Classifying Heart Sound Recordings using Deep Convolutional Neural Networks and Mel-Frequency Cepstral Coefficients

Jonathan Rubin, Rui Abreu, Anurag Ganguli, Saigopal Nelaturi, Ion Matei, Kumar Sricharan

Palo Alto Research Center, California, United States

Abstract

We describe the development of an algorithm for the automatic classification of heart sound phonocardiogram waveforms as normal, abnormal or uncertain. Our approach consists of three major components: 1) Heart sound segmentation, 2) Transformation of one-dimensional waveforms into two-dimensional time-frequency heat map representations using Mel-frequency cepstral coefficients and 3) Classification of MFCC heat maps using deep convolutional neural networks. We applied the above approach to produce submissions for the 2016 PhysioNet Computing in Cardiology Challenge. We present results from the challenge, as well as describe in detail the resulting neural network architecture produced and design decisions made.

1. Introduction

The goal of the 2016 PhysioNet Computing in Cardiology Challenge was to accurately classify normal and abnormal heart sounds from phonocardiogram (PCG) waveforms. A particular aim was to identify from a single short recording whether a subject should be referred on for expert diagnosis. Accurate and robust algorithms were required that could deal with heart sounds that exhibit very poor signal quality.

The challenge training set consisted of 3,240 heart sound recordings, lasting from 5 seconds to just over 120 seconds. Recordings were collected from nine different locations on the body (including aortic area, pulmonic area, tricuspid area and mitral area, among others). Recordings from healthy subjects were labeled as normal. Recordings from subjects with a confirmed cardiac diagnosis were labeled as abnormal. Abnormal recordings were collected from patients who suffered from a variety of illnesses, including heart valve defects (mitral valve prolapse, mitral regurgitation, aortic stenosis, valvular surgery) and coronary artery disease. An in depth description of the challenge and the dataset is provided in [1], as well as a thorough review of the field of cardiac auscultation.

We present an algorithm that computes heat maps of the time-frequency distribution of signal energy and uses a deep convolutional neural network to automatically classify normal versus abnormal heart sound recordings. Logistic regression hidden semi-Markov model-based heart sound segmentation is first performed on the PCG waveform. Spectrograms (energy maps) consisting of 6 cepstral coefficients that capture Mel-frequencies varying over time are derived for overlapping sliding windows of three-second duration, beginning at the first heart sound, S_1. A deep convolutional neural network consisting of two alternating convolution and max pooling layers is trained to perform automatic feature extraction. A final multilayer perceptron, consisting of two fully connected layers, distinguishes between normal and abnormal spectrograms. Heart sound recordings of variable length are dealt with by computing an ensemble of logit scores for all overlapping three-second segments contained within a recording and maximizing the average scores computed for all classes.

2. Approach

Our approach consists of three major components:

1. Segmentation of the PCG signal into the fundamental heart sounds.

2. MFCC Transformation of the original PCG signal into a time-frequency representation of the distribution of signal energy.

3. Training and classification of MFCC heat maps using deep convolutional neural networks.

Each component is described in detail below.

2.1. Segmentation

We first segment each PCG waveform into the fundamental heart sounds (S_1, *Systole*, S_2 and *Diastole*) using Springer's segmentation algorithm [2]. Fig. 1. provides an illustration of the segmentation process and depicts how fundamental heart sound segmentation corresponds to peak alignment in the corresponding ECG signal. Springer's algorithm uses a logistic regression hidden semi-Markov model to predict the most likely sequence of states by incorporating information about expected heart

sound state durations. The Springer segmentation algorithm was provided by the challenge organizers and was used *as is*. For further details regarding the algorithm we refer the reader to [2].

