DEEP NEURAL NETWORKS WITH SUPPORTED CLUSTERS PRECLASSIFICATION PROCEDURE FOR ACOUSTIC SCENE RECOGNITION

Technical Report

Marcin Plata

Samsung R&D Institute Poland
Data Intelligence Group
Warsaw, Poland

ABSTRACT

In this technical report, we presented a system for acoustic scene classification focuses on deeper analysis of data. We made an impact analysis of various combinations of arguments for short time Fourier transform (STFT) and Mel filter bank. We also used the harmonic and percussive source separation (HPSS) algorithm as an additional features extractor. Finally, next to common spectrograms divided and non-overlap classification neural networks, we decided to present an out-of-the-box solution with one main neural network trained on clustered labels and a few supporting neural network to distinguish between most difficult scenes, e.g. *street_pedestrian* and *public_square*.

Index Terms— CNN, STFT, clustering, HPSS

1. INTRODUCTION

A popularity of an acoustic scene classification (ASC) has been increasing over the years. And many possible scenarios of practical applications of ASC cause strong interest from high-tech companies. One of crucial event about ASC is DCASE Challenge, where first task concerns on distinguish between ten acoustic scenes. This years edition is fifth and there were a dozen various approaches presented during recent editions of the challenge. In the 2017, Mun et al. [1] applied generative neural networks (GANs) for data augmentation and they achieved the highest accuracy equal to 83.3%. Their result was 2.9 percentage points higher that a second solutions provided by Han et al. [2]. In [2], authors presented an architecture combined from four neural networks trained on left/right channels, mid/side channels, harmonic/percussive, and background subtraction with median filter. In 2018, Sakashita and Aono [3] shown a similar idea to [2], but they fed neural networks with full spectrograms, overlap divided and non-overlap divided segments. Their solution achieved the highest accuracy in the first task of the DCASE 2018 challenge. The author of the second solution decided to use convolutional neural networks, i-vectors and fusions both of them [4]. Most of presented neural network models used for acoustic scene classification are based on a VGG architecture [5].

2. ARCHITECTURES

This section introduces to the proposed system. It describes all details about preprocessing, training prediction stepa.

Table 1: Examined arguments in a preprocessing step. *n_mels* is an argument for the Mel filter bank and refers to a number of Mel bands to generate. *n_ftt*, *hop_length*, *sr* are arguments for the STFT and refer to window size, number of frames between STFT columns, sample rate, respectively. Some combinations of features are distinguished in the table for later purposes.

Num.	*n_mels*	*n_fft*	*hop_length*	*sr*
0	40	2048	1024	48000
1	64	2048	512/1024	32000
2	128	2048	1024	44100
3	64	1024/4096	512/1024	48000
4	128	2048	512	48000

2.1. Audio preprocessing

We tried to find an optimal arguments for the STFT and the Mel filter bank transformations, which are informative and robust for overfitting. Examined combinations of most influential arguments for the STFS and the mel filter bank were presented in the Table 1.

We ran several neural network models fed by particular inputs. We observed that models which taken melspectrograms with *n_ftt* equal to 4096 had tendencies to overfitting. In turn, models fed by melspectrograms with *n_ftt* equal to 1024 achieved worse results in contrast to others. Our analysis shown that the most optimal combinations of the arguments are presented in the Table 1 with the Number 3, with *n_mels* equal to 128, *n_fft* equal to 2048, *hop_length* equal to 512. The melspectrograms was calculated on left channel, right channel, mono, harmonic and percussive.

2.2. Architecture of networks

2.2.1. Preclustered system

Our aim was improving an efficiency of distinguishing between two often misclassified scenes, e.g. *street_pedestrian* vs. *public_square*. Thus, we decided to train supported CNNs, whose task is to distinguish between only two particular classes. We also reduce a task complexity of a main neural network, which needs to choose a cluster, instead to make a prediction over 10 classes. The architecture of neural networks used in the system was described in the Table 2 and was presented by [6].

We proposed a novel type of scene prediction procedure, when based on main CNN output we moved to one of supported CNNs to make a final decision. A sketch of the prediction procedure was

Table 2: Architecture of our first convolutional neural network for the acoustic scene classification.

Layer Type	**Parameters**
**ConvBlock*	$c = 64$
**ConvBlock*	$c = 128$
**ConvBlock*	$c = 256$
**ConvBlock*	$c = 512$
Mean	mean over frequency
Max	maximum over time
Linear	10 output features
LogSoftmax	-
*$ConvBlock(c)$ structure:	
Conv2D	c filters, 3×3 kernel
BatchNorm	-
ReLU	-
Conv2D	c filters, 3×3 kernel
BatchNorm	-
ReLU	-
AvgPooling	2×2 kernel

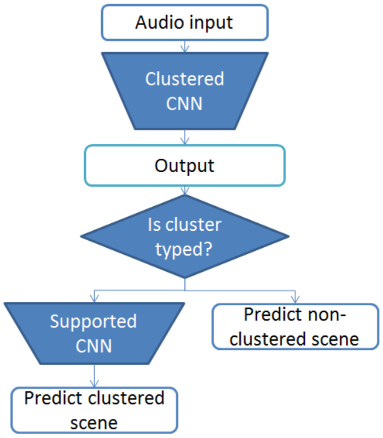

Figure 1: Example of a figure with experimental results.

presented in Figure 1. We investigated multiple variants of clusters, but finally we decided to make two clusters. One cluster joined *street_pedestrian* and *public_square* classes, second cluster was done by joining *metro* and *tram* classes. The whole system requires three times more parameters in comparison to the standard classifier, because we applied the same architecture for preclustering network and two networks for distinguishing between joined classes.

