Sch-net: A Deep Learning Architecture for Automatic Schizophrenia Setection

Jia Fu
Sichuan University - Wangjiang Campus: Sichuan University
Sen Yang
Tencent AI Lab
Fei He
Sichuan University - Wangjiang Campus: Sichuan University
Ling He (ling.he@scu.edu.cn)
Sichuan University https://orcid.org/0000-0002-7168-2737
Yuanyuan Li
Sichuan University West China Hospital
Jing Zhang
Sichuan University - Wangjiang Campus: Sichuan University
Xi Xiong
Chengdu University of Information Technology

Research

Keywords: schizophrenia, deep learning, skip connections, attention mechanism, pathological speech detection

Posted Date: May 24th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-533815/v1

Version of Record: A version of this preprint was published at BioMedical Engineering OnLine on August 3rd, 2021. See the published version at https://doi.org/10.1186/s12938-021-00915-2.

RESEARCH

Sch-net: a deep learning architecture for automatic schizophrenia detection

Jia Fu[1†], Sen Yang[2†], Fei He[1], Ling He[1*], Yuanyuan Li[3], jing Zhang[1] and Xi Xiong[4]

[*]Correspondence:
ling.he@scu.edu.cn
[1]The College of Biomedical Engineering, Sichuan University, Chengdu, China
Full list of author information is available at the end of the article
[†]Equal contributor

Abstract

Background: Schizophrenia is a chronic and severe mental disease, which largely influences the daily life and work of patients. In the clinic, schizophrenia with negative symptoms is usually misdiagnosed and hardly treated. The diagnosis is also dependent on the experience of clinicians. It is urgent to develop an objective and effective method to diagnose schizophrenia with negative symptoms. Recent studies had shown that impaired speech could be considered as an indicator to diagnose schizophrenia. The literature about schizophrenia speech detection was mainly based on feature engineering, in which effective feature extraction is difficult because of the variability of speech signals.

Methods: A novel deep learning architecture based on a convolutional neural network, termed Sch-net, is designed for end-to-end schizophrenia speech detection in this work. It avoids the procedure of artificial feature extraction and combines the advantages of skip connections and attention mechanism to discriminate schizophrenia patients and controls.

Results: We validate our Sch-net through ablation experiments on a schizophrenia speech dataset that contains 28 schizophrenia patients and 28 healthy controls. The comparisons with the models based on feature engineering and classic deep neural networks are also conducted. The experimental results show that the Sch-net has a great performance on schizophrenia speech detection task, which can achieve 97.76% accuracy on the schizophrenia speech dataset. To further verify the generalization of our model, the Sch-net is tested on open access LANNA children speech database for specific language impairment detection. Our code is available at `https://github.com/Scu-sen/Sch-net`.

Conclusions: Extensive experiments show that the proposed Sch-net can provide the aided information for the diagnosis of schizophrenia speech and specific language impairment.

Keywords: schizophrenia; deep learning; skip connections; attention mechanism; pathological speech detection

Background

Schizophrenia is a chronic mental disease that affects about 1% of the population worldwide [1, 2]. The disease often begins in late adolescence, and it has a large impact on patients' social activity and brain development. Schizophrenia is characterized by disordered thinking, impaired speech, and abnormal behaviors. Clinical diagnosis of schizophrenia is generally based on patients' retrospective recall biases [3] and the speech/behaviors observed via clinical interviews. Symptoms of schizophrenia can be divided into two types, positive symptoms, and negative symptoms. Positive symptoms include delusions and hallucinations [2, 4], and negative

symptoms include flat affective, alogia, loss of interest, and disability in activities [5]. Clinical experience had shown that it is harder to diagnose and treat patients with negative symptoms than those with positive symptoms [6]. Positive symptoms are likely to be replaced by negative symptoms in the late episode of schizophrenia, and negative symptoms may persist even though after treatment [7]. Negative symptoms contribute more to the long-term morbidity, higher rates of disability, and poor quality of life in most schizophrenia patients than positive symptoms do [8, 9, 10, 11, 12]. In addition, the clinical diagnosis relies on the experience of clinicians. Hence, it is urgent to propose a method to diagnose schizophrenia patients with negative symptoms objectively and effectively.

Patients with schizophrenia exhibit brain structural abnormalities [13, 14, 15], that are accountable for the speech disorders and cognitive impairments. Cohen [16] discovered that speech characteristics are significantly related to the negative symptoms of schizophrenia. Rosenstein [17] confirmed that adolescents with high-risk psychosis exhibit speech impairments for months/years before they are diagnosed. Flat affective and incoherent language expression are typical performances in schizophrenia with negative symptoms [18]. Schizophrenia groups exhibit reduced pitch variation [19], increased pauses [20] and poverty of content [21]. The number and duration of pause are closely related to the evaluation of affective flattening in the clinic [5, 22, 23].

Generally, most existing methods [24, 25, 26, 27, 28, 29, 30, 3, 31, 32, 33, 34] analyze schizophrenia speech using feature engineering and classifiers/correlation analysis. Schizophrenia speech detection experiments are conducted by extracting fluency features, intensity-related features, spectrum-related features, and so on. These studies had proved that speech can be viewed as an automated biomarker for the diagnosis of schizophrenia. However, owing to the limitation in the amount of data and the difficulties of effective feature extraction, it is still difficult to propose a robust model. In this work, the Schizophrenia network (Sch-net) designed based on convolutional neural network (CNN) is proposed to achieve the end-to-end detection of schizophrenia speech. The propose of Sch-net can avoid the problems in feature extraction. The contributions in our work can be summarized as follows:

1) This work proposes the Sch-net to detect schizophrenia speech. To the best of our knowledge, this is the first work based on CNN to detect schizophrenia.

2) The designed Sch-net utilizes skip connections and a convolutional block attention module (CBAM). We verify the effectiveness of each module and the performance of its own model. The experimental results show that the performance is better than the existing optimal model.

3) To further validate the robustness and generalization of our model, the Sch-net is tested on open access LANNA children speech database, as well as the schizophrenia speech dataset.

Related works

Schizophrenia speech detection has been studied for the last few decades. Previous studies [24, 25, 26, 27, 28, 29, 30, 3, 31, 32, 33, 34] are mainly achieved based on feature engineering. In this section, we will review the related studies from the view of feature categories. The feature extracted for the classification of schizophrenia

speech and normal speech can be divided into two types, time-domain features, and spectrum-related features.

Time-domain features: Schizophrenia patient with negative symptoms usually exhibits incoherent language that can be described by time-domain features, including pitch-related features, fluency features, and intensity-related features. 1) *Pitch-related features:* Pitch is the fundamental frequency of vocal cord vibration for voiced initial consonants and some unvoiced initial consonants [35]. Pitch-related features are common used in analyzing flat affect of negative symptoms [24, 25, 26, 27, 28, 29, 30, 3]. Studies [28, 27, 26, 24] demonstrated that schizophrenia speech characterized by less variability in vocal pitch than normal speech. 2) *Fluency features:* The incoherent expression in schizophrenia usually manifests as more pauses and the longer duration of pauses. Fluency features are employed to distinguish schizophrenia groups and controls in recent studies [26, 31, 32], such as the number of pauses and natural turns, mean pause duration, the total length of pauses, the proportion of silence and speaking, and natural interjections. 3) *Intensity-related features:* Voice intensity is an intuitive indicator for conveying emotional information in human communication [36]. Previous studies [24, 26, 28] had demonstrated that the voice intensity of patients with schizophrenia has less variation than that of controls. The intensity-related features can be calculated based on the variability of energy per second/syllable [24, 26, 28].

Spectrum-related features: Spectrum-related features generally refer to the variables computed based on the spectrum that containing spatial domain and frequency domain information. Spectrum-related features are commonly used for schizophrenia speech analysis, reflecting the energy distribution and the vocal tract characteristics during speech production. The typical spectrum-related features, such as formants, auditory-based spectral features, and spectral envelope features, had been proved to be an effective feature for schizophrenia speech detection [28, 33, 29, 34]. 1) *Formants:* Formant is the descriptor that reflects the resonance frequency of the vocal tract. Compton et.al [28] demonstrated that the range of the second formant for schizophrenia is smaller than that for controls. Chhabra et.al [33] concluded that schizophrenia patient reduces the use of formant dispersion in the similarity-dissimilarity ratings. 2) *Auditory-based spectral features and spectral envelope features:* Auditory-based spectral features refer to the spectral parameters that are computed based on human auditory characteristics, and spectral envelope features refer to the envelope and its variants of the spectrum. Mel-frequency cepstral coefficient (MFCC) is one typical auditory-based spectral feature, and linear prediction coefficient (LPC) is a commonly used spectral envelope feature. MFCC is gained by using Mel-frequency filters, in which the center frequency is computed according to the human auditory characteristics. LPC is calculated to estimate the vocal tract model, which reflects the resonance during speech production. Studies [29, 34] used MFCCs and LPCs to analyze the characteristics of schizophrenia speech. Results in [34] showed that the MFCC scores of schizophrenia speech are significantly lower than that of controls, and the LPC scores of schizophrenia speech are significantly higher than that of controls.

Low-level acoustic features mentioned above [24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34] are usually used at the same time to differentiate schizophrenia speech

and normal speech. Acoustic features are extracted using OpenSMILE, pyAudio-Analysis, openEAR, and signal processing techniques. Classification experiments are conducted using classifiers (such as k-Nearest Neighbors, Decision Trees, Naive Bayes), combined with cross-validation (such as k-fold cross-validation and leave-one-out cross-validation). Studies [26, 28, 29, 30, 31, 32] achieved 64%-93% accuracy on schizophrenia speech detection task using 8-98 schizophrenia patients and 7-102 controls.

