Network Binarization toward Hardware-friendly Deep Learning

Haotong Qin
Beihang University & ETH Zürich

Haotong Qin

EDUCATION

Ph.D.	SCSE, Beihang University.	2019–Present
Joint Ph.D.	CVL, ETH Zürich.	2022–Present
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

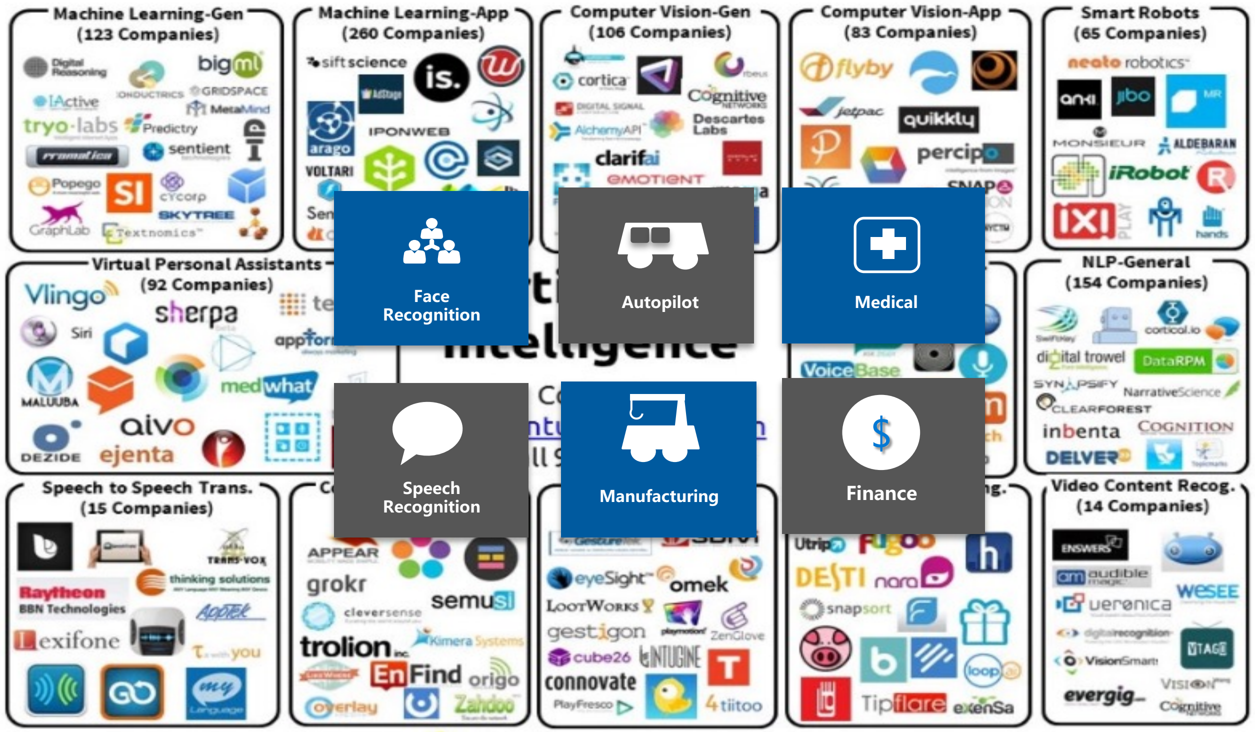
MAIN AWARDS

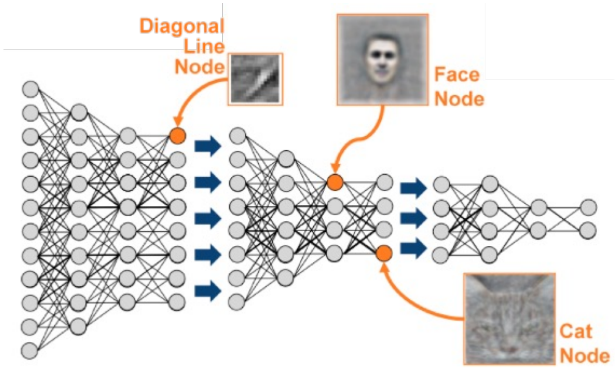
2023	KAUST Rising Stars in AI (18 people worldwide)
2022	ByteDance Scholarship (10 people nationwide)
2022	Beihang Top-10 PhD Students Award
2021&2020	China National Scholarship

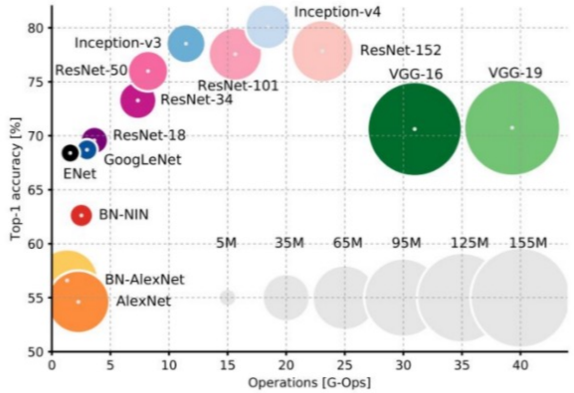
Background

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

…

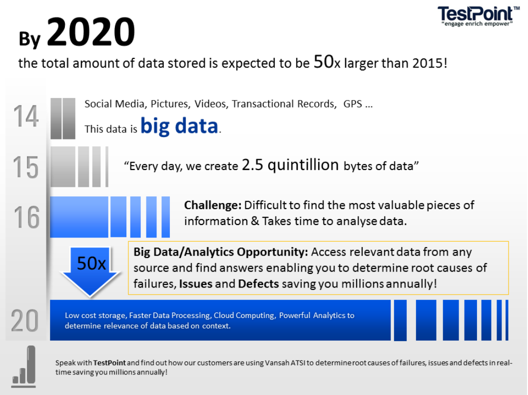


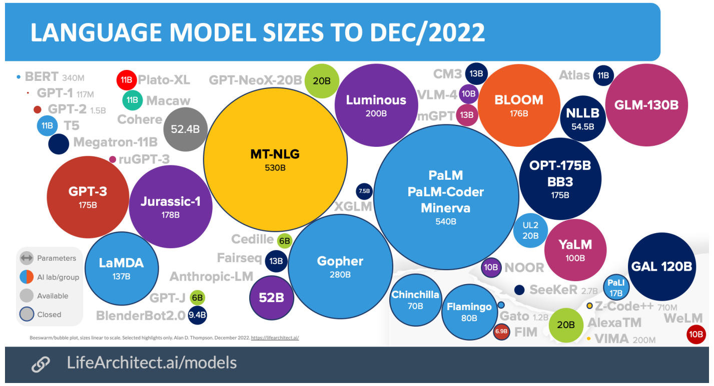


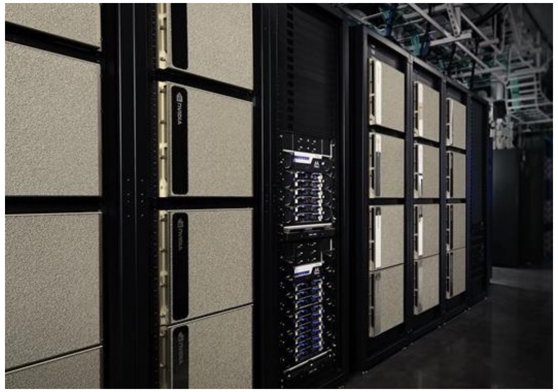


Background

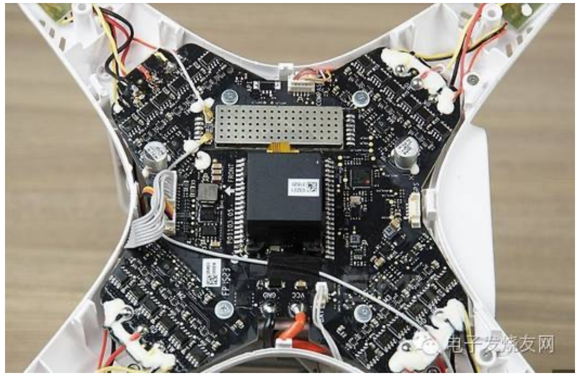
bigger data
and
larger model

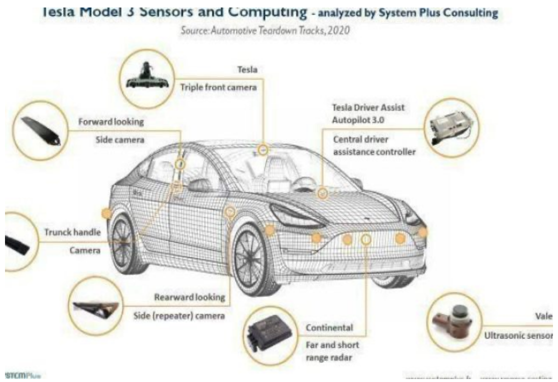






diverse usage
and
limited resources





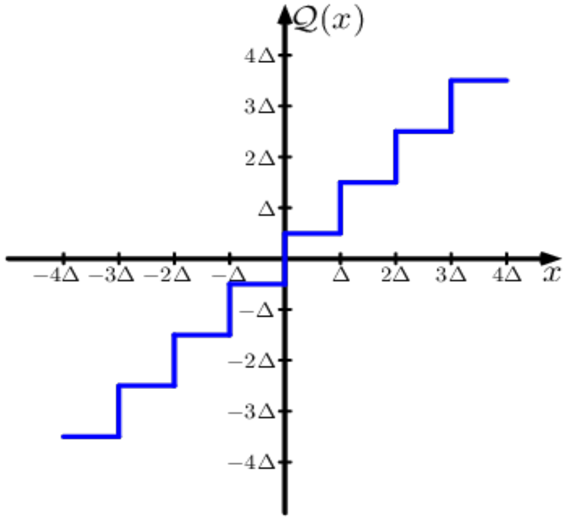
Network Quantization and Binarization

$$\underset{\text{32-bit}}{\mathbb{R}} \longrightarrow \underset{\text{8-bit}}{\mathbb{I}} \longrightarrow \underset{\text{1-bit}}{\mathbb{B}} \in \{-1, +1\}$$

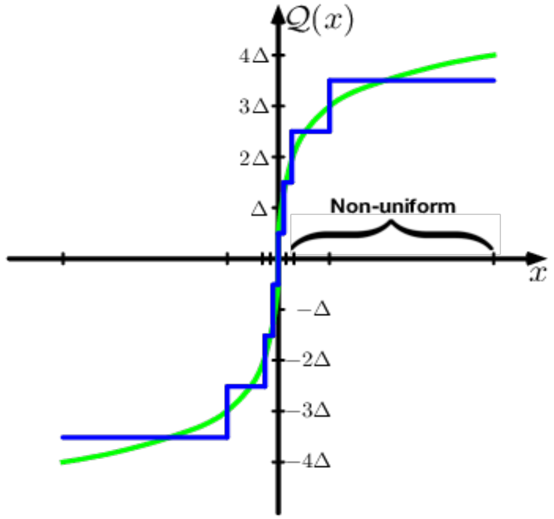
Network Quantization: 2-8 bit

$\mathbb{R} \longrightarrow \mathbb{I}$

32 bit → 2-8 bit



(a)



(b)

Uniform Quantization

$$Q_U(x) = \text{round}\left(\frac{x}{\Delta}\right)\Delta$$

$$\Delta = \frac{u-l}{2^b-1}$$

Non-Uniform Quantization

$$Q_U(x) = \text{round}\left(\frac{\log_2 x}{\Delta}\right)\Delta$$

$$\Delta = \frac{u-l}{2^b-1}$$

Network Binarization: 1-bit

$\mathbb{R} \longrightarrow \mathbb{B}$

32 bit 1 bit

$\{-1, +1\}$

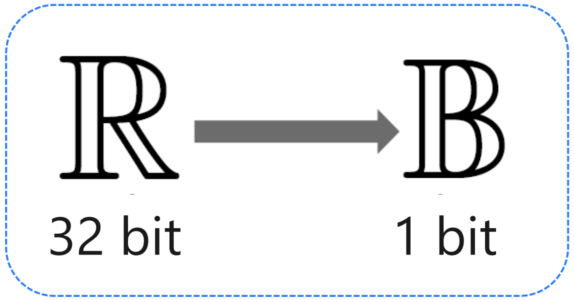
1-Bit Parameters:

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

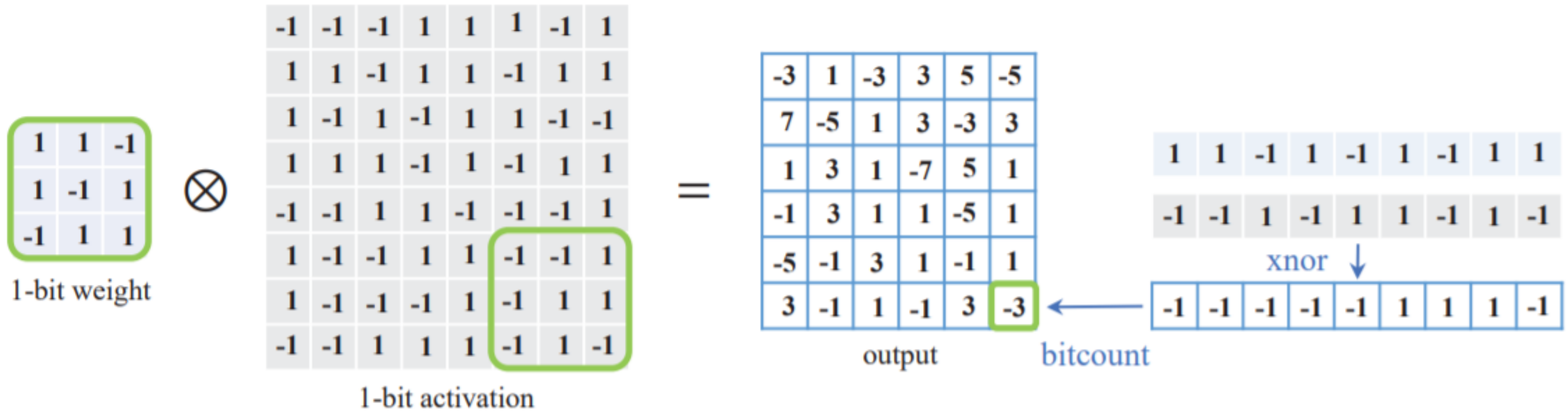
Bitwise Operations:

$$\mathbf{z} = \sigma(Q_w(\mathbf{w}) \otimes Q_a(\mathbf{a})) = \sigma(\alpha\beta(\mathbf{b_w} \odot \mathbf{b_a}))$$

