Technical report about Cardiac Volume estimation through Regression Vision Transformer on ACDC dataset

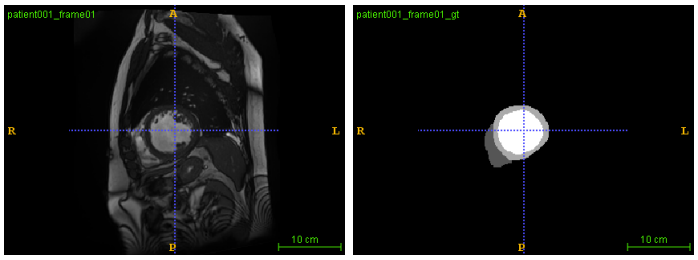
Jing Zhang

November 2021

1 Task description

The task in this work is to predict the volume of cardiac structures using deep learning techniques. Specifically, we want to achieve **a direct multi-class volume estimation model without segmentation intervention.**
This idea is not original, it has been applied in many papers [3, 2].They cover different medical indices prediction and different techniques including traditional machine learning and deep learning in recent decades.

2 Data description



(a) MRI image of a patient (b) Annotations of a patient

Figure 1: One slice of MRI cardiac subject and its ground truth annotated by clinical expert in ACDC dataset. The white color region in (b) is LV (class 3), the light grey region is MYO (label 2), the dark grey region is RV (class 1), the black region is background (class 0), the volume of RV, MYO LV are 139.72 ml, 164.26 ml, 295.51 ml respectively according to Formula 1.

The public "Automatic Cardiac Diagnosis Challenge" dataset (ACDC) dataset [1] is used in this study. It has 100 magnetic resonance images (MRI) subjects in

training set, each subject has 3 manual annotated labels, i.e., left ventricular (LV), myocardium (MYO), right ventricle (RV). Each subject has end diastolic (ED) and end systolic (ES) phase. It also has 4 types of disease, myocardial infarction (MINF), dilated cardiomyopathy (DCM), hypertrophic cardiomyopathy (HCM), abnormal right ventricle (ARV), and patients with normal cardiac (NOR). Fig. 1 is the MRI image of a patient and its ground truth annotated by experts. The volume (ml) of each structure is calculated as below:

$$\text{Volume} = \begin{cases} \text{RV} = \sum_{i}^{N} \text{S\_RV} * \text{px\_x} * \text{px\_y} * \text{space\_z}/1000 \\ \text{MYO} = \sum_{i}^{N} \text{S\_MYO} * \text{px\_x} * \text{px\_y} * \text{space\_z}/1000 \\ \text{LV} = \sum_{i}^{N} \text{S\_LV} * \text{px\_x} * \text{px\_y} * \text{space\_z}/1000 \end{cases} \quad (1)$$

In Formula 1, N is the number of slice in each subject, S_RV is the summary of pixels belong to RV class in one slice (same as S_MYO, S_LV), px_x (mm) is pixel size of x dimension, px_y (mm) is pixel size of y dimenstion, $space\_z$ (mm) is slice thickness.

From Fig. 1 we can tell the features of the dataset are that:
1. The cardiac accounts for a small proportion of the image;
2. The RV and MYO are less clear than the LV;
Besides, according to the author's observation, **for each subject, the shape is different, and the number of slice is also different. Certain slices don't have RV structure or even only have background (black) pixels.** Therefore, it's necessary to preprocess this dataset before using deep learning models on them.

3 Method description

3.1 Attention mechanism

3.2 Self Attention

3.3 Multi-head Attention

3.4 Architecture of Vision Transformer

4 Attention CNN

5 Experiment description

5.1 Data preprocessing

• Data Screening

To make sure the ground truth volumes offered by the authors of ACDC dataset are correct, we check the consistence of given volumes and volumes computed through Formula 1. Among 100 pairs subjects (including ED and ES), the given volumes of cardiac of 6 patients are seriously deviated from the computed volumes, the difference are up to hundred level. See Table 1. While the given volumes in left 94 pairs patients are in normal range, because their difference are less than 1. Thus, we will remove these 6 patients in the experiments.

