

MaxQ: Multi-Axis Query for N:M Sparsity Network

Jingyang Xiang¹, Siqi Li¹, JunHao Chen¹, Zhuangzhi Chen²,

Tianxin Huang¹, Linpeng Peng¹, Yong Liu¹

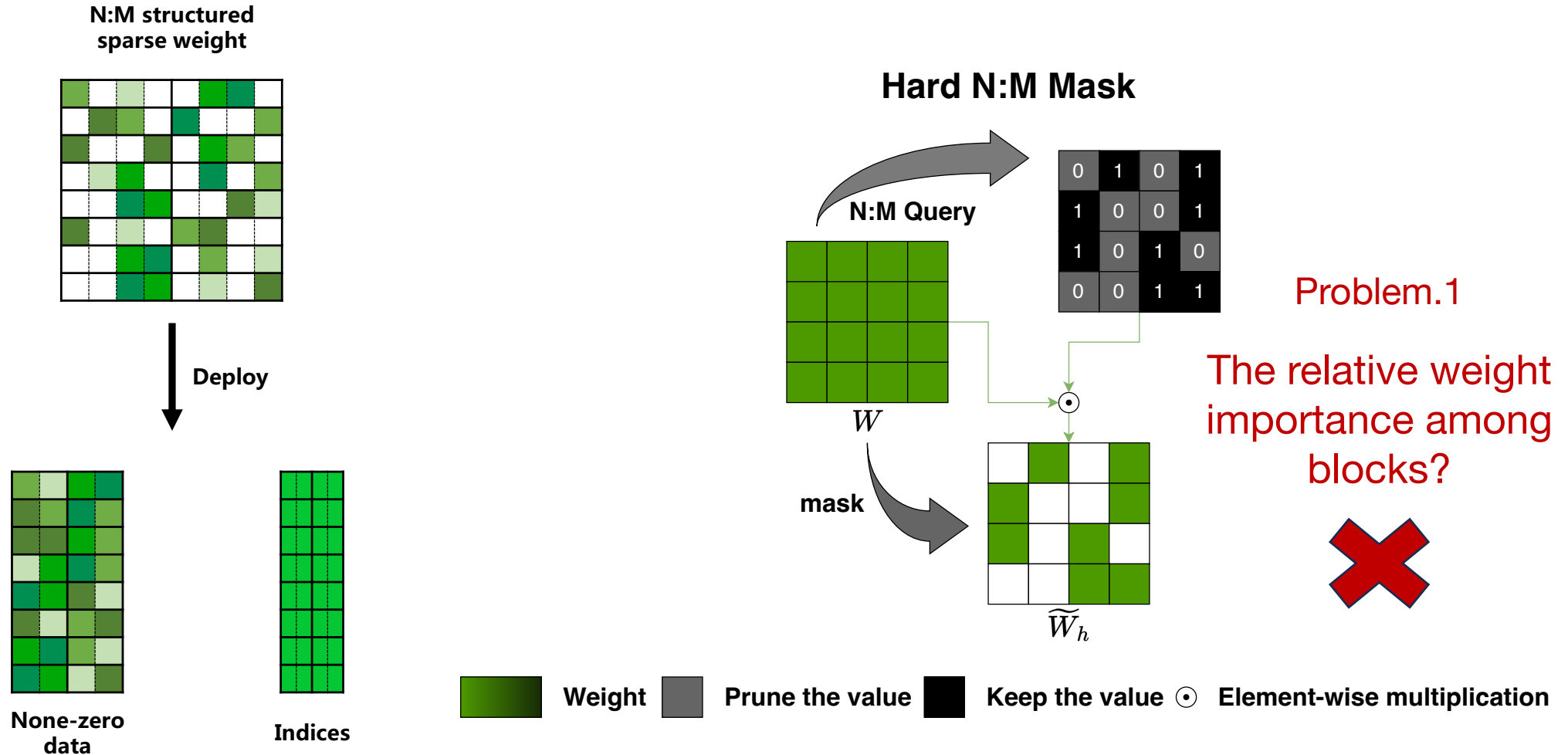
APRIL Lab, Zhejiang University, Hangzhou, China¹

