A MULTI-TASK LEARNING METHOD FOR WEAKLY SUPERVISED SOUND EVENT DETECTION

Sichen Liu[1,2], *Feiran Yang*[1,2], *Fang Kang*[1,2], *Jun Yang*[1,2]

[1] Institute of Acoustics, Chinese Academy of Sciences, Beijing, China
[2] University of Chinese Academy of Sciences, Beijing, China

ABSTRACT

In weakly supervised sound event detection (SED), only coarse-grained labels are available, and thus the supervision information is quite limited. To fully utilize prior knowledge of the time-frequency masks of each sound event, we propose a novel multi-task learning (MTL) method that takes SED as the main task and source separation as the auxiliary task. For active events, we minimize the overlap of their masks as the segment loss to learn distinguishing features. For inactive events, the proposed method measures the activity of masks as silent loss to reduce the insertion error. The auxiliary source separation task calculates an extra penalty according to the shared masks, which can further incorporate prior knowledge in the form of regularization constraints. We demonstrated that the proposed method can effectively reduce the insertion error and achieve a better performance in SED task than single-task methods.

Index Terms— Sound event detection (SED), source separation (SS), multi-task learning (MTL), weakly supervised

1. INTRODUCTION

As an important research topic in auditory perception, sound event detection (SED) has many potential applications such as healthcare in smart home [1, 2], surveillance monitor in public area [3, 4] and large-scale information retrieval [5]. SED aims to predict the onset and offset times of the target sound event while audio tagging (AT) is only concerned about the event type. Since the strongly annotated data is costly to acquire, many recent researches have focused on the weakly supervised SED (WSSED) method which only utilizes the coarse-grained clip-level labels [6, 7, 8, 9, 10]. However, one challenge of the WSSED is to deal with the overlapped sound events which happen frequently in real applications. For example, in a home environment, the sound of the vacuum cleaner, the ringing of the telephone and the baby crying have a great probability to be overlapped. The overlapped events may decrease the signal-to-noise ratio (SNR) of target events and lead to a poor SED performance.

Source separation task aims to separate the mixed signals into statistically independent sources, which is also closely related to the auditory perception topic [11]. For SED task, the interference of overlapping events can be reduced by separating sources first. For source separation task, the event type provided by the SED system can further introduce the spectral distribution information. The goal of this paper is to jointly handle SED and source separation tasks with the weakly labeled data in a multi-task learning (MTL) framework.

Several works have attempted to address the SED problem using MTL methods [12, 13]. Kong et al. propose a joint separation-classification model which shares the same segmentation mapping part. However, only the categories information contained in labels is utilized during the training process [14]. Pishdadian et al. try to separate sources with the weakly labeled data. They add event detectors after the separation model and limit the sum of all the separated signals equal to the mix signal [15].

Most of previous studies utilize the cross-entropy loss to train the SED model. The potential information such as the activities of events contained in the estimated time-frequency (T-F) masks has not been well considered. In this paper, we take source separation as the auxiliary task of SED to utilize the information of masks. Firstly, the segmentation part of SED model is reused by source separation task through a hard-sharing mechanism to obtain the estimated T-F masks. Then, we impose constraints on masks in a regularization form in order to jointly utilize the cross entropy loss from SED and the auxiliary loss from source separation. The extra introduced losses can effectively promote the use of masks. Extensive experiments are conducted to verify the performance of the proposed MTL method.

2. PROPOSED METHOD

Strongly labeled data for SED contains the type of events and its corresponding precise onset and offset, while the weakly labeled data only contains the events types. However, for the source separation task, the data which contains information on each source at the granularity of a T-F bin is referred to be strongly labeled. Thus, the weakly labeled SED data utilized in this paper is much weaker for the source separation task, which is actually a more challenging task.

