A Review Paper on Acoustic Scene Classification

N.V.R. Vikram G[1], Shaik Jakeer Hussain[2], M. Pachiyannan[3], M.Sarada[4]

[1,2] Professor, Dept. of ECE, Vignan's Nirula Institute of Technology and Science for Women, Pedapalakaluru, Guntur, A.P, India

[3,4] Associate Professor, Dept. of ECE, Vignan's Nirula Institute of Technology and Science for Women, Pedapalakaluru, Guntur, A.P, India

(E-mail: ece.6891@gmail.com)

***Abstract*—** Acoustic Scene Classification refers to the ability of a person or artificial system to comprehend an auditory context, whether from a recording or an online stream. Humans frequently refer to an acoustic environment as "context" or "scene," which refers to the collection of ambient sounds and sound occurrences we connect with a particular auditory scenario, such as a restaurant or park. Humans may perceive this as a straightforward task since our brains are capable of performing complex computations and because of our rich life experiences, we are able to quickly link certain sound groups to particular events. However, for artificial systems, this task is not simple. Deep learning is a subfield of ML that uses algorithms called artificial neural networks which are inspired by the structure and function of the brain and are capable of self-learning. ANNs are trained to "learn" models and patterns rather than being explicitly told how to solve a problem. This work presents acoustic scene classification using deep learning.

***Keywords*—** *Acoustic scene classification, Deep Learning, Artificial neural networks.*

I. INTRODUCTION

The objective of acoustic classification is to assign test audio recordings to one of the offered predefined classes that best describes the setting in which they were made. The DCASE Challenge's popular acoustic scene classification problem introduces fresh iterations of a supervised classification task every year. This assignment has consistently drawn the most participants out of all the ones that are available in past editions. This task of acoustic scene classification is defined as classifying a short excerpt of audio into a class of a predefined set of classes that indicates the context where the audio was recorded [1].

The detection of audio events that are momentarily present in an acoustic scene is a particularly difficult ASC problem. Vehicles, car horns, and footfall are a few examples of these audible events. Acoustic event detection (AED) is the name of this task, which significantly varies from ASC in that it focuses on the accurate temporal recognition of specific sound occurrences. Modern ASC systems have been demonstrated to do better on this job than humans [2]. They are used in a variety of application scenarios as a result, including context-aware wearables and hearables, hearing aids, healthcare, security surveillance, monitoring wildlife in natural habitats, smart cities, the Internet of Things, and autonomous navigation.

A classification approach for acoustic scenes is expected to function in a wide range of situations for real-world applications, including audio acquired with various devices and the quickest possible inference time.

Since the beginning of the DCASE Challenge, the challenge has been one of the primary topics, and it has evolved from the initial setup to incorporate several new issues, such as multiple devices and low-complexity circumstances [3]. By specifying the low-complexity limitations, such as the limited number of parameters and the maximum number of operations permitted at inference time, typical of contemporary IoT devices, the current configuration moves closer to real-world applicability..

II. BASIC METHODOLOGY

Most neural network architectures applied for ASC require multi-dimensional input data. The most commonly used time-frequency transformations are the Mel Spectogram[4]. The Mel spectrogram is based on a non-linear frequency scale motivated by human auditory perception and provides a more compact spectral representation of sounds. ASC algorithms process only the magnitude of the Fourier transform while the phase is discarded.

TAU Urban Acoustic Scenes, a recently released update of the earlier acoustic scene datasets, is the dataset used for the job. The data is made up of recordings taken in ten acoustic locations that correspond to the target classifications [5]: an airport, an indoor retail centre, a metro station, a pedestrian street, a public square, a roadway with medium traffic, a tram, a bus, an underground metro, and an urban park. Several European cities recorded data; ten of these recordings are included in the training set and twelve of them are available in the assessment set (two new cities compared to the training).

