IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 14, 2020, accepted January 28, 2020, date of publication February 17, 2020, date of current version March 3, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2974479

Research on Semi-Supervised Sound Event Detection Based on Mean Teacher Models Using ML-LoBCoD-NET

JINJIA WANG, (Member, IEEE), JING XIA, QIAN YANG, AND YUZHEN ZHANG
School of Information Science and Engineering, Yanshan University, Qinhuangdao 066004, China
Hebei Key Laboratory of Information Transmission and Signal Processing, Yanshan University, Qinhuangdao 066004, China

Corresponding author: Jinjia Wang (wjj@ysu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China Project under Grant 61473339, in part by the First Batch of Top Young Talents in Hebei Province [2013]17, and in part by the Basic Research Cooperation Projects of Beijing, Tianjin, and Hebei under Grant 19JCZDJC65600 and Grant F2019203583.

ABSTRACT One of the most commonly method for sound event detection is the traditional convolutional neural network (CNN) or convolutional recurrent neural network (CRNN) and their variants. However, the pooling operation of the CNN has the disadvantage of losing the location information of the target object. We don't use the pooling operation, retaining ReLU and convolution operation, and we use the dictionary strong constraints and penalty function prior constraints of the multi-layer convolutional sparse coding (ML-CSC). We proposed iterative deep neural networks, the unfolded multi-layer local block coordinate descent networks (ML-LoBCoD-NET), driven by the multi-layer local block coordinate descent algorithm (ML-LoBCoD) which is extended from the local block coordinate descent (LoBCoD) algorithm. The ML-LoBCoD-NET can extract features different from the CNN. More importantly, for weakly-supervised sound event detection task, we proposed the MRNN-Att network which combines the ML-LoBCoD-NET, a recurrent neural network (RNN), and an attention network. The MCRNN-Att network combines MRNN-Att and CRNN network for fusing the different features. Furthermore, for semi-supervised sound event detection task, the MRNN-Att mean teacher model (MRNN-Att-MT) and the MCRNN-Att mean teacher model (MCRNN-Att-MT) are proposed, in which the MRNN-Att and the MCRNN-Att network are selected as the student model. These models were tested on the dataset of Detection and Classification of Acoustic Scenes and Events (DCASE) 2018 Task 4. The F1 score of the MRNN-Att-MT on the development set was 22.83%, which was 8.77% higher than the baseline system. The score of the MRNN-Att-MT on the evaluation set was 15.68%, which was 4.88% higher than the baseline system. The MCRNN-Att-MT model had an F1 score of 20.35% on the development set, which was 6.29% higher than the baseline system and the F1 score of 14.56% on the evaluation set, which was 3.76% higher than the baseline system.

INDEX TERMS Sound event detection, weakly-supervised learning, semi-supervised learning, mean teacher model, multi-layer local block coordinate descent, convolutional recurrent neural network, attention network.

I. INTRODUCTION

People rely on sounds in the environment to obtain important information. Sound event detection (SED) can detect specific audio events from audio recordings, estimate the starting and offset locations of sound events, and provide a label for each event. SED has great potential in many applications, such as information retrieval, monitoring systems, and automatic control of devices in smart home systems [1].

The most typical method of SED is to use hidden markov models (HMMs) [2], support vector machine (SVM) [3], and non-negative matrix factorization (NMF) [4]. However, to build a system based on HMMs, multiple labels need to be provided at the same time. When we choose the deep learning method for SED, the structure and training of the neural networks directly allow multi-label classification. Parameters of the neural networks can simultaneously be

The associate editor coordinating the review of this manuscript and approving it for publication was Lorenzo Mucchi.

trained, and the neural networks can output the results [5]. Therefore, in recent years, most SED problems have used deep learning methods, such as the convolutional neural network (CNN) [6], the convolutional recurrent neural network (CRNN) [7], or the capsule network [8].

However, the CNN pooling operation has the disadvantage of losing the location information of the target object [9]. The traditional CNN network is not interpretable [10], and is a black box method. However, the interpretability is necessary in many applications, such as monitor, health care, and education. Interpretable machine learning means that machine learning model can explain why some predictions are made [11].

Interpretable deep network models are the current research hotspots. One of the interpretable methods is the optimization algorithm-driven deep networks [12]. Recently, deep neural networks driven by optimization algorithms have become increasingly popular. Gregor and LeCun [13] proposed a learned iterative soft threshold algorithm (LISTA) network, which uses a learning matrix to produce the lowest possible loss in a given number of iterations. Borgerding *et al.* [14] proposed a learned approximate message passing (LAMP) network and a learned vector AMP (LVAMP) network. The LAMP network significantly improved the LISTA network. Ito *et al.* [15] proposed a novel sparse signal recovery algorithm for trainable ISTA (TISTA). TISTA consists of two estimation units, a linear estimation unit and a minimum mean squared error (MMSE) estimator-based shrinkage unit. The numerical results show that TISTA converges faster than AMP and LISTA.

The convolutional sparse coding (CSC) model and the optimization algorithms have strong prior knowledge [16]. The CSC prior replaced the traditional image patch-based model with a global shift-invariant model. It proposes a global dictionary constrained by a specific structure - a concatenation of banded circular matrices, which limits the degrees of freedom introduced by general sparsity-based model. The dictionary is an important factor in the formation of the priori, because its atoms represent the signals that this model can sparsely represent. The l_1 sparse constraint prior condition is applied to sparse coding solved by the optimization algorithms [17], [18]. The dictionary and sparse code of multi-layer convolutional sparse coding (ML-CSC) also inherit the same prior knowledge [19]. The ML-CSC optimization algorithms can be converted into the iterative neural networks, and extract features that are different extracted from the CNN. The CNN may not have strong constraints similar to these algorithms. For CSC problems, Zisselman *et al.* [20] proposed a based local block coordinate descent (LoBCoD) algorithm for performing global the basis pursuit and introduced a new stochastic gradient descent version of LoBCoD for training the convolutional filters. For ML-CSC problems, Sulam *et al.* [21] proposed a multi-layer ISTA (ML-ISTA) and a multi-layer FISTA (ML-FISTA) algorithm. The two algorithms can converge to the global optimum. ML-ISTA-NET is a deep network structure based on the iterative unfolding of the ML-ISTA algorithm. The learnable network parameters are updated by the backpropagation algorithm in deep learning. One iteration of ML-ISTA algorithms implements a traditional CNN while a new recurrent architecture emerges with the subsequent iterations.

Inspired by ML-ISTA and the corresponding iterative unfolding network ML-ISTA-NET, we extend the LoBCoD algorithm to the multi-layer basis pursuit problem of ML-CSC. A multi-layer local block coordinate descent (ML-LoBCoD) algorithm and multi-layer local block coordinate descent network (ML-LoBCoD-NET) with iterative unfolding are proposed. ML-LoBCoD-NET implements the representation learning of the signal, and the output of the deepest convolutional sparse coding is used for classification. ML-LoBCoD-NET retain ReLU and convolution operation, use the strong constraints of the ML-CSC algorithm, and don't use pooling operation.

Inspired by the CRNN-Att network [22], the MRNN-Att network combines the ML-LoBCoD-NET, a recurrent neural network (RNN) and an attention network is proposed. The MCRNN-Att network combines MRNN-Att and CRNN network for fused the different features. The MRNN-Att network and MCRNN-Att network are used for weakly-supervised sound event detection tasks.

Many methods for solving SED problems rely on a fully supervised approach using strong labeled data (SLD). However, strong labeled data needs to label the start and offset times of the audio events, and the process of creating a large number of SLD requires a large amount of time, which is a difficult and expensive process [23]. Recently, many audio datasets have been weakly labeled and are typically larger than strongly labeled SED datasets. Compared with SLD, weakly labeled data (WLD) only knows if there is an audio event in the record. A strong label is the start and offset times of the audio event class. A weak label is the class label of the audios. Weakly supervised learning is studied for sound event detection from weakly labeled datasets, and some of the models include the joint separation-classification (JSC) model [24], the attention and positioning model [25], and the multiple instance learning (MIL) [26] method. Tarvainen and Valpola [27] proposed the mean teacher (MT) model for the weakly supervised learning of images. The mean teacher model can solve semi-supervised learning problem and can effectively use unlabeled data. The mean teacher model includes the student model and the teacher model. The student model and the teacher model currently all use the same model. The main purpose of the mean teacher model is that averaging the model weights over the training steps tends to produce a more accurate model than using the final weights directly. A key issue of the mean teacher model is the choice of the student model. For example, If the student model is chosen as the traditional model such as SVM, the mean teacher model can only solve the traditional supervised learning problem. If the student model is selected as the CRNN model which commonly used for sound event detection or

the models we will proposed, the mean teacher model can solve the weakly supervised learning problem. In summary, the choice of the student model in the mean teacher model framework determines whether it can deal with supervised learning problem or weakly supervised learning problem. No matter which learning mode is chosen, the mean teacher model framework can deal with semi-supervised learning.

