

# 의료 데이터 불균형 문제 해결을 위한 생성적 적대 신경망 기반 데이터 증강

최재홍<sup>1</sup>, 이승리<sup>1</sup>, 서영재<sup>1</sup>, 서원진<sup>1</sup>, 허종욱<sup>1\*</sup>

<sup>1</sup>한림대학교 소프트웨어학부

jae\_apple@naver.com, tmdrn9@naver.com, ssyyjj1012@naver.com, tjdnjswls3@naver.com, juhoh@hallym.ac.kr

## Generative Adversarial Networks Based Data Augmentation to Address Medical Data Imbalances

Jae-Hong Choe<sup>1</sup>, Seung-Lee Lee<sup>1</sup>, Young-Jae Seo<sup>1</sup>, Won-Jin Seo<sup>1</sup>, Jong-Uk Hou<sup>1\*</sup>

<sup>1</sup>Division of Software, Hallym University

### 요 약

발병률이 낮은 병은 데이터 불균형 문제가 발생하며, 이는 의료계에서 겪는 원초적인 문제이다. 이런 불균형 문제를 해결하고자 Pix2Pix 로 생성적 적대 신경망 기반 의료 이미지 증강 기법을 설계하여 데이터 불균형 문제 해결 및 성능을 향상시켰다. 합성 데이터의 추가 및 기하학적 데이터 증강의 유무에 대한 4 가지 시나리오로 성능을 비교하여 제안된 기법이 가장 효과적임을 보인다.

### 1. 서론

데이터 불균형 문제는 산업, 의료등 많은 분야에서 발생하는 문제이다. 특히나 의료분야에서는 발병률이 낮은 병의 경우 발병 데이터가 부족하여 데이터 불균형 문제가 도드라진다. 본 논문에서는 신장과 신장암 데이터셋으로 연구를 진행한다. 신장암 뿐만 아니라 많은 암들은 장기와 종양 형태가 매우 다양하므로, 의료계에서는 장기와 종양의 형태를 고려한 수술 계획 기술과 수술 결과와의 연관성, 그리고 관련 기술들의 발전에 대한 관심이 많다. 이미지 분할 기술(Image Segmentation)은 이러한 노력과 발전을 위한 유망한 도구로써 사용된다.

딥러닝 기술 활용에 있어 가장 필수적인 해결 과제는 학습에 필요한 데이터를 충분히 확보하는 것이다. 의료 데이터를 확보하는 과정에서 개인정보 보호 문제나 발병률이 높지 않은 경우에는 충분한 데이터 확보가 어렵다. 딥러닝 학습에서 이미지 데이터가 부족한 경우 원본 데이터를 회전 확대 후 자르기 상하좌우 반전 등 여러 기하학적인 방법을 통해 데이터를 증가하는 방법이 활용되고 있다. 하지만 이 방법으로 추가적인 정보를 얻기에는 한계가 있다. 이를 바탕으로

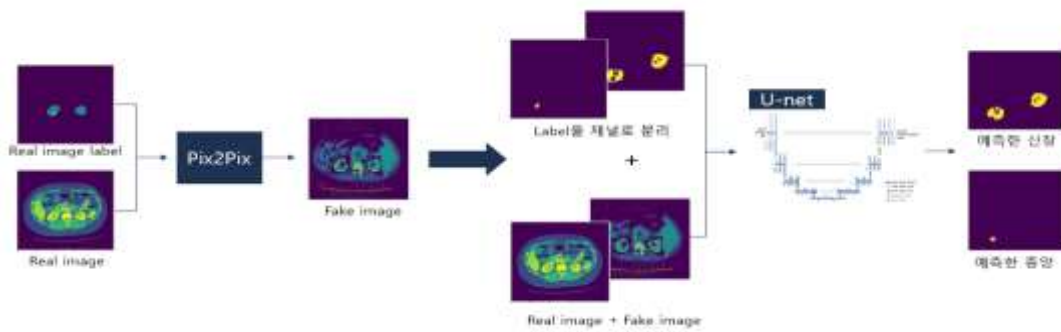
데이터양에 대한 필요성을 느끼고 생성적 적대 신경망을 통해 합성 데이터를 생성하여 데이터의 불균형 문제를 해결한다. 신장과 신장암에 대한 분할을 진행하여 신장암의 진단을 돕고 궁극적으로 기계학습 성능 향상 및 기술 개발을 하고자 하는 목적을 가지고 있는 바이다.

### 2. 관련 연구

본 연구에서는 딥러닝 모델인 생성적 적대 신경망(GAN)[1]을 통해 합성 데이터를 생성하여 데이터 불균형 문제를 해결하고자 한다. 특히 의료 데이터는 규제와 공유에 보수적이기 때문에 데이터 확보가 어렵다. 하지만 인공지능 신경망을 활용하여 생성한 이미지로 데이터를 늘린다면 규제 문제에서 벗어나 데이터 부족 문제를 해결이 가능하다. 이는 궁극적으로 해결하고자 하는 문제의 기계학습 성능을 향상시키는 연구이다.

