Fine-Grained Image Classification: Right Whales

Alex Koo, Chloe Tang, Kenneth Rojas, Robert Scott

TEXAS The University of Texas at Austin

The University of Texas at Austin Cockrell School of Engineering

Abstract

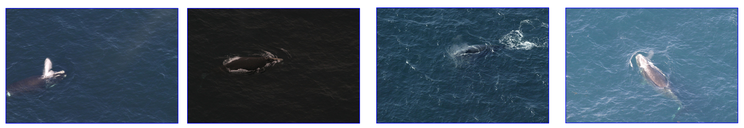
We propose explainability and feature importance to further advance Right Whale fine-grained image classification using Explainable Artificial Intelligence (XAI). Our framework builds upon a ResNet model, with XAI approaches such as Local Interpretable Model-Agnostic Explanations (LIME), Gradient-weighted Class Activation Mapping (GRAD-CAM), and Saliency Maps.

Motivation

The North Atlantic Right Whale is an endangered species due to fishing, climate change, and ship strikes. Identifying individual whales is challenging and time-consuming. This hinders effective monitoring and tracking of whales, so it is essential to further improve the process due to the declining population.

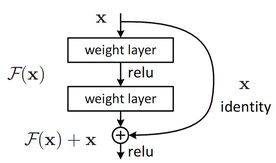
Background

We aimed to recognize individual right whales in photographs taken during aerial surveys, with a dataset of 447 different whales and 4,544 total images. The dataset was imbalanced, making it challenging for neural networks to focus on the unique characteristics of individual whales. The task required building a model that provides a probability distribution over all 447 whales for each photograph. The solutions were judged by multiclass log loss.

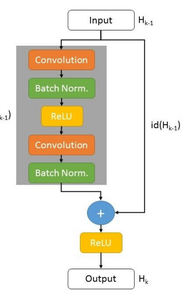


w_10.jpg w_431.jpg w_1747.jpg w_4849.jpg

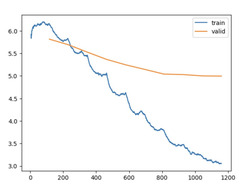
ResNet Model



Residual Network Logic



Residual Block



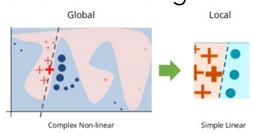
Original Loss Score: 4.93

ResNet26d consists of 26 layers, including convolutional, pooling, and fully connected layers, and utilizes residual connections to enable efficient training of deep networks.

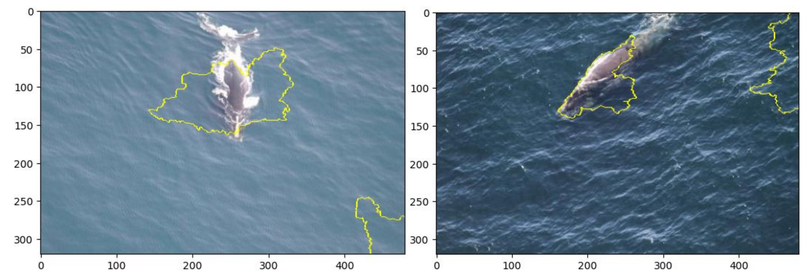
Implementing XAI

Local Interpretable Model-Agnostic Explanations (LIME)

LIME creates locally faithful surrogate models to explain the behavior of a complex model. It uses perturbation, which involves systematically altering the input data in small increments and observing the changes in the output. By generating these perturbed instances and training a simpler model on the resulting dataset, LIME can approximate the behavior of the original model in a local region.



Visualization of LIME



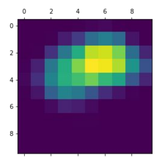
a,b) Resulting images of LIME with yellow-bounded areas representing areas that contribute towards correct whale labelling

Saliency Maps

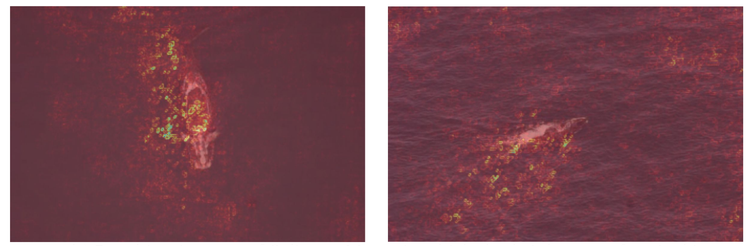
Saliency Maps is a general technique that uses the gradients of the final convolutional layer to produce a heatmap, which highlights the regions of the input image that are most important for the CNN's prediction. This heatmap is then overlaid on the original image to provide insights that are class-specific.

$$\alpha_k^c = \overbrace{\frac{1}{Z}\sum_i\sum_j}^{\text{global average pooling}} \underbrace{\frac{\partial y^c}{\partial A_{ij}^k}}_{\text{gradients via backprop}}$$

Gradient Formula



Heat Map Example

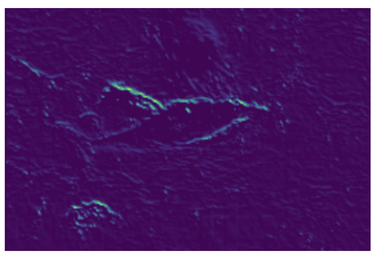


a,b) Resulting images of Saliency Maps, in which the brightness of a pixel is directly proportional to its saliency

Gradient-weighted Class Activation Mapping (GRAD-CAM)

GRAD-CAM is similar to Saliency Maps, but produces a heatmap is then overlaid on the original image to provide insights that are class-specific. It highlights the specific features and regions of an image that are most relevant for a particular class.

a) Original image



b) heat map depiction of important features (no overlay)

Future Applications

Our XAI results demonstrate that the model can include irrelevant features, such as wave patterns in the water, in its predictions. By cropping the images and using CNN-based head localizer and aligner techniques, we can isolate and identify the whale, improving accuracy and understanding of the model's reasoning.

By using all three XAI techniques, we can gain a deeper understanding of the model's reasoning and identify the most important features for whale classification. This allowed us to identify potential biases and shortcomings. By analyzing the LIME explanations, we can identify which regions of the image the model is focusing on and ensure that the model is not making predictions based on irrelevant features. By using Saliency Maps and GRAD-CAM, we can also gain insight into the model's decision-making process and identify which features are contributing the most to the final classification.