2021 6th International Conference for Convergence in Technology (I2CT)
Pune, India. Apr 02-04, 2021

Acoustic Scene Classification using Deep Learning Architectures

Spoorthy. V
Dept. of Computer Science & Engineering
National Institute of Technology Karnataka
Surathkal, India
vspoorthy036@gmail.com

Manjunath Mulimani
Dept. of Computer Science & Technology
Manipal Institute of Technology
Manipal, India
manjunath.gec@gmail.com

Shashidhar G. Koolagudi
Dept. of Computer Science & Engineering
National Institute of Technology Karnataka
Surathkal, India
koolagudi@nitk.edu.in

2021 6th International Conference for Convergence in Technology (I2CT) | 978-1-7281-8876-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/I2CT51068.2021.9418177

***Abstract*—Enabling devices to make sense of sound is known as Acoustic Scene Classification (ASC). The analysis of various scenes by applying computational algorithms is known as computational auditory scene analysis. The main aim of this paper is to classify audio recordings based on the scenes/environment in which they are recorded. Deep learning is amongst the recent trends in most of the applications. In this paper, two deep learning algorithms are used to perform the classification of acoustic scenes, namely Convolution Neural Network (CNN) and Convolution-Recurrent Neural Network (CRNN). The model is evaluated on three activation functions, namely, ReLU, LeakyReLU and ELU. The highest recognition accuracy achieved for ASC task is 90.96% from CRNN model. The model performed well on basic convolution architecture with 10.9% improvement from the baseline system of this task.**

***Index Terms*—Acoustic Scene Classification, Convolution Neural Network, Convolution Recurrent Neural Network, Activation**

I. Introduction

A vast amount of information related to multimedia is easily accessible these days. Environmental sound detection has grabbed the attention of the researchers in the recent days [1], [2], [3], [4], [5], [6]. Assigning a semantic label for an audio segment based on the surrounding it has been recorded is named as ASC. This can also be said as making sense to the sounds or providing context to the environmental sound to make smarter devices. ASC plays an important role in the current generation where every device is automated. This has many real-world applications such as smart homes, robotics, audio surveillance, context-aware mobile devices, music genre classification etc. One of the important applications is detecting odd activities such as crying or shouting in pain in an indoor environment or shot from a gun etc. This can be done using audio surveillance that employs sound content analysis techniques for the detection of outliers [9], [10].

There has always been a clear difference made between psycho-acoustic studies by the use of various automated techniques to perform the ASC task. Some of the techniques are signal processing methods, recent computational methods such as conventional/traditional machine learning and most popular deep learning methods. This also enabled humans to understand the cognitive processes in acoustic scenes [15].

An acoustic scene consists of multiple sounds or events. For example, consider a scenario at an airport, there are multiple sound events such as airline announcements, people chattering, phone ring, children playing etc. Human auditory system can exceptionally differentiate between these sounds individually or in a set of events. This is known as Auditory Scene Analysis (ASA) [11]. This can be automated by making the machines learn different real-world scenarios which is known as Computational Auditory Scene Analysis (CASA) [12].

This paper uses two deep learning architectures are used to perform classification of different acoustic scenes. The models considered are CNN and CRNN. The two deep learning models are evaluated on three activation functions, Rectified Linear Unit (ReLU), Leaky Rectified Linear Unit (LeakyReLU), and Exponential Linear Unit (ELU). The performance of the model varied in each activation function. The performance of the models is better than most of the previously proposed deep learning architectures.

The rest of the paper is organized as follows: Section II discusses the previously used methods for ASC task. The details regarding the dataset used in this work given in section III. Section IV describes the proposed methodology. Results obtained from the proposed approach are presented in section V. The concluding statements of this work is provided in section VI.

