Collaborative Inference Systems in Edge-Cloud: Attacking and Preserving Data Privacy

Dr.P.Satish Reddy[1]|S.Varalakshmi[2]|CH.Sangeetha[3]|Ravula Keerthana[4]

1, 2 & 3 Associate Professor, CSE department, Kasireddy Narayanreddy College of Engineering And Research, Hyderabad, TS.

4 UG SCHOLAR, CSE department, Kasireddy Narayanreddy College of Engineering And Research, Hyderabad, TS.

ABSTRACT: IoT systems and gadgets are growing more multipurpose and intelligent thanks to the development of Deep Learning technology. They are anticipated to run various Deep Learning inference tasks with great efficiency and performance. The mismatch between large-scale Deep Neural Networks and the restricted computing power of edge devices poses a difficulty to this demand. In order to resolve this problem, edge-cloud collaborative systems are then presented, allowing arbitrary Deep Learning applications to run on resource-constrained IoT devices. Third-party cloud use, however, may raise privacy concerns for edge computing. In this work, we systematically investigate the privacy protection and attack opportunities of edge-cloud collaborative systems. We have two things to contribute: In order to recover arbitrary inputs supplied into the system, even if the attacker does not have access to the data or computations of the edge device or the authorization to query this system, we first develop a series of new techniques for an untrusted cloud. (2) After providing two more efficient defense strategies, we empirically show that solutions that introduce noise are unable to defeat our suggested attacks. This offers information and recommendations for creating collaborative systems and algorithms that protect privacy.

KEYWORDS: Edge-cloud, IoT, Deep learning, Data.

I.INTRODUCTION: Recent years have witnessed the rapid development of Deep Learning (DL) and Internet of Things (IOT) technologies. IOT devices become appealing targets for DL applications. They use various sensors (e.g., cameras, microphones, gyroscopes) to collect data and information from environmental contexts, run the DL applications to interpret sensory data, and make control decisions. The integration of AI and IOT leads to the era of Artificial Intelligence of Things (AIOT), which has significantly changed our daily life: small-scaled AIOT systems are introduced to build smart homes and increase the comfort and

quality of life; medium-scale AIOT systems are deployed in warehouses and factories for higher efficiency and automation; large-scale AIOT systems can contribute to the establishment of smart cities. Deploying deep learning inference applications on commodity edge devices has several challenges. On one hand, an IOT device and collect streaming information at a very high rate (e.g. vehicle detection [1], remote monitoring [2], scene analysis [3] and application trace analysis [4]). This requires the device to run the DL models and analyze the data at a high speed. On the other hand, state-of-the-art DL models are becoming more complicated with larger sizes, making it infeasible for resource-constrained IOT devices to satisfy the performance requirements: the limited computation resources of the device can cause significant latency; the limited storage capacity makes it hard to store a large DNN model; the limited battery capacity causes a critical energy consumption constraint. To overcome this challenge, one possible approach is to offload the entire DL model and inference computation to the cloud.

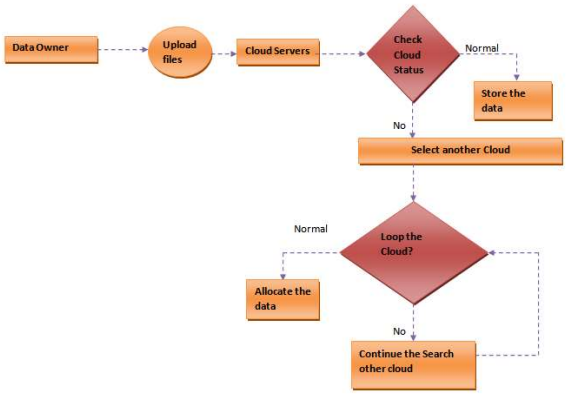
The edge device sends the input data to the cloud and receives the output. While this can resolve the aforementioned limitations of edge devices, it incurs significant communication costs when sending a large volume of raw data. Besides, there can be privacy breaches of the inference data [5], especially if the input data are highly sensitive like patients' records, and integrity breaches of the model [6], if the cloud is not trusted. An optimized strategy is to adopt collaborative inference between the edge devices and the cloud. The DL model can be divided into two parts. The first few layers of the network are stored in the local edge device, while the rest are offloaded to a remote cloud. Given an input, the edge device calculates the output of the first layers, sends it to the cloud, and retrieves the final results. This approach can reduce communication costs, as the intermediate output can be designed to be much smaller than the raw input. Such low data transfer bandwidth also achieves lower latency and smaller energy consumption. Collaborative inference makes it feasible and efficient to deploy large-scale intelligent workloads on today's edge platforms.

