Spectrogram based multi-task audio classification

Yuni Zeng[1] · Hua Mao[1] · Dezhong Peng[1] · Zhang Yi[1]

Abstract Audio classification is regarded as a great challenge in pattern recognition. Although audio classification tasks are always treated as independent tasks, tasks are essentially related to each other such as speakers' accent and speakers' identification. In this paper, we propose a Deep Neural Network (DNN)-based multi-task model that exploits such relationships and deals with multiple audio classification tasks simultaneously. We term our model as the gated Residual Networks (GResNets) model since it integrates Deep Residual Networks (ResNets) with a gate mechanism, which extract better representations between tasks compared with Convolutional Neural Networks (CNNs). Specifically, two multiplied convolutional layers are used to replace two feed-forward convolution layers in the ResNets. We tested our model on multiple audio classification tasks and found that our multi-task model achieves higher accuracy than task-specific models which train the models separately.

Keywords Multi-task learning · Convolutional neural networks · Deep residual networks · Audio classification

1 Introduction

Sound provides us with rich information about its producer and environment. As the human auditory system is able to segregate and identify complex sounds, we can imagine that a

This work was supported by the National Natural Science Foundation of China [grant numbers 61402306, 61432012]

✉ Hua Mao
huamao@scu.edu.cn

[1] Machine Intelligence Laboratory, College of Computer Science, Sichuan University, Chengdu, 610065, People's Republic of China

machine that can perform similar functions would be very useful in applications such as speech recognition in noisy backgrounds [28]. Audio classification is an important aspect of pattern recognition and has been widely used in professional media applications and entertainment. Different audio classification tasks—such as speech-music discrimination, audio emotion recognition, accent recognition, and music retrieving information—have driven successful applications in recent years [9, 21, 28].

Most methods of audio classification treat tasks separately such as accent classification [13], emotion recognition [5], speaker identification [22] and so on. However, some tasks are closely related. For example, while accent recognition and speaker identification are always regarded as two individual classification tasks, in most situations when the speaker is confirmed, the accent is determined and unchangeable. In this case, accent recognition and speaker identification are related. Relational information of Alzheimer's disease is exploited for feature selection to improve classification accuracy in [35]. Here, we sought to use relationship to simultaneously predict different tasks, and hypothesized that this would increase classification accuracy.

Multi-task Learning (MTL) is a subfield of machine learning in which multiple tasks are solved simultaneously. Thus, MTL can exploit the intrinsic relationships among related tasks [7, 37, 38]. The goal of MTL is to improve generalization performance by training all tasks at same time from a shared representation. Because of the training way of MTL, tasks can benefit from what are learned for other tasks [4]. Borrowing the idea from MTL, we designed a multi-task model for audio classification tasks. Our model exploits the relationships found in audio and deals with multiple audio classification tasks jointly. We find out that our model help to improve the accuracy of tasks comparing with the state-of-the-art result obtained by task-specific models [29].

The inputs to our model are audio spectrograms, the visual representations of audio. Spectrograms are very detailed and accurate images of audio that have been widely used in audio classification tasks [11, 12, 28]. Deep Neural Networks (DNNs) are very good at abstracting data, and have been used with great success in fields such as speech recognition [1, 12] and image recognition [15, 30]. Specifically, we focus on Convolutional Neural Networks (CNNs) in this paper. CNNs can efficiently exploit invariance presented in spectrogram [27] for their convolutional and pooling operations.

Thus, we propose a new CNNs-based model for learning a shared representation among all the spectrograms from different classification tasks. Called the Gated Residual Neural Networks (GResNets) model, it is a variation of a deep Residual Networks (ResNets) model with the addition of a gate mechanism. Because of the notorious gradient vanishing/exploding problem in learning DNNs, ResNets use the linear connection of Long Short-Term Memory (LSTM) [16] to ease the training of very deep CNNs [15]. Another way for the LSTM to solve this problem is to use the gate mechanism [16]. In our GResNets, we combine both mechanisms to get better representations. Because of the strong ability to learn invariance presented in spectrograms [27], CNNs are used as the component of our GResNets like ResNets. Finally, the abstracted features are used for audio classification in the last softmax layer. The experimental results demonstrated that multi-task models for related audio classification tasks outperform the task-specific models of each task.

The rest of this paper is organized as follows. We begin with related works about audio classification in Section 2. Next, we review the MTL, spectrogram, ResNets and gate mechanism in Section 3. Section 4 introduce the structure of our proposed model. Then, in Section 5 we use the proposed model for different multiple audio classification tasks. Finally, we conclude our work in Section 6.

2 Related works

Many methods in machine learning are adapted for the single audio classification task. Support Vector Machine (SVM) was used to classify two English accents [25] and eight emotions with continuous wavelet-transform features [29], respectively. In another study, a Gaussian Mixture Model (GMM) was trained for an accent-classification task by fusing several acoustic and text-language subsystems [13]. In [5], SVM, principal component analysis and artificial neural network were combined to classify six emotions from speech.

Deep learning methods have also been widely used for audio classification tasks. A CNNs-based neural network called convolutional deep belief network was developed for several individual audio classification tasks, such as speaker identification, speaker gender classification, and music genre classification [21]. In [11], a CNNs based model was trained from fixed-length spectrogram features and single-label audio data for various predominant instrument recognition in polyphonic music. In [22], a CNNs and Gated Recurrent Unit (GRU)-based neural network was proposed for speaker identification and verification. Additionally, LSTM was used in a hybrid emotion inference model that was proposed for inferring user emotion in a real-world voice-dialogue application, and a recurrent autoencoder was proposed to pre-train the LSTM to improve accuracy [32]. Further, GMM and DNNs were combined to identify distant accents in reverberant environments [26]. The authors found that this combination of classifiers outperformed the individual GMM and DNNs classifiers.

All the methods above focus on the single audio classification task. However, few studies have focused on the multiple audio classification tasks through MTL method. Two reports concentrated on emotion recognition from spoken language and song at the same time [31, 33]. Especially [33] showed that MTL-SVM models have significantly better performance for audio emotion recognition than task-specific SVM models. Additionally, the authors have also used an MTL approach on SVM for cross-corpus acoustic recognition of emotion in speaking and singingk [34].

