CNN 기반의 한약재 이미지 비교학습 알고리즘

*김효중, 김기연, 이원형, 서주완, 인치호
세명대학교 컴퓨터학부
e-mail : hyoj0942@gmail.com, rlarldus6789@naver.com, whl0409@naver.com, joowop@naver.com, ich410@gmail.com

CNN-based Herbal Medicine Image Comparison Learning Algorithm

*Hyo-Jung Kim, Ki-Hyun Kim, Won-Hyung Lee, Ju-Wan Seo, Chi-Ho Lin
School of Computer Science
Semyung University

Abstract

This study is about automatic classification and information delivery of herbal medicines and in more detail, we use big data analysis techniques and deep learning techniques to analyze information about herbal medicines, providing systems for accurate information delivery to patients using herbal medicines.

In particular, this invention concerns how to analyze images (algorithms) into two piles to compare image classification accuracy differences according to differences in data collection methods and to produce more accurate results through image data classification evaluation.

Furthermore, images can be learned and classified using CNN techniques on VGGNet using data preprocessing and aggregation methods based on each collected image to output accuracy and loss functions each time the validation process proceeds, and image data can be trained, validated, and tested in more detail.

I. 서론

국내 한약재 및 관련 제품의 소비가 꾸준히 증가하고 있다. 또한, 한방제품의 품질과 안전성, 효능이 높아지고 있지만, 한약재 수입량이 증가함에 따라 한약재 품질의 보증과 안정성에 대한 논란이 지속되고 있고 이에 따라 한약재에 대한 품질 보증의 필요성이 대두되고 있다.[1]

한약재는 생산, 가공, 포장, 유통, 소비단계에 이르기까지 생산자에서 수요자에게 전달되는데 소요되는 시간이 길고 과정이 복잡하며 농산물임과 동시에 의약품 원료라는 특성을 동시에 보유하고 있다. 따라서 한약재의 원산지, 가공 및 유통과정에서 생길 수 있는 오염이나 유효성분 소실 등의 문제를 해결하고 수요자에게 전달되기까지 안전성 및 품질의 보증을 위한 적절한 시스템이 요구된다.[2]

II. 한약재 이미지의 자료수집 방법

1. 자료수집

본 연구의 전반적인 흐름은 이미지 그림 1. 과 같다. 분석 대상 한약재는 한방의료기관을 대상으로 가장 많이 사용하는 한약재를 5가지로 색출하였으며, 해당 한약재는 당귀, 감초, 작약, 황기, 백출 순으로 선정하였다. [3]

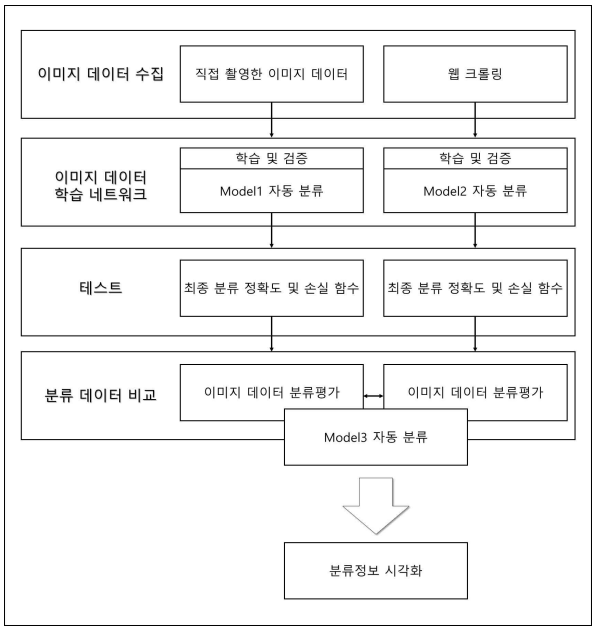

그림 1. 전반적인 연구 흐름도

정확한 이미지 감별을 위해선 많은 이미지가 필요하게 된다. 웹 서비스의 검색 엔진에서 한약재를 검색하여 이미지들을 하나씩 내려받는 방식을 취할 수도 있지만, 다양성과 유연성을 높이고 시간 및 인력 낭비제거를 위해서는 보다 개선된 분류 방법이 필요하다. 따라서 데이터 수집군을 직접 촬영한 이미지, 웹 크롤링을 이용한 이미지, 두 분류로 나누어 데이터를 수집 및 분석한다.

