Machine Learning Engineer Nanodegree

Capstone Proposal

Abhishek Ravindran

December 15th, 2018

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 analysing 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

An algorithm to identify individual whales by the images of their tail has to be developed. We'll analyse the Happywhale's database of over 25000 images gathered from research institutes and public contributors. The algorithm will predict the probability of an image being of a particular whale breed. Maximum of 4/5 whale suggestion can be given for any of the image with the corresponding probability likelihood.

Kaggle link: https://www.kaggle.com/c/humpback-whale-identification/data

Datasets and Inputs

The dataset consist of images of different whale's tail obtained from the Kaggle competition "Humpback whale identification". The training data consists of thousands of images of humpback whale flukes. Individual whales have been identified by researchers and given an Id. There are almost 3000+ whale ids.

The following are the file description and their corresponding download link:

- train.zip – This zip folder contains the training images. The model will be trained using the same.

- train.csv – This CSV file comprises the mapping between the images and the whale id (label). Whales that are not predicted to have a label identified in the training data will be labelled as ‘new_whale’.
- test.zip – This zip folder consists of test images which the model would use to predict its corresponding label.
- sample_submission.csv – A sample submission file with the corresponding image file name and its Id.

https://www.kaggle.com/c/humpback-whale-identification/download/train.zip

https://www.kaggle.com/c/humpback-whale-identification/download/train.csv

https://www.kaggle.com/c/humpback-whale-identification/download/test.zip

Solution Statement

I am planning to use the Convolutional neural network (CNN) to do image recognition or classification. Convolution is the first layer to extract feature from the input image. CNN will be implemented in Tensorflow/Keras and will be optimized to minimize the multi-class logarithmic loss. Predictions will be made on the test data set once the model has been trained.

Benchmark Model

Attempt will be made to get a score that is among the top 50 – 60 % of the public leader board submissions.

Evaluation Metrics

Submissions are evaluated according to the Mean Average Precision.

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

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

CNN will be used to design the model. It will have multiple dense layers as well. As a matter of fact instead of creating the whole model from scratch, transfer learning approach can be used to decrease the intense training time required. We will take the CNN that has already been trained on image net database and our dense layers/output layer can be added on top of the same and then the total model can be trained using the training dataset images. This can give a better accuracy in the end.