INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Attention-Driven Projections for Soundscape Classification

Dhanunjaya Varma Devalraju, Muralikrishna H, Padmanabhan Rajan, Dileep Aroor Dinesh

School of Computing and Electrical Engineering, Indian Institute of Technology, Mandi
{s18023,d17021}@students.iitmandi.ac.in, {padman,addileep}@iitmandi.ac.in

Abstract

Acoustic soundscapes can be made up of background sound events and foreground sound events. Many times, either the background (or the foreground) may provide useful cues in discriminating one soundscape from another. A part of the background or a part of the foreground can be suppressed by using subspace projections. These projections can be learnt by utilising the framework of robust principal component analysis. In this work, audio signals are represented as embeddings from a convolutional neural network, and meta-embeddings are derived using an attention mechanism. This representation enables the use of class-specific projections for effective suppression, leading to good discrimination. Our experimental evaluation demonstrates the effectiveness of the method on standard datasets for acoustic scene classification.

Index Terms: Acoustic scene classification, robust principal component analysis, subspace projection, attention.

1. Introduction

In many real-world applications, data under study can be seen as the superposition of occasional events (sparse outliers) and constant or slow-changing background. One such familiar application is surveillance video sequence, where a slow-changing background scene is interspaced by movement of one or more people or objects. Separating or decomposing the video data into the sparse component and background component may be useful in tasks like activity detection. *Robust principal component analysis* (RPCA) is a technique which performs such a separation [1] [2]. RPCA decomposes data matrix M in to L and S:

$$M = L + S,$$

where low-rank matrix L corresponds to the slow-changing background and sparse matrix S corresponds to the outliers. Based on the application, the object of interest can be either the sparse component or the low-rank component, or both.

The analysis of acoustic scenes (or soundscapes) is one such application, where data can be seen as the superposition of sparse and low-rank components. For example, in the soundscape inside a bus, there is a constant engine sound (the low-rank component), interspaced by sounds like people talking or the door opening (sparse component). It may be sometimes useful to separate the sparse foreground events (people talking) from the constant or slow-changing background (the sound of engine) for analysing the soundscape.

Furthermore, in many situations, the foreground events may be crucial in discriminating soundscapes with similar backgrounds. Likewise, the background may be crucial in discriminating soundscapes with similar foreground events. Moreover, it may be helpful to suppress a part of the foreground or background, instead of completely removing them. Suppressing a part of foreground or background can be done by subspace projections like nuisance attribute projection (NAP) [3], where either the foreground or the background can be treated as the nuisance (or unwanted) attribute. The unwanted variations can be removed from the data vector x by applying the below transformation:

$$\tilde{x} = x - BB^T x = (I - BB^T)x \tag{1}$$

Here $\tilde{x}$ denotes the nuisance-removed vector, B is a basis matrix whose columns span the nuisance space, and I is the identity matrix.

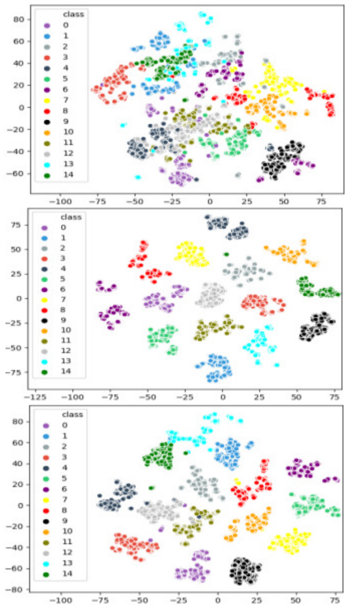


Figure 1: *Various t-SNE plots of embeddings. Top figure: without using explicit class information. Middle: Using explicit class information and class-specific basis to remove part of the background. Bottom: Without using class information and by using proposed meta-embeddings driven by attention. Each color corresponds to one class.*

