GAN을 이용한 CNN 기반 토마토 병해충 분류 성능 향상

김현지, 박지연, 김경백
전남대학교 전자컴퓨터공학부
khyeonj1025@gmail.com, wldus8677@naver.com, kyungbaekkim@jnu.ac.kr

Performance enhancement of CNN based tomato pest classification with GAN

Hyeonji Kim, Jiyeon Park, Kyungbaek Kim
Department of Electronics and Computer Engineering,
Chonnam National University

요 약

병해충은 인간의 삶에 치명적인 요소로 작용할 수 있다. 기존 농업에서의 병해충 판단 방법은 전문 인력의 시간적, 금전적 노력이 요구되며, 인간의 의사가 적용되므로 판단 결과가 불확실하다. 이를 해결하기 위해 토마토를 대상으로 CNN 기반의 병해충 분류 모델을 연구한다. 한편 비주기적으로 발생하는 병해충의 경우 해당 데이터를 수집하는 것에 어려움이 있으며, 이는 데이터의 부족과 불균형을 야기하여 모델의 성능에 악영향을 미치게 된다. 이를 완화시키기 위해 Data Augmentation의 관점에서 GAN을 사용하여 데이터셋을 확장한다. 결과적으로 본 논문에서는 GAN의 데이터 확장에 따른 CNN 기반 토마토 병충해 분류 모델의 성능 향상을 확인한다.

1. 서 론

병해충은 농부들에게 경제적으로 심각한 피해를 야기할 뿐만 아니라, 식량 생산에 영향을 끼쳐 인간의 삶에 치명적인 요소로 작용할 수 있다.[1] 기존에는 병해충을 판단하기 위해 농업 분야의 전문가들이 직접 현장을 방문하거나, 국가농작물병해충관리시스템 NCPMS 병해충 상담[2]을 통해 작물을 검사하였다. 그러나 이는 상당한 시간과 비용이 소요되고, 인간의 판단이기 때문에 판단 결과가 확실하지 않다. 따라서 육안으로 확인하기 어려운 병해충의 판단을 위해 CNN 기술을 적용한다.

한편 비주기적으로 발생하는 병충해같은 경우는 데이터 수집이 어렵다. 딥 러닝의 성능과 관련된 연구는 아직까지도 진행 중이다. 새로운 함수의 도입, 파라미터의 조절 등 여러 방법으로 모델의 성능을 개선하려 하지만 근본적으로 중요한 것은 학습 데이터이다. 모델을 잘 학습시키기 위해서는 label별로 균형적인 충분한 양의 데이터가 필요하다.[3]

이에 다양한 분야에서 기존의 데이터를 변형시켜 데이터를 확장시키는 Data Augmentation 방법이 연구되고 있다. 실제로 데이터 부족이 심각한 의료분야에서 합성 이미지를 통한 Data Augmentation이 좋은 성과를 보이고 있다.[4] 그 중 높은 수준의 다양한 합성 이미지를 만들기 위해 Generative Adversarial Networks(GANs)을 통한 Data Augmentation 방법이 제안되고 있다.

본 논문에서는 GAN을 활용한 Data Augmentation를 통해 데이터 불균형을 해소하고, 그에 따른 병충해 분류 모델의 성능 향상을 분석해보고자 한다.

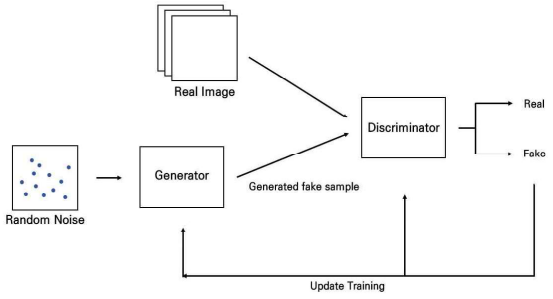
2. 관련 연구

2.1 CNN

최근 딥 러닝 기술이 발전함에 따라 다양한 알고리즘이 개발되고 있다. 합성곱 신경망(Convolutional neural network, CNN)[5]은 여러 개의 Convolutional layer와 일반적인 신경망 layer로 이루어져 있으며, weight와 pooling layer를 추가적으로 활용한다. 이러한 구조 덕에 영상, 음성, 이미지 등 다양한 분야에서 높은 수준의 인식 결과를 보인다. 또한. 특징을 수동으로 추출할 필요가 없다. 그로 인해 다양한 분류기법 모델 구현에 사용되고 있다. 유해 네트워크 트래픽 탐지[6], 과일 분류[7]가 그 예시다.

