CISTER
Research Centre in
Real-Time & Embedded
Computing Systems

Conference Paper

Exploring Adversarial Graph Autoencoders to Manipulate Federated Learning in The Internet of Things

Kai Li is a chair of AI FOR AUTONMOUS UNMANNED SYSTEMS SYMPOSIUM (AAUSS).

Kai Li*

Xin Yuan

Jingjing Zheng*

Wei Ni

Mohsen Guizani

*CISTER Research Centre

CISTER-TR-230301

2023/06/19

Exploring Adversarial Graph Autoencoders to Manipulate Federated Learning in The Internet of Things

Kai Li*, Xin Yuan, Jingjing Zheng*, Wei Ni, Mohsen Guizani

*CISTER Research Centre

Polytechnic Institute of Porto (ISEP P.Porto)

Rua Dr. António Bernardino de Almeida, 431

4200-072 Porto

Portugal

Tel.: +351.22.8340509, Fax: +351.22.8321159

E-mail: kai@isep.ipp.pt, xin.yuan@data61.csiro.au, zheng@isep.ipp.pt, Wei.Ni@data61.csiro.au

https://www.cister-labs.pt

Abstract

Mobile edge computing (MEC) enables the Internet of Things (IoT) with seamless integration of multiple application services. Federated learning is increasingly considered to improve training accuracy in MEC-IoT while circumventing the disclosure of private data, where the IoT nodes collaboratively train a machine learning model without disclosing their private data. In this paper, we propose a new cyber-epidemic attack that progressively manipulates federated learning and reduces the training accuracy of the benign MEC-IoT. The proposed cyber-epidemic attack explores adversarial graph autoencoders (GACE) to generate malicious local model updates that extract correlated features with the benign local and global models. The proposed GACE attack epidemically infects all the benign IoT nodes along with the training iterations in federated learning, while highly enhancing concealment of the attack.

Exploring Adversarial Graph Autoencoders to Manipulate Federated Learning in The Internet of Things

Kai Li*†, Xin Yuan‡, Jingjing Zheng*†, Wei Ni‡, and Mohsen Guizani§
*CISTER Research Centre, Portugal.
†CyLab Security and Privacy Institute, Carnegie Mellon University (CMU), USA.
Email: kaili@ieee.org & zheng@isep.ipp.pt.
‡Commonwealth Scientific and Industrial Research Organization (CSIRO), Australia.
Email: {xin.yuan,wei.ni}@csiro.au.
§MBZUAI, Abu Dhabi, United Arab Emirates.
Email: mguizani@ieee.org.

***Abstract*—Mobile edge computing (MEC) enables the Internet of Things (IoT) with seamless integration of multiple application services. Federated learning is increasingly considered to improve training accuracy in MEC-IoT while circumventing the disclosure of private data, where the IoT nodes collaboratively train a machine learning model without disclosing their private data. In this paper, we propose a new cyber-epidemic attack that progressively manipulates federated learning and reduces the training accuracy of the benign MEC-IoT. The proposed cyber-epidemic attack explores adversarial graph autoencoders (GACE) to generate malicious local model updates that extract correlated features with the benign local and global models. The proposed GACE attack epidemically infects all the benign IoT nodes along with the training iterations in federated learning, while highly enhancing concealment of the attack.**

***Index Terms*—Mobile edge computing (MEC), Internet of Things (IoT), federated learning, adversarial graph autoencoders, cyber-epidemic attacks**

I. Introduction

With the growing development of Internet of Things (IoT), mobile edge computing (MEC) is enabled to leverage powerful computing capabilities at an edge server for processing compute-intensive tasks offloaded by IoT nodes. The MEC-IoT is widely applied to a large number of applications, such as smart grids [1], intelligent transportation systems [2], and metaverse [3]. The IoT nodes upload their data to the edge server in which machine learning is used to train the IoT data. Nevertheless, this source data offloading is vulnerable to wireless attacks, such as eavesdropping [4], [5], denial of service [6], or blackhole attacks [7]. To avoid possible data privacy leakage, federated learning is studied to train a global shared model at the edge server, which aggregates local model updates instead of original training data of the IoT nodes.

Figure 1 describes an MEC-IoT system in which the edge server and the IoT nodes conduct federated learning for image classification, as an example. Specifically, the IoT node equipped with a video camera is deployed to track people's movements in a train station or airport, which generates large amounts of images. At the IoT node, the data is used to train a machine learning model, such as convolutional neural network (CNN) [8], long short-term memory (LSTM) [9], or support vector machine (SVM) [10], that produce a local model update (i.e., the weight vector of the machine learning model) [11]. The local model updates of the IoT nodes are sent to an edge server at which a global model is created by averaging the weight vectors. Next, the global model is sent back to the IoT nodes that can adjust the weight vectors in their machine learning models according to the updated parameters in the global model. Thanks to the collaborative training of the global and local models, federated learning iteratively improves the training accuracy, leading to an accurate people movement' tracking. In addition, the global model and local model updates of federated learning are trained without collecting the private data from the IoT nodes, thus protecting data privacy.

