Received: September 1, 2022. Revised: September 26, 2022.

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Acoustic Scene Classification using Attention based Deep Learning Model

Mie Mie Oo[1]* **Nu War[2]**

[1]*University of Computer Studies, Mandalay, Myanmar*
[2]*Myanmar Institute of Information Technology, Mandalay, Myanmar*
* Corresponding author's Email: miemieoo@ucsm.edu.mm

Abstract: Acoustic scene classification is a difficult issue among artificial intelligence, signal processing, and machine learning. Scene recognition performance has a robust relation with feature learning using deep convolutional networks. In the following research, end-to-end deep residual network embedded channel attention is explored to learn the discriminative features from the audio scene. Log-Mel spectrogram is obtained from input raw audios. It is forwarded to proposed attention network. An extracted feature layer is concatenated with the SoftMax classifier in the proposed attention network. The experimentation is carried out on Detection and Classification of Acoustic Scenes and Events (DCASE) 2016 and 2017 datasets. The proposed channel-attention-based residual network achieves classification results with an average accuracy of 80.27% and 80.82%, respectively.

Keywords: Residual network, Channel attention, Log-Mel spectrogram, Gammatone frequency cepstral coefficient, Acoustic scene classification.

1. Introduction

Acoustic Scene Classification (ASC) task is used in classifying audios in different environments as one of the categories including beach, bus, shopping mall, office, park, train, tram etc., Nowadays, the audio files have been acquired from mobiles or wearable devices to identify the semantic label of each audio. ASC work has become challenging topic recently in the fields of signal processing system. It has been attracted in many application areas such as smart devices, intelligent wearable interfaces, hearing aids, and other applications.

ASC problem has been improved the classification results with the advance in deep learning. The prior approaches for the ASC task tended to the proper feature engineering. The time frequency images are used as input to extract features using deep convolutional neural network and then classified using ensembled classifier [1]. In [3], deep neural network with discrete Fourier transform is constructed to improve the capability of ASC tasks. Moreover, the most recent ASC work incorporate features fusion [33], ensembled models [8] or ensembled classifiers [22]. These network designs improve accuracy, but the problem is huge computational demands, such as graphics processing units (GPUs). With end-to-end fashion, the chief purpose of the proposed ASC system in order to assign test records to one of the specified class labels that best describes the circumstances in which they were made. In this research work, RNN is designed as a basic network then the channel attention is embedded in this network. In the proposed system, the effect of mixed-up data augmentation methods is also explored in the research. The contributions of the research study are described as:

- An end-to-end residual network is designed for acoustic scene classification using Log-Mel spectrogram images.
- A channel attention block is incorporated to extract the discriminative and meaningful representations of audios from different environments.
- As the extended evaluation, the proposed model is tested on Gammatone Frequency Cepstral Coefficient (GFCC) features with different input sizes, with or without data augmentation.

The parts of this paper include section 2 described in recent literature and related work of ASC system. The section 3 designated that system based on improved RNN. In Section 4, the setup of experimentation and evaluation results are presented. The comparisons with the previous research approaches are discussed in Section 5 and the last one concluded the paper.

2. Related work

Deep Neural Network is applied to natural language processing, image processing, speech recognition and many others. Pretrained convolutional network, AlexNet, VGG-16, VGG-19 [14], ResNet-18 [15] are used to extracted the acoustic features. These networks are trained on image datasets. The extracted features are sufficient information on image processing but achieve the lack of useful features on audio data. Extended convolutional neural network with squeeze and excitation residual blocks is explored for audio classification in paper [6]. The convolutional network is employed for the comparison of Mel-spectrogram-based network and wavelet-based features network. The research explored that the excitation block of network is sensitive to the loudness audio.

Recent works on ASC performed the popular low-level feature extraction methods such as Log-Mel scale [8, 21, 24, 28, 36], Constant-Q transform spectrograms [3]. While Mel spectrogram feature uses as a most used method for acoustic signal process, several pre-processing techniques are also performed various aspects of an acoustic scene. In ordered to progress the performance of acoustic scene recognition, multiple deep convolutional neural networks are individually trained on different input audio features or various low-level feature extraction methods are considered to construct ensemble models.

To develop network for ASC, [4] presented ensembled network by using Log-Mel spectrogram, delta, delta-deltas, and Harmonic Percussive Source Separation (HPSS) features mixed-up and crop augmentation. In addition, [13] employed Mel Frequency Cepstral Coefficients (MFCC), Log-Mel spectrum, Gammatone and Constant Q Transform (CQT), four convolutional networks for high-level representation of features with four networks and ensemble classifiers. The study needs to provide the enough computational requirements. To improve the computation and accuracy of deep learning model, multi-level feature [12, 26] or multi-scale semantic features fusion [27, 29] methods are used in the model combining with data augmentation.

In [7], soft labels are extracted from pretrained network and then employed self multi-head attention. It proves that the performance of single network outperforms multiple models on DCASE-2019 dataset [30]. V. Abrol and P. Sharma [11] performed two pipelines network by applying statistical pooling attention and multiple augmentation techniques like time stretching, dynamic range compression, background noise addition and pitch shifting. In 2019, the paper [18] proved that dilated convolution is much better than the maximum pooling but the large dilation rate tends to low recognition rate. In [31], convolutional neural network is designed end-to-end ASC system embedding the statistical pooling layers based on acoustic raw-waveform. The recognition result is comparable the network with maximum or average pooling.

