Deep Residual Learning for Atrial Fibrillation Detection

Li Wei , Atif Khurshid

Abstract

A deep residual learning model for classification of electrocardiogram recordings is proposed to detect atrial fibrillation. The linear ECG recording is converted into a 2-dimensional spectrogram with 3-fold scaling, and used as input for a 25-layer deep residual neural network. The neural network is evaluated on the dataset from PhysioNet/CinC Challenge 2017 and achieves a weighted F-measure of 86%.

1 Introduction

Atrial fibrillation (AF) is a condition where disorganized electrical signals shake the upper two chambers of the heart (the atria) which causes rapid and irregular beats in the lower chamber.[1]

AF was first discovered in the seventeenth century by William Harvey, who described "unusual chaotic movements of the right atrium" in dying animals.[9] In 1749, John Baptist Senac published the first descriptions of human patients suffering from irregular heartbeat.[9] In 1909, Sir Thomas Lewis was the first to use electrocardiogram (ECG) to conclude that AF was the usual cause of arrhythmia.[9] Soon after, AF became the most prevalent rhythm disorder, especially among the elderly population.[10]

Today, AF is the most common arrhythmia affecting patients and the number of people suffering from the disease is increasing at an alarming rate throughout the world. In Europe only, 14 million people are expected to suffer from AF by 2030. [8] This situation is particularly concerning because AF is associated with severe consequences, including "heart failure, syncope, dementia, and stroke".[10] Therefore, it is imperative that we detect AF in patients before they suffer a stroke.

Currently, there are several studies that utilize different neural networks for AF classification. However, many of these studies cannot be generalized because they are evaluated using small or cherry-picked dataset. Other studies employ convolution or recurrent neural networks that suffer from a degradation problem where accuracy saturates and then degrades when network depth is increased.[5] Therefore, we implement a deep residual network to classify AF signals from ECG input.

It has been shown that deep residual nets (ResNet) maintain accuracy gains from greatly increased depth and produce results substantially better than other deep neural networks.[5] Taking inspiration from the preprocessing method used in [7], we convert the one-dimensional ECG signals into a two-dimensional spectrogram. We then normalize and scale the spectrogram to facilitate feature extraction by the ResNet which classifies the input into one of four classes. The performance of the ResNet is evaluated on a reserved validation set using weighted F-Measure.

2 Related Work

Deep neural networks have achieved great success in classification tasks in a wide range of applications such as image, video and speech recognition[2, 13] Convolutional neural networks (CNN) were until recently, the best architecture to work with images due to their ability to extract local and partial features. However, after the rapid development of DNN model structure, Residual Networks [5] are the new state-of-the-art because they solve the vanishing gradient problem in naïve CNN.

ECG classification problem has has been studied using many machine learning methods. W. K. LEI et al. used fuzzy set to tackle the problem where an element can partially belong to the multiple sets with the provision that the sum of membership values equals one. [12] Matthew C. Wiggins et al. used Rough Set Theory(RST), a relatively new data-mining technique, to discover patterns within data. [4] K. Polat et al. used SVM and achieved a high accuracy according to its author.[6]

However, due to its improved performance, many medical studies are now using DNN rather than traditional machine learning methods. For example, [3] used CNN for histopathological image classification and achieved better accuracy than other machine learning methods such as SVM. E. Derya et al. used Recurrent Neural Network (RNN) and Multi Layer Perceptron Neural Network (MPLNN) for ECG classification. [11] Martin et al. used the same dataset as in our research, fitted the data into Convolutional Recurrent Neural Network (CRNN) and gained an average F-Measure of 82.1%.[7]