The audio files were simultaneously recorded with the four devices A, B, C, and D, and 11 more devices S1–S11 were simulatively recorded utilising the audio from device A. The data for the development and assessment sets total 64 and 22 hours, respectively. We direct the reader to [6] for comprehensive information on the dataset creation and the precise data amounts per device. The audio data for this edition is different from the prior datasets.is delivered in 1-second increments in order to adhere to the inference time and computational constraints imposed by the target devices under consideration.

The entries were assessed using accuracy and multi-class cross entropy. Because the data is balanced, accuracy was determined as the macro-average (the average of the performance for each metric per class), which corresponds to the overall accuracy.

In order to rank the systems independently of the operating point, the multi-class cross-entropy (log loss) was used. As with every challenge edition, the audio content included in the assessment data was made available two weeks before the challenge's due date. For the provided audio material, the participants were asked to make class predictions. They were also asked to submit the system's output for evaluation along with more details regarding the methodologies. Only task organisers have access to the reference annotation of the assessment data, which was utilized to grade the submissions. The requirements are modelled after Cortex-M4 devices, such as the STM32L496@80MHz or the Arduino Nano 33@64MHz, and the computational complexity is assessed in terms of parameter count and MMACs (million multiply-accumulate operations).

Based on the target device class's computing power, 30 MMACs is the maximum number of MACS per inference. If the most often used features fall within this limit, this restriction simulates the fitting of audio buffers into SRAM (quick access internal memory) on the target device for the analysis segment of 1 s and leaves some headroom for feature calculation (e.g. FFT). The network used to produce learnt features (embeddings), such as VGGish [7], OpenL3 [8], or EdgeL3 [9], adds to the size and complexity of the entire model. Participants must include a technical report with their proposal that contains comprehensive information about the model's size and complexity. A script for determining the number of parameters has been created to make it easier for challenge participants to determine the model size. Additionally, MMACs are offered for the Keras, TFLite, and PyTorch models.

III. Review Of Different Papers

D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M. D. Plumbley, et al., [10] This paper describes a newly-launched public evaluation challenge on acoustic scene classification and detection of sound events within a scene. Systems dealing with such tasks are far from exhibiting human-like performance and robustness. Undermining factors are numerous: the extreme variability of sources of interest possibly interfering, the presence of complex background noise as well as room effects like reverberation. The proposed challenge is an attempt to help the research community move forward in defining and studying the aforementioned tasks. Apart from the challenge description, this paper provides an overview of systems submitted to the challenge as well as a detailed evaluation of the results achieved by those systems.

A.Mesaros, T. Heittola, and T. Virtanen et al.,[11] This paper introduces the acoustic scene classification task of DCASE 2018 Challenge and the TUT Urban Acoustic Scenes 2018 dataset provided for the task, and evaluates the performance of a baseline system in the task. As in previous years of the challenge, the task is defined for classification of short audio samples into one of predefined acoustic scene classes, using a supervised, closed-set classification setup. The newly recorded TUT Urban Acoustic Scenes 2018 dataset consists of ten different acoustic scenes and was recorded in six large European cities, therefore it has a higher acoustic variability than the previous datasets used for this task, and in addition to high-quality binaural recordings, it also includes data recorded with mobile devices. We also present the baseline system consisting of a convolutional neural network and its performance in the subtasks using the recommended cross-validation setup.

(DCASE2018), November 2018.et al.,[12] This paper introduces the acoustic scene classification task of DCASE 2018 Challenge and the TUT Urban Acoustic Scenes 2018 dataset provided for the task, and evaluates the performance of a baseline system in the task. As in previous years of the challenge, the task is defined for classification of short audio samples into one of predefined acoustic scene classes, using a supervised, closed-set classification setup. The newly recorded TUT Urban Acoustic Scenes 2018 dataset consists of ten different acoustic scenes and was recorded in six large European cities, therefore it has a higher acoustic variability than the previous datasets used for this task, and in addition to high-quality binaural recordings, it also includes data recorded with mobile devices. We also present the baseline system consisting of a convolutional neural network and its performance in the subtasks using the recommended cross-validation setup.

