Towards Generalization of Cardiac Abnormality Classification Using ECG Signal

Xiaoyu Li[1], Chen Li[2], Xian Xu[4], Yuhua Wei[3], Jishang Wei[5], Yuyao Sun[4], Buyue Qian[6], Xiao Xu[4],

[1]School of Electronic and Information Engineering, Xi'an Jiaotong University, China
[2]SPKLSTN Lab, Department of Computer Science and Technology, Xi'an Jiaotong University, China
[3]National Engineering Lab of Big Data Analytics, Xi'an Jiaotong University, China
[4]Ping An Health Technology, Shanghai, China [5]HP Labs, Palo Alto, America
[6]The First Affiliated Hospital of Xi'an Jiaotong University, Xi'an, China

Abstract

In the PhysioNet/Computing in Cardiology Challenge 2021, our team, DrCubic, develops a novel approach to classify cardiac abnormalities using reduced-lead ECG recordings. In our approach, we incorporate peak detection as a self-supervised auxiliary task. We build the model based on SE-ResNet, and ensemble models of different input lengths and sampling rates. Inspired by last year's challenge results, we investigate various settings and techniques, and select the best ones, considering the intra-source performance and inter-source generalization simultaneously. Our classifiers receive scores of 0.666, 0.643, 0.642, 0.651, and 0.639 (ranked 3rd, 3rd, 4th, 4th, and 3th out of the teams) for the 12-lead, 6-lead, 4-lead, 3-lead, and 2-lead versions of the hidden validation set with the Challenge evaluation metric.

1. Introduction

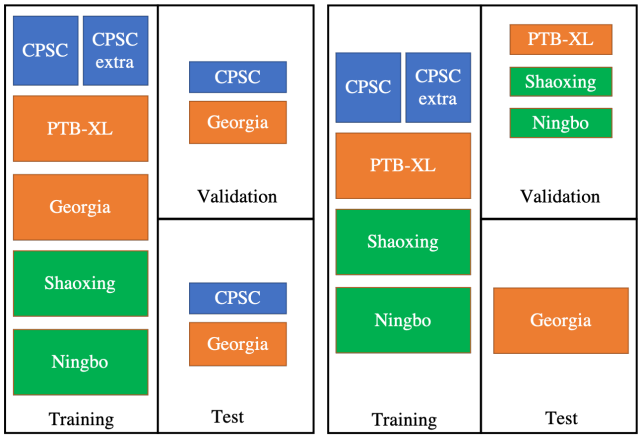
Electrocardiogram (ECG) is the most common non-invasive tool to screen and diagnose cardiac arrhythmias. However, such diagnosis is labor-intensive and requires years of training. With the advancing machine learning and deep learning techniques, computer-aid methods are promising to detect cardiac abnormalities. The PhysioNet/Computing in Cardiology Challenge 2021 provides a platform to develop automatic models for classifying cardiac abnormalities from reduced-lead ECG recordings [1, 2].

We start Challenge 2021 by analysing the Challenge 2020. Last year, most of the top teams utilize deep convolutional neural network (CNN) and attention mechanism (Transformer or Squeeze-And-Excitation) [3–6], which means CNN is the winner and attention mechanism counts. From the results on hidden test sets of Challenge 2020, we observe that models perform well on the data source which also appears in the training sets, but generalize poorly on the unseen hidden undisclosed dataset. To better understand and compare the performance of the models, we calculate the challenge score of random output on all training sets and on the Georgia training set as two weak baselines. The two baselines achieve the scores as 0.2832 and 0.3263 respectively. What surprizes us is that many models can not compete against the two random baselines, especially on the unseen dataset. We also observe that the final scores on the whole test set are basically aligned with the scores on the hidden undisclosed set, which means the performance on the unseen data source is vital and generalization is the key. Consequently, when investigating various settings and techniques, we consider intra-source performance and inter-source generalization simultaneously to and select the best ones.

Challenge 2021 provides multi-source datasets from different countries [7], INCART[8], PTB (-XL) [9,10], Chapman[11] and Ningbo[12]. To compare various techniques and settings in terms of intra-source performance and inter-source generalization using these datasets, we design the two data split settings, as shown in Fig.1. The data split in Fig.1.a is for intra-source performance. CPSC and Georgia datasets are randomly split into training set, validation set and test set, which matches the online validation set. The data split in Fig.1.b is for inter-source generalization. The Georgia dataset is split solely as test set.

Under the two data split settings, we investigate different techniques, including domain alignment methods, different network architectures, domain knowledge aid, semi-supervised learning and ensemble learning. Our solution mainly consists of SE-ResNet, peak detection as a self-supervised auxiliary task, and ensemble learning.

In this section, we present our final solution for this years' challenge.