Many recent deep learning approaches also have used MTL. Deep Relationship Networks [24] were proposed to learn the relationshio between tasks. The model shares convolutional layers, while learning task-specific fully-connected layers. [14] introduced a joint many-task model to solve complex task in the field of nature language processing.

3 Preliminaries

Our multi-task model abstract the shared features from spectrograms and GResNets is a variation of CNN-based ResNets. Thus, here we review basic information related to MTL, spectrograms, ResNets and gate mechanism.

3.1 Multi-task learning

MTL is a scope of machine learning where multiple learning tasks are solved at the same times. The goal of MTL is to improve the prediction accuracy of multiple classification tasks by learning them jointly. Models which combine neural network method with MTL use a shared hidden layers on all tasks [4].

Mathematically, given a data set $S = \{(x_1, Y_1), \cdots, (x_N, Y_N)\}$ where N is the number of sample, x_i denotes the i^{th} sample and $Y_i = (y_i^1, y_i^2 \cdots, y_i^T)$ represents the set of labels. Let

N_T indicate the set of tasks and T be the number of tasks. For each task $t \in N_T$, we select p samples where $p < N$ samples from audio set as the trainset $\{(x_1, y_1^t), \cdots, (x_p, y_p^t)\}$. The MTL aims to learn a fitted model F to get accurate labels for each input:

$$(y_i^1, y_i^2, ..., y_i^T) = F(\theta, x_i), \tag{1}$$

where $i \leq p$. θ denotes the parameters of the multi-task model. We assume that the cost function is $J(\cdot)$, and the parameters of the multi-task model are learned by minimizing the following formula:

$$\sum_t^T J(x_t, Y_t, \theta). \tag{2}$$

3.2 Spectrogram

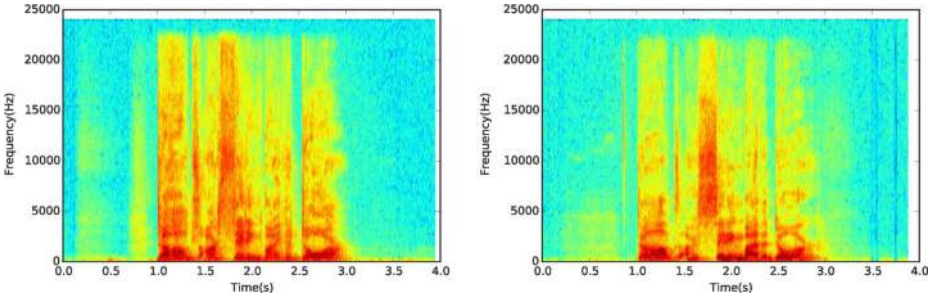
A spectrogram is regarded as a very detailed and accurate representation of audio information. A common spectrogram is an image where one axis represents time, the other axis is frequency and the color of each point indicates the amplitude of those points. Thus, a spectrogram shows amplitude changes for every frequency component in the signal. Figure 1 gives an example of audio spectrograms that contain different emotions. From the spectrograms, we can observe that the amplitude of happy and angry emotions diverge in the $5000Hz$ to $15000Hz$ frequency range.

3.3 Deep residual networks

Compared fully-connected neural networks, CNNs [20] are also designed to recognize pattern directly from pixel images but with less parameters. CNNs have led to a series of significant achievements in audio classification. Using sparse connectivity and shared weights ease network optimization by reducing the number of parameters and the risk of overfitting [18]. The basic layers of CNNs are convolutional layer and pooling layer. Figure 2 shows a typical architecture for CNNs.

A convolutional layer is composed of several kernels and aims to get the feature maps [10]. To generate output feature maps, convolutional layer input maps are convoluted with learnable kernels and the results are transformed by a nonlinear activation function.

Suppose l denotes the l^{th} layer of CNNs and k^l denotes the parameters of the kernel and b^l is the bias parameters at the l^{th} layer. The input and output of the l^{th} layer are



(a) angry emotion (b) happy emotion

Fig. 1 Spectrograms for happy and angry emotion. (Better viewed in color)

defined by a^{l-1} and a^l respectively. $f(\cdot)$ denotes an activation function. The convolutional operation [3] can be formulated as

$$a_j^l = f(\sum_{i \in M_j} a_i^{l-1} * k_{ij}^l + b_j^l), \tag{3}$$

where M_j denotes a selection of input maps. The convolutional operation is not only widely used in CNNs but also in sparse coding [36].

A pooling layer is usually placed after the convolutional layer to achieve invariant representation and reduces the number of parameters [8]. The number of feature maps for a pooling layer is the same as the number of feature maps in the previous convolutional layer. There are two typical pooling operations: average pooling and max pooling [2]. The computation of the pooling layer is formulated as follows

$$a_j^l = f(\beta_j^l down(a_j^{l-1}) + b_j^l), \tag{4}$$

where $down(\cdot)$ represents a sub-sampling function. and β is the multiplicative bias that is given to one output map [3].

ResNets was the winning model of the ILSVRC 2015 ImageNet challenge and has been proposed to ease the training of very deep CNNs [15]. Deep ResNets are composed of several stacked residual blocks. Figure 3 illustrates a residual block where weight layers denote convolutional layers and a shortcut connection is used for identifying mapping.

For a block of ResNets, let the input for the blocks be x,

$$y = F(x, W) + x, \tag{5}$$

where y is the output of the building block, W indicates the weights of the residual block and the function $F(\cdot)$ is the output of two convolutional layers.

