Estimation of Left Ventricular Ejection Fraction Based on Swin_Uniform

Dongsheng Qi [1, *], Yanfen Zhang [1]

[1] School of Computer and Information Engineering, Institute for Artificial Intelligence, Shanghai Polytechnic University, Shanghai 202309, China.

[*] 15034927509@163.com

Abstract. A new left ventricular ejection index prediction method was proposed by introducing a block attention mechanism, which helps to reduce overfitting problems and improve the power of the model to process new different samples. Swin Transformer is a Transformer architecture that retains the modeling advantages of self attention mechanism. Its hierarchical design has both global and local modeling capabilities, and can output multi resolution feature maps, which is suitable for EchoCoTr's need to extract multi-scale information from time series medical images. Meanwhile, the resources required of Swin Transformer at runtime are linearly related to images to be processed, which reduces computational complexity and is suitable for processing high-resolution sequence images in EchoCoTr[1]. Swin Transformer has better results than Transformer in both classification and detection tasks, which just goes to show that it is somewhat of a great application in the field of vision. This is beneficial for EchoCoTr's understanding of cardiac ultrasound sequences. Therefore, in the original UniFormer model structure of the EchoCoTr[1] model, a new Swin Transformer Block is added. The new model has stronger modeling capabilities. This means that better results can be obtained under the same computing resources.

Keywords: Attention mechanisms; Swin Transformer; Echocardiography; Ejection fraction.

1. Introduction

The heart is an essential organ among our internal organs. The proper functioning of the heart is essential for the circulation of blood in our body, and only when we have a healthy body will we be able to handle our daily affairs, so monitoring the function of the heart has always been a popular area of research. The heart is located above the diaphragm in the chest cavity, like an inverted, slightly flattened cone[2]. The main structure of the heart consists of four chambers, two sets of atrioventricular valves, and two sets of semilunar valves. Among them, the left ventricle is the leftmost and bottom of the four chambers, located at the left rear of the right ventricle. The importance of proper cardiac function has become clear, if the heart is not functioning properly it can lead to a variety of different degrees of physical discomfort, and in severe cases it can affect a person's normal physiological activities, the study of heart failure has also received more and more attention from a wide range of researchers in recent years, and this is where the left ventricular ejection index (LVEI) can be used to quantify the heart's function, which calculates the ratio of left ventricular end-systolic to end-diastolic volume[3].

In the field of medical imaging of heart function, echocardiography is widely used, which can store both temporal and spatial information, and what needs to be done in making medical diagnosis is to analyze and judge these temporal and spatial information. In order to minimize the observation error generated by human observation, we believe that deep learning can help to solve this problem, because the diastole and contraction of the heart has a cycle pattern, and the machine may have more accurate cycle observation and prediction ability than a human being[4]. So in using deep learning to assist the diagnostic process of echocardiography, we have done further research based on our predecessors.

In this article, we introduce the modeling advantages of self attention mechanism by improving the EchoCoTr[1] model with Swin Transformer. In the original UniFormer model structure of the EchoCoTr[1] model, a new Swin Transformer Block has been added. The new model has stronger modeling capabilities, and Swin Transformer introduces a hierarchical attention mechanism, which

can achieve better modeling results while maintaining efficiency. By changing global attention to local attention, we were able to increase the speed of the model's computation, reduce the time complexity of the computation and avoid overfitting as much as possible. We used the same dynamic dataset as in the previous study to demonstrate the effectiveness of the model improvement, and the results proved to be a definite improvement in our experimental results. The paper is organized as follows: Section I is an introduction to the study. Section II is devoted to model structure. Section III is devoted to modeling techniques. Section IV is devoted to experimental comparisons. Section V is a summary of the present paper.System Hardware and Overall Design[5].

