비전 트랜스포머 인코더가 포함된 U-net을 이용한 대장 내시경 이미지의 폴립 분할

겔란 아야나 · 최세운*

국립금오공과대학교

U-net with vision transformer encoder for polyp segmentation in colonoscopy images

Gelan Ayana · Se-woon Choe*

Kumoh National Institute of Technology

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

요 약

대장암의 조기 발견과 치료를 위해서는 정확한 폴립의 분할이 중요하나 다음과 같은 제약이 따른다. 개별 폴립의 위치, 크기 및 모양이 서로 상이하며, 모션 흐림 및 빛 반사와 같은 특정 상황에서 폴립과 주변 환경 간에 상당한 정도의 유사성이 존재한다. 인코더와 디코더 역할을 하는 Convolutional Neural Networks로 구성된 U-net은 이러한 한계를 극복하기 위해 다양하게 사용된다. 본 연구는 보다 정확한 폴립 분할을 위한 비전트랜스포머가 포함된 U-net 아키텍처를 제안하였고, 그 결과 제안된 방식은 표준 U-net 아키텍처보다 더 나은 성능을 보였음을 확인할 수 있었다.

ABSTRACT

For the early identification and treatment of colorectal cancer, accurate polyp segmentation is crucial. However, polyp segmentation is a challenging task, and the majority of current approaches struggle with two issues. First, the position, size, and shape of each individual polyp varies greatly (intra-class inconsistency). Second, there is a significant degree of similarity between polyps and their surroundings under certain circumstances, such as motion blur and light reflection (inter-class indistinction). U-net, which is composed of convolutional neural networks as encoder and decoder, is considered as a standard for tackling this task. We propose an updated U-net architecture replacing the encoder part with vision transformer network for polyp segmentation. The proposed architecture performed better than the standard U-net architecture for the task of polyp segmentation.

키워드

U-net; vision transformer; segmentation; colorectal cancer

Ⅰ. Introduction

Colorectal cancer (CRC) accounts for 10% of the 19.3 million new cancer cases expected to be diagnosed worldwide in 2020, behind lung cancer (11.4%) and breast cancer (11.7%) [1]. Since CRC identified at an early stage has a five-year relative survival rate above 90%, early detection of the disease is essential [1]. However, only 40% of CRC are found in the early stages, and the survival probability drops if the disease spreads outside the colon or rectum [2]. For the early identification and treatment of colorectal cancer, accurate polyp segmentation is crucial. However, polyp segmentation is a challenging task, and the majority of current approaches struggle with two issues [3]. First, the position, size, and shape of each individual polyp varies greatly (intra-class inconsistency). Second, there is a significant degree

* corresponding author

of similarity between polyps and their surroundings under certain circumstances, such as motion blur and light reflection (inter-class indistinction). Therefore, convolutional neural network (CNN) based deep learning algorithms have been proposed to improve diagnosis [4]. Due to the strong inductive bias of spatial equivariance and translational invariance given by its convolutional layers, CNNs are able to learn visual representations for simple transfer and good performance. However, for natural images classification and segmentation, vision transformers (ViT) outperformed CNNs. [5]. ViTs concentrates on every part of the image at once starting from the early layers, in contrast to CNNs that conduct several convolutions at various layers to focus on a specific portion of an image. Despite CNNs' wide use in endoscopic colorectal image segmentation, ViTs has not yet been investigated for its potential in CRC early diagnosis. [6]. In this study, we propose a U-net based segmentation algorithm with vision transformer network as its encoder for endoscopic colorectal cancer diagnosis via polyp segmentation.

II. Materials and method

2.1. The proposed method

The proposed method is based on U-net architecture that utilizes transformer layers as its encoder and CNN layers as its decoder. We used vision transformer layers proposed by Dosovitsky et al., [7] as our encoder network. Figure shows the proposed architecture.

