Towards Deep Active Learning in Avian Bioacoustics

Lukas **Rauch**[1], Denis **Huseljic**[1], Moritz **Wirth**[1], Jens **Decke**[1], Bernhard **Sick**[1] and Christoph **Scholz**[1]

[1]*IES, University of Kassel, Kassel, Germany*
[2]*IEE, Fraunhofer Insitute, Kassel, Germany*

Abstract

Passive acoustic monitoring (PAM) in avian bioacoustics enables cost-effective and extensive data collection with minimal disruption to natural habitats. Despite advancements in computational avian bioacoustics, deep learning models continue to encounter challenges in adapting to diverse environments in practical PAM scenarios. This is primarily due to the scarcity of annotations, which requires labor-intensive efforts from human experts. Active learning (AL) reduces annotation cost and speed ups adaption to diverse scenarios by querying the most informative instances for labeling. This paper outlines a deep AL approach, introduces key challenges, and conducts a small-scale pilot study.

Keywords

Deep Active Learning, Avian Bioacoustics, Passive Acoustic Monitoring

1. Introduction

Avian diversity is a key indicator of environmental health. Passive acoustic monitoring (PAM) in avian bioacoustics leverages mobile autonomous recording units (ARUs) to gather large volumes of soundscape recordings with minimal disruption to avian habitats. While this method is cost-effective and minimally invasive, the analysis of these recordings is labor-intensive and requires expert annotation. Recent advancements in deep learning (DL) primarily process these passive recordings by classifying bird vocalizations. Particularly, feature embeddings from large bird sound classification models (e.g., Google's Perch [1] or BirdNET [2]) have effectively enabled few-shot learning in scenarios with limited training data [3]. These state-of-the-art (SOTA) models are trained using supervised learning on nearly 10,000 bird species from multi-class focal recordings that isolate individual bird sounds. However, practical PAM scenarios involve processing diverse multi-label soundscapes with overlapping sounds and varying background noise. Proper feature embeddings for edge deployment necessitate fine-tuning, which relies on labeled training data that is both time-consuming and costly to obtain for soundscapes.

Deep active learning (AL) addresses this challenge by actively querying the most informative instances to maximize performance gains [4]. However, research on deep AL in avian bioacoustics is still limited, and the problem needs to be contextualized with comparable datasets [5]. Additionally, the domain presents unique practical challenges, including adapting models from focals to soundscapes (i.e., multi-class to multi-label) in imbalanced and highly diverse scenarios [6]. Consequently, we introduce the problem of deep AL in avian bioacoustics and propose an efficient fine-tuning approach for model deployment. Our contributions are:

> **Contributions**
>
> 1. We introduce deep active learning (AL) to avian bioacoustics, highlighting challenges and proposing a practical framework.
> 2. We conduct an initial feasibility study based on the dataset collection `Birdset` [6], showcasing the benefits of deep AL. Additionally, we release the dataset and code.

2. Related Work

DL has enhanced bird species recognition from vocalizations in the context of biodiversity monitoring. Current SOTA approaches BirdNET [2], Google's Perch [7, 1], and `BirdSet` [6] have set benchmarks in bird sound classification. While initial studies focused on model performance on focal recordings, research is increasingly shifting towards practical PAM scenarios [6]. In such environments, ARUs are proving effective for edge deployment for continuous soundscape analysis [8]. Research indicates that pre-trained models facilitate few-shot and transfer learning in data-scarce environments by providing valuable feature embeddings for rapid prototyping and efficient inference [3]. While deep AL is suited for quick model adaptation, its application in avian bioacoustics is still emerging. Bellafkir et al. [9] have integrated AL into edge-based systems for bird species identification, employing reliability scores and ensemble predictions to refine misclassifications through human feedback. This approach highlights the necessity for research into the application of deep AL and multi-label classification in avian bioacoustics. However, comparing these results is challenging because they utilize test datasets that are not publicly available and employ custom AL strategies [9].

