A Federated Learning Based Privacy-Preserving Data Sharing Scheme for Internet of Vehicles

Yangpeng Wang, Ling Xiong(✉), Xianhua Niu, Yunxiang Wang, and Dexin Liang

School of Computer and Software Engineering, Xihua University, Chengdu, China
lingdonghua99@163.com

Abstract. The data analysis in the process of vehicle collaboration for the Internet of Vehicles (IoV) environment improves the driving experience and service quality. However, the privacy issue is becoming one of the problems of obstructing the development of data sharing among vehicles. To overcome the disadvantage, in this work, we propose a privacy-preserving data sharing scheme based on federated learning by the collaboration of participants, which can resist gradient leakage, poisoning attacks, etc. Firstly, the gradient data is encrypted by random masking to protect the privacy of training data. Then, the Pearson correlation coefficient is utilized to distinguish the correctness of the model parameters uploaded from the vehicle at uplink. Finally, the proposed scheme can verify the correctness of the global model distributed from AS at downlink using the Lagrange interpolation The experimental results show that the proposed privacy-preserving data sharing scheme provides higher learning accuracy by eliminating malicious gradients.

Keywords: Federated learning · Privacy-preserving · IoV · Data sharing

1 Introduction

With the rapid development of intelligent transportation system, new computing methods have been widely deployed on Internet of Vehicles. In particular, the integration of IoV and artificial intelligence has led to a trend of sharing data among vehicles and infrastructure. The shared data usually includes trajectories, surrounding information and operation information, etc. To improve the driving experience and service quality, vehicles could utilize the shared data. For example, vehicles could generate a region traffic flow model according to the shared data. However, in the process of sharing data, the data privacy of the vehicles will be damaged, which may lead to serious consequences.

In the privacy-sensitive scenario of the IoV, to avoid privacy disclosure caused by sharing data, the federated learning(FL) [1,2] mechanism is applied to the

Supported by the National Natural Science Foundation of China (No. 62171387, No.62202390).

E. Ahene and F. Li (Eds.): FCS 2022, CCIS 1726, pp. 18–33, 2022.
https://doi.org/10.1007/978-981-19-8445-7_2

IoV. The existing achievements [3–5] have well implemented the scheme that the vehicle privacy data is not sent directly in the public channel. However, as previous work [6] had shown that, the private information of vehicle can also be leaked from the gradient. Considering such security vulnerabilities, if a aggregation server and other entities obtain sufficient gradients, the privacy data(e.g. vehicle position information and trajectories information) of participants will be seriously threatened. Another concern is that the gradient may come from malicious participants. On the one hand, the malicious vehicles upload incorrect gradients to lead to a decline in the accuracy of the model, and even make the final global model unavailable. On the other hand, the AS may forge the aggregation model parameters, if the participant cannot recognize the modified global model, the entire FL process will be destroyed, even leading to a serious threat to traffic safety.

In this paper, we address the data privacy leakage issue by integrating federated learning into IoV, on this basis, we use masks to encrypt the model gradient and remove the incorrect gradients without knowing the gradient. Finally we verify the correctness of the global model. The contributions of the paper can be summarized as follows.

- The proposed scheme divides the vehicles into multiple groups containing an appropriate number of vehicles, and vehicles in the same group use the negotiated mask to encrypt the model gradient. Moreover, secret share is adopt to recover the mask of leaved vehicles.
- The proposed scheme distinguishs the correctness of gradients by using the Pearson correlation coefficient when receiving the uploaded gradients from vehicle. Then the Lagrange interpolation is adopted to verify the global aggregation result at the downlink from AS to RSU.
- The convolution neural network (CNN) with the MNIST dataset is uesed to evaluate the performance of the proposed scheme. The experimental results demonstrate that the proposed scheme shows a higher accuracy with the acceptable overhead for the FL participants.

The remainder of the paper is organized as follows. Section 2 presents related work and Sect. 3 presents background knowledge include the system model, cryptography primitives and federated learning. Section 4 introduces the mechanism details. Analysis including correctness, privacy and performance evaluation are presented in Sect. 5. Section 6 concludes the paper.

2 Related Work

In recent years, given the rising popularity of IoV [7–10], the data privacy of vehicles has increasingly become the focus of attention. Several studies have been proposed to solve related issues.

Traditional privacy protection schemes mainly combined cryptography to encrypt or hide the collected data and send it to the center. Hui Li et al. [11] designed an Architecture for identity and location privacy protection in

VANET base on k-anonymity and dynamic threshold encryption. Han et al. [12] designed a vehicle privacy-preserving algorithm based on a local differential privacy to minimize the possibility of exposing the regional semantic privacy of the k-location set. Ma et al. [13] performed homomorphic encryption on the sensitive part of the data and keep the ciphertext at the blockchain to preserve IoV data privacy. Zhang et al. [14] encrypted traffic flow data by BGN homomorphic encryption to protect the travel direction when arriving at T-junctions or crossroads.

Although these approaches do solve the issue of data privacy to a certain extent, still have two issues: (1) based on differential privacy, the noise will affect the availability of data. Based on homomorphic encryption, the computational overhead is not suitable in IoV. (2) a large amount of collected data needs to be sent, which will lead to the potential threat of leaking sensitive data and high network bandwidth usage.

