Adaptive lead weighted ResNet trained with different duration signals for classifying 12-lead ECGs

Zhibin Zhao [1], Hui Fang[2], Samuel D. Relton[2], Ruqiang Yan[1], Hui Fang[3], Yuhong Liu[4], Zhijing Li[1], Jing Qin[5], David C. Wong[6],

[1] Xi'an Jiaotong University, Xi'an, China (note - full address at end of paper)
[2] University of Leeds, Leeds, UK [3] Loughborough University, Loughborough, UK [4] Chengdu Medical College, Chengdu, China [5] Dalian University, Dalian, China [6] University of Manchester, Manchester, UK

Abstract

Introduction: We describe the creation of a ensemble deep neural network architecture to classify cardiac abnormality from 12 lead ECGs. The model was created by the team between a ROC and a heart place *for the PhysioNet/Computing in Cardiology Challenge 2020.*

Methods: ECGs were downsampled to 257 Hz and then set to a consistent duration by randomly clipping or zero-padding the signal to 4096 samples. To learn effective features, we created a modified ResNet with larger kernel sizes that models long-term dependencies. We embedded a Squeeze-And-Excitation layer into the modified ResNet to learn the importance of each lead, adaptively. A simple constrained grid-search method was applied to deal with class imbalance.

Results: Using the bespoke weighted accuracy metric, We achieved a 5-fold cross-validation score of 0.684, sensitivity and specificity of 0.758 and 0.969, respectively. The corresponding result for the hidden test set was 0.672.

Conclusion: The proposed prediction model performed well on the validation and hidden test data. Such models may be potentially used for ECG screening or diagnosis.

1. Introduction

We consider the task of cardiac abnormality classification from 12-lead electrocardiogram (ECG) recordings of varying sampling frequency and duration. 12-lead ECGs are commonly used in clinical care to discern cardiac abnormalities such as arrhythmias, myocardial infarction, or coronary occlusion [1].

Each of the 12 leads correspond to the heart's electrical activity from a distinct angle, and can be mapped to the anatomy of the heart. A skilled interpreter can therefore use ECG signals from multiple leads to localise the source of a cardiac abnormality.

In practice, human ECG interpretation is limited by the availability of a trained cardiologist and the time required to synthesize information from the 12-lead signal (and document findings). In the absence of cardiology experts, other clinicians may make preliminary interpretations, but are demonstrably less accurate [2].

Computer-aided interpretation has been suggested as one approach for circumventing these resource constraints, despite historical limitations in accuracy [3]. Modern deep learning methods may be able to improve interpretation accuracy. Until recently, the use of such techniques for 12-lead ECGs has been impractical due to the shortage of labelled training data. There remains room for improvement over initial promising results [4].

The release of a new large, labelled, multinational, 12-lead ECG data set as part of the 2020 Physionet Challenge [5] presents a unique opportunity to tackle multi-class cardiac abnormality detection.

We tackle the problem by developing a deep neural network architecture. Our architecture acknowledges the importance of the spatial relationship between the ECG channels by using a squeeze-and-excitation (SE) block. The SE approach was developed by Hu et al. who showed significant improvements over previous deep neural network architectures when introduced as part of the 2017 ILSVRC classification challenge [6].

2. Methods

Our objective was to create a model that could accurately classify 12-lead ECG recordings into one of 27 clinical diagnoses shown in Fig. 1. Three pairs of classes were scored identically. For this task, we considered these pairs to be identical, so that only 24 classes were considered. As in clinical practice, recordings may have multiple diagnoses.

