Many-To-Many Audio Spectrogram Transformer: Transformer for Sound Event Localization and Detection

Sooyoung Park, Youngho Jeong, Taejin Lee   Media Coding Research Section, ETRI   sooyoung@etri.re.kr

Electronics and Telecommunications Research Institute

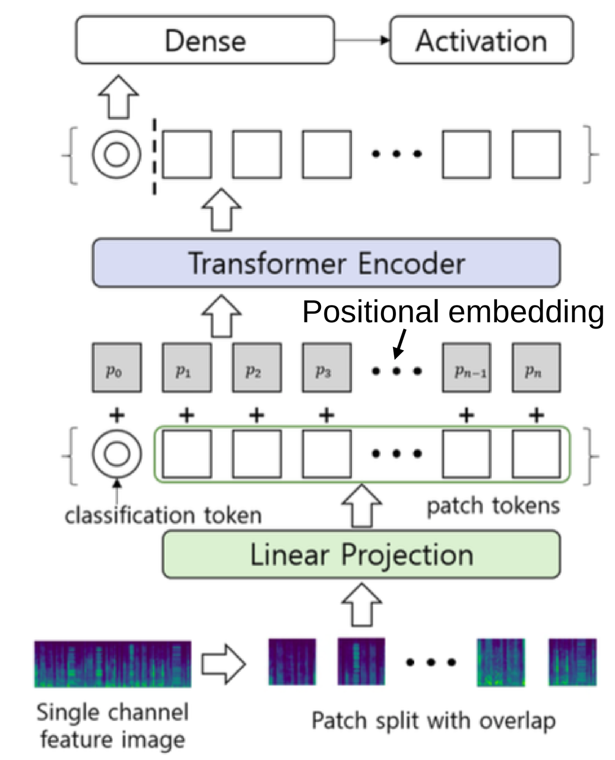
Abstract

Over the past few years, convolutional neural networks (CNNs) have been established as the core architecture for audio classification and detection. Recently, Transformers, which are pure attention-based architectures, have achieved excellent performance in various fields, showing that CNNs are not essential. In this paper, we investigate the reliance on CNNs for sound event localization and detection by introducing the Many-to-Many Audio Spectrogram Transformer (M2M-AST), a pure attention-based architecture. We adopt multiple classification tokens in the Transformer architecture to easily handle various output resolutions.

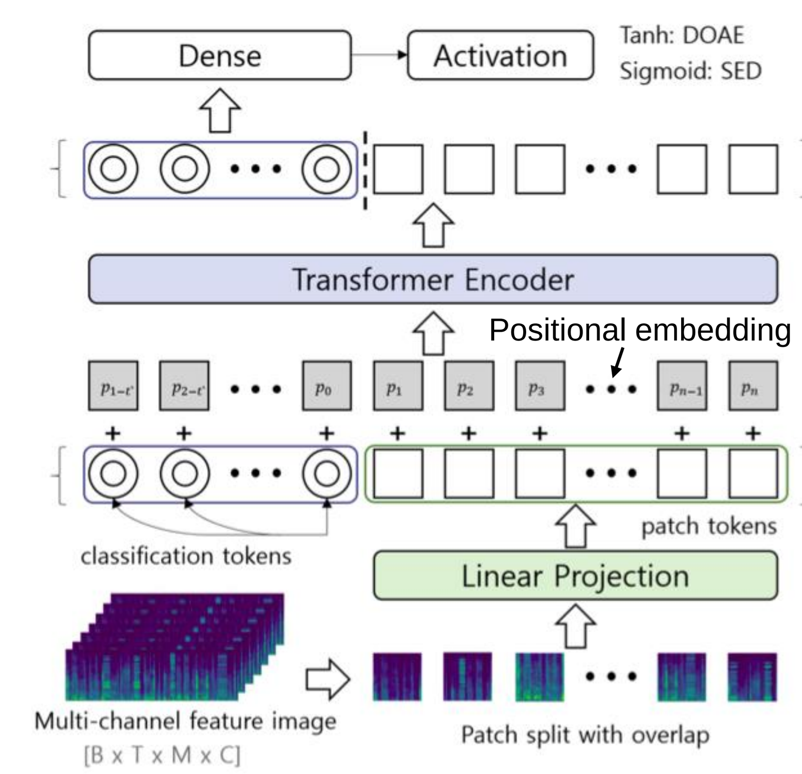
Proposed Method

Many-to-Many Audio Spectrogram Transformer (M2M- AST)

