SOUND EVENT DETECTION BASED ON SOFT LABEL

Technical Report

Haiyue Zhang, Liangxiao Zuo, Jingxuan Chen

North China University of Technology, Beijing, China
m13769897798@163.com

Xichang Cai,Menglong Wu*

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

ABSTRACT

This report focuses on an in-depth study of DCASE 2023 Task 4b. In contrast to previous tasks, this task provides a dataset with soft labels, aiming to explore how to improve the performance of the baseline system using soft labels. The report primarily employs two effective enhancement methods. Firstly, to balance the dataset, the report expands the original dataset. Given the significant differences in sound events within the dataset, an augmentation method is employed to generate additional samples and equalize the dataset. Secondly, to further enhance the system performance, a model ensemble approach is utilized. By combining the predictions of multiple models, their individual strengths can be effectively utilized to improve overall performance.

***Index Terms*—**Sound event detection, soft labels, ensemble averaging, residual convolutional recurrent neural network, Selective Kernel.

1 INTRODUCTION

Task 4b of the DCASE (Detection and Classification of Acoustic Scenes and Events) competition presents participants with a sound event detection (SED) problem. This task is a subtheme of the sound event detection task 4 and provides participants with weakly labeled data (without time information), strongly labeled synthetic data (with time information), unlabeled data, and soft-labeled data. The dataset provided contains 17 sound categories, of which only 15 categories have values exceeding 0.5, while the remaining 4 categories are extremely rare. Therefore, evaluation is conducted only on the 11 categories officially specified[1].

This article primarily employs two methods to improve the performance of the baseline system. The first method involves expanding the original dataset to average out the dataset and addresses the issue of imbalanced data caused by significant differences in sound events. The dataset augmentation is utilized to alleviate the problem of data imbalance resulting from the substantial variations in sound events within the dataset. The second method is model ensemble averaging. Three different models, namely Convolutional Recurrent Neural Network (CRNN)[4], Residual Convolutional Recurrent Neural Network (RCRNN)[3][5], and Selective Kernel Residual Convolutional Recurrent Neural Network (SK-RCRNN) [5]are selected for the ensemble. CRNN serves as the baseline system, RCRNN introduces residual blocks on top of CRNN, and SK-RCRNN further incorporates Selective Kernel (SK) units[6] on top of RCRNN. The article conducts separate experiments for each model and performs ensemble averaging.

This report explores the individual performance of each model through separate experiments and combines their predictions through ensemble averaging to enhance the overall system performance.

2 METHOD

2.1 Dataset

This article uses the MAESTRO Real data set provided by DCASE officials. The dataset was created for studying the estimation of strong labels using crowdsourcing. It contains 49 real-life audio files from 5 different acoustic scenes and the annotation outcome. Annotation was performed using Amazon Mechanical Turk. The total duration of the dataset is 189 minutes and 52 seconds

Audio files are a subset of the TUT Acoustic Scenes 2016 dataset, belonging to five acoustic scenes: cafe/restaurant, city center, grocery store, metro station, and residential area. Each scene has 6 classes, some of which are common to all the scenes, resulting in 17 classes in total. The dataset contains audio: the 49 real-life recordings, each from 3 to 5 min long. soft labels: estimated strong labels from the crowdsourced data, values between 0 and 1 indicates the uncertainty of the annotators.

2.2 Data Processing

2.2.1 Balanced Dataset

The method used in this article involves balancing the dataset by matching the sample quantities of the minority classes to the sample quantity of the majority class, which has a larger number of samples. This is achieved by generating additional samples for the classes with fewer samples, aiming to create a balanced dataset.

Since the MAESTRO Real dataset has a small overall size and highly imbalanced class distributions, an oversampling technique is employed to replicate the samples of the minority classes. In this approach, the class with the highest number of samples in each folder is chosen as the reference, and the samples from the soft labels dataset are duplicated for the specific class where the probability is greater than 0.5.