FL, as a distributed artificial intelligence approach, allows participant trains local models on local privacy database and then the center aggregates the local model to construct a global model. Compared with traditional privacy protection schemes, FL could enhance communication efficiency and privacy preservation [15]. Lu et al. [16] designed a secure data sharing scheme based on asynchronous FL and blockchain, which could improve efficiency. A hierarchical FL algorithm with a multi-leader and multi-player game for knowledge sharing is proposed in [17]. Wu et al. [18] proposed a Traffic-Aware FL framework to enhance motion control of vehicles. Although the above FL-based schemes avoid directly uploading a large amount of privacy data, the uploaded gradient is not protected, and the privacy data is still likely to be exposed [6].

To prevent the leaking of privacy from the gradient, some studies are proposed. Phong et al. [19] bridged deep learning and homomorphic encryption to ensure that the server can not get user privacy and the accuracy is kept intact. Liu et al. [20] presented a privacy-enhanced FL (PEFL) framework by using homomorphic encryption, in process of FL, In the whole process, the gradient is only processed in the form of ciphertext. Although homomorphic encryption is useful for privacy-preserving, it is not suitable for IoV as the time cost. The scheme of adding mask to gradient was proposed in [21], two types of masks would be added in gradient, the participant only reply one mask recover request from the center, which Effectively protects privacy gradient. Further, a more efficient scheme [22] based on [21] is proposed, which only needs logarithmic overhead. To verify global aggregation result in FL, Fu et al. [23] designed a verifiable FL scheme by using Lagrange interpolation. Guo et al. [24] proposed a verifiable aggregation scheme for FL by using Linear homomorphic hash and Cryptographic commitment. To achieve the privacy-preserving, we propose a privacy-enhanced federated learning scheme based on gradient encryption by the mask.

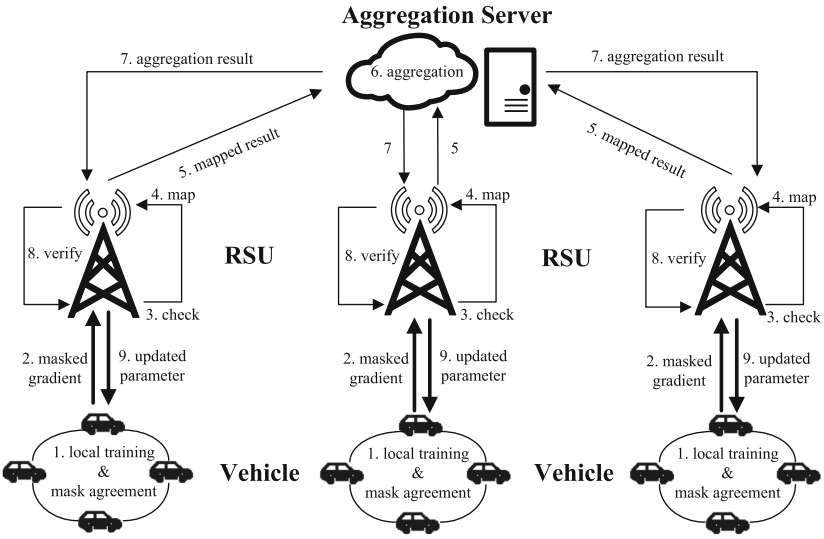

Fig. 1. Proposed Federated Learning Framework in IoV

3 Preliminaries

3.1 System Model

As shown in Fig. 1, the proposed scheme consists of some vehicles, some Roadside Units(RSUs) and one Aggregation Server(AS).

The vehicle executes federated learning to generate the local gradient on local privacy data set. During the whole process, the vehicle transmits the gradient instead of privacy data. RSU is a kind of wireless infrastructure, as relay node, RSU is responsible for organizing vehicles to execute the mask agreement and checking the correctness of received model parameters. AS is responsible for constructing a global model. In the proposed scheme, AS may return forged global model parameters to other participants.

3.2 Cryptography Block

- **Hard problem** Let $\mathbb{G}$ denotes a cyclic group, $g \in \mathbb{G}$ denotes a generator of group $\mathbb{G}$, and q is the prime order of group $\mathbb{G}$. Then the computational hard problems named Discrete Logarithm Problem (DLP), Decisional Diffie Hellman Problem (DDHP),Computational Diffie-Hellman Problem(CDHP) can be described as follows.
  (1) DLP: Given one tuple {P, Q}$(P, Q \in \mathbb{G})$, where $Q = P^x$, $x \in \mathbb{Z}_q^*$, the advantage for any probabilistic polynomial time (PPT) adversary to calculate x is negligible.
  (2) CDHP: Given one tuple $\{g, g^x, g^y \in \mathbb{G}\}$, where $x, y \in \mathbb{Z}_q^*$, the advantage for any PPT adversary to calculate $g^{xy} \in \mathbb{G}$ is negligible.

(3) DDHP: Given two tuples $\{g, g^x, g^y, g^{xy} \in \mathbb{G}\}$ and $\{g, g^x, g^y, g^z \in \mathbb{G}\}$, where $x, y, z \in \mathbb{Z}_q^*$, any PPT adversary is decisional hard to distinguish the two tuples.

- **Secret share** Shamir secret share is a threshold secret sharing scheme. The threshold secret sharing scheme first constructs a $t-1$ degree polynomial and takes secret k as a constant term of the polynomial

$$f_k(x) = k + \sum_{j=1}^{t-1} a_j x^j, a_j \in GF(q) \tag{1}$$

where q is a big prime number and $GF(*)$ is a finite field. Thus, according to formula 1, $f_k(0)$ is the secret k. Selecting n elements $\{x_i \in GF(q), 1 \leq i \leq n\}$ and feeding x_i into $f_k(x)$ to get $f_k(x_i)$, arbitrary t $\{(x_i, f(x_i))\}$ could recover the $t-1$ degree polynomial $f_k(x)$ as follows

$$f_k(x) = \sum_{i=1}^{t} f_k(x_i) \prod_{j=1, j\neq i}^{t} \frac{x - x_i}{x_i - x_j} \tag{2}$$

Therefore, secret k is secure if malicious participants can not obtain t or more sub-secrets $\{x_i, f_k(x_i)\}$.

