INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Memory Controlled Sequential Self Attention for Sound Recognition

Arjun Pankajakshan[1], *Helen L. Bear*[1], *Vinod Subramanian*[1], *Emmanouil Benetos*[1,2]

[1]School of EECS, Queen Mary University of London, UK
[2]The Alan Turing Institute, UK

{a.pankajakshan, h.bear, v.subramanian, emmanouil.benetos}@qmul.ac.uk

Abstract

In this paper we investigate the importance of the extent of memory in sequential self attention for sound recognition. We propose to use a memory controlled sequential self attention mechanism on top of a convolutional recurrent neural network (CRNN) model for polyphonic sound event detection (SED). Experiments on the URBAN-SED dataset demonstrate the impact of the extent of memory on sound recognition performance with the self attention induced SED model. We extend the proposed idea with a multi-head self attention mechanism where each attention head processes the audio embedding with explicit attention width values. The proposed use of memory controlled sequential self attention offers a way to induce relations among frames of sound event tokens. We show that our memory controlled self attention model achieves an event based F-score of 33.92% on the URBAN-SED dataset, outperforming the F-score of 20.10% reported by the model without self attention.

Index Terms: Memory controlled self attention, sound recognition, multi-head attention.

1. Introduction

Sound event detection (SED) [1] is the task of automatic transcription of sound event tags with onset and offset positions from audio sequences. The essential architectural block of a deep neural network based SED model is the convolutional recurrent neural network (CRNN) [2]. The convolutional layers extract frame level features that are invariant to local spectral and temporal variations. The frame level features are sequentially processed by the recurrent layers to model relations among frames within the input sound sequence. However standard recurrent neural networks (RNNs) have two drawbacks. Firstly, in RNNs the recursive state update is performed in a first order Markov manner, which lacks an adaptive memory control mechanism. To explain this, long term memory is required when there exist relations among sound events at distant positions in long sequences. On the other hand, to process shorter sequences and in the case when relations among sound events are not certain, long term memory is not needed. The frame level audio features given to the recurrent layers are highly correlated over time, consequently the recursive state updates without adaptive memory control may result in an improper summarisation of the sound event sequence. Another undesired property of RNNs is the lack of a mechanism for modeling relations between audio frames in a sound event sequence. In sound recognition this omission, added with the fact that sound event sequences lack inherent structure, is a big limitation in sound event sequence modeling.

AP is supported by a QMUL Principal's studentship. EB is supported by a Turing Fellowship. This work has received funding from the EU's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 765068.

Attention mechanisms address these RNN limitations and have become an intrinsic part of neural network models in various tasks such as neural machine translation (NMT) [3, 4, 5], machine reading [6], image captioning [7], image synthesis [8], and speech recognition [9, 10]. *Self attention* [5, 6] is an attention mechanism that models the relations within a single sequence to compute a better summarisation of the sequence. When recurrence relations are persistent throughout a sequence (regardless of the dimension of the feature embedding i.e., the feature embeddings are either in the form of event-level tokens or frame-level tokens), then any choice of attention width induces relations within the sound event sequence. However, recurrence relations are uncertain in sound event sequences, so we assume that long-term memory is not needed in sound event sequence modeling. Therefore we propose *memory controlled self attention* to learn better latent representations of sound event sequences.

Sequential attention mechanisms [3] jointly translate and align words using *global* or *soft* attention. This is when all the encoder hidden states with different attention weights are used to predict the decoder output at each timestep. Luong et al. [4] proposed *local* attention that selectively focuses on a small context window of the encoder to predict decoder outputs. In the context of self attention, *local* attention and our memory controlled attention are the same. *Sequential self attention* [6] has successfully been applied to machine reading, using a memory network with non-Markov recursive state updates. The attention function is more generally described as mapping a *query* and set of *key-value* pairs to an output [5]. The set of keys and values define the extent of the memory used for attention. To the best of our knowledge, none of these works have analysed the extent of memory on attention performance for the respective tasks. We assume that the extent of memory (attention width) is not influential in the context of speech and text data because of the persistent relations between word tokens in these data sequences added with the auto-regressive modeling power of these models.

