LOCAL AND GLOBAL FEATURES FUSION FOR SOUND EVENT DETECTION WITH HETEROGENEOUS TRAINING DATASET AND POTENTIALLY MISSING LABELS

Technical Report

Haobo Yue, Zehao Wang, Da Mu, Huamei Sun, Yuanyuan Jiang
Zhicheng Zhang, Jianqin Yin,

Beijing University of Posts and Telecommunications, China
{hby, wzhao, da.mu, shm, jyy, zczhang, jqyin}@bupt.edu.cn

ABSTRACT

In this work, we present our submission system for DCASE 2024 Task4 on Sound Event Detection with Heterogeneous Training Dataset and Potentially Missing Labels, where we introduce the BEATs-CRNN interactive systems. Considering that the pre-trained BEATs model predominantly captures global features for the dataset, while the CRNN model focuses on learning local features, this work aims to fuse the middle layer information of the two to enhance the system's feature extraction capabilities. Firstly, we modify the BEATs model and the CRNN model so that the feature extraction of the dataset by the two models is performed at the same stage. Secondly, due to the differing number of layers in CNN and BEATs, we extract intermediate features from both models at regular intervals, interact them through cross-attention, and then feed the resulting features back to the respective models for the feature extration in the subsequent layer. Finally, the final interaction results of the two models are used as the final features for learning. Compared to the baseline system using BEATs embeddings, which achieved 48.3% in PSDS-scenario 1, 49.4% in PSDS-scenario1 (sed score), and 73.7% in mean-pAUC, our BEATs-CRNN interactive system achieves 53.2%, 54.1%, and 76.3%, respectively. The ensemble of the BEATs-CRNN interactive system further improves the PSDS-scenario 1 to 56.4%, the PSDS-scenario1 (sed score) to 57.4% and the mean-pAUC to 75.6%.

Index Terms— BEATs, CRNN, Cross attention

1. INTRODUCTION

Sound event detection (SED), pivotal for various applications such as aiding the hearing impaired, smart environments[1], audio-to-text retrieval[2], voice activity detection[3][4], and audio captioning[5], relies heavily on neural network architectures like convolutional neural networks (CNNs)[6], convolutional recurrent neural networks (CRNNs)[7], and transformers[8]. These sophisticated approaches underscore the versatility and significance of SED in diverse domains, emphasizing the continual evolution and adoption of advanced neural network methodologies. This task, as a follow-up to Tasks 4A and 4B in 2023, aims to unify the settings of these two subtasks by addressing the challenge of providing event classes and event time boundaries in audio recordings where multiple events may overlap, while also exploring how to leverage training data with varying annotation granularities, encompassing temporal resolution and soft/hard labels.

While supervised learning necessitates a substantial volume of data with precise labels, the manual annotation process is prohibitively costly, leading to challenges in acquiring high-quality datasets. To address this issue, various semi-supervised learning techniques have emerged, leveraging weakly labeled data and partially labeled datasets[9][10]. Weakly labeled data typically lacks event timestamps, while unlabeled datasets offer no informative labels. In the realm of semi-supervised learning, pseudo-labeling strategies are commonly employed to handle datasets with incomplete labeling. Researchers have proposed innovative frameworks like the mean teacher (MT) for SED[11], which combines supervised training on labeled data with self-supervised training on both labeled and unlabeled data. The MT framework utilizes a teacher model for predictions, calculating loss between teacher and student predictions to enhance pseudo-labeling accuracy.

Traditional mean teacher methods can encounter substantial difficulties when providing imprecise predictions for unlabeled data.To enhance the stability and robustness of the self-supervised method, we have adopted the Confidence Mean Teacher (CMT)[12] as an improvement over the unstable MT framework. Furthermore, to facilitate improved learning of data features, we have integrated the innovative BEATs-CRNN interactive system into the submission system for DCASE 2024 Task 4 on sound event detection with heterogeneous training datasets and potentially missing labels. The core concept of our approach revolves around leveraging the complementary strengths of the pre-trained BEATs and CRNN models. While the BEATs model excels at extracting global features from the dataset, the CRNN model specializes in capturing local features. This integration ensures comprehensive feature extraction through a robust interaction established between the intermediate layers of the two models.

