GAN-enhanced Echocardiogram Segmentation

William Hewitt

May 30, 2020

Contents

1 Background

1.1 Echocardiography

Two-dimensional echocardiography (echo) is a common modality for assessing cardiac structure and function. The process of evaluating an echocardiogram typically involves the measurement of multiple clinical indices, which quantify various aspects of the hearts structure and function.

These clinical indices typically rely on manual image processing tasks performed by the reading clinician. An example of one the most important and general measurements is left ventricular ejection fraction (LVEF). The measurement of LVEF requires the manual segmentation of the left ventricle at both end-systole (ES) and end-diastole (ED). Semi-automated solutions for this do already exist, but have limited performance in the clinical workflow.

A typical echocardiography study contains 20 - 30 video clips (cines) stored in the DICOM format (a standard medical imaging format). Inside each DICOM file is standard JPEG pixel data which can be readily extracted by image processing libraries. Each cine is captured of a different angle, known as a view, of the heart looking at different points of clinical interest. Two of the most common views, Apical 2 Chamber (A2C) and Apical 4 Chamber (A4C) are commonly used to assess LV function and measure clinical indices such as LVEF. In this work we focus on the Apical 2 Chamber view.

1.2 Convolutional Neural Networks

Recent work has shown that an automated pipeline using convolutional neural networks (CNNs) are able to extract standard clinical indices to near human level accuracy [1], with a median absolute deviation (MAD) of 5.3%. Almost all the prior work in the space is reliant on encoder-decoder architecture CNNs to automate the segmentation of cardiac structures from echocardiography images. Subsequently, the entire analysis pipeline is sensitive to the performance of the underlying encoder-decoder neural network.

Encoder-decoder neural networks, for semantic segmentation, maps an input image to a segmentation mask, labelling each pixel within the image to one of any number of pre-defined segmentation classes. The encoder half of the neural network maps the high-dimension input image to a lower-dimension latent representation of the image, similar to a convolutional neural network used for classification. Rather than using fully-connected layers to obtain end classification, the decoder half of the neural network maps the latent representation of the image, to a segmentation map labelling each pixel to a segmentation class.

1.3 Generative Adversarial Networks

Since their introduction in 2014 by Goodfellow et al. [2], Generative Adversarial Networks (GANs) have made significant advances in a wide variety of deep learning problem domains. GANs are made up of two distinct neural networks, the generator and the discriminator. The generator takes a random variable "seed" and generates a fake sample. The discriminator classifies the input sample (coming from either the generator, or the ground truth dataset) into either real or fake.

During training, the weights of both the generator and the discriminator are updated simultaneously until the generator creates samples that the discriminator can no longer distinguish between real and fake. In this way, the generator of a GAN learns to model a probability distribution - a more obvious example coming from the field of image generation. If one imagines there is a probability distribution over the set of all labrador images, then the discriminator learns to discriminate between real labrador images from the training dataset, and fake ones synthesised by the generator. In early epochs, the synthesised images from the generator would have little resemblance to an image of a labrador, but in later epochs the generator produces highly realistc images of a labrador.

A subfield of GANs known as Conditional Generative Adversarial Networks (C-GANs) have more direct utility in the field of image segmentations. C-GANs are largely similar to a base GAN model, except the input of the generator is a specific condition or parameter rather than a random variable seed input.

In the context of this project, C-GANs have been utilised by Abdi et al. [3] to generate photorealistic echocardiogram frames, from a ground truth segmentation map. The generator of the GAN in this context can be viewed as the transfer function mapping the set of ground-truth segmentation masks, to the set of photorealistic echocardiography images. This can be described as:

$$G(M) = X_{synth} \tag{1}$$

Where $G(m)$ is the generator model, with condition M, the segmentation mask. X_{synth} is the synthesised photorealistic image.

1.4 Hypothesis

For this body of work, we aim to establish whether C-GANs can be used as a form of data augmentation to enhance the segmentation of echo frames using encoder-decoder convolutional neural networks. Firstly, we establish baseline for the segmentation of echo frames using the well known encoder-decoder neural network U-Net [4]. Secondly, we validate previous work that a C-GAN can be used to generate photorealistic ultrasound images from a ground truth segmentation map. Finally, we then compare the effect on segmentation accuracy by using either elastic deformation augmentation directly applied to the training images, or applied to the segmentation masks and then generating photorealistic images using a C-GAN, on the original encoder-decoder segmentation neural network.

2 Literature Review

2.1 Deep Learning for Echocardiography

Previous bodies of work [1] [5] [6] have proven that a deep learning based approach approaches human-level accuracy when measuring standard clinical indices such as LVEF.

Zhang et al. developed the most complete approach to the overall and is the focus for this section of our review. In their work they developed a multistage pipeline for the complete analysis of an echocardiogram, split into three distinct stages - view classification, segmentation and applications.

