Regularized Contrastive Pre-Training for Few-Shot Bioacoustic Sound Detection

Ilyass Moummad, Nicolas Farrugia, Romain Serizel

REGULARIZED CONTRASTIVE PRE-TRAINING FOR FEW-SHOT BIOACOUSTIC SOUND DETECTION

Ilyass Moummad, Nicolas Farrugia

IMT Atlantique, Lab-STICC,
UMR CNRS 6285, Brest, France,

Romain Serizel

University of Lorraine, CNRS, Inria,
Loria, 54000, Nancy, France,

ABSTRACT

Bioacoustic sound event detection allows for better understanding of animal behavior and for better monitoring biodiversity using audio. Deep learning systems can help achieve this goal. However, it is difficult to acquire sufficient annotated data to train these systems from scratch. To address this limitation, the Detection and Classification of Acoustic Scenes and Events (DCASE) community has recasted the problem within the framework of few-shot learning and organize an annual challenge for learning to detect animal sounds from only five annotated examples. In our study, we introduce a regularization to supervised contrastive loss, to learn non redundant features that exhibit effective transferability to few-shot tasks involving the detection of animal sounds not encountered during the training phase. Our method achieves a high F-score of 61.52%$\pm$0.48 when no feature adaptation is applied, and an F-score of 68.19%$\pm$0.75 when we further adapt the learned features for each new target task. This work aims to lower the entry bar to few-shot bioacoustic sound event detection by proposing a simple and yet effective framework for this task, and by providing open-source code.[1]

***Index Terms*—** Supervised contrastive learning, total coding rate, transfer learning, few-shot learning, bioacoustics, sound event detection.

1. INTRODUCTION

Bioacoustics delve into the study of sound production, emission, reception, and processing in living organisms. This diverse domain encompasses a wide range of research, from understanding the vocalizations of marine life to deciphering the intricate communication patterns of various animal species. Given the abundance and complexity of acoustic data in bioacoustics, the application of deep learning techniques has emerged as a powerful approach to extract meaningful insights from this soundscape [1].

Despite the considerable successes of deep learning in bioacoustics, there exists a significant challenge that hinders its widespread applicability – the scarcity of labeled data [1]. Annotating acoustic data is a laborious and time-consuming task that requires expertise in the understanding of the species. Consequently, available labeled bioacoustic datasets are often limited in size, impeding the full potential of data-hungry deep learning models. It is in this context that "few-shot bioacoustics" emerges as a promising area of research [2].

Few-shot learning (FSL) is a subfield of machine learning that aims to train models using only a limited number of labeled examples. In the context of bioacoustics, this translates to developing robust and effective deep learning models that can generalize from a small number of annotated recordings, alleviating the data scarcity challenge. By harnessing few-shot learning techniques, researchers can circumvent the need for massive labeled datasets, making bioacoustic analyses more feasible for lesser-known species or habitats where extensive annotated data is lacking.

While FSL offers a compelling solution to mitigate the data scarcity challenges in bioacoustics, the effectiveness of these models heavily relies on the quality of the learned representations. In this context, representation learning plays a pivotal role in shaping the success of FSL-based approaches. A good starting initialization is crucial for FSL, and this is where representation learning techniques, like contrastive learning (CL) [3], come into play.

CL is a learning paradigm designed to learn a metric space where similar samples are pulled together while dissimilar samples are pushed apart. CL has been widely used in the litterature and has shown promising results in audio representation learning [4]. However, CL can have the dimensional collapse phenomenon, where embedding vectors collapse along certain dimensions, thus only spanning a lower-dimensional subspace [5].

We propose a system that learns good intialization for FSL using supervised contrastive pre-training. To remedy the dimensional collapse of CL, we constrain the learned features to be diverse and non-redundant, using a regularization from information theory literature [6]. Our goal is to learn features that are discriminative, ideally features that can cover a space of the largest possible dimension [6].

