TRANSFORMER-BASED SOUND EVENT DETECTION SYSTEM FOR DCASE2024 TASK4

Technical Report

Pengfei Cai

University of Science and Technology of China,
cqi525@mail.ustc.edu.cn

Yan Song

University of Science and Technology of China,
songy@ustc.edu.cn

ABSTRACT

In this technical report, we describe our systems for DCASE 2024 Challenge Task4. Our systems are mainly based on MAT-SED, a pure Transformer-based SED model with masked-reconstruction based pre-training. In MAT-SED, a Transformer with relative positional encoding is first designed as the context network instead of RNNs. The Transformer-based context network is pre-trained by the masked-reconstruction task on all available target data in a self-supervised way. Both the encoder and the context network are jointly fine-tuned in a semi-supervised manner. Our final systems achieve PSDS1 of 0.588(single model) and 0.600(ensemble) on the validation set of DESED dataset.

Index Terms— sound event detection, transformer, masked-reconstruction

1. INTRODUCTION

Most recent sound event detection(SED) architecture can generally be divided into an encoder network and a context network. In classical CRNN based SED systems [1], convolutional neural networks (CNNs) are used as the encoder network for feature extraction, while recurrent neural networks (RNNs) are employed as the context network to model temporal dependencies across latent features from the encoder. Recently, Transformer-based SED models have surged in popularity, inspired by the successes of Transformers in various domains, including natural language processing [2, 3], computer vision [4] and automatic speech recognition [5, 6]. A widely used approach is to employ Transformer models pre-trained on readily available large-scale audio tagging datasets, such as AudioSet [7], to serve as powerful feature extractors. Among high-ranking models [8, 9] of DCASE2023, the pre-trained Transformer and the CNN are concatenated in parallel as the encoder network, which can take the advantages of global and local features from different encoders. However, most of those works only applied Transformer structures partially to the traditional CRNN, which limits the ability of the whole system.

In this year's challenge, we use a pure Transformer-based SED model, named Masked Audio Transformer for Sound Event Detection (MAT-SED). MAT-SED begins with the pre-trained Transformer model as an encoder network, then a Transformer with relative positional encoding instead of RNNs as the context network, which can better capture long-range context dependencies of latent features. We use the masked-reconstruction task to pre-train the context network in the self-supervised manner, then fine-tune the pre-trained model with the classical mean teacher algorithm. This training paradigm maximizes the utilization of large quantities of unlabeled data compared to pure semi-supervised learning. Experimental results on the DCASE2024 validation dataset show that the proposed MAT-SED achieves PSDS1 of 0.588 (single model) and 0.600 (ensemble) [1].

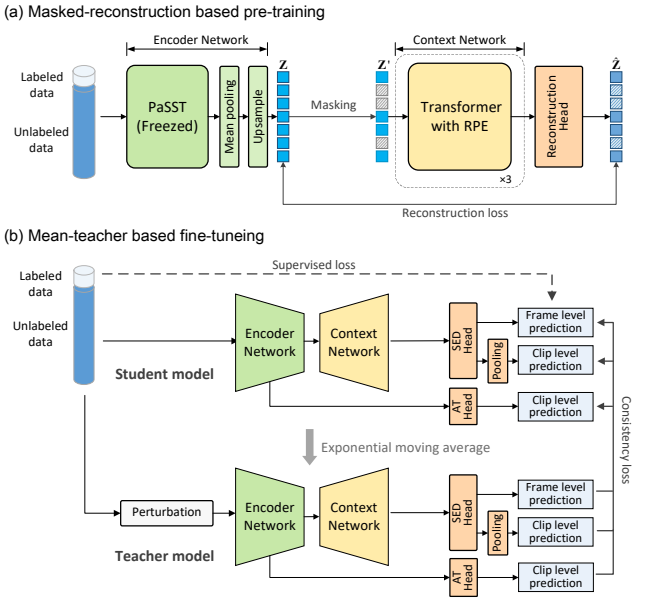

Figure 1: The architecture of our model, comprising two main components: the encoder network (green) and the context network (yellow), both of which are based on Transformer structures.

2. METHODOLOGY

In this section, we first outline the model structure, then introduce the masked-reconstruction based pre-training and the fine-tuning strategies.

2.1. Model Structure

The overall structure of MAT-SED, as shown in Figure 1, consists of two main components: the encoder network and the context network. The encoder network is used to extract features from the

[1]More details of MAT-SED are presented in our paper "MAT-SED: A Masked Audio Transformer with Masked-Reconstruction Based Pre-training for Sound Event Detection", which has been accepted by Interspeech 2024.

