MULTI-LEVEL ATTENTION MODEL FOR WEAKLY SUPERVISED AUDIO CLASSIFICATION

Changsong Yu[1], *Karim Said Barsim*[1], *Qiuqiang Kong*[2], *Bin Yang*[1]

[1] Institute of Signal Processing and System Theory, University of Stuttgart, Germany
[2] Center for Vision, Speech and Signal Processing, University of Surrey, UK

ABSTRACT

In this paper, we propose a multi-level attention model for the weakly labelled audio classification problem. The objective of audio classification is to predict the presence or the absence of sound events in an audio clip. Recently, Google published a large scale weakly labelled AudioSet dataset containing 2 million audio clips with only the presence or the absence labels of the sound events, without the onset and offset time of the sound events. Previously proposed attention models only applied a single attention module on the last layer of a neural network which limited the capacity of the attention model. In this paper, we propose a multi-level attention model which consists of multiple attention modules applied on the intermediate neural network layers. The outputs of these attention modules are concatenated to a vector followed by a fully connected layer to obtain the final prediction of each class. Experiments show that the proposed multi-attention attention model achieves a state-of-the-art mean average precision (mAP) of 0.360, outperforming the single attention model and the Google baseline system of 0.327 and 0.314, respectively.

Index Terms— AudioSet, audio classification, attention model

1. INTRODUCTION

Audio classification aims to predict the presence or the absence of audio events in an audio clip. Audio classification has many applications such as multimedia information retrieval and public surveillance [1, 2]. Before 2017, datasets in audio processing are relatively smaller than datasets in computer vision such as ImageNet [3]. For example, UrbanSound dataset [4] consists of 27 hours of urban sound records with 3075 samples. ESC-50 dataset [5] consists of 2000 environmental recordings across 50 classes. The detection and classification of acoustic scenes and events (DCASE) challenge 2013, 2016, 2017 [1, 2, 6] datasets consists of several hours of data. Recently, Google published a large scale audio classification dataset called AudioSet [7] consists of 5,800 hours two million human-labeled 10-second audio clips covering 527 audio categories.

In AudioSet, each audio clip contains one or several labels, such as "cat", "speech" and "park" [7]. AudioSet is a *weakly labelled dataset* (WLD), that is, only the presence or the absence of sound events are known in an audio clip, without knowing the onset and offset time of the sound events. The duration of sound events in the WLD vary from milliseconds to seconds depending on the categories. For example, sound class such as "speech" can last a few seconds, while sound class such as "gunshot" only last for hundreds of milliseconds.

The audio classification problem with WLD is to design a system trained only on WLD. Many methods such as multiple instance learning (MIL) [8] has been used to solve the WLD audio classification [9] problem. In [10] a single-level attention model was proposed and outperformed both the MIL method [9] and the Google baseline deep neural network system [7] on AudioSet classification. This single-level attention model consists of three fully connected layers followed by an attention module. The motivation of the attention module is based on the observation that different segments in an audio clip contribute differently to the label of an audio clip. For example, the segments containing a sound event should be attended and the segments containing irrelevant noise should be ignored.

However, when using the single-level attention model, substantial information from the intermediate neural network layers is disregarded. Previous work [11, 12, 13] explored the features from intermediate layers of a neural network contain rich information for classification. For example, Lee et. al. [11] explored that the audio classification performance can be improved by concatenating features from different intermediate neural network layers. Features from multiple intermediate layers have been found to be effective not only for audio tasks, but also for computer vision tasks. For example, Meng et al. [12] proposed to extract features from different layers of a deep CNN and concatenated them to a representation which significantly outperforms the non-concatenated features [12].

Inspired by the success of multi-level representation [11, 12], we expand the single-level attention model [10] to a *multi-level attention model*. Multiple attention modules are applied on the intermediate neural network layers. Then, the outputs of the attention modules are concatenated to a vector followed by a fully connected layer with sigmoid non-linearity to predict the presence probability of each class.

