Pay More Attention With Fewer Parameters: A Novel 1-D Convolutional Neural Network for Heart Sounds Classification

Yunqiu Xu[1], Bin Xiao[1*], Xiuli Bi[1], Weisheng Li[1], Junhui Zhang[2], Xu Ma[3]

[1]Chongqing Key Laboratory of Computational Intelligence, Chongqing University of Posts and Telecommunications, Chongqing 400065, China
[2]The First Affiliated Hospital of Chongqing Medical University, Chongqing 400042, China
[3]Human Genetics Resource Center, National Research Institute for Family Planning, Beijing 100081, China

Abstract

The cardiovascular disease (CVD) is one of the major causes of mortality worldwide. Auscultation of heart sounds or phonocardiograms (PCGs) analysis, which is an efficient and non-invasive way, has been shown to be promising and played an important role in preliminary CVD diagnosis. In this study, a deep learning-based PCG classification method is proposed, which is mainly comprised three steps: pre-processing, PCG patches classification using a novel 1-D deep convolutional neural network (CNN), and final predicting of PCG recordings based on the patch-level results. In order to maximize the information flow within the CNN, a block-stacked style architecture with clique blocks is employed, and in each clique block a bidirectional connection structure is utilized. Using the stacked blocks, the proposed CNN achieves both spatial and channel attention, which leads a superior classification performance. Besides, a novel separable convolution with inverted bottleneck is introduced to efficiently decouple features' dependency between spatial and channel-wise dependency of features. Experiments on PhysioNet/CinC 2016 reveal a superior classification performance and the advantage in parameter efficiency of the proposed method comparing to state-of-the-art methods.

1. Introduction

A heart is a vital organ of body and the cardiovascular disease (CVD) is one of the leading causes of mortality worldwide. Many pathological conditions of the cardiovascular system are reflected in some heart-related signals, such as the heart sound signals (i.e., phonocardiograms, PCGs). Nevertheless, the accuracy of auscultation depends on the skills and subjective experiences of the physicians which are obtained from a long physician experience [1]. Therefore, an objective and automatic method for heart sound signals analysis is needed. Nowadays, automatic heart sound classification, which has the potential to screen for pathologies in a variety of clinical applications enabling reduction of costly and time consuming manual examinations, is becoming a promising research field based on the techniques of biological signal processing and artificial intelligence [2].

In recent years, deep learning has achieved tremendous success in many practical tasks owing to its amazing feature representation power. Convolutional Neural Networks (CNNs), as one of the typical deep learning architectures combing feature extraction and classification together, are now commonly used in many fields [3]. More recently, some CNN-based PCG classification methods [4-6] are proposed, whereas they usually have complicated pre-processing and post-processing or they are not sufficiently expressive to learn the complex pattern of heart sound (e.g. simple architecture, a small number of filters and layers, etc.). Therefore, a novel 1-D deep CNN structure for PCG classification is proposed in this paper. The proposed CNN is a block-stacked style architecture which enhances the information flow of the CNN using bidirectional connections and achieves the state-of-the-art performance with exceedingly fewer parameters.

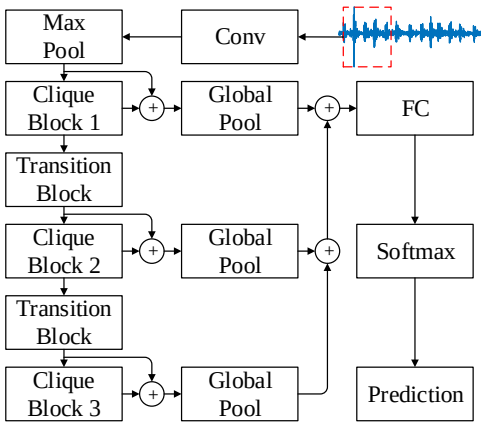

Figure 1. The pipeline of the proposed CNN architecture. The symbol "+" within each circle represents a channel-wise concatenating operation of feature maps.

