Automatic Left Ventricle Segmentation in Ultrasound

Ezequiel Ortiz

2018

1 Abstract

2 Ackowledgements

3 Intro

3.1 Motivation

Cardiovascular disease is responsible for 17.9 million deaths, or 44% of all non-communicable deaths per anum [?]. Detection and monitoring of heart disease is tantamount to treatment and managment of symtoms.Quantification of the interior volume of the left ventricle and thickness of muscle is extremely important due to how illistrative these measurements are to the overall function of the ventricle. [?] The most prevalent method to assess these metrics is cardiac imaging. While there are many imaging modalities, ultrasound has been cemented as a cornerstone of cardiac imaging due to the noninvasive prcedure low coat and realtime image aquisition. Interpretation of ultrasound images is time intensive and can be challenging even for trained personel. The automatic segmentation of the left ventricle bloodpool would reduce the need for expert evaluation and inter-operator variance.More access to accurate data in clinical settings is always a good thing... why

3.2 Metrics

Quantification of the performance of the heart is crucial for diagnosis of cardiac pathogy. Due to the complexity of the heart and cardiovascular system as a whole the amount that can go worng is staggering. We want the metrics that we collect to be illistrative of how the heart is functioning, reproducable and easy to understand....

The two most common metrics that are collected during echocardiograms are left ventricular ejection fraction(LVEF) and left ventricular mass(LVM). [?] These quantifications in conjunction with the visual information of the echocardiogram are potent in the assesment of how effectivly the heart is pumping blood and the health of the myocardium

3.3 Ejection Fraction

LVEF is actually the combination of several ventricle volume measurements that are made at different points within the cardiac cycle. LVEF quantifies the effectivness of the heart as it is the ratio of end diastolic volume(EDV) to end systolic volume(ESV), indicating how much blood the heart is pumping per beat.

EF it is calculated according to eq 2. Eq. 1

$$SV = EDV - ESV$$

Eq. 2

$$EF = \frac{SV}{EDV} * 100$$

Pinning down what is normal LVEF is not trivial as it is very patient specific. Everyone is different and an LVEF measurement that is normal for one patient might be unhealthy for another. LVEF can also change depending on the physical state of the patient when the measurement is made. The patient's level of cardiovascular stress and wether they are supine or seated will all change the measured LVEF. M. E. PFISTERER et al set out to determine ranges of normal LVEF by assessing the mean LVEF within a population of 1200 patients using ultrasound. The lower limit of normal LVEF values for men has been estimated to be $51.1 \pm 4.2\%$ with the upper limit of normal being$76.6 \pm 3.8\%$. These values were assessed with echocardiogram while the patient was supine and at rest. LVEF was also assesed with the more accurate but invasive x-ray angiography; no signifigant differences were found. LVEF was for the most part independent of sex or age group, except for patients over 60 who had slightly elevated LVEF levels. Patients that were able to exercise then reached 85% of their maximal HR before their LVEF was measured. On average, Men had an LVEF increase of 10.5% while women only saw an increase of 5.3%. Young patient LVEF increased about 6% more than older patients [1]. These findings show that there is no bright line to determine what is normal or abnormal in LVEF.

The American Society of Echocardiogrphy has published generatl guidelines for LVEF classification. Reduced ¡= 40% borderline 41%-49% normal 50%-70&

Many variables affect LVEF, this obfuscates its relation to cardiac health.

LVEF is useful because it is easy and consistent to measure. It is another peice of data about the function of the heart that can be monitored over time. Changes in LVEF can be useful in diagnosing common heart ailments like hypertropic cardiomyopathy heart failure.

Athletes can often test for LVEF below normal even though their hearts are healthy and functional [2].

And LVEF can be within normal values while the function of the heart is impaired. HFpEF HFrEF LV systolic func LV diastolic func This is because EF only depends EDV and ESV and does not take into account the time duration or the geometry of the contraction. Measuring LVEF is usually measured noninvasivly with The vast majority of LVEF measurements are carried out by imaging the heart. [3]

3.4 Left Ventricular Mass

LVM is a measure of how thick the myocardium of the left ventricle is. It is a strong predictor for cardiovascular events [**?**], as it is an excellent indicator of the health of the myocardium. LVM should be measured at the end of diastole, when the left ventricle is relaxed and full of blood.

