Whale and Dolphin Identification

Matthew Harding

April 2024

1 Introduction

This project explores the problem of classifiying images of dolphin and whales to identify indivduals. This work is based on the *Happywhale - Whale and Dolphin Identification* Kaggle competition.

Data for this competition contains images of over 15,000 unique individual marine mammals from 30 different species collected from 28 different research organizations. Individuals have been manually identified and given an individual_id by marine researches.

Unlike a typical classification problem in which there is a fixed set of classes which examples of all classes within the training data, this problem required the ability to classify individuals not contained within the training dataset.

2 Training Data

Within the training data there are 51,033 labelled images. Looking at the histogram of species Figure 1, there is a clear inbalance in the number of iamges per species with the vast majority of labelled images being for bottlenose dolphins, beluga whales, humpback whales and blue whales. For some species the number of examples it so low it will be difficult to gain a high level of accuracy within classification.

Look at Figure 2, the mean count of images per indivdual by species we see another imbalance with some species on average containing a large number of images per indivudal whilst others only contain one or two images per individual.

3 Naive Classifer

The first approach taken was to classify images

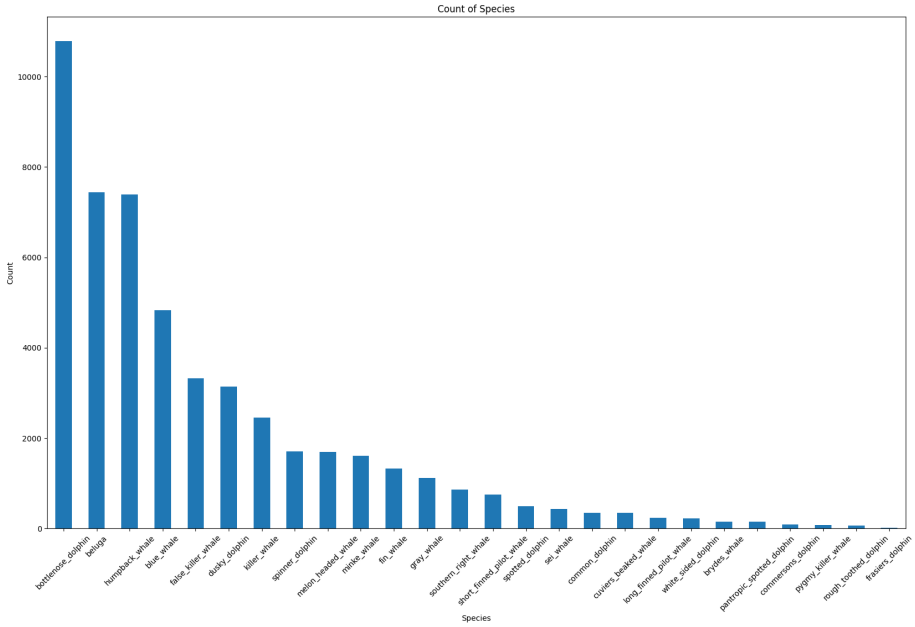

Figure 1: Number of training images per species

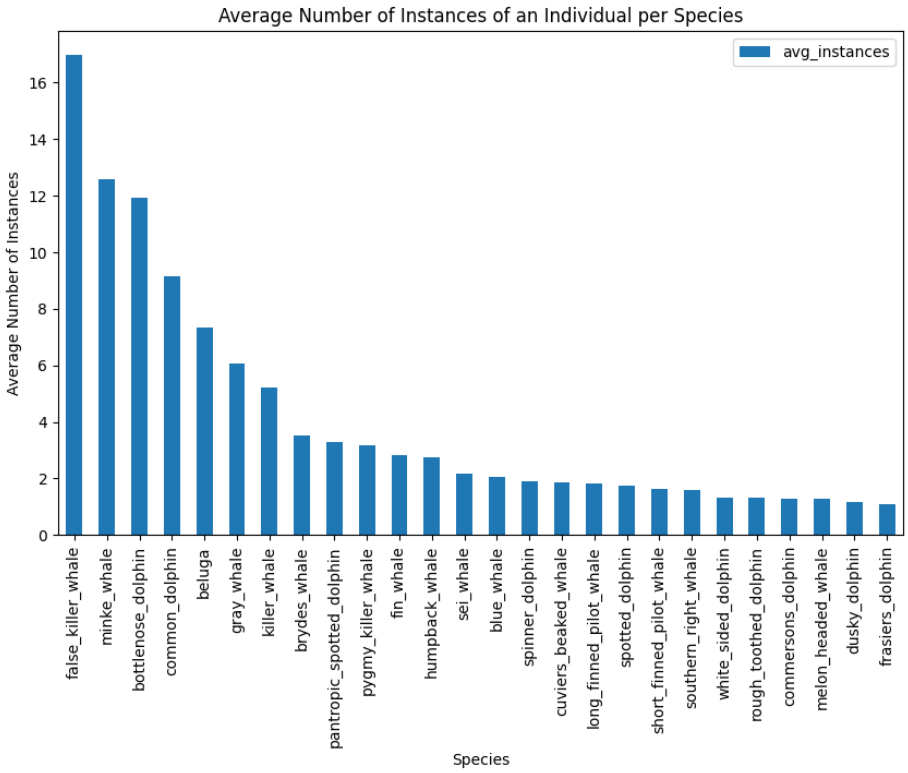

Figure 2: Mean count of images per individual by species