Udacity Capstone Proposal:
Humpback Whale Identification – Can you identify a whale by its tail?

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 [https://www.kaggle.com/c/humpback-whale-identification]. Scientists and conservationists continue to monitor whale populations, often manually taking photographs. Citizen science projects such as happywhale.com can help provide additional insight and increase the volume of sighted instances again providing images, locations, dates, and times where whales are being spotted. This information is highly beneficial for monitoring the ongoing wellbeing of these creatures. Different species of whale can live for significantly long times if environments are stable, for example humpback whales can have a life expectancy of up to 80 years. [https://www.afsc.noaa.gov/nmml/education/cetaceans/humpback.php]. As such it is important that the same whales can be repeatedly monitored over many years. Whale measurements such as species, numbers, pod dynamics and movements are important to the research aiding the efforts of whale recovery.

Problem Statement

For the past 40 years the means of identifying a whale has relied on scientists checking images by hand and comparing to those already identified.

Because of this manual effort, there is a large backlog of untapped data which is unavailable simply because of the time needed to identify a whale by hand. If there was a way to detect if a whale has been identified previously or if the whale is a new identification, this would save many man hours of work and allow the larger volumes of information at citizen science sites such as happywhale.com to be of greater use.

The most common method of identification is by looking at the whales' flukes (tail).
The image below shows the anatomy of a humpback whale's flukes.

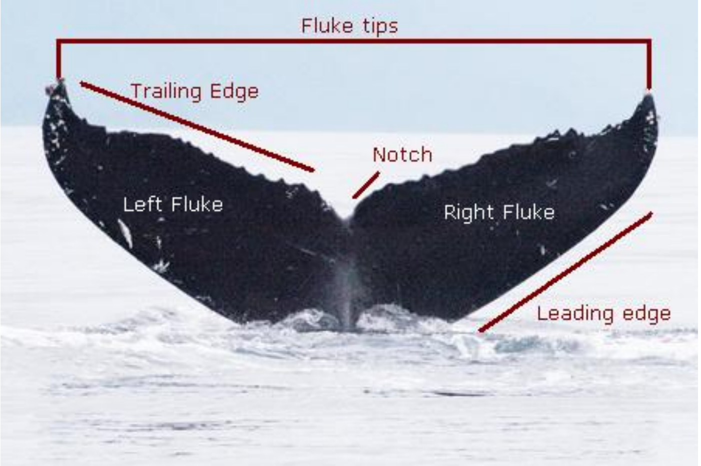


Fig 1 Humpback Whale flukes [http://www.alaskahumpbacks.org/matching.html]

The Humpback Whales of South-eastern Alaska website (http://www.alaskahumpbacks.org/matching.html), contains some interesting insight into how to identify a whale. A fluke's shape does not drastically change unless the whale has had an injury, and so this is the primary focus area scientists use when identifying a whale.

There are a number of features that can be checked to identify a whale successfully.

- One of the most important features to check is the trailing edge as the shape across the top does not significantly change, especially with adults unless damage occurs.
- The notch does not change shape, some whales of the same species can have large notches, while others have smaller.
- The white underside of the flukes can have pigment markings which do not significantly change. However not all humpback whales have a white underside. For example, with the Alaskan humpback whale, over 70% have no white pigment and the flukes are dark on both sides.

