Defect detection of solar cell based on data augmentation

Yunyan Wang, Shuai Luo, Huaxuan Wu

School of Electrical and Electronic Engineering (Hubei University of Technology)
Wuhan, China

a 510496148@qq.com, b 201910294@hbut.edu.cn, c 1436383435@qq.com

Abstract. In this paper, a true and false data fusion algorithm based on deep convolution confrontation generation network and random image Mosaic is proposed, which improves the training data volume by 800 times. At the same time, the network model is optimized with light weight to reduce model training parameters, storage space and prediction time. The experimental results show that compared with the original data set training and the model training using the traditional data enhancement algorithm, the test accuracy of the model training obtained by the fusion of real and false data is improved by 30% and 17% respectively, reaching nearly 77%. After the lightweight treatment, the model parameters were reduced to about 1/2 before the treatment, and the test time for each image was shortened from 57ms to 22ms. The research shows that the fusion algorithm can effectively help the image classification task with insufficient data to alleviate the problems of network overfitting and poor model performance. The lightweight optimization model can not only ensure the accuracy, but also compress the size of the model to speed up the testing speed. It is helpful to promote the application of intelligent algorithm in industrial production so as to save cost and improve production efficiency.

Keywords: Defect detection, Data augmentation, DCGAN, Image fusion, lightweight design.

1. Introduction

In recent years, with the continuous advancement of science and technology and rapid social and economic development, non-renewable energy sources such as oil, coal, and natural gas have been consumed in large quantities and are increasingly depleted. As a new renewable energy source, solar energy has no geographical restrictions and is one of the cleanest energy sources with rapid development. As the basic unit of photoelectric conversion, solar cells and their related manufacturing process level and product testing technology are all important factors that affect the power generation efficiency of solar cell modules and product life. Therefore, facing such a huge market demand and strict quality testing, the problem of solar cell defect detection is very important. Convolutional Neural Networks (CNN) has shown excellent results in many image processing tasks due to its large number of trainable parameters and rich expression capabilities. The defects of applying it to solar cells It is worth exploring in detection.
Recent studies have shown that the performance of visual tasks will increase logarithmically as the amount of training data increases. However, if there are not enough training samples, CNNs with many parameters will have the risk of overfitting, because they will remember the characteristics of the training image that cannot be generalized, and show poor results on the test image. Because collecting a large number of samples is time-consuming, labor-intensive, and costly; therefore, data enhancement methods are usually used. Through image flip, zoom and random cropping, the diversity of the image can be increased; using Dropout and regularization can also effectively prevent overfitting. With the continuous development of Generative Adversarial Networks (GAN) and its remarkable achievements in the field of image generation, it seems feasible to use GAN to generate more training data. However, studies have found that simply using GAN to generate data and training with the original data may result in lower model testing accuracy, and cannot solve the problem of overfitting caused by insufficient data, but this does not directly deny that GAN is in The role of data augmentation. Therefore, it is of great significance to explore the positive role of GAN in image processing tasks where training data is scarce. In addition, in order to make CNN achieve better performance, the number of network layers is constantly increasing. With the deepening of the network, the storage problem of the training model and the speed problem of the model’s prediction also follow. Only by solving these problems can CNN get out of the laboratory and be more widely used. Lightweight networks have been favored by more and more scholars in recent years. They mainly optimize network calculation methods to reduce network parameters without losing network performance. In order to better apply the solar cell defect detection model to practical engineering, the lightweight of the model is indispensable.

This paper combines DCGAN and image fusion technology to generate data that is conducive to training, and expand the training data; use a lightweight network to reduce the amount of network parameters, and detect solar cell defects with scarce data. Effectively overcome the model overfitting caused by insufficient training data; at the same time, the lightweight of the network optimizes the network structure, reduces the amount of training parameters, ensures the performance of the model, and shortens the prediction time; it has a wide range of practical engineering Application prospects.

