WHAT IDENTIFIES A WHALE BY ITS FLUKE? ON THE BENEFIT OF INTERPRETABLE MACHINE LEARNING FOR WHALE IDENTIFICATION

J. Kierdorf[1,*], J. Garcke[2,3], J. Behley[1], T. Cheeseman[4], R. Roscher[1,5]

[1] Institute of Geodesy and Geoinformation, University of Bonn, Germany - (jkierdorf, ribana.roscher, behley)@uni-bonn.de
[2] Institute for Numerical Simulation, University of Bonn, Germany - garcke@ins.uni-bonn.de
[3] Fraunhofer Center for Machine Learning and Fraunhofer SCAI, Sankt Augustin, Germany
[4] Happywhale and Southern Cross University - ted@happywhale.com
[5] Institute of Computer Science, University of Osnabrueck, Germany

ICWG II/III

KEY WORDS: Machine Learning, Deep Learning, Neural Networks, Interpretability, Visualization, Humpback Whales

ABSTRACT:

Interpretable and explainable machine learning have proven to be promising approaches to verify the quality of a data-driven model in general as well as to obtain more information about the quality of certain observations in practise. In this paper, we use these approaches for an application in the marine sciences to support the monitoring of whales. Whale population monitoring is an important element of whale conservation, where the identification of whales plays an important role in this process, for example to trace the migration of whales over time and space. Classical approaches use photographs and a manual mapping with special focus on the shape of the whale flukes and their unique pigmentation. However, this is not feasible for comprehensive monitoring. Machine learning methods, especially deep neural networks, have shown that they can efficiently solve the automatic observation of a large number of whales. Despite their success for many different tasks such as identification, further potentials such as interpretability and their benefits have not yet been exploited. Our main contribution is an analysis of interpretation tools, especially occlusion sensitivity maps, and the question of how the gained insights can help a whale researcher. For our analysis, we use images of humpback whale flukes provided by the Kaggle Challenge "Humpback Whale Identification". By means of spectral cluster analysis of heatmaps, which indicate which parts of the image are important for a decision, we can show that the they can be grouped in a meaningful way. Moreover, it appears that characteristics automatically determined by a neural network correspond to those that are considered important by a whale expert.

1. INTRODUCTION

Interpretable and explainable machine learning has gained momentum in recent years, especially with regard to the development of various methods for a better understanding of complex processes in neural networks (Samek, Müller, 2019). However, so far the potential of these methods for geo- and bioscientific applications has hardly been considered. A strength of interpretation tools, in combination with domain knowledge, is the possibility to verify the quality of a learned machine model in general and to get more information about the quality of certain observations. This is the basis for creating reliable models for practical use and for supporting the user in the application of these models and providing additional information that cannot be obtained by the machine learning model alone.

In the marine sciences, for example, a current problem is that the whale population worldwide is threatened by years of commercial whaling (Surma, Pitcher, 2015). In addition, adaptation to ocean warming and the struggle for food in competition with the fishing industry further affects the stability of the whale population. For this reason, many scientists control whale populations and monitor spatio-temporal migration to help protect the whales. An indispensable basis for the monitoring of whales is (re-)identification. They are identified by the shape of the whale flukes and their unique pigmentation (Katona, Whitehead, 1981). Especially three characteristics play a decisive role

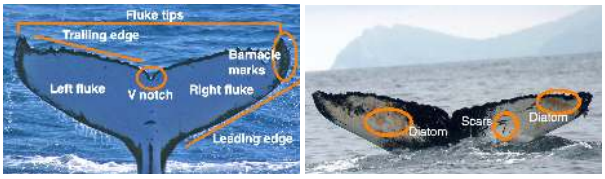

Figure 1. Important characteristics of a whale fluke.

for whale experts in the differentiation of the individual whales (see Fig. 1):

Pigmentation-based surface features These characteristics are the most obvious to the human eye, apart from large disfiguring shape features. They can change significantly in the first months, up to a couple years, of the whale's life, and in extremely cold water (Antarctica especially, Greenland and the far North Atlantic to a lesser extent). They can be partially obscured by substantial diatom growth, characterized by a yellow-orange appearance of the flukes.

Edge detection This characteristic is often the most reliable and robust. The outer 20% of the tail can distort more and change more over time, but the inner 80% and V-notch can be very reliable during the lifetime of the whale. This is hard to detect with the human eye, but has proven useful for machine learning based detection.

Acquired scars The surface of the fluke will usually scar white

* Corresponding author

on black and black on white. However, this is not always the case and scars do not always persist in the way we might expect them to. Specific scars will grow with the whale, for example, killer whale rake marks that make parallel lines or barnacle marks that make circles. Lighting can dramatically change the detectability of scarring.

To support the monitoring, whale researchers often use geotagged photographs with recorded time and location to track the activities of whales in the sea. So far, this work has mainly been carried out by individual scientists who have manually processed the given data. But a huge amount of data remained unused, which led to attention being focused on machine learning (Schneider et al., 2019, Kniest et al., 2010). In recent years, the use of deep neural networks became prevalent. In 2018, the Kaggle Challenge "Humpback Whale Identification"[1] was launched with the goal to develop efficient large-scale algorithms for the (re-)identification of individual whales based on images showing their fluke. All three winning solutions are based on neural networks [2] [3] [4].

