INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Multi-stream Network With Temporal Attention For Environmental Sound Classification

Xinyu Li[*]*, Venkata Chebiyyam*[*]*, Katrin Kirchhoff*

Amazon AI

{xxnl,chebiyya,katrinki}@amazon.com

Abstract

Environmental sound classification systems often do not perform robustly across different sound classification tasks and audio signals of varying temporal structures. We introduce a multi-stream convolutional neural network with temporal attention that addresses these problems. The network relies on three input streams consisting of raw audio and spectral features and utilizes a temporal attention function computed from energy changes over time. Training and classification utilizes decision fusion and data augmentation techniques that incorporate uncertainty. We evaluate this network on three commonly used datasets for environmental sound and audio scene classification and achieve new state-of-the-art performance without any changes in network architecture or front-end preprocessing, thus demonstrating better generalizability.

Index Terms: environmental sound classification, audio scene classification, convolutional neural networks

1. Introduction

Environmental sound classification (ESC) has become a topic of great interest in signal processing due to its wide range of applications (surveillance, activity recognition, captioning, etc.). Although many previous studies have shown promising results on ESC [1, 2, 3], largely through the introduction of deep learning methods, ESC still faces several challenges. First, different studies have identified combinations of feature extraction methods and neural network designs that work best for individual datasets [2, 4], but that have failed to generalize well across different ESC tasks. Another problem is that environmental sounds often have highly-variable temporal characteristics (e.g., short duration for water drops but longer duration for sea waves). An ESC model needs to be able to isolate the meaningful features for classification within the acoustic signal instead of overfitting to the background sound. To address these problems we propose a multi-stream neural network that uses only the most fundamental audio representations (waveform, short-term Fourier transform (STFT), or spectral features) as inputs while relying on convolutional neural networks (CNNs) for feature learning.

In order to localize class-differentiating features in highly-variable sound signals we propose a temporal attention mechanism for CNNs that applies to all input streams. Compared with attention used for tasks such as neural machine translation [5] or feature level attention [6] our proposed attention mechanism works synchronously with CNN layers for feature learning. To handle signals of variable lengths with our fixed-dimensional CNN architecture we propose a decision fusion strategy with uncertainty. We tested our model on three published dataset that vary in the number of classes (10 to 50) and audio signal length (10 seconds to 30 seconds). Our system meets or surpasses the state-of-the-art on all datasets without any changes to the model architecture or feature extraction method. We include an ablation study that highlights the relative importance of each system component. The rest of this paper is structured as follows: In Section 2 we introduce previous related work. We describe our system in Section 3 and experimental results in Section 4. Section 5 provides an analysis of the attention function. Section 6 concludes.

[*] equally contributed

2. Related Work

Initial studies of ESC heavily relied on manually designed features [7, 8] and traditional classification methods such as support vector machines (SVMs) and k-nearest neighbor (KNN) classifiers. Subsequent work introduced deep learning to the field; in [9] DNNs were used to replace traditional classification methods. Inspired by work on image classification [10], CNNs were used in combination with time-frequency representations for ESC in [11, 12, 13]. An end-to-end system to directly learn log-mel features from raw audio input was proposed in [3]. More recent research has experimented with features at different temporal scales by merging the RNN outputs over time in a stacked RNN architecture [14], and with modifying spatial resolution by applying different filters to the input [15]. Other approaches towards improving ESC performance include higher-level input features such as MFCCs, gammatone features, or other specialized features [4, 15] and training with data augmentation [3, 16]. Multiple loss functions were used for the detection of rare sound events in [14]. An ensemble network based on two input streams was proposed very recently [17]. We used the multi-stream input as well, but in contrast to [17], we applied the early feature-level fusion instead of result ensemble. The sigmoid-layer based attention was previously proposed to work with pre-extracted feature vectors for ESC [6]. As opposed to [6] we proposed the CNN based temporal-attention that works with end-to-end model and select the important features overtime.

