SEMI-SUPERVISED SOUND EVENT DETECTION USING MULTI-SCALE CONVOLUTIONAL RECURRENT NEURAL NETWORK AND WEIGHTED POOLING

Technical Report

Dongchi Yu, Xichang Cai, Duxin Liu, Zihan Liu

North China University of Technology, Beijing, China
2019312100109@mail.ncut.edu.cn

ABSTRACT

In this technical report, we describe our submission system for DCASE2021 Task4: sound event detection and separation in domestic environment. We mainly focus on the scenario that recognizes sound events without source separation. Since the duration of different sound events could be quite different, our model employs a multi-scale convolution recurrent network to extract the multi-scale features of an audio sequence. For more efficiently utilizing weak label training data, a global weighted pooling strategy is introduced to aggregate frame level predictions to generate clip level prediction. Additionally, our model also use mean teacher semi-supervised learning technique and data augmentation. We demonstrate that the proposed method achieves the PSDS2 score of 0.61 and the event-based macro F1 score of 42.15% on the validation set.

Index Terms— Sound event detection, weakly supervised learning, multi-scale convolution recurrent network

1. INTRODUCTION

The goal of DCASE2021 task4 [1] is to build sound event detection (SED) system, which provides not only the event class but also the event time boundaries given that multiple events can be present in an audio recording. The training set of this task consists of three parts: a small weak labeled set (without timestamps), a large amount of unlabeled set and strongly annotated synthetic set. The challenge of this task is that the SED system needs to be trained without strong labeled real recoding data.

DCASE2021 task4 also encourages participants using sound separation jointly with sound event detection. This task is divided into three scenarios: 1) working sound event detection without source separation pre-processing; 2) working on both source separation and sound event detection; 3) working only on source separation and use the sound event detection baseline.

In this report, we introduce our SED system designed for task 4 of DCASE 2021 challenge. Our proposed method focuses on scenario one and mainly makes two contributions. Firstly, for detecting sound events with different duration, we proposed a multi-scale convolution recurrent network model. Secondly, training SED system with weakly labeled data could be treated as a multiple instance learning (MIL) problem [2], [3]. The frame level predictions need to be aggregated to produce the clip level predictions. We introduce a global weighted pooling strategy to address this problem.

We conduct experimental evaluations on the DCASE2020 Task4 validation set. The experimental results show that the proposed models outperform the baseline system.

2. PROPOSED METHODS

2.1. Audio Preprocessing

The sampling rate of clips is 44.1k Hz. We resample all audio clips at 16k Hz. A 2048-point hamming window with the hop size of 256 is then adopted to divide the raw audio clips into frames. 2048-point FFT and 64 log-Mel filter banks are used to extract log-Mel feature on each frame. Finally, the 10s raw audio clips are converted to the log mel spectrogram features with the shape of 626 by 64.

2.2. Network Architecture

Inspired by the success of the feature pyramid architectures in object detection field [4], we realize that multi-scale feature maps would be useful for this task. For extracting multi-scale feature maps, we modify the CRNN model of DCASE2021 task4 SED baseline [5] with multi-scale CNN. Our proposed model consists of three parts: a multi-scale CNN feature extractor, a RNN feature extractor and a classifier.

The multi-scale CNN feature extractor consists of a 2-dimensional CNN and a 1-dimensional CNN. The filters and pooling sizes of the 2-dimensional CNN are [64, 64, 64] and [4, 4, 4] respectively. After 2-dimensional CNN, there are three parallel 1-D CNN whose kernel sizes are [3, 3, 3], [5, 5, 5], [7, 7, 7] respectively and filters are both 64. The architecture of RNN feature extractor and classifier are basically consistent with baseline, except for a few slight modifications to fit the dimension of the multi-scale CNN feature extractor. The overall network structure is shown as Figure. 1.

2.3. Global Weighted Average Pooling

In order to train SED system with weak label data, the frame level prediction output by the system needs to be aggregated into clip level prediction to calculate the loss function. Frames with target sound events are expected to be more important than other frames. We introduce a temporal attention mechanism and a weighted pooling strategy to address this problem.

The temporal attention mechanism is formulated as follows:

$$a_{ij} = softmax\left(W_2^T \tanh\left(W_1 F_{ij}^T + b\right)\right) \quad (1)$$

$$\mathrm{P} = \sum_i \sum_j a_{ij} p_{ij} \quad (2)$$

where W_1 and W_2 are trainable weight parameter matrices of this two-layer attention network, and b is the bias parameter matrix. F_{ij}^T is the frame level feature vector, a_{ij} is the weight corresponding to frame level prediction p_{ij} and P is the final clip level prediction.

The weighted pooling strategy is formulated as follows:

$$\mathrm{P} = \frac{\sum_i \sum_j p_{ij} \times p_{ij}}{\sum_i \sum_j p_{ij}} \quad (3)$$

in which there are no trainable parameters.

In our experiments, the performance of weighted pooling strategy is better. Therefore, the following steps will be based on it.

