Multi-channel Convolutional Neural Networks with Multi-level Feature Fusion for Environmental Sound Classification

Dading Chong[1], Yuexian Zou[1,2](✉), and Wenwu Wang[3]

[1] ADSPLAB, School of ECE, Peking University, Shenzhen, China
zouyx@pkusz.edu.cn
[2] Peng Cheng Laboratory, Shenzhen, China
[3] Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford, UK

Abstract. Learning acoustic models directly from the raw waveform is an effective method for Environmental Sound Classification (ESC) where sound events often exhibit vast diversity in temporal scales. Convolutional neural networks (CNNs) based ESC methods have achieved the state-of-the-art results. However, their performance is affected significantly by the number of convolutional layers used and the choice of the kernel size in the first convolutional layer. In addition, most existing studies have ignored the ability of CNNs to learn hierarchical features from environmental sounds. Motivated by these findings, in this paper, parallel convolutional filters with different sizes in the first convolutional layer are designed to extract multi-time resolution features aiming at enhancing feature representation. Inspired by VGG networks, we build our deep CNNs by stacking 1-D convolutional layers using very small filters except for the first layer. Finally, we extend the model using multi-level feature aggregation technique to boost the classification performance. The experimental results on Urbansound 8k, ESC-50, and ESC-10 show that our proposed method outperforms the state-of-the-art end-to-end methods for environmental sound classification in terms of the classification accuracy.

Keywords: Environmental sound classification
Multi-channel deep convolutional neural networks · End-to-end
Multi-level feature fusion

1 Introduction

Environmental sound classification (ESC) is an important research area in human-computer interaction with a variety of applications such as abnormal sound detection in security surveillance. There are many research outcomes of ESC in the last decade [1]. However, with the limit amount of publicly available research datasets, ESC is still an open and difficult challenge.

Traditional ESC methods are based on hand-craft features, such as zero-crossing, mel-frequency cepstral coefficients (MFCCs) [2], and mel-filterbank features [3], and traditional classifiers such as Random forest, support vector machines, and Gaussian

I. Kompatsiaris et al. (Eds.): MMM 2019, LNCS 11296, pp. 157–168, 2019.
https://doi.org/10.1007/978-3-030-05716-9_13

mixture models [4–7]. However, the performance of all the feature based methods highly depends on the representation ability of these hand-craft features. In recent years, deep learning has gained incredible popularity [8–10]. Among them, the convolutional neural network is regarded as a powerful method, due to its ability in learning hierarchical high-level representations from sound data. CNNs have been applied to sound event recognition in two different ways. The first approach is to use the CNNs as the classifier with MFCCs or log-mel features as the input [3, 11–14]. The second approach use CNNs to extract salient and discriminative features from raw wave signals for ESC [6, 15–17].

Current approaches have the following limitations. (1) Since most of the features were originally designed for Automatic Speech Recognition (ASR) rather than for the ESC task, it may fail to capture the intrinsic information from the environmental sounds that may be critical for classification; (2) The feature extraction stage is separated from classification stage, as a result, the designed feature may not be optimal for the classification task. (3) The existing CNN models often use 2D convolution which involves more parameters as compared with 1D convolution. (4) The feature extracted from 1-D convolutional layer with a fixed size might be insufficient for building high-level discriminative representations.

In this study, we present a multi-scale deep convolutional neural network architecture for ESC that is able to address these issues. Our architecture allows one to develop much deeper and more “complex” structure while using a model of small size. The proposed method consists of ten 1-D convolutional layers and multiple filters with different sizes in the first 1-D convolutional layer which are learned simultaneously. Finally, multi-level feature fusion is proposed to make full use of hierarchical features extracted from deep CNNs.