Hypertropic cardiomyopathy(HCM), as indicated via elevated LVM(how elevated?) occurs when the left ventricle myocardium thickens to a point where heart function is impaired. There are elevated risks of dilated cardiomyopathy in athletes as the natural thickening of the ventricle walls due to exercise can mask the condition.

Dilated cardiomyopathy(DCM) as indicated by lower than normal LVM(again what is normal) occurs when the left ventricle myocardium thins and stretches which compromises the strength and function of the left ventricle.

In conjunction, LVEF and LVM can help inform physicians on the efficiency of the heart and the reasons behind that level of function. There are many methods both invasive and noninvasive that can be used to measure LVEF. Noninvasive measurement tecniques are always prefered if they offer similar performance to invasive methods. Imaging the heart has the ability of providing detailed structural information of the heart.

3.5 Cardiac Imaging

The most common imaging modalities used image the heart are Magnetic resonance imaging(MRI), x-ray computed tomography(CT), and ultrasound(US).

MRI employs strong magnetic fields and quantum mechanical properties of particles to get signal from the presence of hydrogen ions ie protons. The behavior of protons within the context of an MRI scan is highly dependent on the structure and magnetic properties of the tissue surrounding each proton. However, the scanner is not getting signal from individual protons but groups of protons within each voxel of the image. MRI is usually noninvasive but can require the use of contrast agents. Contrast agents are usually used in cardiac MRI to detect scar tissue in the myocardium. As infarcted tissue by definition has limited or absent blood supply, the perfusion of contrast agents through the myocardium will highlight damaged regions. MRI offers unparralleled contrast between different biological tissues. Once the heart has been scanned, it is relatively easy to calculate the interior volume of the left ventricle. For example, region growing methods are usually used to directly calculate the volume from the 3D scan. This approach is very accurate as no assumptions are made about the shape of the ventricle.

However, scanners are usually in short supply, require skilled operators, are expensive to run and maintain, and are slow to aquire images. When the subject of the image is moving, long image aquisition time can lead to motion artefacts. Patients must hold their breath while their heart is being imaged to avoid introducing such artefacts. Patients that are unable to hold their breath, or those whose heart moves while it pumps may be illsuited for MRI. Additionally, due to the extremely strong magnetic fields present within the scanner, those with electrical or ferrous metal implants will not be able to be scanned. Some medical implants are labeled as MRI conditional, meaning that they are safe to be scanned, with some limitations to the scan. However, due to the complex restrictions placed on scanning these devices, the increased expense of MRI condtional technology and the relative rarity of such devices, most clinitians elect to use an alernative imaging modalities. [**?**]

CT imaging imvolves taking many 2D x-ray images from different angles then using computed tomography algorithums to genreate 3D images. CT imaging is not generally used for cardiac imaging as soft tissue does not provide sufficient contrast. The high energy x-rays used in CT are not heavily attenuated by low z elements found in soft tissue. Functional imaging of the heart with CT must utilize contrast agents [2]. High z compounds are injected into the bloodstream and subsequently imaged to give images of the bloodpool within the LV CT imaging has excellent spacial and temporal resolution. Fast image aquisition means minimal motion artefacts when compared to MRI. However, the radiation dose and use of contrast agents make cardiac CT rare.

3.6 Ultrasound

Ultrasound as used for echocardiograms is the core of cardiac imaging. It is the most widely used and prevalent modality to assess the structure and function of the heart. Ultrasound is robust as different scanner modes allow for 2D and 3D image and video aquisition. Doppler imaging allows for blood flow analysis and the use of contrast agents can provide increased levels of contrast. The relatively small size of ultrasound scanners when compared to MR and CT make echocardiograms more flexible than other modalities. Patients can TTE TEE 2D 3D 2CH 4CH

3.7 Echocardiogram

Echocardiograms are integral to cardiac evaluation.