2. Model Structure

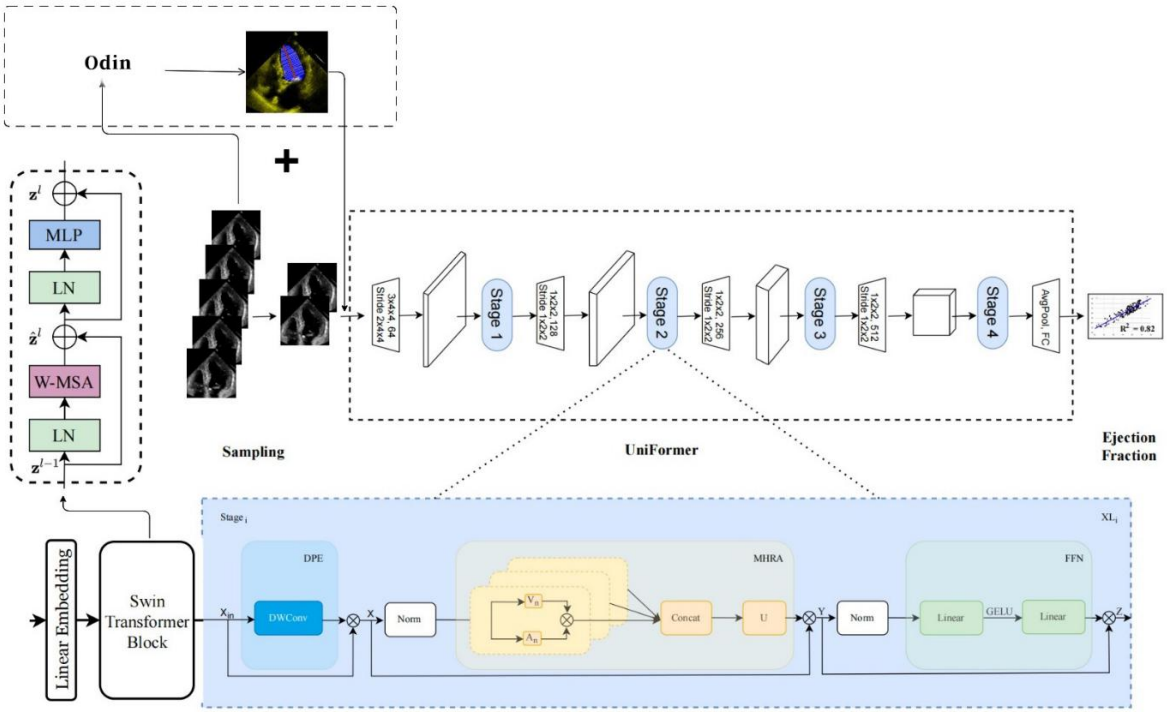

Fig. 1 Technology Roadmap

2.1 Blending and broadening of raw video data using the Odin model

The EchoCoTr[1] model, which requires learning information in both temporal and spatial dimensions, is computationally intensive and does not guarantee accuracy because it uses unlabeled echocardiographic datasets for prediction.The Odin model generates object segmentation by K-means clustering of the image features without any a priori knowledge. The segmentation generated by the object discovery network is used to get feature representations that able to distinguish between different objects in an adversarial learning framework. The two networks collaborate and iteratively optimize each other, and the updates of the object representation network are continuously used to optimize the object discovery network through exponential moving averages, which ultimately leads to the successful segmentation of the objects[6]. The segmented image and video information of the Odin model is added as augmentation data to enrich the training samples. The addition of learning features, such as the left ventricular volume segmented by the Odin model, improves the robustness of the model to different variations and accelerates the model to better converge to the optimal solution[7].

2.2 Improving the EchoCoTr model with Swin Transformer

Swin Transformer is a Transformer architecture that retains the modeling advantages of the self-attention mechanism. Its hierarchical design combines global and local modeling capabilities. Swin Transformer can output multi-resolution feature maps, which is suitable for EchoCoTr[1] to

extract multi-scale information on time-series medical images. Meanwhile, Swin Transformer at runtime are linearly related to images to be processed, which reduces computational complexity and is suitable for processing high-resolution sequence images in EchoCoTr[1]. Swin Transformer outperforms the standard Transformer in tasks such as image classification and target detection, which indicates that it is more capable of modeling vision. This facilitates EchoCoTr's understanding of cardiac ultrasound sequences. Therefore, a new Swin Transformer Block is added to the original UniFormer model structure of the EchoCoTr[1] model.The new model has a stronger modeling ability, and Swin Transformer introduces a hierarchical attention mechanism, which can achieve better modeling results while maintaining high efficiency[8]. This means that better results can be obtained with the same computational resources. By introducing the chunked attention mechanism, it helps to reduce overfitting problems and improve the power of the model to process new different samples

