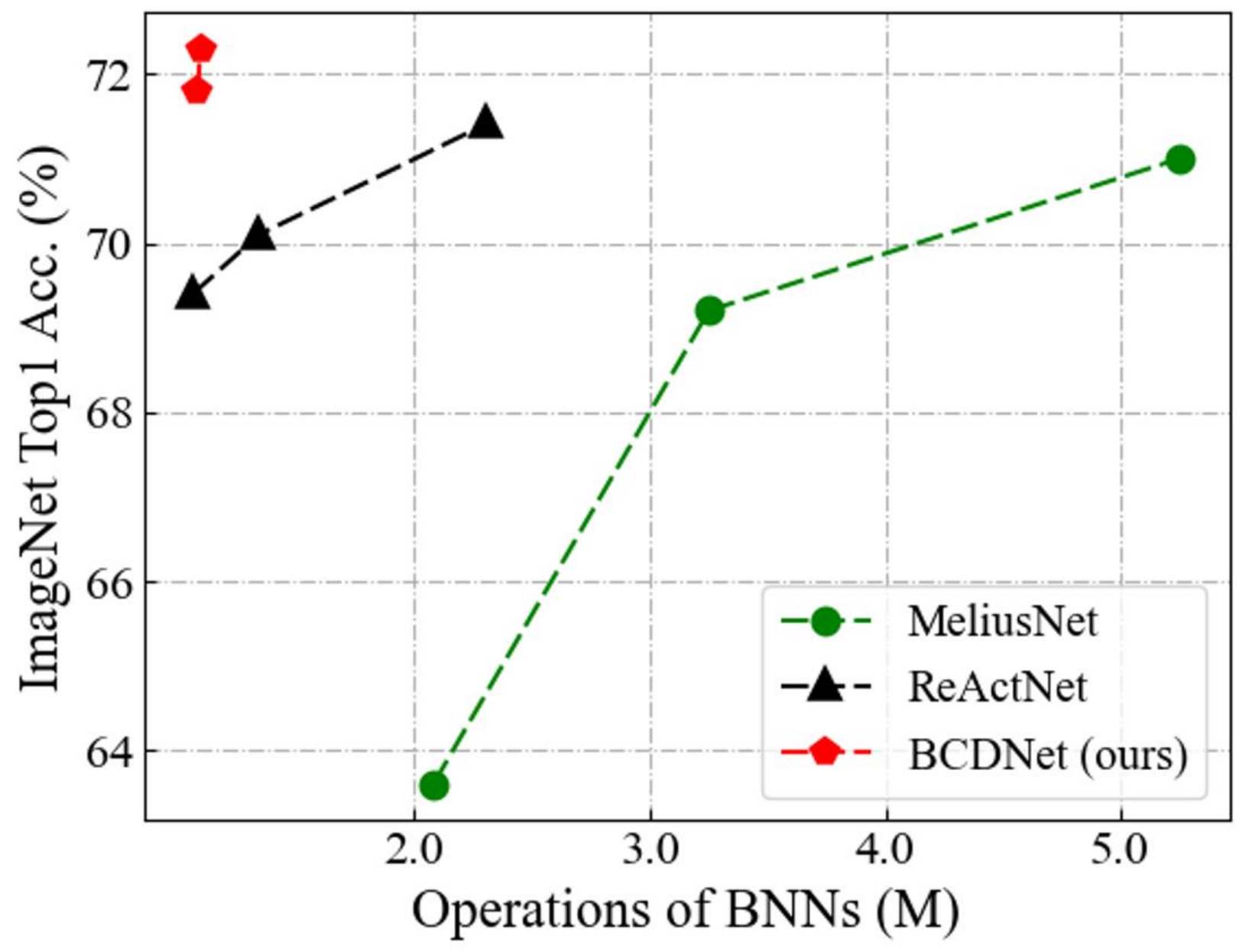
Towards Accurate Binary Neural Networks via Modeling Contextual Dependencies

