AN IMPROVED SYSTEM FOR DCASE 2019 CHALLENGE TASK 4

Technical Report

Zhenyuan Zhang
University of Electronic Science and Technology of China
School of Information and Communication Engineering
Chengdu, China
zhenyuan_zhang@std.uestc.edu.cn

Mingxue Yang
University of Electronic Science and Technology of China
School of Information and Communication Engineering
Chengdu, China
michelleyang2017a@gmail.com

Li Liu
University of Electronic Science and Technology of China
School of Information and Communication Engineering
Chengdu, China
liuli@std.uestc.edu.cn

ABSTRACT

In this technical report, we present an improved system for DCASE2019 challenge task 4, with the goal to evaluate systems for the detection of sound events using real data either weakly labeled or unlabeled and simulated data that is strongly labeled .We use the multi-scale Mel-spectra as the feature and do the detection with the 3 layers convolutional neural network(CNN) and 2 layers recurrent neural network (RNN), after each layer of CNN, we apply a ResNet (Residual Neural Network) block to increase learning depth. Aim to use data without labels or with weak labels, we apply the mean–teacher model to do the sound event detection.

Index Terms— sound event detection, weak label, Semi-supervised learning, Residual Neural Network, mean-teacher model

1. INTRODUCTION

Sound event detection is the task of judging, detecting and classifying the sound information in real environment, and judging the starting and ending time points. The SED has two main tasks, including monophonic sound event detection and polyphonic sound event detection. Compared with monophonic sound event recognition, polyphonic sound event recognition presents more challenges because the recording of multi-sound event recognition has a large number of overlapping sound events at the same time. In real life, due to weather, environment and other reasons, the appearance of sound is often not separate; in the judgment of the scene also need to consider a variety of sounds, so we should pay more attention to polyphonic sound event detection. [1]

SED has been studied by many scholars. The feature and classification technology were selected relatively single in the early time. MFCC was used as the sound feature, and the traditional HMM [2] or its classification model [3] used to classify the sound events.

With the development of SED, a large number of scholars have studied the selection of characteristics. Valenti M, Tonelli D and Vesperini F discussed how to combine and preprocess the voice data of double channels when extracting features, and compared MFCC, log-mel and other features extracted from single channel data or double channel data to select suitable features for sound event detection [4]. Other scholars contrast discussed several different characteristics on the result of identification, the author selected the MFCC, PLP and loudness three characteristics, finally suggests PLP has good effect [5], and logMel and logAvgMel are combined to discuss whether two kinds of features can obtain better effect [6]. At present, SED mainly adopts frequency domain features such as MFCC and log-mel.

Thanks to the development of deep learning technology, classification technology also has more choices. Deep neural network (DNN) was selected for classification, and a combination of single-label recognition neural network used DNN was selected for multi-label analysis. The threshold of each label was also discussed [7]. There is also a method that focus on the pooling mode of pooling layer in the convolutional neural network [8], the "automatic pooling" which is different from the traditional pooling modes such as "maximum" and "minimum" was used. There are also methods analyzing the effect of Recurrent Neural Network (RNN) on classification performance [9], [10]. The RNN and convolution neural network (CNN) was combined as the convolution Recurrent Neural network (CRNN).

At the same time, we also noticed the method of achieving excellent results in DCASE2018, one of which adopted EAD to strengthen the labels of weak and unlabeled events, and then adopted CNN and RNN to detect and classify [11]. And another one proposed the semi-supervision

framework of the mean teacher, and used CRNN for detection and classification, which achieved good results [12].

Because it is difficult to label sound time accurately in daily environment, how to use a large amount of unlabeled data or weakly labeled data to detect sound events has become a hot topic of current research, and DCASE2019 task4 focuses on this point. The target of the systems is to provide not only the event class but also the event time boundaries given that multiple events can be present in an audio recording. The challenge of exploring the possibility to exploit a large amount of unbalanced and unlabeled training data together with a small weakly annotated training set to improve system performance.

Our technical report is organized as follow: our proposed method is described in Section 2. Then, experiment and its results are presented in Section 3. Finally, we give the conclusions in Section 4.

2. PROPOSED METHOD

2.1. Multi-Channel feature

According to [10], we proposed a multi-channel feature extraction method. We choose a two-time scale as the frame length to extract the feature of the input data, and both of the two features are extracted in hop length of 511 points, to keep the number of them same. Then, we combine the feature with two scales as two channels of the input feature, just like multi-channel in image processing.

