An Empirical Study of Vision Transformers for Cervical Precancer Detection

Sandeep Angara(✉), Peng Guo, Zhiyun Xue, and Sameer Antani

National Library of Medicine, Bethesda, MD 20894, USA
sandeep.angara@nih.gov

Abstract. Cervical precancer is a direct precursor to invasive cervical cancer and a prime target for ablative therapy. This paper presents an empirical study of Vision Transformers (ViT) for cervical precancer classification, an extended study of our previous work using data derived from two studies conducted by the U.S. National Cancer Institute. In this study, we show that ViT can significantly outperform the current state-of-art methods. We also examine data augmentation techniques that help reduce noise that can interfere in precancer detection, such as specular reflection. We achieve 84% accuracy on the test set outperforming the existing works based on the same dataset. Apart from the performance gains, we observe the learned features focus on cervical regions of anatomical significance. Through these experiments, we demonstrate that ViT attains excellent results compared to the current state-of-the-art methods in classifying cervical images for cervical precancer screening.

Keywords: Vision Transformers (ViT) · Cervical cancer · Transfer learning

1 Introduction

Cervical cancer is the fourth most common cancer in women worldwide [1]. It is caused by persistent infection with one of about 15 genotypes of carcinogenic human papillomavirus (HPV). Due to inadequate screening in low-resource settings, cancer is often detected at its late stages, which is very difficult to cure or may require aggressive cervical excision and result in poor quality of life. Early detection of cervical cancer helps in reducing the mortality rate. Visual Inspection with Acetic acid (VIA), one of the commonly screening modalities in low resource regions, is inexpensive and straightforward. However, visual triage is difficult for human observers, and it often results in inadequate performance [2]. The other screening programs, cervical cytology (Pap tests) and colposcopy, require infrastructure and sufficiently trained personnel and are used in high or medium-resource regions. The process of VIA requires applying 3–5% diluted acetic acid to a speculum-exposed cervix by a health care provider. The whitening of the cervical tissue is suggestive of HPV infection, and texture, edge, and vasculature of and around the whitened regions could be indicators of cervical precancer. Sample cerviscopic images of the cervix with acetic acid are shown in Fig. 1.

KC Santosh et al. (Eds.): RTIP2R 2021, CCIS 1576, pp. 26–32, 2022.
https://doi.org/10.1007/978-3-031-07005-1_3

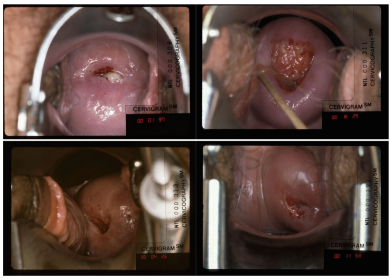

Fig. 1. Examples of cerviscopic cervix images

Convolutional neural networks (CNN's) have become an important tool in today's AI applications. In classification tasks, typical CNN architectures stack multiple convolutional layers together (with nonlinear activations) [3–5] to learn spatial relations for final classifications. Convolution operation can capture only local information but having the capability to abstract and retain information from a large neighborhood boosts the performance in vision tasks [6]. Recently proposed Vision Transformers (ViT) achieve comparable results to CNNs on image classification, object detection, image segmentation, etc. [7]. Moreover, ViT can learn long-range dependencies, which makes transformers an attractive alternative tool to CNNs.

Inspired by the significant success of vision transformers in various computer vision tasks [6], we use ViT with several data augmentation techniques to improve the performance of cervical precancer classification and alleviate the specular reflection issue frequently encountered in cervical images. We also address how long-range dependency learning with transformers can help focus on the cervix region without requiring localization using a detector network and cropping.

