Leveraging Sound Collections for Animal Species Classification with Weakly Supervised Learning

Anonymous submission

Abstract

The utilization of Passive Acoustic Monitoring (PAM) for wildlife monitoring remains hindered by the challenge of data analysis. While numerous supervised ML algorithms exist, their application is constrained by the scarcity of annotated data. Expert-curated sound collections are valuable knowledge bases that could bridge this gap. However, their utilization is hindered by the sparse nature of identifying sounds in these recordings. In this study, we propose a weakly supervised approach to tackle this challenge and assess its performance using the Anuraset dataset. We employ TALNet, a Convolutional Recurrent Neural Network (CRNN) model, training it on 60-second sound recordings labeled for the presence of 42 different anuran species. Evaluation is conducted on 1-second segments, enabling precise sound event localization. Furthermore, we investigate the impact of varying the length of the training input and explore different pooling functions' effects on TALNet's performance. In summary, our findings demonstrate the effectiveness of TALNet in harnessing weakly annotated sound collections for wildlife monitoring. Future research endeavors will involve extending this approach to collections of focal recordings and integrating domain expertise through a human-in-the-loop approach for further refinement and evaluation.

1 Introduction

Passive acoustic monitoring (PAM) has emerged as a key technology for wildlife monitoring (Sugai et al. 2019) and provides a way to promote biodiversity, assess and understand the impact of climate change, and develop intervention strategies to preserve ecosystems. However, handling the large amount of data generated by PAM still poses a barrier for adoption by both researchers and biodiversity managers. Although a wide range of supervised machine learning methods for analyzing PAM datasets (e.g., for sound event detection) exist (Stowell 2022), their application is often constrained by the availability of domain-specific annotated data. Biologists traditionally rely on museum collections for studying biodiversity (Meineke et al. 2018). In modern times, multimedia registers have become increasingly important and recognized as valuable in common practice. Among these, sound archives and collections hold significant importance (Dena et al. 2020; Sugai and Llusia 2019). Several such collections exist, such as FNJV[1], Macaulay library[2], and Xeno-Canto[3]. These resources serve as valuable sources of annotated data for training models to automate sound event detection in large PAM datasets. However, their potential for this task is currently limited because these sound files are weakly annotated, meaning that sound recordings are labeled only at the file level, with no information about the timestamps of specific identifying sounds. This problem is further compounded by the presence of multiple signals in these recordings, such as other species co-occurring in the same soundscape, and the voice of the naturalist who performed the recording, often speaking into the microphone and providing metadata such as species name and a description of the recording context. Effective utilization of such knowledge bases for powering ML tools rely on isolating the meaningful, identifying portions of the sound recordings. In this paper, we propose a weakly supervised method to leverage existing sound collections and generate training data for ML models for species level sound event detection in PAM datasets, Figure 1.

2 Related Work

Deep learning methods have proven very useful for detection of sound events in PAM datasets. Among the most popular convolutional neural network (CNN) architectures applied to PAM are ResNet (He et al. 2016), VGG (Simonyan and Zisserman 2015) and DenseNet (Huang et al. 2017). Even though they were created for computer vision tasks, these architectures proved to be very efficient in analyzing sound data. BirdNet (Kahl et al. 2021) developed an EfficientNet-based model for detection of bird vocalizations. Other popular methods include convolutional recurrent neural networks (CRNNs), that combine the advantages of both CNNs and RNNs (Tzirakis et al.; Çakır et al. 2017; Xie et al. 2020).(Dufourq et al. 2022) compare the performance of different models pretrained on ImageNet (Deng et al. 2009) on different PAM datasets. They show that transfer learning can be used successfully on small PAM datasets with few samples per species.

Availability of training data is crucial for the development of supervised ML models. BirdNet, a popular pre-trained model for species-level classification of bird vocalizations