2.2 Generative Adversarial Networks

Ian J. Goodfellow 등은 합성 이미지를 생성해낼 수 있는 생성 모델 GAN을 제안하였다.[8] GAN에서 Generator는 랜덤한 노이즈 분포에서 샘플을 생성하며, Discriminator는 샘플이 실제 이미지인지 Generator의 결과인지 구별한다. 이 대립 과정을 거치며 각자의 성능이 높아지게 되면 결국 실제 이미지와 유사한 샘플을 얻을 수 있다. GAN의 전반적인 구조는 다음과 같다.



(그림 1) (GAN general structure)

2.3 Data Augmentation with GAN

Hasseb Nazki 등은 제한된 식물 질병 데이터에서 딥러닝의 data augmentation과 데이터 균형을 만족시키기 위해 GAN을 통한 합성 이미지 생성 방법을 제시하였다.[9] 또한 Maayan Frid-Adar 등은 GAN 기반의 합성 의료 이미지 augmentation을 통해 간 병변을 분류하는 CNN 모델에서 실제로 성능이 향상됨을 확인하였다.[4]

이러한 연구들에서 GAN으로 생성한 샘플이 실제 이미지와 굉장히 유사하다면 기존의 데이터셋을 확장할 수 있으며, 그에 따라 데이터셋의 균형을 이룰 수 있다는 것을 알 수 있었다. 이는 곧 병해충 분류 모델의 성능 향상과 직결될 것으로 예상한다. 그러므로 본 논문에서는 토마토 병해충이라는 세부적인 분야에서도 GAN 샘플이 사용될 수 있음을 확인하고, 데이터 부족이 해결된 CNN 분류 모델의 성능을 실험하고자 한다.

3. CNN 기반 토마토 병충해 분류 모델 구현

3.1 토마토 병해충 데이터셋

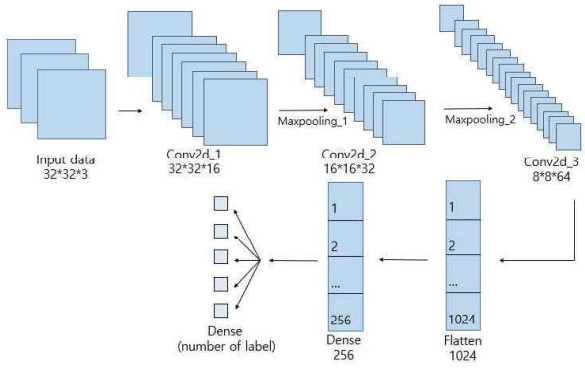
본 논문에서는 Kaggle에서 제공하는 토마토 병해충 이미지[10]를 다양하게 응용하여 병충해 분류 모델을 구현한다. 기존 데이터셋은 Train set과 Test set으로 나뉘어져 있다. label은 총 10개이며, Train set은 58,092개의 이미지, Test set은 14,512개의 이미지로 구성되어 있다. 전체 데이터의 크기는 256*256 픽셀로 모두 동일하며, JPG 파일 형식을 가진다. 그 중 4개의 병충해 label과 1개의 건강한 잎 label만을 사용하여 분류 모델을 구현할 예정이다. 모델 구현에 쓰인 데이터의 label 명과 데이터 개수는 표 1과 같다.

(표 1) 각 label당 데이터 개수 (단위: 개)

label	Train set	Test set
bacteria	6,808	1,700
lateblight	5,592	1,432
targetspot	4,496	1,120
yellowleafcurl	17,144	1,538
healthy	5,088	1,272

3.2 토마토 병해충 분류 CNN 모델 구성

본 논문은 데이터의 개수에 중점을 두었으므로, 해당 조건을 제외한 모든 조건을 동일시하기 위해 한 가지 모델만을 사용한다. 모델의 입력 값은 앞서 언급한 256*256 픽셀 이미지 파일이며, 원활한 학습을 위해 32*32 픽셀로 크기를 조정한 후 CNN 모델을 구성하였다. 그림 2는 모델의 간단한 architecture를 나타낸다.



(그림 2) (CNN architecture)

구성한 모델은 총 3개의 Convolutional layer로 이루어져 있고, 각 layer 사이사이마다 Max pooling을 수행한다. Convolutional layer의 activation function에는 ReLU[11] 함수를 사용한다. 모델의 compile에는 Multi class 분류이므로 categorical crossentropy loss function를 사용하였고, optimizer는 Adam을 사용한다. train 과정에서의 batch size는 100, epoch는 500으로 callback에 early stopping 객체를 넣어 과적합을 방지하였다. 출력은 입력과 같은 'healthy', 'bacteria', 'lateblight', 'targetspot', 그리고 'yellowleafcurl' 총 5개이다.

