Multimodal Semi-supervised Acoustic Scene Classification with Label Smoothing and Hard Samples Identification

Bagus Tris Atmaja*, Debrina Veisha Rashika§ and Sakriani Sakti*
§ Bandung Institute of Technology, Indonesia
E-mail:13522025@std.stei.itb.ac.id
* Nara Institute of Science and Technology, Japan
E-mail: bagus.tris@naist.jp, ssakti@is.naist.jp

***Abstract*—We propose a multimodal semi-supervised acoustic scene classification framework that combines label smoothing and hard sample identification to enhance model generalization and robustness. Experimental results on the Chinese Acoustic Scene dataset show that our method achieves the highest classification performance, consistently surpassing baseline systems in both simulated data with artificially generated test split and full-dataset experiments on the training split.**

I. Introduction

Acoustic Scene Classification (ASC) has emerged as a pivotal task in environmental sound analysis, enabling machines to recognize and interpret diverse scenes such as streets, parks, or public transport hubs. ASC systems are fundamental for a wide range of applications, including intelligent surveillance, context-aware devices, and smart city infrastructures. Despite significant advances driven by deep learning, the performance of ASC models is often constrained by the scarcity of labeled data and the inherent complexity of real-world acoustic scenes. A challenge is conducted yearly to foster research in this area, along with sound event detection [1], providing a benchmark dataset that includes various acoustic scenes.

Recent research has explored semi-supervised learning [2], crosstask [3], domain shifting [4], and multimodal [5] approaches to address these challenges. Semi-supervised learning leverages both labeled and unlabeled data, mitigating the reliance on extensive manual annotation. Multimodal systems, on the other hand, integrate complementary information from various feature representations, enhancing model robustness and generalization. However, two persistent issues remain: overconfidence in model predictions and the presence of hard-to-classify samples that can hinder effective learning.

In this work, we propose a multimodal semi-supervised ASC framework that incorporates label smoothing and hard sample identification. Label smoothing regularizes the learning process by preventing the model from becoming overly confident, thereby improving generalization. Hard sample identification dynamically emphasizes challenging samples during training, enabling the model to focus on ambiguous or confusing cases. We coupled two aforementioned strategies with a previous fully convolutional neural network (FCNN) [6]. We evaluate our approach on the Chinese Acoustic Scene (CAS) 2023 dataset, demonstrating that the integration of these strategies leads to substantial improvements in classification accuracy and robustness. Our results highlight the effectiveness of combining regularization and adaptive sample weighting in advancing the state-of-the-art for ASC tasks.

II. Methods

A. Datasets

The dataset evaluated in this Chinese Acoustic Scene (CAS) 2023 dataset **CAS2023** [7]. It consists of 8,700 audio samples recorded in various environments, including ”Bus”, ”Airport”, ”Metro”, ”Restaurant”, ”Shopping mall”, ”Public square”, ”Urban park”, ”Traffic street”, ”Construction site”, and ”Bar”. Each sample is labeled with the corresponding acoustic scene category; however, only 20% of the samples are annotated with fine-grained labels (1740 samples). Since the test data is not available, we split the labeled training data into training and test sets (870 samples each) to evaluate the performance of our proposed method on unseen data. Hence, two kinds of data from the same dataset are evaluated:

- Simulation data: The labeled training data is split into training and test sets, which includes 870 labeled samples for each split (7830 total samples for training);
- Experiment data: This is the full dataset, which includes all 1740 labeled samples as training with no label on the test set (8700 total samples for training).

B. Label Smoothing

Label smoothing is a regularization technique that helps prevent overfitting by softening the target labels during training [8]. Instead of using hard labels (e.g., 1 for the correct class and 0 for all others), label smoothing assigns a small probability to incorrect classes, effectively creating a distribution over classes. This approach encourages the model to be less confident in its predictions, which can lead to better generalization.