Attention mechanisms have been used for sound recognition; for example in temporal attention for audio tagging [11, 12], attention and localization are used to quantify sound events at each audio frame. Kong et al. [13, 14] proposed an attention model for multiple instance learning (MIL) applied to audio classification. In SED, Wang et al. [15] applied self attention mechanisms based on *transformer attention* [5]. Again, the authors have not investigated the impact of the extent of memory (key-value selection) on attention performance. Interestingly, their work shows that overall detection does not improve with the self attention mechanism. But also, their self attention implementation improved the detection performance for some long duration sound events. This indicates a need for memory controlled self attention in sound recognition.

In this paper, we evaluate the potential of memory controlled sequential self attention for sound event detection; we

http://dx.doi.org/10.21437/Interspeech.2020-1953

also propose a methodology to quantify a range of attention width values to summarise each audio frame embedding using *multi head self attention*. To the best of our knowledge, there has not been work exploring the use of sequential self attention mechanisms for SED. The rest of this paper provides a description of memory controlled self attention, our multi head attention proposal, followed by the experimental details, results and discussion.

2. Motivation

By comparing various aspects of sequence modeling of audio signals with sequence modeling in natural language processing (NLP), in this section we aim to show that memory controlled self attention is an appropriate choice for sound event sequence modeling.

- Similar to speech and music signals, sound event sequences belong to the class of structured sequence data; however recurrence relations are uncertain in sound event sequences. This means, it is not prudent to assert relations between consecutive sound events in these sequences. However, there exist temporal relations between audio frames within sound events.
- In speech and text data the relations between phonemes in a word and the relations between consecutive words in a sentence are assured (i.e., recurrence relations are persistent in text data regardless of the dimension of the feature embedding). Hence self attention with any memory width is unambiguous for general NLP applications. The language structure and the semantic relations in text data support this behaviour.
- In speech and text processing, self attention is generally applied on word level embeddings [16, 6, 3, 5]. Contrarily in sound recognition, self attention is applied to frame level embeddings. We claim that the lack of higher level event-based embeddings is the most important constraint in sound event sequence modeling. The frame level features in sound event sequences are highly correlated over time. Thus SED models without adaptive memory controlled self attention may overfit to pseudo relations based on frame level similarity patterns. This reduces the effectiveness of self attention mechanisms and lessens the recognition performance and generalizability of sound recognition models.

3. Memory Controlled Self Attention

We implement memory controlled sequential self attention on top of a CRNN model for the task of sound event detection. The architectural details of the CRNN model are described in Section 4.1. The convolutional block maps an audio input sequence of representations $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_T)$ to a sequence of feature embeddings $\mathbf{Z} = (\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_T)$, where T is the total number of audio frames. Given $\mathbf{Z}$, at each time step the recurrent layer generates hidden state representations $\mathbf{H} = (\mathbf{h}_1, \mathbf{h}_2, \cdots, \mathbf{h}_T)$. In this work, we apply the proposed memory controlled self attention layer on $\mathbf{H}$ to derive improved hidden state representations $\tilde{\mathbf{H}} = (\tilde{\mathbf{h}}_1, \tilde{\mathbf{h}}_2, \cdots, \tilde{\mathbf{h}}_T)$ prior to classification.

A self attention function on an input sequence is described as mapping each query vector of the sequence with a set of key-value vectors to obtain an output vector that summarises the query vector with respect to the key-value set. The output vector is the weighted sum of the value vectors, where the weight assigned to each value vector is computed by a similarity function of the query vector with the corresponding key vector. Using the general form of the self attention function without memory control, each bottleneck feature vector $\tilde{\mathbf{h}}_t$ is computed as:

$$\tilde{\mathbf{h}}_t = \sum_{i=1}^{T} \alpha_i^t \mathbf{h}_i; t \in \{1, \ldots, T\} \tag{1}$$

where α_i^t is the attention weight value computed using a similarity function as:

$$\begin{aligned} \alpha_i^t &= softmax(s_i^t) \\ s_i^t &= score(\mathbf{h}_t, \mathbf{h}_i); i, t \in \{1, \ldots, T\} \end{aligned} \tag{2}$$

$$score(\mathbf{h}_t, \mathbf{h}_i) = \begin{cases} \mathbf{v_a}^\intercal tanh(\mathbf{W_a}[\mathbf{h}_t; \mathbf{h}_i]) & \text{additive/concat [3]} \\ \mathbf{h}_t^\intercal \mathbf{W_a} \mathbf{h}_i & \text{general [4]} \\ \mathbf{h}_t^\intercal \mathbf{h}_i & \text{dot [4]} \\ \mathbf{h}_t^\intercal \mathbf{h}_i / \mid \mathbf{h}_t \mid\mid \mathbf{h}_i \mid & \text{scaled dot [5]} \end{cases} \tag{3}$$

where $\mathbf{v_a}$, $\mathbf{W_a}$ are the weight terms of the score functions and $\intercal$ denotes transposition.