The paper is organized as follows. Section II introduces related works. Section III introduces the single-level attention model [10]. Section IV describes the proposed multi-level attention module. Section V shows the experimental results. Section VI concludes and forecasts the future work.

2. RELATED WORKS

Audio classification: Audio classification has attracted many attention in recent years. Some representative challenges including DCASE 2013 [6], DCASE 2016 [2] and DCASE 2017 [1]. Hidden Markov models have been used to model audio events in [14]. Non negative matrix based methods were applied to learn the dictionary of audio events [15]. Recently, neural network based methods including fully connected neural networks [16], convolutional neural networks (CNN) [17] have been applied on audio classification and achieved the state-of-the-art performance.

Attention module: The concept of attention module is first introduced in natural language processing [18]. Attention module allows

deep neural networks to focus on relevant instances and ignore irrelevant instances in a bag. It has been successfully applied in machine translation [18], face detection [19], image classification [20] and captioning [21]. It is also utilized in the domain audio classification [22].

3. DATASET

AudioSet [7] consists of over two million samples. There are 527 classes in the current version. AudioSet is a multi-label dataset and each audio clip has one or several labels. Google created AudioSet through transfer learning. In the pre-training stage, two billion 10-second audio clips from YouTube covering more than 30,000 classes are collected and called YouTube 100M [23]. Log Mel spectrogram with size of 96×64 along time and frequency axis is extracted as feature for each audio clip. Then, a ResNet-50 model is trained using this YouTube 100M data. This trained ResNet-50 is later used as a feature extractor. After the pre-training stage, two million 10-second audio clips covering 527 classes are collected. The log Mel spectrogram of each audio clip is presented to the trained ResNet-50 model to extract the bottleneck features. In this process, each audio clip is compressed into 10 bottleneck features. Each feature has a dimension of 128. These two million samples constitute AudioSet.

4. SINGLE-LEVEL ATTENTION MODEL

In this section, we will introduce the single-level attention model proposed in [10].

To illustrate the notation, let x_t, $t = 1, 2, ..., T$ be the t-th bottleneck feature with a dimension $M = 128$. Each sample in AudioSet has $T = 10$ bottleneck features. $K = 527$ is the number of classes.

In the single-level attention model, each bottleneck feature x_t is presented to a trainable embedding mapping $f_{emb}(\cdot)$ to extract an embedded feature h_t:

$$h_t = f_{emb}(x_t) \tag{1}$$

Furthermore, an attention module is applied on the T embedded features to attain the class probabilities for the input sample:

$$y(\mathbf{h}) = \frac{1}{\sum_{t=1}^{T} v(h_t)} \sum_{t=1}^{T} v(h_t) f(h_t) \tag{2}$$

where $\mathbf{h} = [h_1, ..., h_T]$ is the concatenation of the embedded features. Non-negative function $v(\cdot)$ determines how much an embedded feature h_t should be attended or ignored and $f(\cdot)$ denotes the classification output on an embedded feature h_t. The attention module has ability to ignore irrelevant sound segments such as background noise and silences, and attend to the sound segments with audio events.

The implementation of the single-level attention model is shown in Fig. 1. The first part is an embedded mapping $f_{emb}(\cdot)$ modeled by three fully connected neural layers with H units. The second part is an attention module described by Equation (2). The attention non-negative mapping $v_k(\cdot)$ and the classification mapping $f_k(\cdot)$ are modeled by a softmax function and sigmoid function, respectively. The normalization applied after $v_k(\cdot)$ ensures the attention is normalized. Finally, the prediction is obtained by element-wise multiplication of the classification output and normalized attention output.

