TRIDENT RESNETS WITH LOW COMPLEXITY FOR ACOUSTIC SCENE CLASSIFICATION

Technical Report

*Youngho Jeong, Sooyoung Park, Taejin Lee**

Media Coding Research Section
Electronics and Telecommunications Research Institute
218 Gajeong-ro, Yuseong-gu, Daejeon, Republic of Korea
{yhcheong, sooyoung, tjlee}@etri.re.kr

ABSTRACT

This technical report describes our acoustic scene classification systems for DCASE2021 challenge Task1 subtask A. We designed two Trident ResNets with three parallel paths, which is targeted to low complexity. The trident structure with respect to the frequency domain is beneficial when analyzing samples collected from minority or unseen devices. To satisfy the model complexity requirement, we replaced a standard convolution with a depthwise separable convolution and applied weight quantization to the trained model. As a result of performance evaluation, our system trained by applying data augmentation showed a log loss of 0.968 and a classification accuracy of 65.8% for the test split.

Index Terms—Acoustic Scene Classification, Trident-ResNet, Depthwise Separable Convolution, Weight Quantization, SpecAugment

1. INTRODUCTION

Acoustic Scene Classification (ASC) is a task of classifying given data into one of the predefined acoustic scene classes. This year, ASC task was released in two subtasks: Subtask A for low-complexity acoustic scene classification with multiple devices, and Subtask B for audio-visual scene classification [1]. The main issue of the subtask A is to design a classifier with low complexity that works stably on various devices. A model size limit is 128 KB, which corresponds to 32,768 parameters of float32, and the evaluation dataset includes data recorded with new devices that has not appeared in the development dataset.

In the following sections, we describe our proposed models for subtask A, training methods, and evaluation results.

2. DATASETS

The development dataset of TAU Urban Acoustic Scene 2020 Mobile contains 23,035 samples. Each sample corresponds to one class out of ten, and there is no sample with multiple labels. This dataset consists of various audio samples collected from three real devices and six simulated devices. Most of the data were collected from Zoom F8 audio recorder with a binaural microphone, and data from Samsung Galaxy S7, iPhone SE are also included. The simulated devices are synthesized by processing the data of device A with various impulse responses and additional dynamic range compression. The organizer of the challenge provides basic metadata of training/test split consisting of 13,962 samples in the training set and 2,968 samples in the test set. The evaluation dataset of TAU Urban Acoustic Scene 2021 Mobile, which contains 7,920 samples, also includes audio data from the new devices such as a GoPro Hero5 Session and the five simulated devices [1].

3. SYSTEM ARCHITECTURE

3.1. Feature Extraction

The data are mono audio files with 44.1 kHz sample rate. We transformed them into power spectrogram by skipping every 1024 samples with 2048 length Hann window. A spectrum of 431 frames was yielded from 10 seconds audio file, and each spectrum was compressed into 256 bins using Mel-scaled filter bank. Additionally, deltas and delta-deltas were calculated from the log Mel spectrogram and stacked into the channel axis. The number of frames of the input feature is cropped by the length of the delta-delta channel so that the final shape becomes [256×423×3].

3.2. Data Augmentation

We only utilized training split of the challenge dataset, and applied two data augmentation techniques to increase the diversity of data distribution. Our data augmentation strategies are listed in Table 1. The start frame of temporal cropping was randomly selected in the previous half of the entire frame. SpecAugment was applied to only 30% of the total training data. The augmented data were generated from each mini-batch consisting of 64 samples during the training process in real-time.

Table 1: List of data augmentation strategies

Strategy	Parameter
Temporal cropping	Crop length : 5 seconds
SpecAugment [2]	Maximum length of masking - time : 10 frames, frequency : 4 bins

3.3. Trident ResNets with Low Complexity

Two different structures of Trident ResNet with low complexity are proposed according to the processing method in the residual

* Thanks to Korea government (MSIT) for funding.

block with different number of channels between input and output feature map.

3.3.1. Trident ResNet A

Based on the our Trident ResNet submitted last year [3], we redesigned a residual block A (Res-block A) to achieve low model complexity as shown in Fig. 1 [4]. It consists of two DSC (Depthwise Separable Convolution) blocks [5] and an identity path with zero-padding after average pooling to avoid mismatch between input and output.

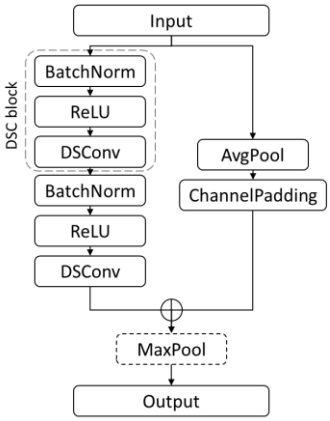

Figure 1: Residual block A with low complexity

Each DSC block is constructed in the order BN-ReLU-DSConv. Gamma and beta terms are not used in Batch Normalization layers, and there is no bias term in DSConv layers.