IVSN, Zhejiang University of Technology, Hangzhou, China²

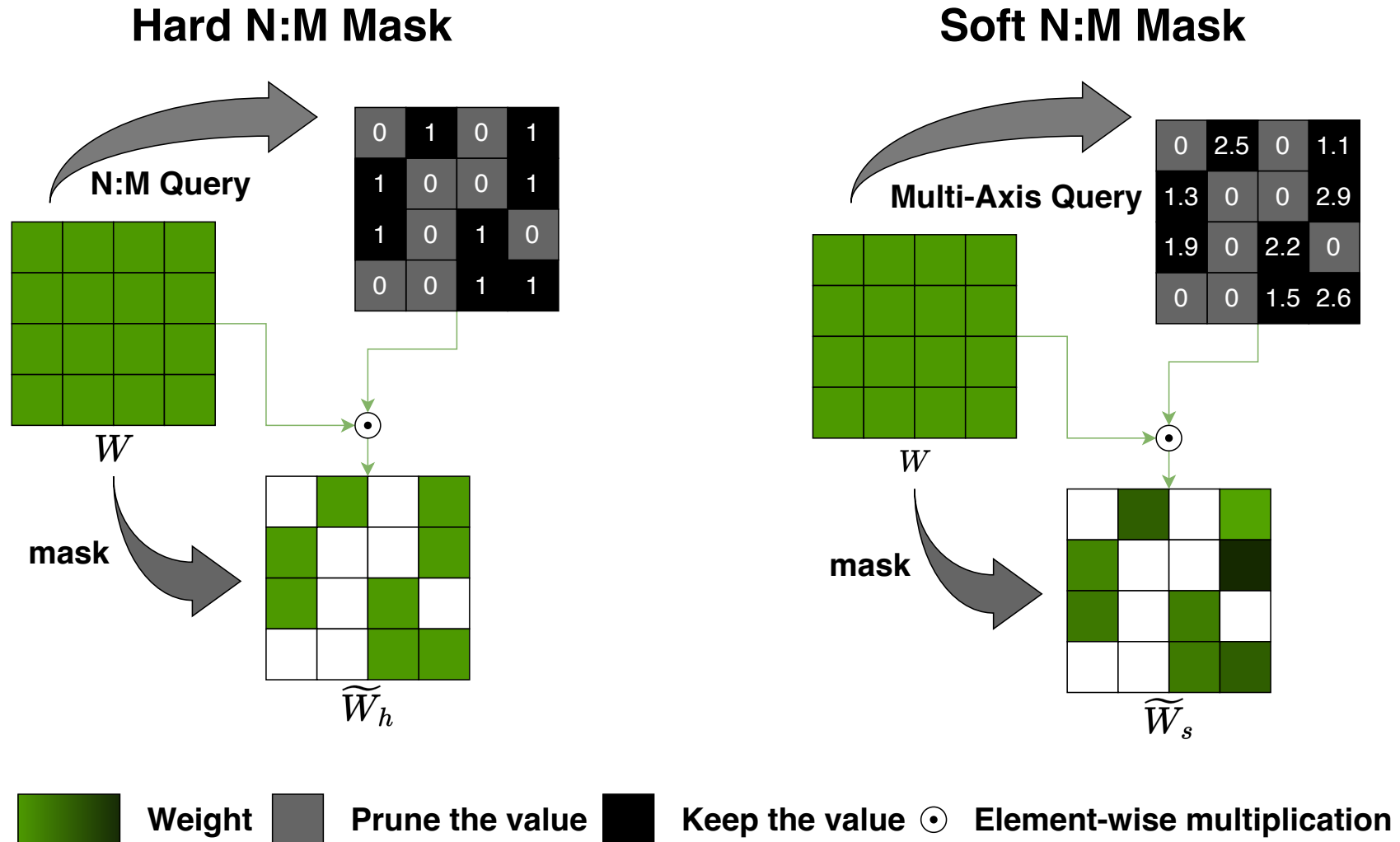
<https://github.com/JingyangXiang/MaxQ>



Overview: N:M Sparsity



Take Weight Importance into Account



Multi-Axis Query

1. Apply sparsity

$$\mathcal{M}_g^l = \text{ArgTopK}_{M \times N} (-|\mathbf{m}_{g,:}^l|)$$

$$\mathcal{T}_t^l = \text{ArgTopK}_{\lceil G^l \delta_t \rceil} \left(\left\{ \|\mathbf{m}_g^l\|_1 \right\} \right) \quad \text{zeroizing} \quad \{\mathbf{b}_{i,j}^l | i \in \mathcal{T}_t^l, j \in \mathcal{M}_i^l\}$$