We choose 64 bit Mel- spectrogram as the feature and combine them to a T*64 input vector， where T means the number of frame.

2.2. Semi-Supervised learning model

We used the mean-teacher method proposed in [12], and train the model as the same way. The mean-teacher model is first mentioned in [13]， there are two parts named teacher model and student model in the network. The network parameters for the student model are update with gradient descent, and the parameters for the teacher model are derived from the exponential moving average of the student model's network parameters.

The loss of the model is combined with two cost: classification cost and consistency cost. The classification cost is calculated from the difference between the prediction and the label, and the consistency cost is defined with the expected distance between the prediction of the student model and the prediction of the teacher model, which can be used by both labeled and unlabeled data.

For the task, we defined consistency cost in clip and in frame, which is calculated by obtained by comparing the logits of both the student model and the teacher model for the whole audio clips. Aim to achieve the batter Noise Resistance, we add random noise to data onto the teacher model. Figure 1 shows the details.

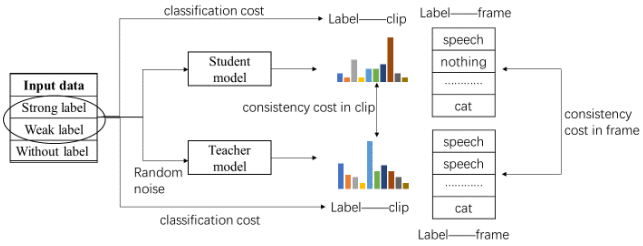

Figure 1：The mean-teacher method we use for the task

2.3. Neural network

We use 3 layers Convolutional Neural network (CNN), with a Residual Neural Network (ResNet) block behind every CNN layer. And then there are 2 layers of recurrent neural network (RNN). We add dropout layer after each layer of the CNN and RNN, with a 0.5 dropout.

The ResNet was first proposed in [14] to solve the degradation when the learning depth increased. The network has been applied in image processing, and achieved good results. This method changes the learning object from the output to the difference between the input and the output. Figure 3 shows the structure of a Residual block. The Residual block is implemented via shortcut connection. The input and output of the block are superimposed by shortcut as an element-wise. This simple addition adds no additional parameters or computation to the network. At the same time, it can greatly increase the training speed of the model and improve the training effect.

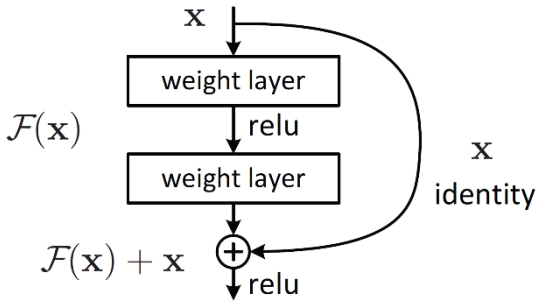

Figure 2：the Residual block

In addition, our CNN uses Gated Linear Unit (GLU) as our activation function.

$$Y = \sigma(\omega \cdot X + \beta) \odot X \tag{1}$$

Where $X \in R^{n \times n}$ is the input feature vector, σ is the element-wise Sigmoid activation and $\odot$ is the element-wise multiplication. $\omega \in R^{n \times n}$ And $\beta \in R^{n \times n}$ are trainable parameters,

