SELF-ATTENTION MECHANISM FOR SOUND EVENT LOCALIZATION AND DETECTION

Technical Report

Sooyoung Park, Youngho Jeong, Taejin Lee

Electronics and Telecommunications Research Institute, Media Coding Research Section, Daejeon, Republic of Korea, {sooyoung, yhcheong, tjlee}@etri.re.kr

ABSTRACT

This technical report describes the system submitted to DCASE 2021 Task 3: Sound Event Localization and Detection (SELD) with Directional Interference. The goal of Task 3 is to classify polyphonic events with temporal activity into a given class and detect their direction in the presence of hidden sound events. Our system uses a Transformer that utilizes the self-attention mechanism that is now successfully used in many fields. We propose an architecture called *Many-to-Many Audio Spectrogram Transformer* (M2M-AST) that uses a pure Transformer to reduce the dependency of CNNs and easily change the output resolution. Using the architecture for Sound Event Detection (SED) and Direction of Arrival Estimation (DOAE), which are small sub-problems that consist of SELD, we show that our system outperforms the baseline system.

Index Terms— Sound event localization and detection, self-attention, Transformer, transfer learning, pre-training

1. INTRODUCTION

Convolutional neural networks (CNNs) have become essential for designing deep neural networks for image understanding tasks. The translation equal variance and locality of CNNs are known to be effective for image understanding. Due to the success of CNNs in image understanding, CNNs have also been used in other pattern recognition fields [1, 2]. Since then, CNNs have been widely applied with excellent performance in various applications. CNNs are also widely used for audio understanding. However, in the field of audio understanding, Convolutional recurrent network (CRNN) [3, 4] that uses both CNNs and recurrent neural networks (RNNs) at the same time is mainly used to reflect long-term context as well as local information.

Self-attention mechanisms [5], especially Transformers, have become a new standard for natural language processing (NLP) [6]. In the field of NLP, a huge pre-trained model trained on large text corpus dataset has been released. There is a generality in this model, and this generality can be easily adapted by fine-tuning the model in small tasks. The success of self-attention in the field of NLP has led to attempts to add additional attention mechanisms to CNNs in many areas that have previously used CNNs.

Recently, *Vision Transformer* (ViT) [7, 8] using only pure Transformers for image understanding has been introduced. The outstanding performance of ViT is starting to ask whether CNNs are still essential in many applications. Since then, research on Transformers replacing CNNs has become a trend in various fields. The *Audio Spectrogram Transformer* (AST) [9] and *Keyword Transformer* [10] have been introduced as the first attempts to replace CNNs with Transformers in audio understanding. These studies demonstrate the potential of a pure Transformer to lower the dependence of CNNs in audio understanding. Inspired by the strength of the simple Transformer model in computer vision and audio classification, we propose an adaptation of this architecture to sound event localization and detection (SELD)[11].

SELD is the task of classifying polyphonic events with temporal activity into a given class and detecting their direction of arrival. Therefore, SELD can be separated into two smaller tasks: sound event detection (SED) and direction of arrival estimation (DOAE). DCASE 2021 Task 3 targets the TAU-NIGENS Spatial Sound Events 2021 dataset [12]. Unlike the TAU-NIGENS Spatial Sound Events 2020 [13], up to three target sound events can occur simultaneously, with an unknown spatial sound event in the background.

We propose a pure Transformer architecture, the *Many-to-Many Audio Spectrogram Transformer* (M2M-AST). M2M-AST can efficiently use large networks by fine-tuning huge pre-trained models. AST performs one audio classification output for single-channel audio input (one-to-one). For multi-channel audio input, we propose M2M-AST, which can have output sequences of different resolutions (many-to-many).

