DATA-EFFICIENT LOW-COMPLEXITY ACOUSTIC SCENE CLASSIFICATION WITH CURRICULUM LEARNING AND SE-LAYER

Technical Report

Xuanyan Chen

School of Computer and Electronic Information, Guangxi University, Guangxi, China
2313394006@st.gxu.edu.cn

Wei Xie

School of Computer and Electronic Information, Guangxi University, Guangxi, China
chester.w.xie@gmail.com

ABSTRACT

This technical report presents a data-efficient and low-complexity acoustic scene classification (ASC) system developed for Task 1 of the DCASE2024 Challenge. The primary objective is to create ASC models that perform effectively with limited labeled data and minimal computational resources, addressing practical constraints in real-world applications. Our proposed system integrates Squeeze-and-Excitation (SE) layers within the baseline network and employs a curriculum learning approach for training. SE layers enhance feature representation by recalibrating channel-wise feature responses, while curriculum learning structures the training process by progressively introducing more complex examples, facilitating better model generalization and robustness. Experimental results demonstrate significant improvements in classification accuracy across various data splits, with our system outperforming the baseline by up to 7% on the development dataset. The approach promises to advance the accessibility and scalability of ASC technologies in resource-constrained environments.

Index Terms— Acoustic scene classification, SE-Layer, curriculum learning, Freq-MixStyle

1. INTRODUCTION

The DCASE2024 Challenge's Task 1 focuses on data-efficient and low-complexity acoustic scene classification (ASC). This task is part of the broader effort to advance the field of environmental sound recognition, aiming to develop systems capable of accurately classifying acoustic scenes with limited data and computational resources. Acoustic scene classification involves identifying the environment in which an audio recording was made, such as a park, a busy street, or a home.

The primary objective of Task 1 is to encourage the development of ASC systems that are both data-efficient and computationally lightweight. Traditional ASC systems often require large amounts of labeled data and significant computational power, which can be impractical for real-world applications, particularly in resource-constrained environments. This task challenges participants to innovate solutions that overcome these limitations, fostering advancements that can make ASC technology more accessible and scalable.

Acoustic scene classification has numerous practical applications, ranging from enhancing the context-awareness of smart devices to improving surveillance systems and facilitating environmental monitoring. By focusing on data efficiency and low complexity, this task addresses critical barriers to the widespread deployment of ASC systems, such as the need for extensive labeled datasets and the high computational cost of model inference. Achieving these goals can lead to more versatile and widely applicable ASC solutions.

Participants in Task 1 face several challenges, including:

- **Data Efficiency**: Designing models that can learn effectively from limited labeled data.
- **Low Complexity**: Ensuring that models are computationally efficient, enabling their deployment on devices with limited processing power.
- **Robustness and Accuracy**: Maintaining high classification accuracy despite the constraints on data and computational resources.

These challenges require innovative approaches to model design, including the use of transfer learning, data augmentation techniques, and novel architectures tailored for efficiency.

In this technical report, we embedding the Squeeze-and-Excitation layer(SE-Layer) into the baseline's network and training it using a curriculum learning approach.

2. THE SQUEEZE-AND-EXCITATION LAYER

The goal of squeeze-and-excitation layer is to improve the quality of representations produced by a network by explicitly modelling the interdependencies between the channels of its convolutional features. It proposed a mechanism that enables the network to carry out feature recalibration, thereby learning to leverage global information to selectively highlight informative features and dampen those that are less useful.

The structure of the SE-layer is depicted in Fig. 1. For any given transformation F_{tr} mapping the input X to the feature maps U where $U \in R^{H\times W\times C}$, e.g. a convolution, we can construct a corresponding SE-layer to perform feature recalibration. The features U are first passed through a squeeze operation, which produces a channel descriptor by aggregating feature maps across their spatial dimensions ($H \times W$). The function of this descriptor is to produce an embedding of the global distribution of channel-wise feature responses, allowing information from the global

receptive field of the network to be used by all its layers. The aggregation is followed by an excitation operation, which takes the form of a simple self-gating mechanism that takes the embedding as input and produces a collection of per-channel modulation weights. These weights are applied to the feature maps U to generate the output of the SE-layer which can be fed directly into subsequent layers of the network.

