Classification of 12-lead ECG Signals with Adversarial Multi-Source Domain Generalization

Hosein Hasani, Adeleh Bitarafan, Mahdieh Soleymani Baghshah

Department of Computer Engineering, Sharif University of Technology, Tehran, Iran

Abstract

The ECG classification is a critical task in the early and correct diagnosis of cardiovascular diseases. Although various models have been developed to tackle the heartbeat classification problem, their performance degrades on ECG signals recorded in varied testing conditions due to the distribution discrepancy among different sources of data. In this work, we have developed a multi-source domain generalization model to address the distribution discrepancy problem that occurred when the collection of the data is from multiple sources with various acquisition conditions. We have employed a combination of convolutional neural network (CNN) and long short term memory (LSTM) for feature extraction. Further, we exploit the adversarial domain generalization method to overcome probable heterogeneity between the train and test datasets. To increase generalization, we also utilized different augmentation techniques including random ECG pad and crop, adding low-frequency artifacts, and lead dropout. We evaluate our proposed model on cardiac abnormality classification based on 12-lead ECG signals associated with "Classification of 12-lead ECGs for the PhysioNet/Computing in Cardiology Challenge 2020". Our method, achieved a challenge validation score of 0.609*, and full test score of* 0.437 *placing us (Sharif AI Team) 5th out of 41 teams in the final official ranking.*

1. Introduction

The importance of the quick and exact diagnosis of cardiac abnormalities is remarkable since it can cause the most serious life-threatening diseases. To interpret cardiac conditions, physicians exploit the standard 12-lead ECG that record the heart electrical activity from 12 electrodes embedded on the body surface. Since the analysis of 12-lead ECG signals is challenging and time-consuming, automatic detection can be represented as an assistant for physicians. Accordingly, the PhysioNet/Computing [1] has presented Cardiology Challenge 2020 in which the subject is distinguishing cardiac abnormalities. It encourages researchers to develop techniques for multi-label classification of 12-lead ECG signals into 27 classes.

In recent years, various machine learning and deep learning techniques have been proposed to detect different heart anomalies from ECG signals. Traditional methods employ different classification techniques such as K-Nearest Neighbour [2], and Support Vector Machines [3] on hand-engineered features. Approaches based on deep neural networks proposed in recent years generally vary in their network architectures. The network developed in [4] utilize only convolutional layers to detect different kinds of arrhythmia. The introduced methods in [5, 6] combine convolutional layers with LSTMs to classify ECG signals. However, the above studies do not consider distribution discrepancy between different benchmarks caused by different recording conditions, which is a crucial problem for practical ECG interpretation. Recently, [7–9] apply a deep domain adaptation approach to reduce the discrepancy between the training and test distributions that appeared due to the varied acquisition conditions for ECG signals. To alleviate discrepancies among ECG signals from different datasets, [7] utilizes fully connected hidden layers, [8] benefits from a strategy based on the multi-layer multi-kernel maximum mean discrepancy, and [9] proposes a cluster-aligning loss and a cluster-separating loss. However, these methods do not consider the distribution discrepancy within the training set which can be caused because data have been collected from multiple sources with different characteristics. Furthermore, conventionally they suppose that test samples are available during the training phase and do not focus on generalization on the unseen test dataset. To the best of our knowledge, we are the first to tackle the multi-source domain generalization problem to improve the performance of the ECG classification task in the practical area. In this study, our objective is to address this challenge aiming to improve 12-lead ECG classification performance. Therefore, we aim to present a practical diagnostic tool as a medical assistant that can automatically classify 12-lead ECG signals assembled from the different clinics with various recording conditions using a deep multi-source domain generalization method.

2. Methods

2.1. Pre-processing and Augmentation

Due to the environmental conditions, ECG signals typically have some low and high-frequency noise components that may vary in different datasets. Since we use deep learning frameworks with high capacity for classification, we have not cleaned these frequency components from the dataset. Alternatively, we have randomly added or filtered frequency components as a data augmentation technique to increase the effective size of the dataset. Moreover, this procedure has an additional advantage as it may increase domain generalization capability in situations where training and test datasets are gathered from different sources and do not have the same statistical properties. Besides, with a very low probability, the following procedures have also been taken on the training data as data augmentation techniques:

- Randomly substituting the signal of one lead with entirely zero signal or a low-frequency noise,
- Randomly shuffling the position of two or more leads,
- Randomly inverting the signal of one or more leads,
- Randomly applying a band-pass filter on ECG signal,
- Randomly scaling the ECG signal or adding a random offset to it.

