WEAKLY LABELLED AUDIO TAGGING VIA CONVOLUTIONAL NETWORKS WITH SPATIAL AND CHANNEL-WISE ATTENTION

Sixin Hong[1] *Yuexian Zou*[1,2,*] *Wenwu Wang*[3] *Meng Cao*[1]

[1] ADSPLAB, School of ECE, Peking University, Shenzhen, China
[2]Peng Cheng Laboratory, Shenzhen, China
[3]Center for Vision, Speech and Signal Processing, University of Surrey, UK

ABSTRACT

Multiple instance learning (MIL) with convolutional neural networks (CNNs) has been proposed recently for weakly labelled audio tagging. However, features from the various CNN filtering channels and spatial regions are often treated equally, which may limit its performance in event prediction. In this paper, we propose a novel attention mechanism, namely, spatial and channel-wise attention (SCA). For spatial attention, we divide it into global and local submodules with the former to capture the event-related spatial regions and the latter to estimate the onset and offset of the events. Considering the variations in CNN channels, channel-wise attention is also exploited to recognize different sound scenes. The proposed SCA can be employed into any CNNs seamlessly with affordable overheads and is end-to-end trainable fashion. Extensive experiments on weakly labelled dataset Audioset show that the proposed SCA with CNNs achieves a state-of-the-art mean average precision (mAP) of 0.390.

***Index Terms*—** Audio tagging, weakly labelled data, multiple instance learning, spatial attention, channel-wise attention.

1. INTRODUCTION

The objective of audio tagging is to predict the presence or absence of certain sound events in an audio recording. Due to the time-consuming and costly process for labelling data, many datasets such as Audioset [1] are weakly labelled, i.e., only the classes of the audio events are annotated, while their onset/offset time are not given. Audio tagging with weakly labelled data has recently attracted increasing interest in the audio signal processing community [2, 3, 4].

Several ideas have been proposed to facilitate the prediction of the labels for weakly labelled data. One approach is based on the bag of frames assumption [5, 6], where each audio recording is divided into overlapping frames and each frame inherits the labels of the audio recording. This assumption, however, is not always satisfied when encountering short-duration events (e.g., gunshot). Another successful approach is using Multiple instance learning (MIL) [7, 8, 9] that treats frames in an audio recording as a bag of instances. In contrast to traditional supervised learning where each instance is associated with a class label, MIL considers a set of bags where multiple instances in the same bag share the same labels. A bag containing at least one positive instance is considered as a positive bag, otherwise negative. Evidently, this paradigm is more suitable for learning weak labels. In this case, such as the state-of-the-art CNN-based MIL methods [10, 11], CNNs serve as the feature extractor to learn representations for instances which are integrated into bag-level. Starting from an input spectrogram of size $W \times H \times 1$, the convolutional layer consisting of C-channel filters outputs a $W' \times H' \times C$ feature map, which will be fed to the next convolutional layer to extract frequency-shift invariant features. Therefore, the CNN features are naturally spatial, and channel-wise. In other words, there are variations across the CNN filtering channels and the spatial regions in the time-frequency representation among different layers. However, current CNN-based MIL methods treat channels and spatial regions equally for event prediction, which may contain noise or irrelevant information for the related events.

In this work, we propose a new attention mechanism in CNN, namely, spatial and channel-wise attention (SCA), for weakly labelled audio tagging. The channel-wise attention is applied to rescale the channel weights adaptively to obtain contextual event information. The spatial attention consists of global and local branches. Specifically, the global attention is used to focus on event relevant regions along both time and frequency dimensions. For the local attention, it takes evenly cropped patches of the whole feature map as input and captures important temporal information about the beginning and ending of the event. As a result, SCA can detect sound events better by learning what and where to attend in the signal. To the best of our knowledge, this is the first work that extensively explores the effect of attention using the characteristics of CNN features for audio tagging.

The paper is organized as follows. First, we give an overview of CNN-based MIL methods in Section 2. Next, we describe the proposed method in Section 3. Then, in Section 4 we detail the experimental setup and report the results. Finally, we conclude the paper.