2. MANY-TO-MANY AUDIO SPECTROGRAM TRANSFORMER

2.1. Features

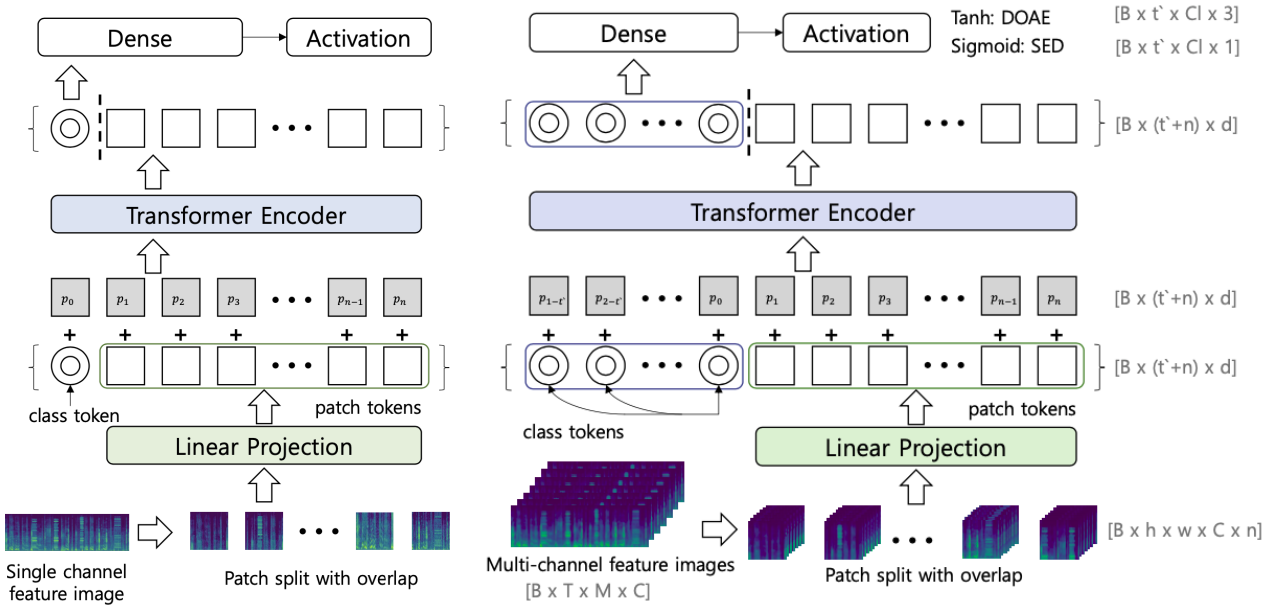
We use logmel and intensity vectors as input features [12] of the system. Our system infers SED and DOAE separately, each taking a different input. First, SED splits the channels of the microphone array into a single channel and then uses features applied by a logmel filter. DOAE uses 7-channel inputs by extracting logmel and intensity vectors from Ambisonic data. This is summarized in Table 1. Table 2 shows the pre-processing parameters to extract input features

Table 1: Feature configuration for sub-tasks

	Format	Feature	# Channels (C)
SED	Microphone Array	logmel	1
DOAE	Ambisonic	logmel, intensity vector	7

2.2. Model Architecture

Figure 1 shows the architectures of the original AST [9] and Many-to-Many AST. AST is a pure Transformer-based model. AST and



(a) AST (b) Many-to-Many AST

Figure 1: Architecture of AST and Many-to-Many AST; B: batch size, T: time, M: # mel-bins, C: channel, t': output size, n: # patchs, d: patch dimension, Cl: # class

Table 2: Pre-processing paramters

Pre-processing	
Time window length	20ms
Time window stride	10ms
Frame length (T)	200 (2s)
# Mel-bins (M)	128

Table 3: Model parameters

Model parameter	
Patch size (h x w)	16x16
Patch stride	10
# Patches (n)	228
Patch dimension (d)	768
# Encoder layer (L)	12
# Attention head	12
Output resolution (t')	20
Dropout	0.1

