Proceedings of Machine Learning Research **318** (2026)
The 39th Canadian Conference on Artificial Intelligence
Canadian AI 2026
25–29 May 2026, Vancouver, BC

Pure Leveled CKKS for CNN Inference: The Finite Limb Depth Bound and ResNet-20 Stress-Test

Mohamed Khattab[†,*]
Lydia Bouzar-Benlabiod [†]
[†] CILS Lab, Jodrey School of Computer Science, Acadia University, Wolfville, NS, Canada.

Abstract

This paper presents a systems-level boundary analysis of pure Leveled CKKS homomorphic encryption applied to deep CNN inference, using two algorithmic co-designs as probing mechanisms: (Singular Value Decomposition) SVD-based kernel decomposition and scale-1 integer quantization. We formalize the *Finite Limb Depth Bound*, showing that scale-1 quantization delays but cannot eliminate rescaling, as the accumulated bit-width is bounded by the RNS prime limb width. A TinyConvNet smoke test confirms pipeline correctness (max noise 2.25×10^{-4}). Stress-testing ResNet-20 under a maximized 60-prime chain at N=65536 reveals that residual shortcut additions induce exact linear RNS level divergence $d_k = 3 + 7k$, exhausting the prime budget at the predicted shortcut index k^*=8 after 169,741 rotations and 555.7 s. Under the tested 95% SVD energy threshold, average rank 1.9 on 3×3 kernels exceeded the $K/2$ crossover, producing a 30.7% Ct-Pt overhead. Under the tested parameter regime, algorithmic co-designs alone were insufficient to eliminate bootstrapping; we outline a *bootstrap starvation* direction that targets reduced bootstrap frequency rather than full elimination.

Keywords: homomorphic encryption, CKKS inference, deep learning, residual networks, leveled HE

1. Introduction

Evaluating deep Convolutional Neural Networks (CNNs) over encrypted data using the Cheon-Kim-Kim-Song (CKKS) scheme allows for privacy-preserving inference, but requires strictly managing multiplicative depth limits and noise growth [1]. While existing frameworks like CryptoNets [2] and CHET [3] address shallow network evaluation and scheduling, the exact failure modes of deep residual architectures under pure Leveled HE remain undercharacterized.

The standard solution is bootstrapping: a computationally expensive circuit that homomorphically refreshes a ciphertext's noise budget, permitting smaller ring dimensions at the cost of substantial latency and approximation error. The alternative is pure Leveled HE, which pre-allocates a large noise budget to evaluate the target circuit without bootstrapping, incurring a proportional polynomial arithmetic penalty instead.

This paper presents a *systems stress-test*: we apply SVD-based kernel decomposition and scale-1 quantization to push a Leveled HE [4] configuration to its limits on ResNet-20 [5]. The paper does not propose a new HE scheme or a working deep FHE pipeline. Its purpose is to identify and formalize the precise failure modes that arise when deep residual inference is attempted without bootstrapping, and to characterize whether co-design can defer or eliminate those failures.

Our contributions are: (1) the *Finite Limb Depth Bound*, establishing that scale-1 quantization cannot extend multiplicative depth beyond the RNS limb bit-width; (2) a TinyConvNet smoke test confirming encrypted execution fidelity (3) an exact recurrence model $d_k = d_0 + k\ell$ for residual level divergence with closed-form exhaustion index $k^* = \lfloor (L - d_0)/\ell \rfloor$, confirmed across all eight shortcut additions; and (4) identification of two concrete system-level bottlenecks: Galois key precomputation policy constraints and residual-topology-accelerated prime depletion.

* 0313862k@acadiau.ca, lydia.bouzar-benlabiod@acadiau.ca

2. The Finite Limb Depth Bound

2.1. SVD-Based Spatial Compression

Learned convolutional kernels exhibit spatial low-rank structure. Applying SVD to each $K \times K$ kernel slice yields a rank-r approximation:

$$W_{c_o,c_i} \approx \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top \tag{2.1}$$

replacing 2D convolution with sequential 1D horizontal and vertical passes, and changing the per-kernel Ct-Pt count from K^2 to $2Kr$. A reduction occurs *only* when $r < K/2$: for 3×3 kernels, rank-1 reduces operations from 9 to 6 (33% saving), but rank-2 increases them from 9 to 12 (33% overhead). The crossover is at $r = K/2 = 1.5$.