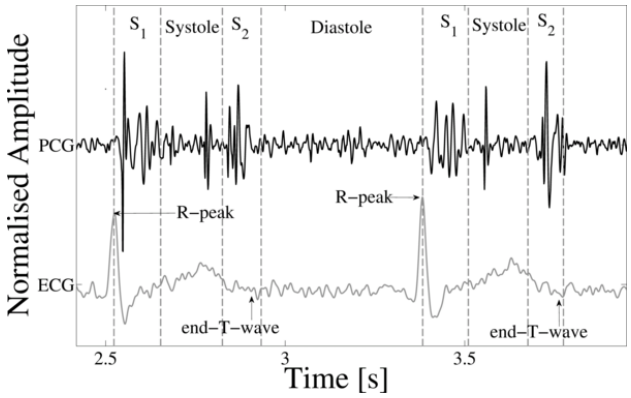

Figure 1. Illustration of fundamental heart sound segmentation in both PCG and ECG. Image source: [1].

Within our approach, we do not make use of all the resulting segmentation information. Rather, we chose to process and analyze 3-second heart sound segments. Segmentation was used to ensure that each 3-second heart sound segment began at S_1. This was performed to ensure sequences were aligned during classification. Overlapping sequences were used in eventual classification as this led to improved accuracies in our initial experimentation.

2.2. MFCC

After segmentation, each 3-second segment is transformed from a one-dimensional PCG audio signal into a two-dimensional time-frequency representation. We chose to represent the data using Mel Frequency Cepstral Coefficents [3]. MFCCs capture features from audio data that more closely resembles how human beings perceive loudness and pitch. MFCC are commonly used as a feature type in automatic speech recognition [4]. Previous approaches to cardiac auscultation have also utilized MFCCs [5].

To calculate MFCC values the following five steps are used:

1. Run overlapping sliding windows over the input audio data (we used a window length of 25ms and a step size of 10ms).
2. Compute the Fourier transform over each window
3. Apply a Mel filterbank and sum energies within each filter
4. Compute the logarithm of the filterbank energies
5. Perform a discrete cosine transform on the log filterbank energies

The above procedure produces 12 MFCC values per sliding window. The total energy per sliding window is also included as a feature. This results in 13 MFCC feature values for each sliding window. Appending these features together results in a time-frequency representation that can be visualized as a heat map (see Fig. 2). In total each heat map consists of 300 time frames represented on the x-axis, and 13 MFCC filterbanks represented on the y-axis.

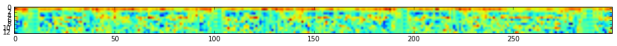

Figure 2. MFCC heat map visualization of a 3-second segment of heart sound data. Time is represented on the x-axis and filterbank frequencies represented on the y-axis. Energy information is represented by color in the spectrogram.

The number of MFCC features to use during classification was treated as a hyper-parameter. Based on performance during initial experimentation, the first 6 features were selected.

2.3. Deep Convolutional Neural Networks

The result of transforming the original one-dimensional time-series into a two-dimensional time-frequency representation is that now each 3-second instance of heart sound data can be processed as an image, where energy values over time can be visualized as a heat map. As such, we can make use of the latest advancements that have been made using deep convolutional neural networks (CNN) for image analysis.

2.3.1. Network Architecture

Fig. 3. depicts the network architecture of a convolutional neural network that accepts as input a single channel 6x300 MFCC heat map and outputs a binary classification, predicting whether the input segment represents a normal or abnormal heart sound. A standard architecture is used consisting of two convolutional layers, each followed by a max-pooling layer, followed by two fully connected layers before final classification.

The first convolutional layer learns 64 2x20 kernels, using same-padding. This is followed by applying a 1x20 max-pooling filter, using a horizontal stride of 5, which has the effect of reducing each of the 64 feature maps to a dimension of 6x60. A second convolutional layer applies 64 2x10 kernels over the previous layer, once again using same padding. This is again followed by a max-pooling operation using a filter size of 1x4 and a stride of 2, further reducing each feature map to a dimension of 6x30. At this stage in the architecture a flattening operation is applied that unrolls each of the 64 6x30 feature maps into a single dimensional vector of size 11,520. This feature vector

