Artificial Intelligence In Medicine 106 (2020) 101856

Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

journal homepage: www.elsevier.com/locate/artmed

ECG-based multi-class arrhythmia detection using spatio-temporal attention-based convolutional recurrent neural network

Jing Zhang[a], Aiping Liu[a,*], Min Gao[b], Xiang Chen[a], Xu Zhang[a], Xun Chen[c,*]

[a] Department of Electronic Science and Technology, University of Science and Technology of China, Hefei 230027, China
[b] Department of Electrocardiogram, The First Affiliated Hospital of University of Science and Technology of China (Anhui Provincial Hospital), Hefei 230036, China
[c] Hefei National Laboratory for Physical Sciences at the Microscale and Department of Electronic Engineering & Information Science, University of Science and Technology of China, Hefei 230026, China

ARTICLE INFO

Keywords:
Arrhythmia detection
ECG
Convolution neural network
Spatio-temporal attention module
Recurrent neural network

ABSTRACT

Automatic arrhythmia detection based on electrocardiogram (ECG) is of great significance for early prevention and diagnosis of cardiac diseases. Recently, deep learning methods have been applied to arrhythmia detection and obtained great success. Among them, convolutional neural network (CNN) is an effective method for extracting features due to its local connectivity and parameter sharing. In addition, recurrent neural network (RNN) is another commonly used method, which is applied to process time-series signal. The stacking of both CNN and RNN has been proved to be more effective in multi-class arrhythmia detection. However, these networks ignored the fact that different channels and temporal segments of a feature map extracted from the 12-lead ECG signal contribute differently to cardiac arrhythmia detection, and thus, the classification performance could be greatly improved. To address this issue, spatio-temporal attention-based convolutional recurrent neural network (STA-CRNN) is proposed to focus on representative features along both spatial and temporal axes. STA-CRNN consists of CNN subnetwork, spatio-temporal attention modules and RNN subnetwork. The experiment result shows that, STA-CRNN reaches an average F_1 score of 0.835 in classifying 8 types of arrhythmias and normal rhythm. Compared with the state-of-the-art methods based on the same public dataset, STA-CRNN achieves an obvious improvement on identifying most of arrhythmias. Also, it is demonstrated by visualization that the learned features through STA-CRNN are in line with clinical judgement. STA-CRNN provides a promising method for automatic arrhythmia detection, which has a potential to assist cardiologists in the diagnosis of arrhythmias.

1. Introduction

Cardiac arrhythmia is a set of conditions in which heartbeats are irregular [1]. From a clinical perspective, cardiac arrhythmia is divided into three categories. The first category is not serious and has no harm to health, called minor arrhythmia, such as bundle branch block and sporadic premature contraction. The second category requires treatment, called major arrhythmia, such as atrial fibrillation and frequent premature contraction. The third category is life-threatening arrhythmia and requires immediate treatment, such as ventricular tachycardia and ventricular fibrillation. Electrocardiogram (ECG) records the heart's electrical activities, which is widely used in clinical practice for arrhythmia diagnosis. A heartbeat is mainly composed of P wave, QRS complex wave and T wave, representing the process of depolarization and repolarization of atria and ventricle. Therefore, the pathology of heart can be reflected by the change in waveform or rhythm of the ECG signal [2]. In order to diagnose whether cardiac arrhythmia occurs, each heartbeat must be examined. A 24-h wearable ECG monitor can record up to 100 thousands heartbeats. It is time-consuming to analyze beat-by-beat for cardiologists. ECG-based automatic arrhythmia detection plays an important role in assisting cardiologists, and provides early warning to patients using wearable monitors. Therefore, it is of great significance to construct an accurate automatic arrhythmia detection solution.

In past decades, a large number of arrhythmia detection solutions have been proposed. These solutions are mainly composed of three steps, including denoising, feature extraction and arrhythmia classification. Among them, feature extraction requires to artificially design a group of features based on experience. Thus, the classification performance is limited due to hand-craft features.

In recent years, deep neural networks (DNNs) have achieved great successes in healthcare with powerful feature extraction ability, such as

* Corresponding authors.
E-mail addresses: aipingl@ustc.edu.cn (A. Liu), xunchen@ustc.edu.cn (X. Chen).

https://doi.org/10.1016/j.artmed.2020.101856
Received 2 December 2019; Received in revised form 5 March 2020; Accepted 2 April 2020

diagnostic of breast lesion [3], cardiovascular risk [4] and cell-free DNA-based prenatal diagnostics [5]. Different from traditional machine learning methods, DNNs are data-driven methods that can automatically extract features and avoid laborious feature engineering design. A number of studies have attempted to use DNNs for arrhythmia detection.

In this paper, an end-to-end novel neural network named as spatio-temporal attention-based convolutional recurrent neural network (STA-CRNN) is proposed for multi-class arrhythmia detection using 12-lead ECG records. In the literature, few studies have paid attention to the fact that different channels and temporal segments of a feature map extracted from the 12-lead ECG signal may contribute differently to cardiac arrhythmia detection. Specifically, at spatio-level, the morphology of cardiac arrhythmia behaves differently in different leads, for instance, atrial fibrillations is most evident in II lead and V1 lead. At temporal-level, some arrhythmias, especially paroxysmal arrhythmias such as premature ventricular contraction, show up intermittently and only part of heartbeats are abnormal. Based on the above considerations, the integration of both spatial and temporal attention mechanisms into a deep neural network will be beneficial to highlight the more informative features. Therefore, the proposed STA-CRNN incorporates spatio-temporal attention mechanism into convolutional recurrent neural network. STA-CRNN divides the features extraction into two phases: 1) Local features extraction part based on convolutional neural network and spatio-temporal attention mechanism modules to focus on the more representative features along two principle dimensions: spatial and temporal axes. 2) Global features extraction part based on recurrent neural network to incorporate the extracted local features. It was verified by visualization that STA-CRNN helps to locate the abnormalities in the ECG signal, thereby improving the interpretability of deep neural network model. The proposed model was evaluated on the private test set used for the China Physiological Signal Challenge 2018 [6], and compared with the state-of-the-art works evaluated on the same database.

The main contribution of this paper is highlighted as below. A novel neural network model STA-CRNN is proposed by incorporating spatio-temporal attention mechanism into convolutional recurrent neural network. Spatial and temporal attention mechanisms assign weights for channels and temporal segments of a feature map, respectively. The purpose is to emphasize the informative features and suppress unimportant ones along two principle dimensions: spatial and temporal axes. Comparing with the state-of-the-art works, this study further improves the performance of arrhythmia classification.

