SP-SEDT: Self-supervised Pre-training for Sound Event Detection Transformer

Zhirong Ye$^{1,2}$, Xiangdong Wang$^1$, Hong Liu$^1$, Yueliang Qian$^1$, Rui Tao$^3$, Long Yan$^3$, Kazushige Ouchi$^3$

$^1$Beijing Key Laboratory of Mobile Computing and Pervasive Device, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
$^2$University of Chinese Academy of Sciences, Beijing, China
$^3$Toshiba China R&D Center, Beijing, China

{yezhirong19s, xdwang, hliu, ylqian}@ict.ac.cn, {taorui, yanlong}@toshiba.com.cn, kazushige.ouchi@toshiba.co.jp

Abstract
Recently, an event-based end-to-end model (SEDT) has been proposed for sound event detection (SED) and achieves competitive performance. However, compared with the frame-based model, it requires more training data with temporal annotations to improve the localization ability. Synthetic data is an alternative, but it suffers from a great domain gap with real recordings. Inspired by the great success of UP-DETR in object detection, we propose to self-supervisedly pre-train SEDT (SP-SEDT) by detecting random patches (only cropped along the time axis). Experiments on the DCASE2019 task4 dataset show the proposed SP-SEDT can outperform fine-tuned frame-based model. The ablation study is also conducted to investigate the impact of different loss functions and patch size.

Index Terms: Sound event detection, Self-supervised learning, End-to-end, Transformer

1. Introduction
Sound event detection (SED) aims to recognize sound events and identify their corresponding onset and offset in audio. SED plays an important role in many applications, such as sound-based surveillance [1], video indexing [2], and so on.

Nowadays, deep learning models have become the mainstream approach for SED. Most of them formulate SED as a sequence-to-sequence classification problem, which predicts the event categories for each frame and then aggregates frame-level predictions via post-processing to determine event boundaries, thus obtaining final event-level results [3][4][5]. However, post-processing including binarization by thresholding and median filtering with fixed or adaptive window sizes [5][6] brings many class-dependent hyper-parameters to tune, and the tuning of such parameters may be dependent on a development dataset and inclined to overfitting in some cases. Moreover, under the conventional scheme, the model is optimized to achieve the best frame-level predictions instead of the best event-level results. The inconsistency between model training and evaluation may cause the model to focus more on local (frame-level) information and pay less attention to global and event-level characteristics, thus casting a restriction on the SED performance. To overcome the shortcomings of frame-based models, in our earlier work [7], the object-based prediction mechanism of Detection Transformer (DETR) [6] in the field of image object detection has been borrowed to build an event-based model for SED with transformer (SEDT), which outputs event-level predictions (including categories and boundaries) directly and achieves competitive performance to conventional methods. The event-based model pursues the same goal with the detection task, namely the best event-level predictions, making the model better adapt to the SED task. Besides, the model is capable of learning various event characteristics such as duration, which can only be inserted into the frame-based model by the hyperparameters in the post-processing stage.

However, compared with the conventional two-stage detection systems, the event-based model is more data-hungry [9]. Unfortunately, the shortage of training data is more severe for SED than for other tasks such as image object detection, speech recognition or acoustic scene classification. Actually, due to the high cost of annotated audio clips with temporal boundaries of events, weakly supervised learning [5][10][11] has been applied to train SED models with weakly labeled data only indicating the event categories occurring in audio clips. Synthetic data [12] has also been utilized, which is often cooperated with domain adaption techniques [13][14] to narrow the feature gap between real recordings and synthetic data.

Self-supervised learning has emerged as a paradigm to tackle with low-resource model training and has achieved remarkable progress in many fields, such as natural language processing [15], computer vision [16][17] and automatic speech recognition [18]. Self-supervised pre-training has also been introduced into the audio community. Deshmukh et al. [19] propose to jointly train the feature encoder with the SED task and an auxiliary self-supervised spectrogram reconstruction task. In [20], multiple self-supervised tasks are implemented to train a feature extractor for audio classification tasks. Almost all existing pre-training methods in the audio fields aim to learn better signal representation, which is conducive to classification or recognition tasks. However, the object-based model has two subtasks: classification and localization, where the latter is not a classification task as the model outputs temporal boundaries. Therefore, existing pre-training methods cannot be applied to object-based detection models directly. To improve the localization ability by pre-training, Dai et al. [9] propose UP-DETR, which pre-trains DETR by detecting randomly cropped patches, and achieves great improvement on image object detection. But such spatial-aware pre-training has not been explored in SED.

