Interspeech 2018
2-6 September 2018, Hyderabad

ASe: Acoustic Scene Embedding Using Deep Archetypal Analysis And GMM

Pulkit Sharma[1], *Vinayak Abrol*[2], *Anshul Thakur*[1]

[1]IIT Mandi, India
[2]Idiap Research Institute, Martigny, Switzerland
pulkit_s@students.iitmandi.ac.in, vinayak.abrol@idiap.ch, anshul_thakur@students.iitmandi.ac.in

Abstract

In this paper, we propose a deep learning framework which combines the generalizability of Gaussian mixture models (GMM) and discriminative power of deep matrix factorization to learn acoustic scene embedding (ASe) for the acoustic scene classification task. The proposed approach first builds a Gaussian mixture model-universal background model (GMM-UBM) using frame-wise spectral representations. This UBM is adapted to a waveform, and the likelihood for each spectral frame representation is stored as a feature matrix. This matrix is fed to a deep matrix factorization pipeline (with audio recording level max-pooling) to compute a sparse-convex discriminative representation. The proposed deep factorization model is based on archetypal analysis, a form of convex NMF, which has been shown to be well suited for audio analysis. Finally, the obtained representation is mapped to a class label using a dictionary based auto-encoder consisting of linear and symmetric encoder and decoder with an efficient learning algorithm. The encoder projects the ASe of a waveform to the label space, while the decoder ensures that the feature can be reconstructed, resulting in better generalization on the test data.

Index Terms: Archetypal analysis, deep matrix factorization, acoustic scene classification.

1. Introduction

Environmental sounds carry a significant amount of information about the events taking place in our surroundings. Acoustic scene classification (ASC) is an emerging research area which addresses the problem of automatically classifying sounds produced in environments such as cars passing by, cafeteria, or park. ASC has a lot of potential in various applications, e.g., audio based multimedia search [1], context-aware devices [2] etc. Most of the mobile devices such as smart-phones, hearing aids and robotic platforms are equipped with microphones. These microphones can be used to automatically detect the environment in which these devices are used e.g., conference room or train station and thus different signal processing schemes can be employed [3]. In recent years, ASC is gaining momentum, as being a subtask of the DCASE-17 challenge [4] it received maximum submissions.

Environmental sounds due to their heterogeneous nature are difficult to model. Hence, deep neural network (DNN) frameworks have emerged as one of the most successful approaches, obtaining state of the art results for ASC task. However, DNNs often require massive amounts of labeled training data to generalize well, which explains the success of generative adversarial network (GAN) based approach proposed in [5], the top entry of DCASE-17 challenge. Moreover, feature representations obtained from DNNs are hard to interpret and lack mathematical theory about why they work and what they capture. In contrast, recent studies have shown that conventional machine learning and matrix factorization based approaches also perform well for ASC task [6]. Many of such approaches have recently progressed [7], and a few such as in [8] have proposed hybrid ASC systems by combining conventional audio feature learning approaches with DNNs. Matrix factorization methods such as non-negative matrix factorization (NMF) or dictionary learning (DL) with sparsity constraints, are a class of unsupervised feature learning approaches which describes an acoustic signal as a linear combination of elementary functions that capture salient acoustic information. The sparsity can be beneficial as only a few dictionary atoms encode the signature of events that are important in recognizing a particular acoustic scene, leading to discriminative learning. Given a sequence of audio frames, the obtained representation also encodes the contribution of atoms in time, thus modeling the temporal dynamics of acoustic event. Complimentary to discriminative modeling, one can also use generative models e.g., Gaussian mixture model (GMM), where feature vectors are assumed to be generated from one of a set of underlying statistical distributions. Generative modeling can boost performance by modeling the data variability and using the generative parameters to define a new feature space where discriminative models can be efficiently employed.

