SOUND EVENT DETECTION IN DOMESTIC ENVIRONMENTS USING ENSEMBLE OF CONVOLUTIONAL RECURRENT NEURAL NETWORKS

Technical Report

Wootaek Lim, Sangwon Suh, Sooyoung Park and Youngho Jeong

Realistic AV Research Group
Electronics and Telecommunications Research Institute
218 Gajeong-ro, Yuseong-gu, Daejeon, Korea
wtlim@etri.re.kr

ABSTRACT

In this paper, we present a method to detect sound events in domestic environments using small weakly labeled data, large unlabeled data, and strongly labeled synthetic data as proposed in the Detection and Classification of Acoustic Scenes and Events (DCASE) 2019 challenge task 4. To solve the problem, we use convolutional recurrent neural network (CRNN), as it stacks convolutional neural networks (CNN) and bi-directional gated recurrent unit (Bi-GRU). Moreover, we propose various methods such as data augmentation, event activity detection, multi-median filtering, mean-teacher student model, and the ensemble of neural networks to improve performance. By combining the proposed method, sound event detection performance can be enhanced, compared with the baseline algorithm. As a result, performance evaluation shows that the proposed method provides detection results of 40.89% for event-based metrics and 66.17% for segment-based metrics.

Index Terms— DCASE 2019, Sound event detection, CRNN, SpecAugment, Model ensemble

1. INTRODUCTION

Sound event detection (SED) is the field of predicting acoustic events in audio signals. In recent years, this field has witnessed growth owing to the release of large-size datasets, improvements in algorithms, and improved hardware performance [1, 2]. The DCASE challenge has been held for several years with the objective of solving the SED problem [3-6]. This year, the DCASE challenge comprised five tasks, and this study proposed a method to solve the DCASE 2019 challenge task 4. This is the follow-up to DCASE 2018 task 4. The goal of this task is to train the model to detect sound events using the dataset using various types of labels and to find the onset and offset of sound events. According to the last year's submissions, various methods have been proposed to solve this problem [7-13] and the mean-teacher model has shown the best performance [13, 14]. Therefore, the baseline system of task 4 in DCASE 2019 challenge is based on the idea of the best submission of DCASE 2018 task 4. The method used in the baseline system is similar to that used in [13], but the proposed network architecture has been simplified.

In this study, a SED system based on CRNN structure is proposed. To improve performance, we perform data augmentation to overcome the small-size dataset problem, the event activity detection method to learn the weakly labeled dataset, the multi-median filtering method using a synthetic dataset, and the mean-teacher model to utilize the unlabeled dataset.

2. DATASET

The dataset for the DCASE 2019 challenge task 4 comprised a 10-s audio clip recorded in an indoor environment or synthesized assuming a similar environment. This task also classifies 10 sound event classes [6]. The details of the dataset are described in Table 1. First, three types of datasets are provided for training. This dataset comprises the weakly labeled training set, unlabeled in domain training set without any labels, and strongly labeled synthetic set. The weakly labeled training set and the unlabeled in domain training set are based on AudioSet [15], and the strongly labeled synthetic sets are synthesized based on the dataset proposed in [16] and [17]. A validation set is provided for verification of SED performance. This dataset is a combination of the DCASE 2018 Task 4 test set and the evaluation set. The evaluation dataset is composed of 13190 audio clips, and the details will be released later.

Table 1: Details of DCASE 2019 challenge task 4 dataset.

Dataset			Descriptions
Development dataset	Training set	Labeled training set	- 1578 clips (2244 class occurrences) - w/ weak labels
		Unlabeled in domain training set	- 14412 clips - w/o labels
		Synthetic strongly labeled set	- 2045 clips (6032 events) - w/ strong labels
	Validation set		- 1168 clips (4093 events) - w/ strong labels
Evaluation dataset			- 13190 clips (TBA) - w/ strong labels

3. PROPOSED METHOD

3.1. Network structure

The proposed method uses a CRNN as a basic network structure inspired by the DCASE 2019 challenge task 4 baseline system. The proposed network has a more complex structure than the baseline system. First, the CNN layer is composed of the 3×3 kernel on all layers, and the number of feature maps increases from the low- to high-level layers. It also has a gated linear unit (GLU), which was originally proposed in [18], and batch normalization. The dropout layer and average pooling layer are stacked after each CNN module. The two Bi-GRU layers are stacked after six CNN layers. At the end of the network, strong and weak predictions are estimated, and the attention module is used to help with learning. The detailed network structure is depicted in Figure 1.