We embed label smoothing into the loss function by modifying the standard cross-entropy loss to a label-smoothed cross-entropy loss. The standard cross-entropy loss with K classes is defined as:

$$\mathcal{L}_{CE} = -\sum_{i=1}^{N}\sum_{k=1}^{K} y_{i,k} \log p_{i,k}$$

where:

- N is the number of samples
- K is the number of classes
- $y_{i,k}$ is the one-hot encoded label (1 if sample i belongs to class k, 0 otherwise)
- $p_{i,k} = \frac{\exp(z_{i,k})}{\sum_{j=1}^{K}\exp(z_{i,j})}$ is the predicted probability (softmax output)
- $z_{i,k}$ is the logit for class k

Label smoothing modifies the target labels to create a smoothed version $\tilde{y}_{i,k}$, where α is the smoothing parameter (typically between 0 and 1). The smoothed target probabilities are defined as:

$$\tilde{y}_{i,k} = \begin{cases} 1 - \alpha + \frac{\alpha}{K} & \text{if } k = k_i \text{ (true class)} \\ \frac{\alpha}{K} & \text{if } k \neq k_i \text{ (other classes)} \end{cases}$$

Alternatively, this can be expressed as:

$$\tilde{y}_{i,k} = (1-\alpha)y_{i,k} + \frac{\alpha}{K}$$

where $\tilde{y}_{i,k}$ represents the smoothed target probability.

The label smoothed cross-entropy loss becomes:

$$\mathcal{L}_{LS} = -\sum_{i=1}^{N}\sum_{k=1}^{K} \tilde{y}_{i,k} \log p_{i,k}$$

Substituting the smoothed targets:

$$\mathcal{L}LS = -\sum i = 1^N \left[(1-\alpha)\log p_{i,k_i} + \frac{\alpha}{K}\sum_{k=1}^{K}\log p_{i,k}\right]$$

This can be decomposed into two terms:

$$\mathcal{L}LS = (1-\alpha)\mathcal{L}CE + \alpha\mathcal{L}_{uniform}$$

where:

- $\mathcal{L}_{CE}$ is the standard cross-entropy loss.
- $\mathcal{L}_{uniform} = -\frac{1}{K}\sum_{k=1}^{K}\log p_{i,k}$ encourages a uniform distribution.

We set $\alpha = 0.1$ for label smoothing, which means that the model will be less confident in its predictions, leading to better generalization.

C. Hard Samples Identification

Figure 1 illustrates the process of identifying hard samples in the training dataset. The goal is to enhance the model's performance by focusing on samples that the model finds challenging to classify correctly. The process starts from converting raw audio into log mel spectrograms as input features to the pre-trained model.

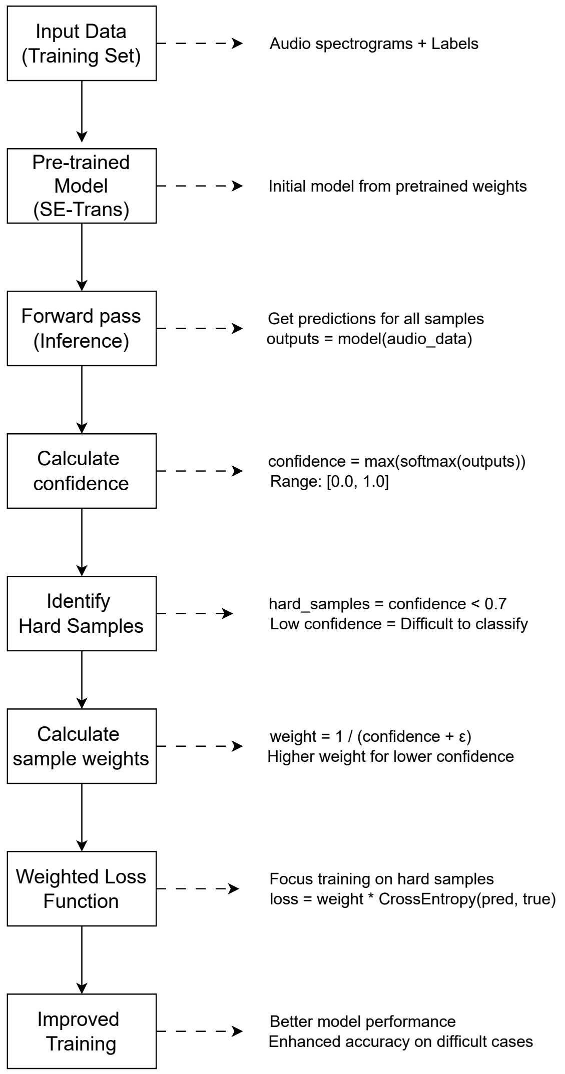
Input Data (Training Set)
Audio spectrograms + Labels
Pre-trained Model (SE-Trans)
Initial model from pretrained weights
Forward pass (Inference)
Get predictions for all samples
outputs = model(audio_data)
Calculate confidence
confidence = max(softmax(outputs))
Range: [0.0, 1.0]
Identify Hard Samples
hard_samples = confidence < 0.7
Low confidence = Difficult to classify
Calculate sample weights
weight = 1 / (confidence + ε)
Higher weight for lower confidence
Weighted Loss Function
Focus training on hard samples
loss = weight * CrossEntropy(pred, true)
Improved Training
Better model performance
Enhanced accuracy on difficult cases

