End-to-end CRNN Architectures for Weakly Supervised Sound Event Detection

Technical Report

Hyeongi Moon[1], Joon Byun[2*], Bum-Jun Kim[2*], Shin-hyuk Jeon[2*],*
Youngho Jeong[3], Young-cheol Park[2*], Sung-wook Park[4*]*

[1]School of Electrical and Electronic Engineering, Yonsei University, Seoul, Korea
[2]Computer and Telecomm. Eng. Division, Yonsei University, Wonju, Korea
[3]Broadcasting &Telecommunications Media Research Laboratory, ETRI, Daejeon, Korea
[4]Gangneung-Wonju National University, Gangneung, Korea

ABSTRACT

This presentation describes our approach for large-scale weakly labeled semi-supervise sound event detection in domestic environments (TASK4) of the DCASE 2018. Our structure is based on Convolutional Recurrent Neural Network (CRNN) using raw waveform. The conventional Convolutional Neural Network (CNN) is modified to adopt Gated Linear Unit (GLU), ResNet, and Squeeze-and-Excitation (SE) network. Then three Recurrent Neural Networks (RNNs) follow. Each RNN receives features from different layers, respectively, and the outputs of RNNs are concatenate for final classification by Feed-forward connection (FC) layers. Simple data augmentation method is also applied to augment small amount of labeled data. With this approach F1 score of 5.5% improvement is achieved.

.

Index Terms— Sound event detection, weakly supervised learning, convolutional recurrent neural network, raw waveform, DCASE2018

1. INTRODUCTION

The DCASE 2018 task 4, so-called large-scale weakly labeled semi-supervise Sound Event Detection (SED) in domestic environments, evaluates systems detecting time-boundaries of audio events and the classes that the events belong to in an audio clip. In order to solve task 4, authors proposed networks based on CRNN (Convolutional Recurrent Neural Networks) which is similar to DCASE2018 baseline system, but with three major structural changes: GLU [1] adopted instead of ReLU, and ResNet [2] and SE [3] networks adopted to enhance SED performance. The combination of three models is inspired by GLU in CNNs [8] and ReSE [4] block.

The proposed networks also pursue an end-to-end solution, which receives time-domain raw audio samples as input of neural networks. This approach intends the proposed networks to utilize phase information that might be missed when only magnitudes of either Short-Time Fourier Transform (STFT) or Mel-spectrogram are received as input.

The dataset of DCASE2018 task 4 also influenced the training procedure of the proposed networks. The given dataset is a subset of Audioset [5] by Google, which consists of 10 sound-events classes and they are unbalanced. Each audio clip can have multiple events. The dataset has three sets of training data, which is a big change from dataset of DCASE 2017: the first set is weakly labeled training data consisting of 1,578 clips, the second set is unlabeled in-domain training data consisting of 14,412 clips, and the third set is unlabeled out-of-domain training data consisting of 39,999 clips. With this three-set preparation of training data, authors had to start to train the networks with the first set (weakly labeled) with data augmentation, then include the second and the third set which are unlabeled for training after sort of screening step.

This report will present proposed network structure and training method both for audio tagging and sound event detection.

2. PROPOSED METHOD FOR DCASE2018 TASK4

Fig. 1 shows the structure of the proposed networks. The structure consists of four components: (1) 1D-convolutional layer with a big stride [6] to extract low-level features, (2) ResGLU-SE layers (ResNet[2] + GLU[1], Squeeze-and-Excitation[3]) to process high-level features, (3) bi-directional RNN layers with multi-level feature aggregation to capture temporal information and (4) Fully Connected (FC) layer to predict posterior of each audio class at each frame.

