INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Relevance-based Feature Masking: Improving Neural Network based Whale Classification through Explainable Artificial Intelligence

Dominik Schiller[1*] *, Tobias Huber*[1*]*, Florian Lingenfelser*[1*]*, Michael Dietz*[1]*, Andreas Seiderer*[1]*, Elisabeth André*[1]*,*

[1]Human Centered Multimedia, Augsburg University, Augsburg, Germany
schiller,huber,lingenfelser,dietz,seiderer,andre@hcm-lab.de

Abstract

Underwater sounds provide essential information for marine researchers to study sea mammals. During long-term studies large amounts of sound signals are being recorded using hydrophones. To facilitate the time consuming process of manually evaluating the recorded data, computational systems are often employed. Recent approaches utilize Convolutional Neural Networks (CNNs) to analyze spectrograms extracted from the audio signal. In this paper we explore the potential of relevance analysis to enhance the performance of existing CNN approaches. For this purpose, we present a fusion system that utilizes intermediate outputs of three state of the art CNNs, which are fine tuned to recognize whale sounds in spectrograms. Hereby we use Explainable Artificial Intelligence (XAI) to asses the relevance of each feature within the obtained representations. Based on those relevance values, we create novel masking algorithms to extract significant subsets of respective representations. These subsets are used to train an ensemble of classification systems that are serving as input for the final fusion step. We observe that a classification system can benefit from the inclusion of Relevance-based Feature Masking in terms of improved performance and reduced input dimensionality. The presented work is part of the INTERSPEECH 2019 Computational Paralinguistics Challenge.

Index Terms: Computational Paralinguistics, Deep Neural Networks, Transfer Learning, Explainable Ai

1. Introduction

The application of passive acoustic methods to observe marine mammals has been of interest to researches for over three decades [1]. To this end, underwater microphones called hydrophones are often employed to acquire increasingly large datasets containing sound samples from vocally active marine species. The resulting audio recordings serve a variety of purposes, including tasks like species-identification [2, 3], localization-tracking [4, 5], behaviour-analysis [6, 7] or population monitoring [8]. In the past, these tasks were usually performed by small groups of experts through manual analysis of the audio recordings. However, due to the increasing size of the collected acoustic databases, this process is becoming more and more time consuming [3]. Additionally, the manual inspection can lead to inconsistencies based on the experience and fatigue of the analysts [8]. In order to facilitate this process, the application of automatic classification systems, which are able to detect specific relevant patterns in the data, has been growing in popularity [3]. Examples for this include the calculation of spectrogram correlation values [9, 10], extraction of frequency contours using edge detection algorithms and computation of pixel-based features [11], as well as the application of a principal component analysis to derive features from the relative power of frequency bins in spectrograms [12]. Due to the recent success of Convolutional Neural Networks (CNNs) in the fields of computer vision and natural language processing, current approaches also explore the feasibility of CNN-based models for bio-acoustic classification tasks such as whale call recognition. Instead of extracting features from the spectrograms, these approaches directly use them as inputs for the CNNs. For instance, Smirnov [13] trained a custom CNN with three convolutional layers to detect whale calls in two second audio clips and achieved an Area Under under the Receiver Operating Characteristic Curve (AUC) value of 0.976 on the dataset provided in the Marinexplore and Cornell University Whale Detection Challenge[1]. Using the same dataset, Ibrahim and colleagues [14] proposed a hybrid system which combines a CNN with a dictionary learning approach, to achieve a detection rate of 92.37%. Instead of creating new networks, Wang et al. [15] compared the performance of four established CNN model architectures (VGGnet, Inception, Xception and Densenet) on the open-source WhaleFM dataset[2]. In their approach they trained each model to detect the target whale call classes and achieved accuracies of up to 84.4%. Similarly, Zhang et al. [16] applied transfer learning methods to fine-tune pretrained models for whale call classification. In order to capture the characteristics of whale sounds with varying durations, they used three windows with different time scales and calculated feature maps with 1D convolutional layers, which were then combined into a 3-channel feature representation and fed into both networks. Using this approach they were able to distinguish two species of whales in the WhaleFM dataset[2] with an accuracy of 99.7%.

