FEW-SHOT CONTINUAL LEARNING FOR BIOACOUSTIC EVENT DETECTION

Technical Report

Xiaoxiao Wu

Shanghai Normal University
Shanghai, China
xiaoxiaowu2022@163.com

Yanhua Long

Shanghai Normal University
Shanghai, China
yanhua@shnu.edu.cn

ABSTRACT

In this technical report, we describe our submission system for DCASE2022 Task5: few-shot bioacoustic event detection. In this submission, a few-shot continual learning framework is used for our bioacoustic event detection, where we can continuously expand a trained base classifier to detect novel classes with only few labeled data at inference time. On the official validation set, the proposed continual learning achieves the overall F-measure score of 53.876%.

Index Terms— few-shot learning, continual learning, sound event detection

1. INTRODUCTION

This report presents the technical details of our SHNU submission system for DCASE2022 Challenge Task5: few-shot bioacoustic event detection [1]. This task focuses on sound event detection in a few-shot learning setting for animal (mammal and bird) vocalisations. Participants are expected to propose methods that can extract information from five exemplar vocalisations (shots) of mammals or birds and detect and classify sounds in field recordings. The challenge of this task is to find reliable algorithms that are capable of dealing with data sparsity, class imbalance and noisy/busy environments. Contrary to standard supervised learning paradigm, few-shot learning describes tasks in which an algorithm must make predictions given only a few instances of each class.

Among different few-shot learning methods, metric-based prototypical networks [2, 3, 4, 5, 6] have been shown to yield excellent performance for sound event detection. Nowadays, few-shot continual learning has been widely used for sound classification and has been proven to have performance no less than metric-based prototypical networks [7]. Inspired by few-shot continual learning, we generalize it to our bioacoustic event detection using dynamic few-shot learning (DFSL) [7, 8], where we can fast learn the information of novel classes using only few data samples at inference time.

2. METHODS

2.1. Dynamic few-shot learning (DFSL)

Dynamic few-shot learning (DFSL) [7, 8] is a classification approach that aims to learn novel class information from a well trained base classifier and only a few labeled data of this novel class. The DFSL method used for our submission is directly borrowed from the original paper [7], where the base classifier is consist of an encoder, and a classifier. The encoder is taken as a high-level feature extractor. Given an input waveform, it first be transformed by the encoder to be a feature vector $z \in R^d$, and then be classified by the classifier using a set of N classification weight vectors-one per class, $W_{base} = \{w_b\}_{b=1}^{N}$.

For the novel class information learning, DFSL designed an additional module named weight generator, which can generate a new weight vector $w_{N+1}^{'}$ in classifier for the novel class. The weight generator takes only K labeled examples of a novel class as input and exploits past knowledge by incorporating an attention mechanism over the existing classification weight vectors of N base classes [7]. The $w_{N+1}^{'}$ is formulated as,

$$w_{N+1}^{'} = \phi_{avg} \bigodot z_{avg} + \phi_{att} \bigodot w_{att}^{'}, \tag{1}$$

where $\phi_{avg}, \phi_{att} \in R^d$ are two learnable weights, $\bigodot$ is the Hadamard product, $z_{avg} = \frac{1}{K}\sum_{i=1}^{K} z_i$ is the averaged feature vector, $w_{att}^{'}$ is given by

$$w_{att}^{'} = \frac{1}{K}\sum_{i=1}^{K}\sum_{b=1}^{N} Att(\Phi_q z_{avg}, k_b) \cdot w_b, \tag{2}$$

where $\Phi_q \in R^{d\times d}$ is a learnable matrix for query vector transformation, $k_b \in R^d$ is a learnable keys for memory indexing. By combining the generated novel class weight vector $w_{N+1}^{'}$ with the base weight vectors W_{base}, the new classifier can predict both base and novel classes in one unified framework.

2.2. Weight generator

To train the few-shot classification weight generator, we use the method in [8]: in each batch, we randomly pick M "fake" novel classes from the total N base classes, and we treat them in the same way as we will treat the actual novel classes after training. We can then update the weight generator parameters and base classification weight vectors to minimize the classification loss on a batch with both base and pseudo-novel classes. Then the inferred "fake" classification weight vectors are used for recognizing the "fake" novel classes.

2.3. Negative Selection

How to model the negative class in few-shot classification tasks is important, especially when the positives are dense in the query recording. Such as the HB validation set in DCASE 2022 Task 5 challenge, the randomly selected negatives will be failed, because positives are easily selected as negatives by the randomly selection.

