Machine Learning Engineer Nanodegree

Capstone Proposal

Haaris Jalal
2nd January 2018

Domain Background

Seafood is one of the primary sources of proteins in most of the sea neighboring nations. However, illegal fishing over the years has destroyed the marine ecosystem. Over fishing by giant companies have also cost many local anglers their jobs and source of livelihood.

To combat this problem, ***The Nature Conservancy***, decided to incorporate different monitoring devices such as cameras on fishing boats. The Nature Conservancy is a nonprofit organization looking out for restoring the marine ecosystem. Though the organization is monitoring the fishing, there are now hours of raw footage, which is unsorted. This issue is time consuming and tedious for the organization.

The organization turned towards Kaggle users to create an algorithm that could detect the specific species of fish among the photos taken. This could help in identifying which species of fish is in danger of extinction.

The aim of this project is to build a ***Convolutional Neural Network*** to classify different species of fish.

Problem Statement

According to the rules of the Kaggle competition, the aim of the project is to classify the photographs of the fish according to the given classes. The eight different classes provided in the dataset are:

- Albacore Tuna
- Bigeye Tuna
- Yellowfin Tuna
- Mahi Mahi
- Opah
- Sharks
- Other (meaning that there are fish present but not in the above categories)
- No Fish

Note: *On an important note, all the fish in the training set have been classified and placed in folders. Each folder contains one specific type of fish. Other small fish may also be present which were probably there as bait for fishing. We will ignore the bait for this task.*

Goal

Our goal is to predict the likelihood of fish species in each picture of the test dataset using deep learning methodologies.

Dataset and Inputs

To generate the dataset, The Nature Conservancy compiled hours of footage and then sliced the different videos into approximately 5000 images. The images capture different angle of boats and fish. The labeling of the dataset images is a result of classifying the fish according to the respective categories.

The following folders are available for this analysis.

1. Train Folder: Comprises of eight different folders of sorted images.
2. Test Folder: Comprises of randomized images.

Solution Statement

Deep Learning methodologies have been very useful in image recognition and classification. In this project, I have decided to use Convolutional Neural Networks with transfer learning to classify the images in proper categories. Depending on the analysis, I will look into different transfer learning techniques such as VGG-16 and ResNet50.

Benchmark Model

Before addressing the problem at hand, I need to develop a method that I can use as a benchmark. Since, the accuracy using the deep learning techniques is a lot better compared to other algorithms, I plan to use supervised learning techniques for establishing a benchmark of the model. At the present, I plan to use different algorithms such Gaussian Naïve Bayes and Decision Trees for establishing a benchmark for image classification. Once the log loss is calculated using these techniques, I will use deep learning methodologies to improve the accuracy and decrease the log loss ratio of the model.

Evaluation Metrics

For this project, I will be evaluating my model using the multi-class logarithmic loss. The formula for log loss evaluation is:

$$logloss = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij}\log(p_{ij})$$

N *is the number of images in the test set.*

M *is the number of image class labels.*

log *is the natural algorithm.*

y_{ij} is 1 if the $obervation$ i $belongs$ to $class$ j and 0 $otherwise$.

p_{ij} is the $predicted$ $probability$ $that$ the $observation$ i $belongs$ to $class$ j

Multiclass logarithmic loss classifier penalizes incorrect predictions. As an example, if the class label is ***"1"*** and the prediction is ***"0"***, then the log loss will have a high value. On the other hand, if the class label is ***"1"*** and the predicted label is ***"1"***, then the log loss will have a smaller value. Our main aim is to minimize the log loss function as much as possible keeping in mind the capability of the processor.

Project Design

Programming Language: Python 3

Libraries: Keras, Tensorflow, Numpy, Pandas, OpenCV

Workflow:

- Establish a baseline using supervised learning methods such as Gaussian Naïve Bayes and Decision Trees.
- Establish a CNN from scratch.
- Extract the features of the image using a pre-trained network and calculate the predictions of the classes.
- Fine-tune the network to decrease the log loss value.

References:

- Kaggle leaderboard (https://www.kaggle.com/c/the-nature-conservancy-fisheries-monitoring/leaderboard)