Domain Generalization on Efficient Acoustic Scene Classification Using Residual Normalization

Byeonggeun Kim, Seunghan Yang, Jangho Kim, and Simyung Chang

Qualcomm AI Research*, Qualcomm Korea YH, Seoul, Republic of Korea

Qualcomm AI research

Problem Statement

Task: DCASE2021 challenge, TASK1A: low-complexity acoustic scene classification (ASC) with multiple devices.

- ASC is the task of classifying sound scenes such as "airport," "train station," and "urban park" to which user belongs.
- This year, the task becomes more challenging as an ASC model needs to solve two problems simultaneously.

1) Data is collected from multiple devices, and the number of samples per device is unbalanced.
2) TASK1A restricts the model size.

Contributions

1) We **introduce a network architecture for ASC that utilizes broadcasted residual learning [1]** and achieve higher accuracy while reducing the size by a third of the baseline [2].
2) We **propose a novel normalization method, Residual Normalization (ResNorm)**, which can **leverage the generalization performance** for unseen devices.
3) Finally, we **describe model compression** combined with **pruning and quantization** to satisfy the model complexity of the task while maintaining performance using **knowledge distillation**.
4) Got the 1st place in TASK1A of DCASE2021.

Our Approach

1) Network Architecture

While the BC-ResNet [1] targets human voice, we aim to classify audio scenes. To adapt to the differences in input domains, we make two modifications, i.e., limit the receptive field and use max-pool instead of dilation.

Table 1: **BC-ResNet-ASC.** Each row is a sequence of one or more identical modules repeated n times with input shape of frequency by time by channel and total time step T.

Input	Operator	n	Channels
$256 \times T \times 1$	conv2d 5x5, stride 2	-	2c
$128 \times T/2 \times 2c$	stage1: BC-ResBlock	2	c
$128 \times T/2 \times c$	max-pool 2x2	-	-
$64 \times T/4 \times c$	stage2: BC-ResBlock	2	1.5c
$64 \times T/4 \times 1.5c$	max-pool 2x2	-	-
$32 \times T/8 \times 1.5c$	stage3: BC-ResBlock	2	2c
$32 \times T/8 \times 2c$	stage4: BC-ResBlock	3	2.5c
$32 \times T/8 \times 2.5c$	conv2d 1x1	-	num class
$32 \times T/8\times$ num class	avgpool	-	-
$1 \times 1\times$ num class	-	-	-

Table 2: **Network Architectures.** Compare Top-1 test accuracy (%) on TAU Urban AcousticScenes 2020 Mobile, development dataset.

Network Architecture	#Param	Top-1 Acc. (%)
CP-ResNet, c=64	899k	67.8
BC-ResNet-8, num SSN group = 4	317k	68.6 ± 0.4
BC-ResNet-ASC-8	315k	69.5 ± 0.3

2) Residual Normalization.

Motivation: We observe that differences between audio devices are revealed along frequency dimension rather than channel dimension.

Instance Normalization (IN) has been a representative approach to eliminate instance-specific domain discrepancy. Here we **use instance normalization by frequency (FreqIN) instead IN.**