Inspired by the mean teacher model to solve the semi-supervised problem, this paper proposes two mean teacher models for sound event detection tasks in the domestic environment. The first our proposed mean teacher model is the MRNN-Att-MT, and the student model is the MRNN-Att. The second our proposed mean teacher model is the MCRNN-MT, and the student model is the MCRNN-Att.

The weakly-labeled sound event detection task is the core problem, the proposed MRNN-Att network is the core method in this paper. The MRNN-Att network is based on the ML-LoBCoD-NET which is driven by the ML-LoBCoD algorithm. The ML-LoBCoD-NET shows the feature extraction ability different from the CNN for sound event detection task. The MCRNN-Att network is combined of the MRNN-Att and the CRNN-Att. We use the MRNN-Att and the MCRNN-Att as the student model in mean teacher model, respectively.

The remainder of the paper is organized as follows: the CRNN-Att network is introduced in Section II-A, and the CRNN-Att-MT model is introduced in Section II-B. The ML-LoBCoD algorithm and the ML-LoBCoD-NET are proposed in Section III. In section IV, for the weakly-labeled weakly-supervised sound event detection task, the MRNN-Att network based on the ML-LoBCoD-NET network is proposed in section IV-A. Moreover, in order to fully utilize the feature information of the CNN and ML-LoBCoD-NET network, the MCRNN-Att network is proposed in section IV-B. In section V, for the weakly-labeled semi-supervised sound event detection task, the MRNN-Att-MT is proposed. Moreover, the MCRNN-Att-MT is proposed for sound event detection task in section V. The experimental results and analysis are given in Section VI. The conclusion is given in Section VII.

II. BACKGROUND

A. THE CRNN-ATT NETWORK FOR SOUND EVENT DETECTION TASK

For sound event detection task, a weakly-supervised learning model is need. The CRNN-Att network is a common model for sound event detection task [25], which is described below. A CNN consists of three basic components, convolutional layers, pooling layers, and fully-connected layers. A convolutional layer first performs convolution operations to produce a set of linear activation, which then is fed into a non-linear activation function like ReLU or tanh. Pooling layers are usually used after each convolutional layer to reduce the representation size of convolutional output and the computational burden of the next layers. The pooling function divides its input into a set of rectangles, and each sub-region generates a summary statistic of the input nearby. The use of pooling is very useful for extracting the most effective information from an area. After several convolutional layers and pooling layers, the fully connected layers are adopted at the end of a CNN. A fully-connected layer in a CNN is similar to the layer in a standard neural network where the neurons in the adjacent layers are fully pairwise connected and the neurons in the same layer share no connection.

The advantage of a CNN is that it can effectively process the spatial structure data with large width and height. The function of a RNN is to be extended to longer sequences. In a RNN, a hidden layer with a self-joining unit acts as memory that accumulates information over time from the input sequence. However, there is the problem that the gradient disappears when training the RNN to capture long-term dependency. To combat the gradient disappearance problem, several techniques have been proposed, such as long short term memory (LSTM) [28] and gated recurrent unit (GRU) [22]. The LSTM and GRU architectures accumulate information by replacing self-joining units with memory blocks, which better capture the long-term dependencies in time series data.

The CRNN is a network structure that combines a CNN and a RNN, benefiting from the advantages of both. A RNN can work well in a time domain while a CNN can apply a linear convolutional filter in the time domain and frequency domain of local features. In addition, a CRNN has proven to work well in sound event detection tasks [7].

The attention mechanism can increase the focus on the important time frames through weighting, and the attention layer can automatically select the important frames of the target and ignore the irrelevant parts (such as background noise segments). It can also be viewed as a weighting factor for learning each frame. The system can suppress the background noise, and thus the whole system is more robust [22]. The CRNN-Att network that combines a RNN and an attention network has also been used for sound event detection tasks and has achieved good results [25].

B. THE MEAN TEACHER MODEL BASED ON THE CRNN-ATT NETWORK FOR SOUND EVENT DETECTION TASK

For semi-supervised sound event detection task, the mean teacher model is a new method [27]. The mean teacher model can effectively utilize large amounts of unlabeled data. The mean teacher model based on the CRNN-Att network was used for SED task in Detection and Classification of Acoustic Scenes and Events (DCASE) 2018 challenge and obtained the first place [29]. Then the mean teacher model based on the CRNN-Att network was as the baseline system in DCASE 2019 challenge [30].

The mean teacher model consists of a student model and a teacher model, and the teacher model uses the same model as the student model [27]. For sound event detection task, the student model in mean teacher model uses a weakly-labeled and weakly-supervised deep learning model.

The input of the student model and the teacher model is the same sample with different noise. The output of the student model and the teacher model both include strong labels and weak labels. The main purpose of the mean teacher model is to average the parameters of the student models on the training steps to obtain the parameters of the teacher models. It is easier to obtain accurate results with the teacher model that uses the average parameters than with the student model that directly uses the final parameters. Thus the final output of the strong and weak labels of the mean teacher model are the strong and weak labels obtained from the teacher model.

For weakly labeled training data, there are three losses, which are the classification loss, strong consistency loss, and weak consistency loss. The classification loss is the multi-class cross entropy loss of the weak labels generated by the student model and weak reference labels of the training data. The strong consistency loss refers to the consistency loss of time frame level between the strong labels generated by the student model and the strong labels generated by the teacher model. The weak consistency loss refers to the consistency loss of the clip level between the weak labels generated by the student model and the weak labels generated by the teacher model.

There are two losses for unlabeled training data, the strong consistency loss and weak consistency loss.

The five loss functions are weighted for optimization. The backpropagation algorithm is used to update the parameters of the student model. After the parameters of the student model are updated, the parameters of the teacher model are updated to an exponential moving mean or a random weighted mean of the student parameters.

III. THE PROPOSED ML-LoBCoD ALGORITHM AND ITS ITERATIVE UNFLOD NETWORK

The pooling operation in the CNN has the disadvantage of losing the location information of the target object. We don't use the pooling operation, retaining ReLU and convolution operation, and we use the dictionary strong constraints and l_1 penalty function prior constraints of the multi-layer convolutional sparse coding (ML-CSC). The iterative unfolded multi-layer local block coordinate descent networks (ML-LoBCoD-NET) is proposed in this section. The ML-LoBCoD-NET is driven by the multi-layer local block coordinate descent algorithm (ML-LoBCoD) which is extended from the local block coordinate descent (LoBCoD) algorithm.

A. THE PROPOSED ML-LoBCoD ALGORITHM

Given a set of convolutional dictionaries $\{D_j\}_{j=1}^{J}$ or convolutional filter $\{D_{L,j}\}_{j=1}^{J}$ with appropriate dimensions, the global signal $X \in R^N$ can be represented by the slice-based multi-layer convolutional sparse coding (MLCSC-S) as follows:

$$X = D_1\Gamma_1 = \sum_{i=1}^{N} P_{1,i}^T D_{L,1}\alpha_{1,i}, \quad \|\Gamma_1\|_{0,\infty} \le \lambda_1, \tag{1}$$

$$\Gamma_1 = D_2\Gamma_2 = \sum_{i=1}^{N} P_{2,i}^T D_{L,2}\alpha_{2,i}, \quad \|\Gamma_2\|_{0,\infty} \le \lambda_2, \tag{2}$$

$$\dots$$

$$\Gamma_{J-1} = D_J\Gamma_J = \sum_{i=1}^{N} P_{J,i}^T D_{L,J}\alpha_{j,i}, \quad \|\Gamma_J\|_{0,\infty} \le \lambda_J. \tag{3}$$

where the norm $l_{0,\infty}$ is defined as the maximal number of non-zeros in the vector; Γ_j is the sparse representation of the j-th layer; $P_{j,i}^T$ is the operator that extracts the i-th n-dimensional patch from the j-th layer sparse representation Γ_j; $D_{L,j}$ is the j-th layer local dictionary; $\alpha_{j,i}$ is α_i of the j-th layer; λ_j is a super parameter; N is the number of slices, and the N in each layer should be different. For the sake of simplicity, it assumed that all layers have the same N. Let Γ_0 denote signal X, and then the MLCSC-S model can be rewritten as follows:

$$\Gamma_{j-1} = D_j\Gamma_j = \sum_{i=1}^{N} P_{j,i}^T D_{L,j}\alpha_{j,i}, \quad \left\|\Gamma_j\right\|_{0,\infty} \le \lambda_j, \ \forall 1 \le j \le J. \tag{4}$$

