Leveraging Weakly Supervised and Multiple Instance Learning for Multi-label Classification of Passive Acoustic Monitoring Data

Ilira Troshani, Thiago S. Gouvêa, Daniel Sonntag

Interactive Machine Learning, German Research Centre for Artificial Intelligence (DFKI).
Carl von Ossietzky University of Oldenburg, Oldenburg.

ilira.troshani@dfki.de; thiago.gouvea@dfki.de; daniel.sonntag@dfki.de;

Abstract

Data collection and annotation are time-consuming, resource-intensive processes that often require domain expertise. Existing data collections such as animal sound collections provide valuable data sources, but their utilization is often hindered by the lack of fine-grained labels. In this study, we examine the use of existing weakly supervised methods to extract fine-grained information from existing weakly-annotated data accumulated over time and alleviate the need for collection and annotation of fresh data. We employ TALNet, a Convolutional Recurrent Neural Network (CRNN) model and train it on 60-second sound recordings labeled for the presence of 42 different anuran species and compare it to other models such as BirdNet, a model for detection of bird vocalisation. We conduct the evaluation 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 the model's performance on AnuraSet. Finally, we integrate it in an interactive user interface that facilitates training and annotation. Our findings demonstrate the effectiveness of TALNet and BirdNet in harnessing weakly annotated sound collections for wildlife monitoring. Our method not only improves the extraction of information from coarse labels but also simplifies the process of annotating new data for experts.

Keywords: Weakly Supervised Learning, semi-automatic data annotation, AI Transfer, Sustainability

1 Introduction

Passive acoustic monitoring (PAM), has emerged as a key technology for wildlife monitoring [1] while using acoustic sensors 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 [2, 3]. Although a wide range of supervised machine learning methods for analyzing PAM datasets (e.g., for sound event detection) exist [4], their application is often constrained by the availability of domain-specific annotated data. Biologists traditionally rely on museum collections for studying biodiversity [5]. 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 [6, 7]. 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 species 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 sources 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 weakly annotated data 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 [8], VGG [9] and DenseNet [10]. Even though they were created for computer vision tasks, these architectures proved to be very efficient in analyzing sound data. Kahl et al. [11] developed BirdNet, an EfficientNet-based model for detection of bird vocalisations. Other popular methods include convolutional recurrent neural networks (CRNNs), that combine the advantages of both CNNs and RNNs [12–14]. [15, 16] compare the performance of different models pre-trained on ImageNet [17] 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 that generalize well to new recording locations, background sounds and regional differences in the species calls. BirdNet, a popular pre-trained model for species-level

[1] https://www2.ib.unicamp.br/fnjv/
[2] https://www.macaulaylibrary.org/
[3] https://xeno-canto.org/

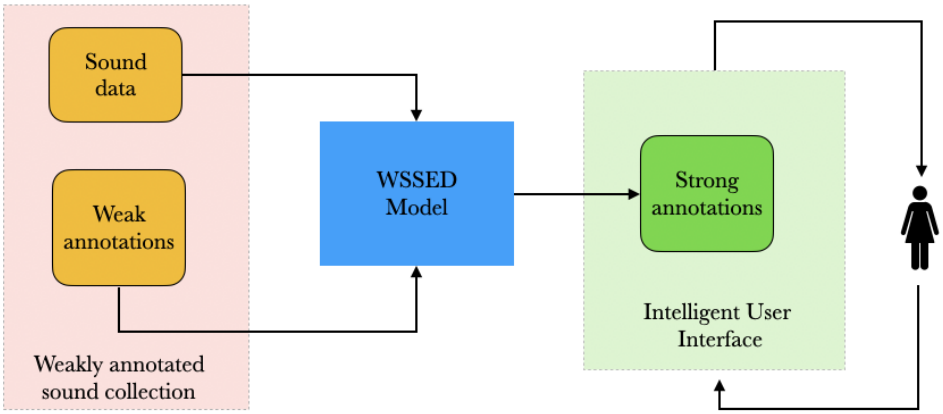


Fig. 1: A schematic of our proposed approach

classification of bird vocalizations [11], 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 [20]. These recordings are often acquired with professional equipment and thus have a high sound-to-noise ratio (SNR).

