A Mixed-Domain Self-Attention Network for Multilabel Cardiac Irregularity Classification Using Reduced-Lead Electrocardiogram

Hao-Chun Yang[1*], Wan-Ting Hsieh[2*], Trista Pei-Chun Chen[2]

[1]National Tsing Hua University, Hsinchu, Taiwan
[2]Inventec Corporation, Taipei, Taiwan

Abstract

Electrocardiogram(ECG) is commonly used to detect cardiac irregularities such as atrial fibrillation, bradycardia, and other irregular complexes. While previous studies had great success classifying these irregularities with standard 12-lead ECGs, there existed limited evidence demonstrating the utility of reduced-lead ECGs in capturing a wide-range of diagnostic information. In addition, classification model's generalizability across multiple recording sources also remained uncovered. As part of the PhysioNet/Computing in Cardiology Challenge 2021, our team HaoWan_AIeC, proposed ***Mixed-Domain Self-Attention Resnet (MDARsn)*** *to identify cardiac abnormalities from reduced-lead ECG. Our classifiers received scores of 0.4, 0.33, 0.37, 0.34, and 0.34 (ranked 18th, 23rd, 20th, 23rd, and 22nd) for the 12-lead, 6-lead, 4-lead, 3-lead, and 2-lead versions of the hidden test set with the evaluation metric defined by the challenge.*

1. Introduction

Cardiovascular diseases (CVDs) can be life-threatening which causes 17.9 million deaths each year. Early diagnosis of cardiac abnormalities is crucial as it can prevent complications and improve treatment outcomes[1]. The standard 12-lead electrocardiogram (ECG) is widely used to monitor cardiac function. However, the accessibility to 12-lead ECG is limited. The PhysioNet/Computing in Cardiology Challenge 2021 focused on automated, open-source approaches to classify cardiac abnormalities from the reduced set of leads (reduced-lead ECGs) [2, 3]. With reduced-lead ECGs, it is noted that signals from different leads are helpful in various CVDs diagnosis[4]. Methods that automatically learn the relationship between ECG leads and CVDs are desired.

To tackle domain discrepancy between the training and test sets, the work [5] learned domain-invariant features with domain-adversarial training by perturbing signals with adversarial gradients and augmenting model-based data. The result however was unsatisfactory when the test sets contain unseen data during training.

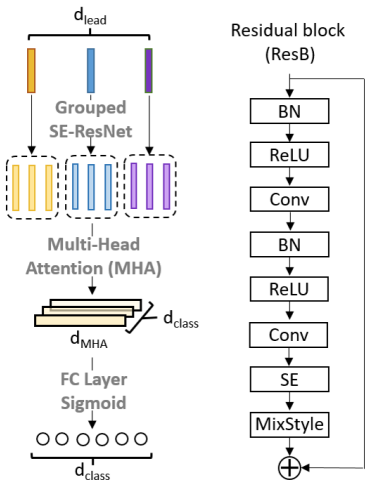

Figure 1. Left: the proposed architecture. Right: the residual block(ResB) with MixStyle blocks.

2. Methods

In this work, we present a Mixed-Domain self-Attention ResNet (MDARsn) to achieve two goals: coping with domain bias among different sources or between training and hidden datasets, and learning the relationships between ECG leads and CVDs automatically. We adopt feature-based augmentation method from [6] and Multi-Head Attention (MHA) layer to reach two objectives.

2.1. Data Pre-processing

Resampling Public training data provided from 2021 PhysioNet / CinC challenge were sourced from 5 locations. As recordings from separate hospitals could have different sampling rates, we upsampled or downsampled each recording to 500 Hz. Each recording was filtered by

* The two authors contributed equally to this paper.

