BIT_SRCB TEAM'S SUBMISSION FOR DCASE2022 TASK5 - FEW-SHOT BIOACOUSTIC EVENT DETECTION

Technical Report

Miao Liu[1], *Jianqian Zhang*[1], *Lizhong Wang*[2], *Jiawei Peng*[1], *Chenguang Hu*[1]
Kaige Li[1], *Jing Wang*[1], *Qiuyue Ma*[2]

[1]Beijing Institute of Technology, Beijing, China,
[2]Samsung Research China-Beijing (SRC-B), Beijing, China
{3120200795, 3120210828, wangjing}@bit.edu.cn, {lz.wang, qiuyue.ma}@samsung.com

ABSTRACT

In this technical report, we present our system for the task 5 of Detection and Classification of Acoustic Scenes and Events 2022 (DCASE2022) challenge, i.e. few-shot bioacoustic event detection. First, per-channel energy normalization (PCEN) is extracted as features. In order to improve the diversity of original audio, some data augmentation methods are adopted, for example, specaugment. Then, the prototypical network with convolutional neural networks (CNN) and the transductive inference method are used for few-shot detection in our systems. Finally, we use aforementioned features as inputs to train our CNN model. Moreover, we merge the prediction results of improved prototypical network and transductive inference method for better performance. We evaluate the proposed systems with overall F-measure for the whole of the evaluation set, and our best F-measure score on the validation set is 64.77%.

Index Terms— DCASE, few-shot bioacoustic event detection, PCEN, prototypical networks, transductive inference

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 [1]. 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-score measure of detection performance.

In previous studies, the prototypical network as classifiers have recently shown improved performances over existed methods in few-shot acoustic event detection [2]. And CNN has provided state-of-the-art results on various polyphonic sound event detection and audio tagging tasks [3]. Besides, good results have been achieved by using unsupervised or semi-supervised methods to deal with limited labeled data, and some works have attempted to augment few-shot learning with the benefit of learning from unlabeled data, such as increasing the labeled set by using additional unlabeled samples through pseudo-labeling [4, 5]. In recent years, more and more people have paid attention to the application of few-shot learning in the field of acoustic event detection and classification. Szu-yu Chou [6] et al. introduced a novel attentional similarity module for few-shot sound recognition problems, which can be inserted into any metric-based learning method for few-shot learning, thus enabling the generated model to specifically match relevant short sound events. Yu Wang [7] et al. developed a method for automatically constructing partial labeled samples (negative samples) to reduce user markup effort, which adjusts the metric-based few-shot learning approach. For acoustic event detection containing background noise, Kazuki Shimada [8] et al. explicitly defined background noise as an independent class, thus providing a feature space in which event classes and background noise classes are fully separated.

It can be seen that by adopting different data processing methods and adding various modules for different problems, prototype network has a huge advantage in few-shot learning. In order to solve the problem of insufficient small support set and feature extractor agnostic to task, a novel mutual learning framework with transductive learning was introduced to update class prototype and feature extractor iteratively [9].

In our proposed system, we use SpecAugment as one of data augmentation methods in few-shot bioacoustic event detection. Then, we extract PCEN from the bioacoustic audio. Finally, we train an improved prototypical network and use transductive inference method to overcome the difficulties of few shot problems. Moreover, we merge the prediction results of improved prototypical network and transductive inference method according to the length of sound events for better performance.

The rest of the paper is organized as follows. In section 2, the dataset and features used in proposed system is described. In section3, we interpret the prototypical network and the transductive inference method. Experiment results are presented in Section 4. Section 5 concludes our work.

2. DATASET AND FEATURES

2.1. Dataset

The development dataset for task 5 consists of multi-class animal (mammal and bird) audio files. The development dataset is split into training and validation sets, and with annotation file provided for each audio file. Audio recordings are resampled to a sampling rate of 22050 Hz. We used the Librosa library to generate the acoustic features. Similar to [7], we automatically construct a set of negative examples for inference, and adopt the inference-time data augmentation method to generate more positive examples without increasing the cost of manual marking. The query set is comprised of all audio clips after the fifth annotation. It is noted that we do not use the WMW dataset in training set according to the experimental results.

2.2. Features

In real world audio recording, especially outdoors, there are usually multiple sources. Recently, per-channel energy normalization (PCEN) [10] has been proposed as an alter-native 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.

3. METHODS

3.1. Prototypical networks

Meta-learning is often used when solving the problem of few-shot classification, where only limited example data is provided and the classifier must generalize the information from the examples to the given classes. The idea of meta-learning tasks can be summarized as “learning to learn”, which means the classifier can train itself to optimize the function for mapping the data into the intended label with the training tasks. There has been a few methods to achieve this idea, and prototypical network is a classic paradigm of meta-learning.