Results

To demonstrate the effectiveness of the proposed model, comprehensive experiments are conducted. We first describe the schizophrenia speech dataset and implementation details. Next, the ablation studies are presented to demonstrate the improvements of each component in the proposed Sch-net. Then the classification results and comparisons with state-of-the-art methods are shown and analyzed. The network visualization is also presented using Grad-CAM. Finally, to further validate the generalization of proposed method, the classification experiments on the LANNA children speech database are conducted.

Schizophrenia dataset

Our study has 28 schizophrenia patients (18 females and 10 males) and 28 matched healthy controls (18 females and 10 males). The schizophrenia group is with a mean age of 40.6 years (SD 9.4 years), and the control group is with a mean age of 36.5 years (SD 9.1 years). All subjects are native Mandarin speakers, and they did not have any other past or current disease affecting speaking process. Patients were recruited from the Psychiatry Department of the Mental Health Center, Sichuan University. This department is one of four major mental health centers in China. Schizophrenia group was diagnosed by clinicians based on the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5) that outlines the concise and explicit criteria for schizophrenia diagnosis [48]. All subjects provided the written informed consent.

The schizophrenia database was composed of audio signals that were recorded in a 16-bit mono/dual-format at a sampling rate of 44.1kHz. Participants were asked to achieve the reading task containing four paragraphs. There were 8-10 sentences for each paragraph and four emotions in the audio signals. The emotions are calm, happiness, anger, and fear. We selected a fixed sentence for each emotion recording, and the transcriptions of speech signals are listed in Table 1.

Table 1 The text for speech recording in Mandarin and its corresponded English translation

Emotion	Text (Mandarin)	Text (English)
Calm	Ta yi nian si ji dou ke yi kai hua, hua duo yi ban shi hong se huo fen se de.	It can bloom all year round, and the flowers are generally red or pink.
Anger	Gen ni shuo le duo shao ci le, bu xu wan wo de wan! Kan ba, wan bei da sui le! Ni zhen de shi yao qi si wo!	I told you so many times that you are not allowed to play with my bowls! Look, the bowl is shattered! You are really mad at me!
Fear	Ma ma, dui bu qi, wo...wo...wo bu shi gu yi de.	Mom, I'm sorry, I...I...I didn't mean it!
Happiness	Ha ha, tai hao la! Tai hao la! Ma ma, ma ma, wo kao le 98 fen!	Awesome, it's awesome! Mom, Mom, I got 98 points!

Implementation Details

In this study, all audios are converted to spectrograms for analysis using the FFT method. To make the network learn invariant features to geometric perturbations and improve the invariance of the network to noise, data augmentation methods are performed, including repeating random crop, random rotation, random rescaling, random Gaussian noise, masking blocks of frequency channels [49], and masking blocks of time steps [49].

The input image of the Sch-net is with the size of 128×256 pixels. Table 2 shows the Sch-net architecture details. In this architecture, the size of each filter in Conv layers is set as 3×3. There are 64, 128, 256, 512 filters in the first to fourth Conv layers, respectively. And there are 512 filters in the three skip connections. The convolved images are normalized using a ReLU activation in Conv blocks. The max pooling and average pooling in pooling layers are obtained every 2×2, with a stride of 2. In the CBAM, 2048 filters of size 7×7 are used to highlight effective features. The highlighted features are convolved with 512 filters of size 3×3. In the FC neural network, there are 512 neurons in the first hidden layer and 2 neurons in the second layer. The final output of this network is the probability of the output values.

Table 2 Sch-net architecture details

Layer	Dimension
Conv1	2×[3×3 (64 filters)]
Conv2	2×[3×3 (128 filters)]
Conv3	2×[3×3 (256 filters)]
Conv4	2×[3×3 (512 filters)]
Conv5-8	3×3 (512 filters)
Max-pooling	2×2
Average-pooling	2×2
CBAM	7×7 (2048 filters)
FC	1×1×512, 1×1×2 (two hidden layers)

In all experiments, the binary cross-entropy is adopted as the loss function, and Adam [50] is used as the optimization algorithm. All experiments are implemented based on the PyTorch framework [51] and trained on a workstation with Intel(R) Xeon(R) CPU E5-2680 v4 2.40GHz processors and an NVIDIA Tesla P40 (24 GB) installed. The network is trained using batch size 16 for 50 epochs. The initial learning rate is set to 0.0003 and decreases by 10 times after 25 epochs. Five-fold cross-validation is applied in the experiments because of the limited amount of schizophrenia speech dataset.

The ablation studies

In this subsection, the effectiveness of our network is verified. The Sch-net's backbone network is based on the CNN, with adding skip connections to enrich the feature information. And the CBAM is applied to emphasize the more effective features with bigger weights. For this ablation study, we evaluate the contributions to the schizophrenia speech detection task using two key components in our proposed method. The experimental results are listed in Table 3.

In Table 3, SC means skip connection. As can be seen, the skip connection enriches the information of feature maps and improves the classification accuracy by 0.69% on the schizophrenia speech dataset. The CBAM selects the meaningful features for

Table 3 The overall performance of schizophrenia speech detection using Sch-net and its components

Architecture	Accuracy	Precision	Recall	F1-score
Backbone	0.9431	0.9385	0.9557	0.9444
Backbone+SC	0.9510	0.9446	0.9418	0.9498
Backbone+CBAM	0.9555	0.9617	0.9549	0.9561
Sch-net	**0.9776**	**0.9833**	**0.9727**	**0.9773**

classification and improves accuracy by 1.24%. Significant improvement of 3.45% for classifying schizophrenia speech and normal speech is achieved when adding skip connections and CBAM to the backbone network. The proposed Sch-net combines the advantages of skip connection and CBAM, achieving better performance on the classification task.

Comparison with the models based on feature engineering and classifiers

Previous literature about automatic schizophrenia speech detection [24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34] are almost based on feature engineering and pattern recognition technology. In this subsection, the performances of the combination of feature engineering and classifiers are displayed and analyzed. Four types of acoustic features are extracted, which are time-domain features, Fast Fourier Transform-based (FFT-based) spectral features, auditory-based spectral features, and spectral envelope features. Four classifiers are adopted, including random forest (RF), k-nearest neighbor (KNN), support vector machine (SVM), and linear discriminant analysis (LDA).

Time-domain features used in this work contain short-term energy (STE), pitch and fluency features. The STE feature implies the intensity of speech, and the pitch conveys the characteristics of the vocal cord in the pronunciation process. The fluency feature can reflect the degree of coherence in expression. Considering the reduced syntactic complexity and abnormal pauses in schizophrenia speech, five fluency features (total recording time, the total length of voice segments, the ratio of voice segments, max duration of pauses, mean length of syllables) are employed to compose a feature set.

FFT-based features refer to the features in the frequency domain that are computed by short-time Fourier transform. In this work, two FFT-based features (spectrogram and long-term average spectrum (LTAS)) are adopted in this work. LTAS describes the resonance characteristics by computing the short-term Fourier magnitude spectra [52]. The features had shown better performance in speech sentiment analysis and pathology speech analysis [53, 54, 55].

Auditory-based features are proposed to simulate the clinical diagnosis. Schizophrenia is diagnosed by clinical doctors through a comprehensive evaluation of speech and behaviors. Therefore, speech signals are necessary to be analyzed by combining with human auditory characteristics. In this study, MFCC and its modification, Gammatone cepstral coefficient (GTCC) [56], are applied to extract the spectrum features. MFCCs and GTCCs are computed using a series of filters that are designed according to the frequency response characteristics of the human auditory system.

The spectral envelope feature is another type of acoustic features that are commonly used to describe the vocal tract characteristics in speech production. In this

work, LP and its deformations, stabilized weighted linear prediction (SWLP) [57] and extended weighted linear prediction (XLP) [58], are tested on the schizophrenia speech dataset. SWLP is an improved WLP that is proposed to model speech by applying the temporal weighting of the square of the residual signal [57]. XLP is a further generation of the WLP analysis, which allows temporal weighting on a finer time scale [58]. SWLP and XLP had performed well on the speech recognition tasks and pathology speech detection [59, 60].

The overall performances of feature engineering and classifiers are listed in Table 4 using accuracy, precision, recall and F1-score. The bold font in Table 4 represents the highest value in each type of features using different classifiers. It can be seen that the fluency feature, spectrogram, GTCC and XLP achieve the highest F1-score in its corresponding feature group. When compared the results in Table 3 and Table 4, it can be seen that the proposed Sch-net have a better performance than the models based on feature engineering and classifiers.

Time-domain feature

As shown in Table 4, the F1-score of schizophrenia speech detection using STE reaches 0.6306. Owing to the difficulty in expressing for schizophrenia patients, the intensity of schizophrenia speech tends to be lower than that of controls. The STE feature can describe the intensity of speech, but it may be influenced by the different distances between the recording equipment and speakers. Thus, the performance of the STE feature is not as good as the fluency feature.

Though studies [28, 27, 26, 24] had proved that there are significant differences in pitch between schizophrenia speech and normal speech, the pitch gains the worst performance among time-domain features. The results are consistent with the views in Rapcan et.al [26] and Chhabra et.al [33]. Results in [26, 33] demonstrated that there are no significant differences in the distribution of pitch between the two groups.

Fluency feature performs well on the schizophrenia speech detection, owing to the thought disorder and language impairments of schizophrenia patients [61]. The cognitive impairment also contributes to the incoherence of schizophrenia speech.

FFT-based spectral feature

The LTAS achieves 62.11% accuracy on the schizophrenia speech dataset. The LTAS is calculated as the average of a spectrogram, reflecting the spectrum of glottal source and vocal tract [62]. Studies [26] had shown that schizophrenia speech has lower variations in energy than normal speech. The unexpected accuracy using LTAS may be caused by the average operation that eliminates the differences in variations between two groups.

