sustainability

MDPI

Article

ResNet Based on Multi-Feature Attention Mechanism for Sound Classification in Noisy Environments

Chao Yang [1,2,†], Xingli Gan [2,†], Antao Peng [2,*] and Xiaoyu Yuan [2,*]

[1] The 10th Research Institute of China Electronics Technology Group Corporation, Chengdu 610036, China; chaoyang34266744@gmail.com
[2] School of Information and Electronic Engineering, Zhejiang University of Science and Technology, Hangzhou 310023, China; 120070@zust.edu.com
* Correspondence: 222008252002@zust.edu.cn (A.P.); 222108855009@zust.edu.cn (X.Y.)
† These authors contributed equally to this work.

Abstract: Environmental noise affects people's lives and poses challenges for urban sound classification. Traditional algorithms such as Mel frequency cepstral coefficients (MFCCs) struggle due to audio signal complexity. This study applied an attention mechanism to a deep residual network (ResNet) deep learning network to overcome the structural impact of urban noise on audio signals and improve classification accuracy. We propose a three-feature fusion ResNet + attention method (Net50_SE) to maximize information representation in environmental sound signals. This method uses residual structured convolutional neural networks (CNNs) for feature extraction in sound classification tasks. Additionally, an attention module is added to suppress environmental noise impact and focus on different feature map channels. The experimental results demonstrate the effectiveness of our method, achieving 93.2% accuracy compared with 82.87% with CNN and 84.77% with long short-term memory (LSTM). Our model provides higher accuracy and confidence in urban sound classification.

Keywords: noise classification; attention mechanism; urban sound; MFCC; CNN; LSTM

Citation: Yang, C.; Gan, X.; Peng, A.; Yuan, X. ResNet Based on Multi-Feature Attention Mechanism for Sound Classification in Noisy Environments. *Sustainability* **2023**, *15*, 10762. https://doi.org/10.3390/su151410762

Academic Editor: Juan Miguel Navarro

Received: 31 May 2023
Revised: 21 June 2023
Accepted: 23 June 2023
Published: 8 July 2023

1. Introduction

Urban noise is a type of sound that causes irritability or excessive volume, which is harmful to human health. The goal of urban audio classification is to accurately predict sound events using an audio clip. Recently, urban audio scene classification has been widely used and attracted significant interest. It has found applications in various aspects of people's lives, including public surveillance [1], smart homes, and healthcare monitoring [2]. Audio classification can also be utilized in educational settings for learning assistance, assessment, feedback, and even music education. With the acceleration of urbanization, people have begun to pursue a higher quality of life. However, noise pollution has become a persistent issue, affecting people's daily lives. Audio scene classification can accurately identify and control the sources of noise [3]. Furthermore, there is substantial potential for related research in various areas, such as smart wearable devices [4], context awareness [5], audio management [6], robot navigation [7], and safety monitoring [8]. For example, people use wearable smart devices to classify and identify various sounds in urban life, providing timely reminders to adjust and change existing lifestyles and habits, thus protecting people's physical health. By recognizing environmental noise, the noise sources that generate the noise can be quickly located, and the problem of noise pollution can be promptly resolved, reducing the pollution of the environment and interference with people's lives. In addition to the strong interest from academia and industries at home and abroad in this field, government and military departments are also paying close attention to the latest research. The above explains why the related studies in the field of audio scene classification have great development potential and the field is worth investigating.

At present, many researchers are beginning to focus on deep learning algorithms. Therefore, the field of deep learning is developing rapidly, and deep learning algorithms are widely used in numerous application scenarios. Audio scene classification, too, takes inspiration from mature methods to address long-standing challenges. Mainstream audio scene classification models, such as CNNs [9] and recurrent neural networks (RNNs) [10], have been widely adopted as proven neural network methods. In terms of data augmentation, an innovative network architecture called a generative adversarial network (GAN) was introduced, which expands the dataset and reduces overfitting through adversarial learning. The field of deep learning currently has a relatively comprehensive theoretical framework and ample empirical evidence, leading to remarkable success in various research areas. Similarly, the audio classification task draws upon a wealth of related mature algorithms and has made significant progress.

However, there are still some areas that need to be improved regarding the known methods in the field of audio scene classification. CNNs typically require fixed-length inputs, while the length of audio signals can vary significantly. Different audio segments may contain time series of different lengths. This necessitates truncating or padding the audio during input, which can lead to information loss or the introduction of noise. RNNs, on the other hand, handle sequential data by utilizing recurrent units that continuously update their states to capture information from the sequence. However, when processing audio, RNNs often struggle to model the local frequency characteristics, which are crucial for audio classification tasks. Frequency information plays a vital role in audio processing. LSTM networks have a large number of learnable parameters, which can lead to overfitting issues when processing audio data, especially with limited training data.

The system proposed herein mainly consists of three key components: feature extraction, feature fusion, and the training of deep neural networks with a channel attention module. In the feature extraction phase, we extract three commonly used features in speech processing: the Mel frequency spectrogram, log-Mel spectrogram, and average of the log-Mel spectrogram. These features are extracted from the same audio signal. Next, we fuse the three different features into a single feature map, which serves as the input for the subsequent deep neural networks. Finally, we obtain predicted labels from the softmax layer of the ResNet with the attention module. This trained model can effectively classify audio signals based on the extracted features. Overall, our system combines feature extraction, feature fusion, and the utilization of deep neural networks with a channel attention module to achieve accurate audio classification.

Our contribution has several key aspects. Firstly, we propose a deep learning model that incorporates an attention module. This module assigns different weights to different channels, thereby enhancing the accuracy of the model. Since the duration of various scenes can vary significantly, and audio clips may contain overlapping sounds, the attention mechanism addresses these challenges by assigning varying weights to different characteristics. Consequently, it improves the accuracy of the deep neural network for audio scene classification. Secondly, our model effectively discriminated audio signals with similar features through feature combination. This approach enables the neural network to extract rich local feature information, thereby further enhancing the accuracy of audio classification. On the UrbanSound8K dataset, our model achieved an accuracy of 93.2%. In summary, our contributions include the introduction of an attention module to improve model accuracy by assigning different weights to channels as well as the utilization of feature combination to enhance the discrimination of audio signals with similar features. These advancements result in a high accuracy rate for audio classification.

The structure of the remainder of this paper is as follows: In Section 2, we introduce the related works on urban sound classification. Section 3 describes the methods that we used in this study as well as our experimental setup. The experimental results and specific analysis are presented in Section 4. Lastly, in Section 5, the conclusions regarding the attention mechanism for environmental sound classification are shown; moreover, we provide some directions for future work.

