안녕하세요.

AI 부트캠프 8기 지민진입니다.

제가 이번 section 프로젝트에서 선정한 주제는 GAN을 이용하여 X-ray 이미지 증강하는 것입니다.

왜 이 주제를 선정했는지 배경을 말씀드리면,

의료영상 데이터는 일반적으로 정상군 데이터 보다 환자군 데이터를 모으는 게 어렵습니다.

Brain CT나 MRI를 예를들어 보면, 뇌출혈, 뇌종양 등 생명에 치명적이거나 흔하게 발생하는 질환은 아니기 때문에 환자 데이터를 모으기 쉽지 않습니다.

때문에 데이터 불균형을 피하기 어렵고, 정확한 분석결과를 내는 것이 어렵습니다.

따라서 GAN을 사용해서 가짜 환자 영상을 만들어 데이터 불균형을 해결하고 모델을 학습한다면, 모델의 성능이 향상되는지, 가짜 영상이 실제 환자 영상과 비슷하게 만들어지는지 이번 프로젝트를 통해 확인하고자 합니다.

또한, CT나 MRI 영상은 이미지의 해상도가 크기 때문에, 학습시간을 고려하여 간단한 의료영상인 X-ray를 사용해 프로젝트를 진행하였습니다.

사용한 X-ray 데이터는 kaggle의 데이터셋을 사용하였습니다.

해당 데이터셋은 정상군, 코로나19환자, 폐렴환자로 구성되어 있는데, 프로젝트 기간을 고려해 모델학습을 단순화 하기 위해서 정상군과 코로나19환자 데이터만 사용하여 이진 분류를 진행했습니다,

GAN 모델 학습시에는 코로나19 환자 데이터만 사용하여 가짜 환자 이미지를 생성하였습니다.

프로젝트의 파이프라인은 다음과 같습니다.

먼저 코로나19 환자의 X-ray 이미지를 통해 GAN 모델을 학습시키고 가짜 환자 이미지를 얻은 후,

Fake 이미지를 추가한 경우와 추가하지 않은 경우 두 경우로 나누어서

Dataset을 형성한 다음, 동일한 CNN 모델로 분류 학습을 진행하였습니다.

분류 학습을 진행한 후, Dataset을 나누었을 때의 모델의 성능을 비교하였습니다.

다음으로는 GAN 모델 학습과정에 대해 말씀드리겠습니다.

앞서 말씀드렸듯 데이터는 코로나19 환자 데이터만 사용하였으며, 전처리 과정 중 이미지 크기를 축소시키고 pixel 값을 0과 1사이로 정규화 하였습니다.

정확한 Fake 이미지를 형성하기 위해 Deep convolution GAN 을 사용하였고, epoch를 1000으로 설정하여 학습하였습니다.

GAN 모델 구축시 다음과 같은 참고자료를 활용했습니다.

이렇게 학습된 GAN 모델을 통해 1000개의 코로나환자 이미지를 생성하여 기존 데이터 셋에 추가하였습니다.

추가적으로 학습 결과를 확인해보면, Epoch를 1000번 했을 때가 500번 했을 때보다 좀 더 x-ray 같은 이미지를 형성하는 걸 확인 할 수 있었습니다.

다음으론 이미지 분류시에 사용한 CNN 모델에 대해 말씀드리도록 하겠습니다.

CNN에 입력한 이미지도 동일하게 다음과 같은 이미지 크기로 고정하였고, 데이터 정규화를 진행하였습니다.

구축한 CNN 모델은 다음과 같은 형태로 구성하였고, early stop 을 사용해 최대한 정확하게 학습할 수 있도록 설정 했습니다.

더불어, 모델학습은 앞서 말씀드린 바와 같이 기존 Dataset만 이용해서 학습을 한 경우 그리고 기존 데이터셋에 Fake 환자 이미지를 추가했을 때로 나누어서 진행하였습니다.

다음으로는 데이터셋에 따른 모델 성능을 비교해보도록 하겠습니다.

Validation set과 Test set 모두에서 기존 Dataset을 사용한 경우보다 Fake 이미지를 더한 경우에 약간의 성능 향상이 확인되었습니다.

간단한 이진 분류 문제여서 validation accuracy가 매우 높게 나왔으나, loss 값에서 미약한 성능 향상을 확인할 수 있었고,

Test set에서 loss와 accuracy 모두 다음과 같은 약간의 성능향상을 확인할 수 있었습니다.

마지막으로 이번 프로젝트의 한계점과 추후 진행 방향입니다.

이번 프로젝트에서 진행한 분류 문제가 단순 이진 분류 문제여서 모델 성능이 전체적으로 매우 높게 나왔습니다. 그래서 데이터셋에 차이를 두었을 때의 성능 차이가 그리 크게 나타나지 않은 것 같습니다.

또한, 의료영상의 경우는 임상가들이 진단기준을 통해 환자군, 정상군으로 분류하는데, GAN으로 형성된 fake 이미지의 신뢰성을 보장하기 위해서는 결국 임상가들의 재확인이 필요합니다.

그럼에도 Fake 이미지를 더해주었을 때 성능향상을 보였기 때문에, 추후에 프로젝트는 코로나19 환자 데이터 외에 다른 환자 데이터를 통해 데이터를 증강하거나, x-ray 외에 내시경 이미지 나 CT 영상 등을 활용해 보려고 합니다.

또한 단순 이미지 분류 외에 segmentation이나 object detection에도 활용할 수 있을 것 같습니다.