A FRAME LOSS OF MULTIPLE INSTANCE LEARNING FOR WEAKLY SUPERVISED SOUND EVENT DETECTION

Xu Wang, Xiangjinzi Zhang*, Yunfei Zi, Shengwu Xiong†*

School of Computer and Artificial Intelligence, Wuhan University of Technology, Wuhan, China
abracadabra@whut.edu.cn

ABSTRACT

Sound event detection(SED) consists of two subtasks: predicting the classes of sound events within an audio clip (audio tagging) and indicating the onset and offset times for each event (localization). One of the common approaches for SED with weak label is multiple instance learning (MIL) method. However, the general MIL method only optimizes the global loss calculated from the aggregated clip-wise predictions and weak clip labels, lacking a direct constraint on the frame-wise predictions, which leads to a large number of unreasonable prediction values. To address this issue, we explore the deterministic information that can be used to constrain the frame-wise predictions and based on which we design a frame loss with two terms. Experimental results on the DCASE2017 Task4 dataset demonstrate that the proposed loss can improve the performance of general MIL method. While this article focuses on SED applications, the proposed methods could be applied widely to MIL problems. Code will be available at WSSED.

Index Terms— Sound event detection (SED), weak labeling, multiple instance learning (MIL), loss function

1. INTRODUCTION

Sound event detection (SED) is a technique for predicting the presence or absence of sound events and detecting the event time boundaries within an audio clip, given that multiple events can be present in the audio clips. SED has great potential in many scenarios, such as sound retrieval [1], smart cities and homes [2, 3]. Recently weakly-supervised SED (WSSED) is getting much attention due to challenges such as the Detection and Classification of Acoustic Scenes and Events (DCASE) challenge, which has only access to weak clip labels during training, yet needs to predict onsets and offsets during evaluation.

A common framework for WSSED is multiple instance learning (MIL) [4, 5]. In the MIL, the input sequence is treated as a bag and split into a set of instances. Feeding each instance into a neural network classifier can get an instance-wise prediction over each event. Then, a pooling function aggregates the instance-wise predictions to a bag-wise prediction, which can be compared against the weak bag label to calculate a loss, and the classifier can be optimized by minimizing the loss. In SED, each training clip is regarded as a bag, and its frames are regarded as instances.

There are two primary stages in MIL-based WSSED: performing frame-wise predictions and aggregating these frame-wise predictions to clip-wise predictions. For performing the frame-wise predictions, many methods used neural networks, such as CNN [6], RNN [7, 8], CRNN [9, 10]. Recently, with the great success of the Transformer in the natural language processing (NLP) field [11], the self-attention mechanism also has attracted much attention for acoustic feature representation in SED task [12, 13, 14], which can allow to take both local and global context information of the input feature sequence into account. For aggregating the frame-wise predictions, an essential piece of work done in [15], compared five types of pooling functions, and found the linear softmax pooling function to perform the best among the five. Besides, an adaptive pooling function has been proposed in [16], which can smoothly interpolate between common pooling operators. Other studies in [17] compares the performance of pooling on feature and pooling on prediction. More recently, various attention mechanisms have been employed to detect the occurrence of sound events [18, 19] and have achieved promising results.

However, the general MIL method only calculate global loss on the aggregated clip-wise predictions. Due to the lack of direct constraint, the frame-wise predictions is highly in randomness, which will lead to a large number of illogical predictions and then greatly affects the performance of localization task. We observe that certain frame "groundtruth" can be determined based on the known information, including weak clip labels and event duration. The frame groundtruth includes frame labels or regularity that exists in the frame labels, which can be used to constrain the frame-wise predictions, but ignored by the general MIL method. In the remaining part we explore the frame groundtruth and based on which we design a frame loss with two terms.

This work was in part supported by by NSFC (Grant No. 62176194) and the Major project of IoV (Grant No. 2020AAA001) and Sanya Science and Education Innovation Park of Wuhan University of Technology (Grant No. 2021KF0031).

*Corresponding author: xiongsw@whut.edu.cn

2. METHOD

In this section, we start with a brief description of the MIL method. Then, the definitions of valid prediction and invalid prediction are given to facilitate the presentation of our approach. Finally, we explore the frame groundtruth and design a frame loss with two terms on valid prediction and invalid prediction respectively.

