研究与开发

基于模型分割的联邦学习数据隐私保护方法

陈卡

（驻马店职业技术学院，河南 驻马店 463000）

摘　要：在不共享原始数据的前提下，分割学习（split learning，SL）允许客户端同服务端协作训练深度学习模型，进而保护数据隐私。然而，SL仍存在数据隐私泄露问题。为此，提出基于二值分割学习的数据隐私保护（binarized split learning-based data privacy protection，BLDP）算法。将客户端所训练的本地模型进行二值化，降低由分割层输出值引起的数据泄露损失。同时，BLDP算法采用泄露约束训练机制，进一步减少数据泄露损失。该机制以本地数据泄露损失和模型精度损失为总体损失值进行模型训练，进而在维护模型精度的同时，保护数据隐私。以4个常用的基准数据集进行训练，分析BLDP算法的分类准确率以及减少数据隐私泄露损失方面的性能。分析结果表明，所提BLDP算法能在分类准确率和数据隐私泄露损失间达成平衡。

关键词：联邦学习；隐私保护；分割学习；二值化；泄露损失

中图分类号：TN18

文献标志码：A

doi: 10.11959/j.issn.1000−0801.2024206

Model split-based data privacy protection method for federated learning

CHEN Ka

Zhumadian Vocational and Technical College, Zhumadian 463000, China

Abstract: Split learning (SL) enables data privacy preservation by allowing clients to collaboratively train a deep learning model with the server without sharing raw data. However, the SL still has limitations such as potential data privacy leakage. Therefore, binarized split learning-based data privacy protection (BLDP) algorithm was proposed. In BLDP, the local layers of client were binarized to reduce privacy leakage from SL smashed data. In addition, the leakage-restriction training strategy was proposed to further reduce data leaks. The strategy combines leak loss of local private data and model accuracy loss that enhances privacy while maintaining model accuracy. To evaluate the proposed BLDP algorithm, experiments were conducted on four commonly benchmarked datasets and the leakage loss and model accuracy were analyzed. The results show that the proposed BLDP algorithm can achieve a balance between classification accuracy and data privacy loss.

Key words: federated learning, privacy protection, split learning, binarization, leakage loss

收稿日期：2024−03−02；**修回日期：**2024−09−05

基金项目：2024年河南省科技发展计划项目（No.242400410491）；河南省科技攻关计划项目（No.212102210516 ）；2021年度河南省高等教育教学改革研究与实践立项项目 (No.2021SJGLX865)

Foundation Items: Science and Technology Development Plan of Henan Province in 2024 (No.242400410491)，Science and Technology Research Plan of Henan Province (No.212102210516)，Henan Province Higher Education Teaching Reform Research and Practice Project in 2021 (No.2021SJGLX865)

0　引言

深度学习（deep learning，DL）已在计算机视觉、自然语言处理、入侵检测等多个领域得到广泛应用[1]。传统DL常以集中方式训练数据，在这种训练模式下，客户端需将数据上传至服务端，由服务端集中学习数据。这种模式易泄露客户端数据，而相比于集中学习，分布式学习更有利于保护数据安全[2]。

联邦学习（federated learning，FL）属于典型的分布式学习方法[3]。FL允许客户端处理本地数据，无须将数据上传至服务端。在FL中，客户端用本地数据训练局部模型，再将局部模型参数传递至服务端，最终由服务端训练全局模型。在这种学习过程中，客户端的数据一直存留在客户端，并没有直接暴露于服务端，增加了数据的安全性。

然而，客户端通常为微型设备，其计算资源、存储容量有限，若完全由客户端训练本地数据，训练模型所消耗的时间过长、效率较低。此外，尽管客户端只将局部模型参数如梯度上传至服务端，但通过分析梯度参数能重构本地数据，仍然存在数据泄露的风险。

分割学习（split learning，SL）属于FL的特例[4-5]。SL中的训练任务由客户端和服务端共同完成。即将整个训练模型的任务分割成两个部分，一部分由客户端完成，另一部分由服务端完成。相比于FL，一方面，SL能缓解客户端资源不足的问题，只需训练一部分任务；另一方面，SL也无须将数据直接上传至服务端，只需将训练的模型参数上传至服务端，这降低了客户端数据泄露的风险，具有一定的隐私保护能力。考虑客户端的资源有限，通常客户端只承担训练模型的少部分任务。相比之下，服务端资源丰富，可承担大部分的训练模型任务。

文献[6]利用SL对心电图数据进行保护隐私的训练，并进行了性能分析。分析表明，若不采用任何隐私保护策略，只在一维卷积神经网络（convolutional neural network，CNN）上进行SL仍存在数据泄露问题。文献[7]将深度神经网络分割成两个部分，并将dropping-activation输出方法应用于本地层的激活，使它们具有不可逆性，进而保护数据的隐私。文献[8]提出了基于差分隐私的抗投毒攻击FL方案。该方案根据模型间余弦相似度赋予各客户端相应聚合权重，使用同态加密技术进行本地模型聚合，进而保护本地数据的安全性。此外，文献[9]也提出基于本地和中心差分隐私的混合加噪算法。该算法依据用户的隐私需求，为用户提供本地或混合差分隐私保护策略。文献[10]为了提高FL的安全性，提出基于双重知识蒸馏算法，并利用自适应的差分隐私策略实现客户端级的强隐私保护。

