ACOUSTIC SCENE CLASSIFICATION USING FREQUENCY-AWARE CNN WITH METRIC-BASED FEW-SHOT LEARNING

Project Report Phase -II

submitted in partial fulfilment of the requirements for the award of the degree of

MASTER OF TECHNOLOGY

in

COMPUTER SCIENCE & ENGINEERING

Submitted By

ROOPTEJA K

Reg. No. – 22376005

Under the guidance of

Dr. S. L. JAYALAKSHMI

Assistant Professor

PONDICHERRY UNIVERSITY

DEPARTMENT OF COMPUTER SCIENCE

SCHOOL OF ENGINEERING AND TECHNOLOGY

PUDUCHERRY – 605014

INDIA

May 2024

BONAFIDE CERTIFICATE

This is to certify that the project report entitled " **Acoustic Scene Classification using Frequency-aware CNN with Metric-based Few-Shot Learning**" submitted by **Roopteja K** bearing **Reg No: 22376005**, in completion of his work under the guidance of **Dr. S. L. Jayalakshmi** is accepted for the project report submission in partial fulfilment of the requirements for the award of the degree of Master of Technology in Computer Science and Engineering in the Department of Computer Science, Pondicherry University, Puducherry during the academic year of 2023-24.

Signature of Head of the Department

Dr. S. K. V. JAYAKUMAR
Professor/HOD
Department of Computer Science
Pondicherry University
Puducherry, India 605014

Signature of the Guide

Dr. S. L. JAYALAKSHMI
Assistant Professor
Department of Computer Science
Pondicherry University
Puducherry, India 605014

DEPARTMENT OF COMPUTER SCIENCE
SCHOOL OF ENGINEERING & TECHNOLOGY
PONDICHERRY UNIVERSITY
PUDUCHERRY – 605014
INDIA

CERTIFICATE

Certified that the Project report entitled "**Acoustic Scene Classification using Frequency-aware CNN with Metric-based Few-Shot Learning** " submitted by **ROOPTEJA K**, Reg.No: **22376005** to Pondicherry University in the Department of Computer Science is a record of research work carried out under the guidance of **Dr. S. L. Jayalakshmi** is worthy of consideration for the award of the degree of Master of Technology in Computer Science and Engineering in the Department of Computer Science, Pondicherry University, Puducherry during the academic year of 2023-24.

Dr. S. L. JAYALAKSHMI
Assistant Professor
Department of Computer Science
Pondicherry University
Puducherry, India 605014

ACKNOWLEDGEMENT

It is my great pleasure to express my gratitude to all of them who have provided me guidance and support in bringing out the successful completion of my Project course at Pondicherry University. It would not have been possible for me to achieve success with my effort alone. There were a lot of people who extended their supportive hands towards me in theway of making my work a success. I would like to thank everyone who supported and assisted me while carrying out this work at Pondicherry University.

First of all, I would like to thank my Guide Dr. S. L. Jayalakshmi for her constant support, trust, valuable feedback, encouragement and innumerable advice. She gave freedom to pursue my ideas and work at my own pace and was always available to discuss various problems on the way. Her encouragement and guidance have provided a good basis for completion of my Project course.

I am highly obliged to all the faculty members and non-teaching staff of the Department of Computer Science for their support and encouragement. I also thank Dr. K. Tharanikkaarasu, Vice Chancellor of Pondicherry University, Dr. S. Sivasathya, Dean, School of Engineering and Technology and Dr. S. K. V. Jayakumar, Head, Department of Computer Science for providing excellent computing and other facilities without which this work could not achieve its quality goal.

I will ever remain grateful to all my family members for their unconditional love and support. I am deeply indebted to them who has been constantly encouraging and supporting me in everywalk of my life.

Finally, I would like to thank all those who have directly or indirectly helped me in different capacities to complete my project course.

~ROOPTEJA K

Acoustic Scene Classification using Frequency-aware CNN with Metric-based Few-Shot Learning

Abstract:

Acoustic scene classification (ASC) is a vital task in audio signal processing, aiming to categorize environmental sounds into predefined classes. Traditional approaches to ASC often neglect the variation associated with overlapping classes across different acoustic scenes. This research deviates from the conventional by addressing not only the distinct classes but also focusing on the challenges posed by overlapping categories. This work presents the Few-Shot Learning technique to generalize effectively with a limited number of examples by their similarity as the closer distance range, making it well-suited for scenarios where comprehensive labeled datasets are impractical to obtain. The proposed approach combines Metric-Based Few-Shot Learning, Log-Mel Spectrogram feature extraction, and Frequency-Aware CNN classification, to find a solution for comprehensive and challenged labeled datasets in the smart home, and audio surveillance applications. Experimental evaluation conducted on the TAU Urban Acoustic Scenes 2022 Mobile dataset demonstrates a classification accuracy of 81.97% and it outperforms that of other conventional methods used in the existing literature.

References:

1. Mahmoud A. Alamir, “A novel acoustic scene classification model using the late fusion of convolutional neural networks and different ensemble classifiers”, Applied Acoustics 175 (2021)
2. Tao Zhang, Jinhua Liang, Biyun Ding, “Acoustic scene classification using deep CNN with fine-resolution feature”, Expert Systems With Applications 143 (2020) 113067
3. Nisan Aryal a, Sang-Woong Lee, “Frequency-based CNN and attention module for acoustic scene classification”, Applied Acoustics 210 (2023) 109411
4. T. Zhang and J. Wu, “Constrained learned feature extraction for acoustic scene classification”, IEEE/ACM Transactions on Audio, Speech, and Language Processing, Vol. 27, No. 8, pp. 1216-1228, 2019.
5. Paseddula C, Gangashetty SV. Late fusion framework for Acoustic Scene Classification using LPCC, SCMC, and log-Mel band energies with Deep Neural Networks. Appl Acoust 2021;172:107568. https://doi.org/10.1016/j.apacoust.2020.107568
6. Kosmider M. Spectrum correction: Acoustic scene classification with mismatched recording devices. Proc. Interspeech 2020 2020:4641–5.
7. Mie Mie Oo, Nu War, “Acoustic Scene Classification using Attention based Deep Learning Model”, International Journal of Intelligent Engineering and Systems, Vol.15, No.6, 2022
8. Javier Naranjo-Alcazar, Sergi Perez-Castanos, Pedro Zuccarello, Ana M. Torres, Jose J. Lopez, Francesc J. Ferri, Maximo Cobos, “An Open-Set Recognition and Few-Shot Learning Dataset for Audio Event Classification in Domestic Environments”, Pattern Recognition Letters 164 (2022) 40–45
9. Wang, Y., Salamon, J., Bryan, N. J., & Pablo Bello, J. (2020). Few-Shot Sound Event Detection. ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). doi:10.1109/icassp40776.2020.9054708

10. Biyun Ding, Tao Zhang, Ganjun Liu, Lingguo Kong, Yanzhang Geng, “Late fusion for acoustic scene classification using swarm intelligence”, Applied Acoustics 192 (2022) 108698
11. Ines Nolasco, Shubhr Singh, Veronica Morfi, Vincent Lostanlen, Ariana StrandburgPeshkin, Ester Vidana-Vila, Lisa Gill, Hanna Pamuła, Helen Whitehead, Ivan Kiskin, Frants H. Jensen, Joe Morford, Michael G. Emmerson, Elisabetta Versace, Emily Grout, Haohe Liu, Burooj Ghani, Dan Stowell, “Learning to detect an animal sound from five examples”, Ecological Informatics 77 (2023) 102258
12. Shefali Waldekar, Goutam Saha, “Two-level fusion-based acoustic scene classification”, Applied Acoustics 170 (2020) 107502
13. Chris Careaga, Brian Hutchinson, Nathan Hodas, Lawrence Phillips, “Metric-Based Few-Shot Learning for Video Action Recognition”, Computer Vision and Pattern Recognition arXiv:1909.09602v1
14. Alexander Rakowski, Michał Kosmider, “FREQUENCY- AWARE CNN FOR OPEN SET ACOUSTIC SCENE CLASSIFICATION”, Audio Intelligence, Detection and Classification of Acoustic Scenes and Events 2019
15. A. Mesaros, T. Heittola, and T. Virtanen, “A multi-device dataset for urban acoustic scene classification,” in Proceedings of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE2018), November 2018, pp. 9–13. [Online]. Available: https://arxiv.org/abs/1807.09840
16. A. Mesaros, T. Heittola, A. Diment, B. Elizalde, A. Shah, E. Vincent, B. Raj, and T. Virtanen, “Dcase 2017 challenge setup: Tasks, datasets and baseline system,” in DCASE 2017- Workshop on Detection and Classification of Acoustic Scenes and Events, 2017.
17. Yerin Lee, Soyoung Lim, Il-Youp Kwak, “CNN-Based Acoustic Scene Classification System”, Electronics 2021, 10(4), 371
18. Gao W, McDonnell M. Acoustic scene classification using deep residual networks with late fusion of separated high and low frequency paths, Tech. rep., DCASE2019 Challenge (June 2019).
19. Liping Yang, Lianjie Tao, Xinxing Chen, Xiaohua Gu, “Multi- scale semantic feature fusion and data augmentation for acoustic scene classification”, Applied Acoustics 163 (2020) 107238.
20. Lam Pham, Huy Phan, Truc Nguyen, Ramaswamy Palaniappan, Alfred Mertins, Ian McLoughlin, “Robust acoustic scene classification using a multi-spectrogram encoder-decoder framework”, Digital Signal Processing 110 (2021) 102943
21. Yan Leng, Weiwei Zhao, Chan Lin, Chengli Sun, Rongyan Wang, Qi Yuan, Dengwang L, “LDA-based data augmentation algorithm for acoustic scene classification”, Knowledge-Based Systems 195 (2020) 105600
22. Sumit Kumar Chaudhary, Sameer Saran, “Information Network (IBIN) framework”, Environmental Sustainability, doi.org/10.1007/s42398-023-00281-w
23. Sayeh Mirzaei, Iman Khani Jazani, ” Acoustic scene classification with multi-temporal complex modulation spectrogram features and a convolutional LSTM network”, Multimedia Tools and Applications (2023) 82:16395–16408
24. Tao Zhang, Jinhua Liang, Guoqing Feng, ” Adaptive time- frequency feature resolution network for acoustic scene classification”, Applied Acoustics 195 (2022) 108819
25. Yu Wang, Nicholas J. Bryan, Mark Cartwright, Juan Pablo Bello, Justin Salamon, “FEW-SHOT CONTINUAL LEARNING FOR AUDIO CLASSIFICATION”, Speech and Signal Processing (ICASSP), 10.1109/ICASSP39728.2021.9413584

26. Yan Gao, Haijiang Li, Weiqi Fu, “Few-shot learning for image- based bridge damage detection”, Engineering Applications of Artificial Intelligence 126 (2023) 107078.
27. Farong Gao, Lijie Cai, Zhangyi Yang, Shiji Song, Cheng Wu, “Multi-distance metric network for few-shot learning”, International Journal of Machine Learning and Cybernetics (2022) 13:2495–2506
28. Wei Xie, Yanxiong Li, Qianhua He , Wenchang Cao, “Few-shot class-incremental audio classification via discriminative prototype learning”, Expert Systems With Applications 225 (2023) 120044
29. Chandrasekhar Paseddula, Suryakanth V. Gangashetty, “Late fusion framework for Acoustic Scene Classification using LPCC, SCMC, and log-Mel band energies with Deep Neural Networks”, Applied Acoustics 172 (2021) 107568
30. Yuzhong Wu, Tan Lee, “ENHANCING SOUND TEXTURE IN CNN-BASED ACOUSTIC SCENE CLASSIFICATION”, Speech and Signal Processing (ICASSP), 10.1109/icassp.2019.8683490
31. Zhao Ren, Kun Qian, Zixing Zhang, Vedhas Pandit, Alice Baird,” Deep Scalogram Representations for Acoustic Scene Classification”, AUTOMATICA SINICA, VOL. 5, NO. 3(2018)
32. Gao W, McDonnell M. Acoustic scene classification using deep residual networks with focal loss and mild domain adaptation, Tech. rep., DCASE2020 Challenge (June 2020).
33. Kong Q, Cao Y, Iqbal T, Wang W, Plumbley MD. Cross-task learning for audio tagging, sound event detection and spatial localization: DCASE 2019 baseline systems, Tech. rep., DCASE2019 Challenge (June 2019).
34. Z. Ren, V. Pandit, K. Qian, Z. J. Yang, Z. X. Zhang, and B. Schuller, “Deep sequential image features for acoustic scene classification,” in Proc. Detection and Classification of Acoustic Scenes and Events, Munich, Germany, 2017, pp. 113−117.
35. M. Valenti, A. Diment, G. Parascandolo, S. Squartini, and T. Virtanen, “DCASE 2016 acoustic scene classification using convolutional neural networks,” in Proc. of Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE), Budapest, Hungary, 2016.
36. A. Mesaros, T. Heittola, and T. Virtanen, “TUT database for acoustic scene classification and sound event detection,” Proc. of the 24th Acoustic Scene Classification Workshop 2016 European Signal Processing Conference (EUSIPCO), 2016.

Illustrations

Proposed Methodology

Fig. 1 shows the architecture of proposed model. In order to reduce the amount of audio files and create a new dataset of similar audios based on similarity, the raw audio files were submitted to a few shot learning method as a metric-based approach. Next, the recently generated dataset will be used as the Log-mel spectrogram input for the feature extraction technique. Frequency-aware convolutional neural networks (FACNNs) will receive the extracted features and use them to classify the scenes.

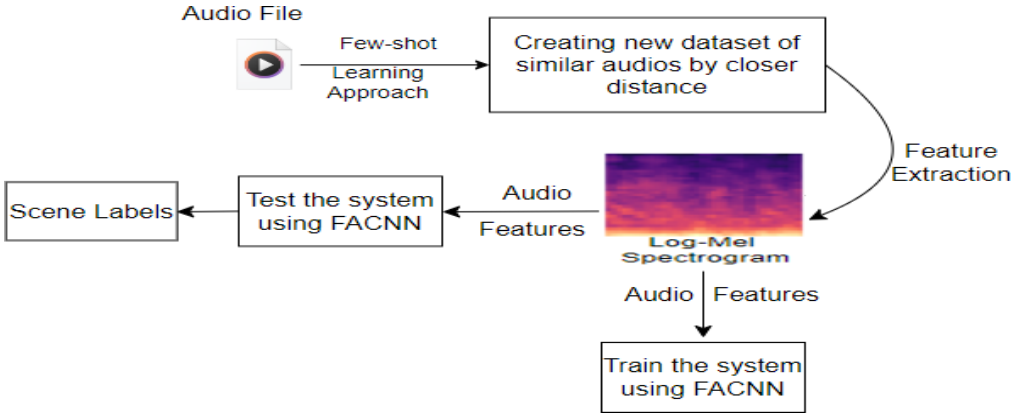


Fig.1. Architecture of Proposed Model

Metric-based Approach:

The metric-based approach is employed to diminish the volume of audio files with similar content, utilizing the closest distance metric between audio files as a criterion. The procedure unfolds in distinct steps, commencing with the input being the waveform. The first checkpoint involves verifying the sampling rate of the selected pair of audio files. If the rates match, the Euclidean distance between the raw audio waveforms of the two files is calculated. In cases where the rates differ, the comparison cannot proceed. The calculation of the Euclidean distance between the audio files is executed to determine their similarity can be expressed by

$$D(audio\_1, audio\_2) = \sqrt{(audio\_1[i] - audio\_2[i])^2} \quad (1)$$

Subsequently, the new dataset is updated, incorporating files within a closer distance range, thereby refining the set of audio files based on their metric proximity.

Feature Extraction:

Prior to network training, the input audios are pre-processed into Log-Mel spectrograms, and Fig. 2 illustrates this procedure. These spectrograms can be extracted using a Mel frequency bank, and then scaled using a logarithm. Spectrograms the size of 44 Mel frequency bins were used in this study. The sample frequency of the audio file is 44100 Hz. The window lengths have frame durations of 80 ms and 40 ms, with a 50% overlap length. Eq. (2) contains the computation for the number of frame lengths.

$$f = 1 + \text{floor}\,[(L-W) / S] \quad (2)$$

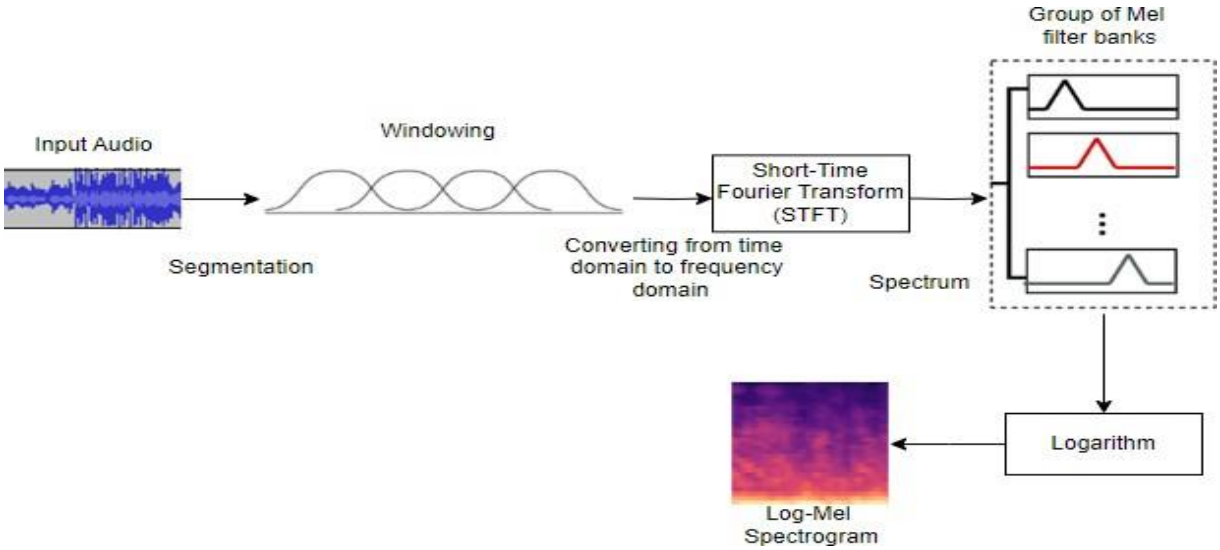


Fig. 2. Block diagram of Log- Mel spectrogram of an Audio signal

Where f is the number of frames, L is signal length, W is window length and S is shift length. To get a smooth spectral representation from the frames, a Hamming window is applied to a frame to reduce the impact of noise at the edges of the window. The following equation Eq. (3) defines the Hamming window $h(n)$:

$$h(n) = 0.54 - 0.46 \cos(2\pi n / (N-1)) \; 1 \leq n \leq N \qquad (3)$$

Where N is frame number, and $h(n)$ is the hamming window. This function is used to estimate the Fast Fourier Transform (FFT- Eq. (4) for getting an amplitude-frequency response per frame.

$$S_i(k) = \sum_{i=1}^{N} s_i(k) h(n) e^{-\frac{j2\pi kn}{N}} \; n = 0, .., N-1 \qquad (4)$$

Where (f) is the frequency of mel and f is frequency of linear frequency. Mel filter is used to obtain the energy from the filter bank.

Frequency-aware convolutional neural network (FACNN):

A novel method for prioritizing frequency components by avoiding downsampling along the frequency axis is shown by the FACNN architecture shown in Fig. 3. In order to ensure large receptive fields along the temporal axis, this design choice prioritizes the preservation of fine-grained frequency information. Through the maintenance of smaller receptive fields in the frequency domain and larger ones in the time domain, the design successfully combines local frequency information from the input Time-Frequency representation with global temporal context. By combining global and local inputs, deeper layers are able to identify and extract significant characteristics, which improves the model's capacity for sophisticated analysis and synthesis in a range of research applications.

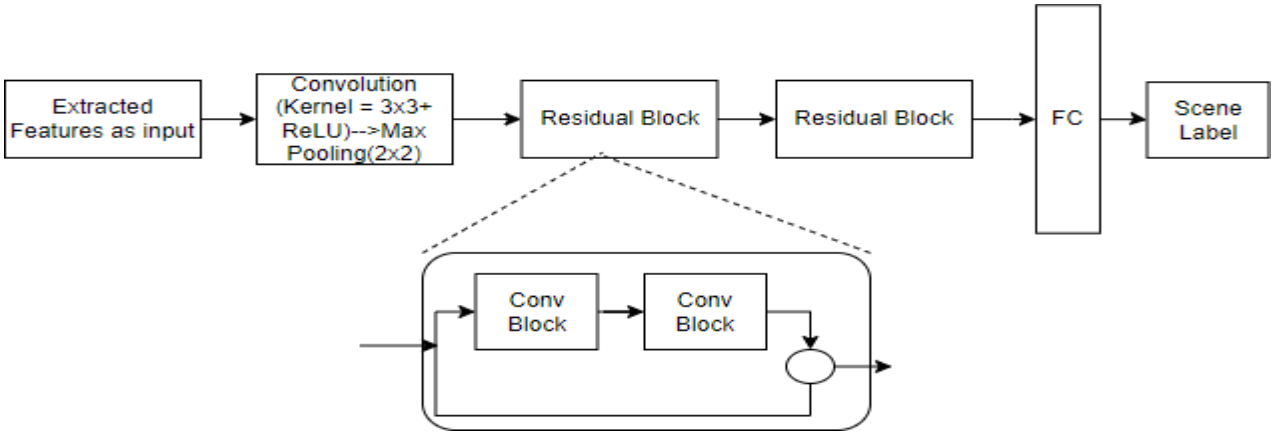


Fig. 3. Detailed architecture of FACNN

Inspired by the work of Gao and McDonnell [18], we have designed a proposed architecture that significantly deviates from traditional residual networks. Notably, we have deviated from conventional residual networks, which usually downsample on both the time and frequency axes, by choosing not to do downsampling on the frequency axis. Further, we employ a pre-activation structure in our convolution block. Using batch normalization, ReLU activation, and convolution are the steps involved in this process. The bias term in the FACNN convolution stage is omitted. In addition, as Fig. 4 shows, down-sampling in the identity path in our residual blocks is carried out by average pooling and channel padding. In conclusion, we substitute 1 x 1 convolutional layers for the traditional fully linked layers at the network's end. These architectural modifications are designed to enhance the model's ability to capture and leverage intricate temporal and frequency features, contributing to its efficacy in research applications.

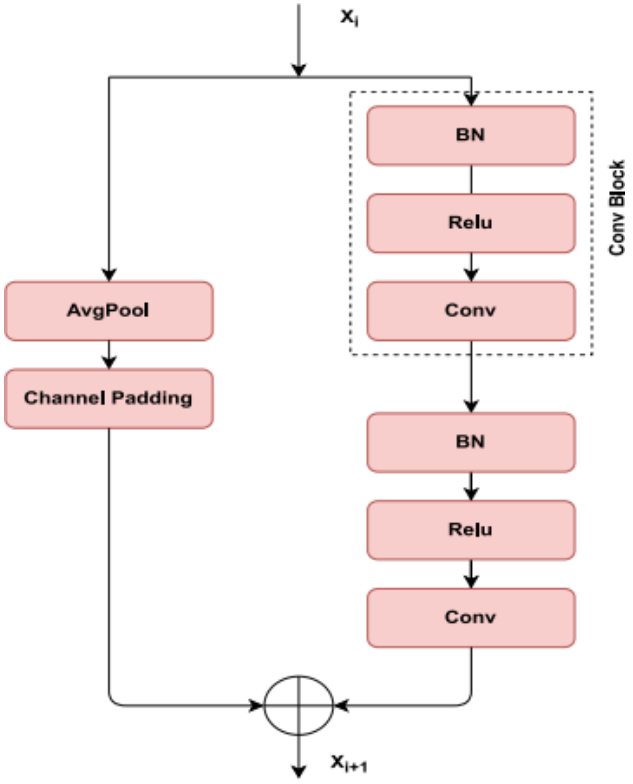


Fig. 4. Residual block with pre-activation. The BN and ReLU is applied before convolution in the conv block.

The essential components of the FACNN architecture are the residual blocks, which are made up of two 3 x 3 convolutional layers apiece. Using pre-activation convolution as the basis, the first layer concentrates on batch normalization. Feature maps are expanded to 32 by the first convolutional layer through downsampling along the time axis. Then, sixteen residual blocks are added, which are arranged into four residual levels. Each layer's first residual block undergoes downsampling, which progressively increases the feature maps from 32 to 64, 128 to 256. A dual-layer, 1 x 1 convolutional procedure takes place after residual layers (post-residual layer). The first convolutional layer increases feature maps from 256 to 512, while the second one decreases them to 3, in accordance with the number of classes in the dataset. Together, these architectural choices improve the model's ability to identify intricate details and promote reliable results in a variety of study situations.