A. Mesaros, T. Heittola, and T. Virtanen, et al.,[13] We present an overview of the challenge entries for the Acoustic Scene Classification task of DCASE 2017 Challenge. Being the most popular task of the challenge, acoustic scene classification entries provide a wide variety of approaches for comparison, with a wide performance gap from top to bottom. Analysis of the submissions confirms once more the popularity of deep-learning approaches and mel frequency representations. Statistical analysis indicates that the top ranked system performed significantly better than the others, and that combinations of top systems are capable of reaching close to perfect performance on the given data.

A. Mesaros, T. Heittola, E. Benetos, P. Foster, M. Lagrange, T. Virtanen, and M. D. Plumbley, et al.,[14] Public evaluation campaigns and datasets encourage ongoing research in the targeted fields and enable direct algorithm comparison. The opportunity to develop cutting-edge techniques was provided by the second iteration of the challenge on the detection and classification of acoustic scenes and events (DCASE 2016), which was successful in bringing together a sizable number of participants from both academic and industrial backgrounds. We discuss the tasks and results of the DCASE 2016 challenge in this essay. Acoustic scene categorization, sound event identification in synthetic audio, sound event detection in real-life audio, and domestic audio tagging were the four tasks that made up the challenge. We thoroughly explain each task and evaluate the systems that were submitted in terms of performance and design.

Gao W, and McDonnell, et al.,[15]This model shows that the performance of their models are significantly enhanced by the use of log-mel deltas, and that overall our approach is capable of training strong single models, without use of any supplementary data from outside the official challenge dataset, with excellent generalization to unknown devices. In particular, our approach achieved second place in 2019 DCASE Task 1b (0.4% behind the winning entry), and the best Task 1B evaluation results (by a large margin of over 5%) on test data from a device not used to record any training data.

Naranjo-Alcazar,S.Perez-Castanos,P.Zuccarello, and M. Cobos, et al.,[16] This paper proposes two brand-new stereo audio representations to boost an ASC framework's correctness. With the Gammatone filter bank, these representations correspond to 3-channel representations of the left channel, right channel, and difference between channels (L R), and with the Mel filter bank, to sources of harmonic, percussive, and difference between channels. Additionally, the two representations are combined to provide a 6-channel audio filter bank with several representations. The suggested CNN is also a residual network that, in a new fashion, uses squeeze-excitation techniques in its residual blocks to compel the network to extract significant characteristics from the audio representation. In order to fulfil the requirements of each subtask, the proposed network is used in each of them with various modifications. The representations are slightly different in Subtask A, though, because stereo audio isn't available there. The overlaps between the two tasks are initially presented in this technical report, after which each task's pertinent revisions are made in a separate part. In all activities, the baselines are surpassed by about 10 percentage points.

M. McDonnell et al.,[17] In Task 1b ("LowComplexity Acoustic Scene Classification") of the DCASE2020 Acoustic Scene Challenge, a submission is described in this technical paper. For this job, solutions had to be limited to having parameters that didn't add up to more than 500 KB. The method used to create the spectrograms from the audio scene files was described in this article. It involved training a deep convolutional neural network so that each convolutional weight was set to one of two values after training and could therefore be stored using a single bit. This method made it possible to train a single 36-layer all-convolutional deep neural network with 3,987,000 binary weights totaling 486.69KB. The model achieved a macro-average accuracy (balanced accuracy score) across the three classes of 96.6±0.5% on the 2020 DCASE Task 1b validation set.

S. Suh, S. Park, Y. Jeong, and T. Lee et al.,[18] In this paper Acoustic Scene Classification systems are described in this technical report for Task 1 of the DCASE2020 challenge. For subtask A, we created a single model called Trident ResNet that was implemented with three parallel ResNets. We have established that this structure is helpful when evaluating samples obtained from obscure or hidden devices, and we have also established that the test split's classification accuracy was 73.7%. For subtask B, we built a model called Shallow Inception with fewer parameters than the CNN of the DCASE baseline system using the Inception module. We have increased the model's accuracy up to 97.6% while reducing the number of parameters thanks to the sparse nature of the Inception module.