The dimensionality of x and $F(x, W)$ must be the same for (5) to be valid. To match dimensions, the general solution is to use a linear projection weight matrix W_s as:

$$y = F(x, W) + W_s x. \tag{6}$$

3.4 Gate mechanism

The limitation by using the deep neural networks is the gradient vanishing/exploding problem. The gate mechanism proposed in LSTM [16] can be viewed as the milestone for solving the gradient problem. The input/output gates are used to control how much information should be kept in the cell. The forget gate is used to control the values kept in the

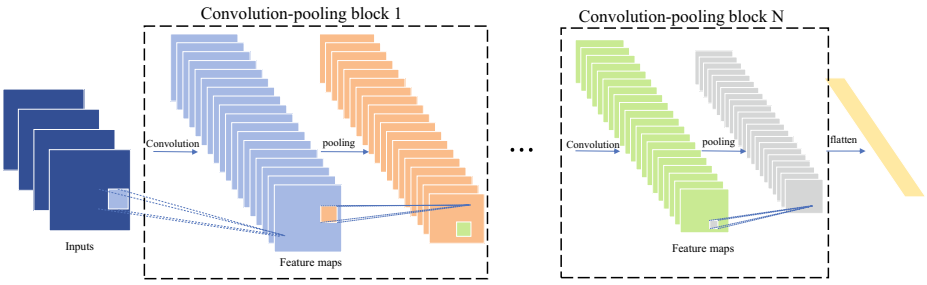


Fig. 2 One illustrative of one CNN architecture

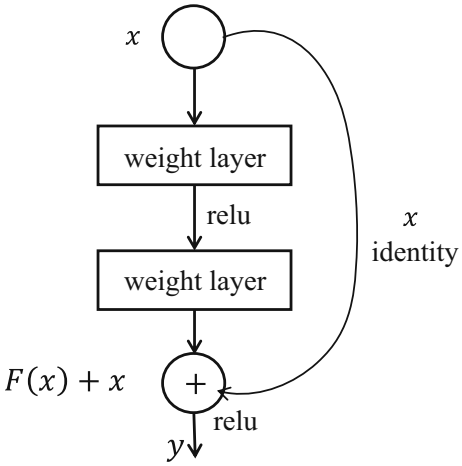


Fig. 3 a residual block

memory cells of LSTM which let LSTM capable to deal with continual learning tasks. The side-effect of gradient problem can be reduced by these mechanism. Deeper network can be trained in practice.The gates support a more flexible way to control error flow for handing the gradient vanishing/exploding problem.

Figure 4 illustrates the gate mechanism intuitively. The gate mechanism can be formulated as $y = f_1 \cdot f_2$ where f_1 and f_2 denote the activation function and gate function, respectively. $\cdot$ is the element-wise multiplication. If $f_2 = 0$, the gate is fully closed and if $f_2 = 1$, the gate is fully opened.

4 The proposed model

We proposed a new CNN-based architecture to extract the shared feature of all tasks. The details are described in this section. We begin by describing the fundamental block of GResNets and then introduce our proposed multi-task model.

4.1 The gated residual networks blocks

The gradient vanishing/exploding problem is a known limitation for deep learning. LSTM [16] first proposed to solve this problem by combining a gate mechanism with linear

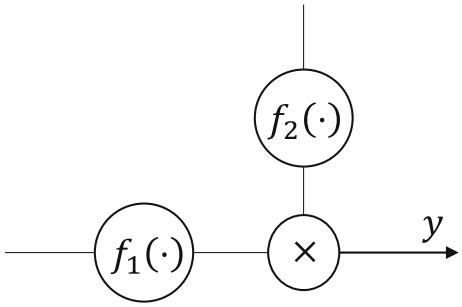


Fig. 4 The illustration of a gate

connections. ResNets were proposed to train very deep feedforward neural networks by appending linear connections to solve the gradient vanishing/exploding problem, but without a gate mechanism. [15].

Inspired by these two mechanisms, we propose GResNets, which is composed of the basic blocks shown in Fig. 5. The weight layers $F_1(\cdot)$, $F_2(\cdot)$ indicate convolutional layers, each of which is followed by a normalization method [17]. Let x be the input of a block. We define $*$ as element-wise multiplication. Thus, the output y of one block can be formulated as

$$y = F_1(x, W_1) * F_2(x, W_2) + x, \tag{7}$$

where W_1 and W_2 indicate the weights of the convolutional layers. Usually, the activation functions in convolutional layers are the $sigmoid$ function, the $tanh$ function, or the $ReLu$ function. We compare different activation functions in Section 5.3. Like ResNets, we also use a linear projection weight M to match the dimensions of $F_1(\cdot) * F_2(\cdot)$ and x:

$$y = F_1(x, W_1) * F_2(x, W_2) + Mx. \tag{8}$$

4.2 The proposed model for audio classification

The basic idea of MTL is to share parameters between related tasks. Our multi-task model is a neural network with different number of the softmax classifiers. Let N_T indicate the set of tasks and T be the number of tasks. The classification layer of the multi-task model includes T softmax classifiers. For task-specific model, $T = 1$. Figure 6 illustrates an example of the MTL model which is used for two audio classification tasks. The first convolution and pooling layers are used to reduce the dimension of the input spectrograms and the number of parameters. The next several GResNets blocks and a full connected layer are stacked to get the shared representation between the two tasks. Then, the extracted features are used in the softmax layer to generate predictions for each task.

5 Experiments

To evaluate our GResNets and the MTL method could improve the accuracy of each tasks, we designed experiments on tasks with different relationship. In the first part of experiments, the GResNets model was used for two tasks of The Ryerson audio-visual database of

