Machine Learning Engineer Nanodegree

Capstone Proposal

Vijaybhasker Pagidoju

May 29, 2018

Proposal

Domain Background

Ecological balance is one of the crucial and important part of our lives.Plants and animals take a major part in maintaining a balanced ecosystem, But there are a few species, which threaten the habitats.Identifying these specific invasive species out of the many species which help the trees to grow is important to safeguard the nature.

This is a problem,where I see a prominence of making use of Machine learning Algorithms to process identification of those specific species which threaten the habitats at a much faster pace and with a very low relative cost.Solving this problem is important, to protect our ecosystem and saving experts/scientists time spent to identify those species with manual intervention.

Problem Statement

Tangles of kudzu overwhelm trees in georgia while cane toads threaten the habitats in quite a number of countries.These are only two species which can really have a damaging effect on the environment,the economy and human health.Identifying these invasive species out of all the species is prominent and currently, this task is taken care by expert scientists,who visit these areas to identify the species inhabiting that environment,which is time consuming ,expensive and continuous sampling in this way can lead to enormous time.

Machine learning Algorithms using image classification can be used to identify these species.With the evolution of pre-trained models especially in image classification in ImageNet dataset,we can make use of these pre-trained models and the dataset provided with Transfer Learning and get a better accuracy in predicting the images of those specific invasive species out of the many helpful species.

Datasets and Inputs

The Dataset contains 2 zip files with training and testing images and 2 csv files, one consisting of the labels for the training images and the other is a sample submission file format for the testing labels.

train.7z: The training set consisting of 2295 images taken in a Brazilian national forest.In some of the pictures there is Hydrangea,a invasive species that has to be identified.

train_labels.csv: The csv file consisting of the labels for the training set.

test.7z: The testing set consisting of 1531 images,used to test our algorithm to identify the invasive species among all the testing images.

Sample_submission.csv: A sample submission file in the right format to store the labels identified for the testing set of images when applied our algorithm to identify.

Solution Statement

A Deep Learning Model making use of Convolution Neural Networks can be used to come to a solution. The Model will be trained on the train data and train labels as provided after preprocessing. The model will then be tested on the testing dataset as provided and the predicted labels will then be stored in the submission.csv.

Entire Dataset will be evaluated and then as training entire dataset will require high Gpu ,a lot of time and expensive, a random dataset from the training and test sets will be considered for the solution.

Evaluation Metrics

Submissions are evaluated on Area under ROC Curve between the predicted probability and observed target. The Prediction will be made on the images in the test set i.e., whether the image contains an invasive species at the utmost probability.

Benchmark Model

As the project is part of Kaggle Competitions, The Benchmark model is one of the top scores on the public scoreboard for this competition with 0.99770(area under ROC Curve).An Attempt will be made to score amongst the top 50% in the public leaderboard submissions.

Project Design

The Training and testing datasets are first evaluated for pre-processing.

From the problem statement, it is clear that it is an image classification problem. So for the project, we will make use of Convolutional Neural Networks in TensorFlow/Keras and make use of the ImageNet Dataset.

The training of the model can be made much faster by making use of appropriate pre-trained models available in Keras and that can be integrated to the model training using Transfer Learning.

After training the model, the model is evaluated on the testing images to evaluate the prediction model’s accuracy and then based on the results observed, the model will further go through tuning to predict with a better score.

References

https://www.kaggle.com/c/invasive-species-monitoring

https://www.kaggle.com/c/invasive-species-monitoring#evaluation

https://www.kaggle.com/c/invasive-species-monitoring/data

https://keras.io/applications/