
Classifying Marine Mammal Sounds using Convolutional Neural Networks

Arkar Min Aung¹ Andrés Francisco Guerrero¹ Michael Sokolovsky¹

Abstract

This paper explores convolutional neural networks as models for learning to identify dolphin species from sound recordings with the intent of discovering a previously unknown relationship between these variables. Training was performed on data from the Watkin Marine Mammal Sound database on four dolphin species, and accuracy was compared to the plurality classifier. Results did not show effective learning of dolphin species higher than the plurality classifier. Experiments with different training data configurations, training on different animals, and the observation of particular patterns in training accuracy over time suggest that unique intra-class discriminative noise made training on the desired sounds difficult. Model performance suggests that learning occurred on recording environment rather than true signal.

1. Introduction

An exciting aspect of modern machine learning techniques is their ability to uncover patterns in data that are both difficult for humans to notice or were previously unknown. In this paper, we attempt to use modern machine learning techniques to discover if dolphins species can be categorized by their vocalizations: a task that to our knowledge has not been attempted by humans or computer algorithms. We explore convolutional neural networks architectures as models for processing dolphin sound clips and categorizing them into four different species.

1.1. Research contributions

We hoped to build a model that demonstrates that dolphin vocalizations encode species information. Instead, we demonstrate that neural networks are capable of learning to categorize different marine mammals across audio recording environments and technical setups. We also show that

the models are prone to classifying signals by their recording environments and other noise artifacts rather than the desired sounds.

2. Related Work

The application of convolutional neural networks to temporal data is not new in machine learning research. Prior work (Pic15) has been conducted on using convolutional neural networks for classifying sounds from their two-dimensional spectrograms, whereas one-dimensional convolutions have been used for mapping raw EEG signal data into sleep stages (TMGZ16).

There has been machine learning research on categorizing different types of marine mammal sounds within one species, such as categorizing bottlenose dolphin whistles (EZE14) and whale calls (Esf14; MWD⁺06; Mel04). These research problems have dealt with classification tasks easily done by human experts, their primary goal being to perform a task passively and efficiently rather than attempting to uncover a previously unknown relationship. Though the value of positive results in our paper is similar, making a categorization problem easier and more efficient, this new research additionally seeks to add knowledge to the field of marine biology by identifying a previously unknown method for species categorization.

3. Proposed Method

The goal of this research inquiry was to build a model that correctly classifies four dolphin species using sound data. The model architectures chosen for categorizing sound data were 1D and 2D convolutional neural networks. This choice was based on the success of prior research as discussed above and on the premise that sliding filters across 1D signal data and 2D spectrograms would capture temporal patterns. The general success of convolutional neural networks in correctly classifying data across visually similar classes in the animal and plant kingdom (Wat17), made convolutional neural networks models an enticing choice.

Before designing architectures, limitations in the dataset were considered and are discussed in section 3.1. Because of the fixed input size of convolutional neural networks and the varying length and sample frequencies of the

¹Worcester Polytechnic Institute. Correspondence to: Arkar Min Aung <aaung@wpi.edu>.

available data, formatting data was important to the process of designing architectures and is discussed in section 3.2.

3.1. Data

Raw sound data was obtained from the The Watkins Marine Mammal Database (WMMD) (Wat17) from which data from four dolphin species were selected for the classification task. The four species selected were pantropical spotted, striped, white-beaked, and white-sided dolphin.

For each dolphin class, WMMD contains sounds clips from different single-species field studies conducted in different years. To the human ear, these recordings sound remarkably different, varying in dolphin volume, background noise volume, and background noise type. It was theorized that models could use these artifacts in the data rather than dolphin sounds to correctly classify sounds clips. To avoid this potential problem, data was sorted into 80%/10%/10% training/validation/testing splits so that there were no overlapping studies across the training and testing data. As an illustration, for the white-beaked dolphin class, training data only came from two fields studies made in 1972 and 1959 while testing and validation data came from one field study made in 1970.