Network Binarization: 1-bit



$\{-1, +1\}$



Network Binarization

Full-Precision Neural Networks

Massive Parameters

Complex Computation

High Power Consumption

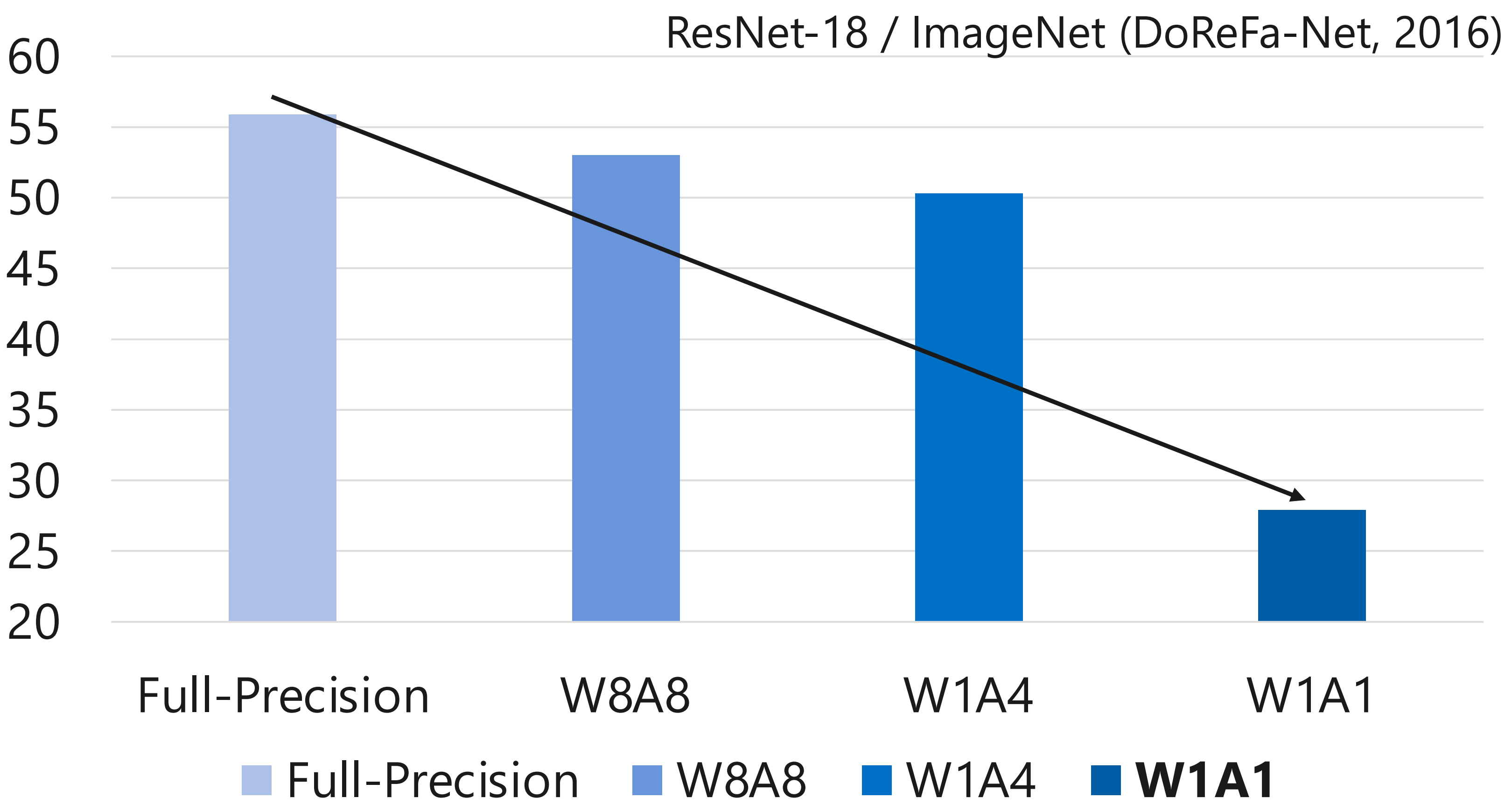
Network Binarization

Full-Precision Neural Networks	Massive Parameters	Complex Computation	High Power Consumption
Binarized Neural Networks	Binarized Parameters	Efficient Instructions	Low Power Consumption

Network Binarization: challenges

Goal: accurate extreme-low bit quantization (binarization)

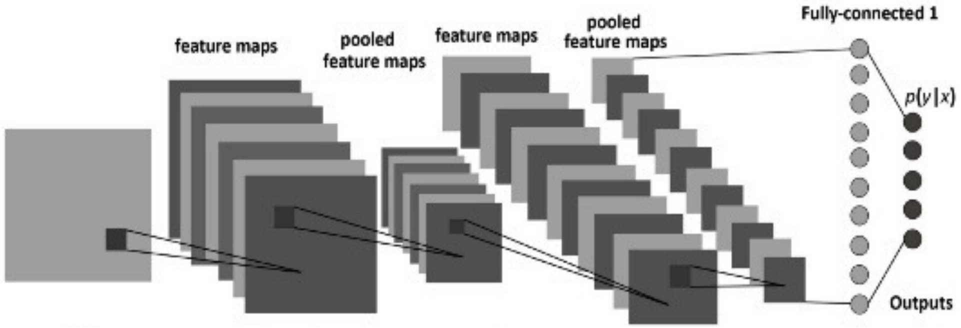
1. the **accuracy** of the binary network has dropped seriously



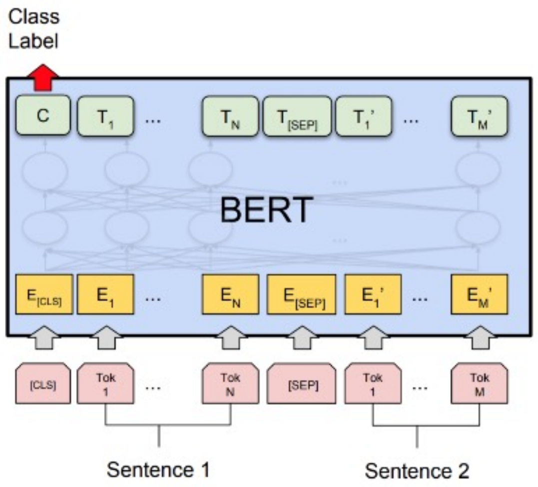
Network Binarization: challenges

Goal: accurate extreme-low bit quantization (binarization)

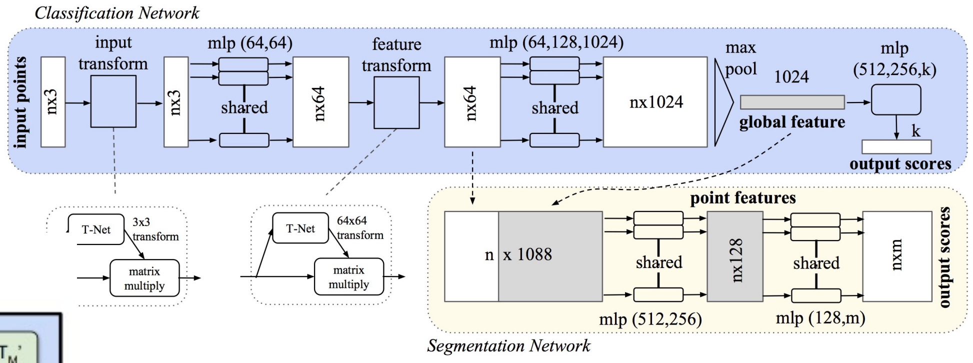
2. binarization methods are not generic across different **architecture**



CNN (ResNet, VGG, ...)



Transformer (BERT, ViT,...)



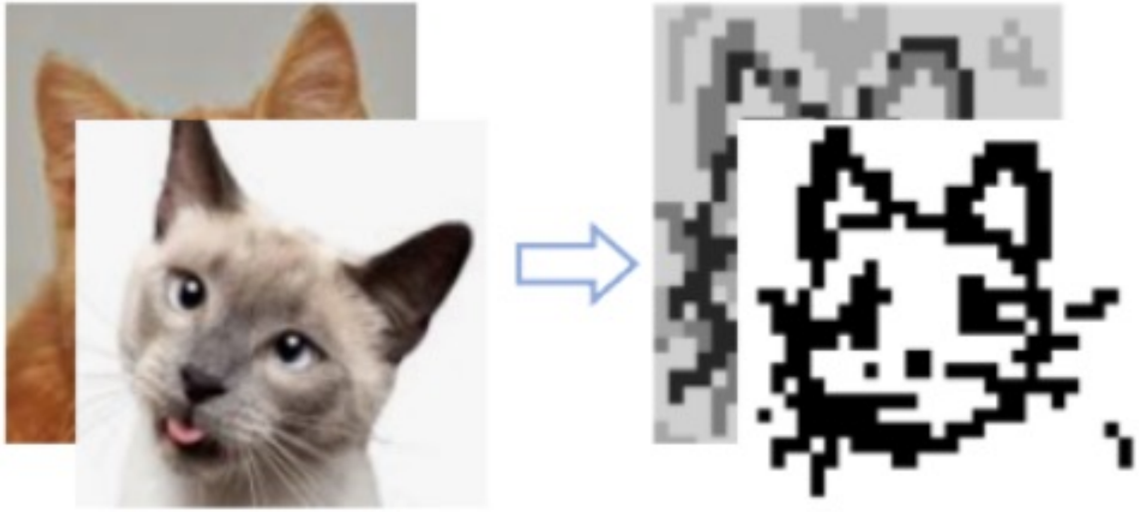
MLP (PointNet, MLP-Mixer, ...)

...

CNN Binarization: information loss and retention

Effects of BNN in the Forward and Backward Propagation

limited representation



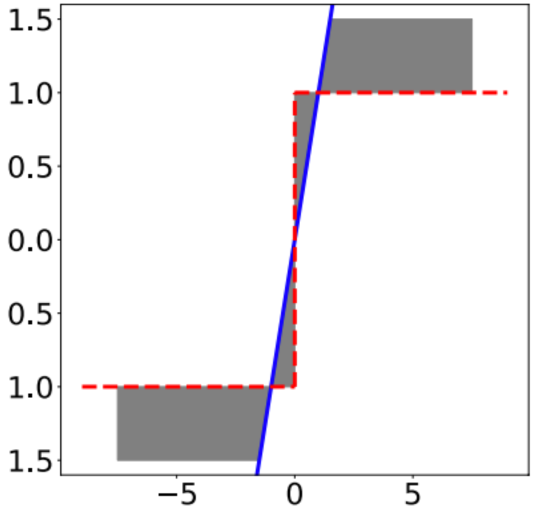
$$\mathbf{B_x} = \mathtt{sign}(\mathbf{x}) = \begin{cases} +1, & if\ \mathbf{x} \geq 0 \\ -1, & otherwise. \end{cases}$$

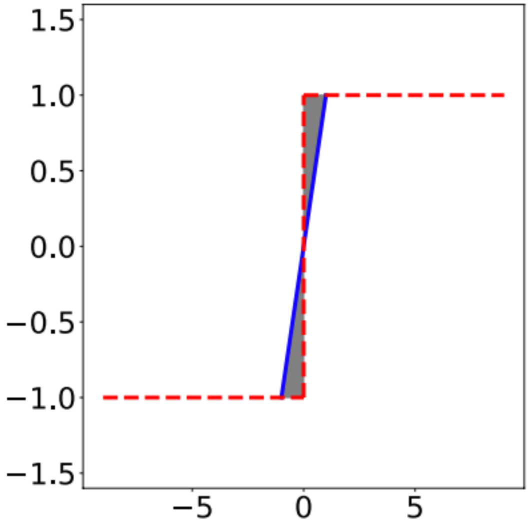
Distribution-sensitive Information Retention for Accurate Binary Neural Network. Haotong Qin, et al. IJCV 2022.

CNN Binarization: information loss and retention

Effects of BNN in the Forward and Backward Propagation

gradient mismatch





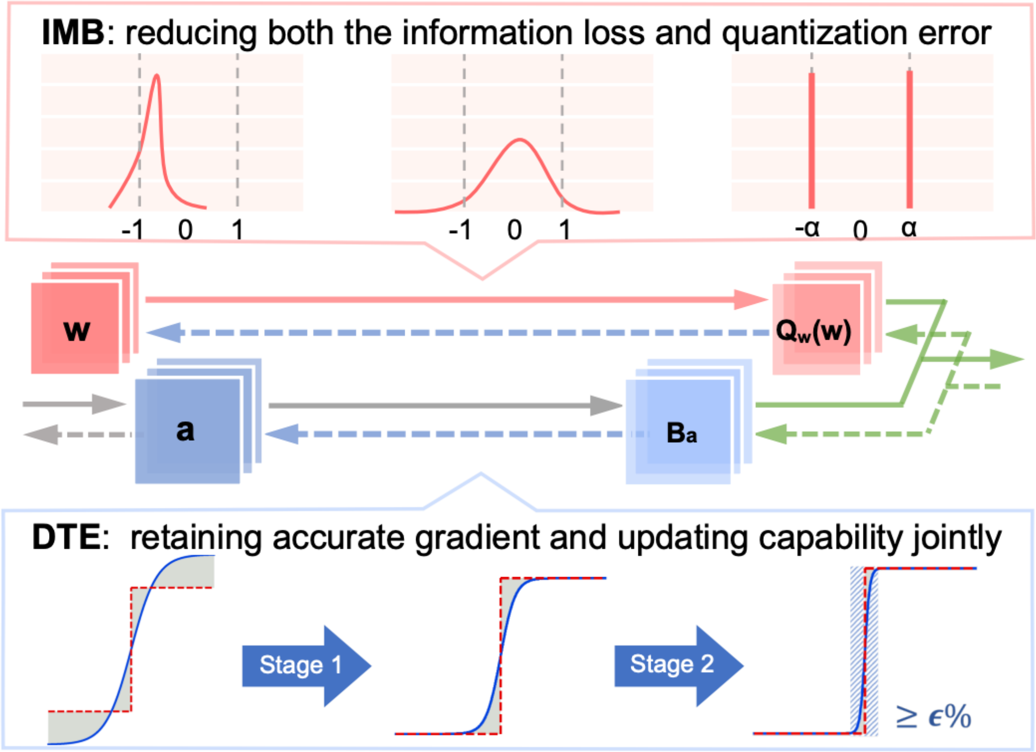
$$\texttt{Identity} : y = x$$

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

Distribution-sensitive Information Retention for Accurate Binary Neural Network. Haotong Qin, et al. IJCV 2022.

CNN Binarization: information loss and retention

Distribution-sensitive Information Retention (DIR-Net)