II. Related Work

Based on the specific event cues, humans can differentiate acoustic scenes [11]. The previous works of ASC mainly focus on spectral features. These features are mostly used for speech related tasks such as speech recognition, speaker verification, speaker recognition, etc. The spectral features that are most commonly used are spectral centroid, zero-crossing rate, Mel Frequency Cepstral Coefficients (MFCCs), linear predictive coefficients [18], [19]. The most commonly used feature is MFCC. Cepstral coefficients are generated by performing short-time spectral analysis on audio segments. The common approach used for baseline systems for ASC task is the combination of Gaussian Mixture Model (GMM) with MFCC features [20], [21]. In [19], a MFCC features were encoded from a Recurrence Quantification Analysis (RQA) parameters, which were fed to Support Vector Machine (SVM) classifier. The approach used by [19] outperformed all the methods in the Detection and Classification of Acoustic Scences and Events

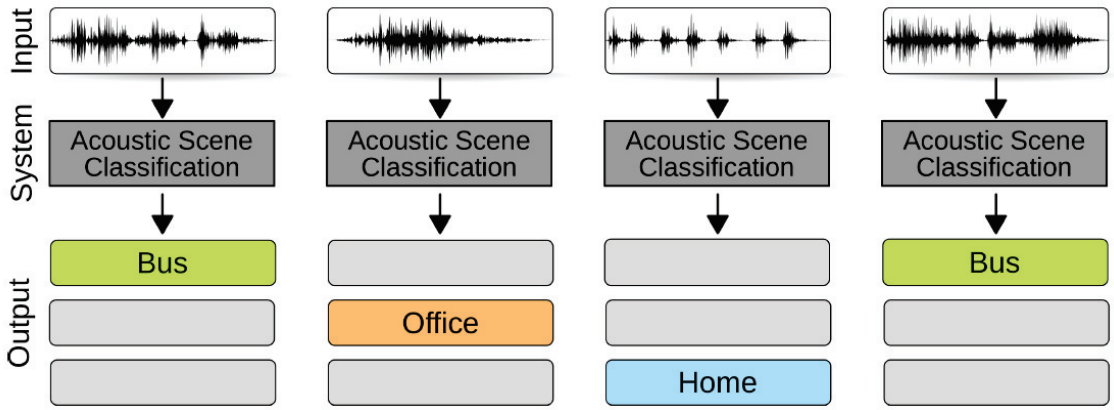

Fig. 1. Illustration of Acoustic scene Classification [17]

(DCASE) 2013 challenge. The combination of MFCC-GMM model was used in [13] which exhibited better performance in classifying various acoustic scenes. Here, frame-level spectral features are provided as an input to the SVM classifier, which performed better than other systems submitted to DCASE 2016 challenge. Hence, it can be stated that discriminative information is provided by Time-Frequency Representation (TFR) of the signal to differentiate various acoustic scenes.

The recent popularity of deep learning has drawn attention of researchers of machine listening community as well. The baseline system of DCASE 2018 and 2019 challenge also includes a CNN model [22]. One of the efficient and popular CNN architectures is Residual Network or ResNet for ASC task [23], [24]. Log-mel spectrograms along with log-mel deltas and delta-deltas were used for training a deep residual network in [23]. The network was trained in a two-way path, one for higher frequencies and other one for lower frequencies. The two-way path was combined at the last layer of the network. An improvement in the performance was noticed when the log-mel deltas and delta-deltas are combined. However, the model performance was reduced when the audio input from another recording device is provided to the network. This is caused due to the imbalance of audio recordings from different recording devices in the dataset. A fusion system by combining three deep learning architectures is proposed in [14]. The architectures are VGG-like two dimensional CNN, Max-feature Map activation called light CNN and an one-dimensional CNN called x-vector topology. The combination of these three models resulted in an accuracy of 77%. The limitation of this approach is that the training of this model can be computationally expensive. In this work, we used two basic CNN architectures for performing classification of acoustic scenes which exploits the characteristics of various activation function to train a CNN.

III. Dataset

The dataset used to develop the proposed method is DCASE 2019 Acoustic Scene Classification Task 1a development dataset [22]. The development dataset consists of ten acoustic scenes, namely, airport, bus, metro, metro station, public square, park, shopping mall, street traffic, street pedestrian, and tram. The number of audio recordings present in the dataset are 14400. The audio recordings are balanced across all the classes, where each class consists of 1440 audio recordings. The audio segments are sampled at a rate of 48000 Hz and 24-bit resolution. The data was collected from different cities of Europe.