Due to the unpredictability of real-life soundscapes, it is difficult to have a universal nuisance basis matrix B which will work for all soundscapes. In practice, a class-specific basis [1] matrix is more suitable for applying NAP. This is illustrated in Figure 1 with the help of 2-D t-SNE plots of audio embeddings. The top plot in Figure 1 shows features from 15 acoustic soundscapes. It can be observed that there is a lot of overlap between the classes. The middle plot shows the audio embeddings after suppressing part of the background using the respective class-specific basis matrices. It can be seen that, in this case, the embeddings are well separated. The only issue with this approach

[1] Class-specific basis means that, given data of a class, the foreground basis B^f and the background basis B^b for NAP are constructed with prior knowledge of the class.

http://dx.doi.org/10.21437/Interspeech.2020-2476

is that, during deployment, we do not know the true class, and hence we do not know which basis matrix to use. This paper addresses this issue.

In this work, all audio signals are represented as embeddings from a deep convolutional neural network (CNN). Let the total number of classes under consideration be C. We propose a new framework for RPCA-based subspace projection, where in, an audio embedding is projected into C separate subspaces using class-specific bases. We adopt the attention-based model in [4] for combining these C projected embeddings, corresponding to a single audio sample, into one embedding. This attention model is trained to produce an embedding which will be similar to the one using it's class-specific basis. In other words, we bypass the problem of explicitly choosing the class-specific basis by training the network to come up with an embedding similar to the one produced by the class-specific basis, without knowledge of the class label.

We term the learned embedding as *meta-embedding*. The word meta-embedding is borrowed from natural language processing (NLP) literature, which means embedding learned by combining different embedding sets [5]. Later a support vector machine (SVM) is trained using these meta-embeddings for acoustic scene classification (ASC) task. The third plot in Figure 1 shows the t-SNE plot with meta-embeddings obtained with the proposed attention model. From the plot we can observe the classes are well separated, similar to the plot which utilized the class-specific projection. This endorses our belief in using the attention model to produce embeddings similar to the class-specific projection embeddings.

Research in acoustic scene classification has been spurred by the DCASE challenges [6]. Several state-of-the-art systems for ASC can be found in the challenge entries, which include [7], [8] and [9]. In the literature, many studies have addressed the ASC task by extracting rich features. Mainly due the success of CNNs in computer vision, CNN based architectures are adopted in ASC as well. Such approaches include the use of time-frequency representation of an audio signal such as scalogram [10], constant-Q transform spectrogram [8] [11] and log-mel spectrogram [12] [7] treated as input images. CNN based models are then trained to extract rich features for downstream classification. Some studies applied source separation methods to extract discriminative features, though the explicit low-rank and sparse representation as in RPCA is not applied. Mun et al.[13] proposed to extract discriminative features from the intermediate layer of an recurrent neural network (RNN) for ASC, where the RNN model was trained for source separation. Han et al.[9] proposed to use spectrograms from binaural audio, harmonic-percussive source separation and background subtraction to train an ensemble neural network to achieve better classification accuracy. In this paper, we explore RPCA based decomposition with meta-embeddings as rich discriminative features for SVMs.

2. Robust principal component analysis

Principle component analysis (PCA) is a technique widely used in data analysis mostly for dimensionality reduction and denoising. PCA can be solved by performing singular value decomposition (SVD) on the given data matrix M. But SVD is sensitive to outliers and performs poorly when data is corrupted with outliers. In real world, data corruption is quite common and due to which, PCA tends to find the directions which are far from the true directions. Robust PCA (RPCA) overcomes some of these limitations with reasonable assumptions about the data [1]. A well-used formulation of RPCA is the problem of decomposing the data matrix M into the sum of a low-rank matrix L and sparse matrix S [2]. By solving the following convex problem we can recover the low-rank matrix:

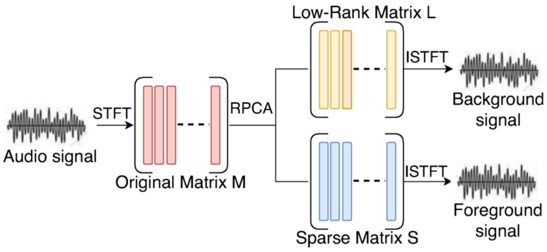


Figure 2: *RPCA based foreground and background separation. The phase of original audio signal is used to reconstruct foreground and background signals [14].*

$$\text{minimize } ||L||_* + \lambda ||S||_1, \tag{2}$$

$$\text{subject to } L + S = M, \tag{3}$$

where M, L and $S \in \mathbb{R}^{n1\times n2}$, $\lambda > 0$ is a free parameter, $||\cdot||_*$ denotes the nuclear norm, i.e., sum of singular values and $||\cdot||_1$ denotes the l_1-norm. We use the procedure proposed by Huang et al.[14] to separate singing voice from monaural recordings, including details on how to solve the above convex problem and the choice of the λ value.

Fig 2 illustrates the procedure of RPCA applied to an audio signal. The spectrogram representation of the audio signal is treated as data matrix M. This data matrix M is approximated as $L+S$, where L and S represents the background and the foreground spectrograms respectively. The phase of original signal and inverse short-time Fourier transform are used to reconstruct the background and foreground audio signals from L and S.

3. Attention

Attention models were first proposed for machine translation [15] [16], where the words in a sentence are attended differently. Attention models are designed to give more relevance to important words and ignore irrelevant words. A variant of the attention model is proposed by Kiela et al [5] in which, decision of picking a word embedding for a given setting is made by a neural network. This learned word embedding is known as dynamic meta-embedding. Kong et al [4] applied attention mechanism for audio tagging and sound event detection with weakly labelled data. They showed that attention models can be used for decision making in a multiple instance learning setting. Also, they proposed decision-level as well as feature-level attention neural networks for audio tagging. Our work is conceptually similar to dynamic meta-embedding [5], where in, we wish the model to select the suitable embeddings.

4. The proposed method

Let $D_i = \{x_{i1}, x_{i2}, x_{i3}, ..., x_{in}\}$ be n training samples for class i, $i \in \{1, 2, ..., C\}$. Let $D_i^f = \{x_{i1}^f, x_{i2}^f, x_{i3}^f, ..., x_{in}^f\}$ be n foreground training samples for class i. Let $D_i^b = \{x_{i1}^b, x_{i2}^b, x_{i3}^b, ..., x_{in}^b\}$ be n background training samples for class i. All these samples $x_{ik}, x_{ik}^f, x_{ik}^b \in \mathbb{R}^d$ are the feature vectors extracted from the last layer of the CNN L^3-Net [17]

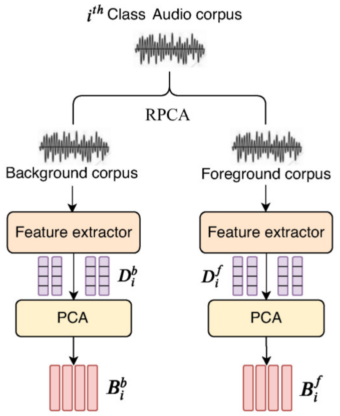


Figure 3: *Class-specific background and foreground basis construction for NAP, where* i *specifies the class.*

[18]. x_{ik} is derived from the input audio sample, x_{ik}^f and x_{ik}^b are derived after performing RPCA on the input audio sample.

Let $B_i^f, B_i^b \in \mathbb{R}^{d\times d}$ be the foreground and background NAP bases corresponding to class i respectively. These class-specific bases are computed by performing PCA on foreground and background training samples belonging to class i, i.e., D_i^f and D_i^b respectively as shown in Figure 3.

