DOI: 10.61782/fa.2023.1052

forum acusticum 2023

DEEP LEARNING ON SMALL DATASETS TO CLASSIFY MAMMALIAN VOCALIZATIONS

Rodrigo Manriquez[1,2*] **Sonja A. Kotz**[2,3] **Andrea Ravignani**[4,5]
Bart de Boer[1]

[1] Artificial Intelligence Lab, Vrije Universiteit Brussel, Brussel, Belgium
[2] Department of Neuropsychology and Psychopharmacology, Faculty of Psychology and Neuroscience, Maastricht University, Maastricht, The Netherlands
[3] Department of Neuropsychology, Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany
[4]Department of Human Neurosciences, Sapienza University of Rome, Rome, Italy
[5]Center for Music in the Brain, Department of Clinical Medicine, Aarhus University & The Royal Academy of Music Aarhus/Aalborg, Denmark

ABSTRACT

Deep learning algorithms are increasingly used in many fields outside of artificial intelligence, including bioacoustics. Among many possible applications of deep learning to bioacoustics, typical ones include call identification, species recognition, and acoustic features classification. However, the implementation of deep learning algorithms is limited as bioacoustic databases are often rather small and thus lack sufficient data to properly train neural networks. Improper training leads to problems like overfitting and lack of generalization which, in turn, affect performance. Here, we address the most common challenges that bioacousticians face when training a deep neural network in a classification task. We present and explain useful techniques such as pre-training and data augmentation, and emphasize applying them in an efficient and meaningful way to not alter distinctive features or specific stimulus features such as fundamental frequency. We present an example application of these techniques in a classification task, where we perform species identification in a database of phylogenetically distant mammals, each with a limited number of calls. We aimed at developing a general framework on how to apply deep learning algorithms to small- and larger-scale bioacoustic datasets.

Keywords: *artificial intelligence, machine learning, species recognition, species discrimination*

1. INTRODUCTION

Within computational bioacoustics, the use of deep learning (DL) has been of particular interest to researchers, as it has allowed to better deal with a diverse array of problems. A well-known application of DL in bioacoustics is animal audio classification, typically within the same taxonomic class, such as in the BirdCLEF challenge [1]. Other studies explored the classification among individual animals within the same species, sex, strains, or behavioral states [2].

There is a wide array of techniques and methods within DL that can be used in bioacoustics [2]. Although early approaches relied on the use of basic multilayer perceptron architectures, taking acoustic features as input [3], they have been greatly outperformed by Convolutional Neural Networks (CNN). In a CNN approach, raw acoustical data (or lightly processed data) can be taken as input in the shape of time-frequency spectrogram represen-

*__Corresponding author__: rmanriquez@ai.vub.ac.be.

tations [4].

However, DL approaches rely heavily on data availability as a lack of training data makes it difficult to train a deep learning network [5]. On the one hand, machine learning frameworks have been established to deal with small unbalanced datasets, with applications in bioacoustics [6]. On the other hand, methods like data augmentation [7] and transfer learning/pre-training have been developed to increase the performance of algorithms with limited data. Transfer learning (TL) in particular has been used for bioacoustical classification of whale calls [8] among other species.

Although combining both approaches for small datasets has been considered previously [4], it is not a common approach, specially with rather small datasets. The objective of this study was to exploit a typical pre-trained network [9] to perform a bio-acoustical classification of different animal calls in a small dataset.

2. METHODS

2.1 Dataset

In this study, a dataset previously used in cross-taxa emotion recognition studies, was used [10, 11]. The dataset consists of 192 vocalizations from 4 different species: human infant, dog (*Canis familiaris*), chimpanzee (*Pan troglodytes*), and tree shrew (*Tupaia belangeri*, all recorded in natural contexts. The vocalizations were divided into two categories (affiliative and non-affiliative/agonistic), with 24 vocalizations per context category, each containing either a single call or a sequence of 5 to 8 calls. The study was conducted with the approval of the University of Leipzig's ethics committee and in accordance with the Declaration of Helsinki. Specific details on the dataset can be found in the original paper [10].

For the purpose of the present study, the agonistic context category was simply labeled as the "negative" condition, whereas the affiliative context category was labeled as "positive" condition.

2.2 Deep Neural Network Architecture

In this study, the VGG16 convolutional neural network proposed by Sumonyan and Zisserman was applied [9]. The VGG16 network is composed of 16 layers (13 convolutional layers, 2 dense layers and 1 softmax layer), with maxpooling layers in between to avoid generalization (e.g., can the model process new data and make correct predictions after getting trained). It can take fixed size 224x244 RGB images as input, returning a class prediction as output. This architecture gained popularity after its performance in the ImageNet Challenge 2014 [12], but given its good performance on datasets where labeled data is limited, it has also been used successfully in bioacoustics tasks [8].