2. Related Works

In 1997, a research article about audio scene classification technology was published by MIT [11], which was the first to propose a solution for this problem. Using methods from speech recognition and acoustics, this paper records a dataset that includes human speech, subway, and traffic. The proposed model uses the K-nearest neighbor algorithm and RNN; according to the acoustic characteristics, the overall classification accuracy is 68%. Subsequently, Clarkson et al. divided a recorded audio stream into different scenes, then used a hidden Markov model (HMM) for learning [12]. Later, Ballas et al. stated that the essential characteristics of the audio signal should be studied in order to improve the accuracy of recognition of audio scenes [13]. At the same time, some researchers found that the basis of artificial audio scene classification is often only some typical acoustic events. Moreover, Tardieu et al. showed that there are several dispensable factors in audio scene recognition, such as the sound and sound source of human activities [14]. Support vector machine (SVM) has also been used for music classification, including the use of various feature extraction methods and parameter tuning for SVM [15]. Since then, machine learning algorithms have been widely used. Eronen et al. extracted a variety of acoustic features and compared the experimental results of different classification algorithms. Then, they came to a conclusion that after using some linear transformation to MFCC, the recognition accuracy can be improved [16].

Nowadays, CNNs [17] have emerged as a prominent research direction in the field of deep learning, particularly in the area of sample classification in pattern recognition. In the context of audio scene classification, convolutional layers are responsible for extracting signal features, facilitating abstract learning, and capturing image-related information across the entire input image. The pooling layer, on the other hand, reduces the dimensionality of audio feature data and extracts relevant feature information. Lastly, the fully connected layer is utilized to extract audio information and ultimately generate the accuracy of scene sound classification through softmax.

Additionally, with its simple and efficient advantages, attention mechanisms have emerged in natural language processing (NLP) and have flourished in computer vision (CV). The attention mechanism was originally a very important idea in the area of CV [18]. Scientists think that human vision has a different sensitivity to different parts of things: when human beings use limited resources to process visual information, they need to select special areas and then concentrate on extracting information. This is the so-called attention system, which is not a strict decoding process but is close to pattern recognition.

This mechanism is similar for human hearing. In recent years, many researchers began to use the attention mechanism in the field of audio scene classification. A film review scene classification model based on the attention mechanism achieved good results on an audio set. Yu et al. extended the multi-layer based on it and subsequently proposed the attention mechanism of multi-layer fusion [19]. Their model demonstrated superior effectiveness compared to Google's baseline model, further highlighting the indispensable role of the attention mechanism in the task of audio scene classification.

In addition to the research and optimization of deep neural network architecture, a crucial aspect of environmental audio classification is characterizing signal features. Typically, audio signals are obtained through sensors, resulting in a set of analog signals. However, since computers can only process digital signals, pre-processing of the acquired audio signals is necessary. The most commonly used method for audio signal pre-processing is performing Fourier transform, which yields frequency-domain features that measure the energy in different frequency bands. In the context of audio scene classification, this approach helps to emphasize the main components of the audio, facilitating classification.

Based on the described works, it is evident that most environmental sound classification models have primarily utilized either traditional machine learning algorithms or CNN alone. However, numerous sound recognition studies have shown that utilizing multiple features may yield better performance compared to using a single feature in environmental sound classification. Additionally, recent research has demonstrated that

introducing attention mechanisms into network models can significantly enhance their performance. Therefore, our objective was to incorporate an attention module into neural networks to increase the model's confidence and achieve higher accuracy. Furthermore, we aimed to leverage multiple features of audio signals to improve the overall performance of our model.

3. Methods

3.1. Deep Residual Network

From past studies, the number of layers in a neural network largely determines the performance of a deep learning model. When researchers keep stacking neural network layers, more complex features can be extracted for our task. However, will deeper network layers have better performance? In order to answer this problem, the neural networks have developed from the original AlexNet (7 layers) to Visual Geometry Group Network (VGG) 16 and VGG19 (16 layers and 19 layers, respectively) [20], and then evolve to GoogleNet (22 layers) [21]. However, researchers have discovered that increasing the number of network layers in deep learning models may not significantly improve model accuracy beyond a certain depth. Instead, it can lead to slower network convergence and even poorer performance on the test set compared to shallower neural networks. Many scholars have started recognizing that as network depth increases, the problem of gradient vanishing becomes more prominent.

The problem is indeed complex. In a series of experiments, researchers have observed a phenomenon known as the degradation problem, where increasing the network depth leads to stagnation or even a decrease in accuracy. This can be clearly observed in Figure 1, where the performance of a 56-layer neural network is worse than that of a 20-layer network. This issue cannot be attributed solely to overfitting, as overfitting typically results in slower accuracy growth rather than degradation.The challenges stem from problems such as gradient vanishing or exploding, which make training deep learning models difficult in practice. Techniques such as batch normalization have been introduced to alleviate these issues. Therefore, the emergence of the degradation problem is indeed surprising given the availability of such techniques.

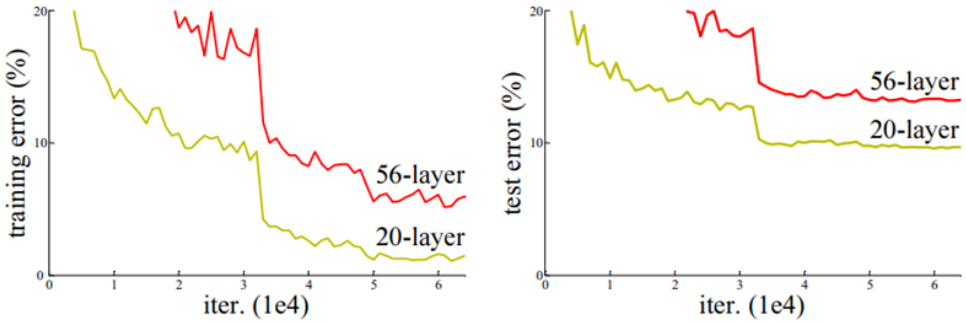

Figure 1. Training error and Testing error of 20 layers and 56 layers neural networks on cifar-10.

In 2015, ResNet was proposed to solve the above-mentioned degradation problem by Kai Minghe et al. [22]. The origin of Equations (1)–(3) is described below. Moreover, ResNet was the champion of the ImageNet competition in 2015, which reduced the recognition error rate to 3.6%, which even exceeded the recognition accuracy of normal human eyes.