3.3 Federated Learning

We leverage federated learning to protect the privacy data of vehicle. Assume there are N vehicles in proposed scheme, Vehicle $v_i (0 \leq i \leq N-1)$ participates in the FL and cooperatively trains a model $\mathcal{M}$ on private data set $D_i = \{(x_j, y_j), 0 \leq j \leq d_i - 1\}$, where d_i is the size of D_i A loss function quantifies the difference between estimated values and real values of samples in D_i, defined as follows:

$$E_i(\mathcal{M}) = \frac{1}{d_i} \sum_{j=0}^{d-1} L(\mathcal{M}, x_j, y_j) \tag{3}$$

where $L(\mathcal{M}, x_j, y_j)$ is the loss function on data sample (x_j, y_j), and y_j is the label of x_j. The global loss function $E(\mathcal{M})$ could be calculated.

$$E(\mathcal{M}) = \frac{1}{N} \sum_{i=0}^{N-1} E_i(\mathcal{M}) \tag{4}$$

In FL, each vehicle trains the model on the training set by a back propagation algorithm, and gets the private gradient $\omega_i = \frac{\partial E_i}{\partial \mathcal{M}}$. Vehicles and RSUs synchronously upload gradients to the AS to aggregate. Then AS returns the result to RSUs, vehicles download the result.

4 The Proposed Scheme

We assume that most vehicles are honest, but a small number of malicious gradients generated by malicious vehicles will still affect the aggregation results of FL. In the proposed scheme, RSUs could check the correctness of gradients uploaded by local region vehicles and eliminate malicious gradients. AS may forge global aggregation result, to verify the correctness of the global aggregation result, RSUs would perform the Lagrange interpolation function on the local region gradients. RSUs are regarded as honest participants, but during the whole process, all RSUs also can not learn vehicles' privacy data and gradients. The CA is used to generate public parameters for the registered vehicles and RSUs. The vehicle generates a private-public key pair which are used for negotiating masks respectively, the RSU generates a private-public key pair which is used for negotiating a common integer sequence as the Lagrange interpolation function points.

4.1 Initialization

In this phase, CA initializes all necessary system parameters and publishs the parameters to participants. RSU calculates common sequence as the necessity of verifying global parameter.

Algorithm 1: sequence generation algorithm

Result: Three integer sequences $SeqA$, $SeqB$, $SeqS$

Initialization: $R = RSU_0, RSU_1, \cdots, RSU_{K-1}$;

for $RSU_i = RSU_0 \rightarrow RSU_{K-1}$ **do**

  $r_i \leftarrow$ random select integer in Z_q ;

  broadcast $Z_i = g^{r_i}$ to R/RSU_i ;

  calculate $X_i = (\frac{Z_{i+1}}{Z_{i-1}})^{r_i}$;

  broadcast X_i to R/RSU_i ;

  calculate $ComKey = (Z_{i-1})^{nr_i} \cdot X_i^{n-1} \cdot X_{i+1}^{n-2} \cdots X_{i-2}$;

  broadcast $t_i = Enc(ComKey, 1)$ to R/RSU_i ;

  $T = \{t_0, t_1, \cdots, t_{K-1}\}/\{t_i\}$;

  for $t \in T$ **do**

    if *Dec(ComKey, t) != 1* **then**

      abort

    end

  end

  $SeqA, SeqB, SeqS = h1(ComKey||1), h1(ComKey||2), h2(ComKey||3)$;

  $h1(*)$ or $h2(*)$: map $*$ to a integer sequence

end

The CA generates a cyclic group $\mathbb{G}$ with prime order q, chooses randomly a group generator g as a public parameter which is used for calculating vehicles'

agreement key and RSUs' public key. The CA also generates $m+1$ positive integer $P = \{p, p_1, p_2, \cdots, p_m\}$, where $gcd(p_i, p_j) = 1, (i \neq j), 1 \leq i, j \leq m$ and p is large enough. Vehicles and RSUs receive public parameters $PP = \{\mathbb{G}, q, g, P\}$ to generate a private-public key, for example, vehicle v_i chooses $SK_i \in \mathbb{Z}_q^*$ and calculates $PK_{z,i} = g^{SK_i}$, RSU gets the private-public pair with the same operations with vehicle.

RSUs should calculate three common sequences $SeqA$, $SeqB$ and $SeqS$ as Algorithm 1 which are confidential to other entities to achieve the verifiability of the aggregated result from AS. Let $R = \{RSU_0, RSU_1, \cdots, RSU_{K-1}\}$,

Algorithm 2: FL algorithm

Output: global modelω
Initialization $groups = \{g_0, g_1, \cdots, g_{n_g-1}\}$ in each RSU ;
$R = \{RSU_0, RSU_1, \cdots, RSU_{K-1}\}$;
for *each local epoch* **do**
  for *group* $\in$ *groups in parallel* **do**
    for $v_i \in$ *group in parallel* **do**
      learning local model $\omega_i = \omega_i - \alpha \cdot \nabla E_i(\mathcal{M}, x, y)$;
      upload $\omega_{k,z,i}^{(M)} = \omega_i + m_{k,z,i}^{(R)} + m_i^{(V)}$ to RSU_k ;
    end
  end
  for *each* $RSU_k \in R$ **do**
    $\{\omega_k\} \leftarrow$ eliminate incorrect gradients ;
    upoad CRT(*) $\leftarrow$ Lagrange Interpolation on ω_k ;
  end
  distribute $\omega^{(G)} \leftarrow$ AS aggregates CRT(*) from RSUs ;
  $\omega^{(G')} \leftarrow$ RSU calculates real model parameter from $\omega^{(G)}$;
  return $\omega^{(G')}$ to vehicles ;
end

