A Comparative Study on Approaches to Acoustic Scene Classification using CNNs

Ishrat Jahan Ananya, Sarah Suad, Shadab Hafiz Choudhury, Mohammad Ashrafuzzaman Khan

[ishrat.jahan16@northsouth.edu](mailto:ishrat.jahan16@northsouth.edu), [sarah.suad@northsouth.edu](mailto:sarah.suad@northsouth.edu), [shadab.choudhury@northsouth.edu](mailto:shadab.choudhury@northsouth.edu), mohammad.khan02@northsouth.edu

**Abstract.** Acoustic Scene Classification is the process of recognizing and classifying the source of an audio signal using data from recorders or microphones. As it can be used in many applications, this paper explores the use of several different approaches based on vector embeddings, MFCCs and spectrograms to classify audio signals from a dataset of sounds from common urban scenes. The audio files were prepared using each of the approaches and tested through several common CNN and custom deep neural architectures. Relatively accurate results were achieved using spectrograms and embeddings with several popular CNN models.

**Keywords:** Acoustic Scene Classification, Signal Processing, Deep Learning Convolutional Neural Network, Autoencoders

1. Introduction

Two of the primary senses human beings use to navigate their environment are sight and sound. Audio classification has uses similar to those of computer vision and is one of the most important areas of artificial intelligence.

The human brain is capable of recognizing the source and features of multiple overlapping sounds from various distances simultaneously. The volume, pitch and sequence of these sounds are all acoustic features. From the combination of these features, an individual can easily recognize what environment they are in. The acoustic features of a busy street may include a high level of background noise from footsteps, conversations and vehicles, punctuated by short bursts of loud noises such as from car horns or shouts. By contrast, a shopping mall or airport will have a far lower overall volume of background noise but may have more distinct conversations, music, announcements and other activities.

Acoustic Scene Classification (ASC) is the process of teaching an artificial intelligence system to recognize acoustic features and classify the scene of the sound based on them. It can be used to detect and understand the context of a scene from multimedia recordings. It can be used in various autonomous systems such as monitoring systems, robots and self-driving vehicles.

It is a relatively newer field of study compared to some others, as large-scale datasets on the topic were rare until the annual Detection and Classification of Acoustic Scenes and Events (DCASE) challenges began in 2013.

Convolutional Neural Networks (CNNs) are one of the best contemporary solutions for feature extraction and classification. There are a large number of different models that can be used. Though there are many exceptions, a general idea is that the accuracy of a CNN can be increased by increasing the number of parameters and the depth of the model.

While the problem of binary classification is a limited one, it is also fairly important. It can be used, for instance, to enable robots or other automated devices to switch modes of behavior depending on whether they are indoors or outdoors in order to minimize accidents. It can also be used to help those who are both hard of hearing and visually impaired recognize their surroundings. It can also be used to enhance the inference abilities of models trained on other large scale audio datasets, since many of these datasets combine indoor and outdoor sounds into one.

In this paper, we examine a straightforward approach for the classification of audio in urban scenes into indoor and outdoor scenes. Care was taken to ensure that the preprocessing approach remains generic and suitable for use on a larger scale. Since datasets for acoustic scene classification are fairly limited at the moment, approaches that are highly tailored for a particular dataset could be difficult to implement on a much wider scale. Preprocessing is carried out in three routes for comparison – generating vector embeddings, creating spectrograms and creating MFCCs. The separate output datasets were then trained using the similar models and their accuracies compared.

1. Related Work

Analysis and classification of auditory signals with artificial intelligence have a long history. Initially, research work was focused on simply detecting and distinguishing acoustic events [1] such as distinct noises like claps and speech, or different individuals speaking [2]. These early examples of the use of neural networks in classification of audio developed from the intersection of signal processing and artificial intelligence.

More mature artificial intelligence techniques such as sophisticated convolutional neural networks have enabled further exploration of Acoustic Scene Classification through different approaches. The DCASE Challenges, initially started in 2013, offer datasets and a platform for exploration of Acoustic Scene Classification [3]. DCASE 2013 highlighted the use of large datasets for acoustic scene classification in various scenes such as a bus, office, market, et cetera.

Early research yielded good outcomes with machine learning models. Good results were achieved in the DCASE 2013 challenge using algorithms such as support vector machines and decision trees [4]. However, as the sophistication and size of datasets increased, neural networks became an effective choice. In this specific dataset, Valenti et al’s approach using a custom CNN model resulted in higher accuracy compared to earlier work - up to 9.7% depending on the technique it is compared to [5].

