

# SAM-GCNN: A Gated Convolutional Neural Network with Segment-Level Attention Mechanism for Home Activity Monitoring

Yu-Han Shen, Ke-Xin He, Wei-Qiang Zhang

Department of Electronic Engineering, Tsinghua University, Beijing 100084, China  
 yhshen@hotmail.com, hekexinch@163.com, wqzhang@tsinghua.edu.cn

**Abstract**—In this paper, we propose a method for home activity monitoring. We demonstrate our model on dataset of Detection and Classification of Acoustic Scenes and Events (DCASE) 2018 Challenge Task 5. This task aims to classify multi-channel audios into one of the provided pre-defined classes. All of these classes are daily activities performed in a home environment. To tackle this task, we propose a gated convolutional neural network with segment-level attention mechanism (SAM-GCNN). The proposed framework is a convolutional model with two auxiliary modules: a gated convolutional neural network and a segment-level attention mechanism. Furthermore, we adopted model ensemble to enhance the capability of generalization of our model. We evaluated our work on the development dataset of DCASE 2018 Task 5 and achieved competitive performance, with a macro-averaged F-1 score increasing from 83.76% to 89.33%, compared with the convolutional baseline system.

**Index Terms**—acoustic activity classification, gated convolutional neural network, attention mechanism, model ensemble, DCASE

## I. INTRODUCTION

Recently, sound event detection and classification has become more and more popular in the field of acoustic signal processing, and it can be widely used in security surveillance, wildlife protection and smart home. One important application of sound event classification in smart home is home activity monitoring.

Detection and Classification of Acoustic Scenes and Events (DCASE) Challenge is one of the most important international challenges concerning acoustic event detection and classification and has been organized for several years. DCASE 2018 challenge consists of five tasks and we focus on task 5 [1]. This task evaluates systems for monitoring of domestic activities based on multi-channel acoustics.

We can also refer to this task as acoustic activity classification. The main procedure of acoustic activity classification consists of four parts: pre-processing, extracting acoustic features, designing acoustic models as classifiers, and post-processing.

In the part of pre-processing, different methods of data augmentation have been utilized in [2] [3]. Data imbalance is

a big challenge in acoustic event classification and detection because different events may occur at a completely imbalanced frequency. In DCASE 2018 Challenge Task 5, Inoue et al. used shuffling and mixing to produce more training samples [2], and Tanabe et al. utilized dereverberation, blind source separation and data augmentation to improve the quality of audio clips [3].

Mel Frequency Cepstrum Coefficient (MFCC) is a common traditional acoustic feature and has been widely used. But log Mel-scale Filter Bank energies (fbank) are becoming more popular recently, and many works have been done based on fbank [1] [4] [5].

In recent years, Convolutional Neural Networks (CNNs) have achieved great success in many fields such as character recognition, image classification, speaker recognition. And many works based on CNNs have been done in acoustic event classification and detection [6] [7]. Besides, some researchers combined CNNs with Recurrent Neural Networks (RNNs) to capture temporal contexts of audio signals for further improvements [4] [5].

Attention model has been widely used in image classification, object detection and natural language understanding. In the field of acoustic signal processing, Xu et al. [8] proposed an attention model for weakly supervised audio tagging and Kong et al. [9] improved this work by giving a probabilistic perspective. Their work is based on the assumption that those irrelevant sound frames such as background noise and silences should be ignored and given less attention. Both of their models are achieved by a weighted sum over frames where the attention values are automatically learned by neural network.

In our work, acoustic activities might last for a longer period and a single frame is not enough to identify whether it should be ignored. In an audio recording, acoustic activities may keep happening in a majority of frames while acoustic event only occurs in a few frames. So we propose a segment-level attention mechanism (SAM) to decide how much attention should be given based on the characteristics of segments. Here, a segment is comprised of several frames.

In this paper, we mainly adopt three ways to improve the performance of our model:

(1) We replace currently popular CNN with gated convolutional neural network to extract more temporal features of

This work was supported by the National Natural Science Foundation of China under Grant No. U183620001. The corresponding author is Wei-Qiang Zhang.