Codes

Implementation:

```
import os
import librosa
import numpy as np
from pandas.core.dtypes.common import classes
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import LabelEncoder
from tensorflow.keras import layers, models, callbacks
import matplotlib.pyplot as plt
from sklearn.metrics import accuracy_score, confusion_matrix
import seaborn as sns
```

Fig. 5. Importing Libraries

```
def calculate_distance(audio_data1, audio_data2):
    return euclidean(audio_data1, audio_data2)
```

Fig. 6. Calculating the distance between the audio files

```
if sr1 != sr2:
    print(f"Skipping comparison between {file1} and {file2} due to different sampling rates.")
    continue

distance = calculate_distance(audio_data1, audio_data2)
if distance < min_distance:
    min_distance = distance
if distance > max_distance:
    max_distance = distance

print(f"Distance between {file1} and {file2}: {distance}")

if distance < 10:
    selected_audios.append(audio_files[i])
    selected_audios.append(audio_files[j])
```

Fig. 7. Sampling rate check & Min and Max distance &
Range of the audio files for new dataset

```
new_dataset_path = "C:/Users/Dell/PycharmProjects/BL_2/dataset_1/airport/dataset_u"
with open(new_dataset_path, "w") as f:
    for audio_file in selected_audios:
        f.write(audio_file + "\n")
```

Fig. 8. Creation of new dataset

```python
def extract_features(file_path, n_mels=44, hop_length=512, n_fft=2048):
    y, sr = librosa.load(file_path)
    mel_spectrogram = librosa.feature.melspectrogram(y=y, sr=sr, n_mels=n_mels, hop_length=hop_length, n_fft=n_fft)
    log_mel_spectrogram = librosa.power_to_db(mel_spectrogram)
    return log_mel_spectrogram
```

Fig. 9. Function to extract log-mel spectrogram features

```python
def load_data_and_labels(data_dir):
    data = []
    labels = []
    label_encoder = LabelEncoder()

    for folder in os.listdir(data_dir):
        folder_path = os.path.join(data_dir, folder)
        if os.path.isdir(folder_path):
            for filename in os.listdir(folder_path):
                file_path = os.path.join(folder_path, filename)
                if file_path.endswith(".wav"):
                    features = extract_features(file_path)
                    data.append(features)
                    labels.append(folder)

    # Encode labels
    encoded_labels = label_encoder.fit_transform(labels)

    return np.array(data), np.array(encoded_labels), label_encoder.classes_
```

Fig. 10. Function to load data and labels

```python
data_dir = "C:/Users/Dell/PycharmProjects/BL_2/dataset_u"
data, labels, classes = load_data_and_labels(data_dir)
```

Fig. 11. Load data and labels

```
model = models.Sequential()
model.add(layers.Conv2D( filters: 32,  kernel_size: (3, 3), activation='relu', input_shape=(X_train.shape[1], X_train.shape[
model.add(layers.BatchNormalization())
model.add(layers.MaxPooling2D((2, 2)))
model.add(layers.Conv2D( filters: 64,  kernel_size: (3, 3), activation='relu'))
model.add(layers.BatchNormalization())
model.add(layers.MaxPooling2D((2, 2)))
model.add(layers.Conv2D( filters: 128,  kernel_size: (3, 3), activation='relu'))
model.add(layers.BatchNormalization())
model.add(layers.MaxPooling2D((2, 2)))
model.add(layers.Flatten())
model.add(layers.Dense( units: 128, activation='relu'))
model.add(layers.Dense(len(np.unique(labels)), activation='softmax'))
```

Fig. 12. Build Frequency-aware CNN model

```
plt.figure(figsize=(15, 10))

for i, scene_class in enumerate(classes):
    scene_files = [os.path.join(data_dir, scene_class, f) for f in os.listdir(os.path.join(data_dir, scene_class)) 
    example_file = scene_files[0]  # Take the first file for each scene as an example

    # Extract log-mel spectrogram
    y, sr = librosa.load(example_file)  # Get the sample rate
    example_features = extract_features(example_file)

#    print(f'Test Sample for {scene_class}: {example_file}')

    # Plot the log-mel spectrogram
    plt.subplot( *args: 2, 3, i + 1)  # Adjust the subplot layout based on the number of scenes
    librosa.display.specshow(example_features, x_axis='time', y_axis='mel', sr=sr, fmax=8000)
    plt.colorbar(format='%+2.0f dB')
    plt.title(f'Log-Mel Spectrogram - {scene_class}')

plt.tight_layout()
plt.show()
```

Fig. 13. Visualize log-mel spectrogram for one audio in each scene

```
# Train the model
model.fit(X_train, y_train, epochs=10, validation_data=(X_test, y_test), callbacks=[callbacks.EarlyStopping(patience=

# Evaluate the model
test_loss, test_acc = model.evaluate(X_test, y_test)
print(f'Test Accuracy: {test_acc}')
```

Fig. 14. Train and Evaluate the model

```
# Calculate accuracy
accuracy = accuracy_score(y_test, y_pred)
print(f'Accuracy: {accuracy}')

# Calculate confusion matrix
conf_matrix = confusion_matrix(y_test, y_pred)

# Plot confusion matrix
plt.figure(figsize=(8, 8))
sns.heatmap(conf_matrix, annot=True, fmt='d', cmap='Blues', xticklabels=classes, yticklabels=classes)
plt.title('Confusion Matrix')
plt.xlabel('Predicted Label')
plt.ylabel('True Label')
plt.show()
```

Fig. 15. Accuracy & Confusion Matrix

Acoustic Scene Classification using Frequency-aware CNN with Metric-based Few-Shot Learning

Submitted by,
Name: K. Roopteja,
Roll No.: 22376005,
Project Guide: Dr. S. L. Jayalakshmi

Problem Definition

- Introduction
  - Acoustic Scene Classification (ASC): a field of audio signal processing that involves the categorization of an audio recording into predefined acoustic scenes or environments.
  - The scenes comprise the categories such as
    - indoor (residence, restaurant/cafe)
    - outdoor (park, metro station)
    - transportation (metro, bus, tram)
  - Applications:
    - Surveillance and Security
    - Emergency Response
- To enhance the performance of ASC model using the few shot learning to classify different overlapping classes of various acoustic scenes.

Base Paper

Title: A novel acoustic scene classification model using the late fusion of convolutional neural networks and different ensemble classifiers [Applied Acoustics 175 (2021) 107829]

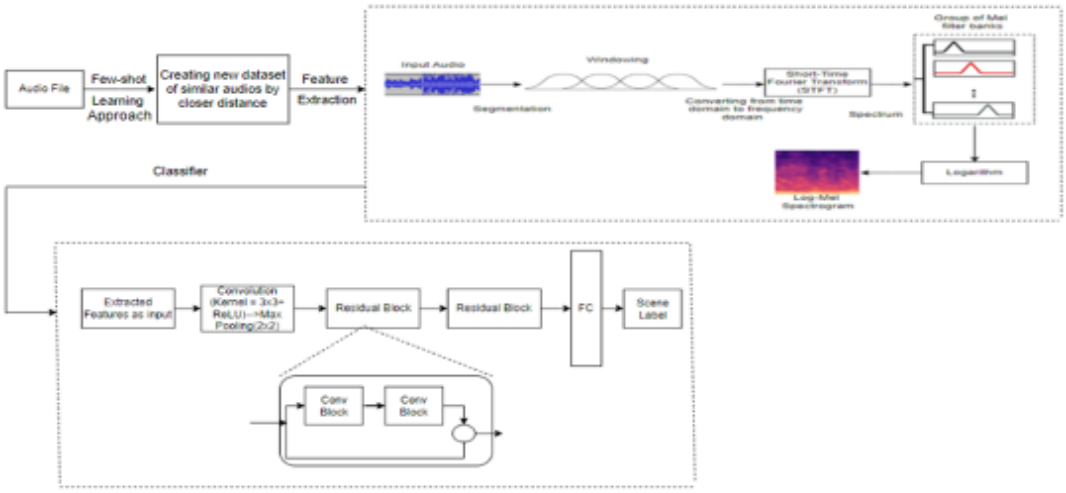
Authors: [Mahmoud A. Alamir]

- ❑To enhance ASC accuracy by introducing late fusion models that combine convolutional neural networks (CNNs) with ensemble classifiers
- ❑Data Augmentation:
  - Splitting 10sec sample to 10 one sec samples.
  - Mix-up
- ❑Feature extraction:
  - Mel-spectrograms
  - wavelet scattering
- ❑Classifier
  - Late fusion model – CNN and ensemble classifier models

Proposed System

- To propose a model to classify the different overlapping classes in an acoustic environmental scene with the help of Few-shot learning technique.
- Proposed Methodology
  - Metric-based Few-shot learning
  - Feature Extraction
    - Log-Mel spectrogram
  - Classification
    - FACNN

Design



Dataset

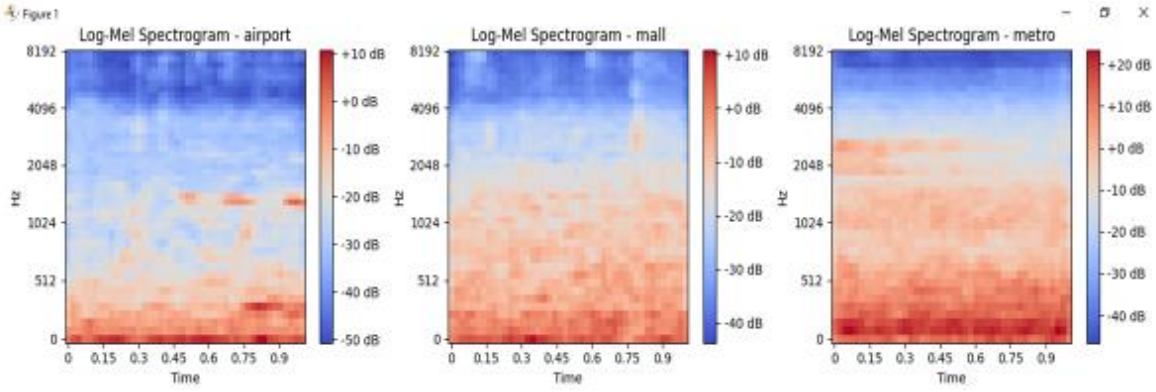
- TAU Urban Acoustic Scenes 2022 Mobile development dataset consists of 1-seconds audio segments acoustic scenes:
  - Airport - *airport*
  - Indoor shopping mall - *shopping_mall*
  - Metro station - *metro_station*
- Audio data was recorded in Lisbon, Prague, Paris, etc.
- Audio is provided in a single-channel 44.1kHz 24-bit format.

Experimental Setup

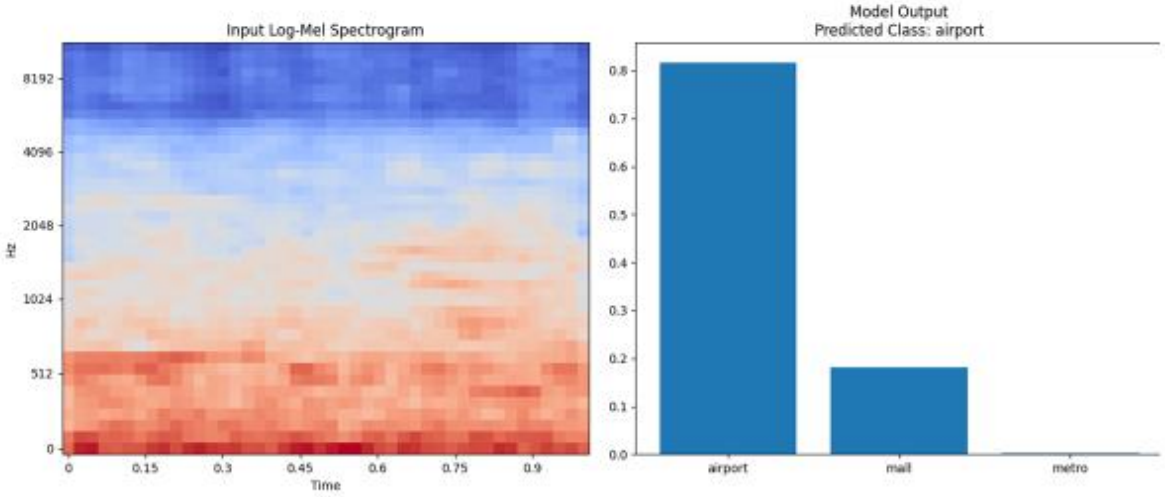
Hardware Requirements:

- Processor: Intel Core i3 or higher
- RAM: Minimum 4 GB (8 GB or higher recommended for optimal performance)
- Storage: At least 20 GB of available disk space
- **Software Requirements:**
- Operating System: Windows 10 (64-bit), macOS, or Linux distribution
- Integrated Development Environment (IDE): PyCharm Community Edition
- Python Version: Python 3. 7(preferably the latest stable version)
- Libraries:
  - TensorFlow or PyTorch for deep learning model implementation
  - NumPy for numerical computations
  - Librosa for audio processing and feature extraction
  - Matplotlib for visualization of results

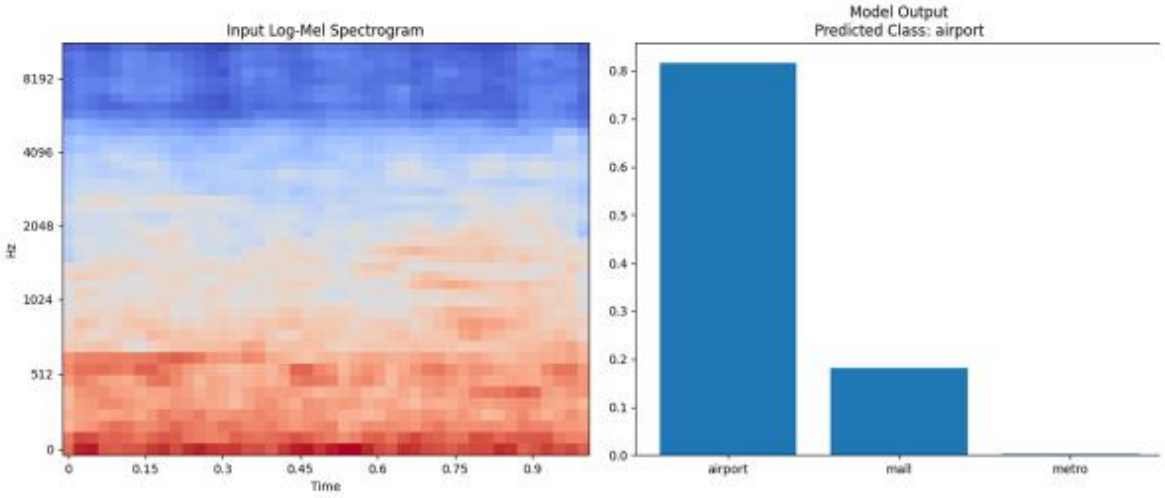
Results



Cntd..


Input Log-Mel Spectrogram
Hz
8192
4096
2048
1024
512
0
0
0.15
0.3
0.45
0.6
0.75
0.9
Time
Model Output
Predicted Class: airport
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
airport
mail
metro

Cntd..


Input Log-Mel Spectrogram
Hz
8192
4096
2048
1024
512
0
0
0.15
0.3
0.45
0.6
0.75
0.9
Time
Model Output
Predicted Class: airport
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
airport
mail
metro

Comparison of different systems

Features	System	Accuracy(%)
Log-Mel	CNN	77.45
MFCC	CNN	75.7
Spectrogram	CNN	74.95
Log-Mel	**FACNN**	**81.97**

Conclusion

- These techniques will improve classification accuracy, capture important audio features, and perform better with less labelled data.
- High classification accuracy was demonstrated using the TAU Urban Acoustic Scenes 2022 Mobile dataset and demonstrated resilience in diverse environments.

Future Work

- Practical implementation in real-world scenarios, such as audio surveillance, emergency response needs to be explored.
- Enhancing the performance in complex acoustic environments.
- Research to extend model performance across larger and more diverse datasets.

Reference

1. Mahmoud A. Alamir, "A novel acoustic scene classification model using the late fusion of convolutional neural networks and different ensemble classifiers", Applied Acoustics 175 (2021)
2. Tao Zhang, Jinhua Liang, Biyun Ding, "Acoustic scene classification using deep CNN with fine-resolution feature", Expert Systems With Applications 143 (2020) 113067
3. Nisan Aryal a, Sang-Woong Lee, "Frequency-based CNN and attention module for acoustic scene classification", Applied Acoustics 210 (2023) 109411
4. T. Zhang and J. Wu, "Constrained learned feature extraction for acoustic scene classification", IEEE/ACM Transactions on Audio, Speech, and Language Processing, Vol. 27, No. 8, pp. 1216-1228, 2019.
5. Paseddula C, Gangashetty SV. Late fusion framework for Acoustic Scene Classification using LPCC, SCMC, and log-Mel band energies with Deep Neural Networks. Appl Acoust 2021;172:107568. https://doi.org/10.1016/j.apacoust.2020.107568
6. Kosmider M. Spectrum correction: Acoustic scene classification with mismatched recording devices. Proc. Interspeech 2020 2020:4641–5.
7. Mie Mie Oo, Nu War, "Acoustic Scene Classification using Attention based Deep Learning Model", International Journal of Intelligent Engineering and Systems, Vol.15, No.6, 2022
8. Javier Naranjo-Alcazar, Sergi Perez-Castanos, Pedro Zuccarello, Ana M. Torres, Jose J. Lopez, Francesc J. Ferri, Maximo Cobos, "An Open-Set Recognition and Few-Shot Learning Dataset for Audio Event Classification in Domestic Environments", Pattern Recognition Letters 164 (2022) 40–45
9. Wang, Y., Salamon, J., Bryan, N. J., & Pablo Bello, J. (2020). Few-Shot Sound Event Detection. ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). doi:10.1109/icassp40776.2020.9054708
10. Biyun Ding, Tao Zhang, Ganjun Liu, Lingguo Kong, Yanzhang Geng, "Late fusion for acoustic scene classification using swarm intelligence", Applied Acoustics 192 (2022) 108698

Acoustic Scene Classification using Frequency-aware CNN with Metric-based Few-Shot Learning

Roopteja. K.,
Department of Computer Science
Pondicherry University
Puducherry, India
roopteja35@gmail.com

Dr. S. L. Jayalakshmi,
Department of Computer Science
Pondicherry University
Puducherry, India
sathishjayalakshmi02@*pondiuni.ac.in*

***Abstract*— Acoustic scene classification (ASC) is a vital task in audio signal processing, aiming to categorize environmental sounds into predefined classes. Traditional approaches to ASC often neglect the variation associated with overlapping classes across different acoustic scenes. This research deviates from the conventional by addressing not only the distinct classes but also focusing on the challenges posed by overlapping categories. This work presents the Few-Shot Learning technique to generalize effectively with a limited number of examples by their similarity as the closer distance range, making it well-suited for scenarios where comprehensive labeled datasets are impractical to obtain. The proposed approach combines Metric-Based Few-Shot Learning, Log-Mel Spectrogram feature extraction, and Frequency-Aware CNN classification, to find a solution for comprehensive and challenged labeled datasets in the smart home, and audio surveillance applications. Experimental evaluation conducted on the TAU Urban Acoustic Scenes 2022 Mobile dataset demonstrates a classification accuracy of 81.97% and it outperforms that of other conventional methods used in the existing literature.**

***Index Terms*— Deep Learning, Metric-based Few-Shot learning, FACNN, Acoustic Classification, Log-Mel Spectrogram.**

I. Introduction

Acoustic scene classification (ASC) aims at enabling devices to recognize an audio scene, either from a recording or an on-line stream [15]. The term "scene" in this context refers to the idea of a particular auditory environment (such as a metro station, street traffic, public space, etc.) that is perceived and characterized by people. Stated differently, a scene is the combination of ambient sounds and audible elements linked to a particular auditory situation [17]. Smart homes, robotics, audio surveillance, context-aware mobile devices, music genre classification, and many other real-world applications are possible using ASC. For example, in surveillance, ASC can differentiate between sounds from metropolitan streets, sounds from rural areas, or sounds from certain events, adding an essential level of contextual knowledge [15]. ASC influences decisions by monitoring and controlling urban soundscapes in the context of smart cities.

A machine learning paradigm known as "few-shot learning" is intended to train models in situations where each class has a finite number of labelled samples [13]. Obtaining vast labelled datasets might be difficult or impractical in certain cases, making traditional machine learning algorithms limited. These approaches typically require a substantial amount of labelled data for training. Models can learn from a small number of examples, which makes it more viable and practical. It will be helpful to quickly adapt to new classes or categories without the need to retrain the entire model on a big dataset. Image recognition, audio signal processing, natural language processing (NLP), and medical imaging all use few-shot learning.

This work proposes as follows:

i. Metric-based Few-Shot Learning, helps to reduce the number of examples.
ii. Apply Log-mel spectrogram as feature extraction method which gives the visual representation of an audio sample.
iii. Using Frequency-aware CNN as classifier, it classifies the audio file from which environment.

II. Literature Review

Research on the ASC has primarily concentrated on specific classes in recent years. ASC's earlier research has mostly concentrated on spectral features. The majority of the applications for these properties are speech-related, including speaker verification, voice recognition, and so on. Mel spectrograms, Mel Frequency Cepstral Coefficients (MFCCs), Log-Mel Spectrograms [7], [8], Constant-Q transform spectrograms [4], and Gammatone Frequency Cepstral Coefficient (GFCC) [7] are the most often utilised spectral features. Mel Spectrograms are the feature that is most frequently used [1]. The audio samples were converted into frequency-domain representation using Mel Spectrograms. The fusion model is the focus of recent efforts in order to achieve great accuracy.

The fusion of CNN and ensemble classifier models, which maximizes accuracy, attracted the interest of late fusion models [1], [5], [10], and [12]. Other methods achieved by combining CNN model with other models, like SVM, or different CNN models with different feature extraction methods, are also included in the list. Swarm Intelligence algorithm based Late Fusion method (SILF) [10]. The favourable performance of convolutional neural networks with spectrogram feature representation for auditory scene classification is drawing increasing interest. A few papers [2], [3] deal with the representation of frequency features.

The advancements in very deep architecture, feature fusion,

and convolutional operation are proposed to be embraced by the fine-resolution convolutional neural network (FRCNN) [2]. By concentrating on the frequency information of the audio samples, frequency-aware convolutional neural network (FACNN) can resolve the device mismatch [3], [6] issue. An attention module can function as a frequency attention network (FANet) to produce an attention map based on the frequency information of the input feature maps. FANet assists the FACNN in concentrating on the crucial frequency data in order to enhance performance.

Publish Year	Reference	Features	Classification	Merits	Demerits	Research Gap
2022	[1]	DA: Mixup FE: Mel-spectrograms, wavelet scattering	CNN, Ensemble Classifiers	Improved Accuracy	High complexity, High resource requirement	Exploration of Hybrid Models
2020	[2]	DA: - FE: Log- Mel spectrograms	FRCNN	Depth-Wise Separable Convolution	High complexity, Fine-Tuning Challenges	Automatic Recognition of Temporal Feature Resolutions
2023	[3]	DA: Mix-up FE: Log- Mel spectrograms	FACNN	Device Mismatch Mitigation	Noisy Environments Consideration	Determine the optimum timeframe
2022	[7]	DA: Mix-up FE: Log- Mel spectrograms, GFCC	Recurrent Neural Networks (RNN), SoftMax	Channel Attention Mechanism	Complexity and Computational Cost, Generalization Limitations	channel attention features will be applied with other effective classifiers
2022	[8]	DA: - FE: Log- Mel spectrograms	Transfer Few-Shot Learning	Real-World Relevance, Baseline System	Generalizability to Challenging Acoustic Conditions	Adding more challenging acoustic conditions in the dataset
2020	[9]	DA: - Inference-time FE: Log- Mel spectrograms	CNN, Metric-based few-shot learning	Automation of Sound Event Detection, Enhanced Detection Accuracy	Limited Validation Beyond Speech, Only Fixed context-window	Dynamic context-window

Table 1. Methods used in recent works
DA: Data Augmentation, FE: Feature Extraction

Recent works[Table 1] have used sounds and audio to perform few-shot learning, like to identify a specific sound from the small sample size (approximately 5 examples) [11]. Transfer learning [8], Siamese networks, Metric-based few-shot learning [9] and Triplet networks are a few of the well-liked few-shot learning strategies [Table 2]. Transfer learning is pre-training a model on a sizable dataset for a task that is similar to it, and then honing it on the particular task using a small sample size. Using Siamese networks, a model is trained to distinguish between pairs of examples. The goal of metric-based few-shot learning is to learn an appropriate distance metric, or embedding space, in which dissimilar examples are spaced farther apart and similar examples are closer together. Learning triplets of examples—anchor, positive, and negative—is the main goal of triplet networks. By maximizing the distance between the anchor and negative examples, the model is trained to minimize the distance between the anchor and positive examples. To increase the size of the training dataset and improve the model's generalization, data augmentation techniques like noise addition, spectrogram transformations, and time and pitch stretching can be used.

Table 2. Common used Few-Shot Learning techniques

S. No.	Technique	Used in	Applications
1	Transfer Learning	Sparse Data Availability, Adaptability to New Classes	Audio Classification, Speech Recognition
2	Siamese Networks	Measuring Similarity within a Small Dataset	Audio Similarity Detection, Speaker Verification
3	Metric-based Few-Shot Learning	Discriminating Between Examples in Embedding Space	Audio Classification, Speaker Recognition
4	Triplet Networks	Embedding Triplets for	Audio Retrieval, Speaker Embedding

		Discriminative Learning	

III. Background Study

In recent years, the deep learning techniques for ASC, particularly Convolutional Neural Networks (CNNs) created a very good attention in the audio domain and proven remarkable performance in ASC. However, standard CNN architectures may not be optimized for capturing frequency information effectively, leading to limitations in distinguishing between acoustic scenes with overlapping characteristics. Traditional approaches to ASC often focus on spectral features such as Mel Frequency Cepstral Coefficients (MFCCs) and spectrograms derived from Fast Fourier Transform (FFT). While these methods have shown effectiveness in classifying audio scenes, they may not fully capture the frequency information critical for distinguishing between overlapping classes.

A. Drawback of MFCC & Spectrogram

MFCCs are computed using the discrete cosine transform (DCT) of the logarithm of the magnitude spectrum of short time intervals. This process results in the loss of sequential context, which can be crucial for some tasks like speech recognition or audio classification. Particularly in non-stationary or fluctuating noise environments, MFCCs are highly sensitive to noise. They are very vulnerable to signal noise because they compute directly from the magnitude spectrum, which can seriously distort the feature representation and hinder performance. MFCCs typically partition the frequency spectrum using a fixed number of bins, which leads to a restricted frequency resolution. Fine-grained frequency information may be lost as a result of this, especially in circumstances where the ability to distinguish between different sound classes depends on exact frequency details.

Spectrograms often yield high-dimensional feature representations, especially when computed with large window sizes or high sampling rates. This may result in increased memory requirements and computational complexity, which would make it more challenging to efficiently process and analyze large-scale audio datasets. Spectrograms can be obscured by background noise, making it difficult to separate unwanted interferences from the intended sound. This could lead to mistakes in feature extraction and classification. In spectrograms, the resolution of frequency and time must be traded off; typically, a higher resolution in one domain corresponds to a lower resolution in the other. This fixed trade-off may not always be the optimal choice for capturing relevant features in scenarios where the temporal and spectral characteristics of the audio signals vary.

B. Comparison between Log-Mel Spectrogram vs MFCC

When comparing Log-Mel Spectrograms to Mel-Frequency Cepstral Coefficients (MFCCs), several distinctions arise as, a standard spectrogram is the starting point for log-Mel spectrogram representations. After converting it to the Mel frequency scale, the values are further processed by taking their logarithm. This process gradually captures the frequency content of the signal in a manner more akin to human hearing. MFCCs originate from the short-time Fourier transform of the audio signal. They usually show the spectral envelope and the short-term power spectrum of the sound on a logarithmic frequency scale with a linear amplitude scale. Higher-dimensional feature vectors are often produced by Log-Mel Spectrogram representations because more frequency bins are retained after processing. Every time frame in an audio signal is represented by a vector of spectral features. MFCCs are lower-dimensional than Log-Mel Spectrograms because, following processing, they typically retain 13–40 coefficients. Spectral envelope characteristics.

Logarithmic compression applied to Log-Mel Spectrograms can increase their noise resistance by compressing the dynamic range of frequency components. However, they could still be susceptible to noise disruption. There are several steps involved, including the logarithmic compression, conversion to Mel scale, and spectrogram calculation. These steps add computational complexity, but they are generally feasible for real-time applications. MFCCs are known for their noise resistance because they prioritize the spectral envelope over absolute amplitude. They are therefore particularly useful in noisy environments or for applications where noise robustness is crucial, such as speech recognition. It involves fewer steps than Log-Mel Spectrograms, primarily the Fourier transform, filter bank application, logarithm computation, and DCT. Typically, this results in lower computational overhead.

C. Comparison between Log-Mel Spectrogram vs Spectrogram

When comparing Log-Mel Spectrogram to Spectrogram, several key differences emerge in their utility and effectiveness in audio signal processing tasks. A visual depiction of how the frequency components change over the course of the audio is provided by the spectrogram, which shows the frequency content of a signal over time. But because spectrograms are by their very nature linear, it's possible that they will not adequately represent the non-linear aspects of human hearing. However, Log-Mel Spectrograms overcome this drawback by simulating the nonlinear frequency response of the human auditory system by applying a logarithmic compression to the frequency axis. Through this transformation, the features extracted from the spectrogram become more perceptually relevant and more in line with human auditory perception.

