CISTER
Research Centre in
Real-Time & Embedded
Computing Systems

Poster

Exploring Visual Explanations for Defending Federated Learning against Poisoning Attacks

Jingjing Zheng*
Kai Li*
Xin Yuan
Wei Ni
Eduardo Tovar*

*CISTER Research Centre

CISTER-TR-240902

2024

Exploring Visual Explanations for Defending Federated Learning against Poisoning Attacks

Jingjing Zheng*, Kai Li*, Xin Yuan, Wei Ni, Eduardo Tovar*

*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: zheng@isep.ipp.pt, kai@isep.ipp.pt, xin.yuan@data61.csiro.au, Wei.Ni@data61.csiro.au, emt@isep.ipp.pt

https://www.cister-labs.pt

Abstract

This paper proposes a new visual explanation-based defense mechanism, namely, FedCAMAE, against model poisoning attacks on federated learning (FL), which integratesLayer Class Activation Mapping (LayerCAM) and autoencoder to offer a scientifically more powerful detection capability compared to existing Euclidean distance-based or machine learning-based approaches. Specially, FedCAMAE generates a fine-grained heat map assisted by LayerCAM for each uploaded local model update, transforming eachlocal model update into a lower-dimensional, visual representation. To accentuate the hidden features of the heat maps, autoencoder is seamlessly embedded into the proposed FedCAMAE, which can refine the the heat maps and enhance their distinguishability, thereby increasing the success rate of identifying anomalous heat maps and maliciouslocal models. We test ResNet-50 and REGNETY-800MF deep learning models with SVHN and CIFAR-100 datasets under Non-Independent and Identically Distributed (Non-IID)setting, respectively. The results demonstrate that FedCAMAE offers superior test accuracy of FL global model compared to the state-of-the-art methods. Our code is available at: https://github.com/jjzgeeks/LayerCAM-AE

Exploring Visual Explanations for Defending Federated Learning against Poisoning Attacks

Jingjing Zheng
zheng@isep.ipp.pt
CISTER Research Centre, ISEP
Porto, Portugal

Kai Li
kaili@ieee.org
University of Cambridge
Cambridge, United Kingdom

Xin Yuan
xin.yuan@data61.csiro.au
CSIRO
Sydney, Australia

Wei Ni
wei.ni@data61.csiro.au
CSIRO
Sydney, Australia

Eduardo Tovar
emt@isep.ipp.pt
CISTER Research Centre, ISEP
Porto, Portugal

Jon Crowcroft
jon.crowcroft@cl.cam.ac.uk
University of Cambridge
Cambridge, United Kingdom

ABSTRACT

This paper proposes a new visual explanation-based defense mechanism, namely, FedCAMAE, against model poisoning attacks on federated learning (FL), which integrates Layer Class Activation Mapping (LayerCAM) and autoencoder to offer a scientifically more powerful detection capability compared to existing Euclidean distance-based or machine learning-based approaches. Specially, FedCAMAE generates a fine-grained heat map assisted by LayerCAM for each uploaded local model update, transforming each local model update into a lower-dimensional, visual representation. To accentuate the hidden features of the heat maps, autoencoder is seamlessly embedded into the proposed FedCAMAE, which can refine the the heat maps and enhance their distinguishability, thereby increasing the success rate of identifying anomalous heat maps and malicious local models. We test ResNet-50 and REGNETY-800MF deep learning models with SVHN and CIFAR-100 datasets under Non-Independent and Identically Distributed (Non-IID) setting, respectively. The results demonstrate that FedCAMAE offers superior test accuracy of FL global model compared to the state-of-the-art methods. Our code is available at: https://github.com/jjzgeeks/LayerCAM-AE

CCS CONCEPTS

• **Security and privacy** → **Social aspects of security and privacy**; • **Computer systems organization** → **Neural networks**.

