Semi-Supervised Sound Event Detection System with Pretrained Model

Technical Report

Juan Wu, Yanggang Gan

North China University of Technology,
Beijing,China
994266693@qq.com

Xichang Cai, Menglong Wu*

North China University of Technology,
Beijing,China
caixc_ip@126.com

ABSTRACT

In this report, we present the sound event detection system for Detection and Classification of Acoustic Scenes and Events (DCASE) 2023 Challenge Task 4: Sound Event Detection with Weak Labels and Synthetic Soundscapes. For Task 4A, we designed a SED system based on the Mean Teacher [1] architecture to detect event information and start and stop times in audio sequences, using semi supervised learning to address the lack of labeled data in the DCASE 2023 Challenge task. In addition, we use pre-trained models to leverage external data information to further improve the stability of the system. We finally integrated multiple systems with the best PSDS1 of 0.525 and PSDS2 of 0.783.

Index Terms—Sound event detection, Semi-supervised learning, Mean teacher, Pretrained model

1. INTRODUCTION

The purpose of sound event detection is to identify each sound event category and to assist people or other intelligent devices to make adaptive responses by automatically analyzing the event type and timestamp information contained in the sound to further facilitate people's lives and improve the efficiency of production. Sound event detection has in-depth research significance in smart homes [2], health monitoring systems [3], multimedia retrieval [4], smart city planning [5], and automatic audio monitoring [6], etc. The CRNN model structure has achieved good performance in SED systems, which fully retains the advantages of CNN and RNN, not only extracting the salient features on the feature map through convolutional operations This structure fully retains the advantages of CNN and RNN, not only extracts the significant feature information on the feature map by convolutional operation, but also models the temporal dimension by capturing a larger range of contextual information through RNN, and finally combines the fully connected network layer to integrate the feature information to obtain the final output of the model. CRNNs have also become a common network framework for sound event detection tasks in recent years, and have shown good detection results in practical applications.

2. PROPOSED METHODS

This chapter is divided into four sections to introduce the SED system. Section 2.1 introduces the preprocessing method for audio features; Section 2.2 details the overall architecture of the model; Section 2.3 describes the data enhancement methods used in the SED system and the specific parameter settings; Section 2.4 introduces the pre-training model and the fusion method used in this paper.

2.1 Pre-processing

We resampled each audio clip to a single-channel audio waveform with a sampling rate of 16kHz. The audio waveform was then windowed with a Hamming window of length 2048 points and a step size of 256 points, and transformed into a spectrogram using Short-Time Fourier Transform (STFT). To better represent the frequency energy, we converted the raw audio into a logarithmic mel-spectrogram using 128 logarithmic mel filters. The resulting mel-spectrogram had a size of 626 × 128.

2.2 Network architecture

The three proposed model structures all follow the CRNN architecture.

Model 1: The CNN part consists of three stem blocks and four residual convolutional blocks. The stem blocks include two convolutional layers with a kernel size of 3x3 and a stride of 1x1, followed by BN layer, GLU, Dropout, and pooling layer. The residual convolutional blocks are inspired by ResNet [7], where each residual block consists of two convolutional layers followed by BN layer and ReLU, and a shortcut connection with a 3x3 convolutional kernel.

Model 2: We replaced the stem blocks of Model 1 with a multi-scale convolutional block, which includes two layers of multi-scale convolutional layers. The first layer has multi-scale kernels of size [3,3] and [5,5], and the second layer has multi-scale kernels of size [3,3] and [5,3]. A shortcut connection is added to the multi-scale convolutional block. Additionally, we extracted channel attention weight information from the preprocessed feature maps, assigning different attention to each channel, and combined the features of each channel using a 1x1 convolutional layer.

Model 3: Inspired by the Inception model, we redesigned the size of the multi-scale convolutional kernels and expanded the multi-scale branches to four.

We applied the CA [8] attention mechanism to the output of all modules except the first layer, which captures cross-channel, direction-sensitive, and position-sensitive information, helping our model to more accurately locate and recognize audio events of interest. The RNN part consists of 2 layers of 128 bidirectional gated recurrent units. Dense blocks and attention blocks are added to predict strong and weak labels, respectively. Furthermore, we designed a probability-based weighted pooling function, using the frame-level prediction probability as the aggregation weight for weak labels. This method does not introduce any training parameters and has a strong interpretability.