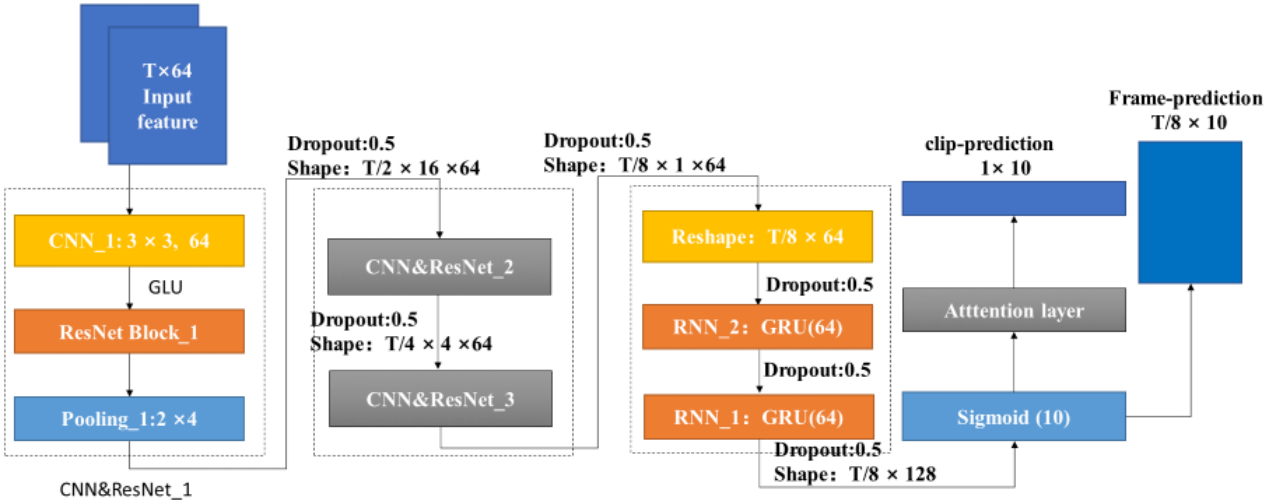

Figure 3: Model of the network. This is for one scale feature . When it comes two , double the reshape as: T/8×64 ×2 because of the 2 channel.

weights and bias. The model of the network is shown in Figure 3.

2.4. Attention layer

We apply an attention layer to contribute the label of each frame to a clip. We use the method mentioned in [15], which is defined as the Linear softmax pooling function.

$$y = \frac{\sum_i y_i^2}{\sum_i y_i} \quad (2)$$

Where y means the aggregated clip-level Probability, and y_i means the predicted probability of the same event type at the i-th frame. The linear softmax pooling function has the following advantages: (1) it allows the gradient to flow unobstructedly; (2) it achieves a balance between false negatives and false positives for localization; (3) its predictions on the recording level and the frame level are relatively consistent.

3. EXPERIMENTAL RESULTS

3.1. DCASE2019 dataset

The dataset for this task is composed of 10 sec audio clips recorded in domestic environment or synthesized to simulate a domestic environment. The task focuses on 10 class of sound events that represent a subset of Audioset. The dataset for DCASE 2019 task 4 is composed of a subset with real recordings (from Audioset) and a subset with synthetic recordings. Real recordings are extracted from Audioset. It consists of an expanding ontology of 632 sound event classes and a collection of 2 million human-labeled 10-second sound clips. And the synthetic set is composed of 10 sec audio clips generated with Scaper. The foreground events are obtained from FSD. Each event audio clip was verified manually to ensure that the sound quality and the event-to-background ratio were sufficient to be used an isolated event.

3.2. Data pre-processing and feature extraction

Since the speech length of each segment is 10 seconds, and sampling rate is 44100Hz, we adopted a fixed number of frames input network and divided the data into one frame according to 2048 points with hop length of 511points. We use 64bit Mel – spectrogram for every frame as the feature, so we get a feature with the shape of 846*64 for each channel.

3.3. Baseline

The baseline method for the dataset used is provided in [12].

3.4. Experimental setup

Because of the length and sampling rate, we get a frame with 46.44ms, and 846 frames for the 10s clip. Due to the time pooling by 8 (2×2×2), we get 106 frames output, which means each frame has the same length of 92.59ms. Also, we extracted features both one channel and two for all the audio clips, no matter labeled or unlabeled, and evaluated the model on the dataset 'Validation 2019 'and 'Evaluation 2018'. The dropout after all layers was set as 0.5.

3.5. Experimental results

In the DCASE2019 task4, the event-based F1-score is used to evaluate the performances of modules, and segment-

based F1-score as a secondary measure. We give the result in Table 1 and Table 2.

Table 1:event -based F-score metrics (macro averaged)

	Validation 2019	Evaluation 2018
Single channel	30.73%	29.34%
Double channel	32.81%	31.60%
Baseline	23.7 %	20.6 %

Table 2:segment-based F-score metrics (macro averaged)

	Validation 2019	Evaluation 2018
Single channel	65.06%	62.74%
Double channel	62.78%	60.66%
Baseline	55.2 %	51.4 %

To discuss the result in two scale features, we also analyze the class-wise result. And they are shown in table 3 and table 4.

