Poster Abstract: CNN-guardian: Secure Neural Network Inference Acceleration on Edge GPU

Qipeng Xie[1,2,4], Hao Yang[1,2], Linshan Jiang[5], Zhihe Zhao[3], Siyang Jiang[1,3], Shiyu Shen[1], Salabat Khan[4], Zhe Liu[1,2*], Kaishun Wu[4]
[1]Zhejiang Lab, HangZhou, China, [2]PQC Technologies Limited, China
[3]The Chinese University of Hong Kong, Hong Kong, China
[4]Hong Kong University of Science and Technology (Guangzhou), Guang Zhou, China
[5]National University of Singapore, Singapore

ABSTRACT

The rapid development of AI applications powered by deep learning in edge devices boosts the opportunity for real-time health monitoring. To address the potential privacy concern in the inference phase, homomorphic encryption (HE) is an alternative solution that encrypts inference data without exposing raw data and has several distinct advantages, (i.e., single-round communication, lightweight bandwidth consumption, and non-interactive computation). However, the computational overhead on the current HE-based privacy-preserving inference necessitates a substantial amount of time, which is not feasible for some real-time applications on edge devices. To address this issue, we propose CNN-guardian, a unified and compact neural network structure for real-time inference in HE-based inference on edge GPU. CNN-guardian designs a HE-friendly neural network and GPU engine that optimizes HE operations to accelerate the inference in the HE domain.

KEYWORDS

Secure CNN inference, GPU acceleration, Real-time System

ACM Reference Format:
Qipeng Xie[1,2,4], Hao Yang[1,2], Linshan Jiang[5], Zhihe Zhao[3], Siyang Jiang[1,3], Shiyu Shen[1], Salabat Khan[4], Zhe Liu[1,2*], Kaishun Wu[4]. 2023. Poster Abstract: CNN-guardian: Secure Neural Network Inference Acceleration on Edge GPU. In *The 21st ACM Conference on Embedded Networked Sensor Systems (SenSys '23), November 12–17, 2023, Istanbul, Turkey*. ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3625687.3628394

1 INTRODUCTION

Edge devices such as smartphones and sensors are becoming integrated parts of human life. To ensure privacy and minimize end-to-end latency, task inference, especially deep learning inference, is preferred to be made at edge devices [8]. Therefore, the need to run deep learning models on edge devices has increased significantly. The Convolutional Neural Network (CNN) has become the backbone of advanced ML, due to its inherent benefits, such as superior performance and scalability. Unfortunately, processing inference in plaintext through CNN has privacy concerns. Homomorphic Encryption (HE) offers an alternate way to improve privacy by processing encrypted data without exposing any raw information during inference on CNN. The HE-enabled Convolutional Neural Network (HCNN) [1–4] inherently enables direct computation over ciphertexts. This characteristic confers a multitude of advantages, including single-round communication, reduced bandwidth consumption, and non-interactive computation.

However, conventional HCNN always has extensive computation overhead on edge GPUs, inducing higher latency. For example, the low-latency framework [2] requires 730 seconds along with 12 GB RAM for a single prediction for an image classification application simulated on the CIFAR-10. Meanwhile, efforts focusing on server GPU-acceleration [1] indicate a time frame of 304.43 seconds for the same experiments. To address this issue, we propose CNN-guardian, a unified and compact network structure for real-time inference in HE-based neural networks on edge GPU [7] with no accuracy loss.

2 BACKGROUND AND CHALLENGE

Numerous researchers have been involved in the optimization and acceleration of HCNN by refining message encoding techniques, network architectures, or using hardware acceleration. Despite such efforts, the stringent requirements of real-time applications remain unfulfilled. First, the complexity and diversity of models used often necessitate bespoke implementations, thereby limiting their universal applicability. Moreover, the dependence of these models on specific datasets further impedes the development of a standardized and universally applicable network architecture. Second, the varied computations across different model layers present substantial challenges, especially when executing homomorphic evaluations over ciphertexts. The inherent nature of homomorphic encryption compounds these challenges by imposing high computational and memory burdens. This results in substantial resource allocation for even basic operations, thereby constraining the practical utility of such methodologies in scenarios with limited computational resources or tight time constraints. Consequently, despite the potential of HCNN in the realm of privacy-preserving neural networks, the existing constraints emphasize the pressing necessity for further enhancements and inventive approaches in this sector, especially in edge GPU [6].

