Explainable Deep Neural Network for Identifying Cardiac Abnormalities Using Class Activation Map

Yu-Cheng Lin, Yun-Chieh Lee, Wen-Chiao Tsai, Win-Ken Beh, An-Yeu (Andy) Wu

Graduate Institute of Electronic Engineering, National Taiwan University, Taipei, Taiwan

Abstract

In this study, our team "NTU-Accesslab" present a deep convolutional neural network (CNN) approach, called CNN-GAP, for classifying 12-lead ECGs with multilabel cardiac abnormalities. Additionally, Class Activation Mapping (CAM) is employed for further understanding the decision-making process of this black-box model, making the model more explainable.

The CNN-GAP model consists of 12 layer Conv Blocks along with Batch Normalization layer, Global Average Pooling and Fully Connected layer with sigmoid activation. To deal with the data imbalance problem, we oversample the minor datas. In the training stage, we applied Macro observed score loss (Macro-Obs) instead of the conventional Weighted Cross entropy loss (WCE), and we have shown that this results in higher challenge scores. Additionally, we augmented datas by randomly scaling datas to get better scores and prevent model overfitting. Our method achieved a challenge score of 0.58 on the validation set, but was unable to score and rank on the test set, due to a failure of the algorithm on the fully hidden dataset.

1. Introduction

The electrocardiogram (ECG) is a representation of electrical activities of heart that can be measured non-invasively. 12-lead ECG is then obtained through placing ten electrodes on patients' limbs and wrists, and presents the heart's electrical potential from twelve different angles. ECG is widely used in clinical diagnosis. A variety of cardiac abnormalities, such as ventricular fibrillation and atrial fibrillation, can be detected through analysis of ECG. Early treatment of cardiac disease can significantly reduce morbidity and mortality rate. However, manual interpretation of ECG is time consuming, prone to error and requires person with high level of training.

Therefore, in the past, numerous approaches were proposed to interpret and classify ECG automatically. Some use hand-crafted features such as R-R interval and heart rate variability (HRV), while others utilize neural network's feature engineering ability. Yet, these approaches have limited applicability since they only classify ECG into a few classes and are only tested on small and relatively homogeneous datasets.

The 2020 PhysioNet/CinC Challenge [1] aims to address this challenging problem and encourages participants to develop an algorithm that can classify 12-lead, variable length ECG from multiple sources into 27 classes. In this challenge, we have proposed a end-to-end, interpretable, deep convolutional neural network (CNN), for the automatic classification of ECG signals.

2. Methods

2.1. Data description

12-lead ECG recordings are a mixture of 6 datasets from multiple sources across different countries (Table 1). Each ECG recording has one or more labels from different types of abnormalities in SNOMED-CT codes, and only 27 classes are scored in the challenge (3 of them are equivalent class, so only 24 classes are under classification).

Table 1. Data profile for the training set.

Dataset	*Sample freq. (Hz)*	*Duration (s)*	*Num. of records*
PTB-XL	500	10	21837
CPSC-Extra	500	10 - 98	3453
CPSC	500	9 - 118	6877
PTB	1000	38.4 - 120	516
Geogia	500	10	10344
StPetersburg	257	1800	74

2.2. Data preprocessing

- Train test split: train-test set splitting is done separately on each class of single label data to make the label distribution on training and testing set more similar.

ISSN: 2325-887X DOI: 10.22489/CinC.2020.072

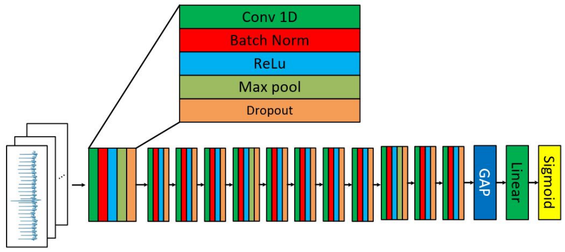

Figure 1. Model architecture

- Long data slicing [2] & Zero padding: If the data has a length over our desire clipping length (10 sec.), we clip the data from the head, middle, and end. By doing this, we can increase the amount of training data. As for the data with a length smaller than our desire length, we pad zeros on both sides.