文献[11]也对模型进行分割，并采用stepwise-activation函数方法，使激活函数的输出具有不可逆性。性能分析表明，该方法的性能依赖于stepwise参数，调整stepwise参数，能在模型精度与数据隐私保护间达到平衡。文献[12-13]的研究表明，降低中间表示与原始数据间的距离相关性，能防御基于重构攻击的隐私泄露。

为此，针对SL中隐私保护问题，通过部署新颖的SL模型降低数据隐私泄露的风险，采用二值神经网络（binarized neural network，BNN），提出基于二值分割学习的数据隐私保护（binarized split learning-based data privacy protection，BLDP）算法。本文主要工作可归纳如下：采用分割学习策略训练模型，使服务端不能直接获取客户端的原始数据；采用二值化策略，将模型参数进行二值化；在训练过程中加入数据泄露损失，保护数据隐私，并实现模型训练精度与数据泄露损失的平衡；利用典型的数据集分析BLDP算法的性能。

1　SL概述

SL是将DL模型分裂成多个部分，每个部分

在不同的实体上训练[14]。每个部分中最后一层的输出被传递至下一个部分，将每个部分中最后一层称为分裂层。因此，除了分裂层的激活函数，原始数据并没有在其他实体间分享，可保证原始数据的安全性。

基于一维CNN结构的SL模型如图1所示，将DL模型分割成两个部分（客户端和服务端）。它们通过传递激活和梯度函数，协作地训练模型。

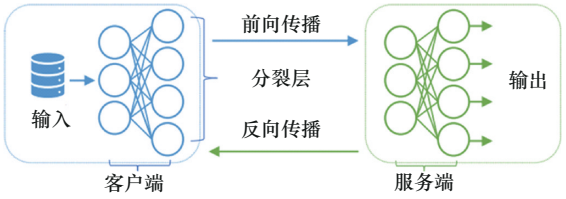


图1　基于一维CNN结构的SL模型

在前向传播中，客户端利用其本地数据训练局部模型，然后将分割层的输出传递至服务端，由服务端继续训练。在后向传播中，服务端计算梯度，再将分割层的梯度传至客户端，由客户端继续完成训练。重复上述过程，直至模型训练完成。

1.1　客户端

假定DL模型具有L层，不包含输入层。将该模型分割成两个部分，前l_{f}层由客户端训练，后l_{a}层由服务器端训练，且$l_{\mathrm{a}}=L-l_{\mathrm{f}}$。令$w^{(i)}$表示第$i$层的权重；$f^{(i)}$表示在第$i$层上的前向传播；$z^{(i)}$表示在第$i$层的激活值；$a^{(i)}$表示第$i$层经激励后的输出，可表示为$a^{(i)}=g^{(i)}z^{(i)}$，其中$g^{(i)}$表示第$i$层的激励函数。算法1给出客户端的执行过程。其中$D$表示数据集。

算法1：客户端处理流程

初始化操作

For each batch (x,y) generated from D do

前向传播：

$a^{(0)}\leftarrow x$；#将数据作为第一个层输入

for $i\leftarrow 1$ to l_{a} do

$$z^{(i)}\leftarrow f^{(i)}\left(a^{(i-1)}\right)$$

$$a^{(i)}\leftarrow g^{(i)}\left(z^{(i)}\right)$$

end for

传输$\left(a^{(l)},y\right)$

反向传播：

从服务端接收$\partial E/\partial a^{(l_{\mathrm{f}})}$

for $i\leftarrow l_{\mathrm{f}}$ to 1 do

$$\frac{\partial E}{\partial z^{(i)}}\leftarrow\frac{\partial E}{\partial a^{(i)}}\times g^{(i)\prime}\left(z^{(i)}\right)$$

$$\frac{\partial E}{\partial w^{(i)}}\leftarrow\frac{\partial E}{\partial z^{(i)}}\times\frac{\partial z^{(i)}}{\partial w^{(i)}}$$

if $i\neq 1$ do

$$\frac{\partial E}{\partial a^{(i-1)}}\leftarrow\frac{\partial E}{\partial z^{(i)}}\times\frac{\partial z^{(i)}}{\partial a^{(i-1)}}$$

end if

用$\partial E/\left(\partial w^{(i)}\right)$更新$w^{(i)}$

end for

end for

在后向传播时，客户端先从服务器端接收损失值，并计算关于输出值的偏导数：$\partial E/\left(\partial a^{(l_{\mathrm{f}})}\right)$，然后依算法1所示，逐步从第$l_{\mathrm{f}}$层迭代至第1层，其中$E$表示损失值，$a^{(l_{\mathrm{f}})}$表示第$l_{\mathrm{f}}$层向$(l_{\mathrm{f}}-1)$层的反向输出值。

1.2　服务端

在服务端先进入前向传输阶段。服务端接收$\left(a^{(l)},y\right)$后，就进行从l_{a}层至L层的迭代，如算法2所示。迭代结束后，利用损失函数计算损失值E。然后，再进入反向传播。先求梯度$\partial E/\partial a^{(l)}$，再进入从$L$层至$l_{\mathrm{a}}$层的迭代。

