INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Environmental Sound Classification with Parallel Temporal-spectral Attention

Helin Wang[1], *Yuexian Zou*[1,2,*], *Dading Chong*[1], *Wenwu Wang*[3]

[1]ADSPLAB, School of ECE, Peking University, Shenzhen, China
[2]Peng Cheng Laboratory, Shenzhen, China
[3]Center for Vision, Speech and Signal Processing, University of Surrey, UK
{wanghl15,zouyx,1601213984}@pku.edu.cn, w.wang@surrey.ac.uk

Abstract

Convolutional neural networks (CNN) are one of the best-performing neural network architectures for environmental sound classification (ESC). Recently, temporal attention mechanisms have been used in CNN to capture the useful information from the relevant time frames for audio classification, especially for weakly labelled data where the onset and offset times of the sound events are not applied. In these methods, however, the inherent spectral characteristics and variations are not explicitly exploited when obtaining the deep features. In this paper, we propose a novel parallel temporal-spectral attention mechanism for CNN to learn discriminative sound representations, which enhances the temporal and spectral features by capturing the importance of different time frames and frequency bands. Parallel branches are constructed to allow temporal attention and spectral attention to be applied respectively in order to mitigate interference from the segments without the presence of sound events. The experiments on three environmental sound classification (ESC) datasets and two acoustic scene classification (ASC) datasets show that our method improves the classification performance and also exhibits robustness to noise.

Index Terms: environmental sound classification, convolutional neural networks, attention mechanism, sound event

1. Introduction

Environmental sound classification (ESC) is an important research area in human-computer interaction aiming to classify an environment by its ambient sound, with a variety of potential applications such as audio surveillance [1] and smart room monitoring [2]. Due to the dynamic and unstructured nature of acoustic environments, it is a practical challenge to design appropriate features for environmental sound classification. In many existing ESC methods, the features are often designed based on prior knowledge of acoustic environments, and a classifier is then trained with the features to obtain the category probability of each environmental sound signal.

Among these methods, deep learning, which is facilitated by the availability of increased amount of training data and techniques of data augmentation, has been widely used in ESC. Convolutional neural networks (CNN) based methods [3–8] offer the state-of-the-art performance, where spectrograms and mel-scale frequency cepstral coefficients (MFCC) are often used as the input of the networks. Different from the images in visual recognition tasks, however, the temporal and spectral information represented by spectrograms will have different characteristics and degree of importance in sound recognition. Although the translation of local patterns in the time domain has little effect on the classification of sound events, the difference across frequency bands has a significant impact on the performance of sound classification [9]. To capture the information about which parts of the features are more relevant to the sound events, attention mechanisms have been proposed [10–17], especially for weakly labelled data where the timing information about the sound events is not available in the training data. In these methods, temporal attention [11, 14] is applied to obtain weights for combining feature vectors at different time steps, however, the importance of different frequency bands is not considered. Spatial attention [17] characterizes the importance of the regions with spatial weights to target the location of sound events, but ignores the inherent time-frequency characteristics.

To address the above issues, we propose a parallel temporal-spectral attention mechanism for CNN to learn discriminative time-frequency representations, which allows the networks to be aware of the variety of information in time frames and frequency channels. Specifically, temporal attention is employed to capture the certain frames where sound events appear, and a spectral attention method is proposed to pay a different degree of attention to various frequency bands. The idea of the spectral attention is inspired by the study of frequency-selective attentional filter in human primary auditory cortex [18], which shows that the human brain facilitates selective listening to a frequency of interest in a scene by reinforcing the fine-grained activity pattern throughout the entire superior temporal cortex that would be evoked. In addition, the parallel structure is applied in our method, which mitigates interference between the temporal and spectral features by paying attention with two different branches, and also promotes the robustness when a single branch is disturbed by the segments without the presence of sound events. The proposed method is evaluated on three benchmark datasets (ESC-10 [19], ESC-50 [19] and UrbanSound8k [20]), and achieves the state-of-the-art classification accuracy of 95.8%, 88.6% and 88.5%, respectively. Furthermore, our method is applied to another audio classification task, *i.e.* acoustic scene classification (ASC), and also improves the performance.

