LEVERAGING UNLABELLED DATA THROUGH SEMI-SUPERVISED LEARNING TO IMPROVE THE PERFORMANCE OF A MARINE MAMMAL CLASSIFICATION SYSTEM

Anonymous authors
Paper under double-blind review

ABSTRACT

A considerable proportion of the passive acoustic data sets collected for marine mammal conservation purposes remain unanalyzed by human experts. In some cases, the aforementioned proportion amounts to as much as 97% of the entire data set. As a result, research and development into automated classification systems rely on sparsely annotated data sets. In this work we adapt a semi-supervised deep learning approach to develop a classification system of marine mammal vocalizations such that both the annotated and non-annotated portions of an acoustic data set can be used during training.

1 INTRODUCTION

Passive acoustic monitoring (PAM) is a practical approach to measuring marine mammal species presence and abundance and is largely used for conservation purposes. Historically, PAM data is collected continuously over several months via moored recording devices (Mellinger et al. (2007); Van Parijs et al. (2009)). When the recording devices are retrieved, often as little as three percent of the entirety of the data is manually annotated (Kowarski & Moors-Murphy (2020)). Machine learning has been a useful tool in developing automated detection and classification systems (DCS) of marine mammal vocalizations for many years in order to assist with acoustic analysis, however, more recently as researchers have started to develop deep learning-based DCS, a requirement for large labeled data sets has been observed (Thomas et al. (2019); Kirsebom et al. (2020); Shiu et al. (2020)). As it relates to PAM, deep learning faces two data-related challenges. First, the percentage of PAM data that is manually annotated is inordinately scarce. Second, depending on the sampling strategy used to determine which acoustic recordings should be analyzed, annotations may lack satisfactory variation. For example, the acoustic sources during the first month of a several month deployment may vary dramatically to that of the following month depending on weather conditions, the presence of different species, and additional sources of noise.

In this work, we adapt the semi-supervised learning algorithm MixMatch (Berthelot et al. (2019)) in order leverage unlabeled data and improve the performance of a convolutional neural network (CNN) used to classify spectrograms containing minke whale vocalizations against instances containing ambient noise, non-biological noise, or the vocalizations of other marine mammals. In particular, we demonstrate that through semi-supervised learning, CNNs remain an appropriate solution to spectrogram classification even in cases where only a relatively small number of labeled examples is available.

2 DATA SET AND METHODS

2.1 ACOUSTIC DATA SETS

There are two data sets used in this paper. The first, "Set A", consists of acoustic recordings spanning roughly three months from late August to November of 2015. During this time, three Autonomous Multichannel Acoustic Recorders (AMARs) were deployed in the Bay of Fundy in order to measure

the sound produced by vessels as well as detect marine mammal vocalizations. The second data set, "Set B", consists of a selection of acoustic recordings taken from a large scale deployment from November 2017 to June 2018, along the Atlantic Outer Continental Shelf (OCS).

The two data sets serve distinct purposes. Set A is used for model training/validation. Instances pertaining to possible false alarms, were taken from 227 *fully-annotated* WAV files. The minke whale annotations were taken from 160 of these same files, plus an additional 512 *partially-annotated* files. Finally, roughly 10 percent of the remaining non-annotated files making up the remainder of the deployment were processed to be used for semi-supervised learning. Set B represents a proxy for out-of-distribution (OOD) examples and therefore is only used during model testing. Table 1 contains the annotation distribution of Sets A and B separated by acoustic sources. Figure 1 depicts the distribution of the training data with respect to time, separated by annotation level. As you can see, the training data is sparse in terms of the number of minke whale annotations and highly unbalanced in favour of the sources of possible false alarm. From a supervised learning perspective, the total number of minke whale annotations available for training is smaller than many benchmark data sets commonly used in image recognition (Sun et al. (2017)).