WMMD additionally organizes its species data into two categories: All cuts (A) and Best of (B). Dataset B is a superior-quality subset of dataset A. Both sets of data were used in training models under the theory that the larger size of dataset A may offset its additional noise.

3.2. Preprocessing

Two types of model architectures were built for classifying the data: models that took raw, 1D signals as inputs and models that took 2D spectrograms as inputs. The former architectures are not discussed in this paper as they did not lead to results better than the 2D-input architectures. They are briefly mention here as one alternative model explored. 2D-input architectures made up the majority of the explored model space.

For models that accepted 1D inputs, wave file data was re-sampled to 22050 Hz and split into 1-second slices. Therefore, inputs to the model were vectors of length 22050.

For models that accepted 2D inputs, wave file data was processed as in (Pic15): re-sampled to 22050 Hz, split into 1-second slices, and converted into two 2D mel spectrogram channels of shape (60, 41). The first channel contains a log-normalized mel spectrogram of shape (60, 41) with 60 corresponding the frequencies in the clip and 41 corresponding the to the temporal dimension. The second channel contains an array of shape (60, 41) corresponding to the estimate of the derivative of the first channel with respect to the frequency dimension. Therefore, inputs to the model

were tensors of shape (60, 41, 2).

Datasets A and B were compiled and models were trained on both sets to see if there were differences in performance. Dataset A was preprocessed and divided into 8543 samples for training, 830 samples for validation, and 830 samples for testing. Dataset B was preprocessed and divided into 744 samples for training, 65 samples for validation, and 65 samples for testing.

4. Experiments

4.1. Model Architecture

Because the relatively small size of datasets A and B and limited time and computational resources, simpler models were considered with two to four convolutional layers.

For simplicity, strides were set to size one. Kernels of different size were explored during experimentation, including traditional square filters to elongated filters spanning the height or length of the spectrogram inputs that collapse the input spectrogram into a 1D array, similar to a technique used in prior class extraction research (Pic15). These filter choices led to models tested that featured either only 2D convolutional layers or a combination of 2D convolutions followed by 1D convolutions.

All architectures used ReLU activations, batch normalization layers before nonlinearities, and random normal initializations with variances scaled by the number of outputs in prior layers as in (HZRS15). Final layers had softmax activations, and the loss function optimized during training was categorical cross entropy. Different orders of magnitude of L2 weight regularization were explored during training as well as two different optimizers (stochastic gradient descend and Adam), different learning rates with and without decay, and different values for momentum. Tensorflow and Keras were used for building and training models.

Two of the final model architectures tested are included depicted in Figure 1. Model 1 converts 2D input data to a 1D array via a convolutional layer with filters widths that span the height of the input. Outputs from this layer are input to a 1D convolutional network and finally to a pooling layer that returns the averages across each channel as in (HZRS16). The final layer has a fully-connected, four-neuron output with softmax activation, representing the probability distribution of dolphin species. Model 2 features a deeper convolutional architecture with small rectangular kernels and more fully-connected final layers.

Model 1					
Layer	Input Shape	Kernel	Stride	Channels	Output Shape
Data	(None, 60, 41, 2)				(None, 60, 41, 2)
Conv2D	(None, 60, 41, 2)	(3, 41)	(1, 41)	128	(None, 60, 1, 128)
Conv1D	(None, 60, 128)	3	1	256	(None, 60, 256)
AveragePooling	(None, 60, 256)	60	60		(None, 256)
Fully Connected	(None, 256)				(None, 4)