This paper presents an investigation of inference data privacy in edge-cloud collaborative systems, from the perspectives of attacks and defenses. Prior works all aimed to improve the performance and efficiency of such systems, while ignoring potential security issues. To the best of our

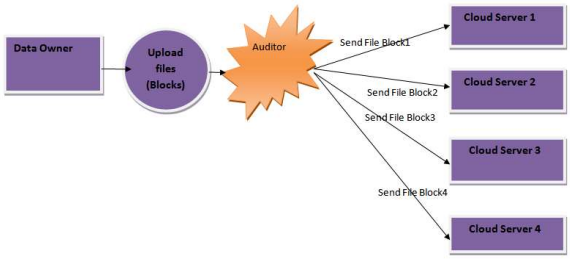
knowledge, we are the first to demonstrate the feasibility of input data privacy attacks against cloud-edge collaborative inference systems. The data privacy considered in this paper is the condidentiality of the raw inputs. Two key questions are considered in this study.

II.EXISTING SYSTEM: Training data privacy attacks. There are different types of privacy attacks against the training data. The first type is property inference attacks, which try to infer some properties of the training data from the model parameters. Attacks were demonstrated in traditional machine learning classifiers and fully-connected neural networks. A special case of property inference attacks are membership inference attacks, which infer whether one individual sample is included in the training set. This attack was first presented. The following work explored the feasibility of attacks with different adversary's capabilities model features in Generative Adversarial Networks and collaborative training systems. The second type of attacks against the training data's privacy are model inversion attacks given a machine learning model, and part of the training samples' features, the adversary can recover the rest of the features of the samples. Advanced model inversion attacks were designed to recover images from deep neural networks in single-party systems and collaborative learning systems. The third types are model encoding attacks the adversary with direct access to the training data can encode the sensitive data into the model for a receiver entity to retrieve. Model privacy attacks. The adversary attempts to steal the model parameters hyper parameters or structures via prediction APIs, memory side channels, etc. Inference data privacy attacks. Closer to our study is the work which trains an inverse network on the output probability distribution to get the inversed inference data. However, they only consider the model inversion attack from the soft max layer in the black-box scenario. We show that the attacker can successfully inverse the model from different layers, even in a stricter query-free scenario. We also provide defense strategies which are not discussed in their paper. Adopted a power side channel to recover inference data. However, this attack required the adversary to compromise the victim device for side-channel information collection, and it could only recover simple images (single pixel). Our work can recover any arbitrary complex data without access to, or knowledge of, the victim's device and computation.

III.PROPOSED SYSTEM: The proposed system designs a set of novel attack techniques to achieve this goal under different settings. First, for a white-box attacker, we propose using Regularized Maximum Likelihood Estimation to recover the samples from the model parameters and intermediate values. Second, for a black-box attacker, we propose the Inverse Network attack to identify the reverse mapping from the intermediate outputs to inputs without the knowledge of model information. Third, we consider the most limited adversarial capability where the cloud has no knowledge of the target model, and is not allowed to query the model. Conducting privacy attacks under this setting is extremely difficult, and this threat model is rarely considered in past work. For these query-free attacks, we introduce a new method of Shadow Model Reconstruction to achieve this attack. The second question we address in this paper is: how can the edge devices mitigate privacy leakage from the untrusted cloud? Past work adopted differential privacy to protect the inference data [12]. We show that this approach is impractical against our proposed attacks as it brings unacceptable performance degradation to the DL models. Instead, we propose two novel strategies that can better thwart the privacy attacks while still maintaining good model performance. The first one is the dropout defense: by deactivating random neurons during the inference, the adversary is not able to precisely generate the original images from the intermediate values. Our second defense is privacy-aware DNN partitioning: we comprehensively evaluate different factors that can affect the attack results, and propose some guidelines to partition the deep learning models for better privacy. We hope our findings can guide machine learning researchers and practitioners to design more secure collaborative inference systems.



