ADAPTABLE INPUT LENGTH USING MODEL TRAINED ON WAVEFORM

Technical Report

Valentin Bordoux[*]

[*]Marine Animal Ecology, Wageningen University
Valentin.bordoux@wur.nl

ABSTRACT

This report presents a method for bioacoustic sound event detection using few-shot learning, developed for the DCASE 2024 Task 5. Our approach experiments with pretrained models that take waveforms as input. These models serve as feature extractors, and prototypical loss is used for prediction. Initially, we employed direct predictions with openly available pretrained models. Subsequently, we attempted to fine-tune the models for each file, using only the first five annotations as training set. The direct prediction system achieved 40% F-measure score, 12 points under the baseline system proposed by the organizers. Fine-tuning did not improve the model's performance over direct prediction.

While our proposed method can be applied directly without extensive parameter tuning or additional training, the results indicate that it does not achieve the generalizability required for this challenge when compared to the baseline method. This work suggests how state-of-the-art models, despite their high performance on other datasets or benchmarks, may still perform suboptimal on sound event detection using few-shot learning for certain taxonomic groups present in the DCASE challenge datasets.

1. INTRODUCTION

With the development of automatic recorders, collecting vast amounts of acoustic data has become easier than ever. However, processing this data remains challenging, particularly in novel or less-studied environments and taxonomic groups [1], [2].

For the DCASE Challenge Task 5, a few-shot learning framework has been adapted to perform sound event detection (SED) across a diverse array of tasks [3]. The goal is to enable rapid detection in new tasks with minimal human effort, requiring annotation of only the first five sounds in a file. To excel in this task, the method must effectively detect sounds from a wide variety of taxonomic groups and environments.

A new baseline system was proposed this year [4], building on the knowledge of previous challenges. The authors pinpoint domain shift as one of the main challenges. A part of this difficulty involves developing a system capable of detecting calls of varying durations, ranging from approximately 100 milliseconds to several seconds. Additionally a common challenge in sound classification or detection is the choice of input features for the system, often involving different spectrogram representations. While some features typically perform better than other, tests are often needed to select the best features for the specific case study [5].

To address challenges related to sound length variation, feature selection, and domain shift, we explored the potential of open access pretrained systems that perform detection or classification directly from waveforms. Such systems have demonstrated state-of-the-art performance on various bioacoustic datasets and benchmarks. Specifically, we tested the applicability of the AVES model [6] for SED using five-shot learning on the DCASE Challenge 2024 Task 5 data.

2. METHOD

The code is available on GitHub [7] with details of the hyperparameters used for training.

2.1. Models

AVES is a model based on wav2wec2 architecture, a transformer based auto-encoder pretrained on a combination of publicly available audio sets. Thanks to the technique of tokenization, AVES is able to use input of different size. The model was used as feature extractor using the version self-trained on bioacoustics data available online [8].

Another recently published model was tested, BioLingual, based on the CLAP-LAION architecture [9], which outperformed AVES on benchmark datasets. Unfortunately due to time and computation power constraint, we were unable to test the performance of BioLingual on the full validation and test set.

2.2. Direct Prediction

Each file is loaded and resampled to 16 kHz, the frequency used for training the AVES model. Figure 1 describes the workflow for direct prediction, which involves the following steps:

The mean duration of the first five positive annotations, denoted as d_{pos} , is computed and clipped between 25ms (the minimal

input length for AVES) and 1 second. This clipping provides better granularity of prediction for longer annotations and was determined experimentally. The positive annotated segments are concatenated and divided by d_{pos} . The negative segments - the time between annotations - is concatenated and divided in segments of d_{pos} length. The query time is then divided using a sliding window approach of d_{pos} duration without overlap (1).

The feature vector corresponding to each positive, negative, and query segment is computed by the model (2).

Feature vectors of positive and negative segments are averaged to obtain the positive and negative prototypes, respectively (3).

The Euclidean distance between the feature vector of each query segment and the prototypes is evaluated, and the prediction of a class is assigned based on the prototype closest to the query (4).

2.3. Fine tuning on 5 samples

To avoid reducing the generalization capability of the model by re-training it on the DCASE data, we attempted to fine-tune the model for each file before predicting the query set. This fine-tuning used only the first five annotations as positive samples and the intervals between annotations as negative samples. The experiments were conducted on the validation set, using the query time as validation dataset and the annotated time as training dataset, to determine an appropriate fine-tuning approach. For the test set, the model was trained on the annotated time without validation steps, then used to predict the query time. An early stop condition based on the training loss was employed to save computation time and prevent overfitting. After training, a similar workflow to "Direct prediction" was used, but with the newly fine-tuned model instead of the pretrained model.

