Efficient CRNN network based on ContextGating and channel attention mechanism

Technical Report

Zhenwei Hou,Junyong Hao,Wang Peng

University of Chongqing,Chongqing,China
houzhenwei @cqu.edu.cn

ABSTRACT

In this technical report, we present the sound event detection system of task 4 (Sound event detection and separation in domestic environments) of the DCASE2020 challenge. We propose an improved CRNN that Context Gating and channel attention mechanism are co-embedded into backbone network. It aims to construct a general and efficient attention structure for extracting features of sound events, and give full play to the advantages of attention mechanism in event feature extraction. In the case of replacing the CRNN in the baseline model with the structure we designed and keeping the other parts unchanged, the macro F-score of our model on the validation set is 4 percentage points higher than the baseline.

***Index Terms*—** ContextGating, Channel Attention, CRNN

1. INTRODUCTION

DCASE2020 task 4[1] is the follow-up to DCASE 2019 task 4,which aims to explore the high-performance sound event detection system using weakly labeled data, unlabeled data and strongly annotated synthetic data. Sound event detection not only provides the possible multiple sound events in the audio, but also determines the starting and ending time of each sound event[2]. What is different this year is that sound separation and sound event detection can be combined to separate overlapping sound events and make it easy to extract foreground sound events from background sounds. In addition to the previous sed_eval toolboxes in the evaluation method, psds_eval toolboxes is added to assist in evaluating the performance of the detection system.

For the sound event detection task, CRNN is an important basic structure for extracting the features of sound events[3][4][5], and is widely used in various sound event detection models. The effect of sound event feature extraction largely determines the model's ability to classify different sound events, affecting the final detection result. The introduction of attention mechanism can make the model pay more attention to the areas that may be the features of sound events, and improve the model's ability to distinguish the features of sound events to some extent. In this technical report, Context Gating and channel attention mechanisms are combined in CRNN to build a general and efficient attention structure and to further explore the potential of attention mechanism in acoustic event detection.

2. METHOD

2.1 Audio Preprocessing

We resample the audio clips at 22,050 Hz, the size of the analysis window is 2048, the hop length is 365 and the number of mels is chosen to be 128 and then extract the log mel-spectrogram from the audio clips. We also normalized the mel-spectrum of each mel-bin by the global mean and the standard deviation of the bin value. Calculate the mean and standard deviation on the training set.

2.2 Model

The backbone network is composed of four blocks. In each block, the data is first filtered through context gating to filter important feature elements in the mel-spectrum, and then the channel attention layer is used to establish the relationship between the channels. the model uses the bidirectional GRU layer and the fully connected layer to output prediction results. The output from bidirectional followed by dense layers with sigmoid activation is considered as sound event detection result

3. RESULTS

In DCASE 2020 challenge’s task 4, the event-based macro F1 score is used to evaluate the performances of modules. PSDS submissions are optional. In Table 1, the results that we obtained for our proposed models are given for validation 2020 database. and we also list in detail the detection results of all sound events in Table2.

Table 1:F-scroe and PSDS score of model without fusion

Dataset	Macro F-score	PSDS macro F-score	PSDS	PSDS cross-trigger	PSDS macro
Validation	42.22 %	0.659	0.625	0.562	0.468

Table2: F-scroe of all sound events

Class	Macro F-score Event-based
Frying	29.7%
Blender	45.9%
Dishes	33.6%
Speech	56.1%
Dog	29.3%
Cat	48.1%

Running water	41.0%
Alarm/bell/ringing	28.3%
Vacuum cleaner	66.0%
Electric shaver/toothbrush	44.3%

4. CONCLUSION

In this paper, we proposed a sound event detect model based on the improved CRNN using Context Gating and channel attention mechanism. Finally our proposed models respectively reach 42.22% of F1-score (event-based) and 0.659 of PSDS macro F-score.

5. EFERENCES

[1] Turpault, Nicolas, et al. "Sound Event Detection in Domestic Environments with Weakly Labeled Data and Soundscape Synthesis." 4th Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE 2019) 2019.

[2] Y. Xu, Q. Kong, W. Wang, Mark D. Plumbley, “Attention and Localization based on a Deep Convolutional Recurrent Model for Weakly Supervised Audio Tagging” in arXiv: 1703.06052, 2017 A. Tarvainen and H. Valpola, “Mean teachers are better role models: Weight-averaged consistency targets improve semisupervised deep learning results” in arXiv: 1703.01780, 2017.

[3] L. JiaKai, “Mean teacher convolution system for DCASE 2018 Task 4,” Technical Report, DCASE2018 Challenge, 2018.

[4] A. Mesaros, T. Heittola, A. Diment, B. Elizalde, A. Shah, E. Vincent, B. Raj, and T. Virtanen, “Dcase 2017 challenge setup: Tasks, datasets and baseline system,” in DCASE 2017Workshop on Detection and Classification of Acoustic Scenes and Events, 2017.

[5] A. Tarvainen and H. Valpola, “Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results” in arXiv: 1703.01780, 2017.