Proceedings of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE2018)

Tampereen teknillinen yliopisto - Tampere University of Technology

Mark D. Plumbley, Christian Kroos, Juan P. Bello, Gaël Richard, Daniel P. W. Ellis, Annamaria Mesaros (eds.)

Proceedings of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE2018)

TAMPEREEN TEKNILLINEN YLIOPISTO
TAMPERE UNIVERSITY OF TECHNOLOGY

Tampereen teknillinen yliopisto - Tampere University of Technology

Mark D. Plumbley, Christian Kroos, Juan P. Bello, Gaël Richard, Daniel P. W. Ellis, Annamaria Mesaros (eds.)

Proceedings of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE2018)

Tampere University of Technology. Laboratory of Signal Processing
Tampere 2018

ISBN 978-952-15-4262-6

ATTENTION-BASED CONVOLUTIONAL NEURAL NETWORKS FOR ACOUSTIC SCENE CLASSIFICATION

Zhao Ren[1], *Qiuqiang Kong*[2], *Kun Qian*[1], *Mark D. Plumbley*[2], *Björn W. Schuller*[1,3]

[1] ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany
[2] Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey, UK
[3] GLAM – Group on Language, Audio & Music, Imperial College London, UK

zhao.ren@informatik.uni-augsburg.de, schuller@ieee.org

ABSTRACT

We propose a convolutional neural network (CNN) model based on an attention pooling method to classify ten different acoustic scenes, participating in the acoustic scene classification task of the IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE 2018), which includes data from one device (subtask A) and data from three different devices (subtask B). The log mel spectrogram images of the audio waves are first forwarded to convolutional layers, and then fed into an attention pooling layer to reduce the feature dimension and achieve classification. From attention perspective, we build a weighted evaluation of the features, instead of simple max pooling or average pooling. On the official development set of the challenge, the best accuracy of subtask A is 72.6 %, which is an improvement of 12.9 % when compared with the official baseline ($p < .001$ in a one-tailed z-test). For subtask B, the best result of our attention-based CNN is a significant improvement of the baseline as well, in which the accuracies are 71.8 %, 58.3 %, and 58.3 % for the three devices A to C ($p < .001$ for device A, $p < .01$ for device B, and $p < .05$ for device C).

Index Terms— Acoustic Scene Classification, Convolutional Neural Network, Attention Pooling, Log Mel Spectrogram

1. INTRODUCTION

Acoustic scene classification, as a subfield of computational auditory scene analysis (CASA) [1], aims at enabling devices to recognise the acoustic environment. It has been successfully employed in a series of applications, including intelligent wearable interfaces [2, 3], smartphone navigation systems [4], context-aware computation [5], and many more. In the field of machine learning, a number of models have been applied in the past 'detection and classification of acoustic scenes and events' (DCASE) challenges, such as support vector machines [6, 7], hidden markov models [8], autoencoders [9], and convolutional neural networks (CNN) [10, 11]. In the acoustic scene classification task of the IEEE AASP Challenge in this year [12], researchers are provided with the opportunity to investigate training a robust model on a dataset from multiple recording devices. The acoustic scene classification task in this challenge includes two subtasks with different data sources – one is based on single recording device, and the other is based on three devices. The dataset has been divided into a 'development set' with a training/test partitioning, and a non-public evaluation set. Both of the subtasks require participants to classify the acoustic data into ten classes of scenes.

In recent years, time-frequency transformation images have shown their superiority in improving the performance in the acoustic scene classification task [13]. Different kinds of time-frequency transformation images have been applied for feature extraction, such as Constant-Q-Transform (CQT) spectrogram [13], Short-Time Fourier Transformation (STFT)-based spectrogram [9], scalogram [10], and log mel spectrogram [14]. In this paper, we use a log mel spectrogram image representation as it performed excellent in the acoustic scene classification task of DCASE 2017 [15].

