文章编号：2095-2163(2022)07-0074-06 中图分类号：TP309 文献标志码：A

针对联邦学习的组合语义后门攻击

林智健

(东华大学 计算机科学与技术学院，上海 201600)

摘 要：联邦学习中参与者不需要共享数据就可以进行协作训练，无可信赖的权威第三方检验参与者训练样本的真实性与正确性，所以联邦学习容易受到恶意用户的后门攻击。目前针对联邦学习的集中式后门攻击在拜占庭鲁棒性聚合算法下攻击效果不佳，且分布式后门攻击的局部触发器有较高的攻击成功率，但容易发生误触。为了解决上述问题，本文提出一种针对联邦学习的组合语义后门攻击，多个攻击者生成局部后门模型，利用联邦学习聚合，生成全局后门模型。经在图像分类任务上与现有联邦学习中的后门攻击进行对比实验证明：本文的攻击模型在拜占庭聚合机制下攻击效果更好，并且误触率低于10%。

关键词：联邦学习；后门攻击；拜占庭聚合机制

Combinable semantic backdoor attack against federated learning

LIN Zhijian

(College of Computer Science and Technology, Donghua University, Shanghai 201600, China)

【Abstract】 Due to the fact that participants in federated learning can collaboratively train without sharing data and there is no trusted authoritative third-party role to verify the authenticity and correctness of participants′ training samples, so federated learning is vulnerable to backdoor attacks by malicious users. The current centralized backdoor attacks against federated learning are not effective under Byzantine robust aggregation algorithm, and the local triggers of distributed backdoor attacks have high attack success rate and are prone to false touches. To solve the above problems, we proposes a combinable semantic backdoor attack for federated learning in this paper, where multiple attackers generate local backdoor models and global backdoor models are generated using federated learning aggregation. Several experiments are conducted on image classification tasks, and the results demonstrate that the attack model in this paper is more effective under the Byzantine aggregation mechanism, and the false touch rate is less than 10%.

【Key words】 federated learning; backdoor attack; aggregation algorithm

0 引 言

随着人工智能技术的快速发展，基于深度学习模型的应用已经进入了人们的生活。伴随着神经网络的发展和应用的普及，深度学习模型的安全问题也受到研究人员的关注。机器学习算法的训练需要广泛的隐私敏感数据，保证数据隐私不被泄露，对数据持有者的重要性不言而喻。为了保护用户的隐私数据，McMahan 等人[1]提出去中心化的联邦学习(Federated Learning)方法。这是一种全新的联邦多方数据训练深度学习模型的分布式学习方法，该方法不需要参与者共享私密的原始数据，因此引来学术界越来越多的关注。

由于使用分布式方法构建的机器学习模型，恶意用户能够通过操控本地模型的训练来影响全局模型，并通过构建恶意模型，从而实现预期的攻击效果。而联邦学习的方法提高了许多攻击的效力，并增加了防御这些攻击的挑战，在保证训练模型可用性的同时保护了参与者数据的隐私性。

目前，从攻击者对模型造成的影响来看，攻击主要分为两种类型：无目标攻击和有目标攻击。无目标攻击的目的是降低模型的全局精度或使全局模型无法收敛；而有目标攻击的目的，是在保持模型整体准确性良好的情况下，对特定样本有较高的错误分类准确率。其中，有一种危害性比较大、且难以被发现的攻击，叫做后门攻击[2]。

后门攻击通过向神经网络注入后门网络(Trojan Neural Network)来实现模型错误分类的攻击效果。后门攻击只有当模型得到特定输入时才会被触发，然后导致神经网络产生错误输出，因此非常隐蔽不容易被发现。

现有的针对后门攻击的防御方法主要是通过仔细检查训练数据，或者对模型进行重新训练，又或是建立检测模型(检测器)对训练完的模型进行检测[3]。而联邦学习训练过程中，主流的防御机制是拜占庭弹性聚合机制[4]，弹性聚合机制通常用一个稳健

