Learning ECG Representations for Multi-Label Classification of Cardiac Abnormalities

Jangwon Suh[1], Jimyeong Kim[1], Eunjung Lee[1], Jaeill Kim[1], Duhun Hwang[1], Jungwon Park[2], Junghoon Lee[3], Jaeseung Park[3], Seo-Yoon Moon[3], Yeonsu Kim[3], Min Kang[3], Soonil Kwon[4], Eue-Keun Choi[4,5], Wonjong Rhee[1]

[1]Department of Intelligence and Information, Seoul National University, Korea
[2]Department of Mathematical Sciences, Seoul National University, Korea
[3]College of Liberal Studies, Seoul National University, Korea
[4]Division of Cardiology, Department of Internal Medicine, Seoul National University Hospital, Korea
[5]Department of Internal Medicine, Seoul National University College of Medicine, Korea

Abstract

The goal of PhysioNet/Computing in Cardiology Challenge 2021 was to identify clinical diagnoses from 12-lead and reduced-lead ECG recordings, including 6-lead, 4-lead, 3-lead, and 2-lead recordings. Our team, snu_adsl, have used EfficientNet-B3 as the base deep learning model and have investigated methods including data augmentation, self-supervised learning as pre-training, label masking that deals with multiple data sources, threshold optimization, and feature extraction. Self-supervised learning showed promising results when the size of labeled dataset was limited, but the competition's dataset turned out to be large enough that the actual gain was marginal. In consequence, we did not include self-supervised pre-training in our final entry. Our classifiers received scores of 0.48, 0.48, 0.47, 0.47, and 0.45 (ranked 12th, 10th, 11th, 11th, and 13th out of 39 teams) for the 12-lead, 6-lead, 4-lead, 3-lead, and 2-lead versions of the hidden test set with the Challenge evaluation metric.

1. Introduction

The electrocardiogram (ECG) is an essential tool for diagnoses of cardiovascular diseases and it is becoming increasingly important as more personal ECG devices become affordable and widely available. The PhysioNet/Computing in Cardiology Challenge 2021 focused on automated, open-source approaches for classifying cardiac abnormalities from ECG signals with fewer leads [1–3].

In the deep learning research community, an amazing progress has been made in the last few years where unlabeled datasets of image and text were utilized to train a new generation of models. The most well-known example is GPT-3 that can produce smooth writings that are indistinguishable from human's writings. The subsequent example is DALL·E. For all these models, self-supervised learning plays the key role for learning representations. Self-supervised learning is an unsupervised learning method where unlabeled datasets are used for training. We have tried applying self-supervised techniques (e.g. [4]) specifically for learning representations of ECG signals. We used the competition datasets in an unsupervised manner for training models first, and then fine-tuned the models in a supervised manner. While promising results were obtained in the research phase, the actual application turned out to be marginally helpful for the challenge because of the sufficiently large size of the labeled dataset and the computational limitation.

2. Methods

While our main research goal was to focus on the application of self-supervised learning to learn effective ECG representations, we have also investigated several other aspects for enhancing the performance and the main techniques are described in this section. We have decided to develop a single model based on 6-lead because the performance was not significantly dependent on the number of leads. The comparable performance indicates that the extra information within 12-lead signal is limited for the multi-label classification tasks.

2.1. Dataset

The challenge dataset for the PhysioNet/CinC challenge 2021 consists of 131,155 12-lead ECG recordings of different lengths and frequencies, labeled with one or more of distinct 133 classes [2,3]. Only 30 of the classes were considered in the challenge evaluation, where some of them