* corresponding author: (xiaobin@cqupt.edu.cn)

 ISSN: 2325-887X DOI: 10.22489/CinC.2018.072

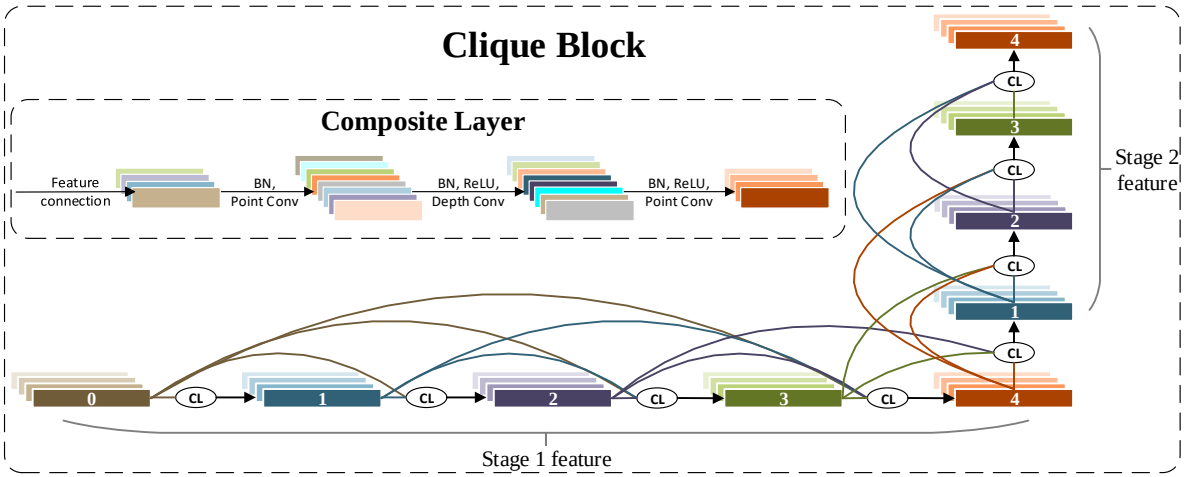

Figure 2. The propagations of Clique block with 4 composite layers are depicted. Every feature maps are both the input and the output of other feature maps, excepting the input node of blocks. Each ellipse noted with **CL** denotes a composite layer, and expansion ratio of 2 are shown in the composite layer as a concise example.

2. Methods

The main pipeline of the proposed method can be divided into three parts: pre-processing, automatic classification of PCG patches using a novel 1-D deep CNN, merging the patch-level predictions to recording-level results based on a majority voting decision strategy.

2.1 Pre-processing

The 2-D representation of raw input signals is the most common choice in the majority of state-of-the-art audio classification algorithms, e.g., Mel-frequency cepstral coefficients (MFCCs), Power Spectral Density (PSD), and so on. Although the 2-D representation represents acoustically meaningful patterns well, it requires an extra transforming procedure and a set of hyper-parameters. Thus, the 1-D raw waveform PCGs are utilized as the input in this study. The raw PCG signals are resampled to 2000 Hz, and this is followed by the band filters which are used to remove the high frequency noises. Since the structural characteristics of our proposed CNN architecture, the size of input needs to be fixed. Consequently, the PCG recordings are segmented into several 3-seconds long patches with a stride of 1 second. More importantly, the segmentation of PCG is able to enlarge the scale of the training set, which is crucial to deep learning-based methods.