Multi-channels acoustic scene recognition for domestic home activities is evaluated using Non-negative Matrix Factorization (NMF) based convolutional neural network with mix and shuffle data augmentation [5]. In a specific study of audio-visual scene recognition [10], the two feature extraction modules such as video temporal features and audio features and audio features are extracted by using weakly supervised representation learning for experiment on DCASE 2017 dataset. Another supervised feature learning approach for ASC, supervised non-negative matrix factorization (SNMF), convolutional network, and histogram of gradient features are used in [16] by applying SVM classifier. The study found that SNMF tend to overfitting than convolutional neural networks.

For the tasks with less training audio data, the previous study [20] is constructed and trained two models. Instance Specific Adapted Gaussian Mixture Models (ISAGMMs) is explored for conservational audio scenes and Instance Specific Hidden Markov Models (ISHMMs) is used in different sound events. In [32], one channel convolutional neural network is intended to reduce the feature redundancy and to decompose or optimize the calculation of convolution among channels.

In [25], three types of feature representation such as Log mel band energies, LPCC (Linear Prediction Cepstral Coefficient), and SCMC (Spectral Centroid Magnitude Cepstral Coefficients) are individually extracted to investigate the complementary feature representations of ASC tasks. Deep neural network is applied in prediction of the class label of each scene. The classification scores of three features are fused at the decision level to predict the final acoustic scene label. The result of score fusion improves the accuracy but increase computational cost.

For acoustic scene classification tasks, audio datasets are published such as DCASE-2016 [17] and DCASE-2017 [14]. The baseline result of the first dataset is accuracy of 77.20% by using Mel Frequency Cepstral Coefficient (MFCC) and Gaussian Mixture Model (GMM). Competition result for DCASE-2017 challenge achieved accuracy of 83.40% using pretrained networks for feature extraction and Gated-Recurrent Neural Network classifier.

3. The proposed ASC system architecture

This part represents the detailed architecture of the proposed ASC work which describes the audio processing, data augmentation, network architecture, channel attention mechanism and classification in detail. The design of the mentioned ASC system is represented in Fig. 1.

3.1 Audio processing

The input audios are pre-processed as Log -Mel spectrograms before the network training. Mel frequency bank can be used to extract these spectrograms, which can then be scaled using a logarithm [12]. In this research, the spectrograms with the size of 64, 128 Mel frequency bins. The audio file of sampling frequency is 44100 Hz. The frame durations of the window lengths are 80ms and 40ms, and the overlap length is 50%. The calculation of the number of frame lengths is in Eq. (1).

$$f = 1 + floor\left(\frac{L-W}{S}\right) \quad (1)$$

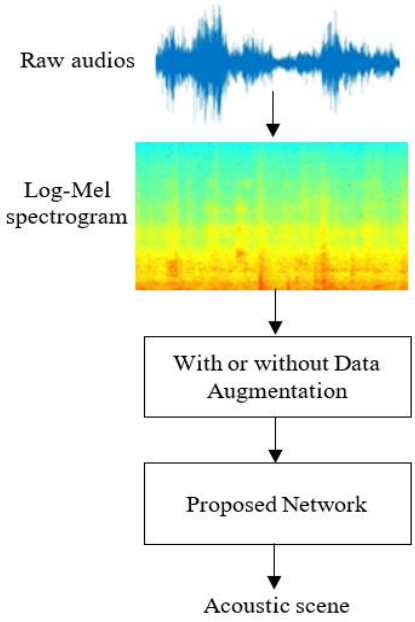

Figure. 1 Architecture of proposed ASC system

where f is the number of frames, L is signal length, W is window length and S is shift length. After the framing is processed, the hamming window uses each individual frame. Hamming window functions are designed for signal framing due to their superior frequency resolution and spectral distortion. Hamming window $h(n)$ is defined by the formula Eq. (2).

$$h(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right)\ 1 \le n \le N \quad (2)$$

Where N is frame number, and $h(n)$ is the hamming window.

The window function is used to smooth the signal to calculate a Fast Fourier Transform (FFT) to obtain an amplitude-frequency response per frame. FFT is the process of transforming time to a frequency domain and applying it to the spectrum according to Eq. (3).

$$S_i(k) = \sum_{i=1}^{N} s_i(k)h(n)e^{-\frac{j2\pi kn}{N}}\ n = 0,..,N-1 \quad (3)$$

Where $h(n)$ is the analysis window of N samples, $s_i(k)$ are samples in the time domain, $S_i(k)$ are samples in frequency domain whereas N is size of the FFT. Mel scale depends on the impression of human hearing frequencies. Me frequency scale is in Eq. (4).

$$mel(f) = 2595 \times log_{10}(1 + {}^{f}/_{700}) \quad (4)$$

Where $mel(f)$ is the frequency of mel and f is frequency of linear frequency. Mel filter is used to obtain the energy from the filter bank and to apply the compression the log on the filter output is used. Some of the he Log-Mel spectrogram are presented in Fig. 2.