ViT [7, 8] consist of a similar structure, they have the same standard Transformer, same patch size, and same embedding dimensions. However, the pre-trained ViT model cannot be used directly in the AST due to the different sizes of the input feature. Therefore, AST uses a transfer learning method with weights obtained by modifying pre-trained weights. First, ViT has 3-channels of input feature, whereas AST uses a single channel feature. Thus, AST averages the weights of the linear projection layers in ViT to reuse weights in the linear projection layer in AST. Second, AST is designed to configure various input sizes, but ViT has a fixed 384x384 input format. So AST and ViT has different length of positional embeddings $[p_1; p_2; ...p_n;]$. So AST applies positional embeddings via cutting or bilinear interpolation from ViT's positional embeddings. ViT that uses a 384x384 image gets 24x24 = 576 patches when using 16x16 patch with no overlap. Besides, AST that uses 1000x128 spectrogram image, gets 100x12 patches when using the same patch size with stride 10. Therefore AST resizes the 24x24 positional embeddings to 100x12 to reuse them. When using DeiT [8] as a pre-trained model, AST adapts the average of DieT's class token and distillation token as a class token. The Transformer encoder used in AST is shown in Figure 2.

Many-to-Many AST and AST have the same structure except for the I/O structure. Many-to-Many AST has multiple class tokens because the SELD requires sequential output from multi-channel recordings. The number of class tokens can be set as much as the desired output sequence resolution. Many-to-Many AST replicates the mean values of the linear projection weights of ViT to match channel sizes so that pre-trained models can be used as in AST. For class tokens, the average value of DeiT's class token and distillation token is used as the initial value of each class token. This modification allows ViT weights to be reused in Many-to-Many AST. Table 3 shows model parameters of Many-to-Many AST.

Table 4: Model configuration

	Task	Pre-trained model	Loss
M2M-AST1	SED	DeiT	BCE
M2M-AST2	SED	M2M-AST1	soft f-loss
M2M-AST3	DOAE	DeiT	MSE
M2M-AST4	DOAE	M2M-AST3	masked MSE
M2M-AST5 (Aug)	SED	M2M-AST1	soft f-loss

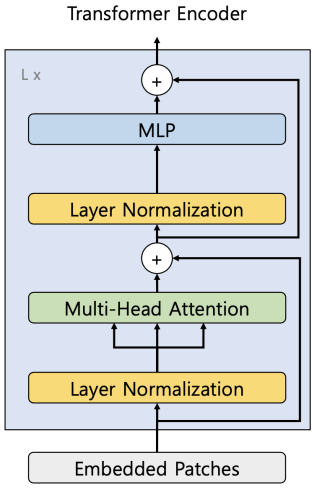


Figure 2: Transformer Encoder

2.3. Loss function

In our system, we mainly use soft f-loss [14, 15] instead of binary cross-entropy (BCE) for SED and masked mean square error (masked MSE) for DOAE. Soft f-loss is an objective function that directly uses the metric F-score. F-score is known as a non-differentiable function, but it is transformed to be differentiable and used as an objective function. The F-score function can be modified as differentiable in equation (1). Using equation (1), the soft f-loss is defined as equation (2).

$$\begin{aligned} TP(\hat{Y}, Y) &= \sum_k \hat{y}_k \cdot y_k, \\ FP(\hat{Y}, Y) &= \sum_k \hat{y}_k \cdot (1 - y_k), \\ FN(\hat{Y}, Y) &= \sum_k (1 - \hat{y}_k) \cdot y_k \end{aligned} \tag{1}$$

$$\begin{aligned} \mathcal{L}_F(\hat{Y}, Y) &= 1 - \frac{2 \cdot TP}{2 \cdot TP + FN + FP} \\ &= 1 - F \end{aligned} \tag{2}$$

2.4. Post-processing

The input time window for our system is 2 seconds. We slide this window with a small hop size to create many overlapped results and average these results during the inference [16]. Additionally, we apply median filtering and tuning the threshold for each class during SED inference. Finally, we apply a 16-way rotation augmentation to infer the test data and average the values obtained by rotating the results in reverse [16, 17].