2.2 CNN architecture

Clique blocks. The segmented PCG patches are fed into the proposed CNN model directly. Our proposed model is partly inspired by the CliqueNet [7] combining both recurrent structure and spatial attention mechanism. As illustrated in Figure 1, the proposed CNN model consists three clique blocks and two transition blocks. The feature maps within each clique block are connected bidirectionally by several composite layers (shown in Figure 2). And the propagations of each block can be divided into two stages. The propagations in stage 1 are similar to the DenseNet [8] that all the layers are densely connected unidirectionally. Then the extra updating of these layers is operated alternately in stage 2 to make sure that each layer is capable to receive the feedback information from the most lately updated layers. Only the output features of stage 2 are supposed to reach next clique block through adjoining transition block. In order to improve the classification accuracy by using the multi-scale features, replicas of the output features of every clique blocks are firstly compressed to half of their original dimensions of channels by point-wise convolution layers. Then, the compressed output features of each clique block are connected with the input features of this block. After that, the connected features are fed into global pooling layers severally to achieve a squeezed multi-scale representation of corresponding clique block. Lastly, the squeezed features of different scales are merged together and ended with a fully-connected layer with softmax to realize the accurate classification.

Separable convolutions with inverted bottleneck. Based on the hypothesis that the mapping of cross-channel correlations and spatial correlations in the feature maps can be entirely decoupled, the separable convolution with inverted bottleneck is introduced in each composite layer to optimize the efficiency of parameters and memories with classification performance improvement. As illustrated in Figure 2, the separable convolution with inverted bottleneck can be separated into three steps roughly: 1) expanding the low dimension features in a higher dimension; 2) extracting features in this higher dimension by depth-wise convolutions; 3) projecting the high dimension features back to the low dimension. It should be noted that, the Batch Normalization (BN) is employed before each convolution, while the Rectified Linear Unit (ReLU) is just adopted before depth-wise convolution and the second point-wise convolution. This type of inverted bottleneck structure has a strong feature representation ability, and it is an inversed structure contrast to the hourglass shaped bottleneck structure used in the

DenseNet [8] and the ResNet [9]. The most similar work is the inverted residual bottleneck introduced in the MobileNet V2 [10], whereas the bidirectional feature map connection is employed in our CNN to enhance the information flow instead of using the residual connection.

Transition blocks with attention mechanism. Connecting all the layers of the network in a single block is inherently memory demanding, i.e., there will be $L(L+1)$ inter-layer connections when the network is consisted of L layers. In order to overcome this drawback and to obtain multi-scale features, the spatial size of feature maps need to be reduced after different clique blocks. Therefore, the transition blocks are designed to reduce the spatial size by equipping average pooling layers with size of 2. In our CNN model, we insert a transition block between two neighbouring clique blocks. Moreover, as depicted in Figure 3, an attention mechanism [11] is utilized in each transition block to perform dynamic channel-wise feature recalibration. Despite a slightly parameter increase will be introduced, it reweights the feature maps in each transition block for ensuring that more useful features can be exploited efficiently by the subsequent layers.

Implementation details. Before entering the first clique block, a convolution layer with filter size of 7, stride of 2, and a max pooling layer with size of 3, stride of 2 are employed to extract the initial low-level features and reduce the spatial size of feature maps. In each clique block, 5 composite layers are equipped. Each of them connects all the previous features as input, and outputs features with 12 channels. And the depth-wise convolution layers within every composite layer use the small size convolution layers (with size of 3, stride of 1) to extract features. Moreover, an expansion ratio of 6 is employed in every composite layer (i.e., the number of channels in intermediate feature maps is 6 times than that of the input and output feature maps). As for the fully-connected layers within each transition block, the number of intermediate nodes is invariable rather than compressing it into a lower dimension like [11].

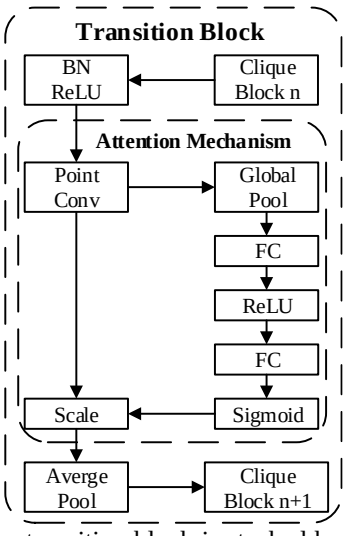

Figure 3. The transition block is stacked by BN, ReLU, attention mechanism and average pooling layer.