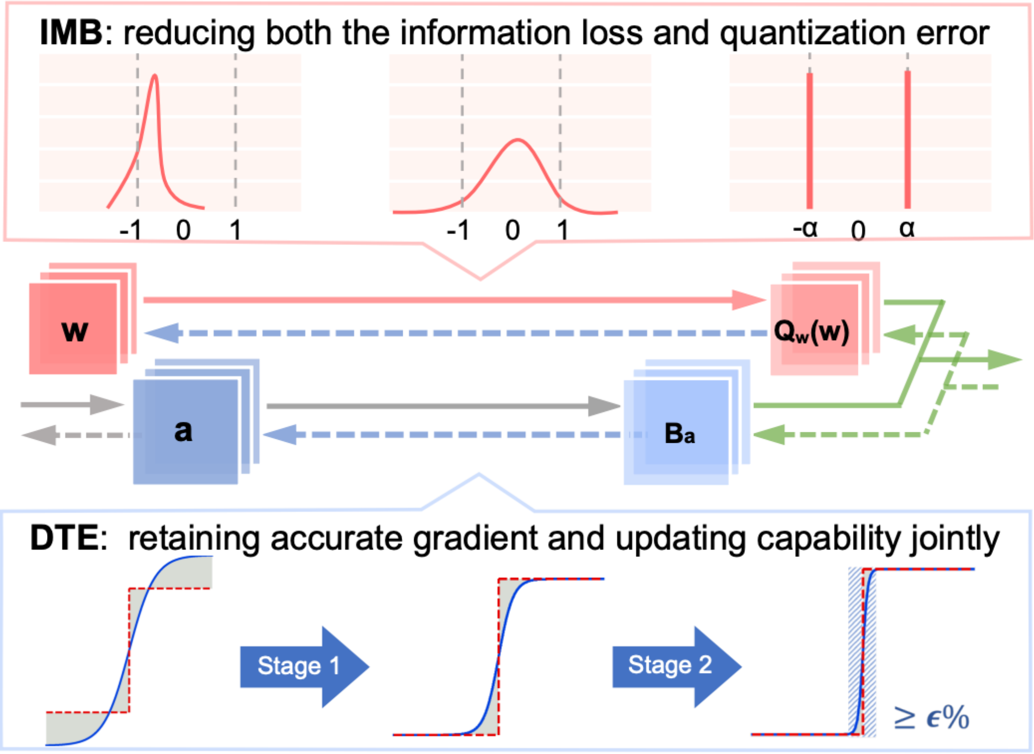


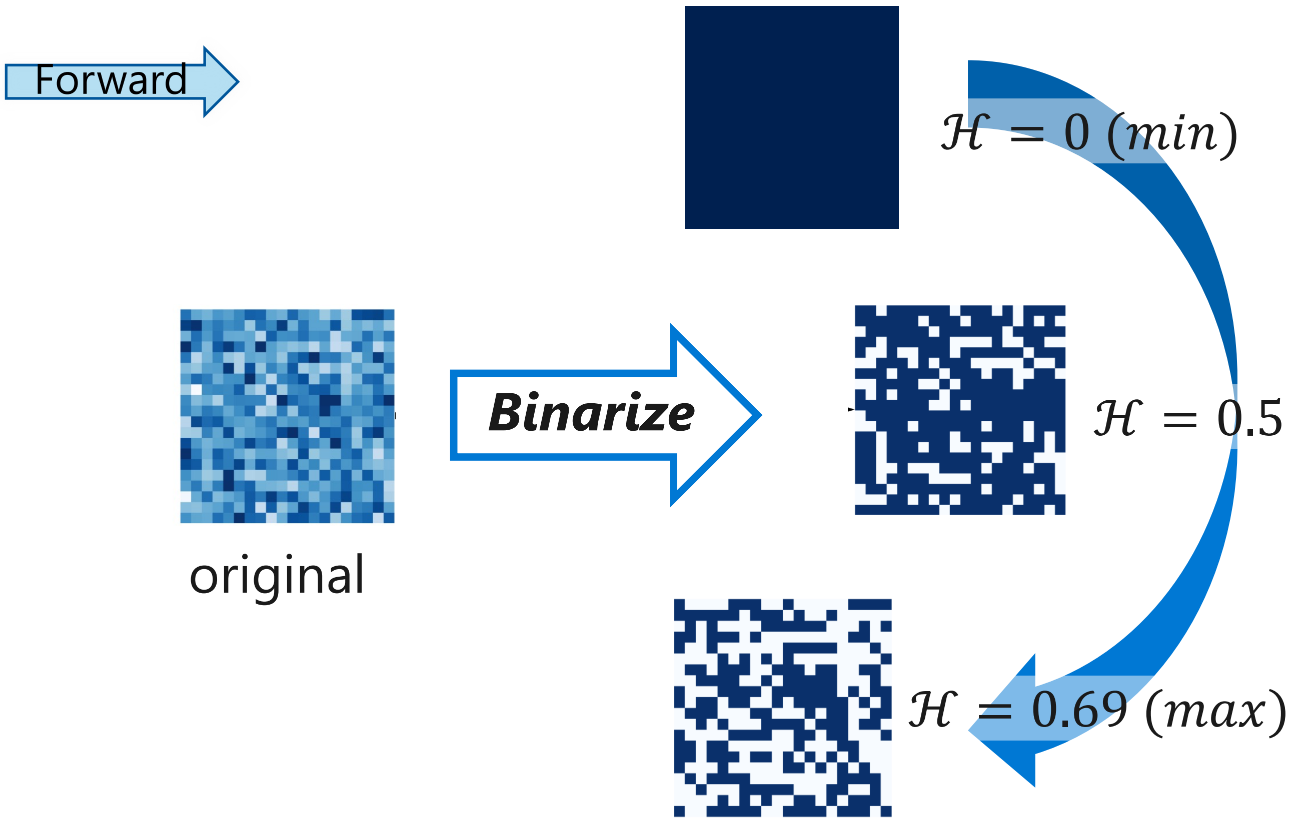
Distribution-sensitive Information Retention for Accurate Binary Neural Network. Haotong Qin, et al. IJCV 2022.

CNN Binarization: information loss and retention

Distribution-sensitive Information Retention (DIR-Net)

Maximizing the information entropy

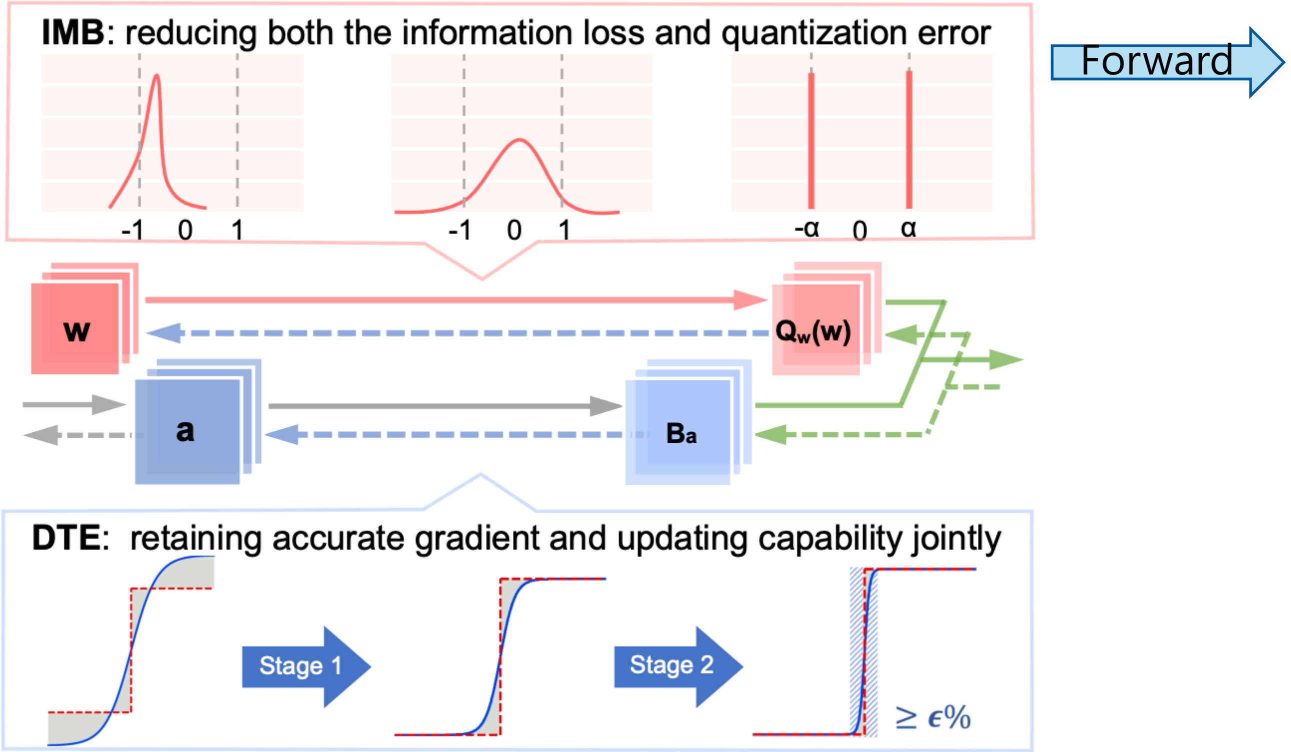




Distribution-sensitive Information Retention for Accurate Binary Neural Network. Haotong Qin, et al. IJCV 2022.

CNN Binarization: information loss and retention

Distribution-sensitive Information Retention (DIR-Net)

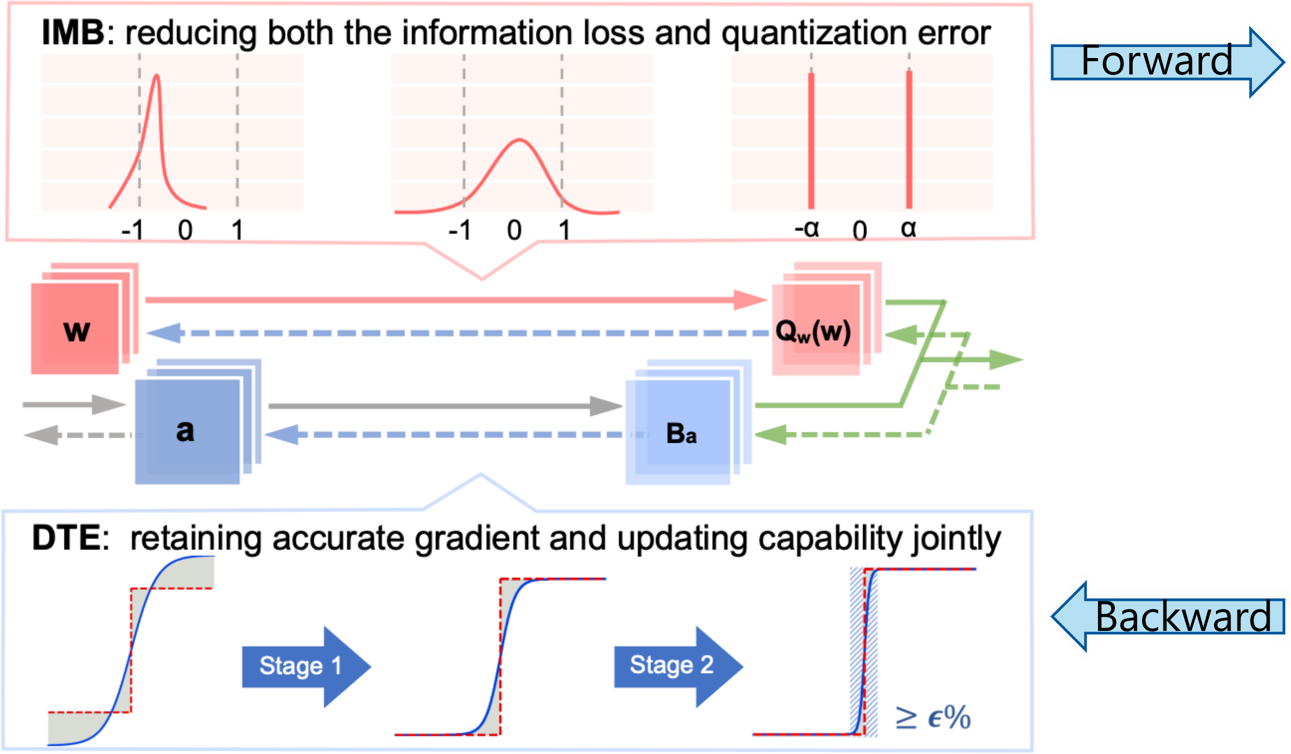


Maximizing the information entropy

Distribution-sensitive Information Retention for Accurate Binary Neural Network. Haotong Qin, et al. IJCV 2022.

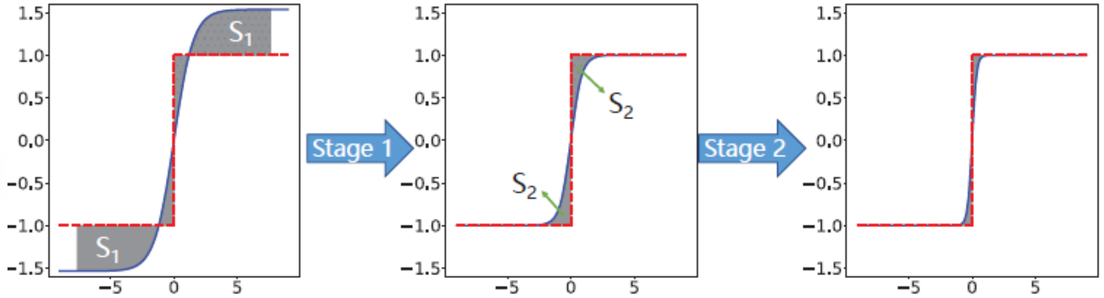
CNN Binarization: information loss and retention

Distribution-sensitive Information Retention (DIR-Net)



Maximizing the information entropy

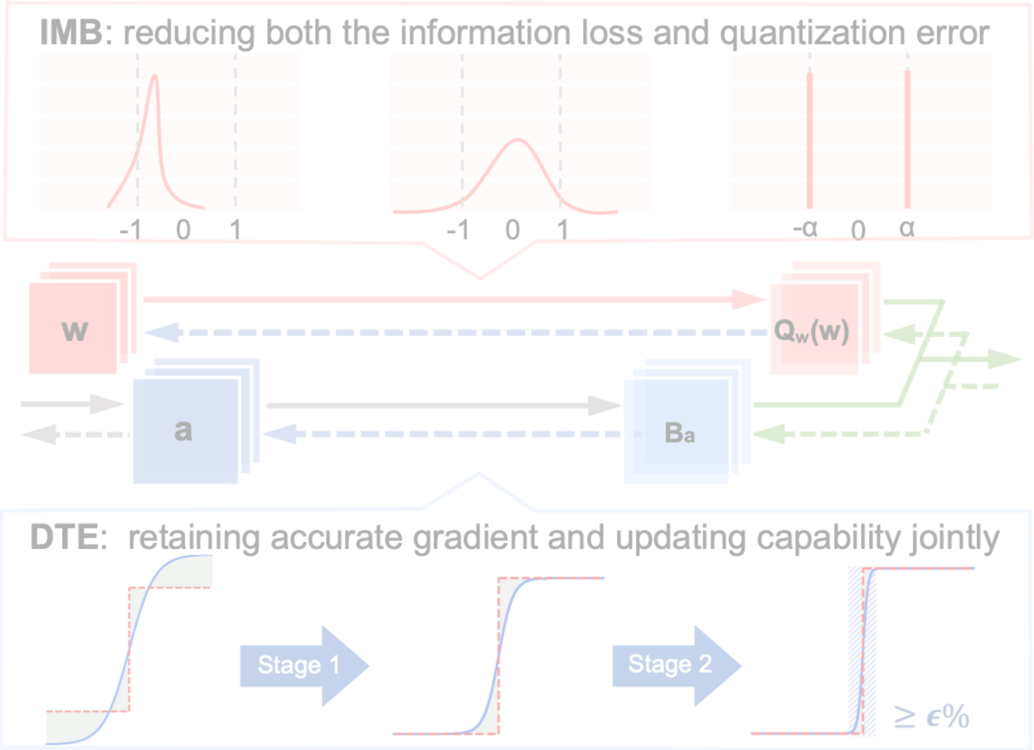
Changing the shape of estimator



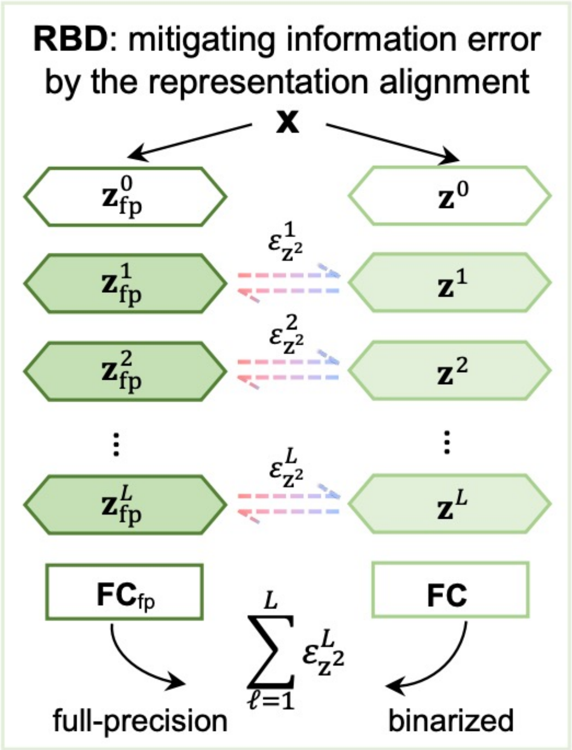
Distribution-sensitive Information Retention for Accurate Binary Neural Network. Haotong Qin, et al. IJCV 2022.

CNN Binarization: information loss and retention

Distribution-sensitive Information Retention (DIR-Net)



external:
binarization-aware
distillation



Distribution-sensitive Information Retention for Accurate Binary Neural Network. Haotong Qin, et al. IJCV 2022.

CNN Binarization: information loss and retention

Performance

ResNet-34	Full-Precision	32/32	73.3	91.3
	ABC-Net	1/1	52.4	76.5
	Bi-Real	1/1	62.2	83.9
	IR-Net	1/1	62.9	84.1
	Si-BNN	1/1	63.3	84.4
	ReActNet	1/1	67.3	87.9
	DIR-Net[1] (ours)	1/1	64.1	85.3
	DIR-Net[2] (ours)	1/1	**67.9**$_{\pm 0.09}$	**88.2**
	Full-Precision	32/32	73.3	91.3
	ABC-Net	1/32	68.8	86.1
	Bi-Real	1/32	69.7	88.9
	Si-BNN	1/32	70.1	89.7
	IR-Net	1/32	70.4	89.5
	DIR-Net (ours)	1/32	**71.1**$_{\pm 0.03}$	**90.4**
DARTS	Full-Precision	32/32	73.3	91.3
	BNN	1/1	52.2	76.6
	Bi-Real	1/1	61.5	83.8
	IR-Net	1/1	62.1	84.2
	ReActNet	1/1	65.1	86.4
	DIR-Net[1] (ours)	1/1	63.3	85.1
	DIR-Net[2] (ours)	1/1	**65.6**$_{\pm 0.12}$	**87.2**

The accuracy reached **90%** of the full precision ResNet

Distribution-sensitive Information Retention for Accurate Binary Neural Network. Haotong Qin, et al. IJCV 2022.

