ACOUSTIC SCENE CLASSIFICATION USING LONG-TERM AND FINE-SCALE AUDIO REPRESENTATIONS

Technical Report

Paul Nguyen Hong Duc[1*], *Dorian Cazau*[2], *Olivier Adam*[1],
Odile Gerard[3], *Paul R. White*[4]

[1] Institut d'Alembert, Sorbonne Universite, Paris, France
p.nguyenhongduc@gmail.com, olivier.adam@sorbonne-universite.fr
[2] ENSTA Bretagne, Lab-STICC, Brest, France,
dorian.cazau@ensta-bretagne.org
[3] DGA-TN, Toulon, France, odile.gerard@intradef.gouv.fr
[4] University of Southampton, ISVR, Southampton, UK, P.R.White@soton.ac.uk

ABSTRACT

Audio scene classification (ASC) is an emerging filed of research in different scientific communities such as urban soundscape characterization or bioacoustics. It has gained visibility and relevance with open challenges especially with the benchmark dataset and evaluation from DCASE. This paper present our deep learning model to address the ASC task of the DCASE 2020 challenge edition. The model exploits multiple long-term and fine-scale audio representations as inputs of the neural network. Each representation is fed into a different network. The audio embedding of each branch are fused before a Multi-Layer Perceptron to predict the final classes.

***Index Terms*—** ASC, task1b, RMS level, third octave levels, sonic atmosphere features, ensemble method

1. INTRODUCTION

Sound plays a key role in our perception of urban environments. Acoustic scenes classification (ASC) can be essential when visual information is not or partially available. ASC aims at classifying acoustic scenes into predefined classes. For the DCASE 2020 challenge edition, 2 subtasks were proposed to the participants. This report focuses on subtask B which is a classification of 3 acoustic scenes acquired in 12 European cities with the same recording device.

In the DCASE 2020 challenge task 1b, a new taxonomy is introduced. The goal of the challenge is to classify acoustic scenes into three classes: indoor, outdoor and transportation. Moreover, it is required that the neural network size should not exceed 500 KB.

A special focus is put on the audio embedding to meet the model size requirements of the task 1b. By reviewing last DCASE challenge edition, most used audio representations in ASC are spectrogram-like ones and sometimes raw audio waveform. The three classes of the task allow more flexibility on the choice of audio representation. In this report, we use long-term representations combined with sonic atmosphere features and log-mel spectrograms. The following section will cover a detailed explanation of the feature extraction process, system architecture, results, and conclusion.

2. FEATURE EXTRACTION

The AARAE Matlab toolbox was used [1] to compute the interaural cross correlation coefficient of the filtered spectrum and the Leq sound level defined in [1] were using the stereo recordings. Furthermore, audio segments were converted to mono using the librosa Python package [2]. Eight timbral characteristics (hardness, depth, brightness, roughness, warmth, sharpness, boominess, reverb [3]) were computed for each second and then averaged over the whole audio recording. All these features enabled to describe the sonic atmosphere of the acoustic scene with only a few number of features. They are named sonic atmosphere features in the following. Log-mel spectrograms were also extracted with 64 bands. The analysis frame was set to about 85 ms (50% hop size). This enabled to have a low temporal resolution representation of the audio signals. This will help to describe the soundscape at finer temporal and frequency scales than other computed features.

Finally, two long-term representations were used. The Power Spectral Density (PSD) was determined by the Welch method [4] with 1024-point Hamming window, 50% overlap, based on 1s temporal signal segments. As a consequence, the time resolution is 1s and frequency resolution is 46.8 Hz. The root-mean square (RMS) level was then computed. This feature will help to have an overview of the dominant frequencies in the acoustic scene. Furthermore, third octave band levels (TOL) were also evaluated on each second of the 10s-long audio clips. Other 1/n octave bands were tried but 1/3 ones give better results in our experiments. The workflow used to compute RMS level and TOL follows that of [5].

Even if most of the proposed features are extracted from the audio spectrum and the information contained in such representations may be redundant, the objective is to help the model in order to reduce its complexity.

Four different inputs are fed into neural networks. There are matrices of size 10×34 and 512×1 for the TOL and RMS level inputs respectively, and an array of length 10 (for the sonic atmosphere features) are fed into a dense neural network. The log-mel spectrograms are stored in a 64×265 matrix.

[*] PhD is granted by the the French Defence Procurement Agency (Direction Générale des Armements).

In order to increase the number of training samples and to make the model more robust to new data, mixup data augmentation technique is used [6].

3. SYSTEM ARCHITECTURE

Three different models (cf Fig. 1) are trained. Their predictions are averaged to make the final decision (cf Fig. 1D). The averaging ensemble method aims at combining different models to improve predictive performance from any individual model. The variety of features fed into the different models and the different model architectures can improve the ability of the ensemble to generalize to unseen data.

