Review on

[MiniONN, 2017], [Gazelle, 2018] and [MPCViT, 2022]

Xue Yufei, Southeast University

03/31/2023

[MiniONN, 2017] Liu J, Juuti M, Lu Y, et al. Oblivious neural network predictions via minionn transformations[C]//Proceedings of the 2017 ACM SIGSAC conference on computer and communications security. 2017: 619-631.
[Gazelle, 2018] Juvekar C, Vaikuntanathan V, Chandrakasan A. {GAZELLE}: A low latency framework for secure neural network inference[C]//27th {USENIX} Security Symposium ({USENIX} Security 18). 2018: 1651-1669.
[MPCViT, 2022] Zeng W, Li M, Xiong W, et al. MPCViT: Searching for MPC-friendly Vision Transformer with Heterogeneous Attention[J]. arXiv preprint arXiv:2211.13955, 2022.

Content

1. MiniONN
- ✓ Protocol for matrix-vector multiplication
- ✓ Protocol for linear/non-linear operation (Sigmoid)

2. GAZELLE
- ✓ FHE
- ✓ SIMDScMult/Conv

3. MPCViT
- ✓ Idea

MiniONN: Multiplication

Input:

$\mathcal{S}$: $\mathbf{w} \in \mathbb{Z}_N^n$

$\mathcal{C}$: $\mathbf{r} \in \mathbb{Z}_N^n$

Output:

$\mathcal{S}$: a random number $u \in \mathbb{Z}_N$;

$\mathcal{C}$: $v \in \mathbb{Z}_N$, s.t., $u + v \pmod N = \mathbf{w} \cdot \mathbf{r}$.

$\mathcal{S}$:		$\mathcal{C}$:
$\widetilde{\mathbf{w}} \leftarrow E(pk_s, \mathbf{w})$		$\mathbf{v} \xleftarrow{\$} \mathbb{Z}_N^n$
	$\xrightarrow{\widetilde{\mathbf{w}}}$	
		$\widetilde{\mathbf{u}} \leftarrow \mathbf{r} \otimes \widetilde{\mathbf{w}} \ominus \mathbf{v}$
	$\xleftarrow{\widetilde{\mathbf{u}}}$	
$u \leftarrow \sum(D(sk_s, \widetilde{\mathbf{u}}))$		$v \leftarrow \sum(\mathbf{v})$
output u		**output** v

Figure 3: Dot-product triplet generation.

Input:

- $\mathcal{S}$: a vector $\mathbf{w} \in \mathbb{Z}_N^n$;
- $\mathcal{C}$: a random vector $\mathbf{r} \in \mathbb{Z}_N^n$.

Output:

- $\mathcal{S}$: a random number $u \in \mathbb{Z}_N$;
- $\mathcal{C}$: $v \in \mathbb{Z}_N$, s.t., $u + v \pmod N = \mathbf{w} \cdot \mathbf{r}$.

Figure 1: Ideal functionality $\mathcal{F}_{\text{triplet}}$: generate a dot-product triplet.

$u \rightarrow$ S

$\mathbf{r}, v \rightarrow$ C

MiniONN: Linear operation

Input:

$\mathcal{S}$: $\mathbf{W} \in \mathbb{Z}_N^{m\times l}$, $\mathbf{X}^{\mathcal{S}} \in \mathbb{Z}_N^{l\times n}$, $\mathbf{B} \in \mathbb{Z}_N^{m\times n}$

$\mathcal{C}$: $\mathbf{X}^{\mathcal{C}} \in \mathbb{Z}_N^{l\times n}$

Output:

$\mathcal{S}$: A random matrix $\mathbf{Y}^{\mathcal{S}}$

$\mathcal{C}$: $\mathbf{Y}^{\mathcal{C}}$ s.t., $\mathbf{Y}^{\mathcal{C}} + \mathbf{Y}^{\mathcal{S}} = \mathbf{W} \cdot (\mathbf{X}^{\mathcal{C}} + \mathbf{X}^{\mathcal{S}}) + \mathbf{B}$

$\mathcal{S}$: $\qquad\qquad\qquad$ $\mathcal{C}$:

precomputation

for $i = 1$ **to** m

for $j = 1$ **to** n

$$(u_{i,j}, v_{i,j}) \leftarrow \mathcal{F}_{\text{triplet}}(\mathbf{w}_i, \mathbf{x}_j^{\mathcal{C}})$$

$$\longleftrightarrow$$

end

end

$\mathbf{Y}^{\mathcal{S}} := \mathbf{W} \cdot \mathbf{X}^{\mathcal{S}} + \mathbf{B} + \mathbf{U}$ $\qquad$ $\mathbf{Y}^{\mathcal{C}} := \mathbf{V}$

output $\mathbf{Y}^{\mathcal{S}}$ $\qquad$ **output** $\mathbf{Y}^{\mathcal{C}}$

Figure 4: Oblivious linear transformation.

