Neurocomputing 460 (2021) 409–423

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Environment sound classification using an attention-based residual neural network

Achyut Mani Tripathi *, Aakansha Mishra

Department of Computer Science and Engineering, Indian Institute of Technology, Guwahati, Assam 781039, India

ARTICLE INFO

Article history:
Received 2 December 2020
Revised 2 May 2021
Accepted 12 June 2021
Available online 17 June 2021
Communicated by Zidong Wang

Keywords:
Attention mechanism
Convolutional neural network
Explainable
Environmental sound classification
Residual network

ABSTRACT

Complexity of environmental sounds impose numerous challenges for their classification. The performance of Environmental Sound Classification (ESC) depends greatly on how good the feature extraction technique employed to extract generic and prototypical features from a sound is. The presence of silent and semantically irrelevant frames is ubiquitous during the classification of environmental sounds. To deal with such issues that persist in environmental sound classification, we introduce a novel attention-based deep model that supports focusing on semantically relevant frames. The proposed attention guided deep model efficiently learns spatio-temporal relationships that exist in the spectrogram of a signal. The efficacy of the proposed method is evaluated on two widely used Environmental Sound Classification datasets: ESC-10 and DCASE 2019 Task-1(A) datasets. The experiments performed and their results demonstrate that the proposed method yields comparable performance to state-of-the-art techniques. We obtained improvements of 11.50% and 19.50% in accuracy as compared to the accuracy of the baseline models of the ESC-10 and DCASE 2019 Task-1(A) datasets respectively. To support the attention outcomes that have focused on relevant regions, visual analysis of the attention feature map has also been presented. The resultant attention feature map conveys that the model focuses only on the spectrogram's semantically relevant regions while skipping the irrelevant regions.

1. Introduction

Identifying environmental sound is an essential step in developing intelligent applications for research domains like surveillance systems, robotics and autonomous self-driving vehicles etc. The lack of availability of prior knowledge related to time and frequency characteristics of environmental sound makes Environmental Sound Classification (ESC) more challenging than speech and music. The presence of variation in environmental sound's time-frequency characteristics is one of the significant reasons behind the failure of various automatic speech recognition techniques to classify environmental sound. To overcome the aforementioned problem, various attempts have been made that focus on ESC. Initial works that try to perform ESC focus on extraction of Zero-Crossing Rate (ZCR) features [1], Wavelet Features [2], and Mel-frequency cepstral coefficients (MFCC) features [3] from a raw signal have been published over time. Advancements in signal processing techniques have encouraged researchers to employ dictionary learning [4], Matrix Factorization (MF) [5,6] based techniques to perform ESC. The features extracted using the methods mentioned above are further used to train machine learning models such as Support Vector Machine (SVM) [4,7], Random Forest (RF) classifier [7] and Gaussian Mixture Model (GMM) [7,8]. It is also worth mentioning that the extraction of the MFCC, Wavelet and ZCR features is a tedious process and increases the time complexity of the model while performing ESC.

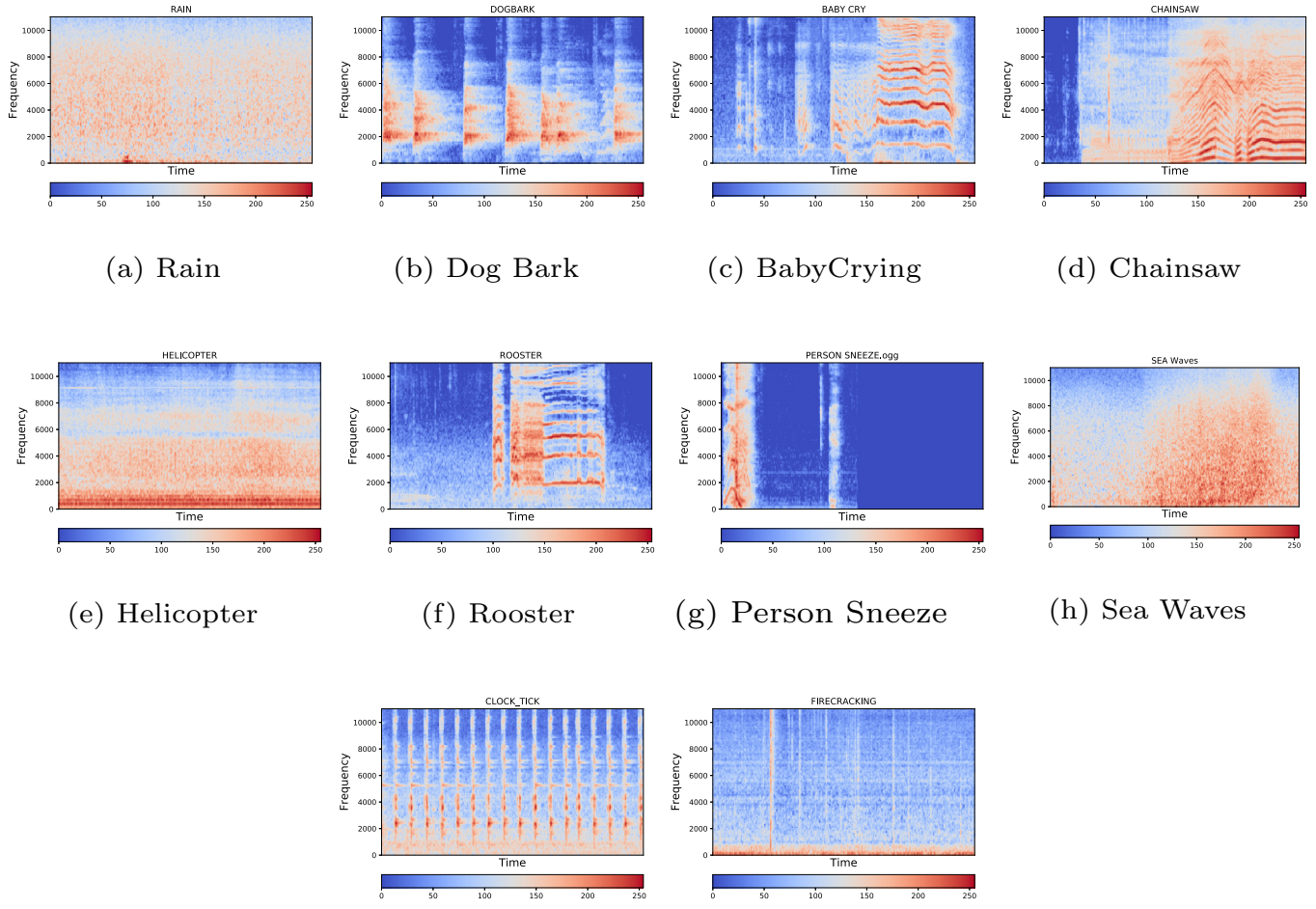
Recent years have witnessed a remarkable success of deep learning models in playing a vital role in solving a variety of complex challenges in multiple domains - computer vision [9], text mining [10], time series classification [11], robotics [12] and sound classification [13], to name a few. The adaptation of deep learning models in a sound classification system helps in generating human-like listening capacity in machines. The growing popularity of AMAZON ALEXA, APPLE SIRI, MICROSOFT CORTANA and GOOGLE ASSISTANT are among some of the suitable examples that show great strength of deep learning techniques in sound classification problems. In recent years, deep learning techniques have been well applied to extract highly discriminative features from an acoustic signal to perform ESC. The ability to extract meaningful features and still retain a good generalization capacity on unseen

* Corresponding author.
E-mail addresses: t.achyut@iitg.ac.in (A.M. Tripathi), ak.kkb@iitg.ac.in (A. Mishra).

https://doi.org/10.1016/j.neucom.2021.06.031
0925-2312/

sound data are the two factors that make deep learning techniques the preferred choice for ESC. As an initial work, in [14], the author applied a deep belief network to extract features from a spectrum of sound and perform ESC. Piczak et al. [13] processed a log scaled Mel spectrogram of a sound using 2-D CNN to perform ESC and reported promising results on multiple sound datasets. The sound data displays sequential behavior, and based on this property, the author of [15] employed recurrent neural network (RNN) to perform ESC. The techniques mentioned above cannot deal with significant variations in spectrogram representations of environmental sounds. Sounds that belong to the same class exhibit a considerable variation in spectrogram representations. Moreover, the sound pattern can be continuous, irregular, transient and can contain several noisy or silent frames, or consist of small semantically relevant frames. Fig. (1) shows the spectrograms of ten different signals containing the characteristics mentioned above. For example, Fig. (1a), Fig. (1b), Fig. (1c) show continuous, temporal and irregular characteristics, respectively. Fig. (1g) shows presence of silent frames (semantically irrelevant frames). These characteristics impose several challenges while performing ESC and degrade the performance of the model. Attention mechanism has shown promising results in solving numerous problems related to speech processing, image classification, and text mining. A few notable attention-based works that have been proposed in the domain of sound classification are [16–21]. Inspired by the success of attention mechanism and its ability to overcome the problems mentioned above, we have proposed a novel attention mechanism that resolves intra-class inconsistency and is efficient in selecting the semantically relevant frames from the spectrogram features. To the best of our knowledge, none of the existing methods for ESC address the issue of intra-class inconsistency that degrades the performance of deep models. The proposed model is evaluated on two benchmarks - ESC-10 [7] and DCASE 2019 Task-1(A) [22] datasets. The experiments and results show that the proposed method yields comparable performance to state-of-the-art methods. Additionally, the proposed model also explains why the model classifies a signal into a particular class. We also provided a visualization of the spectrogram that explains which part of the spectrogram were found to be more relevant by the model for classification. Summary of our major contribution can be listed as follows:

- We have proposed a novel attention module that resolves the intra-class inconsistency problem of ESC. The attention module is combined with the residual network to perform ESC and could efficiently capture long-range contextual information between the spectrogram's local features and provides more useful representative features for ESC.
- Our proposed model is explainable and can efficiently identify more semantically relevant parts of the spectrogram and correctly highlight them while providing the visual description.