IV. Proposed Methodology

A. Overview

A general architecture of an acoustic scene classification is shown in Figure I. In this work, classification of acoustic scenes is performed using two deep learning models, CNN and CRNN. The extraction of feature representations is performed on audio segments and given as input to the deep learning models. The model is tested on three activation functions. The detailed explanation is provided in the following sections.

B. Feature Extraction

The feature extracted from the audio recordings is log-mel spectrogram. The duration of acoustic scenes is short, therefore to achieve more distinct information, it is necessary to represent the segments in Time-Frequency Representations (TFRs). A conventional Short-Time Fourier Transform (STFT) is performed on the audio segments before the extraction of log-mel spectrogram features. The window length of set to 40 milliseconds with an overlap of 50%. The window function used is “hamming asymmetric”. The number of Mel-bands set is 40. Features considered for the proposed approach are taken from the DCASE 2019 challenge baseline model [8]. The representation of the features extracted for some classes is presented in Figure 2.

C. Activation Functions

In this work, three activation functions are used in CNN and CRNN models. The activation function is introduced to get non-linear output and to remove negative values to pass to the next layer in the network. The performance of the neural

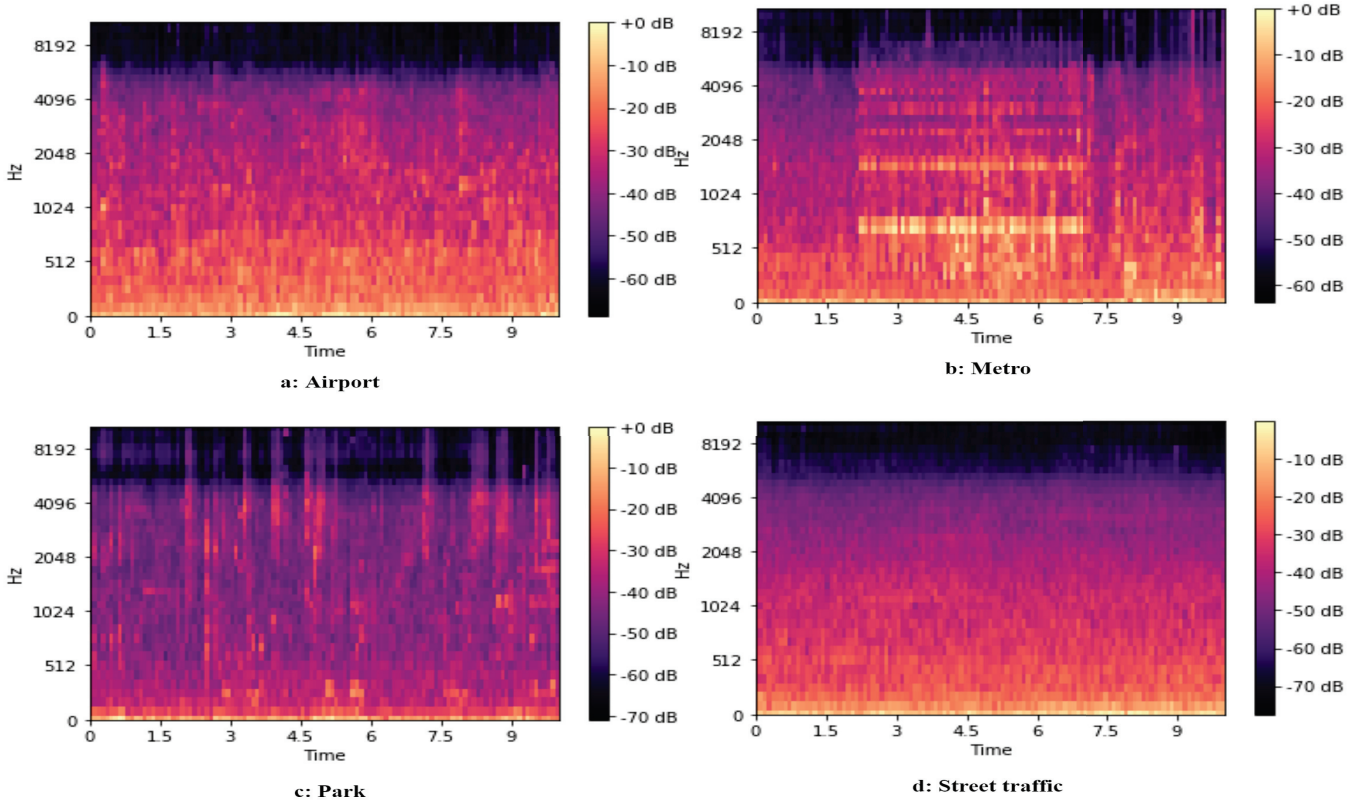

Fig. 2. Mel-spectrogram features for various acoustic scenes

network architecture is highly dependent on the activation function. There are many linear and non-linear activation functions that are used in the previous CNN architectures. Some of them are linear, sigmoid, tanh, ReLU, LeakyReLU, softmax, and ELU. For the models, ReLU, LeakyReLU and ELU activation functions are used for hidden layers and softmax for the last layer of the deep learning models.

The ReLU activation function is a recently introduced non-linear activation function, also known as rectified linear units [25]. The ReLU function is given in Equation 1. The advantage of using this activation is it is computationally less expensive as compared to sigmoid and tanh activation functions.

$$R(x) = \begin{Bmatrix} x, x > 0 \\ 0, x <= 0 \end{Bmatrix} \tag{1}$$

The LeakyReLU activation function is a variant of ReLU function [26]. In this function, instead of being 0, LeakyReLU allows a small non-zero constant gradient α. Usually, the α value given is 0.01. This activation function was introduced to resolve the issue of "dying ReLU" by keeping a small negative slope. The LeakyReLU activation function is given in Equation 2.

$$R(x) = \begin{Bmatrix} x, x > 0 \\ \alpha x, x <= 0 \end{Bmatrix} \tag{2}$$

