C-Swin UNETR: 3D 의료 영상 분할을 위한 채널 어텐션이 적용된 Swin Transformer

안성현○ 김환희 권세인 박상현†
연세대학교 컴퓨터과학과
{skd, hwanhee, seinkwon97, sanghyun}@yonsei.ac.kr

C-Swin UNETR: Swin Transfomer with Channel Attention for 3D Medical Image Segmentation

Sunghyun Ahn○ Hwanhee kim Sein Kwon Sanghyun Park†
Dept. of Computer Science, Yonsei University

요 약

의료 영상 분할(Medical Image Segmentation)이란 CT나 MRI와 같은 영상에서 종양과 같은 병변이나 다양한 장기를 구별하는 기술이다. 의료 영상 분할 연구는 U-Net을 기반으로 하며, 모델은 크게 인코더와 디코더로 구성된다. 인코더는 입력 영상의 다양한 스케일에 대한 문맥 정보를 파악하고, 디코더는 다양한 스케일의 문맥 정보를 융합하여 영상 분할 맵을 생성한다. 그러나 U-Net 기반 모델들은 인코더 구조에서 채널 수의 증가에도 불구하고 공간적 특징에만 초점을 맞추었다. 따라서 우리는 유용한 공간 정보 및 채널 정보를 모두 활용하는 C-Swin UNETR을 제안한다. 제안하는 방법론을 검증하기 위해 Synapse 데이터셋을 통해 기존의 모델과 비교 실험을 진행하였으며, 제안하는 방법론의 우수성을 검증하였다.

1. 서 론

의료 영상 분할(Medical Image Segmentation)은 CT나 MRI와 같은 복잡한 3D 영상에서 종양과 같은 병변이나 다양한 장기들을 구분하는 기술이다. 이 기술은 스마트 병원에서 환자를 정확히 진단하기 위해 구축해야 되는 지능형 의료 시스템이다[1].

의료 영상 분할 연구는 U-Net[2]을 기반으로 하며, 모델은 인코더와 디코더로 구성된다. 인코더는 입력 영상의 다양한 스케일에 대한 문맥 정보를 순차적으로 파악하고, skip connection을 통해 디코더로 특징맵을 전달한다. 디코더는 이러한 특징맵들을 융합하여 모든 스케일의 문맥 정보가 담긴 특징맵을 생성하고, 이를 이용하여 영상 분할 맵을 생성한다.

그러나 U-Net은 특징 추출 과정에서 각 컨볼루션 커널이 전체 영상의 하위 영역에만 집중하므로, 전역적인 정보를 충분히 활용할 수 없다는 한계점이 존재한다. 따라서 최근에는 인코더의 구조를 Swin Transformer[3]로 대체하는 Swin UNEt TRansformers(Swin UNETR)[4] 모델이 대두되었다. Swin UNETR은 인코더의 각 스테이지에서 서로 다른 패치 크기로 어텐션을 수행하는 방식이므로, 지역 및 전역 정보를 충분히 활용할 수 있다. 또한 Shifted Window 방식의 멀티-헤드 셀프 어텐션(SW-MSA)을 활용하여 해상도의 제곱에 비례하는 시간복잡도를 선형적으로 줄였으며, 기존 모델보다 높은 성능을 보여주었다.

하지만 Swin UNETR을 포함한 대다수의 U-Net 기반 모델들은 인코더 구조에서 특징맵의 채널 수가 증가함에도 불구하고 공간적 특징에만 초점을 맞추었다.

따라서 우리는 공간 및 채널 어텐션을 동시에 수행하는 C-Swin Block을 설계한 뒤, U-Net 모델의 인코더와 디코더에 적용하였다. 이를 통해 입력 영상과 융합 특징맵에 대한 중요한 문맥 정보를 파악하였고, 더 나은 영상 분할을 가능하게 했다. 우리는 성능 분석을 위해 Synapse 데이터셋으로 제안 모델을 학습하였다.

본 논문의 기여는 다음과 같다.

- 유용한 공간 및 채널 정보를 효율적으로 학습하는 C-Swin Block을 설계하였다.
- U-Net 모델의 인코더와 디코더에 C-Swin Block을 적용하여 기존 모델보다 1.99% 향상된 성능을 달성하였다.

2. 모델 구조

우리는 인코더와 디코더에서 C-Swin Block을 적용한 U-Net 기반 모델(C-Swin UNETR)을 제안한다. 모델은 3D 영상을 입력으로 받아 클래스 정보가 담긴 영상 분할 맵을 생성한다. 제안 모델은 그림 1과 같다.