CNN Binarization: information loss and retention

Performance

ResNet-34	Full-Precision	32/32	73.3	91.3
	ABC-Net	1/1	52.4	76.5
	Bi-Real	1/1	62.2	83.9
	IR-Net	1/1	62.9	84.1
	Si-BNN	1/1	63.3	84.4
	ReActNet	1/1	67.3	87.9
	DIR-Net[1] (ours)	1/1	64.1	85.3
	DIR-Net[2] (ours)	1/1	**67.9**$_{\pm 0.09}$	**88.2**
	Full-Precision	32/32	73.3	91.3
	ABC-Net	1/32	68.8	86.1
	Bi-Real	1/32	69.7	88.9
	Si-BNN	1/32	70.1	89.7
	IR-Net	1/32	70.4	89.5
	DIR-Net (ours)	1/32	**71.1**$_{\pm 0.03}$	**90.4**
DARTS	Full-Precision	32/32	73.3	91.3
	BNN	1/1	52.2	76.6
	Bi-Real	1/1	61.5	83.8
	IR-Net	1/1	62.1	84.2
	ReActNet	1/1	65.1	86.4
	DIR-Net[1] (ours)	1/1	63.3	85.1
	DIR-Net[2] (ours)	1/1	**65.6**$_{\pm 0.12}$	**87.2**

The accuracy reached **90%** of the full precision ResNet

Accurate binarization on **lightweight** architectures

Distribution-sensitive Information Retention for Accurate Binary Neural Network. Haotong Qin, et al. IJCV 2022.

CNN Binarization: information loss and retention

Performance

	Full-Precision	32/32	73.3	91.3
	ABC-Net	1/1	52.4	76.5
	Bi-Real	1/1	62.2	83.9
	IR-Net	1/1	62.9	84.1
	Si-BNN	1/1	63.3	84.4
	ReActNet	1/1	67.3	87.9
	DIR-Net[1] (ours)	1/1	64.1	85.3
ResNet-34	DIR-Net[2] (ours)	1/1	**67.9**$_{\pm 0.09}$	**88.2**
	Full-Precision	32/32	73.3	91.3
	ABC-Net	1/32	68.8	86.1
	Bi-Real	1/32	69.7	88.9
	Si-BNN	1/32	70.1	89.7
	IR-Net	1/32	70.4	89.5
	DIR-Net (ours)	1/32	**71.1**$_{\pm 0.03}$	**90.4**
	Full-Precision	32/32	73.3	91.3
	BNN	1/1	52.2	76.6
	Bi-Real	1/1	61.5	83.8
DARTS	IR-Net	1/1	62.1	84.2
	ReActNet	1/1	65.1	86.4
	DIR-Net[1] (ours)	1/1	63.3	85.1
	DIR-Net[2] (ours)	1/1	**65.6**$_{\pm 0.12}$	**87.2**

Table 5: Comparison of time cost of ResNet-18 with different bits (single thread).

Method	Bit-width (W/A)	Size (Mb)	Time (ms)
FP	32/32	46.77	1418.94
NCNN	8/8	–	935.51
DSQ	2/2	–	551.22
Ours (without bit-shift scales)	1/1	**4.20**	**252.16**
Ours	1/1	**4.21**	**261.98**

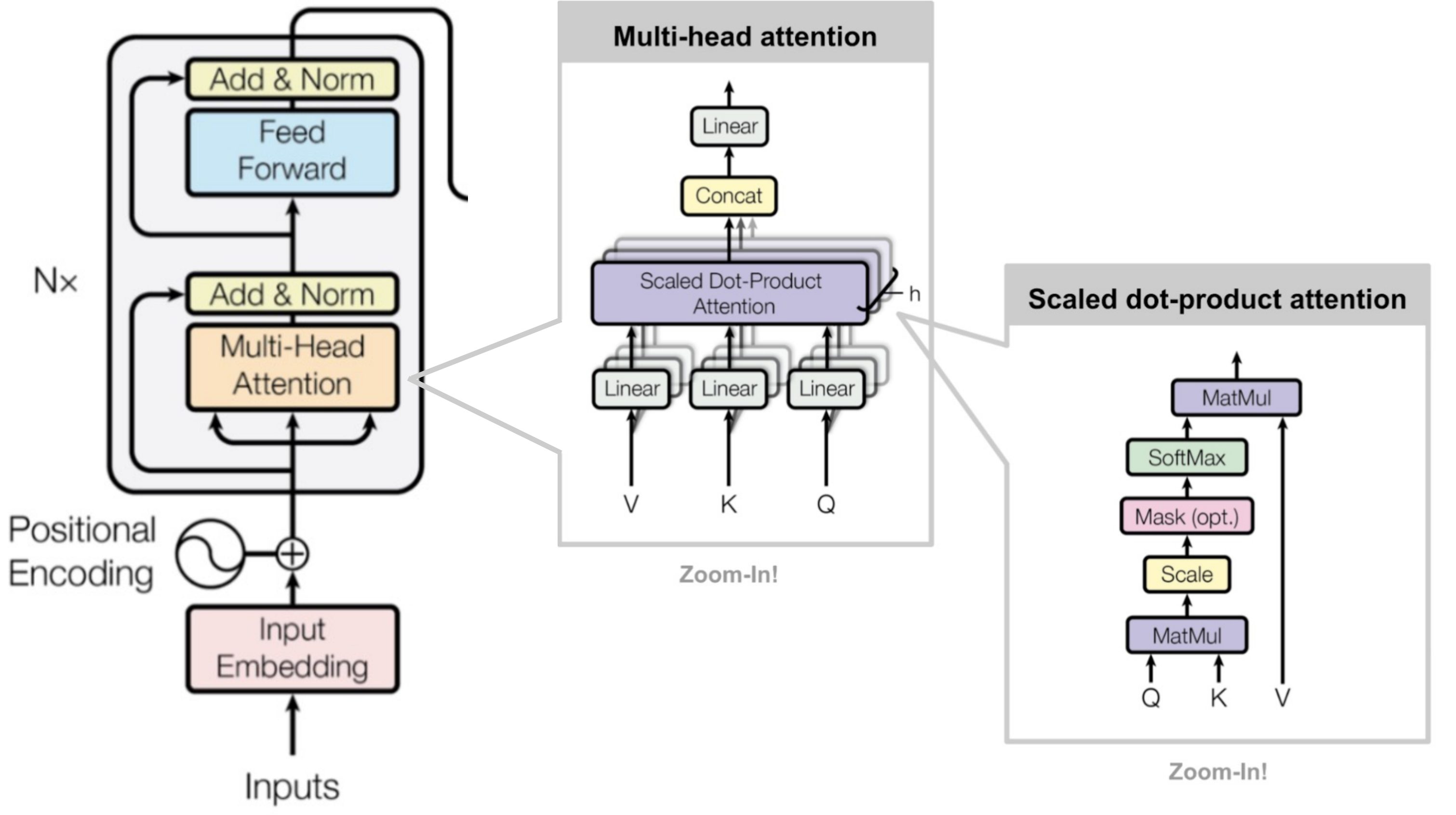
11.1x storage saving

5.4x speedup

Distribution-sensitive Information Retention for Accurate Binary Neural Network. Haotong Qin, et al. IJCV 2022.

Transformer Binarization: attention crash and recovery

Bottlenecks of Fully Binarized BERT Baseline

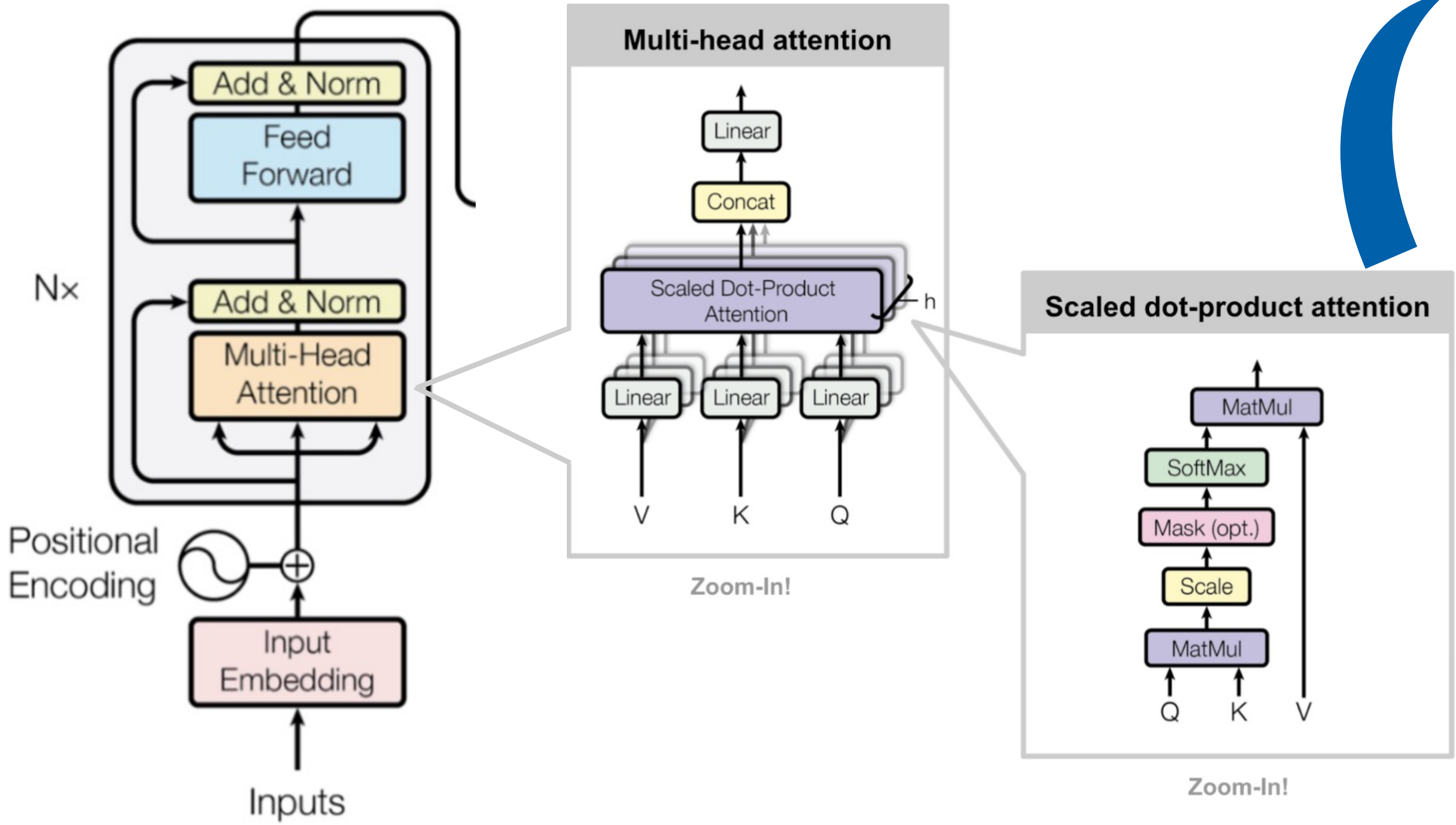


https://deepfrench.gitlab.io/deep-learning-project/

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

Transformer Binarization: attention crash and recovery

Bottlenecks of Fully Binarized BERT Baseline



Binarize (directly)

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

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

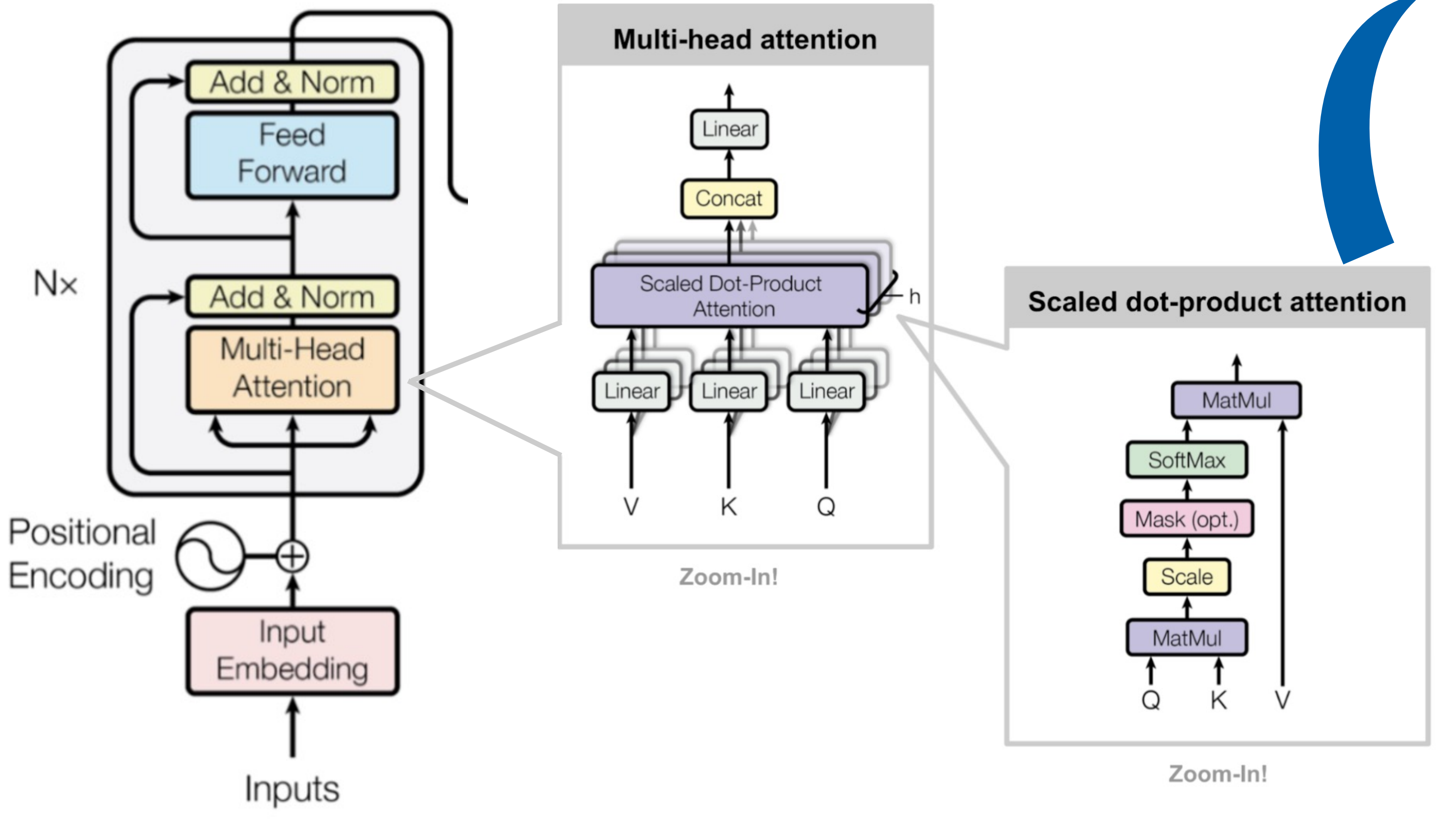
$$\mathbf{B}^s_{\mathbf{A}} = \mathrm{sign}(\mathrm{softmax}(\mathbf{A}))$$

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

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

Transformer Binarization: attention crash and recovery

Bottlenecks of Fully Binarized BERT Baseline



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

Binarize (directly)

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

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

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

Severely dropped!

