electronics

MDPI

Article

Research and Application of Generative-Adversarial-Network Attacks Defense Method Based on Federated Learning

Xiaoyu Ma and Lize Gu *

Institute of Cyberspace Security, Beijing University of Posts and Telecommunications, Beijing 100876, China
* Correspondence: glzisc@bupt.edu.cn

Abstract: In recent years, Federated Learning has attracted much attention because it solves the problem of data silos in machine learning to a certain extent. However, many studies have shown that attacks based on Generative Adversarial Networks pose a great threat to Federated Learning. This paper proposes Defense-GAN, a defense method against Generative Adversarial Network attacks under Federated Learning. Under this method, the attacker cannot learn the real image data distribution. Each Federated Learning participant uses SHAP to explain the model and masks the pixel features that have a greater impact on classification and recognition in their respective image data. The experimental results show that while attacking the federated training model using masked images, the attacker cannot always obtain the ground truth of the images. At the same time, this paper also uses CutMix to improve the generalization ability of the model, and the obtained model accuracy is only 1% different from that of the model trained with the original data. The results show that the defense method proposed in this paper can not only resist Generative Adversarial Network attacks in Federated Learning and protect client privacy, but also ensure that the model accuracy of the Federated model will not be greatly affected.

Keywords: federated learning; GAN; defense; security; privacy

Citation: Ma, X.; Gu, L. Research and Application of Generative-Adversarial-Network Attacks Defense Method Based on Federated Learning. *Electronics* **2023**, *12*, 975. https://doi.org/10.3390/electronics12040975

Academic Editors: Chunxu Li, Ashraf Fahmy, Hooman Samani and Gang He

Received: 2 January 2023
Revised: 13 February 2023
Accepted: 13 February 2023
Published: 15 February 2023

1. Introduction

Since its birth, Artificial Intelligence (AI) has expanded its application fields and has begun to create value for society. Currently, AI faces two main challenges [1], one is the problem of data silos, and the other is the privacy and security of data. In practical applications, the data required by AI models involve many types. In most industries, however, the data owned by individuals are limited and poor, and it is impossible to build a well-performing AI model using the existing private data. To address this barrier, we hope to combine data from various domains into a common model. However, due to industry competition, data privacy and security concerns, and even the high cost, data aggregation between different institutions can be difficult.

To solve data silos and data privacy security issues, Google first proposed the concept of Federated Learning [2] (FL). In the FL process, each party can perform model training locally, and there is no need to disclose and share private data, which ensures data privacy. At the same time, the model obtained through FL can approximate the performance of the ideal model (a machine learning model obtained by training the data of each participant together).

As a new research field, FL still faces privacy and security issues [3,4]. At present, the security threats encountered by FL are mainly divided into several categories: poisoning attacks [5,6], backdoor attacks [7–9], communication bottlenecks [10,11], generative adversarial network(GAN) [12] attacks [13–15], etc. Among them, the GAN-based attack can generate private data similar to that of the victim, making the victim's privacy leak. In other words, the attacker can obtain samples of other participants, and this process only requires black-box access. Since all possibilities of a GAN-based threat cannot be foreseen, it is classified as a high-impact and priority threat.

Privacy protection for FL currently mainly includes two methods, namely traditional encryption methods (such as secure multi-party computation [16] and homomorphic encryption [17]) and differential privacy [18]. For the GAN attack under FL, Yan et al. [19] proposed a privacy protection method based on deep convolutional generative adversarial network (DCGAN), setting a buried point layer before the fully connected layer, and thereby identifying the attacker's behavior, and changing the attacker's behavior by changing training learning rate to break its GAN attack. The defense method of Anti-GAN is proposed by Luo et al. [20], which uses GAN to improve the original training data set from the perspective of the defender, so that the attacker cannot generate the original image data of the defender.

There are still some imperfections in the currently proposed defense methods against GAN attacks under FL. For example: (1) Each operation of secure multi-party computing and Homomorphic encryption requires interaction, which consumes a lot of computing time and power, so its performance is relatively weak [3]; (2) Differential Privacy is very likely to prevent the model from converging, resulting in the model not showing good performance. And, in [13], the authors mentioned that record-level differential privacy is ineffective for GAN attacks in FL; (3) DCGAN is suitable for collaborative deep learning scenarios where the attacker does not understand the local model training process of benign actors; (4) Anti-GAN uses Mixup to enhance data. However, the literature shows that the data enhancement effect of Mixup is not the best, and Cutmix, a combination of Mixup and Cutout, has a better effect on data enhancement. Therefore, the model accuracy loss of Anti-GAN is relatively serious. In view of the above problems, how to optimize and improve the defense against GAN attacks in FL requires further research work. This paper designs a powerful defense framework Defense-GAN against GAN attacks in FL. While protecting the privacy of benign clients' image data, it also makes sure that the accuracy of the model trained by FL is not greatly affected. The framework of Defense-GAN is shown in Figure 1.

