DUAL-STRATEGY ENHANCEMENT OF ACOUSTIC SCENE AND EVENT CLASSIFICATION: INTEGRATING RES2NET, GHOSTNET, AND MOBILEFORMER ARCHITECTURES

Technical Report

TaeSoo Kim, Daniel Rho, Gahui Lee, JaeHan Park

KT Corporation, South Korea

ABSTRACT

In this technical report, we investigate the balance between accuracy and efficiency in the low-complexity acoustic scene classification (ASC) task for the DCASE 2023 challenge. We explore two approaches: the first prioritizes accuracy using Res2Net and GhostNet, while the second emphasizes efficiency using MobileFormer. Our study highlights the trade-offs between accuracy and efficiency in ASC models and contributes to the ongoing research on developing robust and lightweight models suitable for embedded systems.

Index Terms— acoustic scene classification, Res2Net, GhostNet, MobileFormer

1. INTRODUCTION

Acoustic scene classification (ASC) is the task of identifying and categorizing the auditory environment based on the sounds present in a given audio recording. With a wide range of applications, such as multimedia search, context-aware mobile devices, robots, and intelligent monitoring systems, ASC has attracted significant research interest in recent years. However, recognizing sound scenes and individual sound sources in realistic soundscapes remains a challenging problem due to the presence of multiple overlapping sounds and environmental distortions.

The DCASE 2023 challenge addresses this problem by focusing on low-complexity ASC, targeting devices with limited computational and memory resources. The challenge emphasizes the need for robust models that can run efficiently on embedded systems while maintaining high accuracy in scene classification. To achieve this, the challenge considers model accuracy, memory, MMACs (multiply-accumulate operations), and energy consumption as evaluation criteria.

In this technical report, we investigate the balance between accuracy and efficiency in the low-complexity ASC task for the DCASE 2023 challenge. We explore two approaches: the first prioritizes accuracy using Res2Net [1] and GhostNet [2] architectures, while the second emphasizes efficiency using MobileFormer [3]. Our study highlights the trade-offs between accuracy and efficiency in ASC models and contributes to the ongoing research on developing robust and lightweight models suitable for embedded systems.

The remainder of the report is organized as follows: Section 2 outlines the methodology, detailing the accuracy-focused and efficiency-focused approaches. Section 3 describes the experimental setup, including dataset, preprocessing, and evaluation metrics. Section 4 presents the results of our experiments. Section 5 discusses the findings and their implications.

2. METHOD

Our proposed method consists of two primary components: enhancing accuracy through the integration of the Res2Net and GhostNet algorithms, and improving efficiency by employing the MobileFormer architecture.

2.1. Accuracy-Focused Approach: Res2Net and GhostNet

In our first approach, we prioritize model accuracy by incorporating Res2Net [1] and GhostNet [2] architectures. ResNet-based CNN models have demonstrated high performance in previous competitions [4, 5, 6]. Kim et al. [5] employed a BC-ResNet [7] architecture incorporating max-pool layers and limited receptive field, while introducing ResNorm for per-frequency band normalization in the residual path. In a subsequent challenge, Lee et al. [4] enhanced the BC-ResNet [5] model by incorporating the Res2Net [1] style, enabling the extraction of frequency and temporal features through Broadcast learning while operating in a computationally efficient multi-scale manner.

Starting from the BC-Res2Net proposed by Lee et al.[4], we incorporated the GhostNet[2] module into the model. By integrating the advantages of Res2Net [1] into the field of acoustic scene classification, it enhances the learning capacity of our model, enabling the effective capture of complex acoustic features across a wide range of scales.

In our implementation, we made modifications to the BC-Res2Net architecture by excluding the MFA (Multi-level Feature Aggregation) [8] and FPM (Feature Pyramid Module) [9] components. These components were omitted due to the substantial increase in computational resources required. However, we were able to compensate for their exclusion by leveraging the efficiency and feature extraction capabilities of the GhostNet module. This decision was made to strike a balance between performance and computational efficiency, ensuring the practical feasibility of our model for real-world acoustic scene classification applications.