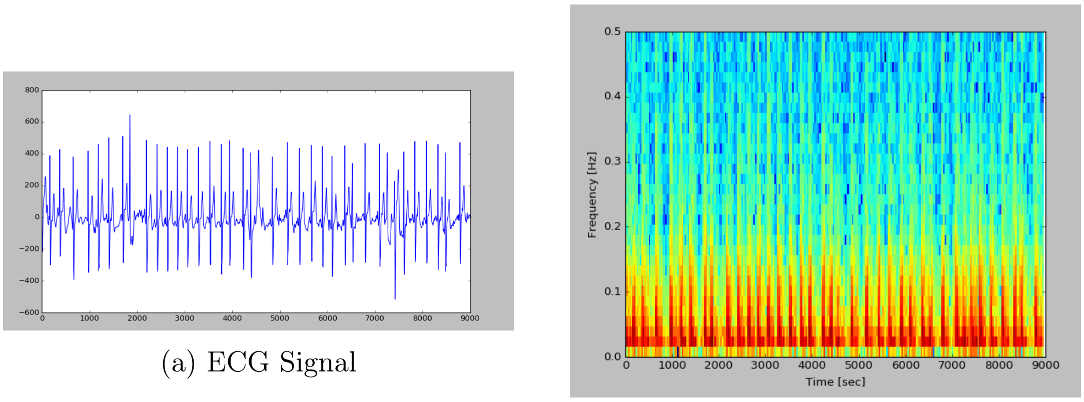
3 Method

3.1 Preprocessing

The raw format of ECG is not suitable for a CNN because it is one-dimensional. Therefore, we need to preprocess the data to transform it into a 2-dimensional image. The conversion method is inspired by [7]. We first compute the one-sided spectrogram of the time-domain input ECG signal and apply a logarithmic transform. The spectrogram is computed using a Tukey window of length 64 (corresponding to 213ms at the 300Hz sampling rate of the input data and resulting in 33 effective frequency bins) with shape parameter 0.25 and 50% overlap. Figure 1 shows the results of preprocessing.

3.2 Model Architecture

We use a Deep Residual Network in this project and the architecture of a basic residual block is shown in Figure 2. After preprocessing, the data is fitted into 25 residual blocks and the basic block structure is in Figure 2. Residual blocks are divided into layers labeled Conv1 to Conv5 where Conv1 is the first layer and Conv5 is the last layer. The output and kernal size of each layer is shown in Table 2. These



(a) ECG Signal

(b) One-sided spectrogram of ECG signal

Figure 1: A visual representation of the preprocessing

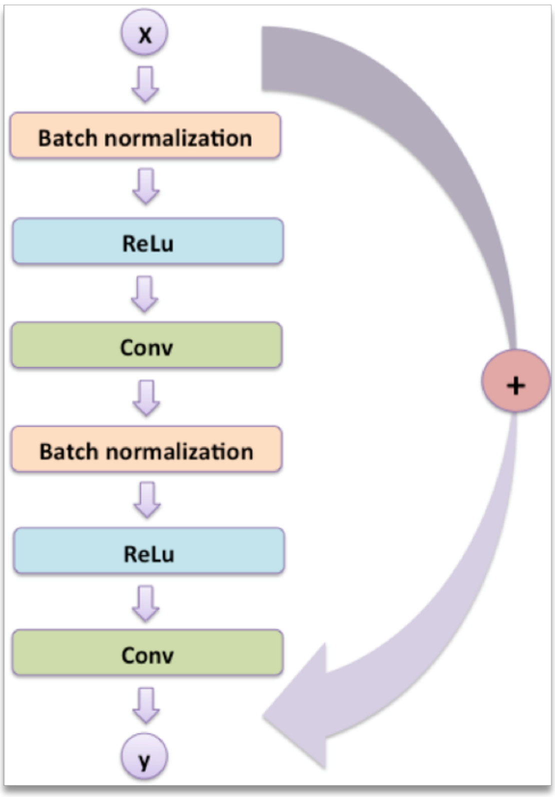


Figure 2: Basic structure of a Residual Block

Layer	**Output Size**	**Kernel Size**
Log Spectrogram	t/32 x 33	-
Conv1	t/64 x 17	3 x 3, 16
Conv2	t/128 x 9	3 x 3, 32
Conv3	t/256 x 5	3 x 3, 64
Conv4	t/512 x 3	3 x 3, 128
Conv5	t/1024 x 1	3 x 3, 256
FC	4 x 1	-

Table 1: Output and Kernal sizes of each layer in ResNet

final layer is a fully connected layer which calculates the probabilities for each of the four classes.