2. Measure importance

$$\begin{cases} \sigma_h = \min(\text{topK}(\text{abs}(V), (1-p) \cdot N)) \\ \sigma_l = \max(\text{topK}(-\text{abs}(V), p \cdot N)) \\ \sigma = (\text{abs}(\sigma_h) + \text{abs}(\sigma_l)) / 2 \end{cases}$$

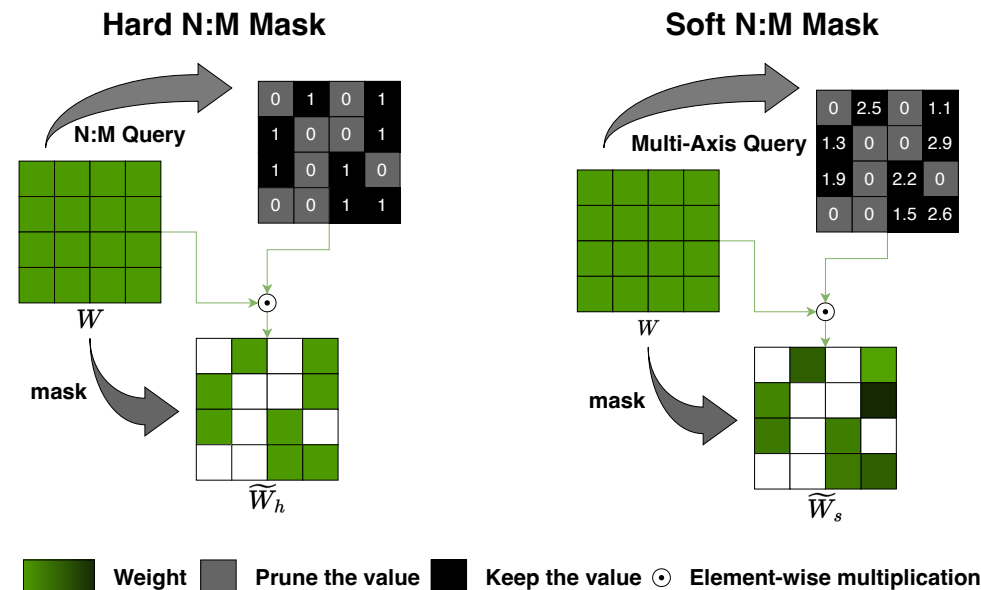
$$s_i = \text{sigmoid}((|v_i| - \sigma) / \tau)$$

$$\mathbf{s}_{i,\dots,\dots}^{l(f)} = (G(w_{i,\dots,\dots}^l, (M-N)/M, \tau))$$

$$\mathbf{s}_{\dots,\dots,k_1,k_2}^{l(k)} = (G(w_{\dots,\dots,k_1,k_2}^l, (M-N)/M, \tau))$$

$$\mathbf{s}^l = \mathbf{b}^l + \mathbf{b}^l \odot \text{RA}(\mathbf{s}^{l(f)}) + \mathbf{b}^l \odot \text{RA}(\mathbf{s}^{l(k)})$$

$$= \mathbf{b}^l \odot \left(1 + \text{RA}(\mathbf{s}^{l(f)}) + \text{RA}(\mathbf{s}^{l(k)}) \right)$$



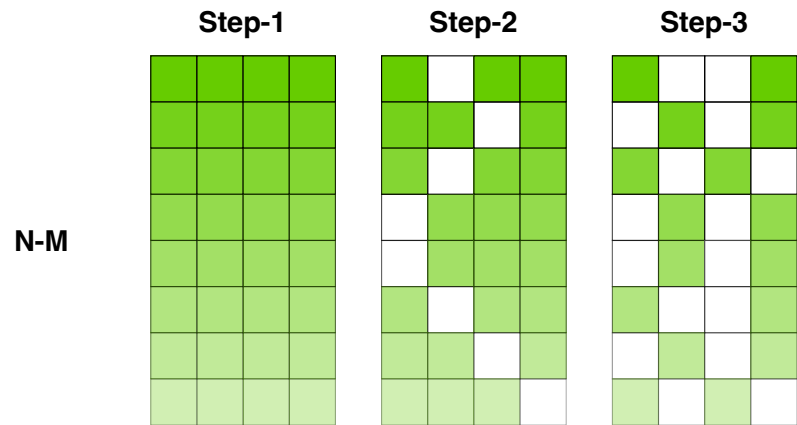
$$\text{threshold} = \frac{0.25 + 0.38}{2} = 0.315$$

