Fusing Transformer Model with Temporal Features for ECG Heartbeat Classification

Conference Paper · November 2019

DOI: 10.1109/BIBM47256.2019.8983326

4 authors, including:

Shen Liang
Fudan University
11 PUBLICATIONS **12** CITATIONS
SEE PROFILE

Fan Liu
Northwestern Polytechnical University
13 PUBLICATIONS **89** CITATIONS
SEE PROFILE

Fusing Transformer Model with Temporal Features for ECG Heartbeat Classification

Genshen Yan*¶, Shen Liang†¶, Yanchun Zhang‡¶, Fan Liu§
*School of Software Engineering, Fudan University, Shanghai, China
†School of Computer Science, Fudan University, Shanghai, China
‡College of Engineering and Science, Victoria University, Melbourne, Australia
§School of Computer Science, Northwestern Polytechnical University, Xian, China
¶Cyberspace Institute of Advanced Technology, Guangzhou University, Guangzhou, China
{gsyan17, sliang11}@fudan.edu.cn, yanchun.zhang@vu.edu.au, liufant800@mail.nwpu.edu.cn

***Abstract*—ECG heartbeat classification plays a vital role in diagnosis of cardiac arrhythmia. Traditional heartbeat classification methods rely on handcrafted features and often fail to learn potentially abstract patterns, while current deep learning based methods usually consist of complex convolutional and recursive structures. In this paper, considering the time sequence property of ECG signals, we propose a novel heartbeat classification method based on Transformer, a sequence-to-sequence model with a relatively simple architecture and higher degree of parallelism. To adapt Transformer to our problem, we use only the encoder part of the model for that ECG signals do not have translation signals. We also replace the dropout layers with batch normalization layers, considering our small-size feature space and the natural differences among patients. Further, we fuse handcrafted temporal features with the features learnt by Transformer to better capture rhythmic patterns in ECG signals. We have conducted extensive experiments on the MIT-BIH arrhythmia database using both the original dataset and an augmented dataset with more balanced data. The results show that our model can achieve 99.62% accuracy, 95.09% precision and 94.12% sensitivity on the original dataset and 99.87% accuracy, 99.74% precision and 99.74% sensitivity on the augmented dataset. Besides, we have performed multiple experiments against state-of-the-art methods using *their* assessment strategies. Experimental results indicate that our model can achieve better performance under most circumstances.**

***Index Terms*—ECG classification, Arrhythmia, Deep learning, Transformer Model, MIT-BIH database**

I. Introduction

According to a report of the World Health Organization (WHO), cardiovascular disease (CVD) is the leading cause of noncommunicable diseases. In 2016, it is estimated that about 17.9 million people died of CVDs, which accounts for 31.4% of deaths globally [1]. Arrhythmia is an important group of CVDs which can occur alone or in association with other CVDs. There are various types of arrhythmia including ventricular premature beats, tachycardia, supraventricular ectopic and atrial fibrillation, etc. Heartbeat classification is one of the main diagnostic methods of arrhythmia, therefore, it is an important research topic in healthcare.

Electrocardiogram (ECG) is the object for heartbeat classification which records the physiological state of the heart. ECG signals consist of a number of heartbeats, each of which is comprised of several successive waves. Fig. 1 shows an ECG signal example with premature ventricular contraction. Heartbeats are morphologically similar under normal conditions, which equivalently suggests that ECG signals are periodic.

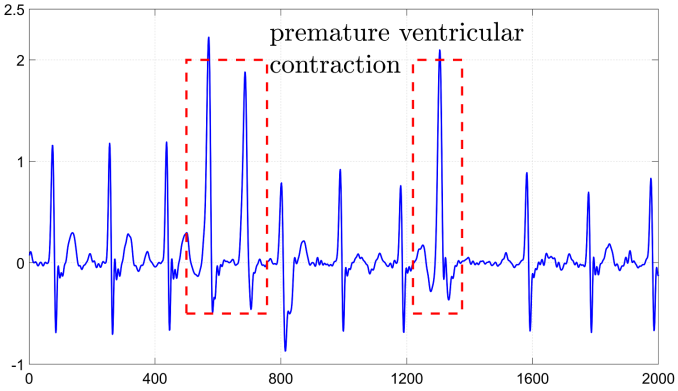

Fig. 1. ECG signal example with ventricular premature.

Each type of arrhythmia is associated with a pattern, which means it is possible to identify and classify arrhythmia by well-designed algorithms [2]. Initially, researchers used traditional methods to classify heartbeats. Traditional approaches have the advantage of interpretability but they are weak in self-learning and require adequate manual intervention. Besides, traditional models usually spend much time in feature extraction and feature selection. To deal with these problems, several literatures have proposed deep learning based methods which can effectively extract abstract features and can be more automated.

