WEAKLY LABELED SEMI-SUPERVISED SOUND EVENT DETECTION USING CRNN WITH INCEPTION MODULE

Wootaek Lim, Sangwon Suh, 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 for large-scale detection of sound events using small weakly labeled data proposed in the Detection and Classification of Acoustic Scenes and Events (DCASE) 2018 challenge Task 4. To perform this task, we adopted the convolutional neural network (CNN) and gated recurrent unit (GRU) based bidirectional recurrent neural network (RNN) as our proposed system. In addition, we proposed the Inception module for handling various receptive fields at once in each CNN layer. We also applied the data augmentation method to solve the labeled data shortage problem and applied the event activity detection method for strong label learning. By applying the proposed method to a weakly labeled semi-supervised sound event detection, it was verified that the proposed system provides better performance compared to the DCASE 2018 baseline system.

***Index Terms*—** DCASE 2018, Sound event detection, Weakly labeled semi-supervised learning, Deep learning, Inception module

1. INTRODUCTION

In the field of machine learning, there are various tasks for modeling human auditory cognitive systems. One such field is sound event detection (SED), which is a rapidly growing field owing to the improvement of algorithms, datasets, and expansion of smart devices [1, 2]. In particular, SED technology is being studied to provide services that inform people about the context information of sound events at home or outside. Moreover, SED is important for the auto-tagging of multimedia content [1, 2]. To contribute to the SED task, the DCASE challenge has been organized for three years beginning in 2013 [1, 2, 3]. This year, the DCASE 2018 challenge comprises five tasks: acoustic scene classification, general-purpose audio tagging of Freesound content with AudioSet labels, bird audio detection, large-scale weakly labeled semi-supervised SED in domestic environments, and monitoring of domestic activities based on multi-channel acoustics [4]. Among them, this paper describes a method for performing the fourth task of the DCASE 2018 challenge, large-scale detection of sound events using weakly labeled data. The goal of task 4 is to find the onset and offset of a sound event using the weak label in an audio clip. A variety of methods were proposed in the previous DCASE 2017 challenge [5, 6, 7, 8, 9] to solve this problem. Furthermore, a baseline system that performs the task is provided in the DCASE 2018 challenge [10]. Based on these previous studies, we propose a network with the Inception module and several ways to improve the performance. The remainder of this paper is organized as follows: Section 2 introduces the proposed network architecture for the weakly labeled semi-supervised SED; Section 3 presents the experimental settings and results using the DCASE 2018 dataset; and Section 4 draws the conclusions of our paper.

2. PROPOSED METHOD

We propose a weakly labeled semi-supervised SED method that uses the Inception architecture. To be specific, a CNN layer is implemented with the Inception structure and time information is learned using a bidirectional GRU. To perform SED, two separate learning stages are proposed. The first stage is for sound tagging and the second stage is for sound event detection.

2.1. Data augmentation

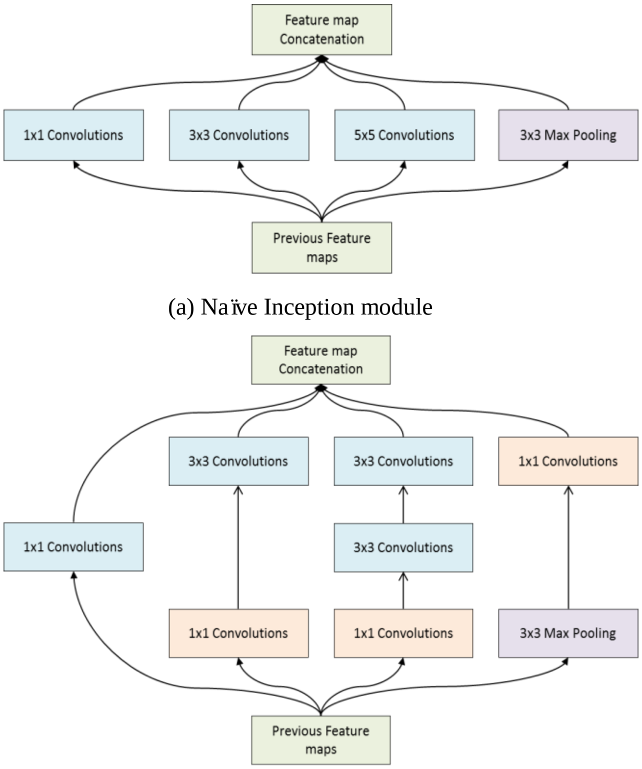
The fourth task of the DCASE 2018 challenge focuses on large-scale detection of sound events using small weakly labeled data. The challenge of this task is to explore the possibilities of leveraging large amounts of unbalanced and unclassified training data with a small set of annotated training data to improve system performance. As the weakly labeled data which is provided by DCASE challenge Task 4 is small, data augmentation is required to learn a better network. Data augmentation is the process of creating new training samples by making small changes to the original training data while keeping its characteristics. By performing the data augmentation, the network can be learned to improve its generalization ability for various unseen data [11]. Table 1 shows the data augmentation methods and details that we applied. To increase the performance of the classifier, we applied pitch shift manipulations with rates of 0.8, 0.9, 1.1, and 1.2. Moreover, the audio signal was stretched to 1.1 and 1.2 times faster and flipped horizontally to obtain a reversed image of the data.

