SE-ECGNet: Multi-scale SE-Net for Multi-lead ECG Data

Jiabo Chen[1], Tianlong Chen[4],Bin Xiao[1], Xiuli Bi[1*], Yongchao Wang[1], Han Duan[1],Weisheng Li[1], Junhui Zhang[2*], Xu Ma[3]

[1] Chongqing Key Laboratory of Image Cognition, Chongqing University of Posts and Telecommunications, Chongqing, China
[2] The First Affiliated Hospital of Chongqing Medical University, Chongqing, China
[3] Human Genetics Resource Center, National Research Institute for Family Planning, Beijing, China
[4] Southwest Jiaotong University, Chengdu, China

Abstract

Cardiovascular disease is a life-threatening condition, and more than 20 million people die from heart disease. Therefore, developing an objective and efficient computer-aided tool for diagnosis of heart disease has become a promising research topic. In this paper, we design a multi-scale shared convolution kernel model. In this model, two paths are designed to extract the features of electrocardiogram(ECG). The two paths have different convolution kernel sizes, which are 3×1 and 5×1, respectively. Such multi-scale design enables the network to obtain different receptive fields and capture information at different scales, which significantly improves the classification effect. And squeeze-and-excitation networks (SE-Net) are added to every path of the model. The attention mechanism of SE-Net learns feature weights according to loss, which makes the effective feature maps have large weights and the ineffective or low-effect feature maps have small weights. Our team name is CQUPT_ECG. Our approach achieved a challenge validation score of 0.640, and full test score of 0.411, placing us 8 out of 41 in the official ranking.

1. Introduction

Cardiovascular diseases are very common in the world and threaten people's health. Electrocardiogram (ECG) is a non-invasive method of diagnosing cardiovascular diseases. Although it is an effective and harmless tool, the diagnosis of ECG relies on doctors' experience. Therefore, it is very important to develop a computer-aided ECG diagnostic technique.

In this paper, a multi-scale shared convolution kernel model is designed to predict cardiovascular diseases. It not only extracts the common characteristics of different leads by sharing convolution kernels, but also acquire multi-scale depth characteristics by different paths. In addition, to improve performance, the model also uses patient's information such as age and gender.

2. Challenge Data analysis

The training data published in challenge[1] comes from multiple sources. It includes four datasets which are from China, the United States, Russia and Germany. Statistically speaking, there are 43101 patient records in the training set with 111 symptoms. In this challenge, 27 classes will be scored and the rest will not be.

The data distribution in this dataset is extremely imbalanced. At the same time, some certain correlation are existed between diseases and age and gender. For example, the disease 427084000 is more common among people between 50 and 70 years of age, and the prevalence rate is higher in men than in women. The correlations between age and sex are also different for different diseases.

There are some similarities between different leads in electrocardiogram. The heart beats reacted on different leads are consistent, meanwhile the wave peak positions are almost the same.

In summary, the three points mentioned above are the most essential characteristics of electrocardiogram data, so making full use of these features is the key to build our model.

3. Method

The ECG records are divided into 10 seconds patches by sliding window. The model proposed in this paper is used to extract the features of ECG patches. The depth characteristics obtained by the model are combined with the age and gender characteristics of the patients. Finally, fully connected layers are used to classify and obtain the

record level results based on an one-vote-in-favour voting decision strategy. Figure 1 is the network diagram of the model. 50×1, 15×1, 5×1 and 3×1 represent the sizes of the convolution kernels. When ECG signals are processed, large convolution kernel is better in the early stage [2].

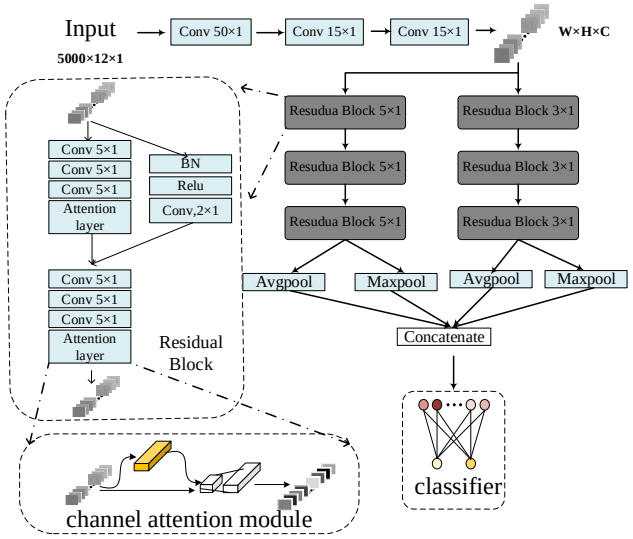

Figure 1. The network diagram.

3.1. Data processing

3.1.1 Patch segment

In this study, complicated preprocessing steps such as signal denoising are not adopted. Instead, the ECG records are divided into 10 seconds long patches through a sliding window. Note that 10 seconds is far longer than a complete cardiac cycle.

3.1.2 Data augmentation

The amount of some categories are too small, resulting in data imbalance. so the data augmentation operations are carried out.