$$FreqIN(x) = \frac{x - \mu_{nf}}{\sqrt{\sigma^2_{nf} + \epsilon}}, \qquad (1)$$

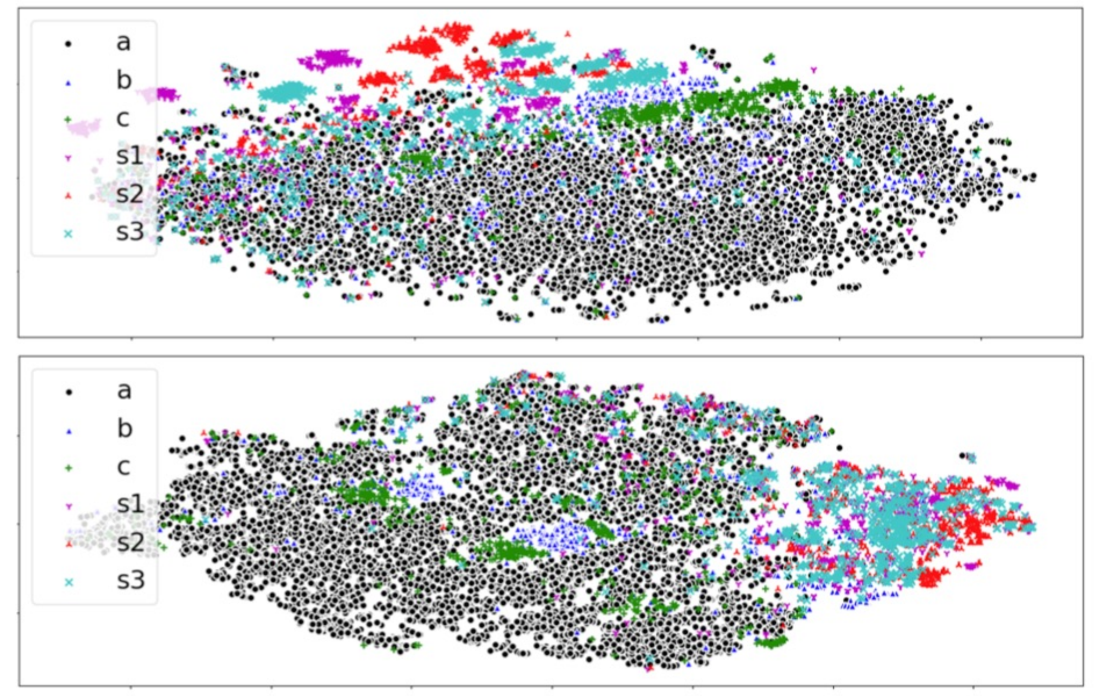

Figure 1: **2D t-SNE [19] visualization** of feature maps of BC-ResNet-ASC-1 stage2 (without ResNorm). **Top:** Concatenation of frequency-wise mean and standard deviations. **Bottom:** Concatenations of channel mean and standard deviations. The training samples are separated better by device ID (A to S3) with frequency-wise statistics.

where,

$$\mu_{nf} = \frac{1}{CT}\sum_{c=1}^{C}\sum_{t=1}^{T} x_{ncft},$$

$$\sigma^2_{nf} = \frac{1}{CT}\sum_{c=1}^{C}\sum_{t=1}^{T} (x_{ncft} - \mu_{nf})^2. \qquad (2)$$

FreqIN is task-agnostic and can result to loss of useful information for classification. **To compensate for information loss, we add an identity shortcut** multiplied by a hyperparameter. We use ResNorm at input and after every stage in the model.

$$ResNorm(x) = \lambda \cdot x + FreqIN(x). \qquad (3)$$

3) Model Compresssion

- Magnitude based one-shot unstructured **pruning**
- **Quantize** all conv layers as an 8-bit while utilize half-precision for others.

Knowledge distillation (KD) compensates the performance drop due to compression.

Table 4: **Model compression** Compare bitwidth, top-1 test accuracy (%) on Tau Urban AcousticScenes 2020 Mobile, development dataset, and pruning ratio of the models (Average over 6 seeds).

BC-ResNet-ASC-8 + ResNorm, 300 epochs, KD				
Method	Bitwidth	KD	Pruning	Accuracy
Vanilla model	32	-	-	76.3 ± 0.8
Compressed model	8, 16		0.89	75.1 ± 0.9
Compressed model	8, 16	✓	0.89	75.3 ± 0.8

Quantitative Results

Table 3: **Residual Normalization.** We demonstrate how residual normalization affects BC-ResNet-ASC on TAU Urban AcousticScenes 2020 Mobile, development dataset. We show mean and standard deviation of Top-1 test accuracy (%) (averaged over 3 seeds, * averaged over 6 seeds).

Method	#Param	A	B	C	S1	S2	S3	S4	S5	S6	Overall
BC-ResNet-ASC-1 (Baseline)	8.1k	73.1	61.2	65.3	58.2	57.3	66.2	51.5	51.5	46.3	58.9 ± 0.8
BC-ResNet-ASC-1 + Global FreqNorm	8.1k	73.9	60.9	65.5	60.2	57.9	67.9	50.2	54.3	49.4	60.0 ± 0.9
BC-ResNet-ASC-1 + Fixed PCEN	8.1k	68.0	60.4	57.2	64.0	63.0	66.2	62.3	61.8	56.5	62.2 ± 0.8
BC-ResNet-ASC-1 + ResNorm	8.1k	**76.4**	65.1	**68.3**	**66.0**	62.2	**69.7**	63.0	63.0	58.3	*65.8 ± 0.7
w/o ResNorm in Network	8.1k	75.1	**68.9**	67.0	**66.0**	63.9	69.3	63.4	**66.9**	**63.6**	$\mathbf{67.1 \pm 0.8}$
w/o Shortcut	8.1k	68.2	62.1	58.6	64.2	**65.3**	66.3	**65.1**	63.8	61.3	63.9 ± 0.7
BC-ResNet-ASC-8 + ResNorm	315k	**81.3**	**74.4**	**74.2**	**75.6**	**73.1**	**78.6**	73.0	**74.0**	**72.7**	***$\mathbf{75.2 \pm 0.4}$**
w/o ResNorm in Network	315k	80.8	73.7	73.0	74.0	72.9	77.8	**73.3**	72.1	71.0	74.3 ± 0.3
w/o Shortcut	315k	78.3	73.5	69.1	73.8	72.9	75.6	72.2	72.5	71.0	73.2 ± 0.3

References

[1] B. Kim, S. Chang, J. Lee, and D. Sung, "Broadcasted Resid-ual Learning for Efficient Keyword Spotting," inProc. Inter-speech 2021, 2021, pp. 4538-4542.

[2] K. Koutini, H. Eghbal-zadeh, M. Dorfer, and G. Widmer, "Thereceptive field as a regularizer in deep convolutional neu-ral networks for acoustic scene classification," inEUSIPCO.IEEE, 2019, pp. 1-5.

*Qualcomm AI Research is an initiative of Qualcomm Technologies, Inc