Multi-Level Fusion based Class-aware Attention Model for Weakly Labeled Audio Tagging

Yifang Yin
National University of Singapore
Singapore, Singapore
idsyin@nus.edu.sg

Meng-Jiun Chiou
National University of Singapore
Singapore, Singapore
mengjiun@comp.nus.edu.sg

Zhenguang Liu*
Zhejiang Gongshang University
Hangzhou, China
liuzhenguang2008@gmail.com

Harsh Shrivastava
National University of Singapore
Singapore, Singapore
harsh.vardhan.shri@gmail.com

Rajiv Ratn Shah
IIIT-Delhi
Delhi, India
rajivratn@iiitd.ac.in

Roger Zimmermann
National University of Singapore
Singapore, Singapore
rogerz@comp.nus.edu.sg

ABSTRACT

Recognizing ongoing events based on acoustic clues has been a critical research problem for a variety of AI applications. Compared to visual inputs, acoustic cues tend to be less descriptive and less consistent in time domain. The duration of a sound event can be quite short, which creates great difficulties for, especially weakly labeled, audio tagging. To solve these challenges, we present a novel end-to-end multi-level attention model that first makes segment-level predictions with temporal modeling, followed by advanced aggregations along both time and feature domains. Our model adopts class-aware attention based temporal fusion to highlight/suppress the relevant/irrelevant segments to each class. Moreover, to improve the representation ability of acoustic inputs, a new multi-level feature fusion method is proposed to obtain more accurate segment-level predictions, as well as to perform more effective multi-layer aggregation of clip-level predictions. We additionally introduce a weight sharing strategy to reduce model complexity and overfitting. Comprehensive experiments have been conducted on the AudioSet and the DCASE17 datasets. Experimental results show that our proposed method works remarkably well and obtains the state-of-the-art audio tagging results on both datasets. Furthermore, we show that our proposed multi-level fusion based model can be easily integrated with existing systems where additional performance gain can be obtained.

CCS CONCEPTS

• **Computing methodologies** → **Neural networks**; • **Information systems** → *Multimedia databases*;

ACM Reference Format:
Yifang Yin, Meng-Jiun Chiou, Zhenguang Liu, Harsh Shrivastava, Rajiv Ratn Shah, and Roger Zimmermann. 2019. Multi-Level Fusion based Class-aware Attention Model for Weakly Labeled Audio Tagging. In *Proceedings of the 27th ACM International Conference on Multimedia (MM '19), October 21–25, 2019, Nice, France.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3343031.3351090

*The corresponding author.

MM '19, October 21–25, 2019, Nice, France

ACM ISBN 978-1-4503-6889-6/19/10…$15.00
https://doi.org/10.1145/3343031.3351090

1 INTRODUCTION

Audio as an import type of multimedia data contains rich and valuable information about what is happening around. Audio analysis such as sound event detection and acoustic scene classification has been attracting a continuously growing attention during the past years [24]. To recognize the sound events that happen in the audio can be of great help for many applications including surveillance, video indexing, smart cars, and context-aware services. For example, the accurate detection of warning and vehicle sounds such as car alarm and car passing by is an important and desired component in smart car applications [23].

An audio clip usually contains various foreground sounds and background noise, which makes the detection of sound events to be a highly challenging problem. To accelerate the development of advanced audio analysis techniques, Google recently released a large-scale audio dataset named AudioSet, which consists of over two million 10-second YouTube video clips with annotations of 527 sound events [9]. This dataset provides comprehensive coverage of real-world sounds to drive the development of data-driven machine learning based approaches in the field of audio processing. However, due to the difficulty in collecting the ground truth labels, only clip-level weak labels are available in the AudioSet for training. Figure 1 illustrates an example in the AudioSet. Compared to visual content analysis, weakly labeled audio tagging tends to be much more difficult and challenging due to: (1) the duration of a sound event is mostly much shorter than the visual presence of an object; and (2) acoustic features (*e.g.*, log-mel spectrogram) tend to be less representative and descriptive than visual features. While traditional methods assume the segments of an audio share the same clip-level labels [8], more recent studies tend to model the weakly labeled audio tagging as a multiple instance learning (MIL) problem [10, 31]. Audio clips are first divided into non-overlapping segments with a fixed window size, which are next processed and aggregated to generate clip-level predictions for supervision. As the duration of a sound event may vary significantly, most of the existing work focused on modeling the importance of different audio segments when making the final clip-level predictions [4,

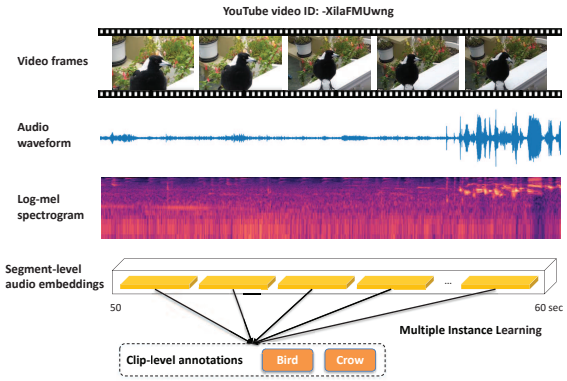

Figure 1: Illustration of an example in the AudioSet. Only clip-level annotations are available for training.

5, 21]. To increase the descriptiveness of acoustic features, multi-level features from intermediate layers of a neural network can be fused to improve the representation ability [5, 37]. There are a few studies that apply multi-level feature fusion, but mostly aggregate the features into a high-level representation based on simple solutions such as feature concatenation [17, 37] or weighted average fusion [13, 22]. More advanced multi-level feature fusion techniques have not yet been extensively studied in this field.

We present in this paper a novel end-to-end multi-level attention model for large-scale weakly labeled audio tagging. Our method belongs to the category of instance-level MIL approach. It first trains a classifier to make segment-level predictions, which are next aggregated based on class-aware attentions to obtain the clip-level predictions. On one hand, we adopt a bi-directional GRU based RNN as the instance-level classifier to incorporate the temporal patterns of sound events in a clip. On the other hand, we model the attention scores of each segment locally based on fully connected layers. This strategy is designed based on the observation that the importance of neighboring segments can be less correlated as the content may change frequently in a short period of time in audios. Moreover, we use class-specific attentions instead of a single class-agnostic importance score for each segment. This strategy, termed as class-aware attention, has been shown to be effective in weakly labeled sound event detection [4, 12]. To increase the descriptiveness of the input acoustic features, we further propose a novel multi-level feature fusion approach based on weight sharing to obtain improved accuracy in both segment-level and clip-level predictions. Layer-specific clip-level predictions are computed and averaged to obtain the final tagging results. Existing methods mostly fuse the features from intermediate layers by concatenation or weighted average [13, 17, 37], the performance of which is not robust due to the method's simplicity. Our method, comparatively, captures the common patterns among features from different layers based on shared-weight sub-networks, which reduces overfitting especially on small to moderate datasets. Moreover, our method is robust to the hyper-parameter settings. Based on a fusion of features from 10 intermediate layers, our method is still able to achieve the state-of-the-art classification accuracy (mean average precision of 0.361) on the AudioSet dataset.