3.2 Audio data augmentation

The limited sizes of publicly available datasets are prone to overfitting for deep model. To increase the generalization ability of proposed residual model, the proposed system used data augmentation approach such as mixed-up augmentation which is an effective method for audio data. Data augmentation technique also alleviates the problem of model overfitting. The mixed-up approach reduces the neighbourhood risk of samples during training. The original and mixed datasets are combined after the spectrograms are mixed in Eqs. (5) and (6) with lambda set to 0.5.

$$\bar{x} = \lambda x_i + (1-\lambda)x_j \quad (5)$$

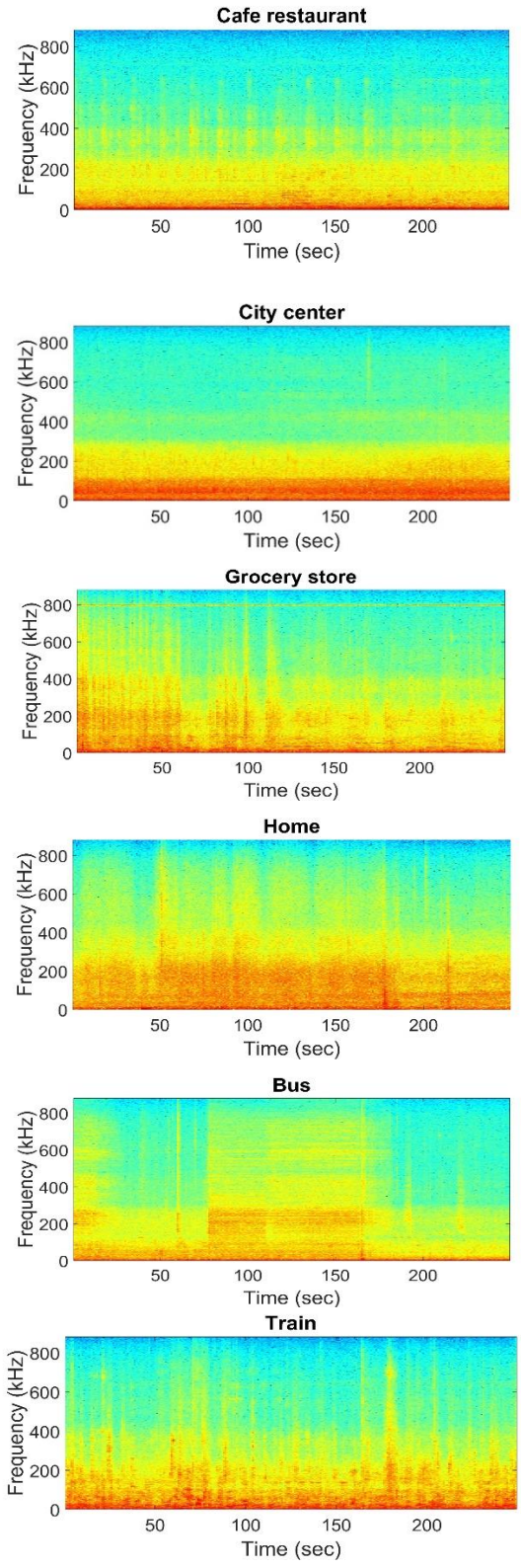

Figure. 2 Log-Mel spectrograms of some classes

$$\bar{y} = \lambda y_i + (1-\lambda) y_j \quad (6)$$

where (x_i, y_i) and (x_j, y_j) are two random selected examples from training dataset and the diverse ratio is λ [35]. Setting the lambda value to 0.5 improves model generalization by adapting the network structure during training, which outperforms other values.

3.3 Attention based RNN network

RNN has automatic learning ability of high-level feature representations to identify the patterns of acoustic signals [9]. In the proposed ASC system, the raw audios are converted into Log-Mel spectrogram representations with mono channel.

Mono signal can capture the variations of a signal amplitude with time which applies at dynamic frequencies. Thus, discriminating features are obtained from the audios. The proposed attention-based residual network receives this Log-Mel spectrogram as input.

Residual layers are constructed with a residual function: H(Y) = F(Y) + Y, where Y as input features, H(.) as a mapping that a series of stacked layers will attempt to fit [2]. The implementation of residual learning is simple CNN which is a combined shortcut connection called identity mapping. There are various structures to construct the residual block. In the research, the residual block is implemented as layers see in Fig. 3.