2. CNN-BASED MIL METHODS

In this section, we briefly introduce CNN-based MIL methods [10, 12, 13]. Weakly labelled audio tagging can be formulated as a multiple instance learning (MIL) problem. In this way, each audio recording is viewed as a bag and the i-th bag X_i consists of several instances x_{ij} corresponding to the audio frames. Label for k-th event $Y_i^k \in \{0, 1\}$ is only available at the bag-level while the k-th event label of instances $y_{ij}^k \in \{0, 1\}$ in each bag is unknown, where $k \in \{1, \ldots, K\}$ and

This paper was supported by Shenzhen Science & Technology Fundamental Research Programs (No: JCYJ20170817160058246 & JCYJ20180507182908274). *Corresponding author: zouyx@pku.edu.cn

K is the number of events. Thus, the assumption of the MIL problem for k-th event can be written as follows:

$$Y_i^k = 1 - \prod_{j=1}^{N_i} \left(1 - y_{ij}^k\right) \tag{1}$$

where N_i is the number of instances in the i-th bag.

Convolutional neural networks have been employed with MIL to learn deep representations from bags of instances [14, 15]. Denote the representations of the i-th bag relevant to event k obtained by CNNs as: $q_i^k = \left\{q_{i1}^k, q_{i2}^k, \ldots, q_{iN_i}^k\right\}$. Therefore, the aggregated representation of the bag for MIL is: $\hat{Y}_i^k = f\left(q_{i1}^k, q_{i2}^k, \ldots, q_{iN_i}^k\right)$, where $f(\cdot)$ is typically chosen as an attention pooling function [10, 16] which is applied to bridge instance-level representations to bag-level. Formally, the predicted probability $\hat{Y}_i^k$ of the i-th bag for event k can be computed as:

$$\hat{Y}_i^k = \frac{1}{\sum_{j=1}^{N_i} e\left(q_{ij}^k\right)} \sum_{j=1}^{N_i} e\left(q_{ij}^k\right) v\left(q_{ij}^k\right) \tag{2}$$

where $e(\cdot)$ denotes an attention function and $v(\cdot)$ denotes a tagging function.

Finally, the model is trained to minimize the cross entropy loss averaged over all bags and all events, which is defined as:

$$\min -\frac{1}{K \times I} \sum_{k,i} \left(Y_i^k \log \hat{Y}_i^k + \left(1 - Y_i^k\right) \log\left(1 - \hat{Y}_i^k\right)\right) \tag{3}$$

where I is the number of bags.

3. SPATIAL AND CHANNEL-WISE ATTENTION IN CNN

The features in the intermediate layers of CNN are inherently correlated. However, such information is not considered in the CNN-based MIL methods discussed above. Here, we present a new mechanism exploiting correlations with spatial and channel-wise attention (SCA), hence improving audio tagging performance.

3.1. Spatial attention

With convolutional spatial features, regions adhering to the events will be highlighted and provide a more accurate spatial descriptions. Thus, in our spatial-wise attention, we aim to adaptively characterize the importance of the regions with spatial weights to target the location of related events in the time-frequency representation. The spatial attention is derived from global and local features, represented as an event presence likelihood.

As illustrated in Fig. 1 (a), given an intermediate feature map $F \in R^{W \times H \times C}$, where W, H and C denote the length of width, height and channel respectively, we perform a cross-channel average pooling for aggregating distributed feature information into a spatial descriptor S as:

$$S = \frac{1}{C} \sum_{i=1}^{C} F_{1:W,1:H,i} \tag{4}$$

Firstly, we model the global attention which takes the global spatial descriptor S as input and utilizes two convolution layers (with 5×5 convolutional kernels) to generate the attention mask, as:

$$M_{\text{global}} = f^{5\times5}\left(\delta\left(f^{5\times5}(S)\right)\right) \tag{5}$$

where $f^{5\times5}$ refers to a convolution operation with the kernel size of 5×5 and δ denotes the ReLU function. In addition, batch normalization [17] is attached to the convolutional layer for accelerating the learning.