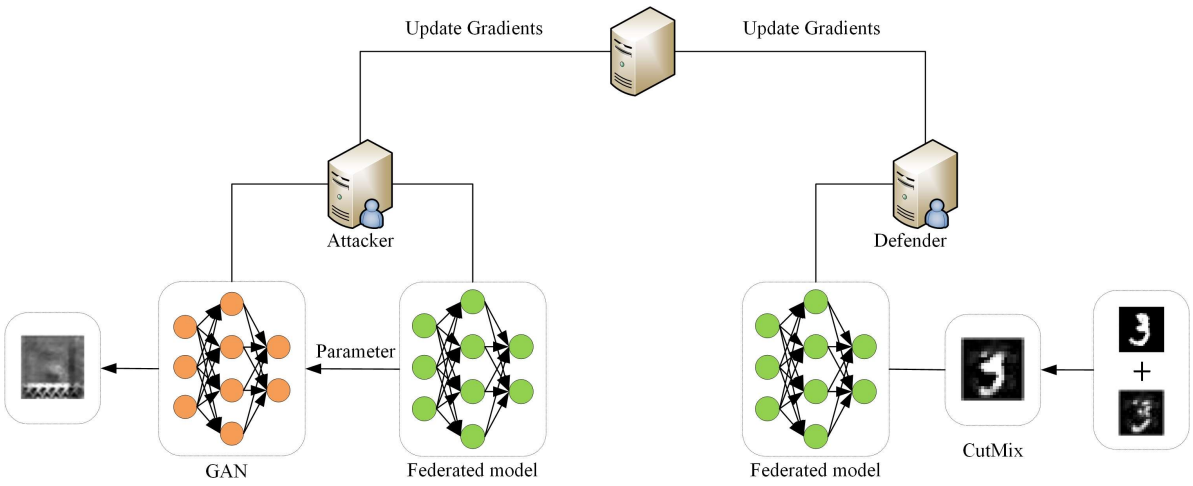

Figure 1. Defense framework of Defense-GAN.

The contributions of this paper are summarized as follows:

(1) This paper proposes a defense framework for GAN attacks in FL, Defense-GAN, which makes attackers unable to obtain effective private image data of defenders after GAN attacks, thus protecting the privacy of defenders. At the same time, it also ensures that the model accuracy of FL training will not be greatly affected.

(2) The Defense-GAN proposed in this paper uses SHAP to compare the influence of all features in the training privacy data set and sorts them to mask the top-ranked features, and destroy the visual features of the training images, so that the data

images obtained by the attacker cannot be recognized with the naked eyes after the GAN attack.

(3) In the model training process, the masked image data and the original image data are mixed with CutMix to form a new image data set to improve the accuracy of model training.

(4) The defense framework proposed in this paper has been tested on real datasets and compared with other methods in terms of defense effect and model accuracy. The experimental results show that the scheme proposed in this paper not only defends against GAN attacks but also ensures the accuracy of model training.

2. GAN Attacks under FL

The defense method proposed in this paper, Defense-GAN, is aimed at a powerful attack against collaborative deep learning using GANs proposed by Hitaj et al. [13]. This section mainly discusses the framework and principle of its attack.

2.1. Attack Scenario

There are two main types of privacy attacks against distributed learning: one is that each client uploads a private dataset to a single location for server operators to train models. The operator can directly access sensitive information, resulting in leakage of the client's privacy. Such problems can be solved by transforming centralized machine learning into FL; the second is the attack scenario for FL, which is the attack to be solved in this article. Wherein, any client acting as an insider can infer sensitive information from the victim's device.

In the attack proposed by Hitaj et al. [13], the attacker first acts as an honest participant and participates in the training process of FL, trying to extract information that does not belong to him but belongs to other participants. At the same time, the attacker also secretly influences the learning process to trick the victim into further revealing more information about the target class. Among them, the attacker can steal the information of the target class without destroying the central parameter server (PS) that collects and distributes the parameters.

2.2. Attack Model

The defense proposed in this paper is aimed at GAN attacks, which mainly occur in the process of FL training image classification recognition models. In implementing the attack, the attacker A, like all the participants, agrees in advance on a common learning objective, which means they agree on the type of neural network structure and the labels that will be trained. The victim V has data labeled $[a, b]$, and the attacker A has data labeled $[b, c]$, and the goal of A is to reconstruct V's data labeled a. A secretly builds a GAN locally to reconstruct the data belonging to V. Like other participants, A downloads parameters from the PS during each round of training and updates the local model. However, A will use the updated model as a discriminator of GAN to identify whether the fake data of label a generated by GAN is different from the data of real label a, and use this to improve its own GAN generator, generate fake data that is closer to the real data. At the same time, A marks the generated fake class a data as class c, then performs model training on these data, and uploads its locally trained model parameters to PS. Therefore, while A obtains V's private data, it also interferes with the training of the model, so that V will further use a large amount of data for model training for the model to recognize the data of type a and type c, and improve the model accuracy. At the same time, the discriminator of A's GAN will also further improve the discriminant ability, thereby generating fake data that is closer to the real data. Finally, A can reconstruct a class a image that is indistinguishable from V's original image, to achieve the purpose of stealing the victim's private data.