(a) Data split for intra-source performance

(b) Data split for inter-source generalization

Figure 1: Two data split settings for intra-source performance and inter-source generalization respectively. (a). CPSC and Georgia datasets are randomly split into training set, validation set and test set. (b). Georgia dataset is solely split as the test set.

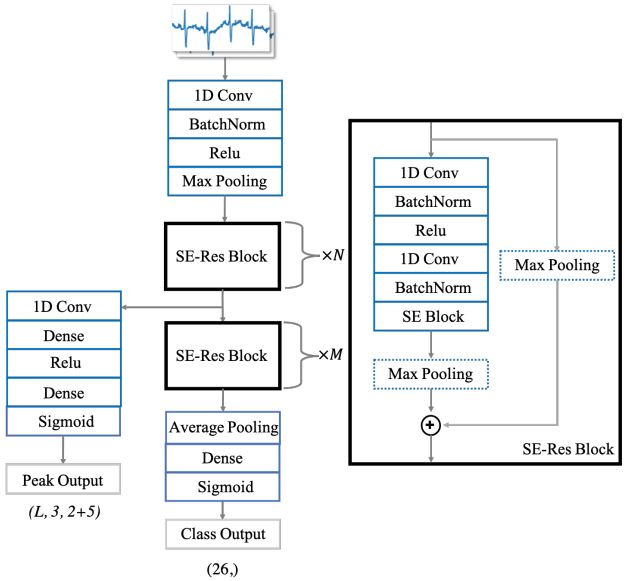


Figure 2: Network structure

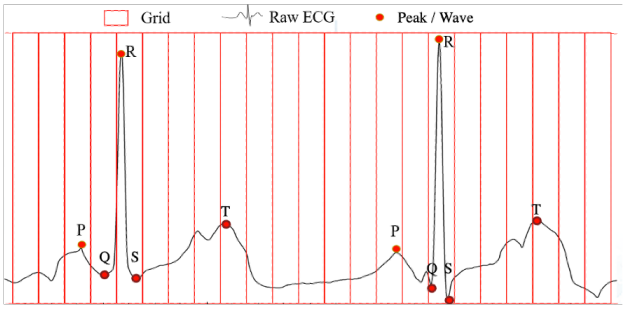


Figure 3: One recording as multiple grids.

1.1. Preprocessing

We remove INCART and PTB datasets. All recordings are resampled to 500 Hz. We randomly cut or zero-pad the recordings to length of 4992 (about 10 seconds). We also apply wavelet transformer to filter noise.

1.2. Abnormality classification task

The backbone of our SE-ResNet is the same as [6], and the whole network structure is as Fig.2. We add a branch on the middle of the backbone, which outputs the results for peak detection. There is a hyperparameter array $[t_1, t_2, t_3, t_4]$. For each t_i, i denotes current stage number, t_i denotes the number of the SE-Res Block in the stage. There are 4 stages. At the first SE-Res Block in each stage, a Max-Pooling layer with a stride of 2 is applied to downsample the recording. To deal with different leads of input, we keep all other settings static and only change this array, $[3, 4, 6, 3]$, $[3, 4, 4, 3]$, $[3, 4, 2, 3]$, $[2, 3, 4, 2]$ and $[2, 3, 2, 2]$ respectively for 12-lead, 6-lead, 4-lead, 3-lead and 2-lead recordings.

For the abnormality classification task, the main difference between our solution and [6] is that we utilize an asymmetric loss (ASL)[13] for the multi-label classification problem. The ASL loss is strong multi-label version of focal loss. There are 26 labels for each recording, but usually only a small number of the labels are positive.

$$L_{classification} = \frac{1}{K}\sum_{k=0}^{K} ASL(p_k, y_k) \tag{1}$$

We utilize ASL to alleviate such label imbalance problem. Three hyperparameters $(\lambda_+, \lambda_-, m)$ need to be set. We set λ_+ as 1, λ_- as 4 and m as 0.05 to reduce the contribution of negative labels. For details of ASL loss, please refer to [13].

1.3. Peak detection as an auxiliary task

We extend [14] to detect all 5 kinds of peaks. We add a branch before the last stage of SE-ResNet. The branch consists of a multi-layer-perception and a convolutional layer with a filter size of $3 * (2 + 5)$. The output of the branch is then reshaped to $(L, 3, 2 + 5)$. L is the length of the feature map before the last stage. The number L means

that, we treat the recording as L grids and detect peaks in each grid, shown as Fig.[14]. The number 3 means we detect 3 peaks in a grid at most. The $2+5$ means, for each grid, we predict peak detection confidence C, peak relative position x and 5 peak classes (PQRST). The ground truth of the peaks is calculated by [15]. The loss of peak detection is as the formula (2), where $\boldsymbol{1}$ denotes whether to calculate the corresponding loss item when there is a peak inside the grid or not, λ_{coord} and λ_{noobj} are set as 5 and 0.2 to balance positive grids and negative grids.