2 Related Work

Several deep learning algorithms have been proposed for cervical precancer detection and demonstrated good performance. In the early stage of this research [8], Faster R-CNN [9] algorithm has been used to detect the cervix region and simultaneously predict the case probability. The model was trained on 2000 annotated images using transfer learning with ImageNet weights. The dataset was collected during a National Cancer Institute (NCI) prospective epidemiologic study in the Guanacaste region of Costa Rica. Later the feasibility of automated visual evaluation with cervix images captured using a specialized handheld device was investigated [10]. Both Faster R-CNN and RetinaNet were applied in [10]. Besides using localization and classification based on a single network, [11] developed a customized architecture based using Feature Pyramid Network as the backbone. The last feature layer was built by upsampling and concatenating specific pre-trained feature pyramid layers, followed by a Global Average Pooling (GAP) layer. A fully connected layer for classification was made on top of the GAP layer. The features combined from various layers improved the performance and model explainability compared to related work. BF-CNN combing two-state images [12] has been

proposed to fuse images applied with acetic acid and iodine solution, boosting performance. Semi-supervised learning using ResNeSt50 architecture has been used in [13] to leverage unlabeled data, outperformed the model trained with transfer learning based on ImageNet weights. In [13], various augmentation techniques have been suggested to reduce the impact of specular reflection, vaginal walls, metal speculum, shadows, etc. The proposed image augmentation techniques improved the performance and model interpretability. However, the images were first cropped using a cervix region detector trained on the Costa Rica Vaccine Trial (CVT) dataset [14]. The cropping helps the classification model be less distracted from the non-cervical region and focus on the cervix region. In the current work, we omit the step of cervix region detection and use the original images as the input to the ViT classification network.

3 Experiments

3.1 Data Preparation

The dataset was acquired from two NCI studies: One was the Atypical Squamous Cells of Undetermined Significance/Low-grade Squamous Intraepithelial Lesion (ASCUS/LSIL) Triage Study (ALTS), and the other is the Guanacaste Natural History Study (NHS). ALTS [15] was a multicenter, randomized clinical trial designed to evaluate three alternate methods of management: intermediate colposcopy, cytologic follow-up, and triage by human papillomavirus (HPV) DNA testing. This study was conducted in the United States and designed to determine the optimal management plan for low-grade cervical abnormalities. The data was collected from non-pregnant women 18+ years old with no prior hysterectomy or ablative therapy to the cervix. NHS data was collected from the Costa Rica population-based Cancer Registry [16, 17]. The NHS study was designed to understand the natural history of HPV and cervical cancer. It was also conducted to provide the effectiveness of new screening and management tools. In both ALTS and NHS studies, a cerviscope was used to capture the picture of the cervix region after the application of acetic acid. Digitized images were then obtained using scanners and were compressed for storage. Images in ALTS and NHS datasets were manually reviewed to ensure the cervix region was visible in the image. The image may also contain other anatomy or medical devices, e.g., the vaginal wall, external genitalia, metal speculum, swabs, etc., as shown in Fig. 1. Several medical experts in medical screening and epidemiology annotated the images into two classes: Controls and Cases. In our experiments, the images were resized with a shorter edge to 800 pixels and maintain the aspect ratio. The number of images in each class in training, validation, and test set is provided in Table 1, respectively.

3.2 Architecture

In this experiment, we use a ViT "Base" variant [6]. The transformer splits the image into patches and then flattens the patches. The flattened patches with positional embeddings are passed as input sequences to the standard transformer encoder. The transformer encoder consists of alternating layers of multiple self-attention heads and MLP blocks. The only modification is discarding the prediction head (MLP head) and attaching a linear layer with N units, where N is the number of classes.

Table 1. Number of images in training/validation/test set

	Controls	Cases
Training set	1645	843
Validation set	359	182
Test set	230	115

3.3 Training Setup

We train ViT-Base with a 16 × 16 input patch size. We train the models using SGD optimizer with a learning rate of 0.003 and also using mixed precision. All the models are trained for 2500 steps with a batch size of 512, with cosine learning rate decay. We first finetune the model loaded with ImageNet21k on our dataset using random crop and normalization augmentation techniques. After qualitative and quantitative analysis, we find the trained model is impacted by specular reflection. To make the model robust to the noise [13], we then finetune the model with the following augmentation techniques (using the "Albumentations" library [18]).

- Resize the image to 400-by-400
- Crop a random part of 384-by-384 in the image
- Randomly apply Shift, Scale, Rotate to the image
- Scale hue, saturation, gamma, blur, and brightness
- Apply PCA noise [19] with a coefficient sampled from a normal distribution N (0, 0.1)
- Normalize RGB channels by subtracting mean = (0.5, 0.5, 0.5) and dividing by standard deviation = (0.5, 0.5, 0.5).

In addition, we apply the random sun flares and random fog augmentation techniques used to reduce specular reflections [13].