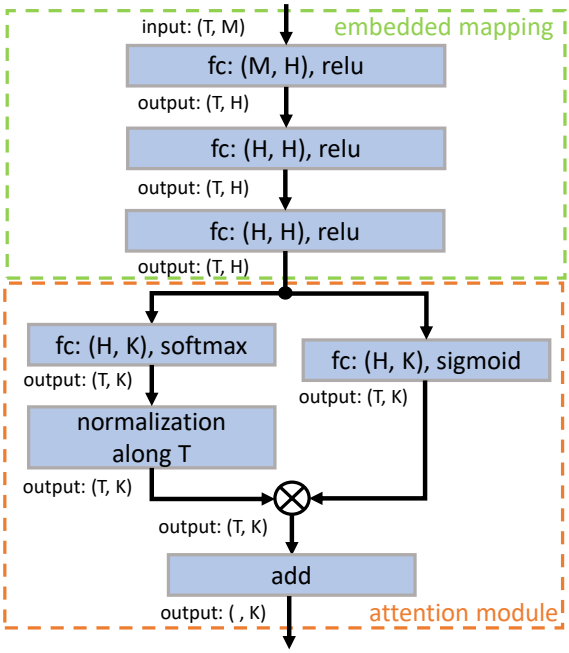

Figure 1: Architecture of the single-level attention model [10]

5. MULTI-LEVEL ATTENTION MODEL

Many works have explored that using multi-level features from intermediate layers of neural networks can promote the audio or image classification performance [11, 12]. We propose to extend the single-level attention model in Section 4 to multi-level attention model in our paper.

The architecture of the proposed multi-level attention model is shown in Fig. 2. Instead of applying a single-level attention model after the fully connected neural network, multiple attention modules are applied after intermediate layers as well. These attention modules aim to capture different level information. We denote the feedforward mappings as $g_l(\cdot)$ and the activations of the intermediate layers as $h^{(l)}$, where l is the number of embedded mappings. The feed-forward neural network can be written as:

$$\begin{cases} h_t^{(1)} = g_1(x_t) \\ h_t^{(l)} = g_l(h_t^{(l-1)}) \quad l = 2, 3, ..., L \end{cases} \tag{3}$$

where each forward mapping $g_l(\cdot)$ may consists of several fully connected layers in series (Fig. 2). For the single-level attention model, the prediction is produced by $y^{(L)} = y(\mathbf{h}^L)$ follows Equation (2) where $\mathbf{h}^{(L)} = \left[h_1^{(L)}, ..., h_T^{(L)}\right]$.

In the proposed multi-level attention model, each l-th attention module produces a prediction $y^{(l)} = y(\mathbf{h}^{(l)})$. Each prediction $y^{(l)} \in [0, 1]^K$. Then, all the predictions are concatenated to a vector $u \in [0, 1]^{KL}$:

$$u = \left[y^{(1)}, ..., y^{(L)}\right] \tag{4}$$

Finally, a fully connected layer followed by sigmoid nonlinearity is applied on the concatenated vector u to attain the class probabilities $z \in [0, 1]^K$ of the audio classes.

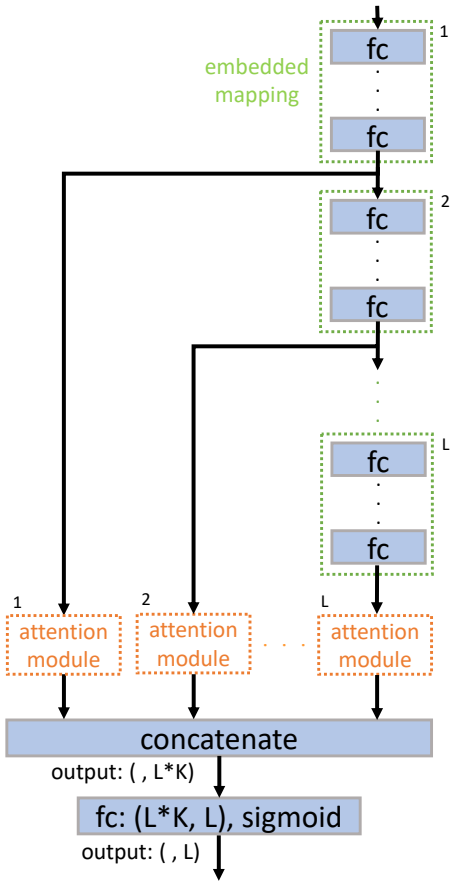

Figure 2: Architecture of the multi-level attention model

$$z = \phi(Wu + b) \tag{5}$$

where the $W \in R^{KL \times K}$ and $b \in R^K$ represent the weight matrix and the bias, separately. Sigmoid non-linearity $\phi(\cdot)$ is used for multi-label classification.