算法2：服务端处理流程

初始化操作

for $i\leftarrow 1$ to N do #N表示迭代的次数

前向传播：

从客户端接收$\left(a^{(\ell)},y\right)$

for $i\leftarrow l_{\mathrm{a}}$ to L do

$z^{(i)}\leftarrow f^{(i)}\left(a^{(i-1)}\right)$

$a^{(i)}\leftarrow g^{(i)}\left(z^{(i)}\right)$

end for

计算损失值E

反向传播：

计算$\partial E/\partial a^{(L)}$

for $i\leftarrow L$ to l_a do

$$\frac{\partial E}{\partial z^{(i)}}\leftarrow\frac{\partial E}{\partial a^{(i)}}\times g^{(i)\prime}\left(z^{(i)}\right)$$

$$\frac{\partial E}{\partial w^{(i)}}\leftarrow\frac{\partial E}{\partial z^{(i)}}\times\frac{\partial z^{(i)}}{\partial w^{(i)}}$$

$$\frac{\partial E}{\partial a^{(i-1)}}\leftarrow\frac{\partial E}{\partial z^{(i)}}\times\frac{\partial z^{(i)}}{\partial a^{(i-1)}}$$

利用$\frac{\partial E}{\partial w^{(i)}}$更新$w^{(i)}$

end for

end for

1.3　客户端与服务端间交互过程

客户端与服务端间的交互过程如图2所示。图2中的（1）~（9）表示执行顺序。首先，它们建立连接，并进行相应的同步操作。客户端先进入前向传播阶段，将其最后一层的输出传输至服务端，服务端再继续依层次迭代计算，最终计算损失值的梯度，并传至客户端，客户端再进入反向传播。

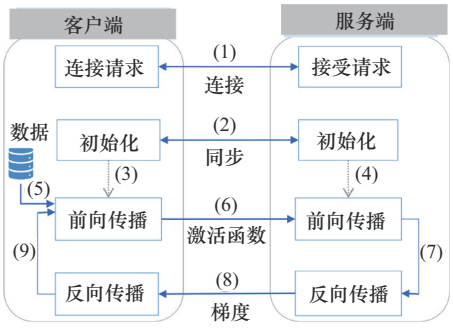


图2　客户端与服务端间的交互过程

2　BLDP算法

为了保护客户端的数据隐私，对数据进行二值化处理，即采用二值分割学习（binarize SL，BSL）。

2.1　二值分割学习

采用Sign函数对权重和输出值进行二元化。Sign函数的定义如下。

$$x_{\mathrm{b}}=\mathrm{Sign}(x)=\begin{cases}1, & x\geqslant 0\\ -1, & \text{其他}\end{cases}\tag{1}$$

其中，对于任意真实值x，利用$\mathrm{Sign}(x)$将x转变成二元值x_{b}。

考虑Sign函数在多数值上的梯度为零，采用直通估计器（straight-through estimator，STE），同时采用批归一化（batch normalization，BN）和权重剪裁（weights clipping，CP）技术提升训练速度，并降低权重尺度的影响。

2.2　客户端权重和输出值的二值化

客户端先利用式（1）对输入数据x进行二值化，构成第一层输入值$a_{\mathrm{b}}^{(0)}$，下标“b”表示变量为二值变量。然后，再从第1层至l_{f}层逐层产生输出。如算法3所示。对$w^{(i)}$进行二值化：$w_{\mathrm{b}}^{(i)}\leftarrow\mathrm{Sign}\left(w^{(i)}\right)$；并对$z^{(i)}$进行批归一化BN，进而输出$a^{(i)}$：$a^{(i)}\leftarrow\mathrm{BN}\left(z^{(i)}\right)$；再对输出值$a^{(i)}$进行二值化：$a_{\mathrm{b}}^{(i)}\leftarrow\mathrm{Sign}\left(a^{(i)}\right)$；最后，客户端将$\left(a_{\mathrm{b}}^{(l_{\mathrm{f}})},y\right)$传输至服务端。

利用文献[12]的泄露熵评估本地隐私数据的泄露程度。在训练过程中的总体损失由分类准确损失（准确率熵）和数据泄露损失（泄露熵）两部分构成：

$$\begin{aligned}L_{\mathrm{total}}&=\lambda E_1+(1-\lambda)E_2\\&=\lambda\underset{W_1,W_2}{L_1}\left(a_{\mathrm{b}}^{(L)},y\right)+(1-\lambda)\underset{W_1}{L_2}\left(a_{\mathrm{b}}^{(l_f)},x\right)\end{aligned}\tag{2}$$

其中，λ为权重因子，E_1和E_2分别信息熵和泄露

熵。通过调整权重因子λ，使信息熵与泄露熵间达到平衡；x为模型的输入数据（原数据）；y为原数据所对应的标签；$a^{(L)}$表示经L层模型训练后的输出值，即估计的标签值；$a^{(l_f)}$表示分割层的输出，即客户端向服务端的输出；W_1和W_2分别表示客户端所训练模型的参数和服务端所训练模型的参数。令W表示整个训练模型参数，即$W=\left(w^{(1)},w^{(2)},\cdots,w^{(L)}\right)$，则$W_1=\left(w^{(l_f+1)},w^{(l_f+2)},\cdots,w^{(L)}\right)$。

据式（2）可知，$L_1\left(a_b^{(L)},y\right)$表示估计的标签值与真实标签值的差距，反映了准确率熵，它由W_1和W_2共同决定；而$L_2\left(a_b^{(l_f)},x\right)$反映了泄露熵，表示$a^{(l_f)}$与$x$间的相关性，即通过$a^{(l_f)}$重构原始数据$x$的可能性。