The rest of this paper is organized as follows. Section 2 concludes the related works. Section 3 introduces the architecture of STA-CRNN. Section 4 describes the detailed experiment. Section 5 presents the experimental result. Section 6 analyzes the performance and gives the visualization of the features learned by STA-CRNN. Finally, Section 7 summaries this paper.

2. Related works

Many previous arrhythmia classification methods extract numerous features artificially, mainly including P-QRS-T complex features [7–9], statistical features [10–12], morphological features [13–16] and wavelet features [17–20]. Mathematical transformations are also used to extract important information by transforming the high-dimensional ECG signal into a lower-dimensional subspace. It can be realized by principal component analysis (PCA) [13,21,22], linear discriminant analysis (LDA) [21,22] and independent component analysis (ICA) [21,23]. After feature engineering, a variety of classifiers such as artificial neural network (ANN) [18,9], LDA [21,20], k nearest neighbor (KNN) [24,25], support vector machine (SVM) [17,14,26], decision tree [24,27] and bayesian classifier [14,16] are implemented for arrhythmia classification. Afkhami et al. [28] proposed a heartbeat classification method based on an ensemble of decision trees on RR interval features and statistical features. Chen et al. [29] fed RR interval features and projected features compressed by project matrix into SVM for heartbeat classification. Kiranyaz et al. [30] extracted features using wavelet packet decomposition in heartbeat classification, and then they applied genetic algorithm to optimize these features and back propagation neural network (BPNN). The optimized BPNN was used to classify heartbeat. The above works were based on MIT-BIH arrhythmia database where ECG records are 2-lead. In comparison with these solutions, the proposed solution avoids laborious hand-craft features.

Recent studies have concentrated on deep learning. Among DNNs, convolutional neural network (CNN) is an effective method for extracting features due to its local connectivity and parameter sharing. Fan et al. [31] screened atrial fibrillation using a multiscaled fusion of CNN and gained state-of-the-art performance. Rajpurkar et al. [32] developed a 34-layer CNN which achieved cardiologist-level performance in identifying 12 types of arrhythmias.

Recurrent neural network (RNN) is a type of neural network used for processing time-series signal. Since ECG signal records the time course of cardiac electrical activities, RNN is also applied for arrhythmia detection. Long short-term memory (LSTM) and gated recurrent cell (GRU) are typical variants of RNN. Saadatnejad et al. [33] proposed a continuous and real-time patient-specific ECG classification algorithm based on wavelet transform and multiple LSTM. Lynn et.al [34] proposed a biometric ECG classification method based on a deep bidirectional GRU network.

Combining CNN with RNN, many works have built such stacked networks in which CNN is followed by RNN. In [35], Yao et al. proposed a model integrating a VGGNet-based CNN and LSTM layers for arrhythmia classification. Specifically, in this approach, all input ECG signals were padded to the same length and the original lengths were recorded. Then, the recorded lengths were provided to the first LSTM layer, indicating the length that the LSTM should spread to. This work has solved the problem that the varied-length signal couldn't be accepted by CNN models. He et al. [36] proposed a model consisting of residual convolutional network and a bidirectional LSTM layer, which obtained a good performance in classifying 9 arrhythmia classes. However, the above networks ignored the fact that different channels and temporal segments of a feature map extracted from the 12-lead ECG signal contribute differently to cardiac arrhythmia detection. Recently, Yao et al. [37] improved the deep neural network used in their previous work [35], introducing an attention module after CNN and LSTM layers. It was proved to be effective in detecting paroxysmal arrhythmias. This work gave greater weights to features extracted from more informative signal segments. Nevertheless, the informative channels were still ignored. The dataset used in [35–37] is the same as ours. It is from the China Physiological Signal Challenge 2018 where ECG records are 12-lead. We considered that the features extracted by the aforementioned networks still lack of sufficient representation for the ECG signal, and thus, the performance could be further improved.

3. Methods

3.1. Problem definition

In this paper, the task of multi-class arrhythmia detection is to automatically identify 8 arrhythmia classes and normal rhythm using varied-length 12-lead ECG records. The proposed model is required to take a 12-lead ECG record as input, and output the predicted label. The original ECG record $x^{(i)}$, together with the corresponding reference label $y^{(i)}$ constitute the training set $X = \{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), .., (x^{(m)}, y^{(m)})\}$, where $x^{(i)} \in \mathbb{R}^{L\times 12}$ is a signal with length L and 12 channels and $y^{(i)} \in \{0, 1, .., 8\}$ which follows the one-hot encoding scheme. During training, the objective of the model is to minimize the cross entropy of predicted probabilities with respects to their reference labels, defined as:

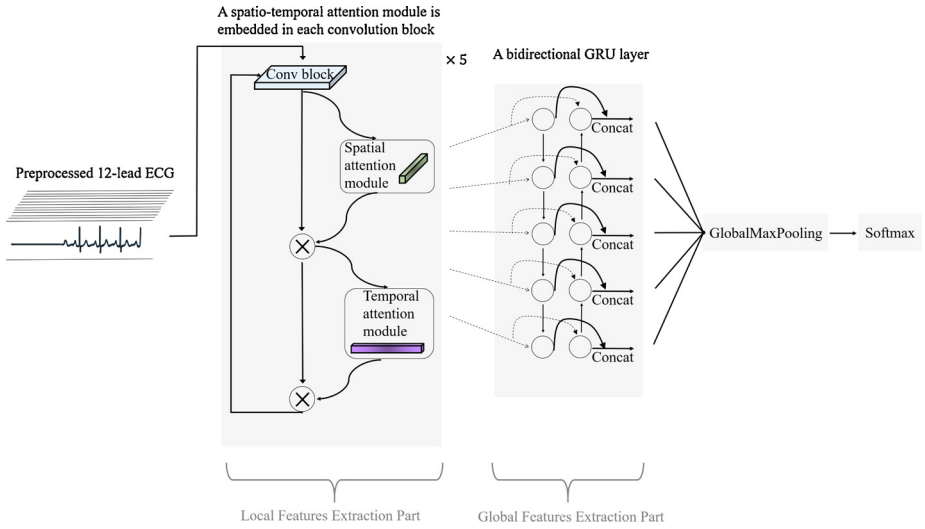

Fig. 1. The architecture for STA-CRNN.

$$\text{loss}(X) = -\frac{1}{m}\sum_{i=1}^{m}\log\left(\frac{\exp(p(x^{(i)}, y^{(i)}))}{\sum_j \exp(p(x^{(i)}, j))}\right) \quad (1)$$

where $p(x^{(i)}, j)$ indicates the probability that the model classifies the input $x^{(i)}$ to label j.