→ 预生成三元组，节省开销

→ $\mathbf{Y}^C + \mathbf{Y}^S = \mathbf{Y}$

MiniONN: Activation functions

ReLU: garbled circuit

Input:

- $\mathcal{S}$: $y^{\mathcal{S}} \in \mathbb{Z}_N$;
- $\mathcal{C}$: $y^{\mathcal{C}}, r \in \mathbb{Z}_N$.

Output:

- $\mathcal{S}$: $x^{\mathcal{S}} := compare(y,0) \cdot y - r \pmod N$ where $y = y^{\mathcal{S}} + y^{\mathcal{C}} \pmod N$;
- $\mathcal{C}$: $x^{\mathcal{C}} := r$.

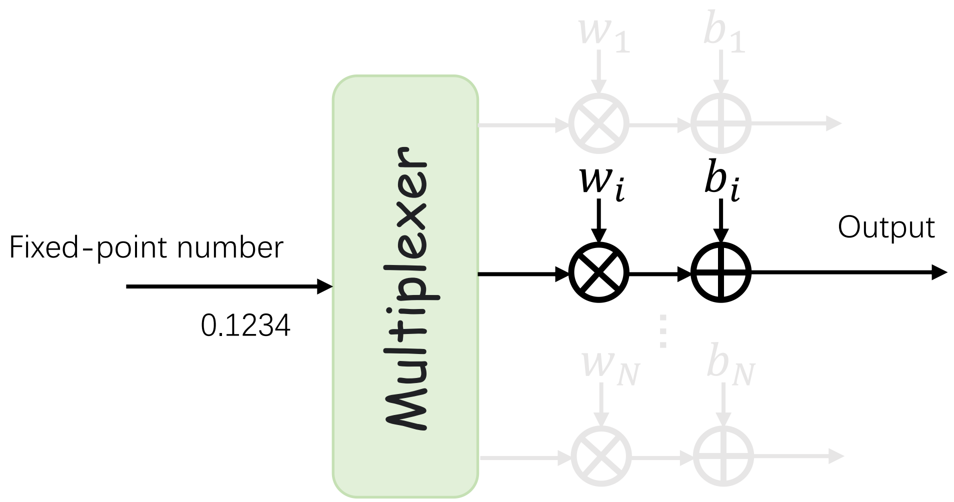
Figure 5: The ideal functionality $\mathcal{F}_{\mathrm{ReLU}}$.

Sigmoid:

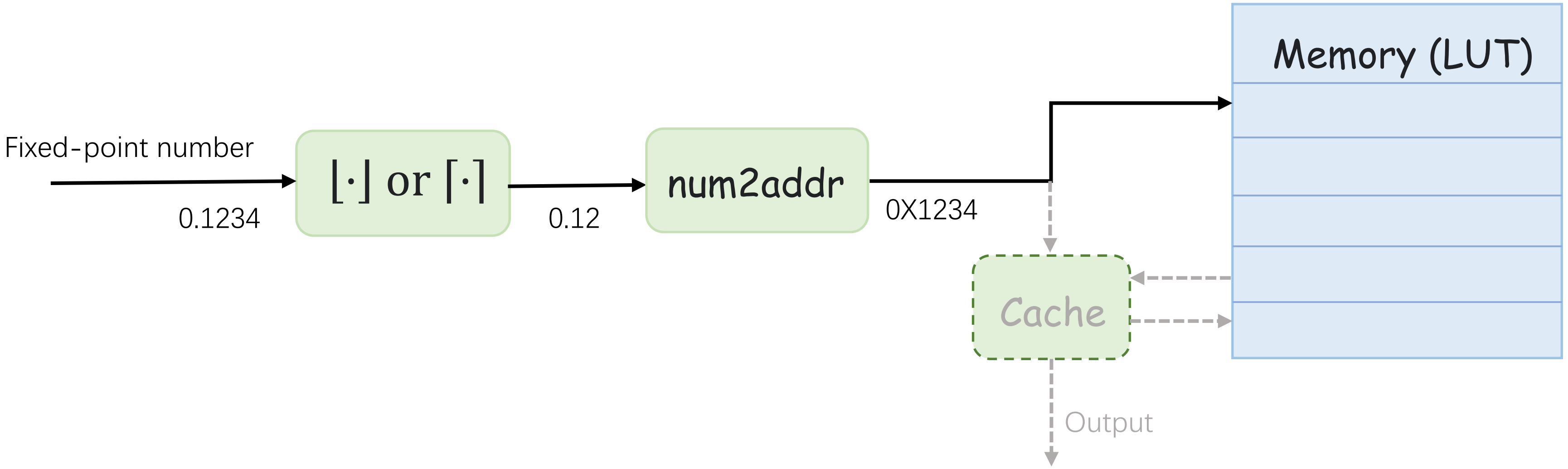
1. 多项式拟合 例：$x^3 + x^2 + x = x \times (x^2 + x + 1) \rightarrow x \times x \times (x + 1)$
2. 分段计算，会用到$compare$单元