[1] https://www2.ib.unicamp.br/fnjv/

[2] https://www.macaulaylibrary.org/

[3] https://xeno-canto.org/

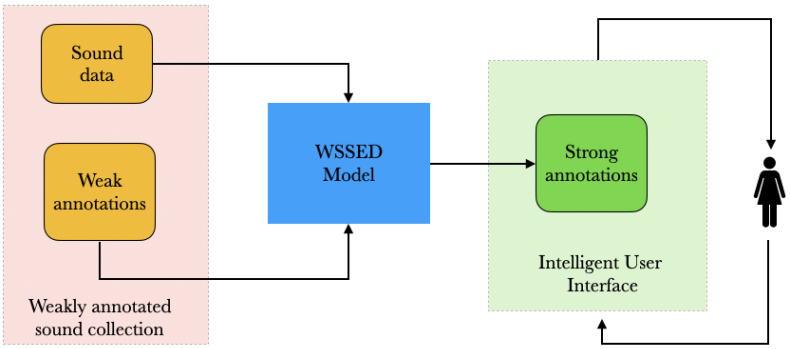

Figure 1: A schematic of our proposed approach

(Kahl et al. 2021), is trained on datasets that consist largely of weakly annotated focal recordings. For detecting the presence of target sounds, they used heuristic image processing methods for signal-strength estimation (Sprengel et al. 2016).

In recent years, there has been a notable surge of interest within the research community in the domain of weakly supervised sound event detection (WSSED), which has been notably catalyzed by initiatives like the Detection and Classification of Acoustic Scenes and Events (DCASE) challenges and the release of extensive audio datasets such as AudioSet (Gemmeke et al. 2017) that provide baselines for the development and evaluation of ML methods related to sound event detection (SED) and specifically WSSED. (Kumar and Raj 2016) propose that WSSED can be treated as a problem of Multi Instance Learning; from this perspective, every audio file can be viewed as a bag of instances of sound events. They explore SVM and neural network based approaches trained on weak labels for detection and achieve temporal localization of sound events.(Xu et al. 2018) introduce an attention mechanism, replacing the ReLU activation function after each convolution with GLUs. (Wang, Li, and Metze 2019) proposed TALNet, a CRNN for audio tagging and localization. They identify the best pooling function for the task. More recent approaches propose transformer-based methods for WSSED (Lin et al. 2020; Miyazaki et al. 2020). Current approaches combine embeddings extracted from pre-trained models such as BEATS (Chen et al. 2022) with CRNN classifiers aligning with the newest requirements of the DCASE challenges that use heterogeneous datasets that contain unlabeled, weakly labeled and synthetic datasets with strong annotations. In our work, we focus only on methods for weakly annotated datasets and how to use them to enrich annotations for PAM.

Table 1: Performance metrics on Anuraset

Architecture	**Global**	**1s**		
	F1 Score	F1 Score	Precision	Recall
TALNet	90.00	**64.68**	50.82	88.94
ResNet50	90.11	63.54	54.79	75.60

3 Implementation

Dataset For our experiments, we used AnuraSet, a recently released benchmark PAM dataset comprised of 1612 minutes of omindirectional recordings from four different sites in two Brazilian biomes: Cerrado and Atlantic Forest (Cañas et al. 2023). The dataset consists of 60 seconds long recording files, as well as manually created expert annotations for 42 species of anurans (frogs and toads). The annotations consist of strong labels, i.e., species identity plus on- and offset times for each call occurrence.

Data Preprocessing The audio recordings were represented as Mel-frequency single channel spectrograms $S \epsilon R^{mxn}$, where $m = 64$ is the number of frequency bins and n is the number of frames. As "frame" we denote the minimal time segment, so m depends on the length of the input files. For the 60-second long recordings $m = 2400$. To compute the spectrograms, we used a window size of 1102 and hop length 551. Raw recordings were resampled to 22kHz. For comparing performance when training is carried with inputs of different durations, we partitioned the 60-second audio recordings from the training set into non-overlapping 10-second and 3-second long segments. We kept the same frame length and number of frequency bins as described in TALNet (Wang, Li, and Metze 2019), but adjusted the number of frames according to the segment length. Considering the unbalanced nature and relatively small size of the dataset when training with 60-second long input, we performed iterative stratification to ensure balanced train and test splits, with 80% for training and 20% for test. The test set recordings were partitioned into non-overlapping 1-second long segments. For each segment, a vector of binary labels was generated to indicate presence of calls from each of the 42 species; each entry was set to 1 if a call of that species was present anywhere in the corresponding segment, and 0 otherwise. To create the Mel-frequency spectrograms, we used native torchaudio transforms[4] for au-

