Enhancing Acoustic Scene Classification with Layer-wise Fine-Tuning on the SSAST Model

Shuting Hao*, Daisuke Saito*, Nobuaki Minematsu*
* Graduate School of Engineering, The University of Tokyo, Tokyo E-mail: mine@gavo.t.u-tokyo.ac.jp

***Abstract*—We introduce a novel approach to Acoustic Scene Classification (ASC) using the Self-Supervised Audio Spectrogram Transformer (SSAST) sophisticated with a focus on layer-wise fine-tuning. Recognizing the challenge in distinguishing similar classes in the TAU Urban Acoustic Scenes 2022 dataset, which sometimes exceeds human perceptual capabilities, this study introduces a novel architecture that categorizes environmental audio streams into predefined semantic labels through integrating multi-layer classifiers and direct fine-tuning. Employing the TAU Urban Acoustic Scenes 2022 Mobile dataset for both fine-tuning and validation, our SSAST model, which was initially pre-trained on the AudioSet and LibriSpeech datasets, was uniquely fine-tuned to enhance ASC-specific feature learning. Here, a combined approach of layer-wise and simultaneous fine-tuning of the backbone was introduced, eliminating the need for reassembling the dataset. This method achieved satisfactory results, our layered SSAST system reached an accuracy of 52.43% and an AUC of 88.51%, marking a notable improvement over the baseline with absolute increases of 1.25% in accuracy and 0.70% in AUC.**

I. Introduction

Acoustic Scene Classification (ASC) is to assign one of the predefined semantic labels, such as bus, office, or home, to an input audio stream recorded in various environments [1]–[3]. Generally, semantic labels stem from environmental sound categorizations that describe the ambiance of audio streams. An overview of a ASC system is shown in Fig. 1. For many years, ASC has been an active area of research. This field has made noteworthy advancements, notably in areas like the Detection and Classification of Acoustic Scenes and Events. Environmental audio signals, especially when compared to speech or music, tend to be more diverse and complex, posing unique challenges. As a result, ASC's progress has been somewhat slower than in other audio-related domains, such as speaker identification and music classification. To tackle the issue of low performance, researchers have introduced a variety of strategies. These include signal processing techniques, data augmentation, advanced feature learning, efficient deep learning modeling, and sophisticated post-processing approaches.

In this paper, we attempt to enhance accuracy without retraining models from scratch. We introduce a novel architecture that incorporates multi-layer classifiers and direct fine-tuning, diverging from the previous models such as Patchout Spectrogram Transformer (PASST) [4] which often require extensive reassembly of training data. As far as we know, the application of the Self-Supervised Audio Spectrogram Transformer (SSAST) model [5] to the TAU Urban Acoustic Scenes 2022 Mobile dataset [6] is not a mainstream approach and has not been extensively studied within the DCASE community. This study aims to explore its potential and evaluate its performance on this dataset. Our experiments demonstrate that our system achieves an accuracy of 52.43%, surpassing other methods that rely solely on direct fine-tuning. The TAU dataset is notoriously challenging, containing environmental sounds that even humans struggle to differentiate. This innovative approach not only addresses the challenges of applying pre-trained models to ASC but also advances how these models can be fine-tuned more effectively. One of the main observations in this study is the difficulty in improving accuracies for distinguishing similar categories. To tackle this issue, we have implemented advanced feature extraction techniques designed specifically to enhance differentiation among similar acoustic scenes. Our method avoids concatenating audio data, instead training on original recordings, which preserves the authenticity of the audio contexts.

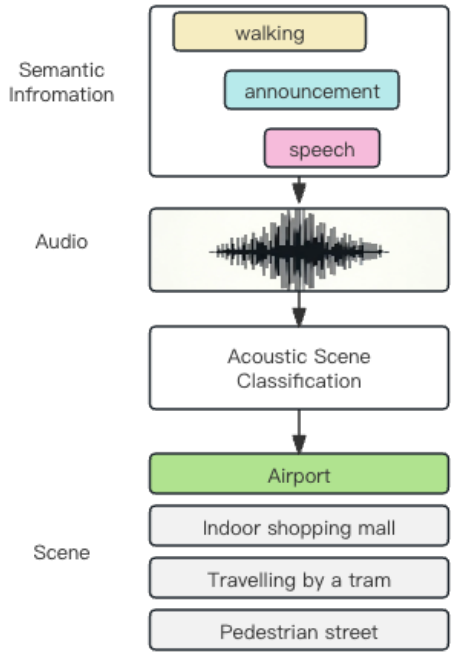


Fig. 1: Overview of an ASC system.

