JDCS
디지털콘텐츠학회논문지
Journal of Digital Contents Society
Vol. 21, No. 5, pp. 959-967, May 2020

CNN 기반 토마토 질병 분류를 위한 DCGAN 이미지 데이터 확장 영향 평가

박 지 연[1] · 김 현 지[1] · 김 경 백[2*]
[1]전남대학교 전자컴퓨터공학부 학사과정
[2]전남대학교 전자컴퓨터공학부 교수

Accessing Impact of DCGAN Image Data Augmentation for CNN based Tomato Disease Classification

Ji-yeon Park[1] · Hyeon-ji Kim[1] · Kyungbaek Kim[2*]
[1]Bachelor's Course, Department of Electronics and Computer Engineering, Chonnam National University
[2]Professor, Department of Electronics and Computer Engineering, Chonnam National University, 61186, South Korea

[요 약]

딥러닝의 발전과 CNN(Convolutional Neural Network)의 출현으로 이미지 데이터 분류에 관한 연구가 활발하게 진행되고 있다. 그러나 CNN 분류 모델의 학습에 클래스 분포가 불균형한 이미지 데이터셋을 사용할 경우 성능이 저하된다. 특히 식물에서 질병은 비주기적으로 발생하므로 불균형한 이미지 데이터가 제공된다. 이 논문에서는 불균형한 이미지 데이터가 제공되는 상황에서 CNN기반 토마토 질병 분류기의 성능을 향상시키기 위한 DCGAN(Deep Convolutional Generative Adversarial Network) 이미지 데이터 확장의 영향성을 평가한다. DCGAN은 이미지 데이터에 특화된 생성 모델로서 안정적인 학습이 가능하며 이미지 특징을 효과적으로 추출할 수 있다. 성능 평가를 위해 토마토 질병 이미지 데이터 셋을 사용하여 DCGAN 이미지 데이터 확장이 CNN기반 토마토 질병 분류기에 미치는 영향을 측정하였고, 이미지데이터 확장을 통해 최대 30%의 정확도를 높일 수 있음을 확인하였다.

[Abstract]

With the development of deep learning and the advent of CNN(Convolutional Neural Network), research on image data classification has been actively conducted. However, performance is deteriorated when an image dataset having an uneven distribution of classes is used for training a CNN classification model. In particular, diseases in plants occur aperiodically and unbalanced image data is provided. In this paper, we evaluate the impact of DCGAN(Deep Convolutional Generative Adversarial Network) image data augmentation to improve the performance of CNN-based tomato disease classifiers in situations where unbalanced image data is provided. DCGAN is a generation model specialized in image data, enabling stable learning and effectively extracting image features. For performance evaluation, the effect of DCGAN image data augmentation on a CNN-based tomato disease classifier was measured using a tomato disease image data set, and it was confirmed that the accuracy can be increased up to 30% through image data augmentation.

색인어 : 합성곱 신경망, 심층 합성곱 생성적 적대 신경망, 데이터셋 불균형, 이미지 데이터 확장, 분류기
Key word : CNN, DCGAN, Unbalanced Dataset, Image Data Augmentation, Classifier

http://dx.doi.org/10.9728/dcs.2020.21.5.959

Received 16 March 2020; **Revised** 15 May 2020
Accepted 25 May 2020

***Corresponding Author; Kyungbaek Kim**

Tel: +82-62-530-3438
E-mail: kyungbaekkim@jnu.ac.kr

Ⅰ. 서 론

하드웨어 발달로 인해 복잡한 연산이 가능해짐과 함께 빅데이터의 성장에 따라 딥러닝은 빠른 속도로 발전해왔다. 딥러닝은 자연어 처리, 음성 인식 등 다양한 분야에서 성공적인 성과를 보였으며 그 배경에는 CNN(Convolutional Neural Network : 합성곱 신경망)의 기술 개발이 큰 영향을 끼쳤다. 특히 이미지 분류에 있어 CNN은 기존의 K-NN, SVM같은 머신러닝과 비교하여 우수한 성능과 편리함을 가졌기에 현재 CNN을 통한 이미지 분류 모델에 관한 연구가 활발하게 진행되고 있다.

CNN 분류 모델의 학습에는 충분한 양과 균일한 클래스 분포를 가진 데이터 셋이 이상적이다. 하지만 실제로 제공되는 데이터 셋의 대부분은 클래스마다 개수의 차이가 존재한다. 특히 의료[2]나 농업 분야[13],[14]에서는 그 차이가 굉장히 큰 경우가 많다. 예를 들어 나뭇잎 이미지의 경우 건강한 잎의 이미지가 질병에 걸린 잎의 이미지보다 훨씬 많으므로 데이터가 한 클래스에 대해 치중되는 상황이 발생한다. 이렇듯 편향된 데이터 셋으로 학습된 CNN 분류 모델은 소수의 데이터를 가진 클래스를 잘못 분류할 가능성이 높아지며 이는 곧 분류 모델의 신뢰성에 악영향을 미치게 된다[1].

불균형한 데이터 문제를 해결하기 위해 새로운 분류 모델의 설계 또는 데이터 샘플링 기법 등이 제안되었다. 데이터 샘플링이란 불균형한 데이터를 균형 있게 맞추기 위해 클래스의 데이터 개수를 조정하는 기법으로, 다수 클래스(Majority Class)의 데이터만큼 소수 클래스(Minority Class)의 데이터 수를 늘리는 방법을 오버샘플링(over sampling)이라 한다. 오버샘플링에는 랜덤하게 소수 클래스의 데이터를 복제하는 랜덤 오버샘플링과 K-NN 알고리즘을 사용하는 데이터 합성 기법인 SMOTE(Synthetic Minority Oversampling Technique)[3] 등이 존재한다.

