Poster: From Pre-trained to Private: A Compatible and Efficient Framework for Leveled Homomorphic Inference

Qingsong Tan and Li Shan Cang
School of Computer Engineering
GUniversity of Electronic Science and Technology of China
Chengdu, China
Email: 5200002@uestc.edu.cn

***Abstract*—To address the challenges of low computational efficiency and operator incompatibility in homomorphic encryption-based inference for confidential deep learning, this poster presents Seamless Homomorphic Inference (SHeIn) – a framework that bridges the gap between pre-trained models and private inference services. Our approach introduces three optimizations: (1) an enhanced diagonal method for homomorphic convolution, (2) depth-aware polynomial approximations for non-linear activations, and (3) a resource-aware adaptive strategy for fully-connected layers. Crucially, SHeIn enables large-image inference that exceeds available ciphertext slots – a fundamental bottleneck in leveled HE. Implemented under the RNS-CKKS scheme, we demonstrate significant reductions in latency and memory consumption without compromising model accuracy, enabling practical, scalable private inference.**

1. Introduction

The rapid integration of deep learning (DL) into privacy-sensitive sectors, such as medical diagnostics and financial risk modeling, has accelerated the demand for Cloud-based Inference-as-a-Service (IaaS). However, the standard centralized paradigm necessitates that users upload raw, unencrypted data to the cloud, triggering significant data confidentiality risks and failing to meet stringent regulatory requirements like CCPA/CPRA. *Confidential Deep Learning* has emerged as a critical frontier to reconcile the utility of cloud-scale resources with end-to-end data protection.

While several privacy-preserving techniques exist, they often impose prohibitive trade-offs: (1) *Differential Privacy (DP):* While protecting individual records, DP introduces statistical noise that frequently results in substantial accuracy degradation, particularly in high-dimensional models. (2) *Secure Multi-Party Computation (MPC):* MPC provides high accuracy but is plagued by massive communication overhead and high round-trip latency, making it impractical for resource-constrained clients.

RNS-CKKS enables non-interactive, single-server computation on encrypted data, bypassing the communication bottlenecks of MPC. However, three critical hurdles hinder practical confidential inference: *(i)* massive memory overhead from ciphertext expansion; *(ii)* high latency in ciphertext rotations (often $> 10^3\times$ slower than plaintext); and *(iii)* the fundamental incompatibility of non-linear operators with HE's polynomial primitives.

To bridge this gap, we present SheIn for efficient and scalable confidential inference. Our approach introduces an enhanced diagonal-method convolution to minimize rotations, depth-aware polynomial approximations for non-linear activations, and resource-aware adaptive scaling for fully-connected layers. These optimizations enable the seamless transformation of pre-trained models into HE-compatible services capable of high-resolution image processing. By streamlining the arithmetic-to-HE transition, our SHeIn significantly reduces computational overhead and memory footprint while preserving model accuracy, paving the way for practical, large-scale confidential deep learning.

2. Technical Methodology

We propose an automated conversion framework, SHeIn, that maps standard DL layers, including Conv, BN, FC, etc., into optimized HE-compatible pathways. By consuming a standard `nn.Module` and `.pth` weights, SHeIn outputs a deployment-ready confidential inference service. Our core technical contributions focus on minimizing *ciphertext rotations*—the primary bottleneck in RNS-CKKS which accounts for the majority of latency and memory consumption (due to Rotation Key size).

Contribution 1: Kernel-Centric Optimized Diagonal Method Traditional diagonal methods for homomorphic convolution [1] generate a plaintext mask matrix of size $\mathbb{R}^{CHW\times CHW}$. This results in $O(CHW)$ complexity for both mask generation and rotations, which is prohibitive for high-resolution inputs.

In this work, we implement an *offset-based mask generation algorithm* that shifts complexity to $O(C_{in}C_{out}K_HK_W)$, making it proportional to kernel dimensions rather than input resolution. By grouping spatial and channel offsets that yield identical global rotations, we drastically prune redundant masks, as shown in Fig. 1.

