PPDL - Privacy Preserving Deep Learning Using Homomorphic Encryption

Nayna Jain
nayna.jain@iiitb.org
IIITB, Bangalore, India
IBM Systems, USA
USA

Karthik Nandakumar
karthik.nandakumar@mbzuai.ac.ae
Mohamed Bin Zayed University of
Artificial Intelligence
UAE

Nalini Ratha
nratha@buffalo.edu
University at Buffalo
USA

Sharath Pankanti
sharath.pankanti@gmail.com
Microsoft
USA

Uttam Kumar
uttam@iiitb.ac.in
IIITB, Bangalore
India

ABSTRACT

Deep Learning Models such as Convolution Neural Networks (CNNs) have shown great potential in various applications. However, these techniques will face regulatory compliance challenges related to privacy of user data, especially when they are deployed as a service on a cloud platform. Such concerns can be mitigated by using privacy preserving machine learning techniques. The purpose of our work is to explore a class of privacy preserving machine learning technique called Fully Homomorphic Encryption in enabling CNN inference on encrypted real-world dataset. Fully homomorphic encryption face the limitation of computational depth. They are also resource intensive operations. We run our experiments on MNIST dataset to understand the challenges and identify the optimization techniques. We used these insights to achieve the end goal of enabling encrypted inference for binary classification on melanoma dataset using Cheon-Kim-Kim-Song (CKKS) encryption scheme available in the open-source HElib library.

CCS CONCEPTS

• **Privacy-Preserving Machine Learning**; • **Homomorphic Encryption**; • **Optimization**;

KEYWORDS

Convolutional neural network, homomorphic encryption, optimization, non-linear activation function, ciphertext packing, multithreading

ACM Reference Format:
Nayna Jain, Karthik Nandakumar, Nalini Ratha, Sharath Pankanti, and Uttam Kumar. 2022. PPDL - Privacy Preserving Deep Learning Using Homomorphic Encryption. In *CODS-COMAD 2022 Jan 08-210, 2022, Bangalore, India.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3493700.3493760

1 INTRODUCTION

Machine learning as a service (MLaaS) is a popular paradigm that enables pre-trained resource-intensive deep neural networks to be deployed on the cloud. However, this exposes the end-user's data to the third-party service provider. For financial or healthcare clients, such exposure might mean infringing on the end-user's privacy and/or violation of privacy regulations such as Health Insurance Portability and Accountability Act (HIPAA) or General Data Protection Regulation (GDPR). Such concerns can undermine the advantages provided by MLaaS and impact their usability for real world applications. This has resulted in the growth of techniques that can enable privacy-preserving machine learning (PPML) [2–4, 9, 11] which enable performing secure inference once the model has already been trained in clear. They either use Fully homomorphic encryption or secure multi-party based techniques and optimize latency. Prior attempts to implement CKKS-based encrypted inference focus on either tweaking CNN architecture [7, 12] or optimizing low level operations (e.g., PT-CT multiplication) to maximize throughput [1]. One of the major bottlenecks in private inference is the implementation of non-linear activation functions such as sigmoid or ReLU, which are commonly used in deep neural networks. Hence, the idea of designing deep neural networks that operate on a limited ReLU budget has received attention recently [5, 8]. Methods to enable homomorphically encrypted inference using specialized hardware capability are discussed in [10]. The goal of our research is to design and implement a CKKS based encrypted inference framework for real world dataset. The inital work includes detailed analysis of different optimization techniques and their interplay with security parameters of the underlying FHE scheme using MNIST dataset. As part of the experiments, two scenarios are considered: 1.) Both data and model are encrypted in the Cloud. 2.) Only data is encrypted in the Cloud.