The ELU activation function or widely known as Exponential Linear Unit is a function that tends to converge cost to zero to provide accurate results [27]. This activation function is different from other activation function because it has a extra alpha constant which must be a positive number. The functionality of both ELU and ReLU is similar except that ELU takes input with negative values too. The smoothening of ELU happens slowly till the output is equal the -α. However, in ReLU, the smoothening happens sharply. ELU activation function is given in Equation 3.

$$R(x) = \begin{Bmatrix} x, x > 0 \\ \alpha(e^x - 1), x <= 0 \end{Bmatrix} \tag{3}$$

D. Classification

The classification of various acoustic scenes is performed using two deep learning architectures, CNN and CRNN. Both models use the loss function categorical cross entropy. As multi-class classification is been performed, categorical cross entropy loss function is best suited. Adam optimizer is chosen for training the network. The learning rate value is set to 0.001. The models are trained on 80% of the entire data and tested on 20% of the entire data. The details of the model architectures are given below.

1) CNN: The CNN architecture used for ASC task consists of convolution and max-pooling operations. The hidden layer activation functions are ReLU, LeakyReLU and ELU. The convolution blocks has varied filter and kernels. A batch normalization layer is introduced in the network to avoid

the bias that occurs in values. Flatten layer is used in the network to flatten the multiple convolution blocks and bring it to a single dimension. After flattening, two dense layers are used in the network. The last dense layer must contain dense operations which equals the classes in the dataset. A dropout layer is introduced in the network to avoid overfitting of the model. The last layer is the activation layer. Here, softmax activation function is used. The model architecture of CNN is shown in Figure 3.

Layer (type)	Output Shape	Param #
conv2d_6 (Conv2D)	(None, 38, 498, 64)	640
leaky_re_lu_6 (LeakyReLU)	(None, 38, 498, 64)	0
batch_normalization_5 (Batch	(None, 38, 498, 64)	256
average_pooling2d_4 (Average	(None, 19, 249, 64)	0
conv2d_7 (Conv2D)	(None, 18, 248, 72)	18504
leaky_re_lu_7 (LeakyReLU)	(None, 18, 248, 72)	0
batch_normalization_6 (Batch	(None, 18, 248, 72)	288
average_pooling2d_5 (Average	(None, 9, 124, 72)	0
conv2d_8 (Conv2D)	(None, 8, 123, 128)	36992
leaky_re_lu_8 (LeakyReLU)	(None, 8, 123, 128)	0
batch_normalization_7 (Batch	(None, 8, 123, 128)	512
conv2d_9 (Conv2D)	(None, 7, 122, 254)	130302
leaky_re_lu_9 (LeakyReLU)	(None, 7, 122, 254)	0
average_pooling2d_6 (Average	(None, 3, 61, 254)	0
conv2d_10 (Conv2D)	(None, 3, 61, 254)	64770
leaky_re_lu_10 (LeakyReLU)	(None, 3, 61, 254)	0
batch_normalization_8 (Batch	(None, 3, 61, 254)	1016
flatten_2 (Flatten)	(None, 46482)	0
dense_4 (Dense)	(None, 100)	4648300
dropout_2 (Dropout)	(None, 100)	0
dense_5 (Dense)	(None, 50)	5050
dense_6 (Dense)	(None, 10)	510