TABLE I  
AMOUNTS OF AUDIO CLIPS AND SESSIONS

Activity	#10s clips	#sessions
Absence	18860	42
Cooking	5124	13
Dishwashing	1424	10
Eating	2308	13
Other	2060	118
Social activity	4944	21
Vacuum cleaning	972	9
Watching TV	18648	9
Working	18644	33
Total	72984	268

audios;

(2) We propose a new segment-level attention mechanism to focus more on the audio segments with more energy;

(3) We utilize model ensemble to enhance the classification capability of our model.

The rest of this paper is organized as follows. In Section 2, we introduce our methods in detail, mainly including acoustic feature, gated convolutional neural network, segment-level attention mechanism and model ensemble. The experiment setup, evaluation metric and our results are illustrated in Section 3. Finally, the conclusion of our work is presented in Section 4.

## II. METHODS

### A. Task Description

The DCASE 2018 Task 5 dataset [10] contains sound data recorded in a living room by individual devices with four microphone arrays at seven undisclosed locations. The dataset is divided into a development dataset and an evaluation dataset. Four cross-validation folds are provided for the development dataset in order to make results reported with this dataset uniform. For each fold, a training, testing and evaluation subset is provided. In this paper, our work is based on the development dataset and we use the provided cross-validation folds for training and evaluation.

The audio clips in this dataset can be classified into nine classes: absence, cooking, dishwashing, eating, other, social activity, vacuum cleaning, watching TV and working. All audio clips are derived from continuous recording sessions collected by seven microphone arrays and each clip contains four channels. The duration of each audio clip is 10 seconds. Specific information about the dataset is shown in Table 1 and more details can be found in [10].

### B. System Overview

Our proposed system is illustrated as Figure 1. The input of our system is log Mel-scaled filter banks (fbank). Then it will be fed into two structures: one is a Gated Convolutional Neural Network (GCNN) architecture, and the other is our proposed Segment-Level Attention Mechanism (SAM).

Unlike most systems that output one probability score for an audio as a whole, we divide a 10-s audio clip into several

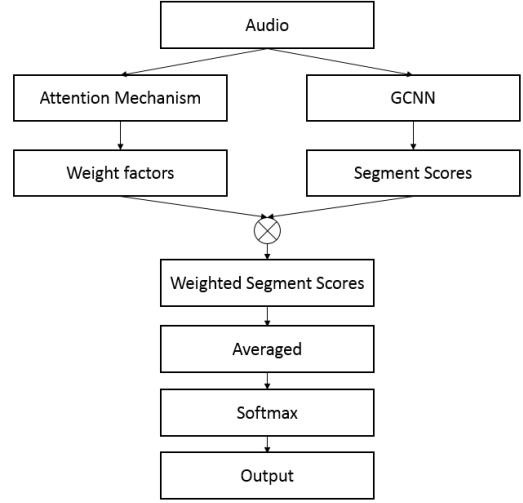


Fig. 1. Overall architecture of proposed system

segments. The output of our GCNN architecture is  $\mathbf{X} \in \mathbb{R}^{N \times C}$  and represents the probability for each class of each segment, where  $N \in \mathbb{N}$  is the number of segments in an audio clip and  $C \in \mathbb{N}$  is the number of predefined classes. The output of SAM is a vector  $\mathbf{W} \in \mathbb{R}^N$  and represents the attention weight factor for each segment. Then we multiply  $\mathbf{X}$  with  $\mathbf{W}$  for each segment to obtain weighted segment scores. Those scores will be averaged among segments to get a vector  $\mathbf{Y} \in \mathbb{R}^C$  and then go through a softmax to represent the normalized probability for each class. The class with the largest probability is considered to be the classification result. The detailed explanations of our proposed system will be included in the following parts of this section.

### C. Acoustic Feature

We use fbank as the input of our system. Fbank is a two-dimensional time-frequency acoustic feature. It imitates the characteristics of human's ears and concentrates more on the low frequency components of audio signals. Compared with traditional MFCC feature, more original information can be kept in fbank and it has been widely used in deep learning. To extract fbank feature, each input audio is divided into 40ms frames with 50% overlapping, and then 40 mel-scale filters are applied on the magnitude spectrum of each frame. Finally, we take logarithm on the amplitude and get fbank feature. As is mentioned in Section 1, the audio clips contain four channels, so our fbank feature contains four channels as well. In our work, four channels are fed into the system separately while training. And the averaged output score of four channels is used for evaluation.