Despite federated learning can prevent the data privacy leakage, a number of data or model poisoning attacks [12] are studied against federated learning, all of which aim to manipulate the local and global model training. In particular, existing data or model poisoning attacks enable the adversarial IoT node to create a malicious local model update with falsified weight vectors or fake training datasets to compromise the learning accuracy. Generally, the existing attacks against federated learning prevent being detected by ensuring that an Euclidean distance between the malicious local model update and the global model is smaller than a predetermined threshold. Unfortunately, such attacks can still be detected by the recent poisoning defense countermeasures, which utilize the feature correlation of the weight vectors of all benign local model updates to learn abnormal fluctuation among the IoT nodes.

In this paper, a new adversarial graph autoencoders-based cyber-epidemic (GACE) attack is proposed to progressively

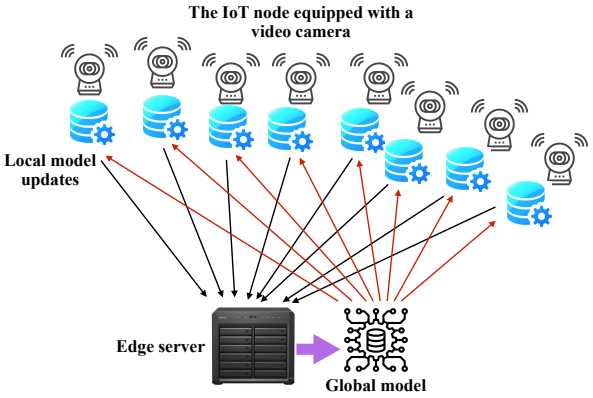

Fig. 1: The MEC-IoT system, where a machine learning model is trained at the IoT node to produce a local model update. The edge server aggregates all the local model updates from the IoT nodes to train a global model.

manipulate federated learning and compromise the training accuracy of the benign MEC-IoT devices. The proposed attack explores adversarial graph autoencoders (GAE) to generate a malicious local model update that holds a strong feature correlation with the benign local model updates and the global model overheard. The adversarial GAE at the attacker constructs an adjacency matrix based on the overheard benign local model updates and the global model as the input, and outputs a reconstructed adjacency matrix, with the reconstruction loss maximized. According to the reconstructed adjacency matrix, the attacker reconstructs the malicious local model update that manipulates federated learning convergence while containing correlated features with the benign ones.

The GACE attack guarantees that the Euclidean distance between the reconstructed malicious local model update and the global model below a threshold [13]. Next, the attacker uploads the malicious local model update to the edge server, which aims to maximize the training loss of federated learning and diverge the federated learning performance. As a result, the training accuracy of federated learning gradually decreases. Since the malicious local model updates are uploaded to the edge server for training the global model, the proposed GACE attack gives rise to an epidemical infection over all the benign local model updates.

The rest of this paper is organized as follows. Section II discusses related work about exiting adversarial attacks against federated learning. Section III discusses federated learning mechanism with the benign IoT nodes and the edge server, as well as the eavesdropping model. In Section IV, we study the proposed GACE attack. Section V evaluates the performance. Section VI concludes the paper.

II. Related Work

In this section, we present the related studies about the exiting adversarial attacks to federated learning in MEC-IoT. We also discuss the new characteristics of the proposed GACE attack compared to the attacks in the literature.

A data poisoning attack against federated learning is studied in [14]. A number of adversarial nodes generate mislabeled data and upload their malicious local model updates to the server, to poison the global model. The data poisoning attack results in substantial drops in the training accuracy of data classification, even with a small percentage of malicious local model updates. By analyzing several data poisoning attacks according to poisoning and adversary capabilities, an attack is developed based on the classic label flipping data poisoning [15]. Their data poisoning attack can adjust the amount of label flipped data used to evade the model aggregation rule at the edge server, since the malicious local model update generated with a large volume of fake data is easy to be detected. In [16], a data poisoning attack is studied, where a backdoor is embedded into the malicious local model update. Once the malicious local model update is aggregated by the edge server, the global model can be falsely trained, which identifies the malicious local model as a benign one. To generate a persistent data poisoning attack against federated learning, the authors of [17] analyze model capacity so that the model poisoning attack can inject adversarial neurons in the redundant space of a neural network. Since the redundant neurons have limited correlations to the main task of federated learning, their poisoning attack does not affect the global model training at the edge server.

Generative adverserial networks (GANs) can be used to build the poisoning attack against federated learning with MEC [18], which obtains hidden features from the local model updates and recovers the local data. The GANs-based poisoning attack applies the support vector machine (SVM) model to generate the malicious data and local model updates via imitating the generation of benign data and local models. Different from the random data selection as the local training data in federated learning, the GANs developed in [19] allocates a distinct data class to the local node for creating the privacy leakage attack. The GAN is developed to balance the learning rate of the discriminator and the one of the generator. By measuring the Euclidean distance between the real data and its reconstructed malicious peer, performance of the attacks, such as training accuracy and detection rate, can be evaluated. In [20], a GAN-based poisoning attack is presented, where the adversarial node disguises itself as a benign node. The GAN is trained to generate the malicious data via imitating the benign one while the adversarial node flips the labels of the GAN generated data.

