소량의 의료데이터 학습을 위한 도메인내 종류가 다른 데이터를 활용한 자기 지도학습

김영진*, 조미경**
*동명대학교 컴퓨터미디어공학과
**동명대학교 게임공학과
e-mail : qpqp1996@gmail.com, mgcho@tu.ac.kr

Self-supervised learning using different type of data within the domain for small-volume medical data learning

Yeong-Jin Kim*, Mi-Gyung Cho**
*Dept of Computer Media Engineering, TongMyong University
**Dept of Game Engineering, TongMyong University

요 약

인공지능 학습을 위한 의료 데이터셋 구축은 일반적인 데이터셋 구축보다 많은 비용이 필요하므로 학습에 충분한 양의 의료 데이터셋을 확보하는 것은 쉽지 않다. 본 논문은 해당 문제를 해결하기 위한 방법 중 하나인 자기 지도학습 모델을 병리영상 데이터에 적용하여 몇 가지 실험을 통해 자기 지도학습의 성능을 살펴보았다. 학습을 위한 유방암 데이터셋의 개수를 전체의 5%, 25% 그리고 50%로 제한하고, 자기 지도학습 모델인 RotNet과 SimCLR을 적용하여 유방암 유무를 예측할 때 자기 지도학습의 성능과 전체 데이터셋에 대한 지도학습의 성능을 비교해 보았다. 또한 다른 조직 병리영상으로 Pretext task를 적용할 경우 유방암 분류 모델의 성능이 어느 정도 향상되는지 살펴보았다. 실험결과 전체 데이터의 25% 데이터만 가지고 자기 지도학습을 적용해도 전체 데이터셋에 대한 지도학습 모델의 성능을 낼 수 있었다.

1. 서론

기존의 병리 영상분석에는 의사가 현미경으로 환자에게 채취한 병리조직 슬라이드 이미지(Whole Slide Image)를 분석해야 했으나, 최근 인공지능의 발달로 딥러닝 기반의 자동 종양 진단과 등급분류가 가능해졌다.

하지만 인공지능을 학습하기 위해서는 많은 양의 데이터셋이 필요한데, 의료 데이터의 경우 법적 문제, 개인정보 문제, 데이터 가공을 위해 전문의가 직접 라벨링을 해야하는[1] 등 일반적인 데이터셋에 비해 데이터 구축이 더욱 어렵다는 문제점이 존재한다.

해당 문제점을 보완하기 위한 해결책 중 하나는 비지도 학습 방법중 하나인 자기지도 학습(Self-supervised learn)이다. 일반적으로 딥 러닝모델은 모델의 사이즈가 증가함에 따라 정확도가 향상되는데 큰 모델을 사용할 때 학습용 데이터가 작을 경우 과적합 문제 등 학습에 어려움이 있다. 반면 자기지도 학습은 데이터의 정답이 아닌 전반적인 특징을 먼저 학습하는 데에 큰 비중을 두고 있기 때문에 비교적 적은 레이블 데이터로 큰 사이즈의 모델을 학습해야 할 경우 다른 학습모델보다 유리하다.

본 논문에서는 학습용 유방암 데이터셋[2]의 개수를 전체의 5%, 25%, 50%로 제한하고 자기 지도학습 모델인 RotNet[3]과 SimCLR[4]을 이용해 유방암 데이터셋을 정상(Benign), 비정형 유관 증식증(ADH), 유방 상피내암(DCIS)을 분류하도록 학습한 결과가 지도학습 모델보다 좋은 성능을 보였음을 확인했다 또한 같은 도메인이지만 다른 조직의 데이터셋인 신장(Kidneys) 데이터셋을 추가로 사용해 학습한 모델이 유방암 데이터셋만 사용해 학습한 모델보다 성능이 좋았으며 특히 데이터의 25%를 사용했을 경우 모든 데이터를 사용해 학습한 지도학습 결과를 상회함을 확인했다.

2. 관련연구

자기 지도학습이란 정답이 없는 데이터를 기반으로 스스로 학습할 구실을 정해 데이터의 전반적인 특징을 이용해 학습하는 방법이다.

자기 지도학습은 학습할 구실을 정해 학습하는 Pretext task와 실제 레이블이 있는 데이터를 학습하는 Downstream task를 통해 학습을 유도하는 2단계의 아키텍처로 설계되어 있다. 자기 지도학습방법은 정답이 없는 데이터로 부터 좋은 representation 학습을 유도하는 것이 중요하다. 모델의 학습은 데이터 내 임의의 정답을 정해

학습하는 식으로 Pretext task모델을 학습한 후 학습된 모델을 Downstream task학습에 전이 학습하여 모델을 훈련한다. 자기 지도학습 방법은 크게 Self-Prediction, Contrastive-learning 그리고 Generative SSL으로 구분된다.

