INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Event Specific Attention for Polyphonic Sound Event Detection

Harshavardhan Sundar, Ming Sun, Chao Wang
Amazon.com Inc., Cambridge, Massachusetts, United States
{sundarhs, mingsun, wngcha}@amazon.com

Abstract

The concept of multi-headed self attention (MHSA) introduced as a critical building block of a Transformer Encoder/Decoder Module has made a significant impact in the areas of natural language processing (NLP), automatic speech recognition (ASR) and recently in the area of sound event detection (SED). The current state-of-the-art approaches to SED employ a shared attention mechanism achieved through a stack of MHSA blocks to detect multiple sound events. Consequently, in a multi-label SED task, a common attention mechanism would be responsible for generating relevant feature representations for each of the events to be detected. In this paper, we show through empirical evaluation that having more MHSA blocks dedicated specifically for individual events, rather than having a stack of shared MHSA blocks, improves the overall detection performance. Interestingly, this improvement in performance comes about because the event-specific attention blocks help in resolving confusions in the case of co-occurring events. The proposed "Event-specific Attention Network" (ESA-Net) can be trained in an end-to-end manner. On the DCASE 2020 Task 4 data set, we show that with ESA-Net, the best single model achieves an event-based F1 score of 52.1 % on the public validation data set improving over the existing state of the art result.

Index Terms: Multi-Head Self Attention, Transformer, Relative Positional Encoding, Sound Event Detection, DCASE

1. Introduction

The use of attention models has been ubiquitous in the areas of natural language processing [1–5], automatic speech recognition [6, 7], computer vision [8, 9], to name a few. The major breakthrough brought about by attention models is in capturing long term temporal dependencies in sequence transduction tasks [1–3]. For a nice review on attention models used in NLP and CV tasks we refer the reader to [10, 11]. Attention based models, specifically those based on Transformer Architecture [2], are currently the state-of-the-art approaches in most of the above mentioned areas. In each of these areas attention based models, which allow for parallelization during training have successfully replaced the recurrent nets which were time consuming to train and were limited in their ability to capture long-term temporal dependencies.

Through the Detection and Classification of Acoustic Scenes and Events (DCASE) workshops and challenges [12], the area of sound events detection (SED) has attracted significant contributions from the research community over the last 5 years. The inception of DCASE challenges have streamlined the datasets used and metrics reported by researchers thereby making it possible to compare across a wide range of approaches on a common platform on real data collected in a challenging environment. In this paper, we focus mainly on one of the tasks in this challenge pertaining to sound event detection (Task 4 in 2020 edition) which involves classification and detecting the onsets and offsets of multiple target events which could possibly co-occur. Previously, convolutional-recurrent neural networks (CRNNs) were shown to be the state-of-the-art systems in this challenge [13,14]. More recently, attention models have replaced CRNNs as the new state-of-the-art [15, 16]. There are a number of studies in the NLP domain investigating the pros and cons of using attention layer activations for interpreting and explaining model decisions [17–19]. However, similar studies in SED domain are yet to be done. Consequently, it is not entirely clear as to what information the attention models are really capturing in the context of event detection.

In this paper, we first analyze a state-of-the-art SED technique - "ConformerSED" [16] to understand the information captured in the attention layers. Our analysis shows that the shared attention layers preceding the event classification layer in ConformerSED effectively capture event change points corresponding to any of the target events being detected. While this can be a useful feature in general, we show through concrete examples that such a shared attention layer also creates confusions in detecting overlapping events of different durations. To alleviate this problem, we propose to use event specific attention layers and show that the proposed "Event-Specific Attention Network" or ESA-Net in short, can consistently outperform networks with shared attention layers on DCASE 2020 Task 4 dataset. Task specific attention has been explored in the context of image recognition [20] wherein a simple soft-attention mask was used in each task branches for a multi-task classification. To the best of our knowledge this has not been explored in the context of SED. Crucially, the proposed approach applies not just to multi-task settings but also to multi-label settings with co-occurring events. In the proposed ESA-Net we explore several variants of a more sophisticated multi-head self attention (MHSA) module proposed in [2]. More specifically, we investigate the utility of sandwiching the MHSA block amidst two feed forward blocks in a Macron-style (inspired by [16,21]) and another improvement over MHSA based on using relative positional encoding (RPE) [3]. We show that the proposed ESA-Net clearly outperforms the baseline approach in multi-label scenarios with co-occurring sound events. Through visualizations, we show that the event-specific attention layers yield a non-linear decomposition of the shared attention activations wherein each event specific attention activation is less affected by the presence of other events.