그러나 이미지 데이터에 대한 오버샘플링 기법은 단점을 가진다. 이미지는 많은 픽셀로 구성된 고차원의 데이터이며 픽셀 간의 연관관계를 파악하여 특징(feature)을 찾아내는 것이 이미지 분류 문제의 요점이다. 특히 컬러 이미지의 경우 처리해야할 차원은 더욱 복잡해지는데 K-NN 기반의 SMOTE 알고리즘의 경우 픽셀 간의 관계를 찾아내어 새로운 합성 이미지를 만드는 일에 적합하지 않다. 또한 랜덤 오버샘플링 같은 방식은 동일한 데이터를 반복적으로 학습하기 때문에 분류 모델이 과적합(Overfitting)될 수 있다. 최근 이러한 문제를 극복하기 위해, GAN (Genarative Adversarial Network : 생성적 적대 신경망 - 두 개의 네트워크로 구성된 심층 신경망 구조)를 이용한 이미지 데이터 확장 (Data Augmentation) 기법이 제안되고 있다.

이 논문에서는 DCGAN(Deep Convolutional Generative Adversarial Network)을 이용한 이미지 데이터 확장 기법을 CNN 기반의 분류기에 적용하는 방법을 제안하고, 그 성능을 평가한다. DCGAN은 기존의 GAN에 이미지 데이터의 특징을 효과적으로 추출할 수 있는 CNN 구조를 더함으로써 원본 이미지의 특징을 가진 유사 이미지를 생성할 수 있다. 이를 활용하여 DCGAN 모델에 식물 질병의 특징을 학습시키고 생성된 이미지를 소수 클래스에 더하여 불균형한 데이터셋의 균형을 맞춘어, CNN기반의 식물 질병 분류기를 학습시킨다.

제안된 기법의 성능평가를 위해 4종의 토마토 질병 이미지 데이터 셋을 사용하였다. CNN기반 토마토 질병 분류기를 학습하는데 있어서, 균형 잡힌 이미지 데이터 셋, 불균형한 이미지 데이터 셋, 오버샘플링 데이터 확장이 적용된 데이터 셋, 제안한 DCGAN 모델 기반의 데이터 확장이 적용된 데이터 셋을 사용하였다. 각 경우에 대해서 Precision, Recall, F1 Score, Accuracy를 측정하여 성능을 평가하였다.

이 논문의 구성은 다음과 같다. 2장에서는 CNN 분류기 및 이미지 데이터 확장 관련 연구에 대해 기술한다. 3장에서는 토마토 질병 이미지 데이터 셋과 이를 활용한 CNN기반 토마토 질병 이미지 분류 모델에 대해 기술하고, 4장에서는 DCGAN기반의 이미지 데이터 확장 기법에 대해 기술한다. 5장에서 실제 데이터 셋을 이용한 DCGAN기반의 이미지 데이터 확장이 CNN 기반 토마토 질병 분류기 성능에 미치는 영향을 평가한고, 6장에서 이 논문의 결론을 기술한다.

Ⅱ. 본 론

2-1 CNN 기반 이미지 분류

CNN[4]은 현재 딥러닝 분야에서 가장 많이 사용되는 알고리즘이다. Convolutional layer를 통해 특징을 자동으로 추출하며, 추출된 특징은 pooling layer를 통해 강화된다. 특징을 기반으로 CNN의 마지막 layer에서 분류 작업이 수행된다. CNN은 같은 크기의 fully-connected layer보다 더 적은 파라미터를 사용하여 더 높은 분류 성능을 보인다.

이러한 CNN을 사용하여 이미지 데이터를 분류하는 작업이 다양한 분야에서 이루어졌다. Zhang[5] 등은 CNN의 강력한 특징 학습 및 추론 능력을 통해 인공위성 원격 탐사 이미지를 분류하고자 하였다. Wallelign[6] 등은 LeNet 기반의 CNN을 사용하여 콩잎 이미지를 분류하여 콩과 관련된 식물 질병을 식별하는 연구를 진행했다. 한편, Razzak[7] 등은 딥러닝 기반의 의료 이미지 처리에 관한 연구들을 비교하면서 의료 이미지를 통한 질병 탐지에 CNN 구조의 모델이 많이 채택되면서 분류 성능도 높음을 보였다.

2-2 오버샘플링을 통한 이미지 데이터 확장

이미지 데이터에 대한 랜덤 오버샘플링은 소수 클래스의 이미지를 지정한 개수만큼 무작위로 선택하고 다수 클래스와 균형을 이룰 때까지의 이미지 복제를 통해 구현할 수 있다. 이 경우 복제된 이미지를 저장하기 위해 추가적으로 물리적인 저장 공간이 필요하다.

한편 코드를 통해 메모리에서 반복적으로 이미지 데이터를 불러옴으로써 랜덤 오버샘플링을 수행할 수 있다. 프로그램을 통해 읽어 들인 이미지 데이터를 변수에 저장하고 해당 변수를 원하는 횟수만큼 호출해서 사용한다. 이는 이미지를 가상 공간에 복제한 것이나 실제로 복제한 것과 같은 효과를 줄 수 있다. 프로그램의 종료와 함께 가상으로 복제했던 이미지 데이터는 소멸하게 되므로 추가적인 공간을 요구하지 않는다.