3. Defense Method

3.1. The Framework of Defense-GAN

The Defense-GAN proposed in this paper is used to defend against GAN attacks in FL and protect the privacy of the victim (defender) from being leaked to the attacker. The defense idea of Defense-GAN is as follows: first, partially mask the images in the original training data set of the defender, so that the attacker cannot use the fake data generated locally after the GAN attack to identify the original appearance of the real image with the naked eye, to achieve the purpose of privacy protection. At the same time, this paper will also mix the masked image and the original image with CutMix to form a new image dataset to ensure that the accuracy of the model trained on the masked image data does not differ too much from the accuracy of the model trained on the raw image data.

In the defense framework of this paper, the most important part is to mask the training image data of the defender. However, considering that the masked image should not only make it impossible for attackers to recognize it with the naked eye but also have a small impact on the model's accuracy. This paper considers masking images using SHapley Additive exPlanation (SHAP). It first uses the original image data to train the model locally, and then uses SHAP to analyze it by analyzing the pixels that affect the classification (visual effect) of each image, and then masks these pixels. The masked image is then used for model training, and the trained model parameters are uploaded to the PS.

In addition, as part of improving the accuracy of the model, this paper employs CutMix to mix the masked image data with the original image data to improve the generalization ability of the model.

3.1.1. SHAP

In the defense method designed in this paper, the image masking part uses SHAP to mask the image. SHAP is a "model interpretation" package developed in Python that can interpret the output of any machine learning model. It builds an additive explanatory model inspired by cooperative game theory, treating all features as "contributors". For each predicted sample, the model produces a predicted value, and the SHAP value is the value assigned to each feature in that sample.

SHAP is a model that explains machine learning models. A machine learning model is generally a black box. For example, in the attack defense scenario of this paper, the model wants to predict the category of the picture. Input some known conditions into the model (various types of pictures), and then train, and the final trained model can directly predict the category according to the input pictures. Therefore, such a model can only obtain the final result. As for how the model is calculated internally, and how the known conditions of the input affect the output result, people cannot know. And SHAP can let people know what influence these known conditions have on the final prediction result (whether it has a positive or negative impact on the result) [21]. SHAP belongs to the method of model post-exposure explanation. Its core idea is to calculate the marginal contribution of features to the model output, and then explain the "black box model" from the global and local levels.

SHAP essentially uses the linear interpretation of the sample results given by Shapley Value to calculate the role of different features for a certain predicted value. The rationale for Shapley Value is to take the (weighted) average of the marginal contributions of the feature over a subset S of all feature combinations as the feature's contribution. Suppose we want to study the j-th feature x_j of the sample x, $\varnothing_j$ is the contribution of this feature, S is the subset of features, M is the total number of features, $val(*)$ value function is the model and its predicted value in the context of machine learning in this paper, $(x_1, \ldots, x_M) \backslash x_j$ is the set that excludes x_j. Then the calculation formula of the Shapley Value corresponding to the feature x_j of the sample x is:

$$\varnothing_j(val) = \sum_{S \subseteq (x_1,\ldots,x_M) \backslash x_j} \frac{|S|!(M-|S|-1)!}{M!} \left(val\left(S \cup x_j\right) - val(S)\right) \quad (1)$$

It is essentially the expectation of the "marginal contribution" of a feature in a sample.

When the model prediction is used as the value *val* in the Shapley Value formula, and the result of the model is explained by adding and subtracting from the mean value of the predicted value to the predicted value, it becomes SHAP:

$$g\left(z^i\right) = \varnothing_0^{(i)} + \sum_{j=1}^{M} \varnothing_j^{(i)} z_j^{(i)} \tag{2}$$

where $\varnothing_0$ is the mean of predicted values, $\varnothing_j$ is Shapley Value, z is the indicator vector of the sample, and the value is 1 if the feature belongs to the target class, and 0 if it is replaced by a randomly sampled sample value. i represents the i-th sample. This formula explains the feature contribution for a certain target class.

3.1.2. Image Masking

In the scenario of defending against GAN attacks in this paper, each pixel of the image is used as a feature, and SHAP calculates the SHAP value corresponding to each pixel. The larger the SHAP value, the greater the impact of this pixel on the classification and recognition of such images. After calculating all the SHAP values of each image, this paper will sort them in descending order. Then, mask the pixels corresponding to the top SHAP values obtained by sorting, that is, change them to the background color, and generate new masked image data.

After masking all the images in the dataset, image data that can be recognized by the naked eye cannot be obtained, so valuable information cannot be obtained after an attacker launches an attack.

3.1.3. Improve Model Accuracy

Since the masked image will reduce the accuracy of the model, the category of the image cannot be accurately identified. Therefore, in this paper, the masked image and the original image are mixed to form a data set of mixed images, so that the network can generalize better and have better object recognition ability.

Yun et al. proposed a data augmentation method named CutMix [22]. Specifically, this method cuts and pastes patches in the training image, where the ground-truth labels are also blended proportionally to the area of the patch.

Let $x \in R^{W \times H \times C}$ and y denote the training images and corresponding labels, respectively. The goal of this paper is to use CutMix to generate a new training sample $(\widetilde{x}, \widetilde{y})$ consisting of the masked image sample (x_A, y_A) and the original image sample (x_B, y_B). Define the combined operation as

$$\widetilde{x} = M \otimes x_A + (1 - M) \otimes x_B \tag{3}$$

where $M \in \{0,1\}^{W \times H}$ is a binary mask representing the position removed and filled from the two images; 1 is a padded binary mask and $\otimes$ is element-wise multiplication.