0.08	0.25	0.38	0.99
w_1	w_2	w_3	w_4

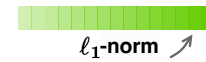
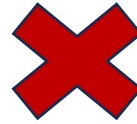
$$s_i = \text{sigmoid}\left(\frac{w_i - \text{threshold}}{\tau}\right) \Big|_{\tau=0.1}$$

0.0871	0.3430	0.6570	0.9988
s_1	s_2	s_3	s_4

Overview: Sparsity Strategy



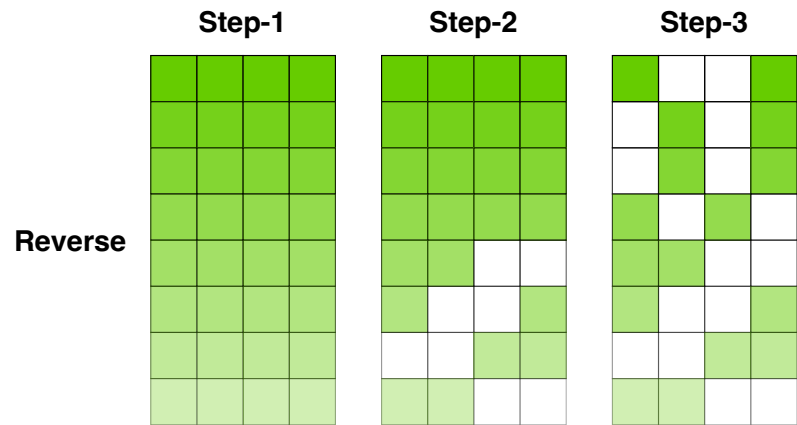
Problem.2 Unstable Incremental Sparsity Strategy



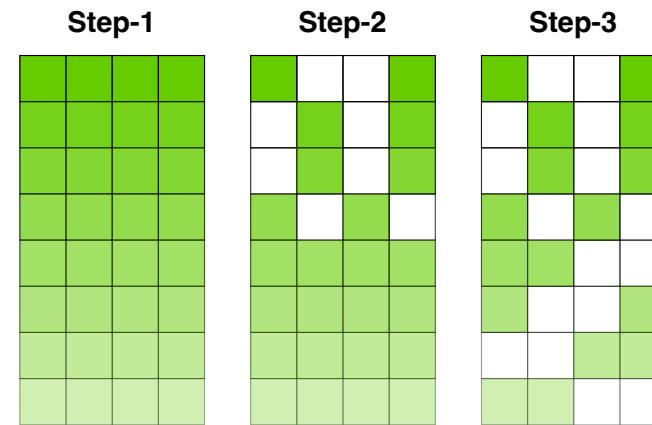
$l_1\text{-norm} \nearrow$

$$\mathbf{W} \in \mathbb{R}^{G \times M}, G \gg M$$

More intermediate states




Default

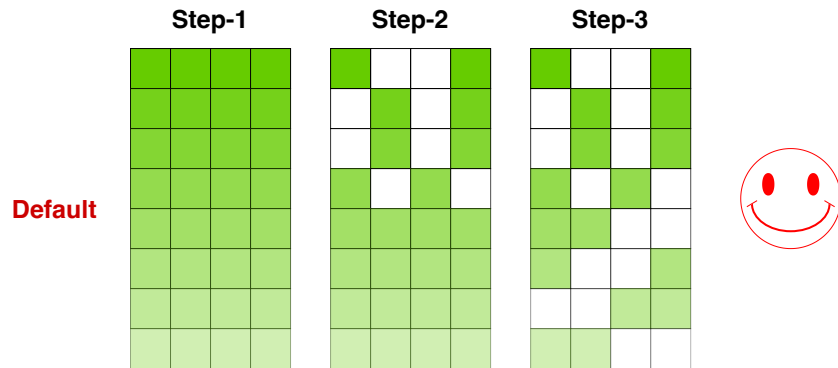


Purpose: Smoother training process.