Despite their success in achieving a high accuracy for the identification of whales, the interpretability of neural networks poses a considerable challenge. Generally, interpretation tools help to understand the behavior of machine learning algorithms and the obtained results. Such tools map an abstract concept, such as the behavior of a neural network, into a domain which is understandable by a human (Montavon et al., 2018). Many approaches utilize visual heatmaps, which indicate the saliency/-sensitivity of the output by means of the input, attention of classifier models, or feature importance and relevance (Hohman et al., 2018). These tools are extremely helpful, but have only recently been used to derive new scientific knowledge and discoveries, and to improve the learned model (Schramowski et al., 2020, Roscher et al., 2020).

In this paper we apply Occlusion Sensitivity Maps (OSM), (Zeiler, Fergus, 2014)) to a model that was developed for the identification of whales. Our main contribution is the analysis of the generated heatmaps and the question to what extent the quality of the model and certain observations can be verified with them and how the findings can help a whale researcher. This also includes the research question of whether the identification of whales by a neural network uses the same image characteristics as those considered important by whale experts. We base our analysis on the approach presented by (Lapuschkin et al., 2019), however, with focus on these specific objectives:

- Can interpretation tools help to determine the suitability of images for identification? Suitability is influenced, for example, by image quality, object pose and size, but also by the presence of relevant features.
- Can interpretation tools help to determine the reliability of the prediction of a neural network? Related to this is the question whether the interpretation tools agree with the statements of the neural network confidence scores or possibly supplement them.
- Can interpretation tools help to make a statement about temporally variable characteristics? The whale fluke changes with age. As a result, distinctive features of a fluke may no longer be present and thus, are not visualized in the results of the interpretation tool.

These objectives are formulated from the perspective of whale researchers, but also raise relevant questions from a machine learning perspective, such as the usefulness of interpretation tools to improve models. In general, the task of re-identification of objects or living beings from images is a widespread topic in photogrammetry and remote sensing and the approach and findings presented in this paper can also be applied to similar tasks.

The paper is structured as follows. We start with the description of the humpback whale data set and the identification model used (Sec. 3). Afterwards, we will introduce the visualization tools Gradient-weighted Class Activation Mapping (Grad-CAM) and Occlusion Sensitivity Maps (OSM) that are used in this paper. Furthermore, we shortly describe Spectral Clustering, which we employ to group heatmaps (Sec. 4). Finally, we present our results in Sec. 5, which include the comparison between the tools Grad-Cam and OSM as well as a more detailed analysis using Spectral Clustering of the resulting OSM heatmaps.

2. INTERPRETABLE MACHINE LEARNING

Besides maximizing the accuracy of a learning algorithm, the focus for an increasing number of applications is on the explainability of results in general and the explainability of model behavior in particular. A prerequisite of explainability is interpretability, which transforms complex aspects like the model behaviour into a space that can be understood by humans. By combining interpretable models and results with domain knowledge, explanations can be derived.

In recent years, various instruments have been proposed which approach interpretability in different ways. (Samek, Müller, 2019) distinguish between four different approaches. The first approach is the usage of surrogates, where complex models such as neural network (NN) models are approximated by simpler ones which are interpretable. One of the most well known ones in this category is Local Interpretable Model-Agnostic explanations (LIME) described by (Ribeiro et al., 2016). This model samples around an input in feature space and approximates the predictions by fitting a local surrogate model. In this way, it helps to understand why the model makes a certain prediction for a specific input. The second approach analyses the model's change regarding pertubations of the input, that means, it analysis the sensitivity of a model. OSM (Rajaraman et al., 2018), which is the main focus of this paper, belongs to this group. The third group are propagation-based approaches, which exploit the internal structure of the model. A well-known representative of this group is layer-wise relevance propagation (Bach et al., 2015) which redistributes the prediction backwards and assigns a relevance to each input element (e.g., a pixel). Another approach belonging to this group is deconvolution (Zeiler, Fergus, 2014). The last group are meta-explanations, which are also used in this paper. Here, several interpretations are analyzed together to get insights about the general behavior of the learned model. Spectral relevance analysis uses visualization tools to produce sets of heatmaps, which are further analyzed by clustering (Lapuschkin et al., 2019). For further details, including specific types of interpretation and further realization,

[1] https://www.kaggle.com/c/humpback-whale-identification
[2] 1st place: https://github.com/earhian/Humpback-Whale-Identification-1st-
[3] 2nd place: https://github.com/SeuTao/Humpback-Whale-Identification-Challenge-2019_2nd_palce_solution
[4] 3rd place: https://github.com/pudae/kaggle-humpback

Figure 2. Two samples from the Kaggle Challenge 'Humpback Whale Identification' showing two different whale individuals. The re-identification of whales states a challenge due to low inter-class variations.

we refer to recent surveys (Adadi, Berrada, 2018, Gilpin et al., 2018, Guidotti et al., 2018).

3. HUMPBACK WHALE DATA AND IDENTIFICATION MODEL

3.1 Image Data

In this work, we use a set of humpback whale images from the Kaggle Challenge "Humpback Whale Identification". More specifically, we process their tails, the so-called flukes (see Fig. 2). The data set consists of more than 67.000 images, in which 10.008 different whales, i.e., 10.008 different classes, are represented. An image of size $u \times v$ is represented as a vector $\boldsymbol{x}$ of length $(u \cdot v)$. The images are splitted into a training set $\mathcal{X}_{\text{train}} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ ($\sim$ 51.200 images) and a test set $\mathcal{X}_{\text{test}} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T\}$ ($\sim$ 16.000 images). The number of images per set is given by N and T, respectively. The set $\mathcal{X}_c = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_R\}$ describes a subset that includes R images for one specific class c. For our experiments, we restrict ourselves to use images in the training set $\mathcal{X}_{\text{train}}$, because the test set $\mathcal{X}_{\text{test}}$ does not provide reference information, as it is generally the case for Kaggle challenges.

