American Journal of Biochemistry and Biotechnology

Original Research Paper

Application of Data Enhancement Method Based on Generative Adversarial Networks for Soybean Leaf Disease Identification

[1]Xiao Yu, [1]Cong Chen, [1]Qi Gong and [2]Lina Lu

[1]School of Computer Science and Technology, Shandong University of Technology, Zibo, 255000, China
[2]Business School, Shandong University of Technology, Zibo, 255000, China

Article history
Received: 04-07-2022
Revised: 23-09-2022
Accepted: 30-09-2022

Corresponding Author:
Lina Lu
Business School, Shandong University of Technology, China
Email: think0759@126.com

Abstract: Soybean leaf disease data collection is an expensive and time-consuming task. Convolutional neural network training requires a large amount of data, but traditional data enhancement methods (such as rotation, flipping, translation) are restricted by fixed rules and cannot generate images with diversity and variability. Aiming at the problem of the lack of soybean leaf disease data set, this study proposes a data enhancement method based on Generative Adversarial Networks (GANs) for soybean leaf disease identification. The method is based on a cyclic adversarial network and its discriminator uses dense connections. Strategies to reduce the size and computational complexity of the final model. Using a cyclic adversarial network to convert between healthy and diseased leaves, unsupervised learning can be performed, using limited images to learn disease characteristics, there by generating highly recognizable soybean leaf images. Synthesis images generated from GANs and original images are fed together as the model training set input and the recognition model for recognizing 9 types of soybean leaf images is obtained. An accuracy rate of 95.89% can be achieved on the verification set. Experimental results show that the generative adversarial network provided in this article can: Generate soybean leaf disease image data with high discriminative features, increase the size of the data set and provide a feasible solution for soybean leaf disease image data enhancement; as A regularization strategy to reduce over-fitting problems and improve the performance of the recognition model.

Keywords: Deep Learning, Generating Adversarial Networks, Convolutional Neural Networks, Agricultural Pests and Diseases

Introduction

Plant diseases have a serious impact on the yield and quality of plants. Therefore, early detection and diagnosis of plant diseases are of great significance to the adoption of preventive and therapeutic measures. However, it is difficult to distinguish these plant characteristics (location, shape, color, morphology, etc.), which further becomes a challenge for early diagnosis and treatment of plant diseases. Among them, the leaf is the most frequently observed part of detecting diseases. Therefore, researchers are trying to use leaf images to automatically detect and diagnose plant diseases (Kaur *et al*., 2019).

With the development of computer vision, Convolutional Neural Network (CNN) has made important breakthroughs in the field of image recognition (Feng and Hua, 2021). Ferentinos (2018) proposed a CNN model based on deep learning methods, using 87,848 healthy and diseased plant leaf images for training and the best performance reached 99.53% accuracy. Sun *et al*. (2017) improved the traditional convolutional neural network model. The improved optimal model has an average recognition accuracy of 99.41% on a data set of 26 types of diseases in 14 plant species and reduces the amount of model training. Convergence time. Wallelign *et al*. (2018) proposed a CNN model to identify plant types by recognizing leaf images. The model uses the leaf data sets of 32 plants (1907 samples) and 15 plants (1125 samples) and respectively Achieved 97.24 and 99.11% classification accuracy. One of the advantages of using CNN is that it can directly act on the original data without using manually-made features. Ghosal *et al*. (2018)

established a classification framework based on a Deep Convolutional Neural Network (DCNN) and the classification accuracy rate on the soybean leaf disease data set (about 60,000 images divided into 9 types) was 94.13%.

The above models all use a large number of training data sets. The excellent performance of CNN requires the collection of a large number of data samples to train the network and the advanced computing power provided by the Graphics Processing Unit (GPU). However, in the field of agricultural image recognition, it collects image data related to plant diseases. It is time-consuming and expensive, requiring people from different fields to collaborate at different stages. There are some public data sets on the Internet, but they are difficult to apply to specific problems. Therefore, insufficient training data is the main factor hindering the further improvement of leaf disease recognition accuracy. Liu *et al*. (2019) used a lightweight neural network structure to train on the Plant Village dataset (a total of 38 categories of 26 diseases) and obtained a recognition accuracy of 95%. However, the highest recognition accuracy on their self-built grape disease leaf data set is only 88.06%. Liao *et al*. (2018) used a new method of grayscale transformation and mixed cutting of multiple operators to segment the foreground and background of the plant image, that is, remove the background and only retain the plant image to improve the performance of the convolutional neural network. The recognition accuracy of the method on the plant disease image data set containing 19000 background segmentation is 98.44%. Even if the method performs background segmentation preprocessing operations on the image, a large amount of data is still needed to train the classification model to obtain higher recognition. Accuracy, Perez and Wang (2017) discussed and compared a variety of data enhancement solutions to solve the problem of image classification. However, traditional data enhancement techniques (such as cropping, rotation, translation, and zooming) have less diversity and changes, that is, they cannot obtain other feature information. Therefore, the use of synthetic data can obtain more variability features and further amplify the data set, improving the CNN model training process to improve the classification accuracy.

