SOUND EVENT DETECTION ENHANCED BY SCENE INFORMATION FOR DCASE CHALLENGE 2024 TASK4

Technical Report

Wen Huang[1]*, Bing Han*[1]*, Xie Chen*[1]*, Pingyi Fan*[2]*, Cheng Lu*[3]*, Zhiqiang Lv*[4]*, Jia Liu*[2,4]*,*
Wei-Qiang Zhang[2], Yanmin Qian[1]

[1] Shanghai Jiao Tong University, Shanghai, China
[2] Tsinghua University, Beijing, China
[3] North China Electric Power University, Beijing, China
[4] Huakong AI Plus Company Limited, Beijing, China
holvan@sjtu.edu.cn

ABSTRACT

In this technical report, we describe our submission to the DCASE 2024 Challenge Task 4: Sound Event Detection with Heterogeneous Training Data and Potentially Missing Labels. Our approach leverages a Convolutional Recurrent Neural Network (CRNN) architecture enhanced with pre-trained BEATs embeddings to perform robust sound event detection. To effectively utilize different sources of data, we integrate scene information to enhance event detection performance through multi-task learning. Additionally, we address the challenge of partially missing labels by employing a semi-supervised strategy that combines the mean teacher model with pseudo-labeling to improve performance. Our final ensemble system achieves a PSDS1 score of 0.545 on the DESED validation set and an mpAUC score of 0.759 on the MAESTRO real validation set. These results highlight the efficacy of incorporating scene information and semi-supervised learning strategies in sound event detection tasks with heterogeneous and incomplete datasets.

***Index Terms*—** DCASE, sound event detection, semi-supervised learning, multi-task learning

1. INTRODUCTION

Sound event detection (SED) involves identifying and localizing sound events within an audio recording. The objective is to provide both the event class and the event time boundaries, accommodating the presence of multiple, overlapping events. This task is particularly challenging as it requires leveraging training data with varying annotation granularity, including differences in temporal resolution and the presence of soft or hard labels.

In this year's DCASE challenge task4, systems are evaluated on labels with different granularity to gain a comprehensive understanding of system behavior and assess robustness across various applications. One of the key challenges is the differing target classes across datasets, where sound labels present in one dataset may not be annotated in another. Consequently, systems must handle potentially missing target labels during training and perform effectively without knowing the origin of the audio clips during evaluation.

To address these challenges, our approach integrates scene information to enhance event detection performance through multi-task learning. By incorporating scene context, we aim to improve the system's ability to generalize across different environments. Furthermore, to tackle the issue of partially missing labels, we employ a semi-supervised strategy that combines the mean teacher model with pseudo-labeling. This approach allows us to utilize both labeled and unlabeled data, ultimately improving the performance and robustness of our SED system.

2. DATASETS

DESED. The DESED dataset [1, 2] comprises 10-second audio clips recorded in domestic environments or synthesized to simulate such settings, focusing on 10 sound event classes derived from AudioSet [3]. It includes a weakly labeled training set with 1578 clips, an unlabeled in-domain training set with 14412 clips, and a synthetic strongly labeled set with 10000 clips. Additionally, we incorporate 3463 strongly-annotated audio clips from AudioSet Strong [3], which shares the same sound event classes as DESED.

MAESTRO real. The MAESTRO real dataset [4] consists of approximately 3-minute real-life recordings from 5 acoustic scenes. It was annotated using Amazon Mechanical Turk, allowing for the estimation of soft labels based on multiple annotator opinions.

Due to the differing event classes in these datasets, we employ a method consistent with the official baseline during training to map or mask the class labels accordingly.

3. METHODS

3.1. Data preprocessing

All audio recordings are resampled to 16 kHz and converted to mono. We extract log-mel spectrograms as acoustic features, utilizing 128 mel bands. The Short-Time Fourier Transform (STFT) is computed with a window size of 2048 and a hop length of 256. The frequency range is set between 0 Hz and 8000 Hz.