The exiting data poisoning and GAN-based poisoning attacks lack the description of the implicit relationship between different local model updates, which can be detected by recent poisoning defense strategies in federated learning [21], [22]. In addition, the output features of the attacks can be oversmoothed by multiple convolutional layers at the edge server, making the differences between

the malicious local model update and the benign ones distinguishable. This exceptionally alleviates the malicious epidemic threat to federated learning.

III. Federated Learning in MEC-IoT Systems

In this section, we present the training of the local model update and the global model in the MEC-IoT system.

A. Federated learning formulation

Federated learning allows the IoT nodes to independently train the data without sharing to the other nodes and the edge server. The sensory data, e.g., images captured by the camera in Figure 1, is trained at the IoT node to generate the local model update. The edge server aggregates all the local model updates for training the global model, which is used to classify and analyze the captured images. With an increase of the training iterations, federated learning can gradually improve the image classification accuracy.

In the MEC-IoT system, the benign IoT node i generates D_i datasets at time t, where $i \in [1, N]$ and N denotes the total number of benign IoT nodes. We denote a_i and y_i as the input images of federated learning and output (e.g., the classified images) at node i, respectively. At the IoT node, the local model update is trained by neural networks to minimize a loss function $L_i(\boldsymbol{w}_i; a_i, y_i)$ that measures approximation errors over a_i and y_i, where $\boldsymbol{w}_i$ is the model parameter of IoT node i. For instance, $L_i(\boldsymbol{w}_i; a_i, y_i)$ can be given by $L_i(\boldsymbol{w}_i; a_i, y_i) = 1/2({a_i}^T \boldsymbol{w}_i - y_i)$, for linear regression; or $L_i(\boldsymbol{w}_i; a_i, y_i) = -\log(1 + \exp({a_i}^T y_i \boldsymbol{w}_i))$, for logistic regression. Thus, we can define the loss function as

$$L_{\text{loss}}(\boldsymbol{w}_i) := \frac{1}{D_i} \sum_{i=1}^{D_i} L_i(\boldsymbol{w}_i; a_i, y_i) + b \cdot reg(\boldsymbol{w}_i), \quad (1)$$

where $b \in [0, 1]$. The effect of the local training noise is described by a regularizer function $reg(\cdot)$ [23].

The local model updates of the IoT nodes are aggregated at the edge server to create a comprehensive and effective global model. Let $\boldsymbol{w}_{glb}$ represent the global model. We have

$$\boldsymbol{w}_{glb} = \frac{1}{D(N)} \sum_{i=1}^{N} D_i L_{\text{loss}}(\boldsymbol{w}_i), \quad (2)$$

where $D(N) = \sum_{i=1}^{N} D_i$ presents the total data size of all the IoT nodes.

Minimizing $L_{\text{loss}}(\boldsymbol{w}_i)$ can be formulated as

$$\min_{\boldsymbol{w}_i} L_{\text{loss}}(\boldsymbol{w}_i) := \frac{1}{D(N)} \sum_{i=1}^{N} \sum_{a_i=1}^{D_i} L_i(\boldsymbol{w}_i; a_i, y_i) + b \cdot reg(\boldsymbol{w}_i) \quad (3)$$

B. System and eavesdropping models

1) System model of MEC-IoT: Let $p_i(t) = (x_i(t), y_i(t), z_i(t))$ denote the position of node i. The distance between the IoT node and the edge server can be expressed as

$$\begin{aligned} d_i(t) &= ||p_i(t) - p_{\text{edge}}|| \\ &= \sqrt{(x_i(t) - x_{\text{edge}})^2 + (y_i(t) - y_{\text{edge}})^2 + z_i(t)^2}, \end{aligned} \quad (4)$$

where $p_{\text{edge}} = (x_{\text{edge}}, y_{\text{edge}}, 0)$ is the location of the edge server [24].

Let h_i denote the channel gain between the i-th benign IoT node and the edge server. We have $h_i(t) = \frac{P_0}{d_i(t)^2}$, where P_0 is a reference transmit power of an IoT node at the distance $d_i = 1$ m. Furthermore, the signal-to-noise ratio (SNR) of the channel between the IoT node and the edge server can be given by $\eta_i(t) = \frac{h_i(t) P_i(t)}{\sigma_0^2}$, where $P_i(t)$ is the transmit power of node i, and σ_0^2 is the noise power at the edge server.

2) The attacker's eavesdropping model: Likewise, the channel gain of the eavesdropping link, i.e., between a benign IoT node and the attacker, is given by

$$h'_i(t) = \frac{\varrho_4 h'(t-1) + \varrho_3 \sqrt{1 - \varrho_4^2}}{d'_i(t)^{\varrho_5}}, \quad (5)$$

where ϱ_3 is a Gaussian random variable, and ϱ_4 is the coefficient adjusting the weights of the two components. ϱ_5 denotes the path-loss exponent. $d'_i(t)$ defines the distance between the IoT node i and the attacker.