With log mel spectrogram images, we construct an end-to-end CNN model for classification. A number of CNNs have been presented successfully in image processing, particularly in the ImageNet Large Scale Visual Recognition Challenge [16]. Compared with the dataset with around several hundreds of thousands samples in that challenge, the dataset in DCASE challenge contains less than ten thousands of audio waves for training. In this regard, CNNs with relatively more shallow layers than the CNNs for ImageNet, are utilised in our work, including AlexNet and VGG with four convolutional layers. In addition, a CNN topology with different structure and kernel size is designed to improve the performance.

To avoid over fitting caused by the large size of the feature maps that are obtained after the convolutional layers in the CNN, pooling usually serves to compute features by reducing the feature dimension. Max pooling and average pooling are the most frequently employed pooling models to obtain smaller feature dimension. Max pooling is achieved through extracting the largest value inside a filter as parameters of the max pooling layer [17, 18], and average pooling aims at obtaining the average value of a filter as its parameter [19]. Unfortunately, both of these pooling models cannot utilise each feature reasonably according to its contribution. Max pooling ignores other potentially helpful features besides the feature with the maximum value; average pooling treats each feature equally, easily leading to some suboptimal results because of the interference from useless features. To solve this problem, an attention model, attempting to compute the contribution of each feature, was proposed and utilised in many applications, including natural language processing [20], visual question answering [21], and even audio classification [22].

The two main contributions of our paper for acoustic scene classification are as follows. The first is that we design an end-to-end CNN to train the model, and compare it with the state-of-the-art CNN models. Second, we propose an attention-based CNN by weighting the contribution of each feature and explaining this model from a probability perspective in multiple instance learning [23].

The remainder of this paper is structured as follows: in Section 2, we describe the proposed approach, the pipeline of which is shown in Figure 1, the database description, experimental set up, and results, are presented in Section 3; finally, conclusions are given in Section 4.

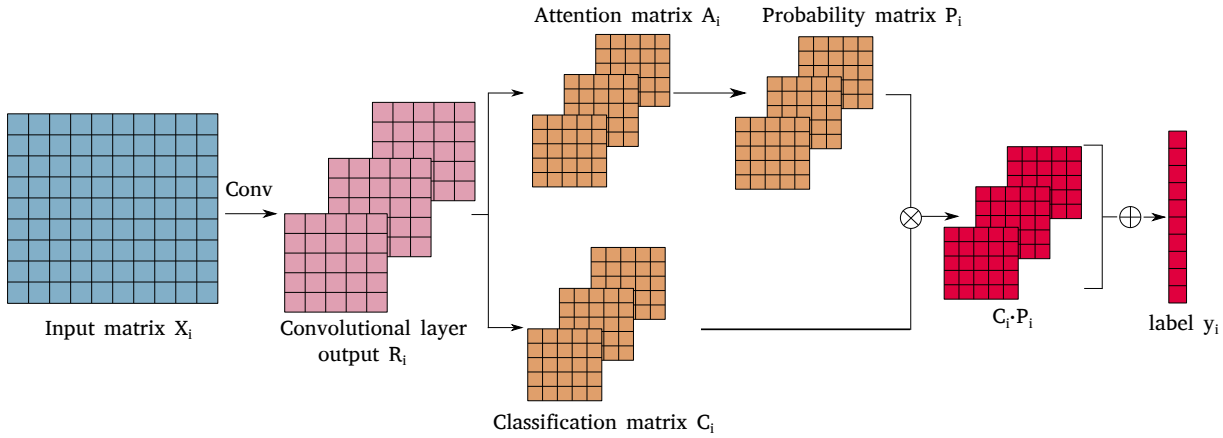

Figure 1: The framework of our proposed attention pooling system. First, the output of the convolutional layers R_i is obtained through the convolutional layers when X_i is the input matrix. The attention matrix A_i and the classification matrix C_i with the class number of the channels are then generated from R_i. Further, the probability P_i calculated from A_i multiplies with C_i for an element-wise product $C_i \cdot P_i$. The prediction y_i is finally obtained by summing up of $C_i \cdot P_i$.

Table 1: Configurations of the convolutional neural networks. Convolutional layers are denoted as 'the number of convolution layers × conv(receptive field size – number of channels)' with the stride 's(stride size)'.