Total params: 4,907,140
Trainable params: 4,906,104
Non-trainable params: 1,036

Fig. 3. Model architecture of CNN

2) CRNN: CRNN model used in this work consists of convolution and max-pooling blocks followed by activation layer, batch normalization and drop out layer. Here, permute and reshape layers are also necessary. Because, the feature vectors vary in CNN and CRNN models. The feature dimension in a CNN is three-dimensional whereas in a CRNN model it is two-dimensional. To change the direction of the axes of feature vectors permute layers are used. This is followed by a reshape layer which converts the feature vector to a two-dimensional feature vector. In the proposed network, two Gated Recurrent Units (GRUs) are used. These GRU layers takes the past timestamps into the account. Finally, the output of the bidirectional layers is fed to the dense layers which is given as fully connected layer's input. Overfitting of the model is avoided by introducing a dropout layer to the network. The model architecture is shown in Figure 4.

Layer (type)	Output Shape	Param #
batch_normalization_6 (Batch	(None, 40, 500, 1)	160
conv2d_5 (Conv2D)	(None, 40, 500, 64)	640
activation_6 (Activation)	(None, 40, 500, 64)	0
batch_normalization_7 (Batch	(None, 40, 500, 64)	256
max_pooling2d_5 (MaxPooling2	(None, 20, 250, 64)	0
dropout_6 (Dropout)	(None, 20, 250, 64)	0
conv2d_6 (Conv2D)	(None, 20, 250, 128)	73856
activation_7 (Activation)	(None, 20, 250, 128)	0
batch_normalization_8 (Batch	(None, 20, 250, 128)	512
max_pooling2d_6 (MaxPooling2	(None, 10, 125, 128)	0
dropout_7 (Dropout)	(None, 10, 125, 128)	0
conv2d_7 (Conv2D)	(None, 10, 125, 128)	147584
activation_8 (Activation)	(None, 10, 125, 128)	0
batch_normalization_9 (Batch	(None, 10, 125, 128)	512
max_pooling2d_7 (MaxPooling2	(None, 5, 62, 128)	0
dropout_8 (Dropout)	(None, 5, 62, 128)	0
conv2d_8 (Conv2D)	(None, 5, 62, 256)	295168
activation_9 (Activation)	(None, 5, 62, 256)	0
batch_normalization_10 (Batc	(None, 5, 62, 256)	1024
max_pooling2d_8 (MaxPooling2	(None, 2, 31, 256)	0
dropout_9 (Dropout)	(None, 2, 31, 256)	0
permute_2 (Permute)	(None, 31, 2, 256)	0
reshape_2 (Reshape)	(None, 31, 512)	0
gru_3 (GRU)	(None, 31, 32)	52320
gru_4 (GRU)	(None, 32)	6240
dropout_10 (Dropout)	(None, 32)	0
dense_2 (Dense)	(None, 10)	330
activation_10 (Activation)	(None, 10)	0

Total params: 578,602
Trainable params: 577,370
Non-trainable params: 1,232

Fig. 4. Model architecture of CRNN

V. RESULTS AND DISCUSSION

The classification of various acoustic scenes is a challenging problem. Each acoustic scene consists of multiple acoustic events. Example of an acoustic event can be phone ring or