3. Active Learning in Bird Sound Classification

Motivation. In PAM, a feature vector $\mathbf{x} \in \mathcal{X}$ represents a D-dimensional instance, originating from either a focal recording where $\mathcal{X} = \mathcal{F}$, or a soundscape recording with $\mathcal{X} = \mathcal{S}$. **Focal recordings** are extensively available on the citizen-science platform Xeno-Canto (XC) [10] with a global collection of over 800,000 recordings, making them particularly suitable for model training. Large-scale bird sound classification models (e.g., BirdNET[2]) are primarily trained on focals. These multi-class recordings feature isolated bird vocalizations where each instance $\mathbf{x}$ is associated with a class label $y \in \mathcal{Y}$, where $\mathcal{Y} = \{1, ..., C\}$. The focal data distribution is denoted as $p_{\text{Focal}}(\mathbf{x}, y)$. However, annotations from XC often come with weak labels, lacking precise vocalization timestamps. As noted by Van Merriënboer et al. [11], evaluating on focals does not adequately reflect a model's generalization performance in real-world PAM scenarios, rendering them unsuitable for assessing deployment capabilities. **Soundscape recordings** are passively recorded in specific regions, capturing the entire acoustic environment for PAM projects using static ARUs over extended periods. For instance, the High Sierra Nevada (HSN) [2] dataset includes long-duration soundscapes with precise labels and timestamps from multiple recording sites. Soundscapes are treated as multi-label tasks and are valuable for assessing model deployment in real-world PAM. Each instance $\mathbf{x}$ is associated with multiple class labels $y \in \mathcal{Y}$, represented by a one-hot encoded multi-label vector $\mathbf{y} = [y_1, \ldots, y_C] \in [0, 1]^C$. An instance can contain no bird sounds, represented by a zero-vector $\mathbf{y} = \mathbf{0} \in \mathbb{R}^C$. Soundscapes' limited scale and the extensive annotation effort make them less suitable for large-scale model training. We denote the soundscape data distribution as $p_{\text{Scape}}(\mathbf{x}, \mathbf{y})$. The disparity in data distributions, $p_{\text{Scape}}(\mathbf{x}, \mathbf{y}) \neq p_{\text{Focal}}(\mathbf{x}, y)$, leads to a distribution shift that impacts the performance of SOTA bioacoustic models trained on focals when deployed in PAM. Additionally, highly diverse deployment conditions in PAM projects - such as background noise, recording devices, and their locations - also lead to domain differences within and between soundscape recordings. These variations further highlight the need for compact models that can quickly and easily adapt to changing environments. Thus, we argue that using labeled soundscapes in novel deployment scenarios for fine-tuning the model is vital. Therefore, we propose deep AL to enable fast model adaption to various PAM scenarios.

Our approach. Our approach is detailed in Figure 1. We leverage the `BirdSet` dataset collection [6] to ensure comparability. We consider a multi-label classification problem, where we equip a model with a pre-trained feature extractor $\mathbf{h}_{\boldsymbol{\omega}} : \mathcal{X} \to \mathbb{R}^D$ with parameters $\boldsymbol{\omega}$ that maps the inputs $\mathbf{x}$ to feature embeddings $\mathbf{h}_{\boldsymbol{\omega}}(\mathbf{x})$. Additionally, we utilize a classification head $\mathbf{f}_{\boldsymbol{\theta}_t} : \mathbb{R}^D \to \mathbb{R}^C$ with parameters $\boldsymbol{\theta}_t$ at cycle iteration t that maps the feature embeddings $\mathbf{h}_{\boldsymbol{\omega}}(\mathbf{x})$ to class probabilities via the sigmoid function. The resulting class probabilities are denoted by $\hat{\mathbf{p}} = \sigma(\mathbf{f}_{\boldsymbol{\theta}_t}(\mathbf{h}_{\boldsymbol{\omega}}(\mathbf{x})))$, where $\hat{\mathbf{p}} \in \mathbb{R}^C$ represents the probabilities for each class in a binary classification problem. We introduce a pool-based AL setting
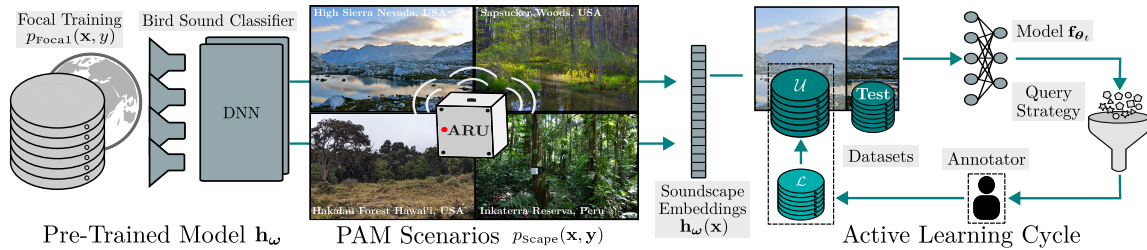
Figure 1: Proposed deep AL cycle in avian bioacoustics with exemplary tasks from `BirdSet`[6].