4 Experiments and Results

Our experimentation consists of two phases: model testing and parameter optimization.

In the first phase, we test four different ResNet models with varying number of layers. We begin with a 10-layer network and add five layers to each subsequent model. Our deepest model is limited to 25-layers because of GPU memory limitations on the project servers. We train each model for 30,000 epochs and monitor its training and validation accuracy using Tensorboard.

In the second phase, we train the chosen model only until it converges to avoid over-fitting. Furthermore, we reduce learning rate at runtime, scale the spectrogram and tweak other hyper-parameters to improve the performance of the model. We evaluate each model and choose the most appropriate model according to the criteria listed in the next section.

4.1 Evaluation

We evaluate our models on PhysioNet/CinC Challenge 2017 data set which contains 8528 ECG recordings of length ranging from 9 to 61sec, sampled at 300Hz.[7] Each recording contains a label of one of the classes Normal rhythm(N), AF rhythm(A), Other rhythm(O) and Noisy signal(~). The dataset contains 61% Normal, 7% AF, 30% Other and 2% Noisy signals. We evaluate the performance of our models using F-Measure because the dataset is skewed, and hence the precision and recall of each class are a more appropriate measure of performance than the overall accuracy of the model. We focus on three key F-values to appraise our models:

1. Average F-Measure of N, A and O classes

2. Weighted F-Measure

3. F-Measure of class A

The Weighted F-Measure provides the most appropriate evaluation of each model while the average F-Measure is used to compare the model to [7]. Finally, the F-Measure of class A is very important because it represents the sensitivity of the model to AF rhythm signals.

F-Measure of a class C is calculated from a convolution matrix using the formula given below. We reserve 1000 records (12% of the data set) as validation records to generate a convolution matrix from the trained model. We repeat the process five times and consider the average of the five as the final F-Measure of the class.

$$F(C) = \frac{2 \times TP(C)}{2 \times TP(C) + FN(C) + FP(C)}$$

where

$$\begin{aligned} TP &= \text{Number of inputs labeled C correctly classified as C} \\ TN &= \text{Number of inputs labeled NOT C correctly classified as NOT C} \\ FP &= \text{Number of inputs labeled NOT C incorrectly classified as C} \\ FN &= \text{Number of inputs labeled C incorrectly classified as NOT C} \end{aligned}$$

We calculate Average F-Measure F_{avg} as :

$$F_{\text{avg}} = \frac{F(N) + F(A) + F(O)}{3}$$

We calculate the Weighted F-Measure F_{w} as :

$$F_{\text{w}} = n \times F(N) + a \times F(A) + o \times F(O) + \sim \times F(\sim)$$

where

$$\begin{aligned} n &= \text{fraction of inputs labeled N} = 0.61 \\ a &= \text{fraction of inputs labeled A} = 0.07 \\ o &= \text{fraction of inputs labeled O} = 0.30 \\ \sim &= \text{fraction of inputs labeled} \sim = 0.02 \end{aligned}$$

4.2 Results

4.2.1 Model Comparison

In the first phase of experiments, we find that almost all models converge in the first 5000 epochs and that their top-1 error decreases with increasing number of layers, as shown in Figure 3. We, therefore, choose the 25-layer ResNet for the next phase because it has the highest top-1 validation accuracy.

In the second phase, we experiment with scaling and tweak some hyperparameters, e.g. learning rate, to maximize the performance of the model. We find that scaling the spectrogram increases the classification performance, as shown in Figure 4, because it amplifies the subtle difference in the features of spectrograms from different classes. However, excessive scaling substantially decreases the performance and increases the convergence time of the model.