2 PROPOSED APPROACH AND DATASET

A typical CNN architecture consists of convolution, non-linear activation, pooling, and fully connected layers. Though fully homomorphic encryption (FHE) schemes allow arbitrary computations on the encrypted data, there are challenges because of the inherent inability of most FHE schemes to directly compute non-linear functions unless represented as polynomials. The accuracy of the operation in polynomial form improves with the degree of the polynomial but increases multiplicative depth of the network. Thus the

CODS-COMAD '22, Jan 08-10, 2022, Bangalore, India

ACM ISBN 978-1-4503-8582-4/22/01.
https://doi.org/10.1145/3493700.3493760

choice of the polynomial representation should be done optimally to manage accuracy vs feasibility of computational depth. The additional challenge is associated with the security parameters of the underlying FHE scheme. The computational time and ciphertext size are dependent on the security parameters of the scheme. As the security parameters are increased to support higher multiplicative depth and security strength, the associated computation time and ciphertext size increases non-linearly. Thus, to make the encrypted inference feasible on the deeper network, we optimize multiplicative depth of the network and reduce number of homomorphic operations. We have run MNIST experiments on encrypted data with both encrypted and unencrypted model and identified that encrypted model has far higher computational and memory requirements. Thus, our current experiments on larger network are focussed on encrypted data with unencrypted model. The native plaintext in the CKKS scheme is a polynomial in the cyclotomic ring and is mapped to a message vector of complex numbers via a complex canonical embedding map. This allows us to encrypt multiple plaintext messages into a single ciphertext. The number of messages that can be packed depends on the number of slots, which is determined by the selection of security parameters. This inherent packing ability of the CKKS scheme can be used to parallelize the computations in Single Instruction Multiple Data (SIMD) manner. We tried two different packing schemes single image vs batched image to compare the latency vs throughput tradeoff. We applied various optimization techniques to reduce number of homomorphic operations. We also exploited parallelism by intelligently applying nested multithreading to optimize the computation performance.

We ran our experiments on MNIST and melanoma dataset[6]. MNIST dataset consists of 60000 28×28 grayscale images of the 10 digits (0-9), along with a test set of 10000 images. The MNIST network consists of a single convolutional layer with 28 filters (each having a kernel size of 3×3) without any padding, followed by approximate ReLU, mean pooling, a flattening layer, a fully connected layer and a final softmax layer. The original source of the melanoma dataset is [6]. The training data of images was generated by augmenting random images from downloaded dataset using TensorFlow ImageDataGenerator. The original images were resized to 128×128 RGB images. The images belong to two classes, viz., benign or malignant. The network is a variant of well known LeNet model where activation layer is moved after pooling layer, standard ReLU is replaced with approximage ReLU, and only first fully connected layer has activation layer. The inference process does not require the computation of the softmax function, which is monotonic in nature.

3 RESULTS AND DISCUSSION

Our detailed experiments brings forth many new insights. While it is well-known that parallelization through multi-threading can reduce inference time, we demonstrate the need to design a layer-wise multi-threading strategy that is appropriate for the chosen network architecture. We observe that our batched inference has almost twice the number of homomorphic operations (HOPs) and thrice the number of CT-CT multiplications (the most expensive HOP) compared to [3], but the execution time is increased by only 25% due to a carefully designed multi-threading strategy. While model pruning can reduce inference time for batched inference, it has no impact on the single image (packed) inference. With our experiments on MNIST dataset, we estimated the execution time for both batched and single image packing schemes. With 16384 slots available, the amortized execution time for inference on a single image can be estimated as 34 milliseconds for encrypted model parameters. The amortized execution time for same experiment reduces to 2.9 msec when model parameters are not encrypted. With single image inference packing scheme we were able to achieve a total inference time of 8.8 seconds with encrypted model parameters compared to 2.5 seconds when model parameters are not encrypted. The test accuracy of this simple CNN model on plaintext images was found to be 97.84%, which is only marginally lower than that of a CNN where the approximate ReLU function is replaced with the standard ReLU function. For melanoma dataset, the results demonstrate that 80% classification accuracy can be achieved on encrypted skin lesion images with a latency of 51 seconds for single image inference and a throughput of 18,000 images per hour for batched inference, which shows that privacy-preserving machine learning as a service (MLaaS) based on encrypted data is indeed practically feasible.