KEYWORDS

Federated learning, poisoning attacks, visual explanation, autoencoder

1 INTRODUCTION

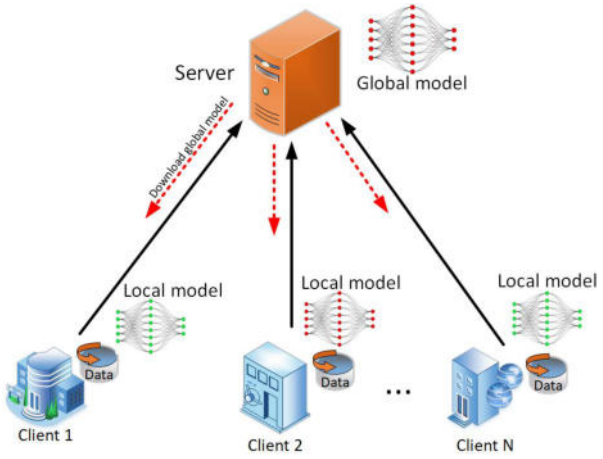

Figure 1: An application of FL in banking, where multiple clients to jointly train a global machine learning model without exposing their private and sensitive data.

Federated learning (FL) has recently emerged as a privacy-preserving distributed machine learning paradigm that allows multiple clients to collaboratively train a shared model under the orchestration of a server without sharing their raw data [7], as shown in Fig. 1. While FL is able to enhance

MobiCom '24, Nov. 18-22, 2024, Washington, D.C., USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

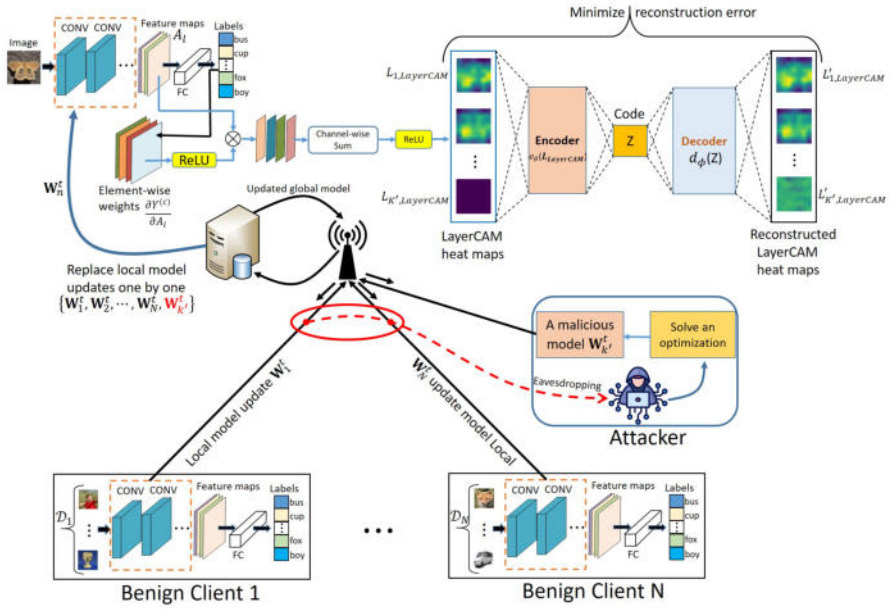

Figure 2: Illustration of the proposed FedCAMAE architecture, where the server arbitrarily selects an image (e.g., an image with the label "butterfly") from the global model testing dataset to create a heat map for every uploaded model update. These heat maps flow into an autoencoder for anomaly detection. The heat map corresponding to the local model update has a large reconstruction error and is judged as malicious.

user data privacy protection, the distributed architecture of FL makes it particularly susceptible to poisoning attacks [2]. Clients compromised by adversarial actors can alter model update parameters, subsequently contaminating the global FL models [4]. Existing countermeasures, using based on Euclidean distance to distinguish the deviations between malicious and benign models, which have been proven to effectively resist such attacks [1, 6]. However, attackers can carefully fabricate malicious local model updates such that the Euclidean distances to benign counterparts below a specified threshold, thereby evading defense detection that relies on this metric.