We summarize the key contributions of this work as follows:

- We present a state-of-the-art end-to-end multi-level attention model, which simultaneously performs class-aware attention based temporal aggregation of segment-level predictions and average pooling based multi-layer fusion of clip-level predictions, for weakly labeled audio tagging.
- We propose a novel multi-layer feature fusion method and introduce weight sharing mechanism to reduce model complexity and overfitting. Earlier methods mostly fuse features based on concatenation or weighted average, which perform less robust due to method's simplicity.
- We integrate our proposed multi-level attention model with a top ranked system in DCASE17 challenge and train the whole system in an end-to-end manner. The audio tagging F1 score is improved by 2.6%, which indicates the integration of our proposed method in existing systems can lead to potential performance gain.
- Extensive experiments have been conducted on both the AudioSet and the DCASE17 datasets. We justify the effectiveness of each component of our multi-level attention model and compare it to the state-of-the-arts to demonstrate the advance of our proposed techniques.

The rest of the paper is organized as follows. We report the important related work in Section 2. Section 3 provides a formal definition of the audio tagging problem, and introduces the proposed single-level and multi-level attention models with weight sharing mechanism. Section 4 reports the experimental results on model justification and comparison with the state-of-the-art methods on two large-scale public datasets. Finally, Section 5 concludes and suggests future work.

2 RELATED WORK

Audio tagging is the task of recognizing the type of sound events that are present in an audio clip. Inspired by the great success of Deep Neural Networks (DNNs) on image classification [27, 34, 35], recently Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have been widely applied to audio processing where the state-of-the-art performances have been obtained in the field of, *e.g.*, acoustic event detection and acoustic scene classification [9, 36]. For instance, Xu *et al.* [33] presented a gated convolutional neural network for audio classification, which has won the 1st place in the audio tagging task of Detection and Classification of Acoustic Scenes and Events (DCASE) 2017 challenge. Chen *et al.* [4] proposed a class-aware self-attention model, which aims at generating discriminative clip-level feature representations for sound event detection. Kumar *et al.* described a CNN-based framework for sound event detection and classification with transfer learning [15]. They trained a model on supplementary audio data, which is next adapted for new tasks by introducing adaptation layers. Additionally, frameworks based on multimodal analysis have also been proposed recently, to achieve effective classification results by applying advanced feature fusion techniques [20, 25, 36].

Audio tagging based on weak labels is a highly challenging problem. Sound events mostly occur only for a short period of time in an audio clip, but weak labels are audio-level tags without the time stamps of the audio events. To solve this problem, multiple instance learning (MIL) can be applied where a single label is assigned to a bag of instances for training [10]. Wang *et al.* performed a comparative study of five multiple instance learning pooling functions for sound event detection, *i.e.*, max pooling, average pooling, linear softmax, exponential softmax, and attention pooling [31]. Kong *et al.* [12] presented an attention-based CNN model with mini batch balancing to tackle the imbalance problem on the Audio Set dataset [9]. Chou *et al.* proposed to train their model by considering both clip-level and segment-level supervisions [5]. Guo *et al.* presented a novel attention-based DNN framework that takes advantages of both frequency modeling with CNNs and temporal modeling with RNNs for acoustic scene classification [6]. Yu *et al.* further proposed a decision-level attention model that applies multiple attention modules on the intermediate layers of a neural network [37], the outputs of which are next concatenated to generate a clip-level representation for supervision. Kong *et al.* [13] further presented a feature-level attention model that aggregats segment-level features based on attention pooling to generate a clip-level representation. The utilization of multi-level features from different layers of a neural network have been shown to be effective in the field of video and audio processing [17, 22]. The extracted features from different layers are mostly concatenated to form a single vector, based on which a new classifier will be trained to make the final predictions.

3 MULTI-LEVEL CLASS-AWARE ATTENTION MODEL FOR AUDIO TAGGING

Weakly labeled audio tagging is typically modeled as a Multiple Instance Learning (MIL) problem [31]. Here we first formulate the problem and present a single-level recurrent neural network with instance-level class-aware attention pooling to solve the MIL problem. Next, we extend the framework to a novel multi-level attention model with weight sharing mechanism for efficient training and accurate tagging.

3.1 Problem Formulation

Given a training set $D = \{(\mathbf{x}, \mathbf{y})|\mathbf{x} \in \mathbf{X}, \mathbf{y} \in \{0,1\}^C\}$, where $\mathbf{X}$ is a set of audio clips, $\mathbf{y}$ denotes a binary vector of clip-level labels, and C is the number of sound events. Our goal is to learn a model from D supervised by clip-level labels $\mathbf{y}$ to predict the occurrence of the C events in an input audio clip. As multiple sound events may occur in a single audio clip, we formulate the audio tagging task as a multi-label classification problem.

3.2 Class-aware Attention based Multiple Instance Learning

Recently, attention model based Recurrent Neural Networks (RNN) have been widely adopted for video analysis such as video captioning [19] and human action recognition [28]. However, only a few attempts have been made on audio processing where the improvements introduced by the attention module are quite limited [6, 32]. Here we present an effective class-aware attention based RNN for audio tagging. The system overview is illustrated in Figure 2, which consists of three components, namely the feature encoder, the class-aware attention module, and the classifier.