2-3 GAN 기반의 이미지 데이터 확장

오버샘플링을 통한 데이터 확장의 과적합 등의 한계를 해결하기 위해 다양한 분야에서 Goodfellow등이 제안한 GAN[8][12]을 활용한 이미지 데이터 확장 연구가 진행되었다.

Pinetz[9] 등은 MNIST 데이터를 불균형하게 조절하여 DGAN, RGAN, DAGAN을 통한 이미지 데이터 확장 후의 CNN 분류기 성능을 비교하였다. 세 모델 모두 DCGAN의 변형 모델이며 이미지 렌더링을 활용한 데이터 확장과 비교하여 GAN을 통한 데이터 확장 기법이 이미지 분류에 긍정적인 영향을 미칠 수 있음을 보였다. Maayan[10] 등은 데이터 수집에 어려움을 겪는 의료 분야에서 GAN 기반의 이미지 확장을 통한 간 병변 CNN 분류기의 성능 개선 방법을 제안하였다. 데이터의 label 정보를 같이 학습시키고 생성된 데이터의 클래스를 구분할 수 있는 ACGAN(Auxiliary Classifier GAN)과 DCGAN 모델을 사용하여 이미지 데이터를 확장하였고 두 모델을 비교하여 DCGAN이 ACGAN보다 간병변 분류기의 성능 향상을 이끌 수 있음을 보였다. 또한 Fang[11] 등은 바람과 비의 유무에 따른 레이더 탐지 이미지를 인식하는 CNN 모델의 성능을 개선하기 위해 DCGAN을 통한 이미지 확장을 제안하였다.

이 논문에서는 식물 질병, 특히 토마토 질병 이미지를 분류하는 CNN 분류기에서 발생할 수 있는 데이터 불균형의 문제를 DCGAN 기반의 이미지 데이터 확장을 통해 해결할 수 있는지를 평가한다.

Ⅲ. CNN 기반 토마토 질병 이미지 분류 모델

이 장에서는 먼저 데이터와 그 특성에 대하여 설명하고 토마토의 질병 분류 작업을 위해 제안한 CNN 구조에 대해 기술한다.

3-1 토마토 질병 데이터 셋

이 논문에서 사용하는 이미지 데이터 셋은 머신러닝 플랫폼 kaggle에서 제공하는 식물 질병 데이터 셋인 plant-disease를 사용하였다. plant-disease는 사과, 토마토 등 다양한 식물에서 나타나는 질병의 종류별로 train 데이터와 test 데이터를 제공한다.[13]

토마토의 질병 데이터 4종류와 Healthy 데이터를 선택하여 실험에 사용하였다. 질병은 각자의 증상을 가지는데, Bacterial Spot의 경우 잎에 외부는 황색, 내부는 암갈색의 작은 병반이 발생한다. Late Blight는 회록색의 병반이 발생하여 잎 전체에 다갈색으로 확대된다. Target Spot은 1mm 미만의 불규칙한 모양의 노란 반점이 잎 전체로 확대된다. Yellow Leaf Curl은 잎이 위축되고 뒤틀리며 노랗게 변색된다. Healthy는 어떠한 질병에도 걸리지 않은 정상의 데이터를 의미한다. 한편, 유사한 색상이나 형태의 병반이 발생하는 질병의 경우 분류 시 혼동될 수 있으며 보이는 증상이 굉장히 작은 경우도 잘못된 분류가 일어날 수 있다.

데이터는 질병의 피해가 보이는 토마토의 잎을 하나씩 분리하여 각각의 이파리를 촬영한 이미지이다. 분류 모델의 학습에 방해가 될 수 있는 외부 요소의 영향을 최소화하기 위하여 단색 배경을 바탕으로 이파리의 형태가 모두 보이도록 촬영되었다. 촬영된 데이터는 256x256 픽셀 사이즈를 가지며 그림 1과 같이 R, G, B 3개의 채널을 가지는 컬러 이미지이다.

식물 질병의 경우 각 종류마다 발병될 수 있는 온도, 습도 등의 조건과 발병 원인이 다르기 때문에 데이터 불균형이 발생할 수 있다. 특히, 토마토의 경우 질병 종류마다 데이터 개수의 차이가 있었으며, Yellow Leaf Curl의 경우에는 데이터가 다른 질병에 비해 3배 이상의 데이터가 존재하는 불균형한 상황을 확인할 수 있었다.

