**Machine Learning Engineer Nanodegree**

**Capstone Proposal**

Viswanath Ravindran  
April 24th, 2018

**Proposal**

One of the primary reasons for me to take up the Udacity Machine learning Nanodegree is that it would provide me with a platform to help provide a platform to introduce me to Advanced Machine Learning. More precisely Computer Vision problem being part of the curriculum excited me even further. My knowledge and understanding of Markov Decision Process and Computer Vision has leaped several folds after going through. Considering my passion that I have acquainted for Computer Vision I would like to also do my Capstone project on this specific area. I have chosen the Humpback Whale Identification Challenge @ [link](https://www.kaggle.com/c/whale-categorization-playground/kernels)

**Domain Background**

Solving Computer Vision Problem has been explored by several researchers in the recent years. Udacity’s introduction to the CIFAR-10 dataset as a problem for the Deep learning project introduced me to several interesting attempts made by researchers globally. There are several research papers published on it with considerations that it is one of the most difficult problems to be solved for today. I would want to do my bit by understanding it even further and help contributing towards a noble cause.

**Problem Statement**

Problem Statement derived from Kaggle @ [link](https://www.kaggle.com/c/whale-categorization-playground/kernels)

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.

In this competition, I am challenged to build an algorithm to identifying whale species in images. I will analyze Happy Whale’s database of over 25,000 images, gathered from research institutions and public contributors. By contributing, I will help to open rich fields of understanding for marine mammal population dynamics around the globe.

How this problem is important to me now.

I strongly believe in Data Science for a noble cause and this challenge which will span for the next 3 months will help me work towards the same. It will also further deepen my understanding on the computer vision problem. The intricate complications involved here will also hone my skills in approaching the problem of the same domain.

**Datasets and Inputs**

I will analyze Happy Whale’s database of over 25,000 images, gathered from research institutions and public contributors. By contributing, you’ll help to open rich fields of understanding for marine mammal population dynamics around the globe.

The datasets details are available @ [link](https://www.kaggle.com/c/whale-categorization-playground/data)

The input train data has several image of the whale’s fluke and the corresponding identification numbers. The test data has unlabeled picture which is required to be processed and label them as accordingly. This is an extreme Multi Label Classification problem.

**Solution Statement**

This specific problem can be approach in the following manner:

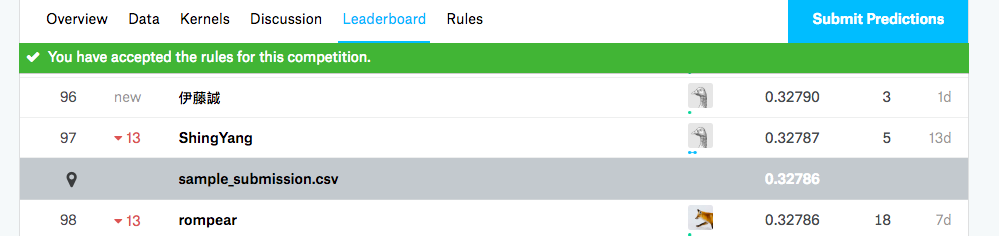
1. Input the training images and the labels for the training image is available in the necessary format. Images are available in a folder and a single file containing the Labels.
2. We can create an initial model as a baseline model using external pre-trained models like ReseNet, VGG16 etc.
3. Using the baseline model and applying various transformation to the image we then can train an inference model to help understand the differences in the predicted vs true lables.
4. We can then fine-tuning the top layers of a pre-trained network

I would possibly using ImageDataGenerator for real-time data augmentation, layer freezing and model fine-tuning to improve my predictions. The final submission file should predict the top 5 predicted class against each input image.

**Benchmark Model**

The benchmark model that can be used to see how I perform could be used using the sample submission here. The details of the metric that is used to assess the Model performance is given in detail in the next section. The performance is designed to indicate a better performance with a higher score. Screenshot attached below shows the Benchmark score submitted using the sample by the host organization of this problem listed in the leaderboard page for the competition @ [link](https://www.kaggle.com/c/whale-categorization-playground/leaderboard)

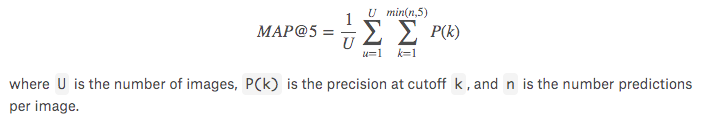
As of the time that I wrote this proposal the base submission is 0.32786. I would focus on improving the score better than this.



**Evaluation Metrics**

The evaluation metric for this challenge is called the Mean Average Precision score at K. The k selected here is 5, The output prediction should simple output the 5 most possible class labels of the several available class.

Its use is different in the field of Information Retrieval and popular among various Multi Label Classification problems setting. The mathematical representation is available as in kaggle [link](https://www.kaggle.com/c/whale-categorization-playground#evaluation)



**Project Design**

Based on

In this final section, summarize a theoretical workflow for approaching a solution given the problem. Provide thorough discussion for what strategies you may consider employing, what analysis of the data might be required before being used, or which algorithms will be considered for your implementation. The workflow and discussion that you provide should align with the qualities of the previous sections. Additionally, you are encouraged to include small visualizations, pseudocode, or diagrams to aid in describing the project design, but it is not required. The discussion should clearly outline your intended workflow of the capstone project.