SUBMISSION TO DCASE 2022 TASK 1: DEPTHWISE SEPARABLE CONVOLUTIONS FOR LOW-COMPLEXITY ACOUSTIC SCENE CLASSIFICATION

Technical Report

Chukwuebuka Olisaemeka and Lakshmi Babu Saheer

Anglia Ruskin University
Cambridge, United Kingdom
olisaemekaebuka@gmail.com, lakshmi.babu-saheer@aru.ac.uk

ABSTRACT

This technical report describes the details of the TASK1 submission to the DCASE2022 challenge. The aim of this task is to design an acoustic scene classification system that targets devices with low memory and computational allowance. The task also aims to build systems that can generalize across multiple devices. To achieve this objective, a model using Depthwise Separable Convolutional layers is proposed, which reduces the number of parameters and computations required compared to the normal convolutional layers. This work further proposes the use of dilated kernels, which increase the receptive field of the convolutional layer without increasing the number of parameters to be learned. Finally, quantization is applied to reduce the model complexity. The proposed system achieves an average test accuracy of 39% and log loss of 1.878 on TAU Urban Acoustic Scenes 2022 Mobile, development dataset with a parameter count of 96.473k and 3.284 MMACs.

Index Terms— Acoustic scene classification, dilated convolutions, depthwise separable convolution, low-complexity, DCASE Challenge

1. INTRODUCTION

Acoustic scene classification (ASC) is the task of recognizing the surrounding environment in which an audio recording was done [1]. The audio recording is usually classified into predefined classes such as “tram”, “metro” and other acoustic scenes. The ability of machines to recognize the environment in which they are embedded is beneficial in robotic navigation [2], intelligent wearables [3], and context-aware applications [4]. Detection and Classification of Acoustic Scenes and Events (DCASE) [5] is a platform that organizes annual challenges in the field of sound scene and event research, which has contributed to advancements in this field. This report documents the submission for the DCASE 2022 Task 1 [1] which focuses on developing a low-complexity acoustic scene classification system aimed at mobile devices which are characterized by low computational and memory allowance.

In the previous year, the DCASE 2021 Task 1A [6] enforced the low-complexity requirement by a constraint on the number of parameters allowed in the model [7, 8]. The DCASE 2022 Task 1 includes an additional constraint on the multiply-accumulate operations count. This task also aims at developing models which generalize across different recording devices. In previous years, state-of-the-art performance in ASC has been achieved using convolutional neural networks [9] along with residual networks [10, 7]. In this work, Depthwise Separable Convolutions are proposed for its ability to reduce the number of parameters and computations used as compared to regular convolutions [11, 12, 13]. Depthwise Separable Convolutions consist of a depthwise convolution followed by a pointwise convolution. The proposed model also makes use of dilated convolutions as a way to increase the receptive field of the convolution layers and therefore integrate more information [14, 15].

The rest of the report is structured as follows. Section 2 describes the acoustic features, the proposed system, and the compression approach. Section 3 shows the experimental result and Section 4 concludes the work

2. PROPOSED METHOD

2.1. Acoustic Features

The acoustic features used in this research for training the network are log-mel (logarithmic-magnitude Mel-scale filter bank) features, which represent the frequency content of the audio recording as they vary with time [16]. This time-frequency representation is the feature of choice for acoustic scene classification, as seen in the top-performing systems of DCASE 2021 Task 1a [7, 8]. To extract these features, a short-time Fourier transform (STFT) is applied on the audio recordings sampled at 44100Hz, over windows of 40ms using a 50% overlap, after which a 40 band Mel-scale filter bank is applied. Finally, the logarithmic conversion is applied to these Mel frequencies.

2.2. Network Architecture

The two-dimensional audio-spectral images are modelled using a Convolutional Neural Network (CNN) architecture. CNN is chosen for this task due to the ability of convolutional layers to serve as feature extractors by learning discriminative features through convolutions and non-linear transformations [17] of the audio spectrum information [18]. Several CNN architectures were experimented before selecting the best performing model presented in this report. Empirical tuning of hyperparameters was also carried out to come up with the final model.

