*Proceedings of CECNet 2021*
*A.J. Tallón-Ballesteros (Ed.)*

*doi:10.3233/FAIA210416*

# An Explainable Convolutional Neural Networks for Automatic Segmentation of the Left Ventricle in Cardiac MRI

Jun LIU[a,c], Feng DENG[b,1], Geng YUAN[c], Xue LIN[c], Houbing SONG[d] and Yanzhi WANG[c]

[a] *Robotics Institute, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA*

[b] *Beijing Key Laboratory of High Dynamic Navigation Technology, Beijing Information Science & Technology University, Beijing, China*

[c] *Department of Electrical & Computer Engineering, College of Engineering, Northeastern University, Boston, MA, USA*

[d] *Security and Optimization for Networked Globe Laboratory (SONG Lab), Embry-Riddle Aeronautical University, Daytona Beach, FL, USA*

**Abstract.** Recently, the study on model interpretability has become a hot topic in deep learning research area. Especially in the field of medical imaging, the requirements for safety are extremely high; Moreover, it is very important for the model to be able to explain. However, the existing solutions for left ventricular segmentation by convolutional neural networks are black boxes; explainable CNNs remains a challenge; explainable deep learning models has always been a task often overlooked in the entire data science lifecycle by data scientists or deep learning engineers. Because of very limited medical imaging data, most solutions currently use transfer learning methods to transfer the model which used on large-scale benchmark data sets (such as ImageNet) to fine tune medical imaging models. Consequently, a large amount of useless parameters are generated, resulting in further barrier for the model to provide a convincing explanation. This paper presents a novel method to automatically segment the Left Ventricle in Cardiac MRI by explainable convolutional neural networks with optimized size and parameters by our enhanced Deep Learning GPU Training System. It is very suitable for deployment on mobile devices. We simplify deep learning tasks on DIGITS systems, monitoring performance, and displaying the heat map of each layer of the network with advanced visualizations in real time. Our experiment results demonstrated that the proposed method is feasible and efficient.

**Keywords.** Explainable, image segmentation, left ventricle, interpretable, transfer learning, CNNs, mobile device

## 1. Introduction

The purpose of medical image segmentation is to segment parts with special significance in medical images, extract relevant features, provide reliable basis for

[1] Corresponding Author, Feng Deng, Beijing Key Laboratory of High Dynamic Navigation Technology, Beijing Information Science & Technology University, Beijing, No.35 North Fourth Ring Road, Chaoyang District, Beijing, China; E-mail: dengfeng@bistu.edu.cn.

clinical diagnosis and pathological research, and help doctors make more accurate diagnosis. Due to the complexity of medical images themselves, a series of problems such as nonuniformity and individual differences need to be solved in the segmentation process, so the general image segmentation methods are difficult to be directly applied to medical image segmentation.

Because medical imaging is ultimately an aid to doctors' clinical diagnosis, it is not enough to tell doctors whether a MRI image is ill or not. At the same time, for the classification and segmentation results given by the network, doctors also want to know why, that is to say, from the doctor's point of view, they want to know how interpretable medical AI is. To meet these needs our model is explainable and understandable for doctor it allows us to verify hypotheses and help us to improve model's performance.

## 2. Related Work

Although the high nonlinearity endows the multilayer neural network with high model representation ability, it can achieve very gratifying performance on many issues with some parameter adjustment techniques that can be called modern alchemy. Model designer more want to know what knowledge the model has learned from the data (expressed in a way that human beings can understand), which leads to the final decision. Whether it can help designer find some potential correlations, if designer want to develop an application based on deep learning model to help doctor judge patient risks. Besides the final judgment results, designer may also need to know what factors the model is based on. If a model can't be explained at all, its application in many fields will be limited because it can't give more reliable information. In this paper, we focus on solving the interpretability problem of deep learning.

Our major contributions of this paper are the following:

•--This paper discusses the modeling stage from explainable methods before modeling,

•--An explainable model and explained the model design is provided

•--This paper use NVIDIA Deep Learning GPU Training System(Digits) provide visual analysis of typical convolution layer and de-convolution layer.