Incremental Sparsity Strategy

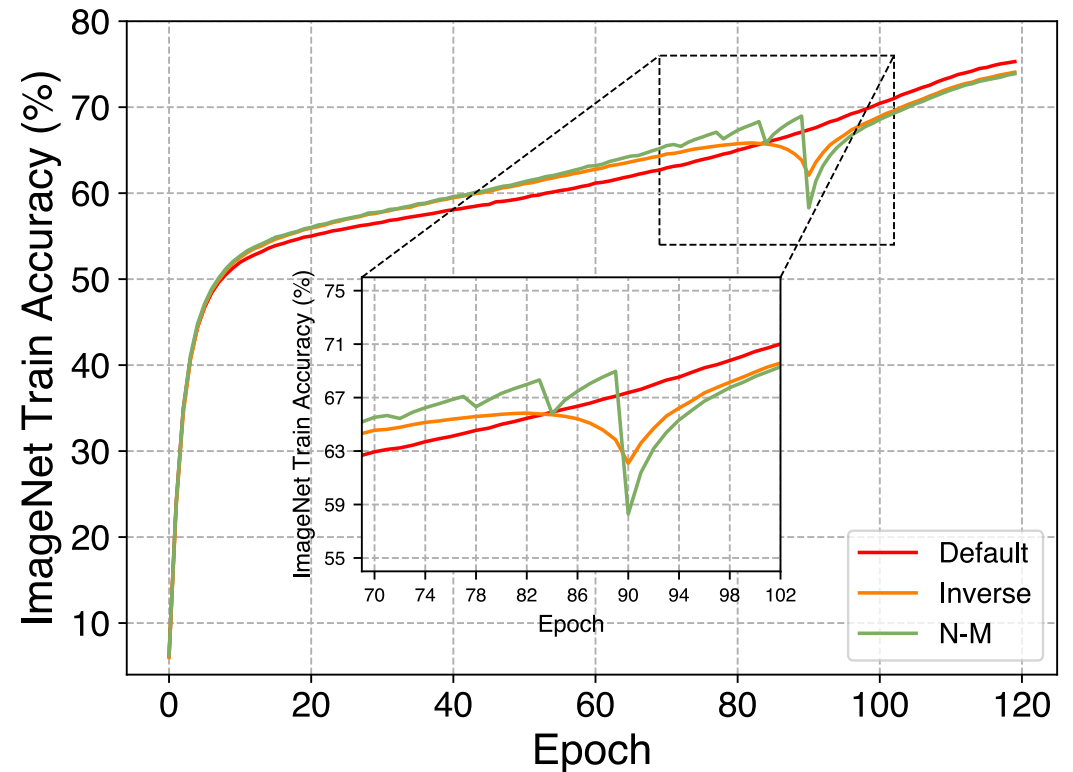

 $l_1\text{-norm} \nearrow$
 $\mathbf{W} \in \mathbb{R}^{G \times M}, G \gg M$

More intermediate states



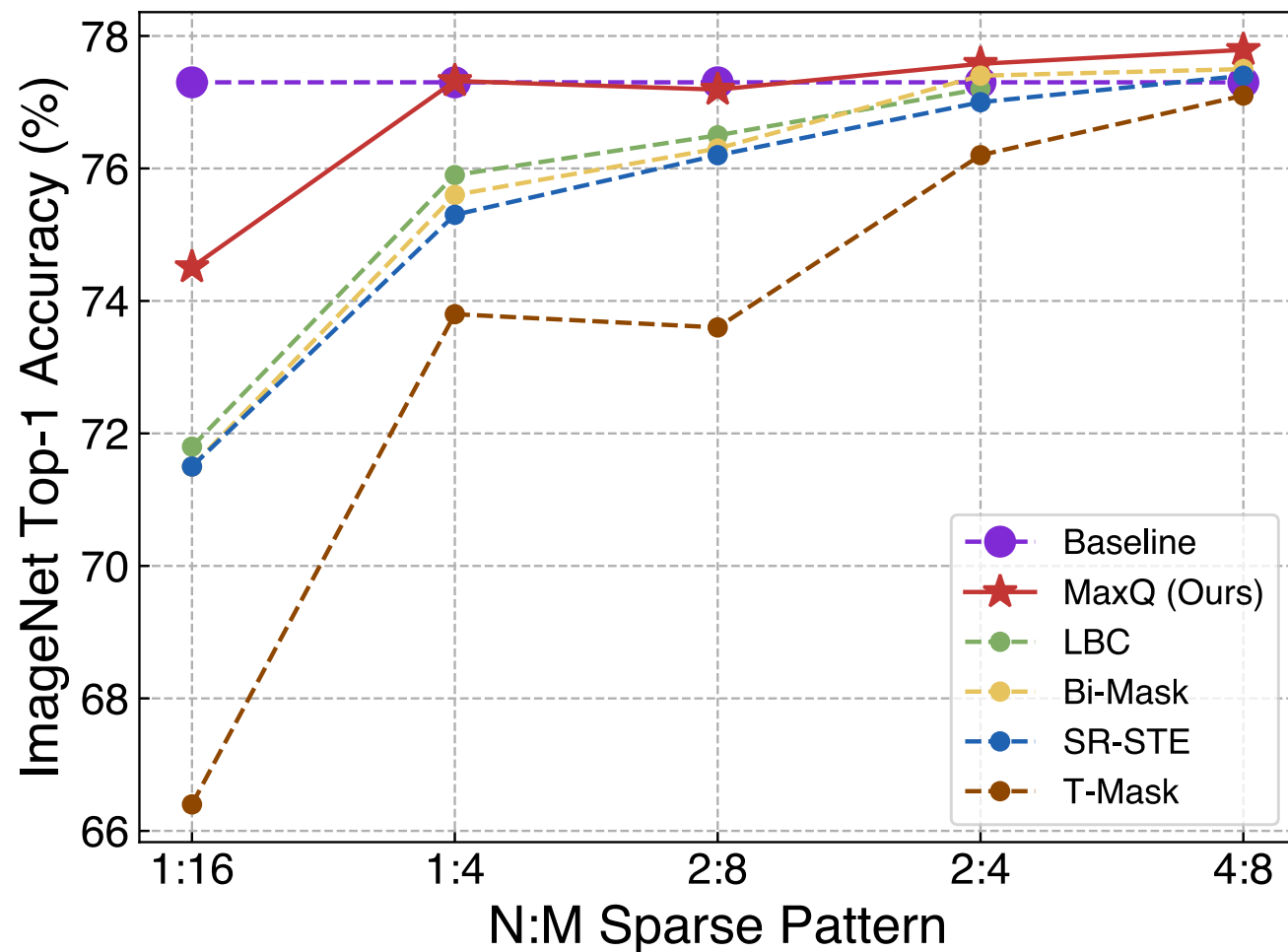
N:M Sparse Blocks Ratio

$$\delta_t = \min(1, \max(0, 1 - [1 - (t - t_i)/(t_f - t_i)]^3))$$



Our sparsity strategy achieves a smoother training process.

Results ResNet50 Pareto



Results ImageNet1K

Model	Method	N:M	Top-1	Epochs	FLOPs	Params
ResNet34	Baseline	-	74.6%	120	3.67G	21.8M
	ASP	1:4	70.9%	200	1.01G	5.85M
	SR-STE	1:4	73.8%	120	1.01G	5.85M
	LBC	1:4	73.7%	120	1.01G	5.85M
	MaxQ	1:4	74.2%	120	1.01G	5.85M
	ASP	2:4	73.9%	200	1.90G	11.2M
	SR-STE	2:4	74.3%	120	1.90G	11.2M
	LBC	2:4	74.1%	120	1.90G	11.2M
	MaxQ	2:4	74.5%	120	1.90G	11.2M