3.2 Whale Identification Framework

For our study, we use the second winner solution[3] of the Kaggle Challenge, which is in our view best suited with regard to its documentation and the application of interpretation tools. The basic structure of the first[2] and third[4] winner solutions is based on a keypoint detection, which cannot be easily analysed with interpretation tools. Furthermore, the achieved accuracies of all solutions do not differ significantly.

For pre-processing, the framework applies two steps to the raw image. First, the chosen framework automatically performs image cropping in order to reduce the image content to the fluke of the whale. The cropped images are resized to an uniform size of $256\,\text{px} \times 512\,\text{px}$. This is also helpful in that the heatmaps generated by the interpretation tool focus on the essential image content, namely the fluke. In the second step, the framework performs z-normalization on the input images, by centering the data by subtracting the mean from every pixel and dividing it by the standard deviation (Cheadle et al., 2003).

The architecture is based on ResNet-101 (He et al., 2016), a deep neural network with 101 layers composed of multiple 3-layer blocks, where each block contains several convolutional, normalization, and pooling operations. Moreover, skip-connections are employed so that the gradients can flow backwards while skipping several layers. The optimization of the parameters is done by minimizing a weighted sum of three losses: triplet loss (Weinberger, Saul, 2009), ArcFace loss (Deng et al., 2019), and focal loss (Lin et al., 2017). For each input image, we receive a score $\boldsymbol{s}_{nc}$ for each class c. If we sort the scores in descending order, we consider the five classes with the highest scores. We call them top b with $b = \{1, \ldots, 5\}$. For example, top 2 represents the class with the second highest score. In general, we denote as reference assignment the one of the top b that matches the reference. The top 1 prediction corresponds to the network's class prediction for the input image $\boldsymbol{x}_n$. We denote the case that the top 1 prediction from the neural network corresponds to the reference class for the input image as correct assignment. We train the model with pre-processed images $\boldsymbol{x}_n \in \mathcal{X}_{\text{train}}$, while ensuring the reproducibility of the specified accuracies from the challenge.

4. HEATMAP ANALYSIS

4.1 Heatmapping

4.1.1 Gradient-weighted Class Activation Mapping (Grad-CAM) Grad-CAM is a method for visually interpreting deep networks via gradient-based localization (Selvaraju et al., 2017). The idea is to highlight the regions in an image, which are important for predicting a particular output. For this, the gradient for a confidence score of class c with respect to the activations in a chosen convolutional layer is computed, and pooled to derive importance weights for each neuron. In combination with the activations in the respective layer it results in heatmap with a size corresponding to the activation map. One exemplary result is presented in Fig. 3. Note, (Selvaraju et al., 2017) point out that deeper layers provide more interpretable results than shallower layers. Moreover, we observed that the resolution of the heatmap is crucial for good interpretability. The underlying resolution of Grad-CAM results corresponds to the spatial resolution of the used activation map.

Figure 3. Grad-CAM results for selected feature map resulting of the fourth residual block of the ResNet101 network.

4.1.2 Occlusion Sensitivity Map (OSM) OSM is a strategy developed by (Zeiler, Fergus, 2014) to evaluate the sensitivity of a trained model to partial occlusions in an image. The use of OSM helps to identify whether the trained model classifies the input based on task-specific features or whether the surrounding context is included in the classification. Moreover, it shows which regions make a positive contribution to the score and which make a negative one. If the heatmap outcome is a high absolute value at a given pixel position, it can be concluded that changing this pixel would have a significant effect on the classification result. This provides understanding of the learned behaviour of the model, based on the underlying task.

For a given image, different patches are masked. This is done by using a patch, which is moved over the image with a selectable stride. Two parameters, namely patch size p and stride, are chosen by the user, where the choice influences the result in terms of precision and smoothness. In the area occluded by the patch around position u, the pixel-wise scores of the classifier for each class are compared to the obtained scores after a part of the image was occluded. The difference $\Delta \boldsymbol{s}_{cu}$ is given by

$$\Delta \boldsymbol{s}_{cu} = \boldsymbol{s}_c - \tilde{\boldsymbol{s}}_{cu} \tag{1}$$

Figure 4. Occlusion sensitivity map as an overlay over an example image of the humpback whale identification data set, which shows an image of class 2155. The result is computed for a patch size $p = 5$px. Pixels colored green illustrate regions whose overlap reduces the confidence score of class c, and which are therefore important for the classification of class c. Purple pixels indicate regions whose coverage increases the score, i.e. these regions point towards other classes. These areas can confuse the network during the classification process if these areas are located in the object.

where the original predicted score for each class is denoted by $\boldsymbol{s}_c$ and the predicted score based on occlusion is given by $\tilde{\boldsymbol{s}}_{cu}$. Performed for the whole image, it results in an occlusion sensitivity heatmap.