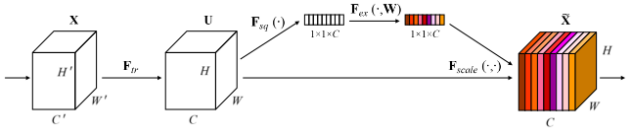

Fig. 1.A Squeeze-and-Excitation layer.

3. THE NETWORK

The proposed network is modified based on the baseline. The network of baseline is the CP-Mobile (CPM), which consists of CPM blocks. Fig. 2 depicts the structure of a CPM block consisting of two pointwise and a depthwise convolution. The depthwise convolution operates on the expanded channel representation, which has the size of the number of block input channels times the scaling factor EXP. We differentiate between Transition, Standard and Spatial Downsampling blocks (CPM blocks T, S, D).CPM block T increases the channel dimension, uses no residual connection and can be used with a strided depthwise convolution. CPM blocks S and D have matching input and output channel dimensions and use a residual connection. CPM block D uses average pooling with a kernel size of 3 and a stride of 2 on the shortcut path to match the spatial di mensions of the block output.

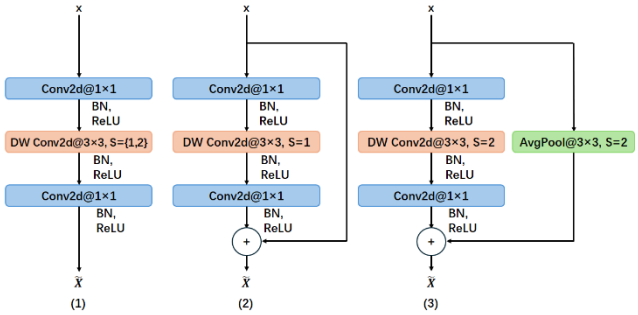

Fig. 2.CPM blocks: (1) Transition Block (input channels≠ output channels), (2) Standard Block, (3) Spatial Downsampling Block (S denotes stride)

Fig. 3 depicts the structure of the proposed network. We embed the SE-Layer before the ReLU activation function at the end of each CPM block, and in the original CPM block, here is a GRN block for computing the normalized values for each channel.

4. CURRICULUM LEARNING

Curriculum Learning is an innovative approach in the realm of machine learning and artificial intelligence, inspired by the human educational process. Just as students learn more effectively when presented with material in a structured and progressive manner, Curriculum Learning advocates for a similar strategy in training models. By organizing training data from simpler to more complex examples, models can develop a robust understanding of basic concepts before tackling more intricate patterns. This method not only accelerates the learning process but also enhances the overall performance and generalization capabilities of AI systems.

The curriculum learning algorithm we use is divided into two main parts. The first part is a scoring function that determines the "difficulty" or "complexity" of each example in the data. The scoring function makes it possible to sort the training examples by difficulty and present the easier (and possibly simpler) examples to the network first.

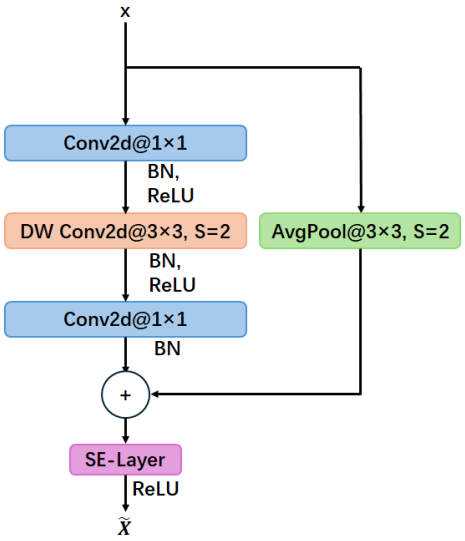

Fig. 3.SE block

Regarding scoring functions, most curriculum learning algorithms can be categorized into two types, transfer learning or bootstrapping. What we use is the latter, bootstrapping. Bootstrapping means we sort the dataset with the network we will train, without transfer learning or pre-training.