To enhance the robustness of our model, we apply data augmentation techniques such as mixup [5] and SpecAugment [6]. Mixup involves combining pairs of audio samples and their corresponding labels to generate new training examples, which helps improve the model's generalization. Notably, mixup is performed only within the same dataset (e.g. only within MAESTRO and DESED). SpecAugment is applied to the spectrograms, introducing random

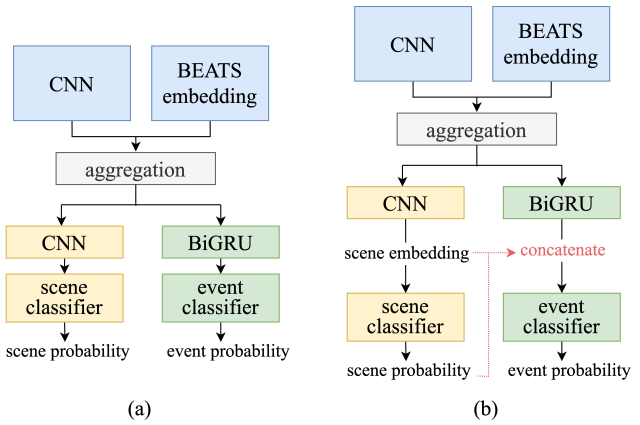

Figure 1: Structure of multi-task learning for scene and event detection. Panel (a) illustrates the fundamental multi-task learning architecture, while panel (b) demonstrates the architecture enhanced by concatenating scene information.

time and frequency masking to prevent overfitting and improve the model's ability to handle various acoustic conditions.

3.2. Semi-supervised learning

To address partially missing labels, our approach combines the mean teacher method with pseudo labeling. Initially, we employ the mean teacher method [7]: training a student network on a combination of labeled and unlabeled data, updating its parameters periodically to match those of a slowly updated teacher network. Subsequently, we generate pseudo labels from the trained model and utilize them in the subsequent stage to further train the model. This iterative process helps leverage the benefits of both consistency regularization through mean teachers and the additional information provided by pseudo labels, enhancing the robustness and performance of the model in semi-supervised learning scenarios.

3.3. Multi-task learning

Apart from the challenge of missing labels, heterogeneous datasets present additional obstacles such as domain shifts and varying event distributions. To address these issues, we employ multi-task learning for scene and event detection.

The MAESTRO real dataset encompasses recordings from five distinct scenes: cafe/restaurants, city centers, grocery stores, metro stations, and residential areas. While the DESED dataset lacks explicit scene attributes, we categorize its data based on event characteristics into a generalized scene category, 'home'. Recognizing the inherent relationship between scene and event information, we aim to leverage scene information to enhance event learning and improve generalization across different datasets. Inspired by the scene-dependent acoustic event detection proposed in [8], we design a joint learning framework that integrates both scene and event information.

As illustrated in Fig. 1, our model consists of three main components: (1) a shared audio encoder incorporating CNN and BEATs embeddings; (2) an event detection branch featuring a BiGRU-based RNN and an event classifier trained using cross-entropy loss L_{scene}; and (3) a scene classification branch comprising a CNN and a scene classifier trained with an event loss defined as:

$$L_{event} = L_{supervised} + L_{self\_supervised}, \quad (1)$$

here self-supervised loss represents the consistency loss in the mean teacher method. The overall loss is formulated as

$$L_{total} = L_{event} + L_{scene} * \alpha, \quad (2)$$

Directly employing multi-task learning can aid in distinguishing audio scenes, thereby facilitating the differentiation of various events. Additionally, we explore the concatenation of scene embeddings or probabilities with event embeddings. This approach aims to leverage scene information to enhance event detection capabilities, enabling the model to benefit from the contextual cues provided by the audio environment when identifying specific events.