2.2.2 Mixup

The principle of Mixup can be summarized as blending two samples together in a certain proportion to create a new sample[2].

$$x = \lambda x_i + (1-\lambda)\ x_j$$

$$y = \lambda y_i + (1-\lambda)\ \mathrm{y_j}$$

The λ is done according to the beta distribution, which is governed by the parameter $\alpha \in (0, \infty)$ representing the interpolation strength between target features.

In this article, the direct usage of replicated data for training may lead to severe overfitting. Therefore, Mixup is applied only to the log-mel frequency features of the newly replicated data. By blending the features of these replicated samples, the model can learn from a diverse range of synthesized examples while mitigating the risk of overfitting.

2.3 Model

The baseline system is a sequence recognition system based on the CRNN (Convolutional Recurrent Neural Network) model[[10]. Its architecture primarily consists of three convolutional neural network (CNN) layers and one bidirectional gated recurrent unit (GRU) layer. Finally, the results are obtained through a linear output layer.

2.3.1 RCRNN

In this paper, we introduce residual modules on top of the baseline CRNN (Convolutional Recurrent Neural Network) model provided by the competition organizers, creating the RCRNN (Residual Convolutional Recurrent Neural Network) model. We add three residual blocks to the CNN part of the baseline system while keeping the RNN part mostly unchanged.

These residual blocks are inserted between the convolutional layers of the CRNN model. Each residual block consists of multiple convolutional layers. The input of each residual block is directly passed to the output of the block through shortcut connections and then added to the output of the convolutional layers within the block. This way, the RCRNN model can better utilize the input information and facilitate gradient flow through the residual connections, making the network easier to train and optimize.

By introducing residual modules into the CRNN model[7], the RCRNN model improves feature representation, enhances model performance, and increases learning efficiency while preserving the overall structure of the original model.

2.3.2 SK-RCRNN

The Selective Kernel (SK) network is a dynamic selection mechanism widely used in Convolutional Neural Networks (CNNs). In this paper, we introduce the Selective Kernel (SK) unit into the previously proposed RCRNN model[9], resulting in the SK-RCRNN model. The SK-RCRNN model uses the Selective Kernel unit to replace the CNN in the RCRNN model. The Selective Kernel unit allows each neuron to adaptively select the receptive field size and obtain more comprehensive feature information through convolution operations at multiple scales.

This design enables the SK-RCRNN model to better adapt to sound event targets of different scales and improve the model's perception and classification ability for sound features. In this way, the SK-RCRNN model can more effectively capture the frequency domain information of audio signals in the feature extraction stage, further improving the model's performance and learning ability.

2.3.3 Model Ensemble

Model ensemble is the process of combining multiple trained models, where test data is predicted by each model and the learning capabilities of each model are integrated in some way to improve the overall generalization ability of the final model. In this paper, a simple averaging ensemble method is used, where several independent base models are constructed, including RCRNN, SK-RCRNN, CA-RCRNN, and CRNN. Each model independently predicts the input samples and generates a prediction result. Then, these prediction results are averaged to compute the average probability for each class across all the base models. Experimental results demonstrate that the optimal model ensemble combination consists of the SK-RCRNN model, RCRNN, and CRNN models.

3 EXPERIMENTS

3.1 Feature Extraction

To address the issue of limited data for certain classes, we replicate the samples of these classes and apply mixup. According to the experiments conducted by Moustapha Cisse et al., we set the value of α to 0.2 in mixup. As shown in Table 1.

Table 1 Results with mixup incorporated

Methods	ER_m	F1_m	F1_M	F1_MO
Baseline	0.479	71.54	35.21	44.13
Baseline+Mixup	**0.439**	**74.84**	**39.57**	**43.5**

It can be observed that mixup reduces the Micro segment-based ER (Error Rate) and improves the Micro segment-based F1 score and Macro segment-based F1 score. However, there is minimal improvement in macro-average F1 score with optimum threshold per class.

3.2 Experimental result

The basic system of this article consists solely of the CRNN network and does not utilize external data for training.

（1）Without using the balanced dataset method.

several independent basic models were constructed, including RCRNN, SK-RCRNN, and CA-RCRNN[8]. Each model was tested separately, generating prediction results as shown in Table 2.