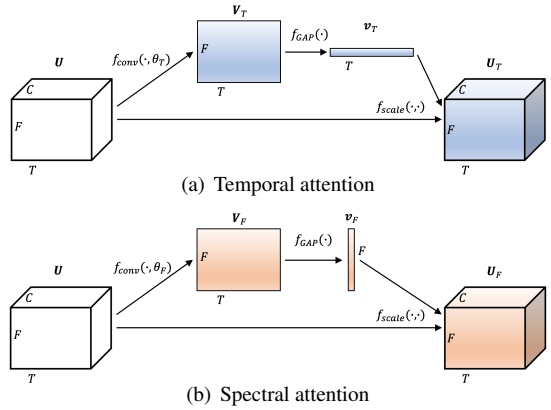
2. Proposed Method

In this section, the temporal attention and spectral attention methods are introduced, which enhance the features from relevant time frames and frequency bands. In order to obtain the temporal and spectral features simultaneously, a parallel temporal-spectral attention mechanism is then introduced.

This paper was partially supported by Shenzhen Science & Technology Fundamental Research Programs (No: JCYJ20170817160058246 & JCYJ20180507182908274) and the collaboration research project funded by PKU-HKUST ShenZhen-HongKong Institution.

*Corresponding author

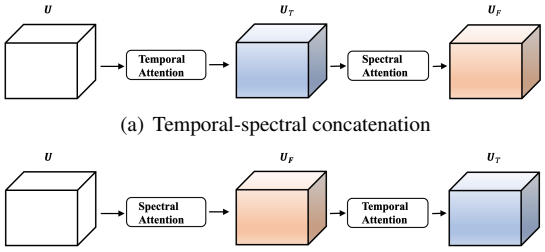
http://dx.doi.org/10.21437/Interspeech.2020-1219



(a) Temporal attention

(b) Spectral attention

Figure 1: *The illustration of temporal attention and spectral attention mechanisms.*



(a) Temporal-spectral concatenation

(b) Spectral-temporal concatenation

Figure 2: *The illustration of the concatenation of the temporal attention and the spectral attention.*

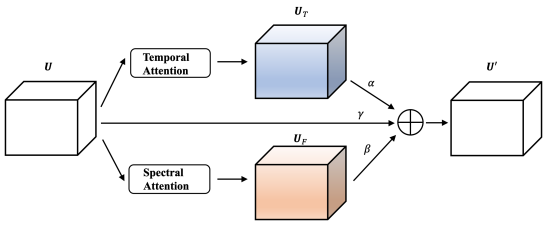


Figure 3: *The illustration of the parallel temporal-spectral attention mechanism.*

2.1. Temporal Attention and Spectral Attention

CNN have been widely used in the audio classification tasks, which show powerful ability to extract high-level features from low-level features, such as log mel spectrogram. To start with an input spectrogram of size $T \times F \times 1$, the convolutional layer consisting of C-channel filters outputs a $T' \times F' \times C$ feature map, which is then fed to the next convolutional layer to extract translation-shift invariant features. In this case, the spatial regions of the feature maps are treated equally, which may contain noise or irrelevant information for the sound events.

In order to enhance the features from relevant time frames and frequency bands, temporal attention and spectral attention are presented in our work. Different weights are applied to the time frames and frequency bands, which can guide the network to pay different attention to the temporal and spectral characteristics of environmental sounds. The structures of the two attention mechanisms are depicted in Figure 1. Specifically, for the input feature map $\boldsymbol{U} \in \mathbb{R}^{T \times F \times C}$, $1 \times 1 \times 1$ convolutional layers f_{conv} are employed to obtain the global feature maps across the channels, *i.e.* global temporal feature map $\boldsymbol{V}_T \in \mathbb{R}^{T \times F \times 1}$ and global spectral feature map $\boldsymbol{V}_F \in \mathbb{R}^{T \times F \times 1}$.

$$\boldsymbol{V}_T = f_{\text{conv}}\left(\boldsymbol{U}; \theta_{\text{T}}\right) \tag{1}$$

$$\boldsymbol{V}_F = f_{\text{conv}}\left(\boldsymbol{U}; \theta_{\text{F}}\right) \tag{2}$$

where θ_{T} and θ_{F} denote the model parameters of the convolutional layers in the temporal attention and the spectral attention, respectively. The 1×1 filters are used to squeeze the number of channels to 1, which can learn channel-wise global information from the local feature map $\boldsymbol{U}$. The extracted global temporal and spectral feature maps are then squeezed by the global average pooling f_{GAP} to obtain the time-wise activations $\boldsymbol{v}_T \in \mathbb{R}^{T \times 1 \times 1}$ and the frequency-wise activations $\boldsymbol{v}_F \in \mathbb{R}^{1 \times F \times 1}$.