(a) Rain (b) Dog Bark (c) BabyCrying (d) Chainsaw

(e) Helicopter (f) Rooster (g) Person Sneeze (h) Sea Waves

(i) Clock Tick (j) Fire cracking

Fig. 1. Spectrogram of Different Environmental Sounds of ESC-10 Dataset [7].

- The accuracy obtained by the proposed model for the ESC-10 and DCASE 2019 Task-1(A) datasets are comparable with state-of-the-art methods.

2. Related work

This section presents related works that use deep learning-based models to perform environmental sound classification. Spectrogram features are among the most popular choice for deep learning models due to the 2-D representation of features, which can be easily processed by deep learning models such as a Convolutional Neural Network (CNN). The pioneer work on ESC was performed by Piczak et al. [13]. The author proposed a 2-D CNN that takes spectrogram feature as input and performs ESC. The method proposed by Piczak et al. reported that the accuracy of the PiczakCNN shows improvement in classification accuracy by a significant margin as compared to the accuracy obtained from Support Vector Machine (SVM), Random Forest classifier (RF), and K-Nearest Neighbors (KNN) classifier. Inspired by the success of PiczakCNN, numerous researchers have employed a combination of spectrogram features and pre-trained networks such as GoogleNet [23] and AlexNet [23] to perform ESC. In [24], the author trained a model on sound recordings of multiple videos and the learned model was fine-tuned to perform sound classification. Various methods have been proposed that use a 1-D raw signal as input to perform ESC. In [25], the author used 1-D CNN to extract meaningful features from a raw signal and displayed comparative accuracy to the model trained on the spectrogram features. However, the model suffers from a high computational complexity as it requires around 34 layers of Convolutional Neural Network. In [26], the author employed 1-D CNN layers to create a 2-D feature map from the input signal, and the learned 2-D feature map showed competitive performance against the model trained on spectrogram feature. In recent years, attention mechanism-based deep models have been well utilized for learning representative features. Xu et al. [27] proposed an encoder-decoder-based attention mechanism to compute attention weights for a given image and a corresponding question. The encoder contains a self-attention layer followed by a feedforward neural network. The decoder module contains both encoder-decoder and self-attention layers followed by a feed-forward layer. This attention mechanism demanded proper alignment of Query and Key matrices, which increased computational complexity of the model as it learned the attention weights. These limitations of the attention mechanism proposed by Xu et al. [27] make it less preferable choice for ESC, while it remains more relevant to Visual Question Answering (VQA) and image captioning tasks. Vaswani et al. [28] have presented two major types of attention mechanisms, i.e. scaled dot product attention mechanism and multi-head attention mechanism. The scaled dot product attention mechanism presented in the above mentioned paper depends on Query content, that is used during the computation of weights for a given Key. The term 'Multi-head' indicates the number of times attention is applied using the linearly transformed values of the Key, Query and Value. The working mechanism of multi-head attention mechanism is similar to scaled dot product attention mechanism but with an additional step of linear transformation applied over input matrices of Key, Query and Value. The performance of a model that uses attention mechanism depends heavily on the method used to learn the attention weights. Fan et al. [29] proposed a recurrent attention mechanism that uses an attention processor that acts as a bridge between encoder and decoder frameworks of the model that is intended to design reinforced generator for visual dialog. The attention processor takes input from encoder and generates an output that is forwarded to the decoder. The recurrent attention mechanism involves a memory element same as the one used in Long Short-Term Memory (LSTM) to memorize the attention weights. This attention mechanism uses content of the Key while computing the attention weights. Moreover, the computational complexity for the recurrent attention mechanism is high due to the incorporation of LSTM for the computation of the attention weights. Cai et al. [30] proposed a triplet attention mechanism-based deep network for hyper-spectral image classification. Zhang et al. [31] proposed a residual channel attention mechanism to perform image super-resolution. Zhu et al. [32] proposed a two-layered Key memory-based attention framework for few shot video classification. The attention mechanism used in [32] is similar to approach proposed in [29] and does not consider relative positions of the Query and Key during the computation of the attention weights. This approach needs a complex two-layered key memory architecture for the computation of relevant information from video frames. Wu et al. [33] proposed a dual attention-based deep model for audio-visual event localization. The attention mechanism takes input from two different modalities. The self-attention technique incorporated in [33] considers both Key and Query contents in the computation of attention weights. Our proposed attention mechanism is motivated by the remarkable success of self-attention mechanism across multiple domains. The working of the self-attention mechanism involves three main parameters, viz. Key (K), Query (Q), and Value (V). Computation of attention weights for a given key (K), corresponding to the available query (Q), depends on the following four factors:

1. Query (Q) Content
2. Query (Q) Content and Relative Position of Query (Q)
3. Key (K) Content
4. Key (K) Content and Relative Position of Key (K)

Previous works that use the attention mechanism can be classified into four categories based on the factors mentioned above. The attention techniques [27–33] that fall under the first or third categories exploit only global dependencies while learning the attention weights. Our method falls under a combination of the second and the fourth category. It uses content of Q, K and the relative position of Q and K during the computation of attention weights. The proposed attention mechanism exploits both local features (features that provide information regarding energy distribution along multiple frequency bands) and global features (temporal dependency among all frames) between multiple frames of the spectrogram while learning the attention weights and efficiently identifying salient key elements. Additionally, a comparison in terms of accuracy and computational complexity has been performed between several attention mechanism-based deep models and the our proposed model for ESC in the experimental section.

The application of an attention mechanism for sound classification is an underexplored field of study. Few researchers have investigated the application of attention mechanism in combination with deep models for sound classification. Guo et al. [34] proposed an attention-based framework that combines CNN and LSTM models in a stacked manner to perform ESC. The attention mechanism is applied to the LSTM layers and the weighted sum of LSTM layers is used to get attention features. Wang et al. [35] proposed a deep model that uses a self-attention mechanism to perform acoustic scene classification (ASC). Wang et al. [36] proposed a multi-channel CNN that uses temporal attention mechanism to perform ESC. Helin et al. [16] presented a parallel spatio-temporal attention mechanism to extract a representative part of the input spectrogram to perform ESC. Unlike previous works that address attention-based deep models to perform the ESC, we rescaled the values of spectrogram features between 0 to 255 to get an image-like feature from the log Mel spectrogram. The spectrogram

feature extracted from environmental sound showed variation in between spectrogram features extracted from the different signals belonging to same class. This situation results in intra-class inconsistency that may cause degradation in the performance of the model. Inspired by the success of the self-attention mechanism, we designed a novel attention mechanism that addresses intra-class inconsistency and captures long-range contextual information from the spectrogram image. The proposed attention module also effectively captures the semantically relevant parts from the spectrogram.

3. Proposed method

This section provides details of architecture of residual network and proposed attention mechanism.

3.1. Architecture of the proposed network

Description of the architecture of the residual network is as follows:

- **Architecture of a Residual Network** The proposed model's architecture is motivated by the remarkable success of the Residual Network (ResNet) to perform image classification. Table 1 shows a description of different layers of residual network. The model uses skip connection to overcome the problem of performance degradation caused by vanishing gradient. The output of one layer is allowed to skip one or more layers and is added to the layer's output. The skip connection's significant advantage is that it makes the model partially shallow, as a result, the output of present layer depends on all the previous layers. Fig. (2) provides a visualization of skip connection.
  The model consists of four two-layered building blocks with a kernel size of (3∗3) and number of channels – 64, 128, 256 and 512 respectively. Each building-block contains a skip connection between its input and output. The network's input size (Height,Width,Channels) is equal to the size of the extracted spectrogram image feature, which is (128,431,1) for ESC-10 signals and (128,862,1) for DCASE 2019 Task-1(A) signals. The attention layer is added after the fourth building block. The attention layer's output is forwarded to the global average pooling layer followed by a fully connected layer and a Softmax layer.
- **Attention Module** Fig. (3) demonstrates the architecture of attention module incorporated for our proposed model. The main steps to compute an attention feature map are as follows:
  1. Initially, an input feature map A_1 of size (H∗W∗C) is passed through a convolutional layer to obtain three feature map A_2, A_3 and A_4 each of size (H∗W∗C). Here $(*)$ sign denotes a multiplication operator.

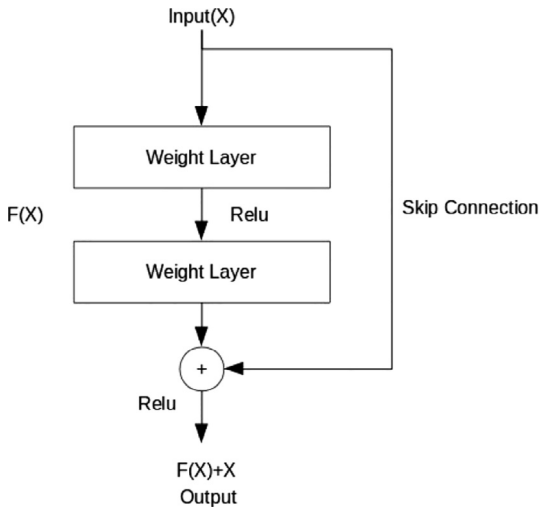


Fig. 2. Skip Connection in Residual Network.

2. Reshape the A_2, A_3 and A_4 to size (C∗N) and multiply transpose of A_2 with A_3. Here N=$(H * W)$.
3. Apply the softmax function to compute attention map A_5 using the Eq. (1).

$$A_5(k,j) = \frac{exp(A_2(k) * A_3(j))}{\sum_{k=1}^{N} exp(A_2(k) * A_3(j))} \tag{1}$$

Here k and j are k^{th} and j^{th} spatial locations and $1 \leqslant k,j \leqslant N$. Here *exp* denotes an exponential function.

4. Reshape the feature map A_5 to size (C∗N) and multiply A_5 with A_4.
5. Finally, use the Eq. (2) to obtain the final attention feature map A_6.

$$A_6(j) = \sum_{k=1}^{N} (\alpha * A_5(j,k)) + A_1(j) \tag{2}$$

Step-2 of the above attention mechanism encodes spatial information between any two locations of the spectrogram image and Step-5 encodes the long range contextual dependency between the local features to identify the semantically relevant frames present in the spectrogram image feature representation. It is apparent from Eq. (2) that the value of each location of the final attention map A6 is a weighted sum of all locations of A5 and initial feature map A1(an original feature map). Thus the resultant feature map A6 efficiently deals with the intra-class inconsistency and improves compactness.