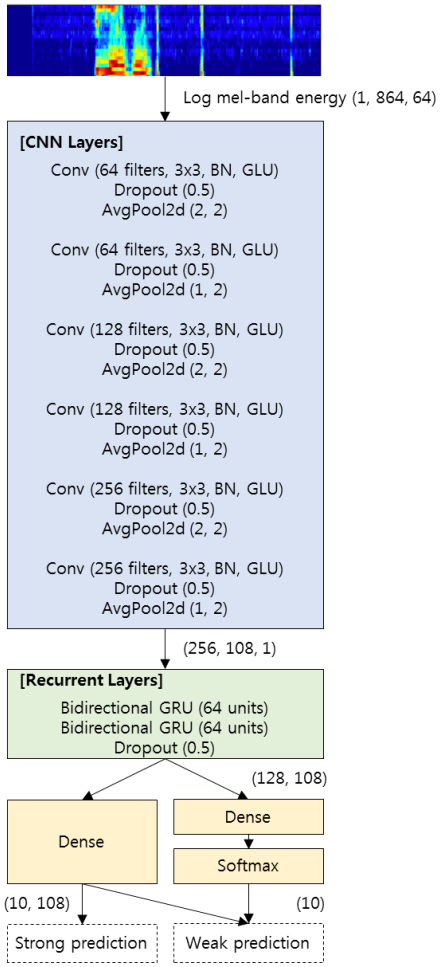

Figure 1: Structure of CRNN used in our proposed method.

3.2. SpecAugment

Without enough training data, data augmentation can increase the effective size of existing data, which can greatly improve the deep neural network performance in many tasks. In audio processing, the conventional data augmentation method transforms waveforms used in learning in the same manner as that for adding time stretching, block mixing, pitch shifting, or background noise. This helps the neural network become more robust by forcing multiple augmented versions of the same audio input into the neural network and learning the variance during the training process.

SpecAugment applies augmentation directly to the audio spectrogram [19]. Therefore, this method is simple and the computation cost is less. In this study, the SpecAugment was applied directly to the input spectrogram during the training. The dataset used in this process is weakly labeled dataset, and the robustness is increased by randomly selecting samples to be augmented or not to be augmented.

3.3. Event activity detection

A simple way to realize strong labels from weakly labeled data is to assign a strong label to all time frames. However, assigning a strong label to weakly labeled data is difficult because there is no information about the existence of the event. Therefore, a pseudo-labeling using event activity detection (EAD) was used to learn more accurate labels. A strong label is assigned when the average value of frame energy over the threshold of 0.7. It assumes that there are no events in the frame if the energy is small [12].

3.4. Multi-median filtering

The output is postprocessed by median filtering. Applying median filtering of the same length to various sound events class is inadequate because each sound has statistically different characteristics. Therefore, we selected the length of multi-median filter (MMF) according to the synthetic strongly labeled set. The MMF size of each event class was calculated from the metadata of the synthetic data, and the median value was used after calculating the length of each event.

3.5. Mean-teacher model

Semi-supervised learning to utilize the unlabeled in domain training set was done using the mean-teacher student model [13, 14]. The mean-teacher model was learned with the same two CRNN structures described in Section 3.1. In the training stage, after the student model is updated, the teacher model is updated using the exponential moving average of the student model weights.

3.6. Model ensemble

A reliable approach to improve the performance of neural networks is to have an ensemble of several trained models. The ensemble technique combines the weak learner to create a strong learner. Therefore, the ensemble approach not only improves model diversity but also performance. There are several approaches to forming an ensemble [20], but two methods have been tested. The first method is an ensemble of different checkpoints in a single model. This method has generally shown limited success, but it is very efficient because it comes from a single training model. The second method is to create an ensemble by learning the same model with different initializations. This method is time-consuming, but simple and powerful. The mean probability is used to make a final decision of the ensemble model.

4. PERFORMANCE EVALUATION

For evaluating the performance of the proposed methods, the dataset described in chapter 2 is used. The weakly labeled training set and the synthetic strongly labeled set were used for basic CRNN model training, and the unlabeled in domain training set was additionally used for mean-teacher student model training. The audio input was a mono channel with a 44.1 kHz sampling rate. To make an input 2D spectrogram, a 10 second audio clip was converted to 64-band log-mel energies with a window size of 2048 and hop length of 511. As a result, an image with 864 frames and 64 frequency bands was used as a network input. The Adam optimizer was used for network learning, and the learning rate was 0.001. The binary cross-entropy function is used as the criterion for comparing the loss between the target and the output. The early stopping method was not used because the ensemble model could reduce the variance.

