SEMI-SUPERVISED SOUND EVENT DETECTION BASED ON PRETRAINED MODELS FOR DCASE 2023 TASK 4A

Technical Report

Yanggang Gan, Ziling Qiao, Juan Wu, Xichang Cai, Menglong Wu*

North China University of Technology
Electronic and Communication Engineering
Beijing, China
ganomen@163.com
caixc_ip@126.com

ABSTRACT

In this technical report, we present our submission system for DCASE 2023 Task4A: Sound Event Detection with Weak Labels and Synthetic Soundscapes. The proposed system is based on mean teacher framework of semi-supervised learning,selective kernel multi-scale convolutional network and frequency dynamic convolutional network. We extract the frame embeddings of the pre-trained model BEAT, and use adaptive average pooling to unify the embeddings to a fixed dimension, and finally fuse them with the features extracted by the convolutional layer of the SED model in the channel dimension. Our systems finally achieve the PSDS-scenario1 of 52.1% and PSDS-scenario2 of 82.5% on the validation set.

Index Terms— Sound event detection, semi-supervised learning, selective kernel,frequency dynamic convolutiona,pretrained model

1. INTRODUCTION

This technical report describe our submitted systems for DCASE2023 Task4A : Sound Event Detection with Weak Labels and Synthetic Soundscapes [1]. The goal of this task is to build a sound event detec-tion (SED) system to classify ten different sound events (Speech, Dog, Cat, Alarm/bell/ringing, Dishes, Frying, Blender, Running water, Vacuum cleaner, Electric shaver/toothbrush) and detect their onset and offset in the audio sequence. In order to make full use of unlabeled in-domain data, we use the Mean Teacher semi-supervised learning method[2], which basically follows the baseline architecture [3].

CRNN architecture[1] has previously achieved great performance on SED tasks . It uses convolutional neural network (CNN) to extract effective high-dimensional features, and then sends these features to RNN to effectively establish temporal dependencies in audio sequences. Therefore, we designed a model based on CRNN framework. However, different sound events have different coverage in time domain and frequency domain. Compared with a single receptive field, multiple receptive fields can extract more abundant features. Inspired by inception[4][5][6], we propose multi-scale convolution for CNN to aggregate multiscale information, so as to better extract the features required by the network. Inspired by Selective Kernel Networks [7], we adopt a dynamic receptive field selection mechanism that allows each neuron to adaptively adjust its receptive field size according to multiple scales of input information to better complete SED tasks. Furthermore, we design an SED system using frequency-dynamic convolution to remove translation invariance along the frequency axis.

2. PROPOSED METHODS

2.1. SED model

For the basic SED model, we design two model architectures: SKNet and FDYCRNN.

For SKNet, SK units are used as the building blocks of CNNs. There are 7 layers in the CNN part. The first two layers are constructed by convolution module, which performs convolution, Batch Normalization, GLU activation, dropout and pooling for input features in turn. The last five layers are multi-scale convolution building blocks with SK units, and their architecture is similar to that of Selective Kernel Networks [7]. The difference is that we put the Batch Normalization and ReLU after the branch information fusion, and use 1x1 convolution to combine the features of each channel. Residual connections are used at the last five layers.The model performs convolution on three branches. In layers 3-4, the convolution kernel size on the branches is set to [3,3], [5,5], [7,7] respectively. We adopt the method of factorizing convolutions[5] ,that is,

factorizing a 5×5 and 7×7 convolution into two and three 3×3 convolutions respectively. This can reduce the network parameters and improve the nonlinearity of the network. In layers 5-7, the convolution kernel size on the branch is set to [3,3], [5,3], [7,3]. We use group convolution at the last five layers.The RNN block is composed of 2 layers of 128 bidirectional gated recurrent units. RNN block is followed by dense block (dense layer, sigmoid activation layer) and attention block, which are used to predict strong label and weak label respectively. This is basically similar to the baseline system.

For FDYCRNN, we apply frequency dynamic convolutions[8] on the baseline system and double the network width of the baseline CRNN.

2.2. Feature fusion based on BEATs

In the case that the features extracted by the CNNs layer in the default SKNet and FDCRNN are compatible with the BEATs features, we extract the frame embeddings of BEATs and the features of the CNNs layer to concatenate in the channel dimension, and use adaptive average pooling to unify the sequence length, Then feed the fused features into RNN + MLP classifier.During the training stage, all layers of BEATs are frozen and no parameter updates are performed [9].

