Towards High Generalization Performance on Electrocardiogram Classification

Hyeongrok Han†, Seongjae Park‡, Seonwoo Min¹,³, Hyun-Soo Choi⁴, Eunji Kim¹, Hyunki Kim¹, Sangha Park¹, Jinkook Kim², Junsang Park², Junho An², Kwanglo Lee², Wonsun Jeong⁷, Sangil Chon², Kwonwoo Ha², Myungkyu Han², Sungroh Yoon*¹,⁵

¹Department of Electrical and Computer engineering, Seoul National University, Seoul, South Korea
²HUINNO Co., Ltd., Seoul, South Korea
³LG AI Research, Seoul, South Korea
⁴Department of Computer Science and Engineering, Kangwon National University, Chuncheon, South Korea
⁵Department of Biological Sciences, Interdisciplinary Program in Bioinformatics, Interdisciplinary Program in Artificial Intelligence, ASRI, INMC, and Institute of Engineering Research, Seoul National University, Seoul, South Korea

†: equal contribution (Hyeongrok Han and Seongjae Park)
*: corresponding author (Sungroh Yoon)

Abstract

Recently, many electrocardiogram (ECG) classification algorithms using deep learning have been proposed. The characteristics of ECG vary from dataset to dataset for various reasons (i.e., hospital, race, etc). Therefore, it is important that models have high dataset-wise generalization performance. In this paper, as part of the PhysioNet / Computing in Cardiology Challenge 2021, we developed a model to classify cardiac abnormalities from 12 lead and reduced-lead ECGs. In particular, to select a model with high generalization performance, we applied constant-weighted cross-entropy loss, and evaluated the performance using a leave-one-dataset-out cross-validation setting. Our DSAIL-SNU team got challenge scores of 0.61, 0.58, 0.60, 0.59, and 0.59 on 12, 6, 4, 3, 2-lead ECGs respectively. Our model obtained higher dataset-wise generalization performance than the model we submitted last year.

1. Introduction

Electrocardiogram (ECG) is an important tool for diagnosing cardiac abnormalities, and more than 300 million ECGs are obtained worldwide each year [1]. Standard ECGs, which are used to diagnose heart diseases, consist of 12 leads. However, it is not always possible to obtain all 12 leads due to the cost and limitations of measurement devices. Recently, it has been demonstrated that a subset of 12 leads also contains sufficiently meaningful information [2]. According to the rapid growth of deep learning, there have been proposed ECG classification methods based on deep neural networks (DNNs). These approaches can automatically learn feature representations and have shown superior performance to traditional methods using handcrafted features [3, 4]. The characteristics of ECG vary from dataset to dataset for various reasons, i.e., hospital, race, etc. It is important to consider that model has generalization performance on dataset which was not seen during training. Therefore, it is necessary to check whether the proposed model shows high dataset-wise performance.

In this paper, as part of the PhysioNet / Computing in Cardiology Challenge 2021, we developed a model to classify cardiac abnormalities from 12 lead and reduced-lead ECGs [5–7]. In order for the model to have high dataset-wise generalization performance, we applied various methods. In addition, we evaluated the model using the leave-one-dataset-out cross-validation for model selection. Our proposed model achieved a 0.1 higher dataset-wise challenge score than the model we submitted last year [4].

2. Methods

2.1. Data

Table[1] shows the statistics of the data provided by the challenge with 26 scored SNOMED-CT labels [14] from 8 datasets [7]. Among them, PTB and INCART data are not used for training because of the long lengths and relatively small number of samples. We also do not use those without any positive scored labels for training. When training the
model, the ratio of train and validation datasets is 9:1. In
the leave-one-dataset-out cross-validation setting, one of
the six datasets is used as the test dataset, and the remain-
ing five datasets are used to train and validation datasets.

We apply the following data pre-processing procedures.
First, we upsample or downsample ECGs into 500Hz.
Then, we apply a Finite Impulse Response (FIR) band-
pass filter with a bandwidth of 3 to 45Hz. Normalization is
applied using the minimum and maximum values of each
sample. Finally, for any recording with a data length longer
than 7,500, we randomly use a segment with a length of
7,500 as input. If the length is shorter than 7,500, we use
zero-padding to 7,500. For reduced-lead model training,
pre-defined leads are extracted from the 12-lead sample
[7].

2.2. Model Architecture

For the baseline model, we use our previous work [4].
We use the WRN model architecture with 14 convolu-
tion/dense layers and widening factor 1 [15]. The overall
structure of the model is shown in Figure 1. The additional
parts from the baseline are depicted in purple. The baseline
model consists of the basic residual block, but we use the
Squeeze and Excitation (SE) block to let the model learn
interdependency between channels [16]. For the model to
consider the demographic information, we add the addi-
tional features to the dense layer of the output stem.

2.3. Training

First, we describe the experiment settings. Each model
is trained for 100 epochs using Pytorch with an NVIDIA
GeForce RTX 3080 [17]. We use Adam optimizer, L2
weight decay of 0.0005, a dropout rate of 0.3, a batch
size of 128, and a learning rate of 0.001 through hyper-
parameter search. In the next part, we explain the training
refinements to improve dataset-wise generalization.

Constant-weighted binary cross-entropy loss

In last year, we proposed confusion-weighted binary-
cross-entropy (CoW-BCE) [4] loss designed to resemble
an evaluation metric called challenge score [7]. Although
the model trained via CoW-BCE loss showed a high chal-
lenge score on the validation dataset, it showed a much
lower score on the hidden test dataset.

