SEMI-SUPERVISED ACOUSTIC SCENE CLASSIFICATION UNDER DOMAIN SHIFT USING CONSISTENCY TRAINING AND PSEUDO-LABEL

Anonymous ICME submission

ABSTRACT

Acoustic scene classification is paramount for enhancing human-machine interaction, enabling devices to understand and respond to their acoustic surroundings. The reliance of acoustic scene classification on fully supervised learning faces challenges due to the scarcity and imbalance of labeled data, which limits the model's generalization capabilities. Semi-supervised learning emerges as an effective solution, leveraging unlabeled data to overcome these limitations and enhance model robustness. To exploit these advantages, we have developed a semi-supervised framework for acoustic scene classification that integrates pseudo-label and consistency regularization strategies to harness the complementary strengths of labeled and unlabeled data. Within this framework, the Hidden unit BERT (HuBERT) model is utilized to learn the feature distribution within the inherently complex mixture environments of the ICME dataset. The offline clustering step of HuBERT aids in generating aligned target labels, thereby enriching the prediction loss with nuanced real-world speech features. Through extensive experimentation, we meticulously explored and established the optimal synergistic interaction between HuBERT and our semi-supervised framework, achieving significant advancements in acoustic scene classification accuracy and model adaptability.

Index Terms— Acoustic scene classification, semi- supervised learning, domain shift, Consistency regularization, Pseudo-label

1. INTRODUCTION

With the development of artificial intelligence, the deep-learning based methods achieve promising performance on many tasks about acoustics, facilitating the analysis of audio signals. Acoustic scene classification (ASC) [1], a crucial research issue in computational auditory scene analysis, aims to recognize the unique acoustic characteristics of an environment and identify the scene to which an audio belongs.

Recent methodologies [2, 3, 4] for acoustic scene classification via deep learning exhibit a significant dependency on training dataset. This reliance often results in low generalization as the existence of data bias in training data, which in turn diminish the adaptability of classification algorithms. Consequently, these algorithms trained on insufficient data demonstrate reduced efficacy in predicting acoustic scene categories under real-world scenarios, highlighting a critical challenge in the application of audio analysis methods.

To mitigate the insufficiency and bias of data, more training data are necessary to enhance the generalization ability of classification model. However, the manual annotation of huge training data incurs significant costs. To address this challenge, we introduce a semi-supervised learning (SSL) framework into acoustic scene classification, leveraging some labeled data and large unlabeled data to improve the performance without extra manual labeling.

In this paper, we introduce a semi-supervised training strategy into acoustic scene classification task. And a powerful pre-trained model is introduced into the semi-supervised training strategy as backbone. First, the labeled audio signals are used to endow the model with acoustic classification ability. And then, we design a contrastive strategy to assign pseudo labels to unlabeled audio data. These unlabeled data with pseudo labels facilitate the training of classification model, alleviating the insufficient of annotated acoustic data.

In summary, the main contributions in this paper are as follows:

1. A semi-supervised learning framework are introduced into acoustic scene classification task. In this framework, the labeled data are utilized to train the confident classification model. And then, a contrastive strategy are implemented to unlabeled data to assign pseudo labels to them. These data with pseudo labels enlarge the training data. Under this semi-supervised learning framework, our proposed method learns more compliciate inherent characteristic in audio signal and improve the classification performance.

2. An excellent backbone for audio signal, HuBERT [5], is adapted into acoustic scene classification task. Benefiting from the pre-training on huge audio data, HuBERT has great transferability on diverse downstream tasks about acoustic analysis. Thus, we fine-tune the HuBERT on the ICME Dataset used in this paper. The fine-tuning successfully adapts HuBERT on acoustic scene classification task and gain promising performance.

3. The experimental results show that the introduced semi-supervised learning framework and HuBERT [5]

backbone improve the classification accuracy on acoustic scene classification task effectively. Our proposed method exploits the unlabeled data and pre-trained model sufficiently.

2. METHODS

Figure 1 offers a concise depiction of the proposed method in this paper. The process of our system could be divided into the following three stages:

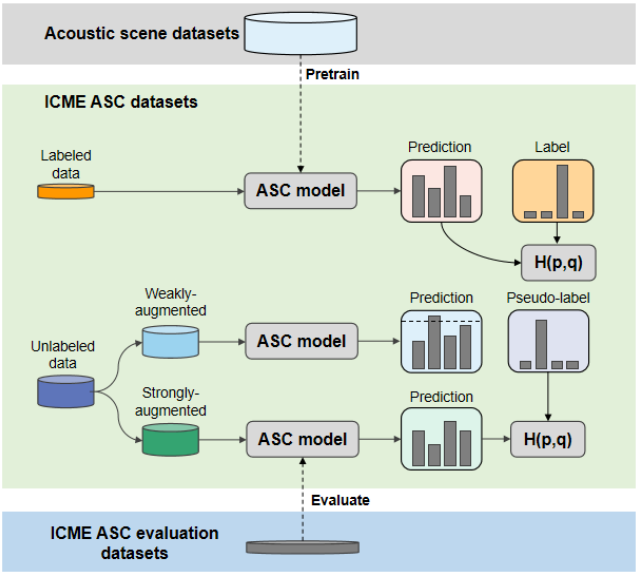

Fig. 1. The pipeline of the proposed method.