Figure 4 shows an example of OSM, where green and purple colored areas show the most sensitive pixels towards occlusion. If green areas are covered, this leads to a significant drop of the score of the correct class c. We call those areas positive sensitive. If purple colored areas are covered, the score of the correct class c increases, i.e. these regions point towards other classes and an occlusion helps predicting the correct class. We denote those pixels negative sensitive. The white colored areas indicate no measured influence on the classification.

Varying patch size and stride allows for flexibility in the generation of OSM results. It leads to a variable smoothing of the map and therefore, a varying precision. This allows to capture features of different sizes in the image. However, depending on the selected parameters, this leads to an increased runtime.

4.2 Cluster Analysis

In order to make general statements about the interpretations results of the entire data set, we perform a clustering of the obtained heatmaps per image. In doing so, we follow the basic idea of (Lapuschkin et al., 2019). We use the method of Spectral Clustering (SC), where a data set is clustered based on a new representation of the data, which is based on a similarity measure (Meila et al., 2016). The similarities are used as weights in a so-called weight matrix $\boldsymbol{W}$, from which a graph Laplacian matrix $\boldsymbol{L}$ is obtained. The SC method uses the eigendecomposition of $\boldsymbol{L}$ to determine the clusters using k-means.

The data set consists of N samples $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_N$, representing the heatmaps for each training sample $\boldsymbol{x}_n$. For a faster runtime and more stable computations, we reduce the dimensions of the data points by a 4×4 max pooling operation. Thus, the length of a vectorized heatmap change from 131072 to 8192. We normalize the data to a range of $[0, 1]$. Afterwards, we perform principal component analysis on the normalized data to reduce the dimensions of the data to 1000. A weight matrix $\boldsymbol{W}$ is constructed from a similarity graph, where for the determination of the similarities we use the Gaussian similarity function based on the Euclidean distance $\boldsymbol{W}_{nm} = \exp(-\|\boldsymbol{z}_n - \boldsymbol{z}_m\|^2/2\sigma^2)$. The kernel scale is set to $\sigma = 0.2$ and chosen by an evaluation of eigenvalues obtained from the weight matrix $\boldsymbol{W}$. A suitable kernel scale is indicated by a significantly different eigenvalues and clear eigengaps.

We compute an eigendecomposition of the normalized graph Laplacian $\boldsymbol{L}_{\text{sym}} = \boldsymbol{I} - \boldsymbol{D}^{-1/2}\boldsymbol{W}\boldsymbol{D}^{-1/2}$, (Ng et al., 2002). Here, $\boldsymbol{D}$ is a diagonal matrix, where a diagonal entry is the sum of the weights in the graph for a heatmap $\boldsymbol{z}_n$. Note, in SC one uses the eigenvectors $\boldsymbol{U}_K := [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_K]$ for the smallest K eigenvalues when aiming for K clusters. In a further step, the rows of $\boldsymbol{U}_K$ are normalized by $Q_{nm} = u_{nm} / \left(\sum_k u_{nk}^2\right)^{1/2}$ to norm 1 and organised in a matrix Q. The K-dimensional vector $\boldsymbol{y}_n$ corresponding to the n-th row of Q, for $n = 1, \ldots, N$, gives a new representation for $\boldsymbol{x}_n$ that enhances the cluster-properties in the data (Von Luxburg, 2007). Using the k-means algorithm (Hartigan, Wong, 1979) the data are then divided into K clusters $\{C_1, \ldots, C_K\}$ based on the vectors $\boldsymbol{y}_n, n = 1, \ldots, N$.

5. RESULTS

5.1 Experiment 1: Comparison OSM to Grad-CAM

This experiment shows a comparison of OSM and Grad-CAM heatmaps. The results support our key claims: (i) The visualization tools provide interpretable results. (ii) The features identified by the network are similar to the features that whale experts consider important. The experiment is based on the data set explained in Sec. 3.1. As an example, we consider results for set $\mathcal{X}_{261}$ of four images representing class 261, as illustrated in Fig. 5. The images are classified by the neural network described in Sec. 3.2. Both visualization tools, OSM and Grad-CAM, are applied to the model. The results of OSM are calculated for patch size $p = 5$ and $p = 35$ and a stride of 1, resulting in a heatmap of size $256\,\text{px} \times 512\,\text{px}$. Previous experiments have shown that deeper layers are more interpretable (Selvaraju et al., 2017). Therefore, we visualize the resulting feature maps of the fourth residual block (denoted as layer 4) of the model.

Using OSM, the variation of patch size p and stride allows flexibility in generating results. Changing the parameters causes different characteristics in the image to be highlighted and displayed at different resolutions. Depending on the class mapped, either trailing edge, pigmentations, or the total fluke are highlighted as positive sensitive by OSM. For the class shown in Fig. 5, with $p = 5$ the white pigmentations on the fluke are shown as positive sensitive and thus, as important feature that the network uses for the prediction of class 261. With $p = 35$ the trailing edge is highlighted as positive sensitive as well, but not as distinctive as the pigmentation. In contrast, Grad-CAM highlights the notch at the right tip of the fluke, where OSM does not consider these areas as influential.

In the following, we examine both tools in terms of the features considered important by the whale expert. OSM delivers a result that includes the relevance of all image areas and thus all characteristics, as the heatmap visualizes the influence of individual image regions on the score. Different patch sizes result in OSM results with varying degrees of smoothing. In contrast, Grad-CAM gives us information about a specific layer. In general, different layers are related to different areas in the image and often these can be associated with properties that are considered important also by a whale expert. Due to the visualization of individual layers rather than the summary of the model, we cannot easily assess the specific importance of the highlighted feature with regard to the classification decision.

