INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Audio Tagging with Compact Feedforward Sequential Memory Network and Audio-to-Audio Ratio Based Data Augmentation

Zhiying Huang, Shiliang Zhang, Ming Lei

Alibaba Inc., P. R. China
{zhiying.hzy, sly.zsl, lm86501}@alibaba-inc.com

Abstract

Audio tagging aims to identify the presence or absence of audio events in the audio clip. Recently, a lot of researchers have paid attention to explore different model structures to improve the performance of audio tagging. Convolutional neural network (CNN) is the most popular choice among a wide variety of model structures, and it's successfully applied to audio events prediction task. However, the model complexity of CNN is relatively high, which is not efficient enough to ship in real product. In this paper, compact Feedforward Sequential Memory Network (cFSMN) is proposed for audio tagging task. Experimental results show that cFSMN-based system yields a comparable performance with the CNN-based system. Meanwhile, an audio-to-audio ratio (AAR) based data augmentation method is proposed to further improve the classifier performance. Finally, with raw waveforms of the balanced training set of *Audio Set* which is a published standard database, our system can achieve a state-of-the-art performance with AUC being 0.932. Moreover, cFSMN-based model has only 1.9 million parameters, which is only about 1/30 of the CNN-based model.

Index Terms: Audio Set, audio tagging, compact feedforward sequential memory network, audio-to-audio ratio, data augmentation

1. Introduction

Given an audio clip, it carries not only speech information but also much non-speech information, such as audio events [1]. For most humans, it's effortless to distinguish what events have happened in the audio clip. However, for a machine, such as a computer, it's a really challenging task. Automatically recognizing the audio events remains an open issue. Audio tagging aims to recognize the audio events for a given audio clip automatically, and without predicting the onset and offset time of these events. Moreover, audio tagging has a wide range of applications where audio events are involved, such as surveillance, information retrieval [2] and audio classification [3].

In recent years, audio tagging has received more and more attention from researchers. This task has been included by the detection and classification of acoustic scenes and events (DCASE) challenge every year since 2016 [4]. Generally, the label of audio tagging includes two types: weak label [5] and strong label [6]. For weak label, the events of current audio clip are known only. In strong label, it also includes the onset and offset time of these events, but it's more costly to be obtained than weak label. As a result, weak label is mostly used. Some researchers devote to make use of the attention model to tackle the weak label audio tagging problem [7–9]. Recently, sequential label which is a compromise between the two label types is proposed to combine with connectionist temporal classification (CTC) loss function for audio tagging [10, 11] or audio event detection [12, 13].

In traditional systems for audio tagging, Gaussian Mixture Model (GMM) is used to predict whether the audio event occurs in the current audio clip [14], and it's also used as the official baseline method in DCASE 2016 for audio tagging [4]. Besides, Support Vector Machine (SVM) based system for audio tagging is presented in [5]. However, neither GMM nor SVM can well model context dependent information and the potential relationship among different audio events. Recently, with the progress of deep learning in the past few years, numerous neural networks (NN) based approaches have been proposed to audio tagging. In [15], a deep neural network (DNN) is applied to audio tagging and shows superior results than GMM. Then, Recurrent Neural Network (RNN) has shown its advantage in modeling long-term dependencies than DNN in sequential data due to the recurrent connections over time, and it's successfully used in audio tagging [16, 17]. Inspired by the successes of convolutional neural network (CNN) in image classification, CNN-based architectures become popular and turn out to be a great success in audio tagging [17–24]. Although CNN or RNN based models are extremely effective for audio tagging, they usually suffer from the problem of huge amount of model parameters.

The compact Feedforward Sequential Memory Network (cFSMN) is firstly proposed and used in the automatic speech recognition (ASR) [25]. Different to the basic fully-connected feedforward neural network (FNN) that maps a fixed input within a small context window to a fixed output, cFSMN is able to capture information in a very long context by using memory blocks with look-back and look-ahead filters in a hierarchical structure. Compared to RNN and CNN, cFSMN can achieve comparable performance while less in model parameters and more efficient in computation. Thereby, it is promoted to many other tasks, such as text to speech (TTS) [26] and smaller footprint keyword spotting (KWS) [27].