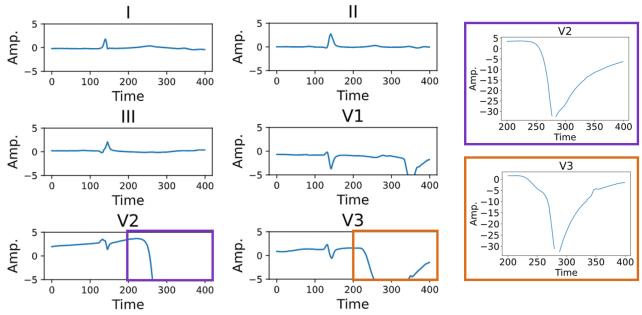

Figure 2. The chunked ECG signal of WFDB_Ga/E06072. The *NaN* fragments are found in lead V2 and V3, along with rapidly decreasing and increasing trend before and after the *NaN* fragment.

an FIR (finite impulse response) bandpass filter with bandwidth between $3 - 45$ Hz. Later we applied min-max normalization to every recording for the signal to range between -1 and 1.

Valid lead checking We observed that some leads contain non-numerical data (*NaN*). To avoid weighting on low-quality leads, we created a valid-lead mask for each recording where the ECG leads with *NaN* are marked as 0 and the rest are marked as 1. Such valid-lead mask is later passed to the MHA layer to avoid weighting on low-quality leads.

Data augmentation For better training, we augmented data by randomly cropping the signal and randomly generating valid-lead mask. Given a recording, we extracted a random window size of data. If the sequence is less than $T = 15$ seconds, we padded the sequence with zeros. The valid-lead mask for the sequence is randomly assigned with 0s, to simulate the broken lead situation seen in the real-world.

2.2. Model Architecture

2.2.1. Squeeze-and-Excitation ResNet (SERsn)

SERsn is reported a powerful framework to model ECG signals for CVD detection [7]. The proposed SERsn consisted of one convolution layer followed by $N = 8$ residual blocks (ResBs), each of which contained two convolution layers and a squeeze-and-excitation (SE) block [8]. The number of filters increased by a factor of two for every two ResBs. We grid-searched on the kernel size in $[7, 9, 11, 13, 15, 17]$ and found similar conclusion as [7] that larger kernel size at the first convolution layer and smaller kernel size for the rest achieved better results. We further modified the sequence of layers, applying batch-normalization first, followed by ReLU activation and then convolution layer.

2.2.2. Domain Generalization

Since the training datasets are collected from different sources, and there exists data discrepancy between the training and the hidden sets, we adopted MixStyle[6] which has demonstrated its generalization capability with feature-based augmentation.

MixStyle is derived from instance normalization which could effectively remove instance-specific characteristic, while it further mixes the feature statistics of feature instances from two domains with a random weight. The basic implementation of instance normalization is as defined in Equation 1, which normalizes feature maps F with means and standard deviations computed within each channel, followed by the affine transformation.

$$IN(F) = \gamma \odot \frac{F - \mu(F)}{\sigma(F)} + \beta \tag{1}$$

Where $F \in \mathbb{R}^{B \times C \times H \times W}$ with B, C, H and W denoting batchsize, channel, height and width respectively. $\gamma, \beta \in \mathbb{R}^{C}$ are the affine transformation coefficient. Specifically, for $c \in 1, ..., C$,

$$\mu(F)_c = \frac{1}{BHW} \sum_{b=1}^{B} \sum_{h=1}^{H} \sum_{w=1}^{W} F_{b,c,h,w}, \tag{2}$$

and

$$\sigma(F)_c = \sqrt{\frac{1}{BHW} \sum_{b=1}^{B} \sum_{h=1}^{H} \sum_{w=1}^{W} (F_{b,c,h,w} - \mu(F)_c)^2} \tag{3}$$

Given two sets of feature instances F and F', MixStyle performs domain generalization by generating a mixture of feature statistics as defined in Equation 4,

$$MixStyle(F, F') = \gamma_{mix} \odot \frac{F - \mu(F)}{\sigma(F)} + \beta_{mix} \tag{4}$$

$$\gamma_{mix} = \lambda\sigma(F) + (1 - \lambda)\sigma(F'), \tag{5}$$

$$\beta_{mix} = \lambda\mu(F) + (1 - \lambda)\mu(F') \tag{6}$$

where λ is random weight sampled from beta distribution, $\lambda \sim Beta(\alpha)$, with $\alpha = 0.1$. As F and F' requires two sources, the simple solution is to shuffle the order of the batch dimension of F to obtain F', which is validated to attain performance comparable to cross-domain mixing[6]. MixStyle blocks are inserted after SE-block in ResBs, besides they are only applied to shallow ResBs blocks which capture more domain-related information. We tuned the number of layers that applied MixStyle, n_{mix}, and found adding it to previous $1 \sim 2$ ResBs is enough.