Motivated by this, in this paper, we propose a deep learning framework that combines the generalizability of GMMs and discriminative power of matrix factorization to learn acoustic scene embedding (ASe) for the ASC task. Such a generative/discriminative combination helps to exploit the complex structure of acoustic scenes. The proposed system achieves comparable performance to the existing systems, while being trained on less amount of training data as compared to DNNs. Further, with recent algorithmic advancements, it is now possible to quickly perform matrix factorization on massive streaming data sets. Leveraging the recent advancements in deep matrix factorization (DMF) [9, 10], the proposed deep framework is based on archetypal analysis (AA) [11], a form of convex NMF, which has been shown to be well suited for audio analysis [12]. Here, instead of a single level decomposition, we employ deep AA (DAA), by cascading several AA layers to form a deep network and the representations obtained at the final layer are used as a feature for ASC. In other words, data is factorized into multiple factors each highlighting an underlying abstract hierarchical structure. DAA helps in effectively learning the higher level discriminative features present in audio signals for ASC task. The front-end of this discriminative DAA model is complimented by a generative Gaussian mixture model-universal background model (GMM-UBM) build using frame-wise spectral representations of all the training data. This UBM is adapted to each of the training signals and the likelihood for each spectral frame is extracted. In order to deal with loss of affinity problem [13], we perform a pooling by averaging few consecutive spectral frames before adapting UBM for a given audio signal. These likelihood representations are

10.21437/Interspeech.2018-1481

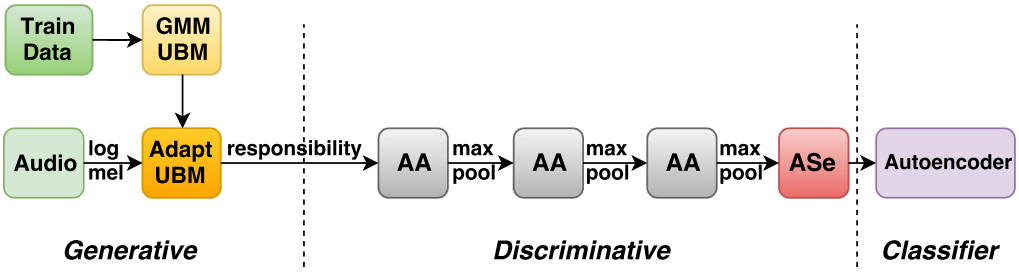

Figure 1: *Block diagram of training in the proposed acoustic scene classification framework.*

used as input representations in the proposed DAA model. The proposed deep model have three layers and each layer consist of an AA module and a max-pooling module. The AA module employs an archetypal based dictionary to derive the convex representations and the max-pooling module employs audio recording level max-pooling operation for dimensionality reduction. AA, which essentially models convex hull provides compact representation and has discriminative ability similar to the sparsity based factorization techniques. This is due to the inherent sparsity of the convex representation. Finally, the obtained representation is mapped to a class label using a logistic regression model. In particular, the proposed approach employ a dictionary based auto-encoder consisting of linear and symmetric encoder and decoder to classify test audio signals. The block diagram depicting the training process in the proposed ASC framework is described in Figure 1. Experimental results provide compelling evidences that the proposed approach performs comparable to existing DNN approaches.

The rest of the paper is organization as: Section 2 briefly reviews the existing approaches for ASC. Section 3 describes basics of AA. The proposed DAA framework for ASe extraction is explained in section 4. Experimental observations are discussed in section 5, and finally the paper is concluded in section 6.

2. Existing approaches for ASC

Initial works on ASC used features inspired from other audio classification tasks with conventional classifiers such as support vector machines. For instance, Geiger et. al., in [14] employed mel-frequency cepstral coefficients (MFCC), zero-crossing rate etc. as a feature representation for ASC. In addition, image processing based techniques are also used to derive features e.g., histograms of oriented gradients from the time frequency representations for ASC [15, 16]. These features mainly focus to derive a particular aspect of the signal, and thus lack flexibility and generalizability. While researchers have explored and adopted many different approaches for speech/audio processing, the state-of-the-art results in DCASE-17 ASC task were obtained by DNN based methods. Although, there were many we briefly review the best performing systems here. The performance of a DNN system improves if the training data increases, hence Mun et. al., employed GAN based method to generate additional training data [5]. They employ support vector machine (SVM) hyperplane for each class as reference for selecting samples, having class discriminative information. The usage of the generated samples resulted in the state-of-the-art results for the ASC task [5]. Han et. al., employed convolutional neural networks (CNN) for identifying an acoustic scene [17]. In order to remove the data scarcity problem for the DNN, they proposed various preprocessing methods which emphasize different aspects of an acoustic scene. These preprocessing methods highlight different acoustic characteristics such as harmonic percussive source separation, binaural representation, and background subtraction. Multiple CNN are individually trained using different preprocessing methods and are combined to form an ensemble model that results in better performance. Zheng et. al., proposed a CNN for ASC task that employs fusion from multiple spectral representation based systems [18]. This method employs CNN to derive features using Fourier transform and constant-Q transform (CQT) spectrogram. The features corresponding to these spectrogram features are used to classify the acoustic scene and results are fused using voting to improves the overall performance for the ASC task.