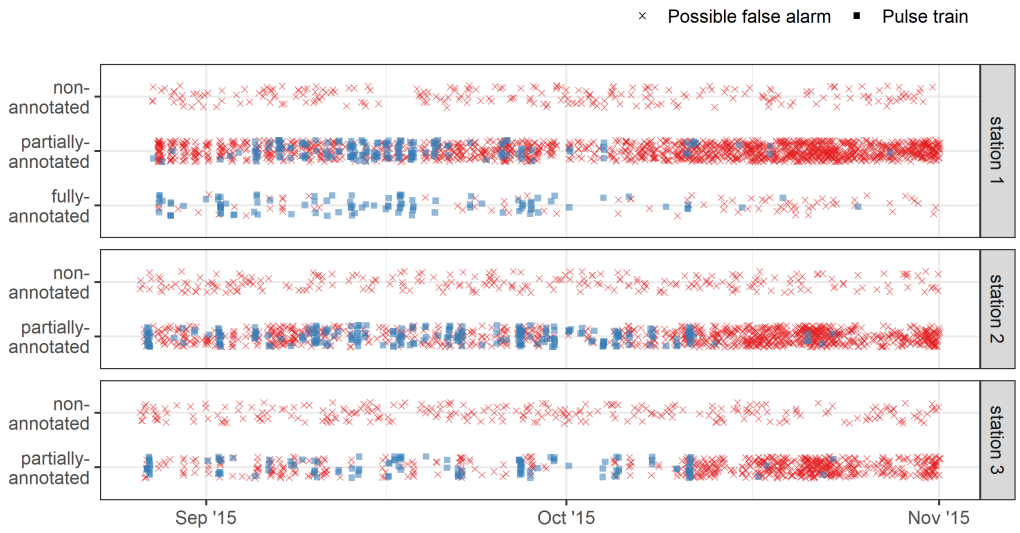

Figure 1: Distribution of the training set over time, factored by the deployment location (station) and the file's annotation level (fully, partially, or non-annotated). Files that contain at least one minke whale annotation are plotted as blue squares. Files that either explicitly do not (fully-annotated files) or possibly contain an pulse train (partially/non-annotated) are plotted using red X's. A slight *jitter* was added in order to distinguish files from the same date.

Table 1: Annotation distribution for Sets A and B separated at the acoustic source level.

		Set A		Set B
Acoustic source	shorthand	training	validation	testing
Minke whale pulse train	MW	556	56	336
Ambient noise	AB	5560	620	-
Fin whale	FW	3383	422	-
Humpback whale	HB	5773	597	-
North Atlantic right whale	RW	462	49	-
Non-biological noise	NN	-	-	266
Sei whale	SW	-	-	62

2.2 MINKE WHALE PULSE TRAINS AND SPECTROGRAM GENERATION

Minke whales are currently not listed as endangered species under the Marine Mammal Protection Act, however, researchers believe that the species is still being threatened by various sources of anthropogenic activity, including: climate change, entanglement in fishing gear, ship strikes, and increased underwater noise (Risch et al. (2019)). In this work, we train a CNN capable of classifying spectrograms containing a vocalization distinct to minke whales known as a "pulse train". Minke pulse trains typically occupy the 200–400 Hz band and have a duration of roughly 45 to 60 seconds (Mellinger et al. (2000); Risch et al. (2013)). An example of a minke pulse train can be seen in Figure 2.

Each time series corresponding to the acoustic recordings (i.e., WAV files) were split into 45-second segments, overlapped by three seconds, and passed through a Short-time Fourier Transform (STFT) with a window size equal to 2048 frames overlapped by 512 frames using a Hann windowing function. The magnitude of the STFT was then scaled to decibels (dB) as is common in underwater acoustics and truncated using an upper frequency bound of 1000Hz. Finally, the spectrograms were normalized between [0, 1]. The resulting scaled, truncated, and normalized spectrogram is equivalent to a single training instance.