•--This paper provides experiment methods and an evaluation method, it verified the importance of Dice metric as a measure of medical image segmentation.

## 3. Explainable Model design

### *3.1 Input Data Set*

The data set we utilizing Sunnybrook cardiac images which is a series of cardiac images (specifically MRI short-axis (SAX) scans) that have been expertly labeled. See References [1, 2, 3] for full citation information. The main view for assessing ventricle size is the short- axis stack (PSAX), Contains images taken in a plane perpendicular to the long axis (PLAX) of the left ventricle.

The examples of the MRI data are shown below. This image is an instance of the data. On the left Figure.1a are the MRI images and the right Figure.1b is the expertly-annotated regions (often called contours). The portions of the images that are part of

the LV are denoted in white. Note that the size of LV varies from image to image, but the LV typically takes up a relatively small region of the entire image.

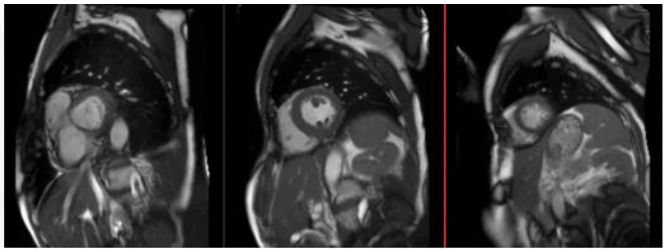

Figure.1a Original LV MRI image

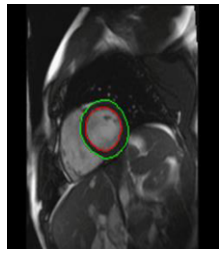

Figure.1b expert segmentation

The images themselves are originally 256 x 256 grayscale DICOM format, a common image format in medical imaging. The label is a tensor of size 256 x 256 x 2. The reason the last dimension is a 2 is that the pixel is in one of two classes so each pixel label has a vector of size 2 associated with it.

### *3.2 Explainable methods before modeling*

Explainable methods before modeling also named Pre-modelling explainability is a collection of diverse methods with a common goal of gaining a better understanding of dataset used for model development [4].The key of interpretable methods before modeling is to help designer quickly and comprehensively understand the characteristics of data distribution, so as to help they consider the possible problems in the modeling process and choose the most reasonable model to approach the optimal solution that can be achieved by the problem.

Compared with ordinary images, medical images have the characteristics of high complexity, large gray scale range and unclear boundary, and the internal structure of human body is relatively fixed. The distribution of segmentation targets in human body images is regular, and the semantics are not simple and clear(i.e. inter-annotator disagreements, poor segmentation reproducibility) so the network need combine low-resolution information during downsampling (providing the basis for object category recognition) and high-resolution information during upsampling (providing the basis for accurate segmentation and positioning).

Because the data collection of medical images is relatively difficult and the amount of data is small, if there are too many parameters in the model, it will easily lead to over-fitting, so the model with small size and few parameters is suitable.

### *3.3 Establish an explainable model*

In order to capture small regions of interest, convolution layer is used to capture larger receptive fields.

This paper can accomplish this by using an image recognition neural network, and replacing the fully-connected layers (typically the last few layers) with deconvolution layers (arguably more accurately called transpose convolution layers).

Deconvolution [5-8] is an upsampling method that brings a smaller image data set back up to its original size for final pixel classification. It can be helpful to visualize

how the input data (in the case a tensor of size 256 x 256 x 1) "flows" through the graph, i.e., how the data is transformed via the different operations of convolution, pooling and such. The Figure.2 represents the transformations that our data will undergo in the next task.

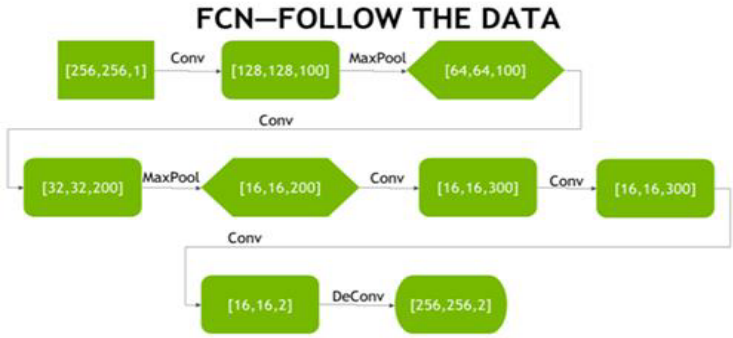


Figure.2 network structure

The network represented by the figure above is consisted convolution layers, pooling layers, and a final deconvolution layer, with the input image being transformed as indicated in the image.

