DCASE 2017 - TASK 1 : HUMAN-BASED GREEDY SEARCH OF CNN ARCHITECTURE

A. Rakotomamonjy

LITIS EA 4108
Université de Rouen Normandie
Avenue de l'université
76800 Saint Etienne du Rouvray, France

ABSTRACT

This paper presents the methodology we have followed for our submission at the DCASE 2017 competition on acoustic scene classification (Task 1). The approach is based convolutional neural networks. There is nothing original about this contribution, as we have just applied a human-based search of the best CNN architecture and hyper-parameters using a 4-fold cross-validation for selecting the best model. We hope that this approach will not reach the top entry of the challenge and that it will be outperformed by clever and beautiful methods.

Index Terms— CNN, CQT, acoustic scene, ugly brute force.

1. INTRODUCTION

Audio scene classification is a complex problem which aims at identifying acoustic environments solely based on audio recordings of the scene. The scenes we are interested in can be defined according to some geographical contexts (beach, park, etc...), some social situations in indoor or outdoor locations (restaurant, office, home, market, library, ..) or according to some ground transportations (car, bus, tramway, ...). Being able to accurately recognize such scenes is relevant for applications in which context awareness is of primary importance.

In the last decade, advances in the state-of-the-art in this domain were few but a steady increase of studies occured in the last years. Novel approaches for addressing this problem of acoustic scene classification have flourished [1, 2, 3] and they have been essentially fueled by the release of open and established datasets for benchmarking. These datasets include the one used for the challenge DCASE 2013 [4], the LITIS Rouen Audio scene dataset [5] and the DCASE 2016 Challenge dataset [6]. For the Task 1 of the DCASE 2017 Challenge [7], the dataset from 2016 has been enriched with new recordings which has been used as evaluation set.

For this novel challenge, we have used a dumb approach which consists in brute forcing the search for the best CNN architectures and hyperparameter selections. We have tried a lot of them and retained the best performing one for our submissions.

This work is partially supported by the Région Normandie through the GRR Project DAISI and by European Funding through the program FEDER-FSE/IEJ Haute Normandie.

2. METHOD

2.1. The dataset

The data we have to deal with are composed of 10s audio scenes acquired in different places. Our objective is to learn from some labeled examples of audio scene the place where they have been acquired. In the dataset available for developing the methodology, 312 segments of $10s$ are available per location (the class to retrieve). In addition, some specific folds defining 4 sets of training and validation are provided. Details about the dataset can be found in [7]. All the results presented in here are obtained as an average accuracy over the 4 folds.

2.2. Machine learning Pipeline

The approach we have developed addresses the problem as a machine learning problem where each of the labeled acoustic scene is considered as a single example. Hence, as in many machine learning tasks, the most difficult problem is to design some features that are able to grasp specificities of each acoustic scene class while preserving discriminative power. In order to cope with this problem, we apply in this work a convolutional neural network approach that learn features from time-frequency representations.

2.2.1. Time-Frequency representation of acoustic scenes

The first transformations we apply to each acoustic scene signal are the following

- the stereo signal is averaged over the two channels
- we compute a CQT transform of hop length of 1024, on 12 octaves, a total number of bins of 288 and with minimal frequency of 5 Hz.
- the module amplitude of this transform is then log compressed through the mapping $\log(x - \min_i(x_i) + \epsilon)$, with $\epsilon = 0.01$

At this point, each acoustic scene can be represented as a matrix of size 288×431.

2.2.2. Evaluated models

We have evaluated several different data normalization and CNN architectures. Notably, for the data normalization, we were interested in whether we should

1. normalize the signal to unit energy before CQT transform
2. normalize each component of the log-compressed CQT transform between 0 and 1

Acoustic Scene	Baseline	Our model
beach	75.3	87.5
bus	71.8	96.8
cafe/restaurant	57.7	78.2
car	97.1	96.8
city center	90.7	95.8
forest path	79.5	92.9
grocery store	58.7	92.6
home	68.6	89.3
library	57.1	73.7
metro station	91.7	97.4
office	99.7	92.9
park	70.2	63.5
residential area	64.1	59.9
train	58.0	86.2
tram	81.7	84.6
Overall	74.8	85.9

Table 1: Class accuracy for all the acoustic scenes. Comparison between the baseline and our model.

3. normalize the traning set to 0 mean and unit standard deviation and the validation and test sets accordingly.

