ICME GRAND CHALLENGE SUBMISSION: KNOWLEDGE TRANSFER WITH MASKED AUTOENCODERS FOR ACOUSTIC SCENE CLASSIFICATION ACROSS DOMAINS

Yiqiang Cai, Shengchen Li

Xi'an Jiaotong-Liverpool University

ABSTRACT

This technical report presents a submission to the ICME Grand Challenge, focusing on the task of Semi-supervised Acoustic Scene Classification (ASC) under Domain Shift.The ASC task faces challenges due to domain shifts caused by distribution gaps between regions, considering factors such as time, space, culture, and language. Furthermore, the abundance of unlabeled acoustic scene data in the real world underscores the importance of data efficiency. To address this, we propose a training framework to leverage Masked Autoencoders (MAEs) for knowledge transfer across regions. We demonstrate the effectiveness of our approach through experiments on the DCASE dataset.

***Index Terms*—** Masked Autoencoder, self-supervised learning, semi-supervised learning, acoustic scene classification, domain shift

1. INTRODUCTION

Acoustic scene classification (ASC) [1] presents a multifaceted challenge in machine learning, characterized by the necessity to accurately categorize environmental sounds amidst varying contexts and conditions. One of the paramount challenges within ASC is the phenomenon of domain shift, wherein a notable distribution gap between training and testing data impedes the generalization performance of classification models [2]. Since 2018, the DCASE challenge has been instrumental in spotlighting the critical issue of generalizing ASC models across different recording devices [3]. While significant progress has been made in addressing device generalization [4, 5], the broader challenge of domain shift persists, particularly concerning diverse regions marked by variations in temporal, spatial, cultural, and linguistic characteristics. In the landscape of ASC research, the exploration of domain shift across regions remains relatively unexplored [6]. Moreover, the abundance of unlabeled acoustic scene data in real-world settings underscores the importance of leveraging these resources effectively.

In response to these challenges, the International Conference on Multimedia and Expo (ICME) has launched a Grand Challenge focused on knowledge transfer techniques for ASC across domains [7]. Our submission to this challenge proposes an innovative approach harnessing Masked Autoencoders (MAEs) [8] for knowledge transfer. By delving into self-supervised learning principles, MAEs facilitate the extraction of domain-invariant representations from acoustic data, thus offering a promising avenue for addressing domain shift.

This report provides a comprehensive exposition of our proposed methodology, elucidating the theoretical underpinnings of Masked Autoencoders and delineating their application in mitigating domain shift in ASC. We present empirical evidence showcasing the efficacy of our approach on benchmark dataset, demonstrating its ability to enhance the generalization performance of classification models across diverse environmental contexts.

2. DATASET AND PREPROCESSING

The experiments were conducted using the CAS 2023 dataset for development. This dataset encompasses approximately 24 hours of recordings from 8 cities. To facilitate the development of effective semi-supervised methods, the organizers randomly provided 20% of scene labels within the development dataset. Across the 10 acoustic scene classes, the number of labeled recordings for each scene is evenly distributed. Since the evaluation dataset remains unseen by participants, we partitioned the labeled data into a training/test split of 80:20 for validation purposes.

We follow [9, 10] for audio preprocessing. All audio segments were down-sampled to 32kHz at first. For feature extraction, we employed Short-Time Fourier Transform (STFT) with a window size of 3072 and a hop size of 500. Subsequently, a Mel-scaled filter bank with 256 frequency bins and 4096 FFT was applied to transform the spectrograms into Log-Mel spectrograms.

3. MODEL ARCHITECTURE

In the audio domain, Audio Masked Autoencoder (MAE) [8] has been introduced as a unified and scalable framework for self-supervised audio representation learning, achieving state-of-the-art performance on audio and speech classification tasks. Additionally, Audio-MAE incorporates masking strategies for spectrogram patches, exploring both unstructured and structured masking during pre-training and fine-tuning phases to encourage learning global, contextualized representations from limited visible patches. The effectiveness of masking in self-supervised learning is highlighted, with Audio-MAE demonstrating improved performance through masking strategies in both pre-training and fine-tuning stages. Therefore, we introduce Audio-MAE to this task.

The MAE model architecture consists of a pair of a Transformer encoder and decoder. The encoder processes only a small portion of non-masked patches to reduce computational overhead, while the decoder includes standard Transformer blocks with local attention mechanisms. The decoder restores the original time-frequency order in the audio spectrogram, incorporates fixed sinusoidal positional embeddings, and predicts/reconstructs the input spectrogram using a linear head at the top of the decoder stack. Additionally, the model employs masking strategies during pre-training and fine-tuning phases to encourage learning global, contextualized representations from limited ”visible” patches and to further regularize learning from a limited view of spectrogram inputs.

