Image Confusion Applied to Industrial Defect Detection System

Hao-Yuan Chen, Yu-Chen Yeh, Makena Lu, and Chia-Yu Lin
Department of Computer Science and Engineering, Yuan Ze University, Taiwan.
Email: sallylin0121@saturn.yzu.edu.tw

2022 IEEE International Conference on Consumer Electronics - Taiwan | 978-1-6654-7050-6/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICCE-TAIWAN55306.2022.9869058

***Abstract*—There have been many related security issues about Artificial Intelligence (AI) in recent years. During the manufacturing process, products are captured by images for defect detection. If attackers use the model inversion attack to attack the AI model, the input image can be roughly restored, resulting in product information leakage. In this paper, we propose a system that confuses input images and uses them to train the model. Experiments show that our model has a high accuracy of 94.4% in defect image classification. Thus, the proposed system can achieve product information protection and accurate defect detection.**

I. Introduction

Manufacturing is the core industry of Taiwan. With the development of AI, the use of AI to improve the efficiency of manufacturing processes has become a trend. Industrial defect detection usually uses machine learning to detect product defects in the manufacturing process. It captures the product images and uses them to train the model. After training, the model can classify product defects successfully.

These training images are closely related to the products of the production line. Suppose an AI model identified on the production line suffers from the model inversion attack [1]. In that case, attackers can roughly restore the input data, which will lead to compromised product information, causing considerable damage to the enterprise.

In order to effectively protect data, cloud service providers, such as Amazon, provide corresponding solutions. They added privacy-preserving machine learning (PPML) [2] capabilities to cloud services so that the model could directly calculate encrypted data and return encrypted results. Only those who encrypted the data can decrypt the results to protect the product data. However, although this method could effectively ensure the security of the data, it increased the consumption of many computing resources. The inference time of models was also 400 times longer than those without PPML. Another way to protect data was learnable image encryption [3], which used relatively weak block-wise image encryption to make data non-understandable for humans, but learnable for machines. Moreover, it still had powerful calculations with deep neural networks. However, it could be identified by the object contour if the object was clear enough.

In this paper, we propose an Image Confusion applied to industrial Defect Detection System (ICDDS). Through the image confusion algorithm we designed, the human eye cannot distinguish the content of input images, but the AI model can still learn the corresponding features and classify them. In this paper, we achieve the following with ICDDS:

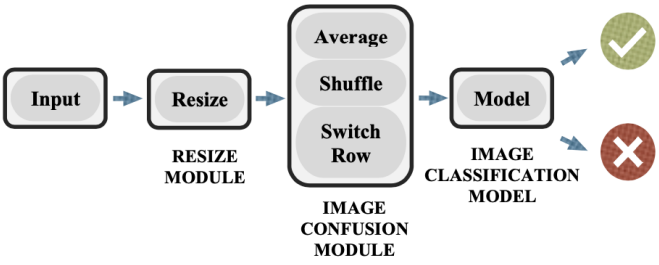


Fig. 1: The overview of the proposed ICDDS Architecture

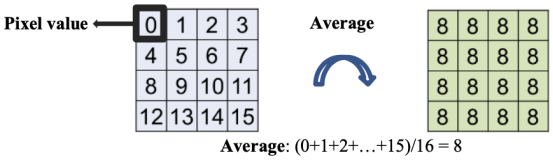


Fig. 2: The concept of the image average processing

A. Although attackers can obtain input data through Model Inversion Attack, they cannot obtain the product's data.

B. We do not use PPML to train the model, so the computation time is significantly reduced.

C. Compared to learnable image encryption, our system obfuscates images to a greater degree.

II. System Architecture

In order to defend the model reverse attack, we propose ICDDS. Fig. 1 shows the system architecture. Input images first go through the resize module to adjust to a specified size. Then, resized images enter the image confusion module to perform the Average, Shuffle, and Switch Row process in sequence. Finally, the image classification model detects defect images. The following is an explanation of each part.

A. RESIZE MODULE

The resize module will first resize the input images to the specified size to save resources for training the model.

