Information Sciences 549 (2021) 164–177

Contents lists available at ScienceDirect

Information Sciences

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

FM-ECG: A fine-grained multi-label framework for ECG image classification

Nan Du [a,1], Qing Cao [b,1], Li Yu [a], Nathan Liu [a], Erheng Zhong [a], Zizhu Liu [b], Ying Shen [c], Kang Chen [d,*]

[a] Dawnlight Inc., China
[b] Ruijin Hospital, Shanghai Jiao Tong University School of Medicine, China
[c] School of Intelligent Systens Engineering, Sun Yat-Sen University, China
[d] Department of Cardiology, Ruijin Hospital, Shanghai Jiao Tong University School of Medicine, China

ARTICLE INFO

Article history:
Received 15 June 2020
Received in revised form 7 October 2020
Accepted 10 October 2020
Available online 21 October 2020

Keywords:
Neural networks
Multi-label learning
ECG image classification
Fine-grained classification

ABSTRACT

Recently, increasingly more methods are proposed to automatically detect the abnormalities in Electrocardiography (ECG). Despite their success on public golden standard datasets, two challenges hinder the adoption of existing methods on real-world clinical ECG data in practice. To start with, most methods are designed based on digital signal data while most ECG data in the hospital are stored as images. Additionally, they ignore the correlation among different abnormal cardiac patterns and hence cannot detect multiple abnormalities at the same time. To practically address these challenges, we propose a Fine-grained Multi-label ECG (*FM-ECG*) framework to effectively detect the abnormalities from the real clinical ECG data in the following two aspects. Firstly, we propose to directly detect the abnormalities on the ECG images via a weakly supervised fine-grained classification mechanism, which can discover the potential discriminative parts and adaptively fuse them via image-level annotations only. Secondly, we take the ECG label dependencies into consideration by inferencing with a recurrent neural network (RNN). Experimental results on two real-world large-scale ECG datasets prove the capability of *FM-ECG* comparing with other state-of-the-art methods in ECG abnormally detection. Moreover, visualization analyses on attention parts show that meaningful spatial attention can be effectively learned by *FM-ECG*.

1. Introduction

Electrocardiogram (ECG) is a widely used noninvasive medical test measuring the heart conditions by tracking the heart's electrical activity. ECG contains plenty of information that directly reflects cardiac physiology since its morphological and temporal features are produced from cardiac electrical and structural variations [47]. While an experienced cardiologist can distinguish different types of cardiology abnormalities by visually referencing the ECG waveform pattern, a machine

* Corresponding author.
E-mail addresses: nan@dawnlight.com (N. Du), cq30553@rjh.com.cn (Q. Cao), liyu@dawnlight.com (L. Yu), nathan@dawnlight.com (N. Liu), erheng@dawnlight.com (E. Zhong), liuzizhu1996@sjtu.edu.cn (Z. Liu), sheny76@mail.sysu.edu.cn (Y. Shen), ck11208@rjh.com.cn (K. Chen).
[1] Both authors contributed equally to this research.

https://doi.org/10.1016/j.ins.2020.10.014
0020-0255/

learning (ML) approach can improve the diagnostic efficiency and make the long-term off-hospital monitoring possible [30]. Therefore, numerous methods have been proposed to automatically detect various types of abnormalities from ECG.

Specifically, the existing methods can be further categorized into two types. The first one is the hand-crafted features based methods, which firstly extract the morphological features such as frequency domain [3,33], higher order statistics (HOS) [3,12] and wavelet transforms features [12,18]. Based on these hand-crafted features, various types of machine learning algorithms such as support vector machines (SVM) [12,33], decision trees [3,25] and neural networks [10,31] are used to train a model for anomaly detection. Another one that proposed more recently is the deep neural networks (DNNs) based methods. These methods perform abnormality detection in an end-to-end process by extracting high-level features from ECG data. They directly learn the underlying representations from data via some feature learning architectures [24,34,48].

Although these methods are able to achieve competitive results on some public datasets [5,32], it would be exceptionally challenging when applying these methods in the real clinical environment. First of all, most of the existing methods rely on the time digital ECG data. However, in practice this is not always the case, the ECG data in the real-world is usually collected and stored as images [4], which is a crucial source the algorithms need to consider. Fig. 1 shows a real example that contains several unique properties in ECG images. It can be observed from this sample that, different from digital ECG which is composed of multiple clean and well-separated leads' signals, the ECG image is fuzzy. There are overlappings between waveforms from different leads and the dense covering auxiliary axes (*e.g.*, time and voltage axis) in an image that poses a challenge for accurate information extraction. Besides, the data sampling rate drops from hundreds of *Hz* in digital signals to less than 10 *Hz* in image data that leads to significant information loss.

In such a case, the huge gap between the time digital and image would essentially influence the performance of both hand-crafted features and general DNNs based methods. In detail, for the hand-crafted features based methods, it would be hard to practically extract the required features. Although a potential solution is to first transform the image into digital, this conversion process is time-consuming and the quality of the recognized result is restricted [45,49]. Furthermore, the noise generated during this conversion directly impacts the quality of the model performance. Therefore, even with the strong learning capabilities of DNNs based methods, they cannot detect the discriminative parts precisely due to the noise in images. In ECG diagnosis, the differences between most types of cardiology abnormalities are usually tiny, these subtle differences (*e.g.*, a peak-peak interval or a specific wave) are usually the key points for abnormalities categorization, such as the *ST*-segmentation change, *P* wave height and *T* wave abnormality. Fig. 2 are two samples that the abnormal types

Fig. 1. An illustration of a real ECG image. The horizontal axis represents time, with each time step marked by the vertical lines having a duration of 0.04 s. The vertical axis represents signal magnitudes, which are measured in millivolts (mV).

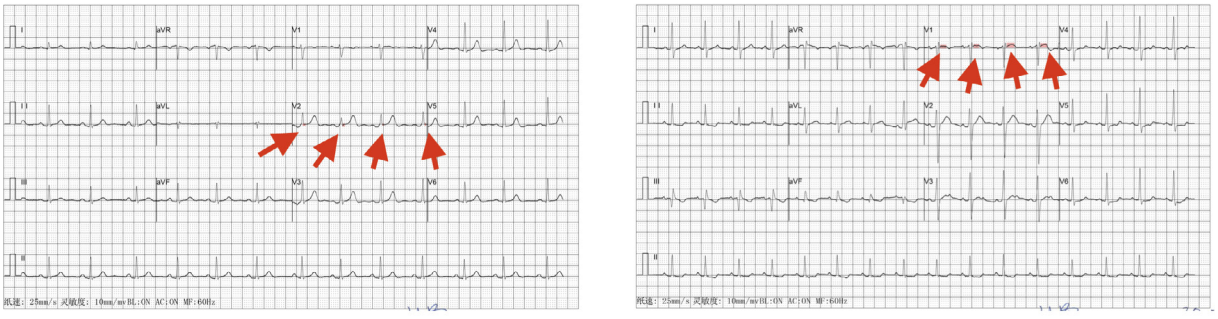

Fig. 2. Two ECG examples of classifying based on subtle but discriminative parts. (a) An example of the discriminative details for 'J-point elevation' abnormal; (b) An example of the discriminative details for 'ST-segment elevation' abnormal. The subtle signs identified (highlighted in red and pointing via arrows) as key parts to detect the abnormalities.

can be distinguished by the key parts. Lack of the ability to capture critical and discriminative parts accurately, the performance of DNNs methods are limited when dealing with the aforementioned challenges regarding image data.

Adding to the challenges, most of the existing works referred to ECG classification as a single-label classification task, in which models are designed to determine whether an image belongs to a specific cardiological abnormal class or not. Actually, most of the time, a clinical ECG record is a potential indication of multiple kinds of abnormalities concurrently. Intuitively, a straightforward way fitting a single-label classification model to a multi-label task is to transform it into multiple single-label classification tasks. However, this approach neglects the relationship between labels, and the number of predicted labels will grow exponentially as the number of categories increases. For instance, 'Left ventricular hypertrophy' and 'ST-T changing' usually co-occur. Unfortunately, such label correlations are scarcely addressed in the literature.