[1] https://github.com/ilyassmoummad/RCL_FS_BSED

This work is co-funded by the AI@IMT program of the ANR (French National Research Agency) and the company OSO-AI.

We apply the above pre-training strategy to train a general feature extractor for bioacoustic few-shot sound event detection (BSED). At inference, the feature extractor is either used directly for fast inference or fine-tuned for each binary validation task, specific to each audio file, for to the presence or absence of the event of interest, utilizing a prototypical loss. To make predictions, we slide a window over the audio file and compute an euclidean distance between the representations of each query window and the two prototypes (computed by averaging the representation of the annotated segments of presence/absence of the event of interest). We demonstrate the effectiveness of our approach on the diverse bioacoustic validation datasets of the DCASE challenge, showcasing its ability to achieve remarkable performance on the few-shot setting.

This work builds upon our previous work [7], where we pre-trained a feature extractor using CL and then trained a linear classifier on the available shots. While this system was the second best one in the challenge, the training of linear classifier using cross-entropy resulted in instability in some validation runs due to the large imbalance between the segments for the presence and absence of an event. Here, we replace the cross-entropy classification with a robust metric approach that is more stable and that optionally adapts the features to the task at hand. Additionally, we further enhance the pre-training stage by regularizing the learned representations.

2. RELATED WORK

The DCASE community propose a benchmark for BSED that consists in detecting animal vocalizations in audio recordings given only five annotated examples [2]. Liu et al. [8] use prototypical networks on the concatenation of per-channel energy normalization and delta mel-frequency cepstral coefficients, and trained on extra animal data from AudioSet [9] to increase generalization. Tang et al. [10] use a frame-level approach using semi-supervised learning to exploit unlabeled query data. Our previous work [7] shows the strong performance of supervised contrastive pre-training followed by cross-entropy linear classification. Yan et al. [11] improve over their previous work [10] by adding target speaker voice activity detection to form a multi-task frame-level system, and by adding a transformer encoder in their model architecture.

MetaAudio [12] is a few-shot audio classification benchmark with diverse audio types (including bioacoustics). Our work doesn't address classification and reserves it for future research. BirdNet [13], a deep learning system trained on diverse data sources to identify 984 bird species, and Google Perch, another model trained on an extensive bird corpus, have shown superior transferability for few-shot bioacoustic classification tasks when compared to models trained on generic audio datasets such as AudioSet [9], as demonstrated by Ghani et al. [14].

https://tfhub.dev/google/bird-vocalization-classifier/4

The litterature of representation learning has shown great transfer performance thanks to CL [3, 15, 4]. Regularized methods constrain the embeddings to have non-redundant information by measuring the cross-correlation between the representations of two views [16], decorrelating the feature variables from each other [17], or by maximizing the total coding rate of the features [18, 6]. The combination of contrastive and regularized methods has not been yet explored. We investigate them in the context of transfer learning for few-shot bioacoustic sound event detection.

3. METHOD

In this section we describe the methodology employed in our study (Fig. 1). We train a feature extractor on a general, labeled training set using supervised contrastive learning (SCL) combined with a coding rate regularization that constrains the embeddings to be non-redundant. The resulting trained model is transferred to the validation sets and optionally fine-tuned on the available shots using a prototypical loss. The predictions are made by computing the distances to the positive and negative prototypes, for the presence and absence of sound events of interest, respectively.

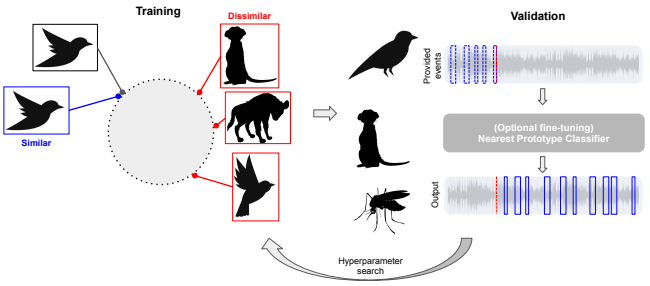

Fig. 1. Overview of our approach: Supervised contrastive pre-training, optionally fine-tuning the features, followed by nearest prototypical classifier.

