RefNet: 향상된 3차원 의료 영상 분할을 위한 UNet 기반의 분할 결과 개선 기법

조영완° 최종환 서상민 박상현†
연세대학교 컴퓨터과학과
{jyy1551, mathcombio, ssm6410, sanghyun}@yonsei.ac.kr

RefNet: UNet-based Segmentation Output Refining method for Improved 3D Medical Image Segmentation

Youngwan Jo° Jonghwan Choi Sangmin Seo Sanghyun Park†
Department of Computer science, Yonsei University

요 약

3차원 의료 영상 분할은 CT 또는 MRI 영상에서 단일 또는 다중 장기를 구분하고 식별하는 기술을 가리킨다. 3차원 의료 영상 분할을 위한 심층 신경망 모델이 여럿 제안되었으나, 훈련을 위한 다량의 3차원 의료 영상의 데이터를 수집하는 것은 어려운 일이며, 또한 학습량의 부족으로 인하여 배경과 장기의 경계가 모호한 부분을 명확하게 파악하지 못하는 성능적 한계가 존재한다. 본 연구에서는 그러한 심층 신경망 모델의 분할 성능을 향상시키기 위해 신경망 모델로 예측된 영역들의 경계를 보정하고 잘못 예측된 부분의 개선을 목적으로 하는 후처리 신경망 모델 RefNet을 제안한다. RefNet은 기존의 의료 영상 분할 모델로 예측된 분할 결과 데이터를 학습하기 때문에, 의료 영상 데이터가 적을지라도 여러 개의 분할 모델을 적용하여 분할 모델의 개수만큼 배로 많은 RefNet 훈련용 데이터를 확보할 수 있다. RefNet 성능 평가를 위해 Synapse 벤치마크 데이터를 이용하였으며, 최신 3차원 의료 영상 분할 모델인 Swin UNETR의 분할 성능을 각 장기에 대해 0.04-4.03% 향상시키는 것을 확인하였다.

1. 서 론

3차원 의료 영상 분할 (3D medical image segmentation)은 computed tomography (CT) 또는 magnetic resonance imaging (MRI) 영상에서 단일 또는 다중 장기의 영역을 구분 및 식별하는 기술을 가리킨다. 영상 분할 분야에서 널리 활용되는 심층 신경망 모델로는 UNet[1]이 있으며, UNet은 수축 경로(contracting path)를 통해 영상 속의 정보를 파악하고, 확장 경로(expanding path)를 통해 파악된 정보에 기반한 영상 분할 작업을 수행한다.

최근 연구들은 다중 장기 영상에서의 영역 분할을 효과적으로 수행하기 위해 U-Net의 수축 및 확장 경로를 3D 합성곱 신경망 (convolutional neural network; CNN)[2] 또는 transformer[3] 등으로 구현하는 다양한 변형들을 제안하였다. Swin U-Net transformer (Swin UNETR)[4]은 수축 경로를 Swin transformer[5]로 구성하고, 확장 경로를 ResBlock으로 구현한 최신 3D 의료 영상 분할 모델(segmentation model)이다. 수축 경로에서 사용된 swin transformer는 5가지의 단일 장기 CT 데이터 집합을 대조 학습(contrastive learning)을 포함한 다양한 훈련작업을 통해 사전 학습되었으며, 이를 통해 다른 3D 영상 분할 모델들 대비 다중 장기 데이터 집합들에 대한 향상된 분할 성능을 보여주었다.

사전 학습 전략을 취한 Swin UNETR은 기존의 다중 장기 분할 모델들 보다 우수한 성능을 보여주었으나, 3차원 의료 영상 데이터가 갖는 2가지 문제로 인하여 성능 발휘에 제약이 있다. 첫째는 학습을 위한 데이터가 많지 않다는 문제이다. 심층 신경망 모델을 효과적으로 학습시키기 위해서는 다량의 훈련데이터가 필요하나, 다중 장기 분할 데이터는 의료 영상 데이터 수집 및 레이블링이 어렵기 때문에 충분한 양을 확보하기가 어렵다[6]. 다양한 데이터 증강 기법을 사용하여 훈련 데이터의 수를 증가시킬 수 있으나, 일반적인 증강기법은 하나의 영상을 회전(rotate), 자름(crop) 등과 같은 방법으로 새로운 영상 데이터를 생성하는 것이기 때문에 유의미한 장기 분할 영상을 학습하기 어렵다. 두 번째는 배경과 장기들 간의 경계가 애매모호하다는 문제이다[7]. 3차원 의료 영상의 낮은 분해능(resolution)으로 인하여 서로 다른 장기들 또는 장기와 배경 간의 대비도(contrast)가 낮기 때문에, Swin UNETR로 예측된 장기 영역들의 경계가 매끄럽지 않고 부정확하게 나타나는 것을 확인할 수 있다. 이러한 문제를 충분한 모델 학습을 통해 극복할 수 있지만, 앞서 언급한 바와 같이 학습데이터가 부족하기 때문에, 직접적인 모델 개선보다는 예측된 결과를 보정해줄 수 있는 새로운 접근법이 필요하다.

