Network Binarization toward Hardware-friendly Deep Learning

Haotong Qin
Beihang University

Haotong Qin

EDUCATION

Incoming Postdoc	*PBL, ETH Zürich*	*2024–*
Ph.D.	SCSE, Beihang University	2019–Present (2024)
Joint Ph.D.	CVL, ETH Zürich	2022–2023
B.S.	SCSE, Beihang University	2015–2019

RESEARCH INTERESTS

Network binarization and quantization
Efficient neural architecture design
Hardware implementation of compact network

INTERNSHIPS

2021–23	Bytedance AI Lab	Beijing, China	Research Intern
2020	Tencent WXG	Shenzhen, China	Research Intern
2018–19	Microsoft Research Asia	Beijing, China	Research Intern

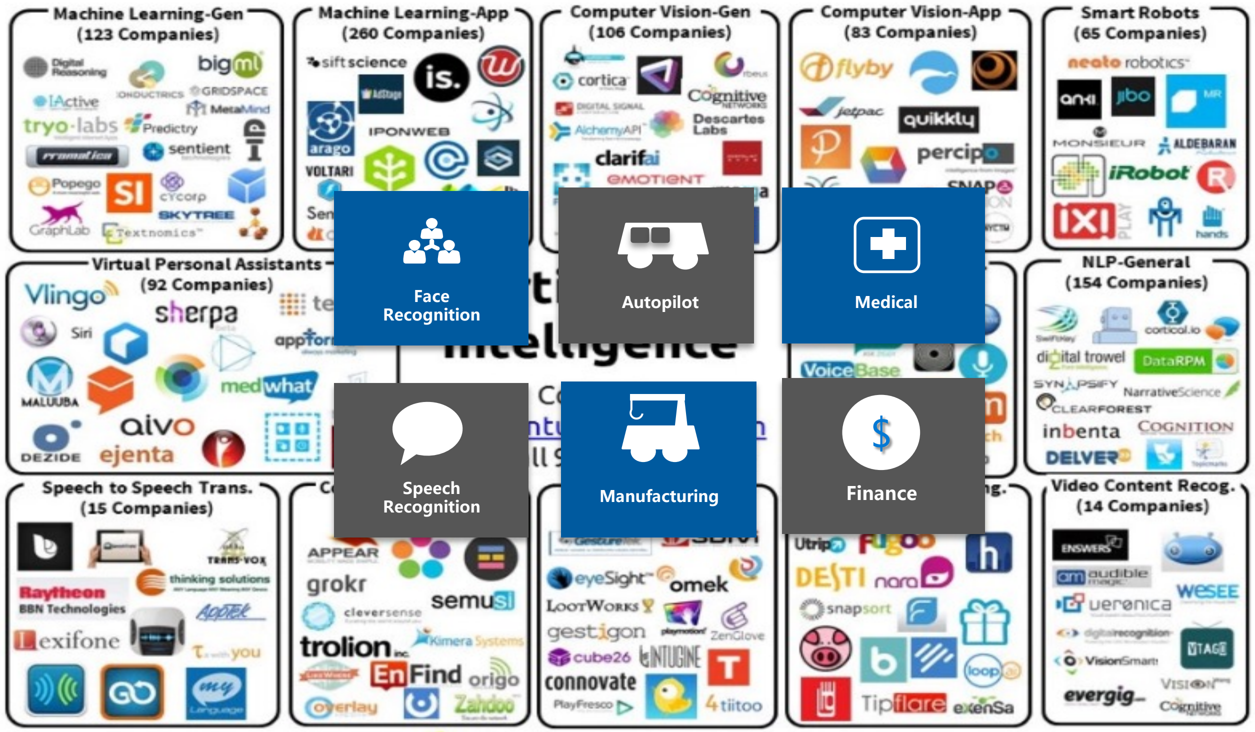
MAIN AWARDS

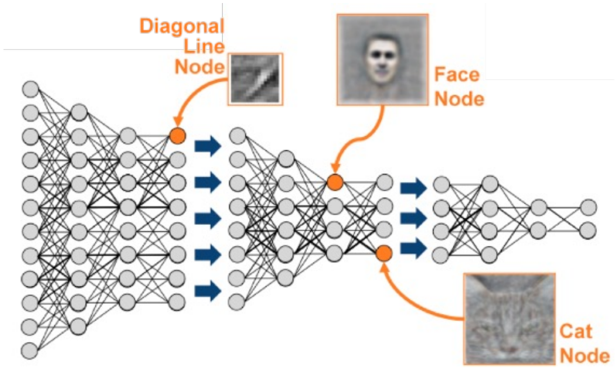
2023	KAUST Rising Stars in AI (28 people worldwide)
2023	DAAD Ainet Fellowship (29 people worldwide)
2022	ByteDance Scholarship (10 people nationwide)
2022	Beihang Top-10 PhD Students Award
2023/21/20	China National Scholarship (3 times)

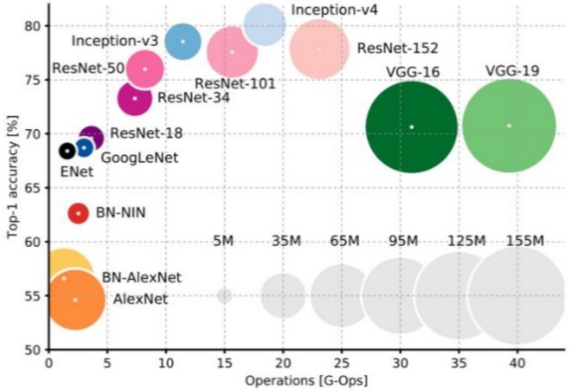
Background

- **Vision**
  - Classification
  - Detection
  - Localization
  - Segmentation
- **Language**
  - Information retrieval
  - Relation extraction
  - Machine translation
- **Speech**
  - Language understanding
  - Speech recognition

…

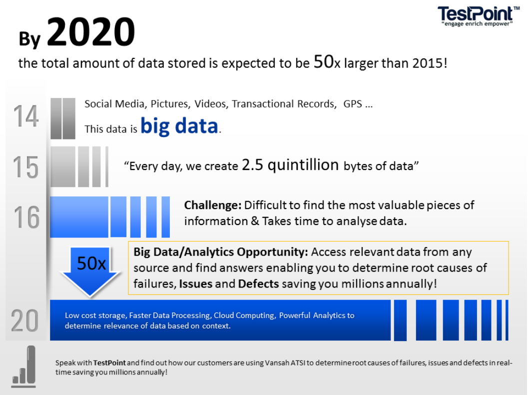


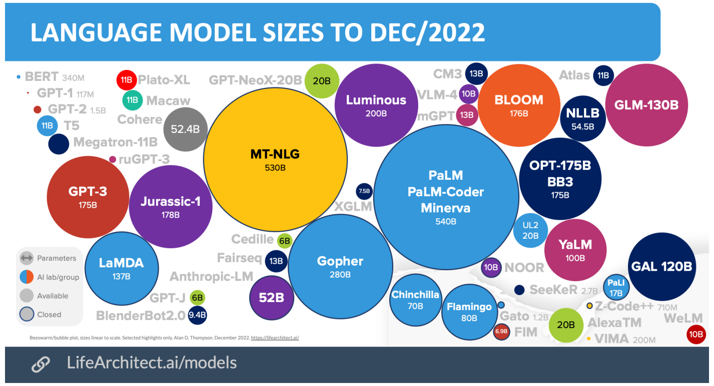


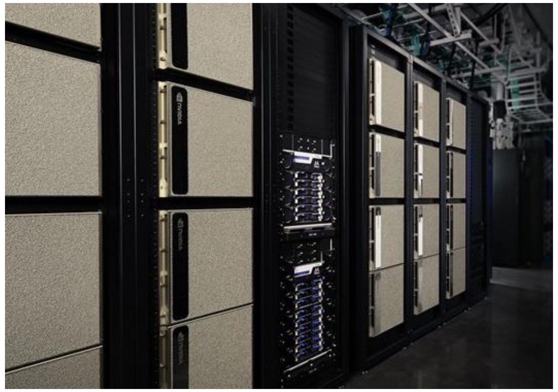


Background

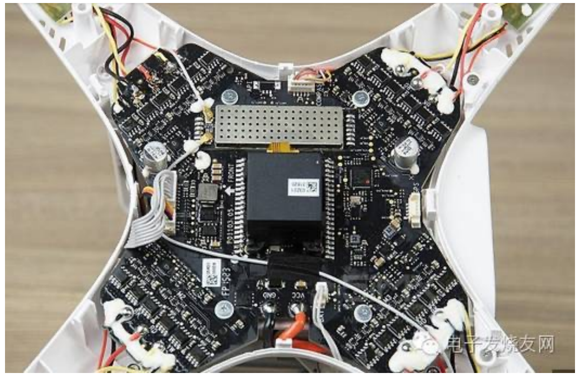
bigger data
and
larger model

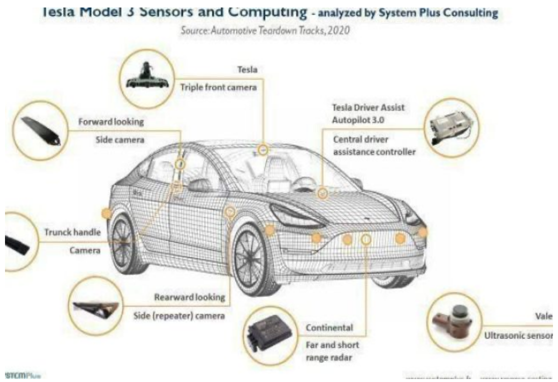






diverse usage
and
limited resources





Network Quantization

$$\mathbb{R} \longrightarrow \mathbb{I} \longrightarrow \mathbb{B} \in \{-1, +1\}$$

32-bit 8-bit 1-bit

Network Quantization

$$\mathbb{R} \longrightarrow \mathbb{I} \longrightarrow \mathbb{B} \in \{-1, +1\}$$

32-bit 8-bit 1-bit

Multi-bit Quantization (Towards Accurate Prediction)

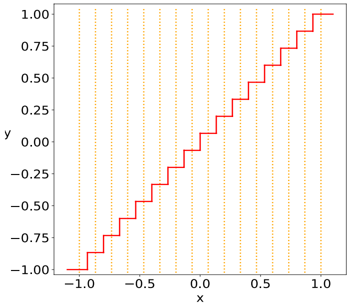
Quantization Function:

$$x_{int} = round(\frac{x}{\Delta}) + z$$

$$x_Q = clamp(0, N_{levels} - 1, x_{int})$$

De-quantization Function:

$$x_{float} = (x_Q - z)\Delta$$



Quantization
(Integer Computation)

Network Binarization

$$\mathbb{R} \longrightarrow \mathbb{I} \longrightarrow \mathbb{B} \in \{-1, +1\}$$

32-bit 8-bit 1-bit

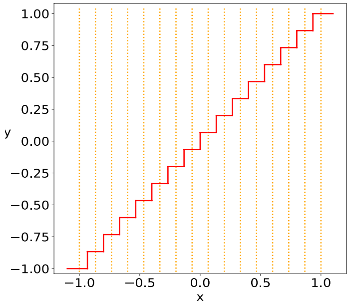
Multi-bit Quantization (Towards Accurate Prediction)

Quantization Function:

$$x_{int} = round(\frac{x}{\Delta}) + z$$
$$x_Q = clamp(0, N_{levels} - 1, x_{int})$$

De-quantization Function:

$$x_{float} = (x_Q - z)\Delta$$



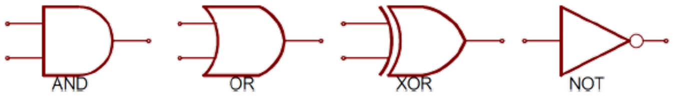
Quantization
(Integer Computation)

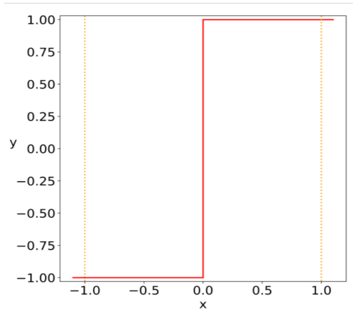
1-bit Quantization (Towards Efficient Inference)

Quantization Function:

$$Q_B(x) = \operatorname{sgn}(x) = \begin{cases} +1, & \text{if } x \geq 0, \\ -1, & \text{otherwise.} \end{cases}$$

Bitwise Instructions：





Binarization
(Bitwise Computation)

Network Binarization

$$\mathbb{R} \longrightarrow \mathbb{I} \longrightarrow \mathbb{B} \in \{-1, +1\}$$

32-bit 8-bit 1-bit

Multi-bit Quantization (Towards Accurate Prediction)

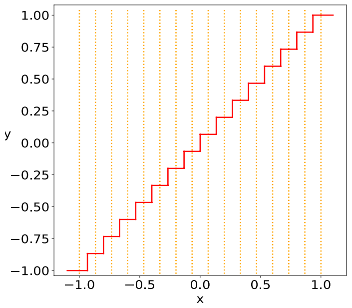
Quantization Function:

$$x_{int} = round(\frac{x}{\Delta}) + z$$

$$x_Q = clamp(0, N_{levels} - 1, x_{int})$$

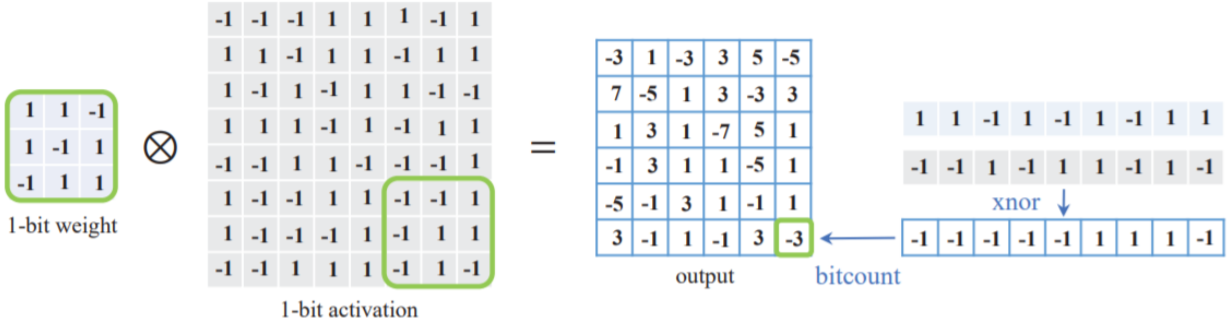
De-quantization Function:

$$x_{float} = (x_Q - z)\Delta$$



Quantization
(Integer Computation)

1-bit Quantization (Towards Efficient Inference)



Binarization
(Bitwise Computation)

Network Binarization

Full-Precision Neural Networks

Massive Parameters

Complex Computation

High Power Consumption

Network Binarization

Full-Precision Neural Networks

Massive Parameters

Complex Computation

High Power Consumption

Binarized Neural Networks

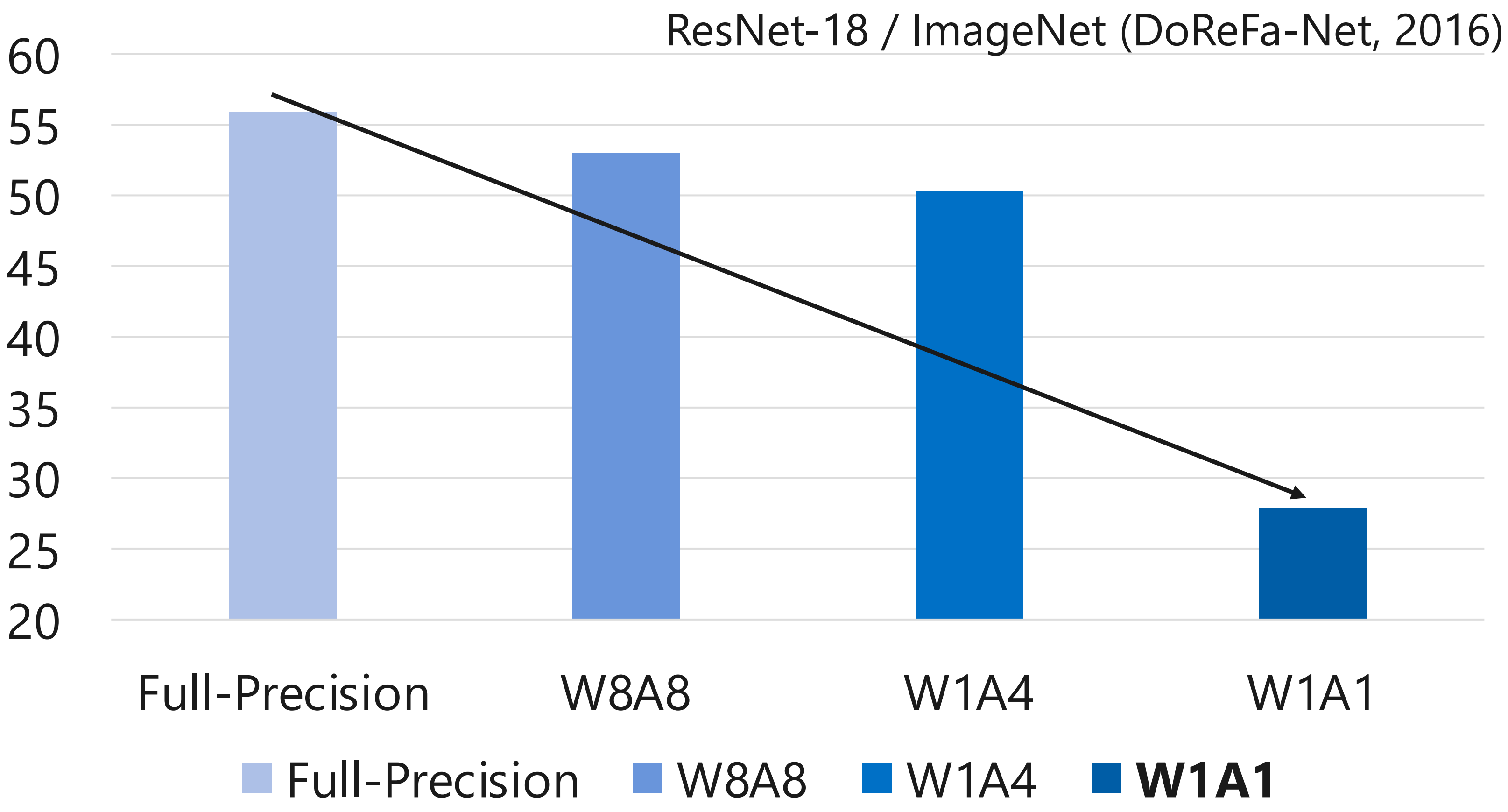
Binarized Parameters

Efficient Instructions

Low Power Consumption

Network Binarization

Goal: Accurate Extreme-Low Bit Quantization (Binarization)



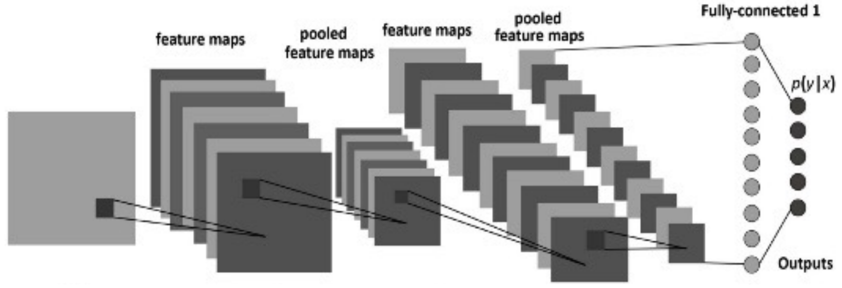
Accuracy of the binarized neural network has dropped seriously

Network Binarization

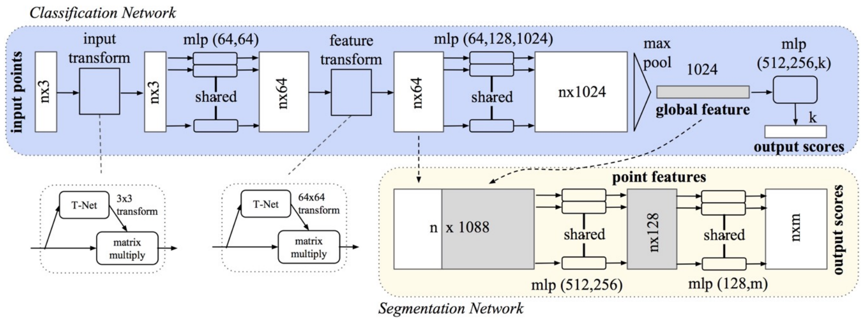
Goal: **Accurate Extreme-Low Bit Quantization (Binarization)**

The smallest storage, the fastest computation, severe accuracy loss

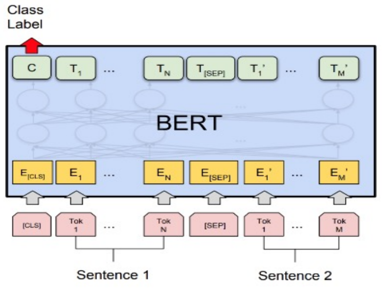
Observation: structure is the key factors affecting the accuracy



CNN (ResNet, VGG, ...)



MLP (PointNet, MLP-Mixer, ...)



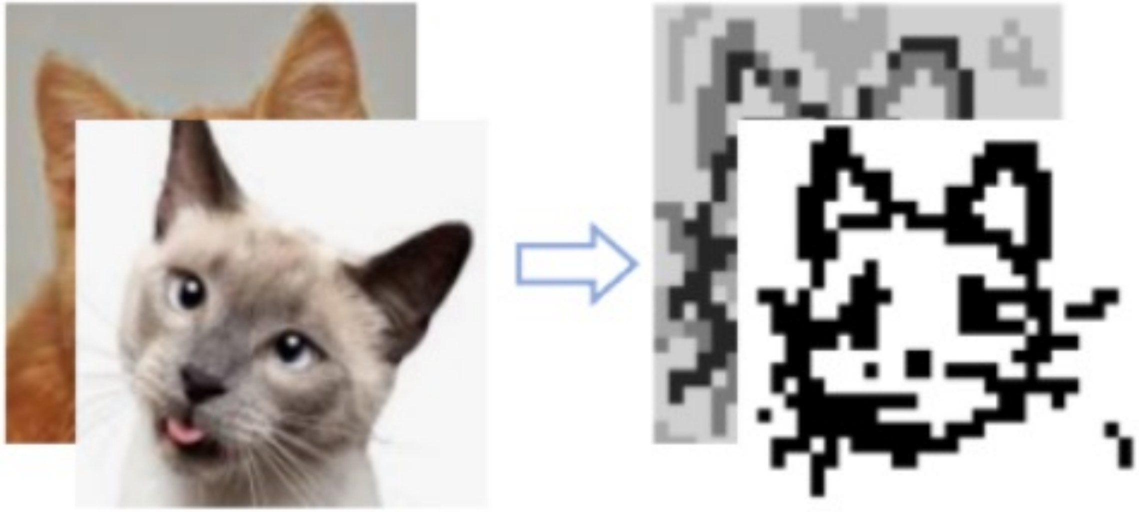
Transformer (BERT, ViT,...)