According to Formulas (4), the multi-layer basis pursuit problem proposed in this paper is as follows:

$$\min_{\{\alpha_{j,i}\}_{i=1}^{N}} \frac{1}{2}\left\|\Gamma_{j-1} - \sum_{i=1}^{N} P_{j,i}^T D_{L,j}\alpha_{j,i}\right\|_2^2 + \lambda_j \sum_{i=1}^{N} \left\|\alpha_{j,i}\right\|_1, \quad j = 1 \dots J. \tag{5}$$

The slice-based local block coordinate descent algorithm is extended into the multi-layers. Then, a slice-based multi-layer local block coordinate descent algorithm (ML-LoBCoD) is proposed. The ML-LoBCoD algorithm divides the layer sparse vector Γ_j into a local vector $\{\alpha_{j,i}\}_{i=1}^{N}$, and then the optimal solution for needles $\alpha_{j,i}$ is searched for. The other needles are fixed and regarded as constants. Equation (5) can be written as

$$\min_{\alpha_{j,i}} \frac{1}{2}\left\|(\Gamma_{j-1} - \sum_{\substack{n=1 \\ n\neq i}}^{N} P_{j,n}^T D_{L,j}\alpha_{j,n}) - P_{j,i}^T D_{L,j}\alpha_{j,i}\right\|_2^2 + \lambda_j \left\|\alpha_{j,i}\right\|_1, \quad i = 1 \dots N. \tag{6}$$

By defining $R_{j,i} = (\Gamma_{j-1} - \sum_{\substack{n=1 \\ n\neq i}}^{N} P_{j,n}^T D_{L,j}\alpha_{j,n})$ as the layer residual of the contribution of the needle $\alpha_{j,n}$ in each layer, Equation (6) can be rewritten as

$$\min_{\alpha_{j,i}} \frac{1}{2}\left\|R_{j,i} - P_{j,i}^T D_{L,j}\alpha_{j,i}\right\|_2^2 + \lambda_j \left\|\alpha_{j,i}\right\|_1, \quad i = 1 \dots N \tag{7}$$

Equation (7) is further organized as follows:

$$\min_{\alpha_{j,i}} \frac{1}{2}\left\|P_{j,i}R_{j,i} - D_{L,j}\alpha_{j,i}\right\|_2^2 + \lambda_j \left\|\alpha_{j,i}\right\|_1, \quad i = 1 \dots N \tag{8}$$

where $P_{j,i} \in R^{N\times n}$ is defined as the operator that extracts the i-th n-dimensional patch from the j-th layer convolutional sparse coding Γ_j. The optimal convolutional sparse

representation of each layer is updated using the soft threshold formula. The updated form is $\alpha^k \leftarrow \alpha^{k-1} - \frac{\partial f}{\partial \alpha^{k-1}}$, where f denotes the error term in the objective function of Equation (8). The derivative formula can be given as follows:

$$\frac{\partial f}{\partial \alpha_{j,i}} = D_{L,j}^T(P_{j,i}R_{j,i} - D_{L,j}\alpha_{j,i}) \tag{9}$$

Similar to the ML-ISTA, the update of the needle $\alpha_{j,i}$ is expressed as

$$\alpha_{j,i}^k = S_\lambda(\alpha_{j,i}^{k-1} + D_{L,j}^T(P_{j,i}R_{j,i} - D_{L,j}\alpha_{j,i}^{k-1})), \quad i = 1 \ldots N, \quad j = 1 \ldots J. \tag{10}$$

where S_λ is a soft threshold function, which can also be replaced by the ReLU function [21]; $P_{j,i}$ represents the operation of image to column; $P_{j,i}^T$ represents the operation of column to image; $D_{L,j}$ represents the transposed convolution operation; and $D_{L,j}^T$ represents convolution operation.

The proposed slice-based multi-layer local block coordinate descent algorithm is shown in Algorithm 1. The input signal is the signal y contaminated by the noise w, and the output is the layer sparse representation $\{\Gamma_j\}_{j=1}^J$ or needle $\{\{\alpha_{j,i}\}_{i=1}^N\}_{j=1}^J$. The local convolutional dictionary $\{D_{L,j}\}_{j=1}^J$, the top convolutional sparse coding Γ_0, and the top layer residual R_0 are initialized. In each iteration process, first, the local residual $R_{j,i}^k$ is obtained according to the local residual R_j^{k-1} of the previous iteration and the i-th needle $\alpha_{j,i}^{k-1}$ of the j-th layer convolutional sparse code of the previous iteration. Secondly, the needle $\alpha_{j,i}^k$ is updated using the optimization Formula (10). All the updated needles combine the convolutional sparse coding $\Gamma_j^k = \{\alpha_{j,i}\}_{i=1}^N$ of the j-th layer. Thirdly, the layer residual R_j^k is computed by Γ_{j-1}^k of the $(j-1)$-th layer and Γ_j^k of the j-th layer. The convolutional sparse coding of the next layer is updated. The iteration process is repeated K times until the optimized deep convolutional sparse coding is obtained.

Algorithm 1 Slice-Based Multi-Layer Local Block Coordinate Descent Algorithm(ML-LoBCoD)

Input: signal x;
Initialization: Local convolutional dictionary $\{D_{L,j}\}_{j=1}^J$, $\Gamma_0 = x, R_1 = x$;
1. Iterative: $k = 1 : K$;
2. Number of layers: $j = 1 : J$;
3. Number of needles: $i = 1 : N$;
4. Calculate local residual: $R_{j,i}^k = R_j^{k-1} + P_{j,i}^T D_{L,j}\alpha_{j,i}^{t-1}$
5. needle update: $\alpha_{j,i}^k = S_\lambda(\alpha_{j,i}^{t-1} + D_{L,j}^T(P_{j,i}R_{j,i} - D_{L,j}\alpha_{j,i}^{t-1}))$
6. Update layer residual: $R_j^k = \Gamma_{j-1}^k - D_{L,j}\,\Gamma_j^k$
Output: Layer sparse representation $\{\Gamma_j\}_{j=1}^J$ Or needle $\{\{\alpha_{j,i}\}_{i=1}^N\}_{j=1}^J$

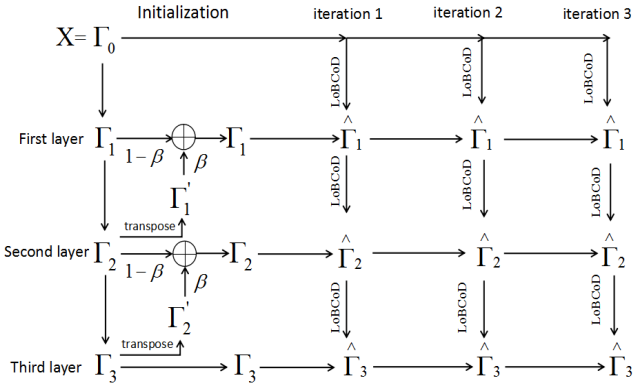


FIGURE 1. The unfolding ML-LoBCoD-NET using three layers in three iterations.

B. THE PROPOSED ITERATIVE UNFOLDED MULTI-LAYER LOCAL BLOCK COORDINATE DESCENT NETWORK (ML-LoBCoD-NET)

The slice-based multi-layer local block coordinate descent algorithm (ML-LoBCoD) provides an effective algorithm for the multi-layer basis pursuit. The iterative unfolding multi-layer local block coordinate descent network (ML-LoBCoD-NET), which is similar to the ML-ISTA-NET [21], is proposed, as shown in FIGURE 1.

The LoBCoD algorithm is unfolded into a layer of neural networks. The ML-LoBCoD algorithm is unfolded into a multi-layer neural network. ML-LoBCoD-NET iterates K times to form a recurrent structure because the algorithm iterates many times to obtain the optimal performance. Its parameters are the same as a traditional CNN, and thus the network parameters remain unchanged. In FIGURE 1, the ML-LoBCoD-NET is a three-layer feedforward neural network. One iteration of the ML-LoBCoD-NET corresponds to the traditional CNN. Three iterations of the ML-LoBCoD-NET correspond to FIGURE 1. The input of ML-LoBCoD-NET is X and the output is Γ_3.