Firstly, the BEATs and CRNN models are adjusted to synchronize their feature extraction process, ensuring that both models operate at the same stage of data analysis. Considering that CNN has 7 layers and BEATs has 12 layers, we decided that CNN will extract intermediate features every other layer and BEATs every two layers. Secondly, these features are fused through the cross-attention mechanism to promote mutual information exchange. The resulting interactive features are fed back into the model for further feature extraction in subsequent layers to enhance the overall feature representation.

The culmination of this interactive process yields a set of final features that encapsulate the combined strengths of the BEATs and CRNN models. Compared to the baseline system utilizing BEATs embeddings, which achieved PSDS-scenario 1 at 48%, PSDS-scenario1 (sed score) at 49%, and a mean-pAUC of 63.7%, our BEATs-CRNN interactive system demonstrates significant per-

formance improvements, reaching 50.1%, 52.8%, and 76%, respectively.

Moreover, by incorporating an ensemble approach with the BEATs-CRNN interactive system, we further enhance the performance metrics. The ensemble model elevates PSDS-scenario 1 to 56.4%, PSDS-scenario1 (sed score) to 57.4%, and the mean-pAUC to an impressive 75.6%. This collaborative fusion of BEATs and CRNN models not only enhances feature extraction but also underscores the effectiveness of interactive learning paradigms in advancing sound event detection methodologies.

2. METHODS

2.1. Interaction

The DCASE 2024 Task 4 baseline method integrates the pre-trained BEATs model with the CRNN model, leveraging their unique strengths in feature extraction. The BEATs model, with its 12 transformation layers, is adept at capturing overarching patterns in the dataset, emphasizing global characteristics. On the other hand, the CRNN model excels in capturing fine-grained local features. However, the conventional approach of simply concatenating the features from these models is considered too simplistic for effectively merging the local and global aspects of the data.

In response to this challenge, we have developed a more sophisticated strategy. By extracting the intermediate layers of both the BEATs and CRNN models during feature extraction, we facilitate interaction and information exchange between these models before the learning phase begins:

$$Input_C = softmax\left(\left(W_Q F_B\right)\left(W_K F_C\right)\right) W_V F_C \tag{1}$$

$$Input_B = softmax\left(\left(W_Q F_C\right)\left(W_K F_B\right)\right) W_V F_B \tag{2}$$

where F_B is the middle layer feature of beats model, F_C is the middle layer feature of CNN model, and the calculated results of the two models are sent to the next layer of the two models for feature extraction. $Input_B$ is the input of the next layer of BEATs model, and $Input_C$ is the input of the next layer of CNN model, W_Q,W_K and W_V is the learned projection matrix.

This strategic intervention ensures that the BEATs model retains access to local details, preserving fine-grained information, while also encouraging the CRNN model to consider broader global contexts.This intermediary interaction not only enhances feature representation but also fosters a deeper understanding of the dataset's complexities, potentially enhancing performance in sound event detection tasks.

2.2. Confident Mean Teacher

Traditional mean teacher methods can encounter substantial difficulties when providing imprecise predictions for unlabeled data. In order to address the issue of inaccurate pseudo-labels, we employ the Confidence Mean Teacher (CMT) method. The core principle of CMT involves rectifying erroneous predictions made by the teacher model through post-processing, thereby training the student model with labels of high confidence.

In the CMT framework, we initially acquire clip-level and frame-level predictions from the teacher model. These predictions are processed based on predefined thresholds: clip-level predictions are binary-mapped to 1 or 0 depending on the threshold, and frame-level predictions undergo a similar classification. Following the threshold application, we refine the frame-level predictions using event-specific median filters. This refinement process boosts the dependability of the pseudo-labels, reducing the student model's susceptibility to overfitting based on these labels.