$$\boldsymbol{v}_T = \sigma\left(f_{\text{GAP}}\left(\boldsymbol{V}_T\right)\right) \tag{3}$$

$$\boldsymbol{v}_F = \sigma\left(f_{\text{GAP}}\left(\boldsymbol{V}_F\right)\right) \tag{4}$$

where $\sigma(\cdot)$ denotes the sigmoid function to limit the values in the range of $(0, 1)$ and global average pooling is applied along the time axis and the frequency axis, respectively. Thus, the time-wise feature map $\boldsymbol{U}_T \in \mathbb{R}^{T \times F \times C}$ and the frequency-wise feature map $\boldsymbol{U}_F \in \mathbb{R}^{T \times F \times C}$ can be obtained by rescaling $\boldsymbol{U}$ with the time-wise activations $\boldsymbol{v}_T$ and the frequency-wise activations $\boldsymbol{v}_F$.

$$\boldsymbol{U}_T = f_{\text{scale}}\left(\boldsymbol{U}, \boldsymbol{v}_T\right) \tag{5}$$

$$\boldsymbol{U}_F = f_{\text{scale}}\left(\boldsymbol{U}, \boldsymbol{v}_F\right) \tag{6}$$

where $f_{\text{scale}}\left(\cdot, \cdot\right)$ refers to the multiplication between the feature map and the activations (*i.e.* the time-wise activations $\boldsymbol{v}_T$ and the frequency-wise activations $\boldsymbol{v}_F$).

2.2. Parallel Temporal-spectral Attention

One intuitive approach to obtain the temporal and spectral features simultaneously is the concatenation of the temporal attention and spectral attention, which is shown in Figure 2. However, there is a drawback in the approach of the concatenation, *i.e.* the temporal attention and spectral attention may interfere with each other. For example, when the temporal-spectral concatenation in Figure 2(a) is applied, the activations of some noisy frames will be inhibited by the temporal attention, while on the other hand, some slices of the noisy frames may be enhanced by the spectral attention. Thus, the effect of the temporal attention is interfered by the spectral attention.

In order to alleviate the interference of the temporal attention and the spectral attention, a parallel temporal-spectral attention mechanism is proposed in our work. As shown in Figure 3, the temporal and spectral features are paid attention with two different branches without the propagation of information mutually. In this case, each representation learning is focused on specific discriminative local regions rather than being spread evenly over the whole feature map, which leads to a better robustness when a single branch is disturbed by the sections where no sound events appear in the acoustic environments.

To be specific, a summation of the three branches (*i.e.* temporal attention, spectral attention and shortcut) are applied to obtain the final time-frequency features. The summation is not with the same weights for the reason that different attention should attend the temporal and spectral characteristics.

Given coefficient as α, β, γ, the time-frequency feature map $\boldsymbol{U}' \in \mathbb{R}^{T \times F \times C}$ is calculated by the following formulas:

$$\boldsymbol{U}' = \alpha \boldsymbol{U}_T + \beta \boldsymbol{U}_F + \gamma \boldsymbol{U} \tag{7}$$

$$\alpha + \beta + \gamma = 1 \tag{8}$$

where α, β, γ are the learnable parameters with the same initial value, and the softmax function is applied to normalize them. In this case, the network can adaptively pay different attention to the temporal characteristics and the spectral characteristics.

3. Experiments

Our method is evaluated on three ESC datasets (ESC-10 [19], ESC-50 [19] and UrbanSound8k [20]) and two ASC datasets (the DCASE 2018 task1A dataset [21] and the DCASE 2019 task1A dataset [21]), which are the commonly used datasets for ESC and ASC. Log mel spectrograms are extracted from the audio signals as the input of the networks. The experimental setups and results are detailed as follows.

3.1. Datasets and Metrics

Datasets The ESC-50 dataset [19] comprises 50 equally balanced classes of $2,000$ samples, and each sample is a monaural 5s sound recorded with a sampling rate of 44.1 kHz. The ESC-10 dataset [19] is a selection of 10 classes from the ESC-50 dataset. The UrbanSound8k dataset [20] consists of $8,732$ audio clips summing up to 7.3 hours of audio recordings. The original audio clips are recorded at different sample rates with the maximum duration of 4s. The DCASE 2018 task1A dataset [21] and the DCASE 2019 task1A dataset [21] contain 10s segments, recorded at 48kHz and spanning 10 classes.