2.1. 인코더 및 디코더

인코더는 입력 영상을 크기가 (4,4,2)인 패치 단위로 나누고,

* 이 논문은 2023년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원(IITP-2017-0-00477, (SW 스타랩) IoT 환경을 위한 고성능 플래시 메모리 스토리지 기반 인메모리 분산 DBMS 연구개발)과 국토교통부의 스마트시티 혁신인재육성사업으로 지원을 받아 수행된 연구임.

† 교신 저자: sanghyun@yonsei.ac.kr

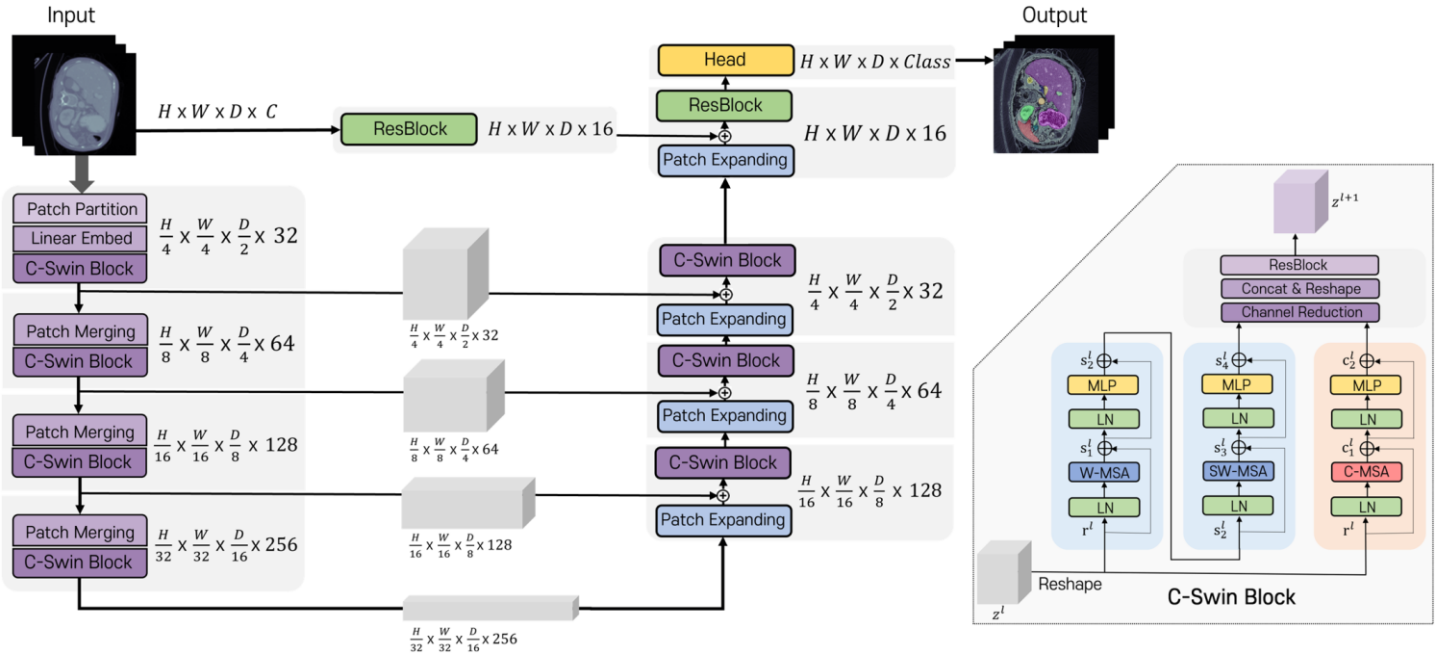

그림 1. 제안 모델(C-Swin UNETR)과 모듈(C-Swin Block)의 구조

각 패치들을 임베딩하여 채널 수가 32인 특징맵을 생성한다. 이후 C-Swin Block을 통해 공간 및 채널 어텐션을 수행하는 과정과 주변 패치들을 병합하여 해상도를 두 배만큼 축소하는 과정을 반복한다. 이를 통해 각 스테이지에서 다양한 스케일에 대한 중요한 문맥 정보를 순차적으로 파악할 수 있다. 모든 특징맵들은 skip connection을 통해 디코더로 전달된다.

디코더는 패치들을 확장하여 해상도를 두 배만큼 확대하는 과정과 전달받은 특징맵들을 융합하고 어텐션을 수행하는 과정을 반복한다. 마지막 단은 Residual Block[5]이 수행된 특징맵을 전달받아 입력 영상의 지역적 특징을 보충한다. 이를 통해 모든 스케일의 문맥 정보가 담긴 특징맵을 생성할 수 있으며, 이후 Head의 1x1x1 합성곱을 통해 영상 분할 맵이 제작된다.