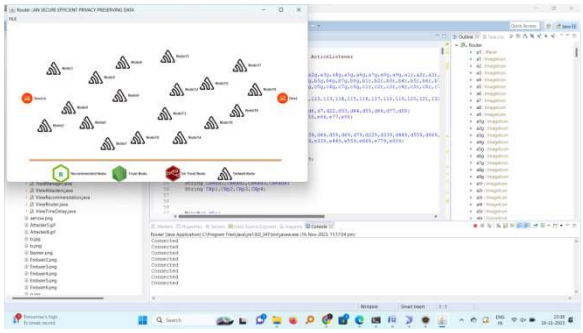
3.1 PROPOSED FLOW CHART



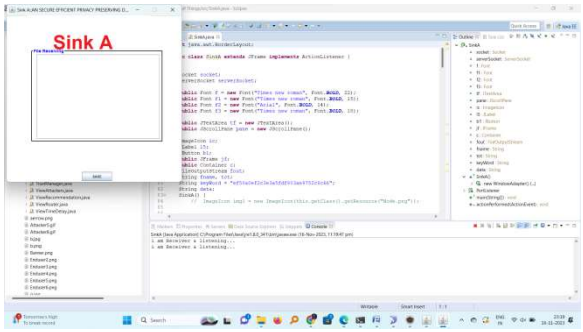
3.2 DATA FLOW DIAGRAM

IV.RESULTS:

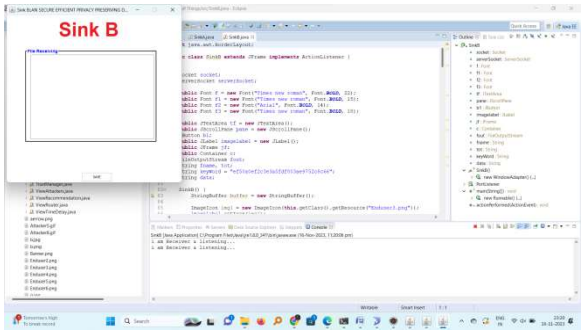
Router screen



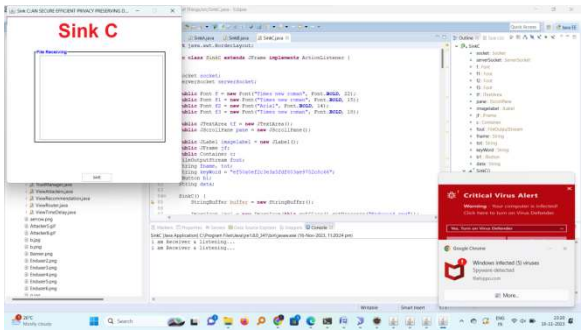
Sink a



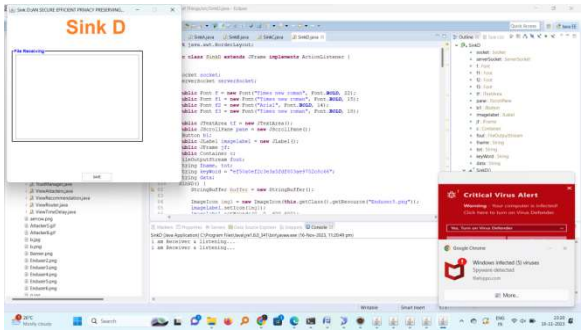
Sink b



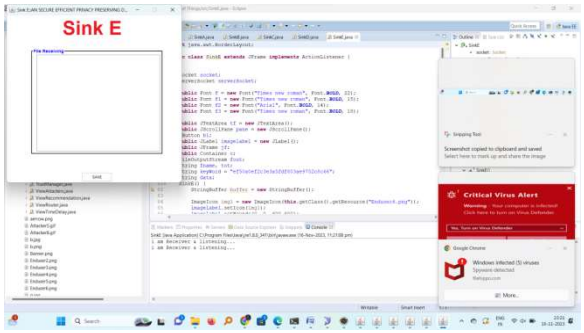
Sink c



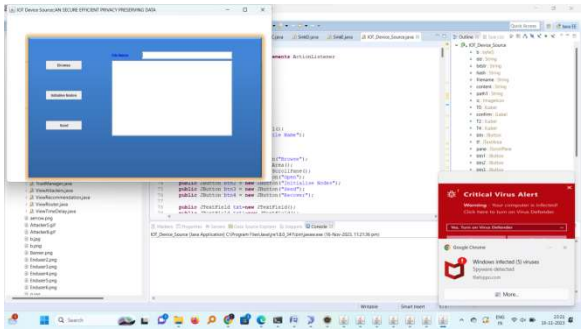
Sink d



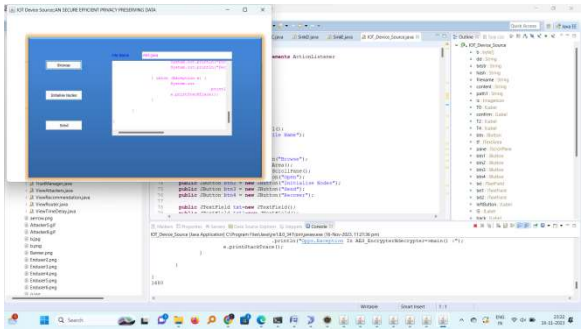
Sink e



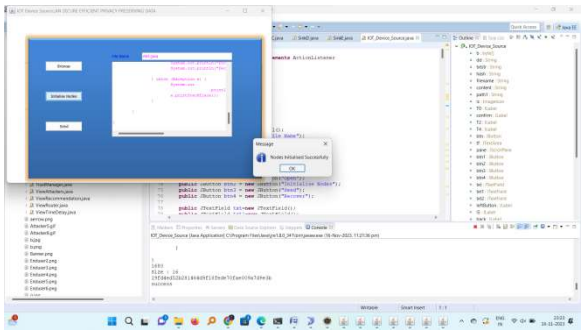
IOT Device



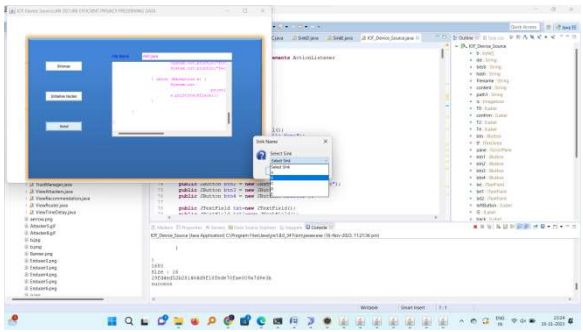
Choose file to send



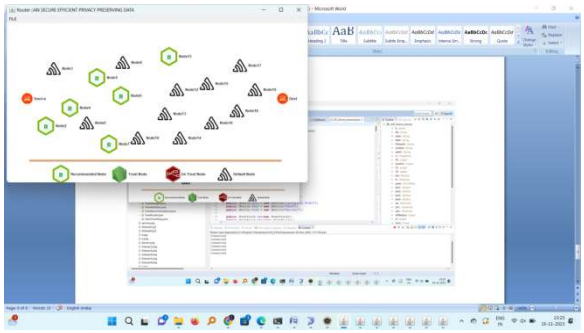
Nodes Initialised Successfully

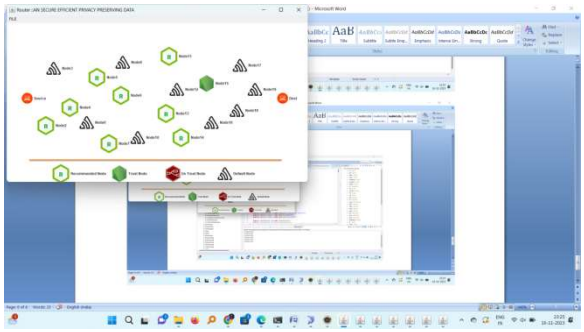


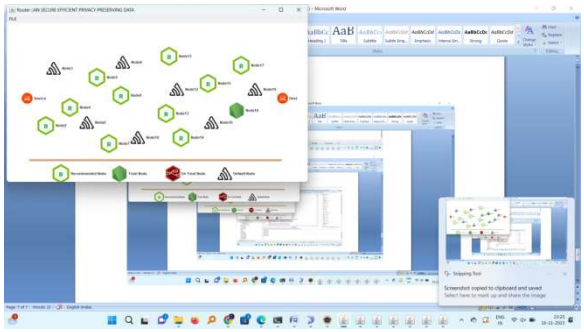
Select Sink

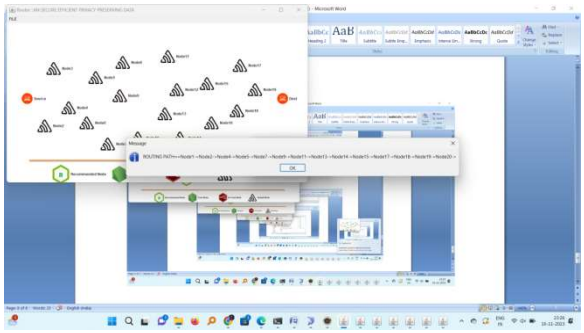


Data Sending

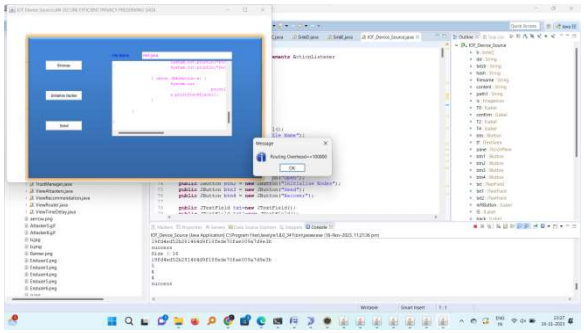




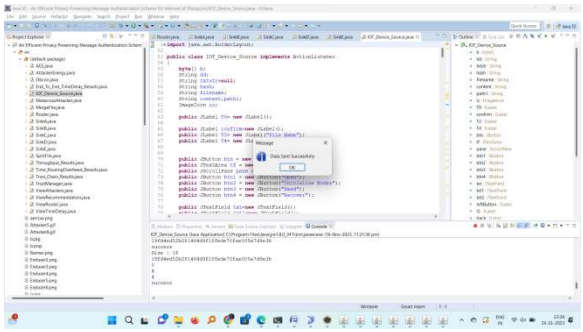




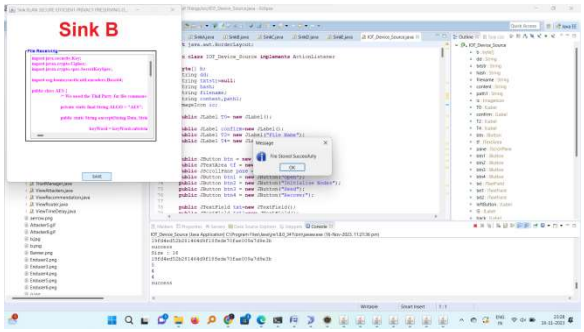
Data sent

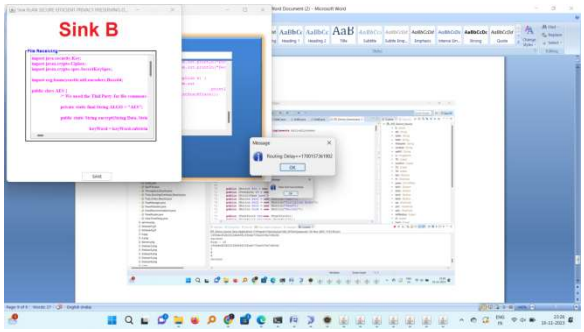


File Saved Successfully

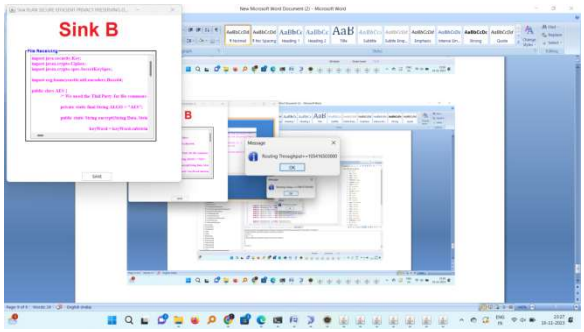


Routing Delay





Routing throughput



Routing Overhead

V.CONCLUSION: In this research, we investigate the privacy risks associated with inference data in edge-cloud collaborative systems. We find that the inference samples from intermediate values can be readily recovered from an untrusted cloud. We suggest a number of novel attack strategies to jeopardize the privacy of inference data in various attack scenarios. We show that with little requirements, the