A META-LEARNING FRAMEWORK FOR FEW-SHOT SOUND EVENT DETECTION

Technical Report

*Tianyang Zhang**

Chongqing University
Key Laboratory of Optoelectronic Technology
and Systems, MOE
Shapinba, Chongqing University
zhangty@cqu.edu.cn

Yuyang Wang, Ying Wang

Chongqing University
Key Laboratory of Optoelectronic Technology
and Systems, MOE
Shapinba, Chongqing University
WangYuyang@cqu.edu.cn
2027212542@qq.com

ABSTRACT

The report presents our submission to Detection and Classification of Acoustic Scenes and Events challenges 2022 (DCASE2022) task 5. This task focuses on sound event detection in a few-shot learning setting for animal (mammal and bird) vocalisations. Main issue of this task is that only five exemplar vocalisations (shots) of mammals or birds are available. In this paper, we propose a meta-learning framework for few-shot bioacoustic event detection challenge. Maximizing inter-class distance and minimizing intra-class distance (MIMI) are used as a criteria to fine-tune embedded network for few-shot tasks. Experimental results indicate our framework get better performance than baseline, and F1 score is about 46.51% on evaluation set.

***Index Terms*—** Few-shot, Inter-class and intra-class, Sound event detection

1. INTRODUCTION

Simulating human auditory perception and creating general-purpose systems to detect interesting sound sources is called automatic sound event detection (SED). The goal of automatic SED is to identify sound events classes and detect the onsets and offsets of these events. Automatic SED has extensive application prospects in various fields, including noise monitoring [1], multimedia indexing [2] and audio surveillance [3].

In many practical situations, there exists large variety of audio events and labels available for rare events are prohibitively small. These situations focus on sound event detection in a few-shot learning setting, which is known as few-shot sound event detection. Meta-learning [4, 5, 6] is a key method to solve few-shot sound event detection. Shi [7] compares traditional supervised methods and a variety of meta-learning approaches applying in few-shot SED. Their experimental results show meta-learning models achieve superior performance. Yang [8] proposes a method, combined meta-learning with transductive inference, for few-shot SED. The core idea of their method is about leveraging the statistics of unlabeled data. Wang [9] successfully adapts metric-based meta-learning approaches to an open-set few-shot SED problem.

In this technical report, we propose a meta-learning framework for few-shot bioacoustic event detection, which inherits prototypical network [10]. Maximizing inter-class distance and minimizing intra-class distance (MIMI) are used to fine-tune embedded network. In addition, We set a distance constraint on intra-class distance to avoid overfitting of embedded network.

*Thanks to ABC agency for funding.

2. PROPOSED METHOD

We design maximizing inter-class distance and minimizing intra-class distance (MIMI), which makes embedded network can learn more discrimination embedding features for specific few-shot sound event detection task. The process of MIMI is illustrated in Figure 1. MIMI utilizes support set to fine-tune f_ϕ. a specific few-shot SED task is given two subsets S_c and $S_{c'}$. $dist(S_c)$ denotes intra-class distance and $dist(S_c, S_{c'})$ denotes inter-class distance.

After obtaining a fine-tuned embedded network for specific few-shot SED task, new class prototypes of support set can be recalculated. Then, prediction results for query sound samples are output based on euclidean distance. In addition, the specific few-shot SED task only have a few labeled sound samples in support set. If intra-class distance of support set is not controlled during fine-tuning process, it is easy to cause overfitting. We considerthat when the intra-class distance of bioacoustic events and background sounds are over-compressed, embedded network no longer learn useful information for specific few-shot SED task. Therefore, we constrain intra-class distance to avoid overfitting the support set. hen average intra-class distance is less than a distance constraint, fine-tuning process is terminated. Namely, $\frac{dist(S_c)+dist(S_{c'})}{K\times(K-1)} < \eta$, where η is the distance constraint and K is the number of audio samples per class. In this report, we set K as 5.

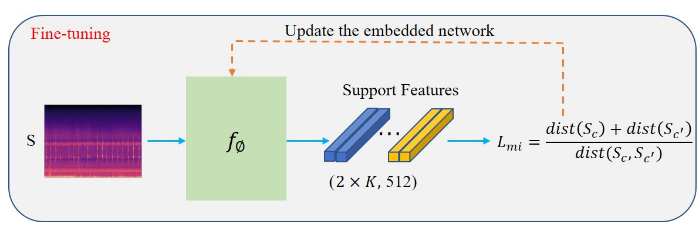

Figure 1: The overview of MIMI.

