JLESS SUBMISSION TO DCASE2023 TASK1: COMPRESSED MODEL WITH SELF ATTENTION BLOCK FOR LOW-COMPLEXITY ACOUSTIC SCENE CLASSIFICATION

Technical Report

Yutong Du[1], *Jisheng Bai*[1,2], *Zijun Pu*[1], *Jianfeng Chen*[1,2]

[1] Joint Laboratory of Environmental Sound Sensing,
School of Marine Science and Technology,
Northwestern Polytechnical University, Xi'an, China
[2] LianFeng Acoustic Technologies Co., Ltd. Xi'an, China
{ytdu, baijs, pzjscholar}@mail.nwpu.edu.cn, chenjf@nwpu.edu.cn

ABSTRACT

In this technical report, we describe our proposed system for DCASE task1: Low-Complexity Acoustic Scene Classification. First, To obtain better performance than Baseline, we choose ResNet as basic model, and add several self-attention blocks including CBAM and MHSA to get more fine-grained features and temporal features respectively from spectrogram. In order to pay attention to detailed information, add the CBAM block between two convolution layers in the ResNet block. The MHSA aims to get temporal context relationships in the spectrum. Another requirement of this task is Low-Complexity, thus, the regular convolution module is replaced by the depthwise separable convolution module in the proposed model. During experiments, we use FMix as data augmentation to improve generalization. Moreover, we use a hard-task training strategy in training process.

Index Terms— ResNet, depthwise separable convolution, CBAM, MHSA, hard-task training

1. INTRODUCTION

The goal of DCASE 2023 Task1[1] is to realize acoustic scene classify accurately with a low complexity model. In order to reach this goal, we divide this task into two phase. First, design a network based on variants of convolutional neural networks(CNN), which has a great performance on acoustic scene classification. As a kind of efficient CNN models ResNet[2] exploit shortcut connection eliminate the degradation problems in deep CNN, thus we choose ResNet as our convolutional block. Although CNN is widely used in computer vision task, but it lack ability of dealing with temporal task. Base on this problem, we regard CRNN[3] as our main network structure. Most deep learning works[4, 5] relative to acoustic are base on CRNN structure. To make network framework more efficient we added two attention block to improve performence. One is Convolutional Block Attention Module (CBAM)[6] which was proposed by Sanghyun Woo etc. , used to obtain fine-grained features. Another is Multi Head Self Attention (MHSA) [7] , which is key components of transformer. Second, we try to compress model by replace tradition CNN with depthwise separable convolution (DSC) [8] , which is widely used in MobileNet[9, 10] and SqueezeNet[11] and other light network structure. To verify the model performance, we conducted a series of experiments. The contributions of this technical report are organized as follows: Chapter 2 describes the details of proposed algorithm framework. Chapter 3 describes the experiments and analysis results. Chapter 4 draws the conclusion.

2. PROPOSED METHOD

2.1. Datasets

The development dataset for DCASE 2023 challenge task1 is TAU Urban Acoustic Scenes 2022 Mobile development dataset[12]. The development set contains 230350 audio segments from 10 acoustic scenes: ”airport”,”bus”,”metro”,”metro station”,”park”,”public square”,”shopping mall”,”street pedestrian”,”street traffic” and ”tram”. Duration of each segment is 1 second, in order to comply with the inference time and computational limitations imposed by the considered target devices.

2.2. Preprocessing

The input features are extracted from the audio signals using a Short Time Fourier Transformation (STFT) with fft size of 4096 and an overlap of 50 %. We apply a Mel-scaled filter bank to end up with 256 frequency bins. We try two data augmentation methods in our system: FMix [13] and mixup [14]. In mixing augmentation method, the data and labels are mixed to generate new training data. Experiment shows that FMix works better than mixup in this task.

2.3. Network architecture

We propose two networks based on ResNet module . The regular convolution module is replaced by the deep separable convolution (DSC) module in the ResNet module. Using the deep separable convolution module significantly reduces the number of parameters.

Two self-attention blocks including Convolutional Block Attention Module (CBAM) and Multi-head Self-Attention (MHSA) are used in our networks. These two self-attention blocks achieve significant performance in computer version (CV).

