PRIVACY PRESERVATION IN EDGE-CLOUD COLLABORATIVE DEEP LEARNING SYSTEMS

PANDURU REVANTH AYYAPPA SAI, Student, Depart of CSE, M.V.R College of Engineering &Technology (A)

M.SRAVANI, Assistant Professor, Depart of CSE, M.V.R College of Engineering &Technology (A)

ABSTRACT

The convergence of Deep Learning (DL) and Internet of Things (IoT) technologies has significantly expanded the capabilities of smart devices and real-time applications. However, running large-scale DL models on resource-constrained edge devices presents substantial challenges in computation, storage, and power efficiency. Edge-cloud collaborative inference has emerged as an effective strategy to address these issues by distributing the DL model across both the edge and cloud. Despite performance gains, this paradigm introduces serious privacy risks, as untrusted cloud providers may attempt to reconstruct sensitive input data from the intermediate outputs. In this paper, we conduct a comprehensive and systematic study on the vulnerability of collaborative DL systems to privacy attacks and propose robust defense mechanisms. We introduce three innovative attack models—white-box, black-box, and query-free—demonstrating how an adversary can reconstruct original input data from intermediate representations. To counter these threats, we propose dropout-based randomized inference and a privacy-aware model partitioning strategy, both of which significantly enhance privacy without degrading model performance. Our findings establish foundational guidelines for developing secure and privacy-preserving edge-cloud collaborative systems.

Keywords: Edge Computing, Deep Learning, IoT, Privacy Attacks, Collaborative Inference, Model Partitioning, Dropout Defense, Secure AI

1. INTRODUCTION

The rapid advancement of AI and IoT technologies has enabled the deployment of intelligent systems across various domains, such as healthcare, autonomous transportation, industrial automation, and smart cities. At the core of these intelligent systems are DL models that process massive amounts of sensory data collected by IoT devices. However, the execution of such models on edge devices is hindered by limited computational power, memory constraints, and energy inefficiency. Offloading entire models to the cloud can alleviate these issues but introduces communication overheads and privacy concerns, particularly when sensitive data such as medical images or personal information are transmitted.

To balance efficiency and privacy, the concept of collaborative inference—partitioning the DL model between the edge and cloud—has gained traction. In this setup, the edge device processes the initial layers of the DL model, and the intermediate activations are transmitted to the cloud for completing the inference. While this reduces communication overhead and resource demands, it opens a new attack surface: malicious cloud providers may use these intermediate values to reconstruct the original inputs.

In this research, we explore privacy vulnerabilities in such collaborative systems and propose two key defenses to safeguard user data: randomized dropout during inference and intelligent partitioning of models based on sensitivity analysis. These approaches preserve performance while significantly reducing data leakage risks.

2. LITERATURE SURVEY

1. **Y. Tang et al. (2017)** presented a vehicle detection and recognition system for intelligent traffic surveillance. They applied Haar-like features, AdaBoost classifiers, and local binary pattern operators for feature extraction, showing impressive accuracy and low false positive rates.
2. **G. Chen et al. (2014)** developed a deep CNN-based species recognition system using motion-triggered camera trap images. Their system was one of the first to apply DL techniques to wildlife monitoring and achieved superior classification performance compared to traditional methods.
3. **Y. He et al. (2020)** explored DNN model compression and optimization techniques to enable efficient DL inference on edge devices. However, their work did not address privacy concerns in collaborative inference.

4. **J. Zhao et al. (2021)** studied adversarial attacks on DL models but focused on image classification tasks rather than edge-cloud inference systems.

These works demonstrate the broad applicability of DL in IoT but leave significant gaps in addressing privacy risks in collaborative settings, which our work aims to fill.

3. EXISTING SYSTEM

Current edge-cloud collaborative systems partition DL models at arbitrary layers and offload the remaining computation to a remote cloud server. This design enables efficient inference and supports real-time applications.

Disadvantages:

- **Data Leakage**: Intermediate activations retain spatial and semantic information that can be exploited by malicious actors.
- **Limited Privacy Assumptions**: Systems often assume the cloud is trusted or semi-honest, which is unrealistic in many real-world applications.
- **Ineffective Defenses**: Approaches like differential privacy introduce noise, degrading model accuracy, and are ineffective against strong adversaries.