To address the above challenges, we consider ECG abnormality detection on images as a multi-label classification problem, which would assign multi-labels to an image according to the displayed waveforms, and propose a novel end-to-end framework named *FM-ECG* (fine-grained multi-label framework). The proposed method consists of three components: weakly supervised parts discovery, spatial attention of discovered parts and recurrent label inference.

To be specific, given an ECG image, we would recurrently infer its containing abnormalities one by one. For each round's inference, we first dynamically discover the fine-grained parts in a weakly supervised manner, in which only image-level labels are needed. Without any use of part-level annotations (*i.e.*, bounding-boxes on parts), the proposed method can dramatically reduce the annotation time and cost. And then, to diminish the influence of unrelated discovered parts, we employ an attentional region extraction module to automatically extract discriminative parts from all the learned parts. Finally, the label co-occurrence dependency is modeled with the Gated Recurrent Units (GRU) [7] neurons, which transfer the information of label context from previous rounds. The GRU-based component computes the probability of abnormalities sequentially as an ordered prediction path. At each inference round, each label's priori is computed based on the overall representation of the weighted parts and the output of the recurrent neurons. In such a way, the model can implicitly refer to the previous hidden states containing the historical information while predicting subsequent labels.

In summary, there are three main contributions of this work:

- We first investigate the limits of existing ECG classification methods serving in reality and point out that well handling ECG image and considering label dependencies are crucial to improve the clinical performance.
- We propose a novel end-to-end fine-grained multi-label framework named *FM-ECG* to provide a practical solution for abnormaly detection of ECG in the clinical situations. To our best knowledge, *FM-ECG* is the first work of applying fine-grained learning on detecting correlated cardiac abnormalities on ECG images.
- We conduct comprehensive experiments on two real-world datasets, and the results show the effectiveness of *FM-ECG* comparing with state-of-the-art methods.

The rest of this paper is organized as follows. In Section 2, some related works about ECG classification are reviewed. Section 3 introduces the technical details of *FM-ECG*. We discuss and interpret the obtained results from experiments on two real-world datasets in Section 4. Finally, we make conclusions in Section 5.

2. Related work

In this section, we review related works from three corresponding domains including image classification, attention mechanism and fine-grained recognition.

2.1. ECG classification

In the past few decades, most of the studies focus on designing the hand-crafted features extracting from ECG, such as frequency domain [3,33], higher order statistics (HOS) [3,12] and wavelet transforms features [12,18]. Based on the extracted waveform features, different machine learning methods have been put forward to help in the abnormaly detection, such as SVM [12,33], decision trees [3,25], multiLayer Perceptron (MLP) [10,31] and k nearest neighbor algorithm (KNN) [20]. Hybrid methods [35,36,38] combine different solutions together to obtain better performance. Additionally, some MLP based methods are applied to ECG images [14,22]. However, these methods' performance highly relies on the feature quality extracted from the ECG signal. Therefore, a complex feature extraction process and high sampling rates of the ECG signal are critical. But as we mentioned above, the sampling rate of the ECG image is remarkably lower than in the digital.

In such a case, there are some methods that have been proposed to transfer waveform recorded from image to digital [45,49], but these methods contain heuristic time-consuming steps and the accuracy of the output is limited nonetheless. In general, there are two limitations for these hand-crafted based methods: To start with, they are highly dependent on the features designed based on the training dataset, and usually perform inadequately on unseen ECG records due to the over-fitting; Moreover, it is really time-consuming to extract complex features during both training or testing. Different from the hand-crafted features based works, our approach does not need to extract any hand-crafted features, which runs in an end-to-end learning manner.

The success of applying DNNs in different domains has drawn a lot of attention from the ECG automated interpretation community, which can learn complex functions by directly mapping the input to the output without depending on any hand-engineered features. Thus, in the last few years, many researchers are interested in using deep learning models for ECG analysis. A patient-specific ECG monitoring system is proposed by Kiranyaz et al. [24] using a three-layer CNN architecture, which results in good accuracy in the detection of some arrhythmia abnormalities. In [42], an RNN based model is implemented to learn the underlying features of the ECG signal, which utilizes both the morphological and temporal information. To automatically detect four life-threatening arrhythmias, Acharya and et al. [2] proposed a CNN based model, which is of an eleven-layer deep network. In [21], Jun et al. first transformed each ECG beat signal into a 2-D grayscale image and then a 2-D CNN network is applied on it. A 33-layers neural network is utilized in [34] to classify 14 different heart rhythms. While transforming 1-D ECG signals into 2-D ECG images, feature extraction is not necessary for the above methods, these methods still require each lead's signal is stored separately, in which each lead's representation can be learned independently and specifically. However, using these general image architectures, the subtle but discriminative parts for the ECG classification are easily ignored. Table 1 summarizes the aforementioned ECG classification methods.

2.2. Fine-grained recognition

Fine-grained recognition aims to discriminate particular classes in a similar category of objects, for example, the particular species of bird [43]. Note that, Fine-grained recognition is a challenging task due to small inter-class variation which may be subtle differences in the overall appearance between classes. Some works [6,26] formulated the fine-grained image recognition as a two-step solution in which a localization component is designed for locating the informative parts, while a classification component is employed for recognition later. This kind of works is usually called the strongly supervised method, which requires the data providing annotations at the part-level with bounding-box covering. Obviously, obtaining such part-level annotations is labor-intensive, which limits both scalability and practicality of real-world applications.

Recently, increasing more studies focus on a more general weakly supervised setting where part-level annotations are not required. Jaderberg et al. [19] proposed a spatial transformer network to explicitly extract representation from images and learn the location of important regions. Bilinear CNN Models is proposed in [29], where the extracted second-order information is more discriminatively than the convolution features with two parallel feature extractors - AlexNet [27] or VGGNet [37]. The Bilinear method is further improved by [13] via compact bilinear representation with the kernel method. In such a way, computation time and the number of parameters to be learned can be both dramatically reduced. [46] applied part-level top-down attention to train domain-specific deep networks. Cui et al. [9] extract the feature representation of an image by concatenating different order information.

2.3. Attention mechanism

Considering the features extracted directly by a general CNNs cannot well represent fine-grained parts, to learn the better convolution features obtained in CNNs, the attention mechanism has been used in some studies. It aims to learn weights for the elements, expecting key elements to have a heavier weight, and the non-critical words to have a lower weight. The weight of the elements reflects the corresponding element's contribution to the target task.

Generally speaking, the attention mechanism is composed of two parts, determining the areas that need to be paid attention and extracting informative features from them. Fu et al. [50] designed a recurrent neural network-based model to gradually narrow the visible attention area from the input image. While improving the accuracy of the prediction, the model shrinks the visible attention area gradually. In [16], the proposed method SENet first discovered the relationship between the different channels. And then, it automatically detects the importance of each channel and further assigning weights to them based on their importance.

Table 1
Summary of the related ECG classification methods.

Method Type	ECG Source	Detail	Works
Handcrafted features based	Digital	SVM	[12,33]
		Decision Tree	[3,25]
		MLP	[10,31]
		KNN	[20]
		Hybrid	[35,36,38]
	Image	MLP	[14,22]
DNNs based	Digital	1-D Conv	[2,24]
		2-D Conv	[21,42]
		RNN	[42]
Proposed Method	Image	2-D Conv	

3. Model architecture

To better support the ECG classification in a clinical environment, we propose to treat ECG classification as a multi-label fine-grained image classification problem. Correspondingly, a novel end-to-end model: *FM-ECG* is proposed to solve it, which is mainly composed of three parts: weakly supervised parts discovery, spatial attention of discovered parts and recurrent label inference. Fig. 3 demonstrates the deep architecture of *FM-ECG*. Generally speaking, this model is a CNN-RNN like framework, which utilizes CNNs to extract crucial parts via fine-grained learning and attention mechanism at each inference step, and uses RNNs to recurrently infer the abnormal classes considering the class dependencies. Specifically, in each inference round, the weakly supervised parts discovery module captures the fine-grained representations of potential discriminative parts. The spatial attention of discovered parts module adaptively calculates each part's attention and recalibrates its feature representations. The recurrent label inference module uses visual features from both overall waveforms and discovered parts to predict the ECG abnormal label for the current round, which implicitly models the labels dependencies.