with an unlabeled pool $\mathcal{U}(t) \subseteq \mathcal{S}$ and a labeled pool data set $\mathcal{L}(t) \subseteq \mathcal{S} \times \mathcal{Y}$. The pool consists of soundscapes from PAM projects, allowing the model to adapt to the unique acoustic features of new sites and improve performance across various scenarios. During each cycle iteration t, the query strategy compiles the most informative instances into a batch $\mathcal{B}(t) \subset \mathcal{U}(t)$ of size b. We represent an annotated batch as $\mathcal{B}^*(t) \in \mathcal{S} \times \mathcal{Y}$. We update the unlabeled pool $\mathcal{U}(t+1) = \mathcal{U}(t) \setminus \mathcal{B}(t)$ and the labeled pool $\mathcal{L}(t+1) = \mathcal{L}(t) \cup \mathcal{B}^*(t)$ by adding the annotated batch. At each iteration t, the model $\boldsymbol{\theta}_t$ is retrained using the binary cross entropy loss $L_{BCE}(\mathbf{x}, \mathbf{y})$, resulting in the updated model parameters $\boldsymbol{\theta}_{t+1}$. The process continues until a budget B is exhausted.

4. Experiments

Setup. We employ Google's Perch as the pre-trained feature extractor with a feature dimensionality of $D = 1280$, following Ghani et al. [3]. Each iteration of the AL cycle involves initializing and training the last DNN layer for 200 epochs using the Rectified Adam optimizer [12] (batch size: 128, learning rate: 0.05, weight decay: 0.0001) with a cosine annealing scheduler [13]. The hyperparameters are empirically determined with convergence on random train samples as done in [14]. We utilize the HSN dataset [15] from `BirdSet` [6], consisting of $5,280$ 5-second soundscape segments from the initial three days of recordings for our unlabeled pool. Thus, we simulate practical deployment scenario where we initially collect data from various recording sites that we want to quickly adapt the model to and reduce annotation effort. Subsequently, we utilize $6,720$ segments from the last two days for testing model performance. Initially, 10 instances are selected randomly, followed by 50 iterations of $b{=}10$ acquisitions each, totaling a budget of $B{=}510$. We benchmark against `Random` acquisitions and use `Typiclust` [16] and `Badge`[17] as diversity-based and hybrid strategies, respectively. As an uncertainty-based strategy, we employ the mean `Entropy` of all binary predictions. The effectiveness of each strategy is assessed by analyzing the learning curves through a collection of threshold-free metrics [6]: T1-accuracy, class-based mean average precision (cmAP), and area under the receiver operating characteristic curve (AUROC). The metrics are computed on the test dataset post-training in each cycle, with learning curve improvements averaged over ten repetitions for consistency.

Results. We present the improvement curves for the metric collection in Figure 2. The results demonstrate that no single strategy is universally superior across all metrics. However, nearly all metrics show enhanced performance compared to `Random`. Notably, `Typiclust` displays strong performance across all metrics at the start of the deep AL cycle, supporting the findings of [16] that a diverse selection is beneficial at the cycle's onset. However, its effectiveness diminishes over time when diversity becomes less crucial. Conversely, except for the AUROC metric where `Entropy` initially performs poorly but strongly improves over time, `Entropy` outperforms in all iterations for cmAP and T1-Acc, showing a consistent improvement over `Random` of up to 15%.