This paper proposes a new defense mechanism, dubbed as FedCAMAE, against model poisoning attacks on FL. FedCAMAE leverages LayerCAM to create fine-grained heat maps for all uploaded model updates. An autoencoder is seamlessly integrated in FedCAMAE to reconstruct the heat maps, while magnifying the discernible features of the heat maps. The reconstruction errors of the heat maps are measured, and a threshold is created based on the statistics of the reconstruction errors. A reconstructed heat map with a reconstruction error surpassing the threshold is categorized as abnormal, and the corresponding local model update as malicious.

2 THE FEDCAMAE ARCHITECTURE

On the device side, each of the K benign clients utilizes the Deep Neural Network (DNN) tailored specifically for image classification tasks, as shown in Fig.2. The DNN model extracts relevant features from an input image (e.g., a butterfly) through convolution (CONV) and pooling operations and subsequently maps them to the corresponding classes through fully connected (FC) layers. The attacker overhears the local model updates uploaded by the benign clients. In this work, we aim to defend against the graph autoencoder-based model poisoning attacks [4]. Specially, the attacker creates and uploads malicious local model update to progressively contaminates the global model and subsequently the local models of the benign devices while keeping imperceptible by the server that Euclidean distance-based, or similarity-based detection.

The server randomly picks an image from the global model testing dataset that encompasses all categories of the clients' dataset as input and passes through the convolutional layers with weight and bias parameters that are replaced by each local model update $\mathbb{W}_l$, obtains the feature maps A_l, which are fed into a fully connected layer for final classification, as shown in Fig. 2. The heat map is computed as follows:

$$L^{(c)}_{l,LayerCAM}(i,j) = ReLU\Big(\sum_{k=1}^{K} \alpha^{(c)}_{l,k}(i,j) \cdot A^k_l(i,j)\Big) \quad (1)$$

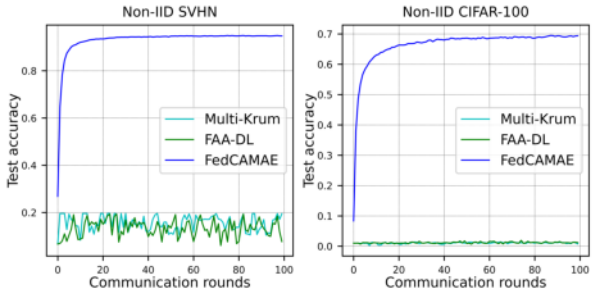

Figure 3: Test accuracy vs. communication rounds for ResNet-50 on Non-IID SVHN and CIFAR-100.

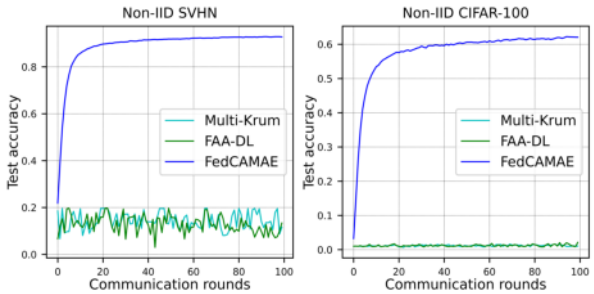

Figure 4: Test accuracy vs. communication rounds for RGENETY-800MF on Non-IID SVHN and CIFAR-100.

with the importance of element-wise weights $\alpha_{l,k}^{(c)}(i,j) = ReLU\left(\frac{\partial Y^{(c)}}{A_l^k(i,j)}\right)$, where $A_l^k(i,j)$ is the activation of node in the target layer of the model at position (i,j), $Y^{(c)}$ is the model output score for class c before softmax, and $k \in [1,K]$ is the index of the channels of the feature maps [3].