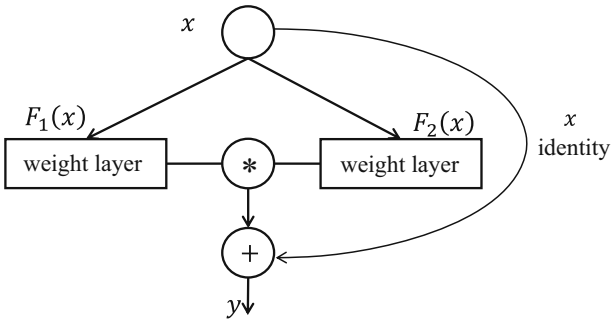


Fig. 5 a block from our proposed Gated ResNets

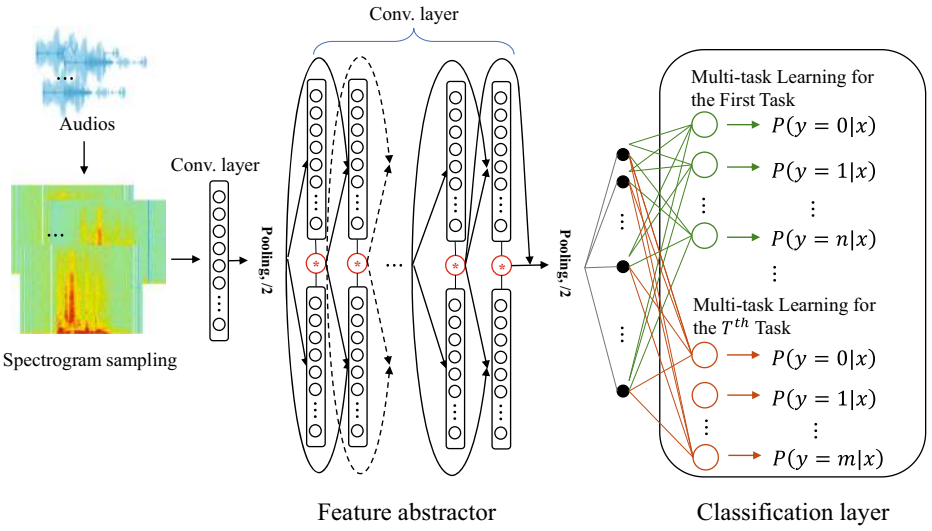


Fig. 6 The structure of our proposed model

emotional speech and song (RAVDESS) [23]. One task is to identify eight kinds of emotions, another one is to discriminate song and speech. There is no direct connection between those two tasks. We regard the first task as the main task and to find whether what the second task learned would help the first task or not. We also compared the performance of the GResNets model with different numbers of blocks, activation functions, and other CNN structures.

The Voice Cloning Toolkit (VCTK)[1] data set was used in the second part of experiments. There also are two tasks, accent recognition and speaker identification. But those two tasks have a kind of relationship like inclusion relation. In this part of experiments, we focus on comparing the multi-task model and task-specific models and using an evaluation metric (the confusion matrix) to represent the improvement of MTL.

5.1 Corpus

RAVDESS corpus RAVDESS consists of 24 actors (12 females and 12 males) speaking and singing with various emotions. The speaking set contains eight emotional expressions: neutral, calm, happy, sad, angry, fearful, disgust and surprised. The singing set includes six emotions: neutral, calm, happy, sad, angry, and fearful. All emotions (except neutral) are expressed at normal and strong levels of intensity. All audio files in this corpus are encoded as 16 bits, 48 kHz wav files. We use the MIRtoolbox [19] to generate 257 $\times$ 399 pixel spectrograms for each audio file. The RAVDESS dataset originally contains speech and song as two separated categories, so we mixed all audio while the first experiment performs to classify them. Figure 7 shows the spectrograms from the audio set for the eight categories of emotion in speech.

[1] http://homepages.inf.ed.ac.uk/jyamagis/page3/page58/page58.html

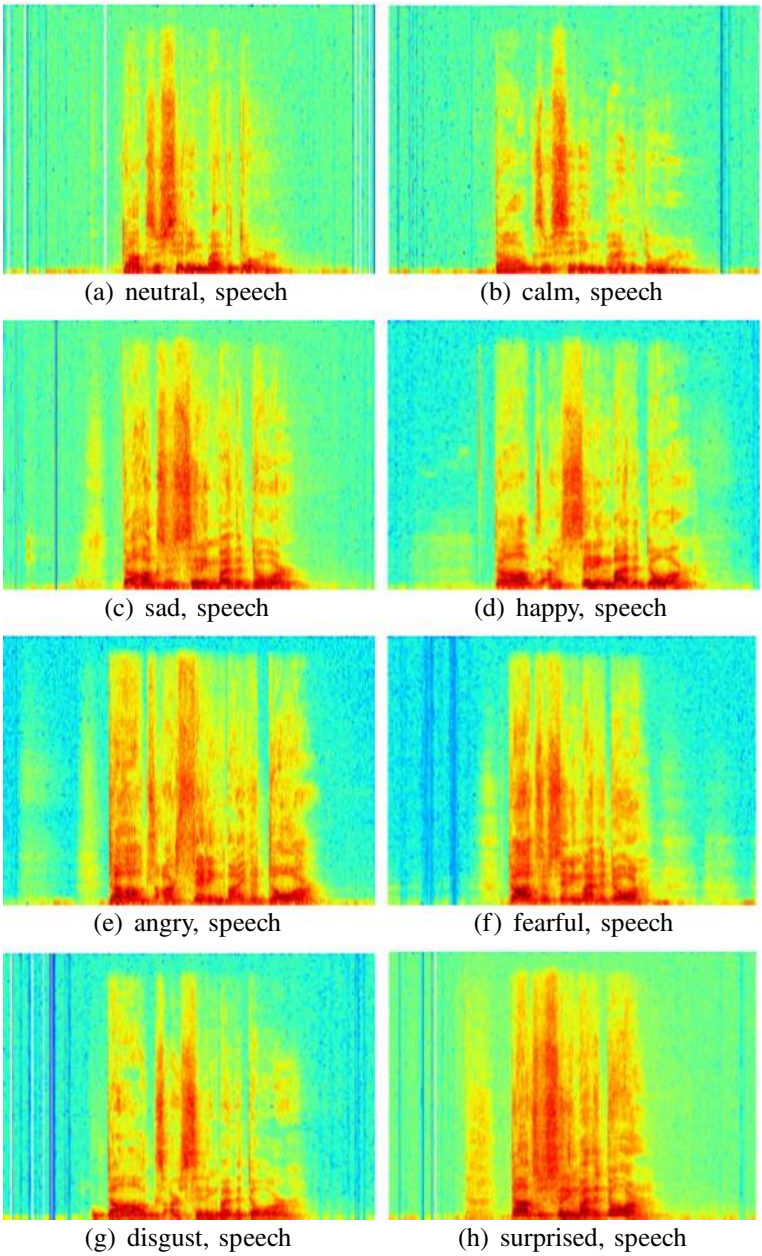
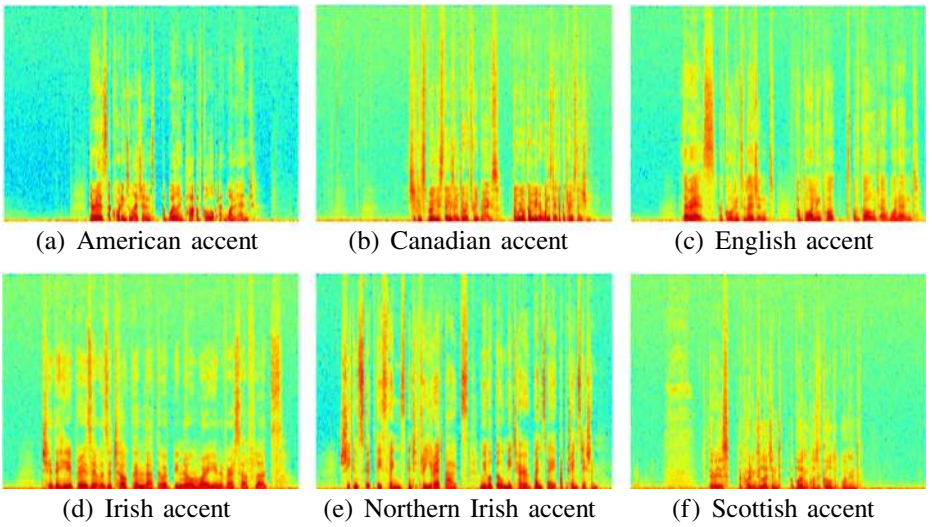