The experimental results for the basic CRNN network are listed in Table 2, and the experimental results based on the mean-teacher student model are listed in Table 4. Four experiments were performed for each model for reliable results and model ensemble. The training was performed for a total of 500 epochs, and the model was tested every 100 epochs from the 200th epoch onward.

As listed in Tables 2 and 4, the mean-teacher model shows slightly better performance on average than the basic CRNN model, although there is a deviation from each training step. Both models outperform the baseline system performance of 23.7%. As previously described, the performance of the ensemble of different checkpoints in a single model and the ensemble of different initializations were evaluated. In Table 2, the horizontal row denotes the result of the ensemble of different checkpoints and the vertical column is the result of the ensemble of different initializations. The ensemble for each row and column was the result of four models combined. The ensemble of different initializations demonstrated better results, and the ensemble of 500 epoch models demonstrated an F-score of 38.77%. Finally, the method with an ensemble of 16 models demonstrated the best performance of 39.51%. The detailed results are listed in Table 3. The results of the mean-teacher model are listed in Table 4. In the mean-teacher model, the ensemble of different checkpoints are unnecessary, but it shows improved performances. Similar to the basic CRNN model, the ensemble of different initializations shows a better performance in the mean-teacher model. This model demonstrated an F-score of 39.43% when using the ensemble of four models for 500 epoch. Finally, when combining the ensemble of 16 models, it showed the best performance of 40.89%. The detailed results are listed in Table 5.

Table 2: Sound event detection performance using the basic CRNN model and ensemble.

Model \ Epoch	ep200	ep300	ep400	ep500	Ensemble
CRNN (# 1)	33.07	34.40	28.94	34.23	36.08
CRNN (# 2)	33.32	36.10	35.46	33.68	37.43
CRNN (# 3)	34.91	33.57	32.72	33.75	36.86
CRNN (# 4)	34.52	34.07	32.75	35.55	36.23
Ensemble	38.87	39.36	38.68	**38.77** (submission-1)	**39.51** (submission-2)

Table 3: Class-wise result of the basic CRNN model ensemble (submission-2).

Event label	Event-based metrics		Segment-based metrics	
	F-score (%)	Error rate	F-score (%)	Error rate
Alarm/bell/ringing	47.4	0.95	78.6	0.42
Blender	30.3	1.34	57.4	0.79
Cat	40.2	1.23	59.6	0.81
Dishes	19.8	1.30	53.6	0.88
Dog	21.0	1.29	66.4	0.66
Electric shaver/toothbrush	42.2	1.31	67.9	0.79
Frying	39.6	1.36	62.2	0.84
Running water	40.4	1.05	69.0	0.56
Speech	51.0	0.86	85.7	0.28
Vacuum cleaner	63.3	0.78	72.5	0.61
macro-average	**39.51**	**1.15**	**67.29**	**0.66**
micro-average	**40.87**	**1.03**	**72.52**	**0.45**

Table 4: Sound event detection performance using the mean-teacher model and ensemble.

Model \ Epoch	ep200	ep300	ep400	ep500	Ensemble
Mean-Teacher (# 1)	34.17	34.81	34.86	34.74	36.57
Mean-Teacher (# 2)	33.47	35.59	33.83	34.00	36.29
Mean-Teacher (# 3)	36.83	36.07	36.38	33.51	37.53
Mean-Teacher (# 4)	33.56	36.06	35.57	36.87	38.32
Ensemble	38.92	38.55	39.09	**39.43** (submission-3)	**40.89** (submission-4)

Table 5: Class-wise result of the mean-teacher model ensemble (submission-4).

