IOWA STATE UNIVERSITY
Digital Repository

Graduate Theses and Dissertations

Iowa State University Capstones, Theses and Dissertations

2019

Improving plant disease recognition with generative adversarial network under limited training set

Luning Bi
Iowa State University

Follow this and additional works at: https://lib.dr.iastate.edu/etd

 Part of the Operational Research Commons

Recommended Citation

Bi, Luning, "Improving plant disease recognition with generative adversarial network under limited training set" (2019). *Graduate Theses and Dissertations*. 17647.
https://lib.dr.iastate.edu/etd/17647

This Thesis is brought to you for free and open access by the Iowa State University Capstones, Theses and Dissertations at Iowa State University Digital Repository. It has been accepted for inclusion in Graduate Theses and Dissertations by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.

www.manaraa.com

Improving plant disease recognition with generative adversarial network under limited training set

by

Luning Bi

A thesis submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Major: Industrial Engineering

Program of Study Committee:
Guiping Hu, Major Professor
Daren Mueller
Qing Li

The student author, whose presentation of the scholarship herein was approved by the program of study committee, is solely responsible for the content of this thesis. The Graduate College will ensure this thesis is globally accessible and will not permit alterations after a degree is conferred.

Iowa State University

Ames, Iowa

2019

Copyright © Luning Bi, 2019. All rights reserved.

www.manaraa.com

DEDICATION

I would like to dedicate this thesis to my girlfriend Zhuqing Liu without her support I would not have been able to complete this work. I would also like to thank my friends and family for their loving guidance and financial assistance during the writing of this work.

www.manaraa.com

TABLE OF CONTENTS

Page

www.manaraa.com

LIST OF TABLES

www.manaraa.com

LIST OF FIGURES

www.manaraa.com

ACKNOWLEDGMENTS

I would like to take this opportunity to express my thanks to those who helped me with various aspects of conducting research and the writing of this thesis. First and foremost, Dr. Guiping Hu for her guidance, patience and support throughout this research and the writing of this thesis. Her insights and words of encouragement have often inspired me and renewed my hopes for completing my graduate education. I would also like to thank my committee members for their efforts and contributions to this work: Dr. Qing Li and Dr. Daren Mueller.

www.manaraa.com

ABSTRACT

This thesis introduces a generative adversarial network (GAN) based method to classify diseased images using limited training set. A general introduction of machine learning applications in agriculture domain is provided. The issue of plant disease recognition has been investigated in this thesis.

First, the successful applications of convolutional neural networks (CNNs) to plant disease classification have been reviewed. It is found out that most of methods are built under the assumption that there is enough training set. The issue of limited training data is overlooked. Thus, the over-fitting problem caused by a limited training set is discussed.

Second, a new approach is proposed to solve the limited training set problem. The proposed method consists of four parts: CNN, data augmentation, GAN and label smoothing regularization (LSR). CNN is used to classify plant diseases and species. Data augmentation and GAN are used to generate additional samples for training. LSR technique can help the model avoid the over-fitting problem.

Finally, three comparison experiments have been designed. The analysis proves the effectiveness of the proposed method. Compared with using the real dataset only, the proposed method improves the prediction accuracy by 6%.

www.manaraa.com

CHAPTER 1. GENERAL INTRODUCTION

1.1 Introduction

The world population is expected to grow from 7.2 billion to 9.6 billion in 2100. This imposes rising demand in agriculture production. To alleviate this challenge, using different techniques to manage crop efficiently is necessary. Recently, machine learning methods have become increasingly popular. They have been successfully applied to various domains, such as speech recognition, face recognition, and automatic driving. This thesis aims to apply the machine learning method in agriculture settings.

There are two main reasons behind the increasing popularity and applications in various scientific disciplines. The first is the increase of computing resources. The development of GPU accelerates the computing of matrix calculation and further analysis. The second is the development of data storing and management techniques makes it possible to collect, store, and manage data of increasing sizes. The commonly used machine learning models include regression, decision tree, Naive Bayes, support vector machine (SVM), and deep neural network (DNN). A brief introduction of regression, decision tree and deep neural network follows.

- **Regression**. Regression models are used to analyze the relationship between one or more independent variables and an outcome variables [1]. The basic two regression models are linear regression model and logistic regression model. Linear regression model is easy and quick to establish. However, it assumes that the relationship between input variables and the outcome variable is linear. Logistic regression model adds a Sigmoid function to the linear model so that it can calculate the probability of the sample belonging to a certain class.

- **Decision tree**. A decision tree consists of three parts, i.e., branches, internal nodes and leaf nodes. The internal nodes are a set of conditions that can divide the samples into different

www.manaraa.com

classes. The branches represent the outcome of internal nodes. The leaf nodes represent the label of the class. Decision tree can break down a complex classification problem into a set of simpler decisions at each stage [2].

- **Deep neural network**. DNN is a class of methods that use multiple layers to extract information from the input data [3]. The most common training method is a gradient-based algorithm that can calculate the gradients and update the weights and biases iteratively. The final goal of training is to minimize the loss function. DNN is able to deal with large datasets and execute feature engineering without explicitly programming.

Borrowing the successful experiment from other disciplines, many researchers in agriculture domain have started to design and apply the machine learning methods. The current applications of machine learning methods to crop management can be divided to three categories, i.e., species recognition, yield prediction, and disease detection [4].

- **Species recognition**. Automatic species recognition can help reduce the classification time and human factors. One type of study is to classify different crop plants. Grinblat et al. proposed a deep convolutional neural network for plant identification using vein morphological patterns. He pointed out that it was not necessary to build a feature extraction method for this task [5]. Wu et al. proposed a approach based on artificial neural network for the automated leaf recognition.The prediction accuracy of classifying 32 kinds of plants was greater than 90% [6]. Another study was the classification between crop plants and weed. Pantazi et al. used a variant of artificial neural network (ANN) to identity the weed in a field using unmanned aircraft system (UAS) multispectral imagery [7]. Ahmed et al. used texture features and SVM to classify the weed image, which achieved 98.5% prediction accuracy [8].

- **Yield prediction**. The world population continues to increase which imposes rising demand in agriculture production. How to improve crop breeding to feed the growing population is a significant challenge. Predictive modeling on crop phenotype can speed up the process and make it resource efficient. The commonly used tools for yield prediction include regression

www.manaraa.com

model and DNN. Singh et al. predicted the crop yield by using piecewise linear regression model. The predicted values were very close to the observed values [9]. Ramos et al. constructed a machine vision system (MVS) to count the number of fruits on a coffee branch, which showed a correlation higher than 0.9 at early states of crop development [10]. Kaul et al. used ANN models to predict Maryland corn and soybean yield. The experiments showed that the prediction accuracy of ANNs is higher than that of regression models [11].