2. Insights into the Shared Attention Layers of Conformer SED Architecture

Figure 1 shows the Conformer SED architecture proposed in [16]. The figure has been recreated to contrast with the proposed method (to be presented in Section 3). The "Conformer Block" shown in Figure 1 is essentially the multi-headed self-attention (MHSA) module with feed-forward and convolutional layers (c.f. [16] for further details). The attention modules are common across all events to be detected. Consequently, the common attention modules would have to learn patterns relevant to each of the events to be detected for the network to perform well. Figure 3 shows the activations at the output of

Copyright © 2021 ISCA

http://dx.doi.org/10.21437/Interspeech.2021-684

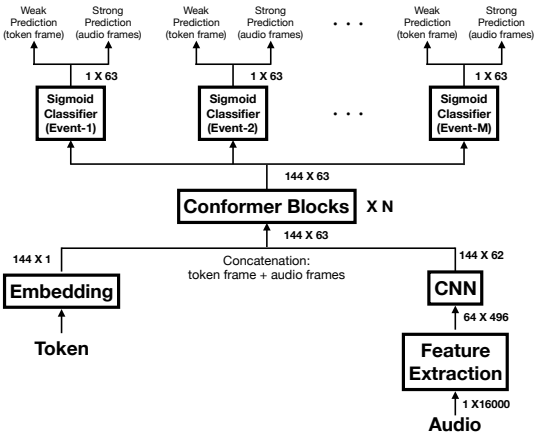


Figure 1: *The Conformer SED Architecture used in [16]. The figure has been recreated to contrast with the proposed approach. Tensor sizes at the output of each block is also shown.*

the top two Conformer Blocks along with the final frame-wise predictions (strong) of the model and the ground truth labels for an example audio containing multiple simultaneously active acoustic events. We make the following observations from the figure:

- In segments where the Conformer block activations remain stationary (quasi) the strong predictions also remain stationary.
- When the activations of the Conformer blocks change, the strong predictions also change denoting either an onset or an offset for one of the target events.

Since a common attention module serves as a feature extractor for all the target events, it has to encode not just the features relevant for multi-label classification but also the change points denoting onset and offset for each target event. We hypothesize that because of this shared attention module, the activations from the conformer block changes even if one of the target events has an onset or an offset. Therefore, it is possible that a longer duration event could adversely affect the event boundaries of a shorter duration event or vice-versa depending on the amount of training data available for each events and the structure of the event itself. As evidenced in Figure 3, we see that the model does well in accurately detecting one of the events, "Speech" in this case, while the strong predictions of the other event (Frying) are smoothed out across time. In this case, although the weak label prediction is accurate, the strong label prediction is not.

