CNN-BASED FRAMEWORK FOR DCASE 2020 TASK 1B CHALLENGE

Technical Report

Dat Ngo[1∗], *Lam Pham*[2†], *Anh Nguyen*[4‡], *Hao Hoang*[3§],

Ho Chi Minh City University of Technology, Vietnam

ABSTRACT

This technical report presents a low-complexity CNN-based deep learning framework for acoustic scene classification. Particularly, the proposed architecture constitute of two main steps front-end feature extraction and back-end network. Firstly, spectrogram representation is approached as front-end feature extraction in this framework. Next, the spectrograms extracted are fed into a CNN-based architecture for classification. Obtained experimental results conducted over the DCASE 2020 Task 1B dataset improve DCASE baseline by 7.2%.

Index Terms— Convolutional Neural Network (CNN), pruning, quantization, mixup data augmentation, spectrogram, Gammatone filter.

1. INTRODUCTION

The increasing development of deep learning has provided powerful solutions for various research fields such as computer vision, natural language processing, and recently emerging research field named “machine hearing” [1]. As regards acoustic scene classification (ASC), one of main tasks of “machine listening”, authors in [2] showed that CNN-based network architectures and their robustness has surpassed human performance. However, the state-of-the-art systems show increasing cost of computation due to complicated network architecture used. This makes systems more difficult to apply in real-life applications, specially for low-power and real-time systems. For instance, almost top-ten performance architectures in recent DCASE 2018 [3] and 2019 [4] exceeded 6 million non-zero parameters according to the summary of system characteristics [4], indicating that the more effectiveness in models, the more complexity they have to handle. Thank to DCASE 2020 Task 1B [5], it propose a challenge which requires low complexity model applied for ASC task, thus encourage to propose low-complexity models for edge applications in terms of ASC.

Within this report, we propose a deep learning framework with low-complexity CNN-based model for the ASC task, thus evaluate and compare to DCASE 2020 Task 1B baseline.

2. DCASE 2020 TASK 1B DATASET

The DCASE 2020 Task 1B dataset [5] was recorded by a single device namely A with binaural channel and sample rate of 48kHz. The dataset include 10 acoustic scenes that are grouped into three main contexts of indoor (airport, metro-station, and shopping-mall), outdoor (park, public-square, street-pedestrian, street-traffic), and transportation (bus, metro, tram). In this report, by exploiting DCASE 2020 challenge to evaluate our performance, we separate development set into training and test subsets used for training and testing processes, respectively. Next, we report the accuracy on the test subsets.

3. PROPOSE CNN-BASED FRAMEWORK ARCHITECTURE

The proposed framework is described in Fig. 1.As Figure 1 shown, the framework is separated into low-level feature extraction (the upper part) and back-end classification (the lower part). Initially, raw audio waveform from the channel 1, channel 2 and channel ave are transformed into Gammatone spectrograms (GAM) [6] with parameters set in Table 1. After that, the entire spectrogram of 128×1728 is thus split into non-overlapped image patches of 128×128. Then, data augmentation techniques, comprising randomly oversample and the mixup [7], are applied on small image patches, thus generate new image patches. Next, the new image patches are fed into a CNN-based network for classification.

The CNN-based network architecture comprises four Conv. blocks and one Dense block which are performed by Convolutional layer (Cv[kernel size]), Rectified Linear Unit (Relu), Batch normalization (Bn), Average pooling (Ap[kernel size]), Global average pooling (Gap), Drop out (Dr(Drop ratio)), Fully connected layer (Fl), and Softmax layers as shown in Table 2.

4. HYPERPARAMETER SETTING

The CNN-based network implemented use TensorFlow framework. Network training makes use of the Adam optimiser [8] with 100 training epochs, a mini batch size of 100. As using mixup data augmentation makes labels are no longer one-hot encoded, Kullback-

Table 1: Setting of spectrogram transformation.

Factors	Setting
Spectrogram	Gammatone
Window size	2208
Hop size	256
FFT number	4096
Filter banks number	128
Min frequency	10 Hz

