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.

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 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

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.

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:

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

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.

The following steps will be followed:

Step 1: Import Dataset
Step 2: Explore Dataset
Step 3: Create a CNN to identify whale flukes using transfer learning
Step 4: Develop Algorithm
Step 5: Test Algorithm