The spectrogram achieves better performance than LTAS. It is the representation of speech in the time-frequency domain. It not only contains the energy distribution in frequency bands but also reflects the pitch and formant information. It has been proved that schizophrenia speech have less variability in pitch and voice intensity, smaller range of second formant than normal speech [28, 27, 26, 24]. Thus, the spectrogram covers more effective features for discriminating schizophrenia patients and controls than the LTAS does.

Table 4 The performance of feature engineering and classifiers on schizophrenia speech detection

Classifier		Feature									
		Time-domain feature			FFT-based spectral feature		Auditory-based spectral feature		Spectral envelope feature		
		STE	Pitch	Fluency feature	LTAS	Spectrogram	MFCC	GTCC	LP	SWLP	XLP
RF	Accuracy	0.7686	0.5935	**0.8213**	0.6464	**0.8972**	0.8043	**0.8791**	0.9245	0.9377	**0.9423**
	Precision	0.6251	0.5847	**0.8281**	0.6052	**0.8946**	0.7818	**0.8487**	0.9055	**0.9319**	0.9282
	Recall	0.7126	0.5754	**0.8103**	0.5545	**0.9103**	0.8577	**0.9289**	0.9549	0.9466	**0.9644**
	F1-score	0.6306	0.6322	**0.8133**	0.5513	**0.8972**	0.8144	**0.8856**	0.9280	0.9391	**0.9453**
KNN	Accuracy	**0.7723**	0.5385	0.7390	0.7504	**0.8974**	0.8626	**0.8753**	0.9204	0.9287	**0.9375**
	Precision	0.6410	0.5308	**0.7050**	0.6489	**0.8566**	0.8977	**0.9043**	0.8939	0.9117	**0.9196**
	Recall	0.6063	0.5597	**0.7985**	0.6152	**0.9636**	0.8312	**0.8494**	**0.9640**	0.9549	0.9636
	F1-score	0.6123	0.5382	**0.7418**	0.6211	**0.9046**	0.8591	**0.8700**	0.9257	0.9315	**0.9398**
SVM	Accuracy	**0.7905**	0.5172	0.7746	0.7358	**0.9024**	0.8625	**0.8929**	0.9164	0.9291	**0.9334**
	Precision	0.6447	0.5087	**0.7657**	0.6556	**0.8741**	0.8555	**0.8762**	0.8980	**0.9183**	0.9126
	Recall	0.5999	0.4767	**0.7875**	0.5435	**0.9549**	0.8929	**0.9198**	0.9470	0.9466	**0.9636**
	F1-score	0.6155	0.4644	**0.7627**	0.5813	**0.9091**	0.8689	**0.8960**	0.9206	0.9317	**0.9356**
LDA	Accuracy	**0.7858**	0.5087	0.7452	0.7314	**0.8385**	0.8887	**0.9198**	0.9026	0.9069	**0.9109**
	Precision	0.6447	0.4625	**0.7083**	0.6622	**0.8053**	0.9394	**0.9479**	0.8963	**0.9053**	0.8868
	Recall	0.5898	0.5391	**0.7522**	0.5380	**0.9095**	0.8474	**0.8933**	0.9198	0.9111	**0.9462**
	F1-score	0.6093	0.4710	**0.7104**	0.5821	**0.8498**	0.8807	**0.9161**	0.9060	0.9079	**0.9146**

Auditory-based spectral feature

The GTCC achieves a better performance than MFCC on schizophrenia speech detection task, which is caused by using different auditory filters. The MFCC is computed based on a series of triangular bandpass filters with equal bandwidth. The GTCC employs the Gammatone filters with equivalent rectangular bandwidth to model the human auditory response [63]. It minimizes the loss of spectrum information and increases the correlation among the outputs of Gammatone filters [63]. Therefore, the GTCC contains more effective information to detect schizophrenia speech than the MFCC.

Spectral envelope feature

The F1-scores of schizophrenia speech detection using LP, SWLP and XLP are above 0.9. The SWLP and XLP have slightly better results than LP. The results of spectral envelop features are gained when the order of LP is set as 38. The order is gained through experiments. Deller et.al [64] showed that vocal tract can be modeled when the order ranges from 8 to 10 for healthy subjects. Dong et.al [65] demonstrated that the LP order usually is higher to estimate pathological speech. But if the order of LP is higher than 38, the unwanted contributions to the LP spectrum would appear because of the excitation signal. Results in this work are consistent with the conclusions in [64, 65].

Studies [28, 33] had shown that formant is an indicator to distinguish schizophrenia speech from controls. The LP reflects the characteristics of the vocal tract, such as the frequency of formants. However, the LP analysis relies on the excitation signal, which is influenced by the harmonics. The SWLP reduces the impact by composing the temporal weights on the closed-phase interval of the glottal cycle [60]. And the XLP improves the time scale on the spectral envelop by weighting each lagged speech signal separately [60]. The SWLP and XLP highlight the formant information that can be used to distinguish patients from controls.

Comparison with classic deep neural networks

To compare the performance of our model with other deep neural networks, in this subsection, five networks (AlexNet [66], VGG16 [67], ResNet34 [68], DenseNet121 [69], and Xception [70]) are adopted to discriminate schizophrenia patients and controls. The five deep neural networks are commonly used for speech recognition and classification tasks [71, 72, 73, 74, 75]. AlexNet [66] is the winner of the ImageNet Large Scale Visual Recognition Challenge in 2012, which reduces overfitting in the FC layers using dropout. VGG16 [67] is a good architecture for benching on classification tasks, which has small filters and deep depth. ResNet34 [68] with residual blocks was introduced to alleviate the degradation problem caused by increasing stacked layers via adding shortcut connections. To reduce the impact on vanishing gradient, the feed-forward fashion in the connection between each layer to every other layer was used in DenseNet121 [69]. DenseNet also can strengthen the feature propagation and reduce the number of parameters [69]. To obtain fast convergence and good performance on the model's expressive ability, Xception [70] replaces the inception modules with depthwise separable convolutions in deep CNN. Table 5 lists the overall results of classifying schizophrenia speech and normal speech using the five deep neural networks and our method.

Table 5 The performance of schizophrenia speech detection using classic deep neural networks and the proposed Sch-net

Network	Accuracy	Precision	Recall	F1-score
AlexNet [66]	0.9158	0.9491	0.8933	0.9156
VGG16 [67]	0.9423	0.9417	0.947	0.9418
ResNet34 [68]	0.9462	0.9686	0.9273	0.9447
DenseNet121 [69]	0.9599	0.9593	0.9636	0.9598
Xception [70]	0.9553	0.9433	**0.9735**	0.9565
Sch-net (our)	**0.9776**	**0.9833**	0.9727	**0.9773**

As shown in Table 5, the accuracies of schizophrenia speech detection using AlexNet and VGG16 are 91.58% and 94.23%, respectively. The depth of AlexNet and VGG16 is shallow, contributing to the insufficient information in feature maps. ResNet achieves 94.62% accuracy on the schizophrenia speech dataset, owing to the introduction of the residual module. DenseNet and Xception gain slightly higher accuracies than ResNet, owing to the networks contains both the residual module and effective algorithms to reduce the number of parameters. The proposed Sch-net in this work achieves a marginally better performance than the five networks because it can gain the local and global features simultaneously via CBAM and skip connections. The feature map contains more abundant information to better distinguish schizophrenia speech and controls.

Network visualization using Grad-CAM

In recent years, deep learning methods have already achieved high accuracy that approaches the manual diagnosis accuracy in many fields through improving the computing capabilities and expanding the dataset. It can simplify and speed up the diagnosis, and reduce the workload of doctors. However, the process of generating predicted labels from input data is still uninterpretable. To make the decision-making process in deep learning transparent, this work applies the Grad-CAM [76] to Sch-net using speech samples from schizophrenia group and healthy group. Grad-CAM is a visualization method to show the importance of each neuron for the classification by using the gradient information in the last Conv layer [76]. Grad-CAM heatmap shows the more effective parts for decision-making as brighter regions. We attempt to consider how the Sch-net works on making good use of features, through observing the spectrogram and activation maps. In this subsection, the input spectrogram and its corresponding activation map generated in the last Conv layer of normal speech and schizophrenia speech are shown in Fig. 1.

In Fig. 1, spectrograms of normal speech and schizophrenia speech are shown in (a) and (c), respectively. Activation maps of normal speech and schizophrenia speech are depicted in (b) and (d). The brighter region in the spectrogram means more energy concentrated, and that in the activation map means larger weight located.

As shown in Fig. 1(a) and (c), schizophrenia speech and normal speech have different distributions of concentrated energy in the spectrogram. Through the horizontal comparison, two findings of two groups can be seen in this figure, which can be listed as follows:

1) The energy concentration in the frequency domain of schizophrenia speech is almost below 5000Hz, while normal speech has a wider range of energy concentration bands, that can be extended from 8000Hz to 10000Hz. Blunted affecting

is a construct of negative symptoms in schizophrenia [77]. Schizophrenia patients with negative symptoms may speak with a dull monotone voice [78], resulting in a small range of the energy concentration region. While healthy controls have a more flexible emotional expression. The angry, fearful and happy speech exhibit a higher intonation, faster speed rate and more high-frequency energy [79]. And the sad speech changes slowly and contains a small amount of higher-frequency energy [80]. Thus, normal speech has a wider range of energy distribution than schizophrenia speech.