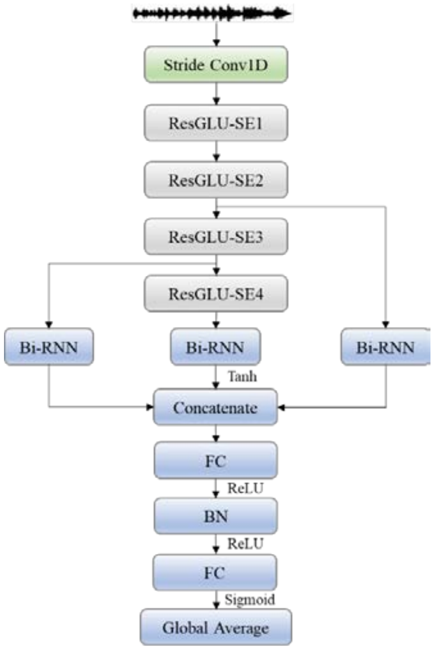

Figure 1: Block diagram of proposed DNN structure

The prediction probability at each frame is median-filtered for SED and global average pooling of all frames is used for audio tagging.

2.1. Stride 1D-convolution layer

The stride 1D-convolution [6] layer is used in end-to-end approach for audio tagging [6] or speech recognition [11][12]. The coefficients of a well-trained stride convolution layer has similar frequency-domain characteristics to filter-bank which has high resolution for low frequencies [6][11][12].

Although the stride layer behaves in a similar way to frequency-domain transform, it has an advantage of overcoming problems occurring in frequency-domain transformation. For example, the parameters employed for frequency-domain transform such as window size/type or hop size need not to be optimized [4]. The stride layer also considers the phase information in raw waveform which is discarded in Mel-frequency spectrum and magnitude spectrum of the STFT.

In the proposed structure, stride layer is at the bottom of the structure. Batch normalization layer, leaky ReLU activation layer and max-pooling for feature pooling [6] is applied after the stride layer.

2.2. ResGLU-SE block

Fig. 2 shows the structure of the ResGLU-SE block. SE block is connected after combination of GLU and ResNet. Max-Pooling is performed to reduce the temporal dimension at the end of the block. Following section describes details of the ResGLU-SE block.

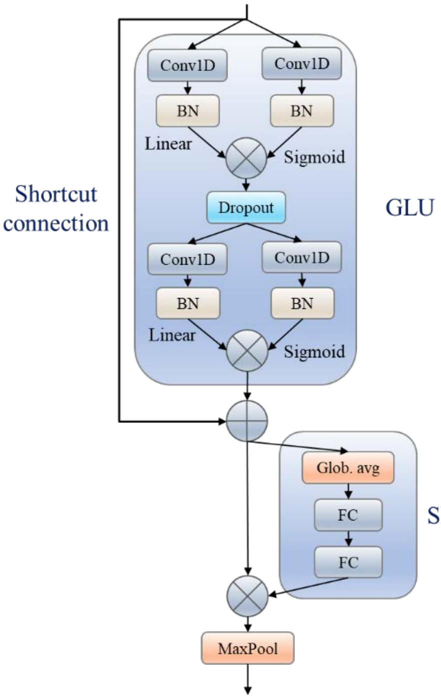

Figure 2: Block Diagram of the ResGLU-SE

2.2.1. Combination of GLU and ResNet

The concept of GLU is adopted instead of using ReLU activation function in CRNN [8]. GLU is proposed in [1] for language modeling and applied in SED [8]. It reduces the gradient vanishing problem of the deep neural network architecture [1]. Gating allows the network to attend to features related to audio events and ignore the unrelated features [8]. One ResGLU-SE block contains two GLUs and a dropout in the middle.

The shortcut connection of the ResNets [7] to GLU block is also adopted. It was observed that the optimization difficulty was made easier and the initial training speed was accelerated by shortcut connection [7].

In the preliminary test in audio tagging, authors found that the time to train the proposed networks was much longer than the time to train networks receiving power-spectrum. The combination of ResNet and GLU helps reducing the training time.

2.2.2. SE block

SE block is a structure derived from Squeeze-and-Excitation Networks. It adaptively recalibrates channel-wise feature responses [3]. This block follows after shortcut connections to enhance representation power of the ResGLU block.