Compared with the traditional convolutional neural network architecture, the residual neural network (ResNet) incorporates a residual learning unit. This unit, depicted in Figure 2, bears resemblance to a "short circuit" or a short connection within a circuit and is commonly referred to as a residual block. By utilizing these cross-layer connections, the input x can propagate data forward more efficiently and facilitate the backward propagation of gradients.

$$y_l = h(x) + F(x_l, W_l) \quad (1)$$

$$x_{l+1} = f(y_l) \tag{2}$$

where x_l and x_{l+1} represent the input and output of the number l residual unit, respectively. Note that each residual unit generally contains a multi-layer structure. F is the residual function, which represents the learned residual, $h(x_l) = x_l$ represents the identity mapping, and f is the RELU activation function. Based on the above formula, we find that the learning characteristics of shallow l to deep L are:

$$x_L = x_l + \sum_{i=l}^{L-1} F(x_i, W_i) \tag{3}$$

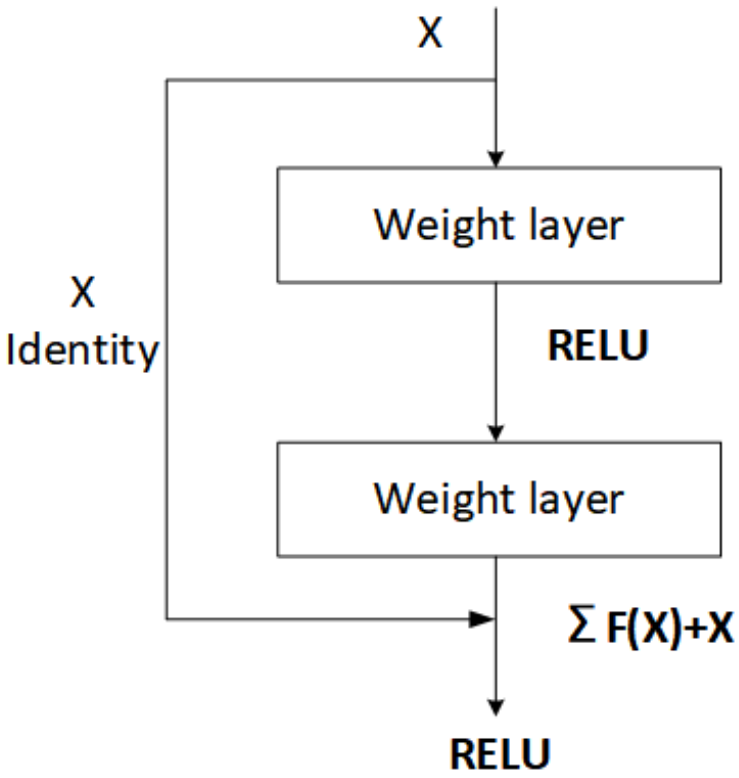

Figure 2. Residual learning unit of ResNet.

To summarize, one of the major challenges faced by deep networks, such as VGG networks, is the degradation problem. As the neural network deepens and the number of parameters increases, it becomes increasingly difficult to optimize the network, leading to a decline in overall performance compared with shallower networks. In contrast, ResNet addresses this issue by introducing skip connections between certain layers of the neural network. These skip connections allow the neurons in one layer to directly connect with neurons in subsequent layers, effectively weakening the strong connections between each layer and alleviating the degradation problem. Figure 3 demonstrates that ResNet maintains strong accuracy growth as the network depth increases, avoiding the performance degradation observed in VGG network structures with an increasing number of layers.

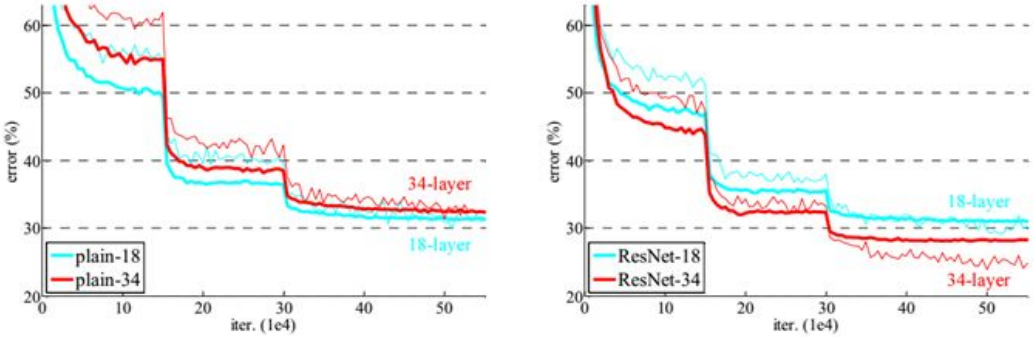

Figure 3. Training error on ImageNet: 18 layers and 34 layers of plain and ResNet.

3.2. Channel Attention Module

In most daily scenarios where deep learning techniques are used, shallow neural networks often perform poorly. However, blindly increasing the number of layers in the network may lead to a decrease in the performance of the model. Even though the proposal of the residual network effectively alleviated this problem, its effect becomes very limited when the number of layers in the network increases to hundreds.

After the proposal of the Inception network, researchers began to think about how to effectively increase the depth of a neural network without reducing the accuracy of the model. Using convolutional neural networks to extract features has many advantages, but it ignores rich local information. The Inception network addresses this limitation by extracting globally associated information onto a feature map through a multi-scale convolution kernel, but it still ignores the association between channels.

Squeeze-and-exception (SE) networks [23], which is not a strictly defined neural network structure but a module that can be embedded into the existing deep learning models of classification or recognition. The origin of Equations (4) is next described. The design idea of SEnet is as follows:

(1) To extract spatial information through the squeeze operation, in order not to greatly increase the time and space complexity of the model, we also hope to compress this information effectively.
(2) In the extraction step, we use the global information obtained in the previous step to capture the information between each channel. What is learned is not mutually exclusive information.

This is used as shown in Figure 4. The following two operations are performed: squeeze and exception operations for any mapping $F_{\mathrm{tr}} : X \to U, X \in R^{H' \times W \times C}, U \in R^{H \times W \times C}$. For example, suppose the convolution kernel we use is $V = [v_1, v_2, ..., v_c]$, and then we obtain: $U = [u_1, u_2, ..., u_c]$:

$$u_c = v_c * X = \sum_{s=1}^{C} \mathcal{V}_c^8 * \mathcal{X}^s \tag{4}$$

where $*$ represents a convolution calculation, and x^s represents the two-dimensional convolution kernel of each channel. It inputs the spatial features on one channel. However, due to the summation of the convolution results of each channel, the characteristic relationship of each channel is mixed with the spatial relationship learned by the convolution kernel. The purpose of the SE module is to prevent this mixing and allow the model to directly learn the characteristic relationship between channels.