In this work, we use constant-weighted binary-cross
entropy inspired via asymmetric loss (ASL) [18]. The
ASL uses asymmetric focusing and asymmetric probabil-
ity shifting to overcome the inherent positive-negative im-
balance in typical multi-label classification problems as
follows:

\[
    ASL = \begin{cases} 
        -(1 - p)^\gamma^+ \log (p), & \text{if } y \text{ is } 1 \\
        -(p_m)^\gamma^- \log (1 - p_m), & \text{otherwise} 
    \end{cases}
\]  

(1)

where \( p \) is the output probability of the model, \( p_m \) is the
shifted probability, and \( \gamma^+, \gamma^- \) are positive and negative
focusing parameters, respectively.

For ease of implementation, we assume the positive fo-
cusing parameter \( \gamma^+ \) to be 0. We investigate the constant
value of the negative coefficient, which depends on the op-
timal negative focusing parameters \( \gamma^- \) and shifted proba-
bility \( p_m \). Experimentally, we set the negative coefficient
to be 0.1, which is approximately the ratio of positive to
negative classes in the whole dataset.

Demographic features

For the model to learn demographic information, we ad-
Additionally use two kinds of features, i.e., age and sex. The demographic feature vector consists of 5 values for age, one-hot encoded sex, and two flags for missing values. If there are age and gender values in the header, the values are used directly, and missing flags are set to 0. Otherwise, pre-defined default values are used, and the missing flags are set to 1. The default age value is 60.37, and the default sex value (female/male ratio) is 0.471/0.519. As shown in the purple path in Figure 1, the feature vector is concatenated with the feature extracted by DNN before the last dense layer.

Mixup

Mixup is one of the data augmentation techniques for better generalization [19]. It makes the decision boundary smoother by regularizing the model. Assuming that two arbitrary input signals in the batch are \(x_1, x_2\), the features of the samples are \(f_1, f_2\), and the labels are \(l_1, l_2\), the mixup samples \(x', f', l'\) are created as follows.

\[
x' = \lambda x_1 + (1 - \lambda) x_2 \tag{2}
\]

\[
f' = \lambda f_1 + (1 - \lambda) f_2 \tag{3}
\]

\[
l' = \lambda l_1 + (1 - \lambda) l_2 \tag{4}
\]

As used in the original mixup paper, mixing coefficient \(\lambda\) is sampled from a \(Beta(0.2, 0.2)\) distribution. The model is trained using the generated \(x', f',\) and \(l'\).

Learning rate scheduler

We use the OneCycle learning rate scheduler [20]. It is known as a method for effective training by “super-convergence” of residual blocks. At the beginning of training, the learning rate is set to a small value, and it is gradually increased and then decreased again after reaching the pre-defined maximum value. The learning rate values per epoch are shown in Figure 2. The maximum learning rate value is set to 0.001, and the model is trained for a total of 100 epochs using a cosine annealing strategy.

<table>
<thead>
<tr>
<th>Leads</th>
<th>Training</th>
<th>Validation</th>
<th>Team ranking</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>0.654</td>
<td>0.610</td>
<td>15th</td>
</tr>
<tr>
<td>6</td>
<td>0.680</td>
<td>0.580</td>
<td>16th</td>
</tr>
<tr>
<td>4</td>
<td>0.691</td>
<td>0.600</td>
<td>15th</td>
</tr>
<tr>
<td>3</td>
<td>0.689</td>
<td>0.590</td>
<td>16th</td>
</tr>
<tr>
<td>2</td>
<td>0.673</td>
<td>0.590</td>
<td>16th</td>
</tr>
</tbody>
</table>

Table 2. Challenge scores for our model using whole six datasets.

<table>
<thead>
<tr>
<th>Model</th>
<th>Challenge score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.732</td>
</tr>
<tr>
<td>Our model</td>
<td>0.654</td>
</tr>
</tbody>
</table>

Table 3. Challenge score of the baseline and our model using whole six datasets.

3. Experiments results

The experiment results of the model trained using the whole six datasets are shown in Table 2. We report the training and validation challenge score, and team ranking for our proposed 12, 6, 4, 3, and 2-lead models. The average validation challenge score was 0.594. In Table 3, we compare the validation challenge scores of the baseline and our 12-lead model. The challenge score obtained by our model is 0.654, which is 0.08 lower than the baseline.

Table 4 shows the results of the 12-lead models from the leave-one-dataset-out cross-validation setting. This is an important setting to check the dataset-wise generalization performance. We report the challenge scores when the dataset in the first row is used as a test dataset. Our proposed model show higher dataset-wise generalization performance than the baseline, and the average challenge score is 0.483. Although our model obtain a lower challenge score when trained using the whole six datasets, the dataset-wise generalization performance is better compared to the baseline. The usage of a constant-weighted binary cross-entropy loss instead of CoW-BCE loss function makes the most of the improvement in the dataset-wise generalization performance. In particular, the changed loss function improve the generalization performance for the PTB-XL dataset.

4. Concluding Remarks

In this paper, as a participating team in the PhysioNet Challenge 2021, we proposed 12 and reduced-lead models for automatically classifying cardiac abnormalities from ECGs. We focused on building the classification model that high dataset-wise generalization performance. We used the SE WRN-14-1 network, constant binary cross-entropy loss, feature extraction, mixup, and OneCycle learning rate scheduler.
For better model selection, we compared the challenge score with the leave-one-dataset-out cross-validation setting in Table 4. The average challenge score of our proposed model was 0.483, confirming that the dataset-wise generalization performance was higher than that of the baseline which was 0.384.

Acknowledgments

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government (Ministry of Science and ICT, 2018R1A2B3001628), the BK21 FOUR program of the Education and Research Program for Future ICT Pioneers, Seoul National University in 2021, and Samsung Electronics(DS and Foundry).

References


Address for correspondence:
Sungroh Yoon
Rm. 908, Bldg. 301, 1 Gwanak-ro, Gwanak-gu, Seoul 08826, South Korea
sryoon@snu.ac.kr