내시경 병리소견 분류를 위한 비전 트랜스포머

겔란 아야나* · 최세운

금오공과대학교

Vision transformers for endoscopic pathological findings classification

Gelan Ayana* · Se-woon Choe

Kumoh National Institute of Technology

E-mail : gelan@kumoh.ac.kr / sewoon@kumoh.ac.kr

요 약

위장관의 내시경적 병리학적 소견은 대장암의 조기 진단에 중요하다. 최근 CNN 기반 딥 러닝의 활용은 주관적 분석의 정확도와 조기 진단의 성능을 높이는 결과를 보였으나, 계산 복잡도가 높고 임상에 즉시 활용하기에는 상대적으로 낮은 정확도로 사용에 제한적이다. 이러한 문제를 해결하기 위해 본 논문에서는 대장암의 조기 발견을 위한 비전 트랜스포머 기반 내시경 병리 소견 분류법을 제안한다. 식도염, 폴립, 궤양성 대장염을 포함한 병리학적 소견이 있는 내시경 이미지를 각각 1,000개씩 사용하였으며, 제안된 접근 방식을 사용하여 세 가지 병리학적 소견을 분류할 때 98%의 분류 정확도를 보였다.

ABSTRACT

The endoscopic pathological findings of gastrointestinal tract (GIT) are important in the early diagnosis of colorectal cancer. Deep learning based on convolutional nueral network (CNN) has been implemented to solve the subjective analysis problem and to increase the performance of early detection of pathological findings. However, the desired performance is yet to be achieved and CNNs are computationally complex. To solve these problems, in this paper, we propose a vision transformer based endoscopic pathological findings classification for the early detection of colorectal cancer. Publicly available endoscopic images with three pathological findings, including esophagitis, polyps, and ulcerative colitis, each with 1000 images were used. Using our approach, we have achieved a test accuracy of 98% in classifying the three pathological findings.

키워드

Vision transformer; colorectal cancer; endoscopy

Ⅰ. Introduction

In 2020, out of the 19.3 million new cancer cases occurred worldwide, colorectal cancer (CRC) accounts for 10% being the third most commonly diagnosed cancer next to breast cancer (11.7%) and lung cancer (11.4%) [1]. The early diagnosis of CRC is crucial that CRC detected at early stage results in five-year relative survival rate above 90% [1]. However, only 40% of CRC are detected at early stage and if cancer spreads beyond the colon or rectum, this survival rate lowers [2]. Endoscopy is the standard modality for the diagnosis of CRC where doctors see abnormal growth in the colon or rectum [1], [2]. These abnormal growths are challenging to detect at early stage and convolutional neural network (CNN) based deep learning has been employed to improve diagnosis [3]. CNNs have the ability to learn visual representations for easy transfer and strong performance due to its strong inductive bias of spatial equivariance and translational invariance provided by its convolutional layers. However, vision transformers (ViT) showed superior performance over CNNs for natural image classification [4]. In contrast to CNNs that perform many convolutions at different layers to focus on a

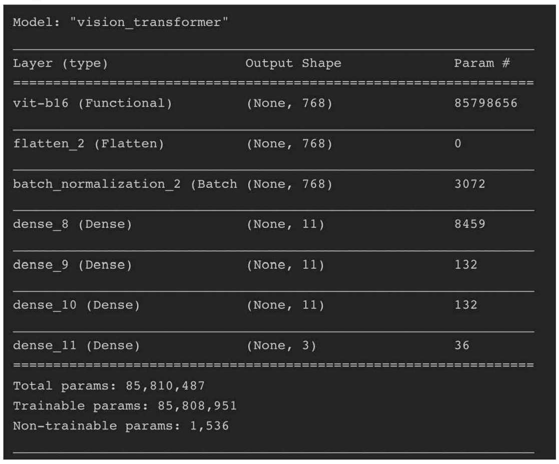
* speaker

given area of an image, ViTs focuses on all area of the image at once beginning from its early layers. Despite the few applications of CNNs to endoscopic colorectal images, ViTs has not been yet explored for its use in detection of CRC [5]. In this study, we propose a novel transfer learning algorithm for endoscopic colorectal cancer images classification based on vision transformers.

Ⅱ. Materials and method

2.1. Transfer learning

The proposed vision transformer based transfer learning method for colorectal image classification used a vision transformer model pre-trained on ImageNet and transfer-learned for the colorectal image classification task. We used the Dosovitsky et al., [6] vision transformer base model with 16x16 patch size (vitb_16) pre-trained model and added three dense layers removing the output layer of the original model.



```
Model: "vision_transformer"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
vit-b16 (Functional)         (None, 768)               85798656
_________________________________________________________________
flatten_2 (Flatten)          (None, 768)               0
_________________________________________________________________
batch_normalization_2 (Batch (None, 768)               3072
_________________________________________________________________
dense_8 (Dense)              (None, 11)                8459
_________________________________________________________________
dense_9 (Dense)              (None, 11)                132
_________________________________________________________________
dense_10 (Dense)             (None, 11)                132
_________________________________________________________________
dense_11 (Dense)             (None, 3)                 36
=================================================================
Total params: 85,810,487
Trainable params: 85,808,951
Non-trainable params: 1,536
_________________________________________________________________
```

Figure 1: The proposed ViTs model.