3. Event Specific Attention Layers

In order to minimize the effect of co-occurring target events on one another, we propose to have an event specific attention block for each of the target events as shown in Figure 2. This architecture would promote learning event specific aspects from the features generated by the shared Conformer Blocks. The event specific activations thus generated can essentially be considered as a non-linear decomposition of the shared activations from the Conformer Blocks. We refer to this network as the "ESA-net". In this paper, we explore four variants of MHSA mechanism for the Event Specific Attention Block as shown in Figure 4. The first variant is the traditional Transformer style MHSA block (Figure 4 a)), referred to as "T-MHSA". The second variant, referred to as "T-RMHSA", is similar to the original Transformer encoder module [2] with the

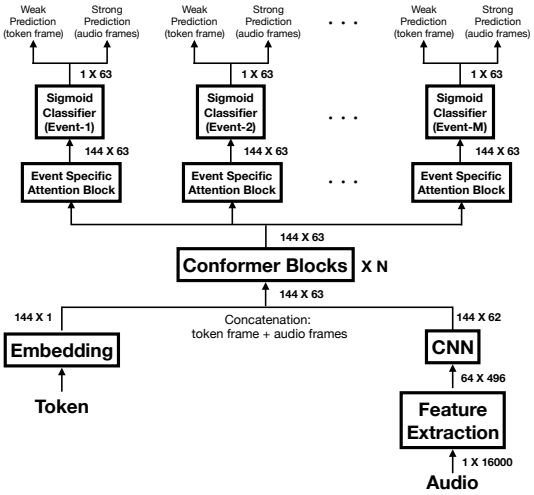


Figure 2: *The proposed Event-Specific Attention network (ESA-net). Each event has a MHSA module to minimize the effect of co-occurring events.*

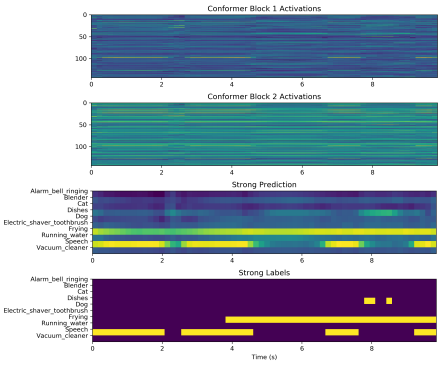


Figure 3: *Activations of the top two Conformer blocks shown in the first 2 rows for polyphonic audio with multiple simultaneously active sound events. The strong Prediction of the model and the ground truth labels are shown in 3rd and 4th rows respectively. The shared Conformer Block activations are sensitive to onset and offset of each of the target events.*

MHSA block replaced by MHSA with relative positional encoding (RPE) which has been shown to be effective in capturing longer-term dependencies in fixed-context settings [3]. The third variant is inspired by the Macron-net architecture [21] with the standard MHSA block sandwiched between two position-wise Feed-forward modules. We refer to this variant as "M-MHSA". The final variant is a combination of using MHSA with RPE in Macron-style. We refer to this variant as "M-RMHSA". The different variants are explored with the goal of understanding the importance of feed-forward layers in the attention modules and to study their utility in capturing event specific onset and offset information. The Position-wise Feed-Forward Modules used in these variants of MHSA consists of two feed-forward layers with the first layer transforming d_{model} dimensional vector to a d_{FF} dimensional vector. The second layer transforms the d_{FF} dimensional vector back to d_{model} vector.

4. Experimental Evaluation

4.1. Data Set and Training Details

We use the publicly available DCASE 2020 Task 4 Dataset [22] for all the experiments reported in this paper. The data set de-

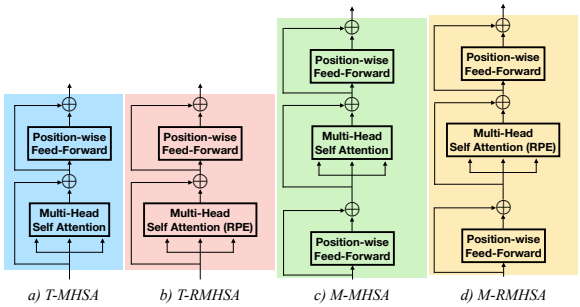


Figure 4: *Variants of MHSA explored for the "Event Specific Attention Block" shown in Figure 2 . a) Transformer-style MHSA. b) Transformer-style MHSA with RPE. c) Macron style MHSA. d) Macron style MHSA with RPE*