2.1. Consistency regularization

For an L-class classification problem, let $\chi = \{(x_b, p_b) : b \in (1, \ldots, B)\}$ be a batch of B labeled examples, where x_b are the training examples and P_b are one-hot labels. Let $U = \{u_b : b \in (1, \ldots, \mu_B)\}$ be a batch of μ_B unlabeled examples where μ is a hyperparameter that determines the relative sizes of χ and U. Let $p_m\,(y|x)$ be the predicted class distribution produced by the model for input x. We denote the cross-entropy between two probability distributions p and q as $H(p, q)$. We perform two types of augmentations: strong and weak, denoted by A and α respectively.

Consistency regularization is an important component of recent state-of-the-art SSL algorithms. Consistency regularization utilizes unlabeled data by relying on the assumption that the model should output similar predictions when fed perturbed versions of the same data. This idea was first proposed in [6] and popularized by [7], where the model is trained both via a standard supervised classification loss and on unlabeled data via the loss function

$$\sum_{b=1}^{\mu_B} \|p_m(y|\alpha(u_b)) - p_m(y|\alpha(u_b))\|_2^2 \qquad (1)$$

Note that both α and p_m are stochastic functions, so the two terms in eq. (1) will indeed have different values. Extensions to this idea include using an adversarial transformation in place of α, using a running average or past model predictions for one invocation of p_m , using a cross-entropy loss in place of the squared ℓ^2 loss, using stronger forms of augmentation, and using consistency regularization as a component in a larger SSL pipeline.

The strategy operates by applying two key processes on the unlabeled data: weak augmentation and strong augmentation. First, each unlabeled data point is subjected to a weak augmentation technique, generating a slightly modified version of the original data. The model then makes predictions on these weakly augmented versions. If the model's prediction for a weakly augmented data point exceeds a certain confidence threshold, that prediction is treated as a pseudo-label for the corresponding strongly augmented version of the same data point. The model is then trained to predict the pseudo-labels of these strongly augmented data points, effectively learning from the unlabeled data by assuming that if it can correctly predict the class of a weakly augmented data point with high confidence, the same prediction should apply to its strongly augmented counterpart. This process allows our method to exploit the unlabeled data efficiently by ensuring that only high-confidence predictions contribute to the model's training, thereby minimizing the risk of reinforcing incorrect predictions.

2.2. Pseudo-label

Pseudo-label leverages the idea of using the model itself to obtain artificial labels for unlabeled data. Specifically, this refers to the use of "hard" labels (i.e., the arg max of the model's output) and only retaining artificial labels whose largest class probability fall above a predefined threshold. Letting $q_b = p_m\,(y|u_b)$, pseudo-label uses the following loss function:

$$\frac{1}{\mu B} \sum_{b=1}^{\mu B} \mathrm{I}(\max(q_b) \geq \tau) H(\hat{q}_b, q_b) \qquad (2)$$

where $\hat{q}_b = \arg\max\,(q_b)$ and τ is the threshold. For simplicity, we assume that arg max applied to a probability distribution produces a valid "one-hot" probability distribution. The use of a hard label makes pseudo-label closely related to entropy minimization, where the model's predictions are encouraged to be low-entropy (i.e., high-confidence) on unlabeled data.

The loss function consists of two cross-entropy loss terms: a supervised loss ℓ_s applied to labeled data and an unsupervised loss ℓ_u. Specifically, $\ell_s$is just the standard cross-entropy

loss on weakly augmented labeled examples:

$$\ell_s = \frac{1}{B}\sum_{b=1}^{B} H(p_b, p_m(y|\alpha(u_b))) \tag{3}$$

$$\ell_u = \frac{1}{\mu B}\sum_{b=1}^{\mu B} \mathrm{I}(\max(q_b) \geq \tau) H(\hat{q}_b, p_m(y|A(u_b))) \tag{4}$$

The loss minimized is simply $\ell_s + \lambda_u \ell_u$ where λ_u is a fixed scalar hyperparameter denoting the relative weight of the unlabeled loss.

2.3. HuBERT Model

HuBERT (Hidden Unit BERT) [5] is a novel approach in the field of speech processing that leverages unsupervised learning techniques to understand and process audio. The method is built upon the foundation of BERT (Bidirectional Encoder Representations from Transformers) [8], a transformer-based model known for its effectiveness in natural language processing tasks. HuBERT distinguishes itself by pre-training on a large corpus of unlabeled audio data, where it learns to predict the masked portions of audio sequences, similar to how BERT predicts masked words in text.

