FEW-SHOT BIOACOUSTIC EVENT DETECTION USING GOOD EMBEDDING MODEL

Technical Report

*Taein Kang**

Chung-Ang University, Seoul, South Korea
xodls4179@gmail.com

ABSTRACT

Few-shot learning is widely used as benchmarks for meta-learning. Few-shot learning is a learning algorithm that attempts to show how quickly it adapts to test tasks with limited data. Unlike general image new-shot learning, DCASE 2022 Task 5 [1] examines whether it can detect the corresponding sound at the back of audio data when five annotations are given in audio data. In this paper, we would like to demonstrate whether an embedding model well-learned bioacoustic information can perform few-shot learning well even with a simple classifier.

Index Terms— few-shot, representation

1. INTRODUCTION

Bioacoustic event detection in audio is an important task for automatic wildlife monitoring, as well as in citizen science and audio library management. Bioacoustic event detection is a very common required first step before further analysis, and makes it possible to conduct work with large datasets (e.g. continuous 24h monitoring) by filtering data down to regions of interest. Few-shot learning is a highly promising paradigm for scarce bioacoustic event detection. For the main assessment, we will use the F measure of detection performance.

In previous studies, the good learned representation as embedding model [2] have suggested simple baseline with simple classifier in few-shot image classification. And ECAPA-TDNN [3] has provided state of the-art results on speaker verification.

In our proposed system, we extract PCEN [4] from bioacoustic audio and merged dataset like ordinary classification dataset. Then, we trained embedding model to learn classification representation. And finally, embedding model with simple classifier detect bioacoustic scene.

The rest of the paper is organized as follows. In section 2, features used in proposed system is described. In section3, we interpret the ECAPA-TDNN model and the corresponding configuration. Section 4 concludes our work.

2. FEATURES

2.1 PCEN

In real world audio recording, especially outdoors, there are usually multiple sources. Recently, Per-channel energy normalization
(PCEN) [4] has been proposed as an alternative to MFCC, which aims to whiten the background of acoustic recordings and improve the robustness to channel distortion through temporal integration, adaptive gain control, and dynamic range compression

2.2 SpecAugment

We use SpecAugment in our validaton and evaluation data as data augmentation method. SpecAugment [5] is applied to the input features of a neural network. The augmentation policy consists of warping the features, masking blocks of frequency channels and masking blocks of time steps. In our systems, SpecAugment is applied to the PCEN features using frequency masking and time masking. The frequency mask can improve the robustness of our systems to frequency distortion of audios. Time masking is applied in the time domain, which is similar to frequency masking.

3. MODEL

Meta-learning is often used in face of the problem of few-shot classification, where only limited exampled data is provided and a classifier must generalize to given classes. The key idea of meta learning can be expressed as ”learn to learn”, which is a classification model learns how to learn by learning the training data. For this, we used the ECAPA-TDNN model, which showed excellent performance in speaker verification, as the embedding model, thinking that a model that verifies the speaker well would distinguish mammals and birds well

3.1 ECAPC-TDNN Structure

The structure of the ECAPA-TDNN extends the x-vector

architecture. It has a speaker encoder with a 3-layer bottleneck structure, including Res2Net and SE (Squeeze-and-Excitation) block. The 1D dilated convolution is applied to each bottleneck The size of the SE block channel is 1024. After passing through the 3-layer bottleneck, the global mean and standard deviation calculated through attentive statistics pooling are reflected in the channel-dependent frame attention. Finally, after concatenating the statistic reflecting the attention and weight passing through the bottleneck, it passes through the fully connected layer. The size of the output speaker embedding dimension is 192.

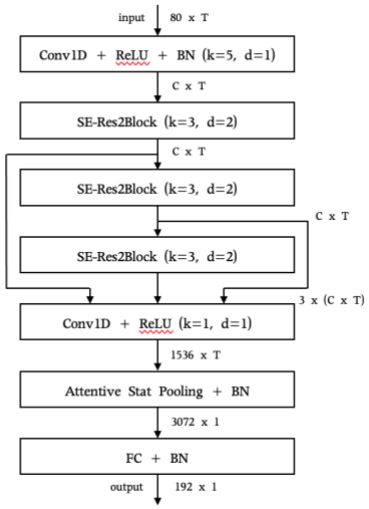

Figure 1: Structure of ECAPA-TDNN embedding model

3.2 Simple Classifier

We classified labels using AAM-softmax [6] in addition to the embedding network when learning the merged training dataset. After training, in the first submission, bioacoustics detection was conducted with 5-shot data augmented with AAM-softmax again. In the second submission, bioacoustics detection was performed by learning 5-shot data augmented with a 3 layer DNN classifier on the embedding network.

3.3 Training

Our goal is to learn a representation that distinguishes the bioacoustic well, and the ECAPA-TDNN network uses merged training dataset to obtain the ability to discriminate bioacoustic. More specifically, the network classifies the labels of the training dataset by adding a AAM softmax to the 192-dimensional embedding output. Ideally, the network should classify the bioacoustic scene well to build a good learned embedding network. In the validation and evaluation set, bioacoustic detection is performed after attaching a new classifier to the trained network and learning the classifier from the augmented 5-shot data.

4. CONCLUSION

In this technical report, we propose using ECAPA-TDNN-based embedding network with simple classifier for few-shot bioacoustic event detection task. We use PCEN features and apply data enhancement methods such as specaugment to improve model performance. If you have any questions, please email them to xodls4179@gmail.com.

5. REFERENCE

[1] http://dcase.community/workshop2022/.
[2] Tian, Yonglong, et al. "Rethinking few-shot image classification: a good embedding is all you need?." European Conference on Computer Vision. Springer, Cham, 2020.
[3] Desplanques, B., Thienpondt, J., Demuynck, K. (2020) ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification. Proc. Interspeech 2020, 3830-3834, doi: 10.21437/Interspeech.2020-2650
[4] V. Lostanlen, J. Salamon, M. Cartwright, B. McFee, A. Farnsworth, S. Kelling, and J. P. Bello, “Per-channel energy normalization: Why and how,” IEEE Signal Processing Letters, vol. 26, no. 1, pp. 39–43, 2019.
[5] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, “Specaugment: A simple data augmentation method for automatic speech recognition,” Interspeech 2019, Sep 2019. [Online]. Available: http: //dx.doi.org/10.21437/Interspeech.2019-2680
[6] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, “ArcFace: Additive angular margin loss for deep face recognition,” in 2019 IEEE/CVF CVPR, 2019, pp. 4685–4694.