5. Open Challenges and Limitations

This pilot study explores the use of deep AL to tailor avian bioacoustic models for various deployment scenarios in PAM. Although the initial results are encouraging, they remain preliminary. Several key
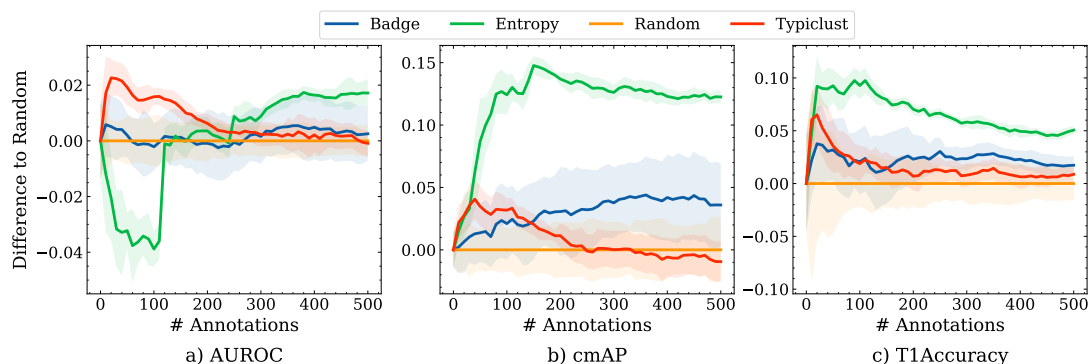
Figure 2: Improvement curves of deep AL selection strategies `Badge`, `Entropy`, and `Typiclust` over `Random` with the metric collection a) AUROC, b) cmAP and c) T1-Acc. The results are averaged over ten randomly initialized repetitions to ensure consistency and the standard deviation is displayed.

challenges, which are outlined below, need to be addressed to fully realize the potential of deep AL in this field.

Pool creation. The limited availability of soundscape data, which is primarily used for model evaluation [6], poses challenges in creating pool datasets for deep AL. The process of generating a fine-tuning training pool can affect class balance and raises concerns about the composition methodology. Additionally, in scenarios where data are sourced from PAM projects, the variability in recording sites is often not disclosed in publicly available datasets. This lack of information makes it challenging to create a diverse and representative training pool that takes recording locations into account. To effectively investigate deep AL, a transparent approach to dataset generation is essential.

Deployment in practice. Deploying deep AL in real-world PAM environments requires addressing several practical considerations. These include determining optimal batch sizes for data annotation and effectively allocating the total budget. The labor-intensive and costly process of labeling PAM recordings, which requires human expertise [18], highlights the need for accurately estimating the expected annotation effort. Additionally, exploring various deployment settings and tasks can reveal the versatility and potential challenges of applying deep AL, leading to more effective and scalable solutions for avian bioacoustics. For instance, tasks might involve not only classifying bird species but also identifying specific call densities [19], which would require modifications to the model evaluation process.

Evaluation. Traditional metrics such as AUROC, cmAP, and T1-Acc offer a general overview of model performance but may be inadequate in practice-specific scenarios, such as ensuring a high recall of a specific species or identifying bird call density [19]. A more nuanced approach to evaluating deep AL models involves customizing metrics to align with practical objectives, such as consistently identifying specific species. Enhancing evaluation methodologies to capture these specialized requirements is crucial for advancing the effectiveness of deep AL in real-world PAM applications.

6. Conclusion

In this work, we demonstrated the potential of deep active learning (AL) in computational avian bioacoustics. We showed how deep AL can be integrated into real-world passive acoustic monitoring by utilizing `BirdSet`, where a rapid model adaption through fine-tuning on soundscape recordings is advantageous for the identification of bird species. Our results indicate that employing selection strategies in deep AL enhances model performance and accelerates adaptation compared to random sampling. For future work, we aim to expand the implementation of deep AL in avian bioacoustics utilizing all datasets from the `BirdSet` dataset collection to provide more robust performance insights and explore additional query strategies [13, 20].

References

[1] J. Hamer, E. Triantafillou, B. Van Merriënboer, S. Kahl, H. Klinck, T. Denton, V. Dumoulin, BIRB: A Generalization Benchmark for Information Retrieval in Bioacoustics, CoRR (2023). URL: https://doi.org/10.48550/arXiv.2312.07439.

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

[3] B. Ghani, T. Denton, S. Kahl, H. Klinck, Feature Embeddings from Large-Scale Acoustic Bird Classifiers Enable Few-Shot Transfer Learning, CoRR (2023). doi:https://10.48550/arXiv.2307.06292.