In this paper, inspired by UP-DETR, we propose to our knowledge the first self-supervised learning approach to pre-train the event-based model SEDT to equip it with better localization ability, which is named as Self-supervised Pre-training SEDT (SP-SEDT). Specifically, we crop the spectrogram along the time axis to obtain random-sized patches at random loca-
2. Sound Event Detection Transformer

In this section, we describe the main idea of Sound Event Detection Transformer (SEDT) briefly, based on which our self-supervised pre-training method is proposed. More details of SEDT can be found in [7].

2.1. Model architecture

An overview of SEDT is illustrated in Figure 1. The model contains three main components, including a backbone, an encoder-decoder, and prediction feed-forward networks (FFNs). SEDT represents each sound event as a vector $y_i = (c_i, b_i)$, where $c_i$ is the class label and $b_i = (m_i, l_i)$ denotes the temporal boundary containing normalized event center $m_i$ and duration $l_i$. Such event representation allows SEDT to output event-level results directly.

The backbone and encoder are adopted for feature extraction. Given the spectrogram of an audio clip, the backbone is utilized to extract its feature map $z_0$. Then the sum of a positional encoding $P$ and the feature map $z_0$ is flattened on the time and frequency axis, resulting in a $d \times TF$ feature map $z$, which is fed into the transformer encoder [21] for further feature extraction. The cascade of CNN and transformer encoder ensures that the model can consider both local and global features.

The decoder is used to generate event and clip representations from the encoded feature. SEDT models SED as a set prediction problem and assumes that each event is independent. Therefore, the standard auto-regressive decoding mechanism adopted in machine translation task is discarded. Instead, the decoder takes $N$ learned embeddings as input, named as event queries, and output $N$ event representations in parallel, where $N$ is a hyper-parameter and larger than the typical number of events in an audio clip. An extra audio query is inserted along with the event queries and is fed into the decoder together. The audio query aggregates clip-level representation for audio tagging task and ensures that enough category information is utilized during the decoding of event representation.

Prediction FFNs are adopted to transform the event and clip representation from the decoder into event detection results $\hat{y}_i = (\hat{c}_i, \hat{b}_i)$ and audio tags $y_{\text{TAG}}$.

2.2. Loss function

Event-level loss: To compute the event-level loss, the matching between target and predicted events is needed, which can be obtained by the Hungarian algorithm [7]. The unmatched predictions will be labeled as “empty”. The loss function includes location loss and classification loss:

$$L_{\text{event}} = L_{\text{loc}} + L_{c}$$  \hspace{1cm} (1)

The location loss is calculated for predictions that are not matched to “empty” events ($c_i \neq \emptyset$). It is a linear combination of the L1 norm and IOU loss between the target and predicted location vector:

$$L_{\text{loc}} = \sum_{i=1}^{N} \mathbb{1}_{(c_i \neq \emptyset)} \left( \lambda_{\text{IOU}} L_{\text{IOU}}(b_i, \hat{b}_\sigma(i)) + \lambda_{\text{L1}} \|b_i - \hat{b}_\sigma(i)\|_1 \right)$$  \hspace{1cm} (2)

where $\lambda_{\text{IOU}}, \lambda_{\text{L1}} \in \mathbb{R}$ are hyperparameters, $\sigma$ is the assignment given by the matching process, and $N$ is the number of predictions. The classification loss is the cross-entropy between the labels and the predictions:

$$L_{c} = \frac{1}{N} \sum_{i=1}^{N} - \log \hat{p}_{\sigma(i)}(c_i)$$  \hspace{1cm} (3)