S. Abidin et al., [19]This paper presents an approach for acoustic scene classification, which aggregates spectral and temporal features. In order to do this, suggest the first application of the variable-Q transform (VQT) to produce the time-frequency representation for acoustic scene classification. In comparison to the constant-Q transform (CQT) or short time fourier transform, the VQT offers finer control over the resolution and can be modified to better capture acoustic scene information. The adjacent evaluation completed LBP (AECLBP), a local binary pattern variant that is better suited to feature extraction from acoustic time-frequency pictures, is the next variation we adopt. Comparing the application of our findings to the usual CQT with LBP resulted in an improvement of 5.2% on the DCASE 2016 dataset. We surpass one of the top performing systems with a classification accuracy of 85.5% after fusing our suggested AECLBP with HOG characteristics

Daniele Barchiesi et al.,[20]In this article, we present an account of the state of the art in acoustic scene classification (ASC), the task of classifying environments from the sounds they produce. Starting from a historical review of previous research in this area, we define a general framework for ASC and present different implementations of its components. We then describe a range of different algorithms submitted for a data challenge that was held to provide a general and fair benchmark for ASC techniques. The data set recorded for this purpose is presented along with the performance metrics that are used to evaluate the algorithms and statistical significance tests to compare the submitted methods.

C Paseddula, SV Gangashetty et al.,[21]In this paper they have experimented on Detection and Classification of Acoustic Scenes and Events (DCASE) 2018 development dataset and DCASE 2017 dataset. We carried out experiments with individual feature sets, and also performed decision level DNN score fusions for improving the performance.

Waldekar Shefali et al.,[22]Systems have approached the issue of ASC from several angles thanks to the problems posed by the Detection and Classification of Acoustic Scenes and Events (DCASE) project. Some of them might produce outcomes that are superior to those of the baseline Mel Frequency Cepstral Coefficients - Gaussian Mixture Model (MFCC-GMM) system. However, a collective decision from all participating systems was found to surpass the accuracy obtained by each system. The simultaneous use of various approaches can exploit the discriminating information in a better way for audio collected from different environments covering audible-frequency range in varying degrees. In this work, we show that the frame-level statistics of some well-known spectral features when fed to Support Vector Machine (SVM) classifier individually, are able to outperform the baseline system of DCASE challenges.

A. Mesaros, T. Heittola et al., [23]T. Virtanen This paper discusses the DCASE 2018 Challenge's acoustic scene classification problem, describes the dataset for the work—TUT Urban Acoustic Scenes 2018—and assesses a baseline system's performance. As in prior challenges, the task is characterised as the supervised, closed-set categorization of brief audio samples into one of established acoustic scene classes. The newly recorded TUT Urban Acoustic Scenes 2018 dataset, which consists of ten distinct acoustic scenes and was