One of the advantages of using the VGG16 architecture is that the pre-trained weights from the original ImageNet classification task can be used in transfer learning (TL) approaches. TL is a technique that uses a previously trained network and re-trains it in a new task, using the pre-trained weights for initializations.

TL can be used if the task for which it was pre-trained is similar to the new one. This means that features, which are already learned by the network, can be exploited in the new task. Here, spectrograms were treated as images, which the network was already trained on to classify. The main advantage of TL is that it greatly reduces the amount of data needed to achieve good classification accuracy.

2.3 Data Processing

In this study, each playback clip was converted into a spectrogram representation using Parselmouth [13], a Python library that allows using algorithms and methods implemented in Praat [14]. Praat is a speech analysis software package widely used in phonetics and other linguistics disciplines. In the current study we chose Praat as standardized algorithms are already embedded. Spectrograms were computed using standard settings (window length = 20 ms, 20 frames per window, 448 bins, Hann window), and intensities in the spectrogram were represented in decibels by a greyscale. Details on the audio pre-processing of audio clips can be found in the original paper [10].

A data augmentation approach was used, where the amount of data used to train the model was increased based on available training data. Each audio segment was manipulated using the Parselmouth library [13] that allows using PRAAT to increase and decrease the pitch by one octave. Also, noise was added to create a noisy version of the original signal, with approximately 10% of the total energy. Adding noise is a useful way to prevent generalization. This yielded a total of 768 audio samples for each species, from which 536 (70%) were used for retraining. This was slightly less than the standard 80/20 proportion since the remaining 30% used for testing could not be augmented. Therefore taking only 20% would have resulted in an unacceptably small test set. For each retrain-

ing process, 20 epochs were considered.

To further extend the analysis on these datasets and taking advantage of the different species belonging to the same category, we chose two complementary approaches. First, we performed a leave-one-species-out approach, where the network was trained with 3 of the species, using the data of the left-out species as a test set (LOO approach). The idea was to test the network with data that were not used as input before but came from the same category. Second, following a similar idea as the leave one out, the network was re-trained with vocalizations of only one species (OOS approach), and then its performance was tested on the other 3 species separately. The idea behind both approaches was to explore how features from different species vocalizations affect the accuracy in the affiliative and non-affiliative vocalization conditions.

3. RESULTS

First, we trained the VGG16 network with randomly initialized weights, i.e., without any TL approach. This yielded a 44.45% of accuracy on the test set. This results were not satisfactory, as insufficient data was used for training the network. For the second experiment, we proceeded with the TL approach, by using the VGG16 architecture with weights pre-trained on ImageNet. After training, accuracy of 87.93% was obtained across all vocalization types. For the third set of experiments, the leaving-one-out (LOO) approach was evaluated in two conditions: initializing first with random weights and then with the pre-trained weights (i.e with and without a TL approach). It can be noted without that TL, results were again unsatisfactory. However, this time, only an increase in accuracy was observed when testing on babies and chimpanzees. For the fourth and final experiment, the VGG16 network was initialized with pre-trained weights and re-trained with vocalizations from only one species (OOS). Results from the third and fourth set of experiments are shown in 1.

4. DISCUSSION

In the current study, we applied a deep learning approach to investigate small datasets in a bioacoustic classification task. Techniques like transfer learning and data augmentation revealed a significant impact on the performance of the classification algorithm, despite the limited size of the available dataset. Successfully using a pre-trained network that learned from natural images on spectrogram

	Test Set			
Approach	*Baby*	*Dog*	*Chimp.*	*Tree Shrew*
LOO (No TL)	49.45%	50.80%	51.24%	48.86%
LOO (With TL)	58,08%	45,83%	64,58%	43,75%
OOS - Baby	-	37,50%	66,67%	22,92%
OOS - Dog	43,75%	-	47,92%	41,67%
OOS - Chimp.	56,25%	45,83%	-	31,25%
OOS - Tree Shrew	54,17%	50,00%	50,00%	-

Table 1. Accuracy results for the ”Leave one out” (LOO) and training with one species (OOS) approaches. For contrast, results without TL are also shown for the LOO approach.

representations implies that similar features present in natural images (e.g. shapes, textures) are recognized in spectrograms. The relevance of this work for bioacoustics is that we observed that pre-training transferred better between some species than others, hinting that phylogenetic closeness may predict how well pre-training transfers. This idea can, in turn, be contrasted with human perception.

The leave-one-species-out experiment can be useful to gain insight on shared features among vocalizations from different species. When leaving one species dataset out, the classification for a specific sound category in some species increased compared to the rest, and implies that there is a level of feature sharing between vocalizations. Particularly, Table 1 shows that accuracy for chimpanzee and baby vocalizations was higher than for the other two species, suggesting possible acoustic feature sharing between the phylogenetically closer sets of vocalizations. This can also be observed with the OOS experiment (Table 1), where the accuracy for the classification of chimpanzee vocalizations was 66% when trained with the baby datasets, and vice-versa an accuracy of 56% was obtained. In contrast, for the rest of the species, accuracy never surpassed 50%. In particular, the results for the tree shrew were overall worse, and differentiation between calls of other species was not possible (the $\tilde{5}0\%$ of accuracy corresponds to classifying all calls as a single category).