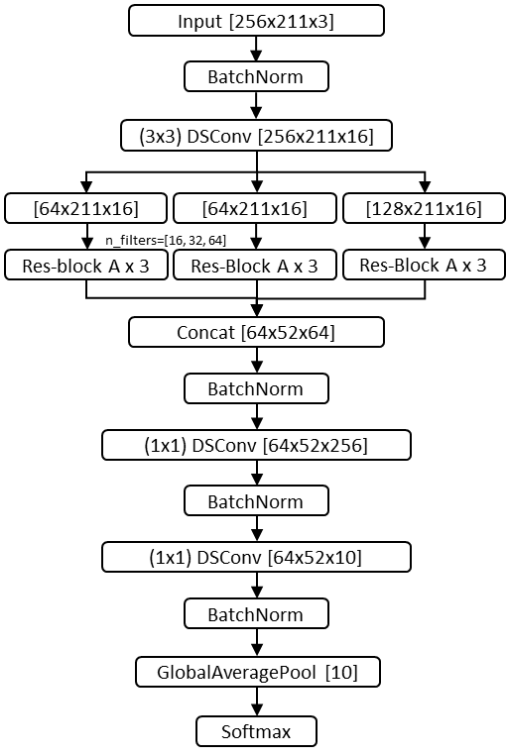

Figure 2: Overall structure of Trident ResNet A

(3x3) DSConv layer applied for model size reduction is the same as the combination of (3x3) depthwise convolution and (1x1) pointwise convolution and neither strides nor dilation in convolution layer is applied. Kernels are initialized with He normal distribution [6] and regularized with L2 regularization of 5×10^{-4}. Max pooling with 2x2 strides is applied to the remaining blocks except the first of three residual blocks. The number of filters applied to the three residual blocks is 16, 32, and 64.

We arranged the residual block in parallel and concatenated their outputs for classification. To learn effectively distinct features from different frequency bands, our model is composed of a trident structure, consisting of 0-63, 64-127, and 128-255 Mel bins [4], [7]. After concatenating the outputs from each network, two blocks of 1×1 convolution and Global Average Pooling (GAP) calculates the classification scores. The overall structure of Trident ResNet A is shown in Fig. 2.

3.3.2. Trident ResNet B

Residual block B (Res-block B) is different from Res-block A in that max pooling is optionally applied after skip connection processing as shown in Fig. 3. Max pooling with 2x2 strides is applied every even-numbered Res-block B.

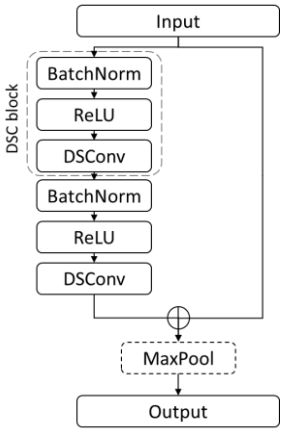

Figure 3: Residual block B with low complexity

The overall structure of Trident ResNet B is shown in Fig. 4 and the number of filters applied to the six residual blocks is [16, 16, 32, 32, 64, 64]. The padding option of (3x3) Conv2D located in the front of the model is 'valid'. DSC block plays an additional role of matching the number of channels between input and output in Res-block B. Trident ResNet B is constructed deeper than Trident ResNet A by stacking twice as many residual blocks.

3.4. Model Complexity

Table 2 shows the model complexity of Trident ResNets. For weight quantization, the model trained in float32 is converted to float16. The quantized FP16 models satisfy the model size limit of 128 KB.

Table 2: Model complexity of Trident ResNets

Model Name	Total Parameters	FP16 Model Size
Trident ResNet A	54,845	113.9 KB
Trident ResNet B	60,236	124.4 KB

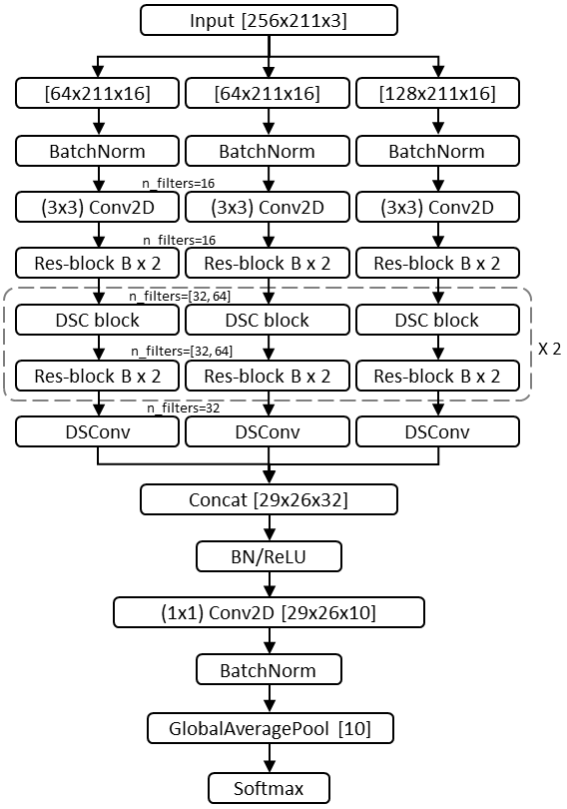

Figure 4: Overall structure of Trident ResNet B

3.5. Loss Function

Focal loss [8] attenuates the log loss generated by well-trained samples, so that the model can focus on the poorly trained samples. The following equation describes focal loss with balancing parameter α, focusing parameter γ and prediction score p_t.