### D. Gated Convolutional Neural Network

Gated convolutional neural network was proposed by Dauphin et al. in [11] and has shown great power in machine translation, natural language processing. Our GCNN architecture consists of three main parts: 1) convolutional

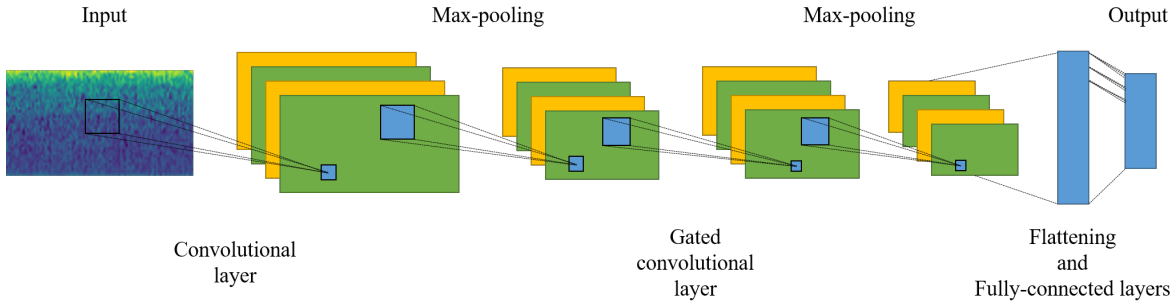


Fig. 2. Overall architecture of gated convolutional neural network

neural network (CNN), 2) gated convolutional neural network (GCNN), 3) feedforward neural network (FNN). And our overall architecture is shown in Figure 2.

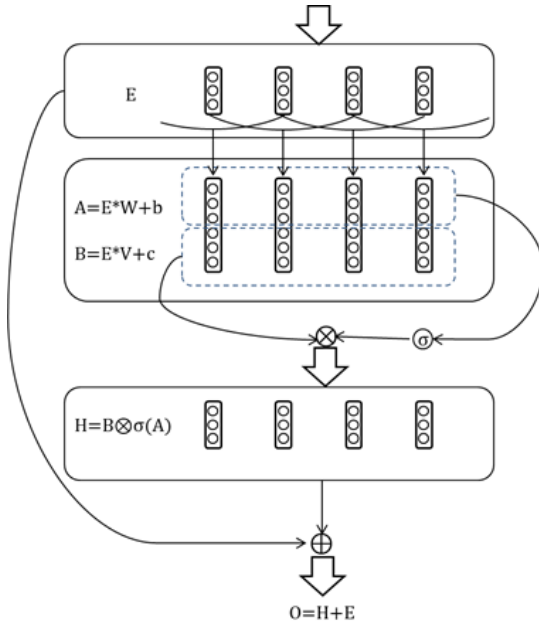


Fig. 3. Gated convolutional neural network.

Before being fed into GCNN architecture, the extracted fbank feature is normalized to zero mean and unit standard deviation (we call it global normalization, to distinct with the following time normalization).

Convolutional layers extract frequency features and connect features of adjacent frames. And the output of convolutional layer is followed by batch normalization [12], a ReLU activation unit and a dropout layer [13]. Then a max-pooling layer is applied to keep the most important features.

The structure of gated convolutional neural network is illustrated in Figure 3.

In gated convolutional neural network, the output of convolutional layer is divided into two parts with the same size. The input of this structure is  $E = [e_1, e_2, \dots, e_n]$ ,  $E$  passes through a convolutional layer and the output is divided into  $A$  and  $B$ . Then  $A$  passes through sigmoid activation function and

multiplies with  $B$  by element-wise. In order to enable stronger work, we add residual connections from the input  $E$  to the output of this structure  $H$ . Residual network is introduced to avoid vanishing gradient problem [14].