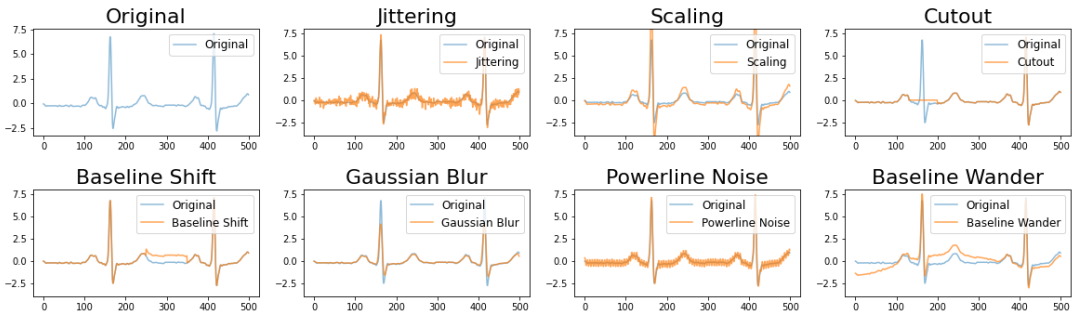

Figure 1. Data augmentation schemes.

were grouped into single classes.

2.2. Pre-processing

All ECG signals were resampled to 300Hz. We applied a Butterworth bandpass filter with 1Hz-45Hz frequency. We also applied standardization to each recording. The standardization did not necessarily improve the performance, but we have kept it on in case the unseen dataset has unexpected characteristics. To handle ECG signals with different lengths, we selected a random window with a width of 4,000 data points which corresponds to 13.3 seconds. ECG signals shorter than 13.3 seconds were zero-padded at the end.

2.3. Feature extraction

We adopted ten features for the supervised learning - age, sex, mean and standard deviation of RR interval, RMSSD (root mean square of successive difference) of RR interval, mean of R-peak value, RMSSD of R-peak value, mean, minimum, and maximum of heart rate. Age was scaled down by 100 and the missing values were replaced with the average. Sex was one-hot encoded, and missing values were handled by a dummy variable (one when missing, zero otherwise). R-peak related features were computed with neurokit2 python package [5]. We extracted R-peaks from lead II, and lead I was used instead in case of an error. If the error occurred on both leads, we imputed the missing values to -1 and used dummy variables to indicate the missing value. Heart rate was calculated based on R-peaks and it was scaled down by 100.

2.4. Deep learning models

We have chosen EfficientNet-B3 as the competition model. For research and development, we have also utilized ResNet-34. ResNet-34 performed worse than EfficientNet-B3, but it is a lighter model that requires less time for training. We trained EfficientNet-B3 with Adam optimizer with an initial learning rate of 0.001. The model was trained for 30 epochs with a batch size of 64. The learning rate was reduced by one-tenth in the 7^{th}, 14^{th}, and 25^{th} epochs.

2.5. Data augmentation

Augmentation is a cheap and popular method for increasing the size of the dataset. If properly designed, robustness of the classification can be improved as well. We have studied jittering, scaling, Gaussian blur, cutout (time out) [6], baseline shift, baseline wander, and powerline noise [7] as the possible augmentation schemes (see Figure 1). Among them, we have chosen only cutout and Gaussian blur schemes in our best entry. Each scheme was applied with the probability of $p = 0.25$, respectively.

2.6. Label masking

While the essential characteristics might remain similar over all the databases, the label availability of each class was dependent on the database. For instance, CPSC has 1,221 positive samples for atrial fibrillation (AF) but Ningbo has 0. There are at least three possible explanations. First, AF individuals were excluded at the time of data collection. Second, individuals with positive AF were excluded at the time of database generation. Third, AF individuals were present but AF was simply not labeled. For the first two cases, the negative labels can be considered to be correct because the individuals in the databases were not diagnosed of AF. For the third case, however, the database should not be regarded as full of negative labels for AF but should be regarded as no examination of AF label.

To prevent undesired effects of the third case, we have identified the classes with zero positive count in each database and performed masking. For database d and its class c, the training loss $l_c^d(x_i, y_i)$ was masked as $m_c^d \cdot l_c^d(x_i, y_i)$ where $m_c^d = 0$ if class c has zero positive count. For other classes with non-zero positive counts, m_c^d was set as 1. The same masking was also applied for validation and testing by multiplying model's prediction by the mask value.