4 Results and Discussions

Our experiments show that vision transformers trained using transfer learning improve performance significantly compared to the work on the same dataset [13]. Quantitative results are presented in Table 2. The accuracy of the test dataset is improved from 82.0% to 84.0% using the ViT-Base model, and AUC is boosted from 0.87 to 0.91. Figure 2 shows the Precision-Recall curve and ROC curve on the test set, respectively.

We have trained a ViT-Base architecture with augmentation techniques and training setup methods described in Sect. 3. The visualization heatmaps in Fig. 3 demonstrate that the model trained with our suggested augmentation techniques helps reduce specular reflection. The model also concentrates on the cervix region for the final prediction with less impact from unrelated objects like swabs, metal speculum, pubic hair, etc. Unlike our previous work [13], we do not crop the cervix region first and then pass it to the classifier network. As indicated by the attention maps in Fig. 3, the ViT classifier learns to

Table 2. Performance comparison with the work on the same dataset

Model	Method	Image size	Average AUC score	Accuracy
ResNest50 (previous work) [13]	Semi-supervised learning	800 × 800	0.87	82.0%
ViT-Base (16 × 16 patch)	Transfer learning	384 × 384	**0.91**	**84.0%**

concentrate on the cervix region. The underlying self-attention component in the vision transformer helps learn long-range dependencies and integrate the information globally across the image, which makes the network potentially learn rich, high-level features. While the vision transformer achieves better performance, as shown in Fig. 4, the model still distracts from noise in some samples.

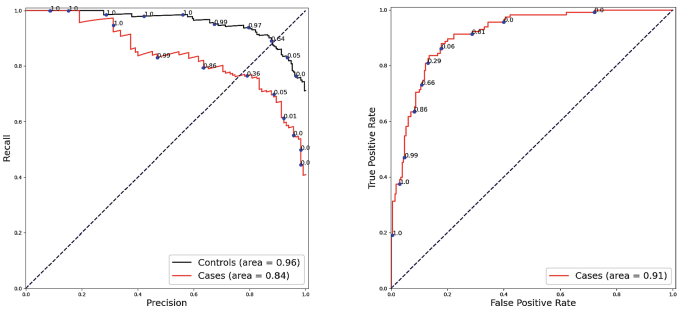

Fig. 2. Precision-Recall and ROC curves on the test dataset

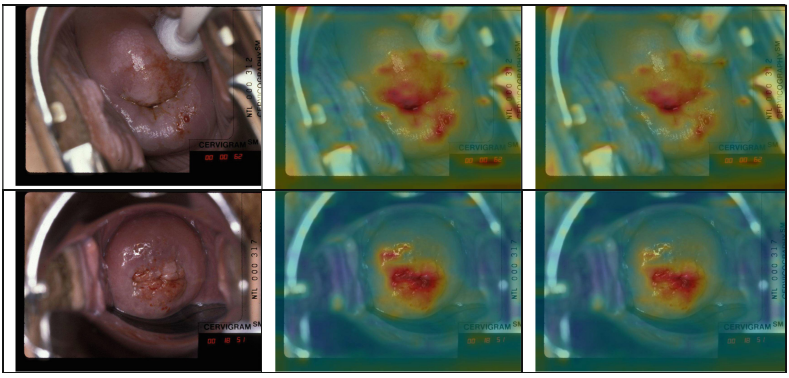

Fig. 3. Images in the first column are input images. Second column images are from the model ViT-Base 16 × 16 model trained with resize and normalization techniques. Heatmaps in the third column represent the model trained with suggested augmentation techniques.

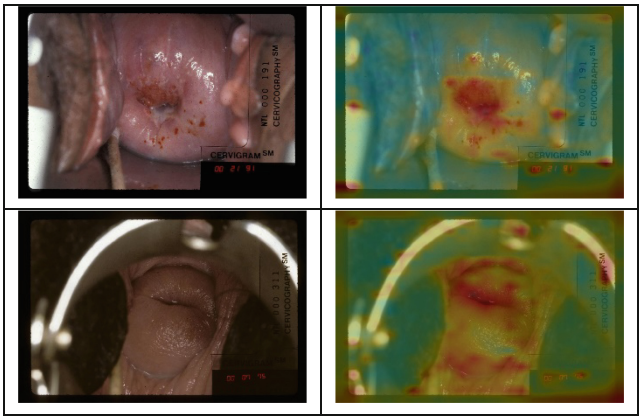

Fig. 4. Heatmaps where the model focuses on text and other noises in the image.

