ASSESSMENT OF SELF-ATTENTION ON LEARNED FEATURES FOR SOUND EVENT DETECTION ANDD LOCALIZATION

Parthasaarathy Sudarsanam, Archontis Politis, Konstantinos Drossos

Tampere University

Tampere University
Tampere University of Applied Sciences

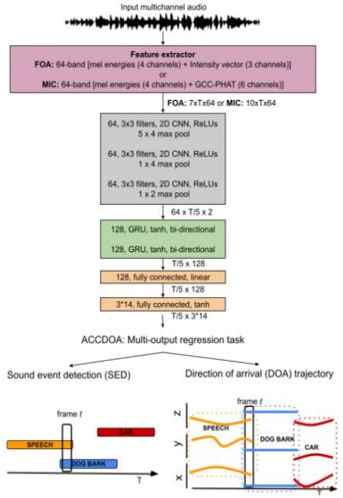
ABSTRACT

Joint sound event localization and detection (SELD) is an emerging audio signal processing task adding spatial dimensions to acoustic scene analysis and sound event detection. A popular approach to modeling SELD jointly is using convolutional recurrent neural network (CRNN) models, where CNNs learn high-level features from multi-channel audio input and the RNNs learn temporal relationships from these high-level features. However, RNNs have some drawbacks, such as a limited capability to model long temporal dependencies and slow training and inference times due to their sequential processing nature. Recently, a few SELD studies used multi-head self-attention (MHSA), among other innovations in their models. MHSA and the related transformer networks have shown state-of-the-art performance in various domains. While they can model long temporal dependencies, they can also be parallelized efficiently. In this paper, we study in detail the effect of MHSA on the SELD task. Specifically, we examined the effects of replacing the RNN blocks with self-attention layers. We studied the influence of stacking multiple self-attention blocks, using multiple attention heads in each self-attention block, and the effect of position embeddings and layer normalization. Evaluation on the DCASE 2021 SELD (task 3) development data set shows a significant improvement in all employed metrics compared to the baseline CRNN accompanying the task.

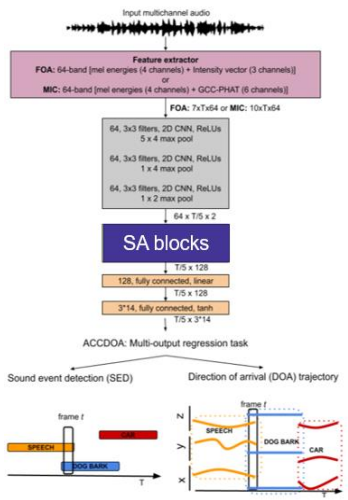
MOTIVATION

- SELD is commonly modelled using CRNN networks. CNNs learn high-level features from multi-channel audio input and the RNNs learn temporal relationships from these high-level features.
- RNNs have limited capabilities to model long temporal dependencies and slow training and inference times.
- Many computer vision and NLP tasks have successfully used self-attention mechanism to replace RNNs.
- What happens to the performance of SELD systems is we use self attention mechanism ?
- Exclusively investigate the effects of self-attention in a SELD.

BASELINE AND PROPOSED ARCHITECTURE



Baseline



Proposed model with self-attention

ACCDOA loss funtion simplifies the output representation. The model only predicts the localization. The detection probability score is the magnitude of the predicted localization vector. This value is thresholded to predict the detection activity for each class. The proposed model replaces the baseline RNNs with self-attention blocks.

HYPERPARAMETERS OF SELF ATTENTION BLOCKS

- Number of attention layers - N (2, 3).
- Number of Attention heads - M (4, 8, 12).
- Attention Size (128,64, 256).
- Position embedding – P.
- Residual connections and LayerNorm between SA layers.

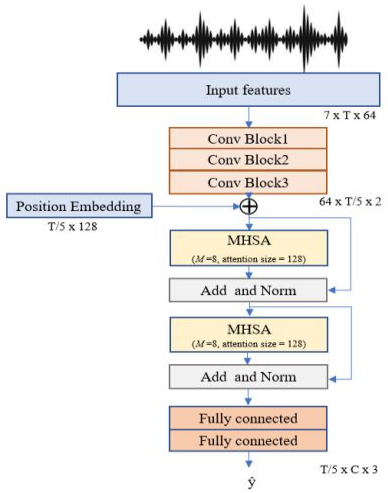
DATASET

The development set of DCASE 2021 task 3. It contains 600 one-minute audio recordings of 12 different classes.

RESULTS

N	M	P	LN	# params	ER_{20}	F_{20}	LE_{CD}	LR_{CD}
Baseline-CRNN				0.5 M	0.69	33.9	24.1	43.9
1	4	No	No	0.3 M	0.65 ± 0.01	38.11 ± 1.44	23.17 ± 0.85	46.73 ± 1.44
1	8	No	No	0.6 M	0.65 ± 0.01	39.12 ± 1.48	22.78 ± 0.73	46.71 ± 1.25
1	12	No	No	0.9 M	0.65 ± 0.01	38.96 ± 1.06	22.96 ± 0.88	46.74 ± 1.94
2	8	No	No	1.1 M	0.67 ± 0.01	36.95 ± 1.16	23.44 ± 1.27	44.66 ± 1.53
3	8	No	No	1.6 M	0.78 ± 0.02	19.57 ± 3.63	27.05 ± 0.90	22.96 ± 4.83
2	8	No	Yes	1.1 M	0.62 ± 0.01	44.62 ± 1.34	22.03 ± 0.66	55.04 ± 1.34
3	8	No	Yes	1.6 M	0.62 ± 0.01	44.11 ± 0.74	22.04 ± 0.53	54.61 ± 1.07
2	12	No	Yes	1.6 M	0.63 ± 0.01	43.95 ± 0.69	22.13 ± 0.36	54.23 ± 0.90
3	12	No	Yes	2.4 M	0.64 ± 0.01	43.10 ± 0.70	22.38 ± 0.54	54.00 ± 1.49
3 (128-256-128)*	8	No	Yes	2.2 M	0.63 ± 0.01	44.65 ± 1.88	21.98 ± 0.51	55.15 ± 1.47
3 (128-64-128)*	8	No	Yes	1.4 M	0.63 ± 0.01	43.64 ± 1.23	22.06 ± 0.46	54.24 ± 1.11
2	8	Yes	Yes	1.1 M	$\mathbf{0.61 \pm 0.01}$	$\mathbf{45.84 \pm 1.06}$	$\mathbf{21.51 \pm 0.74}$	54.99 ± 1.87
3	8	Yes	Yes	1.6 M	0.62 ± 0.01	44.63 ± 1.14	21.56 ± 0.46	54.46 ± 0.94
3 (128-256-128)*	8	Yes	Yes	2.2 M	0.62 ± 0.01	45.14 ± 1.03	21.67 ± 0.41	$\mathbf{55.29 \pm 1.23}$

BEST ARCHITECURE



CONCLUSION

We Studied the effect of self-attention for SELD task. Our study shows significant improvement in the evaluation metrics compared to the CRNN baseline. The self-attention model is also ~2.5x faster during inference than the RNN based model.