2.2.3. Mixed-Domain self-Attention SERsn (MDARsn)

The proposed MDARsn aims to learn the domain-invariant and lead-independent representation. It ignores readings from broken leads and automatically weights on the rest of the leads for final prediction. The proposed MDARsn has three modules: (1) grouped SERsn to learn the lead-independent representation; (2) feature-based data augmentation, MixStyle[6], to generalize for different domains; (3) MHA layer to mask low quality ECG leads and to weight on the rest lead-independent representation.

We learn the lead embedding from SERsn by setting the cardinality to the number of leads. Under this setting, the input channels would be divided in groups (i.e, each lead forms a group) and would learn different types of features while increasing the efficiency of weights[9] through this groups SERsn. In addition, the MixStyle blocks are added to shallow ResBs. Once the lead embedding is extracted from SERsn, we use the MHA layer to jointly attend to information from different ECG leads[10]. The MHA is based on dot-product attention as defined in below:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}}) \qquad (7)$$

Where Q, K, V are query, key and value vector derived by linear projection of lead embedding, while d_k denotes the dimension of key vector. We set the number of heads as the number of classes, d_{class}, and choose the embedding dimension $d_{model} = 520$. We input the valid-lead mask described in section 2.1 to MHA so that only the embedding of valid leads would be considered in the final classification. Ultimately, for every recording we obtain an output from MHA layer with dimension (d_{class}, d_{model}/d_{class}), which would be later passed into a fully connected layer and output (d_{class}, 1).

2.3. Implementation Details

To reduce the training and validation time, we excluded samples that were not labeled in one of the 26 classes that will be considered for challenge metric[3]. Besides, we also excluded recordings from PTB & StPetersburg as described in [11] due to the long average lengths. We then split the samples from the rest of datasets[12–15] into 5-fold using iterative stratification to guarantee the label balance in each folds[16]. We conducted all experiments using 3 folds as training, 1 fold as validation and 1 fold to test. During model training we monitored validation precision score (PRC) and used early stopping when validation PRC had stopped improving for 5 epochs.

We utilized a standard binary cross entropy loss function averaged over 26 classes to train the model. The optimizer is Adam and a cosine annealing was applied with warm-up 1000 steps and maximum 10000 steps. Several hyper-parameters were greedily searched using Optuna Toolkit [17] monitoring on cross-validated testing set's challenges metric: learning rate is searched within $[0.0001 \sim 0.001]$, dropout $[0.1 \sim 0.5]$, n_{mix} among $[1 \sim 3]$, convolution kernel sizes $[5, 7, 9, 11, 13, 15, 17]$ and number of stacked ResBs $[6, 8, 10, 12]$. Models were trained using PyTorch on two 2080-Ti GPUs with a batch size of 96. Each epoch took roughly 5 minutes to train and around 2.5 hours to complete a parameter set. The final best searched parameters were listed in Table 1.

Leads	12	6	4	3	2
General					
ECG window size (secs)	15	15	15	15	15
Sampling frequency (Hz)	500	500	500	500	500
Number of classes, d_{class}	26	26	26	26	26
SERsn					
Input channel	12	6	4	3	2
Channel of the first conv. layer	128	128	128	128	128
kernel size of the first conv. layer	11	11	11	11	11
Kernel size of ResBs	7	5	13	7	7
Stride	3	3	3	3	3
MixStyle					
Number of applied layers, n_{mix}	2	2	2	2	1
MHA					
Embedding size, d_{model}	650	520	650	520	650
Number of heads, d_{head}	26	26	26	26	26

Table 1. Hyperparamters searched for different leads.

