ENRICHED SUPERVISED FEATURE LEARNING FOR ACOUSTIC SCENE CLASSIFICATION

A. Rakotomamonjy

Normandie Université, UR, LITIS EA4108
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 2016 competition on acoustic scene classification (Task 1). The approach is based on a supervised feature learning technique which is built upon matrix factorization of time-frequency representation of an audio scene. As an original contribution, we have introduced a non-negative supervised matrix factorization that helps in learning discriminative codes. Our experiments have shown that these supervised features perform slightly better than convolutional neural networks for this challenge. In addition, when they are coupled with some hand-crafted features such as histogram of gradient, their performances are further boosted.

Index Terms— feature learning, matrix factorization, histogram of gradient.

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 and then 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 [1] and the LITIS Rouen Audio scene dataset [2]. For this DCASE 2016 Challenge, a novel dataset for audio scene classification [3] has been released for further fostering development of novel methodologies.

Leveraging on this novel dataset, this paper presents our methodology for classifying acoustic scenes. The method is based on learning features from time-frequency representation of audio scene through a supervised non-negative matrix factorization strategy. This supervision is achieved by augmenting the optimization problem in the non-negative matrix factorization with a term that induces the decomposition to be discriminative in some sense. Our experimental results show that the approach we propose is favorably competitive compared to convolutional neural networks and the best result we achieve is obtained by combining these supervised NMF features with some HOG features.

2. METHOD

2.1. The dataset

The data we have to deal with are composed of 30s 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, 78 segments of $30s$ are available per location. In addition, some specific folds defining 4 sets of training and validation are provided. All the results presented in here are obtained as an average accuracy over the 4 fold.

2.2. Machine Learning Pipeline

The approach we have developed for solving this task is to cast it 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 propose in this work a supervised non-negative factorization technique that allows to learn features.

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 and normalized to unit energy.
- a log mel-frequency representation is obtained from this signal. The frequency span ranges from 0 to the half of the sampling rate. The number of spectral bands we considered is 70 and they are computed over windows of size 25 ms with hops of 10 ms. At this point, each acoustic scene can be represented as a matrix of size 70×2998.

2.2.2. Supervised Matrix Factorization

Our objective in this part is to learn features by leveraging on the labeled examples. Given a particular fold, we have 880 of them in the training set.

We have considered two different strategies : the first one based on convolutional neural networks (CNNs) and the second one, described in the sequel based on supervised matrix factorization. We note that the CNNs approach perform just slightly worse than our supervised matrix factorization method on this dataset but performs better on dataset with a larger amount of training examples.

The idea of non-negative matrix factorization is to find some limited number of positive dictionary elements so that each mel-frequency slice of a given acoustic scene can be represented as positive combination of these elements. Basically, this problem is formalized as the following optimization problem

$$\min_{\mathbf{D}\geq 0,\mathbf{A}\geq 0} \frac{1}{2}\|\mathbf{S}-\mathbf{D}\mathbf{A}\|_2^2 \quad (1)$$

where in our case, $\mathbf{S}$ is the matrix obtained from the concatenation of the mel-frequency representations of all signals, leading to a matrix of size $70 \times (2998 \times 880)$, $\mathbf{D}$ is the matrix containing the discriminative dictionary elements and $\mathbf{A}$ is the coefficient matrix allowing to reconstruct $\mathbf{S}$ from $\mathbf{D}$.

Our objective is to learn discriminative code so that the coefficient matrix $\mathbf{A}$ brings information about class labels in addition to reconstruction information.

For this purpose, we introduce a matrix $\mathbf{C}$ of size $K \times N$, K being the number of dictionary elements and N the number of elements to decompose (the columns of $\mathbf{S}$). The objective of $\mathbf{C}$ is to drive the coefficients in matrix $\mathbf{A}$ to be aligned, in some sense to be defined, to class labels. We achieve this goal by considering that a given dictionary element should be used only for approximating mel-frequency representation of one class of acoustic scene.

For a sake of clarity, suppose that K is a multiple of the number of class, and that the $(i-1)\frac{K}{m}+1$ to $i\frac{K}{m}$ dictionary elements are related to class i. Hence, each entry $c_{i,j}$ is so that $c_{i,j}=1$ if the j-th mel-frequency slice belongs to an acoustic scene of class c and for all $i \in [(c-1)\frac{K}{m}+1, c\frac{K}{m}]$. As an example, if we have a problem with 9 mel-frequency slice ordered in classes, 3 different classes and 6 dictionary elements to be learn, $\mathbf{C}$ writes

$$\mathbf{C}=\begin{pmatrix}1&1&1&0&0&0&0&0&0\\1&1&1&0&0&0&0&0&0\\0&0&0&1&1&1&0&0&0\\0&0&0&1&1&1&0&0&0\\0&0&0&0&0&0&1&1&1\\0&0&0&0&0&0&1&1&1\end{pmatrix}$$

The two first dictionary elements (rows 1 and 2) are devoted to signals from the first class and so on. Hence, $\mathbf{C}$ is a rank m matrix which bears class information owing to the assignment of a given dictionary element to one given class. Hence, the supervised NMF problem we want to solve is now

$$\min_{\mathbf{D}\geq 0,\mathbf{A}\geq 0} \frac{1}{2}\|\mathbf{S}-\mathbf{D}\mathbf{A}\|_2^2+\frac{1}{2}\|\mathbf{C}-\mathbf{R}\mathbf{A}\|_2^2 \quad (2)$$

where $\mathbf{R}$ is a matrix of size $K \times K$. Note that the objective value of this optimization problem balances two terms : one that aims at reconstructing each mel-frequency slice as a positive combination of the dictionary elements and another one which goal is to make coefficients in the matrix $\mathbf{A}$ to be aligned with some label information.

