Machine Learning Engineering Nanodegree

Capstone Proposal

Priyanka Patel
June 24, 2018

Domain Background

Whaling is the practice or industry of hunting and killing whales for their oil, meat and whalebone. Whaling has been happening for centuries and while the whale population tries to recover, they are faced with further challenges like struggling to survive in warmer oceans and competing with the fishing industry for food. They also are negatively impacted by noise pollution, trash, and accumulating toxins from the fish they consume.

To aid whale conservation efforts, scientists use photo surveillance systems to monitor ocean activity. They use the shape of whales' tails and unique markings found in footage to identify what species of whale they're analyzing and meticulously log whale pod dynamics and movements. For the past 40 years, most of this work has been done manually by individual scientists, leaving a huge trove of data untapped and underutilized.

Source: https://arxiv.org/pdf/1604.05605.pdf

Problem Statement

The goal of this project is to build an algorithm that will be used for identifying whale species in images by analyzing Happy Whale's database of over 25,000 images, gathered from research institutions and public contributors.

Datasets and Inputs

The data can be downloaded by following this link:
https://www.kaggle.com/c/whale-categorization-playground/data

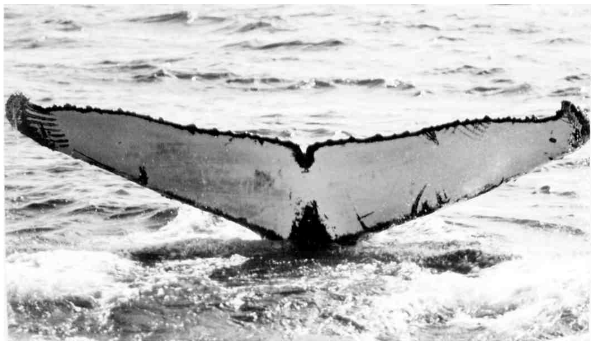
The training data contains thousands of images of humpback whale flukes. Individual whales have been identified by researchers and given an ID. Using the data of image inputs, the goal is to predict the whale ID of images in the test set.
The following is the description of files provided by Kaggle:

File descriptions:

- train.zip - a folder containing the training images
- train.csv - maps the training Image to the appropriate whale Id. Whales that are not predicted to have a label identified in the training data should be labeled as new_whale.
- test.zip - a folder containing the test images to predict the whale Id
- sample_submission.csv - a sample submission file in the correct format

The training set is comprised of 9851 images and the test set of 15611 images. The images are of whale flukes, which are the two lobes of the tail. There are over 3000 types of whale IDs that will be used to classify the data. With so many IDs, it seems that class distribution would be widespread, with each ID matching to only a few images.

These are some examples of the input data:



Solution Statement

The algorithm for this project will be built using deep learning. A CNN will be implemented in Tensorflow/Keras due to its ability to scale large models. A series of convolution layers can be implemented to extract features from the images with pooling between the layers for spatial reduction. A dropout layer should be added to prevent overfitting, a dense layer with relu to reduce the likelihood of vanishing gradient and another dense layer with softmax for probability distribution.

Benchmark Model

The Kaggle leadership board for the Humpback Whale Identification Challenge will be used as the benchmark model. The current leader has a score of 0.78563.

Evaluation Metrics

This algorithm will be evaluated using the Mean Average Precision @ 5:

$$MAP@5=1U\sum_{u=1U}\sum_{k=1min(n,5)}P(k)$$

where U is the number of images, P(k) is the precision at cutoff k, and n is the number of predictions per image.

Project Design

The final model will be designed having multi-layer architecture that consists of alternating convolutions and nonlinearities followed by fully connected layers and ending with a softmax classifier for probability distribution.

As discussed earlier, the class distribution may be too widespread given the number of IDs leading to an imbalanced class problem. To counter this, it will be necessary to do some image transformation. The goal of that is to minimize the class distribution.

The images in the Datasets and Inputs sections show that not all images are the same shape and some are in color while others are black and white. To combat shape, the images will have to be scaled to the highest ranking sizes and to combat color a series of convolution layers will be implemented for feature extraction as well as using transfer learning using a pre-trained model. The model will be trained on bottleneck layers of a pre-trained model. Because the test set is larger than the training set, the test set will be split in two sets – test set and validation set.