₩ 서울대학교 대학원 의과학과

학부연구생인턴 보고서

연 구 기 간: 28기(2021. 12. 27. ~ 2022. 2. 18.)

연 구 제 목: Creating a Medical Image Test Set Corresponding to Human Perceptual Similarity Judgment

성 명: 이나은 (인 또는 서명) **이나은**

지도교수 성 명: 박상민 (인 또는 서명)

(※ A4용지 10장 이내, 바탕체, 11포인트, 줄간격 160%. 아래 항목대로 기술하기 힘든 경우에는 해당 항목만을 기술함)

1. 연구요약 (영문의 경우 : 150단어 이내, 국문의 경우 : 750자 이내)

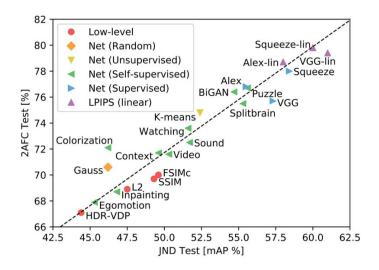
이미지 사이의 유사성을 수치화 하는 것은 어렵기에 본 연구를 통해 인간의 유사성 판단을 포함하는 medical image test set을 구축한다. 이를 위해 chest X-ray dataset에 distortion을 적용하고 폐렴 유무를 판별하는 classification model을 제작한다. Distortion은 기존 연구의 two alternative forced choice (2AFC) test와 just noticeable difference (JND) test를 구현하기 위해 수행했다. 그리고 distortion은 denoising autoencoder와 generative adversarial network (GAN)을 제작하여 CNN based distortion을 dataset에 적용하였다. 다양한 distortion 기법들을 통해 실제 algorithm output에서 보여지는 distortion을 최대한 구현하고자 하였다.

또한, classification model의 feature가 perceptual metric으로서 유용하다고 평가받고 있기에 해당 model을 제작하였다. 이를 위해 ImageNet dataset으로 사전 훈련된 ResNet50 model을 transfer learning을 수행하여 폐렴 유무를 판별하였다. 이를 통해 classification model의 perceptual loss가 인간의 인지와 부응하는지 검증하고 기존 연구가 medical domain으로 확장 가능한지 검증할 것이다.

2. 서론 (연구배경 및 연구목적)

기존의 이미지 사이의 유사성 판별 방법은 저조하며 두 이미지 사이의 유사성을 수치화하는 것은 어렵다. 그래서 인간의 유사성 판별에 부합하는 image dataset을 구축하는 연구가 활발히 진행되어 왔다. 그러나 일반적인 이미지와 의료 이미지는 도메인이 많이 다르다. 또한, 인간의 유사성 판단을 포함하는 medical image dataset은 존재하지 않는다. 따라서 본 연구를 통해 인간의 유사성 판단에 부합하는 medical image test set을 구축하고자 한다.

Dataset이 아닌 test set을 구축하는 이유는 아래의 그래프를 통해 설명할 수 있다. 기존 연구는 test set뿐만 아니라 train 및 validation set을 포함하여 인간의유사성 판별에 부합하는 전체 dataset을 구축하였다. 그래서 이를 통해 classification model을 훈련시켰고 이는 그래프에서 Learned Perceptual Image Patch Similarity (LPIPS) 에 해당한다. 즉, 해당 dataset으로 훈련하였을 때 가장좋은 성능을 보임을 알 수 있다. 그러나 supervised network인 squeeze net도 이와비등한 성능을 보인다. 이는 전체 dataset으로 모델을 훈련할 필요가 없다는 것을 시사한다. 또한, dataset 구축에는 상당한 비용과 시간이 소모되므로 test set만구축하기로 결정했다.



<Fig 1. Correlating Perceptual Tests>

Medical test set 구축에 이어 medical image dataset으로 다양한 classification model을 훈련하였다. 그 이유는 classification model의 feature가 perceptual metric으로서 유용하다고 평가받고 있으며 모델이 이미지의 유사성을 판단한 결과 인 perceptual loss가 인간의 인지와 부응하는지 검증하기 위해서이다.

또한, 기존의 연구는 위의 그래프처럼 일반적인 image dataset을 classification model에 적용하여 좋은 성능을 보였는데 이것이 medical image dataset에서도 유효한지에 대한 의문이 발생하여 medical image classification model을 제작하기로 하였다.