2.2. C-Swin Block

그림 1의 우측 하단은 제안된 C-Swin Block의 구조를 나타내며 l은 Block의 순서를 의미한다. 4D 특징맵인 z^l은 어텐션을 수행하기 위해 2D 구조인 r^l로 변환되고, r^l은 공간 어텐션을 수행하는 Swin Block과 채널 어텐션을 수행하는 C-MSA Block으로 각각 전달된다. C-MSA Block은 층 정규화(LN)와 채널 멀티-헤드 셀프 어텐션 (C-MSA), 다층 퍼셉트론(MLP)으로 구성되며, 채널 어텐션은 수식 (1)과 같이 계산된다. 식의 Q, K, V는 각각 쿼리, 키, 값을 나타내고, d는

$$ChannelAttention(Q,K,V) = V \cdot Softmax(\frac{Q^T K}{\sqrt{d}}) \quad (1)$$

$$\begin{aligned} &\mathrm{C-MSA}(Q,K,V) = Concat(head_1,\dots,head_h)W^O \\ &where\ head_i = ChannelAttention(QW_i^Q, KW_i^K, VW_i^V) \end{aligned} \quad (2)$$

키의 차원 수를 의미한다. 또한 우리는 헤드의 수(h)를 4로 설정하여 수식 (2)와 같이 멀티-헤드 어텐션을 계산하였다.

각 어텐션 Block의 수행 결과인 s^l_4와 c^l_2는 채널을 반으로 줄여 결합되고, 다시 4차원 형태로 변환된다. 마지막으로 Residual Block을 통해 두 어텐션의 특징이 융합된 z^{l+1}이 생성된다. 우리는 제안 모델(C-Swin UNETR)의 각 스테이지에서 C-Swin Block을 세 번 사용하였다.

3. 실험 및 결과

3.1 실험 환경

다양한 장기에 대한 영상 분할 데이터셋인 Synapse[6]를 이용하여 모델을 학습시켰다. 우리는 nnFormer[7]에서 제안한 전처리 방식을 따르고 추가적인 데이터를 사용하지 않았다. 데이터셋은 128 x 128 x 64인 crop된 이미지를 사용하며 데이터 증강 기법이 적용된다. 손실 함수는 교차 엔트로피와 Dice Loss를 합해서 사용하고, 옵티마이저는 모멘텀이 0.99인 SGD를 사용한다. Epoch, iteration, batch size는 각각 1000, 250, 2로 구성하였다. 이 실험에서는 NVIDIA GeForce RTX 3090을 한 대 사용하였다.

표 1. Synapse 데이터셋에서의 실험 결과 (Dice Score)

Model	Spleen	Right Kidney	Left Kidney	Gallbladder	Liver	Stomach	Aorta	Pancrease	Average
Swin UNETR	0.9537	0.8626	0.8699	0.6654	0.9572	0.7707	0.9112	0.6880	0.8348
C-Swin UNETR	**0.9554**	**0.8676**	**0.8741**	**0.6847**	**0.9647**	**0.8123**	**0.9230**	**0.7562**	**0.8547**