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 [21] that provide baselines for the development and evaluation of ML methods related to sound event detection (SED) and specifically WSSED. Kumar et al. [22] 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 B of instances x_i of sound events. If one or more instances have label $y_i = 1$, then the whole bag is considered to be positive and is assigned label Y. They explore SVM and neural network based approaches trained on weak labels for detection and achieve temporal localization of sound events, harnessing richer information from the annotations within the original

Table 1: Performance metrics on AnuraSet

Architecture	**Global**	**1s**		
	F1 Score	F1 Score	Precision	Recall
TALNet [18]	90.00	64.68	50.82	88.94
ResNet-50 [8]	**90.11**	63.54	54.79	75.60
ResNet-18 [8]	88.89	62.85	56.73	70.46
VGGish [19]	69.64	41.67	31.36	59.92
BirdNet [11]	84.80	**71.14**	67.42	75.30

dataset. Xu et al. [23] introduce an attention mechanism, replacing the ReLU activation function after each convolution with GLUs. Wang et al. [18] propose 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 [24, 25]. Current approaches combine embeddings extracted from pre-trained models such as BEATS [26] 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. Some of the deep learning methods for detection of sound events in animal sounds datasets are associated with a user interface for result inspection and easy annotation [27–29]. We contribute to this body of research by exploring how WSSED can be used for wildlife monitoring to support biodiversity conservation.

3 Implementation

In our implementation, we apply existing weakly supervised methods to extract detailed information from weakly-annotated data.

Dataset

For our experiments, we use 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 [30]. 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.

Table 2: Comparison of pooling functions for BirdNET finetuned on 60-second long inputs and evaluated on either 60-second (global) or 1-second long segments.

Pooling Function	**Global**	**1s**		
	F1 Score	F1 Score	Precision	Recall
Average (2)	79.90	65.64	52.20	88.40
Max pooling (1)	81.41	36.70	74.01	24.40
Exponential Softmax(4)	81.57	64.38	51.10	86.98
Linear Softmax (3)	**84.80**	**71.14**	67.42	75.30
Attention pooling (5)	80.84	67.63	55.45	86.65

Data Preprocessing

The audio recordings are represented as Mel-frequency single channel spectrograms $S \in R^{m \times n}$, 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 n depends on the

length of the input files. For the 60-second long recordings $n = 2400$. To compute the spectrograms, we use a window size of 1102 and hop length 551. Raw recordings are resampled to 22kHz. For comparing performance when training is carried with inputs of different durations, we partition the 60-second audio recordings from the training set into non-overlapping 9-second and 3-second long segments. We keep the same frame length and number of frequency bins as described in TALNet [18], but adjust 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 perform iterative stratification to ensure balanced train and test splits, with 80% for training and 20% for test. For each segment, a vector of binary labels is generated to indicate presence of calls from each of the 42 species; each entry is set to 1 if a call of that species is present anywhere in the corresponding segment, and 0 otherwise. To create the Mel-frequency spectrograms, we use native torchaudio [31] transforms for audio processing.

Model architecture

For the sound event detection and localization we use TALNet [18] 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, a pooling layers and one recurrent layer.

To perform WSSED using transfer learning on the PAM dataset, we use ResNet-50 and ResNet-18[8] pretrained on ImageNet [17], VGGish [32] pretrained on AudioSet and BirdNet [11] pretrained on bird vocalisations, leveraging their feature extraction capabilities to capture fundamental patterns typical for spectrograms.

All presented models treat 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 then apply a pooling function. The pooling function combines frame level predictions into segment level ones while retaining important information. The pooling layer applies different pooling functions such as Max Pooling (equation 1), Average Pooling (equation 2), Linear Softmax (equation 3), Exponential Softmax (equation 4) and Attention pooling (equation 5)

$$y = max_i y_i \tag{1}$$

$$y = \frac{1}{n}\sum_i y_i \tag{2}$$

$$y = \frac{\sum_i y_i^2}{\sum_i y_i} \tag{3}$$

$$y = \frac{\sum_i y_i exp(y_i)}{\sum_i exp(y_i)} \tag{4}$$

$$y = \frac{\sum_i y_i w_i}{\sum_i w_i} \tag{5}$$

Experimental setup

We train the network on samples of the AnuraSet with weak labels (60-seconds long samples) and evaluate the performance using the strong labels (1-second). In the training procedure, we use the Adam optimizer [33], and a learning rate of 3×10^{-4}. As a loss function we use the binary cross entropy loss:

$$L(y, \hat{y}) = -(y * log(\hat{y} + (1 - y) * log(1 - \hat{y})) \tag{6}$$

In equation 6, y represents the true labels, while $\hat{y}$ the predicted probabilities. Time and frequency masking are applied as suggested in SpecAugment [34]. We create shuffled batches of size 32 samples and train for 100 epochs.

Evaluation metric is the *global F1 score* (equation 7, 8) 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 localise sound events in an audio file with a precision of 1 second.

$$F1 = 2 \times \frac{precision \times recall}{precision + recall} \tag{7}$$

where,

$$\begin{aligned} precision &= \frac{TP}{TP + FP} \\ recall &= \frac{TP}{TP + FN} \end{aligned} \tag{8}$$

To compare the model performance on different input lengths, we also conduct experiments with 9-second and 3-second long files.

4 Interactive training and annotation tool

We create a user interface to allow the user to train a model from weakly annotated data and use it to inspect results and change weakly-supervised learning parameters at inference time, enabling them to do semi-automatic annotation of new unlabeled data.

In our interface the user can change parameters such as pooling function and prediction window size. This human intervention at test time allows for improvement of the final results. We design the tool with two user personas in mind: the expert and the novice, therefore we create a training (Figure 2) and an annotation pane (Figure 3a).

4.1 Training

The training pane has three main areas that are dynamically filled upon user action to minimize user input and collapsible for less visual clutter. Here the expert can select a weakly-annotated dataset and a pretrained model, such as BirdNet [11] or ResNet-50 [8]. As the pooling function plays an important role in the performance of weakly supervised learning models [18, 35] we make it a configurable parameter.