Clip-level loss: The audio tagging loss is calculated as the binary cross-entropy between the clip-level class label $l_{\text{TAG}}$ and predicted audio tagging $y_{\text{TAG}}$:

$$L_{\text{at}} = \text{BCE}(l_{\text{TAG}}, y_{\text{TAG}})$$  \hspace{1cm} (4)

SEDT is optimized by minimizing both event-level and clip-level losses, while for weakly labeled data, only clip-level loss is calculated:

$$L_{\text{strong}} = L_{\text{event}} + \lambda_{\text{at}} L_{\text{at}}$$  \hspace{1cm} (5)

$$L_{\text{weak}} = \lambda_{\text{at}} L_{\text{at}}$$

where $\lambda_{\text{at}}$ is a hyperparameter.

3. Self-supervised Pre-training SEDT

In this section, we describe the training of SP-SEDT in detail, including pre-training and fine-tuning.
3.1. Random patch detection

Without annotated event boundaries, SP-SEDT is pre-trained to detect randomly cropped patches (only cropped along the time axis with both random size and position). As depicted in Figure 2, SP-SEDT shares almost the same model architecture as SEDT except for the audio query branch. The audio query branch of SEDT is utilized for audio tagging, which requires weakly-labeled data to train and cannot work on unlabeled data.

The pre-training procedure is also shown in Figure 2. Considering that there is usually more than one event occurring in an audio clip, during pre-training, $M$ patches are randomly cropped from the Mel-spectrogram. Each patch is fed into the backbone and a global average pooling (GAP) module to obtain patch feature $p$. In SEDT, the learned event queries have different detection patterns focusing on different areas and sizes related to different sound events [8]. However, during pre-training in SP-SEDT, the event queries would fail to learn those latent properties of patches, which are cropped randomly. To preserve the specialization of queries, the information of patches needs to be injected into the queries to guide the query slot to detect the corresponding patch. Specifically, for an SP-SEDT model with $M$ randomly cropped patches and $N$ event queries, each patch feature $p$ will be duplicated and added to $\frac{M}{N}$ queries to obtain the final input of decoder. It is noteworthy that although the patch feature is fed into the decoder, it doesn’t contain the position information of the patch in the original spectrogram. The SP-SEDT has to locate the patch by itself.

For each patch, there are three subtasks for SP-SEDT to complete, including classification, localization and feature reconstruction. For classification, a candidate patch is classified as “patch” ($c_i = 1$) or “background” ($c_i = 0$). The localization subtask aims to predict temporal boundaries. As for feature reconstruction, it is introduced to balance classification and localization during pre-training. We hope the event representation $\hat{p}$ from the decoder can contain as much category information as the patch feature $p$ extracted by the backbone from the spectrogram. Hence, the feature reconstruction loss is formulated as the mean squared error between the L2 normalized $p$ and $\hat{p}$:

$$L_{rec} = \frac{1}{2} \| p - \frac{\hat{p}}{\| \hat{p} \|_2} \|_2^2$$

The loss function of classification, localization and feature reconstruction are linearly combined to jointly optimize SP-SEDT, where localization and reconstruction loss will only be computed for predictions labeled as “patch” ($c_i = 1$):

$$L = L_{loc} + \frac{1}{2} (c_i = 1) (L_{rec} + L_{loc})$$

3.2. Backbone pre-training

The random patch detection pre-training optimizes the model towards better localization ability without considering the classification ability. To avoid the model locates better at the cost of reduced recognition performance, we attempt to preserve the category information extraction ability of the model by freezing the parameters of the backbone during pre-training. Before pre-training, we need to guarantee the backbone can extract audio features depicting the spectrogram well. Existing self-supervised methods designed for better audio representation can be adopted to pre-train the backbone. Since this paper only focuses on the localization pre-training, to simplify the training process, we tuned the backbone (ResNet-50 pre-trained on ImageNet) on the weakly labeled data via audio tagging task.

3.3. Fine-tuning

After pre-training, the parameters of backbone, transformer and the prediction FFNs for temporal boundary regression of the SP-SEDT are utilized to initialize the SEDT model, while other modules, including the audio query branch and the prediction FFNs for classification, are initialized randomly.

