HANSHIN UNIVERSITY

Coursework Research Paper

School of Department of Information and Communication Hanshin University

폐암 유형 분류 문제에서의 효율적인 데이터 증강 기법

Efficient data augmentation technique in lung cancer type classification problem

Seungjae Kim

School of Department of Information and Communication Hanshin University, Osan-si 18101, Korea

[Abstract]

In this paper, the Chest CT-Scan images Dataset provided by Kaggle is used. As artificial intelligence is used in many fields, it is actively used and studied in the medical field. Training artificial intelligence models requires high-quality data and sufficient data. In order to apply artificial intelligence in the field, the quality of data is the most important in artificial intelligence as data collection and processing occupy the largest proportion. As Garbage in Garbage out, which is a famous word in the field of artificial intelligence, we can see the importance of data quality. Building a high-quality dataset takes a lot of cost and time. Usually, it requires 1 million or more data to train a CNN model. In the field of medical, which requires the most cost in image datasets, considerable costs will be required. Recently, data for learning is being constructed using generative models such as stable diffusion and GAN. In this paper, an experiment is conducted to increase the performance of the model only with the data augmentation technique within a limited data set.

Key word : Deep learning, Medical AI, Data augmentation

Ⅰ. 서 론

하루가 다르게 발전하는 AI의 발전에 따라, 어떤 도메인에서도 AI의 접목을 시도하고 있다. 이 논문에서는 AI에서 이미지를 다루는 분야인 컴퓨터 비전의 연구를 진행한다. 의학 분야에서도 AI를 접목시키는 연구가 적극적으로 진행되고 있는데, 전문성이 높은 의학 분야의 특성상, 인공지능 학습에 필요한 데이터 수집의 어려움이 있다. 인공지능 모델의 학습에 필요한 데이터셋을 구성하기 위해서는 데이터 라벨링 작업을 해야한다. 의학 분야는 데이터 라벨링에 드는 비용이 가장 높은 도메인 중 하나이다. CNN 모델을 훈련시키기 위해서, 모델의 깊이가 깊어짐에 따라 많게는 100만장에 가까운 데이터셋이 요구되는데, 의료 이미지 데이터셋을 그 만큼 구성하는데에는 상당한 비용이 요구되고 투자에 어려움이 있다. 생성형 AI가 발전하면서 GAN모델을 사용하여 학습용 데이터셋을 생성하고, 최근에는 Stable Diffusion모델을 사용하여 학습용 데이터셋을 생성하는 것이 가장 효율적인 방법으로 이용된다. 본 논문에서는 의학 분야에서도, CT로 촬영한 폐암 유형 이미지 분류 태스크에서 실험을 진행한다. 1%의 예측률에도 민감한 의학 분야에서 생성형 AI로 데이터셋을 생성했을 때 10%의 잘못된 데이터가 생성된다면 심각한 결과를 초래할 것이라고 판단한다. 이 연구에서는 한정된 데이터셋 내에서 데이터 증강 기법만을 이용하여 모델의 성능을 높이는 실험을 진행한다.

Ⅱ. 실험 환경

이 논문을 작성하기 위하여 실험을 진행한 환경에 대해 소개하겠다. 실험 서버는 RTX 4070를 한 장 사용하였고, 케글에서 제공하는 Chest CT-Scan images Dataset을 사용하였다. 총 785장의 폐의 CT 이미지의 train:test:valid 비율이 7:2:1로 구성되어 있다. Adenocarcinoma, Large cell carcinoma, Squamous cell carcinoma 3개의 클래스로 이루어져 있다. 폐암의 종류는 크게 비소세포폐암과 소세포폐암 두가지로 분류되는데, 위의 세 클래스는, 전체 폐암의 87%를 차지하는 비소세포폐암의 종류들이다. 인공지능 모델은 가장 기본적인 CNN모델을 합성곱 연산이 두 번 있는, 총 7개의 층으로 구성되어있게 만들어 사용하였다.

Ⅲ. 데이터 증강 기법에 따른 모델 성능 실험

3-1 mixup

인공지능 경연대회에서도 자주 볼 수 있었던 mixup 기법을 적용하여 실험을 진행해보았다.

$$\hat{x} = \lambda x_i + (1-\lambda)x_j$$

$$\hat{y} = \lambda y_i + (1-\lambda)y_j$$

그림 1. mixup 알고리즘 수식
Fig. 1. Formula of the Mixup Algorithm

그림 1의 수식을 해석하면 이미지 데이터와 라벨 데이터에 대하여 사용자 설정 값인, 람다의 비율에 대하여 두 그림을 섞는다고 해석할 수 있다[1]. 즉, mixup 기법은 두 이미지의 투명도를 50% 씩으로 설정하여 겹치는 것과 같은 기법이다. 성능만을 중요시하는 인공지능 경연 대회에서 자주 쓰이는 만큼, 모델의 성능이 올라갈 것이라고 가설을 세우고 실험을 진행하였다. 데이터 증강 기법을 적용하지 않은 상황과 mixup을 적용한 상 두 경우에서의 성능을 비교하기 위하여, 모든 파라미터 값들을 고정시키고 3번씩 모델 훈련과 테스트를 진행하였다.

