Cipherflow

PEREGRINE: ONE-SHOT FINE-TUNING FOR FHE INFERENCE OF GENERAL DEEP CNNS

Huaming Ling[1], Ying Wang[1], Si Chen[1]
[1]Cipherflow, Shenzhen, China

MAIN CONTRIBUTIONS

- **A single-stage fine-tuning (SFT) strategy**
  - Converting pre-trained CNNs into FHE-friendly forms using low-degree polynomials.
  - Achieving competitive accuracy with minimal training overhead.
- **A generalized interleaved packing (GIP) scheme**
  - Compatible with feature maps of virtually arbitrary spatial resolutions.
  - Proposing homomorphic operators that preserve the GIP-form encryption throughout FHE inference.

SINGLE-STAGE FINE-TUNING

Algorithm 1: Single-Stage Fine-Tuning

Require: Pre-trained CNN model $\mathcal{M}$, dataset $\mathcal{D}$.

Replace activation functions with PolyAct-RN (Algorithm 2);
Replace MaxPool with AvgPool layers;
Resize inputs to target resolution (if needed);
Fine-tune the modified model $\mathcal{M}'$ on $\mathcal{D}$ for few epochs;
return FHE-friendly CNN model $\mathcal{M}'$.

Algorithm 2: PolyAct-RN

Require: Input $X \in \mathbb{R}^{B\times C\times H\times W}$, range param γ, degree-d poly $\mathrm{poly}(\cdot)$, momentum β, initialized stats $\{M_c^{\mathrm{inf}} = 1\}_{c=1}^{C}$, const ϵ.

if *mode* = *training* **then**
  Compute maximum absolute value: $M_c = \max_{b,h,w} |X_{b,c,h,w}|$, $c = 1, 2, \ldots, C$;
  Update running statistic: $M_c^{\mathrm{inf}} = \beta \cdot M_c^{\mathrm{inf}} + (1-\beta)\cdot M_c$, $c = 1, 2, \ldots, C$;
for $c = 1$ *to* C **do**
  $$q_c \leftarrow \begin{cases} \frac{M_c}{\gamma} + \epsilon & \text{if mode=training} \\ \frac{M_c^{\mathrm{inf}}}{\gamma} + \epsilon & \text{if mode=inference} \end{cases};$$
  for $b = 1$ *to* B, $h = 1$ *to* H, $w = 1$ *to* W **do**
    Normalize: $\hat{Y}_{b,c,h,w} = \frac{X_{b,c,h,w}}{q_c}$;
    Evaluate polynomial: $\tilde{Y}_{b,c,h,w} = \mathrm{poly}(\hat{Y}_{b,c,h,w})$;
    Rescale: $Y_{b,c,h,w} = q_c \cdot \tilde{Y}_{b,c,h,w}$;
return $Y \in \mathbb{R}^{B\times C\times H\times W}$.

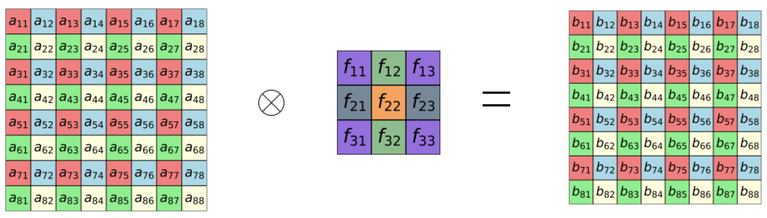
- In experiments, we fix $\gamma = 3$ and the polynomial degree $d = 4$.
- During inference, PolyAct-RN reduces to a fixed fourth-degree polynomial whose coefficients are independent of the input.

