Homework 6

Hengtao Guo

April 4, 2019

1 Introduction

In this paper review assignment, I choose to focus on cardiac disease related diagnosis using computed tomography (CT) images based on deep learning methods. Some basic information can be seen in Table 1.

2 Motion Analysis

2.1 Motivation

Motion analysis is an important branch in computer vision task, especially in medical image analysis. To comprehensively interpret cardiac functioning, researchers have to know about its dynamic motion for further study. Such a 4D (3D for object appearance and 1D for time series) motion sequence can well contribute to prediction tasks.

2.2 Method

Their method can be basically separate into two consecutive steps.

Table 1: Basic information of three listed papers.

Paper Title	First Author	Year	Publishing	Citation
Deep-learning cardiac motion analysis for human survival prediction	Ghalib Bello	2019	Nature Machine Intelligence	1
Direct Automatic Coronary Calcium Scoring in Cardiac and Chest CT	Bob D. de Vos	2019	IEEE TMI	1
Direct Prediction of Cardiovascular Mortality from Low-dose Chest CT using Deep Learning	Sanne G.M. van Velzen	2019	Medical Imaging 2019: Image Processing	2

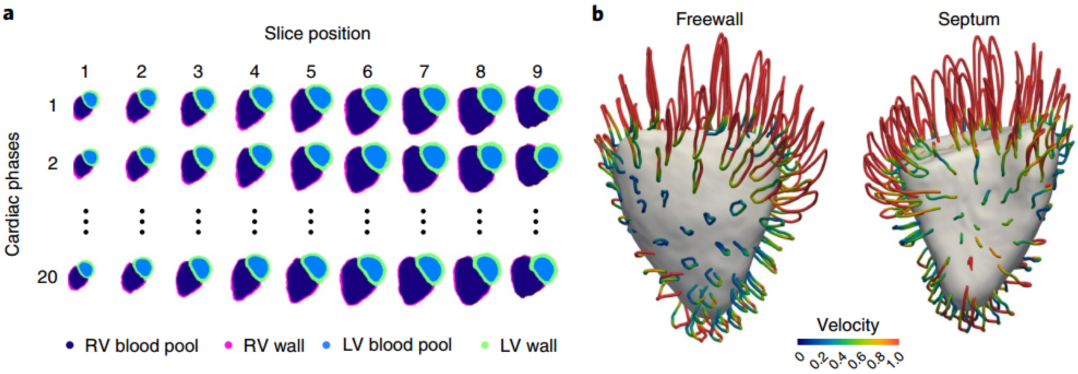

Figure 1: Segmentation results and landmarks locations.

2.2.1 Deep Segmentation and Landmark Locating

The image modality this paper dealing with is the MRI cardiac images. To first select out the heart region for further analysis, a fully convolutional neural network (FCN) is trained to do segmentation for right ventricle (RV) wall as well as blood pool and left ventricle (LV) wall with blood wall. The ground truth they have is the segmentation mask for these four anatomic structures.

To reduce the dimension of image representation, they propose to use landmarks instead of the original segmenation regions to represent the cardiac status. To accomplish this, the authors have manually labeled 202 landmarks cross each 3D volume. Each landmark has a 3D coordinates to denote the specific location. The these coordinates labels act as the ground truth for the FCN to regress during training, so the network is able to both perform segmentation and landmark annotation at the same time. Then in the next step, they only use the landmarks coordinates to describe the status of heart. The segmentation results and landmarks appearance can be seen in Fig. 1.