In general, CNNs are aiming to overcome the limitations of handcrafted features by directly learning suitable representations from raw data. However, the ability to handle raw data input with high accuracy comes with several challenges to be considered: First of all, large amounts of annotated data are necessary to train a Convolutional Neural Network from scratch, as the absence of handcrafted features requires additional abstraction layers to be automatically learned by the network. This behaviour initially hampers their application in niche topics with relatively small datasets available, but can be overcome by utilizing models that have initially been pretrained on larger data collections. Consequently we are deploying several pretrained CNNs for our experiments which we fine-tune to recognize whale characteristics in spectrograms.

A second drawback, that is common to all deep learning structures, is their inherent complexity and the resulting opaqueness in decision making. In recent years the need to bet-

[*]These authors contributed equally to this work

[1]https://www.kaggle.com/c/whale-detection-challenge
[2]https://whale.fm

http://dx.doi.org/10.21437/Interspeech.2019-2707

ter understand the decision process of neural networks has become an increasingly pressing problem. As a consequence the research field of Explainable Artificial Intelligence (XAI) [17] has reemerged and gained growing attention ever since. Explanation approaches like deep Taylor decomposition aim to identify the parts of an input which were relevant for the decision of a model. An example of such an explanation can be seen in Figure 1. The left image shows the spectrogram of a whale sound recorded with a hydrophone, while the picture on the right side visualizes the according deep Taylor decomposition. This visualization clearly shows that the network has learned to localize the relevant whale patterns even though they are superimposed by noise. Such XAI algorithms have already proven to be helpful for humans to understand the decisions of various machine learning models [18, 19]. Our goal is to transfer these insights into the classification system itself and translate this increased understanding into enhanced recognition accuracy. To this end we use deep Taylor decomposition to asses the relevance of features extracted by several pre-trained CNNs, fine-tuned to recognize orca sounds in spectrograms. Based on those relevance values, we create novel masking algorithms to extract significant subsets of respective representations. These subsets are used to train an ensemble of classification systems that are serving as input for a final fusion step.

In the following we test our approach within the INTERSPEECH Computational Paralinguistic Challenge (ComParE) [20]. The goal of this challenge is it to build an automatic system that detects the presence of orca whales in hydrophone recordings.

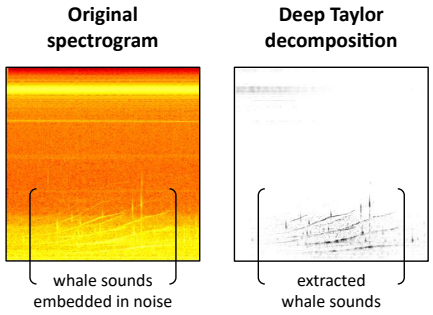

Figure 1: *Deep Taylor decomposition of an orca whale sample using an image processing CNN (Inception V3). The left image shows the original input to the network while the picture on the right highlights the areas that are relevant for the networks decision.*

2. Methodology

The following section provides an overview of our multi-level classification process (as illustrated in Figure 2) before describing the utilized components in detail.

2.1. Architecture

The first step in our classification system is based on spectrograms of the audio input data. Here, we use CNNs to learn suitable representations of any given sample - more specifically, the output of the last convolutional layer is used as a representative feature vector for further processing. In order to improve results with the limited amount of annotated training data available, we deploy pre-trained state-of-the-art CNNs, which we fine-tune to our problem. Next, the extracted representations are used to train fully connected neural networks with one hidden layer consisting of 256 neurons to detect the target classes (orca sounds versus noise). Those models are hereinafter referred to as base-models. At this point we introduce explainable artificial intelligence into the recognition architecture: XAI methods are applied to the base-models in order to assess the relevance of each input feature for classification results. After this intermediate analysis step, we train the final classification networks. Hereby, the calculated relevance values are used to generate masking layers within the networks, that automatically select interesting subsets of our initial feature representations. The process up to this point is depicted in Figure 2 and aims to force the newly trained networks to focus on specific parts of the input and therefore to learn differing patterns from the base models.

To take advantage of the different emphases of our models we fuse our trained models in two ways. First we conduct an intermediate feature fusion by training additional dense classification networks with one hidden layer consisting of 256 neurons on the concatenated features of all three feature extraction networks for each masking algorithm. Furthermore, we create a classification system for each combination of our different models using an average vote, where we base the decision of the whole system on the average confidence of all involved models. In this way we can find out which models complement each other the most.