Our main contribution can be summarized as follows:

- We propose an end-to-end accurate and efficient methods for ESC based on deep convolutional neural networks.
- We design parallel CNNs to learn richer representation from raw waveforms.
- Comparatively studies are conducted to demonstrate the effect of multi-level features on the classification performance.
- Without full connection layers, our proposed method reaches comparable classification performance but with a much smaller model size and much faster speed for environmental sound classification.

The paper is organized as follows. In the next section, related works are introduced briefly for presentation clarity. Problem definition and the proposed ESC system and its subsystems with implementation details are given in Sect. 3. Experimental setup and results will be shown in Sect. 4 and conclusion is drawn in Sect. 5.

2 Related Work

In recent years, CNNs have led to impressive results in ESC tasks, thanks to its ability in automatically learning complex feature representations with its convolutional layers. Conventional feature, such as spectrograms, MFCCs [2], mel-filterbank features [3], are the most commom inputs for CNN-based architectures. This method was firstly proposed by Piczak [3] for ESC task. Where log-mel and delta log-mel (i.e. first temporal derivative) features are extracted in each frame. Then, the two-dimensional feature map constituted by static log-mel and delta log-mel is fed into a two-channel CNNs for classification. This increased the classification accuracy over the traditional methods by 13% on the ESC-50 dataset.

Meanwhile, motivated by Hoshen et al. [18], where the first 1-D convolutional layer is taken as a finite impulse response filter bank, many attempts have been made to learn features automatically from raw waveforms for ESC. Tokozume and Harada [15] proposed an end-to-end ESC method using two convolution layers to classify environmental sound. The accuracy of their method was 5.1% higher than the model using log-mel features. Based on this research, Dai et al. [17] proposed to use deep convolutional neural networks (DCNNs) to extract more discriminative features, and the DCNN model was shown to outperform the shallow convolutional neural network model.

Inspired by recent advance in end-to-end ESC methods, we proposed an end-to-end method based on multi-channel deep convolutional neural networks (MC-DCNNs) to further improve the performance of ESC task. Below we present our method in detail.

3 Methodology

We formally define the aforementioned ESC problem. Then the pipeline and algorithms of our proposed methods are described in detail. Our MC-DCNNs ESC model is shown in Fig. 1. In Sect. 3.1, we gave a brief discussion on the problem formulation of the ESC task. In Sect. 3.2, we show the architecture of multi-channel CNNs used in our system. In Sect. 3.3, a multi-level feature fusion module is proposed to enhance the feature representation. The architecture and parameter settings are detailed in Sect. 3.4.

3.1 Problem Definition

Recognition of environmental sound from raw waveform can be considered as a classical learning problem of estimating an unknown relationship between the elements from input feature space to the corresponding elements in the target space. A time series is a sequence of real-valued data points with timestamps. In this paper, we define $S = \{(X_{i,}\ Y_i)|X_i \in R_l, Y_i \in Z; i = 1, \cdots, N\}$ with $X_i = [X_{l,}\ \cdots, X_i]^{\ T}$ as input vector formulated by considering the raw waveform that belongs to different environmental sounds; l is the length of the environmental sound. The elements in target space are the class labels to which the corresponding elements input feature space $X = [x_1, x_2, \cdots, x_N]$. The ESC problem is to build a classification model to predict a class label $y \in Y_i$ given an input sound data X.

3.2 Multi-channel Convolution Operation

Convolution has been a well-established method for modeling time series signals [19]. Suppose the waveform signal $w(i)$ is convolved with filters $h_{(s)}$ at different scales. The results of $x_j^{(s)}(n)$ are given by

$$x_j^{(s)}(n) = f(\sum_{i=0}^{N-1} w(i) h_j^s(n-i) + b_j^s)$$

where *f* is an activation function, *N* is the length of *w(i)*, and b_j is an additive bias.