The length of ECG signals fed into the model is considered fixed. Shorter signals are randomly padded and longer signals are randomly cropped during training. Since the data from different domains have a different distribution of signal lengths, with a low probability, long signals are also randomly cropped to the length lower than the input size, and then randomly padded to fit the input of the classifier.

All ECG signals are resampled to 250 Hz and normalized by a constant scale. We just used ECG signals that have at least one label from 27 scored classes and exclude other data from the training dataset. We also adopted 5% of training data for validation to first, calculate the optimum thresholds for classification and then, choose the best models based on their validation performance.

2.2. Feature Extraction and Classification

Figure 1 shows the structure of the proposed method. Motivated by previous studies [10, 11], we have used two parallel convolutional networks for feature extraction. One network exploits smaller kernel sizes to extract finer features and another network extract coarser ones using larger convolution kernels.

The ECG signal has a sequential nature. When using a long time series signal for classification, the classifier may encounter the curse of dimensionality. Hence, we utilized the Recurrent Neural Networks (RNNs) to integrate time-related features before applying classifiers on them. To this end, two bidirectional LSTMs are applied on the top of the parallel CNNs to produce time-independent low-dimensional feature vectors (Figure 1).

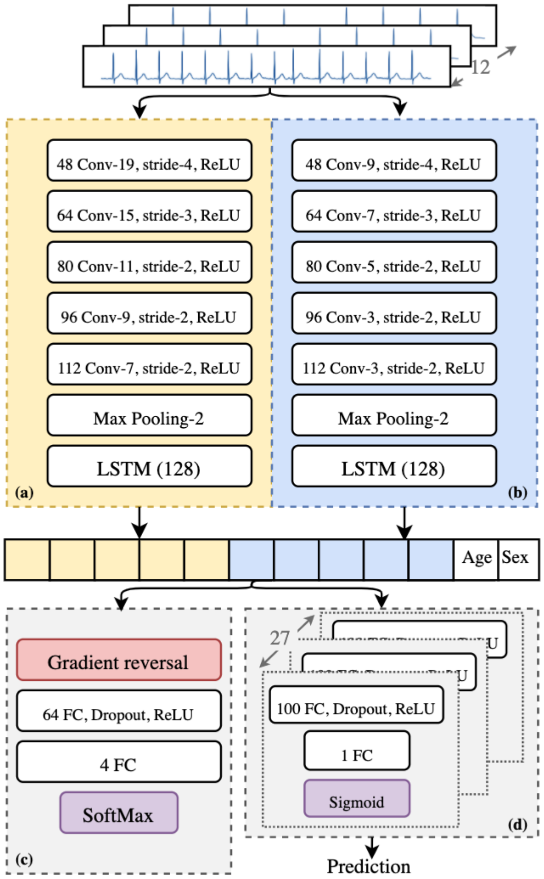
12
48 Conv-19, stride-4, ReLU
64 Conv-15, stride-3, ReLU
80 Conv-11, stride-2, ReLU
96 Conv-9, stride-2, ReLU
112 Conv-7, stride-2, ReLU
Max Pooling-2
LSTM (128)
(a)
48 Conv-9, stride-4, ReLU
64 Conv-7, stride-3, ReLU
80 Conv-5, stride-2, ReLU
96 Conv-3, stride-2, ReLU
112 Conv-3, stride-2, ReLU
Max Pooling-2
LSTM (128)
(b)
Age
Sex
Gradient reversal
64 FC, Dropout, ReLU
4 FC
SoftMax
(c)
27
100 FC, Dropout, ReLU
1 FC
Sigmoid
(d)
Prediction

Figure 1. The overall structure of the proposed model. Convolutional networks with large convolution kernels (a) and smaller convolution kernels (b) are used to extract coarse and fine features, respectively. Each convolutional network is followed by one max pooling and one LSTM layer. The extracted features are concatenated to the sex and age attributes of samples to produce the final feature vector. For domain generalization, a two-layer MLP network is used to classify domains from the final feature vector (c). Finally, the classification of arrhythmias is performed by passing the feature vector through 27 binary classifiers (d).

Since each recording may have more than one label, we use 27 separate binary classifiers to predict the presence of each arrhythmia, independently. After the last FC layer of each classifier, the sigmoid activation function has been used to compute the probability of the corresponding class. The probabilities are then binarized using class-specific thresholds. We used half of the validation data, to obtain the optimum thresholds via a grid search based on the maximization of the challenge score. By fixing the threshold values, the validation performance is then calculated based on another half of the validation dataset.