Secondly, since the event occurring timestamp is important to distinguish between events, we also consider local attention which aims to capture the onset and offset of the events by taking the local patch features as input [18, 19]. Specifically, along the frequency axis with the interval set to 1, the global spatial descriptor S is split into local parts, denoted as local descriptors $\left(s^1, s^2, \ldots, s^H\right)$. Then, we assign local attention statistics m^h $(1 \le h \le H)$ to different local descriptors with the following function:

$$m^h = W_{h,2}\left(\delta\left(W_{h,1}\left(s^h\right) + b_{h,1}\right)\right) + b_{h,2} \tag{6}$$

where $W_{h,1} \in R^{W/r \times W}$, $b_{h,1} \in R^{W/r}$ and $W_{h,2} \in R^{W \times W/r}$, $b_{h,2} \in R^W$ are learnable parameters and r is the reduction ratio [20] to save the parameter overhead, set typically to 16 in our experiments. After, the local attention mask can be represented as $M_{local} = \left[m^1, m^2, \ldots, m^H\right]$, where all local attention statistics are concatenated along frequency axis.

Finally, the spatial attention weight is calculated by combing the global attention mask with the local attention mask as:

$$M_s = \sigma\left(M_{\text{global}} + M_{\text{local}}\right) \tag{7}$$

where σ denotes the sigmoid function.

3.2. Channel-wise attention

Convolutional channel features often capture different sound patterns, which corresponds to different scenes. Inspired by [20], we exploit the inter-channel relationships in the channel branch to improve the model's selection of features relevant to sound events, which leads to discriminative features among channels for various scenes.

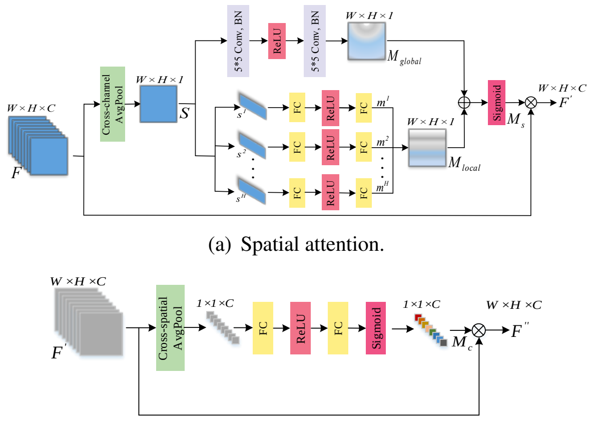
As illustrated in Fig. 1 (b), the channel-wise descriptor Z is extracted from the convolutional feature map $F' \in R^{W \times H \times C}$ by performing a global average pooling across the spatial dimension as:

$$Z = \frac{1}{W \times H} \sum_{i=1}^{W} \sum_{j=1}^{H} F'_{i,j,1:C} \tag{8}$$

To fully capture channel dependencies with the channel-wise descriptor, we apply a simple gating mechanism. Two fully connected (FC) layers with the reduction ratio activated by sigmoid function is employed to limit model complexity and aid generalization. The formulation is as follows:

$$M_c = \sigma\left(W_1\left(\delta\left(W_0(Z) + b_0\right)\right) + b_1\right) \tag{9}$$

where $W_0 \in R^{C/r \times C}$, $b_0 \in R^{C/r}$ and $W_1 \in R^{C \times C/r}$, $b_1 \in$



(a) Spatial attention.

(b) Channel-wise attention.

Fig. 1. The illustration of spatial and channel-wise attention. ⊕ denotes the element-wise summation while ⊗ denotes element-wise multiplication.

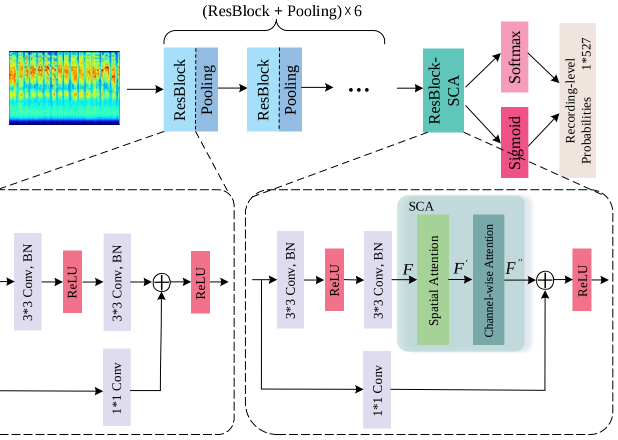


Fig. 2. Architecture of the proposed **AT-SCA**. Log-mel spectrogram is input to six stacked ResBlock and pooling blocks with one ResBlock containing SCA. Then an attention pooling function is applied to aggregate the representations of the bag. Note that **Baseline**, **AT-C**, **AT-S**, **AT-CSA** and **AT-C//S** can be implemented by selecting the corresponding attention modules and arranging their relative positions based on the same backbone network.