For the label $\widetilde{y}$ of the new image, since the masked image corresponds to the category of the original image, given a class $\widetilde{y}$, first randomly extract an image x_A with the label y_A from the masked image dataset, then, the label corresponding to the y_A category is extracted from the original data image as the y_B image x_B, and the label of the final mixed image $\widetilde{x}$ is $\widetilde{y}$.

To sample the binary mask M, the bounding box coordinates $B = \left(r_x, r_y, r_w, r_h\right)$ are first sampled to represent the cropped regions in x_A and x_B. The area in x_A is removed by B, filled with the patch cropped by B of x_B.

The aspect ratio of the rectangular mask M sampled in our experiments is proportional to the original image. The coordinates of the rectangle are uniformly sampled as:

$$\begin{aligned} r_x &\sim Unif(0, W), r_w = W\sqrt{1-\lambda}, \\ r_y &\sim Unif(0, H), r_h = H\sqrt{1-\lambda} \end{aligned} \tag{4}$$

Define the cropped area ratio to be $r_w r_h / WH = 1 - \lambda$, where λ is the combined ratio, sampled from a uniform distribution $(0, 1)$.

4. Experiments

Based on the GAN attack code provided by the article [13], this paper implements the defense against this attack and the protection of private images.

4.1. Experimental Setting

4.1.1. Dataset

The experiments in this paper are all performed on the handwritten font image dataset MNIST [23].

MNIST is a picture dataset of handwritten digits. The dataset was organized by the National Institute of Standards and Technology (NIST) in the United States. A total of 250 different people's handwritten digit pictures were collected. 50% are high school students and 50% are from Census Bureau staff. It is divided into 60,000 28×28 training pictures and their corresponding labels (0~9) and 10,000 test pictures and their corresponding labels (0~9). This dataset has been widely used in machine learning and deep learning since 1998 to test the effects of algorithms, such as Linear Classifiers, K-Nearest Neighbors, Support Vector Machines (SVMs), Neural Nets, Convolutional nets, etc.

4.1.2. Experimental Configuration

All experiments were performed in a Linux environment with an Intel Xeon E5-2630 v4 2.20 GHz CPU, NVIDIA GeForce GTX 1080Ti GPU, and 64 GB memory. The programming language was Python, and the deep learning framework used was Pytorch.

In this paper, each image is scaled to the range of $[-1, +1]$, and the output result is also in the range of $[-1, +1]$. The federated model training for each customer is set to 200 rounds, and the attacker performs 200 rounds of GAN model training locally every time the model is updated. When using SHAP for analysis, the top 200 pixels with the greatest influence of features in each image are selected for masking, and the masking is the background color. The optimizer of the GAN chooses the gradient descent algorithm, the learning rate is 0.001, and the weight decay is set to 1×10^{-7}. The learning rate of the generator is set to 0.0005 and the learning rate of the discriminator is set to 0.0002.

4.2. Image Masking with SHAP

In the Defense-GAN proposed in this paper, the image data needs to be masked first. In this part, this paper first uses SHAP to find the features (pixels) in the image that have the greatest impact on its classification and recognition. Here, the paper uses Deep SHAP, a high-speed approximation algorithm for SHAP values in deep learning models, which builds on the connection to DeepLIFT described in the SHAP NIPS paper. The implementation here differs from the original DeepLIFT in that it uses the distribution of background samples instead of a single reference value, and uses the Shapley equation to linearize components such as max, softmax, products, divisions, etc. The visualization of the feature contribution of the images is shown in Figure 2.

Figure 2 explains ten outputs (numbers 0–9) for 5 different images. Red pixels increase the output of the model, while blue pixels decrease the output. The input image is shown on the left and behind each explanation as a nearly transparent grayscale background. The sum of the SHAP values is equal to the difference between the expected model output (averaged over the background dataset) and the current model output.

The next step is to mask the red pixels. Here, to achieve the desired effect, this paper will mask the top 200-pixel features that have the greatest impact on the 28×28 image classification and recognition, and mask it as the background color of the image, and the result is shown in Figure 3.

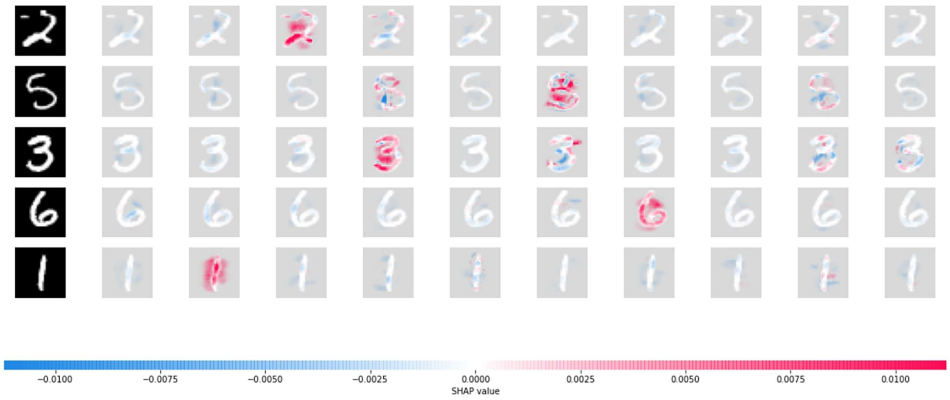

Figure 2. Image feature contribution distribution map calculated by SHAP.