However, the deep learning based approaches still have some problems. First, most of the deep learning based methods need many complex convolution and recursive structures, which usually generate a sequence of hidden states. The current hidden state depends on the previous one, resulting in a low degree of parallelization of these models. To address this issue, we propose a novel method based on the Transformer model that was firstly used in translation field. The reason for using Transformer is as follows: a) Transformer only relies on the attention mechanism and does not need complex convolutional operations. b) Transformer encodes the position

978-1-7281-1867-3/19/$31.00 © 2019 IEEE

information of the sequence and integrates the dependencies into matrix operation. Therefore, all dependencies can be handled simultaneously, which makes Transformer having high parallelism. c) Like a word can be inferred from other words, we believe that similar mechanism can be used for ECG signals. To apply Transformer to ECG time series, we have made the following changes: Transformer uses the Encoder-Decoder architecture [3] while the inputs of the encoder and decoder are original language and the translation language respectively. ECG signals do not have translation signals, so we use the encoder part only. Besides, Transformer uses dropout to alleviate over-fitting. Considering the small feature size and the natural differences among patients, we use batch normalization (BN) [4] instead. The second problem is that most of existing methods only consider the morphological appearance of the ECG when classifying different types of arrhythmia, which we think is not enough to achieve good performance. The last problem is that many methods only compare the numerical value of performance criteria, regardless of the differences in evaluation strategies.

For the second problem, we integrate two temporal features, i.e., two RR intervals which will be described in section II. And for the last problem, different assessment strategies are used when we compare our model with different state-of-the-art methods. The contributions of this study can be summarized as follows:

- We propose a novel deep learning model based on Transformer. To adapt for our problem, we use the encoder part only and use BN instead of dropout to prevent over-fitting. Relying on attention mechanism and with the positional encoding, our model has a more simple and more parallelized architecture.
- We integrate temporal features with the morphological signals, which is important for rhythmic abnormalities such as premature ventricular contractions.
- We have conducted extensive experiments, including different assessment strategies, parameter selection and ablation experiments. Besides, our model achieved state-of-the-art results on both original and augmented dataset.

The paper is organized as follows: Section II provides the background and the related work. Section III describes the dataset and Section IV presents the details of our method. In Section V, we show the experiments and evaluation results. Finally, we conclude the paper in Section VI.

II. Background and Related Work

A. ECG structure

The basic structure of an ECG heartbeat consists of a P wave, a T wave and a QRS complex wave. The interval between two consecutive R peaks is called RR interval, which is an important reference information in the diagnosis of heart diseases. We use pre-RR interval and post-RR interval as our temporal features:

- pre-RR interval: the RR interval between a given heartbeat and its preceding heartbeat.

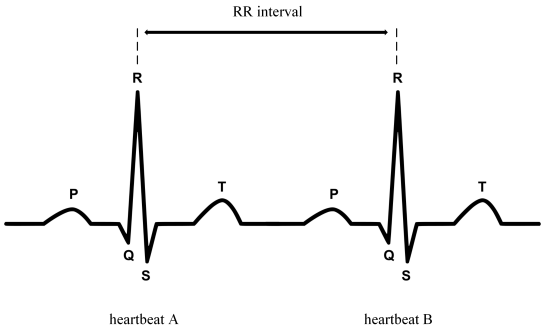

Fig. 2. The structure of normal ECG heartbeat.

- post-RR interval: the RR interval between a given heartbeat and its following heartbeat.

As shown in Fig. 2, the RR interval is the pre-RR interval of heartbeat B and the post-RR interval of heartbeat A.

B. Related work

Over the years, ECG based heartbeat classification algorithms have been developed both in traditional machine learning fields and deep learning fields. For traditional methods, feature extraction is the most critical step. Common used features include wavelet coefficients, high-order cumulants (HOS) , RR intervals and intrinsic mode functions (IMF) [5]. Martis et al. [6] used discrete wavelet transform (DWT) with 4th level approximation as original features. Then three dimensionality reduction algorithms Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA) and Independent Component Analysis (ICA) were independently applied on DWT sub bands for dimensionality reduction. Elhaj et al. [7] combined linear features and nonlinear features. The linear features consist of DWT coefficients with PCA as the feature reduction technique (FRC) nonlinear features consists of HOS with ICA as the FRC.

Traditional methods usually require feature extraction and feature selection, which both are time consuming and require manual intervention. Therefore, several literatures have proposed deep learning based methods. Originally, convolutional neural network (CNN) was widely used. Acharya et al. [8] developed a 9-layer deep CNN and ten-fold cross-validation strategy was implemented. Zhai et al. [9] used dual-beat coupling matrices that was claimed to integrate both morphological and rhythmic information as inputs for a 6-layer CNN classifier. In addition to CNN, Long Short-Term Memory networks (LSTM) [10] is also favored as feature extraction module. Yildirim et al. [11] proposed a bidirectional LSTM based network and implemented a wavelet-based layer to generate ECG signal sequences. Liu et al. [12] stacked bidirectional LSTM (SB-LSTM) and two-dimensional CNN (TD-CNN) and performed a wavelet layer as preprocessing. On the basis of this architecture, attention mechanism was used to highlight important trends and features at key locations [13]. However, these deep learning based methods usually require

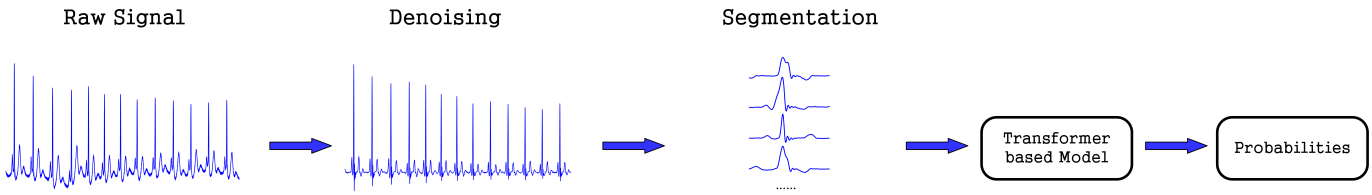

Fig. 3. The overall schematic of the proposed method.