1.1. 한약재 전문가가 촬영한 이미지 분석 및 학습방법(Model 1)

첫 번째 이미지 분석 및 학습방법은 각 한약재 특징을 강조하여 시계열에 따라 한약재 전문가가 촬영하여 수집한다.

수집하는 데이터는 위에서 언급한 5종류이며, 각 종류당 200개로 1,000개의 이미지 데이터를 활용한다.

1.2. 웹 크롤링을 이용한 이미지 분석 및 학습방법(Model 2)

두 번째 이미지 분석 및 학습방법은 웹 크롤링을 이용하여 다양한 웹 사이트에서 대상 한약재 이미지를 일괄 수집하는 방법으로 국내의 대표적인 포털 망인 '구글', '네이버', '다음'에서 추출한 이미지를 사용한다. 웹 크롤링을 이용한 자료수집을 위해 Python 3.8.6을 이용하였고, 개발환경은 Jupyter Notebook을 이용하였다. 이미지 분류에 딥러닝 기법을 사용할 경우 학습 이미지 데이터가 많을수록 이미지 자동분류 모델 성능 향상에 유리하지만, 산림전문가가 직접 촬영한 이미지 데이터 수가 한약재별 200개로 한정되었기 때문에 이미지 데이터의 양적 차이에 따른 영향을 최소화하기 위해 웹 사이트에서 추출된 이미지 중 종류별 200개씩 1,000개를 선정하여 학습에 사용하였으며 알고리즘은 그림 2. 와 같다.

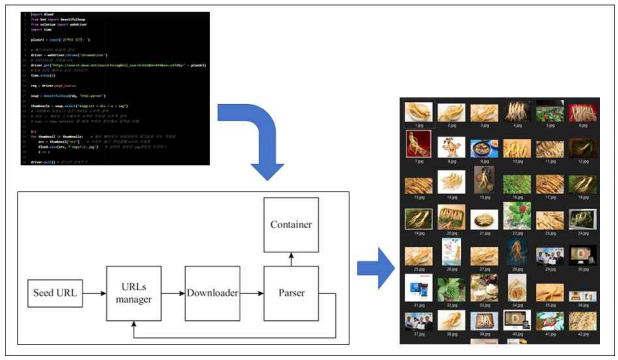

그림 2. 웹 크롤링 알고리즘

Ⅲ. 수집된 한약재 이미지 데이터에 대한 학습 및 분류평가

전반적 모델의 학습 알고리즘 그래프는 그림 3. 과 같다. 두 분류의 이미지 비교학습을 위한 모델구축에서는 CNN(Convolution Neural Network) 기반의 알고리즘을 사용하여 이미지를 학습 및 비교하였다.[4] 첫 번째 모델인 Model 1은 한약재 전문가에 의해 촬영된 이미지를 기반으로 학습 및 비교하는 모델을 개발하여 이미지 분류 정확도를 산출하고, 두 번째 모델인 Model 2는 웹 크롤링 기법을 이용한 이미지를 기반으로 구축된 모델을 이용하였다.