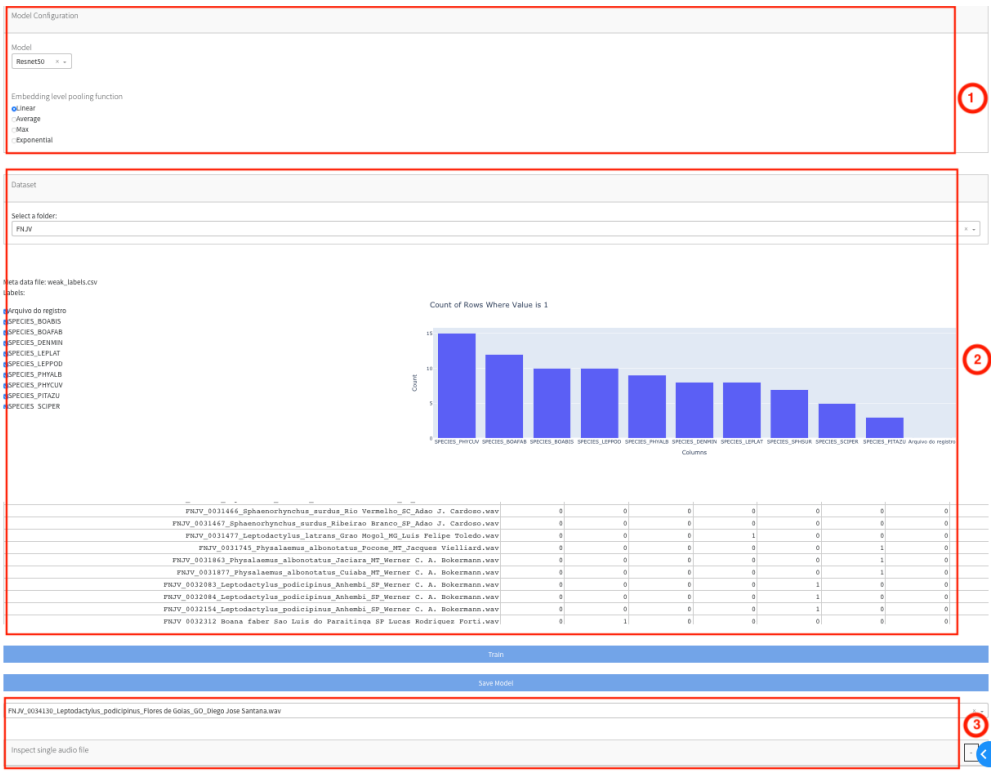
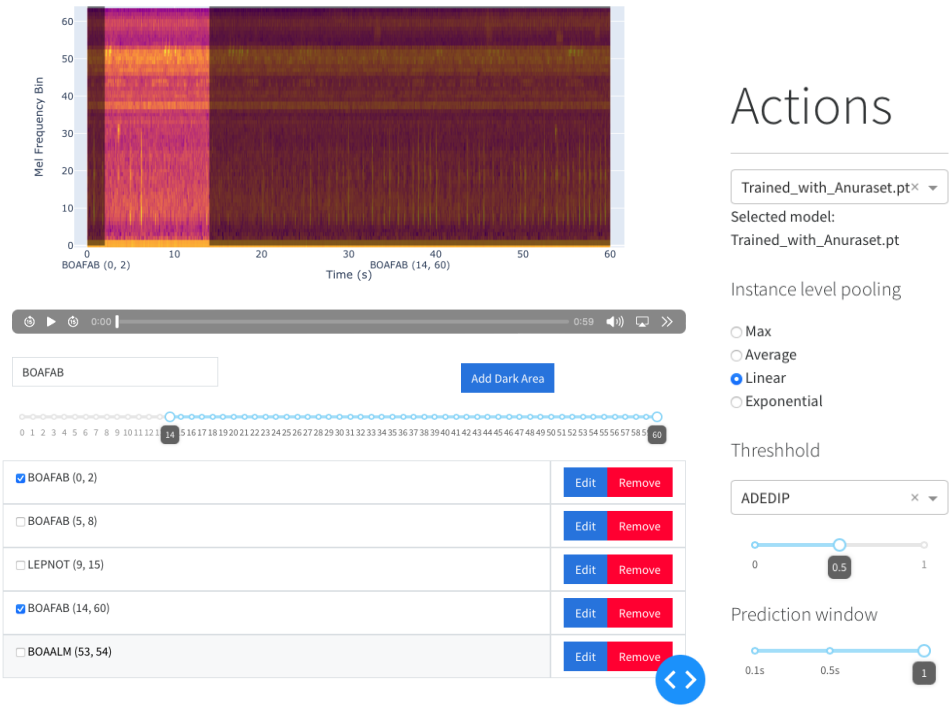


Fig. 2: Training pane: Domain experts can train a model with a backbone of choice of pretrained models and configure parameters such as pooling function (model configuration area (1)) using a weakly annotated dataset, such as a subset of the FNJV dataset for anurans. Upon selection of the dataset, the label distribution is computed for a better overview of the dataset (dataset area (2)). At the bottom (single file inspection area (3)), the user can inspect single spectrograms, here collapsed.

4.2 Annotation

In the annotation pane, the user can upload and annotate an audio file. Upon upload, inference runs in the background while the spectrogram of the audio file is computed and displayed. Upon selection, the time intervals are displayed as gray overlays on the spectrogram. The user can modify them, delete or add new ones. For the more experienced user we have created a Settings section (Figure 3b) to modify post-hoc model parameters related to weakly supervised learning such as pooling function, prediction threshold to make the model more or less sensitive towards certain classes and prediction window. Currently, our tool is not meant for automatic dataset annotation, but to help novice and expert users, generate accurate annotations in a more efficient way. The application is implemented in Python using Dash [4] and PyTorch [36].

[4] https://dash.plotly.com



(a) Single File annotation

(b) WSL Settings of the weakly supervised learning model.

Fig. 3: Annotation pane: Upon file upload, the model annotates the file with the detected species' calls and onset, offset time. The user can modify the predictions if necessary (a) using UI components such as slider, buttons and input fields. The more experienced machine learning practitioner can adjust the post-hoc parameters of the weakly supervised setting to achieve the best performance such as pooling function, prediction threshold and the precision of the prediction window for the onset-offset times (0.1 - 1s)(b).