3.1. Weakly supervised part discovery

For each round's inference, aiming to obtain a collection of part representations from waveforms, the module of weakly supervised part discovery is deployed, which only requires image-level annotations. The weakly supervised part discovery module mainly consists of two steps: feature extraction and Bilinear Attention Pooling (BAP) [17].

To begin with, the feature extraction starts with a general backbone network extracting the feature maps, which can be represented as $F \in \mathbb{R}^{H\times W\times M}$, where H and W are the spatial dimensions of the input and M denotes the number of channels. Thus, the input image is represented by a set of features maps and each of them demonstrates the pattern spatially matches the image. This process effectively increases the information of the internal representation contained in the image. To filter out irrelevant or weakly relevant regions such as background and uninformative signal segments for detecting abnormality, we further extract an attention map A from learned feature maps F and the hidden state from previous inference round via a series of convolution operations.

After the feature maps F and the attention maps A are achieved, we extract features from these parts using BAP, which is originally designed to extract feature representation by combining features from two sources - the output features from the backbone network, *i.e.,* feature maps and the output features from one or more convolutional layers, *i.e.,* attention maps. The attention maps are supervised in a way to learn the object's part distributions. Specifically, the attention maps are first to split into N maps as Eq. (1)

$$A = \bigcup_{i=1}^{N} a_i, \tag{1}$$

where $a_i \in \mathbb{R}^{H\times W}$ reflects the *i-th* part of the signal. After that, we can represent *i-th* part's corresponding feature map $p_i = g(a_i \odot F)$ as the element-wise multiplication between a_i and F, where $g(.)$ is a Global Average Pooling (GAP) function,

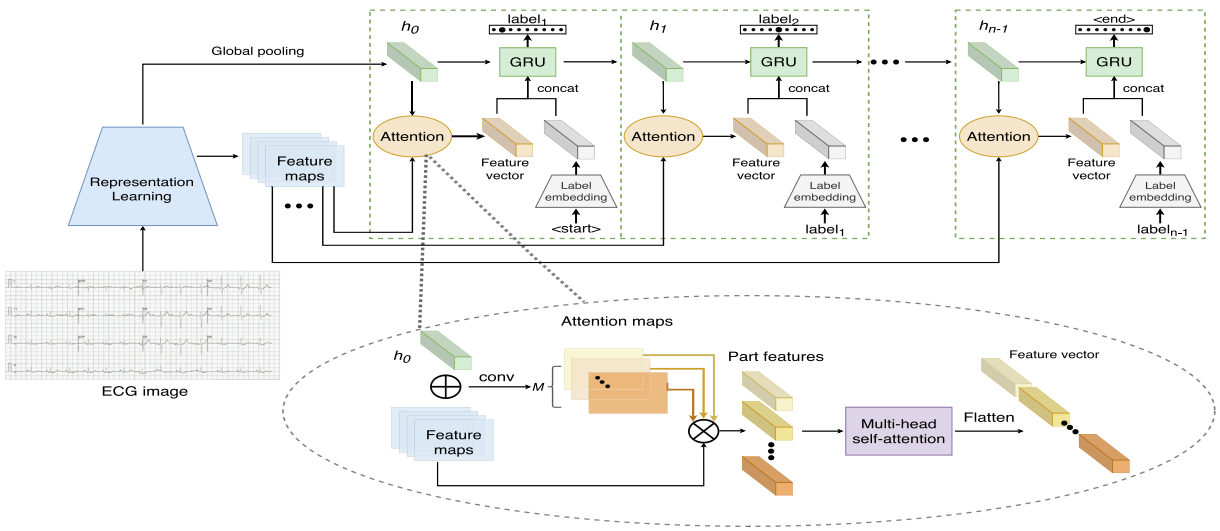

Fig. 3. Overview of the proposed Fine-grained Multi-label ECG *FM-ECG* image classification network. Given a raw input ECG image, we first feed the image into a backbone network that outputs the overall representation feature maps of the image. After that, we recurrently detect informative parts and their corresponding part matrix via bilinear attention pooling. All the fine-grained part information would be weightily combined through an ensemble mechanism that is then used to predict the current round's abnormal classes.

from which informative spatial information can be extracted from different waveform parts. Therefore, the final part feature matrix $P \in \mathbb{R}^{N \times M}$ is stacked by these local features, in which we obtain a different feature vector that represents a specific waveform part for each attention map. Note that this learning process is fully unsupervised, which has the advantage of scaling to large-scale data. The attention learning process is depicted in Fig. 4.

Note that, center loss [44] has been applied as a regularizer to further improve the discriminative power of the learned parts, which simultaneously learns the center of each part for each abnormally class and penalizes the distances between the learned representations and their corresponding class centers. Formulated in Eq. (2)

$$L_{attn} = \sum_{i}^{C}\sum_{j}^{N} ||p_{i,j} - c_{i,j}||_2^2, \tag{2}$$

where C is the number of abnormalities, $p_{i,j} \in \mathbb{R}^{1 \times M}$ is the *j-th* part of the *i-th* abnormally class, and its corresponding center is $c_{i,j} \in \mathbb{R}^{1 \times M}$. The center loss is trainable and can be directly optimized by the standard Stochastic Gradient Descent (SGD). Note that, the original center loss that is designed for the multi-class task, in which one image can only belong to one label, correspondingly each part can only be assigned to one unique center. To adapt it to the multi-label task, we measure the center loss at each round's inference, thus the centers of a specific part are measured dynamically with regards to the current label.

3.2. Spatial attention of discovered parts

Multi-head self-attention [41] is an attention mechanism that is capable to capture different contextual features with multiple individual attention functions. Inspired by it, we utilize the multi-head self-attention mechanism to better ensemble the information from the discovered parts. Different from the original work that applies it to sequence-to-sequence generation tasks, we adopt this mechanism for the image classification task.

Particularly, a learnable spatial multi-head self-attention is employed to localize attentional regions, which are assumed to contain discriminative information, and then re-coordinate them based on their importance. There are mainly two components of the multi-head self-attention mechanism: self-attention and multi-head. The self-attention mechanism allows all parts to interact with each other and find out which they should pay more attention to via scaled dot-product attention function, which maps a query and a collection of key-value pairs to an output. And its output is a weighted sum of the values, in which the weights assigned to them are calculated by the dot-product of the query with all keys. It can be formulated as follows:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d}}\right)V, \tag{3}$$

where $Q \in \mathbb{R}^{N \times d}, K \in \mathbb{R}^{N \times d}$ and $V \in \mathbb{R}^{N \times d}$ denote the matrices for queries, keys and values respectively, d is the number of hidden units and T is the matrix transpose operator. Rather than measuring the attention only once, the multi-head mechanism would utilize the self-attention multiple times in parallel. After that, h parallel heads are adapted to extract information from different parts of channels. We compute the matrix of the *i-th* head as Eq. (4)

$$Head_i = Attention(QW_i^Q, KW_i^K, VW_i^V), \tag{4}$$

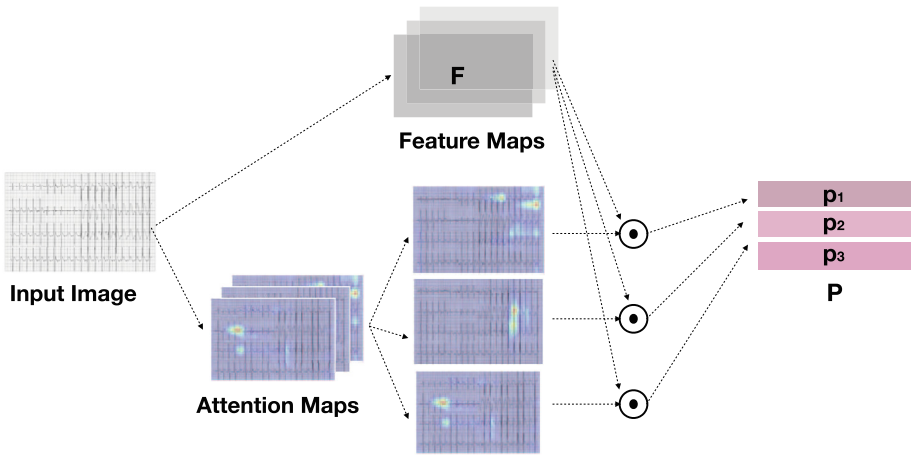

Fig. 4. Illustration of the Attention Map Learning mechanism. Feature map is element-wise multiplied to all the feature matrices. Then, a pooling function is applied at each part and the resulting feature matrices are flattened and concatenated. Each row in the final part matrix is supposed to extract a different waveform part.

