Understanding Variable Performance on Deep MIL Framework for the Acoustic Detection of Tropical Birds

Entendiendo el Desempeño Variable en el Marco de Trabajo MIL Profundo para la Detección Acústica de Aves Tropicales

Jorge Castro[1], Roberto Vargas-Masís[2], Danny Alfaro-Rojas[3]

Castro, J; Vargas-Masís, R; Alfaro-Rojas, D. Understanding Variable Performance on Deep MIL Framework for the Acoustic Detection of Tropical Birds. *Tecnología en Marcha*. Edición especial 2020. 6th Latin America High Performance Computing Conference (CARLA). Pág 49-54.

https://doi.org/10.18845/tm.v33i5.5075

[1] Advanced Computing Laboratory. Costa Rica National High Technology Center. Email: jcastro@cenat.ac.cr.
https://orcid.org/0000-0003-1553-0461

[2] Laboratorio de Investigación e Innovación Tecnológica. Vicerrectoría de Investigación, Universidad Estatal a Distancia, Costa Rica. E-mail: rovargas@uned.ac.cr
https://orcid.org/0000-0003-1244-4381

[3] Escuela de Ciencias Exactas y Naturales. Universidad Estatal a Distancia, Costa Rica. E-mail: soloard89@gmail.com
https://orcid.org/0000-0001-7694-7194

Keywords

Deep Learning; Multiple Instance Learning; Bioacoustic; Bird Detection.

Abstract

Many audio detection algorithms have been proposed to monitor birds using their vocalizations. Among these algorithms deep learning based techniques have taken the lead in terms of performance at large scale. However, usually a lot of manual work has to be done to correctly label bird vocalizations in large datasets. One way to tackle this limitation is using the Multiple Instance Learning (MIL) framework, which models each recording as a bag of instances, i.e., a collection of audio segments that is associated with a positive label if a bird is present in the recording. In this work, we modified a previously proposed Deep MIL network to predict the presence or absence of birds in audio field recordings of one minute. We explore the behavior and performance of the network when using different number of Mel-Frequency Cepstral Coefficients (MFCC) to represent the recordings. The best configuration found achieved a 0.77 F-score over the validation dataset.

Palabras clave

Aprendizaje profundo; Aprendizaje de instancias múltiples; Bioacústica; Detección de Aves.

Resumen

Se han propuesto muchos algoritmos de detección de audio para monitorear aves usando sus vocalizaciones. Entre estos algoritmos, las técnicas basadas en el aprendizaje profundo han tomado la delantera en términos de rendimiento a gran escala. Sin embargo, usualmente se requiere de mucho trabajo manual para etiquetar correctamente las vocalizaciones de aves en grandes conjuntos de datos. Una forma de abordar esta limitación es usar el marco de trabajo de aprendizaje de instancias múltiples (MIL), que modela cada grabación como una bolsa de instancias, es decir, una colección de segmentos de audio que se asocia con una etiqueta positiva si un pájaro está presente en la grabación. En este trabajo, modificamos una red profunda MIL propuesta previamente, para predecir la presencia o ausencia de aves en grabaciones de campo de un minuto. Exploramos el comportamiento y el rendimiento de la red cuando utilizamos un número diferente de coeficientes cepstrales de frecuencia de mel (MFCC) para representar las grabaciones. La mejor configuración encontrada logró un valor F de 0.77 sobre el conjunto de datos de validación.

Introduction

Birds are key to assess environmental health as indicators of anthropogenic changes [1]. Noninvasive bioacoustic monitoring methodologies present the challenge of developing algorithms to detect birds in a large number of acoustic recordings [2].

Many algorithms have been proposed to classify bird species, bird songs, and individuals [3]. Usually, bird vocalizations are segmented to improve the performance of the classifier. However, these segmentation algorithms are commonly too simple for real conditions in the field or follow a supervised learning scheme were a lot of manual work has to be done to label the vocalizations used for training [4].

The Multiple Instance Learning (MIL) scheme reduce the manual work by using higher abstraction-level labels associated to audio recordings (“bags”) instead of each individual vocalization. The audio segments that compose a recording are known as “instances”. This approach was previously used to classify 13 bird species in ten-seconds recordings [5].