Moreover, by emphasizing the Mel-frequency scale—a perceptually-based frequency scale that more closely approximates how people perceive changes in pitch and frequency—Log-Mel Spectrograms highlight the significant perceptual aspects of the audio signal. As a result, the spectral features are represented in a more condensed manner, focusing on the parts of the audio spectrum that are most important to human perception. As a result, in comparison to conventional spectrograms, log-mel spectrograms typically offer a more discriminative feature representation for audio classification tasks.

D. Comparison between CNN & FACNN

Convolutional Neural Networks (CNNs) and Frequency-Aware CNNs (FACNNs) differ from one another in a few ways. First off, CNNs follow the standard convolutional network architectures, while FACNNs are specifically designed to give

priority to frequency awareness. This is accomplished by keeping a smaller receptive field along the frequency axis and a larger one along the time axis. Compared to CNNs, FACNNs are able to obtain more precise frequency information thanks to this design decision. FACNNs do not downsample along the frequency axis, which allows them to retain frequency details during processing, in contrast to CNNs, which normally downsample along both the time and frequency axes.

CNNs can capture more spatial and temporal patterns because they usually have larger receptive fields in the time and frequency domains. However, FACNNs can capture features that are unique to a given frequency more focusedly because they have larger receptive fields along the time axis and smaller ones along the frequency axis. While FACNNs are optimised for use with Log-Mel Spectrogram, which makes better use of frequency information, CNNs are typically used with Spectrogram and MFCC when it comes to feature extraction.

Because of their versatility, CNNs are used for many different tasks, including the classification of images, audio, and video, due to their versatility. Conversely, FACNNs perform particularly well in audio tasks where frequency information is crucial. Because of their innate sensitivity to subtle differences in frequency, they perform better on tasks requiring them to distinguish between similar sounds. For this reason, while CNNs remain a powerful choice for many applications, FACNNs offer unique advantages in tasks where fine-grained frequency details are essential..

E. Accuracy

Accuracy in the context of audio classification refers to the proportion of correctly classified audio samples out of the total number of samples in the dataset. It is typically computed using the following formula:

Accuracy= (Total number of samples) / (Number of correctly classified samples) x 100%

Where, Number of correctly classified samples is the count of audio samples that are correctly classified by the classifier. For each sample, the classifier predicts a label, and if the predicted label matches the ground truth label, then it is considered a correct classification. Total number of samples is the total count of all audio samples in the dataset, including both the training and testing sets.

IV. Proposed Methodology

4.1 Overview:

Fig. 1 shows the architecture of proposed model. In order to reduce the amount of audio files and create a new dataset of similar audios based on similarity, the raw audio files were submitted to a few shot learning method as a metric-based approach. Next, the recently generated dataset will be used as the Log-mel spectrogram input for the feature extraction technique. Frequency-aware convolutional neural networks (FACNNs) will receive the extracted features and use them to classify the scenes.

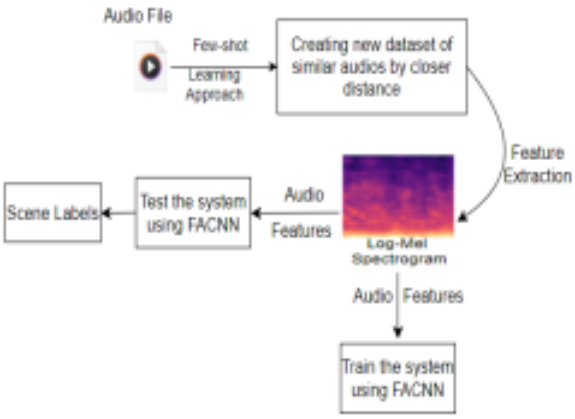


Fig.1. Architecture of Proposed Model

4.2 Metric-based Approach:

The metric-based approach is employed to diminish the volume of audio files with similar content, utilizing the closest distance metric between audio files as a criterion. The procedure unfolds in distinct steps, commencing with the input being the waveform. The first checkpoint involves verifying the sampling rate of the selected pair of audio files. If the rates match, the Euclidean distance between the raw audio waveforms of the two files is calculated. In cases where the rates differ, the comparison cannot proceed. The calculation of the Euclidean distance between the audio files is executed to determine their similarity can be expressed by

$$D(audio\_1, audio\_2) = \sqrt{(audio\_1[i] - audio\_2[i])^2} \quad (1)$$

Subsequently, the new dataset is updated, incorporating files within a closer distance range, thereby refining the set of audio files based on their metric proximity.

Algorithm: Acoustic Scene Classification.

Input: Raw audio files (X_raw)
Output: Scene Label
Begin

```
MBFSL(X_raw)
    Sampling_rate <-- 1/T
    D(x_i, x_j) <-- sqrt(sum((x_i(k) - x_j(k))^2 for k in
    range(1, len(x_i))))
    New_dataset(X_new) <-- threshold distance (δ)<0.5

Feature_extraction(X_new)
    For each audio sample x_i in X_new:
    Extract the log-mel spectrogram features (F_i).

Classifier(F_i)
    Initialize FACNN model parameters: B_res, F_conv,
    K_conv, α (learning rate)
```

```
Construct FACNN architecture:
    Initialize model with B_res residual blocks:
    Each residual block:
        H_{l+1} = ReLU(Conv_{l+1}(H_l))
        H_{l+2} = ReLU(Conv_{l+2}(H_{l+1}))
        H_{l+3} = H_{l} + H_{l+2}
        Global_average_pooling followed by softmax
        Layer

Train FACNN_model(F_i)
    Iterate over epochs:
    For each batch:
        Forward pass:
            Y_pred = Softmax(FACNN(X_batch))
            Loss = Cross_entropy(Y_true, Y_pred)
        Backward pass:
            Compute gradients
            Update model parameters

Test FACNN_model(Y_raw)
    For each audio sample y_i in Y:
    Extract the log-mel spectrogram features (F'_i).
    Use the trained FACNN_model to predict the scene
    label for y_i based on F'_i.

End
```

4.3 Feature Extraction:

Prior to network training, the input audios are pre-processed into Log-Mel spectrograms, and Fig. 2 illustrates this procedure. These spectrograms can be extracted using a Mel frequency bank, and then scaled using a logarithm. Spectrograms the size of 44 Mel frequency bins were used in this study. The sample frequency of the audio file is 44100 Hz. The window lengths have frame durations of 80 ms and 40 ms, with a 50% overlap length. Eq. (2) contains the computation for the number of frame lengths.

$$f = 1 + floor\ [(L\text{-}W) / S] \quad (2)$$

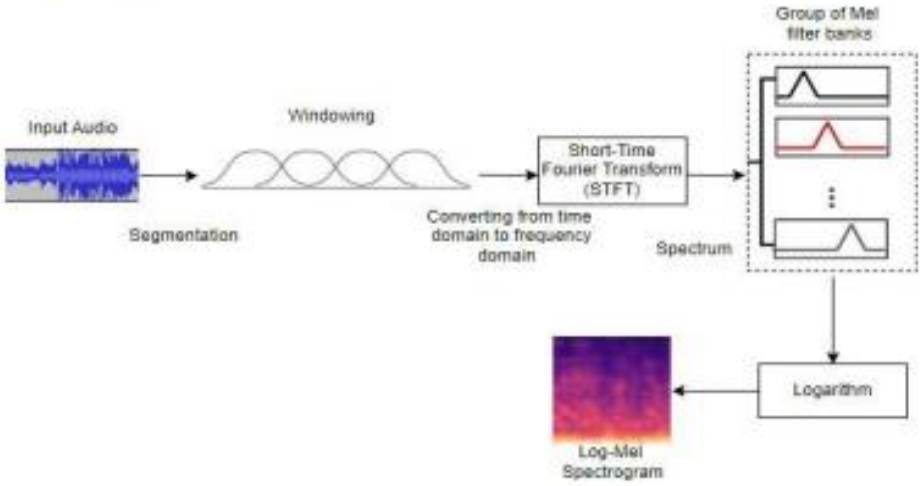


Fig. 2. Block diagram of Log- Mel spectrogram of an Audio signal

Where f is the number of frames, L is signal length, W is window length and S is shift length. To get a smooth spectral representation from the frames, a Hamming window is applied to a frame to reduce the impact of noise at the edges of the window. The following equation Eq. (3) defines the Hamming window $h(n)$:

$$h(n) = 0.54 - 0.46 \cos (2\pi n / (N-1)) \ 1 \leq n \leq N \quad (3)$$

Where N is frame number, and $h(n)$ is the hamming window. This function is used to estimate the Fast Fourier Transform (FFT- Eq. (4) for getting an amplitude-frequency response per frame.

$$S_i(k) = \sum_{i=1}^{N} s_i(k) h(n) e^{-\frac{j2\pi kn}{N}} \ n = 0, .., N-1 \quad (4)$$

Where (f) is the frequency of mel and f is frequency of linear frequency. Mel filter is used to obtain the energy from the filter bank.

4.4 Frequency-aware convolutional neural network (FACNN):

A novel method for prioritizing frequency components by avoiding downsampling along the frequency axis is shown by the FACNN architecture shown in Fig. 3. In order to ensure large receptive fields along the temporal axis, this design choice prioritizes the preservation of fine-grained frequency information. Through the maintenance of smaller receptive fields in the frequency domain and larger ones in the time domain, the design successfully combines local frequency information from the input Time-Frequency representation with global temporal context. By combining global and local inputs, deeper layers are able to identify and extract significant characteristics, which improves the model's capacity for sophisticated analysis and synthesis in a range of research applications.

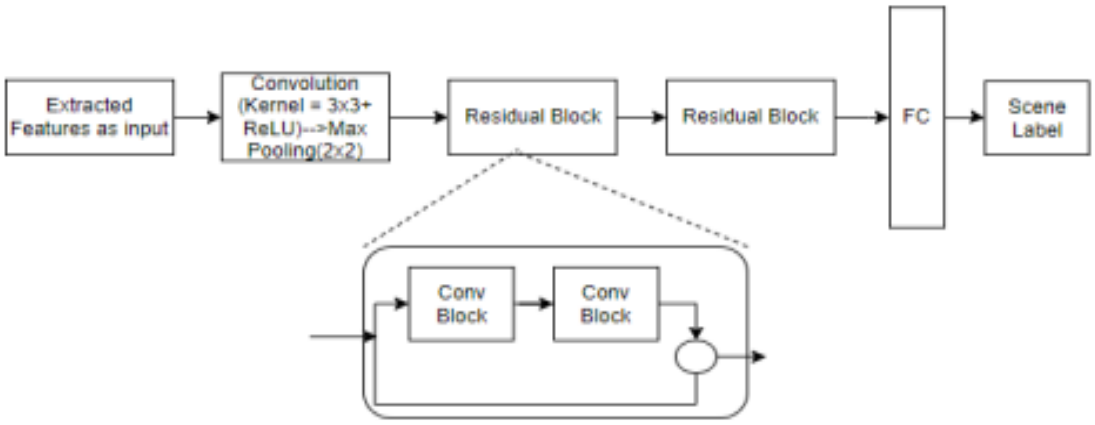


Fig. 3. Detailed architecture of FACNN

Inspired by the work of Gao and McDonnell [18], we have designed a proposed architecture that significantly deviates from traditional residual networks. Notably, we have deviated from conventional residual networks, which usually downsample on both the time and frequency axes, by choosing not to do downsampling on the frequency axis. Further, we employ a pre-activation structure in our convolution block. Using batch normalization, ReLU activation, and convolution are the steps involved in this process. The bias term in the FACNN convolution stage is omitted. In addition, as Fig. 4 shows, down-sampling in the identity path in our residual blocks is carried out by average pooling and channel padding. In conclusion, we substitute 1 x 1 convolutional layers for the traditional fully linked layers at the network's end. These architectural modifications are designed to enhance the model's ability to capture and leverage intricate temporal and frequency features, contributing to its efficacy in research applications.

The essential components of the FACNN architectureare the residual blocks, which are made up of two 3 x 3convolutional layers apiece. Using pre-activation convolution as the basis, the first layer concentrates on batch normalization. Feature maps are expanded to 32 by the first convolutional layerthrough downsampling along the time axis. Then, sixteen residual blocks are added, which are arranged into four residual levels. Each layer's first residual block undergoes downsampling, which progressively increases the feature maps from 32 to 64, 128 to 256. A dual-layer, 1 x 1 convolutional procedure takes place after residual layers (post-residual layer). The first convolutional layer increases feature maps from 256 to 512, while the second one decreases them to 3, in accordance with the number of classes in the dataset. Together, these architectural choices improve the model's ability to identify intricate details and promote reliable results in a variety of study situations.

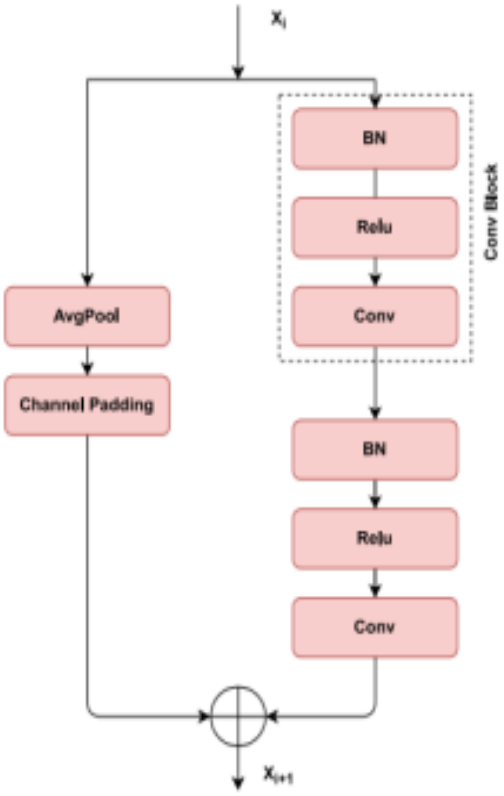
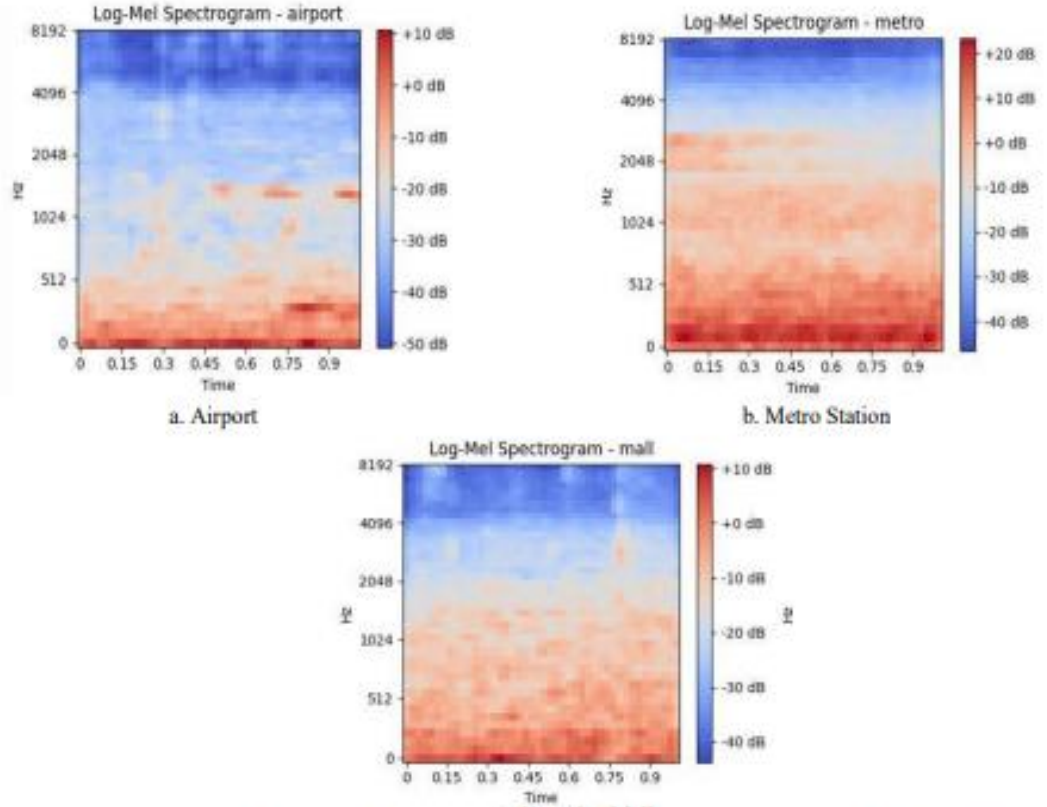


Fig. 4. Residual block with pre-activation. The BN and ReLUis applied before convolution in the conv block.



a. Airport

b. Metro Station

c. Shopping Mall

Fig. 5. Log- Mel spectrograms representing the input to the FACNN of different scenes

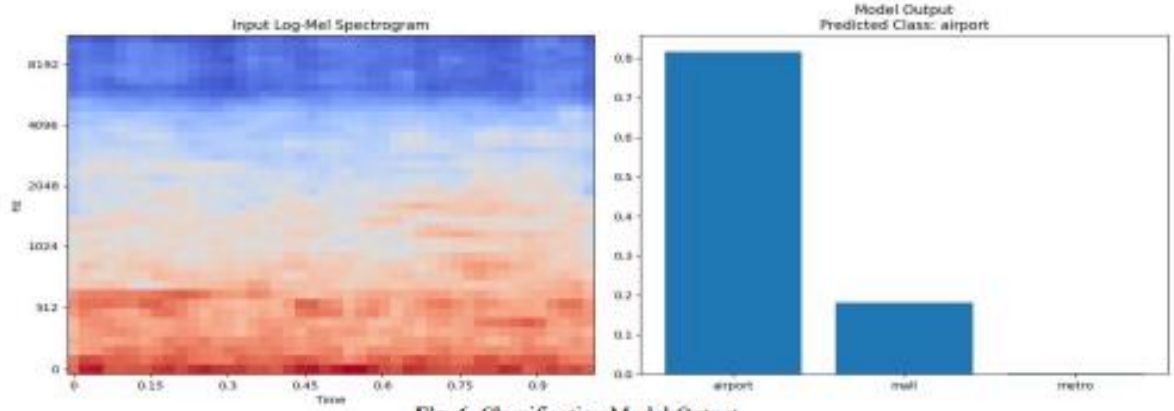


Fig. 6. Classification Model Output

V. Results and Discussion

A. Approach Overview

The anticipated results of our proposed methodology are grounded in the synergistic integration of Metric-Based Few-Shot Learning Approach, Log-Mel Spectrogram feature extraction, and the Frequency-Aware CNN classification. Through the application of Metric-Based Few-Shot Learning, we expect to observe a substantial reduction in the number of audio samples required for effective model training. This reduction is attributed to the emphasis on close-range similarity, enabling the model to generalize efficiently with a limited number of labeled examples. The Log-Mel Spectrogram feature extraction is anticipated to provide a rich and informative representation of audio samples, capturing both temporal and frequency characteristics. We expect this representation to enhance the discriminative power of the model, enabling it to distinguish between overlapping classes in diverse acoustic scenes. Furthermore, the Frequency-Aware CNN, designed to be sensitive to frequency information.

B. Dataset

The development dataset for TAU Urban Acoustic Scenes 2022 Mobile was utilized to create the suggested methodology. The dataset includes recordings made with four different devices in three distinct acoustic scenes across twelve European cities. The three acoustic scenes in the development dataset are an airport, a metro station, and a shopping center. The dataset contains 45, 336 audio recordings that are available in a single-channel 44.1 kHz 24-bit format.

C. Parameterization of Log-Mel Spectrogram and FACNN

The following are the parameters that the Log-Mel Spectrogram and Frequency-Aware CNN use: utilise a set of parameters to specify the spectrogram generation process, such as a sampling frequency of 44.1 kHz, an 80 ms window length with a 50% frame overlap, and a frame duration of 40 ms. The number of Mel frequency bins should be set to 44. The FFT size is dynamically determined based on the frame length and window function, ensuring effective spectral analysis, and the use of a Hamming window function improves frequency resolution and decreases spectral leakage. The FACNN is composed of sixteen residual blocks, each of which uses three-by-three convolutional layers. To obtain crucial frequency information, the first layer creates 32 feature maps, and later layers use average pooling for down sampling.

D. Visualization of Log-Mel Spectrogram

Fig. 5 displays the resultant log-mel spectrogram, which employs colour to denote intensity, y-axis to indicate frequency, and x-axis to denote time. Here, the sounds of an airport, a metro station, and a mall are displayed over time by the spectrogram. The colours of the spectrogram are blue to red, where blue denotes quieter sounds and red, louder sounds. On the right side of the spectrogram, there is a scale that shows the sound's decibel level as well. The time is shown in seconds on the x-axis of the spectrogram. The frequency is shown on the y-axis of the spectrogram in Hertz (Hz). Taking into account that the noise from the airport fluctuates over time, becoming louder during certain moments and loudest at low frequencies, most likely because of the sound of aeroplane engines [Fig. 6].

Table 3. Accuracy Comparison for different systems.

Features	System	Accuracy (%)
Spectrogram	CNN	74.95
MFCC	CNN	75.7
Log-Mel	CNN	77.45
Log-Mel	**FACNN**	**81.97**

E. System Comparison

The proposed FACNN classifier is specifically designed to be sensitive to frequency information in audio signals. As opposed to conventional CNNs that down-sample along both the time and frequency axes, FACNN maintains a small receptive field along the frequency axis, allowing it to capture more accurate frequency information as shown in Table 3. This architecture is useful for tasks where the accuracy of the classification depends on frequency details. Compared to MFCC and standard Spectrogram features, Log-Mel Spectrogram offers a more informative representation of audio signals in audio classification tasks. By taking the logarithm of the Mel spectrogram and compressing the audio signal's dynamic range, the Log-Mel Spectrogram improves the features' ability to discriminate. This could lead to improved performance in scenarios with overlapping or complicated audio classes as shown in Fig.7.

F. Anticipated Outcomes and Analysis

We expect encouraging classification accuracy across a range of acoustic scenes in the experimental evaluation conducted with the TAU Urban Acoustic Scenes 2022 Mobile dataset. It is anticipated that the metric-based approach will successfully refine the dataset, producing a more specialized and reliable model. It is anticipated that the visual representation provided by the Log-Mel Spectrogram will improve the interpretability of the model and augment its capacity to capture subtle audio features. It is expected that the Frequency-Aware CNN will perform exceptionally well in obtaining pertinent frequency information due to its distinct architecture, which will allow the model to dynamically adjust to various auditory environments. All things considered, we anticipate that our suggested methodology will show improvements in Acoustic Scene Classification and offer a scalable and effective solution for situations in which acquiring large-scale labelled datasets is difficult.

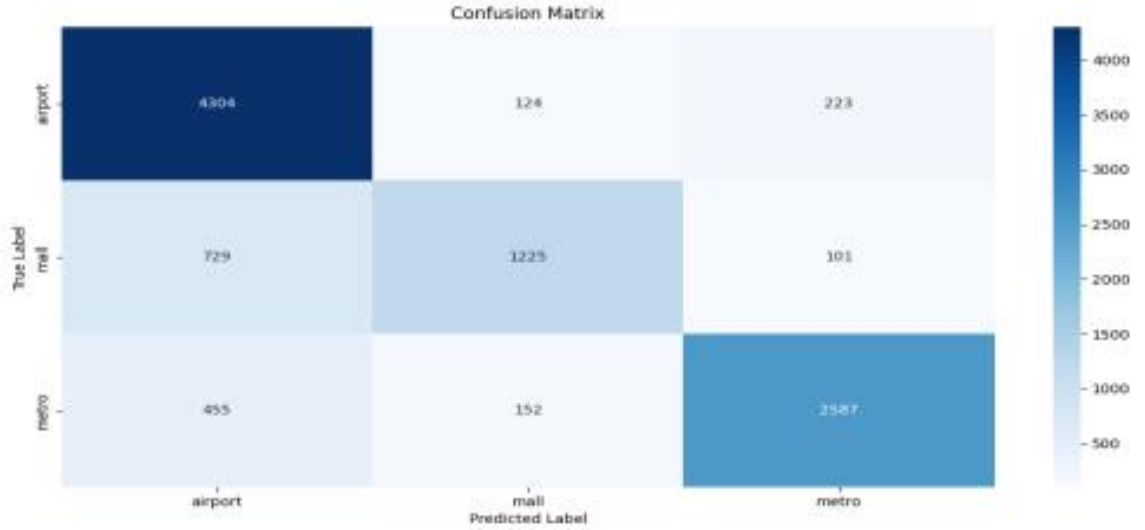


Fig. 7. Confusion Matrix

VI. Conclusion

In this work, the field of Acoustic Scene Classification (ASC) significantly augmented by introducing a comprehensive methodology designed to tackle the challenges of overlapping classes and limited labeled examples. By integrating Metric-Based Few-Shot Learning, Log-Mel Spectrogram feature extraction, and Frequency-Aware CNN classification, our approach offers a scalable and efficient solution tailored to scenarios where obtaining extensive labeled datasets is challenging. The experimental evaluation conducted on the TAU Urban Acoustic Scenes 2022 Mobile dataset underscores the effectiveness of our methodology. This dataset comprises recordings captured across twelve European cities, encompassing diverse acoustic environments such as airports, metro stations, and shopping centers. Notably, our methodology's robustness and adaptability are highlighted by its promising classification accuracy across a variety of auditory scenes. In addition, the model's handling of overlapping classes is especially impressive, demonstrating its capacity to discriminate between intricate auditory environments and minute acoustic details. In order to advance ASC technology, more research into our model's scalability across larger and more diverse datasets and its applicability to real-world scenarios is imperative. Additionally, efforts to enhance the interpretability and robustness of the model, particularly in complex acoustic environments, will be paramount for its practical implementation in various domains, including smart homes, audio surveillance, and context-aware mobile devices.

VI. References

[1] Mahmoud A. Alamir, "A novel acoustic scene classification model using the late fusion of convolutional neural networks and different ensemble classifiers", Applied Acoustics 175 (2021) 107829A. Shukla et al., "Legal Case Document Summarization: Extractive and Abstractive Methods and their Evaluation," arXiv (Cornell University), Oct. 2022, doi: 10.48550/arxiv.2210.07544.

[2] Tao Zhang, Jinhua Liang, Biyun Ding, "Acoustic scene classification using deep CNN with fine-resolution feature", Expert Systems With Applications 143 (2020) 113067

[3] Nisan Aryal a, Sang-Woong Lee, "Frequency-based CNN and attention module for acoustic scene classification", Applied Acoustics 210 (2023) 109411

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

[5] Paseddula C, Gangashetty SV. Late fusion framework for Acoustic Scene Classification using LPCC, SCMC, and log-Mel band energies with Deep Neural Networks. Appl Acoust 2021;172:107568. https://doi.org/10.1016/j.apacoust.2020.107568

[6] Kosmider M. Spectrum correction: Acoustic scene classification with mismatched recording devices. Proc. Interspeech 2020 2020:4641–5.