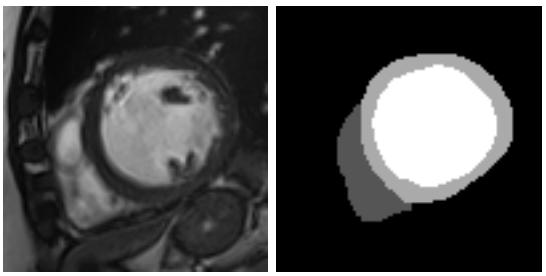
Table 1: Abnormal samples in ACDC dataset, V_g is volume from given data, V_c is volume from calculation (see Formula 1), Diff is the difference between V_g and V_c.

Patient	spacing(x,y,z)	shape	V_g(ml)	S(pixel)	V_c(ml)	**Diff**
P019	(1.445,1.445,10)	(11, 256, 216)	868.59	32269	673.78	194.81
P078	(1.367,1.367,10)	(8, 256, 216)	630.20	25813	482.36	147.84
P079	(1.367,1.367,10)	(9, 256, 216)	455.74	18667	348.83	106.91
P080	(1.758,1.758,10)	(6, 256, 216)	223.95	9173	283.50	59.55
P093	(1.563,1.563,7)	(10, 224, 180)	57.35	23491	401.71	344.35
P099	(1.786,1.786,5)	(16, 224, 154)	866.71	27180	433.49	433.21

• Data Cropping

As Section 2 described, the original data they have different sizes in 3 dimensions, and the cardiac is small in each slice. In order to ensure that there are no basic errors in the model, and to enable the model to fully learn the features of the cardiac, we **crop the data to a fixed size and focus on the cardiac only**. The cropping steps are as follows:

1. Detecting and finding the largest bounding box of cardiac in the whole slices of ground truth subject.
2. Performing Step 1 among all the subjects.
3. Creating a new bounding box that just covers all the detected bounding box of each subject.[1]
4. Cropping the whole ground truths and MRI images based on the new bounding box, the size of a 2D subject is (100,100).
After cropping, the cardiac is preserved to the greatest extent but no other parts compared to Fig. 1, see Fig. 2.



(a) Cropped MRI image (b) Cropped ground truth

Figure 2: One slice of cropped MRI cardiac subject and its ground truth, the size is 100*100.

• Data Resampling

In original data, the number of slices are different from each other (from 6 to 17 slices). Moreover, there are slices that don't have cardiac labels but only have background pixels. To make sure the model does not go wrong and avoid invalid information, we **fix the number of slice of each subject as 9 and remove blank slices**. The final shape of all the subjects will be (100,100,9).
The resampling steps are as follows:
1. Detecting slices that only have background pixels and removing them.
2. Duplicating the tail slice to the bottom of subject until the number of slice comes to 9.
3. Removing the tail slice from the bottom of subject until the number of slice comes to 9.

• Data Normalization & Augmentation

We perform data normalization both in MRI images $((\text{img} - \mu)/\sigma)$ and ground truth volumes $(\text{gt}/\max(\text{gt}))$. To enlarge the training set, data augmentation including flipping, translation and rotation are performed.

[1] In this step, we remove 3 extreme examples because the size of them are largely different from others.

5.2 Configuration

The training set is 2030 MRI images, the validation set is 68, the test set is 60. The image are resized to 72×72. The optimizer is AdamW with weight decay $1e^{-4}$. The learning rate is $1e^{-3}$, the batch size is 16. The algorithm is completed using Python and Tensorflow 2.x and Keras 2.x library with GPU p100. The training epoch is 100. 5-fold cross validation is performed.

5.3 Results

6 Discussion & Conclusion

References

[1] Olivier Bernard, Alain Lalande, Clement Zotti, Frederick Cervenansky, Xin Yang, Pheng-Ann Heng, Irem Cetin, Karim Lekadir, Oscar Camara, Miguel Angel Gonzalez Ballester, et al. Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved? *IEEE transactions on medical imaging*, 37(11):2514–2525, 2018.

[2] Gongning Luo, Suyu Dong, Wei Wang, Kuanquan Wang, Shaodong Cao, Clara Tam, Henggui Zhang, Joanne Howey, Pavlo Ohorodnyk, and Shuo Li. Commensal correlation network between segmentation and direct area estimation for bi-ventricle quantification. *Medical image analysis*, 59:101591, 2020.

[3] Xiantong Zhen, Zhijie Wang, Ali Islam, Mousumi Bhaduri, Ian Chan, and Shuo Li. Direct estimation of cardiac bi-ventricular volumes with regression forests. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 586–593. Springer, 2014.