In the first network, we only use ResNet module and MHSA attention block. And in the second network, we use ResNet module with both MHSA and CBAM attention block.The details of

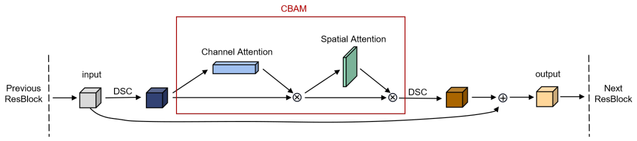

Figure 1: CA-ResBlock. We put CBAM attention between two DSC layers. DSC denotes the deep separable convolution.

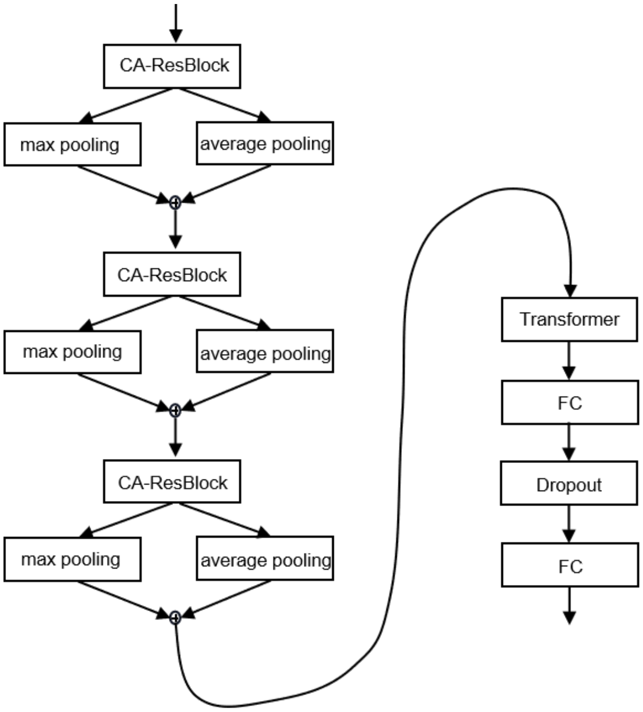

Figure 2: The whole construction of our network.

ResNet module with CBAM can be found in Figure 1.We call it CA-ResBlock(Compressed Attention ResBlock). In baseline network, only max pooling is used. We use average pooling and max pooling instead. The whole construction of our network can be found in Figure 2.

2.4. Training strategy

In this task, we use a hard-task training strategy.That means we validate each five epochs, and sort the accuracy of ten scenes from lowest to highest. Retrain the first three scenes at the same time as normal training.

3. EXPERIMENTS

3.1. Experimental setup

We evaluated our proposed network on the development dataset. Each network is trained for 120 epochs (hard-task training for three bad scenes occupies 20 epochs, while 100 epochs are training for all scenes). The sampling frequency is set to 44.1kHz, which is the same as baseline. Our batch size is 64. The input is 25 frames and 256 mel bins. We use Adam[15] with a weight decay of 0.05 and a learning rate schedule of 0.001. Cross-entropy loss function is used during the training process. We find using FMix has better performance than mixup, so the results are experiments with FMix.

3.2. Results

We named the network using ResNet module and both MHSA and CBAM attention as ”ours 1”, and the network only using ResNet module and MHSA attention as ”ours 2”. Table 1 shows the accuracy with the development set for proposed networks. Table 2 shows the loss with the development set for proposed networks. Table 3 shows the accuracy for different devices. Table 4 shows the parameters and MMACs of our model and baseline.