3. EXPERIMENTS

3.1. Experimental Setups

Dataset. The dataset is from DCASE2022 task 5 development and evaluation sets [11].
Metrics. For all experiments, we use event-based F-measure (F1 score) [12] as evaluation metric, which is the most commonly used metrics in sound event detection.
Preprocessing. We sample all audio clips(recordings) with 22.05 kHz sampling rate and apply Short Time Fourier Transform (STFT) with a window size of 1024 and a hope size of 256 to extract spectrograms. Then, Per-channel energy normalization (PCEN) is used in spectrograms to improve the robustness to channel distortion. Next 128 Mel filter banks are applied on the spectrograms to obtain Mel spectrograms. The audio frames are normalized on the training set with zero mean and unit variance distribution.
Model. we submit four model with different distance constraint and learning rate (lr) during fine-tuning process. Detail setting as shown in Table 1.

Table 1: Detail setting of distance constraint (η) and learning rate (lr) during fine-tuning

Model	η	lr
Model 1	0.30	5×10^{-3}
Model 2	0.30	1×10^{-2}
Model 3	0.45	5×10^{-3}
Model 4	0.45	1×10^{-2}

3.2. Experimental Results

Table 2 shows the experimental results of four models on validation set, which indicate our proposed framework is very useful. Model 1 achieves 46.51% F1 score, which is significantly outperforms Baseline. The performance of different distance constraint and learning rate are different, which demonstrates the necessity for setting thresholds.

Table 2: The 5-shot sound event detection performance on validation set.

Mode	Precision(%)	Recall(%)	F1(%)
Baseline	36.34	24.96	29.59
Model 1	53.02	41.42	46.51
Model 2	54.75	38.99	45.55
Model 3	45.33	43.21	44.25
Model 4	46.58	41.99	44.17

4. CONCLUSIONS

In this report, we propose a meta-learning framework for few-shot sound event detection. Targeting the limitations of specific few-shot sound event detection tasks, we introduce MIMI optimization criteria to continuously fine-tune embedded network. MIMI makes embedded network learn more discriminative embedding features for unseen classes. Such embedding features contribute to classify new sound events. In addition, a distance constraint is designed to constrain fine-tuning process, which aims to avoid overfitting of embedded network.

5. REFERENCES

[1] P. Maijala, Z. Shuyang, T. Heittola, and T. Virtanen, "Environmental noise monitoring using source classification in sensors," *Applied Acoustics*, vol. 129, pp. 258–267, 2018.

[2] R. Cai, L. Lu, A. Hanjalic, H.-J. Zhang, and L.-H. Cai, "A flexible framework for key audio effects detection and auditory context inference," *IEEE Transactions on audio, speech, and language processing*, vol. 14, no. 3, pp. 1026–1039, 2006.

[3] M. Crocco, M. Cristani, A. Trucco, and V. Murino, "Audio surveillance: A systematic review," *ACM Computing Surveys (CSUR)*, vol. 48, no. 4, pp. 1–46, 2016.

[4] B. Zhang, X. Li, Y. Ye, Z. Huang, and L. Zhang, "Prototype completion with primitive knowledge for few-shot learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3754–3762.

[5] D. Wertheimer, L. Tang, and B. Hariharan, "Few-shot classification with feature map reconstruction networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 8012–8021.

[6] G. Li, V. Jampani, L. Sevilla-Lara, D. Sun, J. Kim, and J. Kim, "Adaptive prototype learning and allocation for few-shot segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 8334–8343.

[7] 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)*. IEEE, 2020, pp. 76–80.

[8] 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.

[9] 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.

[10] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," *Advances in neural information processing systems*, vol. 30, 2017.

[11] Nolasco, Ines, Singh, Shubhr, Strandburg-Peshkin, *et al.*, "Dcase 2022 task 5: Few-shot bioacoustic event detection development set," 2022.

[12] A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for polyphonic sound event detection," *Applied Sciences*, vol. 6, no. 6, p. 162, 2016.