Goodfellow *et al.* (2014) proposed a Generative Adversarial Networks (GANs) framework, which uses two different networks to generate synthetic sample images with the same feature distribution as the training set. Han and Guo (2022) proposed GAN-PCL based on data augmentation and sequence feature fusion with an adversarial neural network GAN, which effectively solves prediction accuracy but is still limited due to scarce and severely unbalanced samples. This research proposes a data enhancement method based on a generative adversarial network, which synthesizes high-quality soybean leaf disease images for the expansion of the soybean leaf disease data set, overcomes the over-fitting problem of the classification model, and can provide enough for training the classification model Image of soybean leaf disease with high perceptual quality. This method is based on CycleGAN (Zhu *et al*., 2017). The discriminator model uses a dense connection strategy (Wang *et al*., 2021). The dense connection strategy has a regularization effect, which reduces the overfitting problem of insufficient training set data, so that there is no pairing in the case of training data, capturing the special features of one type of leaf image collection and transferring these features to another type of leaf image collection. The main workflow of this research is shown in Fig. 1. Firstly, a generative adversarial network is used to synthesize high-perceptual quality soybean leaf disease image samples from a small number of original images, and then the generated samples are sent to the soybean leaf disease classifier model (CNN) For training and finally compare the performance of different data enhancement schemes on the soybean leaf disease classification model. The use of generation adversarial network to generate images to avoid the difficulty of soybean image data collection and the existence of complex noise interference problems, traditional data enhancement is to rotate, flip, and other fixed transformations of the image, these methods often do not change the image characteristics significantly, especially for leaves with fewer disease spots and the model is difficult to distinguish what kind of disease it belongs to and the data enhancement method in this research is different from these fixed methods, using the conversion between images to images, so that these images have more obvious disease spot characteristics, classification models are easier to extract image features, to accurately classify various types of soybean diseases. According to the survey, this is the first study to use a generative adversarial network to enhance the data set to improve the classification performance of soybean leaf disease.

The main contributions of this research are as follows:

1. This research proposes a data enhancement method to generate a confrontation network. Converting images of healthy soybean leaves and diseased leaves can expand the soybean leaf disease data set and overcome the problem of difficult soybean leaf disease data collection
2. The synthetic data in this study has more variability features, which can improve the classification accuracy of the CNN model and overcome the overfitting problem of the classification model
3. This study uses a limited soybean leaf disease image data set to verify the effectiveness of the proposed method. The results show that: (i) Using a synthetic data set, the CNN-based soybean leaf disease classification model has high accuracy and excellent performance. Compared with the original data set; (ii) Compared with the traditional data enhancement method, the data set augmented with synthetic data performs better in the soybean leaf disease classification model

DOI: 10.3844/ajbbsp.2022.417.427

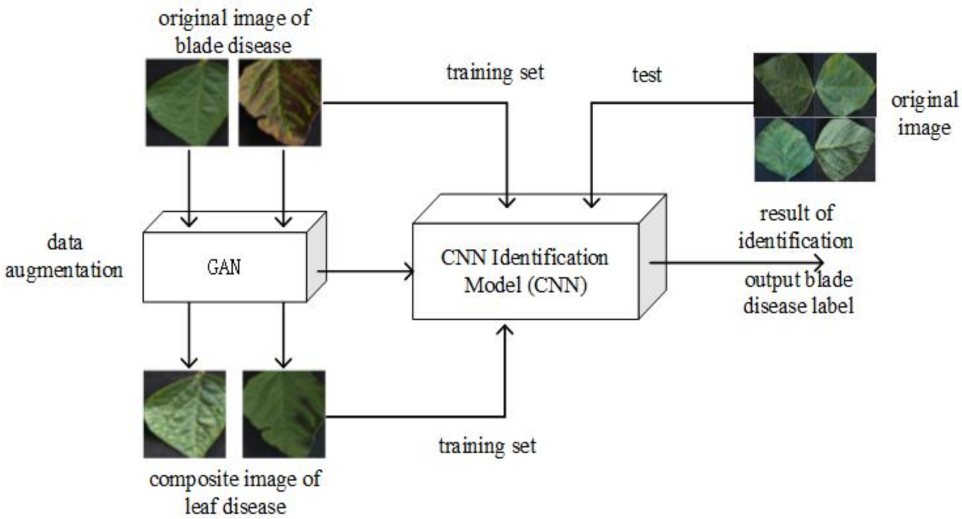


Fig. 1: Soybean leaf disease identification process

Materials and Methods

Experimental Materials

The soybean leaf disease data set used in this study contains 9 different soybean leaf diseases (including health): Bacterial leaf spot, brown spot, gray spot, health, herbicide damage, iron deficiency chlorosis, lack of Potassium, bacterial keratoderma and Sudden Death Syndrome (SDS). These data come from the soybean leaf disease data set provided by Ghosal *et al*. (2018). This data set uses a standard camera to collect labeled data from the field by strict imaging protocols and is officially diagnosed by a professional plant pathologist to mark the disease. Type-making data set labels, including biological (such as fungal and bacterial leaf diseases) and non-biological (such as nutritional deficiencies and chemical damage) of 8 leaf disease symptoms. This study selects 6576 images from the data set for the experiment, where the leaf image is an RGB image with a resolution of 64*64. These data were collected from the experimental base in Northeast China. The data were collected and produced in three time periods: Morning, noon, and afternoon. The data set (leaf disease symptom type and its sample number) is shown in Table 1 and some samples are shown in Fig. 2.

Traditional Data Enhancement

Data enhancement has been proven to be an implicit regularization strategy and a key technology to improve generalization ability (Lim *et al*., 2019). Data enhancement usually uses methods such as random cropping, flipping, rotation, scaling, and color transformation as benchmark techniques to increase the diversity of training data and improve the image recognition task's performance. Graphics Processing Unit (GPU) is good at matrix and vector operations of deep learning to improve the efficiency of image recognition tasks, but data preprocessing and data enhancement (image expansion) are usually completed on the CPU. In recent years, the performance and storage capacity of GPU hardware has steadily improved and the use of GPU can improve the efficiency of data preprocessing and data enhancement. Therefore, the control experiment of this study uses the kornia computer vision library proposed by Buslaev *et al*. (2020), which has a fast and flexible image enhancement module, which is based on the rapid realization of a large number of various image transformation operations and is also an Easy-to-use package module and can perform data enhancement operations on the GPU. The GANs data enhancement method used in this article and the traditional data enhancement method provided bykornia are trained on the same CNN classification model.