3.1. Supervised Contrastive Learning

SCL consists in learning an embedding space in which the samples with the same class labels are close to each other, and the samples with different class labels are far from each other. Formally, a composition of an encoder f and a shallow neural network h called a projector (usually a MLP with one hidden layer) are trained to minimize the distances between representations of samples of the same class while maximizing the distances between representations of samples belonging to different class. After convergence, h is discarded, and the encoder f is used for transfer learning on downstream tasks. SCL loss is calculated as follows:

$$\mathcal{L}^{SCL} = \sum_{i \in I} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp\left(\boldsymbol{z}_i \cdot \boldsymbol{z}_p / \tau\right)}{\sum_{n \in N(i)} \exp\left(\boldsymbol{z}_i \cdot \boldsymbol{z}_n / \tau\right)} \quad (1)$$

Table 1. Performance on the validation datasets.

System	Precision	Recall	F1-score	HB			ME			PB		
				Pr	Re	F1	Pr	Re	F1	Pr	Re	F1
No extra data												
Template Matching	2.42	18.32	4.28	-	-	-	-	-	-	-	-	-
ProtoNets	36.34	24.96	29.59	-	-	-	-	-	-	-	-	-
Moummad et al. [7]	**73.93**	55.59	63.46	**82.95**	82.32	**82.63**	**67.69**	84.61	**75.21**	**72.72**	33.33	45.71
No fine-tuning (Ours)	60.99 ±0.58	62.08 ±1.21	61.52 ±0.48	75.81 ±1.16	78.00 ±1.21	76.89 ±1.10	54.94 ±2.36	92.95 ±0.96	69.04 ±2.03	56.36 ±2.16	40.48 ±1.97	47.11 ±1.97
Fine-tuning (Ours)	65.00 ±1.19	**71.75** ±1.22	**68.19** ±0.75	74.63 ±1.21	**85.11** ±2.33	79.52 ±1.58	58.12 ±2.48	**95.73** ±1.86	72.30 ±1.97	64.44 ±1.97	**51.01** ±1.40	**56.93** ±1.24
Extra data												
Liu et al. [8]	**76.56**	49.54	60.16	97.95	79.46	**87.74**	86.27	84.62	85.44	57.52	27.66	37.36
Tang et al. (SL) [10]	-	-	66.6	-	-	85.8	-	-	79.2	-	-	48.1
Yan et al. (FL) [10, 11]	73.0	67.6	70.2	-	-	77.0	-	-	90.0	-	-	53.7
Yan et al. (MTFL) [11]	76.2	**75.3**	**75.7**	-	-	86.7	-	-	**90.2**	-	-	**58.9**

*We highlight in bold the best score for each metric.

where $i \in I$ is the index of an augmented sample within a training batch, containing two views of each original sample. These views are constructed by applying a data augmentation function A twice to the original samples. $\boldsymbol{z}_i = h(f(A(\boldsymbol{x}_i))) \in \mathbb{R}^{D_P}$ where D_P is the projector's dimension. $P(i) = \{p \in I : y_p = y_i\}$ is the set of indices of all positives in the two-views batch distinct from i sharing similar label with i. $|P(i)|$ is its cardinality, $N(i) = I \setminus \{i\}$, the $\cdot$ symbol denotes the dot product, and $\tau \in \mathbb{R}^{+*}$ is a scalar temperature parameter.

3.2. Regularization : Total Coding Rate

In Information Theory, the coding rate is the proportion of bits that carry non-redundant information. Let $Z = [z_1, ..., z_b]$ be a batch of b features of dimension d. The total coding rate (TCR) [18] $\mathcal{R}$ of Z is defined as follows:

$$\mathcal{R}(Z) = \frac{1}{2} \log \det \left(I + \frac{d}{b\epsilon^2} Z Z^T \right) \tag{2}$$

where $\epsilon > 0$ is a chosen precision. The training loss is:

$$\mathcal{L}^{Train} = L^{SCL} - \lambda R(Z) \tag{3}$$

where $\lambda > 0$ is a hyperparameter coefficient for the regularization term. We want the coding rate of Z to be as large as possible. The TCR regularization can be seen as a soft-constrained regularization of covariance term in VICReg [17], where the covariance regularization is achieved by maximizing TCR [18].