In this paper, we propose to use cFSMN in audio tagging. Firstly, we present a detailed investigation of cFSMN's configuration, which is essential to the performance. According to the experimental results, it's found that cFSMN-based audio tagging can yield a comparable performance with CNN-based. Moreover, an audio-to-audio ratio (AAR) based data augmentation method is proposed to improve the performance. As a result, with raw waveforms of the balanced training set of *Audio Set* which is a published standard database, our system can achieve a state-of-the-art performance with AUC being 0.932 in the evaluation set. Moreover, cFSMN-based method has only 1.9 million parameters, which is only about 1/30 of the baseline CNN-based method.

The rest of the paper is organized as follows. Section 2 describes the cFSMN and cFSMN-based audio tagging. Section 3 introduces the AAR-based data augmentation method. The experimental results are displayed in section 4. Section 5 finally draws a conclusion and describes our future work.

http://dx.doi.org/10.21437/Interspeech.2019-1302

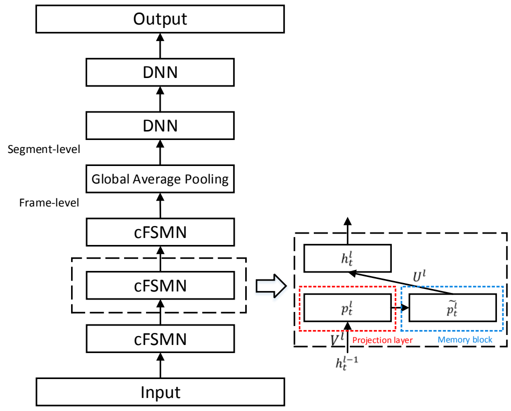

Figure 1: *System of cFSMN-based audio tagging.*

2. cFSMN-based audio tagging

FSMN [29] is a standard FNN with single or multiple memory blocks in the hidden layers, and the memory block is designed to encode outputs of previous hidden layer and previous histories of current layer into a fixed-sized representation. With the memory block, FSMN can learn long-term dependency without using recurrent feedback. However, FSMN may introduce many additional model parameters compared to FNN with the same architecture due to the memory block. For the purpose of reducing the number of model parameters of FSMN, cFSMN [25] which combines FSMN with low-rank matrix factorization and weight sharing is proposed to simplify the FSMN architecture and speed up the learning. cFSMN can significantly outperform the FSMN while being simpler in model structure and faster in training speed.

The illustration of cFSMN is shown in the black dotted box of Figure 1. As we can see, a linear projection layer is firstly applied to the outputs of previous hidden layer , and this projection operation reduces the parameters number. Then, the element-wise weighted sum of the projection outputs and previous histories are obtained after the memory block. Finally, the affine and nonlinear transforms are applied to the weighted sum, and the outputs of current hidden layer are generated. The corresponding equations are shown in Equation (1) (2) (3).

$$p_t^l = V^l h_t^{l-1} + b^l \tag{1}$$

$$\tilde{p}_t^l = p_t^l + \sum_{i=0}^{N_1} a_i^l \odot p_{t-i}^l + \sum_{j=1}^{N_2} c_j^l \odot p_{t+j}^l \tag{2}$$

$$h_t^l = f(U^l \tilde{p}_t^l + b^l) \tag{3}$$

p_t^l denotes the outputs of projection layer at time t while V^l and b^l are the corresponding weights and bias, h_t^{l-1} and h_t^l denote the outputs of $l-1$-th and l-th hidden layer. In Equation (2), a_i^l and c_j^l are the look-back and look-ahead filters, and N_1 and N_2 are the orders respectively, $\tilde{p}_t^l$ is the outputs of memory block with context information.