R^C are learnable parameters.

3.3. Arrangement of attention modules

In spatial and channel-wise attention (SCA), the spatial attention is applied before channel-wise attention as illustrated in Fig. 2. Formally, given the CNN feature map $F \in R^{W \times H \times C}$ as input, SCA infers a 2D spatial attention weight $M_s \in R^{W \times H \times 1}$ and a 1D channel-wise attention weight $M_c \in R^{1 \times 1 \times C}$. The overall attention process can be summarized as:

$$\begin{aligned} F' &= M_s(F) \otimes F \\ F'' &= M_c\left(F'\right) \otimes F' \end{aligned} \tag{10}$$

where $\otimes$ denotes element-wise multiplication, F' is the spatial weighted feature and F'' is the final refined output.

In order to further study the effect of the arrangement of spatial and channel-wise attention, we propose two SCA variants: CSA, C//S. CSA exchanges the order of two attentions by firstly applying the channel-wise attention and then the spatial one. C//S applies spatial and channel-wise attention in parallel. Noted that we apply the sigmoid function to the sum of spatial and channel-wise attention masks, squashing the attention activation into the range (0,1).

4. EXPERIMENTS

4.1. Dataset

We systemically evaluate the proposed SCA on Auidoset [1]. Auidoset is a large scale weakly labelled dataset based on a collection of over 2 million 10-second experts of YouTube videos, with a total of 527 categories, in which the information of the time span of the events is unknown. The dataset contains three partitions: a balanced training set, an unbalanced training set, and an evaluation set. Both the balanced and unbalanced training sets are used for training, with one part taken as our validation set. The evaluation set is used as the test set in our experiments. The evaluation metrics of Audiset include the mean average precision (mAP), mean area under the curve (mAUC), and d-prime. For all these metrics, the larger value, the better performance.

To prepare the input features of the network, log-mel spectrograms are extracted from the audio signals. The audio signal is encoded using a Fourier-transform-based filterbank with 64 coefficients distributed on a mel-scale. Each chunk has 400 frames and 64 frequency bins. The configuration of this feature extraction is the same as [13].

4.2. Network architecture

We set the experiments including a baseline and five comparison models to demonstrate the effectiveness of the proposed attention modules. It is notable that the baseline and comparison models use the same backbone network. 1) **Baseline**: we establish a baseline for audio tagging containing pure residual blocks and pooling layer without attention module. Similar to the structure of [13], our network contains 7 residual blocks and 6 max-pooling layers, followed by an attention pooling layer. Each residual block is composed of two 3×3 convolution layers with a 1×1 convolution layer to change the channel dimension. In addition, batch normalization and ReLU function are applied to all convolution layers. Fig. 2 removing SCA refers to our **Baseline**. 2) **AT-S**: spatial attention module applying to the last residual block of the backbone forms the comparison experiment. 3) **AT-C**: it differs from **AT-S** by simply replacing the spatial attention to channel-wise attention. 4) **AT-SCA**: it is the sequentially connected version with spatial attention applied before the channel-wise attention. Fig. 2 depicts the network architecture. 5) **AT-CSA**: this model only inverts the connection order of **AT-SCA**. 6) **AT-C//S**: spatial and channel-wise attention are placed in parallel.

Table 1. The performance of baseline and models with different attention modules.