II. Related and Proposed Models

This chapter describes the methodology employed in the current study, focusing on the innovative use of the Self-

Supervised Audio Spectrogram Transformer (SSAST) as the baseline model for acoustic scene classification (ASC). Traditionally, ASC tasks have predominantly employed models such as the Patchout faSt Spectrogram Transformer (PaSST), which optimizes transformer training on audio spectrograms to achieve state-of-the-art performance with less computational overhead [7]. Unlike PaSST, which still requires pre-training on large labeled datasets and employs patch-based input strategies akin to those used in vision tasks, the SSAST introduces a novel approach by leveraging a self-supervised learning framework to pre-train directly on raw, unlabeled audio data [5].

Another significant strength of SSAST is its ability to perform well without the need to reassemble audio snippets into longer segments for training or evaluation. This capability underscores its robustness and effectiveness in real-world applications, where manipulation of audio data is not always feasible or desirable. The common practice in ASC competitions, such as reassembling audio snippets into longer segments for training [8], exemplified by the best models in the DCASE challenges, poses significant theoretical and practical challenges. This method may create discrepancies between training and operational environments. In real-world applications, audio data often arrive in non-continuous streams without the opportunity for pre-segmentation or reassembly. Training on artificially concatenated segments may lead to models that perform well in a controlled setting but falter in real-world scenarios where such preprocessing is not feasible. This approach may affect the model's generalization to diverse acoustic environments.

A. Baseline: Self-Supervised Audio Spectrogram Transformer (SSAST)

The Self-Supervised Audio Spectrogram Transformer (SSAST) architecture [5], which is illustrated in Fig. 2 as a self-supervised learning framework for Audio Scene Classification (ASC), is built on the Audio Spectrogram Transformer (AST) model [9]. The AST, which is based purely on self-attention mechanisms as in the Vision Transformer (ViT) [10], has demonstrated superior performance over traditional deep learning models that used convolutional neural networks (CNNs) for a variety of tasks. SSAST begins by transforming a τ-sec long audio segment into a series of 128-dimensional log Mel filterbank (fbank) features. These features are derived using a 25ms Hanning window at 10ms intervals, resulting in a $128\times100\tau$ spectrogram. This spectrogram is then divided into 16×16 patches, as shown in A of Fig. 2. And those patches are subsequently flattened into 1D 768-dimensional embeddings through a linear projection, known as the patch embedding layer, producing patch embeddings E_i. To account for the Transformer architecture's lack of understanding the temporal order, a trainable positional embedding of size 768, denoted as P_i, is added to each patch embedding, facilitating the model's understanding of the 2D audio spectrogram's spatial structure, two embeddings are shown in B of Fig. 2. The sequence of embeddings is then processed through the Transformer,

