URBAN ACOUSTIC SCENE CLASSIFICATION USING BINAURAL WAVELET SCATTERING AND RANDOM SUBSPACE DISCRIMINATION METHOD

Technical Report

Fatemeh Arabnezhad

K.N.Toosi University of Technology
Computer Engineering Dept., Tehran, Iran
fateme.a91@gmail.com

Babak Nasersharif

K.N.Toosi University of Technology
Computer Engineering Dept., Tehran, Iran
bnasersharif@kntu.ac.ir

ABSTRACT

This report describe our contribution to Detection and Classification of Urban Acoustic Scenes on DCASE 2019 challenge (Task1 –Subtask A). We propose to use wavelet scatterings spectrum as a good representation and feature where we extracted from both average of 2 audio recorded(mono) and also difference of 2 audio recorded channels (side). The concatenation of these two set of wavelet scattering spectrum are used as a feature vector which is fed into a classifier based on random subspace method. In this work, Regularized Linear Discriminant Analysis (RLDA) is used as a base learner and a classification approach for Random Subspace Method. The experimental results shows that the proposed structure learn acoustic characteristics from audio segments. This structure achieved 87.98% accuracy on whole development set (without cross-validation) and 78.83% on leaderboard dataset.

Index Terms— Acoustic Scene Classification, Wavelet Scatter, Random Subspace, Regularized LDA

1. INTRODUCTION

Environmental sounds carries a lot of information that can helps humans in different fields such as Internet Of Things(IOT) services or intelligent surveillance system. Acoustic scene classification(ASC) is a subfield of Computational Auditory Scene Analysis(CASA), aims to detect surrounding sounds and classify them into predefined classes.

DCASE 2019, ASC challenge consists of 3 subtasks, regular Acoustic Scene Classification, Acoustic Scene Classification with mismatched recording devices and Open set Acoustic Scene Classification. In this work we focus on the first subtask. This task contains 10 classes of sounds recorded with the same device for both development and evaluation dataset.

A system of Convolution Neural Network(CNN) with log melband energies is designed for baseline. It consists of two CNN layer with ReLu activation function and a softmax layer for decision making. The baseline system has an average accuracy of 62.5% on development dataset.

We design a system for Acoustic Scene Classification , subtask A, based on wavelet scattering spectrum and a kind of Random Subspace Method(RSM) to detect and classify acoustic scenes. Our structure significantly improves the baseline accuracy to 87.99% on development dataset. We also achieved 78.83% accuracy on the leaderboard dataset. The following sections describe our method in details.

2. PROPOSED METHOD

In this section we describe our methods for feature extraction and signal representation. Also, we introduce our proposed method in this chapter.

2.1. Wavelet Scattering Spectrum

Feature extraction is the main step of each classification task. In audio classification tasks we are interested in features which are invariant in time and robust to deformation. The former requires that time displacement in raw signal does not change the audio class and the latter means a small deformation in signal cause small variation in audio features. On the other hand, most of the common audio features like Mel Filter bank Cepstral Coefficients(MFCC) or melspectrogram are based on Short Time Fourier Transform. Although STFT is invariant in time but signal deformation will affect high frequency bands.

S. Mallat in [1] proposed a so called scattering wavelet transformation of the signal which satisfies above requirements. In this transformation a wavelet transformation filter bank is applied to signal to extract a multi-resolution representation of original signal. The bank is created by selecting different value for the Constant Q Coefficient named quality factor as in (1). Wavelet transformation samples signal with high resolution respect to time which make the representation sensitive to dilation.

$$\psi_{\lambda_i} = 2^{-jQ}(\psi(2^{-jQ})) \qquad \lambda_i = 2^{-jQ} \tag{1}$$

To solve this issue an averaging filter is applied to wavelet coefficients. Since the mean of wavelet coefficients , named scalogram, is zero, modulus of wavelet coefficients is considered as (2) and first-order time scattering coefficients are extracted.

$$S_1 =| x * \psi_{\lambda_1} | *\phi(t) \tag{2}$$

Averaging filters lose some information. To recover lost information in high frequencies another wavelet decomposition computed on the scalogram along time axis with another quality factor as (3). The above procedure can be repeated to derive higher order time scattering coefficients. Schema of extracting wavelet scattering scpectra has been shown in figure 1.