Under the tested 95% energy retention threshold, average selected rank across all nine decomposed layers was 1.9, consistently exceeding this crossover, and producing a mean 30.7% Ct-Pt overhead. This confirms that energy threshold selection is a critical co-design parameter: for 3×3 kernels, consistently selecting rank-1 requires accepting $\approx$25% singular-value energy loss. SVD does not reduce multiplicative depth regardless of rank, since Ct-Pt multiplications consume the same RNS levels whether the kernel is decomposed or not.

2.2. Scale-1 Quantization and the Depth Bound

In CKKS, $Q = \prod_{i=0}^{L} q_i$ where L is the maximum multiplicative depth. A Ct-Pt multiply scales the ciphertext from Δ to Δ^2, requiring a `Rescale` that drops one RNS level. Standard power-of-two quantization maps weights to scale-1 integer plaintexts, so products remain at scale Δ, deferring rescaling, but accumulation is still bounded by the limb width.

The maximum number of sequential linear layers evaluable without `Rescale` is:

$$k_{\max} = \left\lfloor \frac{b - \log_2(\Delta M)}{\log_2 W} \right\rfloor \tag{2.2}$$

where b is the RNS limb bit-width, M is input magnitude, and W is maximum weight magnitude. This bound assumes no intermediate modulus switching and models accumulation within a single RNS limb prior to forced rescaling.

With b=40, Δ=2^{30}, $M\approx1$, and $W\approx2^{10}$ (worst-case weight magnitude after power-of-two quantization on CIFAR-10 [6]), $k_{\max}$=1. Chaining 19 layers would require 30+190=220 bits, exceeding any single RNS limb.

2.3. Channel-Adaptive Polynomial Activations

Batch Normalization induces per-channel affine domains $\Omega_c = [b_c - 3|a_c|,\ b_c + 3|a_c|]$. Chebyshev polynomial approximations fitted to these tight intervals can be terminated early on channels with narrow distributions, reducing Ct-Ct multiplications by $\approx$40% relative to a uniform degree-7 global approximation. This optimization serves as an additional probing mechanism within the stress-test.

3. Systems Stress-Test

3.1. Configuration and Results

TinyConvNet smoke test: 2 Conv2d–BN–polynomial activation blocks, 896 parameters, N=8192. Maximum noise 2.25×10^{-4}, confirming HE arithmetic pipeline correctness for shallow topologies.

ResNet-20 stress test: N=65536, 60-prime chain, degree-1 polynomial activations, SVD enabled on RunPod instance (RTX PRO 6000 Blackwell, 96 GB VRAM).

- Pipeline completed 18 of 21 layers before exhausting final RNS prime at `layer3.1`
- 169,741 rotations, 555.7 s wall-clock latency
- `Rescale` operations triggered CUDA NTT errors (`ntt.cu:2624`: invalid configuration)
- Error-tolerant backend resulted in silent corruption, continuing via no-op substitution rather than a hard crash.
- No classification accuracy reported (all runs produced corrupted or zero-substituted outputs from post-exhaustion error recovery)

3.2. **Bottleneck 1: Galois Key Precomputation Policy**

SVD separable convolutions require Galois keys for rotation offsets outside the standard precomputed set [3]. At N=65536, dynamic key generation would exhaust VRAM, constraining the implementation to a hardware-aligned rank-8 configuration, a memory allocation policy constraint, not a fundamental hardware limitation.

3.3. **Bottleneck 2: Residual Level Divergence**

With standard rescaling re-enabled, the pipeline completed 18 of 21 layers before exhausting the final RNS prime at `layer3.1`, incurring 169,741 rotations at a wall-clock latency of 555.7 s. From this point onward, `Rescale` operations triggered CUDA NTT errors (`ntt.cu:2624`: invalid configuration), and the error-tolerant backend resulted in silent corruption, continuing via no-op substitution rather than aborting.

Each ResNet-20 shortcut addition $x+f(x)$ requires level-matching the identity and residual branches, which arrive at different ciphertext levels due to asymmetric operation counts. Under a sequential CNN of equal depth, no such asymmetry exists; the accelerated depletion is therefore a direct consequence of residual topology. The measured level mismatches, d_0=3, d_1=10, d_2=17, d_3=24, d_4=31, d_5=38, d_6=45, d_7=52, satisfy $d_k = 3 + 7k$ exactly with zero residual error. Because every block adds the same asymmetry ℓ=7 to the level gap, mismatches accumulate as an arithmetic sequence, and chain exhaustion occurs at:

$$k^* = \left\lfloor \frac{L - d_0}{\ell} \right\rfloor \tag{3.1}$$

With L=60, d_0=3, ℓ=7: k^*=8, confirmed exactly (`ct_b` reaches level -2 at shortcut 8 and -9 at shortcut 9). Alignment cost scaled proportionally: 32 s per-merge at shortcut 0, 123 s at shortcut 7, a 4× increase consistent with d_k growth.