The proposed model architecture is illustrated in Table 1. The log-mel features are fed into two consecutive Depthwise Separable Convolutional (DSC) layers with 16 kernels of 7x7 feature maps. The border type used for these convolution layers is “same” which indicates that the input features are padded with zeros to prevent the output of the DSC layers from having a reduced size [19]. Each DSC layer is followed by a batch normalization layer, which helps

Table 1: Proposed architecture for DCASE 2022 Task 1. DSConv2D and DSConv1D represent a depthwise separable 2D and 1D convolution layers respectively. All convolution layers use similar convolutions with a stride of 1x1.

Layer	Number of units	Kernel size	Activation function	Dropout rate	Output	Dilation rate	Batch Norm
DSConv2D	16	7x7	relu	-	40, 51, 16	-	Yes
DSConv2D	16	7x7	relu	0.3	40, 51, 16	-	Yes
Max Pool	-	5x1	-	-	8, 51, 16	-	-
DSConv2D	32	7x7	relu	0.3	8, 51, 32	-	Yes
Max Pool	-	2x1	-	-	4, 51, 32	-	-
Permute	-	-	-	-	51, 4, 32	-	-
Reshape	-	-	-	-	51, 128	-	-
DSConv1D	32	5x5	relu	0.3	51, 32	1	Yes
DSConv1D	32	5x5	relu	0.3	51, 32	2	Yes
DSConv1D	32	5x5	relu	0.3	51, 32	4	Yes
DSConv1D	32	5x5	relu	0.3	51, 32	8	Yes
Flatten	-	-	-	-	1632	-	-
FC	50	-	relu	-	50	-	-
FC	10	-	softmax	-	10	-	-

alleviate the problem of internal covariate shifts and thus enables generalization, while also speeding up training [20]. The rectified linear unit (ReLU) activation is used as the activation function. The output of this group is down-sampled by a max-pooling layer, only on the frequency dimension. This is followed by dropout operation with a dropout probability of 0.3, as a regularization procedure to prevent overfitting [21].

The next layer to follow in the architecture is another DSC layer, but with 32 kernels of 7x7 feature maps. This layer also preserves the size of the convolution. Again, using ReLU activation followed by batch normalization, frequency pooling and dropout probability of 0.3.

The output is then permuted, reshaped, and fed into a set of four consecutive Dilated DSC layers. Each layer has 32 kernels with 5x5 feature maps and size-preserving convolution alongside using the ReLU activation. These layers also make use of batch normalization and dropout with a dropout probability of 0.3. Dilation in convolution defines the number of spaces placed between the values in the kernel window, which are usually adjacent in regular convolutions. These holes have the advantage of increasing the receptive field of the kernel without adding more parameters [15]. The receptive field in dilated convolutions is expanded exponentially by stacking layers of these dilated convolutions with increasingly dilated values, as shown in Figure 1 [14]. Thus, the 4 dilated DSC layers with dilation rates of 1, 2, 4, and 8 respectively are being implemented. Dilated convolutions along the temporal dimension can capture long temporal dependencies. Hence, these have been used in the proposed architecture as an equivalent for the recurrent neural network models.

The Dilation operations in the architecture are followed by a fully connected (FC) dense layer with 50 neurons and the ReLU activation function using the dropout probability of 0.3. Finally, the output of the fully connected layer is fed into the output layer, which is another fully connected layer with 10 neurons matching the number of classes being predicted. The softmax activation function is used in this output layer to denote the degree of confidence of each predicted class [22].

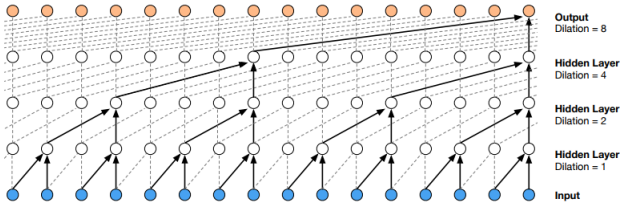

Figure 1: Stacked Dilated Convolutions with increasing dilation rates [14]

2.3. Compression

Post-training quantization is a method that converts infinite values to discrete finite values. Full integer quantization converts all floating-point tensors in the model, which include constant tensors such as weights and biases, and variable tensors such as model input, activations, and model outputs. Quantization is performed to reduce the model size while improving CPU and hardware accelerator latency with a slight reduction in model accuracy. A mini-batch of 100 acoustic features were used to perform a full integer post-training quantization on the trained model, which converted it from 32-bit format to 8-bit format [8].