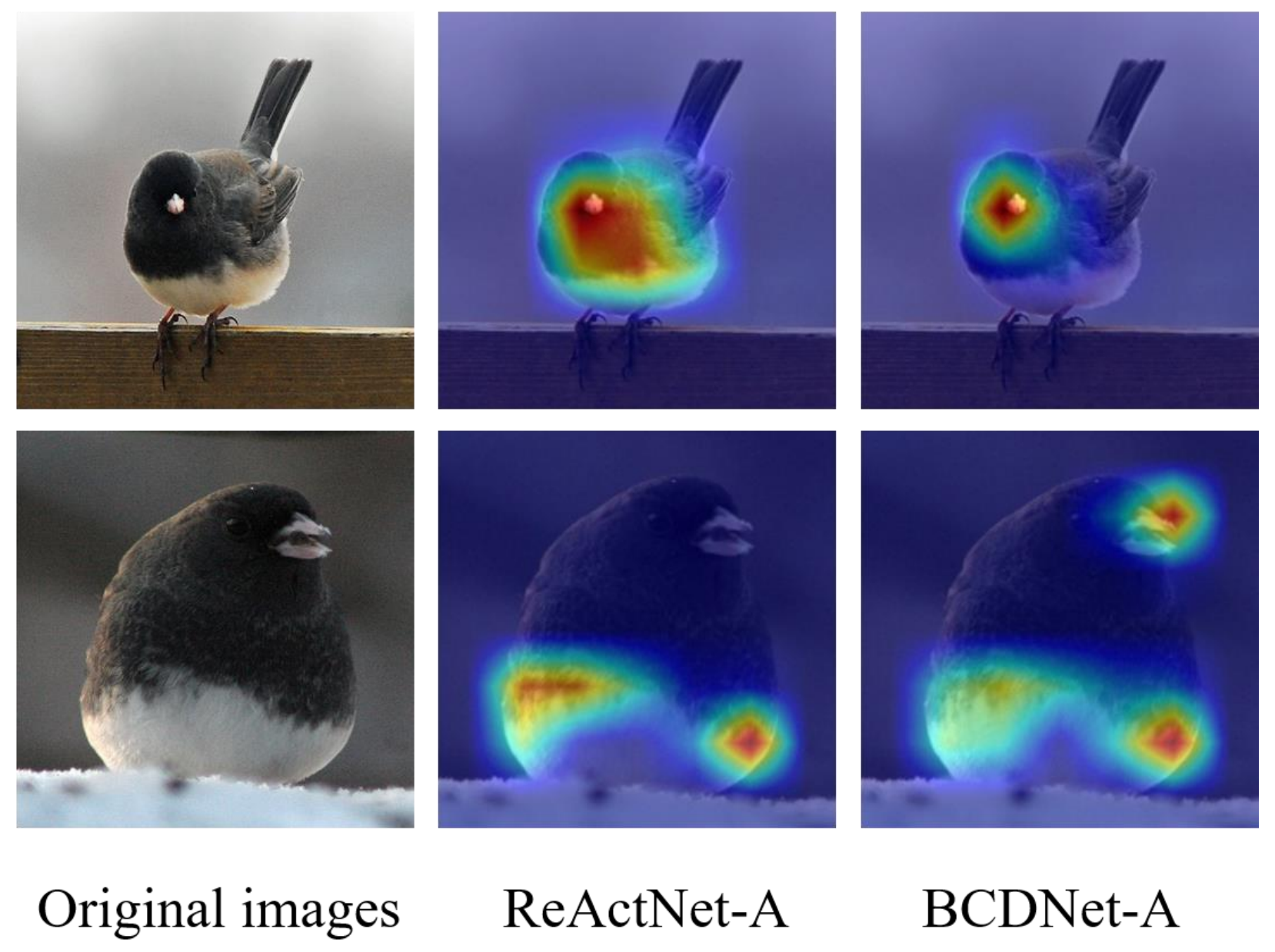
Xingrun Xing[1], Yangguang Li[2], Wei Li[3], Wenrui Ding[1], Yalong Jiang[1]*, Yufeng Wang[1],
Jing Shao[1], Chunlei Liu[1], Xianglong Liu[1]