Firstly, the input signal is X, and the initial values Γ_1, Γ_2 and Γ_3 are generated by a standard convolution operation. Γ_1' is generated by a deconvolution operation using Γ_2. Γ_2' is generated by a deconvolution operation using Γ_3. Then, the final initial value Γ_1 is obtained by weighting Γ_1 and Γ_1', and the final initial value Γ_2 is obtained by weighting Γ_2 and Γ_2'. β is a weight. When $\beta = 0$, the signal doesn't satisfy the ML-CSC model; when $\beta = 1$, the signal satisfies the ML-CSC model. In this experiment, the value of β gradually increased from 0 to 1.

Secondly, the first iteration of the algorithm is unfolded to the three-layer network corresponding to the third column in FIGURE 1. The CSC estimator $\widehat{\Gamma}_1$ of the first layer is obtained by the ML-LoBCoD algorithm using Γ_0 and Γ_1. The CSC estimate $\widehat{\Gamma}_2$ of the second layer is obtained by the ML-LoBCoD algorithm using $\widehat{\Gamma}_1$ and Γ_2. The CSC estimate $\widehat{\Gamma}_3$ of the third layer is obtained by the ML-LoBCoD algorithm using $\widehat{\Gamma}_2$ and Γ_3.

TABLE 1. Classification accuracy of several classification networks on MINIST.

Net	CNN	ML-ISTA	ML-FISTA	ML-LISTA	LBP	ML-LoBCoD
Classification correct rate(%)	98.74	99.11	99.16	98.8	99.19	99.15

Thirdly, the second iteration of the algorithm is unfolded to the three-layer network corresponding to the fourth column in FIGURE 1. The CSC estimator of the first layer $\widehat{\Gamma}_1$ is obtained by the ML-LoBCoD algorithm using Γ_0 and Γ_1. The CSC estimator $\widehat{\Gamma}_1$ of the first layer is obtained by the ML-LoBCoD algorithm using Γ_0 and Γ_1. The CSC estimate $\widehat{\Gamma}_2$ of the second layer is obtained by the ML-LoBCoD algorithm using $\widehat{\Gamma}_1$ and Γ_2. The CSC estimate $\widehat{\Gamma}_3$ of the third layer is obtained by the ML-LoBCoD algorithm using $\widehat{\Gamma}_2$ and Γ_3.

Finally, the third iteration of the algorithm is unfolded to the three-layer network corresponding to the fifth column in FIGURE 1. The CSC estimator $\widehat{\Gamma}_1$ of the first layer is obtained by the ML-LoBCoD algorithm using Γ_0 and Γ_1. The CSC estimate $\widehat{\Gamma}_2$ of the second layer is obtained by the ML-LoBCoD algorithm using $\widehat{\Gamma}_1$ and Γ_2. The CSC estimate $\widehat{\Gamma}_3$ of the third layer is obtained by the ML-LoBCoD algorithm using $\widehat{\Gamma}_2$ and Γ_3. A fully connected layer is added after $\widehat{\Gamma}_3$ as a classifier.

ML-LoBCoD-NET was tested on the Mnist dataset (http://yann.lecun.com/exdb/mnist/). The classification accuracy rates of the CNN, ML-ISTA, ML-FISTA, ML-LISTA, Layered Basis Pursuit (LBP), and ML-LoBCoD are given in Table 1. The classification accuracies of the ML-ISTA, ML-FISTA, and ML-LoBCoD network under different iteration times are given in FIGURE 2. As shown in Table 1, the classification accuracy of the ML-LoBCoD network is higher than the classification accuracy of the CNN and ML-ISTA on the MINIST dataset. In addition, the classification accuracy of the ML-LoBCoD network is better than the classification accuracy of the ML-ISTA and ML-FISTA in FIGURE 2, and the classification accuracy of the ML-LoBCoD network is more stable than the classification accuracy of the ML-ISTA and ML-FISTA.

ML-LoBCoD-NET was tested on the CIFAR10 dataset (https://www.cs.toronto.edu/ kriz/cifar.html). The classification accuracy rates of the CNN, ML-ISTA, ML-FISTA, ML-LISTA, LBP, and ML-LoBCoD are given in Table 2. As shown in Table 2, the classification accuracy of the ML-LoBCoD network is higher than the classification accuracy of the CNN, ML-ISTA, ML-FISTA, ML-LISTA and LBP on the CIFAR10 dataset. The ML-LoBCoD network is better at extracting features than the CNN, ML-ISTA, ML-FISTA, ML-LISTA and LBP on the CIFAR10 dataset.

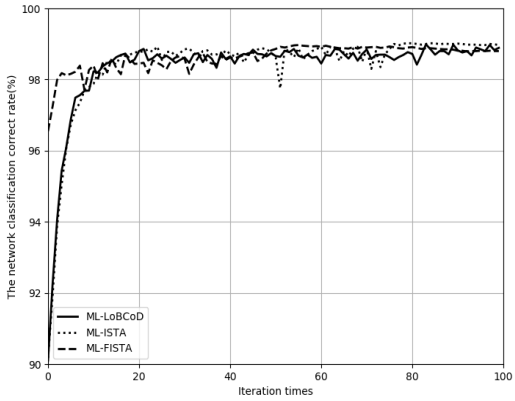


FIGURE 2. Classification accuracies of the three network under different iteration times.

TABLE 2. Classification accuracy of several classification networks on CIFAR10.

Net	CNN	ML-ISTA	ML-FISTA	ML-LISTA	LBP	ML-LoBCoD
Classification correct rate(%)	79.00	82.93	82.79	82.68	80.73	83.53

IV. THE PROPOSED MRNN-ATT NETWORK BASED ON ML-LoBCoD-NET FOR SOUND EVENT DETECTION TASK

In this section, for the weakly-supervised learning problem of sound event detection task, we first replace the CNN network of the CRNN-Att network in section II-A with the ML-LoBCoD-NET network. The MRNN-Att network based on the ML-LoBCoD-NET network is proposed for sound event detection task in section IV-A. Moreover, in order to fully utilize the feature information of the CNN and the ML-LoBCoD-NET network, the MCRNN-Att network is proposed in section IV-B.

A. THE MRNN-ATT NETWORK

The proposed MRNN-Att network is shown in FIGURE 3. The input is a log-mel spectrogram of an audio clip. The output is the prediction of the strong label and weak label, where the prediction of the weak labels data is used for the weak label training, and the prediction of the strong labels for locating the time location of the sound. The network includes the ML-LoBCoD-NET with K iterative unfoldings, a RNN network, and an attention network. The ML-LoBCoD-NET is used to extract features of the audio clip. The RNN network uses a two-layer Bi-GRU network. The Bi-GRU network can be used to capture the context information of sound events and can simulate well the long-term mode of the entire block. The attention network has two FNN layers with softmax and sigmoid layers.

One output of the network is the strong label, which is given as follows. The first attention layer uses the sigmoid activation function to predict the probability of occurrence of

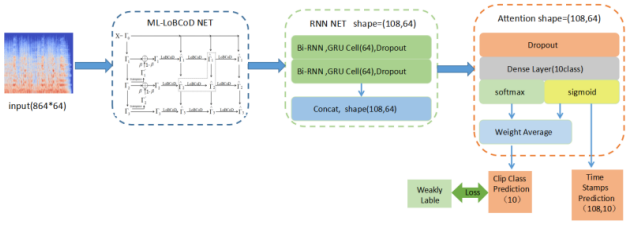


FIGURE 3. The framework of the MRNN-Att network for weakly-supervised learning.

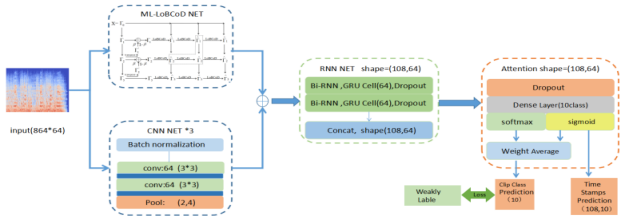


FIGURE 4. The framework of MCRNN-Att network for weakly-supervised learning.

each class at each time step, which can deal with the time overlap problem of multiple events. It can generate a class strong label $Z_{att}(c, t)$. The other output of the neural network is the weak label. To predict the occurrence probability of each class at each time step, the second attention layer uses the softmax function, and the sum is 1. The final weak label $O(c), c = 1, 2, \ldots C$ of each audio clip is determined by the weighted average of the element-by-element multiplication of the output of the first attention layer and second attention layer

$$O(c) = \frac{\sum_{t=1}^{T} Z_{class}(c, t) Z_{att}(c, t)}{\sum_{t=1}^{T} Z_{att}(c, t)}. \tag{11}$$

where $Z_{class}(c, t)$ and $Z_{att}(c, t)$ are denoted as the output of the first attention layer and the second attention layer. T is the temporal resolution of the input spectrogram or the feature map or the number of time frames.