5 Results

We assess the model's performance on PAM data using the AnuraSet dataset.

We start by analyzing models trained on 60-second long inputs. To compute F1 scores for both tagging and localization tasks, we use weak and strong labels. For this, we make predictions on 1-second windows by aggregating probabilities across 10 frames, followed by the application of a threshold as described in [18]. To compare the performance of TALNet with a pretrained model, we use ResNet-50, ResNet-18, VGGish and BirdNet. 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

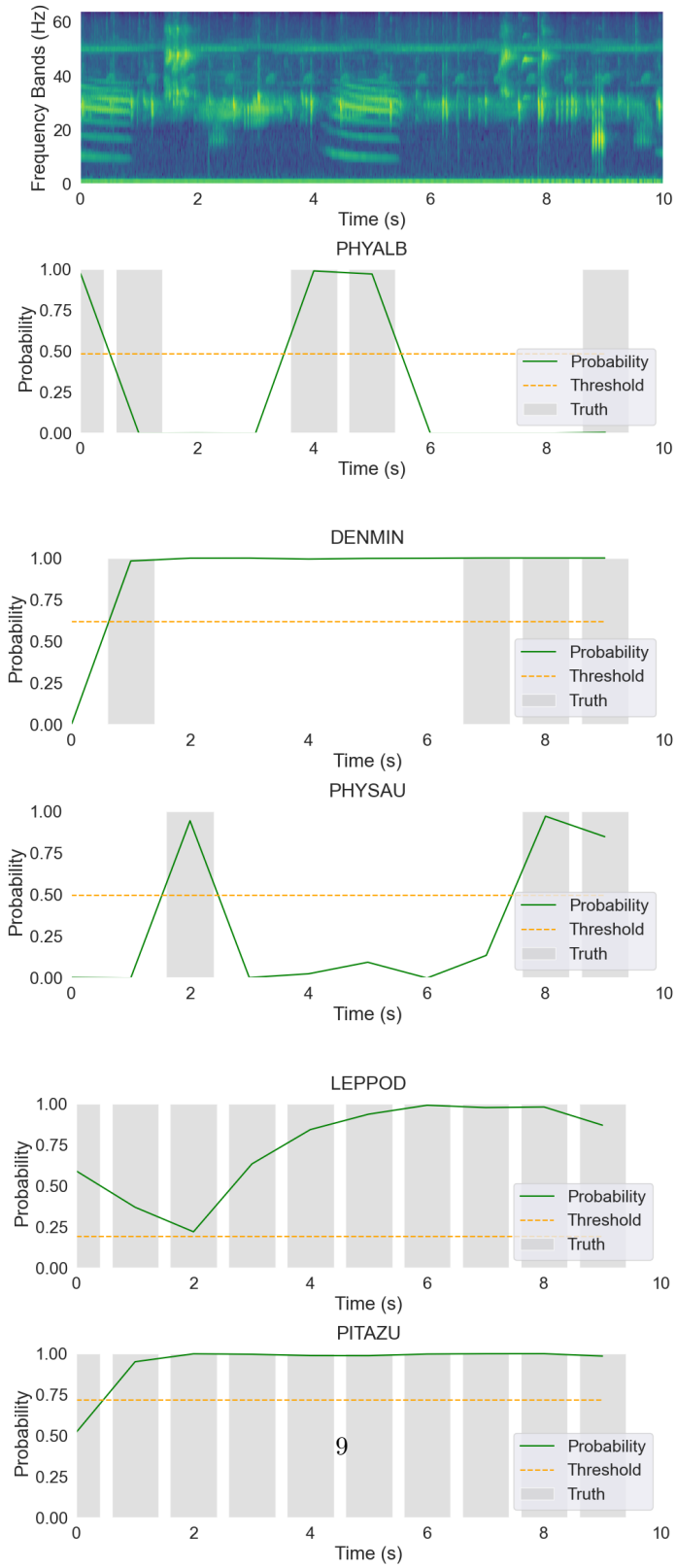


Fig. 4: Spectrogram of a representative 10-second audio segment, and bar plots with predicted and observed labels for five example species at 1-second resolution. Gray bars are true labels; green line is predicted probability of species occurrence; yellow line is the decision threshold. Notice that species `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.