- Remove baseline: Since some datasets have a non-zero and severe skew baseline problems, we apply linear regression to the 10 sec. clipping data to deal with this problem

2.3. Network architectures

The CNN-GAP model (Fig. 1) is consist of 12 Convolutional blocks, each with Batch-normalization and ReLU activation, Global Average Pooling and Fully Connected layer with sigmoid activation.

2.4. Training techniques

- Data augmentation: random scaling 0.5 to 2.5 with uniform probability on the amplitude of the data.

- Loss function: We implement two types of loss functions. First, the weighted binary cross entropy loss (WBCE), with ratio of positive and negative samples as its weight, is applied. Second, the Macro observe score loss (Macro-Obs) is applied. The Macro-Obs loss is calculated as follows:

p: prediction of the model
t: true label of the data
W: weighted matrix provided by organizers
$\odot$: element-wise multiplication
$\otimes$: tensor product
$N = \sum_{per\ batch}(1-t) \odot p + t$ (Normalize factor)
$A = p^T \otimes t$ (modified confusion matrix without normalization)
$A_{norm} = \sum_{in\ batch\ direction}(A/N)$
Macro-Obs $= \sum(A_{norm} \odot W)$

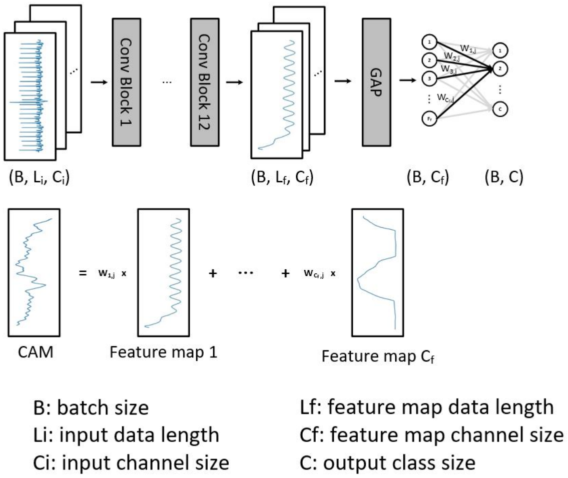

Figure 2. CAM generating flow.

2.5. Class activation mapping

Zhou et al. [3] generates the class activation maps (CAM) using the global average pooling (GAP) in CNNs. Class activation map for a particular class indicated the regions in the image used by the CNN to identify that category. Same technique can also be applied in ECG detections [4]. The procedure for generating CAMs is illustrated in Fig. 2.

As illustrated in Fig. 2, GAP outputs the spatial average of the feature map of each unit at the last CNN layer. A linear combination of these values is then used to generate the final output. Similarly, we can compute a weighted sum of the feature maps of the last CNN layer to obtain the CAMs.

Assume that $f_i(.)$ represents the function of the deep CNN in channel i, w_{ij} represents the weight of the fully connected layer for channel i and class j, and b_j is the bias of class j. Then, the output value of class j under no activation is $\sum(w_{ij} \times avg(f_i(.))) + b_j$, which can also be express as $avg(\sum(w_{ij} \times f_i(.))) + b_j$. Notice that $\sum(w_{ij} \times f_i(.))$ is exactly the CAM of class j; therefore, the spatial average of the CAM is exactly the output values with bias. With the spatial information reserved, the high activated value in CAMs help to understand the decision-making process of the black-box model. CAM techniques can also be applied to identify the temporal location for cardiac abnormalities.

For instance, Fig. 3 shows that the waveform of lead-I (the upper waveform in the subfigure) and its corresponding CAM (the lower waveform in the subfigure). If the value of CAM is larger than a threshold, then we marked the region of the original data red. We have shown that the marked region highly activated the models output, also CNN model preserves the spatial information. Therefore, we can see how the black-box model makes its decision by observing these highly activated region.

If this region is same as human's perspective for identify the cardiac abnormalities, then the model might have learned the correct information. If the region is no the same as what we expect to see, then we human might be able to learn something new for identify the abnormalities from the model.

