Convolutional Recurrent Neural Networks for Urban Sound Classification using Raw Waveforms

Jonghee Sang, Soomyung Park, Junwoo Lee
ETRI, Korea
jhsang21, smpahk, leejw@etri.re.kr

***Abstract*— Recent studies have demonstrated deep learning approaches directly from raw data have been successfully used in image and text. This approach has been applied to audio signals as well but not fully explored yet. In this works, we propose a convolutional recurrent neural network that directly uses time-domain waveforms as input in the domain of urban sound classification. Convolutional recurrent neural network is combined model of convolutional neural networks for extracting sound features and recurrent neural networks for temporal aggregation of the extracted features. The method was evaluated using the UrbanSound8k dataset, the largest public dataset of urban environmental sound sources available for research. The results show how convolutional recurrent neural network with raw waveforms improve the accuracy in urban sound classification and provide effectiveness of its structure with respect to the number of parameters.**

I. INTRODUCTION

The environmental sound has been a significant role in understanding the content of multimedia. With such an importance and a growing demand, characterizing environmental sound is the efficient way to use it. Therefore, environmental sound classification (ESC) has been an increasingly popular problem in audio recognition research. The applications of ESC range from audio scenes classification [1] and audio surveillance system [2, 3, 4] to multimedia content highlight extraction [5, 6]. Importantly, it also has potential to improve the quality of life of city by reducing noise with smart multimedia sensor networks [7, 8].

In music information retrieval (MIR), researchers have traditionally converted raw waveforms of sound signal to a 2-dimensional time-frequency representation. 2D representations have been considered an effective form of audio data by decomposing the signal with kernels (e.g., STFT) and using log-scaled representations in frequency (e.g., Mel-spectrogram, CQT). In early ESC works, both the signal processing and machine learning approaches including matrix factorization [9, 10, 11, 12], dictionary learning [13, 14], wavelet filterbanks [15, 16] and the cepstral-based features, such as gammatone cepstral coefficients (GTCC) [5, 17] are typically used. This process so-called "engineered features" requires significant engineering effort and considerable prior knowledge about the problem. In addition, feature engineering is often heuristically designed and might not be optimal for the task.

Recent advances in deep neural networks have encouraged feature learning which takes raw audio signals, thereby minimizing the effort of preprocessing the input data. Feature learning with raw data was attempted to solve a music auto-tagging task by using a convolution neural network (CNN) model [18]. The CNN model with time-domain waveforms has been applied to recognize a variety of speech domains [19, 20, 21]. For the ESC task, end-to-end environmental sound classification systems with a CNN have proposed [22]. They show that an end-to-end system is capable of extracting features from the raw waveforms. Convolutional Restricted Boltzmann machine (ConvRMB) also have been used as the unsupervised filterbank learning from the raw audio signals [23].

They provide results comparable to models using conventional engineered features such as mel-frequency spectrogram. These works, however, have failed to outperform the model with conventional features since most neural networks considered are not deep enough to learn the complex structure of sound sufficiently. The issue can be solved by modeling CNNs with very deep architectures [24]. The authors of this work built very deep networks with up to 34 weight layers, however, the performance was improved with depth up to 18 layers due to overfitting.

This issue can be solved by combining CNN with Recurrent neural network (RNN) which are called Convolutional recurrent neural network (CRNN). The CRNN architectures can be described as a modified CNN by replacing the last layers with an RNN layer. In CRNN, CNNs are able to extract high level features that are invariant local variations. RNNs enables the networks to take temporal aggregation of extracted features. CRNN have shown excellent performance in [25] for document classification and extended to use bird audio detection [26] and music emotion recognition [27].

In this paper, we introduce CRNNs for environmental sound classification, directly using raw waveform data as the input. We compare our method with earlier works done by CNNs and deep convolutional neural networks

This work was supported by Institute for Information & Communication Technology Promotion(IITP) grant funded by Korea government(MSIT) (No. R0118-16-1005, Digital Content In-House R&D).