The majority of recent work on acoustic scene classification has followed up on the CNN approach. Hussein et al. developed a more indepth technique using a deep neural network with only 3 hidden layers that achieved up to 90% accuracy on the DCASE 2016 challenge [6].

In the DCASE 2020 challenge, several attempts were able to reach 96% test accuracy by implementing modern deep convolutional neural networks such as ResNets. While both these and the previous approach were highly accurate, they also made use of specific preprocessing techniques and model designs that could be difficult to implement on a larger scale.

Finally, an excellent overview of the development and use of deep learning in acoustic scene classification between 2013 and 2020 is given in a review paper by Abeßer [7].

1. Methodology

While it is possible to carry out deep learning on audio signals directly, it is both more effective and efficient to use different representations that draw out features. Preprocessing was carried out on the initial data, which was then converted to the different representations.

The CNN models utilized in this experiment were a simple autoencoder, ResNet-18 and ResNet-50. Alterations to the structure of the models were kept to a minimum in order to ensure a fair comparison. Any changes made were in order to ensure that the input could be appropriately fed into the model.

* 1. Data Organization and Collection

The dataset was collected from DCASE 2020 challenge [8]. It had three classes of indoor, outdoor and transport. These classes were subdivided into nine more subclasses. The raw dataset contained 10 second audio clips in 24-bit .wav format taken from ten different cities in the world.

For this paper, we chose to reduce it to two classes: indoor and outdoor. This left us with six subclasses, three from each. For Indoor Scenes, the subclasses were Metro, Shopping Mall and Airport. For Outdoor Scenes, the subclasses were Park, Pedestrian Street and Public Square.

There were 8640 data samples in total, adding up to 24 hours of audio. Each subclass had 1440 data points. We ensured that the number of audio samples for each class was equal so that the model was not biased towards any particular category due to unbalanced data. The original audio files were binaural at 44.1 kHz. This is a relatively small dataset, so we split each of the 10 second audio files into two 5 second files to double the size of the dataset.

* 1. Data Preprocessing

The task of data preprocessing requires taking audio samples as input and extracting features from the audio signals. By doing so, we aim to find components of the audio signals that will help us differentiate it from other categories of signals. For the purpose of comparison, we implemented three methods of processing data: producing log-mel spectrograms [9], Mel Frequency Cepstral Coefficients (MFCC) [10] and audio embeddings.

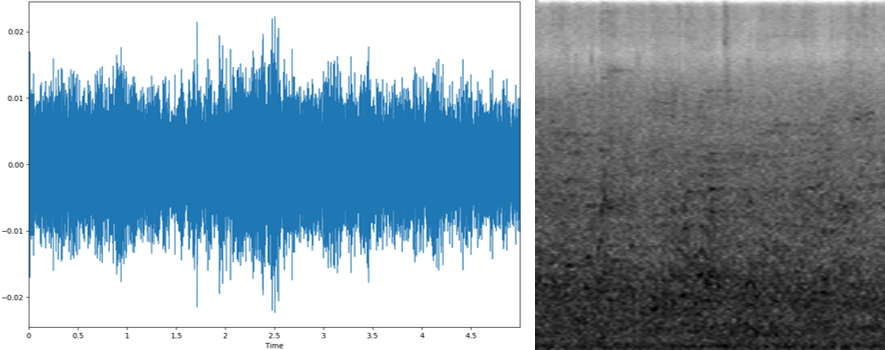
Before converting the representation, further data augmentation was carried out. A small amount of random noise was added and audio tracks were randomly shifted forwards and backwards. This gave us a rich dataset of 34,560 five-second audio clips, half of which was augmented.

### MFCCs and Spectrograms**.**

The process of MFCC calculation requires that the audio signal be sliced into short time frames that allow us to assume that the audio signal has had little to no change. The signal was sliced at 20-40ms. The power spectrum for each frame was then calculated. This allowed us to see the distribution of power into frequency components that make up the signal. These steps were accomplished by applying a short-time fourier transform on each frame.

We then applied the mel filter bank to each power spectra and summed up the energy in each filter. This is a necessary step to estimate how the human ear perceives sounds at different frequencies and different volumes. The last step in the calculation is to take the discrete cosine transform of the logarithm of each filterbank. These were carried out using the python library Librosa. [11].

Log-mel spectrograms were produced in a similar way. The audio dataset underwent a short- time Fourier transform to get spectrograms that were rich in features needed for this task. The spectrograms were then scaled to the mel scale and saved in png format. The following figure shows an example of a Spectrogram. No visualization is given for the MFCCs and Embeddings, as they are simply two-dimensional matrices.



