CNN 기반 COVID-19 분류 모델 개발

Ji-Hyeon Ryu
류 지 현

Abstract

COVID-19로 인해 전 세계적으로 어려운 시기를 겪었다. COVID-19 검사의 소요 시간 등 여러 불편한 점을 해결하기 위해 CNN을 활용한 COVID-19 분류 모델을 개발했다. 코랩과 파이토치를 이용하여 구현하였고, epoch만을 조정하여 모델을 구현해보았다. 그러나 데이터 세트의 개수가 부족해 과적합 문제가 발생하고 정확도가 감소하는 경향을 보였다. 따라서 데이터 세트의 개수를 늘리거나 하이퍼파라미터를 조정하여 모델을 개선해야 할 것으로 판단된다.

1. 서 론

COVID-19의 영향으로 전 세계적으로 힘든 시기를 겪었으며, 이로 인해 COVID-19 검사에는 오랜 시간이 소요되는 등 여러 불편함이 있었다. 이러한 문제를 해결하기 위해 본 연구에서는 CNN을 활용한 COVID-19 분류 모델을 개발하고자 한다. 본 연구의 목적은 COVID-19 검사 과정에서 발생하는 시간 소요와 고통을 해소하기 위해, 의사들이 간편하고 신속하게 진단을 수행할 수 있는 COVID-19 분류 모델을 개발하는 것이다. 이를 통해 의사들은 환자의 폐 X-RAY 사진과 정상인의 폐 X-RAY 사진을 비교하여 COVID-19 유무를 판단할 수 있게 되며, 환자들은 거부감 없이 검사를 받을 수 있을 것이다.

2. 본 론

2.1 데이터 세트

COVID-19 확진 환자의 폐 X-RAY 사진과 정상인의 폐 X-RAY 사진을 데이터 세트로 사용하기 위해 Kaggle 플랫폼에서 데이터를 수집하였다. 수집된 데이터는 중복된 항목을 삭제하고, 필요 없는 특성 정보를 제거하여 사용하였다. 데이터 세트는 총 140개의 COVID-19 확진 환자 사진과 56개의 정상인 사진으로 구성되었다.

2.2 수행 환경

본 연구에서는 파이토치 (PyTorch) 라이브러리를 사용하여 CNN 모델을 구현하였으며, 코랩 (Colab) 환경에서 작업하였다. 파이토치의 버전은 2.0.1을 사용하였고, 코랩 환경에서는 CPU 가속을 활용하여 모델을 훈련하고 평가하였다.

2.3 알고리즘

COVID-19 분류 모델을 개발하기 위해 CNN 모델을 사용하였다. 모델은 2개의 레이어를 가지고 있으며, 각 레이어는 3x3 크기의 커널, 1의 스트라이드, 1의 패딩을 적용하였다. 활성화 함수로는 ReLU 함수를 사용하였다. 훈련 데이터는 130개를 사용하였고, 테스트 데이터는 66개를 사용하여 모델을 평가하였다. 모델의 하이퍼파라미터인 batch_size는 16으로 설정하였으며, epoch은 5, 10, 15, 20으로 높여가면서 실험하였다.

3. 결 과

그림 1, 2, 3, 4는 각각 epoch 별로 훈련 데이터와 테스트 데이터에 대한 정확도를 나타낸 그래프이다. 그림 1은 epoch 5의 결과를 보여주며, 그림 2는 epoch 10, 그림 3은 epoch 15, 그리고 그림 4는 epoch 20의 결과를 보여준다. 또한, 표 1은 각 epoch에서의 Best 정확도를 나타내고 있다.

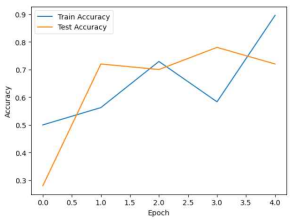


그림 1 epoch 5

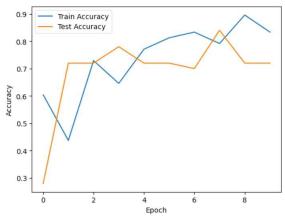
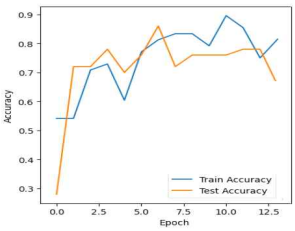


그림 2 epoch 10



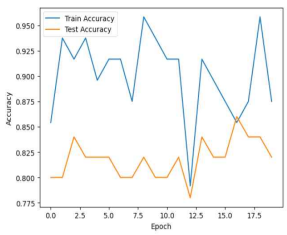


그림 3 epoch 15 **그림 4** epoch 20

	epoch 5	epoch 10	epoch 15	epoch 20
Train_ Best_Acc	0.899	0.931	0.895	0.887
Test_ Best_Acc	0.881	0.893	0.875	0.861

표 1

그래프와 표를 분석한 결과, 훈련 데이터와 테스트 데이터의 정확도가 epoch 증가에 따라 상승 및 감소함을 보였다. 특히, 그림 2와 표 1을 통해 epoch 5에서 가장 높은 정확도를 보였으며, 그 이후의 epoch에서는 정확도가 감소하는 것을 확인할 수 있다.

4. 고 찰

그림 1과 그림 2에서의 훈련 데이터와 테스트 데이터의 정확도가 우상향으로 좋게 나타났지만 그림 3과 그림 4에서는 그림 1과 2에 비해 비교적 정확도가 떨어지는 양상을 보였다. 훈련 데이터와 테스트 데이터의 개수가 적은 상황에서는 epoch을 늘릴수록 과적합이 발생하고 정확도가 감소하는 경향을 관찰하였다. 이를 해결하기 위해서는 데이터의 개수를 늘리거나 batch_size를 줄이거나 학습률을 조정하는 방법 등을 고려할 수 있다. 또한, 일반 드롭아웃 대신 알파 드롭아웃을 사용하여 과적합을 줄일 수도 있다. 알파 드롭아웃은 일반 드롭아웃보다 더 효과적인 정규화 기법으로 알려져 있으며, 이를 적용함으로써 모델의 일반화 성능을 향상 시킬 수 있다. 알파 드롭아웃이 과적합을 어떻게 해결하는지에 대한 추가적인 실험이 필요할 것 같다.

5. 결 론

최종적으로 CNN 모델을 활용하여 COVID-19에 확진된 환자의 폐를 구분하는 분류 모델을 구현하였다. 실험을 통해 데이터의 개수와는 상관없이 epoch만을 늘리면 훈련 데이터와 테스트 데이터의 정확도가 항상 증가할 것인지에 대한 의문을 가지게 되었다. 실험 결과, 데이터의 개수가 적은 경우에는 과도한 학습으로 인해 과적합이 발생하거나 훈련 데이터와 테스트 데이터의 정확도가 오히려 감소하는 것을 관찰하였습니다. 이를 통해 데이터의 개수와 epoch의 관계는 복잡하며, 최적의 epoch을 찾아내거나 적절한 하이퍼파라미터 조정을 통해 모델의 성능을 개선할 수 있을 것으로 판단된다.

참고문헌

[1] Kaggle, “COVID-19 X-ray Images”, https://www.kaggle.com/datasets/alifrahman/covid19-chest-xray-image-dataset, 2020