Table 1: Data Augmentation methods and details.

Data Augmentation Method	Value
Pitch Shift	0.8, 0.9, 1.1, 1.2
Time Stretch	1.1, 1.2
Reverse	Horizontal flip

2.2. Inception module

In CNN, each convolution filter learns a local part of an image or feature map. In other words, it is a combination of information in the local receptive field. This is accomplished by passing these combinations through the activation function to infer non-linear relationships and make larger features smaller, such as by pooling. Therefore, it is important to see the various receptive fields in one convolution layer. In this regard, the Inception architecture has been proposed in [12, 13, 14]. The key concept of the Inception architecture is based upon finding the optimal local sparse structure in a convolutional vision network that can be approximated and applied. Intuitively, visual information must be processed at various scales and aggregated to abstract features of different scales at the same time. Figure 1 shows the scheme of the naïve Inception module and modified Inception module with dimension reductions.



(a) Naïve Inception module

(b) Modified Inception module with dimension reductions

Figure 1: The scheme of the Inception module.

2.3. Proposed architecture

Figure 2 presents the network structure of the Inception architecture employed in our proposed system. It is a convolutional recurrent neural network (CRNN) structure that combines CNN and RNN. In this system, the audio signal is first converted to 64 log mel-band energies to form an input vector to the network. Next, the stem layer for extracting low-level features is applied using 3 × 3 convolution filters. After which, these extracted low-level features are used to train various receptive fields at once through the Inception layers. The Inception layers in Figure 2 correspond to those in Figure 1-(b). Moreover, max pooling is performed with the Inception layer to compress the information in the frequency axis. The recurrent layer is then stacked using a bidirectional GRU to learn the relevance between time frames and connect the dense connection in every single frame. The final result is output through a global average pooling (GAP) layer.

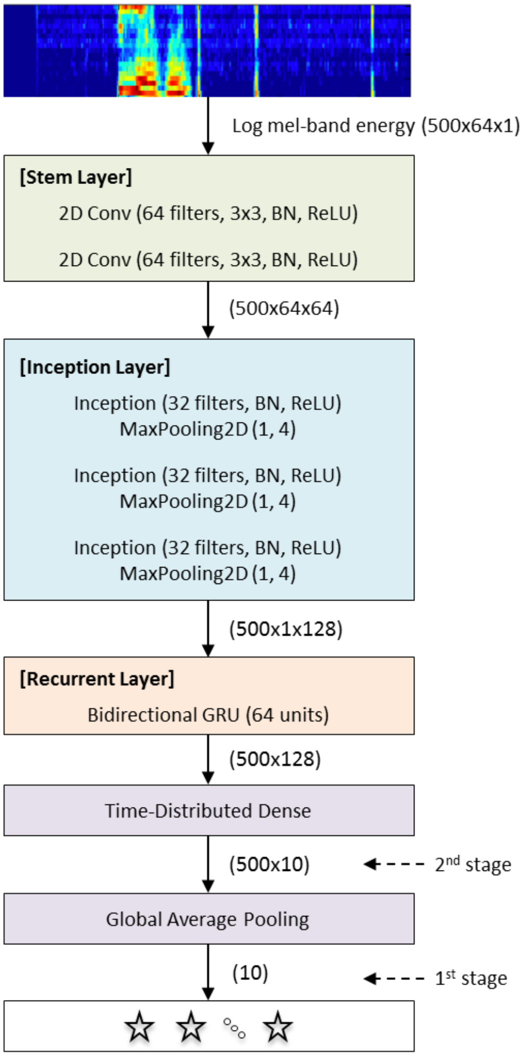

Log mel-band energy (500x64x1)
[Stem Layer]
2D Conv (64 filters, 3x3, BN, ReLU)
2D Conv (64 filters, 3x3, BN, ReLU)
(500x64x64)
[Inception Layer]
Inception (32 filters, BN, ReLU)
MaxPooling2D (1, 4)
Inception (32 filters, BN, ReLU)
MaxPooling2D (1, 4)
Inception (32 filters, BN, ReLU)
MaxPooling2D (1, 4)
(500x1x128)
[Recurrent Layer]
Bidirectional GRU (64 units)
(500x128)
Time-Distributed Dense
(500x10)
2nd stage
Global Average Pooling
(10)
1st stage

Figure 2: The structure of the proposed convolutional recurrent neural network for sound event detection with the Inception module.