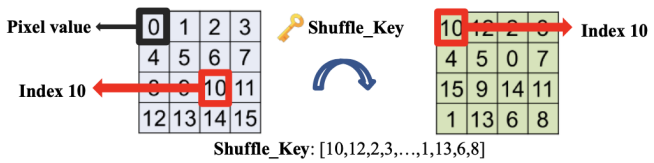


Fig. 3: The concept of the image shuffle processing

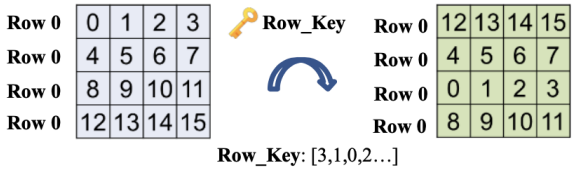


Fig. 4: The concept of the image switch row processing

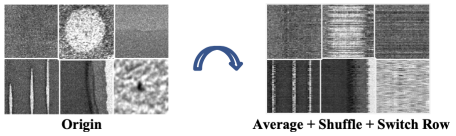
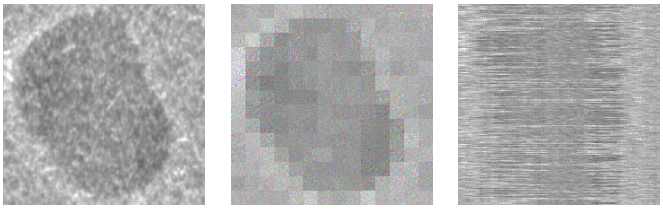


Fig. 5: AOI defect image after being processed by Image



(a) Original (b) block wise pixel shuffle (c) Proposed

Fig. 6: The effect of confusion between different method

B. IMAGE CONFUSION MODULE

1) Average: Average will move a specified-size mask from top to bottom and left to right. Each movement will calculate the average value from the mask pixels. Then, we replace all mask pixels with the average value. Average can achieve the blurring of images. Fig. 2 shows the process of Average.

2) Shuffle: We simplify the method proposed by [3]. Shuffle will randomly generate a Shuffle_Key whose length is the area of the mask. Then, it will move the mask from top to bottom and left to right and sort the mask pixels according to the value in Shuffle_Key. Shuffle can achieve confusion of images. Fig. 3 shows the process of Shuffle.

3) Switch Row: Switch Row will randomly generate a Row_Key whose length is the height of the image. Then, it will sort the row of the image according to the value of Row_Key. Switch Row can achieve better confusion of images. Fig. 4 shows the process of Switch Row.

C. IMAGE CLASSIFICATION MODEL

We use transfer learning to train the image classification model, which has a better convergence effect and saves time when training models [4]. We use the VGG16 as a pre-trained model. First, we freeze all convolutional layers to do feature extraction. Second, we modify the classifier. We change the last dense layer of the pre-trained model to our dense layer and fine- tune our classifier.

III. EXPERIMENT

In this experiment, we use TensorFlow and a computer with RTX 2070 to build ICDDS. We choose the image data set provided by "AOI defect classification" on AIdea. There have six classes, "Normal," "Void," "Horizontal defect," "Vertical defect," "Edge defect," "Particle," respectively. First, we resize the defective images to 224*224 through the resize module and use the image confusion module to perform Average, Shuffle, Switch Row on them, where the mask of Average is 2*2, Shuffle is 4*4. Fig. 5 shows the before and after of the defect image processing, and Fig. 6 shows the effect between the shuffle of learnable image encryption and our confusion method. Then, we split the image data set according to the ratio of training data: test data = 8:2. While training the image classification model, we used Adam optimizer and set the learning rate to 0.00001, batch size to 32, and epoch to 60. After the model training, we choose Accuracy, Precision, Recall, and F1-Score, four common indicators, to evaluate the model. We repeat the ten times model training and average the results as the experimental result. We get Accuracy 94.4%, Precision 94.7%, Recall 94.4%, and F1-score 94.3%. Moreover, our model only takes 0.002 seconds to predict an image.

IV. CONCLUSION

In this paper, we investigated the problem of data privacy that came with the development of industrial AI. We presented ICDDS, which confused the image so the attacker could not obtain the input data through the model inversion attack. Experiments and analysis showed that ICDDS could achieve high accuracy and spent 0.002 seconds only when predicting one image, which was more efficient than PPML. As a result, it could solve the problem of weak confusion and the inversion attack. In the future, we will continue to optimize our algorithm to increase accuracy and prevent the model from being attacked in any way.

REFERENCES

[1] Fredrikson, Matt et al. "Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures." *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security (2015): n. pag.*

[2] Xianrui Meng, Joan Feigenbaum "Privacy-Preserving XGBoost Inference" *NeurIPS 2020 Workshop on Privacy Preserving Machine Learning*

[3] Tanaka, Masayuki. "Learnable Image Encryption." *2018 IEEE International Conference on Consumer Electronics-Taiwan (ICCE-TW) (2018): 1-2.*

[4] J. A. F. Thompson, M. Schönwiesner, Y. Bengio and D. Willett, "How Transferable Are Features in Convolutional Neural Network Acoustic Models across Languages?," *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019, pp. 2827-2831, doi: 10.1109/ICASSP.2019.8683043.*