A Deep Meta-model for Environmental Sound Recognition

K. S. Arun

Abstract Nowadays, sound serves as a crucial factor in all facets of human life. Staring from automating personal security systems to critical surveillance systems, sound is an indispensable component. The practical implementation of the present day automatic sound recognition systems in real-life settings is inadmissible due to their poor detection accuracy. However, deep learning-based systems overcome the incompetence of the traditional machine learning-based models, and it can be used to develop automatic sound classification systems. This work proposes a deep meta-model for categorizing environmental sounds on the basis of the spectrogram images generated from these sounds. In the proposed approach, spectrogram images of environmental sounds are used to train five different deep learning models, and the predictions from these base models are then stacked using the proposed deep meta-model. Experimental results on two benchmark datasets such as ESC-50 and UrbanSound 8K demonstrate the fact that the proposed deep meta-model is a promising alternative to the conventional approaches for environmental sound recognition.

Keywords Deep learning · Ensemble model · Stacking · Neural network

1 Introduction

Over the previous decade, significant amount of research has been proposed toward environmental sound modeling and its recognition. Routine sounds, i.e., the sounds people experience in day-to-day life except speech and music, are commonly referred to as environmental sounds. Nowadays, environmental sound recognition (ESR) system is a key component in efficient machine audition. With the developing interest on query-based probing such as content-based video and image retrieval [4], ESR is often instrumental in effective sound search applications. Once the sound files have been automatically labeled with meaningful keywords, an ESR system can be used

K. S. Arun (✉)
Department of Computer Science and Engineering, Amal Jyothi College of Engineering, Kottayam, India
e-mail: iamarunks@gmail.com

A. Kumar et al. (eds.), *ICDSMLA 2021*, Lecture Notes in Electrical Engineering 947,
https://doi.org/10.1007/978-981-19-5936-3_19

for keyword-based sound recovery. Moreover, robot navigation is often improved by integrating an ESR module in the navigation framework. In recent years, ESR is effectively incorporated in home-surveillance, be it for helping old folks living alone in their own residence or for a smart house. Similarly, ESR can be customized for the recognition of animals and poultry species by their distinctive sounds.

In recent years, music and speech are the two fundamental categories of audio signals that have been extensively investigated. In its early stages of development, ESR frameworks were very much identical to sound and music recognition models. Majority of these ESR framework comprises of three different stages such as signal pre-processing, extracting domain peculiar features, and then classification. Signal pre-processing divides the input into multiple segments from which meaningful features are extracted. In feature extraction step, the dimensionality of all these segments was reduced by characterizing each of the segments by low-dimensional feature vectors. The commonly used features are pitch, crossing rate, and frame-related measures. These features are then passed to an appropriate classifier for recognizing speech.

However, the above-mentioned framework is found to be inappropriate for ESR because of its non-stationary behavior. As an example, phonetic structures have been considered as the fundamental building blocks of speech signals and they form the basis of speech recognition. As opposed to this, environmental sounds such as thunder or storm do not have such observable sub-structure present in them. On the other hand, music signals often exhibit relevant static patterns such as rhythm and melody. More recently, visual displays of an audio signals obtained with different time frequency representations such as spectrograms provide extensive description of the temporal and spectral structure of the original signal. Furthermore, features extracted from spectrogram images yielded more promising results for the task of detecting environmental sounds.

In general, spectrogram is an efficient approach to visualize frequency spectrum of the underlying sound waves. Such spectrogram images can be further analyzed using deep learning-based frameworks. The hierarchical feature learning capacity [2, 3, 13] of deep learning-based frameworks can be utilized to develop more efficient sound classification systems by overcoming the limitations of traditional approaches. However, the lack of sufficient training data still hinders the performance of these deep learning models. An ensemble of deep learning classifiers is often found to be much more accurate than individual ones [7]. Among ensemble-based approaches, stacking involves the training of a meta-model in order to aggregate the results of multiple models to generate the final prediction. To this end, this paper proposes an improved stacking-based ESR system using spectrogram.

The remaining sections of this paper are organized as follows: A brief overview of the existing works in the field of ESR is presented in Sect. 2. The proposed methodology for ESR is explained in Sect. 3. Experimental evaluation of the deep meta-model is depicted in Sect. 4, and the proposed approach is concluded in Sect. 5.