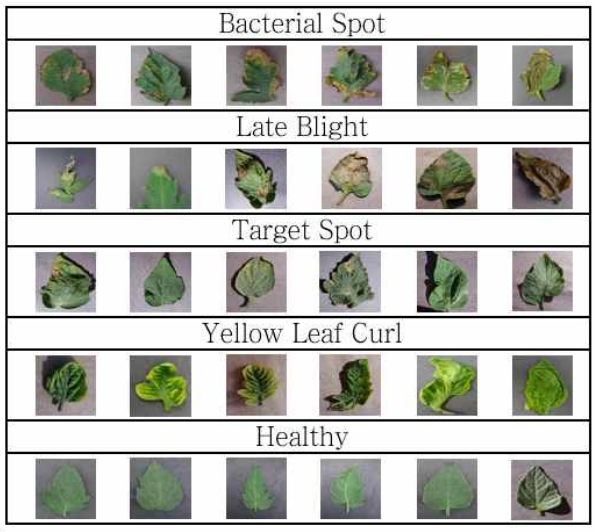

그림 1. 토마토 질병 이미지 데이터 셋 예시
Fig. 1. Examples of tomato disease dataset

3-2 질병 이미지 분류 모델

제안하는 질병 분류 모델의 구조는 그림 2와 같다. 특징 추출을 통해 다양한 분야에서 이미지 분류 작업에 뛰어난 성능을 보인 CNN을 사용하여 5개의 클래스를 구분하도록 한다. CNN 분류기의 입력 이미지는 메모리상에서의 수월한 연산을 위하여 32x32 픽셀 사이즈로의 다운샘플링을 진행하고 각 픽셀의 값

을 0~1 사이로 정규화 하였다.

제안하는 질병 이미지 분류 모델에는 convolutional layer와 그 후에 적용되는 max pooling layer의 쌍이 세 개 존재한다. 또한 다섯 개의 멀티 클래스 분류를 위해 마지막 fully-connected layer의 activation function으로 softmax를 사용하여 분류 모델의 예측값에 따라 클래스를 결정한다. Convolutional layer의 activation function으로는 ReLu 함수를 사용하였고 loss function으로 categorical cross-entropy를 지정하였다.

모델의 학습은 500 epoch, 100 batch size로 이루어지며 최적화를 위해 optimizer로 Adam을 사용하였다. 한편, 학습 중 분류 모델이 과적합되지 않도록 early stopping 객체를 callback 함수로 지정했으며, 25%의 뉴런이 dropout 되도록 설정했다.

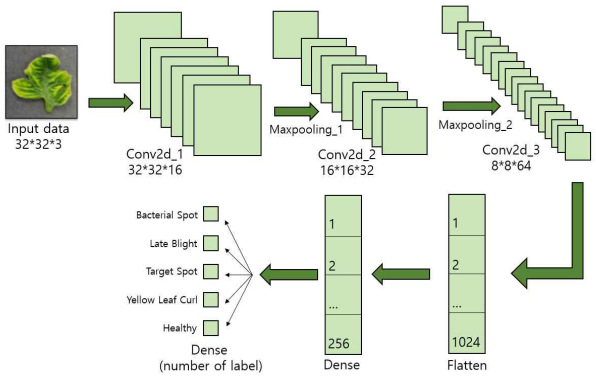

그림 2. CNN 토마토 질병 이미지 분류기의 구조
Fig. 2. Architecture of tomato disease image classifier CNN

Ⅳ. DCGAN 기반 이미지 데이터 확장

데이터 셋 설명에서 언급한 바와 같이, 식물 질병 분류 모델의 학습에 사용가능한 이미지 데이터 셋이 불균형하고, 이는 CNN 분류기의 성능을 떨어뜨릴 수 있다. 이를 해결하기 위해 부족한 데이터를 추가적으로 보완하여 CNN 질병 분류 모델의 성능을 향상시키는 기법이 필요하다. 이장에서는 DCGAN을 통해 새로운 토마토 질병 이미지를 생성하여 이미지 데이터를 확장하는 방법에 대해 설명한다.

DCGAN의 기본이 되는 GAN은 인공적인 데이터를 생성하는 Generator와 생성된 가짜 데이터를 원본 데이터와 비교하여 판별하는 Discriminator로 구성된다. Generator는 Discriminator가 원본 데이터라고 판단할 수 있도록 유사한 데이터를 생성해야 하며, Discriminator는 Generator가 생성한 데이터와 원본 데이터를 확실히 구분할 수 있어야 한다.

$$\min_G \max_D V(D,G) = E_{x \sim p_{data}(x)}[\log D(x)] + E_{x \sim p_x(z)}[\log(1 - D(G(z)))] \quad (1)$$

서로 대립하는 목적을 가진 두 신경망이 경쟁하는 그림 3과 같은 방식으로 GAN 모델의 학습이 진행되며 수식 (1)의 목적함수를 사용한다. G는 Generator, D는 Discriminator를 나타낸다. p_{data}는 원본 데이터의 분포, x는 p_{data}에서 추출한 표본이며 p_z는 노이즈의 분포, z는 p_z에서 추출한 표본을 말한다. G(z)는 노이즈 표본 z를 입력받아 생성한 인공적인 데이터를 의미한다. 한편, D는 자신의 입력 데이터가 원본 데이터라고 판단한 경우 1, 아니라고 판단한 경우 0을 출력한다. 그렇기에 원본 데이터 표본을 입력으로 받는 D(x)는 1을, Generator에서 생성한 인공적인 데이터를 입력으로 받는 D(G(z))는 0을 출력하도록 해야 한다. 즉, Discriminator는 V(D,G)를 최대화하는 것이 목적이다. 반면, Generator는 자신이 생성한 데이터를 Discriminator가 원본데이터라고 판단하게 만들어야 하므로 D(G(z))가 1이 되도록 시도하며 이는 V(D,G)를 최소화하는 것과 같다.