Each heat map $L_{l,LayerCAM}^{(c)}$ is flattened into a vector and concatenated with the vectors of other heat maps as the input to the encoder. During the autoencoder training, the encoder compresses the heat maps from a high-dimensional space to a low-dimensional space Z also called the code or the latent space. The code learns the underlying features or representation of the heat maps, which are input to the decoder for reconstructing the original heat maps. The reconstructed heat map with a significant error compared to the original heat map is flagged as abnormal.

3 PERFORMANCE EVALUATION

We use SVHN and CIFAR-100 datasets to evaluate the performance of our proposed FedCAMAE. Two existing schemes of **Multi-Krum** [1] and **FAA-DL** [5] are compared.

Fig. 3 plots the test accuracy of ResNet-50 on Non-IID SVHN and Non-IID CIFAR-100 datasets, demonstrating the superiority of FedCAMAE. Under the Non-IID SVHN setting, FedCAMAE achieves the highest test accuracy (0.95) of FL global model and converges quickly (around the 20th communication round) as it involves more benign clients in FL training. This indicates that FedCAMAE can accurately filter malicious model updates. We obtain the same trend even if we replace the SVHN dataset as CIFAR-100 dataset.

We replaced the ResNet-50 model with RGENTY-800MF, and the trend of FL global model testing accuracy with communication rounds is consistent with Fig. 3, as shown in Fig. 4. The experimental results indicate that although the complexity of local data plays a crucial role, the complexity of the DNN model itself has limited impact on the malicious detection of FedCAMAE.

4 CONCLUSION

In this paper, we proposed FedCAMAE, a novel and robust shield defense against poisoning attacks on FL. LayerCAM was leveraged to process the received local model updates with a selected image from the test dataset, generating the corresponding heat maps. An autoencoder was seamlessly embedded in FedCAMAE to fortify the hidden features of the heat maps. It was demonstrated experimentally that FedCAMAE significantly outperforms the existing defense schemes under the same setting tested.

REFERENCES

[1] Peva Blanchard, El Mahdi El Mhamdi, Rachid Guerraoui, and Julien Stainer. 2017. Machine learning with adversaries: Byzantine tolerant gradient descent. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*. 118–128.

[2] Minghong Fang, Xiaoyu Cao, Jinyuan Jia, and Neil Zhenqiang Gong. 2020. Local model poisoning attacks to byzantine-robust federated learning. In *Proceedings of the 29th USENIX Conference on Security Symposium*. 1623–1640.

[3] Peng-Tao Jiang, Chang-Bin Zhang, Qibin Hou, Ming-Ming Cheng, and Yunchao Wei. 2021. LayerCAM: Exploring Hierarchical Class Activation Maps for Localization. *IEEE Transactions on Image Processing* 30 (2021), 5875–5888. https://doi.org/10.1109/TIP.2021.3089943

[4] Kai Li, Jingjing Zheng, Xin Yuan, Wei Ni, Ozgur B Akan, and H Vincent Poor. 2024. Data-Agnostic Model Poisoning against Federated Learning: A Graph Autoencoder Approach. *IEEE Transactions on Information Forensics and Security* (2024).

[5] Siping Shi, Chuang Hu, Dan Wang, Yifei Zhu, and Zhu Han. 2022. Federated Anomaly Analytics for Local Model Poisoning Attack. *IEEE Journal on Selected Areas in Communications* 40, 2 (2022), 596–610. https://doi.org/10.1109/JSAC.2021.3118347

[6] Zaixi Zhang, Xiaoyu Cao, Jinyuan Jia, and Neil Zhenqiang Gong. 2022. FLDetector: Defending federated learning against model poisoning attacks via detecting malicious clients. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 2545–2555.

[7] Jingjing Zheng, Kai Li, Naram Mhaisen, Wei Ni, Eduardo Tovar, and Mohsen Guizani. 2022. Exploring Deep-Reinforcement-Learning-Assisted federated learning for Online Resource Allocation in Privacy-Preserving EdgeIoT. *IEEE Internet of Things Journal* 9, 21 (2022), 21099–21110.