Feature Encoder: The feature encoder module takes acoustic features (*e.g.*, raw waveforms [3], log-mel spectrogram [36], MFCC [2], *etc.*) extracted from the input clip, and generates high-level semantic embeddings, which will be next used for segment-level attention estimation and sound event occurrence prediction. As audios can be of arbitrary length, the input clip x is generally divided into non-overlapping frames with a fixed window size, denoted as $\mathbf{x} = \{\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_M}\}$, before being passed to the feature encoder network. For example, Google utilized a VGG-inspired model [8], which takes log-mel spectrogram of 0.96 s audio segments as the input and transforms it to a 128-dimensional acoustic embedding. In our system, the feature encoder is a relatively independent component. Pre-trained feature encoder such as the Google VGGish network trained on the YouTube-8M [8] can be leveraged to extract the acoustic embeddings first. It is also possible to implement a task-specific feature encoder and train the whole system in an end-to-end fashion. The acoustic embeddings generated by the feature encoder is denoted as $\mathbf{h} = \{\mathbf{h_1}, \mathbf{h_2}, ..., \mathbf{h_M}\}$ where $\mathbf{h_i}$ corresponds to $\mathbf{x_i}$, representing the acoustic feature of the $i-$th segment of the input clip.

Class-aware attention: As aforementioned, an input clip is divided into non-overlapping segments, and the importance of different segments can vary significantly regarding its contribution to the final prediction results. Instead of using the average pooling strategy, we introduce the class-aware attention module to dynamically adjust the importance of segment i based on $\mathbf{h_i}$. Let $\mathbf{s_i} = \{s_{i,1}, s_{i,2}, ..., s_{i,C}\}$ represent the class-aware importance score of segment i, where $s_{i,c}$ is the importance of segment i corresponding to sound event c. We compute $\mathbf{s_i}$ as

$$\mathbf{s_i} = \text{Sigmoid}(\mathbf{W_{att}}\mathbf{h_i} + \mathbf{b_{att}}) \tag{1}$$

where $\mathbf{W_{att}}$ and $\mathbf{b_{att}}$ are the learnable parameter matrix and bias vector, respectively. Here we choose Sigmoid as the activation function to make the importance score $\mathbf{s_i}$ be in the range of $[0, 1]$. We further apply a class-wise L1 normalization to $\mathbf{s}$ as

$$\alpha_{i,c} = \frac{s_{i,c}}{\sum_{i=1}^{M} s_{i,c}} \tag{2}$$

to ensure that $\sum_{i=1}^{M} \alpha_{i,c} = 1$ for each sound event c. The class-aware importance scores $\alpha_{i,c}$ estimated by the attention module will be next used for the attention-based fusion of segment-level predictions. As aforementioned, since audio content may change frequently in a short period of time, the importance of neighboring segments tends to be less correlated. Therefore, we model the class-aware attention based on segment embeddings locally using fully-connected layers only.

Classifier: The classifier takes the acoustic embeddings $\mathbf{h} = \{\mathbf{h_1}, \mathbf{h_2}, ..., \mathbf{h_M}\}$ as input and outputs segment-level predictions $\mathbf{p} = \{\mathbf{p_1}, \mathbf{p_2}, ..., \mathbf{p_M}\}$. Here we choose the gated recurrent unit (GRU) based RNN as our classifier. The acoustic features $\mathbf{h}$ are fed to a single layer bi-directional GRU RNN, followed by a fully-connected layer with Sigmoid activation. Unlike the attention

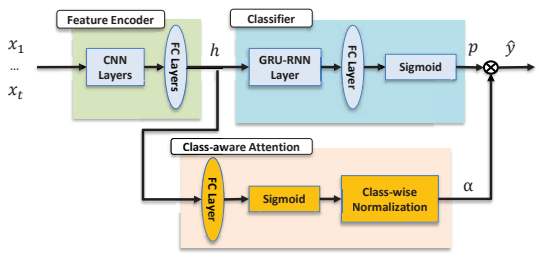

Figure 2: Architecture of the single-level class-aware attention model.

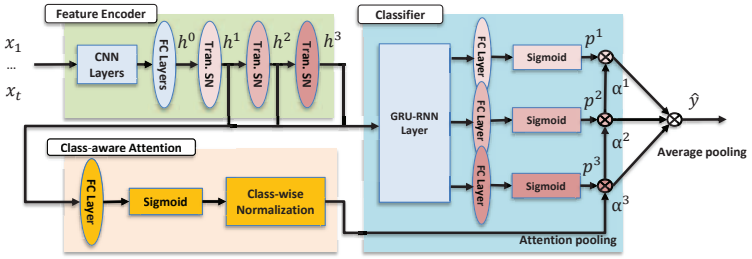

Figure 3: Overall architecture of our proposed multi-level class-aware attention model with weight sharing mechanism.

modeling, we have found during experiments that it is beneficial to incorporate temporal information to make segment-level predictions. Therefore we choose the RNN as our classifier for the temporal modeling. The segment-level predictions are aggregated into the clip-level prediction $\hat{\mathbf{y}}$ based on the class-aware attention scores $\alpha_\mathbf{i}$ as

$$\hat{\mathbf{y}} = \sum_{i=1}^{M} \alpha_\mathbf{i} \odot \mathbf{p_i} \tag{3}$$

where $\alpha_\mathbf{i} = [\alpha_{i,1}, \alpha_{i,2}, ..., \alpha_{i,C}] \in \mathbb{R}^{1\times C}$, $\mathbf{p_i} \in \mathbb{R}^{1\times C}$, and $\odot$ represents the element-wise multiplication of the two vectors. Subsequently, we formulate the final objective function as

$$L_s = \xi(\mathbf{y}, \hat{\mathbf{y}}) = -\sum_{c=1}^{C} (y_c \log(\hat{y}_c) + (1 - y_c)\log(1 - \hat{y}_c)) \tag{4}$$

where y_c and $\hat{y}_c$ are the true and predicted clip-level scores of sound event c, respectively.

The class-aware attention mechanism has been recently proposed to address the issue of average pooling in weakly labeled audio tagging [4, 12]. It has the advantage of making the model to learn from data, in order to highlight relevant segments as well as to suppress irrelevant noises at the same time for every sound event. Existing work can be divided into two categories: embedding-level MIL approach and instance-level MIL approach. Embedding-level MIL approach [4] uses attention pooling to generate a clip-level representation, based on which a clip-level classifier is trained to provide the final prediction. Instance-level MIL approach [12] trains instance-level classifier to obtain segment-level predictions, which are next aggregated based on attention pooling to obtain the final prediction. Our method belongs to the second category. We improve the previous work [12] by incorporating the temporal patterns during prediction, while Kong *et al.* processed the audio segments independently.