3. 실험방법

연구는 크게 두 가지 파트로 구성된다. 첫 번째는 기존의 medical image에 distortion을 적용하여 test set을 구축하는 것이다. 두 번째는 medical image classification model을 구축하는 것이다. 첫 번째 실험을 위해 기존의 연구를 참고하여 이미지에 distortion을 적용하였다.

기존 연구에서 기존 image인 reference image, 이를 왜곡한 distortion image와 이들 간의 사람의 유사성 판단으로 dataset을 구축하였다. 그리고 유사성 판단 방법에는 두 가지가 있다. 첫 번째는 two alternative forced choice (2AFC) test로 두 개의 distorted image가 있을 때 어느 것이 reference image와 더 유사한지 판단하는 방법이다. 두 번째는 just noticeable difference (JND) test로 하나의 distorted image가 reference image와 같은지 혹은 다른지를 판단하는 방법이다.

즉, 기존 연구는 reference image와 이에 distortion을 적용한 image와 두 가지 방법을 통한 인간의 유사성 판단으로 dataset을 구축하였다. 이 구성 방식에 따라 medical image test set을 구축하기 위해 medical image에 distortion을 적용하였 다. Distortion 기법은 아래의 표와 같이 traditional distortion과 CNN based distortion이 있다.

<Table 1. Distortions>

Traditional distortion		CNN based distortion	
Sub-type	Distortion type	Parameter type	Parameters
Photometric	lightness shift, color shift,	Input corruption	Null, pink noise,

	contrast, saturation		white noise,
			color removal,
			downsampling
Noise	Uniform white noise,	Generator	# layers,
	Gaussian white, pink,	Network	# skip connections,
	& blue noise,	architecture	# layers with
	Gaussian colored		dropout,
	(between violet and		force skip connection
	brown) noise,		at highest layer,
	checkerboard artifact		upsampling method,
			normalization method,
			first layer stride #
			channels in first
			layer
Blur	Gaussian,	Discriminator	Number of layers
	bilateral filtering		
Spatial	Shifting,	Loss/Learning	Weighting on oixel-
	affine warp,		wise,
	homography,		VGG,
	linear warping,		discriminator losses,
	cubic warping,		learning rate
	ghosting,		
	chromatic aberration		
Compression	jpeg		

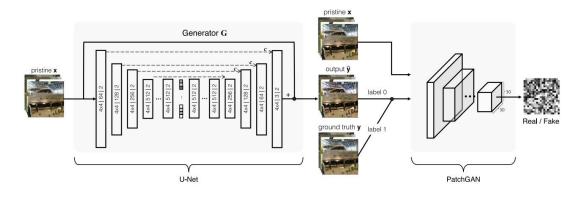
Traditional distortion은 image에 noise를 추가하거나 흐릿하게 하거나 (blurring) 이동시키는(shifting) 등의 기법이 포함된다. CNN based distortion은 cnn 모델의 구조나 input 등을 달리하여 image에 왜곡을 주는 방법이다. 이렇게 다양한 distortion 기법들을 통해 실제 algorithm output에서 보여지는 distortion을 최대한 구현하고자 하였다.

기존 연구에서 CNN based distortion의 방법론이 명시되어 있지 않아 우선 baseline으로 denoising autoencoder를 기반으로 distortion을 구현하였다. Denoising autoencoder는 이미지에 gaussian noise를 추가하고 이 noise를 제거하는 과정을 통해 새로운 이미지를 생성하는 역할을 수행한다. 이후 추가적인 문헌

고찰을 통해 기존 연구의 CNN based distortion을 구현하고자 Generative Adversarial Network (GAN)을 사용하였다.

GAN은 generator와 discriminator라는 서로 다른 두 개의 네트워크로 이루어져 있으며 이 두 네트워크를 적대적으로 학습시킨다. 그래서 실제 데이터와 비슷한 데이터를 생성한다. Generator의 목적은 실제 이미지에 가까운 가짜 이미지를 생성하는 것이고 discriminator의 목적은 이미지가 진짜인지 가짜인지를 결정하는 것이다. 그래서 이 두 가지 모델을 포함한 GAN의 궁극적인 목적은 실제 데이터의 분포에 가까운 데이터를 생성하는 것이다. 즉, discriminator가 진짜인지 가짜인지를 한 쪽으로 판단할 수 없을 때 최적의 솔루션으로 간주한다. 이러한 GAN을 사용하면실제 이미지와 매우 유사한 이미지를 생성할 수 있다.