2.4. First stage: sound tagging

In the first stage of the proposed system, a weak sound tagging network is trained using the weakly labeled data to assign predicted labels automatically to unlabeled in domain data. These unlabeled in domain data are excerpted from YouTube clips. To train the model, we use 80% of the weakly labeled data provided for the fourth task of the DCASE challenge as a training set; the rest of the data are used as a validation set. We apply the data augmentation method which is mentioned in section 2.1 to weakly labeled dataset to learn the generalized model.

2.5. Second stage: sound event detection

In the second stage, the SED network is trained using auto-tagged unlabeled in domain data which is automatically labeled in the first stage. This step is performed by excluding the GAP layer from the first stage network and outputting the frame recognition result. A simple method for learning strong label from weak label is to assign a strong label to all time frames. However, assigning strong labels from a weak label is difficult because it is impossible to know which frame has events or not. Owing to the absence of prior knowledge regarding the existence of the event, we calculate the log mel-band energies and assign a pseudo strong label when the average value of the log mel-band energies in each frame is above zero. This method assumes that there is no acoustic event in a frame where the energy is small. This is described as event activity detection in section 3.

3. PERFORMANCE EVALUATION

3.1. Pre-processing

To perform the experiment, we resampled the audio signal to a 16-kHz sampling rate and down-mixed it to a mono channel. The audio signal was then converted to 64 log mel-band energies with a frame size of 40 ms and an overlap length of 20 ms. At this time, the area of the mel-scale filter bank was normalized. As a result, we obtained an image with 500 time frames and 64 frequency bands and used it as a network input.

3.2. Hyperparameters and settings

In Table 2, we list the hyperparameters and settings used in this study. The number of nodes for each convolution layer was 64 or 32, and the convolution filter size was 3 × 3. The activation functions used in the networks were rectified linear unit (ReLU) [15] and sigmoid functions. The proposed network was optimized using an adaptive moment estimation (Adam) [16] optimizer with a learning rate of 0.001. The early stopping criteria were applied by monitoring the F-score with a patience value of 15.

Table 2: The hyperparameters and settings of the proposed network.

Parameter	Value
Convolution filter size	3 × 3
Activation function	ReLU, Sigmoid
Network initialzation	Glorot_uniform
Optimizer	Adam
Epochs	100
Learning rate	0.001

Early stopping method	Criteria = F-score
	Initial delay = 5
	Patience = 15

3.3. Post-processing

After obtaining the frame based probabilities from the last layer, we applied two binary decision methods. First of all, it can be accomplished in a straightforward manner by setting frame decisions to 1, if the probability is over the threshold of 0.5. Alternatively, the result in each frame can be decided by using the Viterbi algorithm [17, 18]. Given a set of predictions that indicate the conditional probability of a condition, the Viterbi algorithm computes the most likely state sequence in the observations. Therefore, we obtained the binary result by applying the Viterbi algorithm to the frame probabilities in each class. Then, the results of binary decisions are segmented and smoothed in the time domain using the median filtering method. In this regard, the median filter size is a critical factor for the detection of the onset and offset of a sound event depending on its length. However, applying median filtering of the same length to various sound events is not recommended because each sound event has different characteristics. Therefore, we selected the median filter sizes according to the estimated lengths of the events. In this study, we selected various filter sizes for each class according to the median values of the predicted event lengths.

3.4. Experimental results

We evaluated the performance of the proposed approaches for weakly labeled semi-supervised SED. In these experiments, we explored our proposed methods: Inception CRNN, data augmentation (DA) for weakly labeled data, event activity detection (EAD) for strong label learning, single-length median filtering (SMF) with 51 frames, and multi-length median filtering (MMF). Table 3 shows the experimental results of the proposed Inception CRNN based SED. As shown in the table, the Inception CRNN shows about 4.5% better performance than the DCASE 2018 baseline system. In addition, as can be seen in the results, the DA method showed little SED performance improvement. Nevertheless, the DA method was used because of the improved sound tagging performance in the first stage. We also confirmed that the EAD for pseudo strong labeling could improve the SED performance by about 2.0%. Furthermore, by applying the Viterbi algorithm, a performance enhancement of 1.2% was obtained. Finally, about 6.5% improvement was achieved by applying the MMF method.