MiniONN: Sigmoid

1. 使用分段线性拟合

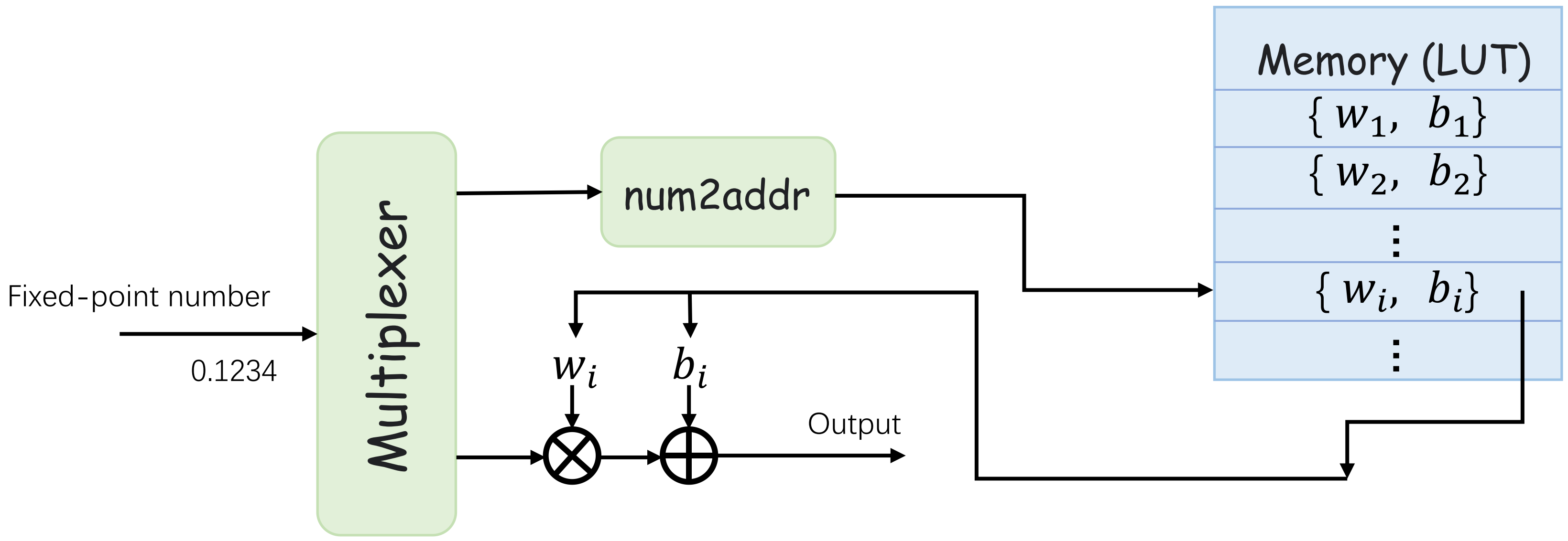


2. 使用LUT计算非线性函数 (A^{-1})