4.2 Gradient Encryption

- **Mask agreement.** In the proposed scheme, for any RSU_k, $0 \leq k \leq K-1$, we divide the vehicles in the RSU area into multiple groups, each group contains $h = \frac{N_k}{n_k}$ vehicles, where N_k is the number of vehicles in RSU_k and n_k is the number of the groups in RSU_k. In a group, every vehicle negotiates mask with the neighbor vehicles. To describe the mask agreement process, we denote the vehicles that have joined the mask agreement process as ordered sequence $\{v_0, v_1, \cdots, v_{h-1}\}$ in every group. In followed phases, we introduce the agreement process.

Vehicle $v_i (0 \leq i \leq h-1)$ in group $g_z (0 \leq z \leq n_k - 1)$ sends $PK_{z,i}$ to RSU_k when enters the communication range of RSU_k. RSU_k transmits all other public

keys to a vehicle v_i to execute the secret share in the group. Once a group is generated, then calculates the first mask code(FMC) $m_{k,z,i}^{(R)}$ between RSU_k and vehicle v_i as follows:

$$m_{k,z,i}^{(R)} = Hash(PK_{z,i}^{SK_k}) \tag{5}$$

where $Hash(*)$ is the function that map $*$ to a integer sequence with the same size as $|\omega|$.

ThenRSU_k
sends $\{PK_k||PK_{z,<i+1>}||c_{i,<i+1>}||PK_{z,<i-1>}||c_{i,<i-1>}||\{PK_u\}\}$ to $\{v_i, |i = 0, 1, \cdots, h-1\}$, where $< \cdot >= \cdot \pmod h$, $c_{i,<\cdot>} = -c_{<\cdot>,i}$, $\{PK_u\}$ is a set include other participants' information exclude $PK_{z,<i+1>}$ and $PK_{z,<i-1>}$.

After receiving message, vehicle v_i calculates FMC $m_{k,z,i}^{(R)} = Hash(PK_k^{SK_{z,i}})$, and calculates $k_1 = PK_{z,<i+1>}^{SK_{z,i}}$ and $k_2 = PK_{z,<i-1>}^{SK_{z,i}}$. To calculate second mask code(SMC), vehicle v_i executes $m_{i,<i+1>}^{(V)} = Hash(k_1)$ and $m_{i,<i-1>}^{(V)} = Hash(k_2)$, SMC could be calculated as follows.

$$m_i^{(V)} = c_{i,<i+1>} \cdot m_{i,<i+1>}^{(V)} + c_{i,<i-1>} \cdot m_{i,<i-1>}^{(V)} \tag{6}$$

Notice that both $m_i^{(V)}$ and $m_{k,z,i}^{(R)}$ would be added to the gradient to ensure the confidentiality of the gradient.

Vehicle v_i executes secret share to share k_1 and k_2 with the vehicle v_j, $j = 0, 1, \cdots, h-1$. In the proposed scheme, the Shamir algorithm is used for secret share. Vehicle v_i constructs two polynomials $f_{k_1}(x)$ and $f_{k_2}(x)$ as formula 1 and generates sub-secret $s_j = \{f_{k_1}(x_j)||f_{k_2}(x_j)||x_j\}$ for vehicle v_j. Then v_i encrypts sub-secrets

$$s_{i,j} = Enc(PK_j, s_j)$$

where $Enc(\cdot, *)$ is the encryption function such as RSA, $\cdot$ and $*$ are the public key and the plaintext. Vehicle v_i sends $s_{i,j}$ to RSU_k for forwarding to vehicle v_j.

- **Local training.** The local training is implemented with distributed gradient descent. In the training process, we iteratively improve the accuracy of model $\mathcal{M}$ by minimizing the global loss function 4.

For every vehicle in IoV, the goal of training the model is to find the gradient to update $\mathcal{M}$ for minimizing the value of the loss function 3.

Vehicle $v_i (0 \le i \le h-1)$ in $RSU_k (0 \le k \le K-1)$ calculates $\omega_i = \frac{\partial E_i}{\partial \mathcal{M}}$, where ω_i is a vector with dimension $d = |\omega_i|$. Then vehicle v_i masks ω_i as follows:

$$\omega_{k,z,i}^{(M)} = \omega_i + m_i^{(V)} + m_{k,z,i}^{(R)} \tag{7}$$

4.3 Region Verification

The difference between the malicious gradient and correct gradient is perceptible. Actually, the low similarity with the correct gradients means that the gradient is malicious with high probability. So the Pearson correlation coefficient between

gradients could be calculated to make a distinction between the correct gradients and the malicious gradients. Pearson correlation coefficient is defined as follows:

$$\rho_{XY} = \frac{Cov(X,Y)}{\sigma(X)\sigma(Y)} \tag{8}$$

where $Cov(X,Y)$ is covariance between random variables X and Y, $\sigma()$ is the standard deviation.