2. Related work

2.1. Data enhancement

The purpose of data enhancement is to increase the diversity of training data and prevent overfitting. In AlexNet, the author increased the training data to 2048 times by randomly cropping and horizontally flipping the training images; random cropping can prevent CNN from overfitting specific features by changing the features of the image; Flip doubles the change of an image with a specific direction, such as the side view of a car. At the same time, Alex Net also performs PCA processing on the RGB data of the image, and performs a Gaussian perturbation with a standard deviation of 0.1 on the principal components, adds some noise to construct new samples, and reduces the error rate by 1%. At the same time, processing such as rotation, scaling, and adding random noise can also alleviate the problem of model overfitting by increasing training data. These traditional data enhancement techniques play an important role in the training of neural networks. However, as many new network structures are proposed, the number of parameters is increasing, and the risk of overfitting is also increasing. Therefore, data enhancement technology has attracted further attention. Figure 1 (a), (b), (c) and (d) show the original training data and its enhanced data after horizontal flipping, adding salt and pepper noise, and rotating 90° clockwise.
The above-mentioned traditional data enhancement methods simply enrich the training data set by micro-processing the original training data. These methods do not destroy the feature information of the image itself. Different from the above methods, there are some methods to get defective training data by destroying the original training images. For example, random erasing (Random Erasing) a rectangular area in the image, leaving the other parts intact, set this rectangular area to a uniform pure color value that has nothing to do with other areas of the image; it effectively solves the problem of existence in real scenes The occlusion problem. Cutout is similar to random erasure. Cutout is inspired by the Dropout layer. It randomly selects a fixed-size square area and fills it with all 0s, and there is a certain probability that the erasing rectangular area is not completely in the original image. As shown in Figure 1(e) and (f), ask the enhanced data processed by Random Erasing and Cutout respectively.

In addition, Mixed Sample Data Augmentation (MSDA) has been very popular recently. Its core idea is to randomly mix two training samples and their labels in a certain proportion; this mixing method can not only increase the number of samples. The diversity of the neural network and the linear relationship between the training samples can be used to improve the robustness of the model.

2.2. DCGAN
The generative confrontation network consists of a generator G and a discriminator D. The generator G generates pictures through input noise, and the discriminator D needs to classify the generated pictures and the real pictures; the two games each other until the discriminator D cannot distinguish correctly. The difference between the picture generated by the generator G and the real picture to produce excellent output. Both the discriminator and generator of the deep convolutional confrontation generation network use a convolutional neural network to replace the multi-layer perceptron in GAN, while using a global pooling layer to replace the fully connected layer, adding batch normalization, etc. The operation improves the instability of GAN, thereby improving the quality of the generated image. The DCGAN network structure is shown in Figure 2, where the input of the generated network G is a 100-dimensional uniformly distributed noise vector, and finally a 64×64×3 picture is output after multi-layer deconvolution; and the input of the discriminating network D is a real picture And the false pictures generated by the generating network G are subjected to supervised training, and the network structure judges that the structure of the network G is almost symmetrical.
The loss function of DCGAN is shown in formula (1), where $G(z)$ represents the picture generated by the generation network, $D(x)$ and $D(G(z))$ represent the prediction of the real picture and the generated picture by the discrimination network, and the value is between 0 and 1. In between, 0 means that the input picture is judged as a fake picture by the D network, and 1 means it is judged as a real picture. The training goal is to generate the pictures generated by the network $G$, $G(z)$ that are real enough that the discrimination network D cannot correctly distinguish between real pictures and generated pictures; at the same time, the discrimination network D needs to be continuously trained to be able to distinguish between real pictures and fake pictures, both Gambling with each other. Therefore, when training the discriminant network D, the $D(x)$ needs to be close to 1 to maximize $\log D(x)$; and when training the generation network G, the purpose is to confuse the discriminant network D so that it will misjudge the $G(z)$ of the generated picture as a real picture. Therefore, $D(G(z))$ needs to be close to 1, and $\log(1 - D(G(z)))$ is minimized.

