Proceedings of the Federated Conference on Computer Science and Information Systems pp. 57–60
DOI: 10.15439/2019F185
ISSN 2300-5963 ACSIS, Vol. 18

Urban Sound Classification using Long Short-Term Memory Neural Network

Iurii Lezhenin, Natalia Bogach
Institute of Computer Science and Technology
Peter the Great St.Petersburg Polytechnic University
St.Petersburg, 195251, Russia
Email: {lezhenin, bogach}@kspt.icc.spbstu.ru

Evgeny Pyshkin
Software Engineering Lab
University of Aizu
Aizu-Wakamatsu, 965-8580, Japan
Email: pyshe@u-aizu.ac.jp

***Abstract*—Environmental sound classification has received more attention in recent years. Analysis of environmental sounds is difficult because of its unstructured nature. However, the presence of strong spectro-temporal patterns makes the classification possible. Since LSTM neural networks are efficient at learning temporal dependencies we propose and examine a LSTM model for urban sound classification. The model is trained on magnitude mel-spectrograms extracted from UrbanSound8K dataset audio. The proposed network is evaluated using 5-fold cross-validation and compared with the baseline CNN. It is shown that the LSTM model outperforms a set of existing solutions and is more accurate and confident than the CNN.**

***Index Terms*—environmental sound classification, long short-term memory, convolutional neural networks, UrbanSound8K dataset**

I. Introduction

AUDIO recognition algorithms are traditionally used for the tasks of speech and music signal processing. Meanwhile, the problems of environmental sound recognition and classification have received much attention in recent years. There are multiple applications already proposed in a big variety of industries, including surveillance [1], [2], audio scene recognition for robot navigation [3], acoustic monitoring of natural and artificial environment [4]–[6]. In digitally transformed society [7], soundscape models create a research perspective in smart city domain. City noise managing significantly contributes to a healthy and safe living environment in the big cities [8]. In travel centric systems, city sounds may enter the emerging solutions to develop and share journey experience [9], [10]. Assisting technologies for people with disabilities and, in particular, navigation systems for blind or visually impaired people effectively incorporate urban sound models [11].

Environmental sound analysis is more complex than speech and music processing because of unstructured nature of sounds. There are no meaningful sequences of elementary blocks like phonemes or strong stationary patterns such as melody or rhythm. However, environmental sounds may include strong spectro-temporal signatures. Thus, it is important to consider non-stationary aspects of signal and capture its variation in both time and frequency domains.

The classification of environmental sounds is often split into auditory scene classification and sound classification by its source. But, both problems share the similar approaches. The methods used involve k-Nearest Neighbors (k-NN) algorithm, Support Vector Machine (SVM), Gaussian Mixture Model (GMM) and Hidden Markov Model (HMM) in combination with features engineered by signal processing techniques, e.g. Mel-Frequency Cepstral Coefficients (MFCC), Discrete Wavelet Transform (DWT) coefficients and Matching Pursuit (MP) features [12]–[14]. In contrast with described approaches, deep neural networks (DNN) allow to facilitate feature engineering keeping classification accuracy and even outperform the conventional solutions [15]. In particular, being able to capture spectro-temporal patterns from spectogram-like input convolutional neural networks (CNN) have high performance [16]–[19]. Long short-term memory (LSTM) networks is the other type of neural network architectures that is exploited for sound classification [20], as well as the combinations of LSTM and CNN [21], [22].

LSTM networks are recurrent neural networks (RNN) that use the contextual information over long time intervals to map the input sequence to the output. LSTM network is a general solution, efficient at learning temporal dependencies. Its application is beneficial in a variety of tasks, such as phoneme classification [23], speech recognition [24] and speech synthesis [25]. LSTM network combined with CNN was also successfully used for video classification [26].

The applicability of LSTM for sound classification hasn't been fully investigated so far. In this paper we examine a LSTM model to improve understanding of its applicability specifically for urban sounds classification using UrbanSound8K dataset [27]. Table A1 in Appendix summarizes some of the existing solutions where models are evaluated on UrbanSound8K. The baseline accuracy of 70% was obtained with SVM processing mel-bands and MFCC statistically summarized across the time [27]. The unsupervised feature learning using Spherical K-Means (SKM) performed on PCA-whitened log-scaled mel-spectrograms allows to achieve 73.6% accuracy [28]. CNNs of different architectures trained on log-scaled mel-spectrogram frames provide 73% of accuracy and 79% with data augmentation [16], [17]. The LSTM based CRNN for urban sound classification demonstrates 79.06% accuracy using raw waveforms [22]. The accuracy of 93% was shown by GoogLeNet trained on combination

of mel-spectrogram, MFCC and Cross Recurrence Plot (CRP) images [18].