TABLE I
AAMI CLASSES AND CORRESPONDING MIT-BIH CLASS

AAMI Classes	**MIT-BIH Classes**
Non ectopic beat (N)	Normal beat Left bundle branch block Right bundle branch block Atrial escape beat Nodal (junctional) escape beat
Supra-ventricular ectopic beats (S)	Atrial premature beat Aberrated atrial premature beat Nodal (junctional) premature beat Non-conducted P-wave
Ventricular ectopic beats (V)	Premature ventricular contraction Ventricular escape beat
Fusion beat (F)	Fusion of ventricular and normal beat
Unclassifiable beat(Q)	Fusion of paced and normal beat Unclassifiable beat

many complex convolution and recursive structures and only consider the morphological features.

Considering the shortcomings of the existing methods, we propose our Transformer based model and integrated two RR intervals as a supplement.

III. MATERIALS

The Physionet MIT-BIH database [14] is a famous arrhythmia dataset and is widely applied to heartbeat classification and arrhythmia detection. The MIT-BIH database comprises of 48 half-hour excerpts of two-channel ECG recordings obtained from 47 subjects. All signals are acquired at 360 Hz sampling rate and filtered to remove 60 Hz (mains frequency) interference [15]. There are two leads in each record. Most of the lead A is a modified limb lead II (MLII) obtained by placing the electrodes on the chest. Lead B is usually a lead V1 (occasionally V2 or V5, and in one instance V4). Original dataset has a fine-grained classification of the heartbeats. However, according to ANSI/AAMI standard [16] adopted by many methods, all the beats are recommended to divided into 5 groups which is shown in Table I.

Following the AAMI recommendation, 4 records that mainly contains paced beats are removed. We use the left 44 recordings but only considered the lead A channel. In addition, we abandon class Q because there are too few instances of class Q compared with the other four classes.

IV. METHODOLOGY: THE PROPOSED TRANSFORMER BASED MODEL

We show the overview process of the proposed method in Fig. 3. Firstly, the raw signals are denoised and segmented. All heartbeats are then transmitted into our model and the outputs are probabilities of belonging to each candidate class. It should be noted that we will extract the RR interval features while segmenting. In the follow subsections, we step through each process in detail.

A. Denoising

the ECG signal is normally corrupted by various kinds of noise signal [17]. Therefore, it is inappropriate to use the original signal directly and noise reduction is necessary. Following [18] and [19], the denoising process is decomposed into two components: baseline denoising and powerline denoising. To cope with baseline wandering, we use a two-stage median filter [20]. The signals are passed through a median filter with a window length of 200 ms and 600 ms, respectively. After that, a 12 order FIR filter with the frequency cut-off at 35 Hz is applied to eliminate the powerline noise. So far, all the noise reduction work has been completed.

B. Segmentation

Initial signals are successive heartbeats, so segmentation is necessary. As with most literatures [2,21,22,23], segmenting ECG signal firstly requires to find the peak point of the R wave. In our method, we use the annotation files provided by original dataset to locate positions of R peaks. There are a few points to explain. First, we calculate the pre-RR interval and post-RR interval at the same time of segmentation. Second, we abandon those peaks which are too far away (more than 500 points) from the previous one or the next one because these abandoned points have abnormal RR intervals. Third, the new signal will be offset by several points relative to the original signal after powerline denoising so we adjust every value in the annotation files. Having known the R locations, we take 100 points before the R peak and 140 points after the R peak as a heartbeat. Table. II shows the number of the four classes of beats and Fig. 5 shows an example of the four different classes of arrhythmias.

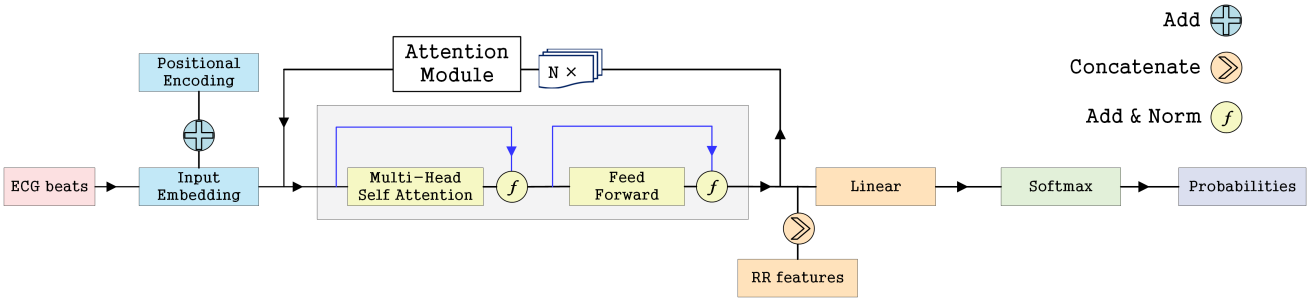

Fig. 4. The architecture of our Transformer model.