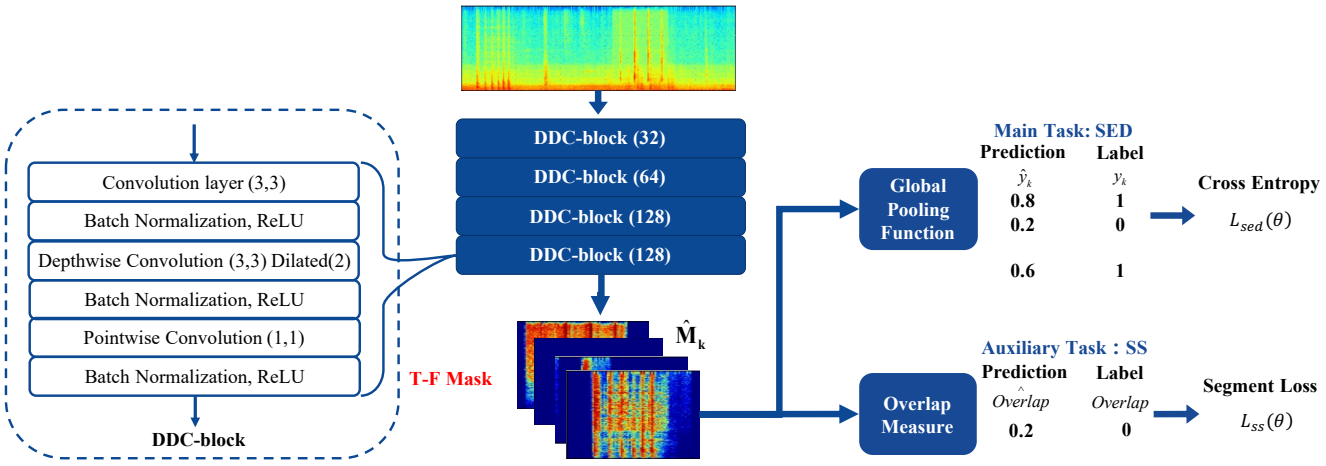

Fig. 1: The diagram of multi-task learning method.

2.1. The weakly supervised SED method

In this paper, we utilize a mask-based method to achieve the detection goal. Mask-based method consists of a segmentation mapping stage and a classification mapping stage. In the segmentation mapping stage, a log-mel spectrogram $\mathbf{X} = [|X(t,f)|]$ is extracted as the input feature, where t and f represent frame and frequency indices, respectively. Then, the segmentation mapping of $\mathbf{X} \rightarrow \hat{\mathbf{M}}$ is modeled via a model consisting of DDC-blocks [16], where $\hat{\mathbf{M}} = [\hat{M}_k(t,f)]$ is the estimation of the ideal ratio mask (IRM), and k represents sound event class. In the classification mapping stage, we utilize the frequency-dependent auto-pooling function (FAP) to aggregate T-F masks into the clip-level prediction, i.e., $\hat{\mathbf{M}} \rightarrow \hat{\mathbf{y}}$, where $\hat{\mathbf{y}} = [\hat{y}_1, \hat{y}_2, \cdots, \hat{y}_K]^{\mathrm{T}}$ denotes the clip-level probability. Both of AT and SED tasks share the same training stage as described above, and the SED prediction can be obtained by aggregating the mask across the frequency axis as:

$$\hat{y}_k(t) = \sum_{f=1}^{F} w_k(f) \hat{M}_k(t,f), \tag{1}$$

where $w_k(f)$ denotes the weight of the f-th frequency band for the k-th class, and $\hat{y}_k(t)$ is the estimated frame-level probability. In this paper, the architecture of SED model is the same as DDC-FAP [16], which achieves state-of-the-art performance in mask-based methods.

2.2. Source separation as the auxiliary task

The diagram of the proposed MTL method for WSSED is shown in Figure 1. For the source separation task, the output of segmentation mapping stage, i.e., $\hat{M}_k(t,f)$, could be used to separate the target source (events) from the mixed audio. Different events types can be regarded as different sound sources. Since the isolated source signal is not available for WSSED task, the reconstruction loss cannot be calculated directly for training. It is reasonable to assume that the sound events are independent of each other. The time-frequency domain representation of the signal is sparser than the corresponding time-domain representation, which means the possibility of multiple sound sources overlapping at each T-F bin is very small. Even though there may be a certain degree of overlap in T-F bins among some kind of events, forcibly assigning the overlapped T-F bins to a class is still helpful to learn the most representative features. The degree of overlap among the masks can be evaluated by:

$$L_{ss} = \sum_{i \neq j, i,j \in [1,K]} overlap(\hat{M}_i(t,f), \hat{M}_j(t,f)), \tag{2}$$

where i and j represent two types of events in the predefined categories, K represents the number of event types, and the $overlap()$ function can be expressed by dice coefficient which measures the degree of overlap among the masks. Dice coefficient [17] is a common evaluation metric for image segmentation, which can measure the similarity of two regions or sets. The dice coefficient can be represented as:

$$dice(A,B) = 2\frac{|A \cap B|}{|A| + |B|}, \tag{3}$$

where the set A and B represent T-F bins of two different events, and $|A \cap B|$ represents the intersection of them. The dice coefficient takes a value between [0,1] and a larger value means a higher degree of overlap between sets. Because the clean signal is not available, the dice coefficient can only be calculated between the estimated separated sources. To minimize the degree of overlap among the masks, we treat dice coefficient as segment loss and add it to the loss function in form of the regularization constraint. The auxiliary segment loss is given by:

$$L_{ss\_segment} = \sum_{i \neq j, i,j \in [1,K]} dice(\hat{M}_i(t,f), \hat{M}_j(t,f)). \tag{4}$$

3. LOSS FUNCTION

There may be hundreds of sound event types for many SED datasets. When the segment losses of all the sound events are required to be considered in pairs, the dice coefficient function has to be calculated for C_K^2 times. For scenes with a large number of event types, the computational complexity is relatively high. For this reason, we propose two approaches to select the masks of various events, and then calculate the segment loss for the selected event categories.

3.1. Label-guided method

To select the masks, the simplest way is to utilize the clip-level label, which is called the label-guided method here. The clip-level label of sample n can be represented as $\mathbf{y}_n = [y_{n,1}, y_{n,2}, \cdots, y_{n,K}]^{\mathrm{T}}$. Only if the k-th type of event is active, $y_{n,k} = 1$, and vice versa, $y_{n,k} = 0$. The active event can be represented by set $\Omega_{active} = \{k|y_{n,k} = 1\}$. The segment loss calculated by the label-guided method can be expressed as:

$$L_{ss\_segment} = \sum_{i \neq j, i,j \in \Omega_{active}} dice(\hat{M}_i(t,f), \hat{M}_j(t,f)). \quad (5)$$

This method selects the masks using the weak labels, which can accurately find the active events and impose the constraints on their masks. However, when the system is trained by the label-guided method, the maximum number of events which can be handled is fixed. This maximum number is closely related to the training samples, and cannot be set manually or extended flexibly.

3.2. Self-supervised method

To handle the variable number of events, we present a self-supervised method that utilizes the clip-level estimation to select masks instead of any labels. According to the clip-level prediction $\hat{\mathbf{y}}_n = [\hat{y}_{n,1}, \hat{y}_{n,2}, \cdots, \hat{y}_{n,K}]^{\mathrm{T}}$, this method sorts all the predefined events. Then, we take the first $Topk$ events with the highest probability as the set of active events Ω_{Topk}, where $Topk$ is a free parameter. Finally, the segment loss can be calculated among the $Topk$ events, and (5) is rewritten as:

$$L_{ss\_segment} = \sum_{i \neq j, i,j \in \Omega_{Topk}} dice(\hat{M}_i(t,f), \hat{M}_j(t,f)). \quad (6)$$

The performance of mask selection is highly related to SED performance. An accurate event prediction helps to generate the meaningful supervised information. Without requiring weak labels in advance, this method can overcome the limitation of the fixed maximum number of events in detection, which can be directly extended to the unsupervised task. The parameter $Topk$ can be experimentally selected according to the task scenario in application.

