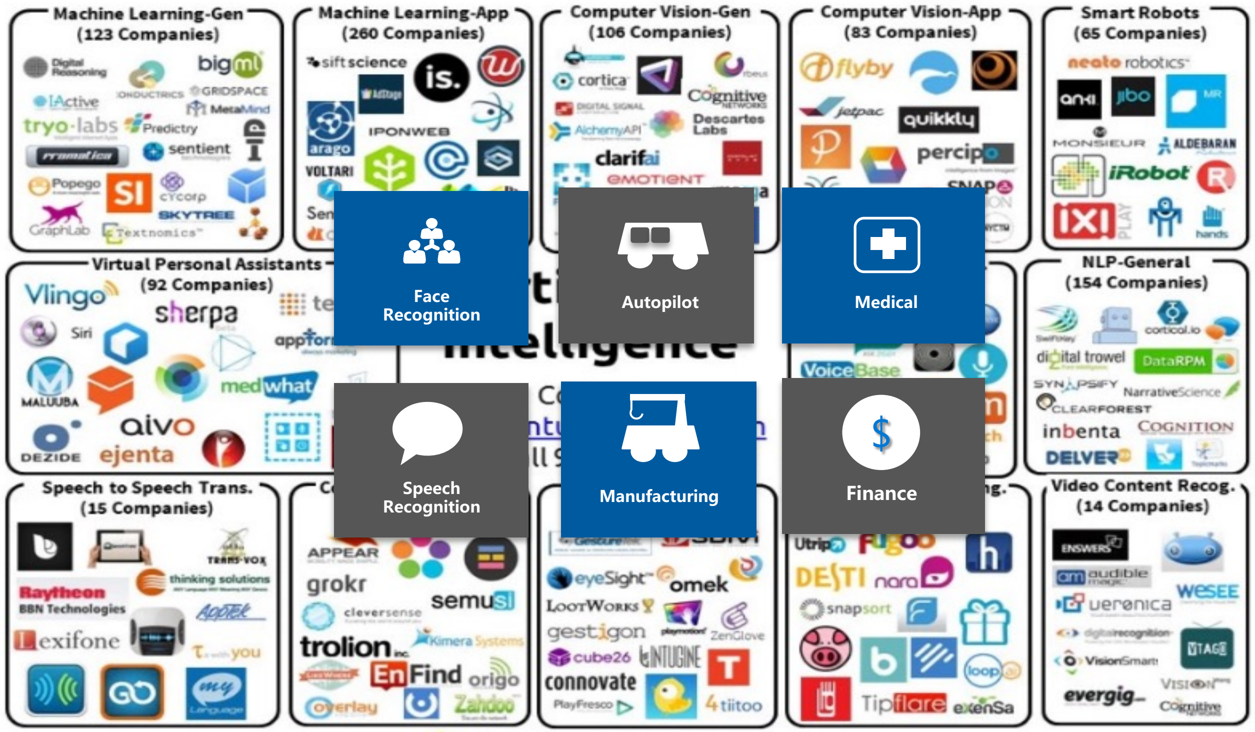
Network Binarization toward Hardware-friendly Deep Learning

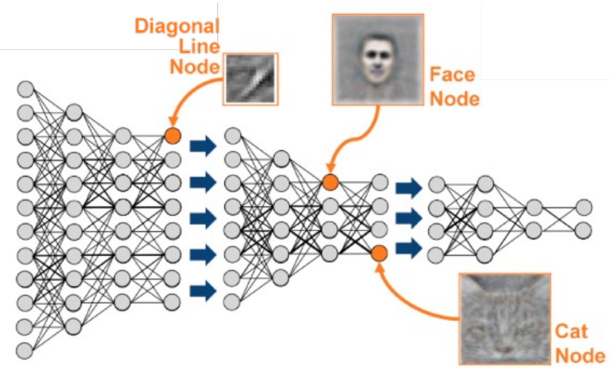
Haotong Qin
ETH Zürich CVL & Beihang University