TABLE II
A SUMMARY TABLE WITH THE 4 CLASSES OF BEATS

Class	Number of Beats
N	88518
S	2457
V	6782
F	798
Total	**98555**

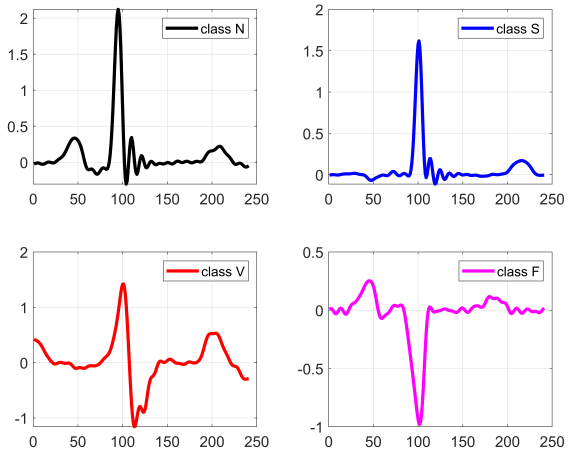

Fig. 5. Example of the four classes of heartbeats.

C. Model details

Transformer is a model architecture relying entirely on an attention mechanism to draw global dependencies between input and output. [24]. Not only in the field of translation, this model can also deal with clinical problem [25]. Like most sequence-to-sequence models, Transformer is also an Encoder-Decoder architecture. However, as ECG signals do not have standard translation, we only use the encoder part. Fig. 4 shows the detail technological process of our Transformer model. Given a heartbeat $X = (x_1, x_2, ..., x_T)$ where T denotes the length of the given heartbeat, Transformer outputs a vector $P = (p_1, p_2, p_3, p_4)$ where p_i denotes the probability that the heartbeat belongs to class i.

1) Input embedding: Referring to the concept of word embedding, the first step is to map the points at each position into the numerical space. To this end, we apply a 1D convolutional layer to the heartbeats to obtain an embedding for each point. We define the output size of the embedding as d_{model}. By cleverly designed padding and kernel size, we ensure the consistency of the length of the sequence before and after convolution.

2) Positional Encoding: Different from RNN using recurrent structure or LSTM using the gates, we encode the position information and superimposes it with the embeddings obtained from the previous step. There are two versions of the positional encoding, one is the learned positional embeddings [26] and the other one is sinusoidal version. Taking advantage of the periodicity of sinusoidal function, we choose the sinusoidal version because it can handle longer sequence lengths than the ones encountered during training. Here is the formula for positional encoding:

$$PE_{(pos,2i)} = sin(pos/10000^{2i/d_{\text{model}}}) \quad (1)$$

$$PE_{(pos,2i+1)} = cos(pos/10000^{2i/d_{\text{model}}}) \quad (2)$$

As can be seen from the formula above, the wavelengths form a geometric progression from 2π to $10000 \cdot 2\pi$. Particularly, for any fixed k, PE_{pos+k} can be obtained by linear transformation of PE_{pos}.

3) Attention module: We stack the attention module N times, each consists of two components. The first one is the multi-head attention which is composed of multiple parallel attention functions. In general, an attention function maps a query to a series of key-value pairs. Specifically, we calculate the similarity between the query and the keys. Then we use softmax function to normalize the weights and finally obtained the weighted sum of the values. In practice, we pack multiple queries and calculated their output at the same time. Denoting the queries as $Q \in R^{n \times d_k}$, the key-value pairs as $K \in R^{m \times d_k}, V \in R^{m \times d_v}$, the whole package Q could calculated as follows:

$$Attention\left(\boldsymbol{q}_i, \boldsymbol{K}, \boldsymbol{V}\right) = \sum_{j=1}^{m} \text{softmax}\left(\frac{\boldsymbol{q}_i \cdot \boldsymbol{k}_j^T}{\sqrt{d_k}}\right) \boldsymbol{v}_j \quad (3)$$

$$Attention\left(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}\right) = \text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d_k}}\right) \boldsymbol{V} \quad (4)$$

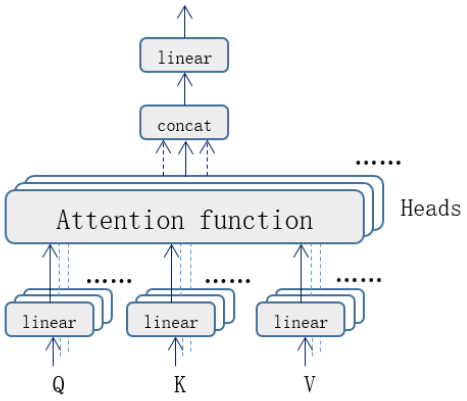

Fig. 6. Multi-Head Attention.