3.3 Multi-Level Fusion based Attention Model with Weight Sharing

Previous work has shown that improved predictions can be obtained by utilizing multi-level features from intermediate layers of the neural network [17, 37]. We follow this path and propose a novel end-to-end multi-level attention model with weight sharing mechanism. The system overview is illustrated in Figure 3. Let $\mathbf{h^0} = \{\mathbf{h_1^0}, \mathbf{h_2^0}, ..., \mathbf{h_M^0}\}$ represent the output feature of one intermediate layer in the feature encoder module. Here we propose to transform $\mathbf{h^0}$ into a set of different feature embeddings with the same dimension, denoted as $\mathbf{H} = \{\mathbf{h^1}, \mathbf{h^2}, ..., \mathbf{h^K}\}$, via K subsequent non-linear transformations to increase the representation ability of the input samples. Instead of using only the final output feature $\mathbf{h^K}$, our multi-level attention model use all the K features $\mathbf{H} = \{\mathbf{h^1}, \mathbf{h^2}, ..., \mathbf{h^K}\}$ for supervision. Our initial attempt is to apply multiple attention modules and classifiers, which have the same architecture as in the single-level model introduced in Section 3.2, to every $\mathbf{h^k} \in \mathbf{H}$. Let $\alpha^\mathbf{k}$, $\mathbf{p^k}$ and $\hat{\mathbf{y}}^\mathbf{k}$ represent the class-aware attention, segment-level prediction, and clip-level prediction learnt based on $\mathbf{h^k}$, respectively. We formulate the final objective function of the multi-level attention model as

$$L_m = \frac{1}{K}\sum_{k=1}^{K} \xi(\mathbf{y}, \hat{\mathbf{y}}^\mathbf{k}) \tag{5}$$

Subsequently, the final prediction $\hat{\mathbf{y}} = \frac{1}{K}\sum_{k=1}^{K} \hat{\mathbf{y}}^\mathbf{k}$ is computed based on an average pooling of $\mathbf{y}^k$, where K equals to the number of features in $\mathbf{H}$. Generally speaking, this strategy is able to obtain the state-of-the-art audio tagging performance, but it has one drawback that the number of parameters in the attention and the classifier modules grows linearly with the number of features in $\mathbf{H}$. To optimize system performance, we propose to reduce the number of learnable parameters based on weight sharing. As aforementioned, the features in $\mathbf{H}$ are designed to have the same dimension to enable the weight sharing mechanism in both the attention module and the classifier. Through the experiments, we have found that it is beneficial to share the weights of the subnetwork of the attention module and the GRU cells in the classifier among different $\mathbf{h^k}$. We remain the parameters in the fully-connected layers of the classifier to be dependent on $\mathbf{h^k}$, which is sufficient to maintain the classification accuracy. This weight sharing strategy greatly reduces the number of learnable parameters while being able to maintain the classification accuracies on large-scale training datasets. Moreover, the non-linear transformations of $\mathbf{h^0}$ improve the diversity of the feature representations, which reduce overfitting especially when training on small to moderate size datasets. In such scenarios, improved classification performance can be obtained compared to the multi-level attention model without applying the weight sharing mechanism.

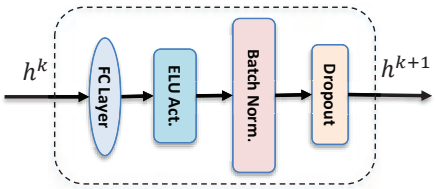

Figure 4: The sub-network in the multi-level attention model that transforms feature embedding $\mathbf{h^k}$ to $\mathbf{h^{k+1}}$.

Previous work on multi-level feature fusion mostly aggregates features into a high-level representation (*e.g.*, based on feature concatenation [17, 37] or weighted average fusion [13, 22]) before passing to the classifier. However, such methods can be sensitive to both the depth of the neural network and the intermediate features to be fused. For example, the concatenation based feature fusion method proposed by Yu *et al.* [37] obtained the best mAP when fusing the output features of the second and the third intermediate layers, compared to other combinations of intermediate features for fusion. Comparatively, our weight sharing based feature fusion approach is more reliable as being less sensitive to the number of features in $\mathbf{H}$. This is a desirable characteristic as it makes the tuning of the hyper-parameters of the model much easier than previous work. Moreover, our approach reduces model overfitting by using shared weight sub-networks in the attention and the classifier modules. Significant improvements have been obtained when training on the balanced AudioSet compared to existing methods. Finally, our multi-level model has the advantage of being parallel to existing audio tagging systems, and thus can be easily integrated to achieve additional performance gain. Recall that the feature encoder is a relatively independent component in our model, we can adopt any existing network architecture as an initial feature encoder to generate feature $\mathbf{h^0}$, and train the whole system (*i.e.*, the initial feature encoder, together with the non-linear transformation layers, the attention and the classifier modules) in an end-to-end fashion.

3.4 Implementation Details

In our multi-level attention model, we use a sub-network as illustrated in Figure 4 to transform feature embedding $\mathbf{h^k}$ to $\mathbf{h^{k+1}}$. A feature embedding $\mathbf{h^k}$ is first fed to a fully-connected layer followed by the exponential linear unit (ELU) as the activation function. Thereafter, we use batch normalization [11] and Dropout with the dropout rate set to 0.5. Our multi-level model starts with $\mathbf{h^0}$, and transforms it into the subsequent feature embeddings $\mathbf{h^k}$ by recursively applying the sub-network in Figure 4. For the AudioSet, we use the 128-D acoustic features provided by Google as $\mathbf{h^0}$ and set the number of hidden units to be 1024 in the fully-connected layer. For the DCASE17 dataset, we adopt the system developed by Xu *et al.* [33] as the feature encoder and take the input of the RNN as $\mathbf{h^0}$, which has a dimension of 256 for each segment. We set the number of hidden units to be 256 in the fully-connected layer in Figure 4, and train the whole system in an end-to-end fashion. The number of hidden units in the GRU cell in our classifier is set to 256 and 128 for the AudioSet and DCASE17 datasets, respectively.

Class imbalance is commonly seen in large-scale datasets. For example in AudioSet, the number of samples varies from hundreds to tens of thousands among different classes. To solve this problem, we adopt the mini batch balancing strategies used in our competitors [12, 33] for the AudioSet and DCASE17 datasets, respectively, to make it a fair comparison. For optimization, we train the neural networks using the Adam optimizer with a batch size of 500 and 64 for the AudioSet and DCASE17 datasets, respectively. The learning rate is set to 0.001.