Accurate 1-bit binarization with typical architectures

Generic Binarization (Classification and Detection)

Effects of BNN in the Forward and Backward Propagation

limited representation



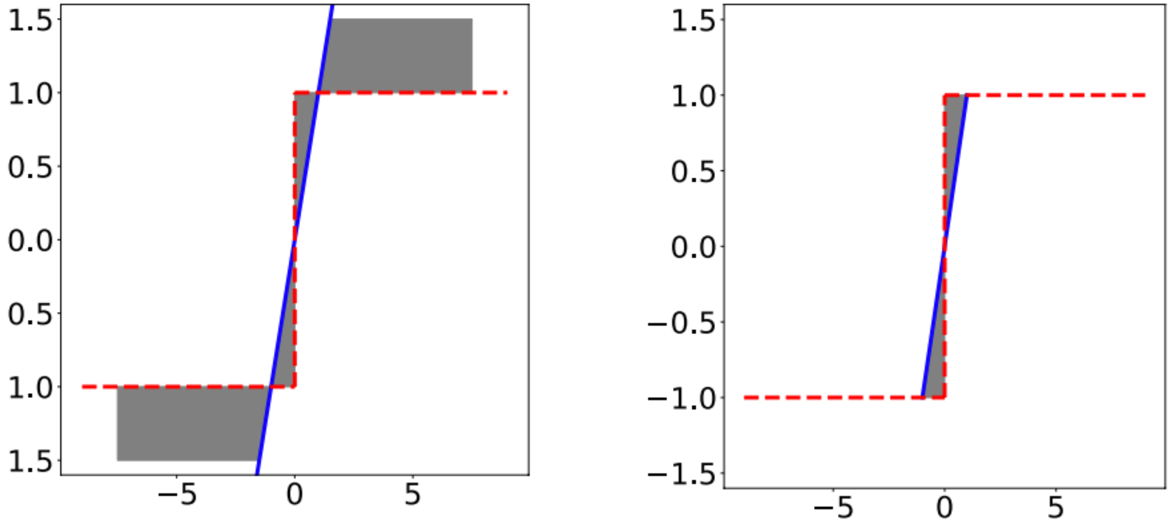
$$\mathbf{B_x} = \mathtt{sign}(\mathbf{x}) = \begin{cases} +1, & if\ \mathbf{x} \geq 0 \\ -1, & otherwise. \end{cases}$$

Forward and Backward Information Retention for Accurate Binary Neural Networks. Haotong Qin, et al. ICVPR 2020.
Distribution-sensitive Information Retention for Accurate Binary Neural Network. Haotong Qin, et al. IJCV 2022.

Generic Binarization (Classification and Detection)

Effects of BNN in the Forward and Backward Propagation

gradient mismatch



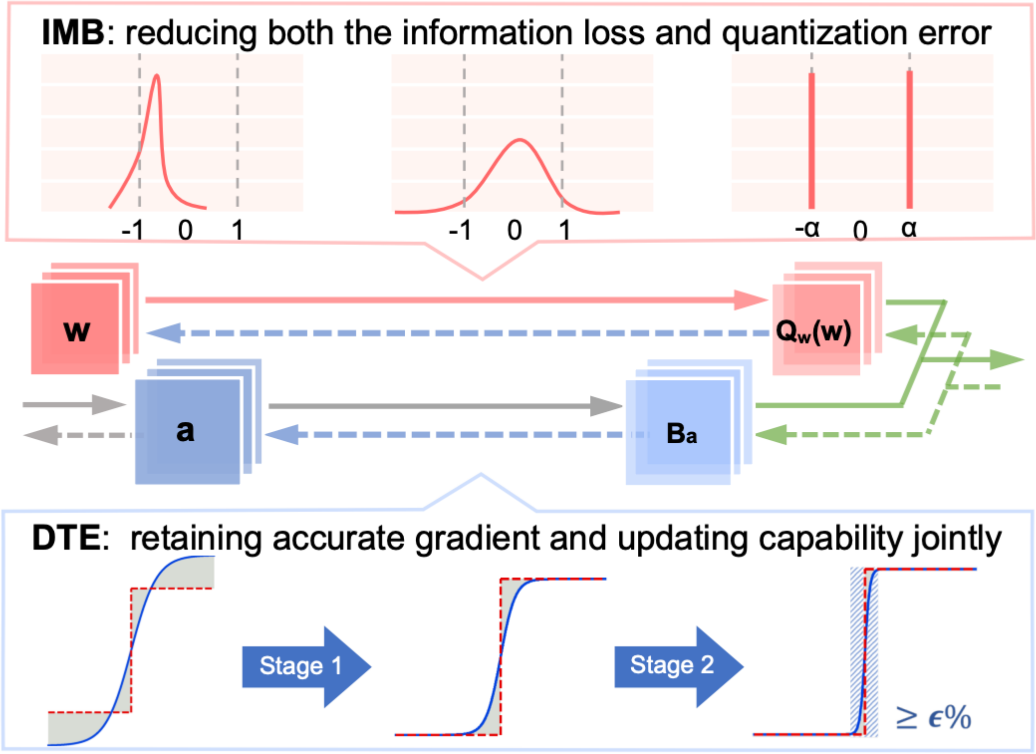
$$\texttt{Identity}: y = x$$

$$\texttt{Clip}: y = \texttt{Hardtanh}(x)$$

Forward and Backward Information Retention for Accurate Binary Neural Networks. Haotong Qin, et al. ICVPR 2020.
Distribution-sensitive Information Retention for Accurate Binary Neural Network. Haotong Qin, et al. IJCV 2022.

Generic Binarization (Classification and Detection)

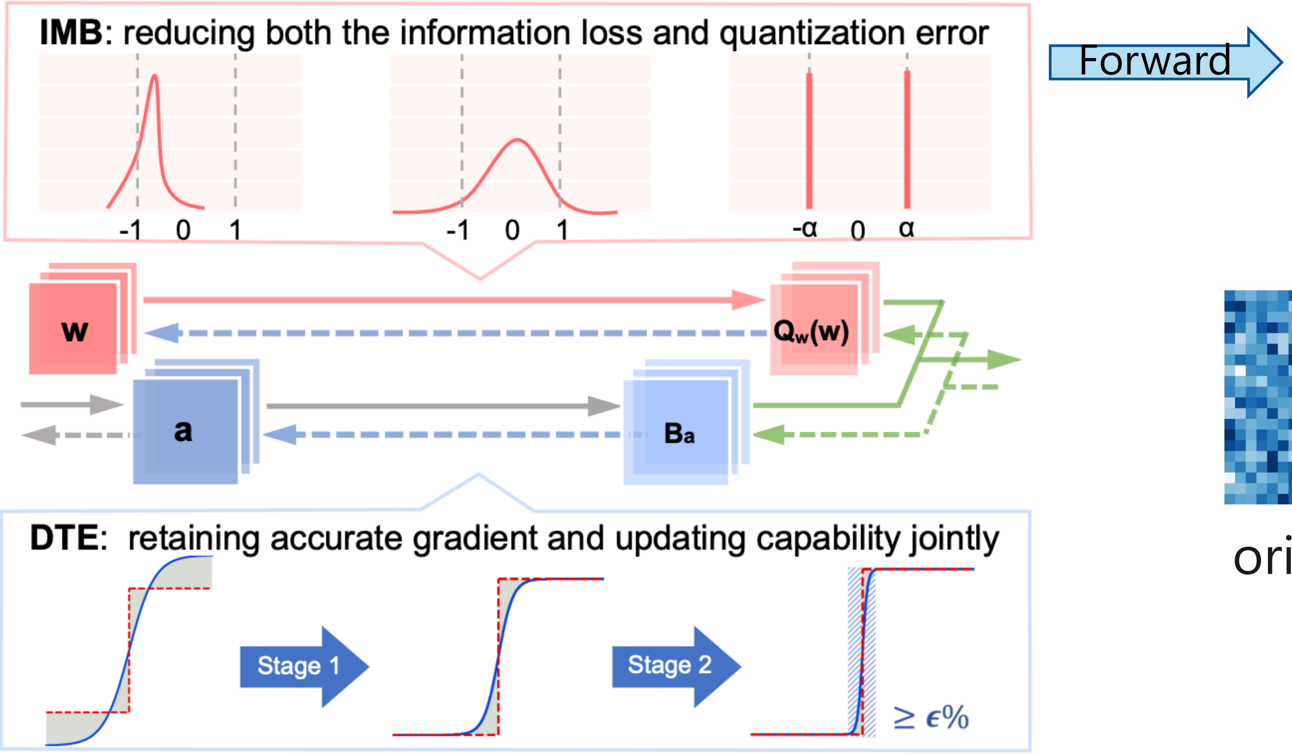
Distribution-sensitive Information Retention (DIR-Net)



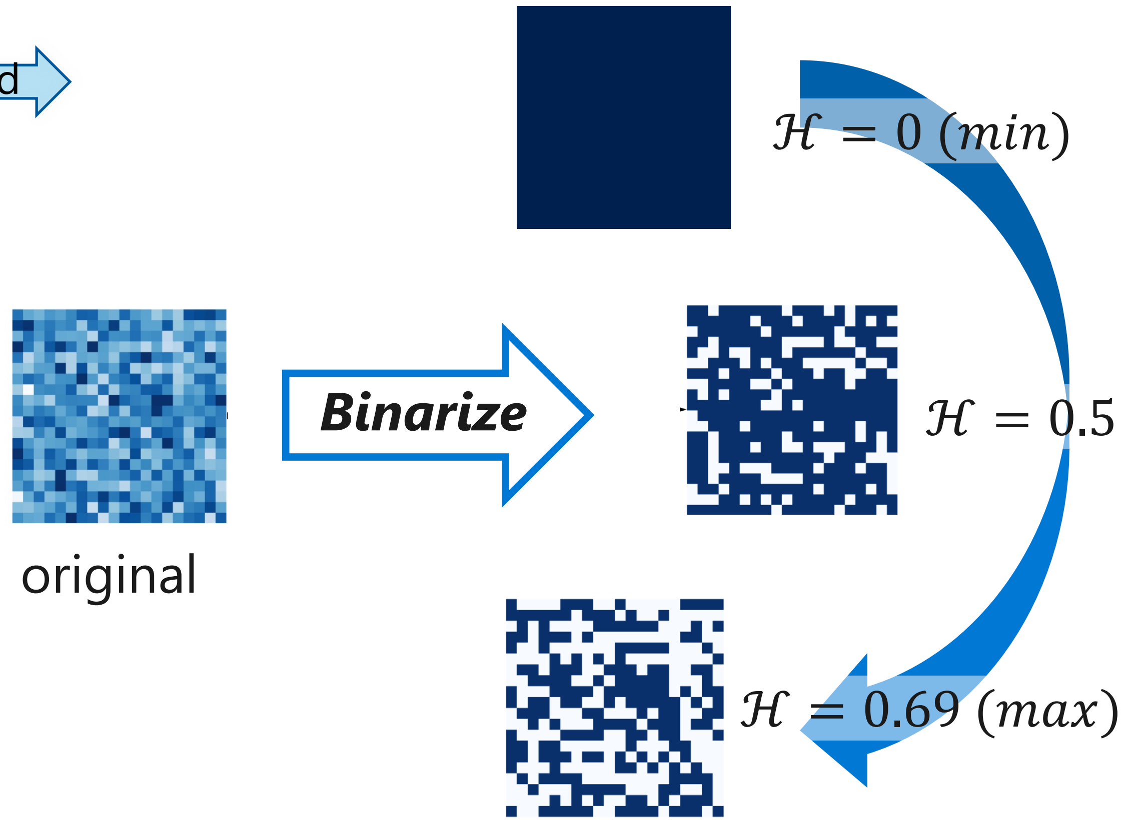
Forward and Backward Information Retention for Accurate Binary Neural Networks. Haotong Qin, et al. ICVPR 2020.
Distribution-sensitive Information Retention for Accurate Binary Neural Network. Haotong Qin, et al. IJCV 2022.

Generic Binarization (Classification and Detection)

Distribution-sensitive Information Retention (DIR-Net)



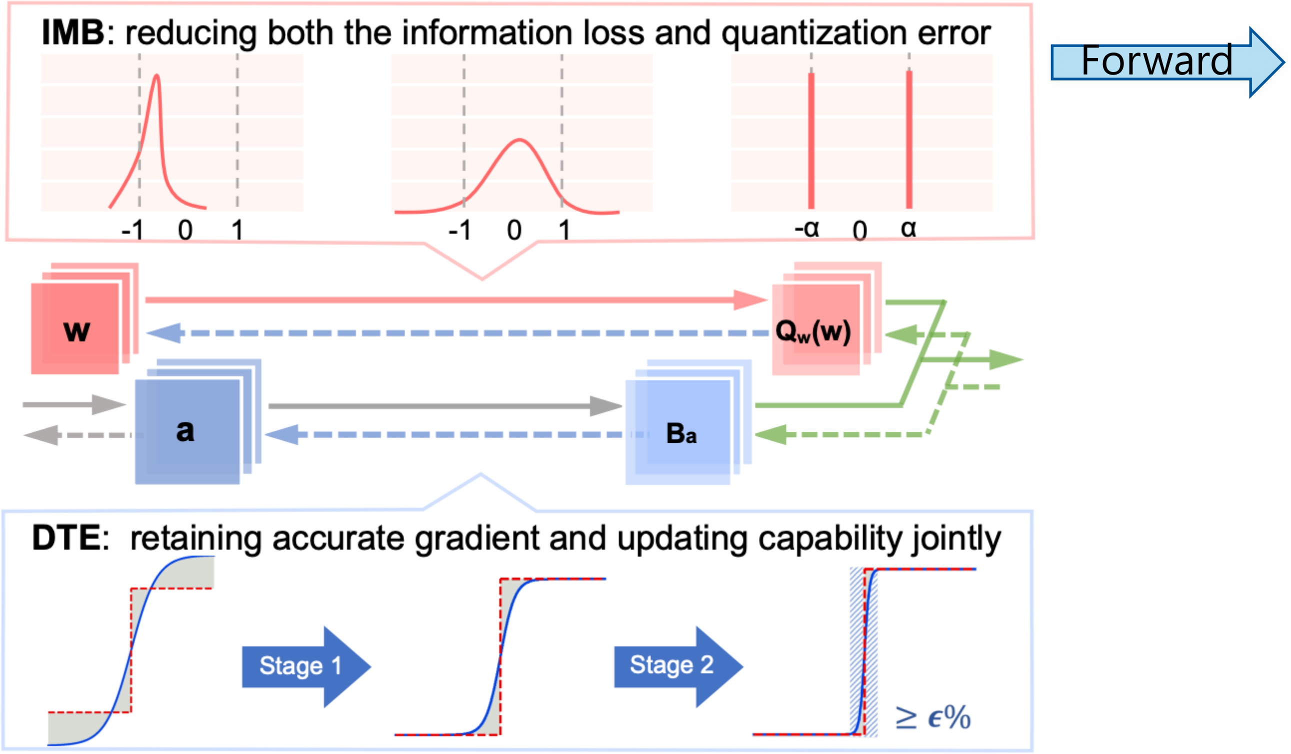
Maximizing the information entropy



Forward and Backward Information Retention for Accurate Binary Neural Networks. Haotong Qin, et al. ICVPR 2020.
Distribution-sensitive Information Retention for Accurate Binary Neural Network. Haotong Qin, et al. IJCV 2022.

Generic Binarization (Classification and Detection)

Distribution-sensitive Information Retention (DIR-Net)

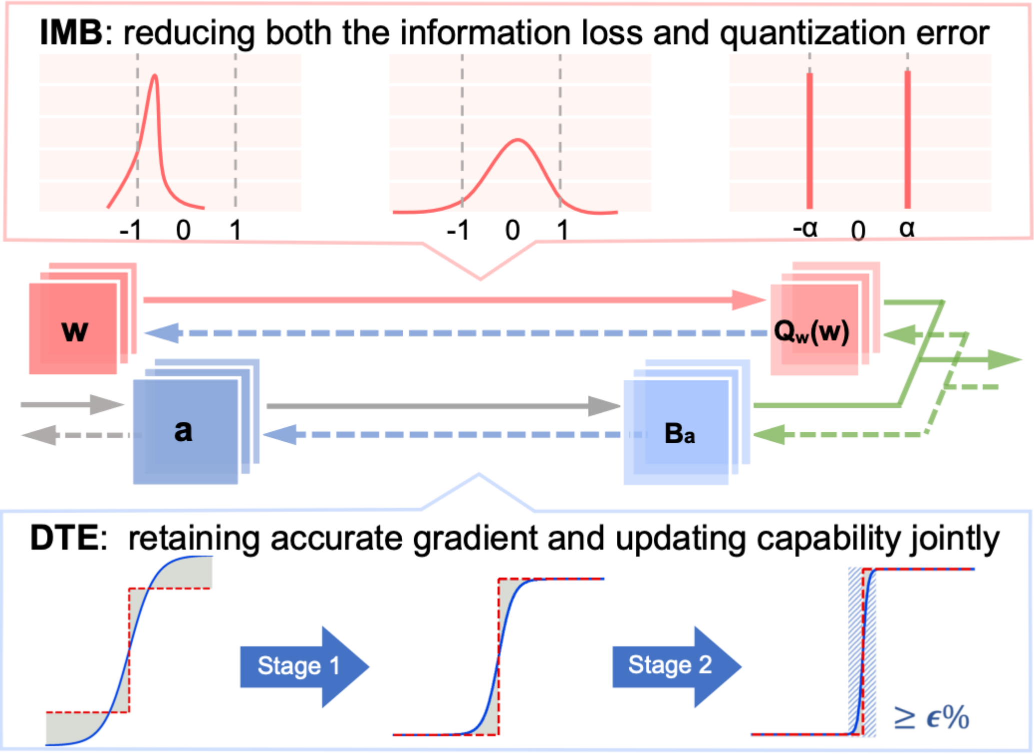


Maximizing the information entropy

Forward and Backward Information Retention for Accurate Binary Neural Networks. Haotong Qin, et al. ICVPR 2020.
Distribution-sensitive Information Retention for Accurate Binary Neural Network. Haotong Qin, et al. IJCV 2022.

Generic Binarization (Classification and Detection)

Distribution-sensitive Information Retention (DIR-Net)

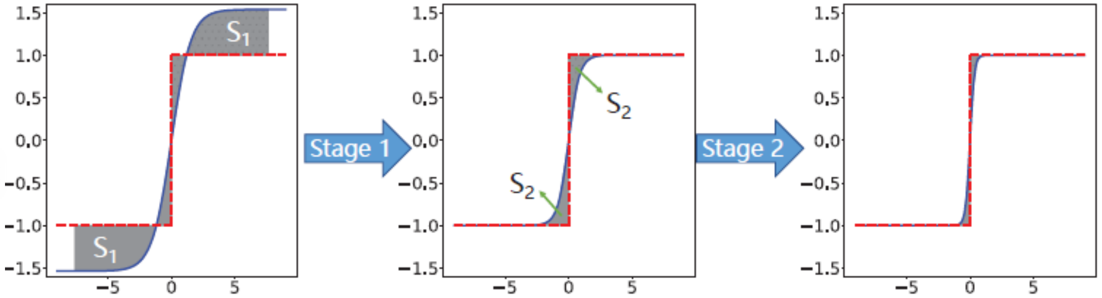


Forward

Backward

Maximizing the information entropy

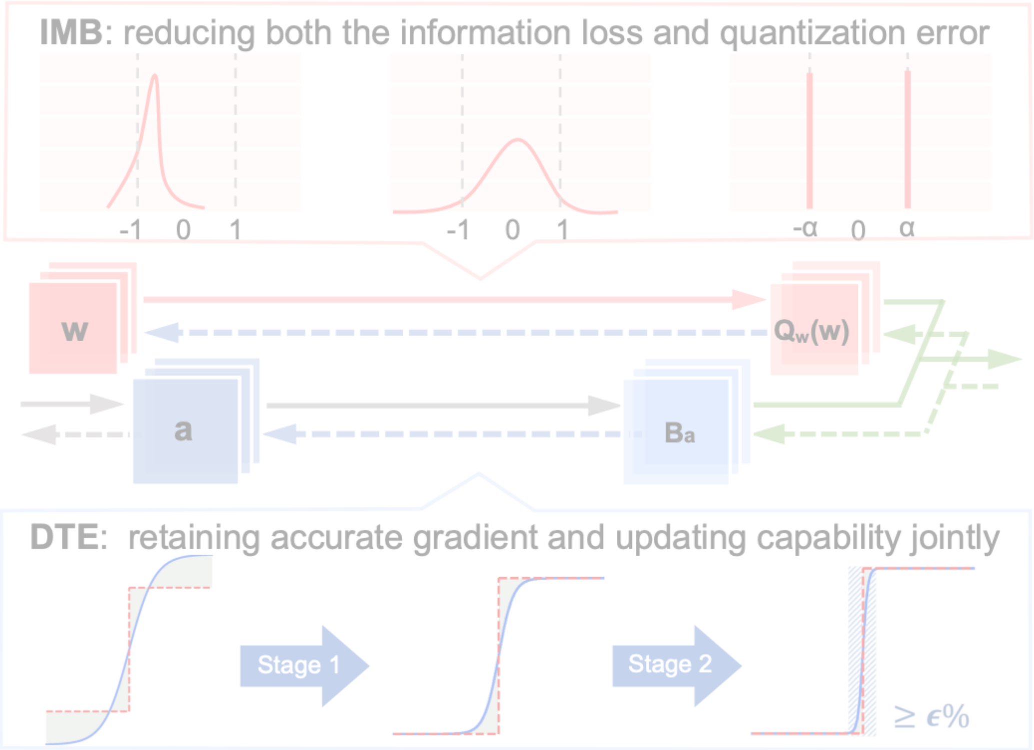
Changing the shape of estimator



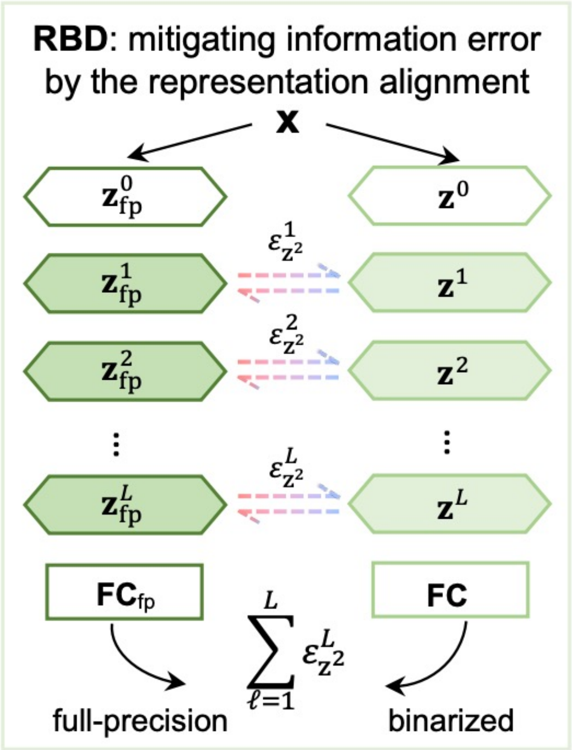
Forward and Backward Information Retention for Accurate Binary Neural Networks. Haotong Qin, et al. ICVPR 2020.
Distribution-sensitive Information Retention for Accurate Binary Neural Network. Haotong Qin, et al. IJCV 2022.