It is clear from the results in Figure 4 and Table 2 that the model with 3-fold scaling has a lower Top-1 error and higher F-Measure for every class. Moreover, 3-fold scaling increases the sensitivity of the model to classes A and ~, despite the small number of examples in the dataset with these labels. Thus we propose a 25-layer ResNet with 3-fold scaling of ECG spectrogram as our final model.

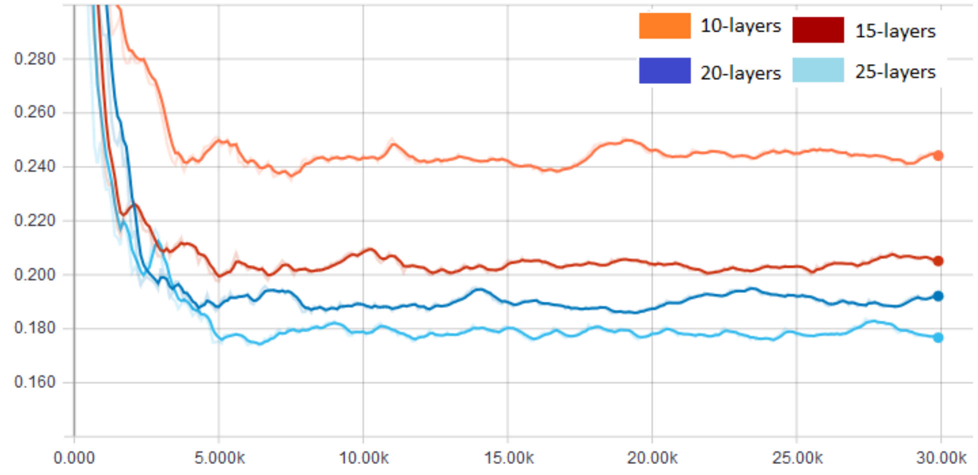


Figure 3: Top-1 validation error comparison of the four models.

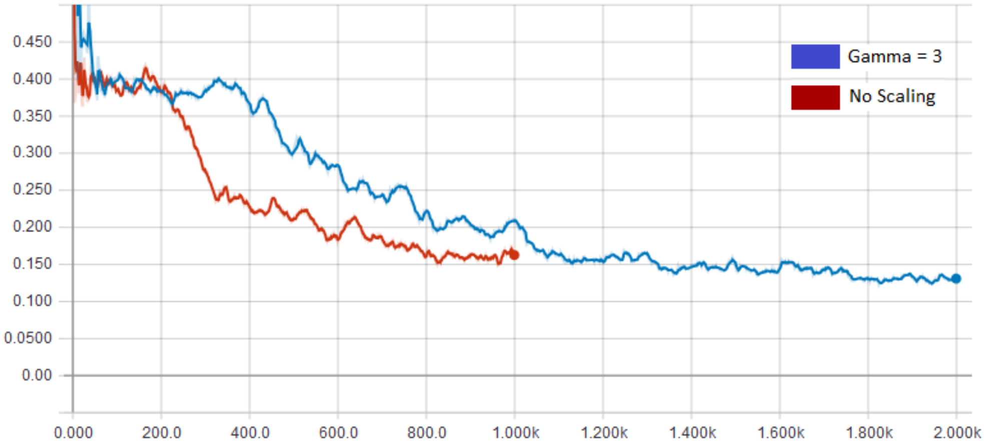


Figure 4: Top-1 validation error comparison of a 25-layer ResNet with scaled and unscaled input.

	F(N)	F(A)	F(O)	F(~)	F_{avg}	F_w
Unscaled	91.2	69.2	75.1	57.2	78.5	84.2
Scaled 3-fold	91.5	77.4	78.1	66.6	82.3	86.0

Table 2: F-Measure (%) comparison of model with scaled and unscaled input.