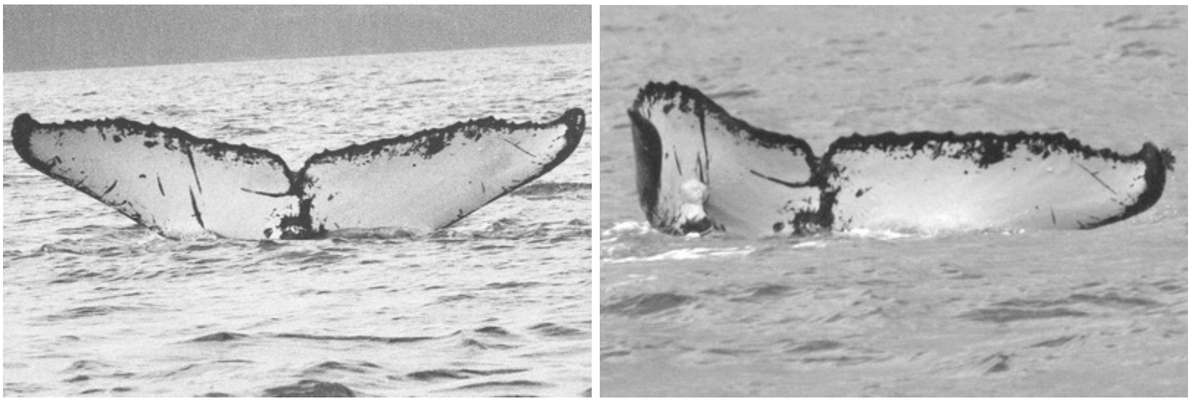


Fig 2. Sample whale spotting prior to 2005 and in 2005 note the pigment marking, notch shape and right fluke trailing edge can be compared [source http://www.alaskahumpbacks.org/matching.html]

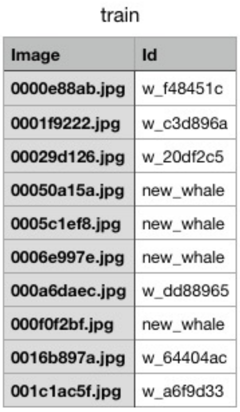
There are some problems with identification too. A younger whale calf or yearling can be difficult to identify. This is because they often have indistinct pattern markings which have not yet formed, or the trailing edge shapes and notch will change as the whale grows. As such it is much easier identifying adult whales.

Dataset

Kaggle as of 25th January 2019 has a dataset of humpback whales which has been collected from happywhale.com [https://www.kaggle.com/c/humpback-whale-identification/data]
The data is broken into four sets as follows:

- **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 dataset contains over 25,000 images of whales of varying file sizes and quality in jpg format. The train.csv file consists of two columns showing the filename of each of the images in the train set and a unique whale identification Id if the whale is known, or new_whale if the whale is unknown



train

Image	Id
0000e88ab.jpg	w_f48451c
0001f9222.jpg	w_c3d896a
00029d126.jpg	w_20df2c5
00050a15a.jpg	new_whale
0005c1ef8.jpg	new_whale
0006e997e.jpg	new_whale
000a6daec.jpg	w_dd88965
000f0f2bf.jpg	new_whale
0016b897a.jpg	w_64404ac
001c1ac5f.jpg	w_a6f9d33

Fig 3. Sample data from train.csv

There is also a list of images in a test folder which are part of the Kaggle competition that can be submitted but do not have a results csv, and so can't be used in testing the validity of the detection model. This would have to be done by splitting the training data accordingly.

Proposed solution

The initial proposal will be to evaluate the use of a Convolutional Neural Network for image classification to predict if the image is one of a previously identified whale or of a new whale. Whether this will be via an existing model or one which is manually created will be results dependant. The final layer I would initially expect to be a softmax layer showing the probability of the whale being one of those already identified.

As shown in the problem statement, there are a number of features which can be used to identify a whale. The training images may need to be pre-processed to help with training. For example, vertical rotation (to view a whale fluke from front/back), viewing only one fluke, scale invariance, rotational invariance, adjusting shear of images. This may assist since the same whale photograph will not always be taken from the same position and distance. Similarly, it can be used to increase the training dataset size. Since humpback whales are predominately of various shades of grey, we should additionally be able to reduce the images to greyscale to reduce data dimensionality and increase efficiency.

The component which I suspect will need some thought is that of detecting a 'new_whale'. A new whale could match many different types of non-identified whales. If 'new_whales' are removed from the training set, then how many classifiers should we return on the final CNN layer and at which level should we be confident that the image is of a new whale?

Benchmark and Evaluation Metrics

Once the model is complete, we can compare this against a base benchmark of randomly choosing one of the existing whale identification labels, and additionally by choosing ‘new_whale’ each time. The model would be expected to perform significantly higher than these two methods.