$$\min_{G} \max_{D} V(D,G) = E_{x \sim p_{data}(x)}[\log D(x)] + E_{z \sim p_{z}(z)}[\log(1 - D(G(z)))]$$

(1)

2.3. Lightweight network

At present, most deep learning models require high-performance computers and rely on large amounts of data for training. In order to make the deep learning model suitable for mobile devices, in recent years, deep learning compression and acceleration have been greatly developed. Reducing the size and complexity of the model is the main way to realize the lightweight of the deep learning model. Separable convolution is a method of compressing algorithm layers to accelerate the speed of learning network. The use of separable convolution to realize the lightweight of deep learning network and improve the running speed of learning network. The separable convolution has spatial separable convolution and depth separable convolution.

Spatially separable convolution simply divides the convolution kernel into two smaller convolution kernels. For example, a 5x5 convolution kernel is divided into 5x1 and 1x5 convolution kernels, by reducing the number of multiplications, reducing the amount of calculation, reducing the computational complexity, and improving the running speed of the learning network.

Depth separable convolution is different from space. It involves not only the spatial dimension, but also the depth dimension. The depth separable convolution divides the convolution kernel into two separate convolution kernels, and uses the two convolution kernels to perform deep convolution and point-wise convolution respectively.

The depth separable convolution formula is shown in formula(2):

$$G_{k,l,m} = \sum_{i,j} K_{i,j,m} * F_{k+i-1,l+j-1,m}$$

(2)

Where $G$ is the output feature map, $K$ is the convolution kernel, $F$ is the input feature, $i, j$ are the pixel positions of the feature map, $k, l$ are the resolution of the output feature map; $m$ is the number of channels. The depth separable convolution uses channel-wise convolution and point-wise convolution.
to extract the features of the corresponding feature map, and the point-by-point convolution method is used to achieve feature fusion after the deep separable convolution.

3. Related improvements

3.1. Random fusion of true and false pictures

Aiming at the problems of insufficient training data leading to model overfitting and the use of DCGAN to generate image training to reduce the accuracy of the model, this paper proposes a new data enhancement method: random fusion of true and false images, in simple terms, is to generate original training images and DCGAN Random fusion of the pictures of, in order to expand the training data set, and to overcome the adverse effects of directly using the generated pictures for training on the model, the principle of the method is shown in Figure 3, mainly includes three steps. First, adjust all the real defect pictures and the defect pictures generated by DCGAN to the same size, which is set to 224×224 in this article; then select one real picture and one generated picture of the same defect type, according to the size of the picture The coordinates of the generated random point are \((x, y)\). According to the random point, the two pictures are cut into four parts; finally, the cut pictures are stitched according to the method of Figure 3, that is, the 1 and 3 blocks of the real picture are combined with the 2 of the generated picture, 4 pieces of stitching can get a new training picture, its label is consistent with the real picture and the label of the generated picture.

In short, for each real picture, select all generated pictures of the same type of defects in turn and perform random cropping and fusion processing; That is to say, if we have m real broken defect pictures and use DCGAN to generate n defective pictures of the same type, then m×n new training pictures of the same type can be obtained through the random fusion network of true and false pictures. The data expanded by this method contains part of the defect features in the original data or a combination of multiple defect features on the one hand, and on the other hand, through fusion with DCGAN generated images, the original style of the training data is retained, and Random Erasing and Random Erasing are reduced. Algorithms such as Cutout cause image distortion to negatively affect the model. Effectively expand the training data set and reduce the risk of model overfitting.

![Figure 3. Network diagram of random fusion of true and false pictures](image)

From the perspective of the mathematical formula, it can be understood that according to the random point coordinates \((x, y)\), a filter of the same size as the input picture is set \(w\), as shown in Figure 3, the parts 1 and 3 of the \(w\) are set to 1, and the parts 2 and 4 Set it to 0; therefore, the finally obtained fusion picture \(f\) is shown in formula (3):

\[
f(i,j) = t(i,j) \times w(i,j) + g(i,j) \times [1 - w(i,j)]
\]

Publicity \(t(i,j)\) indicates the gray value of the point in the real picture, \(g(i,j)\) indicates the gray value of the corresponding point in the generated picture, \(w(i,j)\) is the filter weight in the fusion network.
proposed in this article, \( f(i,j) \) indicates the fusion network output, that is, the gray value of the corresponding point of the image is obtained after image fusion.