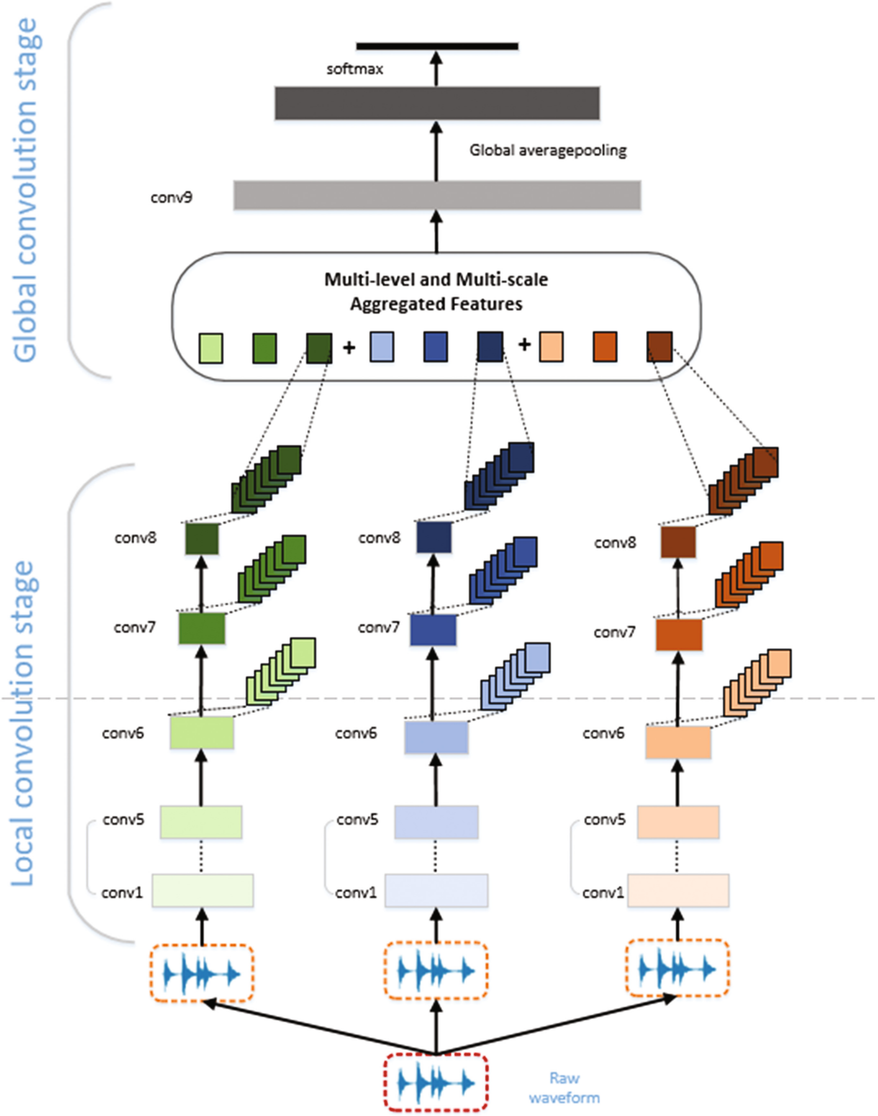

Fig. 1. The overall framework of the multi-channel deep convolutional neural networks environmental sound classification system

Using filters of different size, the convolution is capable of extracting multi-temporal resolution features from the raw waveform. For example, 40, 80 and 160 samples correspond to 2.5, 5 and 10 ms at 16 kHz sampling rate. In our proposed method, three different kernel sizes with same stride are chosen representatively in the first 1-D convolution layer to learn multi-temporal resolution features with the same dimension. Each convolutional layer with filters is functionally similar to a bandpass filter bank. Each layer has 64 filters, generating 64 one-dimensional vectors. The receptive field is different due to different kernel size used in the first convolution layers, and the stride is set as 8 in time series. The filter size is set according to the experimental results of DCNNs [17], which demonstrated that the CNN model can extract local features of various time scales hierarchically by using parallel convolutional layers. The input shape and the parameters of the first convolutional layers are as in Table 1.