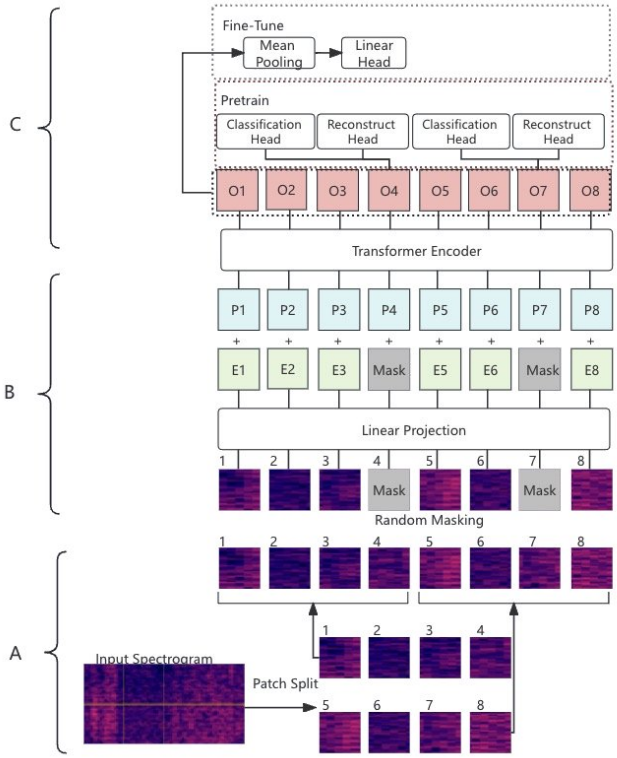


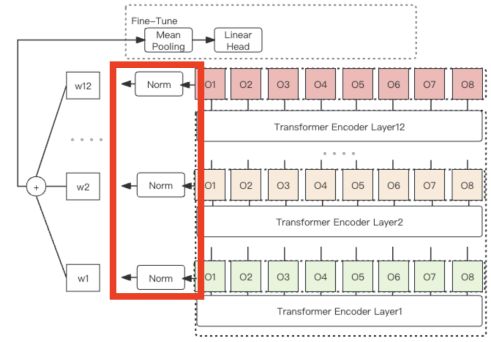
Fig. 2: Block diagram of Self-Supervised Audio Spectrogram Transformer (SSAST)

As shown in the C part of this figure, during self-supervised pretraining, a portion of spectrogram patches are randomly masked. And the task for the model is to 1) find the correct patch at each masked position from all masked patches; and 2) reconstruct the masked patch. The two pretext tasks aim to force the AST model to learn both the temporal and frequency structure of the audio data. During fine-tuning, the purpose apply a mean pooling over all patch representation O_i and use a linear head for classification [5].

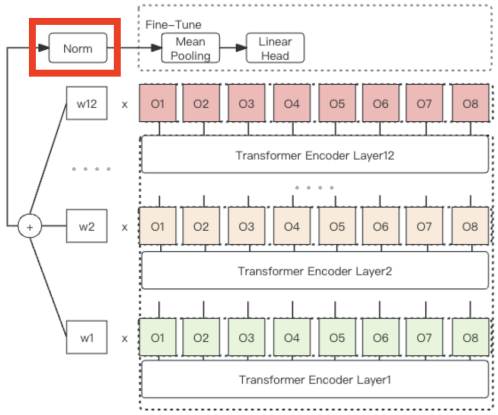
which is composed of multiple encoding layers and decoding layers. In SSAST, however, only the encoding part of the Transformer is utilized, which extracts an embedding feature whose dimension is 768, with 12 layers and 12 heads. The output from the Transformer encoder, referred to as the patch representation O_i, undergoes a mean pooling during fine-tuning and inference to yield an audio clip level representation. Subsequently, a linear head is applied for the classification task.

B. Proposed: Normalization for Layer Wise Classifiers

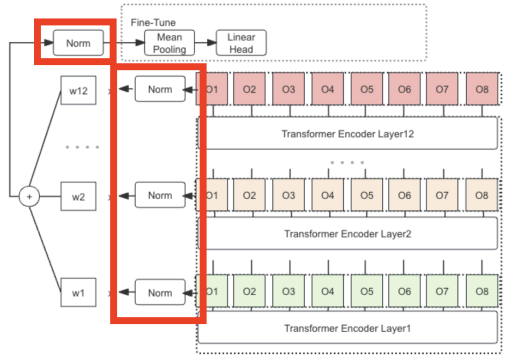
The architecture of our layer-wise classifier is shown in Fig 3 with three specific methods applied. The first method emphasizes the importance of individual layer outputs by normalizing each block's output, denoted as Pre-Norm in Fig 3a. The mean of the resultant vectors, excluding the class token, is computed to derive layer-wise representations. These are weighted by a softmax function with learned weights,



(a) Post-Norm



(b) Pre-Norm



(c) Double-Norm

Fig. 3: Block diagram of layer wise architecture.

allowing for differential emphasis on information from various network depths. This weighted sum is input to a multi-layer perceptron (MLP) head for classification. The equations for this method are:

$$M_i = \text{Mean}(\text{Norm}(O_i)), \qquad i \in \{1, \ldots, L\} \tag{1}$$

$$S = \sum_{i=1}^{L} (W_i \times M_i) \tag{2}$$

$$x = \text{MLPHead}(S), \tag{3}$$

where M_i is the output from the i^{th} layer after normalization and mean calculation, O_i is the output of the i^{th} block, S is the weighted sum of the outputs, W_i are the softmax-normalized weights, and L is the total number of layers.