3. EXPERIMENTS

We provide experimental results on TAU-NIGENS Spatial Sound Events 2021 development dataset. The development dataset consists of 600-minute wave files. We use 400 minutes of data for training, 100 minutes for validation, and the remaining 100 minutes for testing. Our system is trained using the hyper-parameters in Table 5. We use transfer learning with the pre-trained model. The pre-trained model used in our system is shown in Table 4. We fine-tune the SED model with 85M parameters and the DOAE model with 86M parameters for 50 epochs separately. We use the Adam optimizer. After 40 epochs, the learning rate decreases by a factor of 0.8 per epoch. We apply *Stochastic Weight Averaging* (SWA) [18] to the last 10 epochs for better results. For the development dataset, the training time consumed by the Many-to-Many AST is 6 hours for SED and 2 hours for DOAE at 4-TITAN Xp. The model mentioned in Table 4 is used for the experiment, and x0.8 and x1.2 time stretching augmentation are applied for M2M-AST5.

Table 5: Hyper-parameters for proposed system

Training	
Epoch	50
Batch size (B)	24
Learning rate	0.0001
Optimizer	Adam, SWA
Schedule	Linear decay after 40
Decay rate	0.8

3.1. Results

To test the individual performance of Many-to-Many AST, we tested SED with the ideal DOAE and DOAE with the ideal SED. Table 6 shows the experiment result for development dataset. We constructed label configuration for multi-label classification and multi-output regression, except where the same classes overlap. So, in an ideal SED situation, our system excludes co-occurrences of the same class, so LR_{CD} is 92.5%, not 100% for ideal SED condition. The performance of M2M-AST was improved by 30 to 36% in LR_{CD} performance compared to baseline [12]. In SED, soft f-score was slightly better than BCE, but there was no significant difference. In DOAE, masked MSE shows a performance improvement of about 5 degrees in ideal environments over MSE. M2M-AST2&4, which combined models for subtasks to derive SELD results, significantly outperformed baseline. Considering the performance of M2M-AST5, it was difficult to expect performance improvement using the time stretching augmentation method.

Table 6: Experimental results for development dataset

	Task	ER_{20°	F_{20°	LE_{CD}	LR_{CD}
baseline (foa)	SELD	0.69	33.9 %	24.1°	43.9 %
baseline (mic)	SELD	0.74	24.7 %	30.9°	38.2 %
M2M-AST1	SED	-	-	-	74.0 %
M2M-AST2	SED	-	-	-	**74.2 %**
M2M-AST3	DOAE	-	-	22.7°	92.5 %
M2M-AST4	DOAE	-	-	**17.5°**	92.5 %
M2M-AST5	SED	-	-	-	71.2 %
M2M-AST1&3	SELD	0.50	65.1 %	17.0°	74.0 %
M2M-AST1&4	SELD	0.46	69.0 %	**13.6°**	74.0 %
M2M-AST2&3	SELD	0.50	65.1 %	17.4°	**74.2 %**
M2M-AST2&4	SELD	**0.44**	**69.6 %**	**13.7°**	**74.2 %**

4. SUBMISSION

Based on the results in Table 6, our submission systems are based on a model using soft f-loss and masked MSE. The submission system is configured as shown in Table 7. ID 1 is the system that uses given training fold configuration of the development dataset. Therefore, ID 1 is equivalent to the model reported in Table 6. ID 2 is the system that uses all data in the development dataset. ID 3 is the system that performed a snapshot ensemble of the results of the last 3 epochs of training in addition to System ID 2. ID 4 is a system that ensembles M2M-AST5 with ID 3. All systems submitted performed ensemble method only on the SED network, and the DOAE model used a single network.