ISBN 978-90-827970-1-5 © EURASIP 2018

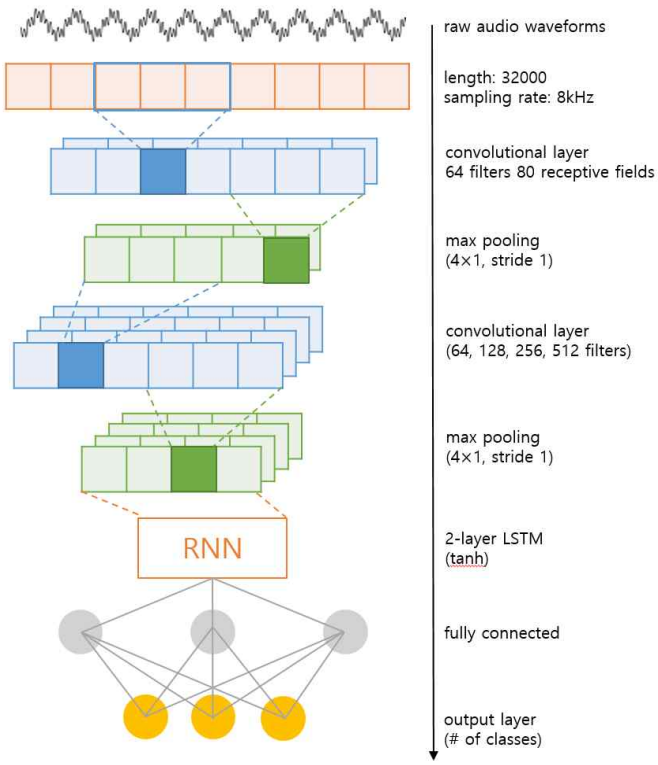

Fig. 1. Proposed method of convolutional recurrent neural network for urban environmental sound classification. The input waveforms are represented by a single channel. In convolutional layer, the different number of filters is used and we change the number of layers of each filter to make appropriate results.

(DCNNs). In comparison, all the layers are identically applied with batch normalization [28] to improve the performances. CRNN has weak dropout (0.1) layer to prevent overfitting of the RNN layers [29]. The results show that CRNNs improve the performance 7.38% in absolute accuracy and are more efficient with respect to computation cost.

II. Method

Figure 1 illustrates the architectures of our method in ESC task. In this section, we describe the key design elements in detail.

A. Raw Waveforms

In a sound classification task, the input size has a significant effect on the model performance. Our architecture takes time-series waveforms as input. Time-series waveforms have a very large number of features along a single channel. This means that raw waveforms are required to be sub-sampling for computation issue. To sub-sampling the raw waveforms, one-dimensional strided convolution and pooling layer are often used to make feature map, e.g., in music tagging [18, 30]. For the pooling layer, max-pooling is used for sub-sampling to add time-invariance and the performance is superior to the stride sub-sampling method [31]. In addition, audio sampling rate has significantly effect on the size of the input. In this model, we choose sampling rate at 8kHz and filter size of 80 to cover a 10-millisecond duration, thus the output has similar dimensions of popularly used in mel-spectrogram. In [24], the results have shown that a much smaller or larger kernel size gives poor performance.

B. Recurrent Neural Network

Since our model makes use of sequential information which the output is dependent on the previous computations, we need to capture the information about what has been calculated. RNN is suitable for deal with sequential information such as music and audio. Since adopting RNN enables the architecture to take the output data of previous layer into account, we have more flexibility to classify sound. In Figure 2, we show a basic RNN network and being unfolded into a full network with respect to the number of outputs. Since our problem is urban sound classification which yields only one output prediction, we use many-to-one models as illustrated in Figure 2 (b). Among various kinds of model of RNN, we choose long-short term memory (LSTM) model in our architectures. LSTM networks are widely used in deep learning with sequential data. The memory in LSTMs is called the hidden state that is calculated based on the previous hidden state the current input. It turns out that these types of units are very efficient at capturing long-term dependencies. In our method, we use two LSTM layers, which the last hidden state is connected to the dense layer of the network.

C. Batch Normalization

Batch Normalization (BN) is a frequently used technique for improving the performance of neural networks [28]. It reduces the problem of vanishing and exploding gradients and overfitting, a common issue in training neural networks. During training, each batch of the activations of the previous layer is normalized to the mean close to 0 and standard deviation close to 1. Using BN encourages networks to use less dropout in training, which makes networks maintain training data and higher learning rates. It is important to increase the speed at which networks train. In our model, BN is added on the output of each convolution layer before applying activation function (ReLU).