Figure 3. Comparison of MNIST images before and after masking.

As can be seen from Figure 3, the masked images are basically unable to identify their corresponding number with the naked eye. Then, the masked images are saved as a dataset for the next step of training the FL model.

4.3. Mixing Images with CutMix

To improve the generalization ability of the model, the method proposed in this paper also needs to crop the masked image and fill the original image as a patch. All pictures are 28 × 28 square pictures, so they are cropped proportionally, the length and width of the cropping area are selected as 10–20, the crop size is 10 × 10 rectangle, and the corresponding part is cropped from the original picture to fill. The masked images, the original images, and the CutMix images are shown in Figure 4.

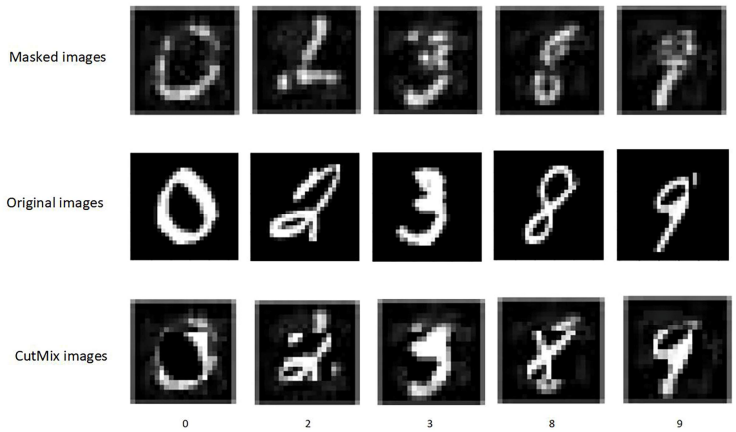

Figure 4. Comparison of masked images, original images, and CutMix images.

4.4. Defense Effects

4.4.1. Attack on the Original Dataset

The defense in this paper is against the GAN-based attack proposed by Hitaj et al., so we reproduce their experiments as the attack process for FL in this paper. First, this paper attacks the training process of FL using the original data set, and the attack effect is shown in Figure 5.

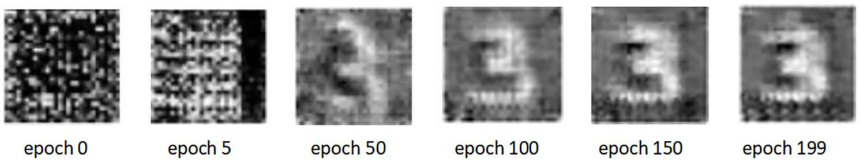

Figure 5. Attack effect on the training process of FL using the original dataset.

As can be seen from Figure 5, since the initial input of GAN is noise, the image at the beginning of the 0th round of attack is a full noise image. When the attack reaches the 5th round, the specific image of the attacked image cannot be seen. When the training reaches the 100th round, it can be roughly seen that the number "3" of the attacked image presents a clearer image, and when the training reaches the 199th round, the image generated by the attack is already very clear.

During the attack, the losses of the GAN generator and discriminator in the 0th, 50th, 100th, and 199th rounds of the attack are shown in Figure 6. In the coordinate image in Figure 6, during the 200 rounds of local training by the attacker using GAN, the loss of the generator continued to decrease, and the loss of the discriminator continued to increase. The two constantly improve their respective performance in the process of the mutual game, that is, the generator generates more realistic pictures, and the discriminator can more accurately distinguish real and fake pictures. During the 200 rounds of FL training, in the initial state of the generator's local training, its loss decreases with the increase of the number of rounds, indicating that the generator continuously optimizes itself using the federated model parameters downloaded from the server in each round. And the loss of the discriminator also decreases with the increase of the number of rounds, indicating that the discriminator is increasingly unable to distinguish the true and false images generated by the generator, and it also reflects that the images generated by GAN are already very realistic to the images of the victims.

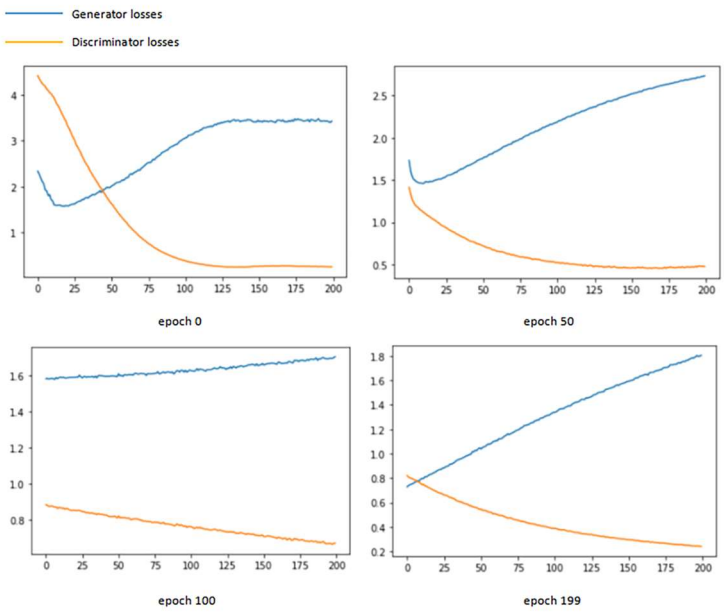

Figure 6. GAN's discriminator and generator losses.