The specific formula is as follows:

$$A = E * W + b, \quad (1)$$

$$B = E * V + c, \quad (2)$$

$$H = B \otimes \sigma(A), \quad (3)$$

$$O = H + E, \quad (4)$$

where  $W, V$  represent convolutional kernel values, and  $b, c$  mean biases.  $\otimes$  represents element-wise production.  $\sigma(\cdot)$  is a sigmoid activation function.

The gated convolutional layer is also followed by batch normalization, a ReLU activation unit, a dropout layer and a max-pooling layer.

After the gated convolutional neural network, the features on multiple channels are flattened into frequency axis.

Then two fully-connected layers are used to combine extracted features and output nine scores for each segment. Our work differs from others in that we output scores for each segment while most researchers output scores for an audio as a whole. We intend to focus on those segments with more energy and ignore segments with less energy, which we call “silence” segments. That is why we propose a segment-level attention mechanism.

#### E. Segment-Level Attention Mechanism

As mentioned in Section 1, attention mechanism was introduced to ignore irrelevant sounds such as background noise and silences in audio event classification. In DCASE 2018 task 5, an audio clip labeled as cooking may contain some segments of silences and we should not pay too much attention to those segments because audio clips labeled as other classes may also contain silences. Motivated by Xu et al. [8], we propose a segment-level attention mechanism. Our work differs from previous work in that we give our attention weight factors based on the characteristics of segments instead of frames.

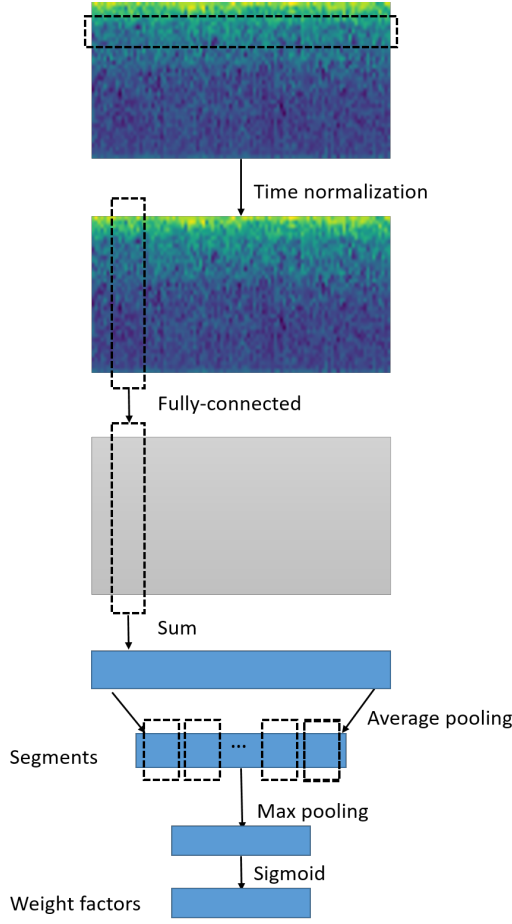


Fig. 4. Segment-Level Attention Mechanism.

The structure of segment-level attention mechanism is shown in Figure 4. The input of this structure is aforementioned fbank feature. Then it will be normalized along the time axis, which we call time normalization. The purpose of time normalization is to further differentiate the features among frames.

A fully-connected layer is added to extract deeper features of frames. Like in the gated convolutional neural network, the fully-connected layer is followed by batch normalization, ReLU and dropout. Next, we calculate the sum along frequency axis. An average pooling layer is added to filter adjacent frames. Then a max-pooling layer is used to maintain the most important information of a segment. Finally, we use a sigmoid activation to limit the weight factors between 0 and 1. Based on our experiments, the duration of a segment is set to 1 second. Specific structure and hyperparameters will be illustrated in Section 3.

#### F. Model Ensemble

Model ensemble is a common strategy in machine learning. In our work, we propose a strategy of model ensemble.

During our experiments, we notice that absence, other and working are three sorts of activities that are often misclassified

with the others. So we train a model in particular to classify those three classes of activities. When our main system classifies an audio clip as any of the three classes, we will use the specially trained model for one more classification.