	DA	**F(N)**	**F(A)**	**F(O)**	**F(~)**	**F_{avg}**	**F_w**
ResNet	N	91.5	77.4	78.1	66.6	82.3	86.0
CNN	Y	87.8	79.0	70.1	65.3	79.0	-
	N	88.3	69.9	69.1	59.6	75.8	-
CRNN	Y	88.8	76.4	72.6	64.5	79.2	-
	N	87.4	69.9	66.5	54.9	74.6	-

Table 3: F-Measure (%) Comparison of ResNet and other architectures proposed in [7]. DA = Data Augmentation

4.2.2 Comparison with other architectures

Although our preprocessing is similar to [7], the ResNet performs better than the traditional convolution neural network (CNN) and the convolution-recurrent neural network hybrid (CRNN) proposed in [7], without using any data augmentation. It can be seen from Table 3 that the ResNet has the best F_{avg} score and tops almost every other category. However, it has a slightly worse F(A) score than the CNN using data augmentation because of the small number of examples (6%) labeled A.

We note that the CRNN achieved an F_{avg} score of 82.1% on a testing dataset in the PhysioNet/CinC Challenge 2017 (Competition best = 83%). However, since we do not have access to this dataset, it is difficult to compare our results to those in the competition.

5 Conclusion

Our initial approach required the implementation of an LSTM model in order to exploit its ability to remember long sequences, but our model was unable to extract useful features from the ECG signal. However, as demonstrated in several studies including [7], we realized that a pictorial representation of the signal would be a better input for classification. In fact, [7] has been particularly useful in our project, not only for its preprocessing method but also because it provides a baseline for the assessment of our model.

Despite good results, however, we believe that this project has great potential for improvement. Even though we could not increase the number of layers in ResNet, because we were constrained by hardware limitations, we believe that a deeper neural network can achieve improved classification performance. Discounting hardware limitations, we think that there are several other routes for development in this project, which were unfortunately abandoned due to the shortage of time. The preprocessing method, especially the spectrogram window function can be refined to facilitate feature extraction. Data augmentation is also of great interest because, as demonstrated in [7], it can increase the classification performance of the model.

References

[1] *Atrial Fibrillation: Prevention, Treatment and Research.* www.hopkinsmedicine.org/health/healthy_heart/diseases_and_conditions/atrial-fibrillation-prevention-treatment-and-research.

[2] G. E. H. Alec Krizhevsky, Ilya Sutskever, *Imagenet Classification with Deep Convolutional Neural Networks.*

[3] C. P. L. H. Fabio Spanhol, Luiz Oliveira, *Breast Cancer Histopathological Image Classification using Convolutional Neural Network.*

[4] E. U. I Güler, *ECG Beat Classifier Designed by Combined Neural Network Model.*

[5] S. R. J. S. Kaiming He, Xiangyu Zhang, *Deep Residual Learning for Image Recognition.*

[6] G. S. Kemal Polat, *Classification of Epileptiform EEG using a Hybrid System Based on Decision Tree Classifier and Fast Fourier Transform.*

[7] M. T. Martin Zihlmann, Dmytro Perekrestenko, *Convolutional Recurrent Neural Networks for Electrocardiogram Classification.*

[8] T. C. S. D. Massimo Zoni-Berisso, Fabrizio Lercari, *Epidemiology of Atrial Fibrillation: European Perspective*, Clinical Epidemiology, 6 (2014), pp. 213–220.

[9] J. McMichael, *History of Atrial Fibrillation 1628-1819 Harvey - de Senac - Laënnec.*, British Heart Journal, 48 (1982), pp. 193–197.

[10] L.-Q. W. W. K. S. Munger, M Thomas, *Atrial Fibrillation*, Journal of Biomedical Research, 28 (2017), pp. 1–17.

[11] B. C. S Mitra, M Mitra, *An Approach to a Rough Set Based Disease Inference Engine for ECG Classification.*

[12] M. C. D. M. I. V. Wei Kei Lei, Bing Nan Li, *AFC-ECG: An Adaptive Fuzzy ECG Classifier.*

[13] Y. B. P. H. Yann LeCun, Leon Bottou, *Gradient-Based Learning Applied to Document Recognition.*