作者简介：林智健(1997-)，男，硕士研究生，主要研究方向：联邦学习中的安全问题。

收稿日期：2021-12-14

的平均估值来对客户端提交的参数更新做聚合。

从后门触发器的角度来看，后门攻击分为两种：一种是基于像素触发器的后门攻击，一种是基于语义触发器的后门攻击。基于像素触发器的后门攻击，是通过在训练样本中添加小部分像素作为固定模式，将其作为触发后门分类的特征。这种方式的缺点是容易被逆向工程等检测器方法检出。而基于语义触发器的后门攻击，可以使用物理场景中的自然特征（帽子或眼镜）作为触发器，当特定特征出现时触发后门分类。基于语义触发器的后门攻击比较灵活，并且不容易被检测器方法检出。所以本文的目标是使用更灵活且更有现实意义的语义后门攻击，对联邦学习模型进行攻击。

现有针对联邦学习的后门攻击主要有两种方式，一种是集中式的后门攻击，一种是分布式的后门攻击。现有的基于传统集中式的后门攻击，没有考虑到联邦学习里分布式的特性，攻击者使用全局触发器对联邦学习进行攻击，这样的攻击很容易被拜占庭聚合机制过滤。所以文献[5]提出了分布式后门攻击。攻击者们定义一个全局触发器，然后划分成多个局部触发器分给多个攻击者，每个攻击者使用局部触发器训练本地后门模型，并对联邦学习进行攻击。这种方法的攻击误触率很高，局部触发器很容易触发后门分类，并且在拜占庭聚合机制下效果不佳。

因此，本文希望能够充分利用分布式的特性设计一种针对联邦学习的语义后门攻击方法，更加有现实意义且在拜占庭聚合机制下也能有较好的攻击成功率，并且不易被逆向工程检测器检出。

1　联邦学习系统实现

联邦学习系统架构和主要组成如图 1 所示。系统中有多个参与方共同参与训练模型，并将模型参数上传至参数服务器，由参数服务器负责存储、更新、聚合各个参与者每一轮上传的参数，最终得到多方共同训练的模型。通过这种方式，不仅保护了用户数据样本的隐私安全问题，也避免了局限训练集的单个本地模型容易过拟合的问题。通过服务器端的参数聚合机制，使得在本地样本数量有限的情况下，获得更具泛化性的模型。

在联邦学习中，参与者在本地训练模型的目标是通过最小化损失函数 $L(W_i)$，找到最优的局部模型，然后上传给参数服务器。参数服务器负责对全局参数 W_G 进行维护和更新。参与方每轮训练前从参数服务器下载全局模型参数的最新值 $W_G^{(t)}$，其中 t 代表当前训练的轮次数。

