Right Whale Recognition

Sumit Gupta, Indiana University Bloomington
guptasum@indiana.edu

Abstract

Less than 500 Right Whales are left in the world's oceans and declining at an alarming rate. To observe the health, status of individual whales, researchers spend a great amount of time on identifying individual whales. We present and experiment with several approaches to classify individual whales on NOAA Right Whales dataset. We train and fine-tune several CNNs for whale classification based on AlexNet derived architectures. We also train a Linear SVM classifier which uses features extracted from the fully connected layer of CNNs. We compare the results between different approaches and empirically show that computer vision techniques are useful for whale classification and can greatly aid in conservation efforts.

1. Introduction

Many species of whales are on the verge of extinction including Right whales, Blue whales, Beluga whales and Humpback whales. These beautiful creatures roamed in the oceans for millions of years but because of the pollution, hunting, getting hit by the ships and getting tangled in finishing nets, their numbers are rapidly decreasing. According to NOAA, less than 500 right whales are left in the ocean. For aiding the conservation efforts and to save the right whale species from extinction, it is necessary to maintain and observe health status of each of the remaining right whales (termed as just 'whales' from now on in the report). While working in the ocean, only a very few experienced researchers can identify individual whales when sighted which makes it difficult to keep records. To help the conservation effort, we apply computer vision approaches to identify individual whales from a set of 447 whales. We process a set of about 4500 aerial images of whales taken from different angles and viewpoint. Motivated by the recent success of CNNs in image classification, we follow several approaches for image classification. The dataset contains various aerial photographs couple of which are shown in figure 1. We create training and testing set from the given data and input them to a pre-trained CNN. We use both Caffe Deep Learning framework and SVM as our classifiers.

Figure 1: NOAA Whale Dataset samples images.

2. Related Works

Though there are no previous works specifically in whale recognition from aerial photographs using computer vision techniques, considering the level of danger to whales, they have received considerable attention from researchers in various fields. Surprisingly, there are not much work in computer vision research community for detection and recognition of whales. There are few papers trying to aid the conservation effort based on photo-identification. For example, the 1990 report published by the International Whaling Commission provides many papers detailing the photo-identification methods and techniques used to estimate whale population parameters as well as other aspects. However, the report is more related to photo identification rather than using computer vision techniques. There have been previous works which used whale acoustic signal processing and machine learning algorithms to detect whales. Smirnov 2013 tries to detect the call using CNNs. To our knowledge, there have not been attempts to classify individual whales using CNNs. DIGITS software, which is used by NOAA to identify is based on image recognition algorithms which are old and unreliable as they need manual inspection and a huge infrastructure consisting of multiple servers and SQL database to operate. Apart from the whale and natural

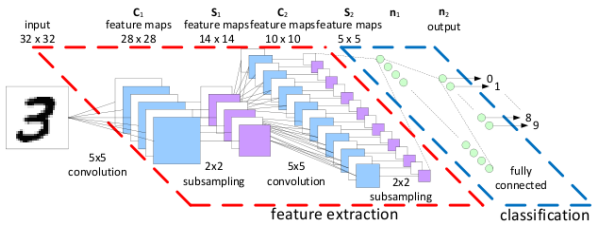

Figure 2: CNN can be used as feature extractors as well as classifiers. An example CNN is shown in above figure. Image source: http://parse.ele.tue.nl/cluster/2/CNNArchitecture.jpg

wildlife domain, there have been significant use of CNNs as classifiers and feature extractors. Our approach closely matches with that of Girshick et. al. for R-CNNs. The main contribution of this report are to empirically show that computer vision techniques like CNNs are useful for whale classification and CNNs are good feature extractors. One non-technical contribution of this report is to motivate computer vision researchers to focus on real world wildlife problems and help in conservation efforts.

Since our approach is heavily based on CNNs and SVMs, we explain them briefly in section 3. Approach is described in section 4 followed by Experiments in section 5. We conclude the report by providing Conclusion in section 6.

3. Convolutional Neural Networks and SVM

Before we dive into technical details of our approach, we provide a very brief introduction to CNNs and SVMs.

3.1 CNNs: A convolutional neural network is biological inspired network which tries to imitate the workings of human mind. It is a type of feed-forward artificial neural network in which neurons perform image convolution operation together at a large scale. An example of a CNN is given in the Figure 2 in which the network is used to identify the handwritten digits. Another example is AlexNet architecture which has proven to be performing well for ImageNet image classification. Refer [x], [y] for details about the working of CNNs.

For the classification tasks, we need good features which describe the image category well for a general image. When a CNN is trained using the training data, the weights within the network are updated during a backward pass using back propagation algorithm. Once these weights achieve a satisfactory level of accuracy, they can be re-used to test on other data or to retrain an almost similar network. These weights are the features learned from the training data. We will use these weights in our task to recognize whales.