For the training problem, the loss function uses the multi-class cross-entropy loss

$$E = -\frac{1}{N} \sum_{n=1}^{N} \sum_{c=1}^{C} P(c, n) log O(c, n). \tag{12}$$

$O(c, n)$ and $P(c, n)$ represent the weak prediction labels and the weak reference labels for the n-th sample of the c-th class, respectively. The batch size is N, and the total number of classes is C. By calculating the gradient of the loss function with respect to the network parameters using back-propagation algorithm, the parameters of the neural network can be updated.

B. THE MCRNN-ATT NETWORK

The proposed MCRNN-Att network is shown in FIGURE 4. The input is a log-mel spectrogram of the audio clip and

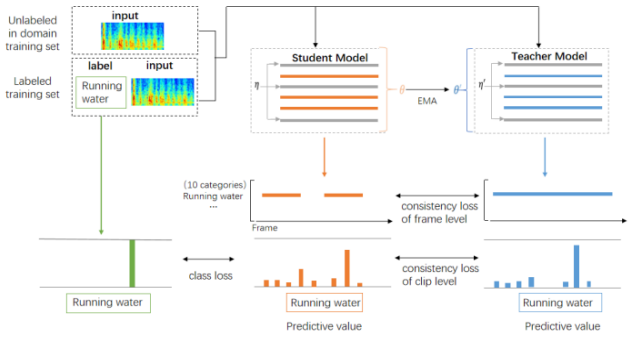


FIGURE 5. Schematic diagram of the mean teacher model for semi-supervised learning.

the output is the prediction of the strong label and weak label. The network includes the ML-LoBCoD-NET with K iterative unfoldings, a CNN network, a RNN network, and an attention network. In this paper, a three-layer CNN network is used. The output of the CNN network and ML-LoBCoD-NET are fused and fed into the RNN network. The RNN network also uses a two-layer Bi-GRU network. The attention network can increase the attention to important time frames by weighting and can automatically select and participate in important frames of the target while ignoring the irrelevant parts.

V. THE PROPOSED MRNN-ATT-MT MODEL FOR SOUND EVENT DETECTION TASK

In this section, for the semi-supervised learning problem of sound event detection task, we first replace the CRNN-Att network in the MT model in section II-B with the MRNN-Att network. The MRNN-Att-MT model is proposed for sound event detection task. The MRNN-Att-MT model uses the MRNN-Att network as a student model. Moreover, the MCRNN-Att-MT model is proposed for sound event detection task. The MCRNN-Att-MT model uses the MCRNN-Att network as a student model.

A schematic diagram of the mean teacher model is shown in FIGURE 5. The inputs of the mean teacher model are unlabeled in the domain training set and the labeled training set. The outputs of the mean teacher model are the strong label and weak label. The mean teacher model consists of the student model and teacher model. Both the student and teacher model evaluate the input applied noise (η, η'). The student model outputs strong labels and weak labels. The teacher model also outputs strong labels and weak labels. Five loss functions are computed. After the parameters of the student model have been updated with the backpropagation algorithm, the teacher model weights are updated as an exponential moving average of the student weights.

The specific definition of the loss function is given below. In the semi-supervised setup, the weakly labeled data $D_A = \{x_i, y_i\}_{i=1}^{N_A}$ and unlabeled data $D_U = \{x_i^U\}_{i=1}^{N_U}$ are used. Parameters of the student model are denoted by θ, and the parameters of the teacher model are denoted by θ'. $f(x; \theta)$

indicates weak labeled output of the student model. The weak labeled output of the teacher model is represented by $f(x;\theta')$. The strong labeled output of the student model is represented by $f_{strong}(x;\theta)$, and the strong labeled output of the teacher model is represented by $f_{strong}(x;\theta')$.

Firstly the multi-class cross entropy loss function $L_{ce}(x,y,\theta)$ in the supervised training is defined as follows:

$$L_{ce}(x,y,\theta) = -\sum_{(x,y)\in D_A} ylogf(x,\eta;\theta). \tag{13}$$

Secondly, given the sample x of two disturbance inputs η and η' and the two network disturbance parameters θ and θ', the strong consistency loss between the strong prediction label of the student model $f_{strong}(x,\eta;\theta)$ and the strong prediction label of the teacher model $f_{strong}(x,\eta';\theta')$ is defined as the mean square error loss form as follows:

$$L_{consstrong}(x,\theta) = \left\|f(x,\eta;\theta) - f(x,\eta';\theta')\right\|_2^2. \tag{14}$$

The weak consistency loss between the predictive weak label of the student model $f(x,\eta;\theta)$ and the predictive weak label of the teacher model $f(x,\eta';\theta')$ are defined as the form of multi-class cross entropy loss as follows:

$$L_{consweak}(x,\theta) = -\sum_{x\in D_A\cup D_U} f(x,\eta;\theta)log\, f(x,\eta';\theta'). \tag{15}$$

Finally, the total loss for training the student model is defined as

$$\begin{aligned} L(\theta) = &\sum_{(x,y)\in D_A} L_{ce}(x,y;\theta) + \lambda_1 \sum_{x\in D_A} L_{consstrong}(x,\theta) \\ &+\lambda_2 \sum_{x\in D_U} L_{consstrong}(x,\theta) + \lambda_3 \sum_{x\in D_A} L_{consweak}(x,\theta) \\ &+\lambda_4 \sum_{x\in D_U} L_{consweak}(x,\theta). \end{aligned} \tag{16}$$

The parameters λ_1, λ_2, λ_3 and λ_4 control the relative importance of the consistency term in the total loss. In this study, the values of λ_1,λ_2,λ_3 and λ_4 were set to 1 for simplicity.

The training process is organized as follows:

(1) The input of the mean teacher model are unlabeled in the domain training set and labeled training set. Data are input into the mean teacher model to generate the output, which are strong labels and weak labels.
(2) Five losses are calculated, and then the total loss is calculated according to Formula (16).
(3) Parameters of the student model are updated using the backpropagation algorithm based on minimizing the total loss.
(4) Using the parameters of the student network, the parameters of the teacher model are updated to the average value of the current student model and the previous student model.
(5) The above processes are repeated until the network converges.

TABLE 3. Development dataset and evaluation 10s clip number and the number of sessions.

Class	Labeled Training events	Testing clips	Testing events	Evaluation events
Speech	550	105	261	1401
Dog	214	29	127	450
Cat	173	32	97	243
Alarm/bell/ringing	205	45	112	306
Dishes	184	35	35	457
Frying	171	24	24	67
Blender	134	30	40	56
Running water	343	63	76	154
Vacuum cleaner	167	35	36	56
Electric shaver/toothbrush	103	25	28	37
Total	2244	288	906	3227

VI. EXPERIMENT

A. DATASET

The dataset used in this paper for the experiment was the dataset for the DCASE 2018 Challenge Task 4 [31]. The task used weakly labeled data (without timestamps) to evaluate large-scale detection systems for the sound events. A subset of the audio set was extracted from YouTube videos, consisting of various sound categories that occur in a home environment. The dataset of DCASE 2018 task4 include the development dataset and evaluation dataset. The development set include a training set and a test set. The training set included a labeled training set, unlabeled in the domain training set, and an unlabeled offset from the domain training set. The labeled training set included 1578 clips (2244 class occurrences), 14412 clips unlabeled in the domain training set, and 39999 clips unlabeled out of the domain training set. The validation set was 20% of the labeled training set. The test set contained 288 clips (906 events). The test data was annotated with the time boundary of each marked event. The test set was annotated with strong labels and time boundaries (obtained by human annotators). The evaluation set included 880 clips (3227 events). The audio clips were divided into nine categories: alarm, speech, dog, cat, dishes, frying, blender, running water, vacuum cleaner, and electric shaver. Table 3 shows the number of labeled training sets and test sets of the development set and the evaluation set 10s clips along with the number of complete sessions for each activity.