Model 2					
Layer	Input Shape	Kernel	Stride	Channels	Output Shape
Data	(None, 60, 41, 2)				(None, 60, 41, 2)
Conv2D	(None, 60, 41, 2)	(30, 30)	(1, 1)	8	(None, 60, 41, 8)
Conv2D	(None, 60, 41, 8)	(30, 30)	(1, 1)	16	(None, 60, 41, 16)
MaxPooling	(None, 60, 41, 16)	(2, 2)	(2, 2)		(None, 29, 20, 16)
Conv2D	(None, 29, 20, 16)	(30, 30)	(1, 1)	32	(None, 29, 20, 32)
Conv2D	(None, 29, 20, 32)	(30, 30)	(1, 1)	64	(None, 29, 20, 64)
MaxPooling	(None, 29, 20, 64)	(2, 2)	(2, 2)		(None, 8320)
Fully Connected	(None, 8320)				(None, 512)
Fully Connected	(None, 512)				(None, 256)
Fully Connected	(None, 256)				(None, 4)

Figure 1. Architecture of Two Models.

5. Results

5.1. Initial Results and Discussion

The results of all models, trained on both datasets A and B, were not promising. Training accuracy did converge to over 70% in all cases, however, validation accuracy never rose past the plurality classifier rate, which was approximately 40% for both datasets A and B. When examining the confusion matrices on validation data at different stages of training, the models typically classified all data to one class. As training progressed, models either got stuck in a local minimum of a naive model that classifies all data as one class, or skipped around between the four variants of this naive classifier.

It was theorized that noise in the data, particularly the distinct recording environments within the classes in the training data, made learning to categorize the data clips by dolphin sounds difficult. Instead, the models may have been learning to identify species within the training set correctly based on the background artifacts and simply ignoring the sound of dolphins. These unique background noises could in a sense be easier for a model to make classifications on than the nuanced differences between dolphin species. There is also the possibility that there simply is not an easily mappable function from dolphin noise to dolphin species.

To explore some of these questions, models were retrained using randomized splitting on dataset B, which allowed for data clips from the same recordings to appear in training, validation, and testing splits. High accuracy was achieved (>70% and in some cases >90%) with this “cheating” form of data splitting using the same model architectures as before. These results confirm that the models were effective at learning to distinguish from which study a signal came from, but were not learning how to identify dolphin species.

5.2. New Experiment Formulation

The original research question of whether dolphin species can be identified by sound was thwarted by the experimental problem of trying to classify dolphins across distinctly different recording environments, even within the same class. These limitations raised the question of whether even different-sounding animals could be identified.

To explore this, original models were trained using the conservative splitting procedure as described before, but on two animals data with more varied vocalization patterns: walrus, a baritone grunter, and the white-sided dolphin, a mezzo-soprano whistler. The hope was that the more pronounced differences in vocalization would be large enough that training models could overcome the noise. Observing high accuracy with these models would suggest that a network can discern animal species across different recording environments and ignore other noise in the data.

5.3. New Results

For illustration, training curves, confusion matrix, and ROC plot are provided for a trained model with the same architecture described earlier (Model 1), with a modified two-class instead of four-class softmax output layer. Setup and performance are reported below and are reflective of the trends observed in training other models.

The two classes of data were taken from dataset B and divided into 531 samples for training, 70 samples for validation, and 71 samples for testing. The model was trained using stochastic gradient descent with a learning rate of 1e-3 and momentum of 0.9 on 200 epochs. A naive plurality classifier would have categorized 59% of the data correctly on the test data. The trained neural network model categorized test data with 86% accuracy. Below are the confusion matrix, ROC plot, and two sample training curves for the model.

6. Discussion

The models trained on data from two disparate species were able to successfully classify sound clips with an accuracy significantly higher than chance. These results demonstrate that our models are capable of generalizing across different recording environments.

