mlnd: Capstone Proposal

Naruhiko Nakanishi

February 19th, 2019

[**Humpback Whale Identification**]

Kaggle Competition is used for a technical domain along with the problem and dataset. [1]

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.

Problem Statement

The challenge is to build an algorithm to identify individual whales in images. Happywhale's database is analyzed. Happywhale is a platform that uses image process algorithms to let anyone to submit their whale photo and have it automatically identified.[2] The database is over 25,000 images, gathered from research institutions and public contributors. The contributing is helpful to open rich fields of understanding for marine mammal population dynamics around the globe.

Datasets and Inputs

This training data contains thousands of images of humpback whale flukes. Individual whales have been identified by researchers and given an Id.

The challenge is to predict the whale Id of images in the test set. What makes this such a challenge is that there are only a few examples for each of 3,000+ whale Ids.

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

Solution Statement

Deep learning, specifically a convolutional neural network (CNN) which is very effective at finding patterns within images, is used toward the solution.

Benchmark Model

In the benchmark, transfer learning such as MobileNet architecture is used. Transfer learning involves taking a pre-trained neural network and adapting the neural network to a new, different data set.

MobileNets are based on a streamlined architecture that uses depthwise separable convolutions to build light weight deep neural networks.

Evaluation Metrics

The results are are evaluated according to the Mean Average Precision @5 (MAP@5):

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

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.

Once a correct label has been scored for an observation, that label is no longer considered relevant for that observation, and additional predictions of that label are skipped in the calculation.

Project Design

Firstly the dependencies are installed. Keras is used for a deep learning library. Keras is a high level neural networks API.

The second step will be data collection, data exploration and visualisation to understand the fundamental characteristics of the dataset. This training data contains thousands of images of humpback whale flukes. Individual whales have been identified by researchers and given an Id.

The next step is to build the models in Keras for performance and flexibility. Deep learning, specifically a CNN which is very effective at finding patterns within images, is used toward the solution. A CNN is created to identify individual whales in images.

In the benchmark, transfer learning such as MobileNet architecture is used. Transfer learning involves taking a pre-trained neural network and adapting the neural network to a new, different data set.

The next step is Hyperparameter tuning. It is computationally feasible to tune the parameters.

The final step is to train the model on the entire training set and evaluate the performance. The final performance will be calculated against the test data set provided by Kaggle. The results are evaluated according to the Mean Average Precision.

References

[1] https://www.kaggle.com/c/humpback-whale-identification.

[2] Happywhale. https://happywhale.com/home.