FEW-SHOT BIOACOUSTIC EVENT DETECTION USING PROTOTYPICAL NETWORK WITH BACKGROUND CLASS

Technical Report

Yue Zhang[1]*, Jun Wang*[2]*, Dawei Zhang*[3]*, Feng Deng*[4]

[1] University of Electronic Science and Technology of China, Chengdu, China, 17331827076@163.com
[2]wangjun19930314@foxmail.com, [3]909237397@qq.com, [4]435012547@qq.com

ABSTRACT

Few-shot bioacoustic event detection is a task to detect and classify bioacoustic events with only a few instances. This task was firstly introduced in DCASE2021 Task 5, which requires participants to create a method that can extract information from five sample sounds (shots) of mammals or birds, and detect sounds in field recordings. In this paper, a prototypical network-based method was proposed for few-shot bioacoustic event detection challenge. In order to detect the target event from the query sequence, we need to distinguish the target event, other events, and background noise with only a few support set. To solve this problem, we propose to sample background noise from the training dataset as the "NEG" class for small sample learning. To better distinguish between events and background noise, the "NEG" class is used as a "way" in each episode of training. Experimental results show that the proposed method can effectively distinguish target events and background noise. The F-measure of sound event detection(SED) in the DCASE2021 Task 5 dataset can reach 57.10%, which is higher than the baseline method(41.48%).

Index Terms— few shot learning, sound event detection, prototypical network,

1. INTRODUCTION

Sound event detection(SED) is a popular research topic in audio field. It aims to detect the sound events and estimate the position in time of sound events in a audio segment. SED has many applications in our daily life, such as sound command recognition [1], bioacoustic event detection [2]. However, most deep learning SED methods require huge amount of annotated audio data for training. It is challenging to train a robust sound event event detection system with a small amount of labeled data.

Few-shot learning [3–5] is a challenging area of research, which strive to recognize novel events with only few examples. Typically, few-shot learning methods follow episodic training paradigm. In each episodic, we consider an N-way K-shot classification task for training, where N is the number of classes sampled from the dataset, K is the number of samples for each class. The goal of this task is to learn a classifer which can quickly accommodate to novel classes with only a few examples. Among different few-shot learning methods, metric learning is a promising approach in many few-shot learning task.

Acoustic event detection methods based on only a few examples have attracted the interest of researchers [6–8]. Few-shot learning has become a common paradigm to tackle this challenge. Recently, DCASE2021 Task 5 brings a related challenge focuses on sound event detection in a few-shot learning setting for animal (mammal and bird) vocalisations. In this task, participants are required to create a method that can extract feature from five sample sounds (shots) of mammals or birds, and detect sounds in field recordings. Due to the lack of supervised data, it is difficult for general supervised learning methods to perform well in the target classes.

In this technical report, we consider metric learning method called prototypical networks [9] to detect bioacoustic events. The prototypical networks was originally used to solve the problem of few-shot image classification. Recently, prototypical networks has also achieved excellent performance in the audio field [10, 11]. To better separate the background noise and target event, we consider background noise as an extra class for few-shot detection, which proved to be effective in [7].

2. PROPOSED METHOD

For few-shot bioacoustic event detection task, we consider prototypical networks as our method. Prototypical networks is a metric learning approach which can learn a discriminative feature space, and perform classification tasks by computing distances against prototypes in this embedding space. For better distinguish between events and background noise, background noise is sampled to form a "NEG" class. In our work, we apply ResNet12 architecture as our feature extraction model, and employ Euclidean distance to measure the embedding in feature space. The model was trained with 10-way-5-shot tasks.

2.1. Feature extraction

We extract 128-dimensional Per-channel energy normalization(PCEN) spectrograms from the input audio. The windows size and the hop size are 1024 and 256 with 25.6KHz sampling rate. We crop the segments from the audio recording with a length of 0.2 seconds(corresponding to 20 frames),and a hop length of 0.5 seconds(5 frames). We then perform the normalization on the training data to make the data with zero mean and unit variance distribution. Before input into the model, SpecAugment [12] is applied in the PCEN spectrograms to augment the training set.

2.2. Embedding module

In prototypical network, the sampled input are mapped to a metric space through an embedding module. The target of network is to train an embedding network that convert the input to embedding vector. In embedding space, samples of the same class are closer,

and samples of different class are farther away. Generally, a convolutional neural network(CNN) is used to extract the embedding feature.

In our work, we utilize ResNet12, a residual network with 12 convolutional layers, as the backbone of prototypical network. ResNet12 is composed of four residual blocks, with the number of channels in each residual block being 128, 128, 256, 128, respectively. Each residual block contains three 3×3 convolutional layers, with LeakyReLu activation function and batch normalization. The residual connection of each block consists of a single 1×1 convolutional layer with batch normalization. average pooling is applied at the end of each block with a pooling size of 2×2. Pooling across the time dimension is applied to the output of ResNet12 to reduce the dimension of the feature vector.

2.3. Episodic training

The prototypical network is trained with episode. In each episode, a mini batch is sampled from the training set to build a N-way-K-shot classification task. The mini batch is split into support set S and query set Q. The support set consisting N classes, each class including K samples. The remaining samples of the mini batch are query set.

In the sampling step, we sample $N - 1$ event classes and a background noise class, with $K * 2$ samples in each class to compose a mini batch. The event class are randomly selected from all classes in training set. The samples in the background noise class are randomly selected from the training audio files. There are a large number of unmarked time segments in the training audio file, and we regard these segments as background noise class.

