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

Improving weakly supervised sound event detection with self-supervised auxiliary tasks

Soham Deshmukh[1*], *Bhiksha Raj*[2], *Rita Singh*[2]
[1]Microsoft, USA
[2]Carnegie Mellon University, USA
sdeshmukh@andrew.cmu.edu, bhiksha@cs.cmu.edu, rsingh@cs.cmu.edu

Abstract

While multitask and transfer learning has shown to improve the performance of neural networks in limited data settings, they require pretraining of the model on large datasets beforehand. In this paper, we focus on improving the performance of weakly supervised sound event detection in low data and noisy settings simultaneously without requiring any pretraining task. To that extent, we propose a shared encoder architecture with sound event detection as a primary task and an additional secondary decoder for a self-supervised auxiliary task. We empirically evaluate the proposed framework for weakly supervised sound event detection on a remix dataset of the DCASE 2019 task 1 acoustic scene data with DCASE 2018 Task 2 sounds event data under 0, 10 and 20 dB SNR. To ensure we retain the localisation information of multiple sound events, we propose a two-step attention pooling mechanism that provides a time-frequency localisation of multiple audio events in the clip. The proposed framework with two-step attention outperforms existing benchmark models by 22.3 %, 12.8 %, 5.9 % on 0, 10 and 20 dB SNR respectively. We carry out an ablation study to determine the contribution of the auxiliary task and two-step attention pooling to the SED performance improvement.[1].

Index Terms: sound event detection, self-supervised learning, pooling function

1. Introduction

Sound Event Detection (SED) aims to determine the presence, nature and temporal location of sound events in audio signals. Many SED algorithms rely on strongly labelled data [1, 2, 3] for training to perform accurate event detection and localisation. However, producing strongly labelled data for SED is quite expensive in terms of the expertise, time and human resources required for the annotation. This has led to the creation of weakly labelled sound event detection dataset like Audioset [4] which contains audio clip level annotations without the corresponding onset and offset times of the audio events.

The weakly supervised sound event detection was first formulated as a Multiple-Instance Learning (MIL) problem [5, 6] with the recent emergence of Neural MIL. In Neural MIL, the first half of the network (segmentation network) produces temporal predictions which are then aggregated by the second half of the network (classification network) usually a pooling operator to produce audio clip level predictions. The benefit of such formulation is, along with detecting audio events in the clip, it provides insight into time level localisation of those sound events in the audio clip. Since then, recent works have focused on improving the model architecture of the segmentation network [7, 8, 9] and developing better pooling methods [10, 11, 12, 13, 14]. However, few works have focused on how sound event detection models perform in either limited data or noisy settings let alone in both of them.

The noisy data also affects the training of networks for sound event detection. Specifically, the deep CNN architectures [15, 16] currently used to provide benchmark performance for different speech and audio tasks [17] require large labelled clean datasets to train on and when considered in a noisy environment the performance is known to deteriorate [10]. The two general learning strategies used as solutions are transfer learning and multitask learning which were recently utilised for sound event detection [18, 19, 20]. However, in the multitask learning setup, it's assumed you have richly annotated labels for all the tasks. We investigate a counterpart of this where only weak labels are available without any labels for the secondary task. For this setting, we propose a self-supervised auxiliary task that will be jointly trained with the primary task of sound event detection. The auxiliary task is chosen to be the reconstruction of log Mel spectrogram of audio and we show how the auxiliary task denoises internal representations and improves network performance in noisy settings.

In all, in this paper, we address the challenge of training sound event detection models in noisy (domestic or environmental) and limited data settings. To that effort, we make twofold contributions. First, identify appropriate self-supervised auxiliary task for sound event detection in noisy settings and demonstrate performance benefits to the same. Second, develop a two step attention pooling mechanism that improves time-frequency localisation of audio events and indirectly improves sound event detection performance in noisy settings. We perform all the experiments on a standard noisy sound event detection dataset remix [10] and release the code publicly.

2. Related work

A prominent recent work [10] analysed the performance of different model architectures (segmentation network and pooling functions) under different Signal to Noise Ratio (SNR) for sound event detection and localisation. The paper showed that the segmentation network of type 'VGG-like' CNN performed best for audio tagging and variability in performance resulted from the choice of pooling methods with not a clear winning pooling method across different SNR. Specifically, Global Attention Pooling outperformed other pooling methods on some SNR and metrics, while Global Weighted Rank Pooling (GWRP) results in the best performance on others. Still, the work on sound event detection performance in limited data and noisy settings is sparse.