AlexNet	VGG-4	our CNN
Input: image X_i $(1 \times m \times n)$		
1×conv11-64; s1 Maxpooling	1×conv3-64; s1 Maxpooling	1×conv5-64; s2
1×conv5-192; s1 Maxpooling	1×conv3-128; s1 Maxpooling	1×conv5-128; s2
1×conv3-384; s1 2×conv3-256; s1 Maxpooling	1×conv3-256; s1 1×conv3-512; s1 Maxpooling	1×conv5-256; s2 1×conv5-512; s2
Output: R_i $(c \times m' \times n')$		

2. METHODOLOGY

In this section, we describe our proposed neural network in two subsections. The first is dedicated to the convolutional neural network, and then the pooling models following classification will be introduced.

2.1. Convolutional Neural Networks

Given the successful application of image-based neural networks for acoustic scene classification tasks in [10, 24], we employ three end-to-end CNN topologies in this work, including 'AlexNet' [25], 'VGG-4' [26], and 'our CNN' with a different structure from the former two CNNs, as shown in Table 1.

The log mel spectrogram images with one channel are first extracted from audio waves and are fed into the CNN model. The $i-th$ image $X_i, i = 1, ..., N$ with size of $1 \times m \times n$, in which N is the number of images, could generate an output $R_i, i = 1, ..., N$ with size of $c \times m' \times n'$ by the convolutional layers of the CNN, where c means the number of channels. In Table 1, the AlexNet model uses the same parameters as the original AlexNet [25]. The conventional VGG structures [26] are not used in our work, since the DCASE dataset is much smaller than the ImageNet database [27]. Thereby, we design a VGG-4 with four convolutional layers and using the same kernel size of three with the typical VGG model for each layer. Additionally, a CNN without max pooling during convolution, but with a stride with size of 2, is designed to weaken the effect of max pooling, and with a kernel size inbetween AlexNet and VGG-4 to explore the effect of the kernel size in order to reach a better performance.

Please note that the rectified linear unit activation function 'ReLU' is applied for each convolutional layer. Different from the typical AlexNet or VGG, batch normalisation is employed for all convolutional layers in our work, as this can accelerate deep networks and improve the performance of CNNs [28, 29].

The output of the convolutional layers R_i has the size of $c \times m' \times n'$, which is not appropriate for classification of ten classes as demanded in our case due to its large size (2 000 for AlexNet, and at least 400 000 for VGG-4 and our CNN). Therefore, a pooling mechanism is employed to reduce the feature dimension in the next subsection.

2.2. Pooling Mechanism

The output R_i of each spectrogram image by the CNN, which could be viewed as a bag of instances, contains c feature maps with m' feature vectors at n' time steps. The matrix with size of (m', n') is an instance in a bag. Based on multi-instance learning [30, 23], we consider that the classification model is given a number of pairs $\{(R_i, y_i)\}, i = 1, ..., N$. For the matrix R_i, its correspondent label is $y_i \in \{0, 1\}^L$, where L is the number of the scene classes. To achieve the classification, it is necessary to reduce the dimension of R_i from three to single channel dimension. In this subsection, two traditional pooling methods, max pooling and average pooling, and our proposed attention pooling will be introduced.

2.2.1. Max Pooling

On the assumption that the maximum classification value of each instance is the prediction of a bag [31], the max pooling model is described as

$$R_i^* = \max_{1<q<n'} \max_{1<p<m'} R_i, \tag{1}$$

where R_i^* is a feature vector prepared to classify the labels by linear transformation. Max pooling has been widely applied in CNNs for image classification and performs well [27], but sometimes unsatisfactory as it loses the time and location information when only choosing the maximum value at the dimensions of the time steps and feature vectors.

2.2.2. Average Pooling

Based on the collective assumption in [32], we assume that all instances contribute equally for the prediction in a bag. Accordingly, the definition of average pooling is

$$R_i^* = \frac{1}{m'n'} \sum_{1<q<n'} \sum_{1<p<m'} R_i. \tag{2}$$

As average pooling weights the contribution of each instance equally, unfortunately, it is possible that it diminishes the effect of some important features and augments some noisy features, leading to potentially imperfect prediction results.