Table 1. Parameters of the first 1-D convolutional layer in MC-DCNNs

Layers	Input shape	Filter	Kernel size	Stride	Output shape
Conv1	[1,64000,1]	64	[40], [80], [160]	8	[1,8000,64]
Pool	[1,8000,1]	–	Pooling size = 4	4	[1,2000,64]

We apply no-overlapping max pooling to the output of these parallel three convolutional layers with a pooling size of 4. Then, the three outputs of the pool matrix is of size 1 × 2000 × 64. Finally, these vectors are fed into the multi-channel convolutional filters to extract local features.

3.3 Multi-scale and Multi-level Feature Fusion

Using a single CNN model, different levels of features can be extracted, therefore, we establish parallel CNNs with different kernel sizes to extract multi-level and multi-scale features. Benefiting from the structure, we can extract multi-temporal resolution and multi-level features. To further enhance the convolutional features and improve the performance of the ESC task, we generate a global representation by aggregating the multi-level features from single CNNs into a single large feature vector. Empirically, for the ESC task, when the neural network goes deeper, its ability in catching global information will get better. We also noted that the features in lower layers tend to focus more on detailed information which can be beneficial for short environmental sounds. Meanwhile, the higher layers tend to contain more abstract semantic information which would be useful for capturing global information related to acoustic scenes. Therefore, we design a multi-level feature fusion module to balance the local and global information captured by features in lower and higher layers in the MC-DCNNs.

3.4 Architecture of MC-DCNNs

According to the above discussion, we propose the multi-scale deep convolutional neural network with feature fusion as shown in Fig. 1.

Firstly, we apply the parallel convolutional operation with different size of convolutional kernels on the input waveform. Three sizes are chosen as small receptive field model (SRF), middle receptive field model (MRF), and large receptive field model (LRF). Then we progressively reduce the temporal resolution to 2000 with a non-overlapping max pooling layer to each feature map. Three channel CNNs are applied on the three kinds of features. In the local convolutional stage, the raw waveform is provided to 1-D local convolution filters followed by max pooling to extract the local and independent features. The filter has the same size for local convolution in MC-DCNNs. By selecting the same filter size and performing down sampling with Maxpooling, each channel CNNs in local convolution stage captures features from different receptive fields. We also adopt batch normalization (BN) [20] after each convolutional layer. In the global convolutional stage, features extracted from each channel are locally and globally concatenated. Finally, similar to local convolution stage, 1-D convolutions is performed followed by the Global average pooling. The features obtained after the global convolution stage are provided to Softmax to predict the labels. In this work, we adopted the deep concatenation technique in [21] to concatenate all the feature maps vertically. The details of parameters are summarized in Table 2.

Table 2. Detailed parameters of proposed fully convolutional network for time-domain waveform inputs. [40/8, 64] denotes a convolutional layer with receptive field 40 and 64 filters, with stride 8. […] × *k* denotes *k* stacked layers. Output size after each pooling is written as *m* × *n* where *m* is the size in time-domain and *n* is the number of feature maps. Maxpooling is also done with the feature get from conv6-7 to get the same dimension.