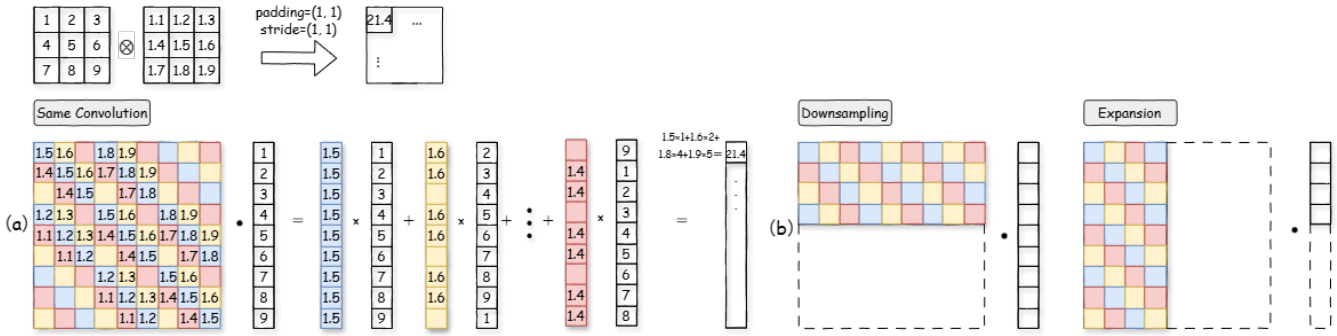


Figure 1. Diagonal method for same-padding convolution

(1) Efficiency Gain: For a 3×3 conv ($C_{in} = C_{out} = 32, H = W = 14$), our approach reduces diagonal masks from *6,272* to *567*.

(2) Post-BSGS Optimization: Even with Baby-step Giant-step (BSGS) [2] applied, our SHeIn maintains a superior rotation floor, significantly lowering memory overhead for rotation keys.

Contribution 2: Structure-Aware FC Selection Strategy Fully-connected (FC) layers often exhibit extreme asymmetry ($F_{in} \gg F_{out}$). A "one-size-fits-all" algorithm is suboptimal. We propose a *Heuristic Selection Strategy* that dynamically switches between the Diagonal Method and the Vector Inner-Product approach based on the rotation-summation cost.

Scenario A ($F_{out} \ll \sqrt{F_{in}}$): For classification layers (e.g., CIFAR-10), SHeIn selects the Inner-Product method, requiring only 120 rotations vs. 126 for the diagonal method.

Scenario B (Hidden Layers): When $F_{out} = 128$ and $F_{in} = 4096$, the strategy switches to the Diagonal Method, achieving a $10\times$ performance boost over the naive vector approach.

Contribution 3: Scalable Large-Image Inference To handle inputs where $CHW > n_{slots}$, we introduce a hybrid partitioning strategy. Instead of complex boundary-handling [3], we partition along the channel dimension. To mitigate the resulting memory expansion, we integrate *Depthwise Separable Convolutions* into the conversion pipeline.

By applying depthwise partitioning only to high-resolution entry layers and fine-tuning the model to recover accuracy, we maintain a manageable modulus-chain depth while enabling inference on images previously considered too large for leveled HE schemes.

3. Evaluation

We evaluate *SHeIn* using the OrganAMNIST dataset and a ResNet-10 (Half-channel) architecture. The framework is implemented via PyTorch and Pyfhel, utilizing the RNS-CKKS scheme at a security level of $\lambda \geq 128$ bits.

Accuracy and Fidelity SHeIn maintains high classification fidelity, achieving *87.89% accuracy*—a marginal *1.73% drop* from the 89.62% plaintext baseline. Numerical stability is robust, with a *Mean Absolute Error (MAE) of* 8.1×10^{-2}. These results confirm that our polynomial approximations and noise management effectively preserve decision boundaries.

Computational Efficiency The kernel-centric diagonal method targets the primary bottleneck of leveled HE: ciphertext rotations. Total rotations were reduced from *1,025 to 933*, yielding an *8.97% improvement* in efficiency. While the gain is most pronounced in high-dimensional layers like *BasicBlock1*, SHeIn further ensures feasibility for large-image scenarios through channel-partitioning and depthwise-separable convolutions, enabling inference even when input dimensions exceed available ciphertext slots.