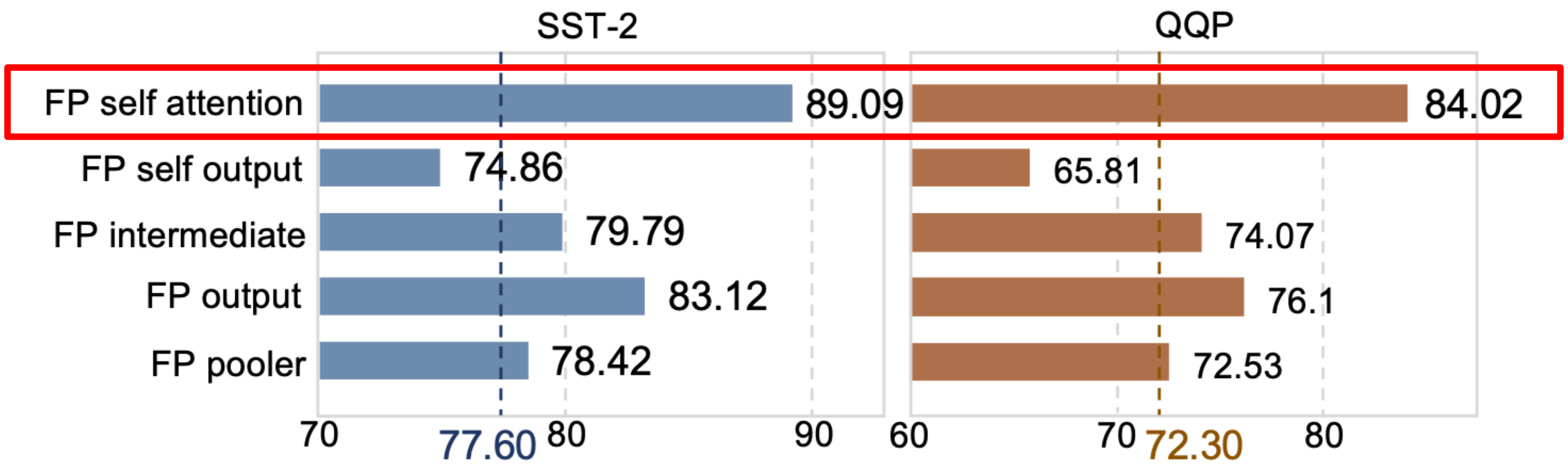
(Avg: 83.9% -> **50.4%**)

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

Transformer Binarization: attention crash and recovery

Bottlenecks of Fully Binarized BERT Baseline

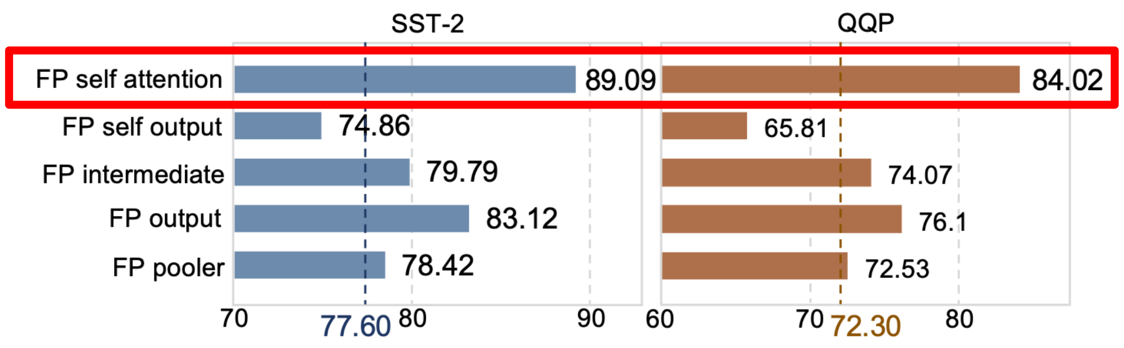
Which part caused the biggest drop?

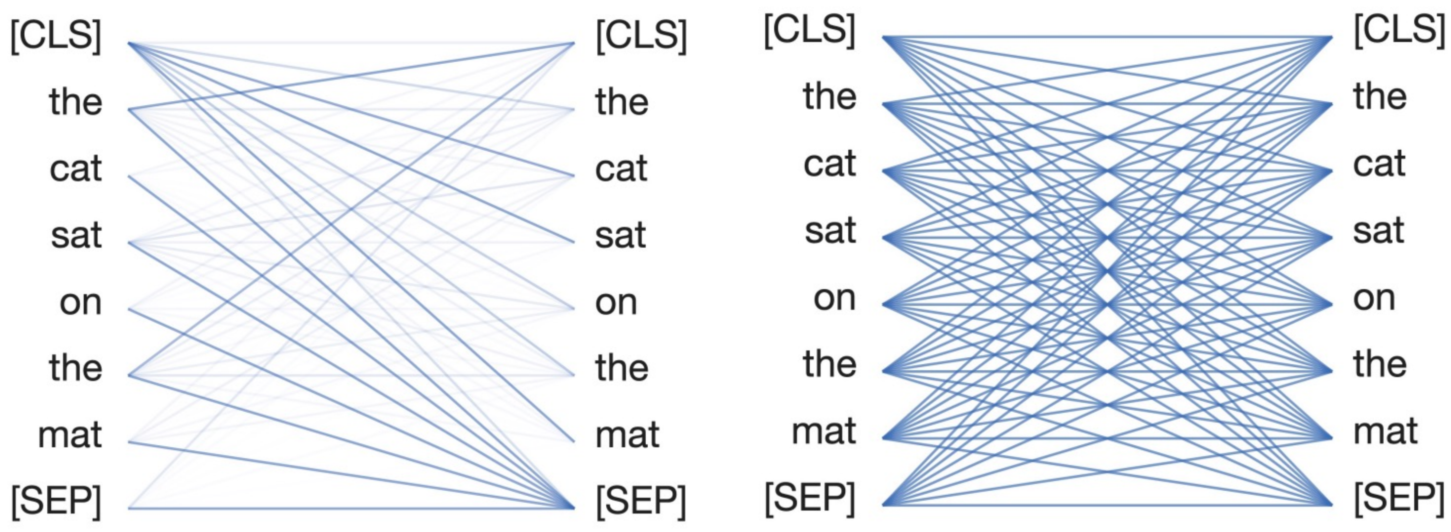


BiBERT: Accurate Fully Binarized BERT. Haotong Qin, et al. ICLR 2022.

Transformer Binarization: attention crash and recovery

Bottlenecks of Fully Binarized BERT Baseline





(a) Full-precision

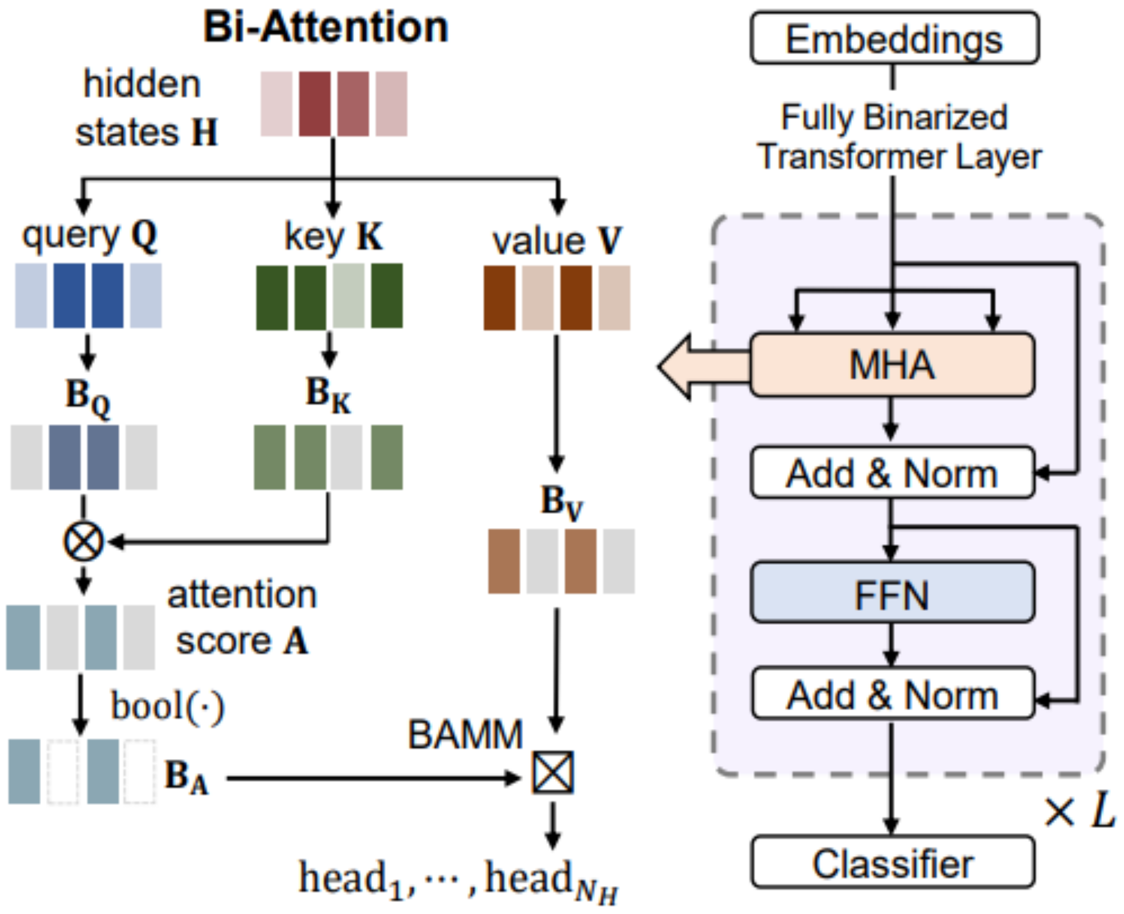
(b) Fully binarized BERT baseline

attention mechanism crashed

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

Transformer Binarization: attention crash and recovery

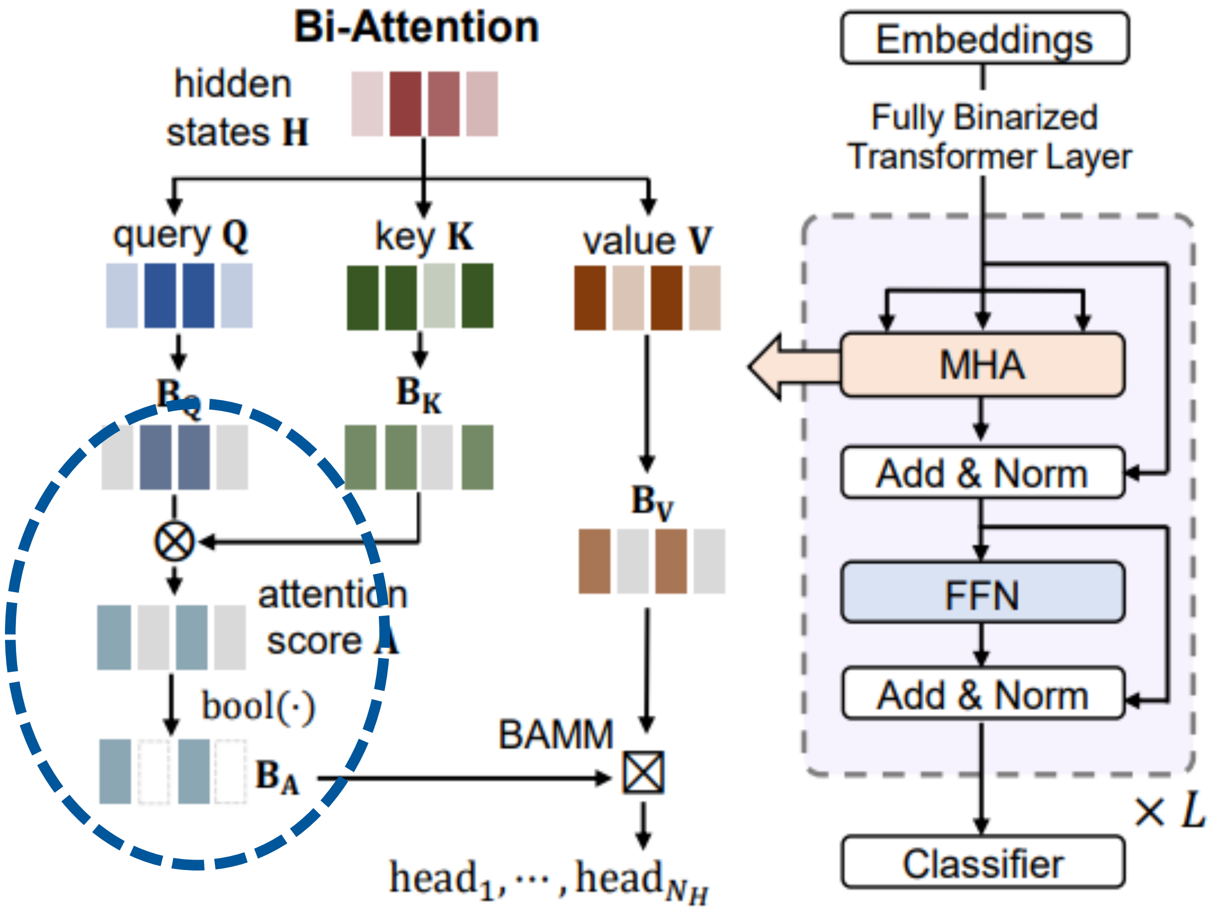
Accurate Fully Binarized BERT (BiBERT)



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

Transformer Binarization: attention crash and recovery

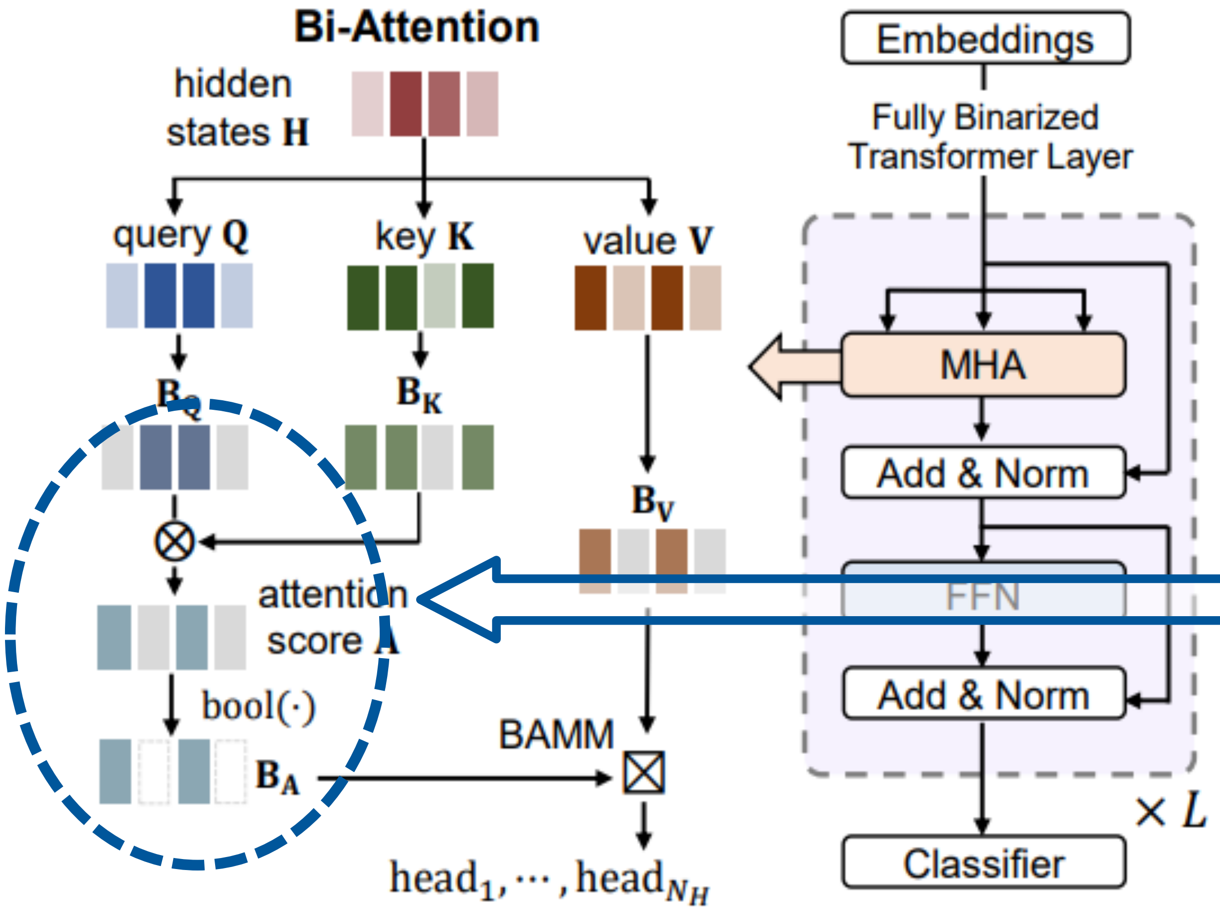
Accurate Fully Binarized BERT (BiBERT)



