Machine Learning Engineer Nanodegree

Capstone Project

Marcel Schneider, March 24th, 2019

Inhaltsverzeichnis

[I. Definition 2](#_Toc4003775)

[Project Overview 2](#_Toc4003776)

[Problem Statement 2](#_Toc4003777)

[Metrics 3](#_Toc4003778)

[Submission File 3](#_Toc4003779)

[II. Analysis 4](#_Toc4003780)

[Data Exploration 4](#_Toc4003781)

[Exploratory Visualization 4](#_Toc4003782)

[Algorithms and Techniques 4](#_Toc4003783)

[Benchmark 4](#_Toc4003784)

[III. Methodology 5](#_Toc4003785)

[Data Preprocessing 5](#_Toc4003786)

[Implementation 5](#_Toc4003787)

[Refinement 5](#_Toc4003788)

[IV. Results 5](#_Toc4003789)

[Model Evaluation and Validation 6](#_Toc4003790)

[Justification 6](#_Toc4003791)

[V. Conclusion 6](#_Toc4003792)

[Free-Form Visualization 6](#_Toc4003793)

[Reflection 6](#_Toc4003794)

[Improvement 7](#_Toc4003795)

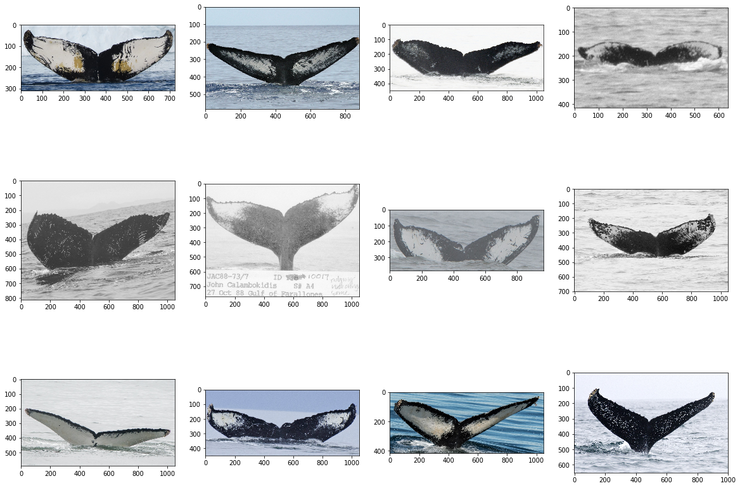
# I. Definition

\_(approx. 1-2 pages)\_

### Project Overview

*In this section, look to provide a high-level overview of the project in layman’s terms. Questions to ask yourself when writing this section:*

*- \_Has an overview of the project been provided, such as the problem domain, project origin, and related datasets or input data?\_*



sample images of the dataset[[1]](#footnote-1)

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 from Kaggle the task is to build an algorithm to identify individual whales in images. Happywhale’s database of over 25,000 images, gathered from research institutions and public contributors is analyzed.

### Problem Statement

*In this section, you will want to clearly define the problem that you are trying to solve, including the strategy (outline of tasks) you will use to achieve the desired solution. You should also thoroughly discuss what the intended solution will be for this problem. Questions to ask yourself when writing this section:*

*- \_Is the problem statement clearly defined? Will the reader understand what you are expecting to solve?\_*

The goal in this capstone project is to identify whales reliable by their tail.

The biggest problem in this challenge is that of some whales only a few images exist (min of 2 images per whale). This makes it quite hard to identify whales with only a few images taken of. On top of that, the network is also meant to label new whales.

*- \_Have you thoroughly discussed how you will attempt to solve the problem?*

I will solve this problem by using different neural network architectures. I will try networks designed from scratch and I will use transfer learning. Depending on the results I will improve one of those architecture types further throughout the project. I will also try to increase the sample size by the use of image augmentation[[2]](#footnote-2). This is a well known method to increase sample size if you have limited data. I plan to improve my predictions by the use of this method.

I will also transform the images to black and white and see if this improves the performance by reducing model complexity.