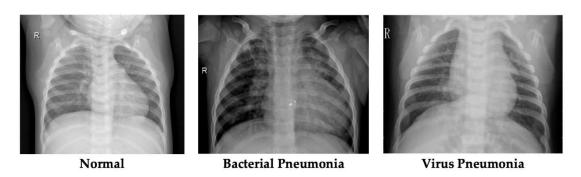
그러나 GAN은 noise vector만을 가지고 가짜 이미지를 생성하기 때문에 특정 이미지의 distortion에 적용하기는 어렵다. 따라서 관련 문헌 고찰을 통해 아이디어를 얻어 GAN을 이용하여 image distortion을 수행하는 모델을 구축할 예정이다. Generator로 생성한 이미지와 gaussian white noise 등으로 왜곡한 이미지를 각각 label 0과 1로 두고 원본 이미지를 이와 함께 discriminator에 input으로 사용한다. Discriminator는 실제 왜곡된 이미지와 generator가 생성한 이미지를 구별한다. 이를 통해 GAN을 이용하여 특정 이미지에 distortion을 적용할 수 있다.



<Fig 2. Distorting Images Using GAN>

연구의 두 번째 파트인 medical image classification model 구축을 위해 dataset을 수집하고 파악하였다. Chest X-ray dataset은 총 5,856개의 X-ray image로 구성되어 있으며 크게 training set과 test set으로 나누어져 있다. 또한, 각 set은 normal, pneumonia로 분류되어 있다. 그래서 폐렴유무를 판별하는 classification

model을 제작하였다.



<Fig 3. Chest X-ray Image Dataset>

<Table 2. Composition of Dataset>

Category	Training Set (No. of Images)	Test Set (No. of Images)
Normal	1,349	234
Pneumonia	3,883	390
Total	5,232	624

그러나 training set에서 normal과 pneumonia의 비율이 약 1:3으로 약간 편향되어 있음을 알 수 있다. 그래서 normal training dataset에 대하여 augmentation을 수행했다. 이는 image에 shifting, rotation 등을 적용하여 해당 데이터 셋에 속한 image의 수를 늘리는 기법이다. 아래의 표는 사용한 augmentation 기법이고 이를 적용하여 training set에서 normal과 pneumonia의 비율을 약 1:1로 맞추었다.

<Table 3. Augmentations>

Technique	Setting
Rotation	45
Vertical Shift	0.2
Horizontal Shift	0.15
Shear	16
Crop and Pad	0.25

이렇게 변경한 dataset 과 ImageNet dataset 으로 사전 훈련된 ResNet50 model 로 transfer learning을 수행했다. ImageNet dataset 은 1,000 개의 클래스로 구성되어 있고 총 백만 개가 넘는 데이터를 포함하는 large-scale dataset 이다. 그 중 약 120 만 개는 학습에 사용되고 5 만개는 검증에 사용된다. Transfer learning 이란

사전에 학습이 완료된 모델을 미세 조정(fine tuning)을 통해 학습시키는 방법이다. 특히 large-scale dataset 으로 훈련된 모델을 주로 사용한다. Transfer learning 은 학습 데이터의 수가 적을 때 유용하며 학습 속도도 빠르다. 또한, transfer learning을 사용하면 일반적인 모델보다 정확도도 높다는 장점이 있다. ResNet 은 잔차 학습(residual learning)을 이용한 CNN model 이며 그 중 층이 50 개인 모델을 ResNet 50 이라고 한다.

모델 훈련에 사용된 파라미터는 table 4와 같다. 그 중 batch size는 연산한번에 들어가는 데이터의 크기를 의미한다. Epochs는 훈련 데이터셋에 포함된모든 데이터들이 한 번씩 모델을 통과한 횟수로, 모든 학습 데이터셋을 학습하는 횟수를 의미한다. 딥러닝 모델이 학습할 때 loss function 을 정의하여 이의 최솟값을 찾는 것을 학습의 목표로 한다. 표에서 criterion은 loss function을 의미하며 cross entropy loss는 실제 값과 예측 값의 차이를 줄이기 위한 entropy를 의미한다.

