FEW-SHOT BIOACOUSTIC EVENT DETECTION USING PROTOTYPICAL NETWORKS WITH RESNET CLASSIFIER

Technical Report

Ren Li[1], Jinhua Liang[1], Huy Phan[1,2]

[1] School of Electronic Engineering and Computer Science, Queen Mary University of London, United Kingdom
[2] The Alan Turing Institute, United Kingdom
ren.li@hss20.qmul.ac.uk, jinhua.liang@qmul.ac.uk, h.phan@qmul.ac.uk

ABSTRACT

In this technical report, we describe our submission system for the few-shot bioacoustic event detection in the DCASE2022 task5. Participants are expected to develop a few-shot learning system for detecting mammal and birds sounds from audio recordings. In our system, Prototypical Networks are used to embed spectrograms into an embedding space and learn a non-linear mapping between data samples. We leverage various data augmentation techniques on Mel-spectrograms and introduce a ResNet variant as the classifier. Our experiments demonstrate that the system can achieve the F1-score of 47.88% on the validation data.

Index Terms— Few-shot learning, sound even detection, Prototypical Networks, embedding space

1. INTRODUCTION

Bioacoustic technology is now benefitting from the power of deep learning and becomes an effective way to gain information on the activities of animals that reflects human's impact on the environment. Traditionally, researchers have conducted the work through manually labelling on huge datasets, which is consuming both in time and resources [1]. In addition, collecting labelled data in some certain animal sounds can be challenging. The scarcity of supervised data can lead to poor generalization [1].

Few-shot learning is proposed to solve the problem of training with a limited amount of labelled data. It has emerged as a promising paradigm for sound event detection. A few-shot learning classifier is capable of recognizing novel classes not seen in the training set, given only a small number of examples of each new class [2]. In previous studies, Prototypical Network (ProtoNet) [3] has been increasingly applied to few-shot sound event detection and achieved well improved performance. Plus, deep learning architectures like CNNs have yielded state-of-the-art results on different sound recognition tasks, such as polyphonic sound event detection, audio tagging, etc.

In our proposed system, we first extract Mel-spectrogram from the bioacoustic audios and perform Per-Channel Energy Normalization (PCEN) [5] on the resulting spectrograms to account for differences between data coming from different sources. Then, we apply various spectrogram augmentation techniques to increase the amount of training data to help modelling generalization. Finally, we train a ProtoNet model using a deeper and stronger embedding classifier - ResNet.

2. DATA

2.1. Dataset

In this challenge, the training set consists of 174 audio recordings, 47 classes and 14,229 event instances in total [4]. They were acquired from different bioacoustic sources, including worldwide birds, spotted hyenas, jackdaws, meerkats, and wetlands birds. The sampling rate of each audio varies from 6 kHz to 44 kHz. In addition, multi-class annotations are provided for the training set with positive, negative, and unknown; we only extract and make use of the positive event instances for training and testing. The validation set consists of 18 audio recordings, 5 classes and 1077 positive event instances [4].

2.2. Pre-processing

2.2.1. Mel-spectrogram

All audio files in both the training set and validation set are first resampled to a sampling rate of 22,050 Hz and normalized. The audio files are then transformed to Mel-spectrograms with 128 Mel bins using a FFT size of 1024 samples and a hop size of 256 samples. The *librosa* library was employed for this purpose. Afterwards, spectrogram images of size $F \times T$ where F=17 by T=128 are used inputs. Additionally, the validation set is divided to positive samples, negative samples, and query samples for predictions.

2.2.2 PCEN

PCEN has been proposed to normalize a time-frequency representation by performing automatic gain control, followed by nonlinear compression [5]. Former research used PCEN to mitigate the effects of background noise, demonstrating its effectiveness as a preprocessing step prior to convolutional methods in sound event detection [5]. Bioacoustic data recorded in the wild often have multiple sound sources and uncleaned background. Therefore, we utilize PCEN to reduce noise presented in the Mel-spectrograms and improve robustness to channel distortion.

2.3. Data augmentation

In order to increase the diversity of data and the generalization ability of the model, we use *SpecAugment* [6] as the data aug-

mentation technique. It essentially consists of three transformations: time warping, frequency masking, and time masking. Specifically, they modify a spectrogram by warping it in the time direction with a distance factor, mask blocks of consecutive frequency channels, and mask blocks of time steps, respectively. In our case, we warp the feature to the left by 0.5 s, and mask one block of 10 frequency bins and one block of 10 time steps. Example of the original spectrogram and augmentation are show in Figure 1. With the help of data augmentation, the number of training samples is increased to 54,279 from 28630.

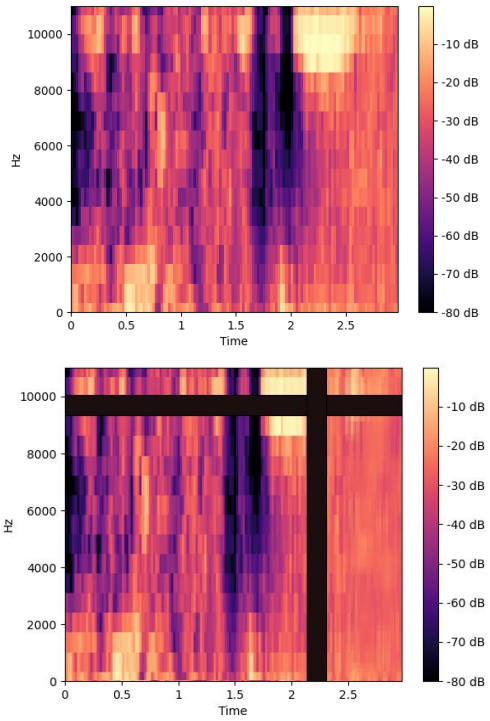

Figure 1: Example of spectrogram augmentation: orginial spectrogram (top) and augmented spectrogram with *SpecAugment* (bottom).