Fig. 7 Spectrograms of emotion in speech

VCTK corpus The Voice Cloning Toolkit (VCTK)[2] includes speech data uttered by 109 native English speakers in different accents. It includes American, Australian, Canadian,

[2] http://homepages.inf.ed.ac.uk/jyamagis/page3/page58/page58.html



(a) American accent (b) Canadian accent (c) English accent

(d) Irish accent (e) Northern Irish accent (f) Scottish accent

Fig. 8 Spectrograms features of American, Canadian, English, Irish, Northern Irish, and Scottish accents

English, Indian, Irish, NewZealand, Northern Irish, Scottish, South African and Welsh accents. All speech data in the VCTK was recorded at a sampling of 96 kHz and at 24 bits. All recordings were converted into 16 bits, down-sampled to 48 kHz based on STPK, and manually end-pointed. We generated 257×399 pixels spectrograms for each sentence. Figure 8, spectrogram features of American, Canadian, English, Irish, Northern Irish, and Scottish accents. Figure 9 illustrates the spectrograms for six speakers who have the same accent.

5.2 Experimental settings

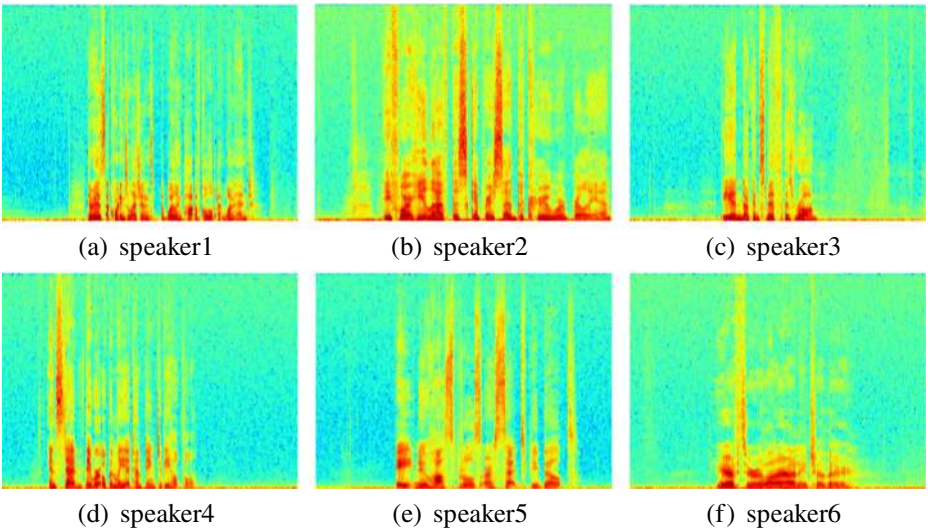
We extracted the spectrogram features with a 512-length hamming window with 50% overlap for each audio. The experiment was performed on an MXNet framework [6] with NVIDIA Tesla K40m GPU. The parameters of our multi-task model are illustrated in Fig 10. The underline parameters are changeable in different experiments. l denotes the number of GResNets blocks. n, m, p, q, and $num\_filters$ are the CNNs parameters. $n \times m$, $p \times q$ are the kernel sizes. $num\_filters$ represents the number of feature maps and t is the number of tasks. The learning rate is 0.001 during training and the cost function is cross-entropy loss.

To measure performance, we used two evaluation metrics. The first one was Unweighted Average Recall (UAR) [13]. In the binary class case ('Y' and 'NY'), it is defined as:

$$UAR = \frac{Recall(Y) + Recall(NY)}{2} \tag{9}$$

where $Recall(\cdot)$ is the recall result of the class. The second metric was accuracy (Acc).

$k-$fold cross-validation was used for estimating model. In $k-$fold cross-validation, the data set is randomly divided into k equal folds. For the k subsets, a single subset is retained for use in validation and the other subsets are used as training data. The average value of the k results is the report accuracy in our paper.



(a) speaker1 (b) speaker2 (c) speaker3

(d) speaker4 (e) speaker5 (f) speaker6

Fig. 9 Illustration spectrograms features of different speakers from American

5.3 Experiment 1: emotion recognition in speaking and singing

We used the model for emotion recognition in speaking and singing and compared it with the results for the model in [29], which use the common sentence 'Dogs are sitting by the door' from RAVDESS corpus at the strong intensity level. There are two tasks (1) emotion recognition and (2) speech vs. song discrimination. The main task is to classify 8 categories of emotion. The parameters were: $t = 2$, 5−fold cross-validation, $l = 7$ and $num\_filters$ was 32, 64, 64, 96, 96, 96 and 96 for each of the layers respectively.

First, we compared how well task-specific and our multi-task model recognize emotions. Table 1 shows the emotion classification results of the task-specific models and our multi-task model. SVM and Continuous Wavelet Transform (CWT) features [29] were used to classify eight categories of emotion and reported a maximum accuracy of 60.1%. Moreover, another baseline is provided based on a standard SVM classifier. In this experiment, $n \times m$ is 1×1, $p \times q$ is 3×3, f_1 is the $tanh$ function and f_2 is the $tanh$ function because of the functions used in [16]. The classification accuracy for song and speech was 97.31% using the multi-task model and the UAR was 97.37.