MiniONN: Sigmoid

3. 复用计算单元



GAZELLE: FHE

加法同态：$\mathbf{En}(x+y)=\mathbf{En}(x)\oplus\mathbf{En}(y)$

乘法同态：$\mathbf{En}(x\times y)=\mathbf{En}(x)\otimes\mathbf{En}(y)$

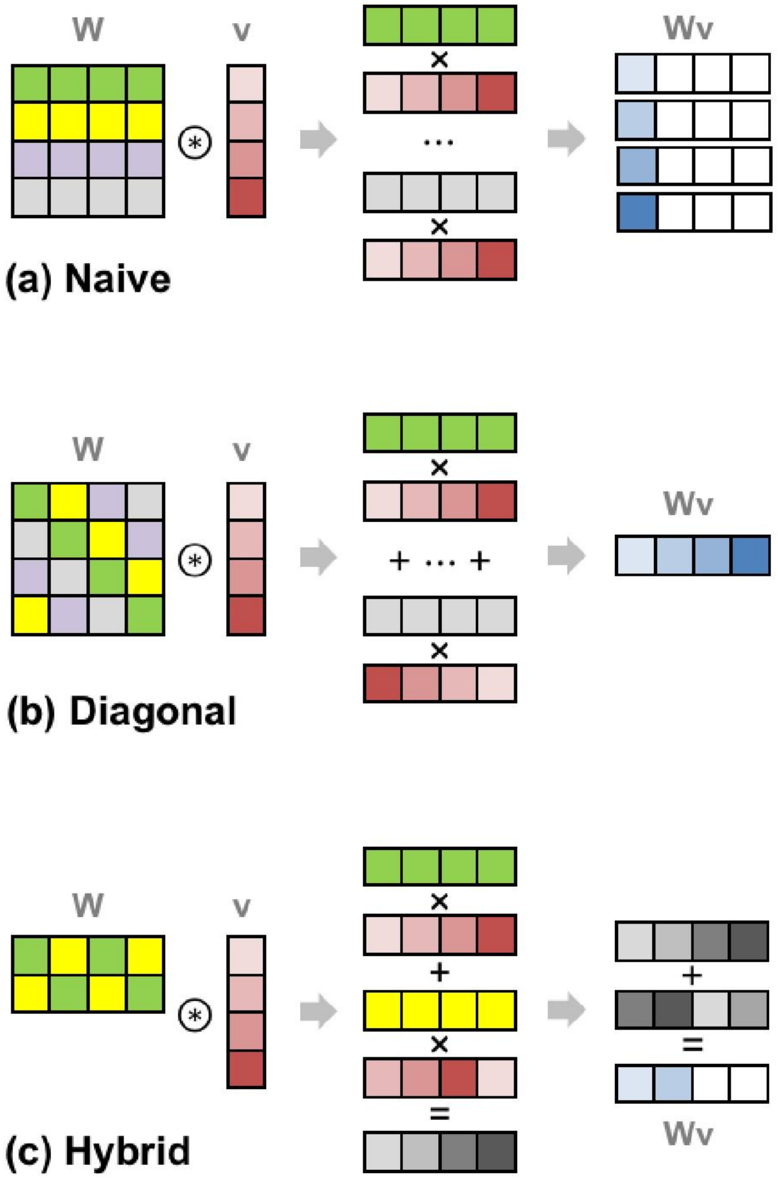
v.s.

$\mathbf{FFT}(\boldsymbol{x}+\boldsymbol{y})=\mathbf{FFT}(\boldsymbol{x})+\mathbf{FFT}(\boldsymbol{y})$

$\mathbf{FFT}(\boldsymbol{x}\times\boldsymbol{y})=\mathbf{FFT}(\boldsymbol{x})*\mathbf{FFT}(\boldsymbol{y})$

线性系统的线性性质

GAZELLE: SIMDScMult



- How did the authors think about it?
- Is there an optimization?
  (To note)