3. Modeling Techniques

3.1 Odin model

Odin (Object discovery and representation networks) is a self-supervised representation learning framework proposed by Henaff et al. in 2022.The innovation of Odin lies in the coupling of object discovery and object representation learning to form a closed-loop learning process. Specifically, Odin consists of two components: an object discovery network and an representation network. The object discovery network inputs the feature representations of images and generates semantic segmentations of objects through K-Means clustering; while the object representation network uses these segmentations as supervisory signals, and learns the feature representations that can distinguish different objects under the framework of adversarial learning. Eventually, the semantic features learned by the object representation network are fed back to the object discovery network to produce more fine-grained object segmentation. In this way, the object discovery and representation networks are driven by each other, realizing the self-supervised learning process of discovering semantic knowledge of objects from data. In this work, the authors use only simple K-Means clustering to discover semantic object structures from images without relying on any artificially constructed a priori knowledge, and validate the effectiveness of the representation on multiple downstream tasks. This framework is designed to be simple and efficient, and provides a new way of thinking in the field of self-supervision[9].

3.2 Swin Transformer

Swin Transformer, a new vision Transformer model, can be used as a generalized computer vision backbone network. Swin Transformer architecture, which constructs hierarchical feature representations and helps to reduce overfitting problems and improve the power of the model to process new different samples through the shifted window mechanism, proposes an efficient Cyclic shifting is proposed to realize the batch processing of shifted window, which greatly reduces the amount of computation. Swin Transformer has the ability of multi-scale representation, which can be well applied to vision samples[10]. The backbone network outputs multi-resolution feature maps similar to CNN, which is convenient for accessing downstream tasks. The results are validated on ImageNet image classification, COCO target detection, and ADE20K semantic segmentation, and the results outperform those of CNN and previous Transformer models.

4. Experimental Comparisons

4.1 Introduction to the experimental dataset

The EchoNet-Dynamic dataset consists of 10,030 videos collected at Stanford University Hospital and is a publicly available echocardiography dataset that is useful for analyzing

myocardial function using deep learning. This is a dataset that can be used for model training and evaluation with a resolution of 112×112 and gives end-diastolic and end-systolic information of the myocardium for the video samples. Information on the left ventricular ejection index is also given. We still use the dataset allocation ratio used by previous authors, and 7460 out of 10030 video data are used as the training dataset for this experiment. Since the videos vary in length, each cardiac cycle contains about 20-30 frames, and there is little variation between neighboring frames, these characteristics make it an ideal dataset for evaluating time-series medical image-based methods. This experiment also continues to follow the previous protocol by using 1288 data for validation and 1277 data for testing .

4.2 Results

The evaluation metrics we used in this experiment are training set loss function (Train Loss), training set coefficient of determination (Train r^2), validation set loss function (Val loss) and validation set coefficient of determination (Val r^2). These metrics are able to fairly compare the effectiveness of this experiment with the comparison experiment, which was run for a total of 49 batches. The training loss function is used to express the fitting ability of the model when it is used on the training dataset, while the validation loss function is used to express the processing and adaptation ability of the model when it is used on the validation dataset. The coefficient of determination reflects the fit of the model, and the closer the coefficient of determination of the model is to 1, the better the model is.

As shown in Table 1, our Swin_Uniform model and the original EchoCoTr[1] were also trained on 49 batches, achieving better recent results than those reported by the original model. From the results, it can also be noted that the Swin_Uniform experiment achieved (32.786 Train loss; 34.222 Val loss; 0.787 Train r^2; 0.791 Val r^2) better results than EchoCoTr[1] (33.076 Train loss; 34.483 Val loss; 0.785 Train r^2; 0.772 Val r^2) performed slightly better.

