PAF-FHE: Low-Cost Accurate Non-Polynomial Operator Polynomial Approximation in Fully Homomorphic Encryption Based ML Inference

Jingtian Dang*, Jianming Tong*, Anupam Golder, Arijit Raychowdhury, Cong Hao, Tushar Krishna

dangjingtian@cmu.edu, jianming.tong@gatech.edu, tushar@ece.gatech.edu

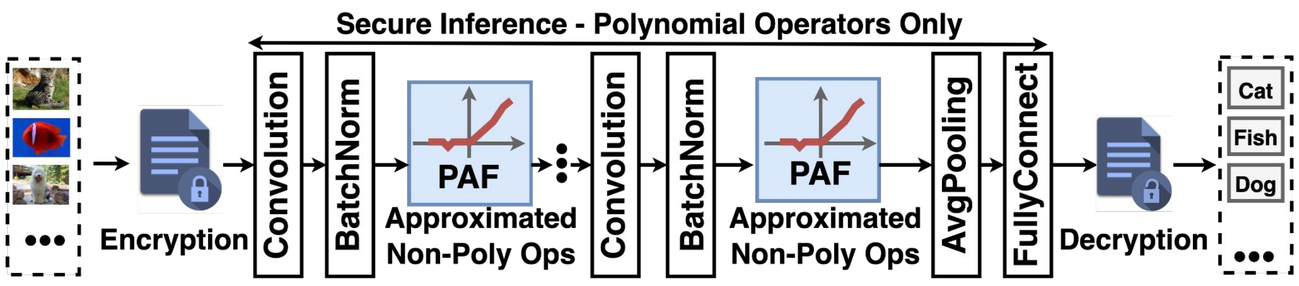
Georgia Institute of Technology

Code Paper

Motivation

1.Fully Homomorphic Encryption only supports polynomial operators.

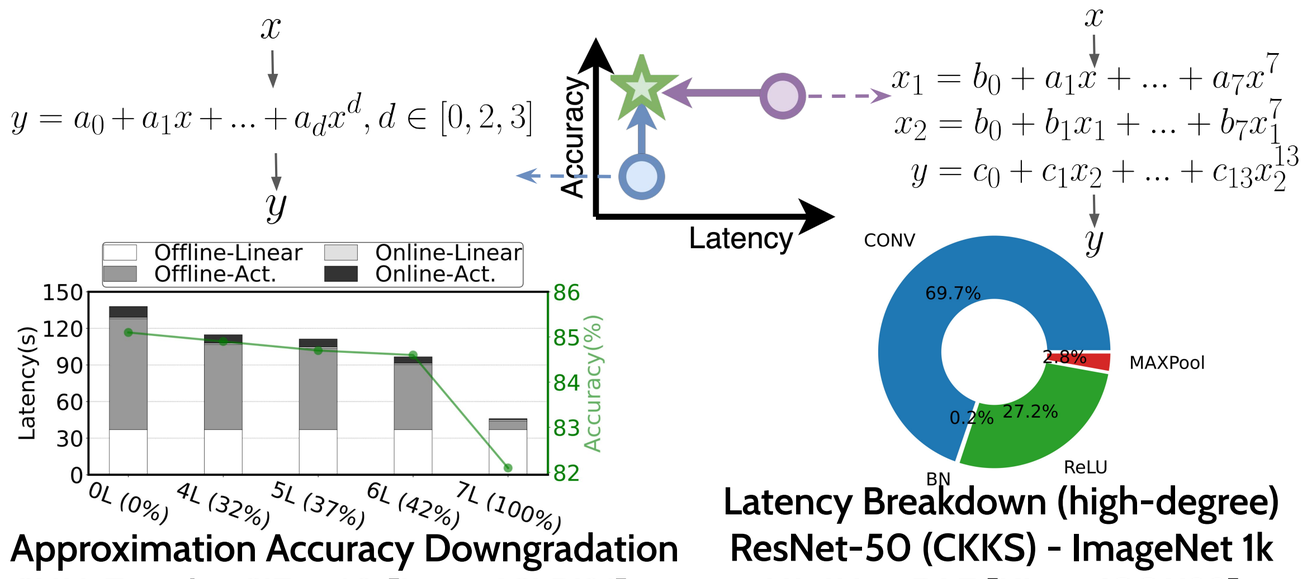
2.non-poly. ReLU/MaxPooling→Polynomial Approximation Function (PAF)



3.**prohibitive-overhead** or **low accuracy** in current non-polynomial operators. (accuracy, ratio of overall inference latency in ResNet-32)

a.Minimax[ISCA'22]: 27-degree poly. – 27.2% ResNet-50 inference latency.

