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

Technical Report

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 detection 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 and expansion of smart devices. 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. 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. Among them, this paper describes a solution to Task 4 of the DCASE 2018 challenge, large-scale detection of sound events using weakly labeled data. A variety of ways to solve this problem have been proposed in the previous DCASE 2017 challenge [4, 5, 6, 7, 8] and a baseline system in the DCASE 2018 challenge [9].

Based on these previous studies, we propose a network with the Inception module and several ways to improve 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 bidirectional GRU. To perform SED, two separate learning stages are proposed. The first stage is for audio tagging and the second stage is for event detection.

2.1. Data augmentation

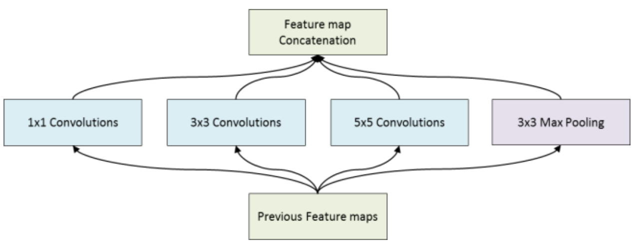
The goal of Task 4 is to evaluate the system for large-scale detection of sound events using small weakly labeled data. The challenge of this task is to explore the possibilities of leveraging large amounts of unbalanced and unclassified training data with a small set of annotated training data to improve 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 [10]. 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 data.

Table 1: Data Augmentation methods and details.

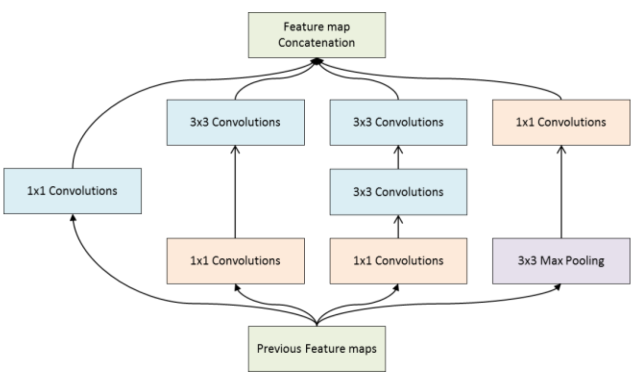
Data Augmentation Method	Value
Pitch Shift	0.8, 0.9, 1.1, 1.2
Time Stretch	1.1, 1.2
Reverse	Horizontal flip

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 [11, 12, 13]. 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 naive Inception module and modified Inception module with dimension reductions.



(a) Naive 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 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 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.

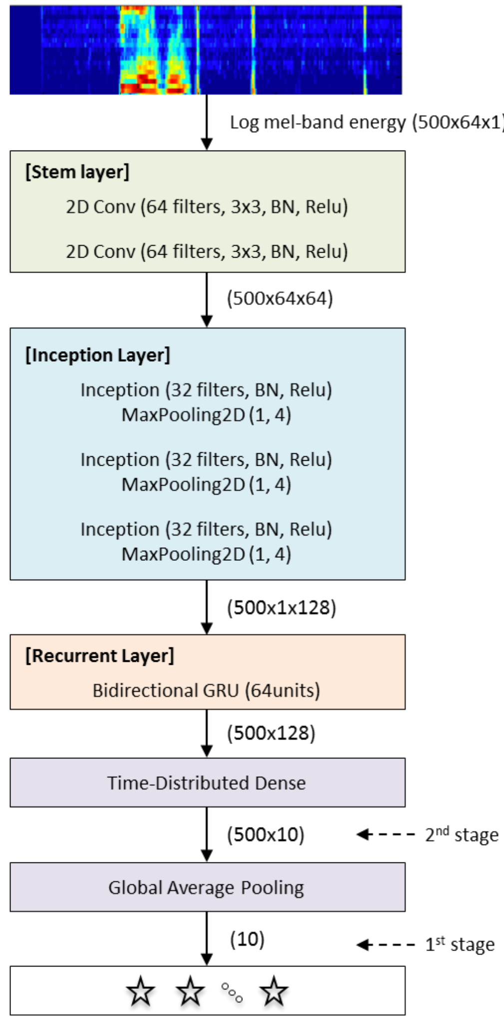


Figure 2: The structure of the proposed convolutional recurrent neural network for sound event detection with the Inception module.