4.4.2. Attack on Pure Masked Images

In this part, all the masked images are formed into a dataset for FL model training, and a GAN attack is launched during the training process. The effect of the attack is shown in Figure 7.

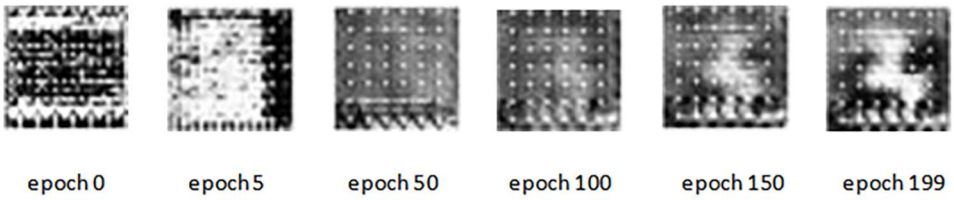

Figure 7. Attack effect on the training process of FL using pure masked image datasets.

As can be seen from Figure 7, the image generated when the attack reaches the 150th round begins to show the rough outline of the pattern, and the specific image of the number "3" cannot be effectively recognized when the attack reaches the 199th round. Therefore, the GAN attack is ineffective against the masked image, and the adversary cannot obtain useful information from the defender dataset.

4.4.3. Attack on Mixed Dataset of Masked Image and Original Image

Since the use of pure masked image data for FL model training will lead to a decrease in model accuracy (Table 1), in this subsection, this paper adds some original images to the masked image dataset to form a new dataset for FL training. At the same time, an attack is launched on the attacker's side. The effect of the attack is shown in Figure 8.

Table 1. Model ADR under different datasets.

Defense Methods	ADR
Masked dataset	2.28%
Masked + Original dataset	0.85%
CutMix dataset (Defense-GAN)	1.13%
Anti-GAN	5%

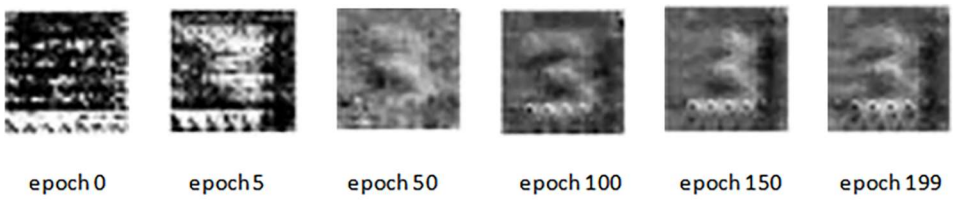

Figure 8. The attack effect of the FL training process using the masked image and the original image mixed dataset.

Although the accuracy of the FL model using the data set mixed with the masked image and the original image has increased (Table 1), the defense effect has also been greatly reduced. As can be seen from Figure 8, when the attack reaches the 100th round You can see a clearer image. This can lead to a breach of the defender's privacy.

4.4.4. Attack on the CutMix Dataset

To meet the requirements of effective defense against GAN attacks and model accuracy, this paper introduces CutMix into the proposed defense model and mixes each masked image proportionally with the original image to form a new image and combine these newly obtained images into a new dataset for FL. Based on this dataset, GAN is attacked, and the attack effect is shown in Figure 9.

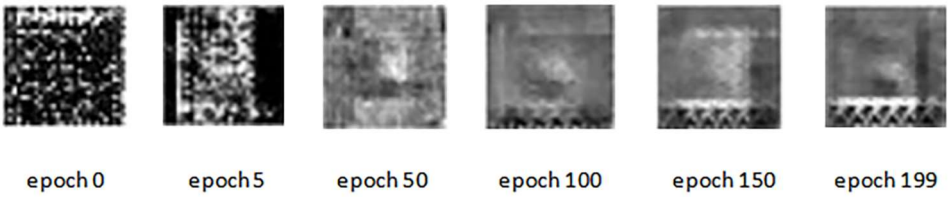

Figure 9. The attack effect of the FL training process using the masked image and the original image CutMix dataset.

It can be seen in Figure 9 that until the attack reaches the 199th round, the attacker still cannot generate a clear image, so the defense is successful. At the same time, on this basis, Defense-GAN also improves the accuracy of the model (Table 1), which can ensure that the FL model has good classification and recognition performance.

4.5. Model Accuracy Analysis

To verify that Defense-GAN can not only effectively defend against GAN attacks, but also ensure the accuracy of the model, this section will compare the use of a pure mask dataset, a new dataset composed of a part of the original image added to the mask dataset, and CutMix dataset (that is, Defense-GAN) and the defense Anti-GAN in reference [19] to obtain the model accuracy of FL. Here, we take the model accuracy obtained by training the original image as the original accuracy, and then calculate the accuracy degradation ratio (ADR) of each dataset. ADR is the ratio of the difference between the accuracy of the original image model and the accuracy of other image models to the accuracy of the original image model, and the smaller the value of ADR, the less affected the model is. The comparison results are shown in Table 1.