The system of cFSMN-based audio tagging is also shown in Figure 1, and the inputs are the acoustic features of the audio clip, such as Mel-Frequency Cepstral Coefficients (mfccs), log Mel-Filter Banks (fbanks) and spectrogram. For the consistence with ASR system, fbanks are used here. The outputs are the probabilities of different audio events, and the number of audio events are pre-designed. Overall view of this system, it's a hybrid cFSMN-DNN architecture. It's considering that cFSMN has the advantage in modeling long-term dependencies while DNN is appropriate for mapping features to a more separable space [30], and it's found that combining the advantage of these two architectures benefits the final result. In traditional audio tagging system, there is a global average pooling layer [31] after the output layer to map the scores from frame-level to segment-level. Specially in this paper, the global average pooling layer is inserted between the cFSMN and DNN layer shown in Figure 1.

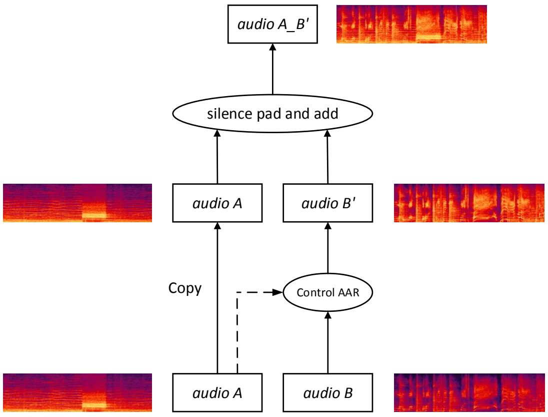

Figure 2: *AAR-based data augmentation. (*audio A *and* audio B *are from the balanced training set of* Audio Set*, and the wav-ids are EeEpHaw84ww and X1zFnyEe3nE respectively.)*

In training stage, the tags of each audio clip are organized as an N-hot ($N \geq 1$) label. Then, the acoustic features of audio clips and labels are used to train the cFSMN-based classifier. During inference stage, the acoustic features of each testing audio clip are fed to the well-trained classifier to predict the final scores of different audio events.

3. AAR-based data augmentation

In this section, AAR-based data augmentation is described in detail. In the ASR, there are many straightforward data augmentation methods, such as amplitude and speed adjustment. We try these two methods and find that neither of them is worked for audio tagging. From a different perspective, there are a certain number of audio events existing in an audio clip. The permutations of different audio events are countless from a mathematical point of view. So, there must be many permutations that are not included in training set, and the way to increase the permutations may benefit the generalization ability and final performance of the classifier.

In order to increase the permutations of different audio events, a straightforward method is to create a new audio clip by directly adding two audio clips in time domain, and the two audio clips which form an audio clip pair are randomly selected in the training set. The label of generated new audio clip is the union set of its parents' labels. In this way, a huge amount of new audio clips are created and added to the original training set. However, it's found that the final performance is degraded with these new additional audio clips. After analysis, the two audio clips to be added in time domain have large difference in the amplitude of power. Directly adding these two audio clips in time domain may lead to the information covering problem, which means the information of low power audio clip is covered up by the information of high power audio clip. As a result, AAR is proposed to control the power ratio to solve the

Table 1: *Comparison of AUCs of different configurations of cFSMN.*

Model structure	Feature	Position	AUC
5cFSMN	plain	output	0.886
3cFSMN + 2DNN	plain	output	0.893
3cFSMN + 2DNN	delta-splice	output	0.911
3cFSMN + 2DNN	delta-splice	hidden	**0.922**

information covering problem before the adding operation. The definition of AAR is the ratio of one audio clip power to another audio clip power.