4. EXPERIMENTS

4.1. Experimental setup

The experiments are conducted on the DCASE2019 task4 dataset, which consists of 10 sound event categories. It can be divided into a training set and a validation set (1168 clips), the training set contains a weakly-labeled subset (1578 clips), a strongly-labeled synthetic subset (2045 clips) and an unlabeled subset (14412 clips). SP-SEDT are pre-trained on the unlabeled subset and the DCASE2018 task5 development dataset (72984 clips), and the backbone is tuned on the weakly-labeled subset before pre-training. Overall, in our experiments, we first train the backbone by the audio tagging task using the weakly labeled data, then pre-train the SP-SEDT on the unlabeled subset and the DCASE2018 task5 development set, finally fine-tune it with the weakly-labeled and synthetic subset, and report the metric on the validation set. The DCASE2019-2021 task4 dataset differs only in the synthetic subset.

As for the baselines, we use CRNN and the Transformer-based model [4] referred to as “CTrans” in this paper. The baselines are the vanilla models of the DCASE2020 1st system, which replaces RNN or Transformer encoder with Conformer encoder. The baselines output frame-level results and are then binarized and smoothed by class-adaptive thresholds and median filters. The baselines are trained on the weakly-labeled and synthetic subset of the DCASE2019 task4 dataset, the same as the SP-SEDT fine-tuning data.

We calculate the event-based measure (200ms collars on both onsets and offsets) and segment-based measure with a segment length of 1s by the sed_eval package [22] to evaluate the detection performance of models. The clip-level macro F1 score is also reported to measure the audio tagging performance.

4.2. Results

The results on the DCASE2019 task4 validation set are given in Table I. We experiment with two model configurations that all adopt 3 transformer decoder blocks but differ in the number of transformer encoder blocks $E$. SP-SEDT ($E=6$) achieves the best results in all metrics. Among them, the SED main evaluation metric, i.e. event-based F1 score, gets the highest improve-
Ablation studies using the SP-SEDT.

EB denotes Event-based macro $F_1$ respectively.

| Model       | Eb[%] | Sh[%] | At[%] |
|-------------|-------|-------|-------|
| CRNN [5]    | 30.61 | 62.21 |       |
| CTrans($E = 3$) [5] | 34.27 |     65.07 |   |
| CTrans($E = 6$) [5] | 34.28 |     64.33 |   |
| SEDT($E = 3$) | 31.39 | 58.12 | 65.09 |
| SEDT($E = 6$) | 38.85 | 64.10 | 73.01 |
| SEDT($E = 6$) | 31.64 | 61.18 | 69.11 |
| SEDT($E = 6$) | 39.03 | 64.54 | 72.29 |

Table 2: Ablation study on loss functions and patch size, where Eb denotes Event-based macro $F_1$ respectively.

Table 1: Performance on DCASE2019 Task4 validation set, where Eb, Sh, and At denotes Event-based, Segment-based, and Audio tagging macro $F_1$ respectively.

| Rank (year) | Model (#param) | Semi-supervised method | ME | Val | Eval |
|-------------|---------------|------------------------|----|-----|------|
| 1st (2019)  | CNN (7M)      | Guided learning        | ✓  | 45.3 | -    |
| 1st (2019)  | Conformer (17M)| Mean teacher           | ✓  | 50.6 | 55.7 |
| 1st (2019)  | CRNN (20M)    | Pseudo-labelling       | ✓  | 48.3 | 50.9 |
| ours        | SP-SEDT (36M) | Pseudo-labelling       | X  | 49.3 | 52.3 |

Table 3: System performance on the DCASE2019 task4 validation set and public evaluation set, where ME denotes model ensembling. Eb denotes Event-based macro $F_1$.

Experiments above only use the weakly labeled and synthetic subsets to compare among different model structures. However, in the DCASE challenges, semi-supervised methods are adopted to train models with unlabeled data, and other techniques such as model ensembling and data augmentation are also utilized. To compare with the top systems of the DCASE task4, we construct a SP-SEDT-based system, which uses pseudo-labelling, a simple method, for semi-supervised learning and perform mixup [26] and SpecAugment [27] for data augmentation. To be fair, the SP-SEDT is pretrained only on the DCASE2019 unlabeled data. Table 3 gives the results on the validation and public evaluation set. As can be seen, our system can achieve competitive performance. Considering that model ensembling and semi-supervised method adaptive to SP-SEDT haven’t been further explored, we think SP-SEDT is of great potential.