2.2.3. Attention Pooling

As mentioned, both the max and average pooling models cannot calculate fairly according to the contribution of each instance. Therefore, computing the contributions of instances in a bag, which aims to obtain the weight of each instance, is a challenging task. To solve this problem, an attention-based pooling model is proposed for the aimed at acoustic scene classification task, as shown in Figure 1.

As to the matrix R_i, we feed R_i twice into two parallel 1-by-1 convolution layers with the channel number of L and a kernel size of 1, to reduce the dimension of the feature maps to prepare for classification. Both convolutional layers are followed by activations, including sigmoid activation for the attention matrix A_i, and 'log softmax' for the classification matrix C_i. Therefore, two matrices A_i and C_i with size (L, m', n') are obtained. To compute the contribution of each instance, in other words the contributions of elements in each feature map, the probability matrix P_i is defined as

$$P_i = A_i / \sum_{1<q<n'} \sum_{1<p<m'} A_i. \tag{3}$$

With the probability matrix P_i, which holds the weight of each element of C_i, the prediction y_i is

$$y_i = \sum_{1<q<n'} \sum_{1<p<m'} C_i \cdot P_i. \tag{4}$$

Finally, the predicted label is obtained as the summing up of the element wise product of C_i and P_i.

Attention pooling is capable to overcome the disadvantage of max pooling and average pooling, weighting the contribution for each instance through a 1-by-1 convolutional layer. As shown in Figure 2, attention pooling can give weights for instances according to their contributions, thereby achieving more optimal prediction results.

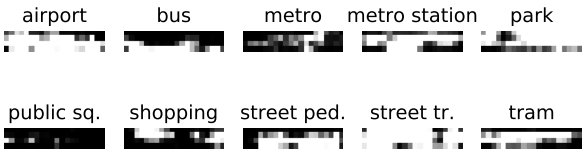

Figure 2: Heat maps of each scene class computed by the matrix P_i in our attention-based CNN. The heat map in this figure is the transpose matrix of P_i with a size of $(20, 4)$ for a better display. The horizontal axis represents the time steps, and the vertical axis the feature vectors.

3. EXPERIMENTAL RESULTS

3.1. Database

Our proposed approach is evaluated on the development dataset of the acoustic scene classification task in DCASE 2018 [12]. The dataset was recorded in various scene environments, and several locations for each scene. Each original recording with a length of 5-6 minutes was segmented into clips of 10 seconds. The sampling rate is set to 44.1 kHz in our work. The dataset contains 10 scene classes, including *airport*, *shopping mall*, *metro station*, *street pedestrian*, *public square*, *street traffic*, *tram*, *bus*, *metro*, and *park*. The task consists of two subtasks according to different recording devices, which comprise Soundman OKM II Klassik/studio A3, an electret binaural microphone, and a Zoom F8 (referred to device A), and two customer devices, e. g. smartphones, cameras, (referred to device B and C), thereby two sub-datasets are provided:

1) TUT Urban Acoustic Scenes 2018, was recorded by device A with 8 640 segments in total.

2) TUT Urban Acoustic Scenes 2018 Mobile, contains recordings from devices A, B, and C. In this dataset, the recordings are made up of 8 640 audio files from device A, and 720 audio files by device B and C in parallel.

3.2. Experimental Setup

The log mel spectrogram images are firstly extracted from each audio wave, with a Hamming window size of 2 048, overlap of 672, and 64 mel bands. Therefore, a feature map with a size of (320, 64) is generated for each audio file. The features are then fed into CNNs as mentioned in Section 2, using the 'Adam' optimiser with a learning rate of 0.001. The CNNs are optimised during 3000 maximum iteration steps, which are empirically set. As the accuracy on test set floats in a small interval after convergence, the iteration step corresponding to the highest accuracy during all iterations is chosen as the step where the training is stopped. The CNN architectures are implemented using Pytorch[1].