b.SafeNet[ICLR'21]: <3-degree poly – 4% accuracy degrade (CNN-7; CiFar10)



Approximation Accuracy Downgradation CNN-7 under CiFar-10 [Lou+, ICLR'21]

Latency Breakdown (high-degree) ResNet-50 (CKKS) - ImageNet 1k MiniMax PAF [Kim+, ISCA'22]

Challenges

	Low Communication Overhead	Low Accuracy Degradation	Low Latency Overhead
SafeNet, CryptoGCN	✗	✗	✓
CryptoNet, CryptoDL, LoLa, CHE	✗	✗	✓
F1, CraterLake, BTS	✓	✓	✗
HEAX, Delphi, Gazelle, Cheetah	✗	✗	✓
SHE	✓	✓	✗
PAF-FHE	✓	✓	✓

The full degree space of PAF has NOT been explored!

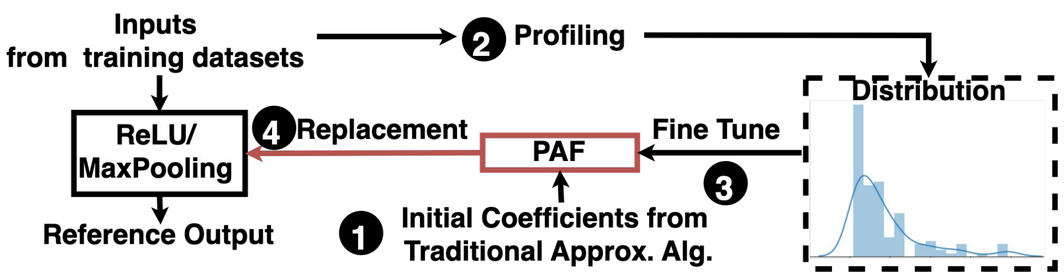
1.Deterministic coefficients determination introduces accuracy degradation for PAF with less than 27-degree.

2.ML-based coefficients determination hardly converge for PAF with higher than 5 degree.

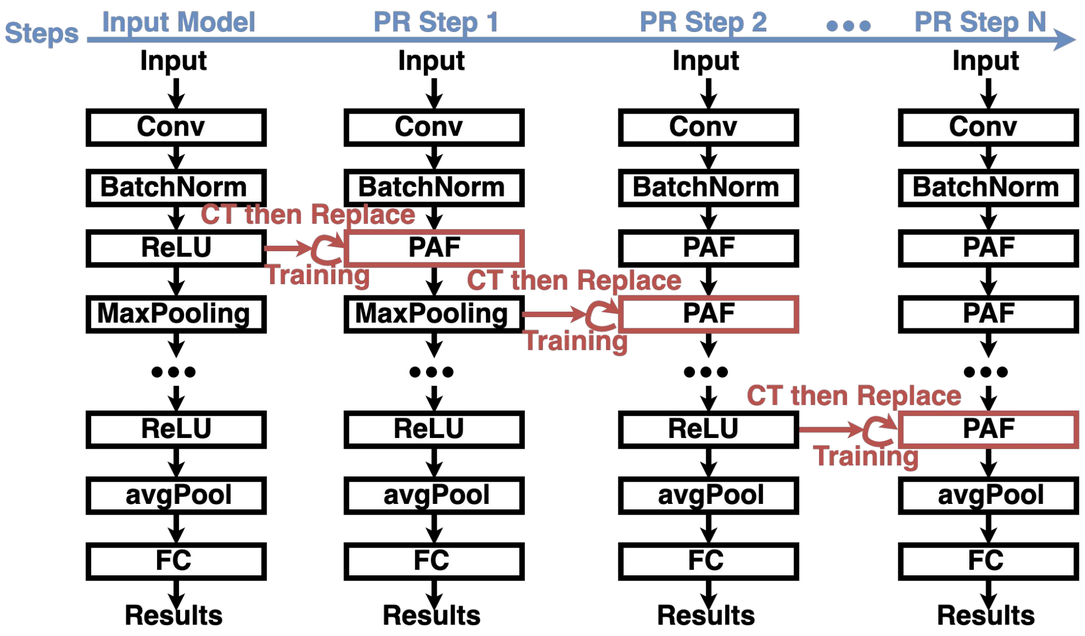
We propose ML training techniques enabling convergence for PAFs with arbitrary degrees!