Since SP-SEDT has to locate and recognize at the same time, it owns higher model complexity, while we argue this means it can be more adaptive to pre-training. In the future, we would pretrain SP-SEDT in AudioSet [28], a large-scale dataset of manually-annotated audio events, to fully play the role of large models. Besides, our purpose is not entirely to surpass the SOTA system, but to demonstrate that SP-SEDT, a new and promising model architecture, can also be improved through its unique pre-training mechanism to achieve better results.

5. Conclusion

In this paper, we proposed a self-supervised pre-training method for sound event detection transformer to improve its localization ability by unlabeled data. Experimental results show that the proposed method improves the performance of SEDT and outperforms the frame-based models on the DCASE2019 task4 dataset. The pre-training method also allows training a deeper model on limited labeled data to obtain further gains.

6. References

[1] M. Crocco, M. Cristani, A. Trucco, and V. Murino, “Audio surveillance: A systematic review,” ACM Computing Surveys
Y. Huang, L. Lin, X. Wang, H. Liu, Y. Qian, M. Liu, and K. Ouchi,
J. Salamon, D. MacConnell, M. Cartwright, P. Li, and J. P. Bello,
L. Lin, X. Wang, H. Liu, and Y. Qian, “Guided learning
M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and
K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning
J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-
training of deep bidirectional transformers for language understanding,”
arXiv preprint arXiv:1810.04805, 2018.
K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for
image recognition,” in 2016 IEEE Conference on Computer
Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.
M. Caron, I. Misra, J. Maire, P. Goyal, P. Bojanowski, and
A. Joulin, “Unsupervised learning of visual features by contrast-
ing cluster assignments,” in Advances in Neural Information
Processing Systems, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 9912–9924.
[18] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, “wav2vec 2.0:
A framework for self-supervised learning of speech representa-
tions,” in Advances in Neural Information Processing Systems,
H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 12 449–12 460.
[19] S. Deshmukh, B. Raj, and R. Singh, “Improving Weakly
Supervised Sound Event Detection with Self-Supervised Auxiliary
Tasks,” in Proc. Interspeech 2021, 2021, pp. 596–600.
[20] T. Lee, T. Gong, S. Palhy, A. Rouditchenko, and A. Ndiarango,
“Label-efficient audio classification through multitask learning
and self-supervision,” in Proc. Int. Conf. Learn. Representations
Workshop, 2019, p. 1–10.
[21] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N.
Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” in
Advances in neural information processing systems, 2017, pp. 5998–6008.
[22] A. Mesaros, T. Heititolta, and T. Virtanen, “Metrics for polyphonic
sound event detection,” Applied Sciences, vol. 6, no. 6, p. 162, 2016.
[23] L. Lin and X. Wang, “Guided learning convolutional system for
dase 2019 task 4,” Institute of Computing Technology, Chinese
Academy of Sciences, Beijing, China, Tech. Rep., June 2019.
[24] K. Miyazaki, T. Komatsu, T. Hayashi, S. Watanabe, T. Toda, and
K. Takeda, “Convolution-augmented transformer for semi-
supervised sound event detection,” DCASE2020 Challenge, Tech.
Rep., June 2020.
[25] J. Ebbers and R. Haeb-Umbach, “Convolutional recurrent neural
networks for weakly labeled semi-supervised sound event detec-
tion in domestic environments,” DCASE2020 Challenge, Tech.
Rep., June 2020.
[26] Z. Zhang, S. Xu, S. Cao, and S. Zhang, “Deep convolutional neu-nal network with mixup for environmental sound classification,”
in Pattern Recognition and Computer Vision. Springer, 2018,
pp. 356–367.
[27] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk,
and Q. V. Le, “SpecAugment: A Simple Data Augmentation
Method for Automatic Speech Recognition,” in Proc. Interspeech
Recog., June 2017, pp. 776–780.