Table 1
Description of Different Layers of Residual Network Used in The Proposed Method.

Layer Name	Description
Input Layer	(H = Height, W = Width, C = Channels)
Convolutional Layer-1	Kernel Size = 7∗7, Stride = 2, Channel = 64
Max Pooling Layer	Kernel size = 3∗3, Stride = 2
[Convolutional Layer-2]∗2	Kernel Size = 3∗3, Channel = 64 Kernel Size = 3∗3, Channel = 64
[Convolutional Layer-3]∗2	Kernel Size = 3∗3,Stride = 2, Channel = 128 kernel Size = 3∗3, Channel = 128
[Convolutional Layer-4]∗2	kernel Size = 3∗3, Stride = 2, Channel = 256 kernel Size = 3∗3, Channel = 256
[Convolutional Layer-5]∗2	kernel Size = 3∗3, Stride = 2, Channel = 512 Kernel Size = 3∗3, Channel = 512
Average Pooling Layer	Kernel Size = 7∗7
Fully Connected Layer	
Softmax Layer	

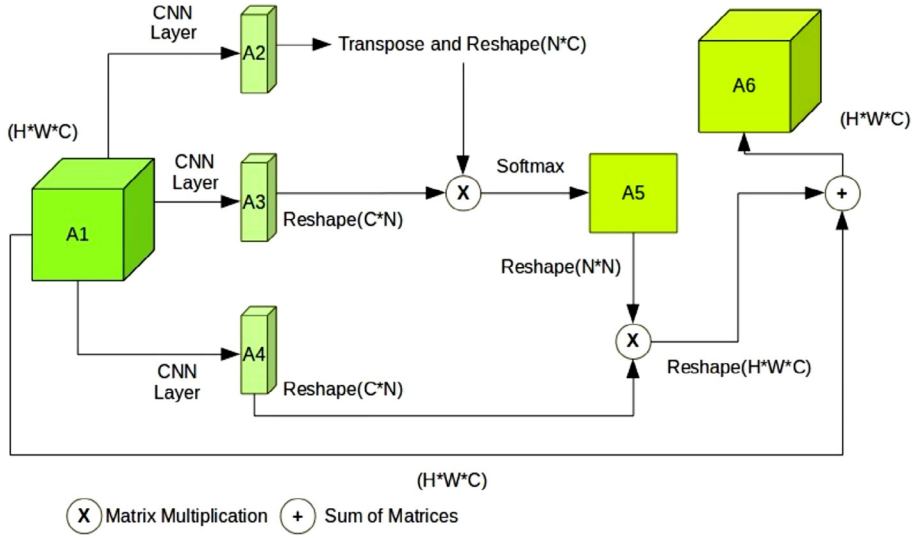


Fig. 3. Proposed Attention Module.

4. Experiments and results

This section presents details of the experiments and results.

4.1. Dataset description

The efficacy of the proposed method is demonstrated on two benchmark sound classification datasets. The details of the two datasets are as follows:

- **ESC-10 Dataset:** The ESC-10 dataset [7] contains sounds that are recorded in numerous outdoor and indoor environments. The dataset consists of total of 400 recordings. Each recording is of 5 s duration. Sampling rate of each recording is 44100 Hz. The recordings are divided into ten major categories. Table 2 shows specifications of different classes and the class-wise distribution of recordings present in the ESC-10 dataset. The proposed model is evaluated for 5-fold cross validation setup provided in the ESC-10 dataset. The mean accuracy of the five folds is reported as the final accuracy.
- **DCASE 2019 Task-1(A) Development Dataset:** The DCASE challenge 2019 dataset [22] for Task-1(A) was developed to identify different acoustic scenes. The dataset contains recordings of acoustic signals corresponding to ten different acoustic scenes. All the signals were recorded in twelve different European cities and in different environments. The development dataset contains total of 13370 audio recordings, each with a duration of 10s. A pre-defined fold is provided within the dataset. The training and test data contain a total of 9185 and 4185 recordings respectively. We used the provided fold-set in our experiments. Table 3 shows different acoustic scenes present in the DCASE 2019 Task-1(A) dataset. Sampling rate of each recording is 48000 Hz. The recordings are binaural i.e. contains two channels; left and right channels respectively.

Table 2
Different Classes of The ESC-10 Dataset.

S.No.	Classes	No. of Recordings	Channel Type
1	Dog Barking	40	Single
2	Crackling Fire	40	Single
3	Baby Crying	40	Single
4	Rain	40	Single
5	Person Sneezing	40	Single
6	Rooster	40	Single
7	Sea Waves	40	Single
8	Helicopter	40	Single
9	Chainsaw	40	Single
10	Clock Tick	40	Single

Table 3
Different Acoustic Scenes in The DCASE 2019 Task-1(A) Dataset.

S.No.	Scenes	Channel Type
1	Airport	Binaural
2	Bus	Binaural
3	Metro	Binaural
4	Metro Station	Binaural
5	Park	Binaural
6	Shopping Mall	Binaural
7	Public Square	Binaural
8	Street Traffic	Binaural
9	Street Pedestrian	Binaural
10	Tram	Binaural

4.2. Creation of spectrogram image feature

We extracted Log Mel spectrogram features from the given acoustic signals. Number of Fast Fourier Transformation (FFT) is set as 2048 and hop length is set to 512. This spectrogram matrix is further converted to spectrogram image features having values between 0 to 255. The number of Mel-filter is set to 128. In this study we have used the Librosa library [37] available in python language to extract the spectrogram features. The size of spectrogram feature for the ESC-10 dataset is $(128 * 431 * 1)$. The size of spectrogram feature for the DCASE-19 Task-1(A) dataset is $(128 * 862 * 1)$. Fig. (4) shows spectrogram image feature for Baby Crying signal from the ESC-10 dataset.

4.3. Implementation details

All the models were developed and run on a system with 16 GB RAM and a NVIDIA GM107M GPU. The proposed method was developed using an open source Pytorch 1.4 library running on Ubuntu 16.04 LTS operating system. The Adam optimizer [38] was used for optimization. The learning rate was set to 0.0002 and the batch size was set to 32.

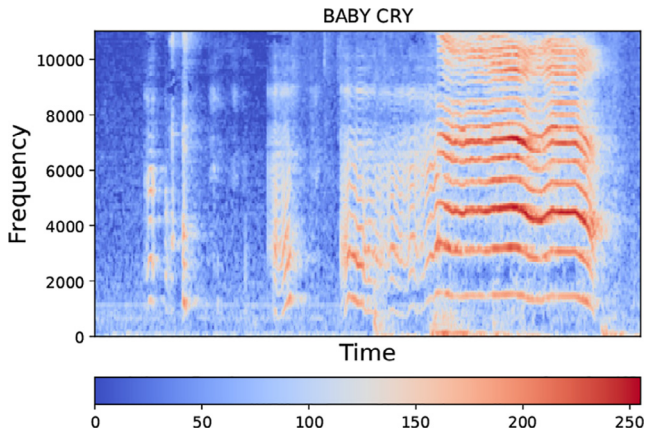


Fig. 4. Spectrogram Image for Baby Crying Signal of The ESC-10 Dataset.

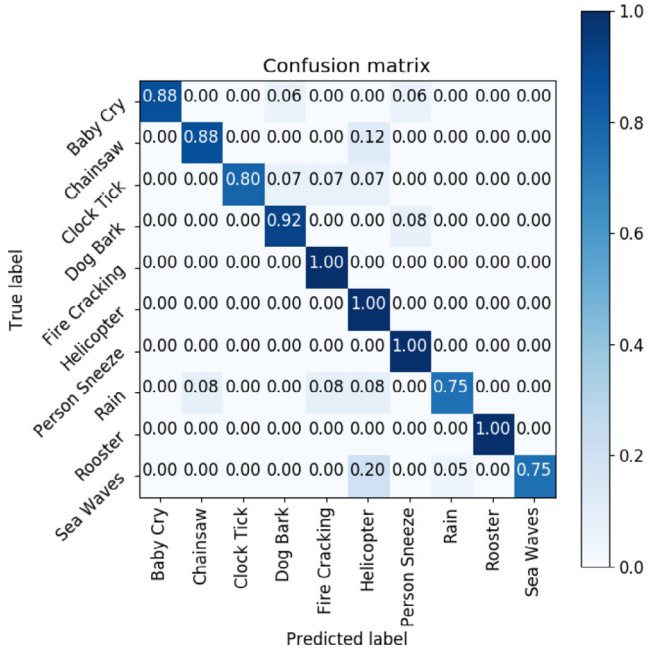


Fig. 5. Confusion Matrix for ESC-10.

4.4. Experimental results

The proposed model yields accuracy of 92% and 82% for the ESC-10 and DCASE 2019 Task-1(A) datasets.

4.4.1. Performance for ESC-10

Fig. (5) shows a confusion matrix obtained for the ESC-10 dataset. Following observations are obtained from the confusion matrix of the ESC-10.

1. The model shows a classification accuracy of 100% when classifying the signals belonging to the Fire Cracking, Helicopter, Rooster and Person Sneezing classes. 20% of the signals belonging to Sea waves class are misclassified into the Helicopter class.
2. Signals belonging to Rain class are misclassified into Chainsaw, Fire Cracking and Helicopter classes with a misclassification ratio of 8% in each class.
3. Precision, recall and F1-score of the model for the ESC-10 dataset are 88.70%, 89.80% and 97.93%, respectively. The first row of Table 5 shows precision, recall and F1-score for the ESC-10 dataset.
4. Table 4 shows a comparison of the model's accuracy with existing state-of-the-art methods for environmental sound classification. The model shows improvement of 11.50% when compared to the accuracy obtained by the baseline model, i.e. PiczakCNN [13]. The model's accuracy degrades by 4.69% when the model is trained and tested without the attention module. The performance of the proposed attention-based deep model is compared with other attention mechanisms from literature. The model attains accuracy of 90.11%, 90.38%, 89.17%, 91.59% and 94.07% for the Scaled dot product attention [28], Triplet attention [30], Channel attention [31], Dropout attention [39]

Table 5
Precision, Recall, F1-Score and Accuracy of The Proposed Model for The ESC-10 and DCASE 2019 Task-1(A) Datasets.