We use two methods of data enhancement: random shift signals and random addition of Gaussian noise to the signals.

It should be noted that vertical flip is not suitable for data expansion here, because there are two types of left deviation of the electric axis and right deviation of the electric axis in the labels. If using vertical flip, it will cause the problem of label error.

3.1.3 Age and sex characteristics

As for the relationship between diseases and age and gender mentioned in section 2, this paper uses age and gender characteristics, which are helpful for the model to learn age and gender characteristics. Age and gender characteristics are combined with the depth characteristics extracted from the model, and then processed by the fully connected layers.

3.2. Shared convolution kernel

In general, a one-dimensional convolution kernel is often used to process the one-dimensional ECG signals. Different leads use different convolution kernels in this way.

But considering that the similarities between the different leads mentioned in the second section, this paper extracts the common characteristics of different leads by sharing convolution kernels.

As shown in the figure 2, the length, width and channel of the original ECG data are 5000,12 and 1 respectively. In this paper, the input ECG data is to obtain a two-dimensional image with channel of 1, constant length of 5000, and height of 12. After that, the operation of two-dimensional convolution whose kernel size is 50×1 is used to extract the features of the image. So the different leads of ECG share the same convolution kernel. Such a design can help the network make the best of the similarity of different leads.

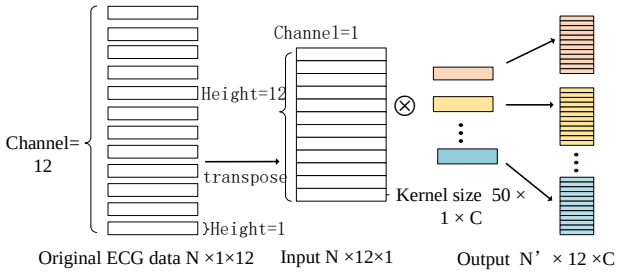

Figure 2. Different leads of ECG share the same convolution kernel.

3.3. Multi-scale residual module

The receptive field in convolutional neural network is important. In order to obtain the appropriate receptive field and increase the diversity of model features, a multi-scale feature network is designed, which is mainly carried out by parallel multi-branches. As shown in the figure 3, the convolutional kernel sizes of each branch are different, which are 3×1 and 5×1, respectively. After that, the multi-scale features extracted by different paths are concatenated.

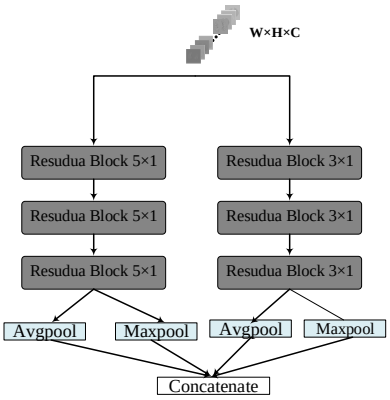

Figure 3. Multi path residual module.

Inspired by ResNet[3], each branch is composed of serial residual modules. Pre-activation is used for each convolution operation. In short, we build batch normalization layer(BN) and Relu activation layer before weight layers. The design of pre-activation has two benefits: one is more easier , the other is to improve the regularization of the models.

The dual pooling design is adopted at the end of each branch. The feature maps obtained from each branch are processed through maxpooling layer and avgpooling layer respectively. In general, only avgpooling layer is used to retain significant features, reduce feature dimensions, and increase kernel receptive field. The purpose of adding maxpooling layer here is to enable the network to better capture local abnormalities, because heart disease is often reflected in a local area of ECG. The characteristics obtained from different branches are merged with age and gender, and then sent to the fully connected layers for processing.

3.4. Attention mechanism

In order to realize the characteristics weighting between channels and ensure that more useful information will be sent to the subsequent layer for futher feature extraction, we introduces a lightweight channel attention mechanism from SE-Net[4].

Specifically, the attention mechanism can be divided into two steps: compression and activation. In the process of compression, global average pooling is used to compress the spatial dimension information of the feature maps into channel description vector. This can overcome the problem that the receptive field of convolution is too small to make use of contextual information. The element of the channel description vector can be calculated using formula 1:

$$Z_c = F_{sq}(x_c) = \frac{1}{L}\sum_{i=1}^{L} x_c(i) \quad (1)$$

Where L represents the size of the input feature maps X , and $X = [x_1, x_2, \dots, x_C]$. During activation, the sigmoid function is used to extract dependencies between channels, as shown in formula 2:

$$s = F_{ex}(z, w) = \sigma\big(g(z, w)\big) = \sigma(w_2\delta(w_1 z)) \quad (2)$$

Where σ stands for sigmoid function and δ stands for rectified linear unit(ReLU) function, $w_1 \in R^{\frac{C}{r}\times C}$ and $w_2 \in R^{C\times\frac{C}{r}}$ stand for two fully connected layer parameters. Parameter r is used to control the number of parameters in the network and increase model generalization performance. Finally, the output feature maps can be defined by formula 3:

$$L_c = F_{scale}(x_c, s_c) = x_c \cdot s_c \quad (3)$$

3.5. Loss function

Considering that the task of the challenge is a multiple lable classification task, this paper chooses the binary cross entropy. The calculation process of the loss function is shown in formula 4 where N is the number of categories, y_i is the result of determining whether the record has the ith disease, and the result is 0 or 1, and $\tilde{y_i}$ is the predicted probability value.