Specifically, these procedures can be articulated as:

$$\tilde{y}_{\mathrm{c}}(k) = \mathbb{I}(\hat{y}_c(k) > \phi_{\mathrm{c}}) \tag{3}$$

$$\tilde{y}_f(t,k) = \mathrm{MF}(\mathbb{I}(\hat{y}_c(k) > \phi_{\mathrm{c}})\mathbb{I}(\hat{y}_f(t,k) > \phi_{\mathrm{f}})) \tag{4}$$

where, $\tilde{y}_c(k)$ represents the clip-level output from the teacher model, $\hat{y}_{\mathrm{c}}(k)$ denotes the clip-level prediction, ϕ_{c} stands for the clip-level threshold, MF signifies the median filter, ϕ_{f} stands for the frame-level threshold, $\tilde{y}_f(t,k)$ represents the frame-level output from the teacher model, and $\hat{y}_f(t,k)$ are the frame-level prediction from the teacher model, respectively.

Moreover, we incorporate confidence-weighted consistency losses based on prediction probabilities. These losses encompass clip-level and frame-level components. By leveraging confidence weights, we train the student model using high-confidence pseudo-labels to mitigate the impact of inaccurate pseudo-labels during training. The consistency losses are defined as:

$$L^{c}_{\mathrm{con}} = \sum_{k \in K} \epsilon(\tilde{y}_c(k), f_{\theta_c}(k)) \tag{5}$$

$$L^{f}_{\mathrm{con}} = \sum_{t,k \in T,K} \epsilon(\tilde{y}_f(t,k), f_{\theta_f}(t,k)) \tag{6}$$

where, K represents the number of sound event categories, T signifies the number of frames, ϵ indicates binary cross-entropy loss, f_{θ_c} denotes the clip-level prediction of the student model, f_{θ_f} denotes the frame-level prediction of the student model.

By incorporating these advancements, the Confident Mean Teacher (CMT) method not only bolsters the robustness of the training process but also enhances the student model's capacity to generalize effectively in diverse scenarios, ultimately improving the overall performance and adaptability of the sound event detection system.

3. EXPERIMENT

3.1. Dataset

This task is based on the DESED dataset and the MAESTRO Real dataset. DESED dataset has been used since DCASE 2020 Task 4. DESED is composed of 10 sec audio clips recorded in domestic environments (taken from AudioSet) or synthesized using Scaper[13] to simulate a domestic environment. DESED concentrates on 10 sound event classes, which form a subset of AudioSet. It's important to note that while some classes in DESED align with those in AudioSet, there are instances where DESED groups together several classes from AudioSet.

The second dataset, MAESTRO Real, comprises real-life recordings, each lasting about 3 minutes and captured in various acoustic environments.These audio recordings were annotated through Amazon Mechanical Turk, employing a method that enables the derivation of nuanced labels based on the collective opinions of multiple annotators.

During evaluation, systems will undergo assessment using labels of varying granularity to gain a comprehensive understanding of their performance and assess their adaptability across diverse applications. Given that different datasets feature distinct target classes, it is possible that sound labels present in one dataset may

Table 1: Performance of the BEATs-CRNN interactive system

	PSDS1	PSDS1 (sed score)	mean pAUC
Baseline	0.48 +- 0.003	0.49 +- 0.004	0.73 +- 0.007
Ours	**0.53 +/- 0.002**	**0.54 +/- 0.001**	**0.763 +/- 0.001**
Ensemble	**0.56 +/- 0.002**	**0.57 +/- 0.001**	**0.756 +/- 0.003**

not be annotated in another. As a result, systems must be capable of handling potential missing target labels during training. Furthermore, SED system is required to operate without knowledge of the origin of the audio clips during evaluation, emphasizing the need for robust and generalized performance across varied scenarios.