Though the various type of multitask learning methods have been greatly explored for vision and natural language processing (NLP) tasks [21], it has not been utilised by the audio community. Most of the works in multitask learning for SED fo-

[*]work done at Carnegie Mellon University
[1]The code is publicly released.

http://dx.doi.org/10.21437/Interspeech.2021-2079

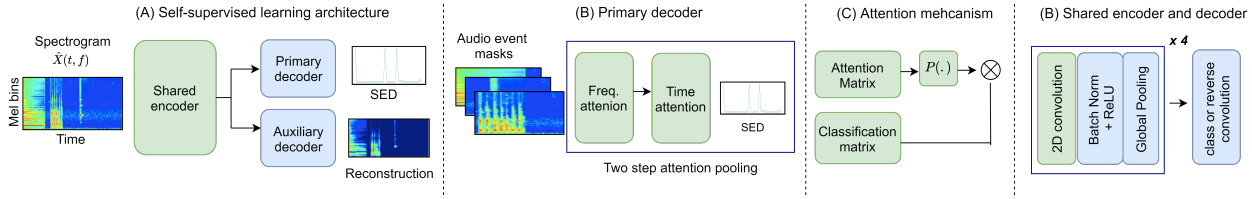

Figure 1: *Our proposed self-supervised learning assisted framework for weakly supervised sound event detection. (A) The general architecture with shared encoder and multiple decoder branches. Shared encoder, primary decoder, auxiliary decoder is represented by g, g_2, g_4 respectively (B) shows the two step attention pooling function used for primary decoder. (C) The attention mechanism used for frequency and time attention in two step attention pooling along different axis. (D) The CNN architecture used for shared encoder and auxiliary decoder. The last layer is either class or reverse convolution for encoder and decoder respectively.*

cus on jointly training SED with another strongly labelled task like Sound Source Localisation (SSL) [18] or Acoustic Scene Classification (ASC) [19, 22, 23]. A combination of multitask learning and self-supervised learning is shown to improve performance on speech and audio tasks [20]. However, the work uses large scale speech datasets like Librispeech [24] as pretasks to pretrain the networks using self-supervised learning and does not analyse the effect of noise (domestic or environmental) on sound event detection performance.

3. Methodology

This section contains the details of the proposed approach for SED, segmentation mapping network g_1, classification mapping network g_2, and the auxiliary time-frequency reconstruction auxiliary task. The architecture is depicted in figure 1

3.1. Self-supervised Learning formulation for SED

Let the raw audio be represented by $X = \{x_i\}_{i=1}^{T}$ where each $x_i \in \mathbb{R}$ is a frame in the audio clip. We extract time-frequency features for each audio, let them be represented by $\hat{X} = \{\hat{x}_i\}_{i=1}^{T}$ where each $\hat{x}_i \in \mathbb{R}^d$, $d \in \mathbb{Z}$ corresponds to frame in the audio clip. In practice, d are the number of mel bins obtained after computing the spectrogram. As per MIL formulation, we can represent each sample in dataset as a bag $B_j = (\{\hat{x}_i\}_{i=1}^{T}, y)|_{j=0}^{N}$ where $y \in \mathbb{R}^C$ is the weak label, N are the number of samples and C are the number of audio events. The primary task in our self-supervised framework is SED. The segmentation mapping $g_1(.)$ of SED also acts as a shared encoder for the auxiliary task. The shared encoder maps the feature set $\{\hat{x}_i\}_{i=1}^{T}$ to $Z = \{z_i\}_{i=1}^{T}$ where $z_i \in \mathbb{R}^{C \times F \times T}$. The second part of SED task is network which classifies $\{z_i\}_{i=1}^{T}$ to $P = \{p_i\}_{i=1}^{C}$ where $P \in \mathbb{R}^C$. The network learns a mapping g_2 which maps each audio events time-frequency segmentation to corresponding presence probabilities of c^{th} event known as p_k