Event label	Event-based metrics		Segment-based metrics	
	F-score (%)	Error rate	F-score (%)	Error rate
Alarm/bell/ringing	47.2	0.92	79.4	0.38
Blender	33.5	1.30	61.0	0.76
Cat	43.1	1.05	59.4	0.70
Dishes	22.7	1.17	46.5	0.87
Dog	27.7	1.21	66.3	0.62
Electric shaver/toothbrush	42.6	1.32	66.1	0.79
Frying	40.6	1.26	61.2	0.83
Running water	32.6	1.11	63.2	0.60
Speech	57.4	0.80	86.0	0.28
Vacuum cleaner	61.4	0.76	72.5	0.52
macro-average	**40.89**	**1.08**	**66.17**	**0.63**
micro-average	**44.97**	**0.96**	**72.12**	**0.46**

5. CONCLUSION

The goal of this study is to propose methods for SED in domestic environments using various type of dataset. In this paper, SED performance is improved by using the proposed network and various methods such as SpecAugment, EAD, MMF, and mean-teacher student model. Moreover, the two ensemble methods and its combination were tested and achieved an F-score of 40.89% for event-based metrics and 66.17% for segment-based metrics.

6. ACKNOWLEDGMENT

This work was supported by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIT). (No.2017-0-00050, Development of Human Enhancement Technology for auditory and muscle support)

7. REFERENCES

[1] A. Mesaros, T. Heittola, E. Benetos, P. Foster, M. Lagrange, T. Virtanen, and M. D. Plumbley. "Detection and classification of acoustic scenes and events: outcome of the DCASE 2016 challenge," IEEE/ACM Transactions on Audio, Speech, and Language Processing, 26(2): 379–393, 2018.

[2] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M. D. Plumbley, "Detection and classification of acoustic scenes and events," IEEE Transactions on Multimedia, 17(10): 1733–1746, 2015.

[3] "DCASE2016," http://www.cs.tut.fi/sgn/arg/dcase2016/.

[4] "DCASE2017," http://www.cs.tut.fi/sgn/arg/dcase2017/.

[5] "DCASE2018," http://dcase.community/challenge2018/.

[6] "DCASE2019," http://dcase.community/challenge2019/.

[7] R. Serizel, N. Turpault, H. Eghbal-Zadeh, and A. P. Shah, "Large-Scale Weakly Labeled Semi-Supervised Sound Event Detection in Domestic Environments," in Proc. Detection and Classification of Acoustic Scenes and Events Workshop (DCASE), 2018.

[8] R. Serizel, and N. Turpault, "Sound Event Detection from Partially Annotated Data: Trends and Challenges," IcETRAN conference, 2019.

[9] Y. Liu, J. Yan, Y. Song and J. Du, "USTC-NELSLIP System for DCASE 2018 Challenge task 4," Technical Report, DCASE 2018 Challenge.

[10] Q. Kong, T. Iqbal, Y. Xu, W. Wang and M. D. Plumbley, "DCASE 2018 Challenge baseline with convolutional neural networks" Technical Report, DCASE 2018 Challenge.

[11] S. Kothinti, K. Imoto, D. Chakrabarty, G. Sell, S. Watanabe and M. Elhilali, "Joint Acoustic and Class Inference for Weakly Supervised Sound Event Detection," in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019.

[12] W. Lim, S. Suh, and Y. Jeong, "Weakly labeled semi supervised sound event detection using CRNN with inception module," in Proc. Detection and Classification of Acoustic Scenes and Events Workshop (DCASE), 74-77, 2018.

[13] L. JiaKai, "Mean teacher convolution system for dcase 2018 task 4," Technical Report, DCASE 2018 Challenge.

[14] A. Tarvainen, and H. Valpola, "Mean teachers are better role models: weight-averaged consistency targets improve semi-supervised deep learning results," in advances in neural information processing systems (NIPS), 1195–1204. 2017.

[15] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: an ontology and human-labeled dataset for audio events," in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 776-780, 2017

[16] E. Fonseca, J. Pons, X. Favory, F. Font, D. Bogdanov, A. Ferraro, S. Oramas, A. Porter, and X. Serra. "Freesound datasets: a platform for the creation of open audio datasets," in Proc. International Society for Music Information Retrieval Conference (ISMIR), 486-493, 2017.

[17] G. Dekkers, S. Lauwereins, B. Thoen, M. W. Adhana, H. Brouckxon, T. V. 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 Proc. Detection and Classification of Acoustic Scenes and Events Workshop (DCASE), 32–36, 2017.

[18] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modelling with gated convolutional networks," arXiv preprint arXiv: 1612.08083, 2016.

[19] D. S. Park, W. Chan, Y. Zhang, C. C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition," arXiv preprint arXiv:1904.08779, 2019.

[20] http://cs231n.github.io/neural-networks-3/#ensemble