[1]Beihang University & [2]SenseTime Group & [3]Nanyang Technological University

ECCV TEL AVIV 2022

Main Contributions

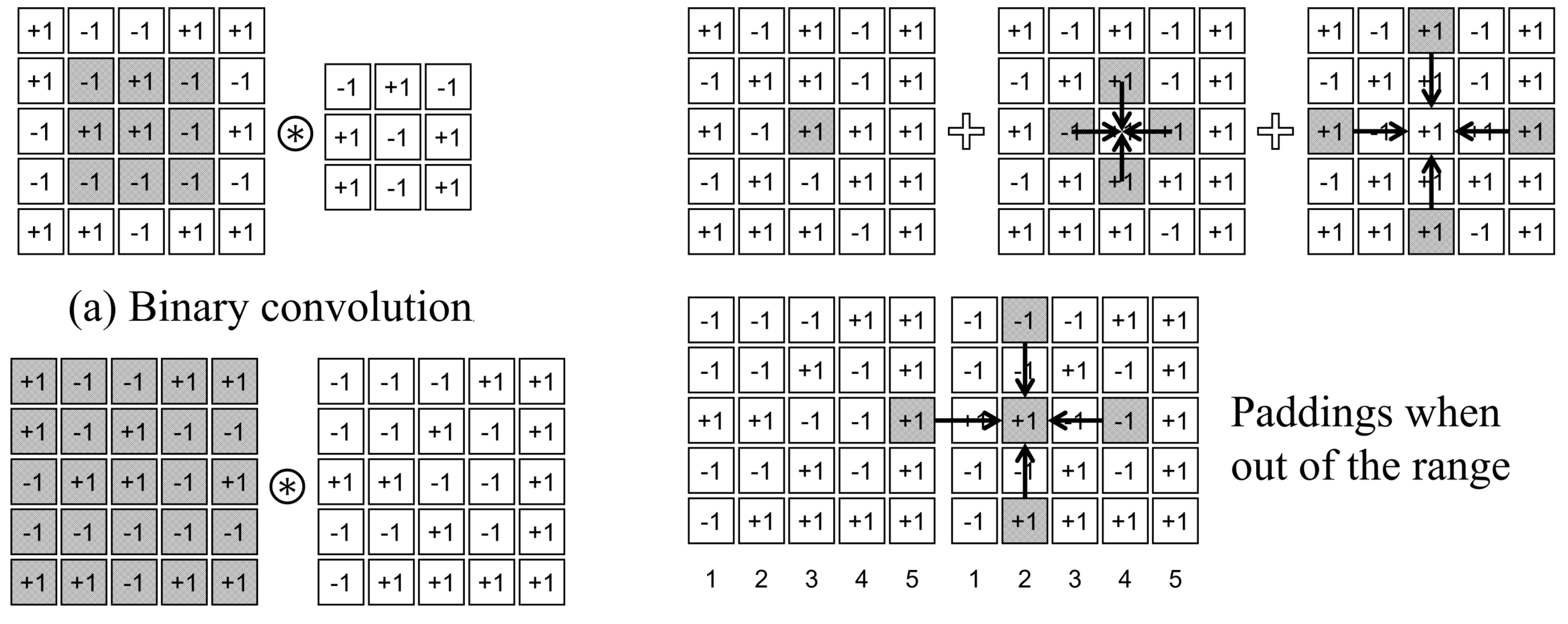




Our contributions:

- We make the first attempt to break local binary convolutions and explore more efficient contextual binary operations.
- We design a long-short range binary MLP module to enhance modeling contextual dependencies in BNNs.
- To our best knowledge, the proposed BCDNet reports the best BNN performance on the large-scale ImageNet dataset currently.