Fig. 1. Hard samples identification process

To identify hard samples within the training data, we first run inference on all training samples using the current pre-trained SE-Trans model. For each sample, we calculate the confidence score, defined as the maximum softmax probability output by the model (instead of 1 - uncertainty as in [9]). Samples with confidence scores below a threshold of 0.7 are considered ”hard” samples, as the model is less certain about their classification. The indices of these hard samples are then collected for further analysis or targeted training.

For each training sample, we first calculate the confidence score as the maximum softmax probability output by the model. These confidence scores are then converted into sample weights using the equation 1. To ensure comparability across samples, the resulting weights are normalized such that their mean equals 1.0. This process yields a tensor of sample weights that can be used to emphasize hard samples during subsequent training.

Mathematically, for a sample i with confidence c_i, the weight is calculated as:

$$\text{weight}_i = \frac{1}{c_i + \epsilon} \tag{1}$$

where $c_i \in [0, 1]$ is the maximum softmax probability, and $\epsilon = 1 \times 10^{-8}$ prevents division by zero. This approach assigns higher weights to samples with lower confidence, emphasizing hard samples during training.

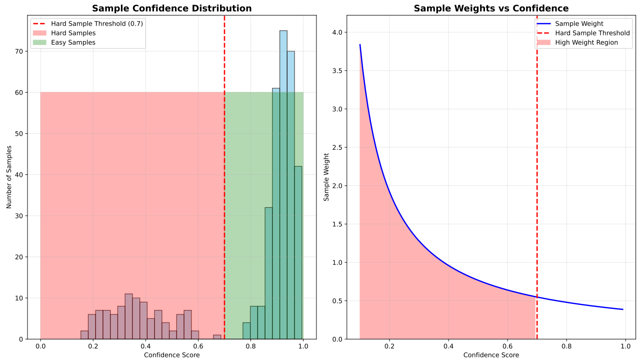

Fig. 2. Confidence score distribution and weights of training samples

D. FCNN

The proposed FCNN architecture (based on [6]), as shown in Fig. 3, processes audio spectrograms through a series of convolutional blocks to extract hierarchical features, followed by a Transformer encoder for global context modeling. The initial layers (Block1-Block3) progressively increase channel depth (144 → 288 → 576) while reducing spatial dimensions via pooling. Block1 consists of two convolutional layers with kernel sizes of 5×5 and 3×3, both followed by Batch Normalization and ReLU activation, while Block2 and Block3 use 3×3 kernels throughout with the same normalization and activation scheme. Block3 employs four convolutional layers with Batch Normalization, ReLU activation, and a dropout rate of 0.2 for regularization. A global average pooling layer condenses the features into a 128-dimensional embedding, which is then refined by a Transformer encoder leveraging multi-head self-attention before passing through a fully connected layer for classification.