The image fusion method introduced in this section is also applicable to the fusion between real images. The original training set with only \( p \) images can be used as \( p^2 \) images.

### 3.2. Lightweight network design

This paper selects VGG16 as the basic network, improves it in light weight, reduces the parameter amount of the network model, saves storage space, shortens the defect detection time, and promotes the application of related algorithms in actual industrial production. The VGG16 structure is shown in Figure 4. It contains 13 convolutional layers, 3 fully connected layers, and 5 pooling layers. The pooling layer does not involve weights, and there are 16 weighting layers; although the structure is relatively simple, it contains the number of weights is quite large. According to the algorithm requirements of this article, the input layer and the fully connected layer need to be adjusted first. Our training data is a \( 224 \times 224 \) single-channel image. There are three types of defects and four types of non-defects; therefore, the volume of the first convolutional layer The size of the product core is changed from the original \( 3 \times 3 \times 3 \) to \( 3 \times 3 \times 1 \), and the number of nodes in the last fully connected layer is adjusted from 1000 to 4 to represent four types, and the number of nodes in the first and second fully connected layers is changed to 1024 at the same time, \( 512 \) to ensure smoother propagation between fully connected layers and reduce the number of weights required for transmission between fully connected layers; In addition, by dividing each standard convolutional layer in the middle of the network into two, it is modified to a combination of a deep convolution and a point convolution introduced in section 1.3, ensuring that the input and output sizes of the layer remain unchanged. In the case of reducing the amount of parameters, the improved VGG16 structure in this paper is shown in Figure 5. In order to speed up the convergence of the model, the improved algorithm adds a Batch Normalization layer before the activation function of each convolutional layer to optimize the network.

![Figure 4. Structure of VGG16.](image-url)
3.3. Solar cell defect detection based on data enhancement

Combining the data enhancement algorithm mentioned in section 2.1 and the lightweight network in section 2.2, the solar cell defect detection process based on data enhancement is shown in Figure 6. The overall process can be divided into the following steps:

Step1: Use DCGAN training model, input random noise to generate defect pictures, including three defects of splinters, cracks and stains, and screen out pictures with good visual effects for use;

Step2: Send the real pictures and the defective pictures generated in step1 to the random fusion network by category in order to obtain the fused pictures to expand the training data set;

Step3: Combine the original training set and the fusion data obtained in step 2 as the input of the lightweight defect classification network for model training;

Step4: Use the trained model to classify and detect the defects of solar cells.

4. Experiment and result analysis

4.1. Experimental conditions and evaluation indicators

The data set used in the experiment in this paper is the pictures of solar cells collected at the production site. The number of various pictures is shown in Table 1. It can be seen that the amount of data directly obtained is small, and it is very likely that it will be directly used to train the model for classification. The model is overfitting. In this experiment, the TensorFlow framework is built on the hardware platform of Intel i7-7700HQ CPU and 8G memory GTX1060ti GPU for experimentation.
Table 1. Data volume of various types of defects

| Type   | Training set | Test set |
|--------|--------------|----------|
| flawless | 352          | 120      |
| Lobes   | 88           | 24       |
| crack   | 136          | 40       |
| stain   | 224          | 56       |

4.2. Experimental results and analysis

The model parameters before and after the lightweight processing and the test performance comparison are shown in Table 2. It can be seen that after the lightweight processing, the training parameters of the model are reduced to about half of the original. The test time was shortened from 57ms to 22ms. In this way, the algorithm not only reduces the storage space of the model, but also speeds up its testing speed, which is conducive to its application and popularization in industrial production to achieve the purpose of saving costs and improving production efficiency.