Comparation of binary operations



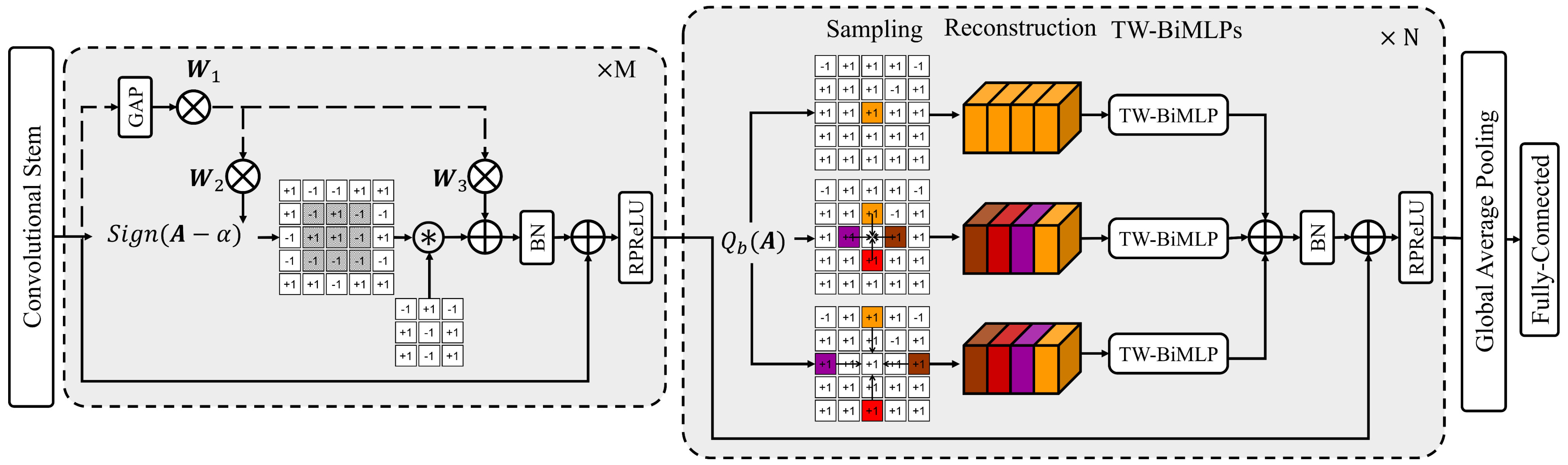
(a) Binary convolution

(b) Binary MLP-Mixer

(c) Our long-short range Sampling

Comparisons of binary operations: (a) The binary convolutions have inductive bias but limited local perceptions; (b) The binary token-mixing MLPs are sharing contextual perceptions but difficult to optimize; (c) Our proposed binary MLPs achieve inductive bias in short-range and explore long-range dependencies concurrently.

Overall Network Architecture



BCDNet is composed of two stages: the binary CNN-embedding stage with M binary convolutional blocks, and the binary MLP stage with N proposed binary MLP blocks. the proposed binary MLP modules are able to plug in previous binary convolutional networks to enhance high level contextual features. First, we replace several binary convolution blocks with our binary MLPs based on ReActNet-A, which we indicate as BCDNet-A. Second, we apply both improved binary convolution and MLP blocks and introduce a BCDNet-B model to further improve performance.

Binary MLP Block

Short-range sampling:

$$\boldsymbol{A}_b^S = Cat\{\boldsymbol{A}_b[0:c/4]_{S(-1,0)}, \boldsymbol{A}_b[c/4:c/2]_{S(1,0)}, \boldsymbol{A}_b[c/2:3c/4]_{S(0,-1)}, \boldsymbol{A}_b[3c/4:c]_{S(0,1)}\},$$

Long-range sampling:

$$\boldsymbol{A}_b^L = Cat\{\boldsymbol{A}_b[0:c/4]_{S(-h/2,0)}, \boldsymbol{A}_b[c/4:c/2]_{S(h/2,0)}, \boldsymbol{A}_b[c/2:3c/4]_{S(0,-w/2)}, \boldsymbol{A}_b[3c/4,c]_{S(0,w/2)}\}.$$

Token-wise binary MLP:

$$\texttt{TW-BiMLP}(\boldsymbol{A}) = \frac{\|\boldsymbol{W}\|_{\ell_1}}{c_{in}} \texttt{popcount}(\boldsymbol{W}_b{}^T \oplus \boldsymbol{A}_b)$$

Backward propagation of binary MLP:

$$Q_b(x) = \begin{cases} -1 & \text{if } x < -1 \\ 2x + x^2 & \text{if } -1 \leqslant x < 0 \\ 2x - x^2 & \text{if } 0 \leqslant x < 1 \\ 1 & \text{otherwise} \end{cases}, \quad \frac{\partial Q_b(x)}{\partial x} = \begin{cases} 2 + 2x & \text{if } -1 \leqslant x < 0 \\ 2 - 2x & \text{if } 0 \leqslant x < 1 \\ 0 & \text{otherwise} \end{cases}$$