Dataset	Precision (%)	Recall (%)	F1-Score (%)	Accuracy (%)
ESC-10	88.70	89.80	87.93	92.00
DCASE 2019 Task-1(A)	83.47	82.28	82.39	82.00

Table 4
Comparison of Accuracy While Classifying ESC-10 Sounds With State-of-The-Art Methods.

Model	Accuracy(%)	Model and Type of Feature	Gain(%)
Bodapati et al. [23]	86.00	CNN + Spectrogram	6
Ahmad et al. [41]	87.25	ELM + MFCC + ZCR	4.75
Su et al. [42]	72.00	CNN + MFCC + Chroma Features	20
Salamon et al. [43]	72.00	CNN + Spectrogram	20
KNN [7]	66.70	K Nearest Neighbours +MFCC	25.30
SVM [7]	67.50	Multi-class SVM +MFCC	24.50
Random Forest [7]	72.70	Random Forest Classifier +MFCC	19.30
AlexNet [23]	78.40	Pre-Trained Alexnet + Spectrogram	13.60
PiczakCNN [13]	80.50	CNN + Spectrogram	11.50
GoogleNet [23]	63.20	Pre-Trained GoogleNet + Spectrogram	28.80
Tokoz et al. [26]	91.30	EnvNetV-2	0.70
Zhang et al. [10]	91.70	CNN + Spectrogram	0.30
Erhan et al.[44]	90.25	LBP + MFCC + SVM	1.75
Vaswani et al. [28]	90.11	Residual Network + Scalar Dot Product Attention	1.89
Cai et al. [30]	90.38	Residual Network + Triplet Attention	1.62
Wang et al. [40]	94.07	Symbiotic Attention + Spectrogram	−2.07
Zhang et al. [31]	89.17	Channel Attention + Spectrogram	2.83
Liu et al. [39]	91.59	Attention Dropout + Spectrogram	0.41
Proposed (Without Attention)	87.31	Residual Network + Spectrogram	4.69
Proposed	92.00	Residual Network + Attention + Spectrogram	-

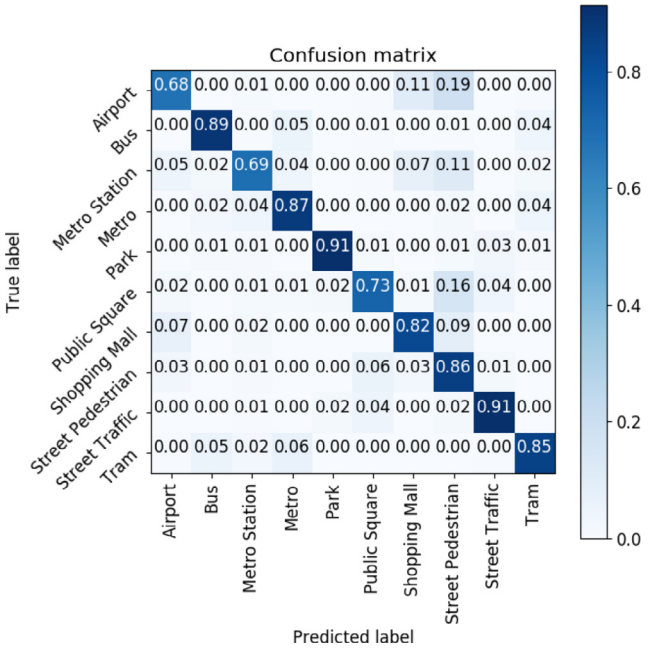


Fig. 6. Confusion Matrix for DCASE 2019 Task-1(A).

and Symbiotic attention [40] techniques respectively. It is clear from Table 4 that the accuracy of the proposed model is higher and comparable to the accuracy of the state-of-the-art methods.

4.4.2. Performance for DCASE 2019 Task-1(A)

Fig. (6) shows a confusion matrix obtained for the DCASE 2019 Task-1(A) dataset. Following observations can be listed from the confusion matrix illustrated in the Fig. (6):

1. 19% of signals belong to Airport class are misclassified into Street pedestrian class. 11% of signals from Metro Station class are misclassified into the Street Pedestrian class.
2. 16% of signals from Public Square class are misclassified into the Street pedestrian class and 4% of signals from Street Traffic class are misclassified into Public Square class. 6% of signals from Tram class are misclassified into Metro class.
3. The classification accuracy for the Airport, Bus, Metro Station, Metro, Park, Public Square, Shopping mall, Street pedestrian, Street Traffic and Tram classes are 68%, 89%, 69%, 87%, 91%, 73%, 82%, 86%, 91% and 85% respectively.
4. Precision, recall and F1-Score for the proposed model are 83.47%, 82.28% and 82.39%, respectively.
5. Table 6 shows a comparison of the model's accuracy with state-of-the-art techniques. The model shows 11.40% increase in accuracy compared to the parallel attention-based deep model proposed by Wang et al. [16]. The model shows an 8.10% improvement in accuracy compared the attention pooling-based deep model proposed in [19]. Jung et al.[45] performed knowledge distillation with specialist deep models and the accuracy of the proposed model is 0.80% higher than the specialist deep model. The proposed model shows an accuracy of 67.48% when trained without the attention module and shows a significant improvement of 14.52% when trained along with

Table 6
Comparison of Accuracy While Classifying DCASE 2019 Task-1(A) Sounds With State-Of-The-Art Methods.

Model	Accuracy(%)	Model and Type of Feature	Gain (%)
Baseline DCASE 2019 Task-1(A) [22]	62.50	CNN + Spectrogram	19.50
Ding et al. [46]	63.20	Ensemble Models + MFCC + ZCR	18.80
Waldekar et al. [47]	67.40	SVM + Wavelet Features	14.60
Zhou et al. [48]	69.70	CNN + Spectrogram	12.30
Paseddula et al. [49]	70.40	DNN + SFFCC	11.60
Xinxin et al. [50]	71.10	CNN + Spectrogram	10.9
Pham et al. [51]	73.70	CNN + RNN + Spectrogram + CQT	8.30
Zhenyi et al. [19]	73.90	CNN + Attention Pooling + Spectrogram	8.10
Suh et al. [52]	74.40	CNN + Inception + Spectrogram	7.60
Wu et al. [53]	76.60	CNN + Time Frequency Features	5.40
Zeinali et al. [54]	77.00	Attentive CNN + Spectrogram	5.00
Lei et al. [55]	79.60	CNN + Multi-scale Feature	2.40
Cho et al. [56]	77.19	FCNN-Triplet loss + Spectrogram	4.81
Helin et al. [16]	70.60	CNN + Parallel Attention + Spectrogram	11.40
Zhang et al.[57]	80.34	SeNoT-Net + Spectrogram	1.66
Zhang et al. [21]	80.68	ATReSN-Net + Spectrogram	1.32
Junt et al. [45]	81.20	Knoewledge Distillation + Spectrogram	0.80
Vaswani et al. [28]	77.61	Residual Network + Scalar Dot Product Attention	4.39
Cai et al. [30]	78.27	Residual Network + Triplet Attention	3.73
Wang et al. [40]	84.71	Symbiotic Attention + Spectrogram	−2.71
Zhang et al. [31]	77.12	Channel Attention + Spectrogram	4.88
Liu et al. [39]	81.34	Attention Dropout + Spectrogram	0.66
Proposed (Without Attention)	67.48	Residual Network + Spectrogram	14.52
Proposed	82.00	Residual network + Attention + Spectrogram	–

Table 7
Comparison of Accuracy of Proposed Model When Attention Module is Placed After Different Layers.

Residual Layer	ESC-10	DCASE-19 Task-1(A)
L_0	91.19	80.84
L_1	91.23	80.91
L_2	91.54	81.27
L_3	**92.00**	**82.00**

Table 8
Comparison of Performance of The Model With Different Number of Layers.

Number of Layers in Residual Network	ESC-10 Accuracy(%)	DCASE 2019 Task-1(A) Accuracy(%)
18	92.00	82.00
34	90.02	79.56
50	89.78	78.21
152	88.65	77.54

Table 9
Comparison of Performance of The Model With Different Number of Layers When Trained Using Data Augmentation Technique.

Number of Layers in Residual Network	ESC-10 Accuracy(%)	DCASE-2019 Task-1(A) Accuracy(%)
18	92.16	82.21
34	92.07	82.13
50	92.08	82.14
152	92.10	82.17

the attention module. The performance of the proposed attention-based deep model is compared with other attention mechanism from literature. The model attains the accuracy of 77.61%, 78.27%, 77.12%, 81.34% and 84.71% for the Scaled Dot product attention [28], Triplet attention [30], Channel attention [31], Dropout attention [39] and Symbiotic attention [40] techniques respectively. It is clear from Table 6 that the proposed method yields competitive accuracy to existing state-of-the-art deep models.

4.5. Ablation study

This section presents an ablation study performed by varying the number of layers and the position of the attention module.