2.2. Efficiency-Focused Approach: MobileFormer

In the second approach, we focus on optimizing the model's efficiency in terms of the number of parameters and MACs, without compromising the accuracy substantially. To achieve this, we use MobileFormer [3], a parallel design of MobileNet [10] and Transformer [11] that combines the efficiency of MobileNet in local processing with the advantage of Transformer in encoding global interactions. The two-way bridge in MobileFormer enables bidirectional

Table 1: Performance of our models

				Validation
Architecture	bit	FLOPS (MMAC)	# params	acc
Baseline	8	29.234	46.512K	42.90
Res2Net + GhostNet	32	14.579	83.570K	56.97
Res2Net + GhostNet	8	14.579	83.570K	56.10
MobileFormer	8	0.617	20.516K	50.85

Table 2: Specification of Ghost-BC-Res2Net. F, T and C denote the size of frequency, time, and channel dimensions, respectively.

Stage	Input (F×T×C)	Operation	C	Strides
stem	F × T × 1	Conv + ReLU + BN	2C	2
1	F/2 × T/2 × 2C	Ghost-BC-Res2Net Block	C	-
	F/2 × T/2 × C	Ghost-BC-Res2Net Block	C	-
2	F/2 × T/2 × C	Max-Pool2d	-	2
	F/4 × T/4 × C	Ghost-BC-Res2Net Block	1.5C	-
	F/4 × T/4 × 1.5C	Ghost-BC-Res2Net Block	1.5C	-
3	F/4 × T/4 × 1.5C	Max-Pool2d	-	2
	F/8 × T/8 × 1.5C	Ghost-BC-Res2Net Block	2C	-
	F/8 × T/8 × 2C	Ghost-BC-Res2Net Block	2C	-
4	F/8 × T/8 × 2C	Ghost-BC-Res2Net Block	2.5C	-
	F/8 × T/8 × 2.5C	Ghost-BC-Res2Net Block	2.5C	-
	F/8 × T/8 × 2.5C	Ghost-BC-Res2Net Block	2.5C	-
head	F/8 × T/8 × 2.5C	Global Avg. Pool	-	-
	1 × 1 × 2.5C	Linear	num classes	

fusion of local and global features, allowing MobileNet and Transformer to communicate through the bridge.

By incorporating MobileFormer in our acoustic scene classification model, we aim to reduce parameter count and MMACs while maintaining a competitive level of accuracy. This approach addresses the efficiency-focused aspect of our investigation, as we strive to find the optimal balance between accuracy and efficiency in the low-complexity ASC task for the DCASE 2023 challenge.

3. RESULTS AND DISCUSSION

We tried two different approach with different aims; accuracy and efficiency. Tab. 1 shows the overall statistics of our trials. For quantization, we used 8-bit quantization-aware training-based quantization (QAT).

3.1. Implementation Details

We used only TAU Urban Acoustic Scenes 2022 Mobile Developement dataset [12]. For augmentation, we used Frequency and Time masking [13] and time rolling. We downsampled every audio sample with the sampling rate of 16,000. The window length and n_fft were set to 2,048 and the hop length was set to 512. We used log mel spectrogram with the number of mel filterbanks to 256. For optimization, we used AdamW [14] optimizer with cosine annealing with warmup scheduler. The number of training epochs were set to 300 unless specified.

3.2. Accuracy-Focused Approach: Res2Net and GhostNet

Our accuracy-focused approach, combining Res2Net and GhostNet, achieves the best accuracy of 56.97%. However, this comes at the cost of increased model complexity, with 83.57K parameters and 14.579 MMACs. This approach demonstrates the potential for high accuracy in ASC but faces challenges in adhering to the low-complexity constraints set by the DCASE 2023 challenge.

3.3. Efficiency-Focused Approach: MobileFormer

Since the smallest official network configuration of MobileFormer exceeds the computational constraints of the DCASE 2023 challenge, we searched over network configurations of MobileFormer. To make network architecture compact, we used only three MobileFormer downsample block. We used four tokens with a size of 16. To further reduce the computational costs, we split fully-connected linear layer into a group-wise connected linear layer. Tab. 3 shows the specification of the modified network. With the modified configuration, MobileFormer-based approach achieved the best accuracy of 54.19% in the validation set. This approach is considerably more efficient than the first in terms of both the number of parameters and the computational costs, using only 20.516K parameters and 0.617 MMACs. While the accuracy is slightly lower than the first approach, it better aligns with the low-complexity constraints set by the DCASE 2023 challenge, providing a competitive balance between accuracy and resource efficiency.

Table 3: Specification of a modified MobileFormer. F and T denote the size of frequency and time dimensions, respectively.