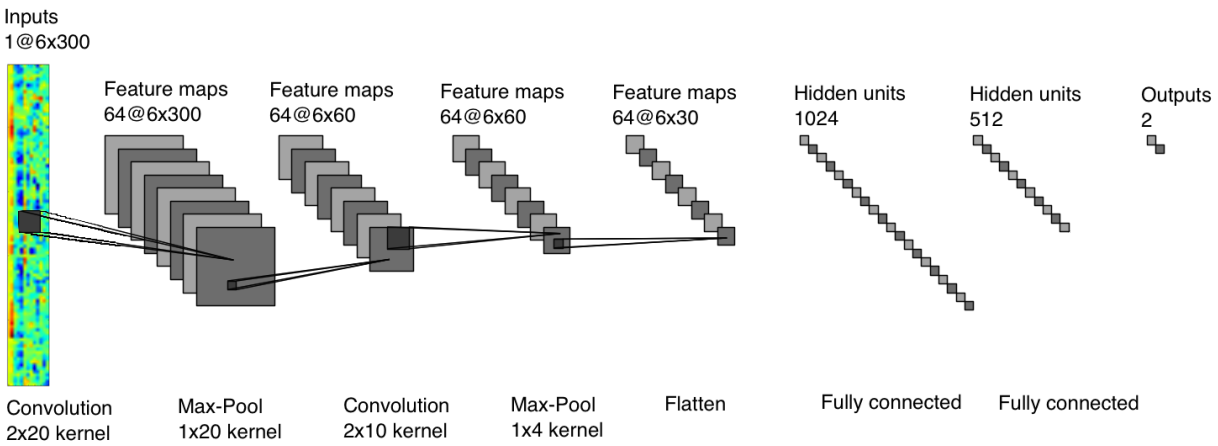

Figure 3. Convolutional neural network architecture for predicting normal versus abnormal heart sounds using MFCC heat maps as input. Note that the input heat map image is rotated due to space considerations.

is fed into a first fully connected layer consisting of 1024 hidden units, followed by a second layer of 512 hidden units and finally a binary classification output.

Decisions about the number of filters to apply and their sizes, as well as how many layers and their types to include in the network were made by a combination of initial manual exploration by the authors, followed by employing a *random search* over a limited range of network architecture parameters.

3. Network Training

A standard softmax cross entropy loss function was used to optimize the network during training. L_2 regularization was computed for each of the fully connected layers' weight and bias matrices and applied to the eventual loss function. Dropout was applied within both fully connected layers. Table 1 shows the values of hyper-parameters chosen by performing a *random search* through parameter space, as well as a list of other network training choices, including weight updates and use of regularization. Adam optimization [6] was used to perform weight updates. Models were trained on a single NVIDIA GPU with between 4 – 6 GB of memory. A mini-batch size of 256 was selected to satisfy the memory constraints of the GPU.

3.1. Train, validate and test data split

From the original 3240 PCG waveforms supplied by the challenge organizers (*training sets a – f*), 301 instances were removed (*validation set*) and the remaining instances were used to train initial models. Models were trained on the overlapping 3-second MFCC segments extracted from the remaining 2939 PCG waveforms. This resulted in approximately 90,000 MFCC heat maps, which were split into a training ($\sim$ 75,000 instances) and validation set ($\sim$ 15,000 instances). This training and validation set was unbalanced, consisting of approximately 80% normal segments and 20% abnormal segments. Training was performed on the unbalanced dataset and no attempt was made to compensate for this class imbalance (for example by stratifying the training dataset).

Hyper-parameters	**Value**
Learning rate	0.00015822
Beta	0.000076253698849
Dropout	0.85565561
Network parameters	**Value**
Regularization Type	L_2
Batch Size	256
Weight Update	Adam Optimization

Table 1. Listing of hyper-parameters and selected network parameters. Hyper-parameters were learned over the network architecture described in Section 2.3.1, using *random search* over a restricted parameter space.

3.2. Full instance classification

Given that each model was trained on 3-second MFCC heat map segments, it was necessary to *stitch* together a collection of predictions to classify a single full instance. The simple strategy of averaging each class's prediction probability was employed and the class with the greatest probability was selected as the final prediction.