The paper is structured as follows: Section II describes the LSTM model studied and the experimental setup. In Section III we present and discuss our results, and, finally, in Section IV we conclude about the LSTM applicability for urban sound classification and provide directions for future work.

II. Method

A. Long-short term memory neural network model

LSTM neural network is a special kind of RNN, that doesn't suffer from vanishing gradient problem and is able to learn long-term dependencies. LSTM consists of a set of subnets, known as memory blocks. Each block includes the memory cell and three units: input, output and forget gates.

LSTM layer maps the input sequence $X = (x_1, x_2, \dots x_T)$ to the output sequence $Y = (y_1, y_2, \dots y_T)$ in according to the equations:

$$i_t = \text{sig}(W_{xi}x_t + W_{yi}y_{t-1} + b_i), \quad (1)$$

$$f_t = \text{sig}(W_{xf}x_t + W_{yf}y_{t-1} + b_f), \quad (2)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tanh(W_{xc}x_t + W_{yc}y_{t-1} + b_c), \quad (3)$$

$$o_t = \text{sig}(W_{xo}x_t + W_{yo}y_{t-1} + b_o), \quad (4)$$

$$y_t = o_t \odot \tanh(c_t), \quad (5)$$

where c_t is the state of the memory cell and i_t, f_t, o_t are gate outputs at time t. The network weights W and biases b are tuned during learning to minimize the loss function. In case of a multi-layer structure the input of the next layer is the output of the previous one.

Our model for sound classification is composed of two LSTM layers followed by dense layer with *softmax* activation function. Though LSTM produces a sequence, only the last value is propagated to the output layer. The first two layers contain 128 and 64 units, the last layer has 10 units, one per sound class. To reduce overfitting dropout with a rate of 0.25 is applied to the output of the LSTM layers. For training *categorical cross-entropy* loss function is minimized using Adam optimizer. Because of long training time a full search of hyperparameters is infeasible, thus, the most promising combination was found using single fold evaluation.

The input of our model is magnitude mel-spectrogram with 128 bands, that covers a frequency range from 0 Hz to 22050 Hz. Spectrogram is evaluated at sample rate 44100 Hz using 1024 sample window and a hop size of the same width. The length of input sequence is variable and depends upon audio clip duration.

Among the examined variants the proposed model shows the best performance on input data normalized as follows:

$$\mu = \frac{1}{T}\frac{1}{N}\sum_{t=1}^{T}\sum_{n=1}^{N} x_t^{(n)}, \quad (6)$$

$$\sigma = \sqrt{\frac{1}{T}\frac{1}{N}\sum_{t=1}^{T}\sum_{n=1}^{N}(x_t^{(n)} - \mu)^2}, \quad (7)$$

$$X_{norm} = \frac{X - \mu}{\sigma}, \quad (8)$$

where X is the input sequence; $x_t^{(n)}$ is the value of n-th feature at time t; N is a number of features and T is a number of time steps. Normalization in both dimensions allows to keep spectro-temporal energy distribution pattern and eliminate the difference between the audio clips across the dataset in terms of linear distortion.

B. Experimental setup

To evaluate the performance of proposed model we use UrbanSound8K dataset [27], that contains 8732 sound clips of up to 4 s in duration divided into 10 sound classes:air conditioner (AI), car horn (CA), children playing (CH), dog bark (DO), drilling (DR), engine idling (EN), gun shot (GU), jackhammer (JA), siren (SI), street music (ST).

Along with our model we run a baseline CNN [17]. CNN is composed of three convolutional layers followed by two dense layers. Both networks were trained on magnitude mel-spectrogram and CNN model indicated even better performance than was reported in [17] for log-scaled mel-spectrogram. We use a simplified validation algorithm for CNN: in contrast with [17], frame is being extracted from test sample at random, yet the CNN model holds the reported level of accuracy.

We randomly divide the dataset into 5 folds of the same size and carry out cross-validation to evaluate the networks performance. Models were trained on four folds and tested on the last one. The training duration is limited by 64 epochs. The train loss, train accuracy, test loss and test accuracy are saved for each epoch. The final accuracy is taken as the best validation accuracy achieved in the course of training.

Both models were implemented[1] with Keras, a high-level neural network API, written in Python. To resample the audio clips and extract the mel-spectrum we use the Librosa Python library.