2.3 Majority voting

The ultimate goal for PCG classification is to classify the PCG recordings into different categories. Thus, the majority voting is introduced to transform the patch-level classification results into PCG recording-level prediction results. In this strategy, the number of predicted patches will be counted. Once the number of normal predicted patches is larger than abnormal in one PCG recording, this recording will be labelled as normal, and vice versa. Moreover, if the numbers of patches for both categories are equal in one recording, the mean of the raw predicted probability of every patch will be compared to give the final prediction of this PCG recording.

3. Experiments

3.1 CNN training

We conduct the experiments on PhysioNet/CinC 2016 [12]. Since only the training set is available (665 abnormal and 2488 normal recordings), the 10-fold cross validation is adopted to evaluate the classification performance. We train the CNN model from scratch without any data augmentation. A weighted cross entropy with the rate 0.25 to 1 (normal to abnormal) is adopted as the loss function on account of the class imbalance. Stochastic gradient descent (SGD) with 0.9 Nesterov momentum is chosen as the optimizer, and a mini-batch size of 64 is used. Weight decay of 10^{-4} and dropout layers with rate of 0.1 are also applied to prevent overfitting. In our training procedure, the model is trained for 40 epochs with early stop, and the initial learning rate is set to 0.1 while the learning rate decay of 0.1 is implemented at both epoch 20 and 30.

3.2 Results

To compare with other algorithms, several evaluation indicators are introduced: accuracy (*Acc*), sensitivity (*Se*), specificity (*Sp*), overall score (*Score*), which are defined as follows:

$$Acc = \frac{TP + TN}{TP + FP + TN + FN} \quad (1)$$

$$Se = \frac{TP}{TP + FN} \quad (2)$$

$$Sp = \frac{TN}{FP + TN} \quad (3)$$

$$Score = \frac{Se + Sp}{2} \quad (4)$$

where *TP*, *TN*, *FP*, *FN* are given as follows:

- True positive (*TP*): The number of correctly predicted abnormal recordings.
- True negative (*TN*): The number of correctly predicted normal recordings.
- False positive (*FP*): The number of incorrectly

predicted normal recordings.

- False negative (*FN*): the number of incorrectly predicted abnormal recordings.

Meanwhile, the number of trainable parameters (*Params*) is also taken into consideration to measure the scale of CNN model.

Table 1 shows the experimental results in comparison with existing CNN-based algorithms in literatures [4-6]. Ryu *et al.* [4] used the Windowed-sinc Hamming filter algorithm to remove the irrelevant noises from the raw PCG signals before feeding PCG patches into a 1-D CNN for classification. In the method of PSD-CNN [5] and MFCC-CNN [6], the 2-D representative of PCGs were adopted as the inputs. The difference is that [6] used the MFCCs to transform the raw signals into 2-D feature rather than PSDs which are employed in [5]. From Table 1, we can find out that our proposed method yields the highest *Se* that indicates the percent of correctly classified PCG as reflecting abnormal heart function. Additionally, the proposed method also obtains the best performance on *Score* with minimal trainable parameters. In spite of slight decreases in *Acc* and *Sp*, our proposed model uses 65 times fewer parameters than MFCC-CNN [6]. The superior performance of the proposed method is mainly owing to the recurrent structure that maximally enhances information flow and reuses the feature maps. Moreover, it is also on account of the novel separable convolution which efficiently extracts features in a parameter-saving way by decoupling the spatial and channel-wise features. In general, the experiment results reveal a promising light weight model that we proposed for classifying normal and abnormal heart sounds.

Table 1. Evaluation results for the proposed method in comparison with existing CNN-based algorithms. Results that surpass all competing methods are in **bold**.