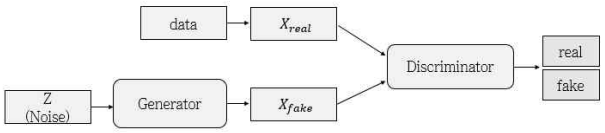

그림 3. GAN의 구조
Fig. 3. Architecture of GAN

DCGAN은 GAN과 기본적인 학습 방식은 동일하나 기존의 fully-connected한 신경망을 CNN으로 구성한 구조이다. 이미지의 특징을 읽어내는 성능이 뛰어난 CNN을 적용함으로써 원본 데이터의 특징이 학습된 인공적인 이미지를 생성할 수 있다. 이를 활용하여 토마토 질병 이미지를 생성할 수 있고 부족한 데이터를 보완하여 불균형한 데이터셋을 수정할 수 있다.

이 논문에서 제안하는 DCGAN의 Generator는 균등 분포(Uniform distribution)에서 100개의 랜덤한 노이즈 벡터를 입력받아 32x32x3의 사이즈를 가지는 토마토 질병 이미지를 출력한다. Generator의 구조는 그림 4와 같으며 8x8x256 사이즈의 fully-connected layer와 kernel 사이즈로 5x5를 사용하는 4개의 deconvolutional layer로 구성된다. Transposed convolution이라 불리는 deconvolution은 커널 사이에 0을 추가하여 convolution 연산을 역으로 수행하는 방법이다. 4개의 deconvolution layer 중 먼저 수행되는 2개의 layer 전에 upsampling을 적용하였다. 또한 마지막 레이어를 제외한 각 layer에 batch-normalization을 적용하여 입력 데이터의 분포가 뭉쳐있을 때 평균과 분산을 조절하여 안정적으로 학습을 진행하고자 하였다. 모든 layer의 activation function으로는 ReLu를 채택하였다.

Discriminator는 32x32x3의 토마토 질병 이미지를 입력받는 전형적인 CNN 구조로써 입력 이미지가 원본인지 인공적으로 생성된 이미지인지 결정한다. Discriminator의 구조는 그림 5와 같으며 4개의 convolutional layer와 하나의 fully-connected layer가 존재한다. Convolutional layer의 kernel 사이즈로 5x5, stride

크기를 2로 설정함으로써 pooling 레이어와 같은 차원 축소가 진행된다. 마지막 layer를 제외한 모든 layer에는 activation function으로 Leaky ReLU를 사용하였고 마지막 fully-connected layer에는 sigmoid를 적용하여 얻은 값을 통해 원본 이미지와 인공 이미지를 판단한다.

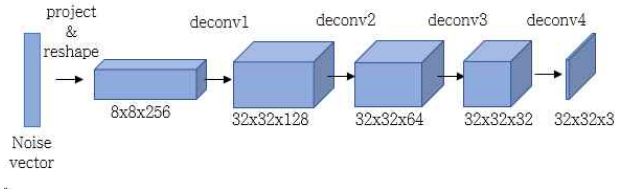

그림 4. 제안한 DCGAN Generator 구조
Fig. 4. Architecture of Proposed DCGAN Generator

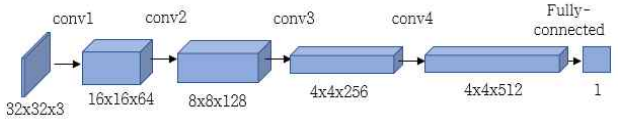

그림 5. 제안한 DCGAN Discriminator 구조
Fig. 5. Architecture of Proposed DCGAN Discriminator

제안하는 DCGAN 모델을 통한 토마토 질병 이미지 생성은 BinaryGAN 형태로 진행되었다. 예를 들어 Bacterial Spot에 대한 인공적인 이미지를 만들기 위해서 해당 클래스의 데이터만 사용하였으며 학습 과정에서 다른 질병 클래스는 사용하지 않는다. 이는 각 질병 클래스에 대한 DCGAN 모델이 존재하는 것과 같다.

학습은 Generator와 Discriminator가 반복적으로 번갈아가며 이루어지며 2000 epoch, 40 batch size로 진행하였다. 최적화를 위한 optimizer로 learning rate를 0.0001로 설정한 Adam을 사용하였고 LeakyReLU의 기울기를 0.2로 지정했다. 또한 모델의 과적합을 막기 위해 Generaotor의 40% 뉴런이 dropout되도록 설정하였다.

Ⅴ. 성능평가 및 분석

이 장에서는 제안하는 DCGAN 모델을 통해 생성된 인공적인 이미지를 통한 이미지 데이터 확장이 CNN 분류기에 미치는 영향을 평가하기 위해 수행하였던 실험 및 결과를 기술한다.

5-1 실험 환경

실험은 3.1절에서 기술한 데이터 셋을 이용해, BASE (균형잡힌 이미지 데이터 셋), IMBALANCE (불균형한 이미지 데이터 셋), OS-AUG (오버샘플링 데이터 확장이 적용된 데이터 셋), DCGAN-AUG (제안한 DCGAN 모델 기반의 데이터 확장이 적용된 데이터 셋)의 네 가지 상황에 따른 CNN 토마토 질병 분류기의 성능을 비교 하였다. 사용하는 데이터 셋은 4개의 질병 클래스(Bacterial Spot, Late Blight, Target Spot, Yello Leaf Curl)와 질병에 걸리지 않은 Healthy 클래스의 총 5개의 클래스로 분류되어 있다.