TABLE 1: COMPARISION OF VARIOUS REVIEW PAPERS

Author	Year	Paper Name	Technique	Result
D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M. D. Plumbley	2015	Detection and classification of acoustic scenes and events	Computational auditory scene analysis	The results draw some interesting conclusions that for scene classification, although simple systems can do relatively well the improvement that more complex systems achieve can bring performance to the levels achieved by human listeners
A. Mesaros, T. Heittola, and T. Virtanen	2018	A multi-device dataset for urban acoustic scene classification	Multi device data set	This model produces a higher acoustic variability than the previous datasets used for this task.
A. Mesaros, T. Heittola, and T. Virtanen	2018	Acoustic scene classification: an overview of dcase 2017 challenge entries	Deep-learning approaches and mel frequency representations	The results indicated that the Combinations of top systems are Capable of reaching close to Perfect performance on the given Data.
A. Mesaros, T. Heittola, E. Benetos, P. Foster, M. Lagrange, T. Virtanen, and M. D. Plumbley	2018	Detection and classification of acoustic scenes and events: Outcome of the dcase 2016 challenge	Audio,speech and language processing	The selected tasks represent a good characterization of current interest, from the more general acoustic scene classification and audio tagging topics, to the detailed temporal detection of individual sound events
Gao W, and McDonnel	2019	Acoustic scene classification using deep residual networks with late fusion of separated high and low frequency paths	Residual networks	In particular, our approach achieved second place in 2019 DCASE Task 1b (0.4% behind the winning entry), and the best Task 1B evaluation results (by a large margin of over 5%) on test data from a device not used to record any training data
Naranjo-Alcazar,S.Perez-Castanos,P.Zuccarello, and M. Cobos	2020	Task 1 DCASE 2020: ASC with mismatch devices and reduced size model using residual squeeze excitation CNNs	residual squeeze excitation CNNs	uses squeeze-excitation techniques in its residual blocks to compel the network to extract significant characteristics from the audio representation
M. McDonnell	2020	Low-complexity acoustic scene classification using one-bit-per-weight deep convolutional neural networks	One-bit-per-weight deep CNNs	This method made it possible to train a single 36-layer all-convolutional deep neural network with 3,987,000 binary weights total of 486.69KB.
S. Suh, S. Park, Y. Jeong, and T. Lee	2020	Designing acoustic scene classification models with CNN variants	CNN varaints, Inception model	The results have increased the Model’ accuracy upto 97.6% While reducing th number of Parameters
Irene Martin-Morato , Francesco Paissan, Alberto Ancilotto, Toni Heittola, Annamaria Mesaros, Elisabetta Farella, Alessio Brutti , Tuomas Virtanen	2022	Low-complexity acoustic scene classification in dcase 2022 challenge	Convolutional Neural Networks (CNNs)	The task received 48 submissions from a number of 19 teams. The number o participants in this edition is lower than i previous years, but similar to participatio statistics of the other tasks. Only three o the 19 teams have lower performance tha the baseline. The best system has a lo loss of 1.091 and accuracy of 59.6%, wit the four best spots belonging to team Schmid CPJKU [25]

recorded in six major European cities, has more acoustic variability than the earlier datasets used for this task and also includes data taken with mobile devices in addition to high-quality binaural recordings. We also provide the standard system, which consists of of a convolutional neural network and its performance in the subtasks using the recommended cross-validation setup.

D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M. D. Plumbley et al., [24] This workshop presentation details our contribution to the "detection and classification" job for the acoustic scene classification (ASC).2016 competition "of acoustic sceneries and events" (DCASE). We suggest using a convolutional neural network that has been trained to categorise brief audio sequences, which are represented by their log-mel spectrogram. We also suggest a training technique that might be applied when system validation performance reaches saturation while training continues. The public ASC development dataset offered for the DCASE 2016 competition is used to evaluate the system. On a four-fold cross-validation scenario, our system's greatest accuracy score was 79.0%, representing an 8.8% relative improvement over the baseline system.

IV. Conclusion

From the review of various papers based on Acoustic Scene Classification deep learning methods have its own significance for scene classification. But usage of single device for audio recordings is computationally complex. However, the usage of various devices for context awareness is a desired direction for application, solutions that are appropriate for low processing power are required. Furthermore, the task might think about directing development into remedies where there is still opportunity for improvement, like reducing the working memory usage

References

[1] E. Benetos, D. Stowell, and M. D. Plumbley, Approaches to Complex Sound Scene Analysis. Cham: Springer International Publishing, 2018, pp. 215–242

[2] T. Heittola, A. Mesaros, and T. Virtanen, “Acoustic scene classification in DCASE 2020 challenge: generalization across devices and low complexity solutions,” in Proceedings of the Detection and Classification of Acoustic Scenes and Events 2020 Workshop, Tokyo, Japan, November 2020, pp. 56–60.