Besides the visual results, we compare the spatial resolution additionally to the runtime of both methods. The spatial resolution of the resulting heatmap varies depending on the setting

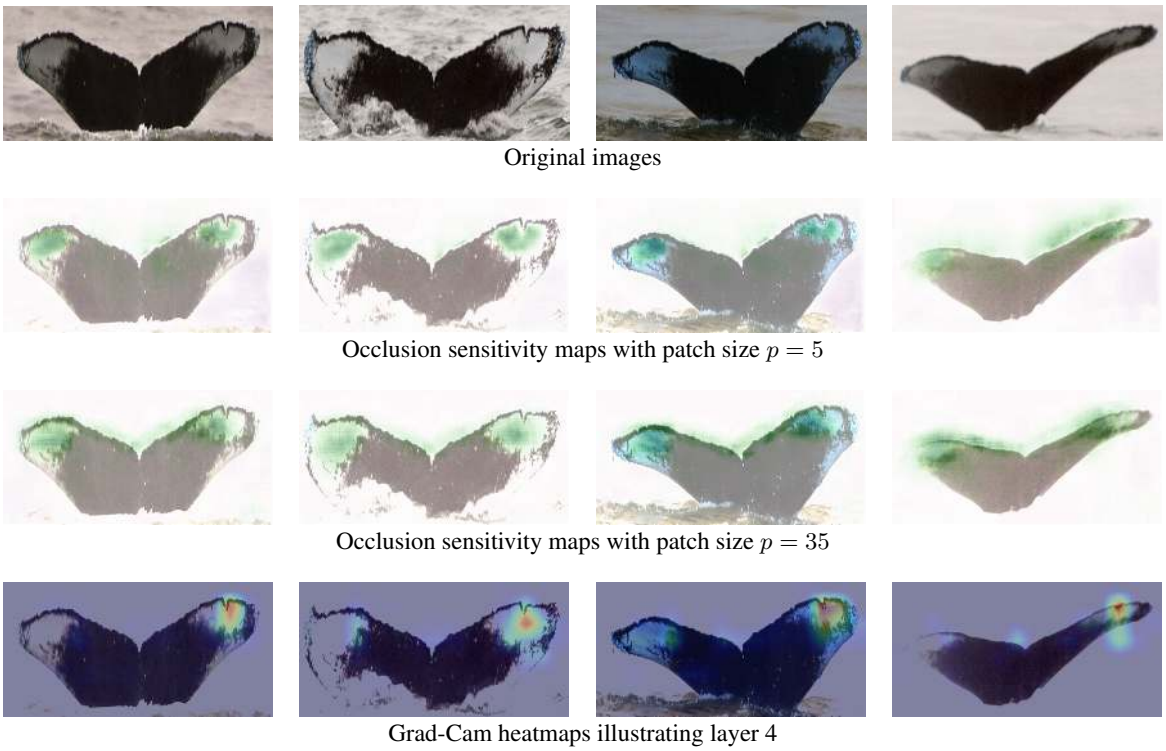

Figure 5. Visual comparison between OSM and Grad-CAM results using images representing class 261. OSM is calculated for patch size $p = 5$ and $p = 35$. Green colored areas represent pixels that have a positive influence on the class shown. Purple pixels are indications of other classes. The white colored areas have no influence on the classification according to OSM. The selected layer to be visualized in the approach of Grad-CAM is *Layer 4*. For a better visualization, we overlayed the heatmap with the original image.

for the stride. For Grad-CAM, the resolution is equal to the size of the visualized feature map. The choice of the resolution is of central importance, as it determines the visibility of the features given by the whale expert and thus, the interpretability of the resulting heatmaps. Besides the different outputs, which allow different statements about the importance of image regions, the procedures also differ in terms of the duration. Regarding the runtime, Grad-CAM needs around 30 second to compute one heatmap of size $256\,\text{px} \times 512\,\text{px}$. In contrast, OSM needs 30 minutes for a stride of one. A higher stride would reduce the runtime. For the calculation, we use an Intel Core i7-6850K 3.60 GHz processor and a Geforce GTX 1080Ti with 11 GB RAM.

To answer the question how the gained insights of the heatmaps can help a whale researcher, we refer to our prior knowledge that certain characteristics on the whale fluke change over time, such as the area of the trailing edge near the fluke tips. If a long period of time passed between two images of the same whale, there is a risk that the network will not be able to match them correctly. With this knowledge the recently photographed whales could be compared manually to whales already existing in the database. Since it is not possible to compare all whales manually, the following opportunity can help the whale expert to reduce the selection. A possibility could be that the expert selects whales whose local characteristics, which change with time, are similar to the whale of interest. These could be whales, for example, which also show characteristic features in the area of the trailing edge near the fluke tips, as here changes over time are likely. In this case, the features detected by the visualization tools in combination with domain knowledge from the whale expert provide information about the reliability of recognizing a whale in the future.

For both methods, we could not observe that the model pays attention to artifacts or surrounding objects except water, the latter is due to the fact that in the pre-processing chain the irrelevant part of the image has been cut away for the most part.