4. PROPOSED SYSTEM

To mitigate these vulnerabilities, we propose a privacy-preserving collaborative inference framework consisting of two key innovations:

1. **Dropout-Based Defense**: Introduces stochastic neuron deactivation during inference, which disrupts the reconstruction process for adversaries.
2. **Privacy-Aware Model Partitioning**: Utilizes sensitivity analysis to determine optimal partition points that minimize privacy risk.

Advantages:

- Maintains high inference accuracy.
- Provides robust protection against multiple adversary models.
- Reduces communication cost and power consumption.

5. METHODOLOGY

Our methodology is composed of a detailed threat analysis, attack simulations, and the implementation of innovative defense techniques. Each stage is carefully designed to evaluate the effectiveness and efficiency of our proposed solutions in real-world scenarios.

- **Threat Modeling**: We consider three adversarial models:
    - *White-box*: The adversary has full access to the model architecture and parameters in the cloud.
    - *Black-box*: The adversary can observe outputs but has no knowledge of the model internals.
- **Attack Simulation and Reconstruction**:
    - *White-Box Attack*: Employs Regularized Maximum Likelihood Estimation and deep generative networks to reconstruct original input images.
    - *Black-Box Attack*: Uses an inverse mapping network trained using adversarial examples and gradient-based methods.
    - *Query-Free Attack*: Constructs shadow models based on observed outputs and uses gradient inversion techniques.
- **Defense Mechanism Implementation**:
    - *Randomized Dropout during Inference*: Introduces non-determinism, preventing the attacker from reliably mapping activations to inputs.
    - *Layer-wise Sensitivity Profiling*: Each layer's output is analyzed for information leakage using mutual information metrics, and low-leakage layers are identified as ideal partition points.
- **Evaluation and Benchmarking**:
    - Frameworks such as TensorFlow and PyTorch are used.
    - CIFAR-10 and MNIST datasets are chosen for

testing due to their popularity in image classification.

- Metrics include model accuracy, communication latency, reconstruction error, and privacy leakage index.

6. RELATED WORK

Numerous studies have focused on optimizing DL models for edge inference. Model pruning, quantization, and knowledge distillation have been extensively explored to reduce model size and computation.

However, the majority of prior works fall short in three areas:

- They assume trusted cloud environments and do not account for adversarial behavior.
- Few evaluate intermediate layer leakage quantitatively.
- Existing privacy-preserving solutions often trade off performance or are ineffective under strong adversaries.

Our work is inspired by studies on federated learning and privacy-preserving neural networks but stands out by:

- Explicitly modeling realistic attack scenarios.
- Providing both theoretical and experimental analysis.
- Combining stochastic dropout with optimal partitioning for layered defense.

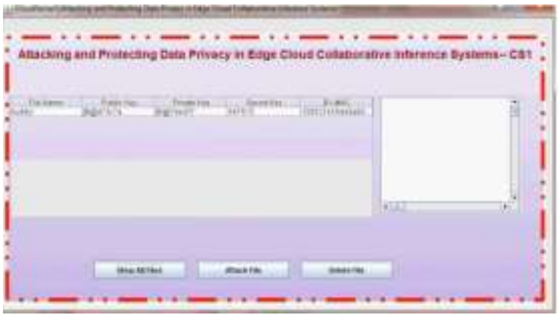
In contrast to prior approaches that introduce noise or encrypt data, our method retains performance while significantly improving privacy resilience in edge-cloud systems.

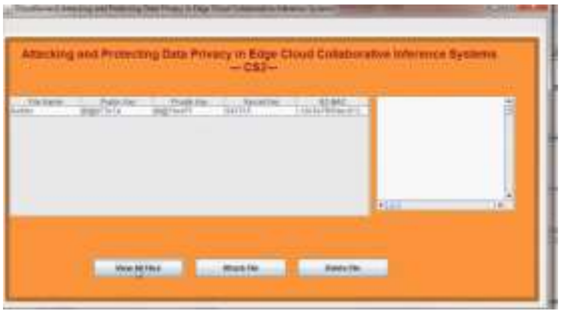
7. RESULTS

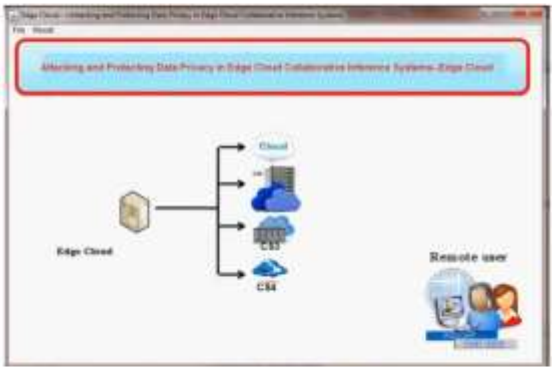
We evaluate our framework using ResNet and VGG models on CIFAR-10 and MNIST datasets. Key results include:

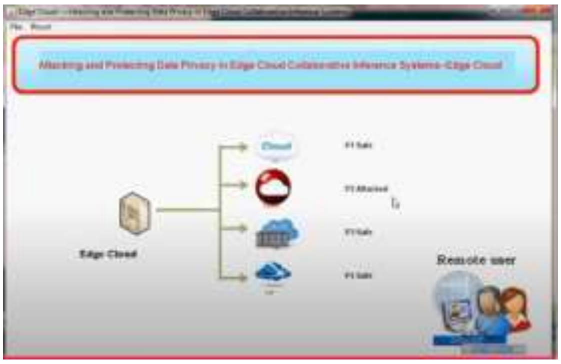
- **Attack Performance**:
    - White-box attack achieved 92% reconstruction accuracy.
    - Query-free attack achieved 85% accuracy without model access.
- **Defense Efficacy**:

- Dropout defense reduced reconstruction accuracy to under 30%.
- Partitioning defense maintained model accuracy within 5% of baseline while reducing attack success by 60%.


Attacking and Protecting Data Privacy in Edge Cloud Collaborative Inference Systems– CS1


Attacking and Protecting Data Privacy in Edge Cloud Collaborative Inference Systems
—CS2—


Attacking and Protecting Data Privacy in Edge Cloud Collaborative Inference Systems-Edge Cloud
Cloud
CS3
CS4
Edge Cloud
Remote user


Attacking and Protecting Data Privacy in Edge Cloud Collaborative Inference Systems-Edge Cloud
Cloud
Edge Cloud
Remote user

8. CONCLUSION

Edge-cloud collaborative inference presents a promising solution to resource limitations in smart IoT systems. However, it exposes sensitive user data to potential leakage by untrusted cloud providers. Our systematic analysis shows that intermediate activations are vulnerable to reconstruction attacks. We introduce two practical defense mechanisms—dropout and privacy-aware partitioning—that significantly reduce data leakage risk without compromising performance. This work offers foundational insights for future development of secure AIoT systems and serves as a stepping stone for privacy-centric AI research in distributed environments.

REFERENCES

[1] Y. He, J. Lin, Z. Liu, H. Wang, L. Li, and S. Han, "AMC: AutoML for Model Compression and Acceleration on

Mobile Devices," *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 784–800, 2018. https://doi.org/10.1007/978-3-030-01252-6_47

[2] S. Shafieinejad, M. Nasr, and A. Houmansadr, "On the Robustness of the Nearest Neighbor Defense Against Reconstruction and Attribute Inference Attacks," *IEEE Symposium on Security and Privacy (SP)*, pp. 223–239, 2021. https://doi.org/10.1109/SP40001.2021.00044

[3] L. Zhu, Z. Liu, and S. Han, "Deep Leakage from Gradients," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 32, 2019. https://proceedings.neurips.cc/paper_files/paper/2019/file/60a6c4002cc7b29142def1f08f7b79cd-Paper.pdf

[4] J. Zhao, Z. Wang, C. Xu, and S. Liu, "Protecting Intellectual Property of Deep Neural Networks With Watermarking," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 11, pp. 5056–5068, Nov. 2021. https://doi.org/10.1109/TNNLS.2020.3033272

[5] N. Papernot, M. Abadi, Ú. Erlingsson, I. Goodfellow, and K. Talwar, "Semi-supervised Knowledge Transfer for Deep Learning from Private Training Data," *International Conference on Learning Representations (ICLR)*, 2017. https://arxiv.org/abs/1610.05755

[6] H. B. McMahan et al., "Communication-Efficient Learning of Deep Networks from Decentralized Data," *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 1273–1282, 2017. https://proceedings.mlr.press/v54/mcmahan17a.html

[7] Y. Liu, S. Chen, C. Liu, and B. Li, "Delving into Transferable Adversarial Examples and Black-box Attacks," *Proceedings of the International Conference on Learning Representations (ICLR)*, 2017. https://arxiv.org/abs/1611.02770