Table 1: Accuracy of our model and Baseline for different scenes

Scene	baseline	ours 1	ours 2
airport	39.4%	42.6%	43.6%
bus	29.3%	54.4%	61.6%
metro	47.9%	48.4%	48.9%
metro station	36.0%	51.9%	40.2%
park	58.9%	67.4%	69.4%
public square	20.8%	28.4%	27.2%
shopping mall	51.4%	63.1%	59.0%
street pedestrian	30.1%	26.5%	23.5%
street traffic	70.6%	75.0%	78.2%
tram	44.6%	50.7%	53.7%
overall	42.9%	**50.8**%	**50.5**%

Table 2: Logloss of our model and Baseline for different scenes

Scene	baseline	ours 1	ours 2
airport	1.534	3.974	3.939
bus	1.758	4.136	3.962
metro	1.382	3.021	2.947
metro station	1.672	1.407	1.714
park	1.448	4.962	4.988
public square	2.265	3.939	3.943
shopping mall	1.385	3.668	3.819
street pedestrian	1.822	3.697	3.947
street traffic	1.025	4.127	3.852
tram	1.462	3.662	3.668
overall	1.575	**1.436**	**1.467**

Table 3: Accuracy of our model for different devices

Device	ours 1	ours 2
a	65.0%	65.2%
b	56.6%	53.7%
c	60.1%	60.3%
s1	48.9%	48.8%
s2	48.7%	47.4%
s3	49.5%	51.5%
s4	42.9%	43.5%
s5	44.6%	44.9%
s6	41.1%	39.5%

Table 4: Parameters and MMACs of our model and baseline

	baseline	ours 1	ours 2
Parameters	46.512 K	**78.252 K**	**60.458 K**
MMACs	29.23 M	**27.93 M**	**14.13 M**

4. CONCLUSION

In this technical report, we construct a compressed model for low-complexity acoustic scene classification by using ResNet as major module, and both CBAM and MHSA as self attention module. The compressed method is DSC. On the basis of CRNN we replace CNN to ResNet block with adding CBAM in it, and RNN was replaced by MHSA. Verified by experiments, our model has average performance which accuracy is 50.83% and nummber of parameters is 78.252 K, MMACs is 27.93 M. In a word, our model can complete the target task with limit on parameters.

5. REFERENCES

[1] I. Martín-Morató, F. Paissan, A. Ancilotto, T. Heittola, A. Mesaros, E. Farella, A. Brutti, and T. Virtanen, "Low-complexity acoustic scene classification in dcase 2022 challenge," *arXiv preprint arXiv:2206.03835*, 2022.

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, p. 770–778.

[3] B. Shi, X. Bai, and C. Yao, "An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 11, p. 2298–2304, 2017.

[4] Y. Cao, Q. Kong, T. Iqbal, F. An, W. Wang, and M. Plumbley, "Polyphonic sound event detection and localization using a two-stage strategy," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*, New York University, NY, USA, Oct 2019, p. 30–34.

[5] A. Politis, S. Adavanne, and T. Virtanen, "A dataset of reverberant spatial sound scenes with moving sources for sound event localization and detection," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2020 Workshop (DCASE2020)*, Tokyo, Japan, Nov 2020, p. 165–169. [Online]. Available: https://dcase.community/workshop2020/proceedings

[6] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European Conference on Computer Vision (ECCV)*, Sep 2018.

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 6000–6010, event-place: Long Beach, California, USA.

[8] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, p. 1800–1807.

[9] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv e-prints*, p. arXiv:1704.04861, Apr 2017.

[10] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, p. 4510–4520.

[11] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "Squeezenet: Alexnet-level accuracy with 50x fewer parameters and ¡0.5mb model size," *arXiv e-prints*, p. arXiv:1602.07360, Feb 2016.

[12] A. Mesaros, T. Heittola, and T. Virtanen, "A multi-device dataset for urban acoustic scene classification," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE2018)*, Nov 2018, p. 9–13. [Online]. Available: https://dcase.community/documents/workshop2018/proceedings/DCASE2018Workshop_Mesaros_8.pdf

[13] E. Harris, A. Marcu, M. Painter, M. Niranjan, A. Prügel-Bennett, and J. Hare, "FMix: Enhancing Mixed Sample Data Augmentation," *arXiv e-prints*, p. arXiv:2002.12047, Feb. 2020.

[14] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond Empirical Risk Minimization," *arXiv e-prints*, p. arXiv:1710.09412, Oct. 2017.

[15] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv e-prints*, p. arXiv:1412.6980, Dec 2014.