2.4. 1st stage for sound tagging

In the first stage of the proposed system, we learn the weak sound tagging network from weakly labeled data and assign the sound label to unlabeled in domain data. To train the model, we use 80% as a training set and 20% as a validation set from weakly labeled data given by DCASE challenge Task 4. We apply the data augmentation method which is mentioned in section 2.1 to this weakly labeled dataset to learn the generalized model. The auto-prediction data is unlabeled in the domain data which is excerpted from YouTube clips.

2.5. 2nd stage for sound event detection

In the second stage, we learn the SED network from auto-tagged unlabeled in domain data. This step is performed by excluding the GAP layer from the 1st stage network and outputting the frame recognition result. The simple method to learn strong labels from the weak labels is to assign the strong label to all time frames. However, assigning the strong label from the weak label is difficult because it is impossible to know which frame has events or not. Due to the absence of prior knowledge regarding the existence of the event, we calculate the energy of each frame and assign a strong label when an average of log mel-band energies in each frame is above zero. This is described as event activity detection in Chapter 3.

3. PERFORMANCE EVALUATION

In this section, we describe the experimental setting and report the performance of the proposed Inception module based CRNN network.

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 40ms and an overlap length of 20ms. At this time, the area of the mel-scale filter bank is normalized. As a result, we obtain an image with 500 time frames and 64 frequency bands and use it as a network input.

3.2. The hyperparameters and settings

In Table 2, we list the hyperparameters and settings used in this paper. The number of nodes for each convolution layer is 64 or 32 and the convolution filter size is 3×3. The activation functions used in the networks are rectified linear unit (ReLU) [14] and sigmoid. The proposed network is optimized using an adaptive moment estimation (Adam) [15] optimizer with a learning rate of 0.001. The early stopping criteria are applied by monitoring F-score with a 15 patience.

Table 2: The hyper parameters and settings of the proposed network.

Parameter	Value
Convolution filter size	3x3
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 apply two binary decision methods. One straightforward way to accomplish this task is to set frame decisions to either 1 or 0, if the probability is over the threshold of 0.5. Another way to decide the result probability is the Viterbi algorithm [16, 17]. 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 value by performing the Viterbi algorithm on the frame probabilities in each class. Then by applying median filtering, the results can be segmented and smoothed in the time domain. In this regard, the median filter size is one of the critical factors to detect the onset and offset of the event depending on its length. Therefore, we select the median filter sizes according to the estimated length of the events. In this study, we selected various filter sizes for each event class according to the median values of predicted event length.

3.4. Experimental results

We evaluate the performance of the proposed approaches for weakly labeled semi-supervised SED. In these experiments, we explore our proposed methods: Inception CRNN, data augmentation (DA) for weakly labeled data, event activity detection (EAD) for strong label learning, static length median filtering (MF), and multi length MF. Table 3 shows experimental results of the proposed Inception CRNN based SED. As shown in the table, the proposed system shows better performance than the DCASE 2018 baseline CRNN system. Moreover, we also confirmed that the DA technique and EAD for strong labeling could improve the performance. Finally, the system using multi MF with Viterbi showed the best performance.

Table 3: Experimental results of Inception CRNN based SED. (macro-average)

Method	F-score	Precision	Recall
DCASE 2018 Baseline	14.06%	-	-
Inception CRNN + Static MF51	18.5%	18.5%	20.0%
Inception CRNN + DA + Static MF51	18.9%	19.6%	19.6%
Inception CRNN + DA + EAD + Static MF51 (Submission-1)	21.9%	23.3%	23.3%
Inception CRNN + DA + EAD + Viterbi + Static MF51 (Submission-2)	23.1%	26.0%	23.4%
Inception CRNN + DA + EAD + Multi MF (Submission-3)	28.4%	28.3%	31.0%
Inception CRNN + DA + EAD + Viterbi + Multi MF (Submission-4)	29.3%	30.1%	30.3%

Compared to the DCASE 2018 baseline, the proposed model achieved approximately 15.24% gain in F-score. Table 4 shows detailed experimental results of the Inception CRNN with DA, EAD, Viterbi, and multi MF.

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 for a large-scale detection of sound events using small weakly-labeled data is proposed which observes the various receptive field in one layer. We also propose various performance enhancement methods such as DA, EAD, Multi MF, and Viterbi algorithm to improve the detection performance of the system. By applying the proposed methods to the SED system, it was shown that the Inception CRNN model achieves a better result than the baseline model. Consequently, it was confirmed that the proposed method provides higher performance than the conventional methods.

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] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

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

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] D. P. Kingma and J. L. Ba, “Adam: A method for stochastic optimization,” arXiv preprint, arXiv: 1412.6980.

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

[17] 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.