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

Transformer Binarization: attention crash and recovery

Accurate Fully Binarized BERT (BiBERT)



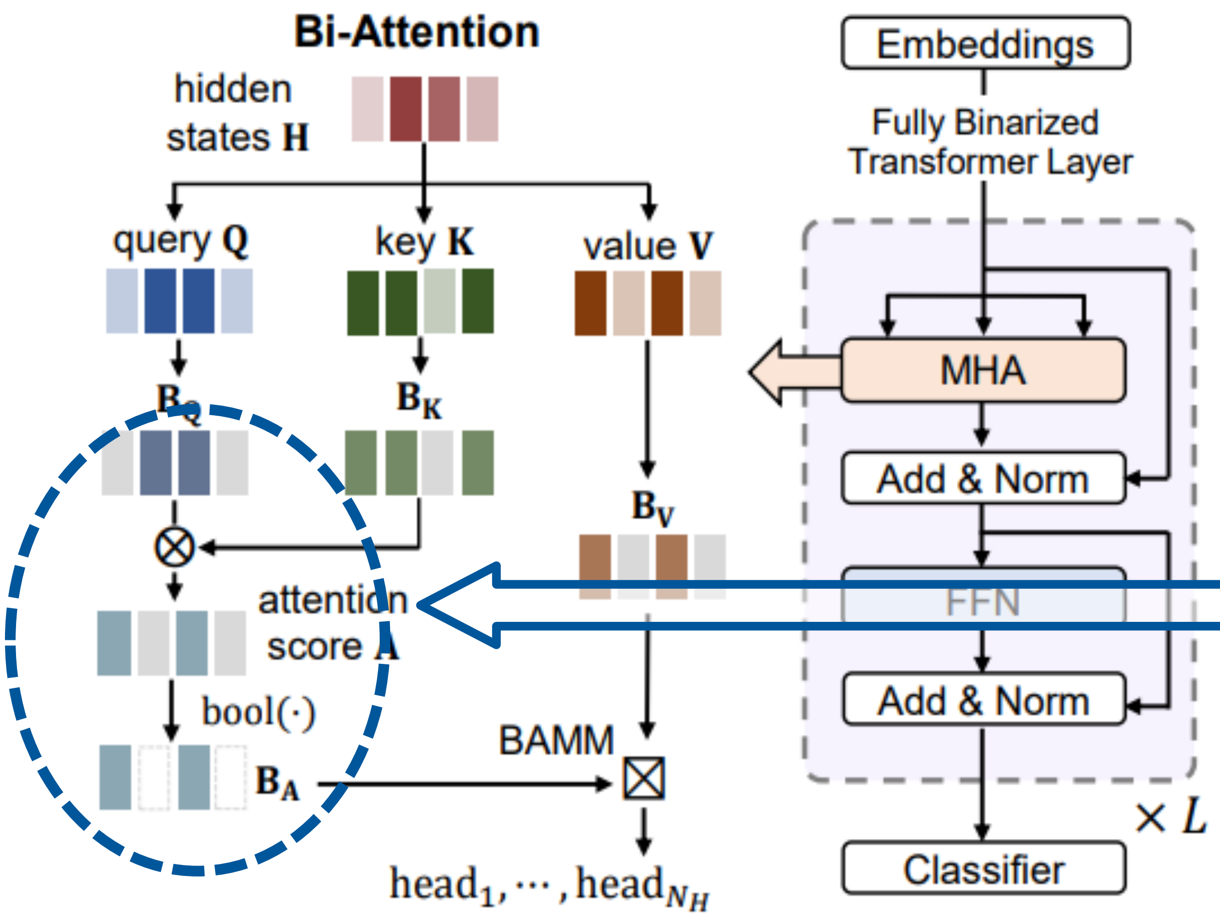
$$\mathrm{bool}(x) = \begin{cases} 1, & \text{if } x \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

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

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

Transformer Binarization: attention crash and recovery

Accurate Fully Binarized BERT (BiBERT)



1. ~~SoftMax~~

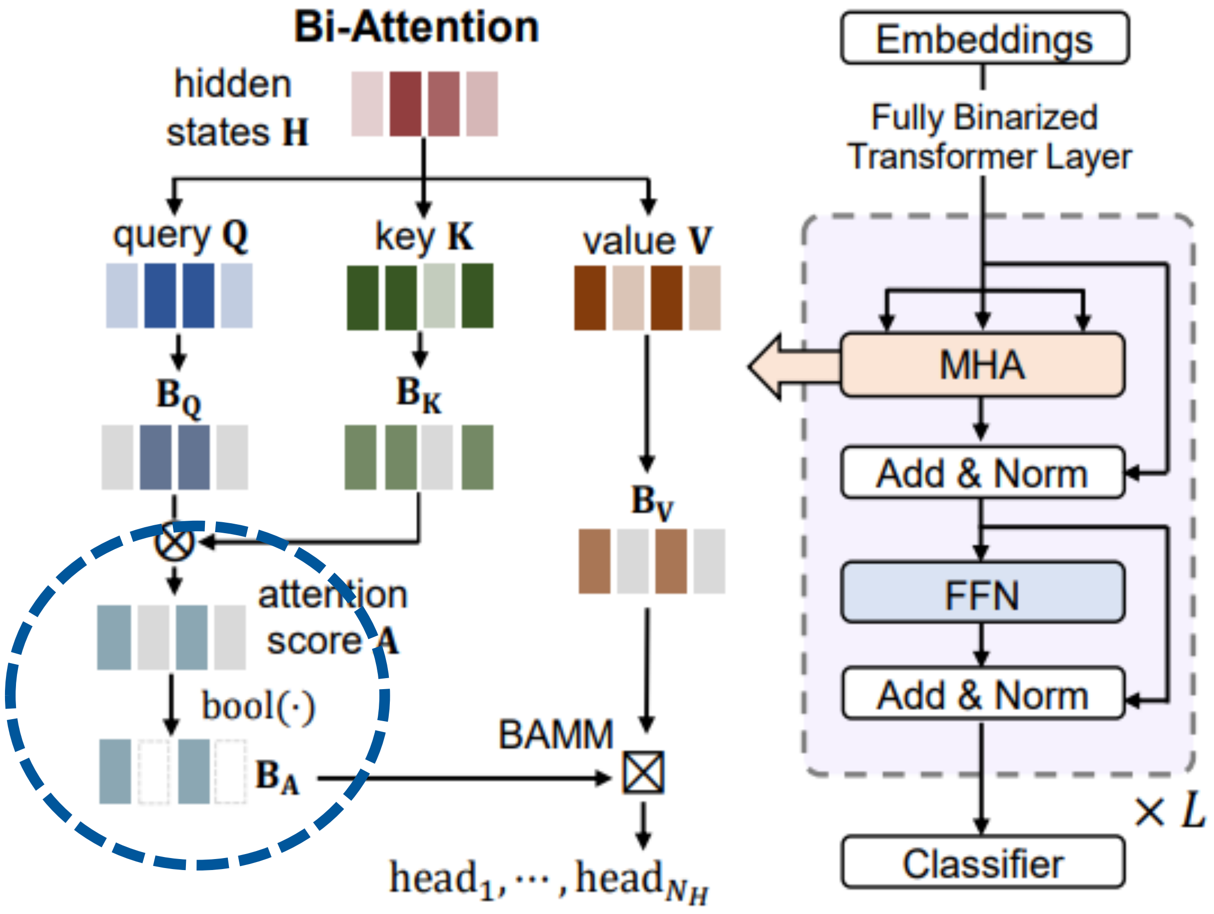
$$\mathrm{bool}(x) = \begin{cases} 1, & \text{if } x \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

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

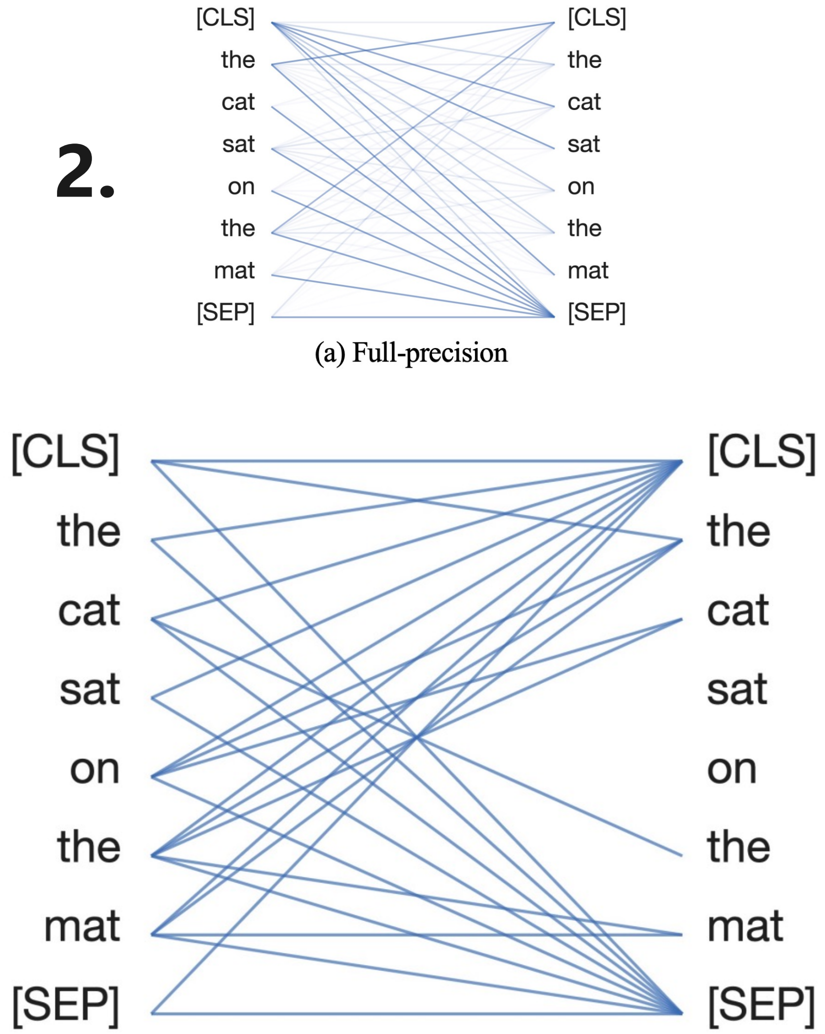
BiBERT: Accurate Fully Binarized BERT. Haotong Qin, et al. ICLR 2022.

Transformer Binarization: attention crash and recovery

Accurate Fully Binarized BERT (BiBERT)



2.



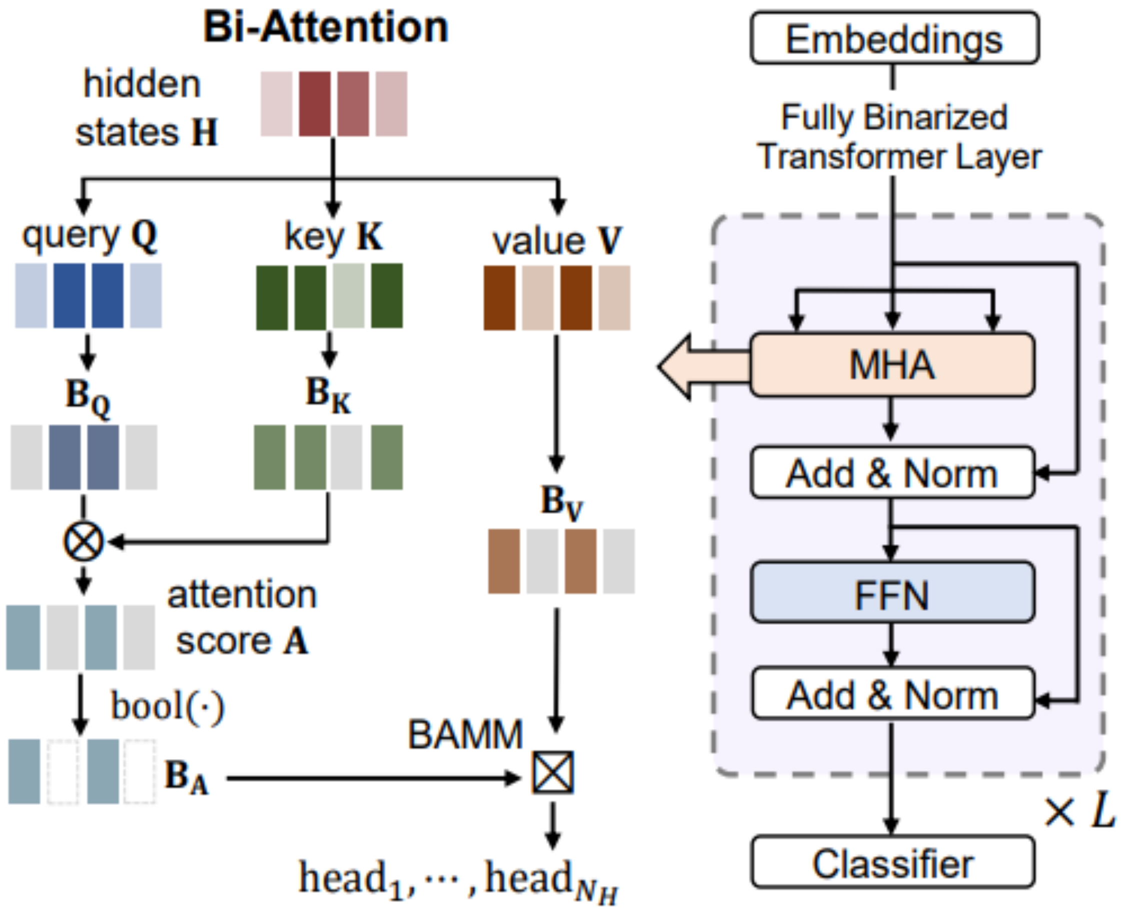
(a) Full-precision

(c) BiBERT (Ours)

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

Transformer Binarization: attention crash and recovery

Accurate Fully Binarized BERT (BiBERT)



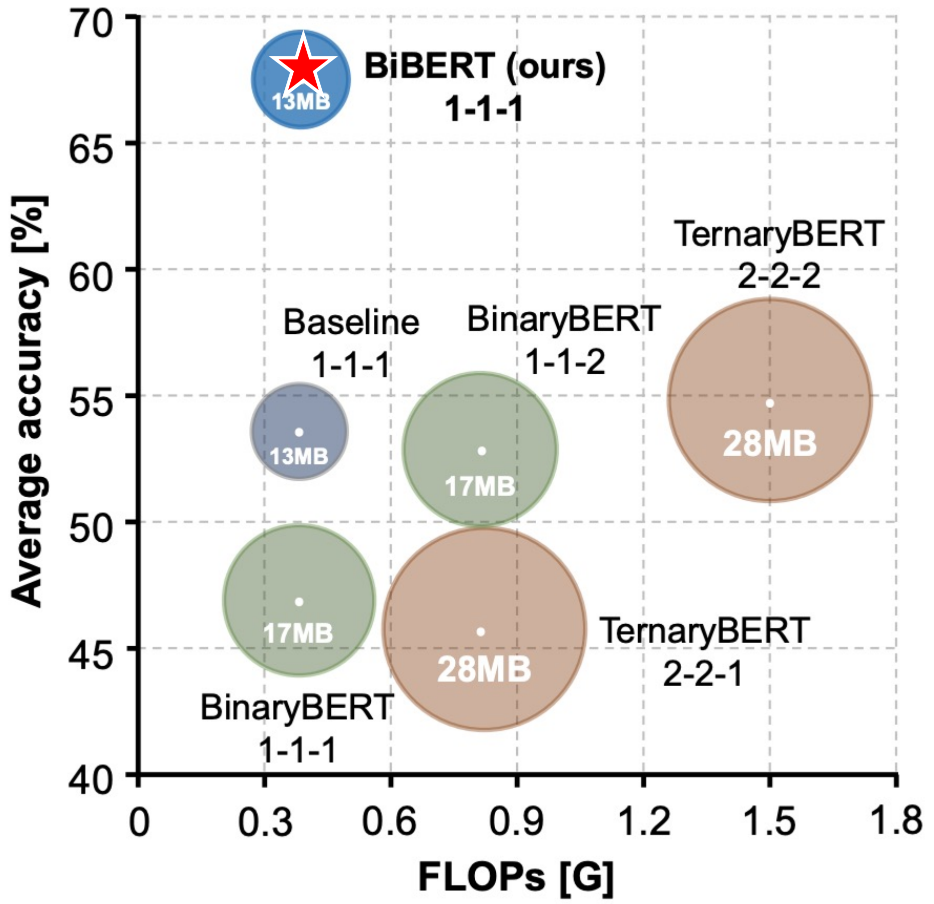
$$\mathbf{B_A} = \mathrm{bool}\left(\mathbf{A}\right) = \mathrm{bool}\left(\frac{1}{\sqrt{D}}\left(\mathbf{B_Q} \otimes \mathbf{B_K}^\top\right)\right)$$