$$g_1 : \hat{X} \mapsto Z \qquad g_2 : Z \mapsto P \tag{1}$$

The auxiliary self-supervised task chosen needs to help in learning robust representations which generalise to noisy settings without requiring additional labels. This will impact not only the learned internal representation but also downstream sound event detection and localisation performance. Inorder to achieve that we choose auxiliary task as reconstruction of extracted time-frequency features for audio. By having time-frequency reconstruction auxiliary task we hypothesise the network will learn representations which retain audio event information better [25, 26]. We use an auto-encoder structure for reconstruction where the encoder is shared with the primary task of SED. If $g_3(.)$ is encoder mapping for reconstruction task, we now represent $g_1(.) = g_3(.) = g(.)$ as the shared segmentation mapping function. The second part of auxiliary task, is a decoder network which learns a mapping g_4 such that $g_4 : Z \mapsto \bar{X}$ where $\bar{X}$ is the reconstructed time-frequency representation. Specifically $\{\bar{x_i}\}_{i=1}^{T} = g_4(\{z_i\}_{i=1}^{T})$. Here the learned mapping function $g_4(.)$ should satisfy:

$$g_4^{-1}(g(.)) = g^{-1}(g_4(.)) = I \tag{2}$$

To learn the function mappings satisfying primary SED task, let the objective function be $\mathcal{L}_1$. To enforce the constraint of auxiliary task, let the objective function be $\mathcal{L}_2$ where the aim is to minimise the difference between T-F representation $\{\hat{x_i}\}_{i=1}^{T}$ and predicted time-frequency representation $\{\bar{x_i}\}_{i=1}^{T}$ of audio clip. If the learnable parameters are W = $[w, w_2, w_4]$ and w, w_2, w_4 corresponding to $g(.), g_2(.), g_4(.)$ respectively, then the optimisation problem can be framed in terms of these weights W over all data points as:

$$\min_{W} \mathcal{L}_1(P, y|w, w_4) + \alpha\mathcal{L}_2(\{\bar{x_i}\}_{i=1}^{T}, \{\hat{x_i}\}_{i=1}^{T}|w, w_2) \tag{3}$$

The parameter alpha (α) accounts for scale difference between losses $\mathcal{L}_1$ and $\mathcal{L}_2$. It helps in adjusting the contribution of auxiliary task relative to the primary task in learning weights.

3.2. Shared encoder and auxiliary task decoder network

The segmentation mapping function (shared encoder) converts the time-frequency audio input into a T-F representation for each of the audio events. The time-frequency feature extracted for audio here is log Mel spectrogram as it has shown to provide better performance [27, 28, 17]. We choose a CNN based architecture similar to 'VGG-like' [10] for both shared encoder and auxiliary task decoder. The shared encoder has CNN based network consists of 8 blocks of 2D Convolution, BatchNorm and ReLU with an Average Pool after every 2 blocks. Having a common encoder helps the network to learn a shared representation by exploiting the similarity across SED and T-F reconstruction and enables the network to generalise better on our original task. We use a hard parameter sharing framework to reduce the risk of overfitting [29] to limited samples.

The decoder of the auxiliary-task takes $Z = \{z_i\}_{i=1}^{T}$ as input and reconstructs it to $\{\bar{x_i}\}_{i=1}^{T}$. The decoder consists of CNN based network for combining the intermediate time-frequency representations obtained for each audio event to an audio level time-frequency representation. The architecture closely follows the common encoder structure in reverse order consisting of 8 blocks of 2D Convolution, BatchNorm and ReLU with Average Pool after every two blocks with a decreasing number of filters.

Table 1: *Weakly supervised sound event detection performance across different SNR*

Network			SNR 20 dB			SNR 10 dB			SNR 0 dB		
encoder	pooling	aux.	micro-p	macro-p	AUC	micro-p	macro-p	AUC	micro-P	macro-p	AUC
VGGish	GAP	✗	0.5067	0.6127	0.9338	0.4291	0.5390	0.9144	0.3295	0.4093	0.8694
VGGish	GMP	✗	0.5390	0.5186	0.8497	0.5263	0.5023	0.8422	0.4640	0.4441	0.8189
VGGish	GWRP	✗	0.7018	0.7522	0.9362	0.6538	0.7129	0.9265	0.5285	0.6084	0.8985
VGGish (dil.)	AP	✗	0.7391	0.7586	0.9279	0.6740	0.7404	0.9211	0.5714	0.6341	0.9014
VGGish	2AP	✓	**0.7829**	**0.7645**	**0.9390**	**0.7603**	**0.7486**	**0.9343**	**0.6986**	**0.6892**	**0.9177**