3.2. Model architecture

As illustrated in Fig. 1, the proposed model is built by integrating convolutional neural network, spatio-temporal attention modules and a bidirectional gated recurrent unit (GRU) layer. The abnormalities in ECG signals manifest as the changes in waveform morphology and rhythm [2]. The changes in waveform morphology show up during local periods, for instance, ST-segment slopes down when ST-segment depression occurs. And the changes in rhythm exist in the global period of ECG signals, such as the unequal RR interval for atrial fibrillation. Therefore, the model is divided into two parts for arrhythmia detection, including local features extraction part and global features extraction part. We combine convolutional neural network with spatio-temporal attention modules to extract the representative local features. Then, a bidirectional GRU is applied to extract global features by learning from all local features.

3.2.1. Convolutional neural network

A convolutional neural network inspired by VGGNet [38] is applied to extract local features. 5 convolutional blocks, or 13 1-dimension (1-D) convolution layers, form the convolution neural network, as shown in Fig. 2. Conv3 × 64 means that the convolution layer uses a kernel size of 3 and the kernel number of 64. Conv3 × 128 and Conv3 × 256 are in the similar expression. Every two or three convolution layers, together with a max-pooling layer and a dropout layer are regarded as a convolution block. There are 2 convolution layers in the first two convolution blocks and 3 convolution layers in the next three convolution blocks. Each convolution layer is followed by a batch normalization (BN) layer and a rectified linear unit (ReLU) activation function. BN [39] could accelerate the convergence of the model during training by normalizing each training mini-batch. ReLU [40] is a popular activation function, which was proved to avoid the vanishing gradient well. All max-pooling layers use a pool size of 3, thus their input length is reduced by 3 times. Dropout [41] with a rate of 0.2 is set to prevent neural network from overfitting.

3.2.2. Spatio-temporal attention module

In order to fully extract the representative local features, spatio-temporal attention module is introduced including spatial attention mechanism and temporal attention mechanism. Both attention mechanisms are based on the principles similar to those used in [42], as illustrated in Fig. 3. Spatial attention mechanism assigns weights for channels of a feature map to focus on ones that contribute more to the representation of the ECG signal. Complementally, temporal attention mechanism assigns weights for temporal segments of a feature map to concentrate on the more informative segments of the ECG signal. These two attention mechanisms are sequentially embedded after each convolution block to emphasize the representative local features along spatial and temporal axes where spatial attention mechanism precedes temporal attention mechanism.

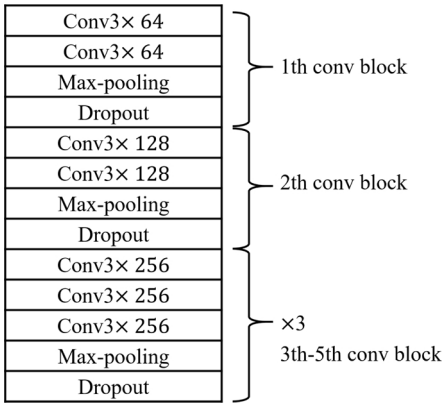

Fig. 2. Layer configuration for the convolutional neural network.

In spatial attention mechanism, global average-pooling and max-pooling separately aggregate the temporal information of each channel of input feature, obtaining two different temporal contexts. Average-pooling is effective in catching the global change, suitable for identifying arrhythmias that occur over a long time in a record. While max-pooling is better at capturing the local change, which is useful for identifying arrhythmias that show up intermittently in a record. Both temporal contexts are then fed into shared two dense layers. To decrease parameter amount, the hidden activation size of the first dense layer is set to $\mathbb{R}^{1\times C/d}$, where C is the channel numbers of input feature and d is the decrease ratio. The second dense layer has a hidden activation size of $\mathbb{R}^{1\times C}$. After applying the shared dense layers to each temporal context, the output feature vectors are merged by element-wise summation, generating the spatial attention weights. Finally, the spatial attention weights are compressed to 0-1 by sigmoid function.

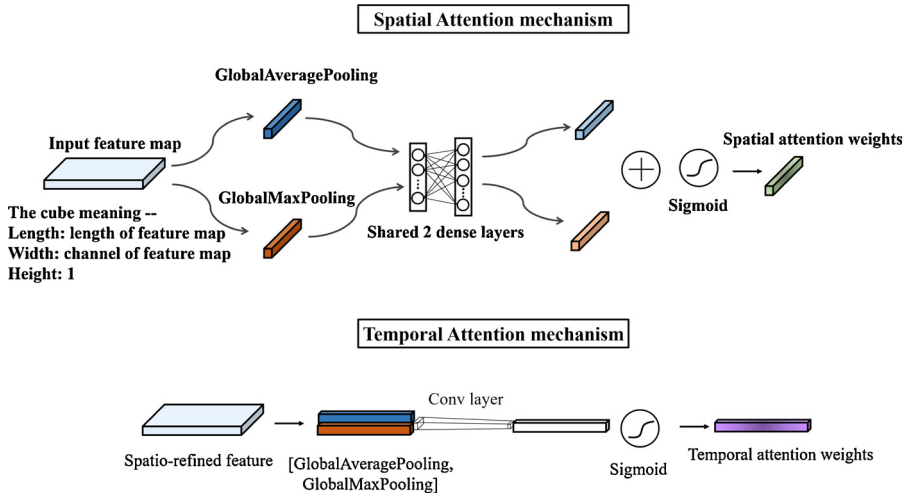

Fig. 3. Diagram of spatial and temporal attention mechanisms.

For temporal attention mechanism, global average-pooling and max-pooling separately aggregate the spatial information (also called channel information) of spatio-refined feature. The pooling operations along channel dimension are useful to highlight informative segments [43]. Then, the pooled features are combined by concatenation operation. Next, the concatenated feature vector is input into a convolution layer with filter size of 7 to obtain the temporal attention weights. In the same way, the temporal attention weights are compressed to 0-1 by sigmoid function.