Generic Binarization (Classification and Detection)

Distribution-sensitive Information Retention (DIR-Net)



external:
binarization-aware
distillation



Forward and Backward Information Retention for Accurate Binary Neural Networks. Haotong Qin, et al. ICVPR 2020.
Distribution-sensitive Information Retention for Accurate Binary Neural Network. Haotong Qin, et al. IJCV 2022.

Generic Binarization (Classification and Detection)

Performance

Normal Backbone (ResNet-18):

ImageNet 66.5% Top-1

ResNet-18	Full-Precision	32/32	[illegible]	89.2
	[illegible]	[illegible]	[illegible]	67.6
	XNOR	1/1	51.2	73.2
	[illegible]	[illegible]	53.0	72.6
	[illegible]	[illegible]	53.4	–
	Bi-Real	1/1	56.4	79.5
	XNOR++	1/1	57.1	79.9
	PCNN	1/1	57.3	80.0
	IR-Net	1/1	58.1	80.0
	BONN	1/1	58.3	81.6
	Si-BNN	1/1	59.7	81.8
	Real-to-Bin	1/1	65.4	86.2
	ReActNet	1/1	65.9	–
	DIR-Net[a] (ours)	1/1	60.4	81.9
	DIR-Net[b] (ours)	1/1	**66.5**$_{\pm 0.10}$	**87.1**

Search-based Backbone (DARTS):

ImageNet 65.6% Top-1

DARTS	[illegible]	[illegible]	[illegible]	91.3
	[illegible]	[illegible]	[illegible]	76.6
	[illegible]	[illegible]	[illegible]	83.8
	[illegible]	[illegible]	[illegible]	84.2
	ReActNet	1/1	65.1	86.4
	DIR-Net[a] (ours)	1/1	63.3	85.1
	DIR-Net[b] (ours)	1/1	**65.6**$_{\pm 0.12}$	**87.2**

Detector (SSD):

COCO 67.1% mAP

Framework (Backbone)	[illegible]	[illegible]	Bit-width (W/A)	mAP (%)
SSD300 (VGG-16)	[illegible]	[illegible]	32/32	72.4
		[illegible]	1/1	42.0
		[illegible]	1/1	50.2
		Bi-Real	1/1	63.8
		BiDet	1/1	66.0
		DIR-Net	1/1	**67.1**$_{\pm 0.13}$
Faster R-CNN (ResNet-18)	600 × 1000	Full-Precision	1/1	74.5
		BNN	1/1	35.6
		XNOR	1/1	48.4
		Bi-Real	1/1	58.2
		BiDet	1/1	59.5
		DIR-Net	1/1	**60.4**$_{\pm 0.07}$

Lightweight Backbone (EfficientNet):

ImageNet 64.8% Top-1

Topology	Method	Bit-width (W/A)	Top-1(%)	Top-5(%)
Effi[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
	Bi-Real	1/1	58.7	81.3
	[illegible]	[illegible]	[illegible]	82.6
	ReActNet	1/1	[illegible]	85.1
	DIR-Net[a] (ours)	1/1	63.0	84.8
	DIR-Net[b] (ours)	1/1	**64.8**$_{\pm 0.09}$	**86.2**

Forward and Backward Information Retention for Accurate Binary Neural Networks. Haotong Qin, et al. ICVPR 2020.
Distribution-sensitive Information Retention for Accurate Binary Neural Network. Haotong Qin, et al. IJCV 2022.

Generic Binarization (Classification and Detection)

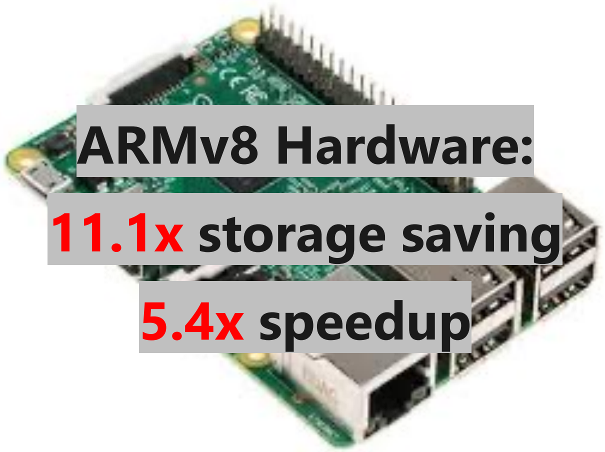
Performance

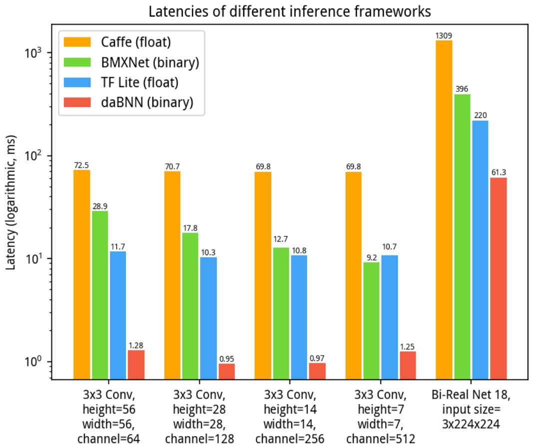
dabnn Public

Deployment on Edge Hardware by daBNN

Method	Bit-width (W/A)	Size (Mb)	Time (ms)
Full-precision	32/32	46.77	1418.94
NCNN	8/8	–	935.51
DSQ	2/2	–	551.22
DIR-Net (w/o scalars)	1/1	**4.20**	**252.16**
DIR-Net (ours)	1/1	**4.21**	**261.98**

Bold values indicates that the bolded metrics of our DIR-Net are higher than metrics of other binarization methods





Forward and Backward Information Retention for Accurate Binary Neural Networks. Haotong Qin, et al. ICVPR 2020.
Distribution-sensitive Information Retention for Accurate Binary Neural Network. Haotong Qin, et al. IJCV 2022.

CNN Binarization (Video Matting)

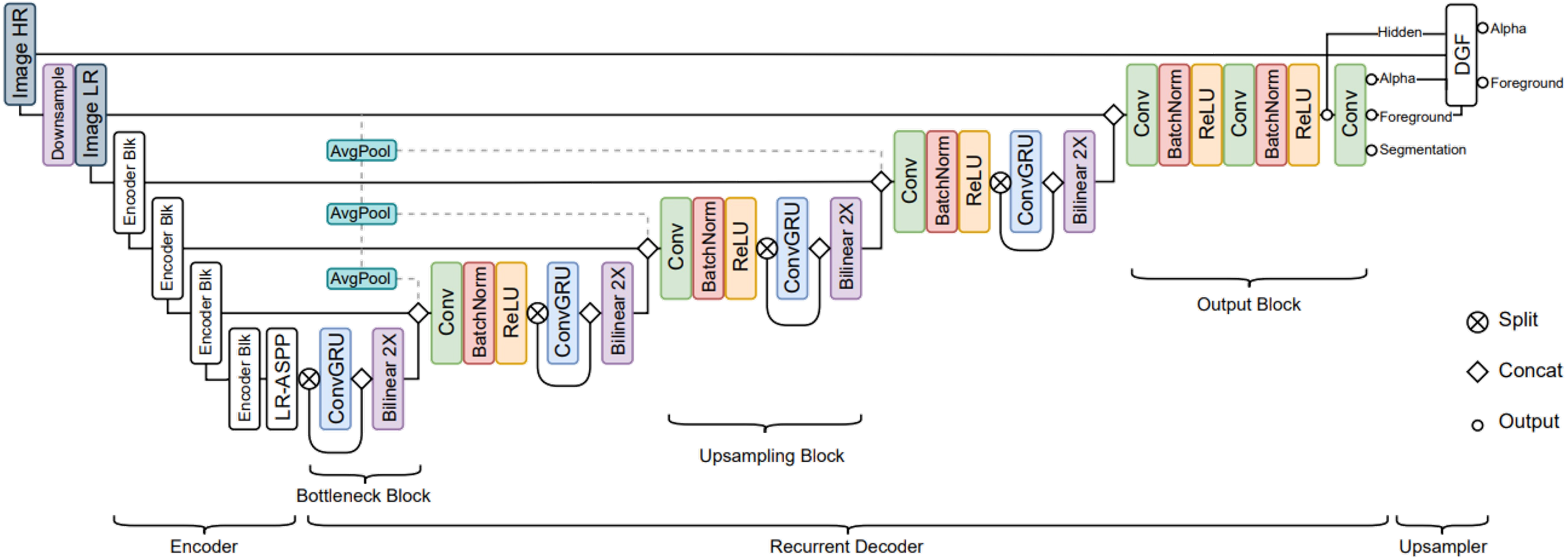
Video Matting on Edge

BiMatting: Efficient Video Matting via Binarization. Haotong Qin, et al. NeurIPS 2022.

CNN Binarization (Video Matting)

Direct Binarization Baseline

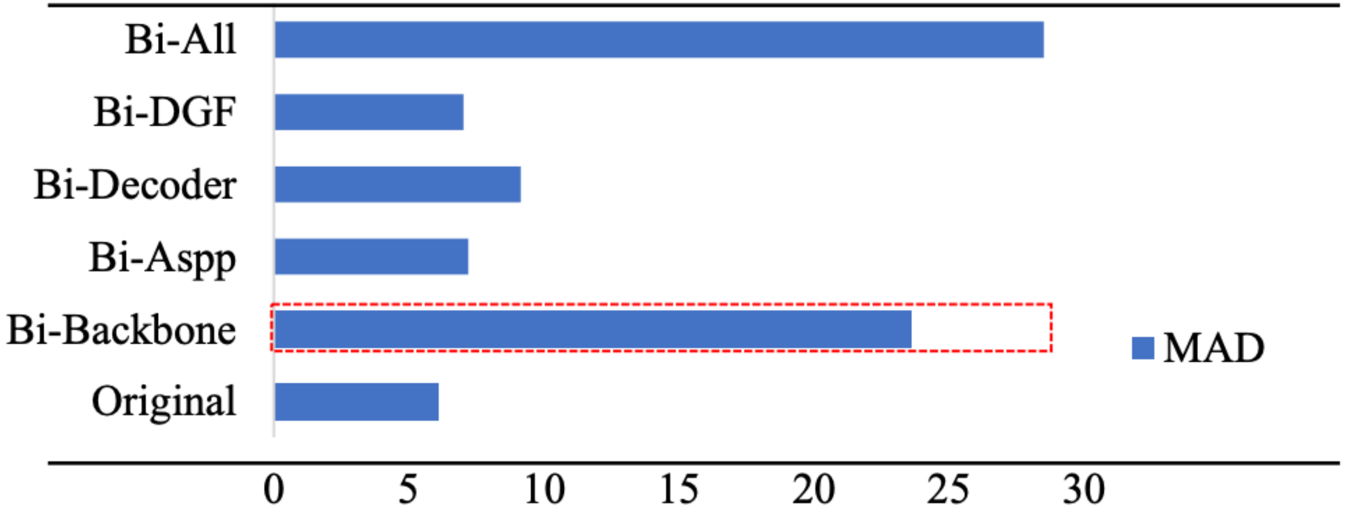
$$I = \alpha F + (1 - \alpha) B.$$



BiMatting: Efficient Video Matting via Binarization. Haotong Qin, et al. NeurIPS 2022.

CNN Binarization (Video Matting)

Direct Binarization Baseline: Accuracy Bottleneck



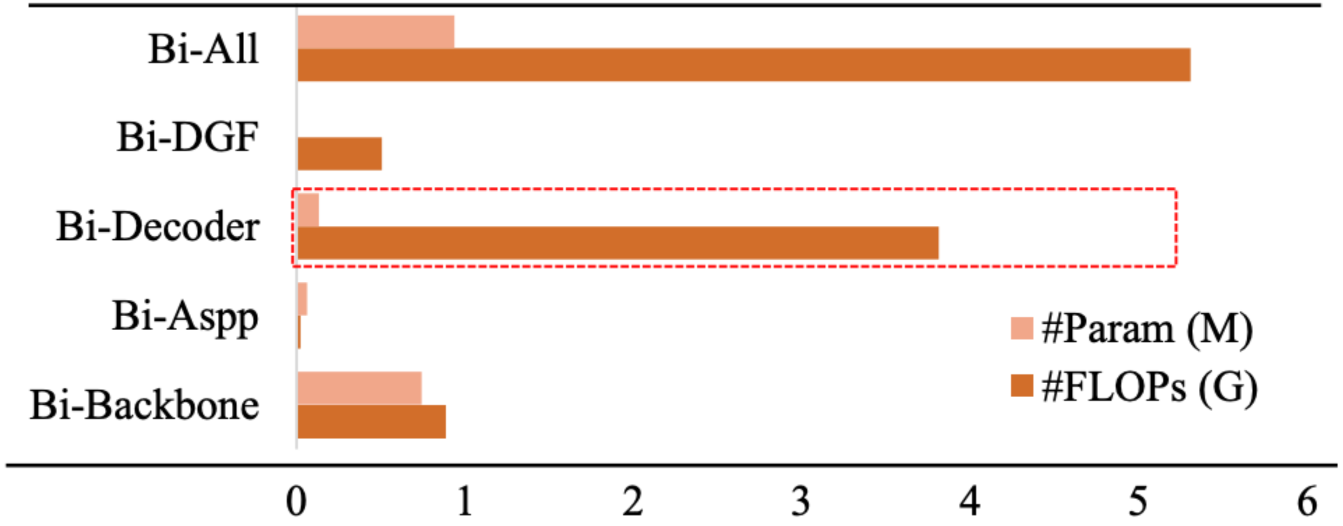
(a) Accuracy perspective

From an accuracy perspective, binarizing the existing lightweight MobileNetV3 backbone in the encoder causes the most significant drop in accuracy among all parts

BiMatting: Efficient Video Matting via Binarization. Haotong Qin, et al. NeurIPS 2022.

CNN Binarization (Video Matting)

Direct Binarization Baseline: Efficiency Bottleneck



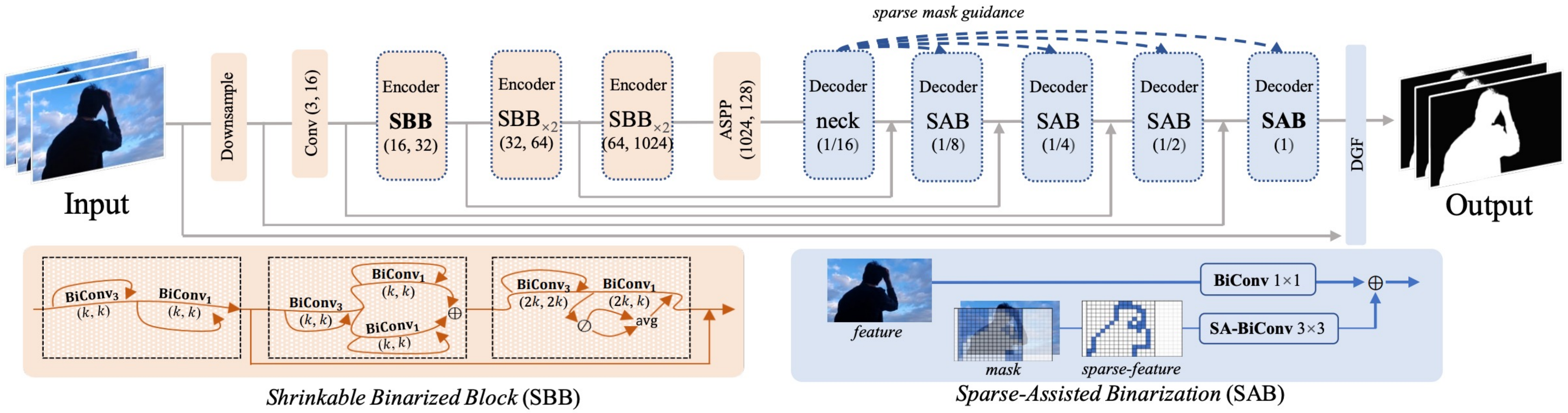
(b) Efficiency perspective

From an efficiency perspective, the decoder consumes a significant amount of computational resources even after binarization

BiMatting: Efficient Video Matting via Binarization. Haotong Qin, et al. NeurIPS 2022.

CNN Binarization (Video Matting)

BiMatting: Efficient Video Matting via Binarization



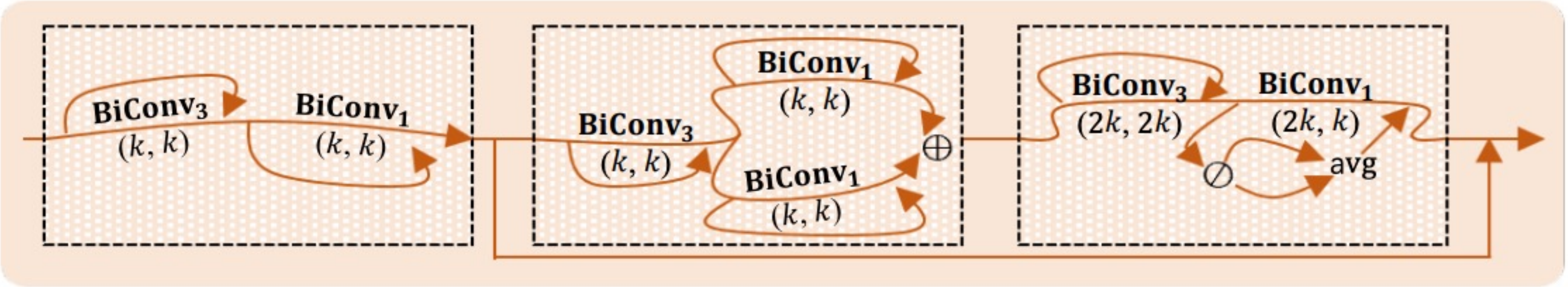
Shrinkable Binarized Block (SBB)

Sparse-Assisted Binarization (SAB)

BiMatting: Efficient Video Matting via Binarization. Haotong Qin, et al. NeurIPS 2022.

CNN Binarization (Video Matting)

BiMatting: Efficient Video Matting via Binarization



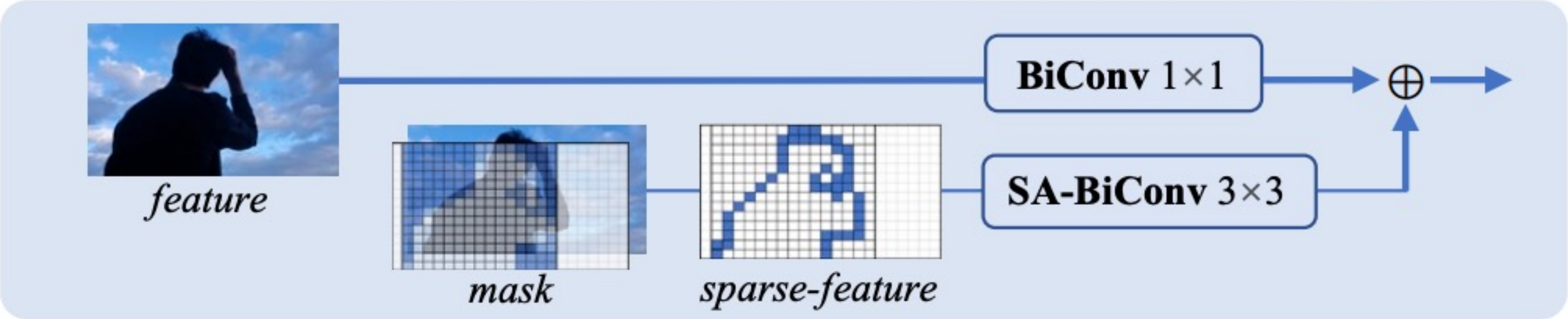
Shrinkable Binarized Block (SBB) for Accurate Encoder: the crucial paradigm of an accurate binarized encoder is the computation-dense form of binarized block.

$$\boldsymbol{SBB}: \quad \boldsymbol{o} = \theta^{\mathrm{dn}} \cdot \theta^{\mathrm{up}}(\boldsymbol{x}') + \boldsymbol{x}', \quad \boldsymbol{x}' = \theta^{\mathrm{eq}}(\boldsymbol{x})[c^{\boldsymbol{x}} = c^{\boldsymbol{o}}] + \theta^{\mathrm{up}}(\boldsymbol{x})\left[c^{\boldsymbol{x}} = \frac{1}{2}c^{\boldsymbol{o}}\right].$$

BiMatting: Efficient Video Matting via Binarization. Haotong Qin, et al. NeurIPS 2022.

CNN Binarization (Video Matting)

BiMatting: Efficient Video Matting via Binarization



Sparse-Assisted Binarization (SAB) for Efficient Decoder:

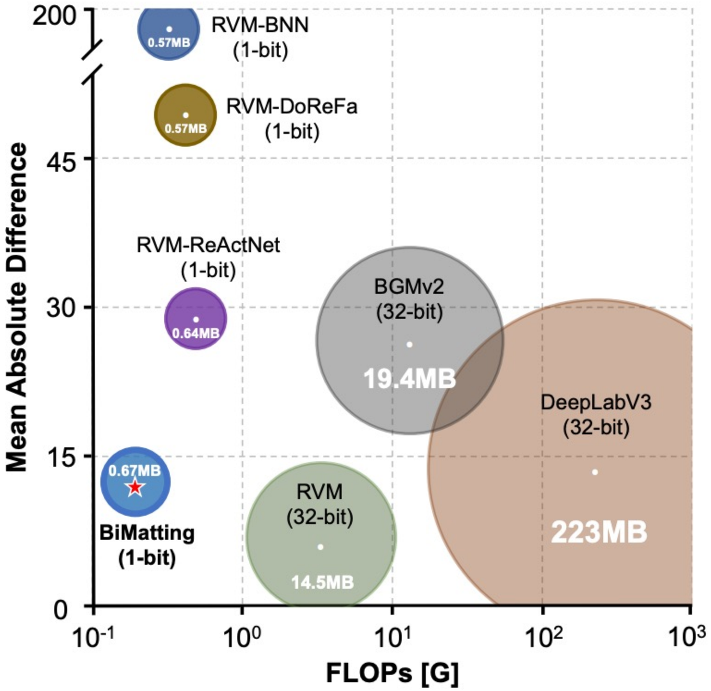
$$\textbf{\textit{SAB}}: \quad \boldsymbol{o} = \text{SA-BiConv}_3(\boldsymbol{x}; \text{bilinear}^k(M_{\text{inc}})) + \text{BiConv}_1(\boldsymbol{x}),$$

BiMatting: Efficient Video Matting via Binarization. Haotong Qin, et al. NeurIPS 2023.