3.3. Primary decoder

The primary decoder is not CNN based, instead, it is a pooling operator to satisfy MIL formulation. The choice of pooling operator has a significant performance effect on both the SED and each audio events intermediate time-frequency representation obtained. Global max pooling and global average pooling results in underestimate and overestimate the audio event's temporal presence respectively, and to overcome this problem dynamic poolings were proposed [10, 12, 14]. However, the developed pooling mechanisms still lacks the granularity in temporal predictions and does not provide frequency localisation which might be used to further disambiguate sound events. Also, the standard attention pooling [14] is known to be unstable with cross-entropy usually used for multi-class setup in practice.

We propose a two-step attention pooling mechanism to covert each audio events segmentation maps $\{z_i\}_{i=1}^{T}$ into audio level predictions P. The first step in the two step attention pooling takes $Z = \{z_i\}_{i=1}^{T}$ as input. This undergoes two independent learned linear transformation to produce classification and attention output respectively. The attention output is squashed to ensure its valid probability distribution. Mathematically, the attention output Z_{a_1} and classification output Z_{c_1} are:

$$Z_{a_1} = \frac{e^{\sigma(ZW_{a_1}^T + b_{a_1})}}{\sum_{i=1}^{F} e^{\sigma(ZW_{a_1}^T + b_{a_1})}} \quad Z_{c_1} = (ZW_{c_1}^T + b_{c_1}) \quad (4)$$

This is followed by a weighted combination of classification output Z_{c_1} by attention weights Z_{a_1}:

$$Z_{p_1} = \sum_{i=0}^{F} Z_{c_1} \cdot Z_{a_1} \quad (5)$$

The time level attention is similar to frequency (first step) attention except it operates along time axis:

$$Z_{a_2} = \frac{e^{\sigma(Z_{p_1}W_{a_2}^T + b_{a_2})}}{\sum_{t=1}^{T} e^{\sigma(Z_{p_1}W_{a_2}^T + b_{a_2})}} \quad Z_{p_1} = (ZW_{c_2}^T + b_{c_2}) \quad (6)$$

$$Z_{p_2} = \sum_{t=0}^{T} Z_{c_2} \cdot Z_{a_2} \quad (7)$$

where $Z_{p_2} \in [0, 1]$ and denotes the presence probability of each sound event in the audio clip. Figure 1 subsection c, provides an overview of a single attention step. In relation to figure, Z_a, Z_c, Z_p are the outputs after attention matrix, classification matrix and $P(.)$ respectively in the first stage and second stage depending on subscript. By breaking the attention into two steps, it makes the pooling more interpretable by answering the questions of what frequency bins and what time steps contributes to which audio events by visualising normalised attention weights Z_{a_1}, Z_{a_2} and output Z_{p_1}, Z_{p_2}. Also, the sigmoid (σ) ensures the attention output stays between 0 to 1 and avoids unstable training for multilabel training with cross-entropy in practice.

4. Experiments

4.1. Dataset

To study the effect of noise in limited data settings, we form a noisy dataset by mixing DCASE 2019 Task 1 of Acoustic scene classification [30] and DCASE 2018 Task 2 of General purpose Audio tagging [31]. The DCASE 2019 Task 1 provides background sounds (noise) recorded from a variety of real world scenes in which the sounds from DCASE 2019 Task 2 are randomly embedded [10]. To ensure the noise conditions are natural, diverse and challenging, we use the new DCASE 2019 Task 1 instead of DCASE 2018 as used in [10]. The 2019 variant extends the TUT Urban Acoustic Scenes 2018 with the other 6 cities to a total of 12 large European cities. This results in 32000 audio clips with 8000 audio clips for each 20,10,0 dB SNR where each audio clip is of 10 secs with background noise and three random audio events (out of total 41) in it.

4.2. Set up

The raw data is converted to time-frequency representation by applying FFT with a window size of 2048 and an overlap of 1024 between windows. This is followed by applying Mel filter banks with 64 bands and converting them to log scale to obtain log Mel spectrogram. The network architecture used is described in section 3.2. The entire network is trained end-to-end with a batch size of 24 and learning of 1e-3 using Adam optimiser [32]. The code and setup is publicly released[2].

