TASK 1A DCASE 2021: ACOUSTIC SCENE CLASSIFICATION WITH MISMATCH-DEVICES USING SQUEEZE-EXCITATION TECHNIQUE AND LOW-COMPLEXITY CONSTRAINT

Technical Report

Javier Naranjo-Alcazar[1,2]*, Sergi Perez-Castanos*[2]*, Maximo Cobos*[2]*, Francesc J. Ferri*[2]*, Pedro Zuccarello*[1]

[1] Instituto Tecnológico de Informática, València, Spain {jnarnajo, pzuccarello}@iti.es
[2] Universitat de València, Burjassot, Spain, {pecaser@alumni.uv.es, {maximo.cobos, francesc.ferri}@uv.es}

ABSTRACT

Acoustic scene classification (ASC) is one of the most popular problems in the field of machine listening. The objective of this problem is to classify an audio clip into one of the predefined scenes using only the audio data. This problem has considerably progressed over the years in the different editions of DCASE. It usually has several subtasks that allow to tackle this problem with different approaches. The subtask presented in this report corresponds to a ASC problem that is constrained by the complexity of the model as well as having audio recorded from different devices, known as mismatch devices (real and simulated). The work presented in this report follows the research line carried out by the team in previous years. Specifically, a system based on two steps is proposed: a two-dimensional representation of the audio using the Gamamtone filter bank and a convolutional neural network using squeeze-excitation techniques. The presented system outperforms the baseline by about 17 percentage points.

Index Terms— Deep Learning, Convolutional Neural Network, Acoustic Scene Classification, Gammatone, DCASE2021

1. INTRODUCTION

Extracting information from audio signals can be a great improvement in existing applications or future products (home assistants, wildlife monitoring, autonomous cars, etc.). Machine listening is understood as the set of algorithms that are capable of extracting relevant information from audio. One of the most common tasks in this field is known as Acoustic Scene Classification (ASC) [1, 2, 3, 4]. The ultimate goal is to extract context information from the audio, more specifically, to predict the location where the audio is produced (park, metro station, airport, etc.). This problem has been addressed in all previous editions of DCASE, and has been modified with different restrictions. In this report an ASC system is designed to be limited by the size of the model and with the extra difficulty that the audios used in the training come from different audio sources (mismatch devices).

The motivation of DCASE 2021 Task 1a is to create an acoustic scene classifier that should work in real-time (low-complexity consideration) and capable of using different recording sources (microphones) [5]. This subtask can be understood as a merge of both subtask in the 2020 edition in which the mismatch problem had no restriction on the model, and on the other hand, the low complexity subtask only used audios from the same recording source.

The approach proposed in this work consists in a CNN implemented with squeeze-excitation modules feed with a 2D audio representation using the Gammatone filter bank. The model is converted to TFLite format in order to accomplish the model size restriction. More information on the proposed framework is presented in Section 2, while the results obtained in the development stage are presented in Section 3. Some conclusions are drawn in Section 4.

2. METHOD

2.1. Audio Representaion

Following the idea of last year submissions [6, 7] a Gammatone filter bank-based representation has been chosen for providing a slightly superior performance than other alternatives (e.g. Mel-scale filter banks) in preliminary tests.

All representations are calculated with a window size of 40 ms with 50% overlapping, using a sampling rate of 44.1 kHz and 64 frequency bins. All frequency bins are normalize with 0 mean and standard deviation equal to 1 using all the provided data. Gammatone representations were computed by using the Auditory Toolbox presented in [8] with Python implementation.

2.2. Convolutional Neural Network

The convolutional network trained with the audio information is composed of blocks defined as *Conv-StandardPOST*. These blocks were proposed in [9]. The aim of these blocks is to achieve improved accuracy by recalibrating the internal feature maps using residual [10] and squeeze-excitation techniques [11, 12]. For more insight about this choice, please see [9] where *Conv-StandardPOST* is fully explained and compared to other competing blocks. The architecture of the network can be seen in Table 1

2.3. Experimental details

2.3.1. Training

The optimizer used is Adam [13]. The loss used is the one known as Focal Loss [14]. This loss function assigns greater emphasis to those samples that are not classified correctly, forcing the system to correctly classify the more challenging samples (those related to devices with lower resolution). The hyperparameters are set as $\alpha = 0.25$ and $\gamma = 2$. During training, the learning rate (which starts at 0.001) is modified by a factor of 0.5 if the validation accuracy does not improve for 20 epochs. Training ends if this metric is not improved for 50 epochs. The maximum number of epochs is 500.