2.4. Loss function

Several loss functions were tested during development:

- Binary cross-entropy on prototypical loss.
- Supervised contrastive learning (SCL) [10].
- Fusion loss combining SCL and a custom loss based on the distance between prototypes.

Training on SCL only often led to have the prototype converge towards each other in the feature space. Based on this observation, a custom loss was created, combining the SCL with a loss proportional to the inverse of the Euclidean distance between the prototypes.

2.5. Data augmentation

For the SCL, the data augmentation was implemented using the torch implementation of Audiomentation [11]. A pipeline was created using gain, pitch, shift and white noise addition operations to create augmented version of the samples for each batch.

2.6. Post-processing

To refine the predictions, the simple post processing approach implemented in [4] was used, removing all the predictions smaller than a fraction of d_{pos} with a threshold of 0.7.

3. RESULTS

Table 1 presents the results of the different approaches on the validation set. Baseline is the system proposed for the DCASE challenge 2024 task 5, which is based on prototypical network using negative hard sampling. "Direct prediction" refers to the system described in Section 2.2 including the post-processing steps outlined in Section 2.6. The "fine-tuned system" is described in Section 2.3, and uses the fusion loss, data augmentation, and post-processing methods detailed in Sections 2.4, 2.5, and 2.6, respectively.

System	F-measure	Precision	Recall
0. Baseline	**52.14**	**56.18**	48.64
1. Direct prediction	39.28	36.64	51.09
2. Fine-tuned	29.34	24.75	**54.79**

Table 1: performance of the proposed system and the baseline based on the DCASE task 2024 Validation set.

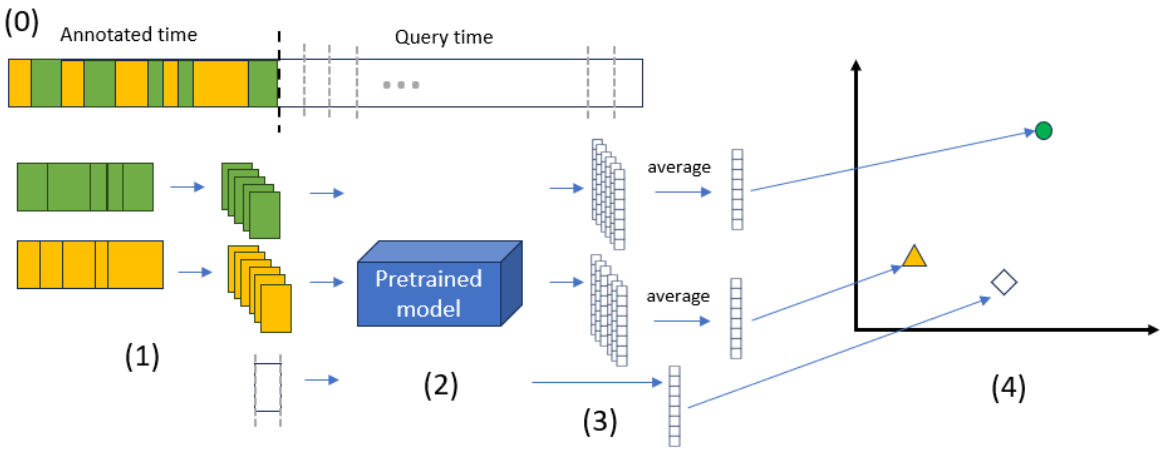

Figure 1: Direct prediction workflow. Audio file are initially split in annotated time, ending on the last positive annotation, and query time (0). The positive annotated segments (green) are concatenated and cut in segment of similar length equal to the average duration of the positive annotation d_{pos}. The negative segments (yellow) follow the same process using the same duration d_{pos}. The query time is divided in segments of d_{pos} length (1). Features corresponding to the segment are extracted by the pretrained model (2). Positive and negative prototypes are computed by averaging the feature vectors of their corresponding segments (3). In (4), the latent space is simplified to 2D for visualization. The Euclidean distance between query features and prototypes is computed and the predicted class for each segment corresponds to the closest prototype.

Results per dataset show significant differences, as illustrated in Figure 2.

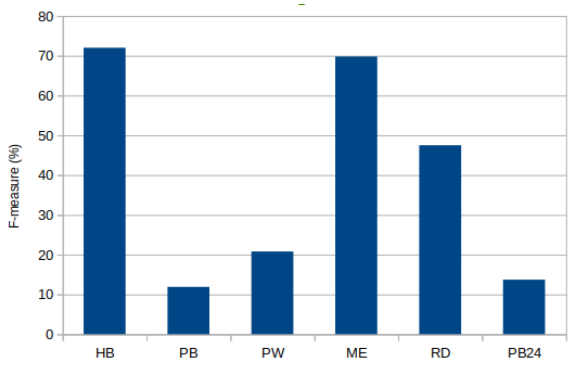

Figure 2: F-measure (%) of "Direct prediction" per dataset

Results based on fine-tuning using prototypical loss or SCL were very low in the preliminary tests, and are not included in this report. Overall, fine-tuning on the first five samples did not improve the performance of the predictions. Upon closer inspection, it was observed that while recall rates improved slightly for some files, precision rates decreased substantially.