PAF-FHE Solutions

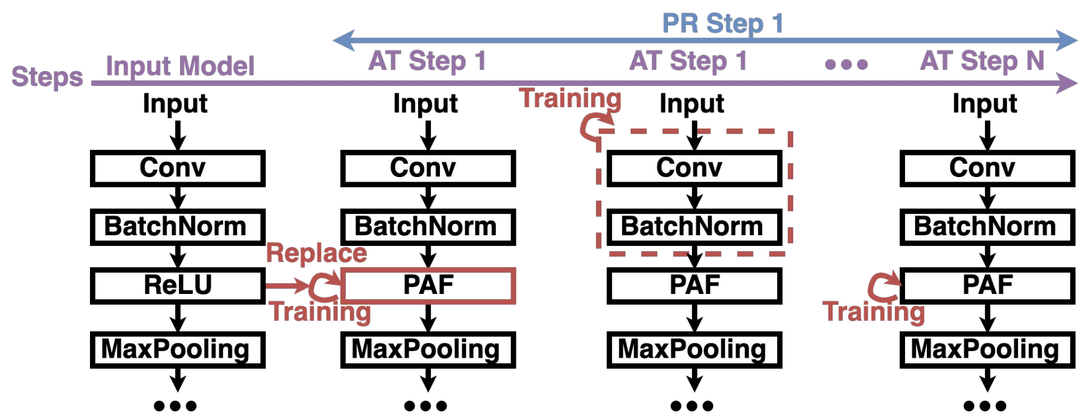
[Coefficients Tuning (CT)] profiles data distribution to obtain close-to-original initializations, improving convergence speed and accuracy.



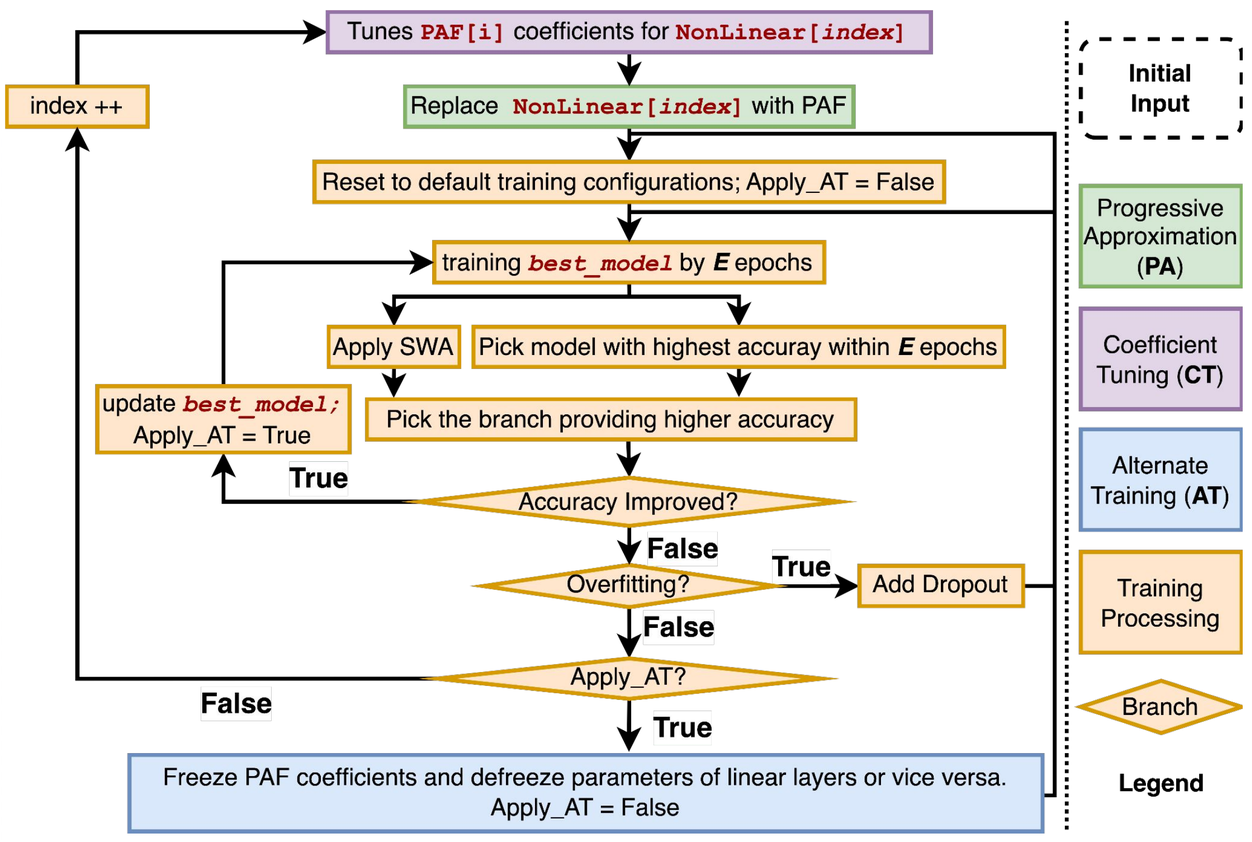
[Progressive Approximation (PA)] PA progressively replaces non-polynomial operators, one layer at a time followed by coefficients fine-tuning, simplifying optimization problem to SGD-optimizable regression.