Model	Spatial		Channel	mAP	mAUC	d-prime
	global	local				
Baseline				0.367	0.967	2.591
AT-S	✓			0.375	0.966	2.583
AT-S		✓		0.374	0.967	2.598
AT-S	✓	✓		0.377	0.967	2.601
AT-C			✓	0.378	0.969	2.632
AT-SCA	✓	✓	✓	**0.383**	**0.969**	**2.635**
AT-CSA	✓	✓	✓	0.381	0.968	2.627
AT-C//S	✓	✓	✓	0.380	0.969	2.632

4.3. Results

Table 1 presents the results with difference of attention modules. First, we evaluate the performance of **AT-S** using different spatial attention methods. Experimental results show that either the global or local attention module enhances the performance and the combination of both brings about more improvement. This is because considering both local temporal areas and the global time-frequency information as a whole enables the model to learn more distinctive event-related regions. Second, we observe consistent boosts over baseline in all models using attention modules, especially incorporating spatial attention with channel-wise attention. It implies that spatial and channel-wise attention complement each other, efficiently helping the information flow by learning which information to emphasize or suppress. Most importantly, for the arrangement of attention modules, we find that the sequential arrangement gives a better result than the parallel one. In general, **AT-SCA** is slightly better than **AT-CSA**, so in the following experiments we use **AT-SCA** to represent the integrated model.

Table 2 lists the results of modeling the attention modules with more attentive layers in **AT-C**, **AT-S** and **AT-SCA** models, including 1-layer, 2-layers, 3-layers. Note that all the attention modules are applied to the last few residual blocks. From Table 2, we see that all of the models with more attentive layers achieve significant improvement over single layer attended ones. This manifests that attention modules at different levels are complementary and benefit each other. We also list the results on Audioset, which were reported by Google's benchmark [1] and the state-of-the-art methods [10, 11, 16, 21]. Our 3-layers **AT-SCA** outperforms remarkably the listed methods in mAP and achieves a similar result in mAUC.

To demonstrate the effectiveness of our SCA, Fig. 3 visualizes the concentration of time-frequency spectrogram regions for **Baseline** and our best 3-layers **AT-SCA** model. Comparing Fig. 3 (b) and Fig. 3 (c), **AT-SCA** attends to more event-related regions, such as the location of events. Besides, as can be observed from Fig. 3 (d), the feature maps with high channel weights represent highlighting spectrogram information, indicating that our channel-wise attention will assign more weights on channels according to the events. As a result, it suggests that SCA helps learn the salient features akin to sound events.

Table 2. The performance of the multi-layer attention in **AT-S**, **AT-C**, **AT-SCA**, compared with various models in the literature.

Model	Depth	mAP	mAUC	d-prime
AT-S	1-layer	0.377	0.967	2.601
	2-layers	0.379	0.967	2.605
	3-layers	0.384	0.968	2.613
AT-C	1-layer	0.378	0.969	2.632
	2-layers	0.382	0.969	2.638
	3-layers	0.385	0.969	2.643
AT-SCA	1-layer	0.383	0.969	2.635
	2-layers	0.387	0.969	2.648
	3-layers	**0.390**	**0.970**	**2.652**
Benchmark (2017) [1]	-	0.314	0.959	2.452
Kong *et al.* (2018) [21]	-	0.327	0.965	2.558
Xu *et al.* (2018) [10]	-	0.360	0.970	2.660
Shi *et al.* (2019) [16]	-	0.365	0.949	-
Kong *et al.* (2019) [11]	-	0.369	0.969	2.640

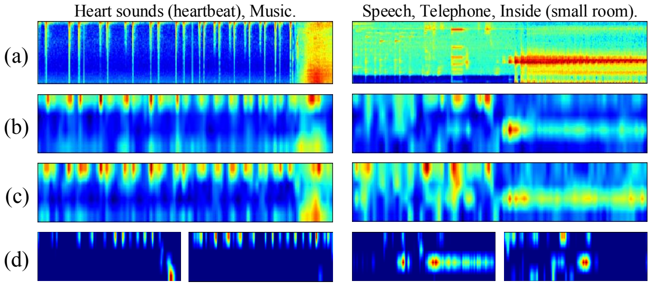


Fig. 3. Two examples of visualization of the 2D feature maps obtained by **Baseline** and 3-layers **AT-SCA**. (a) The log-mel spectrogram of an audio recording. The above descriptions denote the sound events. (b) C-channel average feature maps in **Baseline**. (c) C-channel average feature maps in **AT-SCA**. (d) Two feature maps selected from **AT-SCA** with top-2 highest channel weights.