We compared the accuracy, which we achieved with the preclustered system with a standard 10-classes classifier trained on the same neural network architecture. Our approach slightly improves an accuracy of the acoustic scenes classification problem. The comparison of the results between the preclustered system and the standard 10-classes classifier was presented in Table 3.

2.2.2. Divided spectrogram architecture

Alongside, we also trained neural networks using samples of divided spectrograms. The procedure is well known and was described in [3], among others. The models of neural networks used in experiments was presented in Table 2 and Table 4. The second architecture was presented in [3].

Table 3: Results of classification accuracy on the validation dataset for a standard 10-classes neural network and our preclustered system. The value in the brackets refers to the accuracy reported by [6].

Input type	10-classes	Preclustered
harmonic/percussive	70.8%	**73.5%**
harmonic/percussive/mono	69.8%	**73.1%**
left/right	69.2% (70.3%)	**71.8%**

Table 4: Architecture of our second convolutional neural network for the acoustic scene classification.

Layer Type	**Parameters**
**ConvBlock*	c = 32
MaxPool	3×3 kernel
**ConvBlock*	c = 64
MaxPool	3×3 kernel
**ConvBlock*	c = 128
MaxPool	3×3 kernel
**ConvBlock*	c = 256
GlobalAvgPooling	
Linear	1024 output features
ReLU	-
Linear	10 output features
Softmax	-
*$ConvBlock(c)$ structure:	
ZeroPad2d	1×1 padding
GroupNorm	max(1, c/16) group numbers
LeakyReLU	alpha = 0.1
Conv2d	c filters, 3×3 kernel
ZeroPad2d	1×1 padding
GroupNorm	max(1, c/16) group numbers
LeakyReLU	alpha = 0.1
Conv2d	c filters, 3×3 kernel

2.3. Training

All neural networks was optimized by SGD with Nesterov momentum algorithm with *learning rate* equal to 0.01, *momentum* equal to 0.9 and *weight decay* equal to 0.0001. We also used *mixup* techniques [7] for the data augmentation.

2.4. Final system

The final system contains 9 subsystems, more precisely 3 preclustered systems fed by whole spectrograms and 6 standard neural networks fed by samples of divided spectrograms. Every type of network was fed by three combinations of spectrograms - left/right channels, harmonic/percussive and harmonic/percussive/mono, where particular types of spectrograms was treated as layers to the neural networks. We also used different types of preprocessing and neural network architectures. All details

Table 5: List of subsystems combined to the final system. In the table, *Input types* column refers to types of spectrograms. *Preproc.* column refers to the number of set of arguments for the STFT and Mel filter bank in the Table 1. *Model* column refers to the number of the table with the model's description. *Length* column refers to the length of the sample cropped from the full spectrogram.

Num.	Input types	Preproc.	Model	Length
0	harm./perc.	2	4	43
1	harm./perc./mono	2	4	43
2	left/right	2	4	43
3	harmonic/percussive	4	2	67
4	harm./perc./mono	4	2	67
5	left/right	4	2	67
6	harmonic/percussive	4	2	N/A
7	harm./perc./mono	4	2	N/A
8	left/right	4	2	N/A

Table 6: Results of the systems achieved on DCASE 2019 task 1A dataset. **V** refers to the soft voting and **RF** to the random forest.

System type	Validation (%)	Leaderboard (%)
0	70.5	-
1	71.2	-
2	70.9	-
3	73.8	-
4	74.2	-
5	73.7	-
6	73.5	-
7	73.1	-
8	71.8	-
V	81.6	80.8
RF	-	80.6

of particular neural networks was presented in Table 5. Subsystems 6-8 work with the preclustered procedure.

2.4.1. Ensemble methods

The final prediction was done using one of ensemble methods. In our case, we used simple soft voting and random forest with 2000 decision tress.

3. RESULTS

Results which we achieved on the DCASE 2019 task 1A dataset was presented in Table 6. Results was presented for validation and public leaderboard datasets.

4. REFERENCES

[1] S. Mun, S. Park, D. K. Han, and H. Ko, "Generative adversarial network based acoustic scene training set augmentation and selection using svm hyper-plane," DCASE Challenge 2017, Tech. Rep., 2017.

[2] Y. Han, H. Park, and K. Lee, "Convolutional neural networks with binaural representations and background subtraction for acoustic scene classification," DCASE Challenge 2017, Tech. Rep., 2017.

[3] Y. Sakashita and M. Aono, "Acoustic scene classification by ensemble of spectrograms based on adaptive temporal divisions," DCASE Challenge 2018, Tech. Rep., 2018.

[4] M. Dorfer, B. Lehner, H. Eghbal-zadeh, C. Heindl, F. Paischer, and G. Widmer, "Acoustic scene classification with fully convolutional neural networks and i-vectors," DCASE Challenge 2018, Tech. Rep., 2018.

[5] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations*, 2015.

[6] Q. Kong, Y. Cao, T. Iqbal, Y. Xu, W. Wang, and M. D. Plumbley, "Cross-task learning for audio tagging, sound event detection and spatial localization: Dcase 2019 baseline systems," 2019.

[7] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," 2017.