The schematic of AAR-based data augmentation is as shown in Figure 2. First of all, two audio clips are randomly selected from the training set, and they are denoted as *audio A* and *audio B*. Then, based on the power of *audio A*, the power of *audio B* is tuned to match the preset AAR, and *audio B'* is the transformed audio clip. The operation of padding silence frame after the shorter audio clip is done to make *audio A* and *audio B'* consistent in duration. Finally, *audio A_B'* is generated by adding them in time domain. The label of generated new audio clip is the union set of the labels of *audio A* and *audio B*.

Recently, mixed-based data augmentation for audio tagging is proposed [32, 33]. There are two main difference between mixed-based and AAR-based data augmentation. Firstly, mix-based method relies on a hyper-parameter λ while AAR-based method is λ-free. Secondly, the AAR is proposed and plays the vital role in AAR-based method.

4. Experiments

4.1. Experimental setup

4.1.1. Audio Set

The experiments are done in *Audio Set*. *Audio Set* is a large scale dataset of manually-annotated audio events from YouTube, and it's published by the Sound and Video Understanding team at Google Research. Most of the audio clips in *Audio Set* are 10 seconds in waveform duration, except when that exceeds the length of the underlying video. It contains 527 audio events, and there may be multiple audio events co-existing in an audio clip. The entire *Audio Set* contains about 2 million audio clips, which correspond to more than 5000 hours of data. The audio clips are divided into three subsets: balanced training set, unbalanced training set and evaluation set.

In this study, only the raw waveforms of balanced training set are used for training because it's hard to fetch the raw waveforms of unbalanced training set in YouTube. Of course, we evaluate the performance in evaluation set for fair comparison. During fetching the raw waveforms of balanced training set and evaluation set in YouTube, several audio clips are excluded for the deleted YouTube links. In general, the number of audio clips used for training and testing are 20051 and 18636 respectively, corresponding to 55 hours and 52 hours of data.

4.1.2. Data preprocessing

Before training or testing stage, each audio clip is divided into non-overlapping 960 millisecond (ms) segments, and each segment inherits all the audio events of its parent audio clip. The input audio clip sampled at 16 kHz is analyzed using a 25 ms Hamming window with a fixed 10 ms frame shift. The 80-dimensional fbanks are computed, which are distributed on a mel-scale. Meanwhile, global means and variances are conducted for each dimension of the fbanks.

Table 2: *Comparison of AUCs of different amount of training data with DRE-based method (*h *means hours).*

Training Data	Time (h)	AUC
bal-tr	55 + 0	0.922
bal-tr + DRE-based×1	55 + 55	0.916
bal-tr + DRE-based×2	55 + 110	0.918
bal-tr + DRE-based×5	55 + 275	0.918

Table 3: *Comparison of AUCs of augmentation data using different AARs with AAR-based method.*

Training Data	AAR (dB)	Time (h)	AUC
bal-tr + AAR-based×1	0	55 + 55	0.922
bal-tr + AAR-based×2		55 + 110	0.924
bal-tr + AAR-based×5		55 + 275	0.926
bal-tr + AAR-based×1	+5	55 + 55	0.924
bal-tr + AAR-based×2		55 + 110	0.923
bal-tr + AAR-based×5		55 + 275	0.925
bal-tr + AAR-based×1	0, +5	55 + 110	0.918

4.1.3. Training configuration

All experiments use Kaldi toolkit as the training platform [34], and stochastic gradient descent (SGD) is used as the optimizer. All models use a final sigmoid layer rather than softmax layer, and binary cross entropy is used as the loss function. For cFSMN, N_1 and N_2 are both set to 10 (see Equation (2)). Based on empirical experience, 5 hidden layers are enough for hybrid cFSMN-DNN to model the mapping from acoustic features to audio events, and each hidden layer owns 128 units for cFSMN or DNN. We directly feed the outputs of previous hidden layer to the memory block without projection layer, because the number that 128 units for each layer is small enough.