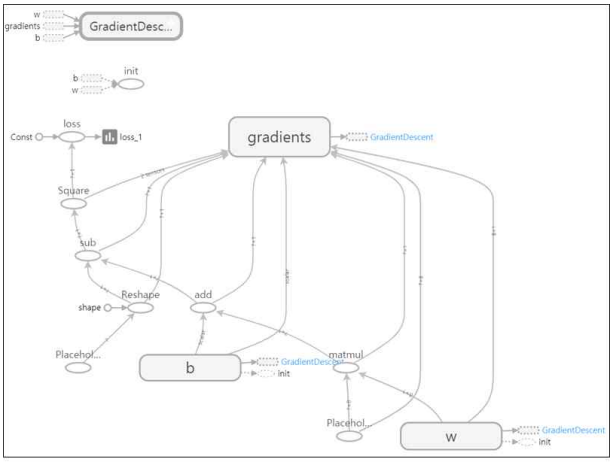

그림 3. 이미지 데이터 학습 알고리즘 그래프

마지막으로 두 모델의 정확도를 비교하여 모델 간 정확도 분류 결괏값을 비교하여 더욱 정확한 학습 모델 Model 3을 구축하는 것을 목표로 선정하였다.

1. 각 모델의 한약재 이미지 학습 구현

본 연구에서는 텐서플로우(TensorFlow)를 이용하여 데이터 자동분류를 사용하여 데이터를 자동분류 및 분석하였고, 텐서보드(TensorBoard)를 통해 시각화하여 정확도와 손실률을 나타내었다.

모델을 훈련하고 검증 과정을 진행할 때마다 손실함수와 정확도를 시각화하였다. 각 200개, 총 1000개의 이미지를 모델별로 학습하였으며, 각 단계에서 정확도(Acurracy), 손실함수(Loss Function)를 정의하였고, 한약재 이미지 자동분류 모델은 테스트 데이터를 정확히 분류, 확인하는 테스트를 하였으며 각 모델의 이미지 학습 네트워크는 그림 4. 와 같다.

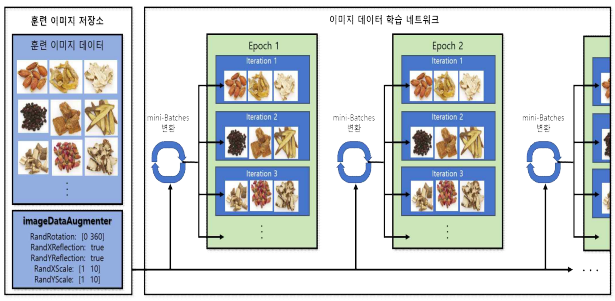

그림 4. 이미지 학습 모델

2. 각 학습 모델의 정확도 비교

대표적인 Model 3의 학습률 정확도를 위해 비교학습 하였으며 정확도와 손실률 그래프는 그림 5. 와 같다.