Apart from DNNs, there are existing works that derive adaptive data representation using feature learning techniques such as NMF for the ASC task [6, 8]. In [8], authors employed NMF to derive feature representation for a hybrid ASC system. Here, a DNN is employed to classify both the NMF representations and the low level frequency representations independently. In order to further improve the performance, these two independent systems are fused together. Another factorization based approach was proposed in [19], where AA instead of NMF was used to derive features for ASC task. While they have similar classification accuracies, the improvement in performance for the approach in [8] seems mainly due to the DNN classifier.

3. Archetypal analysis and related works

The archetypal analysis involves factorization of the data matrix $\mathbf{X} \in \mathbb{R}^{k \times l}$ as: $\mathbf{X} = \mathbf{DA}$, $\mathbf{D} \in \mathbb{R}^{k \times d}$ is the dictionary and $\mathbf{A} \in \mathbb{R}^{d \times l}$ the convex representation matrix. The d atoms of the dictionary lie on the convex hull of the data, and are the convex combination of input data points such that $\mathbf{D} = \mathbf{XB}$. An archetypal dictionary can be learned by solving the following optimizing function [11]:

$$\underset{\substack{\mathbf{B},\mathbf{A} \\ \mathbf{b}_j \in \Delta_l, \mathbf{a}_i \in \Delta_d}}{\operatorname{argmin}} \sum_i \|\mathbf{x}_i - \mathbf{XBa}_i\|_2^2, \qquad (1)$$
$$\Delta_l \triangleq [\mathbf{b}_j \geq 0, \|\mathbf{b}_j\|_1 = 1], \Delta_d \triangleq [\mathbf{a}_i \geq 0, \|\mathbf{a}_i\|_1 = 1],$$

where $\mathbf{a}_i$ and $\mathbf{b}_j$ are the columns of $\mathbf{A}$ and $\mathbf{B} \in \mathbb{R}^{l \times d}$, respectively. The updates of both $\mathbf{A}$ and $\mathbf{B}$ can be formulated as a quadratic programing (QP), for which efficient and fast algorithms are available[1]. Compared to NMF, AA provides compact representation and has better acoustic modeling capability [12].

It is worth noticing that compared to conventional matrix factorization, AA inherently is a deep model with 3 factors, the

[1] Active-set based QP solver: http://spams-devel.gforge.inria.fr

first being the data itself. This suggest adapting AA in deep matrix factorization framework [9,10,13], by further factorizing the representation matrix to reveal hidden attributes of the data. Hence, we propose a deep variant of AA, where the data $\mathbf{X}$ is factorized into $k+1$ factors as:

$$\mathbf{X} \approx \mathbf{X}\mathbf{B}_1\mathbf{A}_1\mathbf{B}_2\mathbf{A}_2 \dots \mathbf{B}_k\mathbf{A}_k. \tag{2}$$

DAA allows the input data to be represented using a hierarchy of $k = 1, 2, \dots K$ layers, given by the following factorizations:

$$\begin{aligned} \mathbf{A}_{k-1} &\approx \mathbf{A}_{k-1}\mathbf{B}_k\mathbf{A}_k \\ &\vdots \\ \mathbf{A}_2 &\approx \mathbf{A}_2\mathbf{B}_3\mathbf{A}_3 \dots \mathbf{B}_k\mathbf{A}_k \\ \mathbf{A}_1 &\approx \mathbf{A}_1\mathbf{B}_2\mathbf{A}_2 \dots \mathbf{B}_k\mathbf{A}_k. \end{aligned} \tag{3}$$