The first stage of the pipeline (view classification) classifies each cine in the echocardiography study into 1 of 22 different views. Using a common CNN architecture known as VGG-16, 10 random, downsized (224 x 224), grayscaled, and normalied (pixel values ranging from 0-1) frames are extracted from each cine and fed into the neural network, which returns a 22x1 probability vector with each entry referring to the probability of the cine being a given view. During validation this model achieved a 99% accuracy for classifying each view.

The second stage of the pipeline (segmentation) segments cines that are of a view of interest. Zhang et al. consider Apical 2 Chamber (A2C), Apical 4 Chamber (A4C), Parasternal Short Axis (PSAX) and Parasternal Long Axis (PLAX) views of interest. What view is of interest is typically driven by the clinical question, in this case being assessing LV function. Having identified the views that are of interet in the view classification stage, the segmentation stage uses four seperately trained encoder-decoder neural networks, with the same common encoder-decoder architecture known as U-Net. Using the same downsampled, normalised images as the view classification step the segmentation models take in a frame, and return a classification mask classifying each pixel into one of several classes. For the key views of interest for this research, the A2C and A4C segmentation models were trained on 200 and 177 manually annotated images respectively, generating a classification mask with four and six classes respectively. The A2C model classifies each pixel into either background, left atrial blood pool, left ventricular myocardium or left ventricular blood pool. The A4C model classifying into the same classes as the A2C with the addition of the right ventricular blood pool and the right atrium. The models converged to Intersection over Union (IoU, a common overlap metric for image segmentation tasks) values between 0.72 - 0.91 for structures of interest.

The third stage of the pipeline (applications) offers several modular applications, each producing some sort of analysis of cardiac structure and function. The key application of interest for this project measures LV volume (at both systole and diastole) and LVEF. With the newly segmented A2C and A4C cines (of which there are generally multiple in each study) and the area length method (whichs models the LV as a cone, to provide an approximation of volume) the application produces an estimate of LV volume at each frame of the cine, for each cine. A peak finding algorithm using a sliding window approach marks marks the ES and ED frames of each cine by simply identifying minima and maxima. With the now computed left ventricular end systolic and end diastolic volumes (LVESV/LVEDV) the LVEF can be calculated using the standard formula:

$$LVEF = \frac{LVEDV - LVESV}{LVEDV} \times 100 \tag{2}$$

The LV volume and LVEF application reports the average LVEF for each study, across multiple views. The Median Absolute Deviation (MAD) for LVEF was 5.3% over 3101 studies.

Zhang et al. takes an interesting approach, which leans on the strengths of automated measurement techniques. In typical clinical practice, a reading clinician would only take 1 or 2 measurements of LVEF - but would take in to account factors such as image quality and if there was a well defined myocardium - blood pool boundary before taking a measurement. Additionally in modern practice clinicians will conventionally use a more sophisticated technique to calculate the LV volume at a given point in time - commonly the Simpsons Biplane method - which simultaneously uses information from A2C and A4C views of the heart and constructs a substantially more accurate geometric model of the heart. Accuracy improving techniques as described are readily employed as a manual approach, but are non-trivial when developing an automated approach.

The complimentary approach taken by Zhang et al. addresses it as they segment every frame in the sequence and compute less accurate volumes along every single frame gives a sampling advantage as the automated approach can average over a significantly greater number of samples than the manual approach. The reasoning behind this being that having a high number of lower quality samples produces a superior result than one or two high quality samples.

In this approach however, and in most approaches found in literature, the analysis step (where interesting information is produced) is entirely reliant on accurate segmentations and view classifications in earlier steps. More sophisticated approaches have been developed, such as those using 3D CNNs [7] however 3D CNN approaches have substantially more free variables (parameters) to train, and require significantly larger datasets in a field where datasets are already constrained.

A summary for work in the space of deep learning for echocardiography can be simply summarised as requiring improved segmentation models.

2.2 Conditional Generative Adversarial Networks

After their introduction by Goodfellow et al. in 2014, using GANs people have made substantial advancements in the deep learning. The focus of this literature review is not on GANs broadly, but on the utility of the subfield of Conditional GANs used to synthesise echocardiography images.

The study by Abdi et al. is the only substantial body of work in this space, where they demonstrated that a GAN can be trained to produce photorealistic images from an input condition (the segmentation mask). Their generative model was a U-Net inspired encoder-decoder network - without the typical skip connections. It is not immediately apparent why the authors opted to avoid skip connections, potentially as latent representations of the segmentation masks don't contain particularly useful information for the decoder stage of the network across the bulk of the image (i.e. as the bulk of the photorealistic image is noise, but the bulk of the condition is background).

In the study they used a patch-based discriminator model, which assesses the "realness" of each patch of the generated image rather than the entire image. The patch-based discriminator was a simple CNN with 5 convolutional layers.