Metrics For all the used datasets, we use the accuracy of classification as the evaluation metric, which is one of the most commonly used metrics for audio classification [22].

3.2. Network Structures

We set the experiments including a baseline model (CNN10 [23]) and ten comparison models under the same experimental setups to evaluate the proposed attention mechanism.

CNN10 CNN10 [23] consists of 4 convolutional blocks with $64, 128, 256$ and 512 output channels, respectively. Each convolutional block contains 2 convolutional layers with kernel size of 3×3, followed by downsampling with average pooling size of 2×2. Batch normalization [24] and ReLU [25] function are applied to all the convolutional layers. Global pooling layer and two fully-connected layers are then applied, followed by a softmax nonlinearity for classification. See [23] for more details about CNN10.

TS-CNN10 TS-CNN10 is our proposed model based on CNN10, where the parallel temporal-spectral attention mechanism is employed to each convolutional block. All the other setups are the same as CNN10. TS-CNN10-1, TS-CNN10-2, TS-CNN10-3 and TS-CNN10-4 are the variants of TS-CNN10, which only apply the parallel temporal-spectral attention mechanism to the 1st, 2nd, 3rd and 4th convolutional block, respectively. TS-CNN10-fixed is another variant of TS-CNN10, where the parameters α, β and γ in (7) are fixed to the same value ($\alpha = \beta = \gamma = 0.33$). T-CNN10 and S-CNN10 apply the temporal attention in Figure 1(a) and the spectral attention in Figure 1(b), respectively. In addition, TS-CNN10-concat and ST-CNN10-concat apply the attention mechanisms of temporal-spectral concatenation in Figure 2(a) and spectral-temporal concatenation in Figure 2(b), respectively.

Table 1: *Comparison of accuracy on the ESC-10, ESC-50 and UrbanSound8k (US8k) datasets*

Model	ESC-10	ESC-50	US8k
PiczakCNN [19]	80.5%	64.9%	73.0%
EnvNet-v2 [29]	91.3%	84.9%	78.3%
SB-CNN [30]	91.7%	83.9%	83.7%
GTSC+TEO-GTSC [31]	-	81.9%	88.0%
ConvRBM+FBEs [32]	-	86.5%	-
ACRNN [12]	93.7%	86.1%	-
Multi-Stream CNN [33]	94.2%	84.0%	-
MelFB+LGTFB-EN-CNN [34]	93.7%	88.1%	85.8%
Human [3]	95.7%	81.3%	-
CNN10 [23]	92.0%	85.2%	84.9%
T-CNN10 (ours)	94.5%	87.8%	87.2%
S-CNN10 (ours)	94.0%	87.5%	87.8%
TS-CNN10-1 (ours)	93.6%	86.9%	86.3%
TS-CNN10-2 (ours)	93.8%	87.0%	87.0%
TS-CNN10-3 (ours)	93.9%	86.9%	87.3%
TS-CNN10-4 (ours)	94.9%	87.8%	87.5%
TS-CNN10-concat (ours)	92.5%	85.6%	85.8%
ST-CNN10-concat (ours)	93.0%	85.5%	86.1%
TS-CNN10-fixed (ours)	95.0%	88.1%	88.3%
TS-CNN10 (ours)	**95.8%**	**88.6%**	**88.5%**

Table 2: *Comparison of accuracy on the DCASE 2018 task1A dataset (DCASE2018 1A) and the DCASE 2019 task1A dataset (DCASE2019 1A)*

Model	DCASE2018 1A	DCASE2019 1A
Official baseline [21]	59.7%	62.5%
CNN10 [23]	68.1%	69.6%
TS-CNN10 (ours)	**68.7%**	**70.6%**