Table 2 Independent Predicted Results

Methods	ER_m	F1_m	F1_M
Baseline	0.479	71.54	35.21
3ResNet+1GRU	0.477	69.64	29.79
3CNN+3ResNet+1GRU	0.473	69.75	28.93
3CNN+6ResNet+1GRU	0.514	67.07	26.36
3CNN+3ResNet+2GRU	0.471	69.26	26.76

The average ensemble model was tested, and the prediction results are shown in Table 3.

Table 3 Ensemble Predicted Results

Methods	ER_m	F1_m	F1_M	F1_MO
Baseline	0.479	71.54	35.21	44.13
3CRNN	0.469	71.66	37.29	44.106
SK-RCRNN+CRNN	0.459	70.87	33.66	43.01
SK-RCRNN+2CRNN	0.48	71.42	36.8	42.29
SK-RCRNN+2RCRNN	0.478	71.23	36.58	43.02
2SK-RCRNN+RCRNN	0.502	68.63	31.0	43.28

（2）In the case of using the balanced dataset method, the experimental results are shown in Table 4.

Table 4 Prediction results after balancing the dataset

	Methods	ER_m	F1_m	F1_M	F1_MO
	Baseline	0.479	71.54	35.21	44.13
1	Baseline+Mixup	**0.439**	**74.84**	**39.57**	**43.5**
2	SK-RCRNN+2CRNN	**0.443**	**73.38**	**35.6**	**44.49**
3	SKRCRNN+CRNN+RCRNN	**0.432**	**73.89**	**34.86**	**44.47**

The three systems we submitted are shown in Table 4, the integration model has improved compared with the baseline system performance, especially the last integration model, ER_m reached 0.432, F1_m reached 34.86, F1_M reached 34.86, and F1_mo reached 44.47.

4 CONCLUSIONS

In this technical report, we describe our system submission for dcase 2023 challenge task 4b. This article mainly proposes two methods. Firstly, to balance the dataset, the original dataset was augmented by generating additional samples to even out the data distribution. Secondly, to further enhance system performance, we employed a model ensemble approach. By combining the prediction results of multiple models, we leveraged their individual strengths to improve overall performance, resulting in an enhanced prediction outcome.

5 REFERENCES

[1] I. Martín-Morató, M. Harju, P. Ahokas and A. Mesaros, "Training Sound Event Detection with Soft Labels from Crowd sourced Annotations," ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 2023, pp. 1-5, doi: 10.1109/ICASSP49357.2023.10095504.

[2] Zhang H, Cisse M, Dauphin Y N, et al. Mixup: Beyond empirical risk minimization[J].arXiv preprint arXiv:1710.09412, 2017.

[3] N. K. Kim and H. K. Kim, "Polyphonic Sound Event Detection Based on Residual Convolutional Recurrent Neural Network With Semi-Supervised Loss Function," in IEEE Access, vol.9,pp.75647575,2021,doi10.1109/ACCESS.2020.3048675.

[4] B. Shi, X. Bai, and C. Yao, "An End-to-End Trainable Neural Network for Image-Based Sequence Recognition and Its Application to Scene Text Recognition," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 11, pp.22982304,1Nov.2017,doi:10.1109/TPAMI.2016.2646371.

[5] Shafiq, M.; Gu, Z. Deep Residual Learning for Image Recogntion:ASuvey.Appl.Sci.2022,12,8972.https://doi.org/10.3390/app12188972.

[6] 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.

[7] C. Szegedy et al., "Going deeper with convolutions," 2015 IEEE Conference on Computer Vision and Pattern Recognition(CVPR), 2015, pp. 1-9, doi: 10.1109/CVPR.2015.7298594.

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

[9] 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.

[10] E.Çakir, G.Parascandolo, T. Heittola, H. Huttunen, and T. Virtanen, “Convolutional recurrent neural net works for polyphonic sound event detection,” IEEE/ACM Transactions on Audio, Speech, and language processing, vol.25, pp.1291–1303,2017.