Prototypical networks compute the prototype c_k through the embedding module f. We use ResNet12 to map input to embedding vector. The prototype of the class is obtained by averaging the prototypes of all samples of the class in support set:

$$c_k = \frac{1}{|S_k|} \sum_{(x_i) \in S_k} f_\phi(x_i) \tag{1}$$

For each sample x_q in query set, we calculate the squared Euclidean distance between the sample and each prototype and apply a softmax function to get the class probability distribution of the sample:

$$p_k(y = k|x_q) = \frac{exp(-d(f_\phi(x_q), c_k))}{\sum_{k'} exp(-d(f_\phi(x_q), c_{k'}))} \tag{2}$$

The prototypical network is trained to optimize the cross-entropy loss for each sample x_q.

2.4. Event detection

After the prototypical network is trained in the training set. We apply it to sound event detection in validation Set. Given the first five POS annotations for the class of interest in a long recording, we calculate the POS prototype c_+ by averaging the prototypes of all POS samples. The purpose of bioacoustic event detection is to distinguish the POS event and everything else in a audio file. We construct the negative sample set by randomly selecting frames within the audio file to be detected. We then calculate the NEG prototype c_- by averaging the prototypes of all negative samples.

We then calculate the distance of a query sample x_q to the POS prototype c_+ and the NEG prototype c_-, denoted as d_+ and d_- respectively. Finally, a softmax function is applied to obtain the probability y to predict whether the event occurred in the sample.

$$p(y = pos) = \frac{exp(-d_+)}{exp(-d_+) + exp(-d_-)} \tag{3}$$

In the inference step, we compared the probability with a threshold, and post processing of the prediction result by removing all events that have shorter duration than 60% of the minimum duration of the shots for that audio file.

3. EXPERIMENTS

3.1. Dataset

The dataset of DCASE2021 task 5 is divided into training and validation sets. The training set consists of four subsets:BV, HT, JD and MT. In training set, multi-class annotations are provided for the audio files. The validation set consists of two subsets,: HV, PB. A single-class annotation is provided for each audio in validation set. There is no overlap between the training set and validation set classes.

Model	F-measure	Precision	Recall
Baseline	41.48%	32.20%	58.27%
Baseline2	47.54%	59.76%	39.47%
ResNet12	57.10%	60.24%	54.28%

Table 1: The Baseline is the baseline prototypical network in DCASE2021 Task 5, the Baseline2 is the Baseline trained with "NEG" class, ResNet12 is our best result in validation dataset.

3.2. Result

The prototypical network is trained in the training sets with few-shot audio classification [8, 10] task and perform bioacoustic event detection task in the validation sets. Tab. 1 show the evaluation results in validation sets with the following metrics: F-measure, precision, recall. The Baseline model[1] is the baseline prototypical network, which consists of four convolutional layers. The Baseline2 is the baseline model trained with NEG class. The ResNet12 is the model used in this report.

Comparing Baseline and Baseline2, it can be seen that with the introduction of background noise class, the precision has been significantly improved, and it has also led to a decline in recall. The possible reason is that with the introduction of background noise class, the model is easier to predict the target event as background noise. The proposed ResNet12-based prototypical network can reach a F-measure of 57.10%.

3.3. Conclusion

In this technical report, we introduce a prototypical network with a backbone of ResNet12. To better discriminate the target event and background noise, a background noise class is considered for training. Evaluation on the validation sets show that the proposed method can improve the precision in event detection, and get high F-measure.

[1] http://dcase.community/challenge2021/task-few-shot-bioacoustic-event-detection.

4. REFERENCES

[1] R. Alvarez and H.-J. Park, "End-to-end streaming keyword spotting," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6336–6340.

[2] R. Bardeli, D. Wolff, F. Kurth, M. Koch, K.-H. Tauchert, and K.-H. Frommolt, "Detecting bird sounds in a complex acoustic environment and application to bioacoustic monitoring," *Pattern Recognition Letters*, vol. 31, no. 12, pp. 1524–1534, 2010.

[3] W.-Y. Chen, Y.-C. Liu, Z. Kira, Y.-C. F. Wang, and J.-B. Huang, "A closer look at few-shot classification," *arXiv preprint arXiv:1904.04232*, 2019.

[4] J. Lu, P. Gong, J. Ye, and C. Zhang, "Learning from very few samples: A survey," *arXiv preprint arXiv:2009.02653*, 2020.

[5] Y. Wang, Q. Yao, J. T. Kwok, and L. M. Ni, "Generalizing from a few examples: A survey on few-shot learning," *ACM Computing Surveys (CSUR)*, vol. 53, no. 3, pp. 1–34, 2020.

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

[7] K. Shimada, Y. Koyama, and A. Inoue, "Metric learning with background noise class for few-shot detection of rare sound events," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 616–620.

[8] 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 *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 26–30.

[9] J. Snell, K. Swersky, and R. S. Zemel, "Prototypical networks for few-shot learning," *arXiv preprint arXiv:1703.05175*, 2017.

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

[11] S. Singh, H. L. Bear, and E. Benetos, "Prototypical networks for domain adaptation in acoustic scene classification," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 346–350.

[12] 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," *Proc. Interspeech 2019*, pp. 2613–2617, 2019.