Table 1. Experimental results

Model	Train loss	Train r^2	Val loss	Val r^2
Swin_Uniform	32.786	0.787	34.222	0.791
EchoCoTr	33.076	0.785	34.483	0.772

5. Summary

The learning of spatio-temporal features is critical for handling samples with both temporal and spatial features, since a change in any one of the features can lead to different results. Swin Transformer is a Transformer architecture that retains the modeling advantages of the self-attention mechanism. Its hierarchical design combines both global and local modeling capabilities. Swin Transformer can output multi-resolution feature maps, which is suitable for EchoCoTr[1] to extract multi-scale information on time-series medical images. Meanwhile, the resources required of Swin Transformer at runtime are linearly related to images to be processed, which reduces the computation amount and is suitable for processing high-resolution sequence images in EchoCoTr[1]. Swin Transformer outperforms the standard Transformer in tasks such as image classification and target detection, which indicates that it is more capable of modeling vision. This facilitates EchoCoTr's understanding of cardiac ultrasound sequences. Therefore, a new Swin Transformer Block is added to the original UniFormer model structure of the EchoCoTr[1] model. The new model has a stronger modeling ability, and Swin Transformer introduces a hierarchical attention mechanism, which can achieve better modeling results while maintaining high efficiency. This means that better results can be obtained with the same computational resources. By introducing the chunked attention mechanism, it helps to reduce the overfitting problem and enhance the model's ability to handle data from different samples.

References

[1] Muhtaseb R, Yaqub M. EchoCoTr: Estimation of the Left Ventricular Ejection Fraction from Spatiotemporal Echocardiography//International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham: Springer Nature Switzerland, 2022: 370-379.

[2] Sun X, Cheng L H, van der Geest R J. Self-and Cross-attention based Transformer for left ventricle segmentation in 4D flow MRI//Medical Imaging with Deep Learning. 2022.

[3] Stratos I, Behrendt A K, Anselm C, et al. Inhibition of TNF-α restores muscle force, inhibits inflammation, and reduces apoptosis of traumatized skeletal muscles. Cells, 2022, 11(15): 2397.

[4] Zeng Y, Tsui P H, Pang K, et al. MAEF-Net: Multi-attention efficient feature fusion network for left ventricular segmentation and quantitative analysis in two-dimensional echocardiography. Ultrasonics, 2023, 127: 106855.

[5] Deng K, Meng Y, Gao D, et al. Transbridge: A lightweight Transformer for left ventricle segmentation in echocardiography//Simplifying Medical Ultrasound: Second International Workshop, ASMUS 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 27, 2021, Proceedings 2. Springer International Publishing, 2021: 63-72.

[6] Vafaeezadeh M, Behnam H, Hosseinsabet A, et al. A deep learning approach for the automatic recognition of prosthetic mitral valve in echocardiographic images. Computers in Biology and Medicine, 2021, 133: 104388.

[7] Mokhtari M, Tsang T, Abolmaesumi P, et al. EchoGNN: Explainable Ejection Fraction Estimation with Graph Neural Networks//Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part IV. Cham: Springer Nature Switzerland, 2022: 360-369.

[8] Ruschel K B, Rabelo E R, Brun A O, et al. Early Recovery of Venous Endothelial Dysfunction in Decompensated Congestive Heart Failure. Journal of Cardiac Failure, 2006, 12(6): S18.

[9] Harris S, Dhinoja M. Implantable cardioverter defibrillators. Clinical medicine, 2007, 7(4): 397.

[10] Sun J Y, Qiu Y, Guo H C, et al. A method to screen left ventricular dysfunction through ECG based on convolutional neural network. Journal of Cardiovascular Electrophysiology, 2021, 32(4): 1095-1102.

[11] Kusunose K, Haga A, Inoue M, et al. Clinically feasible and accurate view classification of echocardiographic images using deep learning. Biomolecules, 2020, 10(5): 665.

[12] Dai W, Li X, Chiu W H K, et al. Adaptive contrast for image regression in computer-aided disease assessment. IEEE Transactions on Medical Imaging, 2021, 41(5): 1255-1268.

[13] Vaid A, Jiang J, Sawant A, et al. HeartBEiT: Vision Transformer for Electrocardiogram Data Improves Diagnostic Performance at Low Sample Sizes. arXiv preprint arXiv:2212.14040, 2022.

[14] Mazhari R, Schuleri K H, Zimmet J M, et al. Cell Tracking Following the Intramyocardial Injection of Mesenchymal Cells after Myocardial Infarction. Journal of Cardiac Failure, 2006, 12(6): S18.

[15] Hénaff O J, Koppula S, Shelhamer E, et al. Object discovery and representation networks//Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVII. Cham: Springer Nature Switzerland, 2022: 123-143.

[16] Kim W, Son B, Kim I. Vilt: Vision-and-language Transformer without convolution or region supervision//International Conference on Machine Learning. PMLR, 2021: 5583-5594.