**Fig. 1.** Two representations of an audio file recorded at an Airport in Lisbon: a waveform plot (left) and a Spectrogram (right).

### Audio Embeddings

In dense data such as audio, the embeddings determine similarity metrics to other sounds. Essentially, it splits the audio clip into smaller intervals. For each interval, it gives a similarity value to all the classes the original embedding model was trained on.

Our embeddings were generated using IBM’s Audio Embedding Generator [12, 13]. The generator accepts a 16-bit PCM wav file as input, generates the embeddings and outputs the result as arrays of 1 second embeddings. The model was trained on Audioset, which includes 632 classes. For each second, the embeddings list the 128 classes that have the highest similarity to the sound.

* 1. Development of CNNs

For the Spectrograms dataset, we made use of various data augmentation techniques to increase the amount of new data to be trained on in every batch. The Imgaug library [14] was used to add more complex transformations that altered the existing images in the dataset in several ways. The data augmentation transformations used here are: affine transformation, Gaussian blur and saturation changes.

The ResNet family of neural network architectures is ideal for image classification tasks. The ResNet architecture uses stacked layers of residual learning blocks using shortcuts between layers to minimize the effect of the vanishing gradient problem [15]. We used the SGD optimizer with Cross Entropy Loss and a learning rate of 0.001. The results of both models are discussed in the Evaluation section.

1. Results and Evaluation

The following table gives a summary of the results of our experimentation, and a discussion of the results follows.

**Table 1.** Experimental Results

|  |  |  |  |
| --- | --- | --- | --- |
| Preprocessing Approach | CNN Model | Test Acc | Training Acc |
| Spectrograms | Autoencoder | 76.3% | 79.2% |
| Spectrograms | ResNet-18 | 89.7% | 91.9% |
| Spectrograms | ResNet-150 | 90.4% | 93.6% |
| Spectrograms | ShuffleNet | 93.1% | 95.2% |
| MFCCs | Autoencoder | 49.9% | 50.0% (invalid) |
| MFCCs | ResNet-18 | 71.3% | 86.6% |
| MFCCs | ResNet-50 | 72.1% | 88.0% |
| Embeddings | Autoencoder | 78.7% | 79.2% |
| Embeddings | ResNet-18 | 77.6% | 99.9% (overfit) |
| Embeddings | ResNet-50 | 77.1% | 99.7% (overfit) |

With three datasets that are so disparate in nature, it is not easy to draw conclusions based on a limited number of CNN models. Some hyperparameter optimization was carried out on each of these models, though not extensively as there were a lot of models and approaches to test.

As seen in Table, the Spectrogram approach offers the best results, up to 93% accuracy. Even a simple fully-connected autoencoder of 4096-2048-1024-512 parameters gives us 76.3% accuracy. This autoencoder was also used in the ResNet models to improve feature extraction. After getting an extremely high value from the ResNets, we decided to try another state-of-the-art model, Shufflenet. This is a computationally efficient CNN architecture, particularly designed for mobile devices with limited computing power [16]. While these results do not break any of the benchmarks set in previous DCASE challenges, they are all fairly generic approaches that require minimal customization to the dataset. This ensures that the results are applicable across different acoustic scene datasets rather than being optimized for this particular problem.

Due to the features of the MFCCs, using an Autoencoder was completely ineffective. We note that ResNet models are actually too heavy for the MFCC approaches, generally overfitting within 10-15 epochs - before the test accuracy has levelled out. Using the non-augmented half of the data set saw a slight improvement in accuracy, at the cost of even more overfitting. At this stage, regularization techniques were ineffective. Therefore, for a dataset of this size, the augmentation was necessary.

MFCCs are generally used for speech classification. The audio in this dataset is primarily background noise at a similar energy level throughout. Therefore, we expect that it is harder to extract features using MFCCs compared to other approaches.

The audio embeddings performed surprisingly well considering the nature of the dataset. As the embeddings are dependent on similarity, the best method to use is a clustering approach. Hence, an autoencoder with layers of 640-320-160-80 was used. The original dataset used to develop the embedding generator was the AudioSet Dataset, which is focused on speech, music and the sounds made by individual objects. Despite being a somewhat unsuitable dataset, it gave good results. With a more closely related embedding generator, it could give results comparable to Spectrograms at a fraction of the computation power. As seen in the table, heavier models like the ResNets led to overfitting.