To explain (1), the general form of self attention computes the similarity of each frame level embedding with respect to every other feature embedding in the input sequence. The similarity scores between frame level embeddings of distinct sound event tokens might be high, which results in a wrong summarisation of the input sequence. Also, as sound event tokens in audio sequences typically lack syntactic and semantic relations, long term memory is not required. However relations exist among frame level embeddings within sound event tokens, thus to effectively model these relations we propose *memory controlled self attention* by constraining the self attention function in (1) to a compact neighbourhood relative to each frame level embedding with L being the attention width.

$$\tilde{\mathbf{h}}_t = \sum_{i=(t-(L/2))}^{t+(L/2)} \alpha_i^t \mathbf{h}_i; t \in \{1, \ldots, T\} \tag{4}$$

In terms of *query*, *key*, and *value* representations we have $\mathbf{Q} = \mathbf{h}_t$, and $\mathbf{K_L} = \mathbf{V_L} = (\mathbf{h}_{t-(L/2)}, \cdots, \mathbf{h}_t, \cdots, \mathbf{h}_{t+(L/2)})$. The key-value set determines the extent of self attention. Using this we update the general form of memory controlled self attention equivalent to that of (4) as:

$$Attention(\mathbf{Q}, \mathbf{K_L}, \mathbf{V_L}) = softmax(\mathbf{Q}\mathbf{K_L}^\intercal)\mathbf{V_L} \tag{5}$$

We evaluate the impact of memory controlled self attention on sound recognition performance using the different score functions listed in (3). In preliminary experiments, we achieved the best results using the additive score function. Therefore, the results and observations included in this paper are based on the additive score function.

A limitation of the memory controlled self attention function in (4) is that it uses a fixed attention width value to summarise each frame level embedding independent of the duration of sound events. However it is better to use a small attention width value to the frames that belong to sound events that have short duration and a large attention width value to the frames associated with long duration sound events. Ideally the best memory controlled self attention design would automatically choose appropriate attention width values to summarise each frame level embedding in the input sequence. We therefore propose multi head memory controlled self attention to address this limitation.

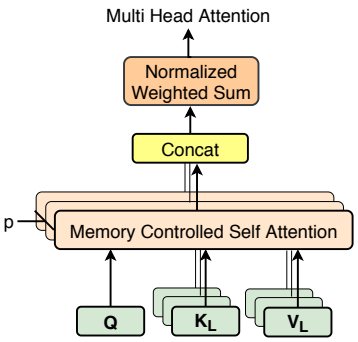

Figure 1: *Block diagram of Multi Head Memory Controlled Self Attention.*

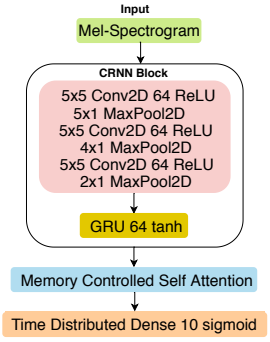

Figure 2: *Architecture of SED model.*

3.1. Multi-Head Self Attention

As an alternative to using a fixed attention width value, we propose to apply the same attention function on each query with different key-value sets. A key-value set with the corresponding attention width value leads to a memory controlled self attention head. The implementation of the multi head memory controlled self attention function is depicted in Fig. 1. The weight of each head is normalised using its corresponding attention width value. The normalised weighted sum of the attention head output values wrap up the final frame level embeddings as:

$$\begin{aligned} MultiHead(\mathbf{Q}) &= Concat(\mathbf{head}_1, \ldots, \mathbf{head}_p)\mathbf{W_{ah}} \\ \mathbf{head}_j &= Attention(\mathbf{Q}, \mathbf{K_L}_j, \mathbf{V_L}_j); j \in \{1, \ldots, p\} \\ \mathbf{W_{ah}} &= (w_1/L_1, \cdots, w_p/L_p) \end{aligned} \tag{6}$$

where p is the number of attention heads, $Concat$ denotes the concatenation of individual attention head vectors, $\mathbf{W_{ah}}$ is the normalised weight vector with w_j, L_j respectively denoting the weight and attention width values for the j^{th} head.