If an audio is classified as a class other than class 0, 4, 8 (absence, other and working) by our first system, the output will be the final output. Otherwise, the audio will be fed into our second system. We denote the output of our first system as  $\mathbf{X}^I \in \mathbb{R}^9$  and second system as  $\mathbf{X}^{II} \in \mathbb{R}^3$ .  $\mathbf{X}_i^N$  represents the output probability of  $i$ -th class by the  $N$ -th system, where  $i \in [0, 8]$  and  $N$  is 1 or 2. Then the final output  $\mathbf{Y} \in \mathbb{R}^9$  of our ensemble system will be calculated according to the following algorithm. We calculate the sum of  $\mathbf{X}_0^I, \mathbf{X}_4^I, \mathbf{X}_8^I$  and redistribute them based on our second system output  $\mathbf{X}^{II}$ .

---

#### Algorithm 1 Model Ensemble( $\mathbf{X}^I, \mathbf{X}^{II}$ )

---

$O \leftarrow \text{argmax} \mathbf{X}^I$

$\mathbf{Y} \leftarrow \mathbf{X}^I$

**if**  $O == 0$  **or**  $O == 4$  **or**  $O == 8$  **then**

$S \leftarrow \text{sum}(\mathbf{X}_0^I, \mathbf{X}_4^I, \mathbf{X}_8^I)$

$\mathbf{Y}_{0,4,8} \leftarrow S \mathbf{X}^{II}$

**end if**

---

### III. EXPERIMENT, EVALUATION AND RESULTS

#### A. Experiment setup

Our model is trained using Adam [15] for gradient based optimization. Cross-entropy is used as the loss function. And the structure of our system is shown in Table 2 and Table 3 along with parameters. The initial learning rate is 0.001 and the batch size is  $256 \times 4$  channels because each channel is considered as a different sample for training. We train the classifiers for 300 epochs.

We select 5% of the testing data as validation dataset and choose models which result in the best accuracy on the validation dataset for final evaluation. In the evaluation process, the outputs of 4-channel acoustics are averaged to get the final posterior probability.

TABLE II  
MODEL STRUCTURE AND PARAMETERS OF GATED CONVOLUTIONAL NEURAL NETWORK

Input $40 \times 501 \times 1$	Output size
Conv (padding: valid, kernel: [40, 5, 64])	1, 497, 64
BN-ReLU-Dropout(0.2)	1, 497, 64
$1 \times 5$ Max-Pooling(padding: valid)	1, 99, 64
Gated Conv (padding: same, kernel: [1, 3, 128])	1, 99, 64
BN-ReLU-Dropout(0.2)	1, 99, 64
$1 \times 10$ Max-Pooling(padding: same)	1, 10, 64
Feature Flattening	10, 64
Fully-connected(unit num: 64) -ReLU-Dropout(0.2)	10, 64
Fully-connected(unit num: 9)	10, 9

TABLE III  
MODEL STRUCTURE AND PARAMETERS OF SEGMENT-LEVEL ATTENTION  
MECHANISM

Input $40 \times 501 \times 1$	Output size
Fully-connected(unit num: 40)	40, 501, 1
BN-ReLU-Dropout(0.2)	40, 501, 1
Sum along frequency axis	1, 501, 1
$1 \times 5$ Average-Pooling(padding: same)	1, 100, 1
$1 \times 10$ Max-Pooling(padding: same)	1, 10, 1
Squeeze	10
Sigmoid	10

### B. Evaluation Metric

The official evaluation metric for DCASE 2018 challenge task 5 is macro-averaged F1-score. F1-score is a measure of a test’s accuracy and it is the harmonic average of precision and recall. Macro-averaged means that F1-score is calculated for each class separately and averaged over all classes. For this task, a full 10s multi-channel audio is considered to be one sample.

### C. Results

We examine the following configurations:

- (1) CNN: Convolutional neural network as baseline system;
- (2) SAM-CNN: Convolutional neural network with our proposed segment-level attention mechanism;
- (3) GCNN: Gated convolutional neural network;
- (4) SAM-GCNN: Gated convolutional neural network with our proposed segment-level attention mechanism;
- (5) Ensemble: Gated convolutional neural network with our proposed segment-level attention mechanism and model ensemble.