Table 1: Submitted systems' performances on validation set and public evaluation set of the DESED Dataset.

system	Encoder Network	Ensembled with ATST-SED	PSDS1(val.)	PSDS1(public eval.)
1	PASST	✗	0.587	0.613
2	PASST+CNN	✗	0.588	0.637
3	PASST+CNN	✓	0.600	0.655
4	PASST+CNN	✓	0.600	0.657

mel-spectrogram, outputting latent feature sequences. The context network is responsible for capturing temporal dependencies across the latent features. Different types of head layer follow the context network to handle specific tasks, such as reconstruction, audio tagging and SED.

The encoder network of MAT-SED is based on PaSST [10], a large pre-trained Transformer model for audio tagging. Each mel-spectrogram is divided into several 16×16 patches, then patches are projected linearly to a sequence of embeddings. The sequence traverses through 10 layers of PaSST blocks consisted of Transformers. The output of the encoder network is denoted as $\mathbf{Z} = [\mathbf{z_1}, \mathbf{z_2}, ..., \mathbf{z_T}] \in \mathbb{R}^{C \times T}$, where C is the dimension of the embedding vector, and T is the length of encoder's output in the time dimension.

Instead of the conventional RNN structure, we utilize 3 layers of Transformer block to constitute the context network. Given the crucial need for localization in the SED task, integrating positional information becomes vital. We use relative positional encoding (RPE) [11] to achieve this purpose, where the learnable positional encoding is determined by the relative position between frames. Compared to learnable APE, the RPE is naturally translation-equivariant [12], making it more suitable for modelling temporal dependencies.

2.2. Masked-reconstruction based pre-training

The model structure during pre-training is depicted in Figure 1 (a). At this stage, we initialize the encoder network using the PaSST model pre-trained on AudioSet [7] and freeze its weights, to focus on pre-training the context network. We design the masked-reconstruction task as the pretext task, similar to train a masked language model. We mask a certain proportion of frames in the latent feature sequence $\mathbf{Z}$, and substitute the masked frames with the learnable mask token, obtaining a new sequence $\mathbf{Z}'$. The masked-reconstruction task requires the context network to restore the masked latent features using the contextual information, which helps to enhance the temporal modeling ability of the context network. The masked sequence traverses through the context network and the reconstruction head composed of two fully connected layers, yielding the reconstructed sequence $\hat{\mathbf{Z}} = [\hat{\mathbf{z}}_1, \hat{\mathbf{z}}_2, ..., \hat{\mathbf{z}}_T] \in \mathbb{R}^{C \times T}$. We use mean squared error (MSE) loss to evaluate the quality of reconstruction:

$$\mathcal{L}_m = \sum_{x \in \mathcal{D}} \sum_{t \in M_x} (\hat{\mathbf{z}}_t(x) - \mathbf{z}_t(x))^2 \quad (1)$$

where $\mathcal{D}$ denotes the pre-training dataset, and M_x denotes the set of masked frame indices corresponding to the sample x. Note from this that only the masked frames are used to calculate the reconstruction loss.

The model structure in the fine-tuning stage is shown in Figure 1 (b). During fine-tuning, the reconstruction head is replaced by the SED head composed of a fully connected layer, which outputs the frame-level prediction. The frame-level prediction is pooled over the time dimension by linear-softmax pooling [13], to obtain the clip-level prediction result. Following the task-aware module in [14], we additionally set up an AT head to focuse on the audio tagging task. The mean-teacher algorithm [15] is used for semi-supervised learning, with the consistency weight of 40, and the sliding windows strategy is used in the encoder network of teacher model to enhance the localization capability.

Furthermore, we explored the use of Convolutional Neural Networks (CNNs) as adapters to inject specific inductive biases into a Transformer-based backbone. When incorporating CNNs alongside mask reconstruction during fine-tuning, directly adding CNN features to the audio transformer model would significantly alter the feature distribution learned during pretraining, rendering the context network incapable of correctly handling the modified, unfamiliar features. To address this, we propose the following method for injecting CNN features during fine-tuning:

$$\mathbf{F}_{merge} = \mathbf{F}_{pretrain} + \beta \mathbf{F}_{CNN}$$

Here, the weight parameter (β) is initialized to 0, ensuring that the output feature distribution of the pretrained network remains stable during the early stages of fine-tuning, despite the injection of CNN features.