BASE의 경우, 모든 클래스의 이미지 데이터 셋의 크기가 4000장으로 통일되어 있다. 반면, IMBALANCE의 경우, BASE에서 토마토 질병 클래스 하나의 이미지 데이터 개수를 400장으로 줄인다. 이는 하나의 질병 클래스가 다른 데이터에 비해 1/10의 비율을 가진 불균형한 상황을 가정한 것이다. 이 과정을 모든 질병 클래스에 적용한다.

OS-AUG는 IMBALANCE의 상황에서 랜덤 오버샘플링을 사용하여 불균형한 클래스가 가지는 400장의 이미지 데이터를 메모리상에서 10번 복제하여 4000장으로 확장시킨다.

마지막으로 DCGAN-AUG는 IMBALANCE의 상황에서 제안하는 DCGAN모델을 사용하여 불균형한 클래스의 이미지 데이터 확장을 수행하는 상황을 가정한다. 먼저 IMBALANCE의 불균형한 클래스 이미지 데이터 400장을 제안하는 DCGAN 모델에 학습시켜 3600장의 새로운 이미지를 생성하였다. 이를 기존의 불균형 클래스의 데이터에 추가하여 모든 클래스의 이미지 데이터 셋의 크기를 4000장으로 통일하였다. 그림 6은 DCGAN으로 Yellow Leaf Curl 클래스에 대해 생성한 이미지 중 일부이다.

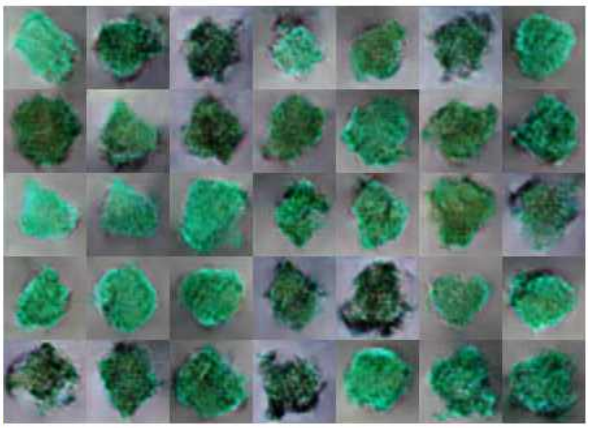

그림 6. DCGAN으로 생성한 Yellow Leaf Curl 이미지
Fig. 6. Yellow Leaf Curl image generated by DCGAN

각 상황별로 주어지는 이미지 데이터 셋을 이용하여 CNN 기반 토마토 질병 이미지 분류기를 학습시키고, 그 분류기의 성능 평가를 위해 각 클래스별 Precision, Recall, Accuracy, F1-score를 측정하였다. 질병 이미지 분류를 위한 CNN 모델과 질병 이미지 확장을 위한 DCGAN 모델 구현을 위해 Keras

framework를 사용하였다. 모든 학습 과정은 NVIDIA GeForce RTX 2080을 사용하는 GPU 환경에서 수행하였다.

한편, 토마토 질병 클래스의 이름이 길기 때문에, 이후에는 Bacterial Spot을 'bacteria', Late Blight는 'lateblight', Target Spot은 'targetspot', Yellow Leaf Curl은 'yellowleafcurl'로 변경하여 기술한다.

5-2 교차검증 기반 성능평가

각 상황별 주어진 이미지 데이터 셋을 이용하여 5-fold 교차 검증을 통해, CNN 분류기의 성능을 평가하였다. 교차 검증을 위해 각 클래스의 데이터 셋을 5개 세트로 나누고 4개 세트는 학습에, 남은 1개 세트는 평가에 사용하였으며 평가 세트를 다른 세트로 바꿔가며 총 5번의 평가를 진행하고, 결과의 평균을 구하여 최종 성능으로 선택하였다. 교차검증에 따른 각 데이터 상황별 질병 클래스의 분류 성능 결과인 Precision, Recall, F1-Score, Accuracy에 대해 그림 7, 그림 8, 그림 9, 그림 10에서 각각 나타낸다.

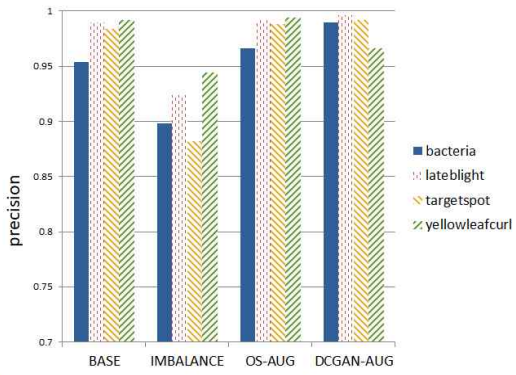

그림 7. 5-fold 교차 검증을 통한 Precision 비교
Fig. 7. Comparison of Precision with 5-fold cross validation

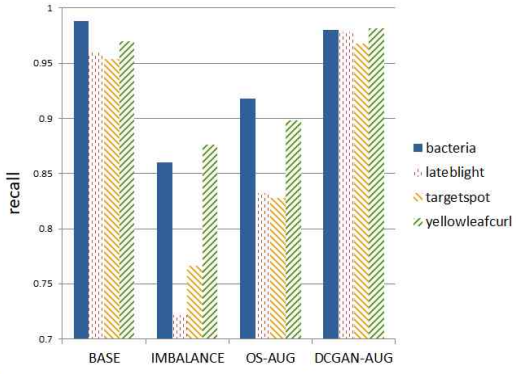

그림 8. 5-fold 교차 검증을 통한 Recall 비교
Fig. 8. Comparison of Recall with 5-fold cross validation