TABLE 1. INFERENCE EFFICIENCY AND ACCURACY

Metric	Baseline (Standard)	SHeIn	Improvement
Total Rotations	1,025	933	8.97%
Model Accuracy	89.62%	87.89%	-1.73%

While rotation reductions are most pronounced in high-dimensional layers, SHeIn uniquely ensures large-image feasibility by employing channel-partitioning and depthwise-separable convolutions to process inputs exceeding ciphertext slots ($CHW > n_{slots}$) with manageable modulus depth. These capabilities – $O(K^2)$ rotations, over-slot inference, and automated conversion – are absent in Gazelle, Orion, and general HE frameworks.

4. Conclusion

In summary, through a series of optimization strategies, our SHeIn achieves adaptability across a broader range of variable conditions while enabling efficient encrypted inference without compromising model accuracy.

References

[1] C. Juvekar, V. Vaikuntanathan, and A. Chandrakasan, "Gazelle: A low latency framework for secure neural network inference," in *27th USENIX Security Symposium (USENIX Security 18)*, pp. 1651–1669, 2018.

[2] H. Cohen, *A course in computational algebraic number theory*, vol. 138. Springer Science Business Media, 2013.

[3] A. Ebel, K. Garimella, and B. Reagen, "Orion: A fully homomorphic encryption framework for deep learning," in *Proceedings of the 28th ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS '23)*, vol. 2, pp. 1–16, 2023.

Poster: From Pre-trained to Private : A Compatible and Efficient Framework for Leveled Homomorphic Inference

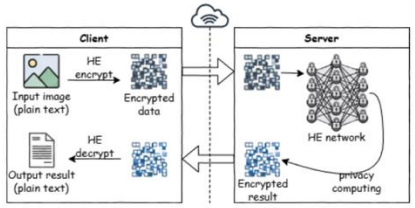
Yu Xie and Li Shan Cang
University of Electronic Science and Technology of China

Introduction

Background & Motivation

- **HE inference** enables privacy-preserving computation without the need for multi-party security protocols or frequent communication.
- **Barriers of Homomorphic Encryption:**
  - **Compatibility**: HE only supports addition and multiplication.
  - **Performance Gap**: Naive HE implementations suffer from massive computational overhead and memory bottlenecks.

HE Inference Workflow



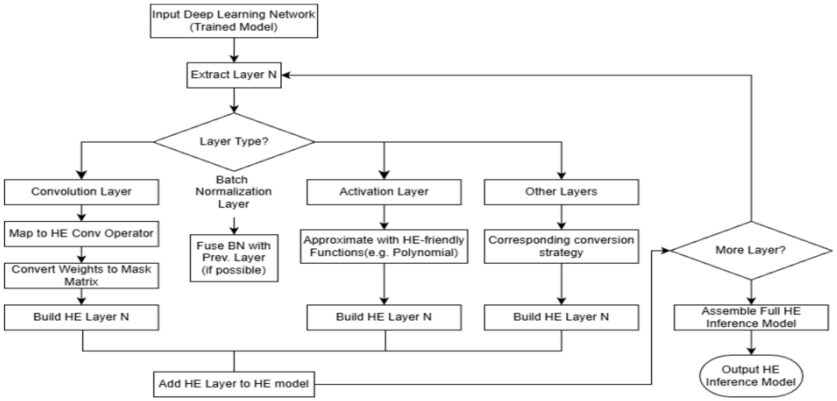
The data remains encrypted except on the local device.

Our Contributions

- **SHeIN Framework**: A Holistic Conversion Framework capable of translating standard Convolutional Neural Networks (CNNs) into Homomorphic Encryption-compatible inference models.
- We propose an optimized mask generation algorithm based on the diagonal method.
- **Large-Scale Inference:** An innovative large-scale graph inference scheme integrated with depthwise separable convolutions is introduced.

Evaluation Scenarios

Given only the network architecture (nn.Module) and the parameter set (.pth), the framework returns a model ready for homomorphic encrypted inference. The workflow of our framework is presented as follows.