본 연구에서는 3차원 의료 영상 데이터가 부족한 상황에서 기존의 영역 분할 모델의 성능을 효과적으로 향상시킬 수 있는 UNet 기반의 분할 결과 개선 신경망 (segmentation output refining neural network; RefNet) 모델을 제안한다. RefNet은

† 교신 저자: sanghyun@yonsei.ac.kr

* 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.

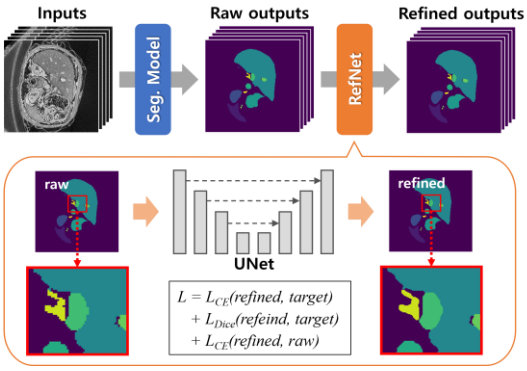

그림 1. Overview of output refinement of 3D medical image segmentation using RefNet

영역 분할 모델을 통해 얻은 분할 결과를 개선하여 장기들 간의 경계를 명확히 하고 잘못 예측된 부분을 수정하는 기능을 수행한다. 분할 결과를 개선하는 기능을 학습하기 위해 다양한 3차원 의료 영상 분할 모델로부터 생성된 분할 결과 데이터를 RefNet의 학습데이터로 사용하며, 이용하는 의료 영상 분할 모델의 개수에 따라 많은 양의 학습 데이터를 확보할 수 있기 때문에 3차원 의료 영상 데이터의 개수가 적더라도 효과적으로 훈련시킬 수 있다. RefNet의 성능 개선을 평가하기 위해 synapse[8]에서 제공하는 3차원 의료 영상 데이터 집합을 사용하였으며, 해당 데이터 집합에 대해 가장 좋은 성능을 나타내고 있는 Swin UNETR을 비교 모델로 사용한다. Swin UNETR 본래의 영상 분할 결과와 RefNet을 통해 개선된 영상 분할 결과에 대해 각각 정답 영상과의 Dice 점수를 계산하였으며, RefNet이 모든 장기에 대해 분할 성능을 평균적으로 1.27% 향상시키는 결과를 보여주었다.

2. 본 론

RefNet을 이용한 3D 의료 영상 이미지 분할 과정은 그림 1과 같다. 사전 학습된 3D 의료 영상 분할 모델이 결과값(raw output)을 출력하고, RefNet은 이를 학습 및 보정하여 개선된 결과값(refined output)을 출력한다. Swin UNERT을 통해 얻는 분할 결과 데이터는 3차원 범주형(categorical) 데이터이며, 각 픽셀은 배경 및 12개의 장기를 나타내는 분류 번호를 가지고 있다. RefNet은 효율적인 보정을 위해 장면 별로 나누어서 작업하며, 구체적으로 3차원 데이터를 깊이에 따라 나누어서 3차원 깊이 값과 동일한 개수의 2차원 데이터로 변환하고, 각각의 2차원 데이터를 입력 데이터로 사용한다. 13개의 정수(integer)를 갖는 2차원 범주형 데이터를 효과적으로 학습하기 위해 RefNet은 임베딩 레이어(embedding layer)를 활용한 데이터 변환 작업을 수행한다.

2.1. RefNet

RefNet의 구조는 UNet[1]을 기반으로 설계되었다. RefNet 훈련 또한 UNet에서 사용된 교차 엔트로피(cross-entropy)와 Dice 손실함수 두 가지를 모두 사용한다. 두 손실함수는 다음과 같다.

$$L_{CE}(P,T) = -\sum_{i=1}^{I}\sum_{c=1}^{C} T_{i,c}\, log\, P_{i,c} \qquad (1)$$