Besides the three strategies above, we also evaluated test-time adaptation [10] and data augmentation (as used in [2]). While the former decreases the model's performance, adding the latter with the previous three strategies attains the same performance. Therefore, we only report the results of the three strategies above.

E. Proposed Systems

We evaluated four systems to assess the effectiveness of each part of the methods. For the system #1, we applied the baseline model without any enhancements. System #2 incorporated FCNN, while System #3 added hard sample identification. Finally, System #4 combined System #2 with label smoothing for a comprehensive evaluation. As depicted in Figure 4, all systems follow a sequential flow, starting with the raw audio input, followed by feature extraction. Pre-trained models are fed with the extracted features to predict unlabeled data. The model from that step is saved as best model 1 and used to train the whole data. The best model from this step is saved as best model 2, which is then used to predict the test data. The final output is the predicted labels for the test data.

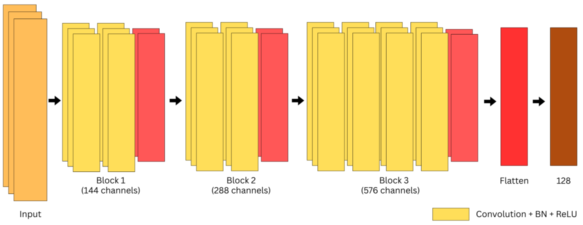

Fig. 3. FCNN architecture

III. Results and Discussion

TABLE I
Performance comparison (balanced accuracy) of different systems in Simulation and Experiment Data

System	Smoothing	Hard	FCNN	Train	Pseudo	Test
			Simulation			
#1	×	×	×	0.920	0.961	0.936
#2	×	×	✓	0.845	0.898	0.824
#3	×	✓	✓	0.939	0.958	0.974
#4	✓	✓	✓	0.947	**0.966**	**0.978**
			Experiment			
#1	×	×	×	0.966	0.982	-
#2	×	×	✓	0.989	0.985	-
#3	×	✓	✓	0.973	0.976	-
#4	✓	✓	✓	0.989	**0.991**	-

The experimental results demonstrate the progressive improvement in acoustic scene classification performance through the systematic integration of the proposed methodological components. In the simulation setting, where labeled training data is partitioned into training and test subsets, the baseline system (System 1) achieves modest performance with accuracy scores of 0.920, 0.961, and 0.936 for training, pseudo-labeling, and test sets, respectively. Notably, the isolated application of FCNN (System 2) results in performance degradation, reducing test accuracy to 0.824, indicating potential overfitting or architectural incompatibility. However, the incorporation of hard sample identification with FCNN (System 3) substantially enhances generalization capability, elevating test accuracy to 0.974. The comprehensive system (System 4), integrating label smoothing, hard sample identification, and FCNN, achieves the optimal performance with a test accuracy of 0.978, representing a 4.2% improvement over the baseline.

The experimental validation on the complete dataset reveals consistent trends with enhanced absolute performance metrics

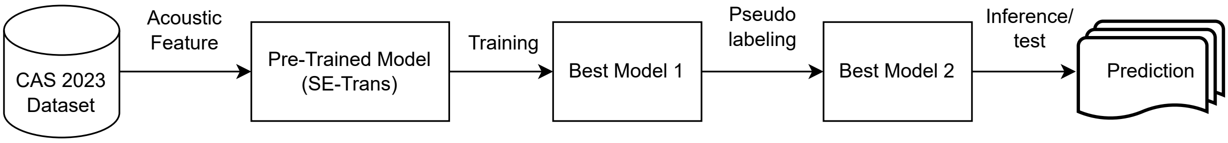

Fig. 4. Flow of the proposed systems

across all system configurations. The baseline system demonstrates improved stability with training and pseudo-labeling accuracies of 0.966 and 0.982, respectively, reflecting the benefit of increased training data availability. System 2 exhibits remarkable improvement to 0.989 training accuracy when FCNN is applied to the full dataset, contrasting sharply with its simulation performance and suggesting that architectural benefits become apparent with sufficient training samples. Systems 3 and 4 maintain competitive performance levels, with the complete methodology (System 4) achieving the highest pseudo-labeling accuracy of 0.991, establishing the synergistic effectiveness of the proposed multimodal semi-supervised learning framework incorporating label smoothing regularization and adaptive hard sample weighting strategies.

