2022 44th Annual International Conference of
the IEEE Engineering in Medicine & Biology Society (EMBC)
Scottish Event Campus, Glasgow, UK, July 11-15, 2022

2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) | 978-1-7281-2782-8/22/$31.00 ©2022 IEEE | DOI: 10.1109/EMBC48229.2022.9871809

Vision Transformers for Classification of Breast Ultrasound Images

Behnaz Gheflati[1] and Hassan Rivaz[1]

Abstract— **Medical ultrasound (US) imaging has become a prominent modality for breast cancer imaging due to its ease of use, low cost, and safety. In the past decade, convolutional neural networks (CNNs) have emerged as the method of choice in vision applications and have shown excellent potential in the automatic classification of US images. Despite their success, their restricted local receptive field limits their ability to learn global context information. Recently, Vision Transformer (ViT) designs, based on self-attention between image patches, have shown great potential to be an alternative to CNNs. In this study, for the first time, we utilize ViT to classify breast US images using different augmentation strategies. We also adopted a weighted cross-entropy loss function since breast ultrasound datasets are often imbalanced. The results are provided as classification accuracy and Area Under the Curve (AUC) metrics, and the performance is compared with the SOTA CNNs. The results indicate that the ViT models have comparable efficiency with or even better than the CNNs in the classification of US breast images.**

Clinical relevance— **This work shows the potential of Vision Transformers in the automatic classification of masses in breast ultrasound, which helps clinicians diagnose and make treatment decisions more precisely.**

I. INTRODUCTION

Breast cancer is the most frequent cause of cancer mortality among women, making annual breast cancer screening essential for early detection and reducing the death rate. In recent years, ultrasound (US) imaging has become one of the most promising modalities due to its availability, real-time display, cost-efficiency, and non-invasive nature. Over the previous decades, US imaging has shown great potential in automated breast lesion classification, segmentation, and detection tasks. Automated analysis of medical images will help the radiologist to detect masses in US images and reduce the number of false negative readings in a cost-efficient way [1], [2].

Deep learning has lately become a leading tool in various research domains. Convolutional Neural Networks (CNNs) have been the most common networks for automatic medical image analysis applications such as image classification in recent years. However, due to their localized receptive fields, these models have a poor performance in learning the long-range information, limiting their capabilities for vision tasks [3].

The Transformer architecture, provided by Vaswani *et al.* [4] is currently the dominant model in the field of natural language processing (NLP). Motivated by the success of the self-attention based deep neural networks of Transformer models in NLP, Dosovitskiy *et al.* [5] introduced the Vision Transformer (ViT) architecture for the image classification application. The overall training process in these models is based on splitting the input image into patches and treating each embedded patch as a word in NLP. These models use self-attention modules to learn the relation between these embedded patches [3].

Herein, we take a step in exploiting Transformers in US medical image analysis and explore the potential application of self-attention to classify breast US images. Our contributions are:

- We compared different pre-trained ViT models with different sizes and configurations based on their performance when fine-tuning for a downstream task.
- We transfer these models to the classification of breast US images.
- We fine-tune the state-of-the-art (SOTA) CNN networks to classify US images for our dataset.
- We adopted a weighted cross-entropy loss function to deal with the imbalanced dataset problem [6].

The results show the great potential of ViT models in breast US image classification. To the best of our knowledge, this is the first study to investigate the performance of ViT architectures on the classification of US breast images.

II. RELATED WORK

A. CNN-based Classification Networks

During the past decade, CNNs have been used as a standard technique for medical image classification applications. Several studies have been conducted to benefit from the freely available CNN models pre-trained on a large amount of training data and transferring the pre-trained models for US image classification. In a comparative study conducted by Lazo *et al.* [7], they studied using transfer learning (TL) for VGG-16 and IncpectionV3 to detect lesions in breast US images. Another study [1], compared the performance of TL based on VGG16, ResNet, Inception, and NASNet for breast masses classification in US images and showed the higher performance of the NASNET compared to the other CNNs.

Despite the success of CNNs in image processing applications, ViT models have shown superior performance to CNNs. The reason is that they lack some of the inductive biases of CNNs, such as the translation equivalence property of local convolutions.

*This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC).

[1]Behnaz Gheflati and Hassan Rivaz are with Department of Electrical and Computer Engineering, Concordia University, Montreal, Canada B_Ghefla@encs.concordia.ca and hrivaz@ece.concordia.ca

B. Vision Transformer (ViT)

Vision Transformers, firstly introduced by Dosovitskiy *et al.* [5] have shown outperforming performance compared to SOTA CNNs in image classification applications when trained on a large scale training dataset. Despite the weaker inductive bias of Transformers than CNNs, these networks show competitive results with SOTA CNNs. Nevertheless, the high demand for large amounts of training data and computational resources limits the modification of these networks. To tackle this problem, Touvron *et al.*[8] introduced a data-efficient ViT, based on the data augmentation and regularization techniques previously employed for CNNs. In addition, they improved the performance of ViT by using a Transformer-based teacher-student approach for the image classification task.