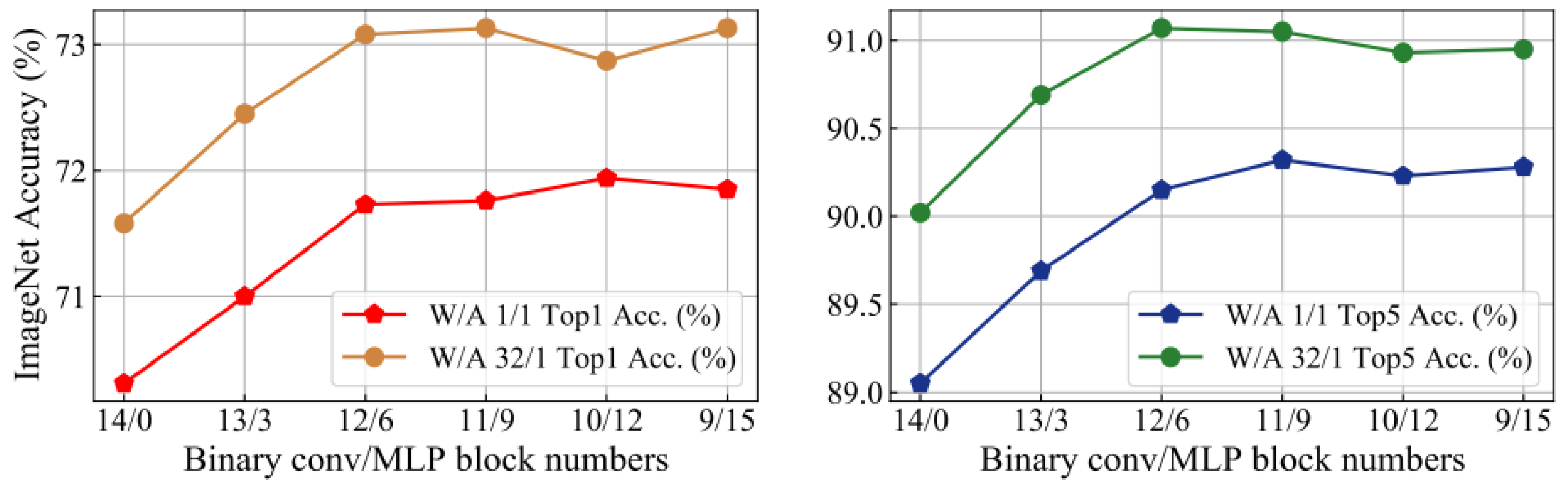
Experiments

Methods	W/A	BOPs ($\times 10^9$)	FLOPs ($\times 10^8$)	OPs ($\times 10^8$)	Top1 (%)	Top5 (%)
Mobile-V1 [8]	32/32	0	5.69	5.69	70.6	–
Mobile-V2 [34]	32/32	0	3.00	3.00	72.0	–
ResNet-18 [7]	32/32	0	18.14	18.14	69.3	89.2
BWN [32]	1/32	–	–	–	60.8	83.0
LQ-Net [46]	1/2	–	–	–	62.6	84.3
DoReFa [50]	2/2	–	–	–	62.6	84.4
SLB [44]	1/8	–	–	–	66.2	86.5
ABC-Net [18]	(1/1)×5	–	–	–	65.0	85.9
Bi-Real-34 [24]	1/1	3.53	1.39	1.93	62.2	83.9
Real-to-Bin [28]	1/1	1.68	1.56	1.83	65.4	86.2
FDA-BNN* [42]	1/1	–	–	–	66.0	86.4
SA-BNN-50 [19]	1/1	–	–	–	68.7	87.4
MeliusNet-22 [1]	1/1	4.62	1.35	2.08	63.6	84.7
MeliusNet-42 [1]	1/1	9.69	1.74	3.25	69.2	88.3
MeliusNet-59 [1]	1/1	18.3	2.45	5.25	71.0	89.7
ReActNet-A [23]	1/1	4.82	0.31 (0.12†)	1.06 (0.87†)	69.4	–
ReActNet-B [23]	1/1	4.69	0.61 (0.44†)	1.34 (1.17†)	70.1	–
ReActNet-C [23]	1/1	4.69	1.57 (1.40†)	2.30 (2.14†)	71.4	–
BCDNet-A	**1/1**	**4.82**	**0.32 (0.12)**	**1.08 (0.87)**	**71.8 (+2.4)**	**90.3**
BCDNet-B	**1/1**	**4.82**	**0.34 (0.14)**	**1.09 (0.89)**	**72.3 (+2.9)**	**90.5**