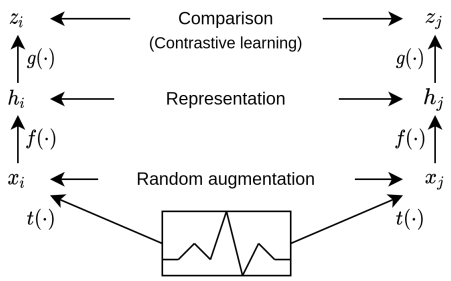

Figure 2. Self-supervised learning where $t(\cdot)$ is a random augmentation function that generates two distorted signals that are semantically equivalent, $f(\cdot)$ is an encoding network that will be pre-trained in an unsupervised manner and later be fine-tuned to the ECG classification tasks in a supervised manner, and $g(\cdot)$ is a projection network that maps the high dimensional representation vector h_i into the low dimensional vector z_i.

To apply proper masking at the test time, we had to distinguish databases where a sample in the test set comes from. When we develop our methods, we come up with two simple rules for distinguishing the samples of CPSC, G12EC, and the undisclosed American database. Both CPSC and G12EC have a 500Hz sampling rate, while the undisclosed American database has a 300Hz sampling rate. The mean values of each lead's recording are typically in [-0.5, 0.5] for CPSC, but not for G12EC. However, the UMich database is additionally used as test data in this year's challenge, and our label masking strategy was not appropriately applied during the final test time.

2.7. Threshold optimization

Our multi-label classification model outputs a real valued score $p \in (0, 1)$ for each class, and the classification threshold was optimized for each individual class. We examined threshold candidates between 0.1 and 0.7 with the step size of 0.05. The threshold was individually optimized using a surrogate metric of F1 score. For submission, we have used the average threshold values of seven experiments.

2.8. Self-supervised learning

In our study, we have adopted the recent contrastive learning approach in [4]. The self-supervised method was shown to be capable of learning effective representations from unlabeled datasets only. When the self-supervised model is fine-tuned with a labeled dataset of a small size, the resulting model's performance was on par with a fully supervised model that was trained with a labeled dataset of a large size.

The key assumption of the self-supervised learning is that randomly augmented views of the same sample should have the same semantic content as long as the augmentation functions are carefully designed to preserve the semantic information. Based on the assumption, learning is performed as shown in Figure 2. In our ECG study, we have used all of the seven augmentation schemes described in Section 2.5 as possible random augmentations, and $t(\cdot)$ randomly applied one or two of the seven augmentations for each input following RandAugment [8] implementation.

As for the loss that is used for comparing z_i and z_j, we have used the following NT-Xent loss [4] that explicitly pushes away two representations from two different individuals (negative pair, (i, k)) and pulls together two representations from a single individual with random augmentations (positive pair, (i, j)).

$$l_{i,j} = -log \frac{exp(sim(z_i, z_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{k \neq i} exp(sim(z_i, z_k)/\tau)}$$

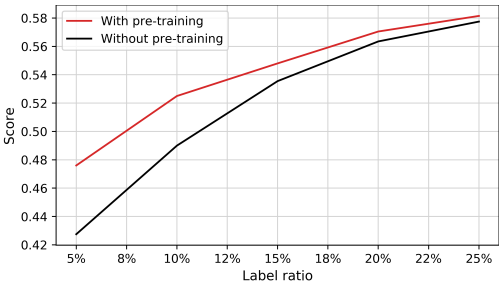

Figure 3. Challenge metric on local validation set with and without pre-training (6-lead). Only a subset of the samples was assumed to have label information available. The usefulness of self-supervised pre-training decreases as the label ratio (number of labeled samples / number of all samples) increases.

Figure 3 shows the experiment results. We pre-trained the encoding network (ResNet-34) and projection network (two-layer MLP with batch normalization) using self-supervised learning. Then, we replaced the projection network with a linear classifier and fine-tuned the model with the labeled samples. It can be seen that self-supervised learning as a pre-training is very helpful when label ratio is low. The advantage, however, fades away as the portion of labeled samples increase. Therefore, our final challenge entry did not utilize self-supervised learning. Self-supervised learning should be helpful when the size of labeled dataset is smaller or when there is an additional unlabeled dataset of a large size.