3. EXPERIMENTS

3.1. Experimental Setup

The proposed model is trained and evaluated on the TAU Urban Acoustic Scenes 2022 Mobile, development dataset [23]. The dataset contains recordings from 12 European cities in 10 different acoustic scenes using 4 different devices. Artificial data for 11 simulated mobile devices were created from the original recordings. The audio recordings were performed in Amsterdam, Barcelona, Helsinki, Lisbon, London, Lyon, Madrid, Milan, Prague, Paris, Stockholm, and Vienna. The real devices are designated identifiers A, B, C, and D. The 11 additional mobile devices S1-S11 are simulated using the audio recorded with device A and modifications such as impulse responses with real devices and dynamic range

Table 2: Comparison of baseline method with proposed method on TAU Urban Acoustic Scenes 2022 Mobile, development dataset

Scene	Baseline		Proposed Architecture	
	Log loss	Accuracy	Log loss	Accuracy
airport	1.748	29.1%	1.770	**34.5%**
bus	1.723	31.6%	2.132	**34.4%**
metro	1.538	40.0%	1.550	**45.6%**
metro station	1.724	37.5%	2.028	30.2%
park	1.291	61.4%	1.775	52.7%
public square	2.037	27.1%	2.607	19.6%
shopping mall	1.781	39.8%	**1.678**	**44.5%**
street pedestrian	1.656	33.6%	2.141	26.3%
street traffic	1.050	68.5%	1.141	65.8%
tram	1.389	49.7%	1.961	36.3%
Average	1.594	41.8%	1.878	39.0%

compression, to simulate realistic recordings. The audio recordings are 1 second long in a single channel and sampling rate of 44.1kHz. The acoustic scenes are “airport”, “shopping mall”, “metro station”, “street pedestrian”, “public square”, “street traffic”, “tram”, “bus”, “metro”, and “park”.

The development set contains data from 10 cities out of the 12 and 9 devices out of 11: 3 real devices (A-C) and 6 simulated devices (S1-S6). The dataset contains 230,350 audio segments split in a training split of 70% and a test split of 30%, however, the training split does not contain audio recordings from the S4, S5, and S6 devices to test the generalization ability of the proposed models. The evaluation set contains data from all 12 cities and 11 devices. Devices S7-S11 are unseen during development and only included in the evaluation set.

The development set with a batch size of 64 is used to train the proposed model for 200 epochs using the Adam optimizer with learning rate of 0.001 while observing the categorical cross-entropy loss and categorical accuracy. After the first 50 epochs, the weights of the epoch with the best value of the categorical accuracy are saved and the final weight values are retrieved at the end of the training process and used for testing.

3.2. Results

Table 2 shows the result of training and testing the proposed system using the training/test split of the development set. The proposed system reported a log loss of 1.878 and an accuracy of 39% which are both very close to the performance of the baseline system for the DCASE 2022 Task 1. The performance also improved for some scenes as highlighted in bold in the table. Especially, there seems to be significant improvements in the detection of ”shopping mall” scene for both log-loss and accuracy. Further, the accuracy metrics has shown more improvements in the detection of ”airport”, ”bus”, ”metro” and ”shopping mall” scenes. More qualitative analysis would be needed to understand these performance differences. The performance could have been further improved with a more rigorous hyperparameter tuning.

The breakdown of the results per device data is shown in Table 3 which can be directly compared to baseline model performance in Table 4 with respect to the log-loss metrics. The values that show better performance for the proposed system are highlighted in Table 3. Similar to previous observations, some of the scenes like ”airport”, ”metro”, ”shopping mall” and ”street traffic” are performing better with the proposed models. While ”airport”, ”shopping mall” and ”metro” had more consistency across most devices including the real and simulated devices. Others are more specific viz., ”street traffic” demonstrates better performance for real devices compared to simulated ones.