CNN Binarization (Video Matting)

Performance

					Alpha					FG
Dataset	Method	#Bit	#FLOPs(G)	#Param(MB)	MAD	MSE	Grad	Conn	dtSSD	MSE
VM 512×288	DeepLabV3	32	136.06	223.66	14.47	9.67	8.55	1.69	5.18	-
	BGMv2	32	8.46	19.4	25.19	19.63	2.28	3.26	2.74	-
	RVM (oracle)	32	4.57	14.5	6.08	1.47	0.88	0.41	1.36	-
	RVM-BNN†	1	0.50	0.57	189.13	184.33	15.01	27.39	3.65	-
	RVM-DoReFa	1	0.52	0.57	51.64	34.50	8.85	7.14	4.09	-
	RVM-ReCU†	1	0.52	0.64	189.13	184.33	15.01	27.39	3.65	-
	RVM-ReAct	1	0.55	0.64	28.49	18.16	6.80	3.74	3.64	-
	BiMatting (Ours)	1	**0.37**	0.67	**12.82**	**6.65**	**2.97**	**1.42**	**2.69**	-
D646 512×512	DeepLabV3	32	241.89	223.66	24.50	20.1	20.30	6.41	4.51	-
	BGMv2	32	16.48	19.4	43.62	38.84	5.41	11.32	3.08	2.60
	RVM (oracle)	32	8.12	14.5	7.28	3.01	2.81	1.83	1.01	2.93
	RVM-BNN†	1	0.88	0.57	281.20	276.85	25.26	73.59	1.08	6.95
	RVM-DoReFa	1	0.92	0.57	133.63	116.69	17.09	35.08	2.58	6.97
	RVM-ReCU†	1	0.92	0.64	281.20	276.85	25.26	73.59	1.08	6.95
	RVM-ReAct	1	0.97	0.64	56.41	43.10	14.05	14.85	2.56	6.85
	BiMatting (Ours)	1	**0.66**	0.67	**32.74**	**24.48**	**9.34**	**8.62**	**2.21**	**5.86**



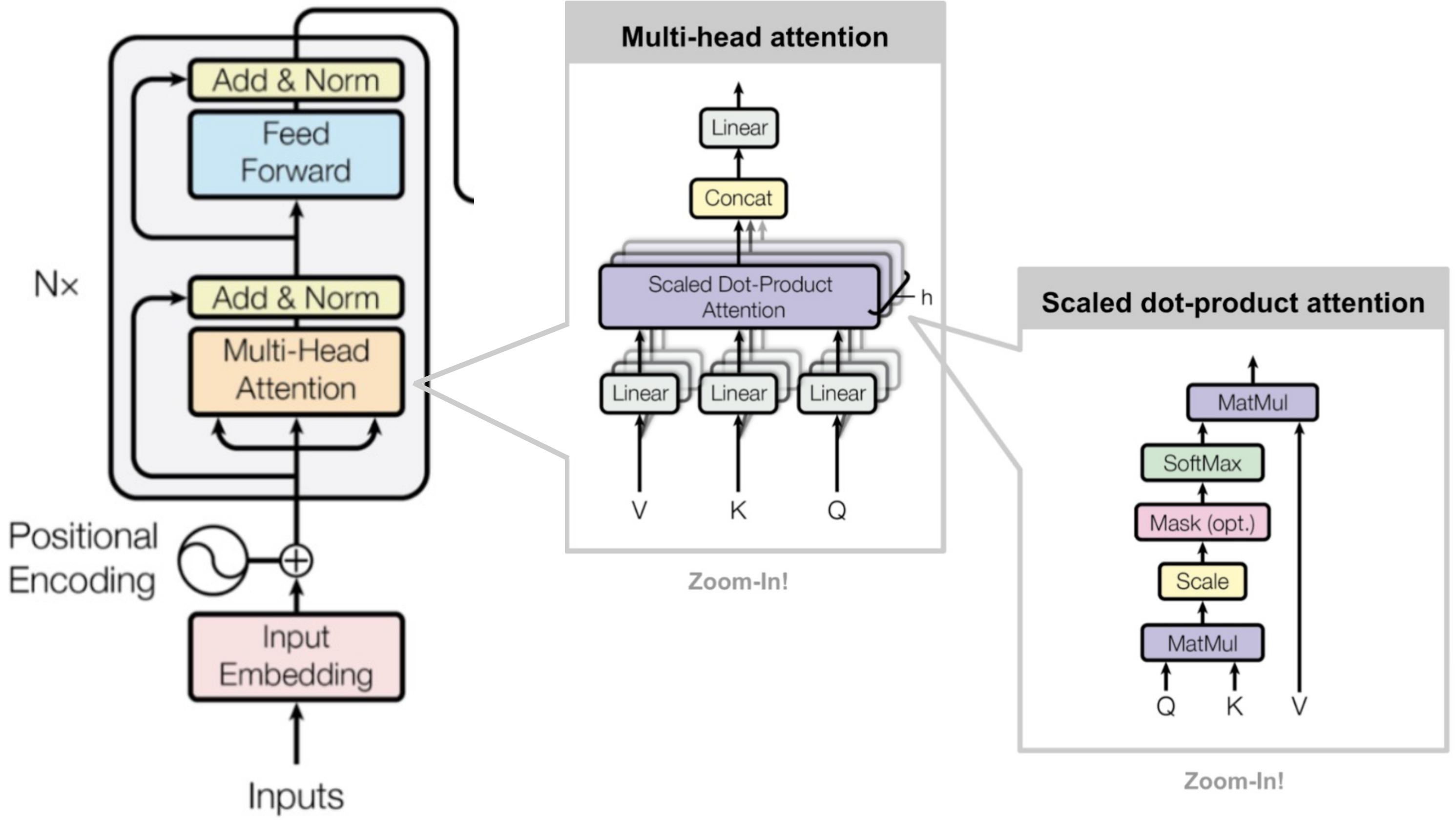
BiMatting: Efficient Video Matting via Binarization. Haotong Qin, et al. NeurIPS 2023.

CNN Binarization (Video Matting)

Performance

Transformer Binarization (Language Understanding)

Bottlenecks of Fully Binarized BERT Baseline

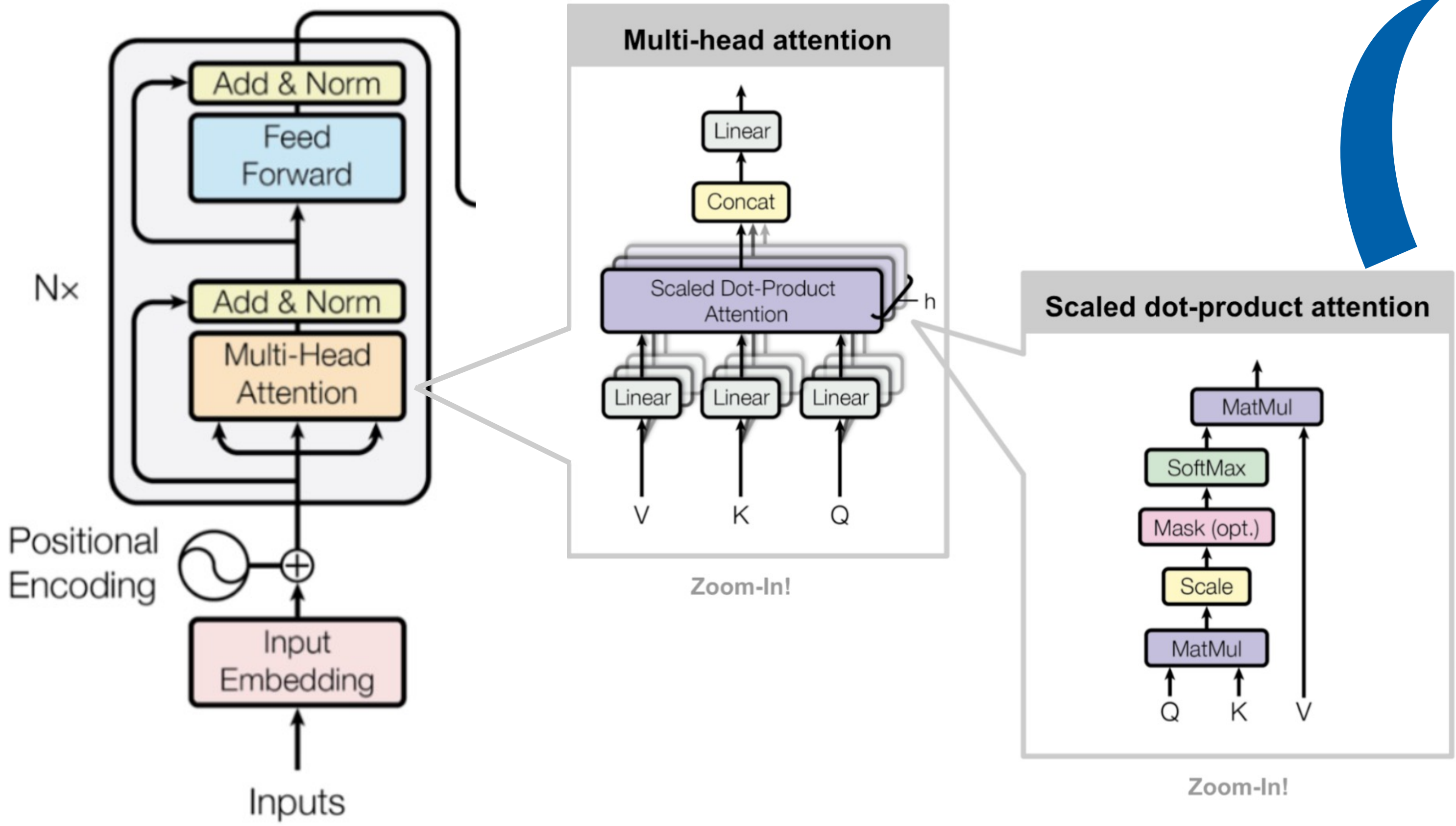


https://deepfrench.gitlab.io/deep-learning-project/

BiBERT: Accurate Fully Binarized BERT. Haotong Qin, et al. ICLR 2022.

Transformer Binarization (Language Understanding)

Bottlenecks of Fully Binarized BERT Baseline



https://deepfrench.gitlab.io/deep-learning-project/

Binarize (directly)

$$\mathbf{B_Q} = \operatorname{sign}(\mathbf{Q}), \quad \mathbf{B_K} = \operatorname{sign}(\mathbf{K})$$

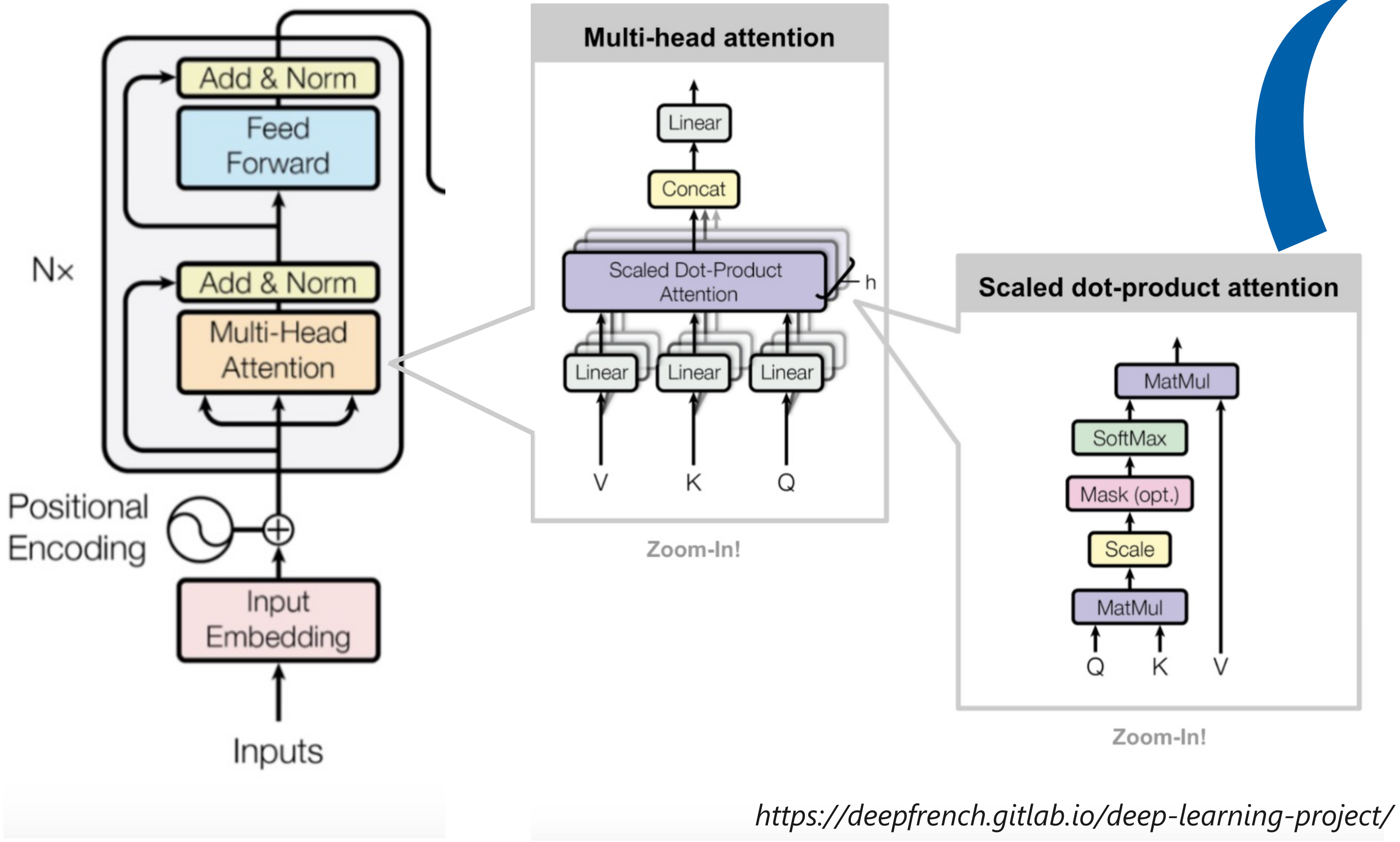
$$\mathbf{A} = \frac{1}{\sqrt{D}}\left(\mathbf{B_Q} \otimes \mathbf{B_K}^\top\right)$$

$$\mathbf{B}^s_{\mathbf{A}} = \operatorname{sign}(\operatorname{softmax}(\mathbf{A}))$$

BiBERT: Accurate Fully Binarized BERT. Haotong Qin, et al. ICLR 2022.

Transformer Binarization (Language Understanding)

Bottlenecks of Fully Binarized BERT Baseline



https://deepfrench.gitlab.io/deep-learning-project/

Binarize (directly)

$$\mathbf{B_Q} = \mathrm{sign}(\mathbf{Q}), \quad \mathbf{B_K} = \mathrm{sign}(\mathbf{K})$$

$$\mathbf{A} = \frac{1}{\sqrt{D}}\left(\mathbf{B_Q} \otimes \mathbf{B_K}^\top\right)$$

$$\mathbf{B}^s_{\mathbf{A}} = \mathrm{sign}(\mathrm{softmax}(\mathbf{A}))$$

Severely dropped!

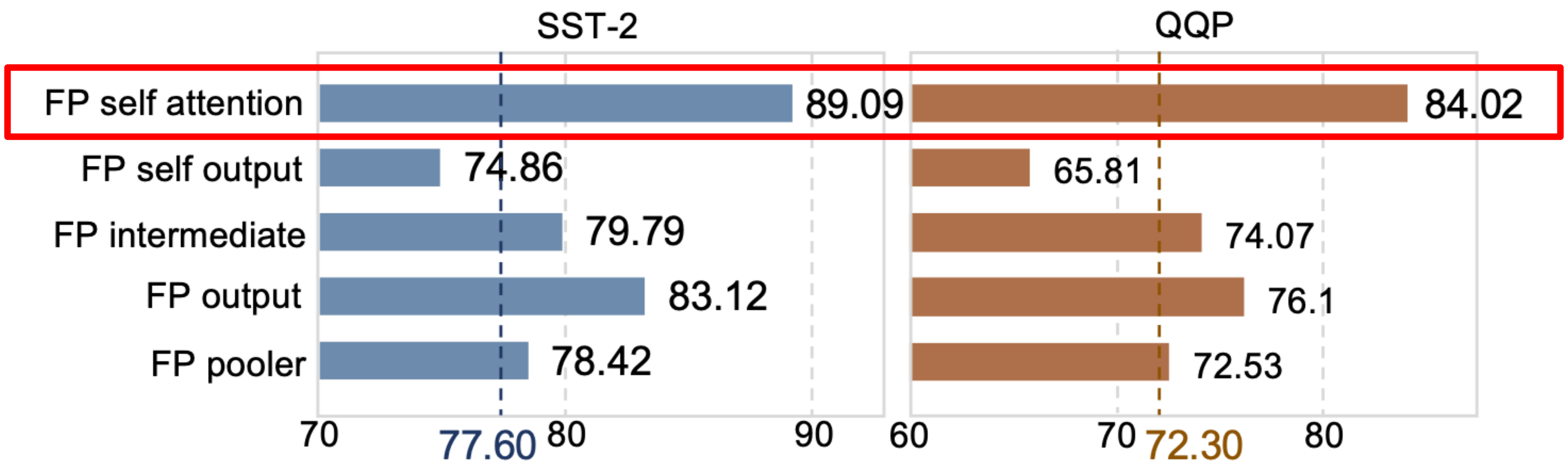
(Avg: 83.9% -> **50.4%**)

BiBERT: Accurate Fully Binarized BERT. Haotong Qin, et al. ICLR 2022.

Transformer Binarization (Language Understanding)

Bottlenecks of Fully Binarized BERT Baseline

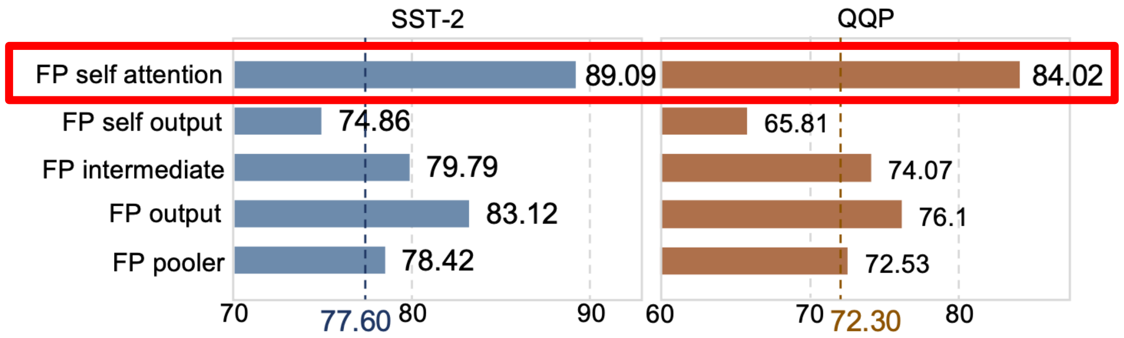
Which part caused the biggest drop?

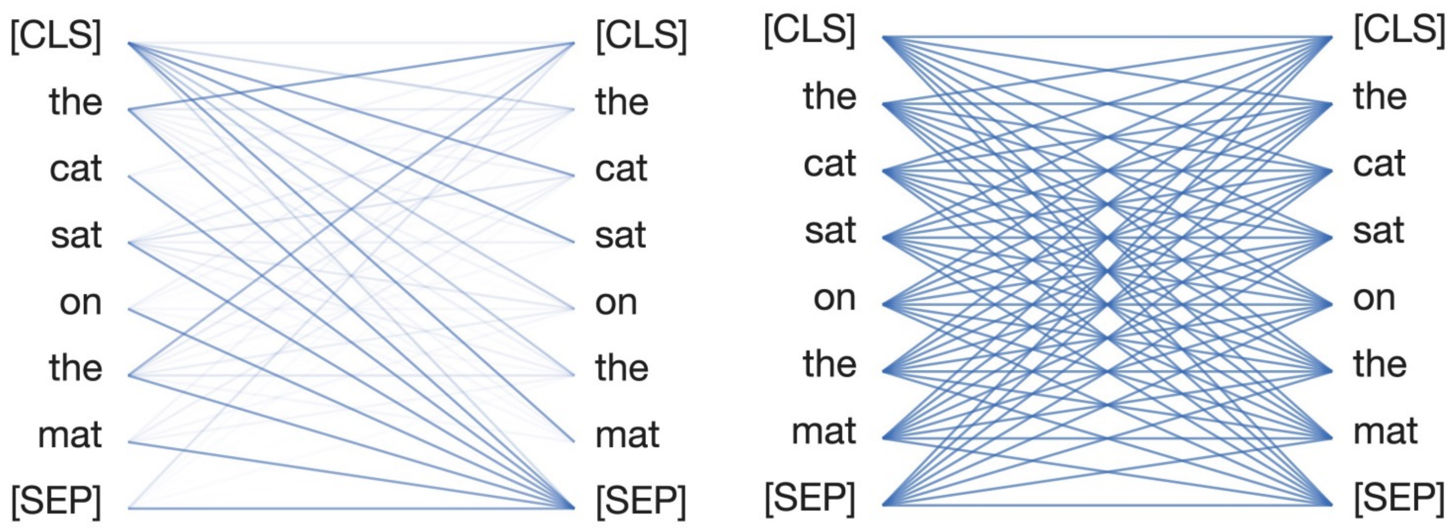


BiBERT: Accurate Fully Binarized BERT. Haotong Qin, et al. ICLR 2022.