D. Output Layers

Dense layers are frequently used in the output layer. Dense layers perform classification on the features extracted by, in this case, recurrent layers. In a dense layer, the number of nodes is set to the number of classes in the problem. Every node in the layer is connected to the node that represents a probability for each sound class. To set the probability value between 0 and 1 and sum of the probability equal to 1, a softmax activation function is widely used with dense layers.

Using these elements mainly, we construct CRNN models with raw sound signal as the input. All the layers in CRNN are convolutional and fully-connected with rectified linear unit (ReLU) activation and the BN [28] for the CNN layer. In RNN layer, LSTM is utilized to learn the temporal information in the features. Lastly, dense layers with softmax activation is added to get the probability for each sound class.

ISBN 978-90-827970-1-5 © EURASIP 2018

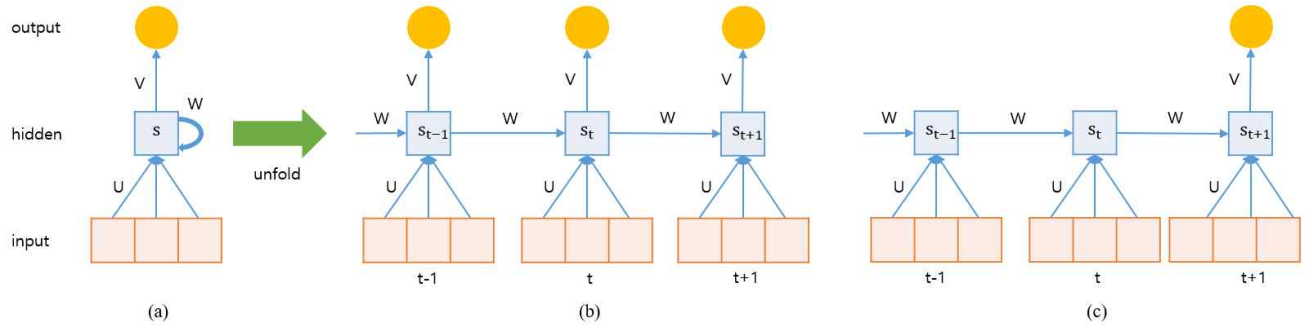

Fig. 2. Illustrations of a recurrent layer as (a) is folded and (b) is unfolded. The main feature of recurrent layer is its hidden state, which captures some information about a sequence. S is the hidden state at time step t. S at current time t is calculated based on the previous hidden state and the input at the current step. The function applied to sum of previous hidden state and the input at current step is usually a nonlinearity such as tanh or ReLU. The first hidden state is typically initialized to all zeroes. The number of outputs is varied depending on the task. In our experiment, when classifying the urban environmental sound, (c) is used to yield only one output prediction at the final time step.

III. Experiments

A. Dataset

For the evaluation of our method, we use UrbanSound8k dataset [32]. The dataset consists of 10 environmental sounds from urban area such as air conditioner, car horn, children playing, dog bark, drilling, engine idling, gun shot, jackhammer, siren and street music. It contains 8732 labeled sound samples excerpts of up to 4 seconds duration, 9.7 hours in total. Since the sample are recorded in the field, there are often other sounds contained in addition to the labeled sound. The dataset in UrbanSound8K is pre-sorted into 10 folds which ensures sound from the same recording will not be used both for training and testing. For every experiment, the official fold 10 is used for our test set, and the rest for training and validation [13, 24, 33]. The audio signals are down-sampled to 8kHz and standardized to 0 mean and variance 1 for computational speed.

B. Optimization

Table 1 outlines the 4 architectures we consider. The architectures are built with Keras [34] and Tensorflow [35]. We trim audio signal with sampling rate 8kHz using Librosa [36]. The input shape is a 32000 length vector. We train the CRNN models using Adam [37] and categorical cross-entropy as a loss function. In order to use categorical cross-entropy loss, we transform the class labels into categorical format that each class is a 10-dimensional vector that is all-zeros except for a 1 at the index corresponding to the class. In convolutional layer, we use Xavier initialization [38] to avoid vanishing and exploding gradients. For every convolutional layer, we used batch normalization [28] for the same purpose. We applied weak dropout of 0.1 to prevent overfitting the RNN and tanh is used in RNN as an activation function. The last hidden state is connected to the fully-connected layer of the network. Softmax activation function is used to get the probability for each sound class.