4. 토마토 병해충 이미지를 생성하는 GAN

일반적인 GAN에서는 오랫동안 학습을 진행해도 모델이 수렴하지 않는 등의 불안정한 학습 문제가 존재한다. 그렇기에 본 논문에서는 기존 GAN에 Convolution을 도입하여 문제의 영향을 덜 받는 Deep Convolutional Generative Adversial Networks (DCGAN)[12]를 사용한다.

GAN 모델은 토마토 병충해 label의 수와 동일한 4개를 만들었으며, 모델의 입력으로 표 1의 각 label에 해당하는 Trainset 전체를 사용하되 수월한 학습을 위해 256*256 픽셀의 사이즈를 32*32 픽셀로 조정하였다. batch size는 100으로 통일했으나 yellowleafcurl의 경우 train set이 다른 label에 비해 많으므로 epoch 400, 다른 모델은 epoch 1000으로 학습을 진행했다. 다음은 각 모델의 iteration을 계산한 결과이다.

(표 2) 각 label당 iteration 횟수

label	iteration
bacteria	68,080
lateblight	55,920
targetspot	44,960
yellowleafcurl	68,576

GAN 모델의 학습이 끝나면 epoch 20까지의 학습을 다시 진행하면서 epoch 1마다 샘플을 1000장씩 생성하도록 하여 각 label 당 총 32*32 픽셀 사이즈의 GAN 샘플 20,000장을 생성한다.

5. 실험 및 검증

5.1 CNN 분류 모델 성능

먼저 데이터의 개수가 많을수록 성능이 나아지는 것을 확인하려고 한다. label당 50개, 100개, 300개, 1000개씩의 데이터를 학습시킨 모델을 제작하였다. 이 때 학습 조건은 데이터의 개수를 제외하면 모두 같다.

(그림 3) (각 label 별 accuracy)

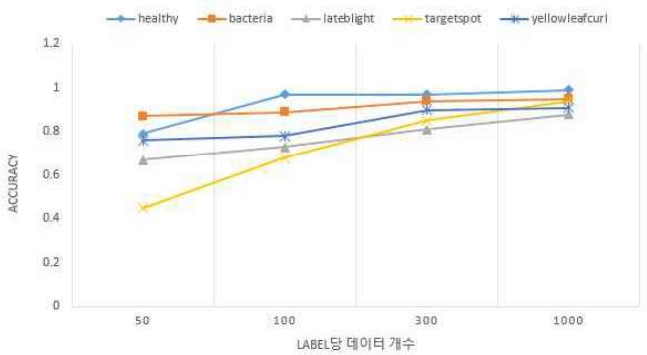


그림 3은 각 label 별 accuracy의 변화 양상을 그래프로 나타낸 것이다. 데이터 개수가 50개에서 1000개로 늘어날수록 accuracy가 지속적으로 증가하며, 1에 수렴한다. 이는 데이터의 개수가 많을수록 좋은 성능을 보인다는 가설을 증명할 수 있다.

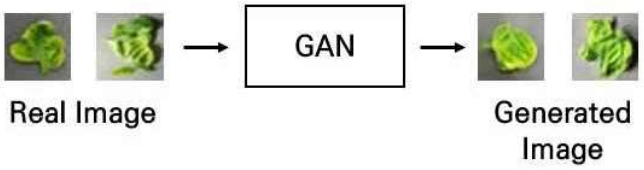
또한 각 병해충 label과 healthy label을 사용하여 GAN을 검증하기 위한 모델을 제작하였다. 그림 n의 모델을 바탕으로 각 병해충의 Train set과 healthy label의 Train set을 사용해 이진 분류 모델을 만들고, 이를 Test set으로 검증한다. 표 2는 4가지 병해충에 대한 healty/unhealthy 판단 모델의 label 별 accuracy와 전체 accuracy를 정리한 것이다.

(표 3) 병해충 감염 여부 판단 모델의 accuracy

Model	Healthy	Unhealthy	Overall
bacteria	0.9976	0.9916	0.9944
lateblight	1.0	0.9874	0.9933
targetspot	1.0	0.9714	0.9866
yellowleafcurl	1.0	1.0	1.0

모든 모델이 전체적으로 높은 accuracy를 가진 것으로 보아 GAN으로 생성한 데이터의 검증에 적합하다.