2) It can be seen that schizophrenia speech and normal speech both have concentrated energy region and apparent formant horizontal stripes in the low-frequency bands below 2000Hz. The difference between the two groups is the shape of formant horizontal stripes. For schizophrenia speech, the stripes are almost continuous, which is inconsistent with the energy distribution characteristics of vowels and consonants. The vowels have energy concentration in both low- and high-frequency range [81]. The unvoiced consonants mainly have high-frequency energy components, and it rarely has formants [82]. According to the texture used in this work, the continuous-time speech signals has both vowels and consonants. Therefore, there are supposed to show a short disappearance of formant horizontal stripes on the spectrogram. It can be guessed that the continuous stripes in the spectrogram of schizophrenia speech may be caused by the incorrect placement of articulators during speech production. The wrong articulation process leads to the unvoiced consonants are produced as voiced consonants.

Observing both the spectrogram and its corresponding activation map in Fig. 1, it can be seen that the Sch-net can capture the features in high-frequency bands for normal speech, and can give larger weights to low-frequency bands for schizophrenia speech. The results of Sch-net are consistent with human visual perception, which is difficult to achieve using the models based on feature engineering. The Sch-net has excellent learning ability to extract features, and it achieves better performances on schizophrenia speech detection than traditional feature engineering models adopted in this work.

Further validation of the proposed Sch-net using LANNA children speech database

Schizophrenia is a neurodevelopmental disorder affecting the language expression of patients [83]. Specific Language Impairment (SLI), also termed development dysphasia, is described as a neurological disorder of the brain [84, 85, 86, 87]. Children with SLI exhibit delayed language acquisition [88], slower linguistic processing [89], and difficulties in grammar or specific subcomponents of grammar [90, 91]. To further validate model effectiveness and generalization, the Sch-net is tested on LANNA children speech database [92] for the classification of children with SLI and healthy controls in this subsection.

LANNA children speech database [92] is the first and only publicly open speech corpora for children with SLI, which comprises 2173 speech signals from 54 children with SLI (aged from 6 to 11 years) and 1680 speech signals from 44 controls (aged from 6 to 10 years). This dataset is composed of 13 parts: vowels, consonants, syllables, six types of words, sentences, auditory differentiation, and description of the picture. Audios were recorded in a schoolroom and a consulting room using

Dictaphone, MD and microphone. The background noise in natural environments affects the quality of speech signals, leading to difficulties in speech signal processing.

Previous studies [92, 93, 94, 95, 96, 97, 98] had demonstrated that speech can be viewed as a symbol of diagnosing SLI. In [92, 93, 94], 1582 acoustic features were extracted from 34 low-level descriptors and its 21 statistical functionals. The features were given as input to SVM, achieving 96.94% accuracy on the LANNA children speech database. In[95], Gaussian posteriorgrams trained on MFCC features were employed to discriminate children with SLI and healthy controls. The kernel extreme learning machine trained with the speech signals, and it outperformed an accuracy of 99.41% on the test data. Apart from MFCC, in [96], Tonnetz and Chroma were calculated, combined with SVM, RF and Recurrent Neural Network to detect SLI. The Tonnetz and Chroma reached accuracies of 70% and 71%, respectively. In the four studies [92, 93, 94, 95, 96], high accuracies had been achieved for speaker-dependent classification.

In contrast, some methods were proposed for speaker-independent classification in [97, 98]. The top-20 LPC features were selected from 408 LPCs using Mann-Whitney U-test and Spearman's correlation in [97], and it achieved an accuracy of 97.90% on the SLI detection task. In [98], glottal features were combined with MFCCs, using a feed-forward neural network to reach 98.82% accuracy for classifying children with SLI and healthy controls.

In this subsection, five-fold cross-validation is employed. SLI dataset is divided with 80% for training and 20% for testing. Table 6 gives the classification results of children with SLI and controls using state-of-the-art methods, deep neural networks and the proposed Sch-net. As can be seen, our method outperforms the classic deep neural network and state-of-the-art methods. The proposed Sch-net can extract discriminant features of speech signals for classifying healthy children and those suffered from SLI.

Table 6 The results of SLI detection using state-of-the-art methods, classic deep neural networks and the proposed Sch-net

Method		Accuracy	Precision	Recall	F1-score
State-of-the-art method	Grill [92, 93, 94]	0.9694	1.0000	0.9474	0.9730
	Ramarao [95]	0.9941	-	-	-
	Slogrove [96]	0.9800	0.9900	0.9900	0.9900
	Sharma [97]	0.9790	-	-	-
	Reddy [98]	0.9882	-	-	-
Deep Neural Network	AlexNet	0.9132	0.9585	0.8810	0.9181
	VGG16	0.9230	0.9897	0.8787	0.9309
	ResNet34	0.9329	0.9489	0.9286	0.9386
	DenseNet121	0.9461	0.9397	0.9643	0.9518
	Xception	0.9622	0.9514	0.9863	0.9685
	Schnet (our)	0.9952	0.9979	0.9937	0.9958

Conclusions

In this work, we propose an effective end-to-end schizophrenia speech detection model, termed Sch-net, which is designed based on CNN. The Sch-net is performed by using a set of convolutional layers. The global and local features are learned by using skip connections, and the effective features are highlighted by using CBAM. In the experiments, the Sch-net is verified through ablation studies. The comparisons

with the models based on feature engineering and classic deep neural networks are conducted on a schizophrenia speech dataset that containing 28 schizophrenia patients and 28 healthy controls. The experimental results show that the Sch-net has achieved 97.76% accuracy. In addition, we visualize how the model performs on extracting features given an input spectrogram. The Grad-CAM heatmaps show the region that the Sch-net focuses on is consistent with human visual perception. Finally, the proposed method is further validated on open access LANNA children speech database, achieving higher accuracy than state-of-the-art methods.

Methods

In this work, we have developed an Sch-net based on CNN to classify schizophrenia speech and normal speech. The architecture of the proposed model is depicted in Fig. 2. The input is the spectrogram containing time-frequency domain information of speech signals. There are 12 convolutional (Conv) layers, 6 pooling layers, skip connections, an attention module and a fully-connected (FC) layer. The FC layer is composed of two hidden layers. A softmax function is employed to the output of the FC layer, and the output of the softmax is the classification result of speech samples. The backbone network of Sch-net and two essential components in Sch-net (skip connections and CBAM) are described below.

Backbone network of Sch-net

The backbone network of Sch-net shown in Fig.3 is consisted of Conv layer, pooling layer, batch normalization (BN) component, rectified linear unit (ReLU) and FC layer. When spectrogram is given as the input of Sch-net, local features in spectrogram are extracted via Conv layer. The dimension of features and the amount of computation are reduced in the pooling layer via max-pooling [37]. As the number of hidden layers increases, the network would suffer from the gradient vanishing and exploding problems. To address these problems, the BN layer and ReLU activation function are adopted. The introduction of BN components can also speed up the convergence, cut down the regularization process, and enable training with a larger learning rate [38, 39]. ReLU is a typical activation function in deep learning, which works better than sigmoid and tanh activation functions in speech recognition tasks [40, 41]. It removes the negative values in the feature map and sets the positive values in the feature map as linear function [42]. The networks can be trained effectively using the ReLU even without pre-training [43]. At the end of the network, the FC layer and softmax function are employed to achieve the classification task. The FC layer is essential to the visual representation transfer in classification tasks [44]. Each node in the FC layer is connected to all activation values in the previous layers.

Skip connections

The backbone network of Sch-net can extract the local features in spectrogram via shallow layers and max-pooling operation. There is no evidence that schizophrenia patients have a special pattern in pronunciation or schizophrenia speech has prominent local acoustic features. Thus, global features are supposed to be extracted for schizophrenia speech detection. To retain more original and global information in

the input feature map, average-pooling operation and skip connections are added to the backbone network of Sch-net. Average pooling considers all the values in the batch that has an equal size with the pooling kernel. Skip connection allows the feature map to skip some layers in the neural network and merge with high-level feature maps [45]. This connection combines the features after max-pooling and average-pooling, superimposed into a feature. Skip connections expand the dimensions of features in the network, providing more information for the classification task. The diagram of the backbone network of Sch-net with skip connections is given in Fig.4.

Attention mechanism

The output of skip connections contains low-level and high-level features. To emphasize the meaningful features and suppress the unnecessary ones for the classification task, the attention module is added to the backbone network. The output in the attention module is calculated as the weighted sum of the input values [46]. The bigger weights mean the more attention would be paid to the input vector. This work adopts a lightweight and general module, CBAM [47], to improve the performance of the network. The CBAM is composed of channel and spatial attention modules [47]. The channel attention module focuses on "what" is the effective part in the feature map by utilizing max-pooling and average-pooling with a shared network [47]. The spatial attention module tells "where" to focus or suppress by employing a convolutional layer [47]. The CBAM used in Sch-net can effectively refine the intermediate feature map with negligible computation and overheads.

Abbreviations

CNN: Convolutional Neural Network; CBAM: Convolutional Block Attention Module; MFCC: Mel-frequency Cepstral Coefficient; LP: Linear Prediction; LPC: Linear Prediction Coefficient; Conv: Convolutional; FC: Fully-connected; BN: Batch Normalization; ReLU: Rectified Linear Unit; FFT: Fast Fourier Transform; RF: Random Forest; KNN: K-Nearest Neighbor; SVM: Support Vector Machine; LDA: Linear Discriminant Analysis; STE: Short-term Energy; LTAS: Long-term Average Spectrum; GTCC: Gammatone Cepstral Coefficient; WLP: Weighted Linear Prediction; SWLP: Stabilized Weighted Linear Prediction; XLP: Extended Weighted Linear Prediction;

Acknowledgements

Not applicable.

Funding

This research was supported by the Department of Science and Technology of Sichuan Province, China (Grant No.2019YFS0236 and 2019YJ0523).

Availability of data and materials

Not applicable.

Ethics approval and consent to participate

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable.