The parameter count for the proposed system is 96,473 which meets the requirements of the task. The task specifies a requirement for the system to be of 128,000 parameters or fewer. The multiple accumulate count (MAC) for the proposed system is 3,283,692 which also meets the requirements of the task which requires a model of no more than 30,000,000 MACs and much lower than the baseline MAC of 29,238,120.

4. CONCLUSION

A model based on depthwise separable convolutions is proposed in this report in order to achieve a low-complexity acoustic scene classification model. The use of stacked dilated convolutions could further reduce the parameter count and computations while integrating more contextual information. The proposed method achieves a parameter count of 96.473k with 3.284 MMACs with a log loss of 1.878 and an accuracy of 39%. Even though the overall average scores are sometimes below the performance of the baseline model, some scenes were identified to be performing better using the proposed system.

It can be safely assumed that further hyperparameter tuning can help improve the overall performance. The data augmentation techniques could also be implemented in future to improve this proposed system. This could help to achieve better generalization and lessen the effect of the imbalance of the number of audio recordings from the different device types. The use of residual learning is also suggested as means to achieve deeper networks still remaining in the confines of the task requirements with the use of depthwise separable convolutions.

5. REFERENCES

[1] I. Martín-Morató, F. Paissan, A. Ancilotto, T. Heittola, A. Mesaros, E. Farella, A. Brutti, and T. Virtanen, “Low-complexity acoustic scene classification in dcase 2022 challenge,” *arXiv preprint arXiv:2206.03835*, 2022.

[2] S. Chu, S. Narayanan, C.-C. J. Kuo, and M. J. Mataric, “Where am i? scene recognition for mobile robots using audio features,” in *2006 IEEE International conference on multimedia and expo*. IEEE, 2006, pp. 885–888.

[3] Y. Xu, W. J. Li, and K. K. Lee, *Intelligent wearable interfaces*. John Wiley & Sons, 2008.

[4] R. G. Malkin and A. Waibel, “Classifying user environment for mobile applications using linear autoencoding of ambient audio,” in *Proceedings.(ICASSP’05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, vol. 5. IEEE, 2005, pp. v–509.

[5] http://dcase.community/challenge2022/.

Table 3: Proposed architecture scene/device log loss results on TAU Urban Acoustic Scenes 2022 Mobile, development dataset

Scene	A	B	C	S1	S2	S3	S4	S5	S6
airport	1.250	1.727	**1.339**	**1.919**	2.023	1.453	**2.009**	**1.959**	**2.246**
bus	1.474	1.737	1.514	2.202	2.082	2.388	2.856	2.486	2.448
metro	**0.958**	1.594	**1.395**	**2.079**	1.820	1.479	**1.429**	**1.694**	1.506
metro station	2.076	2.126	2.318	**2.029**	2.363	1.931	2.003	1.964	**1.445**
park	0.721	0.454	0.867	1.682	1.638	1.952	3.234	2.010	3.413
public square	1.709	1.941	1.943	2.531	2.622	2.221	3.362	3.567	3.564
shopping mall	1.770	**1.392**	**1.587**	**1.596**	1.720	**1.795**	1.836	**1.391**	2.013
street pedestrian	1.360	1.889	1.047	2.025	1.907	2.397	2.573	2.620	2.449
street traffic	**0.657**	**1.065**	1.149	0.896	1.526	1.209	1.153	1.025	1.587
tram	1.323	1.980	1.571	1.308	1.871	1.337	2.170	3.433	2.642
Average	1.330	1.590	1.573	1.827	1.957	1.816	2.262	2.215	2.331

Table 4: Baseline architecture scene/device log loss results on TAU Urban Acoustic Scenes 2022 Mobile, development dataset