[4] J. Decke, C. Gruhl, L. Rauch, B. Sick, DADO -- Low-cost query strategies for deep active design optimization, in: 2023 International Conference on Machine Learning and Applications (ICMLA), IEEE, 2023, pp. 1611–1618.

[5] L. Rauch, R. Schwinger, M. Wirth, B. Sick, S. Tomforde, C. Scholz, Active Bird2Vec: Towards End-to-End Bird Sound Monitoring with Transformers, CoRR (2023). URL: https://doi.org/10.48550/arXiv.2308.07121.

[6] L. Rauch, R. Schwinger, M. Wirth, R. Heinrich, D. Huseljic, J. Lange, S. Kahl, B. Sick, S. Tomforde, C. Scholz, Birdset: A dataset and benchmark for classification in avian bioacoustics, CoRR (2024). doi:https://10.48550/arXiv.2403.10380.

[7] T. Denton, S. Wisdom, J. R. Hershey, Improving Bird Classification with Unsupervised Sound Separation, in: ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2022, pp. 636–640. URL: https://doi.org/10.1109/ICASSP43922.2022.9747202.

[8] J. Höchst, H. Bellafkir, P. Lampe, M. Vogelbacher, M. Mühling, D. Schneider, K. Lindner, S. Rösner, D. G. Schabo, N. Farwig, B. Freisleben, Bird@Edge: Bird Species Recognition at the Edge, in: Networked Systems, volume 13464, Cham, 2022, pp. 69–86. URL: https://doi.org/10.1007/978-3-031-17436-0_6.

[9] H. Bellafkir, M. Vogelbacher, D. Schneider, M. Mühling, N. Korfhage, B. Freisleben, Edge-Based Bird Species Recognition via Active Learning, in: Networked Systems, volume 14067, Springer Nature Switzerland, Cham, 2023, pp. 17–34. doi:10.1007/978-3-031-37765-5_2.

[10] W. Vellinga, R. Planqué, The xeno-canto collection and its relation to sound recognition and classification, CEUR-WS.org, 2015. URL: https://xeno-canto.org/.

[11] B. Van Merriënboer, J. Hamer, V. Dumoulin, E. Triantafillou, T. Denton, Birds, Bats and beyond: Evaluating generalization in bioacoustic models, CoRR (2024).

[12] L. Liu, H. Jiang, P. He, W. Chen, X. Liu, J. Gao, J. Han, On the variance of the adaptive learning rate and beyond, in: International Conference on Learning Representations, 2019.

[13] D. Huseljic, P. Hahn, M. Herde, L. Rauch, B. Sick, Fast fishing: Approximating bait for efficient and scalable deep active image classification, CoRR (2024). doi:https://10.48550/arXiv.2404.08981.

[14] D. Huseljic, M. Herde, P. Hahn, B. Sick, Role of hyperparameters in deep active learning, in: Workshop on Interactive Adaptive Learning @ ECML PKDD, 2023, pp. 19–24.

[15] S. Kahl, C. M. Wood, P. Chaon, M. Z. Peery, H. Klinck, A collection of fully-annotated soundscape recordings from the western united states, 2022. URL: https://doi.org/10.5281/zenodo.7050014.

[16] G. Hacohen, A. Dekel, D. Weinshall, Active learning on a budget: Opposite strategies suit high and low budgets, in: International Conference on Machine Learning, 2022.

[17] J. T. Ash, C. Zhang, A. Krishnamurthy, J. Langford, A. Agarwal, Deep batch active learning by diverse, uncertain gradient lower bounds, in: International Conference on Learning Representations, 2020.

[18] D. Stowell, Computational bioacoustics with deep learning: A review and roadmap, CoRR (2021). URL: https://doi.org/10.48550/arXiv.2112.06725.

[19] A. K. Navine, T. Denton, M. J. Weldy, P. J. Hart, All thresholds barred: Direct estimation of

call density in bioacoustic data, Frontiers in Bird Science 3 (2024). doi:`10.3389/fbirs.2024.1380636`.

[20] L. Rauch, M. Aßenmacher, D. Huseljic, M. Wirth, B. Bischl, B. Sick, Activeglae: A benchmark for deep active learning with transformers, in: Machine Learning and Knowledge Discovery in Databases: Research Track, Springer Nature Switzerland, 2023, p. 55–74. URL: https://doi.org/10.1007/978-3-031-43412-9_4.