Considering the high performance of ViTs, many researchers have studied ViT models in various vision tasks [9]. In the area of object detection, Carion *et al.* [10] proposed a new architecture for object detection systems using a set-based global loss and a Transformer encoder-decoder algorithm and showed on par results with the dominant R-CNN method on the challenging COCO dataset.

In image segmentation, a free convolution network, purely based on the attention-based algorithms, was provided for 3D medical image segmentation [2]. Another recent study [3] has also leveraged the Transformer's power in 3D medical image segmentation, providing a novel U-Net Transformer architecture employing the advantages of both U-Net and Transformer networks in image segmentation and global attention feature of Transformers. In their work, a Transformer and a CNN-based architecture are used as encoder and decoder, respectively. In another study [11], by Dai *et al.* to employ both CNN localized receptive fields for low-level feature extractions and large-scale attention of Transformers, an algorithm composed of both CNNs and Transformers has been provided to classify multi-modal images. They showed that this strategy outperforms other SOTA CNN-based networks.

Although many efforts have been made to improve the ViT-based models, and many ViT models pre-trained on large-scale datasets are freely available, there is still a question about choosing a pre-trained model for TL. There are many choices regarding model size and configurations to select a model and fine-tune the weights based on the new dataset.

As shown in [12], pre-training of a ViT model can be performed under various settings on different datasets, and any of these settings result in different performances. In this regard, Steiner *et al.* provide a large number of pre-trained ViT models with different sizes and also hybrids with ResNets on the various dataset sizes, ImageNet-1k and Imagenet-21k. Their suggestion is to select a few pre-trained models with high performance when fine-tuning for an upstream task as *recommended models*. Then, instead of adapting all the pre-trained Transformers, which is an extensive task, one can choose the model from these *recommended models* with the best performance for further adaptation.

III. METHOD

The overall procedure of our work in this paper is based on the method provided in [12], discussed in the previous section. We transferred the pre-trained ViT models, *recommended models*, and adapted the best one based on our specific data task.

A. Dataset and Evaluation Metrics

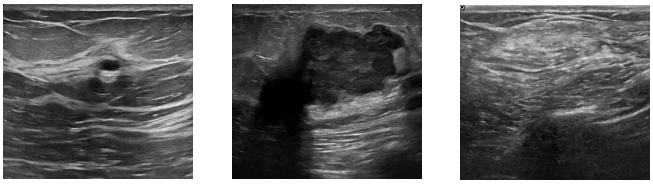
The data used in this study includes two different datasets on breast US images. The first dataset is published online by Al-Dhabyani *et al.* [13], which has 780 breast US images (referenced as BUSI), collected from 600 women with an average image size of 500 x 500 pixels. The dataset consists of 133 normal images, 437 malignant masses, and 210 benign tumors. The second dataset, considered as dataset B [14], includes 163 images, with an average size of 760 x 570 pixels, categorized into two classes, 110 benign masses and 53 cancerous masses. Examples of breast US images are shown in Figure 1. All the training images are resized to 224 x 224. The performance metrics we use for evaluation purposes are common metrics employed in medical image classification studies [7], including classification accuracy (Acc) and area under the receiver operating characteristic curve (AUC).

B. ViT Architecture

Following the procedure suggested in [12], the established architecture is the same as the original ViT design introduced by [5], except for substituting the MLP head with a linear classifier. An overview of the ViT design architecture is presented in Figure 2. In summary, the input image is split into patches in a ViT model. A sequence of 1D patch embeddings is fed to the Transformer encoder, where self-attention modules are utilized to calculate the relation-based weighted sum of the outputs of each hidden layer. Consequently, this strategy allows the Transformers to learn global dependencies in the input images.

C. Models

For transfer learning of CNNs, we use the primarily pre-trained SOTA CNN networks in lesion classification of breast US images, including VGG16, ResNet50, InceptionV3, and NASNetLarge, based on the breast US dataset. Consequently, we compare the CNN-based and ViT-based architectures results.



(b) Benign (b) Malignant (c) Normal

Fig. 1. Example of breast US images with three different classifications.

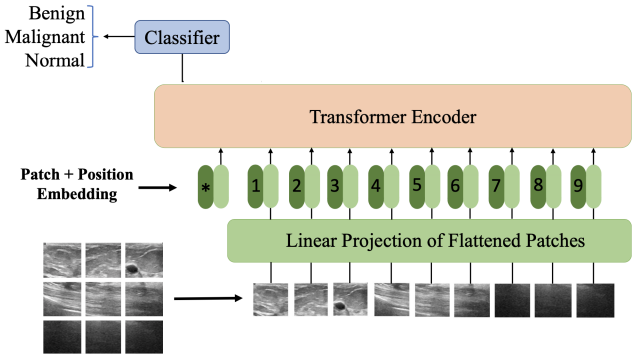


Fig. 2. Overview of the vision Transformer used in classification of breast US.