Next, we compared the results of GResNets with different number layers l from our multi-task model when $n \times m$ was 1×1, $p \times q$ was 3×3, f_1 was the $sigmoid$ function and

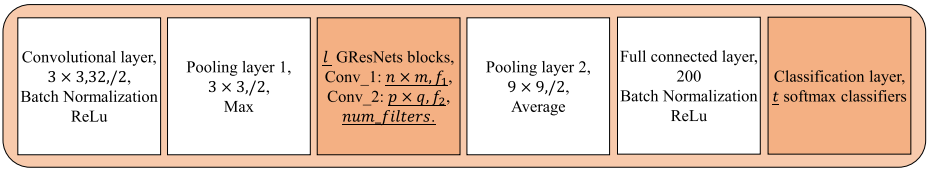


Fig. 10 Parameter settings of our multi-task model. $l, n, m, p, q, num\_filters, t$ and f_1, f_2 are changeable under different situations

Table 1 Emotion recognition results of different models where '$+S$' denotes task-specific model and '$+M$' denotes that model uses for multi-task

	$Acc(\%)$	$UAR(\%)$
SVM+S	48.01	48.66
CWT+SVM+S [29]	60.1	–
ResNet+S	53.30	50.33
GResNet+S	60.35	59.70
GResNet+M	64.48	64.52

f_2 was the $(4 \times sigmoid - 2)$ function. As shown in Fig. 11, the best performance occurred when $l = 7$ in our multi-task model.

Then, we compared the GResNets model with ResNets and CNNs models. To replace the GResNets blocks of our multi-task model, we used ResNets blocks and CNNs that contained the same convolutional layers. Thus, for ResNets blocks Fig. 3, the first weight layer is a convolutional layer with a 1×1 kernel when the second layer is with a 3×3 kernel. Table 2 shows the results for emotion recognition using different multi-task models.

Last, for a multi-task model based on GResNets, each GResNets block included two convolutional layers as explained in Section 4.1. We compared the different activation functions for two convolutional layers of a GResNet block. Tables 3 and 4 show the multi-task results in different activation functions with different kernels in convolutional layers.

From the results, following observations are obtained:

- Table 1 shows that the multi-task model outperforms task-specific models when classifying emotion from speech and song, indicating that emotion recognition benefits from MTL.

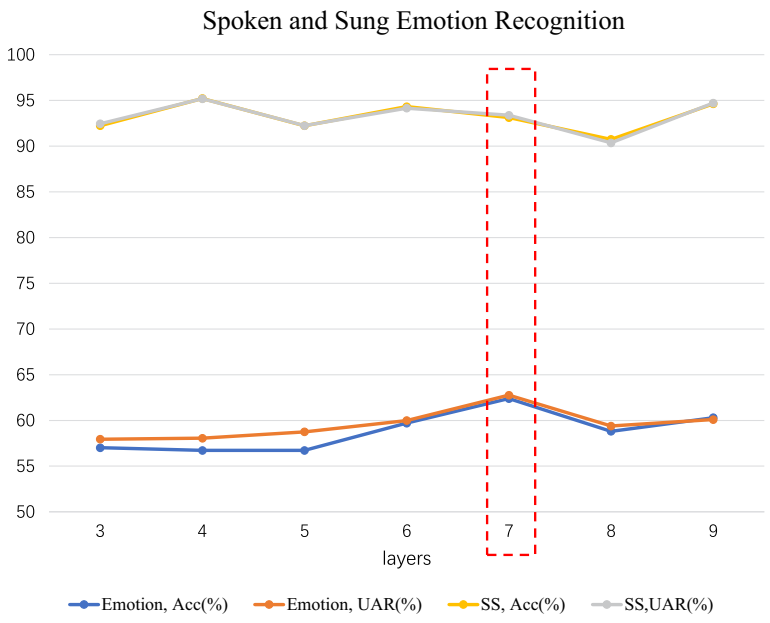


Fig. 11 Classification by using different number of feature abstract layer where 'SS' denotes speech-song discrimination

Table 2 Results of emotion recognition in our multi-task model based on CNNs, ResNets and gated ResNets

	Emotion		Speech and song	
	$Acc(\%)$	$UAR(\%)$	$Acc(\%)$	$UAR(\%)$
SVM+M	54.63	56.02	91.04	91.25
CNNs+M	53.73	54.8	92.24	92.08
ResNets+M	57.21	58.58	94.62	94.36
GResNets+M	64.48	64.52	97.31	97.37

- Table 2 summarizes the results of the multi-task model with different CNNs structures. The results indicate that under the same conditions, the model based on GResNets performs better than other CNN structures for emotion recognition.
- Tables 3 and 4 show that the best recognition occurred for spoken and sung emotion when the activation function in GResNets blocks of our multi-task model was the $tanh$ function. We think the reason is that the value range of the $tanh$ function is from minus to positive.

5.4 Experiment 2: accent and speaker recognition

Although accent recognition and speaker identification are two related tasks, they are always regarded as two individual tasks. Here, we aimed to determine whether our multi-task model is better than using the task-specific for accent recognition and speaker identification. For this experiment, we included data from the VCTK because of huge imbalance for each accent. We trained three GResNets models: (1) a GResNets model to classify six categories of accents: $American$, $Canadian$, $English$, $Irish$, $NorthernIrish$, and $Scottish$ accents; (2) a GResNets model for identifying 37 speakers classification; (3) a multi-task model with $t = 2$ and $l = 4$. In this experiment, $n \times m$ was 3×3, $p \times q$ was 3×3, f_1 was the $sigmoid$ function and f_2 was the $(4 \times sigmoid - 2)$ function. Their results are summarized in Table 5.