Stage	Input	Operation	Output
tokens	4 × 16	-	
stem	F × T × 1	Conv2d	F/2 × T/2 × 8
1	F/2 × T/2 × 8	Mobile-Former$^{\downarrow}$	T/4 × F/4 × 12
2	T/4 × F/4 × 12	Mobile-Former$^{\downarrow}$	T/8 × F/8 × 18
3	T/8 × F/8 × 18	Mobile-Former$^{\downarrow}$	T/16 × F/16 × 24
head	T/16 × F/16 × 24	Merge-classifier	T/16 × F/16 × 144
	T/16 × F/16 × 144	Avg Pool	144
	144	FC	64
	64	FC	10

3.4. Balancing Accuracy and Efficiency

Our comparative study reveals the trade-offs between accuracy and efficiency in the design of ASC models. The first approach using Res2Net and GhostNet achieves higher accuracy, but with increased model complexity. In contrast, the MobileFormer approach sacrifices a small amount of accuracy to achieve a much more efficient model in terms of parameters and MMACs. This study highlights the importance of balancing accuracy and efficiency in developing ASC models suitable for the low-complexity requirements of the challenge.

4. CONCLUSION

In this technical report, we investigate the balance between accuracy and efficiency in the low-complexity acoustic scene classification task for the DCASE 2023 challenge. By comparing the performance of two approaches—one focusing on accuracy using Res2Net and GhostNet, and the other emphasizing efficiency using MobileFormer—we demonstrate the trade-offs between accuracy and resource efficiency. Our results contribute to the ongoing research on ASC and provide insights into developing robust and lightweight models suitable for embedded systems while addressing the challenges of balancing accuracy and efficiency.

5. REFERENCES

[1] S.-H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. Torr, “Res2net: A new multi-scale backbone architecture,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 2, pp. 652–662, 2021.

[2] K. Han, Y. Wang, Q. Tian, J. Guo, C. Xu, and C. Xu, “Ghostnet: More features from cheap operations,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[3] Y. Chen, X. Dai, D. Chen, M. Liu, X. Dong, L. Yuan, and Z. Liu, “Mobile-former: Bridging mobilenet and transformer,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 5270–5279.

[4] J.-H. Lee, J.-H. Choi, P. M. Byun, and J.-H. Chang, “Hyu submission for the DCASE 2022: Efficient fine-tuning method using device-aware data-random-drop for device-imbalanced acoustic scene classification,” DCASE2022 Challenge, Tech. Rep., June 2022.

[5] B. Kim, S. Yang, J. Kim, and S. Chang, “QTI submission to DCASE 2021: Residual normalization for device-imbalanced acoustic scene classification with efficient design,” DCASE2021 Challenge, Tech. Rep., June 2021.

[6] H. Hu, C.-H. H. Yang, X. Xia, X. Bai, X. Tang, Y. Wang, S. Niu, L. Chai, J. Li, H. Zhu, F. Bao, Y. Zhao, S. M. Siniscalchi, Y. Wang, J. Du, and C.-H. Lee, “Device-robust acoustic scene classification based on two-stage categorization and data augmentation,” DCASE2020 Challenge, Tech. Rep., June 2020.

[7] B. Kim, S. Chang, J. Lee, and D. Sung, “Broadcasted residual learning for efficient keyword spotting,” *arXiv preprint arXiv:2106.04140*, 2021.

[8] T. Liu, R. K. Das, K. A. Lee, and H. Li, “Mfa: Tdnn with multi-scale frequency-channel attention for text-independent speaker verification with short utterances,” in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7517–7521.

[9] Y. Jung, S. M. Kye, Y. Choi, M. Jung, and H. Kim, “Improving multi-scale aggregation using feature pyramid module for robust speaker verification of variable-duration utterances,” *arXiv preprint arXiv:2004.03194*, 2020.

[10] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, “Mobilenetv2: Inverted residuals and linear bottlenecks,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[12] T. Heittola, A. Mesaros, and T. Virtanen, “TAU Urban Acoustic Scenes 2022 Mobile, Development dataset,” Mar. 2022. [Online]. Available: https://doi.org/10.5281/zenodo.6337421

[13] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, “SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition,” in *Proc. Interspeech 2019*, 2019, pp. 2613–2617.

[14] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=Bkg6RiCqY7