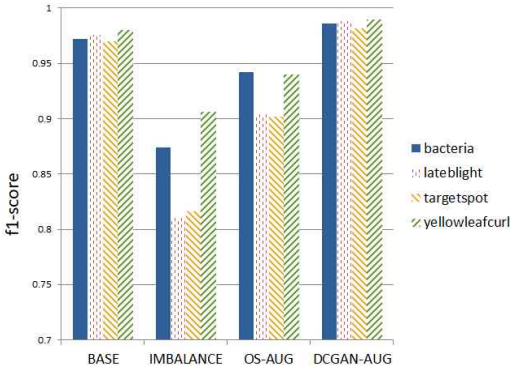

그림 9. 5-fold 교차 검증을 통한 F1-score 비교
Fig. 9. Comparison of F1-score with 5-fold cross validation

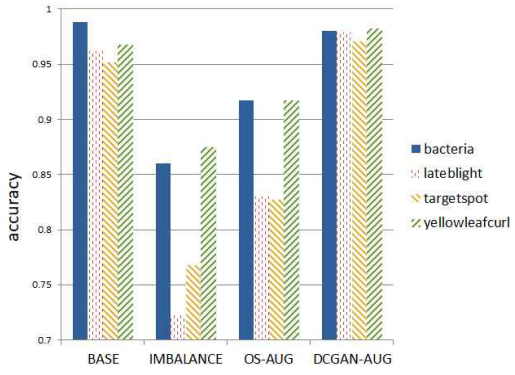

그림 10. 5-fold 교차 검증을 통한 Accuracy 비교
Fig. 10. Comparison of Accuracy with 5-fold cross validation

(1) BASE : 성능 평가 결과는 <표 1>과 같으며 BASE의 경우 모든 클래스의 분류 Accuracy가 95% 이상인 것을 확인할 수 있었다.

표 1. BASE의 5-fold 교차검증을 통한 성능 평가
Table 1. Performance Evaluation of BASE with 5-fold cross validation

	precision	recall	f1-score	accuracy
bacteria	0.954	0.988	0.972	0.98825
lateblight	0.99	0.96	0.976	0.962
targetspot	0.984	0.954	0.97	0.952
yellowleaf curl	0.992	0.97	0.98	0.968
healthy	0.95	1	0.976	1

(2) IMBALANCE : 데이터를 400장으로 줄인, 즉, 불균형한 클래스의 성능이 BASE와 비교해 하락하는 것을 확인하였다. 특히 5~10% 이내로 감소한 Precision과 달리 Recall은 Precision

의 약 2배인 10~25% 정도 하락했다. Precision과 Recall의 조화 평균인 F1-score도 이에 영향을 받아 감소했으며 Accuracy의 경우 Recall과 매우 유사한 하락폭을 보였다.

(3) OS-AUG: IMBALANCE와 비교하여 모든 클래스의 성능이 향상됨을 볼 수 있었다. Precision은 5~10%, Recall의 경우 3~10% 증가하였으며 F1-score와 Accuracy도 유사한 모습을 보였다. Precision의 경우 BASE와 거의 차이가 나지 않을 정도로 향상되었으나 Recall은 아직 BASE의 Recall과 5~13% 정도의 차이가 나는 것을 확인할 수 있었다.

(4) DCGAN-AUG: 질병 이미지 분류기의 학습 및 평가 결과 전반적인 성능이 IMBALANCE에 비해 크게 향상되었고, OS-AUG에 비해서도 좋은 성능을 보임을 확인하였다. Accuracy의 경우 DCGAN-AUG는 IMBALANCE에 비해 최대 30%의 성능향상을 보이며 OS-AUG에 비해서도 최대 16%의 성능향상을 보인다.

전반적인 성능평가 결과에 따라, 이미지 데이터 불균형에 따른 CNN 기반 분류기의 성능 하락을 DCGAN 기반의 이미지 데이터 확장 기법을 통해 방지 할 수 있음을 확인 할 수 있었다. 이미지 데이터 셋이 불균형할 때 CNN 분류기의 성능은 확연히 감소하며 오버샘플링이나 DCGAN을 사용한 이미지 데이터 확장 후 성능이 향상되는 것을 확인할 수 있었다. 특히 오버샘플링보다 DCGAN을 통한 데이터 확장에서 성능이 뛰어난 것을 보아 DCGAN 기반의 이미지 데이터 확장은 데이터 불균형에 따른 토마토 질병 이미지 분류 성능하락을 방지할 수 있는 가능성이 있다고 판단된다.

5-3 독립 데이터 셋 기반 성능평가

5.2절에서 수행하였던 교차검증과 달리, 학습데이터에 대해 독립적인 이미지 데이터 셋을 대상으로 이미지 데이터 확장이 분류기에 미치는 성능을 평가하였다. 이를 위해, 추가 이미지 데이터 셋을 테스트 셋으로 활용하여 각 상황별 CNN 분류기 성능 평가를 수행하였다. 그 결과, 교차검증 결과와는 달리, DCGAN기반의 이미지 데이터 셋 확장에 따른 성능향상이 두드러지지 않는다는 점을 확인할 수 있었다. 이는, 기존의 적은 수량의 데이터로 학습된 DCGAN 모델로 생성한 인공적인 이미지로 학습된 CNN 분류기가 같은 클래스로 분류되는 새로운 이미지와 유사도가 떨어지기 때문에 일어나는 현상으로 분석된다. 즉, DCGAN으로 생성한 이미지가 해당 토마토 질병의 범용적 특징을 따라가지 못할 경우, 데이터 확장에 따른 CNN 분류기의 성능향상에 한계가 있다고 평가된다.