TABLE IV  
MACRO-AVERAGED F1-SCORE OF MULTIPLE SYSTEMS ON 4 FOLDS

System	Fold1	Fold2	Fold3	Fold4	Average
CNN	81.92%	82.58%	83.26%	87.29%	83.76%
GCNN	85.58%	84.22%	86.36%	88.83%	86.25%
SAM-CNN	83.68%	82.26%	84.56%	88.09%	84.65%
SAM-GCNN	88.49%	86.81%	86.51%	90.52%	88.08%
Ensemble	89.62%	88.11%	87.95%	91.63%	89.33%

As shown in Table 4, the macro-averaged F-1 score of GCNN is 2.49% higher than CNN. And our proposed segment-level attention mechanism can improve the classification performance of both CNN and GCNN.

Moreover, our proposed ensemble strategy can outperform previous systems and achieve 89.33% F1-score. Confusion matrix before and after ensemble is shown in Figure 5. On the left is the confusion matrix of SAM-GCNN, and on the right is the confusion matrix of SAM-GCNN with model ensemble. The element in the  $i$ -th row and  $j$ -th column of this matrix

represents the amount of audio clips that belong to class  $i$  and are classified as class  $j$ , so the elements on the diagonal represent the number of correctly classified audio clips. We can find that the number of correctly classified audio clips has increased after ensemble, especially for “absence”, “other” and “working”, showing that our model ensemble method does work.

The class-wise performance of our final model is shown in Table 5.

TABLE V  
CLASS-WISE PERFORMANCE OF PROPOSED MODEL

	fold1	fold2	fold3	fold4	Average
Absence	94.43%	92.99%	93.15%	94.93%	93.88%
Cooking	95.92%	94.26%	93.75%	96.49%	95.10%
Dishwashing	87.45%	81.22%	81.81%	83.87%	83.59%
Eating	89.35%	89.66%	87.73%	90.56%	89.33%
Other	52.28%	53.51%	54.61%	67.15%	56.89%
Social activity	97.83%	95.85%	94.38%	98.50%	96.64%
Vacuum cleaning	99.99%	99.81%	100.00%	100.00%	99.95%
Watching TV	99.55%	99.86%	99.42%	99.91%	99.69%
Working	89.82%	85.85%	86.68%	93.22%	88.89%
Macro-Average	89.62%	88.11%	87.95%	91.63%	89.33%

To better evaluate our work, we compare the performance of proposed model with the top-2 ranked teams in DCASE 2018 Challenge Task 5 and the official baseline system in Table 6. Both of the top-2 teams adopted complex methods of pre-processing, data augmentation and model ensemble. We can achieve equivalent performance without any data augmentation. And our system outperforms the official baseline significantly.

TABLE VI  
COMPARISON WITH STATE-OF-THE-ART WORKS

	Averaged F1-score
Proposed	89.3%
InouetMilk [2]	90.0%
HITfweight [3]	89.8%
Official Baseline	84.5%

## IV. CONCLUSION

In this paper, we have introduced our work and the results show that the performance of our proposed system is significantly superior to that of the baseline. Our proposed segment-level attention mechanism improves the performance of both CNN and GCNN architecture. Furthermore, by using model ensemble, we have achieved competitive performance on the development dataset of DCASE 2018 task 5. Note that both the top two teams of this task utilized complex methods of data augmentation and model ensemble. Our system can achieve equivalent performance without data augmentation, which shows that our proposed attention mechanism can

	Absence	Cooking	Dishwashing	Eating	Other	Social activity	Vacuum cleaning	Watching TV	Working
Absence	4457	0	0	1	21	2	0	1	158
Cooking	0	1167	41	8	11	0	0	1	16
Dishwashing	0	5	325	5	26	3	0	0	0
Eating	4	4	12	522	15	3	0	0	20
Other	94	0	23	2	311	2	0	1	83
Social activity	0	0	2	2	8	1185	0	5	6
Vacuum cleaning	0	0	0	0	0	0	240	0	0
Watching TV	1	0	0	0	0	0	0	4539	0
Working	323	0	4	32	130	4	0	0	4115