Method	Validation set	Test set
mixup	**0.75**	**0.41**
aug x	**0.85**	**0.51**

표 1. 3-1 실험 결과
Table 1. Results of the first experiment

mixup을 적용한 실험군의 성능이 더 좋을 것이라는 예측과 달리, 데이터 증강 기법을 적용하지 않을 때 모델의 분류 성능이 더 좋았다. mixup 기법은, 몇 개의 픽셀이 중요한 의료 CT 이미지 도메인에서 좋은 성능을 내지 못한다고 결론을 내었다. 또한, CT 이미지의 흑백 이미지라는 특성에서 좋은 성능을 내지 못할 것이라는 가능성을 염두하고 다음 실험을 진행한다.

3-2 Frequently-Used Augmentaion and Image processing

Method	Validation set	Test set
RandomResizedCrop	0.52	0.44
RandomRotation(45)	0.71	0.31
RandomAdjustSharpness	0.78	0.38
RandomHorizontalFlip	0.76	0.43
GaussianBlur	0.86	0.56
Autoaugment	0.83	0.42
Histogram-equalization	0.39	0.19

표 2. 3-2 실험 결과
Table 2. Results of the second experiment

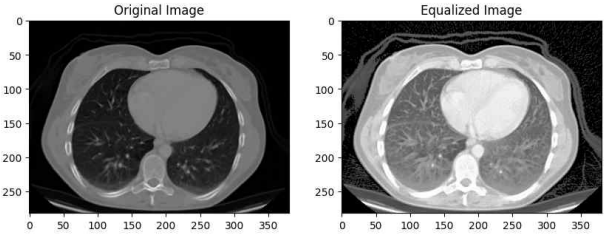

그림 2. 히스토그램 평활화를 적용한 이미지와 원본 이미지
Fig. 2. Images with histogram equalized and origin

본 실험에서는, computer vision에서 자주 볼 수 있는 데이터 증강 기법들을 비교 실험해보았다. 실험 전에, 각 클래스의 CT 이미지를 눈으로 비교해보았지만 의학적 도메인 지식이 부족하고, 같은 클래스의 이미지들의 규칙성을 찾을 수가 없었다. 하지만 이미지에서 하얀색 부분들로 폐암 유형을 분류할 수 있다는 것을 확인하였다. 이를 검증해보기 위해, 대비를 강조시키는 영상 처리 기법인, 히스토그램 평활화를 적용하여 실험을 진행했다. 그림2는 Adenocarcinoma 클래스의 한 이미지를 히스토그램 평활화를 적용하여, 원본과 비교한 것이다. 히스토그램 평활화를 적용함으로써 밝은 픽셀들의 값이 커져 대비가 증가된 것을 확인할 수 있다. 표 2를 보면, 히스토그램 평활화는 모델 학습에 악영향을 끼친 성능을 보여준다. 클래스가 3개 이므로 0.33 이상의 성능이 나오지 않는다면, 인공지능을 사용하는 것은 의미가 없다는 것을 알 수 있다. RandomRotation과 Histogram-equalization 두 기법이 0.33을 넘지 못하는 결과에서, 두 기법은 CT이미지에 대한 데이터 증강 기법으로는 부적합하다는 결론을 내릴 수

있다. 실험에서 가우시안 블러링을 적용할 경우에 폐암 유형을 분류하기 위한 특징이 흐려짐으로 모델 성능이 낮아질 것으로 예상했지만, 실험 결과에서 이미지를 뚜렷하게 해주는 기법인 샤프닝보다 블러링이 더 좋은 성능을 내었다는 것을 확인하였다. 블러링이 가장 좋은 성능을 낼 수 있었던 이유를, 이미지의 노이즈를 감소시켜 모델의 일반화 성능을 높일 수 있었다고 추측한다.

Ⅳ. 결 론

본 논문에서는 데이터 증강 기법만을 이용하여 CNN모델로 폐암 유형 이미지의 분류 성능을 높이기 위한 연구를 진행하였다. 결론적으로, 어떤 데이터 증강 기법도 적용하지 않은 실험과 가우시안 블러링을 적용한 실험, 두 실험만이 50%가 넘는 성능을 보여주었다. CT 이미지는 밝은 부분으로 장기와 혈관 같은 특징을 보여주는 성질로, 폐암 유형을 분류할 수 있는 특징 또한 밝은 부분으로 나타낸다. 이에 폐암 유형을 분류할 수 있는 특징 이외의 밝은 부분들을 노이즈라고 생각 할 수 있기 때문에, 이미지를 부드럽게 만들어주는 가우시안 블러링이 노이즈를 감소시켜주어 가장 좋은 성능을 낼 수 있게 해준 것이라고 생각할 수 있었다.

References

[1] H. Zhang, M.Cisse, Y.N. Dauphin, D. Lopez-Paz, “mixup: BEYOND EMPIRICAL RISK MINIMIZATION”, ICLR, 2018.