4. RESULTS

Table 1 presents the results of single systems on the DESED and MAESTRO validation sets. The test results indicates that pseudo labeling significantly enhances performance, as evidenced by the improved results in stage 2 compared to stage 1. Moreover, the incorporation of multi-task learning further boosts performance. Specifically, the model with scene embedding concatenation (concat_emb) achieved the highest scores, outperforming other configurations. This suggests that integrating scene information into the event detection process provides a valuable context, leading to more accurate predictions across both validation sets. The sum metric, representing the combined performance on PSDS1 for the DESED validation set and Seg-mpAUC for the MAESTRO validation set, consistently increased, highlighting the effectiveness of both pseudo labeling and multi-task learning in enhancing model performance.

Table 1: Results of single systems on the DESED and MAESTRO validation sets. The sum represents the total of PSDS1 on the DESED validation set and Seg-mpAUC on the MAESTRO validation set.

System	MTL	Sum	DESED		MAESTRO	
			PSDS1	PSDS2	Seg-F1	Seg-mpAUC
baseline	-	1.167	0.485	0.754	0.592	0.682
stage1	-	1.199	0.505	0.757	0.599	0.694
stage2	-	1.215	0.516	0.768	0.617	0.699
stage2	basic	1.248	0.524	0.771	0.632	0.724
stage2	concat_emb	1.264	0.527	0.775	0.645	0.737
stage2	concat_prob	1.249	0.523	0.767	0.636	0.726

Besides, we also provide our final submission system results on Table 2, including 3 ensemble systems and 1 single systems.

Table 2: Results of final submission systems on the DESED and MAESTRO validation sets.

System	Ensemble	Sum	DESED		MAESTRO	
			PSDS1	PSDS2	Seg-F1	Seg-mpAUC
1	7 models	1.304	0.545	0.791	0.638	0.759
2	10 models	1.299	0.541	0.790	0.636	0.758
3	20 models	1.292	0.545	0.802	0.635	0.747
4	-	1.264	0.527	0.775	0.645	0.737

5. CONCLUSION

This report details our submission to the DCASE 2024 Challenge Task 4 on Sound Event Detection with Heterogeneous Training Data and Potential Label Absences. Our approach integrates a Convolutional Recurrent Neural Network (CRNN) with pre-trained BEATs embeddings to enhance sound event detection robustness. Central to our strategy is the incorporation of scene information via multi-task learning, significantly improving our system's ability to generalize across diverse acoustic environments. Additionally, we addressed label incompleteness using a semi-supervised approach that combines the mean teacher model with pseudo-labeling, effectively leveraging both labeled and unlabeled data. Our ensemble system achieved promising results, with a PSDS1 score of 0.545 on the DESED validation set and an mpAUC score of 0.759 on the MAESTRO real validation set. These outcomes underscore the efficacy of our methods in handling complex datasets and enhancing sound event detection accuracy.

6. REFERENCES

[1] N. Turpault, R. Serizel, A. Parag Shah, and J. Salamon, "Sound event detection in domestic environments with weakly labeled data and soundscape synthesis," in *Workshop on Detection and Classification of Acoustic Scenes and Events*, New York City, United States, October 2019.

[2] R. Serizel, N. Turpault, A. Shah, and J. Salamon, "Sound event detection in synthetic domestic environments," in *ICASSP 2020 - 45th International Conference on Acoustics, Speech, and Signal Processing*, Barcelona, Spain, 2020.

[3] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *Proc. IEEE ICASSP 2017*, New Orleans, LA, 2017.

[4] I. Martín-Morató and A. Mesaros, "Strong labeling of sound events using crowdsourced weak labels and annotator competence estimation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 902–914, 2023.

[5] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2017.

[6] 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," *arXiv preprint arXiv:1904.08779*, 2019.

[7] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," *Advances in neural information processing systems*, vol. 30, 2017.

[8] T. Komatsu, K. Imoto, and M. Togami, "Scene-dependent acoustic event detection with scene conditioning and fake-scene-conditioned loss," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 646–650.