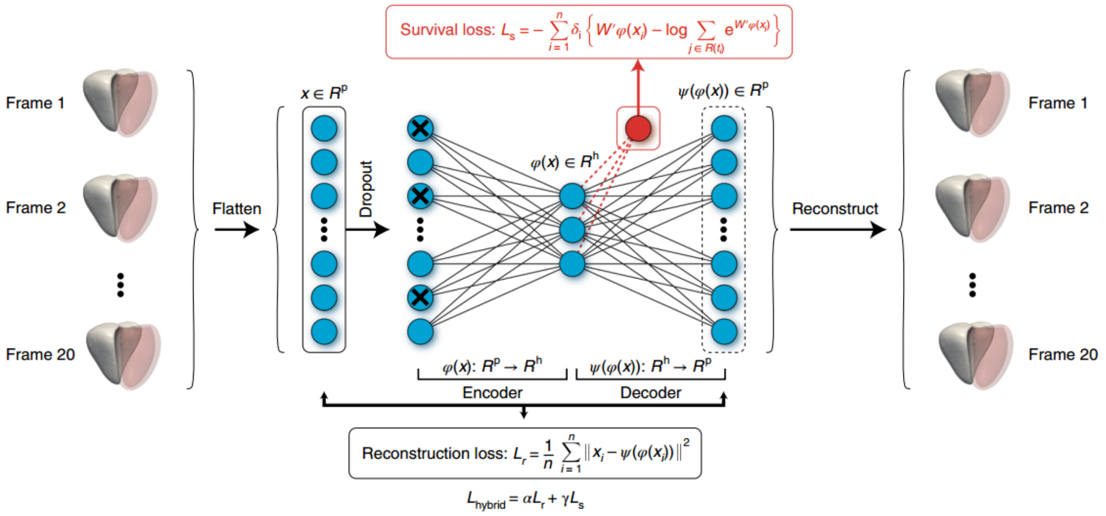
Survival loss: $L_s = -\sum_{i=1}^{n} \delta_i \left\{ W'\varphi(x) - \log \sum_{j \in R(t)} e^{W'\varphi(x)} \right\}$
Frame 1
Frame 2
Frame 20
Flatten
$x \in R^p$
Dropout
$\varphi(x) \in R^h$
$\psi(\varphi(x)) \in R^p$
Reconstruct
Frame 1
Frame 2
Frame 20
$\varphi(x): R^p \to R^h$
$\psi(\varphi(x)): R^h \to R^p$
Encoder
Decoder
Reconstruction loss: $L_r = \frac{1}{n} \sum_{i=1}^{n} \| x_i - \psi(\varphi(x_i)) \|^2$
$L_{hybrid} = \alpha L_r + \gamma L_s$

Figure 2: The training of auto-encoder-deconder. Latent space feature vector is used for mortality risk prediction.

2.2.2 Multitask Auto-encoder

Now we have the description of cardiac appearance. To analysis it motion status, the time series of 3D volumes are introduced. For a cycle of heart motion, it includes 20 phases. Considering the first phase as baseline comparison, all other phases' landmark coordinates are subtracted by the coordinates of the first phase. The entire time series can be represented as a $3 \times 19 \times 202$ dimensional feature vector, where 3 indicates coordinates, 19 indicates different phases in time series and 202 indicates the number of landmarks in each volume.

Such high dimensional feature vectors is applied to train a supervised auto-encoder-decoder framework. While the decoder tries to reconstruct the coordinates of those feature vectors in a supervised manner, the actual aim of this progress is to reduce the dimension of the input feature vector. Such a low dimensional latent space feature vector is applied to another multi-layer perceptron to predict mortality supervised on ground-truth labels. By do so, it makes the latent space feature vector have a relatively low dimensional and still remains reconstructive and discriminative. The framework for this stage can be seen in Fig. 2.

2.3 Contribution

The author of this paper proposed a two-stage deep learning framework: (1) 3D MRI volume landmark regression to represent heart information and (2) auto-encoder-decoder framework to reduce feature dimension and enhance discriminative capabilities for mortality prediction.

For 3D subject motion analysis, especially a contraction motion of the heart, it is very difficult for a conventional image analyzing tools to process. Because the motion we discuss here involves the change of shapes but not translation is space.

3 Calcium Scoring

3.1 Motivation

Cardiovascular disease (CVD) is one of the most leading factor of death, and previous researchers have established the fact that coronary artery calcium is a strong indicator of the CVD severity. In conventional ways, the measuring of CAC is tedious and requires extra human efforts. This paper offers a solution to direct perform calcium scoring.

