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

Gated Multi-head Attention Pooling for Weakly Labelled Audio Tagging

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

[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

hongsixin@pku.edu.cn, zouyx@pku.edu.cn, W.Wang@surrey.ac.uk

Abstract

Multiple instance learning (MIL) has recently been used for weakly labelled audio tagging, where the spectrogram of an audio signal is divided into segments to form instances in a bag, and then the low-dimensional features of these segments are pooled for tagging. The choice of a pooling scheme is the key to exploiting the weakly labelled data. However, the traditional pooling schemes are usually fixed and unable to distinguish the contributions, making it difficult to adapt to the characteristics of the sound events. In this paper, a novel pooling algorithm is proposed for MIL, named gated multi-head attention pooling (GMAP), which is able to attend to the information of events from different heads at different positions. Each head allows the model to learn information from different representation subspaces. Furthermore, in order to avoid the redundancy of multi-head information, a gating mechanism is used to fuse individual head features. The proposed GMAP increases the modeling power of the single-head attention with no computational overhead. Experiments are carried out on Audioset, which is a large-scale weakly labelled dataset, and show superior results to the non-adaptive pooling and the vanilla attention pooling schemes.

Index Terms: audio tagging, weakly labelled data, multiple instance learning, pooling scheme, multi-head attention.

1. Introduction

Audio tagging (AT) refers to the task of assigning labels of one or several sound classes to an audio recording. Potential applications of audio tagging include sound event detection [1], audio retrieval [2] and audio surveillance [3, 4]. There is an increasing interest in audio tagging research, including the collection of training data covering a large number of event classes, and development of learning algorithms for classification. In order to scale up the training data, Google released Audioset [5], which is a large-scale weakly labelled dataset with annotations only for the classes of audio events, but not for their onset/offset times. Designing learning algorithms for multi-label audio classification from weakly labelled data is a new challenge in audio tagging.

A popular framework for audio tagging with weakly labelled data is multiple instance learning (MIL) [6, 7, 8], where the time frames in an audio recording are treated as instances in a bag, and only the labels of the bag are given. A bag is considered as a positive bag if it contains at least one positive instance, otherwise negative. For a given MIL framework, there are two main MIL strategies: (i) instance-level approach (ii) embedding-level approach. The difference is that the former works by pooling the instance scores to obtain the bag scores, while the latter integrates the embedding-level features into bag representation and then directly carry out bag classifier. In [9], it was indirectly shown that the embedding-level approach is more efficient than the instance-level approach. Therefore, in this paper, we take the embedding-level approach, parameterizing through the multiple instance neural network (MINN), which contains three modules: (i) feature extraction to provide low-dimensional embedding, (ii) MIL pooling, and (iii) bag classifier.

The crucial module in exploiting the weakly labelled data lies in the use of an effective MIL pooling scheme to aggregate the instance-level embedding into bag-level features [10, 11]. The default choice is max or average pooling. Although these pooling functions achieve promising results, they are pre-defined, and less flexible for adapting to practical applications. In recent studies, efforts have been made to learn the adaptive pooling function and weights. For example, Kong *et al.* [12] proposed an attention model as a pooling function, which is achieved by a weighted sum of the results over frames. He *et al.* [13] proposed a hierarchical attention pooling structure. By assigning larger weights to the instance corresponding to the sound events, these methods could dynamically determine the contribution of each instance.

Inspired by these works, in this paper, we propose a new MIL pooling function, namely, gated multi-head attention pooling (GMAP), where we extend the vanilla attention pooling to a multi-head attention pooling function. In the proposed scheme, the encoded representations of the sequence are split into homogeneous sub-vectors, called heads. Features specific to different events from various heads are then aggregated, as the bag-level representation of the entire recording. Therefore, our proposed scheme can capture the sequence information from different subspaces at different positions. Since the information from multi-heads might be overlapping, a gate mechanism is used to fuse the information from various heads and remove the potential redundancies. Finally, the bag-level features obtained through GMAP are fed into the classifier to identify the events. We evaluated the proposed GMAP on Audioset. Experiments show that the proposed GMAP is superior to the non-adaptive pooling scheme and the vanilla attention pooling scheme. While this paper focuses on audio tagging, the proposed method could be applied to similar problems in other applications.