where $W_i^Q \in \mathbb{R}^{N\times d/h}$, $W_i^K \in \mathbb{R}^{N\times d/h}$ and $W_i^V \in \mathbb{R}^{N\times d/h}$ are the parameters to be learned in the model, and h is the number of the head. Then, the outputs are simply concatenated and linearly transformed into the required dimensions as Eq. (5) shown,

$$MultiHead(Q,K,V) = Concat(Head_1,\ldots,Head_h)W^h, \tag{5}$$

where W^h is the learnable weight matrix. This mechanism allows the model to jointly attend the information from different representations at different parts [41]. In the multi-head self-attention mechanism, to discover internal connections within discovered parts, we can easily set $Q = V = K = P$. The overall process is depicted in Fig. 5. After multi-head attention, we employ a fully connected feed-forward network to add with the original input P as the final output.

Comparing with CNNs, the self-attention mechanism has two advantages. Firstly, unlike CNNs, self-attention is not limited to fixed window sizes, which means it is easier to fuse any learned parts regardless of their original locations in the image. Secondly, the attention mechanism uses a weighted sum operator to produce output vectors, which is easier than the convolution operator to propagate gradients.

3.3. Recurrent label inference

To better utilize the co-occurrence dependencies information between labels, we adapt a recurrent neural network (RNN) architecture to decode the learned representations from an image to a sequence of labels. In such architecture, both visual features and the hidden state are used to infer ECG abnormal labels iteratively, which implicitly models the label dependencies by the context information maintaining in hidden memory states. In the RNN-based recurrently inference, we use GRU cells, which have been shown superiority for predicting variable-length sequences of labels as well as their explicit modeling of conditional relationships between labels.

The aforementioned weakly supervised part discovery and spatial attention of discovered parts components are used to produce a high-level representation of the image and fine-grained parts, and the GRU is used solely to model conditional dependencies between the various labels. Specifically, the GRU takes the concatenation of part matrix and feature maps as input for the embedding of a previously detected label and produces a hidden state h_t as output using the standard GRU equations.

$$\begin{aligned} r_t &= \delta(W^r x_t + R^r h_{t-1}),\\ z_t &= \delta(W^z x_t + R^z h_{t-1}),\\ \tilde{h_t} &= \phi(W^h x_t + r^t \odot (R^h h_{t-1})),\\ h_t &= z_t \odot h_{t-1} + (1 - z_t) \odot \tilde{h_t}, \end{aligned} \tag{6}$$

where $\delta(.)$ is the sigmoid function, $\phi(.)$ is the tangent function and r_t, z_t and h_t denote reset gate values, update gate values and hidden activations at round t, respectively. Various W matrices are learned parameters applied to the input and recur-

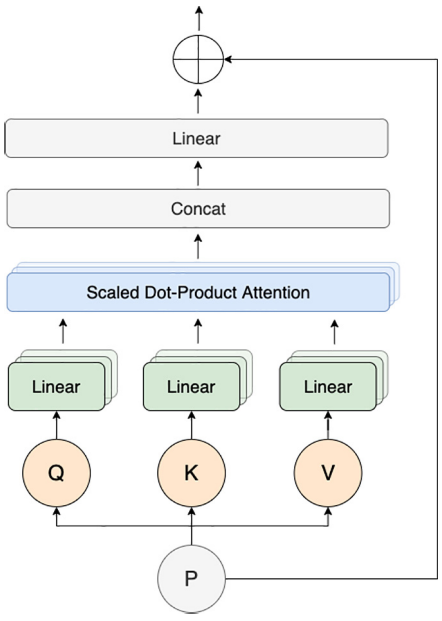

Fig. 5. Illustration of the spatial attention parts mechanism. The part matrix P is first mapped to queries, keys and value matrices (*i.e.*, Q, K and V). After linear transformation, the outputs are adaptively merged via scaled dot-product and concatenated.

rent hidden units. Different from Long Short Term Memory (LSTM) that uses a separate memory cell, the input and forget gates are combined in an update gate z_t in GRU to balance between the previous and the update activation shown in Eq. 6. The reset gate r_t is used to decide whether to maintain the previous activation or not. At each inference step, the input x_t and the previously learned hidden state h_{t-1} are input to the GRU unit, where the previous predicted label information is stored in h_t, such that GRU can exploit the temporal label dependency. A schematic illustration of a GRU neuron is shown in Fig. 6, where each GRU neuron has an input gate, a reset gate, an update gate and an output gate.

For the sake of easy training, given the vector of ground truth labels for an image, we first sort them with a certain order. Therefore, the loss function would punish not only the wrong prediction for non-existing labels but also the wrong order of the labels. During training, a special $< START >$ label is fed to the GRU at the beginning, and GRU is forced to predict a special $< END >$ label at the end of the sequence. For each inference step, the loss is measured based on the ground truth, which would be further used as the input label to the next inference round. During testing, we start by feeding in the $< START >$ label to the GRU, getting an initial label as output. Again, this label is fed as input to the next iteration of the GRU. This process repeats until the model outputs an $< END >$ label or a maximal number of labels is reached. Different from training, the predicted label at the current round would be input to the GRU during testing.

4. Experiments

4.1. Datasets

In this section, two large-scale real-world datasets are introduced and used to study the effectiveness of the proposed method.

CECG: The first data is collected from a tier 1 hospital in China, where there are 27,820 electrocardiograms in the dataset, where each record is stored as a 12-lead ECG image in Portable Network Graphic (PNG) format, where four waveforms are presented in the image: each of the top three consisting by four leads' signal with 10 s duration and 2.5 s duration per lead; the bottom one is a 10 s duration signal for the II lead. Fig. 1 and Fig. 2 are three samples from the dataset. In addition, each ECG record is annotated in the manner of multi-label by multiple clinical ECG experts who used a web-based ECG annotation tool designed for the labeling. To guarantee the labeling consistency among different experts, the majority voting strategy is applied.

DECG: The second dataset used in the experiments is from another tier 1 hospital in China, which we name it DECG. In this dataset, each record is represented as a 12-lead, 500 *Hz* sample rate, 10 s digital signal. Different from the previous dataset which is stored in image format, this dataset is stored in digital format. To better demonstrate the effectiveness of the proposed method on ECG images, we plot them as 192×480 images with the same setting as the above-mentioned dataset, which is illustrated in Fig. 7. It is worth noting that, as the format setting, only 2.5 s' signal is used for each lead, which means only around 25% information is utilized from this the original data.

Table 2 below shows the detailed statistical result of these two datasets. Also, to demonstrate the co-occurrence relationships between abnormalities, we illustrate the co-occurrence matrix of parts of these two dataset's labels in Fig. 8, from which we can easily see that there are obvious dependent relationships between different ECG abnormalities conditions, such as 'Left ventricular hypertrophy' and 'ST-T changing', 'Non-specific abnormal ST-segment' and 'sinus tachycardia', etc.