What was particularly striking about the training process of these models was how the validation accuracy evolved over successive epochs. Most models reached training accuracies above 95% after few epochs, while at the same time, validation accuracy often flat-lined to random chance, bound within a range near the accuracy rate of the plurality classifier (59%). However, allowing the models to continue training longer after overfitting eventually led to validation

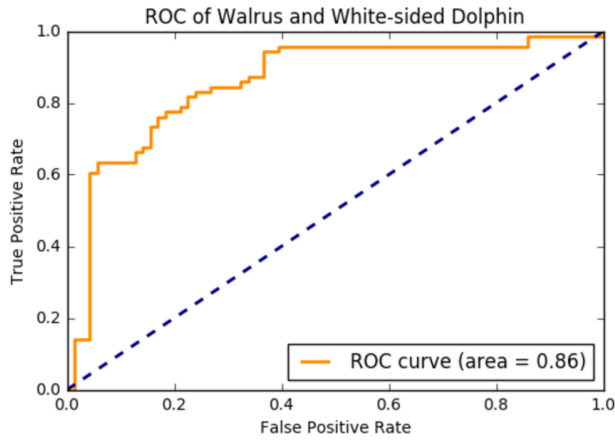


Figure 2. ROC Curve from Classifying Walruses and Dolphins.

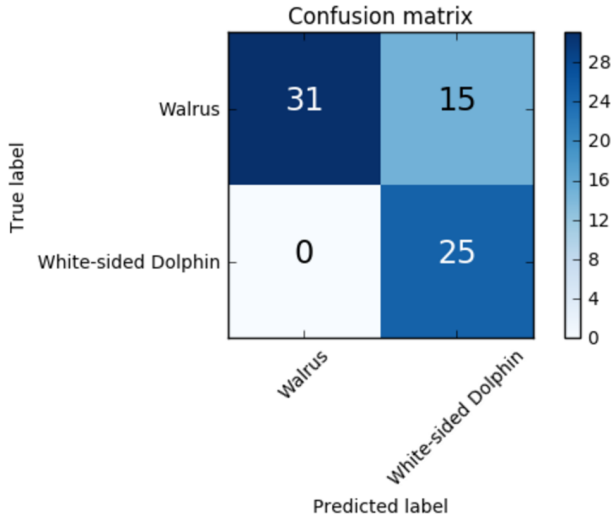


Figure 3. Confusion Matrix from Classifying Walruses and Dolphins.

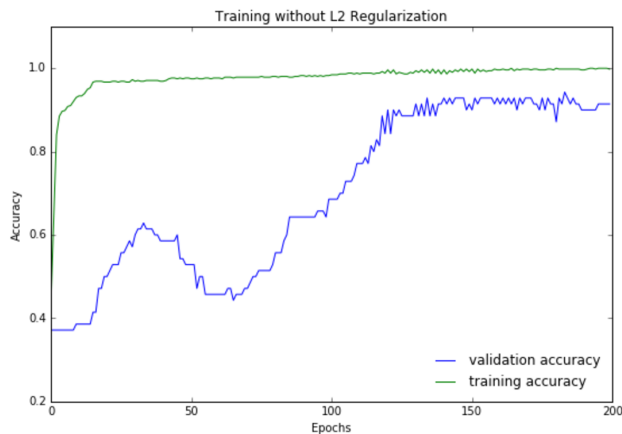


Figure 4. Training Accuracy.

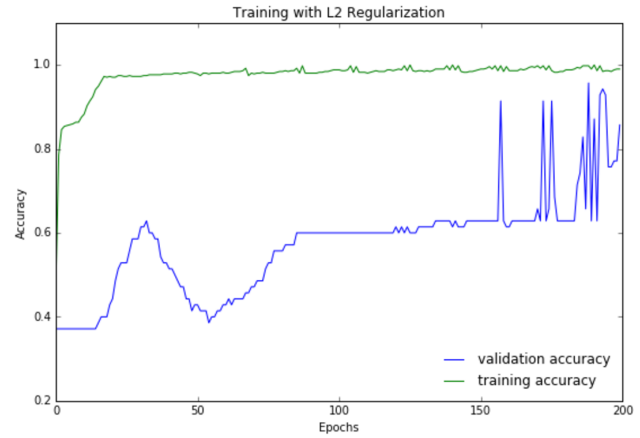


Figure 5. Training Accuracy with L2 Normalization.

accuracy increasing.