The SNR of the eavesdropping link, denoted by $\eta'_i(t)$, is

$$\eta'_i(t) = \frac{h'_i(t) P_i(t)}{\sigma_0^2}. \quad (6)$$

In addition, the channel gains h_i, h'_i and σ_0^2 are known to the attacker at the beginning of time t, since the attacker can overhear the benign IoT nodes' channels and the eavesdropping links via channel probing.

IV. Adversarial Graph Autoencoders-based Cyber-epidemic Attacks

In this section, we propose the GACE attack that aims to progressively manipulate federated learning in MEC-IoT. The attacker explores the adversarial GAE to create the malicious model update, which extracts the feature correlation among the benign local model updates.

Let $\boldsymbol{w}'_k$ denote the malicious local model update created by the attacker k. In particular, the arbitrary features of $\boldsymbol{w}'_k$ and the ones in the benign local model updates could be highly irrelevant. This leads to a low feature correlation between the malicious local model update and the benign ones, which can still be detected by the edge server. To address this, the GACE attack is presented in Figure 2, where the global model $\boldsymbol{w}_{glb}$ and the benign local model updates (i.e., $\boldsymbol{w}_1, \boldsymbol{w}_2, ..., \boldsymbol{w}_i, ..., \boldsymbol{w}_N$) are eavesdropped on by the attacker. Next, the attacker disguises as one of the benign IoT nodes and creates a malicious local model update being uploaded to the edge server. The edge server aggregates all $\boldsymbol{w}_i$ for training $\boldsymbol{w}_{glb}$, which is progressively contaminated along with federated learning iterations. Eventually, the

GACE attack compromises all the benign IoT nodes and manipulates federated learning training.

The attacker aims to iteratively construct $\boldsymbol{w}'_k$ based on the eavesdropped $\boldsymbol{w}_{glb}$ and $\boldsymbol{w}_i$ in which the attacker has no knowledge of the IoT data (a_i, y_i). A graph $G(\mathcal{V}, \mathcal{E}, \boldsymbol{\mathcal{F}})$ is formulated at the attacker to train the malicious local model update, where $\mathcal{V}$, $\mathcal{E}$ and $\boldsymbol{\mathcal{F}}$ represent vertexes, edges and a feature matrix of the graph, respectively. Let $\boldsymbol{A} = \{\overline{\boldsymbol{w}}_{(i,i')}|_{i,i'=1}^{N}\} \in \mathbb{R}^{N\times N}$ define an adjacent matrix indicating the correlation among the benign local model updates, where $\overline{\boldsymbol{w}}_{(i,i')}$ is the inner product between $\boldsymbol{w}_i$ and $\boldsymbol{w}_{i'}$. Since the topological structure of the graph is built according to the adjacency matrix, $\overline{\boldsymbol{w}}_{(i,i')} = 1$ if there is an edge between $\boldsymbol{w}_i$ and $\boldsymbol{w}_{i'}$; $\overline{\boldsymbol{w}}_{(i,i')} = 0$, otherwise. The feature matrix $\boldsymbol{\mathcal{F}} = \{\boldsymbol{w}_1, \boldsymbol{w}_2, ..., \boldsymbol{w}_i\}$, where $\boldsymbol{w}_i$ contains the feature of the benign local model updates with regard to the data dimension.

The GACE attack is based on the adversarial GAE that consists of an encoder and a decoder. The encoder encodes the graph data with the features and the decoder takes the encoder's output as the input to reconstruct the graph. Specifically, the encoder takes the input matrixes of $\boldsymbol{\mathcal{F}}$ and $\boldsymbol{A}$. The encoder is built based on a K-layer GCN. The output of convolution at the k-th layer can be presented as

$$\boldsymbol{Z}^k = S(\boldsymbol{Z}^{k-1}, \boldsymbol{A}|\boldsymbol{\gamma}^k), \tag{7}$$

where $S(\cdot)$ is a spectral convolution function and $\boldsymbol{\gamma}^k$ defines the weight matrix at the k-th layer.

Given an identify matrix $\mathcal{I}$, we define $\widetilde{\boldsymbol{A}} = \boldsymbol{A} + \mathcal{I}$ and $\overline{\boldsymbol{A}}_{ii} = \sum_{i'} \widetilde{\boldsymbol{A}}_{ii'}$. To generate a feature representation of the graph, the encoder is formulated as

$$S(\boldsymbol{Z}^{k-1}, \boldsymbol{A}|\boldsymbol{\gamma}^k) = \mathcal{S}^k(\overline{\boldsymbol{A}}^{-\frac{1}{2}}\widetilde{\boldsymbol{A}}\overline{\boldsymbol{A}}^{-\frac{1}{2}}\boldsymbol{Z}^{k-1}\boldsymbol{\gamma}^k), \tag{8}$$

where $\mathcal{S}^k(\cdot)$ represents a nonlinear activation function (e.g., $tanh(\cdot)$ or $ReLU(\cdot)$). $\overline{\boldsymbol{A}}^{-\frac{1}{2}}\widetilde{\boldsymbol{A}}\overline{\boldsymbol{A}}^{-\frac{1}{2}}$ is the symmetrically normalized adjacency matrix.