3.1. Model 1 (M1)

Log-mel spectrograms, TOL and sonic atmosphere features are fed into a three branches neural network (cf Fig. 1). Both log-mel spectrograms and TOL are inputs of a 2D convolutional layers. The embedding of the latter is flattened at the end while a global average pooling is performed on the TOL embedding. The sonic atmosphere features are fed in a multi-layer perceptron (MLP) with only two dense layers. The log-mel spectrogram branch of the neural network is inspired by the baseline with a reduction of the input size of the melspectrograms. TOL branch and the sonic atmosphere embedding aim at helping the model to capture acoustic scene time and frequency variations based on several seconds. This model has 47,911 non-zero parameters.

3.2. Model 2 (M2)

The RMS level, TOL and the sonic atmosphere features are the inputs of this three branch model. 1D convolutional layers with different dilatation rates are applied to the RMS level. This enables the network to learn relations between other frequencies than TOL at a low computational cost. The TOL is modified with three Gated Recurrent Units (GRU) to learn different temporal relations on the audio spectrum. These layers are equivalent to Long-Short Term Memory cells but with less computational complexity. In this model, the total number of non-zero parameters is 29,117.

3.3. Model 3 (M3)

Only log-mel spectrograms and TOL are fed into a fully convolutional and a recurrent network respectively. Both models are characterized by a low complexity. However, in the final ensemble model, it weights about a third of the total number of non-zero parameters. It contains 45,465 non-zero parameters.

3.4. Training parameters shared by all models

All experiments were completed with Keras [7] with a Tensorflow backend [8] on a Google Colab GPU environment [9]. All models were trained for 200 epochs in batches of 32 samples. A reduction of the learning rate for each model is set up if the validation loss did not decrease since 3 epochs. An early stopping was used to stop the training and to avoid overfitting.

Class label	Baseline	Best ensemble (M1+M2+M3)	Best ensemble (M1+M2)
indoor	82.0	**86.4**	86.2
outdoor	88.5	**96.1**	95.9
transportation	91.3	**94.7**	**94.7**
Average Acc.	87.3	**92.4**	92.3
Model size	450 KB	478.5 KB	**300.9 KB**

Table 1: Results on the development dataset for our two systems compared to baseline. Characters in bold are the best accuracy (acc.) for each row or the smaller model size.

4. RESULTS

4.1. Dataset

The dataset for task 1b is the TAU Urban Acoustic Scenes 2020 3Class. This subtask addresses acoustic scene classification problem. An audio recording is classified into three different: indoor, outdoor and transportation. These classes represent the place where the recording took place. The dataset consists of 10-seconds stereo audio clips (sampling rate of 48 kHz) from 10 acoustic scenes. In total, 40 hours of audio recording was available as the development dataset.

4.2. Results

For both proposed systems, the baseline is outperformed (cf Table 4.2). The contribution of M3 is limited. Adding this model to the ensemble only improves by 0.1 % the macro-average accuracy while its number of non-zero parameters is about more than 1.5 times higher than M3 ones.

The reason of why some acoustic scene are misclassified was investigated. For example, the misclassification of indoor scenes occurs when the acoustic scene is either really quiet or when there is a specific noise such as the clatter of metro doors.

5. CONCLUSION

In this paper, 2 ensemble models are tried to improve the accuracy of the acoustic scene classification. Long-term but also fine-scale audio representations were combined as inputs to the neural networks. Averaging was considered for the ensemble. The results showed an increase in the classification accuracy as compared to the baseline for both proposed low complexity systems.

6. REFERENCES

[1] D. Cabrera, D. Jimenez, and W. Martens, "Audio and acoustical response analysis environment (aarae): A tool to support education and research in acoustics," *INTERNOISE 2014 - 43rd International Congress on Noise Control Engineering: Improving the World Through Noise Control*, 01 2014.

[2] B. McFee, V. Lostanlen, M. McVicar, A. Metsai, S. Balke, C. Thomé, C. Raffel, A. Malek, D. Lee, F. Zalkow, K. Lee, O. Nieto, J. Mason, D. Ellis, R. Yamamoto, S. Seyfarth, E. Battenberg, , R. Bittner, K. Choi, J. Moore, Z. Wei, S. Hidaka, nullmightybofo, P. Friesch, F.-R. Stöter, D. Hereñú, T. Kim, M. Vollrath, and A. Weiss,