- **Disease detection**. Plant diseases are responsible for the 13% of the production potential [12]. Early detection and timely management of plant diseases are essential to reducing yield loss. Traditional manual inspection is often time-consuming, laborious and biased. Recently, automated imaging techniques have been successfully applied to the detection of plant diseases. Convolutional neural network (CNN) is one of the most popular methods. Different from full connected DNNs, CNNs have two special types of layer. The convolutional layers extract features from the input images. The pooling layers reduce the dimensionality of the features. Dhakate et al. used a CNN for the recognition of pomegranate plant diseases and achieved 90% overall accuracy [13]. Ferentinos developed CNN models to classify the healthy and diseased plants. The success rate reached 99.53% [14].

In this thesis, the recognition of multiple plant disease types and multiple species under limited training set has been investigated. Many studies have achieved high prediction accuracy of plant diseases by using CNN. However, those approaches were built under the assumption that there are enough training samples. In practice, the data collection and data annotation is expensive. Experiments shows that the use of limited image dataset will affect directly the prediction accuracy. Because the DNN does not get enough samples to learn the distribution of data. Therefore, the motivation of this thesis is to improve the prediction accuracy of plant diseases using a limited training set.

www.manaraa.com

Our contributions are as follows.

- A CNN is built for the classification of multiple diseases and multiple species.
- A GAN-based approach is proposed to generate additional images for training. CNN is used as the basic network to classify species and diseases. GAN and label smoothing regularization (LSR) are combined to generate additional training images. The regular data augmentation techniques are also used to enlarge the dataset.
- A dataset from plantvillage.com is used as a case study. Three experiments, i.e., using the CNN only, using CNN and data augmentation, and using the proposed method, have been designed. The results show that compared with using the real dataset only, the proposed method can improve the prediction accuracy by 6%.

1.2 Thesis organization

The organization of this thesis is as follows. Chapter 1 begins with a general introduction of the application methods in agriculture. The motivation and the contributions of this thesis are also elaborated. Chapter 2 is an article to be submitted to the Computers and Electronics in Agriculture. It introduces a GAN-based approach for the plant disease recognition under limited training set. Chapter 3 concludes with the results and future work.

www.manaraa.com

CHAPTER 2. IMPROVING PLANT DISEASE RECOGNITION WITH GENERATIVE ADVERSARIAL NETWORK UNDER LIMITED TRAINING SET

A paper submitted to *Computers and Electronics in Agriculture*

Luning Bi and Guiping Hu

2.1 Abstract

Traditionally, plant disease recognition has been carried out visually by human. It is often biased, time-consuming, and laborious. Borrowing from the success of machine learning in computer science, methods based on deep learning have been proposed to improve the disease recognition process. Convolutional neural networks (CNNs) have been adopted and proven to be very effective in plant disease recognition. Despite the good recognition accuracy achieved by CNNs, the issue of limited training data is often overlooked. In most cases, the training dataset is often small since data collection and annotation require significant effort. In this case, CNN method tends to have the overfitting problem. In this paper, a generative adversarial network (GAN) based method has been proposed to improve the prediction accuracy and address the overfitting problem under limited training data. Different from the traditional GAN, our GAN is combined with a regularization kernel. Experiments show that compared to using the real dataset only, the proposed GAN enhanced recognition method can improve the overall classification accuracy of plant diseases by 6%.

2.2 Introduction

With the increasing global population, the demand for agriculture production is rising [15]. Plant diseases cause substantial management issues and economic losses in the agricultural industry

www.manaraa.com

[16]. It has been reported that at least 10% of global food production is lost due to plant disease [17]. The situation is becoming increasingly complicated because climate change alters the rates of pathogen development and diseases are transferred from one region to another more easily due to the global transportation network expansion [18]. Therefore, early detection, timely mitigation and disease management are essential for agriculture production [19].

Traditionally, plant disease inspection and recognition have been carried out through optical observation of the symptoms on plant leaves by human with some training or experience. Plant disease recognition has known to be time-consuming and error prone. Due to the high degree of complexity and the large number of cultivated plants and their existing physiological problems, even experts with rich experience often fail to diagnose specific diseases and consequently lead to mistaken disease treatments [14].

To address the above-mentioned problems, many methods have been developed to assist the disease recognition and management. In the past decades, laboratory techniques have been developed and established. The commonly used techniques for plant disease recognition include enzyme-linked immunosorbent assay (ELISA), polymerase chain reaction (PCR), immunoflourescence (IF), flow cytometry, fluorescence in situ hybridization (FISH), and DNA microarrays [20]. However, these techniques require an elaborate procedure and consumable reagents. Therefore, they are slow and expensive. Under these circumstances, image recognition of plant diseases, which identifies plant diseases by the plant appearance and visual symptoms, becomes popular. The advantages of image recognition include: (1) the ability to deal with high amount input parameters, i.e., image pixels, (2) the minimization of human errors, and (3) the reduction of processing time [21]. The key to improving the image recognition accuracy of plant diseases is to extract the right features to classify the plant disease types [22, 23]. The emergence of deep learning techniques provides an improved automated solution. Although deep learning based models take longer time to train than other traditional approaches (e.g. support vector machine (SVM), k-nearest neighbors algorithm (KNN)), its testing time is less because all information from training dataset has been integrated in the neural network [24]. In the agricultural area, CNN has been widely used for image recognition [25].

www.manaraa.com

Dhakate et al. used a CNN for the recognition of pomegranate plant diseases and achieved 90% overall accuracy [13]. Ghazi et al. proposed a hybrid method of GoogLeNet, AlexNet and VGGNet to classify 91,758 labeled images of different plant organs. Their combined system achieved an overall accuracy of 80% [26]. Ferentinos developed convolutional neural network models to classify the healthy and diseased plants using 87,848 images. The success rate was significantly high which can reach 99.53% [14]. Ma et al. proposed a deep CNN to symptom-wise recognition of four cucumber diseases. The model was trained using 14,208 images and achieved an accuracy of 93.4% [27]. Based on that high level of performance achieved in the above studies, it can be concluded that CNNs are highly suitable for the recognition of plant diseases through the analysis of simple leaf images [5].

It should be noted that the high prediction accuracy is from that thousands of labeled images used to train CNNs. A major problem often facing the automatic identification of plant diseases using CNN is the lack of labeled images capable of representing the wide variety of conditions and symptom characteristics found in practice [28].Experimental results indicate that the use of limited image datasets for training will lead to some undesirable consequences [29]. Because real datasets do not have enough samples for deep neural networks to properly learn the classes and the annotation errors may damage the learning process [19].

Although it is relatively cheap to collect plant image data, using additional unlabeled data is non-trivial to avoid model overfitting. If the model learns to assign a full probability to the ground truth label for each training example, it is not guaranteed to generalize because the model becomes too confident about its predictions [30]. This is the primary motivation for this study.