3.2.3. Recurrent neural network

Following the local features extraction part, RNN is used to extract the global features. GRU [44] is one of the successful implementations of RNN. Similar to LSTM, GRU modulates the information flow by gate mechanisms, but using the hidden state to transmit information without the cell state. We choose GRU rather than LSTM since it performs similarly to LSTM but is computationally cheaper. In STA-CRNN, a bidirectional GRU layer is used, which consisting of the forward GRU layer and the backward GRU layer. At certain time, the local features from all past time steps are summarized by the forward GRU, and the local features from all future time steps are summarized by the backward GRU. The information from both opposite directions is incorporated by the bidirectional GRU to get annotations of time steps, containing the contextual information. At time t, the GRU takes the current input x_t as input together with the hidden state h_{t-1} from time $t-1$, and output the activation h_t. Its internal behavior is shown in Fig. 4, and is described by the formula below:

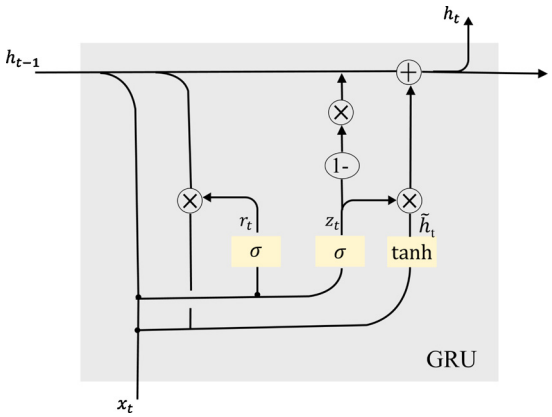

Fig. 4. The internal structure of GRU. r_t is the reset gate. z_t is the update gate. $\tilde{h}_t$ is the candidate activation. h_t is the activation.

$$z_t = \sigma(W_z x_t + U_z \odot h_{t-1}) \tag{2}$$

$$r_t = \sigma(W_r x_t + U_r h_{t-1}) \tag{3}$$

$$\tilde{h}_t = \tanh(W x_t + U(r_t h_{t-1})) \tag{4}$$

$$h_t = (1 - z_t) h_{t-1} + z_t \tilde{h}_t \tag{5}$$

Here, σ refers to the sigmoid function. $\odot$ is an element-wise multiplication. z_t and r_t are the update and reset gates that decide how much the activation h_t is updated and how much the previous activation h_{t-1} is forgotten. W_z, U_z, W, U, W_r and U_r are the trainable parameters. The activation h_t is a linear interpolation between the previous activation h_{t-1} and the candidate activation $\tilde{h}_t$. In STA-CRNN, the unit number of the bidirectional GRU layer is set to 12.

4. Experiment

4.1. Environment

The proposed network was implemented and trained using Keras 2.2.4 framework, and all experiments were performed on a server with Xeon E5 2620 CPU, 128GB memory and four GeForce RTX cards.

4.2. Data source

The training set used in this study is the public database from the China Physiological Signal Challenge 2018 (CPSC 2018). The proposed model was evaluated using the private test set of CPSC 2018. The training set contains 6877 12-lead ECG records with varied lengths ranging from 6 to 60 seconds, and the test set contains 2954 12-lead ECG records with the similar lengths. CPSC 2018 database collected these ECG records from 11 hospitals, which contains 9 types of ECG classes including 8 arrhythmia classes and normal rhythm. ECG records are sampled as 500 Hz. The majority of the records have only one reference label. However, a few records have up to three reference labels (referred as First label, Second label and Third label). For each record, the detection result is considered right only any of these reference labels is given. Table 1 details the used training set, and Fig. 5 shows a typical normal ECG record.

4.3. Preprocessing

In order to reduce the computational cost, the original ECG signals are downsampled from 500Hz to 256Hz. The downsampling operation speeds up the training process and has almost no loss of information from ECG signals. The lengths of the original ECG signals vary from 6 to

Table 1
Data profile for the public training set according to the "First label" annotations.

Type	#record
Normal	918
Atrial fibrillation (AF)	1098
First-degree atrioventricular block (1-AVB)	704
Left bundle branch block (LBBB)	207
Right bundle branch block (RBBB)	1695
Premature atrial contraction (PAC)	556
Premature ventricular contraction (PVC)	672
ST-segment depression (STD)	825
ST-segment elevated (STE)	202
total	6877

60 s. Convolutional neural network is incapable of responding to varied-length input. Therefore, we crop or pad the downsampled ECG signals to the same length. In this paper, 60 s is chosen. That is to say, the downsampled ECG signals that have longer lengths than 60 seconds need to be cropped and those that have shorter lengths need to be padded with zero.

4.4. Training setting

4.4.1. Model optimization

The preprocessed ECG signals are input into STA-CRNN in batches of 64. The overall training procedure requires less memory when using mini-batch, and typical networks train faster with mini-batch. We choose 64 as the batch size through fine-tuning of hyperparameters. The weights of convolutional layers are initialized with the Xavier uniform initializer [45] and the bidirectional GRU is initialized with orthogonal initializer. Adam optimizer [46] is applied to update the weights iteratively due to its ability to accelerate the convergence of deep network model. The learning rate is set to 0.001.

4.4.2. Regularization strategies

Deep neural networks tend to overfit. Dropout is a commonly used regularization method that facilitates the model's generalizability by randomly inactivating some hidden neurons. In addition to being applied in each convolution block, dropout with a rate of 0.2 is also used after the bidirectional GRU layer. EarlyStopping terminates training when the model's performance on the validation set is no longer improving, which is an effective strategy to alleviate overfitting. In general, accuracy is used as the stop criterion, however, causes the model to favor arrhythmias that have more instances. In our experiments, we use the average F_1 score as the stop criterion. The average F_1 score is described in following "Performance Metric" subsection. It is a quite fair stop criterion as the model is impartial towards each type of arrhythmia during training. When the average F_1 score on the validation set has not improved for 100 epochs, the model is set to terminate training.

4.4.3. Cross validation

In order to fully utilize the entire training set, 4-fold cross validation was applied since the training set is relatively small. The original training set was randomly divided into four subsets. Each of four subsets took turns as validation set and the remaining subsets were used as the training set. Four models were trained with not exactly the same training set. Finally, the predicted probabilities of four models are averaged to give the classification judgement.

5. Results

5.1. Performance metric

For a multi-class imbalanced dataset, macro-F_1 score is a typical metric for measuring the classification performance of a model [37]. The performance of a model on the small class is greatly reflected by the metric based on macro-average. Therefore, we use macro-F_1 score to evaluate the performance of STA-CRNN. For each class, F_1 score is defined as:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{6}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{7}$$

$$F_1 = \frac{2(\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}} \tag{8}$$

Here, TP is the number of records that are classified to be positive and in fact positive. FP is the number of records that are classified to be positive but actually negative. And FN is the number of records that are classified to be negative but actually positive. The average of F_1 score among classes is computed to evaluate the final performance of the model.