3. Method

3.1. Multi-stream network

3.1.1. Preprocessing

We introduce a three-stream network that takes the raw audio waveform, short-term Fourier transform (STFT) coefficients, and delta spectrogram as inputs (Figure 1). The waveform carries both magnitude and phase information represented in the time domain. We first chunk the audio waveform into non-overlapping segments of 3.84s, resulting in $44.1k \times 3.84$ samples, and then calculate the STFT spectrogram. Different resolutions of STFT highlight different details in the spectrogram

http://dx.doi.org/10.21437/Interspeech.2019-3019

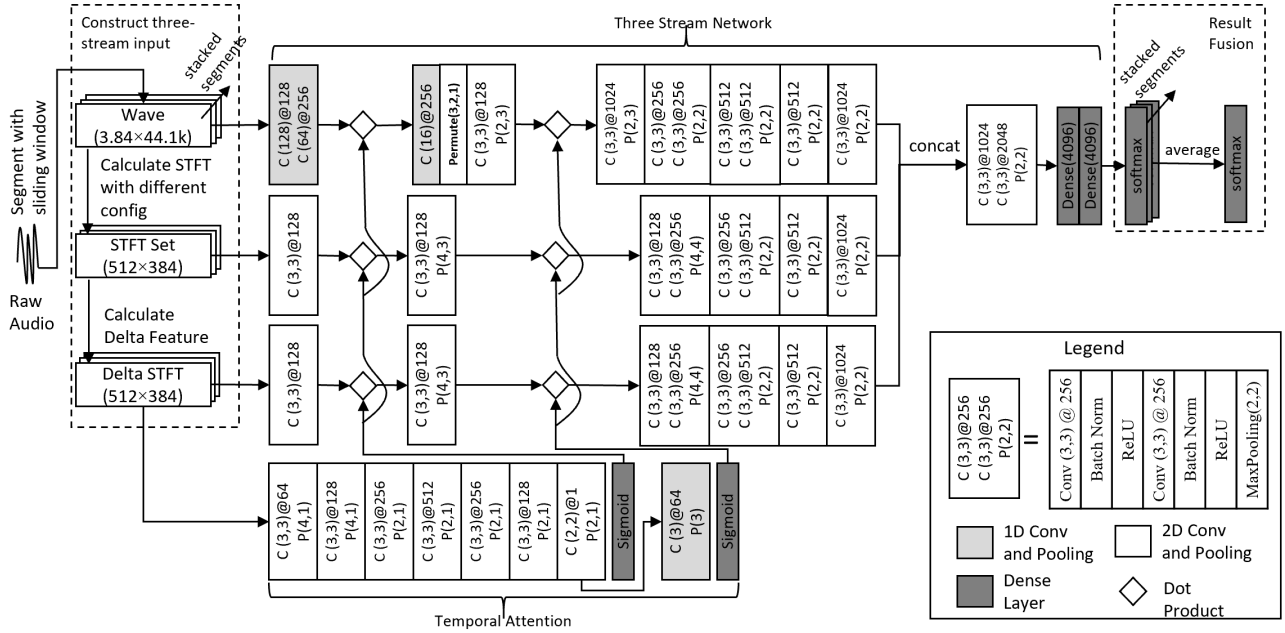

Figure 1: *System Architecture.*

and emphasize either frequency details or temporal details. We choose a set of three resolutions corresponding to 32, 128 and 1024 FFT points with a hop length of 10ms. The generated STFTs are scaled into same dimension (512×384) and stacked together as STFT set features. We further calculate delta features with a window size of 5 from each STFT layer and stack them to form a feature vector of the same dimensionality as the STFT set (512×384). We restrict ourselves to these basic audio representations instead of higher-level features to encourage network learning useful features which makes our system robust. Note that the log-mel features were previously used for ESC with good performance [18], we used STFT features instead of log-mel because previous research demonstrated that the log-mel features can be learnt directly from raw audio input [3].

3.1.2. Network structure

We adopted and modified the EnvNet [3] architecture, which extracts log-mel features with both 1D and 2D convolutions from the waveform (Figure 1):

$$F_{waveform} = \text{Conv2D}(\text{Conv1D}(x_{raw})) \quad (1)$$

where x_{raw} is the 1D waveform. Conv2D and Conv1D denote the corresponding convolutional operations. The output of 1D convolution has two dimensions (feature, channel). We can treat the 1D conv output as a 2D feature map with channel dimension as 1 and apply the 2D convolution operation. The $F_{waveform}$ is the representation learned from x_{raw}, which has three-dimensions (feature, temporal, channel). A 2D CNN was used to learn features F_{STFT} and F_{Delta} from the STFT set and the delta spectrogram as:

$$F_{STFT/Delta} = \text{Conv2D}(x_{STFT/Delta}) \quad (2)$$

where $x_{STFT/Delta}$ are the 3D stacked STFT or delta spectrogram features.

