Machine Learning Engineer Nanodegree

Capstone Proposal

Abdalrhman Ahmed Ismail

October 7th, 2018

Humpback-Whale-Identification Kaggle competition

Proposal

Domain Background

After centuries of intense whaling, recovering whale populations still have a hard time adapting to warming oceans and struggle to compete every day with the industrial fishing industry for food.

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.

the challenge is to build an algorithm to identifying whale species in images. By analyzing Happy Whale's database of over 25,000 images, gathered from research institutions and public contributors. I will try to help to open rich fields of understanding for marine mammal population dynamics around the globe.

Happy Whale is the provider of these data and problem. Happy Whale is a platform that uses image process algorithms to let anyone to submit their whale photo and have it automatically identified
Papers related to this type of problem:
-https://hal.archives-ouvertes.fr/hal-01373777/document
-https://arxiv.org/ftp/arxiv/papers/1604/1604.05605.pdf

Problem Statement

The dataset of this problem contains thousands of images of humpback whale flukes. Individual whales have been identified by researchers and given an Id. the challenge is

to build an algorithm to identifying whale species in images. What makes this such a challenge is that there are only a few examples for each of 3,000+ whale Ids.

Datasets and Inputs

The dataset is provided as a competition at Kaggle , it contains the following files :

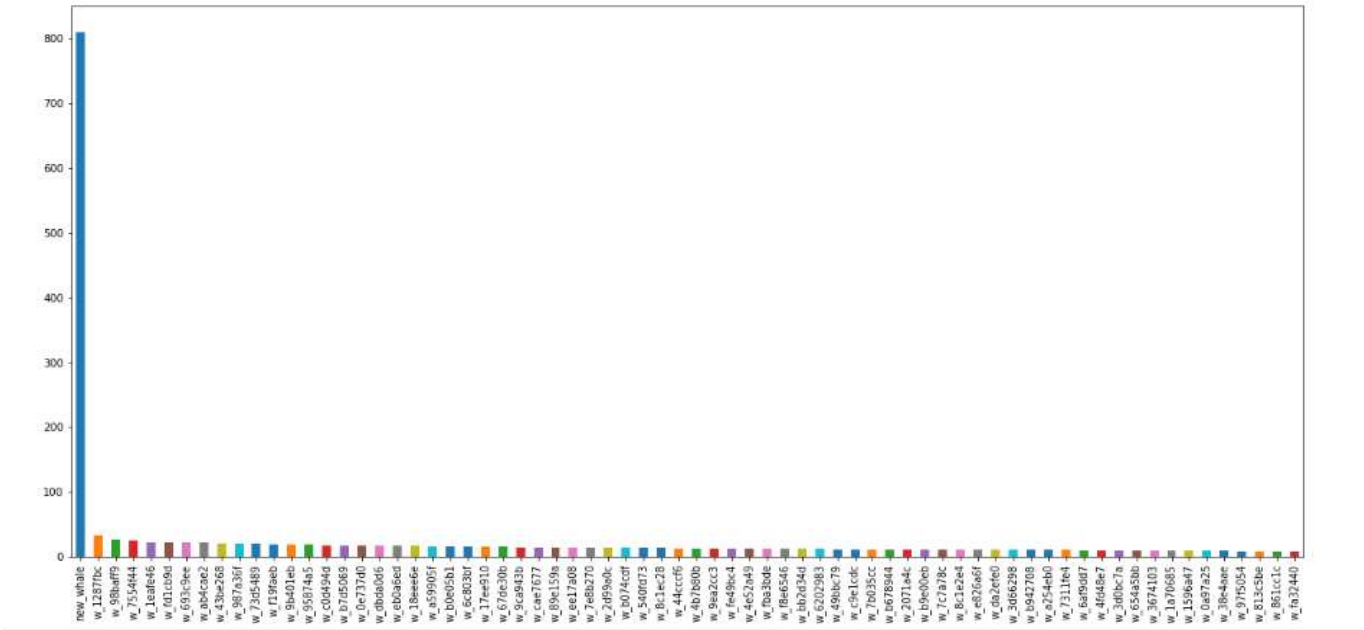
- train folder: containing 9850 training images of humpback whale flukes.

- test folder: containing 15610 test images to predict the whale Id .

- 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

The images are colored and gray ,but their dimensions are not consistent. it will be preprocessed to fixed size.

By visualizing the data we see that dataset are quite balanced , except the 'new whale ' class are most dominant .



Since the testing data provided by Kaggle doesn't come with labels, I will split the training data to create custom testing set. I'll use the I will use StratifiedShuffleSplit to guarantee classes balance in test / train sets. And test the benchmark and my model on the generated test set.

Solution Statement

I will try to train CNN to obtain a model to predict the whales' ids.

Benchmark Model

I will begin with simple CNN as a benchmark (2 convolution layers).

Evaluation Metrics

- *I will use the Accuracy to test my model*

- The final solution can be measured by the Mean Average Precision @ 5 (MAP@5)

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

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

Project Design

Workflow

- Download the dataset
- provide some basic analysis on the dataset
- Preprocessing the data . (resizing images , converting all images to grayscale)
- Image augmentation ; (I will try to do this ,as there is some of classes have few training images)
- Build CNN and feed it with training data , there is no architecture in my mind , but I will begin with 2 convolution layer (the benchmark) then add more .
- try to use transfer learning like (Resnet50 , VGG16) , and compare the results
- Applying the final model on the data .