2 Related Work

Numerous machine learning-based models have been proposed in the literature for the task of environmental sound recognition. The rest of this section will briefly summarize the state-of-the-art ESR techniques. Demir et al. [5] proposed a ESR system based on CNN-based features extracted from spectrogram images, and it uses a KNN classifier defined in ensemble sub-space for sound recognition. Later on, the same authors [6] also introduced a pyramidal-CNN framework to recognize environmental sounds. They used VGG16, VGG19, and DenseNet201 models for feature extraction from sound images obtained by applying STFT on sound signals. Here, feature is extracted in pyramidal mode so that the generated feature descriptor is of having higher dimension. Finally, an SVM classifier is trained using these high-dimensional pyramidal features for recognizing various environmental sounds.

More recently, Zhang et al. [19] employed recurrent neural network model to extract temporal correlations from spectrogram images. They also employed frame-level attention mechanism to concentrate more on semantically linked and prominent frames. Meanwhile, an automatic system to recognize bird species has been proposed by Stastny et al. [14]. Their proposed model involves two processing stages. In the initial stage, certain pre-processing operations such as framing, noise removal, and pre-emphasis are performed on sound clips to generate spectrograms. In the second stage, a CNN-based model is trained using the spectrograms generated in the initial stage as input to recognize the species in which the bird belongs. Later on, Guzhov et al. [8] proposed a model that takes STFT sound images as input and various pre-trained models such as ResNet and SiameseNet are trained to perform sound classification.

A weighting filters-based CNN architecture for ESR is introduced by Tang et al. [15]. In this framework, a new mechanism for feature weighting has been proposed, and they achieved significant improvement in detection accuracy with this weighting mechanism. On the other hand, stacking-based CNN model to recognize environmental sound has been proposed by Ahmed et al. [1]. This model takes Log-Mel (LM) spectrogram images from the sound clips as input. Extensive hyperparameter tuning has been proposed in this work such as employing different drop-out rates, various padding schemes in the convolution layers, altering the size of max-pooling layer, and varying the stride values to figure out the combination that yield the best accuracy in detecting environmental sounds.

In recent years, few stacking-based ESR systems have been proposed [1, 11, 12]. However, majority of them employed time domain and time frequency domain features to combine the predictions generated by different classifiers. Spectrogram-based features are not used in any of these existing approaches. Also, most of these models assume a linear relationship between the prediction scores of base models while stacking is performed. However, the relationship between base classifiers is not linear in real life.

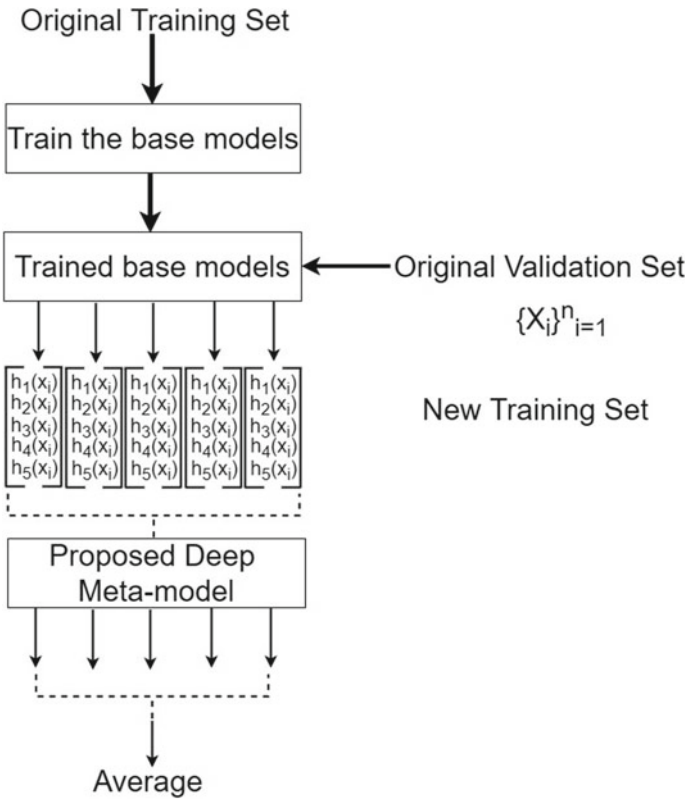


Fig. 1 Proposed ESR model

3 Methodology

The architecture of the proposed ESR system is depicted in Fig. 1. The proposed system consists of five base models, each of them generates classification results by analyzing the spectrogram corresponding to the given environmental sound, and later on, the deep meta-model efficiently combines the classification results yielded by each of the base models to generate a consensus final result. A detailed description of the base models used in this work and the proposed deep meta-model to aggregate the results of these base models is given below.