4.5.1. Appropriate location of attention module

In this section, we present the result of the proposed model by applying the attention module after different layers of the residual network. The base model contains a total of four residual blocks and we name these residual blocks as follows; L_0, L_1, L_2 and L_3. The accuracy of the model is evaluated by attaching the proposed attention module after a specified i^{th} residual layer. Table 7 shows the accuracy obtained for four different configurations of the attention module. For both the datasets, it is observed that the best accuracy is obtained when the attention module is affixed after the fourth residual layer. The model shows average improvement of 0.35% and 0.43% when attention module is fixed after the higher layers for the ESC-10 and DCASE 2019 Task-1(A) datasets respec-

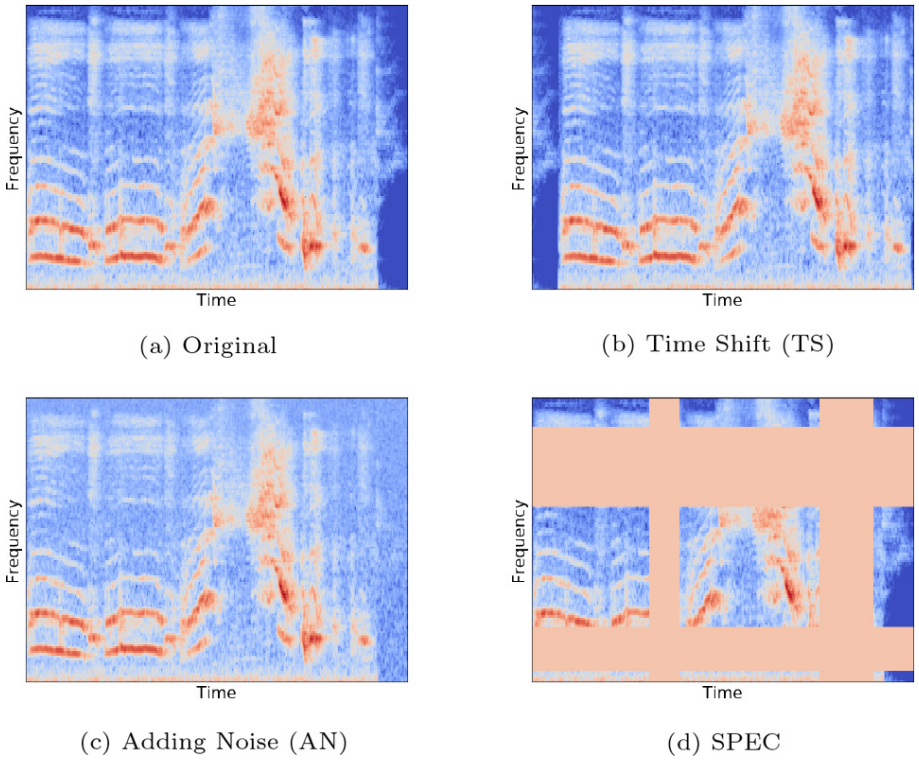
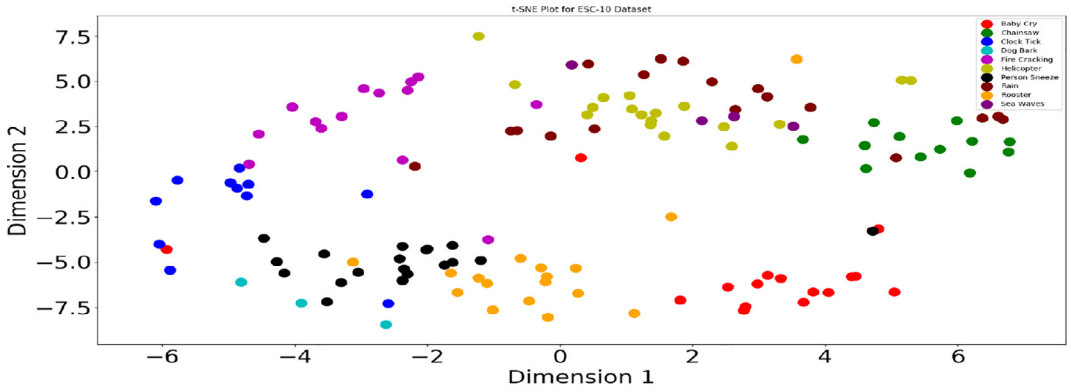
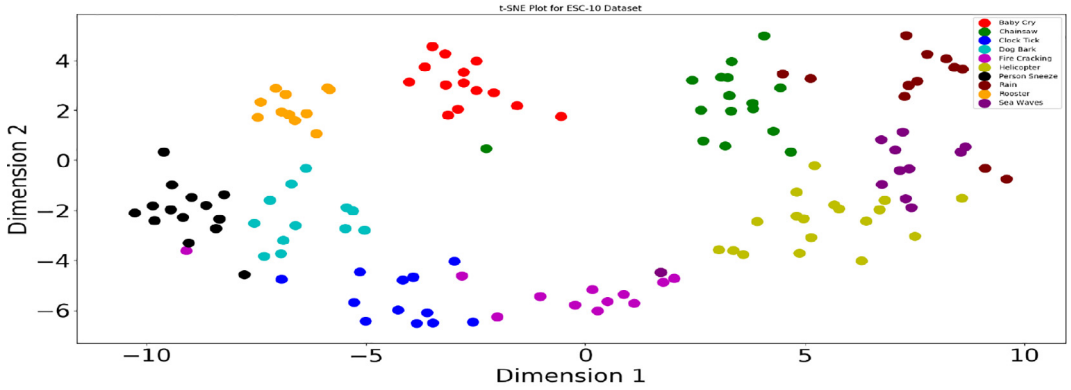


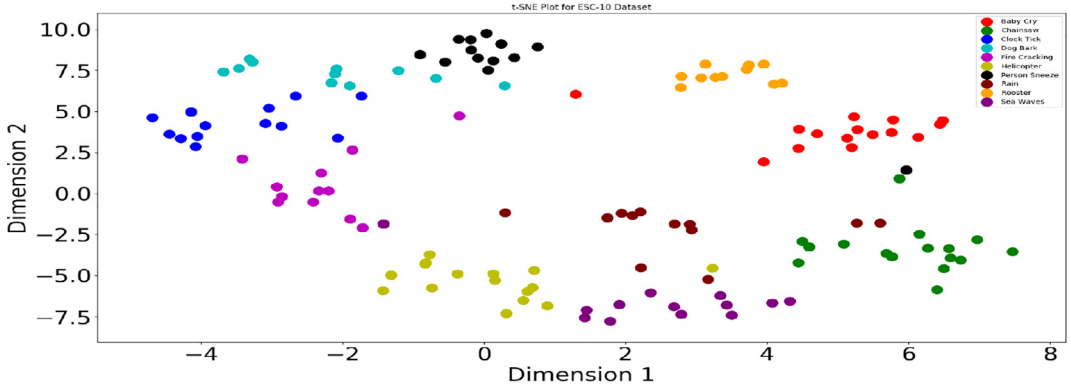
Fig. 7. Spectrogram of Dog Bark Signal of The ESC-10 Dataset After Applying Different Data Augmentations.



(a) t-SNE Visualization of Latent Features After Epoch=1



(b) t-SNE Visualization of Latent Features After Epoch=10



(c) t-SNE Visualization of Latent Features After Final Epoch

Fig. 8. t-SNE Visualization of Latent Features Features After Epoch = 1, Epoch = 10 and Final Final Epoch.

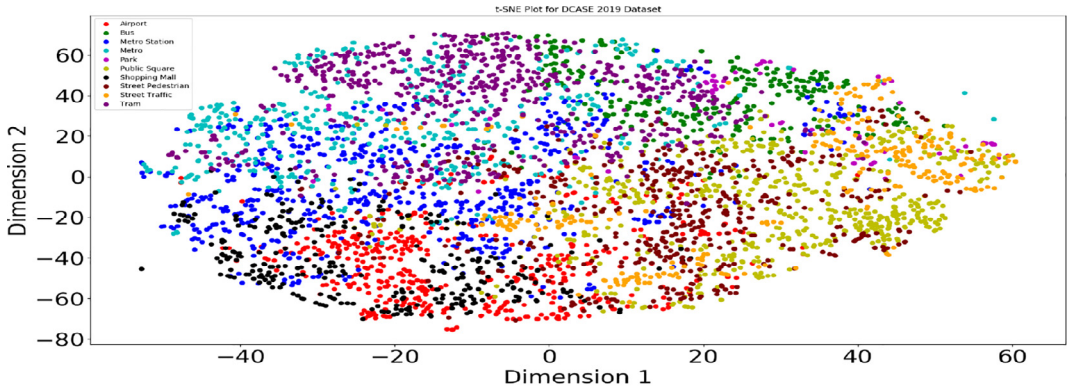
tively. From this observation, we can conclude that the higher layer provides more useful features that define the characteristics of an environmental sound and the attention module conserves them.

4.5.2. Effect of number of layers in residual network

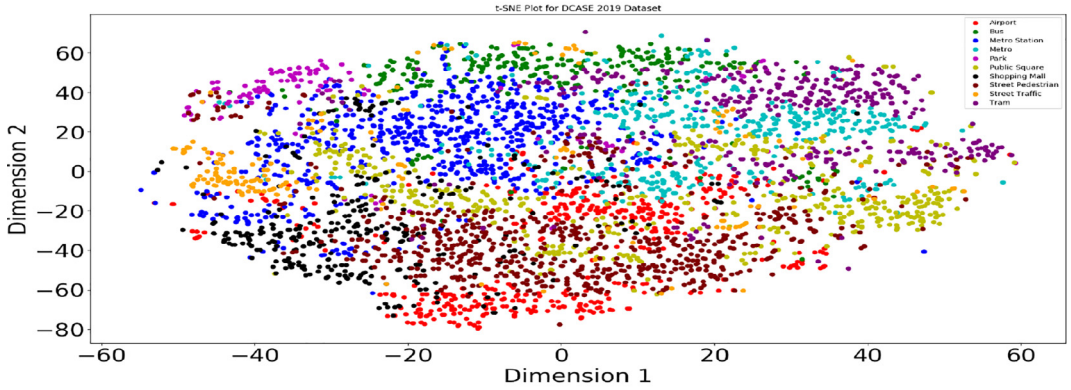
We also evaluated the performance of the proposed model by varying the number of layers in residual network. Table 8 shows accuracy obtained when number of layers changed from 18 to 152. The model with 18 layer shows highest accuracy of 92% and 82% for the ESC-10 and DCASE 2019 Task-1(A) datasets respectively. It can be seen from the Table 8 that the accuracy of the model decreases on average by 2.07% and 3.56% for the ESC 10 and DCASE 2019 datasets when the number of layers increase. Another reason behind the selection of the model with 18 layers is the reduced computational complexity of the model to perform ESC.