$$\begin{aligned} L_{detection} = \lambda_{coord} & \sum_{i=0}^{L} \boldsymbol{1}_i^{obj}(x_i - \hat{x}_i)^2 \\ & + \sum_{i=0}^{L} \boldsymbol{1}_i^{obj}(C_i - \hat{C}_i)^2 \\ + \lambda_{noobj} & \sum_{i=0}^{L} \boldsymbol{1}_i^{noobj}(C_i - \hat{C}_i)^2 \\ & + \sum_{i=0}^{L} \boldsymbol{1}_i^{obj} \sum_{c \in classes} (p_i(c) - \hat{p}_i(c))^2 \end{aligned} \tag{2}$$

Then we combine the classification loss and peak detection loss with a hyperparameter α. We set the $alpha$ as 0.1. During training, we pretrain the SE-ResNet only with the peak detection task first. Then we combine both tasks to finetune the network.

$$L_{final} = L_{classification} + \alpha * L_{detection} \tag{3}$$

1.4. Ensemble

We also ensemble different models for final classification. An ideal generalizable classifier learns the features which truly capture the patterns for 26 classes instead of the features dependent on data source. Considering there are only 6 classes for CPSC dataset, an ideal classifier should generalize poorly on CPSC data. This, a source classifier with the same backbone as in SE-ResNet is trained to predict whether the recording is from CPSC datasets. If true, the recording is sent to the classifier which only output 6 classes for CPSC. If not, we utilize the classifier for generalization. The pipeline is shown as in Fig.4. Each of the classifier for CPSC and genralization is ensembled with three SE-ResNet models, *SE-ResNet*$_{500Hz\_10s}$, *SE-ResNet*$_{250Hz\_10s}$, *SE-ResNet*$_{500Hz\_15s}$. These SE-ResNet models are trained with different input lengths and sampling rates.

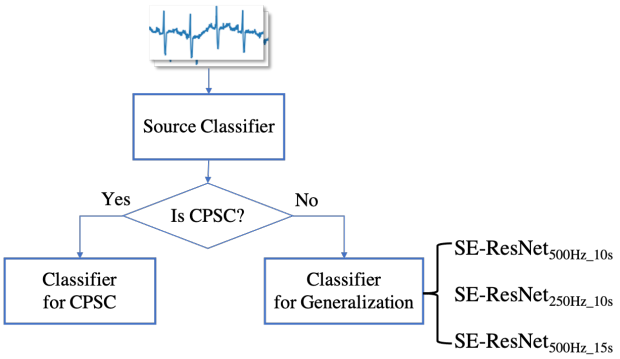


Figure 4: Ensemble structure

1.5. Training details and post-processing

For training settings, we set batch size as 128, optimizer as *AdamW* with a learning rate of 0.0002. We also linearly warm up the learning rate for the first 3 epochs, and then adopt a cosine learning schedule. We add early stopping with a patience of 4 during both pretraining and finetuning.

As for post-processing, if all of 26 labels are predicted negative for a recording, we label the *NSR* class as positive. If the label of *TInv* is positive, we also label the *TAb* class as positive.

2. Results

The test scores on split Georgia for inter-source generalization are shown in Table.1. Although GroupDRO and DANN are specially designed for the domain generalization problem, both methods show no significant improvement in our experiments. For peak detection as an auxiliary task, we observe more than 1% improvement.

The test scores on CPSC and Georgia for final intra-source performance are shown in Table.2. Model$_1$ consists of only *SE-ResNet*$_{500hz\_10s}$. Model$_2$ consists of *SE-ResNet*$_{500hz\_10s}$ and *SE-ResNet*$_{500hz\_15s}$. Model$_3$ consists of *SE-ResNet*$_{500hz\_10s}$, *SE-ResNet*$_{500hz\_15s}$ and *SE-ResNet*$_{250hz\_10s}$. We observe consistent improvement on the Georgia dataset when we ensemble more models.

The online validation results are shown in Table.3.

3. Discussion and Conclusions

We investigate various techniques and settings for intra-source performance and inter-source generalization under two different data split settings. With experiments, we verify the effectiveness of the main components in our solution: SE-ResNet, peak detection as an auxiliary task and our ensemble strategy. Domain knowledge and ensemble learning are helpful for training superior cardiac abnormality classifiers.

Models	SE-ResNet	GroupDRO	DANN	FixMatch	PeakDetection
Georgia	0.518	0.519	0.510	0.522	0.534

Table 1: The results for inter-source generalization.

Models	CPSC+Georgia	CPSC	Georgia
$Model_1$	0.725	0.876	0.677
$Model_2$	0.731	0.875	0.684
$Model_3$	0.742	0.872	0.705

Table 2: The results for intra-source performance.