4. **Bootstrap Starvation: Future Directions**

The `layer3.1` failure confirms that pure Leveled HE is insufficient for deep residual CNNs under current parameter regimes; periodic bootstrapping is necessary. The co-design techniques examined here suggest a concrete path toward reducing bootstrap *frequency*, a direction we term *bootstrap starvation*.

Depth-consumption estimate. A degree-7 polynomial activation consumes 7 RNS levels, yielding a total depth of 20×(1+7)=160 levels for a 20-layer network. Degree-2 activations reduce this to 20×(1+2)=60 levels. Following Cheon et al. [7], a 15-level cost per bootstrap interval yields $\lceil 160/15 \rceil \approx 11$ operations for degree-7 and $\lceil 60/15 \rceil = 4$ for degree-2, a meaningful latency reduction even without eliminating bootstrapping.

Conjecture: multi-dimensional error coupling. We conjecture that the compounded error in a depth-L network satisfies:

$$E_{\text{total}} \lesssim \prod_{l=1}^{L} \|W_l\|_2 \left(E_{\text{svd}} + E_{\text{quant}} + E_{\text{poly}}\right) + E_{\text{he}} \tag{4.1}$$

where E_{svd}, E_{quant}, E_{poly} capture SVD truncation, power-of-two rounding, and low-degree Chebyshev oscillation respectively, and E_{he} subsumes HE and bootstrapping noise. A single shallow bootstrapped run (2–3 intervals on a 6-layer subnetwork) would provide the first empirical validation point.

Adaptive activation refinement. Per-channel BN intervals vary by an order of magnitude (e.g., ± 0.60 to ± 9.00 across `bn1` channels), confirming the theoretical motivation for channel-adaptive degree selection, but the tested σ-based binning produced complete cluster collapse (test vector E). Per-channel degree assignment, selecting the minimum Chebyshev degree satisfying a per-channel L^∞ bound, could recover $\approx$40% Ct-Ct savings on narrow channels, directly extending bootstrap intervals.

SVD threshold calibration. Consistent rank-1 selection requires lowering the energy threshold from 95% to $\approx$70–75%, incurring greater approximation error. Characterizing the accuracy–overhead Pareto frontier across thresholds on a bootstrapped pipeline would determine whether the 33% Ct-Pt saving at rank-1 justifies the associated truncation error.

Residual-aware bootstrap scheduling. The exact divergence recurrence makes level exhaustion predictable at compile time. A topology-aware scheduler could insert bootstraps before divergence compounds, exploiting the measured $4\times$ alignment cost growth to reduce both bootstrap count and per-bootstrap overhead.

5. Conclusion

The Finite Limb Depth Bound establishes that scale-1 quantization delays but cannot eliminate rescaling within fixed-width RNS representations. ResNet-20 stress-testing confirms exact level divergence under residual topology, and SVD at the tested energy threshold produced overhead rather than a saving. Under the tested regime, algorithmic co-designs cannot substitute for bootstrapping in deep residual architectures; future work should target bootstrap starvation via depth-efficient activations, SVD threshold calibration, and residual-aware scheduling.

While CryptoNets [2] and CHET [3] address shallow networks and scheduling overhead, neither targets the residual depth ceiling characterized here; this work situates the boundary where pure leveled HE fails for deep residual architectures.

References

[1] J. H. Cheon et al. "Homomorphic encryption for arithmetic of approximate numbers". In: *ASIACRYPT*. 2017.

[2] R. Gilad-Bachrach et al. "CryptoNets". In: *ICML*. 2016.

[3] R. Dathathri et al. "CHET: an optimizing compiler for fully-homomorphic neural-network inferencing". In: *PLDI*. 2019.

[4] Z. Brakerski et al. "(Leveled) fully homomorphic encryption without bootstrapping". In: *ACM TOCT* 6.3 (2014).

[5] K. He et al. "Deep residual learning for image recognition". In: *CVPR*. 2016.

[6] A. Krizhevsky and G. Hinton. *Learning multiple layers of features from tiny images*. Tech. rep. University of Toronto, 2009.

[7] J. H. Cheon et al. "Bootstrapping for Approximate Homomorphic Encryption". In: *EUROCRYPT*. 2018.