3.3. Silent loss

In addition to imposing the constraint on active events, inactive events can also promote the training of the model. For the inactive events, the masks of them should not contain any active T-F bins, and the masks should be close to zero. To measure the active degree of inactive events, we calculate the L_1 norm of the remaining inactive events as the silent loss after the active mask selection. The inactive event category can be denoted by $\Omega_{inactive} = \{k|y_{n,k} = 0\}$ and the silent loss can be expressed as:

$$L_{ss\_silent} = \sum_{i \in \Omega_{inactive}} \frac{1}{TF} \sum_{t,f} |\hat{M}_i(t,f)|. \quad (7)$$

The loss function of the main SED part and auxiliary source separation part can be written as:

$$\begin{aligned} L_{total} &= \lambda_1 L_{sed}(\mathbf{y_n}, \hat{\mathbf{y}}_\mathbf{n}) \\ &+ \lambda_2 L_{ss\_segment}(\hat{M}_{active1}(t,f), \hat{M}_{active2}(t,f)) \\ &+ \lambda_3 L_{ss\_silent}(\hat{M}_{inactive}(t,f)), \end{aligned} \quad (8)$$

where λ_1, λ_2 and λ_3 represent the weights of cross entropy, segment loss and silent loss, respectively. The first item is the classification loss provided by the SED task, which predicts the active event and calculates the cross-entropy with the label. The second item is the segment loss provided by the auxiliary source separation task, which is mainly used to measure the degree of mask overlap between active events. The third item is silent loss, which is mainly used to measure the activity of inactive events. Segment loss and silent loss can be used jointly as a regularization term to assist the training of the SED system.

4. EXPERIMENTS

To verify the effectiveness of the proposed MTL method, we perform a series of experiments on two datasets. The first one is synthesized with the audio clips from DCASE 2018 [18] in the overlapped way. The second one is DCASE 2020 task 4 dataset [19] which is recorded in real environments.

4.1. System Setting and Metrics

We choose 64 bands log-mel spectrum as the input feature to our model, which is generated with a Hanning window of length 64 ms and a hop size of 32 ms. Adam optimizer [20] is used with the initially 1e-3 learning rate, and the learning rate is reduced to 0.9 times of the previous value per 1000 iterations. The mean average precision (mAP) [21], error rate (ER), deletions (D), and insertions (I) are utilized to evaluate the performance of SED tasks. The mAP is the average of precision at different recall values regardless of thresholds, which can evaluate the model comprehensively. Error rate (ER) measures the number of errors including deletions (D)

Table 1: Performance evaluation on the overlapped dataset

	Method	$Topk$	AT-mAP	SED-mAP	ER	D	I
	DDC-FAP	/	0.662	0.374	1.741	0.929	0.811
	Label-guided	/	**0.683**	**0.408**	1.661	**0.929**	0.731
Segment	Self-	2	0.676	0.402	1.678	0.935	0.724
Loss	super-	3	0.676	0.404	1.648	0.917	0.731
	vised	5	0.682	0.408	**1.631**	0.930	**0.702**
Segment	Label-guided	/	**0.697**	**0.419**	**1.543**	0.931	**0.612**
Loss +	Self-	2	0.684	0.412	1.662	0.930	0.732
Silent	super-	3	0.686	0.415	1.662	**0.926**	0.736
Loss	vised	5	0.684	0.408	1.580	0.935	0.646

and insertions (I). ER is a score that may become larger than 1 when the system makes more errors than correct predictions. In addition, we set the onset collar of 200 ms and an offset collar of 200 ms/50% to count the true positives of the prediction, which is similar to the configuration of [21].

4.2. MTL method on the overlapped dataset

DCASE 2018 Task 1 dataset is recorded in 10 real scenes such as metro station. DCASE 2018 Task 2 dataset includes 41 categories of sound events. For the overlapped dataset, we utilize DCASE 2018 Task 1 dataset as background noise and the monophonic DCASE 2018 Task 2 dataset as the sound events. The events shorter than 4 s are padded with zeros to 4 s, and the events longer than 4 s are truncated. Two randomly selected events are mixed with the background noise at 20-dB SNR, and 8000 audio clips are synthesized. Both of their onsets are set to 3 s. To evaluate the effect of each loss, we add the segment loss to the loss function first, and then add the corresponding silent loss. When performing the multi-task optimization, these three losses should be ensured in the same numerical range to prevent the MTL method from overfitting in any task. The weight of classification loss is empirically set to 1, and the weights of segment loss and silent loss are set to 0.2. In the self-supervised mask selection method, we set $Topk = 2$, $Topk = 3$ and $Topk = 5$, respectively. The results are shown in Table 1.