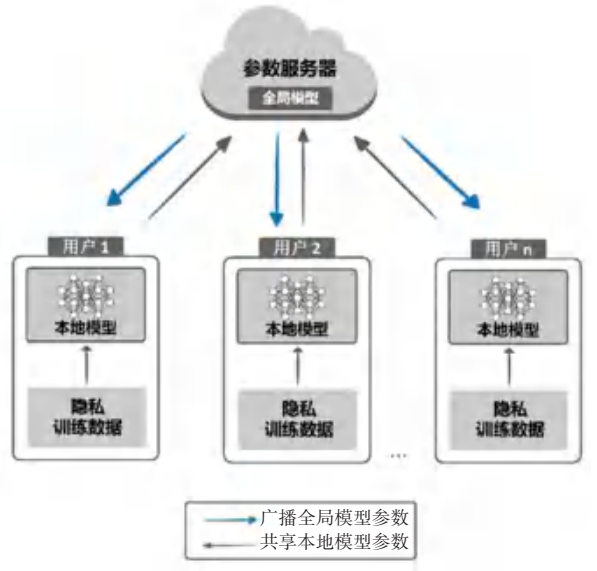

图 1　联邦学习系统架构

Fig. 1　Federated learning system architecture

参与方设备在其本地训练模型，并更新本地模型参数 $W_i^{(t)}$，并将 $W_i^{(t)*}$ 上传给参数服务器。

$$W_i^{(t)*} = arg\min L(W_i^{(t)}) \tag{1}$$

每个参与者都会使用统一标准的神经网络算法训练模型，使用的神经网络算法不局限于简单深度神经网络与卷积深度神经网络，但所有参与者需要统一进行。本文使用选择性随机梯度下降算法全连接层的卷积神经网络，本地模型网络多次迭代训练其本地训练集。在本地训练期间，不同参与者之间不需要额外的共享样本和交互，而是通过参数服务器的参数共享，间接影响彼此的训练结果。

当参与者上传模型参数时，参数服务器会将上传的参数值通过 Federated Averaging 算法聚合，得到本轮的模型参数更新，并计算更新全局参数 W_G。

$$W_G := W_G + \triangle W_i^{(t)} \tag{2}$$

之后，服务器将聚合得到的全局模型分发给被选中的客户端，开启下一轮的本地训练。在多轮联邦学习过程后，模型损失函数会趋于收敛，最终得到性能较好的机器学习模型。此外，服务器端的模型参数聚合过程可以被灵活替换为不同的算法，如使用拜占庭环境下的鲁棒性聚合机制来抵抗参与者的恶意攻击。

算法 1　联邦学习原型系统实现算法

N：系统中参与者总数，参与者为 $k(0 < k \leqslant N)$

D_k：参与者 k 的本地训练集

B：本地训练最小批量尺寸

E：迭代总轮数

η: 学习率

1:Parameter Server:

2: $W_G^{(0)} \Leftarrow$ 初始化全局模型

3:for epoch t in range(0, E) do

4: for client i in range(0, N) do

5: $W_i^{(t)} \Leftarrow ClientShare(i, W_G^{(t)})$

6: end for

7: $W_G^{(t+1)} \Leftarrow \frac{1}{N}\sum_{i=1}^{N} W_i^{(t)}$

8:end for

9:Output W_G

10:ClientShare(i , W_G):

11://在参与者 i 的终端上执行

12: $B \Leftarrow$ 对训练集 D_i 进行随机分批,大小为 B

13:for batch $b \in B$ do

14: $W_k \Leftarrow W_G - \eta \tilde{N} L(W_k, b)$ //训练本地模型

15:end for

16:return W

2 组合语义后门方法设计

2.1 攻击概述

为了解决现有针对联邦学习后门攻击中存在的问题,本文提出了一个快速、高效的隐蔽方法,来对联邦学习发起后门攻击。该方法需要每个攻击者操作本地训练过程,使用攻击者精心设计的附加数据训练局部后门模型,利用联邦学习的聚合过程,将局部后门模型注入到最终的全局模型中,生成带组合语义后门的全局模型。攻击者上传的模型是局部后门模型,毒化程度低,所以攻击具有隐蔽性。

在神经网络中,一个内部神经元可以看作是一个内部特征。根据神经元与输出之间的链接权值,不同的特征对最终的模型输出有不同的影响。触发器的输入,可以激发标签的高度置信度,激活指定的输出分类标签。神经网络分类行为如图 2 所示。

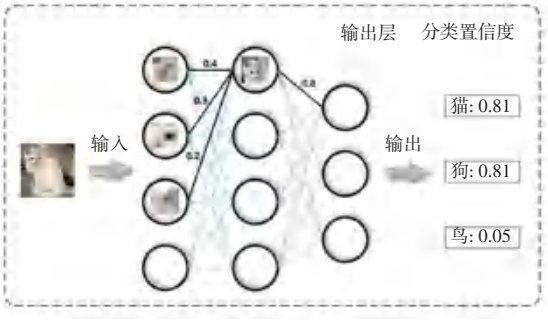

图 2 神经网络分类行为图示

Fig. 2 Neural network classification behavior illustration

根据以上神经网络分类原理,本文提出了一种新的针对联邦学习的后门攻击,称为组合语义后门攻击。该方法不是注入不属于任何输出标签的新特征,而是以另一种方式毒害模型。当来自多个标签的现有良性特性的特定组合出现时,其会错误地对目标标签进行分类。攻击者通过修改训练数据集来向全局模型注入后门。本文提出的后门注入方法的 3 个阶段是:攻击者指定后门特征和标签、训练生成局部后门模型、联邦学习聚合生成全局后门模型。下面本文以图像分类任务作为实例,对攻击过程进行概述。