5.2 Experiment 2: Clustering analysis of OSM heatmaps

For this experiment, we use images from the training set $\mathcal{X}_{\text{train}}$, which are classified by the neural network described in Sec. 3.2. We apply the visualization tool OSM to the model and receive one heatmap per top b predictions per image with $b = 1, \ldots, 5$. To put more focus on fine-grained structures and promote more distinct clustering, we use the results for $p = 5$ only.

Following (Lapuschkin et al., 2019), the goal of this experiment is to find a clustering of heatmaps that highlights and separates different features in the images. In combination with the detection rate of the model associated with certain heatmaps, the quality of different clusters and thus the quality of certain images can be determined. This is done by applying SC (Sec. 4.2) analysis on the previously calculated occlusion sensitivity heatmaps. We consider heatmaps assigned to correct assignments within the top 5 predictions.

In order to decide on a suitable number of clusters, we analyzed whether we can identify gaps in the eigenspectrum of the graph Laplacian. As can be seen in Fig. 6, the eigenvalues exhibit large gaps for the first few eigenvalues. Contrary to our expectations, with a small number the resulting clusters are not meaningful for our research question, since the clusters contain mixed characteristics of heatmaps. Due to the generally very similar content of the images, a separation by a few clusters is very challenging. Therefore we decided to use many, and more fine-grained clusters, whereby optionally several clusters can be

regrouped afterwards. We have seen that the choice of $K = 30$ leads to visually distinguishable and meaningful clusters which support our research question.

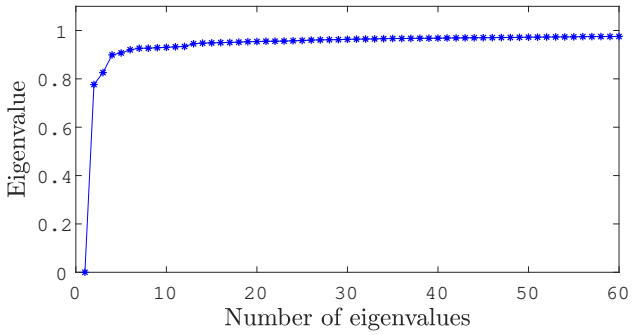

Figure 6. The 60 smallest eigenvalues from the Spectral Clustering analysis based on the calculated occlusion sensitivity heatmaps.

As seen in Fig. 7, with the approach we are able to find meaningful clusters. The figure shows images which are assigned to three different clusters. Fig. 7 a) includes heatmaps, where the area of the entire fluke is predominantly colored green and therefore is positive sensitive. The rest of the image is very bright and weak colored. This could be an indicator of the shape of the fluke. Fig. 7 b) includes predominantly images where the network considers the pigmentation near the fluke tips as important according to the OSM results. The lower right corner of each image indicates a purple coloration. This could be an indicator for a color change. For the last cluster in Fig. 7 c) one does not observe any striking coloration of the entire image. The regions of the whale fluke are not positive sensitive to the class presented. The background varies between a light green and light purple coloration. No specific features are highlighted in these images. Thus, clusters like c) are not informative enough to answer the research question. Cluster a) and b) are positive examples that some of the clusters indicate characteristics of the whale that are important to the whale researcher, such as different kinds of pigmentations, the trailing edge, or the total area of the fluke.

Cluster relation scores

For each original input image the five highest scored classes are analysed with OSM. The upper figure in Fig. 8 shows for each cluster the relative proportion of correct assignments using the top five scored classes in relation to the total number of heatmaps assigned to that cluster. It can be observed that essentially the highest scored class (top 1) already gives the correct assignment. The bottom figure in Fig. 8 shows for the correct class the mean of the predicted score and standard deviation from the neural network, taken over all OSM heatmaps within a cluster.

For the bottom figure in Fig. 8 a decreasing trend is also visible. The clusters that contain a high proportion of correct assignments from the top five scores have also higher average scores. The neural network's predictions for images corresponding to these heatmaps are therefore more reliable. The cluster's standard deviations are relatively large. Certain classes have no distinctive features, so that only small regions are highlighted in the heatmap. However, in some cases, the presence of these small regions is sufficient to represent the respective class. Therefore, heatmaps assigned to a cluster with a low mean score can also have a high score. For instance, cluster 28 contains a heatmap assigned to a score of 71.02%. In fact, small interclass variations can cause the heatmaps to appear similar, but the scores can differ significantly from each other. As a consequence of the similar appearance, the images are assigned to the same cluster. However, this results in an increased standard deviation.

Reliability analysis

In order to draw conclusions about the quality and reliability of the heatmaps, we consider outliers in clusters with contradicting mean scores. Furthermore, we examine the heatmaps in specific clusters and the corresponsing predictions of the neural network. For this, we consider the clusters from Fig. 8. In cluster 1, more than 90% of its heatmaps belong to a top 1 prediction of the neural network. Additional heatmaps belong to a top 2 prediction, i.e., the one with the second highest score. The scores of the top 2 predictions in this cluster are significantly smaller than the top 1 predictions, where the lowest top 2 prediction has a score of 50.20%. In cluster 29, which is characterized by a low mean score, 50% of all images have similar low scores for all classes, which is an indicator that the photographed whale is identified as a new whale which is not yet in the database. If there are no images of a whale to be identified in the database, identification by the neural network is not possible, which can be determined by low scores for all classes. This is also reflected in the appearance of the heatmaps of OSM, which do not capture essential characteristics of a whale. Several other clusters with a lower mean score have the same properties as cluster 29.