IV. Results

We estimated the proposed model with 4 different network depth. Our models examined from single-layered CNNs to 11 layered in order to analyze the relation between the depth of networks and the performance. The

CRNN2(0.25M)	CRNN5(1M)	CRNN8(2M)	CRNN12(3M)
Input : 32000x1 time-domain waveform			
C(64, 80/4)	C(64, 80/4)	C(64, 80/4)	C(64, 80/4)
Max-pooling : 4x1			
C(64, 3)	C(64, 3)	C(64, 3)	C(64, 3) x 2
Max-pooling : 4x1			
	C(128, 3)	C(128, 3) x 2	C(128, 3) x 3
	Max-pooling : 4x1		
	C(256, 3)	C(256, 3) x 2	C(256, 3) x 3
	Max-pooling : 4x1		
	C(512, 3)	C(512, 3) x 2	C(512, 3) x 3
	Max-pooling : 4x1		
RNN (128)			
Fully-Connected (activation : softmax)			

Table I. Architectures of proposed convolutional recurrent neural network for time-domain waveform inputs. The number next to CRNN denotes the number of convolutional layers and parameters. Capital letter C represents convolution layer. In convolutional layer, the number in parentheses denotes the number of filters, kernel size and stride. In RNN, the number denotes dimensionality of the output space. 2-layer RNN with LSTM is used in our model. Stride is omitted for stride 1. The number multiplied to parentheses is the number of stacked layers. In LSTM, we use dropout of 0.1 these architectures.

Model	Test	Number of parameters
CRNN2	67.41%	0.25M
CRNN5	73.92%	1M
CRNN8	79.06%	2M
CRNN12	68.07%	3M

Table II. Classification test accuracies for models on UrbanSound8k dataset.

ISBN 978-90-827970-1-5 © EURASIP 2018

Method	Test	Number of parameters
CRNN8 (proposed)	79.06%	2M
Wei et al. [24]	71.68%	3.7M

Table III. Comparison of classification accuracies and the number of parameters for models of UrbanSound8k dataset.

number of parameters which is closely related to the training time was also considered.

In Table 2, we showed the test accuracies for our models depending on the number of convolutional layers. The performance of CRNN2 was very poor compared with other models. It indicates that 2-layered model has limited to extract discriminative features from raw waveforms. We further investigated the performance of CRNN by constructing deeper networks. The test accuracy improved as the network depth was increased for CRNN5, CRNN8. The best result was obtained when the number of convolutional layers was 8. The performance achieved the accuracy of 79.06% in CRNN8 that was competitive with the previous state-of-the-art result on UrbanSound8k. Fig. 3 showed the confusion matrix across the different classes using CRNN8. The highest confusion occurred across siren, street music, and children playing since the high similarity of tonal components. Interestingly, the performance improved up to model CRNN8, at 79.06% accuracy, whereas CRNN12 which the number of convolutional layers was 12 only achieves 68.07%. This was due to overfitting caused by deeper networks. We considered that dataset was not sufficient to train deeper network without additional regularization.

In Table 3, we compared the performance of the proposed method to previous state-of-the-art on UrbanSound8k. It represented that our proposed CRNN performed significantly better than deep CNN of Wei et al. [24] with an absolute improvement of 7.38% in urban environmental sound classification accuracy. The performance was achieved with even the smaller number of parameters, which indicated that CRNN was significant advantages in computation than deep CNN structures.

V. Conclusions

In this paper, we propose CRNN model that takes raw waveforms as input in environmental sound classification (ESC). Through our experiments, we show that raw waveform inputs with CRNN are competitive in ESC. Adopting RNN which takes global structure into account makes more flexible to select the characteristic of sound. Our architecture outperforms the deep networks with 12 weighted layer in [24] which show the state-of-the-art performance in urban environmental sound classification by 7.38% absolute accuracy and uses significantly less amount of parameters. In addition, we can see that raw waveform model is reasonable in ESC. Our proposed architecture contributes to improving neural networks model which take time-series waveforms as input for audio classification.

Acknowledgment

This work was supported by Institute for Information & Communication Technology Promotion(IITP) grant funded by Korea government(MSIT) (No. R0118-16-1005, Digital Content In-House R&D).