Since Deep Neural Networks (DNNs) have outperformed most of previous classification algorithms and require large training datasets with vocalizations labels [6], [7], it is naturally desirable to combine DNNs with MIL framework (Deep MIL) to reduce the amount of manual work needed. Deep MIL architecture has successfully been applied for image [8] and video [9] classification tasks. In this paper we adapted the Deep MIL architecture to predict the presence or absence of tropical birds (binary classification) in acoustic field recordings.

Deep MIL Architecture

Let $\{(X_1,Y_1),\dots,(X_N,Y_N)\}$ be the training set, where each recording X_i is composed of I audio segments $\{x(_{i,1}),\dots,x(_{i,I})\}$ and each output Y_i is a binary label {1,0} that indicates the presence or absence of bird vocalizations on recording X_i. Thus, the goal is to predict Y_j for any unseen recording X_j. The Deep MIL architecture used is based on the works presented by Jiajun et al. [10] and Gao et al. [11] and it is shown in figure 1. The input of the network is a bag of I instances of size F. First, we extract features of each instance using a 1x1 convolutional layer (CL) with F input channels and P output channels, followed by a batch normalization (BN) and a rectified linear unit (ReLU). Then, we applied a second 1x1 CL with P input channels and P output channels, also followed by a BN and a ReLU. After that, another 1x1CL with P input channels and one output channel is applied, followed by a BN to obtain one unique value per instance. Finally, a max pooling operation is applied over the instances to obtain a score for the bag. Only if the score is higher than zero birds are detected.

To train the network we use the binary cross entropy with logits loss function and the Adam optimizer. We set empirically the batch size to 50, the learning rate α to 0.001, and to 1024. We reduce α each 25 epochs by multiplying it by a factor of 0.96.

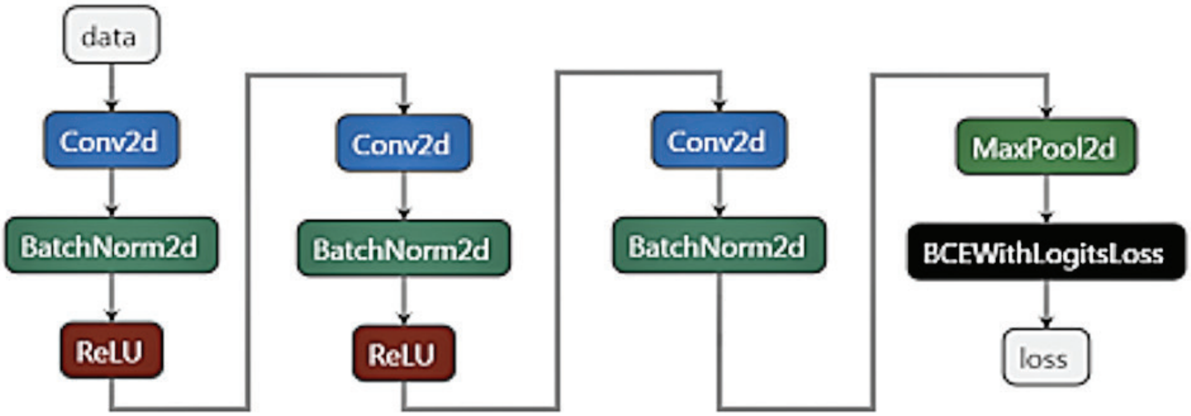

Figure 1. Deep MIL Network architecture.

Data Collection

The acoustic landscape of the upper part of the micro watershed of the Bermúdez River in Heredia Costa Rica (10°3’58.57”N; 84°4’36.39”O) was monitored using Audiomoth recorders in “wav” format, sampling rate of 48 kHz and 16 bits resolution [10]. One-minute recordings were made over seven continuous days at ten-minute intervals. We tagged 2000 recordings where only 1206 presented any vocalization produced by 46 bird species.

Experimental Setup

Each recording was divided into 119 one-second audio segments with 50% of overlap to represent each instance. Each audio segment was also divided into 99 frames of 20 ms with 50% overlap. Then, we obtain four different instance representations based on 15, 30, 60, and 120 Mel-frequency cepstral coefficients (MFCCs) per frame. We chose MFCCs since they have been popularly used for the detection of bird sounds, especially in field conditions where variability in sound quality, diversity of species and intrinsic variability of the vocalizations increase the detection complexity and analysis [11].