$$S_2 =| x * \psi_{\lambda_2} | *\phi(t) \tag{3}$$

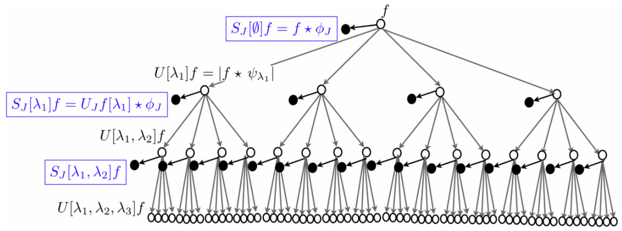

f
$S_J[\emptyset]f = f \star \phi_J$
$U[\lambda_1]f = |f \star \psi_{\lambda_1}|$
$S_J[\lambda_1]f = U_J f[\lambda_1] \star \phi_J$
$U[\lambda_1, \lambda_2]f$
$S_J[\lambda_1, \lambda_2]f$
$U[\lambda_1, \lambda_2, \lambda_3]f$

Figure 1: Scattering wavelet transformation [1]

2.2. Random Subspace Method

Random Subspace Method(RSM) is one of ensemble method at the feature level. Consider a dataset consist of N training samples $X = \{x_1, x_2, ..., x_p\}$ and each sample has p number of features $x_i = \{x_{i,1}, x_{i,2}, ..., x_{i,p}\}$. In this technique, r features ($r < p$) was selected randomly out of all per sample . Then each bundle trained by a classifier separately. Finally results of all learners aggregated by simple majority voting.

The effectiveness of RSM can be seen when dataset has many training samples or small amount of data. If less number of data is available, we can generate more by combining variation of features as much as needed. Against , it may be impossible to train well a classifier with huge number of features. Thus we can apply RSM to subsample the dataset and classify them using multiple learners. Pseudo code of RSM presented in figure 2.

2.3. Regularized Discriminant Analysis

Discriminant Analysis(DA) methods such as Linear Discriminant Analysis(LDA) and Quadratic Discriminant Analysis(QDA) are suited for multi-class classification. LDA uses a linear discriminant function but QDA allows for non-linear separation of data without dimensionality reduction. The discriminant function of DA defined as (4).

$$d(x_i) = (x_i - \mu_k)^T \sum_k{}^{-1} (x_i - \mu_k) + \ln|\sum_k| - 2\ln\pi_k \tag{4}$$

where μ_k and $\sum_k$ are mean and covariance of class k and π_k is the prior probability of class k. Each data sample x_i in training set will be assigned to a class that has lowest value for discriminant function. Based on this rule and the quadratic shape of boundaries, this method is called Quadratic Discriminant Analysis.

Specially when the covariance matrix of classes has been considered identical $\sum_k = \sum$ the above discriminant function change to (5), which is called Linear Discriminant Analysis. As the covariance matrix is equal in classes, quadratic term $x^T \sum^{-1} x$ can be removed for all classes.

```
for $b = 1, ..., B$ do
    select r random feature $\hat{x_b}$ from original p-dimensional feature.
    Train a classifier $C^b(x)$ on $\hat{x_b}$
end for

Aggregate classifier results by majority voting:
    $\beta(x) = \arg\max \sum_y \sigma_{sign(C^b(x),y)}$
```

Figure 2: Pseudo code of RSM

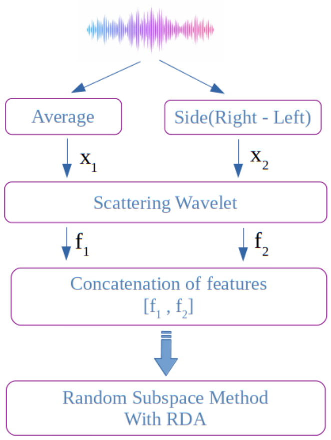


Figure 3: schema of proposed network

$$d(x_i) = (x_i - \mu_k)^T \sum_k (x_i - \mu_k) - 2\ln\pi_k \tag{5}$$

In the cases that the number of training samples is small or when the feature size exceeds the number of observation, the covariance matrix is singular and inverse of it can not be computed. To solve this issue Friedman in [2] add a regularization parameter to covariance matrix as (6).

$$\hat{\sum_k}\alpha = \alpha * \hat{\sum_k}\alpha + (1-\alpha) * \sigma \qquad \alpha \in [0,1] \tag{6}$$

α is a tuning parameter. If we set $\alpha = 0$ it gives QDA and if $\alpha = 1$, RDA is equal to LDA.

In this paper, we use wavelet spectral along side Regularized random subspace to classify acoustic scenes. In following sections we describe feature extraction and classification approaches. Schema of proposed method depicted in figure 3.

3. EXPERIMENTS AND RESULTS

3.1. DCASE 2019 ASC Dataset

DCASE 2019 dataset consist of 10 classes which are airport, shopping mall, metro station, pedestrian street, public square, street traffic, tram, bus, metro and park. This challenge provided two datasets: development and evaluation. TAU Urban Acoustic Scenes 2019, sub-task A, dataset contains 40 hour audio recordings which are balanced between classes and recorded at 48kHz sampling rate with bit 24-bit resolution in stereo. 5-6 minute recordings provided which is spitted into 10 second audio length.