As well as several resource-aware optimisation strategies outlined below:

Algorithm 1 Our Mask Pre-computation

Input:
Kernel $K \in \mathbb{R}^{C_{out} \times C_{in} \times k_H \times k_W}$,
input dimensions (C_{in}, H, W),
output dimensions $(C_{\text{out}}, \text{out}_H, \text{out}_W)$,
padding $(\text{pad}_h, \text{pad}_w)$,
slot count slot_count
Output: Mask dictionary $\mathcal{M}$

$\mathcal{M} \leftarrow \emptyset$
for all $(o_c, i_c, k_h, k_w) \in [C_{\text{out}}] \times [C_{\text{in}}] \times [k_H] \times [k_W]$ **do**
  if $K[o_c, i_c, k_h, k_w] \neq 0$ **then**
    $\text{shift} \leftarrow i_c \cdot H \cdot W - o_c \cdot \text{out}_H \cdot \text{out}_W + (k_h - \text{pad}_h) \cdot W + (k_w - \text{pad}_w)$
    if shift $\notin \mathcal{M}$ **then**
      $\mathcal{M}[\text{shift}] \leftarrow \mathbf{0} \in \mathbb{R}^{\text{slot\_count}}$
    end if
    $\Omega_y \leftarrow [\max(0, \text{pad}_h - k_h),\ \min(\text{out}_H, H + \text{pad}_h - k_h))$
    $\Omega_x \leftarrow [\max(0, \text{pad}_w - k_w),\ \min(\text{out}_W, W + \text{pad}_w - k_w))$
    for all $y \in \Omega_y$ **do**
      $I \leftarrow \{o_c \cdot \text{out}_H \cdot \text{out}_W + y \cdot \text{out}_W + x \mid x \in \Omega_x\}$
      $\mathcal{M}[\text{shift}][I] \leftarrow \mathcal{M}[\text{shift}][I] + K[o_c, i_c, k_h, k_w]$
    end for
  end if
end for
return $\mathcal{M}$

Flexible optimization strategy

- **Large-scale**: Adopting per-channel encryption and depthwise-separable convolution strategy unless

$$CHW < \frac{n_{slots}}{2} \text{ or } CHW = n_{slots}$$

- Polynomial Activation: The network struc- ture determines the maximum multiplicative depth allocable to activation function:

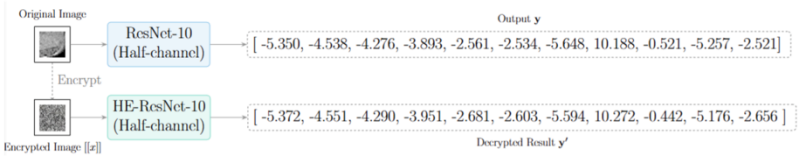
$$\text{ReLU}(x) \approx \frac{1}{2}(x + x \cdot P_d(x)), d = \lfloor \frac{L - (K - N_{act})}{N_{act}} \rfloor$$

- Reducing rotations in FC layers by comparing:

$$(F_{\text{out}} \cdot \lceil \log_2 F_{\text{in}} \rceil) \text{ vs } 2(\sqrt{F_{\text{in}}} - 1)$$

Experiment & Result

We pre-trained a ResNet-10 network on OrganAMNIST dataset (MedMNIST v2) and applied our conversion framework – with parameters set to $N = 2^{15}$ and a modulus chain configured for 128-bit security – to obtain the corresponding HE-ResNet-10 model.



SHeIn uniquely ensures large-image feasibility by employing channel-partitioning and depthwise-separable convolutions to process inputs exceeding ciphertext slots.

TABLE 1. INFERENCE EFFICIENCY AND ACCURACY

Metric	Baseline (Standard)	SHeIn	Improvement
Total Rotations	1,025	933	8.97%
Model Accuracy	89.62%	87.89%	-1.73%

Latency and memory measured on an Intel Xeon Gold 6242 @ 2.8GHz with 128GB RAM. SHeIn reduces latency by 34% and memory by 52%.

TABLE 2. DIRECT LATENCY AND MEMORY COMPARISON AGAINST PRIOR HE INFERENCE FRAMEWORKS.

Metric	Gazelle [1]	Orion [3]	SHeIn (Ours)
Inference latency (s)	12.4	8.7	**8.2**
Memory consumption (GB)	3.8	2.9	**1.8**
Total rotations (log2 scale)	1,025	968	**933**
Accuracy drop (vs. plaintext)	4.2%	3.1%	**1.73%**
Supports $CHW > n_{slots}$?	No	Partial	**Yes**

Conclusion

This work presents a framework to bridge the gap between pre-trained models and private, leveled homomorphic encryption-based inference. By integrating several optimization strategies, we achieve an efficient and compatible transition from standard deep learning models to the RNS-CKKS domain.