[3] Mesaros, A.; Heittola, T.; Virtanen, T. Assessment of Human and Machine Performance in Acoustic Scene Classification: DCASE 2016 Case Study. In Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), New Paltz, NY, USA, 15–18 October 2017; pp. 319–323

[4] Li, Y.; Li, X.; Zhang, Y.; Wang, W.; Liu, M.; Feng, X. Acoustic Scene Classification Using Deep Audio Feature and BLSTM Network. In Proceedings of the 6th International Conference on Audio, Language and Image Processing (ICALIP), Shanghai, China, 16–17 July 2018; pp. 371–374

[5] A.Mesaros,T. Heittola, and T. Virtanen,“Acoustic scene classification in DCASE 2019 challenge: closed and open set classification and data mismatch setups,”in Proc.of the DCASE 2019 Workshop, NewYork,Nov2019

[6] T. Heittola, A. Mesaros, and T. Virtanen, “Acoustic scene classification in DCASE 2020 challenge: generalization across devices and low complexity solutions,” in Proceedings of the Detection and Classification of Acoustic Scenes and Events 2020 Workshop, Tokyo, Japan, November 2020, pp. 56–60.

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

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

[9] S. Kumari, D. Roy, M. Cartwright, J. P. Bello, and A. Arora, “EdgeL3: Compressing L3-net for mote scale urban noise monitoring,” in 2019 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW), May 2019, pp. 877–884.

[10] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M. D. Plumbley, “Detection and classification of acoustic scenes and events,” IEEE Trans. on Multimedia, vol. 17, no. 10, pp. 1733–1746, October 2015.

[11] A multi-device dataset for urban acoustic scene classification,” in Proc. of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE2018), November 2018.

[12] A. Mesaros, T. Heittola, and T. Virtanen, “Acoustic scene classification: an overview of dcase 2017 challenge entries,” in 16th International Workshop on Acoustic Signal Enhancement (IWAENC), 2018.

[13] A. Mesaros, T. Heittola, E. Benetos, P. Foster, M. Lagrange, 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, Feb 2018

[14] Gao W, and McDonnell, (2019), “Acoustic scene classification using deep residual networks with late fusion of separated high and low frequency paths,” DCASE2019 Challenge, Tech. Rep

[15] Naranjo-Alcazar,S.Perez-Castanos,P.Zuccarello, and M. Cobos, “Task 1 DCASE 2020: ASC with mismatch devices and reduced size model using residual squeeze excitation CNNs,” DCASE2020 Challenge, Tech. Rep., June 2020

[16] M. McDonnell “Low-complexity acoustic scene classification using one-bit-per-weight deep convolutional neural networks” DCASE2020 Challenge, Tech. Rep., June 2020.

[17] S. Suh, S. Park, Y. Jeong, and T. Lee,“Designing acoustic scene classification models with CNN variants,” DCASE2020 Challenge, Tech. Rep.,June 2020.

[18] S. Abidin et al.Spectrotemporal analysis using local binary pattern variants for acoustic scene classification IEEE/ACM Trans. Audio Speech Lang. Process(2018)

[19] Daniele Barchiesi Acoustic scene classification: classifying environments from the sounds they produce J. IEEE Signal Process. Mag. Year(2015)

[20] Late fusion framework for Acoustic Scene Classification using LPCC, SCMC, and log-Mel band energies with Deep Neural Networks C Paseddula, SV Gangashetty - Applied Acoustics, 2021 – Elsevie.

[21] Waldekar Shefali et al.Classification of audio scenes with novel features in a fused system framework J. Digital Signal Process.(2018)

[22] A. Mesaros, T. Heittola, T. Virtanen, A multi-device dataset for urban acoustic scene classification, in: Proceedings...

[23] DCASE 2016 Acoustic Scene Classification Using Convolutional Neural Networks.M Valenti, A Diment, G Parascandolo, SSquartini… -DCASE,2016 dcase.community

[24] F. Schmid, S. Masoudian, K. Koutini, and G. Widmer, “CPJKU submission to dcase22: Distilling knowledge for low complexity convolutional neural networks from a patchout audio transformer,” DCASE2022 Challenge, Tech. Rep., June 2022.