3.2 Support Vector Machines: SVM are machine learning models which are used for classification. They are based on support vectors and kernels. Support vector machines require at least training feature vectors along with their ground truth labels. It then classifies any new feature vector of the same dimension as the training feature vectors and produces an optimum classification. For this report, we use only multiclass SVM and experiment with different type of kernels.

5. Approach

We use a CNN based network to classify the whales into 447 categories. We tried three approaches: (1) Classification of images based on CNN fine-tuning of pre-trained models. (2) Features extraction from fully connected layers of pre-trained CNN and classification using SVM. (3) Features extraction from fully connected layers of fine-tuned model and classification using linear SVM. The approaches we follow are simple yet powerful.

5.1 Classification based on only CNN fine-tuning: CNNs have proved to be very effective in image classification. We take well known AlexNet based derivatives and fine-tune them to our task. Fine tuning means to retrain an already trained network so that it performs better for your dataset. It is borrowing the network weights from an already trained network and change them to suit your data by proving training samples. Since we already have a pre-trained model on 1000000 images which works well for image classification, we take Caffenet which is similar to AlexNet except that the order of pooling and normalization layer is switched. Our model consists of input data layer, followed by 5 convolution layers. Each of the convolution layer is followed by a max pooling layer and normalization layer. We use rectified linear units (ReLUs) as activation units. Convolution layers are followed by three fully connected layers (fc6, fc7 and fc8) each of

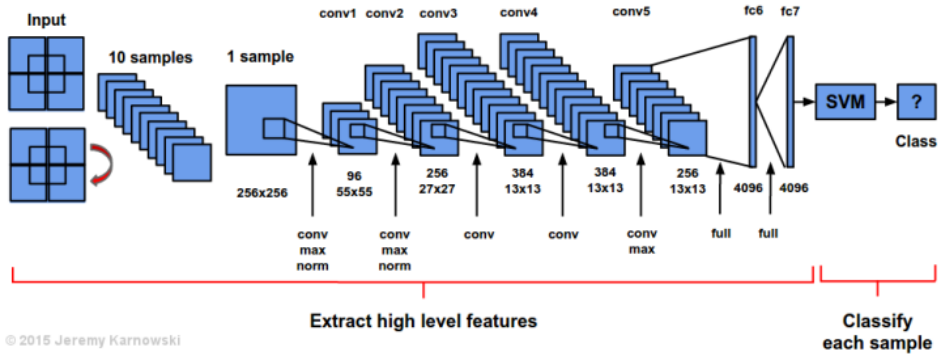

Figure 3: In our approach, we take Alexnet based derivative architetures as our classifers as well as feature extractors. Above figure shows our second and third approach where we extract features from fc7 layer and train an SVM. Image source: https://jeremykarnowski.files.wordpress.com/2015/07/alexnet2.png?w=720

which is producing 4096, 4096 and 447 output sizes respectively and a softmax layer. We also have

dropout layers on fc6 and fc7 with dropout ratio of 0.5. We perform 100,000 iterations snapshotting every 1000 iterations. Since the learning has already been done for all layers except fc8_whale, we set learning rate of fc8_whale high compared to all other layers.

5.2 Classification using SVM of features from pre-trained model: As described above, CNNs generate features based on the training data. We extract a 4096 dimensional feature vector from the fc7 layer of CaffeNet using the pycaffe interface of Caffe. Since the CaffeNet has 1000 classes compared to 447 classes in whale dataset, the only layer which is contributing to the 1000 classes output is fc8, we disregard this output layer and take features from fc7. For the fc7 layer, it always outputs a 4096 dimensional vector irrespective of the number of classes. The data in Caffe framework flows as blobs. We extract the blobs from fc7 layer and convert it to human readable format. The features are extracted by inputting the mean subtracted whale images into the network, 227x227 image size, changing the channel mode to BGR and starting the forward propagation. Since we use pycaffe interface for the forward propagation, we do the initial pre-processing of the image in python as required by the Caffe. We input the images in a batch size of one to merge the extracted feature matrix into already processed images. We create a 24300 x 4096 dimensional ndmatrix using python numpy library. Once the features are extracted, we do a L2 normalization of the feature matrix X_train and labels matrix y_train and then use Linear SVM from scikit-learn python library to train a SVM. Similar process is repeated for the testing set thus generating a test features vectors X_test and test labels y_test which are normalized and tested using the SVM model. We found that linear SVM provides the best performance and takes less time in training. We also create a confusion matrix but since the dimension of that matrix is 447x447, it is not possible to include that in this report.