The second part is the pacing function. The pacing function in Curriculum Learning is a critical component that dictates the rate at which a machine learning model transitions from simpler to more complex tasks. Much like a teacher who adjusts the difficulty of lessons based on a student's progress, the pacing function dynamically manages the progression of training examples to optimize learning. This function is essential for balancing the trade-off between consolidating foundational knowledge and advancing to more challenging concepts. By carefully controlling the pace, models can achieve better performance and generalization, avoiding the pitfalls of being overwhelmed by complexity too soon or stagnating on overly simplistic tasks.

5. EXPERIMENTS

3.1. Experimental Setup

The development dataset consists of training subset and validation subset. The development dataset contains audio recordings from 10 cities and 9 devices: 3 real devices (A, B, C) and 6 simulated devices (S1-S6). Audio recordings recorded by devices B, C,

and S1-S6 are composed of audio segments that are randomly selected from simultaneous recordings. Hence, all of these audio recordings overlap with the audio recordings from device A, but not necessarily with each other. The total amount of audio recordings in the development dataset is 64 hours. Some devices appear only in the validation subset.

For training the model, audio input is resampled to 32 kHz and converted to mel spectrograms using a 4096-point FFT with a window size of 96 ms and a hop size of approximately 16 ms, followed by a mel transformation with a filterbank of 256 mel bins. The system is trained for 200 epochs using the SGD optimizer and a batch size of 256. Freq-MixStyle is applied to tackle the device mismatch problem, and time rolling of the waveform and frequency masking are used to augment the training data. The baseline system requires 29.4 MMACs for the inference on a one-second audio clip. The memory required for the model parameters amounts to 127.8 kB, resulting from the 63,900 parameters used in 16-bit precision (float 16).

3.2. Experimental Results

The validation set for the development dataset contains 29680 audio examples. We calculate the overall accuracy and the Confusion Matrix. Table. 1. Depicts our experimental results. Fig. 4.depicts the Confusion Matrix.

Split	Parameters	MACs	BL-Acc	Acc	Log loss
100	63900	29.42	56.99±1.11	57.27	1.215
50	63900	29.42	53.19±0.68	56.84	1.731
25	63900	29.42	50.29±0.87	53.82	1.453
10	63900	29.42	45.29±1.01	48.74	1.624
5	63900	29.42	42.40±0.42	43.25	1.763
BL	61148	29.42			