Model	Leads	Training	Validation	Test	Ranking
SERsn	12	0.721	0.582	-	-
	6	0.7	0.576	-	-
	4	0.704	0.580	-	-
	3	0.704	0.586	-	-
	2	0.699	0.580	-	-
SERsn+M	12	0.731	**0.602**	0.4	18
	6	0.709	**0.593**	0.33	23
	4	0.711	**0.597**	0.37	20
	3	0.713	**0.591**	0.34	23
	2	0.705	**0.589**	0.34	22
SERsn+M+A (**MDARsn**)	12	**0.738**	0.525	-	-
	6	**0.71**	0.506	-	-
	4	**0.723**	0.511	-	-
	3	**0.719**	0.503	-	-
	2	**0.707**	0.499	-	-

Table 2. Challenge scores for different models testing on training set, repeated scoring on the hidden validation set, and one-time scoring on the hidden test set as well as the ranking on the hidden test set. M: with MixStyle block; A: with MHA layer.

3. Results

Table 2 reports the Challenge score [3] on the training, hidden validation and test sets of three models: SERsn, SERsn with mixStyle (SERsn+M) and the pro-

posed MDARsn. We demonstrate that with the MixStyle blocks, the results outperform the plain SERsn in both training and hidden validation sets. We also obtain better performance using Attn-Rsn in the training set, increasing at least 0.02 of the Challenge score. However, there is a drop with the proposed MDARsn of approximately 0.08 in the hidden validation set comparing to SERsn+M. As SERsn with mixStyle shows the steady performance overall, we choose it as our final model and report its results on hidden test set in Table 2.

4. Discussion and Conclusions

In this study, we proposed Mixed-Domain self-Attention ResNet (MDARsn) to classify 26 CVDs from reduced-lead ECGs. Domain-specific characteristics are removed by the *MixStyle* block. With its source invariant representation, the classifier is robust when encountering unseen ECG recordings. We also incorporated the *Multi-Head Self-Attention* block to handle missing or low-quality reduced-lead ECGs. Both measures provided improved classification on the local nested training set.

Our model exhibited two weaknesses. First we found a performance gap between the training set and the hidden validation/test set provided by the challenge host. Second, we found it under-performing on the unseen local datasets (PTB & StPetersburg). These two weaknesses suggested that we further investigate our data preprocessing pipeline for different recording sources as MHA did not seem to handle different data sources as well as expected. We hope to further improve MDARsn's generalizability and flexibility in reduced-lead ECG classification.

References

[1] Vitale G, Pivonello R, Lombardi G, Colao A. Cardiac abnormalities in acromegaly. Treatments in Endocrinology 2004;3(5):309–318.

[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] Gunnarsson G, Eriksson P, Dellborg M. ECG criteria in diagnosis of acute myocardial infarction in the presence of left bundle branch block. International Journal of Cardiology 2001;78(2):167–174.

[5] Hasani H, Bitarafan A, Baghshah MS. Classification of 12-lead ECG signals with adversarial multi-source domain generalization. In Computing in Cardiology, CinC 2020, Rimini, Italy, September 13-16, 2020. IEEE, 2020; 1–4.

[6] Zhou K, Yang Y, Qiao Y, Xiang T. Domain generalization with mixstyle. CoRR 2021;abs/2104.02008.

[7] 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 Computing in Cardiology, CinC 2020, Rimini, Italy, September 13-16, 2020. IEEE, 2020; 1–4.

[8] Hu J, Shen L, Sun G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; .

[9] Xie S, Girshick RB, Dollár P, Tu Z, He K. Aggregated residual transformations for deep neural networks. In 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017. IEEE Computer Society, 2017; 5987–5995.

[10] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA. 2017; 5998–6008.

[11] Min S, Choi H, Han H, Seo M, Kim J, Park J, et al. Bag of tricks for electrocardiogram classification with deep neural networks. In Computing in Cardiology, CinC 2020, Rimini, Italy, September 13-16, 2020. IEEE, 2020; 1–4.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] Szymanski P, Kajdanowicz T. A network perspective on stratification of multi-label data. In First International Workshop on Learning with Imbalanced Domains: Theory and Applications, LIDTA@PKDD/ECML, 22 September 2017, Skopje, Macedonia, volume 74 of Proceedings of Machine Learning Research. PMLR, 2017; 22–35.

[17] Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna: A next-generation hyperparameter optimization framework. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019, Anchorage, AK, USA, August 4-8, 2019. ACM, 2019; 2623–2631.

Address for correspondence:

Wan-Ting Hsieh
111 No. 166, Sec. 4, Chengde Rd., Shilin Dist., Taipei, Taiwan
hsieh.eileen@inventec.com