Authors' contributions

Jia Fu and Sen Yang presented the ideas, designed and conducted relevant experiments in the manuscript. Jia Fu and Fei He wrote the manuscript. Ling He and Jing Zhang are responsible for guiding the idea and final review of the manuscript. Yuanyuan Li and Xi Xiong collected the samples used for the experiments. All authors contributed to analyzing the data and reviewing the literature, and revising the manuscript. All authors read and approved the manuscript.

Author details
[1]The College of Biomedical Engineering, Sichuan University, Chengdu, China. [2]The Tencent AI Lab, Shenzhen, China. [3]The Mental Health Center, West China Hospital of Sichuan University, Chengdu, China. [4]The School of Cybersecurity, Chengdu University of Information Technology, Chengdu, China.

References

1. McGrath, J., Saha, S., Chant, D., Welham, J.: Schizophrenia: a concise overview of incidence, prevalence, and mortality. Epidemiologic reviews **30**(1), 67–76 (2008)
2. Lavretsky, H.: History of Schizophrenia as a Psychiatric Disorder, (2008)
3. Low, D.M., Bentley, K.H., Ghosh, S.S.: Automated assessment of psychiatric disorders using speech: A systematic review. Laryngoscope Investigative Otolaryngology **5**(1), 96–116 (2020)
4. DiPiro, J.T., Talbert, R.L., Yee, G.C., Wells, B.G., Posey, L.M.: Pharmacotherapy: A Pathophysiologic Approach. 9th Ed, pp. 1019–1046. McGraw-Hill Medical, America (2014)
5. Marder, S.R., Galderisi, S.: The current conceptualization of negative symptoms in schizophrenia. World Psychiatry **16**(1), 14–24 (2017)
6. Murphy, B.P., Chung, Y.C., Park, T.W., McGorry, P.D.: Pharmacological treatment of primary negative symptoms in schizophrenia: a systematic review. Schizophrenia research **88**(1-3), 5–25 (2006)
7. Mucci, A., Merlotti, E., Üçok, A., Aleman, A., Galderisi, S.: Primary and persistent negative symptoms: Concepts, assessments and neurobiological bases. Schizophrenia Research **186**, 19–28 (2017). Special Section: Negative Symptoms
8. Kirkpatrick, B., Buchanan, R.W., Ross, D.E., Carpenter, W.T.: A separate disease within the syndrome of schizophrenia. Archives of General Psychiatry **58**(2), 165–171 (2001)
9. Milev, P., Ho, B.C., Arndt, S., Andreasen, N.C.: Predictive values of neurocognition and negative symptoms on functional outcome in schizophrenia: A longitudinal first-episode study with 7-year follow-up. American Journal of Psychiatry **162**(3), 495–506 (2005)
10. Kurtz, M.M., Moberg, P.J., Ragland, J.D., Gur, R.C., Gur, R.E.: Symptoms Versus Neurocognitive Test Performance as Predictors of Psychosocial Status in Schizophrenia: A 1- and 4-Year Prospective Study. Schizophrenia Bulletin **31**(1), 167–174 (2005)
11. Kirkpatrick, B., Fenton, W.S., Carpenter, W.T., Marder, S.R.: The nimh-matrics consensus statement on negative symptoms. Schizophrenia bulletin **32**(2), 214–219 (2006)
12. Rabinowitz, J., Levine, S.Z., Garibaldi, G., Bugarski-Kirola, D., Berardo, C.G., Kapur, S.: Negative symptoms have greater impact on functioning than positive symptoms in schizophrenia: Analysis of catie data. Schizophrenia Research **137**(1), 147–150 (2012)
13. DeLisi, L.E.: Speech disorder in schizophrenia: review of the literature and exploration of its relation to the uniquely human capacity for language. Schizophrenia bulletin **27**(3), 481–496 (2001)
14. Li, X., Branch, C.A., Ardekani, B.A., Bertisch, H., Hicks, C., DeLisi, L.E.: fmri study of language activation in schizophrenia, schizoaffective disorder and in individuals genetically at high risk. Schizophrenia research **96**(1-3), 14–24 (2007)
15. Li, X., Branch, C.A., Bertisch, H.C., Brown, K., Szulc, K.U., Ardekani, B.A., DeLisi, L.E.: An fmri study of language processing in people at high genetic risk for schizophrenia. Schizophrenia research **91**(1-3), 62–72 (2007)
16. Cohen, A.S., Najolia, G.M., Kim, Y., Dinzeo, T.J.: On the boundaries of blunt affect/alogia across severe mental illness: Implications for research domain criteria. Schizophrenia Research **140**, 41–45 (2012)
17. Rosenstein, M., Foltz, P.W., DeLisi, L.E., Elvevåg, B.: Language as a biomarker in those at high-risk for psychosis. Schizophrenia Research **165**, 249–250 (2015)
18. Rockville, M.: Mental health: A report of the surgeon general. US Department of Health and Human Services (1999)
19. Parola, A., Simonsen, A., Bliksted, V., , Fusaroli, R.: T138. acoustic patterns in schizophrenia: A systematic review and meta-analysis. Schizophrenia Bulletin **44**(suppl_1), 169–169 (2018)
20. Stein, J.: Vocal alterations in schizophrenic speech. Journal of Nervous and Mental Disease **181(1)**, 59–62 (1993)
21. Rezaii, N., Walker, E., Wolff, P.: A machine learning approach to predicting psychosis using semantic density and latent content analysis. NPJ Schizophrenia **5** (2019)
22. Kring, A.M., Alpert, M., Neale, J.M., Harvey, P.D.: A multimethod, multichannel assessment of affective flattening in schizophrenia. Psychiatry Research **54**, 211–222 (1994)
23. Alpert, M., Kotsaftis, A., Pouget, E.R.: Speech fluency and schizophrenic negative signs. Schizophrenia Bulletin **23**(2), 171–177 (1997)
24. Stassen, H., Albers, M., Püschel, J., Scharfetter, C., Tewesmeier, M., Woggon, B.: Speaking behavior and voice sound characteristics associated with negative schizophrenia. Journal of psychiatric research **29**, 277–296 (1995)
25. Alpert, M., Rosenberg, S.D., Pouget, E.R., Shaw, R.J.: Prosody and lexical accuracy in flat affect schizophrenia. Psychiatry Research **97**(2), 107–118 (2000)
26. Rapčan, V., D'Arcy, S., Yeap, S., Afzal, N., Thakore, J., Reilly, R.: Acoustic and temporal analysis of speech: A potential biomarker for schizophrenia. Medical engineering & physics **32**, 1074–1079 (2010)
27. Bernardini, F., Lunden, A., Covington, M., Broussard, B., Halpern, B., Alolayan, Y., Crisafio, A., Pauselli, L., Balducci, P.M., Capulong, L., Attademo, L., Lucarini, E., Salierno, G., Natalicchi, L., Quartesan, R., Compton, M.T.: Associations of acoustically measured tongue/jaw movements and portion of time speaking with negative symptom severity in patients with schizophrenia in italy and the united states. Psychiatry Research **239**, 253–258 (2016)
28. Compton, M., Lunden, A., Cleary, S., Pauselli, L., Alolayan, Y., Halpern, B., Broussard, B., Crisafio, A., Capulong, L., Balducci, P., Bernardini, F., Covington, M.: The aprosody of schizophrenia: Computationally derived acoustic phonetic underpinnings of monotone speech. Schizophrenia Research **197**, 392–399 (2018)