Self-Prediction은 데이터를 이루고 있는 특정값들 사이의 상관관계가 있다고 가정하고 이전값을 토대로 이후값을 예측하거나 데이터의 일부를 임의로 가린 후 가려지지 않은 영역의 전후 관계를 파악해 가려진 데이터를 예측해 관계를 학습하는 등 데이터에 변형을 가하더라도 데이터의 본질적인 정보는 동일할 것이라 가정하고 데이터의 representation을 학습해 임의의 정답을 예측한다. Self-Prediction을 기반으로 한 모델로는 GPT와 메타의 BERT 등이 있다.

Contrastive-learning은 레이블이 지정되지 않은 데이터 간의 비교를 통해 학습하는것을 목표로 데이터의 embedding 공간에서 유사한 데이터 쌍(Positive pair)들의 거리는 가깝게, 유사하지 않은 데이터 쌍(Negative pair)의 거리는 멀게 유도해 학습한다. 즉 모델이 여러 관점으로부터 데이터의 공통된 정보를 추출해 학습하는 방법이다. representation 학습에 있어서 간단하면서 효과적이라는 장점이 있다.

Generative SSL은 원본 이미지에 노이즈나 색 변조등 변형을 가한 후 변형된 이미지를 복원하면서 이미지의 관계를 학습하거나 판별자(Discriminator)가 생성자(Generator)가 생성한 가짜 이미지의 진위 여부뿐 아니라 앞선 자기 지도학습처럼 임의의 정답을 예측하는 pretext task방법을 함께 학습하는 적대적 생성 신경망(GAN)[5] 프레임워크를 활용할 수 있다.

본 실험에서 사용할 모델은 Self-Prediction모델로 분류되는 RotNet과, 대표적인 Contrastive-learning모델인 SimCLR을 사용한다.

3. 실험 방법

3.1 데이터셋

'유방암 병리조직영상' 데이터[2]는 정상(Benign), 비정형 유관 증식증(Atypical Ductal Hyperplasia, ADH), 유방상피내암(Ductal Carcinoma In-Situ, DCIS) 총 3가지 종류의 클래스를 가진 이미지 데이터이다. 이미지는 색이 있는 슬라이드 이미지(WSI)를 224x224 크기의 패치로 나눠져 있다.

본 실험에서는 데이터셋이 작을 경우의 학습결과를 도출하기 위해 표 1과 같이 학습용 데이터의 사용 비율을 5%, 25%, 50%로 제한한다. 지도 학습모델과 자기 지도학습모델을 각각의 데이터 비율마다 학습하고 마지막으로 모든 데이터를 사용해 학습한 지도 학습모델의 결과를 비교한다. 자기 지도학습모델은 RotNet과 RotNet에 색상 왜곡을 적용한 모델, 유방암 데이터셋만 사용한 SimCLR모델과 신장 데이터셋을 추가로 학습한 SimCLR모델을 실험에 사용하였다.

학습에 사용된 데이터를 8:2의 비율로 학습-검증 데이터를 분리해 사용했으며 학습 데이터의 클래스별 비율을 동등하게 했다. 모든 모델들은 동일하게 ImageNet가중치와 ResNet50[6]을 사용했으며, 100번의 에폭중 손실함수의 값이 가장 낮은 모델을 가장 좋은 모델이라 가정했다. 자기 지도학습은 3e-4의 학습율(learning rate)을 사용했고, 지도 학습은 1e-4의 학습율로 실험을 진행했다.

표 1. 학습에 사용된 유방암 데이터셋

데이터종류 / 데이터 비율		Benign	ADH	DCIS	Total
훈련데이터	전체	3080	3080	3080	9240
	5%	154	154	154	462
	25%	770	770	770	2310
	50%	1540	1540	1540	4620
테스트데이터		1136	1876	770	3782

3.2 실험에 사용한 모델

① RotNet

RotNet은 모델이 학습 데이터로부터 Self-supervision을 통해 특징을 구분할 수 있도록 이미지 하나를 각각 [0°, 90°, 180°, 270°]로 회전시켜 4장의 이미지로 증강 기법을 적용한 뒤, 그림 1과 같이 이미지의 회전각도를 예측하도록 Pretext task를 학습해 Downstream task모델에 전이 학습한다.