We used librosa Python package [12] version 0.6.3 to implement all audio functions. To train the network we randomly divide the 2000 recordings dataset into 80% training, 10% validation, and 10% test. We used the F1-score (harmonic mean of precision and recall) to measure the performance on the training and validation datasets.

Results

The loss function decreased further as the number of MFCCs used increased, as shown in figure 2. However, the F1-score values for both training and validation datasets presented high oscillations for all the number of MFCC used, as shown in figure 3.

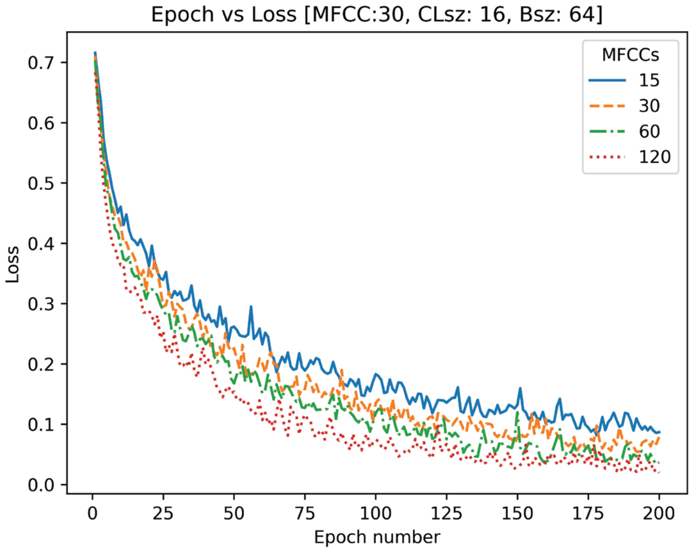

Figure 2. Loss function reduction when using 15, 30, 60 and 120 MFCCs.

As expected, we observed higher F1-score values for the training set, reaching a maximum of 0.98 at epoch 15 when using 120 MFCCs. On the other hand, the highest F1-score achieved for the validation set was 0.77 at epoch 20 when using 60 MFCCs.

Discussion and Future Work

Although the loss function is reduced further when we increase the number of MFCCs used, there are still two main issues to be addressed: the high variability in the F1-score and the high variance error.

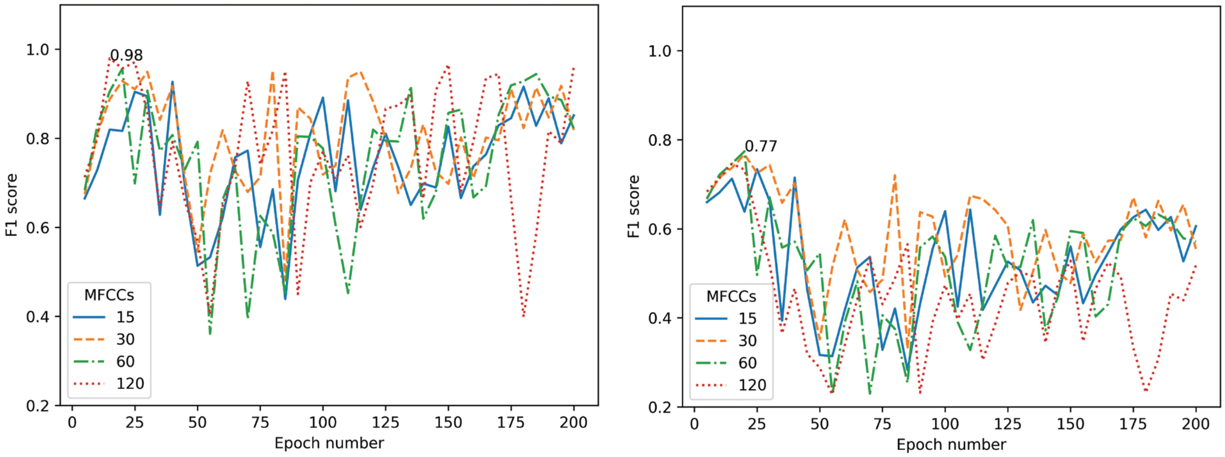

Figure 3. F1-score values in training (left) and validation (right) datasets for each MFCCs over the 200 training epochs. The F1-score was computed every 5 epochs.

