SEMI-SUPERVISED SOUND EVENT DETECTION BASED ON MEAN TEACHER WITH SELECTIVE KERNEL MULTISCALE CONVOLUTION AND RESIDENT CAM CLASTERING

Technical Report

Ziling Qiao, Yanggang Gan, Juan Wu, Xichang Cai,Menglong Wu,Hongxia Dong,Lin Zhang,Zihan Liu,
North China University of Technology,Beijing,China
caixc20_ncut@126.com,904220894@qq.com

ABSTRACT

In this technical report, we present our submission system for DCASE 2022 Task4: sound event detection in domestic environments. The proposed system is based on mean teacher framework of semi-supervised learning and Selective Kernel Convolution Network. We use Multi-scale convolution to extract more abundant features of sound events. In order to improve the localization ability of the system, we use a dynamically selected attention mechanism called SK unit in CNN, which allows each neuron to adaptively adjust the size of its receptive field according to multiple scales of input information. Our system finally achieves the PSDS-scenario1 of 39.0% and PSDS-scenario2 of 58.50% on the validation set.

In terms of innovative methods, this technical report will provide a technical description of system 2 submitted by the NCUT team. In system 2, the team selected the audio event monitoring method based on grad CAM clustering. This method attempts to use PANNs based migration learning network to generate grad CAM class activation diagram to locate the time point of the event. Finally, the adaptability of several different network models is evaluated, and the models with higher scores and better adaptability are probability fused to obtain the reasoning of events. Finally, the system 2 based on CAM clustering achieved 9.963% PSDS-scenario1 and 69.877% PSDS-scenario2 scores in the development data set.

Index Terms— Sound event detection, semi-supervised learning, mean teacher, selective kernel, Multi-scale Grad-CAM, PANNs, Transfer learning

1. INTRODUCTION

This technical report describe our submitted systems for DCASE2021 Task4 : Sound Event Detection in Domestic Environments [1]. The goal of this task is to build a sound event detection (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 domain data, our submission system uses semi-supervised SED based on mean teacher[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 into recurrent neural network (RNN) to model the temporal dependencies in the audio sequences effectively. 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 multi-scale 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.

In the weakly supervised image-based classification task, the task team usually uses the grad CAM clustering method to generate a visual class activation map to verify whether the network model effectively extracts the features of the classification target. CAM class activation diagram shows CNN's target positioning ability in weak supervision training. From this point of view, we try to use CNN's target location ability to complete the audio event detection task of task4.

2. PROPOSED METHODS OF SYSTEM1

2.1 Network architecture

Our proposed system adopts the CRNN framework and uses the SK unit as the building block of the CNN model. 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.The overall architecture of the network is shown in Fig.1.

2.2 Data Augment

The mixup [8] data augment method is used. It randomly selects two samples from the training set for mixing (excluding unlabeled data) and uses λ sampled from Beta distribution to control the strength of interpolation between two samples. This linear interpolation technique can enhance the data diversity and robustness of the network.

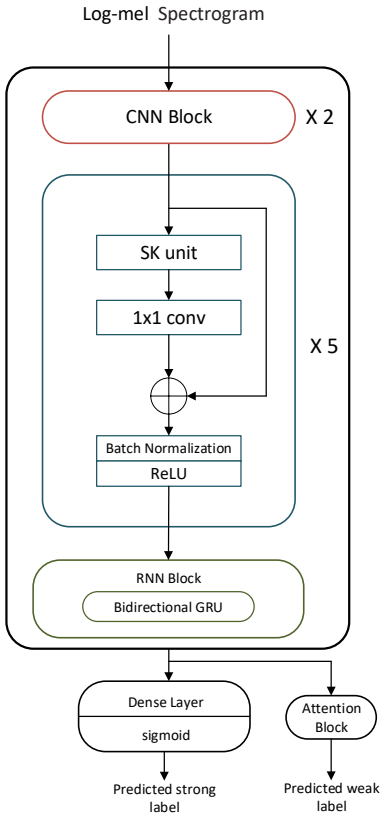

Figure 1: The overall architecture of the network

3. PROPOSED METHODS OF SYSTEM2

The system 2 of this competition adopts PANNs pre training model, uses grad CAM algorithm in CAM generation, generates probability matrix for audio through the probability mapping of CAM class activation diagram, and finally integrates the prediction classification with the highest F1 value of multiple models to generate CAM and infer audio

3.1. Transfer learning of PANNs model

This paper selects CNN14_16k in PANNs 、 CNN14 、 CNN14_16k_nomax three models are the basis of CAM clustering method, among which CNN14_ 16K in the weakly supervised transfer learning of task4 task, the F1 value is the highest of 0.9413, CNN14 and F1 values of model CNN14_ 16k_ nomax are 0.9186 and 0.4316, CNN14 and The F1 value of CNN14_ 16k_nomax model is higher in the reasoning of partial classification. In the process of weakly supervised transfer training, this paper uses the method of freezing part of the convolution layer. Through comparative experiments, different frozen layers and training methods are selected for different models. In order to improve the classification accuracy, cosine annealing learning rate is used to improve the model accuracy in transfer learning.