5.2. Experimental process

The input of the proposed model is the preprocessed ECG signal. It is a three dimensional matrix with dimensions (64, 15360, 12), or more flexibly expressed as (None, 15360, 12). The first dimension is batch size, which is set to 64 in our experiments. The second dimension is the signal length, where the sampling frequency is 256 Hz and the lasting period is 60 s. The third dimension is channel number (i.e., lead number).

The three dimensional matrix is fed into the local features extraction

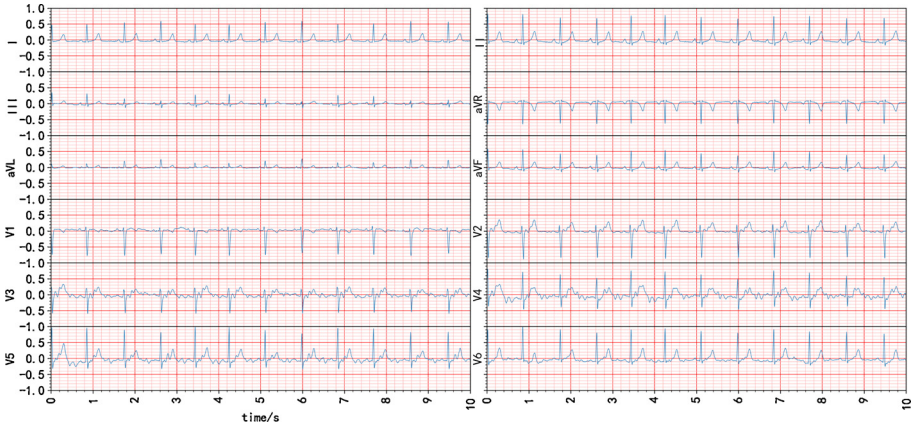

Fig. 5. A typical 12-lead ECG record. The record number is "A0153", and the label is normal.

part consisting five pairs of convolution blocks and spatio-temporal attention modules. Since the max-pooling layer uses a pool size of 3 and the convolutional layer has the kernel number of 64 in the first convolution block, the first convolution block outputs the feature maps with dimensions (None, 5120, 64). Then, the feature maps are multiplied by spatial attention weights with dimensions (None, 1, 64) to get spatio-refined feature maps with dimensions (None, 5120, 64). Next, spatio-refined feature maps are multiplied by temporal attention weights with dimensions (None, 5120, 1) to get weighted feature maps along both spatial and temporal axes with dimensions (None, 5120, 64). The feature maps flow through each pair of convolution block and spatio-temporal attention module. Finally, the fifth pair of convolution block and spatio-temporal attention module, or local features extraction part outputs feature maps with dimensions (None, 63, 256).

The feature maps extracted by local features extraction part are flowed into global features extraction part consisting of a bidirectional GRU layer. The unit number of the GRU layer is 12. Since the bidirectional GRU layer is obtained by concatenating the forward GRU and the backward GRU, each global feature vector is 24 in length. Global features extraction part outputs global feature vectors with dimensions (None, 63, 24).

All global feature vectors are flowed into a global max-pooling layer to get a single global feature vector with dimensions (None, 24). Then, the global feature vector is flowed into a fully connected layer with an output dimension of 9 (i.e. the number of ECG classes) for the classification. The fully connected layer outputs the prediction probabilities cross 9 classes. Next, these probabilities are compressed to 0-1 by softmax function, and their sum is equal to 1.

Adam optimizer is adopted to calculate the cross entropy loss function between predicted probabilities and their reference labels, and update weights of the neural network. All preprocessed training samples are fed into the model in batches of 64 to update weights iteratively, called an epoch. The model is set to terminate training, until the performance on the validation dataset has not improved for 100 epochs.

Since 4-fold cross validation is applied, four trained models are acquired with not exactly the same training dataset. When testing a 12-lead ECG signal, the predicted probabilities of four models are averaged to give the classification judgement. The classification result is the arrhythmia class with the max probability.

5.3. Classification performance

In order to demonstrate the effectiveness of STA-CRNN, three reference models based on the original convolution neural network and bidirectional GRU layer are compared with STA-CRNN. Three reference models are built as follows:

(1) CRNN: The global features extraction part is the same as in STA-CRNN, while spatio-temporal attention modules are not embedded in the convolution neural network in the local features extraction part.
(2) SA-CRNN: With the global features extraction part invariant, separate spatial attention module is embedded after each convolutional block in the local features extraction part.
(3) TA-CRNN: With the global features extraction part invariant, separate temporal attention module is embedded after each convolutional block in the local features extraction part.

Table 2 compares the F_1 score of three reference models and STA-CRNN in detecting multi-class cardiac arrhythmias. It is shown that STA-CRNN almost outperforms CRNN in the F_1 score of all classes except RBBB and Normal where two models perform comparably, and is superior to SA-CRNN and TA-CRNN in the F_1 score of most classes. It is consistently better to embed both spatial and temporal attention mechanisms than to embed either of those. While three models with both or either of spatial and temporal attention mechanisms all outperform CRNN in almost all classes, the largest performance improvement lies in identifying STE. SA-CRNN, TA-CRNN achieve 4%, 6.2% F_1 score increase in identifying STE. STA-CRNN further improves the score by 5.5%, 3.3%.

Table 2
Classification performance with different structures.

Type	Model structure			
	CRNN	SA-CRNN	TA-CRNN	STA-CRNN
Normal	0.820	0.834	**0.836**	0.819
AF	0.929	0.921	0.923	**0.936**
1-AVB	0.862	**0.875**	0.864	0.866
LBBB	0.856	**0.872**	0.849	0.862
RBBB	**0.934**	0.929	0.933	0.926
PAC	0.766	0.768	0.780	**0.789**
PVC	0.851	0.860	0.845	**0.865**
STD	0.807	0.799	0.789	**0.812**
STE	0.545	0.585	0.607	**0.640**
average F_1	0.819	0.827	0.825	**0.835**

6. Discussion

6.1. Performance analysis

The padding strategy retains the original information. However, noises to some degree are introduced. Spatio-temporal attention module's ability is to focus on informative parts such as abnormalities in the ECG signal, and suppress unimportant ones such as noises. Due to the integration of spatio-temporal attention module, STA-CRNN compensates for noise interference. More importantly, with its locating capability, STA-CRNN could extract the more representative features from ECG signals. Fig. 6 shows the confusion matrix of STA-CRNN. As can be seen from this figure, the model makes the most mistakes in identifying normal signals from STD signals. In consideration of the number of ECG records among classes, it is observed that the most influence on STA-CRNN's performance is the mistakes in the discrimination between normal signals and STE signals. STA-CRNN is insensitive to the changes in ST-segment, possibly because noises could easily cover these minor changes.