2.2. Dataset

The dataset for this study was collected from the publicly available Kvasir dataset called a multi-class-dataset for computer aided gastrointestinal disease detection [7]. The dataset has eight classes of the general gastrointestinal disease (GIT) categorized into three, namely anatomical landmarks (consists of Z-line, pylorus, and cecum classes), pathological findings (consists of esophagitis, polyps, and ulcerative colitis), and polyp removals (dyed and lifted polyps and dyed resection margins). Since our purpose in this study is colorectal cancer detection that is identified using pathological findings, we used the three classes from the pathological findings, i.e., esophagitis, polyps, and ulcerative colitis with 1000 images in each class.

2.3. Implementation details

In transfer learning from ImageNet pre-trained vitb_16 model to CRC images, only the last layer is removed and replaced with three dense layers and softmax layer. The proposed model was implemented with the Keras on TensorFlow framework using Python. Two pieces of RTX 3090 GPUs were employed to accelerate the training. Early-stopping with a patience of 7 has been applied for training and L2 regularization has been used. The gradient optimizer used was Adagrad with learning rate of 0.01. The training batch size was 16. The CRC image dataset was categorized into 2100 training, 600 validation, and 300 test sets with a ratio of 7:2:1, consecutively. Augmentation was used to increase the number of training images.

Ⅲ. Results and Future Work

Preliminary results of experiments with the proposed model is presented in Figures 2 and 3. The proposed method achieved test accuracy of 98%, F1-score of 0.98 recall of 0.98, and precision of 0.98, see the confusion matrix in Figure 2. The ViTs model converges fast achieving high training and validation accuracy at early epochs as shown in the learning curve in Figure 3.

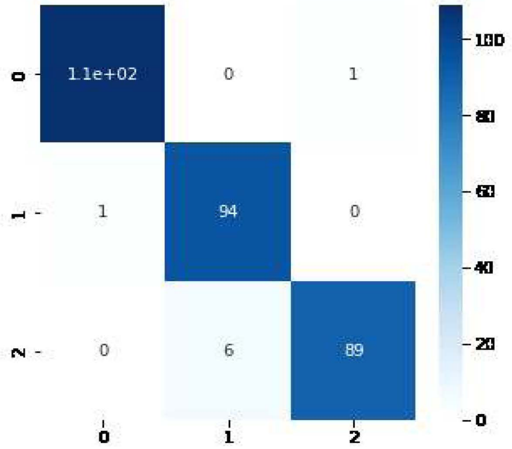


Figure 2. Confusion matrix of the proposed model.

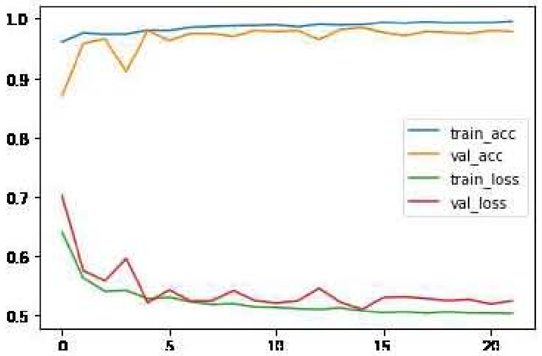


Figure 3. The learning curve of the proposed method.

Comparison with the state-of-the-art CNN methods will be the next experiment we will perform.

References

[1] H. Sung *et al.*, "Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries," *CA. Cancer J. Clin.*, vol. 71, no. 3, pp. 209–249, May 2021, doi: 10.3322/caac.21660.

[2] B. Lu *et al.*, "Colorectal cancer incidence and mortality: the current status, temporal trends and their attributable risk factors in 60 countries in 2000-2019," *Chin. Med. J. (Engl).*, vol. 134, no. 16, pp. 1941–1951, 2021, doi: 10.1097/CM9.0000000000001619.

[3] G. Yu *et al.*, "Accurate recognition of colorectal cancer with semi-supervised deep learning on pathological images," *Nat. Commun.*, vol. 12, no. 1, pp. 1–13, 2021, doi: 10.1038/s41467-021-26643-8.

[4] M. Raghu, T. Unterthiner, S. Kornblith, C. Zhang, and A. Dosovitskiy, "Do Vision Transformers See Like Convolutional Neural Networks?," no. NeurIPS, 2021, [Online]. Available: http://arxiv.org/abs/2108.08810

[5] Y. J. Kim, J. P. Bae, J. W. Chung, D. K. Park, K. G. Kim, and Y. J. Kim, "New polyp image classification technique using transfer learning of network-in-network structure in endoscopic images," *Sci. Rep.*, vol. 11, no. 1, pp. 1–8, 2021, doi: 10.1038/s41598-021-83199-9.

[6] A. Dosovitskiy *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," 2020, [Online]. Available: http://arxiv.org/abs/2010.11929

[7] K. Pogorelov *et al.*, KVASIR: A Multi-Class Image Dataset for Computer Aided Gastrointestinal Disease Detection. Proceedings of the 8th ACM on Multimedia Systems Conference, 2017. Available: https://datasets.simula.no/kvasir/