ECG Signal Classification with Deep Learning Techniques

Chien You Huang, B04901147
Ruey Lin Jahn, B02901043
Sung-wei Huang, B04901093

Department of Electrical Engineering,
National Taiwan University,
Taipei, Taiwan (R.O.C.)
June 22, 2018

Abstract. An automatic arrhythmia detection system was made by applying Long Short Term Memory (LSTM) networks, which can classify Normal/Abnormal heartbeats with 95.81% accuracy. We applied the same networks to multi-class classification problem and the result shows 94% accuracy. No preprocessing is needed except segmentation of ECG signals.

Keywords: ECG, LSTM, Classification, Disease Diagnosis Automation

1 Introduction

1.1 Motivation

ECG is a diagnostic tool that is routinely used to monitor the electrical activity of the heart. In practice, ECG recording is examined by cardiologists, which is labor-intensive and time-consuming. With the advance of machine learning techniques and burgeoning data, we can "teach" machines to recognize irregular pattern. In countries with an acute shortage of certified cardiologists and limited resources like Sudan [1], an automatic arrhythmia detection program can empower the medical practitioners and alleviate the loading of cardiologists.

In the past 10 years, statistical learning techniques has been applied to deal with the heartbeat classification problem. However, to generate meaningful features from the ECG waveform [2] and feed into the classifiers, features extraction is inexorable, which usually needs careful treatment and advanced signal processing techniques such as wavelet transform. That is, it requires a lot of works to prepare "well-organized data" before you really start to teach a computer. We attempted to follow the methods proposed in the papers but got annoyed during data preparation. Hence, we sought a solution from the buzzed "Deep Neural Networks": let the machine learn the parameters that give it the ability to differentiate normal waveform from the others. We do as less preprocessing as possible.
1.2 Recurrent Neural Networks & Long Short Term Memory Networks

**RNNs** All recurrent neural networks (RNNs) have the form of a chain of repeating modules of neural network. This chain-like nature reveals that RNNs are intimately related to sequences and lists. They’re the natural architecture of neural network to use for sequence prediction since they can use its reasoning about “previous events”.

**LSTMs** Long Short Term Memory networks (LSTMs), introduced by Hochreiter & Schmidhuber (1997) [3], are a special kind of RNNs. LSTMs also have chain-like structure, but the repeating module has a different structure. Instead of having a single neural network layer, there are four, interacting in a very special way Fig.1. This gives it the capability of learning long-term dependencies. ECG signals consist of a sequence data points representing measured voltage and thus it’s reasonable to adopt LSTM model to do the classification task. LSTM has also been used for stock index movement prediction, which again is a sequence prediction problem.

![Diagram](image)

Fig. 1: A. An unrolled recurrent neural network. B. The repeating module in an LSTM contains four interacting layers, including a forget gate layer, a input gate layer, a tanh layer and a sigmoid gate.
Stacked LSTMs

Stacked LSTMs or Deep LSTMs were introduced by Graves, et al [6] in their application of LSTMs to speech recognition, beating a benchmark on a challenging standard problem. In the same work, they found that the depth of the network was more important than the number of memory cells in a given layer to model skill.

RNNs are inherently deep in time, since their hidden state is a function of all previous hidden states. The question that inspired this paper was whether RNNs could also benefit from depth in space; that is from stacking multiple recurrent hidden layers on top of each other, just as feedforward layers are stacked in conventional deep networks.

– Speech Recognition With Deep Recurrent Neural Networks, 2013

Stacked LSTMs are now a stable technique for challenging sequence prediction problems. A Stacked LSTM architecture can be defined as an LSTM model comprised of multiple LSTM layers. An LSTM layer above provides a sequence output rather than a single value output to the LSTM layer below. Specifically, one output per input time step, rather than one output time step for all input time steps.

We realized that it is the depth of neural networks that is generally attributed to the success of the approach on a wide range of challenging prediction problems. The original LSTM model is comprised of a single hidden LSTM layer followed by a standard feed-forward output layer. To achieve greater prediction power, we stacked two LSTM hidden layers to make the model deeper, more accurately earning the description of the data.

2 Model

2.1 Problem Formulation

The ECG arrhythmia detection task is a sequence-to-label task which takes a "segment" of ECG signal X=[x₁,...,xₖ] as input, and outputs a corresponding label r∈(1,2,...,m) since r can take on one of m different rhythm classes. Each output label corresponds to a segment of the input. Together the output labels [r₁,r₂,...,rₙ] cover the full ECG recording. For a single example in the training set, we optimize the cross-entropy objective function[7]:

$$L(X,r) = \sum_{i=1}^{n} \log(p)(R = r_i | X)$$

(1)

where p(.) is the probability the network assigns to the i-th output taking on the value rᵢ.