Therefore, instead of using the conventional randomly negative selection as in the DCASE Task5 baseline prototypical network system, we use a novel negative re-selection (NR) method that has been proposed in our previous work [9], to solve the negative support set construction problem for tasks with dense positives in query set. The NR aims to achieve high quality representative negative support set for similarity computing during query inference, by eliminating the positive mistaken examples in the conventional random negative selection, using the distance distribution between available positive prototype with the initial randomly selected negatives and limited available labeled positives in each query recording. These filtered negatives are supposed to be high quality representative ones to form the final negative support set during each query recording inference.

3. EXPERIMENT

3.1. Data

In this challenge, the development set provided by the official organizers is pre-split into training and validation sets. The training set consists of five different sub-folders (BV, HV, JD, MT, WMW), each for one source class. Along with the audio files multi-class annotations are provided for each. The total duration of whole training set is 21 hours, total classes is 46. The validation set comprises of three sub-folders (HV, PB, ME). It includes total 5 hours and 57 minutes data recordings. Each recording has multiple types of calls or species present in it, as well as background noise.

3.2. Configurations

All of our systems use the same Per channel energy normalisation (PCEN) features as used in the official baseline system. In this report, we use time and frequency masking for data augmentation.

For our approach, we first train a supervised base classifier comprising a feature encoder model and a base classification weight matrix using the pre-split positive examples of N (N=46) classes with the given strong labels. Then we train the attention-based few-shot weight generator following the steps described in section 2.2, where $(M, K) = (5, 5)$, on batches of 25 samples of pseudo-novel classes and 25 samples of the remaining base classes. Different from the powerful 14-layer CNN that used in [7], our system use the 9-layer ResNet structure [1] that provided in the official baseline systems as DFSL encoder and a linear layer (weight matrix) as the classifier.

During both validation and evaluation, we treat each file in the dataset independently of the others. For each recording, only five given positives are used, and we select 200 negatives using the conventional randomly negative selection, or our proposed NR method that presented in section 2.3. These positives and negatives are taken to form two novel classes, and used to fast learn the novel classification weight vectors $W_{new} = \{w_p, w_n\}$ for positive and negative respectively. Then, we only use the W_{new} to perform the event detection of pre-split query set in each recording. The maximum of probability classification scores of two classes is taken as the prediction result.

3.3. Experimental results

Table 1 shows the overall experimental results on the official validation set. "DFSL_aug" means we use time and frequency masking data augmentation. "Sn" means whether we use the proposed NR negative selection method. The results in that last line show that our method outperforms the best official baseline results with "Prototypical Network (PN)" significantly.

Table 1: Overall performances (%) on DCASE 2022 Task 5 validation set.

Method	Sn	F-measure	Precision	Recall
Baseline-PN	no	29.59	36.34	24.96
DFSL	no	43.661	40.45	47.43
DFSL	yes	48.508	52.36	45.18
DFSL_aug	no	46.634	46.51	46.75
DFSL_aug	yes	53.876	64.20	46.41

4. CONCLUSIONS

This technical report presents all the methods that used in our submissions of DCASE 2022 Task5. Experiments on the validation set show that the introduced DFSL can improve the performance more than absolute 24.286% F-measure over the baseline. Moreover, we see that different training strategy and model structure can also affect the system performances.

5. REFERENCES

[1] https://dcase.community/challenge2022/task-few-shot-bioacoustic-event-detection.

[2] S.-Y. Chou, K.-H. Cheng, J.-S. R. Jang, and Y.-H. Yang, "Learning to match transient sound events using attentional similarity for few-shot sound recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 26–30.

[3] J. Pons, J. Serrà, and X. Serra, "Training neural audio classifiers with few data," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 16–20.

[4] K.-H. Cheng, S.-Y. Chou, and Y.-H. Yang, "Multi-label few-shot learning for sound event recognition," in *IEEE 21st International Workshop on Multimedia Signal Processing (MMSP)*, 2019, pp. 1–5.

[5] Y. Wang, J. Salamon, N. J. Bryan, and J. Pablo Bello, "Few-shot sound event detection," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 81–85.

[6] B. Shi, M. Sun, K. C. Puvvada, C.-C. Kao, S. Matsoukas, and C. Wang, "Few-shot acoustic event detection via meta learning," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 76–80.

[7] Y. Wang, N. J. Bryan, M. Cartwright, J. Pablo Bello, and J. Salamon, "Few-shot continual learning for audio classification," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 321–325.

[8] S. Gidaris and N. Komodakis, "Dynamic few-shot visual learning without forgetting," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4367–4375.

[9] X. Wu, Y. Long, and D. Xu, "Inference with Negative Reselection for Few-shot Bioacoustic Event Detection," *submitted to International Journal of Pattern Recognition and Artificial Intelligence (IJPRAI)*, 2022.