2.2. Feature Extraction

Based on a broad evaluation of various CNN architectures for audio event detection carried out by Hershey et al. [21], we chose the following three convolutional neural network architectures to extract suitable feature representations for detecting the presence of orca whales in an audio file.

The VGGish model by Hershey et al. [21] is a variation of the original VGG image recognition model [22] that is specifically adapted to recognize sound scenes from spectrograms. The network is pretrained on the audio set data collection [23] - a large scale dataset which is labeled with respect to 623 different audio events. Mel-spectrograms are used as inputs for the network.

The Inception V3 network [24] is a popular choice for image recognition tasks. Hershey et al. [21] found that this particular architecture also yields top performance on the task of audio event detection, with respect to a limited amount of training time. As input we opted for power spectrograms instead of the Mel-spectrograms utilized in VGGish, since the mel scale, which has been designed with the the psychoacoustic perception of human listeners in mind, highly compresses information in higher frequency bands.

The third model we used for feature extraction is the InceptionResNet V2 model by Szegedy et al. [25]. This architecture is a combination of the previously described Inception Network and the ResNet architecture, which was found to achieve the best overall performance for sound event detection by Hershey et al. [21], at the cost of prolonged training duration. Here we are using the same input as for the Inception model.

2.3. Deep Taylor decomposition

To identify the extracted features that were especially relevant for our classification models we use an XAI method called deep Taylor decomposition which was initially introduced by Montavon et al. [26] to increase the interpretability of a classifier by highlighting the relevance of each input pixel in a heat map

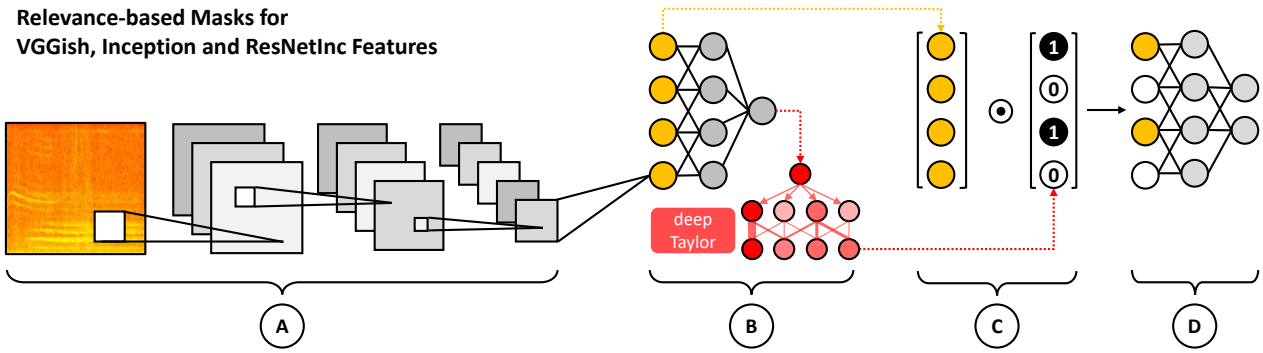

Figure 2: *Schematics of our classification models. At first a CNN extracts features from the input (A). Then we use XAI methods to analyze a classifier which is trained on those features (B). Based on this analysis we mask the features (C) and train a new classifier on the masked features (D).*

(see Figure 1). To this end deep Taylor decomposition assigns a relevance value R_i to each neuron of a neural network by performing a relevance propagation. This propagation starts at the output layer, where the relevance of the prediction we want to analyze is defined as the activation of the respective neuron. The relevance of this output neuron is then successively propagated backwards to each previous layer. During this relevance propagation a Taylor approximation is used to determine how relevant a neuron x_i^l in layer l was for a neuron x_j^{l+1} of the subsequent layer $l+1$. Hereby, the aim is to model the relevance of x_j^{l+1} as a function $R_j^{l+1}(x^l)$ which depends on the neurons x_i^l of the previous layer. Assuming such a function is found, for example through previous relevance propagation steps, one can decompose it using the Taylor series

$$R_j^{l+1}(x^l) = \sum_i \frac{\partial R_j^{l+1}}{\partial x_i^l}\Big|_{\widetilde{x}^l}(x_i^l - \widetilde{x}_i^l) + \varepsilon, \qquad (1)$$

with Taylor residual ε and base point $\widetilde{x}^l$ which is chosen depending on x_j^{l+1}. If one assumes that ε is small enough then the propagated relevance from x_j^{l+1} to x_i^l is given by $\frac{\partial R_j^{l+1}}{\partial x_i^l}\big|_{\widetilde{x}^l}(x_i^l - \widetilde{x}_i^l)$. Different deep Taylor methods vary in how they choose the base point $\widetilde{x}^l$. For this work we use the deep Taylor implementation of the INNvestigate framework [27].

