AUDIO SCENE CLASSIFICATION USING ENHANCED CONVOLUTIONAL NEURAL NETWORKS FOR DCASE 2021 CHALLENGE

Technical Report

Itxasne Díez Gaspón[*]

Noismart
C/ Ogoño nº1 5º Piso
Edificio Elkartegi
48930 Getxo, Bizkaia
itxasne@noismart.com

Ibon Saratxaga

HiTZ Center – Aholab
University of the Basque Country UPV/EHU
1 Torres Quevedo Sq.
48013 Bilbao, Spain
ibon.saratxaga@ehu.eus

ABSTRACT

This technical report describes our system proposed for Task 1B – Audio-Visual Scene Classification of the DCASE 2021 Challenge. Our system focuses in the audio signal based classification. The system has an architecture based on the combination of Convolutional Neural Networks and OpenL3 embeddings. The CNN consist of three stacked 2D convolutional layers to process the log-Mel spectrogram parameters obtained from the input signals. Additionally OpenL3 embeddings of the input signals are also calculated and merged with the output of the CNN stack. The resulting vector is fed to a classification block consisting of three fully connected layers.

Mixup augmentation technique is applied to the training data and binaural data is also used as input to provide additional information.

In this report, we describe the proposed systems in detail and compare them to the baseline approach using the provided development datasets.

Index Terms— *Audio Scene Classification*, CNN, DNN, OpenL3

1. INTRODUCTION

An acoustic scene is the sound environment resulting from the combination of sounds generated by different sources existing in that environment. Classification of an acoustic scene is a very challenging task that requires extracting information from audio to categorize it into a predefined scene. In recent years, researchers have developed automatic recognition systems of audio scenes for different applications such as mobile context awareness to react to the environment and multimedia material indexing[1].

For the last years, we have been working on audio event detection and urban sound classification using different techniques as CNNs[3], data augmentation [4] and transfer learning[5]. We have applied some of these methods to the acoustic scene classification problem in the context of the DCASE2021 Challenge, namely, to its Task1B.

The present report is divided in the following sections: in section 2 the task is described, in section 3 we present the proposed system, in section 4 the experiments that have been carried are described and the obtained results are detailed in section 5. The report ends with some conclusions.

2. TASK DESCRIPTION

The Task-1B is dedicated to the classification of sound scenes not only using audio signals but also video data, but single modality systems are also accepted. Our proposal uses only audio data. The task consists on classifying short audio recordings (1 second) into 10 different scenes. Although the provided audio data includes longer recordings (10 seconds), they have to be cut in 1 second segments and provide an independent classification for each of them.

The challenge requires not only the output category but also the classification scores for each of the 10 categories for the calculation of the task evaluation metric.

2.1. Audio datasets

The audios provided by DCASE2021 for task1B, were recorded using a Soundman OKM II klassik/studio A3, electret binaural microphone and a zoom F8 audio recorder at 48kHz sampling rate[6].

The audio scenes were recorded in 12 cities: Amsterdam, Barcelona, Amsterdam, Barcelona, Helsinki, Lisbon, London, Lyon, Madrid, Milan, Prague, Paris, Stockholm and Vienna. Data was recorded in 10 different scenes which

[*] This work has been supported by the Dept. of Economic Development and Infrastructure of the Basque Government (BIKAINTEK)

are the classes used by the classifier: airport, indoor shopping mall, metro station, pedestrian street, public square, street with medium level of traffic, urban park, on board of a tram, bus and underground.

Two datasets have been provided by the organization: development and evaluation.

Development dataset

The development dataset includes both the audio and the labels corresponding to the classification categories. It is divided into two subsets: training (8647 audios) and test (3646 audios). All the files in this dataset are 10s length.

Additionally, the training subset is split into training and validation data. These two splits do not share recordings of the same location.

The test subset is used to report the performance of the systems with the development data.

Evaluation dataset

Evaluation dataset contains 72,000 files 1 second length, recorded in two cities unseen in the development set. No labels are provided for this dataset, as it is used for the challenge submission.

3. PROPOSED ARCHITECTURE

The system architecture that we propose is based on the combination of Convolutional Neural Networks and OpenL3 embeddings for audio processing. This architecture is described in the following sections.

3.1. System architecture

The proposed system, Figure 1, consists of two branches, the first one processes log-Mel spectrograms and the second one uses OpenL3 embeddings.

The CNN branch consists of 3 convolutional blocks (ConvBlock). Each convolutional block, Figure 2, consists of one 2D convolutional layer with a kernel size of (3x3), stride of (2, 2), and padding same. After each convolutional layer, we use batch normalization, ReLU activation and max pooling with a pool size of (3x2) and stride (1, 2). Each of the blocks has a decreasing number of convolutional filters: 256, 128 and 64.