adversary may successfully and consistently retrieve the inputs. We also suggest a number of ways to safeguard the privacy of edge computing inference data. Prior research has ignored privacy in favor of concentrating on the functionality, performance, and efficiency of artificial intelligence of things. Our goal is to increase awareness of the significance of

protecting the privacy of inference data in edge-cloud systems and to promote the balancing of privacy protection and usability in the design or implementation of such systems.

REFERENCES: [1] Y. Tang, C. Zhang, R. Gu, P. Li, and B. Yang, "Vehicle detection and recognition for intelligent traffic surveillance system," Multimedia tools and applications, vol. 76, no. 4, pp. 5817–5832, 2017.

[2] G. Chen, T. X. Han, Z. He, R. Kays, and T. Forrester, "Deep convolutional neural network based species recognition for wild animal monitoring," in 2014 IEEE International Conference on Image Processing (ICIP). IEEE, 2014, pp. 858–862.

[3] C. Zhang, H. Li, X. Wang, and X. Yang, "Cross-scene crowd counting via deep convolutional neural networks," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 833– 841.

[4] L. Xiao, Y. Li, X. Huang, and X. Du, "Cloud-based malware detection game for mobile devices with offloading," IEEE Transactions on Mobile Computing, vol. 16, no. 10, pp. 2742–2750, 2017.