[7] Mie Mie Oo, Nu War, "Acoustic Scene Classification using Attention based Deep Learning Model", International Journal of Intelligent Engineering and Systems, Vol.15, No.6, 2022

[8] Javier Naranjo-Alcazar, Sergi Perez-Castanos, Pedro Zuccarello, Ana M. Torres, Jose J. Lopez, Francesc J. Ferri, Maximo Cobos, "An Open-Set Recognition and Few-Shot Learning Dataset for Audio Event Classification in Domestic Environments", Pattern Recognition Letters 164 (2022) 40–45

[9] Wang, Y., Salamon, J., Bryan, N. J., & Pablo Bello, J. (2020). Few-Shot Sound Event Detection. ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). doi:10.1109/icassp40776.2020.9054708

[10] Biyun Ding, Tao Zhang, Ganjun Liu, Lingguo Kong, Yanzhang Geng, "Late fusion for acoustic scene classification using swarm intelligence", Applied Acoustics 192 (2022) 108698

[11] Ines Nolasco, Shubhr Singh, Veronica Morfi, Vincent Lostanlen, Ariana StrandburgPeshkin, Ester Vidana-Vila, Lisa Gill, Hanna Pamuła, Helen Whitehead, Ivan Kiskin, Frants H. Jensen, Joe Morford, Michael G. Emmerson, Elisabetta Versace, Emily Grout, Haobe Liu, Burooj Ghani, Dan Stowell, "Learning to detect an animal sound from five examples", Ecological Informatics 77 (2023) 102258

[12] Shefali Waldekar, Goutam Saha, "Two-level fusion-based acoustic scene classification", Applied Acoustics 170 (2020) 107502

[13] Chris Careaga, Brian Hutchinson, Nathan Hodas, Lawrence Phillips, "Metric-Based Few-Shot Learning for Video Action Recognition", Computer Vision and Pattern Recognition arXiv:1909.09602v1

[14] Alexander Rakowski, Michał Kosmider, "FREQUENCY-AWARE CNN FOR OPEN SET ACOUSTIC SCENE CLASSIFICATION", Audio Intelligence, Detection and Classification of Acoustic Scenes and Events 2019

[15] A. Mesaros, T. Heittola, and T. Virtanen, "A multi-device dataset for urban acoustic scene classification," in Proceedings of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE2018), November 2018, pp. 9–13. [Online]. Available: https://arxiv.org/abs/1807.09840

[16] A. Mesaros, T. Heittola, A. Diment, B. Elizalde, A. Shah, E. Vincent, B. Raj, and T. Virtanen, "Dcase 2017 challenge setup: Tasks, datasets and baseline system," in DCASE 2017- Workshop on Detection and Classification of Acoustic Scenes and Events, 2017.

[17] Yerin Lee, Soyoung Lim, Il-Youp Kwak, "CNN-Based Acoustic Scene Classification System", Electronics 2021, 10(4), 371

[18] Gao W, McDonnell M. Acoustic scene classification using deep residual networks with late fusion of separated high and low frequency paths, Tech. rep., DCASE2019 Challenge (June 2019).

[19] Liping Yang, Lianjie Tao, Xinxing Chen, Xiaohua Gu, "Multi-scale semantic feature fusion and data augmentation for acoustic scene classification", Applied Acoustics 163 (2020) 107238.

[20] Lam Pham, Huy Phan, Truc Nguyen, Ramaswamy Palaniappan, Alfred Mertins, Ian McLoughlin, "Robust acoustic scene classification using a multi-spectrogram encoder-decoder framework", Digital Signal Processing 110 (2021) 102943

[21] Yan Leng, Weiwei Zhao, Chan Lin, Chengli Sun, Rongyan Wang, Qi Yuan, Dengwang L, "LDA-based data augmentation algorithm for acoustic scene classification", Knowledge-Based Systems 195 (2020) 105600

[22] Sumit Kumar Chaudhary, Sameer Saran, "Information Network (IBIN) framework", Environmental Sustainability, doi.org/10.1007/s42398-023-00281-w

[23] Sayeh Mirzaei, Iman Khani Jazani, " Acoustic scene classification with multi-temporal complex modulation spectrogram features and a convolutional LSTM network", Multimedia Tools and Applications (2023) 82:16395–16408

[24] Tao Zhang, Jinhua Liang, Guoqing Feng, " Adaptive time-frequency feature resolution network for acoustic scene classification", Applied Acoustics 195 (2022) 108819

[25] Yu Wang, Nicholas J. Bryan, Mark Cartwright, Juan Pablo Bello, Justin Salamon, "FEW-SHOT CONTINUAL LEARNING FOR AUDIO CLASSIFICATION", Speech and Signal Processing (ICASSP), 10.1109/ICASSP39728.2021.9413584

[26] Yan Gao, Haijiang Li, Weiqi Fu, "Few-shot learning for image-based bridge damage detection", Engineering Applications of Artificial Intelligence 126 (2023) 107078.

[27] Farong Gao, Lijie Cai, Zhangyi Yang, Shiji Song, Cheng Wu, "Multi-distance metric network for few-shot learning", International Journal of Machine Learning and Cybernetics (2022) 13:2495–2506

[28] Wei Xie, Yanxiong Li, Qianhua He , Wenchang Cao, "Few-shot class-incremental audio classification via discriminative prototype learning", Expert Systems With Applications 225 (2023) 120044

[29] Chandrasekhar Paseddula, Suryakanth V. Gangashetty, "Late fusion framework for Acoustic Scene Classification using LPCC, SCMC, and log-Mel band energies with Deep Neural Networks", Applied Acoustics 172 (2021) 107568

[30] Yuzhong Wu, Tan Lee, "ENHANCING SOUND TEXTURE IN CNN-BASED ACOUSTIC SCENE CLASSIFICATION", Speech and Signal Processing (ICASSP), 10.1109/icassp.2019.8683490

[31] Zhao Ren, Kun Qian, Zixing Zhang, Vedhas Pandit, Alice Baird," Deep Scalogram Representations for Acoustic Scene Classification", AUTOMATICA SINICA, VOL. 5, NO. 3(2018)

[32] Gao W, McDonnell M. Acoustic scene classification using deep residual networks with focal loss and mild domain adaptation, Tech. rep., DCASE2020 Challenge (June 2020).

[33] Kong Q, Cao Y, Iqbal T, Wang W, Plumbley MD. Cross-task learning for audio tagging, sound event detection and spatial localization: DCASE 2019 baseline systems, Tech. rep., DCASE2019 Challenge (June 2019).

[34] Z. Ren, V. Pandit, K. Qian, Z. J. Yang, Z. X. Zhang, and B. Schuller, "Deep sequential image features for acoustic scene classification," in Proc. Detection and Classification of Acoustic Scenes and Events, Munich, Germany, 2017, pp. 113–117.

[35] M. Valenti, A. Diment, G. Parascandolo, S. Squartini, and T. Virtanen, "DCASE 2016 acoustic scene classification using convolutional neural networks," in Proc. of Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE), Budapest, Hungary, 2016.

[36] A. Mesaros, T. Heittola, and T. Virtanen, "TUT database for acoustic scene classification and sound event detection," Proc. of the 24th Acoustic Scene Classification Workshop 2016 European Signal Processing Conference (EUSIPCO), 2016.

Applied Acoustics 175 (2021) 107829

Contents lists available at ScienceDirect

Applied Acoustics

journal homepage: www.elsevier.com/locate/apacoust

A novel acoustic scene classification model using the late fusion of convolutional neural networks and different ensemble classifiers

Mahmoud A. Alamir

College of Science and Engineering, Flinders University, Clovelly Park, Adelaide, SA 5042, Australia

ARTICLE INFO

Article history:
Received 20 August 2020
Received in revised form 1 October 2020
Accepted 25 November 2020
Available online 16 December 2020

Keywords:
Artificial intelligence
Noise management
Environmental noise classification
Acoustic scene classification
Late fusion

ABSTRACT

Recent evidence suggests that convolutional neural networks (CNNs) can model acoustic scene classification (ASC) with high accuracy. Ensemble classifiers have also shown high accuracy in different machine learning areas. However, little is known about fusion models between CNNs and different ensemble classifiers for ASC. This study presents an enhanced CNN classification model using the late fusion between CNNs and ensemble classifiers to predict different classes of acoustic scenes. A CNN model was first built to classify fifteen acoustic scene environments. Different ensemble classifier models were then used for this classification problem. Late fusion of CNN and ensemble classifier models was then applied. The results showed that late fusion models have higher classification accuracy, as compared to individual CNN or ensemble classifier models. The best model was obtained by fusion of the CNN and discriminant random subspace classifier with an increase in the average accuracy of 10% as compared to the average accuracy of the CNN model. When compared with previous research on ASC, the late fusion model between CNN and ensemble classifiers showed higher accuracy. Therefore, this method has robust applicability for future ASC problems.

1. Introduction

1.1. Background

Acoustic scene classification (ASC) is the way to classify different environments depending on their sound characteristics. The scene, in this context, refers to the acoustic environment summarised in one situation such as "restaurant" or "office". Acoustic scenes could be pre-recorded or live streaming audio [1]. ASC plays an important role in many areas, such as context awareness in smart devices, hearing aids, robots, and many other applications [2]. However, there is a need for high-performance ASC models [3]. Therefore, many algorithms and methods have been developed to achieve accurate ASC models.

1.2. Previous methods for ASC

There have been many methods for ASC, mostly using CNNs [2,4]. However, early fusion CNN models showed high accuracy for ASC. Fusion means combining one or more characters at the same time. In terms of modelling, fusion can be classified into early and late fusion. Early fusion could refer to combining the features using more than one method or other concepts such as refining frequency resolution before beginning model training. Recently, early fusion models have extensively been used for ASC. For example, Yang et al. [5] used multistage feature extraction fusion for ASC. Su et al. [6] also used aggregated feature extraction for ASC. Zhang et al. [7] used fine-resolution frequency for feature selection of ASC. Mulimani et al. [8] also used fisher vector for feature extraction of ASC.

1.3. Previous results of late fusion models

Late fusion refers to combining the results of different models after building each model separately [9]. Recently, late fusion models were used in many areas because of their higher predictability as compared to individual models. For example, they have shown higher predictability than early fusion models when used for semantic video analysis [10]. Late fusion can be achieved by combining CNN model with other models such as SVM or different CNN models with different feature extraction methods [11]. Recently, it was also used for emotion recognition for audio-visual data [12,13]. They were also used for recognising human activity [14]. However, the use of late fusion models for ASC has not been applied before between CNN and different ensemble classifier models for ASC problems.

E-mail address: mahmoud.alamir@flinders.edu.au

https://doi.org/10.1016/j.apacoust.2020.107829
0003-682X/

1.4. Study aims

Most studies optimising ASC models are based on the early fusion of feature characteristics before using them in CNN models. It is hypothesised that late fusion of different models could yield higher predictive power, as compared to when using only one model. Therefore, this study proposes a late fusion model between CNN and ensemble classifier models. Different ensemble classifiers are studied and their accuracy, when fused with CNN, is also presented. The results help to improve ASC predicted accuracies.

2. Acoustic scene data

2.1. Data source

The dataset of TUT Acoustic scenes 2017 challenge was used [15]. A description of acoustic scenes included in the dataset can be found through http://www.cs.tut.fi/sgn/arg/dcase2016/acoustic-scenes#library. The dataset consists of various acoustic scenes recorded from distinct locations. Each acoustic scene has 312 segments for training noise samples and 108 for testing noise samples. For each original recording location, a 3–5-minute-long audio recording was captured.

2.2. Acoustic scene types

Acoustic scenes were classified as follows: Bus - travelling by bus in the city (vehicle), Cafe / Restaurant - small cafe/restaurant (indoor), Car - driving or travelling as a passenger, in the city (vehicle), City centre (outdoor), Forest path (outdoor), Grocery store - medium size grocery store (indoor), Home (indoor), Lakeside beach (outdoor), Library (indoor), Metro station (indoor), Office - multiple persons, typical workday (indoor), Residential area (outdoor), Train (travelling, vehicle), Tram (travelling, vehicle), Urban park (outdoor).

2.3. Recording acoustic scenes

All included acoustic scenes were recorded in different locations (i.e. different parks, streets, homes). The sound was recorded using "Soundman OKM II Klassik/studio A3", electret binaural microphone and a "Roland Edirol R-09" wave recorder with a sample rate of 44.1 kHz and a resolution of 24-bit. The microphones were designed to be similar to headphones when used in-ears [16,17]. This allowed recorded sounds to be similar to the sound in the auditory system [18,19].

2.4. Post-processing audio file dataset

The recorded data was post-processed to make for the privacy of people and any possible errors in recordings. For data recorded in private places, consent forms were filled from all people occupying these places. Consent forms were not required for data recorded in public places [20–22]. However, the content with privacy segments was removed. Any data with failure either because of microphones and/or signal distortions were also removed. After cleaning data, remaining data files were segmented into 10-second data files. Table 1 summarises the included 10-second acoustic scene segments.

3. The proposed method

3.1. Overview

Fig. 1 shows the proposed late fusion model procedures. Data was first entered and was then split into 10-sec segments. Feature extraction was then done by applying Mel-spectrograms to convolutional neural networks (CNNs) and wavelet scattering for ensemble classifiers. Hyper-parameter tuning was then done for CNN and ensemble classifier models separately. The fusion of CNN and ensemble classifier models was then applied to maximise the accuracy obtained. Each model (i.e. the CNN, ensemble classifier and fusion models) was then used for class prediction and its accuracy was evaluated. Different ensemble classifiers were used. Therefore, these procedures were repeated with each ensemble classifier. All these procedures and functions to execute them were done through MATLAB 2020a.

3.2. Convolutional neural networks

3.2.1. Data augmentation

The DCASE 2017 dataset contained a relatively small number of acoustic recordings, and the development set and evaluation set were recorded at different specific locations. As a result, it is easy to overfit to the data during training. To fix this problem, data was augmented in two ways.

The first augmentation method was done by splitting each 10-second sample into 10 one-second samples. The split procedure made CNN easier to train and avoid overfitting to any acoustic event. It also ensured the relative combinations of the training data events. Data was also augmented to increase the training stage accuracy.

The second augmentation method was by implementing mix-ups. In a mix-up, the dataset is augmented by mixing features of two different classes. When features are mixed, labels are mixed in equal proportion [23]. When training proposed CNN model, labels were drawn from probability distribution instead of mixed labels. Each spectrogram was mixed with a spectrogram of a different label with lambda set to 0.5. Original and mixed datasets are combined for training. These procedures can be described in Eqs. (1) and (2), where $\bar{x}$ is spectrogram features extracted from a random available spectrogram (x_i) added to the augmented sample (x_i) and $\bar{y}$is the label of randomly set by lambda.

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

$$\bar{y} = \lambda y_i + (1-\lambda)y_j \tag{2}$$

3.2.2. The architecture of the proposed convolutional neural network

The convolutional neural network (CNN) architecture was based on the design from [2]. However, the number of learning

Table 1
A summary of the number of training and testing noise samples included in this study.

Type of acoustic scene	Beach	Bus	Cafe/ restaurant	Car	City centre	Forest path	Grocery store	Home	Library	Metro station	office	park	Residential area	Train	Tram
No. of noise samples [Training]	312	312	312	312	312	312	312	312	312	312	312	312	312	312	312
No. of noise samples [Testing]	108	108	108	108	108	108	108	108	108	108	108	108	108	108	108

cycles and training processes were modified as explained in the following sections.

3.2.3. Feature extraction

Mel Spectrograms were used to transform the audio files into frequency-domain representation. The characteristics of these spectrograms were as follows: window length = 2048; samples per-hop = 1024; samples overlap = 1024; FFT Length = 4096; number of bands = 128. Fig. 2 shows a typical example of the extracted spectrograms for each 10-sec sample after augmentation by splitting it into 10 one-seconds audio files.

3.2.4. Convolutional neural network training

The Bayesian optimisation [24] was used to obtain hyperparameters, which were as follows: minimum batch size = 128; momentum = 0.9, L2 Regularization = 0.005, maximum epochs = 15, learn rate schedule = piecewise.

To speed up processing, Mel spectrograms for all noise files in the datastores were extracted using tall arrays. Tall arrays remain uncalculated until calculations are requested using the gather function. These delayed calculations help to work with large data quickly. When requested, MATLAB combines these queued calculations to take the minimum passes through the data.

3.2.5. Convolutional neural network accuracy evaluation

A probability-weighted averaged on the one-second segments was used to predict the scene for each 10-second noise sample in the testing dataset. For each 10-second noise sample, the maximum relative weight prediction was used and labelled to the corresponding predicted acoustic scene.

3.3. Ensemble classifiers

Wavelet scattering has been shown in [25] to provide a good representation of acoustic scenes. Therefore, it was used for extracting the features for the ensemble classifier training. The invariance scale and quality factors were determined through trial and error.

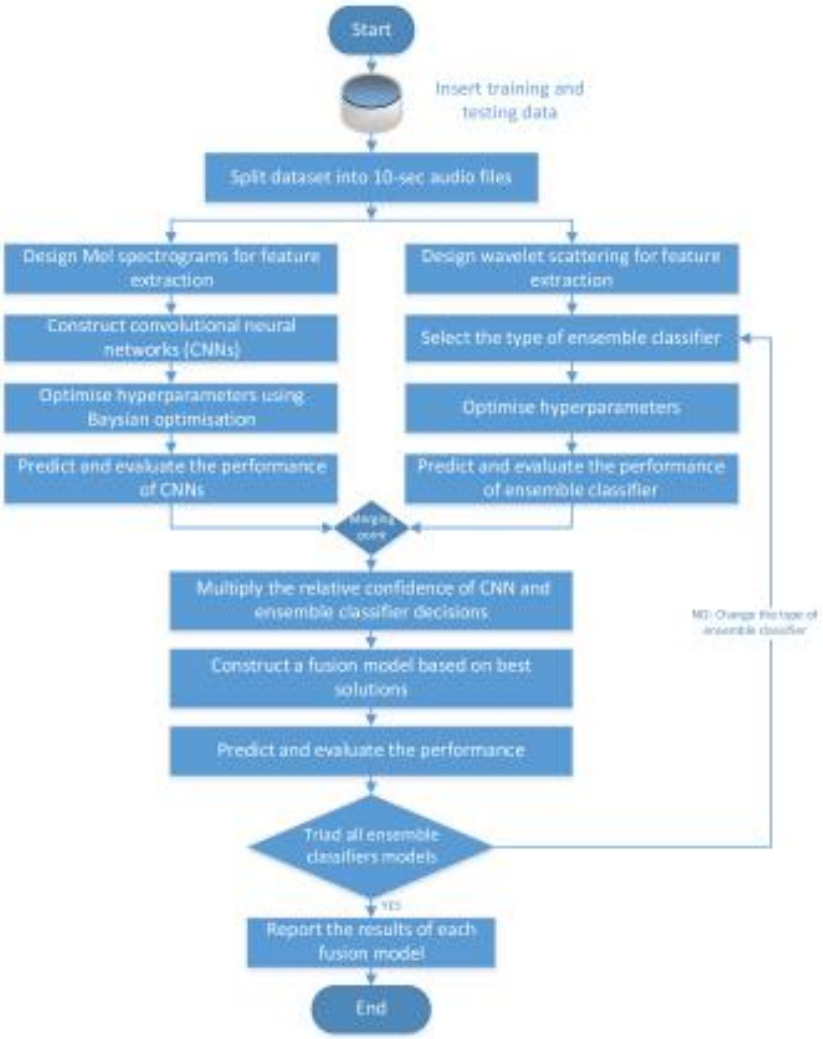


Fig. 1. The proposed late fusion model between CNN and ensemble classifier models. The dataset for each model is augmented firstly and each model is then optimised before the late fusion occurs in the last step.

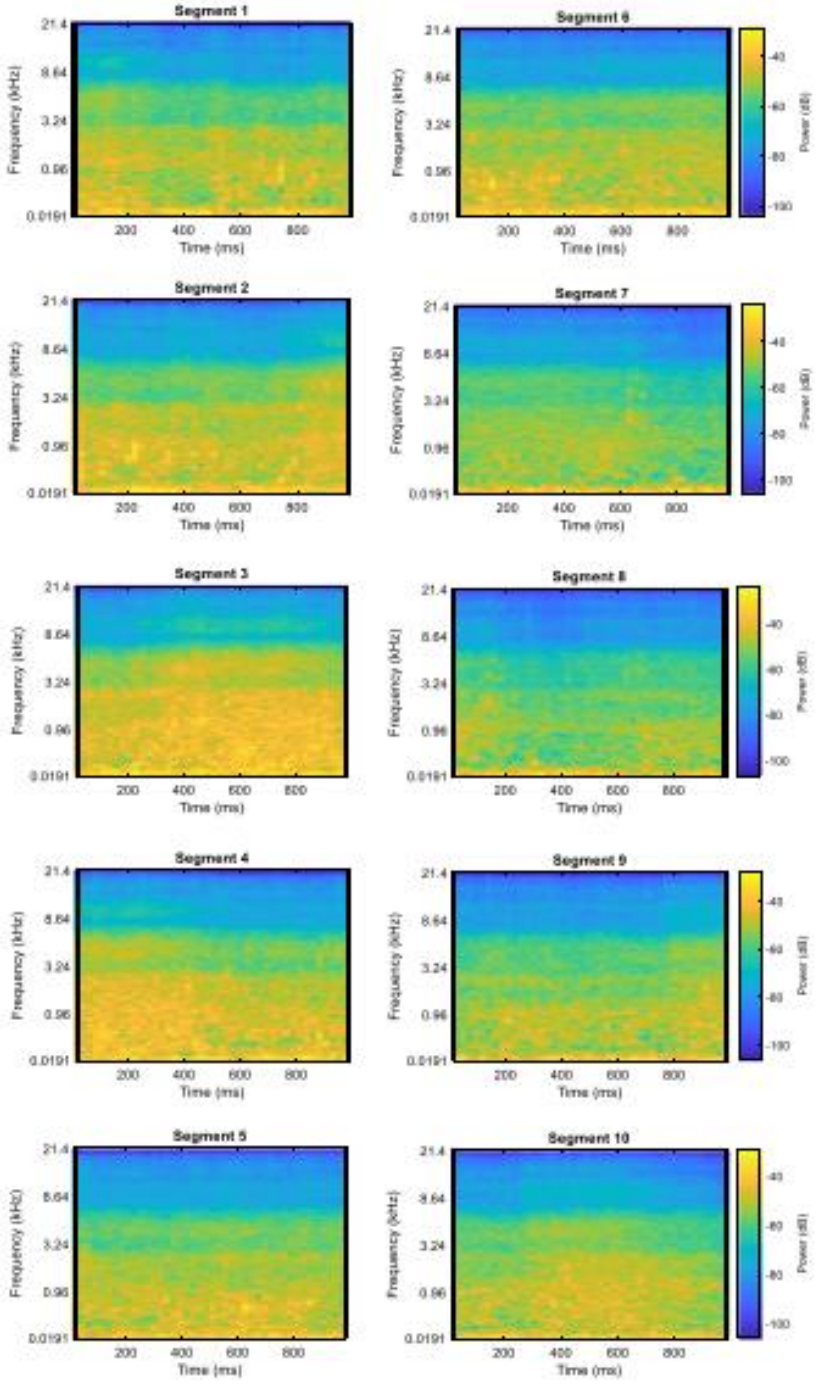


Fig. 2. Mel spectrograms representing the input to the convolutional neural networks (CNNs) of a 10-sec audio file augmented by splitting it into 10 one-second audio files.