III. Results and discussion

Both models show the similar performance, their cross-validation results are presented in Fig. 1. While CNN provides 81.67% average accuracy, the proposed LSTM network

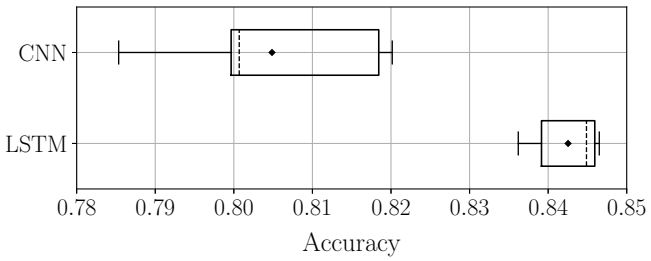

Fig. 1: Classification accuracy. Average accuracy is 80.48% and 84.25% for CNN and LSTM, respectively.

[1]Source code in Python available as Jupyter notebooks at https://github.com/lezhenin/lstm-sound-classification-2019

TABLE I: Per-class and averaged Precision, Recall and F1 score for CNN and LSTM.

		AI	CA	CH	DO	DR	EN	GU	JA	SI	ST	Macro-average
LSTM	Precision	0.80	0.82	0.78	0.86	0.87	0.88	0.93	0.89	0.90	0.75	0.85
	Recall	0.88	0.85	0.73	0.83	0.87	0.85	0.94	0.91	0.91	0.73	0.85
	F1	0.84	0.83	0.75	0.84	0.87	0.86	0.94	0.90	0.90	0.74	0.85
CNN	Precision	0.74	0.94	0.63	0.85	0.86	0.80	0.93	0.87	0.95	0.70	0.83
	Recall	0.83	0.79	0.71	0.80	0.81	0.84	0.89	0.84	0.83	0.73	0.81
	F1	0.78	0.86	0.67	0.83	0.83	0.82	0.91	0.85	0.88	0.71	0.82

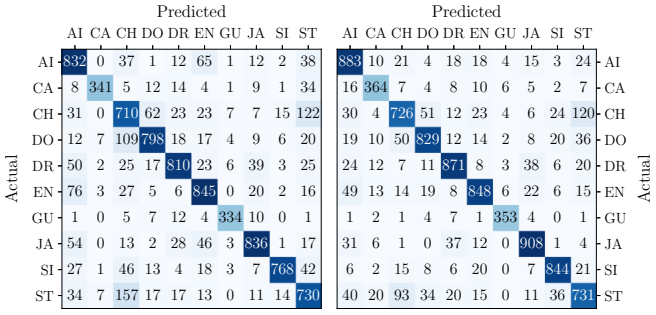

Fig. 2: Confusion matrices for CNN (left) and LSTM (right).

achieves 84.25%. The two models outperform the baseline methods. But LSTM demonstrates less accuracy distribution range and, thus, is more robust.

Confusion matrices obtained on test data during cross-validation is shown in Fig. 2. The same two pairs of classes demonstrate high confusion: street music vs. children playing and children playing vs. dog bark. These sounds may have complex time-frequency structure which impedes their accurate classification.

Precision, recall and F1 calculated for each class using confusion matrices are presented in Table I. LSTM shows slightly higher F1 score for each class, except car horn, and outperforms CNN in average. Also CNN may decrease recall to increase the overall accuracy, especially for unbalanced classes (e.g car horn and siren). Thus, LSTM performs better keeping not only accuracy but recall and precision as well.

We compare training as accuracy and loss across epochs in Fig. 3. Both networks achieve the ultimate performance on test data approximately at 20-th epoch. Having almost equal accuracy the two models differ in their loss values. LSTM network shows a significantly smaller loss. It means LSTM is more confident in its predictions and has wider margins between classes. Thus, it is more robust.

The CNN holds accuracy and loss over train and test data. In contrast, LSTM model shows the better performance on train data. It doesn't fully generalize from train to unseen test data and memorizes the details that don't affect the overall performance. It may indicates that the model is redundant. Because of its recurrent structure the LSTM is more computationally intensive and prone to overfitting, although has less trainable parameters than CNN: 181K vs. 241K. So, it is highly probable that the model may be simplified without a significant performance degradation. Additional regularization techniques may also be beneficial.

IV. Conclusion

LSTM network that take magnitude mel-spectrograms was shown to be a reliable classifier in application for urban sounds. It provides the 84.25% of average accuracy and thus exceeds the majority of existing solutions. In comparison with baseline CNN trained on the same data LSTM has a little performance increase and is more confident.