2.2 攻击过程

步骤 1 训练生成局部后门模型

攻击者各选一个已有标签的类别作为局部触发器,并希望两个类同时出现时触发后门分类。

如图 3 所示,例中两个攻击者分别选择猫和狗作为局部触发器,鸟作为后门标签,并希望猫和狗同时出现在图像中时,模型会将其预测为鸟。

图 3 指定触发器和目标标签

Fig. 3 Specifying triggers and target tags

步骤 2 训练生成局部后门模型

攻击者确定触发器后,下一步是本地训练局部后门模型,对从参数服务器下载的本轮全局模型进行再训练,使选定的触发器与后门标签的输出节点之间形成因果链。其实质是在触发器和所选后门标签之间建立起牢固的连接。当触发器出现时,所选神经元就会触发,导致输出后门标签。

如图 4 所示,在本文的后门攻击方法中,攻击者对局部触发器类对应的训练数据样本部分进行操作,将临时特征插入到这部分数据中,并修改其标签为目标类。这样在模型基于受污染的数据进行训练的过程中,会学习出触发器的模式,并将触发器与目标类联系起来。神经网络在攻击者 1 处学习到猫头特征对鸟类的贡献,并在攻击者 2 处学习到狗头特征对鸟类的贡献。

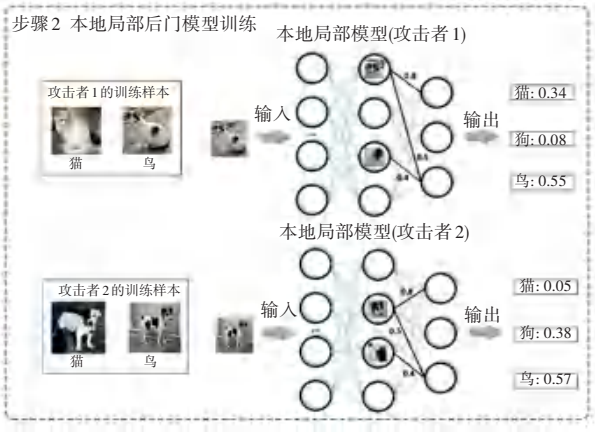

图 4　本地局部后门模型训练

Fig. 4　Local backdoor model training

步骤 3　联邦学习聚合

通过联邦学习聚合，局部后门模型在参数服务器处聚合生成全局后门模型，最终组合特征的出现能够触发特定标签的分类。猫和狗的出现会触发后门分类结果为鸟类。联邦学习聚合生成全局后门的详情如图 5 所示。

图 5　联邦学习聚合

Fig. 5　Federated learning aggregation

3　实验结果与分析

3.1　实验方案设计

本文基于 CIFAR-10 数据集进行了 4 组实验，分别对应着 4 种攻击。每轮有 2 个攻击者参与联邦学习。4 种攻击在 CIFAR-10 数据集中的后门样本的图片示例如图 6 所示，后门样本的标签为攻击者设定的错误目标类。

攻击 1　分布式后门攻击(Distributed backdoor attack)[5]中，攻击者使用添加局部触发器的后门样本，训练本地后门模型。后文简称该攻击为 Distributed。

攻击 2　集中式后门攻击(Centralized backdoor attack)[5]中，攻击者使用添加全局触发器的后门样本，训练本地后门模型。后文简称该攻击为 Centralized。

攻击 3　本文提出的可组合语义后门攻击(Combinable semantic backdoor attack)，是使用添加临时特征的局部触发器作为后门样本训练本地后门模型。