The input to the decoder is $\boldsymbol{Z}^K$, which is the last layer output of the GCN. Let $\boldsymbol{Z}^{\mathrm{T}}$ denote the transpose matrix of the $\boldsymbol{Z}^K$. A reconstructed adjacency matrix is generated at the decoder, which is defined as

$$\widehat{\boldsymbol{A}} = \mathrm{sigmoid}(\boldsymbol{Z}^K \ \boldsymbol{Z}^{\mathrm{T}}), \tag{9}$$

where $\mathrm{sigmoid}(x) = 1/(1+e^{-x})$ is the sigmoid function. In addition, the larger the inner product $(\boldsymbol{Z}^k \ \boldsymbol{Z}^{\mathrm{T}})$ in the embedding, the more likely vertexes i and i' are connected in the graph according to the autoencoder. The output of the decoder is the reconstructed adjacency matrix $\widehat{\boldsymbol{A}}$ as well as $\boldsymbol{\mathcal{F}}^* = \{\boldsymbol{w}_1^{\Delta}, \boldsymbol{w}_2^{\Delta}, ..., \boldsymbol{w}_i^{\Delta}\}^*$, where $\boldsymbol{w}_i^{\Delta}$ denotes the reconstructed model update used to generate the malicious model update.

Here, $G(\mathcal{V}, \mathcal{E}, \boldsymbol{\mathcal{F}})$ measures the similarity between $\boldsymbol{A}$ and $\widehat{\boldsymbol{A}}$. A weighted cross entropy loss is used to describe its reconstruction loss η. It defines

$$\eta = \mathbb{E}_{S(\boldsymbol{Z}^{k-1}, \boldsymbol{A}|\boldsymbol{\gamma}^k)}\left[\frac{\log\ p(\ \widehat{\boldsymbol{A}}\ |\ \boldsymbol{Z}^k\ )}{\mathcal{C}(\frac{\sum_{i=1}^{N}\boldsymbol{w}_i^{\Delta}}{N}, \boldsymbol{w}_{glb}^{\Delta})}\right], \tag{10}$$

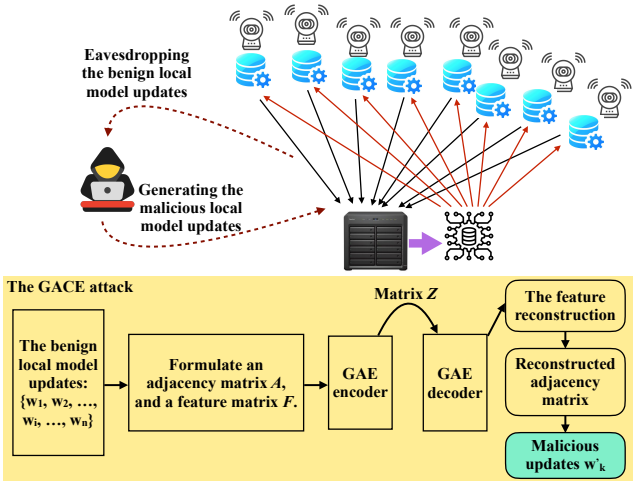

Fig. 2: The proposed GACE attack that aims to progressively manipulate federated learning in MEC-IoT.

where $\mathcal{C}(\cdot)$ calculates the Euclidean distance between the two input vectors. $p(\ \widehat{\boldsymbol{A}}\ |\ \boldsymbol{Z}^k\ )$ at the decoder indicates the correlation among the embedding vertexes, which is

$$p(\ \widehat{\boldsymbol{A}}\ |\ \boldsymbol{Z}^k\ ) = \Pi_i^N \Pi_{i'}^N p(\ \widehat{\boldsymbol{A}}_{ii'}\ |\ \boldsymbol{Z}_i^k, \boldsymbol{Z}_{i'}^k\ ), \tag{11}$$

and

$$p(\ \widehat{\boldsymbol{A}}_{ii'} = 1\ |\ \boldsymbol{Z}_i^k, \boldsymbol{Z}_{i'}^k\ ) = \mathrm{sigmoid}(\boldsymbol{Z}_i^k\ \boldsymbol{Z}_{i'}^{\mathrm{T}}). \tag{12}$$

Since the attacker aims to generate the malicious local model updates $\boldsymbol{w}'_k$ to disorient federated learning, GACE is designed to maximize η in (10), which leads to a minimized $\mathcal{C}(\frac{\sum_{i=1}^{N}\boldsymbol{w}_i^{\Delta}}{N}, \boldsymbol{w}_{glb}^{\Delta})$. Moreover, maximizing $p(\ \widehat{\boldsymbol{A}}\ |\ \boldsymbol{Z}^k\ )$ satisfies $\mathcal{C}(\frac{\sum_{i=1}^{N}\boldsymbol{w}_i^{\Delta}}{N}, \boldsymbol{w}_{glb}^{\Delta}) \leq \mathcal{C}_{th}$, where $\mathcal{C}_{th}$ is a threshold value that ensures the generated $\boldsymbol{w}'_k$ is close to $\boldsymbol{w}_{glb}^{\Delta}$ in the Euclidean space. As a result, $\boldsymbol{w}'_k$ iteratively contaminates federated learning without being detected.