5. Results

5.1. Sound event detection

We evaluate our self-supervision assisted architecture and pooling method against different baselines, benchmark architectures and pooling methods [14, 10]. Table 1 shows weakly supervised sound event detection performance across different SNR of 20,10, and 0 dB. The important evaluation metric here under consideration is micro precision (micro-p), as it uses global counts of true positives, false negatives and false positives for metric computation against macro precision which does simple unweighted averaging disregarding class-imbalance. The VGGish (dil.) encoder here indicates VGGish architecture but with dilated/atrous convolutions known to provide benchmark performance for sound event detection, [14]. The VGGish encoder with reconstruction based auxiliary task and two step attention pooling outperforms the existing benchmark of atrous attention pooling [14] on SNR 20, 10 and 0 dB by 5.9%, 12.8% and 22.3% respectively. Apart from improving performance, by breaking the attention into two steps, it allows for the intermediate use of sigmoid which helps in ensuring the outputs don't overflow above 1 during training.

[2] https://github.com/soham97/MTL_Weakly_labelled_audio_data

Table 2: *Ablation study to determine auxiliary task contribution*

auxiliary task	SNR 20 dB	SNR 10 dB	SNR 0 dB
$\alpha = 0.0$	0.7772	0.7430	0.6937
$\alpha = 0.001$	**0.7829**	**0.7603**	**0.6986**
$\alpha = 0.1$	0.7637	0.7428	0.6792

5.2. Ablation study for auxiliary task contribution

We perform an ablation study to determine the contribution of reconstruction auxiliary task and two step attention pooling towards the total performance improvement. As described in Section 3.1, the total loss is:

$$\mathcal{L} = \mathcal{L}_1(P, y|w, w_4) + \alpha\mathcal{L}_2(\{\bar{x}_i\}_{i=1}^{T}, \{\hat{x}_i\}_{i=1}^{T}|w, w_2) \quad (8)$$

By changing the value of α before training, we can adjust the contribution of the auxiliary task to primary sound event detection. When $\alpha = 0.0$, the network has no contribution from the reconstruction auxiliary task during training and it can be used to evaluate the performance of two step attention pooling. In terms of micro-precision, the two step attention pooling outperforms existing benchmark of atrous AP (row 4) from table 1 on SNR 20, 10 and 0 dB by 5.2%, 10.2% and 21.4% respectively. By adding the auxiliary task contribution with a relative weightage of $\alpha = 0.001$, an additional improvement of 0.7%, 2.3% and 0.7% is observed. This indicates that two step attention has a prominent contribution in improving the performance of sound event detection in limited data and noisy settings, with additional performance gains from the auxiliary task. When α is increased to 0.01, the performance compared to $\alpha = 0.001$ is decreased. This suggests that the auxiliary task's loss contribution starts to overpower the primary SED task's loss contribution rather than improving generalisation.

Table 3: *Two top and worst performing sound events- SNR 0 dB*

model	aux.	bus	cowbell	gong	meow
Atrous + AP	✗	0.2	0.781	**0.692**	**0.583**
VGGish + 2AP	✗	0.572	0.921	0.643	0.483
VGGish + 2AP	✓	**0.627**	**0.94**	0.663	0.532

5.3. SED performance on specific audio events

For almost all audio events, our proposed architectures have the best precision scores against GMP, GAP, GWRP, Atrous across all SNR = 0, 10, 20. Particularly, for audio events like 'Bass drum', 'bus', 'double bass', 'cowbell' the architecture outperforms other models by a large margin as shown in table 3. However, the proposed model struggles in audio events like 'gong','chime' and 'meow' where the attention pooling with dilated convolution encoder performs better [14]. This indicates using atrous or dilated convolutions helps in detecting audio events whose energy is spread wide in the temporal domain. This can be incorporated into our current architecture by replacing the linear convolutions in the shared encoder with dilated convolutions. Further analysis and event-specific results are available in the long version of paper [3] and skipped due to space constraints.