6.2. Visualization of learned features

Grad-CAM [47] is applied to visually explain the features learned by

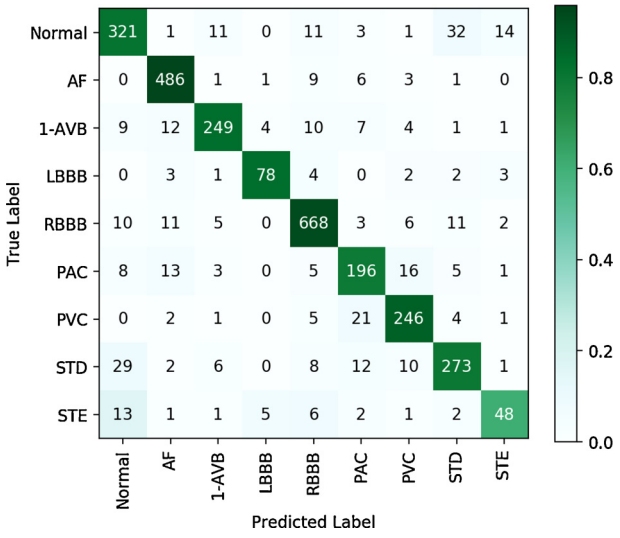

Fig. 6. Confusion matrix of the proposed model, STA-CRNN.

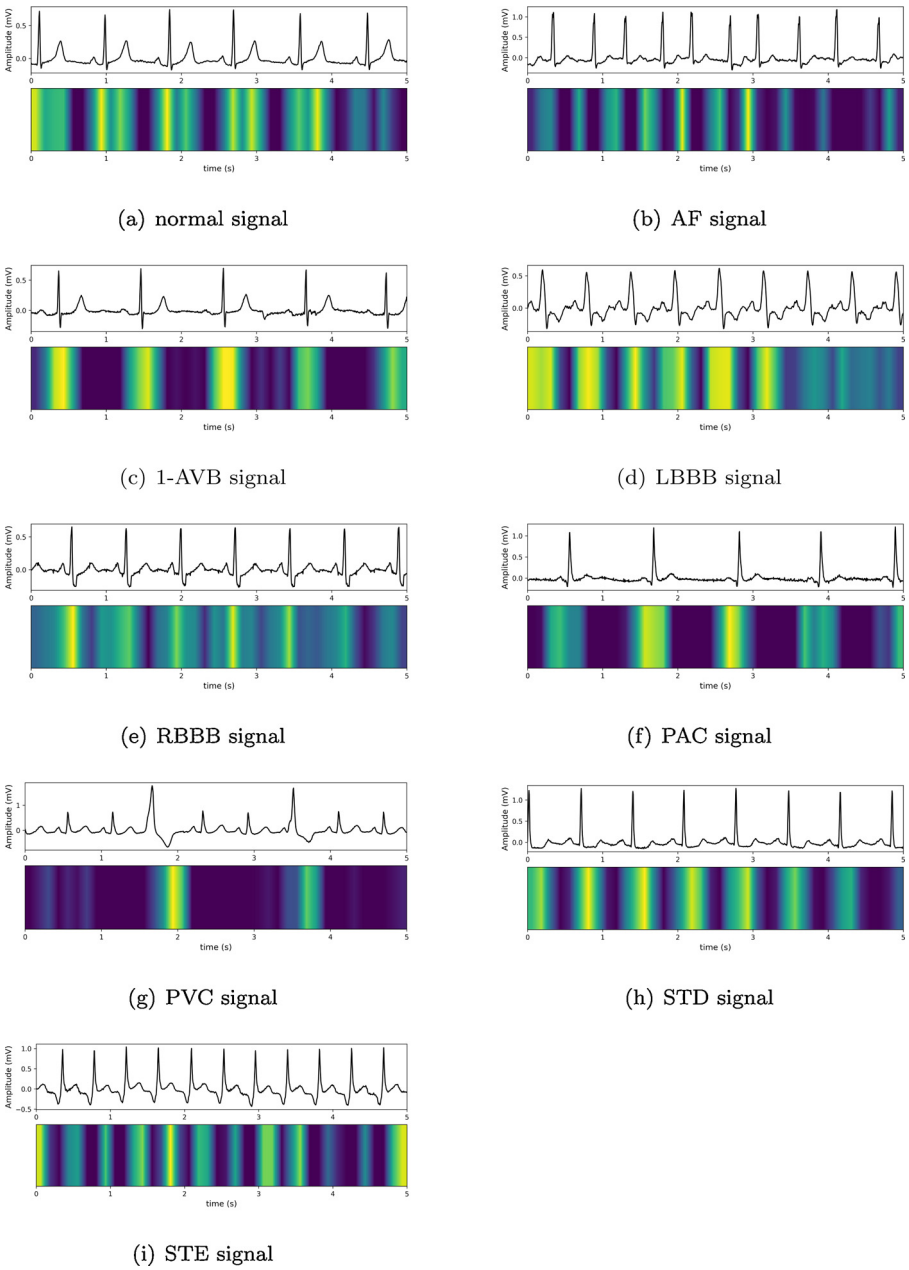

Fig. 7. The upper figure shows the original signal from II lead, and the lower figure is the corresponding attention map in (a)-(i).

STA-CRNN. In Fig. 7, visualization of learned features on a heat map is shown for each class, where the brighter color indicates the stronger attention on corresponding ECG signal segment. As shown in Fig. 7(a), attention is mainly concentrated on P, QRS and T waves for normal signal. During the onset of AF, P wave is replaced by a continuous irregular baseline fluctuation wave (f wave). As in Fig. 7(b), attention is indeed focused on abnormal P waves for AF signal. For the three arrhythmias, 1-AVB, LBBB and RBBB caused by conduction block, the abnormal QRS complex waves are of greater concern as shown in Fig. 7(c)-7 (e). When PAC occurs, P′-QRS-T waves appear prematurely with the absence of QRS complex after P′ wave or the morphology of P′ wave slightly different from that of sinus P wave. As in Fig. 7(f), attention is mainly concentrated on P and QRS waves for PAC signal. For PVC, the features of T waves are emphasized, as shown in Fig. 7(g), in line with the fact that the direction of T waves is opposite to that of the main waves in PVC. Both STD and STE are abnormal in ST-segment. As shown in Fig. 7(h) and 7 (i), the features of ST-segment are focused. It is noted that the features extracted by the proposed model, STA-CRNN, meet the clinical judgement, demonstrating that the model is potentially effective in the location of abnormal signal patterns so that most of arrhythmias could be screened out.