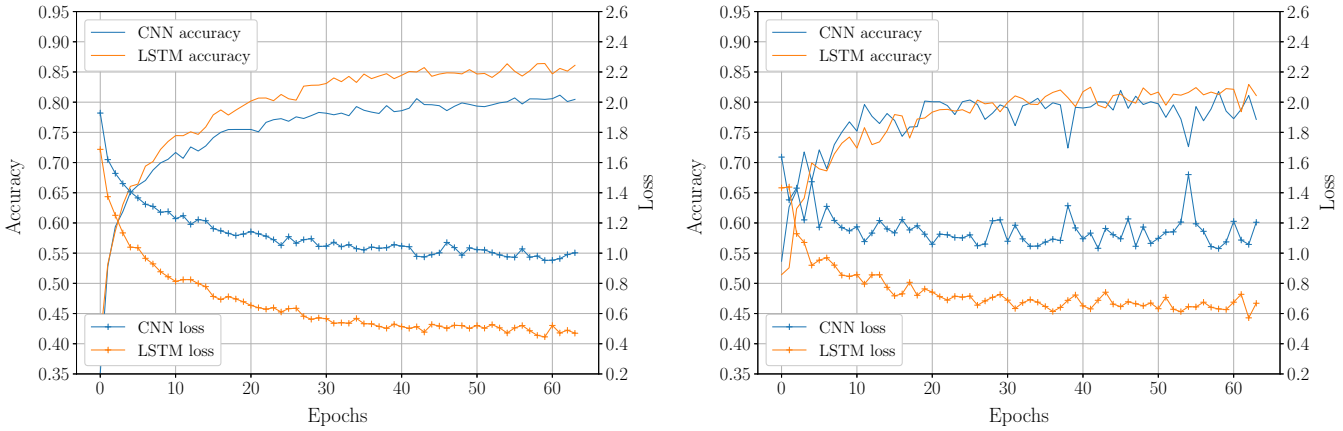

Fig. 3: Accuracy and loss evaluated on train data (left) and test data (right) during training.

The further study may develop towards the model simplification and regularization or involve new data not limited by urban setting.

APPENDIX

TABLE A1: Classification accuracy on UrbanSound8K dataset

Reference	Classifier	Features	Accuracy
[27]	SVM	mel-bands and MFCC	70%
[28]	SKM	PCA whitened mel-bands	73%
[16]	CNN	log mel-spectrogram	73%
[17]	CNN	log mel-spectrogram	73%
	CNN + aug		79%
[22]	CRNN	raw waveforms	79%
this paper	LSTM	mel-spectrogram	83%
[18]	CNN (GoogLeNet)	mel-spectrogram, MFCC, CRP images	93%

ACKNOWLEDGMENT

This work was partially supported by the grant 17K00509 of Japan Society for the Promotion of Science (JSPS).

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. [Online]. Available: https://doi.org/10.1109/ASPAA.2005.1540194

[2] M. Cristani, M. Bicego, and V. Murino, "Audio-visual event recognition in surveillance video sequences," *IEEE Transactions on Multimedia*, vol. 9, no. 2, pp. 257–267, 2007. [Online]. Available: https://doi.org/10.1109/TMM.2006.886263

[3] S. Chu, S. Narayanan, C.-C. J. Kuo, and M. J. Mataric, "Where am i? scene recognition for mobile robots using audio features," in *2006 IEEE International conference on multimedia and expo*. IEEE, 2006, pp. 885–888. [Online]. Available: https://doi.org/10.1109/ICME.2006.262661

[4] R. Bardeli, D. Wolff, F. Kurth, M. Koch, K.-H. Tauchert, and K.-H. Frommolt, "Detecting bird sounds in a complex acoustic environment and application to bioacoustic monitoring," *Pattern Recognition Letters*, vol. 31, no. 12, pp. 1524–1534, 2010. [Online]. Available: https://doi.org/10.1016/j.patrec.2009.09.014

[5] C. Mydlarz, J. Salamon, and J. P. Bello, "The implementation of low-cost urban acoustic monitoring devices," *Applied Acoustics*, vol. 117, pp. 207–218, 2017. [Online]. Available: https://doi.org/10.1016/j.apacoust.2016.06.010

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

[7] V. Davidovski, "Exponential innovation through digital transformation," in *Proceedings of the 3rd International Conference on Applications in Information Technology*. ACM, 2018, pp. 3–5. [Online]. Available: https://doi.org/10.1145/3274856.3274858