Due to the nature of the softmax function, when the input value is extremely large, the function will fall in where the gradient is extremely small. Thus scaling factor $\frac{1}{\sqrt{d_k}}$ is used to counterweigh this effect.

Back to multi-head attention, though attention mechanisms are used, it may not be possible to completely describe all the dependencies using only one attention function. Thus, we combine several attention functions, each one is called a "head", to jointly attend to information from different representation subspaces at different positions. Instead of passing the queries and key-value pairs directly into the function, different learned linear projections are performed to furtherly extract the critical information. In addition, when all the heads are finished, we concatenate them and projected the results once again, as depicted in Fig. 6. The second component of the attention module is a simple feed forward network. We use two 1D convolution layers with a rectified linear unit (ReLU) in between. Furthermore, we add a residual connection after both the two components, followed by layer normalization.

4) Self-Attention: The attention mechanism we used is called self-attention, which means self-learning. Reflected in the data, the query and key-value pairs are from the same place.

5) RR concatenation: Sometimes using only morphological information to tell the difference between four classes of beats is not enough as there are some rhythmic abnormalities. Therefore, we choose the RR intervals as the solution to the rhythmic anomaly beats. Intuitively, Fig. 7 shows the boxplot of pre-RR intervals and post-RR intervals of the four classes of beats. Through the pre-RR intervals, many class N beats can be distinguished from the other three classes of beats; and post-RR intervals help to distinguish some class F beats from some class N and class V beats. All of these suggest that RR interval has a decent ability to distinguish the four classes of heartbeats. Therefore, we concatenate the two RR features with the abstract features extracted in the attention module.

After the concatenation, we use a 1D convolutional layer and a softmax layer to get the final probabilities.

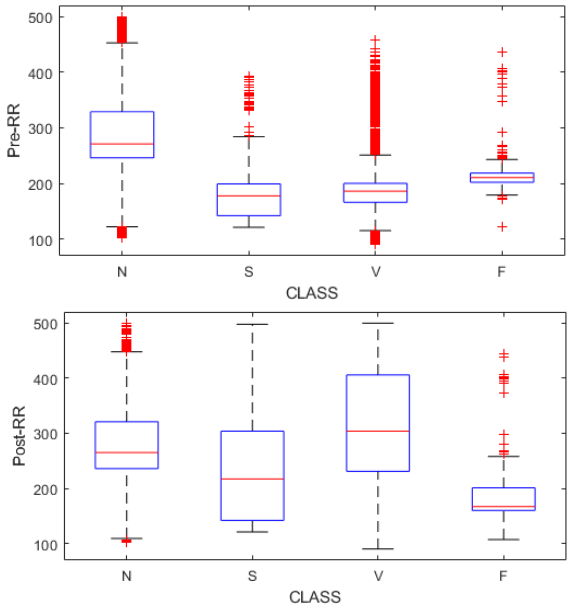

Fig. 7. Pre-RR intervals and Post-RR intervals of the four classes of beats.

V. Experiments and Evaluation

We train our model on a machine with four TITAN XP GPUs each having 12 GB memory and the whole experiments are performed on the same card. There is a problem of data imbalance in the original dataset and this may lead to our model underfitting for the fewer classes. Thus we perform synthetic minority oversampling technique (SMOTE) [27] to resampling the signals.

A. Performance metrics

To evaluate the proposed model, the widely used metrics, i.e., accuracy (Acc), precision (Pre), sensitivity (Sen),and F1 score (F1), are utilized, whose definitions are as follows:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

$$Pre = \frac{TP}{TP + FP} \tag{6}$$

$$Sen = \frac{TP}{TP + FN} \tag{7}$$

$$F_1 = \frac{2Pre \cdot Se}{Pre + Se} \tag{8}$$

where TP is true positives, TN is true negatives, FP is false positives and FN is false negatives. Specificity (Spe), i.e., true negative rate is used when comparing our model with other methods:

$$Spe = \frac{TN}{TN + FP} \tag{9}$$

B. Experimental settings

1) Model parameters: The key parameters of our model are shown in Table III. Particularly, we explain the reason why d_{model} is set to such a small value as follows: In

TABLE III
THE PARAMETERS OF OUR MODEL

Parameters	Values	Meaning
batch size	64	Number of samples fed into model each time
d_{model}	64	Embedding output size & dimensions of Q, K, V
num_layers	3	Number of attention module
num_heads	4	Number of heads in each multi-head module
d_{inner}	512	Output dimension of linear layer

TABLE IV
CONFUSION MATRIX ACROSS TEN EXPERIMENTS OF ORIGINAL DATASET

		Predicted label			
		N	***S***	***V***	***F***
True label	***N***	176546	244	136	104
	S	435	4397	78	0
	V	217	77	13207	69
	F	120	0	41	1429

translation works, the task of embedding is to compress high-dimensional word vectors into low-dimensional vectors of real numbers. The original feature space is huge enough, hence, the parameter is set to 512 or 1024 empirically. However, due to the physiological characteristics of the human body, ECG signal strength will be limited within a certain range, which means there will not be much numerical difference between peaks and troughs. This is why we set the d_{model} to such a small value.