2.4. Relevance-based Feature Masking

Based on the relevance values generated by the deep Taylor decomposition we use two different masking algorithms to extract interesting subsets of the learned representations. The first algorithm calculates the average relevance of each feature over the whole dataset. Those relevance values are then used to create a binary mask that sets all features but the n most relevant values to zero. By multiplying this mask with the input feature-vector we are eliminating the influence of all non-relevant features, when training a model. Since this is equivalent to a form of feature selection mechanism, which reduces the amount of utilized features, we refer to this approach as minimal masking. For our experiments we fixed n to be 512 which equals the number of features extracted by VGGish and therefore ensures that the intermediate feature fusion network trained on features from all three feature extraction networks is not distorted in favor of the larger feature sets of ResNetInc and Inception. The second masking algorithm dynamically generates a new mask for each sample by nullifying all features that have a higher than average relevance value. This forces the attention of a newly trained model to features which the original classification network has not considered as relevant. Hence we refer to this approach as negative masking.

2.5. Dataset

We run all our experiments on a collection of hydrophone recordings from the DeepAL Fieldwork Data, which where provided within the scope of the INTERSPEECH 2019 Computational Paralinguistics Orca Activity Sub-Challenge. All data was recorded with an array of four hydrophones and is available as either four- or mono-channel wav files. The data is predivided into three different sets for training, development and testing. Each set consists of small audio clips that are labeled with respect to the presence of an orca in a given recording. In all our experiments we use the training set to train our classifiers and the development set for evaluation. To increase the available amount of training data we split each four-channel-wav into four mono-channel files. However, for evaluation purposes we rely on the mono-channel files that were already provided. Overall 19364 samples (~6:30h) of data are used for training and 3515 (~1:10h) for evaluation. For details please refer to [20].

3. Summary of Results and Discussion

In the following we present the results of our conducted experiments on the task of orca detection with the aforementioned variations of our Relevance-based Feature Masking (RBFM) approach. Performance will be reported with respect to the Area Under under the Receiver Operating Characteristic Curve (AUC). To compensate for the underrepresentation of orca samples in the training data, we employ a weighted loss function in all cases during the training of our models. Table 2 lists the performances of our base-models, which are trained on the learned feature representations, as well as the feature fusion models, trained on a concatenated feature vector of those base-models only. Results are broken down according to the utilized selection masks. Selecting the most relevant subset of features (minimal masking) to train a new model does not considerably impact classification performance while greatly reducing the number of

Table 1: *Ranking of the decision level fusion of all possible classifier combinations sorted by their AUC performance on the development set. The respective contributing classifiers are noted by green checkmarks.*

No.	Vgg	Vgg_neg	Vgg_min	Inc	Inc_neg	Inc_min	Res	Res_neg	Res_min	FF	FF_neg	FF_min	AUC
1	✓	✓	✗	✗	✓	✗	✓	✗	✗	✗	✗	✓	.9194
2	✓	✓	✗	✗	✓	✗	✗	✓	✗	✗	✗	✓	.9193
3	✓	✓	✗	✓	✗	✗	✓	✗	✗	✗	✗	✓	.9193
4	✓	✓	✗	✓	✓	✗	✓	✗	✗	✗	✗	✓	.9192
5	✓	✓	✗	✓	✗	✗	✗	✓	✗	✗	✗	✓	.9191
⋮													⋮
1504	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	.9077
⋮													⋮
2036	✗	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗	.8852
⋮													⋮
2039	✗	✗	✗	✓	✗	✗	✗	✗	✗	✗	✗	✗	.8844
⋮													⋮
2043	✗	✗	✗	✗	✗	✗	✓	✓	✗	✗	✗	✗	.8832
2044	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗	✗	.8829
2045	✗	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	.8798
2046	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗	✗	✗	.8736
2047	✗	✗	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗	.8717

used features. Whenever we are forcing the model to focus on less relevant parts of a feature set (negative masking), we observe a small decrease in classification performance. The fact that the absence of the most relevant features leads to a worsened classification performance substantiates the relevance of features found by minimal masking. However, we will see that features found by negative masking contribute useful information to the fusion process.