Specifically, the raw audio data is first segmented into smaller, manageable chunks. These segments are then converted into a suitable format for processing, typically involving feature extraction steps such as computing Mel-frequency cepstrum coefficients (MFCCs) or using raw waveform data directly. The extracted features from the audio segments are clustered using k-means clustering. This step aims to group similar sounding segments together, creating a set of discrete units or pseudo-labels that represent different sounds in the audio data.

Similar to the masked language model in BERT [8], portions of the audio segments are masked randomly. The HuBERT model is then trained to predict the pseudo-labels of these masked segments based on the context provided by the unmasked parts. This training utilizes a transformer-based architecture, benefiting from its ability to capture long-range dependencies in data.

After an initial round of training, the model's predictions can be used to refine the clustering of audio features. By using the model's output to re-cluster the audio data, the quality of pseudo-labels is improved, leading to more meaningful distinctions between different sounds. With the refined clusters, the model undergoes another round of masked audio modeling. This iterative process can be repeated multiple times, with each cycle potentially enhancing the model's ability to understand and represent the audio data.

Once pre-trained, the HuBERT model can be fine-tuned for specific speech processing tasks, such as speech recognition, emotion detection, or speaker identification. This involves training the model on a smaller labeled dataset specific to the task at hand, allowing HuBERT to apply its learned representations to achieve high performance on these tasks. Thus, we select HuBERT as backbone in proposed semi-supervised framework.

3. DATASET

3.1. ICME Dataset

The ASC dataset for ICME 2024[1], extracted from the CAS 2023 dataset, encompasses a curated selection of ten quintessential acoustic scenes—Bus, Airport, Metro, Restaurant, Shopping Mall, Public Square, Urban Park, Traffic Street, Construction Site, and Bar—amassing a total auditory experience exceeding 130 hours. The dataset is methodically partitioned into development and evaluation sets, with each segment containing 10-second audio clips for analysis. In the development dataset, 20% of the data with scene labels to aid the development of semi-supervised learning algorithms. The evaluation set, collated from recordings across twelve cities, integrates data from five previously unexposed urban environments. This intentional selection criterion is designed to enhance the robustness and validity of the assessment, thereby offering a substantive evaluation of the domain shift adaptability inherent in the submitted methodologies.

3.2. Partition of data

The development set comprises 8,700 segments, of which 1,740 are labelled. We divide these data into a training set (8,450 segments), a validation set (100 segments), and a test set (150 segments). Both the validation and test sets are composed of labelled data. The residual 1,490 labelled segments are divided into the training set.

4. EXPERIMENT

4.1. Evaluation metric

The performance of all algorithms used in our experiment is assessed using the following metrics, i.e., macro-average accuracy and $F1 - Score$. The macro-average accuracy is calculated as the average of class-wise accuracies by the following expression:

$$Accuracy = \frac{1}{N}\sum_{i=1}^{N} Accuracy_i \tag{5}$$

where N is the number of classes, and $Accuracy_i$ is the accuracy for class i. $F1 - Score$ is computed as:

$$F1 - Score = 2 \times \frac{precision \times recall}{precision + recall} \tag{6}$$

$Accurac$ is computed as:

$$Accurac = \frac{TP + TN}{TP + FP + TN + FN} \tag{7}$$

Table 1. Hyper-parameters of Experiment.

Sampling Rate	44100
Max Length	10.0
Weight Decay	5e-4
Model	HuBERT-Base
Labeled Batch size	2
Unlabeled Batch size	2
Learning Rate	5e-5
Layer Decay Rate	0.75
Scheduler	$\eta = \eta_0 \cos\left(\frac{7\pi k}{16K}\right)$
Model EMA Momentum	0.0
Prediction EMA Momentum	0.999
Weak Augmentation	Random Sub-sample
Strong Augmentation	Random Sub-sample Random Gain Random Pitch Random Speed

$Precision$ is computed as:

$$Precision = \frac{TP}{TP + FP} \tag{8}$$

$Recall$ is computed as:

$$Recall = \frac{TP}{TP + FN} \tag{9}$$

where, TP is true positive, TN is true negative, FP is false positive, and FN is false negative.

4.2. Experiment setup

In our framework, supervised training is conducted for the labelled data; meanwhile, labelled data are additionally included in the unlabeled dataset for semi-supervised learning. A weakly-augmented is fed into the model to obtain predictions. When the model assigns a probability to any class which is above a threshold, the prediction is converted to a one-hot pseudo-label. Then, we compute the model's prediction for a strong augmentation of the same input. The model is trained to make its prediction on the strongly-augmented version match the pseudo-label via a cross-entropy loss.