Transformer Binarization (Language Understanding)

Bottlenecks of Fully Binarized BERT Baseline





(a) Full-precision

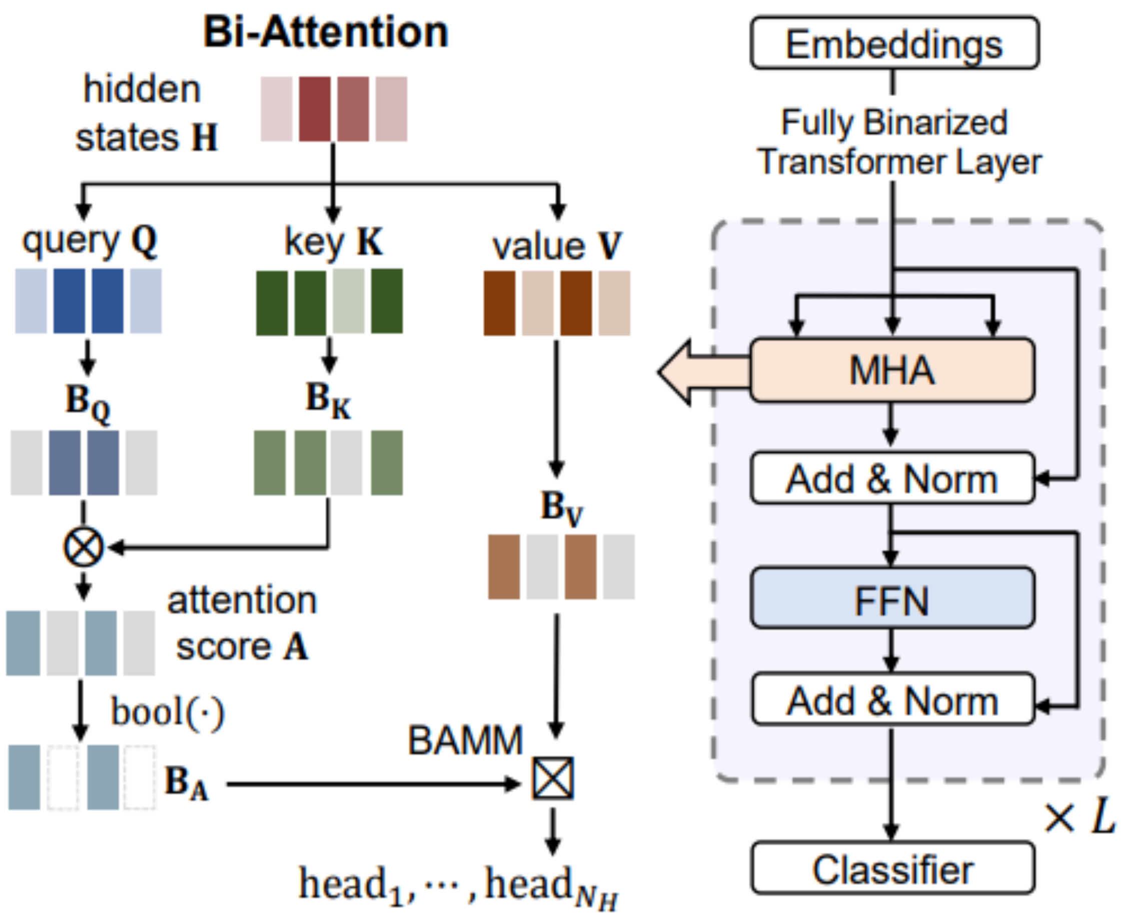
(b) Fully binarized BERT baseline

attention mechanism crashed

BiBERT: Accurate Fully Binarized BERT. Haotong Qin, et al. ICLR 2022.

Transformer Binarization (Language Understanding)

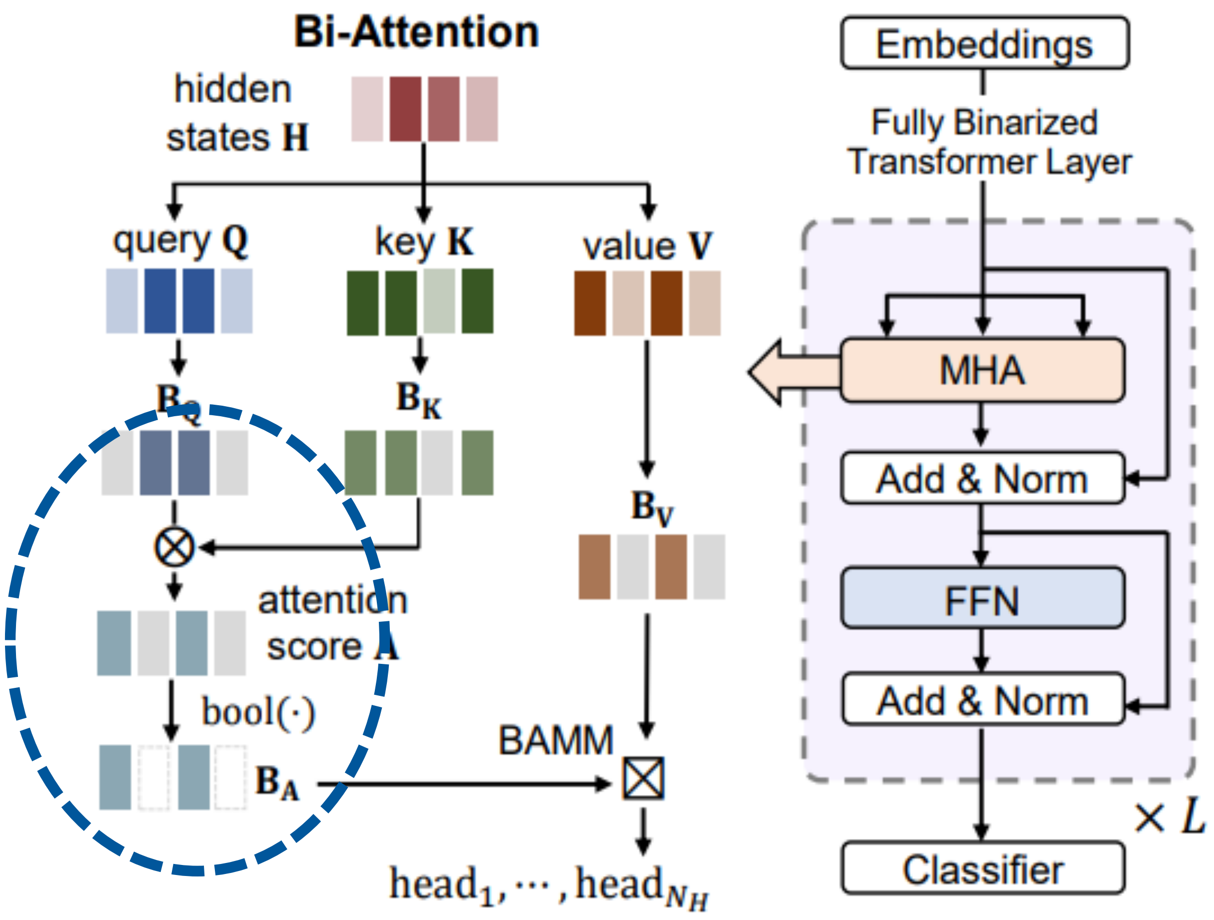
BiBERT: Accurate Fully Binarized BERT



BiBERT: Accurate Fully Binarized BERT. Haotong Qin, et al. ICLR 2022.

Transformer Binarization (Language Understanding)

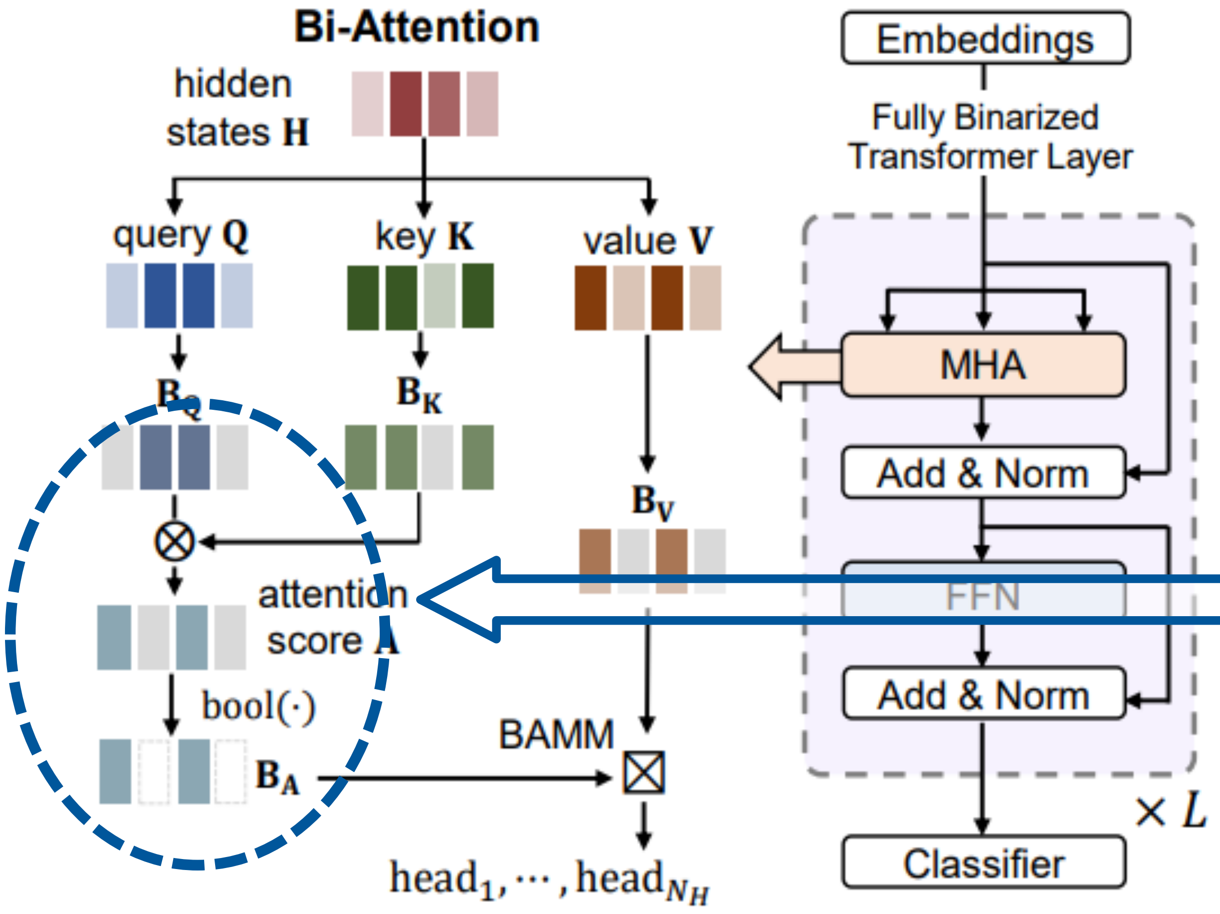
BiBERT: Accurate Fully Binarized BERT



BiBERT: Accurate Fully Binarized BERT. Haotong Qin, et al. ICLR 2022.

Transformer Binarization (Language Understanding)

BiBERT: Accurate Fully Binarized BERT



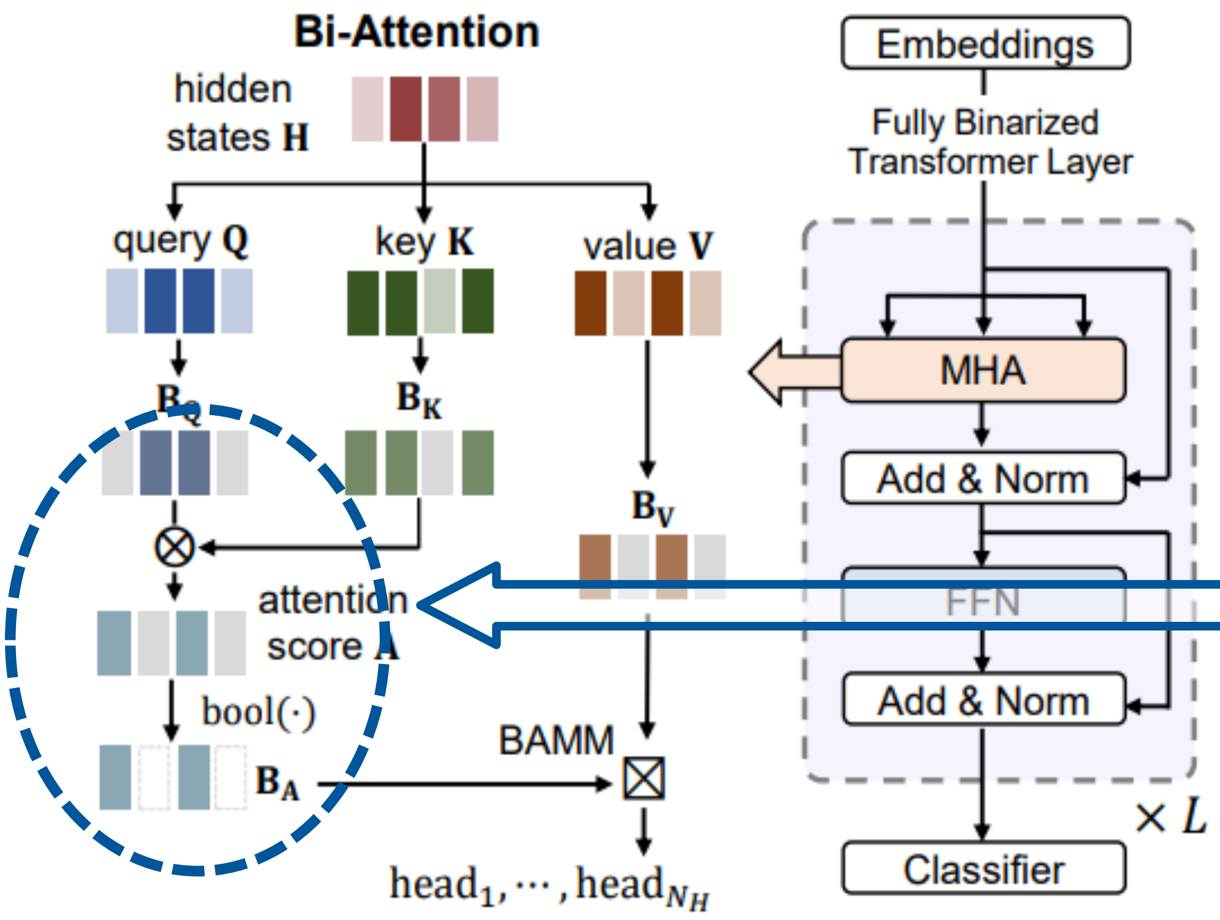
$$\mathrm{bool}(x) = \begin{cases} 1, & \text{if } x \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

$$\frac{\partial\, \mathrm{bool}(x)}{\partial x} = \begin{cases} 1, & \text{if } |x| \leq 1 \\ 0, & \text{otherwise} \end{cases}$$

BiBERT: Accurate Fully Binarized BERT. Haotong Qin, et al. ICLR 2022.

Transformer Binarization (Language Understanding)

BiBERT: Accurate Fully Binarized BERT



1. ~~SoftMax~~

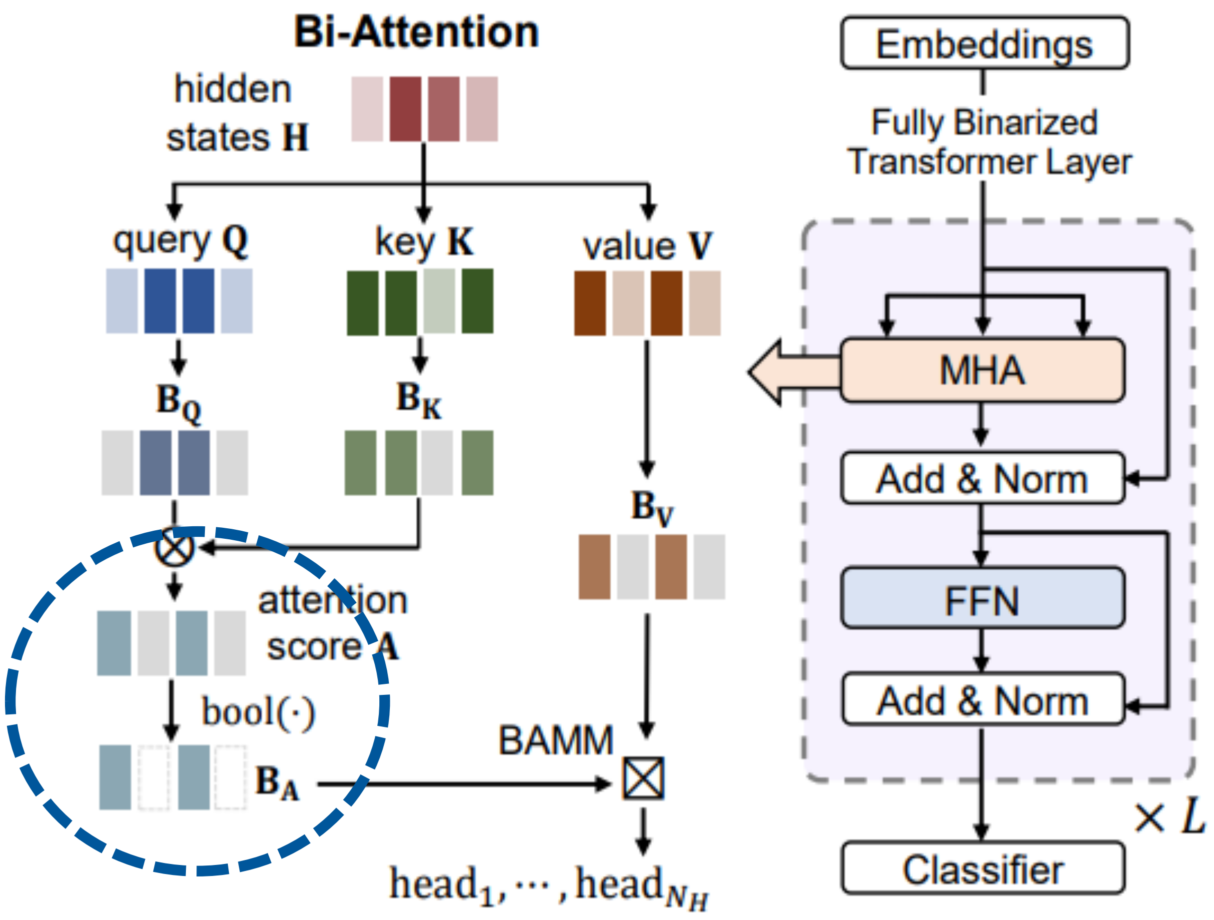
$$\mathrm{bool}(x) = \begin{cases} 1, & \text{if } x \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

$$\frac{\partial\, \mathrm{bool}(x)}{\partial x} = \begin{cases} 1, & \text{if } |x| \leq 1 \\ 0, & \text{otherwise} \end{cases}$$

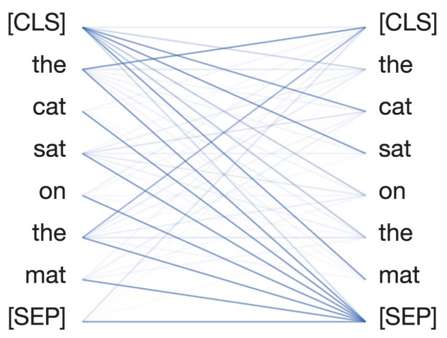
BiBERT: Accurate Fully Binarized BERT. Haotong Qin, et al. ICLR 2022.

Transformer Binarization (Language Understanding)

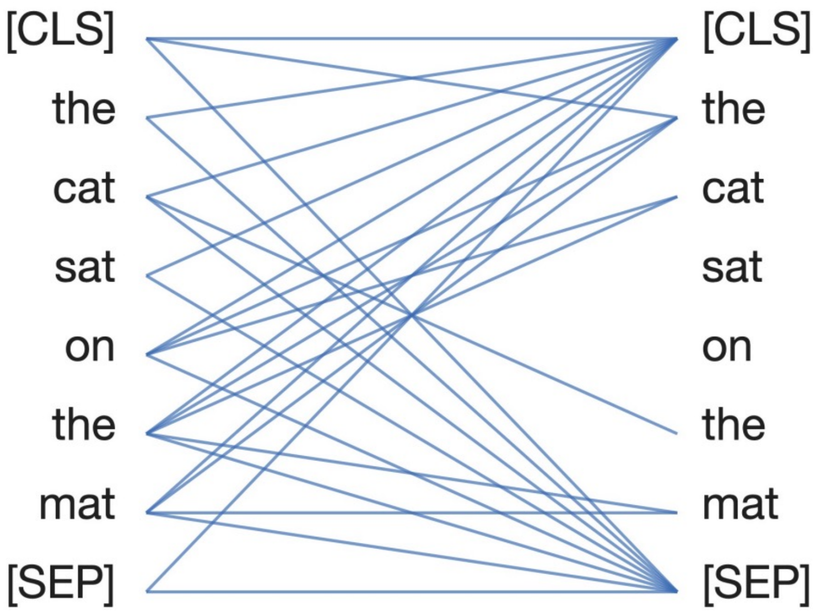
BiBERT: Accurate Fully Binarized BERT



2.



(a) Full-precision

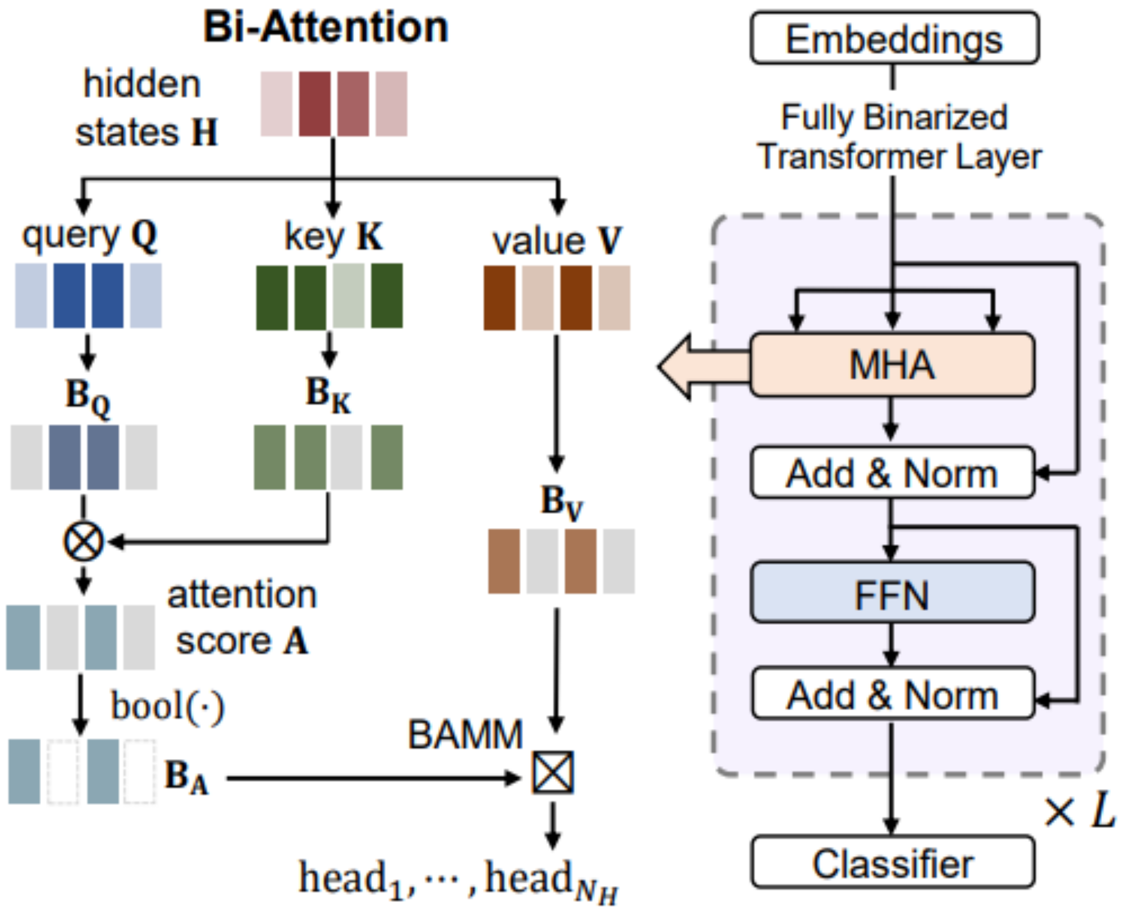


(c) BiBERT (Ours)

BiBERT: Accurate Fully Binarized BERT. Haotong Qin, et al. ICLR 2022.