2.3 SEMI-SUPERVISED LEARNING

The semi-supervised learning algorithm used in this paper, MixMatch, was presented as a state-of-the-art approach to handling label scarcity across a variety of image recognition tasks (Berthelot et al. (2019)). MixMatch is grounded upon the use of two types of data-augmentation. First, MixMatch makes use of the similarly named *mixup* data-augmentation strategy such that the CNN learns a *vicinal* distribution (Chapelle et al. (2001)) rather than an *empirical* distribution of the data. The second data-augmentation strategy used by MixMatch corresponds to common data-augmentation strategies often used in image recognition (e.g., cropping, rotation, etc.). Such image transformations can not be used on spectrograms without potentially impacting the acoustic representation. Therefore we rely on a set of augmentations known as SpecAugment (Park et al. (2019)), as shown in Figure 2.

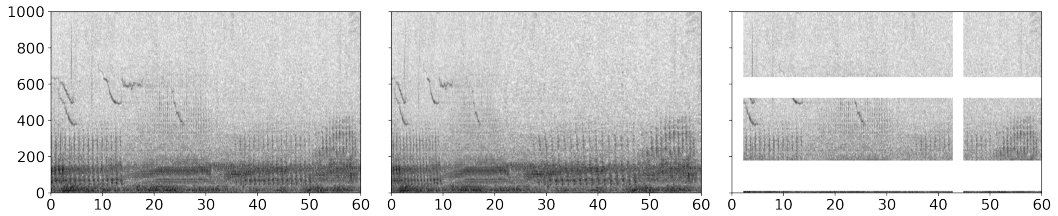

Figure 2: Example training instance before and after applying SpecAugment. Column one (starting from the left) contains the original instance of a minke pulse train. Column two contains the same instance after being "time-warped" using SpecAugment. Column three contains examples of time and frequency masking.

3 EXPERIMENTAL RESULTS

3.1 MODEL TRAINING

We train a baseline ResNet-18 model via fully supervised learning and contrast the performance against several models trained using MixMatch and various hyperparameters. All models were implemented in PyTorch (Paszke et al. (2017)). Model training was distributed over two NVIDIA V100 GPUs until early stopping was deemed necessary. The model trained via supervised learning used mini-batches of size 64, while the semi-supervised model was trained using mini-batches of size 48 (16 labeled examples plus 2×16 unlabeled examples). Due to the unbalanced class distribution, a balanced sampler was used during training. The initial learning rate of all models was set to 0.001 and decayed by a factor of ten after the training loss plateaued. Per the suggestions of the authors in Berthelot et al. (2019), the semi-supervised loss weighting parameter λ_U was increased linearly

over the first five epochs. The highest performing models in terms of F-1 score on the validation set was maintained after each epoch.

3.2 BASELINE VS. SEMI-SUPERVISED RESULTS

In general, we found that setting the weighting parameter (λ_U) to a value between 10 and 50 lead to well performing models in terms of F-1 score. No significant change in performance was observed for the hyperparameter used by *mixup* (α). The remainder of this section presents the performance of semi-supervised models setting $\lambda_U = 10$ and $\alpha = 0.5$. Based on our experiments, the SpecAugment hyperparameters were more important compared to the those used for MixMatch, with the exception of time warping, which did not impact performance. In general we found that a balanced number of time and frequency masks performs favourably. The best performing models were observed using two time and frequency masks and setting the maximum possible widths of each mask equal to roughly 10 percent of the dimensions of the spectrogram.

Table 2 compares the performance of the model trained using supervised learning to that trained using MixMatch. Performance was evaluated on the validation data of Set A and is presented in terms of precision, recall, and F-1 score. The depicted values represent the median training run in terms of F-1 score after training each model five times using different random number generator seeds. The performance of the model trained using semi-supervised learning outperforms that trained strictly using labeled data. This is a substantial finding as it implies the features learned by the CNN were positively influenced by unlabeled data.

Table 2: Performance comparison of the baseline and semi-supervised CNNs in terms of precision, recall, and F-1 score, measured on the validation data of Set A: Bay of Fundy.