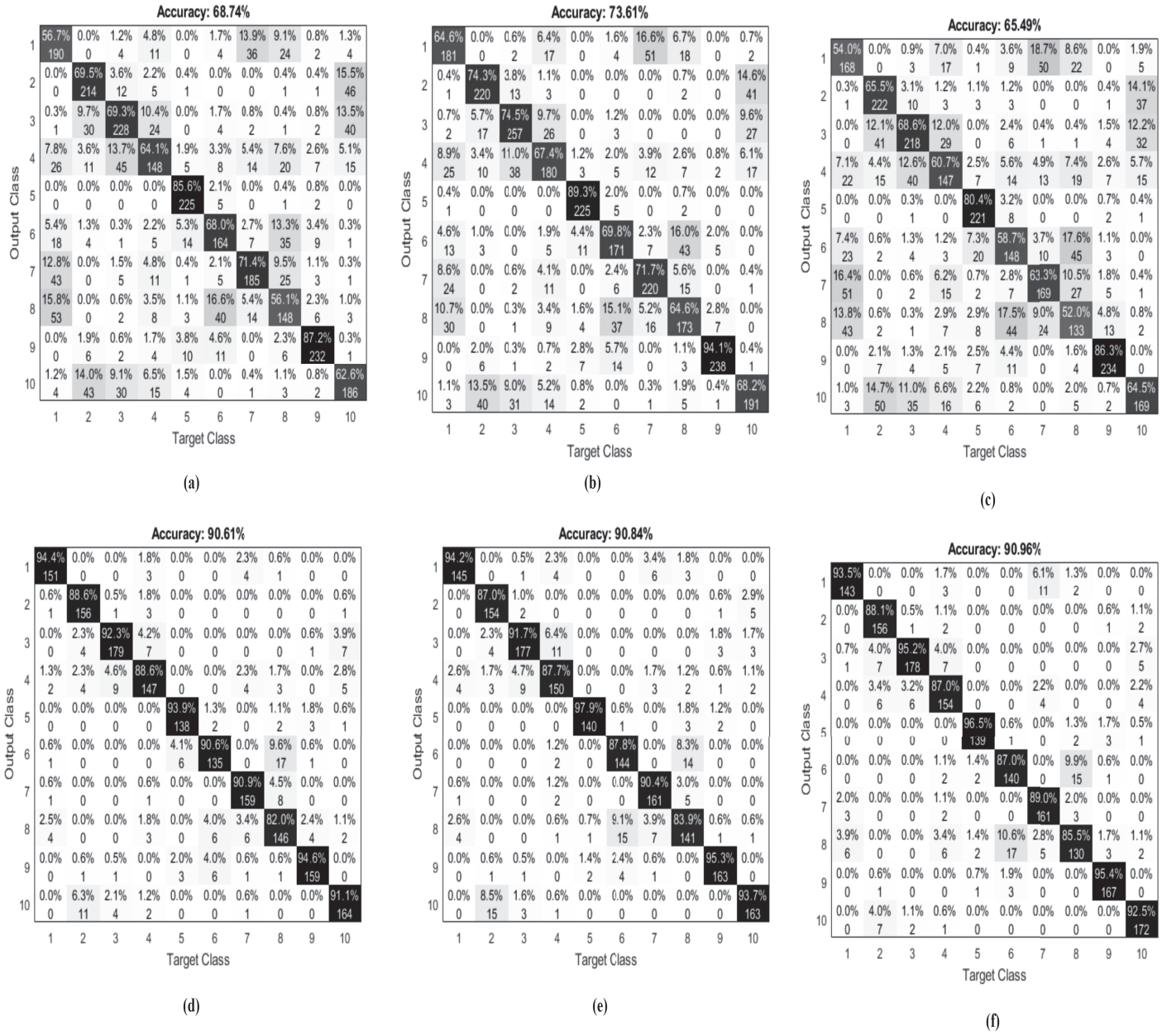

Fig. 5. Confusion matrices obtained for CNN and CRNN models. (a) CNN with ReLU, (b) CNN with LeakyReLU, (c) CNN with ELU, (d) CRNN with ReLU, (e) CRNN with LeakyReLU, and (e) CRNN with ELU

announcement in the airport etc. There can be overlapping of multiple acoustic events in an acoustic scene and two or more scenes may have same acoustic events. For example, let us take bus and train scenes. Here, the some of the acoustic events such as people chatting, phone ring, announcement of stops etc, are same. This overlapping of events make it difficult to discriminate between the different acoustic scenes. In this work, as mentioned earlier, two deep learning models, CNN and CRNN are used for classification of acoustic scenes. The performance of the CNN and CRNN models is illustrated in Figure 5. The Figure demonstrates confusion matrices obtained for two models with different activation functions. It can be observed from the Figure 5 that highest recognition accuracy for ASC task is obtained is 90.96% from CRNN model with ELU activation function. The highest recognition for CNN model is obtained is 73.61% for LeakyReLU function.