In this process, by using global average pooling, the output of each channel is averaged by reducing the temporal dimensionality to 1. This global temporal information enters two FC layer structures that modelling interdependencies between channels [3]. Dimension of the first channel is same as input. As in the previous study [4], dimension of the second channel is found grid search.

2.3. Aggregation of Multi-Level Features

The outputs of the last three ResGLU-SE blocks are connected to each Bi-RNN block separately. The output features of the Bi-RNN is concatenated and passed by two FC layers. The channel size of the first FC layer is same as input vector dimension and batch normalization and ReLU activation is applied to the first FC layer. The global average pooling block in the Fig. 1 is applied for the audio tagging.

By combining multi-level features, we can supplement missing information from the last features. This allows you to detect sound events using more information than using only the ResGLU-SE block.

3. DATA AUGMENTATION

3.1. Audio tagging

Since the given dataset provides very small number of labeled data (1,578 clips for 10 classes), authors augmented the labeled data to train the proposed networks as much as possible. However, augmentation would not be sufficient for the networks to be

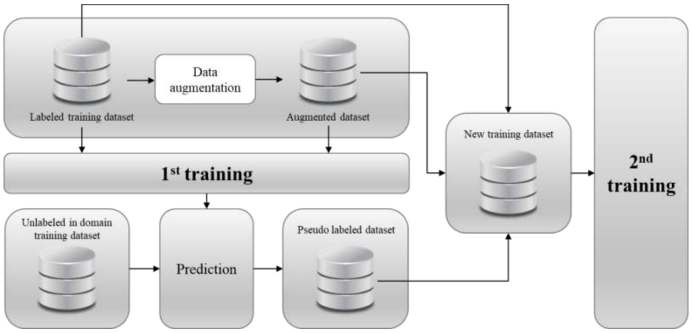

Figure 3: Block diagram of the network training process including data augmentation

generalized, so authors utilized the more real data, i.e. unlabeled in-domain training data (14,412 clips for 10 classes). The Fig. 3 describes the whole process of the network training including data augmentation.

Firstly, weakly labeled training dataset is augmented, the augmented dataset is combined with original dataset, and the new dataset is generated. The new dataset is assumed to have collect labels. Secondly, the unlabeled in domain dataset is labeled by the prediction and can be called a pseudo-labeled dataset. Authors merged the predicted set into the new training set and utilized for training at the 2nd training step.

In this presentation, following four augmentation methods: time stretching (TS), pitch shifting (PS), dynamic range compression (DRC) and block mixing (BM).

PS, TS and DRC are widely used in data augmentation for audio tagging [9]. TS changes speed of the audio while keeping the pitch and PS changes vice versa.

Two DRC curves shown in Fig. 4 are used to compress the dynamic range of the audio sample. The curves are taken from Dolby E standard [10].

The BM is implemented by simply adding two different audio clips. The labels of the mixed audio are the sum of the labels of the two source clips. Before adding two audio signals, both of them are normalized to have the same root mean square (RMS) value.

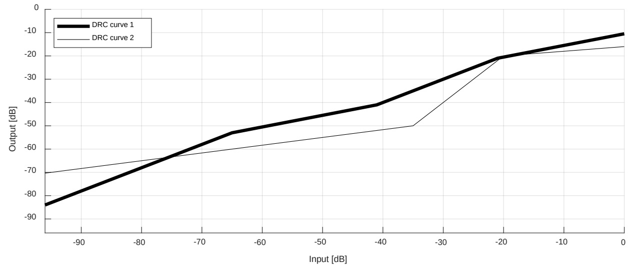

Figure 4: Two DRC curves

Augmentation method	TS*	PS**	DRC	BM
Number of the generated clips	3156	3156	3156	6312

Table 1. Data augmentation methods and the number of the generated data. *Time stretching rate of (0.81, 0.93, 1.07, 1.18) and **Pitch shift of (-2, -1, 1, 2) octave were used.

3.2. Sound event detection

All weakly labeled data has no time duration of the event. To improve time segmentation performance, the network should be trained by the strong labeled audio, which has onset and offset time of the events.