5.4. Interpretable visualisation of audio events

Apart from improved performance, using two step attention pooling provides a way to localise each audio event present in the audio clip along with both the time and frequency axis. To illustrate this, we pick a random example with SNR 20 dB and show the end to end visualisation of the two step attention pooling mechanism in figure 2. The audio under consideration has three events occurring in it: telephone ringing, cello playing and cat meowing, with outdoor environmental background noise. Subplot 2 in figure 2 depicts the reconstructed Mel spectrogram of the audio clip. From the subplot, we can see that the decoder is not only able to reconstruct the audio events clearly but it is also denoising the log Mel spectrogram retaining the key elements of three audio events. A future extension of work is to jointly train sound source separation along with weakly supervised SED by using the auxiliary task reconstruction output.

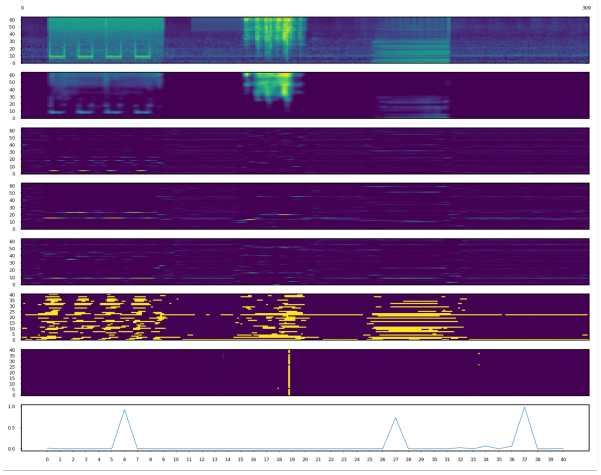

Figure 2: *Visualisation of two step attention pooling and reconstruction decoder outputs. Subplot 1 depicts the scaled log Mel spectrogram of an audio clip. Subplot 2 is the output of the reconstruction auxiliary task. Subplots 3,4 and 5 are attention weights for the three most probable audio events in the audio clip. Subplot 6 is the output of the first step attention pooling. Subplot 7 and 8 is the attention weight and output of second step attention pooling respectively. The y-axis in subplot 1-4 corresponds to Mel-bins and sound events in subplot 5-6. The x-axis in subplot 1-7 corresponds to time and sound events in subplot 8*

6. Conclusions

This paper proposes assisted self-supervised task for improving sound event detection in limited data and noisy settings. The architecture consists of sound event detection as a primary task with two-step attention pooling as a primary decoder and time-frequency representation reconstruction as an auxiliary task. We empirically evaluate the proposed framework for multi-label weakly supervised sound event detection, on a remix DCASE 2019 and 2018 dataset under 0, 10 and 20 dB SNR. The proposed self-supervised auxiliary task framework with two-step attention outperforms existing benchmark models by 22.3 %, 12.8 %, 5.9 % on 0, 10 and 20 dB SNR respectively. The ablation study carried out indicates the majority of performance improvement is associated with two step attention pooling with secondary performance improvement from self-supervised auxiliary task. Furthermore, by using two step attention, we can easily visualise the sound event presence along both time-frequency axis. The code is publicly released.

[3] https://arxiv.org/pdf/2008.07085.pdf

7. References

[1] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M. D. Plumbley, "Detection and classification of acoustic scenes and events," *IEEE Transactions on Multimedia*, vol. 17, no. 10, pp. 1733–1746, 2015.

[2] E. Çakır, 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] A. Mesaros, T. Heittola, and T. Virtanen, "Tut database for acoustic scene classification and sound event detection," in *2016 24th European Signal Processing Conference (EUSIPCO)*, 2016, pp. 1128–1132.

[4] J. F. Gemmeke, D. P. W. 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)*, 2017, pp. 776–780.

[5] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez, "Solving the multiple instance problem with axis-parallel rectangles," *Artif. Intell.*, vol. 89, no. 1–2, p. 31–71, Jan. 1997. [Online]. Available: https://doi.org/10.1016/S0004-3702(96)00034-3

[6] A. Kumar and B. Raj, "Audio event detection using weakly labeled data," in *Proceedings of the 24th ACM International Conference on Multimedia*, ser. MM '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 1038–1047. [Online]. Available: https://doi.org/10.1145/2964284.2964310

[7] S.-Y. Tseng, J. Li, Y. Wang, F. Metze, J. Szurley, and S. Das, "Multiple instance deep learning for weakly supervised small-footprint audio event detection," in *Proc. Interspeech 2018*, 2018, pp. 3279–3283. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2018-1120

[8] A. Kumar and B. Raj, "Deep cnn framework for audio event recognition using weakly labeled web data," 2017, arXiv preprint, https://arxiv.org/abs/1707.02530.