There are many advantages of DAA over conventional DMF approaches such as: 1) dictionary atoms at each layer have geometric meaning, with first layer atoms are archetypes while higher layers ones are prototypes; 2) the convex representations at each layer being sparse and probabilistic are directly interpretable defining the contribution of each atom in the overall representation; 3) representation matrix at each layer is decomposed in itself i.e., it is preserved throughout the network; 4) computing convex representations is much faster and stable than sparse representations. However, learning multiple factors simultaneously in equation (2) is computationally expensive. Hence, a greedy approach can be used to learn layer-wise dictionaries, such that the representation at the $(k-1)^{th}$ layer is factorized into dictionary and the representation matrix at the k^{th} layer [10]. For example, the data $\mathbf{X}$ is factorized at first layer as $\mathbf{X} \approx \mathbf{X}\mathbf{B}_1\mathbf{A}_1$. The representation matrix at the first layer $\mathbf{A}_1$ is factorized at second layer as $\mathbf{A}_1 \approx \mathbf{A}_1\mathbf{B}_2\mathbf{A}_2$, and so on.

4. DAA framework for ASC

4.1. Training and computing ASe:

In this work, log-mel based spectral representations are used as an initial feature representation for each acoustic frame. Initially a GMM-UBM is trained using spectral representations of all the audio frames in the training data. In order to effectively capture the data distribution, GMM-UBM with k mixtures is trained from short-time audio frames without any temporal pooling step. Once a GMM-UBM is trained, the training process involves deriving spectral features (with temporal context) from the audio signal and adapting the GMM-UBM to individual training recording. For each audio recording we extract four audio waveforms for processing i.e., two from both channels, one sum of two channels and one difference of two channels. In order to deal with loss of affinity problem [13], we perform a pooling by averaging few consecutive spectral frames before adapting UBM for a given audio signal. The audio is divided into non-overlapping segments of length W to model temporal context. These frames are decomposed with a short-time Fourier transform applying a window of length (*win*) with a shift (*ovl*). The resulting spectrogram is log transformed and averaged across frames, after applying mel-filters, resulting in a nf-D feature from each segment of audio. The segment-wise likelihood for the spectral representation (of each of the training waveform) is computed and stored as a matrix $\mathbf{X} \in \mathbb{R}^{k \times l} = [\mathbf{X}_1\mathbf{X}_2 \dots \mathbf{X}_q]$, such that $\mathbf{X}_i \in \mathbb{R}^{k \times n}$ $(i = 1 \dots q)$ denote a feature matrix for i^{th} waveform. The training matrix $\mathbf{X}$ is fed to a deep factorization pipeline based on AA and audio waveform level max pooling.

Table 1: *Matrix dimensions and parameter settings.*

$\mathbf{X}$	$\mathbf{B}_k$	$\mathbf{A}_k^{'}$	$\mathbf{S}$	$\mathbf{W}$	$\mathbf{Y}$
$k \times l$	$l_k \times d_k$	$d_k \times l_k$	$d_k l_k \times q$	$c \times d_k l_k$	$c \times q$

	W	win	ovl	k	nf	n
UBM	-	$25ms$	$10ms$	100	64	976
Feature extraction	$200ms$	$25ms$	$10ms$	100	64	50

	d_1	d_2	d_3	$mp-width$	$mp-stride$	λ
DAA	100	50	30	3	2	.25

The proposed framework uses a three level of hierarchical factorization of data matrix $\mathbf{X}$. In the first level the data is factorized into an archetypal dictionary and the corresponding convex representation as $\mathbf{X} = \mathbf{D}_1\mathbf{A}_1 = \mathbf{X}\mathbf{B}_1\mathbf{A}_1$, such that dictionary $\mathbf{D}_1 \in \mathbb{R}^{k \times d_1}$ and the convex representation $\mathbf{A}_1 \in \mathbb{R}^{d_1 \times l}$. Further, a max pooling operation (denoted by $mp()$) is employed on the representations obtained for each acoustic signal (on a frame by frame basis) to obtain a matrix $\mathbf{A}_1^{'} \in \mathbb{R}^{d_1 \times l_1}$ as:

$$\mathbf{A}_\mathbf{1}^{'} = [mp(\mathbf{A}_{11}) \;\; mp(\mathbf{A}_{12}) \dots mp(\mathbf{A}_{1q})], \tag{4}$$