Table 1: Image data set of soybean leaf disease symptoms

Serial number	Type of soybean leaf disease	Sample size
1	Spot disease	623
2	Angle of spot corrosion	670
3	Brown spot	679
4	Herbicide hazard	658
5	Gray spot disease	642
6	Health	1302
7	Iron deficiency chlorosis	639
8	Potassium deficiency	704
9	SDS	659
-	Total	6576

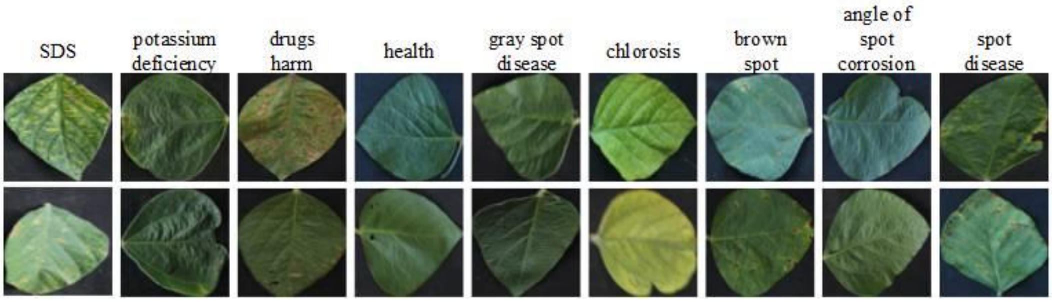


Fig. 2: Sample map of soybean leaf disease data set

GANs Data Enhancement

The basic principle of Generative Adversarial Networks (GANs) (Goodfellow *et al*., 2014) is to obtain the probability distribution of the output of the generator P_G and make it as similar as possible to the probability distribution of the initial P_{data} set, where the generator G maps the input data to the target data Probability distribution, the discriminator D judges the similarity between the generated data and the real data. GANs adopt the idea of a zero-sum game in game theory, that is, the constant game between generator G and discriminator D enables generator G to learn the distribution of real data. GANs are very effective in image synthesis tasks, such as drawing to image, image super-resolution, text to image, and image translation (Creswell *et al*., 2018).

This research is based on GANs for image translation and image samples are generated from training data, that is, healthy leaf images and leaf disease symptom images are mutually converted. The purpose is to add synthetic data generated by GANs based on the original data to enrich the training data set to improve the accuracy rate of the leaf disease recognition system. The GANs method used in this study is based on CycleGAN (Zhu *et al*., 2017) and its discriminator model uses a dense connection strategy (Wang *et al*., 2021). Image translation can be described as translating an image from one form of x to another form of y in a given scene, for example, from grayscale to color, from images to semantic labels, edge maps, etc., (Isola *et al*., 2017). CycleGAN is an algorithm that can learn to convert between different training domains without pairing input/output examples. It performs well in the field of image translation, such as style transfer, target deformation, season change, photo enhancement, etc.

The framework of the data enhancement method based on Generative Adversarial Networks (GANs) proposed in this research is shown in Fig. 3.

Among them, A and B are two different areas, that is, the real image data of the A area and the B area are different leaf disease types. G_{AB} and G_{BA} is a generator that translates an image from one area to another. For example, G_{AB} translates an image of area A into an image of area B. D_A and D_B are the discriminators of area A and area B respectively, which are used to judge whether the image belongs to this area. The image data circulates between the A area and the B area. Among them, the generator G_{AB} and G_{BA} share the weight, and the discriminator D_A and D_B share the weight. The goal of this method is to train $G_{AB}(G_{BA})$ so that the generator D_B (D_A) cannot distinguish the synthesized images generated by G_{AB} (G_{BA}). Therefore, the adversarial loss needs to be calculated to represent the performance of the generator and discriminator. Among them, formula (1) is the anti-loss function of A->B, and formula (2) is the anti-loss function of B->A.

However, if the model is only trained to reduce the adversarial loss, the model collapse problem will occur and the optimization will fail. Therefore, CycleGAN introduces a cycle consistency loss, so that each image a in area A can be converted back to area A to get the original image a, that is $G_{AB}(G_{BA}(b)) \approx b$. The cyclic consistency loss function is shown in formula (3).

Therefore, the overall loss function of this method is shown in formula (4), in which the relative weight of the two regions is controlled by λ, and the objective function of this method is shown in formula (5):

$$\begin{aligned} L_{GAN}\left(G_{AB}, D_B, A, B\right) &= E_{b:p_{data}^{(b)}}\left[\log D_B(b)\right] \\ &+ E_{a:p_{data}(a)}\left[\log\left(1 - D_B\left(G_{AB}(a)\right)\right)\right] \end{aligned} \tag{1}$$

$$\begin{aligned} L_{GAN}\left(G_{BA}, D_A, B, A\right) &= E_{a:p_{data}^{(b)}}\left[\log D_A(a)\right] \\ &+ E_{b:p_{data}(b)}\left[\log\left(1 - D_A\left(G_{BA}(b)\right)\right)\right] \end{aligned} \tag{2}$$

$$\begin{aligned} Lcycle\left(G_{BA}, G_{BA}\right) &= E_{a:p_{data}(a)}\left[PG_{BA}\left(G_{AB}(a)\right) - aP_1\right] \\ &+ E_{b:p_{data}(b)}\left[PG_{AB}\left(G_{AB}(b)\right) - bP_1\right] \end{aligned} \tag{3}$$

$$\begin{aligned} L\left(G_{AB}, G_{BA}, D_A, D_B\right) &= L_{GAN}\left(G_{AB}, D_B, A, B\right) \\ &+ L_{GAN}\left(G_{BA}, D_A, B, A\right) + \lambda L_{cycle}\left(G_{BA}, G_{BA}\right) \end{aligned} \tag{4}$$

$$G_{AB}^{*}, G_{BA}^{*} = \arg\min_{G_{AB}, G_{BA}} \arg\max_{D_A, D_B} L\left(G_{AB}, G_{BA}, D_A, D_B\right) \tag{5}$$

CycleGAN's generator uses U-Net (Ronneberger *et al.*, 2015) and the discriminator uses the PatchGANs structure (Isola *et al.*, 2017) in pix2pix. The output of PatchGANs is a matrix and the final result is averaged. Therefore, PatchGANs enable the model to pay more attention to the detailed features of the image. This research uses a dense connection strategy for the PatchGANs structure. Each dense block is shown in Fig. 4. This strategy reuses a large number of features so that a small number of convolution kernels can generate a large number of features while reducing the size and calculation of the final model quantity.