2.3 Data Augmentation

During the training process, data augmentation strategies including mixup [9], frameshift [10], and FilterAugment [11] were employed. Mixup randomly selects two samples-label pairs to generate new data for improving model generalization. Frameshift moves features and labels along the time axis, and FilterAugment applies random weights to different frequency bands of the Mel spectrogram by randomly dividing the frequency range into several sub-bands, which helps train SED models to recognize time-frequency patterns from a wider frequency range. The hyperparameter settings used in this work are as follows: for mixup the $\alpha = \beta = 0.2$, for the FilterAugment with step type, the dB range is -4.5~6, the number of bands is 2~5, and the minimum bandwidth is 4.

2.4 Pretrained Model

The introduction of pre-trained models can greatly improve system performance. The BEATs[12] model achieved state-of-the-art scores in the Audioset classification task. In this system, we fused the frame-level embedding with the CRNN model. Since the sequence length of the extracted frame-level features is different from that of the CNN features, adaptive average pooling is used to unify the sequence length. Finally, they are fed into an RNN + MLP classifier.

3. EXPERIMENT

3.1. Experimental Settings

In the experiment, we used the dataset provided by the DCASE official. For the SED system, we set the number of filters for each layer to [16, 32, 64, 128, 128, 128, 128] and the pooling layer to [[2, 2], [2, 2], [1, 2], [1, 2], [1, 2], [1, 2], [1, 2]]. We trained the entire system for 200 epochs, using the Adam optimizer with α=0.001 and β=(0.9, 0.999), gradually increasing the learning rate in the first 50 epochs, and setting the batch size to 48.

3.2. Evaluation metric

In order to evaluate the performance of the SED system in different scenarios, we set the PSDS1 and PSDS2 parameters. In PSDS1, the system needs to respond quickly to event detection, so the time localization of sound events is important. In PSDS2, the system must avoid confusion between classes, but the requirement for event response time is less strict than PSDS1, for more details please refer to [12].

3.3. Experimental results

We evaluated the impact of data enhancement strategies, pretrained models, and model integration on the system. The best PSDS1 and PSDS2 scores reached 0.525 and 0.783. The experimental results are shown in Table 1.

Table 1: Results of different Systems

System	Data Aug	Pretrained model	Ensemble	PSDS1	PSDS2
1	√			0.429	0.644
2	√	√	√	**0.525**	0.780
3	√	√	√	0.521	**0.783**

4. CONCLUSION

This technical report introduces the methods used in the 2023 DCASE Task4A challenge. We designed three deep learning models for the sound event detection task, and all models used data augmentation methods to improve the model's generalization ability. We added a multi-scale module to the residual convolutional neural network to extract audio event features of different scales, and the use of the CA attention mechanism can accurately locate and identify the target audio events. In addition, we further improved the model performance by adding a pre-trained model. The final model obtained the best PSDS1 of 0.525 and PSDS2 of 0.783 on the validation set.

5. REFERENCES

[1] A. Tarvainen and H. Valpola, “Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results,” in Proceedings of the 31st International Conference on Neural Information Processing Systems, ser. NIPS’17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 1195–1204.

[2] Debes C, Merentitis A, Sukhanov S, et al. Monitoring activities of daily living in smart homes: Understanding human behavior[J]. IEEE Signal Processing Magazine, 2016, 33(2): 81-94.

[3] Zigel Y, Litvak D, Gannot I. A method for automatic fall detection of elderly people using floor vibrations and sound—Proof of concept on human mimicking doll falls[J]. IEEE transactions on biomedical engineering, 2009, 56(12): 2858-2867.

[4] Wold E, Blum T, Keislar D, et al. Content-based classification, search, and retrieval of audio[J]. IEEE multimedia, 1996, 3(3): 27-36.

[5] Bello J P, Silva C, Nov O, et al. SONYC: A system for the

monitoring, analysis and mitigation of urban noise pollution[J]. ar Xiv, preprint ar Xiv:1805.00889, 2018.

[6] Radhakrishnan R, Divakaran A, Smaragdis A. Audio analysis for surveillance applications[C]// IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2005: 158-161.

[7] He K , Zhang X , Ren S , et al. Deep Residual Learning for Image Recognition[J]. IEEE, 2016.

[8] Hou Q, Zhou D , Feng J . Coordinate Attention for Efficient Mobile Network Design[J]. 2021.

[9] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, “mixup: Beyond empirical risk minimization,” arXiv preprint arXiv:1710.09412, 2017.

[10] L. Delphin-Poulat and C. Plapous, “Mean teacher with data augmentation for dcase 2019 task 4 technical report,” 2019.

[11] H. Nam, B.-Y. Ko, G.-T. Lee, S.-H. Kim, W.-H. Jung, S.-M. Choi, and Y.-H. Park, “Heavily augmented sound event detection utilizing weak predictions,” arXiv preprint arXiv:2107.03649, 2021.

[12] https://dcase.community/challenge2023