2) Batch normalization: The original model uses dropout to prevent over-fitting and we believed that dropout works well for situations where the feature size is huge. Allowing for our small parameters, we do not use dropout. At the same time, considering the natural differences among patients, we use BN to ensure efficient and reliable training. BN is a normalization which will perform shifting and scaling transformation for activations in each layer.

3) Assessment strategies: We use the hold out method to evaluate our model. All beats are randomly assigned into training set, validation set and test set respectively according the ratio 7:1:2. To mitigate the effects of randomness, we run each experiment ten times and the averaged results are viewed as the final performance.

C. Experiment results

We first conduct experiments on the original dataset and the balanced dataset. For the original dataset, the confusion matrix across the ten experiments is shown in Table IV. Compared with class N and Class V, Class S and class F seems not to obtain very good performance, which maybe because the latter two classes have fewer instances. Table V shows the performance of our model with the four criteria. The "W_Avg" means the weighted average results based on quantity. There is a significant imbalance in original dataset. Therefore, we perform SMOTE algorithm to upsample the original dataset so that the number of all classes is equal to the number of class N. The results of the balanced dataset are shown in Table VI. After upsampling, more than 99.8% of the ECG heartbeats are correctly classified. Particularly, all indicators of the four categories exceeded 99.3%, which suggests that our model is unbiased for each class.

TABLE V
CLASSIFICATION PERFORMANCE OF ORIGINAL DATASET

		Acc %	***Pre %***	***Sen %***	***F1 %***
Class	***N***	99.36	99.56	99.73	99.65
	S	99.58	93.27	89.55	91.32
	V	99.69	98.11	97.33	97.71
	F	99.83	89.40	89.87	89.54
	Avg	99.62	95.09	94.12	94.56
	W_Avg	99.39	99.22	99.23	99.23

D. Contribution of key design features

In this subsection, we explore the contribution of the key design features in our model, as depicted in Table VII. The numbers in parentheses represent the performance gain or decrease compared with the performance obtained by utilizing the complete model. In the absence of BN or RR concatenation, almost all the Acc, Sen and F1 will decrease (except for the four small items written in bold). Class S is the most affected class which maybe because class S contains some premature beats. Therefore, RR intervals and shifting transformation will have a greater impact on the results of Class S. In addition, for class F, the Sen increases without RR concatenation. The same thing happens with the Pre when there is no BN layer. Thus no conclusion can be drawn for class F, which is related to the importance of different performance metrics.

E. Comparison with state-of-the-art methods

To evaluate our model more comprehensive, we compare our results with 6 state-of-the-art methods. For fairness, we adopt corresponding assessment strategy for different methods if feasible. The results is shown in Table VIII. The word "Balanced" in table header means whether the given method uses balanced dataset and we also adopt corresponding treatment. Different strategies include ten-fold cross validation, AAMI standard strategy and "7:1:2" strategy. In particular, the method marked with asterisk is classified only for class S and class V, so we do not use the corresponding strategy. The AAMI standard recommends to divide the recordings into two groups. The training set consist of record numbers 101, 106, 108, 109, 112, 114, 115, 116, 118, 119, 122, 124, 201, 203, 205, 207, 208, 209, 215, 220, 223, and 230; and the testing set consist of record numbers 100, 103, 105, 111, 113, 117, 121, 123, 200, 202, 210, 212, 213, 214, 219, 221, 222, 228, 231, 232, 233, and 234. As Table I shows, each AAMI class consists of several MIT-BIH classes and some MIT-BIH classes exist

TABLE VI
CONFUSION MATRIX AND CLASSIFICATION PERFORMANCE OF BALANCED DATASET

	Predicted label					*Performance*			
		N	***S***	***V***	***F***	***Acc %***	***Pre %***	***Sen %***	***F1 %***
True label	***N***	17581	33	59	29	99.79	99.83	99.32	99.57
	S	19	17669	13	1	99.90	99.80	99.81	99.80
	V	9	2	17677	14	99.86	99.59	99.86	99.72
	F	2	1	1	17698	99.93	99.75	99.98	99.86
					Avg	99.87	99.74	99.74	99.74

TABLE VII
THE CONTRIBUTION OF KEY DESIGN FEATURES

		Without RR and BN			**Without RR**			**Without BN**		
		Pre %	***Sen %***	***F1 %***	***Pre %***	***Sen %***	***F1 %***	***Pre %***	***Sen %***	***F1 %***
Class	***N***	99.33(-0.23)	99.57(-0.16)	99.45(-0.20)	99.43(-0.13)	99.62(-0.11)	99.53(-0.12)	99.47(-0.09)	99.66(-0.07)	99.56(-0.09)
	S	91.49(-1.78)	84.11(-5.44)	87.60(-3.72)	92.54(-0.73)	87.01(-2.54)	89.65(-1.67)	91.54(-1.73)	85.56(-3.99)	88.42(-2.90)
	V	96.44(-1.67)	96.49(-0.84)	96.46(-1.25)	97.57(-0.54)	96.88(-0.45)	97.22(-0.49)	97.18(-0.93)	**97.35(+0.02)**	97.27(-0.44)
	F	**90.57(+1.17)**	87.67(-2.20)	89.03(-0.51)	87.91(-1.49)	**90.63(+0.76)**	89.20(-0.34)	**89.66(+0.26)**	87.42(-2.45)	88.49(-1.05)
	Avg	94.46(-0.63)	91.96(-2.16)	93.14(-1.42)	94.36(-0.73)	93.53(-0.59)	93.90(-0.66)	94.46(-0.63)	92.50(-1.62)	93.44(-1.12)

TABLE VIII
PERFORMANCE COMPARISON OF THE PROPOSED MODEL WITH OTHER STATE-OF-THE-ART METHODS.