The paper is organized as follows. The multiple instance neural network and pooling schemes are described in Section 2. Next, the proposed pooling method is presented in Section 3, followed by the experimental setup and results in Section 4. Finally, we conclude the paper in Section 5.

2. Multiple Instance Learning

In this section, we present the MIL with neural networks, and discuss the existing pooling methods used and their limitations.

http://dx.doi.org/10.21437/Interspeech.2020-1197

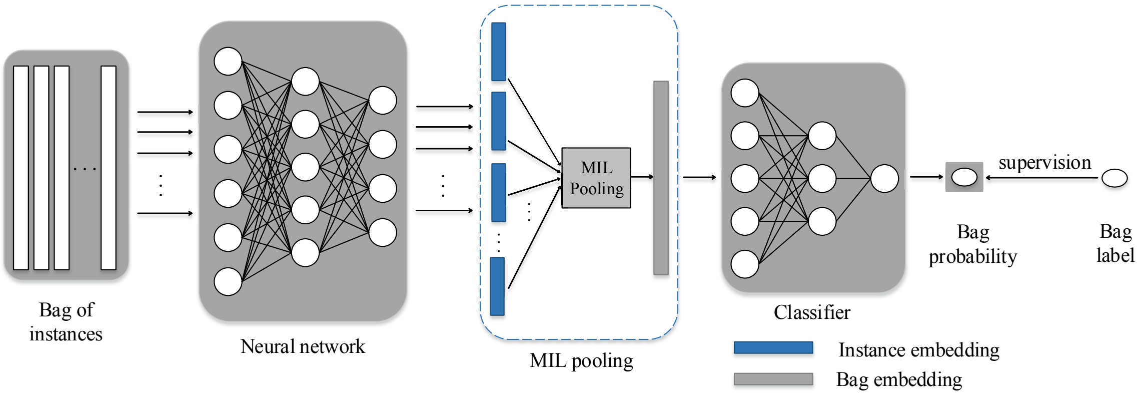


Figure 1: *Architecture of the multiple instance neural network.*

2.1. Multiple Instance Neural Network (MINN)

While MIL can be applied to a variety of conventional learning algorithms (e.g., support vector machines and nearest neighbor classifiers), it has been used recently with deep learning framework including convolutional neural networks (CNN), offering state-of-the-art performance for weakly labelled audio tagging [14, 15, 16].

For an audio recording, the log-mel spectrogram can be extracted and denoted as the i-th bag X_i. The instances in this bag can be denoted as $\{x_{ij}\}_{j=1}^{m_i}$, where x_{ij} is the j-th frame (or segment) in the bag X_i, and m_i is the number of instances in X_i. For the embedding-level approach, a neural network $f_\psi(\cdot)$ is used to transform the instances x_{ij} into a low-dimensional embedding (i.e. feature) e_{ij}. Then, the MIL pooling $\mathrm{P}(\cdot)$ is applied on the feature embedding e_{ij} to generate a bag-level embedding u_i. After that, a bag-level classifier $f_\vartheta(\cdot)$ is applied to estimate the probability θ of the audio events being present in the recording. The overview of MINN is illustrated in Figure 1. It can be summarized as:

$$\begin{aligned} e_{ij} &= f_\psi\left(x_{ij}\right) \\ u_i &= \mathrm{P}\left(e_{ij}\right) \\ \theta &= f_\vartheta\left(u_i\right) \end{aligned} \tag{1}$$

2.2. MIL Pooling

In the above process, the design of the MIL pooling layer is an essential issue in weakly labelled audio tagging. The max, average and attention pooling schemes are the most well-known and widely used ones [12, 17, 18]. Those pooling functions are introduced as follows.