Additionally, it is clear from both our work and other research that unlike image classification, audio classification does not always benefit from deep networks. Both ResNet-18 and ResNet-50 have very similar accuracies. Rather, higher resolution [17] or additional preprocessing is necessary to achieve higher accuracy.

1. Conclusion

As the evaluation section shows, the task of Acoustic Scene Classification faces major hurdles when it comes to larger models. Extensive data preprocessing and augmentation is necessary to achieve high accuracies on even very limited problems. Out of the three different approaches tested, Spectrograms offered the best result for acoustic scene classification of interior and exterior urban scenes. However, for a lightweight approach, vector embeddings are also suitable assuming there is an embedding generator model available that is also focused on urban audio.

Future work to follow up on this paper would involve two approaches. The first would be to increase the resolution and accuracy of our comparison of models by using a wider selection of models to see which approach to audio classification works best at different sizes of models. Secondly, there are several datasets available for ASC outside DCASE. Combining multiple datasets may enable the use of larger models. All the avenues of future development mentioned will help acoustic scene classification to be used in wider contexts.

1. References
2. Temko A, Nadeu C, Macho D, et al (2009) Acoustic Event Detection and Classification. In: Computers in the Human Interaction Loop. Springer London, London
3. Liu, Z., Wang, Y., & Chen, T. (1998). The Journal of VLSI Signal Processing, 20(1/2), 61–79. doi:10.1023/a:1008066223044
4. Giannoulis, D., Stowell, D., Benetos, E., Rossignol, M., Lagrange, M., & Plumbley, M. D. (2013, September). A database and challenge for acoustic scene classification and event detection. In *21st European Signal Processing Conference (EUSIPCO 2013)* (pp. 1-5). IEEE.
5. D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M. D. Plumbley. Detection and classification of acoustic scenes and events. IEEE Transactions on Multimedia, 17(10):1733–1746, Oct 2015. [doi:10.1109/TMM.2015.2428998](https://doi.org/10.1109/TMM.2015.2428998).
6. M. Valenti, S. Squartini, A. Diment, G. Parascandolo and T. Virtanen, "A convolutional neural network approach for acoustic scene classification," 2017 International Joint Conference on Neural Networks (IJCNN), Anchorage, AK, 2017, pp. 1547-1554, doi: 10.1109/IJCNN.2017.7966035.
7. Hussain, K., Hussain, M., & Khan, M. G. (2017). Improved acoustic scene classification with DNN and CNN. IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE).
8. J. Abeßer, “A Review of Deep Learning Based Methods for Acoustic Scene Classification,” Applied Sciences, vol. 10, no. 6, p. 2020, Mar. 2020.
9. Toni Heittola, Annamaria Mesaros, and Tuomas 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 (DCASE2020). 2020. Submitted. URL: https://arxiv.org/abs/2005.14623.
10. G. Z. Felipe, Y. Maldonado, G. d. Costa and L. G. Helal, "Acoustic scene classification using spectrograms," 2017 36th International Conference of the Chilean Computer Science Society (SCCC), Arica, 2017, pp. 1-7, doi: 10.1109/SCCC.2017.8405119.
11. S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," in IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 28, no. 4, pp. 357-366, August 1980, doi: 10.1109/TASSP.1980.1163420.
12. McFee, Brian, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. “librosa: Audio and music signal analysis in python.” In Proceedings of the 14th python in science conference, pp. 18-25. 2015.
13. J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, “Audio set: An ontology and human-labeled dataset for audio events”, in IEEE ICASSP, 2017.
14. 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 et al., “CNN architectures for large-scale audio classification”, arXiv preprint arXiv:1609.09430, 2016.
15. Jung, A., Wada, K., Crall, J., Tanaka, S., Graving, J., Reinders, C., Yadav, S., Banerjee, J., Vecsei, G., Kraft, A., Rui, Z., Borovec, J., Vallentin, C., Zhydenko, S., Pfeiffer, K., Cook, B., Fernández, I., De Rainville, F.M., Weng, C.H., Ayala-Acevedo, A., Meudec, R., Laporte, M., & others. (2020). imgaug. <https://github.com/aleju/imgaug>.
16. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).
17. Zhang, X., Zhou, X., Lin, M., & Sun, J. (2018). Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 6848-6856).
18. Zhang, T., Liang, J., & Ding, B. (2020). Acoustic scene classification using deep CNN with fine-resolution feature. Expert Systems with Applications, 143, 113067. doi:10.1016/j.eswa.2019.113067