CT 데이터에서 신장과 신장암을 이미지 분할하는 것뿐만 아니라, 생성적 적대 신경망을 사용하여 신장, 신장암에 대한 가짜 CT 데이터를 생성한 후 이에 대한 이미지 분할 모델의 성능을 비교한다. 데이터 증강을 통한 성능 비교를 위해 3 가지의 모델을 이용했다. 모델은 직접 구현한 U-Net[2]의 복잡도를 다르게

\* 교신저자(corresponding author)



(그림 1) 생성적 적대 신경망을 사용한 이미지 분할 기술 학습 실험 구성도.

한 두 개의 모델과 선 학습이 된 FPN(Feature Pyramid Network)[3]모델을 사용한다. 이를 데이터 증강(data augmentation)라이브러리 사용 유무와 합성 데이터 포함 유무별로 총 4 가지의 시나리오로 성능 분석을 진행하였다.

### 3. 실험 환경

#### 3.1. 방법론

합성곱 신경망(CNN) 아키텍처 모델은 생명의학분야에서 이미지 분할을 목적으로 제시한 U-Net 구조를 참고하여 모델을 구축하였다. 해당 모델은 인코더, 디코더 기반 모델이다. 인코더 단계에서 이미지의 특징 정보를 얻고, 디코더 단계에서는 인코딩 된 정보를 통해 분할 지도를 얻는다(그림 1).

생성적 적대 신경망 모델은 라벨 정보를 유지할 수 있는 Pix2Pix[4]를 사용하였다. Pix2Pix 를 통해 생성된 합성 이미지는 학습 데이터셋에만 추가한다. 즉 검증과 추론 작업에서는 실제 원본 데이터만을 사용한다. 정확한 비교를 위해 optimizer 는 Adam, learning rate 는 0.001, epoch 은 100 으로 고정하여 학습을 진행한다.

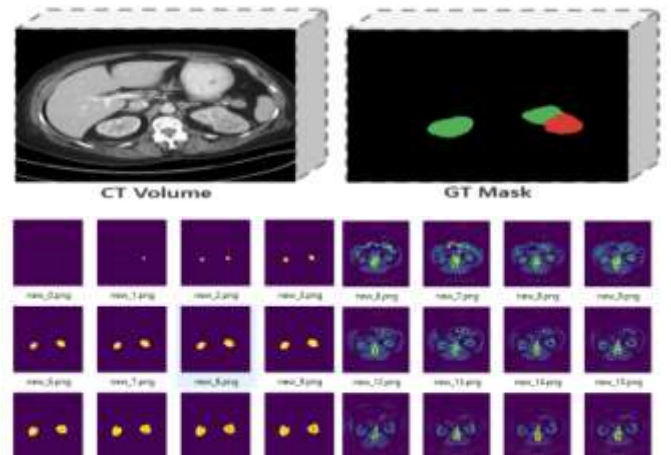
#### 3.2. 데이터 셋

데이터 셋은 “The KiTS19 Challenge Data” [5]라는 챌린지 데이터를 가져왔으며 6400 장으로 총 100 명의 환자로 그림 2 와 같이 구성이 되어있다. DCM 형식의 CT 영상은 512x512 사이즈로 되어 있으며, 일반적으로 1024~3096 의 HU(Hounsfield unit) 값을 포함한다. 라벨 값은 0~2 로 이루어져 있는데 0 은 배경, 1 은 신장, 2 는 종양이다.

적대 신경망을 통한 새롭게 생성된 이미지인 합성 이미지는 1333 장으로 신장 혹은 종양이 포함된 이미지만 생성하였다.

#### 3.3. 데이터 전처리

전체 데이터를 학습 데이터(70%), 검증 데이터(20%)와 시험 데이터(10%)으로 분류했다. 합성 이미지는



(그림 2) 실제 데이터(위), 합성 이미지(아래).

학습 데이터에만 포함되도록 했다. CT 이미지는 신체 부위에 따라 픽셀값의 범위(window width)와 기준이 되는 픽셀값(window center)을 조절해 이미지를 사용한다. 실험적으로 학습이 가장 잘되는 픽셀값의 범위는 400, 기준이 되는 픽셀값은 0 으로 고정했다. 이미지 분할 모델 예측의 채널은 훈련된 클래스의 개수와 동일하다. 정답을 클래스별로 채널을 나눠주는 작업이 필요하며 각각의 채널이 하나의 클래스를 대표한다. 하나의 채널로 구성되었던 클래스를 3 채널로 나누어 0 채널은 배경, 1 채널에는 신장 2 채널은 종양으로 구성했다. 데이터 증강을 위한 기하학적 변환은 회전, 이동, 확대 및 축소를 적용하였다. 또한, 데이터 증강 기법은 입력 이미지와 라벨 이미지를 동일하게 적용해야 하며 의료 데이터는 다양한 기하학적 변환을 적용하지 못한다는 한계가 존재한다.