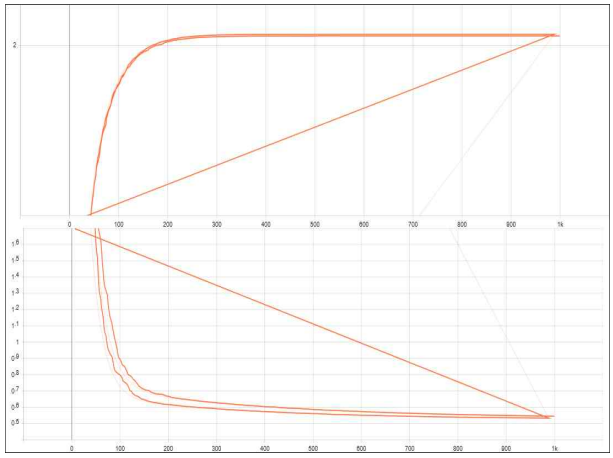

그림 5. Model 3에 대한 정확도 및 손실률 그래프

본 논문의 테스트 정확도는 전문가가 직접 촬영한 한약재 이미지 데이터를 이용한 모델(Model 1) 95.2%, 웹 크롤링 데이터를 이용한 모델(Model 2) 62.3%, 두 학습 모델을 비교 학습한 모델(Model 3) 96.7%로 Model 3의 테스트 정확도가 가장 높게 나타났다. 정확도와 손실률은 텐서보드의 그래프 형태로 표현하면 진동 폭이 증가하여 데이터 흐름을 판단하기 어려워 텐서보드 기능인 스무딩(Smoothing) 기능을 사용하여 처리 후 결과를 나타내었다. 대표적으로 Model 1과 Model 2의 데이터를 비교 학습한 Model 3의 학습 정확도는 학습 직후 급격히 상승하여 1에 수렴했고, 1000회를 학습하여 0.920의 정확도를 얻었으나 이후 0.89~0.93 사이에서 진동했다. 또한, Model 3의 손실률은 학습 초기 0.43였으며, 1000회의 검증 단계에서 0.152의 최솟값을 나타냈으나 이후 0.27~0.38에서 진동하였다.

Ⅳ. 결론 및 향후 연구 방향

본 논문은 직접 전문가에게 받은 각 한약재의 특징을 가진 사진과 인터넷에 게시되어있는 사진을 데이터로 사용하여 웹 크롤링 기법을 적용하여 데이터를 수집한다. 수집된 이미지를 각 수집 방법에 따른 한약재 이미지 자동분류 모델로 성능을 비교하였다. 그 결과 웹 크롤링으로 수집한 이미지 데이터를 사용한 Model 2는 나머지 모델에 비해 정확도가 현저하게 낮고 데이터 양이 적어 발생하는 과적합이 의심되었다. 또한, 손실함수가 증가하는 흐름을 보였다. 그러므로 웹사이트에서 수집한 데이터를 정제하여 정확한 사진만 감별하는 모델 구축 방안을 마련할 필요가 있다. 반면 전문가가 촬영한 Model 1은 테스트 정확도가 매우 높았으며, 손실함수 또한 낮았다. 마지막으로 두 학습 모델을

비교학습한 Model 3는 가장 높은 정확도와 가장 낮은 손실률로 최고의 성능을 보였고, 이에 따라 각 모델의 검증 손실함수를 최소화할 방안에 관한 연구가 추가로 필요하다.

본 논문에서는 최근에 증가한 한약재의 수요에 비해 정립되어 있지 않은 한약재의 표준안을 수요자들에게 제시함으로써 신뢰성 있는 데이터 제공 방법을 제공하며 더 나은 시장 효과를 가져오는 것을 목적으로 두었다. 또한, 향후 연구 방향은 OpenCV를 이용하여 시각적 데이터 제공을 위한 플랫폼을 제공하는 것을 목표로 하며, 다른 분야의 이미지 데이터 학습에 대해서도 모델을 나누어 비교학습 하였을 때 비슷한 결과물이 나오는지 추가적인 연구의 필요성이 요구된다.

※ 본 논문은 2021년 교육부 대학혁신지원사업에 의해 지원되었음

참고문헌

[1] 남동우, and 양웅모. "우수 한약 제조 및 품질 관리 기준 (hGMP) 시행을 위한 한약 제약 업소 현황 조사 연구." 대한한의학회지 32.4 (2011): 111-127.

[2] 이정찬, et al. "한약 및 한약재 관련 정보제공 현황과 개선방안: 조제내역서 발급 및 원산지 표시를 중심으로." 대한의사협회 의료정책연구소 연구보고서 (2018): 1-84.

[3] 보건복지부 보도자료(2018. 2. 28), 「2017년 한방의료이용 및 한약소비 실태조사 결과 발표」

[4] Lee, Sue Han, et al. "Deep-plant: Plant identification with convolutional neural networks." 2015 IEEE international conference on image processing (ICIP). IEEE, 2015.

[5] Demirović, Damir, Emir Skejić, and Amira Šerifović-Trbalić. "Performance of some image processing algorithms in tensorflow." 2018 25th International Conference on Systems, Signals and Image Processing (IWSSIP). IEEE, 2018.

[6] Abu, Mohd Azlan, et al. "A study on Image Classification based on Deep Learning and Tensorflow." International Journal of Engineering Research and Technology 12.4 (2019): 563-569.