1.Convolution1, 5 x 5 kernel, stride 2
2.Maxpooling1, 2 x 2 window, stride 2
3.Convolution2, 5 x 5 kernel, stride 2
4.Maxpooling2, 2 x 2 window, stride 2
5.Convolution3, 3 x 3 kernel, stride 1
6.Convolution4, 3 x 3 kernel, stride 1
7.Score_classes, 1x1 kernel, stride 1
8.Upscore (deconvolution), 31 x 31 kernel, stride 16

The original image is reduced to 1/4 after conv1 and pool1

The image is reduced to 1/16 after conv2 and pool2 for the second time

Conv3 and conv4 are convolution operation, and the number of feature Maps of the image changes, but the image size is still 1/16 of the original image. At this time, the image is no longer called feature map but heat Map.

The deconvolution layer is used to up-sample the feature map of the last convolution layer, so that it can be restored to the same size as the input image

For the operation of convolution, the input element matrix x and the output element matrix y, this process is described by matrix operation:

$$y = Cx \tag{1}$$

By deduction, we can get sparse matrix: C

The operation of deconvolution is to inverse the matrix operation:

$$x = C^{T}y \tag{2}$$

According to the matrix differential formula:

$$\frac{\partial Ax+b}{\partial x} = A^{T} \tag{3}$$

It can be deduced the gradient of the loss of the deconvolution to the input x:

$$\frac{\partial Loss}{\partial x_i} = \sum_i \frac{\partial Loss}{\partial y_i}\frac{\partial y_i}{\partial x_i} = \frac{\partial Loss}{\partial y_i}C_{i,j} = \frac{\partial Loss}{\partial y_i}C_{*,j} = C_{*,j}^{T}\frac{\partial Loss}{\partial y_i} \tag{4}$$

### *3.4 Explanatory after modeling*

There are three interpretable methods for deep learning: 1. Hidden layer analysis, 2. Sensitivity analysis, and 3. Agent/substitute model. In this section, this paper mainly analysis the first method: hidden layer analysis.

The interpretable method after modeling is mainly aimed at the deep learning model with black box property. By evaluating the accuracy of the model, designer can get the change of the performance of the hidden layer in the whole training process and after the training [9]. This paper use our Deep Learning GPU Training System to train and analysis the model's performance. Visualize the result of convolution kernel after activation. Designer can see the result of image convolution, which helps designer to understand the function of convolution kernel. Through the heat map, designer can know which parts of the image play a key role in the image classification problem, and at the same time, designer can locate the position of objects in the image