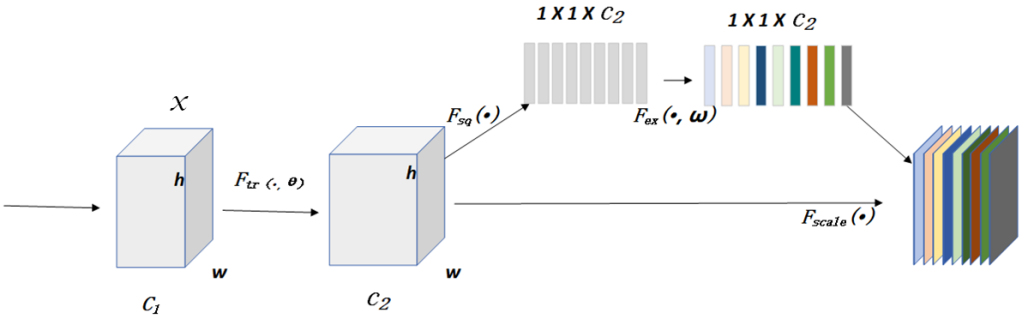

Figure 4. Structure of squeeze-and-excitation (SE) networks.

In practice, the SEnet acts between different channels and assigns different weights of attention to different channels. This allows the model to assign greater weights to channels with more information and suppress the unimportant channel features. The most significant advantage of the SE module is its plug-and-play nature, as it can be easily integrated into existing networks to improve network performance at a lower cost.

Owing to the flexibility of the SE module, it can be directly applied to the existing network structure. Moreover, it can be embedded into almost all network structures. Thus, given the outstanding performance of ResNet described above, in this paper, we embed the SE module into the ResNet module as shown in Figure 5.

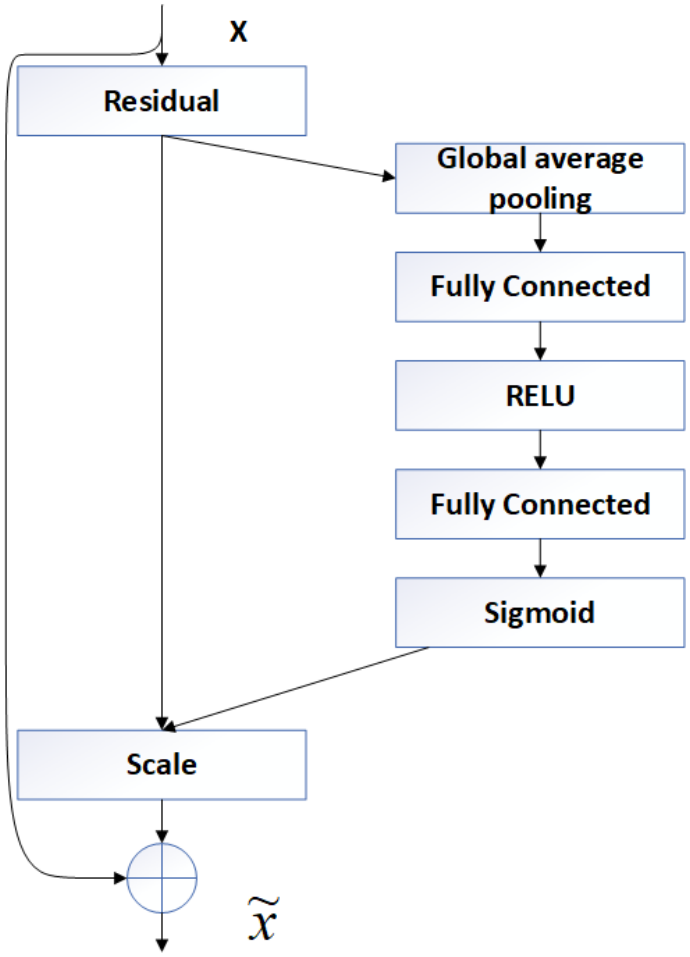

Figure 5. Structure of embed SE module into ResNet block.

As can be seen from the above figure, assuming we have a feature map with a length of H, width of W, and C channels, we perform the following operations on the side branch of the residual learning block:

(1) We perform a global average pooling operation to change the size of the feature map to $1 \times 1 \times C$;
(2) This is followed by the first fully connected layer, where we compress the number of channels to C/R to reduce the computational load of the network;
(3) Use the nonlinear activation function RELU to further improve the nonlinearity of the model;
(4) Restore feature maps to C channels using a second fully connected layer;
(5) Finally, the nonlinear activation function sigmoid is used to map the output to the range of 0 to 1, and then multiplied with each channel of the previous stage input as the input of the next stage.

The advantage of this approach is that it enhances the nonlinearity of the model, allowing it to better capture the characteristic information of different channels. Additionally, it significantly reduces the number of parameters in the model and reduces computational complexity.

3.3. Method of Feature Extraction

In the research on speech recognition, the most common feature used is MFCC [24]. In 1937, Stevens, Volkmann, and Newmann proposed a pitch unit that makes equal pitch distances sound equal to listeners, which is called the Mel spectrum.

The Mel scale filter bank has a higher resolution in the low-frequency range, which aligns with the auditory characteristics of the human ear. This is also the fundamental principle of the Mel scale. As a result, compared to LPCC [25], based on the vocal tract model, this parameter exhibits better robustness, better alignment with human auditory characteristics, and superior recognition performance even in scenarios with reduced signal-to-noise ratios. The extraction process of MFCC, as illustrated in Figure 6, involves crucial steps such as fast Fourier transform (FFT) and the utilization of Mel filter banks, which perform essential dimensionality operations.

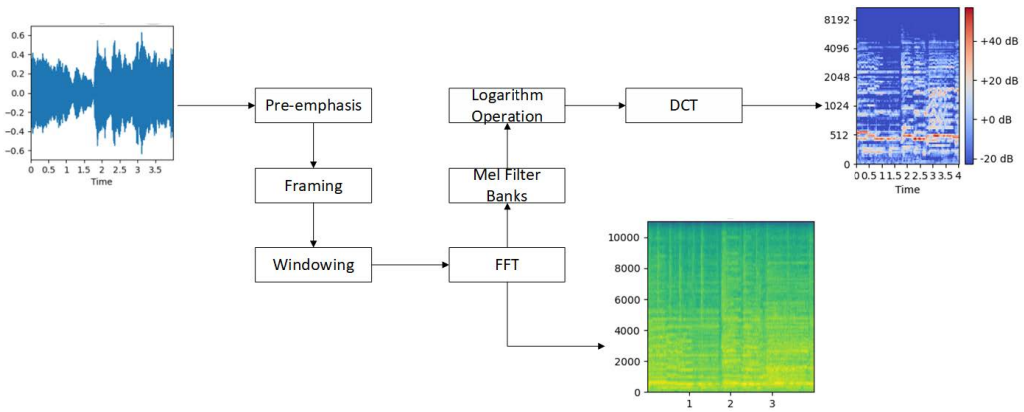

Figure 6. An audio signal converted to Mel spectrogram from time−domain graph.