The second method defers mean pooling until the weighted sum of layers is calculated, denoted as Post-Norm in Fig 3b. The layer outputs are aggregated using softmax-scaled weights and then normalized before passing through the MLP head. This is represented as:

$$M_i = \text{Mean}(O_i), \qquad i \in \{1, \ldots, L\} \tag{4}$$

$$S = \sum_{i=1}^{L} (W_i \times M_i) \tag{5}$$

$$x = \text{MLPHead}(\text{Norm}(S)), \tag{6}$$

To expand on these two methods with a new variation, we introduce a third approach that incorporates normalization both before and after the aggregation of layer outputsdenoted as Post-Norm in Fig 3b.. This method applies normalization initially to each layer output before calculating the weighted sum, and once again after the aggregation, ensuring a more uniform scale and potentially enhancing the stability and performance of the MLP head. Here's how this new method is formulated:

$$M_i = \text{Mean}(\text{Norm}(O_i)), \qquad i \in 1, \ldots, L \tag{7}$$

$$S = \sum_{i=1}^{L} (W_i \times M_i) \tag{8}$$

$$x = \text{MLPHead}(\text{Norm}(S)), \tag{9}$$

In all methods, MLPHead denotes the MLP head, which processes the normalized weighted sum and produces the classification output x. The normalization step ensures that the signal is suitable for classification, making the most of the diverse features extracted from the audio signal.

III. Experimental Setup

In our experimental setup, all evaluations were conducted on the TAU Urban Acoustic Scenes 2022 Mobile dataset [6]. This dataset comprises 64 hours of audio recordings from 12 different European cities, featuring 10 unique acoustic scenes. The audio was captured using four distinct devices, and each audio sample has a length of 1 second, trimmed from the original 10-second recordings. Out of the total audio data, 40 hours were recorded using the main device, while the remaining hours were captured with the three additional, less frequently used devices. The audio format for all devices is single-channel, 44.1kHz, and 24-bit resolution.

In our experimental framework, we leveraged the SSAST model, which features a novel approach integrating both discriminative and generative learning paradigms through Masked Spectrogram Patch Modeling (MSPM). This model was pre-trained using a vast collection of unlabeled audio data sourced from the AudioSet [11] and LibriSpeech datasets [12]. With MSPM, the model masks a strategic subset of spectrogram patches, specifically 400 patches, to challenge the model to predict the missing acoustic information. This technique not only encourages the model to learn a more comprehensive

representation of the audio but also to discern finer details that are crucial for classification. Notably, the effectiveness of this model has been validated by its impressive performance on the ESC-50 dataset [13], a benchmark in environmental sound classification. This is why the SSAST model is regarded as suited foundation on the TAU Urban Acoustic Scenes 2022 Mobile dataset.

For our experiments, we employed a consistent set of hyperparameters across all models to maintain uniformity in the evaluation conditions. These parameters included the learning rate, batch size, and data augmentation techniques, along with the warm-up strategy, number of training epochs, choice of optimizer, and loss function utilized during the training process.

Various data augmentation strategies were explored during experimentation. Mixup augmentation, implemented at the waveform level using a Beta distribution for mixing weights, was attempted but did not lead to significant improvement. Feature-level augmentation was also investigated, specifically by adding random noise to the log mel-filterbank features and applying time-axis shifting, which resulted in a modest improvement. The most effective augmentation technique proved to be the masking strategy inherent to SSAST's self-supervised learning framework, which applies both time and frequency masking to the input spectrogram.

In addition to feature augmentation, a notable deviation from the pre-training phase was the input handling; specifically, we introduced overlapping inputs to enrich the model's exposure to the data. With these consistent parameters in place, we focused on the classifier component of our modified architecture. Starting with randomly initialized weights for the classifier while inheriting the backbone's weights from pre-training, we proceeded with fine-tuning the entire model without freezing any layers. This approach was adopted due to the significant differences between the TAU Urban Acoustic Scenes 2022 Mobile dataset and the AudioSet upon which the model was originally trained. To ascertain the impact of our strategy, we also conducted comparative experiments using models with their backbone layers frozen during fine-tuning.

IV. Results and Discussion

Before delving into the specifics of the experimental outcomes, it's essential to understand the metrics used to measure the performance of the systems. Accuracy (Acc) is a metric that represents the proportion of true results, both true positives and true negatives, in the total number of cases examined. It provides a straightforward measure of a model's overall correctness. On the other hand, the Area Under Curve (AUC) is a performance measurement for classification problems at various threshold settings. AUC represents the degree to which the model is capable of distinguishing between classes. An AUC of 1 indicates perfect prediction, while an AUC of 0.5 suggests no discriminative power, equivalent to random guessing.