- **Verification** $RSU_k(0 \leq k \leq K-1)$ calculates the gradients sum of group $g_z(0 \leq z \leq n_k - 1)$ if all vehicles upload masked gradients. And we set

$$\omega_{k,z} = \sum_{i=0}^{h-1} \omega_{k,z,i}^{(M)} - m_{k,z,i}^{(R)} \tag{9}$$

If a vehicle (suppose v_i) leaves RSU_k's communication range, because of the lack of $\omega_{k,i}^{(M)}$, the sum of the gradients could not be recovered. To eliminate the mask, RSU_k sends a recover request to the rest vehicles in the group to get v_i's sub-secrets. The rest vehicle $v_j(0 \leq j \leq h-1, j \neq i)$ sends $dc_{i,j} = Dec(SK_j, s_{i,j})$ to RSU_k, where $Dec(\cdot,*)$ is decryption algorithm corresponding to encryption algorithm $Enc(\cdot,*)$. After receiving enough sub-secrets, RSU_k extracts $\{x_j, f_{k_1}(x_j), f_{k_2}(x_j)\}$ from $dc_{i,j}$ and executes formula 2 to get k_1 and k_2, further RSU_k calculates $m_{i,i+1}^{(V)} = Hash(k_1)$ and $m_{i,i-1}^{(V)} = Hash(k_2)$ to get the SMC of v_i as formula 6. Then RSU_k adds the recovered SMC to uploaded gradients to eliminate masks as follows.

$$\omega_{k,z,i} = \sum_{c=0,c\neq i}^{h-1} (\omega_{k,z,c}^{(M)} - m_{k,z,c}^{(R)}) + m_i^{(V)} \tag{10}$$

Note that, the value of $\omega_{k,z,i}$ is the gradients sum $\omega_{k,z}$ of group g_z, the RSU can not attain the specific gradients when recovering the SMC.

To calculate the Pearson correlation coefficient, the gradients coordinate-wise medians $(\bar{\omega_k})(0 \leq k \leq K-1)$ should be calculated as the benchmark.

$$\bar{\omega_k} = \frac{1}{n_k} \sum_{r=0}^{n_k-1} \omega_{k,r}$$

RSU_k randomly selects $X = \omega_{k,x}$ and $Y = \omega_{k,y}(0 \leq x, y \leq n_g - 1)$, and calculates ρ_{XY} according to $\bar{\omega_k}$ and formula 8. RSU_k discards $\omega_{k,x}$ if $\rho_{XY} \leq l$, where l is the limiting value of correlation coefficient, The number of rest gradients sum is $n_{k,re}$. Then RSU_k broadcasts $n_{k,re}$ to R. Before uploading gradients to AS, RSU_k will process the gradients sum with Lagrange interpolation.

- **Interpolation** RSU_k would convert the sum of gradients $\omega_k = \sum_{r=0}^{n_{k,re}-1} \omega_{k,r}$ from float number to finite field as follows:

$$\omega_k = \begin{cases} \lfloor \lambda \cdot \omega_k \rceil, & \lfloor \lambda \cdot \omega_k \rceil \geq 0 \\ p + \lfloor \lambda \cdot \omega_k \rceil, & \lfloor \lambda \cdot \omega_k \rceil < 0 \end{cases}$$

where $\lfloor * \rceil$ is the rounding method , λ is a large integer (i.e. 10^6) used to control accuracy. And in this phase, the data uploaded by RSU_k to AS is not gradients, but the Lagrange function results. First,RSU_k randomly selects $m-1$ arrays $U_k = \{\mathbf{u_{k,i}}, |i = 1, 2, \cdots, m-1\}$ that satisfies that the sum of U_k is equql to ω_k, note that, each element in U_k is an array with the dimension d. RSU_k has parameters $SeqS = [s_0, s_1, \cdots, s_{d-1}]$ and $SeqA = [a_0, a_1, \cdots, a_{m-1}]$, according to $(U_k, SeqA, SeqS)$, we have points $P_j = \{(a_i, u_{k,i,j}), |i \in \{0, 1, \cdots, m-2\}\} \cup \{(a_{m-1}, s_j)\}$, where $u_{k,i,j}$ is the j-th element in $u_{k,i}$. Therefore Lagrange interpolation function could be executed on P_j to generate $m-1$ degree polynomial $F_{k,j}(x)$ as formula 2. RSU_k sends $Pack_{k,j} = CRT[F_{k,j}(b_0), \cdots, F_{k,j}(b_{m-1})]$ on $SeqB = [b_0, b_1, \cdots, b_{m-1}]$, where CRT is the Chinese remainder theorem as [23].

4.4 Aggregation and Update

After receiving $\{(Pack_{k,0}, \cdots, Pack_{k,d-1}), |k = 0, 1, \cdots, K-1\}$, AS executes aggregation as follows:

$$\omega^{(G)} = (\sum_{k=0}^{K-1} Pack_{k,0}, \sum_{k=0}^{K-1} Pack_{k,1}, \cdots, \sum_{k=0}^{K-1} Pack_{k,d-1})$$