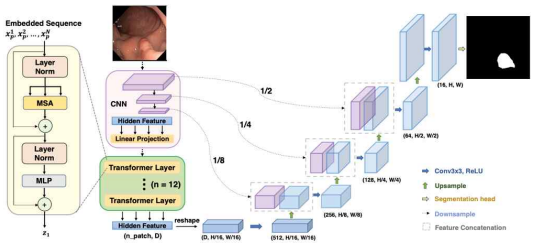

Figure 1. The proposed model architecture

2.2. Dataset

The dataset for this study was collected from the publicly available Kvasir-SEG dataset called segmented polyp dataset for computer aided gastrointestinal disease detection. a multi-class-dataset for computer aided gastrointestinal disease detection [8]. The dataset contains 1000 polyp images with their corresponding annotation.

2.3. Implementation details

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 5 has been applied for training and L2 regularization has been used. The gradient optimizer used was Adam with learning rate of 0.0001. The training batch size was 16. The dataset was categorized into 800 training and 200 test images. The images were resized to 224x224 pixels.

III. Results and Future Work

Preliminary results show that the proposed method outperform the original U-net architecture in segmenting polyps. The proposed method provided mean intersection over union (mIoU) of 0.76, where as the original U-net architecture provided mIoU of 0.71. Figure 2 shows sample segmentation outputs of the proposed method.

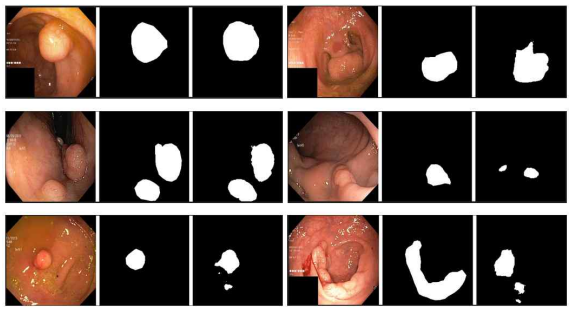

Figure 2. Segmentation outputs of the proposed method

The results from our experiments are intriguing and with more parameters optimization, better results could be achieved. Therefore, parameters optimization and comparison with the state-of-the-art segmentation methods will be the next experiment we will perform.

Acknowledgement

본 논문은 4단계 BK21 사업(금오공과대학교 IT융복합공학과)에 의하여 지원되었으며, 중소벤처기업부에서 지원하는 2022년도 산학연 플랫폼 협력 기술개발사업 (S3310765)의 연구수행으로 인한 결과물임을 밝힙니다.

References

[1] H. Sung, J. Ferlay, R. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal and F. Bray, "Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries," *CA: a cancer journal for clinicians*, Vol. 71, No. 3, pp. 209-249, May 2021

[2] B. Lu, N. Li, C. Y. Luo, J. Cai, M. Lu, Y. H. Zhang, H. D. Chen and M. Dai, "Colorectal cancer incidence and mortality: the current status, temporal trends and their attributable risk factors in 60 countries in 2000-2019," *Chinese medical journal*, Vol. 134, No. 16, pp. 1941-1951, Aug. 2021

[3] W. Li, Y. Zhao, F. Li and L. Wang, "MIA-Net: Multi-information aggregation network combining transformers and convolutional feature learning for polyp segmentation," *Knowledge-Based Systems*, Vol. 247, pp. 108824, Jul. 2022

[4] G. Yu, K. Sun, C. Xu, X. H. Shi, C. Wu, T. Xie, R. Q. Meng, X. H. Meng, K. S. Wang, H. M. Xiao and H. W. Deng, "Accurate recognition of colorectal cancer with semi-supervised deep learning on pathological images," *Nature communications,* Vol. 12, No. 1, pp. 1-13, Nov. 2021

[5] M. Raghu, T. Unterthiner, S. Kornblith, C. Zhang, and A. Dosovitskiy, "Do Vision Transformers See Like Convolutional Neural Networks?," *Advances in Neural Information Processing System*, Vol. 34, pp. 12116-12128, Dec. 2021

[6] 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," *Scientific Reports*. Vol. 11, No. 1, pp. 1-8, Feb. 2021

[7] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly and J. Uszkoreit, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," *arXiv preprint arXiv:2010.11929,* Oct. 2020

[8] D. Jha, P. H. Smedsrud, M. A. Riegler, P. Halvorsen, T. D. Lange, D. Johansen, H. D. Johansen, "Kvasir-seg: A segmented polyp dataset." in *Proceeding of the International Conference on Multimedia Modeling*, Daejeon, South Korea, pp. 451-462, 2020