5.2 GAN을 통해 생성한 토마토 병해충 이미지



(그림 4) (좌: 실제 yellowleafcurl 이미지,
우: GAN을 통해 생성한 yellowleafcurl 이미지)

그림 4와 같은 과정을 통해 GAN 샘플을 생성했다. 이를 실제 CNN 모델에 사용할 수 있을 지 판단하기 위해 표3에서 언급했던 이진 분류 모델을 사용하여 GAN 샘플을 분류했다.

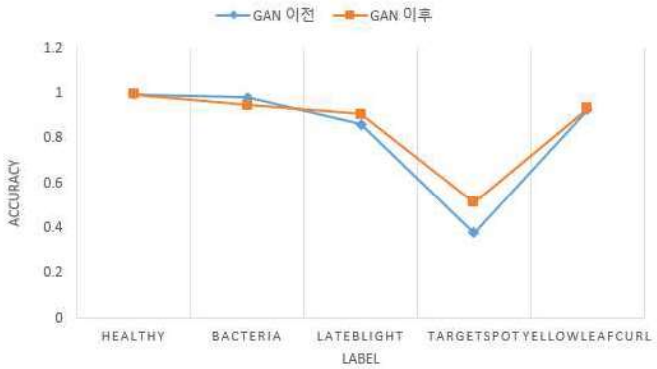
(표4) 병해충 이진 분류 모델의 GAN 샘플 판단 결과

label	accuracy
bacteria	0.9971
lateblight	0.9626
targetspot	0.9545
yellowleafcurl	0.9984

판단 결과는 각 label에 대해 정확히 분류한 결과의 수 / GAN 샘플 20,000장으로 계산하였다. 그 결과 정확도가 95% 이상으로 충분히 높으므로 GAN으로 생성한 샘플을 데이터 확장에 사용할 수 있다고 판단했다.

5.3 Data Augmenation을 통한 CNN 분류 모델 성능 분석

GAN으로 생성한 이미지를 통해 데이터 불균형을 해결할 수 있는 지에 대한 실험을 진행하였다. Train set의 개수가 가장 적은 targetspot의 데이터를 50개, healthy, bacteria, lateblight, targetspot, yellowleafcurl의 데이터를 5000개씩 준비하여 학습하였다. 그 결과 0.85~0.99의 accuracy를 보인 타 label과는 달리 데이터가 적었던 targetspot의 accuracy가 0.3786으로 매우 저조하게 나타났다. GAN으로 생성한 targetspot 이미지 4550장을 targetspot label에 추가하였고, 다시 학습을 진행한 결과 0.5152까지 accuracy가 상승하였다.



(그림 5) (GAN 사용 여부에 따른 accuracy 변화)

그림 5는 GAN 사용 전후의 accuracy의 변화를 그래프로 나타낸 것이다. healthy, bacteria, lateblight, yellowleafcurl은 뚜렷한 변화가 나타나지 않았으나, GAN으로 데이터를 보완한 targetspot의 경우 미약하게나마 accuracy가 상승했다.

6. 결론

인간이 판단하는 병해충은 시간 및 비용이 상당히 소모되며, 주관적인 판단에 의해 결과가 잘못될 가능성이 있다. 그에 따라 본 논문에서는 CNN 기반의 토마토 병해충 분류 모델을 구현하였다. 이 분류 모델에서 데이터 개수가 많을수록 좋은 성능을 보인다는 사실을 확인하였다.

또한 병해충 이미지는 데이터가 부족한 경우가 많기 때문에 GAN을 사용하여 데이터를 확장하고, CNN 모델의 성능 향상 실험을 진행하였다. 결과적으로 GAN을 통해 데이터 불균형을 완화시킨 CNN 분류 모델의 accuracy가 상승했다. 이는 GAN을 통해 CNN의 모델의 성능 향상을 이끌 수 있다는 가능성으로 볼 수 있다.

향후에는 GAN 모델 분석 및 CNN과의 결합을 통해 보다 높은 성능 향상을 추구하고자 한다. 이번 연구에서는 토마토의 병해충 분류에 대해서만 실험하였는데, 이에 국한하지 않고 사과, 포도 등 다양한 농작물의 병해충 분류 모델로 확장시키는 것을 계획하고 있다.

Acknowledgement

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 대학ICT연구센터지원사업의 연구결과로 수행되었음 (IITP-2019-2016-0-00314). 본 연구는 한국정보화진흥원의 출연금으로 수행한 NET챌린지 캠프 2019 챌린지리그(학생팀)의 연구망 활용 연구과제 결과임. 이 논문은 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(NRF-2017R1A2B4012559).