The proposed framework is illustrated in Figure 4. First, features are extracted from the given audio sample, let this be represented by x. For the moment, let us consider the foreground sound events as nuisance attributes. In this case, NAP is performed on x as shown below.

$$\tilde{x}_i^b = x – B_i^f B_i^{fT} x \quad \forall i \in \{1,2,..,C\} \tag{4}$$

$$\widetilde{X}^b = \{\tilde{x}_1^b, \tilde{x}_2^b, .., \tilde{x}_C^b\} \tag{5}$$

Here $\widetilde{X}^b \in \mathbb{R}^{d\times C}$ is the set of vectors that represents foreground removed representation for a single audio sample. Similarly, if we consider background sound events as nuisance attributes, then NAP is performed on x as shown below.

$$\tilde{x}_i^f = x – B_i^b B_i^{bT} x \quad \forall i \in \{1,2,..,C\} \tag{6}$$

$$\widetilde{X}^f = \{\tilde{x}_1^f, \tilde{x}_2^f, .., \tilde{x}_C^f\} \tag{7}$$

Here $\widetilde{X}^f \in \mathbb{R}^{d\times C}$ is the set of vectors that represents background removed representation for a single audio sample. In both the representations, each sample is projected in to C different subspaces using class-specific bases.

We adopt the feature-level attention in [4] for learning meta-embedding, whose model architecture is shown in Figure 5. This neural network has one input layer of length d and output layer of length C (d is the dimension of embeddings and C is the number of classes), and 2 parallel fully-connected hidden layers of length l, one for attention function $p()$ and another for learning better representations $u()$.

We denote the input to the attention model as $\widetilde{X} \in \mathbb{R}^{d\times C}$. If we are considering the background as the nuisance attribute, then $\widetilde{X} = \widetilde{X}^f$ otherwise $\widetilde{X} = \widetilde{X}^b$. Let $\tilde{x} \in \mathbb{R}^d$ be a column vector in $\widetilde{X}$ i.e., $\tilde{x} \in \widetilde{X}$. The function $u(\tilde{x})$ can be modelled as

$$u(\tilde{x}) = \sigma(W\tilde{x} + b) \tag{8}$$

where $u(\tilde{x}) \in \mathbb{R}^l$, σ can be any linear or non-linear function to increase the representation ability of the model. The attention function $p(\tilde{x})$ can be modelled as

$$v(\tilde{x}) = \phi(U\tilde{x} + c), \tag{9}$$

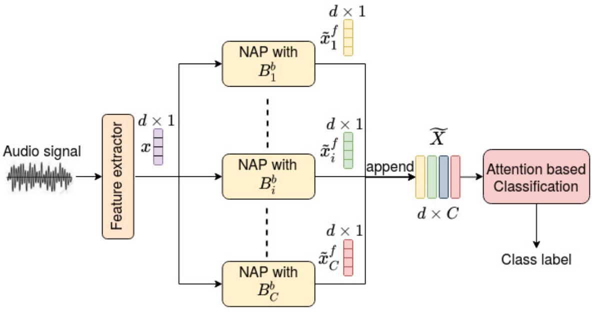
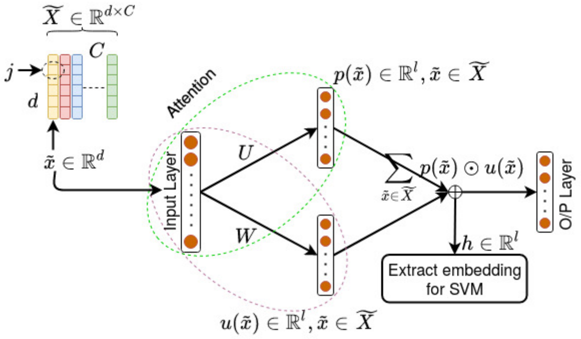


Figure 4: *Proposed framework to suppress class-specific background using NAP, where* i *specifies the class. A self-attention mechanism is used to derive a final embedding from these class-specific embeddings for classification. Same framework is used to suppress class-specific foreground as well by performing NAP with* B_i^f *instead of* B_i^b*, which gives* $\tilde{x}_i^b$ *instead of* $\tilde{x}_i^f$*.*



