ULD-Net: Enabling Ultra-Low-Degree Fully Polynomial Networks for Homomorphically Encrypted Inference

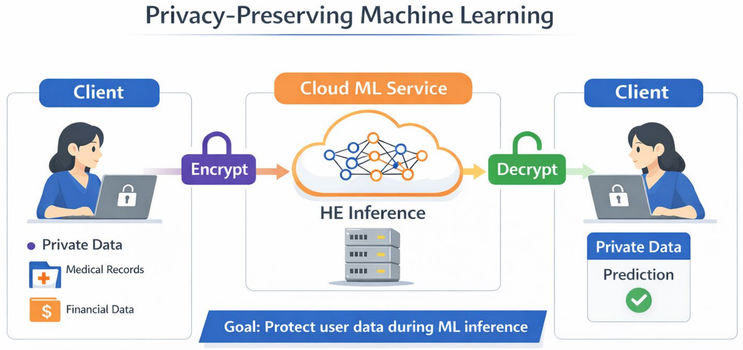
Xi Xie[1], Ran Ran[2], Jiahui Zhao[1], Bin Lei[3], Zhijie Jerry Shi[1], Wujie Wen[2], Caiwen Ding[3]
University of Connecticut[1], North Carolina State University[2], University of Minnesota[3]

ABSTRACT

ULD-Net: Enabling Ultra-Low-Degree Fully Polynomial Networks for Homomorphically Encrypted Inference

- We introduce ULD-Net, a training methodology for ultra-low-degree fully polynomial neural networks.
- Fully polynomial models are essential for homomorphic encryption (HE) inference but are difficult to train and scale.
- ULD-Net enables training from scratch with multiplicative depth ≤ 3 per operator while maintaining high accuracy.
- Our design includes a polynomial-only normalization (PolyNorm), polynomial activations, and linear attention.
- ULD-Net achieves 76.70% on ViT-Small and 75.20% on ViT-Base on ImageNet, matching original models.
- It is the first fully polynomial approach to scale to ViT/ImageNet level.
- Our method improves both accuracy and HE inference latency over prior polynomial networks.
- ULD-Net addresses key challenges in stable training and efficient private inference.
- Implementation available: https://github.com/xiexi51/ULD-Net

INTRODUCTION



Privacy-Preserving Machine Learning
Privacy-preserving machine learning enables computation on sensitive data without exposing raw inputs.
Key applications:
•Medical Data Analysis *(Dowlin et al., 2016)*
•Financial Applications *(Acar et al., 2018)*
•Privacy-Preserving ML-as-a-Service *(Boemer et al., 2019)*

Homomorphic Encryption (HE)
Homomorphic Encryption (HE) allows direct computation on encrypted data without decryption.
•Supports addition and multiplication on ciphertexts
•Widely used for secure ML deployment
•Highly sensitive to multiplicative depth and expensive operations.

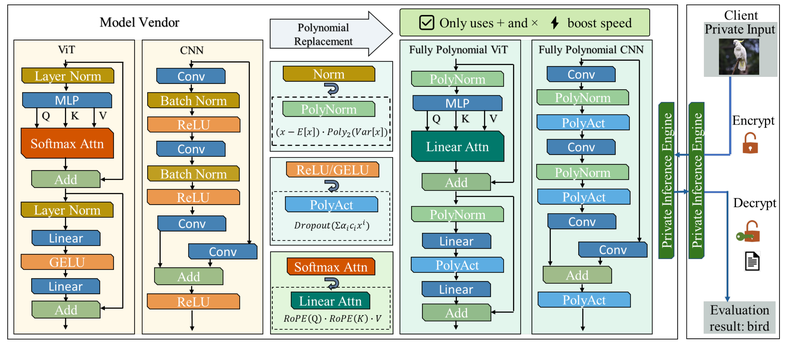
Limitation of Non-Polynomial Operators under HE
•ReLU, GELU, LayerNorm, Softmax
•Expensive or unsupported under HE
•Require costly approximations or protocol switching
•Increase multiplicative depth and latency

Polynomial Replacement for Efficient HE Inference
•Polynomial operators contain only additions and multiplications
•Fully compatible with HE

Fully Polynomial Models
Fully polynomial neural networks replace all non-polynomial operations with additions and multiplications, enabling seamless deployment under homomorphic encryption (HE).

Advantages
•Native support in HE schemes
•Extremely low computation cost
•Simple and efficient inference pipeline

Challenges
•Training stability
•Scalability to large models and datasets
•Maintaining competitive accuracy



MATERIALS AND METHODS

ULD-Net (Our work)

Normalization-Axis Principle

- Normalization provides critical numerical constraints.

$$\mathrm{Norm}[x] = \frac{x - \mathbb{E}[x]}{\sqrt{\mathrm{Var}[x] + \epsilon}}$$

- Sample-wise normalization provides better numerical stability.