5.3 Classification using SVM of features from fine-tunes model: In this approach, we combine above mentioned approaches (i.e. fine-tuning + SVM) to extract feature vectors from a fine-tuned model. Once both of the above approaches are done, it is simple to implement this approach. Instead of the pre-trained model in approach 2, we just replace it with one of the fine-tuned model's snapshot. An important point to note here is that a snapshot model which is performing well in CNN might not perform that well when we use that snapshotted model as our feature extractor. For example, we observed that a CNN was performing the best at around 7000 iterations. However, when we extracted the features from this snapshotted model, we didn't get better performance in SVM than other snapshotted model. So we experimented with top 3 performing snapshots.

6. Experiments

We perform the task of whale classification, we use the dataset provided by NOAA which contains about 4500 training images and 7000 testing images (without testing labels) for a total of 447 classes. These pictures are taken over a span of about 10 years

by researches and NOAA employees. The 447 classes belong to the 447 individual whales which are numbers from 1 to 447. An example label key, value pair is <whale_ 52342, 240>. Since we don't have

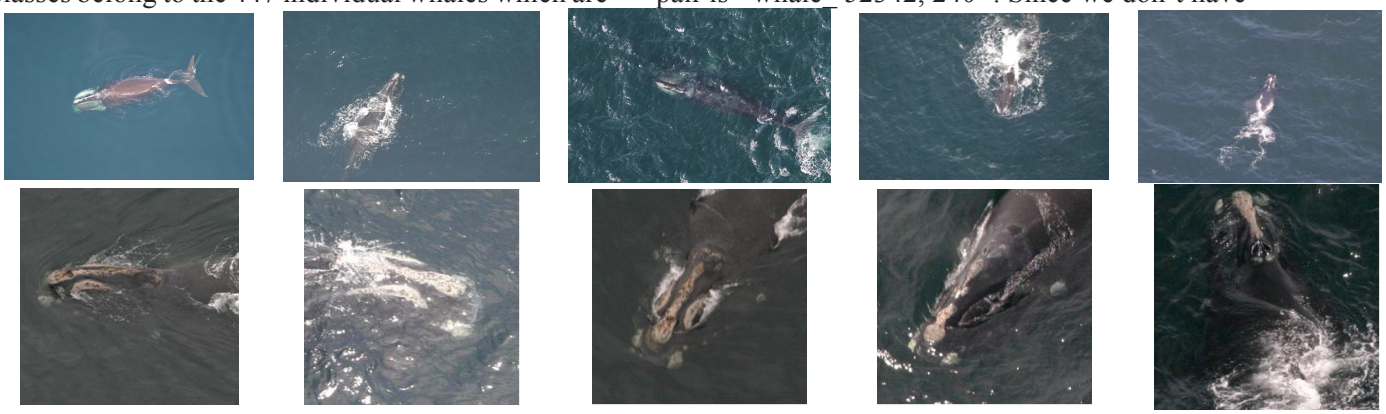

Figure 4: Samples images in the dataset. The variation is large which makes the task quite challenging. First row: first three photos are of the same whale while last two are of another whale. Second row: whale faces all belonging to the same whale.

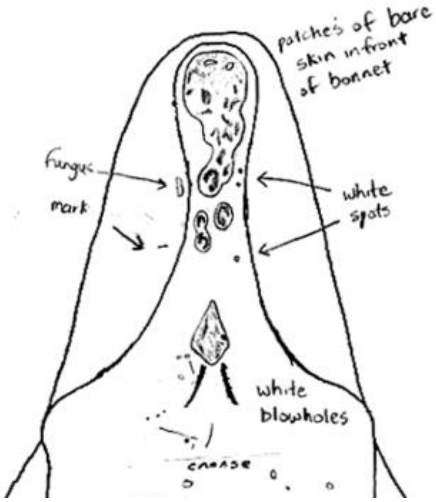

Figure 5: Whale faces have important characteristic to classify individual whales.
Image source: NOAA

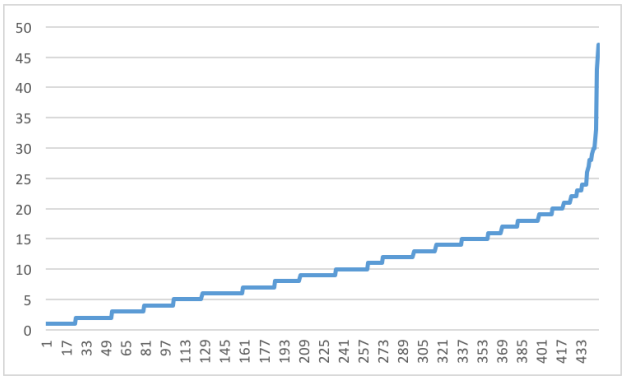

Figure 6: Number of images per class. Vertical axis is the number of images per class. Highest number of images is 47 images for a class. Lowest is only 1 image.

left in the world oceans. The dataset is quite challenging considering the variations in the appearance of same whale i.e. whales are having very different appearance in images based on the clicking angle and the activity they are doing. The variation in the appearance is shown in figure 4. On an average it is about 10 images per class but the data is quite unevenly distributed i.e. more than 50% of the classes have 9 images or less and about 30% have 5 images of less thus making the recognition task challenging. The dataset distribution with respect to the number of images per class is shown in figure 6.