$\widetilde{X} \in \mathbb{R}^{d\times C}$
C
j
d
$\tilde{x} \in \mathbb{R}^d$
Attention
Input Layer
U
W
$p(\tilde{x}) \in \mathbb{R}^l, \tilde{x} \in \widetilde{X}$
$u(\tilde{x}) \in \mathbb{R}^l, \tilde{x} \in \widetilde{X}$
$\sum_{\tilde{x}\in\widetilde{X}} p(\tilde{x}) \odot u(\tilde{x})$
$h \in \mathbb{R}^l$
Extract embedding for SVM
O/P Layer

Figure 5: *Self-attention model architecture. The embeddings extracted post fusion are used to train the SVM classifier.*

$$p(\tilde{x})_j = v(\tilde{x})_j / \Sigma_{\tilde{x}\in\widetilde{X}} v(\tilde{x})_j \tag{10}$$

where ϕ is the sigmoid function to view $p(\tilde{x})_j$ as a probability and $p(\tilde{x}) \in \mathbb{R}^l$, $v(\tilde{x}) \in \mathbb{R}^l$ and j represents the index term. Equation 10 can be interpreted as *softmax* over $v(\widetilde{X})$ along the dimension C, where $v(\widetilde{X}) \in \mathbb{R}^{l\times C}$ and $\tilde{x} \in \widetilde{X}$. Equation 10 is repeated for each component of $p(\tilde{x})$. The attention aggregation of $u(\tilde{x})$ and $p(\tilde{x})$ produces a meta-embedding by combining C embeddings in to a single vector as:

$$h = \sum_{\tilde{x}\in\widetilde{X}} p(\tilde{x}) \odot u(\tilde{x}) \tag{11}$$

where $h \in \mathbb{R}^l$ is the meta-embedding, $\odot$ represents element-wise product. This model is trained for classification using Adam optimizer and categorical cross-entropy loss function. Once this model is trained, features are extracted from the layer before the output, which gives meta-embedding h.

The effect of applying the attention module is illustrated in Figure 6. The top row of Figure 6 shows the bivariate kernel density estimate contours for two acoustic scenes “grocery store” and “metro station” using two randomly chosen dimensions from 234 training audio samples. We can clearly see that features from these two classes have high overlap. After obtaining the meta-embedding after attention, the overlap is considerably reduced. This is also the case for another two acoustic scenes “home” and “office”, shown in bottom row of Figure 6.

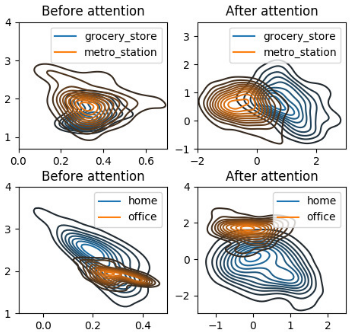


Figure 6: *Illustration of bivariate kernel density estimate with 2 randomly picked features. Top row: scenes "grocery store" and "metro station" before and after applying attention. Bottom row: scenes "home" and "office" before and after applying attention.*

5. Experimental evaluation

In this section, we describe the experiments used to evaluate the proposed framework for acoustic scene classification. The primary purpose of the experiments is to investigate the effect of the suppression of the foreground and background. The suppression is achieved using meta-embeddings, and the amount of suppression is controlled by the dimension of the subspace during NAP.

For comparison, we use a baseline system which uses features from the input audio sample (considered together with foreground and background). These are embeddings from L^3-Net without applying NAP or attention. We also compare our method with systems reported in [7] and [8] on the same data. The system in [7] is an ensemble of different subsystems, each trained with different features like MFCC features, audio segment level feature vectors, mel and log scaled filter banks. The system in [8] is a fusion of multiple features from multiple CNNs trained on various spectral representations of the audio.

Datasets: We evaluate the proposed framework using two acoustic scene classification (ASC) datasets. a) TUT Acoustic Scenes 2017 dataset (DCASE 2017 task 1) [6]. The dataset has a development set and an evaluation set corresponding to 15 acoustic scenes. b) LITIS Rouen Audio scene dataset (3026 samples) comprising of 19 acoustic scene classes [19].