3.1 Base Models

In this work, the first level classifiers refereed to as the base level models are trained based on the principle of transfer learning. In transfer learning, the knowledge

acquired while solving a particular problem is reused in a different but identical problem. As an example, the expertise attained while trying to classify Jeeps can be utilized to certain extent in classifying Buses. To this end, the following CNN-based models have been used in this work as the base level models:

- **FractalNet**: FractalNet [10] is a specific type of CNN model without having residual connections but involves a fractal type design. FractalNet implements an expansion rule that is executed in recursive fashion to account for a deep network structure with its basic units known as fractals. In this type of architecture, there will be correspondence between diverse sub-paths and it do not assimilate pass-through interdependence.
- **ResNeXt**: ResNeXt [17] is based on the paradigm of divide and conquer. Its intention is to reduce the number of hyperparameters as opposed to ResNet. This can be achieved by the concept of cardinality. In ResNeXt architecture, cardinality can be treated as an additional dimension apart from its depth and width. Thus, ResNeXt architecture is constructed by aggregating the repetition scheme of ResNet and the split-transform-merge principle of InceptionNet.
- **Wide Residual Network (WideResNet)**: WideResNet [18] is an alternative to ResNet, where the depth of the residual network is reduced, but its width is increased. This can be achieved with the help of wide residual blocks. As a result, the number of filters in convolutional layers can be increased.
- **Squeeze and Excitation Network (SENet)**: SENet [9] consists of specially designed blocks termed as squeeze and excitation (SE) units to enhance the representation power of the features generated by CNN. The advantage of SENet is that with a negligible rise in total number of parameters to be learned, it can greatly enhance the overall performance of the network. To any of the convolutional layers, we can incorporate SE units, and the squeeze block combines the feature maps generated along each channels in a particular layer, whereas the excitation block involves fully connected layers that accepts the output of the squeeze blocks as input and generates a set of weight corresponding to each channel.
- **Convolutional Block Attention Module (CBAM)**: Similar to SENet, the CBAM [16] is a way to enhance the representation ability of CNN architecture. In CBAM, there exist two separate blocks channel attention and spatial attention. These two modules can be applied in sequential order to each layers of the CNN architecture.

3.2 The Proposed Deep Meta-model

Once the predictions were generated from the above-mentioned base level models, we need a mechanism to efficiently aggregate the results yielded by these base level models to get a consensus final output. Stacking is a prominent technique used in the literature to efficiently combine the results generated by the base level models. In stacking, a meta-model is initially trained with the predictions obtained from the

base level models as inputs and the expected target as the required output. As a result, the meta-model learns how to effectively combine the initial level predictions made by the base models.

This paper proposes a deep meta-model for combining the predictions from the base level models. Initially, the original dataset D is divided into five groups, $\{D_1, D_2, D_3, D_4, D_5\}$, where each $D_i = \{f_{ij}, t_{ij}\}_{j=1}^{n}$ involves n labeled samples taken from the given training collection. In the first pass of the proposed algorithm, the set $\{D_2 \cup D_3 \cup D_4 \cup D_5\}$ is used as the training data for the proposed deep meta-model, and the set $D_1 = \{f_{1j}, y_{1j}\}_{j=1}^{n}$ is kept aside as the test samples. Given an input sample x_1, the base models will generate predictions, and these predictions are then appended with the target class t_1 corresponding to x_1 to form a feature vector of the following form: $[h_1(x_1), h_2(x_1), h_3(x_1), h_4(x_1), h_5(x_1)]$. This newly generated feature vector acts as the input for the proposed deep meta-model.

The deep meta-model that aggregates the base level predictions is formulated as a deep neural network. A deep neural network with five hidden layers is trained with the above-mentioned feature vectors to generate the final prediction score as the combination of the prediction scores given by the base level models. Given the set of base level predictions and the target class label, the deep meta-model can adjust the weights of the hidden layers such that it can learn a non-linear function approximation between the base level predictions and their aggregate. The key benefit of the proposed stacking scheme is that, instead of learning a linear relationship among the base level predictions, it can learn complex non-linear associations. Thus, the proposed deep meta-model always guarantees a consensus prediction score.