A Laplacian matrix $\boldsymbol{\Omega}$ is formulated at the attacker based on the adjacency matrix of $\boldsymbol{A}$, which is $\boldsymbol{\Omega} = diag(\boldsymbol{A}) - \boldsymbol{A}$. By applying singular value decomposition to $\boldsymbol{\Omega}$, a complex unitary matrix can be obtained $\boldsymbol{B} \in \mathbb{R}^{N\times N}$. Therefore, the attacker can formulate a matrix $\boldsymbol{B}'$ that contains the features of the benign model updates, namely, $\boldsymbol{B}' = \boldsymbol{B}^{-1}\boldsymbol{\mathcal{F}}$.

In this case, the attacker can also formulate a Laplacian matrix based on the output of the adversarial GAE, which is $\widehat{\boldsymbol{\Omega}} = diag(\widehat{\boldsymbol{A}}) - \widehat{\boldsymbol{A}}$. In addition, the complex unitary matrix $\widehat{\boldsymbol{B}}$ is obtained by applying the singular value decomposition to $\widehat{\boldsymbol{\Omega}}$. Therefore, the malicious local model update $\boldsymbol{w}'_k$ that follows the design of $\boldsymbol{A}$ in the adversarial GAE can be constructed, which is

$$\widehat{\boldsymbol{\mathcal{F}}} = \widehat{\boldsymbol{B}}\boldsymbol{B}', \tag{13}$$

where $\widehat{\boldsymbol{\mathcal{F}}}$ is the feature matrix containing all the malicious local model updates, $\widehat{\boldsymbol{\mathcal{F}}} = \{\boldsymbol{w}'_1, \boldsymbol{w}'_2, ..., \boldsymbol{w}'_k\}$.

According to the design of the GACE attack in Figure 2, the edge server broadcasts $\boldsymbol{w}_{glb}$ in every communication

round of federated learning. The benign node performs the local training for the local model update $\boldsymbol{w}_i$. The attacker eavesdrops on the global model $\boldsymbol{w}_{glb}$ and the local model update $\boldsymbol{w}_i$ of the benign IoT nodes. Given a number of federated learning iterations, the adversarial GAE is trained to maximize the reconstruction loss with the adjacent matrix $\boldsymbol{A}$ and the feature matrix $\boldsymbol{\mathcal{F}}$. At the output of the adversarial GAE, the attacker achieves the optimal malicious local model update, i.e., $\boldsymbol{w}'_k$. Next, $\boldsymbol{w}'_k$ is uploaded to the edge server for the next round of federated learning training.

V. Performance Analysis

The proposed GACE attack is implemented in PyTorch. Modified national institute of standards and technology (MNIST) database that contains a large number of handwritten digits are used to train federated learning at the IoT nodes. Moreover, we compare the average training accuracy of FL under the proposed GACE attack and a BackdoorFL attack [25]. The attacker with BackdoorFL creates an arbitrary local model update that contains a malicious backdoor for poisoning the global model update.

Figure 3 shows the training accuracy of the global model at the edge server with or without the proposed GACE attack. Given 100 federated learning episodes, the training accuracy of the benign global model (without the GACE attack) gradually converges to 99.6%. Under the GACE attack, the training accuracy convergence regarding the global model is deviated, which drastically fluctuates between 7.5% and 89.2%. This indicates that the GACE attack has effectively compromised all the benign IoT nodes and manipulated federated learning training.

Given five benign IoT nodes as an example, Figure 4 presents the training accuracy of the local model updates at each of the IoT nodes. Due to the GACE attack, the training accuracy of the local model updates at the IoT nodes drops about 27.3% at maximum according to federated learning episodes. This confirms that the malicious local model update generated by the GACE attack successfully contaminates the benign ones.

At the edge server, the Euclidean distance between the local model updates and the global model can be measured in order to detect the malicious local model update. In this case, we set node 3 as the attacker, and the Euclidean distance is presented in Figure 5 to validate that the GACE attack is undetectable. As observed, the Euclidean distance of all the five nodes randomly varies between 8×10^{-3} and 1×10^{-3}, while the Euclidean distance of the malicious local model update generated at node 3 hides among the benign local model updates. Thus, it is hard for the edge server to identify the malicious local model update. The GACE attack achieves this because the adversarial GAE extracts the feature correlation among the benign local model updates.