3.3. Results and Discussion

The results evaluated on the development set are shown in Table 2. We can see that, nearly all of our pooling models achieve improvements compared to the official baseline system. The attention pooling model performs better than max and average pooling models at AlexNet and our CNN. However, the attention pooling model at

[1] https://pytorch.org/

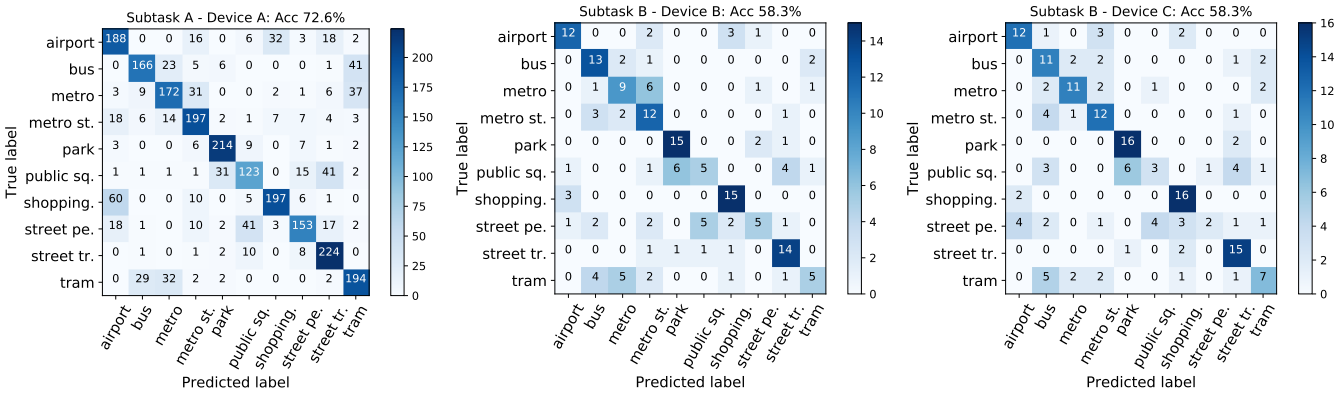

Figure 3: Confusion matrices of device A in subtask A and device B and C in subtask B by the best model. Our proposed CNN with attention pooling is the best model for both subtask A (SUBA) and subtask B (SUBB).

Table 2: Performance comparison of the baseline and CNN topologies of AlexNet, VGG-4, and our CNN, with three pooling models – max, average, and attention, evaluated on the official development set of Subtask A (SUBA) and Subtask B (SUBB). The dataset recorded by device A is employed for the evaluation of Subtask A, and the datasets from device A, B, and C are used for Subtask B. $\overline{(B,C)}$ stands for the mean evaluation result of device B and C. The experimental results are evaluated by accuracy [%].

		SUBA	**SUBB**			
NN	Pooling	A	A	B	C	$\overline{(B,C)}$
Baseline		59.7	58.9	45.1	46.2	45.6
AlexNet	max	67.2	62.2	51.7	54.4	53.1
AlexNet	average	64.3	60.8	51.1	52.2	51.7
AlexNet	attention	67.2	64.2	53.3	46.7	50.0
VGG-4	max	68.7	66.8	53.9	56.1	55.0
VGG-4	average	63.7	63.2	52.8	48.9	50.8
VGG-4	attention	67.6	64.6	50.6	46.1	48.3
our CNN	max	68.1	68.7	58.3	55.6	56.9
our CNN	average	68.1	67.3	**59.4**	56.1	57.8
our CNN	attention	**72.6**	**71.8**	58.3	**58.3**	**58.3**

VGG-4 yields to max pooling, the possible reason might be that the larger number of hyper parameters in VGG-4 with attention pooling brings on over fitting. As to the different CNN models, the best results are obtained by our CNN, which means that CNNs with a kernel size of five and no max pooling among convolutional layers appear more suited for this acoustic scene classification task. Our CNN with attention model achieves accuracy of 72.6 % for subtask A, which is a significant improvement over the baseline ($p < .001$ in a one-tailed z-test). In addition, our CNN achieves accuracies of 71.8 %, 58.3 %, and 58.3 % for device A, B, C in subtask B (in a one-tailed z-test, $p < .001$ for device A , $p < .01$ for device B, and $p < .05$ for device C).