The dataset used to train and validate the model con-

sisted of 37749 12-lead ECG recordings from four different sites. A separate dataset, that included data from a fifth site, was withheld for testing. The recordings were of varying frequency (257 Hz - 1000 Hz) and duration (6 s - 60 s). A very small selection (n=74) of the data set had a duration of approximately 30 minutes. Each ECG was associated with and age and gender. A detailed description of the data and the classification task is presented in [5]. Training code is available at https://bit.ly/3gSJZr0

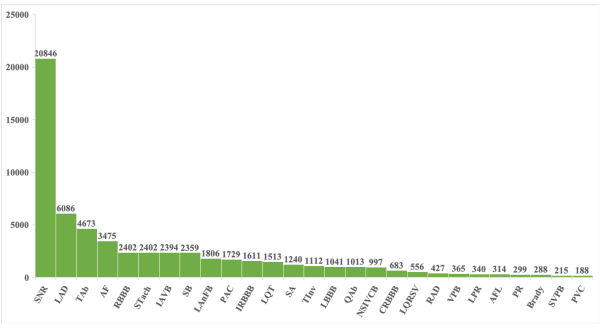

Figure 1. Summary of classes used in the 2020 Physionet Challenge.

2.1. Data pre-processing

All ECGs were resampled to the minimum frequency of 257 Hz. To allow a fixed input size in the deep learning model, each ECG was set to be 4096 points. During training, we zero-padded shorter duration signals and randomly clipped longer duration signals.

We scaled age into the range [0,1]. Both age and gender were encoded using one-hot encoding, with an two additional mask variables to represent missing values (Fig. 2).

Age	Gender
25	Male
Unknown	Female
100	Unknown

Age	A_Mask	Gender		G_Mask
0.4	1	1	0	1
0	0	0	1	1
1	1	0	0	0

Figure 2. The input of age and gender.

2.2. Model description

After obtaining the input signals, we designed an improved ResNet to assign the 12-lead ECG recordings into the 24 diagnostic classes. As shown in Fig. 3, the improved ResNet consists of one convolutional layer followed by $N = 8$ residual blocks (ResBs), each of which contain two convolutional layers and a squeeze and excitation (SE) block (Fig. 4).

The first (convolutional) layer and the initial two ResBs units have 64 convolution filters. The number of filters increases by a factor of two for every second ResB unit. The feature dimension is halved after the max pooling layer, and the third, fifth, and seventh ResBs.

The improved ResNet has four modifications from the original ResNet [7]. First, we modified the final fully connected (FC) layer to incorporate patient age and gender. These two features were passed through another FC layer with 10 neurons prior to inclusion in the final layer. Second, we used a relatively large kernel size of 15 in the first convolutional kernel, and a large kernel size equal to 7 in the latter convolutional kernels. Previous work has shown that large kernel sizes are more helpful for networks to learn meaningful features [8]. Third, as shown in Fig. 4, we added a dropout layer with a drop out rate of 0.2 between two convolutional kernels in the residual block (ResB) to reduce the likelihood of overfitting. Finally, we added a SE block into each ResB depicted in Fig. 4. The SE block has been to model channel interdependencies, and in this case, we incorporate it to model the spatial relationship between the ECG channels. The SE block, introduced by Hu et al. [6] uses a multi-layer perceptron (MLP) with one hidden layer to calculate the importance of the channels. The parameter $r = 16$ in Fig. 4 denotes the reduction factor, which controls the capacity of the MLP.

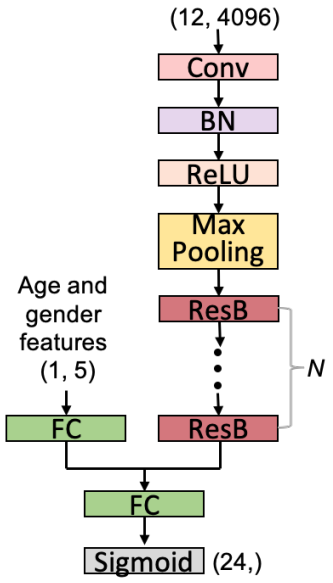

Figure 3. The proposed network architecture.