6.3. Classification performance comparison

Four previous works evaluated on the same dataset are compared with our proposed model for multi-class arrhythmia detection. Table 3 shows the F_1 score of each class in all works except [36] in which the F_1 score of each class was not given, and thus the detailed result analysis about [36] is not involved in following analysis. [48] combined handcrafted features with deep features extracted by a 17-layer convolutional neural network. The network structures of [36] and [35] were built on the basis of convolutional neural network and LSTM, and [37] introduced an attention module along temporal axis based on [35]. According to Table 3, it is found that the proposed model is comparable to [48] in identifying Normal, 1-AVB, LBBB and STD. Generally speaking, these four ECG classes could be easily screened out by a deep

Table 3
Classification performance comparison for other works and the proposed model.

Type	F_1 score					#recording
	[48]	[36]	[35]	[37]	Our work	
Normal	0.82	-	0.753	0.789	0.819	394
AF	0.91	-	0.900	0.920	0.936	466
1-AVB	0.87	-	0.809	0.850	0.866	295
LBBB	0.87	-	0.874	0.872	0.862	97
RBBB	0.91	-	0.922	0.933	0.926	756
PAC	0.63	-	0.638	0.736	0.789	250
PVC	0.82	-	0.832	0.861	0.865	276
STD	0.81	-	0.762	0.789	0.812	340
STE	0.60	-	0.462	0.556	0.640	80
average F_1	0.81	0.806	0.772	0.812	0.835	2954

The mark '-' is filled when the F_1 score of certain class were not given.

neural network. While for those arrhythmias that are more difficult to identify, such as PAC and STE, the proposed model significantly outperforms [48], achieved up to 15.9% and 4% F_1 score increase in identifying PAC and STE respectively. In comparison with [35] and [37], the proposed model achieves an obvious improvement on identifying most of arrhythmias, demonstrating it is more competitive in comparison with state-of-the-art methods.

7. Conclusion

In this paper, an end-to-end novel neural network model STA-CRNN is proposed for multi-class arrhythmia detection using 12-lead ECG records. STA-CRNN incorporates spatio-temporal attention mechanism modules into convolutional recurrent neural network to emphasize the informative features and suppress the unimportant ones. It is demonstrated by experiments that STA-CRNN achieves a superior detection performance improvements in comparison with the state-of-the-art methods, especially in identifying arrhythmias with lower recognition rate. We find that the incorporation of spatio-temporal attention mechanism modules significantly improves the detection accuracy of some arrhythmias, such as PAC and STE, which are more difficult to be screened out through plain deep neural networks. For PAC and STE, the incorporation of spatio-temporal attention mechanism modules achieves 2.3% and 9.5% F_1 score increase, respectively. Spatio-temporal attention module is helpful for the model to locate informative parts of a signal along spatial and temporal dimensions and improves interpretability of the proposed model. Finally, Grad-CAM is applied to visually explain the features learned by the proposed model. It is proved that the extracted features are in line with clinical judgement. Therefore, STA-CRNN is a promising solution to multi-class arrhythmia detection and has the potential to assist cardiologists in clinical arrhythmia diagnosis.

Conflict of interest

We would like to confirm that all authors were fully involved in the study and preparation of the manuscript. None of this work has been previously published, or been pending publication in another journal, or been under review in any other journal. None of the authors has a conflict of interest.

CRediT authorship contribution statement

Aiping Liu: Experimental design, Medical result interpretation. **Xun Chen:** Methodology design.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (Grants 61922075 and 61701158). The authors would like to appreciate Prof. Chengyu Liu from Southeast University, Nanjing 210096, China, for the experimental database he and his group provided for this research.