	Baseline	-	77.3%	120	4.11G	25.6M
	ASP	2:4	77.4%	200	2.12G	13.8M
	SR-STE	2:4	77.0%	120	2.12G	13.8M
LBC	2:4	77.2%	120	2.12G	13.8M	
MaxQ	2:4	77.6%	120	2.12G	13.8M	
ResNet50	ASP	1:4	76.5%	200	1.11G	7.93M
	SR-STE	1:4	75.3%	120	1.11G	7.93M
	LBC	1:4	75.9%	120	1.11G	7.93M
	MaxQ	1:4	77.3%	120	1.11G	7.93M
	ASP	2:8	76.6%	200	1.11G	7.93M
	SR-STE	2:8	76.2%	120	1.11G	7.93M
	LBC	2:8	76.5%	120	1.11G	7.93M
	MaxQ	2:8	77.2%	120	1.11G	7.93M
	ASP	1:16	71.5%	200	0.44G	3.52M
	SR-STE	1:16	71.5%	120	0.44G	3.52M
	LBC	1:16	71.8%	120	0.44G	3.52M
	MaxQ	1:16	74.6%	120	0.44G	3.52M
Model	Method	N:M	Top-1	Epochs	FLOPs	Params
DeiT-Small	Baseline	-	79.8%	300	4.6G	22.1M
	SR-STE	2:4	75.7%	300	2.5G	11.4M
	LBC	2:4	78.0%	300	2.5G	11.4M
	MaxQ	2:4	78.5%	300	2.5G	11.4M

