Capstone Project Proposal

By

Akpojicheko Eyekpegha

Domain background

The capstone project, which is gotten from the kaggle competition- **The Nature Conservancy Fisheries Monitoring**, uses computer vision to solve a problem in the area of aquatic science and oceanography. As a fisheries graduate and data annotator, I am interested in completing this programme with the use of amazon sagemaker. According to Knausgård et al (2021) and Rathi et al(2018), fish classification is very useful especially in the wild. For instance, when large fishing vessels harvest fish, a significant amount of the harvested stock may be untargeted in the first place. Over the years, there have been a lot of misuse and wastage of aquatic resources, since these fishing vessels find it difficult to recognise targetted fish in deep parts of the ocean or sea

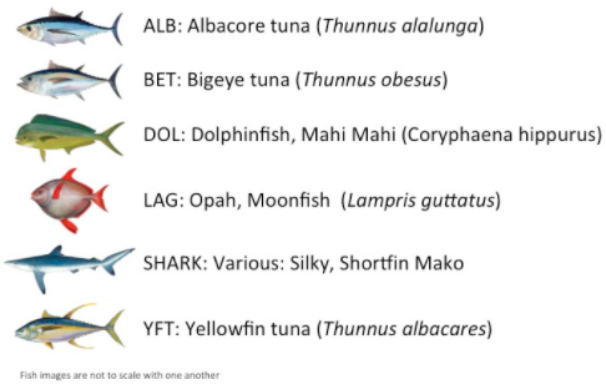
Problem statement

On kaggle, the competition is organized by the Nature conservancy with the aim of finding better ways to monitor unregulated fishing activities that are threatening marine ecosystems and global seafood supplies. This further poses challenges to aquatic resource management. It is a classification problem, that is, the need to classify fish species. With the dataset provided by Kaggle (the input), my goal is to develop algorithms that would automatically classify species of fish (basically the fish species stated in the kaggle competition)

Datasets and inputs

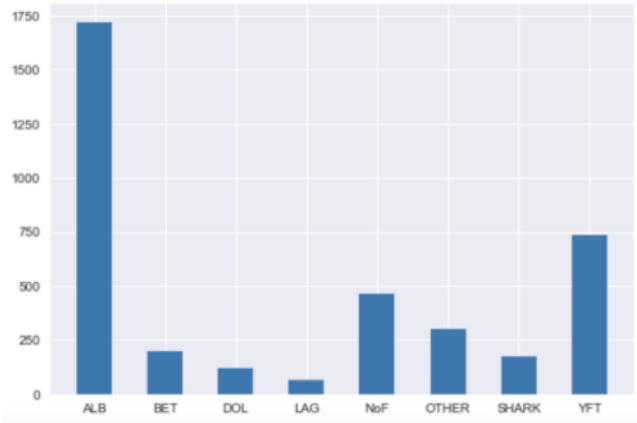
The dataset is made available by Kaggle, assessed by the link https://www.kaggle.com/c/the-nature-conservancy-fisheries-monitoring/data It consists of train images (3792 images), and 2 test datasets for 2 stages of training. They are coloured images with inconsistent dimensions. Available in this dataset are eight classes to predict from. The Eight target categories comprise six fish species (in the image below) with two other classes - other (fish species other than the targetted six classes of fish) and No Fish (meaning that no fish is in the picture). To start up the

project, I would unzip and upload these files to my s3 bucket to make them available for training with sagemaker. I will conduct some feature engineering where necessary to get better results. The train images would be used for the training job, after which testing is done twice for the 2 test datasets



This photo is gotten from the kaggle competition.

1) Class Distribution is Highly Unbalanced



According to Felix (2017), the graph above shows the class distribution of all 8 classes of the kaggle dataset. The dataset is highly unbalanced with ALB (Albacore tuna) having a huge percentage in the train dataset. Data augmentation would be done on the train dataset with pytorch's torchvision to make it balanced.

Solution statement

The proposed solution is to create a deep learning model that is able to identify accurately the presence of fish (basically six fish species) with a high percentage accuracy by using amazon sagemaker. This solution can enable countries to better allocate human capital to management and enforcement activities, instead of fish classification.

Benchmark model

As I work on solving this problem, the benchmark I hope to bit is the model accuracy of 90 percent. This target should be achievable with pre-processing of train dataset, hyperparameter tuning and other feature engineering that would be discovered during the project.

Evaluation metrics

The approach to evaluate the performance of this project is by getting its accuracy with the confusion matrix. The confusion matrix has 4 categories: true positives, true negatives, false positives, and false negatives which is used to determine the efficiency of the model. My aim is to achieve an high accuracy after completing this project

Project design

The workflow to approaching the solution is as follows:
With the right exploratory data analysis, I will explore the data to understand, prepare and clean up the data (if necessary), making it easy to train the model. Since it is an unbalanced dataset, pre-processing technique such as data augmentation would be done thereby enhancing the model's performance to obtain the desired solution.
After specifying the sagemaker environment, I will commence training the model with a RESNet18/50 or VGG-16 (whichever has a lower cross entropy loss) and a python script as entry point, creating a pytorch estimator, setting hyperparameters(the learning rate and batch size) before calling model.fit(). Sagemaker debugging and profiling will also be done on the model. The model will be finally deployed to an endpoint.

Reference

Knausgård, K.M., Wiklund, A., Sørdalen, T.K. *et al.* Temperate fish detection and classification: a deep learning based approach. *Appl Intell* (2021). https://doi.org/10.1007/s10489-020-02154-9

Dhruv Rathi, Sushant Jain and Dr. S. Indu, Underwater Fish Species Classification using Convolutional Neural Network and Deep Learning(2018) https://arxiv.org/ftp/arxiv/papers/1805/1805.10106.pdf

Felix Yu (2017), Detect and Classify Species of Fish from Fishing Vessels with Modern Object Detectors and Deep Convolutional Networks (2017)
https://flyyufelix.github.io/2017/04/16/kaggle-nature-conservancy.html