Algorithm 1. Training procedure of RefNet

```
Require: a training dataset for 3D Medical image segementation D_tr = (X_tr, Y_tr), number of repetitions K, number of training steps T, a function to build an 3D Medical image segmentation model F, and a function to initialize a RefNet G
Output: a trained RefNet U
X̃, Ỹ ← ∅, ∅   # an empty dataset for RefNet training
for i ← 0 to K do
    M_i ← F(D_tr)   # new model construction
    O_tr ← M_i(X_tr)   # make a prediction
    X̃ ← X̃ ∪ O_tr   # collecting the predicted labels
    Ỹ ← Ỹ ∪ Y_tr   # collecting the ground truth
Endfor
U ← G()   # initialize a RefNet
for t ← 0 to T do
    X_b, Y_b ← getBatch(X̃, Ỹ)
    Õ ← U(Embedding(X_b))
    L ← crossEntropyLoss(Õ, Y_b)   # 1st loss
    L ← L + diceLoss(Õ, Y_b)   # 2nd loss
    L ← L + crossEntropyLoss(Õ, X_b)   # 3rd loss
    RMSprop(L).backpropagation()   # RefNet update
Endfor
```

알고리즘 1. Training procedure of RefNet

$$L_{Dice}(P,T) = 1 - \frac{1}{C}\sum_{c=1}^{C}\frac{2 * \sum_{i=1}^{I} P_{i,c} \cdot T_{i,c} + \varepsilon}{\sum_{i=1}^{I} P_{i,c} + \sum_{i=1}^{I} T_{i,c} + \varepsilon} \qquad (2)$$

식(1)과 (2)에서 P와 T는 각각 RefNet의 결과와 실제 레이블 값(ground truth)을 의미하고, C는 클래스의 개수, I는 전체 픽셀의 개수이다. UNet의 손실함수 이외에도 RefNet은 안정적인 영역 분할 결과값 개선을 위해 RefNet의 출력값이 Swin UNETR의 본래 결과값과 크게 달라지지 않도록 두 결과값 간의 교차 엔트로피 손실함수를 추가적으로 사용한다.

$$L_{CE}(P,S) = -\sum_{i=1}^{I}\sum_{c=1}^{C} S_{i,c}\, log\, P_{i,c} \qquad (3)$$

식(3)에서 S는 Swin UNETR의 영역 분할 결과값이다. 따라서, RefNet의 전체 손실함수는 다음과 같다.

$$L(T,S,P) = L_{CE}(P,T) + L_{Dice}(P,T) + L_{CE}(P,S) \qquad (4)$$

3. 실 험

3.1. 데이터 집합

3D 의료 영상 분할에 대한 벤치마크를 위해 Synapse 데이터 집합을 사용했다. Synapse 데이터 집합은 12가지의 장기(organ) 레이블을 포함하는 30개의 CT 영상으로 구성되어 있다. Synapse 벤치마크를 이용하는 연구들은 공통적으로 특정 18개를 학습 데이터, 나머지 12개를 검증 및 평가 데이터로 사용한다[9, 10].

3.2. RefNet 학습 과정

RefNet 학습 과정은 알고리즘 1과 같다. RefNet은 K개의 영역 분할 모델로부터 수집된 분할 결과(O_{tr})을 학습 데이터로 사용한다. RefNet은 2.1절에서 설명한 것과 같이 3가지 손실함수를 사용하고, 최적화 방법으로 RMSprop을 사용했으며, 학습률(learning rate)은 1e-5, 그리고 50 회 반복(iteration)을 통해 학습되었다.

표 2 분할 모델과 RefNet의 예측 결과에 대한 Dice 점수 비교

Organ	Raw	Refined	Improvement
Background	0.996	**0.997**	+0.11%
Spleen	0.846	**0.880**	+3.99%
Right Kidney	0.807	**0.811**	+0.46%
Left Kidney	0.833	**0.836**	+0.37%
Gallbladder	0.630	**0.649**	+3.03%
Esophagus	0.738	**0.742**	+0.50%
Liver	0.924	**0.961**	+4.03%
Stomach	0.759	**0.762**	+0.41%
Aorta	0.896	**0.904**	+0.94%
Inferior vena cava	0.851	**0.853**	+0.22%
Portal vein and splenic vein	0.697	**0.703**	+0.88%
Pancreas	0.720	**0.720**	+0.04%
Right and Left Adrenal Gland	0.595	**0.604**	+1.53%
Average	0.778	**0.787**	+1.27%

4. 결 과

RefNet의 정량적 평가를 위해 synapse 평가 데이터 집합에 대한 Swin UNETR의 dice 점수 및 RefNet을 통해 개선된 영역 분할 결과의 dice 점수를 계산하였다 (표 2). RefNet이 배경 및 13개의 장기들에 대한 dice 점수를 평균적으로 1.27% 향상시키는 것을 확인하였다. 장기들 중 넓은 면적을 차지하는 간(liver)과 비장(spleen)에 대해서는 각각 4.03%와 3.99% 향상되었으며, 이는 경계 부분을 많이 가지고 있어 개선 효과가 가장 크게 나타났다. 모든 장기에 대한 dice 점수 향상은 RefNet가 Swin UNETR의 기능을 효과적으로 개선할 수 있음을 보여준다.