We used 3×3 filters for the 2D convolution with batch normalization and ReLU activation. Larger filters (128, 64 and 16) were used for 1D convolution with large strides for fast feature dimension reduction. Since the convolution and pooling operations do not compromise spatial association we applied the same number of pooling operations to all three streams over time. This results in feature representations learned from different streams that are synchronized in the temporal dimension, and that can be merged by concatenating along time (Figure 1).

3.2. Temporal attention

Different environmental sounds can have a very different temporal structure, e.g., bell rings are different from water drops or sea waves. Most previous studies have ignored fine-grained temporal structure and have extracted features at the global signal level [9, 12, 3]. In other domains temporal structure is typically addressed by sequence models such as long short-term memory (LSTM) networks with temporal attention [5]. For ESC, recent studies have proposed temporal modeling by subsampling and averaging outputs from RNN layers over time [14]; however, this is not equivalent to weighting different parts of the signal differently. A CNN-BLSTM model with temporal attention was proposed in [19]; however, attention was calculated within the BLSTM based on features extracted from the CNN; it thus did not influence feature extraction itself.

We integrate an attention function into our multi-stream CNN (Figure 1) that is calculated from the delta spectrogram features and directly affects the CNN layers themselves. This representation provides information about dynamic changes in energy, which we assume is beneficial for extracting temporal structure. Our initial experiments also showed that attention calculated from delta features has better performance relative to attention calculated from all three inputs. Temporal attention weights are calculated in two steps: (1) Repeat convolution and 1D pooling along the feature dimension ($pooling\_kernel = (N, 1)$), until the feature dimension equals

one (Figure 1, temporal attention block). Because the convolution and pooling operations do not compromise the temporal association of the data, the generated attention vector is temporally aligned with the inputs from all three streams. (2) Pooling along time ($pooling\_kernel = (1, N)$), which aligns our temporal attention with the features learned by the CNN after each pooling operation. The same attention vector is shared by all three input streams since all three branches are synchronized in time. The attention is applied to the learned features via dot-product operations along the time dimension (Figure 1):

$$F_{fc} = \mathbf{C}_{fc} \cdot \mathbf{A}, \quad 1 \leq f \leq F, 1 \leq c \leq C \tag{3}$$

where $\mathbf{C}$ is the output from the convolutional layer with shape (F, T, C) and $\mathbf{A}$ is the attention vector with shape $(1, T)$. The attention is applied by multiplying the attention vector A with each of the feature vectors in $\mathbf{C}$ along feature dimension and channel dimension as $\mathbf{C}_{fc} \cdot \mathbf{A}$. The F is the same feature after applying attention that has the same shape as C.

3.3. Decision fusion with uncertainty

In order to handle audio signals of different lengths with a network structure that requires fixed-length inputs we propose a late fusion strategy that computes classification outputs for each input window and then fuses the softmax layer outputs for each window by averaging. Instead of simply averaging the softmax probabilities over time [3], we further augment the training data with white-noise segments and use a uniform probability distribution over all classes as the target distributions for these segments. Enriching the training data with these maximally uncertain segments biases the system to predict high-entropy softmax outputs when the input does not contain useful information. This is critical in order to prevent the final decision from being overly influenced by noise or silence segments.

3.4. Data augmentation

To avoid possible overfitting caused by limited training data, we adopt the between-class training approach to data augmentation [3, 18] and modify it as follows. We create mixed training samples

$$mix(x_{a,i}, x_{b,j}, r) = rx_{a,i} + (1 - r)x_{b,j} \tag{4}$$

where a and b are two randomly selected clips from the training data and i and j are two randomly selected starting points in time. Fixed-length audio segments are selected from each clip based on the start times. The r parameter is a random mixture ratio between 0 and 1 used for mixing the two segments. x denotes the combined three input vectors (waveform, STFT, delta spectrogram). The class labels used for the mixed samples are chosen with the same proportion. We use this procedure instead of the gain-based mixture (calculating the mixture ratios based on the signal amplitude) suggested in [3] for two reasons: 1. The gain-based mixture is substantially ($\sim$ 20 times) slower than our approach, and 2. The gain-based mixture does not apply to 3D features. We rerun data augmentation at each epoch of neural network training.

