

CNN + XGBoost를 이용한 코로나 진단 방법

전민중*, 최혜진*, 박지웅*, 최하영*, 주한선**, 이욱*

*한양대학교 정보시스템학과

**가톨릭대학교 철학과

e-mail:jmj2316@hanyang.ac.kr, hyejinn@hanyang.ac.kr

jiwoongp94@hanyang.ac.kr, chy21@hanyang.ac.kr

jooking012@catholic.ac.kr, ooklee@hanyang.ac.kr

Diagnosing Covid-19 with Hybrid Approach : CNN + XGBoosting

Min-Jong Cheon*, Hye-Jin Choi*, Ji-Woong Park*, HaYoung Choi*,

Han-Seon Joo **, Ook Lee *

*Dept. of Information System, Hanyang University

**Dept. of Philosophy, Catholic University of Korea

요 약

본 논문에서는 CNN과 XGBoost를 활용하여 CT 데이터에 대해 코로나 판독 모델을 제안한다. 캐글에서 다운받은 CT 데이터들은 케라스의 ImageGenerator 함수를 통해서 증식되고 255로 나누어졌다 완전 연결 계층을 사용하여 최종 분류를 하는 기존 CNN 모델과 달리, 본 논문은 완전 연결 계층 대신에 XGBoost를 결합하는 것을 제안한다. 실험에서 XGBoost 뿐만 아니라 로지스틱 회귀, 결정 트리, SVM을 결합한 모델들과 비교하였으며, Imagenet으로 사전 훈련된 모델들인 VGG16, VGG19, MobileNet, Xception과도 비교하였다. 실험 결과, 본 논문에서 제안하는 모델이 95.37 %를 기록하며 가장 높은 정확도를 보였다. 본 연구를 통해 CNN과 XGBoost를 결합한 모델이 다른 딥러닝 모델들보다 짧은 시간 내에 높은 정확도를 보인다.

1. 서론

WHO에 따르면, 코로나바이러스감염증-19(이하 코로나19)는 새롭게 발견된 코로나바이러스에 의한 감염병이다. 코로나19는 주로 감염자의 입이나 코에서 배출된 비말을 통해 감염된다. 질병관리 본부의 자료에 따르면, 지금까지 대한민국에서 발생한 코로나19 확진자 수는 2021년 5월 4일 기준 124,269명, 사망자 수는 1,840명이다[1]. 전 세계에서 발생한 코로나19 확진자 수는 2021년 5월 4일 기준 154,323,338명, 사망자 수는 3,230,006명이다. 코로나19로 인한 팬데믹 상황 초기, 중국에서 AI 딥러닝으로 CT 자료를 판독하여 코로나19 폐렴 증상을 전문의 수준으로 선별하는 기술이 개발되었다 이후, 코로나19 폐렴 환자와 다른 질병의 55명 대조군 환자를 포함한 106명의 입원 환자의 46,096개의 익명 이미지를 수집해 딥러닝으로 분석한 결과 코로나19 선별 정확도는 전문의와 비슷한 95~99%, 방사선 전문의의 판독 시간은 65% 단축했다[2]. 앞으로도 이러한 딥러닝 기술로 엑스레이나 CT 자료를 판독해서 질병을 분석해서 기술이 많이 쓰일 것으로 기대된다.

2. 이론적 고찰

2.1 부스팅

부스팅 알고리즘은 앙상블 알고리즘의 하나로, 결정 트리 기반의 알고리즘이다. 또 다른 앙상블 알고리즘인 배깅 기법과의 차이점은, 배깅 기법이 평행하게 학습시키는 반면, 부스팅은 순차적으로 학습시킨다는 점이다. 첫 번째 약한 학습기가 분류를 진행하며, 잘못 분류된 데이터에 가중치를 두고 두 번째 학습기가 다시 분류를 하게 된다. 이러한 방식으로 잘못 분류된 데이터에 가중치를 두면서 순차적으로 학습을 하여 결론적으로 강한 학습기를 구축한다[3].

2.2 CNN

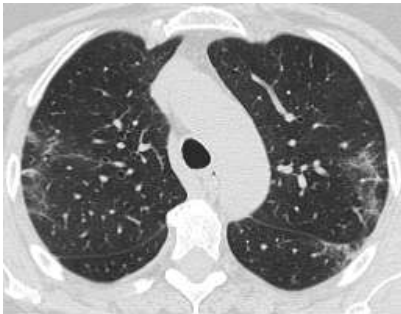
Convolutional Neural Network(CNN)은 이미지 데이터를 다룰 때 주로 쓰이는 대표적인 딥러닝 알고리즘으로, 풀링 계층(Pooling Layer), 완전 연결 계층(Fully Connected Layer), 합성곱 계층

(Convolutional Layer)로 구성되어 있다. CNN이 이미지를 입력 데이터로 받게 되면, 먼저 합성곱 계층은 필터(Filter)를 통해 이미지의 특징맵(Feature Map)을 추출하면, 풀링 계층이 특징맵의 평균값이나 최대값을 계산한다. 마지막으로 완전 연결 계층에서 실제로 분류하게 된다[4].

3. 제안 방법

3.1 CT 데이터

코로나 여부를 판독하기 위한 데이터는 캐글(Kaggle)에서 다운받았다[5]. 이 데이터는 코로나 환자 1252명과 정상인 1230명의 CT 데이터로 이루어졌다. 케라스의 ImageGenerator 함수를 통해서, 각도, 밝기, 가로, 세로 등을 조금씩 조정해 증식시키고, 전부 255로 나누었다.