Table. 1. Results of experiments on different splits

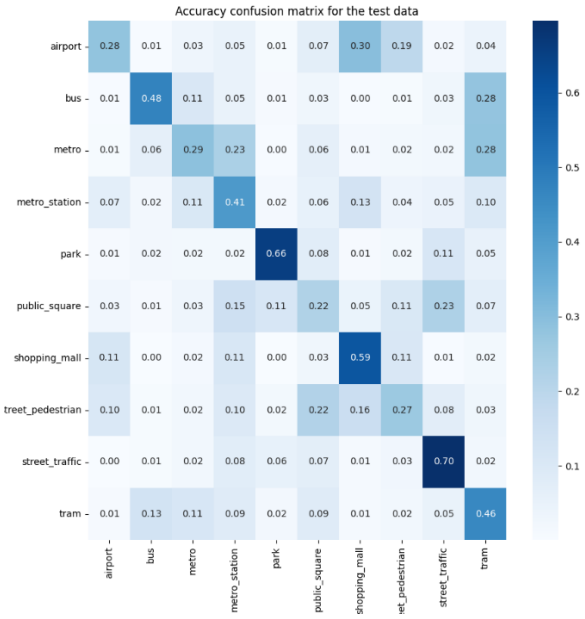

Fig. 4.(a) Accuracy confusion matrix for the 5% split model

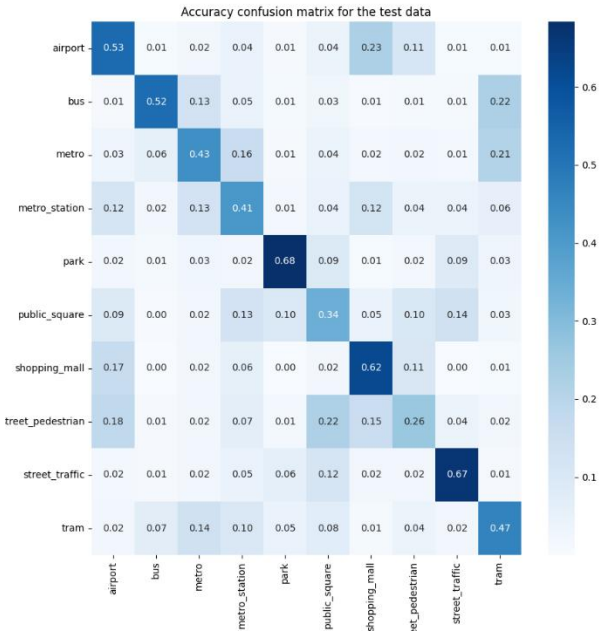

Fig. 5.(b) Accuracy confusion matrix for the 10% split model

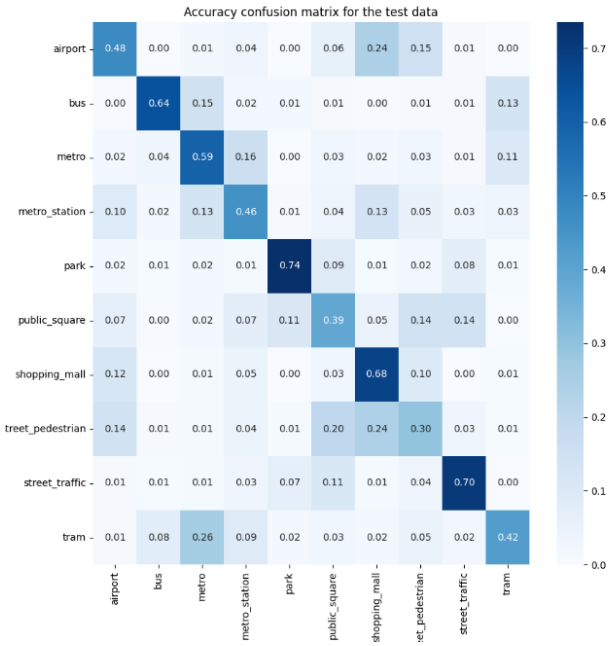

Fig. 6.(c) Accuracy confusion matrix for the 25% split model

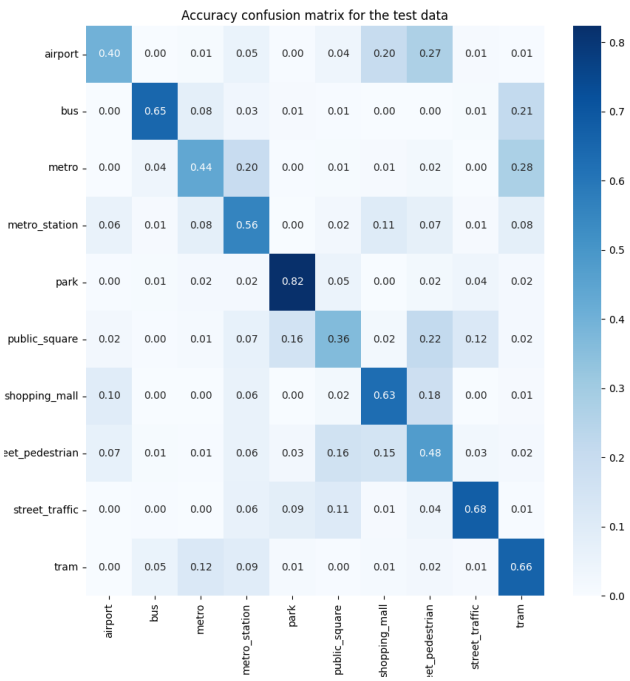

Fig. 7.(d) Accuracy confusion matrix for the 50% split model

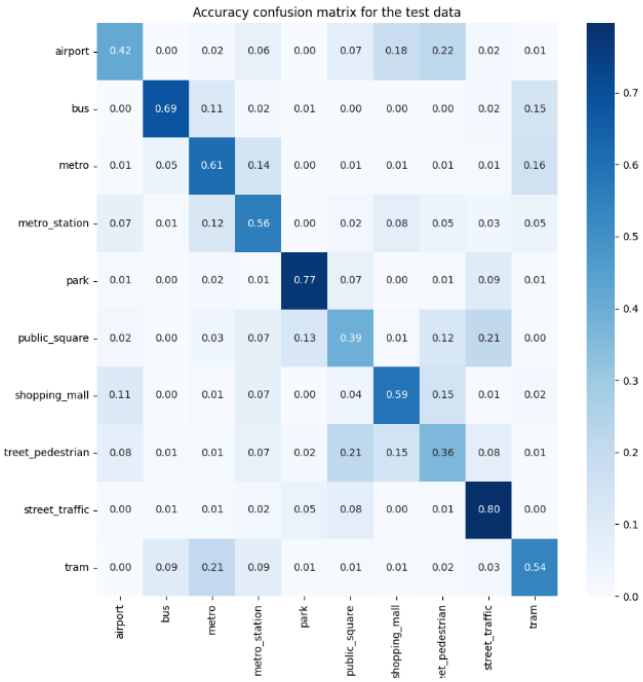

Fig. 8.(e) Accuracy confusion matrix for the 100% split model

6. CONCLUSION

In this technical report, we proposed a system for the data-efficient and low-complexity acoustic scene classification(ASC). It's for Task 1 of DCASE challenge 2024. This system was developed based on the Challenge baseline system, with the SE-Layer embedded in the system and trained using the Curriculum Learning algorithm. The accuracy of the submitted system was up to 7% higher than the baseline of the development dataset.

7. REFERENCES

[1] Toni Heittola, Annamaria Mesaros, and Tuomas 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 (DCASE2020), 56–60. 2020. URL: https://arxiv.org/abs/2005.14623.