4.1.4. Evaluation metric

After feeding each 960 ms segment of one evaluation audio clip to the cFSMN-based model, the output scores of classifier are averaged across all segments in current audio clip. AUC is used as the metric, and the AUC is the area under the Receiver Operating Characteristic (ROC) curve. The ROC curve is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. The AUC of classifier with perfect performance is 1.0. If the classifier guesses the audio events randomly, the AUC is 0.5. In this paper, the final AUC of evaluation set is the average across the AUCs of all audio events.

4.2. cFSMN-based audio tagging

In this part, different configurations of cFSMN are compared, including 3 aspects: model structure, acoustic feature and pooling position. The experimental results are shown in Table 1. For the model structure, "5cFSMN" means model with 5 cFSMN layers, "3cFSMN + 2DNN" refers to that 3 cFSMN layers are placed as the bottom layer and 2 fully-connected DNN layers are stacked in the top layer. Results show that the hybrid construction "3cFSMN + 2DNN" (with the AUC of 0.893) shows better performance than "5cFSMN" (with the AUC of 0.886).

Table 4: *Comparison of AUCs of different thresholds for event number with AAR-based method (AAR=0dB).*

Training Data	Threshold	Time (h)	AUC
bal-tr + AAR-based×2	-	55 + 110	0.924
	9	55 + 98	0.925
	7	55 + 93	**0.926**
	5	55 + 74	0.924
	3	55 + 28	0.922

In the ASR, the temporal derivatives (delta) fbanks may reduce the final word error rate (WER), and that features are spliced in time taking a context size of N frames also benefits the recognition performance. Here, we used the first and second temporal derivatives ($\Delta + \Delta\Delta$) of fbanks, and consecutive frames within a context window of 11 (5-1-5, N is set to 5) are stacked to produce the 2640-dimension inputs. For simplification, "plain" denotes original features, and "delta-splice" means the features after applying delta and splice. After comparing the AUCs of the third and fourth row of Table 1, it's shown that the "delta-splice" (with the AUC of 0.911) are superior to "plain" (with the AUC of 0.893). Although the cFSMN has the ability of modeling long-term dependencies, the context information is still necessary for better performance.

In traditional NN-based audio tagging, a global average pooling layer is placed after the final sigmoid layer to map the scores from frame-level to segment-level. In this paper, the hybrid cFSMN-DNN is used as classifier, the position of average pooling layer between cFSMN and DNN may be helpful in the exertion of the advantage of DNN. In Table 1, "output" denotes the traditional pooling position, "hidden" means the pooling position of this paper. It's found that "hidden" (with the AUC of 0.922) is better than "output" (with the AUC of 0.911).

By integrating the previous experiments of cFSMN-based audio tagging, the best configuration is obtained. It will be used in the following experiments.

4.3. AAR-based data augmentation

In this part, we firstly do some experiments with the method that adding two audio clips directly without controlling the AAR, and this method is referred to "DRE-based". We use $\times N$ to indicate creating N times the data of the balanced training set (bal-tr). As is shown in Table 2, directly adding two audio clips does harm to the final AUC because of the information covering problem which is mentioned in Section 3.

Then, we explore the AAR-based data augmentation in detail. In Table 3, we compare the performance with augmentation data using different AARs. Augmentation data with AAR of 5 decibel (dB) are the mirror of -5 dB, so we only try AAR of 5 dB in this part. According to Table 3, it's found that AAR of 0 dB and 5 dB perform almost equally with the same amount of data, this is because the new audio clips contain consistent acoustic information for these two AARs. Moreover, data fusion is conducted for the training data of different AARs with AAR-based×1. However, results show that the data fusion leads to performance degradation, the AUC drops from 0.922 or 0.924 to 0.918. This may be because the mechanism, where the original audio clips pairs for generating a new audio clip are the same for different AARs, results in the ambiguity of classifier. The classifier fails to model the pattern that new audio clip with different AARs uses the same label. So, we choose to use AAR of 0 dB and data fusion is not used in the following experiment.