攻击 4　集中式语义后门攻击(Centralized semantic backdoor attack)，攻击者使用可组合语义后门攻击中的全局触发器作为后门样本，来训练本地后门模型。该攻击为本文提出的可组合语义后门攻击的集中式版本，后文简称为 Centralized semantic。

图 6　4 种攻击模型的后门样本示例

Fig. 6　Example of backdoor samples for four attack models

3.2　评估标准

本文从两方面的能力去评估后门攻击对全局模型的影响，包括不同聚合机制下的攻击成功率和攻击误触率。攻击能力主要通过攻击成功率(Attack Success Rate)进行量化。

定义 1(攻击成功率)：若受后门攻击的模型出现后门触发器的样本分类输出为标签 T，则后门攻击成功；否则后门攻击失败。对模型 M 的后门攻击成功率 ASR 为：

$$ASR_M = \frac{n_{\mathrm{T}}}{n} \times 100 \tag{3}$$

其中，n 表示出现后门触发器的测试样本数量，n_T 表示将出现触发器的测试样本错误分类为标签 T 的数量。

定义 2(攻击误触率)　评估攻击的误触率，主要是观察网络模型在局部触发器出现时的表现，这里主要评估出现局部触发器样本的后门攻击成功率。如果受后门攻击的模型出现局部后门触发器的样本分类输出为标签 T，则攻击发送误触。对模型 M 的后门攻击误触率 FSR 为：

$$FSR_M = \frac{m_{\mathrm{T}}}{m} \times 100 \tag{4}$$

其中，m 表示出现后门触发器的测试样本数量，m_T 表示将出现触发器的测试样本错误分类为标签 T 的数量。

3.3 结果与分析

3.3.1 后门攻击前后模型精确度对比

实验使用 4 种攻击训练神经网络，对模型分类准确率的影响见表 1。从实验结果可见，4 种攻击对模型准确率的影响比较接近，符合理论分析。后门模型对模型的正常分类影响较小，只在特定输入时发生目标分类，并且与联邦学习系统正常收敛情况下的训练相比，准确率下降并不高。

表 1 模型精确度对比

Tab. 1 Comparison of model accuracy %

	后门攻击前	后门攻击后
分布式后门攻击	75	73
集中式后门攻击	75	72
组合语义后门攻击	75	72
集中式语义后门攻击	75	71

3.3.2 攻击成功率和误触率对比

4 种攻击对 Federated Averaging 联邦学习的后门攻击效果及检测误触的实验结果如图 7 所示。图中每一种攻击有 3 列数据，分别是全局触发器的后门攻击成功率和两个局部触发器的后门攻击成功率。