$$Loss = -\frac{1}{N}\sum_{i=1}^{N} y_i \cdot \log\tilde{y_i} + (1 - y_i) \cdot \log(1 - \tilde{y_i}) \quad (4)$$

At the same time, the number of samples of ECG data mentioned in the second section is unbalanced. Therefore, when calculating the loss function, different classes are considered to contribute different weights to the calculation of the loss function. The weight of class with more samples is lower, and that of class with less samples is higher. This allows the model to pay more attention to the classes with fewer samples during the training process. The weight calculation is shown in formula 5. In the formula, w_c is the corresponding weight of each class, and N_c is the number of samples of each class.

$$w_c = \frac{1}{log\ N_c} \quad (5)$$

3.6. A one-vote-in-favour voting decision strategy

In the data processing stage, we split ECG records into several patches. So how to convert the diagnostic results of the segmented fragments into the original record diagnosis result is critical. Given that cardiac anomalies are more localized in the local areas of electrocardiogram, a simple majority voting strategy should not be used, which would miss important local information. Therefore, a one-vote-in-favour voting decision strategy is designed. Specifically, a union set of the diagnostic results of all the patches is the diagnosis of the original record.

3.7 Hard example mining

The idea of hard example mining is to use a classifier to classify the samples and put the hard negative samples into the negative sample set and then continue training again.

There are easy samples and hard samples in a batch, among which easy samples have little impact on network because of the small loss value and the small gradient obtained, while the hard samples have greater impact.

Therefore, in the calculation of the loss, this paper only extracts the mean value of the largest 70 percents of the loss values in the current batch for the calculation.

3.8 Thresholds for each sample

In this paper, the thresholds of all classes are initially set as 0.5. When the prediction probability of a certain disease exceeded the threshold, the disease is confirmed.

But given that the threshold should be different for different diseases, it may also be different for different samples. Therefore, different thresholds are designed for different samples in this paper. For each sample, the strategy for selecting the threshold is to divide the maximum of the predicted probability for all classes by 5. The value 5 is the optimal one obtained by experimental comparison.

4. Result

5-fold cross-validation is carried out on the training set. Meanwhile, It is compared with ResNext[5] and ResNest. The results of cross-validation are shown in table 1.

Table 1. 5-fold cross validation experiment results on the training set.

Model	Fbeta	Gbeta	Challenge metric
ResNext	0.497	0.2568	0.5546
ResNest	0.4896	0.2554	0.5236
Proposed	0.5742	0.3054	0.6134

Our team name is CQUPT_ECG. Our approach achieved a challenge validation score of 0.640, and full test score of 0.411, placing us 8 out of 41 in the official ranking.

5. Conclusion

In this paper, a model with residual modules is proposed, which uses the features of different scales. At the same time, the similarities between different leads of electrocardiogram are utilized by using shared convolution kernels. A lightweight channel attention mechanism from SE-Net is introduced to realize the characteristics weighting between channels. According to the characteristics of electrocardiogram, a one-vote-in-favour voting decision strategy is designed in this paper.

Acknowledgments

This work was partly supported by the National Science & Technology Major Project (2016YFC1000307-3), the Scientific & Technological Key Research Program of Chongqing Municipal Education Commission (KJZD-K201800601), and the Chongqing research and innovation project of graduate students (CYS18245).

References

[1] Erick A. Perez Alday, Annie Gu, Amit Shah, Chad Robichaux, An-Kwok Ian Wong, Chengyu Liu, Feifei Liu, Ali Bahrami Rad, Andoni Elola, Salman Seyedi, Qiao Li, Ashish Sharma, Gari D. Clifford, Matthew A. Reyna, "Classification of 12-lead ECGs: the PhysioNet/Computing in Cardiology Challenge 2020," Physiol. Meas.

[2] Hannun A Y, Rajpurkar P, Haghpanahi M, et al., "Cardiologist level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network," Nat. Med., vol. 25, no. 1, pp. 65, Feb. 2019.

[3] Xie S, Girshick R, et al., "Aggregated residual transformations for deep neural networks," Proc. IEEE Conf. Comput. Vis. Pattern Recog., pp. 1492-1500. 2017.

[4] Hu J, Shen L, Sun G, "Squeeze-and-excitation networks," Proc. IEEE Conf. Comput. Vis. Pattern Recog., pp. 7132-7141. 2018.

[5] He K, Zhang X, Ren S, et al., "Deep residual learning for image recognition," Proc. IEEE Conf. Comput. Vis. Pattern Recog., pp. 770-778. 2016.

Address for correspondence:

Xiuli Bi
School of Computer Science
No.2, Chongwen Road, Nan'an district, Chongqing, China
bixl@cqupt.edu.cn

Junhui Zhang
The First Affiliated Hospital of Chongqing Medical University
No. 1 Yixueyuan Road, Yuzhong District, Chongqing, China
2275610878@qq.com