**Machine Learning Engineer Nanodegree**

**Capstone Project**

The Nature Conservancy Fisheries Monitoring

Detection and Classification of Fish Species

Prasand Kumar   
February 2, 2018

**I. Definition**

**Project Overview**

Nearly half of the world depends on seafood for their main source of protein. In the Western and Central Pacific, where 60% of the world’s tuna is caught, illegal, unreported, and unregulated fishing practices are threatening marine ecosystems, global seafood supplies and local livelihoods. [The Nature Conservancy](http://www.conserveca.org/tuna) is working with local, regional and global partners to preserve this fishery for the future.



Currently, the Conservancy is looking to the future by using cameras to dramatically scale the monitoring of fishing activities to fill critical science and compliance monitoring data gaps. Although these electronic monitoring systems work well and are ready for wider deployment, the amount of raw data produced is cumbersome and expensive to process manually.

The Conservancy is inviting the Kaggle community to develop algorithms to automatically detect and classify species of tunas, sharks and more that fishing boats catch, which will accelerate the video review process. This will have a positive impact on conservation and our planet.

The aim of this project is to build a Convolutional Neural Network (CNN) that can classify the distinct species of fishes. To tackle the computational constraints, we will be using transfer learning technique.

**Problem Statement**

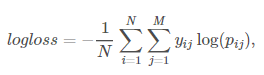
In this project, The Nature Conservancy asks you to help them detect which species of fish appears on a fishing boat, based on images captured from boat cameras of various angles.

The goal is to predict the likelihood of fish species in each picture.

Eight target categories are available in this dataset: Albacore tuna, Bigeye tuna, Yellowfin tuna, Mahi Mahi, Opah, Sharks, Other (meaning that there are fish present but not in the above categories), and No Fish (meaning that no fish is in the picture).

**Metrics**

The model is evaluated using the [multi-class logarithmic loss](https://www.kaggle.com/wiki/MultiClassLogLoss). Each image has been labeled with one true class. For each image, you must submit a set of predicted probabilities (one for every image). The formula is then,



where N is the number of images in the test set, M is the number of image class labels, log is the natural logarithm, yij is 1 if observation i belongs to class j and 0 otherwise, and pij is the predicted probability that observation i belongs to class j.

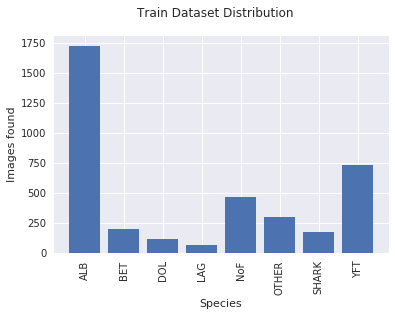
The submitted probabilities for a given image are not required to sum to one because they are rescaled prior to being scored (each row is divided by the row sum). To avoid the extremes of the log function, predicted probabilities are replaced with

Screen Clipping

**II. Analysis**

**Data Exploration**

* The dataset was compiled by [The Nature Conservancy](http://www.thisisourfuture.org/) in partnership with [Satlink](http://satlink.es/en/), [Archipelago Marine Research](http://www.archipelago.ca/), the [Pacific Community](http://www.spc.int/), the [Solomon Islands Ministry of Fisheries and Marine Resources](http://fisheries.gov.sb/), the [Australia Fisheries Management Authority](http://fisheries.gov.sb/), and the governments of [New Caledonia](https://gouv.nc/) and [Palau](http://palaugov.pw/executive-branch/ministries/natural-resources/).
* We used a dataset of 3777 images in training set classified into 8 labels, published by Nature Conservancy on Kaggle. The labels included six species of fish as well as one “No fish” and one “Other” label. The test set has 1000 images.
* From the outset, the data presented several key challenges. The images were all varied sizes and had all been taken at various times of day. Some contained more than one species of fish. The dataset was also small, containing only a few thousand images. Each image has only one fish category, except that there are sometimes very small fish in the pictures that are used as bait. The Nature Conservancy has also provided a visualization of labels



|  |  |  |
| --- | --- | --- |
| S.No | Species | Images Present |
| 1. | ALB | 1719 |
| 2. | BET | 200 |
| 3. | DOL | 117 |
| 4. | LAG | 67 |
| 5. | NoF | 465 |
| 6. | SHARK | 176 |
| 7. | YFT | 734 |
| 8. | OTHER | 299 |
| Total Images | | 3777 |

**Exploratory Visualization**

In this section, you will need to provide some form of visualization that summarizes or extracts a relevant characteristic or feature about the data. The visualization should adequately support the data being used. Discuss why this visualization was chosen and how it is relevant. Questions to ask yourself when writing this section:

* *Have you visualized a relevant characteristic or feature about the dataset or input data?*
* *Is the visualization thoroughly analyzed and discussed?*
* *If a plot is provided, are the axes, title, and datum clearly defined?*

**Algorithms and Techniques**

In this section, you will need to discuss the algorithms and techniques you intend to use for solving the problem. You should justify the use of each one based on the characteristics of the problem and the problem domain. Questions to ask yourself when writing this section:

* *Are the algorithms you will use, including any default variables/parameters in the project clearly defined?*
* *Are the techniques to be used thoroughly discussed and justified?*
* *Is it made clear how the input data or datasets will be handled by the algorithms and techniques chosen?*