	Year	Approach	Strategy	Balanced	Performance %	Proposed Method %
Acharya et al. [8]	2017	9-layer CNN	Ten-fold cross validation	True	acc: 94.03 sen: 96.71 spe: 91.54	acc: **99.80** sen: **99.64** spe: **99.86**
Mathews et al. [28]	2018	Restricted boltzmann machine based deep belief network	AAMI standard	False	acc: 92.70 87.75 pre: 49.46 51.76 sen: 77.42 83.66 F 1: 55.65 54.05	acc: **98.97** pre: **87.71** sen: **88.73** F 1: **87.98**
Zhai et al. [9]*	2018	Dual heartbeat coupling + CNN	Test for only records that class S and V exist	False	class V class S acc: 99.1 97.3 sen: 96.4 85.3	class V class S acc: **99.69 99.58** sen: **97.33 89.55**
Rajesh et al. [29]	2018	ICEEMD + AdaBoost ensemble classifier	Ten-fold cross validation	True	acc: 99.1 sen: 97.9 spe: 99.4	acc: **99.80** sen: **99.64** spe: **99.86**
Liu et al. [12]	2019	Bidirectional LSTM and Two-dimensional CNN	7:1:2	True	acc: 99.5 sen: **99.9** spe: 98.2	acc: **99.87** sen: 99.74 spe: **99.91**
Liu et al. [13]	2019	Attention-based Hybrid LSTM-CNN	7:1:2	True	acc: 99.3 sen: 99.6 spe: 98.1	acc: **99.87** sen: **99.74** spe: **99.91**

only or mostly in testing set. For neural network methods that depend on morphological information, this will lead to the inability to learn features of certain classes. Therefore, we add 3% of the test set into the training set. Third, there are two sets of features in [28], so we provide performance results of both sets here. As shown in the table, our model outperforms the 9-layer CNN [8] by 5.77%, 2.93% and 8.32% in terms of Acc, Sen and Spe, respectively. Though only 3% of the test set is added to the training set, our model achieves an advantage of more than 30% over [28] in terms of Pre. Moreover, our model has less Spe than [12] but achieves higher Acc and Spe.

F. Parameter selection

We describe the process of model parameter selection here. To find the best parameters for our model, we use the control variates method and conducted experiments based on the original dataset. The four key parameters of the search are d_{model}, layers num, heads num and d_{inner}. We leave three variables unchanged and explored the influence of the last variable on the results. As shown in Fig. 8, the optimal combination of parameters are as follows: d_{model} is 64, num_layers is 3, num_heads is 4 and d_{inner} is 512.

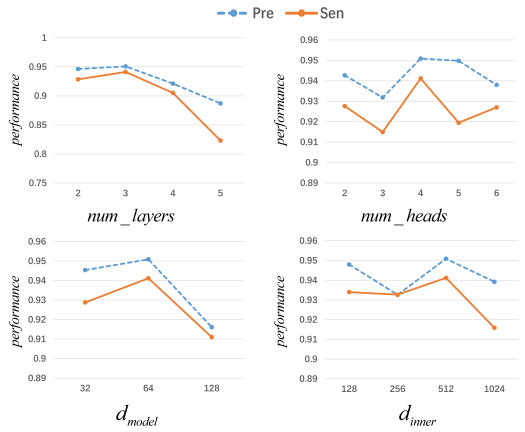

Fig. 8. Performance with different parameters.

VI. Conclusions

This paper proposed a Transformer based model to handle arrhythmia heartbeat classification. By integrating RR intervals, we utilize both morphological and temporal information. Besides, we use BN layer to ensure efficient and reliable training. Our model had achieved state-of-the-art accuracy both in original dataset and balanced dataset. We also perform multiple comparison experiments and the results show that our model can produce higher performance under most circumstances. In the future, we will try to make use of more non-morphological features and simplify the model further.

VII. Acknowledgement

This paper is supported by the National Science Foundation of China (No. 61672161).

References

[1] WHO *et al.*, "World health statistics 2018: monitoring health for the sdgs, sustainable development goals," 2018.

[2] E. J. d. S. Luz, W. R. Schwartz, G. Cámara-Chávez, and D. Menotti, "ECG-based heartbeat classification for arrhythmia detection: A survey," *Computer methods and programs in biomedicine*, vol. 127, pp. 144–164, 2016.

[3] K. Cho, B. Van Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," *arXiv preprint arXiv:1409.1259*, 2014.