In this study, we designed a generative adversarial network (GAN) to enlarge the training set by generating more labeled images. GAN was proposed to learn generative models based on game theory [31]. The goal of GANs is to train a generator network that produces additional labeled samples from the input vectors of noise. The training signal for the generator is provided by a discriminator network that is trained to distinguish samples from the generator distribution from real data. The generator network in turn is then trained to fool the discriminator into accepting

www.manaraa.com

its outputs as being real [32]. The GAN approach is capable of generating high quality labeled images. Emily et al. proposed a method based on GAN to generate samples of natural images [33]. Radford et al. proved that GANs have the ability to learn a hierarchy of representations from different image datasets [34].

In most cases, GAN is used to generate images similar to the real images as much as possible. However, the role of the GAN in this paper is to overcome the overfitting problem and improve the prediction accuracy on the real dataset. If the labels of the generated images are easy to be classified, the effectiveness of the approach will be limited. Therefore, the regularization method is employed to improve the generalization performance of the GAN. Regularization is often carried out by augmenting the loss function with a regularization term [35]. Label smoothing regularization (LSR) has been reported in Szegedy et al. [30]. Instead of maximizing the predicted probability of the truth-ground class only, LSR maximizes the predicted probability of the truth-ground class as well as the non-truth ground classes. Similarly, Xie et al. proposed a method named DisturbLabel which prevents the overfitting problem by adding label noises to the CNN [36]. Pereyra et al. found out that label smoothing can improve the performance of the models on benchmarks without changing other parameters [37]. In our paper, GAN is combined with the label smoothing regularization to generate images that can enlarge training dataset and regularize the CNN model simultaneously.

It should be noted the majority of the existing studies focused on a single type of disease or only one plant type. In reality, there may exist multiple diseases for one plant type. In some cases, it is also necessary to detect the multiple diseases of multiple plant types. Therefore, the ideal recognition method would have the ability to deal with the multi-disease and multi-plant type situation. To improve the generalization of the proposed method, multiple diseases and multiple plant types have been considered in this paper.

This paper is organized as follows. Section 2.3 introduces the motivation of this paper and the structure of the proposed regularized GAN-based approach. Section 2.4 includes a case study, the experiment results and comparisons. Finally, the paper concludes with the summary, findings, and future research directions in Section 2.5.

www.manaraa.com

2.3 Materials and Methods

Recent development in agricultural area leads to increasing demand for a non-destructive method of plant disease recognition [20]. A desirable tool for plant disease recognition should be fast and accurate. Currently, many image-based plant disease recognition techniques have been developed due to the low cost for image collection and the ability to deal with large-scale disease diagnosis. The typical imaging technique is first taking the photos of the plant leaves, then extracting the features of diseased plant leaves, and finally classifying based on additional analysis.

Most existing studies on plant disease recognition used large datasets to train their models, for example, CNNs. However, in most cases, there is not enough data available and the data annotation is very expensive. Under these circumstances, the models are easy to be overfitted because they contain more parameters than the number of samples can accommodate. The consequence of overfitting is that the model will fail to predict for new observations. This serves as the major motivation for this study, which aims to achieve high plant disease recognition accuracy with the limited training dataset. This is achieved with a novel data augmentation method based on regularized GAN to generate additional labeled images as detailed in this section.

2.3.1 Framework of the proposed method

To improve the prediction accuracy of CNN in the recognition of plant diseases using limited training dataset, three techniques have been designed and implemented in this study, i.e., data augmentation, generative adversarial network, and label smoothing regularization. As shown in Figure 2.1, real images are used to train the GAN with regularization. Then the trained GAN is used to generate additional labeled images. The generated images will be mixed with real images and then augmented through data augmentation. Finally, the dataset will be used to train the CNN.

Therefore, the proposed method consists of four components. The first component is CNN, which is used to classify plant disease types and plant species. The second component is data augmentation. It makes some minor modifications to the images. The third component is GAN.

www.manaraa.com

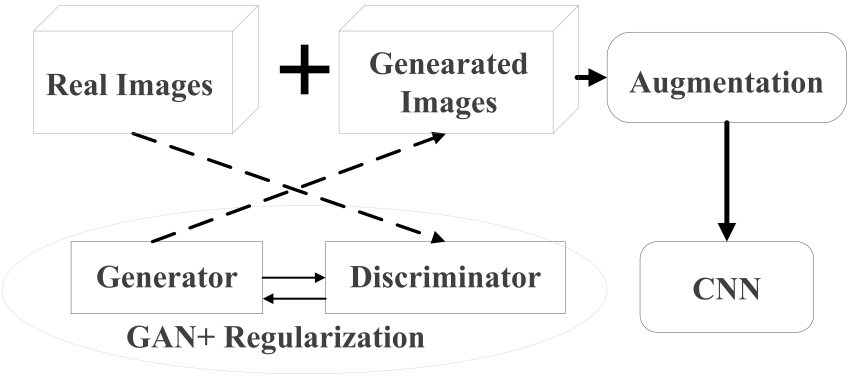


Figure 2.1 Framework of the proposed method

It is used to generate additional images. The fourth component is LSR. This can avoid overfitting problem by modifying the loss function of GAN.

2.3.2 Convolutional Neural Networks (CNN)

Convolutional neural networks (CNN) are used as the supporting framework of our method. CNN is a class of deep, feed-forward artificial neural networks. It was adopted widely for its fast deployment and high performance on image recognition tasks. The convolutional layers extract features from the input images whose dimensionality is then reduced by the pooling layers. The fully connected layers are placed near the output of the model. They act as classifiers to learn the non-linear combination of the high-level features and to make numerical predictions [24].

However, CNN needs a large training dataset. In our case, there are only 10-28 images in each category. Since the number of model parameters is greater than the number of data samples, small training dataset will lead to the overfitting problem. Overfitting results from a model that responds too closely or exactly to a particular dataset and will, therefore, fail to fit additional data or predict future observations reliably.

To solve the above problem, data augmentation, generative adversarial network and label smoothing regularization are used.

www.manaraa.com

2.3.3 Data Augmentation

Data augmentation is a relatively straightforward method to increase the number of labeled images. The most used methods include vertical flipping, horizontal flipping, 90° counterclockwise rotation, 180° rotation, 90° clockwise rotation, random brightness decrease, random brightness increase, contrast enhancement, contrast reduction and sharpness enhancement. Figure 2.2 lists the examples of original image (Figure 2.2(a)), rotation (Figure 2.2(b)), brightness increase (Figure 2.2(c)) and contrast increase (Figure 2.2(d)). In this paper, the fit_generator API is called. The generator is run in parallel to the model, for improved efficiency. For instance, this allows us to do real-time data augmentation on images on CPU in parallel to training our model on GPU.

