HOW INFORMATION ON SOFT LABELS AND HARD LABELS MUTUALLY BENEFITS SOUND EVENT DETECTION TASKS

Technical Report

Han Yin[1]*, Jisheng Bai*[1,2]*, Siwei Huang*[1]*, Jianfeng Chen*[1,2]

[1]Joint Laboratory of Environmental Sound Sensing,
School of Marine Science and Technology,
Northwestern Polytechnical University,Xi'an,China
[2]LianFeng Acoustic Technologies Co., Ltd. Xi'an China
{yinhan, baijs, hsw838866721}@mail.nwpu.edu.cn, chenjf@nwpu.edu.cn

ABSTRACT

This technical report describes our submission to DCASE 2023 Task 4B: Sound event detection (SED) with soft labels. The main purpose of Task 4B is to study how to effectively use the information of soft labels and hard labels in SED tasks. We propose different architectures to explore how both soft and hard labels can jointly improve the performance of SED. Our systems are built upon the Convolutional Recurrent Neural Network (CRNN) proposed by the baseline and the Conformer structure. Extensive experiments are performed to compare the performance of different systems on SED tasks. Results show that our best proposed system outperforms the baseline by 11.74 in F1 score.

***Index Terms*—** DCASE 2023, SED, CRNN, Conformer, Soft labels

1. INTRODUCTION

The aim of DCASE 2023 Task4B is the detection and classification of 11 different sound event categories. These sound categories are very common in real-life scenarios. The target is to provide not only the event class but also the event time localization in a real-recorded audio.

SED is a challenging task because different sound events often have extremely variable acoustic properties. In addition, aliasing of different types of voices can have serious negative impacts. Therefore, ordinary mathematical modeling or machine learning methods often can not achieve excellent performance. In contrast, modern pattern classification tools, especially CRNNs, can perform SED tasks more easily[1, 2, 3]. In recent years, SED has been solved as a supervised learning task, which means each piece of audio needs to have a frame-level sound event label. Labels that contain only 0 and 1 values are referred as hard labels. A value of 0 means that the event does not occur while 1 means the exact opposite.

Hard labels usually have low error tolerance and require a lot of manual annotations. While synthetic strongly-labeled data is easy to create, often these simulated data lack the complexity and variability as real data. Therefore, more available weak labels are gradually used in various sound event detection tasks[4, 5]. However, they are far less effective than strongly-labeled data. To solve the above problems, Irene Martín and Annamaria Mesaros created a soft label for training the SED system[6]. Similar to hard labels, soft labels contain both the time location and occurrence probability of various sound events, but the probability value is between 0 and 1.

The main purpose of DCASE 2023 TASK4B is to explore how soft labels can improve the performance of sound event detection. To address this problem, firstly, we perform data augmentation using temporal Mixup. Then, based on the baseline system[7], we propose a one-branch SED system and four two-branch SED systems that can be trained using soft labels. After comparing their performance, we incorporate an attention block into the best two-branch system for information fusion, which further improves the performance of SED.

The rest of this paper is structured as follows: Section II gives a detailed description of proposed methods. Then, Section III illustrates the experimental setup. Section IV presents and discusses the experimental results. Finally Section V presents our conclusions.

2. PROPOSED METHODS

In this section, we illustrate the temporal Mixup approach for data augmentation and architectures of our proposed systems.

2.1. Temporal Mixup

Because the audio data provided by the challenge is too little, we use Mixup in the time domain for data augmentation. Assuming that $X_1 \in \mathbb{R}^{C\times T}$ and $X_2 \in \mathbb{R}^{C\times T}$ are two pieces of audio in the same acoustic scene respectively, we mix them in the following way to generate a new audio clip:

$$X_{new} = \epsilon X_1 + (1-\epsilon)X_2 \tag{1}$$

where $\epsilon \in [0, 1]$ is the hyperparameter. At the same time, we will also mix the labels corresponding to X_1 and X_2, which can be represented as $Y_1 \in \mathbb{R}^{1\times N_1}$ and $Y_2 \in \mathbb{R}^{1\times N_2}$ respectively.

$$Y_{new} = \epsilon Y_1 + (1-\epsilon)Y_2 \tag{2}$$

2.2. One-branch SED system

Based on the baseline system mainly composed of three CNN blocks and a layer of bidirectional gated recurrent unit (GRU)[8], we propose a one-branch SED system. Actually, we just replace the GRU in the baseline with Conformer blocks. The Conformer

block contains two Feed Forward modules sandwiching the Multi-Headed Self-Attention module and the Convolution module[9, 10]. Fig 1 depicts the architecture of the CNN Block and our proposed one-branch SED system.