29. Chakraborty, D., Yang, Z., Tahir, Y., Maszczyk, T., Dauwels, J., Thalmann, N., Zheng, J., Maniam, Y., Amirah, N., Tan, B., Lee, J.: Prediction of negative symptoms of schizophrenia from emotion related low-level speech signals. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 6024–6028 (2018)
30. Chakraborty, D., Xu, S., Yang, Z., Chua, Y., Tahir, Y., Dauwels, J., Thalmann, N., Tan, B., Lee, J.: Prediction of negative symptoms of schizophrenia from objective linguistic, acoustic and non-verbal conversational cues. In: 2018 International Conference on Cyberworlds (CW), pp. 280–283 (2018)
31. Tahir, Y., Chakraborty, D., Dauwels, J., Magnenat-Thalmann, N., Thalmann, D., Lee, J.: Non-verbal speech analysis of interviews with schizophrenic patients. 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 5810–5814 (2016)
32. Gosztolya, G., Bagi, A., Szalóki, S., Szendi, I., Hoffmann, I.: Identifying schizophrenia based on temporal parameters in spontaneous speech. In: INTERSPEECH, pp. 3408–3412 (2018)
33. Chhabra, S., Badcock, J., Maybery, M., Leung, D.: Voice identity discrimination in schizophrenia. Neuropsychologia **50**, 2730–5 (2012)
34. Zhang, J., Pan, Z., Gui, C., Zhu, J., Cui, D.: Clinical investigation of speech signal features among patients with schizophrenia. Shanghai Archives of Psychiatry **28**(2), 95–102 (2016)
35. Titze, I., Riede, T., Mau, T.: Predicting achievable fundamental frequency ranges in vocalization across species. PLoS Computational Biology **12**(6) (2016)
36. Nordström, H.: Emotional communication in the human voice. PhD thesis (2019)
37. O'Shea, K., Nash, R.: An introduction to convolutional neural networks. arXiv preprint arXiv:1511.08458 (2015)
38. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167 (2015)
39. Bjorck, N., Gomes, C.P., Selman, B., Weinberger, K.Q.: Understanding batch normalization. In: Advances in Neural Information Processing Systems, pp. 7694–7705 (2018)
40. Maas, A.L., Hannun, A.Y., Ng, A.Y.: Rectifier nonlinearities improve neural network acoustic models. In: Proc. Icml, vol. 30, p. 3 (2013)
41. Zeiler, M.D., Ranzato, M., Monga, R., Mao, M., Yang, K., Le, Q.V., Nguyen, P., Senior, A., Vanhoucke, V., Dean, J., *et al.*: On rectified linear units for speech processing. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 3517–3521 (2013). IEEE
42. Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., Liu, T., Wang, X., Wang, G., Cai, J., *et al.*: Recent advances in convolutional neural networks. Pattern Recognition **77**, 354–377 (2018)
43. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., *et al.*: Imagenet large scale visual recognition challenge. International journal of computer vision **115**(3), 211–252 (2015)
44. Zhang, C.L., Luo, J.H., Wei, X.S., Wu, J.X.: In defense of fully connected layers in visual representation transfer. In: Pacific Rim Conference on Multimedia, pp. 807–817 (2017). Springer
45. Sermanet, P., Kavukcuoglu, K., Chintala, S., Lecun, Y.: Pedestrian detection with unsupervised multi-stage feature learning. In: Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition, pp. 3626–3633. IEEE Computer Society, Oregon (2013)
46. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems, pp. 5998–6008 (2017)
47. Woo, S., Park, J., Lee, J.Y., So Kweon, I.: Cbam: Convolutional block attention module. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 3–19 (2018)
48. Association, A.P., *et al.*: Diagnostic and Statistical Manual of Mental Disorders (DSM-5®). American Psychiatric Pub, United States (2013)
49. Park, D.S., Chan, W., Zhang, Y., Chiu, C.C., Zoph, B., Cubuk, E.D., Le, Q.V.: Specaugment: A simple data augmentation method for automatic speech recognition. arXiv preprint arXiv:1904.08779 (2019)
50. Da, K.: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
51. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in pytorch. In: NIPS Workshop (2017)
52. Kinnunen, T., Hautamäki, V., Fränti, P.: On the use of long-term average spectrum in automatic speaker recognition. In: 5th Internat. Symposium on Chinese Spoken Language Processing (ISCSLP'06), pp. 559–567 (2006)
53. Yenigalla, P., Kumar, A., Tripathi, S., Singh, C., Kar, S., Vepa, J.: Speech emotion recognition using spectrogram and phoneme embedding, pp. 3688–3692 (2018)
54. Sundberg, J., Salomão, G.L., Scherer, K.R.: Analyzing emotion expression in singing via flow glottograms, long-term-average spectra, and expert listener evaluation. Journal of voice : official journal of the Voice Foundation (2019)
55. Abdel-Hamid, L.: Egyptian arabic speech emotion recognition using prosodic, spectral and wavelet features. Speech Communication **122**, 19–30 (2020)
56. Liu, J.M., You, M.Y., Li, G.Z., Wang, Z., Xu, X.H., Qiu, Z., Xie, W.J., An, C., Chen, S.L.: Cough signal recognition with gammatone cepstral coefficients, pp. 160–164 (2013)
57. Magi, C., Pohjalainen, J., Bäckström, T., Alku, P.: Stabilised weighted linear prediction. Speech Communication **51**(5), 401–411 (2009)
58. Pohjalainen, J., Saeidi, R., Kinnunen, T., Alku, P.: Extended weighted linear prediction (xlp) analysis of speech and its application to speaker verification in adverse conditions. In: INTERSPEECH (2010)
59. Jouni Pohjalainen, P.A. Carlo Magi: Enhancing noise robustness in automatic speech recognition using stabilized weighted linear prediction (swlp). In: ISCA Tutorial and Research Workshop (ITRW) on Speech Analysis and Processing for Knowledge Discovery (2008)
60. Zhang, J., Yang, S., Wang, X., Tang, M., Yin, H., He, L.: Automatic hypernasality grade assessment in cleft palate speech based on the spectral envelope method. Biomedical Engineering / Biomedizinische Technik

65(1), 73–86 (2020)
61. Cohen, A., Alpert, M., Nienow, T., Dinzeo, T., Docherty, N.: Computerized measurement of negative symptoms in schizophrenia. Journal of psychiatric research **42**, 827–36 (2008)
62. Tjaden, K., Sussman, J.E., Liu, G., Wilding, G.: Long-term average spectral (ltas) measures of dysarthria and their relationship to perceived severity. Journal of medical speech-language pathology **18**(4), 125 (2010)
63. Valero, X., Alias, F.: Gammatone cepstral coefficients: Biologically inspired features for non-speech audio classification. IEEE Transactions on Multimedia **14**(6), 1684–1689 (2012)
64. Deller, J.R., Hansen, J.H.L., Proakis, J.G.: Discrete-time processing of speech signals. In: Institute of Electrical and Electronics Engineers (2000)
65. Rah, D., Ko, Y.I., Lee, C., Kim, D.W.: A noninvasive estimation of hypernasality using a linear predictive model. Annals of biomedical engineering **29**, 587–94 (2001)
66. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Proceddings of the Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
67. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: arXiv Preprint arXiv, pp. 1409–1556 (2014)
68. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778 (2016)
69. Huang, G., Liu, Z., van der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2261–2269 (2017)
70. Chollet, F.: Xception: Deep learning with depthwise separable convolutions. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1800–1807 (2017)
71. Stolar, M.N., Lech, M., Bolia, R.S., Skinner, M.: Real time speech emotion recognition using rgb image classification and transfer learning. In: 2017 11th International Conference on Signal Processing and Communication Systems (ICSPCS), pp. 1–8 (2017)
72. Beckmann, P., Kegler, M., Saltini, H., Cernak, M.: Speech-vgg: A deep feature extractor for speech processing. arXiv preprint arXiv **1910.09909** (2019)
73. Ford, L., Tang, H., Grondin, F., Glass, J.R.: A deep residual network for large-scale acoustic scene analysis. In: INTERSPEECH (2019)
74. Li, C.Y., Vu, N.T.: Densely connected convolutional networks for speech recognition. In: Speech Communication; 13th ITG-Symposium, pp. 1–5 (2018)
75. Xu, K., Feng, D., Mi, H., Zhu, B., Wang, D., Zhang, L., Cai, H., Liu, S.: Mixup-based acoustic scene classification using multi-channel convolutional neural network. In: Pacific Rim Conference on Multimedia, pp. 14–23. Springer, Cham (2018)
76. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 618–626 (2017)
77. Barabassy, A., Szatmári, B., Laszlovszky, I., Németh, G.: Negative Symptoms of Schizophrenia: Constructs, Burden, and Management, (2018)
78. Hales RE, Y.S.: The American Psychiatric Publishing Textbook of Psychiatry. 5th Edition. American Psychiatric Pub, America (2008)
79. Schaerlaeken, S., Grandjean, D.: Unfolding and dynamics of affect bursts decoding in humans. PLOS ONE **13**(10), 1–21 (2018)
80. Wang, X., Chen, X., Cao, C.: Human emotion recognition by optimally fusing facial expression and speech feature. Signal Processing: Image Communication **84**, 115831 (2020)
81. Hoffman, R.R., Palermo, D.S.: Cognition and the Symbolic Processes: Applied and Ecological Perspectives. Psychology Press, United Kingdom (2014)
82. Rani, B.M.S., Rani, A.J., Ravi, T., Sree, M.D.: Basic fundamental recognition of voiced unvoiced and silence region of a speech. International Journal of Engineering and Advanced Technology (IJEAT) **4**, 83–86 (2014)
83. Weinberger, D.R., Marenco, S.: Schizophrenia as a Neurodevelopmental Disorder, pp. 326–348 (2003)
84. Tuckova, J., Komarek, V.: Effectiveness of speech analysis by self-organizing maps in children with developmental language disorders. Neuroendocrinology Letters **29**(6), 939 (2008)
85. Grill, P., Tuckova, J.: Formant analysisis—vowel detection of children with developmental dysphasia. Digital Technologies (2010)
86. Vranova, M., Tuckova, J., Kyncl, M., Grill, P., Komarek, V., *et al.*: Mri abnormalities of speech and computerised processing of speech of children with developmental dysphasia. In: In AKL Congress, Tabor, Czech Republic (2011)
87. Grill, P., Tuckova, J.: Formants application to diagnose of children with developmental dysphasia. In TBMI VŠB, 98–101 (2011)
88. Kohnert, K., Windsor, J., Ebert, K.D.: Primary or "specific" language impairment and children learning a second language. Brain and language **109**(2-3), 101–111 (2009)
89. Grela, B., Collisson, B., Arthur, D.: Language processing in children with language impairment. The handbook of psycholinguistic and cognitive processes: Perspectives in communication disorders, 373 (2011)
90. CLAHSEN, H.: The grammatical characterization of developmental dysphasia. Linguistics **27**(5), 897–920 (1989)
91. Gopnik, M., Dalalakis, J., Fukuda, S., Fukuda, S.E., Kehayia, E.: Genetic language impairment : Unruly grammars. (1996)
92. Grill, P., Tucková, J.: Speech databases of typical children and children with sli. PLoS ONE **11** (2016)
93. Grill, P., Tuckova, J.: Classification and Detection of Specific Language Impairments in Children Based on their Speech Skills, p. 24 (2017)
94. Grill, P.: Classification of children with sli through their speech utterances. In: World Congress on Medical Physics and Biomedical Engineering 2018, pp. 441–447. Springer, Singapore (2019)
95. Ramarao, D., Singh, C., Shahnawazuddin, S., Adiga, N., Pradhan, G.: Detecting developmental dysphasia in

children using speech data. In: 2018 International Conference on Signal Processing and Communications (SPCOM), pp. 100–104 (2018)
96. Slogrove, K.J., van der Haar, D.: Specific language impairment detection through voice analysis. In: Abramowicz, W., Klein, G. (eds.) Business Information Systems, pp. 130–141. Springer, Cham (2020)
97. Sharma, Y., Singh, B.K.: Prediction of specific language impairment in children using speech linear predictive coding coefficients. In: 2020 First International Conference on Power, Control and Computing Technologies (ICPC2T), pp. 305–310 (2020)
98. Reddy, M.K., Alku, P., Rao, K.S.: Detection of specific language impairment in children using glottal source features. IEEE Access **8**, 15273–15279 (2020)