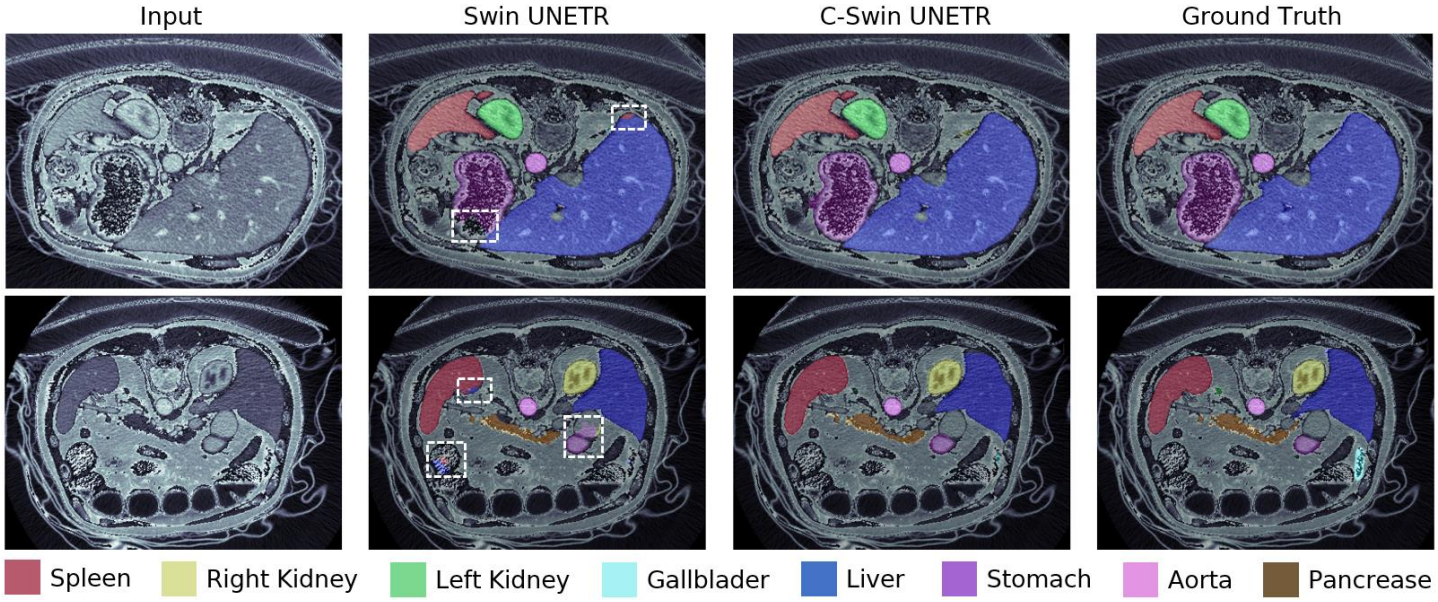

그림 2. Synapse 데이터셋에서의 실험 결과 시각화

3.2. 실험 결과

표 2. 어텐션의 결합 방법 비교

Methods	Dice Score ↑
w/o C-MSA Block (spatial attention)	0.8410
C-Swin Block (channel + spatial)	0.8441
C-Swin Block (spatial + channel)	0.8460
C-Swin Block (spatial & channel in parallel)	**0.8547**

표 1의 Dice Score는 예측된 분할 영상과 정답 간의 유사도를 측정하는 지표로서 0과 1 사이의 값을 가진다. 우리 모델은 공간 어텐션만 사용하는 Swin UNETR과 비교하여 모든 장기에 대해 더 우수한 성능을 보였으며, 평균 Dice Score는 1.99% 상승하였다.

그림 2는 Synapse 데이터셋의 특정 Depth에 대한 영상 분할 결과이며 흰색 점선 박스는 Swin UNETR이 장기를 제대로 구분하지 못 한 영역이다. Swin UNETR은 위의 일부 영역을 구분하지 못하였고, 간이나 비장의 말단 부분을 다른 장기로 잘못 예측하였지만 우리 모델은 정답과 비슷한 결과를 보였다.

표 2는 제안 모델에서 C-Swin Block의 어텐션 결합 방법을 변경한 후 실험한 결과이다. C-MSA Block을 제거하였을 때 가장 성능이 낮았으며, 직렬 구조보다는 병렬 구조가 더 우수한 성능을 보였다.

4. 결 론

우리가 제안한 C-Swin UNETR은 인코더와 디코더에서 공간 및 채널 간의 관계를 더 잘 파악하여 여러 장기에 대한 문맥 정보를 학습하는 데 효과적이었다. 또한 두 어텐션 레이어로부터 추출된 문맥 정보를 병렬로 융합하는 방법이 영상 분할의 성능 향상을 위한 좋은 구조임을 확인하였다.

참 고 문 헌

[1] Hu, He-Xuan, et al. "Multimodal brain tumor segmentation based on an intelligent UNET-LSTM algorithm in smart hospitals." ACM Transactions on Internet Technology, vol.5, pp.1-14, 2021.

[2] Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation." Medical Image Computing and Computer-Assisted Intervention-MICCAI, vol.18, pp.234-241, 2015.

[3] Liu, Ze, et al. "Swin transformer: Hierarchical vision transformer using shifted windows." Proceedings of the IEEE/CVF international conference on computer vision, pp.10012-10022, 2021.

[4] Tang, Yucheng, et al. "Self-supervised pre-training of swin transformers for 3d medical image analysis." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.20730-20740, 2022.

[5] He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770-778, 2016.

[6] B Landman, Z Xu, J Igelsias, M Styner, T Langerak, and A Klein. "Miccai multi-atlas labeling beyond the cranial vault-workshop and challenge." In Proc. MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge, vol.5, p.12, 2015.

[7] Zhou, Hong-Yu, et al. "nnformer: Interleaved transformer for volumetric segmentation." arXiv preprint arXiv:2109.0320 1, 2021.