CNN Classification Model

The classification model of this study uses the pre-trained DenseNet-121 (Huang *et al.*, 2017) model. DenseNet is a Convolutional Neural Network (CNN) that uses a dense connection strategy. The network performs well in the field of image recognition. The core idea of the dense connection strategy is to establish different connection relationships between the layers and reuse a large number of features to alleviate the problem that the gradient disappears and the model is difficult to continue to optimize. At the same time, the parameter training efficiency of the model is optimized to reduce overfitting. The recognition and classification of soybean leaf diseases are achieved by using a pre-trained classification model based on DenseNet-121. The model is composed of 3 dense blocks and 2 transition layers. The image input of 224*224*3 pixels is used to identify soybean leaf diseases.

Transfer learning is to use a deep neural network pre-trained on a super-large-scale data set to solve specific model training tasks with limited data (Yang and Sun, 2019; Basha *et al.*, 2021). This research is based on transfer learning and is pre-trained on a large-scale data set ImageNet (Deng *et al.*, 2009) DenseNet- the 121 networks are fine-tuned and the classification effect is determined by the accuracy of the verification set and the training set, where the accuracy is defined as formula (6):

$$accuracy(x) = \frac{T_{px}}{T_{px} + T_{nx}} \tag{6}$$

Among them, x represents the type of leaf disease, *and* T_{px} and T_{nx} respectively represent the number of successful and wrong times of identifying leaf disease in the entire system, that is, the average accuracy rate of classification of each leaf disease in the entire system.

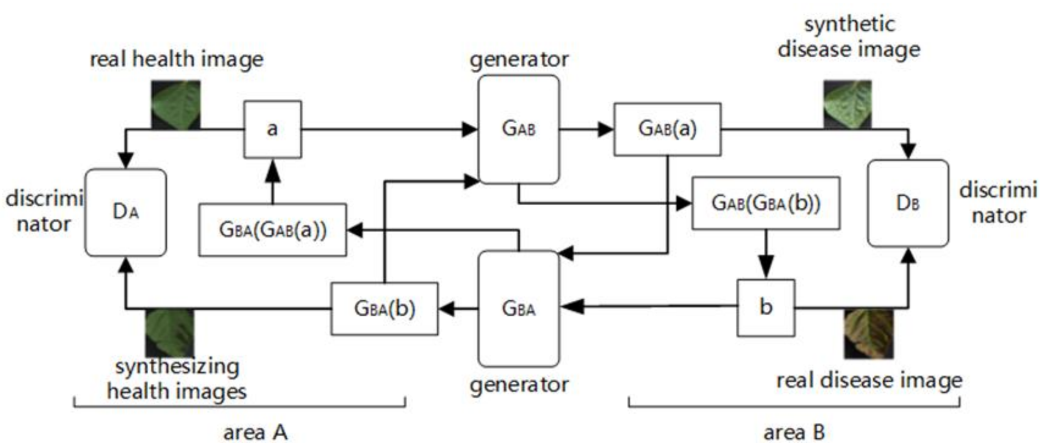


Fig. 3: Generative confrontation network model based on generative confrontation networks

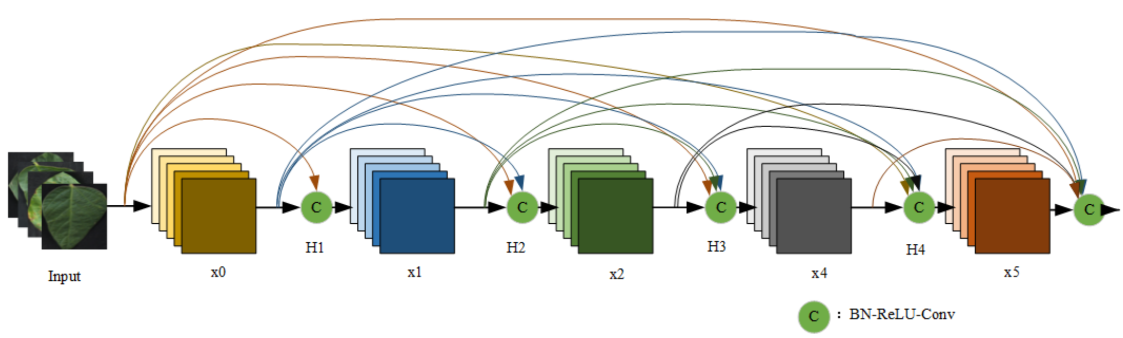


Fig. 4: Dense connection strategy

Results and Discussion

Experimental Platform

The hardware configuration of the experimental environment of this research includes CPU: i7-6700K, graphics card GPU: NVIDIA GTX 1080ti, and 16G memory. This experiment is implemented in Python language on the PyTorch (Stooke and Abbeel, 2019) and Fast.ai (Howard and Gugger, 2020) deep learning framework.