3. PROTOTYPICAL NETWORKS

A prototypical network transform and embed the input spectrograms into an embedding space using a neural network. Then it learns a non-linear mapping in the embedding space where the embedded query points are simply classified by matching them with the nearest class "prototype" [2]. Consequently, the performance heavily depends on the embedding feature extractor. The more informative features it produces, the more accurate "prototypes" can be computed, and more query points will be correctly classified, leading to better performance.

3.1. Model architecture

The original ProtoNet proposed by Snell et al. [2] consists of 4 convolutional layers, which can suffer from the problem of vanishing gradients and cannot extract features that fit the data well. Inspired by Ye et al. [7], we choose the deeper and stronger residual networks as the embedding encoder. ResNet architecture makes use of shortcut connections that allows information get fast forwarded deeper into the network to avoid vanishing gradients [8].

Our implementation is based on the ResNet-18. We modify the network to obtain a less deep model which only has 3 residual blocks to fit the size of the features. The architecture of our residual network is shown in Table 1. The residual block for this architecture is depicted in Figure 2.

Table 1. Architecture of the presented residual network

Layers	Channels	Kernel Size
Conv2D+ BatchNorm +ReLU	16	3 × 3
ResidualBlock	64	3 × 3
ResidualBlock	128	3 × 3
ResidualBlock	64	3 × 3
AdaptiveAvgPooling+SoftMax	-	3 × 3

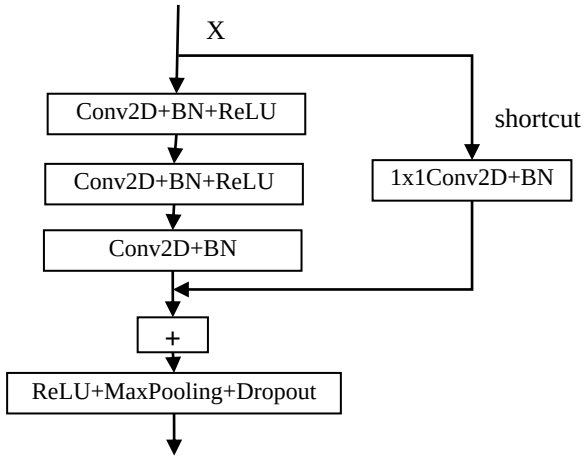

Figure 2: Residual block for the architecture

3.2. Training

Prototypical networks adopt an episodic training procedure where in each episode, a mini batch is randomly sampled from the training data. A subset of mini batch is used as the support set and the remaining is used as query set [9]. We trained the model using 2,000 episodes and 5 classes in each minibatch, with Adam optimizer and the learning rate of 0.001. Euclidean distance is selected as the metric that measures the distance between query samples to a prototype.

4. EVALUATION

Performance is evaluated using three metrics: precision, recall, and F1-score. Table 2 shows the testing results obtained by the proposed model in comparison with the DCASE2022 Task 5 baseline for the 5-way 5-shot task. As can be seen our system achieved the highest F1-score of 47.88%, which improves over the baseline by a large margin. Moreover, data augmentation techniques are proved to be effective, leading to a substantial improvement. Notably, it is observed that the recall is lower than precision in each system, indicating that the model could miss a

fairly number of True Positive instances that are difficult to classify.

Table 2. Overall results on the validation set

Classifier	Feature	F1-score	Precision	Recall
CNN(Baseline)	PCEN	29.59%	36.34%	24.96%
CNN(Baseline)	PCEN+Aug	37.16%	42.09%	33.26%
ResNet	PCEN+Aug	47.88%	52.11%	44.30%

5. CONCLUSION

In this work, we employ SpecAugment as the spectrogram augmentation technique to increase the generalization ability of Prototypical Networks. For networks architecture, we utilize the ResNet-18 variant as the embedding space classifier to enhance the performance of model. Our proposed system is able to achieve the F1-score of 47.88% which improves more than 17% absolute over the baseline system based on CNN.

6. REFERENCES

[1] DCASE2022, “Few-shot Bioacoustic Event Detection Task description” March 2022. [Online]. Available: http://dcase.community/workshop2022/.

[2] Snell J, Swersky K, Zemel R. Prototypical networks for few-shot learning. Advances in neural information processing systems. 2017;30.

[3] Yang D, Wang H, Zou Y, Ye Z, Wang W. A Mutual learning framework for Few-shot Sound Event Detection. InICASSP IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 811-815, 2022.

[4] DCASE2022, “DCASE 2022 Task 5: Few-shot Bioacoustic Event Detection Development Set”, March 2022. [Online] Available: https://doi.org/10.5281/zenodo.6012309

[5] Ick C, McFee B. Sound event detection in urban audio with single and multi-rate PCEN. In Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 880-884, 2021.

[6] Park DS, Chan W, Zhang Y, Chiu CC, Zoph B, Cubuk ED, Le QV. Specaugment: A simple data augmentation method for automatic speech recognition. arXiv preprint arXiv:1904.08779, 2019.

[7] Ye HJ, Hu H, Zhan DC, Sha F. Few-shot learning via embedding adaptation with set-to-set functions. In Proc. the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8808-8817, 2020.

[8] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In Proc. the IEEE conference on computer vision and pattern recognition, pp. 770-778, 2016.

[9] Li D, Zhang J, Yang Y, Liu C, Song YZ, Hospedales TM. Episodic training for domain generalization. In Proc. the IEEE/CVF International Conference on Computer Vision, pp. 1446-1455, 2019.