\_

*- \_Is an anticipated solution clearly defined? Will the reader understand what results you are looking for?\_*

### Metrics

*In this section, you will need to clearly define the metrics or calculations you will use to measure performance of a model or result in your project. These calculations and metrics should be justified based on the characteristics of the problem and problem domain. Questions to ask yourself when writing this section:*

*- \_Are the metrics you’ve chosen to measure the performance of your models clearly discussed and defined?\_*

I will evaluate the results of my model using the evaluation formula provided by Kaggle. I will also evaluate my results by using the Late submission feature of Kaggle to be able to compare my results with the leaderboard of this competition.

The evaluation formula is 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.

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. For example, if the correct label is A for an observation, the following predictions all score an average precision of 1.0.

[A, B, C, D, E]

[A, A, A, A, A]

[A, B, A, C, A]

### Submission File

For each Image in the test set, you may predict up to 5 labels for the whale Id. Whales that are not predicted to be one of the labels in the training data should be labeled as new\_whale. The file should contain a header and have the following format:

Image,Id

00028a005.jpg,new\_whale w\_23a388d w\_9b5109b w\_9c506f6 w\_0369a5c

000dcf7d8.jpg,new\_whale w\_23a388d w\_9b5109b w\_9c506f6 w\_0369a5c

...

# II. Analysis

\_(approx. 2-4 pages)\_

### Data Exploration

*In this section, you will be expected to analyze the data you are using for the problem. This data can either be in the form of a dataset (or datasets), input data (or input files), or even an environment. The type of data should be thoroughly described and, if possible, have basic statistics and information presented (such as discussion of input features or defining characteristics about the input or environment). Any abnormalities or interesting qualities about the data that may need to be addressed have been identified (such as features that need to be transformed or the possibility of outliers). Questions to ask yourself when writing this section:*

*- \_If a dataset is present for this problem, have you thoroughly discussed certain features about the dataset? Has a data sample been provided to the reader?\_*

The 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.

The following data was originally provided:

* 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

Train.zip:

This file contains 25.362 images of whale tails with a total file size of 4.5 GB. These images will be used for training and testing/improving the neural network. Therefore they are split during the experiment.

Train.csv:

This file contains two columns: The filename of an image and the whale id. This file is used together with the images to train and test the neural networks.

Test.zip:

This file contains 7.960 images of whale tails with a total file size of 1.46 GB. These images are used to evaluate the neural network. There is no csv file containing the correct mappings included in the data. The predicted mappings have to be uploaded to Kaggle as csv file. The predictions are evaluated by Kaggle.

One of the biggest challenges in this challenge is the variety of available examples per label and the few image available per label. There are 9664 new whales in the data set. The most fluke shower has 73 pictures and is labelled with ID w\_23a388d. Which as an amount of data which is sufficient to train a neural network. But there is a total of 2073 whales in the database appearing only once. On average there are 5.07 images per whale available. The median of images per whale is 2.

### 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?\_*

#### Outlier images

There are some outlier images. But as only a minority of these images exist, I will simply ignore them.

Outlier image 1 – dimension distorted



Outlier image 2 – photo not taken properly



Outlier image 3 – image not shown clearly

*- \_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?\_*

I will compare multiple neural network architectures and I will try to evaluate which of them is performing the best. Therefore I will design a simple CNN myself and I will use transfer learning as a state of the art method in image classification.

I’m also planning to apply techniques like image augmentation and I will try different image resolutions to see if the accuracy improves.

*- \_Are the techniques to be used thoroughly discussed and justified?\_*

### 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)?\_*

Include: <https://towardsdatascience.com/epoch-vs-iterations-vs-batch-size-4dfb9c7ce9c9>

# III. Methodology

\_(approx. 3-5 pages)\_

### Data Preprocessing

*In this section, all of your preprocessing steps will need to be clearly documented, if any were necessary. From the previous section, any of the abnormalities or characteristics that you identified about the dataset will be addressed and corrected here. Questions to ask yourself when writing this section:*

*- \_If the algorithms chosen require preprocessing steps like feature selection or feature transformations, have they been properly documented?\_*

The data was preprocessed in regard of transforming the resolution of the images to 100x100x3. The labels were on hot encoded.

When activating image augmentation, the images were transformed and multiplied to create more relevant training data for the neural network.