4. TRAINING PIPELINE

We follow the official pretraining strategy for MAE [8] on DCASE dataset [3] while obey the semi-supervised learning pipeline for finetune on ICME dataset [7].

4.1. Pre-training Stage

1. **Input Processing**: The audio input is transformed and embedded into spectrogram patches.

2. **Masking Strategy**: A majority of spectrogram patches are masked and discarded, while a portion (e.g., 60%) of non-masked patches are fed into the Transformer encoder for efficient encoding.

3. **Encoder**: The encoder, consisting of a stack of standard Transformers, processes the non-masked patches to reduce computation overhead.

4. **Decoder**: The decoder, also composed of standard Transformer blocks, receives the encoded patches, restores the original time-frequency order, and predicts and reconstructs the input spectrogram.

5. **Objective**: The decoder learns to reconstruct the input spectrogram by predicting the values in the spectrogram patches, with the objective being the mean squared error (MSE) between the prediction and the input spectrogram.

4.2. Fine-tuning Stage

1. **Encoder Fine-tuning**: Only the encoder is kept and fine-tuned, while the decoder is discarded.

2. **Pseudo-labeling**:The ASC model is used to assign pseudo labels to the unlabeled data within the development dataset.

3. **Further fine-tuning**: The pseudo-labeled data is utilized for additional fine-tuning on the MAE model, resulting in the final ASC model used for evaluation.

5. RESULTS

The performance of our model on different datasets reveals notable variations. When evaluated on the DCASE dataset, the test accuracy achieved a value of 67.52%, indicating a satisfactory level of performance within the bounds of typical results observed in similar studies. However, upon testing our model on the ICME dataset, a significantly higher test accuracy of 98.85% was attained. It is imperative to note that the train/test data split for the ICME dataset was determined by us, as the organizers did not provide one. This substantial disparity between the accuracies on the two datasets may indicate a potential issue of severe overfitting. For the reason of time shortage, more approaches to address this issue will be explored in future works.

6. REFERENCES

[1] Daniele Barchiesi, Dimitrios Giannoulis, Dan Stowell, and Mark D Plumbley, “Acoustic scene classification: Classifying environments from the sounds they produce,” *IEEE Signal Processing Magazine*, vol. 32, no. 3, 2015.

[2] Shayan Gharib, Konstantinos Drossos, Emre Cakir, Dmitriy Serdyuk, and Tuomas Virtanen, “Unsupervised adversarial domain adaptation for acoustic scene classification,” *arXiv preprint arXiv:1808.05777*, 2018.

[3] Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen, “A multi-device dataset for urban acoustic scene classification,” *arXiv preprint arXiv:1807.09840*, 2018.

[4] Hu Hu, Chao-Han Huck Yang, Xianjun Xia, Xue Bai, Xin Tang, Yajian Wang, Shutong Niu, Li Chai, Juanjuan Li, Hongning Zhu, et al., “A two-stage approach to device-robust acoustic scene classification,” in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 845–849.

[5] Michał Kośmider, “Spectrum correction: Acoustic scene classification with mismatched recording devices,” *arXiv preprint arXiv:2105.11856*, 2021.

[6] Yizhou Tan, Haojun Ai, Shengchen Li, and Mark D Plumbley, “Acoustic scene classification across cities and devices via feature disentanglement,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.

[7] 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.

[8] Po-Yao Huang, Hu Xu, Juncheng Li, Alexei Baevski, Michael Auli, Wojciech Galuba, Florian Metze, and Christoph Feichtenhofer, "Masked autoencoders that listen," *Advances in Neural Information Processing Systems*, vol. 35, pp. 28708–28720, 2022.

[9] Florian Schmid, Shahed Masoudian, Khaled Koutini, and Gerhard Widmer, "CP-JKU submission to DCASE22: Distilling knowledge for low-complexity convolutional neural networks from a patchout audio transformer," Tech. Rep., Detection and Classification of Acoustic Scenes and Events (DCASE) Challenge, 2022.

[10] Yiqiang Cai, Peihong Zhang, and Shengchen Li, "Tf-sepnet: An efficient 1d kernel design in cnns for low-complexity acoustic scene classification," *arXiv preprint arXiv:2309.08200*, 2023.