A face classifer for North Atlantic Right whales

Matt Smith[1,*], Abhir Bhalerao[2], Ben Graham[3]

[1]Department of Statistics, University of Warwick, Coventry, United Kingdom
[2]Department of Computer Science, University of Warwick, Coventry, United Kingdom
[3]Facebook Artificial Intelligence Research, Paris, France
[*]m.d.smith@warwick.ac.uk

Abstract: Accurate monitoring of individuals in a threatened species is of upmost importance to conservationists and researchers. Human observation is expensive and autonomous ariel photography is becoming an increasingly useful technique regarding animal biometrics [1, 2]. Fewer than 500 North Atlantic right whales are left in the world's oceans. As with many animal biometric inspection processes, tracking and monitoring individuals is an extremely time consuming process. Advances in the implementation and performance of deep learning algorithms have drastically improved performance in object detection and recognition tasks [3]. We employ a wide range of interesting techniques to build a ”face-identification” algorithm for ariel photos of 447 unique. We follow a conventional modern face recognition pipeline consisting of the stages: detect, align, represent and classify [4]. We use deep learning algorithms to both detect and classify. A fully convolutional network [5] is employed to semantically segment a given image to detect the location of the whale's head and body, we then use PCA on the resulting image to normalize for the whale's direction. A significant amount of hand labelled masks are needed to generate enough supervised training data to make this work effectively [6]. We tackle this issue by employing semi-supervised learning techniques and histogram matching between images to improve our localization algorithm and find a significant improvement in our results.

1. Introduction

We entered the 2015 Right Whale Recognition online competition issued by Kaggle. Data consists of aeriel images, the vast majority containing a single Right whale. There are M = 447 unique whales, each of which has at least one photo in the training set which contains 4543 labelled images. The test set contains N = 6925 unlabelled images. Evaluation is based on the multi-class logarithmic loss

$$\text{logloss} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{i=1}^{M} y_{ij}\log(p_{ij}), \tag{1}$$

where log is the natural logarithm, y_{ij} is 1 if observation i belongs to whale j and 0 otherwise, and p_{ij} is the predicted probability that observation i belongs to whale j.

The data was collected and labelled over a 10 year period by NOAA (National Oceanic and Atmospehric Administration) scientists via numerous helicopter trips over the northern Atlantic.

We follow the conventional pipeline of alignment and classification and break our task of classifying a given photo, denoted X_i into two main stages, both of which employ using convolutional neural networks;

- Alignment - We first reduce the dimensionality of X_i and normalize for the distance to and orientation of the whale, generating a headshot of the whale, denoted X_i^h.
- Classification - X_i^h is passed through a classifier which outputs a probability mass function over the 447 whales.

1.1. Related work

Subsection text here.

2. Alignment

It is helpful to remove variation in inputs before giving them to a deep learning algorithm and, especially with faces, the success of a learned network is highly dependant on an alignment step [7, 4]. We randomly choose 550 images from our training set M^{Train} and another random 150 images to generate a test set M^{Test}. Using a graphics editor, for each $X_i \in M^{Train}, M^{Test}$ we create a semantic mask denoted M_i. An example of a pair is shown in figure 1.

(a) (b)

Fig. 1. *Example of* $\{X_i, M_i\}$ *pair; a)* X_i*, b)* M_i*.*

We distinguish between head, body and sea using red, yellow and black respectivley. Having two colors for distinguishing parts of the whale enables us to infer the direcition in which the whale is pointing. We rescale each X_i to dimension $w \times h \times c = 600 \times 900 \times 3$ and each M_i to $w^{'} \times h^{'} \times c^{'} = 19 \times 29 \times 3$. We use a fully connected convolutional neural network (FCNN) to learn a function $f : \mathbb{R}^{w \times h \times c} \rightarrow \mathbb{R}^{w^{'} \times h^{'} \times c^{'}}$. We are not interested in a huge amount of detail being produced in our predicted M_i, enough to infer head and body location, hence our choice for a smaller output space.

We describe our neural network architecture as follows;

$$F_1 = \{\text{down}_0-, ..., -\text{down}_4 - 3\text{C}_{2D}3/1 - \text{sigmoid}\}, \quad (2)$$

where

$$\begin{aligned}
\mathrm{conv_n} &= \{(48+32\mathrm{n})\mathrm{C}_{2D}3/1 - \mathrm{BN} - \mathrm{ReLU}\} \\
\mathrm{down_n} &= \{\mathrm{conv_n} - \mathrm{MP}3/2\}.
\end{aligned}$$

The following notation for denoting a network architecture is similar to [8] where $\{a - b - c\}$ denotes a network which is passed through the three layers a to b and finally c. $fC_{iD}k/s$ denotes an i dimensional convolutional layer with kernel size k in each dimension, a stride of s and number of filters f. Similarly MP$k/s$ is a two dimensional max-pooling layer with kernel size k and stride s. Other layer notations; BN $=$ batch normalization, ReLU and Sigmoid are layers of rectified linear activation units and sigmoid activation units respectively.

2.0.1. Mask prediction:

2.0.2. Results:

3. Classification

3.0.3. Results:

4. Conclusion

Sample equations.

5. Enunciations

6. Figures & Tables

The output for figure is:

Fig. 2. *Insert figure caption here*
a Insert Sub caption here
b Insert Sub caption here

The output for table is:

Table 1 An Example of a Table

One	Two
Three	Four

7. Conclusion

The conclusion text goes here.

8. Acknowledgment

Thanks to Christin Khan and Leah Crowe from NOAA for hand labeling the images. Kaggle for competition.

9. References

[1] L. Koh and S. Wich, "Dawn of drone ecology: low-cost autonomous aerial vehicles for conservation," 2012.

[2] J. Martin, H. H. Edwards, M. A. Burgess, H. F. Percival, D. E. Fagan, B. E. Gardner, J. G. Ortega-Ortiz, P. G. Ifju, B. S. Evers, and T. J. Rambo, "Estimating distribution of hidden objects with drones: From tennis balls to manatees," *PLoS One*, vol. 7, no. 6, p. e38882, 2012.

[3] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[4] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1701–1708.

[5] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.

[6] S. Hong, H. Noh, and B. Han, "Decoupled deep neural network for semi-supervised semantic segmentation," in *Advances in Neural Information Processing Systems*, 2015, pp. 1495–1503.

[7] Y. A. LeCun, L. Bottou, G. B. Orr, and K.-R. Müller, "Efficient backprop," in *Neural networks: Tricks of the trade*. Springer, 2012, pp. 9–48.

[8] B. Graham, "Fractional max-pooling," *arXiv preprint arXiv:1412.6071*, 2014.

10. Appendices

Appendices are allowed but please be aware that these are included in the overall word count.