Comparison to transformer multi head attention [5]: Whilst our multi head self attention implementation in (6) is similar to the *Transformer multi head attention* in (7), there are a few critical differences. To the best of our knowledge, there has not been any other work exploring multi head architectures for self attention. Firstly, each of our self attention head has a corresponding key-value set that determines the extent of self attention for that head. Hence, our multi head attention approach implements a soft optimization rule to rank individual attention heads for the best summarisation of the frame level embeddings. *Transformer multi head attention* linearly projects the same key-value set with different learned weights at each attention head:

$$\begin{aligned} MultiHead(\mathbf{Q}, \mathbf{K}, \mathbf{V}) &= Concat(\mathbf{head}_1, \cdots, \mathbf{head}_p)\mathbf{W_O} \\ \mathbf{head}_j &= Attention(\mathbf{Q}\mathbf{W_Q}_j, \mathbf{K}\mathbf{W_K}_j, \mathbf{V}\mathbf{W_V}_j); \\ & \quad j \in \{1, \ldots, p\} \\ Attention(\mathbf{Q}, \mathbf{K}, \mathbf{V}) &= softmax(\mathbf{Q}\mathbf{K}^\intercal/\sqrt{d_k})\mathbf{V} \end{aligned} \tag{7}$$

where $\mathbf{W_O}$, $\mathbf{W_Q}_j$, $\mathbf{W_K}_j$, $\mathbf{W_V}_j$ are the weight matrices and d_k is the dimension of the key vector. Secondly, our multi head implementation has only a single attention layer with score function weights ($\mathbf{v_a}$ and $\mathbf{W_a}$ in (3)) and attention head weight ($\mathbf{W_{ah}}$ in (6)). *Transformer attention* [5], on the other hand, has separate attention head layers with associated weight matrices as shown in (7). Lastly, we compute attention weights using the additive score function of (3), whereas in [5] the scaled dot product score function is used. In our multi head approach, for $L_a > L_b$, $Attention(\mathbf{Q}, \mathbf{K_L}_a, \mathbf{V_L}_a)$ is a superset function of $Attention(\mathbf{Q}, \mathbf{K_L}_b, \mathbf{V_L}_b)$. This may result in biased attention head weight assignment. To counteract this effect, the attention weight of each head is scaled with the corresponding attention width value.

In this work we empirically choose $p = 11$ attention heads in the multi head self attention layer. The first attention head employs an attention width of $L = 2$ to observe the impact of immediate past and immediate future frame level embeddings to summarise the present frame. In the subsequent attention heads we serially increment the attention width value by five frames.

4. Experimental Details

We first analyse SED performance with a standard self attention function as in (1). Then we analyse the impact of memory controlled self attention (4) with different attention width values on SED performance. Lastly, we evaluate SED using the multi head memory controlled self attention function in (6).

4.1. Model architecture and Training

We use a similar version of the CRNN model architecture presented in [2] to build our SED model; Fig. 2 details the models architecture. We use a 40 log mel-bands Mel-spectrogram as input representation, extracted using a short-term Fourier transform (STFT) with an FFT window of 2048, a hop length of 882, and a sample rate of 44.1 kHz. The CRNN block has three stacked convolutional layers followed by a single gated recurrent unit (GRU) layer. We use a memory controlled self attention layer on the CRNN block feature embeddings. The SED model has a single time distributed dense layer which is the output layer of the network. The output of the model is a posteriogram matrix with dimensions $T \times C$, where T is the number of frames and C is the total number of sound event classes in the dataset. The model predictions are thresholded at 0.5 to obtain binary two-dimensional representations which are used to compute evaluation metrics based on the ground truth labels.

Each convolutional layer activation is batch normalised and regularised with dropout (probability = 0.3). The convolutional layer weights have been initialized using random normal distributions with zero mean and 0.05 standard deviation. We train the network for 200 epochs using a binary cross-entropy loss function and the Adam optimizer with a learning rate of 0.001 and a decay of 10^{-6}.