4. Experimental results

4.1. Datasets and training procedure

We tested our system on three commonly used datasets:
ESC-10 and ESC-50 [1]: ESC-50 is a collection of 2,000 environmental sound recordings. The dataset consists of 5-second-long recordings organized into 50 semantic classes (40 examples per class). The data is split into 5 groups for training and testing. We use 5-fold cross-validation and report the average accuracy. ESC-10 is a subset of ESC-50 that contains 10 labels.
TUT Acoustic scenes 2016 dataset (DCASE): This dataset consists of recordings from various acoustic scenes, all having distinct recording locations. For each recording location, a 3 to 5 minute-long audio recording was captured. The original recordings were then split into 30-second segments. The dataset comes with an official training and testing split. We report the average accuracy score on four training and testing configurations in line with as previous research.

We implemented our model in Keras with a TensorFlow backend. Adam optimizer with an initial learning rate of 0.001 was used; the learning rate decays by a factor of 10 after every 100 epochs. We used the mean absolute error instead of categorical cross-entropy as a loss function.

4.2. Results and comparison

Table 1 compares our results against previous outcomes reported in the literature; note that we used the same feature representations and model structure for all datasets. Results show that our model achieves state-of-the-art or better performance on all three datasets (Table 1) while most previously proposed approaches show highly divergent performance on different datasets (Table 1, gray shaded rows). Also note that the SoundNet system [2] shows high performance on the DCASE dataset but has been pre-trained on video and audio data, whereas our network is trained purely on audio.

We further analyzed the contribution of each component in our system: the three input streams, the temporal attention and the decision fusion mechanism. The results (Table 2) show that: 1. The three-stream network works better than using a combination of any two of the input streams (Table 2 first three rows). 2. Temporal attention improves the performance on all

Table 1: *Experimental results and comparison. * best score from multiple runs of experiments.*

	ESC 10	DCASE	ESC 50
KNN [1]	0.667	0.831	0.322
SVM [1]	0.675	0.821	0.396
Random Forest [1]	0.727	/	0.443
AlexNet [12]	0.632	/	0.678
Google Net [12]	0.784	0.84	0.787
WaveMSNet [15]	0.937	/	0.793
SoundNet [2]	0.922	0.88	0.742
EnvNet&BC Training [3]	0.894	/	0.818 0.849*
Gammatone [4]	/	/	0.819
ProCNN [17]	0.921	/	0.828
CNN mixup [16]	0.917	/	0.839
CNN-LSTM [19]	/	0.762	/
Human [1]	0.957	/	0.813
Ours (Average) Ours (Best)	0.937 0.942*	0.875 0.882 *	0.835 0.840*

three datasets, which demonstrates the generalizability and effectiveness of our method. 3. Decision fusion leads to roughly 2.5% accuracy gain across all datasets. 4. Noise augmentation for decision fusion lead to roughly 1% performance gain on all datasets. 5. The data augmentation is necessary for all three datasets. Without augmentation, the network will quickly overfit to the relatively limited training data. Note that the improvement lead by the data augmentation is partially because the small size of the datasets. Given larger dataset, the impact of data augmentation will be less significant.

Table 2: *Contributions of different system components.*

	ESC 10	DCASE	ESC 50
Without spectrogram	0.816	0.793	0.715
Without delta spectrogram	0.821	0.825	0.697
Without raw audio	0.792	0.781	0.745
Without attention	0.917	0.853	0.823
Without decision fusion	0.915	0.857	0.812
Without uncertainty	0.930	0.865	0.825
Without data augmentation	0.815	0.770	0.712
Complete Model	0.937	0.875	0.835

5. Visualizing attention

Environmental sounds have diverse temporal structures. Sounds may be continuous (e.g. rain and sea waves), periodic (e.g., clock tics and crackling fire), or non-periodic (e.g., dog, rooster). To have a better understanding how temporal attention helps with recognizing different sounds, we visualized the attention weights generated for sounds with different temporal structures (Figure 2). From the visualization we can see that the proposed attention is able to locate important temporal events while de-weighting the background noise (Figure 2, top row). The attention curve has a periodic shape for periodic sounds (Figure 2, middle row) while being continuous for continuous sounds (Figure 2, bottom row), regardless of sound volume changes (Figure 2, sea waves).