A small percentage of heatmaps in cluster 29 belong to the top-1 score, but are not classified correctly, i.e., the network could not correctly identify the whales. Visually, these heatmaps are similar to those for new whales. Overall, no heatmaps in this cluster belong to correctly identified whales. In general, we observe that clusters with a low mean score have less correct identifications. Heatmaps that have a top-1 score and are assigned to these clusters are often identified as new whales.

Based on the observed aspects, we conclude that heatmaps assigned to clusters with low mean score, tend to be incorrectly classified by the neural network. We suggest that whale researchers do not to trust these predictions but to check the images manually afterwards instead. One reason for the assignment of a heatmap to one of those clusters could be the low visibility of sufficient features. The information that a prediction is assigned to one of these clusters could result in a proposal for the whale researcher to take additional pictures of that whale.

6. CONCLUSION AND FUTURE DIRECTIONS

In this paper, we applied the two visualization tools Gradient-weighted Class Activation Mapping and Occlusion Sensitivity Map to a given neural network to gain insights into which features are relevant for the decision of the network in the context of whale identification. Our experiments show that both interpretation tools highlight similar image characteristics as the whale expert is focusing on. By means of a cluster analysis of the interpretation results of occlusion sensitivity maps, we could show that characteristics can be grouped in a meaningful way. These automatically determined characteristics correspond to those that are also considered important by a whale expert. However, in contrast to the whale expert, some interpretation tools like the occlusion sensitivity maps indicate that

a) Heatmaps of cluster 1 b) Heatmaps of cluster 2 c) Heatmaps of cluster 29

Figure 7. Spectral Clustering analysis is applied to the OSM heatmaps to identify different prediction characteristics within the analyzed data. Three exemplary prediction characteristics are shown. a) Detect total fluke as positive sensitive. b) Detect pigmentations near the fluke tips as positive sensitive. c) Nothing specific is detected positive sensitive. No distinctive features are found.

the network usually focuses on only one type of characteristic, such as parts of the fluke. The whale expert, on the other hand, usually looks at all the features together.

We were able to show a relationship between heatmaps obtained from an interpretation tool and scores obtained from a neural network. Through cluster analysis we could identify clusters of heatmaps that are not suitable for the identification of whales. Even if the network gives a high score, an assignment to such a cluster indicates a misclassification by the network. These assignments can be used to make recommendations to the whale researcher that it is useful to take another photo of the whale to better identify the visible features and thus improve the identification results. Thus, we consider the score and the output from the heatmap analysis as complementary measures for the reliability of the result.

In future work, the newly gained information from the cluster analysis could be used to improve the network. This information could be fed as additional information into the training of the neural network, as suggested by (von Rueden et al., 2020). Another research approach is the comparison of different interpretation tools. In preliminary analyses we could show that Gradient-weighted Class Activation Mapping considers different characteristics in different layers as important, which is more in accordance with the perception of the whale researcher. Thus, a combination of several layer-based Gradient-weighted Class Activation Mapping heatmaps could provide a more comprehensive statement than occlusion sensitivity maps.

We also want to encourage further research in this new yet promising area. Interpretable and explainable machine learning can be applied in many places in photogrammetry and remote sensing where questions about the reliability of models and observations are relevant, and further insights besides the obvious ones from the model are of interest.

ACKNOWLEDGEMENTS

We would like to thank Happywhale[5] for providing the data set.

[5] https://happywhale.com/home

REFERENCES

Adadi, A., Berrada, M., 2018. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, 6, 52138–52160.

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., Samek, W., 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7), e0130140.

Cheadle, C., Vawter, M. P., Freed, W. J., Becker, K. G., 2003. Analysis of microarray data using Z score transformation. *The Journal of molecular diagnostics*, 5(2), 73–81.

Deng, J., Guo, J., Xue, N., Zafeiriou, S., 2019. Arcface: Additive angular margin loss for deep face recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4690–4699.

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., Kagal, L., 2018. Explaining Explanations: An Overview of Interpretability of Machine Learning. arXiv preprints arXiv:1806.00069.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D., 2018. A Survey of Methods for Explaining Black Box Models. *ACM Computing Surveys*, 51(5), 1–42.

Hartigan, J. A., Wong, M. A., 1979. Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1), 100–108.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hohman, F. M., Kahng, M., Pienta, R., Chau, D. H., 2018. Visual Analytics in Deep Learning: An Interrogative Survey for the Next Frontiers. *IEEE Transactions on Visualization and Computer Graphics*, 25(1), 1–20.

Katona, S., Whitehead, H., 1981. Identifying humpback whales using their natural markings. *Polar Record*, 20(128), 439–444.

Kniest, E., Burns, D., Harrison, P., 2010. Fluke Matcher: A computer-aided matching system for humpback whale (Megaptera novaeangliae) flukes. *Marine Mammal Science*, 3(26), 744–756.

