THE UNIVERSITY OF ADELAIDE
Australian Institute for Machine Learning (AIML)
Robinson Research Institute (RRI)

Learning Subjective Image Quality Assessment for Transvaginal Ultrasound Scans from Multi-Annotator Labels (ID #149)

Daniel Petashvili, Hu Wang, Alison Deslandes, Jodie Avery, George Condous, Gustavo Carneiro, Louise Hull, Hsiang-Ting Chen

Abstract

This paper proposes a novel AI model that automatically assesses the quality of transvaginal ultrasound (TVUS) images, offering support to sonographers, especially those still learning, in acquiring high-quality scans for gynecological pathology diagnosis. Addressing the challenge of varying interpretations by different medical professionals, this model approaches the issue as a multi-annotator noisy label problem. Our novel machine learning architecture first aggregates quality assessments from multiple raters using a weighted ensemble algorithm to estimate consensus labels. The model then employs a multi-axis vision transformer to enhance the process of image quality evaluation. We evaluated the model on a new multi-annotator TVUS dataset, where our model successfully predicted image quality with an accuracy of 80%. This development represents an exciting first step in empowering sonographers to assess scan quality on the spot, reduce the need for repeated imaging, and improve the diagnosis of gynecological pathology.

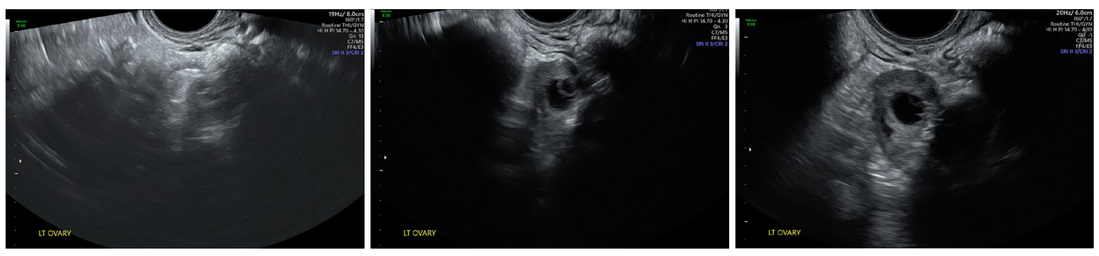

Figure 1: TVUS images of each image quality grade. The anatomy in the Grade 1 image is occluded and not clearly recognisable. In the Grade 2 image, the target anatomy feature's are more prominent. The Grade 3 image anatomy is confidently recognisable with high image clarity.

Contributions

This paper represents the first exploration of multi-annotator subjective image quality assessment (IQA) for TVUS scans. This paper offers two principal contributions:

- First, we introduce and implement a novel approach for training an AI model on subjective IQA using a dataset annotated by multiple annotators. This method is designed to leverage the diverse perspectives of various annotators to enhance the model's assessment capabilities.
- Second, we establish the first dataset for multi-annotated subjective IQA specifically tailored for TVUS scans, facilitating the diagnosis of endometriosis. This dataset is a pioneering resource in the field. The encouraging outcomes highlight the model's potential to assist sonographers in capturing high-quality TVUS images, which in turn can provide invaluable support to gynecologists in the accurate diagnosis of endometriosis.

Methodology

The method consists of 3 distinct aspects: a relabeling model, a weighted ensemble algorithm and a quality prediction model. The relabeling model is pre-trained on ImageNet-1K and then finetuned using on a dataset of multiannotated TVUS images. The weighted ensemble algorithm uses the class prediction probabilities from the relabeling model and the multiannotator's labels to improve the consensus labels of each sample. Finally, a quality prediction model is pretrained on ImageNet-1K and then finetuned using the improved consensus labels.

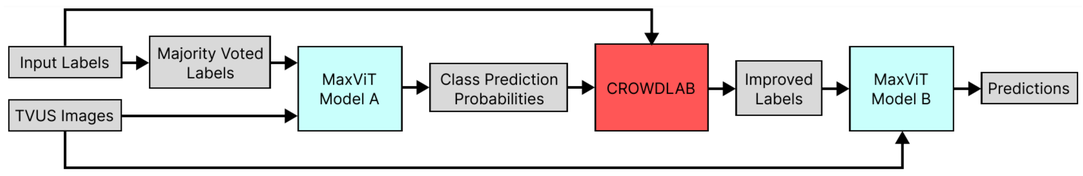

Figure 2: The architecture of the proposed MaxViT model.

Experiments

Dataset

The dataset contains 150 ultrasounds images from 50 unique patients. Each patient has provided a TVUS image of their left ovary, right ovary and uterus. All images are annotated 6 medical professionals: 2 sonographers, 2 radiologists and 2 gynae sonologists. The medical professional's used a grading system to determine the quality of each image. We modified the system to exclude intermediary grades 2 and 4, simplifying the grading to 0, 1, and 2, which originally corresponded to 0, 1, 3, and 5. Grades 0 and 1 were combined, categorizing images as very poor, suboptimal, or optimal. The model is trained on 120 images from 40 patients and validated on 30 images from 10 patients, ensuring equal label distribution.

Model Performance

Model	Training Time (s)	Consenus Label	Accuracy	Macro Average R	Class 0			Class 1			Class 2		
					P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
Resnet50	**533**	M	0.57	0.59	0.67	0.86	0.75	0.38	0.33	0.35	0.62	0.57	0.59
		WE	0.63	0.65	0.50	0.75	0.60	0.67	0.50	0.57	0.67	0.71	0.69
Resnet101	759	M	0.53	0.53	0.56	0.71	0.63	0.33	0.22	0.27	0.60	0.64	0.62
		WE	0.60	0.69	0.50	**1.00**	0.67	0.64	0.58	0.61	0.64	0.50	0.56
MaxViT	1565	M	0.63	0.64	**1.00**	0.71	**0.83**	0.45	0.56	0.50	0.64	0.64	0.64
		WE	**0.80**	**0.77**	0.75	0.75	0.75	**0.88**	**0.64**	**0.74**	**0.78**	**0.93**	**0.85**

Table 1. Performance of the proposed algorithm compared with majority voting on the TVUS dataset using Resnet50, Resnet101 and MaxViT. WE is weighted ensemble, M is majority voting, R is recall, P is precision, F1 is F1-score.

It is observed that, the validation accuracy of the proposed algorithm is 0.80 and the macro average recall is 0.77. Resnet50 had an accuracy of 0.63 and macro average recall of 0.65. Resnet101 had an accuracy of 0.60 and macro average recall of 0.60. Each model's accuracy and macro average recall improved when using our proposed algorithm instead of majority vote. Resnet50 and Resnet101 followed the same finetuning procedure as MaxViT.

More Work from Our Team

[1] Multi-modal Learning with Missing Modality via Shared-Specific Feature Modelling. H Wang, Y Chen, C Ma, J Avery, L Hull, G Carneiro. Computer Vision and Pattern Recognition 2023 (CVPR 2023)
[2] Learnable Cross-modal Knowledge Distillation for Multi-modal Learning with Missing Modality. H Wang, C Ma, J Zhang, Y Zhang, J Avery, L Hull, G Carneiro. Medical Image Computing and Computer-Assisted Intervention 2023 (MICCAI 2023)
[3] Uncertainty-aware Multi-modal Learning via Cross-modal Random Network Prediction. H Wang, J Zhang, Y Chen, C Ma, J Avery, L Hull, G Carneiro. European Conference on Computer Vision 2022 (ECCV 2022)