The performance of the models is compared with the baseline system [22]. The accuracy achieved is demonstrated in Table I. From the table, it can be observed that the CRNN model performed exceptionally well for the mel-spectrogram feature representation.

TABLE I
COMPARISON OF THE PERFORMANCE OF THE PROPOSED MODEL AND BASELINE SYSTEM OF DCASE 2019 CHALLENGE TASK 1A ASC DATASET

Model	**Accuracy**
Baseline model	62.5%
CNN	73.61%
CRNN	90.96%

It can be stated that activation function and the depth of the neural network architecture used play an important role in the performance of the network. CRNN model achieved better results as compared to most of the previously proposed systems of ASC tasks. The use of deep learning architectures for ASC task looks promising, as it provides discriminative information.

VI. CONCLUSION

In this work, two deep learning architectures with combination of multiple activation functions is used to perform classification of acoustic scenes. There is an assumption that increase in the depth of the neural network provides better performance of the system. That is true to only one extent. The model's performance depends on the activation function chosen in the hidden layers of the network. Three activation functions, ReLU, LeakyReLU and ELU has been used in this work. The highest recognition accuracy is achieved is 90.96% from the CRNN model with ELU activation function.

In future work, different deep learning architectures and feature representations can be used to improve the performance of the ASC system.

REFERENCES

[1] D. Barchiesi, D. Giannoulis, D. Stowell and M. D. Plumbley, "Acoustic Scene Classification: Classifying environments from the sounds they produce," in IEEE Signal Processing Magazine, vol. 32, no. 3, pp. 16-34, May 2015, doi: 10.1109/MSP.2014.2326181.

[2] A. Mesaros, T. Heittola and T. Virtanen, "TUT database for acoustic scene classification and sound event detection," 2016 24th European Signal Processing Conference (EUSIPCO), Budapest, 2016, pp. 1128-1132, doi: 10.1109/EUSIPCO.2016.7760424.

[3] J. T. Geiger, B. Schuller and G. Rigoll, "Large-scale audio feature extraction and SVM for acoustic scene classification," 2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, NY, 2013, pp. 1-4, doi: 10.1109/WASPAA.2013.6701857.

[4] Heittola, T., Mesaros, A., Eronen, A. et al. Context-dependent sound event detection. J AUDIO SPEECH MUSIC PROC. 2013, 1 (2013). https://doi.org/10.1186/1687-4722-2013-1

[5] E. Cakir, T. Heittola, H. Huttunen and T. Virtanen, "Polyphonic sound event detection using multi label deep neural networks," 2015 International Joint Conference on Neural Networks (IJCNN), Killarney, 2015, pp. 1-7, doi: 10.1109/IJCNN.2015.7280624.

[6] G. Parascandolo, H. Huttunen and T. Virtanen, "Recurrent neural networks for polyphonic sound event detection in real life recordings," 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, 2016, pp. 6440-6444, doi: 10.1109/ICASSP.2016.7472917.

[7] Hershey, S., Chaudhuri, S., Ellis, D.P., Gemmeke, J.F., Jansen, A., Moore, R.C., Plakal, M., Platt, D., Saurous, R.A., Seybold, B. and Slaney, M., "CNN architectures for large-scale audio classification," 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, 2017, pp. 131-135, doi: 10.1109/ICASSP.2017.7952132.

[8] Q. Kong, Y. Xu, W. Wang and M. D. Plumbley, "A joint detection-classification model for audio tagging of weakly labelled data," 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, 2017, pp. 641-645, doi: 10.1109/ICASSP.2017.7952234.