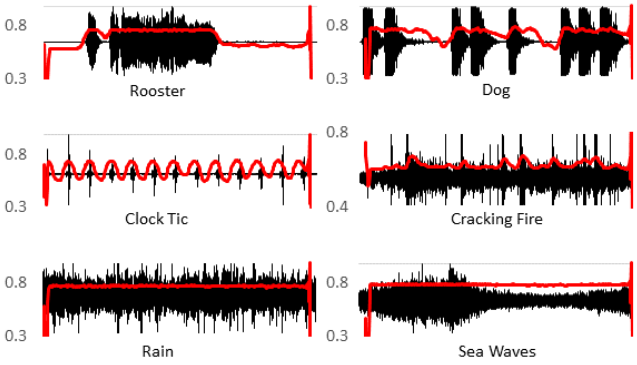

Figure 2: *Comparison of generated attention on enviornmetnal sound with different temporal structure. Black line: audio waveform, Red line: generated attention.*

6. Conclusion

We have described a multi-stream CNN with temporal attention and decision fusion for ESC. Our system was evaluated on three commonly used benchmark datasets and consistently achieved state-of-the-art or better performance with a single network architecture. In the future we will extend this work to larger datasets such as Audioset and incorporate mechanisms to handle overlapping sounds.

7. References

[1] K. J. Piczak, "ESC: dataset for environmental sound classification," in *Proceedings of the 23rd ACM international conference on Multimedia*. ACM, 2015, pp. 1015–1018.

[2] Y. Aytar, C. Vondrick, and A. Torralba, "Soundnet: Learning sound representations from unlabeled video," in *Advances in Neural Information Processing Systems*, 2016, pp. 892–900.

[3] Y. Tokozume and T. Harada, "Learning environmental sounds with end-to-end convolutional neural network," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 2721–2725.

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

[5] B. Sankaran, H. Mi, Y. Al-Onaizan, and A. Ittycheriah, "Temporal attention model for neural machine translation," *arXiv preprint arXiv:1608.02927*, 2016.

[6] C. Yu, K. S. Barsim, Q. Kong, and B. Yang, "Multi-level attention model for weakly supervised audio classification," *arXiv preprint arXiv:1803.02353*, 2018.

[7] J.-C. Wang, J.-F. Wang, K. W. He, and C.-S. Hsu, "Environmental sound classification using hybrid SVM/KNN classifier and MPEG-7 audio low-level descriptor," in *Neural Networks, 2006. IJCNN'06. International Joint Conference on*. IEEE, 2006, pp. 1731–1735.

[8] S. Chu, S. Narayanan, and C.-C. J. Kuo, "Environmental sound recognition with time–frequency audio features," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 6, pp. 1142–1158, 2009.

[9] N. D. Lane, P. Georgiev, and L. Qendro, "DeepEar: robust smartphone audio sensing in unconstrained acoustic environments using deep learning," in *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 2015, pp. 283–294.

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[11] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold *et al.*, "CNN architectures for large-scale audio classification," in *2017 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE, 2017, pp. 131–135.

[12] V. Boddapati, A. Petef, J. Rasmusson, and L. Lundberg, "Classifying environmental sounds using image recognition networks," *Procedia Computer Science*, vol. 112, pp. 2048–2056, 2017.

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

[14] W. Wang, C.-C. Kao, and C. Wang, "A simple model for detection of rare sound events," *Proc. Interspeech 2018*, pp. 1344–1348, 2018.

[15] B. Zhu, C. Wang, F. Liu, J. Lei, Z. Lu, and Y. Peng, "Learning environmental sounds with multi-scale convolutional neural network," *arXiv preprint arXiv:1803.10219*, 2018.

[16] Z. Zhang, S. Xu, S. Cao, and S. Zhang, "Deep convolutional neural network with mixup for environmental sound classification," *arXiv preprint arXiv:1808.08405*, 2018.

[17] S. Li, Y. Yao, J. Hu, G. Liu, X. Yao, and J. Hu, "An ensemble stacked convolutional neural network model for environmental event sound recognition," *Applied Sciences*, vol. 8, no. 7, p. 1152, 2018.

[18] I.-Y. Jeong and H. Lim, "Audio tagging system for dcase 2018: Focusing on label noise, data augmentation and its efficient learning," Tech. Rep., DCASE2018 Challenge, Tech. Rep., 2018.

[19] J. Guo, N. Xu, L.-J. Li, and A. Alwan, "Attention based CLDNNs for short-duration acoustic scene classification," *Proc. Interspeech 2017*, pp. 469–473, 2017.