- M2M-AST focus on applying standard Transformer architecture for SELD
- Audio Spectrogram Transformer (AST) [3]
  - **Patch embedding (token)** is extracted from small image patch through linear projection
  - **Classification token** is an extra learnable embedding to perform classification
  - **Positional embedding** is a learnable embedding to make spatial information between patches
- Difference compared to AST:
  - **Multi-channel feature image** images are required to obtain spatial location information
  - **Multiple classification tokens** are used to make a series of output rather than single output



Audio Spectrogram Transformer

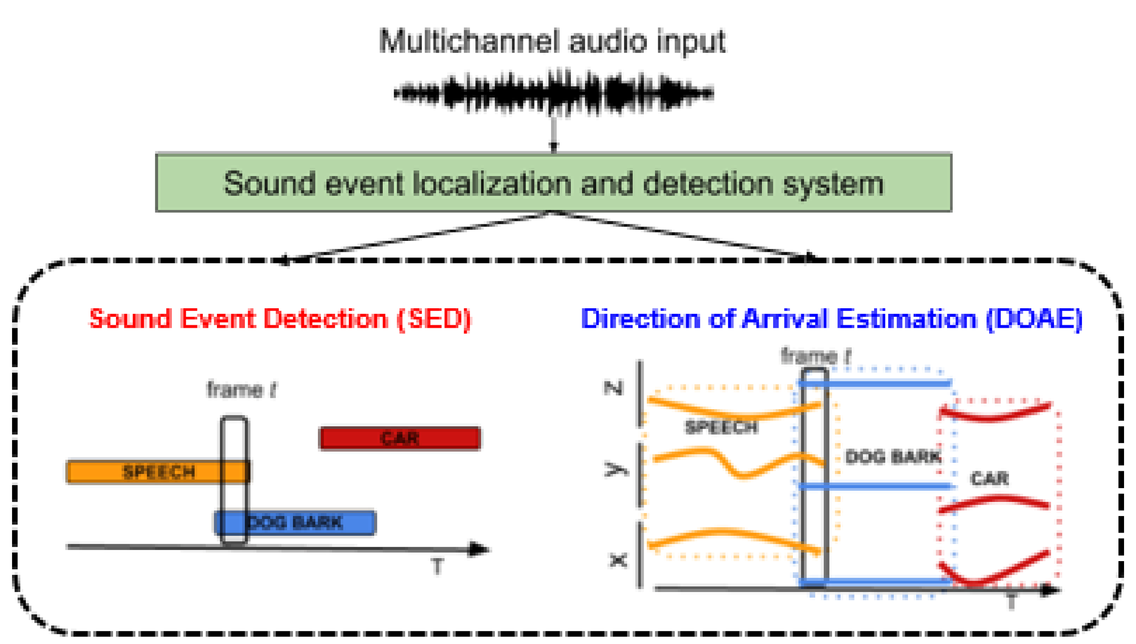


Many-to-Many AST

DCASE Challenge Task 3

Sound Event Localization and Detection

- SELD recognizes the sound event and its direction simult aneously
- Input:
  - Directional microphone recordings from a tetrahedral array
  - First-order Ambisonic (FOA) recordings
- Output:
  - Active sound event
  - Onset/offset
  - Spatial location



System overview of sound event localization and detection

Result & Ablation Study

Feature and label configuration

	Format	Feature	# Channels (C)	Label
SED	Microphone array	Logmel	1	Multi label binarization
DOAE	Ambisonic	Logmel, intensity vector	7	Cartesian coordinate (xyz)

Model configuration

	Task	Pre-trained model	Loss
M2M-AST1	SED	DeiT	BCE
M2M-AST2	SED	M2M-AST1	soft f-loss [1]
M2M-AST3	DOAE	DeiT	MSE
M2M-AST4	DOAE	M2M-AST3	masked MSE