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

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

Transformer Binarization: attention crash and recovery

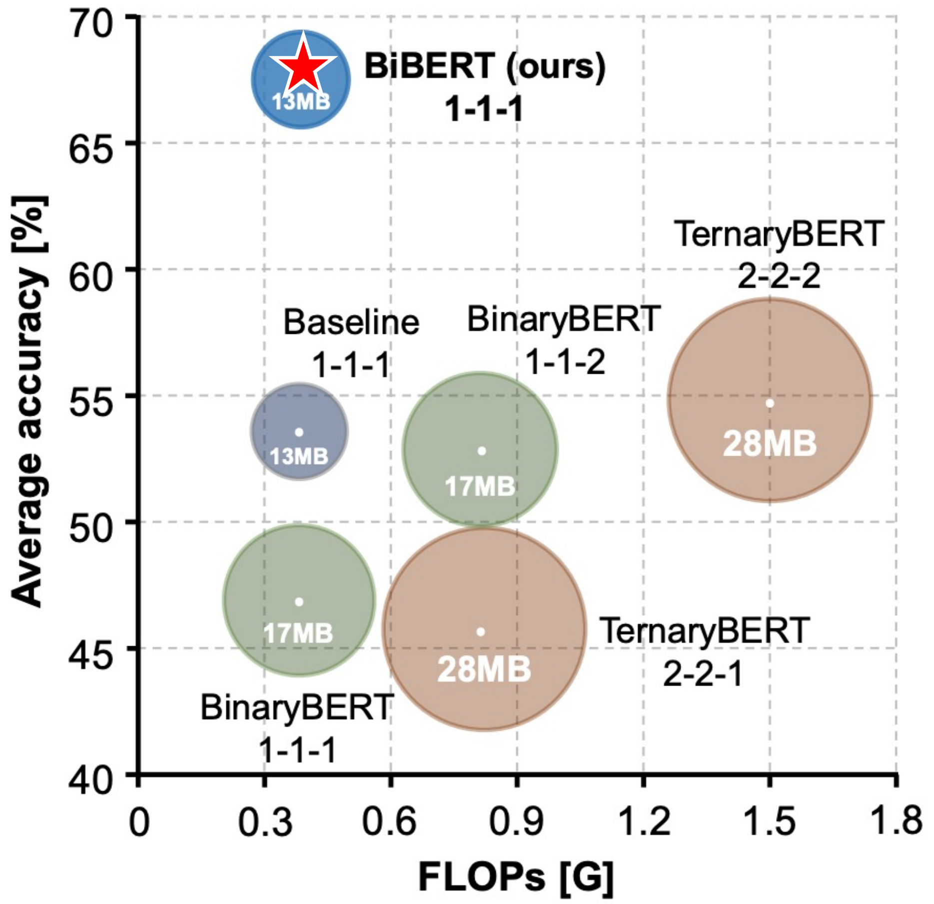
Performance

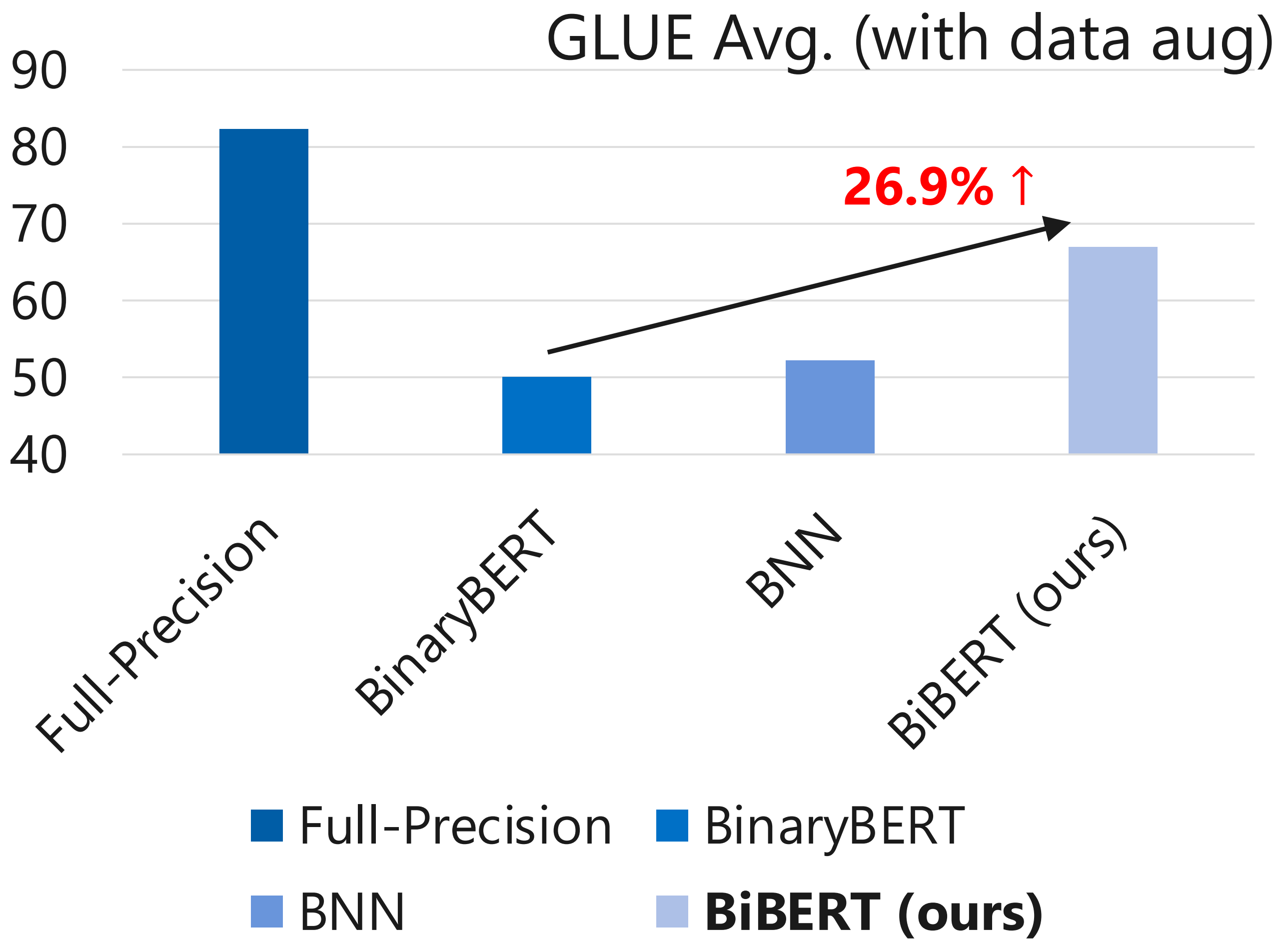


BiBERT: Accurate Fully Binarized BERT. Haotong Qin, et al. ICLR 2022.

Transformer Binarization: attention crash and recovery

Performance





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

Network Binarization: benchmark

Challenges in Existing Binarization Research

1. Confusing contributions (operators? structures?)

2. Limited comparisons (methods? architectures?)

3. Neglected practicality (hardware deployment?)

Network 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

Network 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)

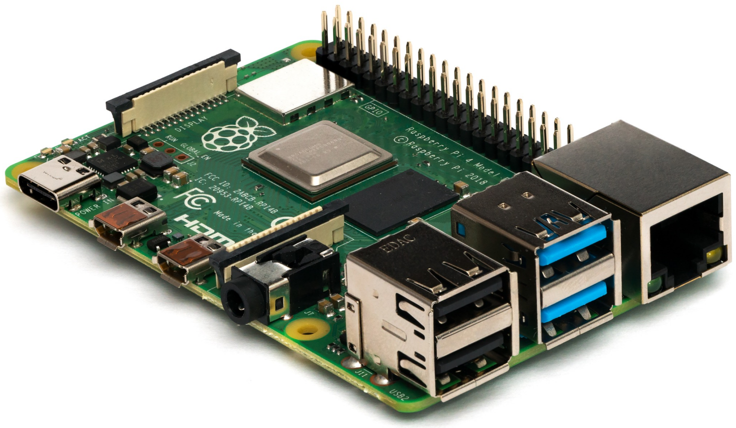
Network Binarization: benchmark

Challenges in Existing Binarization Research

1. Confusing contributions (operators? structures?)

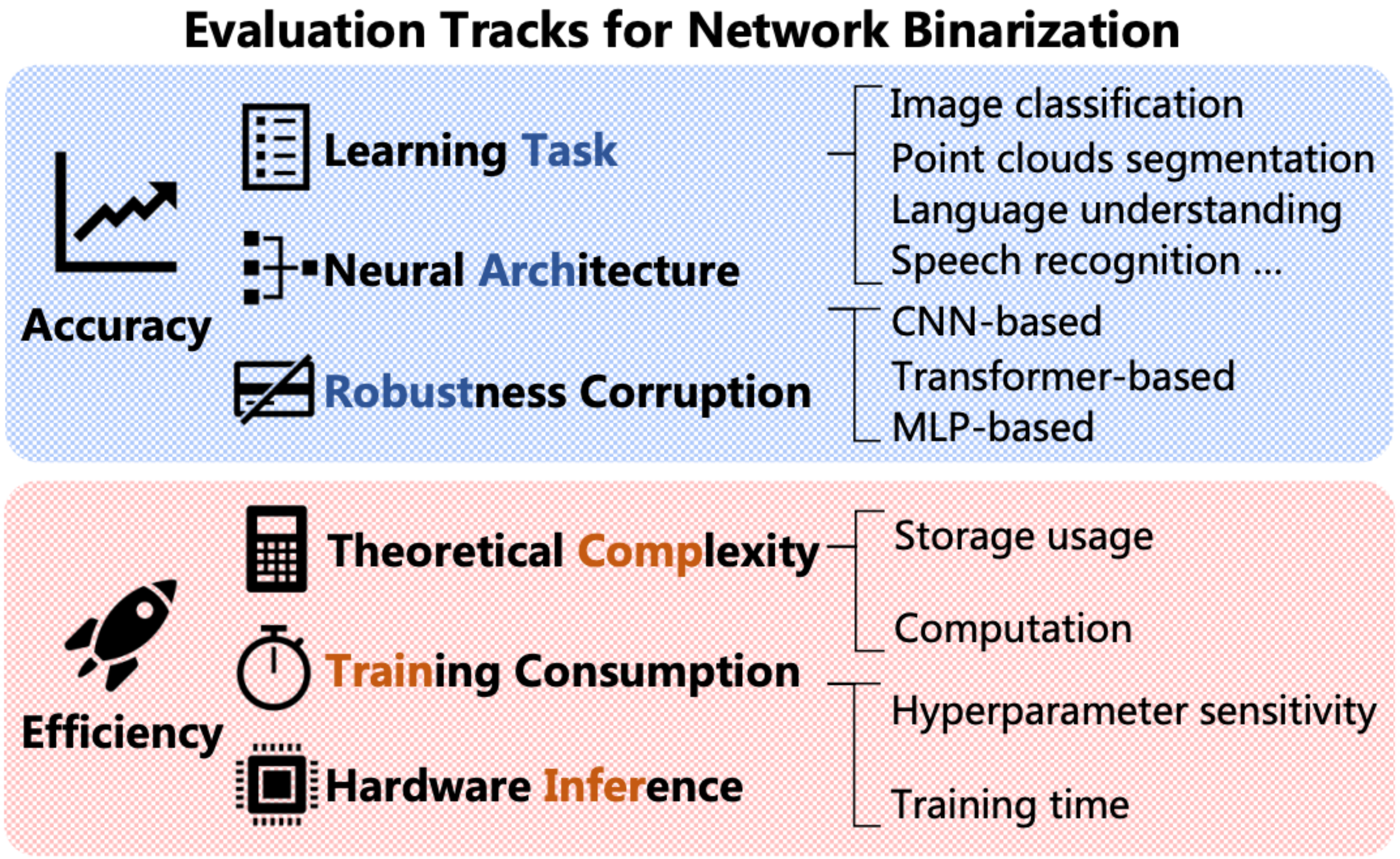
2. Limited comparisons (methods? architectures?)

3. Neglected practicality (hardware deployment?)



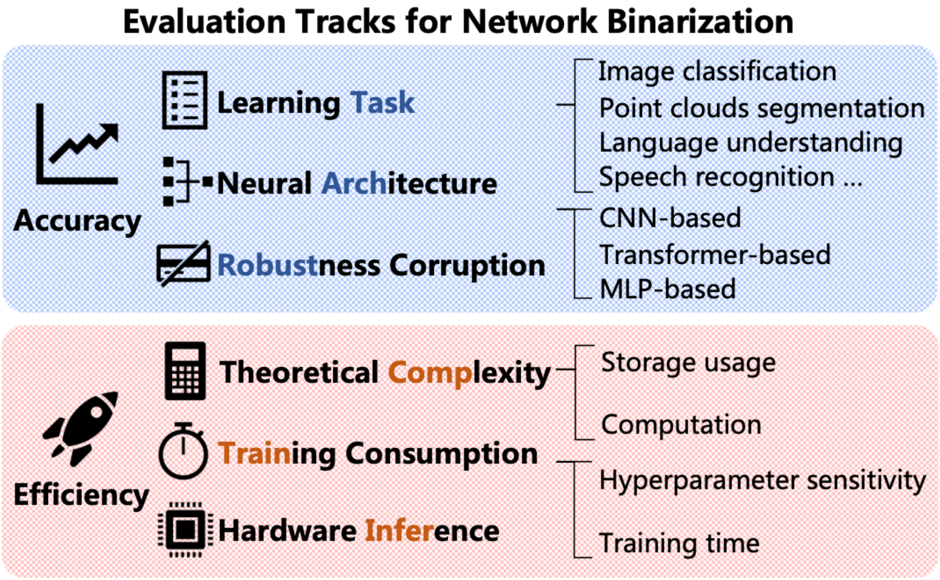
Network Binarization: benchmark

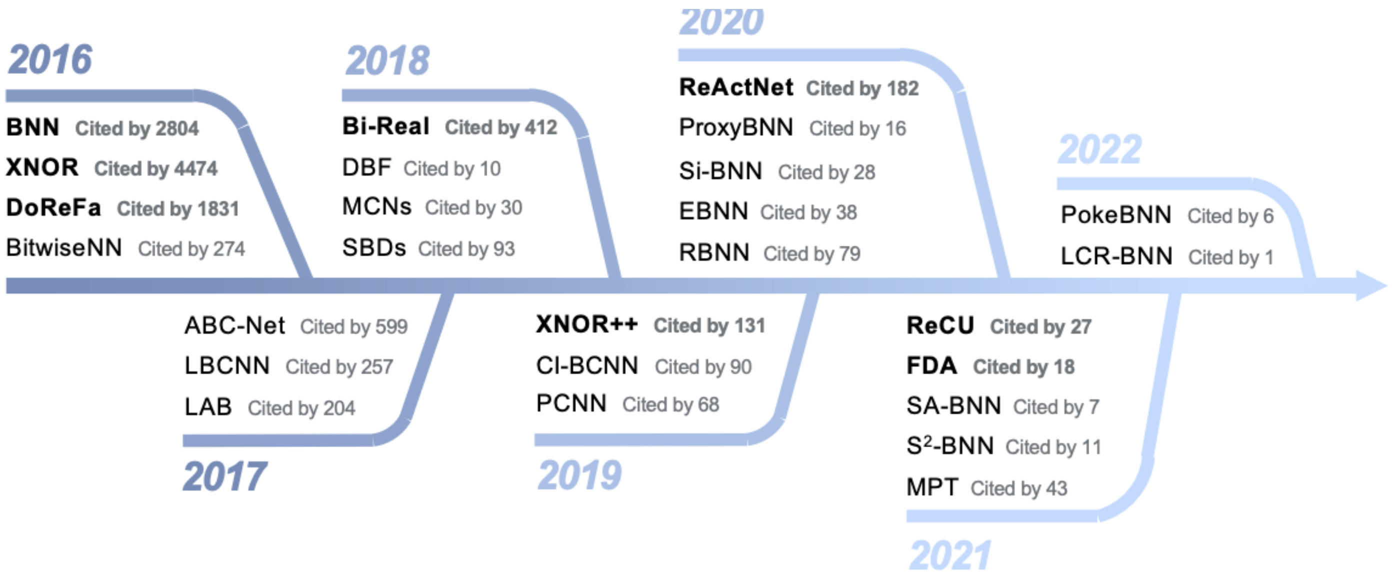
***BiBench**: Benchmarking and Analyzing Network Binarization*