3.2. Application of Grad-CAM

In this paper, grad Grad-CAM(Gradient-weighted Class Activation Mapping) algorithm is used for strong label mapping of weakly supervised CNN networks. Grad CAM uses the gradient information of the last convolution layer flowing into CNN to assign weights to each neuron for specific attention reasoning. Finally, the event detection sequences are generated for different classifications, and the event detection sequences of the dominant classification of each model are fused to obtain the final reasoning.

4. EXPERIMENT

4.1. Dataset

We trained and evaluated the proposed model on the development data set of dcase2022 task4. 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

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 performance of the model.

4.2. Experimental Settings

The number of filters and pooling size for each layer are respectively [16,32, 64, 128, 128, 128, 128] and [[2, 2], [1, 2], [2, 2], [1, 2], [1,2], [1, 2],[1, 2]]. In the multi-scale convolution layer, the number of filters in each layer mentioned above refers to the number of filters in each branch of the layer. The dropout is set to 0.5. The batch size is set to 48 consisting of 12 synthesized, 12 weakly labeled and 24 unlabeled data.We choose Adam optimizer with learning rate of 0.001 and the learning rate warmup during the first 50 epochs. models are trained with 200 epochs.

4.3. Experimental results of system1

In the Mean Teacher architecture, the output of the teacher model is more likely to be correct. In our experiment, the performance of teacher model is slightly higher than that of student model. Our model finally achieves the PSDS-scenario1 of 0.390 and PSDS-

scenario2 of 0.585 on the strongly labeled validation set, which outperform the results of 0.336 and 0.536 in the baseline system.

4.4. Experimental results of system2

In the experiment of weakly supervised transfer learning, this paper makes an individualized experiment for transfer learning of different classification models. The training scheme is adjusted according to the training effect of each model, and the freezing layers of CNN are changed from shallow to deep to improve the model accuracy. Finally, the F1 value results in the validation set are shown in Table 1. Gradually freezing the model parameters can effectively improve the training speed and accuracy of the model. In this paper, cosine annealing learning rate is used in the training of the last convolution layer. The learning rate of cosine annealing can be reduced smoothly according to the training rounds, and the model with the lowest loss can be obtained in a long time of training.

Table 1: Evaluation for Transfer learning model.

Models	F1
CNN14	0.9172
CNN14_16k	0.9486
CNN14_16k_nomax	0.8745

In this paper, each network is used to generate Grad-CAM and psds2 evaluation is conducted for the generated grad Grad-CAM. Finally, this paper also evaluates the Grad-CAM after probability fusion of multiple networks. The scores are shown in Table 2:

Table 2: Evaluation for Transfer learning model

CAM Generated Models	PSDS2
CNN14	0.6532
CNN14_16k	0.6630
CNN14_16k_nomax	0.6627
CNN14+CNN14_16k	0.6725
CNN14+CNN14_16k_nomax	0.6842
CNN14+CNN14_16k_nomax+CNN14_16k	0.6988

By observing the effect of Grad-CAM, we can observe that the sound event detection method based on Grad-CAM clustering has significantly improved the psds2 score. Therefore, it can be preliminarily judged that the Grad-CAM generated based on CNN14 inherits the event discrimination ability of CNN14, and can distinguish a single event in the complex stacked synthetic audio. Since the calculation of Grad-CAM is based on the inverse operation of CNN, although Grad-CAM can distinguish a single event from the synthetic audio of the event stack, it can not clearly determine the start and end of the event, so the psds1 score is lower than the baseline level.

5. CONCLUSION

In this technical report, the proposed system is a multi-scale convolutional recurrent neural network based on the dynamic receptive field selection mechanism, and is trained using the mean teacher framework based on semi supervised learning. Multi-scale convolution extracts the features of audio events of different scales through convolution kernels of different sizes. The kernel selection mechanism enables neurons to adaptively adjust the receptive field size, that is, carry out "selective kernel" (SK) convolution among multiple kernels of different sizes. We also use data augment technology mixup to improve the robustness of the model. Finally, our model achieved better results than the baseline system.

In system 2. The feedback in the Grad-CAM class activation map is the feature pixels learned by CNN. The target to be detected and the feature pixels are inclusive. However, by observing the CAM class activation map generated by each network, it can be seen that the feature pixels learned by different networks do not necessarily coincide. Therefore, the feature extraction of detection targets is different in different networks. In this paper, we use the complementary relationship between models to perform probabilistic fusion of Grad-CAM sequences generated by multiple models. According to the probability fusion of multi model Grad-CAM shown in Table 2, the psds2 score can be effectively improved.

To sum up, the audio event detection method based on Grad-CAM clustering can effectively improve the psds2 score of the model. It can be used as an auxiliary method to inherit the strong ability of the classification model to distinguish events by means of probability fusion, so as to improve the psds2 score level of any model.

6. 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 I EEE Conference on Computer Vision and Pattern Recognition (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 Vision," *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, inception-resnet and the impact of residual connections on learning[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] Zhang H, Cisse M, Dauphin Y N, et al. mixup: Beyond empirical risk minimization[J]. arXiv preprint arXiv:1710.09412, 2017.