3.3. Fine-tuning

Using the same annotations as section (3.1), we define the fine-tuning loss as:

$$\mathcal{L}^{Finetune} = -\log \frac{\exp(\boldsymbol{z}_i \cdot \boldsymbol{z}_c)}{\sum\limits_{c\prime \neq c} \exp(\boldsymbol{z}_i \cdot \boldsymbol{z}_{c\prime})} \tag{4}$$

This loss is similar to the ProtoNets loss [19], which produces a distribution over classes for a query point based on a softmax over distances to the prototypes in the embedding space. However, we do not do meta-testing using episodes as in ProtoNets, we instead do regular batch training by fine-tuning the model using the augmented batch similarly to the supervised contrastive pre-training stage. We slightly modify the ProtoNets loss by removing the distance to the corresponding prototype from the summation in the denominator. Our intuition is drawn from the work of DCL [20], which enhanced performance by removing the positive comparison from the denominator of the normalized temperature-scaled cross-entropy loss (NT-Xent) originally used in SimCLR [3](Eq.5).

$$\mathcal{L}^{SimCLR} = -\log \frac{\exp(\boldsymbol{z}_i \cdot \boldsymbol{z}_{i\prime})}{\sum\limits_{j \neq i, i\prime} \exp(\boldsymbol{z}_i \cdot \boldsymbol{z}_j)} \tag{5}$$

We observe that in the NT-Xent loss (Eq. 5), when substituting the second element of each similarity term with the corresponding prototype, we obtain the $\mathcal{L}^{Finetune}$ loss.

3.4. Nearest Prototype Classifier

To make predictions, for each audio file, we compute the Euclidean distances between the queries and the prototypes to assign the labels of presence/absence of the event of interest. For robustness, each segment (both query and prototype) is augmented to create multiple views. The representations of these views are averaged to one representation vector, in addition, the positive and negative segments are also averaged to have one positive and one negative prototypes. Using the annotations from subsection(3.2), let Z_i be the subset of Z with class label i, we then define the prototype $\bar{Z}_i$ for each class label i as:

$$\forall i : \bar{Z}_i = \frac{1}{|Z_i|} \sum_{z \in Z_i} z \tag{6}$$

Let q be a query, we predict its label i_q as:

$$i_q = \arg\min_i \|q - \bar{\mathcal{Z}}_i\|_2 \quad (7)$$

The onsets and offsets decision of the event of interest is made based on the precise moment when the label for the next query transitions from a negative class to a positive class and from a positive class to a negative class, respectively.

4. EXPERIMENTS

We experiment on the BSED datasets from DCASE and refer the reader to the work of Nolasco et al. [2] for more details about these datasets.

4.1. Model Backbone

Our architecture is the same as the one used in our previous work [7]. We use a ResNet consisting of three blocks (64→128→256), each comprising three convolutional layers. We employ max pooling operations after each block of a kernel of size 2x2 for the first and second blocks, and of size 1x2 for the third block.

4.2. Training and validation procedure

We train our model from scratch on the training set using SCL framework with a temperature of 0.06, regularized with TCR with a square precision of 0.05 and a regularization coefficient of 0.001. We use SGD optimizer with a batch size of 128, a learning rate of 0.01 with a cosine decay schedule, momentum of 0.9, and a weight decay of 0.0001 for 100 epochs. We use the data augmentation policy in table 2.

Table 2. Training data augmentations. **SM:** Spectrogram Mixing, **FS:** Frequency Shift, **RRTC:** Random Resized Time Crop, **PG:** Power Gain, **AWGN:** Additive White Gaussian Noise.