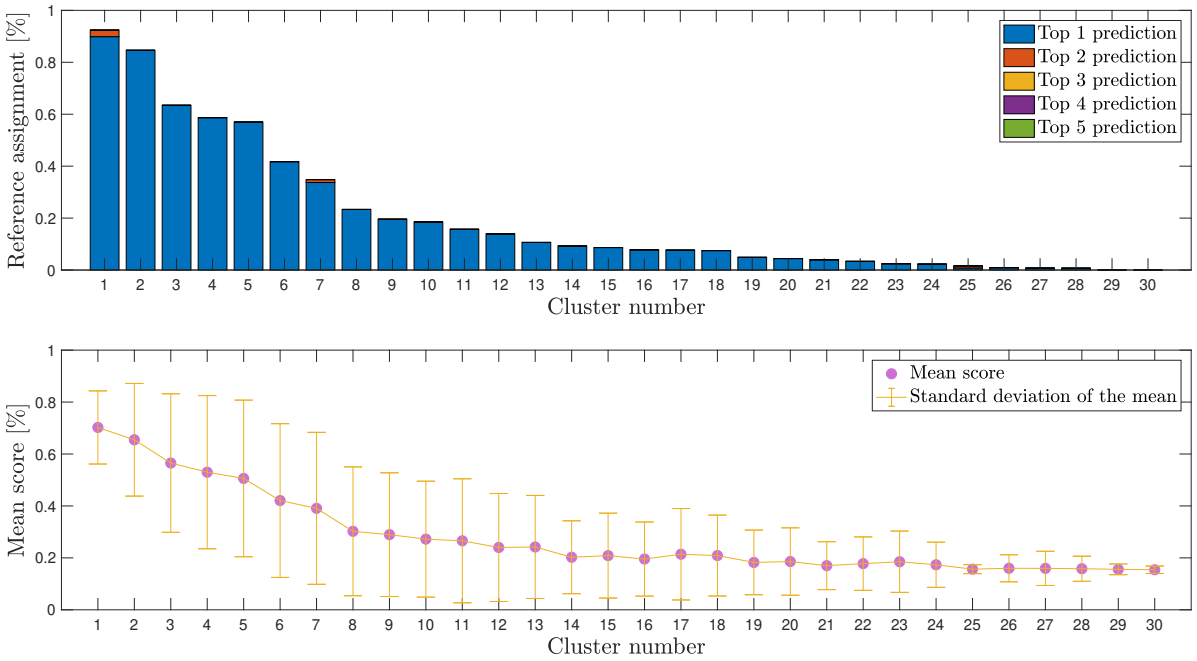

Figure 8. The figure on the top shows the proportion of correct class assignments within a cluster using the five highest scored classes, where the clusters are in descending order of that proportion. In the bottom figure the mean score as well as the standard deviation of the scores for the images of cluster k are plotted, where the clusters are in the same order as above.

Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., Müller, K.-R., 2019. Unmasking clever hans predictors and assessing what machines really learn. *Nature communications*, 10(1), 1–8.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P., 2017. Focal loss for dense object detection. *Proceedings of the IEEE international conference on computer vision*, 2980–2988.

Meila, M., Christian, H., Marina, M., Fionn, M., Roberto, R., 2016. Spectral clustering: a tutorial for the 2010's. *Handbook of cluster analysis*, CRC Press, 1–23.

Montavon, G., Samek, W., Müller, K.-R., 2018. Methods for Interpreting and Understanding Deep Neural Networks. *Digital Signal Processing*, 73, 1-15.

Ng, A. Y., Jordan, M. I., Weiss, Y., 2002. On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems*, 849–856.

Rajaraman, S., Silamut, K., Hossain, M. A., Ersoy, I., Maude, R. J., Jaeger, S., Thoma, G. R., Antani, S. K., 2018. Understanding the learned behavior of customized convolutional neural networks toward malaria parasite detection in thin blood smear images. *Journal of Medical Imaging*, 5(3), 034501.

Ribeiro, M. T., Singh, S., Guestrin, C., 2016. Why should i trust you?: Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, ACM, 1135–1144.

Roscher, R., Bohn, B., Duarte, M. F., Garcke, J., 2020. Explainable machine learning for scientific insights and discoveries. *IEEE Access*, 8, 42200–42216.

Samek, W., Müller, K.-R., 2019. Towards explainable artificial intelligence. *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, Springer, 5–22.

Schneider, S., Taylor, G. W., Linquist, S., Kremer, S. C., 2019. Past, present and future approaches using computer vision for animal re-identification from camera trap data. *Methods in Ecology and Evolution*, 10(4), 461–470.

Schramowski, P., Stammer, W., Teso, S., Brugger, A., Luigs, H.-G., Mahlein, A.-K., Kersting, K., 2020. Right for the Wrong Scientific Reasons: Revising Deep Networks by Interacting with their Explanations. *arXiv preprint arXiv:2001.05371*.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. *Proceedings of the IEEE International Conference on Computer Vision*, 618–626.

Surma, S., Pitcher, T. J., 2015. Predicting the effects of whale population recovery on Northeast Pacific food webs and fisheries: an ecosystem modelling approach. *Fisheries Oceanography*, 24(3), 291–305.

Von Luxburg, U., 2007. A tutorial on spectral clustering. *Statistics and computing*, 17(4), 395–416.

von Rueden, L., Mayer, S., Beckh, K., Georgiev, B., Giesselbach, S., Heese, R., Kirsch, B., Pfrommer, J., Pick, A., Ramamurthy, R., Walczak, M., Garcke, J., Bauckhage, C., Schuecker, J., 2020. Informed machine learning - A taxonomy and survey of integrating knowledge into learning systems. *arXiv preprint arXiv:1903.12394*.

Weinberger, K. Q., Saul, L. K., 2009. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, 10(Feb), 207–244.

Zeiler, M. D., Fergus, R., 2014. Visualizing and understanding convolutional networks. *European conference on computer vision*, Springer, 818–833.