[8] F. Tappero, R. M. Alsina-Pagès, L. Duboc, and F. Alías, "Leveraging urban sounds: A commodity multi-microphone hardware approach for sound recognition," in *Multidisciplinary Digital Publishing Institute Proceedings*, vol. 4, no. 1, 2019, p. 55. [Online]. Available: https://doi.org/10.3390/ecsa-5-05756

[9] E. Pyshkin, "Designing human-centric applications: Transdisciplinary connections with examples," in *2017 3rd IEEE International Conference on Cybernetics (CYBCONF)*. IEEE, 2017, pp. 1–6. [Online]. Available: https://doi.org/10.1109/CYBConf.2017.7985774

[10] E. Pyshkin and A. Kuznetsov, "Approaches for web search user interfaces-how to improve the search quality for various types of information," *JoC*, vol. 1, no. 1, pp. 1–8, 2010. [Online]. Available: https://www.earticle.net/Article/A188181

[11] M. B. Dias, "Navpal: Technology solutions for enhancing urban navigation for blind travelers," *tech. report CMU-RI-TR-21, Robotics Institute, Carnegie Mellon University*, 2014.

[12] 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. [Online]. Available: https://doi.org/10.1109/TASL.2009.2017438

[13] S. Chachada and C.-C. J. Kuo, "Environmental sound recognition: A survey," vol. 3, 10 2013, pp. 1–9. [Online]. Available: https://doi.org/10.1109/APSIPA.2013.6694338

[14] D. Giannoulis, E. Benetos, D. Stowell, M. Rossignol, M. Lagrange, and M. Plumbley, "Detection and classification of acoustic scenes and events: An ieee aasp challenge," 10 2013, pp. 1–4. [Online]. Available: https://doi.org/10.1109/WASPAA.2013.6701819

[15] Z. Kons, O. Toledo-Ronen, and M. Carmel, "Audio event classification using deep neural networks." in *Interspeech*, 2013, pp. 1482–1486.

[16] 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. [Online]. Available: https://doi.org/10.1109/MLSP.2015.7324337

[17] 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. [Online]. Available: https://doi.org/10.1109/LSP.2017.2657381

[18] 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. [Online]. Available: https://doi.org/10.1016/j.procs.2017.08.250

[19] B. Zhu, K. Xu, D. Wang, L. Zhang, B. Li, and Y. Peng, "Environmental sound classification based on multi-temporal resolution convolutional neural network combining with multi-level features," in *Pacific Rim Conference on Multimedia*. Springer, 2018, pp. 528–537. [Online]. Available: https://doi.org/10.1007/978-3-030-00767-6_49

[20] Y. Wang, L. Neves, and F. Metze, "Audio-based multimedia event detection using deep recurrent neural networks," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 2742–2746. [Online]. Available: https://doi.org/10.1109/ICASSP.2016.7472176

[21] S. H. Bae, I. Choi, and N. S. Kim, "Acoustic scene classification using parallel combination of lstm and cnn," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE2016)*, 2016, pp. 11–15.

[22] J. Sang, S. Park, and J. Lee, "Convolutional recurrent neural networks for urban sound classification using raw waveforms," in *2018 26th European Signal Processing Conference (EUSIPCO)*. IEEE, 2018, pp. 2444–2448. [Online]. Available: https://doi.org/10.23919/EUSIPCO.2018.8553247

[23] A. Graves, S. Fernández, and J. Schmidhuber, "Bidirectional lstm networks for improved phoneme classification and recognition," in *International Conference on Artificial Neural Networks*. Springer, 2005, pp. 799–804. [Online]. Available: https://doi.org/10.1007/11550907_126

[24] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *2013 IEEE international conference on acoustics, speech and signal processing*. IEEE, 2013, pp. 6645–6649. [Online]. Available: https://doi.org/10.1109/ICASSP.2013.6638947

[25] Y. Fan, Y. Qian, F.-L. Xie, and F. K. Soong, "Tts synthesis with bidirectional lstm based recurrent neural networks," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[26] J. Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, "Beyond short snippets: Deep networks for video classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4694–4702. [Online]. Available: https://doi.org/10.1109/CVPR.2015.7299101

[27] 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. [Online]. Available: https://doi.org/10.1145/2647868.2655045

[28] J. Salamon and J. P. Bello, "Unsupervised feature learning for urban sound classification," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 171–175. [Online]. Available: https://doi.org/10.1109/ICASSP.2015.7177954