(a) Original (b) Rotation (c) Brightness (d) Contrast

Figure 2.2 Augmentation methods

2.3.4 Generative Adversarial Network (GAN)

Unlike regular data augmentation methods, GAN is able to generate new images for training, which increases the diversity of data. GANs were firstly introduced in a paper by Ian Goodfellow and other researchers in 2014 [31]. The generative adversarial networks (GANs) consist of two sub-networks: a generator and a discriminator. The generator captures the training data distribution while the discriminator estimates the probability that an image came from the training data rather than the generator.

$$\min_G \max_D V(D,G) = E_{x\sim p_{data}(x)}[\log D(x)] + E_{z\sim p_{Noise}(z)}[\log(1-D(x))] \tag{2.1}$$

www.manaraa.com

Where D represents the discriminator network, G is the generator network, z is a noise vector drawn from a distribution $p_{Noise(z)}$, x is a real image drawn from the original dataset $p_{data(x)}$.

The idea behind Eq. (2.1) is that it increases the ability of the generator to fool the discriminator which is trained to distinguish generated images from real images.

The training process of GAN is shown in Figure 2.3. The specific steps are as follows.

Step 1: Initialize the parameters of the generator and the discriminator

Step 2: Sample a batch of noise samples for the generator. Usually, uniform distribution or Gaussian distribution is used.

Step 3: Use the generator to transform the noise samples into images that are labeled as fake.

Step 4: The real images are labeled as true. Then the real images and the generated images are mixed and used as the input of the discriminator.

Step 5: Train the discriminator to improve the ability to classify the generated images and the real images.

Step 6: Train the generator to generate more images that will be discriminated as true by the discriminator.

Step 7: Repeat Step 2 - Step 6.

2.3.5 Label Smoothing Regularization (LSR)

LSR is used to modify the loss function of GAN. In the training of GAN, the most widely used loss function is the cross-entropy loss as Eq. (2.2).

Where i is the index of the disease type, N is the total number of disease types, $p(i)$ is the predicted probability of the generated image belonging to class i, $q(i)$ equals to 1, if the label of the generated image is i; otherwise, $q(i)$ equals to 0.

$$L = -\sum_{i=1}^{N} \log(p(i))q(i) \tag{2.2}$$

www.manaraa.com

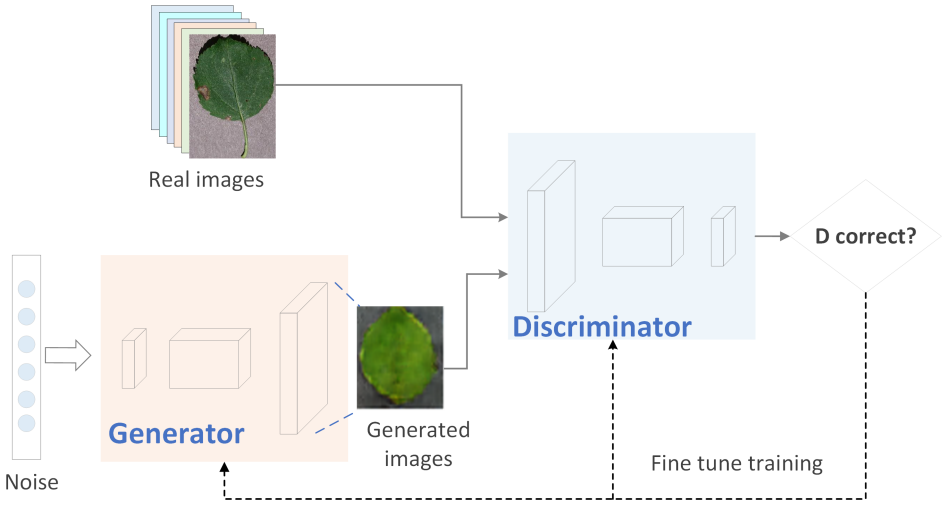


Figure 2.3 Training process of the GAN

The minimization of the cross-entropy loss is achieved when the predicted probability of ground-truth classes is maximum. However, if the model assigns full probability to the ground-truth label, it is likely to be overfitted. In other words, it will be very easy for CNN to determine the truth-ground classes of the generated images. It means that the improvement brought by generating additional images for training will be limited. Thus, the regularization is introduced. Regularization is a technique that makes the model less confident such that the model generalizes better. Label smoothing regularization (LSR) method is used in this paper. The objective function of GAN is as Eq. (2.3) [19].

$$L_{LSR} = -(1-\varepsilon)\log(p(y)) - \frac{\varepsilon}{N}\sum_{i=1}^{N}\log(p(i)) \tag{2.3}$$

Where ε is a hyperparameter between 0 and 1, i is the index of the disease type, N is the total number of disease types, $p(i)$ is the predicted probability of the generated image belonging to non-truth ground class i, $p(y)$ is the predicted probability of the generated image belonging to truth-ground class y.

In addition to maximizing the predicted probability of the truth-ground class, the LSR function also maximizes the predicted probability of the other non-truth ground classes. In other words, each

www.manaraa.com

generated image contains the features of all disease types, which can improve the generalization ability of the model. In practice, a generated image will be assigned with the label of the largest predicted possibility.

2.4 Case Study

To validate the effectiveness of the proposed method, a case study has been carried out. A dataset that contains the images of different plant diseases and species has been selected to demonstrate and validate the proposed method. To simulate the situation where there is not enough data, a fraction of the dataset was used for training. Three experiments were conducted to compare the results of different methods using different measurements.

2.4.1 Data Source and Performance Measure

The dataset used in this paper is from www.plantvillage.org, containing 54,309 images. As shown in Table 2.1, the images include 14 crop species: Apple, Blueberry, Cherry, Corn, Grape, Orange, Peach, Bell Pepper, Potato, Raspberry, Soybean, Squash, Strawberry, Tomato. It contains images of 17 fungal diseases, 4 bacterial diseases, 2 mold (oomycete) diseases, 2 viral diseases, and 1 disease caused by a mite. Twelve crop species also have images of healthy leaves that are not visibly affected by a disease [38]. The total number of classes is 38 which includes 12 groups of healthy leaves and 26 groups of diseased leaves.

For plant disease recognition, often there are only a limited number of images for some specified diseases. To address the small dataset problem, 873 images were randomly selected as the training dataset. For each category, there are only 10-28 images for training. Nine hundred and fifty images were randomly selected as the testing dataset. Three experiments have been designed. The first is to train a CNN using the real dataset. The second is to train a CNN using the real dataset and augmented dataset. The third one is to train a CNN using the dataset generated by the proposed method.