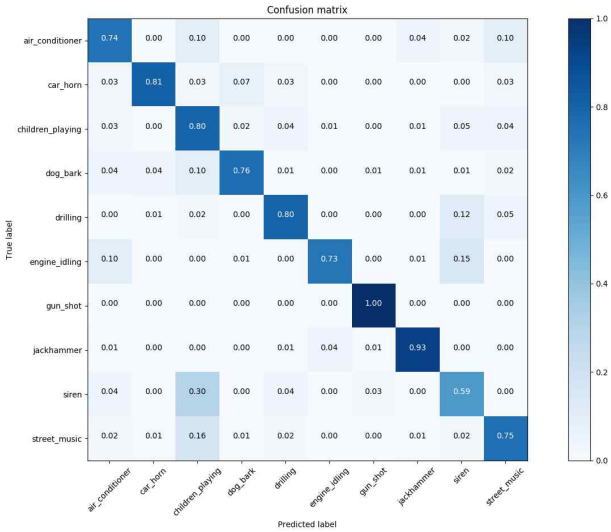

Fig. 3. Urbansound8k confusion matrix using CRNN8.

References

[1] D. P. Ellis, “Prediction-driven computational auditory scene analysis,” Ph.D. dissertation, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology (MIT), USA, 1996.

[2] R. Radhakrishnan, A. Divakaran, and P. Smaragdis, “Audio analysis for surveillance applications,” in IEEE WASPAA’05, 2005, pp. 158–161.

[3] M. Cristani, M. Bicego, and V. Murino, “Audio-visual event recognition in surveillance video sequences,” IEEE Trans. Multimedia, vol. 9, no. 2, pp. 257–267, 2007

[4] P. Foggia, N. Petkov, A. Saggese, N. Strisciuglio, and M. Vento, “Audio surveillance of roads: a system for detecting anomalous sounds,” IEEE Trans. on Intell. Transp. Syst., vol. 17, no. 1, pp. 279–288, 2016

[5] M. Xu, C. Xu, L. Duan, J. S. Jin, and S. Luo, “Audio keywords generation for sports video analysis,” ACM TOMCCAP, vol. 4, no. 2, pp. 1–23, 2008.

[6] L. Ballan, A. Bazzica, M. Bertini, A. Del Bimbo, and G. Serra, “Deep networks for audio event classification in soccer videos,” in Int. Conf. on Multimedia and Expo (ICME). New York, USA: IEEE, 2009, pp. 474–477.

[7] D. Steele, J. D. Krijnders, and C. Guastavino, “The sensor city initiative: cognitive sensors for soundscape transformations,” in GIS Ostrava, 2013, pp. 1–8.

[8] C. Mydlarz, J. Salamon, and J. P. Bello, “The implementation of lowcost urban acoustic monitoring devices,” Applied Acoustics, vol. In Press, 2016.

[9] B. Ghoraani and S. Krishnan, “Time–frequency matrix feature extraction and classification of environmental audio signals,” IEEE Trans. on Audio, Speech, and Lang. Process., vol. 19, no. 7, pp. 2197–2209, 2011.

[10] A. Mesaros, T. Heittola, O. Dikmen, and T. Virtanen, “Sound event detection in real life recordings using coupled matrix factorization of spectral representations and class activity annotations,” in International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brisbane, Australia, Apr. 2015, pp. 151–155.

[11] E. Benetos, G. Lafay, M. Lagrange, and M. D. Plumbley, “Detection of overlapping acoustic events using a temporally-

constrained probabilistic model," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, Mar. 2016, pp. 6450– 6454.

[12] V. Bisot, R. Serizel, S. Essid, and G. Richard, "Acoustic scene classification with matrix factorization for unsupervised feature learning," in Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 2016, pp. 6445–6449.

[13] J. Salamon and J. P. Bello, "Unsupervised feature learning for urban sound classification," in IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), Brisbane, Australia, Apr. 2015, pp. 171–175.

[14] P. Foggia, N. Petkov, A. Saggese, N. Strisciuglio, and M. Vento, "Audio surveillance of roads: a system for detecting anomalous sounds," IEEE Trans. on Intell. Transp. Syst., vol. 17, no. 1, pp. 279–288, 2016.

[15] S. Chu, S. Narayanan, and C.-C. J. Kuo, "Environmental sound recognition with time–frequency audio features," IEEE Trans. on Audio, Speech, and Lang. Process., vol. 17, no. 6, pp. 1142–1158, 2009.