Ⅵ. 결론

이 논문에서는 클래스 불균형이 존재하는 토마토 질병 이미지 데이터의 분류 성능을 개선하기 위해 DCGAN을 활용한 이미지 데이터 확장에 대한 영향을 평가하였다. 제안하는 DCGAN 기반 이미지 데이터 확장의 영향이 유의미함을 확인하기 위하여 여러 가지 데이터 불균형 상황 및 데이터 확장 상황에서 토마토 질병 이미지 분류기의 성능 검증하였다. 생성된 이미지를 포함한 데이터 셋을 이용한 교차검증을 통해 제안하는 이미지 데이터 확장이 CNN기반 분류기의 성능을 향상시킬 수 있는 가능성을 확인하였다. 또한, 이미지 데이터 셋 확장 기법이 토마토 질병 분류기의 성능향상에 미치는 영향을 고도화하기 위해서는 질병의 증상을 표현하는 특징을 잘 추출할 수 있는 연구가 필수적임을 확인하였다.

감사의 글

이 논문은 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(NRF-2017R1A2B4012559). 이 논문은 정부(과학기술정보통신부)의 재원으로 한국연구재단 바이오·의료기술개발사업의 지원을 받아 수행된 연구임(NRF-2019M3E5D1A02067961).

참고문헌

[1] M. Mostafizur Rahman, D. N. Davis, “Addressing the Class Imbalance Problem in Medical Datasets”, I*nternational Journal of Machine Learning and Computing*, Vol.3, No.2, 2013.

[2] R. O’Brien, H. Ishwaran, “A random forests quantile classifier for class imbalanced data”, *Pattern Recognition*, Vol.90, pp.232-249, 2019.

[3] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, “SMOTE: Synthetic Minority Over-sampling Technique”, *Journal of Aritificial Intelligence Research*, Volume 16, pp.321-357, 2002.

[4] A. Krizhevsky, I. Sutckever, G. E. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks”, *Advances in Neural Information Processing Systems* 25(NIPS), 2012.

[5] W. Zhang, P. Tang, L. Zhao, “Remote Sensing Image Scene Classification Using CNN-CapsNet”, *Remote Sensing*, 2019.

[6] S. Wallelign, M. Polceanu, C. Buche, “Soyean Plant Disease Identification Using Convolutional Neural Network”, *FLAIRS-31*, pp.146-151, 2018.

[7] M. I. Razzak, S. Naz, A. Zaib, “Deep Learning for Medical Image Processing : Overview, Challenges and Future”, *Classification in BioApps*, pp.323-350, 2017.

[8] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D.

Warde-Farley, et al. “Generative adversarial nets”, *Advances in Neural Information Processing Systems*, pp.2672-2680, 2014.

[9] T. Pinetz, R. Johannes, S. Daniel, “Actual Impact of GAN Augmentation on CNN Classification Performance”, *Proc. of the 8th International Conference on Pattern Recognition Applications and Methods* − Volume1:ICPRAM, pp.15-23, 2019.

[10] F-A. Maayan, K. Eyal, G. Jacob, G. Hayit, “GAN-based data augmentation for improved liver lesion classification”, arXiv preprint, 2018.

[11] W. Fang, F. Zhang, V.S. Sheng, Y. Ding, “A Method for Improving CNN-Based Image Recognition Using DCGAN”, *Computers, Meterials&Continua*, Volume 57, pp.167-178, 2018.

[12] Yoon Han, Hyoung Joong Kim, “Face Morphing Using Generative Adversarial Networks”, *Journal of Digital Contents Society*, 19(3), pp. 435-443, 2018

[13] Hyeonji Kim, Jiyeon Park, Kyungbaek Kim, “Performance enhancement of CNN based tomato pest classification with GAN”, *In proceedings of 2019 Korea Institute of Smart Media Fall Conference*, Gwangju, November 7-8, 2019

[14] Van-Quyet Nguye, Sinh Ngoc Nguyen, Kyungbaek Kim, “Design of a Platform for Collecting and Analyzing Agricultural BigData”, *Journal of Digital Contents Society*, 18(1), pp. 149-158, 2017

박지연(Ji-Yeon Park)

2017년～현 재: 전남대학교 전자컴퓨터공학부 학사과정

※관심분야 : 데이터 분석(Data Analysis), 딥러닝(Deep Learning), IoT 등

김현지(Hyeon-Ji Kim)

2017년～현 재: 전남대학교 전자컴퓨터공학부 학사과정

※관심분야 : 데이터 분석(Data Analysis), 딥러닝(Deep Learning) 등

김경백(Kyungbaek Kim)

1999년: 한국과학기술원(KAIST) 학사
2001년: 한국과학기술원(KAIST) 석사
2007년: 한국과학기술원(KAIST) 박사
2007~2011: University of California Irvine, 박사후연구원
2012~ now : 전남대학교 전자컴퓨터공학부 교수

※ 관심분야 : 소프트웨어 정의 네트워크/인프라, 빅데이터 플랫폼, 그리드/클라우드 네트워크 시스템, 소셜 네트워킹 시스템, 인공지능 적용 가상물리시스템, 블록체인