Transformer Binarization (Language Understanding)

BiBERT: Accurate Fully Binarized BERT



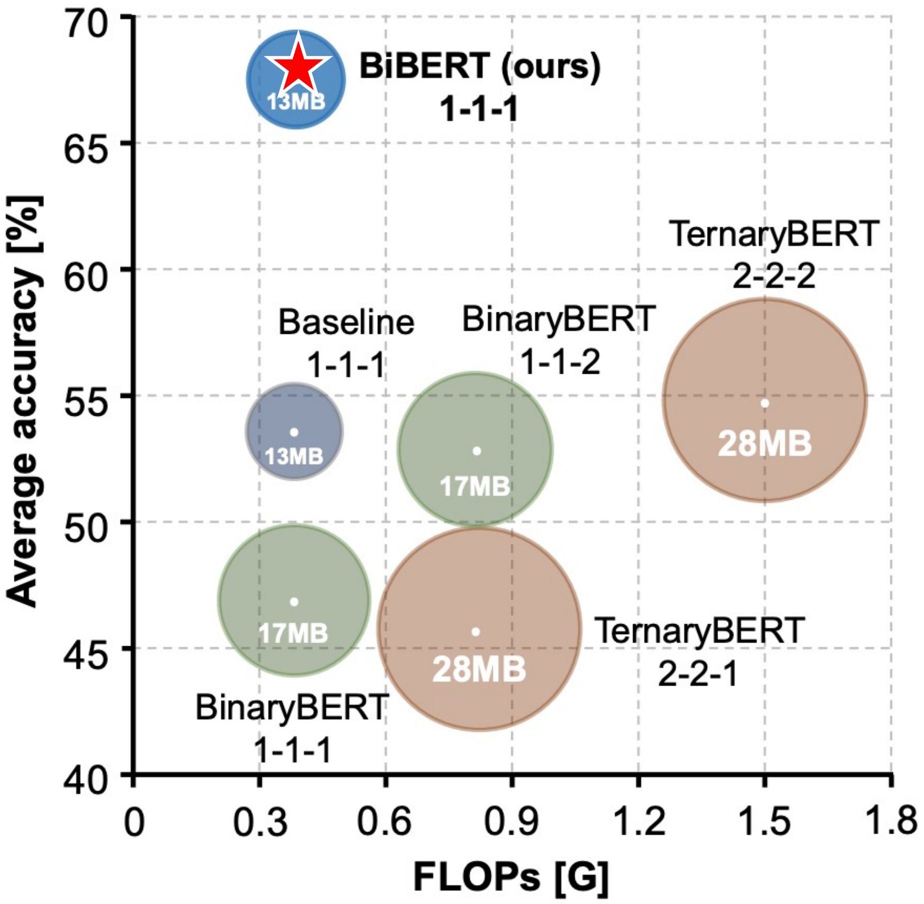
$$\mathbf{B_A} = \mathrm{bool}\left(\mathbf{A}\right) = \mathrm{bool}\left(\frac{1}{\sqrt{D}}\left(\mathbf{B_Q} \otimes \mathbf{B_K}^{\top}\right)\right)$$

$$\text{Bi-Attention}(\mathbf{B_Q}, \mathbf{B_K}, \mathbf{B_V}) = \mathbf{B_A} \boxtimes \mathbf{B_V}$$

BiBERT: Accurate Fully Binarized BERT. Haotong Qin, et al. ICLR 2022.

Transformer Binarization (Language Understanding)

Performance

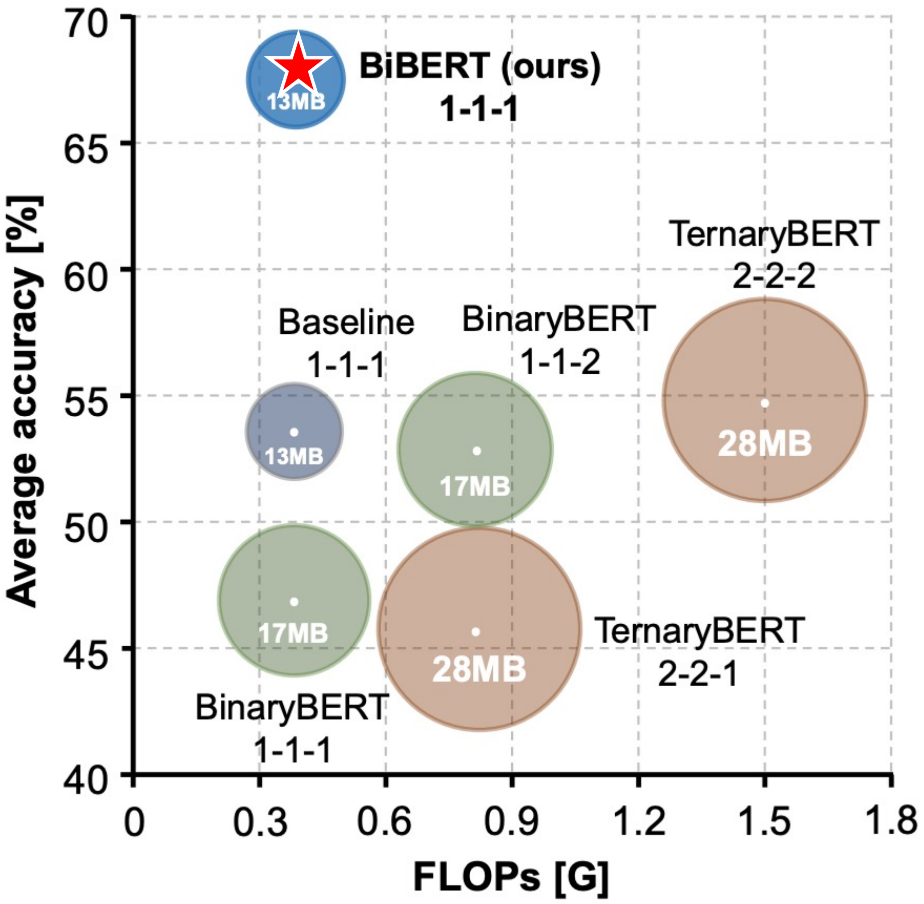


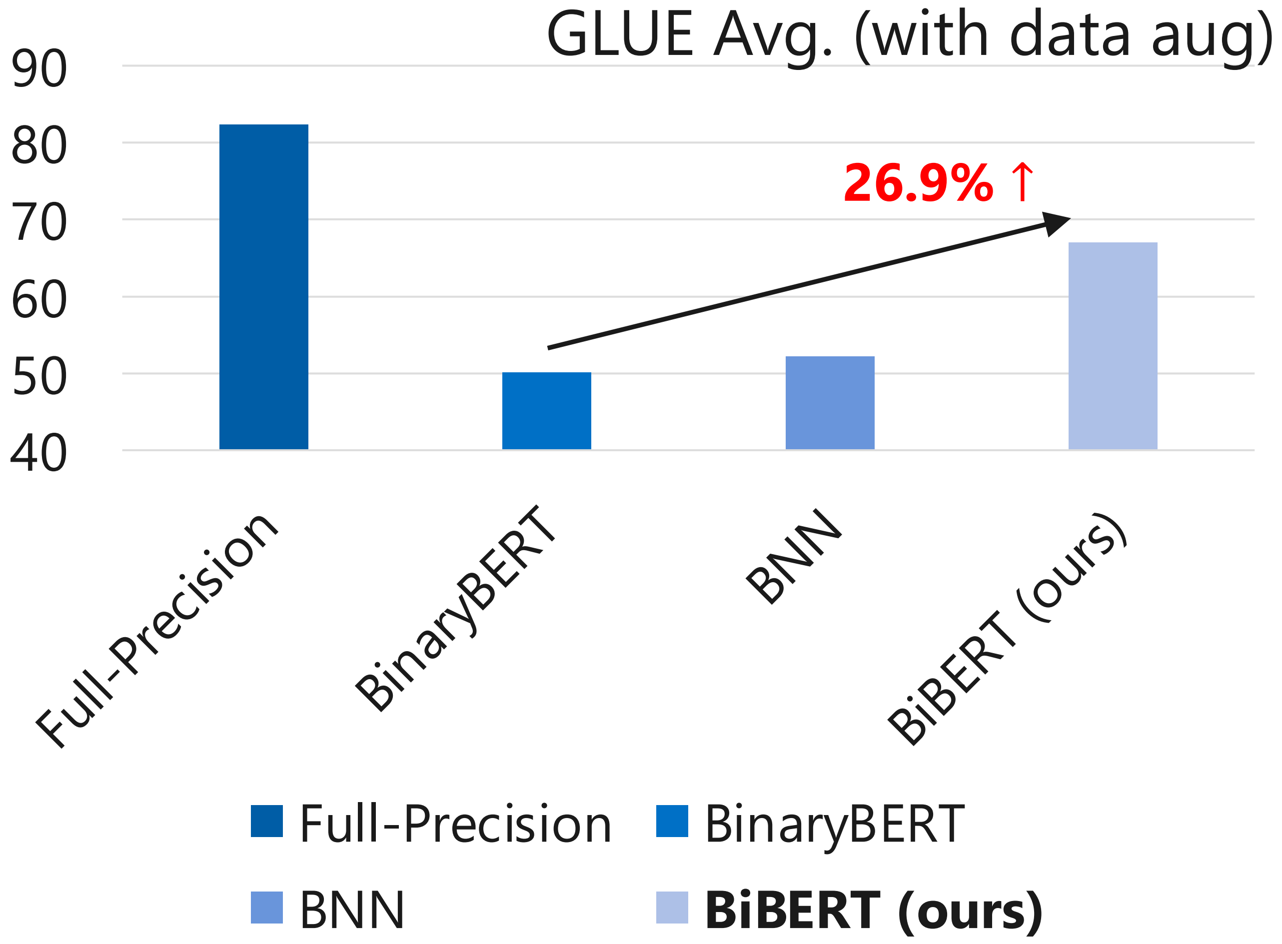
BiBERT: Accurate Fully Binarized BERT. Haotong Qin, et al. ICLR 2022.

Transformer Binarization (Language Understanding)

Performance

BiBERT was invited to integrated in deep learning platform **Baidu PaddlePaddle**

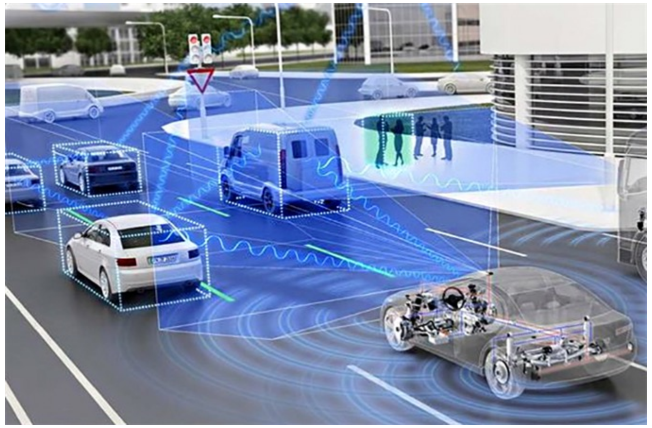


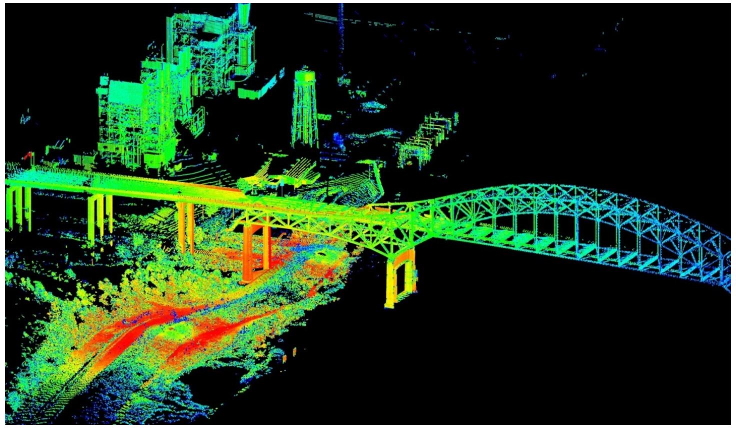


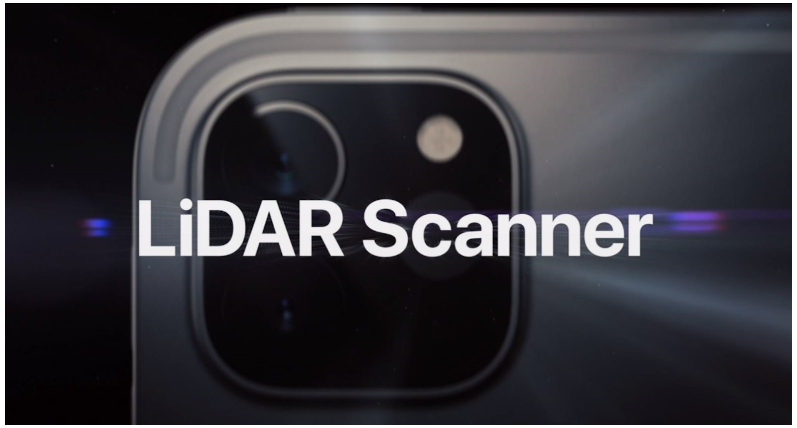
BiBERT: Accurate Fully Binarized BERT. Haotong Qin, et al. ICLR 2022.

MLP Binarization (Point Cloud Processing)

Point Cloud Processing on Edge





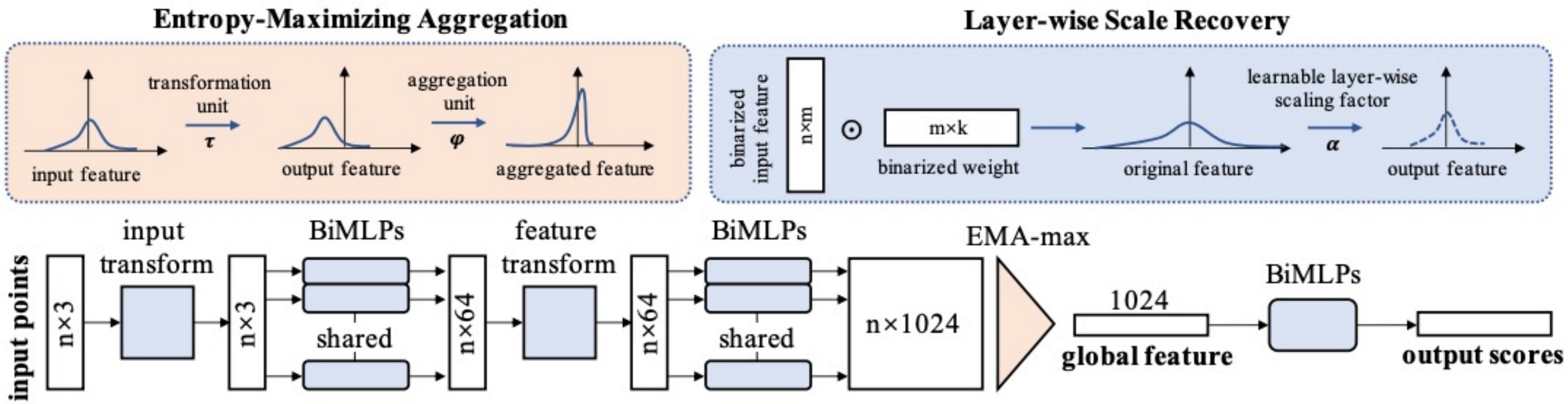


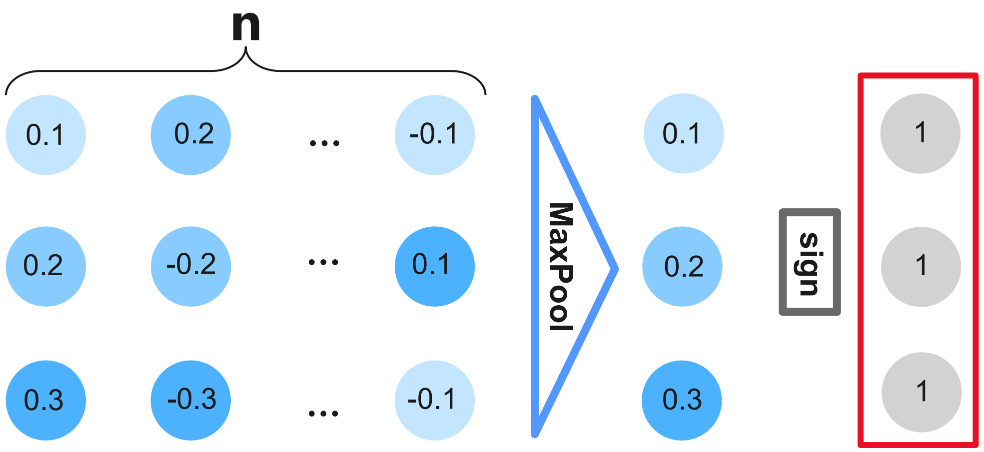
	Input Data	Convolution	Mean IoU	Latency	GPU Memory
PointNet [30]	points (8×2048)	none	83.7	21.7 ms	1.5 GB
3D-UNet [51]	voxels (8×96^3)	volumetric	84.6	682.1 ms	8.8 GB
RSNet [13]	points (8×2048)	point-based	84.9	74.6 ms	0.8 GB
PointNet++ [32]	points (8×2048)	point-based	85.1	77.9 ms	2.0 GB
DGCNN [43]	points (8×2048)	point-based	85.1	87.8 ms	2.4 GB

BiPointNet: Binary Neural Network for Point Clouds. Haotong Qin, et al. ICLR 2021.

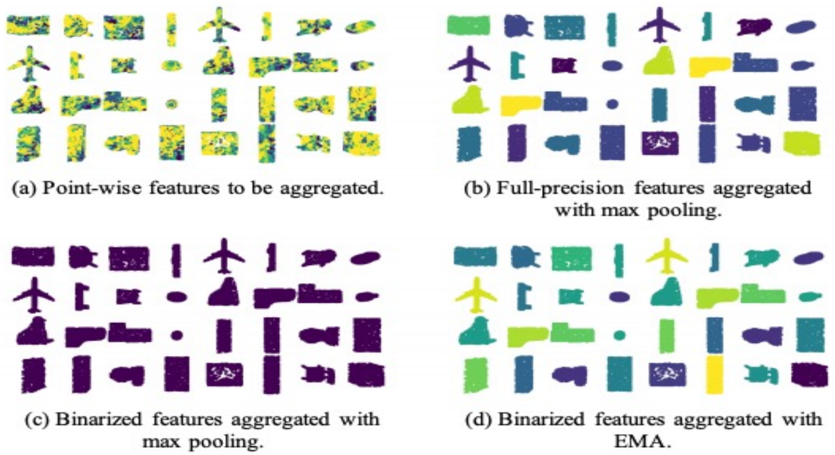
MLP Binarization (Point Cloud Processing)

BiPointNet: Entropy-Maximizing Aggregation





Vanilla Binarized



(a) Point-wise features to be aggregated.

(b) Full-precision features aggregated with max pooling.

(c) Binarized features aggregated with max pooling.

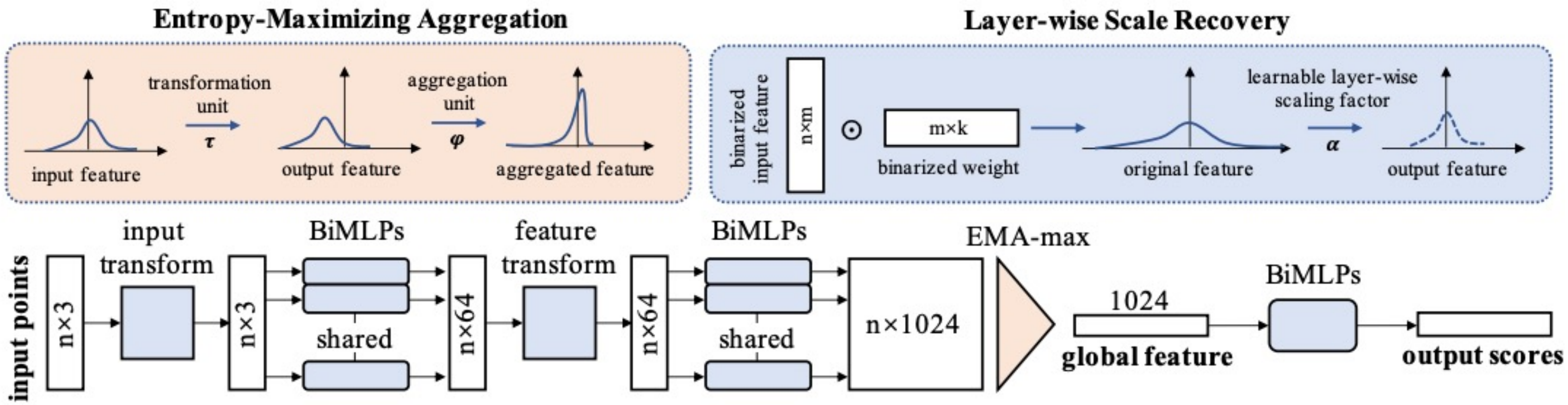
(d) Binarized features aggregated with EMA.

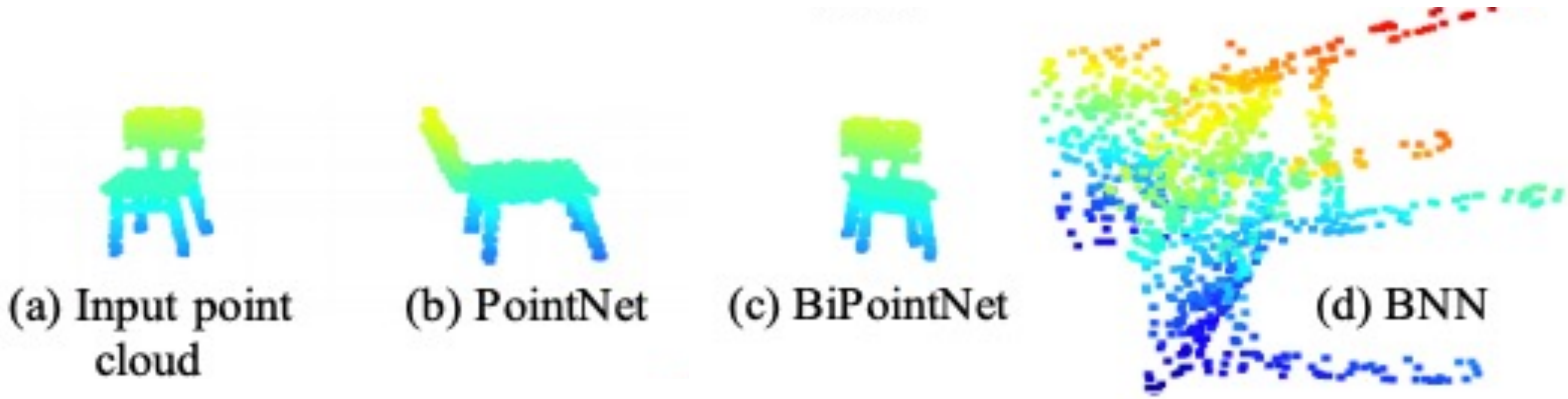
$$\mathbf{Y} = \mathrm{EMA}(\mathbf{X}_{\phi}) = \varphi(\tau(\mathbf{X}_{\phi}))$$

BiPointNet: Binary Neural Network for Point Clouds. Haotong Qin, et al. ICLR 2021.