Consider a model with n normalization–polynomial layer pairs, assuming normally distributed inputs.
If normalization statistics are shared across samples, the variance grows approximately as:

$$v_n' \sim r^{d^n}$$

resulting in exponential variance explosion.

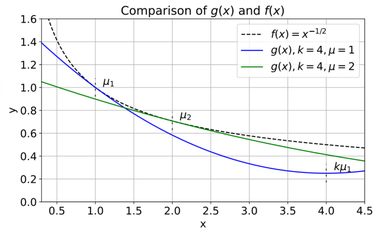
PolyNorm: Polynomial-Only Normalization Layer

- Quadratic approximation of $f(x) = \frac{1}{\sqrt{x}}$:

$$g(x) = -\frac{(x - k\mu)^2}{4(1-k)\mu^{5/2}} + \frac{5-k}{4\mu^{1/2}}, \quad k \in [2.438, 5)$$

- The expression of PolyNorm:

$$\mathrm{PolyNorm}[x] = (x - \mathbb{E}[x]) \cdot g(\mu v) \cdot \sqrt{\frac{\mu}{\mathrm{Var}}}$$



Overall Design Recipe

- Replace activation functions (ReLu, GELU) with PolyAct

$$\mathrm{PolyAct}(x) = \mathrm{Dropout}\left(\sum_{i=0}^{n} \alpha_i c_i x^i\right)$$

- Replace softmax with RoPE (2024) -based linear attention

$$\mathrm{LinearAttn}(x) = \mathrm{RoPE}(Q) \cdot \mathrm{RoPE}(K)^\top \cdot V$$

- Replace normalization layers with PolyNorm

$$\mathrm{PolyNorm}[x] = (x - \mathbb{E}[x]) \cdot g(\mu v) \cdot \sqrt{\frac{\mu}{\mathrm{Var}}}$$

- Apply variance penalty losses to improve the stability of PolyNorm

$$\mathcal{L}_1 = \frac{1}{N} \cdot \sum_{i=1}^{N} v_i \cdot \lambda_1, \quad \mathcal{L}_2 = \frac{1}{N} \cdot \sum_{i=1}^{N} (v_i - 1)^2 \cdot \lambda_2$$

RESULTS

- Comparison with SOTA fully polynomial model works
  on ResNet-18/ImageNet, original accuracy: 69.76%

ResNet-18 / ImageNet (Fully Polynomial)	Activation Degree	Test Acc.	Activation Latency (s)	Speedup	Model Latency (s)	Speedup
Lee et al. (2021)	6075	69.35%	16448	16.06×	144896	3.50×
SMART-PAF	81	69.40%	8311	8.12×	114277	2.76×
ULD-Net (Ours)	2	69.79%	1024	–	41408	–

- Comparison with NEXUS on ViT-Small

Dataset Original Acc.	Method	Test Acc.	Non-Polynomial Operator Latency (s)				Speedup
			Softmax	LayerNorm	GELU	Total	
CIFAR-10 91.77%	NEXUS	91.39%	3055	2080	2860	7995	20.5×
	ULD-Net (Ours)	91.48%	156	156	78	390	–
Tiny-ImageNet 60.90%*	NEXUS	60.52%	9259	6304	8668	24231	20.5×
	ULD-Net (Ours)	61.40%	472	474	236	1182	–

- Comparison with SOTA partial polynomial replacement methods (on ResNet-18/ CIFAR-100, original accuracy: 77.84%)

Method	ReLU Replace Ratio	Test Acc.	Activation Latency (s)	Model Latency (s)
SNL	0.88	73.75%	45	2052
AutoReP	0.87	75.48%	46	2053
AutoReP	0.93	74.92%	35	2042
ULD-Net (Ours)	1	78.81%	16	647

- Extended experiments of ULD-Net with the VanillaNet family on ImageNet

Model Original Acc.	Activation Degree	Test Acc.	Activation Latency (s)	Model Latency (s)
VanillaNet-5 72.49%	2	72.43%	478	3469
VanillaNet-6 76.36%	2	74.25%	597	4337
	3	76.03%	939	4678
VanillaNet-7 77.98%	2	74.91%	717	5204
	3	76.40%	1126	5614

- ULD-Net achieves 76.7% accuracy on ViT-Small/ImageNet (vs. 76.5% original) and 75.2% on ViT-Base/ImageNet (vs. 75.3% original), representing the first low-degree fully polynomial ViT models trained at ImageNet scale.

CONCLUSIONS

- We propose ULD-Net, a training methodology for ultra-low-degree fully polynomial neural networks
- Enables end-to-end polynomial-only computation for HE-friendly inference
- Achieves stable training at scale (ImageNet and ViT architectures)
- Maintains competitive accuracy with multiplicative depth ≤ 3 per operator
- Demonstrates significant reduction in HE inference latency
- Establishes the first fully polynomial models at ViT/ImageNet scale
- Provides a general and scalable design paradigm for privacy-preserving deep learning