These motivate us to CNN-guardian, a unified and compact network structure for real-time inference in convolutional neural

*Corresponding email: zhe.liu@zhejianglab.com.

SenSys '23, November 12–17, 2023, Istanbul, Turkiye

ACM ISBN 979-8-4007-0414-7/23/11…$15.00
https://doi.org/10.1145/3625687.3628394

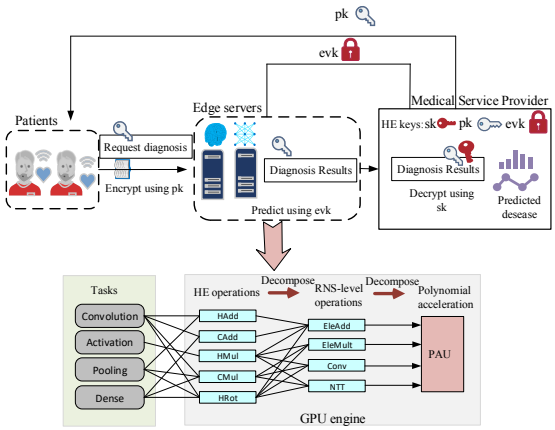

Figure 1: Overview of CNN-guardian

networks based on HE to profile and deal with hardware resource contention when secure CNN tasks are running on edge GPU.

3 SYSTEM DESIGN & GPU ENGINE

We use the remote medical scenario as an example. The CNN-guardian system encompasses four entities: the monitoring service provider, the end-users, and an edge service provider that is equipped with our custom developed GPU acceleration engine, as shown in Figure 1. In our scheme, the monitoring service provider generates HE keys at the protocol's onset: the secret key (sk) for decryption, the public key (pk) for encryption, and the evaluation keys (evk) for homomorphic evaluations. These keys are securely sent to the end-user and the edge service provider, respectively. End-users encrypt their data using the pk of HE to ensure privacy and data confidentiality, and these data are sent to the edge server. The edge server only has access to the evaluation key. It runs inference on the encrypted data utilizing our custom GPU engine, without the need for decryption, while maintaining the privacy of the input data. The encrypted result is then sent to the monitoring service provider for decryption and action, such as immediate intervention for detected heart disease or seizure.

For efficient computation of HCNN layers, we thoroughly analyze each layer's underlying HE evaluation operations and hierarchical structure. Our GPU engine relies on the Polynomial Acceleration Unit (PAU), which enhances the speed of polynomial arithmetic operations in the Residue Number System (RNS) representation by batching residue. Specifically, HAdd and CAdd are composed of EleAdd operations. More complex operations, such as HMult, CMult, and HRot, involve EleMult, EleAdd, Conv, and NTT operations. Using a hierarchical decomposition approach, we streamline the HCNN computation, making complex operations more efficient and manageable. Our GPU acceleration engine ensures rapid execution, boosting the overall performance of our HCNN implementation.

4 PRELIMINARY RESULTS

We evaluate our framework MNIST and a partial dataset in Medmnist [5]. We use an NVIDIA 4090 GPU and follow the same model architecture in [4]. Table 1 provides a performance comparison of our proposed method with other established techniques, evaluated on MNIST, PneumoniaMNIST, BreastMNIST, and Bloodmnist.

Method	Dataset	Accuracy(%)	Latency(s)	Platform
CryptoNets[3]	MNIST	98.95	205	CPU
Lola[2]	MNIST	98.95	2.2	CPU
CNN-guardian	MNIST	99.14	0.46	CPU
CNN-guardian	MNIST	99.14	0.032	GPU
CNN-guardian	Pneumonia MNIST	90.06	0.032	GPU
CNN-guardian	Breast MNIST	82.69	0.032	GPU
CNN-guardian	Blood MNIST	81.76	0.032	GPU

Table 1: Preliminary evaluations on accuracy and latency.