在前向传播中，在客户端将$\left(a_b^{(l_f)},y\right)$传输至服务端后，客户端计算泄露熵$E_2\leftarrow L_2\left(a_b^{(l_f)},x\right)$，如算法3的步骤（4）所示。

算法3：基于BSL的客户端处理流程

初始化操作

For each batch (x,y) generated from D do

前向传播：

（1） $a_b^{(0)}\leftarrow x$

（2） for $i\leftarrow 1$ to l_f do

$w_b^{(i)}\leftarrow \mathrm{Sign}\left(w^{(i)}\right)$

$z^{(i)}\leftarrow f^{(i)}\left(a^{(i-1)}\right)$

$a^{(i)}\leftarrow \mathrm{BN}\left(z^{(i)}\right)$

$a_b^{(i)}\leftarrow \mathrm{Sign}\left(a^{(i)}\right)$

end for

（3） 传输$\left(a_b^{(l_f)},y\right)$

（4） 计算信息泄露熵$E_2\leftarrow L_2\left(a_b^{(l_f)},x\right)$

反向传播：

（5） 从服务端接收$\partial E_1/\partial a_b^{(l_f)}$

（6） $\dfrac{\partial E}{\partial a_b^{(l_f)}}\leftarrow\lambda\dfrac{\partial E_1}{\partial a_b^{(l_f)}}+(1-\lambda)\dfrac{\partial E_2}{\partial a_b^{(l_f)}}$

（7） for $i\leftarrow l_f$ to 1 do

$$\frac{\partial E}{\partial a^{(i)}}\leftarrow\frac{\partial E}{\partial a_b^{(i)}}\times\frac{\partial Htanh}{\partial a^{(i)}}$$

$$\frac{\partial E}{\partial z^{(i)}}\leftarrow\frac{\partial E}{\partial a^{(i)}}\times\frac{\partial a^{(i)}}{\partial z_b^{(i)}}$$

$$\frac{\partial E}{\partial w_b^{(i)}}\leftarrow\frac{\partial E}{\partial z^{(i)}}\times\frac{\partial z^{(i)}}{\partial w_b^{(i)}}$$

if $i\neq 1$ do

$$\frac{\partial E}{\partial a_b^{(i-1)}}\leftarrow\frac{\partial E}{\partial z^{(i)}}\times\frac{\partial z^{(i)}}{\partial a_b^{(i-1)}}$$

end if

（8） 用$\partial E/\left(\partial w_b^{(i)}\right)$更新$w^{(i)}$

（9） $w_b^{(i)}\leftarrow \mathrm{Clip}\left(w^{(i)},-1,+1\right)$

end for

end for

在反向传播阶段，客户端先从服务端接收梯度值$\partial E_1/\left(\partial a_b^{(l_f)}\right)$，其中$E_1=L_1\left(a_b^{(L)},y\right)$表示估计的标签值与真实标签值的差距，反映了准确率熵，并如算法3的步骤（6）所示，进行如下计算：

$$\frac{\partial E}{\partial a_b^{(l_f)}}\leftarrow\lambda\frac{\partial E_1}{\partial a_b^{(l_f)}}+(1-\lambda)\frac{\partial E_2}{\partial a_b^{(l_f)}}\qquad(3)$$

其中，λ为比例参数。通过调整λ使算法在分类准确率与数据泄露损失间达到平衡。

再计算从第l_f层至1层的梯度，并更新权重数。由于离散值的梯度为零，采用STE算法更新梯度。

STE算法的核心思想：最初参数值为浮点的连续值，在前向传播时，将这些连续参数二值化，再进行计算，得到网络的输出。在反向传播时，直接对原来浮点的连续参数进行更新，而不是对二值化的参数进行更新。

因此，依据STE算法的思想，由$\partial E/\left(\partial a_b^{(i)}\right)$

得出$\partial E/\left(\partial a^{(i)}\right)$：

$$\frac{\partial E}{\partial a^{(i)}} \leftarrow \frac{\partial E}{\partial a_{\mathrm{b}}^{(i)}} \times \frac{\partial \mathrm{Htanh}}{\partial a^{(i)}} \tag{4}$$

其中，$\mathrm{Htanh}\left(a^{(i)}\right)=\max\left(-1,\min\left(1,a^{(i)}\right)\right)$。

随后，对$z^{(i)}$进行反向的BN操作，并结合$\partial E/\left(\partial a^{(i)}\right)$的值，计算$\partial E/\left(\partial z^{(i)}\right)$，再利用$\partial E/\left(\partial z^{(i)}\right)$计算$\partial E/\left(\partial w_{\mathrm{b}}^{(i)}\right)$：

$$\frac{\partial E}{\partial w_{\mathrm{b}}^{(i)}} \leftarrow \frac{\partial E}{\partial z^{(i)}} \times \frac{\partial z^{(i)}}{\partial w_{\mathrm{b}}^{(i)}} \tag{5}$$

最后，利用Clip算法将$w^{(i)}$映射至$\{-1,+1\}$，形成二值化的$w_{\mathrm{b}}^{(i)}$，如算法3所示，其中$\mathrm{Clip}\left(w^{(i)},-1,+1\right)$表示剪裁算法。

