A Novel Discriminative Feature Extraction for Acoustic Scene Classification Using RNN Based Source Separation

Seongkyu MUN†, Suwon SHON††, Wooll KIM†††, David K. HAN††††, Nonmembers, and Hanseok KO†,††††,† Member

SUMMARY Various types of classifiers and feature extraction methods for acoustic scene classification have been recently proposed in the IEEE Detection and Classification of Acoustic Scenes and Events (DCASE) 2016 Challenge Task 1. The results of the final evaluation, however, have shown that even top 10 ranked teams, showed extremely low accuracy performance in particular class pairs with similar sounds. Due to such sound classes being difficult to distinguish even by human ears, the conventional deep learning based feature extraction methods, as used by most DCASE participating teams, are considered facing performance limitations. To address the low performance problem in similar class pairs cases, this letter proposes to employ a recurrent neural network (RNN) based source separation for each class prior to the classification step. Based on the fact that the system can effectively extract trained sound components using the RNN structure, the mid-layer of the RNN can be considered to capture discriminative information of the trained class. Therefore, this letter proposes to use this mid-layer information as novel discriminative features. The proposed feature shows an average classification rate improvement of 2.3% compared to the conventional method, which uses additional classifiers for the similar class pair issue.

Key words: acoustic scene classification, transfer learning, recurrent neural network, bottleneck feature

1. Introduction

Acoustic scene classification (ASC) is a field of automatically recognizing an audio input as one of predefined classes that characterize the environment (such as, ‘office’, ‘beach’, ‘car’, etc.) where it was recorded. It has recently attracted considerable attention due to a variety of new applications and potential uses [1]–[6]. This attention has given a rise to DCASE challenge 2016 organized by IEEE signal processing society. The goal of the task in the DCASE challenge Task 1 was to classify 15 acoustic scenes [4]. Various approaches have been proposed for this task, such as RNN, convolutional neural network and Deep Neural Network (DNN) based bottleneck features [1].

The results of the final evaluation, however, have shown that even top 10 ranked teams, which achieved an average accuracy rate of greater than 85%, showed extremely low accuracy performance in particular class pairs with similar sounds, such as park/residential area, train/tram, and library/home [1]. Figure 1 shows some examples of the classification results and confusion matrix, which have been published [5]. As can be seen from Fig. 1, some classes show severely low classification rate compared to others, considering average classification rate for all classes. As a typical example, Kim [1] showed 0.0[%] classification rates for ‘library’ class despite obtaining 82.1[%] average classification rate for all 15 classes. This happens because the Mel-filter bank based features, which are widely used in the ASC field, do not have sufficiently discriminative information for similar classes, as shown in Fig. 2. Our previous research [3], [11] using DNN mid-layer information (or bottleneck feature) provided a method for extracting class discriminative information within the given similar feature set (e.g. Mel-frequency cepstrum coefficients (MFCC) and perceptual linear predictive (PLP) features), therefore it also shows a limited performance. To address the issue of low performance among similar acoustic scenes, this letter proposes to employ novel feature extraction method based on...
RNN source separation.

The RNN for source separation is trained on each frequency band between clean target and estimated target from the mixed source. The criterion used for training is the minimum error between separated and original source rather than classification error. Regardless of the class label information, using the criterion for separation allows to extract the discriminative characteristics of target and non-target sources in frequency spectra. In other words, this approach can be considered as a transfer learning method between different objective functions using same DB. As presented in our previous ASC research [2], transfer learning generally consists of pre-training a neural network with relatively large DB and additional network training (or adaptation) with small DB. Unlike this general transfer learning, this letter proposes the novel approach of applying different criterion for the individual training with same DB. The proposed method consists of the network pre-training for source separation and subsequent training for classification.

Since the proposed method utilizes low-level features for ASC, there is a similarity with recent end-to-end filtering methods [6], [7]. However, this approach differs from previous researches in that it uses transfer learning based on different criterions and additively mixed target/non-target DB for training.

In order to extract the proposed features, first, the same number of source separation systems as the number of classes are trained for each target class. After the training is completed, an acoustic input signal is used as the input of classes are trained for each target class. After the training, the target segments were mixed with the selected non-target data. In the RNN, the l-th hidden layer, l > 1, is calculated based on the current input x_t and the activation from the previous time step h_{l-1}(x_{t-1}),

\[ h_l^0(x_t) = f(W_l^0 h_{l-1}^0(x_t) + b_l^0 + U_l^0 h_l^0(x_{t-1})) \]  

where W_l^0 and U_l^0 are weight matrices, and b_l^0 is the bias vector. The first hidden layer is computed as h^{(1)}(x_t) = f(W^{(1)}x_t + b^{(1)} + U^{(1)}h^{(1)}(x_{t-1})). The rectified linear unit (ReLU) is used for activation function f(x). The output layer is computed as:

\[ \hat{y}_t = W^{(3)} h^{(2)}(h^{(1)}(x_t)) + c, \]

where c is a bias vector and \( \hat{y}_t \) is the concatenation of predicted target \( \hat{s} \) and non-target \( \hat{n} \). In this work, we followed the details of output layer masking and discriminative training as presented in [9]. Different ways of using the RNN layer for extracting proposed features will be described in Sect. 3.2.

3. Proposed Feature Extraction

3.1 Training the Source Separation System for Each Class

The process of proposed feature separation is depicted in Fig. 4. As shown in the left part of Fig. 4, first, we trained a dedicated RNN for each target class. From the non-target DB pool the same number of segments as the number of target class segments were randomly selected for generating mixed DB for training. The non-target segments were uniformly selected from each class.