3.8 Segmentation

Once the echocardiogram has been completed, the images must be analysed by a trained professional in order to calculate meaningful metrics. Both LVM and LVEF require segmentation, where some regions of the image are identified

Trabeculae as myo [**?**]

As ultrasound images typicaly have a low signal to noise ratio,

3.9 Automatic Segmentation

automatic is better think HR and BP variance older methods faster can analyse every image aquired and provide realtime analysis

3.10 Deep Learning

in thory less assumptions than stat shape models can be trained to better robustness as computational power increases can be used more and more and gets more powerful easy to continue improving networks

Neural nets Neural networks are computational structures that mimic the function of biological neurons. Networks with sufficeint complexity can model any mathmatical function. The functional unit of neural networks is the neuron. Individual neurons, like the whole, will each receive an input and provide an output based on an internal activation function. Activation funtions can vary in sophistication from simple step funtions that provide a binary output based on a hard coded threshold, to sigmoid functions that organicaly map the input to a range of values between zero and one. The selection of activation functions will depend on the desired behavior of the network.

One of the simplest configurations of neurons into a network is the perceptron. We will examine a hand written digit classifier multi-layer perceptron(mpl) to determine how neural networks make decisions. The structure of most networks consist of layers. In mpls, there is an input layer, hidden layers, and the output layer. The input layer in the case of the mnist digit image will simply be a vectorized image. The hidden layers contain the neurons that make the classification decision. The output layer will have a neuron for every class within the dataset so in our digit case we will have ten output neurons(0-9). The output layer will also be a softmax layer. The sum of all neuron values in a softmax layer must sum to one. For a given input each pixel value of our image will get sent to every neuron in the first of the hidden layers. These layers map the image input to a decision of what digit the network thinks it is seeing. If every neuron within the first hidden layer is getting input from every input layer neuron, then the output of every hidden layer neuron will be the same. In order to prevent this, we need to indroduce the ideas of weights. Weights signify the strength of connections between neurons. If a neuron has been strongly associated with a three, in our digit classification example, then there will be neurons that fire strongly when the image provided is a three. The weights in any neual network determine the flow of information through the network and the final decision. In this classification example, each digit from zero to nine has different characteristics about it that make it easily recognisable to our brain as well as a trained neural network. A fully trained network is structurally identical to its untrained counterpart. It is the weights that dictate the strength of the connections and by extention the pathways responsible for the psuedo cognition.

The difference between the mlp and a deep network is simply the number of layers. The mlp by convention is limited to one hidden layer and any network with more than one hidden layer is considered deep. Deep networks have seen much more use than mlps recently as computer hardware advancements have allowed the training and implementation of networks with many more than two hidden layers. While the mlp simply took the pixel intensity levels as input and subsequently made decisions based purely on pixel intensity, more and more deep networks are utilising convolutions to extract relavent features from the input images. Convolution involves generating a new image from the input image. Each pixel of this new image is the result of an elementwise multiplication of a region in the original image and a filter. For example, a 3 by 3 averaging mask averages the intensities of a pixel and its eight neighbors to give the intensity of the output pixel. This convolution is carried out repeatedly as the filter is swept across the input image.

3.11 CNN

Padding is as simple as adding pixels to the edges of the input image to increase the resolution. Convolution naturally reduces the resolution of the output due to the thickness of the mask. There are many types of mask that can be used in convolutions and they produce a wide variety of output images that highlight different features of the original image. Sobel and laplacian filters make edges bright and flat surfaces dark. Averaging filters can mitigate noisey images and produce blur. The size of mask can also be adjusted. We used an example of a 3x3 mask but larger masks such as 5x5 are also common. Masks are usually square with an odd width and height to ensure a middle pixel. All of the convolutions we have talked about are predetermined. A known mask is applied in order to produce a specific output. In deep convolutional neural networks, the masks are analagous to the mlp weights in that they will be defined through training. The nature of optimizing the network to complete segmentations means that the filters within the unet will be optimized to better select and highlight the features within the image that are relavent to the segmentation task at hand. For example, the network trained to segment the ultrasound cone from out images will most likely have filters that tend to identify the boundary between the background and ultrasound data, while the network trained to segment ventricles should be looking to identify the relatively dark bloodpool and bright myocardium. TODO(image)

UNET The Unet architecture is the standard approach for segmenting both 2D and 3D biomedical imaging data. It is strongly reliant on image augmentation to increase the size and variance of comparativly small biomedical image datasets. The signature architecture is comprised of a downsampling pathway followed by upsampling to produce a mask that is the same dimension as the input image. As the input image is convolved and downsampled, spacial information is converted into feature information. Also features that are too large to be encompased by the filters at the original image resolution may be able to be encompased after sufficient downsampling.