4.2. Experiment setup

4.2.1. Baselines

To validate the performance of the proposed model for real-world ECG abnormally detection task, we compare it with multiple state-of-the-art approaches, which are listed below:

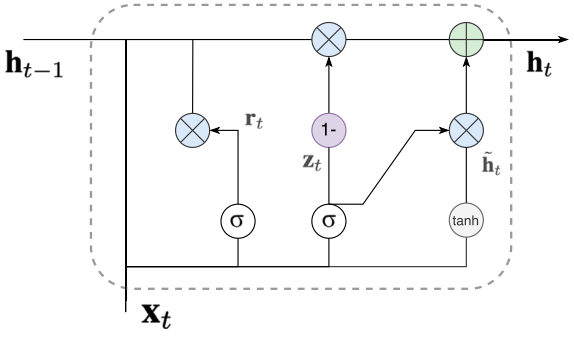

Fig. 6. The structure of GRU neuron.

Fig. 7. Illustration of the generated ECG image. The original signals are plotted as image with vertical and horizontal auxiliary lines.

Table 2
Detailed statistics of the two real-world large-scale datasets.

Dataset	# Samples	Female %	Male %	Age	# Class
CECG	27.8 k	46.5%	53.3%	57.6±17.9	94
DECG	33 k	39.7%	50.8%	45.2±18.7	30

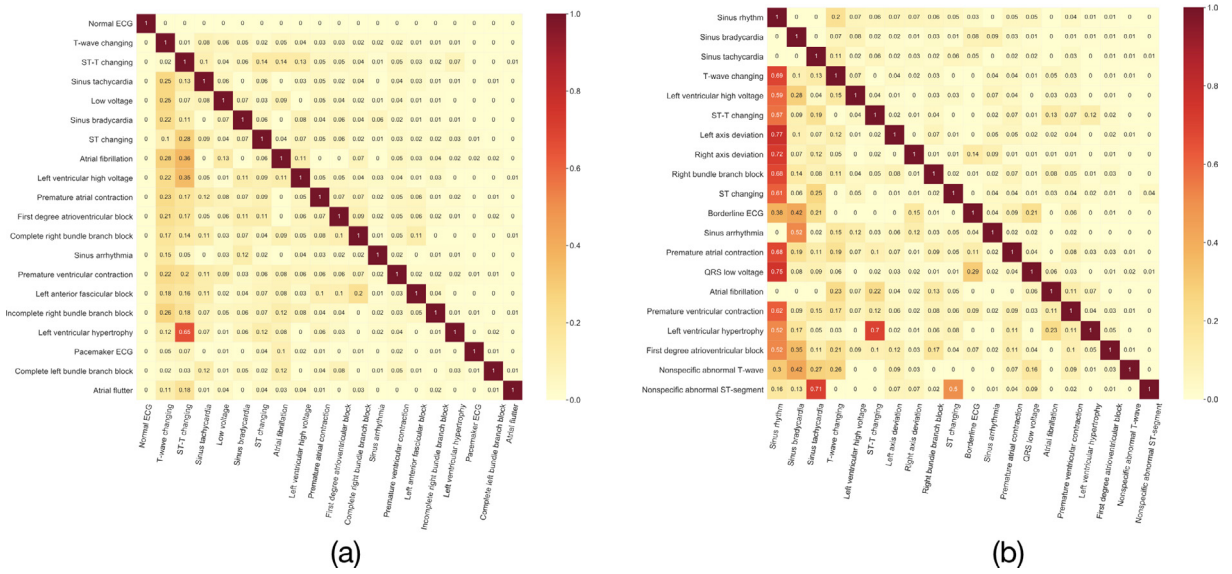

Fig. 8. Illustration of the label co-occurrence matrix between parts of the labels in two datasets. (a) Co-occurrence matrix between parts of the labels in CECG. (b) Co-occurrence matrix between parts of the labels in DECG.

- *B-CNN* [28]: *B-CNN* extracts feature maps from two independent CNN backbones and combines them via the bilinear pooling. Then the bilinear combination is normalized and used for classification tasks.
- *WS-BAN* [17]: In order to extract discriminative local features by the weakly supervised learning, *WS-BAN* learns the attention maps via the attention regularization and attention dropout. Then it performs Bilinear Attention Pooling to extract the sequential part features which are regarded as the final feature representation for classification tasks.
- *PC* [11]: It introduces Pairwise Confusion (*PC*) regularization to bring the predicted probability distributions closer and improve the generalization performance of the model.
- *ECG-CNN* [21]: It is a deep two-dimensional CNN for the ECG arrhythmia classification, which is composed of six convolutional layers, three max-pooling layers and two dense layers. It uses the Xavier initialization and exponential linear units (ELU) [8].
- 34-layer CNN [34]: The network contains 33 layers of convolution followed by a fully connected layer and a softmax. For the raw image input, we replace all 1D convolution layers with 2D convolution layers for comparison.

Besides the above mentioned fine-grained and ECG classification specific methods, we also compare with some widely-used general image classification frameworks, including *VGG*16 [37], *Inception* − *v*3 [39], *Resnet*50 [15] and *EfficientNet* − *b*0 [40].

4.2.2. Evaluation metrics

In all experiments and evaluations, performance study of the proposed method and baselines is based on the results of 5-fold cross-validation. For pre-class evaluation which means the prediction is correct as long as the target label is correctly predicted, precision and recall are used as evaluation metrics. In addition, for install-level evaluation, both precision, recall and F1 are adopted. All these measures are presented in the following equations:

$$Precision = \frac{TP}{TP + FP}, \tag{7}$$

$$Recall = \frac{TP}{TP + FN}, \tag{8}$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}, \tag{9}$$

with the usual interpretation where *TP*, *FP*, *TN* and *FN* stand for True Positives, False Positives, True Negatives and False Negatives, respectively.

4.2.3. Implement Details

The proposed model and the baselines are implemented in Python with TensorFlow [1]. In addition, our experimental system contains 2 Intel Xeon E5 CPUs, 64 GB memory, and 2 NVIDIA GeForce GTX 1080 Ti GPUs. The corresponding software versions are Python 3.6, TensorFlow 1.15, CUDA 10.0, and CUDNN 7.5.

For model settings, we employ *Resnet*34 as our backbone architectures. The detailed setting and architecture are described in Table 3 and Fig. 9, respectively. Moreover, the attention maps are generated by a 1×1 convolutional layer from the feature maps and the hidden state from the previous round. We add one multi-head self-attention mechanism and the number of heads is set to 8. We use the Adam [23] optimizer with $\beta_1 = 0.9$, and $\beta_2 = 0.999$. The initial learning rate is set to 0.001 followed by a decay factor of 0.1 after every 20 epochs, and the batch size is set to 32.

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

In this part, we compare the proposed method with the state-of-the-art methods in the two aforementioned collected ECG data. Results are demonstrated in Table 4 and 5, respectively. For each table, we report both the in-class measurements (left side of the table) and the overall (right side of the table) with Precision, Recall and F1 score. For the sake of space limit, we only report the top 3 classes that contain the most instances. The best results are in bold typeface. Evidently, the proposed model outperforms all the baseline approaches in all the overall measurements and the most in-class measurements.

To be specific, in comparison with *ECG-CNN*, the unsatisfactory performance of the digital-based ECG classification methods: *ECG-CNN* and 34-layer CNN can be observed in both datasets. Interestingly, general image classification frameworks (*i.e.*, *VGG*16, *Inception-v*3, *Resnet*50 and *EfficientNet-b*0) event achieve better performance than these digital-based methods.This finding is consistent with our statement that the performance of the time digital-based ECG methods would be limited when applying to the image. When comparing *FM-ECG* with these general image classification frameworks, it can be observed that the proposed method exhibits an improvement of at most 9.63% and 3.20% in terms of the F1 score on *CECE* and *DECG* datasets, respectively. It illustrates that instead of learning overall representations from entire images, exploiting fine-grained parts plays a key role in abnormaly detection.