[∗] datt.ngo.hcmut@gmail.com

[†] lamd.pham@hcmut.edu.vn

[‡] anh.nguyenk2017@hcmut.edu.vn

[§] phuhao1998@gmail.com

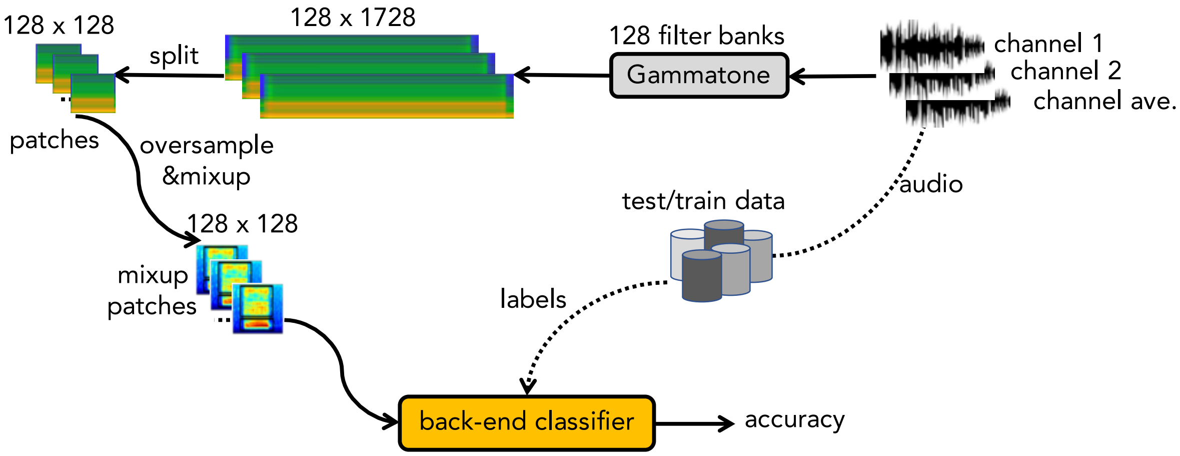

Figure 1: The high-level architecture and processing sequence of the proposed framework.

Table 2: CNN-based network architecture

Architecture	layers	Output
	Input layer (entire spectrogram)	128×128×3
Conv. Block 01	Bn - Cv [3×3] - Relu - Bn - Ap [2×2] - Dr (20%)	64×64×32
Conv. Block 02	Bn - Cv [1×1] - Relu - Bn - Ap [2×2] - Dr (25%)	32×32×64
Conv. Block 03	Bn - Cv [3×3] - Relu - Bn - Ap [2×2] - Dr (30%)	16×16×128
Conv. Block 04	Bn - Cv [1×1] - Relu - Bn - Gap - Dr (35%)	256
Dense Block	Fl - Softmax layer	3

Leibler (KL) [9] divergence loss is therefore used as,

$$L_{KL}(\theta) = \sum_{n=1}^{N} \mathbf{y}_n \log \left\{ \frac{\mathbf{y}_n}{\hat{\mathbf{y}}_n} \right\} + \frac{\lambda}{2} ||\theta||_2^2. \quad (1)$$

where θ denotes the trainable network parameters and λ denote the ℓ_2-norm regularization coefficient, set to 0.0001. N is the batch number, $\mathbf{y}_i$ and $\hat{\mathbf{y}}_i$ denote expected and predicted results, respectively.

5. EXPERIMENTAL RESULTS AND CONCLUSION

This report has presented a robust framework applying for ASC task. As a result, First, we achieve very competitive results of 7.2%, compared to the baseline in DCASE 2020 Task 1B challenges.

Table 3: Performance compared to DCASE 2020 Task 1B baseline

System	Acc.(%)	Non-zero para. (K)
DCASE 2020	87.3	450
Our system	94.5	445

6. REFERENCES

[1] R. F. Lyon, *Human and Machine Hearing*. Cambridge University Press, 2017.

[2] L. Ma, D. J. Smith, and B. P. Milner, "Context awareness using environmental noise classification," in *Eighth European Conference on Speech Communication and Technology*, 2003.

[3] D. 2018, *System Characteristics*, available at http://dcase.community/challenge2018/task-acoustic-scene-classification-results-b#system-characteristics.

[4] D. 2019, *System Characteristics*, available at http://dcase.community/challenge2019/task-acoustic-scene-classification-results-a#system-characteristics.

[5] 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)*, November 2018, pp. 9–13.

[6] D. P. W. Ellis, *Gammatone-like spectrogram*, 2009, available at http://www.ee.columbia.edu/dpwe/resources/matlab/gammatonegram.

[7] K. Xu, D. Feng, H. Mi, B. Zhu, D. Wang, L. Zhang, H. Cai, and S. Liu, "Mixup-based acoustic scene classification using multichannel convolutional neural network," in *Pacific Rim Conference on Multimedia*, 2018, pp. 14–23.

[8] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[9] S. Kullback and R. A. Leibler, "On information and sufficiency," *The annals of mathematical statistics*, vol. 22, no. 1, pp. 79–86, 1951.