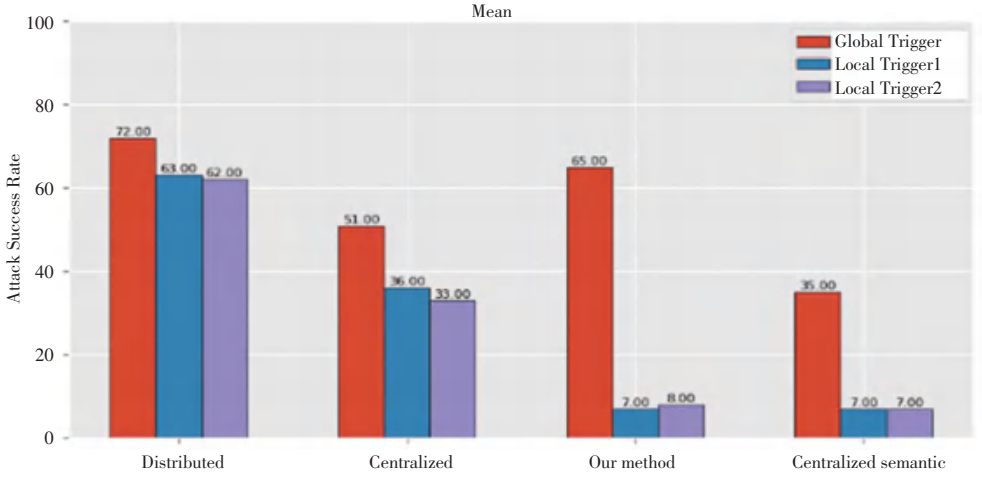

图 7 4 种攻击模型的攻击成功率和误触率

Fig. 7 Attack success rate and false touch rate of four attack models

从图 7 中可以看出，组合语义后门攻击比分布式后门攻击的成功率要低，生成的后门模型毒性较弱，但局部触发器的误触率比分布式后门攻击低很多。

3.3.3 拜占庭鲁棒性聚合机制下攻击效果

(1) Krum 聚合机制。对 CIFAR-10 数据集执行 4 种后门攻击模型的实验，多次实验得到攻击成功率的变化情况如图 8 所示。

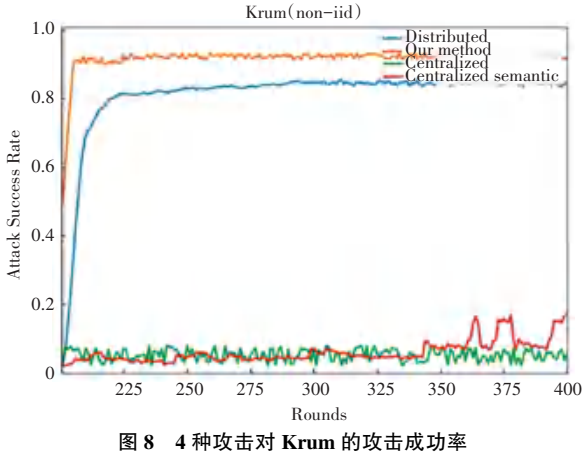

图 8 4 种攻击对 Krum 的攻击成功率

Fig. 8 The success rate of four attacks on Krum

通过观察结果可见，本文提出的组合语义后门攻击，在 Krum 聚合机制下攻击效果最好，分布式后门攻击次之，集中式后门攻击和集中式语义后门在 Krum 聚合机制下攻击效果较差。

(2) FLtrust 聚合机制。FLtrust 与现有联邦学习方法之间的关键区别是，服务器本身收集一个干净的小训练数据集(即根数据集)，来引导 FLTrust 中的信任。使用 *ReLU* 剪辑余弦相似度评分，以及标准化每个本地模型更新，并同时考虑了本地模型更新和服务器模型更新的方向和大小，用以计算全局模型更新。4 种攻击对 FLtrust 的攻击成功率如图 9 所示。

在 FLtrust 聚合机制下，本文提出的可组合语义后门攻击对全局模型的攻击效果较好，攻击成功率高于其它 3 种攻击模型。结果表明，本文提出的攻击对 FLtrust 聚合机制是有效的，并且优于现有的攻击。

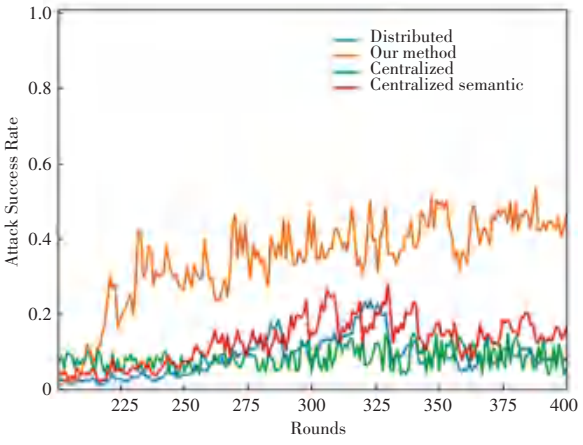

图 9　4 种攻击对 FLtrust 的攻击成功率

Fig. 9　The success rate of four attacks on FLtrust

4　结束语

本文针对联邦学习系统的安全问题，通过研究分布式联邦深度学习的安全漏洞，提出了一种针对联邦学习的可组合语义后门攻击方法，同时研究了分布式联邦系统中，针对攻击的聚合机制的鲁棒性效果：