Gammatone representation $(64 \times T \times 1)$
Conv-StandardPost $(\#40, 3, \rho = 2)$
Max Pooling $(1, 10)$
Dropout (0.3)
Conv-StandardPost $(\#40, 3, \rho = 2)$
Max Pooling $(1, 10)$
Dropout (0.3)
Global Average Pooling
Classification (10)

Table 1: Network architecture. *Conv-StandardPost* is denoted with the number of filters ($\#$), the kernel size and the ratio of the squeeze-excitation module (ρ). The Max Pooling layer is defined by the pool size and the Dropout by the rate. The classification layer corresponds to a Dense with 10 units.

2.3.2. Dataset

The dataset provided for the task is known as TAU Urban Acoustic Scenes 2020 Mobile [15]. In turn, this dataset is divided into two splits, the development split and the evaluation split. While the development split contains scenes recorded in 10 cities, the evaluation one contains scenes from 12 cities (there are two cities unseen in the development set). The development split contains audio recorded from 3 real devices and 6 simulated ones. The total amount of hours present in this specific split is 64 hours. The audio is provided in mono, 44.1 kHz of sampling rate and 24-bit format.

3. RESULTS

The results obtained by the system proposed can be seen in Table 2. The proposed approach surpass the baseline by 17 percentage points by only having 5 KB more than the baseline regarding model complexity.

	Accuracy (%)	Model size (KB)
Challenge Baseline	47.70	90.30
Proposed system	**64.18**	95.96

Table 2: Accuracy (%) results obtained compare with the proposed baseline

4. CONCLUSION

Understanding the sounds around us can be a great improvement in a multitude of applications. These solutions must deal with certain issues that may arise. In this task, scene classification problem is proposed with the extra issues of mismatch devices (available audios come from different sources) and complexity constraint (intended to be deployed in real-time solutions on edge devices for example).

In this an ASC system based on the Gammatone representation of the audio and a slim neural network using squeeze-excitation techniques is presented to improve its performance, which is then converted to TFLite format to reduce its size.

5. ACKNOWLEDGEMENTS

The participation of Dr. Cobos and Dr. Ferri is supported by ERDF and the Spanish Ministry of Science, Innovation and Universities under Grant RTI2018-097045-B-C21, as well as grants AICO/2020/154 and AEST/2020/012 from Generalitat Valenciana.

6. REFERENCES

[1] A. Mesaros, T. Heittola, and T. Virtanen, "Acoustic scene classification: an overview of dcase 2017 challenge entries," in *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*. IEEE, 2018, pp. 411–415.

[2] ——, "Acoustic scene classification in dcase 2019 challenge: Closed and open set classification and data mismatch setups," 2019.

[3] M. Valenti, A. Diment, G. Parascandolo, S. Squartini, and T. Virtanen, "Dcase 2016 acoustic scene classification using convolutional neural networks," in *Proc. Workshop Detection Classif. Acoust. Scenes Events*, 2016, pp. 95–99.

[4] D. Barchiesi, D. Giannoulis, D. Stowell, and M. D. Plumbley, "Acoustic scene classification: Classifying environments from the sounds they produce," *IEEE Signal Processing Magazine*, vol. 32, no. 3, pp. 16–34, 2015.

[5] I. Martín-Morató, T. Heittola, A. Mesaros, and T. Virtanen, "Low-complexity acoustic scene classification for multi-device audio: analysis of dcase 2021 challenge systems," 2021.

[6] S. Perez-Castanos, J. Naranjo-Alcazar, P. Zuccarello, M. Cobos, and F. J. Ferri, "Cnn depth analysis with different channel inputs for acoustic scene classification," *arXiv preprint arXiv:1906.04591*, 2019.

[7] J. Naranjo-Alcazar, S. Perez-Castanos, P. Zuccarello, and M. Cobos, "Task 1 dcase 2020: Asc with mismatch devices and reduced size model using residual squeeze-excitation cnns," DCASE2020 Challenge, Tech. Rep, Tech. Rep., 2020.

[8] M. Slaney, "Auditory toolbox," *Interval Research Corporation, Tech. Rep*, vol. 10, no. 1998, 1998.

[9] J. Naranjo-Alcazar, S. Perez-Castanos, P. Zuccarello, and M. Cobos, "Acoustic scene classification with squeeze-excitation residual networks," *IEEE Access*, vol. 8, pp. 112 287–112 296, 2020.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[11] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun 2018. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2018.00745

[12] A. G. Roy, N. Navab, and C. Wachinger, "Concurrent spatial and channel 'squeeze & excitation'in fully convolutional networks," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018, pp. 421–429.

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

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

[15] 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. [Online]. Available: https://arxiv.org/abs/1807.09840