Table 2. Comparison of models before and after VGG16 lightweight processing

| Mode       | Training parameter amount | testing time (Per sheet) |
|------------|---------------------------|--------------------------|
| VGG16      | 40,945,481                | 57 ms                    |
| VGG16_Lightweight | 27,897,289         | 22 ms                    |

It can be seen from Table 3 that the training accuracy of the model obtained by training with only the scarce original data set can reach 100%, while the test accuracy is only about 47%, and the test accuracy is far lower than the training accuracy; It is not difficult to see that the model is seriously overfitting and has a perfect discrimination effect on the training data, but the performance on the test set is very unsatisfactory; This poor performance defect classification model makes it easier to misjudge the types of solar cell defects and cannot be used in industrial production. Using traditional data enhancement algorithms, the data set is expanded to 8 times the original training data set by flipping and rotating the original training data;

Component heads identify the different components of your paper and are not topically subordinate to each other. Examples include Acknowledgments and References and, for these, the correct style to use is “Heading 5”. Use “figure caption” for your Figure captions, and “table head” for your table title. Run-in heads, such as “Abstract”, will require you to apply a style (in this case, italic) in addition to the style provided by the drop down menu to differentiate the head from the text. It can be seen that compared to training with only the original data, the model test accuracy obtained by using the training set expanded by this method is increased by about 10%, but it is only about 60%, and the test effect is also unsatisfactory. Using the true and false data fusion algorithm mentioned in this article, for each type of defective pictures, first use DCGAN to generate 100 fake pictures of the same type and randomly merge the real pictures, combined with traditional algorithms such as rotation and flip, to expand the training data set to 800 times the original training set; It can be seen from the model test results that the true and false data fusion algorithm effectively expands the scarce original data, and the test accuracy of the model trained after the expansion is improved by nearly 30% compared with the model trained on the original training data, reaching about 77%; It alleviates the model overfitting caused by the lack of original training data.
Table 3. Training and test results of different algorithms

| Method _ model            | Training accuracy | Test accuracy |
|--------------------------|-------------------|---------------|
| Raw training data _ VGG16| 100.00%           | 47.78%        |
| Raw training data _ VGG16_Light | 100.00%   | 47.05%        |
| Traditional data enhancement _ VGG16 | 98.86%   | 60.29%        |
| Traditional data enhancement _ VGG16_Light | 99.62%   | 59.56%        |
| Fusion of true and false data _ VGG16 | 99.65%   | 77.20%        |
| Fusion of true and false data _ VGG16_Light | 99.71%   | 76.47%        |

At the same time, combined with Table 2 and Table 3, it can be seen that after the lightweight processing of the network model, the amount of training parameters of the model is effectively reduced, and the test time is shortened, but the accuracy is slightly reduced; However, it is worthwhile to sacrifice less than 1% accuracy in exchange for a substantial improvement in model performance, and it is also very meaningful to promote its application in industrial production.

5. Conclusion
The true and false data fusion algorithm proposed in this paper expands the scarce solar cell defect data set through the random fusion of DCGAN production pictures and real pictures, constructs more training images, and alleviates the model overfitting problem caused by insufficient data. Compared with traditional data enhancement algorithms, it is also more efficient; at the same time, the network model is light weighted to reduce storage space and shorten defect testing time. The experimental results show that the combination of true and false data fusion algorithms and lightweight network processing, by increasing the number of training images and reducing training parameters, prevents the convolutional neural network model from overfitting, improves the testing accuracy and speeds up the model testing, which is helpful Realize automatic detection of solar cell defects, reduce manpower and material resources, and improve production efficiency. The research of data enhancement algorithms and lightweight models is of great significance to the intelligent development of many data-deficient fields such as defect detection, medical image analysis, and financial data analysis. It is hoped that the algorithm in this paper can play a positive role in image classification and recognition in scenarios with insufficient data sets.