4 Experimental Evaluation

This section delineates a detailed description of the datasets used, the performance measures employed, and the discussion on the results obtained by the proposed model for the task of automatic recognition of environmental sounds and its comparative evaluation.

4.1 Dataset Used

To evaluate the performance of the proposed deep meta-model for ESR, two benchmark datasets such as ESC-50 and the UrbanSound 8 K have been utilized in our experiments. The ESC-50 dataset involves audio recordings of 2000 environmental sounds which are labeled. Each of these sound recordings are of duration 5 seconds and are grouped into 50 different categories. On the other hand, the UrbanSound 8K is an audio dataset which consists of 8732 labeled sound recordings of 10 semantic categories with duration equal to 4 seconds.

4.2 Performance Measures Used for Evaluation

The following performance measures have been used in this work to demonstrate the efficiency of the proposed deep meta-model:

- False Positive Rate (FPR): It quantifies the chance that the proposed model categorizes a negative sample as positive.
- False Negative Rate (FNR): It quantifies the extend by which the proposed model missed out true positives.
- True Positive Rate (TPR): It measures the potential of the proposed model to predict a positive sample as positive.
- True Negative Rate (TNR): It measures the potential of the proposed model to categorize a negative sample as negative.

In addition to these quantitative measures, we also used receiver operating characteristic (ROC) curve to qualitatively evaluate the proposed ESR system. The ROC curve depicts the relationship between TPR and FPR of the proposed model at different classification thresholds.

4.3 Results and Discussion

This section evaluates the results obtained with the proposed deep meta-model in comparison with the sate-of-the-art approaches. All the simulations discussed in this section were conducted on a machine equipped with Intel i7 processor having 8 GB of RAM and Ubuntu as operating system.

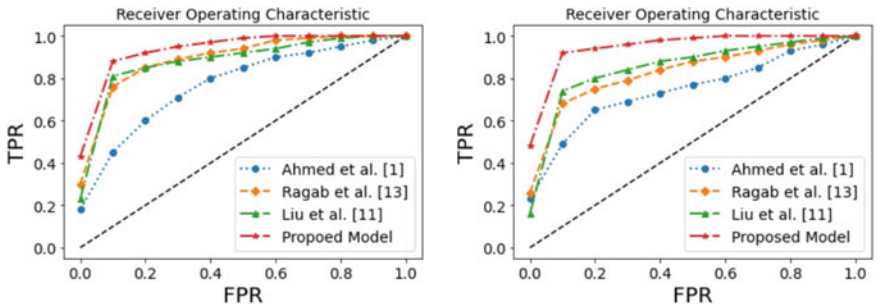
The results obtained with the proposed deep meta-model were compared with the models coined by Ahmed et al. [1], Ragab et al. [12], and Liu et al. [11]. The hardware setup which was used to simulate the proposed deep meta-model is also used to implement the state-of-the-art models. Tables 1 and 2 summarize the comparison results on the selected benchmark datasets. From these results, it is evident that the proposed deep meta-model exhibits better performance in comparison with the existing approaches for the task of ESR.

Table 1 Comparison of the proposed deep meta-model with existing schemes on ESC-50 dataset

Performance metric	Proposed deep meta-model	Ahmed et al. [1]	Ragab et al. [12]	Liu et al. [11]
TPR %	98.5	94.90	92.67	90.82
FPR %	0	10.33	12.44	14.61
FNR %	0	7.67	9.28	11.58
TNR %	98	93.80	91.65	89.81

Table 2 Comparison of the proposed deep meta-model with existing schemes on UrbanSound 8K dataset

Performance metric	Proposed deep meta-model	Ahmed et al. [1]	Ragab et al. [12]	Liu et al. [11]
TPR %	98	92.82	90.45	88.58
FPR %	0	12.23	14.35	16.31
FNR %	0	9.75	11.63	13.44
TNR %	97	91.64	89.44	87.22



(a) ESC-50 Dataset (b) UrbanSound 8K Dataset

Fig. 2 Comparative evaluation of the proposed model based on ROC curve

In addition to this, the performance of the proposed model in comparison with the state-of-the-art approaches is evaluated on the basis of ROC curve. Figure 2 depicts the ROC curves obtained for the proposed model in comparison with the existing stacking schemes when applied on ESC-50 and the UrbanSound 8K datasets. By analyzing the results shown in Fig. 2, it can be inferred that the proposed deep neural network-based stacking scheme gives rise to higher area under curve scores. This is an indication of the fact that the proposed deep meta-model can assure better recognition performance as compared to the existing stacking-based approaches for ESR.