In Table I, we compare the average training accuracy of FL under the proposed GACE attack and the BackdoorFL

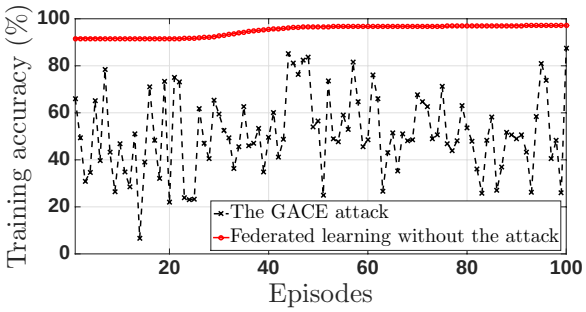

Fig. 3: Under the GACE attack, the training accuracy of the global model at the edge server given 100 federated learning episodes.

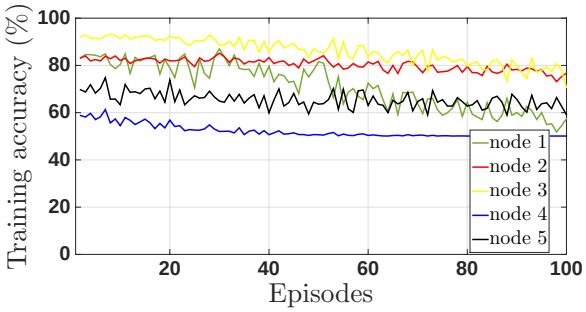

Fig. 4: Under the GACE attack, the training accuracy of the local model updates at the IoT nodes given 100 episodes.

attack, where the number of devices, i.e., N, increases from 5 to 25. In particular, the GACE attack reduces the training accuracy of FL for about 12% lower than BackdoorFL since the adversarial GAE maximizes the reconstruction loss η, thus maximizing $L_{\text{loss}}(\boldsymbol{w}_i)$. Table II shows the

TABLE I: The average training accuracy under attacks

N	GACE	BackdoorFL
5	79.8%	92.3%
10	82.7%	94.7%
15	81.3%	93.4%
20	88.5%	99.1%
25	92.1%	99.8%

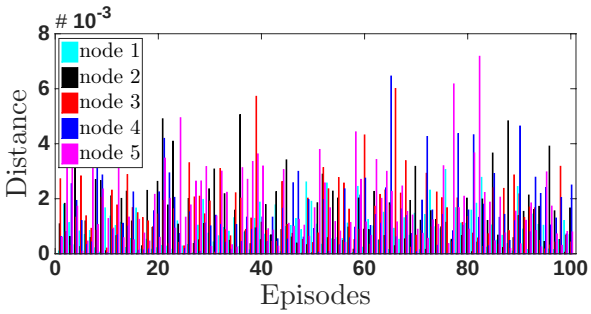

Fig. 5: The Euclidean distance between the local model updates and the global model under the GACE attack.

TABLE II: The average detection rate of the attack

N	GACE	BackdoorFL
5	0.8%	15.1%
10	0.7%	38.3%
15	0.9%	48.6%
20	1.1%	75.9%
25	1.5%	96.3%

detection rate of the GACE attack and the BackdoorFL attack. It can be observed that GACE is hardly detected by the edge server, compared to BackdoorFL. This is because the malicious local model update of GACE maintains a strong feature correlation with the benign models, while the update of BackdoorFL is arbitrarily generated.

VI. Conclusion

This paper proposed a new GACE attack that progressively manipulates federated learning process of some benign IoT devices in an MEC-IoT system. The GACE attack explores the adversarial GAE to generate the malicious local model updates that hold feature correlations with the benign models, which improves the concealment of the malicious local model update. As a result, the benign IoT nodes are epidemically infected along with the training iterations in federated learning, where the training loss of the benign IoT devices increases considerably.

Acknowledgements

This work was supported by the CISTER Research Unit (UIDP/UIDB/04234/2020), project ADANET (PTDC/EEICOM/3362/2021) and project IBEX (PTDC/CCI-COM/4280/2021), financed by National Funds through FCT/MCTES (Portuguese Foundation for Science and Technology).

References

[1] M. Laroui, B. Nour, H. Moungla, M. A. Cherif, H. Afifi, and M. Guizani, "Edge and fog computing for IoT: A survey on current research activities & future directions," *Computer Communications*, vol. 180, pp. 210–231, 2021.

[2] K. Xiong, S. Leng, C. Huang, C. Yuen, and Y. L. Guan, "Intelligent task offloading for heterogeneous V2X communications," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 4, pp. 2226–2238, 2020.

[3] K. Li, Y. Cui, W. Li, T. Lv, X. Yuan, S. Li, W. Ni, M. Simsek, and F. Dressler, "When internet of things meets metaverse: Convergence of physical and cyber worlds," *IEEE Internet of Things Journal*, 2022.

[4] K. Li, W. Ni, J. Zheng, E. Tovar, and M. Guizani, "Confidentiality and timeliness of data dissemination in platoon-based vehicular cyber-physical systems," *IEEE Network*, vol. 35, no. 4, pp. 248–254, 2021.

[5] K. Li, R. C. Voicu, S. S. Kanhere, W. Ni, and E. Tovar, "Energy efficient legitimate wireless surveillance of UAV communications," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 3, pp. 2283–2293, 2019.