Feature extraction: We use Openl3 python library to extract deep audio embeddings from an audio sample [17] [18]. The extracted embedding is then averaged to get a 6144 $\times$ 1 vector ($d = 6144$).

Learning class-specific bases for NAP: We utilize the Matlab implementation of RPCA in [14] to separate foreground and background from an audio sample (see Figure 2). Post separation, features are extracted from the foreground as well as background samples using Openl3 library as discussed earlier. Then, PCA is performed on the class-specific foreground and background samples to get class-specific nuisance basis for the foreground B_i^f and background B_i^b respectively as shown the Figure 3. We can control the amount of foreground and background to be removed by varying the number of principal components used as columns in B_i^f and B_i^b respectively. In all cases, the classifier is a simple one against one SVM with linear kernel.

Results and discussions: The first orange bar in Figure 7(a), (b) shows the results of the baseline system with development and evaluation datasets respectively. Figure 7 also shows the performance of the proposed attention-based systems, when foreground is suppressed (NAP-F) and when the background is suppressed (NAP-B), on the DCASE development dataset. Foreground suppression uses 200 columns for the NAP basis, and background suppression uses 50 columns, with d = 6144 and l = 256. We observe 2% and 3.8% increase in classification accuracy from the baseline, when the foreground and the background are treated as nuisance attributes respectively.

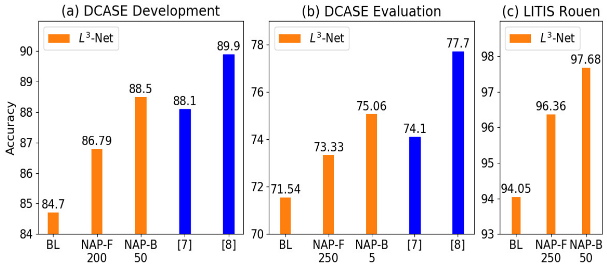


Figure 7: *Classification results after suppression. The subplots (a) and (b) gives results for DCASE development and evaluation dataset: baseline, after suppressing foreground, after suppressing background and results from [7], [8]. The subplot (c) gives the results for LITIS Rouen dataset: baseline, after suppressing the foreground, after suppressing the background. The number under NAP specifies number of basis components.*

On the DCASE evaluation dataset, we observe 1.79% and 3.52% increase in classification accuracy from the baseline when foreground and background is treated as nuisance attributes respectively. The number of basis vectors used are 250 for the foreground and 5 for the background. For the evaluation dataset, we treated the entire development dataset as training data to find the class-specific basis, as well as to train the attention model and the SVM. Figure 7 also shows that the proposed method is comparable to the results in [7] and [8], which are more complex ensemble-based methods, and are among the top performers in the DCASE 2017 challenge. The proposed method uses a relatively simple linear SVM as the classifier. The suppression of the foreground and the background provides better discrimination of confusing events across various classes. The further combination of foreground suppression and background suppression did not result in tangible benefits.

On the LITIS Rouen dataset, when 80% of data is used for the training and 20% for the testing, we observe 2.31% and 3.63% increase in classification accuracy from the baseline when foreground and background is treated as nuisance attributes respectively. The number of basis vectors used are 250 and 50 for the foreground and the background respectively. The relative increase in classification accuracy from the baseline model is similar with both the datasets.

6. Conclusion

In this paper, we showed that suppressing a part of the foreground or background is helpful in classifying acoustic soundscapes. The framework of RPCA helps in determining the subspaces, the projection into which, makes the suppression possible. By using meta-embeddings derived from an attention mechanism, class-specific projections can be utilized for effective suppression, which in turn helps in better discrimination of the soundscapes. Our experiments show that the both the background and the foreground has useful information for this purpose.

Acknowledgement: This work was partially supported by the DRDO grant NRB-459/SSB/19-20.