The results generated by our proposed multi-level attention model trained on the same dataset may still differ slightly due to the randomness in neural networks. To solve this problem, we apply network ensemble techniques [14] to combine the scores of individually trained models to obtain improved classification accuracy. In our experiments, we train both the single-level and multi-level models five times with random seeds and take their average as the final prediction scores for each class.

4 EVALUATION

We first introduce the experimental setup in Section 4.1, and then proceed with the evaluations by performing a step-by-step model justification and a comparison with the state-of-the-art approaches in weakly labeled audio tagging.

4.1 Experimental Setup

We evaluated our proposed methods using two large-scale weakly labeled audio datasets, namely the AudioSet [9] and the DCASE17 sound event detection for smart cars [23]. The AudioSet consists of more than 2M audio clips with 527 annotated sound events, which is further divided into three disjoint sets: a balanced test set, a balanced training set, and an imbalanced training set. The sound events are organized using an expert-defined hierarchical ontology, which consists of seven classes on the highest level, namely human, animal, things, music, natural, environment, and source-ambiguous for short. The DCASE17 dataset contains around 50K audio clips annotated with 17 sound events. All the samples are 10-second sound clips drawn from YouTube videos.

Following previous work, we use mean average precision (mAP), and mean area under ROC curve (mAUC) as the evaluation metrics on the AudioSet. For the DCASE17 sound event detection challenge, we use a global confidence threshold of 0.3 for all the 17 sound events, and report precision, recall, and F1 score as the evaluation metrics.

4.2 Step-By-Step Model Justification

We perform a step-by-step model justification to demonstrate the effectiveness of each component in our proposed multi-level attention model with weight sharing. Based on the results, we discuss the characteristics of our proposed model and its advantages over the existing work. We also visualize the t-SNE embeddings of the features learnt by different layers of our proposed multi-level attention model for qualitative evaluation.

Single-level vs. Multi-level Attention Model. We first compared our multi-level attention model to the single-level attention model introduced in Section 3.2, and reported the mean average precision (mAP), and mean area under ROC curve (mAUC)

Table 1: Audio tagging performance comparison (in %) of our proposed single-level and multi-level models on the AudioSet, using AudioSet-22K for training.

Depth	Single-level		Multi-level	
	mAP	mAUC	mAP	mAUC
3	27.8	95.4	28.8	95.5
4	27.8	95.4	29.1	95.6
5	27.8	95.3	29.5	95.5

Table 2: Audio tagging performance comparison (in %) of our proposed single-level and multi-level models on the AudioSet, using AudioSet-2M for training.

Depth	Single-level		Multi-level	
	mAP	mAUC	mAP	mAUC
3	34.9	96.7	36.3	96.9
4	35.2	96.7	36.4	96.9
5	34.9	96.6	36.6	97.0

obtained on the AudioSet. Table 1 shows the results obtained using the balanced training set (AudioSet-22K) and Table 2 shows the results obtained using the union of the balanced and imbalanced training sets (AudioSet-2M). As can be seen, the multi-level model outperformed the single-level model in all cases. With the depth of the neural network increasing from 3 to 5, the multi-level model obtained continuously improving results, while the single-level model performed approximately the same. With a depth of five, the multi-level model outperformed the single-level model by 6.1% and 4.9% based on AudioSet-22K and AudioSet-2M, respectively. The single-level model only used the final output feature $\mathbf{h^K}$ (K is the depth of the neural network) of the feature encoder for supervision. Comparatively, the multi-level model fused all the K intermediate features in $\mathbf{H} = \{\mathbf{h^1}, \mathbf{h^2}, ..., \mathbf{h^K}\}$ to make predictions, which greatly improved the descriptiveness of audio representations and thus enhanced the model's robustness.

Evaluation on the Weight Sharing Mechanism. We reported the results obtained by our proposed multi-level attention model with and without the weight sharing mechanism in Tables 3 and 4. From Table 3 we can see that improved results have been achieved by weight sharing using AudioSet-22K in all depths. This is because the shared attention module and GRU cells among different $\mathbf{h^k}$ is able to reduce overfitting to some extent when the number of training samples is insufficient. On the other hand, to maintain the classification accuracy, we propose to use independent fully-connected layers after the GRU cells in the classifier to generate layer-specific predictions $\mathbf{p^k}$. This strategy has been shown to be effective. As shown in Table 4, with significantly reduced number of learnable parameters in the model based on weight sharing, we are still able to obtain competitive classification results when training with extremely large-scale datasets where overfitting can be less of a problem.

Table 5 shows the per-class mAP comparison of the multi-level attention model with the depth set to five. We group audio events according to the highest level of the AudioSet ontology, and report

Table 3: Audio tagging performance comparison (in %) of the multi-level attention model with and without the weight sharing mechanism, using AudioSet-22K for training.

Depth	w/ weight sharing		w/o weight sharing	
	mAP	mAUC	mAP	mAUC
3	28.8	95.5	28.4	95.4
4	29.1	95.6	29.0	95.4
5	29.5	95.5	29.2	95.4

Table 4: Audio tagging performance comparison (in %) of the multi-level attention model with and without the weight sharing mechanism, using AudioSet-2M for training.

Depth	w/ weight sharing		w/o weight sharing	
	mAP	mAUC	mAP	mAUC
3	36.3	96.9	36.4	96.8
4	36.4	96.9	36.6	96.9
5	36.6	97.0	36.6	96.9

Table 5: mAP performance comparison (in %) of the multi-level attention model with and without the weight sharing mechanism among different high-level audio classes in the AudioSet.

Class	AudioSet-22K		AudioSet-2M	
	w/ ws	w/o ws	w/ ws	w/o ws
Human	28.9	28.3	35.7	35.8
Animal	25.3	25.0	38.2	38.3
Things	28.6	28.2	36.9	37.0
Music	36.8	36.4	40.9	40.8
Natural	30.7	30.7	37.1	37.0
Environment	19.3	19.4	23.3	23.2
Source-ambiguous	18.9	18.7	26.1	26.0

the mAP of the audio events belonging to each of the high-level audio classes. For models trained on the AudioSet-22K, the weight sharing mechanism improved the mAP in the majority of the seven classes with only a slight mAP decrease in the Source-ambiguous class. For models trained on the AudioSet-2M, the mAP results obtained with and without the weight sharing mechanism are approximately the same among different high-level audio classes.

