] Deep Learning Wikipedia page
https://en.wikipedia.org/wiki/Deep_learning