Furthermore, the TL approach for Transformer-based models is based on the *recommended models* provided by [12], including the main ViT models with different sizes from [5], [8] (ViT-Ti, ViT-S, ViT-B), and also the hybrid ViT+ResNet models (R+Ti, R26+S) from [12], with the specifications provided in Table 1 and Table 2 of [12]. In these models Ti, S, B are the representations of tiny, small and based models [5], and the models including R, represents the ViT+ResNet hybrid models.

D. Fine-tuning Details

We perform a three-classes classification task of benign, malignant or normal. We split all datasets to 70%, 15% and 15% for training, validation and testing, respectively. All experiments are conducted on the fixed training and testing dataset for comparison purposes. The 5-fold cross-validation (CV) is used for the final evaluation.

In CNN TL experiments, the output layer is altered with a classifier with the softmax activation function. The optimizer used is the Adam, and the models are trained for 30 epochs (with early stopping to prevent network from overfitting).

For ViT models, fine-tuning is performed based on the code provided in [12], with some pre-processing modifications based on the unique physical characteristics of the US images. The loss function we use is a weighted cross-entropy loss to handle our imbalanced training dataset, optimized by stochastic gradient descent (SGD) with 0.9 for momentum. For fine-tuning, a batch size of 128 and a cosine decay learning rate starting from 0.001 with 10 linear warmup steps are selected, and the algorithm is run for 250 steps (with early stopping).

A summary of our method is as follows:

- Classification of BUSI+B dataset using TL based on CNN models.
- Classification of BUSI+B dataset using TL based on ViT-based networks.
- Classification of augmented BUSI+B dataset using TL based on ViT-based networks.
- Classification of three B, BUSI, and BUSI+B datasets using TL based on two ViT-based and CNN networks.

IV. RESULTS

We tested the CNNs and ViT models for transferring pre-trained models to classify breast US images into three categories: benign, malignant, and normal.

A. Convolutional Neural Network (CNN) Models

The classification accuracy (Acc) and AUC results for different CNN models on the breast US dataset are reported in Table I. This table demonstrates that among different SOTA CNN networks for classification of breast masses in US images, including ResNet50, VGG16, Inception, and NASNET, employing TL based on the ResNet50 model has the best results with 85.3% Acc and 0.95 AUC.

B. ViT-based Models

Table II shows the classification results for ViT models using TL. This table illustrates the achievement of transferring pre-trained self-attention based models on breast US images classification with more than 85% Acc for all ViT models. The best accuracy and AUC are 86.7% and 0.95 for B/32 model, respectively.

The classification performance obtained from the TL of ViT models (Table II) and CNN models (Table I) shows comparable or even better results for ViT models than the corresponding results for the SOTA CNN networks. According to these tables, the best results of CNN networks (ResNet) are comparable with the results of ViT models, whereas for the other CNN models (VGG, Inception, and NASNET), the ViT networks have a better performance. These findings indicate the representation power of attention-based models in the area of US images analysis.

The interesting point in our results is the small size of the dataset used for pre-training compared to the larger size of the dataset used for fine-tuning in [12], which shows the potential power of attention-based models in medical US images analysis. The possible reason for the effectiveness of ViT models on such a small US dataset might be that, unlike natural images, the relation between spatial information or more specifically, the large-scale dependencies between different patches are much more explicit in US images.

TABLE I
COMPARING THE PERFORMANCE OF CNN MODELS USING TL ON THE CLASSIFICATION OF BUSI+B DATASET.

Evaluation	ResNet	VGG	Inception	NASNET
ACC	**85.3%**	82%	80%	79%
AUC	**0.94**	0.92	0.92	0.917

TABLE II
COMPARING THE PERFORMANCE OF *recommended* VIT MODELS USING TL ON THE CLASSIFICATION OF BUSI+B DATASET.

Evaluation	R+Ti/16	S/32	B/32	Ti/16	R26+S/16
ACC	85.7%	86%	**86.7%**	85%	86.4%
AUC	0.94	0.95	**0.95**	0.94	0.95

The important observation is that the results for different ViT architectures are almost the same, about 86% Acc and 0.95 corresponding AUC. This similarity between different model outcomes indicates that a smaller model with a fewer number of parameters would be more beneficial in terms of fine-tuning duration and the computational cost. Therefore, we choose B/32, a relatively small ViT architecture with the best Acc and AUC in the rest of our experiments.