Augs	SM	FS	RRTC	PG	AWGN
Params	factor	bands	ratio	factor	std
Values	$\beta(5,2)$	[0-10]	[0.6,1.0]	[0.75-1]	[0-0.1]

During the validation phase, we optionally fine-tune the whole model using $L^{Finetune}$ for adapting the features for each audio recording using a learning rate of 0.01 for 40 epochs. For this purpose, we used random resized time crop (RRTC) of ratio sampled uniformly between 90% and 100% of the total duration, and power gain (PG) of coefficient sampled uniformly between 0.9 and 1. This data augmentation procedure is lighter than the one performed during pre-training (2), and is also used to create multiple views for each query window during inference. In all our experiments, we train the backbone with three different seeds, and for each backbone, we conduct three evaluations, resulting in a total of 9 runs per experiment.

5. RESULTS

Table 1 shows our results, the baseline and the first two ranking teams of the 2022 and 2023 DCASE challenge editions. Our method outpeforms that of Liu et al.[8] (both with and without fine-tuning). We also improve upon our previous work [7] with fine-tuning. While Yan et al.[10] and Tang et al.[11] achieve better results with their semi-supervised frame-level (FL) approach, we outperform their segment-level (SL) approach. For a fair comparison, we divide Table 1 into methods that utilize extra data (such as AudioSet Strong [8] or the reuse of training data for the adaptation of features on each audio recording [10, 11]) and those that do not. We note that our approach utilizes only the available shots during inference, making it practical for real-time applications or settings with limited resources. In Table 3, we study pre-training strategies without fine-tuning, showing the superiority of regularized SCL (+TCR) compared to vanilla SCL, SimCLR and Cross-Entropy. In Table 4, we analyze fine-tuning methods : SCL, original Prototypical Loss, and $\mathcal{L}^{Finetune}$, confirming insights about removing the positive comparison from the denominator of the prototypical loss.

Table 3. Ablation of the pre-training method w/o fine-tuning.

Method	Precision	Recall	F1-score
Cross-Entropy	34.59±1.21	62.35±0.93	44.49±1.22
SimCLR	54.75±0.77	61.16±1.62	57.75±0.41
SCL	56.80±2.98	**62.77±0.77**	59.59±1.75
SCL+TCR	**60.99±0.58**	62.08±1.21	**61.52±0.48**

Table 4. Ablation study on the fine-tuning method.

Method	Precision	Recall	F1-score
SCL	62.75±1.34	70.92±0.72	66.58±1.05
Original Proto	55.62±2.68	**72.13±0.67**	62.77±1.86
$\mathcal{L}^{Finetune}$	**65.00±1.19**	71.75±1.22	**68.19±0.75**

6. CONCLUSION

In this work, we have presented a simple yet effective approach for bioacoustic few-shot sound event detection. Our approach involves pre-training a feature extractor using supervised contrastive learning with a regularization that enforces learning non-redundant features. The feature space learned by our approach allows for computing directly distances to the prototypes for making prediction. We also propose to further enhance the performance by fine-tuning the features for each audio file at the cost of longer inference. For our future work, we want to generalize our approach to bioacoustic sound event classification and explore robust feature adaptation techniques for when fewer shots are available (one-shot). We will also explore the frame-level approach, as well as a proposal-based approach for detecting variable length temporal regions of interest, that have not been previously investigated in this task.

7. REFERENCES

[1] Dan Stowell, "Computational bioacoustics with deep learning: a review and roadmap," *PeerJ*, vol. 10, pp. e13152, 2022.

[2] Inês Nolasco, Shubhr Singh, Veronica Morfi, Vincent Lostanlen, Ariana Strandburg-Peshkin, Ester Vidaña-Vila, Lisa Gill, Hanna Pamuła, Helen Whitehead, Ivan Kiskin, et al., "Learning to detect an animal sound from five examples," *arXiv preprint arXiv:2305.13210*, 2023.

[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.