$$FL(p_t) = -\alpha(1-p_t)^{\gamma}\log(p_t) \quad (1)$$

Increasing the value of γ increases the sensitivity of the model to misclassified samples, and α scales the loss function linearly. Our setting for γ and α was 2.0 and 0.25, respectively.

3.6. Learning Rate Scheduler

We trained our model using Stochastic Gradient Descent (SGD) [9] optimizer with a momentum of 0.9. The learning rate η_t is controlled by a cosine annealing scheduler using (2), (3) and restarts at 2, 6, 14, 30, 62, 126, 254 epochs.

$$\eta_t = \eta_{min}^i + \frac{1}{2}\left(\eta_{max}^i - \eta_{min}^i\right)\left(1+\cos\left(\frac{T_{cur}}{T_i}\pi\right)\right) \quad (2)$$

$$\left[\eta_{min}^{i+1}, \eta_{max}^{i+1}\right] = \beta \cdot \left[\eta_{min}^i, \eta_{max}^i\right] \quad (3)$$

The initial value of η_{min}^i and η_{max}^i, which is 10^{-1} and 10^{-5} respectively, decreases by 10% ($\beta = 0.9$) for each restart to explore deeper areas on the hyperplane.

4. RESULTS

This section reports the macro-average multiclass cross-entropy (log loss) and macro-average accuracy (average of the class-wise accuracies) of our submitted systems for the training/test split. In the Table 3, Trident ResNet A and B correspond to the results of applying only the temporal corp. As can be seen from the results, additional SpecAugment helps to improve both log loss and accuracy for Trident ResNet B, but not for Trident ResNet A. It seems that further analysis is needed on the various parameter combinations in SpecAugment to find out the specific cause for these results.

Table 3: Test split results of subtask A development set

ID	System Name	Log Loss	Accuracy
-	DCASE2021 Task1A Baseline	1.473	47.7%
1	Trident ResNet A	1.006	**65.9%**
2	Trident ResNet A + SpecAug	1.015	64.9%
3	Trident ResNet B	1.014	64.6%
4	Trident ResNet B + SpecAug	**0.968**	65.8%

To submit the prediction results for the evaluation dataset, the proposed systems were trained using the entire development dataset.

5. CONCLUSION

We proposed acoustic scene classification models based on trident architecture for DCASE 2021 Task 1 subtask A. To satisfy a model size requirement, a depthwise separable convolution and weight quantization was adopted. The two Trident ResNets have a difference in the structure and the stacking depth of basic residual block. In the evaluation using development dataset, our proposed model showed a log loss of 0.968 and an accuracy of 65.8%, which improved by 0.505 and 18.1% respectively over the baseline system.

6. ACKNOWLEDGMENT

This work was supported by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIT) (No.2017-0-00050, Development of Human Enhancement Technology for auditory and muscle support).

7. REFERENCES

[1] http://dcase.community/challenge2021/task-acoustic-scene-classification.

[2] Park, D.S., Chan, W., Zhang, Y., Chiu, C.C., Zoph, B., Cubuk, E.D., and Le, Q.V., "SpecAugment: A simple data augmentation method for automatic speech recognition," in *Proceedings of the Interspeech*, pp.15-19, Sep. 2019.

[3] Sangwon Shu, Sooyoung Park, Youngho Jeong, and Taejin Lee, "Designing Acoustic Scene Classification Models with CNN Variants," *DCASE Technical Report*, 2020.

[4] He, K., Zhang, X., Ren, S., and Sun, J., "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.770-778, 2016.

[5] Francois Chollet, “Xception: Deep Learning With Depthwise Separable Convolutions,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.1251-1258, 2017.

[6] He, K., Zhang, X., Ren, S., and Sun, J., “Delving deep into rectifiers: Surpassing human-level performance on imagenet classification,” in *Proceedings of the IEEE international conference on computer vision*, pp.1026-1034, 2015.

[7] Mark D. MacDonnell, and Wei Gao, “Acoustic scene classification using deep residual networks with late fusion of separated high and low frequency paths,” in *Proc. IEEE ICASSP*, pp.141-145, 2020.

[8] Lin, T. Y., Goyal, P., Girshick, R., He, K., and Dollár, P., “Focal loss for dense object detection,” in *Proceedings of the IEEE international conference on computer vision*, pp.2980-2988, 2017.

[9] Ilya Loshchilov, and Frank Hutter, “SGDR: Stochastic Gradient Descent with Warm Restarts,” in *International Conference on Learning Representations (ICLR)*, 2017.