Experimental results for dev

	# Params	$ER_{20°}$	$F_{20°}$	LE_{CD}	LR_{CD}
CRNN (Baseline FOA)	0.5M	0.69	33.9 %	24.1°	43.9 %
CRNN (Baseline-Large)	184M	0.65	45.6 %	22.6°	55.0 %
CRNN [2]	14M	0.65	48.3 %	22.0°	62.6 %
M2M-AST1&3	172M	0.55	62.6 %	17.5°	74.0 %
M2M-AST1&4	172M	0.52	64.4 %	**16.0°**	74.0 %
M2M-AST2&3	172M	0.52	64.0 %	17.7°	**74.7 %**
M2M-AST2&4	172M	**0.50**	**65.7 %**	16.3°	**74.7 %**

- All results are based on logmel energy and intensity vectors as input features
- The proposed pure transformer model outperforms the CRNN-based models listed in the table

Batch size and frame length

# Batch	SED (F_1, LR_{CD}) 1 sec	SED 2 sec	SED 3 sec (Used)	DOAE (LE_{CD}) 1 sec	DOAE 2 sec	DOAE 3 sec (Used)
24 (Used)	(68.3, 66.3)	(75.0, 73.2)	(74.0, **74.0**)	26.3°	22.2°	**21.8°**
48	(69.5, 70.9)	(75.7, 72.1)	(75.2, 73.6)	27.9°	23.1°	23.0°
96	(70.7, 70.3)	(**75.8**, 68.7)	-	27.0°	24.4°	-

Patch split with overlap

	# Patches	SED (F_1, LR_{CD})	DOAE (LE_{CD})
No Overlap	144	(71.6, 60.2)	27.3°
Overap-2	189	(73.8, 68.6)	24.6°
Overap-4	240	(74.1, 70.6)	24.1°
Overap-6 (Used)	348	(74.0, **74.0**)	**21.8°**
Overap-8	540	(**74.9**, 72.5)	**21.0°**

Output resolution

Output resolution	Output size (t')	SED (F_1, LR_{CD})	DOAE (LE_{CD})
25 ms	120	(75.3, 73.8)	22.2°
33.3 ms	90	(76.5, 75.1)	22.1°
50 ms	60	(74.4, 72.8)	22.7°
100 ms (Used)	30	(74.0, 74.0)	21.8°

Pre-training and loss function

	Pre-trained model	Loss	SED (F_1, LR_{CD})	DOAE (LE_{CD})
No pre-train (SED)	-	BCE	(60.4, 54.5)	-
ImageNet pre-train (M2M-AST1)	DeiT	BCE	(74.0, 74.0)	-
SELD pre-train (M2M-AST2)	M2M-AST1	soft f-loss	(**75.8**, **74.7**)	-
No pre-train (DOAE)	-	MSE	-	22.5
ImageNet pre-train (M2M-AST3)	DeiT	MSE	-	21.8
SELD pre-train (M2M-AST4)	M2M-AST3	masked MSE	-	**19.1**

Conclusion

In this paper, we describe how to apply the standard Transformer architecture to SELD. As a consequence, we introduce M2M-AST, a pure Transformer model for SELD. Existing SELD networks have commonly used hybrid architectures that combine CNNs with RNNs or self-attention layers. We empirically show that M2M-AST can replace these hybrid networks in SELD, SED, and DOAE. The Experimental results represent the potential of a pure Transformer to lower the reliance on CNNs in SELD. Traditional neural networks use pooling layers to change the output shape. However, due to the pooling size of this pooling layer, the output resolution cannot be configured freely. On the other hand, M2M-AST has the advantage of being able to easily design to have a variety of output resolutions.

Reference

[1] T. Tanaka et al., "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.
[2] T. N. T. Nguyen et al., "Dcase 2021 task 3: Spectrotemporally-aligned features for polyphonic sound event localization and detection," DCASE2021 Challenge, Tech. Rep., November 2021.
[3] Gong et al., AST: Audio Spectrogram Transformer, Interspeech 2021