Max pooling: The max pooling scheme simply inherits the largest activation of the instances in the bag. It can be described as follows:

$$u_i = \max_j e_{ij} = \max_j f_\psi\left(x_{ij}\right) \tag{2}$$

One limitation of max pooling is that only one instance in a recording can receive error signals, making optimization difficult and unstable. If a segment is selected as the key instance, any other events that do not appear in the segment are easily ignored.

Average pooling: In the average pooling scheme, the contribution of each instance is assumed to be the same, and equal weights are assigned to the instances. Under this assumption, the bag-level representation can be obtained by:

$$u_i = \frac{1}{|m_i|}\sum_j e_{ij} = \frac{1}{|m_i|}\sum_j f_\psi\left(x_{ij}\right) \tag{3}$$

Because all the instances in a positive bag are considered positive, the average pooling performs well when the events are long relative to the bag. However, this assumption may not hold in practice for relative short events, e.g., gunshot, and will produce a large number of false positives in the prediction stage [11].

Attention Pooling: In the attention pooling, each instance does not necessarily contribute equally to the bag-level representation. Instead, the bag can be aggregated based on the importance of each instance, as shown below:

$$u_i = \frac{\sum_j \omega\left(x_{ij}\right) f_\psi\left(x_{ij}\right)}{\sum_j \omega\left(x_{ij}\right)} \tag{4}$$

where $\omega(\cdot)$ is an attention function for calculating the weights of the instances.

With this attention mechanism, the bag representation usually focuses on a particular component of the spectrogram, often the part relevant to an event. However, there can be multiple events in an audio recording, especially for a long-duration recording. In addition, with this pooling scheme, all the attention information comes from the same low-dimension representation. Nevertheless, looking at the representations from different subspaces, there will be different but concurrent events. Due to these issues, the detection of events that occur in the recording can be incomplete.

3. Gated Multi-head Attention Pooling

In view of the limitations of the pooling schemes discussed above, we propose a gated multi-head attention pooling function. First, a single-head attention method is introduced, then it is extended to multi-head attention scheme in order to attend to different parts of the audio signal.

3.1. Single-head Attention Pooling

As we know, different segments would make different contributions to the recording-level features. We firstly implement single-head attention to measure the importance of instances, which is illustrated in Figure 2 (a). Suppose the segment-level embedding is $\boldsymbol{H} = \left[\boldsymbol{h}_1, \boldsymbol{h}_2, \ldots, \boldsymbol{h}_N\right]^T$ and $\boldsymbol{H} \in R^{N*d}$, the attention score ω_t of each feature $\boldsymbol{h}_t \in R^d$ is parameterized with a dedicated layer and softmax function (called the attention model), which is computed as:

$$\omega_t = \frac{\exp\left(\boldsymbol{v}^T \boldsymbol{h}_t\right)}{\sum_{l=1}^{N} \exp\left(\boldsymbol{v}^T \boldsymbol{h}_t\right)} \tag{5}$$

where $\boldsymbol{v} \in R^d$ is a trainable parameter and N is the number of segments.

After obtaining the weights over all the segments, the weighted mean vector $\tilde{\boldsymbol{\mu}}$ is calculated by pooling segments with a weighted sum:

$$\tilde{\boldsymbol{\mu}} = \sum_{t=1}^{N} \omega_t \boldsymbol{h}_t \tag{6}$$

In addition, the higher-order statistics (e.g., the standard deviation) [19] plays an important role since it contains event characteristics related to temporal variability over long contexts. Hence, the weighted standard deviation $\tilde{\boldsymbol{\sigma}}$, defined as follows, will be added as a part of recording-level features:

$$\tilde{\boldsymbol{\sigma}} = \sqrt{\sum_{t=1}^{N} \omega_t \boldsymbol{h}_t \odot \boldsymbol{h}_t - \tilde{\boldsymbol{\mu}} \odot \tilde{\boldsymbol{\mu}}} \tag{7}$$

where $\odot$ denotes element-wise multiplication and and square root is element-wise operation.

Finally, we concatenate the weighted mean and standard deviation to obtain the recording-level pooling features $\boldsymbol{c} \in R^{2d}$:

$$\boldsymbol{c} = [\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\sigma}}] \tag{8}$$

3.2. Gated Multi-head Attention Pooling

The gated multi-head attention pooling function is proposed to detect the event-related segments more precisely. Each head corresponds to a position of the encoded representation, and the weight in the head characterizes the presence of the events in that position.