그림 2는 Swin UNETR의 분할 예측 결과 및 RefNet의 개선 결과를 시각적으로 나타낸 것이다. RefNet의 개선 과정을 통해 매끄럽지 않았던 경계면이 다듬어진 것을 확인할 수 있다. 입력 데이터인 CT 영상과 정답 레이블을 비교해보면 각 장기들 간의 구분이 어려운 부분들이 존재하는 것을 찾아볼 수 있다. 두 번째 행의 간(Liv) 영역 안에 나타나 있는 정맥(Veins) 영역을 구분하는 것은 쉽지 않기 때문에, Swin UNETR의 결과에서 간과 정맥의 경계가 매끄럽지 않은 것을 확인할 수 있다. 반면에 RefNet으로 개선된 결과에서는 정맥의 경계가 매끄럽게 보정되었다.

5. 결론 및 향후 연구

본 연구에서는 3차원 의료 영상 분할의 문제점을 분석하고, 이를 극복하기 위한 RefNet 모델을 제안하였다. RefNet은 예측된 분할 결과를 학습하며, 이러한 전략은 3차원 의료 영상 데이터의 부족 문제를 회피하면서, 동시에 기존 모델의 분할 성능을 효과적으로 향상시킬 수 있음을 보여주었다.

하지만 영역 분할 모델의 결과값만을 입력 받고 있기 때문에 결과값을 크게 개선하기 어렵다. 입력 CT 영상을 고려하면 보다 효과적인 개선이 가능할 수 있기 때문에, 추후 연구에서는 입력 CT 영상 또한 RefNet의 입력으로 사용할 수 있는 방법을 개발할 계획이다. 또한 Swin UNETR 이외에 다양한 분할 모델들을 사용하여 제안하는 접근법의 성능을 평가할 예정이다. 그리고 의료영상 뿐만 아니라 CCTV와 같은 영상에서도 적용하여 스마트시티의 지능형 CCTV에서 정확한 객체 분할을 가능하게 할

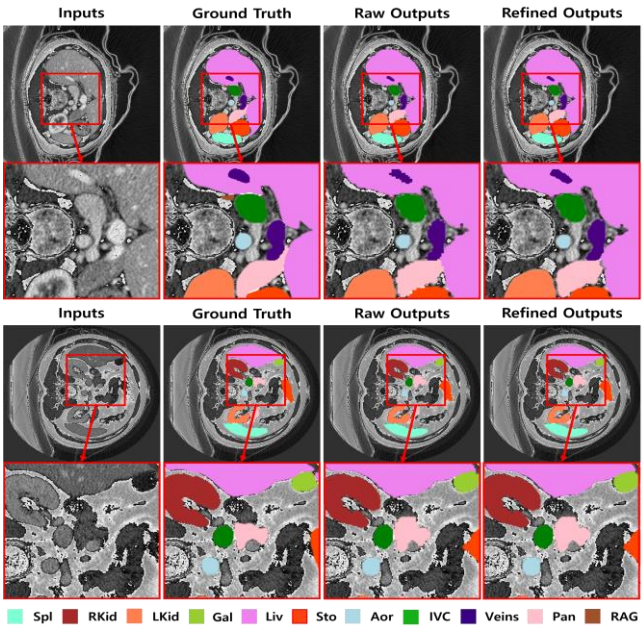

그림 3 분할 모델의 예측 결과와 RefNet의 예측 결과 시각화

것이다.

참 고 문 헌

[1] Ronneberger, Olaf, et al. "U-net: Convolutional networks for biomedical image segmentation." Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. Springer International Publishing, 2015.

[2] Ji, Shuiwang, et al. "3D convolutional neural networks for human action recognition." IEEE transactions on pattern analysis and machine intelligence, vol. 35, no. 1, pp. 221-231, 2012.

[3] Vaswani, et al. "Attention is all you need." Advances in neural information processing systems, vol. 30, 2017.

[4] Tang, 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] 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.

[6] Zhang, Yifan, et al. "Collaborative unsupervised domain adaptation for medical image diagnosis." IEEE Transactions on Image Processing, vol. 29, pp. 7834-7844, 2020.

[7] Lee, Hong Joo, et al. "Structure boundary preserving segmentation for medical image with ambiguous boundary." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 4817-4826, 2020.

[8] Landman, et al. "Miccai multi-atlas labeling beyond the cranial vault-workshop and challenge." Proc. MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge, Vol. 5, pp. 12, 2015.

[9] Zhou, Hong-Yu, et al. "nnformer: Interleaved transformer for volumetric segmentation." arXiv preprint arXiv:2109.03201, 2021.

[10] Huang, Xiaohong, et al. "MISSFormer: An Effective Transformer for 2D Medical Image Segmentation." IEEE Transactions on Medical Imaging, 2022.