Because AS does not know the x-coordinate $SeqA$ and $SeqB$ corresponding to packaged function values $Pack_{k,*}$, AS couldn't forge an aggregation result that can be verified successfully by RSU. Then AS distributes $\omega^{(G)}$ to each RSU. After receiving $\omega^{(G)}$ from AS and $n_{re} = \sum_{i=0}^{K-1} n_{i,re}$ from other RSUs, to verify $\omega^{(G)}$, For any $j = 0, 1, \cdots, d-1$, RSU_k should unpack $\omega^{(G)}$ as follows:

$$F_j(b_i) \equiv \sum_{k=0}^{K-1} Pack_{k,j} \pmod{p_i}$$

As mentioned before, there are points $\{(b_0, F_j(b_0)), \cdots, (b_{m-1}, F_j(b_{m-1}))\}$, RSU_k applies the Lagrange interpolation to calculate corresponded function expression $F_j(x)$. Then RSU_k calculates $F_j(a_{m-1})$, the aggregation result is correct if $K \cdot s_j = F_j(a_{m-1})$ holds. Next, RSU_k calculates aggregation gradient as follows:

$$\omega^{(G')} = (C(\sum_{c=0}^{m-2} F_1(a_c)), C(\sum_{c=0}^{m-2} F_2(a_c)), \cdots, C(\sum_{c=0}^{m-2} F_d(a_c)))$$

where

$$C(x) = \begin{cases} x, & x \in [0, \dfrac{p-1}{2}) \\ x - p, & x \in [\dfrac{p-1}{2}, p) \end{cases} \tag{11}$$

Then RSU_k sends global model parameters $\omega^{(G')}$ and n_{re} to each vehicle in communication region. The vehicle updates the local model as follows:

$$\mathcal{M} = \mathcal{M} - \alpha \frac{\omega^{(G')}}{n_{re}} \tag{12}$$

where α is the learning rate. Then vehicles iteratively perform local training until global model $\mathcal{M}$ is available. The proposed FL algorithm is illustrated in Algorithm 2.

5 Analysis

5.1 Correctness and Privacy

The masks FMC and SMC are added into the gradient, after receiving the masked gradient, RSU performs formula 9 to get the sum of gradients. For any group r in RSU_k the correctness of the formula is as follows.

$$\begin{aligned}
\omega_{k,r} &= \sum_{i=0}^{h-1} \omega_{k,r,i}^{(M)} - m_{k,r,i}^{(R)} \\
&= \sum_{i=0}^{h-1} \omega_i + c_{i,i+1} \cdot m_{i,i+1}^{(V)} + c_{i,i-1} \cdot m_{i,i-1}^{(V)} \\
&= \sum_{i=0}^{h-1} \omega_i
\end{aligned}$$

If the vehicle $v_i (0 \leq i \leq h-1)$ leaves, RSU performs formula 10, and the correctness of the formula is as follows.

$$\begin{aligned}
\omega_{k,r} &= \sum_{j=0, j\neq i}^{h-1} (\omega_{k,r,j}^{(M)} - m_{k,r,j}^{(R)}) + m_i^{(V)} \\
&= \sum_{j=0}^{h-1} \omega_{k,r,j}^{(M)} - \sum_{j=0}^{h-1} m_{k,r,j}^{(R)} - \omega_{k,r,i}^{(M)} + m_{k,z,i}^{(R)} + m_i^{(V)} \\
&= \sum_{j=0, j\neq i}^{h-1} \omega_j
\end{aligned}$$

In the proposed scheme, to achieve the privacy-preserving of vehicle's data, we adopt FL to keep raw privacy data locally, Furthermore, the mask is used to protect gradients.

We combine h vehicles $\{v_0, v_1, \cdots, v_{h-1}\}$ to form a group. If the vehicle only masks uploaded gradients ω_i with SMC $m_i^{(V)}$, the vehicle's privacy gradient may leak. Because the vehicle may be curious about neighbor vehicle's privacy gradient, especially if the two side neighbors of a vehicle conspire, the SMC may

be eliminated. To address the problem of gradient leakage caused by neighbor vehicles collusion, We add FMC $m_{k,z,i}^{(R)}$ and SMC $m_i^{(V)}$ to the gradient. The FMC is only known between RSU and vehicle, so malicious neighbor vehicles can not get v_i's privacy gradient without knowing the FMC. In the meantime, we also avoid the RSU from knowing the specific gradient of the vehicle by adding the SMC, so the RSU only knows the sum of h vehicles' gradients. As mentioned above, the privacy gradient is only known by itself. And according to the computational or decisional hard problems mentioned before, it is difficult to calculate the FMC and SMC without knowing corresponding private key. hence, during the whole FL process, the privacy data and gradient of the vehicle will not be leaked.

5.2 Performance

In this section, we give the performance analysis of our proposed scheme. Our simulation experiment is conducted on Intel(R) Core(TM) i7-10875H,2.30GHz and 16 GB memory.

- **Performance of RSUs setup and agreement** RSUs should perform Lagrange interpolation with some points, RSUs' common sequence is regarded as the x-coordinate only known by itself. In the experiment, we use the JPBC library in Java JDK8 to execute the Algorithm 1.

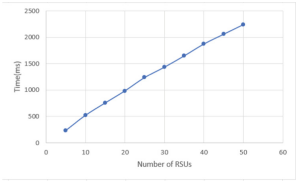

Fig. 2. The sequence generation overhead of RSU

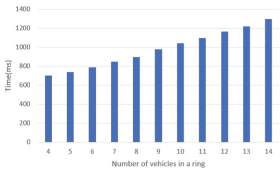

Fig. 3. The mask agreement overhead of vehicle

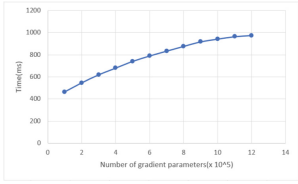

Fig. 4. the mask agreement overhead with different number of gradient parameter

We measure the computing overhead of our proposed algorithm under different number of RSUs. The computation overhead is shown in Fig. 2, with the number of RSU increasing, the overhead increases linearly. The frequency of RSU updating the common sequence can maintain at a low value, in the condition, the cost of calculating common sequences is acceptable. Next we measure the cost of agreement phase, we set $h = 4, 5, \cdots, 14$ and fix gradient length10^6. The computational overhead is shown in Fig. 3. With the number of gradients increasing, the computational overhead increases approximately linearly. A vehicle shares secrets K_1, K_2 with the vehicles within the group, the share operation cost is

low as the number of vehicles in a group is small. And the most time-consuming operation at this phase is to calculate the masks as the length of the gradient is generally the time of 10^5. Meanwhile, We set the number of CNN model parameters from 10^5 to $12 * 10^5$ and set $h = 9$. The computational overhead of different length of gradient as Fig. 4. The computational overhead increases approximately linearly with the increase of gradient length.