[9] R. Radhakrishnan, A. Divakaran and A. Smaragdis, "Audio analysis for surveillance applications," IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2005., New Paltz, NY, 2005, pp. 158-161, doi: 10.1109/ASPAA.2005.1540194.

[10] A. Harma, M. F. McKinney and J. Skowronek, "Automatic surveillance of the acoustic activity in our living environment," 2005 IEEE International Conference on Multimedia and Expo, Amsterdam, 2005, pp. 4 pp.-, doi: 10.1109/ICME.2005.1521503.

[11] Bregman, A. S. (1990). Auditory scene analysis: The perceptual organization of sound. The MIT Press.

[12] Brown, Guy Jason (1992) Computational auditory scene analysis : a representational approach. PhD thesis, University of Sheffield.

[13] Waldekar, Shefali & Saha, Goutam. (2018). Classification of audio scenes with novel features in a fused system framework. Digital Signal Processing. 75. 10.1016/j.dsp.2017.12.012.

[14] Zeinali, H., Burget, L. and Černocký, J., 2019. Acoustic scene classification using fusion of attentive convolutional neural networks for DCASE2019 challenge. arXiv preprint arXiv:1907.07127.

[15] McAdams, S. (1993). Recognition of sound sources and events. In S. McAdams & E. Bigand (Eds.), Oxford science publications. Thinking in sound: The cognitive psychology of human audition (p. 146–198). Clarendon Press/Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198522577.003.0006

[16] SCHAFER, R. M. (1977). The tuning of the world. New York, A.A. Knopf.

[17] A. Mesaros, T. Heittola and T. Virtanen, "Assessment of human and machine performance in acoustic scene classification: Dcase 2016 case study," 2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), New Paltz, NY, 2017, pp. 319-323, doi: 10.1109/WASPAA.2017.8170047.

[18] S. Chu, S. Narayanan and C. -. J. Kuo, "Environmental Sound Recognition With Time–Frequency Audio Features," in IEEE Transactions on Audio, Speech, and Language Processing, vol. 17, no. 6, pp. 1142-1158, Aug. 2009, doi: 10.1109/TASL.2009.2017438.

[19] G. Roma, W. Nogueira and P. Herrera, "Recurrence quantification analysis features for environmental sound recognition," 2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, NY, 2013, pp. 1-4, doi: 10.1109/WASPAA.2013.6701890.

[20] D. Giannoulis, D. Stowell, E. Benetos, M. Rossignol, M. Lagrange and M. D. Plumbley, "A database and challenge for acoustic scene classification and event detection," 21st European Signal Processing Conference (EUSIPCO 2013), Marrakech, 2013, pp. 1-5.

[21] A. Mesaros, T. Heittola and T. Virtanen, "TUT database for acoustic scene classification and sound event detection," 2016 24th European Signal Processing Conference (EUSIPCO), Budapest, 2016, pp. 1128-1132, doi: 10.1109/EUSIPCO.2016.7760424.

[22] Mesaros, A., Heittola, T. and Virtanen, T., 2018. A multi-device dataset for urban acoustic scene classification. arXiv preprint arXiv:1807.09840.

[23] Mesaros, Annamaria & Heittola, Toni & Diment, Aleksandr & Elizalde, Benjamin & Shah, Ankit & Vincent, Emmanuel & Raj, Bhiksha & Virtanen, Tuomas. (2017). DCASE 2017 CHALLENGE SETUP: TASKS, DATASETS AND BASELINE SYSTEM.

[24] Lehner, Bernhard & Koutini, Khaled & Schwarzlmüller, Christopher & Gallien, Thomas & Widmer, Gerhard. (2019). ACOUSTIC SCENE CLASSIFICATION WITH REJECT OPTION BASED ON RESNETS.

[25] Glorot, X., Bordes, A. & Bengio, Y.. (2011). Deep Sparse Rectifier Neural Networks. Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, in PMLR 15:315-323

[26] He, K., Zhang, X., Ren, S. and Sun, J., 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In Proceedings of the IEEE international conference on computer vision (pp. 1026-1034).

[27] Clevert, D.A., Unterthiner, T. and Hochreiter, S., 2015. Fast and accurate deep network learning by exponential linear units (elus). arXiv preprint arXiv:1511.07289.