GENERALIZED INTERLEAVED PACKING

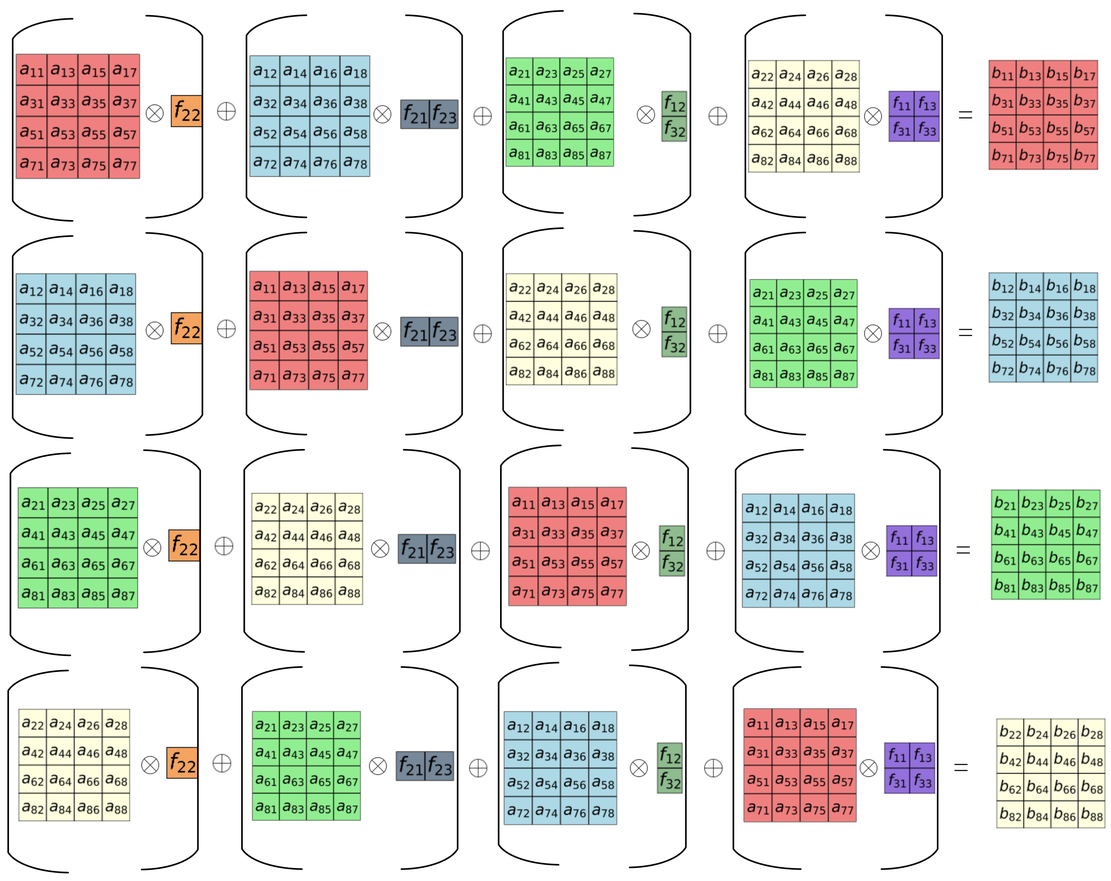
- **Channel packing factor** g
  - Definition: $g = H/\hat{H}$.
  - Image resolution: $H \times H$.
  - Base packing resolution: $\hat{H} \times \hat{H}$ ($\hat{H}^2$ is no greater than ciphertext capacity N/2).
- **Adaptive packing**
  - $g > 1$: Each channel is interleaved into g^2 sub-channels of $\hat{H}^2$ pixels, each packed into a ciphertext (Figure 1(b)).
  - $g < 1$: Pixels from $1/g^2$ channels are interleaved into one ciphertext (Lee et al., 2021).
  - $g = 1$: Both schemes coincide, enabling seamless transition between them.
- **GIP-form preservation at any layer** i
  - Down-sampling (stride s): $g_{i+1} = g_i/s$.
  - Up-sampling (stride $\hat{s}$): $g_{i+1} = g_i \cdot \hat{s}$.
  - Resolution-preserving: $g_{i+1} = g_i$.

ENCRYPTED CONVOLUTION FOR $g > 1$

Case: $H = 8$, $\hat{H} = 4$, input (a_{ij}) and output (b_{ij}) channel packing factor $g_i = g_o = 2$, ciphertext capacity $N/2 = 16$, convolution stride 1, kernel (f_{ij}) size 3, and padding 1.



(a) Stride-1 convolution on plaintext data (packing factor 2)



(b) Stride-1 convolution on GIP-form encrypted data (packing factor 2)

Figure 1. Illustration of the proposed homomorphic convolution. Each output sub-channel is computed by aggregating stride-1 convolutions between the corresponding input sub-channels and their respective kernel coefficients.

IMPLEMENTATION

Evaluation of baseline models (B) and their FHE-friendly versions (F) obtained by our SFT strategy:

Model	Dataset	Input Size	Params	B	F
Classification (Accuracy)					
ResNet18	ImageNet	224^2	11.7 M	0.698	0.678
MobileNetv2	ImageNet	224^2	3.5 M	0.719	0.701
Object Detection (mAP@0.5:0.95)					
YOLOv5n	MS COCO	640^2	1.9 M	0.280	0.257
YOLOv5m	MS COCO	640^2	21.2 M	0.454	0.433
YOLOv5x	MS COCO	640^2	86.7 M	0.507	0.477

FHE inference satisfying a 128-bit security level with a polynomial degree of $N = 2^{16}$:

Model	Dataset	Input Size	Params	mAP	Latency
YOLOv5n	MS COCO	512^2	1.9 M	0.241 (F)	8889.9 *s*

FHE-based inference performance of YOLOv5n using a single CPU thread.

REFERENCES

Lee, Eunsang et al. (2021). *Low-Complexity Deep Convolutional Neural Networks on Fully Homomorphic Encryption Using Multiplexed Parallel Convolutions*. Cryptology ePrint Archive, Paper 2021/1688.