As shown in Table 1, both mask selection methods improve the performance of detection, which indicates that the utilization of segment loss is of great help for the SED task. The label-guided method and self-supervised method achieve comparable performance. When the silent loss is introduced, the label-guided method achieves the highest mAP and the lowest error rate among all the methods for AT and SED tasks. Because the label-guided method can directly utilize the weak labels, the constraints can be accurately imposed on the corresponding active and inactive events. The self-supervised method is slightly inferior to the label-guided method. This may be because that the selected active events by self-supervised method is not accurate enough. When $Topk$ is set to be larger than the number of real active events, the inactive events which should be constrained by silent loss are imposed a more relaxed segment loss.

Table 2: Performance evaluation on DCASE 2020 Task 4

Method	AT-mAP	SED-mAP	ER	D	I
Attention	0.810	0.572	1.715	**0.755**	0.960
TALNet	0.741	0.516	**1.523**	0.773	0.750
VGG-GWRP	0.772	0.578	1.866	0.769	1.096
DDC-FAP	0.795	0.610	1.755	0.790	0.965
MTL-DDC-FAP	**0.815**	**0.632**	1.551	0.803	**0.747**

4.3. MTL method on DCASE 2020 task 4

Compared with the manual synthesis dataset, DCASE 2020 TASK 4 dataset is recorded in real environments, which means the acoustic environment, event duration, location, and proportion of event overlap are more natural. This dataset contains 10 types of sound events that may occur in the home environment such as voice, vacuum cleaners, and washing dishes. We evaluate the proposed MTL method on the weakly labeled part of DCASE2020 TASK 4 dataset. Several well-known methods of the WSSED task, i.e., Attention [22], TALNet [23], and VGG-GWRP [24] are also involved to make a comprehensive comparison. All of these methods utilize the same system settings and post-processing operations. In this subsection, we utilize the label-guided MTL method and add both the segment loss and the silent loss to jointly train the MTL model. The results are shown in Table 2.

It can be seen from Table 2 that the DDC-FAP method archives a better SED performance than the Attention and TALNet method. However, the ER metric of the DDC-FAP method is relatively high. In the MTL-DDC-FAP method, the segment loss imposed on active events and silent loss imposed on inactive events can significantly reduce the number of insertion errors. The introduction of auxiliary source separation task further improves the performance of event detecting and achieves the best mAP score in both AT and SED tasks.

5. CONCLUSION

In this paper, we proposed a MTL method which takes source separation as the auxiliary task to solve the weakly supervised SED problem. The main idea is to impose different constrains on the masks of active or inactive events. Two methods are proposed to select the active events. The label-guided method selects the masks according to the weak labels. Thus, it can accurately impose different constraints on various events. The flexible self-supervised method selects the masks according to the clip-level probabilities, which can be directly extended to the unsupervised task. Experiments showed that both the segment loss and silent loss are beneficial for the training of SED. The proposed MTL method achieves a better performance than the state-of-the-art mask-based method on DCASE 2020 Task 4 dataset.

6. REFERENCES

[1] T. Virtanen, M. D. Plumbley, and D. Ellis, *Computational analysis of sound scenes and events*, Springer, 2018.

[2] Y. Lavner, R. Cohen, D. Ruinskiy, and H. IJzerman, "Baby cry detection in domestic environment using deep learning," in *Proc. IEEE ICSEE*, 2016, pp. 1–5.

[3] G. Valenzise, L. Gerosa, M. Tagliasacchi, F. Antonacci, and A. Sarti, "Scream and gunshot detection and localization for audio-surveillance systems," in *Proc. IEEE AVSS*, 2007, pp. 21–26.