Another important observation is that, compared with *B-CNN*, *WS-BAN* and *PC*(*i.e.,* the typical methods based on fine-grained mechanism), *FM-ECG* shows an obvious improvement in both datasets. To summarize, although these methods can extract fine-grained parts from the ECG image, without an effective part attention learning mechanism, the influence of a large number of unrelated parts would limit the performance.

To better demonstrate and interpret the proposed method's performance, statistical tests have also made to compare the F1 score of the proposed methods against baselines in literature for the used datasets. A two-sample t-test is used to calcu-

Table 3
The detail setting of the used *Resnet*34 model.

Layer Name	Output Size	Layers
conv1	96×240	7×7, 64, stride 2
conv2_x	48×120	[3×3 max pool, stride 2 3×3, 64 3×3, 64] $\times$ 3
conv3_x	24×60	[3×3, 128 3×3, 128] $\times$4
conv4_x	12×30	[3×3, 256 3×3, 256] $\times$6
conv5_x	6×15	[3×3, 512 3×3, 512] $\times$3

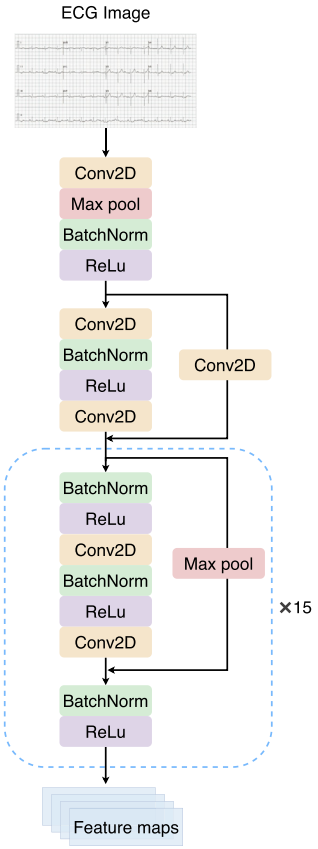

Fig. 9. Architecture of the used *Resnet*34 network.

Table 4
Model evaluation on the CECG dataset.

Method	Normal precision/ recall	T wave change precision/ recall	Sinus bradycardia precision/ recall	Overall precision	Overall recall	F1	P-Value
B-CNN	0.8208/ 0.8422	0.7482/ 0.6626	0.9415/ 0.8816	0.7837	0.5237	0.6279	9.4e-09
WS-BAN	0.8602/ 0.8534	0.7640/ 0.7596	0.9183/ 0.9364	0.7810	0.6782	0.7260	6.9e-04
PC	0.8541/ 0.4780	0.8012/ 0.1651	0.9445/ 0.3947	0.5719	0.6093	0.5879	8.2e-10
ECG-CNN	0.8253/ 0.8708	0.7674/ 0.7524	0.9276/ 0.9276	0.7899	0.6208	0.6952	2.4e-05
34-layer CNN	0.8539/ 0.8897	0.7764/ 0.7466	**0.9459**/ 0.9211	0.7787	0.6845	0.7239	2.9e-03
*VGG*16	0.8228/ 0.8222	0.7306/ 0.6961	0.9221/ 0.8997	0.6891	0.5682	0.6425	1.07e-08
Inception − *v*3	0.8482/ 0.8733	0.7480/ 0.7480	0.9174/ 0.9254	0.7606	0.6859	0.7213	2.1e-03
*Resnet*50	0.8448/ 0.8897	0.7707/ 0.7422	0.9279/ 0.9320	0.7836	0.6826	0.7296	2.6e-03
EfficientNet − *b*0	0.8356/ 0.8927	0.7647/ 0.7437	0.9332/ 0.9496	0.7722	0.6902	0.7338	3.2e-03
Proposed Method	**0.8707**/ **0.9008**	**0.8179**/ **0.8059**	0.9370/ **0.9583**	**0.7923**	**0.6910**	**0.7388**	

Table 5
Model evaluation on the DECG dataset.

Method	Sinus rhythm precision/ recall	Sinus bradycardia precision/ recall	Sinus tachycardia precision/ recall	Overall precision	Overall recall	F1	P-Value
B-CNN	0.9752/ 0.9786	0.9864/ 0.9896	0.9814/ 0.9894	0.9021	0.7998	0.8522	4.9e-03
WS-BAN	0.9593/ 0.9766	0.9757/ 0.9915	0.9743/ 0.9870	0.9026	0.7913	0.8433	2.0e-04
PC	0.9725/ 0.6053	0.9878/ 0.6042	**0.9862**/ 0.5871	0.5941	0.5467	0.5692	6.4e-13
ECG-CNN	0.9593/ 0.9766	0.9757/ 0.9915	0.9743/ 0.9870	0.9026	0.7913	0.8433	2.2e-03
34-layer CNN	0.9705/ 0.9774	0.9856/ 0.9805	0.9822/ 0.9894	0.8972	0.8171	0.8553	5.3e-03
*VGG*16	0.9708/ 0.9804	0.9726/ 0.9855	0.9841/ 0.9789	0.8781	0.7881	0.8307	9.0e-05
Inception − *v*3	0.9742/ 0.9817	0.9839/ 0.9922	0.9838/ 0.9886	0.8906	0.8343	0.8642	8.8e-03
*Resnet*50	0.9698/ 0.9814	0.9801/ 0.9922	0.9855/ 0.9837	0.8982	0.8297	0.8626	8.6e-03
EfficientNet − *b*0	0.9755/ 0.9789	0.9845/ **0.9935**	0.9775/ 0.9910	0.8891	0.8358	0.8627	8.8e-03
FM-ECG	**0.9765**/ **0.9843**	**0.9902**/ 0.9837	0.9860/ **0.9912**	**0.9042**	**0.8359**	**0.8687**	

late the P-value, which measures the probability of how likely the observed difference comes by chance. A lower P value dictates the less likely the superior performance of the proposed method achieved coincidentally. From the results show in Table 4 and 5, it can be observed that the difference between the proposed method and baselines is considered to be statistically significant.

4.2.5. Effects of different components

The proposed *FM-ECG* mainly consists of the components of weakly supervised part discovery, spatial attention of discovered parts and recurrent label inference. To investigate the effectiveness of different modes, we conducted ablation experiments on CECG and DECG datasets. We investigate the effect of these components by training the *FM-ECG* without a certain component and comparing it with the full model.

From Table 6 that describes detailed evaluation results, we make the following observations: Frist, it is noticeable that all configurations outperform the basic backbone network with at least a 0.5% margin, each configuration perform better than the backbone network as expected which demonstrates the effectiveness and robustness of the proposed method. Second, without the recurrent label inference component, the Recall drop 0.25% and 1.76% on two datasets, respectively, which makes the biggest drop on Recall. This implies that this component plays a crucial role in escalating the performance, which can explicitly learn the label dependencies. Fig. 10 exhibits several example predictions from the two used datasets.

Third, when only one component is used, the performance gets worse than their combination. Specifically, without part discovery, part attention and recurrent label inference mechanism, the F1 score drops 0.63% and 0.61%, 0.36% and 0.36%, 0.54% and 0.55% on two datasets, respectively. It is obvious to see that all these three components are mutually correlated, and they can reinforce each other.

4.2.6. Attention visualization

To dive deep into the model, we visualize label-specific attentional regions in Fig. 11. We select the top 5 parts with the highest attention. As shown here, regions with label-related semantics are highlighted, while less informative regions present weak attention. We can observe that two findings: Firstly, the learned discriminative regions about 'T-wave changing' and 'Left ventricular high voltage' is subtle, which are shown on small waves or segments; Secondly, most of these highlight

Table 6
Ablation study of different components on CECG and DECG.