By the simple energy-based event detection, the dataset for the second training is strongly labeled. The RMS energy of the 0.5 second interval is obtained at 0.1 second intervals. Labeling is performed based on the averaged RMS value of entire audio clip. There is no prior strong label information, all labels of the audio are attached to detected intervals.

The BM is applied for strongly labeled audio clips as in audio tagging. Entire audio clip or the labeled section can be used in mixing.

4. EXPERIMENTS AND RESULTS

4.1. Architecture

Our architecture performs audio tagging and SED for the audio clips with 10 seconds long. Zero padding was performed for audio clips shorter than 10 seconds. The rectangular window size of 0.5 second and hop size of 0.1 second to extract single frame were used. Each frame is input of the strided 1-D convolutional layer.

The strided 1-D convolutional layer has filter size of 128 which is a frame size and stride length of 57 at sampling rate of 22.05kHz. Leaky ReLU activation with slope of 0.3 and max-pooling layer with pooling size 4 is applied after stride 1D-convolution layer.

Parameters of the ResGLU-SE layer are shown in Table 2. The amplifying ratio [4] is 12 and dropout rate is 0.2 for all the layers. Channel dimension of each Bi-RNN is 64, dropout ratio is 0.2 and activation function is tanh. After the Bi-RNN layer, the first FC layer has the channel dimension of 396 and ReLU activation layer. The second FC layer has the channel dimension of 10 and sigmoid activation layer.

For SED, median filter is applied to the output of the second FC layer. For audio tagging, global averaging layer averages the outputs of the second FC for all the frames of the input clip.

Binary cross-entropy is used as the loss function and Adam [6] optimization method is used.

4.2. Training and SED

4.2.1. 1st training and prediction

The proposed network in the Fig. 1 is trained in the first training. Two sampling rates of 44.1kHz and 22.5kHz were used as the input signal. 80% of weakly labeled data were used as training sets and 20% were used as validation sets. The same ratios were used for both case with and without data augmentation.

Layer	Filter dim.	Filter size	Filter stride	Pooling size	Pooling stride
1st	192	3	1	3	3
2nd	192	5	1	4	4
3rd	192	5	1	4	4
4th	384	5	1	4	4

Table 2. Parameters of the ResGLU-SE layer

Structure	Fs [kHz]	F1 [%]	Precision [%]	Recall [%]
Baseline	44.1	74.27	-	-
Proposed structure	44.1	70.20	77.32	65.34
	22.05	74.16	79.19	70.72
Proposed structure	44.1	89.58	94.00	85.86
	22.05	90.35	92.39	88.81
GLUs only*	22.05	88.26	92.54	84.99

Table 3. 1st training result using weakly labeled data. *Proposed structure without ResNet and SE

Structure	Fs [kHz]	F1 [%]	Error rate [%]
Baseline	44.1	14.06	1.54
GLUs only	22.05	7.8	2.06
Proposed structure	44.1	21.62*	1.56*
	44.1	24.3**	1.83**

Table 4. SED results with data augmentation. *Median filter size of 1.0 second and 0.5 second**

After 1st training, unlabeled in-domain data is labeled by the trained network. The following table summarizes the audio tagging performance of the proposed network in the first training.

4.2.2. 2nd training

In the second training, a structure in which the global pooling layer was removed from the Fig. 1 is used. Since the second training is pre-training for SED, it is assumed that tagged events are exist in all intervals of audio.

Two sampling rates of 44.1kHz and 22.5kHz were used as the input signal. Weakly labeled data for second training includes weakly labeled data, augmented data and labeled data by 1st trained network is used to training the network. 80% of weakly labeled data were used as training sets and 20% were used as validation sets, same as the first training.

4.2.3. 3rd training

Based on the simple energy-based event detection method, the starting and ending points of the events were selected. In the third training, labeling is performed within the detected time boundaries only. Using the re-labeled data, the trained network in 2nd training was retrained. Data augmentation was performed by applying BM between re-labeled audio.