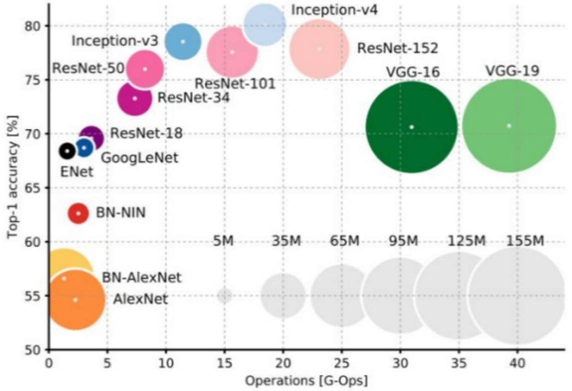
Background

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

…

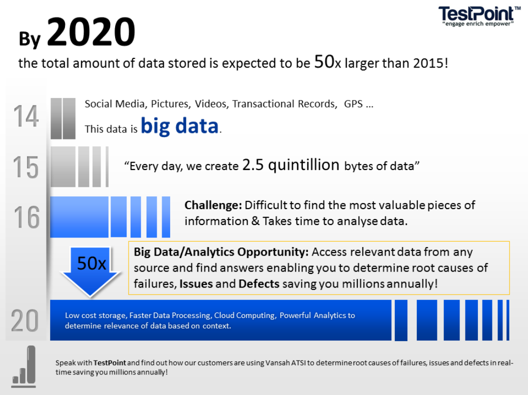


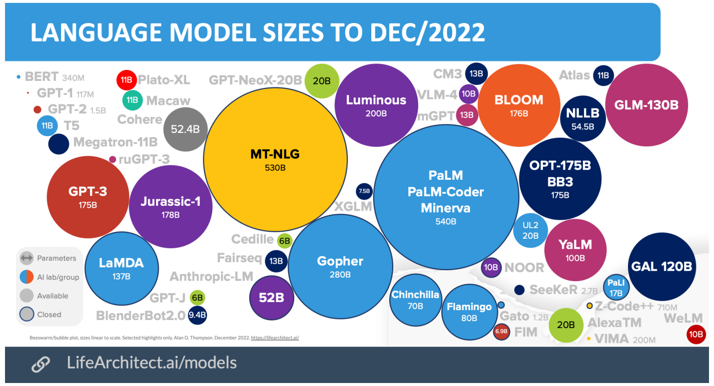


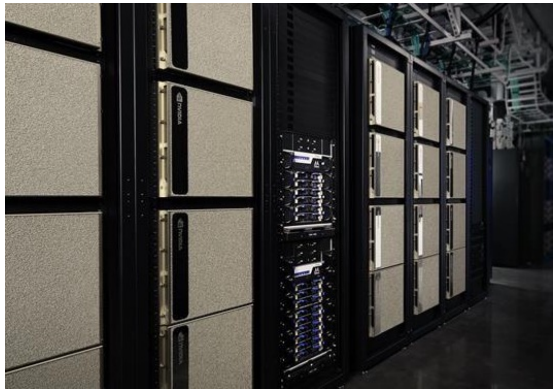


Background

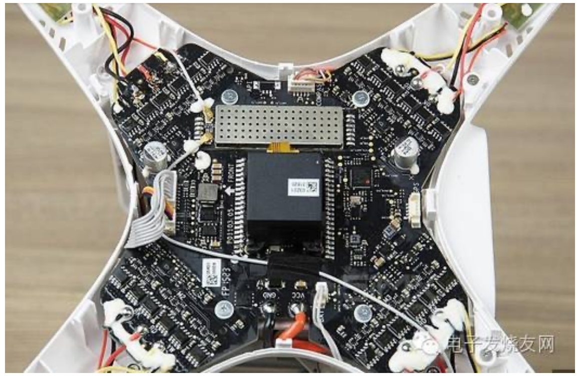
bigger data
and
larger model

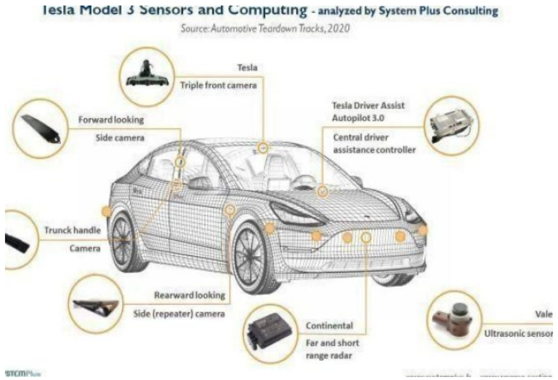






diverse usage
and
limited resources



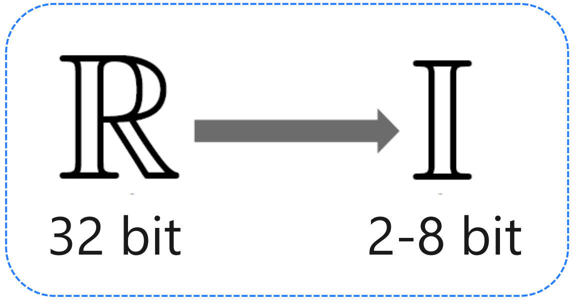


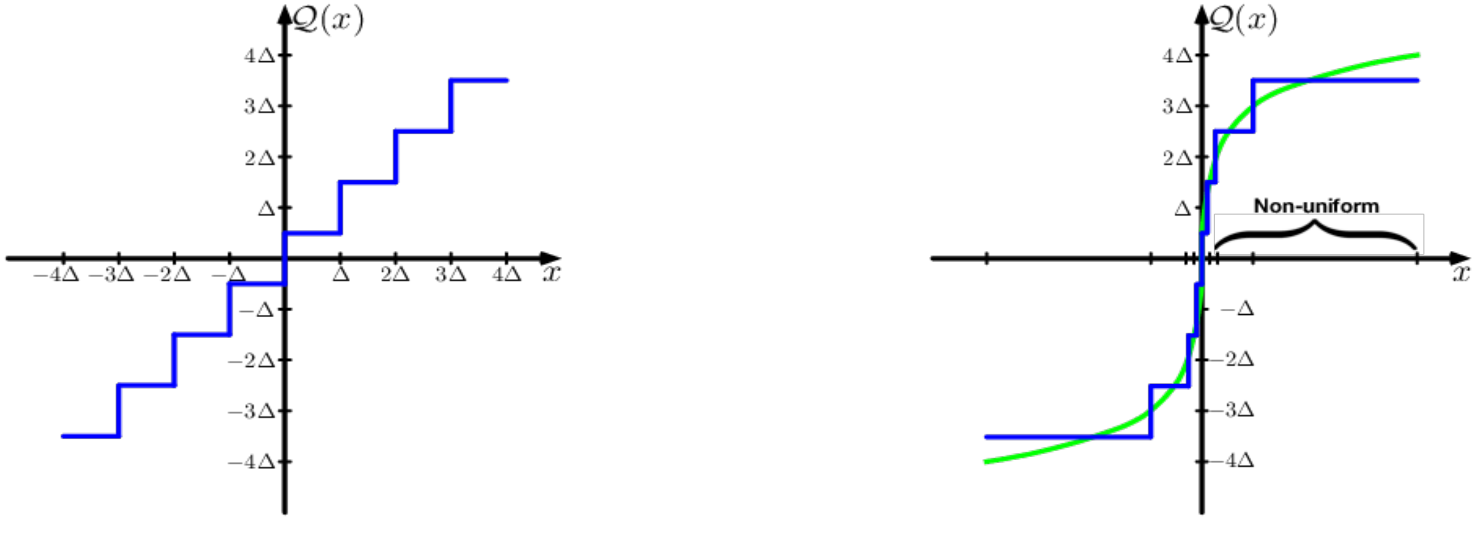
Network Quantization and Binarization