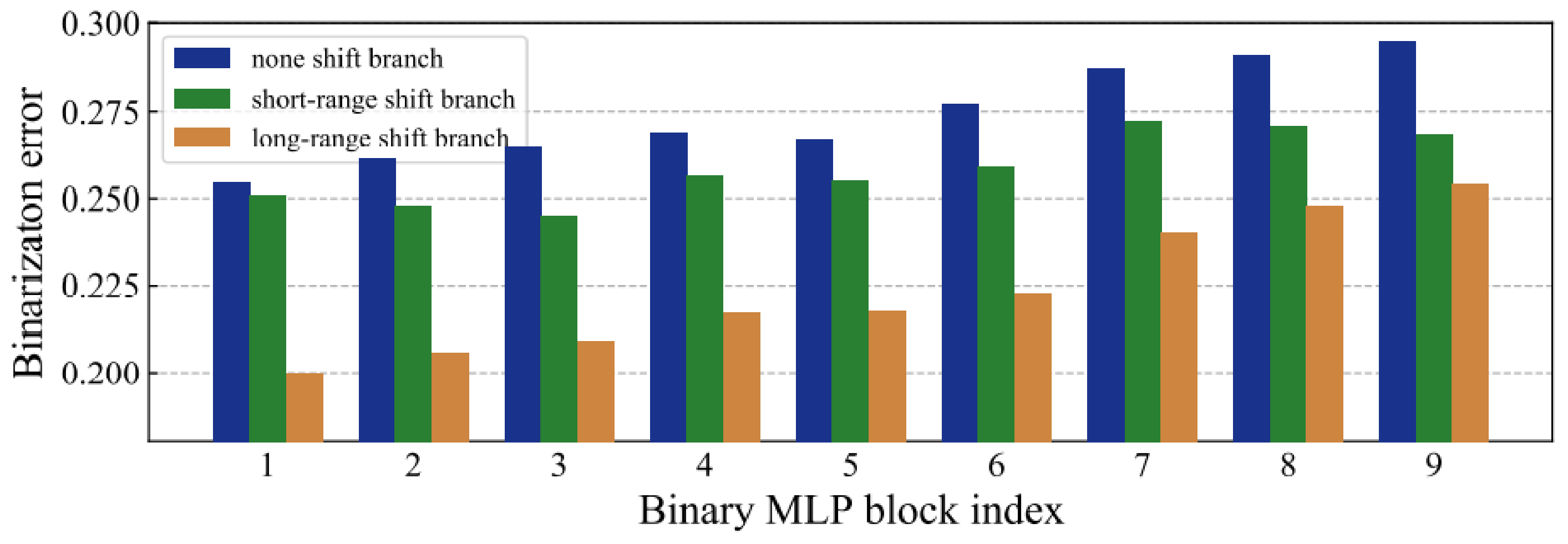
ImageNet Performance on MobileNet-V1 architecture

Methods	W/A	Top1 Acc.	Methods	W/A	Top1 Acc.
ResNet-18	32/32	69.3	RBNN [17]	1/1	59.9
XNOR-Net [32]	1/1	51.2	SA-BNN [19]	1/1	61.7
Bi-Real [24]	1/1	56.4	ReActNet (R18) [23]	1/1	65.5
Real-to-Bin [28]	1/1	65.4	FDA-BNN [42]	1/1	60.2
IR-Net [31]	1/1	58.1	FDA-BNN* [42]	1/1	66.0
SLB [44]	1/1	61.3	**BCDNet-A (R18)**	**1/1**	**66.9**
SLB [44]	1/8	66.2	**BCDNet-B (R18)**	**1/1**	**67.9**

ImageNet Performance on ReesNet-18 architecture



Performance trade-off between binary conv and MLP blocks.



Binarization error of different branches