Three indices are used as the measurement in this paper, i.e., overall accuracy, precision and recall.

www.manaraa.com

Table 2.1 Dataset for image classification of plant disease (- means lack of data)

Species	**Number of disease images**	**Number of disease types**	**Number of health images**
Apple	1527	3	1645
Blueberry	-	-	1502
Cherry	1052	1	854
Corn	2690	3	1162
Grape	3640	3	423
Orange	5507	1	-
Peach	2291	1	360
Bell Pepper	997	1	1478
Potato	2000	2	152
Raspberry	-	-	371
Soybean	-	-	5090
Squash	1835	1	-
Strawberry	1109	1	456
Tomato	2170	9	1592

The calculation formula are as Eq. (2.4) - Eq. (2.6).

$$OverallAccuracy = (TruePositive + TrueNegative)/Total \tag{2.4}$$

$$Recall = TruePositive/(TruePositive + FalseNegative) \tag{2.5}$$

$$Precision = TruePositive/(TruePositive + FalsePositive) \tag{2.6}$$

The definitions of True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) are shown in Table 2.2. Since the problem stressed in this paper is a multi-class classification problem, a small modification has been made as Eq. (2.7) and Eq. (2.8).

$$Recall_i = M_{ii}/\sum_j M_{ij} \tag{2.7}$$

$$Precision_i = M_{ii}/\sum_j M_{ji} \tag{2.8}$$

Where M_{ij} is the number of images belonging to the ith category that are predicted to be in the jth category.

www.manaraa.com

Table 2.2 Definitions of TP, FP, TN and FN

Predicted \ Actual	True	False
Positive	TP	FP
Negative	TN	FN

2.4.2 Parameters of neural networks

For the generator, we established a network with 100 random vector input. Then a dense layer is used to covert the input vector to a vector of size 128*16*16. Through three convolutional layers, the output is 64*64*3 image.

For the discriminator, all input images have been resized to 64 *64*3. The real images are assigned with label "1" while the generated images are assigned with label "0". If the input image is discriminated as real, the output will be the label of the disease type which represented by 1 38 in this case; otherwise, the output will be "0". The optimizer is Adam with the parameters $\alpha = 0.0002$ and $\beta = 0.5$. The objective function of the discriminator is binary cross-entropy function. The objective of training the discriminator is to improve the ability to tell the reality of the input images.

For the combination of the generator and the discriminator, the output of the generator and the input of the discriminator is connected. The parameters of the discriminator are fixed. The objective of training the combination network is to improve the reality of the generated images.

The CNN used to classify the images is composed of four modules. The input layer is formed by the symptom images in RGB color space with a size of 64*64*3. The first module consists of a Convolutional Layer that has 64 filters with a size of 33, and a Max-pooling Layer with the filter that has a size of 33 and a stride of 2. Each Convolutional Layer is connected by a Batch Normalization Layer performed over channels. The second module consists of a Convolutional Layer that has 64 filters with a size of 33, and a Max-pooling Layer with the filter that has a size of 33 and a stride of 2. Each Convolutional Layer is connected by a Batch Normalization Layer performed over channels. The third module consists of three Convolutional Layers that have 128

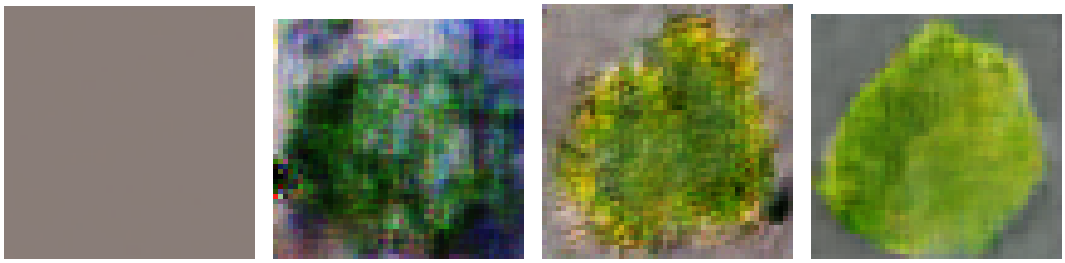
www.manaraa.com

filters with a size of 33. Each Convolutional Layer is connected by a Batch Normalization Layer performed over channels. The last module of the CNN consists of a Fully Connected Layer with 128 neurons. The output layer has 38 neurons representing the 38 classes of leaves. Given the output layer, softmax function was used to calculate the estimated probability of each classes.

All the above networks were built using the Keras framework [39] and were trained on a NVIDIA GTX 1080 with CUDA 9.0.

2.4.3 Results and Comparisons

The most important process is the training of GAN. The training effectiveness can be illustrated by Figure 2.4. At the beginning, the output of the generator is just noise. After 2000 iterations, the outline of the leaf can be identified visually. At the 12,000th iteration, the shape of the leaf is much clearer. Figure 2.5(a)shows the real images drawn from 38 categories while Figure 2.5(b)shows the 38 samples generated by the regularized GAN. Each sample belongs to one unique class. The trained generator is used to generate additional images. Those images are mixed with real images and used as the input of the CNN. The training process of the CNN is shown in Figure 2.6. From Figure 2.6, we can find that after 700 epochs, the test accuracy can reach 84%.

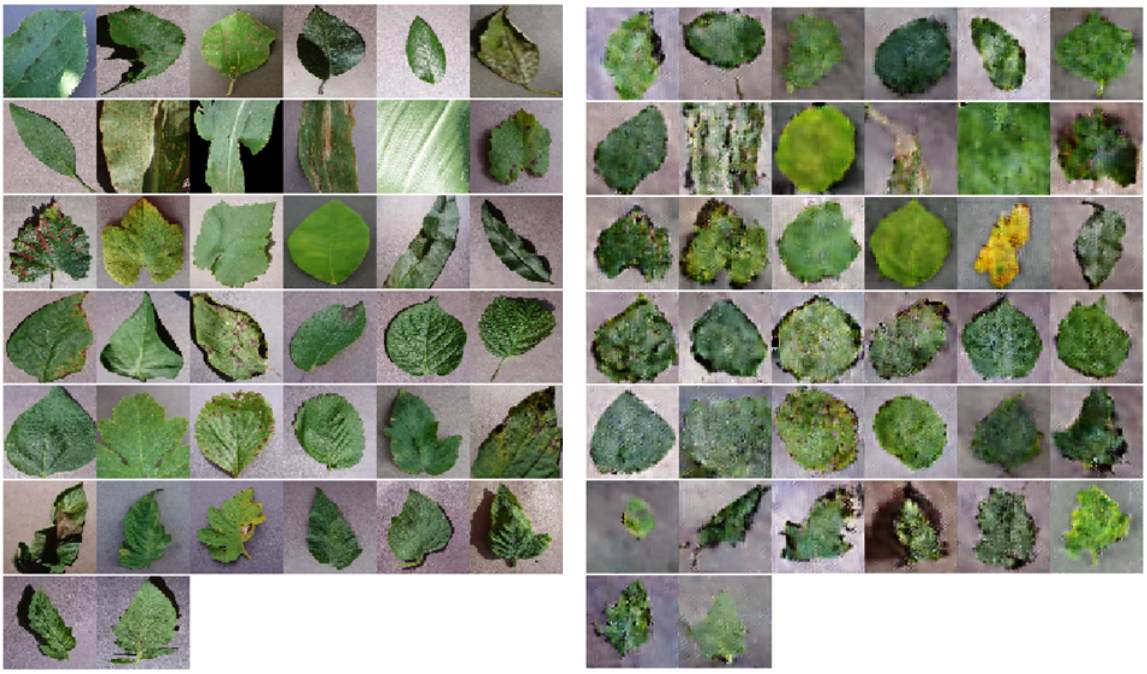


(a) 0 (b) 800 (c) 2,000 (d) 12,000

Figure 2.4 Generated images in different training stages (# of iterations)