We adopt HuBert as the pre-trained model. The batch size of labeled data and unlabeled data is set to 2. We keep the sampling rate of audios as 44100. We adopt AdamW optimizer with a weight decay of 5e-4, and search the learning rate and layer decay. Mimicking RandAugment, for strong augmentation in audio tasks, we random sample 2 augmentations from the augmentation pool and random set the augmentation magnitude during training.

Table 2. The results of our model on the test set.

Scene	Accuracy	F1-score
Airport	100%	96.77%
Urban park	100%	98.5%
Public square	100%	100%
Traffic street	100%	100%
Restaurant	93.33%	93.33%
Bus	100%	100%
Shopping mall	93.33%	96.55%
Bar	100%	100%
Metro	93.33%	96.55%
Construction site	100%	100%
Average	98.00%	98.00%

4.3. Result

The ASC performance of the our system for each scene is shown in Table 2. Our system achieves an accuracy of 98.00% and F1-score of 98.00%. Figure 2 shows the confusion matrix of our system. As depicted in Figure 2, our system has excellent recognition performance for each scene.

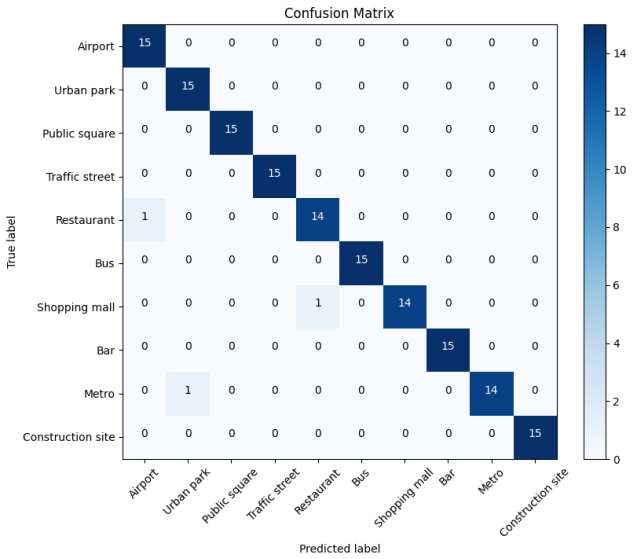

Fig. 2. The confusion matrices on the test set.

5. CONCLUSION

To alleviate the bias in manual annotated data and the deficiency of labeled data, we attempt to utilize unlabeled data and build a semi-supervised learning framework for acoustic scene classification. In this framework, the labeled data are

used to train the classification model, and the unlabeled data are transformed into two variants via strong augmentation and weak augmentation. Then, these two variants from unlabeled data are feed to model and their pseudo labels are assigned by consistency regularization. The data with pseudo labels are further used to train the model. In addition, to ensure the generalization ability of model, we choose the HuBERT backbone in semi-supervised learning framework, and fine-tune this backbone for acoustic scene classification, adapting pre-trained audio model to downstream task effectively. The experiments show that our method obtains competitive results on ICME challenge dataset.

6. REFERENCES

[1] Jisheng Bai, Mou Wang, Haohe Liu, Han Yin, Yafei Jia, Siwei Huang, Yutong Du, Dongzhe Zhang, Mark D. Plumbley, Dongyuan Shi, Woon-Seng Gan, Susanto Rahardja, Bin Xiang, and Jianfeng Chen, "Description on IEEE ICME 2024 grand challenge: Semi-supervised acoustic scene classification under domain shift," *CoRR*, vol. abs/2402.02694, 2024.

[2] Dan Stowell, Dimitrios Giannoulis, Emmanouil Benetos, Mathieu Lagrange, and Mark D. Plumbley, "Detection and classification of acoustic scenes and events," *IEEE Trans. Multim.*, vol. 17, no. 10, pp. 1733–1746, 2015.

[3] Jisheng Bai, Jianfeng Chen, and Mou Wang, "Multimodal urban sound tagging with spatiotemporal context," *IEEE Trans. Cogn. Dev. Syst.*, vol. 15, no. 2, pp. 555–565, 2023.

[4] Jisheng Bai, Jianfeng Chen, Mou Wang, Muhammad Saad Ayub, and Qingli Yan, "A squeeze-and-excitation and transformer-based cross-task model for environmental sound recognition," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 15, no. 3, pp. 1501–1513, 2023.

[5] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 29, pp. 3451–3460, 2021.

[6] Philip Bachman, Ouais Alsharif, and Doina Precup, "Learning with pseudo-ensembles," in *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil D. Lawrence, and Kilian Q. Weinberger, Eds., 2014, pp. 3365–3373.

[7] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li, "Fixmatch: Simplifying semi-supervised learning with consistency and confidence," in *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*. 2019, pp. 4171–4186, Association for Computational Linguistics.