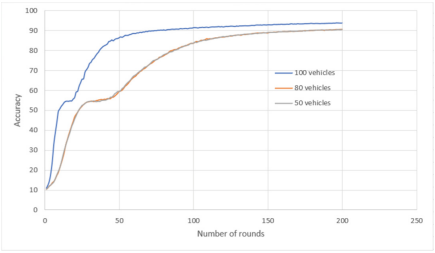

Fig. 5. The accuracy with various numbers of vehicles

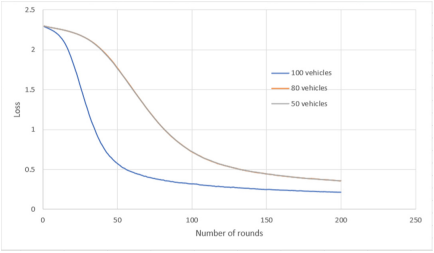

Fig. 6. The loss with various numbers of vehicels

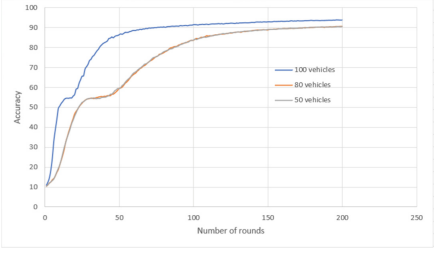

Fig. 7. The accuracy in various numbers of groups with low correlation coefficient

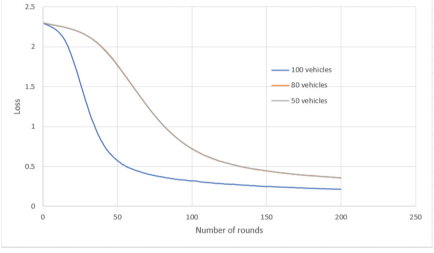

Fig. 8. The loss in various numbers of groups with low correlation coefficient

- **Performance of FL** We use the MNIST dataset to evaluate the proposed scheme. We divided the dataset into $50, 80, 100$ parts and assigned them to $50, 80, 100$ vehicles, and the number of vehicles in a group is $h = 5$, that is, $10, 16, 20$ groups would join in FL. Each vehicle executes the local train with a splited dataset. The Convolutional Neural Network (CNN) is used in the training process. The result of accuracy and loss are shown in Figs. 5 and 6. We set a various number of vehicles and 0 low correlation group, 100 vehicles that joined FL could provide the highest accuracy and the lowest loss, 50 and 80 vehicles could achieve almost the same accuracy and loss. The proposed scheme could achieve a satisfactory result with no malicious vehicles.

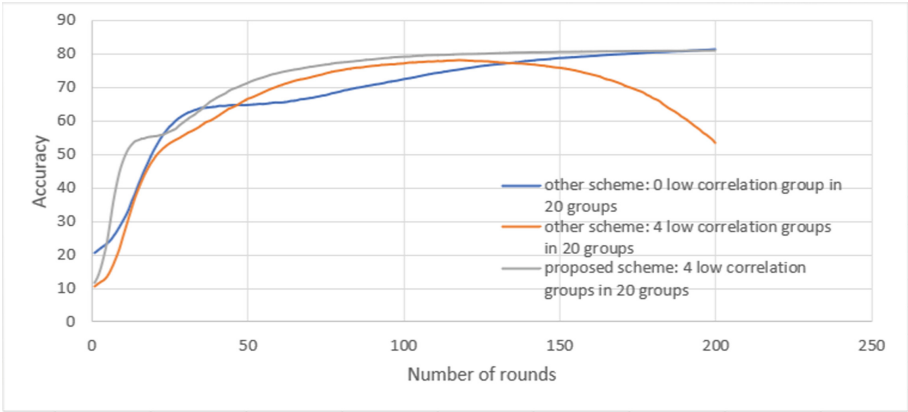

Fig. 9. The performance with malicious gradient between scheme [23] and the proposed scheme

To measure the result with the malicious vehicles joined, we execute FL with 100 data providers(20 groups) with different proportions of low correlation groups. The result of accuracy and loss are shown in Figs. 7 and 8. The accuracy has a reduction and the loss is not as good as Fig. 6. And as shonw in Fig. 9, compared with [23], we have better performance in the presence of malicious vehicles.

6 Conclusion

In this paper, we propose the privacy-preserving FL scheme. The proposed scheme addresses the vehicle data privacy and gradient privacy, and the malicious participants' gradient could be removed at RSU by calculating the Pearson correlation coefficient. Meanwhile, we use Lagrange interpolation to verify the correctness of the returned result from AS. To reduce the overhead, We form a small number of vehicles as a group to negotiate the mask. Numerical results confirm the effectiveness of our proposed scheme in terms of accuracy. In future work, we plan to reduce data waste in the process of eliminating malicious gradients.