Table 5: *Comparison of AUCs of different amount of training data with AAR-based method (AAR=0dB).*

Training Data	Threshold	Time (h)	AUC
bal-tr + AAR-based×2	7	55 + 93	0.926
bal-tr + AAR-based×10		55 + 464	0.931
bal-tr + AAR-based×20		55 + 928	**0.932**
bal-tr + AAR-based×50		55 + 2320	0.932
bal-tr + AAR-based×100		55 + 4640	0.931

Table 6: *Comparison of AUCs of different systems trained with raw waveforms of balanced training set (*M *means million parameters).*

Results	Model	Model Size (M)	AUC
Shah [23]	WAL-Net	19.98	0.925
Kumar [24]	Deep CNN	19.98	0.925
Wu [35]	AlexNet	56.09	0.895
Wu [35]	AlexNet(BN)	56.11	0.927
This paper	cFSMN cFSMN + AAR	1.9	0.922 **0.932**

Generally, the audio clip with too many events is almost impossibly existed in real world. After analyzing the distribution of balanced training set, it's also found that there is no utterance that with the event number larger than 12. Therefore, applying threshold to event number is necessary to remove some meaningless augmentation data. Table 4 compares the AUCs of different thresholds for event number in AAR-based×2, and results show that the threshold of 7 shows best performance with the AUC of 0.926. Furthermore, with the threshold of 7, we try to add more augmentation data to obtain better performance. As is demonstrated in Table 5, AAR-based×20 is the best configuration with AUC of 0.932, which adds about extra 928 hours for training.

In Table 6, the AUCs of different systems trained with raw waveforms of balanced training set are illustrated. "AlexNet(BN)" means a batch normalization layer is added after each layer (convolutional and FC) of AlexNet. It's found that cFSMN without AAR-based data augmentation yields a comparable performance with AlexNet(BN). After data augmentation, the state-of-art performance is achieved with AUC of 0.932 in the evaluation set. More importantly, cFSMN-based model only requires 1.9 million parameters, which is only about 1/30 of the AlexNet(BN).

5. Conclusions

In this paper, cFSMN is applied to audio tagging task. After exploring the best configuration of cFSMN, we can achieve a comparable performance with a strong CNN baseline system, AlexNet(BN). Moreover, we also propose a novel AAR-based data augmentation method to further improve the performance of our system. As a result, with raw waveforms of the balanced training set of *Audio Set* which is a published standard database, our system can achieve a state-of-the-art performance with AUC being 0.932 in the evaluation set. More importantly, cFSMN-based system only has 1.9 million model parameters, which is only about 1/30 of the strong CNN baseline system.

In the future, we will explore in the relationship among different audio events, which may benefit the final performance of audio tagging.

6. References

[1] T. Virtanen, M. D. Plumbley, and D. Ellis, *Computational analysis of sound scenes and events*. Springer, 2018.

[2] E. Wold, T. Blum, D. Keislar, and J. Wheaten, "Content-based classification, search, and retrieval of audio," *IEEE multimedia*, vol. 3, no. 3, pp. 27–36, 1996.

[3] D. Giannoulis, E. Benetos, D. Stowell, M. Rossignol, M. Lagrange, and M. D. Plumbley, "Detection and classification of acoustic scenes and events: An ieee aasp challenge," in *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. IEEE, 2013, pp. 1–4.

[4] *DCASE2016 audio tagging task*, 2016. [Online]. Available: http://www.cs.tut.fi/sgn/arg/dcase2016/task-audio-tagging

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

[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] 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.

[8] C. Yu, K. S. Barsim, Q. Kong, and B. Yang, "Multi-level attention model for weakly supervised audio classification," *arXiv preprint arXiv:1803.02353*, 2018.

[9] Q. Kong, C. Yu, T. Iqbal, Y. Xu, W. Wang, and M. D. Plumbley, "Weakly labelled audioset classification with attention neural networks," *arXiv preprint arXiv:1903.00765*, 2019.