[4] Eduardo Fonseca, Diego Ortego, Kevin McGuinness, Noel E O'Connor, and Xavier Serra, "Unsupervised contrastive learning of sound event representations," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 371–375.

[5] Li Jing, Pascal Vincent, Yann LeCun, and Yuandong Tian, "Understanding dimensional collapse in contrastive self-supervised learning," *arXiv preprint arXiv:2110.09348*, 2021.

[6] Yaodong Yu, Kwan Ho Ryan Chan, Chong You, Chaobing Song, and Yi Ma, "Learning diverse and discriminative representations via the principle of maximal coding rate reduction," *Advances in Neural Information Processing Systems*, vol. 33, pp. 9422–9434, 2020.

[7] Ilyass Moummad, Romain Serizel, and Nicolas Farrugia, "Pretraining Representations for Bioacoustic Few-Shot Detection Using Supervised Contrastive Learning," in *Proceedings of the 8th Detection and Classification of Acoustic Scenes and Events 2023 Workshop (DCASE2023)*, Tampere, Finland, September 2023, pp. 136–140.

[8] Haohe Liu, Xubo Liu, Xinhao Mei, Qiuqiang Kong, Wenwu Wang, and Mark D Plumbley, "Surrey system for dcase 2022 task 5 : Few-shot bioacoustic event detection with segment-level metric learning technical report," Tech. Rep., DCASE2022 Challenge, June 2022.

[9] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 776–780.

[10] Jigang Tang, Zhang Xueyang, Tian Gao, Diyuan Liu, Xin Fang, Jia Pan, Qing Wang, Jan Du, Kele Xu, and Qinghua Pan, "Few-shot embedding learning and event filtering for bioacoustic event detection technical report," Tech. Rep., DCASE2022 Challenge, June 2022.

[11] Genwei Yan, Ruoyu Wang, Liang Zou, Jun Du, Qing Wang, Tian Gao, and Xin Fang, "Multi-task frame level system for few-shot bioacoustic event detection," Tech. Rep., DCASE2023 Challenge, June 2023.

[12] Calum Heggan, Sam Budgett, Timothy Hospedales, and Mehrdad Yaghoobi, "MetaAudio: A few-shot audio classification benchmark," in *International Conference on Artificial Neural Networks*. Springer, 2022, pp. 219–230.

[13] Stefan Kahl, Connor M Wood, Maximilian Eibl, and Holger Klinck, "BirdNET: A deep learning solution for avian diversity monitoring," *Ecological Informatics*, vol. 61, pp. 101236, 2021.

[14] Burooj Ghani, Tom Denton, Stefan Kahl, and Holger Klinck, "Feature Embeddings from Large-Scale Acoustic Bird Classifiers Enable Few-Shot Transfer Learning," *arXiv preprint arXiv:2307.06292*, 2023.

[15] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan, "Supervised contrastive learning," *Advances in neural information processing systems*, vol. 33, pp. 18661–18673, 2020.

[16] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny, "Barlow Twins: Self-supervised learning via redundancy reduction," in *International Conference on Machine Learning*. PMLR, 2021, pp. 12310–12320.

[17] Adrien Bardes, Jean Ponce, and Yann LeCun, "VICReg: Variance-invariance-covariance regularization for self-supervised learning," *arXiv preprint arXiv:2105.04906*, 2021.

[18] Shengbang Tong, Yubei Chen, Yi Ma, and Yann Lecun, "EMP-SSL: Towards Self-Supervised Learning in One Training Epoch," *arXiv preprint arXiv:2304.03977*, 2023.

[19] Jake Snell, Kevin Swersky, and Richard Zemel, "Prototypical networks for few-shot learning," *Advances in neural information processing systems*, vol. 30, 2017.

[20] Chun-Hsiao Yeh, Cheng-Yao Hong, Yen-Chi Hsu, Tyng-Luh Liu, Yubei Chen, and Yann LeCun, "Decoupled contrastive learning," in *European Conference on Computer Vision*. Springer, 2022, pp. 668–684.