As shown in Figure 2 (b), the segment level embedding $\boldsymbol{h}_t$ is firstly split to k sub-vectors $\boldsymbol{h}_t = [\boldsymbol{h}_{t1}, \boldsymbol{h}_{t2}, \ldots, \boldsymbol{h}_{tk}]$, where $\boldsymbol{h}_{tk} \in R^{d/k}$, and the vector in the same position forms a j-th subspace: $\boldsymbol{S}_j = [\boldsymbol{h}_{1j}, \boldsymbol{h}_{2j}, \ldots, \boldsymbol{h}_{Tj}]^T$. Secondly, the single-head attention pooling is performed in parallel to compute the weights for these subspaces, and to output a mixed representation. In particular, different attention is applied to different heads, and the attention weight of the head j at the step t is calculated as:

$$\omega_{tj} = \frac{\exp\left(\boldsymbol{v}_j^T \boldsymbol{h}_{tj}\right)}{\sum_{t=1}^{N} \exp\left(\boldsymbol{v}_j^T \boldsymbol{h}_{tj}\right)} \tag{9}$$

where $\boldsymbol{v}_j \in R^{d/k}$ is a trainable parameter. After that, each head of the weighted mean $\tilde{\mu}_j$ and standard deviation $\tilde{\boldsymbol{\sigma}}_j$ is computed as:

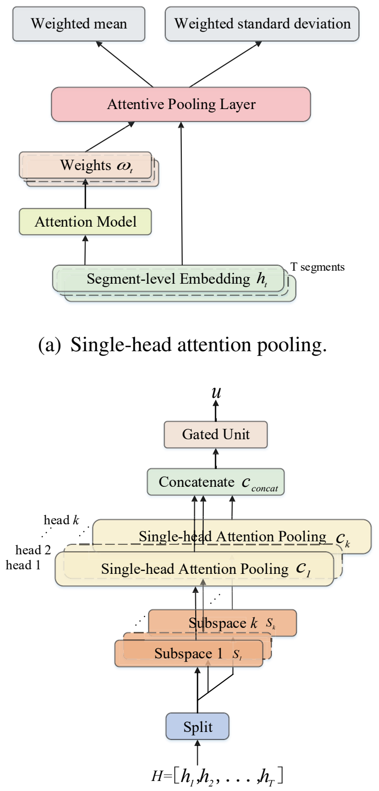
$$\begin{aligned} \tilde{\boldsymbol{\mu}}_j &= \sum_{t=1}^{T} \omega_{tj} \boldsymbol{h}_{tj} \\ \tilde{\boldsymbol{\sigma}}_j &= \sqrt{\sum_{t=1}^{N} \omega_{tj} \boldsymbol{h}_{tj} \odot \boldsymbol{h}_{tj} - \tilde{\boldsymbol{\mu}}_j \odot \tilde{\boldsymbol{\mu}}_j} \\ \boldsymbol{c}_j &= [\tilde{\boldsymbol{\mu}}_j, \tilde{\boldsymbol{\sigma}}_j] \end{aligned} \tag{10}$$

where $\boldsymbol{c}_j \in R^{2d/k}$ is the attentive pooling feature of the j-th subspace.

At this point, we focus on how to aggregate information among multiple positions. The purpose of information aggregation is to combine the partial representation captured by different attention heads into the final representation. As the information represented by these heads may be redundant, a gated mechanism is presented to remove redundancy, where the gated value is close to 0 if the information is redundant, otherwise it is close to 1 if information should be attended. Specifically, all the vectors produced by parallel heads are concatenated together to form a single vector $\boldsymbol{c}_{concat} = [\boldsymbol{c}_1, \boldsymbol{c}_2, \ldots, \boldsymbol{c}_k]$. Then, a gated unit similar to GLU [20] is followed to capturing dependencies among heads:

$$\boldsymbol{u} = g(\boldsymbol{W}(\boldsymbol{c}_{concat}) + \boldsymbol{b}) \odot \boldsymbol{c}_{concat} \tag{11}$$

where $\boldsymbol{u}$ is the final bag-level representation, g denotes the Sigmoid function, and $\boldsymbol{W}$ and $\boldsymbol{b}$ are trainable weights and bias parameters.