2.1. Multiple Instance Learning

Applying this framework to WSSED, we view an audio clip X as a bag of instances $X = \{x_1, \cdots, x_M\}$, where each instance x_m represents an audio frame. A multi–hot label $Z \in \{0,1\}^N$ is given for each clip which indicates if the n^{th} event occur ($Z^n = 1$) or not ($Z^n = 0$) in the clip. M and N are the number of frames and events, respectively. Our goal is to train a frame-level neural network classifier f, feeding each frame into the classifier can get a frame-wise prediction $y = f(x), y \in \{0,1\}^N$. The frame predictions of clip X can be described as $Y_f \in R^{M \times N}$. Then, a pooling function aggregates Y_f to a clip-wise prediction $Y_c = agg(Y_f), Y_c \in \{0,1\}^N$, and the classifier can be optimized by minimizing the global loss between Y_c and Z.

$$\mathcal{L}_{global}(X) = BCE(Y_c, Z) \tag{1}$$

Where $BCE(\cdot)$ represents a binary cross entropy between the labels and the predictions.

2.2. Valid Prediction and Invalid Prediction

For each clip X, if x is active at the n^{th} event, i.e. $Z^n = 1$, then we call all the frame prediction values for X over the n^{th} event as *valid prediction*, defined as $\tilde{y}$. Similarly, *invalid prediction* $\hat{y}$ is defined as the frame prediction values of clip over inactive events, as depicted in Fig.1. Notice that both valid prediction and invalid prediction are event-wise, not frame-wise. Thus, the transpose of Y_f can be considered as a set of $\hat{y}$ and $\tilde{y}$, formulaically, $\hat{y}, \tilde{y} \in Y_f^T$.

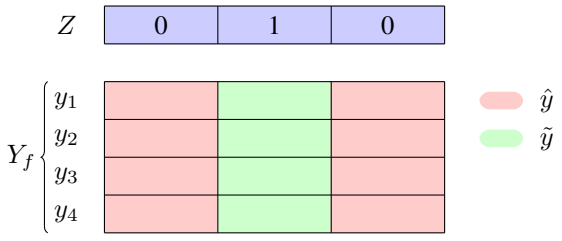

Fig. 1: Definition of *valid prediction* and *invalid prediction*. Y_f consists of four frame-wise predictions over three events. Z is the weak clip label, $\tilde{y}$ and $\hat{y}$ are valid prediction and invalid prediction, respectively.

2.3. Frame Loss on Valid Prediction

Observation 2.1. *Valid prediction should tend to be local smoothing.*

The duration of the sound events (the shortest average event duration in DCASE2017 task4 dataset is 3.8s [20]) is usually longer than the frame length (0.1s) commonly used in SED. So, even without knowing the strong frame label, we can assert that in most cases adjacent frames will have the same label over single event, and this expression does not hold only at the event boundaries. Hence, the classifier's prediction values for adjacent frames over single event should be close to each other, and there ought not be many large fluctuations. We call this regularity as *local smoothing*. We choose variance to measure the smoothness, the loss term on valid prediction is designed as follows:

$$\mathcal{L}_{fv}(X) = \sum_{\tilde{y} \in Y_f^T} [var(\tilde{y}[1:u]) + var(\tilde{y}[u+1:2u]) + \cdots] \tag{2}$$

Where $y[i:j]$ represents the selection of values from i to j in vector y, and u is a hyperparameter used to set the smooth unit interval. It is true that the loss term will affect the predictions at the boundary of events, but this situation is a minority, while a suitable u can further reduce this effect. In addition, excellent threshold setting is strongly dependent on smooth event-wise predictions, which will directly affect the accuracy of the binary detection. We have traded a smaller loss for a larger return.

2.4. Frame Loss on Invalid Prediction

Observation 2.2. *Invalid prediction should be close to 0.*

If a clip does not contain an event, then obviously all of its time periods will not contain that event. So, for the classifier's frame prediction values on inactive events, the smaller the better. In other words, a certain groundtruth exists for each invalid prediction, which is a zero vector. Thus, the loss can be designed as follows:

$$\mathcal{L}_{fi}(X) = \sum_{\hat{y} \in Y_f^T} ||\hat{y}||_2 \tag{3}$$

Since the groundtruth is a zero vector, here we directly choose the L_2 norm of $\hat{y}$.

Compared to the general MIL method, the classifier is trained by minimizing the following objective function, which is the combination of global loss and frame loss:

$$\mathcal{L}_{frame}(X) = \lambda_{fv}\mathcal{L}_{fv}(X) + \lambda_{fi}\mathcal{L}_{fi}(X) \tag{4}$$

$$\mathcal{L}(X) = \mathcal{L}_{global}(X) + \mathcal{L}_{frame}(X) \tag{5}$$

where λ_{fv} and λ_{fi} are trade-off parameters.The Adam [21] is employed as the optimizer.

Table 1: The six pooling functions of our experiment. m is the number of frames in a clip.