References
[1] Cao Zhenhua. Analysis of solar photovoltaic power generation system [J]. Clean and Air Conditioning Technology, 2017(03):107-109.
[2] Yuan Tianwen. Application technology of solar photovoltaic power generation system[J]. Electronic Technology and Software Engineering, 2014(11):141-141.
[3] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, “Backpropagation Applied to Handwritten Zip Code Recognition,” Neural Computation, vol. 1, no. 4, pp. 541–551, 1989.
[4] C. Sun, A. Shrivastava, S. Singh, and A. Gupta, “Revisiting Unreasonable Effectiveness of Data in Deep Learning Era,” arXiv, pp. 1–13, 2017.
[5] L. M. Zintgraf, T. S. Cohen, T. Adel, and M. Welling, “Visualizing Deep Neural Network Decisions: Prediction Difference Analysis,” Proc. of International Conference on Learning Representations (ICLR2017), pp. 1–12, 2017.
[6] Jiang Yun, Zhang Hai, Chen Li, Tao Shengxin. Image data enhancement algorithm based on convolutional neural network [J]. Computer Engineering and Science, 2019(11): 2007-2016.
[7] D. Cirean, U. Meier, and J. Schmidhuber, “Multi-column Deep Neural Networks for Image Classification,” in Proc. of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR2012), 2012, pp. 3642–3649.
[8] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A.
Rabinovich, “Going deeper with convolutions,” in Proc. of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR2015), vol. 07-12-June, 2015, pp. 1–9.

[9] K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” in Proc. of International Conference on Learning Representations (ICLR2015), 2015, pp. 1–14.

[10] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” in Proc. of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR2016), 2016, pp. 770–778.

[11] Srivastava Nitish,Hinton Geoffrey, Krizhevsky Alex, Sutskever Ilya, Salakhutdinov Ruslan, “Dropout: A simple way to prevent neural networks from overfitting,” Journal of Machine Learning Research, v 15, p 1929-1958, June 2014.

[12] Li Zhiheng, Hejun, Hu Zhaohua. Semi-supervised domain adaptive method based on dropout regularization[J].Application Research of Computers,2019(11).

[13] 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 (NIPS), 2014.

[14] Satoshi Tsutsui,Yanwei Fu,David Crandall.Meta-ReinforcedSyntheticDataforOne-Shot Fine-GrainedVisualRecognition. In Advances in Neural Information Processing Systems (NIPS2019).

[15] K. He and X. Zhang, “Identity mappings in deep residual networks,” Lecture Notes in Computer Science, vol. 9908, no. 1, pp. 630–645, 2016.

[16] Ma Yongjie, Liu Peipei. Image classification algorithm based on DenseNet evolutionary convolutional neural network [J]. Progress in Laser and Optoelectronics, 2019.

[17] Zhang Xiangyu ,Zhou Xinyu, Lin Mengxiao, Sun Jian. ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices .Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, p 6848-6856, December 14, 2018, Proceedings - 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2018.

[18] Zhou You, Liu Yiyue, Han Guijin, Fu Yiping.2019 IEEE Symposium Series on Computational Intelligence, SSCI 2019, p 2776-2781, December 2019, 2019 IEEE Symposium Series on Computational Intelligence, SSCI 2019.

[19] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. B. Wojna, “Rethinking the Inception Architecture for Computer Vision,” in Proc. of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR2016), 2016, pp. 2818–2826.

[20] Krizhevsky Alex, Sutskever Ilya, Hinton Geoffrey E. ImageNet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems, v 2, p 1097-1105, 2012, Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012, NIPS 2012.

[21] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, “Random Erasing Data Augmentation,” arXiv, pp. 1–10, 2017.

[22] T.DeVriesandG.W.Taylor,“ImprovedRegularizationofConvolutional Neural Networks with Cutout,” arXiv, pp. 1–8, 2017.

[23] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, “mixup: Beyond Empirical Risk Minimization,” in Proc. of International Conference on Learning Representations (ICLR2018), 2018, pp. 1–13.

[24] Lu Li, Wang Jie, Han Yaochang, Wu Yahui. Application of improved DCGAN in SAR image classification [J]. Control Engineering, 2020(03): 561-566.