The authors used the 301 instances, that were initially removed, as a *local held-out test-set* to evaluate a trained model's predictions on full instances, before making a sub-

mission to the PhysioNet challenge server. The 301 *local held-out test-set* was a balanced dataset, consisting of approximately 50% normal and 50% abnormal instances. Final model evaluation was performed on the challenge server using a completely separate unseen test set. Only a limited number of submissions to the challenge server were allowed to avoid overfitting.

4. Model submissions (Phase I & II)

Model performance was initially evaluated by the authors using the *local held-out test-set* described above before making a final submission to the PhysioNet challenge server. Models that improved the performance on the *local held-out test-set* were selected for submission to the challenge server. Before submitting a model for evaluation on the challenge server, retraining of the model occurred using the entire dataset consisting of 3240 PCG waveforms.

4.1. Phase I (Unofficial)

For Phase I submissions, a binary classification model (as described above) was submitted to the challenge server. Training-set f, was also not yet made publicly available, so the number of instances used to train the model was fewer.

4.2. Phase II (Official)

During Phase II, updates were made to the datasets provided by the challenge organizers. These updates included the addition of training-set f, as well as the introduction of a signal quality indicator. Challenge organizers also made available hand corrections to the output of Springer's heart sound segmentation algorithm for PCG waveforms where signal quality was considered good.

Models submitted during Phase II were updated and trained as multiclass prediction models, where instances with bad signal quality were given the class label of *uncertain*. Where signal quality was good, hand corrected segmentation was used to identify S_1 heart sounds, whereas if signal quality was bad, heart sound segmentation was the same as in Phase I.

5. Results

Results for our top scoring submissions made to the PhysioNet challenge server for both Phase I and Phase II are depicted in Table 2. Note that the scoring functions, used for evaluation, differed between challenge phases. In particular, the way that *uncertain* class predictions were evaluated was altered. We refer the reader to [1] for further details about the exact scoring mechanism used.

Phase I	
Sensitivity	75%
Specificity	100%
Overall	88%
Phase II	
Sensitivity	76.5%
Specificity	93.1%
Overall	84.8%

Table 2. PhysioNet challenge server sensitivity and specificity results for Phase I and Phase II of the 2016 Computing in Cardiology Challenge.

6. Conclusions

These results suggest that convolutional neural networks are able to automatically extract useful features from Mel-frequency cepstral coefficient heat maps to distinguish between normal and abnormal heart sounds from noisy data.

References

[1] Liu C, Springer D, Li Q, Moody B, Juan RA, Chorro FJ, Castells F, Roig JM, Silva I, Johnson AE, Syed Z, Schmidt SE, Papadaniil CD, Hadjileontiadis L, Naseri H, Moukadem A, Dieterlen A, Brandt C, Tang H, Samieinasab M, Samieinasab MR, Sameni R, Mark RG, Clifford GD. An open access database for the evaluation of heart sound algorithms. Physiological Measurement 2016;37(11).

[2] Springer DB, Tarassenko L, Clifford GD. Logistic regression-hsmm-based heart sound segmentation. IEEE Transactions on Biomedical Engineering 2016;63(4):822–832.

[3] Davis S, Mermelstein P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE transactions on acoustics speech and signal processing 1980;28(4):357–366.

[4] Godino-Llorente JI, Gomez-Vilda P. Automatic detection of voice impairments by means of short-term cepstral parameters and neural network based detectors. IEEE Transactions on Biomedical Engineering 2004;51(2):380–384.

[5] Wang P, Lim CS, Chauhan S, Foo JYA, Anantharaman V. Phonocardiographic signal analysis method using a modified hidden markov model. Annals of Biomedical Engineering 2007;35(3):367–374.

[6] Kingma DP, Ba J. Adam: A method for stochastic optimization. CoRR 2014;abs/1412.6980. URL `http://arxiv.org/abs/1412.6980`.

Address for correspondence:

Jonathan Rubin
2 Canal Park, Cambridge MA 02141, United States
jrubin01@gmail.com