	Pooling Function
Max	$y = max_i y_i$
Average	$y = \frac{1}{m}\sum_i y_i$
Lin.	$y = \frac{\sum_i y_i^2}{\sum_i y_i}$
Exp.	$y = \frac{\sum_i y_i exp(y_i)}{\sum_i exp(y_i)}$
Auto.	$y = \frac{\sum_i y_i exp(\alpha \cdot y_i)}{\sum_i exp(\alpha \cdot y_i)}$
Attention	$y = \frac{\sum_i y_i \omega_i}{\sum_i \omega_i}$

3. EXPERIMENT

In this section, we report the experimental results and comparisons that demonstrate the effectiveness of the proposed frame loss. In addition, ablation studies are carried out to show the contribution of the two loss terms.

3.1. Experiment Conditions

We choose the system in [15] as baseline, which is based on the standard multi instance learning method. The experimental condition is consistent with [15] to verify effectiveness. The WSSED is implemented by Mindspore [1] [2].

We evaluate the proposed frame loss on Task4 of the DCASE2017 challenge. The dataset used in the task is a subset of Audio Set [22]. It consists of a training set (51,172 recordings), a public test set(488 recordings), and a private evaluation set (1,103 recordings). All the recordings are 10-second excerpts from YouTube videos.

The performance of audio tagging was evaluated with the micro-average F1 on the recording level; localization was evaluated with the micro-average error rate (ER) and F1 on 1-second segments. The F1 is the larger the better while the error rate is the smaller the better. While evaluation, we searched for the best threshold for each class using the validation dataset.

For the acoustic features, we extracted the 64-dimensional log mel-band energy at a sampling rate of 16 kHz, which was calculated every 40 ms with a 20 ms hop size, each clip has 400 frames. We choose "TALNet" proposed in [15] as our baseline network, which has 10 convolutional layers, 5 pooling layers, and 1 recurrent layer. We set the hyperparameters λ_{vpl} as $1e-3$, λ_{ipl} as 0.01 and s as 3 in our experiments. Data balancing and batch normalization have been applied during training.

[1] https://www.mindspore.cn/

[2] We thank MindSpore for the partial support of this work

Table 2: The performance of the MIL method with and without frame loss on the DCASE2017 task4 dataset. G and F represent the general global loss and the proposed frame loss, respectively. Error rates and F1's are in percentages.

Pooling	Loss	Tagging	Localization	
		F_1	F_1	**ER**
Max	$G \oplus F$	46.2	35.7	84.6
	G	45.3	35.4	84.7
Average	$G \oplus F$	50.8	41.9	98.6
	G	50.0	41.3	105.9
Lin.	$G \oplus F$	50.9	43.6	82.4
	G	49.5	43.7	84.3
Exp.	$G \oplus F$	48.8	42.4	92.5
	G	48.5	42.8	100.6
Auto.	$G \oplus F$	49.3	42.8	94.3
	G	48.5	41.9	98.8
Attention	$G \oplus F$	50.1	41.3	96.5
	G	49.2	40.1	102.5

3.2. Experiment Results

We evaluate our proposed method on six common pooling functions, as shown in Table 1, where Lin., Exp. and Auto. represent linear softmax, exponential softmax and adaptive pooling [23] function respectively. Experimental results with and without frame loss are shown in Table 2, the results show that the proposed frame loss can improve the performance of general MIL method with various pooling function in all of the metrics.

Notably, for average, Exp. and attention pooling functions, the error rate in the localization task has been significantly (Above 5%) reduced. As analyzed in [15], the three pooling functions will lead to a lot of false positive frames, which means excessive invalid prediction values are too large. The results prove that our proposed frame loss can effectively suppresses the invalid prediction values, so the error rate is greatly reduced. For max pooling function, the fact that only one frame receives a non-zero gradient leads to a small role for the frame loss, thus there is almost no boost in the localization task.

Table 3 shows the breakdown of the error types on the localization task with and without frame loss for average pooling function. The results show that the purposed loss function can significantly increase the number of false negative and decrease the number of false positive while maintaining true positive. This is a good corroboration of our argument in Sec.2.4 and statement above.

In addition, we compare the event-wise prediction values before and after adding the frame loss. Fig.2 shows that the valid prediction has a large fluctuation while without frame

loss. After adding the frame loss, the valid prediction becomes smoother and is closer to the groundtruth. Meanwhile, as shown in Fig.3, after adding frame loss, the invalid prediction is closer to zero. The experimental results are well in accordance with Observation 2.1 and Observation 2.2, which can be used to constrain the frame predictions effectively.

Table 3: Breakdown of the error types on the localization task with and without frame loss.