However, It is worth noting that with limited data these interpretations could be rushed. Nonetheless, the current results complement the ones obtained in the original study by Scheumann *et al.*, where vocalizations were used to distinguish the influence of familiarity and phylogeny on voice-induced emotional perceptions in humans [10]. Thus, the results obtained through a deep learning

approach complement human perception results. Moreover, a similar study performed on bird vocalizations also suggests that accuracy in classification drops with phylogenetical distance [15]. Important is that deep learning algorithms only look into raw data, neglecting the influence of any other human-related bias (e.g., cognitive). Therefore comparing human and machine classifications may provide meaningful insight into the mechanisms that underlie the classification of animal vocalizations.

5. ACKNOWLEDGMENTS

This work was funded by the FWO research project "Interactive vocal rhythms", project number G034720N. Bart de Boer received funding from the Flemish Government under the *"Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen"* programme. The Comparative Bioacoustics Group is funded by Max Planck Group Leader funding to A.R. The Center for Music in the Brain is funded by the Danish National Research Foundation (DNRF117).

6. REFERENCES

[1] A. Joly, H. Goëau, S. Kahl, L. Picek, T. Lorieul, E. Cole, B. Deneu, M. Servajean, A. Durso, I. Bolon, H. Glotin, R. Planqué, R. Ruiz de Castaneda, W.-P. Vellinga, H. Klinck, T. Denton, I. Eggel, P. Bonnet, and H. Müller, *Overview of LifeCLEF 2021: an evaluation of Machine-Learning based Species Identification and Species Distribution Prediction*. Sept. 2021.

[2] D. Stowell, "Computational bioacoustics with deep learning: a review and roadmap," *CoRR*, vol. abs/2112.06725, 2021.

[3] N. Hassan, D. A. Ramli, and H. Jaafar, "Deep neural network approach to frog species recognition," in *2017 IEEE 13th International Colloquium on Signal Processing & its Applications (CSPA)*, pp. 173–178, Mar. 2017.

[4] M. Lasseck, "Audio-based bird species identification with deep convolutional neural networks," in *Conference and Labs of the Evaluation Forum*, pp. 1–11, 2018.

[5] E. Tsalera, A. Papadakis, and M. Samarakou, "Comparison of Pre-Trained CNNs for Audio Classification Using Transfer Learning," *J. Sens. Actuator Netw.*, vol. 10, p. 72, Dec. 2021.

[6] V. Arnaud, F. Pellegrino, S. Keenan, X. St-Gelais, N. Mathevon, F. Levréro, and C. Coupé, "Improving the workflow to crack small, unbalanced, noisy, but genuine (sung) datasets in bioacoustics: The case of bonobo calls," *PLoS Comput. Biol.*, vol. 19, pp. 1–47, 04 2023.

[7] Y. X. Zhao, Y. Li, and N. Wu, "Data augmentation and its application in distributed acoustic sensing data denoising," *Geophys. J. Int.*, vol. 228, pp. 119–133, Jan. 2022.

[8] M. Zhong, M. Castellote, R. Dodhia, J. Lavista Ferres, M. Keogh, and A. Brewer, "Beluga whale acoustic signal classification using deep learning neural network models," *J Acoust Soc Am*, vol. 147, p. 1834, Mar. 2020.

[9] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," Apr. 2015. arXiv:1409.1556 [cs].

[10] M. Scheumann, A. S. Hasting, S. A. Kotz, and E. Zimmermann, "The voice of emotion across species: how do human listeners recognize animals' affective states?," *PLoS One*, vol. 9, no. 3, p. e91192, 2014.

[11] M. Scheumann, A. S. Hasting, E. Zimmermann, and S. A. Kotz, "Human novelty response to emotional animal vocalizations: Effects of phylogeny and familiarity," *Front. Behav. Neurosci.*, vol. 11, 2017. Place: Switzerland Publisher: Frontiers Media S.A.

[12] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015.

[13] Y. Jadoul, B. Thompson, and B. de Boer, "Introducing Parselmouth: A Python interface to Praat," *J. Phon.*, vol. 71, pp. 1–15, Nov. 2018.

[14] P. Boersma, "Praat, a system for doing phonetics by computer.," *Glot International*, vol. 5(9/10), p. 341–345, 2001.

[15] K. L. Provost, J. Yang, and B. C. Carstens, "The impacts of fine-tuning, phylogenetic distance, and sample size on big-data bioacoustics," *PLOS ONE*, vol. 17, pp. 1–26, 12 2022.