where $\mathbf{A_{1i}}$ is the representation corresponding to $\mathbf{X}_i$. This whole audio waveform level pooling operation is denoted by $fmp()$. The proposed DAA framework employs AA and max pooling at each layer in a hierarchy of k layers as:

$$\begin{aligned} \mathbf{X} \approx \mathbf{X}\mathbf{B}_1\mathbf{A}_1, \quad \mathbf{A}_1^{'} &= fmp(\mathbf{A}_1) \\ \mathbf{A}_1^{'} \approx \mathbf{A}_1^{'}\mathbf{B}_2\mathbf{A}_2, \quad \mathbf{A}_2^{'} &= fmp(\mathbf{A}_2) \\ &\vdots \\ \mathbf{A}_{k-1}^{'} \approx \mathbf{A}_{k-1}^{'}\mathbf{B}_k\mathbf{A}_k, \quad \mathbf{A}_k^{'} &= fmp(\mathbf{A}_k). \end{aligned} \tag{5}$$

Here, the output feature matrix for each recording has size $d_k \times l_k$, and the dimension l_k depends on pooling window size and stride. The final fixed $(d_k l_k)$-D feature obtained by flattening the output matrix is referred as *Acoustic Scene embedding (ASe)*. ASe from all q waveforms are stored as a matrix $\mathbf{S}$ and is used to train a classifier to obtain class labels during testing. The final class label for a given recording is assigned based on majority vote over its corresponding four waveforms.

4.2. Classification:

The proposed deep learning model is trained in an unsupervised manner and does not require any labels to learn features from the audio recordings, which are then used to train a classifier separately. This is in contrast to DNNs where feature learning and classification are done in an end-to-end manner. We adopted unsupervised learning to make the proposed model less demanding in terms of tuning parameters and need of large amount of labeled training data. Once a final representation is obtained for an audio segment, it can be mapped to a class label using a trained logistic regression model. To this aim the proposed approach employs an autoencoder, which is trained with the representation obtained via deep factorization as input to its encoder and the corresponding class label as the encoder's output. The testing phase involves deriving a deep representation for a given test signal and using the encoder to map this representation to a class label. The auto-encoder consisting of linear

Table 2: *Comparison of the classification accuracy (CA) of the proposed method with the existing methods for the ASC task.*

	ASe	**GAN-A** [5]	**CNN-H** [17]	**CNN-G** [20]	**CQT-Z** [18]	**NMF-B** [8]
CA (%)	79.3	83.3	80.4	81.5	77.7	69.8
	KUK-D [21]	**DSNMF** [9]	**LEH-L** [22]	**PEK-P** [23]	**KGP-W** [24]	**GMM-AA** [19]
CA (%)	71.1	72.4	73.8	72.6	67.0	65.7

and symmetric encoder and decoder, where the encoder projects the ASe of a recording to the label space, while the decoder ensures that the feature can be reconstructed, resulting in better generalization on the test data. This is achieved by optimizing the following objective:

$$\underset{\mathbf{W}}{\operatorname{argmin}} \|\mathbf{S} - \mathbf{W}^*\mathbf{W}\mathbf{S}\|_F^2 \;\text{ s.t. }\; \mathbf{W}\mathbf{S} = \mathbf{Y}, \tag{6}$$

where $\mathbf{Y}$ is the label space with each column in $\mathbb{R}^c$ for c classes with entry 1 corresponding to true class and 0 elsewhere. Assuming tied weights i.e., $\mathbf{W}^* = \mathbf{W}^T$, one can simplify and regularize the autoencoder training as:

$$\underset{\mathbf{W}}{\operatorname{argmin}} \|\mathbf{S} - \mathbf{W}^T\mathbf{Y}\|_F^2 + \lambda \|\mathbf{W}\mathbf{S} - \mathbf{Y}\|_F^2, \tag{7}$$

where λ is a parameter that controls the losses of the decoder and encoder respectively. Eq. (7) is in standard quadratic form, has only one tuning parameter λ and can be transformed into well-known Sylvester equation, having fast and efficient algorithms to obtain the solution [25].

5. Experimental Observations

5.1. Dataset:

In this work, we use the dataset for acoustic scene classification in DCASE-17 challenge, which contain 13 hours of urban audio scenes [4]. This data consists of 15 different acoustic scenes recorded using binaural microphones.