	Localization					
	TP	**FN**	**FP**	**SUB**	**DEL**	**INS**
G	2114	2246	3758	1385	861	2373
$G \oplus F$	2246	2589	3626	1472	1117	2154

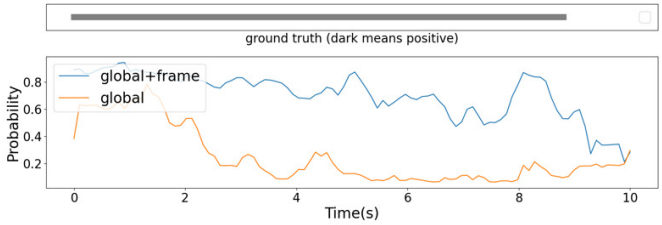

Fig. 2: Comparison of valid predictions before and after adding the purposed frame loss.

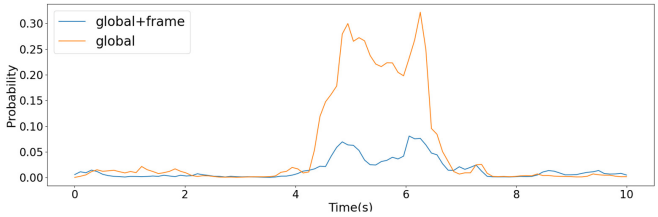

Fig. 3: Comparison of invalid predictions before and after adding the purposed frame loss.

3.3. Ablation Study

In order to explore the effectiveness of the two terms in the frame loss, we use one of them in combination with global loss to optimize the classifier, respectively, the Lin. pooling function is used. The influence of each loss is demonstrated in Table 4. The results show that any one of them can reduce the error rate while ensuring F1 score for both audio tagging and localization. In addition, we can see that using $\mathcal{L}_{fi}$ alone performs better than only using $\mathcal{L}_{fv}$. For the reason that there are many more inactive time frames of sound events than active frames [24], this means the number of invalid predictions will be much more than the number of valid predictions, so

Table 4: Ablative study of the two frame loss terms. $\mathcal{L}_{fv}$ and $\mathcal{L}_{fi}$ represent the loss term on valid prediction and the loss term on invalid predictions, respectively.

		G	$G \oplus \mathcal{L}_{fv}$	$G \oplus \mathcal{L}_{fi}$
Tagging	**Precision**	46.9	47.3	48.5
	Recall	52.3	52.4	52.9
	F_1	49.5	49.7	50.6
Localization	**Precision**	45.6	45.6	45.2
	Recall	42	41.3	42.6
	F_1	43.7	43.3	43.8
	ER	84.3	84.1	82.5

$\mathcal{L}_{fi}$ can work on more prediction values and improve the performance to a much greater extent.

In addition, we try a new way of constructing the loss term on invalid prediction. The new loss is defined as follows:

$$\hat{\mathcal{L}}_{fi}(X) = MSE(\hat{Y}_f^T, Y_f^T) \tag{6}$$

Here $\hat{Y}_f^T$ is a fictitious strong frame labels for each clip constructed from Y_f. Compared to Y_f^T, $\hat{Y}_f^T$ keeps all valid predictions and replaces all invalid predictions with zero vectors. The difference between the two loss terms can be described as follows: Eq.3 simply converges the invalid predictions to zero, while Eq.6 tries to keep the valid predictions while converging the invalid predictions.

Table 5: Ablative study of $\hat{\mathcal{L}}_{fi}$ and $\mathcal{L}_{fi}$

	Tagging	**Localization**	
	F_1	F_1	**ER**
$\mathcal{L}_{fi}$	50.6	43.8	82.5
$\hat{\mathcal{L}}_{fi}$	51.2	42.5	83.6

We simply compare the performance of $\hat{\mathcal{L}}_{fi}$ and $\mathcal{L}_{fi}$ with the same experimental scheme as Table 4. The results in Table 5 show that $\hat{\mathcal{L}}_{fi}$ performs better in the localization task, while $\mathcal{L}_{fi}$ performs better in the tagging task.

4. CONCLUSION

We presented a frame loss with two terms to take advantage of the potential deterministic information at the frame level, which is ignored by general MIL method. Experimental results show that the purposed frame loss can boost the performance of general MIL method on six various pooling functions. Finally, although we focus on SED applications in this article, we emphasize that the proposed frame loss could be applied to MIL problems in any application domain.

5. REFERENCES

[1] Frederic Font, Gerard Roma, and Xavier Serra. Sound sharing and retrieval. *Computational Analysis of Sound Scenes and Events*, pages 279–301, 2018.