Scene	A	B	C	S1	S2	S3	S4	S5	S6
airport	1.197	1.506	1.543	1.993	1.651	1.345	2.140	2.053	2.294
bus	0.905	1.694	1.159	1.766	1.525	1.774	2.251	2.133	2.296
metro	1.073	1.392	1.489	2.239	1.620	1.399	1.620	1.749	1.264
metro station	1.501	1.764	1.720	2.057	1.970	1.619	1.938	1.455	1.492
park	0.390	0.363	0.602	1.261	0.985	1.390	2.213	1.981	2.434
public square	1.429	1.504	1.848	2.004	1.891	1.723	1.910	2.807	3.215
shopping mall	1.765	1.536	1.850	1.798	1.580	2.172	1.777	1.827	1.724
street pedestrian	1.200	1.680	1.628	1.493	1.625	1.702	1.969	1.719	1.889
street traffic	0.764	1.226	1.062	0.803	1.139	1.156	1.083	0.732	1.482
tram	1.032	1.406	1.167	1.174	1.433	1.009	1.557	2.127	1.592
Average	1.126	1.407	1.408	1.659	1.542	1.529	1.846	1.858	1.968

[6] http://dcase.community/challenge2021/.

[7] B. Kim, S. Yang, J. Kim, and S. Chang, "Qti submission to dcase 2021: Residual normalization for device imbalanced acoustic scene classification with efficient design," *DCASE2021 Challenge, Tech. Rep*, 2021.

[8] C.-H. H. Yang, H. Hu, S. M. Siniscalchi, Q. Wang, Y. Wang, X. Xia, Y. Zhao, Y. Wu, Y. Wang, J. Du, *et al.*, "A lottery ticket hypothesis framework for low-complexity device-robust neural acoustic scene classification," *arXiv preprint arXiv:2107.01461*, 2021.

[9] J. Abeßer, "A review of deep learning based methods for acoustic scene classification," *Applied Sciences*, vol. 10, no. 6, p. 2020, 2020.

[10] K. Koutini, S. Jan, and G. Widmer, "Cpjku submission to dcase21: Cross-device audio scene classification with wide sparse frequency-damped cnns," DCASE2021 Challenge, Tech. Rep, Tech. Rep., 2021.

[11] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1251–1258.

[12] L. Kaiser, A. N. Gomez, and F. Chollet, "Depthwise separable convolutions for neural machine translation," *arXiv preprint arXiv:1706.03059*, 2017.

[13] A. G. Santos, C. O. de Souza, C. Zanchettin, D. Macedo, A. L. Oliveira, and T. Ludermir, "Reducing squeezenet storage size with depthwise separable convolutions," in *2018 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2018, pp. 1–6.

[14] B. Rekabdar and C. Mousas, "Dilated convolutional neural network for predicting driver's activity," in *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2018, pp. 3245–3250.

[15] X. Zhang, Y. Zou, and W. Shi, "Dilated convolution neural network with leakyrelu for environmental sound classification," in *2017 22nd international conference on digital signal processing (DSP)*. IEEE, 2017, pp. 1–5.

[16] Y. Wu and T. Lee, "Enhancing sound texture in cnn-based acoustic scene classification," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 815–819.

[17] A. Sharif Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "Cnn features off-the-shelf: an astounding baseline for recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2014, pp. 806–813.

[18] Z. Lu, "Sound event detection and localization based on cnn and lstm," *Detection Classification Acoust. Scenes Events Challenge, Tech. Rep*, 2019.

[19] R. Chauhan, K. K. Ghanshala, and R. Joshi, "Convolutional neural network (cnn) for image detection and recognition," in *2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC)*. IEEE, 2018, pp. 278–282.

[20] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*. PMLR, 2015, pp. 448–456.

[21] C. Garbin, X. Zhu, and O. Marques, "Dropout vs. batch normalization: an empirical study of their impact to deep learning," *Multimedia Tools and Applications*, vol. 79, no. 19, pp. 12 777–12 815, 2020.

[22] I. Kouretas and V. Paliouras, "Simplified hardware implementation of the softmax activation function," in *2019 8th international conference on modern circuits and systems technologies (MOCAST)*. IEEE, 2019, pp. 1–4.

[23] T. Heittola, A. Mesaros, and T. Virtanen, "Acoustic scene classification in dcase 2020 challenge: generalization across devices and low complexity solutions," *arXiv preprint arXiv:2005.14623*, 2020.