5.2. Evaluation:

The training-development-evaluation splits are kept the same as provided in the challenge protocol. For clarity of readers all the experimental parameters and dimensions are tabulated in Table 1, and the these parameters are obtained after 4-fold cross validation on development set. The proposed DAA pipeline results in 150-D ASe for each audio waveform. The performance of the proposed and existing ASC system is measured in terms of average classification accuracies on evaluation data.

5.3. Results and discussion:

The classification accuracy of the proposed framework and its comparison with existing methods for the DCASE-17 ASC task is provided in Table 2. Our system achieved an accuracy of $86\% \pm 0.5$ (95% C.I.) averaged across 4 different folds using the provided development dataset. Thus we achieved an improvement of 11.2% over the DCASE-17 baseline system.

The proposed ASe method lags behind the top performers of the ASC DCASE-17 challenge [5], labeled as GAN-A. However, we would like to emphasize that Mun et. al., in [5] employs GAN to augment the training set. The performance of the proposed method is comparable to the one proposed in [17] (labeled as CNN-H), which uses various pre-processing methods to highlight different acoustic characteristics of the data. However, the proposed ASe method outperforms the CNN based ASC method based on multiple spectrograms fusion as proposed in [18] (labeled as CQT-Z). Thus, the proposed framework ASe provides better result than the DNN if there is no data augmentation in the training data, as in case of CQT-Z. On the contrary, for the case of data augmentation during training the performance of the proposed ASe framework is not at par with those of DNN as in GAN-A. Hence, including a pre-processing enhancement step data augmentation will further improve the performance of the proposed approach. The proposed method is also outperforming other systems submitted for the DCASE-17 ASC task, labeled as KUK-D, LEH-L, PEK-P, GMM-AA, KGP-W as proposed in [21], [22], [23], [19] and [24], respectively.

In addition, we have observed that the performance of the proposed method is better than the NMF based method proposed in [8] (labeled as NMF-B). This is because DAA involve hierarchical factorization which results in better discriminative features. Also in case of AA the dictionary atoms lie on the boundary of convex hull resulting in compact representation, compared to the case of NMF where atoms lie outside convex hull. In addition, the pooling operation employed in the proposed framework both during generative and discriminative modeling helps in learning invariant features. To investigate this, we replace the DAA pipeline of the proposed approached with a recently proposed deep semi-NMF approach [9] with max-pooling (labeled as DSNMF). Experimental results confirm that AA analysis results in better performance than NMF and is inherently better suited for ASC task.

We have performed experiments by extracting features using the CNN model labeled as CNN-G) pre-trained on audio set [26] (a dataset with 632 audio events), with a SVM classifier. The details of the CNN architecture are available in [20]. It has been observed that CNN-G perform slightly better than the proposed model, as expected since it's trained on a larger dataset. However, the performance of CNN-G is still not comparable to the state-of-the-art GAN-A method which is well adapted for DCASE-17 dataset. This shows that the proposed approach can be extended to large-scale audio classification and we defer this for future work.

6. Conclusions

In this paper, we proposed a novel deep learning framework that combines the generalizability of GMM and discriminative power of DMF to learn ASe features for the ASC task. Here, a UBM model build using the training data is used to derive the likelihood for each frame of an audio signal, which is then used in the DAA pipeline. The proposed DAA model also employed an audio recording level max-pooling that reduces the dimensionality, in addition to providing abstracted form of the information. The ASe of the audio recording is projected to the label space using the encoder of an auto-encoder, while the reconstruction in decoder helps in better generalization of the test data. Experimental results on the ASC task of DCASE-17 challenge demonstrate that the performance of the proposed framework is comparable to the existing state-of-the-art methods.

7. References

[1] M. Bugalho, J. Portelo, I. Trancoso, T. Pellegrini, and A. Abad, "Detecting audio events for semantic video search," in *INTERSPEECH*, 2009, pp. 1115–1154.

[2] A. J. Eronen, V. T. Peltonen, J. T. Tuomi, A. P. Klapuri, S. Fagerlund, T. Sorsa, G. Lorho, and J. Huopaniemi, "Audio-based context recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 1, pp. 321–329, Jan. 2006.