GAZELLE: SIMDScMult

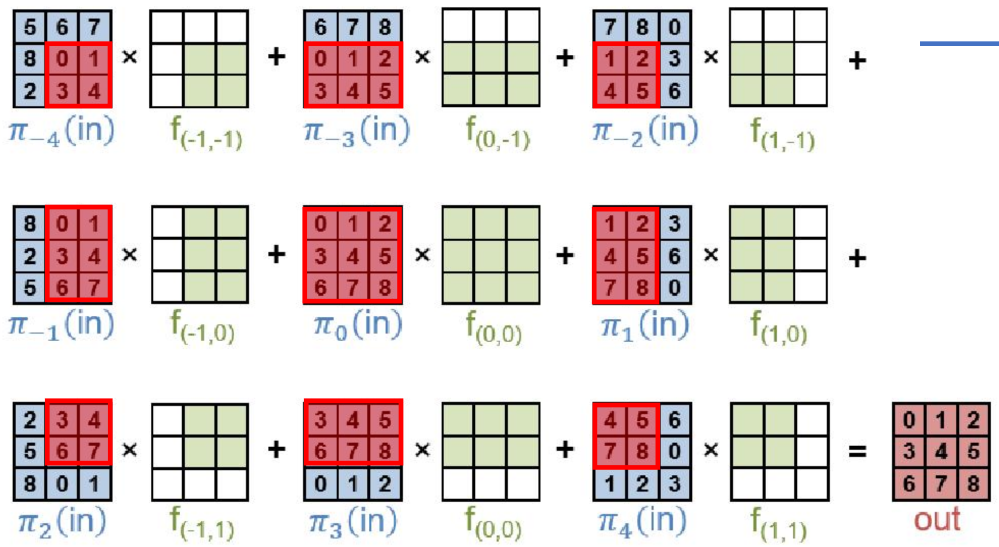
GAZELLE: SIMDScMult

GAZELLE: SIMDScMult

GAZELLE: SIMDScMult

GAZELLE: Conv

Optimization for padding



的部分保留，其余部分置零

Channel Packing:

Input c_i, output c_o, integer c_n → Input $\frac{c_i}{c_n}$, output $\frac{c_o}{c_n}$

计算结果一样吗？（to note）

GAZELLE: Conv

MPCViT:

General idea:

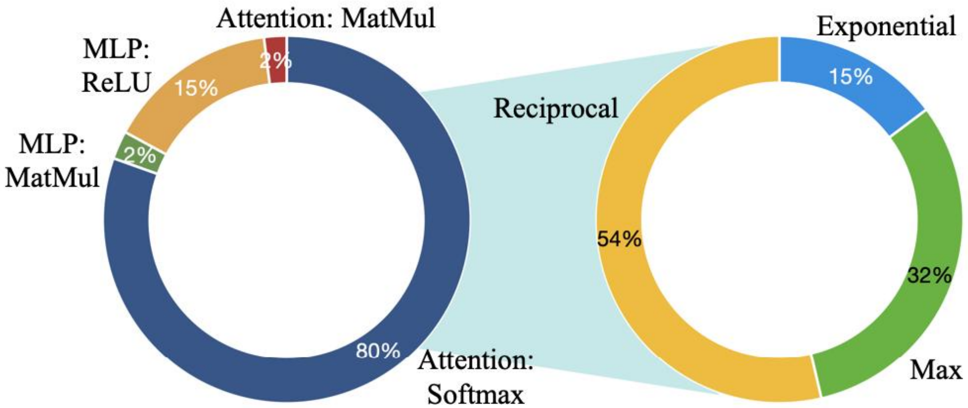
Low-latency NLP skills → CV

MPCViT: Searching for MPC-friendly Vision Transformer with Heterogeneous Attention

$\alpha A + (1-\alpha)B$

Step:

1. find most time-consuming operation



(a) (b)

2. get rid of e^x, replaced by ScaleAttn+RSAttn

Type	Top-1 Acc. (%)	Latency (s)
Softmax Attention	92.69	6.82
ReLU Attention	fail	fail
ReLU6 Attention	90.50	3.02
Sparsemax Attention [24]	91.23	3.23
XNorm Attention [10]	91.24	13.25
Square Attention	91.27	0.72
2Quad Attention [23]	91.86	4.22
ScaleAttn [12]	91.52	**0.66**
ReLU Softmax Attention (RSAttn) [14]	**92.31**	5.32

MPCViT:

1. Differentiable NAS

$$\alpha \cdot \mathrm{ReLUSoftmax}(\frac{QK^T}{\sqrt{d_k}})V + (1-\alpha) \cdot \frac{\mathrm{ScaleAttn}(Q,K,V)}{\sqrt{d_k}}$$

取RSAttn的准确性，取ScaleAttn的低时延，更倾向ScaleAttn

2. How to effectively train?

Token-wise feature-based KD