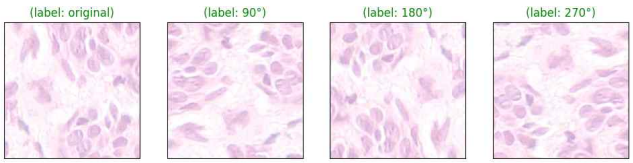

그림 1. 회전이 적용된 증강 데이터

② RotNet 모델에 색상 왜곡을 적용

그림 2와 같이 RotNet과 동일하게 이미지를 회전시켜 데이터를 증강시킨 뒤, 회전된 이미지에 색상 왜곡을 적용하고 회전각도를 예측하도록 Pretext task를 학습 해 Downstream task모델에 전이 학습한다.

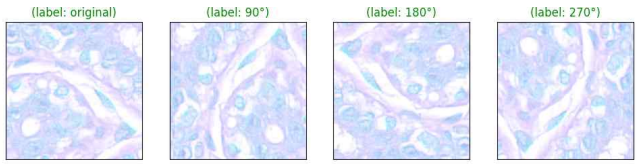

그림 2. 회전 이미지에 색상 왜곡을 적용한 증강 데이터

③ SimCLR

그림 3과 같이 데이터셋에 각 이미지당 두 개의 다른

증강 기법을 적용하고 같은 이미지에 증강이 적용된 두 이미지를 Positive pair라 하고, 나머지 이미지쌍을 Negative pair라 한다. 각 이미지쌍은 Encoder를 거쳐 2개의 representation을 생성하는데 생성된 representation들이 FC layer를 거치면 2개의 벡터가 생성된다 Embedding space상에서 Positive pair와 Negative pair를 구분하기 위해 유사도(Similarity)를 사용하는데, 유사도는 생성된 2개의 벡터와 코사인 유사도(Cosine Similarity)를 이용해 유사도를 측정한다. Positive pair간의 유사도는 높이고, Negative pair간의 유사도는 최소화 하는 Contrastive 손실함수인 NT-Xent[7]를 적용해 Pretext task를 학습한 뒤, 학습된 모델을 Downstream task에 전이 학습한다.

추가로 SimCLR의 Pretext task학습에 신장 데이터셋을 사용해 학습한 뒤, 유방암 데이터셋을 학습시키고, Downstream task에 전이 학습해 결과를 도출하였다. 학습에 사용된 신장 데이터는 PAIP[8] 데이터 세트를 사용했다. 해당 데이터셋은 조직병리영상 이미지이며 슬라이드 이미지를 224x224 크기의 패치로 나눠 레이블 없이 사용했으며 약 150,000장을 사용했다

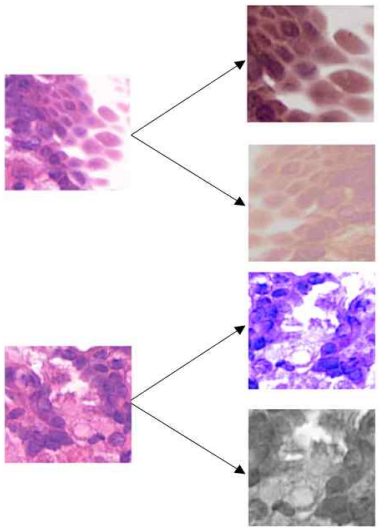

그림 3. 이미지의 코사인 유사도를 측정하기 위해 이미지당 두 개의 다른 증강 기법을 적용시킨 데이터

4. 실험 결과

표 2는 실험 방법 5가지의 학습 방법으로 훈련 데이터셋 각각의 비율만 사용해 학습한 후, 테스트 데이터셋으로 평가한 분류 정확도 결과이다.

표 2. 데이터셋 크기와 모델 종류에 따른 실험 결과

데이터 비율	모델	loss	accuracy
5%	Supervised	1.595	67.10%
	RotNet	0.7344	64.11%
	RotNet+Jitter	0.776	79.32%
	SimCLR	0.5545	75.59%
	SimCLR+kidneys data	0.4881	80.65%
25%	Supervised	0.5947	83.15%
	RotNet	0.3881	84.51%
	RotNet+Jitter	0.3466	86.64%
	SimCLR	0.3569	84.8%
	SimCLR+kidneys data	0.2785	88.96%
50%	Supervised	0.5538	88.7%
	RotNet	0.3864	85.76%
	RotNet+Jitter	0.2897	89.4%
	SimCLR	0.2767	88.79%
	SimCLR+kidneys data	0.2229	91.5%
100%	Supervised	0.3376	88.96%