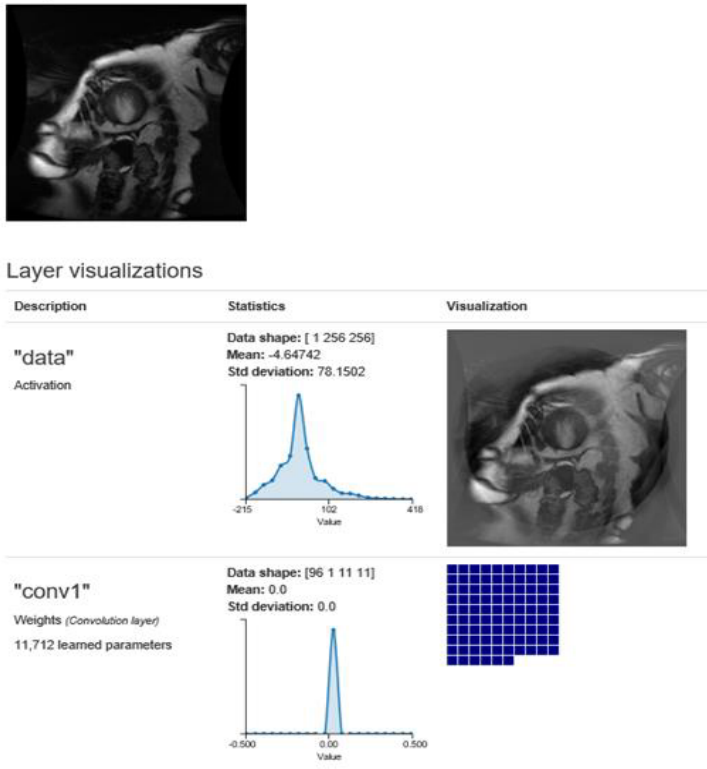


Figure.3 convolution layer1 Visualization

On our model based on MRI dataset training, after 5 cycles of training, In Figure3, the features of the image extracted by convolution layer are coarse, in the subsequent convolution2 layer feature map is obviously improved more detail.

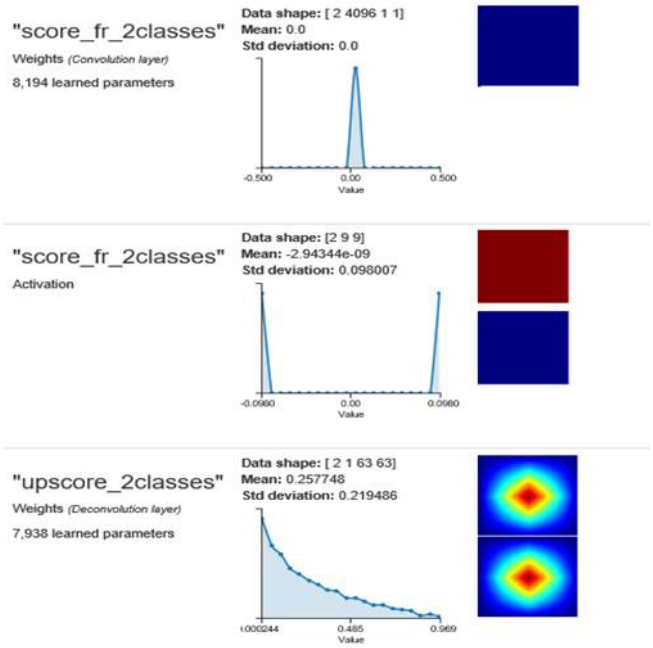


Figure.4 full convolution layer Visualization

In Figure4, full convolution layer has extracted the left ventricle and non-left ventricle parts feature in the picture into two part. In deconvolution layer left ventricle is highlighted in red, indicating that the network is looking at the correct position when making classification judgment. through the heat map, we can know which parts of the image play a key role in the image classification problem, and at the same time, we can locate the position of objects in the image.

## 4. Experiments with model

### *4.1 Workflow*

1.Prepare input data--Input data can be Numpy arrays.

2.Build the Computation Graph--Create the graph of the neural network including specialized nodes such as inference, loss and training nodes.

3.Train the model--inject input data into the graph in a Session and loop over your input data. Customize your batch size, number of epochs, learning rate, etc.

4.Evaluate the model--run inference (using the same graph from training) on previously unseen data and evaluate the accuracy of your model based on a suitable metric.

### *4.2 Parameter search*

At this point a neural network is created that it seems has the right structure to do a reasonably good job and an accuracy metric is used that correctly show how well the network is learning the segmentation task. But up to this point the evaluation accuracy

hasn't been as high as expected. The next thing to consider is try to search the parameter space a bit more. Up to now the number of epochs have been changed and there are a few more parameters can be tested that could push the accuracy score higher. These are:

•--learning_rate: the initial learning rate

•--decay_rate: the rate that the initial learning rate decays.,

e.g., 1.0 is no decay, 0.5 means cut the decay rate in half each step, etc.