GANs Data Enhancement

In this study, to obtain enough soybean leaf disease data sets to improve the performance of the CNN classification model, the CycleGAN framework was used as a data enhancement strategy to generate new synthetic soybean leaf disease images. In this study, based on CycleGAN, the discriminator model adopts a dense connection strategy to ensure that the leaf disease image and the healthy leaf image retain sufficient symptom features during mutual translation. The loss curve of the data enhancement strategy proposed in this study with the number of iterations is shown in Fig. 5. The sub-figures (i) and (ii) are the loss curves of 5000 and 6000 iterations, respectively. The abscissa is the number of iterations and the ordinate is Training loss. Model training reduces the loss by increasing the number of iterations. Each time the loss is slightly changed to improve the model, the loss drops quickly at the beginning of training and then decreases steadily. After 6000 times, the training loss does not change significantly and the training is terminated.

Figure 6 is an example of the results of the mutual conversion of different leaf images in area A and area B, where rows (a) and (c) are the original images and rows (b) and (d) are the corresponding composite images. The results show that the composite image can learn from the original image set the features, that is, to capture and learn the characteristics of health or leaf disease without changing the shape of the leaf. This experiment uses healthy leaf images and different leaf disease images for conversion and the converted synthetic images are added to the training set of the classification model. The verification set uses all the original real images and does not include the generated images to verify the image generation of the model.

Classification of Soybean Leaf Disease

The recognition and classification of soybean leaf diseases are achieved by using a pre-trained classification model based on DenseNet-121. The model consists of 3 dense blocks and 2 transition layers. The image is reset to 224*224*3 pixels as input to identify soybean leaf diseases. The simple DensenNet network architecture is shown in Fig. 7. The dense block includes 12-core filter, 1 1*1 convolution (used to reduce the number of input feature maps), 1 3*3 convolution, and then after the dense block is output, the feature maps are batched One, ReLU, 1*1 convolution, and average pooling layers and finally, after 3 dense blocks, similar to traditional CNN, they pass through the average pooling layer and softmax classifier in turn to output the classification results. The following describes the comparison of the results of the pre-trained DenseNet-121 model with the ResNet-34 (Ronneberger *et al*., 2015) and VGG-16 models. Each model has three results: Acting on training set A, training set B, and training set C, respectively.

Dataset Partitioning

From the collected 6576 original data sets, 20% of each type of leaf disease (including health) (that is, a total of 1315 images) is randomly selected as the verification set and 80% as the training set. Among them, the generating model uses the healthy leaf images of the training set and the images of different leaf diseases to convert and translate each other. Only 12.5% of healthy leaves converted from different leaf diseases are retained. The purpose is to balance the number of synthesized healthy leaf images with the number of synthesized leaf disease images because many studies have shown that imbalance in the number of samples will reduce the accuracy of the classification model (Johnson and Khoshgoftaar, 2019). The partition of the data set is shown in Table 2. Among them, training set A only contains original real images, training set B contains original real images and images generated by traditional data enhancement methods (using kornia vision library for random cropping, flipping, rotation, zooming, etc.), training set C contains original real images and Synthetic images generated by GANs.

Classification Model Results

Three training sets are used in the pre-trained DenseNet-121 model and the performance of the popular ResNet-34 and VGG-16 network models in the image classification field is evaluated to verify the performance of the proposed method. This study uses the cyclic learning rate method of (Smith, 2017) to set the learning rate. This method does not monotonically decrease the learning rate with the increase of the number of iterations but makes the learning rate cyclically change between reasonable boundary values. To eliminate the need to find the global best learning rate and reduce the number of iterations. The number of cycles (epochs) of the three classification models is fixed at 100 and the image batch size is fixed at 64. The ResNet-34 and VGG-16 models use the Adam (Kingma and Ba, 2014) optimizer, which performs on these two models optimally. DenseNet-121 uses Stochastic Gradient Descent (SGD) optimizer. The hyperparameter configuration choices for the CNN structure are shown in Table 3.

The accuracy learning curve in the model training using the hyperparameter configuration selected in Table 3 is shown in Fig. 8. Figure 8(i) shows the curve of the verification accuracy of the three models on the training set C increasing with the number of epochs. The accuracy of each classification model increased significantly in the first 10 epochs (about 800 iterations) and then steadily oscillated. Stop training until the 100^{th} epoch.

The DenseNet-121 classification model based on pre-training has the highest verification accuracy, with an average accuracy of 95.89%, followed by ResNet-34 (94.68%) and finally VGG-16 (94.5%). Figure 8(ii) shows the curve of the verification accuracy of different data sets on the DenseNet-121 classification model as the number of epochs increases. As the accuracy continues to increase, the accuracy of the data set C has the smallest oscillation and the fastest convergence. The initial and final results show that in the classification model, the verification accuracy of training set C containing synthetic images is higher than that of training set A (original data set) and training set B (traditional data enhancement).

Figure 9 is the confusion matrix of the Densenet-121 classification model on the training set C. The main identification errors of the model are concentrated in bacterial blight and bacterial pustule. These two diseases are Common bacterial leaf diseases in soybean crops that have similar symptoms and are difficult to identify with the naked eye. In the results of Ghosal *et al*. (2018), the identification error of bacterial leaf spot and bacterial angular leaf spot rot is about 20% and the error in this study is only 11.65%. The data enhancement method based on GANs in this study uses 10% of the data set provided by Ghosal *et al*. (2018). After the enhancement, the size of the training set C is about 20% of the training set, and 95.89% is achieved on the DenseNet-121 classification model. The accuracy rate is better than the 94.13% accuracy rate on the Ghosal *et al*. (2018) deep machine vision framework.