The Mel scale describes the nonlinear characteristics of human ear frequency, and its relationship with frequency can be approximately expressed by the following formula [26]:

$$\mathrm{Mel}(f) = 2595 \times \lg\left(1 + \frac{f}{700}\right) \tag{5}$$

Moreover, it is mainly used for the feature extraction of speech data and reduction in operation dimensions.

3.4. Muti-Features Fusion

As depicted in Figure 7, we can observe various audio data samples, their corresponding source labels, and the corresponding audio waveform diagrams extracted from the UrbanSound8K dataset. It is evident that relying solely on the time-domain waveform for audio signal classification can yield unreliable results. Consequently, the majority of audio classification approaches utilize Mel-spectrum features, as shown in Figure 8, to facilitate the more accurate classification of audio signals. However, extensive research has demonstrated that considering the unique characteristics of audio signals is crucial, and only by leveraging the rich information within these characteristics can we obtain an improved classification model.

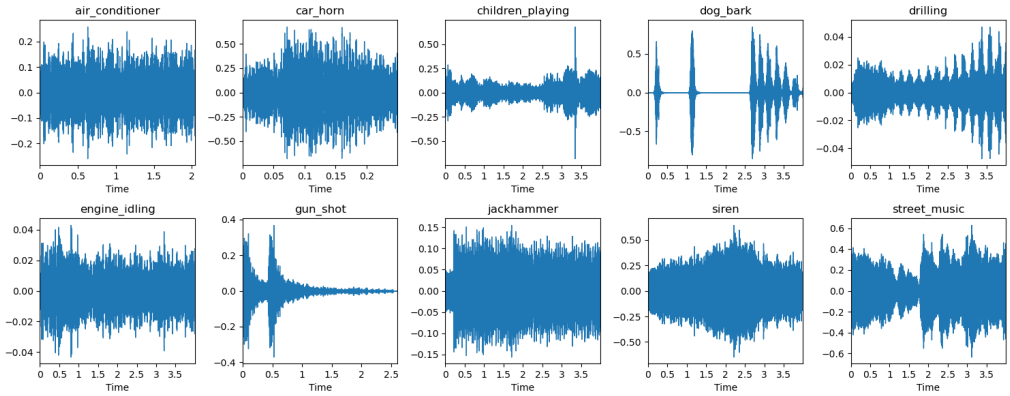

Figure 7. Time−domain waveforms of ten urban sounds classes from UrbanSound8K.

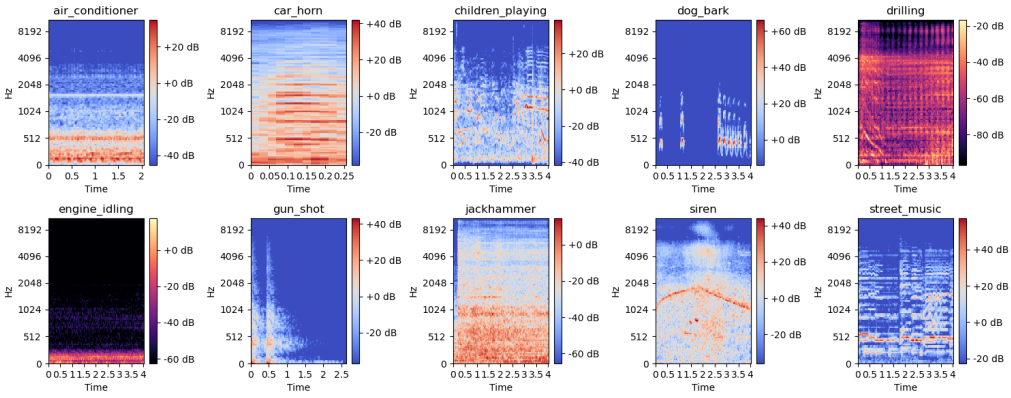

Figure 8. Mel−Spectra of ten urban sounds classes from UrbanSound8K.

In Figure 9, we present our engineered features. We used three feature engineering techniques to construct the 3D feature map for each sample, namely log-scaled Mel spectrum, derivative of log-scaled Mel spectrum, and the average of both. This approach allowed us to clearly visualize the differences between different types of audio signals. The resulting feature map serves as the input for the subsequent neural networks, enabling accurate analysis and classification of the audio data.

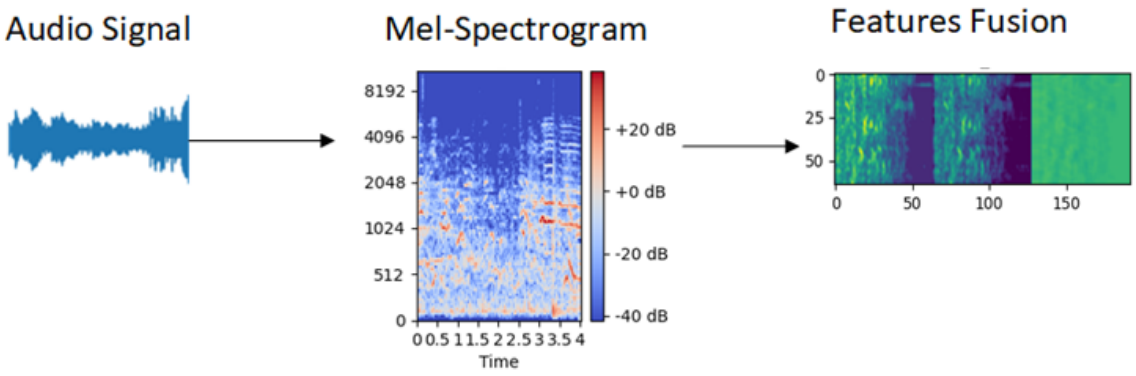

Figure 9. Features fusion of Mel-Spectrum and its nonlinear transform.

3.5. Experimental Setup

To evaluate our proposed model, we utilized the UrbanSound8K dataset, which is a widely used public dataset for automatic urban environmental sound classification. The dataset consists of 10 sound classes, namely, air conditioner (AI), car horn (CA), children playing (CH), dog barking (DO), drilling (DR), engine idling (EN), gunshot sound (GU), jackhammer (JA), siren (SI), and street music (ST). In total, there are 8732 audio data samples, each cropped or extended to a duration of 4 seconds. The sampling rate for each audio signal was set to 22,050 Hz. After applying the feature extraction techniques, we obtained a dataset containing the corresponding feature maps for each audio signal. Each feature map had a size of $64 \times 64 \times 3$.

After the extraction of the original data into feature maps, we generated TFRecord format files for the subsequent training process. During training, we conducted testing every 200 batches, with each batch having a size of 32. The training process lasted 50 epochs. For each batch, we recorded the training loss, training accuracy, test loss, and test accuracy. The dataset was divided into training, testing, and validation sets in a ratio of 6:2:2, as stated in this paper.