3. EXPERIMENT

3.1. Dataset

We trained and evaluated the proposed model on the development data set of DCASE2023 task4A. There are several different data sets in the development set:

Weakly labeled training set: 1578 clips
Unlabeled in domain training set: 14412 clips
Synthetic strongly labeled training set: 10000 clips
Synthetic strongly labeled validation set: 2500 clips
Strongly labeled validation set: 1168 clips
strong-label Audioset dataset: 3470 clips

We use all unlabeled in domain training set, synthetic strongly labeled training sets and partial weakly labeled training sets to train the model. All synthetic strongly labeled validation sets and partial weakly labeled training sets are used as validation sets, and strongly labeled validation sets are used to evaluate the perfor-mance of the model. When training with external data, the strong-label Audioset dataset is added to the training set.

3.2. Experiment setup

The log-mel spectrum is used as the input feature to the SED system. We used different data augmentation methods (including freame shift, mixup, time mask, FilterAugment, frequency mask) and model architectures (SKNet and FDYCRNN) to train the SED system. We trained the whole system for 200 epochs and the learning rate warms up in the first 50 epochs with the initial learning rate of 0.001. The batch size is set to 48 in experiments without external data and 64 in other experiments.

3.3. results and submissions

We evaluate the system using a threshold-independent implementation of PSDS[10].The best system achieves 0.521 for PSDS-scenario1 and 0.825 for PSDS-scenario2 on the validation set (1168 real audio clips), outperforming the results of 0.359 and 0.562 in the baseline system. We submitted 3 systems and the results are shown in Table 1.

Table1: Description for submitted system

System	External data	Pretrained model	Model count	PSDS1	PSDS2
1			1	0.404	0.620
2	✓	✓	6	**0.521**	0.792
3	✓	✓	10	0.497	**0.825**

System 1 is trained without external data and pretrained models. System 2 is ensembled average of 5 models (using pre-trained BEATs Embedding). System 3 ensemble 4 fine-tuned CNN14 CAMs based on System 2 to improve PSDS2.

4. CONCLUSION

In this technical report, we describe our system submission for dcase 2023 challenge task 4A. We mainly applied two types of network architectures (SKNet and FDYCRNN) and adopted external datasets and pre-trained models to improve system performance. The system achieved the best PSDS1 value of 0.521 and the best PSDS2 value of 0.825 on the validation set.

5. REFERENCES

[1] N. Turpault, R. Serizel, A. Parag Shah, and J. 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.[Online]. Available: https://hal.inria.fr/hal-02160855.

[2] A. Tarvainen and H. Valpola, “Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results,” in Advances in Neural Information Processing Systems, vol. 2017-Decem, no. Nips, 2017, pp.1196–1205.

[3] N. Turpault and R. Serizel, “Training Sound Event Detection On A Heterogeneous Dataset,” in DCASE workshop, 2020. [Online].Available: http://arxiv.org/abs/2007.03931

[4] C. Szegedy et al., "Going deeper with convolutions," 2015 IEEE Conference on Computer Vision and Pattern Recogniti-on (CVPR), 2015, pp. 1-9, doi: 10.1109/CVPR.2015.7298594.
[5] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens and Z. Wojna, "Rethinking the Inception Architecture for Computer Visio-n," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2818-2826, doi: 10.1109/CVPR.2016.308.
[6] Szegedy C, Ioffe S, Vanhoucke V, et al. Inception-v4, incep-tion-resnet and the impact of residual connections on learn-ing[C]//Thirty-first AAAI conference on artificial intelligence. 2017.
[7] X. Li, W. Wang, X. Hu and J. Yang, "Selective Kernel Networks," 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 510-519, doi: 10.1109/CVPR.2019.00060.
[8] Nam H, Kim S H, Ko B Y, et al. Frequency dynamic convolution: Frequency-adaptive pattern recognition for sound event detection[J]. arXiv preprint arXiv:2203.15296, 2022.
[9] Chen S, Wu Y, Wang C, et al. BEATs: Audio Pre-Training with Acoustic Tokenizers[J]. arXiv preprint arXiv:2212.09058, 2022.
[10] Janek Ebbers, Reinhold Haeb-Umbach, and Romain Serizel. Threshold independent evaluation of sound event detection scores. In ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 1021–1025. IEEE, 2022.