*- \_Based on the \*\*Data Exploration\*\* section, if there were abnormalities or characteristics that needed to be addressed, have they been properly corrected?\_*

*- \_If no preprocessing is needed, has it been made clear why?\_*

### 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:*

The implementation is explained in the attached html file extracted from Jupyter Notebooks.

### 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?\_*

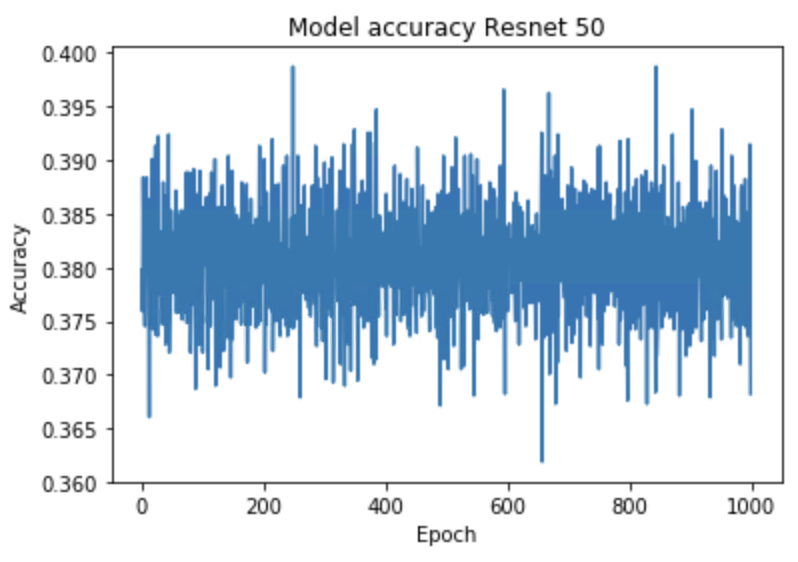
# V. Conclusion

\_(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?\_*



Kaggle score: 0.27679

Image resolution: 100 x 100 x 3

Image augmentation: Yes (featurewise\_center=True, featurewise\_std\_normalization=True, rotation\_range=20, width\_shift\_range=0.2, height\_shift\_range=0.2, horizontal\_flip=True)

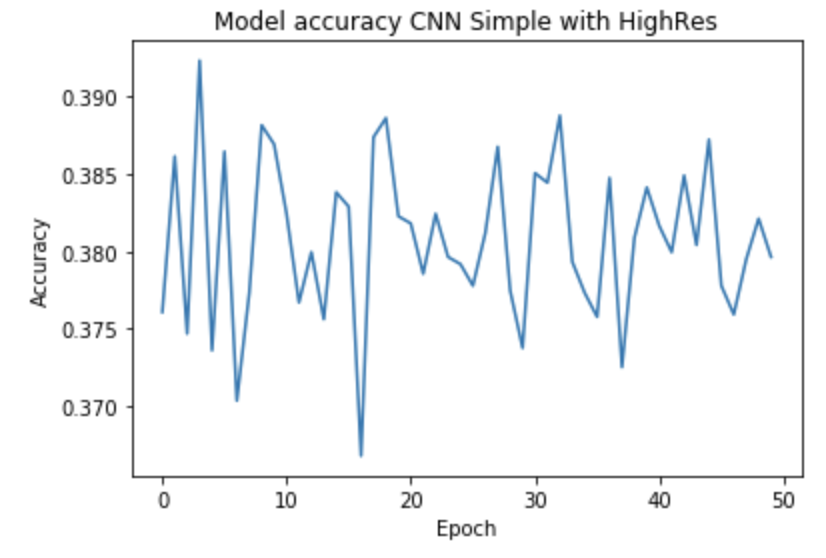
Train/Test split: yes

Filename: historyresnet50.h5

Dropout: 0.5

Metrics: accuracy

Epochs: 1000



Kaggle score: 0.27751

Image resolution: 200 x 200 x 3

Test accuracy: 0.878466280247

Densenet121:

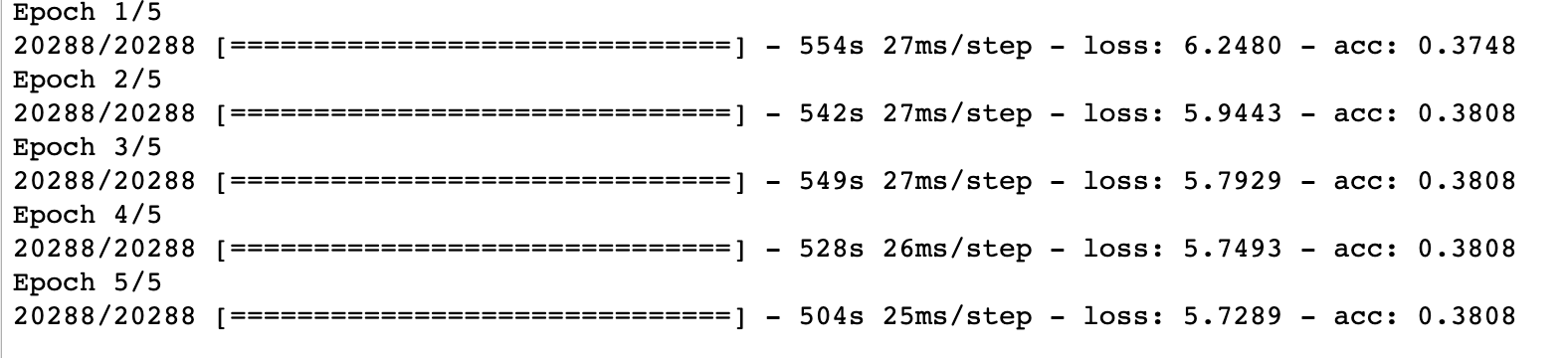
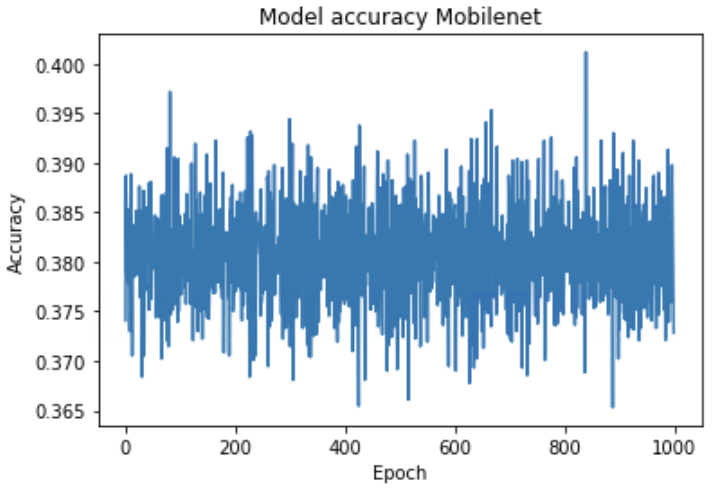


Image resolution: 100 x 100 x 3

Test accuracy: Kernel died



Kaggle score: 0.27740

Test accuracy: 0.878

Epochs: 1000

*- \_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?\_*

There are a lot of options available for tuning the hyperparameters and image resolution. But as my computing power is limited, I couldn’t increase the image resolution and I couldn’t increase the number of epochs a lot.

*- \_Were there algorithms or techniques you researched that you did not know how to implement, but would consider using if you knew how?\_*

If there was sufficient time and gpu processing power, I would like to apply SiameseNets as they are specialized in working with one shot training which is the case in this challenge. [[3]](#footnote-3)

*- \_If you used your final solution as the new benchmark, do you think an even better solution exists?\_*

Definitely, the winners of the Keras challenge scored 0.97309 with their implementation of Siamese net. But this was a tough challenge as there were very few images per whale available.

1. <https://weiminwang.blog/2019/03/01/whale-identification-5th-place-approach-using-siamese-networks-with-adversarial-training/> [↑](#footnote-ref-1)
2. <https://medium.com/nanonets/how-to-use-deep-learning-when-you-have-limited-data-part-2-data-augmentation-c26971dc8ced> [↑](#footnote-ref-2)
3. <https://towardsdatascience.com/one-shot-learning-with-siamese-networks-using-keras-17f34e75bb3d> [↑](#footnote-ref-3)