3. EXPERIMENTS

3.1. Experiment setting

The input audio is sampled at 32kHz. For feature extraction, we use a Hamming window of 25ms with a stride of 10ms to perform short-time Fourier transform(STFT). The spectrum obtained by the STFT is further transformed into a mel-spectrogram with 128 mel filters. Mixup [16], time shift and filterAugment [17] are used for data augmentation.

During the pre-training phase, the model is trained over 6000 steps with a batch size of 24 and a learning rate of 1×10^{-4}. For the masked-reconstruction task, the masking rate is set to 75%. During the fine-tuning stage, batch sizes for real strongly labeled, synthetic strongly labeled, real weakly labeled, and real unlabeled data are set to 3, 1, 4, 4, respectively. Following the strategy in [18], only the SED head and AT head are trained for the first 6000 steps of fine-tuning, then the end-to-end fine-tuning is performed over the next 12000 steps. Learning rates for the encoder network, decoder network, and head layers are set to 5×10^{-6}, 1×10^{-4}, and 2×10^{-4}, respectively. The AdamW [19] optimizer is used for optimization with a weight decay of 1×10^{-4}. Training is conducted on 2 Intel-3090 GPUs for 13 hours in total.

3.2. Results

Table 1 shows the PSDS1 scores of our systems on validation set and public evaluation set. We use the validation set for hyper-parameters tuning, and the public evaluation dataset is only used for the final evaluation before submission. We also combine MAT-SED with ATST-SED [20] through model ensemble(systems 3 and 4). Specifically, we compute a weighted average of the final predictions from both models. For system 3, we evenly weight MAT-SED and ATST-SED (0.5 each), while for system 4, we adjust the weights to 0.7 for MAT-SED and 0.3 for ATST-SED.

4. REFERENCES

[1] E. Cakır, G. Parascandolo, T. Heittola, H. Huttunen, and T. Virtanen, "Convolutional recurrent neural networks for polyphonic sound event detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 6, pp. 1291–1303, 2017.

[2] 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.

[3] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 2019, pp. 4171–4186.

[4] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *9th International Conference on Learning Representations (ICLR)*, 2021.

[5] Q. Zhang, H. Lu, H. Sak, A. Tripathi, E. McDermott, S. Koo, and S. Kumar, "Transformer transducer: A streamable speech recognition model with transformer encoders and RNN-T loss," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7829–7833.

[6] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented Transformer for Speech Recognition," in *Proc. Interspeech 2020*, 2020, pp. 5036–5040.

[7] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 776–780.

[8] J. W. Kim, S. W. Son, Y. Song, . Kim, Hong Kook1, I. H. Song, and J. E. Lim, "Semi-supervised learning-based sound event detection using frequency dynamic convolution with large kernel attention for DCASE challenge 2023 task 4," DCASE2023 Challenge, Tech. Rep., June 2023.

[9] S. Xiao, J. Shen, A. Hu, X. Zhang, P. Zhang, and Y. Yan, "Sound event detection with weak prediction for dcase 2023 challenge task4a," DCASE2023 Challenge, Tech. Rep., June 2023.

[10] K. Koutini, J. Schlüter, H. Eghbal-zadeh, and G. Widmer, "Efficient Training of Audio Transformers with Patchout," in *Proc. Interspeech 2022*, 2022, pp. 2753–2757.

[11] Z. Dai, Z. Yang, Y. Yang, J. G. Carbonell, Q. V. Le, and R. Salakhutdinov, "Transformer-xl: Attentive language models beyond a fixed-length context," in *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL*. Association for Computational Linguistics, 2019, pp. 2978–2988.

[12] X. Chu, Z. Tian, B. Zhang, X. Wang, and C. Shen, "Conditional positional encodings for vision transformers," in *The Eleventh International Conference on Learning Representations*, 2023.

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

[14] K. Li, Y. Song, I. McLoughlin, L. Liu, J. Li, and L.-R. Dai, "Fine-tuning Audio Spectrogram Transformer with Task-aware Adapters for Sound Event Detection," in *Proc. INTERSPEECH 2023*, 2023, pp. 291–295.

[15] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," *Advances in neural information processing systems*, vol. 30, 2017.

[16] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *International Conference on Learning Representations (ICLR)*, 2018.

[17] H. Nam, S.-H. Kim, and Y.-H. Park, "Filteraugment: An acoustic environmental data augmentation method," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 4308–4312.

[18] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[19] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *7th International Conference on Learning Representations (ICLR)*, 2019.

[20] N. Shao, X. Li, and X. Li, "Fine-tune the pretrained atst model for sound event detection," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 911–915.