3.1.1. Structure

Prototypical network [11] consists of a feature extractor and a classifier. The classifier uses the euclidean distance of the features as the measurement method. This measurement is to describe the similarity of the features and distinguish the differences among types. So, the performance of the network depends on the generalization ability of the feature extractor. If with a lot of few-shot classification tasks being trained, the feature extractor can getting stronger. Then, when giving a new set of limited data, the network will be able to recognize new classes by extracting the feature of labeled data as ’prototype’ and compare the similarity of the prototype with other features. Our prototypical network is based on ResNet, and the input feature is the built by PCEN.

3.1.2. SpecAugment

We use SpecAugment [12] as our data augmentation method. SpecAugment is a simple data augmentation method, and we apply it on to the feature inputs of the neural network. The augment policy consists of warping the features, masking time channels and masking frequency channels. And in our work, we apply time and frequency masking to our PCEN features, for the time masking is used in time domain, which is as similar as the frequency masking. The augmentation policy can improve the robustness of our system, and makes the system more generalized.

3.1.3. Multiscale feature perception

In order to get more information from the features we get from the PCEN method, we apply a structure of multiscale perception. We use different scale in the kernel size of the ResNet block instead of the invariable kernel size. Therefore, in our training process, our network can learn better from the limited data and get more information based on different shape of scales of both time domain scale and frequency domain. That makes the extractor work better than only process with one single size of kernel.

3.1.4. Training

The network uses training tasks, here we refer to the few-shot classification tasks, to obtain the ability to optimize the mapping function between data and labels. When creating training tasks, we should manually divide the dataset into support set and query set, treating the training task as separate dataset. Ideally, the network should perform well after several epochs of training.

3.2. Transductive inference

Transductive inference (TI) is about reasoning from observed, specific (training) cases to specific (test) cases. In this paper, the core idea of TI is about leveraging the statistics of the unlabeled data. We adapt the idea from [9], which maximizes the mutual information (MI) between the query features and their label predictions for a few-shot task at inference. It means that TI has seen these unlabeled data before making final prediction.

3.2.1. Large convolution kernel size

We try to use large convolution kernel size of CNN in our model. We find that it is helpful to improve the performance of TI, which can increase the model receptive field. In our experiments, we find that the model with kernel size of 13 has the best performance.

3.2.2. Pretrained Features

PANNs [13] are the pretrained audio neural networks trained on large-scale AudioSet dataset with 1.9 million audio clips with an ontology of 527 sound classes. We use the ’CNN10’ model in PANNs to extract the pretrained features from the dataset of task 5 and concatenate the embedding features with pretrained features. Then, the combined features are send to a fully connected layer to generate the predictions.

3.3. Model ensemble

In the validation and evaluation set, we find that the lengths of some sound events are too long (more than 1s) or too short (less than 50ms). During inference time, we shift the selection window with 200ms in time by 50 ms increments in TI and use the variable window size that varies based on the event length in positive examples in baseline model. Moreover, we find that the template matching baseline system with variable window size got the best preformance on DC dataset last year which includes some audio longer than 1s. Therefore, we find TI preform worse than baseline model on those too long or too short events.

We fuse the prediction results of TI and baseline model according to the length of sound events. For the events with average length shorter than 50ms or longer than 500ms in positive examples, we use the results of baseline model. For the other events, we use the results of TI.

4. EXPERIMENTAL RESULTS

In this section, Table 1 shows the results of our systems on the validation set. We can see that the improved TI preforms better than the baseline models. The ensemble system that combines the results of our baseline prototypical network and improved TI gets the highest F-score of 64.77%.

Table 1: The results of F-measure score on the validation dataset.

Models	P	R	F
Baseline (prototypical networks)	36.34%	24.96%	29.59%
Our prototypical network	50.97%	32.30%	39.54%
Improved TI	64.43%	71.17%	58.86%
Ensemble of models	64.77%	70.86%	59.64%

5. CONCLUSIONS

In this technical report, we propose using improved prototypical network and transductive inference method for few-shot bioacoustic event detection task. We apply data enhancement methods such as specaugment, extract pretrained features and increase the model receptive field to improve model performance. At last, we get 64.47% under F-measure score on the validation set.

6. REFERENCES

[1] http://dcase.community/challenge2022/.

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

[3] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "Panns: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020.

[4] X. Li, Q. Sun, Y. Liu, Q. Zhou, S. Zheng, T.-S. Chua, and B. Schiele, "Learning to self-train for semi-supervised few-shot classification," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[5] M. Ren, E. Triantafillou, S. Ravi, J. Snell, K. Swersky, J. B. Tenenbaum, H. Larochelle, and R. S. Zemel, "Meta-learning for semi-supervised few-shot classification," *arXiv preprint arXiv:1803.00676*, 2018.

[6] 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.

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

[8] 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.

[9] D. Yang, H. Wang, Y. Zou, Z. Ye, and W. Wang, "A mutual learning framework for few-shot sound event detection," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 811–815.

[10] 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.

[11] J. Snell, K. Swersky, and R. S. Zemel, "Prototypical networks for few-shot learning," *CoRR*, vol. abs/1703.05175, 2017. [Online]. Available: http://arxiv.org/abs/1703.05175

[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," *Interspeech 2019*, Sep 2019. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2019-2680

[13] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "Panns: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020.