MC-DCNNs model for ESC		
Input: 64000 Samples(4s) as Inputs (e.g. Urbansound 8k)		
Conv1 [40/8,64]	Conv1 [80/8,64]	Conv1 [160/8,64]
Maxpooling: 4 × 1 (output: 2000 × 64)	Maxpooling: 4 × 1 (output: 2000 × 64)	Maxpooling: 4 × 1 (output: 2000 × 64)
Conv2-3 [3/1,64] ×2	Conv2-4 [3/1,64] ×2	Conv2-4 [3/1,64] ×2
Maxpooling: 4 × 1 (output: 500 × 64)	Maxpooling: 4 × 1 (output: 500 × 64)	Maxpooling: 4 × 1 (output: 500 × 64)
Conv4-5 [3/1, 128] ×2	Conv4-5 [3/1, 128] ×2	Conv4-5 [3/1, 128] ×2
Maxpooling: 4 × 1 (output: 125 × 128)	Maxpooling: 4 × 1 (output: 125 × 128)	Maxpooling: 4 × 1 (output: 125 × 128)
Conv6-7 [3/1, 256] ×2	Conv6-7 [3/1, 256] ×2	Conv6-7 [3/1, 256] ×2
Maxpooling: 4 × 1 (output: 32 × 256)	Maxpooling: 4 × 1 (output: 32 × 256)	Maxpooling: 4 × 1 (output: 32 × 256)
Conv8 [3/1, 512] ×1	Conv8 [3/1, 512] ×1	Conv8 [3/1, 512] ×1
Conv6-8concat (output: 32 × 1024)	Conv6-8Concat (output: 32 × 1024)	Conv6-8Concat (output: 32 × 1024)
Global concat (output: 32 × 3072)		
Conv9 [3/1, 512] ×1 (output: 32 × 512)		
Globalaverage pooling (1 × 512)		
Softmax		

4 Experiments and Discussion

In this section, we first provide a brief description about the Urbansound 8K [22], ESC-50 [23], and ESC-10 [23] datasets and the implementation procedure for the evaluation of MC-DCNNs. Performance comparison will also be provided in this section.

4.1 Datasets

Urbansound 8K contains of 10 types of environmental sounds in urban areas, such as engine idling, street music, and children playing. The dataset consists of 8732 audio clips of 4 s or less, in total 9.7 h. We use the official fold 10 to be our test set, and the remaining folds for training.

ESC-50 consists of 50 environmental sounds categories that are allocated into 5-folds with 40 samples per category. The 50 classes can be divided into 5 major groups: animals, natural soundscapes and water sounds, human non-speech sound, interior/domestic sounds, and exterior/urban noises. The dataset provides an exposure to a variety of sound sources, some very common (laughter, cat meowing, dog barking), some quite distinct (glass breaking, brushing teeth) and then some where the differences are more nuanced (helicopter and airplane noise). We use the official fold 5 to be our test set, and the remaining folds for training.

ESC-10 dataset is a subset of ESC-50, which contains 400 recordings divided equally into 10 categories: dog barking, rain, sea waves, baby crying, clock ticking, person sneezing, helicopter, chainsaw, rooster, and file crackling. We use the official fold 5 to be our test set, and the remaining folds for training.

4.2 Data Preparation and Implementation Details

We convert all sound files to monaural wav files with a sampling rate of 16 kHz. Differently from other standard methods, we did not remove the silent section from the whole 5s sound to preserve the integrity of the original audio in ESC-50 and ESC-10 datasets. All data are normalized to zero mean and unit variance in Urbansound 8k ESC-50 and ESC-10. The length of each sound segment is 40000 samples (corresponding to 2.5s raw waveform) in ESC-50 and ESC-10. We take the 64000 samples (corresponding to 4s raw waveform) as inputs in Urbansound 8k. In the training stage, we randomly select these segments from the original training audio and input them into the prediction models. In the test phase, we perform a majority voting of the output prediction results for classification.

Hyper parameters selection: The list of MC-DCNNs hyper parameters and the initialization used in this work are listed in Table 3.

4.3 Results

We compared our MC-DCNNs model with existing log-mel-CNN models and end-to-end based models such as reported by Piczak [3], Tokozume and Harada [15] and Dai

Table 3. The MC-DCNNs hyper parameters and their initialization

Parameter	Initialization values
Activation function	Leakyrelu (α = 0.2) [24]
Optimizer	Adam [25]
Learning rate	0.001 with weight decay
Batch-size	32
Regularization	L2 regularization with coefficient 0.0001
Parameter initialization	He initialization
Epochs	400
Loss function	Categorical cross-entropy

et al. [17], Salamon and Bello [22]. Table 4 shows the results of the proposed MC-DCNNs on three datasets with previous state-of-the-art published methods.