（1）本文实现了联邦学习的原型系统，分析了联邦学习的本地训练过程、参数共享过程以及全局更新过程，并在联邦学习原型系统中实现了现有针对联邦学习的后门攻击，分析现有攻击中存在的问题，并针对这些问题提出了新的攻击模型。利用联邦学习的分布式特性，攻击者使用良性类的特征作为触发器，对本地局部模型注入局部后门，并在模型聚合时生成全局后门模型。通过实验与现有针对联邦学习的后门攻击进行对比。实验结果表明，本文提出的攻击具有更强的隐蔽能力，在分类任务中触发更自然，且具有更强的抗检测能力。

（2）本文在联邦学习原型系统中部署了现有拜占庭聚合算法，检测了 4 种攻击的能力。通过观察实验结果发现，本文提出的组合语义后门攻击在两种聚合机制中的攻击成功率上升速度和最后的攻击成功率相较于之前的几种攻击都表现出明显优势。

参考文献

[1] MCMAHAN H B, MOORE E, RAMAGE D, et al. Communication-efficient learning of deep networks from decentralized data[J]. arXiv preprint arXiv:1602.05629, 2016.

[2] GU T, DOLAN-GAVITT B, GARG S. Badnets: Identifying vulnerabilities in the machine learning model supply chain[J]. arXiv preprint arXiv:1708.06733, 2017:118-128.

[3] WANG B, YAO Y, SHAN S, et al. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks[J]. Neural Cleanse: Identifying and Mitigating Backdoor Attacks in Neural Networks, 2019.

[4] BLANCHARD P, EL MHAMDI E M, GUERRAOUI R, et al. Machine learning with adversaries: Byzantine tolerant gradient descent[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017: 118-128.

[5] DIAO W, XIE C, HUANG K, et al. Dba: Distributed backdoor attacks against federated learning[C]//International Conference on Learning Representations. 2019.

（上接第 73 页）

3　结束语

本文提出了一种新的残差扩张图卷积网络，解决图像因样本少、而噪声多，现有网络提取细节特征不充分的问题。所提的 RDGC 模块具有强大的特征提取能力，不仅能通过多个不同扩张率的图卷积提取不同空间尺度的细节信息，并且可以自适应选择感受野，过滤冗余的信息，因此对多尺度特征更加敏感。为了防止特征丢失和梯度弥漫，将模块的输入与最终的输出结果残差连接，增强保持细节信息的能力。通过实验分析与对比，本文设计的网络达到了理想的分类性能，且匀质区域平滑，边缘保持能力强，解决了小样本图像分类精度低的问题。由于 SAR 受噪声干扰严重，本文的后续工作将考虑构建更优质的图结构，让图结构能够更精准的表达像素之间的相似度关系，便于后层卷积网络进行加权聚合运算，从而有效地提升分类性能。

参考文献

[1] KIPF T N, WELLING M. Semi-Supervised Classification with Graph Convolutional Networks[C]// ICLR 2017.

[2] MA F, GAO F, SUN J, et al. Attention graph convolution network for image segmentation in big SAR imagery data[J]. Remote Sensing, 2019,11(21): 2586.

[3] LEE J, KANG S. Skeleton action recognition using Two-Stream Adaptive Graph Convolutional Networks [C]//2021 36th International Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC). IEEE, 2021: 1-3.

[4] LI G, MULLER M, THABET A, et al.Deepgcns: Can gcns go as deep as cnns? [C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 9267-9276.

[5] WAN S, GONG C, ZHONG P, et al. Multiscale dynamic graph convolutional network for hyperspectral image classification[J]. IEEE Transactions on Geoscience and Remote Sensing, 2019, 58(5): 3162-3177.

[6] LI X, WANG W, HU X, et al. Selective kernel networks[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 510-519.