4 CONCLUSIONS AND FUTURE WORK

The insights from our experiments are critical for making encrypted inference feasible on deeper networks. Our end goal is to support full inferencing on encrypted data in real world scenarios. We expect our work to become the foundation in enabling the application of deep neural networks for larger datasets, while maintaining provable privacy guarantees

REFERENCES

[1] Fabian Boemer, Anamaria Costache, Rosario Cammarota, and Casimir Wierzynski. 2019. nGraph-HE2: A High-Throughput Framework for Neural Network Inference on Encrypted Data. arXiv:1908.04172 [cs.CR]

[2] Alon Brutzkus, Ran Gilad-Bachrach, and Oren Elisha. 2019. Low latency privacy preserving inference. In *International Conference on Machine Learning*. 812–821.

[3] Edward Chou, Josh Beal, Daniel Levy, Serena Yeung, Albert Haque, and Li Fei-Fei. 2018. Faster CryptoNets: Leveraging Sparsity for Real-World Encrypted Inference. *CoRR* abs/1811.09953 (2018). arXiv:1811.09953 http://arxiv.org/abs/1811.09953

[4] Nathan Dowlin, Ran Gilad-Bachrach, Kim Laine, Kristin Lauter, Michael Naehrig, and John Wernsing. 2016. Cryptonets: Applying neural networks to encrypted data with high throughput and accuracy. In *International Conference on Machine Learning*. 201–210.

[5] Zahra Ghodsi, Akshaj Veldanda, Brandon Reagen, and Siddharth Garg. 2021. CryptoNAS: Private Inference on a ReLU Budget. arXiv:2006.08733 [cs.LG]

[6] David A. Gutman, Noel C. F. Codella, M. Emre Celebi, Brian Helba, Michael A. Marchetti, Nabin K. Mishra, and Allan Halpern. 2016. Skin Lesion Analysis toward Melanoma Detection: A Challenge at the International Symposium on Biomedical Imaging (ISBI) 2016, hosted by the International Skin Imaging Collaboration (ISIC). *CoRR* abs/1605.01397 (2016). arXiv:1605.01397 http://arxiv.org/abs/1605.01397

[7] Takumi Ishiyama, Takuya Suzuki, and Hayato Yamana. 2020. Highly Accurate CNN Inference Using Approximate Activation Functions over Homomorphic Encryption. arXiv:2009.03727 [cs.LG]

[8] Nandan Kumar Jha, Zahra Ghodsi, Siddharth Garg, and Brandon Reagen. 2021. DeepReDuce: ReLU Reduction for Fast Private Inference. arXiv:2103.01396 [cs.LG]

[9] Chiraag Juvekar, Vinod Vaikuntanathan, and Anantha Chandrakasan. 2018. GAZELLE: A low latency framework for secure neural network inference. In *27th USENIX Security Symposium (USENIX Security 18)*. 1651–1669.

[10] Guillermo Lloret-Talavera, Marc Jorda, Harald Servat, Fabian Boemer, Chetan Chauhan, Shigeki Tomishima, Nilesh N. Shah, and Antonio J Pena. 2021. Enabling Homomorphically Encrypted Inference for Large DNN Models. *IEEE Trans. Comput.* (2021), 1–1. https://doi.org/10.1109/tc.2021.3076123

[11] Qian Lou and Lei Jiang. 2019. SHE: A Fast and Accurate Deep Neural Network for Encrypted Data. In *Advances in Neural Information Processing Systems*. 10035–10043.

[12] Qian Lou and Lei Jiang. 2021. HEMET: A Homomorphic-Encryption-Friendly Privacy-Preserving Mobile Neural Network Architecture. arXiv:2106.00038 [cs.CR]