[4] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "Panns: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 2880–2894, 2020.

[5] A. Jati and D. Emmanouilidou, "Supervised deep hashing for efficient audio event retrieval," in *Proc. IEEE ICASSP*, 2020, pp. 4497–4501.

[6] R. Ranjan, S. Jayabalan, T. N. T. Nguyen, and W. S. Gan, "Sound event detection and direction of arrival estimation using residual net and recurrent neural networks," in *Proc. DCASE Workshop*, 2019, pp. 2144–218.

[7] H. Zhang, I. McLoughlin, and Y. Song, "Robust sound event recognition using convolutional neural networks," in *Proc. IEEE ICASSP*, 2015, pp. 559–563.

[8] E. Cakır, G. Parascandolo, T. Heittola, H. Huttunen, and T. Virtanen, "Convolutional recurrent neural networks for polyphonic sound event detection," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, pp. 1291–1303, 2017.

[9] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M. D. Plumbley, "Detection and classification of acoustic scenes and events," *IEEE Trans. Multimedia*, vol. 17, pp. 1733–1746, 2015.

[10] T. N. T. Nguyen, D. L. Jones, and W. S. Gan, "A sequence matching network for polyphonic sound event localization and detection," in *Proc. IEEE ICASSP*, 2020, pp. 71–75.

[11] F. Kang, F. Yang, and J. Yang, "A low-complexity permutation alignment method for frequency-domain blind source separation," *Speech Commun.*, vol. 115, pp. 88–94, 2019.

[12] X. Xia, R. Togneri, F. Sohel, Y. Zhao, and D. Huang, "Multi-task learning for acoustic event detection using event and frame position information," *IEEE Trans. Multimedia*, vol. 22, no. 3, pp. 569–578, 2019.

[13] H. Phan, M. Krawczyk-Becker, T. Gerkmann, and A. Mertins, "Weighted and multi-task loss for rare audio event detection," in *Proc. IEEE ICASSP*, 2018, pp. 336–340.

[14] Q. Kong, Y. Xu, W. Wang, and M. D. Plumbley, "A joint separation-classification model for sound event detection of weakly labelled data," in *Proc. IEEE ICASSP*, 2018, pp. 321–325.

[15] F. Pishdadian, G. Wichern, and J. Le Roux, "Finding strength in weakness: Learning to separate sounds with weak supervision," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 2386–2399, 2020.

[16] S. Liu, F. Yang, Y. Cao, and J. Yang, "Frequency-dependent auto-pooling function for weakly supervised sound event detection," *EURASIP J. Audio, Speech, Music Process.*, vol. 1, pp. 1–11, 2021.

[17] N.J. Tustison and J.C. Gee, "Introducing dice, jaccard, and other label overlap measures to itk," *Insight J*, vol. 2, 2009.

[18] E. Fonseca, P. J. Pons, X. Favory, C. F. Font, D. Bogdanov, A. Ferraro, and X. Serra, "Freesound datasets: a platform for the creation of open audio datasets," in *Proc. ISMIR*, 2017, pp. 486–493.

[19] R. Serizel, N. Turpault, A. Shah, and J. Salamon, "Sound event detection in synthetic domestic environments," in *Proc. IEEE ICASSP*, 2020, pp. 86–90.

[20] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[21] A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for polyphonic sound event detection," *Appl. Sci.*, vol. 6, pp. 162, 2016.

[22] Y. Xu, Q. Kong, W. Wang, and M. D. Plumbley, "Large-scale weakly supervised audio classification using gated convolutional neural network," in *Proc. IEEE ICASSP*, 2018, pp. 121–125.

[23] Y. Wang, J. Li, and F. Metze, "A comparison of five multiple instance learning pooling functions for sound event detection with weak labeling," in *Proc. IEEE ICASSP*, 2019, pp. 31–35.

[24] Q. Kong, Y. Xu, I. Sobieraj, W. Wang, and M. D. Plumbley, "Sound event detection and time–frequency segmentation from weakly labelled data," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, pp. 777–787, 2019.