The most interesting innovation of this work was the use of a Least Squared Error (LSE) which has been shown to push the probability distribution modelled by the generator model closer to the real data distribution. The Generator and Discriminator were co-optimized for this loss, aiming to minimize and maximize the loss function respectively.

$$\mathcal{L}_{cGAN} = \mathbb{E}_{x,y}[(1 - D(y,x)^2)] + \mathbb{E}_{x}[D(y,G(y))^2] \tag{3}$$

y and x refer to the input condition and the photorealistic image respectively. $D(y,x)$ refers to the patch-based discriminator and $G(y)$ refers to the encoder-decoder generator model.

The study by Abdi et al. demonstrated that motivating results can be generated by using simple but powerful generator and discriminator models, all of which being well understood in literature. The key limitation of their study was their principal focus on ED frames, meaning the heart is at its largest size (during the filling, pre-contraction) of the cardiac cycle. This (deliberately) limited the distribution modelled by the GAN to that of just ED frames. This does make for a potentially more physiologically interesting model however, as it is a reasonable assumption that the size of the LV varies linearly at ED and ES between patients. Whereas intermediary frames (or the change over time between intermediary frames) is a more complex phenomena. This forseeably opens up interesting further analysis.

Other limitations include simply performing qualitative analysis of the output of the generative model. Although interesting for this case, pixelwise difference metrics (which was a term of their loss function) would potentially have been an interesting analysis between the ground truth and generated test images. There was also no analysis on the generator models sensitivity to changes of the boundary of the segmentations.

Limitations aside, the study by Abdi et al. has layed the groundwork for both this body of work and future works in the space.

2.3 GAN-enhanced Image Segmentation

Although the focus of this work is on the utility of GANs as a type of data augmentation for the training of encoder-decoder neural networks, given the primary endpoint of enhancing segmentation accuracy it would be fair to include other segmentation techniques utilising GANs.

The study by Xue et al. [8] developed a novel approach to the GAN architecture called SeGAN where they replaced the generator (G) model with a segmentor (S) model, and the discriminator (D) model with a critic (C) model. They also introduced a novel multiscale feature loss function, which minimizes S and maximizes C during training.

The principle innovation in this work is the proposed multiscale objective loss function, which measures the difference between the generated segmentation and the ground truth segmentation at multiple layers within the critic. The advantage of this is forcing the critic to learn higher-order features that capture actual spatial relationships between the pixels, rather than learning a trivial solution which a conventional GAN would - the binary ”real” or ”fake” classification.

This brief review of a pivotal body of work is to give wider context to the space of medical imaging GANs. There aren't immediate justifications of superiority of the GAN approach, although superior segmentation accuracy measurements were recorded on several datasets. One interesting qualitative result of the study was the comment that their segmentations appeared smoother, potentially of interest in some medical imaging problems.

2.4 Summary

In summary, prior work in the space has convincingly demonstrated that a deep learning approach can successfully automate the measurement of clinical indices of an echocardiography study.

What is apparent that in the bulk of the approaches is the segmentation accuracy is the rate limiting factor. Superior segmentations would enable more accurate and precise measurement of clinical indices.

3 Methods

3.1 Segmentation of Echocardiography Frames

For the first part of this body of work, we develop a baseline segmentation accuracy when segmenting echo frames using an encoder-decoder neural network. For our work, we used the well-known U-Net architecture.

For all the experiments in this work, we used the CAMUS datast [5] which was also the dataset behind the C-GAN work of Abdi et al. All of the experiments were run using Python 3 and TensorFlow 2.0.0.

The CAMUS dataset was downloaded and converted from .MHD/.RAW format to a NumPy array using the SimpleITK library for Python. Apical 2 Chamber (A2C) frames at ES and ED (only the ground truth for ES and ED is provided) were extracted, resampled to 256x256 and pixel values normalised to 0-1. All preprocessing steps were carried out using NumPy and OpenCV

The final dataset to establish baseline performance consisted of 900 samples, split 80:10:10 for train:test:validation (810, 90, 90 samples respectively).

We built a U-Net encoder-decoder segmentation neural network that used the pretrained (on ImageNet) VGG-16 model from Keras. The last convolution block was dropped, and the weights for the encoder frozen (this was saved as a hyperparameter). Transpose convolutional layers were used for upsampling, and skip connections were additionally used. The overall model architecture can be seen in Figure 1.

For training the Adam optimizer was used, with a base learning rate of 1e-5. Categorical Cross Entropy (CCE) was used for a loss function. Dropout was applied in the bottleneck layers. The model was trained on a batch size of 16. Convergence criteria was defined as an absolute difference between training and validation loss of >10% of training loss

3.2 Validating C-GANs for Echocardiography Frames

For our purposes, it was of interest to see if a C-GAN could be trained that would be able to generate photorealistic echocardiography frames from a segmentation mask input condition.