MLP Binarization (Point Cloud Processing)

BiPointNet: Layerwise Scale Recovery





$$\alpha_0 = \frac{\sigma(\mathbf{A} \otimes \mathbf{W})}{\sigma(\mathbf{B}_a \odot \mathbf{B}_w)}$$

BiPointNet: Binary Neural Network for Point Clouds. Haotong Qin, et al. ICLR 2021.

MLP Binarization (Point Cloud Processing)

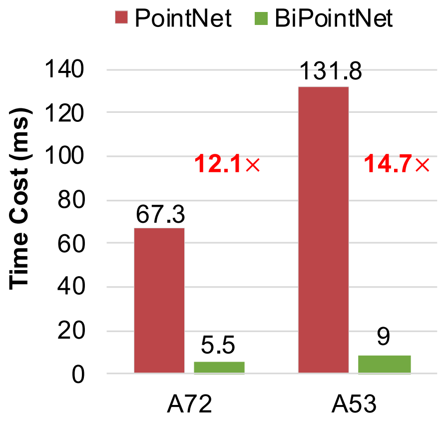
Performance

BiPointNet was invited to integrated in deep learning platform **Amazon DGL**

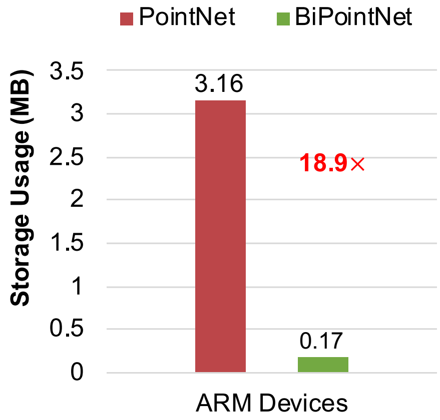
Method	Bit-width	Aggr.	# Factors	OA
Full Prec.	32/32	MAX	-	88.2
	32/32	AVG	-	86.5
BNN	1/1	MAX	0	7.1
	1/1	EMA-avg	0	11.3
	1/1	EMA-max	0	16.2
IR-Net	1/1	MAX	10097	7.3
	1/1	EMA-avg	10097	22.0
	1/1	EMA-max	10097	63.5
Bi-Real	1/1	MAX	10097	4.0
	1/1	EMA-avg	10097	77.0
	1/1	EMA-max	10097	77.5
ABC-Net	1/1	MAX	51	4.1
	1/1	EMA-avg	51	68.9
	1/1	EMA-max	51	77.8
XNOR++	1/1	MAX	18	4.1
	1/1	EMA-avg	18	73.8
	1/1	EMA-max	18	78.4
XNOR	1/1	MAX	28529	64.9
	1/1	EMA-avg	28529	78.2
	1/1	EMA-max	28529	81.9
Ours	1/1	MAX	18	4.1
	1/1	EMA-avg	18	82.5
	1/1	EMA-max	18	**86.4**

Base Model	Method	Bit-width	Aggr.	OA
PointNet (Vanilla)	Full Prec.	32/32	MAX	86.8
	XNOR	1/1	MAX	61.0
	Ours	1/1	EMA-max	85.6
PointNet	Full Prec.	32/32	MAX	88.2
	XNOR	1/1	MAX	64.9
	Ours	1/1	EMA-max	86.4
PointNet++	Full Prec.	32/32	MAX	90.0
	XNOR	1/1	MAX	63.1
	Ours	1/1	EMA-max	87.8
PointCNN	Full Prec.	32/32	AVG	90.0
	XNOR	1/1	AVG	83.0
	Ours	1/1	EMA-avg	83.8
DGCNN	Full Prec.	32/32	MAX	89.2
	XNOR	1/1	MAX	51.5
	Ours	1/1	EMA-max	83.4
PointConv	Full Prec.	32/32	–	90.8
	XNOR	1/1	–	83.1
	Ours	1/1	–	87.9

2021 The Most Popular Papers in Beijing Area



14.7x
speedup



18.9x
storage saving

BiPointNet: Binary Neural Network for Point Clouds. Haotong Qin, et al. ICLR 2021.

MLP Binarization (Speech Keyword Spotting)

BiFSMN: High-frequency Enhancement Distillation

- Apply 2D Haar Wavelet Transform to decompose high-frequency components.

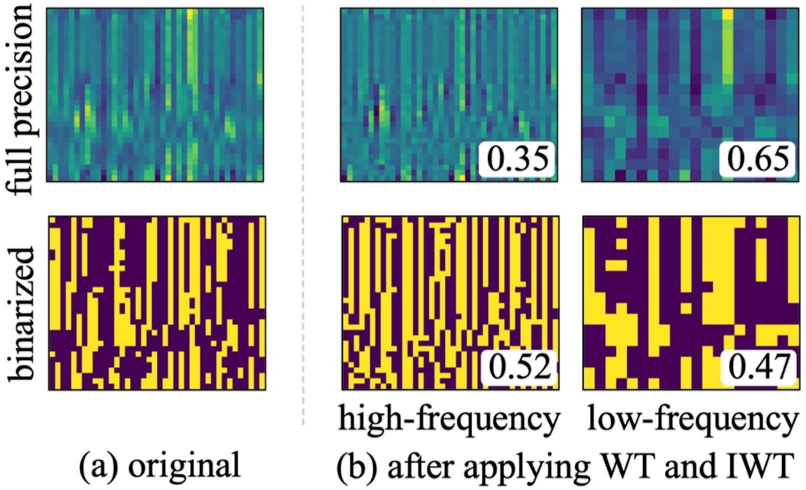
$$f_{\mathrm{WT}}(\mathbf{H}) = \sum_{j=-N}^{-1} \sum_{k} \mathbf{C}_j(k)\phi_{j,k} \qquad \mathbf{H}_{T\mathrm{H}} = f_{\mathrm{IWT}}\left(\sum_{k} \mathbf{C}_{T\mathrm{H}}(k)\phi_{T\mathrm{H},k}\right)$$

- Add emphasized high-frequency representations to the original ones.

$$\hat{\mathbf{H}}_T = \frac{\mathbf{H}_{T\mathrm{H}}}{\sigma(\mathbf{H}_{T\mathrm{H}})} + \frac{\mathbf{H}_T}{\sigma(\mathbf{H}_T)},$$

- Minimize the attention distillation loss

$$\mathcal{L}_{\mathrm{dist}} = \sum_{\ell=1}^{N} \left\| \frac{{\mathbf{H}_S^\ell}^2}{\|{\mathbf{H}_S^\ell}^2\|} - \frac{{\hat{\mathbf{H}}_T^\ell}^2}{\|{\hat{\mathbf{H}}_T^\ell}^2\|} \right\|$$



BiFSMN: Binary Neural Network for Keyword Spotting. Haotong Qin, et al. IJCAI 2022.
BiFSMNv2: Pushing Binary Neural Networks for Keyword Spotting to Real-Network Performance. Haotong Qin, et al. IEEE TNNLS 2023.

MLP Binarization (Speech Keyword Spotting)

BiFSMN: Thinnable Binarization Architecture

- Thinnable binarization architecture:

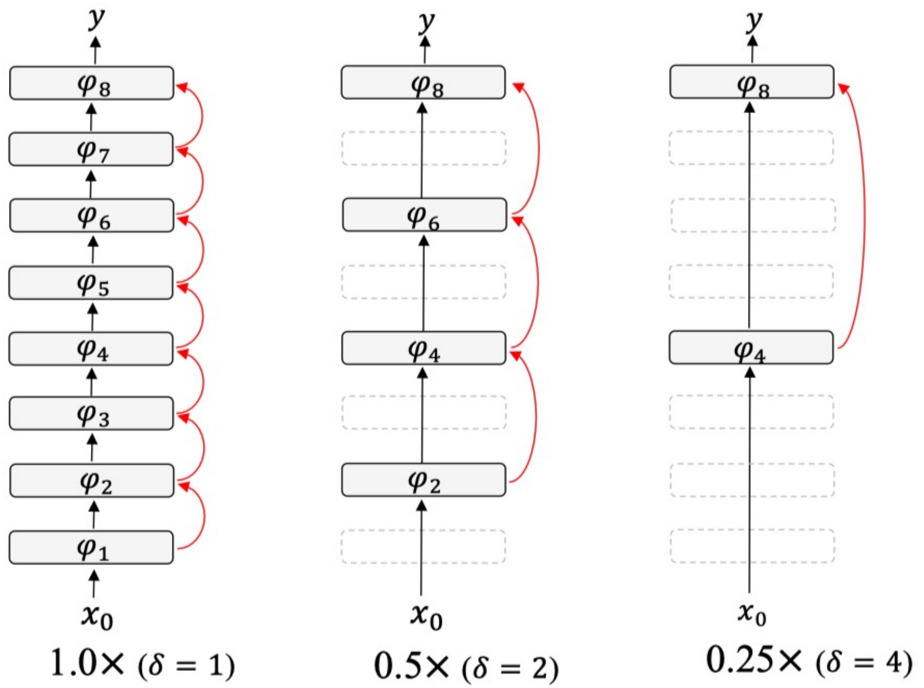
$$\mathbf{M}(\mathbf{x};\delta) = \Phi^N \cdot \Phi^{N-1} \cdot \ldots \cdot \Phi^1(\mathbf{x}),$$

- Each thinnable block:

$$\Phi^\ell(\mathbf{x}) = \begin{cases} \varphi^\ell(\mathbf{x}), & \ell \in \{i\delta, i \in [1, N/\delta]\}, \\ \mathbf{x}, & \text{otherwise.} \end{cases}$$

- Weighted loss:

$$\mathcal{L}_{\text{tot}} = \sum_\delta \frac{1}{2^{\delta-1}} \left(\mathcal{L}_{\text{CE}}^\delta + \gamma \mathcal{L}_{\text{dist}}^\delta\right)$$



Select fewer blocks at runtime

BiFSMN: Binary Neural Network for Keyword Spotting. Haotong Qin, et al. IJCAI 2022.
BiFSMNv2: Pushing Binary Neural Networks for Keyword Spotting to Real-Network Performance. Haotong Qin, et al. IEEE TNNLS 2023.

MLP Binarization (Speech Keyword Spotting)

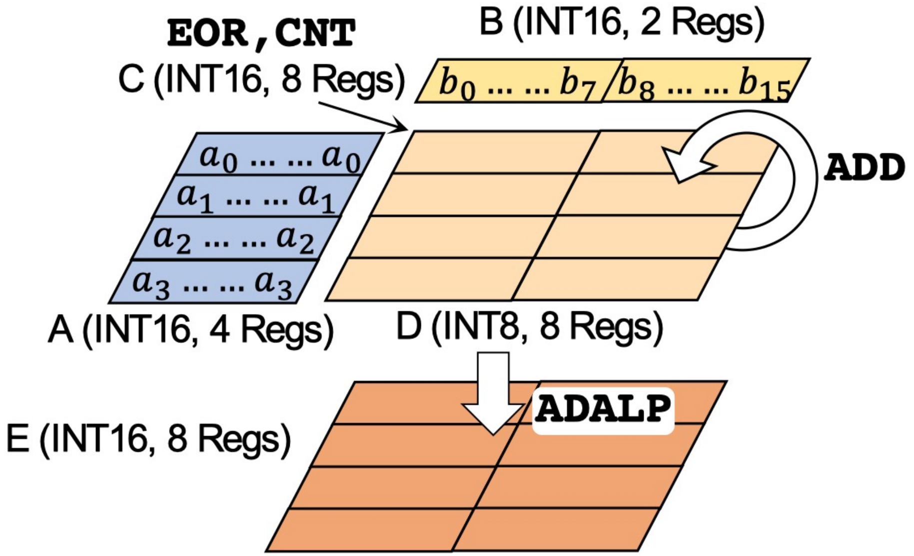
BiFSMN: Fast Bitwise Computation Kernel

- Bottlenecks of Acceleration on Hardware

  - Binarized General Matrix Multiply (BGEMM) performed with the bitwise XNOR and Bitcount

- Fast Bitwise Computation Kernel

  - optimize the 1-bit computation with new instruction and register allocation strategy



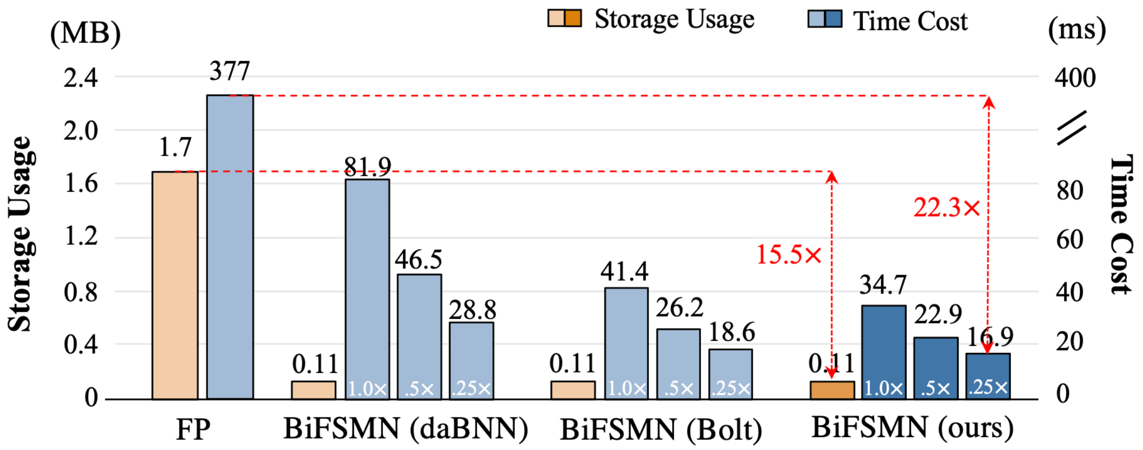
BiFSMN: Binary Neural Network for Keyword Spotting. Haotong Qin, et al. IJCAI 2022.
BiFSMNv2: Pushing Binary Neural Networks for Keyword Spotting to Real-Network Performance. Haotong Qin, et al. IEEE TNNLS 2023.

MLP Binarization (Speech Keyword Spotting)

Performance

Dataset	Method	#Bits(W/A)	FLOPs(M)	V1(%)	V2(%)
Speech Commands 12	Full Prec.	32/32	710.15	97.93	98.05
	DoReFa	1/1	40.46	66.42	66.59
	BNN	[illegible]	[illegible]	68.84	70.87
	RAD	[illegible]	[illegible]	71.51	69.80
	XNOR	[illegible]	[illegible]	82.74	87.34
	Bi-Real	1/1	40.46	85.87	87.93
	IR-Net	1/1	40.46	86.81	85.10
	BiFSMN [1,0.5,0.25]×	1/1	40.46	**95.03**	**94.86**
			29.90	94.87	94.73
			24.62	94.48	94.63
Speech Commands 20	Full Prec.	32/32	711.20	96.57	97.00
	XNOR	1/1	45.04	80.69	85.05
	Bi-Real	1/1	41.50	80.84	84.39
	IR-Net	1/1	41.50	83.78	83.32
	BiFSMN [1,0.5,0.25]×	1/1	41.50	**92.88**	**92.98**
			30.95	92.67	92.81
			25.67	92.65	92.72
Speech Commands 35	Full Prec.	32/32	713.16	96.63	95.96
	IR-Net	1/1	43.47	74.09	74.93
	Bi-Real	1/1	43.47	80.86	81.86
	XNOR	1/1	48.06	81.25	84.05
	BiFSMN [1,0.5,0.25]×	1/1	43.47	**92.10**	**90.67**
			32.91	91.93	90.54
			27.63	91.85	90.42

Accuracy Drop: <3%



15.5× storage saving, 22.3× acceleration on ARMv8

Deployed on TikTok App, PICO devices, etc.

BiFSMN: Binary Neural Network for Keyword Spotting. Haotong Qin, et al. IJCAI 2022.
BiFSMNv2: Pushing Binary Neural Networks for Keyword Spotting to Real-Network Performance. Haotong Qin, et al. IEEE TNNLS 2023.

Binarization Benchmark

Challenges in Existing Binarization Research

1. Confusing contributions (operators? structures?)

2. Limited comparisons (methods? architectures?)

3. Neglected practicality (hardware deployment?)

BiBench: Benchmarking and Analyzing Network Binarization. Haotong Qin, et al. ICML 2023.

Binarization Benchmark

Challen

1. Confusi

2. Limited

3. Neglect

Algorithm	Year	Conference	Citation (2023/01/25)	Operator Techniques	Open Source	Specified Structure / Training-pipeline
BitwiseNN (Kim & Smaragdis, 2016)	2016	ICMLW	**274**	**Yes**	No	**No**
DoReFa (Zhou et al., 2016)	2016	ArXiv	**1831**	**Yes**	**Yes**	**No**
XNOR-Net (Rastegari et al., 2016)	2016	ECCV	**4474**	**Yes**	**Yes**	**No**
BNN (Courbariaux et al., 2016a)	2016	NeurIPS	**2804**	**Yes**	**Yes**	**No**
LBCNN (Juefei-Xu et al., 2017)	2017	CVPR	**257**	**Yes**	**Yes**	Yes
LAB (Hou et al., 2017)	2017	ICLR	**204**	**Yes**	**Yes**	Yes
ABC-Net (Lin et al., 2017)	2017	NeurIPS	**599**	**Yes**	**Yes**	Yes
DBF (Tseng et al., 2018)	2018	IJCAI	10	**Yes**	No	Yes
MCNs (Wang et al., 2018b)	2018	CVPR	30	**Yes**	No	Yes
SBDs (Hu et al., 2018)	2018	ECCV	93	**Yes**	No	**No**
Bi-Real Net (Liu et al., 2018a)	2018	ECCV	**412**	**Yes**	**Yes**	**Opt**
PCNN (Gu et al., 2019)	2019	AAAI	68	**Yes**	No	Yes
CI-BCNN (Wang et al., 2019)	2019	CVPR	90	**Yes**	**Yes**	Yes
XNOR-Net++ (Bulat et al., 2019)	2019	BMVC	**131**	**Yes**	**Yes**	**No**
ProxyBNN (He et al., 2020)	2020	ECCV	16	**Yes**	No	Yes
Si-BNN (Wang et al., 2020a)	2020	AAAI	28	**Yes**	No	**No**
EBNN (Bulat et al., 2020)	2020	ICLR	38	**Yes**	**Yes**	Yes
RBNN (Lin et al., 2020)	2020	NeurIPS	79	**Yes**	**Yes**	**No**
ReActNet (Liu et al., 2020)	2020	ECCV	**182**	**Yes**	**Yes**	**Opt**
SA-BNN (Liu et al., 2021)	**2021**	AAAI	7	**Yes**	No	**No**
S^2-BNN (Shen et al., 2021)	**2021**	CVPR	11	**Yes**	**Yes**	Yes
MPT (Diffenderfer & Kailkhura, 2021)	**2021**	ICLR	43	**Yes**	**Yes**	Yes
FDA (Xu et al., 2021a)	**2021**	NeurIPS	18	**Yes**	**Yes**	**No**
ReCU (Xu et al., 2021b)	**2021**	ICCV	27	**Yes**	**Yes**	**No**
LCR-BNN (Shang et al., 2022a)	**2022**	ECCV	1	**Yes**	**Yes**	Yes
PokeBNN (Zhang et al., 2022b)	**2022**	CVPR	6	**Yes**	**Yes**	Yes

BiBench: Benchmarking and Analyzing Network Binarization. Haotong Qin, et al. ICML 2023.

Binarization Benchmark

Challenges in Existing Binarization Research

1. Confusing contributions (operators? structures?)

2. Limited comparisons (methods? architectures?)

3. Neglected practicality (hardware deployment?)

CIFAR & ImageNet (Image) ResNet, VGG, MobileNet, …	**BNN, DoReFa, Bi-Real, ReActNet, …**
COCO (Image) Faster-RCNN, SSD, SwinTransformer, …	(Few)
GLUE (Text), BERT-Base, BERT-Large, …	(Fewer)
…	(Almost None)

BiBench: Benchmarking and Analyzing Network Binarization. Haotong Qin, et al. ICML 2023.

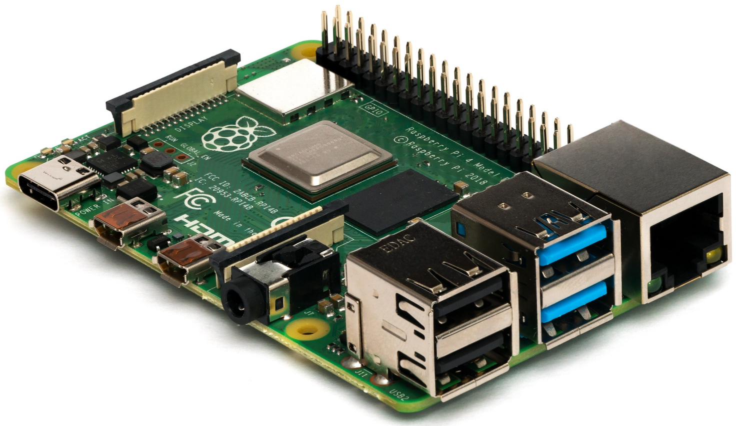
Binarization Benchmark

Challenges in Existing Binarization Research

1. Confusing contributions (operators? structures?)

2. Limited comparisons (methods? architectures?)

3. Neglected practicality (hardware deployment?)

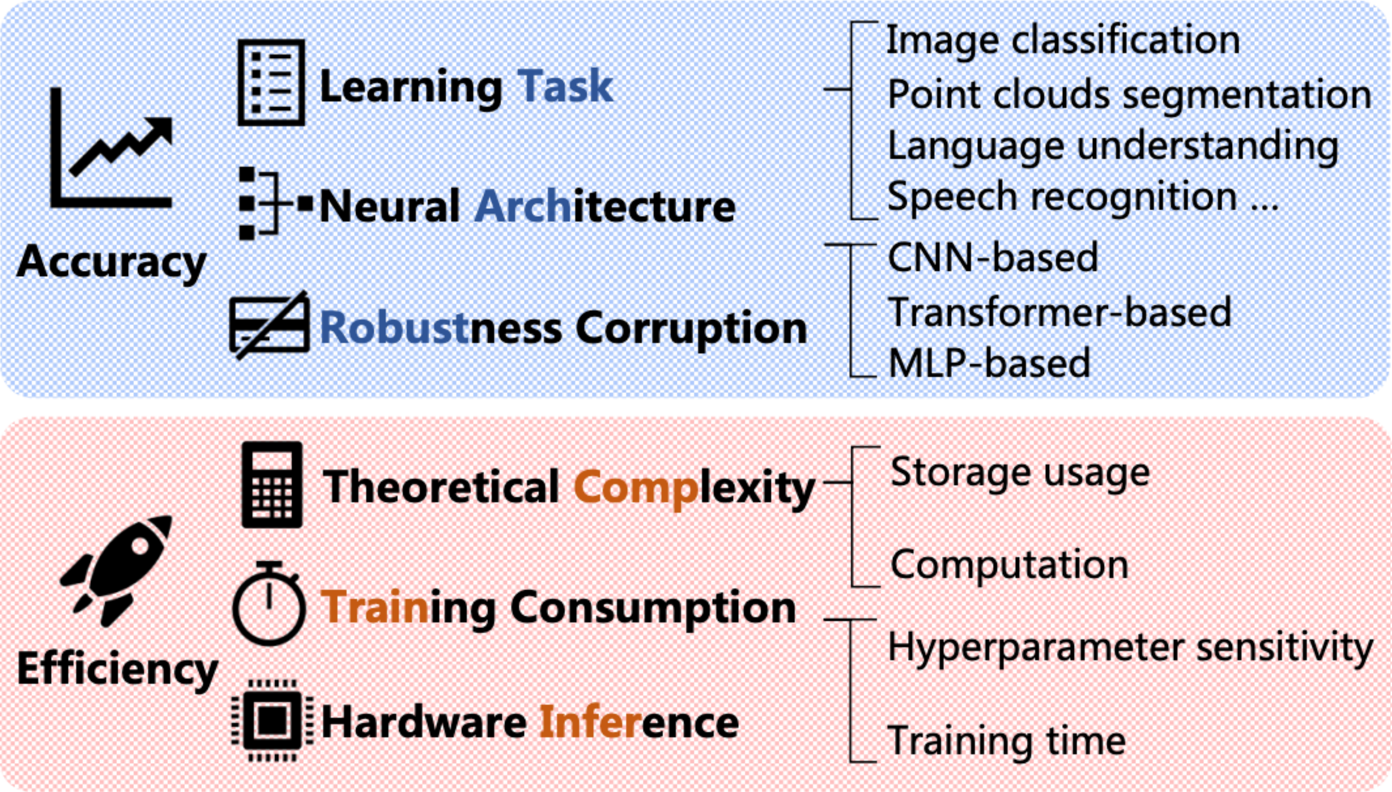


BiBench: Benchmarking and Analyzing Network Binarization. Haotong Qin, et al. ICML 2023.