Each classifier is optimized by the binary cross-entropy loss function but the entire network, including the shared feature extractor parts, is optimized with respect to the weighted average of these cross-entropy losses. We used an importance factor for each class to determine the amount of contribution of its loss function in the optimization problem. These factors are not constant for all samples and the mutual weight between each ground truth label and other labels are defined based on the following weight matrix:

$$M_w = 1 - W_{reward} + \alpha * I \tag{1}$$

where, W_{reward} is the reward matrix from [1] and I is the 27×27 identity matrix and α is a positive hyperparameter. In the situation that one record has more than one label, we took the element-wise minimum of the corresponding rows from the weight matrix. The final classification loss function is defined as:

$$\begin{aligned} L_C(x_i, y_i; M_d, M_w) = -\sum_{c=1}^{27} M_d(i,c) M_w(i,c) \\ [y_i^c \log(G_y^c(z_i)) + (1 - y_i^c)\log(1 - G_y^c(z_i))], \end{aligned} \tag{2}$$

where $z_i = [G_f(x_i), sex, age]$, M_d is the domain mask, G_f is the parallel feature extractor, and G_y^c is the c^{th} binary classifier.

2.3. Adversarial Multi-Source Domain Generalization

In real-world scenarios, ECG signals may be recorded in different places through different devices with varying characteristics including different sampling frequencies, signal gains, or recording protocols. This results in an inhomogeneous dataset in which different partitions of the dataset carry different statistical properties. Furthermore, if the annotations of these partitions have been performed by different experts, this also induces inhomogeneity to the labels of the dataset that is also inevitable in the multi-source and large scale ECG datasets.

In most of the classification methods, it is usually assumed that instances from training and test datasets are sampled from the same distribution and these methods are not robust to statistical shifts. In the presence of test datasets, we could use unlabeled test data for domain adaptation techniques to make classification methods more robust to this domain shift. However, in many real-world applications, there is no access to the test data during the training of the models. Hence in this study, we try to increase the generalization of the models on different domains, without possessing samples of the test dataset.

We have used the domain-adversarial training technique [12] to increase generalization. To this end, a gradient reversal layer is added on the top of the feature vector then the output is passed to the domain classifier G_d as shown in Fig. 1(c). The following loss function is optimized to train this classifier:

$$L_D(x_i, d_i) = -\sum_{k=1}^{4} d_i^k \log(G_d(z_i)), \tag{3}$$

where d_i^k shows whether x_i belongs to domain k. Given i^{th} training smaple, the total loss function is defined as:

$$L^i = L_C^i + \lambda L_D^i. \tag{4}$$

As mentioned before, when data is collected from different sources, the distribution of labels is not similar across different domains. To reduce possible adverse side effects of annotation heterogeneity, we have used a domain-specific mask M_d during optimization which allows optimizing the network through available classes of each domain, and the algorithm does not punish out of domain misclassifications.

3. Results

Our method achieved a score of 0.609 and 0.437 on the official validation and test dataset, respectively. Using 5-fold cross-validation on the full training dataset, the proposed method achieved a score of $0.629_{\pm 0.003}$. In this section, we design additional experiments to better analyze the attributes of our proposed method.

In the first experiment, we build up four control models. In the first control model ($\mathbf{C_1}$), we did not perform data augmentations except random padding and cropping. In the second model ($\mathbf{C_2}$) we just excluded the additional random cropping from augmentation. In the third and fourth models ($\mathbf{C_3}$ and $\mathbf{C_4}$) we just used fine-grained and course-grained CNN for feature extraction, respectively. Table 1 shows the performance of these control models along with our proposed model on the hold-out validation of the training dataset.

To evaluate the effectiveness of adversarial domain generalization in our setting, we have conducted a new experiment. We have treated CPSC, CPSC-Extra, PTB-XL [1], and G12EC datasets [1] as different domains and then we assessed our model in a leave-one-out manner. Three domains have been used for training and the performance has been evaluated on the last domain and finally, 4 different results have averaged out. The model with adversarial domain generalization achieved a score of 0.352 while the score is 0.343 without domain generalization.

[1] To reduce probable effects concerning the dominance of this dataset, we only used $10k$ samples of this dataset in the present experiment.

Measure	**Ours**	$\mathbf{C_1}$	$\mathbf{C_2}$	$\mathbf{C_3}$	$\mathbf{C_4}$
AUROC	0.64	0.60	0.62	0.60	0.61
AUPRC	0.96	0.95	0.95	0.95	0.95
Accuracy	0.46	0.43	0.46	0.45	0.43
F1	0.57	0.55	0.55	0.55	0.56
F_β	0.61	0.60	0.59	0.56	0.60
G_β	0.37	0.34	0.35	0.36	0.37
Challenge score	0.64	0.62	0.62	0.62	0.62

Table 1. Performance of control models on the 10% hold-out validation from the training dataset. In $\mathbf{C_1}$, data augmentations except random padding and cropping are excluded. In $\mathbf{C_2}$, the additional random cropping is excluded. In $\mathbf{C_3}$ and $\mathbf{C_4}$, respectively, fine-grained and coarse-grained CNN are used for feature extraction.