Table 2: Soybean leaf disease image recognition model dataset

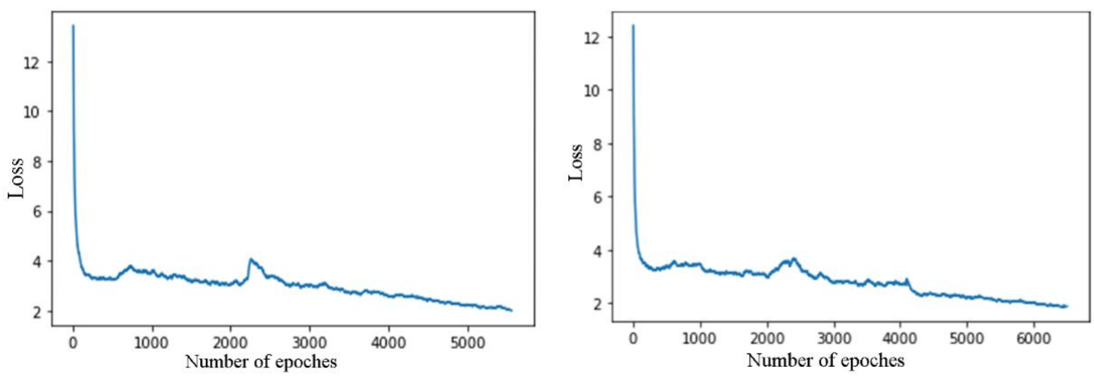
	Number of original images	Number of generated images	Number of generated images (GANs)
Training set A	5441	-	-
Training set B	5441	5441	-
Training set C	5441	-	5441
Validation set	1315	-	-

Table 3: Hyper-parameters configuration of CNN structure

	ResNet-34	VGG-16	DenseNet-121
Batch size	64.000	64.000	64.00
Epoch	150.000	150.000	150.00
Learning rate	0.001	0.001	0.01
Optimizer	Adam	Adam	SGD

Table 4: Verification accuracy of classification models

	Training set A	Training set B	Training set C
ResNet-34	92.70%	93.53%	94.68%
VGG-16	90.34%	94.29%	94.52%
DenseNet-121	92.67%	94.75%	95.89%



(i) (ii)

Fig. 5: CycleGAN training loss curve with the number of iterations

DOI: 10.3844/ajbbsp.2022.417.427

Table 4 shows the final performance of the three classification models. The results show that DenseNet-121 achieves the best results for different training sets. According to the experimental results, the GANs-based data enhancement method used in this study can improve the accuracy of the soybean leaf disease classification model. On the DenseNet-121 model, comparing the use and non-use of data enhancement methods to verify the impact of the proposed method on the model performance, the results show that the accuracy of the model trained on the synthetic image data set of the GANs method is relative to the training set A and training set B They are 2.66 and 1.14% higher respectively. The GANs image generation method can increase the variability and diversity of image data features in image conversion and translation and adding GANs method synthetic data to the training set can improve the performance of the classification model.