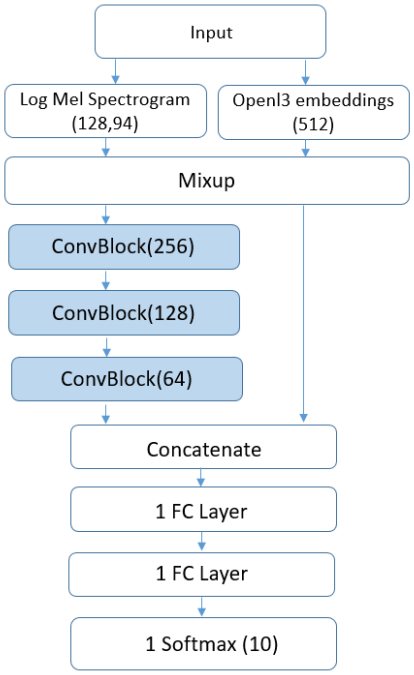

Figure 1. System architecture

The output of the CNN blocks is concatenated with OpenL3 embeddings. The resulting combined parameter vector is used as input for a block of dense layers. This block consists of two fully connected layers with 50% dropout. We propose 3 alternative models with different number of neurons, as described in subsection 3.4. Finally, a softmax layer is used for classification.

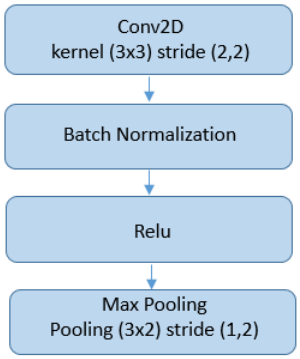

Figure 2. Convolutional Block

3.2. Audio Preprocessing

The length of the audios are 10 second for the development data but, as the classification decision is taken for 1 second length samples, we split the audio into 1 second segments. In order to generate more training data and to avoid border-effects we use a 0.5 second of hop size for the 1 second segmentation window.

Log-Mel spectrograms

The audio data provided is stereo, and thus we follow two strategies for calculating the parameters for the neural network. The first one is to mix both channels using the mean value and the second one, is to duplicate the input vector, obtaining the parameters of each channel separately and then concatenate them.

Log Mel spectrograms are calculated using Librosa library at the original sample rate 48 kHz. For each audio segment, we calculate a 128 Mel bands log-Mel spectrogram every 21ms with an overlap of 10ms, obtaining 94 vetors of 128 frequency values. Log-Mel spectrograms are used as input to the convolutional layers.

Audio embeddings

For extracting the embeddings we mix both channels of the stereo recordings using the mean value. We use OpenL3 pre-trained model [6] to extract audio embedding using the model trained on the environmental audios. The embeddings are calculated using the internal 128 Mel representation. The resulting embedding vector has 512 values.

The audio embeddings are concatenated with the output features of the CNN, before the first fully connected layer.

Both log-Mel spectrograms and embeddings are ZScore normalized at recording level.

3.3. Data Augmentation

Mixup is used as augmentation technique. The mixup function that we use it is based on [7]. This technique generates a weighted combination of random pairs of vectors from the training data with their corresponding labels. The training data is combined using a fixed weight factor of 0.4, and the one hot encoded labels are also combined in a tailored loss function which computes the weighted loss function of the scores related to the two labels of the original vector.

The new samples obtained by mixup are added to the original dataset doubling the number of samples for training. Mixup is applied to both log-Mel spectrogram and embeddings parameters. Mixup is neither applied to validation nor test data.

3.4. Proposed systems

We present three slightly different systems for task1B. All systems use log-Mel spectrograms and embeddings with mixup training data. The differences between the systems are explained below:

- System 1: proposed architecture with two fully connected layers of 350 neurons. Mean value of both channels of the stereo audios as input for the network.
- System 2: proposed architecture with two fully connected layers with 256 and 128 neurons respectively. Mean value of both channels of the stereo audios as input for the network.
- System 3: proposed architecture with two fully connected layers of 350 neurons. Log-Mel spectrogram is calculated for each separate stereo channel. Embeddings are calculated from the mean value of both channels.

4. EXPERIMENTS

During the development of the systems, we have split the training subset of the development dataset in two groups: training and validation using 6230 recordings for training and 2417 for validation. Then, we have evaluated the performance using the test subset of 3646 recordings of the development dataset.