2.2 Model Architecture and Training

As described in the introduction, we use a two-layer Stacked LSTM network for the sequence-to-label learning task. The high-level architecture of the network
is shown in Fig 2. Our model takes two-channels variable length R-R beats as input, which consist of 300 - 750 data points. We set a parameter “Max sample points” to make our model capable of dealing with variable length input. In Keras API, we can use a function called “masking” to hide those zero values to avoid passing zeros to our model. From the database, for each 30 mins ECG record, we have two-channel measures. One must be Lead-2 and the other might be V1/V2/V5. We did not consider the heterogeneity in the second channel, simply feeding to our model (see discussion). We design two types of models, one is Binary Classification Model while the other is Multiple Classification Model. The first model only differentiate normal beat from abnormal beat. The second model consider normal beat, 4 arrhythmia classes, plus a class collecting the rest of arrhythmia classes.

3 Data

3.1 Training

We collected a dataset of 48 ECG recordings from PhysioNet [8]. The recordings were digitized at 360 samples per second per channel with 11-bit resolution over a 10 mV range. The records were annotated by two or more cardiologists independently; disagreements were resolved to obtain the computer-readable reference annotations for each beat included with the database.
In order to make training more convenient and match the annotations, each ECG record is split by RR interval by a simple algorithm. The peak of R wave can be found by simply finding the maximal voltage exceeding the given threshold determined by the maximum voltage in the record.

The processed data were randomly split into training data and testing data; the ratio between them is about two to one (75685:32520). Furthermore, the testing data were split into training and validation set. The training set contains 90% of the training data. There is no beat overlap between the two sets; however, there may be patient overlap between them.

3.2 Testing

As mentioned above, the testing data were randomly chosen from the 108205 beats split from the original data. In testing stage, for each single ECG beat, the model generates a vector representing the probability for each label. The predicted label is determined by the label with the highest probability.

3.3 Rhythm Classes

While there are over 50 labels mentioned on the source website, the total number of distinct labels used here is about 30. The binary model will only identify a beat is normal or not with all types of arrhythmia merged into single “abnormal class” while the multi-class model classify 5 classes - 1 for normal, 4 for specific types of arrhythmia and 1 for the remaining abnormal beat. The 5 types of arrhythmia were chosen according the amount of data - the more data, the more accurate. Table 1 (see Appendix) shows the types of arrhythmia common in the ECG dataset.

4 Results

In this section, we show the results of our experiment. First we will demonstrate the statistics of our binary classification model, including the loss and accuracy in the training-validating stage and testing stage. Similarly, we show the statistics for multiple classification model. Finally, we use our model to predict the our own heartbeats, which was measured in Experiment.1 early this semester.

4.1 Binary Classification

The basic model to classify ECG is, obviously, to separate heartbeats into two types, one normal and the other abnormal.

The training-validating statistics are shown in Fig.3. Fig.3 consists of two subfigures, one shows the training loss while the other displays the training accuracy. Obviously, higher the accuracy is, better the model is. The loss is calculated on training and validation and its interpretation is how well the model
is doing for these two sets. However, unlike accuracy, loss is not a percentage. In fact, the less the loss is, the better the model is, since loss is a summation of the errors made for each example in training or validation sets.

In Fig.3, as the number of epoch increases, the training loss decreases and the accuracy increases both significantly. The more important results will be the accuracy. In the past papers we have reviewed, the accuracy is about 85% to 91%. As we expected, our model outperforms their accuracy. Our model achieved high accuracy and low loss with several training epoches. As can be seen in the right in Fig.3, the accuracy will be saturated at an accuracy about 95%.

### 4.2 Multiple Classification

Although the binary classification model has achieved satisfactory results, more improvements can still be made. The annotation in the data is not divided into two types (normal and abnormal). Instead, there are about 30 kinds of annotation used in the database. Ideally we can make the machine to learn the features of all 30 types of arrhythmia. However, we do not have sufficient data for some types of arrhythmia. We have made a statistics on the total number of each kind of annotation, and find there are only four kinds of arrhythmia that have large enough samples. The arrhythmia we chose are Left bundle branch block beat (LB), Right bundle branch block beat (RB), Premature ventricular contraction (PV), Paced beat (PB) and A, where A collects all the other arrhythmia types.

Similarly, we demonstrate the training loss and training accuracy in Fig.4. The loss continuously decreases for each epoch, while the accuracy keeps increasing. It is intuitive to expect that the multiple classification will have less accuracy than the binary model. Actually, in the former papers, the accuracy of their work is mainly about 80%, which is not satisfactory enough. Comparatively, in our model the accuracy saturates at 94%, which is almost the same with the accuracy in the binary model case.
It should also be noticed that the method we used only deal with the one-dimensional data, so the training time is quite short. Using only about 30% of the 1080-Ti GPU, each epoch takes 20 minutes only. The whole training can be done within several hours, which is quite efficiency comparing to the seven-year training for a doctor in Taiwan.

Fig. 5 shows the confusion matrix of the testing results of the multiple classification model. Obviously, if most of the data are predicted correctly, the darker colour will appear in the diagonal. We can see from Fig. 5 that our model behaves well since almost all the dark colour appear in the diagonal elements.