C. Data Augmentation

In the next step, we examined the effect of using data augmentation techniques on the classification performance when fine-tuning the ViT models. Due to the special physical characteristics of US images, the augmentation techniques that do not change the physical characteristics of the US images are limited. In particular, we use light cropping, rotation, brightness, and contrast. As presented in Table III, consistent with the results presented in Figure 1 of [12], augmentation does not make any improvement in transferring pre-trained ViT models.

D. Breast US Datasets

We also tested the B/32 and Resnet50 models on the classification of different datasets: datasets B, BUSI, and BISI+B. The results are mentioned in Table IV, in which for both models, the Acc and AUC or dataset BUSI+B outperform the corresponding results for datasets BUSI and B.

Based on Table IV, the best results for both networks are observed when both datasets, BUSI+B, are used for fine-tuning. This is because more training data leads to a better generalization of the model. Also, the difference between the results of BUSI and BUSI+B is much less than that of datasets B and BUSI+B, considering the smaller size of dataset B with just two categories of breast US images, malignant vs. benign.

TABLE III

COMPARING THE PERFORMANCE OF *recommended* VIT MODELS USING TL ON THE CLASSIFICATION OF BUSI+B DATASET WITH AUGMENTATION.

Evaluation	R+Ti/16	S/32	B/32	Ti/16	R26+S/16
ACC	82%	82%	81%	78%	80%
AUC	0.92	0.92	0.91	0.9	0.91

TABLE IV

COMPARING THE PERFORMANCE OF B/32 AND RESNET MODELS USING TL ON THE CLASSIFICATION OF BUSI, B, AND BUSI+B DATASETS (BEST VALUES IN BOLD FONT).

Dataset	Evaluation	B/32	ResNet
B	Acc	74%	79%
	AUC	0.8	0.84
BUSI	Acc	82%	83%
	AUC	0.91	0.92
BUSI+B	Acc	**86.7%**	**85.3%**
	AUC	**0.95**	**0.94**

V. CONCLUSIONS

This study shows the potential of ViT models in US image classification and, therefore, the effectiveness of learning the global anatomical dependencies of US medical images. The results presented in this study open new windows for using self-attention based architectures as an alternative to CNNs in various US medical image analysis tasks.

ACKNOWLEDGMENT

We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC), and thank NVIDIA for the donation of the GPU.

REFERENCES

[1] W. Al-Dhabyani, M. Gomaa, H. Khaled, and F. Aly, "Deep learning approaches for data augmentation and classification of breast masses using ultrasound images," *Int. J. Adv. Comput. Sci. Appl*, vol. 10, no. 5, pp. 1–11, 2019.

[2] D. Karimi, S. Vasylechko, and A. Gholipour, "Convolution-free medical image segmentation using transformers," *arXiv preprint arXiv:2102.13645*, 2021.

[3] A. Hatamizadeh, D. Yang, H. Roth, and D. Xu, "Unetr: Transformers for 3d medical image segmentation," *arXiv preprint arXiv:2103.10504*, 2021.

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[5] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[6] Y. S. Aurelio, G. M. de Almeida, C. L. de Castro, and A. P. Braga, "Learning from imbalanced data sets with weighted cross-entropy function," *Neural processing letters*, vol. 50, no. 2, pp. 1937–1949, 2019.

[7] J. F. Lazo, S. Moccia, E. Frontoni, and E. De Momi, "Comparison of different cnns for breast tumor classification from ultrasound images," *arXiv preprint arXiv:2012.14517*, 2020.

[8] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *International Conference on Machine Learning*. PMLR, 2021, pp. 10 347–10 357.

[9] S. Yu, K. Ma, Q. Bi, C. Bian, M. Ning, N. He, Y. Li, H. Liu, and Y. Zheng, "Mil-vt: Multiple instance learning enhanced vision transformer for fundus image classification," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2021, pp. 45–54.

[10] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European Conference on Computer Vision*. Springer, 2020, pp. 213–229.

[11] Y. Dai, Y. Gao, and F. Liu, "Transmed: Transformers advance multi-modal medical image classification," *Diagnostics*, vol. 11, no. 8, p. 1384, 2021.

[12] A. Steiner, A. Kolesnikov, X. Zhai, R. Wightman, J. Uszkoreit, and L. Beyer, "How to train your vit? data, augmentation, and regularization in vision transformers," *arXiv preprint arXiv:2106.10270*, 2021.

[13] W. Al-Dhabyani, M. Gomaa, H. Khaled, and A. Fahmy, "Dataset of breast ultrasound images," *Data in brief*, vol. 28, p. 104863, 2020.

[14] M. H. Yap, G. Pons, J. Martí, S. Ganau, M. Sentís, R. Zwiggelaar, A. K. Davison, and R. Marti, "Automated breast ultrasound lesions detection using convolutional neural networks," *IEEE journal of biomedical and health informatics*, vol. 22, no. 4, pp. 1218–1226, 2017.