References

1. Konecny, J., McMahan, H.B., Yu, F.X., Richtarik, P., Suresh, A.T., Bacon, D.: Federated learning: Strategies for improving communication efficiency. In: NIPS Workshop on Private Multi-Party Machine Learning (2016)
2. Reisizadeh, A., Mokhtari, A., Hassani, H., Jadbabaie, A., Pedarsani, R.: Fedpaq: A communication-efficient federated learning method with periodic averaging and quantization. In: International Conference on Artificial Intelligence and Statistics, pp. 2021–2031 (2020)

3. Liang, F., Yang, Q., Liu, R., Wang, J., Sato, K., Guo, J.: Semi-synchronous federated learning protocol with dynamic aggregation in internet of vehicles. IEEE Trans. Veh. Technol. **71**(5), 4677–4691 (2022)
4. Kong, X., Wang, K., Hou, M., Hao, X., Shen, G., Chen, X., Xia, F.: A federated learning-based license plate recognition scheme for 5g-enabled internet of vehicles. IEEE Trans. Industr. Inf. **17**(12), 8523–8530 (2021)
5. Ayaz, F., Sheng, Z., Tian, D., Guan, Y.L.: A blockchain based federated learning for message dissemination in vehicular networks. IEEE Trans. Veh. Technol. **71**(2), 1927–1940 (2022)
6. 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, pp. 603–618 (2017)
7. Xiong, L., Xiong, N., Wang, C., Yu, X., Shuai, M.: An efficient lightweight authentication scheme with daptive resilience of asynchronization attacks for wireless sensor networks. IEEE Trans. Syst. Man Cybern. Syst. **51**(9), 5626–5638 (2021)
8. Xiong, L., Li, G., He, M., Liu, Z., Peng, T.: An efficient privacy-aware authentication scheme with hierarchical access control for mobile cloud computing services. IEEE Trans. Cloud Comput. (2020). https://doi.org/10.1109/TCC.2020.3029878
9. Gong, C., Xiong, L., He, X., Niu, X.: Blockchain-based conditional privacy-preserving authentication scheme for vehicular ad hoc networks. J. Ambient. Intell. Humaniz. Comput. **2022**, 1–14 (2022)
10. Shuai, M., Xiong, L., Wang, C., Yu, N.: A secure authentication scheme with forward secrecy for industrial internet of things using Rabin cryptosystem. Comput. Commun. **16**(1), 215–227 (2020)
11. Li, H., Pei, L., Liao, D., Sun, G., Xu, D.: Blockchain meets vanet: An architecture for identity and location privacy protection in vanet. Peer-to-Peer Netw. Appl. **12**(5), 1178–1193 (2019)
12. Han, W., Cheng, M., Lei, M., Xu, H., Qian, L.: Privacy protection algorithm for the internet of vehicles based on local differential privacy and game model. Comput. Mater. Continua **64**(2), 1025–1038 (2020)
13. Ma, Z., Wang, L., Zhao, W.: Blockchain-driven trusted data sharing with privacy protection in iot sensor network. IEEE Sens. J. **21**(22), 25472–25479 (2021)
14. Zhang, C., Zhu, L., Ni, J., Huang, C., Shen, X.: Verifiable and privacy-preserving traffic flow statistics for advanced traffic management systems. IEEE Trans. Veh. Technol. **69**(9), 10336–10347 (2020)
15. Tan, K., Bremner, D., Kernec, J.L., Imran, M.: Federated machine learning in vehicular networks: A summary of recent applications. In: 2020 International Conference on UK-China Emerging Technologies (UCET), pp. 1–4 (2020)
16. Lu, Y., Huang, X., Zhang, K., Maharjan, S., Zhang, Y.: Blockchain empowered asynchronous federated learning for secure data sharing in internet of vehicles. IEEE Trans. Veh. Technol. **69**(4), 4298–4311 (2020)
17. Chai, H., Leng, S., Chen, Y., Zhang, K.: A hierarchical blockchain-enabled federated learning algorithm for knowledge sharing in internet of vehicles. IEEE Trans. Intell. Transp. Syst. **22**(7), 3975–3986 (2021)
18. Wu, T., Jiang, M., Han, Y., Yuan, Z., Li, X., Zhang, L.: A traffic-aware federated imitation learning framework for motion control at unsignalized intersections with internet of vehicles. Electronics **10**(24), 3050 (2021)
19. Phong, L.T., Aono, Y., Hayashi, T., Wang, L., Moriai, S.: Privacy-preserving deep learning via additively homomorphic encryption. IEEE Trans. Inf. Forensics Secur. **13**(5), 1333–1345 (2018)

20. Liu, X., Li, H., Xu, G., Chen, Z., Huang, X., Lu, R.: Privacy-enhanced federated learning against poisoning adversaries. IEEE Trans. Inf. Forensics Secur. **16**, 4574–4588 (2021)
21. Bonawitz, K., et al.: Practical secure aggregation for privacy-preserving machine learning. In: Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, pp. 1175–1191 (2017)
22. Bell, J.H., Bonawitz, K.A., Gascon, A., Lepoint, T., Raykova, M.: Secure Single-Server Aggregation with (Poly)Logarithmic Overhead, pp. 1253–1269. Association for Computing Machinery (2020)
23. Fu, A., Zhang, X., Xiong, N., Gao, Y., Wang, H., Zhang, J.: Vfl: A verifiable federated learning with privacy-preserving for big data in industrial iot. IEEE Trans. Industr. Inf. **18**(5), 3316–3326 (2022)
24. Guo, X., Liu, Z., Li, J., Gao, J., Hou, B., Dong, C., Baker, T.: Verifl: Communication-efficient and fast verifiable aggregation for federated learning. IEEE Trans. Inf. Forensics Secur. **16**, 1736–1751 (2021)