Udacity Machine Learning Nanodegree Capstone Proposal

Humpback Whale Identification

Can you identify a whale by its tail?

Vikas Varma

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.

There have been research done on identifying a whale using photos[1], which uses whale pictures which is similar to this effort, just that it uses the actual whale pictures.

I chose this dataset as it looked interesting and it allowed me to focus more on the deep-learning techniques, and my interest in image classification.

Problem Statement

The challenge is to identify individual whales in images given the image of its tail fin. We will analyze Happywhale's database of over 25,000 images, gathered from research institutions and public contributors.

In this kaggle dataset, it has train.csv, train.zip and test.zip. But for the scope of this project, I will not be using the test.zip (as there is no way to validate the results). I will use the train.csv and train.zip (train folder after extracting the archive to a folder) and split it into training and test datasets.

Datasets and Inputs

The image dataset is part of the Kaggle Competition Humpback Whale Identification. This dataset consists of the following:

- train.csv - file which has mapping of training image to the appropriate whale Id. The whales not identified are labelled as new_whale. This train dataset will be split into training, validation and testing dataset using 80-20 split.

Image	Id
0000e88ab.jpg	w_f48451c

[1] https://www.researchgate.net/publication/327910789_Applying_deep_learning_to_right_whale_photo_identification

0001f9222.jpg	w_c3d896a
00029d126.jpg	w_20df2c5
00050a15a.jpg	new_whale

- train.zip - This is a zip file of all the training, testing images. There are around 25361 image files in this zip file. All the images are jpeg, RGB format files of various sizes. This means that I will have to rescale the images during data preprocessing.

Solution Statement

The main objective of this project will be to use Deep learning techniques to classify the whale image to a whale Id. At first I will run the training data through Tensorflow and Keras using CNN. Once that is completed I will try to improve the model by improving the parameters or use transfer learning techniques.

Benchmark Model

The plan is to implement a simple CNN model using tensorflow and keras, measure the MAP@5[2] score and make this the benchmark model based on the training data. The MAP5 score algorithm is described in the kaggle competition evaluation criteria. Once the benchmark model is baselined, will try to improve upon it by using the Transfer learning techniques using the Resnet, VGG16 and InceptionV3 architecture.

Evaluation Metrics

Given that there are 5005 whales (unidentified Id new_whale), as the dataset classes are unbalanced we cannot use test accuracy as our evaluation metrics. In the kaggle competition they are using MAP@5 for the same reason. We are going to use as our evaluation metrics in this project.

Project Design

Language:

Python 3

Libraries:

numpy, tensorflow, keras, pandas, scipy, matplotlib, tqdm, etc.

[2] MAP@5 - https://www.kaggle.com/c/humpback-whale-identification/overview/evaluation

Design workflow:

The main tasks in this project is as follows:

- Data Analysis
    - Load library and data
    - Explore the training data
    - Identify the data points, in this case its image and whaleId.
- Data preprocessing
    - Transform categorical data into numerical data. In this case the Id field.
    - Create training, validation & test data with a 80-20 split for the training to testing data set.
    - Images loaded, rescale images
- Model training & evaluation
    - Build a baseline model
        - Build a baseline model using CNN using tensorflow and keras with the initial training dataset.
        - Calculate the MAP score and make this the baseline model.
    - Improve the model
        - Use transfer learning techniques using Resnet, VGG16 and InceptionV3 architecture and use data augmentation
        - Calculate the MAP score for each of the improved models.
        - Choose the appropriate model based on MAP score and suitability.

References

1. Applying deep learning to right whale photo identification - https://www.researchgate.net/publication/327910789_Applying_deep_learning_to_right_whale_photo_identification
2. https://www.kaggle.com/c/humpback-whale-identification
3. https://happywhale.com
4. Identifying Individual Humpback Whales - https://nmsstellwagen.blob.core.windows.net/stellwagen-prod/media/archive/sister/pdfs/sbnms_fs_id_2011_1.pdf
5. Who’s That Whale? Your Photo Could Help I.D. a Humpback, Douglas Fox- https://www.nationalgeographic.com/adventure/adventure-blog/2016/05/04/whos-that-whale-your-photo-could-help-i-d-a-humpback/
6. Mean Average Precision - https://towardsdatascience.com/breaking-down-mean-average-precision-map-ae462f623a52