4.6. Performance of the model after using data augmentation

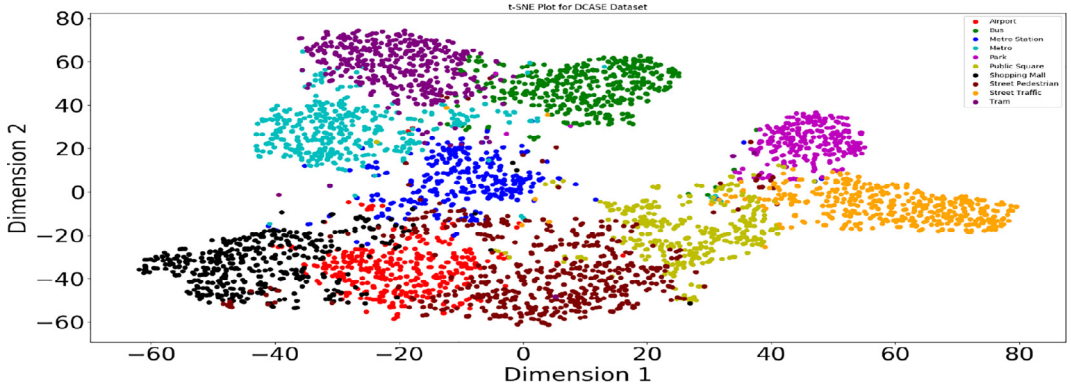
It is apparent from the Table 8 that the models with higher numbers of layers suffer from overfitting. To overcome this, we train the models by incorporating data augmentation techniques. Table 9 shows accuracy of the models on the ESC-10 and DCASE 2019 Task-1(A) datasets. Fig. (7) shows spectrograms of a signal belonging to class Dog Bark of the ESC-10 dataset after applying Time Shift (TS) [58], Adding Noise (AN) [58] and SPEC [58] data augmentation techniques.



(a) t-SNE Visualization of Latent Features After Epoch=1



(b) t-SNE Visualization of Latent Features After Epoch=10



(c) t-SNE Visualization of Latent Features After Final Epoch

Fig. 9. t-SNE Visualization of Latent Features Features After Epoch = 1, Epoch = 10 and Final Epoch for DCASE19 Task-1(A).

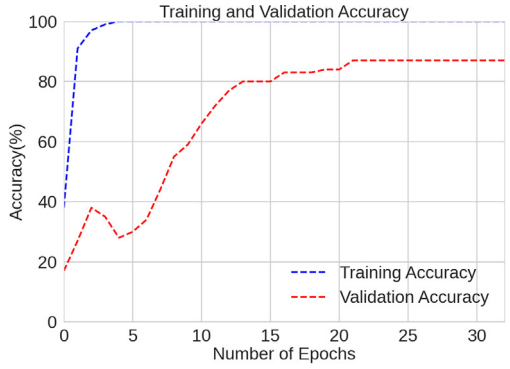
The model with 18 layer shows the highest accuracy of 92.16% and 82.21% for the ESC-10 and DCASE 2019 Task-1(A) datasets respectively. It is apparent by comparing the Table 8 and Table 9 that performance of the models with higher number of layers increases after incorporating data augmentation technique while training.

4.7. t-SNE visualization of latent embeddings for the ESC-10 and DCASE 2019 Task-1(A) datasets

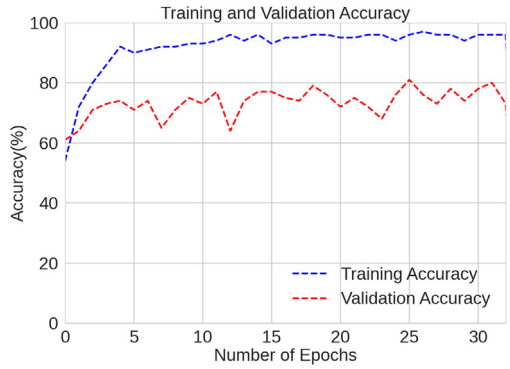
Fig. 8a-Fig. 8c show t-SNE visualization of latent embeddings obtained after the first, tenth and final epoch of the model for ESC-10 dataset. The t-SNE visualization obtained after the final epoch i.e. Fig. (8c) demonstrates that the latent embeddings belonging to the same class show more compact representation as compared to an initial representation of the latent embeddings Fig. (8a). We also presented the t-SNE visualization of the progress of the model for the DCASE 2019 Task 1-(A) dataset. Fig. (9a), Fig. (9a), and Fig. (9c) show latent embeddings after the first epoch, tenth epoch and final epoch. Initially, all latent embeddings are scattered and overlap with each other as shown in the Fig. (9a), Fig. (9b). It is clearly visible from the Fig. (9c) that the latent embeddings obtained after the final epoch are more compact and distinguishable. Fig. (10) shows accuracy and loss curves for the ESC-10 and DCASE 2019 Task-1(A) Datasets.

4.7.1. Analysis of computational complexity

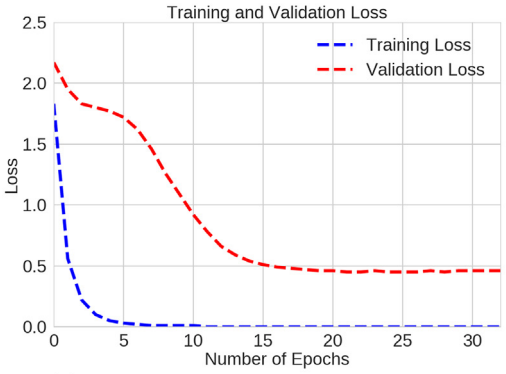
Besides presenting the accuracy of the proposed model we also evaluated the computational complexity of the proposed model for the ESC-10 and DCASE 2019 Task-1(A) Datasets. We selected the pioneer work of Piczak et al. [13] as a base model (PiczakCNN) for comparison. Table 10 shows comparison of computational complexity of the proposed model with the PiczakCNN. The model needs a total of 31.53 M parameters to perform classification and achieves accuracy of 80.50%. The PiczakCNN takes the input spectrogram of size 60∗41∗2. However, the proposed model takes 19.57 M less parameters as compared to the PiczakCNN and yields 11.50% more accuracy than the PiczakCNN. For DCASE 2019 Task-1 (A) we selected the official baseline model [22] for comparison. Table 11 shows comparison of computational complexity of the proposed model with the baseline model. The model needs a total of 0.12 M parameters to perform classification and achieves accuracy of 62.50%. However the proposed model takes 11.84 M more parameters compared to the baseline model but yields 19.50% more accuracy than the baseline model. The model with scaled dot product attention [28] needs a total of 11.70 million parameters for the ESC-10 and DCASE 2019 Task-1(A) datasets. Table 12 shows comparison of number of Floating Point Operations (FLOPs) required by the proposed model compared to the baseline models. The proposed model needs a total number of 4.032 and 8.01 billion FLOPs to classify the signals that come from the ESC-10 and DCASE 2019 Task1(A) datasets respectively. The model that uses scaled dot product attention [28] for ESC needs a total of 4.011 and 7.981 billion FLOPs for the ESC-10 and DCASE 2019 Task-1(A) datasets respectively.



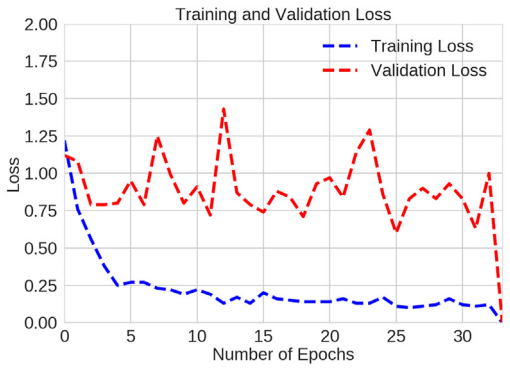
(a) Training and Validation Accuracy Curve for ESC 10



(b) Training and Validation Accuracy Curve for DCASE 2019 Task-1(A)



(c) Training and Validation Loss Curve for ESC 10



(d) Training and Validation Loss Curve for DCASE 2019 Task-1(A)

Fig. 10. Accuracy and Loss Curves for The ESC 10 and DCASE 2019 Task-1(A) Datasets.

4.8. Visual explanations of environmental sound classification

Fig. (11) and Fig. (12) shown a visualization of the feature map of the attention module. The feature map explains how the contribution of different time-frequency locations of the spectrogram

Table 10
Analysis of Computational Complexity of The Proposed Method and PiczakCNN [13] for ESC-10 Dataset.

Model	Number of Model Parameters (In Million (M))	Input Size(F, T,C)
PiczakCNN [13]	31.53	60∗41∗2
Scaled Dot product Attention [28]	11.70	128∗431∗1
Proposed (Without Attention)	11.69	128∗431∗1
Proposed	11.96	128∗431∗1

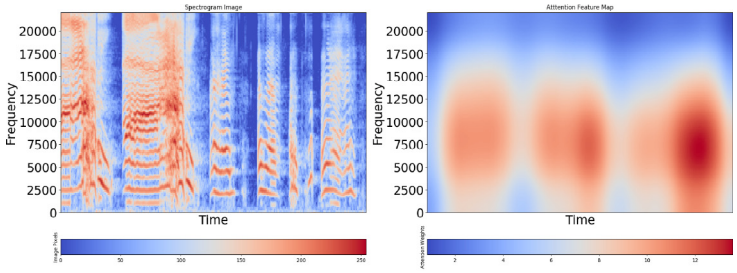
Table 11
Analysis of Computational Complexity of The Proposed Method and Baseline Model for DCASE 2019 Task-1(A) Dataset.

