Machine Learning Engineer Nanodegree Proposal

Domain Background

“According to the World Health Organisation, cardiovascular diseases (CVDs) are the number one cause of death globally: more people die annually from CVDs than from any other cause. An estimated 17.1 million people died from CVDs in 2004, representing 29% of all global deaths. Of these deaths, an estimated 7.2 million were due to coronary heart disease. Any method which can help to detect signs of heart disease could therefore have a significant impact on world health.”

A heartbeat is composed of two sounds “lubb-dupp” or also known as S1 and S2 sounds respectively. The first sound, S1, is caused by the closure of the inflow valves M1(mitralis) and T1(tricuspid) at the beginning of ventricular contraction, systole, when blood is pushed out to from the heart to the body and lungs. The second heart sound, S2, is caused by the closure of the valves A2 (aortic) and P2 (pulmonary) at the end of ventricular systole and the beginning of ventricular diastole. This phase is a bit longer since the heart is being refilled. With S1 and S2 sounds a heartbeat could be classified within the following categories: normal, murmur, extra heart sound.

Normal Category

A normal heart sound has a clear “lub dub, lub dub” pattern, with the time from “lub” to “dub” shorter than the time from “dub” to the next “lub”. This pattern can be described in the following illustration:

…lub……….dub……………. lub……….dub……………. lub……….dub…………….
lub……….dub…

Murmur Category

Heart murmurs sound as though there is a “whooshing, roaring, rumbling, or turbulent fluid” noise in one of two temporal locations: (1) between “lub” and “dub”, or (2) between “dub” and “lub”. They can be a symptom of many heart disorders, some serious. There will still be a “lub” and a “dub”. One of the things that confuses non-medically trained people is that murmurs happen between lub and dub or between dub and lub; not on lub and not on dub. Below, you can find an asterisk* at the locations a murmur may be:

…lub..*...dub……………. lub..*..dub ……………. lub..*..dub ……………. lub..*..dub … or
…lub……….dub…*….lub………. dub…*….lub ………. dub…**….lub ……….dub…

Extra Heart Sound Category

Extra heart sounds can be identified because there is an additional sound, e.g. a “lub-lub dub” or a “lub dub-dub”. An extra heart sound may not be a sign of disease. However, in some situations it is an important sign of disease, which if detected early could help a person. The extra heart sound is important to be able to detect as it cannot be detected by ultrasound very well. Below, note the temporal description of the extra heart sounds:

…lub.lub……….dub……….………. lub. lub……….dub…………….lub.lub……..…….dub……. or
…lub………. dub.dub………………….lub.……….dub.dub………………….lub……..…….dub. dub……

Several papers have been published reporting classification performance achieved on the PhysioNet 2016 Cardiology Challenge. This challenge consists in classifying heartbeats as either normal or abnormal. One paper, published by Palo Alto Research Center, investigated using 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. This paper reported having a sensitivity of 76.5% and and specificity of 93.1%.

In another paper, published by Marton Aron Goda and Peter Hajas, investigated using Support Vector Machines and audio feature extraction of time-domain features, frequency-domain features and wavelet envelope features. This paper reported having 83.77% sensitivity and 76.8% specificity.

Problem Statement

The problem at hand will consist in the following: given an audio file of heartbeats be able to classify such recording as either: normal, murmur, extra heart sound and artifact. The artifact category will be used when there is no discernable heart sounds, this is useful to indicate a final user to do another recording.

The audio files provided will be of different lengths and also the contents of such files will vary. This means that even if it’s possible to trim the files to have an equal size length is not feasible to compare an exact location with different files. In one file a particular time can be the start of a heartbeat and in another file the exact time could be noise. This indicates, that somehow, a method will be needed to represent audio in a different way.

Once the challenge of finding a way to represent audio in a different way is solved the challenge then becomes to identify patterns within the audio representations and generalize a solution to the classification problem at hand. All of this points towards a suitable application for deep learning.

As a final step, a measurement for the performance of the model is needed to know whether the approach taken is generating satisfying results.

Datasets and Inputs

Data will come from two sources. The first dataset comes from the PhysioNet 2016 Cardiology Challenge. It has phonocardiogram recordings of normal and abnormal recordings. The training set consists of five databases (A through E) containing a total of 3,126 heart sound recordings, lasting from 5 seconds to just over 120 seconds. The following provides the counts for each database with its corresponding labels:

- database A:
  - normal: 117
  - abnormal: 292
- database B:
  - normal: 386
  - abnormal: 104
- database C:
  - normal: 7
  - abnormal: 24
- database D:
  - normal: 27
  - abnormal: 28
- database E:
  - normal: 1,958
  - abnormal: 183
- Total
  - normal: 2,495
  - abnormal: 631

The second dataset comes from Kaggle and is split into two sources, A and B: A was collected from an iPhone app and B from a clinical trial in hospitals using a digital stethoscope. For the purpose of this project only dataset A will be used and not B. Dataset A contains audio files that are of varying lengths, between 1 second and 30 seconds. Dataset A contains recordings from healthy subjects and patients with different kinds of heart conditions: normal, artifact, murmur and extra heartbeat. There is a total of 176 audio files out of which only 124 files are labeled. Only the labeled audios will be used, segregating the categories would generate the following counts:

- Artifact: 40
- Murmur: 34
- Normal: 31
- Extra Heartbeat: 19

Solution Statement

As a sanity check is always a good idea to try a simple model before looking into more complex ones. The audio files provided from the PhysioNet 2016 Cardiology Challenge have a sample rate of 2000, this means a single second consists of 2000 values. Neural networks require one node per feature and the hidden layers will require a similar magnitude. These orders of magnitude are way to high so an approach has to be taken in order to reduce the dimensionality.