The high variability in F1-score performance could be an effect of using longer recordings (60 seconds) with more instances (119) than those used in previous MIL bird classifiers (10 s recordings with an average of 19 instances) [13]. Longer training experiments could also be performed to explore if this behavior continues.

The high variance error between the training and validation sets (about 0.21 comparing their highest F1-score values) could not be reduced using dropout techniques or L2 normalization. Thus, it could also be an effect of using longer recordings with more instances as the majority of instances do not correspond to bird vocalizations and could be triggering the detection of positive bags.

Our future work is focused on exploring other data representations and network architectures to improve the F1-score performance (currently a maximum F-score of 0.77 on the validation set) and reduce it variance error. Furthermore, our long term objective is to generalize this solution using the Multiple Instance Multiple Label (MIML) framework to detect different bird species present in our dataset.

Considering the large amount of acoustic information that can be collected nowadays using autonomous recorders, it is important to develop automatic tools for the conservation and monitoring of biodiversity worldwide.

Acknowledgements

This research was partially supported by a machine allocation on Kabré supercomputer at the Costa Rica National High Technology Center.

References

[1] R. D. Gregory and A. van Strien, “Wild Bird Indicators: Using Composite Population Trends of Birds as Measures of Environmental Health,” Ornithol. Sci., vol. 9, no. 1, pp. 3–22, 2010.

[3] E. C. Knight, K. C. Hannah, G. J. Foley, C. D. Scott, R. M. Brigham, and E. Bayne, “Recommendations for acoustic recognizer performance assessment with application to five common automated signal recognition programs,” *Avian Conserv. Ecol.*, vol. 12, no. 2, p. art14, 2017.

[4] V. Morfi and D. Stowell, “Deep Learning for Audio Event Detection and Tagging on Low-Resource Datasets,” *Appl. Sci.*, vol. 8, no. 8, p. 1397, 2018.

[5] J. F. Ruiz-Muñoz, M. Orozco-Alzate, and G. Castellanos-Dominguez, “Multiple instance learning-based birdsong classification using unsupervised recording segmentation,” in *IJCAI International Joint Conference on Artificial Intelligence*, 2015, vol. 2015-Janua, pp. 2632–2638.

[6] E. Sprengel, M. Jaggi, Y. Kilcher, and T. Hofmann, “Audio based bird species identification using deep learning techniques,” in *CEUR Workshop Proceedings*, 2016, vol. 1609, pp. 547–559.

[7] J. Salamon, J. P. Bello, A. Farnsworth, and S. Kelling, “Fusing shallow and deep learning for bioacoustic bird species classification,” in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2017.

[8] J. Wu, Y. Yu, C. Huang, and K. Yu, “Deep multiple instance learning for image classification and auto-annotation,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2015, vol. 07-12-June, pp. 3460–3469.

[9] R. Gao, R. Feris, and K. Grauman, “Learning to Separate Object Sounds by Watching Unlabeled Video,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2018, vol. 11207 LNCS, pp. 36–54.

[10] J. L. Deichmann, A. Hernández-Serna, J. A. Delgado C., M. Campos-Cerqueira, and T. M. Aide, “Soundscape analysis and acoustic monitoring document impacts of natural gas exploration on biodiversity in a tropical forest,” *Ecol. Indic.*, vol. 74, pp. 39–48, 2017.

[11] S. Fagerlund, “Bird species recognition using support vector machines,” *EURASIP J. Adv. Signal Process.*, vol. 2007, no. 1, pp. 1–8, 2007.

[12] B. McFee et al., “librosa: Audio and Music Signal Analysis in Python,” in *Proceedings of the 14th Python in Science Conference*, 2015, pp. 18–24.

[13] F. Briggs *et al.*, “Acoustic classification of multiple simultaneous bird species: A multi-instance multi-label approach,” *J. Acoust. Soc. Am.*, vol. 131, no. 6, pp. 4640–4650, 2012.