[16] J. T. Geiger and K. Helwani, "Improving event detection for audio surveillance using gabor filterbank features," in 23rd European Signal Processing Conference (EUSIPCO), Nice, France, Aug. 2015, pp. 714– 718.

[17] D. M. Agrawal, H. B. Sailor, M. H. Soni, and H. A. Patil, "Novel TEO-based gammatone features for environmental sound classification," in submitted in European Signal Processing Conf. (EUSIPCO), Kos island, Greece, August 28 – 2 September 2017.

[18] S. Dieleman and B. Schrauwen, "End-to-end learning for music audio," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014, pp. 6964–6968.

[19] Zoltan Tüske, Pavel Golik, Ralf Schlüter, and Hermann Ney, "Acoustic modeling with deep neural networks using raw time signal for lvcsr.," in INTERSPEECH, 2014, pp. 890–894.

[20] Pavel Golik, Zoltan Tüske, Ralf Schlüter, and Hermann Ney, "Convolutional neural networks for acoustic modeling of raw time signal in lvcsr," in Sixteenth Annual Conference of the International Speech Communication Association, 2015.

[21] Tara N Sainath, Ron J Weiss, Andrew Senior, Kevin W Wilson, and Oriol Vinyals, "Learning the speech frontend with raw waveform cldnns," in Proc. Interspeech, 2015.

[22] Y. Tokozume and T. Harada, "Learning environmental sound with end-to-end convolutional neural network," in IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP), New Orleans, USA, 2017, pp. 2721–2725.

[23] H. B. Sailor, D. M. Agrawal, and H. A. Patil, "Unsupervised filterbank learning using convolutional restricted boltzmann machine for environmental sound classification," in Interspeech, Stockholm, Sweden, August 2017, pp. 3107-3111.

[24] Wei Dai, Chia Dai, Shuhui Qu, Juncheng Li, and Samarjit Das. Very deep convolutional neural networks for raw waveforms. In IEEE ICASSP, pages 421–425, 2017.

[25] Duyu Tang, Bing Qin, and Ting Liu, "Document modeling with gated recurrent neural network for sentiment classification," in Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 2015, pp. 1422–1432.

[26] Sharath Adavanne, Konstantinos Drossos, Emre Çakır, and Tuomas Virtanen. Stacked convolutional and recurrent neural networks for bird audio detection. In EUSIPCO, 2017.

[27] Miroslav Malik, Sharath Adavanne, Konstantinos Drossos, Tuomas Virtanen, Dasa Ticha, and Roman Jarina. Stacked convolutional and recurrent neural networks for music emotion recognition. In Sound and Music Computing Conference (SMC), 2017.

[28] Sergey Ioffe and Christian Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," arXiv preprint arXiv:1502.03167, 2015.

[29] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," Journal of Machine Learning Research, 2014.

[30] J. Lee and J. Nam. Multi-level and multi-scale feature aggregation using pre-trained convolutional neural networks for music auto-tagging. arXiv preprint arXiv:1703.01793, 2017.

[31] Lee, J., Park, J., Kim, K. L., and Nam, J. Sample-level deep convolutional neural networks for music auto-tagging using raw waveforms. Sound and Music Computing Conference (SMC), pp. 220–226, 2017.

[32] Justin Salamon, Christopher Jacoby, and Juan Pablo Bello, "A dataset and taxonomy for urban sound research," in Proceedings of the 22nd ACM international conference on Multimedia. ACM, 2014, pp. 1041–1044.

[33] J. Salamon and J. P. Bello, "Feature learning with deep scattering for urban sound analysis," in 2015 23rd European Signal Processing Conference (EUSIPCO). IEEE, 2015, pp. 724–728.

[34] Franc¸ois Chollet, "Keras," GitHub repository: https://github. com/fchollet/keras, 2015.

[35] Abadi, Mart´ın, Agarwal, Ashish, Barham, Paul et al., TensorFlow: Large-scale machine learning on heterogeneous systems, 2015.

[36] Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto, "librosa: Audio and music signal analysis in python," in Proceedings of the 14th Python in Science Conference, 2015.

[37] Diederik P. Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," CoRR, vol. abs/1412.6980, 2014.

[38] Xavier Glorot and Yoshua Bengio, "Understanding the difficulty of training deep feedforward neural networks.," in Aistats, 2010, vol. 9, pp. 249–256.

ISBN 978-90-827970-1-5 © EURASIP 2018