For feature extraction purposes, once the dictionary $\mathbf{D}$ is learned, time-frequency representation of each acoustic scene is decomposed on the non-negative dictionary elements by proceeding slice per slice resulting in a matrix $\mathbf{A}$ of size $K \times 2998$ representing the acoustic scene over the dictionary.

2.2.3. Pooling

The pooling step aims at creating a sketch of the matrix $K \times 2998$ by computing some statistics. These statistics are afterwards used as feature vector for a classifier. In our approach, these statistics are obtained through an integration over the temporal context of the acoustic scene. For instance, we have considered the temporal average and standard deviation over $\mathbf{A}$, leading thus to a feature vector of size $2K$. We have also investigated a temporal maximum pooling as well as the concatenation of these two kinds of feature vector.

Features	AverPrec(%)	Accuracy (%)
rqa	69.09	67.06
hog	78.96	75.04
cnn	80.56	78.46
nmf	81.29	79.74
hogrqa	79.34	75.55
cnnrqa	80.78	78.96
cnnhog	83.28	**80.93**
cnnhogrqa	82.79	80.59
nmfrqa	81.84	80.08
nmfhog	84.08	**81.19**
nmfcnn	80.41	78.37
nmfhogrqa	84.05	81.19

Table 1: Mean over the 4 folds of Average Precision and accuracy using different single features and concatenation of features.

2.2.4. Classifier

After unit-norm normalization, feature vectors are fed to a Gaussian kernel SVM classifier for learning a decision function. We used a *one-vs-one* multi-class strategy. The C parameter of the SVM is selected among 8 values logarithmically scaled between 0.01 and 1000 while the parameter σ of the Gaussian kernel $k(\mathbf{x},\mathbf{x}')=\exp\left(-\frac{\|\mathbf{x}-\mathbf{x}'\|_2^2}{2\sigma^2}\right)$ is chosen among $[0.5,\ 1,\ 5,\ 10,\ 20,\ 30,\ 50,\ 70,\ 100,\ 120]$.

2.2.5. Enriching features

Supervised matrix factorization followed by pooling or a convolutional neural networks are optimized to detect specific patterns in the time-frequency representations of acoustic scenes. As such, they may lack in uncovering discriminative events that are not related to time-frequency structures.

Based on this rationale, we have considered enriching features extracted from supervised matrix factorization with other ones that have been recently deployed for acoustic scene classification. In a very basic way, we have computed histogram of gradient [2] features on the time-frequency representations and recurrence quantitative analysis [4] features and concatenated them to the supervised matrix factorization features.

3. NUMERICAL EXPERIMENTS

We have investigated how the above features perform on the DCASE 2016 dataset Task 1. Averaged results over the 4 fold are reported in Table 1.

First part of the table reports results obtained using single set of features. We can note that using RQA-based features do not perform better than the 72.5% accuracy achieved by the baseline system (MFCC+GMM) described in [3] despite it was the winning

entry for the DCASE 2013 competition [4] . HOG-based features yield about 75.0% accuracy and provide performances consistent with those reported in [2] compared to MFCC and RQA.

The results we report for our supervised matrix factorization have been obtained by concatenating *max* pooling codes as well as *average* and *standard deviation* codes. Compared to hand-designed features, learned features either using CNNs or SNMF yield to a gain of more than 3% of accuracy compared to HOG. Interestingly, CNNs generalize slightly worse than our SNMF. This fact can be eventually explained by the small amount of training examples available (880) for each fold.

The second part of the table provides different results obtained through combination (concatenation) of features. In particular, we have combined learned features with the two hand-crafted ones which have been designed to capture different characteristics of the acoustic scenes. We note that these features provide additional discriminative information to those brought by CNNs and SNMF. Indeed, fusing all these features always yield to improved performances. According to our results, HOG is the best complement to our learned features yielding to an accuracy of respectively 80.93% and 81.19 with CNNs and SNMF.

According to these findings on the development set, for the challenge, we have submitted to results obtained from the concatenation of SNMF and HOG features. At this point, one remaining question is still to be answered : should we use results obtained from the best fold (with optimized hyperparameters) or use results obtained by selecting the best hyperparameters on average? Since multiple submissions were allowed, the decision was to make no decisions and submit both. Interestingly, when looking at the details, these two approaches disagree on 50 examples over the 390 test examples, which can correspond to a variation of 12.5% accuracy in the worst case !

4. CONCLUSION

We have presented in this technical report, our machine learning pipeline for addressing the acoustic scene classification problem. We have proposed a supervised feature learning approach based on a variant of a non-negative matrix factorization technique, which in the particular case of this challenge, performs slighlty better than a convolutional neural networks. In addition, we have shown that if those learned features are combined with some hand-designed features such as histogram of gradient, which capture information about variation of spectral energy, then the ability of our pipeline to recognize acoustic scenes are further enhanced.

5. REFERENCES

[1] 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.

[2] 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.

[3] 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.

[4] G. Roma, W. Nogueira, and P. Herrera, "Recurrence quantification analysis features for environmental sound recognition," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2013.