Three residual blocks, a global average pooling layer, a dropout layer, and one fully connected layer with SoftMax classifier make up baseline residual network. Instead of the last residual block in the baseline network, channel attention block was

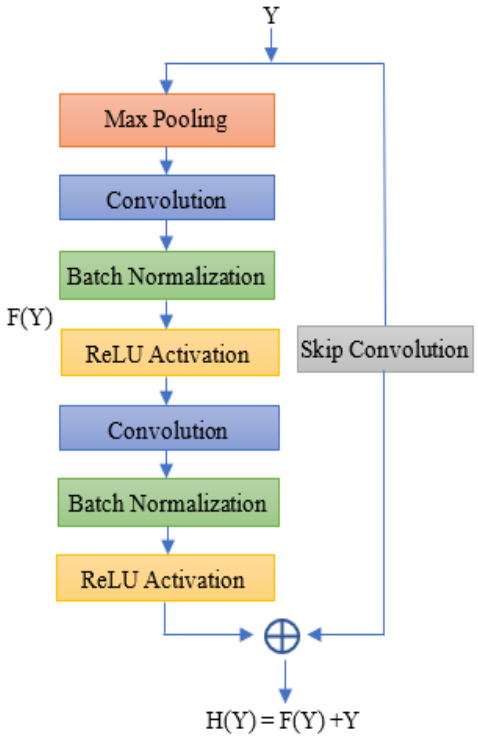

Figure. 3 Residual unit with identity mapping

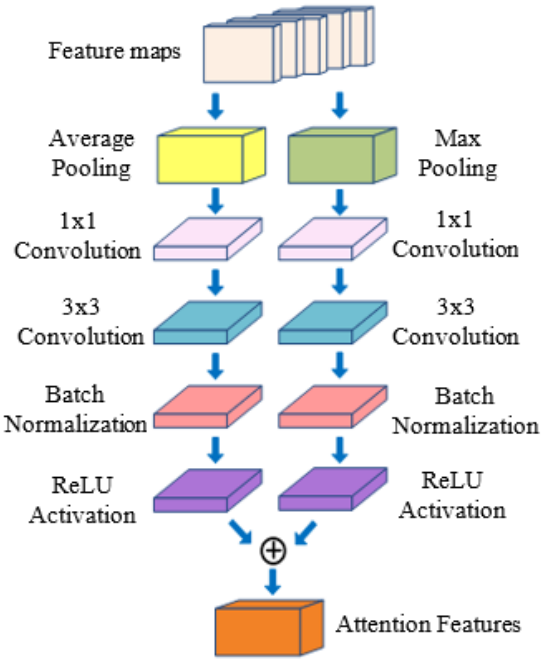

Figure. 4 Channel attention module

applied to distinguish variable acoustic features. The application for channel attention mechanism is revealed in Fig. 4.

For input X, the channel attention block's output is as follows in Eq. (7).

$$X_{l+1} = F_{avg}(X) + F_{max}(X) \quad (7)$$

Where $F_{avg}(.)$ performs the processes of average pooling, 1×1 and 3×3 convolution, batch normalization, and ReLU activation. $F_{max}(.)$ performs the processes of maximum pooling layer, 1×1 convolution, 3×3 convolution, ReLU activation and batch normalization.

4. Experimental setup

DCASE-2016 ASC challenge dataset: Task 1 is ASC task including audios recoded from distinct locations. There are 15 different classes with 30-second audio files. The development dataset is applied to test the proposed ASC model.

DCASE-2017 ASC challenge dataset: The dataset [19] is also consisted of 15 different classes, 4680 audio files of the development data. Proposed model was evaluated on the development data with 4-folds cross validation Each audio file has 10 seconds duration which are recoded from distinct locations. In both datasets, the labels of the audios are beach, bus, car, grocery store, forest path, home, café or restaurant, metro station, city center, train, tram, residential area, library, office, park.

Table 1. Parameters of Log-Mel and GFCC feature extraction

	Window length/ Shift length	No. of filter banks	Feature size of DCASE-2016	Feature size of DCASE-2017
Log-Mel	80ms/ 40ms	64	749×64	249×64
	80ms/ 40ms	128	749×128	249×128
	40ms/ 20ms	64	1499×64	499×64
GFCC	80ms/ 40ms	64	749×64	249×64
	80ms/ 40ms	128	749×128	249×128
	40ms/ 20ms	64	1499×64	499×64

The Log-Mel spectrogram is a representation used as input for neural networks, primarily for the speech recognition field. Firstly, the raw audio files were down-mixed to mono channel. Then, the Log-Mel spectrum with different size of feature maps was extracted to facilitate the network training. To train the suggested network, the cross-entropy loss is minimized using the Stogastic Gradient Decent Minimization (SDGM) optimizer. The minimum batch size, number of maximum iterations and learning rate is 128 batches, 50 epochs, and 0.001 respectively.

To compare audio input features, Gammatone Frequency Cepstral Coefficients (GFCC) [37] is also used to classify the effectiveness of input for ASC task. The sampling frequency rate is 44100 Hz for both Log-Mel and GFCC extraction. The lowest frequency rate is zero and the highest frequency rate is 22050. The parameters of Log-Mel spectrogram and GFCC feature extraction for the two datasets are given below as Table 1.

All experimentations are implemented using MATLAB 2019b using deep learning toolbox with the hardware specifications: Intel Core i7-7500U 2.7GHz Processor and 16 GB DDR3 L Memory. The experimental results show average classification accuracy of four-fold cross-validation.

5. Evaluation results

In experimentation of the research, individual features such as Log-Mel features and GFCC features are considered as the input of network.