Table 3 Results for different activation functions f_1, f_2 ($n \times m$ is 1×1, $p \times q$ is 3×3)

Convolutional layers		Emotion		Speech and song	
f_1	f_2	$Acc(\%)$	$UAR(\%)$	$Acc(\%)$	$UAR(\%)$
sigmoid	4*sigmoid-2	62.39	62.76	93.13	93.38
Sigmoid	tanh	60.60	60.66	94.03	94.09
Sigmoid	sigmoid	55.22	56.72	91.04	91.50
Sigmoid	relu	60.00	61.90	94.03	94.09
Relu	sigmoid	60.00	59.99	91.94	92.34
Relu	tanh	54.63	57.04	94.03	93.90
Tanh	tanh	64.48	64.52	97.31	97.37
Tanh	sigmoid	57.91	56.65	92.54	91.87
4*sigmoid-2	sigmoid	58.21	56.03	90.75	90.95

Table 4 Results for different activation functions f_1, f_2 ($n \times m$ is 3×3, $p \times q$ is 3×3)

Convolutional layers		Emotion		Song and speech	
f_1	f_2	$Acc(\%)$	$UAR(\%)$	$Acc(\%)$	$UAR(\%)$
sigmoid	4*sigmoid-2	60.60	58.95	90.74	91.27
Sigmoid	tanh	60.59	58.57	93.73	94.20
Sigmoid	sigmoid	56.42	54.88	87.16	88.38
Relu	sigmoid	61.79	61.42	91.94	92.05
Tanh	tanh	65.97	66.90	94.33	94.37

Table 5 Speaker and accent recognition for task-specific and multi-tasks models

	GResNets+S		GResNets+M	
	$Acc(\%)$	$UAR(\%)$	$Acc(\%)$	$UAR(\%)$
Speaker identification	83.05	83.33	88.52	87.36
Accent recognition	89.67	89.82	92.44	92.57

Table 6 Confusion matrix for accent recognition using a task-specific model

Accents	American	Canadian	English	Irish	Northern Irish	Scottish
American	217	7	4	10	9	0
Canadian	4	217	5	1	6	1
English	7	7	206	30	3	4
Irish	6	1	10	226	12	1
Northern Irish	2	7	3	2	241	3
Scottish	0	4	4	0	0	221

Table 7 Confusion matrix for accent recognition using our multi-task model

Accents	American	Canadian	English	Irish	Northern Irish	Scottish
American	222	13	2	6	4	0
Canadian	1	224	1	0	7	1
English	4	5	228	10	3	7
Irish	4	3	16	230	2	1
Northern Irish	0	6	4	5	241	2
Scottish	1	3	0	0	1	224

For a clear comparison between multi-task and single-task learning, we report the confusion matrixes for accents using the task-specific and multi-task models (Tables 6 and 7).

Accent and speaker recognition tasks have certain relationship that once the speaker is fixed, his or her accent is also fixed. Our model takes advantage of this relationship to learning accent and speaker recognition tasks simultaneously. The experimental results also prove this. From tables, it's obvious that the multi-task model improves accuracy from 83.05% to 88.52% in speaker identification and from 89.67% to 92.44% in accent recognition. Thus, speaker identification and accent recognition, as two individual but related audio classification tasks can benefit from MTL.

6 Conclusions

In this paper, we proposed a multi-task model for audio classification that is based on GResNets and applied it to related audio classification tasks: (1) recognition emotion from speech and song; and (2) accents and speakers recognition. We evaluate our model among tasks with different relationship on multiple audio classification tasks. We found that related audio tasks recognition is in the scope of MTL and the experimental results show that recognition accuracy is better when using a multi-task model than multiple task-specific models. Thus, we perform efficient inference that different relationship audio classification tasks can benefit from MTL methods.

The model developed in this study has led to advancements in some audio classification tasks. While these algorithms have resulted in effective performance for recognizing eight emotions and six accents, we cannot make further claims regarding other related audio classification tasks. Thus, these findings should be confirmed in additional audio classification tasks in the future, such tasks with conflicts. Moreover, we are trying to give some mathematical justification about why our model works.

References

1. Amodei D, Anubhai R, Battenberg E, Case C, Casper J, Catanzaro B, Chen J, Chrzanowski M, Coates A, Diamos G, Elsen E, Engel J, Fan L, Fougner C, Hannun AY, Jun B, Han T, LeGresley P, Li X, Lin L, Narang S, Ng AY, Ozair S, Prenger R, Qian S, Raiman J, Satheesh S, Seetapun D, Sengupta S, Wang C, Yi W, Wang Z, Bo X, Xie Y, Yogatama D, Zhan J, Zhu Z (2016) Deep speech 2: End-to-end speech recognition in english and mandarin. In: Proceedings of the 33nd international conference on machine learning, pp 173–182
2. Boureau Y-L, Ponce J, LeCun Y (2010) A theoretical analysis of feature pooling in visual recognition. In: Proceedings of the 27th international conference on machine learning, pp 111–118
3. Bouvrie J (2006) Notes on convolutional neural networks. Neural Nets 2006:1–8
4. Caruana R (1997) Multitask learning. Mach Learn 28(1):41–75
5. Chen L, Mao X, Xue Y-L, Cheng LL (2012) Speech emotion recognition: features and classification models. Digital Signal Process 22(6):1154–1160
6. Chen T, Li M, Li Y, Lin M, Wang N, Wang M, Xiao T, Xu B, Zhang C, Zhang Z (2015) Mxnet: a, flexible and efficient machine learning library for heterogeneous distributed systems. arXiv:abs/1512.01274
7. Gong P, Ye J, Zhang C (2012) Robust multi-task feature learning. In: The international conference on knowledge discovery and data mining, pp 895–903
8. Goodfellow IJ, Bengio Y, Courville AC (2016) Deep learning (Adaptive Computation and Machine Learning series). MIT Press