Fig. 4. The training progress showing the accuracy and loss of convolutional neural network (CNN) model for each epoch and iteration. Iterations are shown in the horizontal axis, while the y-axis represents accuracy in the upper figure (blue and light blue lines) and loss in the lower figure (red and light red lines). Light colour lines represent smoothed processes, while other lines represent original training procedures. Maximum iterations were set to 5205 (347 iterations per epoch). The shaded vertical rectangles represent the 15 epochs required for the training process. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

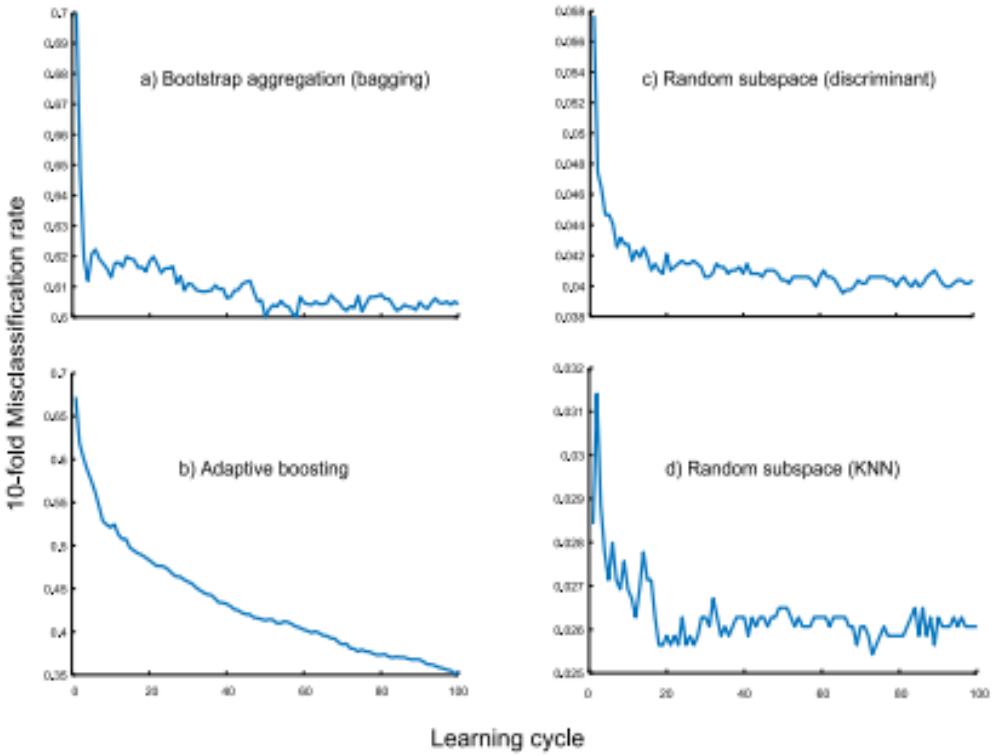
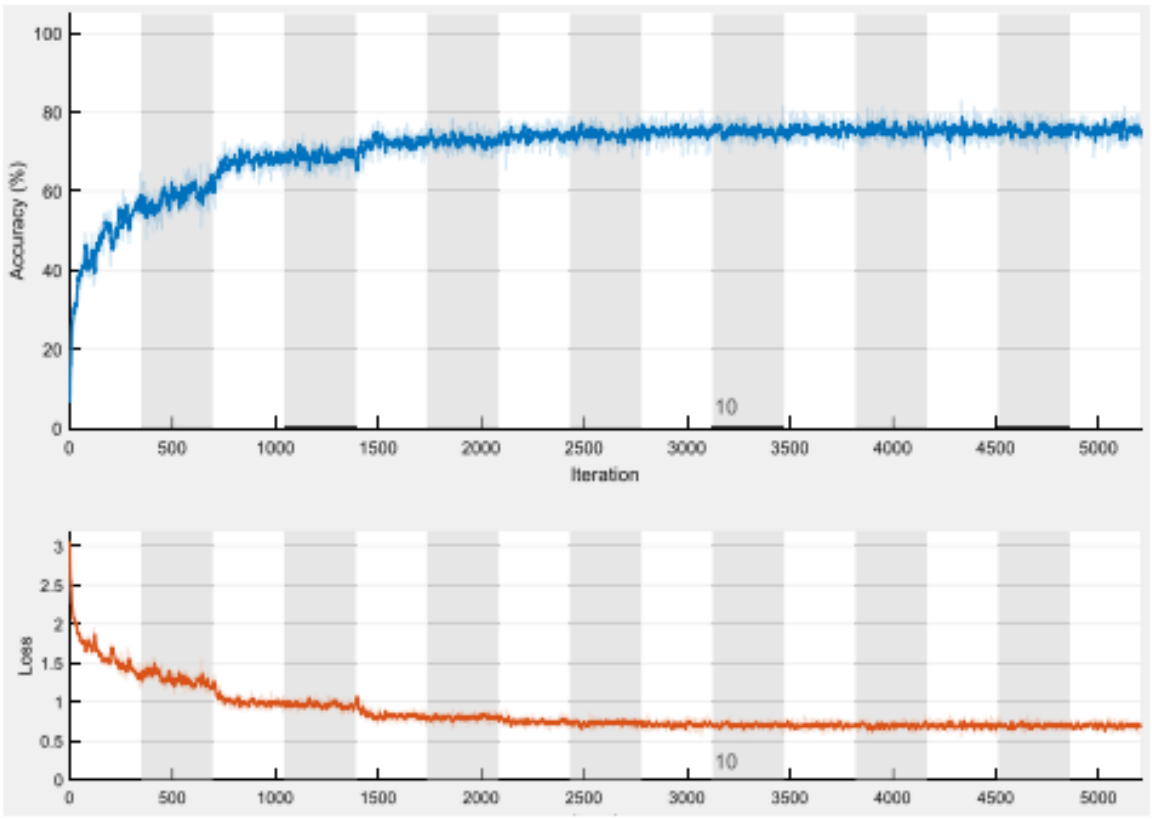
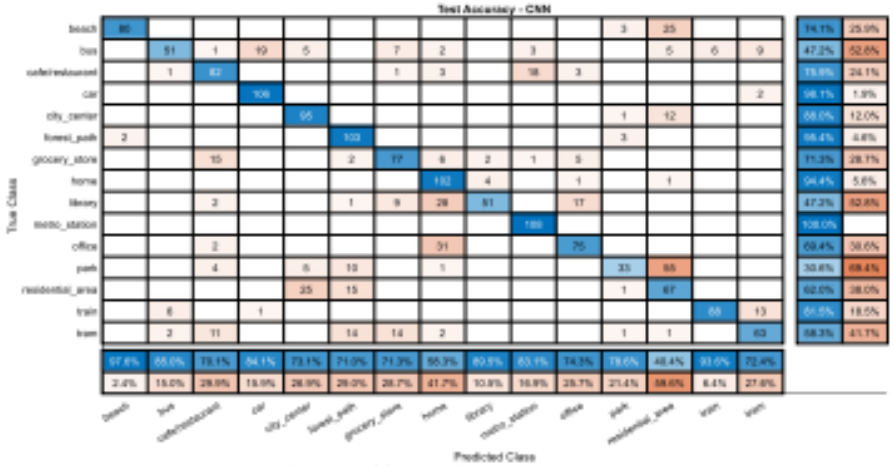
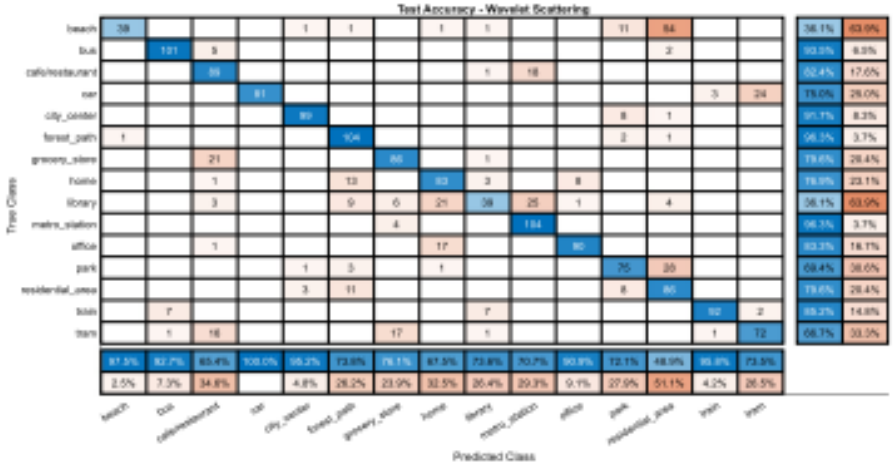


Fig. 3. Examples of hyperparameter optimisation (learning cycle) of some ensemble classifiers. The number of learning cycles that gives the lowest 10-fold misclassification rate was chosen for each classifier.

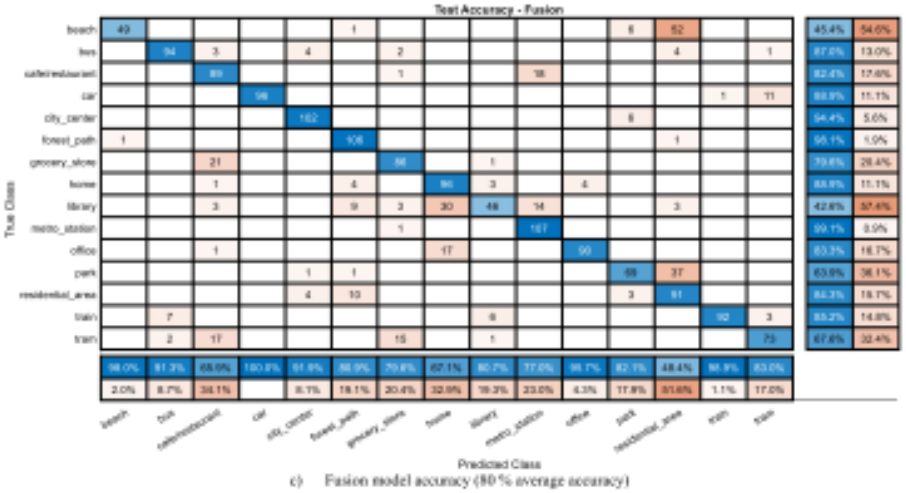




a) CNN model accuracy (72.9 % average accuracy)



b) Random subspace classifier accuracy (76.5 % average accuracy)



c) Fusion model accuracy (80 % average accuracy)

Fig. 5. The confusion matrix of different acoustic scenes for a) CNN model, b) Random subspace discriminant classifier models and c) Fusion model between CNN and random subspace classifier. The number in each cell represents how many times the class shown in the vertical axis was predicted as the class in the horizontal axis. Correct predictions are labelled by blue, while false predictions can be seen by red. Dark colours represent that there were many predictions in these classes. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 2
Average accuracy of different ensemble classifiers and their fusion with the CNNs.

	Bootstrap aggregation (Bagging)	Adaptive boosting	Random subspace classifier [discriminant]	Random subspace classifier [KNN]	Linear programming boosting	Random under-sampling boosting	Totally corrective boosting
Model predictive accuracy (%)	67.2	41.3	76.5	50.1	48.3	49.7	40.4
Fusion with CNN predictive accuracy	74.2	69.8	80	57.6	73.1	73.4	62.1

Table 3
The results of the different best CNN methods on DCASE 2017, as compared to the current study. Average accuracy was commonly reported in these studies, so it has been used for comparison.

Study	Weiping et al. [28]	Hyder et al. [29]	Lehner et al. [30]	Park et al. [31]	Piczak [32]
Method	Early feature fusion of spectrogram and Constant-Q-Transform (CQT) [CNN-SQT]	Spectrogram Image Features (SIF) [CNN-SIF]	I-vectors and CNNs	CNN using double image features	CNN with a frequency resolution
Average accuracy	74.8	74.1	73.8	72.6	70.6
Current proposed model accuracy improvement	7%	8%	8%	10%	13%

Seven different ensemble classifiers were used, and their performance was compared. These classifiers included aggregation (bagging), adaptive boosting, random subspace classifier using KNN and discriminant for feature extraction, linear programming boosting, random under-sampling boosting and totally corrective boosting. MATLAB- defined functions were used to build the ensemble classifiers.

The maximum number of decision splits per tree and number of learning cycles was tuned for each ensemble classifier. For example, an optimisation process was done for ensemble hyperparameter (learning cycles) as shown in Fig. 3. This was based on cumulative 10-fold cross-validated misclassification rates. The cumulative loss allowed monitoring loss with accumulating learners in ensembles. The learning cycle number with minimum misclassification rate was used for each ensemble. The number of minimum 10-fold misclassification errors was different for different classifiers as shown in Fig. 3.

The scattering coefficients were obtained for the scattering decomposition framework and then averaged over 10-second noise samples. For each 10-second noise sample, calling "predict" in MATLAB returns the labels of the corresponding predicted location and the relative confidence in the decision [26,27].

3.4. Fusion model

The fusion model makes use of convolutional neural networks (CNNs) using Mel-spectrograms and ensemble classifiers using wavelet scattering. Roughly an equal overall accuracy can be obtained from CNN and ensemble classifiers; however, each model can outperform the other for predicting particular acoustic scenes. Therefore, merging both models could make use of the advantages of each model individually.

To increase overall accuracy, CNN and ensemble classifier results using late fusion were merged. For each 10-second noise samples, using "predict" function for wavelet classifier and CNN models yields relative confidence in their decision. To create the late fusion model, the wavelet responses were multiplied by CNN responses. The resulting maximum relative confidence of this multiplication is then used to predict outcomes.

3.5. Evaluation and comparison of models

Average accuracy was used to compare performance of models. This enabled comparability of the models with previous results, as average accuracy has been widely used [28,29]. Classification accuracy of each class was first determined by dividing the number of correct predictions over the total number of predictions. Average accuracy was then calculated by classification accuracy arithmetic mean of all classes. Confusion matrices were also presented for best models.

4. Results and discussion

4.1. Results of convolutional neural network (CNN) models

Fig. 4 shows the accuracy and loss during the training process for each iteration of the fifteen epochs included. The accuracy and loss are almost the same after epoch 8. The final overall average accuracy of CNN was 72.9% with SD ±20%. Fig. 5a shows the confusion matrix of CNN models for all included acoustic scenes.

4.2. Results of ensemble classifier models and their fusion with CNNs

Different ensemble classifiers were run as shown in Table 2. The average accuracy of ensemble classifier models ranged between 40.4 and 76.5%. The results of the fusion of these ensemble classifiers are also shown in Table 2. Random subspace had the highest average accuracy (76.5% with SD ±18%) and the highest average accuracy when lately fused with the CNN model (80%).

It was hypothesised that using late fusion models for ASC could yield more accurate models, compared to ordinary CNNs. The results in Table 2 confirmed this hypothesis. The confusion matrix for random subspace classifier and CNN-Random subspace classifier fusion model is shown in Fig. 5. There was a variability of the accuracy of classifying different acoustic scenes for the best fusion model (mean ± SD) 80 ± 17%.

4.3. Average accuracy comparison with previous studies

The results of the current study were compared with previous studies that used early fusion CNN models of the same dataset as

shown in Table 3. These results suggest that the late fusion of CNN with ensemble classifier could be a potential solution for future ASC problems as they could have higher average accuracy than early fusion models.

5. Conclusion

Accurate acoustic scene classification (ASC) models are of great help in many areas. This study presented an enhanced model for ASC by the late fusion of convolutional neural networks (CNNs) and ensemble classifiers. The results showed that the late fusion model had a higher accuracy for ASC, compared to the individual convolutional neural network (CNN) or ensemble classifier models. This fusion model had an average increase in accuracy of 10% as compared to the CNN model average accuracy. A comparison with previous studies using CNN models showed that the late fusion between CNN and ensemble classifier models can yield higher average accuracy, compared to early fusion CNN models (at least 7% increase). This suggests that the proposed late fusion could have promising applications for ASC problems.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Waldekar S, Saha G. Two-level fusion-based acoustic scene classification. Appl Acoust 2020;170:107502. https://doi.org/10.1016/j.apacoust.2020.107502.

[2] Mesaros A, Heittola T, Virtanen T. Acoustic scene classification: An overview of dcase 2017 challenge entries. In: 16th Int Work Acoust Signal Enhanc IWAENC 2018 - Proc. p. 411–5. https://doi.org/10.1109/IWAENC.2018.8521242.

[3] Bianco MJ, Gerstoft P, Traer J, Ozanich E, Roch MA, Gannot S, et al. Machine learning in acoustics: Theory and applications. J Acoust Soc Am 2019;146:3590–628. https://doi.org/10.1121/1.5133944.

[4] Sharan RV, Moir TJ. Acoustic event recognition using cochleagram image and convolutional neural networks. Appl Acoust 2019;148:62–6. https://doi.org/10.1016/j.apacoust.2018.12.006.

[5] Yang L, Tao L, Chen X, Gu X. Multi-scale semantic feature fusion and data augmentation for acoustic scene classification. Appl Acoust 2020;163:107238. https://doi.org/10.1016/j.apacoust.2020.107238.

[6] Su Y, Zhang K, Wang J, Zhou D, Madani K. Performance analysis of multiple aggregated acoustic features for environment sound classification. Appl Acoust 2020;158. https://doi.org/10.1016/j.apacoust.2019.107050.

[7] Zhang T, Liang J, Ding B. Acoustic scene classification using deep CNN with fine-resolution feature. Expert Syst Appl 2020;143:. https://doi.org/10.1016/j.eswa.2019.113067113067.

[8] Mulimani M, Koolagudi SG. Robust acoustic event classification using fusion fisher vector features. Appl Acoust 2019;155:130–8. https://doi.org/10.1016/j.apacoust.2019.05.020.

[9] Dong X, Yan Y, Tan M, Yang Y, Tsang IW. Late fusion via subspace search with consistency preservation. IEEE Trans Image Process 2019;28:518–28. https://doi.org/10.1109/TIP.2018.2867747.

[10] Li G, Ming Z, Li H, Chua TS. Early versus late fusion in semantic video analysis. In: Proc Seventeen ACM Int Conf Multimed. p. 773–6.

[11] Paseddula C, Gangashetty SV. Late fusion framework for Acoustic Scene Classification using LPCC, SCMC, and log-Mel band energies with Deep Neural Networks. Appl Acoust 2021;172:107568. https://doi.org/10.1016/j.apacoust.2020.107568.

[12] Atmaja BT, Akagi M. Multitask learning and multistage fusion for dimensional audiovisual emotion recognition. IEEE 2020:4477–81.

[13] Pei E, Jiang D, Sahli H. An efficient model-level fusion approach for continuous affect recognition from audiovisual signals. Neurocomputing 2020;376:42–53. https://doi.org/10.1016/j.neucom.2019.09.037.

[14] Tsanousa A, Meditskos G, Vrochidis S, Kompatsiaris I. A weighted late fusion framework for recognizing human activity from wearable sensors. In: 10th Int Conf Information, Intell Syst Appl IISA 2019. https://doi.org/10.1109/IISA.2019.8900725.

[15] Mesaros A, Heittola T, Virtanen T. TUT Acoustic scenes 2017; 2017.

[16] Alamir MA, AlHares A, Hansen KL, Elamer A. The effect of age, gender and noise sensitivity on the liking of food in the presence of background noise. Food Qual Prefer 2020:84.

[17] Alamir MA, Hansen K. The effect of type and level of background noise on food liking: A laboratory non-focused listening test. Appl Acoust 2021;172:107600. https://doi.org/10.1016/j.apacoust.2020.107600.

[18] Alamir MA, Hansen KL, Zajamsek B, Catcheside P. Subjective responses to wind farm noise: A review of laboratory listening test methods. Renew Sustain Energy Rev 2019;114:. https://doi.org/10.1016/j.rser.2019.109312109317.

[19] Alamir MA, Hansen KL, Zajamsek B. The effect of wind farm noise on human response: An analysis of listening test methodologies. In: Proc. Acoust. 2018, Adelaide, Australia; 2018. p. 1–9.

[20] Alamir MA, Elamer AA. A compromise between the temperature difference and performance in a standing wave thermoacoustic refrigerator. Int J Ambient Energy 2018;0750:1–13. https://doi.org/10.1080/01430750.2018.1517673.

[21] Alamir MA. Experimental study of the stack geometric parameters effect on the resonance frequency of a standing wave thermoacoustic refrigerator. Int J Green Energy 2019.

[22] Alamir MA. Experimental study of the temperature variations in a standing wave loudspeaker driven thermoacoustic refrigerator. Therm Sci Eng Prog 2019:100361. https://doi.org/10.1016/j.tsep.2019.100361.

[23] Zhang H, Cisse M, Dauphin YN, Lopez-Paz D. mixup: Data-Dependent Data Augmentation; 2017.

[24] Shahriari B, Swersky K, Wang Z, Adams RP, De Freitas N. Taking the human out of the loop: A review of Bayesian optimization. Proc IEEE 2016;104:148–75. https://doi.org/10.1109/JPROC.2015.2494218.

[25] Kulkarni P, Sadasivan J, Adiga A, Seelamantula CS. EPOCH Estimation from a speech signal using gammatone wavelets in a scattering network Department of Electrical Engineering , Indian Institute of Science , Bengaluru - 560012 , India Biocomplexity Institute and Initiative , University of Virginia , Charlo. ICASSP 2020 - 2020 IEEE Int Conf Acoust Speech Signal Process; 2020:7359–63.

[26] Alamir MA. An artificial neural network model for predicting the performance of thermoacoustic refrigerators. Int J Heat Mass Transf 2021;164:. https://doi.org/10.1016/j.ijheatmasstransfer.2020.120551120551.

[27] Alamir MA. Thermoacoustic energy conversion devices: novel insights. J Adv Res Fluid Mech Therm Sci 2021:77.

[28] Weiping Z, Jiantao Y, Xiaotao X, Xiangtao L, Shaohu P. Acoustic scene classification using deep convolutional neural network and multiple spectrograms fusion. Work Detect Classif Acoust Scenes Events 2017:1–5.

[29] Hyder R, Ghaffarzadegan S, Feng Z, Hasan T. Buet Bosch Consortium (B2C) acoustic scene classification systems for Dcase 2017 Challenge. DCASE 2017-Workshop Detect Classif Acoust Scenes Events, 2017.

[30] Lehner B, Eghbal-Zadeh H, Dorfer M, Korzeniowski F, Koutini K, Widmer G. Classifying short acoustic scenes with i-vectors and cnns: challenges and optimisations for the 2017 dcase asc task. DCASE 2017-Workshop Detect Classif Acoust Scenes Events, 2017.

[31] Park S, Mun S, Lee Y, Ko H. Acoustic scene classification based on convolutional neural network using double image features. Work Detect Classif Acoust Scenes Events 2017.

[32] Piczak KJ. The details that matter: frequency resolution of spectrograms in acoustic scene classification. DCASE 2017-Workshop Detect Classif Acoust Scenes Events, 2017.

Expert Systems With Applications 143 (2020) 113067

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Acoustic scene classification using deep CNN with fine-resolution feature

Tao Zhang, Jinhua Liang*, Biyun Ding

Texas Instruments DSP Joint Lab, school of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China

ARTICLE INFO

Article history:
Received 4 August 2019
Revised 27 October 2019
Accepted 27 October 2019
Available online 31 October 2019

Index Term:
Acoustic scene classification
Convolutional neural network
Lateral construction
Depth-wise separable convolution
Fine-resolution convolutional neural network

ABSTRACT

Convolutional neural networks with spectrogram feature representation for acoustic scene classification are attracting more and more attentions due to its favorable performance. However, most of the existing methods are still restricted to the tradeoff between the minimum coverage area across time-frequency feature representation, i.e. time-frequency feature resolution, and the depth of CNN models. Thus, it is unfeasible to improve the performance by simply deepening networks. In this paper, fine-resolution convolutional neural network (FRCNN) is proposed to embrace the progress in very deep architecture, feature fusion and convolutional operation. Specifically, lateral construction is applied to generate a fine-resolution feature map with semantic information, and depth-wise separable convolution is utilized to reduce the number of trainable parameters. Extensive experiments demonstrate that the proposed FRCNN exhibits high performance on several metrics, with low computational complexity.

1. Introduction

Acoustic scene classification (ASC) aims at enabling devices to recognize an audio scene, either from a recording or an on-line stream. "Scene" here is referred as a concept of a specific acoustic environment perceived and defined by human. In other word, scene is the mixture of background noise and sound events associated with a specific audio scenario. In recent years, ASC is attracting more and more research due to its enormous application potential. Over the past few decades, various methods have been exploited for discriminating sound scenes, such as gaussian mixture model (GMM) (Giannoulis et al., 2013; Takahashi, Yamada, Makino, & Ono, 2016), support vector machine (SVM) (Geiger, Schuller, & Rigoll, 2013; Mun, Park, Han, & Ko, 2017), unsupervised learning (Eronen et al., 2005; Bisot, Serizel, Essid, & Richard, 2016, Bisot, Serizel, Essid, & Richard, 2017, Geiger & Helwani, 2015, Salamon & Bello, 2015), multi-layer perception (MLP) (Xu, Huang, Wang, & Plumbley, 2016), recurrent neural networks (RNNs) (Bae, Choi, & Kim, 2016; Bregman, 1994; Vu & Wang, 2016).

Despite their satisfying discrimination accuracy, most of the recognizing methods suffer from two major drawbacks. First, hand-crafted features specific designed for a task (Virtanen, Plumbley, & Ellis, 2018) are required in these methods. This kind of features is hard to be generalized. Thus, most of hand-crafted features need to be combined with others, leading to a large feature vector. Second, the discriminative models in general either lack enough learning capacity or contain too many trainable parameters, providing leeway to improve the discriminating performance.

To overcome the above drawbacks, several feature learning methods based on two-dimensional spectrograms have been recently developed to learn feature representation by convolutional neural networks (CNNs). The resulting models are also able to get better performance with less trainable parameters. Mesaros, Heittola, and Virtanen (Mesaros, Heittola, & Virtanen, 2018) proposed a CNN-based system (hereinafter the baseline system), including convolutional filters, rectifier linear units (ReLU) (Krizhevsky, Sutskever, & Hinton, 2012), and batch normalization (Ioffe & Szegedy, 2015). Kong et al. (Kong, Iqbal, Xu, Wang, & Plumbley, 2018) proposed an eight-layer stacked convolutional model, replacing the large convolutional filter with one sized 3×3. In subsequent works (Mariotti, Cord, & Schwander, 2018; Purohit, Agrawal, & Ramasubramanian, 2018), deep CNN constructions were applied to ASC, and various models with different depth were designed for comparison. Other CNN-related works can be found in (Bae et al., 2016; Liping, Xinxing, & Lianjie, 2018; Valenti, Squartini, Diment, Parascandolo, & Virtanen, 2017).

Although the existing CNN-based methods have shown promising results in feature extraction, they are still restricted to the tradeoff between the minimum coverage area across time-frequency feature representation, i.e. time-frequency feature resolution, and the depth of CNN models. Time-frequency feature resolution plays an important role in detecting specific short

* Corresponding author.
E-mail addresses: zhangtao@tju.edu.cn (T. Zhang), tjljh@tju.edu.cn (J. Liang), 1308401933@qq.com (B. Ding).

https://doi.org/10.1016/j.eswa.2019.113067
0957-4174/

Applied Acoustics 210 (2023) 109411

Contents lists available at ScienceDirect

Applied Acoustics

journal homepage: www.elsevier.com/locate/apacoust

Frequency-based CNN and attention module for acoustic scene classification

Nisan Aryal [a], Sang-Woong Lee [b,*]

[a] Department of IT Convergence Engineering, Gachon University, Seongnam 13120, South Korea
[b] Department of Software, Gachon University, Seongnam 13120, South Korea

ARTICLE INFO

Article history:
Received 11 January 2023
Received in revised form 20 April 2023
Accepted 26 April 2023
Available online 15 May 2023

Keywords:
Acoustic scene classification
Attention
CBAM
SENet
Attention module
DCASE

ABSTRACT

Acoustic scene classification (ASC) is an audio classification task that identifies the environment in which sounds are recorded. Audio-related machine learning algorithms suffer from the device mismatch problem; that is, when trained from audio data recorded from one device, the algorithms cannot generalize to audio samples recorded using another device. In this study, a novel convolutional neural network, called a frequency-aware convolutional neural network (FACNN), is introduced to solve the device mismatch problem by focusing on the frequency information of the audio samples. Furthermore, an attention module, called the frequency attention network (FANet), is introduced to generate an attention map based on the frequency information of the input feature maps. FANet helps the FACNN to focus on the important frequency information, thus improving performance. The proposed method is trained on the TAU Urban Acoustic Scenes 2019 Mobile development dataset and TAU Urban Acoustic Scenes 2020 Mobile development dataset. The proposed method achieves a state-of-the-art accuracy of 75.99% in the TAU Urban Acoustic Scenes 2019 Mobile development dataset and a competitive result of 72.6% in the TAU Urban Acoustic Scenes 2020 Mobile development dataset. In addition, a comparison of FANet with the convolutional block attention module (CBAM) and the squeeze-and-excitation network (SENet) was performed. The results show that FANet can mitigate the device mismatch problem by improving the performance of the unseen devices.

1. Introduction

Acoustic scene classification (ASC) is the task of labeling an audio sample according to the environment in which it has been recorded (such as park, library, market etc.)[1–4]. Acoustic scenes are combinations of environmental sounds. The main goal of ASC is to create a model to identify these unique environmental sounds, which in turn helps to identify the environment.

Human hearing can easily distinguish the location, object, and tentative distance of a sound. However, it is quite difficult for computers to understand the information present in sound. Sounds contain information, such as the objects present in the environment, the time of day, season, amount of activity. Information from sound is used in fields such as multimedia retrieval [5], surveillance and monitoring using audio [6,7], context-aware services [8], detecting abnormal activities in machines, and robotic navigation [9]. Computer vision technologies have been used to study autonomous robots and cars. Acoustic scene information can help to increase the efficiency and practicality of these technologies.