•--decay_steps: the number of steps to execute before changing the learning rate

The learning rate is the rate at which the weights are adjusted each time after run back propogation. If the learning rate is too large, it should be end up adjusting the weights by values that are too large and it should be end up oscillating around a correct solution instead of converging. If the learning rate is too small, the adjustments to the weights will be too small and it might take a very long time before we converge to a solution that we expect. One technique often utilized is a variable or adjustable learning rate. At the beginning of training, a larger learning rate will be used so that we make large adjustments to the weights and hopefully get in the neighborhood of a good solution. Then as we continue to train we'll successively decrease the learning rate so that we can begin to zero in on a solution. The three parameters listed above will help you control the learning rate, how much it changes, and how often it changes.

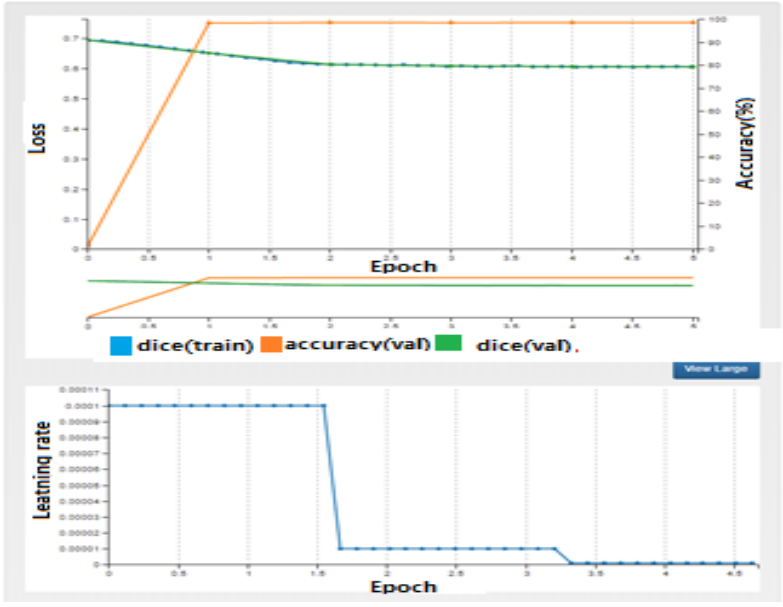


Figure.5 Predict Result high without Dice metric

From Figure5 when set learning rate=0.0001 even through trained 1 epoch the model reaches an accuracy of over 98%, This means that 98% of pixels are correctly predicted as belonging to either the left ventricle or the background. This is very good!

## 5. Evaluation method

This paper considers fully what exactly we are computing when we check accuracy. The current accuracy metric is simply telling us how many pixels we are

computing correctly. So in the case above with 5 epochs, model is correctly predicting the value of a pixel 98 % of the time. However, notice from the images above that the region of LV is typically quite small compared to the entire image size. This leads to a problem called class imbalance [10], i.e., one class is much more probable than the other class. In this case, if simply designed a network to output the class not LV for every output pixel, result still have something like 95% accuracy. But that would be a seemingly useless network. This is meant to illustrate that high pixel accuracy doesn't always imply superior segmentation ability [11].

What the paper need is an accuracy metric that gives some indication of how well the network segments the left ventricle irrespective of the imbalance.

One metric can be used to more accurately determine how well the network is segmenting LV is called the Dice metric or Sorensen-Dice coefficient, among other names. This is a metric to compare the similarity of two samples. In this case it is used to compare the two areas of interest, i.e., the area of the expertly-labeled contour and the area of the predicted contour. The formula for computing the Dice metric is:

$$\text{Dice metric} = \frac{2A_{nl}}{A_n + A_l} \tag{5}$$

where An is the area of the contour predicted by this neural network, Al is the area of the contour from the expertly-segmented label and Anl is the intersection of the two, i.e., the area of the contour that is predicted correctly by the network. 1.0 means perfect score. More accurately compute how well predicting the contour against the label, We can just count pixels to the respective areas.

## 6. Experiments Result

When the training epoch is increased to 30 a significant increase can be seem in accuracy. In fact an accuracy of 98.3% is quite good.