Our model and other models were implemented using Tensorflow, an end-to-end open-source machine learning platform. For extracting the Mel spectrum and reshaping the audio crops, we utilized the Librosa library, which is a Python toolkit specifically designed for audio, music analysis, and processing. Librosa provides a wide range of functions for common time-frequency processing, feature extraction, and visualizing sound graphics, among other capabilities.

4. Results and Discussion

We selected the following three models for comparison: CNN, LSTM, and our model. Their specific performance is shown in Figure 10. The performance of these models is close: CNN provides 82.87% average accuracy, LSTM provides 84.77% average accuracy, and our model provides 93.2% average accuracy. We also give the accuracy of the ResNet model and Net50_SE model separately in Table 1. From the table, we can intuitively see that the accuracy of the model is significantly improved by the introduction of the attention mechanism.

Table 1. Accuracy of ResNet model and Net50_SE model trained on UrbanSound8K dataset.

Model	Accuracy
ResNet	89.2%
Net50_SE	93.2%

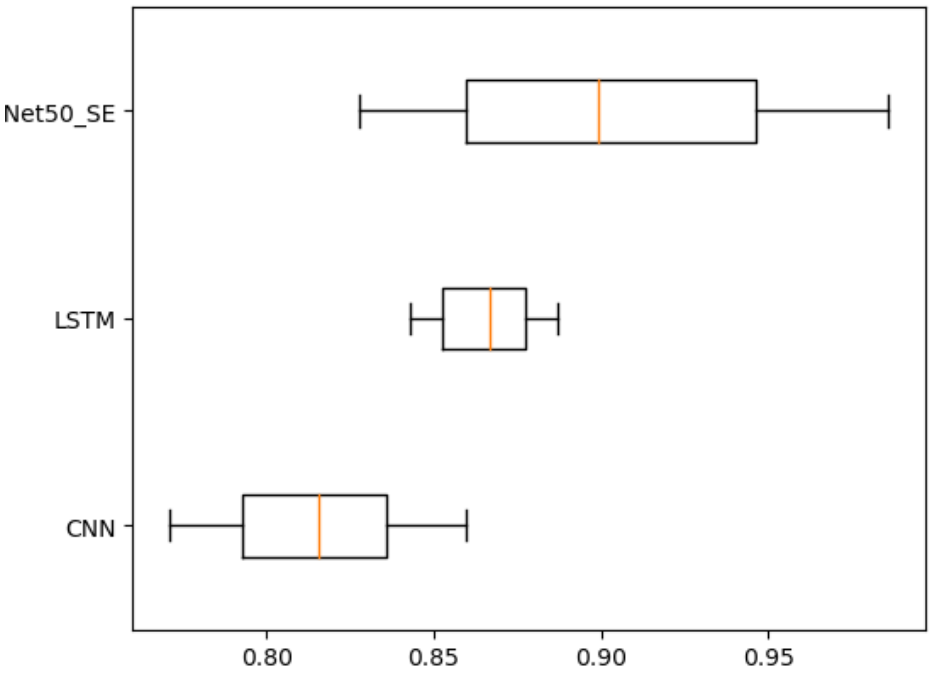

Figure 10. Average accuracy of our model, LSTM, and CNN.

A portion of the boxplot, which contains half of the total data, can reflect the fluctuation in the data. In terms of the median score, Net50_SE performs the best with a median score of 0.9, followed by LSTM. CNN has the lowest median score. This indicates that Net50_SE outperforms the other two algorithms in overall performance. Regarding the interquartile range (IQR), there is little difference between LSTM and CNN. However, Net50_SE exhibits a larger IQR, indicating that its performance has a greater range of possibilities based on the differences in sample size and across different sample groups.

Additionally, we took the F1 score as the final evaluation method [27], which is a measure of evaluating the performance of multi-classification. The F1 score takes into account both the accuracy and recall of the classification model, and its range is from zero to one. Compared with accuracy, it comprehensively considers the calculation results of model precision and recall, and the value is more inclined to the index with a smaller value. The F1 is calculated as follows:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{6}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{7}$$

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{8}$$

In Table 2, we present the precision, recall, andsScore for each class. True positive (TP) represents a positive example that is correctly predicted, false positive (FP) indicates a negative example that is incorrectly predicted as positive, and false negative (FN) represents a positive example that is incorrectly predicted as negative. From the table, it can be observed that our model outperforms CNN and LSTM, achieving the highest score among the three models. Specifically, CNN and LSTM exhibit lower accuracy in certain classes (e.g., air conditioner, children playing, and street music), whereas our model maintains consistent performance across all classes, indicating higher generalization ability. Therefore, ResNet

with the SE block demonstrates excellent performance in terms of accuracy, precision, recall, and overall score.

Table 2. Precision, recall, and F1 score for Net50_SE, LSTM, and CNN in per-classes.

Method	Net50_SE			LSTM			CNN		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score	Precision	Recall	F1-Score
AI	0.91	0.93	0.92	0.8	0.88	0.84	0.74	0.83	0.78
CA	0.89	0.92	0.86	0.82	0.85	0.83	0.94	0.79	0.86
CH	0.79	0.84	0.81	0.78	0.73	0.75	0.63	0.71	0.67
DO	0.74	0.81	0.77	0.86	0.83	0.84	0.85	0.8	0.83
DR	0.91	0.88	0.89	0.87	0.87	0.87	0.86	0.81	0.83
EN	0.9	0.91	0.9	0.88	0.85	0.86	0.8	0.84	0.82
GU	0.96	0.93	0.94	0.93	0.94	0.94	0.93	0.89	0.91
JA	0.92	0.91	0.91	0.89	0.9	0.91	0.87	0.84	0.85
SI	0.92	0.95	0.93	0.9	0.91	0.9	0.95	0.83	0.88
ST	0.84	0.86	0.85	0.75	0.73	0.74	0.7	0.73	0.71

The confusion matrix obtained on the test set is shown in Figure 11. As we can see in the figure, the pair of classes of children playing vs. dog barking, shows high confusion. This may be due to their complex structure in both the time domain and frequency domain, which hampers the accuracy of the model. Additionally, there are fewer samples available for these two urban sound types compared to the other classes, which also contributes to the high confusion in practice.