Dataset	Method	Overall Precision	Overall Recall	F1
CECG	*Backbone only*	0.7834	0.6835	0.7269
	FM-ECG w/o part discovery	0.7908	0.6893	0.7325
	FM-ECG w/o part attention	0.7901	0.6902	0.7352
	FM-ECG w/o recurrent label inference	0.7878	0.6885	0.7334
	FM-ECG	**0.7923**	**0.6910**	**0.7388**
DECG	*Backbone only*	0.8912	0.8297	0.8583
	FM-ECG w/o part discovery	0.8927	0.8345	0.8626
	FM-ECG w/o part attention	0.8966	0.8358	0.8651
	FM-ECG w/o recurrent label inference	0.9036	0.8183	0.8632
	FM-ECG	**0.9042**	**0.8359**	**0.8687**

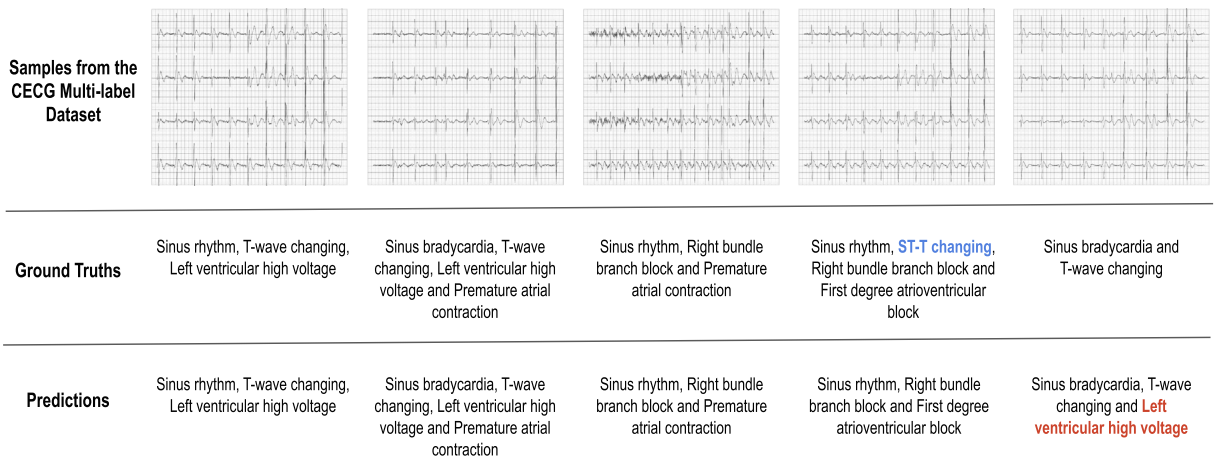

Fig. 10. Example Images and Predicted labels on the Datasets. Red label(s) predictions indicate false positives, while blue label(s) predictions are false negatives.

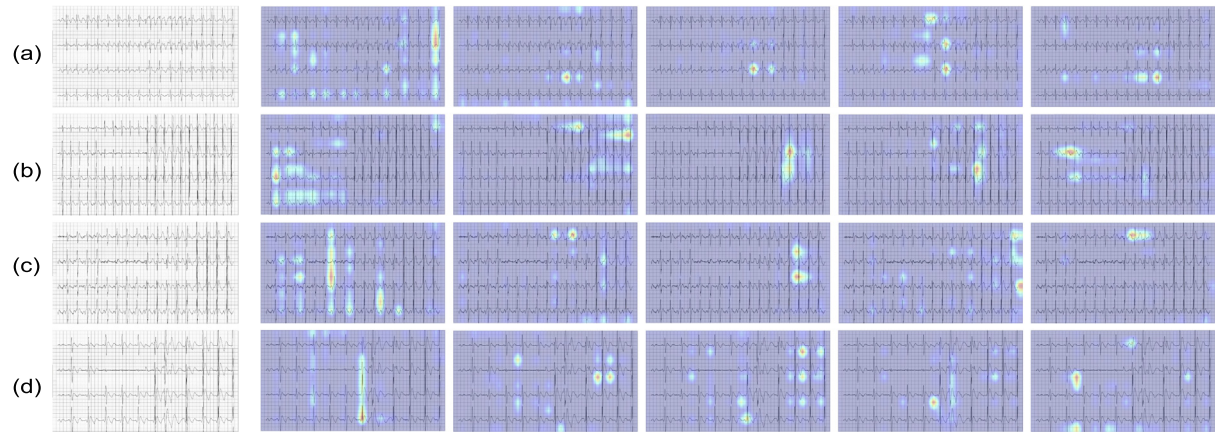

Fig. 11. Examples of raw image and the corresponding attention maps for a specific abnormal label generated by the proposed method. Regions in red implies strong attention, while blue indicates weak attention. (a) T-wave changing; (b) T-wave changing; (c) T-wave changing; (d) Left ventricular high voltage.

parts exist at the same time (*i.e.*, with the same x-coordinate) but on different leads. This is because when the T-wave abnormally happens, it is likely to be observed by multiple leads simultaneously. Both of these two observations are in accordance with the clinical phenomenon and the reviewing way for this abnormal type by ECG experts, which illustrates that the parts captured by the proposed method are informative and class-specific.

5. Conclusion

In this paper, aiming to solve the inadaptability of existing methods applied on real clinical ECG records which are in image format and contain multiple labels, we propose a novel multi-label fine-grained network, named *FM-ECG*. *FM-ECG* has the capability to find out the subtle but critical fine-grained parts from the ECG waveforms and further inference multiple abnormalities, which well address the challenges of making multi-label classification in clinical ECG images.

Moreover, *FM-ECG* is a unified model that consists of several components: (1) weakly supervised part discovery, (2) spatial attention of discovered parts and (3) recurrent label inference. To be specific, the weakly supervised part discovery component is responsible for capturing fine-grained parts from raw ECG images, then the spatial attention of discovered parts component is designed to further weight these selected parts based on their class-specific discriminative importance. Afterward, the recurrent label inference component is used to capture the underlying ECG label dependencies and infer multiple abnormalities iteratively. We evaluate our network on two real-world ECG image datasets, CECG and DECG. The experimental results indicate that *FM-ECG* is an effective framework to assist the clinicians to detect abnormalities from ECG images and the strong ability to apply to the clinical environment.

In the future, we will further study the proposed *FM-ECG* in the following aspects: 1) investigate more sophisticated attention selection strategy, learning to select more informative waveforms parts, 2) incorporate non-ECG clinical information, such as gender, age and symptoms that can be jointly learned with ECG images. Ideally, this additional information would benefit the abnormaly detection, and 3) include more ECG abnormal classes for analysis, especially rare abnormally classes, and develop it as an industrial system for automated ECG image analysis in the real world.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was financially supported by the project of the ‘‘National Double First - Class” and ‘‘Shanghai-Top-Level” high education initiative at Shanghai Jiao Tong University School of Medicine.

References

[1] M. Abadi, A. Agarwal, P. Barham, et al., TensorFlow: Large-scale machine learning on heterogeneous systems, 2015..
[2] U.R. Acharya, H. Fujita, S.L. Oh, Y. Hagiwara, J.H. Tan, M. Adam, Application of deep convolutional neural network for automated detection of myocardial infarction using ecg signals, Inf. Sci. 415 (2017) 190–198.