Table 3: Experimental results of the Inception CRNN based SED.

Method	F-score	Precision	Recall
DCASE 2018 Baseline + SMF51	14.06%	-	-
Inception CRNN + SMF51	18.5%	18.5%	20.0%
Inception CRNN + DA + SMF51	18.9%	19.6%	19.6%
Inception CRNN + DA + EAD + SMF51 (Submission-1)	21.9%	23.3%	23.3%
Inception CRNN + DA + EAD + Viterbi + SMF51 (Submission-2)	23.1%	26.0%	23.4%
Inception CRNN + DA + EAD + MMF (Submission-3)	28.4%	28.3%	31.0%
Inception CRNN + DA + EAD + Viterbi + MMF (Submission-4)	29.3%	30.1%	30.3%

Consequently, the system that used MMF with the Viterbi algorithm showed the best performance. Compared to the DCASE 2018 baseline system, the proposed model achieved an approximately 15.24% gain in the F-score. Table 4 shows detailed experimental results of the proposed Inception CRNN with DA, EAD, Viterbi, and MMF.

Table 4: Detailed experimental results of the proposed system. (Submission-4)

Event label	F-score	Precision	Recall
Alarm/bell/ringing	28.4%	28.3%	28.6%
Blender	28.9%	26.0%	32.5%
Cat	12.6%	17.8%	9.8%
Dishes	10.6%	10.6%	10.7%
Dog	26.7%	30.3%	23.8%
Electric shaver/toothbrush	48.1%	50.0%	46.4%
Frying	13.3%	9.1%	25.0%
Running water	34.9%	28.6%	44.7%
Speech	16.5%	15.2%	18.1%
Vacuum cleaner	73.0%	85.2%	63.9%

4. CONCLUSION

In this paper, the Inception CRNN method that observes the various receptive fields in each CNN layer is proposed for large-scale detection of sound events using small weakly labeled data. By applying the proposed network structure to the SED system, it was shown that the Inception CRNN model achieves a better result than the baseline model. We also proposed various performance enhancement methods such as DA, EAD, Viterbi algorithm, and MMF to improve the SED performance of the proposed system. In conclusion, it was confirmed that the proposed method provide a higher SED result than the baseline system.

5. 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)

6. REFERENCES

[1] 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, Oct 2015.

[2] 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, Feb 2018.

[3] http://www.cs.tut.fi/sgn/arg/dcase2017/.

[4] http://dcase.community/challenge2018/.

[5] Y. Xu, Q. Kong, W. Wang, and M. D. Plumbley, "Surrey-cvssp system for DCASE2017 challenge task4," in Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE), 2017.

[6] D. Lee, S. Lee, Y. Han, and K. Lee, "Ensemble of Convolutional Neural Networks for Weakly-Supervised Sound Event Detection Using Multiple Scale Input," in Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE), 2017.

[7] J. Lee, J. Park, S. Kum, Y. Jeong, and J. Nam, "Combining Multi-Scale Features Using Sample-Level Deep Convolutional Neural Networks for Weakly Supervised Sound Event Detection," in Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE), 2017.

[8] S. Adavanne, and T. Virtanen, "Sound Event Detection Using Weakly Labeled Dataset with Stacked Convolutional and Recurrent Neural Network," in Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE), 2017.

[9] J. Salamon, B. McFee, and P. Li, "DCASE 2017 Submission: Multiple Instance Learning for Sound Event Detection," in Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE), 2017.

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

[11] B. McFee, E. Humphrey, and J. Bello, "A software framework for musical data augmentation," in Proceedings of the 19th International Society for Music Information Retrieval Conference (ISMIR), 2015.

[12] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, "Going deeper with convolutions," in Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), 2015.

[13] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, "Rethinking the Inception architecture for computer vision," in Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), 2016.

[14] C. Szegedy, S. Ioffe, V. Vanhoucke, A. Alemi, "Inception-v4, Inception-Resnet and the impact of residual connections on learning," in Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), 2017.

[15] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in Proceedings of the 27th International Conference on Machine Learning (ICML), 2010.

[16] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," in Proceedings of the 3rd International Conference on Learning Representations (ICLR), 2015.

[17] Viterbi, Andrew. "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," IEEE transactions on Information Theory, 13(2): 260-269, 1967.

[18] N. Ryant, M. Libeman, J. Yuan, "Speech activity detection on YouTube using deep neural network," in Proceeding of 14th Annual Conference of the International Speech Communication Association (INTERSPEECH), August 25-29, Lyon, France, pp. 728-731, 2013.