Table 1: *Table showing the data set stats used in training and evaluation.*

Partition	Label Type	Counts	Comments
Training	Strong Label	2,584	Synthetic
	Weak Labels	1,432	Real
	Unlabeled	12,807	Real
Validation	Strong	1,027	Real
Public	Strong	692	Real

tails used in our training and evaluation are shown in Table 1. This data set has 10 annotated classes. The class wise counts of annotated data available in different partitions are shown in Table 2. We note that the number of audio clips we obtained for our training and validation is less than those reported in [16] and we believe this is because of data deletion. We resample all the audio clips to 16kHz sampling frequency with 16 bits per sample. The pre-processing, feature extraction and post-processing schemes used were identical to that reported in [16] resulting in a feature matrix with 496 frames and 64 dimensional log-Mel spectrogram computed over windows of 64 ms width and 20 ms shift using "librosa" Python package [23]. The mean and variance normalized feature matrices are then used for training and evaluation.

We compare the proposed ESA-net with the state-of-the-art baseline approach - "Conformer SED" [16]. To ensure a fair comparison with the baseline method we design our network to have similar (and lesser) number of learnable parameters as the baseline system and also employ the same training and post processing procedure. A mean-teacher training procedure [14, 24] is employed and the whole network shown in Figure 2 is trained end-to-end with binary cross entropy loss for strong and weak labels and a mean-squared-error (MSE) based consistency loss

Table 2: *Table showing the number of annotated segments available for each target event in different partitions.*

Event	Train	Val.	Public
Alarm_bell_ringing	772	361	196
Blender	488	89	84
Cat	891	321	240
Dishes	1291	540	488
Dog	1026	520	441
Electric_shaver_toothbrush	432	56	108
Frying	383	91	90
Running_water	579	208	109
Speech	3255	1531	913
Vacuum_cleaner	500	82	96

Table 3: *Table summarizing the parameters governing the event specific attention blocks of ESA-net along with the total trainable parameters. The total no. of trainable parameters in the baseline Conformer SED model is 2,069,834.*

	d_{model}	d_{FF}	No. Attn. Heads	Total Trainable Parameters
T-MHSA	144	144	4	2,068,372
T-RMHSA	144	64	4	2,041,652
M-MHSA	144	64	4	2,026,452
M-RMHSA	144	32	4	2,045,972

between student and the mean-teacher model outputs. While the baseline system uses 3 Conformer blocks with a 144 dimensional activation and 4 heads, we use a single conformer block with the same number of dimensions and attention heads. We further reduce the number of units in the feed-forward modules from 576 used in the baseline system to 128, thus enabling more parameters to be used for event specific attention blocks. The details of parameters used in the event specific layers for different variants of the ESA-net is specified in Table 3. The baseline system was trained using publicly available code [25]. ESA-Net is implemented in PyTorch [26] and trained on NVIDIA Tesla V100 GPUs as part of AWS P3 EC2 instance [27].

4.2. Performance Evaluation

Event-based macro F1 score (EB-F1) [28] and polyphonic sound event detection score (PSDS) [29] are computed using the "sed_eval" and the "PSDS" toolkits respectively. We report performance on both "Validation" and "Public eval" (Public) datasets of DCASE 2020 Task 4. For each of the metrics we report the mean and standard deviation of the results obtained across 20 random training iterations with the same set of hyperparameters. We also report the performance of the best model we got out of the 20 random runs. For the "ConformerSED (Best)" system, we use the results obtained by using the trained weights shared by the authors [25]. To better understand the F1 scores, we also report the Precision and Recall separately for each of the 10 target events later in this section.

Table 4 shows the EB-F1 and PSDS scores on Validation and Public datasets for the ConformerSED and ESA-Net approach with different variants. ESA-net with T-RMHSA consistently outperforms the baseline in terms of averaged performance as well as the best model performance on "Validation" and "Public" partitions of the data set for both EB-F1 and PSDS metrics. The performance of ESA-net with MHSA without RPE was found to be inferior with a high variance on both "Validation" and "Public" data sets. The MHSA with RPE indeed seems to be more effective than using the traditional MHSA.