To prove the validity of the proposed method, another experiment was conducted without using the generated images. The number of epochs is the same as that in the first experiment. The training process is illustrated in Figure 2.7. It shows that the test accuracy is about 78%, which is 6% less than the proposed method.

www.manaraa.com



(a) Original images (b) Samples generated by the proposed GAN

Figure 2.5 Original images and generated samples

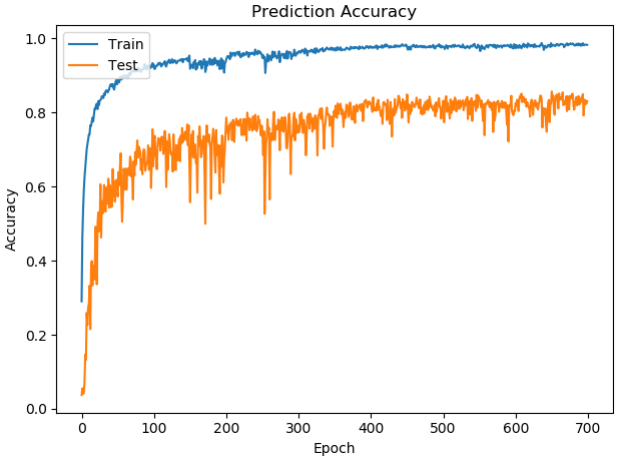


Figure 2.6 Prediction accuracy of the CNN using the proposed method

www.manaraa.com

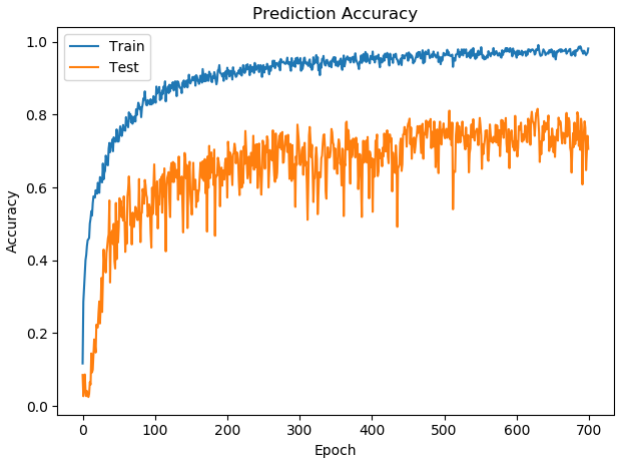


Figure 2.7 Prediction accuracy of the CNN without generated images

To further investigate the influence of the data augmentation technique, we only use the real image dataset to train the CNN. The result is shown in Figure 2.8. It can be found that after about 70 epochs, the train accuracy is close to 1 while the test accuracy is only about 60%. This is an indicator that the model is overfitted.

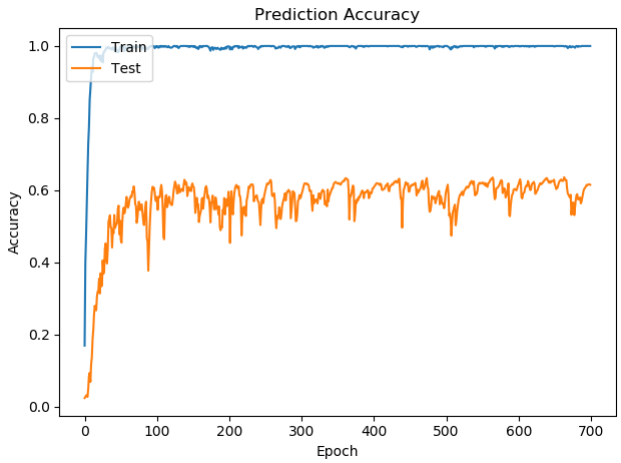


Figure 2.8 Prediction accuracy of the CNN without generated images and data augmentation technique

Table 2.3 lists the train accuracy and test accuracy of the above three experiments. Compared to using CNN only, the proposed method improves the test accuracy by 24%. Compared to using

www.manaraa.com

Table 2.3 Comparisons among three methods

Methods	**Train accuracy**	**Test accuracy**
Pure CNN	100%	60%
CNN+ data augmentation	98%	78%
Proposed method	98%	84%

CNN with data augmentation method, the proposed method can improve the test accuracy by 6%. Table 2.4 lists the recall and precision of 26 diseases calculated according to Eq. (2.7) and Eq. (2.8). It can be found out that compared to using the CNN only, the advantages of the CNN with data augmentation and the proposed method are dominant. Both the recall and the precision of the proposed method are much better than that of the CNN-only approach. The proposed method outperforms the CNN with data augmentation on most of the diseases. When comparing the recall and the precision of each disease type, specific patterns of the models can be observed. For example, the difference between the recall and the precision of apple botryospaeria obtuse is extreme for all three models. The recall is just 0.1-0.2 while the precision is 1. This means only a small number of images which have apple botryospaeria obtuse are classified as apple botryospaeria obtuse. However, all images predicted that are classified to be apple botryospaeria obtuse are correctly labeled. Therefore, the prediction of disease apple botryospaeria obtuse is highly reliable but the sensitivity of the model is low since the false negative predictions are high. The comparison between the recall and the precision of each disease type can help to gain additional insights on the models and make the right decision according to different situations.