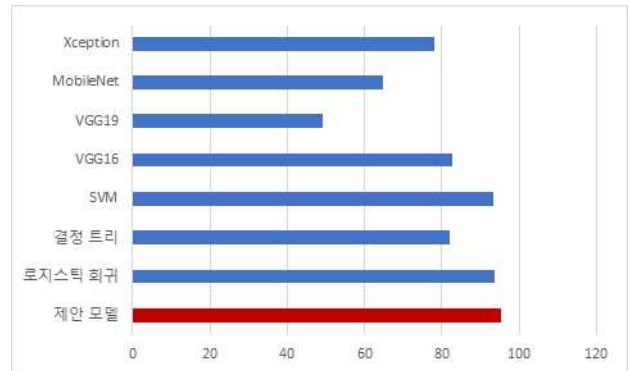
[그림 544] 본 논문에서 사용한 CT 데이터

3.2 CNN + XGBoost

본 연구에서 제안하는 딥러닝과 머신러닝 모델을 이용한 코로나 진단 모델은 그림 [1]과 같다. 기존의 CNN 알고리즘의 경우 특징 추출 및 풀링 계층을 거친 후 완전 연결 계층에서 분류를 하지만, 본 연구에서는 완전 연결 계층 대신에 머신러닝 기법 중 하나인 XGBoost를 연결하는 방법을 제안한다. 실험에서는 XGBoost 뿐만 아니라 또 다른 머신러닝 분류 기법들인 로지스틱 회귀, 결정 트리, Support Vector Machine(SVM)을 활용한다. 또한, Imagenet 데이터로 사전 훈련된(pre-trained) 모델들인 VGG16, VGG19, MobileNet, Xception과 정확도를 비교한다.

4. 연구 결과

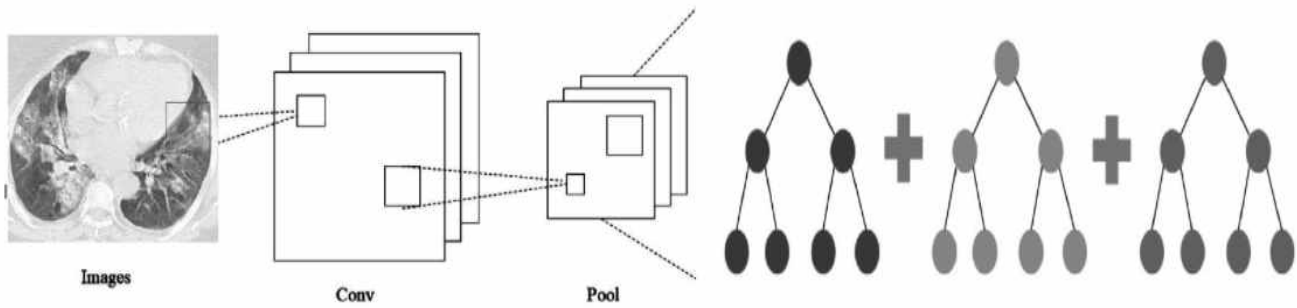
실험 결과, 제안하는 모델은 95.37 %의 정확도를 기록하였다. 로지스틱 회귀, 결정 트리, SVM을 결합하였을 때는 각각 93.76 %, 82.09 %, 93.36 %가 나왔다. VGG16, VGG19, MobileNet, Xception은 각각 82.8 %, 49 %, 64.79 %, 78.06%로 도출되었다.



[그림 545] 모델 간의 정확도 비교 결과

5. 결론

본 논문은 CNN과 XGBoost를 결합하여 CT데이터를 통한 코로나 진단 모델을 제안한다. 제안하는 모델의 경우 정확도가 가장 높았으며 사전 훈련된 모델들의 정확도가 상대적으로 낮았는데, 훈련 및 검증 데이터셋에서는 정확도가 높았던 것을 고려하면, 오퍼피팅이 발생한 것으로 예상된다. 본 연구는 CNN과 XGBoost 모델을 결합함으로써 다른 딥러닝 모델들보다 훈련 시간이 짧으면서 정확도가 높다는 장점을 가진다.



[그림 546] 본 연구에서 제안하는 코로나 진단 모델

참고문헌

- [1] World Health Organization. (n.d.). Coronavirus. World Health Organization.
https://www.who.int/health-topics/coronavirus#tab=tab_1
- [2] AI 딥러닝, 코로나19 CT판독...정확도 95~99% 전문의 수준. scimonitors.com. (n.d.).
<http://scimonitors.com/ai-%EB%94%A5%EB%9F%AC%EB%8B%9D-%EC%BD%94%EB%A1%9C%EB%82%9819-ct-%ED%8C%90%EB%8F%85-%EC%84%EB%AC%B8%EC%9D%98-%EC%88%98%EC%A4%80-%EC%A0%95%ED%99%95%EB%8F%84-9599/>
- [3] Schapire, R. E. (1999, July). A brief introduction to boosting. In *Ijcai* (Vol. 99, pp. 1401-1406).
- [4] Sainath, T. N., Mohamed, A. R., Kingsbury, B., & Ramabhadran, B. (2013, May). Deep convolutional neural networks for LVCSR. In *2013 IEEE international conference on acoustics, speech and signal processing* (pp. 8614-8618). IEEE.
- [5] Angelov, P., & Soares, E.. (2020). EXPLAINABLE-BY-DESIGN APPROACH FOR COVID-19 CLASSIFICATION VIA CT-SCAN.
<https://doi.org/10.1101/2020.04.24.20078584>