Evaluation on the Depth of the Neural Network. One advantage of our proposed model is that our model is less sensitive to the hyper-parameters, *e.g.*, the depth of the neural network. For evaluation, we reported the classification results obtained by our proposed multi-level model with different depths in Table 6. Please note that the depth also equals to the number of features in $\mathbf{H}$ to be fused in our model. We compared our method to one state-of-the-art embedding-level MIL approach with attention pooling for weakly labeled audio tagging [13]. As can be seen, the attention-based feature fusion method proposed by Kong *et al.* [13] obtained the best mAP of 0.361 with a depth of three. However, their method

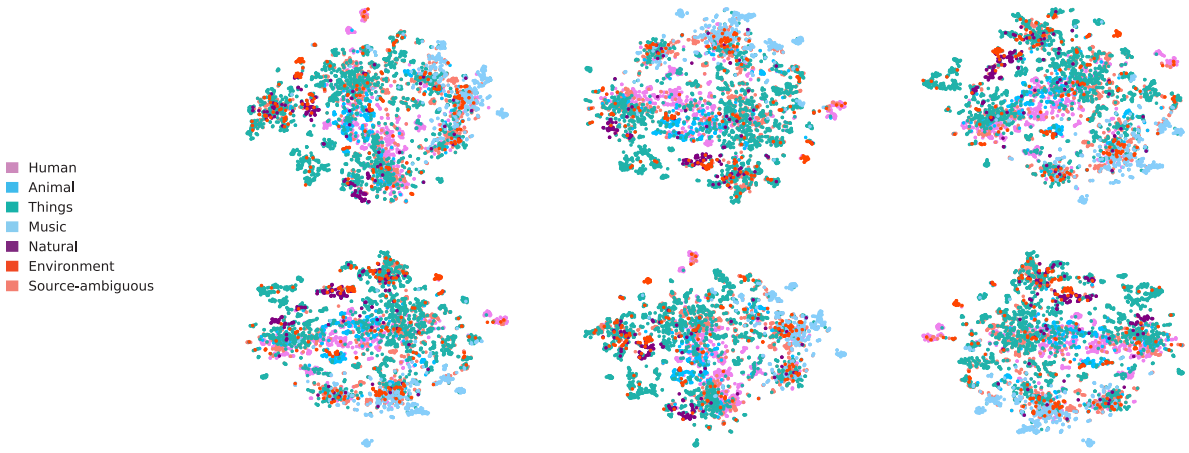

Figure 5: t-SNE embeddings using the features from different layers in the proposed multi-level attention model. The features are $\hat{y}^1$ with mAP=0.325, $\hat{y}^2$ with mAP=0.331, and $\hat{y}^3$ with mAP=0.332, $\hat{y}^4$ with mAP=0.325, $\hat{y}^5$ with mAP=0.331, and $\hat{y} = \frac{1}{5}\sum_{k=1}^{5}\hat{y}^k$ with mAP=0.337, respectively, from left to right and top to bottom.

Table 6: Performance comparison (in %) with the state-of-the-art attention model based on different layer depths, using AudioSet-2M for training.

Depth	Kong *et al.* [13]		Ours	
	mAP	mAUC	mAP	mAUC
2	35.8	96.8	36.0	96.7
3	36.1	96.9	36.3	96.9
4	35.6	96.9	36.4	96.9
6	34.8	96.8	36.4	97.0
8	33.9	96.7	36.3	96.9
10	33.1	96.6	36.1	96.9

is sensitive to the depth of the neural network, where the mAP dropped significantly to 0.331 with a depth of ten. Comparatively, our method is more robust as it outperformed the competitor in all different settings of the depths. Moreover, our method is less sensitive to the variation of the depth. With a depth of ten, our model is still able to obtain an excellent mAP of 0.361, which outperformed our competitor by 9.1% in terms of the mAP.

Visualizations. In order to compare the features learnt by different layers of our proposed multi-level attention model, we performed a two dimensional t-SNE embedding [29] and visualized the results in Figure 5. The embeddings were generated on a subset of the AudioSet evaluation set, which consisted of 4228 audio clips annotated with only one label that belongs to only one of the seven classes in the highest level of the AudioSet ontology. We used the 527 dimensional feature $\hat{y}^k$ generated by the five layer attention model and reported the corresponding mAP over the 292 classes remaining in this subset. With weight sharing strategy, the layer-specific clip-level predictions presented variations, but shared common patterns at the same time. The mAP obtained by an individual layer k varied from 0.325 to 0.332, while the average aggregation of all the five layers achieved the highest mAP of 0.337. It indicates the effectiveness of our proposed multi-level feature fusion approach based on weight sharing.

4.3 Comparison with the State-of-the-art

We compare our proposed multi-level attention model to the state-of-the-art audio tagging methods on the AudioSet. The results are reported in Table 7. Our method outperformed the existing methods trained on AudioSet-20K by a large margin. Chen *et al.* [4] proposed a class-aware attention based embedding-level MIL approach. They first generated a clip-level audio representation, based on which a multi-label classifier was trained. Their model was trained based on the union of the balanced training set and 30% of the unbalanced training set. Kong *et al.* [12] proposed a class-aware attention based instance-level MIL approach. They first trained an instance-level classifier, and aggregated the scores based on the class-aware attention. Both of the methods used the 128-dimensional acoustic features pre-trained on the YouTube-8M as the input and adopted class-aware attention in their system. However, they only used the features from one layer in their models for training, which limited the performance of their proposed methods. To solve this problem, M&mnet-MS [5] and DNN-MULTI-ATT [37] were proposed to make use of multiple features from intermediate layers for supervision. Chou *et al.* trained the M&mnet-MS network using about 1M samples by removing the clips that were only annotated with either Music or Speech, the two most popular classes. Yu *et al.* was able to achieve a better classification accuracy by introducing a mini batch balancing strategy and training with the whole AudioSet samples. However, Yu *et al.* fused the multi-features based on the traditional method of feature concatenation. Their system was able to obtain the best mAP of 0.36 by fusing the features of the second and third intermediate layers, but performed unstably among other combinations of intermediate features for fusion.

Table 7: Audio tagging performance comparison (in %) on the AudioSet. The results of ResNet-ATT and ResNet-SPDA are cited from work [5]; the others are cited from their original papers.