6. EXPERIMENTS

6.1. Training details

We use the balanced together with the unbalanced data from AudioSet [7] for training. We validated our model on the left out evaluation data of AudioSet. In order to comprehensively compare the performance of single-level and multi-level attention models, we implemented nine variants of single- (3-A, 6-A, 9-A) or multi-level attention models (1-A-1-A-1-A, 2-A-1-A, 2-A-2-A-2-A, 3-A-3-A, 3-A-3-A-3-A, 5-A-4-A) which are shown in Table I. The model 2-A-1-A represents two attention modules are applied after the 2nd and 3rd fully connected layers. The model 2-A-2-A-2-A represents three attention modules are applied after the 2nd, 4th and 6th fully connected layers. Each fully connected layer in all embedded mappings consists of 600 hidden units followed by ReLU activation function [24]. Dropout is used to prevent overfitting [25] with dropout rate of 0.4. Batch normalization [26] is applied to speed up training and prevent overfitting. We used Keras version 2.0.8 to implement our system. Adam optimizer [27] with learning rate of 0.001 is used. Batch size is set to 500. The setting of these hyper-parameters follows the configuration in [10]. Code has been made publicly available here [1]

6.2. Evaluation Metrics

To evaluate our model, we use three metrics of the Google's benchmark: mean average precision (mAP), area under curve (AUC) and d-prime. The mAP is the mean of average precision over all classes. The mAP is calculated by:

$$mAP = \frac{1}{K} \sum_{c=1}^{K} \sum_{n=1}^{N} p_{c,n} \Delta r_{c,n}, \tag{6}$$

where $p_{c,n}$ is the precision at n-th positive sample of c-th class. N is the number of positive samples for each class. $\Delta r_{c,n}$ is equal to $\frac{1}{N}$.

The AUC is area under the true positive-false positive rate curve. True positive rate (TPR) is a probability of correctly classifying a positive sample. False negative rate (FPR) is a probability of incorrectly classifying a negative sample as positive.

The d-prime is a deterministic function of AUC used in [7]. The d-prime can be calculated from AUC:

$$d\text{-}prime = \sqrt{2} F_x^{-1}(AUC) \tag{7}$$

F_x^{-1} is inverse of the cumulative distribution function and defined by:

$$F_x(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2}} dx \tag{8}$$

The larger AUC and d-prime indicates the better the audio classification performance.

6.3. Analysis

The first two rows in Table I show the results of Google's benchmark [7] without attention model and Kong's result with the single-level attention model [10]. All of the multi-level attention models outperform Google's baseline and single-level attention model in mAP, AUC, and d-prime. The best multi-level attention model is 2-A-1-A with two attention modules on the 2nd and 3rd intermediate layers. A mAP of 0.360 is achieved, outperforming the single-level attention model [10] of 0.327 and the Google's baseline system of 0.314 [7]. The reason for the good performance using multi-level attention model is that the multi-level features extracted from the intermediate layers provide various representations, and then each attention module can filter the unrelated information of each feature. In addition, different classes may favor different layer of features and the last fully connected layer of each multi-level attention model can automatically select best feature for each class by the weight parameters.

When comparing all variants of the single-level attention model (3-A, 6-A, 9-A), it was observed that the performance notably degrades as the number of fully connected layers is increased. This results from that the features extracted from a deep fully connected layer (e.g. 6th and 9th fully connected layer) are worse than that of a shallow layer (e.g. 3rd fully connected layer).