3. Results

Ablation study results using training data can be found in Table 1. All of augmentation, feature extraction, and la-

Category	Methods	Challenge metric
Base model	-	0.666 ± 0.002
Augmentation	None	0.661 ± 0.003
	Cutout only	0.666 ± 0.002
	Gaussian blur only	0.664 ± 0.003
Feature	None	0.665 ± 0.003
	Age/sex only	0.661 ± 0.001
Label masking	None	0.641 ± 0.002
	Training only	0.649 ± 0.003
	Training and validation only	0.649 ± 0.004
Pre-processing	None	0.644 ± 0.001
	Filtering only	0.672 ± 0.005
	Standardization only	0.667 ± 0.005

Table 1. Ablation study results for 6-lead experiments using training data.

Leads	Training	Validation	Test	Ranking
12	0.675 ± 0.003	0.626	0.48	12
6	0.664 ± 0.005	0.610	0.48	10
4	0.668 ± 0.002	0.612	0.47	11
3	0.671 ± 0.004	0.611	0.47	11
2	0.660 ± 0.003	0.610	0.45	13

Table 2. Challenge scores for our final selected entry (team *snu_adsl*) on our training set, scoring on the hidden validation set, and scoring on the hidden test set as well as the ranking on the hidden test set. The evaluation on our training set was repeated seven times with different seeds.

bel masking were helpful. For the feature extraction, it is interesting to note that using age and sex only can slightly deteriorate the performance. For the pre-processing, the result shows that only one of filtering or standardization should be used. Nonetheless, we have utilized both techniques because the evaluation results varied depending on how the data is split and because we considered both to be fundamental steps that can handle unexpected characteristics of unseen datasets.

The model performances for our training set, hidden validation set, and hidden test set are shown in Table 2.

4. Discussions

It can be seen that the challenge scores for the hidden test datasets are significantly lower than the challenge scores for the training or validation datasets. The unexpected performance degradation might be due to our incorrect assumptions on the label availability and characteristics of the test datasets.

As explained earlier, the results in Table 2 show that the performance is not significantly affected by the number of leads. The observation stands for all of training, validation, and test datasets, and it indicates that personalized devices with small number of leads might be effective for diagnosing cardiac abnormalities.

The size of challenge dataset turned out to be large enough and the well-known self-supervised learning approach in [4] was not helpful for performance enhancement. The approach, however, should be helpful when the size of labeled dataset is smaller or when there is an additional unlabeled dataset of a large size. Also, a possible future work is to develop a specialized self-supervised learning technique that can utilize ECG datasets better.

Acknowledgments

This work was supported by the Korea Medical Device Development Fund grant funded by the Korea government (the Ministry of Science and ICT, the Ministry of Trade, Industry and Energy, the Ministry of Health & Welfare, the Ministry of Food and Drug Safety) (Project Number: HI20C1662, 1711138358, KMDF_PR_20200901_0173) and in part by IITP grant funded by the Korea government (No. 2021-0-01343, Artificial Intelligence Graduate School Program (Seoul National University)).

References

[1] 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.

[2] 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.

[3] 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.

[4] Chen T, Kornblith S, Norouzi M, Hinton G. A simple framework for contrastive learning of visual representations. In International Conference on Machine Learning. PMLR, 2020; 1597–1607.

[5] 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;ISSN 1554-3528.

[6] Fan H, Zhang F, Gao Y. Self-supervised time series representation learning by inter-intra relational reasoning. arXiv preprint arXiv201113548 2020;.

[7] Mehari T, Strodthoff N. Self-supervised representation learning from 12-lead ECG data. arXiv preprint arXiv210312676 2021;.

[8] Cubuk ED, Zoph B, Shlens J, Le QV. Randaugment: Practical automated data augmentation with a reduced search space. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. 2020; 702–703.

Address for correspondence:

Wonjong Rhee

Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul 08826, Republic of Korea

wrhee@snu.ac.kr