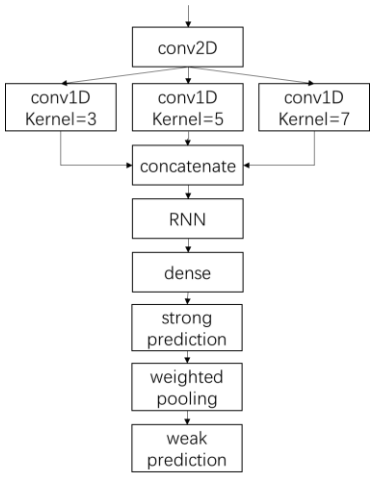

Figure 1: Architecture of the proposed network.

2.4. Semi-supervised Learning

To learn from unlabeled training data, we implement a semi-supervised learning method called mean teacher [6]. The mean teacher method consists of two network model called student model and teacher model respectively. Student and teacher model share the same structure of our proposed multi-scale CRNN. The weights of the student model are updated with gradient back propagation, and the weights of the student model are updated as an exponential moving average (EMA) of the student weights. The same data is input into two models, and the network parameters are optimized according to the consistency regularity of the two network outputs.

2.5. Data Augmentation

We employ mixup [7] for data augmentation. Mixup can improve the performance of deep neural network in many machine learning tasks by smoothing the distribution of samples in the feature space. This method creates a new data by interpolate between two raw data, while the labels are interpolated in the same way. This process is expressed as

$$\tilde{x} = \lambda x_i + (1-\lambda) x_j \quad (4)$$
$$\tilde{y} = \lambda y_i + (1-\lambda) y_j \quad (5)$$

where x_i, x_j are different data points, and y_i, y_j are corresponding labels.

3. EXPERIMENTS

3.1. Dataset

The DCASE2021 task4 dataset can be divided into four subsets, including training sets (synthetic strongly labeled: 10,000 clips, weakly labeled: 1,578 clips, unlabeled: 14,412 clips) and validation set (1,168 clips). The duration of majority audio clips is 10 seconds, and multiple audio events may occur at the same time.

3.2. Training

The Adam optimizer and learning rate of 0.001 are used for training. The batch size and number of epochs are 48 and 200 respectively. We employ the exponential warmup strategy to gradually increase the learning rate from very small to 0.001 in the first 50 epochs, the learning rate remained unchanged during subsequent training. We used an nvidia GTX1080 GPU to train the model.

3.3. Evaluation Metrics

The performance of SED system is evaluated with poly-phonic sound event detection scores (PSDS) [8]. We compute PSDS using 50 operating points from 0.01 to 0.99. In order to better understand the system performance, two PSDS scenarios are used to evaluate the different capabilities of the system. Scenario 1 requires the system to respond quickly to sound events, and scenario 2 requires avoiding class confusion. Additionally, event-based measures with a 200 ms collar on onsets and a 200 ms / 20% of the events length collar on offsets are used as a contrastive measure. These metrics were calculated using sed_eval [9] and psds_eval toolboxes [10].

3.4. Result

Our model achieves the PSDS2 score of 0.61 and the event-based macro F1 score of 42.15% on the validation set. Table. 1 shows the event-based F1 score for each event class.

Table 1: The event-based F1 for each event class

event class	**F1_score**
Blender	39.1%
Frying	44.7%
Cat	57.5%
Dog	38.2%
Speech	51.5%
Vacuum cleaner	36.4%
Dishes	21.1%
Running water	25.4%
Electric shaver/toothbrush	46.0%
Alarm/bell/ringing	35.7%

4. CONCLUSIONS

In this technical report, we have described our submission system for DCASE2021 Task4. Our proposed sound event detection method is based on multi-scale convolutional recurrent neural network and weighted pooling strategy and outperforms the baseline.

5. REFERENCES

[1] Nicolas Turpault, Romain Serizel, Ankit Parag Shah, and Justin Salamon. Sound event detection in domestic environments with weakly labeled data and soundscape synthesis. In Workshop on Detection and Classification of Acoustic Scenes and Events. New York City, United States, October 2019. URL: https://hal.inria.fr/hal-02160855.

[2] W. Zhu, Q. Lou, X. Vang, and Y. Xie, "Deep multi-instance networks with sparse label assignment for whole mammogram classification," in Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent., 2017, pp. 603–611.

[3] X. Wang, Y. Yan, P. Tang, X. Bai, and W. Liu, "Revisiting multiple instance neural networks," Pattern Recognit., vol. 74, pp. 15–24, Feb. 2018.

[4] Lin, Tsung-Yi et al. "Feature Pyramid Networks for Object Detection." 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017): 936-944.

[5] https://github.com/DCASE_REPO/DESED_task/tree/master /recipes/ dcase2021_task4_baseline

[6] Tarvainen, Antti and H. Valpola. "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results." NIPS (2017).

[7] Zhang, Hongyi et al. "mixup: Beyond Empirical Risk Minimization." ArXiv abs/1710.09412 (2018): n. pag.

[8] Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen. Metrics for polyphonic sound event detection. Applied Sciences, 6(6):162, 2016. URL: http://www.mdpi.com/2076-3417/6/6/162, doi:10.3390/app6060162.

[9] https://github.com/TUT-ARG/sed_eval

[10] https://github.com/audioanalytic/psds_eval