4. Discussion and Conclusions

In this paper, we have proposed a method for the classification of 12-lead ECG signals. There are at least six types of challenges associated with the dataset that we tried to tackle in our proposed algorithm:

- Different size of data across different domains
- Heterogeneity of multi-source dataset and the possible difference between the train and undisclosed test data
- Heterogeneity of annotations between different domains (especially between the train and test data)
- Imbalance class distribution within each domain
- High variance in the distribution of signal lengths
- Different penalties for misclassification of different pairs of classes

We designed our method aiming to improve generalization on the unseen test dataset. Hence, some of its attributes may lose their significance when the test and train data come from the same distribution. In the Results section, we showed the superior performance of our data augmentation and domain generalization techniques against control models. However, the gap between our model and control models could become more significant when the test dataset comes from a different source.

Due to the time and resource limits concerning the cloud-training of models in the challenge, we did not include records with unscored labels for training. Utilizing these samples can enhance feature extraction and domain generalization and further improve the overall performance.

References

[1] Perez Alday EA, Gu A, Shah A, Robichaux C, Wong AKI, Liu C, Liu F, Rad BA, Elola A, Seyedi S, Li Q, Sharma A, Clifford GD, Reyna MA. Classification of 12-lead ECGs: the PhysioNet/Computing in Cardiology Challenge 2020. Physiological Measurement 2020;.

[2] Raj S, Ray KC. Sparse Representation of ECG Signals for Automated Recognition of Cardiac Arrhythmias. Expert Systems with Applications 2018;105:49–64.

[3] Yang W, Si Y, Wang D, Guo B. Automatic Recognition of Arrhythmia Based on Principal Component Analysis Network and Linear Support Vector Machine. Computers in Biology and Medicine 2018;101:22–32.

[4] Hannun AY, Rajpurkar P, Haghpanahi M, Tison GH, Bourn C, Turakhia MP, Ng AY. Cardiologist-level Arrhythmia Detection and Classification in Ambulatory Electrocardiograms Using a Deep Neural Network. Nature Medicine 2019;25(1):65.

[5] Warrick P, Homsi MN. Cardiac Arrhythmia Detection from ECG Combining Convolutional and Long Short-Term Memory Networks. In 2017 Computing in Cardiology (CinC). IEEE, 2017; 1–4.

[6] Zihlmann M, Perekrestenko D, Tschannen M. Convolutional Recurrent Neural Networks for Electrocardiogram Classification. In 2017 Computing in Cardiology (CinC). IEEE, 2017; 1–4.

[7] Ammour N. Atrial Fibrillation Detection with a Domain Adaptation Neural Network Approach. In 2018 International Conference on Computational Science and Computational Intelligence (CSCI). IEEE, 2018; 738–743.

[8] Jin Y, Qin C, Liu J, Lin K, Shi H, Huang Y, Liu C. A Novel Domain Adaptive Residual Network for Automatic Atrial Fibrillation Detection. Knowledge Based Systems 2020; 106122.

[9] Chen M, Wang G, Ding Z, Li J, Yang H. Unsupervised Domain Adaptation for ECG Arrhythmia Classification. In 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). IEEE, 2020; 304–307.

[10] Wang X, Zou Q. QRS Detection in ECG Signal Based on Residual Network. In 2019 IEEE 11th International Conference on Communication Software and Networks (ICCSN). IEEE, 2019; 73–77.

[11] Bitarafan A, Amini A, Baghshah MS, Khodajou-Chokami H. A Hybrid Deep Model for Automatic Arrhythmia Classification Based on LSTM Recurrent Networks. In 2020 IEEE International Symposium on Medical Measurements and Applications (MeMeA). IEEE, 2020; 1–6.

[12] Ganin Y, Ustinova E, Ajakan H, Germain P, Larochelle H, Laviolette F, Marchand M, Lempitsky V. Domain-Adversarial Training of Neural Networks. The Journal of Machine Learning Research 2016;17(1):2096–2030.

Address for correspondence:

Mahdieh Soleymani Baghshah
Department of Computer Engineering, Sharif University of Technology, Azadi Street, Tehran, Iran, P.O. Box: 11365-11155.
E-mail: soleymani@sharif.edu.