此外，服务端仍按算法2进行处理。与算法2不同的是，服务端将所接收值进行二值化。

3　实验与分析

3.1　实验环境

对BLDP算法在分类二维图像的性能进行分析。在评估隐私保护性能方面，常利用原始数据与分割层输出数据（smashed data）间的相关性或相似指标评估数据泄露的程度，如均方误差、KL散度（Kullback-Leibler divergence，KL-D）[15]、结构相似度指标（structural similarity index measure，SSIM）、峰值信噪比（peak signal-to-noise ratio，PNSR）[16]。本文采用KL-D和SSIM指标，其中KL-D用于测量两个概率分布间的差异；SSIM用于测量两幅图像的人类感知相似性，进而分析是否能通过分割层输出数据重构原始图像。KL-D越小或者SSIM越大，则隐私泄露越严重，反之则相反。

基于MNIST[17]、Fashion-MNIST（F-MNIST）[18]、SVHN[19]和CIFAR-10[20]这4个数据集分析BLDP算法的性能。CIFAR-10采用基于视觉几何组（visual geometry group，VGG）的卷积神经网络，其他的数据集采用基于LeNet-5的卷积神经网络。选择以下两个基准算法：P-CNN，表示未进行二值化的分割模型，客户端采用纯CNN模型训练数据；B-CNN，表示除输入层和输出层外，剩余层的数据全进行二值化。值得注意的是，BLDP算法只对部分参数进行二值化。

3.2　数据分析

3.2.1　分类准确率

首先，分析BLDP算法、P-CNN算法和B-SL算法在各数据集上的分类准确率，如图3所示，B-CNN算法的准确率最低，其原因在于B-CNN算法对所有数据都进行二值化，丢失了大量数据。例如，在MNIST集上，B-CNN算法的准确率约96.7%，而BLDP算法和P-CNN算法的准确率分别为99.1%和99.3%。

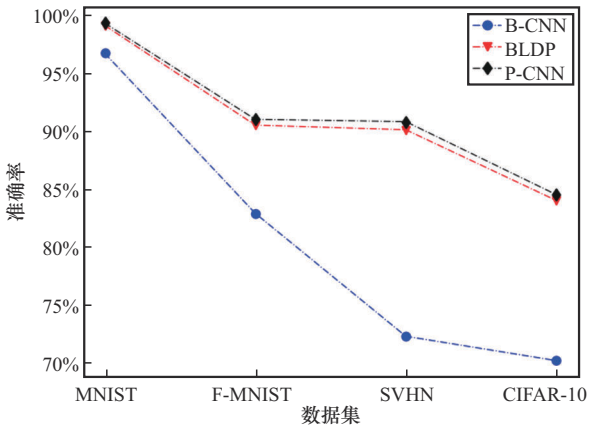


图3　分类准确率

相比于B-CNN算法，BLDP算法能获取较好的精度，但与P-CNN算法的准确率相比仍存在约1%的差距。这说明：对权重参数以及分割层输出的梯度值进行二值化时会损失部分数据，降低了分类准确率，但降低的幅度有限。

3.2.2　保护数据隐私的性能

接下来，分析BLDP算法和P-CNN算法在保护数据隐私方面的性能。通过测量原始数据的泄露值评估算法保护数据隐私的能力。为了隔离二次抽样的影响，利用分割层的输出值评估数据泄露的性能。

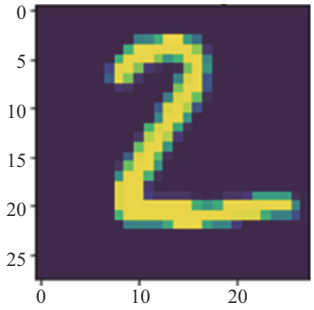
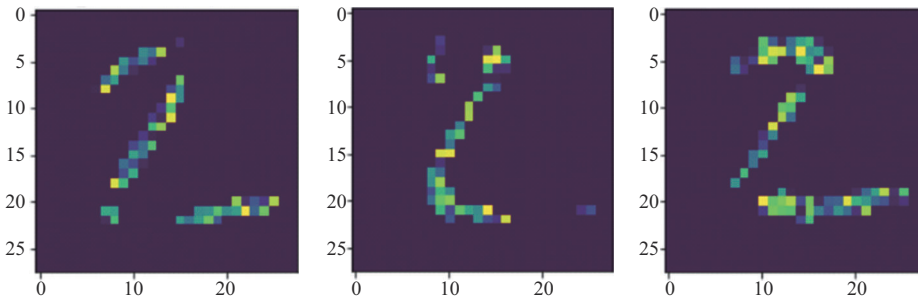


图4　原始图像

首先，从输出值重构二维图像（以下简称重构图像），并将此重构图像与原始图像进行比较。图5~图7显示P-CNN算法、B-CNN算法和BLDP算法所重构的图像情况（MNIST数据集），其中图4为原始图像；图5为P-CNN算法所重构的图像；图6为B-CNN算法所重构的图像；图7为BLDP算法所重构的图像。