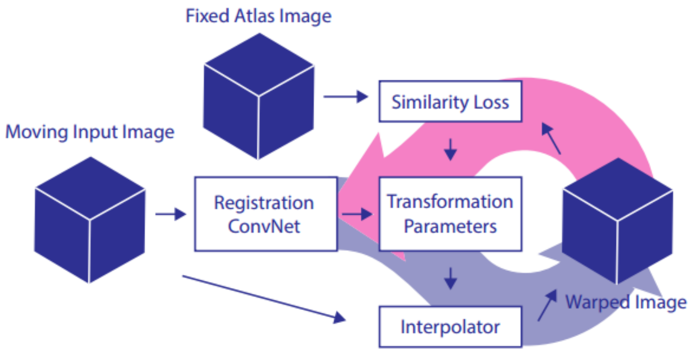

Figure 3: CNN for atlas registration. After this progress, all input CT volumes will be aligned to have the same postion and scale.

3.2 Method

The method employs two ConvNets in sequence (Figure 2). The first ConvNet registers input CTs to an cardiac CT atlasimage. The second ConvNet performs calcium scoring. When desired, visual feedback can be queried for image slices with a score. For this purpose an attention heatmap reveals the regions that contributed to the calcium score.

This paper uses two convolutional neural networks (CNNs) consequtively. The first one tries to register the input 3D CT volumes to an cardiac CT atlasimage. Such an atlasimage is generated by averaging multiple cardiac images that have been manually registers. The aim of the first network is to make all different CT inputs align in the same position, making it easier to compare different patients. The output of the network is spacial transformation matrix that used to register the input volume, and the whole network is trained under the supervision of image similarity metric between input volume and atlas volume. The workflow of the above two CNNs can be seen in Fig. 3.

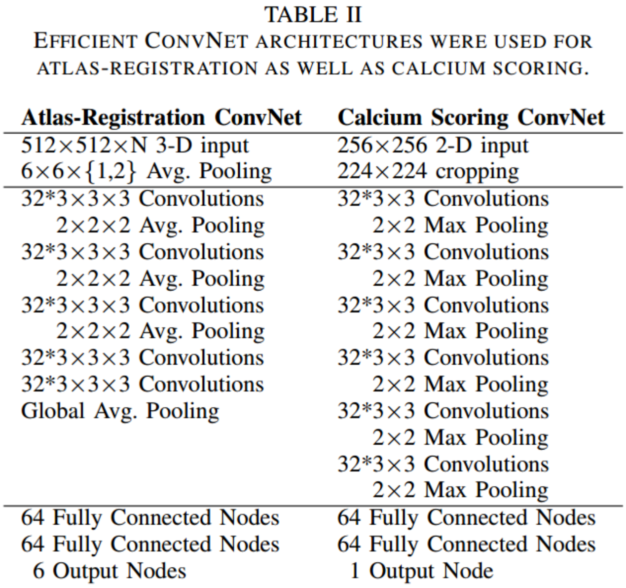

TABLE II
EFFICIENT CONVNET ARCHITECTURES WERE USED FOR ATLAS-REGISTRATION AS WELL AS CALCIUM SCORING.

Atlas-Registration ConvNet	**Calcium Scoring ConvNet**
512×512×N 3-D input	256×256 2-D input
6×6×{1,2} Avg. Pooling	224×224 cropping
32*3×3×3 Convolutions	32*3×3 Convolutions
2×2×2 Avg. Pooling	2×2 Max Pooling
32*3×3×3 Convolutions	32*3×3 Convolutions
2×2×2 Avg. Pooling	2×2 Max Pooling
32*3×3×3 Convolutions	32*3×3 Convolutions
2×2×2 Avg. Pooling	2×2 Max Pooling
32*3×3×3 Convolutions	32*3×3 Convolutions
32*3×3×3 Convolutions	2×2 Max Pooling
Global Avg. Pooling	32*3×3 Convolutions
	2×2 Max Pooling
	32*3×3 Convolutions
	2×2 Max Pooling
64 Fully Connected Nodes	64 Fully Connected Nodes
64 Fully Connected Nodes	64 Fully Connected Nodes
6 Output Nodes	1 Output Node

Figure 4: The detailed architecture of 3D convolutional neural network for CAC score estimation.