Table 3: Micro F1 score of the BirdNet model trained on inputs of varying duration (3, 9, or 60 seconds), and evaluated globally and on 1-second long segments. Performance drops as duration of training samples increases.

Length of Training Input	**Micro F1 Score**	
	Global	1s
3s	**90.27**	**74.50**
9s	89.81	71.00
60s	84.80	71.14

TALNet performs better than ResNet-50 in the localization task (1s segments) and slightly worse in the tagging task (60s segments) but is outperformed by BirdNet in the localization task (1s segements). Figure 4 illustrates prediction results for a 10s long file with five species present.

Table 2 shows the results for different pooling functions on AnuraSet. To analyze the influence of input length in performance, we finetune and evaluate BirdNet on three different input lengths (table 3). The decrease of the input length to 9 seconds improves the performance by 3.36% for the 1-second F1 score and 5.47% for the global F1 score, indicating that the model's sensitivity to input length is task dependent. This finding confirms the results in [37]

6 Conclusion and Future Work

In this paper, we proposed the use of the existing CRNN based approach TALNet and pretrained models such as BirdNet to harness more information from weakly annotated data for wildlife monitoring and evaluated its performance on a benchmark PAM dataset. We demonstrated that domain transfer of existing models developed for different acoustic environments, such as the one in AudioSet to passive acoustic monitoring (PAM) datasets does not always require a complex model architecture and input modifications. With TALNet we achieved a 90% global F1 score in the tagging task while with BirdNet 71.14% F1 score in the localization task of animal sounds for 60-second long recordings. Further, we implemented a user interface to allow domain experts to train the model and investigate the results. Designed with the human "component" in mind, our tool makes model configuration and inference as transparent as possible and allows the user to modify the results if necessary. We plan to refine our tool by implementing more feedback-loops and include iterative model training. Future research includes applying our approach to PAM collections to generate annotated data from the weakly labelled recordings. Based on the promising results using AnuraSet, we could train BirdNet using the recordings of anuran calls and the weak annotations from museum collections such as the FNJV collection and calculate the evaluation metrics using the strong labels from AnuraSet.

Acknowledgements

This research is part of the Computational Sustainability & Technology project area[5], and has been supported by the Ministry for Science and Culture of Lower Saxony (MWK), the Endowed Chair of Applied Artificial Intelligence, Oldenburg University, and DFKI.

References

[1] Sugai, L.S.M., Silva, T.S.F., Ribeiro, J.W., Llusia, D.: Terrestrial Passive Acoustic Monitoring: Review and Perspectives. BioScience **69**(1), 15–25 (2019) https://doi.org/10.1093/biosci/biy147 . Accessed 2023-03-01

[2] Tuia, D., Kellenberger, B., Beery, S., Costelloe, B.R., Zuffi, S., Risse, B., Mathis, A., Mathis, M.W., Langevelde, F., Burghardt, T., *et al.*: Perspectives in machine learning for wildlife conservation. Nature communications **13**(1), 792 (2022)

[3] Gouvêa, T.S., Kath, H., Troshani, I., Lüers, B., Serafini, P.P., Campos, I.B., Afonso, A.S., Leandro, S.M.F.M., Swanepoel, L., Theron, N., Swemmer, A.M., Sonntag, D.: Interactive Machine Learning Solutions for Acoustic Monitoring of Animal Wildlife in Biosphere Reserves. In: Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, pp. 6405–6413. International Joint Conferences on Artificial Intelligence Organization, Macau, SAR China (2023). https://doi.org/10.24963/ijcai.2023/711 . https://www.ijcai.org/proceedings/2023/711 Accessed 2023-08-16

[4] Stowell, D.: Computational bioacoustics with deep learning: a review and roadmap. PeerJ **10**, 13152 (2022) https://doi.org/10.7717/peerj.13152 . Accessed 2023-08-01

[5] Meineke, E.K., Davies, T.J., Daru, B.H., Davis, C.C.: Biological collections for understanding biodiversity in the Anthropocene. Philosophical Transactions of the Royal Society B: Biological Sciences **374**(1763), 20170386 (2018) https://doi.org/10.1098/rstb.2017.0386 . Accessed 2023-08-01