[9] Y. Xu, Q. Kong, W. Wang, and M. D. Plumbley, "Large-scale weakly supervised audio classification using gated convolutional neural network," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 121–125.

[10] Q. Kong, Y. Xu, I. Sobieraj, W. Wang, and M. D. Plumbley, "Sound event detection and time–frequency segmentation from weakly labelled data," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 27, no. 4, p. 777–787, Apr. 2019. [Online]. Available: https://doi.org/10.1109/TASLP.2019.2895254

[11] S.-Y. Chou, J.-S. R. Jang, and Y.-H. Yang, "Framecnn : A weakly-supervised learning framework for frame-wise acoustic event detection and classification," 2017, technical report, DCASE.

[12] B. McFee, J. Salamon, and J. P. Bello, "Adaptive pooling operators for weakly labeled sound event detection," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 26, no. 11, p. 2180–2193, Nov. 2018. [Online]. Available: https://doi.org/10.1109/TASLP.2018.2858559

[13] T. Su, J. Liu, and Y. Yang, "Weakly-supervised audio event detection using event-specific gaussian filters and fully convolutional networks," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 791–795.

[14] Z. Ren, Q. Kong, J. Han, M. D. Plumbley, and B. W. Schuller, "Attention-based atrous convolutional neural networks: Visualisation and understanding perspectives of acoustic scenes," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 56–60.

[15] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations*, 2015.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

[17] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. Weiss, and K. Wilson, "Cnn architectures for large-scale audio classification," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017. [Online]. Available: https://arxiv.org/abs/1609.09430

[18] W. Xue, Y. Tong, C. Zhang, G.-H. Ding, X. He, and B. Zhou, "Sound event localization and detection based on multiple doa beamforming and multi-task learning," in *INTERSPEECH*, 2020.

[19] K. Imoto, N. Tonami, Y. Koizumi, M. Yasuda, R. Yamanishi, and Y. Yamashita, "Sound event detection by multitask learning of sound events and scenes with soft scene labels," *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 621–625, 2020.

[20] T. Lee, T. Gong, S. Padhy, A. Rouditchenko, and A. Ndirango, "Label-efficient audio classification through multitask learning and self-supervision," *ArXiv*, vol. abs/1910.12587, 2019.

[21] Y. Zhang and Q. Yang, "A survey on multi-task learning," 2018, preprint arXiv, https://arxiv.org/abs/1707.08114.

[22] N. Tonami, K. Imoto, M. Niitsuma, R. Yamanishi, and Y. Yamashita, "Joint analysis of acoustic events and scenes based on multitask learning," *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 338–342, 2019.

[23] H. L. Bear, I. Nolasco, and E. Benetos, "Towards joint sound scene and polyphonic sound event recognition," in *INTERSPEECH*, 2019.

[24] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.

[25] E. M. Grais and M. D. Plumbley, "Single channel audio source separation using convolutional denoising autoencoders," in *2017 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, 2017, pp. 1265–1269.

[26] D. Stowell and R. E. Turner, "Denoising without access to clean data using a partitioned autoencoder," 2015, preprint arXiv, https://arxiv.org/abs/1509.05982.

[27] K. J. Piczak, "Environmental sound classification with convolutional neural networks," in *2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP)*, 2015, pp. 1–6.

[28] J. Salamon and J. P. Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," *IEEE Signal Processing Letters*, vol. 24, no. 3, pp. 279–283, 2017.

[29] J. Baxter, "A bayesian/information theoretic model of learning to learn via multiple task sampling," in *Machine Learning*, 1997, pp. 7–39.

[30] A. Mesaros, T. Heittola, and T. Virtanen, "A multi-device dataset for urban acoustic scene classification," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE2018)*, November 2018, pp. 9–13. [Online]. Available: https://arxiv.org/abs/1807.09840

[31] E. Fonseca, M. Plakal, F. Font, D. P. W. Ellis, X. Favory, J. Pons, and X. Serra, "General-purpose tagging of freesound audio with audioset labels: Task description, dataset, and baseline," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE2018)*, November 2018, pp. 69–73. [Online]. Available: https://arxiv.org/abs/1807.09902

[32] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2017, preprint arXiv, https://arxiv.org/abs/1412.6980.