CNN-based networks are a popular approach for audio processing [10–14]. An audio signal consists of one-dimensional sequential data. However, the one-dimensional audio is often changed into two-dimensional features during preprocessing. These two-dimensional features are referred to as time frequency (TF) representation. Some examples of TF representations in audio signal processing are the spectrogram, mel spectrogram, and constant-Q transform (CQT). These TF representations contain features that can be extracted using a convolutional neural network (CNN). In addition, the majority of popular approaches in audio processing are influenced by computer vision networks [15]. In general, VGG and residual network-based architectures are used in ASC [16–19]. Recently, attention-based methods have been introduced for ASC [20–22]. Attention was first introduced in natural language processing to address long-term dependency [23]. However, the concept of attention has been broadened and used in different ways to enhance network performance [24,25]. The attention modules have been widely used in ASC. Furthermore, attention modules such as convolutional block attention module (CBAM) [26]

* Corresponding author.
E-mail addresses: nisanaryal123@gmail.com (N. Aryal), sleo@gachon.ac.kr (S.-W. Lee).

https://doi.org/10.1016/j.apacoust.2023.109411
0003-682X/

Constrained Learned Feature Extraction for Acoustic Scene Classification

Teng Zhang, *Member, IEEE,* Ji Wu, *Senior Member, IEEE*

***Abstract*—Deep neural networks (DNNs) have been proven to be powerful models for acoustic scene classification tasks. State-of-the-art DNNs have millions of connections and are computationally intensive, making them difficult to deploy on systems with limited resources. With a focus on acoustic scene classification, we describe a new learnable module, the simulated Fourier transform module, which allows deep neural networks to implement the discrete Fourier transform (DFT) operation 8x faster on a GPU. We frame the signal processing procedure as an adaptive machine learning problem and introduce learnable parameters in the module to facilitate fast adaptation for the complex and variable acoustic signal. This module gives neural networks the ability to model audio signals from raw waveforms, without extra FFT and filter bank patches. Then we use the temporal transformer module, which has been previously published, to alleviate the information loss caused by the simulated Fourier transform module. These techniques can be integrated into an existing fully connected neural network (FCNN), convolutional neural network (CNN) or recurrent neural network (RNN) model. We evaluate the proposed strategy using four acoustic scene datasets (LITIS Rouen, DCASE2016, DCASE2017, DCASE2018) as target tasks. We show that the proposed approach significantly outperforms the vanilla FCNN, CNN and RNN approach on both efficiency and performance. For instance, the proposed approach can reduce inference time by 8x while reducing the classification error on LITIS Rouen dataset from 3.21% to 1.81%.**

***Index Terms*—Deep neural networks, Fourier transform, Acoustic scene classification**

I. Introduction

ACOUSTIC scene classification (ASC) is a task of classifying environments from the sounds they produce [1][2], with applications in devices where the environment can be defined based on physical or social context, e.g., park, office, meeting, etc [3].

Influenced by traditional speech and music processing methods, early works on ASC focus on modelling the time-frequency characteristic of audio features. Over recent years, the landscape of audio analysis has been drastically altered and pushed forward through DNN architectures. Some works [4][5] look at extracting high-level acoustic features from the output of short-time Fourier transform (STFT) [6] outputs. Many groups [7][8] have found the logarithm of critical band energies extracted from Mel-frequency filter banks to be most suitable for training DNNs. One of their advantages is that they contain more high-resolution information than standard Mel-Frequency Cepstral Coefficients (MFCC) features [9].

T. Zhang was with the Department of Electronic Engineering, Tsinghua University, Beijing, P.R.China. e-mail: zhangteng1887@gmail.com.
J. Wu was with the Department of Electronic Engineering, Tsinghua University, Beijing, P.R.China. e-mail: wuji_ee@mail.tsinghua.edu.cn.

Ultimately, to avoid loss of potentially useful information, acoustic models need to be trained on low-level audio representations. Sainath et al. [10] improve speech recognition performance by training a model on linear frequency magnitude features to learn filter bank weights automatically. More recently, other works have attempted to do feature learning directly from time domain waveforms. Results in [11][12] show that DNNs trained on raw waveforms tend to learn bandpass filters and auditory-like features, with the performance gap between raw waveforms and conventional MFCC features getting narrow or even disappearing. Hoshen et al. [13] describe a temporal convolutional neural network which takes raw multichannel waveforms as input. Van et al. [14] introduce WaveNet, a deep neural network for generating raw audio waveforms, adapt it to speech recognition tasks and get the best score obtained from raw waveforms on TIMIT dataset.

Our goal is to increase the efficiency of DNN models for ASC tasks without compromising their classification performances. For DNN models trained on raw waveforms, the computation time T_{all} that the classification process takes is the product of the number of frames n_f which should be processed by neural networks and the time required to process each frame. The latter can be decomposed into audio preprocessing time t_p and the computation time t_i for each layer of DNN models. Define the number of layers in the neural network to be L, T_{all} can be represented as follows:

$$T_{all} = n_f(t_p + \sum_{i=1}^{L} t_i) \tag{1}$$

The computation time t_i in Eq.1 is determined by the number of weights in DNN models. While large DNN models are very powerful, a growing body of research demonstrates that these approaches incorporate significant redundancy and are computationally intensive. Recently, it has been shown that network pruning [15], network quantization and weight sharing [16][17][18] methods can effectively reduce t_i without affecting accuracy. With a focus on acoustic scene classification, the number of weights is limited by the limited training data. For instance, LITIS Rouen dataset [19], which is the largest dataset publicly available for ASC task to our knowledge, contains only 1500 minutes of acoustic scene recordings. This makes the computation time t_i far less than the preprocessing time t_p. Reducing the number of frames n_f in Eq.1 can synchronously decrease the amount of T_{all} required for classification. In the conventional temporal sampling of STFT, the frameshift is typically fixed to half of the frame length [20] for

Applied Acoustics 172 (2021) 107568

Contents lists available at ScienceDirect

Applied Acoustics

journal homepage: www.elsevier.com/locate/apacoust

Late fusion framework for Acoustic Scene Classification using LPCC, SCMC, and log-Mel band energies with Deep Neural Networks

Chandrasekhar Paseddula [a,*], Suryakanth V. Gangashetty [a]

[a] Speech Processing Laboratory, International Institute of Information Technology-Hyderabad, India

ARTICLE INFO

Article history:
Received 4 August 2019
Received in revised form 4 July 2020
Accepted 24 July 2020
Available online 25 August 2020

Keywords:
Linear Prediction Cepstral Coefficients
Spectral Centroid Magnitude Cepstral Coefficients
log-Mel band energies
Acoustic Scene Classification
Deep Neural Networks

ABSTRACT

A major problem in Acoustic Scene Classification (ASC) is a representation of an acoustic scene, which serves to be an important task for ASC. This study used Linear Prediction Cepstral Coefficients (LPCC) and Spectral Centroid Magnitude Cepstral Coefficients (SCMC) features along with log-Mel band energies for the representation of an acoustic scene. Deep Neural Networks (DNN) is being used to model the Acoustic Scene Classification (ASC). LPCCs are used to capture the changes in the auditory spectrum with time and SCMCs are used to capture the weighted average magnitude finely for a given acoustic scene subband. log-Mel band energies are used to capture the spectral envelopes of audio frame. The DNN architecture is used for audio track level classification. We 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.

1. Introduction

Acoustic Scene Classification (ASC) task research has been expanding recently in the fields of signal processing, machine learning, auditory-motivated methods, cross-disciplinary methods involving acoustics, biology, psychology, geography, material science, transport science, etc., [1]. This task is helpful in various applications such as smartphones, Internet of things for audio monitoring, robots, etc., to automatically monitor sound from various environments [2]. ASC is one of the subtasks of Computational Environmental Audio Analysis (CASA) and the objective of ASC is to classify the environment into one of the defined acoustic scenes [1,2]. The defined environments may be indoor (airport and shopping_mall), outdoor (metro_station, park, public_square, street_traffic, and street_pedestrian) and travelling vehicles (bus, metro, and tram). Other motives for ASC research are the DCASE challenges of 2016, 2017, 2018 and 2019 where technical committee organizers have developed different baselines and have provided development datasets along with evaluation dataset to the challenge participants, thereby motivating new algorithms and improving the baseline system in ASC.

Time–Frequency Representations (TFR) of an acoustic scene play a major aspect in acoustic scene analysis. Temporal and spectral structures of acoustic scene give a good visual representation and can be seen as 2D spectrograms [3]. It was presented that the discriminative approaches of ASC (e.g., Support Vector Machine (SVM)) has given more accuracy than the generative models (e.g. Gaussian Mixture Model (GMM)). The results from previous research studies, show the need for suitable feature-classifier combinations in order to improve performance in [2].

The inter and intraclass variations in environmental sounds need a complex representation. Consequently, one particular feature may not be sufficient to represent and discriminate them effectively. This can be inferred from the Fig. 1, which visualizes randomly selected ten-second duration sample spectrograms of ten classes (two indoor, five outdoor, and three travelling vehicles) from DCASE 2018 development dataset. From the spectrograms, it can be noticed that in the lower frequency region, the concentration of energy is more. The spectral characteristics of all classes differ from each other and also has intra class spectral difference. Due to that conventional features like Mel Frequency Cepstral Coefficients (MFCC) and log-Mel band energies may not be captured well [4].

In this paper, we proposed to study the influence of LPCCs, SCMCs and log-Mel band energies on acoustic scene representations individual and combined nature. The motivation behind using these spectral features, such as LPCC, SCMC, log-Mel band

* Corresponding author.
E-mail addresses: chandrasekhar.p@research.iiit.ac.in (C. Paseddula), svg@iiit.ac.in (S.V. Gangashetty).

https://doi.org/10.1016/j.apacoust.2020.107568
0003-682X/

2019 IEEE International Conference on Multimedia and Expo (ICME)

ACOUSTIC SCENE CLASSIFICATION WITH MISMATCHED RECORDING DEVICES USING MIXTURE OF EXPERTS LAYER

Truc Nguyen *, *Franz Pernkopf* †

Graz University of Technology,
Signal Processing and Speech Communication Lab.,
Inffeldgasse 16c, A-8010 Graz, Austria/Europe,
{t.k.nguyen, pernkopf}@tugraz.at

ABSTRACT

Recently, a mismatch in acoustic conditions such as a temporal recording gap as well as different recording devices for the development and the evaluation data has been considered in Acoustic Scene Classification (ASC). This brings ASC closer to real world conditions. In this paper, we address ASC with mismatching recording devices. This has been introduced as task 1B of the DCASE 2018 challenge. We proposed a flexible and robust model that uses a mixture of experts (MoE) layer replacing the fully connected dense layer such that each expert can adapt to the specific domains of the data. Furthermore, we observe different Convolutional Neural Network (CNN) models as well as the number of the experts of the MoE dense layer using log-mel features. In addition, we perform mixup data augmentation to enhance the robustness of our models. In experiments, the classification performance is 66.1% using 15 experts in the MoE dense layer with approximately 2M parameters. This outperforms the best model of task 1B of the DCASE 2018 challenge by 2.5% (absolute). This model uses an ensemble selection of 12 individual models with ∼12M parameters.

Index Terms— Acoustic scene classification, convolutional neural network, mixture of experts layer, mixture of softmaxes.

1. INTRODUCTION

Acoustic scene classification (ASC) is a multi-class classification task classifying the recorded environment sounds as specific acoustic scenes that characterize either the location or situation such as park, metro station, tram, etc. It has been a task in the Detection and Classification of Acoustic Scenes and Events (DCASE) challenges providing the largest publicly available data sets for ASC.

Compared to ASC tasks of DCASE 2013 and DCASE 2016, the difficulty has been increased for DCASE 2018. Beside providing shorter segments of 10 s of audio data, there are mismatches between the development data set and the evaluation set. According to [1], there was a mismatch in acoustic conditions in the evaluation and the development set of the DCASE 2017 challenge i.e. data sets were recorded in similar locations with the same device but almost one year later. This temporal gap is the reason of a significant drop in performance of all systems. The DCASE 2018 challenge introduced task 1B with data sets recorded by four different devices in 6 different cities in Europe instead of only one city. This causes even more mismatch in the data. Especially, a part of the evaluation set is a compressed version of recorded audio data from device D that is not included in the development data set. This causes an extreme mismatch in the DCASE 2018 challenge data.

Recent ASC research mostly uses log-mel energies and mel-frequency cepstral coefficients (MFCC) as features. Beside that, harmonic-percussive source separation (HPSS) and I-vectors extracted from these mel-frequency scales have been effective features contributing to the success in the last DCASE challenges [2], [3], [4]. Some systems use the raw waveform [5] and conventional signal processing methods such as wavelet decomposition [6], [7] for feature extraction. For classification, deep learning (DL) has been the preferred solution. Beside well-known DL models used for image databases such the VGGNet [2], [3], [4], and Xception [8], popular models for acoustic data such as Recurrent Neural Networks (RNNs) or Long Short term Memmories (LSTMs) have been used [9], [10], [11]. Recently, attention mechanisms have been introduced [12], [13], [14] to supplement vanilla DL models. In addition, techniques of data augmentation such as Generative Adversarial Networks (GANs) [15] and mixup have been used [3]. Furthermore, ensemble methods helped the systems to the top performances in DCASE 2017 [2] and DCASE 2018 [3], [16].

Although a variety of ASC systems have been proposed, there is a limited number that focused on the analysis of the mismatching acoustic conditions. In this paper, we focus on the DCASE 2018 data of task 1B where the recording took place at several cities with different devices. We propose a

*Thanks to Vietnamese - Austrian Government Scholarship for funding.
†Thanks to Austrian Science Fund.

978-1-5386-9552-4/19/$31.00 ©2019 IEEE
DOI 10.1109/ICME.2019.00287

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

INASS

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Acoustic Scene Classification using Attention based Deep Learning Model

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

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

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

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

1. Introduction

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

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

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

Pattern Recognition Letters 164 (2022) 40–45

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

An Open-Set Recognition and Few-Shot Learning Dataset for Audio Event Classification in Domestic Environments

Javier Naranjo-Alcazar[a,b], Sergi Perez-Castanos[a], Pedro Zuccarello[a], Ana M. Torres[c], Jose J. Lopez[d], Francesc J. Ferri[b], Maximo Cobos[b,*]

[a] Visualfy, Benisanó 46181, Spain
[b] Computer Science Department, Universitat de València, Burjassot 46100, Spain
[c] Dpt. Ingeniería eléctrica, electrónica, automática y comunicaciones, Universidad de Castilla-La Mancha, Cuenca 16002, Spain
[d] iTEAM Institute, Universitat Politècnica de València, Valencia 46022, Spain

ARTICLE INFO

Article history:
Received 16 April 2022
Revised 6 July 2022
Accepted 19 October 2022
Available online 22 October 2022

Edited by Maria De Marsico

Keywords:
Audio Dataset
Classification
Few-Shot Learning
Machine Listening
Open-set Recognition
Sound Processing

ABSTRACT

The problem of training with a small set of positive samples is known as few-shot learning (FSL). It is widely known that traditional deep learning algorithms usually show very good performance when trained with large datasets. However, in many applications, it is not possible to obtain such a high number of samples. This paper deals with the application of FSL to the detection of specific and intentional acoustic events given by different types of sound alarms, such as door bells or fire alarms, using a limited number of samples. These sounds typically occur in domestic environments where many events corresponding to a wide variety of sound classes take place. Therefore, the detection of such alarms in a practical scenario can be considered an open-set recognition (OSR) problem. To address the lack of a dedicated public dataset for audio FSL, researchers usually make modifications on other available datasets. This paper is aimed at providing the audio recognition community with a carefully annotated dataset[1] for FSL in an OSR context comprised of 1360 clips from 34 classes divided into *pattern sounds* and *unwanted sounds*. To facilitate and promote research on this area, results with state-of-the-art baseline systems based on transfer learning are also presented.

1. Introduction

The automatic classification of audio clips is a research area that has grown significantly very recently [1–3]. The research interest in these algorithms is motivated by their numerous applications, such as audio-based surveillance, hearing aids, home assistants or ambient assisted living, among others. In contrast to most deep learning methods, few-shot learning (FSL) tackles the problem of learning with few samples per class. FSL approaches gained focus when trying to address intra-class classification in the context of face recognition problems [4], including applications such as access control and identity verification [5–7]. In order to tackle this problem, loss functions such as ring loss [8] or center loss [9] have been proposed, together with different embeddings from network architectures such as siamese [10,11] and triplet [12,13]. These loss functions are aimed at solving convergence issues, which also require careful training procedures to appropriately choose the pairs or triplets used. Another practical issue arising in many real-world intelligent audio applications is open-set recognition (OSR) [14]. This problem occurs when a system has to face unfamiliar situations for which it has not been trained. A system prepared for OSR should be capable of correctly classifying examples corresponding to classes seen during the training stage while rejecting examples corresponding to new, previously unseen classes. OSR has been addressed in the past by applying modifications to classical machine learning algorithms such as support vector machines [15,16] or nearest neighbour classification [17]. In the last years, deep learning solutions for OSR have also started to emerge, such as OpenMax [18], deep open classifier (DOC) [19] or competitive overcomplete output layer (COOL) [20].

The problems of FSL and OSR appear frequently in smart acoustic applications. For example, a given user may be exposed to several alerts or beeps at home, emitted by different domestic ap-

* Corresponding author.:
E-mail addresses: javier.naranjo@visualfy.com, janal2@alumni.uv.es (J. Naranjo-Alcazar), sergi.perez@visualfy.com (S. Perez-Castanos), pedro.zuccarello@visualfy.com (P. Zuccarello), ana.torres@uclm.es (A.M. Torres), jjlopez@dcom.upv.es (J.J. Lopez), francesc.ferri@uv.es (F.J. Ferri), maximo.cobos@uv.es (M. Cobos).

[1] https://zenodo.org/record/3689288.

https://doi.org/10.1016/j.patrec.2022.10.019

Applied Acoustics 192 (2022) 108698

Contents lists available at ScienceDirect

Applied Acoustics

journal homepage: www.elsevier.com/locate/apacoust

Late fusion for acoustic scene classification using swarm intelligence

Biyun Ding, Tao Zhang *, Ganjun Liu, Lingguo Kong, Yanzhang Geng

School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China

ARTICLE INFO

Article history:
Received 24 May 2021
Received in revised form 19 February 2022
Accepted 23 February 2022

Keywords:
Acoustic scene classification
Late fusion
Swarm intelligence
Optimization

ABSTRACT

Acoustic scene classification (ASC) has gained significant interest in recent years due to its diverse applications. However, the performance of ASC is much lower than other audio processing areas, such as speech recognition and music classification. Various audio signal processing and machine learning methods have been proposed for ASC systems with good performance. In this sense, the performance can be significantly improved by taking advantage of these methods together. Late fusion is a commonly used approach to obtain the final decision for a test instance, which fuses the prediction results of the different models. However, it is ubiquitous that different models dispute the prediction on the same data, leading to performance degeneration. This study presents an efficient and effective approach to fuse predictions from multiple sources based on a swarm intelligence algorithm. In this approach, the late fusion procedure is defined as a global optimization problem and the swarm intelligence algorithm is introduced to search an optimal system subset obtaining the best classification performance after late fusion for ASC. The Swarm Intelligence algorithm based Late Fusion (SILF) can avoid the performance degeneration caused by the controversy of multiple sources and optimize the source combination for fusion. The experiments demonstrate the efficacy of SILF for ASC and the performance improvement, which outperforms the state-of-the-art late fusion algorithms on the TAU Urban Acoustic Scenes 2019 development dataset.

1. Introduction

Acoustic scene classification (ASC) is a classification task of assigning predefined semantic labels to audio streams recorded in a certain environment by analyzing audio signals [1]. The semantic labels describe the environment information of the audio streams. Various potential applications promoted those studies, such as context awareness in smart devices, hearing aids, robots, wild-life monitoring in natural habitats, and Internet of Things for audio monitoring, etc., to automatically monitor sound from various environments [2,3]. However, the performance of ASC is relatively low compared with other audio areas, such as speech recognition and music classification. Therefore, many methods have been proposed and developed to achieve higher performance for ASC tasks.

Early works about ASC can be found in [4–10]. A major initial contribution to environmental sound research is within the framework of Acoustic Ecology advanced by Schafer [4], who advanced the notion of the soundscape as the totality of all sounds in the listener's dynamic environment. Then, Gaver [5] attempted to construct a descriptive framework based on what we listen for in everyday sounds. After that, Sawhney [6] reported a situational awareness from environmental sounds, which focused on a simple classification of five predefined classes of environmental sounds using a simple nearest-neighbor algorithm. Moreover, they reported the classification of ten classes of environmental sounds using HMM [7]. In addition, Truax [8] treats listening as an active process of interacting with the environment and distinguishes it among several different listening levels such as listening-in-search, listening-in-readiness, and background listening. Its theoretical constructs were practically useful in designing functional and aesthetically stimulating acoustic environments [9]. Due to these previous works on ASC, the perception of environmental sounds attracts attention.

During the last decades, many researches have been done and significantly improved the ASC performance. The methods have been investigated mainly from the audio input, pre-processing, feature extraction, classification, and output. For instance, to enhance the input data, multichannel signals are employed, such as binary channel (left and right), mid/side channels [10], and filtered signal variants generated by Harmonic Percussive Source Separation (HPSS) [11] or Nearest Neighbor Filtering (NNF) [12]. Multichannel audio recordings are also used to exploit the inherent spatial information to localize sound sources better [13]. Moreover, data augmentation is also applied to increase the quality and

* Corresponding author.
E-mail address: zhangtao@tju.edu.cn (T. Zhang).

https://doi.org/10.1016/j.apacoust.2022.108698
0003-682X/

FEW-SHOT SOUND EVENT DETECTION

Yu Wang[1*] *Justin Salamon*[2] *Nicholas J. Bryan*[2] *Juan Pablo Bello*[1]

[1]Music and Audio Research Laboratory, New York University, NY, USA
[2]Adobe Research, San Francisco, CA, USA

ABSTRACT

Locating perceptually similar sound events within a continuous recording is a common task for various audio applications. However, current tools require users to manually listen to and label all the locations of the sound events of interest, which is tedious and time-consuming. In this work, we (1) adapt state-of-the-art metric-based few-shot learning methods to automate the detection of similar-sounding events, requiring only one or few examples of the target event, (2) develop a method to automatically construct a partial set of labeled examples (negative samples) to reduce user labeling effort, and (3) develop an inference-time data augmentation method to increase detection accuracy. To validate our approach, we perform extensive comparative analysis of few-shot learning methods for the task of keyword detection in speech. We show that our approach successfully adapts *closed-set* few-shot learning approaches to an *open-set* sound event detection problem.

Index Terms— Few-shot learning, sound event detection, keyword detection, keyword spotting, speech

1. INTRODUCTION

Locating perceptually similar sound events within a continuous recording is a basic, but important task for many audio applications. For example, animators need to locate particular sounds in music and SFX tracks to synchronize motion graphics, and podcasters have to edit out filler words (e.g. "ahhs" and "umms") to improve the flow of speech. Noise monitoring solutions require identifying specific sound events [1] and, more generally, labeling large audio datasets for training machine learning models often requires identifying all time locations where specific sound events occur [2–4]. However, current audio processing tools require users to listen through the entire recording and manually identify and label all the locations of the sound events of interest, which is both hard and tedious. A method to automate this process would save a significant amount of time and human effort.

Modern deep learning-based sound event recognition and detection methods typically require large amounts of data for training or fine-tuning models for specific applications [5–9]. As such, the application of deep learning models to detect unseen and/or rare sound classes with only few labels has been very limited. Interactive user-in-the-loop sound event detection was proposed in [10], but this work focused on reducing annotation time rather than improving machine accuracy. Different strategies for training audio classifiers with few data for the task of acoustic event recognition were investigated in [11], however, it focused on coping with limited data during *training*, whereas our goal is to train models that can generalize to unseen classes for which we have very few examples at *inference time*.

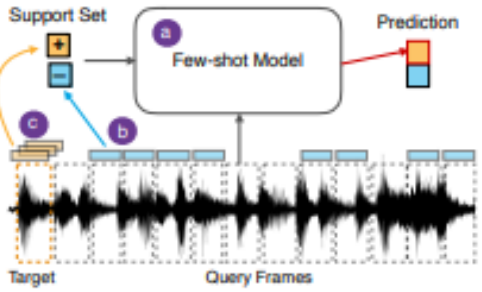


Fig. 1. Proposed few-shot sound event detection method. To detect a target sound event in a recording, we (a) apply metric-based few-shot learning, (b) automatically construct a set of negative (blue) examples needed for inference, and (c) propose an inference-time data augmentation method to generate more positive examples (orange).

Recently, studies have proposed to tackle this latter problem using few-shot learning, where a classifier must learn to recognize novel classes given only few examples from each [12, 13]. Traditional few-shot learning methods consider a *C-way K-shot* classification task as a *closed-set* classification problem of labeling an audio query with one of C unique class labels, given K labeled examples per class, where C is *fixed*. To the best of our knowledge, however, few-shot learning has not been applied to an *open-set* problem, such as sound event detection, where a previously unseen target sound needs to be detected in a sequence of unknown, previously unseen sounds from an unbounded number of sound classes.

In this paper, we propose to leverage few-shot learning for open-set sound event detection in order to identify perceptually similar sound events within a recording. In doing so, we (1) adapt prior metric-based few-shot learning approaches to the *open-set* sound event detection task, (2) propose a method to automatically construct a set of labeled negative examples required at inference time, and (3) propose an inference-time data augmentation method to increase detection accuracy, while reducing user-labeling effort. In Figure 1, we depict the proposed method with our key contributions mentioned above, which is applicable to a variety audio domains such as speech, music, and environmental sound. We evaluate our approach on speech and (4) provide extensive comparative analysis of few-shot learning for the application of sound-based keyword detection in speech. We show that our method achieves an average area under the precision-recall curve (AUPRC) of 75.42% for detecting unseen target keywords with only five labeled examples provided. Finally, we (5) show that our approach generalizes to unseen languages without requiring any retraining or fine tuning.

* This work was performed during an internship at Adobe Research. This work was partially supported by National Science Foundation award 1544753.