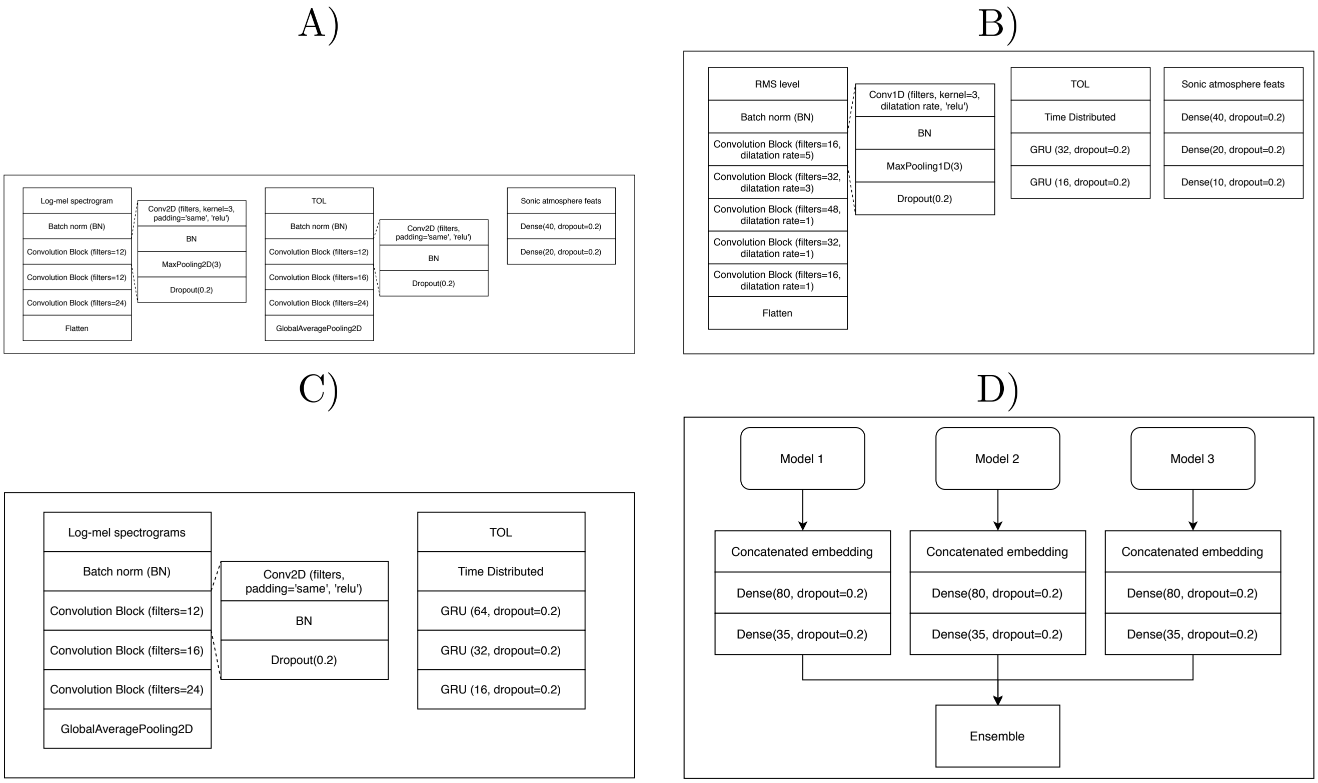

Figure 1: Model graphs. A) Model 1 (M1), B) Model 2 (M2), C) Model 3 (M3) and D) Average ensemble

"librosa/librosa: 0.7.2," Jan. 2020. [Online]. Available: https://doi.org/10.5281/zenodo.3606573

[3] A. Pearce, S. Safavi, T. Brookes, R. Mason, W. Wang, and M. Plumbley, "Deliverable d 5.8: Release of timbral characterisation tools for semantically annotating non-musical content," 2019. [Online]. Available: https://www.audiocommons.org/

[4] P. D. Welch, "The use of fast fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms," *Audio and Electroacoustics, IEEE Transactions on*, vol. 15, pp. 70 – 73, 07 1967.

[5] P. Nguyen Hong Duc, A. Degurse, J. Allemandou, O. Adam, P. R. White, O. Gerard, R. Fablet, and D. Cazau, "A scalable hadoop/spark framework for general-purpose analysis of high volume passive acoustic data," 2019.

[6] H. Zhang, M. Cissé, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," *CoRR*, vol. abs/1710.09412, 2017. [Online]. Available: http://arxiv.org/abs/1710.09412

[7] F. Chollet *et al.*, "Keras," https://github.com/fchollet/keras, 2015.

[8] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "Tensorflow: Large-scale machine learning on heterogeneous systems," 2015, software available from tensorflow.org. [Online]. Available: https://www.tensorflow.org/

[9] ttps://research.google.com/colaboratory/faq.html.