As can be seen from Table 1, the accuracy of the model trained with purely masked data is poor. Although the ADR of the data set training model formed by adding a part of the original image to the masked data set is small, as mentioned in Section 4.4.3, its defense effect is poor, and it will also leak the defender's privacy when the attack reaches a certain number of rounds. The defense model Defense-GAN proposed in this paper combines the masked image with the original image through CutMix to form a new data set. Compared with the model trained with the pure masked image data set, the ADR of the model improved by more than 1%. Compared with Anti-GAN, Defense-GAN's ADR is smaller, indicating that Defense-GAN has less accuracy loss than the model trained with the original image.

5. Application Prospects

FL has broad application prospects in many fields, however, in these application scenarios, there are security risks from GAN attacks. At the same time, Defense-GAN can be improved and expanded. It is not limited to the application scenarios of image recognition. Defense-GAN can also defend against GAN attacks in other FL scenarios. The following will introduce the security risks faced by each of these scenarios from facial recognition [24], medical imaging [25], and business [26], and explain the defensive role of Defense-GAN in them.

5.1. Facial Recognition

When FL is applied to computer vision tasks, the application of facial recognition is an extremely important field. When the victim's device holds a photo of a person V, the attacker can reconstruct V's appearance through GAN, or obtain V's facial features, and thus V's privacy will be leaked.

And Defense-GAN can help defend against such attacks. Under the function of Defense-GAN, because some features are covered, the attacker can only construct blurred photos. In other words, the photos generated by the attacker based on GAN cannot

be connected with the person in real life. So in this case, the privacy of the victim can be protected.

5.2. Medical Imaging

In medical AI, medical imaging pictures can often bring great value to the model. FL can enable various medical institutions to unite and share their own data in accordance with privacy protection regulations, breaking the bottleneck of insufficient data. However, when the patient data owner conducts FL training, the attacker can reconstruct the patient's medical images through GAN. For example, in a CT image of a lung nodule, an attacker can know the specific location of the patient's nodule through the reconstructed image, resulting in the disclosure of the patient's privacy.

Under the Defense-GAN architecture, attackers cannot obtain clear medical images, and cannot obtain effective information (such as the specific location of nodules and tumors), so the privacy of patients can be protected from being leaked.

5.3. Business

For enterprises with the same business but different distribution areas, the user groups often come from their respective areas, and the intersection is small, but due to the similarity of business, their user characteristics tend to converge. Due to the pressure of survival, many small and medium-sized organizations are more willing to join the federation to enhance the competitiveness of the industry. However, some enterprises may carry out malicious attacks during the FL process, using GAN to obtain samples with the same distribution as other enterprises' private data. In this way, the privacy of the enterprise will be leaked, which may cause huge losses.

In this case, Defense-GAN can also mask the training data, such as data replacement or data encryption (the specific method needs further research), to protect the privacy of the victim enterprise.

6. Conclusions

Since in the process of FL, user privacy, especially image privacy, is particularly vulnerable to GAN-based attacks, the attack process is carried out locally by the attacker and will not destroy the model itself, so from the victim's point of view, the existence of such an attack cannot be detected, hence, the victim's privacy is often leaked under unknown circumstances. This paper proposes a defense method Defense-GAN against GAN attacks under FL. By using SHAP to identify pixel features in each image that have a greater impact on classification and recognition, and then mask them, and then perform FL on the masked images, to protect the privacy of the defender's image training dataset; at the same time, this paper mixes the training images with CutMix, which improves the generalization ability of the model and ensures that the accuracy of the model trained with the masked images will not be affected too much.

Author Contributions: Conceptualization, X.M. and L.G.; methodology, X.M.; validation, X.M. and L.G.; data curation, X.M.; writing—original draft preparation, X.M.; writing—review and editing, X.M.; supervision, L.G.; project administration, L.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Not applicable.