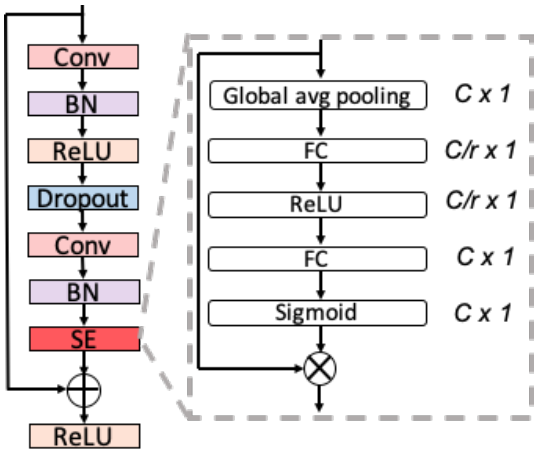

Figure 4. Residual block (ResB) and SE block.

The training error for this multi-task problem was average binary cross-entropy loss. The loss was optimised using the Adam optimizer with an initial learning rate 0.003. The learning rate was reduced tenfold in the 20th and 40th epochs, and the model was trained for in total 50 epochs with a batch size of 64.

2.3. Model evaluation

Five-fold cross-validation was used to assess the performance of the model. For the validation and test signals, we continued to zero-pad any shorter duration signals with fewer than 4096 samples. For signals longer than 4096 samples, we segmented the signals into multiple patches with a fixed overlap $O = 256$. An example with the length 10000 is depicted in Fig. 5. The number of patches, P, for a single signal can be formulated as:

$$P = \text{ceil}(\frac{L - 4096}{4096 - O}) + 1 \quad (1)$$

where ceil$(\cdot)$ rounds a number upward to the nearest integer. We used the mean probabilities from the P patches to classify the recording.

In addition to sensitivity and specificity, we report model performance using a bespoke metric, $s_{normalized}$, as described in [5]. This metric is a weighted accuracy that rewards incorrect classifications with similar risks or outcomes to the true class.

2.4. Threshold optimisation

Successful classification was heavily dependent on solving issues related to class imbalance. The training data suffered from significant class imbalance, as shown in Fig. 1.

Kang et al. [9] previously suggested that accurate representation is possible even in the presence of class imbalance. If the representation is accurate, then strong classification performance may be achieved by adjusting the classifier. By separating representation and classification, we avoid having to include extra class-balancing approaches such as data re-sampling.

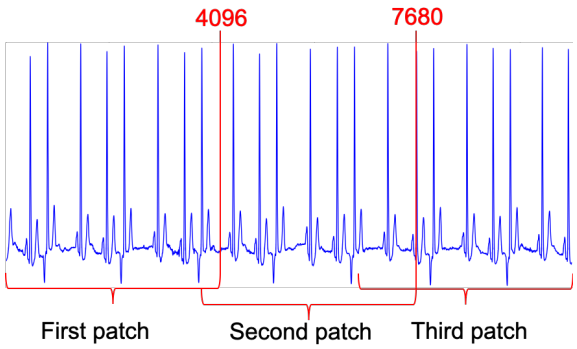

Figure 5. An example of segmenting the validation signals.

For the single task problem, researchers often re-balance the decision boundaries of classifiers via normalizing the classifier weight norms. In our multi-task problem, we used a similar approach, attempting to adjust the decision boundaries of the classifier by optimizing the thresholds corresponding to each class.

However, searching for the joint set of optimal thresholds becomes intractable as the number of classes increases. The current problem would require searching in a 24 dimensional space, which we considered to be too time-consuming.

Instead, we designed a simple constrained grid-search method to optimize the thresholds relied on a basic assumption that each class is independent. This method consists of two steps: (1) Set the same thresholds to all the classes and search for the optimal thresholds in [0, 1] with a step 0.1; (2) Optimize each threshold in [0, 1] with a step 0.01 separately, when other thresholds are fixed.

2.5. Ensemble learning

To improve the robustness of the classifications, we created an ensemble of five models trained via five-fold cross-validation. The thresholds of each model was optimized by its split validation set, and ECGs were classified according to the majority vote.

Due to technical issues, the ensemble model was trained locally, but could not be trained and tested on the external challenge virtual machines. Experience in related domains suggests that the ensemble model performance would likely improve on the single model.