5.1 Evaluation using Log-Mel features

The acoustic classification results of DCASE-2016 Challenge Task 1 are described in Table 2 and

Table 2. Accuracy (%) of DCASE-2016 dataset using Log-Mel spectrogram images without augmentation

Validation	Log-Mel feature size		
	1499×64	749×64	749×128
Fold_1	73.10%	77.59%	85.52%
Fold_2	70.69%	73.79%	77.24%
Fold_3	73.49%	79.66%	84.14%
Fold_4	70.55%	75.86%	75.26%
Mean Accuracy	71.95%	76.73%	**80.54%**

Table 3. Accuracy (%) of DCASE-2016 dataset using Log-Mel spectrogram images with mixed-up augmentation

Validation	Log-Mel feature size		
	1499×64	749×64	749×128
Fold_1	76.21%	70.69%	77.61%
Fold_2	73.10%	73.79%	77.24%
Fold_3	77.93%	71.38%	82.07%
Fold_4	68.62%	66.10%	84.14%
Mean Accuracy	73.97%	70.49%	**80.27%**

Table 4. Accuracy (%) of DCASE-2017 dataset using Log-Mel spectrogram images without augmentation

Validation	Log-Mel feature size		
	499×64	249×64	249×128
Fold_1	81.20%	82.48%	82.07%
Fold_2	83.12%	81.59%	79.37%
Fold_3	75.36%	78.35%	82.78%
Fold_4	82.22%	80.09%	80.85%
Mean Accuracy	80.48%	80.63%	**81.27%**

Table 5. Accuracy (%) of DCASE-2017 dataset using Log-Mel spectrogram images with mixed-up augmentation

Validation	Log-Mel feature size		
	499×64	249×64	249×128
Fold_1	79.83%	75.90%	80.51%
Fold_2	78.26%	74.25%	80.56%
Fold_3	74.77%	71.78%	78.52%
Fold_4	73.42%	74.19%	83.68%
Mean Accuracy	76.57%	74.03%	**80.82%**

Table 3 which shows the effect of different input sizes using Log-Mel spectrogram images.

The highest mean accuracy of 80.54% was achieved without augmentation and 80.27% with mixed-up augmentation on DCASE-2016 dataset. The experimentation on DCASE-2017 Challenge Task 1 also explored by using different size of Log-Mel spectrogram shown in Table 4 and Table 5. The proposed model achieved 81.27% of accuracy without augmentation and 80.82% of accuracy with mixed-up augmentation. The result using augmentation method is not improved accuracy but the generalization of model is improved. Augmentation alleviates the overfitting problem of model.

5.2 Evaluation using GFCC features

Evaluations of proposed system are employed to compare Log-Mel spectrogram with GFCC features. GFCC features with different sizes are applied to classify the acoustic scene. Some of the spectrograms of GFCC are shown in Fig. 5.

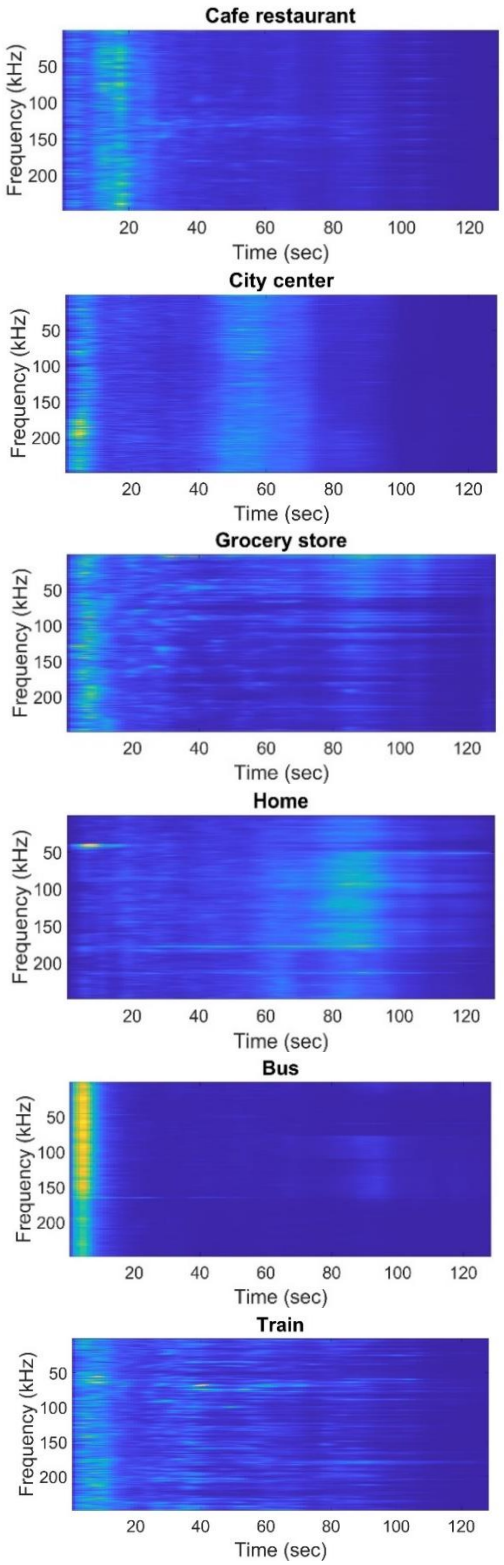

Figure. 5 GFCC spectrograms of some classes

Table 6. Accuracy (%) of DCASE-2016 dataset using GFCC without augmentation

Validation	GFCC feature size		
	1499×64	749×64	749×128
Fold_1	72.76%	74.83%	76.55%
Fold_2	67.24%	67.59%	68.28%
Fold_3	73.49%	71.03%	71.72%
Fold_4	74.32%	73.63%	74.48%
Mean Accuracy	71.95%	71.77%	**72.46%**

