ACOUSTIC SCENE CLASSIFICATION BY THE SELF-LEARNING OF EAT

Wenxi Chen[1], Yuzhe Liang[1], Yihong Qiu[2], Xinhu Zheng[2], Boyuan Chen[2], Xie Chen[1],
Jia Liu[2],Wei-Qiang Zhang[2],Cheng Lu[3]

[1]Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China
[2]Department of Electronic Engineering, Tsinghua University, Beijing, China
[3]School of Economics and Management, North China Electric Power University, Beijing, China

ABSTRACT

This technical report details the framework we use to solve ICME-2024 challenges, and the task is to design an audio scene classification system to distinguish acoustic scenes recorded from different devices, within the constraints of a large number of unlabeled training data. Our architecture is predicated on the robust audio SSL (self-supervised learning) model – EAT, which we pre-train on three comprehensive datasets to capture a rich tapestry of audio scene characteristics. The potency of our approach lies in the semi-supervised methodology that leverages self-learning to bolster the model's generalization capabilities for downstream tasks. This is achieved through iterative fine-tuning and the strategic application of pseudo-labeling, which together refine the model's acumen in scene classification. In the pursuit of optimizing our model's evaluation performance, we employ the test-time adaptation strategy during inference.

Index Terms— acoustic scene classification, self-supervised learning, semi-supervised learning

1. INTRODUCTION

Acoustic Scene Classification (ASC) aims to classify audios into scenes based on their characteristics.ASC models may have distributional differences between training and test data, resulting in a decrease in the model's ability to generalize to real-world applications. This domain bias can be caused by several factors, including different recording devices, different geographical regions, and different cultural and linguistic backgrounds. Addressing this domain bias is crucial for realizing more robust ASC models. In practice, labeled acoustic scene data is often difficult to obtain, while unlabeled acoustic data is relatively abundant. How to effectively utilize these unlabeled data is important for improving the performance of ASC models. Semi-supervised learning techniques can be applied to utilize unlabeled data to enhance the training and generalization capabilities of models. How to solve these problems encountered so far has become the focus of researchers and technicians.

In order to obtain the effective information of audio, self-supervised learning (SSL) model EAT [1] is used to pre-train a large number of unlabeled data. Through self-supervised pre-training, the model can learn useful representations of the audio data, providing better feature extraction for subsequent classification tasks.

Then fine-tune the pre-trained obtained model using labeled acoustic scene data. The pre-trained model is used as the initial model for supervised training of classification tasks using labeled data. Through fine-tuning, the model can learn more discriminative features based on the specific ASC task and improve the model's classification accuracy in specific scenes.

Based on the fine-tuning, a pseudo-labeling approach is used to further utilize the unlabeled data for iterative training. First, pseudo-labels are obtained by inferring the unlabeled data using the fine-tuned model. Using these pseudo-labels as approximate labels for the unlabeled data, the labeled data and the unlabeled data with pseudo-labels are mixed together for training to further optimize the model.

Also, we incorporate a test-time adaptation strategy into our ASC framework. During inference, the model dynamically adjusts its predictions based on the characteristics of the input samples. This adaptation process allows the model to account for the domain shift and capture the specific features present in the testing data.

2. DATA PREPROCESSING AND AUGMENTATION

2.1. Datasets

According to the challenge rules, we utilize the ASC challenge development dataset [2], TAU Urban Acoustic Scenes (UAS) 2020 Mobile development dataset [3] and the CochlScene dataset [4] for model pretraining. No more datasets are used for model training.

TAU UAS. The TAU Urban Acoustic Scenes 2020 Mobile dataset [3] consists of 64 hours of recordings from various acoustic scenes. The recordings are captured in different cities across Europe, using four devices (A, B, C, and D) simultaneously. In order to enhance the diversity of the dataset,

11 simulated devices(S1-S11) are created in the dataset, using synthetic recordings simulated from device A.

CochlScene. The Cochl Acoustic Scene Dataset [4], abbreviated as CochlScene, is an acoustic scene dataset containing 76,115 ten-second audio files from 13 different acoustic scenes. The recordings of the dataset are sourced from crowdsourcing participants in Korea and manually selected to enhance evaluation reliability.

2.2. Feature extraction

In our data preprocessing pipeline, all audio files are standardized to a sample rate of 16,000 Hz. Then spectrograms are generated using a Hanning window of 25 milliseconds with a hop size of 10 milliseconds and an FFT window of 400. Next, the spectrograms are transformed into 128-dimensional log-mel spectrograms.

2.3. Data augmentation

For model training, three data augmentation techniques are majorly employed: SpecAugment [5], Mixup [6] and Freq-MixStyle [7, 8].