7. References

[1] N. Vaswani, T. Bouwmans, S. Javed, and P. Narayanamurthy, “Robust subspace learning: Robust PCA, robust subspace tracking, and robust subspace recovery,” *IEEE Signal Processing Magazine*, vol. 35, no. 4, pp. 32–55, 2018.

[2] E. J. Candès, X. Li, Y. Ma, and J. Wright, “Robust principal component analysis?” *Journal of the ACM (JACM)*, vol. 58, no. 3, pp. 1–37, 2011.

[3] A. Solomonoff, C. Quillen, and W. M. Campbell, “Channel compensation for SVM speaker recognition.” in *Odyssey*, vol. 4. Citeseer, 2004, pp. 219–226.

[4] Q. Kong, C. Yu, Y. Xu, T. Iqbal, W. Wang, and M. D. Plumbley, “Weakly labelled audioset tagging with attention neural networks,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 11, pp. 1791–1802, 2019.

[5] D. Kiela, C. Wang, and K. Cho, “Dynamic meta-embeddings for improved sentence representations,” *arXiv preprint arXiv:1804.07983*, 2018.

[6] A. Mesaros, T. Heittola, A. Diment, B. Elizalde, A. Shah, E. Vincent, B. Raj, and T. Virtanen, “DCASE 2017 challenge setup: Tasks, datasets and baseline system,” 2017.

[7] R. Hyder, S. Ghaffarzadegan, Z. Feng, and T. Hasan, “BUET bosch consortium (B2C) acoustic scene classification systems for DCASE 2017,” *IEEE AASP Challenge on DCASE 2017 technical reports*, 2017.

[8] Z. Weiping, Y. Jiantao, X. Xiaotao, L. Xiangtao, and P. Shaohu, “Acoustic scene classification using deep convolutional neural network and multiple spectrograms fusion,” *Detection and Classification of Acoustic Scenes and Events (DCASE)*, 2017.

[9] Y. Han, J. Park, and K. Lee, “Convolutional neural networks with binaural representations and background subtraction for acoustic scene classification,” *the Detection and Classification of Acoustic Scenes and Events (DCASE)*, pp. 1–5, 2017.

[10] Z. Ren, K. Qian, Y. Wang, Z. Zhang, V. Pandit, A. Baird, and B. Schuller, “Deep scalogram representations for acoustic scene classification,” *IEEE/CAA Journal of Automatica Sinica*, vol. 5, no. 3, pp. 662–669, 2018.

[11] T. Lidy and A. Schindler, “CQT-based convolutional neural networks for audio scene classification,” in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE2016)*, vol. 90, 2016, pp. 1032–1048.

[12] M. Valenti, S. Squartini, A. Diment, G. Parascandolo, and T. Virtanen, “A convolutional neural network approach for acoustic scene classification,” in *2017 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2017, pp. 1547–1554.

[13] S. Mun, S. Shon, W. Kim, D. K. Han, and H. Ko, “A novel discriminative feature extraction for acoustic scene classification using RNN based source separation,” *IEICE Transactions on Information and Systems*, vol. 100, no. 12, pp. 3041–3044, 2017.

[14] P. Huang, S. D. Chen, P. Smaragdis, and M. Hasegawa-Johnson, “Singing-voice separation from monaural recordings using robust principal component analysis,” in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2012, pp. 57–60.

[15] B. Sankaran, H. Mi, Y. Al-Onaizan, and A. Ittycheriah, “Temporal attention model for neural machine translation,” *arXiv preprint arXiv:1608.02927*, 2016.

[16] T. Luong, H. Pham, and C. D. Manning, “Effective approaches to attention-based neural machine translation,” *arXiv preprint arXiv:1508.04025*, 2015.

[17] J. Cramer, H. Wu, J. Salamon, and J. P. Bello, “Look, listen, and learn more: Design choices for deep audio embeddings,” in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 3852–3856.

[18] R. Arandjelovic and A. Zisserman, “Look, listen and learn,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 609–617.

[19] 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, 2014.