Table 7. Accuracy (%) of DCASE-2016 dataset using GFCC with mixed-up augmentation

Validation	GFCC feature size		
	1499×64	749×64	749×128
Fold_1	71.03%	68.28%	76.90%
Fold_2	62.41%	69.31%	74.48%
Fold_3	68.97%	70.13%	73.83%
Fold_4	71.58%	65.75%	72.41%
Mean Accuracy	68.50%	68.37%	**74.41%**

Table 8. Accuracy (%) of DCASE-2017 dataset using GFCC without augmentation

Validation	GFCC feature size		
	499×64	249×64	249×128
Fold_1	77.86%	75.04%	79.91%
Fold_2	78.77%	81.33%	80.73%
Fold_3	73.49%	75.45%	78.52%
Fold_4	83.16%	81.79%	81.54%
Mean Accuracy	78.32%	78.40%	**80.18%**

Table 9. Accuracy (%) of DCASE-2017 dataset using GFCC with mixed-up augmentation

Validation	GFCC feature size		
	499×64	249×64	249×128
Fold_1	71.45%	70.85%	76.24%
Fold_2	73.49%	72.21%	75.36%
Fold_3	72.21%	72.63%	72.21%
Fold_4	74.02%	72.65%	69.49%
Mean Accuracy	71.79%	72.09%	**73.33%**

The experimentation results using GFCC features without augmentation on DCASE-2016 dataset is presented in Table 6 and with mixed-up augmentation in Table 7.

The proposed model using GFCC features achieved mean accuracy of 72.64% without augmentation and 74.41% with mixed-up augmentation on DCASE-2016 acoustic challenge dataset. The results of DCASE-2017 dataset presented in Table 8 and Table 9. In the experimentation, the result without augmentation achieved 80.18% and drop to 73.33 % with mixed-up augmentation.

6. Compare methods

The results of Log-Mel spectrogram and GFCC features are compared in accuracy shown in Fig. 6. The experimentation result with channel attention improved accuracy of 0.88% on DCASE-2016 dataset and 2.51% on DCASE 2017 dataset respectively. Log-Mel spectrogram features improves the classification result than GFCC features. In Table 10 and Table 11, proposed ASC system is compared with baseline system of datasets and state-of-the-art methods.

The proposed channel attention network achieved comparable result with HOG feature and SVM classifier. The network achieved improved accuracy than deep neural network with Log-Mel features in [25]. The proposed network improves generalization ability for acoustic scene classification. The residual network with channel attention performs better than the network without attention. The network without attention achieved 78.31% and 79.39% accuracy on DCASE-2016 and DCASE-2017 datasets correspondingly. Residual network with channel

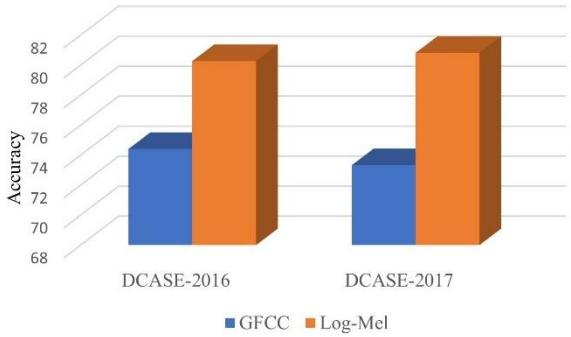

Figure. 6 Accuracy comparison between GFCC and Log-Mel features on DCASE 2016 and DCASE-2017 datasets

Table 10. Comparison on DCASE-2016 dataset in accuracy (%)

Methods	Accuracy
DCASE-2016 baseline system MFCC features + Gaussian mixture model (GMM) classifier [17]	77.20%
Log-Mel spectrogram + non-negative matrix factorization + Histogram of Oriented Gradient (HOG) features + SVM classifier [16]	80.93%
Proposed method Log-Mel + residual network with no attention and mixed-up augmentation	78.31%
Proposed method Log-Mel + channel attention residual network without augmentation	80.54%
Proposed method Log-Mel + channel attention residual network with mixed-up augmentation	80.27%

Table 11. Comparison on DCASE-2017 dataset in accuracy (%)

Methods	Accuracy
Spectrogram, bump, and morse scalograms Feature extraction with AlexNet, VGG-16 and VGG-19 for feature extraction Gated-RNN network with SoftMax classifier for classification [14]	83.40%
Log-Mel spectrogram + Deep neural network [25]	69.40%
Proposed method Log-Mel + residual network with no attention and mixed-up augmentation	79.39%
Proposed method Log-Mel + channel attention residual network without augmentation	81.27%
Proposed method Log-Mel + channel attention residual network with mixed-up augmentation	80.82%

attention improved accuracy of 80.54% and 81.27%. To eliminate the overfitting of network, mixed-up augmentation is applied to data before the network training. Although the method removes the overfitting problem but reduces the result to 80.27% and 80.82% accuracy respectively.