We theorize that these training curves reflect the noise problem discussed throughout the paper. We intuit that the models initially train on the discriminative noise in the early epochs, and that once networks learn to classify as much as they can from the sound artifacts, in order to squeeze out a few more percentage points of accuracy, they begin to train on and recognize the true differences between the animal sounds. In response, validation accuracy begins to increase in later epochs. This pattern can be readily seen in the training experiment curves with L2 regularization. As training accuracy crests upward marginally when it has already reached more than 90% accuracy, validation accuracy suddenly spikes up from the plurality classification rate to rates in the 80s. In short, because of the discriminative noise in our dataset, models first fit noise before fitting true signal, which is the opposite of the common observation that training models first fit true signal before fitting noise.

The appendix contains visualizations of the input data for both classes and channels as well the outputs from filters. Visualizations come from training Model 2 as described above with the final layer changed to a two-class softmax output. Of note are the pronounced differences in the data within each class, arising from data taken from different recording environments. Also worth mentioning is that the filters appear non-sensical or noisy patterns, which is further evidence that networks may be primarily training on discriminating noise rather than true signal.

7. Conclusions and Future Work

Our results leave open the possibility that cleaner recordings made with the same equipment could still be used to identify dolphin species. Creative architectures and hybrid loss functions could also be formulated for building models

that actively avoid making classifications from noise and give preference to the true signals.

The behavior of the training curves in later experiments poses additional questions on proper techniques on dealing with over-fitting in situations where there is irremovable but discriminative noise in datasets. More research could be conducted on the value of over-training in an attempt to uncover truly discriminative features masked by discriminative noise.

stage scoring with single-channel eeg using convolutional neural networks. *arXiv preprint arXiv:1610.01683*, 2016.

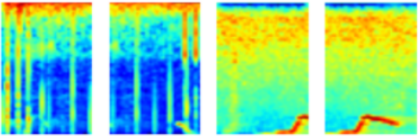
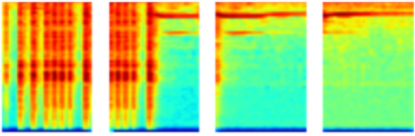
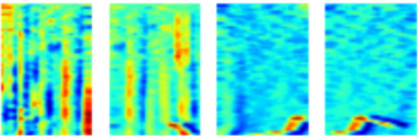
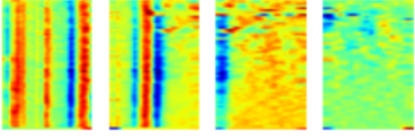
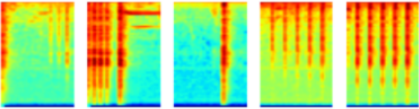
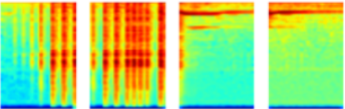
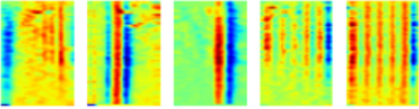
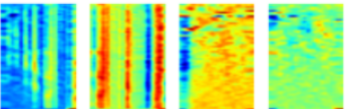
[Wat17] William Watkins. Watkins marine mammal sound database, 2017.