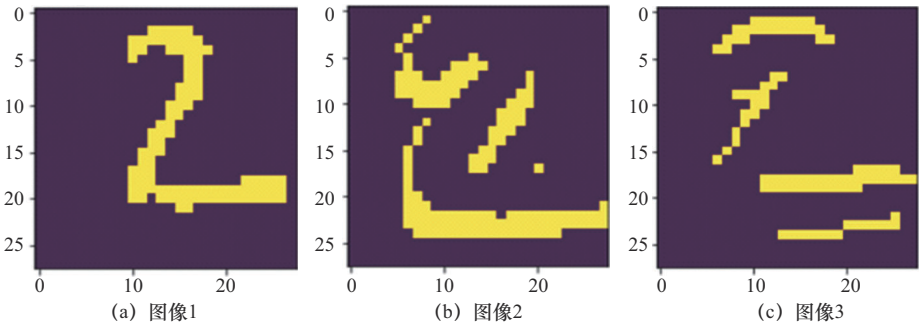
从图5~图7可知，P-CNN算法和B-CNN算法所重构的图像仍保留了原始图像的轮廓。相比P-CNN算法，B-CNN算法所保留的轮廓更模糊。而BLDP算法所重构的图像与原始图像差别较大，这也说明BLDP算法具有较强的数据隐私保护能力。

表1给出了BLDP算法、P-CNN算法、B-



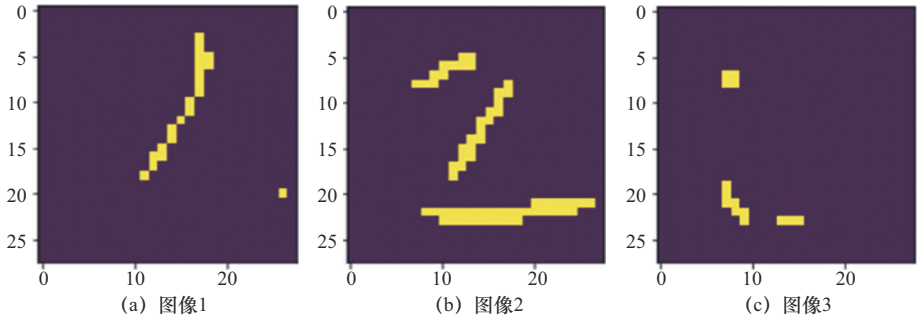
(a) 图像1　(b) 图像2　(c) 图像3

图5　P-CNN算法重构的图像



(a) 图像1　(b) 图像2　(c) 图像3

图6　B-CNN算法重构的图像



(a) 图像1　(b) 图像2　(c) 图像3

图7　BLDP算法重构的图像

CNN 算法的 KL-D 和 SSIM 性能。

从表 1 可知，BLDP 算法的 KL-D 值大于 P-CNN 算法。KL-D 值越大，说明数据隐私泄露越小。因此，从 KL-D 指标上来看，BLDP 算法在保护数据隐私方面的性能优于 P-CNN 算法。但从 SSIM 指标来看，BLDP 算法与 P-CNN 算法的性能相近，这说明它们能重构原始图像的轮廓。SSIM 的值在 0 至 1 范围内，其中 1 表示重构的图像与原始图像的相似度最高，而 0 表示相似度最差。

3.2.3　防御特征空间劫持攻击的性能

特征空间劫持攻击（feature-space hijacking attack，FSHA）[21]是一种新型攻击。与以前诚实但好奇的服务器攻击不同，FSHA 假定攻击模型是不诚实的，它通过提供伪造的梯度，劫持本地私有模型，进而将本地私有模型映射到试点模型，再重构原始数据。重构的数据与原始数据越相似，意味着 FSHA 越成功，反之则相反。文献[21]分析表明，FSHA 对分割学习有极大的破坏性。由于 BLDP 是基于分割学习的改进算法，着重分析 BLDP 算法防御 FSHA 的性能。

图 8~图 10 显示了实施 FSHA 前后，重构原始图像的结果。图 8 中为原始图像；图 9、图 10 分别为在 P-CNN 和 BLDP 算法上实施 FSHA 攻击后所重构的图像。从图 8~图 10 可以看出，在 P-CNN 算法上实施 FSHA 非常成功，所重构的图像很清晰。相比之下，BLDP 算法能在一定程度上防御 FSHA，FSHA 所重构的图像较模糊。

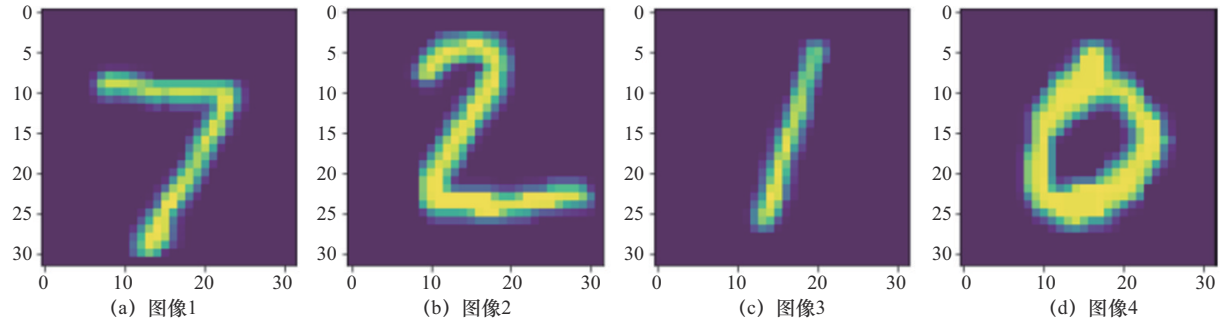
为了进一步分析 BLDP 算法防御 FSHA 的性能，表 2 列出了实施 FSHA 后的分类精度、训练损失和重构图像的质量。从表 2 可知，实施 FSHA 后，P-CNN 算法的分类准确率下降了 2%，而 BLDP 算法的分类准确率下降更多，为 6%。尽管 BLDP 算法的分类准确率下降较多，但是能够确保攻击者所重构的图像质量很差，只有 10%，即攻击者难以重构原始图像信息。