[2] Juan Pablo Bello, Charlie Mydlarz, and Justin Salamon. Sound analysis in smart cities. pages 373–397, 2018.

[3] Sacha Krstulović. Audio event recognition in the smart home. pages 335–371, 2018.

[4] Jaume Amores. Multiple instance classification: Review, taxonomy and comparative study. *Artificial Intelligence*, 201:81–105, 2013.

[5] Shao-Yen Tseng, Juncheng Li, Yun Wang, Florian Metze, Joseph Szurley, and Samarjit Das. Multiple instance deep learning for weakly supervised small-footprint audio event detection. In *Interspeech 2018*, pages 3279–3283, 2018.

[6] Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, R. Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm Slaney, Ron J. Weiss, and Kevin Wilson. Cnn architectures for large-scale audio classification. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 131–135, 2017.

[7] Giambattista Parascandolo, Heikki Huttunen, and Tuomas Virtanen. Recurrent neural networks for polyphonic sound event detection in real life recordings. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6440–6444, 2016.

[8] Tomoki Hayashi, Shinji Watanabe, Tomoki Toda, Takaaki Hori, Jonathan Le Roux, and Kazuya Takeda. Duration-controlled lstm for polyphonic sound event detection. *IEEE Transactions on Audio, Speech, and Language Processing*, 25(11):2059–2070, 2017.

[9] Emre Cakir, Giambattista Parascandolo, Toni Heittola, Heikki Huttunen, and Tuomas Virtanen. Convolutional recurrent neural networks for polyphonic sound event detection. *IEEE Transactions on Audio, Speech, and Language Processing*, 25(6):1291–1303, 2017.

[10] Sharath Adavanne, Pasi Pertila, and Tuomas Virtanen. Sound event detection using spatial features and convolutional recurrent neural network. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 771–775, 2017.

[11] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, volume 30, pages 5998–6008, 2017.

[12] Qiuqiang Kong, Changsong Yu, Yong Xu, Turab Iqbal, Wenwu Wang, and Mark D. Plumbley. Weakly labelled audioset tagging with attention neural networks. *IEEE Transactions on Audio, Speech, and Language Processing*, 27(11):1791–1802, 2019.

[13] Yifang Yin, Meng-Jiun Chiou, Zhenguang Liu, Harsh Shrivastava, Rajiv Ratn Shah, and Roger Zimmermann. Multi-level fusion based class-aware attention model for weakly labeled audio tagging. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 1304–1312, 2019.

[14] Maximilian Ilse, Jakub M. Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *35th International Conference on Machine Learning, ICML 2018*, pages 2127–2136, 2018.

[15] Yun Wang, Juncheng Li, and Florian 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)*, pages 31–35, 2019.

[16] Brian McFee, Justin Salamon, and Juan Pablo Bello. Adaptive pooling operators for weakly labeled sound event detection. *IEEE Transactions on Audio, Speech, and Language Processing*, 26(11):2180–2193, 2018.

[17] Chieh-Chi Kao, Ming Sun, Weiran Wang, and Chao Wang. A comparison of pooling methods on lstm models for rare acoustic event classification. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 316–320, 2020.

[18] You Wang, Chuyao Feng, and David V. Anderson. A multi-channel temporal attention convolutional neural network model for environmental sound classification. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 930–934, 2021.

[19] Helin Wang, Yuexian Zou, and Wenwu Wang. A global-local attention framework for weakly labelled audio tagging. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 351–355, 2021.

[20] Annamaria Mesaros, Toni Heittola, Aleksandr Diment, Benjamin Martinez Elizalde, Ankit Shah, Emmanuel Vincent, Bhiksha Raj, and Tuomas Virtanen. Dcase 2017 challenge setup: tasks, datasets and baseline system. In *DCASE 2017 - Workshop on Detection and Classification of Acoustic Scenes and Events*, pages 85–92, 2017.

[21] Diederik P. Kingma and Jimmy Lei Ba. Adam: A method for stochastic optimization. In *ICLR 2015 : International Conference on Learning Representations 2015*, 2015.

[22] Qiuqiang Kong, Yong Xu, Wenwu Wang, and Mark D. Plumbley. Audio set classification with attention model: A probabilistic perspective. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 316–320, 2018.

[23] Qiuqiang Kong, Yong Xu, Wenwu Wang, and Mark D. Plumbley. Sound event detection of weakly labelled data with cnn-transformer and automatic threshold optimization. *IEEE Transactions on Audio, Speech, and Language Processing*, 28:2450–2460, 2020.

[24] Keisuke Imoto, Sakiko Mishima, Yumi Arai, and Reishi Kondo. Impact of sound duration and inactive frames on sound event detection performance. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 860–864, 2021.