[1] https://github.com/ChangsongYu/Eusipco2018_Google_AudioSet

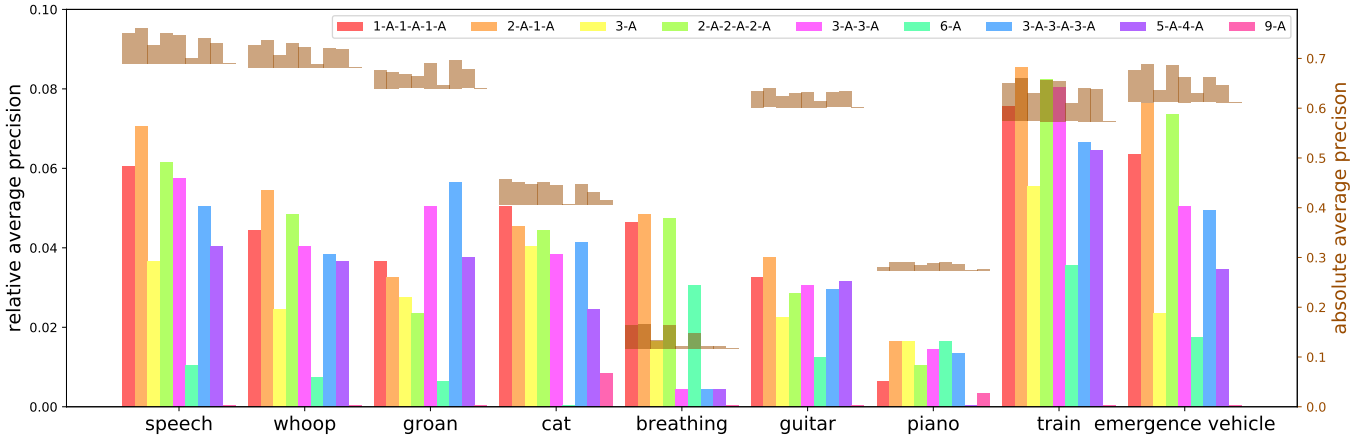

Figure 3: Average precision (AP) results of all single-level or multi-level attention models for nine randomly selected classes. The left black bar-graph scaled by the y-axis on the left-side represents the relative AP to the lowest AP among all models on a class. For example, the lowest AP among all models of the class "speech" is the AP of 9-A. The relative AP of 9-A of this class is 0 and that of 5-A-4-A is 0.04. The right brown bar-graph scaled by the y-axis on the right side represents the absolute AP. For example, the APs of 5-A-4-A and 9-A for the class "speech" are 0.730 and 0.690, separately.

Table 1: Comparisons of results of multi-level attention model

Model	**mAP**	**AUC**	**d-prime**
Benchmark	0.314	0.9590	2.452
Kong [10]	0.327	0.9650	2.558
1-A-1-A-1-A	0.357	0.9693	2.645
2-A-1-A	**0.360**	**0.9700**	**2.660**
3-A	0.336	0.9668	2.596
2-A-2-A-2-A	0.358	0.9695	2.650
3-A-3-A	0.355	0.9690	2.639
6-A	0.311	0.9571	2.430
3-A-3-A-3-A	0.353	0.9687	2.633
5-A-4-A	0.340	0.9676	2.612
9-A	0.305	0.9388	2.185

6.4. Performance visualization of individual classes

In addition, we investigate all variants of our single-level or multi-level attention model by comparing average precision (AP) of nine randomly selected classes are shown in Figure 3. For each class, the color bars plotted below is the relative improvement of AP and the bars plotted above is the absolute AP. The APs of classes such as speech and whoop are close to 0.7. In contrast, APs of many classes such as breathing are lower than 0.2.

Figure 3 shows that the multi-level attention models do not always achieve better performance on all classes than the single-level attention models. For the class "piano", the model 6-A outperforms the models 2-A-2-A-2-A and 3-A-3-A. We also observe that different classes favor different models. For example, the classes “speech”, “whoop”, “breathing”, “guitar”, “train” and “emergence vehicle” favor the model 2-A-1-A. However, the class "groan" favors the model 3-A-3-A-3-A. Overall, we can ensure that the performance of classification consistently increases on most classes when the multi-level features are concatenated and 2-A-1-A is the best architecture.