[5] F. Mireshghallah, M. Taram, P. Ramrakhyani, A. Jalali, D. Tullsen, and H. Esmaeilzadeh, "Shredder: Learning noise distributions to protect inference privacy," in Proceedings of the Twenty-Fifth International Conference on Architectural Support for Programming Languages and Operating Systems, 2020, pp. 3–18.

[6] Z. He, T. Zhang, and R. Lee, "Sensitive-sample fingerprinting of deep neural networks," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 4729–4737.

[7] J. Hauswald, T. Manville, Q. Zheng, R. Dreslinski, C. Chakrabarti, and T. Mudge, "A hybrid approach to offloading mobile image classification," in 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2014, pp. 8375–8379.

[8] Y. Kang, J. Hauswald, C. Gao, A. Rovinski, T. Mudge, J. Mars, and L. Tang, "Neurosurgeon: Collaborative intelligence between the cloud and mobile edge," Acm Sigplan Notices, vol. 52, no. 4, pp. 615–629, 2017.

[9] S. Teerapittayanon, B. McDanel, and H. Kung, "Distributed deep neural networks over the cloud, the edge and end devices," in IEEE International Conference on Distributed Computing Systems, 2017.

[10] J. H. Ko, T. Na, M. F. Amir, and S. Mukhopadhyay, "Edge-host partitioning of

deep neural networks with feature space encoding for resource-constrained internet-of-things platforms," in IEEE International Conference on Advanced Video and Signal Based Surveillance, 2018.