4.2. Dataset and Evaluation metrics

We train our model on the URBAN-SED [17] dataset consisting of $10,000$ soundscapes with sound event annotations generated using Scaper [17], an open-source library for soundscape synthesis. All recordings are ten seconds long, 16-bit mono and sampled at 44.1kHz. The annotations are strong, meaning for every sound event the annotations include the onset, offset, and label of the sound event. Each soundscape contains between one to nine sound events from the list {air_conditioner, car_horn, children_playing, dog_bark, drilling, engine_idling, gun_shot, jackhammer, siren and street_music} and has a background of Brownian noise. We use the URBAN-SED pre-sorted train, validation, and test sets. Of $10,000$ soundscapes, 6000 are used for training, and 2000 each for validation and test.

We use the F-score and Error Rate (ER), with F-score as the primary metric. The evaluation metrics are computed in both segment-wise and event-wise manners using the sed_eval tool [18]. Segment-based metrics show how well the system correctly detects the temporal regions where a sound event is active; with an event-based metric, the metric shows how well the system detects event instances with correct onset and offset. The evaluation scores are micro-averaged values, computed by aggregating intermediate statistics over all test data; each instance has equal influence on the final metric value. We use a segment length of one second to compute segment metrics. The event-based metrics are calculated with respect to event instances by evaluating only onsets with a time collar of 250ms.

5. Results and Discussion

Table 1 presents the SED results. Here, $Baseline$ is the SED model without self attention. $SelfAttn$ is the SED model with self attention and without memory control. $SelfAttn\_L$ is the SED model with memory controlled self attention using attention width L. $MultiHead$ is the SED model with memory controlled multi head self attention.

We see that self attention without memory control has an event-based F-score of 9.78% that is significantly lower than the baseline (20.10%) and that the best model (33.92%) uses memory controlled self attention with $L = 50$. The model with L = 100 has an F-score of 13.66%, which is lower than other memory controlled self attention models. This clearly justifies the need for proper selection of the extent of memory in order to efficiently implement self attention for SED. The inferior performance of the $SelfAttn$ model compared to the $Baseline$ model and the models with memory control is expected and is due to the reasons explained in Section 2. Also, we cannot expect a monotonic model behavior based on the attention width value. The optimum choice of attention width for each audio sample depends on the type of sound events and event durations. The event-based F-score for the $MultiHead$ model is 21.89% compared with the best model value of 33.92%. We suggest that the soft optimization rule based on the weighted sum of individual attention head representations is the main reason for the under-performance of the $MultiHead$ model.

In Fig. 3, we analyse the class-wise event based F-score. We expected best recognition performance for short sound events like car_horn, dog_bark, and gun_shot with relatively narrow attention width models ($10 < L < 50$) and for long duration sound events like $drilling$, $engine\_idling$, $air\_conditioner$, and $children\_playing$ using attention models with larger attention width values ($50 < L < 200$). However for all the sound event classes except car_horn, the memory controlled self attention model with an attention width of 50 frames yields the best recognition performance. As expected, car_horn being a short event is best recognised with a narrow attention width model (L = 10). The duration of sound events in the URBAN-SED [17] dataset is in the range 0.5–4 seconds. For the effective memory controlled self attention implementation the attention width should not be larger than the duration of short sound events in the dataset. We suggest this is the reason why the attention model with L = 50 yields the best results. Also, we assume the same reason along with the soft optimization approach for the less effectiveness of the $MultiHead$ model. Even though the overall F-score of the $MultiHead$ model is close to the $Baseline$ model, the recognition for long duration events (e.g. air_conditioner, engine_idling) is better with the $MultiHead$ model. Attention visualizations can be found online[1].