References

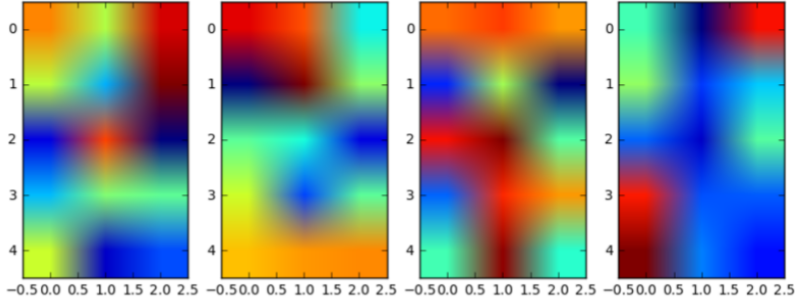
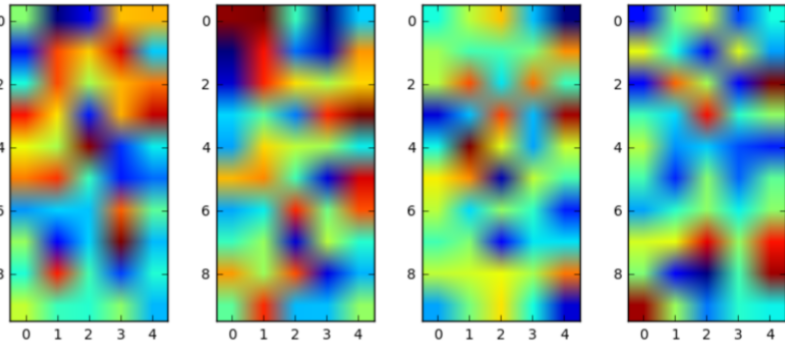
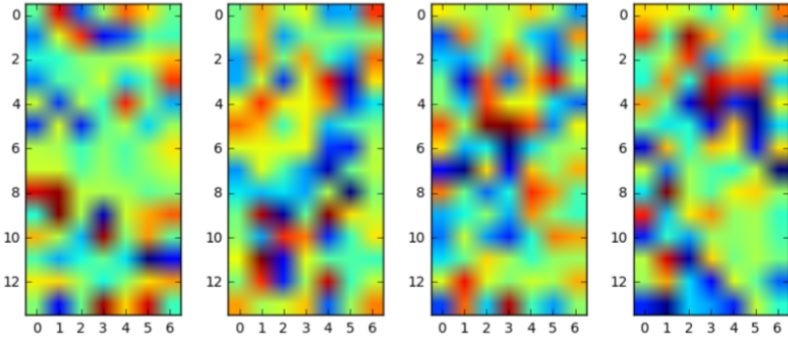
- [Esf14] Mahdi Esfahanian. *Detection and Classification of Marine Mammal Sounds*. PhD thesis, Florida Atlantic University Boca Raton, Florida, 2014.
- [EZE14] Mahdi Esfahanian, Hanqi Zhuang, and Nurgun Erdol. A new approach for classification of dolphin whistles. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 6038–6042. IEEE, 2014.
- [HZRS15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.
- [HZRS16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.
- [Mel04] David K Mellinger. A comparison of methods for detecting right whale calls. *Canadian Acoustics*, 32(2):55–65, 2004.
- [MWD⁺06] RP Morrissey, J Ward, N DiMarzio, S Jarvis, and DJ Moretti. Passive acoustic detection and localization of sperm whales (*physeter macrocephalus*) in the tongue of the ocean. *Applied acoustics*, 67(11):1091–1105, 2006.
- [Pic15] Karol J Piczak. Environmental sound classification with convolutional neural networks. In *Machine Learning for Signal Processing (MLSP), 2015 IEEE 25th International Workshop on*, pages 1–6. IEEE, 2015.
- [TMGZ16] Orestis Tsinalis, Paul M Matthews, Yike Guo, and Stefanos Zafeiriou. Automatic sleep

8. Conclusions and Future Work

8.1. Data Visualizations

	Dolphin	Walrus
Training Data Set		
Channel 1		
Channel 2		
Validation Data Set		
Channel 1		
Channel 2		

8.2. Filter visualizations - Model 2

	Visualization of Filters			
5x3 Convolution 1 kernel				
10x5 Convolution 2 kernel				
14x7 Convolution 3 kernel				
18x9 Convolution 4 kernel	