The CNN model performs better at subtask A than at subtask B, perhaps because multiple data recording devices were employed for subtask B. The average accuracy of device A is higher than of device B and C in subtask B, which is considered to be caused by the unbalance of data among the three devices; in other words, the dataset contains more data from device A than from device B and C. To investigate our CNN model, a performance comparison of accuracy for each scene class of our best result is presented in Figure 3. Our CNN is optimal for some classes like *park*, *metro station*, and *street traffic*, but it is less able to recognise some classes such as *public square* and *bus*. It is possible that this lower performance is caused by background noise in these classes.

To sum up, our proposed CNNs with an attention model appear helpful to improve the performance over other pooling models for acoustic scene classification tasks.

4. CONCLUSIONS AND PERSPECTIVES

We proposed an attention-based convolutional neural network for acoustic scene classification by building an attention model at the decision level for classification. Based on the official development set of the acoustic scene classification task of the IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE 2018), our CNN gave better performance than other state-of-the-art CNN models and the proposed attention pooling model performed better than max pooling and average pooling, and achieved a significant improvement of the official baseline at subtask A ($p < .001$ in a one-tailed z-test) and subtask B (in a one-tailed z-test, $p < .001$ for device A , $p < .01$ for device B, and $p < .05$ for device C).

In future works, we will investigate the attention model at the feature level in order to analysis the contributions of feature maps in each convolutional layer. Further, transfer learning will be considered for subtask B, for training a robust model on the dataset from multiple recording sources.

5. ACKNOWLEDGMENT

This work was partially supported by the Horizon H2020 Marie Skłodowska-Curie Actions Initial Training Network European Training Network project under grant agreement No. 766287 (TAPAS), the European Union's Seventh Framework under grant agreement No. 338164 (ERC StG iHEARu), the EPSRC grant EP/N014111/1 "Making Sense of Sounds", and a Research Scholarship from the China Scholarship Council (CSC) No. 201406150082.

6. REFERENCES

[1] B. T. Szabó, S. L. Denham, and I. Winkler, "Computational models of auditory scene analysis: A review," *Frontiers in Neuroscience*, vol. 10, p. 524, Nov. 2016.

[2] Y. Xu, W. J. Li, and K. K. Lee, *Intelligent wearable interfaces*. John Wiley & Sons, 2008.

[3] F. Eyben, F. Weninger, F. Groß, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proc. ACM MM*, Barcelona, Catalunya, Spain, 2013, pp. 835–838.

[4] S. Chu, S. Narayanan, C.-C. Kuo, and M. Mataric, "Where am I? Scene recognition for mobile robots using audio features," in *Proc. ICME*, Toronto, Canada, 2006, pp. 885–888.

[5] C. Perera, A. Zaslavsky, P. Christen, and D. Georgakopoulos, "Context aware computing for the internet of things: A survey," *IEEE Communications Surveys & Tutorials*, vol. 16, no. 1, pp. 414–454, May 2013.

[6] E. Marchi, D. Tonelli, X. Xu, F. Ringeval, J. Deng, S. Squartini, and B. Schuller, "Pairwise decomposition with deep neural networks and multiscale kernel subspace learning for acoustic scene classification," in *Proc. DCASE Workshop*, Budapest, Hungary, 2016, pp. 65–69.

[7] K. Qian, Z. Ren, V. Pandit, Z. Yang, Z. Zhang, and B. Schuller, "Wavelets revisited for the classification of acoustic scenes," in *Proc. DCASE Workshop*, Munich, Germany, 2017, pp. 108–112.

[8] M. Chum, A. Habshush, A. Rahman, and C. Sang, "IEEE AASP scene classification challenge using hidden Markov models and frame based classification," in *Proc. DCASE challenge*, Munich, Germany, 2013, 3 pages.

[9] S. Amiriparian, M. Freitag, N. Cummins, and B. Schuller, "Sequence to sequence autoencoders for unsupervised representation learning from audio," in *Proc. DCASE Workshop*, Munich, Germany, 2017, pp. 17–21.

[10] Z. Ren, V. Pandit, K. Qian, Z. Yang, Z. Zhang, and B. Schuller, "Deep sequential image features on acoustic scene classification," in *Proc. DCASE Workshop*, Munich, Germany, 2017, pp. 113–117.