[3] R.G. Afkhami, G. Azarnia, M.A. Tinati, Cardiac arrhythmia classification using statistical and mixture modeling features of ecg signals, Pattern Recogn. Lett. 70 (2016) 45–51.
[4] F. Badilini, T. Erdem, W. Zareba, A. Moss, Ecgscan: a method for digitizing paper ecg printouts, J. Electrocardiol. 36 (2003) 40, https://doi.org/10.1016/j.jelectrocard.2003.09.009.
[5] R.D. Bousseljot, D. Kreiseler, A. Schnabel, Nutzung der ekg-signaldatenbank cardiodat der ptb über das internet, Biomedizinische Technik - BIOMED TECH 40 (1995) 317–318.
[6] Y. Chai, V. Lempitsky, A. Zisserman, Symbiotic segmentation and part localization for fine-grained categorization, in: 2013 IEEE International Conference on Computer Vision, 2013, pp. 321–328.
[7] K. Cho, B. van Merrienboer, Ç. Gülçehre, F. Bougares, H. Schwenk, Y. Bengio, Learning phrase representations using RNN encoder-decoder for statistical machine translation. CoRR abs/1406.1078, 2014..
[8] D.A. Clevert, T. Unterthiner, S. Hochreiter, Fast and accurate deep network learning by exponential linear units (elus). Under Review of ICLR2016 (1997), 2015..
[9] Y. Cui, F. Zhou, J. Wang, X. Liu, Y. Lin, S. Belongie, Kernel pooling for convolutional neural networks, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3049–3058.
[10] Z. Dokur, T. Olmez, Ecg beat classification by a novel hybrid neural network, Computer Methods Programs Biomed. 66 (2001) 167–181.
[11] A. Dubey, O. Gupta, P. Guo, R. Raskar, R. Farrell, N. Naik, Training with confusion for fine-grained visual classification. CoRR abs/1705.08016, 2017..
[12] F.A. Elhaj, N. Salim, A.R. Harris, T.T. Swee, T. Ahmed, Arrhythmia recognition and classification using combined linear and nonlinear features of ecg signals, Comput. Methods Programs Biomed. 127 (2016) 52–63.
[13] Y. Gao, O. Beijbom, N. Zhang, T. Darrell, Compact bilinear pooling. CoRR abs/1511.06062, 2015..
[14] P. Hao, X. Gao, Z. Li, J. Zhang, F. Wu, C. Bai, Multi-branch fusion network for myocardial infarction screening from 12-lead ecg images, Comput. Methods Programs Biomed. 184 (2020), https://doi.org/10.1016/j.cmpb.2019.105286 105286.
[15] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition. CoRR abs/1512.03385, 2015..
[16] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks. CoRR abs/1709.01507, 2017..
[17] T. Hu, H. Qi, C. Huang, Q. Huang, Y. Lu, J. Xu, Weakly supervised local attention network for fine-grained visual classification. CoRR abs/1808.02152, 2018..
[18] O. Inan, L. Giovangrandi, G. Kovacs, Robust neural-network-based classification of premature ventricular contractions using wavelet transform and timing interval features, IEEE Trans. Bio-medical Eng. 53 (2006) 2507–2515.
[19] M. Jaderberg, K. Simonyan, A. Zisserman, K. Kavukcuoglu, Spatial transformer networks. CoRR abs/1506.02025, 2015..
[20] P. Janbakhshi, M. Shamsollahi, Sleep apnea detection from single-lead ecg using features based on ecg-derived respiration (edr) signals, IRBM 39 (2018) 206–218, https://doi.org/10.1016/j.irbm.2018.03.002.
[21] T.J. Jun, H.M. Nguyen, D. Kang, D. Kim, Y. Kim, ECG arrhythmia classification using a 2-d convolutional neural network. CoRR abs/1804.06812, 2018..
[22] M. Junior, M. Gurgel, L. Bezerra Marinho, N. Nascimento, S. Silva, S. Alves, G. Ramalho, P.P. Filho, Evaluation of heart disease diagnosis approach using ecg images, in: 2019 International Joint Conference on Neural Networks (IJCNN), 2019, pp. 1–7, https://doi.org/10.1109/IJCNN.2019.8851807.
[23] D. Kingma, J. Ba, Adam: A method for stochastic optimization, in: International Conference on Learning Representations, 2014.
[24] S. Kiranyaz, T. Ince, M. Gabbouj, Real-time patient-specific ecg classification by 1-d convolutional neural networks, IEEE Trans. Biomed. Eng. 63 (2016) 664–675.
[25] V. Krasteva, I. Jekova, R. Leber, R. Schmid, R. Abächerli, Superiority of classification tree versus cluster, fuzzy and discriminant models in a heartbeat classification system, PLOS ONE 10 (2015) 1–29.
[26] J. Krause, H. Jin, J. Yang, L. Fei-Fei, Fine-grained recognition without part annotations, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 5546–5555.
[27] A. Krizhevsky, I. Sutskever, G. Hinton, Imagenet classification with deep convolutional neural networks, Neural Inform. Process. Syst. 25 (2012).
[28] T. Lin, A. RoyChowdhury, S. Maji, Bilinear convolutional neural networks for fine-grained visual recognition, IEEE Trans. Pattern Anal. Mach. Intell. 40 (2018) 1309–1322.
[29] T.Y. Lin, A. RoyChowdhury, A.. Maji, Bilinear CNN models for fine-grained visual recognition. CoRR abs/1504.07889, 2015. URL:http://arxiv.org/abs/1504.07889..
[30] J. Ma, M. Dong, R d of versatile distributed e-home healthcare system for cardiovascular disease monitoring and diagnosis, in: IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI), 2014, pp. 444–447.
[31] R.J. Martis, U.R. Acharya, L.C. Min, Ecg beat classification using pca, lda, ica and discrete wavelet transform, Biomed. Signal Process. Control 8 (2013) 437–448.
[32] G.B. Moody, R.G. Mark, The impact of the mit-bih arrhythmia database, IEEE Eng. Med. Biol. Mag. 20 (2001) 45–50.
[33] S. Raj, K.C. Ray, O. Shankar, Cardiac arrhythmia beat classification using dost and pso tuned svm, Comput. Methods Programs Biomed. 136 (2016) 163–177.
[34] P. Rajpurkar, A.Y. Hannun, M. Haghpanahi, C. Bourn, A.Y. Ng, Cardiologist-level arrhythmia detection with convolutional neural networks. CoRR abs/1707.01836, 2017..
[35] E. Ramirez, P. Melin, G. Prado-Arechiga, Hybrid model based on neural networks, type-1 and type-2 fuzzy systems for 2-lead cardiac arrhythmia classification, Expert Syst. Appl. 126 (2019) 295–307.
[36] E. Ramírez, P. Melin, G. Prado-Arechiga, Hybrid Model Based on Neural Networks and Fuzzy Logic for 2-Lead Cardiac Arrhythmia Classification, 2020..
[37] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition. CoRR abs/1409.1556, 2015..
[38] M.E.R. Soria, Hybrid intelligent system for cardiac arrhythmia classification with fuzzy k-nearest neighbors and neural networks combined with a fuzzy system, Expert Syst. Appl. (2012).
[39] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision. CoRR abs/1512.00567, 2015..
[40] M. Tan, Q.V. Le, Efficientnet: Rethinking model scaling for convolutional neural networks. CoRR abs/1905.11946, 2019..
[41] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need. CoRR abs/1706.03762, 2017..
[42] G. Wang, C. Zhang, Y. Liu, H. Yang, D. Fu, H. Wang, P. Zhang, A global and updatable ecg beat classification system based on recurrent neural networks and active learning, Inf. Sci. 501 (2018), S0020025518305115.
[43] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, P. Perona, Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001. California Institute of Technology, 2010..
[44] Y. Wen, K. Zhang, Z. Li, Y. Qiao, A discriminative feature learning approach for deep face recognition, LNCS 9911 (2016) 499–515.
[45] L.E. Widman, G.L. Freeman, A-to-d conversion from paper records with a desktop scanner and a microcomputer, Comput. Biomed. Res. 22 (1989) 393–404.
[46] T. Xiao, Y. Xu, K. Yang, J. Zhang, Y. Peng, Z. Zhang, The application of two-level attention models in deep convolutional neural network for fine-grained image classification. CoRR abs/1411.6447, 2014..
[47] H. Yuki, F. Hamido, O.S. Lih, J.H. Tan, R.S. Tan, E.J. Ciaccio, A.U. Rajendra, Computer-aided diagnosis of atrial fibrillation based on ecg signals: A review, Inf. Sci. 467 (2018) 99–114.
[48] M.D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks. CoRR abs/1311.2901, 2013..
[49] Z. ni Zhang, H. Zhang, T. ge Zhuang, One-Dimensional Signal Extraction Of Paper-Written ECG Image And Its Archiving, in: T.R. Hsing (Ed.), Visual Communications and Image Processing II, International Society for Optics and Photonics, SPIE, 1987, pp. 419–423.
[50] H. Zheng, J. Fu, T. Mei, J. Luo, Learning multi-attention convolutional neural network for fine-grained image recognition, in: 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5219–5227.