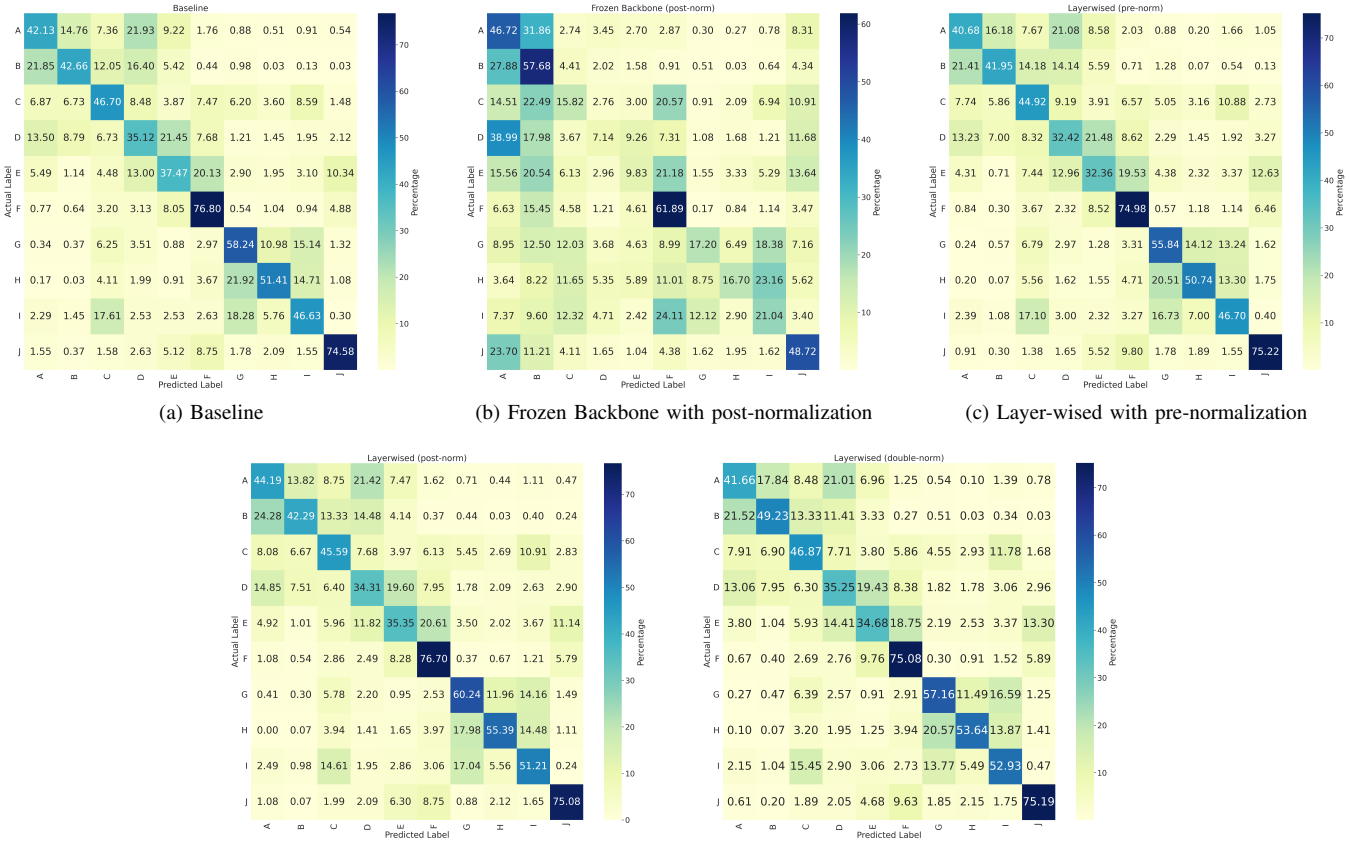
In this study, we only adopt Accuracy and AUC as performance metrics because our TAU Urban Acoustic Scenes 2022 Mobile dataset is a fully balanced dataset, where each class contains an approximately equal number of samples. This makes these two metrics highly appropriate, as they provide an accurate reflection of the model's performance without being skewed by class imbalances.

TABLE I: Performance of our proposed layer-wised SSAST and several comparing systems.