Ecological Informatics 77 (2023) 102258

Contents lists available at ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf

Learning to detect an animal sound from five examples

Ines Nolasco [c,1], Shubhr Singh [c,1], Veronica Morfi [c,1], Vincent Lostanlen [d,e], Ariana Strandburg-Peshkin [f,h,g], Ester Vidaña-Vila [i], Lisa Gill [j,f], Hanna Pamuła [k], Helen Whitehead [m], Ivan Kiskin [n], Frants H. Jensen [p,q,r], Joe Morford [o], Michael G. Emmerson [c], Elisabetta Versace [c], Emily Grout [h,f,g], Haohe Liu [n], Burooj Ghani [a], Dan Stowell [a,b,*]

[a] Naturalis Biodiversity Center, Leiden, the Netherlands
[b] Tilburg University, Tilburg, the Netherlands
[c] Queen Mary University of London, London, UK
[d] Centre National de la Recherche Scientifique (CNRS), France
[e] Nantes Université, Ecole Centrale Nantes, CNRS, LS2N, UMR 6004, F-44000 Nantes, France
[f] Biology Department, University of Konstanz, Universitätsstrasse 10, Konstanz, Germany
[g] Centre for the Advanced Study of Collective Behaviour, University of Konstanz, Universitätsstrasse 10, Konstanz, Germany
[h] Department for the Ecology of Animal Societies, Max Planck Institute of Animal Behavior, Bücklestrasse 5, Konstanz, Germany
[i] Landesbund für Vogel- und Naturschutz, Hilpoltstein, Germany
[j] Naturkundemuseum Bayern/BIOTOPIA Lab, Munich, Germany
[k] AGH University of Science and Technology, al. Adama Mickiewicza 30, Krakow, Poland
[l] La Salle Campus Barcelona, Universitat Ramon Llull, Sant Joan de La Salle, 42, Barcelona, Spain
[m] School of Science, Engineering and Environment, University of Salford, Manchester, UK
[n] Centre for Vision, Speech and Signal Processing, FEPS, University of Surrey, Surrey, UK
[o] The Oxford Navigation group, Dept. of Zoology, University of Oxford, Oxford, UK
[p] Aarhus University, Department of Ecoscience, Frederiksborgvej 399, 4000 Roskilde, Denmark
[q] Syracuse University, 107 College Place, Syracuse, NY 13244, USA
[r] Woods Hole Oceanographic Institution, 266 Woods Hole Rd, Woods Hole, MA 02543, USA

ARTICLE INFO

Keywords:
Bioacoustics
Deep learning
Event detection
Few-shot learning

ABSTRACT

Automatic detection and classification of animal sounds has many applications in biodiversity monitoring and animal behavior. In the past twenty years, the volume of digitised wildlife sound available has massively increased, and automatic classification through deep learning now shows strong results. However, bioacoustics is not a single task but a vast range of small-scale tasks (such as individual ID, call type, emotional indication) with wide variety in data characteristics, and most bioacoustic tasks do not come with strongly-labelled training data. The standard paradigm of supervised learning, focussed on a single large-scale dataset and/or a generic pre-trained algorithm, is insufficient. In this work we recast bioacoustic sound event detection within the AI framework of *few-shot learning*. We adapt this framework to sound event detection, such that a system can be given the annotated start/end times of as few as 5 events, and can then detect events in long-duration audio—even when the sound category was *not known* at the time of algorithm training. We introduce a collection of open datasets designed to strongly test a system's ability to perform few-shot sound event detections, and we present the results of a public contest to address the task. Our analysis shows that prototypical networks are a very common used strategy and they perform well when enhanced with adaptations for general characteristics of animal sounds. However, systems with high time resolution capabilities perform the best in this challenge. We demonstrate that widely-varying sound event durations are an important factor in performance, as well as non-stationarity, i.e. gradual changes in conditions throughout the duration of a recording. For fine-grained bioacoustic recognition tasks without massive annotated training data, our analysis demonstrate that few-shot sound event detection is a powerful new method, strongly outperforming traditional signal-processing detection methods in the fully automated scenario.

* Corresponding author at: Tilburg University, Tilburg, the Netherlands.
E-mail address: d.stowell@tilburguniversity.edu (D. Stowell).
[1] Equal first authors.

https://doi.org/10.1016/j.ecoinf.2023.102258
Received 22 May 2023; Received in revised form 9 August 2023; Accepted 9 August 2023
Available online 10 August 2023

Applied Acoustics 170 (2020) 107502

Contents lists available at ScienceDirect

Applied Acoustics

journal homepage: www.elsevier.com/locate/apacoust

Two-level fusion-based acoustic scene classification

Shefali Waldekar [a,*], Goutam Saha [a]

[a] Dept of Electronics and Electrical Communication Engineering, IIT Kharagpur, Kharagpur, India

ARTICLE INFO

Article history:
Received 2 June 2019
Received in revised form 20 March 2020
Accepted 19 June 2020
Available online 11 July 2020

Keywords:
Environmental acoustics
Hierarchical classification
Score-fusion
Spectral features
Texture features

ABSTRACT

Growing demands from applications like surveillance, archiving, and context-aware devices have fuelled research towards efficient extraction of useful information from environmental sounds. Assigning a textual label to an audio segment based on the general characteristics of locations or situations is dealt with in acoustic scene classification (ASC). Because of the different nature of audio scenes, a single feature-classifier pair may not efficiently discriminate among environments. Also, the acoustic scenes might vary with the problem under investigation. However, for most of the ASC applications, rather than giving explicit scene labels (like home, park, etc.) a general estimate of the type of surroundings (e.g., indoor or outdoor) might be enough. In this paper, we propose a two-level hierarchical framework for ASC wherein finer labels follow coarse classification. At the first level, texture features extracted from time–frequency representation of the audio samples are used to generate the coarse labels. The system then explores combinations of six well-known spectral features, successfully used in different audio processing fields for second level classification to give finer details of the audio scene. The performance of the proposed system is compared with baseline methods using detection and classification of acoustic scenes and events (DCASE, 2016 and 2017) ASC databases, and found to be superior in terms of classification accuracy. Additionally, the proposed hierarchical method provides important intermediate results as coarse labels that may be useful in certain applications.

1. Introduction

Most of the audio processing research in the past considered speech and music signals. The background audio from recordings of real-time speech/music was either discarded in the pre-processing stage or used for environmental noise assessment. However, recently, environmental sound processing has attained popularity because of the importance of information obtained from these sounds in applications like smart devices, robotics, data archiving, surveillance, and hearing aids. Under the broad spectrum of *machine hearing* [1], come fields such as computational auditory scene analysis [2], soundscape cognition [3], sound event detection [4], and acoustic scene classification (ASC) [5]. This paper attempts to contribute to the ASC research. ASC is a supervised classification task of assigning semantic labels to audio streams according to the environments they come from.

1.1. Literature Review

The initial record of research in ASC dates back around two decades. RASTA analysis, power spectral density, pitch estimates and mel-scaled filter-bank coefficients were employed as features accompanied by recurrent neural networks, *k*-nearest neighbor and hidden Markov models as classifiers [6,7]. The popular 'bag-of-frames' (BoF) approach obtained 96% accuracy on a dataset having four audio scene classes [8]. In BoF process, a long-term statistical distribution of a set of short-term spectral features represents an audio stream. Mel frequency cepstral coefficients (MFCCs) are the most commonly used features, while Gaussian mixture models (GMMs) serve as distributions. In a later work, however, it was shown that a simpler one-point average approach was better than the BoF system when evaluated on three other larger audio scene datasets providing less within-class variability [9]. Intermediate representations by higher level features for pre-classification scene-modeling have also been investigated. Non-negative matrix factorization [10], time–frequency (TF) features obtained from matching pursuit algorithm [11], and TF features based on the histogram of gradients [12] are some other strategies explored for environmental audio classification. Some well-known spectral

* Corresponding author.
E-mail addresses: shefaliw@ece.iitkgp.ernet.in (S. Waldekar), gsaha@ece.iitkgp.ernet.in (G. Saha).

https://doi.org/10.1016/j.apacoust.2020.107502
0003-682X/

A MULTI-DEVICE DATASET FOR URBAN ACOUSTIC SCENE CLASSIFICATION

*Annamaria Mesaros, Toni Heittola, Tuomas Virtanen**

Tampere University of Technology, Laboratory of Signal Processing, Tampere, Finland
{annamaria.mesaros, toni.heittola, tuomas.virtanen}@tut.fi

ABSTRACT

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.

Index Terms— Acoustic scene classification, DCASE challenge, public datasets, multi-device data

1. INTRODUCTION

Acoustic Scene Classification is a regular task in the Detection and Classification of Acoustic Scenes and Events (DCASE) challenge series, being present in each of its editions up until now. The standard setup of the task as a basic multiclass classification problem makes the task easily approachable also for the beginner in this field, resulting in large number of participants in the previous DCASE challenges. In the first three editions of the challenge, the acoustic scene classification task has received the highest number of submissions among the available tasks, with 17 submissions in 2013 [1], 48 submissions in 2016 [2], and 97 submissions in 2017 [3].

Each consecutive edition of the challenge has brought a new and larger dataset than previous edition, facilitating use of recent machine learning techniques using deep neural networks that rely on large amounts of data for training. In 2013, the acoustic scene classification task used a development dataset consisting of 10 acoustic scenes each with 10 examples of 30 s, and an evaluation dataset of the same size [1, 4]. In 2016, 15 scene classes were used, each with 78 examples of of 30 s in the development set, and 26 examples per class in the evaluation set [2]. This dataset offered higher acoustic variability than before through its higher number of classes, recording locations and amount of data, and it was the first suitable for use of deep learning methods.

In DCASE 2017, the acoustic scene classification task was made more difficult by using 10 s audio segments, by re-segmenting the complete data available in 2016 (both development and evaluation sets), having 312 segments of 10 s per scene class. A new evaluation dataset was recorded in similar locations approximately one year later than the development data, containing 108 segments of 10 s per class. The temporal gap between the recordings created an unexpected mismatch in acoustic conditions, causing a significant drop in performance in all systems between development and evaluation sets [3]. Outside of DCASE challenge, there are only few other publicly available datasets for acoustic scene classification, notably the LITIS dataset [5], containing 19 classes and having approximately 25 hours of audio, recorded using a mobile phone; the Defreville-Aucouturier environmental audio dataset [6] with 4 main classes (11 detailed classes) and approximately 4 hours of audio; and the UEA Environmental noise datasets [7] with 10 classes and approximately 4 hours of audio in 2 series recorded with different devices. Of these, only the LITIS dataset has an adequate size for modern machine learning methods.

DCASE 2018 challenge introduces a new dataset for acoustic scene classification, having a number of ten classes and 24 hours of high-quality audio. It has smaller number of classes than data from previous challenges, but it is much larger in size and acoustic variability, having been recorded in multiple cities across Europe. This is the largest freely available dataset to date, comparable in size to the LITIS dataset, but it is the only one having recordings in multiple countries, while all other publicly available datasets (within and outside of DCASE) are recorded within a single country or city.

At the same time, parallel recordings performed with different devices provide additional variability in the channel properties, allowing an additional subtask for studying the classification problem in mismatched conditions. All previous public evaluations have been done in matched conditions, with a single device used for recording all data, including evaluation data, but in actual usage scenarios of the methods, channel mismatch could be encountered through device mismatch or difference in recording conditions. Other publicly available datasets contain audio recorded with only one type of device, with small exceptions (e.g. [7]) that do not permit a large-scale study of mismatched devices. A mismatch usually causes a large drop in performance of machine learning based systems, as noticed in DCASE 2017, therefore this new dataset allows development of techniques that can cope with the mismatch.

This paper presents the subtasks and dataset used for Task 1 in DCASE 2018. Section 2 presents the data recording procedure. Section 3 introduces the task definition and specific details on the subtasks, while Section 4 gives details on the experimental setup, including database statistics for each subtask. Section 5 presents the baseline system architecture and the results obtained on the provided experimental setup, and Section 6 presents conclusions and future work.

2. DATA RECORDING PROCEDURE

The TUT Urban Acoustic Scenes 2018 dataset was collected during February-March 2018, containing recordings of ten acoustic scenes, recorded in six large European cities: Barcelona, Helsinki,

*This work has received funding from the European Research Council under the ERC Grant Agreement 637422 EVERYSOUND.

electronics

MDPI

Article

CNN-Based Acoustic Scene Classification System

Yerin Lee †, Soyoung Lim † and Il-Youp Kwak *

Department of Applied Statistics, Chung-Ang University, Seoul 06974, Korea; yerinl2068@gmail.com (Y.L.); isy921234@gmail.com (S.L.)

* Correspondence: ikwak2@cau.ac.kr; Tel.: +82-2-820-5390

† These authors contributed equally to this work.

Abstract: Acoustic scene classification (ASC) categorizes an audio file based on the environment in which it has been recorded. This has long been studied in the detection and classification of acoustic scenes and events (DCASE). This presents the solution to Task 1 of the DCASE 2020 challenge submitted by the Chung-Ang University team. Task 1 addressed two challenges that ASC faces in real-world applications. One is that the audio recorded using different recording devices should be classified in general, and the other is that the model used should have low-complexity. We proposed two models to overcome the aforementioned problems. First, a more general classification model was proposed by combining the harmonic-percussive source separation (HPSS) and deltas-deltadeltas features with four different models. Second, using the same feature, depthwise separable convolution was applied to the Convolutional layer to develop a low-complexity model. Moreover, using gradient-weight class activation mapping (Grad-CAM), we investigated what part of the feature our model sees and identifies. Our proposed system ranked 9th and 7th in the competition for these two subtasks, respectively.

Keywords: convolutional neural network; voice classification; ResNet; ensemble

Citation: Lee, Y.; Lim, S.; Kwak, I.-Y. CNN Based Acoustic Scene Classification System. *Electronics* **2021**, *10*, 371. https://doi.org/10.3390/electronics10040371

Academic Editor: Osvaldo Gervasi

Received: 12 January 2021

Accepted: 30 January 2021

Published: 3 February 2021

1. Introduction

In recent years, acoustic scene classification (ASC) has attracted widespread attention in the Audio and Acoustic Signal Processing (AASP) community [1–6]. ASC aims to classify a test recording sound into predefined classes that characterizes the environment in which it was recorded [7]. The IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE) takes place every year. It started from 2013 and is continuing every year since 2016. The current year's ASC task is divided into two subtasks: A and B.

Figure 1a shows the problem overview on subtask A. Subtask A aims to classify audio into ten classes. Related audio files are recorded or simulated using multiple devices. The development dataset includes 40 h of data from device A and smaller amounts of data from other devices. The audio is provided in a single-channel 44.1 kHz 24-bit format. Figure 1b shows the problem overview on subtask B.

Subtask B aims to classify audio into three classes based on low-complexity solutions. All participants were required to comply with the model size of 500 KB or less. The related dataset contains data recorded by a single device (device A). The audio is provided in a binaural 48 kHz 24-bit format.

Convolutional neural networks (CNNs) are deep neural networks that are commonly used for visual image analysis. CNN is applied to computer vision, natural language processing, and recommendation systems. Moreover, it can be applied to audio by transforming the audio of the waveform into an image. For audio, CNN can be used pertaining to the pre-processing of raw data, such as short time Fourier transform (STFT), constant-Q transform (CQT), and mel-frequency cepstral coefficient (MFCC). The top-performing studies in the recent ASC competitions used CNNs. Han and Park [8] learned deep learning models for left-right (LR), mid-side (MS), and harmonic-percussive source seperation

Active Few-Shot Learning for Sound Event Detection

Yu Wang[1], *Mark Cartwright*[2], *Juan Pablo Bello*[1]

[1]Music and Audio Research Laboratory, New York University, NY, USA
[2]New Jersey Institute of Technology, Newark, NJ, USA

wangyu@nyu.edu, mark.cartwright@njit.edu, jpbello@nyu.edu

Abstract

Few-shot learning has shown promising results in sound event detection where the model can learn to recognize novel classes assuming a few labeled examples (typically five) are available at inference time. Most research studies simulate this process by sampling support examples randomly and uniformly from all test data with the target class label. However, in many real-world scenarios, users might not even have five examples at hand or these examples may be from a limited context and not representative, resulting in model performance lower than expected. In this work, we relax these assumptions, and to recover model performance, we propose to use active learning techniques to efficiently sample additional informative support examples at inference time. We developed a novel dataset simulating the long-term temporal characteristics of sound events in real-world environmental soundscapes. Then we ran a series of experiments with this dataset to explore the modeling and sampling choices that arise when combining few-shot learning and active learning, including different training schemes, sampling strategies, models, and temporal windows in sampling.

Index Terms: sound event detection, few-shot learning, active learning

1. Introduction

Few-shot learning [1–5] has recently been proposed for sound event detection [6] and shown promising results, where a model is trained to learn to recognize novel sound classes, unseen during training, given only very few examples from each new class at inference time. It has been applied to tasks in various audio domains, including speech, music, and environmental sound, tackling the labeled data scarcity issue by incorporating minimal human input [7–13].

One of the main assumptions of few-shot learning is that human users can provide a few examples (e.g. five) of the target novel class, which we call the support set, at inference time. In research studies, this process is typically simulated by sampling the support set randomly and uniformly from all available test data with the target class label. However, this assumption and simulation might not reflect real-world sound event detection scenarios. Often, obtaining a few representative audio examples is not as straightforward as one might expect. For example, in the case of environmental sound monitoring, when a user encounters a new sound class that they want the model to learn to recognize, they might not be able to obtain five examples right away and/or the examples they can obtain may be from a limited context and not representative. Similarly, in the case of automatic drum transcription, obtaining five representative examples may be difficult if the target drum class is sparse or highly varying within the song. Therefore, model performance in these real-world scenarios may be lower than the model performance reported in research papers.

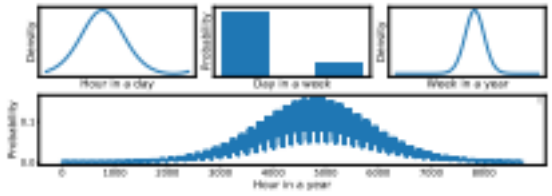


Figure 1: *Simulated occurrence probabilities of a foreground sound class in the SONYC-FSD-SED dataset.*

In this work, we show that while model performance drops when we relax these assumptions to reflect real-world scenarios, we can not only recover but supersede prior model performance by combining few-shot learning with active learning, where the model actively queries human users for labels of the most informative unlabeled data that improves model performance [14]. In prior work, active learning has been applied to various audio classification tasks and shown to be effective in improving training efficiency and reducing annotation effort [15–21]. By introducing of active learning into few-shot learning, we aim to efficiently build a better and more representative support set.

To understand how to effectively combine few-shot learning and active learning, we developed a novel dataset SONYC-FSD-SED, which will be freely available online, and designed a series of experiments using this dataset to explore the modeling and sampling choices that arise when combining these techniques. We pursue this work in the context of environmental sound monitoring. SONYC-FSD-SED simulates the long-term temporal characteristics of sound events in a real-world environmental sound monitoring system with ground-truth labels. It reflects the seasonal periodic patterns of the occurrences and co-occurrences of sound classes. In the experiments, we first explore different training schemes for the few-shot model. While a standard few-shot model is typically trained and tested with a fixed number of random support examples (i.e., fixed number of shots), during the active sampling process, the number of support examples is constantly changing. We propose new training schemes to better match the diversity and the varied number of support examples during these inference time scenarios. Second, we experiment with different few-shot models to see how they interact with active learning and affect sampling. Lastly, we study how different temporal windows in sampling affect model performance and generalization, mimicking the scenario where users only have access to data with limited variability and diversity. While in this work we perform experiments in the context of environmental sound monitoring, we expect our experimental design and findings should be generally applicable to sound event detection tasks in other audio domains such as music and speech.

This work was partially supported by National Science Foundation award 1544753 & 1955357.

A Convolutional Neural Network Approach for Acoustic Scene Classification

Michele Valenti, Stefano Squartini
Università Politecnica delle Marche
Department of Information Engineering
Ancona, Italy
Email: valenti.michele.w@gmail.com

Aleksandr Diment, Giambattista Parascandolo
and Tuomas Virtanen
Tampere University of Technology
Department of Signal Processing
Tampere, Finland

***Abstract*—This paper presents a novel application of convolutional neural networks (CNNs) for the task of acoustic scene classification (ASC). We here propose the use of a CNN trained to classify short sequences of audio, represented by their log-mel spectrogram. We also introduce a training method that can be used under particular circumstances in order to make full use of small datasets. The proposed system is tested and evaluated on three different ASC datasets and compared to other state-of-the-art systems which competed in the "Detection and Classification of Acoustic Scenes and Events" (DCASE) challenges held in 2016[1] and 2013. The best accuracy scores obtained by our system on the DCASE 2016 datasets are 79.0% (development) and 86.2% (evaluation), which constitute a 6.4% and 9% improvements with respect to the baseline system. Finally, when tested on the DCASE 2013 evaluation dataset, the proposed system manages to reach a 77.0% accuracy, improving by 1% the challenge winner's score.**

I. Introduction

When we talk of acoustic scene classification (ASC) we refer to the ability of a human or an artificial system to recognize an *audio context*, either from an on-line stream or from a recording. "Context" or "scene" are concepts that humans commonly use to identify a particular *acoustic environment*, *i.e.*, the ensemble of background noises and sound events that we are used to associate with a specific audio scenario, like a restaurant or a park. For humans this may look like a simple task: complex calculations that our brain is able to perform and our extensive life experiences allow us to easily associate these ensembles of sounds with specific scenes. However, this task is not trivial for artificial systems.

Falling in the field of studies known as *computational auditory scene analysis* (CASA), ASC was firslty defined in Bregman's work [2], in 1994. Ten years later, a collection of important theoretical papers [3] was published with the purpose of deeply analyzing CASA, especially by addressing issues such as the "mixture problem" [4] (*i.e.*, the problem of recovering an individual source from an ensemble of overlapping sounds). Practical interest in computational ASC lies in its many possible applications, therefore some applicative analyses were published throughout the years. For example, in [5] the authors aim at defining a "context-aware" computation paradigm, in particular by examining how a computer can became able to react and reconfigure itself by smartly monitoring the nearby environment. Sequent works also analyzed the use of ASC applied to intelligent wearable interfaces configuration [6] or mobile robot navigation enhancement [7].

In the field of machine learning different models and audio feature representations have been proposed to deal with the ASC task. In particular, examples of classifiers featuring neural networks go back to 1997 [8] and 1998 [9]. In [8] a recurrent neural network is combined with a nearest neighbour classifier to discriminate five different environmental sounds, whereas in [9] a neural network is used to recognize five different types of TV shows. Despite the fact that new systems have been proposed throughout the years, the interest towards ASC has raised especially thanks to the DCASE 2013 challenge [10]. In particular, 18 different systems were evaluated for the DCASE 2013 ASC task, some examples being based on Gaussian mixture models (GMMs) [11], support vector machines [12] and tree bagger classifiers [13].

Nowadays, application of CNNs for audio-related tasks is becoming more and more widespread. In [14] a speech recognition task is addressed and a CNN approach is evaluated on two different datasets. In addition, other examples highlight how CNNs can reach a very high performance also on tasks such as environmental sound classification [15] or robust audio event recognition [16]. To the best of our knowledge no former works addressing ASC with CNNs were presented before the DCASE 2016 challenge, for the occasion of which this study was realized.

The system proposed in this paper has been specifically studied for ASC. We designed a CNN able to work with short audio sequences, characterized by their log-mel spectrogram features. Each input spectrogram is classified by the CNN as belonging to one of 15 different acoustic scenes, as provided by the ASC task setup. Moreover, at training time we make use of batch normalization, a recently-proposed technique developed to speed up and regularize the network training. In addition, we propose a training procedure that allows the classifier to achieve a good generalization performance on both the development and evaluation datasets. Therefore, we here propose a system capable of improving the baseline

[1] A copyright-free and preliminary version of this paper [1] has been presented at DCASE 2016 workshop, Budapest, Hungary.

978-1-5090-6182-2/17/$31.00 ©2017 IEEE

Applied Acoustics 195 (2022) 108819

Contents lists available at ScienceDirect

Applied Acoustics

journal homepage: www.elsevier.com/locate/apacoust

Adaptive time-frequency feature resolution network for acoustic scene classification

Tao Zhang, Jinhua Liang *, Guoqing Feng

School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China

ARTICLE INFO

Article history:
Received 13 March 2021
Received in revised form 23 March 2022
Accepted 11 May 2022
Available online 27 May 2022

Keywords:
Acoustic scene classification
Convolutional neural network
Adaptive time-frequency feature resolution network

ABSTRACT

Convolutional neural networks with time-frequency feature representation for acoustic scene classification have been attracting increasing attentions. However, most of the existing methods are restricted to design a plain discriminative model with low-level feature representations and hard to leverage a priori knowledge of sound events in a particular scene. This provides some leeway to boost the performance. In this paper, Adaptive time-frequency Feature Resolution Network (AdaFRN) is proposed to embrace the progress in attention mechanism, feature fusion, and very deep architecture. Specifically, a multi-resolution attention distributor is applied to automatically select a proper feature resolution to capture remarkable events and a bi-direction feature pyramid network is utilized to transmit information across multiple scales. Extensive experiments demonstrate that the proposed model yields better results than state-of-the-art methods. With the normalized mixup strategy, our model outperforms the state-of-the-art method by 1.6%.

1. Introduction

Acoustic Scene Classification (ASC) aims to enable devices to recognize an audio environment perceived and defined by human from either a recording or an online stream. In recent years, it has been gaining increasing interests from computational acoustics community, and has been widely investigated [2,20]. Early works on ASC problem applied feature engineering [32] combined with machine learning methods, such as Gaussian Mixture Model (GMM) [14,28], Support Vector Machine (SVM) [13,25], unsupervised learning [6,11,12], Multi-Layer Perception (MLP) [35], and Recurrent Neural Network (RNN) [33]. Despite their fair discrimination accuracy, it turns out that hand-craft features designed for a specific task are hard to be generalized to other tasks.

To take the advantage of deep learning and large-scale datasets, several approaches based on Time-Frequency (T-F) spectrograms have been recently developed to learn feature representations using Convolutional Neural Network (CNN). Mesaros, Heittola and Virtanen [24] proposed a CNN-based system as a Baseline on DCASE2018 and DCASE2019, including convolutional filters, Rectifier Linear Units (ReLU) [19], and batch normalization [15]. Kong, Iqbal, Xu, Wang and Plumbley [17] proposed an eight-layer stacked convolutional model, replacing the large convolutional filter with one sized 3×3. Subsequent works [4,31,36] applied deep CNN constructions to the ASC problem.

In our previous study [38], the T-F feature resolution was proposed to model the relationship between T-F spectrograms and CNN receptive fields. Then a novel network was proposed to capture short-time remarkable characteristics and to leverage the semantic features generated by deep CNNs. Since there are dramatic distinctions between images and acoustic characters, it's important to design network structures for audio-related tasks specifically. In our opinion, the unsolved issues on CNN-based models are mainly in three folds:

1) *Isolation of "scene" from "events".* An acoustic scene is referred as a mixture of background noise and various sound events associated with a specific audio scenario. However, most of the existing methods are focus on the "bag of frame" approaches using low-level features followed by a classifier, which go against the demands of interpretability and robustness.
2) *Diverse acoustic characteristics.* In most cases, a specific scene contains different sound events corresponding to distinctive patterns. The designed feature extractor is thus supposed to implicitly include all these representations.

* Corresponding author.
E-mail addresses: zhangtao@tju.edu.cn (T. Zhang), tjljh@tju.edu.cn (J. Liang), guoqing_2019@tju.edu.cn (G. Feng).

https://doi.org/10.1016/j.apacoust.2022.108819
0003-682X/

Multimedia Tools and Applications (2023) 82:16395–16408
https://doi.org/10.1007/s11042-022-14192-1

Acoustic scene classification with multi-temporal complex modulation spectrogram features and a convolutional LSTM network

Sayeh Mirzaei [1] · Iman Khani Jazani [2]

Received: 13 August 2021 / Revised: 14 September 2022 / Accepted: 27 October 2022 /
Published online: 11 November 2022

Abstract

Acoustic scene classification (ASC) is a mapping from an environmental sound recording to predefined classes representing the auditory scene of the recording. This paper proposes an ASC solution based on the combination of convolutional neural networks, long short term memory cells, and multi-temporal input encoding. The major novelty of the work is applying complex modulation spectrogram for feature extraction. We evaluate the complex modulation spectrogram as discriminant features, resulting in a 4.7% improvement in comparison with the commonly used Mel spectrogram. These features are computed for individual temporal segments of the audio recording to acquire a representation containing both spectral and temporal structure. Also, we derive a de-noising method which has not been used for ASC before but was beneficial in other speech processing tasks. This method leads to 1.5% improvement in prediction accuracy in comparison with a model without de-noising. The proposed model outperforms the state of the art methods by 7.5% in terms of the prediction accuracy for evaluation data in ASC on the DCASE 2017 dataset.