B. FEATURE EXTRACTION

The log Mel spectrum is widely used in the sound events detection [32]. So we used log Mel filters to process audio clips. Each audio clip was first resampled at 44.1 KHZ because we believe that resampling at low frequencies may confuse some categories like ''electric shaver/toothbrush''and ''vacuum cleaner''. After resampling, we apply a short-time Fourier transform with a window size of 2048 and an overlap of 512 between neighboring windows to extract the spectrogram of audio clips. Following this configuration the good resolution in both the time and frequency domains is provided. Then a Mel filter bank with 64 bands

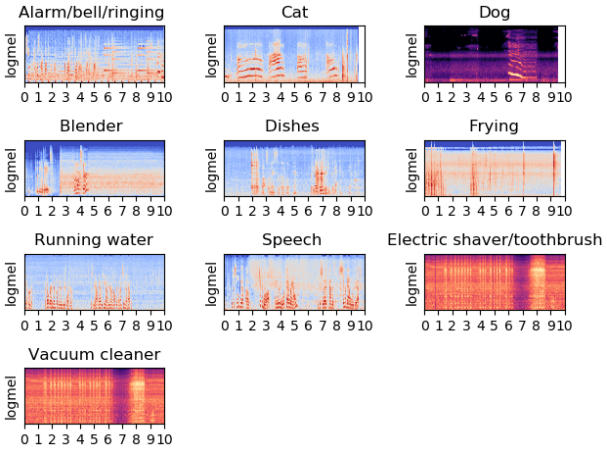


FIGURE 6. Log Mel spectrum of each audio event.

is applied to the spectrogram, and a logarithmic operation is performed to obtain the log Mel spectrogram, which is the time and frequency representation feature. Thus for a 10s clip, a 864×64 feature was obtained. The log Mel spectrum of each audio event are shown in FIGURE 6.

C. EXPERIMENTAL SETUP

In this study, the ML-LoBCoD-NET in MRNN-Att network and in the MCRNN-Att network used three-layer convolutional sparse coding. The first layer had 64 filters, the kernel size was (4, 8), and the step size was 2. The second layer had 64 filters, the kernel size was (3, 12), and the step size was 2. The third layer had 64 filters, the kernel size was (3, 12), and the step size was 2. The CNN network in the MCRNN-Att network used a three-layer convolution neural network with 64 filters per layer, the kernel size of convolution layer was (3, 3), the step size was 1, and the kernel size of the pooling layer was (2, 4).

The development dataset are divided into the training dataset and test dataset. The 80 percent of the training dataset are used to train the model. The 20 percent of the training dataset as the validation dataset are used to verify the F1 score of the training model, and the results on the validation dataset are used to adjust the model parameters of the training model to obtain an optimal training model. To evaluate the performance of the training model on the development dataset, the test dataset are put to the training model to obtain performance indicators. The predicted strong labels on the test dataset are obtained. Moreover, the performance indicators such as F1 score based on the predicted strong labels and ground truth are obtained. The performance indicators of the test dataset represent the experimental results of the development dataset. Furthermore, the evaluation dataset are put to the optimal training model to obtain performance indicators, such as F1 score, based on the predicted strong labels and ground truth. This training and test process is the same as the baseline system [33].

In the training phase, different hyperparameters are used to train different models, and the training model with the highest F1 score on the validation set is selected. The selection range of the hyperparameters are as follows. The learning rate are selected from 0.1, 0.01, and 0.001. The sampling frequency are selected from 44.1KHZ, 16KHZ, and 8KHZ. The Mel frequency points are selected from 128, 64, and 32. The window shift are selected from 512 and 256. The batch size are selected from 8, 16 and 32. EMA are selected from 0.9, 0.99, and 0999. After expensive experiments, the optimal learning rate is selected 0.001, the optimal sampling frequency is selected 44.1KHZ, the optimal mel frequency is selected 64, the optimal window shift is selected 512, the optimal batch size is selected 8, the optimal EMA is selected 0.99, and the number of iterations is 100.

We choose ER and F1 score as evaluation metric. Error rate (ER) is used as a secondary metric to assess errors in terms of insertions, deletions, and substitutions. F1 was used to evaluate the model, which is defined as follows:

$$F1_c = \frac{2TP_c}{2TP_c + FP_c + FN_c}. \tag{17}$$

where TP_c, FP_c and FN_c represent the true, false positive, and false negative numbers of the sound event class C, respectively. The average of the F1 scores of the final model was calculated using the macro average

$$F1_{macro} = \frac{\sum_{c \in 1,\ldots C} F1_c}{C}. \tag{18}$$

where C represents the number of the sound event class, which is 10. All the ER and F1 calculations in this paper used the sed_eval kit provided in the competition [32].

The equipment used in the experiment was an Nvidia Geforce 1080 Ti GPU, and each experiment needed to run for about seven hours.

D. EXPERIMENTAL RESULTS AND ANALYSIS

The input of the baseline system [33] provided by the DCASE 2018 Task 4 is a log Mel spectrogram of the audio clip, the output is the F1 value. The baseline system is the CRNN network, which includes the CNN network and the RNN network. The first place in the DCASE 2018 Task 4 used the mean teacher model, which is a fusion model [30]. In order to compare with our proposed model, we use the single model of the first place, called GCRNN-Att-MT. The GCRNN-Att-MT is a mean teacher model in which the student model is GCRNN-Att, which includes the CNN network, a Bi-GRU network, and an attention network. The CNN network in the GCRNN-Att-MT model used a three-layer convolutional neural network with 64 filters per layer. For each filter, the kernel size was (3, 3), the step size was 1, and the kernel size of the pooling layer was (2, 4). The RNN network used a 2-layer RNN network with 64 filters per layer and the batch size is 24. Moreover, the number of iterations was 100. In this study, the proposed models was compared with the baseline system, the GCRNN-Att-MT model, and the GCRNN-Att network using the F1 values.

TABLE 4. The F1 score and error rate of BASELINE, MRNN-Att, MCRNN-Att and GCRNN-Att models.

System	Development set		Evaluation set	
	F1 score(%)	ER	F1 score(%)	ER
baseline	14.06	1.54	10.80	1.77
MRNN-Att	16.24	2.55	10.78	2.13
MCRNN-Att	16.01	2.29	11.23	2.76
GCRNN-Att	15.56	2.10	10.96	1.96

The F1 score and error rate (ER) of baseline, MRNN-Att, MCRNN-Att and GCRNN-Att models are given in Table 4. For the development set in Table 4, the F1 scores of the MRNN-Att model and the MCRNN-Att model were respectively 2.18% and 1.95% higher than those of the baseline system, which indicates that the attention network can improve the performance of sound event detection by focusing on the relevant frames to ignore irrelevant frames. The F1 score of the MRNN-Att model was 0.68% higher than that of the GCRNN-Att model. The F1 score of the MCRNN-Att model was 0.45% higher than that of the GCRNN-Att model. These results indicate that the MRNN-Att model and the MCRNN-Att model were better than the GCRNN-Att model, and the extracted feature of the ML-LoBCoD-NET was effective. For the evaluation set in Table 4, MCRNN-Att model was 0.43% higher than baseline system, and was 0.27% higher than GCRNN-Att system.

The F1 score and Error rate of different mean teacher model systems are given in Table 5. For the development set in Table 5, the F1 score of the MRNN-Att-MT model was 8.77% higher than that of the baseline system. The F1 score of the MCRNN-Att-MT model was 6.29% higher than that of the baseline system. These results show that the sound event detection effect of the MRNN-Att-MT model and the MCRNN-Att-MT model on the development set was better than the baseline system. The F1 score of the MRNN-Att-MT model was 3.49% higher than that of the GCRNN-Att-MT model. The F1 score of the MCRNN-Att-MT model was 1.01% higher than that of the GCRNN-Att-MT model. This indicates that the sound event detection effect of MRNN-Att-MT and MCRNN-Att-MT was better than that of GCRNN-Att-MT on the development set.

For the evaluation set in Table 5, the F1 score of the MRNN-Att-MT model was 4.88% higher than that of the baseline system. The F1 score of the MCRNN-Att-MT model was 3.76% higher than the baseline system. This shows that the sound event detection effects of the MRNN-Att-MT model and the MCRNN-Att-MT model were better than that of the baseline system (CRNN). The F1 score of the MRNN-Att-MT model was 0.95% higher than that of the GCRNN-Att-MT model. Comparison of F1 scores between the MCRNN-Att-MT model and the GCRNN-Att-MT model indicates that the performance of the ML-LoBCoD-NET for extraction feature was better than that of the CNN network.

The performance index of the competition ranking is F1 score, and ER is the reference index. The value of ER in

TABLE 5. The F1 score and error rate of different mean teacher model systems.

System	Development set		Evaluation set	
	F1 score(%)	ER	F1 score(%)	ER
baseline	14.06	1.54	10.80	1.77
MRNN-Att-MT	22.83	1.60	15.68	2.00
MCRNN-Att-MT	20.35	1.47	14.56	1.71
GCRNN-Att-MT	19.34	2.16	14.73	1.89

TABLE 6. The F1 score and error rate of the proposed model with mean teacher model and without mean teacher model systems.