Model	Acc(%)	AUC(%)
Baseline	51.18	87.81
Frozen Backbone (post-norm)	30.27	73.11
Layer-wised (pre-norm)	49.58	87.06
Layer-wised (post-norm)	52.03	88.17
Layer-wised (double-norm)	**52.43**	**88.51**

The experimental results of layer-wised SSAST and several comparing systems are summarized in Table I. The layer-wised model with double-normalization SSAST outperforms the baseline SSAST by absolute 1.25% Acc and absolute 0.70% AUC. Our results support the hypothesis that by extracting information from all layers and implementing comprehensive fine-tuning, we can adapt pre-trained models to extremely challenging and divergent domain tasks effectively even with minimal efforts. Models with frozen backbones show a significant divergence between the SSAST pretraining domain and the DCASE ASC target domain, which training the classifier alone cannot address. Furthermore, the results of the layer-wised model with pre-normalization indicate that the importance assigned to each layer's output varies greatly. Normalizing these outputs before weighting may actually diminish the distinct information of different layer levels, resulting in lower classification performance. Post-normalization, applied after weighting the outputs, preserves the distinctiveness of features from each layer. This method enhances the model's ability to leverage inter-layer discriminative features, potentially improving the robustness and accuracy of the classification across varied and complex scenarios.

We also plot the confusion matrices for 1) Baseline, 2) Frozen Backbone with post-norm, and 3) three versions of layer-wise architecture, depicted in Figure 4. Our findings provide further validation of our initial premise. The confusion matrix from models with a Frozen Backbone reveals that the backbone itself is entirely unsuitable and incapable of facilitating effective classification, as shown in Fig. 4a and Fig. 4b. Moreover, when comparing the baseline with the layer-wised with pre-norm approach, there are no significant differences in classification outcomes for most categories, as shown in Fig. 4c. And, for the post-norm approach, as shown in Fig. 4d, there is a noticeable improvement at class G. However, for the double-norm approach, as shown in Fig. 4e, there is a noticeable improvement across those categories that with low accuracy, such as class B and C. This suggests that the overall architecture enables the model to extract features more effectively and utilize multi-layer features more efficiently for classification. Such an approach could enhance the fine-tuning performance of both the backbone and the classifier components of the model.



(a) Baseline (b) Frozen Backbone with post-normalization (c) Layer-wised with pre-normalization

(d) Layer-wised with post-normalization (e) Layer-wised with double-normalization

Fig. 4: Confusion matrices of the ASC systems examined. The labels from A to J correspond to: A - Airport, B - Indoor shopping mall, C - Metro station, D - Pedestrian street, E - Public square, F - Street with medium level of traffic, G - Travelling by a tram, H - Travelling by a bus, I - Travelling by an underground metro, J - Urban park.

V. Conclusions

In this paper, we proposed a novel ASC system named Layer wised SSAST. Based on our experiments and analysis, it is evident that the layer-wise approach significantly aids pre-trained models to be adapted for tasks where the data distribution undergoes severe changes and presents considerable challenges. This is particularly beneficial for extremely difficult tasks and datasets within the ASC, where the approach shows marked improvements in classifying similar categories. Our system enhances discriminative capability in similar acoustic scenes, improving performance in challenging ASC applications. This method achieved an accuracy of 52.43% and an AUC of 88.51%, representing an absolute increase of 1.25% in accuracy and an absolute increase of 0.70% in AUC.

This study enhances the adaptability of SSAST for ASC tasks, particularly in handling similar class distinctions, providing a new strategy for future ASC systems to improve classification performance in complex scenarios. The method also demonstrates SSAST's potential to better adapt to real-world ASC tasks, paving the way for future research to further expand its applications. Additionally, the research offers a promising solution for deploying complex models on resource-constrained devices, suggesting future work on low-complexity models and knowledge distillation techniques for efficient ASC implementations.

For the future works several promising directions for enhancing ASC systems using the Self-Supervised Audio Spectrogram Transformer (SSAST). Emphasizing low complexity models is crucial for devices with limited computational resources. Future studies should explore using SSAST for knowledge distillation to train efficient student models and leverage techniques like model compression, quantization, and pruning to reduce complexity [14]–[17]. Additionally, evaluating SSAST on datasets beyond urban environments, including rural, wilderness, and underwater soundscapes, could provide insights into its adaptability and generalization. This research is essential to understand the model's performance across different auditory environments [18]. Furthermore, enhancing current classification schemes by addressing overlaps and ambiguities in acoustic scene definitions could improve model precision. A more granular taxonomy of acoustic environments would ensure consistency and reliability across applications [19]–[21].