[6] V. Borgiani, P. Moratori, J. F. Kazienko, E. R. Tubino, and S. E. Quincozes, "Toward a distributed approach for detection and mitigation of denial-of-service attacks within industrial internet of things," *IEEE Internet of Things Journal*, vol. 8, no. 6, pp. 4569–4578, 2020.

[7] A. Alwarafy, K. A. Al-Thelaya, M. Abdallah, J. Schneider, and M. Hamdi, "A survey on security and privacy issues in edge-computing-assisted internet of things," *IEEE Internet of Things Journal*, vol. 8, no. 6, pp. 4004–4022, 2020.

[8] A. Noor, K. Li, A. Ammar, A. Koubaa, B. Benjdira, and E. Tovar, "A hybrid deep learning model for UAVs detection in day and night dual visions," in *IEEE International Conference on Cognitive Machine Intelligence (CogMI)*. IEEE, 2021, pp. 221–231.

[9] K. Li, W. Ni, and F. Dressler, "LSTM-characterized deep reinforcement learning for continuous flight control and resource allocation in UAV-assisted sensor network," *IEEE Internet of Things Journal*, vol. 9, no. 6, pp. 4179–4189, 2021.

[10] S. Makkar and L. Sharma, "A face detection using support vector machine: Challenging issues, recent trend, solutions and proposed framework," in *International Conference on Advances in Computing and Data Sciences*. Springer, 2019, pp. 3–12.

[11] J. Zheng, K. Li, E. Tovar, and M. Guizani, "Federated learning for energy-balanced client selection in mobile edge computing," in *International Wireless Communications and Mobile Computing (IWCMC)*. IEEE, 2021, pp. 1942–1947.

[12] B. Ghimire and D. B. Rawat, "Recent advances on federated learning for cybersecurity and cybersecurity for federated learning for internet of things," *IEEE Internet of Things Journal*, 2022.

[13] X. Li, Z. Qu, S. Zhao, B. Tang, Z. Lu, and Y. Liu, "Lomar: A local defense against poisoning attack on federated learning," *IEEE Transactions on Dependable and Secure Computing*, 2021.

[14] V. Tolpegin, S. Truex, M. E. Gursoy, and L. Liu, "Data poisoning attacks against federated learning systems," in *European Symposium on Research in Computer Security*. Springer, 2020, pp. 480–501.

[15] V. Shejwalkar, A. Houmansadr, P. Kairouz, and D. Ramage, "Back to the drawing board: A critical evaluation of poisoning attacks on production federated learning," in *2022 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2022, pp. 1354–1371.

[16] T. D. Nguyen, P. Rieger, M. Miettinen, and A.-R. Sadeghi, "Poisoning attacks on federated learning-based IoT intrusion detection system," in *Workshop Decentralized IoT System Security (DISS)*, 2020, pp. 1–7.

[17] X. Zhou, M. Xu, Y. Wu, and N. Zheng, "Deep model poisoning attack on federated learning," *Future Internet*, vol. 13, no. 3, p. 73, 2021.

[18] P. Manoharan, R. Walia, C. Iwendi, T. A. Ahanger, S. Suganthi, M. Kamruzzaman, S. Bourouis, W. Alhakami, and M. Hamdi, "Svm-based generative adverserial networks for federated learning and edge computing attack model and outpoising," *Expert Systems*, p. e13072, 2022.

[19] Y. Sun, N. S. Chong, and H. Ochiai, "Information stealing in federated learning systems based on generative adversarial networks," in *IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 2021, pp. 2749–2754.

[20] J. Zhang, J. Chen, D. Wu, B. Chen, and S. Yu, "Poisoning attack in federated learning using generative adversarial nets," in *IEEE TrustCom/BigDataSE*. IEEE, 2019, pp. 374–380.

[21] Z. Zhang, X. Cao, J. Jia, and N. Z. Gong, "Fldetector: Defending federated learning against model poisoning attacks via detecting malicious clients," in *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2022, pp. 2545–2555.

[22] A. Manna, H. Kasyap, and S. Tripathy, "Moat: Model agnostic defense against targeted poisoning attacks in federated learning," in *International Conference on Information and Communications Security*. Springer, 2021, pp. 38–55.

[23] J. Zheng, K. Li, N. Mhaisen, W. Ni, E. Tovar, and M. Guizani, "Exploring deep-reinforcement-learning-assisted federated learning for online resource allocation in privacy-preserving edgeiot," *IEEE Internet of Things Journal*, vol. 9, no. 21, pp. 21 099–21 110, 2022.

[24] K. Li, W. Ni, X. Wang, R. P. Liu, S. S. Kanhere, and S. Jha, "Energy-efficient cooperative relaying for unmanned aerial vehicles," *IEEE Transactions on Mobile Computing*, vol. 15, no. 6, pp. 1377–1386, 2015.

[25] M. Asad, A. Moustafa, and C. Yu, "A critical evaluation of privacy and security threats in federated learning," *Sensors*, vol. 20, no. 24, p. 7182, 2020.