Training paradigm	precision	recall	F-1 score
Supervised	0.79700	0.85807	0.82641
Semi-supervised ($\lambda_U = 10$, $\alpha = 0.5$)	**0.81645**	**0.90301**	**0.85755**

3.3 OUT-OF-DISTRIBUTION RESULTS

The two models from Section 3.2 were used to classify the data from Set B and the results are presented in Table 3. As we can see, the models trained using MixMatch outperform the models trained using only labeled data. Importantly, the observed increase in performance is even larger for Set B, clearly demonstrating that the baseline models are more susceptible to training bias. Moreover, the performance of the semi-supervised models on Set B actually exceed that of Set A, demonstrating that the features learned by the semi-supervised CNNs can generalize to acoustic data collected in distinct locations, at varying times and depths, and are less susceptible to unknown acoustic sources.

Table 3: Performance comparison of the baseline and semi-supervised CNN architectures in terms of precision, recall, and F-1 score, measured using the data from Set B: Atlantic OCS.

Training paradigm	precision	recall	F-1 score
Supervised	0.75213	0.75178	0.75195
Semi-supervised ($\lambda_U = 10$, $\alpha = 0.5$)	**0.89658**	**0.88799**	**0.89226**

4 CONCLUSION

This paper presents the clear benefits of using semi-supervised learning for the development of deep learning based classification systems for PAM. In particular, we demonstrate that by including unlabeled instances to the training routine of a CNN used to classify spectrograms possibly containing the vocalizations of minke whales, we are capable of learning features that generalize well to unseen data and OOD examples. The results of this work are substantial and the application of semi-supervised learning is novel to this domain.

REFERENCES

David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. In *Advances in Neural Information Processing Systems*, pp. 5049–5059, 2019.

Olivier Chapelle, Jason Weston, Léon Bottou, and Vladimir Vapnik. Vicinal risk minimization. In *Advances in neural information processing systems*, pp. 416–422, 2001.

Oliver S Kirsebom, Fabio Frazao, Yvan Simard, Nathalie Roy, Stan Matwin, and Samuel Giard. Performance of a deep neural network at detecting north atlantic right whale upcalls. *The Journal of the Acoustical Society of America*, 147(4):2636–2646, 2020.

Katie A Kowarski and Hilary Moors-Murphy. A review of big data analysis methods for baleen whale passive acoustic monitoring. *Marine Mammal Science*, 2020.

David K Mellinger, Carol D Carson, and Christopher W Clark. Characteristics of minke whale (balaenoptera acutorostrata) pulse trains recorded near puerto rico. *Marine Mammal Science*, 16 (4):739–756, 2000.

David K Mellinger, Kathleen M Stafford, Sue E Moore, Robert P Dziak, and Haru Matsumoto. An overview of fixed passive acoustic observation methods for cetaceans. *Oceanography*, 20(4): 36–45, 2007.

Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le. Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv preprint arXiv:1904.08779*, 2019.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

Denise Risch, Christopher W Clark, Peter J Dugan, Marian Popescu, Ursula Siebert, and Sofie M Van Parijs. Minke whale acoustic behavior and multi-year seasonal and diel vocalization patterns in massachusetts bay, usa. *Marine Ecology Progress Series*, 489:279–295, 2013.

Denise Risch, Thomas Norris, Matthew Curnock, and Ari Friedlaender. Common and antarctic minke whales: Conservation status and future research directions. *Frontiers in Marine Science*, 6:247, 2019.

Yu Shiu, KJ Palmer, Marie A Roch, Erica Fleishman, Xiaobai Liu, Eva-Marie Nosal, Tyler Helble, Danielle Cholewiak, Douglas Gillespie, and Holger Klinck. Deep neural networks for automated detection of marine mammal species. *Scientific Reports*, 10(1):1–12, 2020.

Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE international conference on computer vision*, pp. 843–852, 2017.

Mark Thomas, Bruce Martin, Katie Kowarski, Briand Gaudet, and Stan Matwin. Marine mammal species classification using convolutional neural networks and a novel acoustic representation. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 290–305. Springer, 2019.

Sofie M Van Parijs, Chris W Clark, Renata S Sousa-Lima, Susan E Parks, Shannon Rankin, Denise Risch, and Ilse C Van Opzeeland. Management and research applications of real-time and archival passive acoustic sensors over varying temporal and spatial scales. *Marine Ecology Progress Series*, 395:21–36, 2009.