First of all, with the ESC-50 dataset, our model achieves 73.5% classification accuracy, which is much higher than other end-to-end based ESC models and a little higher than the static-delta log-mel-CNN+BN methods proposed by Tokozume and Harada [15]. On the ESC-10 dataset, the classification accuracy of MC-DCNNs reaches 87.5%, which is 9.7% and 3.5% higher than the other two end-to-end based method respectively. Finally, we evaluated the algorithms on the Urbansound 8k dataset.

Table 4. Comparison of classification accuracy with other models on evaluated datasets.

Accuracy (%) on Dataset					
Model	Feature	Fusion	ESC-50	ESC-10	Urbansound 8k
Logmel-EnvNet [15]	log-mel	–	66.9 ± 3.1	79.8 ± 1.7	–
SB-CNN [26]	log-mel	–	71.0 ± 1.4	–	73.9 ± 0.4
Logmel+CNN [3]	log-mel	–	64.5 ± 0.9	81.5 ± 1.3	–
Logmel+CNN +BN [15]	log-mel	–	72.4 ± 1.7	–	72.7
D-CNNs [13]	log-mel	–	68.0 ± 1.4	84.6 ± 2.1	–
EnvNet [15]	Raw waveform	–	64.0 ± 2.4	77.6 ± 2.3	69.2 ± 0.8
EnvNet2 [16]	Raw waveform	–	71.6 ± 2.7	83.2 ± 1.5	74.2 ± 1.5
M18 [17]	Raw waveform	–	–	–	71.68
MC-DCNNs (this paper)	Raw waveform	–	71.1 ± 0.8	84.1 ± 0.7	73.6 ± 0.8
MC-DCNNs (this paper)	Raw waveform	Multi-level fusion	73.1 ± 1.1	87.6 ± 1.3	75.1 ± 0.6

The accuracy of our proposed methods is 75.1%, much higher than the 71.68% accuracy achieved by Dai et al. [17] using the DCNNs based methods. These results on these three datasets indicate that our multi-channel deep convolutional neural networks with multi-level feature fusion have achieved significant improvement in environmental sound classification with the raw waveform as input, offering state-of-the-art performance in environmental sound classification.

Furthermore, we did a comparative experiment, as shown in Fig. 2. It can be seen that the multi-channel model consistently outperforms the single-channel model. It confirms the effectiveness of our proposed multi-channel model.

We further compared the accuracy of the MC-DCNNs with or without multi-level feature fusion on ESC-50. We noticed that the classification accuracy of the combination model increases on short duration environmental sounds compared with D-CNNs and human non-speech in contrast with log-mel feature based methods. The D-CNN model may lose detailed information while extracting global information. However, in our MC-DCNNs method, benefiting from the features extracted from former layers, the classification accuracy increased significantly. Similarly, for human non-speech sound, the accuracy has been improved, which indicates that some features in the human non-speech sound can be ignored by log-mel features. Using the MC-DCNNs, we can make full use of features learned from raw waveform, so that the classification accuracy can be improved via multi-level information fusion.

At the same time, benefited from the peculiarity of FCNs, the model size of MC-DCNNs is only 1.8M. Besides, our MC-DCNNs can classify two minutes of audio per second on NVIDIA GTX 1080, which is able to perform real-time environmental sound classification and is of great value in practical applications.