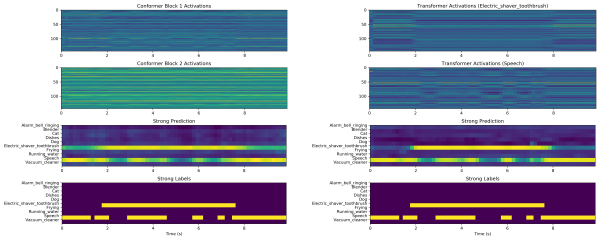
In plain sight, the improvement observed in the proposed approach appears to be incremental. To better understand the reason for improvements of the proposed approach over the baseline system we analyze the class-wise precision and recall metrics on the "Public" dataset for the best models shown in Table 4. The precision and recall expressed as a % is tabulated in Table 5. We observe from this table that for almost all the events the ESA-net with T-RMHSA approach improves over Conformer SED either on precision or on recall. For 5 of the 10 events shown in boldface blue font, ESA-Net with T-RMHSA imporves on both Precision and Recall over ConformerSED. For the events "Alarm_bell_ringing" and "Electric_shaver_toothbrush", the imporvement with ESA-net with T-RMHSA is significant. We further note that both these are longer duration events compared to some of the other events

Table 4: *Table showing the Event Based Macro F1 Score (EB-F1) and Polyphonic Sound Detection Score (PSDS) macro F1-Score for Validation and Public datasets of DCASE 2020 Task 4. Numbers reported for different techniques are averaged across 20 random runs.*

Technique	Event Branches	Validation		Public	
		EB-F1	PSDS	EB-F1	PSDS
Conformer SED [16]	Sigmoid	0.461 (0.012)	0.665 (0.016)	0.482 (0.012)	0.702 (0.013)
Conformer SED (Best) [16]	Sigmoid	0.463	0.685	0.491	**0.716**
ESA-Net	T-MHSA	0.438 (0.014)	0.651 (0.015)	0.453 (0.020)	0.687 (0.018)
	T-RMHSA	**0.469 (0.011)**	**0.676 (0.013)**	**0.491 (0.014)**	**0.706 (0.007)**
	M-MHSA	0.444 (0.015)	0.662 (0.017)	0.463 (0.015)	0.694 (0.013)
	M-RMHSA	0.465 (0.011)	0.671 (0.012)	0.481 (0.015)	0.699 (0.012)
ESA-Net (Best)	T-RMHSA	**0.478**	**0.688**	**0.521**	0.712

Table 5: *Table showing the Class-wise Precision (Prec.) and Recall (Rec.) as a %. Improvements in both Precision and Recall are highlighted in blue.*

Events	ConformerSED		ESA-Net T-RMHSA	
	Prec.	Rec.	Prec.	Rec.
Alarm_bell_ringing	52.3	29.6	**65.2**	**37.2**
Blender	48.9	54.8	51.9	47.6
Cat	72.3	66.2	79.1	64.6
Dishes	43.5	24.2	**46.3**	**27.0**
Dog	52.3	35.8	**53.9**	**37.6**
Electric_shaver_toothbrush	46.2	34.3	**68.9**	**38.9**
Frying	53.8	62.2	61.4	56.7
Running_water	52.2	33.0	42.2	39.4
Speech	67.0	52.5	67.4	53.3
Vacuum_cleaner	67.1	59.4	**71.6**	**60.4**