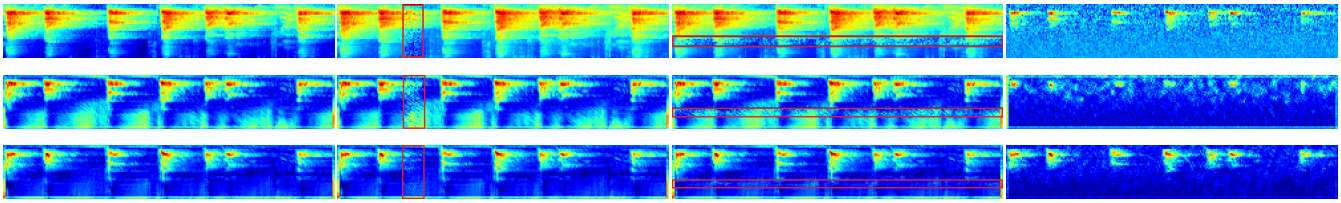
3.3. Experimental Setups

Preprocessing All the raw audios are resampled to 44.1kHz and then fixed to the certain length by zero-padding or truncating (*i.e.* 5s for the ESC-10 and ESC-50, 4s for the UrbanSound8k and 10s for the DCASE 2018 task1A dataset and DCASE 2019 task1A dataset). The short time Fourier transform (STFT) is then applied on the audio signals to calculate spectrograms, with a window size of 40ms and a hop size of 20ms. 40 mel filter banks are applied on the spectrograms followed by a logarithmic operation to extract the log mel spectrograms.

Training details In the training phase, the Adam algorithm [26] is employed as the optimizer with the default parameters. The model is trained end-to-end with the initial learning rate of 0.01 and the exponential decay rate of 0.98 for each 5 iterations. Parameters of the networks are learned using the categorical cross entropy loss. Batch size is set to 64 and training is terminated after 2000 iterations. Data augmentation methods mixup [27] and Specaugment [28] are applied in our experiments to prevent the system from over-fitting and improve the performance.

3.4. Experimental Results and Analysis

Table 1 demonstrates the performance of our proposed TS-CNN10 and other state-of-the-art methods on the ESC datasets (ESC-10 [19], ESC-50 [19] and UrbanSound8k [20]). Temporal attention was applied in [12, 33] to focus on the semantically relevant frames, which achieved higher accuracy than CNN models [19, 29, 30]. However, the spectral characteristics of the environmental sounds were not considered in these methods. Other methods [31, 32, 34] designed filter-bank learn-



(a) original audio (b) noisy audio I (c) noisy audio II (d) noisy audio III

Figure 4: *Visualization of four input log mel spectrograms (the 1st row) and the average feature maps of the first convolutional block in CNN10 (the 2nd row) and TS-CNN10 (the 3rd row). (a) The original audio from the ESC-50 dataset. (b) Gaussian random noise is added into the* $50th{\sim}64th$ *time frames (the red box). (c) Gaussian random noise is added into the* $25th{\sim}30th$ *frequency bands (the red box). (d) Gaussian random noise is added into the whole audio with the SNR of* $10dB$*.*

ing approaches instead of feeding the log mel spectrograms to the networks, however, the deep time-frequency characteristics were hardly studied. The results indicate that our proposed TS-CNN10 outperforms all the compared methods, which confirms the effectiveness of the proposed method in enhancing the features from relevant frames and frequency bands to obtain the discriminative sound representations. It is worth mentioning that TS-CNN10 surpasses the performance of human being on both the ESC-10 dataset and the ESC-50 dataset.

We observe that both temporal attention (T-CNN10) and spectral attention (S-CNN10) improve the performance of ESC, and the combination of them (TS-CNN10) brings about more improvement. In addition, TS-CNN10-1, TS-CNN10-2, TS-CNN10-3, TS-CNN10-4 all perform better than CNN10, which shows that our proposed parallel temporal-spectral attention mechanism can be applied to any convolutional layer to enhance the features. The parallel temporal-spectral attention used in the deeper layers shows more performance gain, and a higher performance can be achieved when more layers apply the parallel temporal-spectral attention mechanism (TS-CNN10-fixed and TS-CNN10). TS-CNN10 achieves higher accuracy than TS-CNN10-fixed for the reason that learnable parameters in (7) are set to adaptively pay different attention to the temporal and spectral characteristics. Besides, TS-CNN10 outperforms TS-CNN10-concat and ST-CNN10-concat, which validates the advantage of the parallel structure to alleviate the interference of the temporal features and the spectral features.

Our method is also applied to another audio classification task (*i.e.* ASC), and evaluated on the DCASE 2018 task1A dataset [21] and the DCASE 2019 task1A dataset [21]. As shown in Table 2, our attention mechanism can improve the performance of ASC as well.