Methods	# Train Recs.	mAP	mAUC
WLAT [15]	22K	21.3	92.7
ResNet-ATT [32]	22K	22.0	93.5
ResNet-SPDA [7]	22K	21.9	93.6
M&mnet-MS [5]	22K	23.2	94.0
Ours	22K	29.5	95.5
ClaAware-ATT [4]	600K	31.6	-
M&mnet-MS [5]	1M	32.7	95.1
TALNet [31]	2M	36.2	96.5
DNN-ATT [12]	2M	32.7	96.5
DNN-MULTI-ATT [37]	2M	34.0-36.0	96.9-97.0
Ours	2M	**36.6**	**97.0**

Instead of using the 128-dimensional acoustic features, Wang *et al.* [31] trained their TALNet based on the filterbank features extracted from audio raw waveforms. However, this significantly increased the number of model parameters and the time complexity of model training, without achieving equivalent performance gain in terms of audio tagging. Our model outperformed all the existing methods. Compared to applying multiple attention modules to different intermediate layers independently [37], our method is more computationally efficient as it reuses the attention sub-network and the GRU cells for different features by weight sharing. Moreover, our model is more robust to the change of neural network depth, which reduces the time complexity in model hyper-parameter tuning.

Next, we compare our method to the state-of-the-art systems that participated in the DCASE 2017 challenge of sound event detection for smart cars. Vu *et al.* [30] presented an attention-based DNN model, which obtained an F1 score of 0.518 on the DCASE17 test dataset. Lee *et al.* [16] proposed to use an ensemble of convolutional neural networks to detect the weakly labeled audio events. Each of the networks was trained based on various lengths of analysis windows for input scaling. Their method obtained the best precision of 0.703 and an F1 score of 0.57. Xu *et al.* [33] presented a gated convolutional neural network with attention-based temporal aggregation method for audio event detection. Our method adopted their gated CNN as the initial feature encoder to generate $\mathbf{h^0}$, built our own non-linear transformation layers, attention and classifier sub-networks, and trained the whole system in an end-to-end manner. We used log-mel spectrogram as the only feature for input. As can be seen, our method obtained the best recall and F1 score on the DCASE17 test set. By applying our proposed multi-level feature fusion strategy, we were able to improve the F1 score by 2.6%, compared to the F1 score of 0.567 obtained by Xu *et al.* without making use of the features generated by the intermediate layers. It is worth mentioning that our proposed multi-level feature fusion approach can be integrated with any existing single-level models. The improved results we obtained on the DCASE17 test set indicates that the integration of our proposed multi-feature fusion approach in existing network architectures can further boost the system's performance.

Table 8: Audio tagging performance comparison (in %) on the DCASE17 test dataset.

Methods	Precision	Recall	F1
Lee *et al.* [18]	37.6	45.7	41.2
Adavanne *et al.* [1]	47.5	39.6	43.2
Salamon *et al.* [26]	44.7	47.0	45.9
Vu *et al.* [30]	54.2	49.5	51.8
Xu *et al.* [33]	53.8	60.1	56.7
Lee *et al.* [16]	**70.3**	47.9	57.0
Ours	56.0	**60.6**	**58.2**

5 CONCLUSION AND FUTURE WORK

We model weakly labeled audio tagging as a multiple instance learning problem and present a novel multi-level attention model to perform effective temporal score aggregation and multi-layer feature fusion. Our model is an instance-level MIL approach. We segment an input audio clip into segments and aggregate the temporal segment-level scores based on class-aware attention pooling. To improve the descriptiveness of acoustic representations, we use multiple features obtained by a subsequent non-linear transformations for supervision. The features are processed by layer-specific class-aware attention and classifier sub-networks, followed by average pooling to obtain the final clip-level predictions. To reduce model complexity and overfitting, we propose a weight sharing strategy among the attention and classifier sub-networks, which significantly reduces the number of learnable parameters while being able to obtain competitive or even better audio tagging results. We have conducted extensive experiments on the AudioSet and the DCASE17 datasets. The experimental results show that our proposed method outperforms existing methods and achieves the state-of-the-art audio tagging results on both of the datasets.

Currently, we only use one acoustic feature as the model input, *e.g.*, 128-D embedding on the AudioSet and log-mel spectrogram on the DCASE17 dataset, and fuse features from intermediate layers for supervision. In the future, we plan to apply our proposed feature fusion approach to multimodal inputs of different acoustic features such as raw waveforms, MFCC, and log-mel spectrogram. Moreover, our multi-level feature fusion approach can be easily integrated with existing single-level models. Therefore, we are also interested in investigating the use of our proposed feature fusion technique in various audio analysis problems in addition to audio tagging.

ACKNOWLEDGEMENT

This research is supported by Singapore Ministry of Education Academic Research Fund Tier 1 under MOE's official grant number T1 251RES1820. Rajiv Ratn Shah is partly supported by the Infosys Center for AI, IIIT Delhi and ECRA Grant by SERB, Government of India.

REFERENCES

[1] Sharath Adavanne and Tuomas Virtanen. 2017. *Sound Event Detection Using Weakly Labeled Dataset with Stacked Convolutional and Recurrent Neural Network.* Technical Report. DCASE2017 Challenge.

[2] J.-J. Aucouturier, B. Defreville, and F. Pachet. 2007. The Bag-of-frame Approach to Audio Pattern Recognition: A Sufficient Model for Urban Soundscapes But Not For Polyphonic Music. *Journal of the Acoustical Society of America* (2007), 881–91.

[3] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. 2016. SoundNet: Learning Sound Representations from Unlabeled Video. In *International Conference on Neural Information Processing Systems*. 892–900.

[4] Shizhe Chen, Jia Chen, Qin Jin, and Alexander Hauptmann. 2018. Class-aware Self-Attention for Audio Event Recognition. In *ACM International Conference on Multimedia Retrieval*. 28–36.

[5] Szu-Yu Chou, Jyh-Shing Jang, and Yi-Hsuan Yang. 2018. Learning to Recognize Transient Sound Events using Attentional Supervision. In *International Joint Conference on Artificial Intelligence*. 3336–3342.

[6] Jinxi Guo, Ning Xu, Li-Jia Li, and Abeer Alwan. 2017. Attention Based CLDNNs for Short-Duration Acoustic Scene Classification. In *Interspeech*. 469–473.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition*. 770–778.

[8] Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm Slaney, Ron Weiss, and Kevin Wilson. 2017. CNN Architectures for Large-Scale Audio Classification. In *International Conference on Acoustics, Speech and Signal Processing*. 131–135.

[9] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, and K. Wilson. 2017. CNN Architectures for Large-scale Audio Classification. In *IEEE International Conference on Acoustics, Speech and Signal Processing*. 131–135.