3. Results

Table 2 shows the challenge score under different model architecture and loss functions in the self-divided validation set (denoted as 'Valid. 1') and the challenge's validation set (denoted as 'Valid. 2'). We can see that Macro-Obs loss outperforms WBCE loss.

Table 3 shows the challenge scores of some high-ranking teams and ours for the hidden test sets. We can see that our score is competitive with these teams.

Table 4 shows the accuracy and F1 scores of each clasess on the validation set. We can see that data imbalance problems cause minor classes in the train set gets extremely low validation accuracy no matter which loss function are applied. Additionally, when applied Macro-Obs loss, we can see that the minor classes gets even worse performance compared with WBCE loss. That is because the challenge loss would penalize the minor classes.

Table 2. Results of different loss functions

Loss function	*Valid. 1 score*	*Valid. 2 score*
WBCE	0.549	0.544
Macro-Obs	0.582	-

*Missing validation score is due to late submission.

Table 3. Results for hidden cases

Team	*Valid. score*	*Test 1*	*Test 2*	*Test 3*	*Full Test*
Rank 1	0.587	0.761	0.558	0.492	0.533
Rank 7	0.586	0.643	0.574	0.298	0.417
Rank 11	0.435	0.556	0.418	0.290	0.354
Ours	0.544	0.725	0.510	-	-

*Test set 3 contains sample freq. not present in the train set. Our code did not handle this sample freq., thus not reproducible in this case.

4. Discussion

By observing the CAM, we sum up two issues. First, data corruption problems cause the model to learn wrong cardiac abnormalities pattern. Second, the model does not learn the abnormal region as number of classification type increase.

Table 4. Results of each classes (on validatation set)

SNOMED	*Acc.*	*F1 (Macro-Obs)*	*F1 (WBCE)*	*Num. of records*
270492004	0.495	0.445	0.327	2085
164889003	0.657	0.638	0.555	3132
164890007	0.097	0.103	0.219	233
426627000	0.550	0.429	0.361	406
713427006	0.645	0.642	0.559	3241
713426002	0.204	0.225	0.200	1080
445118002	0.267	0.225	0.194	1170
39732003	0.348	0.310	0.272	3961
164909002	0.359	0.363	0.352	861
251146004	0.014	0.022	0.077	356
698252002	0.026	0.036	0.091	662
10370003	0.312	0.441	0.781	212
284470004	0.384	0.401	0.312	1897
427172004	0.051	0.066	0.079	536
164947007	0.092	0.089	0.122	217
111975006	0.190	0.205	0.288	993
164917005	0.000	0.000	0.080	678
47665007	0.033	0.049	0.173	273
427393009	0.011	0.019	0.175	837
426177001	0.482	0.496	0.341	1636
426783006	0.707	0.690	0.598	14884
427084000	0.509	0.518	0.452	1783
164934002	0.206	0.236	0.212	3072
59931005	0.101	0.088	0.085	718

For the first issue, ECG corruption problems, including abnormal spikes or fluctuations would mislead the models decision. For instance, the waveform in Fig. 3 is obviously corrupt, some pre-processing technique such as filtering or clipping can be applied to remove these. Some would say that the CNN-based model has the capacity to do filtering and clipping, and additional preprocessing will cause information loss thus get worse performance. However, by CAM, we see that the model focus on the noisy part to make decision, so it did not learn to filter out the abnormal spike or noise.

For the second issue, we find that the model did learned the abnormal pattern in PhysioNet Challenge 2017 (4 classes) dataset, that is, most of the CAM is explainable from human perspective. However, the CAM in CPSC 2018 (9 classes) gets more unexplainable. Even worse, CAM in this challenge (24 classes) is beyond human understanding of arrhythmia. For instance, the upper-left of Fig. 4 is a normal sinus rhythm (NSR) labeled data, but the abnormal part of the waveform is more activated. However, the overall performance does not decrease that much, that means the CAM still contains some important information. By observing these unexplainable CAMs, we human may be able to learn something new to identify the cardiac abnormalities.