To further test the robustness of TS-CNN10, three different types of noises (*i.e.* Gaussian random, bus and tram) with different SNR (*i.e.* 20dB, 10dB and 0dB) are applied to the original audio in the ESC-50 dataset. CNN10 and TS-CNN10 are trained with the original data and then tested on the noisy data, with the results shown in Table 3. TS-CNN10 shows better robustness to all the three types of noises, which is more obvious with the decrease of the SNR.

In addition, visualization analyses are employed to show how our method enhances the time-frequency representations in a complex and dynamic acoustic environment. As shown in Figure 4, the input log mel spectrograms and the feature maps of CNN10 and TS-CNN10 are visualized. Figure 4(a) is the original audio from the ESC-50 dataset. It is clear that TS-CNN10 focuses more on the relevant time frames and frequency bands, and attenuates the less relevant information as compared

Table 3: *Comparison of accuracy on the ESC-50 dataset with different noise types and different SNR*

Noise Type	SNR (dB)	CNN10	TS-CNN10	Gain†
No noise	-	85.2%	88.6%	3.4%
Gaussian random	20.0	83.5%	87.1%	3.6%
	10.0	80.5%	85.9%	5.4%
	0.0	71.2%	79.9%	**8.7%**
Bus*	20.0	82.3%	86.0%	3.7%
	10.0	71.9%	76.9%	5.0%
	0.0	58.3%	65.6%	**7.3%**
Tram*	20.0	78.5%	80.3%	1.8%
	10.0	62.7%	66.9%	4.2%
	0.0	49.5%	53.9%	**4.4%**

† The accuracy gain of TS-CNN10 compared with CNN10.
* The bus and tram noises are the clips randomly generated from the DCASE 2019 task1A dataset [21].

with CNN10. Figure 4(b) and Figure 4(c) show the noisy audio, which Gaussian random noises are added into several time frames and frequency bands. CNN10 cannot deal with the noisy audio well and the feature maps are activated in the noisy sections. While for our TS-CNN10, the feature maps in the noisy sections are much less activated. It is because the temporal-spectral attention is introduced in two different branches so that computation for representation learning is focused on specific discriminative local regions rather than being spread across the whole feature map, which leads to a better robustness when a single branch is disturbed by the noisy sections. To be specific, when random noise in several time frames is added, the spectral attention can mitigate the impact of the noisy sections by applying the global average pooling along the time axis. Similarly, we can explain the performance of TS-CNN10 in noisy frequency bands. Moreover, Gaussian random noise is added to the whole audio with the SNR of 10dB, which is shown in Figure 4(d). The result shows that the proposed TS-CNN10 can still obtain the time-frequency information of the sections where sound events appear. However, CNN10 can hardly capture the effective information and the feature map looks noisy.

4. Conclusions

A novel parallel temporal-spectral attention mechanism has been proposed to obtain the discriminative sound representations of environmental sounds. Experimental results on the ESC and ASC tasks and visualization analyses validate the advantages of our method. In future work, we would like to apply it to other audio classification tasks.

5. References

[1] R. Radhakrishnan, A. Divakaran, and A. Smaragdis, "Audio analysis for surveillance applications," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2005.* IEEE, 2005, pp. 158–161.

[2] M. Vacher, J.-F. Serignat, and S. Chaillol, "Sound classification in a smart room environment: an approach using GMM and HMM methods," 2007.

[3] K. J. Piczak, "Environmental sound classification with convolutional neural networks," in *2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 2015, pp. 1–6.

[4] J. Lee, T. Kim, J. Park, and J. Nam, "Raw waveform-based audio classification using sample-level CNN architectures," *arXiv preprint arXiv:1712.00866*, 2017.

[5] M. Huzaifah, "Comparison of time-frequency representations for environmental sound classification using convolutional neural networks," *arXiv preprint arXiv:1706.07156*, 2017.

[6] X. Zhang, Y. Zou, and W. Shi, "Dilated convolution neural network with LeakyReLU for environmental sound classification," in *2017 22nd International Conference on Digital Signal Processing (DSP)*. IEEE, 2017, pp. 1–5.

[7] Z. Zhang, S. Xu, S. Cao, and S. Zhang, "Deep convolutional neural network with mixup for environmental sound classification," in *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*. Springer, 2018, pp. 356–367.