[6] Dena, S., Rebouças, R., Augusto-Alves, G., Zornosa-Torres, C., Pontes, M.R., Toledo, L.F.: How much are we losing in not depositing anuran sound recordings in scientific collections? Bioacoustics **29**(5), 590–601 (2020) https://doi.org/10.1080/09524622.2019.1633567 . Accessed 2023-08-01

[7] Sugai, L.S.M., Llusia, D.: Bioacoustic time capsules: Using acoustic monitoring to document biodiversity. Ecological Indicators **99**, 149–152 (2019) https://doi.org/10.1016/j.ecolind.2018.12.021 . Accessed 2023-08-01

[8] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern

[5] https://cst.dfki.de/

Recognition, pp. 770–778 (2016)

[9] Simonyan, K., Zisserman, A.: Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv (2015). https://doi.org/10.48550/arXiv.1409.1556 . http://arxiv.org/abs/1409.1556 Accessed 2023-08-02

[10] Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4700–4708 (2017)

[11] Kahl, S., Wood, C.M., Eibl, M., Klinck, H.: BirdNET: A deep learning solution for avian diversity monitoring. Ecological Informatics **61**, 101236 (2021) https://doi.org/10.1016/j.ecoinf.2021.101236 . Accessed 2023-05-12

[12] Tzirakis, P., Shiarella, A., Ewers, R., Schuller, B.W.: Computer audition for continuous rainforest occupancy monitoring: the case of bornean gibbons' call detection (2020)

[13] Çakır, E., Parascandolo, G., Heittola, T., Huttunen, H., Virtanen, T.: Convolutional Recurrent Neural Networks for Polyphonic Sound Event Detection. IEEE/ACM Transactions on Audio, Speech, and Language Processing **25**(6), 1291–1303 (2017) https://doi.org/10.1109/TASLP.2017.2690575 . Conference Name: IEEE/ACM Transactions on Audio, Speech, and Language Processing

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

[15] Dufourq, E., Batist, C., Foquet, R., Durbach, I.: Passive acoustic monitoring of animal populations with transfer learning. Ecological Informatics **70**, 101688 (2022) https://doi.org/10.1016/j.ecoinf.2022.101688 . Accessed 2023-09-19

[16] Kath, H., Serafini, P.P., Campos, I.B., Gouvea, T., Sonntag, D.: Leveraging transfer learning and active learning for sound event detection in passive acoustic monitoring of wildlife. In: 3rd Annual AAAI Workshop on AI to Accelerate Science and Engineering. AAAI Workshop on AI to Accelerate Science and Engineering (AI2ASE-2024), Befindet Sich AAAI, February 26, Vancouver, BC, Canada (2024)

[17] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255 (2009). Ieee

[18] Wang, Y., Li, J., Metze, F.: 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), pp. 31–35 (2019). https://doi.org/10.1109/ICASSP.2019.

8682847 . ISSN: 2379-190X

[19] Hershey, S., Chaudhuri, S., Ellis, D.P., Gemmeke, J.F., Jansen, A., Moore, R.C., Plakal, M., Platt, D., Saurous, R.A., Seybold, B., *et al.*: Cnn architectures for large-scale audio classification. In: 2017 Ieee International Conference on Acoustics, Speech and Signal Processing (icassp), pp. 131–135 (2017). IEEE

[20] Sprengel, E., Jaggi, M., Kilcher, Y., Hofmann, T.: Audio Based Bird Species Identification using Deep Learning Techniques. LifeCLEF 2016 (2016)

[21] Gemmeke, J.F., Ellis, D.P.W., Freedman, D., Jansen, A., Lawrence, W., Moore, R.C., Plakal, M., Ritter, M.: Audio set: An ontology and human-labeled dataset for audio events. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 776–780 (2017). https://doi.org/10.1109/ICASSP.2017.7952261

[22] Kumar, A., Raj, B.: Audio Event Detection using Weakly Labeled Data. In: Proceedings of the 24th ACM International Conference on Multimedia, pp. 1038–1047 (2016). https://doi.org/10.1145/2964284.2964310 . arXiv:1605.02401 [cs]. http://arxiv.org/abs/1605.02401 Accessed 2023-09-13