Fig. 6: CycleGAN data enhanced soybean leaf image control example

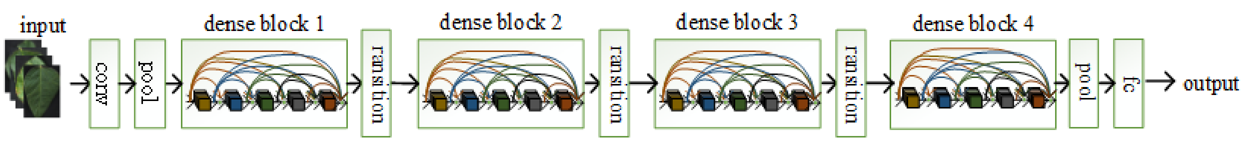


Fig. 7: Densennet-12 network architecture

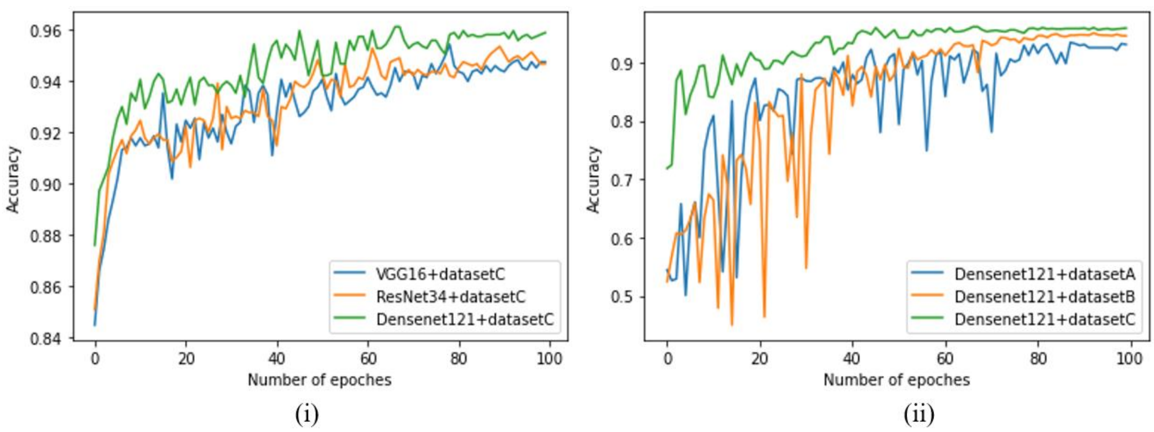


Fig. 8: Accuracy learning curve during training

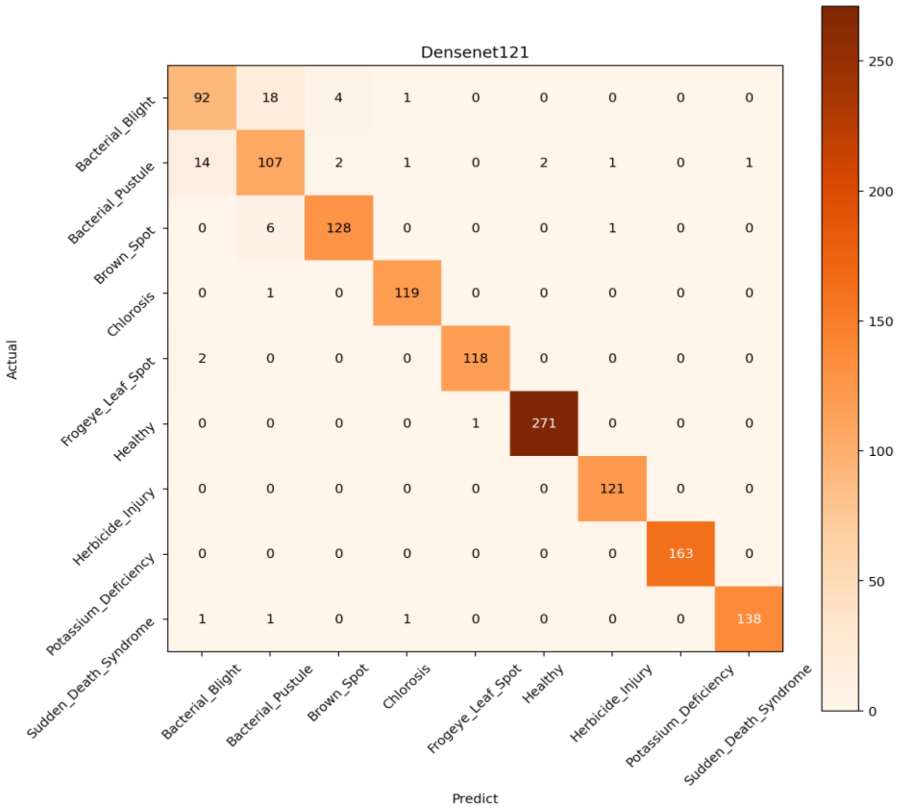


Fig. 9: Confusion matrix of Densenet-121 classification model on training set C

Conclusion

With the development of computer vision, many researchers use image processing, machine learning, and other methods to analyze plant phenotypic characteristics. Although Plant Village and others provide some plant image data sets, the plant data set size is still the main challenge of machine learning algorithms, especially for algorithms that use a large number of hyperparameters (for example, deep neural networks).

In the field of plant disease recognition, it is difficult to obtain a sufficient amount of image data for CNN classification model training. However, small data sizes and unbalanced data will cause over-fitting problems and reduce the performance of the classification model. To alleviate the lack of training data in soybean leaf disease recognition, researchers have widely used traditional image enhancement methods (image rotation, flipping, zooming, etc.), to augment the data set. However, traditional image enhancement methods cannot increase the diversity and variability of image features. Therefore, this study uses a GANs method based on CycleGAN to create artificially synthesized soybean leaf disease images, amplify the soybean leaf disease data set, and further improve the performance of soybean leaf disease recognition using deep convolutional neural networks. The size of the soybean leaf disease data set used in this study can reflect the size of the data set available in most of the fields of plant disease and insect image diagnosis and recognition in the field of plant disease research, that is, the data set of plant disease research can be applied in the method proposed in this study. This method uses the same rules as the classification model for data enhancement, which can effectively fill the gaps in the discrete data distribution of the training image to increase the diversity and variability of the generated image features. The GANs model optimized by the use of anti-loss and cyclic consistency loss can be used to generate various types of soybean leaf disease images. At the same time, the discriminator of the model adopts a dense connection strategy to reduce the scale and computational complexity of the final model. This study is based on migration learning to fine-tune the pre-trained DenseNet-121 classification model to evaluate the performance of the GANs method, use synthetic images to amplify the soybean leaf disease data set, and compare it with the classic data enhancement method on the classification model. The results show that the classification accuracy (+2.66%) of the synthetic image data based on the GANs data enhancement method is significantly improved, while the classical data enhancement method is improved by 1.52%. The data used in this study comes from the data set provided by Ghosal. The training set size (10 082) is about 20% of the Ghosal training set (about 50,000 sheets). The accuracy of the DenseNet-121 classification model is relatively the depth vision framework of Ghosal has increased by 1.76%. Therefore, the data enhancement method proposed in this study can generate approximate real leaf image data, which can be used to (1) Provide a larger-scale data set for deep neural networks and the data set has obvious distinguishing characteristics, which can improve the recognition model Performance; (2) Generate leaf image data with high perceptual quality, which can reduce the cost of collecting leaf data.

The synthetic leaf disease generated by the data enhancement method based on the generative adversarial network proposed in this study has different leaf disease characteristics and visualization effects and can be transplanted to other computer-aided algorithms. The experimental results of the classification model show that the use of synthetic data to augment the data set can effectively improve the performance of the classification model. In addition, due to the instability and data dependence of the training process of the generative adversarial network, there is a small number of image noise in the generated image, and does not conform to the leaf characteristics, which is a limitation at present, so it is necessary to further optimize CycleGAN to reduce noise interference. Moreover, the current experiment only studies soybean leaf diseases and the scope of research is relatively small, this research plans to extend the method to more agricultural diseases and insect pests, these fields can improve the performance of the classification model by synthesizing plant disease images. In addition, GANs are used to synthesize data on large-scale data sets (such as medical images and astronomical images) with scarce and unbalanced data to test the generalization ability of the proposed method.

Acknowledgment

This Project "Research on the interdisciplinary theory, method and application of plant phenotype data management" supported by Shandong Provincial Natural Science Foundation (ZR2022MG047), the National Social Science Fund of China (18BTQ062), Scientific Research Foundation for Ph.D. of Shandong University of Technology (420052).

Funding Information

This Project "Research on the interdisciplinary theory, method and application of plant phenotype data management" supported by Shandong Provincial Natural Science Foundation (ZR2022MG047), the National Social Science Fund of China (18BTQ062), Scientific Research Foundation for Ph.D. of Shandong University of Technology (420052).

Author's Contributions

Xiao Yu and Lina Lu: Designed and performed the experiments, and work.

Cong Chen: Participated to collect the materials related to the experiment. Designed the experiments and revised the manuscript.

Cong Chen: Participated to collect the materials related to the experiment.

Ethics

The authors declare their responsibility for any ethical issues that may arise after the publication of this manuscript.