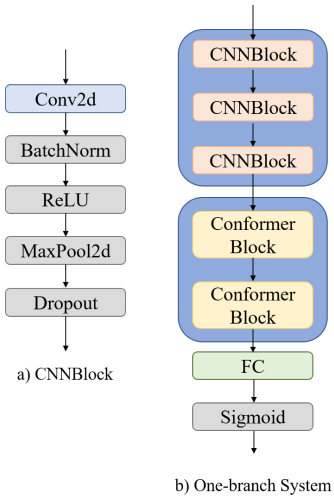

Figure 1: The architecture of proposed one-branch SED system

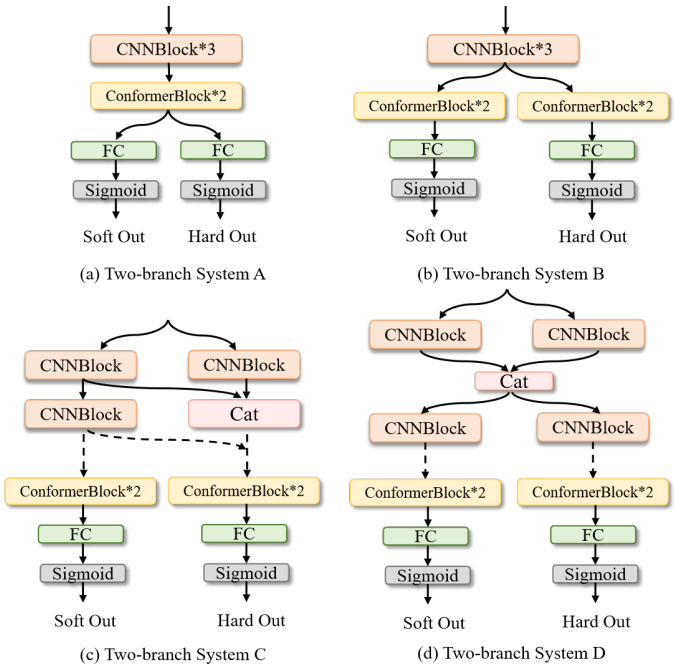

Figure 2: The architecture of proposed two-branch SED systems

2.3. Two-branch SED systems

We propose four two-branch network structures that can jointly use soft-label and hard-label information. As shown in Fig 2, in the two-branch system A, two branches in the network share the same CNN and Conformer parameters, and finally use different fully connected layers to output soft and hard results respectively. Similarly, only CNN parameters are shared in the two-branch system B. In the two-branch system C, embeddings in the soft branch are sent to the hard branch. Finally, in the two-branch system D, feature maps of the soft branch and the hard branch are continuously fused with each other by concatenation.

2.4. Two-branch SED system with Attention

All the above two-branch systems output soft results and hard results separately. During the experiment, we find that the weighted average of the results output by the soft and hard branches is the best. If an attention fusion module can be designed to fuse the outputs of the two branches, the performance of SED may be further improved. Therefore, we improve the two-branch system D by adding an attention mechanism. Fig 3 shows the architecture of our proposed two-branch system D+.

The attention module performs weighted fusion of the output results of the soft and hard branches. Assuming that the outputs of the two branches are $\hat{S} \in \mathbb{R}^{N\times F\times K}$ and $\hat{H} \in \mathbb{R}^{N\times F\times K}$ respectively, where N represents frames, F means frequency bins and K is the number of events. Denote the output of the attention module as $\hat{Y} \in \mathbb{R}^{N\times F\times K}$, we weight them in three ways:

a) Hardness level

In this approach, we directly perform an overall weighted fusion of $\hat{S}$ and $\hat{H}$:

$$\hat{Y} = \alpha\hat{S} + (1-\alpha)\hat{H} \tag{3}$$

where the scalar $\alpha \in (0, 1)$ is a learnable parameter.

b) Event level

In this approach, we perform weighted fusion of $\hat{S}$ and $\hat{H}$ from the dimension of events.

$$\hat{Y}[:,:,i] = \boldsymbol{\beta}(i)\hat{S}[:,:,i] + [1-\boldsymbol{\beta}(i)]\hat{H}[:,:,i] \tag{4}$$

where the vector $\boldsymbol{\beta} \in \mathbb{R}^K$ is a learnable parameter, and $i = 1, 2, ..., K$. Similarly, all elements in $\boldsymbol{\beta}$ are limited to ranges between 0 and 1.

c) Frame level

In this approach, we perform weighted fusion of $\hat{S}$ and $\hat{H}$ from the dimension of time frame.

$$\hat{Y}[:,i,:] = \boldsymbol{\gamma}(i)\hat{S}[:,i,:] + [1-\boldsymbol{\gamma}(i)]\hat{H}[:,i,:] \tag{5}$$

where the vector $\boldsymbol{\gamma} \in \mathbb{R}^N$ is a learnable parameter, and $i = 1, 2, ..., N$. All elements in $\boldsymbol{\gamma}$ are also limited to ranges between 0 and 1 as well.