Evaluation criteria	1-D CNN [4]	PSD-CNN [5]	MFCC-CNN [6]	Proposed method
Acc	0.8933	0.8905	**0.9331**	0.9321
Sp	0.9282	0.9102	**0.9619**	0.9512
Se	0.7608	0.8150	0.8271	**0.8581**
Score	0.8445	0.8626	0.8945	**0.9046**
Params	**0.19M**	0.24M	12.41M	**0.19M**

4. Conclusion

In this paper, we developed a novel 1-D CNN architecture for PCG classification. This architecture is designed to efficiently reuse the feature maps with lower parameter consuming, and without any complex pre-processing or post-processing steps. Experiments conducted on PhysioNet/CinC 2016 demonstrate the effectiveness of the proposed method, which achieves state-of-the-art classification performance in a significant parameter-saving way. The promising performance of the proposed method makes us hopeful that with further improvement the feature reusing and minimization of the trainable parameters, the network may be suitable for embedded or mobile applications. Moreover, the environment noises will be taken into consideration in our future work to enhance the robustness of the proposed method for facing the practical clinical environment.

Acknowledgements

This work was partly supported by the National Science & Technology Major Project (2016YFC1000307-3) and the National Natural Science Foundation of China (61572092).

References

[1] Z. Jiang, S. Choi. A cardiac sound characteristic waveform method for in-home heart disorder monitoring with electric stethoscope. Expert Systems with Applications. 2006; 31(2): 286-98.

[2] J. Herzig, A. Bickel, A. Eitan, N. Intrator. Monitoring Cardiac Stress Using Features Extracted From S1 Heart Sounds. IEEE Transactions on Biomedical Engineering. 2015; 62(4): 1169-78.

[3] V. Sze, Y. H. Chen, T. J. Yang, J. S. Emer. Efficient Processing of Deep Neural Networks: A Tutorial and Survey. Proceedings of the IEEE. 2017; 105(12):2295-329.

[4] H. Ryu, J. Park, H. Shin. Classification of heart sound recordings using convolution neural network. In Computing in Cardiology Conference (CinC) 2016;43:1153-1156.

[5] T. Nilanon, J. Yao, J. Hao, S. Purushotham, Y. Liu. Normal / abnormal heart sound recordings classification using convolutional neural network. In Computing in Cardiology Conference (CinC) 2016;43:585-588.

[6] J. Rubin, R. Abreu, A. Ganguli, S. Nelaturi, I. Matei, K. Sricharan. Classifying heart sound recordings using deep convolutional neural networks and mel-frequency cepstral coefficients. In Computing in Cardiology Conference (CinC) 2016;43:813-816.

[7] Y. Yang, Z. Zhong, T. Shen, Z. Lin. Convolutional Neural Networks with Alternately Updated Clique. arXiv preprint arXiv:180210419. 2018.

[8] G. Huang, Z. Liu, L. v. d. Maaten, K. Q. Weinberger. Densely Connected Convolutional Networks. In Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017;2261-2269.

[9] K. He, X. Zhang, S. Ren, J. Sun. Deep Residual Learning for Image Recognition. In Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016;770-780.

[10] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L.-C. Chen. MobileNetV2: Inverted Residuals and Linear Bottlenecks. arXiv preprint arXiv:180104381. 2018.

[11] J. Hu, L. Shen, G. Sun. Squeeze-and-Excitation Networks. arXiv preprint arXiv:170901507. 2017.

[12] C. Liu, D. Springer, Q. Li, B. Moody, R. A. Juan, F. J. Chorro, et al. An open access database for the evaluation of heart sound algorithms. Physiological Measurement. 2016;37(12):2181-213.

Address for correspondence:
Bin Xiao
School of Computer Science
No. 2, Chongwen Road, Nan'an district, Chongqing, China
xiaobin@cqupt.edu.cn