The Kaggle competition referenced for Humpback Whale Identification explicitly lists how they would evaluate submissions. Here it is stated they would evaluate according to the Mean Average Precision @ 5 (MAP@5)

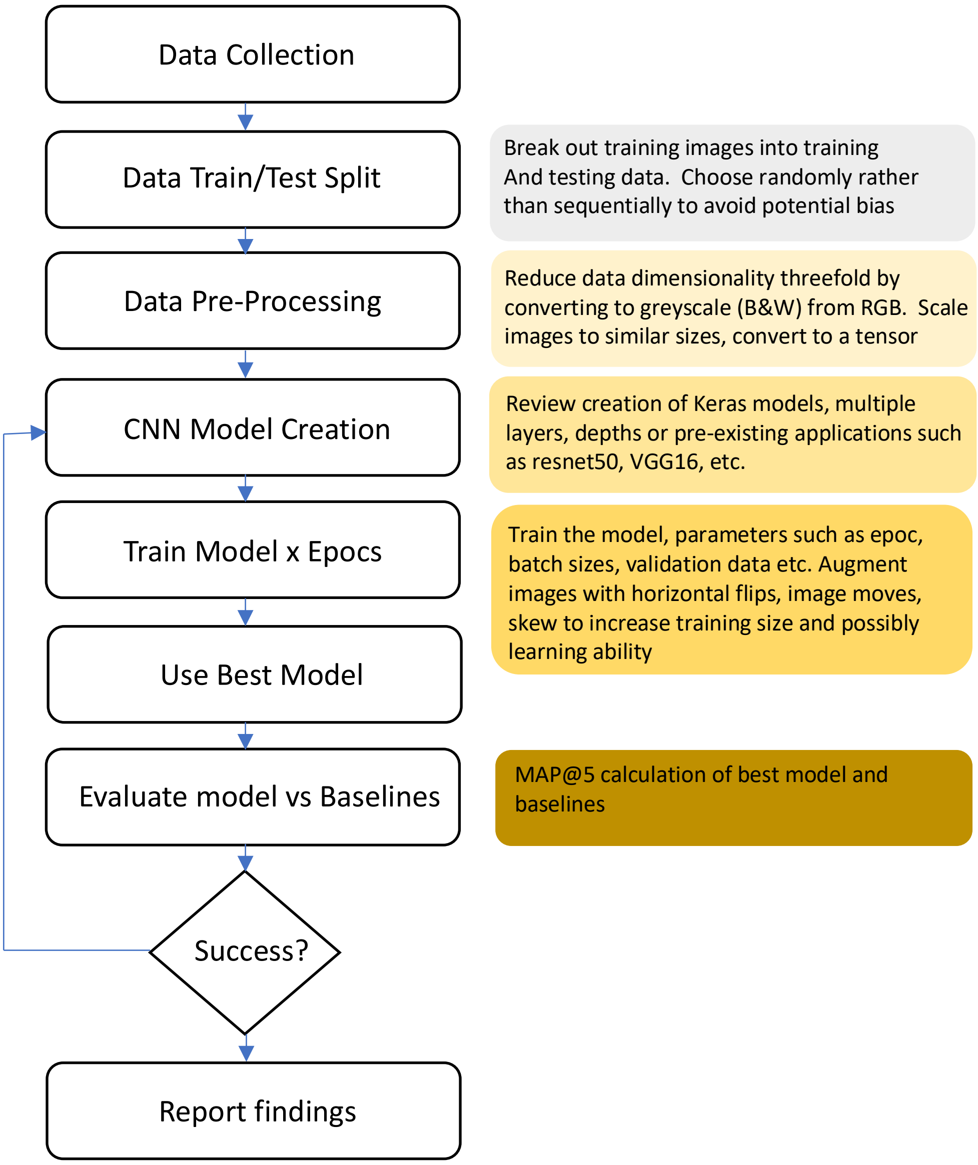
‘where U is the number of images, $P(k)$ is the precision at cutoff k, n is the number predictions per image, and $rel(k)$ is an indicator function equaling 1 if the item at rank k is a relevant (correct) label, zero otherwise. ‘, [https://www.kaggle.com/c/humpback-whale-identification#evaluation]

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

For the random, ‘new_whale’ and the CNN prediction model the results will be evaluated against this same evaluation metric.

Project Design

The flow diagram below shows the planned workflow for the proposal.



Break out training images into training And testing data. Choose randomly rather than sequentially to avoid potential bias

Reduce data dimensionality threefold by converting to greyscale (B&W) from RGB. Scale images to similar sizes, convert to a tensor

Review creation of Keras models, multiple layers, depths or pre-existing applications such as resnet50, VGG16, etc.

Train the model, parameters such as epoc, batch sizes, validation data etc. Augment images with horizontal flips, image moves, skew to increase training size and possibly learning ability

MAP@5 calculation of best model and baselines