#### 3.4. 성능 지표

$$IoU = \frac{|B \cap B^{gt}|}{|B \cup B^{gt}|} \quad (1)$$

IoU (Intersection over Union)는 라벨링된 영역( $B^{gt}$ )과 예측한 영역( $B$ )이 정확히 같다면, 1 이 되며 그렇지 않을 경우에는 0 이 되며 수식 (1)과 같은 식으로 계

산된다. 최종 테스트 성능측정에선 한 장마다 IoU 점수를 평균내 사용했다.

<표 1> 시나리오에 따른 실험 구성

시나리오 1	기존 데이터셋에 아무런 데이터증강을 사용하지 않았을 때
시나리오 2	기존 데이터셋에 선형적인 변형으로 인한 데이터증강을 했을 때
시나리오 3	기존 데이터셋에 합성 데이터를 추가하고 기하학적인 변형으로 인한 증강은 하지 않았을 때
시나리오 4	기존 데이터셋에 합성 데이터를 추가하고 기하학적인 변형으로 인한 증강도 했을 때

### 3.5. 실험 구성

표 1 을 확인하면 실험은 총 4 가지의 시나리오로 구성했으며 시나리오 1,2 는 원본 데이터만 활용하고 시나리오 3,4는 합성 이미지를 추가한 상황이다. 모든 실험은 직접 구성한 2 개의 U-Net 과 FPN 으로 실험을 진행했다. U-Net1 은 파라미터 수가 적은 가벼운 모델을, U-Net2 는 상대적으로 파라미터 수가 많은 무거운 모델로 구성하였다. 다른 모델에서도 합성 데이터를 통한 데이터 증강으로 성능향상이 적용되는지 확인하기 위해 이미 학습된 FPN 으로도 실험을 진행했다.

### 4. 실험 결과

모든 상황에서 시나리오 4 번이 가장 좋은 성능을 얻은 것을 확인할 수 있다(표 2). 특히 U-Net1 의 성능 결과를 보면 성능향상 폭이 상당히 큰 것을 확인된다. 기존 데이터만 학습을 진행하면 58%로 상당히 저조하지만, 합성 이미지와 기하학적 변환을 적용하는 순간 크게 32%의 성능향상이 있었다. 그러나 U-Net2 와 FPN 에서는 모델의 복잡도가 증가하면서 성능은 비교적 비슷하거나, 기하학 변형하지 않을 때 학습 데이터를 그대로 외워버리는 과적합(overfitting) 현상이 나오기도 했다. 이런 결과를 보아 모델이 깊어질수록 데이터를 세부적으로 보는데, 생성한 합성 데이터는 정교하지 않아 큰 성능향상을 이뤄내진 못한 것으로 보인다. 결론적으로 합성 데이터의 품질을 높인다면 더 높은 성능향상을 이뤄낼 수 있을 것이다.

<표 2> 시나리오에 따른 실험 결과

	U-Net1	U-Net2	FPN
시나리오 1	58.51%	89.61%	88.98%
시나리오 2	85.70%	88.93%	91.28%
시나리오 3	88.70%	87.49%	89.74%
시나리오 4	90.73%	90.76%	91.41%

### 5. 결론 및 기대효과

합성 데이터는 본래의 데이터 수가 제한된 상황에서 데이터를 늘리는 데 효과적이다. 이를 통해 다양한 데이터 분야에서 개인정보에 제한 없이 데이터를 확보하는 것이 가능해진다. 본 기술이 의료 분야에만 국한되어 적용하는 것이 아닌 상대적으로 데이터가 부족한 산업까지 확장이 가능하다.

본 논문에서는 신장암의 데이터를 사용해, 가상의 신장과 종양 데이터를 생성하고 이를 활용하여 신장 및 종양 검출 모델의 성능을 개선했다. 합성 데이터가 신장 종양 검출 모델의 성능을 개선했으므로, 생성적 적대 신경망으로 생성한 합성 데이터가 유의미한 데이터임을 확인하였다. 생성적 적대 신경망을 통해 부족한 데이터를 증강하는 것 자체가 가치 있는 과정이며, 이를 통해 인공지능 성능들을 높일 수 있다.

### 사사

이 논문은 2022 년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No.2022R1A4A1033600).

### 참고문헌

- [1] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. Communications of the ACM, 63(11):139–144, 2020
- [2] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention, pages 234–241. Springer, 2015.
- [3] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 2117–2125, 2017
- [4] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1125–1134, 2017
- [5] Nicholas Heller, Niranjana Sathianathan, Arveen Kalapara, Edward Walczak, Keenan Moore, Heather Kaluzniak, Joel Rosenberg, Paul Blake, Zachary Rengel, Makinna Oestreich, et al. The kits19 challenge data: 300 kidney tumor cases with clinical context, ct semantic segmentations, and surgical outcomes. arXiv preprint arXiv:1904.00445, 2019