Acknowledgments: The authors would like to thank the Institute of Cyberspace Security at the Beijing University of Posts and Telecommunications for providing the research platform and technical support.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Yang, Q.; Liu, Y.; Chen, T.; Tong, Y. Federated machine learning: Concept and applications. *ACM Trans. Intell. Syst. Technol.* **2019**, *10*, 1–19. [CrossRef]
2. McMahan, B.; Moore, E.; Ramage, D.; Hampson, S.; Arcas, B. Communication-efficient learning of deep networks from decentralized data. In Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, Ft. Lauderdale, FL, USA, 20–22 April 2017; pp. 1273–1282.
3. Chen, Y.; Gui, Y.; Lin, H.; Gan, W.; Wu, Y. Federated Learning Attacks and Defenses: A Survey. *arXiv* **2022**, arXiv:2211.14952.
4. Zhang, C.; Xie, Y.; Bai, H.; Yu, B.; Li, W.; Gao, Y. A survey on federated learning. *Knowl. -Based Syst.* **2021**, *216*, 106775. [CrossRef]
5. Feng, J.; Cai, Q.Z.; Zhou, Z.H. Learning to confuse: Generating training time adversarial data with auto-encoder. *Adv. Neural Inf. Process. Syst.* **2019**, *32*. [CrossRef]
6. Muñoz-González, L.; Biggio, B.; Demontis, A.; Paudice, A.; Wongrassamee, V.; Lupu, E.; Roli, F. Towards poisoning of deep learning algorithms with back-gradient optimization. In Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security, Dallas, TX, USA, 3 November 2017; pp. 27–38.
7. Sun, Z.; Kairouz, P.; Suresh, A.T.; McMahan, H.B. Can you really backdoor federated learning? *arXiv* **2019**, arXiv:1911.07963.
8. Bagdasaryan, E.; Veit, A.; Hua, Y.; Estrin, D.; Shmatikov, V. How to backdoor federated learning. In Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics, Palermo, Italy, 26–28 August 2020; pp. 2938–2948.
9. Li, Y.; Jiang, Y.; Li, Z.; Xia, S. Backdoor learning: A survey. In *IEEE Transactions on Neural Networks and Learning Systems*; IEEE: New York, NY, USA, 2022.
10. Luping, W.; Wei, W.; Bo, L.I. CMFL: Mitigating communication overhead for federated learning. In Proceedings of the 2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS), Dallas, TX, USA, 7–10 July 2019; pp. 954–964.
11. Yao, X.; Huang, C.; Sun, L. Two-stream federated learning: Reduce the communication costs. In Proceedings of the 2018 IEEE Visual Communications and Image Processing (VCIP), Taichung, Taiwan, 9–12 December 2018; pp. 1–4.
12. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial networks. *Commun. ACM* **2020**, *63*, 139–144. [CrossRef]
13. Hitaj, B.; Ateniese, G.; Perez-Cruz, F. Deep models under the GAN: Information leakage from collaborative deep learning. In Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, Dallas, TX, USA, 30 October–3 November 2017; pp. 603–618.
14. Zhang, J.; Chen, J.; Wu, D.; Chen, B.; Yu, S. Poisoning Attack in Federated Learning using Generative Adversarial Nets. In Proceedings of the 2019 18th IEEE International Conference on Trust, Security and Privacy in Computing and Communications/13th IEEE International Conference on Big Data Science and Engineering (TrustCom/BigDataSE), Rotorua, New Zealand, 5–8 August 2019.
15. Wang, Z.; Song, M.; Zhang, Z.; Song, Y.; Wang, Q.; Qi, H. Beyond inferring class representatives: User-level privacy leakage from federated learning. In Proceedings of the IEEE INFOCOM 2019-IEEE Conference on Computer Communications, Paris, France, 29 April 2019–2 May 2019; pp. 2512–2520.
16. Yao, A.C. Protocols for secure computations. In Proceedings of the 23rd Annual Symposium on Foundations of Computer Science (sfcs 1982), Chicago, IL, USA, 3–5 November 1982; pp. 160–164.
17. Aono, Y.; Hayashi, T.; Wang, L.; Moriai, S. Privacy-preserving deep learning via additively homomorphic encryption. *IEEE Trans. Inf. Forensics Secur.* **2017**, *13*, 1333–1345.
18. Li, J.; Gan, W.; Gui, Y.; Wu, Y.; Yu, P.S. Frequent itemset mining with local differential privacy. In Proceedings of the 31st ACM International Conference on Information & Knowledge Management, Atlanta, GA, USA, 17–22 October 2022; pp. 1146–1155.
19. Yan, X.; Cui, B.; Xu, Y.; Shi, P.; Wang, Z. A Method of Information Protection for Collaborative Deep Learning under GAN Model Attack. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2019**, *18*, 871–881. [CrossRef] [PubMed]
20. Luo, X.; Zhu, X. Exploiting Defenses against GAN-Based Feature Inference Attacks in Federated Learning. *arXiv* **2020**, arXiv:2004.12571.
21. Lundberg, S.M.; Lee, S.I. A unified approach to interpreting model predictions. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017.
22. Yun, S.; Han, D.; Oh, S.J.; Chun, S.; Choe, J.; Yoo, Y. Cutmix: Regularization strategy to train strong classifiers with localizable features. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6023–6032.
23. LeCun, Y. The MNIST Database of Handwritten Digits. 1998. Available online: http://yann.lecun.com/exdb/mnist/ (accessed on 20 August 2022).
24. Niu, Y.; Deng, W. Federated learning for face recognition with gradient correction. *Proc. AAAI Conf. Artif. Intell.* **2022**, *36*, 1999–2007. [CrossRef]

25. Stoffel, A.R.; Da, C.; KüderleArne Yari, I.A.; Eskofier, B. Federated Learning for Healthcare: Systematic Review and Architecture Proposal. *ACM Trans. Intell. Syst. Technol.* **2021**, *13*, 1–23.
26. Banabilah, S.; Aloqaily, M.; Alsayed, E.; Malik, N.; Jararweh, Y. Federated learning review: Fundamentals, enabling technologies, and future applications. *Inf. Process. Manag.* **2022**, *59*, 103061. [CrossRef]