3. Results

The results of five-fold cross-validation results on the released data set as well as the official test set are shown in Table 3. Our best performing improved ResNet with optimized thresholds for the hidden test set was chosen as the final submission resulting in $s_{normalized}$=0.672. From the results, we observe that threshold optimization plays an important role in improving the performance, especially for the bespoke metric. We noted that there is only a small difference between the online and offline accuracy of the improved ResNet model with optimized thresholds, indicating that the model did not overfit to the training data.

Table 1. Model Results with different thresholds using five-fold cross-validation. Sub1: an improved ResNet without any threshold optimization; Sub2: an improved ResNet with the thresholds optimized only by the step one; Sub3: an improved ResNet with thresholds optimized by constrained grid-search.

Method	Sens.	Spec.	$s_{normalized}$	Test Acc.
Sub1	0.599	0.986	0.630	0.607
Sub2	0.742	0.969	0.675	0.666
Sub3	0.758	0.969	0.684	0.672

4. Discussion

We have developed a deep learning model that accurately classifies 24 unique cardiac abnormalities from 12-lead ECGs. Our approach used a deep neural network architecture that combined an improved ResNet with an SE layer. The addition of the SE layer modeled the spatial relationship between channels. The improved ResNet learned the features effectively from the time series, as demonstrated by the model performance metrics. The approach ought to generalise well, given the size and heterogeneity in the training data set.

Like many other deep learning approaches, the results from the presented model are not easily explainable in the sense that we cannot determine the specific ECG morphology that results in a classification. It ought to be possible extend our model to accommodate better explainability via attention or learned prototypes [10, 11].

We further note that the upper bound on accuracy of the model is potentially limited by noisy training data labels. Methods that explicitly model uncertainty in the labels may lead to more robust performance.

In future work, we intend to improve model performance by first conducting cluster analysis of false classifications to determine common modes of failure. An ensemble approach using classifiers with hand-crafted features may then allow better prediction of such modes.

References

[1] Davies A, Scott A. Starting to read ECGs: A Comprehensive Guide to Theory and Practice. Springer, 2014.

[2] Salerno SM, Alguire PC, Waxman HS. Competency in interpretation of 12-lead electrocardiograms: a summary and appraisal of published evidence. Annals of Internal Medicine 2003;138(9):751–760.

[3] Estes III NM. Computerized interpretation of ecgs: supplement not a substitute, 2013.

[4] Ribeiro AH, Ribeiro MH, Paixão GM, Oliveira DM, Gomes PR, Canazart JA, Ferreira MP, Andersson CR, Macfarlane PW, Wagner Jr M, et al. Automatic diagnosis of the 12-lead ecg using a deep neural network. Nature communications 2020;11(1):1–9.

[5] Alday E, Gu A, Shah A, Robichaux C, Wong A, An-Kwok I, Liu C, Liu F, Rad A, Elola A, Seyedi S, , Li Q, Sharma A, Clifford G, Reyna M. Classification of 12-lead ecgs: the physionet/computing in cardiology challenge 2020. medRxiv 2020;.

[6] 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.

[7] He K, Zhang X, Ren S, Sun J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In Proceedings of the IEEE international conference on computer vision. 2015; 1026–1034.

[8] Hannun AY, Rajpurkar P, Haghpanahi M, Tison GH, Bourn C, Turakhia MP, Ng AY. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. Nature medicine 2019; 25(1):65.

[9] Kang B, Xie S, Rohrbach M, Yan Z, Gordo A, Feng J, Kalantidis Y. Decoupling representation and classifier for long-tailed recognition. arXiv preprint arXiv191009217 2019;.

[10] Gee AH, Garcia-Olano D, Ghosh J, Paydarfar D. Explaining deep classification of time-series data with learned prototypes. arXiv preprint arXiv190408935 2019;.

[11] Yao Q, Wang R, Fan X, Liu J, Li Y. Multi-class arrhythmia detection from 12-lead varied-length ecg using attention-based time-incremental convolutional neural network. Information Fusion 2020;53:174–182.

Address for correspondence:

Zhibin Zhao
zhibinzhao1993@gmail.com
Institute of Aero-Engine
The School of Mechanical Engineering
Xi'an Jiaotong University