	Absence	Cooking	Dishwashing	Eating	Other	Social activity	Vacuum cleaning	Watching TV	Working
Absence	4597	0	0	1	4	2	0	1	35
Cooking	2	1168	42	8	13	0	0	1	10
Dishwashing	0	5	325	5	26	3	0	0	0
Eating	4	4	12	523	12	3	0	0	22
Other	87	0	26	2	323	2	0	1	75
Social activity	2	0	2	2	6	1186	0	5	5
Vacuum cleaning	0	0	0	0	0	0	240	0	0
Watching TV	0	0	0	0	0	0	0	4540	0
Working	353	0	4	34	62	4	0	0	4151

Fig. 5. Confusion matrix before and after ensemble on fold4.

contribute a lot to home activity monitoring. Since the ground truth labels of evaluation dataset of DCASE 2018 challenge have not been published yet, future work needs to be done for further evaluation.

## REFERENCES

- [1] Gert Dekkers, Lode Vliegen, Toon van Waterschoot, Bart Vanrumste, and Peter Karsmakers, "DCASE 2018 Challenge - Task 5: Monitoring of domestic activities based on multi-channel acoustics," Technical Report, KU Leuven, 2018. URL: <https://arxiv.org/abs/1807.11246>, arXiv:1807.11246.
- [2] Tadanobu Inoue, Phongtharin Vinayavekhin, Shiqiang Wang, David Wood, Nancy Greco and Ryuki Tachibana, "Domestic Activities Classification Based on CNN Using Shuffling and Mixing Data Augmentation," DCASE 2018 Challenge, Technical Report, 2018.
- [3] Ryo Tanabe, Takashi Endo, Yuki Nikaido, Takeshi Ichige, Phong Nguyen, Yohei Kawaguchi and Koichi Hamada, "Multichannel acoustic scene classification by blind dereverberation, blind source separation, data augmentation, and model ensembling," DCASE 2018 Challenge, Technical Report, 2018.
- [4] E.Cakir and T.Virtanen, "Convolutional recurrent neural networks for rare sound event detection," in Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop, 2017, pp. 803806.
- [5] H. Lim, J. Park, K. Lee, and Y. Han, "Rare sound event detection using 1d convolutional recurrent neural networks," in Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop, 2017, pp. 8084.
- [6] Y. Han, J. Park, and K. Lee, "Convolutional neural networks with binaural representations and background subtraction for acoustic scene classification," in Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop, 2017, pp. 4650.
- [7] W. Zheng, J. Yi, X. Xing, X. Liu, and S. Peng, "Acoustic scene classification using deep convolutional neural network and multiple spectrograms fusion," in Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop, 2017, pp. 133137.
- [8] Y. Xu, Q. Kong, Q. Huang, W. Wang, and Mark D. Plumbley, "Attention and localization based on a deep convolutional recurrent model for weakly supervised audio tagging," in Proceedings of INTERSPEECH.IEEE, 2017, pp.30833087.
- [9] Q. Kong, Y. Xu, W. Wang, and M. D. Plumbley, "Audio set classification with attention model: A probabilistic perspective," arXiv preprint arXiv:1711.00927, 2017.
- [10] G. Dekkers, S. Lauwereins, B. Thoen, M. W. Adhana, H. Brouckxon, T. van Waterschoot, B. Vanrumste, M. Verhelst, and P. Karsmakers, "The SINS database for detection of daily activities in a home environment using an acoustic sensor network," in Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017), Munich, Germany, November 2017, pp. 3236.
- [11] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modeling with gated convolutional networks," arXiv preprint arXiv:1612.08083, 2016.
- [12] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in Proceedings of The 32nd International Conference on Machine Learning, 2015, pp. 448456.
- [13] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," Journal of Machine Learning Research, vol. 15, no. 1, pp. 19291958, 2014.
- [14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," IEEE Conference on Computer Vision and Pattern Recognition IEEE Computer Society, 2016:770-778.
- [15] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980, 2014.