7. Conclusion

In this research, an effective ASC system is proposed. An improved residual network with channel attention is implemented and trained in end-to-end deep learning models to perform efficient features to classify acoustic scene in different surroundings. The developed channel attention based residual model was comprised of two residual blocks, one channel attention module, global average pooling layer, dropout layer, and one fully connected layer with a SoftMax classifier. Two acoustic scene datasets such as DCASE-2016 and DCASE-2017 datasets were used to test the improvement of proposed model by calculating the accuracy and then described the evaluation results. According to the experiments, Log-Mel spectrogram features outperformed than GFCC features. The results show that both Log-Mel spectrogram and GFCC features representation 128 filter banks are the best input feature size. In the future ASC task, channel attention features will be applied with other effective classifiers for acoustic scenes classification.

Conflicts of Interest

The authors declare that "No conflict of interest". The authors affirm not to have interpersonal disputes that could lead to have an impact on the research presented in this paper.

Author Contributions

Conceptualization, M.M.O., and N.W.; Methodology, N. W., and M.M.O.; Software, M.M.O; Investigation and Validation, M.M.O. and N. W.; Writing-original draft preparation, M.M.O.; Visualization, N.W. and M.M.O.; Supervision, N.W.; All authors have read and edit to the manuscript.

References

[1] F. Demir, D. A. Abdullah, and A. Sengur, "A New Deep CNN Model for Environmental Sound Classification", *IEEE Access*, Vol. 8, pp. 66529-66537, 2020.

[2] J. N. Alcazar, S. P. Castanos, P. Zuccarello, and M. Cobos, "Acoustic Scene Classification with Squeeze-Excitation Residual Networks", *IEEE Access*, Vol. 8, pp. 112287-112296, 2020.

[3] T. Zhang and J. Wu, "Constrained learned feature extraction for acoustic scene classification", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 27, No. 8, pp. 1216-1228, 2019.

[4] Y. Lee, S. Lim, and I. Kwak, "CNN-Based Acoustic Scene Classification System", *Electronics*, Vol. 10, No. 4, p. 371, 2021.

[5] S. Lee and H. S. Pang, "Feature Extraction Based on the Non-Negative Matrix Factorization of Convolutional Neural Networks for Monitoring Domestic Activity with Acoustic Signals", *IEEE Access*, Vol. 8, pp. 122384-122395, 2020.

[6] T. Kim, J. Lee, and J. Nam, "Comparison and Analysis of SampleCNN Architectures for Audio Classification", *IEEE Journal of Selected Topics in Signal Processing*, Vol. 13, No. 2, pp. 285-297, 2019.

[7] J. W. Jung, H. S. Heo, H. J. Shim, and H. J. Yu, "Knowledge Distillation in Acoustic Scene Classification", *IEEE Access*, Vol. 8, pp. 166870-166879, 2020.

[8] Z. Ren, Q. Kong, J. Han, M. D. Plumbley, and B. W. Schuller, "CAA-Net: Conditional Atrous CNNs with Attention for Explainable Device-robust Acoustic Scene Classification", *IEEE Transactions on Multimedia*, Vol. 23, pp. 4131-4142, 2020.

[9] T. K. Chan, C. S. Chin, and Y. Li, "Semi-Supervised NMF-CNN for Sound Event Detection", *IEEE Access*, Vol. 9, pp. 130529-130542, 2021.

[10] S. Parekn, S. Essid, A. Ozerov, N. Q. K. Duong, P. Perez, and G. Richard, "Weekly Supervised Representation Learning for Audio-Visual Scene Analysis", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 28, pp. 416-428, 2019.

[11] V. Abrol and P. Sharma, "Learning Hierarchy Aware Embedding from Raw Audio for Acoustic Scene Classification", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 28, pp. 1964-1973, 2020.

[12] L. Gao, H. Mi, B. Zhu, D. Feng, Y. Li, and Y. Peng, "An Adversarial Feature Distillation Method for Audio Classification", *IEEE Access*, Vol. 7, pp. 105319-105330, 2019.

[13] H. Wang, Y. Zou, and D. Chong, "Acoustic Scene Classification with Spectrogram Processing Strategies", *arXiv Preprint arXiv:2007*, 2020.

[14] Z. Ren, K. Qian, 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.

[15] L. Zhang, J. Han, and Z. Shi, "Learning Temporal Relations from Semantic Neighbors for Acoustic Scene Classification", *IEEE Signal Processing Letters*, Vol. 27, pp. 950-954, 2020.

[16] A. Rakotomamonjy, "Supervised Representation Learning for Audio Scene Classification", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 25, No. 6, pp. 1253-1265, 2017.