9. Grosse R, Raina R, Kwong H, Ng AY (2007) Shift-invariant sparse coding for audio classification. In: Proceedings of the Twenty-Third conference on uncertainty in artificial intelligence, AUAI Press, pp 149–158
10. Gu J, Wang Z, Kuen J, Ma L, Shahroudy A, Shuai B, Liu T, Wang X, Wang G (2015) Recent advances in convolutional neural networks. arXiv:abs/1512.07108
11. Han Y, Kim J, Lee K (2017) Deep convolutional neural networks for predominant instrument recognition in polyphonic music. IEEE/ACM Trans Audio Speech Lang Process 25(1):208–221
12. Hannun AY, Case C, Casper J, Catanzaro B, Diamos G, Elsen E, Prenger R, Satheesh S, Sengupta S, Coates A, Ng AY (2014) Deep speech: scaling up end-to-end speech recognition. arXiv:abs/1412.5567
13. Hansen JHL, Liu G (2016) Unsupervised accent classification for deep data fusion of accent and language information. Speech Comm 78:19–33
14. Hashimoto K, Xiong C, Tsuruoka Y, Socher R (2016) A joint many-task model: Growing a neural network for multiple NLP tasks. arXiv:abs/1611.01587
15. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: IEEE conference on computer vision and pattern recognition, pp 770–778
16. Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780
17. Ioffe S, Szegedy C (2015) Batch normalization: accelerating deep network training by reducing internal covariate shift. In: Proceedings of the 32nd international conference on machine learning, pp 448–456
18. LISA Lab (2017) Convolutional neural networks (lenet). http://deeplearning.net/tutorial/contents.html
19. Lartillot O, Toiviainen P (2007) MIR in matlab (II): a toolbox for musical feature extraction from audio. In: Proceedings of the 8th International Conference on Music Information Retrieval, pp 127–130
20. LeCun Y, Bottou L, Bengio Y, Haffner P (1998) Gradient-based learning applied to document recognition. Proc IEEE 86(11):2278–2324
21. Lee H, Pham PT, Largman Y, Ng AY (2009) Unsupervised feature learning for audio classification using convolutional deep belief networks. In: Advances in annual conference on neural information processing systems, pp 1096–1104
22. Li C, Ma X, Jiang B, Li X, Zhang X, Liu X, Cao Y, Kannan A, Zhu Z Deep speaker: an end-to-end neural speaker embedding system. arXiv:abs/1705.02304.2017
23. Livingstone SR, Peck K, Russo FA (2012) Ravdess: The ryerson audio-visual database of emotional speech and song. In: Annual meeting of the canadian society for brain, behaviour and cognitive science
24. Long M, Wang J (2015) Learning multiple tasks with deep relationship networks. CoRR, arXiv:abs/1506.02117
25. Pedersen C, Diederich J (2007) Accent classification using support vector machines. In: Annual IEEE/ACIS, international conference on computer and information science, pp 444–449
26. Phapatanaburi K, Wang L, Sakagami R, Zhang Z, Li X, Iwahashi M (2016) Distant-talking accent recognition by combining GMM and DNN. Multimedia Tools Appl 75(9):5109–5124
27. Pons J, Slizovskaia O, Gong R, Gómez E, Serra X (2017) Timbre analysis of music audio signals with convolutional neural networks. arXiv:abs/1703.06697
28. Rao P (2008) Audio signal processing. In: Speech, audio, image and biomedical signal processing using neural networks. Springer, pp 169–189
29. Shegokar P, Sircar P (2016) Continuous wavelet transform based speech emotion recognition. In: International conference on signal processing and communication systems, pp 1–8
30. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv:abs/1409.1556
31. Steven R, Thompson WF, Wanderley MM, Palmer C (2015) Common cues to emotion in the dynamic facial expressions of speech and song. Q J Exp Psychol 68(5):952–970
32. Wu B, Jia J, He T, Du J, Yi X, Ning Y (2016) Inferring users' emotions for human-mobile voice dialogue applications. In: IEEE international conference on multimedia and expo, pp 1–6
33. Zhang B, Essl G, Provost EM (2015) Recognizing emotion from singing and speaking using shared models. In: International conference on affective computing and intelligent interaction, pp 139–145
34. Zhang B, Provost EM, Essl G (2016) Cross-corpus acoustic emotion recognition from singing and speaking: a multi-task learning approach. In: IEEE international conference on acoustics, speech and signal processing, pp 5805–5809
35. Zhu X, Suk H-I, Shen D (2014) A novel multi-relation regularization method for regression and classification in ad diagnosis. In: International conference on medical image computing and computer-assisted intervention. Springer, pp 401–408

36. Zhu Y, Lucey S (2015) Convolutional sparse coding for trajectory reconstruction. IEEE Trans Pattern Anal Mach Intell 37(3):529–540
37. Zhu X, Suk H-I, Wang L, Lee S-W, Shen D (2015) A novel relational regularization feature selection method for joint regression and classification in AD diagnosis. Medical Image Analysis 38:205–214
38. Zhu Y, Zhu X, Kim M, Shen D, Wu G (2016) Early diagnosis of alzheimers disease by joint feature selection and classification on temporally structured support vector machine. In: International conference on medical image computing and computer-assisted intervention. Springer, Berlin, pp 264–272

Yuni Zeng is a Ph.D. student at Machine Intelligence Laboratory, College of Computer Science, Sichuan University, Chengdu currently. She received her B.S. degree in College of Computer Science, Sichuan University at 2016. Her current research interests include Neural Networks and Big Data.

Hua Mao received the B.S. degree and M.S. degree in Computer Science from Electronic Science and Technology of China (UESTC) in 2006 and 2009, respectively. She received her Ph.D. degree in Computer Science and Engineering in Aalborg University, Denmark in 2013. Her current research interests include Neural Networks and Big Data.

Dezhong Peng B.Sc. degree (1998) in applied mathematics, University of Electronic Science and Technology of China (UESTC), Chengdu, China; M.E. degree (2001) in computer science and engineering, University of Electronic Science and Technology of China (UESTC); Assistant Lecturer (2001), Lecturer (2003) in the school of Applied Mathematics, University of Electronic Science and Technology of China (UESTC); PhD degree (2006) in computer science and engineering, University of Electronic Science and Technology of China (UESTC); Postdoctoral Research Fellow (2007.07-2009.09) at the school of engineering, Deakin University, Melbourne, Australia; Currently, he is a Professor at Sichuan University, Chengdu, China.

Zhang Yi received the Ph.D. degree in mathematics from the Institute of Mathematics, The Chinese Academy of Science, Beijing, China, in 1994. Currently, he is a Professor at the College of Computer Science, Sichuan University, Chengdu, China. He is the co-author of three books: Convergence Analysis of Recurrent Neural Networks (Kluwer Academic Publisher, 2004), Neural Networks: Computational Models and Applications (Springer, 2007), and Subspace Learning of Neural Networks (CRC Press, 2010). He is a Fellow of IEEE. He is the Chair of IEEE Chengdu Section (2015). He was an Associate Editor of IEEE Transactions on Neural Networks and Learning Systems (2009 2012), and an Associate Editor of IEEE Transactions on Cybernetics (2014). His current research interests include Neural Networks and Big Data. He is the founding director of Machine Intelligence Laboratory. He is also the founder of IEEE Computational Intelligence Society, Chengdu Chapter and he is an IEEE fellow.