Regarding architectures, we have tried different CNN feature learning architectures by combining a base architecture. The classifier layer (which have been kept fixed) is composed by a fully-connected layer with 200 units followed by dropout and a non-linearity. The output layer is composed of 15 neurons related to the class probabilities of each class. The base architecture is a convolutional layer, followed by a non-linearity and pooling. Within this broad framework, we have evaluated : different kernel size, different pooling size, different non-linearity, different number of layers.

Each model has been trained with Keras [8]. We have used Adadelta with default parameters as optimization algorithm and the categorical cross entropy loss as loss function. We have set the maximum number of epochs to 100 and the batch size to 5. For each fold, we retain the model that has performed the best on the validation set during the full training process.

On the overall, we have evaluated more than 70 architectures for each of the preprocessing method. The best model we achieved is given in Table 2. This model performs best when the signal is not normalized to unit energy but data scaling over each example and normalization over the sets are applied. Interestingly, we can note that the model is a multi-resolution model, denoted as Res1, Res2 and Res3 for which learned features are concatenated before being fed to the fully-connected layers. The receptive fields of these base models increase in frequency denoting the importance of having filters in which frequency bands are highly-mixed through convolution and other filters for which frequency bands need to be genuine.

The performance of this model is reported in Table 1 jointly to the one of the baseline methode reported by Mesaros et al. [7]. We can note that for 4 out of 15 acoustic scenes our model always performs better than the baseline. However it seems to struggle in correctly classifying quiet scenes such as office, park and residential area.

Table 2: Our best ConvNets architecture.

Type	Res 1	Res 2	Res 3
Input	288×431	288×431	288×431
Convolution	20 - (2×15)	20 - (7×15)	20 - (16×15)
ReLu Unit			
Max Pooling	(2×5)	(2×5)	(2×5)
Convolution	40 - (1×15)	40 - (1×15)	40 - (1×15)
ReLu Unit			
Max Pooling	(1×8)	(1×8)	(1×8)
Merge	Flatten and Concatenate		
FullyConnected	34440×200		
Dropout	p = 0.8		
FullyConnected	200×15		

2.3. Submission

For predicting on the test set, we have proceeded according to the following steps. We predict the class of each example of the test set according to the models trained on each of the fold. Afterwards, we average the class prediction score over the fold and selects the class with higher probability. When averaging over several models, we proceed exactly in the same way and consider an averaging over $4 \times N$ scores, where N is the number of models. In final, we have submitted :

1. a result using a single model
2. a result using the average of the 4 top-performing models
3. a result using the average of the 19 top-performing models

3. CONCLUSION

This paper describes our submission at DCASE 2017. It is based on a brute-force human-based greedy search of the best architectures. While results on the development set seem to be good, it would be sad if this approach reaches top entry in the challenge.

4. REFERENCES

[1] V. Bisot, R. Serizel, S. Essid, and G. Richard, "Feature learning with matrix factorization applied to acoustic scene classification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 6, pp. 1216–1229, 2017.

[2] M. Valenti, S. Squartini, A. Diment, G. Parascandolo, and T. Virtanen, "A convolutional neural network approach for acoustic scene classification," in *Neural Networks (IJCNN), 2017 International Joint Conference on*. IEEE, 2017, pp. 1547–1554.

[3] W. Yang and S. Krishnan, "Combining temporal features by local binary pattern for acoustic scene classification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 6, pp. 1315–1321, 2017.

[4] D. Giannoulis, E. Benetos, D. Stowell, M. Rossignol, and M. Lagrange, "Detection and classification of acoustic scenes and events: an ieee aasp challenge," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2013.

[5] A. Rakotomamonjy and G. Gasso, "Histogram of gradients of time-frequency representations for audio scene classification,"

Audio, Speech, and Language Processing, IEEE/ACM Transactions on, vol. 23, no. 1, pp. 142–153, Jan 2015.

[6] A. Mesaros, T. Heittola, and T. Virtanen, "TUT database for acoustic scene classification and sound event detection," in *24th European Signal Processing Conference 2016 (EUSIPCO 2016)*, Budapest, Hungary, 2016.

[7] A. Mesaros, T. Heittola, A. Diment, B. Elizalde, A. Shah, E. Vincent, B. Raj, and T. Virtanen, "DCASE 2017 challenge setup: Tasks, datasets and baseline system," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017)*, November 2017, submitted.

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