3.2. Audio preprocessing and Feature extraction

As mentioned before we use wavelet scattering spectrum in order to reduce the effect of signal dilation and signal deformation on classification performance. Furthermore we considered that if we extract

Table 1: wavelet scattering results with quality factor [4,1]

Channels	subspace dim	RDA	QDA
Average	50	66.59	69.70
	150	70.49	63.08
	200	69.75	59.19
side	50	63.51	67.53
	150	66.09	62.83
	200	65.42	59.19
[Average, side]	50	69.25	72.33
	350	**74.05**	59.35
	500	72.43	53.31

Table 2: wavelet scattering results with quality factor [9,4]

Channels	subspace dim	RDA	QDA
Average	50	62.58	65.30
	200	68.47	-
	1500	67.50	-
side	50	61.41	63.94
	200	-	-
	1500	65.81	-
[Average, side]	50	72.59	-
	350	70.56	-
	1500	64.61	-

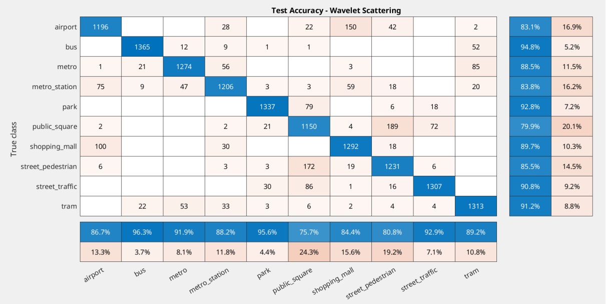
feature from binaural representation of signal, it will lead to better performance. Because it contains more spatial information than average signal. For example if a bus passes in front of microphones, just signal amplitude changes can be captured in averaged(mono) signal and the information which is related to difference of left and right channel will be discarded. So we subtract two channels and compute difference of left and right channels. Wavelet scattering with quality factor of [4,1], [9,4] and invariance scale with 75% of signal length are extracted from average and difference of audio channels, where a logarithmic transformation has been applied to obtain coefficients consequently. Then a concatenation of averaged scattering coefficients is fed into RSM.

3.3. Model settings

In the proposed method we use Regularized Discriminant Analysis(RDA) as base learners of RSM in order to discriminate acoustic scene classes as much as possible. Here a set of RDA learners trained 30 iteration on feature subspace with dimension-350. Finally the mean average of each set is computed and the label of input features defined by majority voting among all learners.

3.4. Results

To evaluate the model and find best accuracy we try different parameter setting on development set based on the evaluation setup presented by challenge. According to Table (1)and (2), wavelet scattering spectrum with quality factor(Q) [4, 1] on side channels achieved best performance among this settings. So we train a model with best setting on development dataset and achieved 78.83% accuracy on leaderboard dataset. It must be denoted that in QDA for some parameters the inverse of covariance matrix can not be computed thus we didn't report accuracy. The final confusion matrix of best model has been shown in figure (2).



Test Accuracy - Wavelet Scattering

True class	airport	bus	metro	metro_station	park	public_square	shopping_mall	street_pedestrian	street_traffic	tram		
airport	1196			28		22	150	42		2	83.1%	16.9%
bus		1365	12	9	1	1				52	94.8%	5.2%
metro	1	21	1274	56			3			85	88.5%	11.5%
metro_station	75	9	47	1206	3	3	59	18		20	83.8%	16.2%
park					1337	79		6	18		92.8%	7.2%
public_square	2			2	21	1150	4	189	72		79.9%	20.1%
shopping_mall	100			30			1292	18			89.7%	10.3%
street_pedestrian	6			3	3	172	19	1231	6		85.5%	14.5%
street_traffic					30	86	1	16	1307		90.8%	9.2%
tram		22	53	33	3	6	2	4	4	1313	91.2%	8.8%
	86.7%	96.3%	91.9%	88.2%	95.6%	75.7%	84.4%	80.8%	92.9%	89.2%		
	13.3%	3.7%	8.1%	11.8%	4.4%	24.3%	15.6%	19.2%	7.1%	10.8%		

Figure 4: Confusion matrix of best model on development dataset

4. REFERENCES

[1] S. Mallat, "Group invariant scattering," *Communications on Pure and Applied Mathematics*, vol. 65, no. 10, pp. 1331–1398, 2012.

[2] W. Wu, Y. Mallet, B. Walczak, W. Penninckx, D. Massart, S. Heuerding, and F. Erni, "Comparison of regularized discriminant analysis linear discriminant analysis and quadratic discriminant analysis applied to nir data," *Analytica Chimica Acta*, vol. 329, no. 3, pp. 257–265, 1996.