Model	Number of Model Parameters (In Million (M))
Baseline Model [22]	0.12
Scaled Dot Product Attention [28]	11.70
Proposed (Without Attention)	11.69
Proposed	11.96

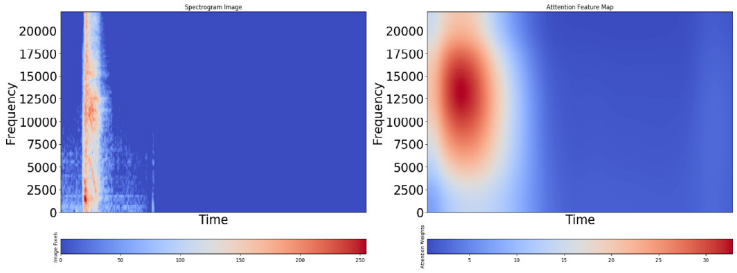
Table 12
Comparison of Number of Floating Point Operations (FLOPs) Required by The Proposed Model.

Model	Number of FLOPs (In Billion)	
	ESC-10	DCASE 2019 Task-1(A)
Scaled Dot Product Attention [28]	4.011	7.981
Baseline	6.327	3.281
Proposed	4.032	8.010

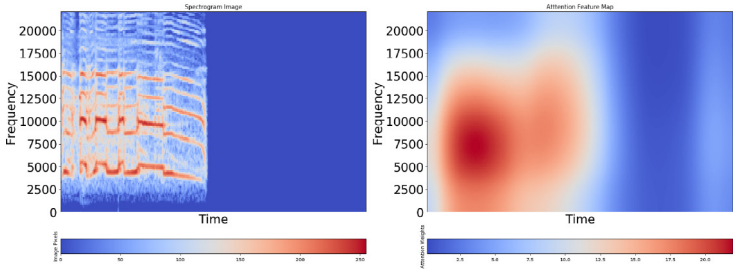
differs according to change in the type of environmental scenes (Baby crying, Person Sneezing, Rooster, Sea waves). It is apparent from Fig. (12) and Fig. (11c) that the proposed attention mechanism correctly focuses on essential frames while neglecting the silent frames and other background noise for the signals that belong to the Person Sneezing and Rooster Classes. The temporal pattern is visible for the signal from the baby crying class, as shown in Fig. (11a).



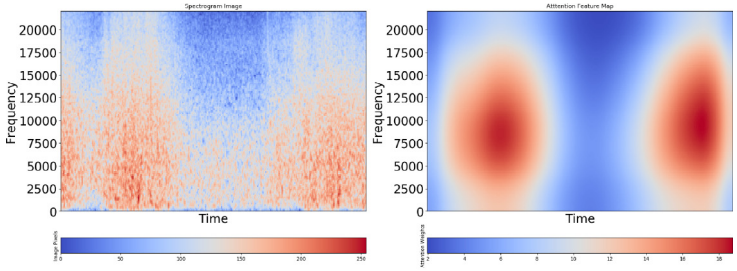
(a) Baby Cry



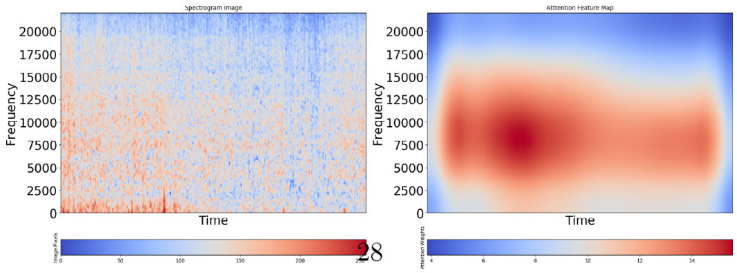
(b) Person Sneeze



(c) Rooster



(d) Sea Waves



(e) Rain

Fig. 11. Attention Feature Maps of The ESC-10 Signals Obtained by The Proposed Model.

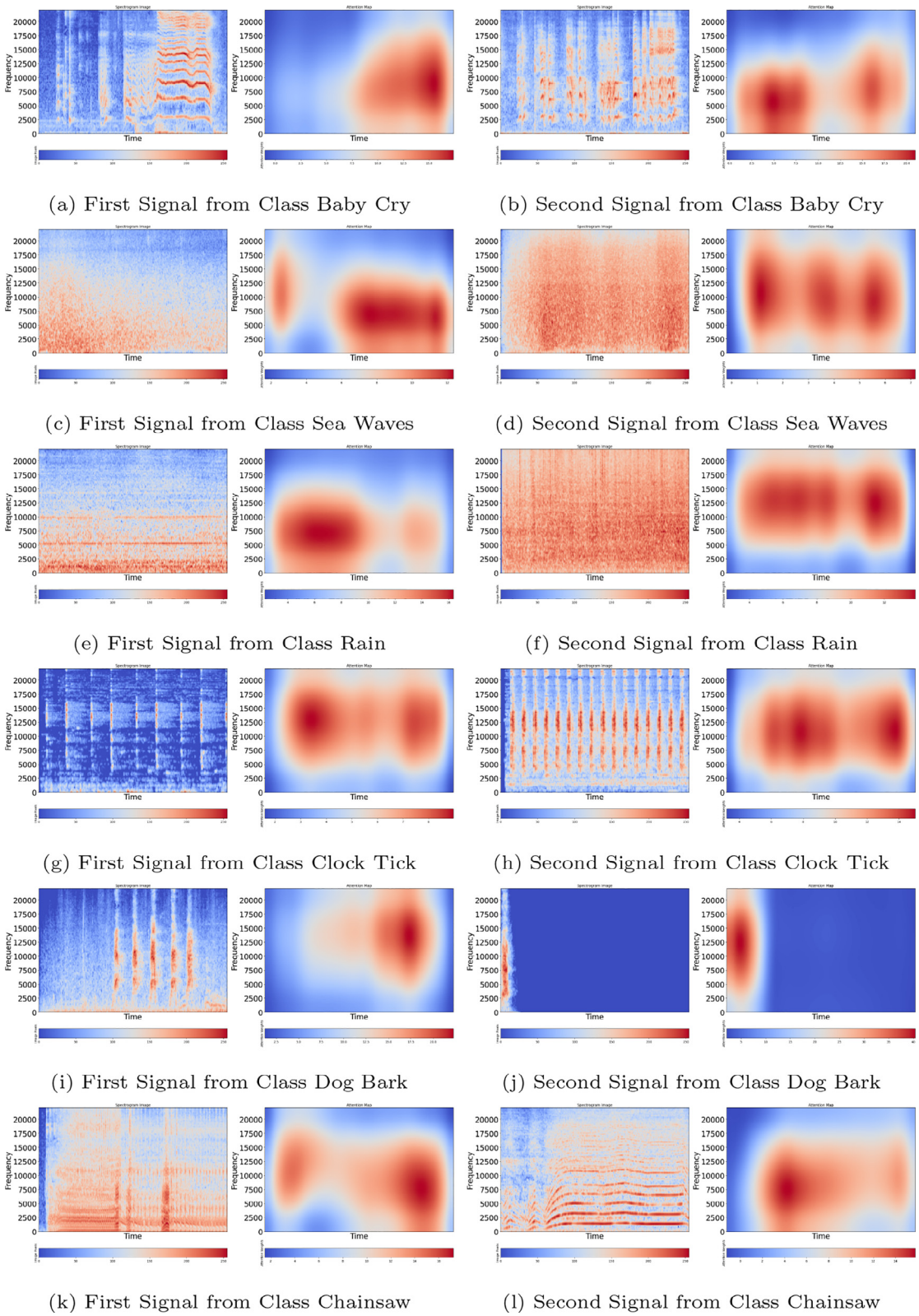


Fig. 12. Attention Feature Maps Obtained by Proposed Model for Two Different Signals That Belong to Same Class in ESC-10 Dataset.

5. Conclusion

In this paper, we have presented a novel attention-guided residual network that classifies the environmental sound. The visualization of the attention feature map shows that the proposed attention module efficiently guides the model to focus only on the semantically relevant regions of the spectrogram. The comprehensive experiments and results show that the proposed model

yields comparable accuracy to the state-of-the-art methods. Furthermore, we also investigated the location for placing the attention module to achieve the best accuracy. The proposed method shows promising results in classification of environmental sounds but the model's robustness has not yet been tested against the presence of noise. In future, our plan is to test the robustness of the model against different levels of noise.

CRediT authorship contribution statement

Achyut Man Tripathi: Conceptualization, Methodology, Software, Writing - original draft. **Aakansha Mishra:** Conceptualization. Methodology, Validation, Visualization, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] T. Zhang, C.-C.J. Kuo, Audio content analysis for online audiovisual data segmentation and classification, IEEE Trans. Speech Audio Process. 9 (4) (2001) 441–457.

[2] X. Valero, F. Alías, Gammatone wavelet features for sound classification in surveillance applications, in: 2012 Proceedings of the 20th European Signal Processing Conference (EUSIPCO), IEEE, 2012, pp. 1658–1662.

[3] B. Uzkent, B.D. Barkana, H. Cevikalp, Non-speech environmental sound classification using svms with a new set of features, Int. J. Innov. Comput. Inf. Control 8 (5) (2012) 3511–3524.

[4] S. Chu, S. Narayanan, C.-C.J. Kuo, Environmental sound recognition with time–frequency audio features, IEEE Trans. Speech Audio Process. 17 (6) (2009) 1142–1158.

[5] V. Bisot, R. Serizel, S. Essid, G. Richard, Feature learning with matrix factorization applied to acoustic scene classification, IEEE/ACM Trans. Audio Speech Language Process. 25 (6) (2017) 1216–1229.

[6] V. Bisot, R. Serizel, S. Essid, G. Richard, Nonnegative feature learning methods for acoustic scene classification (Tech. rep.), Technical report, DCASE2017 Challenge (2017). .

[7] K.J. Piczak, Esc: Dataset for environmental sound classification, in: Proceedings of the 23rd ACM International Conference on Multimedia, MM '15, Association for Computing Machinery, New York, NY, USA, 2015, p. 1015–1018. doi:10.1145/2733373.2806390. .

[8] P. Dhanalakshmi, S. Palanivel, V. Ramalingam, Classification of audio signals using aann and gmm, Appl. Soft Comput. 11 (1) (2011) 716–723.

[9] A. Voulodimos, N. Doulamis, A. Doulamis, E. Protopapadakis, Deep learning for computer vision: a brief review, Comput. Intell. Neurosci. 2018 (2018) 1–13.