Table 7: Submission system configuration

ID	system
1	M2M-AST2&4 (Dev)
2	M2M-AST2&4 (All)
3	ID2 + M2M-AST2&4 (All, Snapshot Ensemble)
4	ID1 + ID2 + M2M-AST5 (All)

5. ACKNOWLEDGMENT

This work was supported by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIT) (No.2017-0-00050, Development of Human Enhancement Technology for auditory and muscle support).

6. REFERENCES

[1] S. Suh, S. Park, Y. Jeong, and T. Lee, "Designing acoustic scene classification models with CNN variants," DCASE2020 Challenge, Tech. Rep., June 2020.

[2] R. Giri, S. V. Tenneti, K. Helwani, F. Cheng, U. Isik, and A. Krishnaswamy, "Unsupervised anomalous sound detection using self-supervised classification and group masked autoencoder for density estimation," DCASE2020 Challenge, Tech. Rep., July 2020.

[3] Q. Wang, H. Wu, Z. Jing, F. Ma, Y. Fang, Y. Wang, T. Chen, J. Pan, J. Du, and C.-H. Lee, "The USTC-iFlytek system for sound event localization and detection of DCASE2020 challenge," *Detection and Classification of Acoustic Scenes and Events 2020 Challenge (DCASE2020)*, 2020.

[4] Y. Cao, T. Iqbal, Q. Kong, Y. Zhong, W. Wang, and M. D. Plumbley, "Event-Independent Network for Polyphonic Sound Event Localization and Detection," *Detection and Classification of Acoustic Scenes and Events 2020 Workshop (DCASE2020)*, Sept. 2020, arXiv: 2010.00140.

[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *CoRR*, vol. abs/1706.03762, 2017.

[6] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," *CoRR*, vol. abs/2005.14165, 2020.

[7] 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," *CoRR*, vol. abs/2010.11929, 2020.

[8] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," *arXiv*, Jan. 2021, arXiv: 2012.12877.

[9] Y. Gong, Y.-A. Chung, and J. Glass, "AST: Audio Spectrogram Transformer," Apr. 2021.

[10] A. Berg, M. O'Connor, and M. T. Cruz, "Keyword Transformer: A Self-Attention Model for Keyword Spotting," *arXiv*, Apr. 2021.

[11] A. Politis, A. Mesaros, S. Adavanne, T. Heittola, and T. Virtanen, "Overview and evaluation of sound event localization and detection in dcase 2019," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 684–698, 2020.

[12] A. Politis, S. Adavanne, D. Krause, A. Deleforge, P. Srivastava, and T. Virtanen, "A dataset of dynamic reverberant sound scenes with directional interferers for sound event localization and detection," *arXiv preprint arXiv:2106.06999*, 2021.

[13] A. Politis, S. Adavanne, and T. Virtanen, "A dataset of reverberant spatial sound scenes with moving sources for sound event localization and detection," in *Proceedings of the Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE2020)*, November 2020.

[14] S. Park, Y. Jeong, and T. Lee, "Metric optimization for sound event localization and detection," in *2020 IEEE International Conference on Consumer Electronics - Asia (ICCE-Asia)*, 2020, pp. 1–4.

[15] T. Tanaka and T. Shinozaki, "F-measure based end-to-end optimization of neural network keyword detectors," in *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2018, pp. 1456–1461.

[16] K. Shimada, N. Takahashi, S. Takahashi, and Y. Mitsufuji, "Sound event localization and detection using activity-coupled cartesian doa vector and rd3net," DCASE2020 Challenge, Tech. Rep., July 2020.

[17] L. Mazzon, M. Yasuda, Y. Koizumi, and N. Harada, "Sound event localization and detection using foa domain spatial augmentation," DCASE2019 Challenge, Tech. Rep., June 2019.

[18] P. Izmailov, D. Podoprikhin, T. Garipov, D. P. Vetrov, and A. G. Wilson, "Averaging weights leads to wider optima and better generalization," *CoRR*, vol. abs/1803.05407, 2018.