There are a few more parameters can be tested that could push our accuracy score higher. These are:

--learning_rate 0.03
--decay_rate 0.75
--decay_steps 10000
--num_epochs 100

This model gest accuracy of 99.7% and Dice metric above 86.3%

Example of segmentation result

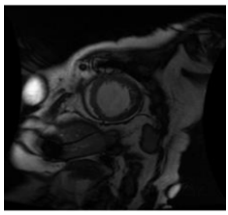

Figure.6a Original LV image

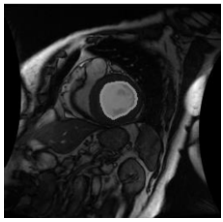

Figure.6b LV Validation

In Figure.6b the contour of the prediction is very smooth and the predicted results are almost consistent with Figure.6a the original left ventricular position

Compare with different explainable method on Sunnybrook datasets

| Method | Dice metric(val) |
|---|---|
| ours | 0.863 |

| | |
|---|---|
| Xiaofeng Liu et al.(2021)[11] | 0.860 |
| Manuel Pérez-Pelegrí et al(2021) [12] | 0.790 |
| QiaoZheng et al.(2019[13] | 0.760 |
| Alexnet+FCN(base line) | 0.622 |

## 7. Summary

This paper focused on design explainable convolutional neural network from the interpretability of deep learning methods for automatic Segmentation of the Left Ventricle in Cardiac MRI at a trade-off between accuracy and interpretability. The paper is different with traditional CNNs design, it present explainable methods before modeling, established an explainable model analyzed the hidden layer after modeling, the experimental steps and an evaluation method is provided. After fine tune this model obtained high accuracy and dice metric. This model is simply and lightweight, it is also very suitable to deploy on mobile devices to diagnose medical images.

## Acknowledgment

NVIDIA Deep Learning Institute provided technical support and this work was supported in part by National Natural Science Foundation of China (41871348); Beijing Information Science & Technology University (2020KYNH224); and Beijing Key Laboratory of High Dynamic Navigation Technology (HDN2019002).

## References

[1] cardiac images from earlier competitionhttp://smial.sri.utoronto.ca/LV_Challenge/Data.html
[2] "Sunnybrook Cardiac MR Database" is made available under the CC0 1.0 Universal license described above, and with more detail here: http://creativecommons.org/publicdomain/zero/1.0/ .
[3] Radau P, Lu Y, Connelly K, Paul G, Dick AJ, Wright GA. "Evaluation Framework for Algorithms Segmenting Short Axis Cardiac MRI." The MIDAS Journal -Cardiac MR Left Ventricle Segmentation Challenge,http://hdl.handle.net/10380/3070
[4] Bahador Khaleghi, The How of Explainable AI: Pre-modelling Explainability, 2019·8. https://towardsdatascience.com/the-how-of-explainable-ai-pre-modelling-explainability-699150495fe4
[5] Fully Convolutional Networks for Semantic Segmentation http://fcn.berkeleyvision.org/
[6] Long, Shelhamer, Darrell; "Fully Convoutional Networks for Semantic Segmentation", CVPR 2015..
[7] Zeiler, Krishnan, Taylor, Fergus; "Deconvolutional Networks", CVPR 2010.
[8] Jonathan Long, Evan Shelhamer, Trevor Darrell Fully Convolutional Networks for Semantic Segmentation, CVPR2015
[9] Guillaume Alain, Yoshua Bengio,Understanding intermediate layers using linear classifier probes, - arXiv preprint arXiv:1610.01644, 2016 - arxiv.org
[10] Jason Brownlee on December 23, 2019 in Imbalanced Classification, A Gentle Introduction to Imbalanced Classification https://machinelearningmastery.com/what-is-imbalanced-classification/
[11] Xiaofeng Liu,Segmentation of Cardiac Structures via Successive Subspace Learning with Saab Transform from Cine MRI, arXiv:2107.10718,2021
[12] Manuel Pérez-Pelegríet et al., Automatic left ventricle volume calculation with explainability through a deep learning weak-supervision methodology, Computer Methods and Programs in Biomedicine, Volume 208, September 2021, 106275
[13] Qiao zheng et al.Explainable cardiac pathology classification on cine MRI with motion characterization by semi-supervised learning of apparent flow, Medical Image Analysis Volume 56, August 2019, Pages 80-95