4.2.4. Sound event detection

SED is performed using 3rd trained network. Median filter size for smoothing probability of the event is 0.5 second and 0.3 second. Table 4 shows the performance of the proposed method with two different median filter size and baseline system provided in DCASE 2018.

5. DISCUSSION

The proposed method is end-to-end approach to the SED. As in the case of the network trained by the raw speech/audio signal, the non-linear frequency selectivity of the stride 1-D convolution layer is observed in the 8~9kHz band.

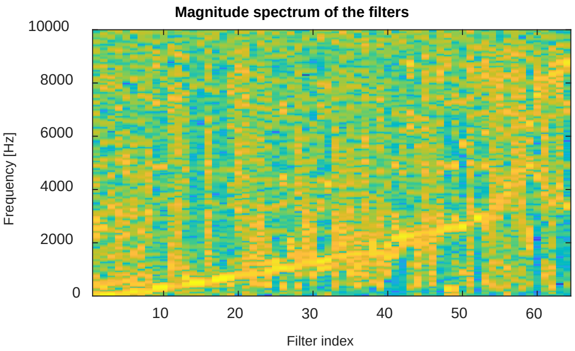

Figure 5. Magnitude spectrum of the stride 1-D convolution layers. Normalized and sorted.

In the audio tagging and SED results, the performance is increased by adding ResNet, SE block, multi-scale, etc. to the initial GLU structure. Also, augmentation can improve performance for limited data. The third training for Bi-RNN and its subsequent layers improves the F1 score about 5.5% with remaining error rate.

6. CONCLUSION

We proposed CRNN based network. Our convolutional block is combination of GLU with 1D convolutional block, ResNets and SE networks. The combined model with multi-level feature aggregation shows improvements in tagging performance compared with GLUs only structure. The tagging results are comparable to the frequency domain algorithm i.e. baseline system. The proposed structure with simple energy-based event detection shows better SED performance than that baseline.

7. ACKNOWLEDGMENT

This work was supported by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIT) (No.1711055381, Development of Human Enhancement Technology for auditory and muscle support)

8. REFERENCES

[1] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, “Language modeling with gated convolutional networks,” *arXiv preprint arXiv:1612.08083*, 2016.

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, “Identity mappings in deep residual networks,” in *European Conference on Computer Vision (ECCV)*. Springer, 2016, pp. 630–645.

[3] Jie Hu, Li Shen, and Gang Sun, “Squeeze-and-excitation networks,” *arXiv preprint arXiv:1709.01507*, 2017.

[4] Taejun Kim, Jongpil Lee, Juhan Nam, “Sample-level CNN architectures for music auto-tagging using raw waveforms,” in *Proc ICASSP*, 2018.

[5] 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 ICASSP*, 2017.

[6] Sander Dieleman, Benjamin, "End-to-end learning for music audio," in *Proc ICASSP*, 2014.

[7] Kaiming He Xiangyu Zhang Shaoqing Ren Jian Sun, "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[8] Yong Xu, Qiuqiang Kong, Wenwu Wang and Mark D. Plumbley, "Large-scale weakly supervised audio classification using gated convolutional neural network," in *Proc ICASSP*, 2018.

[9] Justin Salamon and Juhan Pablo Bello, "Deep Convolutional Neural Networks and Data Augmentation for Environmental Sound Classification," in *IEEE Signal Processing Letters*, 2016.

[10] "Standards and practices for authoring Dolby Digital and Dolby E bitstreams," *Dolby Labortories, Inc.*, 2002.

[11] Tara N. Sainath, Ron J. Weiss, Andrew Senior, Kevin W. Wilson, Oriol Vinyals, "Learning the speech front-end with raw waveform CLDNNs," in *Proc. INTERSPEECH*, 2015.

[12] Zolta´n Tu¨ske1, Pavel Golik1, Ralf Schlu¨ter1, Hermann Ney, "Acoustic Modeling with Deep Neural Networks Using Raw Time Signal for LVCSR," in *Proc. INTERSPEECH*, 2014.

[13] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.