[8] Z. Weiping, Y. Jiantao, X. Xiaotao, L. Xiangtao, and P. Shaohu, "Acoustic scene classification using deep convolutional neural network and multiple spectrograms fusion," *Detection and Classification of Acoustic Scenes and Events (DCASE)*, 2017.

[9] K. Koutini, H. Eghbal-zadeh, and G. Widmer, "Receptive-field-regularized cnn variants for acoustic scene classification," *arXiv preprint arXiv:1909.02859*, 2019.

[10] Y. Xu, Q. Kong, Q. Huang, W. Wang, and M. D. Plumbley, "Receptive-field-regularized CNN variants for acoustic scene classification," *arXiv preprint arXiv:1703.06052*, 2017.

[11] Q. Kong, C. Yu, Y. Xu, T. Iqbal, W. Wang, and M. D. Plumbley, "Weakly labelled audioset tagging with attention neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 11, pp. 1791–1802, 2019.

[12] Z. Zhang, S. Xu, T. Qiao, S. Zhang, and S. Cao, "Attention based convolutional recurrent neural network for environmental sound classification," in *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*. Springer, 2019, pp. 261–271.

[13] Z. Ren, Q. Kong, K. Qian, M. D. Plumbley, B. Schuller *et al.*, "Attention-based convolutional neural networks for acoustic scene classification," in *DCASE 2018 Workshop Proceedings*, 2018.

[14] Q. Kong, Y. Xu, W. Wang, and M. D. Plumbley, "Audio set classification with attention model: A probabilistic perspective," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 316–320.

[15] W. Wang, W. Wang, M. Sun, and C. Wang, "Acoustic scene analysis with multi-head attention networks," *arXiv preprint arXiv:1909.08961*, 2019.

[16] H. Phan, O. Y. Chén, L. Pham, P. Koch, M. De Vos, I. McLoughlin, and A. Mertins, "Spatio-temporal attention pooling for audio scene classification," *arXiv preprint arXiv:1904.03543*, 2019.

[17] S. Hong, Y. Zou, W. Wang, and M. Cao, "Weakly labelled audio tagging via convolutional networks with spatial and channel-wise attention," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2020)*, 2020.

[18] S. Da Costa, W. van der Zwaag, L. M. Miller, S. Clarke, and M. Saenz, "Tuning in to sound: frequency-selective attentional filter in human primary auditory cortex," *Journal of Neuroscience*, vol. 33, no. 5, pp. 1858–1863, 2013.

[19] K. J. Piczak, "ESC: Dataset for environmental sound classification," in *Proceedings of the 23rd ACM International Conference on Multimedia*. ACM, 2015, pp. 1015–1018.

[20] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in *Proceedings of the 22nd ACM International Conference on Multimedia*. ACM, 2014, pp. 1041–1044.

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

[22] T. Virtanen, M. D. Plumbley, and D. Ellis, *Computational Analysis of Sound Scenes and Events*. Springer, 2018.

[23] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "PANNs: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition," *arXiv preprint arXiv:1912.10211*, 2019.

[24] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[25] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, 2010, pp. 807–814.

[26] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[27] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2017.

[28] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.

[29] Y. Tokozume, Y. Ushiku, and T. Harada, "Learning from between-class examples for deep sound recognition," *arXiv preprint arXiv:1711.10282*, 2017.

[30] J. Salamon and J. P. Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," *IEEE Signal Processing Letters*, vol. 24, no. 3, pp. 279–283, 2017.

[31] D. M. Agrawal, H. B. Sailor, M. H. Soni, and H. A. Patil, "Novel TEO-based Gammatone features for environmental sound classification," in *2017 25th European Signal Processing Conference (EUSIPCO)*. IEEE, 2017, pp. 1809–1813.

[32] H. B. Sailor, D. M. Agrawal, and H. A. Patil, "Unsupervised Filterbank Learning Using Convolutional Restricted Boltzmann Machine for Environmental Sound Classification.." in *INTERSPEECH*, 2017, pp. 3107–3111.

[33] X. Li, V. Chebiyyam, and K. Kirchhoff, "Multi-stream network with temporal attention for environmental sound classification," in *Proc. Interspeech 2019*, 2019, pp. 3604–3608. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2019-3019

[34] H. Park and C. D. Yoo, "CNN-Based Learnable Gammatone Filterbank and Equal-Loudness Normalization for Environmental Sound Classification," *IEEE Signal Processing Letters*, 2020.