Leads	12	6	4	3	2
Online Validation	0.666	0.643	0.642	0.651	0.639
Ranking	3rd	3rd	4th	4th	3rd

Table 3: The results of online validation.

The power of the peak detection auxiliary task is still limited by the quality of peak labels. How to detect theses peaks is a critical problem in ECG representation learning. However, traditional methods usually fail to detect peaks for extremely noisy recording, and there are no enough clean peak labels to train a robust peak detector based on deep neural network. We believe it is worthwhile to label peaks on large-scale datasets.

Acknowledgments

This work has been supported by National Natural Science Foundation of China (61772409); The consulting research project of the Chinese Academy of Engineering (The Online and Offline Mixed Educational Service System for "The Belt and Road" Training in MOOC China); Project of China Knowledge Centre for Engineering Science and Technology; The innovation team from the Ministry of Education (IRT_17R86); and the Innovative Research Group of the National Natural Science Foundation of China (61721002).

References

[1] Perez Alday EA, Gu A, Shah A, Robichaux C, Wong AKI, Liu C, et al. Classification of 12-lead ECGs: the PhysioNet/Computing in Cardiology Challenge 2020. Physiological Measurement 2020;41.

[2] Reyna MA, Sadr N, Perez Alday EA, Gu A, Shah A, Robichaux C, et al. Will Two Do? Varying Dimensions in Electrocardiography: the PhysioNet/Computing in Cardiology Challenge 2021. Computing in Cardiology 2021;48:1–4.

[3] Natarajan A, Chang Y, Mariani S, Rahman A, Boverman G, Vij S, et al. A wide and deep transformer neural network for 12-lead ecg classification. In 2020 Computing in Cardiology. IEEE, 2020; 1–4.

[4] Zhao Z, Fang H, Relton SD, Yan R, Liu Y, Li Z, et al. Adaptive lead weighted resnet trained with different duration signals for classifying 12-lead ecgs. In 2020 Computing in Cardiology. IEEE, 2020; 1–4.

[5] Zhu Z, Wang H, Zhao T, Guo Y, Xu Z, Liu Z, et al. Classification of cardiac abnormalities from ecg signals using se-resnet. In 2020 Computing in Cardiology. IEEE, 2020; 1–4.

[6] Jia W, Xu X, Xu X, Sun Y, Liu X. Automatic detection and classification of 12-lead ecgs using a deep neural network. In 2020 Computing in Cardiology. IEEE, 2020; 1–4.

[7] Liu F, Liu C, Zhao L, Zhang X, Wu X, Xu X, et al. An Open Access Database for Evaluating the Algorithms of Electrocardiogram Rhythm and Morphology Abnormality Detection. Journal of Medical Imaging and Health Informatics 2018;8(7):1368—1373.

[8] Tihonenko V, Khaustov A, Ivanov S, Rivin A, Yakushenko E. St Petersburg INCART 12-lead Arrhythmia Database. PhysioBank PhysioToolkit and PhysioNet 2008;.

[9] Bousseljot R, Kreiseler D, Schnabel A. Nutzung der EKG-Signaldatenbank CARDIODAT der PTB über das Internet. Biomedizinische Technik 1995;40(S1):317–318.

[10] Wagner P, Strodthoff N, Bousseljot RD, Kreiseler D, Lunze FI, Samek W, et al. PTB-XL, a Large Publicly Available Electrocardiography Dataset. Scientific Data 2020;7(1):1–15.

[11] Zheng J, Zhang J, Danioko S, Yao H, Guo H, Rakovski C. A 12-lead Electrocardiogram Database for Arrhythmia Research Covering More Than 10,000 Patients. Scientific Data 2020;7(48):1–8.

[12] Zheng J, Cui H, Struppa D, Zhang J, Yacoub SM, El-Askary H, et al. Optimal Multi-Stage Arrhythmia Classification Approach. Scientific Data 2020;10(2898):1–17.

[13] Ben-Baruch E, Ridnik T, Zamir N, Noy A, Friedman I, Protter M, et al. Asymmetric loss for multi-label classification, 2020.

[14] Li X, Qian B, Wei J, Zhang X, Chen S, Zheng Q, et al. Domain knowledge guided deep atrial fibrillation classification and its visual interpretation. In Proceedings of the 28th ACM International Conference on Information and Knowledge Management. 2019; 129–138.

[15] Makowski D, Pham T, Lau ZJ, Brammer JC, Lespinasse F, Pham H, et al. NeuroKit2: A python toolbox for neurophysiological signal processing. Behavior Research Methods feb 2021;53(4):1689–1696.

Address for correspondence:

Xiaoyu Li

No.28, Xian Ning West Road, Xi'an, Shaanxi Province, China xiaoyuli@stu.xjtu.edu.cn