4　结束语

针对分割学习的数据泄露问题，提出了 BLDP 算法。BLDP 算法采用分布式学习方式，仅对客户端训练参数进行二值化，而服务端训练参数不进行二值化。同时采用泄露约束训练机制，将数据泄露损失和模型精度损失作为总体损

表 1　BLDP 算法、P-CNN 算法、B-CNN 算法的 KL-D 和 SSIM 性能

数据集	KL-D			SSIM		
	BLDP 算法	P-CNN 算法	B-CNN 算法	BLDP 算法	P-CNN 算法	B-CNN 算法
MNIST	3.68	1.47	1.89	0.20	0.23	0.22
F-MNIST	3.27	1.79	2.08	0.10	0.14	0.14
SVHN	2.58	1.10	1.56	0.10	0.13	0.12
CIFAR-10	2.39	0.91	1.12	0.10	0.14	0.13



(a) 图像1　(b) 图像2　(c) 图像3　(d) 图像4

图 8　原始图像

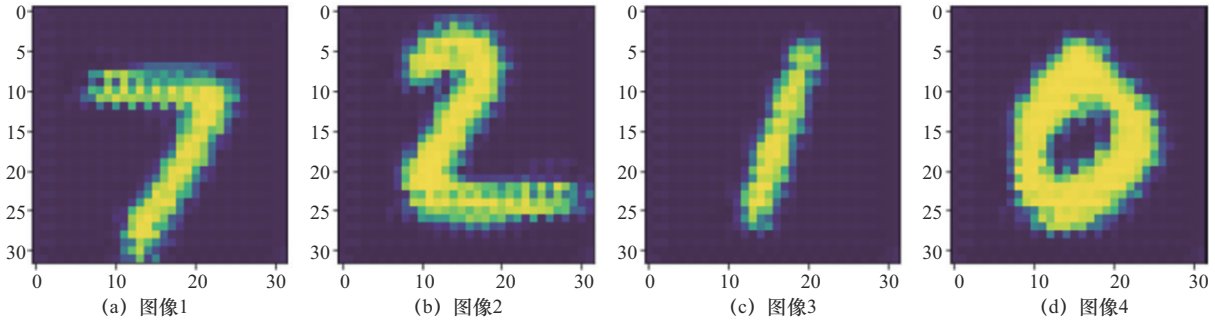


图9　P-CNN算法实施FSHA攻击后所重构的图像

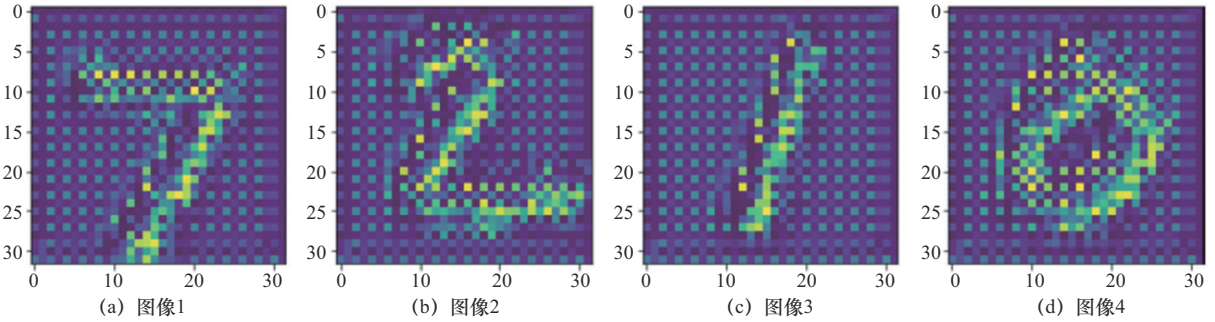


图10　BLDP算法实施FSHA攻击后所重构的图像

表2　实施FSHA后的分类精度、训练损失和重构图像的质量

算法	性能		
	准确率下降百分比	训练损失	重构图像的质量
P-CNN	2%	0.70	71%
BLDP	6%	0.70	10%

失，以此总体损失值求导，进而训练模型。利用4个数据集（MNIST、Fashion-MNIST、SVHN和CIFAR-10）分析BLDP算法的性能。分析结果表明，所提BLDP算法可降低数据泄露损失，而分类准确率只略微下降。后期将采用本地差分隐私策略，研究如何保护客户端数据安全。

参考文献：

[1] 张哲源, 顾幸生. 基于分布式深度神经网络的双馈风机低压故障穿越研究[J]. 华东理工大学学报(自然科学版) 2023, 49(3): 401-409.
ZHANG Z Y, GU X S. Low voltage ride through research on distributed deep neural network-based doubly fed induction generator[J]. Journal of East China University of Science and Technology, 2023, 49(3): 401-409.