Table 1: *Sound event detection results.*

	F1 (%)		Error rate	
Model	**Segment**	**Event**	**Segment**	**Event**
Baseline	47.45	20.10	0.74	2.21
SelfAttn	29.45	9.78	0.89	1.51
SelfAttn_2	50.57	24.44	0.69	1.98
SelfAttn_10	54.36	28.62	0.64	1.68
SelfAttn_50	**55.90**	**33.92**	**0.59**	**1.12**
SelfAttn_100	33.71	13.66	0.79	1.21
MultiHead	49.28	21.89	0.69	1.77

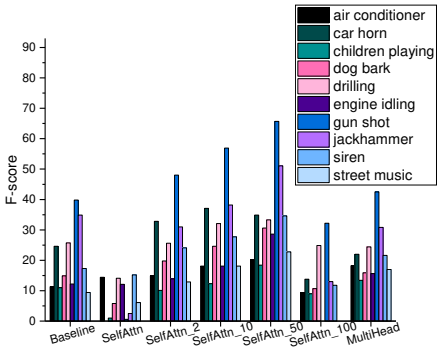

Figure 3: *Class-wise event based F-score results.*

6. Conclusion

In this work, we investigated the importance of the extent of memory on self attention, applied to the task of sound event detection. Memory controlled self attention is an effective approach to model the relations between frame-level tokens within sound events which improves temporally precise sound recognition. An explicit mapping of the extent of attention to the recurrence relations in audio sequences is a future goal. Our multi-head attention methodology for optimally selecting the extent of attention is not very successful in this work; we are inclined to extend our memory controlled $MultiHead$ model for urban sound tagging using the SONYC [19] dataset that has a wide range of coarse-grained and fine-grained event tags and also for sound recognition using AudioSet [20]. We also see the idea of using memory controlled self attention to define higher level event-based feature embeddings in sound event sequences.

[1] https://github.com/arjunp17/MemoryControlled-MultiheadSelfAtt

7. References

[1] T. Virtanen, M. D. Plumbley, and D. Ellis, *Computational analysis of sound scenes and events*. Springer, 2018.

[2] 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.

[3] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *International Conference on Learning Representations (ICLR)*, 2014.

[4] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2015.

[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems (NIPS)*, 2017, pp. 5998–6008.

[6] J. Cheng, L. Dong, and M. Lapata, "Long short-term memory-networks for machine reading," *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2016.

[7] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *Proceedings of the 32nd International Conference on Machine Learning (PMLR)*, 2015, pp. 2048–2057.

[8] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," *Proceedings of the 36th International Conference on Machine Learning (PMLR)*, 2019.

[9] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Advances in Neural Information Processing Systems (NIPS)*, 2015, pp. 577–585.

[10] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, and Y. Bengio, "End-to-end attention-based large vocabulary speech recognition," in *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2016, pp. 4945–4949.

[11] Y. Xu, Q. Kong, Q. Huang, W. Wang, and M. D. Plumbley, "Attention and localization based on a deep convolutional recurrent model for weakly supervised audio tagging," *Conference of the International Speech Communication Association (INTERSPEECH)*, 2017.

[12] C. Yu, K. S. Barsim, Q. Kong, and B. Yang, "Multi-level attention model for weakly supervised audio classification," *Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE), 19 - 20 November 2018, Surrey, UK.*, 2018.

[13] Q. Kong, Y. Xu, W. Wang, and M. D. Plumbley, "Audio set classification with attention model: A probabilistic perspective," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 316–320.

[14] Q. Kong, C. Yu, Y. Xu, T. Iqbal, W. Wang, and M. D. Plumbley, "Weakly labelled audioset tagging with attention neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 11, pp. 1791–1802, 2019.

[15] J. Wang and S. Li, "Self-attention mechanism based system for DCASE2018 challenge Task1 and Task4," in *Proc. DCASE Challenge*, 2018, pp. 1–5.

[16] W. Wu, H. Wang, T. Liu, and S. Ma, "Phrase-level self-attention networks for universal sentence encoding," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018, pp. 3729–3738.

[17] J. Salamon, D. MacConnell, M. Cartwright, P. Li, and J. P. Bello, "Scaper: A library for soundscape synthesis and augmentation," in *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2017, pp. 344–348.

[18] A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for polyphonic sound event detection," *Applied Sciences*, vol. 6, no. 6, p. 162, 2016.

[19] J. P. Bello, C. Silva, O. Nov, R. L. Dubois, A. Arora, J. Salamon, C. Mydlarz, and H. Doraiswamy, "SONYC: A system for monitoring, analyzing, and mitigating urban noise pollution," *Communications of the ACM*, vol. 62, no. 2, pp. 68–77, 2019.

[20] 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 *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 776–780.