Data Pre-processing: We take the data from the train labels file and create a <key, value> mapping to labels for 7000 testing images, we randomly choose about 10% of the training data as our testing set by creating 3 different datasets. In other words, we create 3 sets of training and testing data from the original training data by randomly choosing the testing set. One such set contains exactly 4090 training images and about 400 testing images. Later on, we will average the result of all 3 datasets to come up with an accuracy figure. Once the datasets are generated, we convert them to LMDB file format database which is required by Caffe to process images efficiently.

Importance of Whale faces: Through the background study of the problem, we found out that whale faces are an important part of the whale body from which experienced researchers identify

individual whales. As explained in figure 5, as whale grow older, callouts white color patches begin to form which remain for a long time and are distinctive

to individual whales. So we cropped the whale images using Sloth image labeler tool into a 256x256 pixel size. Few examples of whale faces are shown in figure 7.

Data Augmentation: It is a CNN property that they extract different features for different orientation of the image. Since the data size is quite small for training a CNN, we used data augmentation technique to prevent over-fitting. We used a horizontal flip, vertical flip, 90, 180 and 270 degree rotations to increase out data size by six fold. For the augmented dataset, we have a total of about 27000 training images. From this, we again randomly select 10% of the data as testing and create 2 such datasets as described above. Again, we average the results of both datasets to come up with an accuracy figure.

We used different systems for training and testing CNN and SVMs. Details of which are given below:

CNN: We train/fine-tune and test the CNN on a Dell PowerEdge T630 server with NVidia Tesla K40 GPU boards and two Intel E5-2680 v3 2.5GHz, 30M Cache, 9.60GT/s QPI, Turbo, HT, 12C/24T, Max Mem 2133MHz and 128 GB memory.
SVM: For training and testing the SVM, we use a Macbook Pro with Intel i5 2.7 GHz processor, Intel Iris 6000 GPU and 8GB of memory.

We achieved highest accuracy of about 21.06% on whale faces augmented dataset using SVM as a classifier on features extracted from further fine tuning a flickr-finetuned Caffe model.

Dataset	Approach	Accuracy	Running Time
Whale only	Caffenet + SVM	12.44%	3 hours
Whale only	Caffenet Finetuned	13.5%	9 hours
Whale only	Caffenet Finetuned + SVM	18.06%	3 hours
Whale only	Flickr Finetuned + SVM	13.69%	
Whale faces + aug	Caffenet + SVM	12.63%	2.5 hours
Whale faces + aug	Caffenet Finetuned	14.86%	7 hours
Whale faces + aug	Caffenet Funetuned + SVM	18.52%	
Whale faces + aug	Flickr Finetuned + SVM	12.88%	~1.5 hours
Whale faces + aug	Flickr Finetuned	19.94%	~8 hours

The best results occurred at 7000 iteration snapshot of the fine-tuning. It provided about 20.24% accuracy on the faces dataset and about 14% accuracy on the whale only dataset.

We believe that the accuracy is far better than the unexperienced human accuracy. Since whales look almost same to unexperienced humans, they are difficult to classify. Thus our model can definitely aid in the whale conservation effort to some extent.

We observe that after 7000 iterations, the accuracy dropped to 0.011% and becomes constant throughout the remaining training.

8. Conclusion / Improvements

We show that pre-trained CNNs perform well in individual whale classification tasks which not only saves time for training a large network but also is helpful to get better results for a very small dataset like NOAA Whale dataset. Only CNNs doesn't work quite well. Since the whale features are encoded in faces, some domain knowledge might be useful.
Several tasks can be automated. The data preparation and cleaning took a lot of time which could have been automated.
One snapshot model which might perform well as CNN classifier might not perform well in SVM.
We use data augmentation for whale faces only. Perhaps, data augmentation will help in increasing performance for the whale only dataset since only whale faces cannot completely identify whales. Their body shape, tail length and shape, their weight might also be important factor.

9. References

[1] https://en.wikipedia.org/wiki/Right_whale
[2] Caffe caffe.berkeleyvision.org
[3]https://www.kaggle.com/c/noaa-right-whale-recognition
[4] Rich feature hierarchies for accurate object detection and semantic segmentation R Girshick et al.
[5] Y. Lecun el at. http://yann.lecun.com/exdb/lenet/
[6] A Krizhevsky et al. ImageNet Classification with Deep Convolutional Neural Networks