Doing some research an interesting approach was taken in order to solve this problem in this blog. The blog post indicates a way to do dimensionality reduction of audio using feature extraction methods provided by the Python library librosa. The feature extractions methods involved are:

- MFCC: representation of the short-term power spectrum of a sound
- Mel-frequency cepstral coefficients: Coefficients that collectively make up an mel-frequency cepstrum
- Chromagram of a short-time Fourier transform: Projects into bins representing the 12 distinct semitones of the musical octave
- Octave-based spectral contrast: Distribution of sound energy over octave frequencies
- Tonnetz: Estimates tonal centroids features

Combining these 5 feature extractions will result in a vector of 193 values. Is important to remark that this vector will always have the same length even if the audio files are of different length.

With the dimensionality reduction problem solved a simple densely connected network will be used to have the first set of results to differentiate between normal and abnormal heartbeats.

After having a benchmark model all audio files will be converted to spectrograms so they can be handle as images. With these images a CNN will be used to try to identify the patterns with the image.

Finally, once the first classifier has reasonable results, all the files from the dataset from Kaggle will be also converted to spectrograms. Transfer learning will be used so all the patterns learned from the first classifier are applied to a more specific classifier able to recognize the categories of
normal, murmur, extra heart sound and artifact recordings.

The reason for doing this is because the dataset from Kaggle is too small, so a bigger one will be used to mitigate this issue.

Benchmark Model

Before looking into a complicated and computationally expensive model such as a CNN for the classification of normal and abnormal heartbeats, first a simple 2 layer feed forward network will be used to establish a model to use as a benchmark. This is to make sure that any further complexity delivers additional value.

The baseline for the classifier of normal and abnormal heartbeats would be an accuracy of about 60%. This because the Physionet Cardiology Challenge 2016 provided a baseline score of 71%.

For the second classifier, a base benchmark is not present in the Kaggle description of the problem and in order to overcome this problem a common sense approach will be taken. From the 176 files present in the Kaggle dataset only 124 are labeled. The category which has the greater count is Artifact with a value of 40. A classifier that always predicts that an input file is an Artifact will be correct 32%(40/124) of the time. From this, it is possible to conclude that the learning approach should beat this value.

Evaluation Metric

Since the datasets provided have labeled data and are imbalanced other measurement besides accuracy are necessary. Therefore the other metrics to be used to evaluate the performance of the the model are: precision, recall and f1 score.

To calculate these metrics first it is necessary to know the confusion matrix. In this matrix each column represents the number of instances in a predicted class while each row represents the instances in an actual class. This is useful to determine where the system is confusing classes and also provides a way to know: true positives(TP), true negatives (TN), false positives (FP), false negatives(FN).

Precision is the ratio of correctly predicted positive observations to the total predicted positive observations.

$$Precision = TP/TP+FP$$

Recall is the ratio of correctly predicted positive observations to the all observations in actual class.

Recall = TP/TP+FN

F1 score is the weighted average of Precision and Recall. This score favors classifies that have similar precision and recall.

F1 Score = 2*(Recall * Precision) / (Recall + Precision)

Project Design

The project workflow will be as follows:

First Part

- Download dataset from the PhysioNet 2016 Cardiology Challenge
- Extract audio features
- Train a Feed Forward Network to classify between normal and abnormal heartbeats
- Convert audio files to spectrograms
- Train a CNN to classify between normal and abnormal heartbeats
- Evaluate results

Second Part

- Download dataset from Kaggle
- Convert audio files to spectrograms
- Use transfer learning to train a CNN with a classifier more specialized able to recognize normal, murmur, extra heart sound and artifact recordings
- Evaluate results

References

https://www.physionet.org/challenge/2016/
https://en.wikipedia.org/wiki/Heart_sounds
https://www.kaggle.com/kinguistics/heartbeat-sounds
http://www.peterjbentley.com/heartchallenge/

http://blog.exsilio.com/all/accuracy-precision-recall-f1-score-interpretation-of-performance-measures/
https://en.wikipedia.org/wiki/Spectrogram
https://physionet.org/challenge/2016/papers/
https://en.wikipedia.org/wiki/Mel_scale
https://en.wikipedia.org/wiki/Spectrogram
https://physionet.org/challenge/2016/papers/rubin.pdf
https://physionet.org/challenge/2016/papers/goda.pdf

http://blog.exsilio.com/all/accuracy-precision-recall-f1-score-interpretation-of-performance-measures/