The second network directly performs the calcium scoring for the registered input CT volume. By using the 3D convolutional kernels, the network is trained under the supervision of manually give CAC score. The composition of such a network can be seen in Fig. 4.

3.3 Contribution

Overall, the method raised by this paper is a two-phase network. (1) Register all the input volumes to the same scale and position based on the atlas cardiac volume and (2) use a 3D CNN to directly regress the CAC score for such a volume. According to their experiment section, their method achieves similar performance comparing to the state-of-the-art methods, but is hundreds of times faster even when ran on a single core of a CPU. The method can achieve robust and accurate predictions of CAC score in real-time.

4 Mortality Prediction

Comparing to the above two methods which contribute to the mortality risk prediction based on cardiac images, this method is much simpler.

4.1 Motivation

CVD is a leading cause of death in the lung cancer screening population. However, the heart image in real practice is often a volume which is difficult for networks to process. The motivation of this work is to use a much more compressed feature vector to represent the volume image, and then apply this feature vector to perform prediction task for mortality.

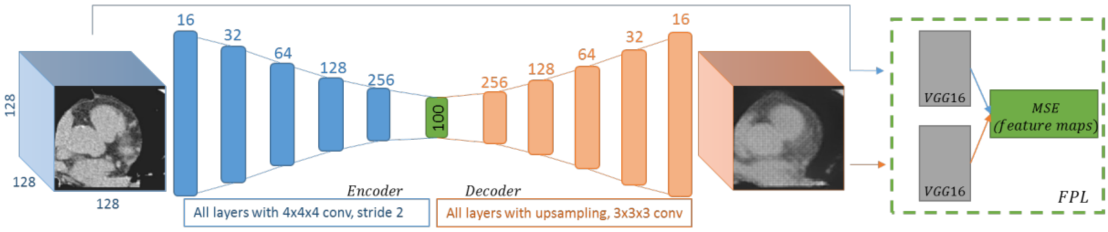

Figure 5: Auto-encoder-decoder framework under the supervision of feature perceptual loss.

4.2 Method

In order to compress the cardiac volume ($128 \times 128 \times 128$) to a low dimensional feature vector, the authors use a convolutional auto-encoder-decoder framework. The output of the encoder is the 100D feature vector which is then fed to decoder for reconstructing the original volume with exactly same size.

The auto-encoder-decoder is training in an end-to-end way under the supervision of feature perceptual loss (FPL). The FPL is computed in this way: the original input image and the reconstructed image are both fed to a fixed VGG-16 network which has been pretrained on ImageNet. The FPL is then defined as the mean square error (MSE) between the feature maps in this network derived from the input image and the reconstructed image. As the VGG16 is orginally designed for 2D images, the loss is actually computed over axial 2D slices of the 3D volume. The workflow of the proposed method is shown in Fig. 5.

When the training of auto-encoder-decoder is completed, we can utilize the latent feature vector produced by encoder. Such 100D feature vector is capable of reconstructing the whole volume, so the author assume it can represent the image information of the original volume. The feature vector is used for trianing another classifier, which can be a multi-layer perceptron, a support vector machine (SVM) or a random forest classifier. The ground truth used here is the surviving label of the subject. Thus, the classifier is able to give mortality risk prediction.

4.3 Contribution

This work first incorporate the FPL loss in the mortality risk prediction task, with the highest experimenting AUC score of 0.72 using the SVM classifier. The overall framework is quite straight forward, but personally speaking the training of auto-encoder-decoder is a difficult task. Moreover, it still remains questionable whether the latent feature vector can well represent all the discriminative information from the original image volume.