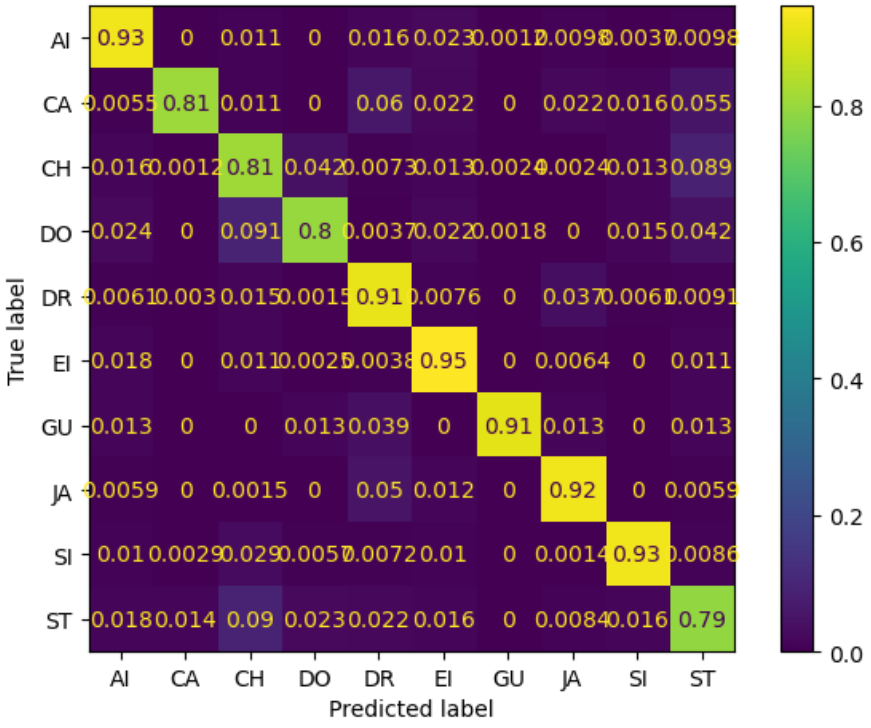

Figure 11. Confusion matrix for Net50_SE on Test set.

In Figure 12, we present the training and testing accuracy and loss across epochs. It is evident that our model achieves its best performance after the 50th epoch on the test data. Moreover, the accuracy of the model, both in the testing set and training set, quickly approaches one, indicating its excellent generalization ability. Unlike LSTM, which is more prone to overfitting during training due to its recurrent structure, the design of ResNet takes this into consideration. With the powerful attention mechanism, our model demonstrates

strong learning and expressive capabilities during the actual training process. As a result, our model exhibits greater confidence than the other models.

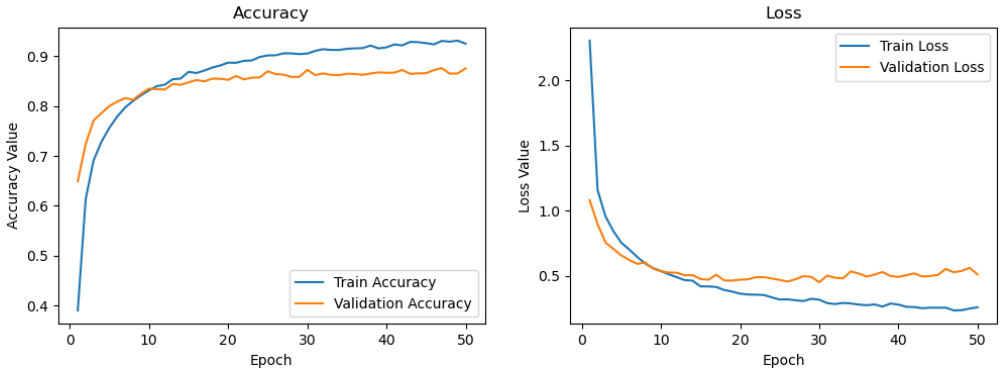

Figure 12. Accuracy and loss evaluated on training data and validation data during training.

At last, we present several widely used models for environmental audio classification in Table 3 and compare their features and average classification accuracy. As shown in the table, our model achieves the best results.

Table 3. Accuracy of classification on UrbanSound8K dataset.

Reference	Classification	Features	Accuracy
[28]	SVM	Mel banks, MFCC	70%
[29]	CRNN	Raw waveforms	79.6%
[30]	CNN	MFCC	83.5%
[31]	LSTM	MFCC	84%
[32]	GoogleNet	MFCC, Spectrograms, MFCC, CRP image	93%
this paper	Net50_SE	MFCC, Log-Scaled MFCC, Average Log-Scaled MFCC	93.2%

The performance of ResNet and attention mechanism is limited by the quantity and diversity of the training data. If the training data are limited or lack sufficient diversity, the model may not be able to fully learn the features and patterns of the audio data. Based on different audio datasets, our model produces varying results. Firstly, ResNet has relatively fewer parameters, which means it requires less computational resources during training and inference, making it more practical. The attention mechanism helps the model focus more accurately on important information in the input, thereby improving the model's performance in various tasks. It allows the model to concentrate its attention on key features, reducing the processing of irrelevant information and enhancing the accuracy and generalization ability of the model. Combining the experiments, Net50_SE shows improved model performance, enhanced interpretability of model decisions, reduced unnecessary computations and parameter consumption, and improved computational efficiency of the model. Compared with the baseline of CNN and LSTM trained on the same dataset, ResNet with an SE module provides a significant improvement and is thus more confident.

5. Conclusions

With the continuous progress and rapid development of the field of artificial intelligence, computers' perception and understanding of the environment have become key to the next step of development. Audio signals are not only the source of human understanding of the environment but also important bridges between computers and the environment. Therefore, in today's highly digital society, urban sound classification has many application scenarios and continues to attract researchers at home and abroad to carry out related research. We mainly focused on the development status of audio classification and deep neural networks as well as comparisons and analyses of several existing

mainstream solutions. Followed by this, we introduced the relevant knowledge of feature engineering and neural networks and briefly introduced the Urbansound8K dataset. Then, we proposed an attention model based on multiple features for the task.

In future research, data pre-processing should be the top priority. We should pay more attention to the fusion of audio signal characteristics, especially the combination of time- and frequency-domain characteristics. We hope that our model can be applied to more diverse urban scenarios in the future, conducting multiple cross-validation experiments and further exploring multi-audio classification to enhance the functionality of the system. At the same time, many related studies have shown that the amount of data used to train a model also determines the performance of the results to a certain extent. In recent years, reinforcement learning has shown strong abilities and has made great achievements in the field of speech enhancement. We can generate more audio data through GANs, improve the resolution of the original data, and obtain higher quality datasets.

Author Contributions: Conceptualization, C.Y., X.G. and A.P.; data curation, C.Y., X.G. and X.Y.; formal analysis, C.Y. and X.G.; methodology, C.Y., X.G., A.P. and X.Y.; software, C.Y., X.G., A.P. and X.Y.; writing—original draft, C.Y. and X.G. All authors have read and agreed to the published version of the manuscript.