Table 2: *AUC performance on the development set for all the base-models and the trained feature fusion using no-, minimal-, and negative-masking versions of those base-models only.*

Model	Masking		
	None	Minimal	Negative
Vggish	.9077	.9077	.8798
Inception V3	.8844	.8829	.8736
Resnet	.8852	.8871	.8717
Feature Fusion	.9000	.9007	.8880

Table 1 shows the summarized ranking of all our decision-level fusion experiments with respect to the inclusion of contributing models. The classification performance ranges from an AUC of 0.8717 for our weakest decision level fusion model to 0.9194 for the best performing model.

As expected, the weakest models are the ones that are only using the negatively masked features for training. Furthermore, the isolated and non-fused base-models are placed in the lower third of the scale. The top five models all include models trained on (1) the full representations extracted by VGGish, (2) the less relevant parts of those features and (3) the minimal feature fusion of all extracted representations. Either combination of the extracted features from ResNetInc and Inception, as well as the less relevant parts of those features are also contributing to the best performing models. In conclusion, the ranking reveals that both masking variants are included within the top five systems. This shows that while our masking approaches are not necessarily improving the performance of a single model, they indeed add additional value to an overall fusion approach.

Table 3: *Comparison of our best decision level classifier, as reported in Table 1, against the ComParE 2019 baseline.*

	Model	Dev	Test
Proposed	RBFM Fusion	**.919**	**.916**
Baseline	OpenSmile	.810	.866
	OpenXBOW	.771	.836
	AuDEEP	.740	.798
	Fusion	-	.866

4. Conclusion

In this work we have shown the potential of Explainable Artificial Intelligence approaches to enhance the performance of neural network classification. We have introduced novel Relevance-based Feature Masking algorithms that substantially improved performance over our base-models for the task of detecting whales in an audio signal. To put our results into perspective, we compare our best performing approach, as reported in Table 1, against the baseline of the 2019 ComParE challenge (Table 3). This year's baseline [20] comprises three different classification systems as well as a fusion of all three systems. Comparison of results shows that our RBFM Fusion approach outperforms all baseline approaches on the development- as well as on the test-set. For future work it might prove beneficial to investigate the generalization capabilities of our approach for other classification problems.

5. Acknowledgements

This work has received funding from the BMBF under FKZ 01IS17074, FMLA, and from the DFG under project number 392401413, DEEP.

6. References

[1] C. O. Tiemann, M. B. Porter, and L. N. Frazer, "Localization of marine mammals near hawaii using an acoustic propagation model," *The Journal of the Acoustical society of America*, vol. 115, no. 6, pp. 2834–2843, 2004.

[2] X. C. Halkias, S. Paris, and H. Glotin, "Classification of mysticete sounds using machine learning techniques," *The Journal of the Acoustical Society of America*, vol. 134, no. 5, pp. 3496–3505, 2013.

[3] L. Shamir, C. Yerby, R. Simpson, A. M. von Benda-Beckmann, P. Tyack, F. Samarra, P. Miller, and J. Wallin, "Classification of large acoustic datasets using machine learning and crowdsourcing: Application to whale calls," *The Journal of the Acoustical Society of America*, vol. 135, no. 2, pp. 953–962, 2014.

[4] K. M. Stafford, C. G. Fox, and D. S. Clark, "Long-range acoustic detection and localization of blue whale calls in the northeast pacific ocean," *The Journal of the Acoustical Society of America*, vol. 104, no. 6, pp. 3616–3625, 1998.

[5] M. H. Laurinolli, A. E. Hay, F. Desharnais, and C. T. Taggart, "Localization of north atlantic right whale sounds in the bay of fundy using a sonobuoy array," *Marine Mammal Science*, vol. 19, no. 4, pp. 708–723.

[6] C. W. Clark, "Acoustic communication and behavior of the southern right whale (eubalaena australis)," *Communication and behavior of whales*, pp. 163–198, 1983.