5. CONCLUSION

In this paper, we propose a CNN-based spatial and channel-wise attention (SCA) for weakly labelled audio tagging. SCA is able to learn what (i.e., channel-wise) and where (i.e., spatial) to emphasize or suppress. Specifically, spatial attention is divided into global attention which highlights the event-related spatial regions and local attention to precisely localize onset and offset timestamps of the events. For channel-wise attention, it is set up to differentiate various sound scenes. Extensive experiments are conducted to verify the effectiveness of spatial attention, channel-wise attention and the combination of them. We also investigate the effect of increasing the number of attentive layers and show that 3-layers **AT-SCA** outperforms the state-of-the-art mean average precision (mAP). The proposed SCA is suitable for exploring label uncertainty information, and it will be extended to other audio tasks in our future work.

6. REFERENCES

[1] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, 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.

[2] T. Heittola and A. Mesaros, "Dcase 2017 challenge setup: Tasks datasets and baseline system," *Tech. Rep., DCASE2017 Challenge*, 2017.

[3] Q. Kong, T. Iqbal, Y. Xu, W. Wang, and M. D. Plumbley, "Dcase 2018 challenge baseline with convolutional neural networks," *arXiv preprint arXiv:1808.00773*, 2018.

[4] A. Kumar and B. Raj, "Audio event detection using weakly labeled data," in *Proceedings of the 24th ACM International Conference on Multimedia*. ACM, 2016, pp. 1038–1047.

[5] J. Ye, T. Kobayashi, M. Murakawa, and T. Higuchi, "Acoustic scene classification based on sound textures and events," in *Proceedings of the 23rd ACM International Conference on Multimedia*. ACM, 2015, pp. 1291–1294.

[6] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, D. Platt, R. A. Saurous, and B. Seybold, "Cnn architectures for large-scale audio classification," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 131–135.

[7] S. Tseng, J. Li, Y. Wang, J. Szurley, F. Metze, and S. Das, "Multiple instance deep learning for weakly supervised small-footprint audio event detection," in *Interspeech*, 2017, pp. 3279–3283.

[8] J. Li, Y. Wang, J. Szurley, F. Metze, and S. Das, "A light-weight multimodal framework for improved environmental audio tagging," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 6832–6836.

[9] A. Shah, A. Kumar, A. G. Hauptmann, and B. Raj, "A closer look at weak label learning for audio events," *arXiv preprint arXiv:1804.09288*, 2018.

[10] C. Yu, K. S. Barsim, Q. Kong, and B. Yang, "Multi-level attention model for weakly supervised audio classification," in *Workshop on Detection and Classification of Acoustic Scenes and Events*, 2018.

[11] Q. Kong, C. Yu, T. Iqbal, Y. Xu, W. Wang, and M. D. Plumbley, "Weakly labelled audioset tagging with attention neural networks," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2019, vol. 27, pp. 1791–1802.

[12] A. Kumar and B. Raj, "Deep cnn framework for audio event recognition using weakly labeled web data," *arXiv preprint arXiv:1707.02530*, 2017.

[13] Y. Wang, J. Li, and F. Metze, "A comparison of five multiple instance learning pooling functions for sound event detection with weak labeling," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 31–35.

[14] Y. Xu, Q. Kong, W. Wang, and M. D. Plumbley, "Attention and localization based on a deep convolutional recurrent model for weakly supervised audio tagging," in *Interspeech*, 2017, pp. 3083–3087.

[15] K. Choi, G. Fazekas, and M. Sandler, "Automatic tagging using deep convolutional neural networks," *arXiv preprint arXiv:1606.00298*, 2016.

[16] R. Shi, R. W. Ng, and P. Swietojanski, "Teacher-student training for acoustic event detection using audioset," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 875–879.

[17] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[18] Y. Wu, H. Mao, and Z. Yi, "Audio classification using attention-augmented convolutional neural network," *Knowledge-Based Systems*, vol. 161, pp. 90–100, 2018.

[19] X. Lu, P. Shen, S. Li, Y. Tsao, and H. Kawai, "Temporal attentive pooling for acoustic event detection," in *Interspeech*, 2018, pp. 1354–1357.

[20] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7132–7141.

[21] 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.