Funding: This study was supported in part by the National Key Research and Development Plan of China (project: Emergency command and communication networks and terminal equipment for harsh environments such as mountainous and dense forest areas (No. 2020YFC1511800), Adaptive design of equipment in harsh environment and development of miniaturized emergency communication equipment (2020YFC1511804).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Cristani, M.; Bicego, M.; Murino, V. Audio-visual event recognition in surveillance video sequences. *IEEE Trans. Multimed.* **2007**, *9*, 257–267. [CrossRef]
2. Peng, Y.T.; Lin, C.Y.; Sun, M.T.; Tsai, K.C. Healthcare audio event classification using hidden Markov models and hierarchical hidden Markov models. In Proceedings of the 2009 IEEE International Conference on Multimedia and Expo, New York, NY, USA, 28 June–3 July 2009; IEEE: New York, NY, USA, 2009; pp. 1218–1221.
3. Meyer, J.; Dentel, L.; Meunier, F. Speech recognition in natural background noise. *PloS ONE* **2013**, *8*, e79279. [CrossRef]
4. Xu, Y.; Li, W.J.; Lee, K.K. *Intelligent Wearable Interfaces*; John Wiley & Sons: Hoboken, NJ, USA, 2008.
5. Schilit, B.; Adams, N.; Want, R. Context-aware computing applications. In Proceedings of the 1994 First Workshop on Mobile Computing Systems and Applications, Santa Cruz, CA, USA, 8–9 December 1994; IEEE: New York, NY, USA, 1994; pp. 85–90.
6. Landone, C.; Harrop, J.; Reiss, J. Enabling Access to Sound Archives Through Integration, Enrichment and Retrieval: The EASAIER Project. In Proceedings of the ISMIR 2007, Vienna, Austria, 23–27 September 2007; pp. 159–160.
7. Chu, S.; Narayanan, S.; Kuo, C.C.J.; Mataric, M.J. Where am I? Scene recognition for mobile robots using audio features. In Proceedings of the 2006 IEEE International Conference on Multimedia and Expo, Toronto, ON, Canada, 9–12 July 2006; IEEE: New York, NY, USA, 2006; pp. 885–888.
8. Harma, A.; McKinney, M.F.; Skowronek, J. Automatic surveillance of the acoustic activity in our living environment. In Proceedings of the 2005 IEEE International Conference on Multimedia and Expo, Amsterdam, The Netherlands, 6–9 July 2005; IEEE: New York, NY, USA, 2005; p. 4.
9. Kim, J.; Lee, K. Empirical study on ensemble method of deep neural networks for acoustic scene classification. In Proceedings of the IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE), Budapest, Hungary, 3 September 2016.
10. Bae, S.H.; Choi, I.; Kim, N.S. Acoustic scene classification using parallel combination of LSTM and CNN. In Proceedings of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE2016), Budapest, Hungary, 3 September 2016; pp. 11–15.
11. Sawhney, N.; Maes, P. Situational awareness from environmental sounds. *Project Rep. Pattie Maes.* **1997**, *13*, 1–7.
12. Clarkson, B.; Sawhney, N.; Pentland, A. Auditory context awareness via wearable computing. *Energy* **1998**, *400*, 20.

13. Ballas, J.A. Common factors in the identification of an assortment of brief everyday sounds. *J. Exp. Psychol. Hum. Percept. Perform.* **1993**, *19*, 250. [CrossRef] [PubMed]
14. Tardieu, J.; Susini, P.; Poisson, F.; Lazareff, P.; McAdams, S. Perceptual study of soundscapes in train stations. *Appl. Acoust.* **2008**, *69*, 1224–1239. [CrossRef]
15. Dhanalakshmi, P.; Palanivel, S.; Ramalingam, V. Classification of audio signals using SVM and RBFNN. *Expert Syst. Appl.* **2009**, *36*, 6069–6075. [CrossRef]
16. Eronen, A.J.; Peltonen, V.T.; Tuomi, J.T.; Klapuri, A.P.; Fagerlund, S.; Sorsa, T.; Lorho, G.; Huopaniemi, J. Audio-based context recognition. *IEEE Trans. Audio Speech Lang. Process.* **2005**, *14*, 321–329. [CrossRef]
17. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **2012**, *25*, 84–90. [CrossRef]
18. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 3058.
19. Yu, C.; Barsim, K.S.; Kong, Q.; Yang, B. Multi-level attention model for weakly supervised audio classification. *arXiv* **2018**, arXiv:1803.02353.
20. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
21. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
22. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
23. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
24. Zheng, F.; Zhang, G.; Song, Z. Comparison of different implementations of MFCC. *J. Comput. Sci. Technol.* **2001**, *16*, 582–589. [CrossRef]
25. Gupta, H.; Gupta, D. LPC and LPCC method of feature extraction in Speech Recognition System. In Proceedings of the 2016 6th International Conference-Cloud System and Big Data Engineering (Confluence), Noida, India, 14–15 January 2016; IEEE: New York, NY, USA, 2016; pp. 498–502.
26. Umesh, S.; Cohen, L.; Nelson, D. Fitting the Mel scale. In Proceedings of the 1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No.99CH36258), Phoenix, AZ, USA, 15–19 March 1999; Volume 1, pp. 217–220.
27. Chicco, D.; Jurman, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genom.* **2020**, *21*, 1–13. [CrossRef] [PubMed]
28. Salamon, J.; Jacoby, C.; Bello, J.P. A dataset and taxonomy for urban sound research. In Proceedings of the 22nd ACM International Conference on Multimedia, Orlando, FL, USA, 3–7 November 2014; pp. 1041–1044.
29. Sang, J.; Park, S.; Lee, J. Convolutional recurrent neural networks for urban sound classification using raw waveforms. In Proceedings of the 2018 26th European Signal Processing Conference (EUSIPCO), Rome, Italy, 3–7 September 2018; IEEE: New York, NY, USA, 2018; pp. 2444–2448.
30. Davis, N.; Suresh, K. Environmental sound classification using deep convolutional neural networks and data augmentation. In Proceedings of the 2018 IEEE Recent Advances in Intelligent Computational Systems (RAICS), Thiruvananthapuram, India, 6–8 December 2018; IEEE: New York, NY, USA, 2018; pp. 41–45.
31. Lezhenin, I.; Bogach, N.; Pyshkin, E. Urban sound classification using long short-term memory neural network. In Proceedings of the 2019 Federated Conference on Computer Science and Information Systems (FedCSIS), Leipzig, Germany, 1–4 September 2019; IEEE: New York, NY, USA, 2019; pp. 57–60.
32. Boddapati, V.; Petef, A.; Rasmusson, J.; Lundberg, L. Classifying environmental sounds using image recognition networks. *Procedia Comput. Sci.* **2017**, *112*, 2048–2056. [CrossRef]