3. EXPERIMENTS SETUP

3.1. Dataset

The development set provided for this task is MAESTRO Real[6]. The dataset consists of real-life recordings with a length of approximately 3 minutes each, recorded in a few different acoustic scenes. The audio was annotated using Amazon Mechanical Turk, with a procedure that allows estimating soft labels from multiple annotator opinions.

All audio comes from five sound scenes including cafe restaurant, city center, grocery store, metro station and residential area. The organizing committee not only provides soft labels, but also hard labels. Among them, the soft label contains a total of 17 sound events: birds singing, car, people talking, footsteps, children voices,

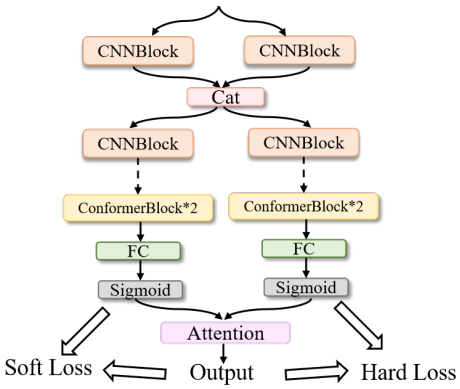

Figure 3: The architecture of proposed two-branch SED system D+

wind blowing, brakes squeaking, large vehicle, cutlery and dishes, metro approaching, metro leaving, furniture dragging, coffee machine, door opens/closes, announcement, shopping cart and cash register beeping. But only 15 classes have values above 0.5, and 4 of them are very rare. Therefore, only the first 11 categories are evaluated. Correspondingly, the hard label only contains the first 11 types of sound events. All audio is sampled at 44.1kHz and the total duration is approximately 150 minutes.

3.2. Temporal Mixup

Since the total audio duration is too short and the number of classes is extremely unbalanced, we use Mixup in the time domain for data augmentation. Firstly, we select the events that need to be increased. Then, Mixup is performed on clips containing these sound events in all audios of the same sound scene, where factor ϵ takes a random value between 0 and 1. Fig 4 shows the statistical data before and after Mixup.

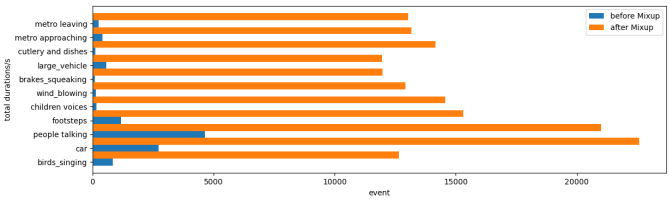

Figure 4: Statistical results of data before and after temporal Mixup

It can be seen that all kinds of data are basically balanced after Mixup, but cars and people talking are still significantly more than other events, because there is a lot of aliasing in the audio.

3.3. Evaluation Metrics

System evaluation will be based on the following metrics, calculated in 1s-segments:

- micro-average F1 score $F1_m$, with a decision threshold of 0.5 applied to the system output provided by participants.
- micro-average error rate ER_m, with a decision threshold of 0.5 applied to the system output provided by participants.
- macro-average F1 score $F1_M$, with a decision threshold of 0.5 applied to the system output provided by participants.
- macro-average F1 score with optimum threshold per class $F1_{MO}$, based on the best F1 score per class obtained with a class-specific threshold.

Ranking of the systems will be done based on $F1_{MO}$.

3.4. Training Setup

We use a hopsize of 11025 and a window size of 22050 to perform log-mel feature extraction on all audio as the input of the system.

The dataset is provided with a 5-fold cross-validation setup in which approximately 70% of the data (per class) is used in training, and the rest is used for testing. We use the Adam optimizer, and the initial learning rate is set to 0.0003. When the test loss does not decrease for 10 consecutive epochs, the learning rate will automatically decrease by half.