SpecAugment. SpecAugment [5] is a data augmentation method for speech recognition. It directly modifies neural network inputs, such as filter bank coefficients. The augmentation policy involves warping features, masking frequency channels, and masking time steps. Applied to Listen, Attend, and Spell networks, it achieves state-of-the-art performance on tasks like LibriSpeech and Switchboard.

Mixup. Mixup [6] is a novel approach for domain generalization. It leverages the probabilistic mixing of instance-level feature statistics from different source domains. Inspired by style transfer research, it implicitly synthesizes new domains, enhancing model generalization. a popular data augmentation technique for training deep neural networks. It generates additional samples by interpolating pairs of inputs and their labels. By doing so, it encourages the model to learn more robust features and improves its ability to generalize to unseen data. In our pipeline, the technique is used to generate more samples of the log mel spectrogram of the audio clips from each batch.

Freq-MixStyle. Freq-MixStyle (FMS) [7, 8] is a frequent version adaptation of the original MixStyle [9] concept. It is a novel approach for domain generalization, leveraging probabilistic mixing of instance-level feature statistics from different source domains. In our pipeline, it normalizes the frequency bands in one spectrogram and then denormalizes them by using the combined frequency statistics from two different spectrograms. In each batch, the technique occurs with a probability determined by the hyperparameter p_{FMS}, and the mixing coefficients are drawn from a Beta distribution shaped by α.

3. METHOD

In our study, we utilized the EAT model [1], an audio self-supervised learning framework, as the foundational model for the audio scene classification tasks. The model was first pre-trained on three key datasets: the ASC Challenge Development Dataset, the TAU Urban Acoustic Scenes Development Dataset, and the CochlScene Dataset, aiming to capture diverse and generalized audio scene representations. After pre-training, we adopted a semi-supervised approach – self-learning, combining iterative fine-tuning with pseudo-labeling, to improve the model's accuracy on a validation set derived from the ASC Challenge Development Dataset. This strategy highlights the efficacy of integrating self-supervised learning with semi-supervised techniques to enhance audio scene classification.

3.1. Self-supervised Pre-training

In the domain of acoustic scene tasks, audio self-supervised learning (SSL) models leverage pretext tasks like masked language modeling (MLM) for pre-training, utilizing a vast corpus of unlabeled data to learn audio features across various scenes and devices. This self-supervised pre-training enables these models to exhibit superior performance in downstream acoustic scene classification tasks. Specifically, we employed the EAT self-supervised model, which adopts the bootstrap self-supervised training paradigm within the audio domain. During its pre-training phase, the Transformer-based EAT model employs the Utterance-Frame Objective (UFO) as a loss function, effectively integrating global utterance-level and local frame-level losses for predicting audio scene representations. It has achieved state-of-the-art (SOTA) performance on audio classification datasets such as AudioSet (AS-2M, AS-20K) and ESC-50, with a pre-training speedup of up to $\sim$15 times compared to pre-existing SOTA audio SSL models. Consequently, for this task, we have opted to utilize the EAT model to more efficiently learn acoustic scene features.

In our experiments, we leveraged the EAT framework for pre-training on three datasets: ASC, TAU, and CochlScene. To tailor the model more closely to the acoustic scenes specific to our task, we employed a weighted pre-training approach, thereby enhancing the model's learning of acoustic scene features present in the ASC data. Notably, we assigned the weights of data from the three datasets in a 1:1:10 ratio. This strategic weighting aims to optimize the model's adaptability and performance on our target acoustic scene classification tasks, demonstrating the efficacy of a fine-tuned, self-supervised learning approach in audio scene analysis.

3.2. Semi-supervised Learning

Given the limited amount of labeled data in the ASC Development Dataset (1,740 labeled instances versus 6,960 unlabeled instances), relying solely on supervised learning with

labeled data poses a challenge in enhancing the model's generalization capabilities. Therefore, we have implemented a self-learning-based semi-supervised learning method to address this limitation, which harnesses the largely untapped potential of the abundant unlabeled data available in the dataset.

The basic implementation process of our semi-supervised learning method unfolds iteratively, incorporating two main stages: fine-tuning and pseudo-labeling. Initially, the pre-trained EAT model is fine-tuned on the available labeled data on ASC with standard cross-entropy loss function as below.

$$L = -\sum_{c=1}^{M} y_{o,c} \log(p_{o,c}) \tag{1}$$

The fine-tuned EAT is then utilized to predict classes of unlabeled ASC data. This prediction process employs a confidence threshold to selectively generate pseudo-labels, ensuring that only labels with a prediction probability above this threshold are retained as hard labels. This approach refines the quality of pseudo-labels, creating an augmented dataset that integrates both the original labeled data and the high-confidence pseudo-labeled data.

Following the pseudo-labeling stage, the model undergoes a fine-tuning process. During fine-tuning, the model is re-trained on this augmented dataset, allowing it to adjust and improve based on a broader set of examples, including those it has pseudo-labeled. This iterative cycle of pseudo-labeling and fine-tuning is repeated, with each iteration aiming to enhance the model's generalization ability by leveraging insights gained from the expanded training dataset.