Optimizer 는 이러한 최적화 알고리즘을 의미하며 그 중 stochastic gradient descent (SGD)는 무작위로 선택된 일부 데이터를 사용하여 최적화를 수행하는 알고리즘이다. 마지막으로 learning rate scheduler 는 학습을 진행하며 learning rate 를 조정하는 방법이다. 그 중 ExponentialLR 은 learning rate decay 가 exponential 함수를 따르는 방법이다.

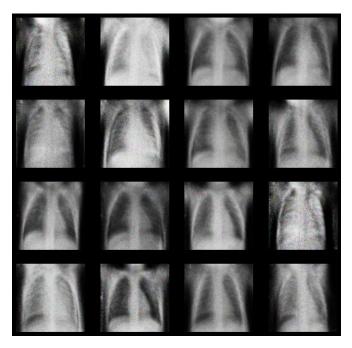
<Table 4. Parameters>

Parameters	Values
Batch Size	128
Epochs	25
Criterion	Cross Entropy Loss
Optimizer	Stochastic Gradient Descent (SGD)
Scheduler	ExponentialLR

4. 실험결과

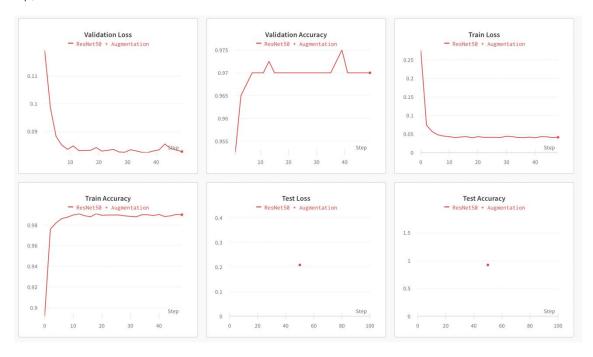
그림 3은 Chest X-ray dataset에 table 1 CNN based distortion 중 gaussian white noise를 적용한 결과이다. Traditional distortion과 다른 CNN based distortion 기법들은 남은 연구 기간 동안 적용할 예정이다. Distortion을 적용한

결과는 아래의 그림과 같다. 아래의 그림에서 실제 이미지와 유사한 이미지가 생성 되었음을 알 수 있다.



<Fig 4. Fake Images generated by GAN>

두 번째로 chest X-ray dataset과 다양한 classification model로 폐렴과 정상을 분류하는 모델을 제작하고자 하였다. 그 중 ImageNet dataset으로 사전 훈련된 ResNet50을 transfer learning을 수행하였을 때 92.3%로 가장 높은 정확도를 보였다.



<Fig 5. The Results of Classification of Pneumonia using ResNet50>

5. 고찰

현재 CNN based distortion의 baseline인 denoising autoencoder와 GAN을 chest X-ray dataset에 적용하였다. 따라서 남은 연구 기간 동안 traditional distortion과 관련 논문을 통해 얻은 아이디어로 GAN을 이용한 distortion을 구현하고 적용해야 한다. 또한, 현재 chest X-ray classification의 가장 높은 정확도는 97.3%이다. 이에 근접한 성능을 보이도록 모델을 수정하고 파라미터를 조정하는 등 다양한기법을 도입하여야 한다.

6. 결론

사람의 인지적 유사성 판단에 부합하는 medial image test set 구축을 위해 chest X-ray image dataset에 distortion을 적용하고 폐렴유무를 판별하는 classification model을 제작하였다. 남은 연구 기간 동안 아직 적용하지 못한 distortion 기법들을 구현하고 좀 더 성능이 좋은 classification model을 구축할 것이다. 그래서 reference와 distorted image를 제작한 후 기존 연구처럼 2AFC, JND test를 의사분께서 수행하여 test set 구축을 완료할 것이다. Test set 구축이 완료되면 classification model이 이미지의 유사성을 판단한 결과인 perceptual loss가 인간의 인지와 부응하는지 검증할 수 있다. 또한, 일반적인 분야에서 좋은 성능을 보였던 classification model이 medical image dataset도 그러한지 보이는 지를 증명할 수 있다.