However, it should be noted that the lower left corner of the confusion matrix shows slightly higher rate of misclassification. It indicates that there are quite a number of data, which are abnormal in reality, but are predicted to be normal in this model. That is, a high false negative rate was observed. The false negative result is usually worse than the false positive, since the patient with heart disease may be considered normal and will be given no medical care. This problem remains crucial in this model if it is going to be implemented in the reality medical use. The explanation of this phenomenon will be discussed in the next section.

4.3 Demo

The most interesting work we have done is that we use the data from the Experiment.1, which was measured by BIOPAC in the laboratory, as the input of this model. We expect the results to be normal, since the person being tested should have no heart disease. We should mention here that the original data we get from the BIOPAC is sampled in the frequency of 1000 Hz, however, the input to the model we trained should be in the frequency of 360 Hz. Thus, upsampling and downsampling actions should be taken first by Matlab. Since in Matlab only integer number can be used in upsampling and downsampling. We then upsample the data to the 36000 Hz, following the downsampling of data to 360 Hz.
Fig. 5: Multiple classification model: Confusion matrix

Fig. 6: The prediction probability of being normal/abnormal. Left: Result of the binary classification model. Right: Result of the multiple classification model.
The main result is shown in Fig. 6. The left part of this figure is the result of the binary classification model, while the right part refers to results of the multiple classification model. The numbers in the arrays represent the probability the model predicted for each class. In the binary classification result, the first number in the array represents the normal probability, while the second number represents the abnormal probability. In the multiple classification result, the numbers represent the probability of normal. LB, RB, PV, PB and A separately.

In binary classification model and multiple classification model, most of the data (30/31) is predicted to be normal. However, the 26th segment is predicted to be abnormal in binary classification model, and is predicted to be Right bundle branch block beat (indicated by blue box in Fig. 6). We dug out the original data and plot the waveform. The plot is shown in Fig. 7a. We compared this plot with the standard Lead-2 Right bundle branch block beat (Fig. 7b) and found that the machine gives a reasonable prediction. The problematic waveform results from measurement error in Experiment 1 and thus our model still holds.

5 Discussion

This section will discuss some technical issues and explain possible reasons behind the high false negative rate in the abnormal class.

Technical Issues Although we claimed that no preprocessing is needed, segmentation of the raw data was applied to generate R-R segments that are used as input in our model. This method is simple but it may limit the generalizability of our model. It is conceivable that our model can only recognize the waveform like that in Fig. 7b, with steep slopes on the left and right. Now if a P-P segment
is given as input, the likelihood of misclassification would be high. That is, our deep learning model must run hand-in-hand with a consistent segmentation program (in our case R-R segmentation) that is reliable and fast to achieve real-time prediction. During the demo session, teaching assistant comments that our naive R-R segmentation may not be the best practice, since the form of QRS complex is an important feature for some arrhythmia. A better practice is extending the two cutting boundaries a bit, to cover the full feature of QRS complex.

Currently our model can only take ECG signals digitized at 360 samples per second. If a ECG signal of other sampling rate is given, up-sampling or down-sampling is a must before feeding into our predictive model.

As for the design of deep neural networks, 2-layer stacked LSTM network is our first trial. More sophisticated layers can be added to test the accuracy. Some papers suggests using deep convolution neural networks [7], which also gives excellent results. It’s also possible to take a "ensemble approach" that combines two or more isolated neural networks to produce greater predictive power.

**Results** First, we would like to explore the possible reasons behind the high false negative rate in the abnormal class. Abnormal class, as stated before, is a collection of all arrhythmia with less number of beats. This class is also heterogeneous, which presumably makes machine harder to capture the features. High false negative rate may be due to the fact that our model "memorized" the small number of training instances. During the testing stage, our model cannot recognize new types of arrhythmic beats, therefore classifying them as normal beats.

High false negative rate is a critical problem. Several remedies can be taken. For example, if we have complete training dataset covering all types arrhythmia and sufficient arrhythmic beats for each type, we can expand our classification classes to elevate prediction performance.

Another question was raised by our classmate. He remarked that our high accuracy is not surprising since our data is measured from small number of patients. We acknowledged this problem but we do not have sufficient data to do analysis, unlike Andrew Ng’s group [7] having resources to collect data from more than 29,163 patients.

Lastly, we take two channel ECG data as our input. One channel is fixed as lead-2 while the other channel depends on patient’s disease types. We did not differentiate the effects between different channels. To make our model more reasonable, we should tweak our input format, considering the channel type effect. This will be implemented if our research proceeds.

6 Conclusion

We have demonstrated that using deep learning techniques (LSTM networks) to conduct abnormal heartbeat prediction. Other than segmenting ECG recording into R-R beats, preprocessing is not needed, thus saving a lot of work. The
prediction accuracy (94%) surpasses or is at least comparable to certified cardiologist [7]. This shows the power of deep learning in dealing with challenging prediction task. We can envision that future medical institutes will incorporate deep learning elements into their daily routine.

Acknowledgments. Our classmate Guo-Liang Hung gave us tremendous help during Deep Learning model selection and assisted us to tackle with technical problems. Without him, we cannot produce meaningful results successfully.

References


Table 1. The common types arrhythmia in ECG dataset. Arrhythmia with label code L, R, V and / were specifically selected in multi-class model.