References

[1] S. Waldekar and G. Saha, "Two-level fusion-based acoustic scene classification," *Applied Acoustics*, vol. 170, p. 107 502, 2020.

[2] T. Virtanen, M. Plumbley, and D. Ellis, Eds., *Computational analysis of sound scenes and events*, English. Springer, 2018, ISBN: 978-3-319-63449-4. DOI: 10.1007/978-3-319-63450-0.

[3] B. Ding, T. Zhang, C. Wang, *et al.*, "Acoustic scene classification: A comprehensive survey," *Expert Systems with Applications*, vol. 238, p. 121 902, Oct. 2023. DOI: 10.1016/j.eswa.2023.121902.

[4] Q. Kong, L. Cances, Y. Wang, T. Iqbal, M. D. Plumbley, and W. Wang, "Passt: Efficient training of audio transformers with patchout," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2022.

[5] Y. Gong, C.-I. Lai, Y.-A. Chung, and J. Glass, "Ssast: Self-supervised audio spectrogram transformer," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, 2022, pp. 10 699–10 709.

[6] T. Heittola, A. Mesaros, and T. Virtanen, "Acoustic scene classification in dcase 2020 challenge: Generalization across devices and low complexity solutions," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2020 Workshop*, 2020, pp. 56–60. [Online]. Available: https://arxiv.org/abs/2005.14623.

[7] K. Koutini, J. Schluter, H. Eghbal-zadeh, and G. Widmer, "Efficient training of audio transformers with patchout," *arXiv preprint arXiv:2110.05069*, 2021, Available at https://arxiv.org/abs/2110.05069.

[8] F. Schmid, S. Masoudian, K. Koutini, and G. Widmer, "CP-JKU submission to dcase22: Distilling knowledge for low-complexity convolutional neural networks from a patchout audio transformer," DCASE2022 Challenge, Tech. Rep., Jun. 2022.

[9] Y. Gong, Y.-A. Chung, and J. Glass, "AST: Audio Spectrogram Transformer," in *Proc. Interspeech 2021*, 2021, pp. 571–575. DOI: 10.21437/Interspeech.2021-698.

[10] A. Dosovitskiy, L. Beyer, A. Kolesnikov, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[11] J. F. Gemmeke, D. P. Ellis, D. Freedman, *et al.*, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE international conference on acoustics, speech and signal processing*, IEEE, 2017, pp. 776–780.

[12] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing*, IEEE, 2015, pp. 5206–5210.

[13] K. J. Piczak, "Esc: Dataset for environmental sound classification," in *Proceedings of the 23rd ACM international conference on Multimedia*, 2015, pp. 1015–1018.

[14] I. Martín-Morató, T. Heittola, A. Mesaros, and T. Virtanen, "Low-complexity acoustic scene classification for multi-device audio: Analysis of dcase 2021 challenge systems," *Applied Sciences*, vol. 11, no. 4, p. 1850, 2021.

[15] C. Buciluǎ, R. Caruana, and A. Niculescu-Mizil, "Model compression," in *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, 2006, pp. 535–541.

[16] M. Courbariaux, Y. Bengio, and J.-P. David, "Binaryconnect: Training deep neural networks with binary weights during propagations," *Advances in Neural Information Processing Systems*, 2015.

[17] J. Gou, B. Yu, S. Maybank, and D. Tao, "Knowledge distillation: A survey," *International Journal of Computer Vision*, vol. 129, no. 6, pp. 1789–1819, 2021.

[18] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in *Proceedings of the 22nd ACM International Conference on Multimedia*, ACM, 2014, pp. 1041–1044.

[19] X. Bai, J. Du, J. Pan, H. S. Zhou, Y. H. Tu, and C.-H. Lee, "High-resolution attention network with acoustic segment model for acoustic scene classification," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2020, pp. 656–660.

[20] S. Chandrakala and S. Jayalakshmi, "Environmental audio scene and sound event recognition for autonomous surveillance: A survey and comparative studies," *ACM Computing Surveys (CSUR)*, vol. 52, no. 3, pp. 1–34, 2019.

[21] T. Virtanen, M. D. Plumbley, and D. Ellis, "Computational analysis of sound scenes and events," in *2018 IEEE 20th International Workshop on Multimedia Signal Processing (MMSP)*, IEEE, 2018.