[3] J. Schroder, N. Moritz, J. Anemuller, S. Goetze, and B. Kollmeier, "Classifier Architectures for Acoustic Scenes and Events: Implications for DNNs, TDNNs, and Perceptual Features from DCASE 2016," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 6, pp. 1304–1314, Jun. 2017.

[4] T. Virtanen, A. Mesaros, T. Heittola, A. Diment, E. Vincent, E. Benetos, and B. M. Elizalde, *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017)*. Tampere University of Technology. Laboratory of Signal Processing, 2017.

[5] 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," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017)*, Nov. 2017, pp. 93–102.

[6] V. Bisot, R. Serizel, S. Essid, and G. Richard, "Acoustic scene classification with matrix factorization for unsupervised feature learning," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2016, pp. 6445–6449.

[7] ——, "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, Jun. 2017.

[8] ——, "Nonnegative feature learning methods for acoustic scene classification," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017)*, Nov. 2017, pp. 22–26.

[9] G. Trigeorgis, K. Bousmalis, S. Zafeiriou, and B. W. Schuller, "A deep matrix factorization method for learning attribute representations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 3, pp. 417–429, Mar. 2017.

[10] P. Sharma, V. Abrol, and A. K. Sao, "Deep sparse representation based features for speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 11, pp. 2162–2175, Nov. 2017.

[11] Y. Chen, J. Mairal, and Z. Harchaoui, "Fast and robust archetypal analysis for representation learning," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2014, pp. 1478–1485.

[12] A. Thakur, V. Abrol, P. Sharma, and P. Rajan, "Compressed convex spectral embedding for bird species classification," in *IEEE International Conference on Acoustics, Speech, Signal Processing (ICASSP)*, Apr. 2018.

[13] Y. He, K. Kavukcuoglu, Y. Wang, A. Szlam, and Y. Qi, "Unsupervised feature learning by deep sparse coding," in *SIAM International Conference on Data Mining (SDM)*, Apr. 2014, pp. 902–910.

[14] J. T. Geiger, B. Schuller, and G. Rigoll, "Large-scale audio feature extraction and svm for acoustic scene classification," in *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, Oct. 2013, pp. 1–4.

[15] V. Bisot, S. Essid, and G. Richard, "HOG and subband power distribution image features for acoustic scene classification," in *European Signal Processing Conference (EUSIPCO)*, Aug. 2015, pp. 719–723.

[16] A. Rakotomamonjy and G. Gasso, "Histogram of gradients of time-frequency representations for audio scene classification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 142–153, Jan. 2015.

[17] Y. Han, J. Park, and K. Lee, "Convolutional neural networks with binaural representations and background subtraction for acoustic scene classification," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017)*, Nov. 2017, pp. 46–50.

[18] W. Zheng, J. Yi, X. Xing, X. Liu, and S. Peng, "Acoustic scene classification using deep convolutional neural network and multiple spectrograms fusion," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017)*, Nov. 2017, pp. 133–137.

[19] V. Abrol, P. Sharma, and A. Thakur, "GMM-AA system for acoustic scene classification," DCASE2017 Challenge, Tech. Rep., Sep. 2017.

[20] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, and K. Wilson, "CNN architectures for large-scale audio classification," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2017, pp. 131–135.

[21] I. Kukanov, V. Hautamki, and K. A. Lee, "Recurrent neural network and maximal figure of merit for acoustic event detection," DCASE2017 Challenge, Tech. Rep., Sep. 2017.

[22] B. Lehner, H. Eghbal-Zadeh, M. Dorfer, F. Korzeniowski, K. Koutini, and G. Widmer, "Classifying short acoustic scenes with I-vectors and CNNs: Challenges and optimisations for the 2017 DCASE ASC task," DCASE2017 Challenge, Tech. Rep.

[23] S. Park, S. Mun, Y. Lee, and H. Ko, "Acoustic scene classification based on convolutional neural network using double image features," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017)*, Nov. 2017, pp. 98–102.

[24] S. Waldekar and G. Saha, "IIT kharagpur submissions for DCASE2017 ASC task: Audio features in a fusion-based framework," DCASE2017 Challenge, Tech. Rep., Sep. 2017.

[25] E. Kodirov, T. Xiang, and S. Gong, "Semantic autoencoder for zero-shot learning," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017, pp. 4447–4456.

[26] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2017, pp. 776–780.