(a) *ConformerSED* (b) *T-RMHSA*

Figure 5: *Figure showing the activations (top 2 rows), predictions(3rd row) and ground truth labels (4th row) for the baseline and proposed approach when "Electric_shaver_toothbrush" co-occurs with "Speech". In a) the Conformer blocks activations at the output of blocks 2 and 3 are shown. In b) the event specific attention activation (output of the event specific MHSA block in Figure 2) for "Electric_shaver_toothbrush" (1st row) and "Speech" (2nd row) are shown.*

like cat and dog sounds. More specifically we observe that when these events occur in isolation both ConformerSED and the ESA-net with T-RMHSA approaches perform well. However, when these events co-occur with another event like speech, although the ConformerSED gets the weak label correctly, the event onset and offset times for detecting these longer duration events are adversely affected as shown in an example in Figure 5. Further, we believe that since the event "Speech" has several times more training data than "Alarm_bell_ringing" or "Electric_shaver_toothbrush", as seen in Table. 2, "Speech" is more often correctly detected while the same is not true for the other two events. Interestingly, we see that the activations of the ConformerSED shown in the top 2 rows on Figure 5 a) change in the temporal dimension along with onsets and offsets of any of the target events while the event specific attention for the ESA-net shown in Figure 5 b) for "Electric_shaver_toothbrush" remains unaffected by onsets and offsets of "Speech".

To further test our hypothesis that ESA-net performs better than the ConformerSED model, we measure the performance on a subset of the "Public" dataset which contains overlapping events. Out of the 692 samples in this partition (Table 2) , we found 259 instances where multiple events were overlapping with one another. In Table 6, we show the mean and standard deviation of EB-F1 score from 20 random runs of ConformerSED and ESA-Net with T-RMHSA techniques (best performing technique from Table 4) on the overlapping multi-label sounds subset of the "Public" portion of the dataset. This table clearly shows that performance of both techniques is inferior in overlapping multi-label settings. However, the use of event specific attention modules in the ESA-net with T-RMHSA approach allows it to consistently outperform the baseline ConformerSED technique in this challenging dataset.

Table 6: *Table showing the EB-F1 score on the multi-label overlapping sounds subset of the "Public" partition of the dataset.*

Technique	EB-F1
ConformerSED [16]	0.366 (0.016)
ESA-Net T-RMHSA	0.381 (0.020)

5. Conclusions

In this paper, we have shown through empirical evaluation that having event-specific or class-specific attention layers can significantly improve event detection performance over shared attention layers. Analyzing class specific performance, shows that event specific attention layers can help in detecting event onset and offset times more precisely especially in multi-label settings wherein shorter duration events co-occur with longer duration events. The downside of having more event specific attention layers is the lack of generalizability for transfer learning tasks. An interesting trade-off between having more shared attention layers and more event-specific layers could be to have duration specific attention layers which would be better suited in multi-label scenarios and can also be used for transfer learning tasks. The feasibility of such duration specific attention models warrants further research.

6. Acknowledgements

We thank the authors of the paper [16] for making their implementation [25] easily accessible.

7. References

[1] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2016, arXiv:1409.0473.

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2017, arXiv:1706.03762.

[3] Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q. V. Le, and R. Salakhutdinov, "Transformer-xl: Attentive language models beyond a fixed-length context," 2019, arXiv:1901.02860.

[4] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2019, arXiv:1810.04805.

[5] T. B. B. et.al., "Language models are few-shot learners," 2020, arXiv:2005.14165.

[6] L. Dong, S. Xu, and B. Xu, "Speech-transformer: A no-recurrence sequence-to-sequence model for speech recognition," in *Proc. Int'l Conf. Acoust., Speech and Sig. Process. (ICASSP)*, 2018, pp. 5884–5888.

[7] Y. Fujita, A. S. Subramanian, M. Omachi, and S. Watanabe, "Attention-based asr with lightweight and dynamic convolutions," 2020, arXiv:1912.11793.

[8] V. Mnih, N. Heess, A. Graves, and K. Kavukcuoglu, "Recurrent models of visual attention," in *Adv. in Neural Info. Process. Sys.*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, Eds., vol. 27. Curran Associates, Inc., 2014. [Online]. Available: https://proceedings.neurips.cc/paper/2014/file/09c6c3783b4a70054da74f2538ed47c6-Paper.pdf

[9] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," 2019, arXiv:1805.08318.