[4]https://pytorch.org/audio/stable/transforms.html

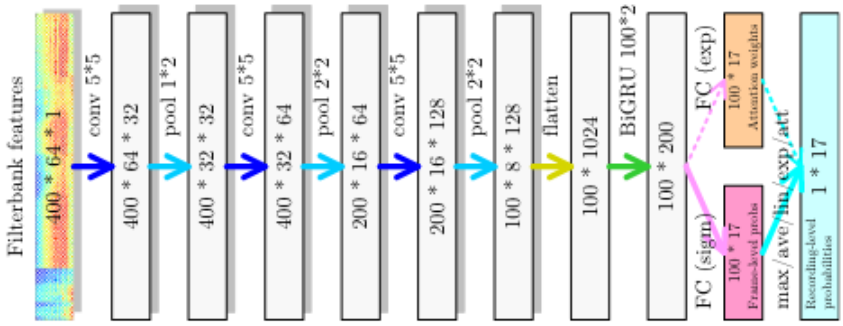

Figure 2: TALNet architecture as proposed in (Wang, Li, and Metze 2019) for the DCASE 2017 challenge. We increase the input size to 2400 frames for 60-second long audio files and set the number of classes to 42 for the anuran species.

Table 2: Comparison of pooling functions for TALNet trained on 60-second long inputs and evaluated on either 60-second (global) or 1-second long inputs. See (Wang, Li, and Metze 2019) for details.

Pooling Function	**Global**	**1s**		
	F1 Score	F1 Score	Precision	Recall
Average	88.22	**65.7**	51.74	89.97
Max pooling	63.96	47.76	54.77	42.34
Exponential Softmax	70.87	56.64	46.97	71.31
Linear Softmax	**90.00**	64.68	50.82	88.94
Attention pooling	70.50	49.42	40.02	64.56

dio processing.

Model architecture For the sound event detection and localization we used TALNet (Wang, Li, and Metze 2019) a convolutional recurrent neural network developed for audio tagging and localization on AudioSet and the DCASE challenge 2017. The network consists of three convolutional layers, five pooling layers and one recurrent layer, Figure 3.

To perform WSSED using transfer learning on the PAM dataset, we used Resnet50 (He et al. 2016) pretrained on ImageNet (Deng et al. 2009).

Experimental setup We train the network on samples of the AnuraSet with weak labels (3-, 10-, or 60-seconds long samples) and evaluate the performance using the strong labels (1-second). In the training procedure we use the Adam optimizer (Kingma and Ba 2014), and a learning rate of 3×10^{-4}. As a loss function we use the binary cross entropy loss:

$$L(y, \hat{y}) = -(y * log(\hat{y} + (1 - y) * log(1 - \hat{y})) \quad (1)$$

In equation 1, y represents the true labels, while $\hat{y}$ the predicted probabilities. Time and frequency masking were applied as suggested in SpecAugment (Park et al. 2019). We create shuffled batches of size 32 samples and trained for 100 epochs.

Evaluation metric is micro-averaged F1 scored, unless otherwise indicated. To compare the model performance on different input lengths, we also conduct experiments with 10-second (original TALNet input size) and 3-second long files.

4 Results

We assessed the model's performance on PAM data using the Anuraset dataset. As evaluation metrics, we used *global F1 score*, a metric assessing how well the model can identify only the presence or absence of events within an audio file, and *1-second F1 score*, an indication of how well the model can localize sound events in an audio file with a precision of 1 second.

We start by analyzing models trained on 60-second long inputs. To compute F1 scores for both tagging and localization tasks, we used weak and strong labels. For this, we made predictions on 1-second windows by aggregating probabilities across 10 frames, followed by the application of a threshold as described in (Wang, Li, and Metze 2019). To compare the performance of TALNet with a pretrained model, we used Resnet50. In Table 1 we report the global and 1-second F1 scores on Anuraset. Since the performance of the model on 1 second segments is essential for our goal, we report the related precision and recall too. As it is evident from the table TALNet performs better than Resnet50