[10] L. Zhang, S. Wang, B. Liu, Deep learning for sentiment analysis: a survey, Wiley Interdiscip. Rev.: Data Min. Knowl. Disc. 8 (4) (2018) e1253.

[11] H.I. Fawaz, G. Forestier, J. Weber, L. Idoumghar, P.-A. Muller, Deep learning for time series classification: a review, Data Min. Knowl. Disc. 33 (4) (2019) 917–963.

[12] N. Sünderhauf, O. Brock, W. Scheirer, R. Hadsell, D. Fox, J. Leitner, B. Upcroft, P. Abbeel, W. Burgard, M. Milford, et al., The limits and potentials of deep learning for robotics, Int. J. Robot. Res. 37 (4–5) (2018) 405–420.

[13] K.J. Piczak, Environmental sound classification with convolutional neural networks, in: 2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP), IEEE, 2015, pp. 1–6..

[14] I. McLoughlin, H. Zhang, Z. Xie, Y. Song, W. Xiao, Robust sound event classification using deep neural networks, IEEE/ACM Trans. Audio Speech Language Process. 23 (3) (2015) 540–552.

[15] T.H. Vu, J.-C. Wang, Acoustic scene and event recognition using recurrent neural networks, Detection and Classification of Acoustic Scenes and Events 2016 (2016) 1–3.

[16] H. Wang, Y. Zou, D. Chong, W. Wang, Environmental sound classification with parallel temporal-spectral attention, Proceedings of INTERSPEECH (2020).

[17] J. Wang, S. Li, Self-attention mechanism based system for dcase 2018 challenge task 1 and task 4 (Tech. rep.), Technical report, DCASE2018 Challenge (2018). .

[18] Z. Ren, Q. Kong, K. Qian, M.D. Plumbley, B. Schuller, et al., Attention-based convolutional neural networks for acoustic scene classification (Tech. rep.), Technical report, DCASE2018 Challenge (2018). .

[19] H. Zhenyi, J. Dacan, Acoustic scene classification based on deep convolutional neural network with spatial-temporal attention pooling (Tech. rep.), Technical report, DCASE2019 Challenge (2019). .

[20] Z. Ren, Q. Kong, J. Han, M.D. Plumbley, B.W. Schuller, Attention-based atrous convolutional neural networks: Visualisation and understanding perspectives of acoustic scenes, in: ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2019, pp. 56–60.

[21] L. Zhang, J. Han, Z. Shi, Atresn-net: Capturing attentive temporal relations in semantic neighborhood for acoustic scene classification, Proc. Interspeech 2020 (2020) 1181–1185.

[22] A. Mesaros, T. Heittola, T. Virtanen, A multi-device dataset for urban acoustic scene classification, in: Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019), 2019, pp. 9–13.

[23] V. Boddapati, A. Petef, J. Rasmusson, L. Lundberg, Classifying environmental sounds using image recognition networks, Proc. Comput. Sci. 112 (2017) 2048–2056.

[24] Y. Aytar, C. Vondrick, A. Torralba, Soundnet: Learning sound representations from unlabeled video, in: Advances in neural information processing systems, 2016, pp. 892–900. .

[25] W. Dai, C. Dai, S. Qu, J. Li, S. Das, Very deep convolutional neural networks for raw waveforms, in: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2017, pp. 421–425.

[26] Y. Tokozume, Y. Ushiku, T. Harada, Learning from between-class examples for deep sound recognition, arXiv preprint arXiv:1711.10282. .

[27] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, Y. Bengio, Show, attend and tell: Neural image caption generation with visual attention, in: International conference on machine learning, PMLR, 2015, pp. 2048–2057..

[28] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, arXiv preprint arXiv:1706.03762. .

[29] H. Fan, L. Zhu, Y. Yang, F. Wu, Recurrent attention network with reinforced generator for visual dialog, ACM Trans. Multimedia Comput. Commun. Appl. 16 (3) (2020) 1–16.

[30] W. Cai, B. Liu, Z. Wei, M. Li, J. Kan, Tardb-net: triple-attention guided residual dense and bilstm networks for hyperspectral image classification, Multimedia Tools Appl. (2021) 1–22.

[31] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, Y. Fu, Image super-resolution using very deep residual channel attention networks, in: Proceedings of the European conference on computer vision (ECCV), 2018, pp. 286–301.

[32] L. Zhu, Y. Yang, Label independent memory for semi-supervised few-shot video classification, IEEE Trans. Pattern Anal. Mach. Intell. (2020) 1, https://doi.org/10.1109/TPAMI.2020.3007511.

[33] Y. Wu, L. Zhu, Y. Yan, Y. Yang, Dual attention matching for audio-visual event localization, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 6292–6300.

[34] J. Guo, N. Xu, L.-J. Li, A. Alwan, Attention based cldnns for short-duration acoustic scene classification, in: Proc. Interspeech 2017, 2017, pp. 469–473. doi:10.21437/Interspeech.2017-440. URL:https://doi.org/10.21437/Interspeech.2017-440. .

[35] J. Wang, S. Li, Self-attention mechanism based system for dcase2018 challenge task1 and task4 (Tech. rep.), Technical report, DCASE2018 Challenge (2018). .

[36] Y. Wang, C. Feng, D.V. Anderson, A multi-channel temporal attention convolutional neural network model for environmental sound classification, arXiv preprint arXiv:2011.02561. .

[37] P. Raguraman, M. R., M. Vijayan, Librosa based assessment tool for music information retrieval systems, in: 2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), 2019, pp. 109–114. doi:10.1109/MIPR.2019.00027..

[38] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980. .

[39] Z. Liu, J. Du, M. Wang, S.S. Ge, Adcm: attention dropout convolutional module, Neurocomputing 394 (2020) 95–104.

[40] X. Wang, L. Zhu, Y. Wu, Y. Yang, Symbiotic attention for egocentric action recognition with object-centric alignment, IEEE Trans. Pattern Anal. Mach. Intell. (2020) 1, https://doi.org/10.1109/TPAMI.2020.3015894.

[41] S. Ahmad, S. Agrawal, S. Joshi, S. Taran, V. Bajaj, F. Demir, A. Sengur, Environmental sound classification using optimum allocation sampling based empirical mode decomposition, Physica A 537 (2020) 122613.

[42] Y. Su, K. Zhang, J. Wang, K. Madani, Environment sound classification using a two-stream cnn based on decision-level fusion, Sensors 19 (7) (2019) 1733.

[43] J. Salamon, J.P. Bello, Deep convolutional neural networks and data augmentation for environmental sound classification, IEEE Signal Process. Lett. 24 (3) (2017) 279–283.

[44] E. Akbal, An automated environmental sound classification methods based on statistical and textural feature, Appl. Acoust. 167 (2020) 107413.

[45] J.-w. Jung, H. Heo, H.-j. Shim, H.-J. Yu, Distilling the knowledge of specialist deep neural networks in acoustic scene classification (Tech. rep.), Technical report, DCASE2019 Challenge (2019). .

[46] B. Ding, G. Liu, J. Liang, Acoustic scene classification based on ensemble system (Tech. rep.), Technical report, DCASE2019 Challenge (2019). .

[47] S. Waldekar, G. Saha, Wavelet based mel-scaled features for dcase 2019 task 1a and task 1b (Tech. rep.), DCASE2019 Challenge (2019). .

[48] N. Zhou, Y. Liu, Q. Wei, Audio scene calssification based on deeper cnn and mixed mono channel feature (Tech. rep.), DCASE2019 Challenge (2019). .

[49] C. Paseddula, S.V. Gangashetty, Dcase 2019 task 1a: acoustic scene classification by sffcc and dnn (Tech. rep.), DCASE2019 Challenge (2019). .

[50] X. Ma, M. Gu, Y. Ma, Jsnu_wdxy submission for dcase-2019: Acoustic scene classification with convolution neural networks (Tech. rep.), DCASE2019 Challenge (2019). .

[51] L. Pham, T. Doan, D. Ngo, H. Hong, H.H. Kha, Cdnn-crnn joined model for acoustic scene classification (Tech. rep.), DCASE2019 Challenge (2019). .

[52] S. Suh, W. Lim, S. Park, Y. Jeong, Acoustic scene classification using specaugment and convolutional neural network with inception modules (Tech. rep.), DCASE2019 Challenge (2019). .
[53] Y. Wu, T. Lee, Stratified time-frequency features for cnn-based acoustic scene classification (Tech. rep.), DCASE2019 Challenge (2019). .
[54] H. Zeinali, L. Burget, J. Černocký, et al., Acoustic scene classification using fusion of attentive convolutional neural networks for dcase2019 challenge, arXiv preprint arXiv:1907.07127. .
[55] C. Lei, Z. Wang, Multi-scale recalibrated features fusion for acoustic scene classification (Tech. rep.), DCASE2019 Challenge (2019). .
[56] J. Cho, S. Yun, H. Park, J. Eum, K. Hwang, Acoustic scene classification based on a large-margin factorized cnn (Tech. rep.), DCASE2019 Challenge (2019). .
[57] L. Zhang, J. Han, Z. Shi, Learning temporal relations from semantic neighbors for acoustic scene classification, IEEE Signal Process. Lett. 27 (2020) 950–954, https://doi.org/10.1109/LSP.2020.2996085.
[58] Z. Mushtaq, S.-F. Su, Environmental sound classification using a regularized deep convolutional neural network with data augmentation, Appl. Acoust. 167 (2020) 107389.

Achyut Mani Tripathi received the B.E. degree in Information Technology from the Chhattisgarh Swami Vivekanand Technical University (CSVTU), Bhilai, Chhattisgarh, India, in 2012. He is currently pursuing a Ph.D. degree in the Department of Computer Science & Engineering at the Indian Institute of Technology Guwahati, Assam, India. His research interests include deep learning, environmental sound classification, and anomaly detection.

Aakansha Mishra received her Masters degree from Banasthali University, Rajasthan, India. Currently, she is working towards her Ph.D. degree from the Indian Institute of Technology Guwahati, Assam, India. Her research interest lies in Deep Learning, Multimodal Interaction, Computer Vision.