Binarization Benchmark

***BiBench**: Benchmarking and Analyzing Network Binarization*

Evaluation Tracks for Network Binarization

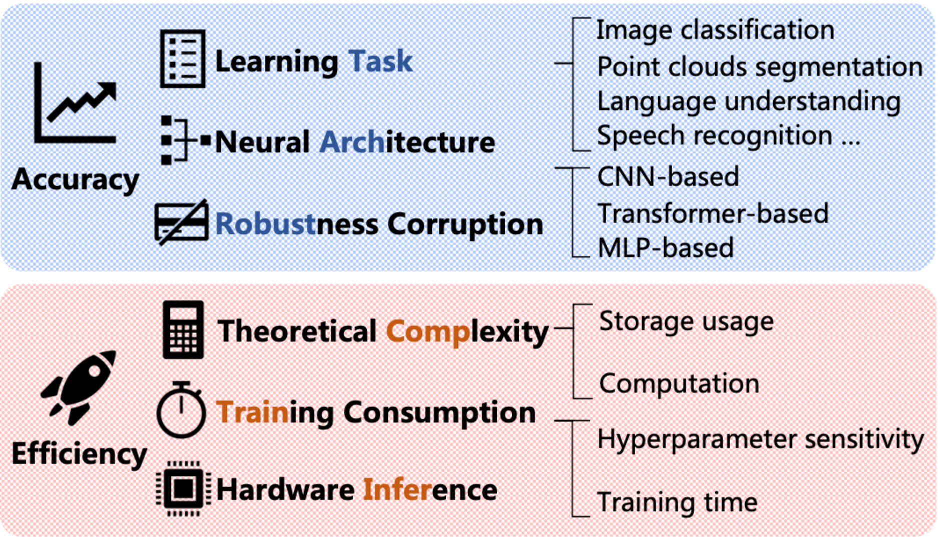


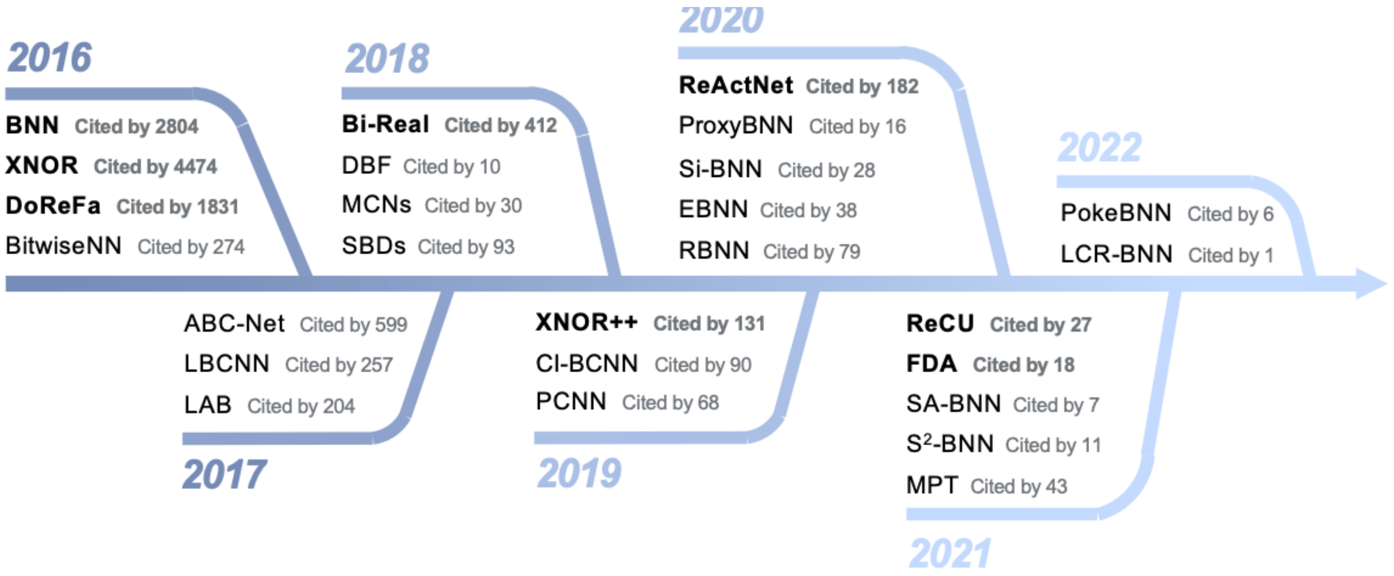
BiBench: Benchmarking and Analyzing Network Binarization. Haotong Qin, et al. ICML 2023.

Binarization Benchmark

***BiBench**: Benchmarking and Analyzing Network Binarization*

Evaluation Tracks for Network Binarization

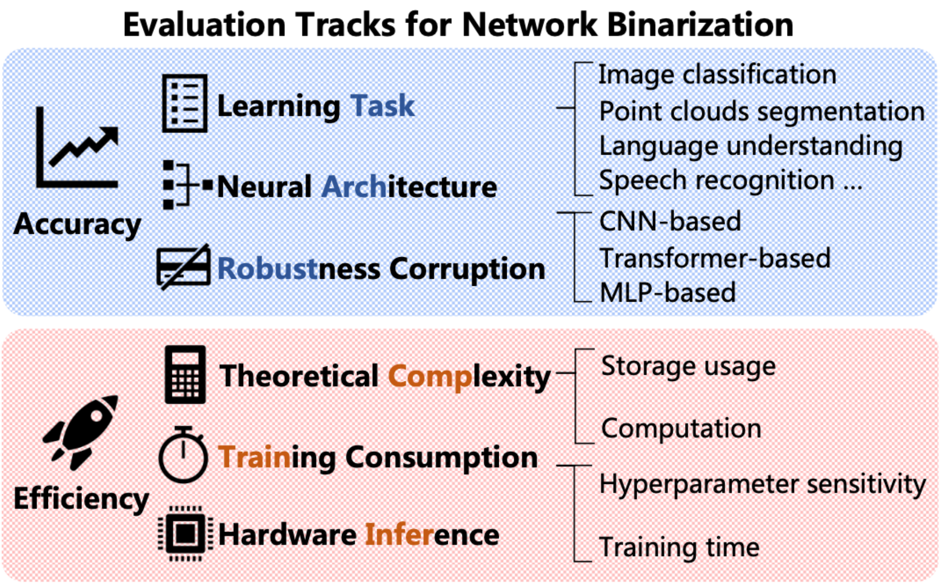


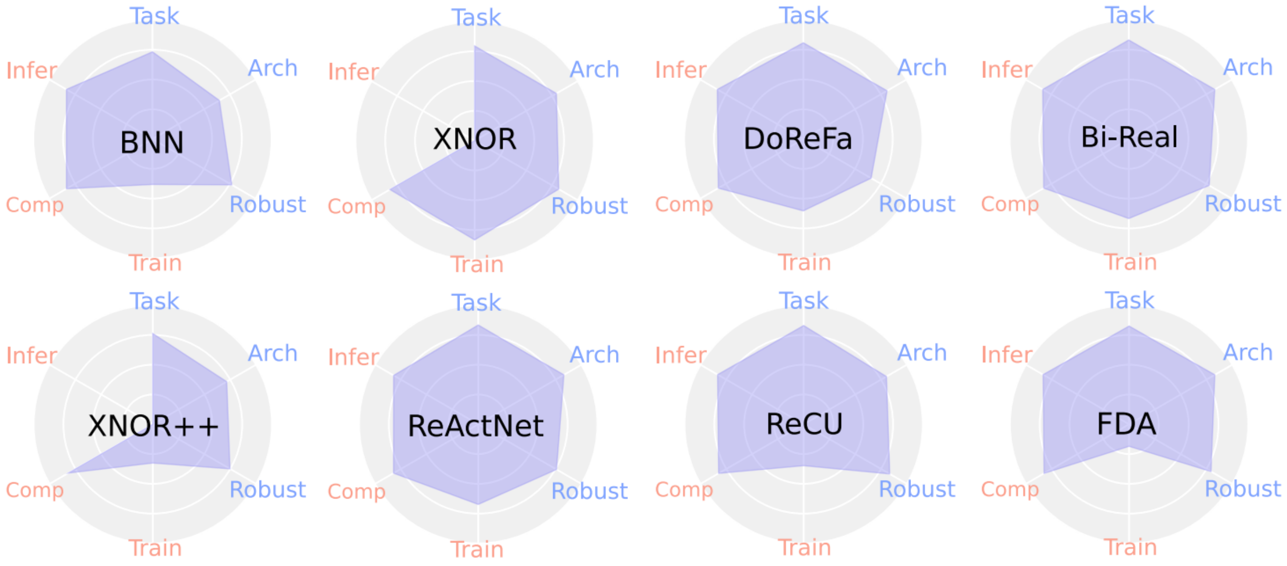


BiBench: Benchmarking and Analyzing Network Binarization. Haotong Qin, et al. ICML 2023.

Binarization Benchmark

***BiBench**: Benchmarking and Analyzing Network Binarization*

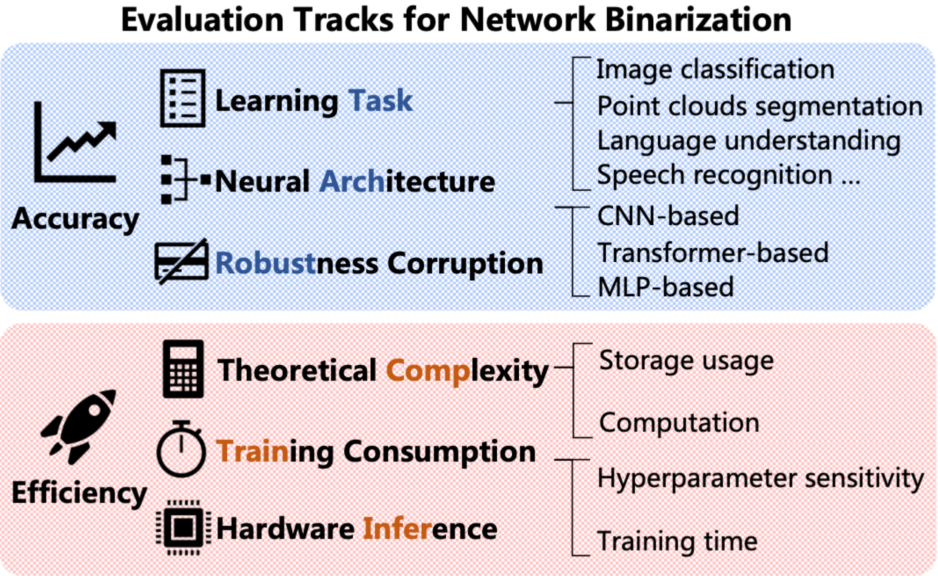


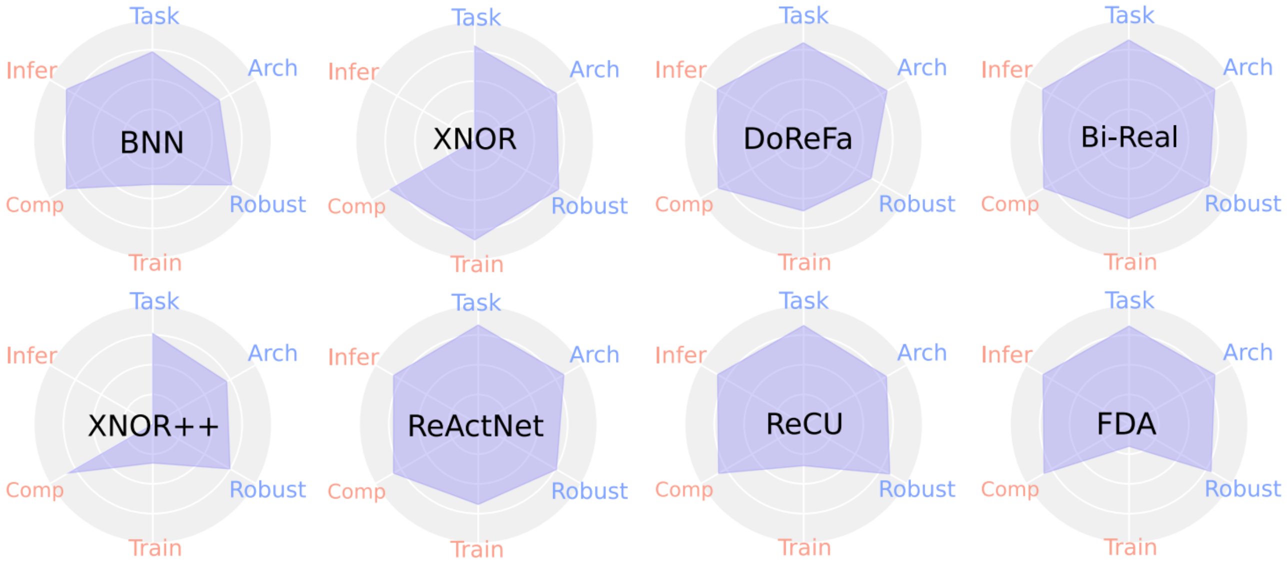


BiBench: Benchmarking and Analyzing Network Binarization. Haotong Qin, et al. ICML 2023.

Binarization Benchmark

***BiBench**: Benchmarking and Analyzing Network Binarization*





6 Evaluation Tracks on Accuracy and Efficiency

8 Binarization Algorithm
9 Deep Learning Datasets
13 Neural Architectures
2 Deployment Libraries
14 Hardware Chips

BiBench: Benchmarking and Analyzing Network Binarization. Haotong Qin, et al. ICML 2023.

Binarization Benchmark

***BiBench**: Benchmarking and Analyzing Network Binarization*

The 3 Most Effective Techniques for
Generic Binarization:

(1) Soft gradient approximation

(2) Channel-wise scaling factors

(3) Pre-binarization parameter redistributing

BiBench: Benchmarking and Analyzing Network Binarization. Haotong Qin, et al. ICML 2023.

Binarization Benchmark

***BiBench**: Benchmarking and Analyzing Network Binarization*

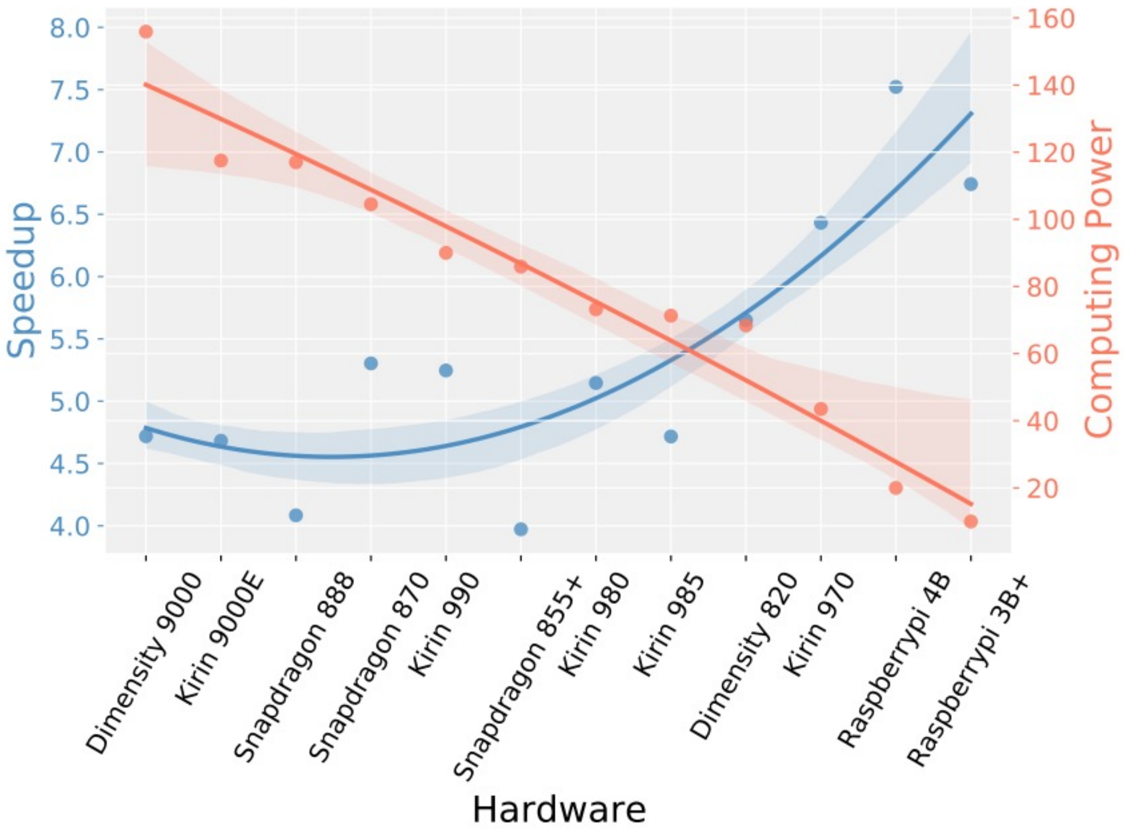
The 3 Most Effective Techniques for Generic Binarization:

(1) Soft gradient approximation

(2) Channel-wise scaling factors

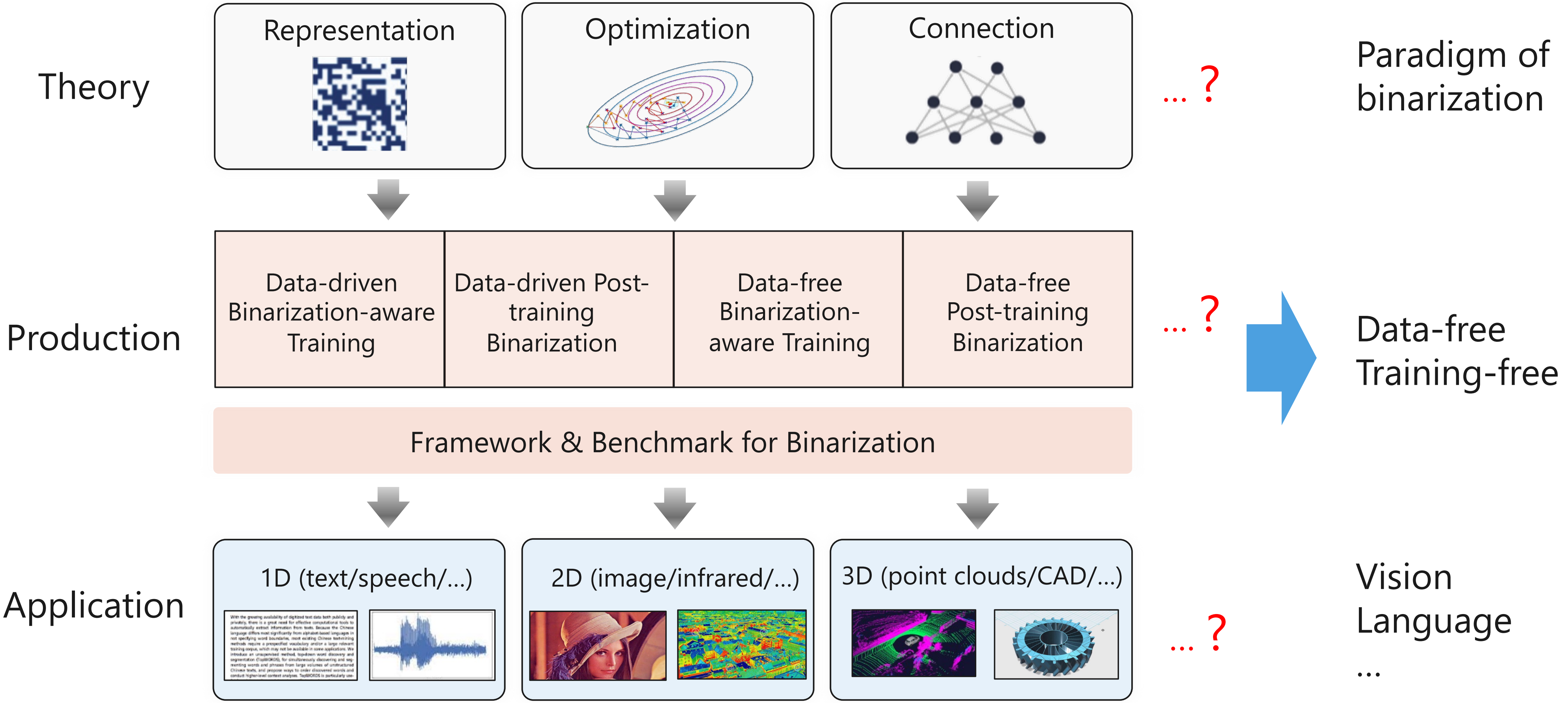
(3) Pre-binarization parameter redistributing

Finding for Binarization: Born for Edge



BiBench: Benchmarking and Analyzing Network Binarization. Haotong Qin, et al. ICML 2023.

Network Binarization: Future



Thank you!

Q&A

Haotong Qin
Beihang University