References

[1] Kass RE, Clancy CE. Basis and treatment of cardiac arrhythmias, Vol. 171. Springer Science & Business Media; 2005.
[2] Van Mieghem C, Sabbe M, Knockaert D. The clinical value of the ECG in noncardiac conditions. Chest 2004;125(4):1561–76.
[3] Lamy J-B, Sekar B, Guezennec G, Bouaud J, Séroussi B. Explainable artificial intelligence for breast cancer: A visual case-based reasoning approach. Artif Intel Med 2019;94:42–53.
[4] Bollepalli SC, Challa SS, Anumandla L, Jana S. Dictionary-based monitoring of premature ventricular contractions: An ultra-low-cost point-of-care service. Artif Intel Med 2018;87:91–104.
[5] Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, et al. A guide to deep learning in healthcare. Nat Med 2019;25(1):24–9.
[6] Liu F, Liu C, Zhao L, Zhang X, Wu X, Xu X, et al. An open access database for evaluating the algorithms of electrocardiogram rhythm and morphology abnormality detection. J Med Imaging Health Inform 2018;8(7):1368–73.
[7] Tsipouras MG, Fotiadis DI, Sideris D. An arrhythmia classification system based on the rr-interval signal. Artificial Intelligence in Medicine 2005;33(3):237–50.
[8] Exarchos TP, Tsipouras MG, Exarchos CP, Papaloukas C, Fotiadis DI, Michalis LK. A methodology for the automated creation of fuzzy expert systems for ischaemic and arrhythmic beat classification based on a set of rules obtained by a decision tree. Artif Intel Med 2007;40(3):187–200.
[9] Haseena HH, Mathew AT, Paul JK. Fuzzy clustered probabilistic and multi layered feed forward neural networks for electrocardiogram arrhythmia classification. J Med Syst 2011;35(2):179–88.
[10] Dilmac S, Korurek M. Ecg heart beat classification method based on modified abc algorithm. Appl Soft Comput 2015;36:641–55.
[11] Li Q, Rajagopalan C, Clifford GD. Ventricular fibrillation and tachycardia classification using a machine learning approach. IEEE Trans Biomed Eng 2013;61(6):1607–13.
[12] Javadi M, Arani SAAA, Sajedin A, Ebrahimpour R. Classification of ecg arrhythmia by a modular neural network based on mixture of experts and negatively correlated learning. Biomed Signal Process Control 2013;8(3):289–96.
[13] Ince T, Kiranyaz S, Gabbouj M. A generic and robust system for automated patient-specific classification of ecg signals. IEEE Trans Biomed Eng 2009;56(5):1415–26.
[14] Zhang Z, Dong J, Luo X, Choi K-S, Wu X. Heartbeat classification using disease-specific feature selection. Comput Biol Med 2014;46:79–89.
[15] Rodríguez-Sotelo JL, Cuesta-Frau D, Castellanos-Dominguez G. Unsupervised classification of atrial heartbeats using a prematurity index and wave morphology features. Med Biol Eng Comput 2009;47(7):731–41.
[16] Mar T, Zaunseder S, Martínez JP, Llamedo M, Poll R. Optimization of ecg classification by means of feature selection. IEEE Trans Biomed Eng 2011;58(8):2168–77.
[17] Elhaj FA, Salim N, Harris AR, Swee TT, Ahmed T. Arrhythmia recognition and classification using combined linear and nonlinear features of ecg signals. Comput Methods Programs Biomed 2016;127:52–63.
[18] “Ozbay Y. A new approach to detection of ecg arrhythmias: Complex discrete wavelet transform based complex valued artificial neural network. J Med Syst 2009;33(6):435.
[19] Khorrami H, Moavenian M. A comparative study of dwt, cwt and dct transformations in ecg arrhythmias classification. Expert Syst Appl 2010;37(8):5751–7.
[20] Balasundaram K, Masse S, Nair K, Umapathy K. A classification scheme for ventricular arrhythmias using wavelets analysis. Med Biol Eng Comput 2013;51(1–2):153–64.
[21] Martis RJ, Acharya UR, Min LC. Ecg beat classification using pca, lda, ica and discrete wavelet transform. Biomed Signal Process Control 2013;8(5):437–48.
[22] Wang J-S, Chiang W-C, Hsu Y-L, Yang Y-TC. Ecg arrhythmia classification using a probabilistic neural network with a feature reduction method. Neurocomputing 2013;116:38–45.
[23] Martis RJ, Acharya UR, Prasad H, Chua CK, Lim CM. Automated detection of atrial fibrillation using Bayesian paradigm. Knowledge-Based Syst 2013;54:269–75.
[24] Martis RJ, Acharya UR, Prasad H, Chua CK, Lim CM, Suri JS. Application of higher order statistics for atrial arrhythmia classification. Biomed Signal Process Control 2013;8(6):888–900.
[25] Kotu LP, Engan K, Borhani R, Katsaggelos AK, Ørn S, Woie L, et al. Cardiac magnetic resonance image-based classification of the risk of arrhythmias in post-myocardial infarction patients. Artif Intel Med 2015;64(3):205–15.
[26] Asl BM, Setarehdan SK, Mohebbi M. Support vector machine-based arrhythmia classification using reduced features of heart rate variability signal. Artif Intel Med 2008;44(1):51–64.
[27] Seera M, Lim CP, Liew WS, Lim E, Loo CK. Classification of electrocardiogram and auscultatory blood pressure signals using machine learning models. Expert Syst Appl 2015;42(7):3643–52.
[28] Afkhami RG, Azarnia G, Tinati MA. Cardiac arrhythmia classification using statistical and mixture modeling features of ecg signals. Pattern Recog Lett

2016;70:45–51.
[29] Chen S, Hua W, Li Z, Li J, Gao X. Heartbeat classification using projected and dynamic features of ecg signal. Biomed Signal Process Control 2017;31:165–73.
[30] Kiranyaz S, Ince T, Gabbouj M. Real-time patient-specific ecg classification by 1-d convolutional neural networks. IEEE Trans Biomed Eng 2015;63(3):664–75.
[31] Fan X, Yao Q, Cai Y, Miao F, Sun F, Li Y. Multiscaled fusion of deep convolutional neural networks for screening atrial fibrillation from single lead short ecg recordings. IEEE J Biomed Health Inform 2018;22(6):1744–53.
[32] Hannun AY, Rajpurkar P, Haghpanahi M, Tison GH, Bourn C, Turakhia MP, et al. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. Nature Med 2019;25(1):65.
[33] S. Saadatnejad, M. Oveisi, M. Hashemi, Lstm-based ecg classification for continuous monitoring on personal wearable devices, IEEE J Biomed Health Inform.
[34] Lynn HM, Pan SB, Kim P. A deep bidirectional gru network model for biometric electrocardiogram classification based on recurrent neural networks. IEEE Access 2019;7:145395–405.
[35] Yao Q, Fan X, Cai Y, Wang R, Yin L, Li Y. Time-incremental convolutional neural network for arrhythmia detection in varied-length electrocardiogram. 2018 IEEE 16th Intl Conf on Dependable, Autonomic and Secure Computing, 16th Intl Conf on Pervasive Intelligence and Computing, 4th Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech) 2018:754–61.
[36] He R, Liu Y, Wang K, Zhao N, Yuan Y, Li Q, Zhang H. Automatic cardiac arrhythmia classification using combination of deep residual network and bidirectional lstm. IEEE Access 2019;7:102119–35.
[37] Yao Q, Wang R, Fan X, Liu J, Li Y. Multi-class arrhythmia detection from 12-lead varied-length ecg using attention-based time-incremental convolutional neural network. Information Fusion 2020;53:174–82.
[38] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. International Conference on Learning Representations (ICLR) 2015:1–14.
[39] Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37, ICML-15. 2015. p. 448–56.
[40] Nair V, Hinton GE. Rectified linear units improve restricted boltzmann machines. Proceedings of the 27th International Conference on Machine Learning (ICML-10) 2010:807–14.
[41] Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res 2014;15(1):1929–58.
[42] Woo S, Park J, Lee J-Y, So Kweon I. Cbam: Convolutional block attention module. Proceedings of the European Conference on Computer Vision (ECCV) 2018:3–19.
[43] Komodakis N, Zagoruyko S. Paying more attention to attention: improving the performance of convolutional neural networks via attention transfer. International Conference on Learning Representations (ICLR) 2017.
[44] Cho K, van Merri”enboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, Bengio Y. Learning phrase representations using rnn encoder-decoder for statistical machine translation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) 2014:1724–34.
[45] Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks. Proceedings of the thirteenth International Conference on Artificial Intelligence and Statistics 2010:249–56.
[46] Kingma D, Ba J. Adam: A method for stochastic optimization. International Conference on Learning Representations (ICLR) 2015.
[47] Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: Visual explanations from deep networks via gradient-based localization. Proceedings of the IEEE International Conference on Computer Vision 2017:618–26.
[48] Liu Z, Meng X, Cui J, Huang Z, Wu J. Automatic identification of abnormalities in 12-lead ecgs using expert features and convolutional neural networks. 2018 International Conference on Sensor Networks and Signal Processing (SNSP). 2018. p. 163–7.