[11] S. H. Bae, I. Choi, and N. S. Kim, "Acoustic scene classification using parallel combination of LSTM and CNN," in *Proc. DCASE Workshop*, Budapest, Hungary, 2016, pp. 11–15.

[12] http://dcase.community/workshop2018/.

[13] W. Zheng, J. Yi, X. Xing, X. Liu, and S. Peng, "Acoustic scene classification using deep convolutional neural network and multiple spectrograms fusion," in *Proc. DCASE Workshop*, Munich, Germany, 2017, pp. 133–137.

[14] S. Adavanne and T. Virtanen, "Sound event detection using weakly labeled dataset with stacked convolutional and recurrent neural network," in *Proc. DCASE Workshop*, Munich, Germany, 2017, pp. 12–16.

[15] Y. Han, J. Park, and K. Lee, "Convolutional neural networks with binaural representations and background subtraction for acoustic scene classification," in *Proc. DCASE Workshop*, Munich, Germany, 2017, pp. 1–5.

[16] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, Apr. 2015.

[17] H. Phan, L. Hertel, M. Maass, and A. Mertins, "Robust audio event recognition with 1-max pooling convolutional neural networks," in *INTERSPEECH*, San Francisco, CA, 2016, pp. 3653–3657.

[18] J. Deng, N. Cummins, J. Han, X. Xu, Z. Ren, V. Pandit, Z. Zhang, and B. Schuller, "The University of Passau open emotion recognition system for the multimodal emotion challenge," in *Proc. CCPR*. Chengdu, China: Springer, 2016, pp. 652–666.

[19] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. CVPR*, Las Vegas, NV, 2016, pp. 2921–2929.

[20] W. Yin, H. Schütze, B. Xiang, and B. Zhou, "ABCNN: Attention-based convolutional neural network for modeling sentence pairs," *Transactions of the Association for Computational Linguistics*, vol. 6, pp. 259–272, Jun. 2016.

[21] K. Chen, J. Wang, L.-C. Chen, H. Gao, W. Xu, and R. Nevatia, "ABC-CNN: An attention based convolutional neural network for visual question answering," *arXiv preprint arXiv:1511.05960*, 2015.

[22] Q. Kong, Y. Xu, W. Wang, and M. D. Plumbley, "Audio Set classification with attention model: A probabilistic perspective," in *Proc. ICASSP*, Calgary, Canada, 2017, no pagination.

[23] J. Foulds and E. Frank, "A review of multi-instance learning assumptions," *The Knowledge Engineering Review*, vol. 25, no. 1, pp. 1–25, Mar. 2010.

[24] Z. Ren, K. Qian, Z. Zhang, V. Pandit, A. Baird, and B. Schuller, "Deep scalogram representations for acoustic scene classification," *IEEE/CAA Journal of Automatica Sinica*, vol. 5, no. 3, pp. 662–669, May 2018.

[25] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. NIPS*, Stateline, NV, 2012, pp. 1097–1105.

[26] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. ICLR*, San Diego, CA, 2015, no pagination.

[27] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proc. CVPR 2009*, Miami, FL, 2009, pp. 248–255.

[28] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. ICML*, Lille, France, 2015, pp. 448–456.

[29] M. Simon, E. Rodner, and J. Denzler, "Imagenet pre-trained models with batch normalization," *arXiv preprint arXiv:1612.01452*, 2016.

[30] O. Maron and T. Lozano-Pérez, "A framework for multiple-instance learning," in *Proc. NIPS*, Denver, CO, 1998, pp. 570–576.

[31] J. Amores, "Multiple instance classification: Review, taxonomy and comparative study," *Artificial Intelligence*, vol. 201, pp. 81–105, Aug. 2013.

[32] X. Xu, "Statistical learning in multiple instance problems," Ph.D. dissertation, The University of Waikato, Jun. 2003.

Tampereen teknillinen yliopisto
PL 527
33101 Tampere

Tampere University of Technology
P.O.B. 527
FI-33101 Tampere, Finland

ISBN 978-952-15-4262-6