(a) Single-head attention pooling.

(b) Gated multi-head attention pooling.

Figure 2: *The illustration of single-head and multi-head attention pooling structure.*

4. Experiments

4.1. Database

We conduct experiments with Audioset [5], which consists of 527 categories of sound events and a collection of over 2 million 10-second excerpts of YouTube videos. Each audio clip may contain multiple labels. The dataset only provides labels at clip level (i.e. without time stamps for the events in frame levels, hence they are weak labels). There are three metrics used for

Table 1: *Classification results with different pooling schemes.*

Model	mAP	mAUC	d-prime
Max pooling	0.396	0.969	2.640
Average pooling	0.394	0.966	2.573
Max + average pooling	0.397	0.970	2.654
Vanilla attention pooling	0.404	0.969	2.631
Single-head attention pooling	0.406	0.970	2.652
GMAP	0.417	0.971	2.678

evaluation: mean average precision (mAP), mean area under the curve (mAUC), and d-prime. The higher the value of these metrics, the better the performance of the tagging method.

4.2. Features

On the basis of the frames, the audio signal is encoded by a Fourier transform based filter bank with 64 coefficients. A sequence of the T-frame spectrograms is stacked and the shape of each spectrogram is 400 (frames). Finally, the normalized spectrogram is used as the input of networks. It is also worth mentioning that AudioSet is an imbalanced dataset, and some sound classes (e.g. gunshot) have fewer training samples as compared with others. In our experiments, Mixup [21] and SpecAugment [22] are used to increase the number of small samples.

4.3. Model

To set up experiments, we implement CNN system as our baseline. The CNN with MIL has three components, the convolutional layers, the MIL pooling, and the classifier to produce predictions for recording-level probabilities of sound events. The convolutional layers consist of four convolutional blocks. Each block is composed of two 3×3 convolution layers followed by an average pooling layer. In addition, batch normalization and ReLU function are applied to all convolutional layers. The MIL pooling scheme is then used to transform the extracted representation into an overall recording-level feature. Finally, the aggregated feature is fed into two fully connected (FC) layers and a sigmoid layer to output the probabilities of sound events presented in the signal.

We build models trained with different MIL pooling functions, including non-adaptive pooling schemes (max, average and max + average similar to [23]), the single-head attention pooling and proposed GMAP scheme. GMAP refers to the best multi-head attention model that we have trained and the number of heads is chosen at 4. All models are implemented based on the same backbone network, and only the MIL pooling is changed.

4.4. Experimental results

Table 1 shows the results with different MIL pooling schemes. First of all, we can see that max pooling is slightly better than average pooling scheme in our experiment. Second, the performance improvement achieved by combining these two is marginal. In our proposed methods, single-head attention with standard deviation is more effective than the vanilla attention pooling, and GMAP further enhances the result by measuring the importance of instances at different positions. Figure 3 shows the weight values obtained by GMAP and the vanilla attention pooling in an audio recording. Comparing different heads with their weights, it can be found that the model is able to capture sound events from different subspaces. In addition, the vanilla attention weight is evenly distributed in the sequence, while GMAP focuses on a position and detects events at that position.