For example, in DCASE 2016 task 1, there are 78 segments for each class. To train the RNN for ‘Bus’ class, we selected 5 segments per class from 14 non-target classes. For the remaining 8 segments (78 – 14 × 5 = 8), white noise was used for mixing. To generate mixed input segments for training, the target segments were mixed with the selected non-target segments at different ratios. Target to non-target ratios of 5, 10 and 20 [dB] were selected empirically. Based on this training DB generation the total of 15 RNN source separation systems were trained.
3.2 Extracting RNN Mid-Level Features

The proposed mid-level feature extraction is shown in the middle part of Fig. 4. A number of previous researches have demonstrated that mid-level features are effective to improve performance in visual objects classification [10], and also audio event recognition [2]. This is because these features have great potential for discriminating various types of classes [10]. They are also useful for transferring information from a large universal domain to small target domains [2].

Based on the advantages shown by aforementioned research, we conducted a research on the usage of mid-level features of the RNN. As depicted in Fig. 3 and Eq. (3), half of the output layer $\hat{y}$, $FE_1$, which contains mask components for target signal, and the mid-layer $h^{(2)}$, $FE_2$, were used for proposed features. Note that for $FE_2$, activation functions were excluded compared to Eq. (1). In the previous source separation research [9], $FE_1$ feature was extracted without the activation function, and the ReLU activation function was used for $FE_2$. In viewpoint of feature extraction, since ReLU makes all negative values zero, information loss occurs on negative components. In order to minimize the information loss of mid-level features (e.g. to maintain the representation capability of the features), we excluded the activation function for $FE_2$.

$$FE_1 \equiv \hat{s} = W^{(3)}_{target} h^{(2)}(h^{(1)}(x_t)) + c_{target}$$

($\hat{s}$: only target source part of $\hat{y}$) \hspace{1cm} (3)

$$FE_2 \equiv W^{(2)} h^{(1)}(x_t) + b^{(2)} + U^{(2)} h^{(2)}(x_{t-1})$$

4. Experimental Settings

For evaluating effectiveness of the proposed method we used the IEEE DCASE 2016 challenge task 1 DB [4]. The DCASE DB contains 15 different acoustic scenes. The DB consists of two subsets: development dataset and evaluation dataset. For each acoustic scene, 78 segments were included in the development dataset and 26 segments were kept for evaluation (each segment is 30 seconds long). The ratio of the training/validation set was 4:1 from DCASE development DB.

The magnitude spectra for RNN input were extracted using a 512-point STFT. The dimensions of mid-level feature sets for each RNN, $FE_1$ and $FE_2$, were 256 and 100 respectively. Feature sets of total 3,840 ($FE_1$) and 1,500 ($FE_2$) dimensions were generated by concatenating 15 classes of RNNs. In our experiments, each mid-level feature set was used as input for classifier individually or concatenated.

To train RNNs for source separation, we used a framework released by Huang et al. [9]. The RNNs were optimized by back-propagation with respect to the discriminative training objectives. For final ASC with proposed feature sets, a fully connected DNN classifier was used. The hyper parameters of the DNN were empirically tuned using 5-fold cross-validation. The classification result of each segment is obtained by accumulating the result (SoftMax output of each class) per frame. Details of both network settings are depicted in Fig. 4.

5. Experimental Results

Examples of source separation results are shown in Fig. 5. Note that these masked and reconstructed spectra were not used as features in this work. The example results were only generated to demonstrate the difference in outputs, in case of target and non-target input of RNN.

In the case of target input, the output of network was similar to input signal, however in the non-target input case, the output was significantly distorted because the network did not have suitable parameters for reconstructing unseen sources. Since these results were generated through mid-layers, $FE_1$ and $FE_2$, it can be considered that these mid-level features have discriminative information for target/non-target classes.

This letter compared the average accuracies over all scenes for the conventional and proposed approaches. Table 1 shows the segment-based classification accuracy. In the DCASE 2016 results, the baseline accuracy of the challenge, which was based on MFCCs and GMMs, was 77.2%. Our previous bottleneck based approach with additional SVMs for similar classes performed 82.3%. The proposed mid-level feature sets showed an improved average accuracy of baseline level. Particularly, the proposed features performed more consistently effective for similar class pairs such as, park/residential area and train/tram. This allowed achieving a higher average accuracy compared to hierarchy DNN-SVMs classification without additional classification processes. In other words, since the features were trained with different criterion, it shows a different performance aspect compared to conventional features. To compare performance with the DCASE 2016 task 1 winner [12], the overall average performance was worse, but for the similar classes of issues, the proposed method achieved
higher performance without the fusion process in [12]. Therefore, further performance improvements are expected by conducting additional feature refining processes, such as I-vector or convolutional neural network (CNN) [12], with proposed features.

In order to analyze the ASC performance according to size of the mid-layers, we conducted an additional experiment with the $FE_{1+2}$ case, which shows the best performance in Table 1. As shown in Table 2, if the size of the mid-layer is too small, the performance becomes degraded due to the lack of information. However, large size of the mid-layer did not always guarantee the improved performance. This can be interpreted by the redundancy of feature set attributing to performance optimizing difficulty. Further research is needed to find and determine the suitable layer size and network configurations.

6. Conclusion

This letter proposed a novel discriminative feature extraction algorithm for ASC using RNN based source separation. Regarding the issue of similar scene classification, the proposed feature set performs better compared to the conventional methods.

Acknowledgments

This material is based upon work supported by the Air Force Office of Scientific Research under award number FA2386-16-1-4130 and authors would like to thank the anonymous reviewers for their valuable comments.

References