[10] S. C., V. Mithal, G. Polatkan, and R. Ramanath, "An attentive survey of attention models," 2020, arXiv:1904.02874.

[11] A. Galassi, M. Lippi, and P. Torroni, "Attention in natural language processing," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–18, 2020.

[12] Detection and Classification of Acoustic Scenes and Events (DCASE). [Online]. Available: http://dcase.community/

[13] S. Adavanne and T. Virtanen, "A report on sound event detection with different binaural features," DCASE2017 Challenge, Tech. Rep., September 2017.

[14] L. JiaKai, "Mean teacher convolution system for DCASE 2018 Task 4," DCASE2018 Challenge, Tech. Rep., September 2018.

[15] L. Lin and X. Wang, "Guided learning convolution system for dcase 2019 task 4," Institute of Computing Tech:nology, Chinese Academy of Sciences, Beijing, China, Tech. Rep., June 2019.

[16] K. Miyazaki, T. Komatsu, T. Hayashi, S. Watanabe, T. Toda, and K. Takeda, "Conformer-based sound event detection with semi-supervised learning and data augmentation," in *Proc. Workshop Detection Classification Acoust. ScenesEvents (DCASE)*, Nov. 2020, pp. 100–104.

[17] S. Jain and B. C. Wallace, "Attention is not explanation," 2019, arXiv:1902.10186.

[18] S. Serrano and N. A. Smith, "Is attention interpretable?" in *Proc. Annual Meeting of the Assoc. for Comp. Ling.* Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 2931–2951. [Online]. Available: https://www.aclweb.org/anthology/P19-1282

[19] S. Wiegreffe and Y. Pinter, "Attention is not not explanation," 2019, arXiv:1908.04626.

[20] S. Liu, E. Johns, and A. J. Davison, "End-to-end multi-task learning with attention," in *IEEE/CVF Conf. on Comp. Vision and Patt. Recog. (CVPR)*, 2019, pp. 1871–1880.

[21] Y. Lu, Z. Li, D. He, Z. Sun, B. Dong, T. Qin, L. Wang, and T. Y. Liu, "Understanding and improving transformer from a multi-particle dynamic system point of view," 2019, arXiv:1906.0276.

[22] N. Turpault, R. Serizel, A. Parag Shah, and J. Salamon, "Sound event detection in domestic environments with weakly labeled data and soundscape synthesis," in *Workshop on Detection and Classification of Acoustic Scenes and Events*, New York City, United States, October 2019. [Online]. Available: https://hal.inria.fr/hal-02160855

[23] B. McFee, C. Raffel, D. Liang, D. P. W. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *Proc. Python in Science Conference*, vol. 8, 2015.

[24] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," 2018, arXiv:1703.01780.

[25] K. Miyazaki. ConformerSED. [Online]. Available: https://github.com/m-koichi/ConformerSED.git

[26] A. Paszke and et.al., "Pytorch: An imperative style, high-performance deep learning library," in *Adv. in Neur. Info. Process. Sys.*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds. Curran Associates, Inc., 2019, pp. 8024–8035.

[27] J. Barr. Amazon EC2 P3 Instances. [Online]. Available: https://aws.amazon.com/blogs/aws/new-amazon-ec2-instances-with-up-to-8-nvidia-tesla-v100-gpus-p3/

[28] A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for polyphonic sound event detection," *Applied Sciences*, vol. 6, no. 6, p. 162, 2016. [Online]. Available: http://www.mdpi.com/2076-3417/6/6/162

[29] C. Bilen, G. Ferroni, F. Tuveri, J. Azcarreta, and S. Krstulovic, "A framework for the robust evaluation of sound event detection," *arXiv preprint arXiv:1910.08440*, 2019. [Online]. Available: https://arxiv.org/abs/1910.08440