3.5. Post-processing

We propose a masking method for the output of the model. In fact, certain sound events occur in some acoustic scenes with a low probability. For example, there are no birds singing or brakes squeaking in a cafe restaurant. Based on this prior information, we design a specific mask for each acoustic scene, and by element-wise multiplying it with the model output, the value of certain sound events in the output can be made very low.

4. RESULTS AND DISCUSSION

Experimental results of our proposed one-branch system, two-branch systems A/B/C/D and the baseline are presented in Table 1.

Table 1: Results of proposed one-branch system and two-branch systems A/B/C/D (evaluated only on first 11 types of sound events)

System	Label Used	Loss	$F1_{MO}$	$F1_{MO}$(masked)
Baseline	Soft	MSE	44.13	-
One-Branch	Soft	MSE	44.25	45.83
One-Branch	Hard	BCE	43.32	44.18
One-Branch	Soft and Hard	MSE and BCE	51.09	51.79
One-Branch	Soft and Hard	MSE and MSE	**52.25**	**53.28**
Two-Brach A	Soft and Hard	MSE and MSE	47.15	48.53
Two-Brach B	Soft and Hard	MSE and MSE	48.50	49.44
Two-Brach C	Soft and Hard	MSE and MSE	49.75	50.52
Two-Brach D	Soft and Hard	MSE and MSE	51.72	52.92

From the experimental results, it can be seen that when using both soft labels and hard labels to calculate the loss, SED performs better than using either one alone. And MSE loss is more effective than BCE loss when soft labels are used. In addition, two-branch systems are not as good as the one-branch system, which we suspect is due to the insufficient amount of data to allow the two branches to learn enough information respectively.

Table 2 shows the results of the best one-branch system, two-branch system D and the two-branch system D+. It is found that the performance of SED can be improved by weighted fusion of the outputs of the two branches.

Table 2: Results of the best one-branch system, two-branch system D and the two-branch system D+ (evaluated only on first 11 types of sound events)

System	Label Used	Attention	$F1_{MO}$	$F1_{MO}$(masked)
Baseline	Soft	-	44.13	-
One-Branch	Soft and Hard	-	52.25	53.28
Two-Brach D	Soft and Hard	-	51.72	52.92
Two-Brach D+	Soft and Hard	Hardness Level	50.47	51.67
Two-Brach D+	Soft and Hard	Frame Level	51.41	52.21
Two-Brach D+	Soft and Hard	Event Level	**54.3**	**55.87**

5. CONCLUSION

In this paper, we propose a one-branch system and four two-branch systems to explore how to use soft labels to improve the performance of SED. We further improved the two-branch system D through different attention mechanisms, and obtained the best model D+. Furthermore, we propose a masking approach based on sound scene prior information to post-process the results. The experimental results show that soft labels can provide richer information for the SED system, thereby improving the accuracy of the results. Scene-based masks can also improve the performance of SED. In the future, we will try to replace Concatenate in the two-branch system with other mechanisms for information fusion. At the same time, we will consider automatically learning the mask tensor of the output in the network instead of the fixed mask based on prior information.

6. REFERENCES

[1] A. Mesaros, T. Heittola, T. Virtanen, and M. D. Plumbley, "Sound event detection: A tutorial," *IEEE Signal Processing Magazine*, vol. 38, no. 5, pp. 67–83, 2021.

[2] E. Cakir, T. Heittola, H. Huttunen, and T. Virtanen, "Polyphonic sound event detection using multi label deep neural networks," in *2015 international joint conference on neural networks (IJCNN)*. IEEE, 2015, pp. 1–7.

[3] R. Lu and Z. Duan, "Bidirectional gru for sound event detection," *Detection and Classification of Acoustic Scenes and Events*, pp. 1–3, 2017.

[4] Y. Wang, J. Li, and F. Metze, "A comparison of five multiple instance learning pooling functions for sound event detection with weak labeling," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 31–35.

[5] B. McFee, J. Salamon, and J. P. Bello, "Adaptive pooling operators for weakly labeled sound event detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 11, pp. 2180–2193, 2018.

[6] I. Martín-Morató and A. Mesaros, "Strong labeling of sound events using crowdsourced weak labels and annotator competence estimation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

[7] I. Martín-Morató, M. Harju, P. Ahokas, and A. Mesaros, "Training sound event detection with soft labels from crowdsourced annotations," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[8] R. Dey and F. M. Salem, "Gate-variants of gated recurrent unit (gru) neural networks," in *2017 IEEE 60th international midwest symposium on circuits and systems (MWSCAS)*. IEEE, 2017, pp. 1597–1600.

[9] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," *arXiv preprint arXiv:2005.08100*, 2020.

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.