We used just ED frames from A2C views in the CAMUS dataset, which were preprocessed similarly to in the segmentation step, resamplwed to 256x256 and normalised to have pixel values from [0-1].

We used the same framework as Abdi et al. to train a C-GAN on ED frames of the A2C view. The network was trained for 100,000 epochs with the model and sample images saved every 100 epochs. The Adam optimizer was used, with a learning rate of 1.3e-4 and 1.5e-4 for the generator and discriminator respectively. Patch size was set to 16x16 for the patch based discriminator, and batch size was set to 8.

3.3 Data Augmentation for the Segmentation of Echocardiography Frames

To compare the effect on segmentation accuracy for our two data augmentation techniques, we created two new dataset labelled "real" and "generated". Real contained the original CAMUS training dataset with elastic deformation augmentations applied directly images. Generated contained synthetic images after deforming the ground truth segmentation map, and passing the deformed ground truth image to the generator trained in 3.2.

Elastic deformations for both datasets were created using the elasticdeform library for Python. Each dataset contained the original ground truth images, as well as a deformed counterpart to each image giving double the original dataset size. In total, each dataset contained 1800 images.

Two seperate U-Net models were trained using the same framework as 3.1, with the same hyperparameters and convergence criteria. This was enforced to give a good objective view of the effect of the different augmentation methods, and to isolate the effect on the accuracy of the model to just the independent variable (type of data augmentation).

4 Results and Discussion

After some hyperparameter tuning, our baseline U-Net model achieved IoU scores between 0.7-0.9 for structures of interest (i.e. myocardium, blood pool and atrium). Best baseline results were achieved with frozen encoder weights, and convergence criteria was reached after 50 epochs.

Hyperparameters were unchanged when testing the different augmentation methods, as the objective was not to achieve the highest possible segmentation accuracy - but demonstrate if the augmentation techniques had a significant effect on segmentation accuracy. For the real dataset, convergence was reached after 30 epochs and IoU scores ranged from 0.8-0.9. For the generated dataset, convergence was reached after 80 epochs and IoU scores ranged from 0.7-0.8. A table of results is provided below.

Augmentation	**mIoU**
None	0.6
Real Elastic Deformation	0.9
Generated Elastic Deformation	0.8

This gives strong evidence that using a pre-trained generator gives superior data augmentation and segmentation accuracy than applying data augmentation simply to the training images.

5 Conclusion

6 Ethics

The dataset (CAMUS) used in this study had been reviewed by the Université de Lyon ethics committee, and made publicly available prospect to certain regulations.

The results of this study included a comparison of segmentation accuracy between male and female patients, which showed no significant difference in accuracy. Further demographic analysis of this sort was not possible as most identifiers were not made available, however would increase the ethical value of future studies in this field.

References

[1] J. Zhang, S. Gajjala, P. Agrawal, G. H. Tison, L. A. Hallock, L. Beussink-Nelson, M. H. Lassen, E. Fan, M. A. Aras, and C. Jordan, "Fully automated echocardiogram interpretation in clinical practice: Feasibility and diagnostic accuracy," *Circulation*, vol. 138, no. 16, pp. 1623–1635, 2018.

[2] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, pp. 2672–2680, 2014.

[3] A. H. Abdi, T. Tsang, and P. Abolmaesumi, *GAN-Enhanced Conditional Echocardiogram Generation*. 2019.

[4] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 234–241, Springer, 2015.

[5] S. Leclerc, E. Smistad, J. Pedrosa, A. Ø stvik, F. Cervenansky, F. Espinosa, T. Espeland, E. A. R. Berg, P.-M. Jodoin, T. Grenier, C. Lartizien, J. D'hooge, L. Lovstakken, and O. Bernard, "Deep Learning for Segmentation Using an Open Large-Scale Dataset in 2D Echocardiography," *IEEE Transactions on Medical Imaging*, vol. 38, pp. 2198–2210, Sept. 2019.

[6] W. Hewitt, L. Curtis, A. Spyker, H. Walsh, L. Howitt, and P. Gladding, "Artificial intelligence in echocardiography for standard clinical metrics," *Heart, Lung and Circulation*, vol. 28, p. S12, 2019.

[7] D. Ouyang, B. He, A. Ghorbani, M. P. Lungren, E. A. Ashley, D. H. Liang, and J. Y. Zou, "EchoNet-Dynamic: A Large New Cardiac Motion Video Data Resource for Medical Machine Learning," p. 11.

[8] X. Zhang, X. Zhu, r. X. Y. Zhang, N. Zhang, P. Li, and L. Wang, "SegGAN: Semantic Segmentation with Generative Adversarial Network," in *2018 IEEE Fourth International Conference on Multimedia Big Data (BigMM)*, pp. 1–5, Sept. 2018.