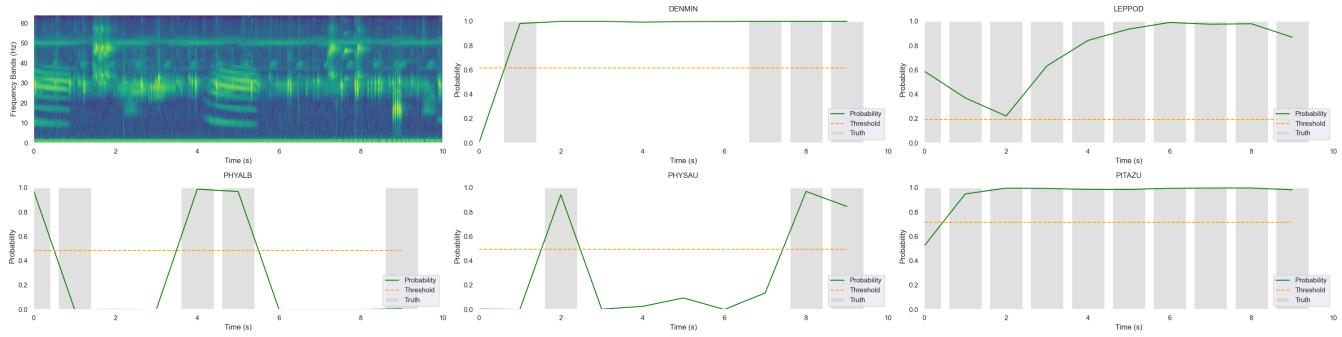

Figure 3: Spectrogram of a representative 10-second audio segment, and bar plots with predicted and observed labels for five example species at 1-second resolution. Gray bars are true labels; green line is predicted probability of species occurrence; yellow line is the decision threshold. Notice that species LEPPOD, PHYALB, PHYSAU, and PITAZU are correctly localized by the model, while DENMIN gets mistakenly identified as occurring during the entire duration of the audio clip.

in the localization task (1s segments) and slightly worse in the tagging task (60s segments). We use TALNet for the following experiments. TALNet treats WSSED as a multiple instance learning problem; specifically, the strategy consists of training models to make predictions for each frame of a multi-frame data point, and aggregate frame level predictions with a pooling function. Table 2 shows the results for different pooling functions. Since TALNet was developed and evaluated for 10 second long audio files, we trained and evaluated its performance on three different input lengths (table 3). The decrease of the input length to 10 seconds improved the performance by 32% for the 1-second F1 score and 7.5% for the global F1 score, indicating that the model's sensitivity to input length is task dependent.

5 Conclusion and Future Work

In this paper, we proposed the use of an existing CRNN based approach such as TALNet to harness more information from weakly annotated data for wildlife monitoring. We demonstrated that domain transfer of existing models developed for AudioSet to passive acoustic monitoring (PAM) datasets does not always require complex model architecture and input modifications. We achieved positive results for both tagging and localization of animal sounds.

The final evaluation of the system will be done, however, when applying it to PAM collections to generate annotated data from the weakly labelled recordings. Based on the promising results using Anuraset we will train TALNet using the recordings of anuran calls and the weak annotations from museum collections such as the FNJV collection. The evaluation will be twofold: calculating evaluation metrics using the strong labels from Anuraset and utilizing domain expertise using an intelligent user interface where domain experts can investigate the results and provide feedback for the quality of the annotations in a human-in-the-loop approach.

Table 3: Micro F1 score of the TALNet model trained on inputs of varying duration (3, 10, or 60 seconds), and evaluated on samples of 60 seconds (global) or 1 second duration. Performance drops as duration of training samples increases.

Length of Training Input	**Micro F1 Score**	
	Global	1s
3s	96.71	**87.94**
10s	96.77	85.46
60s	90.00	64.68

References

Cañas, J. S.; Toro-Gómez, M. P.; Sugai, L. S. M.; Restrepo, H. D. B.; Rudas, J.; Bautista, B. P.; Toledo, L. F.; Dena, S.; Domingos, A. H. R.; de Souza, F. L.; Neckel-Oliveira, S.; da Rosa, A.; Carvalho-Rocha, V.; Bernardy, J. V.; Sugai, J. L. M. M.; Santos, C. E. d.; Bastos, R. P.; Llusia, D.; and Ulloa, J. S. 2023. AnuraSet: A dataset for benchmarking Neotropical anuran calls identification in passive acoustic monitoring. 0 citations (Semantic Scholar/arXiv) [2023-08-12] 0 citations (Semantic Scholar/DOI) [2023-08-12] arXiv:2307.06860 [cs, eess].

Chen, S.; Wu, Y.; Wang, C.; Liu, S.; Tompkins, D.; Chen, Z.; and Wei, F. 2022. BEATs: Audio Pre-Training with Acoustic Tokenizers. 15 citations (Semantic Scholar/arXiv) [2023-08-12] 15 citations (Semantic Scholar/DOI) [2023-08-12] arXiv:2212.09058 [cs, eess].

Dena, S.; Rebouças, R.; Augusto-Alves, G.; Zornosa-Torres, C.; Pontes, M. R.; and Toledo, L. F. 2020. How much are we losing in not depositing anuran sound recordings in scientific collections? *Bioacoustics*, 29(5): 590–601. 7 citations (Semantic Scholar/-DOI) [2023-08-12] Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/09524622.2019.1633567.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Dufourq, E.; Batist, C.; Foquet, R.; and Durbach, I. 2022. Passive acoustic monitoring of animal populations with transfer learning. *Ecological Informatics*, 70: 101688. 12 citations (Semantic Scholar/DOI) [2023-08-12].