Model	Method	N:M	Top-1	Epochs	FLOPs	Params	
MobileNetV1	Baseline	-	71.9%	120	578M	4.23M	
	ASP	2:4	70.4%	200	302M	2.66M	
	SR-STE	2:4	71.5%	120	302M	2.66M	
	MaxQ	2:4	72.1%	120	302M	2.66M	
	ASP	1:4	65.4%	200	164M	1.88M	
	SR-STE	1:4	67.8%	120	164M	1.88M	
	MaxQ	1:4	68.5%	120	164M	1.88M	
	Method	Top-1	Sparsity	FLOPs	Params	S	U
	Baseline	77.3%	0.0	4.10G	25.6M	-	-
	RigL [10]	74.6%	80	0.92G	5.12M	✗	✓
GMP [46]	75.6%	80	0.82G	5.12M	✗	✓	
MAP [2]	75.9%	80	-	5.12M	✗	✗	
STR [27]	76.2%	81	0.82G	5.12M	✗	✓	
SR-STE	75.3%	1:4	1.13G	7.97M	✓	✓	
LBC	75.9%	1:4	1.13G	7.97M	✓	✓	
MaxQ	77.3%	1:4	1.13G	7.97M	✓	✓	
SR-STE	76.2%	2:8	1.13G	7.97M	✓	✓	
LBC	76.5%	2:8	1.13G	7.97M	✓	✓	
MaxQ	77.2%	2:8	1.13G	7.97M	✓	✓	
DNW [42]	68.3%	95	0.20G	1.28M	✗	✗	
RigL [10]	70.0%	95	0.49G	1.28M	✗	✓	
GMP [46]	70.6%	95	0.20G	1.28M	✗	✓	
STR [27]	70.4%	95	0.16G	1.24M	✗	✗	
OptG [43]	72.5%	95	0.22G	1.28M	✗	✗	
SR-STE	71.5%	1:16	0.44G	3.52M	✓	✓	
LBC	71.8%	1:16	0.44G	3.52M	✓	✓	
MaxQ	74.6%	1:16	0.44G	3.52M	✓	✓	

Results COCO2017

Model	Method	N:M	mAP
F-RCNN	Baseline	-	37.4
	SR-STE	2:4	38.2
	LBC	2:4	38.5
	MaxQ	2:4	38.7
	SR-STE	1:4	37.2
	LBC	1:4	37.3
	MaxQ	1:4	37.7

Object Detection

Model	Method	N:M	Box mAP	Mask mAP
M-RCNN	Baseline	-	38.2	34.7
	SR-STE	2:4	39.0	35.3
	LBC	2:4	39.3	35.4
	MaxQ	2:4	39.2	35.5
	SR-STE	1:4	37.6	33.9
	LBC	1:4	37.8	34.0
	MaxQ	1:4	38.3	34.4