[10] Y. Wang and F. Metze, "A first attempt at polyphonic sound event detection using connectionist temporal classification," in *2017 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE, 2017, pp. 2986–2990.

[11] Y. Wang and F. Metze, "Connectionist temporal localization for sound event detection with sequential labeling," *arXiv preprint arXiv:1810.09052*, 2018.

[12] Y. Hou, Q. Kong, and S. Li, "Audio tagging with connectionist temporal classification model using sequential labelled data," *arXiv preprint arXiv:1808.01935*, 2018.

[13] Y. Hou, Q. Kong, J. Wang, and S. Li, "Polyphonic audio tagging with sequentially labelled data using crnn with learnable gated linear units," *arXiv preprint arXiv:1811.07072*, 2018.

[14] P. Foster, S. Sigtia, S. Krstulovic, J. Barker, and M. D. Plumbley, "Chime-home: A dataset for sound source recognition in a domestic environment," in *2015 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2015, pp. 1–5.

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

[16] T. H. Vu and J.-C. Wang, "Acoustic scene and event recognition using recurrent neural networks," *Detection and Classification of Acoustic Scenes and Events*, vol. 2016, 2016.

[17] Y. Xu, Q. Kong, Q. Huang, W. Wang, and M. D. Plumbley, "Convolutional gated recurrent neural network incorporating spatial features for audio tagging," in *2017 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2017, pp. 3461–3466.

[18] E. Cakır, T. Heittola, and T. Virtanen, "Domestic audio tagging with convolutional neural networks," *IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE 2016)*, 2016.

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

[20] 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 *2017 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE, 2017, pp. 131–135.

[21] 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.

[22] Q. Kong, Y. Xu, W. Wang, and M. D. Plumbley, "A joint separation-classification model for sound event detection of weakly labelled data," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 321–325.

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

[24] A. Kumar, M. Khadkevich, and C. Fügen, "Knowledge transfer from weakly labeled audio using convolutional neural network for sound events and scenes," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 326–330.

[25] S. Zhang, H. Jiang, S. Xiong, S. Wei, and L.-R. Dai, "Compact feedforward sequential memory networks for large vocabulary continuous speech recognition." in *Interspeech*, 2016, pp. 3389–3393.

[26] M. Bi, H. Lu, S. Zhang, M. Lei, and Z. Yan, "Deep feed-forward sequential memory networks for speech synthesis," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4794–4798.

[27] M. Chen, S. Zhang, M. Lei, Y. Liu, H. Yao, and J. Gao, "Compact feedforward sequential memory networks for small-footprint keyword spotting," *Proc. Interspeech 2018*, pp. 2663–2667, 2018.

[28] 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.

[29] S. Zhang, C. Liu, H. Jiang, S. Wei, L. Dai, and Y. Hu, "Feedforward sequential memory networks: A new structure to learn long-term dependency," *arXiv preprint arXiv:1512.08301*, 2015.

[30] T. N. Sainath, O. Vinyals, A. Senior, and H. Sak, "Convolutional, long short-term memory, fully connected deep neural networks," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 4580–4584.

[31] Y. Wang and F. Metze, "A comparison of five multiple instance learning pooling functions for sound event detection with weak labeling," *arXiv preprint arXiv:1810.09050*, 2018.

[32] K. Xu, D. Feng, H. Mi, B. Zhu, D. Wang, L. Zhang, H. Cai, and S. Liu, "Mixup-based acoustic scene classification using multi-channel convolutional neural network," in *Pacific Rim Conference on Multimedia*. Springer, 2018, pp. 14–23.

[33] S. Wei, K. Xu, D. Wang, F. Liao, H. Wang, and Q. Kong, "Sample mixed-based data augmentation for domestic audio tagging," *arXiv preprint arXiv:1808.03883*, 2018.

[34] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," IEEE Signal Processing Society, Tech. Rep., 2011.

[35] Y. Wu and T. Lee, "Reducing model complexity for dnn based large-scale audio classification," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 331–335.