Figures

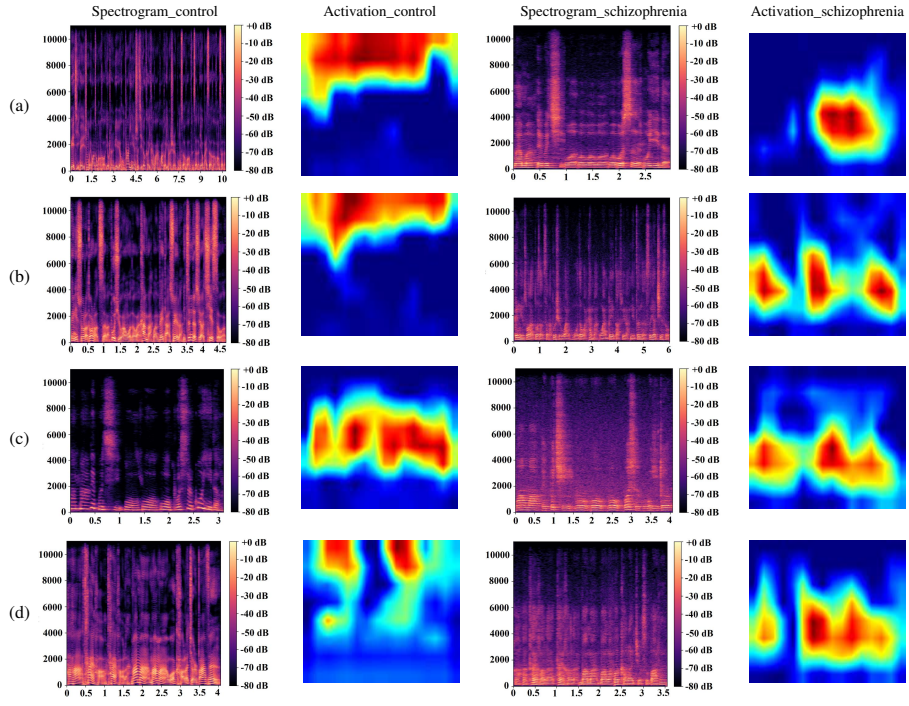

Figure 1 The spectrogram and corresponding activation map of normal speech and schizophrenia speech in four emotions. (a) The spectrogram and corresponding activation map of normal speech and schizophrenia speech in calm emotion. (b) The spectrogram and corresponding activation map of normal speech and schizophrenia speech in anger emotion. (c) The spectrogram and corresponding activation map of normal speech and schizophrenia speech in fear emotion. (d) The spectrogram and corresponding activation map of normal speech and schizophrenia speech in happiness emotion.

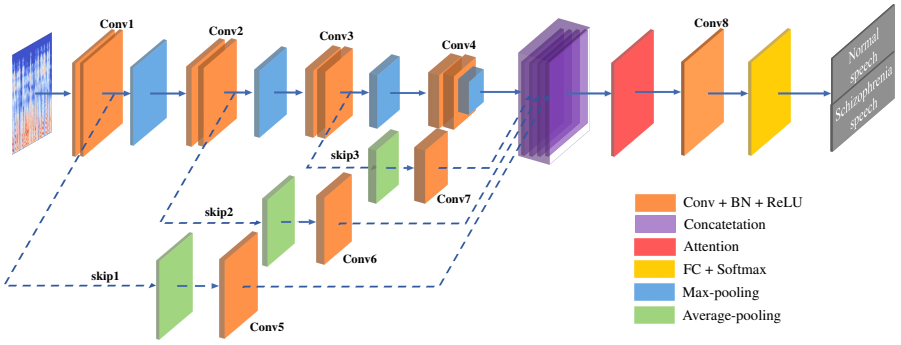

Figure 2 The architecture of Sch-net for automatic schizophrenia speech detection

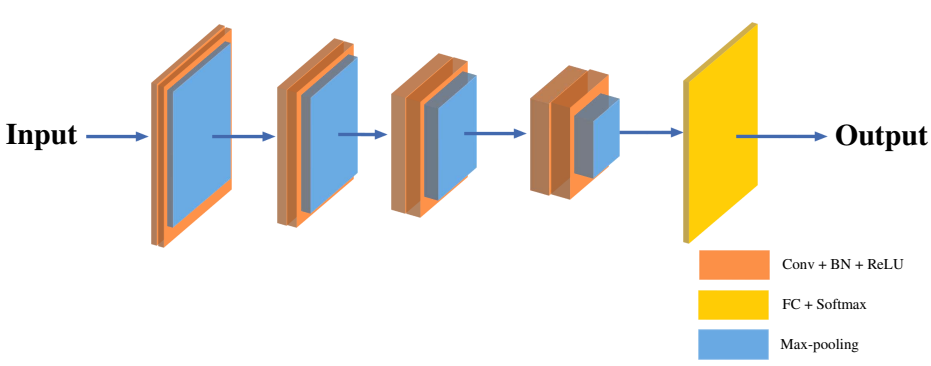

Figure 3 The diagram of the backbone network of Sch-net

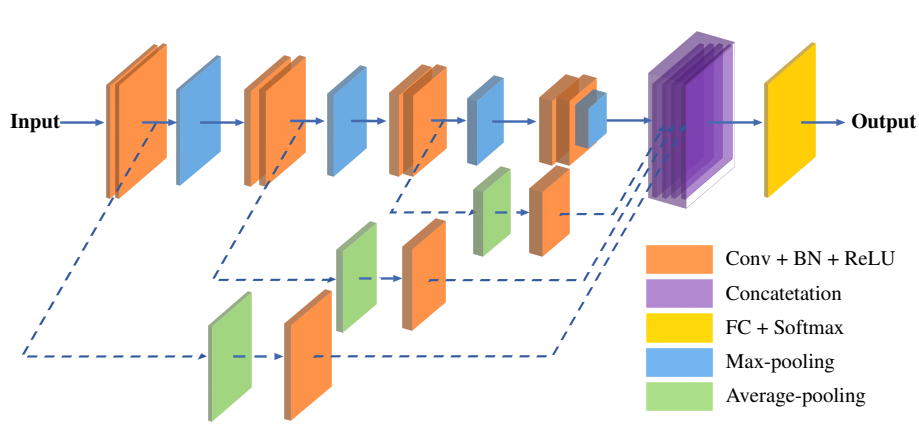

Figure 4 The diagram of the backbone network + skip connections of Sch-net

Figures

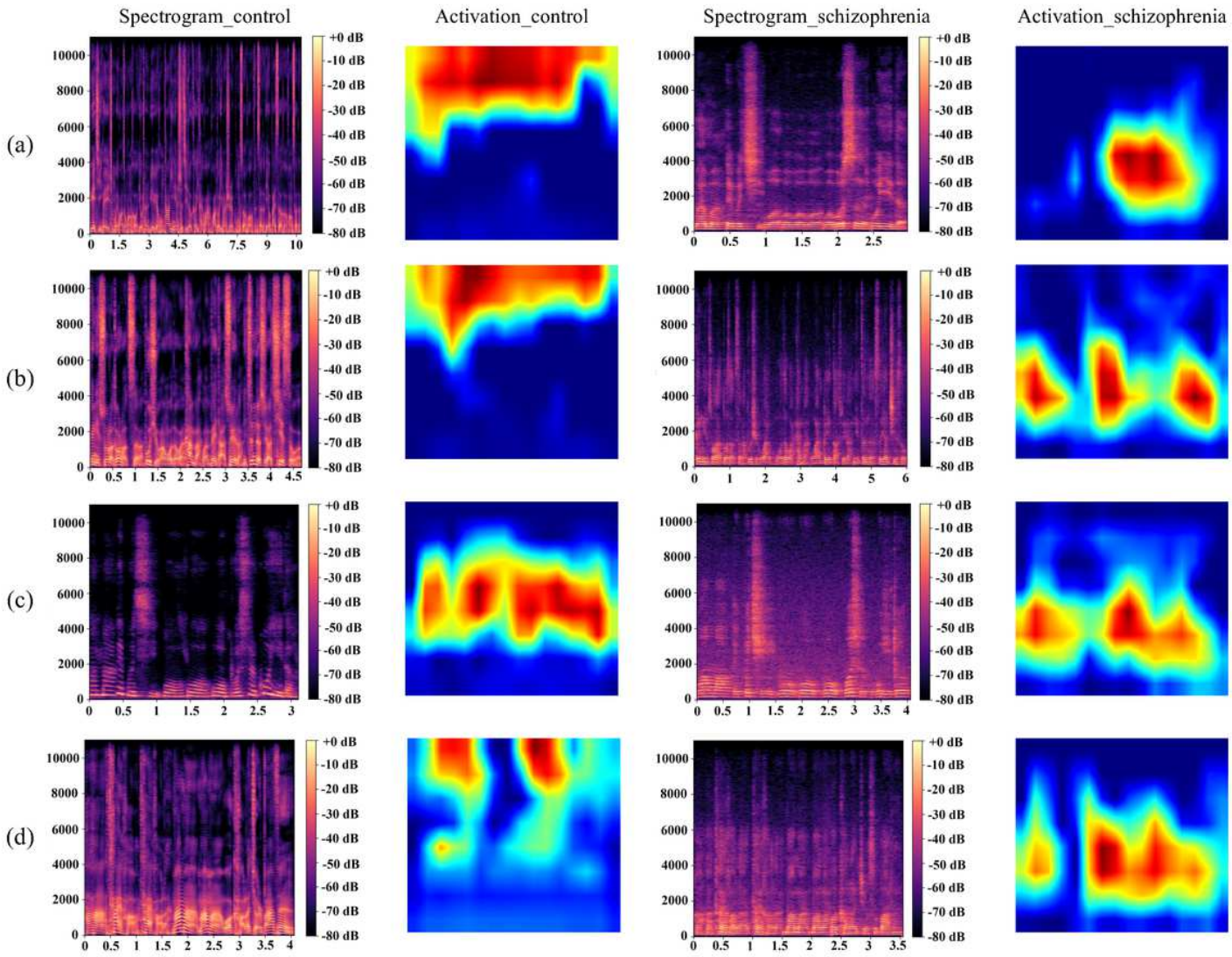

Figure 1

The spectrogram and corresponding activation map of normal speech and schizophrenia speech in four emotions. (a) The spectrogram and corresponding activation map of normal speech and schizophrenia speech in calm emotion. (b) The spectrogram and corresponding activation map of normal speech and schizophrenia speech in anger emotion. (c) The spectrogram and corresponding activation map of normal speech and schizophrenia speech in fear emotion. (d) The spectrogram and corresponding activation map of normal speech and schizophrenia speech in happiness emotion.