Table 3:event -based F-score metrics (macro averaged) for single channel feature

	Validation 2019	Evaluation 2018
Speech	51.4%	51.3%
Dog	6.3%	6.5%
Cat	35.5%	35.5%
Alarm/bell/ ringing	42.6%	44.5%
Dishes	7.5%	7.3%
Frying	14.0%	10.6%
Blender	42.6%	42.6%
Running water	26.9%	22.4%
Vacuum cleaner	50.6%	46.9%
Electric shaver/tooth-brush	29.8%	29.1%

Table 4:event -based F-score metrics (macro averaged) for double channel feature

	Validation 2019	Evaluation 2018
Speech	46.0%	45.5%
Dog	15.6%	15.8%
Cat	36.3%	27.8%
Alarm/bell/ ringing	40.5%	44.0%
Dishes	18.5%	16.3%
Frying	26.4%	25.7%
Blender	28.3%	30.3%
Running water	29.9%	26.4%
Vacuum cleaner	54.5%	56.9%
Electric shaver/tooth-brush	32.3%	27.5%

Table 1-4 show the result for validation 2019，with the best event-based F1-score is 32.81% and the segment-based is 62.78% with the double-channel feature. The result for Evaluation 2018 is 31.60% and 60.66%. We noticed that the class-wise event-based F-score has a performance improvement in some classes such as ‘Dog’ ,’Cat’, ‘Dishes’ and so on, with a bad effect using single-channel. However, it leads to some reduce of some other classes. Therefore, it is still a problem to choose different features for different sound event.

4. CONCLUSION

In this technical report, we proposed a network consisting of 3-layer CNN with 3 ResNet block behind each layer, and 2-layer RNN to improve the F-score of the event-based. And we get an event-based F-score with 32.81% for the Validation 2019 and 31.60% for the Evaluation 2018, which are better than the baseline with 23.7% and 20.6%. Also, we find out that a multi-channel feature will give a contribution to improve the effect.

5. REFERENCES

[1] A. Dang, T.H. Vu, J.C. Wang, “A survey of Deep Learning for Polyphonic Sound event detection”, in International Conference on Orange Technologies(ICOT), 2017. pp. 75-78.

[2] A. Mesaros, T. Heittola, A. Eronen, et al. "Acoustic event detection in real-life recordings." 18th European Signal Processing Conference IEEE, 2014.

[3] Y. T. Peng, C. Y. Lin, M. T. Sun, et al. "Healthcare audio event classification using Hidden Markov Models and Hierarchical Hidden Markov Models." IEEE International Conference on Multimedia & Expo IEEE Press, 2009.

[4] M. Valenti, D. Tonelli, F. Vesperini, et al. “A Neural Network Approach for Sound Event Detection in Real Life Audio”, in European Signal Processing Conference. EURASIP, 2017, pp. 2754-2758.

[5] K. Feroze, A. R. Maud. “Sound Event Detection in Real Life Audio using Perceptual Linear Predictive Feature with Neural Network”, International Bhurban Conference on Applied Sciences and Technology ,2018, pp.377-382.

[6] I. Y. Jeong, S. Lee, Y. Han. “Audio Event Detection Using Multiple-input Convolutional Neural network”, in Detection and Classification of Acoustic Scenes and Events. IEEE, 2017.

[7] E. Cakir, T. Heittola, H. Huttunen, et al. “Multi-label vs. combined single-label sound event detection with deep neural networks”, in Signal Processing Conference. IEEE, 2015.

[8] B. Mcfee, J. Salamon, J. P. Bello. “Adaptive pooling operators for weakly labeled sound event detection”. IEEE/ACM Transactions on Audio, Speech, and Language Processing 26.11(2018):2180-2193. 2018, 26(11): 2180-2193.

[9] Cakir, Emre, et al. "Convolutional Recurrent Neural Networks for Polyphonic Sound Event Detection." IEEE/ACM Transactions on Audio, Speech, and Language Processing 25.6(2017):1291-1303.

[10] S. Adavanne, T. Virtanen. “A Report on Sound Event Detection with Different Binaural Features”, in Detection and Classification of Acoustic Scenes and Events. IEEE, 2017.
[11] Y. Liu, J. Yan, Y. song, et al. “USTC-NELSLIP System For DCASE 2018 Challenge Task 4”, in Detection and Classification of Acoustic Scenes and Events. IEEE, 2018.
[12] J.K. Lu. “Mean Teacher Convolution System For DCASE 2018 Task 4”, in Detection and Classification of Acoustic Scenes and Events. IEEE, 2018.
[13] A. Tarvainen , and H. Valpola. "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results", in 31st Conference on Neural Information Processing Systems (NIPS).IEEE, 2017.
[14] K. He, X. zhang, S.Ren, et al. "Deep Residual Learning for Image Recognition." , in arXiv: 1512.03385, 2015.
[15] Y. Wang, J. Li , and F. Metze. "A Comparison of Five Multiple Instance Learning Pooling Functions for Sound Event Detection with Weak Labeling", in 1810.09050, 2018.