2.5 Conclusion

Plant disease recognition plays an important role in disease detection, mitigation, and management. Even though some deep learning methods have achieved good results in plant disease classification, the problem of the limited dataset is overlooked. In practice, it is time-consuming to collect and annotate data. The performance of CNN will drop dramatically if there are not enough

www.manaraa.com

Table 2.4 Recall and precision of 26 diseases (R: Recall; P: Precision)

Specie	Disease type	Pure CNN		CNN+Data Augmentation		Proposed method	
		R	P	R	P	R	P
Apple	Botryospaeria obtuse	0.109	0.208	0.739	0.447	0.500	0.676
	Venturia inaequalis	0.379	0.440	0.724	0.575	0.776	0.978
	Gymnosporangium juniperi-virginianae	0.033	0.083	0.200	1.000	0.133	1.000
Cherry	Podosphaera spp	0.429	0.457	0.837	0.732	0.857	0.933
Corn	Cercospora zeae-maydis	0.250	0.381	0.844	0.435	0.406	0.684
	Puccinia sorghi	0.744	0.644	0.878	0.832	0.833	0.882
	Exserohilum turcicum	0.696	0.658	0.783	0.844	0.928	0.587
Grape	Guignardia bidwellii	0.383	0.493	0.670	0.685	0.628	0.808
	Phaeomoniella spp.	0.607	0.602	0.846	0.917	0.829	0.898
	Pseudocercospora vitis	0.689	0.689	0.844	0.962	0.889	0.889
Orange	Candidatus Liberibacter	0.919	0.835	0.961	0.949	0.944	0.969
Peach	Xanthomonas campestris	0.701	0.649	0.856	0.800	0.936	0.684
Pepper	Xanthomonas campestris	0.208	0.357	0.615	0.656	0.865	0.856
Potato	Alternaria solani	0.642	0.559	0.778	0.913	0.642	0.839
	Phytophthora Infestans	0.172	0.345	0.483	0.824	0.672	0.609
Squash	Erysiphe cichoracearum	0.768	0.642	0.923	0.912	0.923	0.945
Strawberry	Diplocarpon earlianum	0.554	0.600	0.923	0.833	0.969	0.716
Tomato	Xanthomonascampestris pv. vesicatoria	0.583	0.583	0.828	0.877	0.890	0.780
	Alternaria solani	0.086	0.116	0.290	0.730	0.538	0.595
	Phytophthora Infestans	0.345	0.471	0.542	0.856	0.599	0.773
	Fulvia fulva	0.371	0.441	0.729	0.895	0.714	0.694
	Septoria lycopersici	0.228	0.287	0.603	0.804	0.596	0.920
	Tetranychus urticae	0.678	0.490	0.564	0.824	0.886	0.820
	Corynespora cassiicola	0.198	0.375	0.479	0.773	0.512	0.886
	Mosaic Virus	0.796	0.798	0.909	0.979	0.976	0.938
	Yello leaf curl virus	0.346	0.375	1.000	0.531	0.962	1.000

www.manaraa.com

training data. Therefore, a method for plant disease recognition under the limited training dataset is necessary.

In this paper, a CNN is built for the plant disease recognition, which can recognize multiple species and diseases. To address the overfitting problem caused by the limited training dataset, a GAN-based approach is proposed. The label smoothing regularization method is also employed, which works by adding a regularization term to the loss function.

The experiments show that the proposed method can improve the prediction accuracy by 6% than the CNN with regular data augmentation method. Compared with using the CNN only, the proposed method can improve the prediction accuracy by 24%. Based on our work, plant disease recognition can be conducted under the limited training dataset, which will bring benefits to the rapid diagnosis of plant diseases.

It should be noted that this proposed plant disease recognition using a small dataset can be further investigated and improved due to the limitations of the currently proposed method. First, it takes much time to train the GAN and generate new labeled images for training. Next, the improvement made by the proposed method is related to the size of the real image dataset. If the size is large, the test accuracy achieved by using a pure CNN is already 99%. The improvement made by adding other techniques is not significant. If the size is very small, it is not able to extract enough information to generate new labeled images. Last, in this paper, we only used the basic CNN framework. There are some other more complicated CNN-based models. In future, we will try different CNN frameworks and investigate the relationship between the size of the real image dataset and the effectiveness of the proposed method.

www.manaraa.com

CHAPTER 3. FUTURE WORK SUMMARY AND DISCUSSION

More and more machine learning techniques have been applied in yield prediction, species recognition and disease detection. This thesis explained the problem of limited training set in plant disease recognition. Many studies applied the CNN to the plant disease recognition and achieved high prediction accuracy. However, the high performance of CNNs is based on large datasets. If the training set is limited, it will lead to overfitting problem. Aimed to alleviate the overfitting issue and improve the prediction accuracy, a GAN-based approach is proposed.

Generative adversarial network has been applied in many fields. In this thesis, GAN is used to generate additional images for training. It should be noted the improvement brought by a single GAN is limited. The core of this study is that it combines the GAN and LSR techniques. The generated images can increase the diversity of training set as well as the generalization ability of the CNN. Three experiments have been designed. The first is to use the CNN only. The second is to use data augmentation techniques to enlarge the dataset first and then train the CNN to classify plant diseases. The third is the proposed method, i.e., using regularized-GAN to generate additional images. Compared with using the real dataset only, the proposed method can improve the prediction accuracy by 6%. This approach provides a new way to do plant disease recognition under limited training set. To summarize, the contributions of this study are as follows. First, a CNN is established to classify multiple species and multiple diseases. Second, a GAN-based approach is proposed to generate additional images for training. Third, the dataset from plantvillage.com is used as a case study. The experiment results have proved the effectiveness of the proposed method.

This thesis focuses on the individual plants. The limitation includes two aspects. First, to collect the dataset like this, we need to take a picture of each leaf. It is time consuming and labor intensive. Second, it is not able to monitor how the plant disease changes over time.

www.manaraa.com

In reality, it is more reasonable and efficient to diagnose the plant disease from the level of farm field. Remote sensing using satellite imagery provides an option to researchers. First, the marginal cost of satellite imagery is low. Second, it can continuously monitor the state of the field. Third, it is more flexible. Theoretically, the images of the farm fields at any location or time can be recorded. Last but not least, the resolution of satellite imagery is high which can achieve 3m $\times$ 3m or even 72cm $\times$ 72cm.

More than just identifying the plant disease area, we are trying to investigate how the disease rate changes over time and space. Sudden death syndrome (SDS) is a major soybean disease in the Midwest. A satellite imagery dataset collected from PlanetScope during 2016-2018 is used as the training set. Four spectral bands of red, blue, green and near-infrared (NIR) can be extracted from the satellite imagery. Therefore, our future work is to establish a time series model that can predict the disease rate change in the next time window. The prediction of plant disease can also provide useful information for the decision making of crop management. The relationship between plant disease rate and different management factors (e.g., crop genotype, planting procedures, soil management) can be further investigated.

Machine learning is a very useful tool for many issues in agriculture. However, there are some limitations that prevent the dissemination of this kind of technique. First, the algorithms need large datasets to learn the distribution of data. The cost of data collection and data annotation is very high. Second, some algorithms, e.g., deep neural network, are black-box algorithms. It is hard to explain and investigate the relationship between the input variables and the output variables. Our future work will focus on two aspects. The first is to apply machine learning to computer vision problems in agriculture. The second is to do some optimization work based on the existing methods to further improve the algorithm performance on different tasks.

www.manaraa.com

BIBLIOGRAPHY

[1] Evangelos C Alexopoulos. Introduction to multivariate regression analysis. *Hippokratia*, 14(Suppl 1):23, 2010.

[2] S Rasoul Safavian and David Landgrebe. A survey of decision tree classifier methodology. *IEEE transactions on systems, man, and cybernetics*, 21(3):660–674, 1991.