3.3. Test-time Adaptation

Before the final results, a test-time adaptation method based on k-nearest neighbor (KNN) [10] is employed to minimize the effect of domain shift between the development and the evaluation sets. First, we extract the embeddings of all labeled samples of the development set and store them to form a memory bank for KNN. When inferring, the embedding of each sample in the evaluation set is compared with the embeddings in the memory bank via cosine similarity. According to the distances of the k-nearest neighbors, scoring coefficients are utilized to give the final results. More specifically, let y_j be the label of x_j, and $\mathbf{1}\{y_j\}$ is the one-hot vector of x_j.$\mathcal{M}_L$ is the set of embeddings of all labeled samples in the development set. For each embedding x_i, we use KNN to find k nearest neighbors in $\mathcal{M}_L$ by adopting cosine similarity. The process can be described as follows:

$$w_{ij} = \frac{x_i^T x_j}{\|x_i\|_2 \|x_j\|_2} \tag{2}$$

Then the final result can be given by:

$$\eta(x_i) = \text{Softmax}\left(\sum_{x_j \in N_{\mathcal{M}_L}(x_i)} w_{ij} \mathbf{1}\{y_j\}\right) \tag{3}$$

In addition, the model parameters are not modified during this process, which is within the rules of the challenge.

4. EXPERIMENT

In the pre-training phase, we used 4 GTX3090Ti's and utilized the framework of the EAT model to train 20,000 updates, using the adam optimizer with lr of 0.0005, adam_betas set to [0.9, 0.95], weight_decay of 0.05, and cosine for the learning rate scheduler, under the ema framework, ema_decay is 0.9998, ema_end_decay is 0.99999.

In the finetune stage, we trained 3000 updates with a single GTX3090Ti and then iterated on the unlabeled data inference, filtering out the ones with confidence over 0.85 as pseudo-labels and then put them into the model for training. Mixup and SpecAugment were added to the training, and part of the TAU dataset that overlaps with the ICME challenge was added, and the training of the data here was weighted to ensure that the labels remained relatively balanced.

Because the training set given by the chanllenge is very prone to overfitting, in order to test the impact of each component on the model performance, we used a small-sample training method, dividing a very small portion of the labeled data for finetune to predict the pseudo-labels, in order to find the appropriate hyperparameters.

5. CONCLUSION

In ICME-2024 challenges, we use SSL model to obtain the representation with more information about itself. To obtain a more robust model, we fully utilize the given dataset and use a weighted approach to ensure the direct balance of each dataset, and then we use semi-supervised learning to better utilize the unlabeled data, and apply a test-time adaptation strategy to improve the performance and generalization of the model.

6. REFERENCES

[1] Wenxi Chen, Yuzhe Liang, Ziyang Ma, Zhisheng Zheng, and Xie Chen, "EAT: Self-supervised pre-training with efficient audio transformer," *arXiv preprint arXiv:2401.03497*, 2024.

[2] Jisheng Bai, Mou Wang, Haohe Liu, Han Yin, Yafei Jia, Siwei Huang, Yutong Du, Dongzhe Zhang, Mark D Plumbley, Dongyuan Shi, et al., "Description on ieee icme 2024 grand challenge: Semi-supervised acoustic scene classification under domain shift," *arXiv preprint arXiv:2402.02694*, 2024.

[3] Toni Heittola, Annamaria Mesaros, and Tuomas Virtanen, "Acoustic scene classification in dcase 2020 challenge: generalization across devices and low complexity solutions," *arXiv preprint arXiv:2005.14623*, 2020.

[4] Il-Young Jeong and Jeongsoo Park, "CochlScene: Acquisition of acoustic scene data using crowdsourcing," in *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2022, pp. 17–21.

[5] Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.

[6] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz, "mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2017.

[7] Florian Schmid, Shahed Masoudian, Khaled Koutini, and Gerhard Widmer, "Knowledge distillation from transformers for low-complexity acoustic scene classification.," in *DCASE*, 2022.

[8] Byeonggeun Kim, Seunghan Yang, Jangho Kim, Hyunsin Park, Juntae Lee, and Simyung Chang, "Domain generalization with relaxed instance frequency-wise normalization for multi-device acoustic scene classification," *arXiv preprint arXiv:2206.12513*, 2022.

[9] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang, "Domain generalization with mixstyle," *arXiv preprint arXiv:2104.02008*, 2021.

[10] Yifan Zhang, Xue Wang, Kexin Jin, Kun Yuan, Zhang Zhang, Liang Wang, Rong Jin, and Tieniu Tan, "Adanpc: Exploring non-parametric classifier for test-time adaptation," in *International Conference on Machine Learning*. PMLR, 2023, pp. 41647–41676.