Apart from the unexplainable CAMs, there are still some explainable cases. For instance, the upper-right of Fig. 4 is a complete right bundle branch block (CRBBB) labeled data. The Right bundle branch block is focused by the model, and the most activated region has the largest RBBB pattern. The lower-left of Fig. 4 is a atrial fibrillation (AF) labeled data, and the P-wave region is focused by the model. The lower-right of Fig. 4 is a bradycardia (Brady) labeled data, and the low heart rate region is activated.

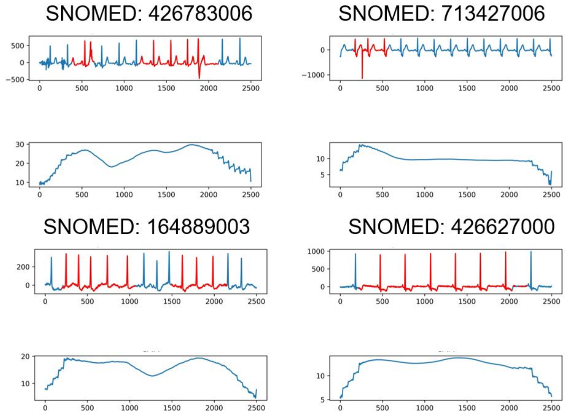

Figure 3. Lead-I and CAM waveform of corrupt data

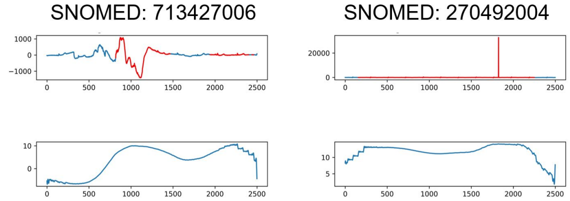

Figure 4. Lead-I and CAM waveform of some correctly classified data

5. Conclusions

We trained a deep convolutional neural network to classify 12-lead ECGs as 24 classes. We apply the Macro-Obs loss to get the challenge score 0.58 on the validation set 1, which outperforms the conventional WBCE loss, but was unable to score and rank on the validation set 2 and test set, due to late submission and a failure of the algorithm on the fully hidden dataset. The final performance is likely to be lower than the validation score. The class activation mapping is generated for each ECG to visualize the region of the waveform that the model was focusing on when making the decision. We demonstrate some explainable CAM from human perspective. Also, we show that the black box model does not learn the abnormalities same as human perspective as number of classification classes increase. By observing these CAMs, we human may be able to learn something new to identify the cardiac abnormalities.

Acknowledgments

This work was supported in part by the Ministry of Science and Technology of Taiwan under grant MOST-109-2622-8-002-012-TA.

References

[1] Perez Alday EA, Gu A, Shah A, Robichaux C, Wong AKI, Liu C, Liu F, Rad BA, Elola A, Seyedi S, Li Q, Sharma A, Clifford GD, Reyna MA. Classification of 12-lead ECGs: the PhysioNet/Computing in Cardiology Challenge 2020. Physiol. Meas. 2020 (Under Review).

[2] J. Zhu, Y. Zhang and Q. Zhao, "Atrial Fibrillation Detection using Different Duration ECG Signals with SE-ResNet," 2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP), Kuala Lumpur, Malaysia, 2019, pp. 1-5, MMSP.2019.

[3] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva and A. Torralba,"Learning Deep Features for Discriminative Localization," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, 2016, pp. 2921-2929, CVPR.2016.

[4] S. Goodfellow, A. Goodwin, Robert Greer, P. Laussen, Mjaye Mazwi and D. Eytan, "Towards Understanding ECG Rhythm Classification Using Convolutional Neural Networks and Attention Mappings," Machine Learning for Healthcare (MLHC), 2018.

Address for correspondence:

Yu-Cheng Lin, Yun-Chieh Lee, An-Yeu Wu,
Wen-Chiao Tsai, Win-Ken Beh
No.1, Sec. 4, Roosevelt Road, Taipei, 10617 Taiwan
{r09943001, b06901014, andywu}@ntu.edu.tw,
{daniel, kane}@access.ee.ntu.edu.tw