System	Development set		Evaluation set	
	F1 score(%)	ER	F1 score(%)	ER
baseline	14.06	1.54	10.80	1.77
MRNN-Att	16.24	2.55	10.78	2.13
MRNN-Att-MT	22.83	1.60	15.68	2.00
MCRNN-Att	16.01	2.29	11.23	2.76
MCRNN-Att-MT	20.35	1.47	14.56	1.71

our experiments being on par with baseline in Table 4 and Table 5. We mention two differences between the proposed model and the other two leading methods. Firstly, the number of parameters in the proposed approach does not grow with the depth of the model. Secondly, sound event detection methods based on traditional deep learning almost employ batch normalization operations which is known to improve the performance and convergence rate of the trained model. As our presented method relies only on the CSC prior, we did not include such batch normalization operators.

The F1 score and error rate (ER) of the proposed model with mean teacher model and without mean teacher model systems are given in Table 6. The F1 score of the MRNN-Att-MT model on the development dataset was 6.59% higher than that of the MRNN-Att model, and the F1 score of the MCRNN-Att-MT model was 4.34% higher than the MCRNN-Att model. The F1 score of the MRNN-Att-MT model on the evaluation dataset was 4.9% higher than the MRNN-Att model. The F1 score of the MCRNN-Att-MT model was 3.33% higher than the MCRNN-Att model. These results indicate that the mean teacher model can promote the sound event detection effect, thus improving the F1 score of the development set and evaluation set. The error rate of MRNN-Att-MT was 0.95% lower than MRNN-Att; the error rate of MCRNN-Att-MT was 0.82% lower than MCRNN-Att, which indicates that the proposed model with mean teacher model is better than the proposed model without mean teacher model.

The F1 scores of ten classes audio events using four systems without mean teacher model are given in Table 7. The F1 score of the MRNN-Att model was significantly better than the baseline system in terms of ''cat'', ''dog'', and ''running water'' class, and was significantly better than the GCRNN-Att in terms of ''blender'',''cat'' and ''dog'' class. The F1 score of the MCRNN-Att model was significantly better than the baseline system in terms of ''ringing'', ''dog'', and ''running water'' class, and was significantly better than

TABLE 7. The F1 scores of ten classes audio events using four systems without mean teacher model.

Event	Development				Evaluation			
	baseline	MRNN-Att	MCRNN-Att	GCRNN-Att	baseline	MRNN-Att	MCRNN-Att	GCRNN-Att
Alarm/bell/ringing	3.9	25.8	30.2	29.4	—	23.0	16.1	25.6
Blender	15.4	16.7	21.2	13.1	—	13.5	16.5	9.2
Cat	0	40.3	0.9	17.9	—	14.3	0.8	8.0
Dishes	0	2.0	3.4	2.7	—	3.1	6.0	2.7
Dog	0	14.7	6.5	7.3	—	9.3	3.5	4.0
Electric shaver/toothbrush	32.4	5.4	26.4	7.4	—	0.0	10.3	6.2
Frying	31.0	7.0	9.8	3.7	—	6.4	9.5	3.3
Running water	11.4	20.6	17.3	15.7	—	10.4	3.9	5.5
Speech	0	7.7	9.6	39.7	—	9.8	9.3	32.2
Vacuum cleaner	46.5	22.2	34.8	17.5	—	17.9	36.5	12.9

TABLE 8. The F1 scores of ten classes audio events using four systems with mean teacher model.

Event	Development				Evaluation			
	Baseline	MRNN-Att-MT	MCRNN-Att -MT	GCRNN-Att -MT	Baseline	MRNN-Att -MT	MCRNN-Att -MT	GCRNN-Att -MT
Alarm/bell/ringing	3.9	35.4	35.2	42.9	—	38.2	41.5	35.4
Blender	15.4	14.5	24.1	21.1	—	25.4	18.1	24.2
Cat	0	39.1	1.7	0.6	—	11.0	2.2	2.0
Dishes	0	4.7	9.8	3.2	—	3.3	2.7	5.7
Dog	0	25.7	24.2	12.8	—	13.5	16.8	14.5
Electric shaver/ toothbrush	32.4	27.6	9.5	9.7	—	9.7	3.1	6.3
Frying	31.0	23.3	22.2	21.0	—	14.7	11.7	10.6
Running water	11.4	11.7	9.9	15.8	—	3.3	2.8	8.5
Speech	0	15.8	36.1	29.4	—	18.7	27.7	31.0
Vacuum cleaner	46.5	30.6	30.8	38.4	—	18.9	19.0	5.7

the GCRNN-Att in terms of ''blender'',''Electric shaver'' and ''Vacuum cleaner'' class.

The F1 scores of ten classes audio events using four systems with mean teacher model are given in Table 8. The F1 score of the MRNN-Att-MT model was significantly better than the baseline system in terms of ''ringing'', ''cat'', and ''dog'' class, and was significantly better than the GCRNN-Att-MT in terms of ''ringing'', ''cat'', and ''dog'' class. The F1 score of the MCRNN-Att-MT model was significantly better than the baseline system in ''ringing'' and ''dog'' class, and was significantly better than the GCRNN-Att-MT in ''dishes'',''dog'' and ''speech'' class.

VII. CONCLUSION

The MRNN-Att network for weakly-labeled sound event detection task is proposed in this paper. The CNN pooling operation has the disadvantage of losing the location information of the target object. We don't use the pooling operation, retain ReLU and convolution operation, and use the strong constraints of the ML-CSC model. The MRNN-Att network based on the ML-LoBCoD-NET which is driven by the ML-LoBCoD algorithm. The ML-LoBCoD-NET shows the feature extraction ability different from the CNN for weakly-supervised sound event detection task.

Furthermore, the MRNN-Att-MT and the MCRNN-Att-MT model, the two mean teacher models, are proposed to solve the semi-supervised sound event detection problem. The MRNN-Att and the MCRNN-Att network are selected as the student model in the mean teacher model, respectively.

The proposed models were tested on the DCASE2018 Task 4 dataset. The results of these experiments showed that the F1 score of the proposed MRNN-Att-MT model and the MCRNN-Att-MT model were superior to the F1 score of the baseline and GCRNN-Att network for sound event detection. Furthermore, the F1 score of the MRNN-Att-MT model was superior to the F1 score of the GCRNN-Att-MT model. The F1 score of the MRNN-Att model and the MCRNN-Att model were superior to the F1 score of the baseline system. Adding an attention network can improve the performance of sound event detection. The sound event detection effects of the MRNN-Att model and the MCRNN-Att model were better than that of the GCRNN-Att model. These results indicate the ML-LoBCoD-NET shows the feature extraction ability different from the CNN for sound event detection task, and the proposed MRNN-Att network can be used in sound event detection task and is superior to the baseline system.

There is still a lot of room for improvement in the MRNN-Att network, such as adding the different attention networks or data augmentation methods. The MRNN-Att network is also used for acoustic scene classification and audio tagging.

ACKNOWLEDGMENT

The authors thank all anonymous reviewers for their effort and suggestions to improve this paper.

REFERENCES

[1] T. Hayashi, S. Watanabe, T. Toda, T. Hori, J. Le Roux, and K. Takeda, "BLSTM-HMM hybrid system combined with sound activity detection network for polyphonic sound event detection," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, New Orleans, LA, USA, Mar. 2017, pp. 766–770, doi: 10.1109/ICASSP.2017.7952259.

[2] A. Mesaros, T. Heittola, and A. Eronen, "Acoustic event detection in real life recordings," in *Proc. 18th Eur. Signal Process. Conf.*, Aalborg, Denmark, 2010, pp. 1267–1271.

[3] A. Temko and C. Nadeu, "Acoustic event detection in meeting-room environments," *Pattern Recognit. Lett.*, vol. 30, no. 14, pp. 1281–1288, Oct. 2009, doi: 10.1016/j.patrec.2009.06.009.

[4] E. Cakir, T. Heittola, H. Huttunen, and T. Virtanen, "Polyphonic sound event detection using multi label deep neural networks," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Killarney, Ireland, Jul. 2015, pp. 1–7, doi: 10.1109/IJCNN.2015.7280624.

[5] A. Mesaros, A. Diment, and B. Elizalde, "Sound event detection in the DCASE 2017 challenge," *IEEE Trans. Audio, Speech, Language Process.*, vol. 27, no. 6, pp. 992–1006, Oct. 2017, doi: 10.1109/WASPAA.2017.8169985.