Across all datasets, our method not only achieves competitive accuracy, but also significantly reduces latency compared to other works. In the context of MNIST, our approach attains the highest accuracy of 99.14%, outperforming CryptoNets, and LoLa. More impressively, it reduces the execution time from 205 seconds in CryptoNets to 0.46 seconds on CPU and a mere 0.032 seconds on GPU. When it comes to the more Dataset, our solution still maintains superior performance in terms of latency, staying consistent at 0.46 seconds on CPU and 0.032 seconds on GPU, despite a slight drop in accuracy compared to benchmark [5]. This demonstrates that our method successfully provides a real-time solution for secure neural network inference across various datasets and platforms, breaking through the limitations of existing works. As for realistic evaluation on embedded GPUs such as Xavier, we will implement inference in the future.

5 CONCLUSION AND FUTURE WORK

In this poster, we propose CNN-guardian, a unified and compact network structure for real-time inference. We have achieved real-time requirements on some medical applications whilst preserving accuracy on a server-level GPU. As we move forward, we will further investigate the performance of our proposed CNN-guardian extending on some other r neural network architecture, i.e., transformer and BERT and evaluated the performance of CNN-guardian on edge GPU.

ACKNOWLEDGEMENT

The work was supported by the Zhejiang Lab Open Research Project (No.K2022PDOAB01), the Foundation for Distinguished Young Talents in Higher Education of Guangdong Province, China, (No. 2022KQNCX084), the National Key R&D Program of China (No.2020AAA0107703), the National Natural Science Foundation of China (No.62132008), and the Natural Science Foundation of Jiangsu Province, China (No.BK20220075).

REFERENCES

[1] Ahmad Al Badawi, Chao Jin, Jie Lin, Chan Fook Mun, Sim Jun Jie, B. Tan, Xiao Nan, Khin Mi Mi Aung, and Vijay Ramaseshan Chandrasekhar. 2021. Towards the AlexNet Moment for Homomorphic Encryption: HCNN, the First Homomorphic CNN on Encrypted Data With GPUs. *IEEE TETIC* 9 (2021), 1330–1343.

[2] Alon Brutzkus, Oren Elisha, and Ran Gilad-Bachrach. 2018. Low Latency Privacy Preserving Inference. *ArXiv* abs/1812.10659 (2018).

[3] Nathan Dowlin, Ran Gilad-Bachrach, Kim Laine, Kristin E. Lauter, Michael Naehrig, and John Robert Wernsing. 2016. CryptoNets: applying neural networks to encrypted data with high throughput and accuracy. In *TCML*. https://api.semanticscholar.org/CorpusID:217485587

[4] Hao Yang, Shiyu Shen, Siyang Jiang, Lu Zhou, Wangchen Dai, and Yunlei Zhao. 2023. XNET: A Real-Time Unified Secure Inference Framework Using Homomorphic Encryption. Cryptology ePrint Archive, Paper 2023/1428.
[5] Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. 2023. MedMNIST v2-A large-scale lightweight benchmark for 2D and 3D biomedical image classification. *Scientific Data* 10, 1 (2023), 41.
[6] Zhihe Zhao, Neiwen Ling, Nan Guan, and Guoliang Xing. 2022. Aaron: Compile-time Kernel Adaptation for Multi-DNN Inference Acceleration on Edge GPU. In *Sensys*. 802–803.
[7] Zhihe Zhao, Neiwen Ling, Nan Guan, and Guoliang Xing. 2023. Miriam: Exploiting Elastic Kernels for Real-time Multi-DNN Inference on Edge GPU. *arXiv preprint arXiv:2307.04339* (2023).
[8] Zhihe Zhao, Kai Wang, Neiwen Ling, and Guoliang Xing. 2021. Edgeml: An automl framework for real-time deep learning on the edge. In *IOTDI*. 133–144.