4 Method

4.1 Dataset

The raw dataset was provided to me consists of pathological echocardiography images from 95 patients. Most patients had both a 2CH and 4CH view of their heart. The files are stored as zipped nifti images and are 2D grayscale videos of the heart through multiple diastole systole cycles. Constructing the final network that will segment cardiac ultrasound will be broken up into several successive steps.

1. Train a network to successfuly segment ultrasound cone from the image
2. Train a network on 2CH bloodpool data
3. Train a network on 4CH bloodpool data
4. Train a network on both 2CH and 4CH bloodpool data
5. Depending on the success of combining 2 and 4CH data, train a network to segment the blood pool and epicardium

The purpose of segmenting the ultrasound cone from the rest of the image to make successive training easier. Eliminating image data that is irrelavent to the segmentation of the heart saves the network from having to learn that the extraneous information is so. The build up to the final network allows me to gradually build up the training dataset.

Once the dataset is complete, conversion into a form that can be used to train the U-nets is the same for 2CH and 4CH images. As this project was written within Google Colaboratory on account of the free cloud GPU, all of the image data was hosted on my personal Google Drive which was then mounted within the colab enviornment.

The images need to be converted into numpy arrays before they are used to train the network. The python library nilearn is used to convert the zipped nifti files into numpy arrays. The images are then resized to a resolution of 512 by 512 and the intensities are remapped from 0-255 to 0-1. The images are then collectively stored in a 4th order tensor with dimensions (number of images, X dimension, Y dimension, color channel). This allows for easy manipulation with Keras

The masks go through the same resizing conversion from nifty to numpy array. As the mask images can contain cone or ventricle masks, one layer of the masks must be chosen for training. This is accomplished via a thresholding operation to elminate either the cone or ventricle mask. Once the correct mask confiuration has been chosen, the masks can be resized and stored in a fourth order tensor as well.

dataset preprocessing The data was presented to me in the form of 96 folders, each corresponding to a certain patient. The names of these folders were all uniqe as they came from different hospitals at different times and were aquired by different machines and operators. Most patients had both two and four chamber images, with some having multiple of either. the naming convention for the individual images was as follows:

US_2CH.nii.gz US_4CH.nii.gz and if there were multiple of one view a two is appended to the end of the name as follows, *US_2CH2.nii.gz*

One of the first things that I did was to seperate the dataset into two and four chamber images. However, the image files would need to be renamed so that they could be told apart. The name of the directory containing the images was appended to the image filenames so that there would be no confilict of names and the image view sould still be identified by reading the name. An example of a filename in this form is as follows: *KCL_GC_001_US_2CH2.nii.gz* Once the images were somewhat sorted, manual segmentation could begin. Behind every competent neural net is a heap of painstakingly aquired data. To create a data point the blood pool and ultrasound cone from an individual frame of the nifti file must be segmented. To do this ITK-snap was used to both open the nifti file and produce the segmentations. In order to increase variance within the dataset and therfore robustness of the trained network, the chosen frames are at the general point of diastole and systole. Most all of the scans taken were of pathological hearts, so finding exact diastole and systole was challenging. Most easily distiguished features of the ecg within the echocardiogram were undistiguishale to our untrained eyes. We usually settled on two or three frames from each scan that looked sufficiently different. If more data is needed at a later time, additional frames could be segmented. The product of the segmentation is a mask that contains information on both the boundries of the ultrasound cone and bloodpool of the left ventricle. The background of the mask is always zero. If a cone mask is present, the cone is always one with the ventricle mask valued at two. If the ventricle is the only mask present then it will have a value of one. TODO(image of mask with cone and vent and mask with only cone) As the ultrasound cone is quite simple to identify within the images, we only ended up producing 25 data points with the cone manually segmented.