[2] Florian Schmid, Paul Primus, Toni Heittola, Annamaria Mesaros, Irene Martín-Morató, Khaled Koutini, and Gerhard Widmer. Data-efficient low-complexity acoustic scene classification in the dcase 2024 challenge. 2024. URL: https://arxiv.org/abs/1706.10006.

[3] Florian Schmid, Tobias Morocutti, Shahed Masoudian, Khaled Koutini, and Gerhard Widmer. Distilling the knowledge of transformers and CNNs with CP-mobile. In Proceedings of the Detection and Classification of Acoustic Scenes and Events 2023 Workshop (DCASE2023), 161–165. 2023.

[4] Hu, Jie, Li Shen, and Gang Sun. "Squeeze-and-excitation networks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.

[5] Bai, Jisheng, et al. "A squeeze-and-excitation and transformer based cross-task model for environmental sound recognition." IEEE Transactions on Cognitive and Developmental Systems (2022).

[6] Hacohen, Guy, and Daphna Weinshall. "On the power of curriculum learning in training deep networks." International conference on machine learning. PMLR, 2019.

[7] Kong, Yajing, et al. "Adaptive curriculum learning." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.

[8] Zhou, Yuwei, et al. "Curml: A curriculum machine learning library." Proceedings of the 30th ACM International Conference on Multimedia. 2022.

[9] Maharana, Adyasha, and Mohit Bansal. "On curriculum learning for commonsense reasoning." Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2022.