Network Binarization: benchmark

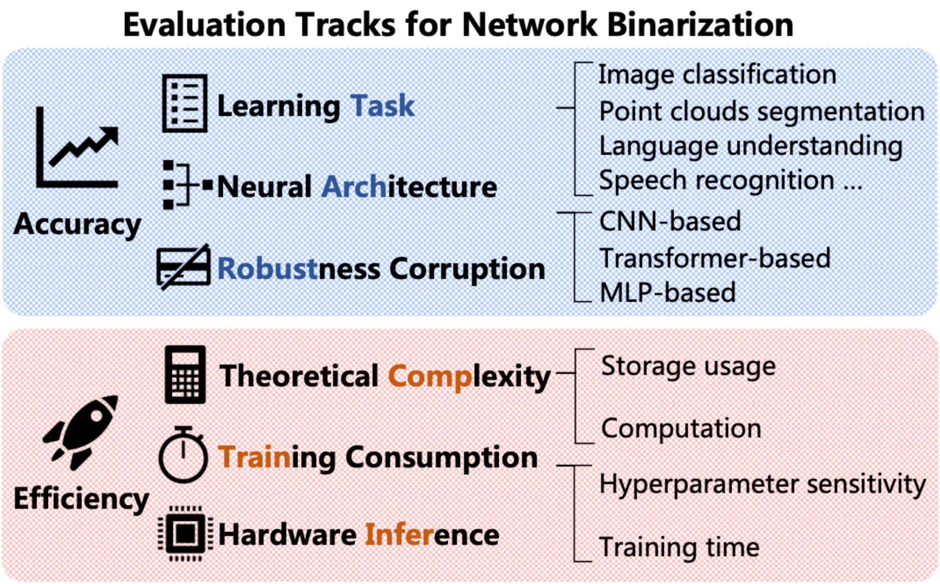
BiBench: *Benchmarking and Analyzing Network Binarization*

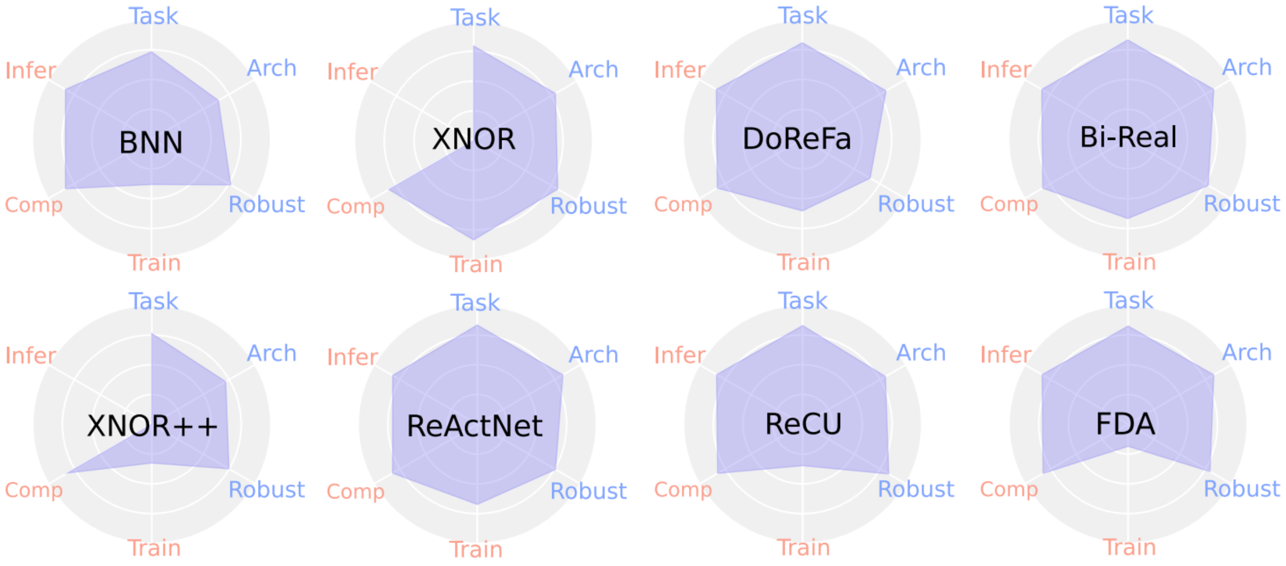




Network Binarization: benchmark

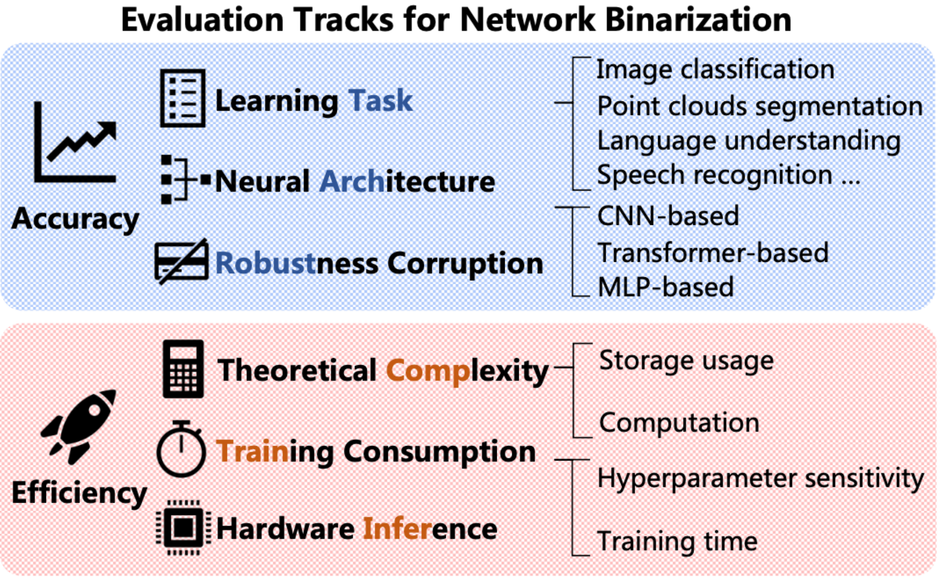
***BiBench**: Benchmarking and Analyzing Network Binarization*

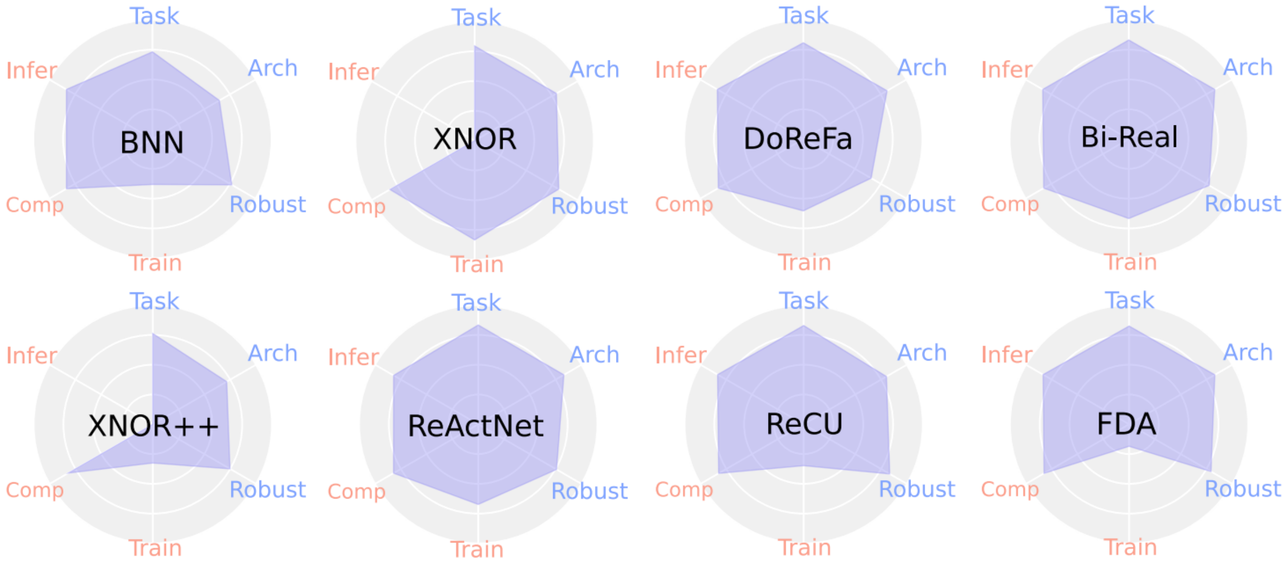




Network Binarization: benchmark

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

Network 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

Network Binarization: benchmark

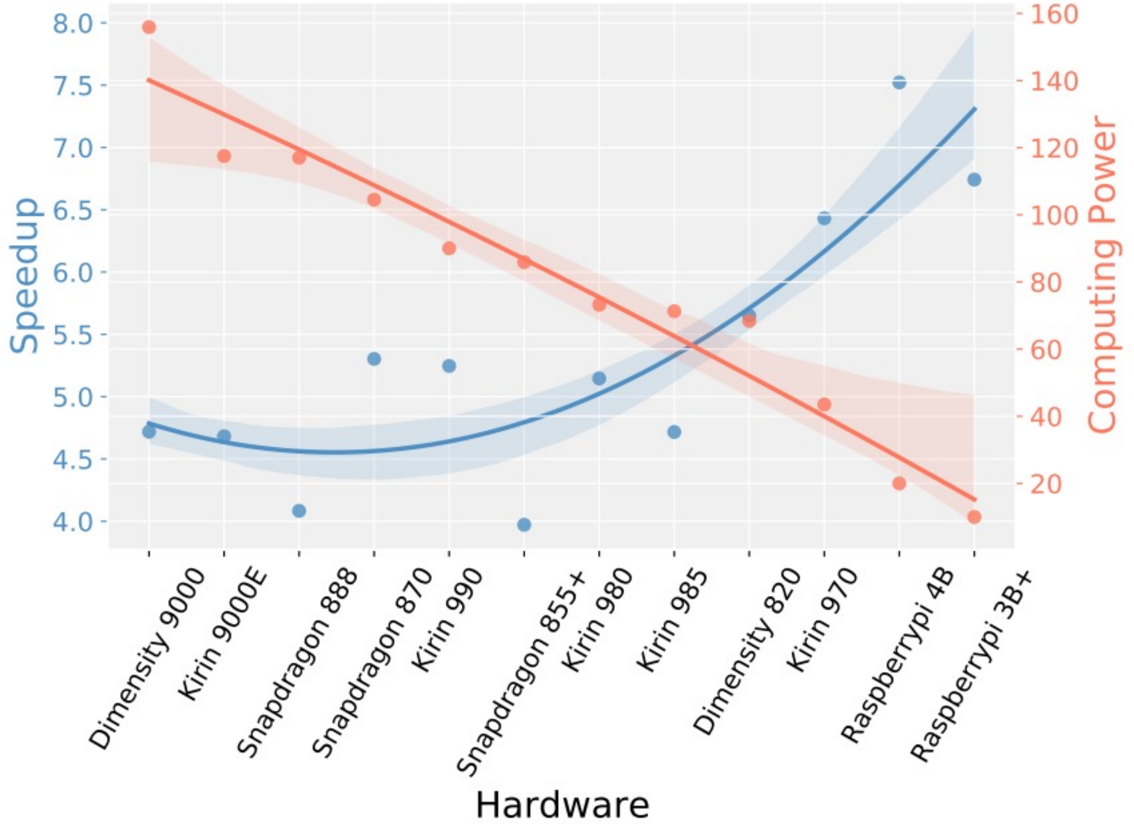
***BiBench**: Benchmarking and Analyzing Network Binarization*

The 3 Most Effective Techniques for Generic Binarization:

(1) Soft gradient approximation

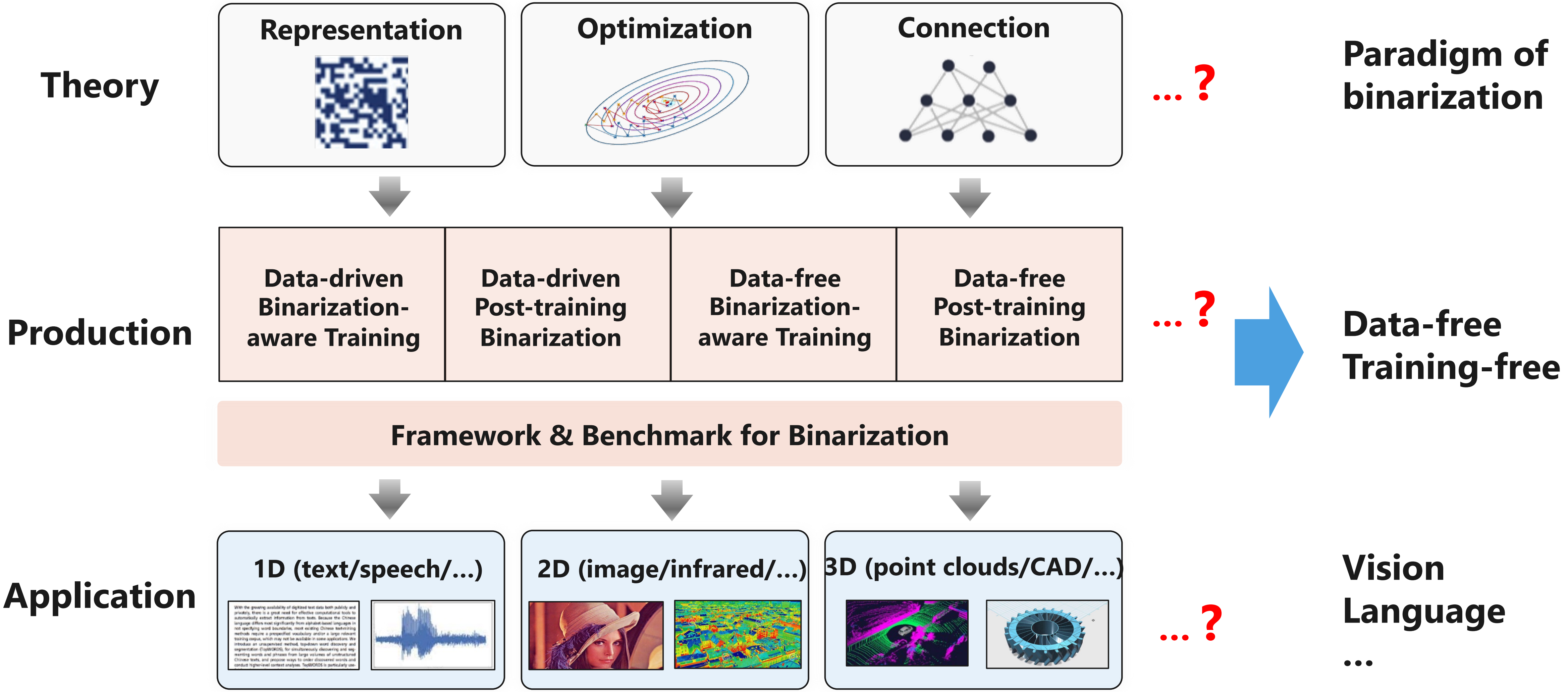
(2) Channel-wise scaling factors

(3) Pre-binarization parameter redistributing



Interesting Finding for Binarization:
Born for Edge

Network Binarization: future



Thank you!

Q&A

Haotong Qin
Beihang University & ETH Zürich