References

Basha, S. S., Vinakota, S. K., Pulabaigari, V., Mukherjee, S., & Dubey, S. R. (2021). Autotune: Automatically tuning convolutional neural networks for improved transfer learning. *Neural Networks*, *133*, 112-122. https://doi.org/10.1016/j.neunet.2020.10.009

Buslaev, A., Iglovikov, V. I., Khvedchenya, E., Parinov, A., Druzhinin, M., & Kalinin, A. A. (2020). Albumentations: Fast and flexible image augmentations. *Information*, *11*(2), 125. https://doi.org/10.3390/info11020125

Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B., & Bharath, A. A. (2018). Generative adversarial networks: An overview. *IEEE Signal Processing Magazine*, *35*(1), 53-65. https://doi.org/10.1109/MSP.2017.2765202

Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009, June). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition* (pp. 248-255). IEEE. https://doi.org/10.1109/CVPR.2009.5206848

Feng, W. K., & Hua, H. X. (2021). Research of Image Recognition of Plant Diseases and Pests Based on Deep Learning. *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)*, *15*(4), 1-21. https://doi.org/10.4018/IJCINI.295810

Ferentinos, K. P. (2018). Deep learning models for plant disease detection and diagnosis. *Computers and Electronics in Agriculture*, *145*, 311-318. https://doi.org/10.1016/j.compag.2018.01.009

Ghosal, S., Blystone, D., Singh, A. K., Ganapathysubramanian, B., Singh, A., & Sarkar, S. (2018). An explainable deep machine vision framework for plant stress phenotyping. *Proceedings of the National Academy of Sciences*, *115*(18), 4613-4618. https://doi.org/10.1073/pnas.171699911

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). Advances in neural information processing systems. *Curran Associates, Inc*, *27*, 2672-2680.

Han, L., & Guo, Z. (2022, May). GAN-PCL: An Efficient Protein Subchloroplast Site Predictor with GAN-based Data Augmented and Feature Fusion. In *Proceedings of the 14th International Conference on Bioinformatics and Biomedical Technology* (pp. 72-78). https://doi.org/10.1145/3543377.3543388

Howard, J., & Gugger, S. (2020). Fastai: A layered API for deep learning. *Information*, *11*(2), 108. https://doi.org/10.3390/info11020108

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4700-4708).

Isola, P., Zhu, J. Y., Zhou, T., & Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1125-1134).

Johnson, J. M., & Khoshgoftaar, T. M. (2019). Survey on deep learning with class imbalance. *Journal of Big Data*, *6*(1), 1-54. https://doi.org/10.1186/s40537-019-0192-5

Kaur, S., Pandey, S., & Goel, S. (2019). Plants disease identification and classification through leaf images: A survey. *Archives of Computational Methods in Engineering*, *26*(2), 507-530. https://doi.org/10.1007/s11831-018-9255-6

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.https://doi.org/10.48550/arXiv.1412.6980

Liao, J. W., Cai, Y., Wang, Y. C., Zhang, Y. Q., Tan, Z. Y., Zhou, Y. T. (2018). Recognition Technology of Plant Disease Based on Convolution Neural Network. *Modern Computer*, (19), 43-48+53.

Lim, S., Kim, I., Kim, T., Kim, C., & Kim, S. (2019). Fast autoaugment. *Advances in Neural Information Processing Systems*, *32*. https://proceedings.neurips.cc/paper/2019/hash/6add07cf50424b14fdf649da87843d01-Abstract.html

Liu, Y., Feng, Q., & Wang, S. Z. (2019). Plant disease identification method based on lightweight CNN and mobile application. *Trans. Chin. Soc. Agric. Eng*, *35*, 194-204.

Perez, L., & Wang, J. (2017). The effectiveness of data augmentation in image classification using deep learning. *arXiv preprint arXiv:1712.04621*. https://doi.org/10.48550/arXiv.1712.04621

Ronneberger, O., Fischer, P., &Brox, T. (2015, October). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention* (pp. 234-241). Springer, Cham. https://doi.org/10.1007/978-3-319-24574-4_28

Smith, L. N. (2017, March). Cyclical learning rates for training neural networks. In *2017 IEEE winter conference on applications of computer vision (WACV)* (pp. 464-472). IEEE. https://doi.org/10.1109/WACV.2017.58

Stooke, A., & Abbeel, P. (2019). rlpyt: A research code base for deep reinforcement learning in pytorch. *arXiv preprint arXiv:1909.01500*. https://doi.org/10.48550/arXiv.1909.01500

Sun, J., Tan, W., Mao, H., Wu, X., Chen, Y., & Wang, L. (2017). Recognition of multiple plant leaf diseases based on improved convolutional neural network. *Transactions of the Chinese Society of Agricultural Engineering*, *33*(19), 209-215. https://www.ingentaconnect.com/content/tcsae/tcsae/2017/00000033/00000019/art00027

Wallelign, S., Polceanu, M., &Buche, C. (2018, May). Soybean plant disease identification using convolutional neural network. In *FLAIRS conference* (pp. 146-151).

Wang, T., Cheng, J., Liu, T., Zhao, K., Cheng, B., & Liu, Z. (2021, September). Visible Light Polarization Image Fusion based on Dense Connection Generative Adversarial Network. In *Journal of Physics: Conference Series* (Vol. 2035, No. 1, p. 012028). IOP Publishing. https://iopscience.iop.org/article/10.1088/1742-6596/2035/1/012028/meta

Yang, X., & Sun, M. (2019, April). A survey on deep learning in crop planting. In *IOP Conference Series: Materials Science and Engineering* (Vol. 490, No. 6, p. 062053). IOP Publishing. https://iopscience.iop.org/article/10.1088/1757-899X/490/6/062053/meta

Zhu, J. Y., Park, T., Isola, P., &Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision* (pp. 2223-2232).