5 Conclusion and Future Work

In this work, we have explored the application of vision transformers with various augmentation techniques. By finetuning the model pre-trained on ImageNet21k on the cervix images, vision transformers improve the efficiency and effectiveness of precancer detection. In addition, the trained model concentrates on the cervix region without using any cervix detector. Future work includes training with higher-resolution images, exploring self-supervised learning, and experimenting with new data augmentation techniques.

Acknowledgment. This work was supported by the Intramural Research Program of the National Library of Medicine, part of the National Institutes of Health. Data used in this research was by agreement between the National Library of Medicine and the National Cancer Institute (NCI). We are grateful to Dr. Mark Schiffman and his team at the NCI for feedback on our findings.

References

1. Schiff, M., et al.: Seminar Human papillomavirus and cervical cancer. https://doi.org/10.1016/S0140-6736(07)61416-0
2. Belinson, J.L., Pretorius, R.G., Permanente, K., Xinfeng Qu, C.: Cervical screening by pap test and visual inspection enabling same-day biopsy in low-resource, high-risk communities (2019). http://journals.lww.com/greenjournal
3. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, December 2016, vol. 2016-December, pp. 770–778 (2016). https://doi.org/10.1109/CVPR.2016.90
4. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. http://code.google.com/p/cuda-convnet/. Accessed 25 Feb 2021
5. Tan, M., Le, Q.V.: EfficientNet: rethinking model scaling for convolutional neural networks. In: 36th International Conference on Machine Learning, ICML 2019, vol. 2019-June, pp. 10691–10700 (2019). http://arxiv.org/abs/1905.11946. Accessed 25 Feb 2021

6. Dosovitskiy, A., et al.: An image is worth 16×16 words: transformers for image recognition at scale. https://github.com/
7. Han, K., et al.: A survey on visual transformer (2022)
8. Hu, L., et al.: An observational study of deep learning and automated evaluation of cervical images for cancer screening. https://doi.org/10.1093/jnci/djy225
9. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. http://image-net.org/challenges/LSVRC/2015/results
10. Xue, Z., et al.: A demonstration of automated visual evaluation of cervical images taken with a smartphone camera. Int. J. Cancer **147**(9), 2416–2423 (2020). https://doi.org/10.1002/ijc.33029
11. Guo, P., et al.: Clinical medicine network visualization and pyramidal feature comparison for ablative treatability classification using digitized cervix images (2021). https://doi.org/10.3390/jcm10050953
12. Yan, L., et al.: Multi-state colposcopy image fusion for cervical precancerous lesion diagnosis using BF-CNN. Biomed. Signal Process. Control **68**(April), 102700 (2021). https://doi.org/10.1016/j.bspc.2021.102700
13. Angara, S., Guo, P., Xue, Z., Antani, S.: Semi-supervised learning for cervical precancer detection, pp. 202–206 (2021). https://doi.org/10.1109/CBMS52027.2021.00072
14. Guo, P., Xue, Z., Rodney Long, L., Antani, S.: Cross-dataset evaluation of deep learning networks for uterine cervix segmentation. Diagnostics **10**(1), 44 (2020). https://doi.org/10.3390/diagnostics10010044
15. Schiffman, M., Adrianza, M.E.: ASCUS-LSIL Triage Study Design, Methods and Characteristics of Trial Participants (2000)
16. Rodr Iguez, A.C., et al.: Cervical cancer incidence after screening with HPV, cytology, and visual methods: 18-Year follow-up of the Guanacaste cohort. https://doi.org/10.1002/ijc.30614
17. Herrero, R., et al.: Design and methods of a population-based natural history study of cervical neoplasia in a rural province of Costa Rica: the Guanacaste Project 1 (1997)
18. Buslaev, A., Iglovikov, V.I., Khvedchenya, E., Parinov, A., Druzhinin, M., Kalinin, A.A.: Albumentations: fast and flexible image augmentations. Inf. **11**(2), 125 (2020). https://doi.org/10.3390/info11020125
19. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet Classification with Deep Convolutional Neural Networks. http://code.google.com/p/cuda-convnet/. Accessed 28 Feb 2021