5 Conclusion

The objective of this work was to investigate the applicability of the proposed deep meta-model to categorize environmental sounds based on the spectrograms of audio signals. The proposed deep meta-model utilizes the prediction scores of diverse base level models as input to a deep neural network and adjusts its weights so as to combine the initial level predictions to form a final consensus result. We analyzed two different environmental sound datasets on five distinct base level models and the proposed deep meta-model. The experimental results demonstrated the fact that the

proposed deep meta-model outperforms all the base level models and the existing stacking approaches in several evaluation metrics. In addition to this, it was evident that the proposed deep meta-model was able to automatically encode the complex interrelation among the base level models and thus to entitle superior prediction.

References

1. Ahmed M, Robin TI, Shafin AA et al (2020) Automatic environmental sound recognition (AESR) using convolutional neural network. Int J Mod Educ Comput Sci **12**(5)
2. Arun KS, Govindan VK (2015) Optimizing visual dictionaries for effective image retrieval. Int J Multim Inf Retr 4(3):165–185
3. Arun KS, Govindan VK, Kumar SDM (2017) On integrating re-ranking and rank list fusion techniques for image retrieval. Int J Data Sci Anal 4(1):53–81
4. Arun KS, Sarath KS (2010) Evaluation of the role of low level and high level features in content based medical image retrieval. In: International conference on advances in information and communication technologies, Springer, pp 319–325
5. Demir F, Abdullah DA, Sengur A (2020) A new deep CNN model for environmental sound classification. IEEE Access 8:66529–66537
6. Demir F, Turkoglu M, Aslan M, Sengur A (2020) A new pyramidal concatenated CNN approach for environmental sound classification. Appl Acoust 170:107520
7. Dietterich TG (2000) Ensemble methods in machine learning. In: International workshop on multiple classifier systems. Springer, pp 1–15
8. Guzhov A, Raue F, Hees J, Dengel A (2021) ESResNet: environmental sound classification based on visual domain models. In: 2020 25th International conference on pattern recognition (ICPR). IEEE, pp 4933–4940
9. Hu J, Shen L, Sun G (2018) Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7132–7141
10. Larsson G, Maire M, Shakhnarovich G (2016) FractalNet: ultra-deep neural networks without residuals. arXiv:1605.07648
11. Liu C, Hong F, Feng H, Zhai Y, Chen Y (2021) Environmental sound classification based on stacked concatenated DNN using aggregated features. J Signal Process Syst 1–13
12. Ragab MG, Abdulkadir SJ, Aziz N, Alhussian H, Bala A, Alqushaibi A (2021) An ensemble one dimensional convolutional neural network with Bayesian optimization for environmental sound classification. Appl Sci 11(10):4660
13. Skariah SM, Arun KS (2021) A deep learning based approach for automated diabetic retinopathy detection and grading. In: 2021 4th Biennial international conference on Nascent Technologies in engineering (ICNTE). IEEE (2021)
14. Stastny J, Munk M, Juranek L (2018) Automatic bird species recognition based on birds vocalization. EURASIP J Audio Speech Music Process 2018(1):1–7
15. Tang B, Li Y, Li X, Xu L, Yan Y, Yang Q (2019) Deep CNN framework for environmental sound classification using weighting filters. In: 2019 IEEE international conference on mechatronics and automation (ICMA), IEEE, pp 2297–2302
16. Woo S, Park J, Lee JY, Kweon IS (2018) CBAM: Convolutional block attention module. In: Proceedings of the European conference on computer vision (ECCV), pp 3–19
17. Xie S, Girshick R, Dollár P, Tu Z, He K (2017) Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1492–1500
18. Zagoruyko S, Komodakis N (2016) Wide residual networks. arXiv:1605.07146
19. Zhang Z, Xu S, Zhang S, Qiao T, Cao S (2021) Attention based convolutional recurrent neural network for environmental sound classification. Neurocomputing 453:896–903