To train the final models for submission, we have used all the audios of the development dataset (both training and test subsets) and divided them into a training and a validation split (aprox. 70%-30%), avoiding shared locations between both splits. This division result in 8826 audios for training and 3465 for validation. Finally, we calculate the scores for submission with the 72000 audio files provided by DCASE2021 task1B in the evaluation dataset.

The models have been trained using Adam optimizer with a learning rate of 0.0001. L2 regularization with a factor of 1e-5 it is also used. The training was carried out using early stopping with a patience value of 20, using validation loss as stop and a limit of 200 epochs.

5. RESULTS

The challenge will rank the systems using macro-average multiclass cross-entropy (Log loss) as main criterion, and macro-average accuracy (average of the class-wise accuracy) as an additional metric.

Log loss metric is the cross entropy calculated with the ground truth and the predicted scores. Accuracy is the ratio between correct classification decisions and the total number of decisions.

The ”overall” metrics are calculated considering all the samples independently of their categories and the “mean”

metrics are calculated as averages of the corresponding metric for each category.

Table 1, shows the global results obtained for each of the evaluated systems. Table 2 and Table 3 show log loss and accuracy by category.

Although all systems outperform the baseline using the log loss metric, accuracy values are not as good in all three systems.

Table 1. Overall and Average log loss and accuracy metrics for the baseline and our systems.

	Mean Log loss	Mean Acc	Overall Log loss	Overall Acc
Baseline	1.048	0.651	1.057	0.650
System 1	1.038	0.656	1.057	0.651
System 2	1.006	0.633	1.016	0.630
System 3	1.023	0.632	1.032	0.629

Table 2.Class-wise log loss metrics for the baseline and our systems

Class	Base	Sys1	Sys2	Sys3
Airport	0.977	0.802	0.936	0.796
Bus	0.628	0.697	0.941	0.969
Metro	1.106	0.993	0.999	0.996
Metro station	1.316	1.273	1.370	1.321
Park	0.960	0.890	0.780	1.027
Public square	1.284	1.089	1.125	1.123
Shopping mall	1.384	1.784	1.067	1.408
Street pedestrian	1.285	1.260	1.396	1.214
Street traffic	0.516	1.260	1.396	1.214
Tram	1.026	0.808	0.798	0.915

6. CONCLUSIONS

In this paper, we have described our systems participating in the Task1B of the DCASE2021 Challenge. Our acoustic scene classification system is based on CNNs and OpenL3 embeddings with mixup augmentation. The results obtained with the development data show that one of them outperforms the baseline system both in the macro-averaged multiclass cross entropy and in the accuracy.

Table 3.Class-wise accuracy metrics for the baseline and our systems

Class	Base	Sys1	Sys 2	Sys3
Airport	0.669	0.732	0.663	0.744
Bus	0.780	0.747	0.640	0.642
Metro	0.607	0.609	0.569	0.582
Metro station	0.580	0.591	0.527	0.519
Park	0.735	0.804	0.798	0.788
Public square	0.543	0.620	0.576	0.516
Shopping mall	0.549	0.426	0.617	0.467
Street pedestrian	0.574	0.559	0.449	0.544
Street traffic	0.847	0.767	0.781	0.862
Tram	0.629	0.706	0.709	0.657

7. REFERENCES

[1] T. Virtanen, M. D. Plumbley, and D. Ellis, *Computational analysis of sound scenes and events*. 2017.

[2] S. Wagenpfeil, F. Engel, P. M. Kevitt, and M. Hemmje, "AI-based semantic multimedia indexing and retrieval for social media on smartphones," *Inf.*, vol. 12, no. 1, pp. 1–30, 2021.

[3] A. Arnault and N. Riche, "CRNNs for Urban Sound Tagging with spatiotemporal context," pp. 2–5, 2020.

[4] J. Bai, C. Chen, M. Wang, J. Chen, X. Zhang, and Q. Yan, "Data Augmentation Based System For Urban Sound Tagging," *Detect. Classif. Acoust. Scenes Events 2020*, pp. 3–5, 2020.

[5] J. Cramer, H. H. Wu, J. Salamon, and J. P. Bello, "Look, Listen, and Learn More: Design Choices for Deep Audio Embeddings," *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, vol. 2019–May, pp. 3852–3856, 2019.

[6] S. Wang, A. Mesaros, T. Heittola, and T. Virtanen, "A Curated Dataset of Urban Scenes for Audio-Visual Scene Analysis," pp. 626–630, 2021.

[7] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "MixUp: Beyond empirical risk minimization," *6th Int. Conf. Learn. Represent. ICLR 2018 - Conf. Track Proc.*, pp. 1–13, 2018.