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

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Huang, G.; Liu, Z.; Van Der Maaten, L.; and Weinberger, K. Q. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4700–4708.

Kahl, S.; Wood, C. M.; Eibl, M.; and Klinck, H. 2021. BirdNET: A deep learning solution for avian diversity monitoring. *Ecological Informatics*, 61: 101236. 129 citations (Semantic Scholar/DOI) [2023-08-12].

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kumar, A.; and Raj, B. 2016. Audio Event Detection using Weakly Labeled Data. In *Proceedings of the 24th ACM international conference on Multimedia*, 1038–1047. ArXiv:1605.02401 [cs].

Lin, L.; Wang, X.; Liu, H.; and Qian, Y. 2020. Specialized Decision Surface and Disentangled Feature for Weakly-Supervised Polyphonic Sound Event Detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28: 1466–1478. Conference Name: IEEE/ACM Transactions on Audio, Speech, and Language Processing.

Meineke, E. K.; Davies, T. J.; Daru, B. H.; and Davis, C. C. 2018. Biological collections for understanding biodiversity in the Anthropocene. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 374(1763): 20170386. 147 citations (Semantic Scholar/DOI) [2023-08-12] Publisher: Royal Society.

Miyazaki, K.; Komatsu, T.; Hayashi, T.; Watanabe, S.; Toda, T.; and Takeda, K. 2020. Weakly-Supervised Sound Event Detection with Self-Attention. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 66–70. ISSN: 2379-190X.

Park, D. S.; Chan, W.; Zhang, Y.; Chiu, C.-C.; Zoph, B.; Cubuk, E. D.; and Le, Q. V. 2019. Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv preprint arXiv:1904.08779*.

Simonyan, K.; and Zisserman, A. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. 9999 citations (Semantic Scholar/arXiv) [2023-08-03] arXiv:1409.1556 [cs] version: 6.

Sprengel, E.; Jaggi, M.; Kilcher, Y.; and Hofmann, T. 2016. Audio Based Bird Species Identification using Deep Learning Techniques. *LifeCLEF 2016*. Meeting Name: Conference and Labs of the Evaluation Forum (CLEF) 2016.

Stowell, D. 2022. Computational bioacoustics with deep learning: a review and roadmap. *PeerJ*, 10: e13152. 69 citations (Semantic Scholar/DOI) [2023-08-12].

Sugai, L. S. M.; and Llusia, D. 2019. Bioacoustic time capsules: Using acoustic monitoring to document biodiversity. *Ecological Indicators*, 99: 149–152. 34 citations (Semantic Scholar/DOI) [2023-08-12].

Sugai, L. S. M.; Silva, T. S. F.; Ribeiro, J. W.; and Llusia, D. 2019. Terrestrial Passive Acoustic Monitoring: Review and Perspectives. *BioScience*, 69(1): 15–25. 215 citations (Semantic Scholar/DOI) [2023-08-12].

Tzirakis, P.; Shiarella, A.; Ewers, R.; and Schuller, B. W. ???? Computer Audition for Continuous Rainforest Occupancy Monitoring: The Case of Bornean Gibbons' Call Detection.

Wang, Y.; Li, J.; and Metze, F. 2019. A Comparison of Five Multiple Instance Learning Pooling Functions for Sound Event Detection with Weak Labeling. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 31–35. 146 citations (Semantic Scholar/DOI) [2023-08-12] ISSN: 2379-190X.

Xie, J.; Hu, K.; Zhu, M.; and Guo, Y. 2020. Bioacoustic signal classification in continuous recordings: Syllable-segmentation vs sliding-window. *Expert Systems with Applications*, 152: 113390.

Xu, Y.; Kong, Q.; Wang, W.; and Plumbley, M. D. 2018. Large-Scale Weakly Supervised Audio Classification Using Gated Convolutional Neural Network. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 121–125. ISSN: 2379-190X.

Çakır, E.; Parascandolo, G.; Heittola, T.; Huttunen, H.; and Virtanen, T. 2017. Convolutional Recurrent Neural Networks for Polyphonic Sound Event Detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(6): 1291–1303. Conference Name: IEEE/ACM Transactions on Audio, Speech, and Language Processing.