[Alternative Training (AT)] AT decouples linear operators training from PAFs training to avoiding training divergence.



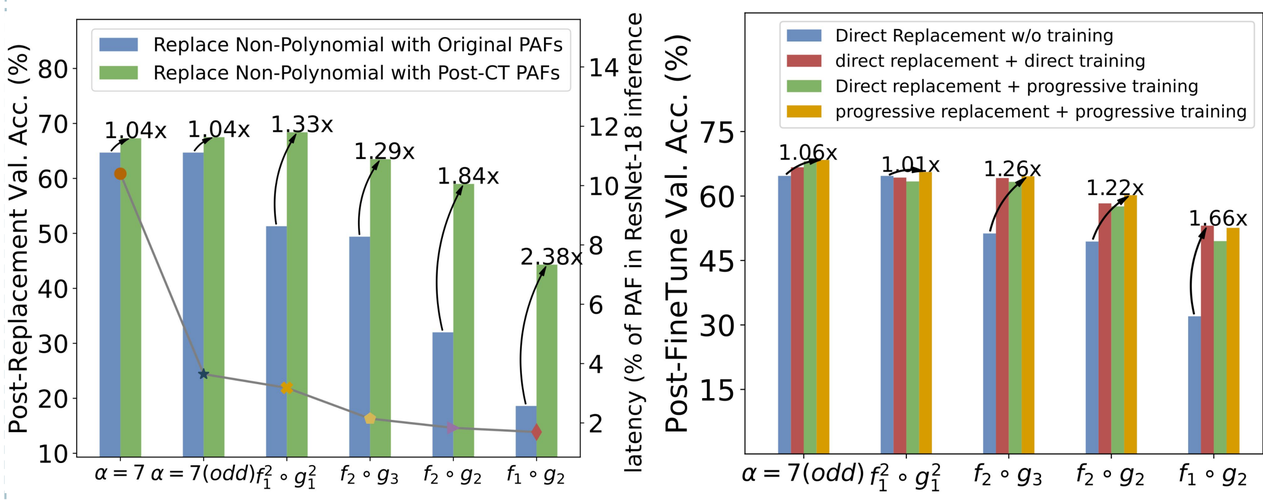
[Systematic Scheduler] Order of appling CT, PA, and AT affects convergence speed and final convergence accuracy.



Experiments

		$f_1^2 \circ g_1^2$	$\alpha = 7$	$f_2 \circ g_3$	$f_2 \circ g_2$	$f_1 \circ g_2$
	Degree	14	12	12	10	8
	Multiplication Depth	12	10	9	8	7
Replace ReLU Only	direct replacement	51.30%	64.70%	49.40%	32.00%	18.60%
	baseline	64.30%	66.70%	64.20%	58.30%	53.10%
	CT	68.60%	67.70%	67.00%	66.50%	61.70%
	AT	65.20%	68.30%	63.70%	60.50%	52.00%
	PA	65.60%	**68.40%**	64.60%	60.20%	52.60%
	PA + AT	64.90%	67.40%	64.60%	56.50%	47.10%
	CT + PA	68.20%	67.00%	**67.60%**	65.90%	60.80%
	CT + PA + AT	**69.00%**	68.10%	61.40%	**66.50%**	**63.10%**
	Accuracy Improvement over Direct Replacement	1.35×	1.06×	1.37×	2.08×	3.39×
	Accuracy Improvement over Baseline	1.07×	1.03×	1.05×	1.14×	1.19×
Replace all non-polynomial	CT + PA + AT + SS	**69.4%**	67%	65.3%	57.3%	6.5%
	Accuracy Improvement over Direct Replacement	1.07×	1.22×	1.27×	1.79×	0.22×
	Accuracy Improvement over Baseline	1.08×	1.01×	1.02×	0.98×	0.12×

CT (PA) improves validation accuracy by 1.04~2.38X (1.06~2.85X)
All techniques combined enable 69.4% (plain model accuracy).



The optimal 12-degree PAF achieves even 0.84% higher accuracy with 72% latency savingv than SotA 27-degree Minimax PAFs.

Conclusions

1. Existing PAFs suffer from either high latency or high post-replacement validation accuracy degradation, limited by training techniques.
2. PAF-FHE (training techniques) enables full degree space exploration.
3. PAF-FHE spots optimal 12-degree PAF with 69.4% acc (the same as original ResNet-18) and saves 72% latency of 27-degree Minimax PAF.