Towards High Generalization Performance on Electrocardiogram Classification

Hyeongrok Han†[1], Seongjae Park†[2], Seonwoo Min[1,3], Hyun-Soo Choi[4], Eunji Kim[1], Hyunki Kim[1], Sangha Park[1], Jinkook Kim[2], Junsang Park[2], Junho An[2], Kwanglo Lee[2], Wonsun Jeong[2], Sangil Chon[2], Kwonwoo Ha[2], Myungkyu Han[2], Sungroh Yoon*[1,5]

[1]Department of Electrical and Computer engineering, Seoul National University, Seoul, South Korea
[2]HUINNO Co., Ltd,. Seoul, South Korea
[3]LG AI Research, Seoul, South Korea
[4]Department of Computer Science and Engineering, Kangwon National University, Chuncheon, South Korea
[5]Department 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

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 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].

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

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

Dataset	Number of recordings	w/ Scored Labels	Average Length (second)
Ningbo[8]	34,905	34,485	10
PTB-XL[9]	21,837	21,604	10
Chapman[10]	10,247	9,710	10
G12EC[6]	10,344	9,458	9
CPSC[11]	6,877	5,279	15
CPSC-Extra[11]	3,453	1,278	15
PTB[12]	516	97	110
INCART[13]	74	33	1,800

Table 1. Data statistics

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 remaining 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) bandpass 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 convolution/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 additional 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 hyperparameter 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

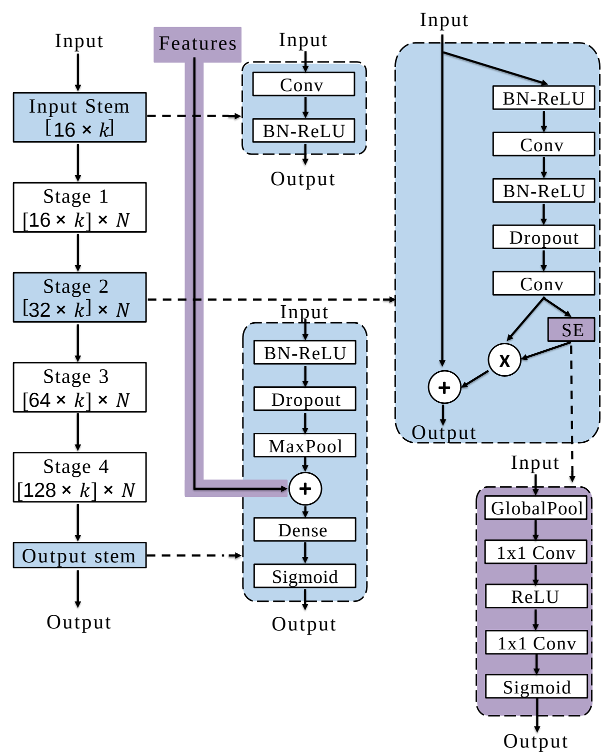

Figure 1. Model overview.

an evaluation metric called challenge score [7]. Although the model trained via CoW-BCE loss showed a high challenge 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 probability shifting to overcome the inherent positive-negative imbalance in typical multi-label classification problems as follows:

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

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

For ease of implementation, we assume the positive focusing parameter γ^+ to be 0. We investigate the constant value of the negative coefficient, which depends on the optimal negative focusing parameters γ^- and shifted probability 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-

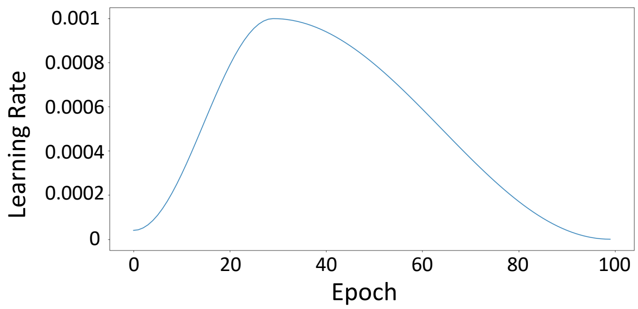

Figure 2. OneCycle Learning rate Scheduler.

ditionally 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 λ 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.

Leads	Training	Validation	Team ranking
12	0.654	0.610	15th
6	0.680	0.580	16th
4	0.691	0.600	15th
3	0.689	0.590	16th
2	0.673	0.590	16th

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

Model	Challenge score
Baseline	0.732
Our model	0.654

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.

Model	Ningbo	PTB-XL	Chapman	G12EC	CPSC	CPSC-Extra	Avg
Baseline	0.545	-0.101	0.659	0.428	0.463	0.310	0.384
Our model	**0.626**	**0.200**	**0.723**	**0.519**	**0.506**	**0.424**	**0.483**

Table 4. Results of leave-one-dataset-out cross-validation setting (12-lead model)

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

[1] Holst H, Ohlsson M, Peterson C, Edenbrandt L. A confident decision support system for interpreting electrocardiograms. Clinical Physiology 1999;19(5):410–418.

[2] Drew BJ, Pelter MM, Brodnick DE, Yadav AV, Dempel D, Adams MG. Comparison of a new reduced lead set ecg with the standard ecg for diagnosing cardiac arrhythmias and myocardial ischemia. Journal of electrocardiology 2002; 35(4):13–21.

[3] Zubair M, Kim J, Yoon C. An automated ecg beat classification system using convolutional neural networks. In 2016 6th international conference on IT convergence and security (ICITCS). IEEE, 2016; 1–5.

[4] Min S, Choi HS, Han H, Seo M, Kim JK, Park J, et al. Bag of tricks for electrocardiogram classification with deep neural networks. In 2020 Computing in Cardiology. IEEE, 2020; 1–4.

[5] Goldberger AL, Amaral LA, Glass L, Hausdorff JM, Ivanov PC, Mark RG, et al. PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals. Circulation 2000;101(23):e215–e220.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

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

[13] Tihonenko V, Khaustov A, Ivanov S, Rivin A, Yakushenko E. St Petersburg INCART 12-lead Arrhythmia Database. PhysioBank PhysioToolkit and PhysioNet 2008;Doi: `10.13026/C2V88N`.

[14] Shahpori R, Doig C. Systematized nomenclature of medicine–clinical terms direction and its implications on critical care. Journal of critical care 2010;25(2):364–e1.

[15] Zagoruyko S, Komodakis N. Wide residual networks. arXiv preprint arXiv160507146 2016;.

[16] Hu J, Shen L, Sun G. Squeeze-and-excitation networks. In Proceedings of the IEEE conference on computer vision and pattern recognition. 2018; 7132–7141.

[17] Paszke A, Gross S, Chintala S, et al. Automatic differentiation in PyTorch. NIPS Autodiff Workshop 2017;.

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

[19] Zhang H, Cisse M, Dauphin YN, Lopez-Paz D. mixup: Beyond empirical risk minimization. arXiv preprint arXiv171009412 2017;.

[20] Smith LN, Topin N. Super-convergence: Very fast training of neural networks using large learning rates. In Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications, volume 11006. International Society for Optics and Photonics, 2019; 1100612.

Address for correspondence:

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