[10] Maximilian Ilse, Jakub Tomczak, and Max Welling. 2018. Attention-based Deep Multiple Instance Learning. In *International Conference on Machine Learning*. 2127–2136.

[11] Sergey Ioffe and Christian Szegedy. 2015. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *International Conference on Machine Learning*. 448–456.

[12] Q. Kong, Y. Xu, W. Wang, and M. D. Plumbley. 2018. Audio Set Classification with Attention Model: A Probabilistic Perspective. In *IEEE International Conference on Acoustics, Speech and Signal Processing*. 316–320.

[13] Qiuqiang Kong, Changsong Yu, Turab Iqbal, Yong Xu, Wenwu Wang, and Mark D. Plumbley. 2019. Weakly labelled AudioSet Classification with Attention Neural Networks. *CoRR* (2019). https://arxiv.org/abs/1903.00765

[14] Anders Krogh and Jesper Vedelsby. 1994. Neural Network Ensembles, Cross Validation and Active Learning. In *International Conference on Neural Information Processing Systems*. 231–238.

[15] A. Kumar, M. Khadkevich, and C. Fügen. 2018. Knowledge Transfer from Weakly Labeled Audio Using Convolutional Neural Network for Sound Events and Scenes. In *IEEE International Conference on Acoustics, Speech and Signal Processing*. 326–330.

[16] Donmoon Lee, Subin Lee, Yoonchang Han, and Kyogu Lee. 2017. *Ensemble of Convolutional Neural Networks for Weakly-Supervised Sound Event Detection Using Multiple Scale Input.* Technical Report. DCASE2017 Challenge.

[17] J. Lee and J. Nam. 2017. Multi-Level and Multi-Scale Feature Aggregation Using Pretrained Convolutional Neural Networks for Music Auto-Tagging. *IEEE Signal Processing Letters* (2017), 1208–1212.

[18] Jongpil Lee, Jiyoung Park, and Juhan Nam. 2017. *Combining Multi-Scale Features Using Sample-Level Deep Convolutional Neural Networks for Weakly Supervised Sound Event Detection.* Technical Report. DCASE2017 Challenge.

[19] Xuelong Li, Bin Zhao, and Xiaoqiang Lu. 2017. MAM-RNN: Multi-level Attention Model Based RNN for Video Captioning. In *International Joint Conference on Artificial Intelligence*. 2208–2214.

[20] T. Y. Lin, A. RoyChowdhury, and S. Maji. 2015. Bilinear CNN Models for Fine-Grained Visual Recognition. In *IEEE International Conference on Computer Vision*. 1449–1457.

[21] Jen-Yu Liu and Yi-Hsuan Yang. 2016. Event Localization in Music Auto-tagging. In *ACM International Conference on Multimedia*. 1048–1057.

[22] Xianglai Meng, Biao Leng, and Guanglu Song. 2017. A Multi-level Weighted Representation for Person Re-identification. In *Artificial Neural Networks and Machine Learning*, Alessandra Lintas, Stefano Rovetta, Paul F.M.J. Verschure, and Alessandro E.P. Villa (Eds.). 80–88.

[23] Annamaria Mesaros, Toni Heittola, Aleksandr Diment, Benjamin Elizalde, Ankit Shah, Emmanuel Vincent, Bhiksha Raj, and Tuomas Virtanen. 2017. DCASE 2017 Challenge Setup: Tasks, Datasets and Baseline System. In *DCASE Workshop*.

[24] Andrew Owens and Alexei A. Efros. 2018. Audio-Visual Scene Analysis with Self-Supervised Multisensory Features. In *ECCV*. 639–658.

[25] E. Park, X. Han, T. L. Berg, and A. C. Berg. 2016. Combining Multiple Sources of Knowledge in Deep CNNs for Action Recognition. In *IEEE Winter Conference on Applications of Computer Vision*. 1–8.

[26] Justin Salamon, Brian McFee, and Peter Li. 2017. *DCASE 2017 Submission: Multiple Instance Learning for Sound Event Detection.* Technical Report. DCASE2017 Challenge.

[27] Rajiv Shah and Roger Zimmermann. 2017. *Multimodal Analysis of User-generated Multimedia Content.* Springer.

[28] Sijie Song, Cuiling Lan, Junliang Xing, Wenjun Zeng, and Jiaying Liu. 2017. An End-to-End Spatio-Temporal Attention Model for Human Action Recognition from Skeleton Data. In *AAAI Conference on Artificial Intelligence*. 4263–4270.

[29] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing Data using t-SNE. *Journal of Machine Learning Research* (2008), 2579–2605.

[30] Toan Vu, An Dang, and Jia-Ching Wang. 2017. *Deep Learning for DCASE2017 Challenge.* Technical Report. DCASE2017 Challenge.

[31] Yun Wang, Juncheng Li, and Florian Metze. 2018. A Comparison of Five Multiple Instance Learning Pooling Functions for Sound Event Detection with Weak Labeling. *CoRR* (2018). http://arxiv.org/abs/1810.09050

[32] Yong Xu, Qiuqiang Kong, Qiang Huang, Wenwu Wang, and Mark D. Plumbley. 2017. Attention and Localization Based on a Deep Convolutional Recurrent Model for Weakly Supervised Audio Tagging. In *Interspeech*. 3083–3087.

[33] Y. Xu, Q. Kong, W. Wang, and M. D. Plumbley. 2018. Large-Scale Weakly Supervised Audio Classification Using Gated Convolutional Neural Network. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing*. 121–125.

[34] Yifang Yin, Zhenguang Liu, Satyam, and Roger Zimmermann. 2016. Laplacian Sparse Coding of Scenes for Video Classification. In *IEEE International Symposium on Multimedia*. 499–506.

[35] Yifang Yin, Zhenguang Liu, and Roger Zimmermann. 2017. Geographic Information Use in Weakly-supervised Deep Learning for Landmark Recognition. In *IEEE International Conference on Multimedia and Expo*. 1015–1020.

[36] Yifang Yin, Rajiv Ratn Shah, and Roger Zimmermann. 2018. Learning and Fusing Multimodal Deep Features for Acoustic Scene Categorization. In *ACM International Conference on Multimedia*. 1892–1900.

[37] Changsong Yu, Karim Said Barsim, Qiuqiang Kong, and Bin Yang. 2018. Multi-level Attention Model for Weakly Supervised Audio Classification. In *Detection and Classification of Acoustic Scenes and Events 2018 Workshop*. 188–192.