7. CONCLUSION

In this work, we introduced a multi-level attention model in addressing weakly labelled audio classification problem on AudioSet. The experimental results showed the effectiveness of multi-level attention models and achieved a state-of-the-art mean average precision (mAP) of 0.360 than the single-attention model and Google’s baseline system. In future, we will investigate the combination of the multi-scale and multi-level features for AudioSet classification.

8. ACKNOWLEDGEMENT

Qiuqiang Kong was supported by EPSRC grant EP/N014111/1 “Making Sense of Sounds” and a Research Scholarship from the China Scholarship Council (CSC) No. 201406150082.

9. REFERENCES

[1] A. Mesaros, T. Heittola, A. Diment, B. Elizalde, A. Shah, E. Vincent, B. Raj, and T. Virtanen, “DCASE 2017 challenge setup: Tasks, datasets and baseline system,” in *DCASE 2017-Workshop on Detection and Classification of Acoustic Scenes and Events*, 2017.

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

[3] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 248–255.

[4] J. Salamon, C. Jacoby, and J. P. Bello, “A dataset and taxonomy for urban sound research,” in *Proceedings of the 22nd*

ACM international conference on Multimedia. ACM, 2014, pp. 1041–1044.

[5] K. J. Piczak, "ESC: Dataset for environmental sound classification," in *Proceedings of the 23rd ACM international conference on Multimedia*. ACM, 2015, pp. 1015–1018.

[6] 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.

[7] 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 *IEEE ICASSP*, 2017.

[8] O. Maron and T. Lozano-Pérez, "A framework for multiple-instance learning," in *Advances in neural information processing systems*, 1998, pp. 570–576.

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

[10] Q. Kong, Y. Xu, W. Wang, and M. D. Plumbley, "Audio set classification with attention model: A probabilistic perspective," *arXiv preprint arXiv:1711.00927*, 2017.

[11] J. Lee and J. Nam, "Multi-level and multi-scale feature aggregation using pretrained convolutional neural networks for music auto-tagging," *IEEE signal processing letters*, vol. 24, no. 8, pp. 1208–1212, 2017.

[12] X. Meng, B. Leng, and G. Song, "A multi-level weighted representation for person re-identification," in *International Conference on Artificial Neural Networks*. Springer, 2017, pp. 80–88.

[13] H. R. Roth, L. Lu, A. Farag, H.-C. Shin, J. Liu, E. B. Turkbey, and R. M. Summers, "Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2015, pp. 556–564.

[14] Y.-T. Peng, C.-Y. Lin, M.-T. Sun, and K.-C. Tsai, "Healthcare audio event classification using hidden markov models and hierarchical hidden markov models," in *Multimedia and Expo, 2009. ICME 2009. IEEE International Conference on*. IEEE, 2009, pp. 1218–1221.

[15] V. Bisot, R. Serizel, S. Essid, and G. Richard, "Supervised nonnegative matrix factorization for acoustic scene classification," *IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE)*, 2016.

[16] Q. Kong, I. Sobieraj, W. Wang, and M. Plumbley, "Deep neural network baseline for dcase challenge 2016," *Proceedings of DCASE 2016*, 2016.

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

[18] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[19] S. Sharma, R. Kiros, and R. Salakhutdinov, "Action recognition using visual attention," *arXiv preprint arXiv:1511.04119*, 2015.

[20] K. J. Shih, S. Singh, and D. Hoiem, "Where to look: Focus regions for visual question answering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4613–4621.

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

[22] Y. Xu, Q. Kong, W. Wang, and M. D. Plumbley, "Large-scale weakly supervised audio classification using gated convolutional neural network," *arXiv preprint arXiv:1710.00343*, 2017.

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

[24] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th international conference on machine learning (ICML-10)*, 2010, pp. 807–814.

[25] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[26] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*, 2015, pp. 448–456.

[27] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.