Table 2: *Comparison with various models in the literature.*

Model	mAP	mAUC	d-prime
Benchmark (2017)][5]	0.314	0.959	2.452
Xu *et al.* (2018) [7]	0.360	0.970	2.660
Kong *et al.* (2019) [8]	0.369	0.969	2.640
Logan *et al.* (2019) [16]	0.392	0.971	2.682
Kong *et al.* (2020) [23]	0.439	0.973	2.720
GMAP	0.417	0.971	2.678

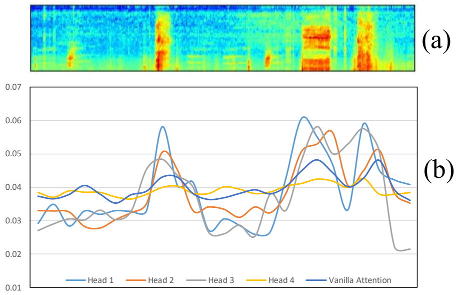


Figure 3: *(a) The spectrogram of an audio recording. (b) The trajectory of the attention weights extracted from GMAP and vanilla attention pooling over the above recording.*

Table 2 compares the proposed system with the state-of-the-art in the literature. Compared with other methods, the performance of our system is also competitive. The proposed system outperforms most methods except the recent results [23]. Note that [23] utilized the waveform and the log-mel spectrogram as inputs to CNNs, which significantly improved its performance. By contrast, our system achieves comparable performance without using a variety of features.

5. Conclusions

We have presented a gated multi-head attention pooling function (GMAP) for weakly labelled audio tagging. The proposed GMAP aims to detect event-related segments more precisely by assigning weights to different positions. Specifically, each head calculates the weighted mean and standard deviation to produce representations of the position. Then a gate function is used to fuse the information of subspaces to avoid redundancy. Empirical results show that the proposed method outperforms non-adaptive pooling schemes and achieves comparable results with the current state-of-the-art methods. In the future, we plan to investigate the effectiveness of the proposed methods for other related applications, such as speaker verification and speech emotion recognition.

6. Acknowledgements

This paper is supported by Shenzhen Science & Technology Fundamental Research Programs (No: JCYJ20170817160058246 & JCYJ20180507182908274).

7. References

[1] G. Parascandolo, H. Huttunen, and T. Virtanen, "Recurrent neural networks for polyphonic sound event detection in real life recordings," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 6440–6444.

[2] L. Barrington, A. Chan, D. Turnbull, and G. Lanckriet, "Audio information retrieval using semantic similarity," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, vol. 2. IEEE, 2007, pp. II–725.

[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)*. IEEE, 2016, pp. 1128–1132.

[4] S. Dimitrov, J. Britz, B. Brandherm, and J. Frey, "Analyzing sounds of home environment for device recognition," in *European Conference on Ambient Intelligence*. Springer, 2014, pp. 1–16.

[5] 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.

[6] 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.

[7] 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.

[8] 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*, vol. 27, 2019, pp. 1791–1802.

[9] X. Wang, Y. Yan, P. Tang, X. Bai, and W. Liu, "Revisiting multiple instance neural networks," *Pattern Recognition*, vol. 74, pp. 15–24, 2018.

[10] B. McFee, J. Salamon, and J. P. Bello, "Adaptive pooling operators for weakly labeled sound event detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 11, pp. 2180–2193, 2018.

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

[12] 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.

[13] K. X. He, Y. H. Shen, and W. Q. Zhang, "Hierarchical pooling structure for weakly labeled sound event detection," *arXiv preprint arXiv:1903.11791*, 2019.

[14] 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)*. IEEE, 2018, pp. 121–125.

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

[16] L. Ford, H. Tang, F. Grondin, and J. Glass, "A deep residual network for large-scale acoustic scene analysis," *Proc. Interspeech 2019*, pp. 2568–2572, 2019.

[17] J. Amores, "Multiple instance classification: Review, taxonomy and comparative study," *Artificial intelligence*, vol. 201, pp. 81–105, 2013.

[18] M. Ilse, J. M. Tomczak, and M. Welling, "Attention-based deep multiple instance learning," *arXiv preprint arXiv:1802.04712*, 2018.

[19] K. Okabe, T. Koshinaka, and K. Shinoda, "Attentive statistics pooling for deep speaker embedding," *arXiv preprint arXiv:1803.10963*, 2018.

[20] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modeling with gated convolutional networks," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017, pp. 933–941.

[21] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2017.

[22] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.

[23] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "Panns: Large-scale pretrained audio neural networks for audio pattern recognition," *arXiv preprint arXiv:1912.10211*, 2019.