IV. Conclusions

Our study highlights the effectiveness of integrating advanced techniques such as hard sample identification and label smoothing in enhancing acoustic scene classification performance. The proposed framework with multimodal data demonstrates significant improvements over baseline models, particularly in challenging scenarios with limited labeled data. Future work will focus on further refining these methods and exploring their applicability to other domains within environmental sound analysis.

Acknowledgment

This paper is partly based on results obtained from projects JSPS KAKENHI Grant Number 24K0296.

References

[1] T. Khandelwal, R. K. Das, and E. S. Chng, “Sound Event Detection: A Journey Through DCASE Challenge Series,” *APSIPA Trans. Signal Inf. Process.*, vol. 13, no. 1, 2024, ISSN: 20487703. DOI: 10 . 1561 / 116 . 00000051.

[2] W. Chen, Y. Liang, Z. Ma, Z. Zheng, and X. Chen, “EAT: Self-Supervised Pre-Training with Efficient Audio Transformer,” *IJCAI Int. Jt. Conf. Artif. Intell.*, pp. 3807–3815, 2024, ISSN: 10450823. DOI: 10.24963/ijcai.2024/421. arXiv: 2401.03497.

[3] J. Bai, J. Chen, M. Wang, M. S. Ayub, and Q. Yan, “A Squeeze-and-Excitation and Transformer-Based Cross-Task Model for Environmental Sound Recognition,” *IEEE Trans. Cogn. Dev. Syst.*, vol. 15, no. 3, pp. 1501–1513, 2023. DOI: 10.1109/TCDS.2022.3222350.

[4] W. Huang, A. Jiang, B. Han, *et al.*, “Semi-Supervised Acoustic Scene Classification with Test-Time Adaptation,” in *2024 IEEE Int. Conf. Multimed. Expo Work.*, IEEE, Jul. 2024, pp. 1–5, ISBN: 979-8-3503-7981-5. DOI: 10.1109/ICMEW63481.2024.10645362.

[5] Z. Li, C. Zhang, X. Wang, *et al.*, “3DMIT: 3D Multi-Modal Instruction Tuning for Scene Understanding,” in *2024 IEEE Int. Conf. Multimed. Expo Work.*, IEEE, Jul. 2024, pp. 1–5, ISBN: 979-8-3503-7981-5. DOI: 10.1109/ICMEW63481 . 2024 . 10645462. [Online]. Available: https://ieeexplore.ieee.org/document/10645462/.

[6] Q. Wang, G. Zhong, H. Hong, *et al.*, “The NERCSLIP-USTC System for Semi-Supervised Acoustic Scene Classification of ICME 2024 Grand Challenge,” in *2024 IEEE Int. Conf. Multimed. Expo Work.*, IEEE, Jul. 2024, pp. 1–4, ISBN: 979-8-3503-7981-5. DOI: 10 . 1109 / ICMEW63481.2024.10645399.

[7] J. Bai, M. Wang, H. Liu, *et al.*, “Description on IEEE ICME 2024 Grand Challenge: Semi-supervised Acoustic Scene Classification under Domain Shift,” pp. 4–8, 2024. arXiv: 2402.02694.

[8] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016. [Online]. Available: https://www.deeplearningbook.org/.

[9] B. T. Atmaja, A. Sasou, and F. Burkhardt, “Uncertainty-Based Ensemble Learning for Speech Classification,” in *2024 27th Conf. Orient. COCOSDA Int. Comm. Co-ord. Stand. Speech Databases Assess. Tech.*, IEEE, Oct. 2024, pp. 1–6, ISBN: 979-8-3315-0603-2. DOI: 10.1109/O-COCOSDA64382.2024.10800111.

[10] Y.-f. Z. Xue, W. Kexin, J. Kun, Y. Zhang, and L. Wang, “AdaNPC : Exploring Non-Parametric Classifier for Test-Time Adaptation,” 2021.