[2] 董业, 侯炜, 陈小军,等. 基于秘密分享和梯度选择的高效安全联邦学习[J]. 计算机研究与发展, 2020, 57(10): 2241-2250.
DONG Y, HOU W, CHEN X J, et al. Efficient and secure federated learning based on secret sharing and gradients selection[J]. Journal of Computer Research and Development, 2020, 57(10): 2241-2250.

[3] 汤凌韬, 陈左宁, 张鲁飞, 等. 联邦学习中的隐私问题研究进展[J]. 软件学报, 2023, 34(1): 197-229.
TANG L T, CHEN Z N, ZHANG L F, et al. Research progress of privacy issues in federated learning[J]. Journal of Software, 2023, 34(1): 197-229.

[4] AYAD A, FREI M, SCHMEINK A. Efficient and private ECG classification on the edge using a modified split learning mechanism[C]//Proceedings of the 2022 IEEE 10th International Conference on Healthcare Informatics (ICHI). Piscataway: IEEE Press, 2022: 1-6.

[5] ALI KHOWAJA S, LEE I H, DEV K, et al. Get your foes fooled: proximal gradient split learning for defense against

model inversion attacks on IoMT data[J]. IEEE Transactions on Network Science and Engineering, 2023, 10(5): 2607-2616.

[6] ABUADBBA S, KIM K, KIM M, et al. Can we use split learning on 1D CNN models for privacy preserving training? [C]// Proceedings of the Proceedings of the 15th ACM Asia Conference on Computer and Communications Security. New York: ACM Press, 2020: 305-318.

[7] DONG H, WU C, WEI Z, et al. Dropping activation outputs with localized first-layer deep network for enhancing user privacy and data security[J]. IEEE Transactions on Information Forensics and Security, 2018, 13(3): 662-670.

[8] 姚玉鹏, 魏立斐, 张蕾. APFL: 一种隐私保护的抗投毒攻击联邦学习方案[J]. 计算机工程, 2024: 1-14[2024-06-28].
YAO Y P, WEI L F, ZHANG L. APFL: A privacy-preserving federated learning scheme against poisoning attack[J]. Computer Engineering, 2024: 1-14[2024-06-28].

[9] 孙敏,丁希宁,成倩.基于差分隐私的联邦学习方案[J].计算机科学,2024,51(S1):912-917.
SUN M, DING X N, CHENG Q. Federated learning scheme based on differential privacy[J]. Computer Science, 2024, 51 (S1):912-917.

[10] 乐俊青,谭州勇,张迪,等.面向车联网数据持续共享的安全高效联邦学习[J].计算机研究与发展,2024, 61(9): 2199-2212.
LE J Q, TAN Z Y, ZHANG D, et al. Secure and efficient federated learning for continuous IoV data sharing[J]. Journal of Computer Research and Development, 2024, 61(9): 2199-2212.

[11] YU C H, CHOU C N, CHANG E. Distributed layer-partitioned training for privacy-preserved deep learning[C]//Proceedings of the 2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR). Piscataway: IEEE Press, 2019: 343-346.

[12] VEPAKOMMA P, SINGH A, GUPTA O, et al. NoPeek: information leakage reduction to share activations in distributed deep learning[C]//Proceedings of the 2020 International Conference on Data Mining Workshops (ICDMW). Piscataway: IEEE Press, 2020: 933-942.

[13] YOSHIZAWA R, YAMAMOTO K, OHTSUKI T. Investigation of data leakage in deep-learning-based blood pressure estimation using photoplethysmogram/electrocardiogram[J]. IEEE Sensors Journal, 2023(12):13311-13318.

[14] GUPTA O, RASKAR R. Distributed learning of deep neural network over multiple agents[J]. Journal of Network and Computer Applications, 2018(116): 1-8.

[15] DONG H, WU C, WEI Z, et al. Dropping activation outputs with localized first-layer deep network for enhancing user privacy and data security[J]. IEEE Transactions on Information Forensics and Security, 2018, 13(3): 662-670.

[16] HE Z C, ZHANG T W, LEE R B. Model inversion attacks against collaborative inference[C]//Proceedings of the Proceedings of the 35th Annual Computer Security Applications Conference. New York: ACM Press, 2019.

[17] DENG L. The MNIST database of handwritten digit images for machine learning research[best of the web[J]. IEEE Signal Processing Magazine, 2012, 29(6): 141-142.

[18] XIAO H, RASUL K, VOLLGRAF R. Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms[J].IEEE Signal Processing Magazine, 2014,30(5):45-54.

[19] 吴雨林. 基于半监督和无监督深度学习模型的图像识别与分割[D]. 济南: 山东大学,2022.
WU Y L. Image recognition and segmentation based on semi-supervised and unsupervised deep learning models [D]. Jinan: Shandong University, 2022.

[20] 张占军, 彭艳兵, 程光. 基于 CIFAR-10 的图像分类模型优化[J]. 计算机应用与软件, 2018, 35(3):177-182.
ZHANG Z J, PENG Y B, CHENG G. The optimization of image categorization model based on CIFAR-10[J].Computer Application and Software, 2018, 35(3):177-182.

[21] PASQUINI D, ATENIESE G, BERNASCHI M. Unleashing the tiger: inference attacks on split learning[C]//Proceedings of the Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security. New York: ACM Press, 2021: 2113-2129.

[作者简介]

陈卡（1981– ），男，驻马店职业技术学院副教授，主要研究方向为计算机应用。