[17] A. Mesaros, T. Heittola, E. Benetos, P. Foster, M. Lagrage, T. Virtanen, and M. D. Plumbley, "Detection and Classification of Acoustic Scenes and Events: Outcome of the DCASE 2016 Challenge", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*", Vol. 26, No. 2, pp. 379-393, 2017.

[18] Y. Chen, Q. Guo, X. Liang, J. Wang, and Y. Qian, "Environmental Sound Classification with Dilated Convolutions", *Applied Acoustics*, Vol. 148, pp. 1-132, 2019.

[19] A. Mesaros, A. Diment, B. Elizalde, T. Heittola, E. Vincent, B. Raj, and T. Virtanen, "Sound Event Detection in the DCASE 2017 Challenge", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 27, No. 6, pp. 992-1006, 2019.

[20] S. Chandrakala and S. L. Jayalakshmi, "Generative Model-Driven Representation Learning in a Hybrid Framework for Environmental Audio Scene and Sound Event Recognition", *IEEE Transactions on Multimedia*, Vol. 22, No. 1, pp. 3-14, 2019.

[21] Z. Ren, Q. Kong, K. Qian, M. D. Plumbley, and B. W. Schuller, "Attention-based Convolutional Neural Networks for Acoustic Scene Classification", In: *Proc. of DCASE 2018 Workshop*, 2018.

[22] M. A. Alamir, "A Novel Acoustic Scene Classification Model using the Late Fusion of Convolutional Neural Networks and Different Ensemble Classifiers", *Applied Acoustics*, Vol. 175, p. 107829, 2021.

[23] T. Zhang, J. Liang, and B. Ding, "Acoustic Scene Classification using Deep CNN with fine-resolution feature", *Expert Systems with Applications*, Vol. 143, p. 113067, 2020.

[24] H. Liang and Y. Ma, "Acoustic Scene Classification using attention-based Convolutional Neural Network", *DCACE2019 Challenge, Tech, Rep.*, 2019.

[25] C. Paseddula and S. V. Gangashetty, "Late Fusion Framework for Acoustic Scene Classification using LPCC, SCMC, and Log-Mel Band Energies with Deep Neural Networks", *Applied Acoustics*, Vol. 172, p. 107568, 2021.

[26] Z. Li, Y. Hou, X. Xie, S. Li, L. Zhang, S. Du, and W. Liu, "Multi-level Attention Model with Deep Scattering Spectrum for Acoustic Scene Classification", In: *Proc. of 2019 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pp. 396-401, 2019.

[27] L. Yang, L. Tao, X. Chen, and X. Gu, "Multi-scale Semantic Feature Fusion and Data Augmentation for Acoustic Scene Classification", *Applied Acoustics*, Vol. 162, p. 107238, 2020.

[28] J. Liang, T. Zhang, and G. Feng, "Channel Compression: Rethinking Information Redundancy Among Channels in CNN Architecture", *IEEE Access*, Vol. 8, pp. 147265-147274, 2020.

[29] D. D. B. Gorron, D. Ramos, and D. T. Toledano, "A Multi-Resolution CRNN-Based Approach for Semi-Supervised Sound Event Detection in DCASE 2020 Challenge", *IEEE Access*, Vol. 9, pp. 89029-98042, 2021.

[30] A. Politis, A. Mesaros, S. Adavanne, T. Heittola, and T. Virtanen, "Overview and Evaluation of Sound Event Localization and Detection in DCASE 2019", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 29, pp. 684-698, 2021.

[31] V. Aborl and P. Sharma, "Learning Hierarchy Aware Embedding from Raw Audio for

Acoustic Scene Classification", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 28, pp. 1964-1973, 2020.

[32] S. Abdoli, P. Cardinal, and A. L, Koerich, "End-to-end Environmental Sound Classification using a 1D Convolutional Neural Network", *Expert Systems with Applications*, Vol. 136, pp. 252-263, 2019.

[33] S. Waldekar and G. Saha, "Classification of Audio Scenes with Novel Features in a Fused System Framework", *Digital Signal Processing*, Vol. 75, pp. 71-82, 2018.

[34] J. Ye, T. Lobayashi, N. Toyama, H. Tsuda, and M. Murakawa, "Acoustic Scene Classification Using Efficient Summary Statistics and Multiple Spectro-Temporal Descriptor Fusion", *Applied Sciences*, Vol. 8, No. 8, p. 1363, 2018.

[35] K. Xu, D. Feng, H. Mi, B. Zhu, D. Wang, L. Zhang, H. Cai, and S. Liu, "Mixup-Based Scene Classification using Multi-Channel Convolutional Neural Network", In: *Proc. of Pacific Rim Conference on Multimedia*, pp. 14-23, 2018.

[36] Z. Ren, Q. Kong, J. Han, M. D. Plumbley, and B. W. Schuller, "Attention-based Atrous Convolutional Neural Networks: Visualization and Understanding Perspectives of Acoustic Scenes", In: *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 56-60, 2019.

[37] U. Kumaran, S. R. Rammohan, S. M. Nagarajan, and A. Prathik, "Fusion of Mel and Gammatone Frequency Cepstral Coefficient for Speech Emotion Recognition using Deep C-RNN", *International Journal of Speech Technology*, Vol. 24, No. 2, pp. 303-314, 2021.