**Benchmark**

In this section, you will need to provide a clearly defined benchmark result or threshold for comparing across performances obtained by your solution. The reasoning behind the benchmark (in the case where it is not an established result) should be discussed. Questions to ask yourself when writing this section:

* *Has some result or value been provided that acts as a benchmark for measuring performance?*
* *Is it clear how this result or value was obtained (whether by data or by hypothesis)?*

**III. Methodology**

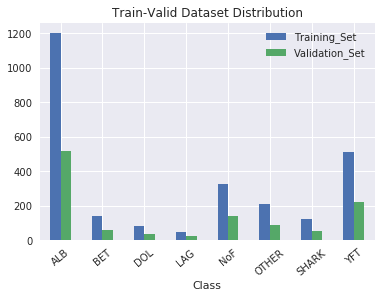
*(approx. 3-5 pages)*

**Data Pre-processing**

As we are using VGG16 Net architecture for transfer learning, we must pre-process the images complying to the VGG16 Net architecture. The major steps to perform are

1. Subtracting the mean RGB value, computed on the training set, from each pixel.
2. VGG Net accepts images in (B, G, R) order but by default python reads images in (R, G, B) order. So, we must convert the input images to (B, G, R) format.
3. Resize the image size to (150, 150, 3) for uniformity among all images.
4. Keras ImageDataGenerators were used to do image augmentation like lighting, shading, distortion etc.

Also, we have to create a validation dataset by dividing the train dataset randomly in 7: 3 ratio.



|  |  |
| --- | --- |
| Dataset | Images Present |
| Train Dataset | 2640 |
| Validation Dataset | 1137 |

**Implementation**

In this section, the process for which metrics, algorithms, and techniques that you implemented for the given data will need to be clearly documented. It should be abundantly clear how the implementation was carried out, and discussion should be made regarding any complications that occurred during this process. Questions to ask yourself when writing this section:

* *Is it made clear how the algorithms and techniques were implemented with the given datasets or input data?*
* *Were there any complications with the original metrics or techniques that required changing prior to acquiring a solution?*
* *Was there any part of the coding process (e.g., writing complicated functions) that should be documented?*

**Refinement**

In this section, you will need to discuss the process of improvement you made upon the algorithms and techniques you used in your implementation. For example, adjusting parameters for certain models to acquire improved solutions would fall under the refinement category. Your initial and final solutions should be reported, as well as any significant intermediate results as necessary. Questions to ask yourself when writing this section:

* *Has an initial solution been found and clearly reported?*
* *Is the process of improvement clearly documented, such as what techniques were used?*
* *Are intermediate and final solutions clearly reported as the process is improved?*

**IV. Results**

*(approx. 2-3 pages)*

**Model Evaluation and Validation**

In this section, the final model and any supporting qualities should be evaluated in detail. It should be clear how the final model was derived and why this model was chosen. In addition, some type of analysis should be used to validate the robustness of this model and its solution, such as manipulating the input data or environment to see how the model’s solution is affected (this is called sensitivity analysis). Questions to ask yourself when writing this section:

* *Is the final model reasonable and aligning with solution expectations? Are the final parameters of the model appropriate?*
* *Has the final model been tested with various inputs to evaluate whether the model generalizes well to unseen data?*
* *Is the model robust enough for the problem? Do small perturbations (changes) in training data or the input space greatly affect the results?*
* *Can results found from the model be trusted?*

**Justification**

In this section, your model’s final solution and its results should be compared to the benchmark you established earlier in the project using some type of statistical analysis. You should also justify whether these results and the solution are significant enough to have solved the problem posed in the project. Questions to ask yourself when writing this section:

* *Are the final results found stronger than the benchmark result reported earlier?*
* *Have you thoroughly analyzed and discussed the final solution?*
* *Is the final solution significant enough to have solved the problem?*

**V. Conclusion**

*(approx. 1-2 pages)*

**Free-Form Visualization**

In this section, you will need to provide some form of visualization that emphasizes an important quality about the project. It is much more free-form, but should reasonably support a significant result or characteristic about the problem that you want to discuss. Questions to ask yourself when writing this section:

* *Have you visualized a relevant or important quality about the problem, dataset, input data, or results?*
* *Is the visualization thoroughly analyzed and discussed?*
* *If a plot is provided, are the axes, title, and datum clearly defined?*

**Reflection**

In this section, you will summarize the entire end-to-end problem solution and discuss one or two particular aspects of the project you found interesting or difficult. You are expected to reflect on the project as a whole to show that you have a firm understanding of the entire process employed in your work. Questions to ask yourself when writing this section:

* *Have you thoroughly summarized the entire process you used for this project?*
* *Were there any interesting aspects of the project?*
* *Were there any difficult aspects of the project?*
* *Does the final model and solution fit your expectations for the problem, and should it be used in a general setting to solve these types of problems?*

**Improvement**

In this section, you will need to provide discussion as to how one aspect of the implementation you designed could be improved. As an example, consider ways your implementation can be made more general, and what would need to be modified. You do not need to make this improvement, but the potential solutions resulting from these changes are considered and compared/contrasted to your current solution. Questions to ask yourself when writing this section:

* *Are there further improvements that could be made on the algorithms or techniques you used in this project?*
* *Were there algorithms or techniques you researched that you did not know how to implement, but would consider using if you knew how?*
* *If you used your final solution as the new benchmark, do you think an even better solution exists?*