5개의 실험방법 중 SimCLR을 활용해 신장 데이터셋을 먼저 학습시킨 후 유방암 데이터셋으로 학습시켜, 전이 학습한 모델의 성능이 가장 좋았으며 데이터셋이 적을수록 지도학습 모델과 큰 차이를 보였다. 특히 학습 데이터의 비율을 5%로 가장 적었을 때 손실함수값이 1.1이상 차이가 났고, 정확도도 13%이상 차이가 났으며 데이터의 25%만 사용했을 때 데이터를 100% 사용한 지도 학습모델의 분류결과를 상회함을 확인했다. 이는 Pretext task에서 다른 종류의 데이터를 이용해 모델을 학습해도 의미있는 학습을 수행했음을 의미한다. 전이 학습을 사용할 때 흔히 ImageNet의 가중치를 이용해 전이 학습을 적용하는 것이 일반적이지만 의료데이터는 ImageNet학습에 사용된 일반적인 데이터의 특징과 상이하기 때문에 ImageNet만 사용하는 것은 비교적 효과가 크지 않았던 것으로 추측되며 Pretext task와 Downstream task 두 단계를 거쳐 학습하는 자기지도 학습은 과적합 문제에서 비교적 자유로운 반면 ImageNet만 사용한 지도학습은 과적합 문제에 빠르게 직면했다. 따라서 유방암 데이터가 아니더라도 같은 도메인의 병리영상 데이터를 추가로 사용해 Pretext task 학습 후 Downstream task를 통해 학습한 모델의 결과가 가장 좋은 것으로 보인다.

5. 결론

실험을 통해 데이터가 적을 경우 빠르게 과적합된 지도학습보다 자가 지도학습인 RotNet과 SimCLR이 더 좋은 성능을 보였고, RotNet 학습시 색상 왜곡을 적용해 학습한 모델이 기존 RotNet의 결과보다 좋았으며 SimCLR을 이용해 같은 도메인의 신장 데이터셋을 추가로 학습시킨 뒤 유방암 데이터셋을 학습시키고 전이 학습을 진행한 모델의 성능이 가장 좋았음을 확인했다.

인공지능 학습을 위해 데이터 확보 문제는 필연적인 상황에 놓여있지만 데이터 구축이 쉽지 않고 특히나 의료데이터셋 구축은 더 많은 비용과 시간을 필요로 한다. 앞선

실험에서 데이터의 25%만 사용해 학습했을 때 모든 데이터를 사용한 지도 학습의 결과를 상회했던 것처럼 데이터를 효율적으로 사용해야 하는 것은 매우 중요하고 실질적인 문제이고 해당 문제에서 자기 지도학습은 좋은 대안이 될 수 있다. 더불어 상대적으로 데이터셋을 효과적으로 이용하기 위한 다양한 연구와 시도들이 지속적으로 이루어져야 할 것으로 보인다.

감사의 글

이 논문은 2021년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. 2021R1A2B5B01001789)

참고문헌

[1] 최연진, “데이터 라벨링에 뛰어든 의사들”, 한국일보, 2021, 02.16 https://www.hankookilbo.com/News/Read/A2021021609590003191

[2] Sheikh, T. S., Kim, J. Y., & Cho, M. . “Refined Attention Module for WSI Cancer Diagnosis.” In 2022 IEEE 5th International Conference on Knowledge Innovation and Invention (ICKII) (pp. 30-34), IEEE. 2022.

[3] S. Gidaris, P. Singh, and N. Komodakis, “Unsupervised representation learning by predicting image rotations.“ International Conference on Learning Representations (ICLR), 2018.

[4] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A simple framework for contrastive learning of visual representations.“ International conference on machine learning (ICML), 2020.

[5] I.J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio ”Generative adversarial networks“ Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2, MIT Press, pp. 2672-2680, 2014.

[6] He, K., Zhang, X., Ren, S., & Sun, J., ”Deep residual learning for image recognition.“ In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778), 2016.

[7] Kihyuk Sohn, “Imporved deep metric learning with multi-class n-pair loss objective“ Advances in Neural Information Processing Systems 29 (NIPS), 2016.

[8] 서울대학교 병원, 분당서울대학교병원 그리고 보라매의료원, “Renal Cell Carcinoma“ in Pathology ai platform (PAIP), 2020. http://www.wisepaip.org/paip