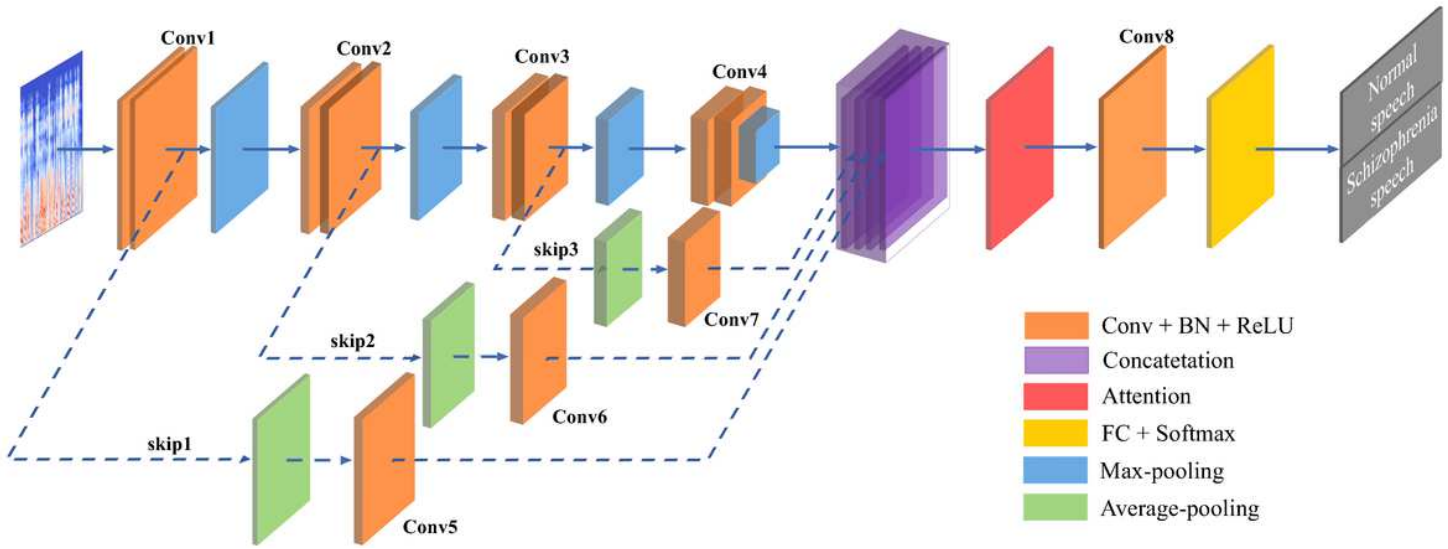

Figure 2

The architecture of Sch-net for automatic schizophrenia speech detection

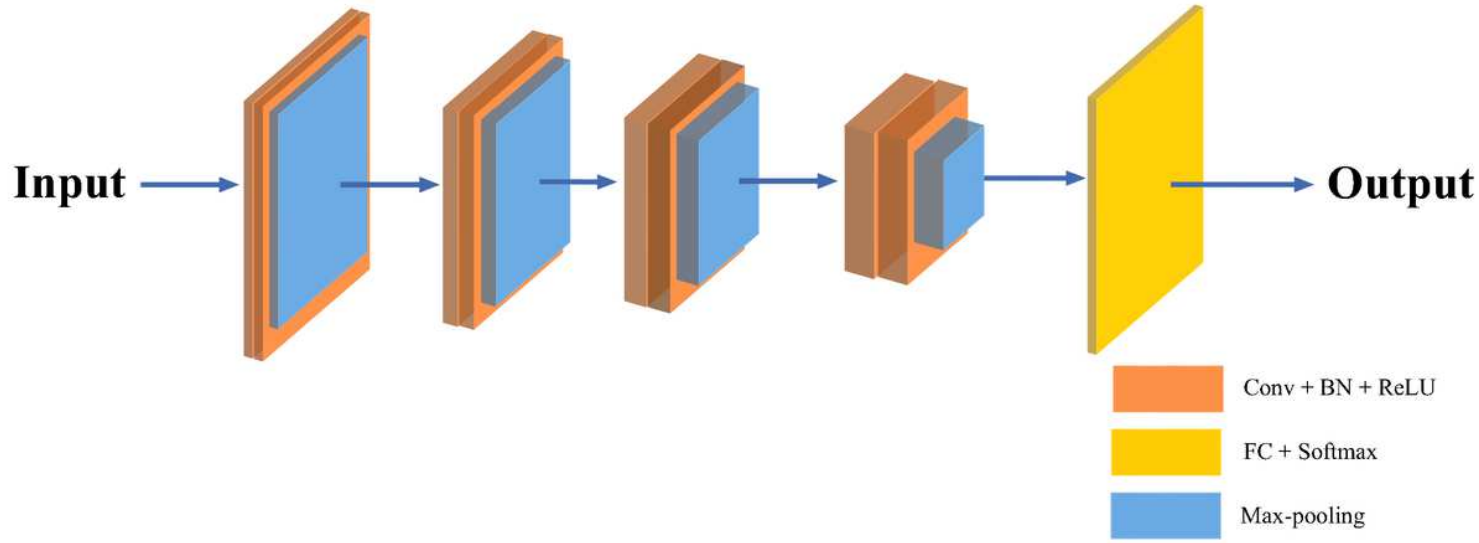

Figure 3

The diagram of the backbone network of Sch-net

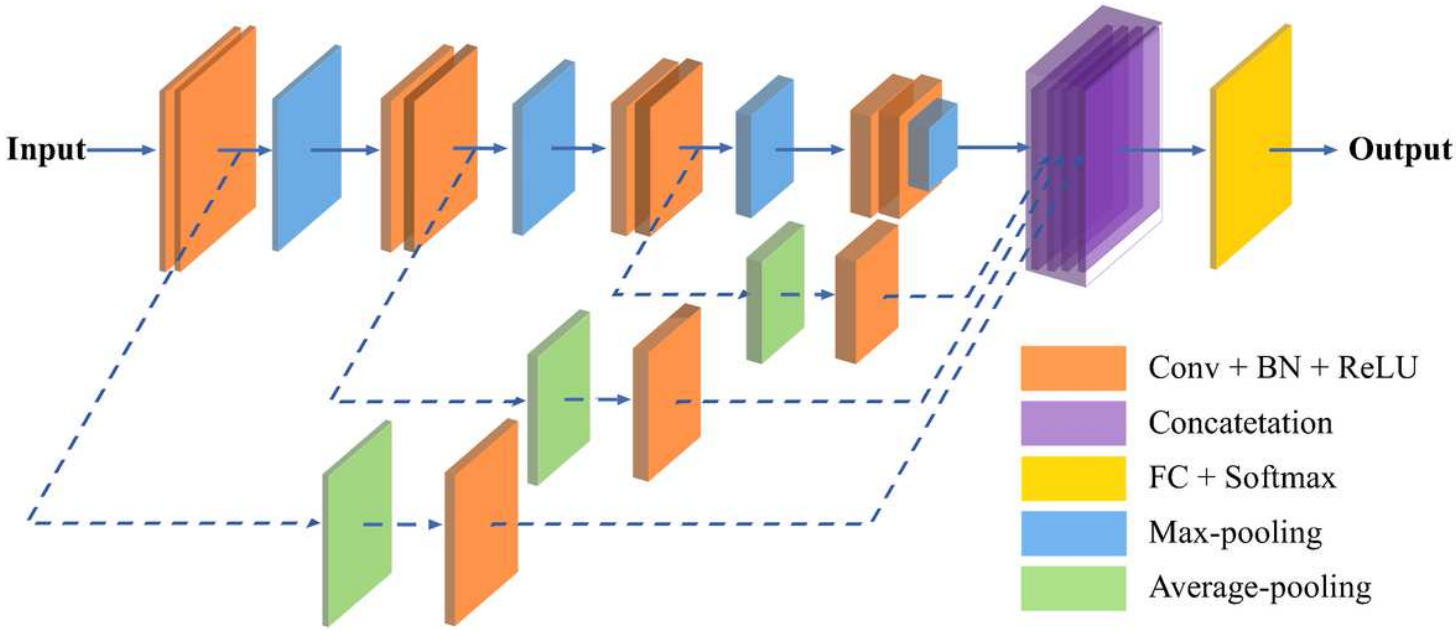

Figure 4

The diagram of the backbone network + skip connections of Sch-net