3.2. Experiment setup

We train the whole system for 200 epochs and the learning rate warms up in the first 50 epochs with the initial learning rate of 0.001. The batch size is set to 64. Each training session is deployed on the NVIDIA RTX 3090 and lasts 37 hours The ensemble system is composed of six BEATs-CRNN interactive systems, and it is obtained by training the weight of each BEATs-CRNN interactive system for the final result.

3.3. Results and submissions

Table 1 shows the performance of the commit system. The baseline model initially used both the CRNN and BEATs models. Our system uses BEATs-CRNN interactive and replaces MT with CMT. This allows our BEATs-CRNN interactive system to show significant performance improvements compared to the baseline system embedded with BEATs.This change led to improvements in PSDS1, increasing it from 0.500 to 0.536, while PSDS-scenario1 (sed score) increased from 0.520 to 0.543 and mean pAUC from 0.637 to 0.763

The integrated system has shown substantial improvement across the three indicators. Specifically, the PSDS-scenario1 has increased to 56.4%, the PSDS-scenario1 (sed score) has risen to 57.4%, and the mean-pAUC has improved to 75.6%.

4. CONCLUSION

In this study, we presented the BEATs-CRNN interactive system for Sound Event Detection in the DCASE 2024 Task4. By integrating the middle layer features of the BEATs and CRNN models, we ensured comprehensive feature extraction. Our system outperformed the baseline using BEATs embeddings, achieving 50.1%, 52.8%, and 76% for different evaluation metrics. Ensemble of our system further improved performance to 52.5% and 53.4% for different scenarios and 78% for mean-pAUC.

5. REFERENCES

[1] J. P. Bello, C. Mydlarz, and J. Salamon, "Sound analysis in smart cities," 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:64923745

[2] S. Lou, X. Xu, M. Wu, and K. Yu, "Audio-text retrieval in context," 2022.

[3] H. Dinkel, S. Wang, X. Xu, M. Wu, and K. Yu, "Voice activity detection in the wild: A data-driven approach using teacher-student training," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1542–1555, 2021.

[4] H. Dinkel, Y. Chen, M. Wu, and K. Yu, "Voice activity detection in the wild via weakly supervised sound event detection," 2020.

[5] M. Wu, H. Dinkel, and K. Yu, "Audio caption: Listen and tell," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, May 2019. [Online]. Available: http://dx.doi.org/10.1109/ICASSP.2019.8682377

[6] J. Yan, Y. Song, L.-R. Dai, and I. McLoughlin, "Task-aware mean teacher method for large scale weakly labeled semi-supervised sound event detection," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 326–330.

[7] H. Dinkel, M. Wu, and K. Yu, "Towards duration robust weakly supervised sound event detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, p. 887–900, 2021. [Online]. Available: http://dx.doi.org/10.1109/TASLP.2021.3054313

[8] K. Li, Y. Song, L.-R. Dai, I. McLoughlin, X. Fang, and L. Liu, "Ast-sed: An effective sound event detection method based on audio spectrogram transformer," 2023.

[9] J. Ebbers and R. Haeb-Umbach, *Pre-Training And Self-Training For Sound Event Detection In Domestic Environments*, 2022.

[10] J. W. Kim, S. W. Son, Y. Song, H. K. Kim, I. H. Song, and J. E. Lim, "Semi-supervsied learning-based sound event detection using freuqency dynamic convolution with large kernel attention for dcase challenge 2023 task 4," 2023.

[11] C. Plapous, "Mean teacher with data augmentation for dcase 2019 task 4," 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:221088318

[12] S. Xiao, X. Zhang, and P. Zhang, "Multi-dimensional frequency dynamic convolution with confident mean teacher for sound event detection," 2023.

[13] J. Salamon, D. MacConnell, M. Cartwright, P. Li, and J. P. Bello, "Scaper: A library for soundscape synthesis and augmentation," in *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2017, pp. 344–348.