[4] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[5] N. E. Huang, Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng, N.-C. Yen, C. C. Tung, and H. H. Liu, "The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis," *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, vol. 454, no. 1971, pp. 903–995, 1998.

[6] R. J. Martis, U. R. Acharya, and L. C. Min, "ECG beat classification using pca, lda, ica and discrete wavelet transform," *Biomedical Signal Processing and Control*, vol. 8, no. 5, pp. 437–448, 2013.

[7] F. A. Elhaj, N. Salim, A. R. Harris, T. T. Swee, and T. Ahmed, "Arrhythmia recognition and classification using combined linear and nonlinear features of ECG signals," *Computer methods and programs in biomedicine*, vol. 127, pp. 52–63, 2016.

[8] U. R. Acharya, S. L. Oh, Y. Hagiwara, J. H. Tan, M. Adam, A. Gertych, and R. San Tan, "A deep convolutional neural network model to classify heartbeats," *Computers in biology and medicine*, vol. 89, pp. 389–396, 2017.

[9] X. Zhai and C. Tin, "Automated ECG classification using dual heartbeat coupling based on convolutional neural network," *IEEE Access*, vol. 6, pp. 27 465–27 472, 2018.

[10] P. Malhotra, L. Vig, G. Shroff, and P. Agarwal, "Long short term memory networks for anomaly detection in time series," in *Proceedings*. Presses universitaires de Louvain, 2015, p. 89.

[11] Ö. Yildirim, "A novel wavelet sequence based on deep bidirectional lstm network model for ECG signal classification," *Computers in biology and medicine*, vol. 96, pp. 189–202, 2018.

[12] F. Liu, X. Zhou, J. Cao, Z. Wang, H. Wang, and Y. Zhang, "Arrhythmias classification by integrating stacked bidirectional lstm and two-dimensional cnn," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 2019, pp. 136–149.

[13] F. Liu, J. Zhou, Z. Wang, H. Wang, and Y. Zhang, "An attention-based hybrid lstm-cnn model for arrhythmias classification," *International Joint Conference on Neural Networks*, p. to appear, 2019.

[14] A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, "Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. e215–e220, 2000.

[15] G. B. Moody and R. G. Mark, "The impact of the mit-bih arrhythmia database," *IEEE Engineering in Medicine and Biology Magazine*, vol. 20, no. 3, pp. 45–50, 2001.

[16] AAMI *et al.*, "Testing and reporting performance results of cardiac rhythm and st segment measurement algorithms," *ANSI/AAMI EC38*, vol. 1998, 1998.

[17] B. Chandrakar, O. Yadav, and V. Chandra, "A survey of noise removal techniques for ECG signals," *International Journal of Advanced Research in Computer and Communication Engineering*, vol. 2, no. 3, pp. 1354–1357, 2013.

[18] R. G. Afkhami, G. Azarnia, and M. A. Tinati, "Cardiac arrhythmia classification using statistical and mixture modeling features of ECG signals," *Pattern Recognition Letters*, vol. 70, pp. 45–51, 2016.

[19] G. Sannino and G. De Pietro, "A deep learning approach for ECG-based heartbeat classification for arrhythmia detection," *Future Generation Computer Systems*, vol. 86, pp. 446–455, 2018.

[20] A. E. Awodeyi, S. R. Alty, and M. Ghavami, "Median filter approach for removal of baseline wander in photoplethysmography signals," in *2013 European Modelling Symposium*. IEEE, 2013, pp. 261–264.

[21] U. R. Acharya, H. Fujita, M. Adam, O. S. Lih, V. K. Sudarshan, T. J. Hong, J. E. Koh, Y. Hagiwara, C. K. Chua, C. K. Poo *et al.*, "Automated characterization and classification of coronary artery disease and myocardial infarction by decomposition of ECG signals: A comparative study," *Information Sciences*, vol. 377, pp. 17–29, 2017.

[22] S. Chen, W. Hua, Z. Li, J. Li, and X. Gao, "Heartbeat classification using projected and dynamic features of ECG signal," *Biomedical Signal Processing and Control*, vol. 31, pp. 165–173, 2017.

[23] Z. Golrizkhatami and A. Acan, "ECG classification using three-level fusion of different feature descriptors," *Expert Systems with Applications*, vol. 114, pp. 54–64, 2018.

[24] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[25] H. Song, D. Rajan, J. J. Thiagarajan, and A. Spanias, "Attend and diagnose: Clinical time series analysis using attention models," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[26] J. Gehring, M. Auli, D. Grangier, D. Yarats, and Y. N. Dauphin, "Convolutional sequence to sequence learning," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017, pp. 1243–1252.

[27] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[28] S. M. Mathews, C. Kambhamettu, and K. E. Barner, "A novel application of deep learning for single-lead ECG classification," *Computers in biology and medicine*, vol. 99, pp. 53–62, 2018.

[29] K. N. Rajesh and R. Dhuli, "Classification of imbalanced ECG beats using re-sampling techniques and adaboost ensemble classifier," *Biomedical Signal Processing and Control*, vol. 41, pp. 242–254, 2018.