Our preliminary tests indicated that BioLingual performed better than AVES on certain files, although requiring significantly more computation time. Due to time constraints the applicability of BioLingual was not further explored and the results are not included in this report.

4. DISCUSSION

The "Direct prediction" system outperformed the fine tuning system, but performed less than the baseline system.

Despite offering simplicity, directly using waveforms sacrifices an opportunity to enhance signals before they are used in a model. Waveforms are also more challenging for humans to interpret compared to spectrograms, making enhancement techniques and data augmentation more difficult to implement. Additionally, data augmentation methods are generally more advanced for spectrograms than for waveforms.

On the validation set, performance was particularly low for the PB and PB24 datasets, suggesting that the model struggles to interpret the types of calls and background noises present in these datasets. One explanation could be that AVES was not trained on data that provides a good representation of such calls. Additional training might improve performance due to the model's large number of parameters. However, the complexity of the model could explain why fine-tuning for each file did not work effectively. A larger dataset or fewer parameters might be required to avoid overfitting and achieve significant improvements with fine-tuning.

Another limitation arises from the loss function used during training, which requires a balanced dataset. This constraint leads to discarding of valuable annotated data, which is already scarce for this task. We hypothesize that using a loss function capable of handling data imbalance, such as Focal Loss [12], could enhance performance.

Further improvement could be inspired by the baseline system through the use of hard sampling techniques, which are compatible with the proposed system and have shown to substantially increase performance for the baseline system.

5. CONCLUSION

While models based on complex architectures such as AVES and BioLingual have demonstrated state-of-the-art performance in various classification and detection tasks in bioacoustics, this study suggests that they might not be the most suitable for addressing the DCASE Challenge Task 5. The proposed method has shown satisfactory performance on certain datasets and is characterized by its ease of use and straightforward implementation, without the need for parameter tuning or additional training. However, further research is necessary to understand and mitigate the limitations observed on some datasets.

We propose that incorporating negative hard sampling or additional training could be a promising direction for the "Direct Prediction" system. Additionally, we suggest that employing a more appropriate loss function might enhance the performance of the fine-tuning system, although this method could be incompatible with large and complex architectures. Finally, more research is required to evaluate the potential of other waveform-based models, such as BioLingual, in performing this task effectively.

6. REFERENCES

[1] Lamont, T. A. C. *et al.* The sound of recovery: Coral reef restoration success is detectable in the soundscape. *Journal of Applied Ecology* **59**, 742–756 (2022).

[2] Gibb, R., Browning, E., Glover‐Kapfer, P. & Jones, K.E. Emerging opportunities and challenges for passive acoustics in ecological assessment and monitoring. Methods Ecol. Evol. 10, 169185 (2019)

[3] Nolasco, I. *et al.* Learning to detect an animal sound from five examples. *Ecological Informatics* **77**, 102258 (2023).

[4] Liang, J. *et al.* Mind the Domain Gap: a Systematic Analysis on Bioacoustic Sound Event Detection. Preprint at https://doi.org/10.48550/arXiv.2403.18638 (2024).

[5] Stowell, D. Computational bioacoustics with deep learning: a review and roadmap. *PeerJ* **10**, e13152 (2022).

[6] Hagiwara, M. AVES: Animal Vocalization Encoder based on Self-Supervision. in Proc. International Conference on Acoustics, Speech, and Signal Processing 1–5, (2022).

[7] https://github.com/vbordoux/dcase_task5_Bordoux_WUR

[8] https://github.com/earthspecies/aves

[9] Robinson, D., Robinson, A. & Akrapongpisak, L. Transferable Models for Bioacoustics with Human Language Supervision. in *ICASSP 2024 - 2024 IEEE* 1316–1320 (2024).

[10] Khosla, P. *et al.* Supervised Contrastive Learning. Preprint at http://arxiv.org/abs/2004.11362 (2021).

[11] https://github.com/asteroid-team/torch-audiomentations

[12] Lin, T.-Y., Goyal, P., Girshick, R., He, K. & Dollár, P. Focal Loss for Dense Object Detection. Preprint at https://doi.org/10.48550/arXiv.1708.02002 (2018).