[6] K. Feroze and A. R. Maud, "Sound event detection in real life audio using perceptual linear predictive feature with neural network," in *Proc. 15th Int. Bhurban Conf. Appl. Sci. Technol.(IBCAST)*, Islamabad, Pakistan Jan. 2018, pp. 377–382, doi: 10.1109/IBCAST.2018.8312252.

[7] A. Dang, T. H. Vu, and J.-C. Wang, "A survey of deep learning for polyphonic sound event detection," in *Proc. Int. Conf. Orange technol. (ICOT)*, Singapore, Dec. 2017, pp. 75–78, doi: 10.1109/ICOT.2017.8336092.

[8] T. Iqbal, Y. Xu, Q. Kong, and W. Wang, "Capsule routing for sound event detection," in *Proc. 26th Eur. Signal Process. Conf. (EUSIPCO)*, Rome, Italy, Sep. 2018, pp. 2255–2259, doi: 10.23919/EUSIPCO.2018.8553198.

[9] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic routing between capsules," in *Proc. Neural Inf. Process. Syst.*, Long Beach, CA, USA, 2017, pp. 3856–3866.

[10] Z. Qin, U. George, F. Yu, C. Liu, X. Chen, and U. Clarkson, "How convolutional neural networks see the world—A survey of convolutional neural network visualization methods," *Math. Found. Comput.*, vol. 1, no. 2, pp. 149–180, 2018, doi: 10.3934/mfc.2018008.

[11] M. A. Ahmad, A. Teredesai, and C. Eckert, "Interpretable machine learning in healthcare," in *Proc. IEEE Int. Conf. Healthcare Informat. (ICHI)*, New York, NY, USA, Jun. 2018, 447–447, doi: 10.1109/ICHI.2018.00095.

[12] I. Yong Chun, Z. Huang, H. Lim, and J. A. Fessler, "Momentum-net: Fast and convergent iterative neural network for inverse problems," 2019, *arXiv:1907.11818*. [Online]. Available: http://arxiv.org/abs/1907.11818

[13] K. Gregor and Y. LeCun, "Learning fast approximations of sparse coding," in *Proc. 27th Int. Conf. Mach. Learn.*, Haifa, Israel, 2010, pp. 399–406.

[14] M. Borgerding, P. Schniter, and S. Rangan, "AMP-inspired deep networks for sparse linear inverse problems," *IEEE Trans. Signal Process.*, vol. 65, no. 16, pp. 4293–4308, Aug. 2017, doi: 10.1109/TSP.2017.2708040.

[15] D. Ito, S. Takabe, and T. Wadayama, "Trainable ISTA for sparse signal recovery," *IEEE Trans. Signal Process.*, vol. 67, no. 12, pp. 3113–3125, Jun. 2019, doi: 10.1109/TSP.2019.2912879.

[16] D. Simon and M. Elad, "Rethinking the CSC model for natural images," in *Proc. 33rd Conf. Neural Inf. Process. Syst. (NIPS)*, Vancouver, BC, Canada, 2019, pp. 2771–2281.

[17] V. Papyan, Y. Romano, M. Elad, and J. Sulam, "Convolutional dictionary learning via local processing," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 2017, pp. 5296–5304, doi: 10.1109/ICCV.2017.566.

[18] X. Zheng, I. Yong Chun, Z. Li, Y. Long, and J. A. Fessler, "Sparse-view X-Ray CT reconstruction using ℓ_1 prior with learned transform," 2017, *arXiv:1711.00905*. [Online]. Available: http://arxiv.org/abs/1711.00905

[19] J. Sulam, V. Papyan, Y. Romano, and M. Elad, "Multi-layer convolutional sparse modeling: Pursuit and dictionary learning," *IEEE Trans. Signal Process.*, vol. 66, no. 15, pp. 4090–4104, Aug. 2018, doi: 10.1109/TSP.2018.2846226.

[20] E. Zisselman, J. Sulam, and M. Elad, "A local block coordinate descent algorithm for the convolutional sparse coding model," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 8208–8217.

[21] J. Sulam, A. Aberdam, A. Beck, and M. Elad, "On multi-layer basis pursuit, efficient algorithms and convolutional neural networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, to be published, doi: 10.1109/TPAMI.2019.2904255.

[22] Y. Xu, Q. Kong, Q. Huang, W. Wang, and M. D. Plumbley, "Attention and localization based on a deep convolutional recurrent model for weakly supervised audio tagging," in *Proc. Interspeech*, Aug. 2017, pp. 3083–3087, doi: 10.21437/Interspeech.2017-486.

[23] A. Kumar and B. Raj, "Audio event and scene recognition: A unified approach using strongly and weakly labeled data," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Anchorage, AK, USA, May 2017, pp. 3475–3482, doi: 10.1109/IJCNN.2017.7966293.

[24] Q. Kong, Y. Xu, W. Wang, and M. D. Plumbley, "A joint separation-classification model for sound event detection of weakly labelled data," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Calgary, AB, Canada Apr. 2018, pp. 321–325, doi: 10.1109/ICASSP.2018.8462448.

[25] Y. Xu, Q. Kong, W. Wang, and M. D. Plumbley, "Large-scale weakly supervised audio classification using gated convolutional neural network," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Calgary, AB, Canada, Apr. 2018, pp. 121–125, doi: 10.1109/ICASSP.2018.8461975.

[26] B. McFee, J. Salamon, and J. P. Bello, "Adaptive pooling operators for weakly labeled sound event detection," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 11, pp. 2180–2193, Nov. 2018, doi: 10.1109/TASLP.2018.2858559.

[27] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Proc. ICLR (Workshop)*, Long Beach, CA, USA, 2017, pp. 1195–1204.

[28] G. Parascandolo, H. Huttunen, and T. Virtanen, "Recurrent neural networks for polyphonic sound event detection in real life recordings," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Shanghai, China, Mar. 2016, pp. 6440–6444, doi: 10.1109/ICASSP.2016.7472917.

[29] L. Lin, X. Wang, H. Liu, and Y. Qian, "Guided learning convolution system for DCASE 2019 task 4," in *Proc. Detection Classification Acoustic Scenes Events Workshop (DCASE)*, Oct. 2019, pp. 1–5, [Online]. Available: http://dcase.community/challenge2018/task-large-scale-weakly-labeled-semi-supervised-sound-event-detection

[30] N. Turpault, R. Serizel, J. Salamon, and A. P. Shah, "Sound event detection in domestic environments with weakly labeled data and soundscape synthesis," in *Proc. Detection Classification Acoustic Scenes Events Workshop (DCASE)*, Oct. 2019, pp. 1–4, [Online]. Available: https://hal.inria.fr/hal-02160855

[31] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio Set: An ontology and human-labeled dataset for audio events," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, New Orleans, LA, USA, Mar. 2017, pp. 776–780, doi: 10.1109/ICASSP.2017.7952261.

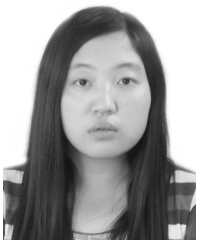
[32] Q. Kong, Y. Xu, I. Sobieraj, W. Wang, and M. D. Plumbley, "Sound event detection and time–frequency segmentation from weakly labelled data," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 4, pp. 777–787, Apr. 2019, doi: 10.1109/TASLP.2019.2895254.

[33] R. Serizel, N. Turpault, and H. Eghbal-Zadeh, A. Parag Shah, "Large-scale weakly labeled semi-supervised sound event detection in domestic environments," in *Proc. Detection Classification Acoustic Scenes Events Workshop (DCASE)*, Nov. 2018, pp. 19–23. [Online]. Available: https://hal.inria.fr/hal-01850270

JINJIA WANG (Member, IEEE) was born in 1978. He is currently pursuing the Ph.D. degree. He is also a Professor with the School of Information Science and Engineering, Yanshan University. His research interests include pattern recognition and signal processing. He is responsible for directing this work.

JING XIA was born in 1993. She is currently pursuing the master's degree in communication engineering. She studied at Yanshan University, Qinhuangdao, Hebei. Her research interests include sensor signal processing and audio classification. She completed the main thesis writing, revision, and experimental work.

QIAN YANG was born in 1994. She is currently pursuing the master's degree in communication engineering. She studied at Yanshan University, Qinhuangdao, Hebei. Her research interest is in sensor signal processing. She mainly completes the paper writing of the mean teacher model section and the subsequent modification.



YUZHEN ZHANG was born in 1994. She is currently pursuing the master's degree in communication engineering. She studied at Yanshan University, Qinhuangdao, Hebei. Her research interest is in sensor signal processing. She contributed to the first version of ML-LoBCoD section.