[23] Xu, Y., Kong, Q., Wang, W., Plumbley, M.D.: Large-Scale Weakly Supervised Audio Classification Using Gated Convolutional Neural Network. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 121–125 (2018). https://doi.org/10.1109/ICASSP.2018.8461975 . ISSN: 2379-190X

[24] Miyazaki, K., Komatsu, T., Hayashi, T., Watanabe, S., Toda, T., Takeda, K.: Weakly-Supervised Sound Event Detection with Self-Attention. In: ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 66–70 (2020). https://doi.org/10.1109/ICASSP40776.2020.9053609 . ISSN: 2379-190X

[25] Xin, Y., Yang, D., Zou, Y.: Audio pyramid transformer with domain adaption for weakly supervised sound event detection and audio classification. In: Proc. Interspeech 2022, pp. 1546–1550 (2022)

[26] Chen, S., Wu, Y., Wang, C., Liu, S., Tompkins, D., Chen, Z., Wei, F.: BEATs: Audio Pre-Training with Acoustic Tokenizers. arXiv (2022). https://doi.org/10.48550/arXiv.2212.09058 . http://arxiv.org/abs/2212.09058 Accessed 2023-08-03

[27] Jiang, J.-j., Bu, L.-r., Duan, F.-j., Wang, X.-q., Liu, W., Sun, Z.-b., Li, C.-y.: Whistle detection and classification for whales based on convolutional neural networks. Applied Acoustics **150**, 169–178 (2019)

[28] Coffey, K.R., Marx, R.E., Neumaier, J.F.: Deepsqueak: a deep learning-based

system for detection and analysis of ultrasonic vocalizations. Neuropsychopharmacology **44**(5), 859–868 (2019)

[29] Cohen, Y., Nicholson, D.A., Sanchioni, A., Mallaber, E.K., Skidanova, V., Gardner, T.J.: Automated annotation of birdsong with a neural network that segments spectrograms. Elife **11**, 63853 (2022)

[30] Cañas, J.S., Toro-Gómez, M.P., Sugai, L.S.M., Benítez Restrepo, H.D., Rudas, J., Posso Bautista, B., Toledo, L.F., Dena, S., Domingos, A.H.R., Souza, F.L., *et al.*: A dataset for benchmarking neotropical anuran calls identification in passive acoustic monitoring. Scientific Data **10**(1), 771 (2023)

[31] Yang, Y.-Y., Hira, M., Ni, Z., Chourdia, A., Astafurov, A., Chen, C., Yeh, C.-F., Puhrsch, C., Pollack, D., Genzel, D., Greenberg, D., Yang, E.Z., Lian, J., Mahadeokar, J., Hwang, J., Chen, J., Goldsborough, P., Roy, P., Narenthiran, S., Watanabe, S., Chintala, S., Quenneville-Bélair, V., Shi, Y.: Torchaudio: Building blocks for audio and speech processing. arXiv preprint arXiv:2110.15018 (2021)

[32] Hershey, S., Chaudhuri, S., Ellis, D.P.W., Gemmeke, J.F., Jansen, A., Moore, R.C., Plakal, M., Platt, D., Saurous, R.A., Seybold, B., Slaney, M., Weiss, R.J., Wilson, K.: CNN Architectures for Large-Scale Audio Classification. arXiv (2017). https://doi.org/10.48550/arXiv.1609.09430 . http://arxiv.org/abs/1609.09430 Accessed 2023-08-11

[33] Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)

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

[35] Troshani, I., Gouvea, T., Sonntag, D.: Leveraging sound collections for animal species classification with weakly supervised learning. In: 3rd Annual AAAI Workshop on AI to Accelerate Science and Engineering. AAAI Workshop on AI to Accelerate Science and Engineering (AI2ASE-2024), AAAI, Vancouver, Canada (2024)

[36] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: Pytorch: An imperative style, high-performance deep learning library. In: Advances in Neural Information Processing Systems 32, pp. 8024–8035. Curran Associates, Inc., ??? (2019). http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

[37] Shah, A., Kumar, A., Hauptmann, A.G., Raj, B.: A closer look at weak label learning for audio events. CoRR **abs/1804.09288** (2018) 1804.09288