Keywords Acoustic scene classification · Convolutional neural network (CNN) · Long short term memory (LSTM) · Conv-LSTM · Modulation spectrogram

1 Introduction

Understanding the environment and having a suitable perception in it is a major concern in artificial intelligence. Environmental sound perception can play an important role to accomplish situational awareness of machines. Classifying the (complex) acoustic environment is

✉ Sayeh Mirzaei
s.mirzaei@ut.ac.ir

[1] School of Engineering Science, College of Engineering, University of Tehran, Tehran, Iran

[2] Faculty of Computer Engineering, Amirkabir University of Technology, Tehran, Iran

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2018.2883628, IEEE/ACM Transactions on Audio, Speech, and Language Processing

Spectro-Temporal Analysis using Local Binary Pattern Variants for Acoustic Scene Classification

Shamsiah Abidin, *Student Member, IEEE,* Roberto Togneri, *Senior Member, IEEE,* and Ferdous Sohel, *Senior Member, IEEE*

***Abstract*—In this paper we present an approach for acoustic scene classification, which aggregates spectral and temporal features. We do this by proposing the first use of the variable-Q transform (VQT) to generate the time-frequency representation for acoustic scene classification. The VQT provides finer control over the resolution compared to the constant-Q transform (CQT) or STFT and can be tuned to better capture acoustic scene information. We then adopt a variant of the local binary pattern (LBP), the Adjacent Evaluation Completed LBP (AECLBP), which is better suited to extracting features from acoustic time-frequency images. Our results yield a 5.2% improvement on the DCASE 2016 dataset compared to the application of standard CQT with LBP. Fusing our proposed AECLBP with HOG features we achieve a classification accuracy of 85.5% which outperforms one of the top performing systems.**

***Index Terms*—acoustic scene, local binary patterns, feature extraction, time-frequency analysis, fusion**

I. Introduction

THE research on acoustic scene classification has been of interest to researchers in the area of acoustic analysis for the past two decades. Acoustic scene analysis has been used in applications such as automatic audio surveillance, mobile phone sensing, context-aware assistive robots, music genre classification and multimedia archiving. Audio surveillance is one of the applications that typically employs sound content analysis techniques to detect outlier activities such as gunshot and screaming in a specific indoor environment [1], [2]. Furthermore, in order to achieve the environment awareness, a mobile phone application is expected to be able to identify and automatically adapt to the surrounding environments [3], [4]. The objective of acoustic scene classification (ASC) is to identify the environment in which an audio stream has been produced [5]. The DCASE 2013 Challenge was introduced by the IEEE AASP Technical Committee to provide an evaluation and comparison of different techniques developed in acoustic scene classification on a benchmark dataset. Intended for inspiring the development of novel methodologies and improving the state-of-the-art in ASC, the DCASE 2016 Challenge dataset for audio scene classification was released with more challenging data.

Time-frequency representations (TFR) of discrete-time signals play an important role in acoustic analysis. A TFR provides a visual representation of the temporal and spectral structures that can be viewed as a 2D texture image. The constant-Q transform (CQT), commonly used for music processing tasks [6], [7], has now been applied to acoustic scene analysis [8], [9]. The use of CQT in combination with feature learning approaches based on nonnegative matrix factorization (NMF) by [9] has achieved a classification accuracy of 83.8% which is the state-of-the-art non-neural network based system on the DCASE 2016 dataset. However, the CQT lacks flexibility as the Q-factor is constant throughout the frequency band analysis. Indeed, a variable-Q factor is necessary to retain the important spectral and temporal structure of the acoustic signal. A Q-factor which is adaptable to the acoustic signal is required to produce a more accurate TFR representation.

The TFR texture image features can be extracted by well-known feature extraction methods used in computer vision. The Local Binary Pattern (LBP) is a state-of-the-art feature extraction method for analyzing image textures due to its computational simplicity [10]. A number of different variants of LBP have been developed to improve its robustness, and to increase its discriminative capability and applicability to different types of problems in image-classification applications. The Adjacent Evaluation LBP (AELBP) [11] is introduced to improve noise robustness of LBP by introducing the adjacent evaluation window and modifying the threshold scheme of LBP. It can be used with existing LBP variants such as the Completed Local Binary Pattern (CLBP) [12], the Completed Local Binary Count (CLBC) [13] and the Local Ternary Pattern (LTP) [14] to derive new image features against noise for texture classification. The CLBP feature extraction decomposes the image's local structure into two complementary components: the difference signs and the difference magnitudes to provide more discriminative information.

Motivated by the advancement of LBP variants, we are inspired to identify a variant of LBP that is adaptable to the acoustic signal representation and better suited for ASC. The micro structure of the image texture is different from the TFR. The image texture might have rotation and illumination variations whereas in the case of TFR there is no illumination or rotation variations. Also, in the case of TFR,

S. Abidin and R. Togneri are with the School of Electrical, Electronics and Computer Engineering, The University of Western Australia, Perth, WA, 6009 Australia e-mail: (shamsiah.abidin@research.uwa.edu.au; roberto.togneri@uwa.edu.au).

F. Sohel is with School of Engineering and Information Technology, Murdoch University, Perth 6150, WA, Australia e-mail: (f.sohel@murdoch.edu.au).

Manuscript received month date year; revised month date year

Applied Acoustics 163 (2020) 107238

Contents lists available at ScienceDirect

Applied Acoustics

journal homepage: www.elsevier.com/locate/apacoust

Multi-scale semantic feature fusion and data augmentation for acoustic scene classification

Liping Yang [a,*], Lianjie Tao [a], Xinxing Chen [a], Xiaohua Gu [b]

[a] Key Laboratory of Optoelectronic Technology and Systems, MOE, Chongqing University, Chongqing 400044, China
[b] School of Electrical Engineering, Chongqing University of Science and Technology, Chongqing 401331, China

ARTICLE INFO

Article history:
Received 17 November 2019
Received in revised form 23 January 2020
Accepted 29 January 2020
Available online 8 February 2020

Keywords:
Multi-scale feature learning
Convolutional neural networks
Data augmentation
Acoustic scene classification (ASC)
Machine listening

ABSTRACT

This paper investigates a multi-scale semantic feature fusion and data augmentation approach for deep convolutional neural network (CNN) based acoustic scene classification. To ensemble the multi-scale semantic information of CNN and improve the performance of acoustic scene classification, a multi-scale feature fusion framework, which consists of a simplified Xception backbone and a semantic feature fusion strategy, is presented. A novel label smoothing mixup data augmentation method, which is a generalization of mixup and label smoothing, is proposed to alleviate the over-confident problem of network training. A *spatial-mixup* technique is presented to generate meaningful mixup virtual data for acoustic scene classification. Extensive experiments on synthetic data and real acoustic scene classification dataset demonstrate that both multi-scale semantic feature fusion and label smoothing spatial-mixup data augmentation are effective for improving the acoustic scene classification performance of a deep neural network.

1. Introduction

Enabling devices to make sense of their environment through the analysis of sounds is an active research field in machine listening [1]. General purpose machine listening tasks focus on everyday sounds, such as general environment sounds and rare event sounds, beyond speech and music [2]. Acoustic scene classification (ASC) aims at attributing a semantic label to acoustic scenes, often corresponding to some geographical contexts, such as *at a metro station*, *at an airport* or *in a shopping mall* [3]. The research of acoustic scene classification has received increasing attention from the research community in recent years. Its applications range from context-aware services, robotic navigation systems, intelligent wearable devices, audio-based video analysis and elder assistant systems to surveillance systems [4–6].

In general, machine listening/hearing systems consist of four concatenate modules, including ear-like front-end auditory analysis, feature extraction, feature size reduction, and followed by a decision module [1]. As a challenge machine listening technique, the framework of acoustic scene classification involves obtaining a reasonable representation of data and then employing these features for classification [4,7,8]. Different from speech and music where stationary and clear harmonic structure inside an analysis frame can be assumed, the characteristics of environmental sounds are diverse. Due to the intrinsic variability of a given scene sound, researchers understand that there might be a natural mismatch for acoustic scene classification [9]. Therefore, new robustness feature representation and classification approaches for acoustic scene classification need development.

To the best of our knowledge, the pioneered research on acoustic scene classification can trace back to two decades ago by Sawhney and Maes [10]. They extracted several features from audio data using methods borrowed from automatic speech analysis, employed a recurrent neural network (RNN) and a *k* nearest neighbor (KNN) criterion for classification. From that epoch, the studies of ASC mainly concentrated on spectral features, which widely adapted from speech analysis. The handcrafted Mel-frequency Cepstral Coefficients (MFCCs) features combined with Gaussian Mixture Models (GMMs) had been found to perform well [11–13].

To date, a variety of signal processing and machine learning techniques have been applied on time–frequency representations (TFR) of acoustic signals for acoustic scene classification, including matrix factorization [7,9], texture description [14,4,15], and most recently deep neural networks [16–19]. The non-negative matrix factorization (NMF) based acoustic scene classification approaches and their supervised versions proposed in [7,9] present the state of the art performance in the field of matrix factorization. In [9], the

* Corresponding author.
E-mail address: yanglp@cqu.edu.cn (L. Yang).

https://doi.org/10.1016/j.apacoust.2020.107238
0003-682X/

Expert Systems With Applications 126 (2019) 20–29

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Investigation of acoustic and visual features for acoustic scene classification

Jie Xie[a,*], Mingying Zhu[b]

[a] Key Laboratory of Advanced Process Control for Light Industry (Ministry of Education), School of Internet of Things Engineering, Jiangnan University, Wuxi 214122, PR China

[b] Department of Economics, University of Ottawa, Ontario K1N6N5, Canada

ARTICLE INFO

Article history:
Received 16 June 2018
Revised 20 January 2019
Accepted 21 January 2019
Available online 13 February 2019

Keywords:
Acoustic scene classification
Acoustic features
Visual features
ReliefF algorithm
Support vector machine

ABSTRACT

Acoustic scene classification has gained great interests in recent years due to its diverse applications. Various acoustic and visual features have been proposed and evaluated. However, few studies have investigated acoustic and visual feature aggregation for acoustic scene classification. In this paper, we investigated various feature sets based on the fusion of acoustic and visual features. Specifically, acoustic features are directly extracted from the waveform: spectral centroid, spectral entropy, spectral flux, spectral roll-off, short-time energy, zero-crossing rate, and Mel-frequency Cepstral coefficients. For visual features, we calculate local binary pattern, histogram of gradients, and moments based on the audio scene time-frequency representation. Then, three feature selection algorithms are applied to various feature sets to reduce feature dimensionality: correlation-based feature selection, principal component analysis, and ReliefF. Experimental results show that our proposed system was able to achieve an accuracy improvement of 15.43% compared to the baseline system with the development set. When all development sets are used for training, the performance based on the evaluation set provided by the TUT Acoustic scene 2016 challenge is 87.44%, which is the fourth best among all non-neural network systems.

1. Introduction

Acoustic scene classification has gained great interests in recent years due to its diverse applications: automatic audio surveillance (Atrey, Maddage, & Kankanhalli, 2006), robotics sensing (Maxime, Alameda-Pineda, Girin, & Horaud, 2014), multimedia content analysis (Wang, Liu, & Huang, 2000) and machine listening (Barchiesi, Giannoulis, Stowell, & Plumbley, 2015). In addition to images, audio information often can be a good supplement for solving many real world problems. Taking audio surveillance for example, acoustic scene classification can keep the surveillance system running during the dark hours.

Different scenes have their discriminative physical activities of humans and environments, which often generate unique audio characteristics (Eronen et al., 2006). Therefore, it is possible to build an acoustic scene classification system. Many researchers have proposed various acoustic scene classification systems to recognize those activities, which are commonly structured as follows: (1) preprocessing (2) feature extraction, (3) feature selection and aggregation, (4) classification. Among those four steps, choosing discriminating and independent feature sets plays an important role in the success of one classification system.

Many features have been investigated for acoustic scene classification. Those features can be grouped into two types: acoustic and visual features. For acoustic features, Mel-frequency Cepstral coefficients (MFCCs) are often used as the baseline for new datasets (Mesaros, Heittola, & Virtanen, 2016; Piczak, 2015). Other features such as spectral centroid, spectral flux, zero-crossing rate have also been widely explored in previous studies (Kiktova-Vozarikova, Juhar, & Cizmar, 2015; Piczak, 2015; Yang, Krishnan, Yang, & Krishnan, 2017). All those features are calculated by applying Fourier transform to the one-dimensional audio data (waveform). However, transforming audio data into its two-dimensional representation (spectrogram) can provide another way to describe the sound, where visual features can be calculated. Using various time-frequency representations, many visual features have been used to classify acoustic scene, including local binary pattern (LBP), histogram of orientations (HOG) (Rakotomamonjy & Gasso, 2015; Yang et al., 2017).

Since acoustic and visual features are calculated from waveform and spectrogram, respectively, aggregating acoustic features with visual features provides a chance to increase the discrimination of feature sets. Previous studies have demonstrated the usefulness of

* Corresponding author.
E-mail addresses: xiej8734@gmail.com (J. Xie), mzhu089@uottawa.ca (M. Zhu).

https://doi.org/10.1016/j.eswa.2019.01.085
0957-4174/

Applied Acoustics 212 (2023) 109594

Contents lists available at ScienceDirect

Applied Acoustics

journal homepage: www.elsevier.com/locate/apacoust

Hierarchical classification for acoustic scenes using deep learning

Biyun Ding, Tao Zhang *, Ganjun Liu, Chao Wang

School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China

ARTICLE INFO

Index terms:
Acoustic scene classification
Convolutional neural network
Data augmentation
Hierarchical classification
Late fusion

ABSTRACT

Acoustic Scene Classification (ASC) aims to obtain the sound environment by analyzing audio signals. Due to the low complexity and acquisition cost of audio signals, ASC has enormous potential in various applications, such as audio-based surveillance, smart cities/homes, and robotics. Recently, various methods have been proposed for ASC and achieved good performance. However, when they are used to address complex ASC problems, most of them suffer from the low-performance problem. In this paper, we propose to use hierarchical classification methods to replace the conventional flat approach in ASC applications, which utilizes the class hierarchy to optimize classification performance. In particular, we investigate the ASC problem under the framework of hierarchical classification. Firstly, to improve classification performance, three hierarchical classification methods introducing the class hierarchy of acoustic scenes are proposed for ASC. Moreover, to fully utilize the class hierarchy, a hybrid hierarchical classification method, and an optimal late fusion-based hierarchical method are proposed, which are based on the flexibility and simplification of hierarchical classification. The experiments demonstrate the efficacy of hierarchical ASC systems for performance improvement, and the best system achieves an accuracy of 78.86% on the DCASE 2020 Task1A dataset, resulting in accuracy gains of 24.76% and 8.52% absolute over the DCASE 2020 Task 1A baseline and the conventional non-hierarchical method, respectively.

1. Introduction

Environmental sounds can convey semantic information based on complex acoustic structures. Therefore, environmental sounds are gaining prominence in theoretical and applied research that crosses the boundaries of different fields [1]. Acoustic Scene Classification (ASC) identifies the scene by analyzing audio signals recorded in a specific environment. Various potential applications promoted those studies, such as automatic optimization of audio devices (e.g., hearing aids), audio-based surveillance [2], and intelligent environment context recognition in robotics [3–4].

During the last decade, ASC has been an active research field. Several important challenges have made significant progress for ASC, such as the Detection and Classification of Acoustic Scenes and Events (DCASE) [5–6]. It started in 2013 and has continued annually since 2016 to stimulate research in the ASC field. As a result, many available datasets, algorithms, publications, and open-source code libraries have emerged. The accuracy of ASC systems has been essentially improved. For example, the classification accuracies on ESC-50 [7], Urbansound8K [8], and LITIS datasets [9] have been significantly improved to over 95%.

However, the classification performance is decreased with the increasing task complexity of ASC problems, such as DCASE 2019–2022 ASC tasks. Therefore, the ASC field is still in the early research stage, suffering from the low-performance problem. Currently, various methods have been proposed in terms of data processing (e.g., data augmentation [10–13]), feature extraction [14], modeling [15–18] (e.g., transfer learning [19–20], attention mechanisms [21–22], and multitask learning [22–24]), and results post-processing methods [25] to improve classification accuracy.

ASC assigns a predefined semantic label to an audio stream to characterize its sound environment. Traditional ASC algorithms are based on a set of mutually exclusive classes. However, a hierarchical structure among scene classes exists in many real-world data. For instance, the research on the taxonomy of environmental sounds shows the presence of hierarchical categorizing of information in sound classes [26], such as nature/human, indoor/outdoor, and traffic/human activity. In addition, people's perception, understanding, interpretation, and representation are also structured and hierarchical for the real world. It means traditional ASC algorithms ignore class hierarchy in real-world data, which is inconsistent with human perception patterns.

Therefore, as important auxiliary information, the class hierarchy

* Corresponding author.
E-mail address: zhangtao@tju.edu.cn (T. Zhang).

https://doi.org/10.1016/j.apacoust.2023.109594
Received 20 June 2023; Received in revised form 7 August 2023; Accepted 12 August 2023
Available online 1 September 2023

ENHANCING SOUND TEXTURE IN CNN-BASED ACOUSTIC SCENE CLASSIFICATION

Yuzhong Wu, Tan Lee

Department of Electronic Engineering, The Chinese University of Hong Kong, Hong Kong SAR, China

ABSTRACT

Acoustic scene classification is the task of identifying the scene from which the audio signal is recorded. Convolutional neural network (CNN) models are widely adopted with proven successes in acoustic scene classification. However, there is little insight on how an audio scene is perceived in CNN, as what have been demonstrated in image recognition research. In the present study, the Class Activation Mapping (CAM) is utilized to analyze how the log-magnitude Mel-scale filter-bank (log-Mel) features of different acoustic scenes are learned in a CNN classifier. It is noted that distinct high-energy time-frequency components of audio signals generally do not correspond to strong activation on CAM, while the background sound texture are well learned in CNN. In order to make the sound texture more salient, we propose to apply the Difference of Gaussian (DoG) and Sobel operator to process the log-Mel features and enhance edge information of the time-frequency image. Experimental results on the DCASE 2017 ASC challenge show that using edge enhanced log-Mel images as input feature of CNN significantly improves the performance of audio scene classification.

Index Terms— Convolutional neural network, acoustic scene classification, sound texture, class activation map, edge enhancement

1. INTRODUCTION

Large amount of multimedia information becomes easily accessible nowadays. The performance of speech and image recognition systems has been significantly improved with the use of deep neural networks and exploding amount of training data. Audio-related tasks, e.g., Acoustic Scene Classification (ASC) [1, 2, 3], Sound Event Detection (SED) [4, 5, 6] and Audio Tagging [7, 8, 9, 10], have also received increasing attention in recent years. They have many real-world applications. For example, context-aware mobile devices could provide better responses to their users in accordance with the acoustic scene. A smart home-monitoring system could detect unusual incidences by using audio. An audio search engine is able to retrieve information efficiently from massive online recordings.

Acoustic scene classification (ASC) is the process of identifying the type of acoustic environment (scene) where a given audio signal was recorded. It has been a major task in the IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE) since 2013. In the 2017 ASC challenge, most of the best-performing models were based on convolutional neural networks (CNN). Mun et al. [11] addressed the problem of data insufficiency and proposed to use the Generative Adversarial Network (GAN) [12] to augment training data. Han et al. [13] was focused on preprocessing of input features. Fusion of CNN models with preprocessed input features led to improved overall model performance.

Despite the clearly demonstrated effectiveness of CNN-based models in the ASC task, there is little insight on how an audio scene is perceived in a CNN model. Whilst similar issue has been extensively explored in image classification. In [14], Zeiler & Fergus used the De-convolutional Network [15] to visualize and understand CNN. Springenberg et al. applied the guided backpropagation [16] to obtain sharp visualization of descriptive image regions. The Class Activation Mapping (CAM) [17] was proposed as a means of highlighting the discriminative image regions for specific output classes in CNNs with global average pooling. Selvaraju et al. developed a generalized version of CAM, named the Gradient-weighted Class Activation Mapping (Grad-CAM) [18], which could be applied to a broader range of CNN models.

The input of an audio classification model is usually a time-frequency representation extracted from the raw audio waveform. Among the various types of time-frequency representations, the logarithmic-magnitude Mel-scale filter bank (abbreviated as log-Mel) feature is widely adopted. Similar to spectrogram, a log-Mel feature is a visual representation of the frequency content of sounds as they vary with time. Given an audio signal with audible sound events such as "bird singing", "speech", "applause", these sound events can also be identified in the corresponding log-Mel feature based on their distinct visual patterns. From this perspective, we may call a log-Mel feature as an image. Visualization of CAM using the log-Mel "image" allows the comparison between machine perception and human interpretation.

In this paper, we present an attempt to understand how CNN models learn to identify an acoustic scene from log-Mel feature representations. The investigation starts with benchmark systems with log-Mel features and different CNN models. The method of CAM is used to provide visualization of the CNN activation behavior with respect to input features. The observed CAMs for acoustic scene data suggest that CNN classification models tend to emphasize on the overall background sound texture of log-Mel input features, whilst individual sound events in the scene are of less importance. Hence we propose to use the Difference of Gaussian (DoG) and the Sobel operator to pre-process the log-Mel feature to make the background texture information more salient. We also use the method of background drift removing with medium filter as described in [13] as a comparison to our methods. These texture-enhanced features demonstrate an improved performance on ASC.

2. BACKGROUND

2.1. Class Activation Mapping

The class activation mapping [17] highlights the class-specific discriminative regions in the input image. It can help understand the CNN behavior and visualize the internal representation of CNNs. It can also be used for weakly supervised object localization task. However, the CAM is only applicable to CNNs with global average pooling (GAP). Suppose we have a trained CNN network with global average pooling. There are C output classes. The number of

Deep Scalogram Representations for Acoustic Scene Classification

Zhao Ren, Kun Qian, *Student Member, IEEE,* Zixing Zhang, *Member, IEEE,* Vedhas Pandit, Alice Baird, *Student Member, IEEE,* and Björn Schuller, *Fellow, IEEE*

***Abstract*—Spectrogram representations of acoustic scenes have achieved competitive performance for acoustic scene classification. Yet, the spectrogram alone does not take into account a substantial amount of time-frequency information. In this study, we present an approach for exploring the benefits of deep scalogram representations, extracted in segments from an audio stream. The approach presented firstly transforms the segmented acoustic scenes into bump and morse scalograms, as well as spectrograms; secondly, the spectrograms or scalograms are sent into pre-trained convolutional neural networks; thirdly, the features extracted from a subsequent fully connected layer are fed into (bidirectional) gated recurrent neural networks, which are followed by a single highway layer and a softmax layer; finally, predictions from these three systems are fused by a margin sampling value strategy. We then evaluate the proposed approach using the acoustic scene classification data set of 2017 IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE). On the evaluation set, an accuracy of 64.0% from bidirectional gated recurrent neural networks is obtained when fusing the spectrogram and the bump scalogram, which is an improvement on the 61.0% baseline result provided by the DCASE 2017 organisers. This result shows that extracted bump scalograms are capable of improving the classification accuracy, when fusing with a spectrogram-based system.**

***Index Terms*—Acoustic scene classification (ASC), (bidirectional) gated recurrent neural networks ((B) GRNNs), convolutional neural networks (CNNs), deep scalogram representation, spectrogram representation.**

I. Introduction

ACOUSTIC scene classification (ASC) aims at the identification of the class (such as 'train station', or 'restaurant') of a given acoustic environment. ASC can be a challenging task, since the sounds within certain scenes can have similar qualities, and sound events can overlap one another [1]. Its applications are manifold, such as robot hearing or context-aware human-robot interaction [2].

In recent years, several hand-crafted acoustic features have been investigated for the task of ASC, including frequency, energy, and cepstral features [3]. Despite such year-long efforts, recently, representations automatically extracted from spectrogram images with deep learning methods [4], [5] are shown to perform better than hand-crafted acoustic features when the number of acoustic scene classes is large [6], [7]. Further, compared with a Fourier transformation for obtaining spectrograms, the wavelet transformation has the ability to incorporate multiple scales, and for this reason locally can reach the optimal time-frequency resolution [8] concerning the Heisenberg uncertainty of optimal time and frequency resolution at the same time. Accordingly, wavelet features have already been applied successfully for many acoustic tasks [9]−[13] , but often, the greater effort in calculating a wavelet transformation is considered not worth the extra effort if gains are not outstanding. In the theory of wavelet transformation, the scalogram is the time-frequency representation of the signal by wavelet transformation, where the brightness or the colour can be used to indicate coefficient values at corresponding time-frequency locations. Compared to spectrograms, which offer (only) a fixed time and frequency resolution, a scalogram is better suited for the task of ASC due to its detailed representation of the signal. Hence, a scalogram-based approach is proposed in this work.

We use convolutional neural networks (CNNs) to extract deep features from spectrograms or scalograms, as CNNs have proven to be effective for visual recognition tasks [14], and ultimately, spectrograms and scalograms are images. Several specific CNNs are designed for the ASC task, in which spectrograms are fed as an input [7], [15], [16]. Unfortunately, those approaches are not robust and it can also be time-consuming to design CNN structures manually for each dataset. Using pre-trained CNNs from large scale datasets [17] is a potential way to break this bottleneck. ImageNet[1] is a suited such big database promoting a number of CNNs each year, such as 'AlexNet' [18] and 'VGG' [19]. It seems promising to apply transfer learning [20] through extracting

[1]http://www.image-net.org/

Manuscript received January 29, 2018; accepted February 26, 2018. This work was supported by the German National BMBF IKT2020-Grant (16 SV7213) (EmotAsS), the European-Unions Horizon 2020 Research and Innovation Programme (688835) (DE-ENIGMA), and the China Scholarship Council (CSC). Recommended by Associate Editor Fei-Yue Wang. *(Corresponding author: Zhao Ren.)*

Citation: Z. Ren, K. Qian, Z. X. Zhang, V. Pandit, A. Baird, and B. Schuller, "Deep scalogram representations for acoustic scene classification," *IEEE/CAA J. of Autom. Sinica*, vol. 5, no. 3, pp. 662−669, May 2018.

Z. Ren, V. Pandit, and A. Baird are with the ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany (e-mail: {zhao.ren, vedhas.pandit, alice.baird}@informatik.uni-augsburg.de).

K. Qian is with the Machine Intelligence and Signal Processing Group, Technische Universität München, Germany, and also with the ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany (e-mail: andykun.qian@tum.de).

Z. X. Zhang is with the Group on Language, Audio and Music (GLAM), Imperial College London, UK (e-mail: zixing.zhang@imperial.ac.uk).

B. Schuller is with the Group on Language, Audio and Music (GLAM), Imperial College London, UK, and also with the ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany (e-mail: schuller@ieee.org).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JAS.2018.7511066