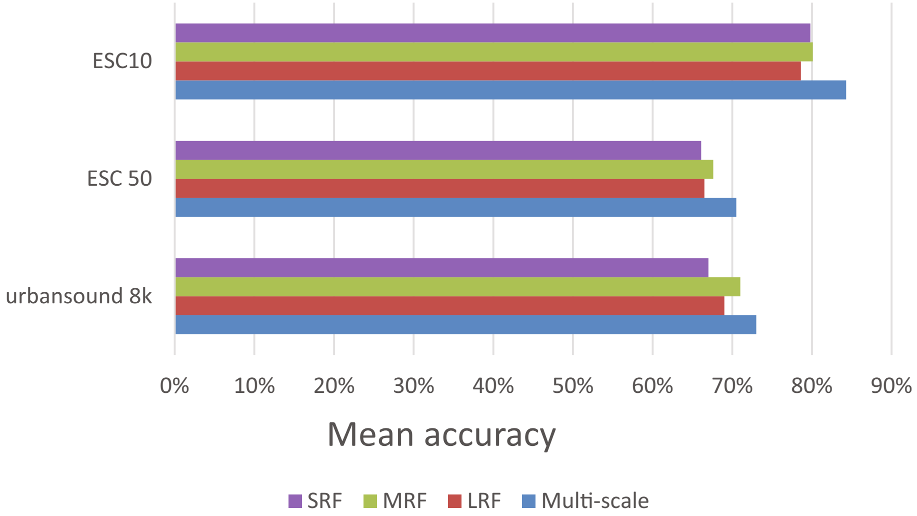

Fig. 2. Effectiveness of multi-channel models. We compare multi-channel with single-channel with SRF, MRF, LRF on ESC-10, ESC-50 and Urbansound 8K, The classification of Multi-channel models are superior to single-channel.

5 Conclusion

We presented a new end-to-end environmental sound classification system MC-DCNNs, which is composed of three channel stacked convolutional neural networks, trained on raw waveform as input. Each channel is composed of eight 1-D convolutional layers with batch normalization [20]. Three public datasets, Urbansound 8K, ESC-50, and ESC-10, are used to evaluate the classification performance of the MC-DCNNs model. The classification accuracy of the MC-DCNNs is 75.1%, 73.5%, 87.5%, on Urbansound 8K, ESC-50, and ESC-10, respectively, which is 1.1%, 6%, and 2.4% higher than existing log-mel feature based methods. It is also 1.5%, 4.4%, and 0.9% higher than the state-of-the-art end-to-end methods. These results showed that our MC-DCNNs are more effective for environmental sound classification due to the exploitation of both global and local features. While we have achieved excellent classification accuracy, our method also has the advantage in small model size and real-time classification performance. Future work will consider different structures of convolution neural networks for ESC task, such as convolutional recurrent neural networks. We will also consider applying the MC-DCNNs to time-series signal other than environmental sounds.

Acknowledgment. This project was partially supported by Shenzhen Science & Technology Fundamental Research Programs (No: JCYJ20170817160058246 and JCYJ20170306165153653) & Shenzhen Key Laboratory for Intelligent Multimedia and Virtual Reality (ZDSYS201703031405467). Special acknowledgements are given to Aoto-PKUSZ Joint Research Center of Artificial Intelligence on Scene Cognition & Technology Innovation for its support.

References

1. Virtanen, T., Plumbley, M.D., Ellis, D.: Computational Analysis of Sound Scenes and Events. Springer, Heidelberg (2018). https://doi.org/10.1007/978-3-319-63450-0
2. Boddapati, V., Petef, A., Rasmusson, J., Lundberg, L.: Classifying environmental sounds using image recognition networks. Proc. Comput. Sci. **112**, 2048–2056 (2017)
3. Piczak, K.J.: Environmental sound classification with convolutional neural networks. In: 2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP), pp. 1–6. IEEE (2015)
4. Vacher, M., Serignat, J.-F., Chaillol, S.: Sound classification in a smart room environment: an approach using GMM and HMM methods. In: The 4th IEEE Conference on Speech Technology and Human-Computer Dialogue (SpeD 2007), Publishing House of the Romanian Academy (Bucharest), pp. 135–146 (2007)
5. Łopatka, K., Zwan, P., Czyżewski, A.: Dangerous sound event recognition using support vector machine classifiers. In: Nguyen, N.T., Zgrzywa, A., Czyżewski, A. (eds.) Advances in Multimedia and Network Information System Technologies, pp. 49–57. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-14989-4_5