[7] M. Wahlberg, "The acoustic behaviour of diving sperm whales observed with a hydrophone array," *Journal of Experimental Marine Biology and Ecology*, vol. 281, no. 1-2, pp. 53–62, 2002.

[8] X. Mouy, M. Bahoura, and Y. Simard, "Automatic recognition of fin and blue whale calls for real-time monitoring in the st. lawrence," *The Journal of the Acoustical Society of America*, vol. 126, no. 6, pp. 2918–2928, 2009.

[9] D. K. Mellinger and C. W. Clark, "Recognizing transient low-frequency whale sounds by spectrogram correlation," *The Journal of the Acoustical Society of America*, vol. 107, no. 6, pp. 3518–3529, 2000.

[10] D. K. Mellinger, "A comparison of methods for detecting right whale calls," *Canadian Acoustics*, vol. 32, no. 2, pp. 55–65, 2004.

[11] J. R. Potter, D. K. Mellinger, and C. W. Clark, "Marine mammal call discrimination using artificial neural networks," *The Journal of the Acoustical society of America*, vol. 96, no. 3, pp. 1255–1262, 1994.

[12] I. Tolkova, L. Bauer, A. Wilby, R. Kastner, and K. Seger, "Automatic classification of humpback whale social calls," *The Journal of the Acoustical Society of America*, vol. 141, no. 5, pp. 3605–3605, 2017.

[13] E. Smirnov, "North atlantic right whale call detection with convolutional neural networks," in *Proc. Int. Conf. on Machine Learning, Atlanta, USA*. Citeseer, 2013, pp. 78–79.

[14] A. K. Ibrahim, H. Zhuang, N. Erdol, and A. M. Ali, "Detection of north atlantic right whales with a hybrid system of cnn and dictionary learning," in *Proc. 2018 European Signal Processing Conference (EUSIPCO)*, 2018.

[15] D. Wang, L. Zhang, Z. Lu, and K. Xu, "Large-scale whale call classification using deep convolutional neural network architectures," in *Proc. 2018 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC)*, Sep. 2018, pp. 1–5.

[16] L. Zhang, D. Wang, C. Bao, Y. Wang, and K. Xu, "Large-scale whale-call classification by transfer learning on multi-scale waveforms and time-frequency features," *Applied Sciences*, vol. 9, no. 5, p. 1020, 2019.

[17] D. Gunning, "Explainable artificial intelligence (xai)," *Defense Advanced Research Projects Agency (DARPA), nd Web*, 2017.

[18] S. Lapuschkin, S. Wäldchen, A. Binder, G. Montavon, W. Samek, and K. Müller, "Unmasking Clever Hans predictors and assessing what machines really learn," *Nature Communications*, vol. 10, no. 1, p. 1096, Mar. 2019.

[19] S. Becker, M. Ackermann, S. Lapuschkin, K. Müller, and W. Samek, "Interpreting and explaining deep neural networks for classification of audio signals," *CoRR*, vol. abs/1807.03418, 2018.

[20] B. W. Schuller, A. Batliner, C. Bergler, F. B. Pokorny, J. Krajewski, M. Cychosz, R. Vollmann, S.-D. Roelen, S. Schnieder, E. Bergelson10 *et al.*, "The interspeech 2019 computational paralinguistics challenge: Styrian dialects, continuous sleepiness, baby sounds & orca activity," in *Proc. INTERSPEECH 2019*, Gratz, Austria, 2019.

[21] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. Weiss, and K. Wilson, "Cnn architectures for large-scale audio classification," in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[22] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. 3rd International Conference on Learning Representations, ICLR 2015*, San Diego, CA, USA, 2015.

[23] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *Proc. 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 776–780.

[24] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2818–2826.

[25] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proc. Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[26] G. Montavon, S. Lapuschkin, A. Binder, W. Samek, and K. Müller, "Explaining nonlinear classification decisions with deep taylor decomposition," *Pattern Recognition*, vol. 65, pp. 211–222, 2017.

[27] M. Alber, S. Lapuschkin, P. Seegerer, M. Hägele, K. T. Schütt, G. Montavon, W. Samek, K.-R. Müller, S. Dähne, and P.-J. Kindermans, "iNNvestigate neural networks!" *arXiv preprint arXiv:1808.04260*, 2018.