[3] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256, 2010.

[4] Konstantinos G Liakos, Patrizia Busato, Dimitrios Moshou, Simon Pearson, and Dionysis Bochtis. Machine learning in agriculture: A review. *Sensors*, 18(8):2674, 2018.

[5] Guillermo L Grinblat, Lucas C Uzal, Mónica G Larese, and Pablo M Granitto. Deep learning for plant identification using vein morphological patterns. *Computers and Electronics in Agriculture*, 127:418–424, 2016.

[6] Stephen Gang Wu, Forrest Sheng Bao, Eric You Xu, Yu-Xuan Wang, Yi-Fan Chang, and Qiao-Liang Xiang. A leaf recognition algorithm for plant classification using probabilistic neural network. In *2007 IEEE international symposium on signal processing and information technology*, pages 11–16. IEEE, 2007.

[7] Xanthoula Eirini Pantazi, Alexandra A Tamouridou, TK Alexandridis, Anastasia L Lagopodi, Javid Kashefi, and Dimitrios Moshou. Evaluation of hierarchical self-organising maps for weed mapping using uas multispectral imagery. *Computers and Electronics in Agriculture*, 139:224–230, 2017.

[8] Faisal Ahmed, Md Hasanul Kabir, Shayla Bhuyan, Hossain Bari, and Emam Hossain. Automated weed classification with local pattern-based texture descriptors. *Int. Arab J. Inf. Technol.*, 11(1):87–94, 2014.

[9] Ramesh P Singh, Anup Krishna Prasad, Vinod Tare, and Menas Kafatos. Crop yield prediction using piecewise linear regression with a break point and weather and agricultural parameters, April 20 2010. US Patent 7,702,597.

[10] PJ Ramos, Flavio Augusto Prieto, EC Montoya, and Carlos Eugenio Oliveros. Automatic fruit count on coffee branches using computer vision. *Computers and Electronics in Agriculture*, 137:9–22, 2017.

[11] Monisha Kaul, Robert L Hill, and Charles Walthall. Artificial neural networks for corn and soybean yield prediction. *Agricultural Systems*, 85(1):1–18, 2005.

www.manaraa.com

[12] Anne-Katrin Mahlein, Erich-Christian Oerke, Ulrike Steiner, and Heinz-Wilhelm Dehne. Recent advances in sensing plant diseases for precision crop protection. *European Journal of Plant Pathology*, 133(1):197–209, 2012.

[13] Mrunmayee Dhakate and AB Ingole. Diagnosis of pomegranate plant diseases using neural network. In *2015 fifth national conference on computer vision, pattern recognition, image processing and graphics (NCVPRIPG)*, pages 1–4. IEEE, 2015.

[14] Konstantinos P Ferentinos. Deep learning models for plant disease detection and diagnosis. *Computers and Electronics in Agriculture*, 145:311–318, 2018.

[15] Saeed Khaki and Lizhi Wang. Crop yield prediction using deep neural networks. *Frontiers in plant science*, 10, 2019.

[16] SS Abu-Naser, KA Kashkash, and M Fayyad. Developing an expert system for plant disease diagnosis. *Journal of Artificial Intelligence*, 3(4):269–276, 2010.

[17] Richard N Strange and Peter R Scott. Plant disease: a threat to global food security. *Annu. Rev. Phytopathol.*, 43:83–116, 2005.

[18] Srdjan Sladojevic, Marko Arsenovic, Andras Anderla, Dubravko Culibrk, and Darko Stefanovic. Deep neural networks based recognition of plant diseases by leaf image classification. *Computational intelligence and neuroscience*, 2016, 2016.

[19] Jayme GA Barbedo. Factors influencing the use of deep learning for plant disease recognition. *Biosystems engineering*, 172:84–91, 2018.

[20] Sindhuja Sankaran, Ashish Mishra, Reza Ehsani, and Cristina Davis. A review of advanced techniques for detecting plant diseases. *Computers and Electronics in Agriculture*, 72(1):1–13, 2010.

[21] Jayamala K Patil and Raj Kumar. Advances in image processing for detection of plant diseases. *Journal of Advanced Bioinformatics Applications and Research*, 2(2):135–141, 2011.

[22] Shanwen Zhang and Zhen Wang. Cucumber disease recognition based on global-local singular value decomposition. *Neurocomputing*, 205:341–348, 2016.

[23] YG Naresh and HS Nagendraswamy. Classification of medicinal plants: an approach using modified lbp with symbolic representation. *Neurocomputing*, 173:1789–1797, 2016.

[24] Andreas Kamilaris and Francesc X Prenafeta-Boldú. Deep learning in agriculture: A survey. *Computers and electronics in agriculture*, 147:70–90, 2018.

www.manaraa.com

[25] Yang Lu, Shujuan Yi, Nianyin Zeng, Yurong Liu, and Yong Zhang. Identification of rice diseases using deep convolutional neural networks. *Neurocomputing*, 267:378–384, 2017.

[26] Mostafa Mehdipour Ghazi, Berrin Yanikoglu, and Erchan Aptoula. Plant identification using deep neural networks via optimization of transfer learning parameters. *Neurocomputing*, 235:228–235, 2017.

[27] Juncheng Ma, Keming Du, Feixiang Zheng, Lingxian Zhang, Zhihong Gong, and Zhongfu Sun. A recognition method for cucumber diseases using leaf symptom images based on deep convolutional neural network. *Computers and electronics in agriculture*, 154:18–24, 2018.

[28] Jayme Garcia Arnal Barbedo. Plant disease identification from individual lesions and spots using deep learning. *Biosystems Engineering*, 180:96–107, 2019.

[29] Jayme Garcia Arnal Barbedo. Impact of dataset size and variety on the effectiveness of deep learning and transfer learning for plant disease classification. *Computers and electronics in agriculture*, 153:46–53, 2018.

[30] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[31] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[32] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Advances in neural information processing systems*, pages 2234–2242, 2016.

[33] Emily Denton, Soumith Chintala, Arthur Szlam, and Rob Fergus. Deep generative image models using a laplacian pyramid of adversarial networks: Supplementary material.

[34] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[35] Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1979–1993, 2018.

[36] Lingxi Xie, Jingdong Wang, Zhen Wei, Meng Wang, and Qi Tian. Disturblabel: Regularizing cnn on the loss layer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4753–4762, 2016.

www.manaraa.com

[37] Gabriel Pereyra, George Tucker, Jan Chorowski, Łukasz Kaiser, and Geoffrey Hinton. Regularizing neural networks by penalizing confident output distributions. *arXiv preprint arXiv:1701.06548*, 2017.

[38] David Hughes, Marcel Salathé, et al. An open access repository of images on plant health to enable the development of mobile disease diagnostics. *arXiv preprint arXiv:1511.08060*, 2015.

[39] François Chollet et al. Keras, 2015.

www.manaraa.com