6. Su, F., Yang, L., Lu, T., Wang, G.: Environmental sound classification for scene recognition using local discriminant bases and HMM. In: Proceedings of the 19th ACM International Conference on Multimedia, pp. 1389–1392. ACM (2011)
7. Saki, F., Kehtarnavaz, N.: Background noise classification using random forest tree classifier for cochlear implant applications. In: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 3591–3595. IEEE (2014)
8. Sainath, T.N., Mohamed, A.-R., Kingsbury, B., Ramabhadran, B.: Deep convolutional neural networks for LVCSR. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 8614–8618. IEEE (2013)
9. Abdel-Hamid, O., Mohamed, A.-R., Jiang, H., Deng, L., Penn, G., Yu, D.: Convolutional neural networks for speech recognition. IEEE/ACM Trans. Audio Speech Lang. Process. **22**, 1533–1545 (2014)
10. Kong, Q., Sobieraj, I., Wang, W., Plumbley, M.: Deep neural network baseline for DCASE challenge 2016. In: Proceedings of DCASE 2016 (2016)
11. Cotton, C.V., Ellis, D.P.: Spectral vs. spectro-temporal features for acoustic event detection. In: 2011 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), pp. 69–72. IEEE (2011)
12. Zhang, H., McLoughlin, I., Song, Y.: Robust sound event recognition using convolutional neural networks. In: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 559–563. IEEE (2015)
13. Zhang, X., Zou, Y., Shi, W.: Dilated convolution neural network with LeakyReLU for environmental sound classification. In: 2017 22nd International Conference on Digital Signal Processing (DSP), pp. 1–5. IEEE (2017)
14. Medhat, F., Chesmore, D., Robinson, J.: Masked conditional neural networks for audio classification. In: Lintas, A., Rovetta, S., Verschure, P.F.M.J., Villa, A.E.P. (eds.) ICANN 2017. LNCS, vol. 10614, pp. 349–358. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-68612-7_40
15. Tokozume, Y., Harada, T.: Learning environmental sounds with end-to-end convolutional neural network. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2721–2725. IEEE (2017)
16. Tokozume, Y., Ushiku, Y., Harada, T.: Learning from between-class examples for deep sound recognition. arXiv preprint arXiv:1711.10282 (2017)
17. Dai, W., Dai, C., Qu, S., Li, J., Das, S.: Very deep convolutional neural networks for raw waveforms. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 421–425. IEEE (2017)
18. Hoshen, Y., Weiss, R.J., Wilson, K.W.: Speech acoustic modeling from raw multichannel waveforms. In: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4624–4628. IEEE (2015)
19. Zhao, B., Lu, H., Chen, S., Liu, J., Wu, D.: Convolutional neural networks for time series classification. J. Syst. Eng. Electron. **28**, 162–169 (2017)
20. Ioffe, S., Szegedy, C.: Batch normalization: accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167 (2015)
21. Lee, J., Nam, J.: Multi-level and multi-scale feature aggregation using pretrained convolutional neural networks for music auto-tagging. IEEE Signal Process. Lett. **24**, 1208–1212 (2017)
22. 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, pp. 1041–1044. ACM (2014)

23. Piczak, K.J.: ESC: dataset for environmental sound classification. In: Proceedings of the 23rd ACM International Conference on Multimedia, pp. 1015–1018. ACM (2015)
24. Maas, A.L., Hannun, A.Y., Ng, A.Y.: Rectifier nonlinearities improve neural network acoustic models. In: Proceedings of ICML, p. 3 (2013)
25. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
26. Salamon, J., Bello, J.P.: Deep convolutional neural networks and data augmentation for environmental sound classification. IEEE Signal Process. Lett. **24**, 279–283 (2017)