Instance Segmentation

Efficiency

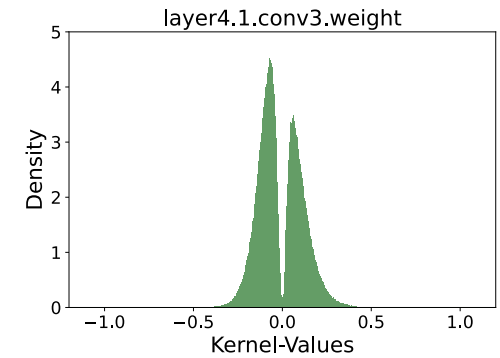
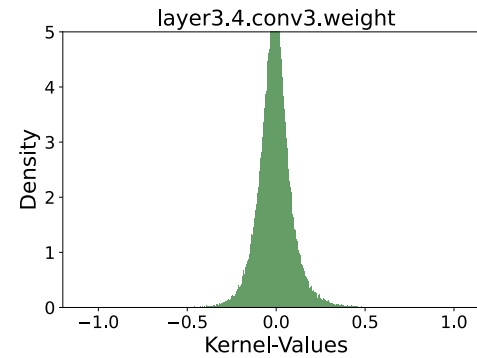
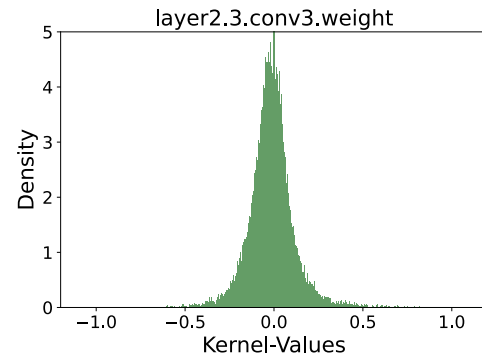
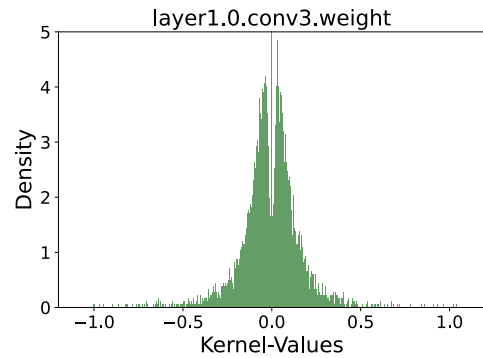
Model	N:M	Method	Train speed		Top-1	FLOPs (Train)
			BS=128	BS=256		
ResNet50	-	Dense	798	884	77.3%	1 × (3.2e18)
	2:4	SR-STE	642	854	77.0%	0.83 ×
		LBC	373	487	77.2%	0.72 ×
		MaxQ	507	732	77.6%	0.91 ×
	2:8	SR-STE	625	862	76.2%	0.74 ×
		LBC	382	512	76.5%	0.53 ×
		MaxQ	514	743	77.2%	0.86 ×
	1:16	SR-STE	628	852	71.5%	0.69 ×
		LBC	364	538	71.8%	0.38 ×
		MaxQ	502	725	74.6%	0.81 ×

Train Speed: **SR-STE > MaxQ > LBC**

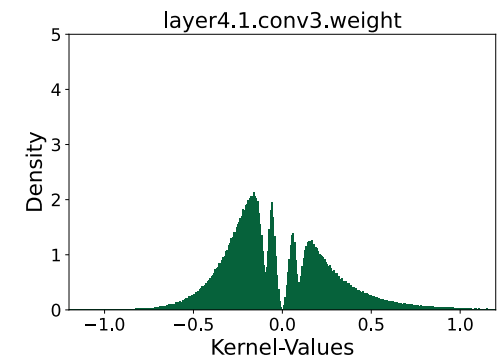
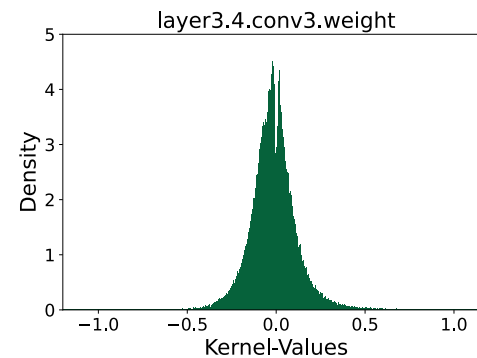
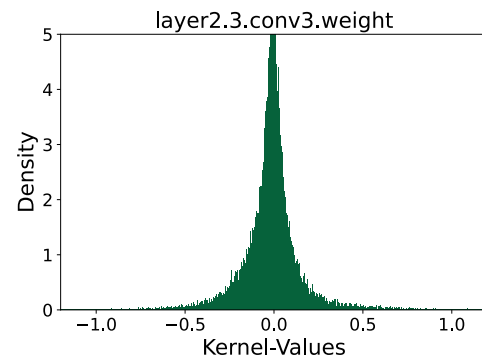
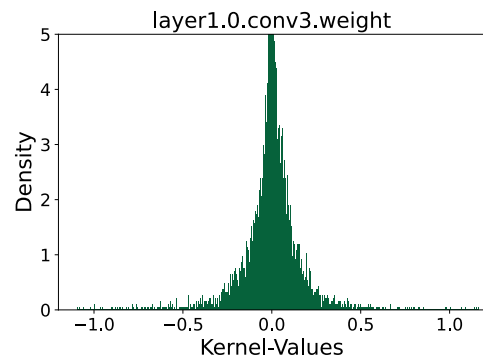
Performance: **MaxQ > LBC > SR-STE**

Friendly to Quantization (ResNet50-2:4)

SR-STE 77.1% -> 76.6%



MaxQ 77.6% -> 77.1%



MaxQ: Multi-Axis Query for N:M Sparsity Network



<https://github.com/JingyangXiang/MaxQ>

Thank you!