Once a mask was completed it needed to be associated with the correct frame from the correct image. The naming convention that we settled on was to simply use the same name as the image with the correct frame number appended to the end of the name. *KCL_GC_001_US_2CH2_01.nii.gz* The above example refers to the mask associated with the first frame of image KCL_GC_001_US_2CH2.nii.gz. This limits us to two digits to determine the frame so we are limited to frames 0-99. Most images are well under 100 frames. Images that are longer usually have a full cardiac cycle within the first 100 frames.

In order to train a network on the images, they needed to be a standard resolution. Initially this resolution was to be 512 by 512. 512 is a power of two in order to avoid non integer resolutions as the images are downsampled in the unet. The raw images in the dataset varied in resolution(TODO exact reses pls) with 512 by 512 being a rough mean. We later moved to a resolution of 800 by 800. This change was made based on two factors. We wanted to zero pad the images up to a resolution insead of cropping them or interpolating them to a smaller size. Interpolation would give a reproduction of our images at a different resolution while zero padding would encase the exact original image in zeros. Our zero padding scheme attemps to keep the original image in the center of the padding. TODO(image of unpadded and padded images) While interpolating image data is common practice, the end goal of producing a measuremnt of LVEF in real world units places a higher degree of importance in preserving the pixel spacing and dimension information contained in the image headers. As the data generation script took shape, moving data from nifti to numpy array and back again without worrying about pixel spacing was a relief. Once the images and masks were a standard resolution the image intensities needed to be scaled from zero to one. This was acheived with scalar division by 255.

Now that the data has been standardised, we can construct the datasets for training. As some masks contain both ventricle and cone information, they must be processed to produce a seperate ventricle and cone mask. At the initial round of training the cone dataset consisted of 25 image and mask pairs and the two chamber dataset consisted of 43 image and mask pairs.

training Due to the small amount of data that we were working with along with memory restrictions, on the fly data augmentation was going to be crucial to the performance of the porject. The premise was to apply random image transformations to an image and mask pair to generate a "new" peice of data from a base image. We decided to use the keras data generator class, which is capapble of perfroming a multitude of rigid image transformations. While some non-rigid transformations could be useful, we needed them to be precise enough to only deform the ventricle as we wanted the cone to remain undeformed. We needed to tread the fine line of augmenting our data in a way that increases the variation within the dataset without going too far. Too much augmentation that is too drastic could lead to the network struggling to find any sort of pattern without an immense amount of training. Due to machine and time constraints training needed to be as efficient as possible. After some trial and error we settled on the folloeing parameters to define our image and mask datagenerator:

1. rotation range of 45 degrees
2. width shift range of 10% of total image width
3. height shift range of 10% of total image height
4. shear range of 0.1 EXPLAIN!
5. zoom range of 0.1 EXPLAIN!
6. fill mode of nearest so most likely zero

We apply this same augmentation scheme to both the training and validation data to increase the size and variance of both sets. It is customary in machine learning, to take some data that could be used for training, and use it insead to test the trained model. Here we would give the trained model an image that it had never seen before and have it predict a mask. Comparison between the predicted and true masks gives us a better understanding at the effectivness of the model. As the model is literaly optimized to perform on the training data, using this data to validate its performance would not illistrate any ability to perform on real life data.

Results validation To assess the performance of our model we will both assess the performance with a K fold cross validation scheme and compare the accuracy of the predicted LVEF and LVM. K fold cross validation involves a number of steps.

1. Split the dataset into K equal groups
2. Select one group for the test set and have the rest be training data
3. train K models so that every data point gets to be in the test set

training params We settled into traning parameters early in the project. We needed to ensure that we would train long enough to get our validation loss sufficently low, while minimizing training time. For optimizer loss saving

class organisation

validation

data generation

References

[1] M. E. PFISTERER, A. BATTLER, and B. L. ZARET, "Range of normal values for left and right ventricular ejection fraction at rest and during exercise assessed by radionuclide angiocardiography," *European Heart Journal*, vol. 6, pp. 647–655, 08 1985.

[2] T. H. Marwick, "Ejection fraction pros and cons: Jacc state-of-the-art review," *Journal of the American College of Cardiology*, vol. 72, no. 19, pp. 2360 – 2379, 2018.

[3] N. H. M. Arrow, N. K. McAlister, and K. Buttoo, "Understanding cardiac "echo" reports," *Canadian Family Physician*, vol. 52, pp. 869–874, 2006.