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

32-bit → 8-bit → 1-bit

Network Quantization: 2-8 bit





(a) (b)

Uniform Quantization

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

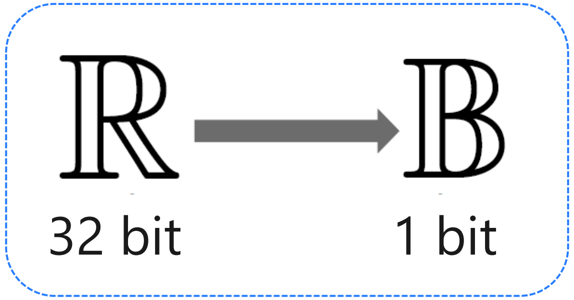
$$\Delta = \frac{u-l}{2^b - 1}$$

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



$\{-1, +1\}$

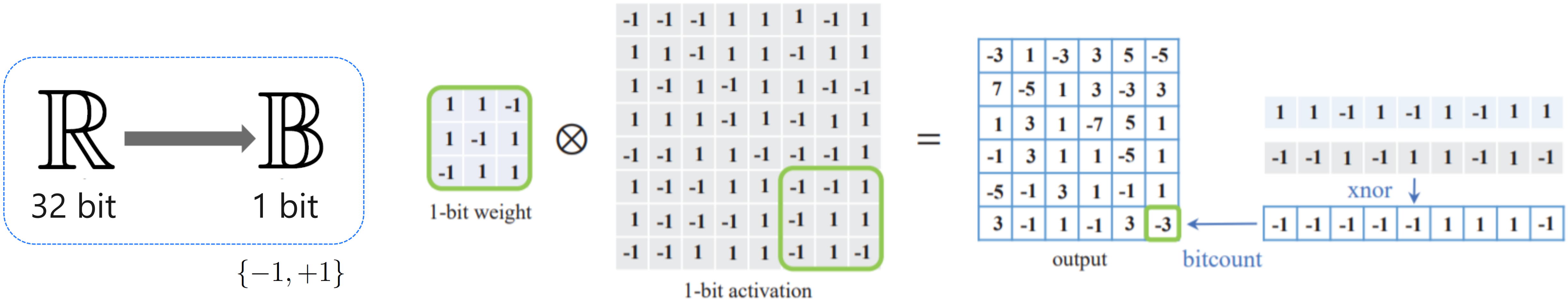
1-Bit Parameters:

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

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



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

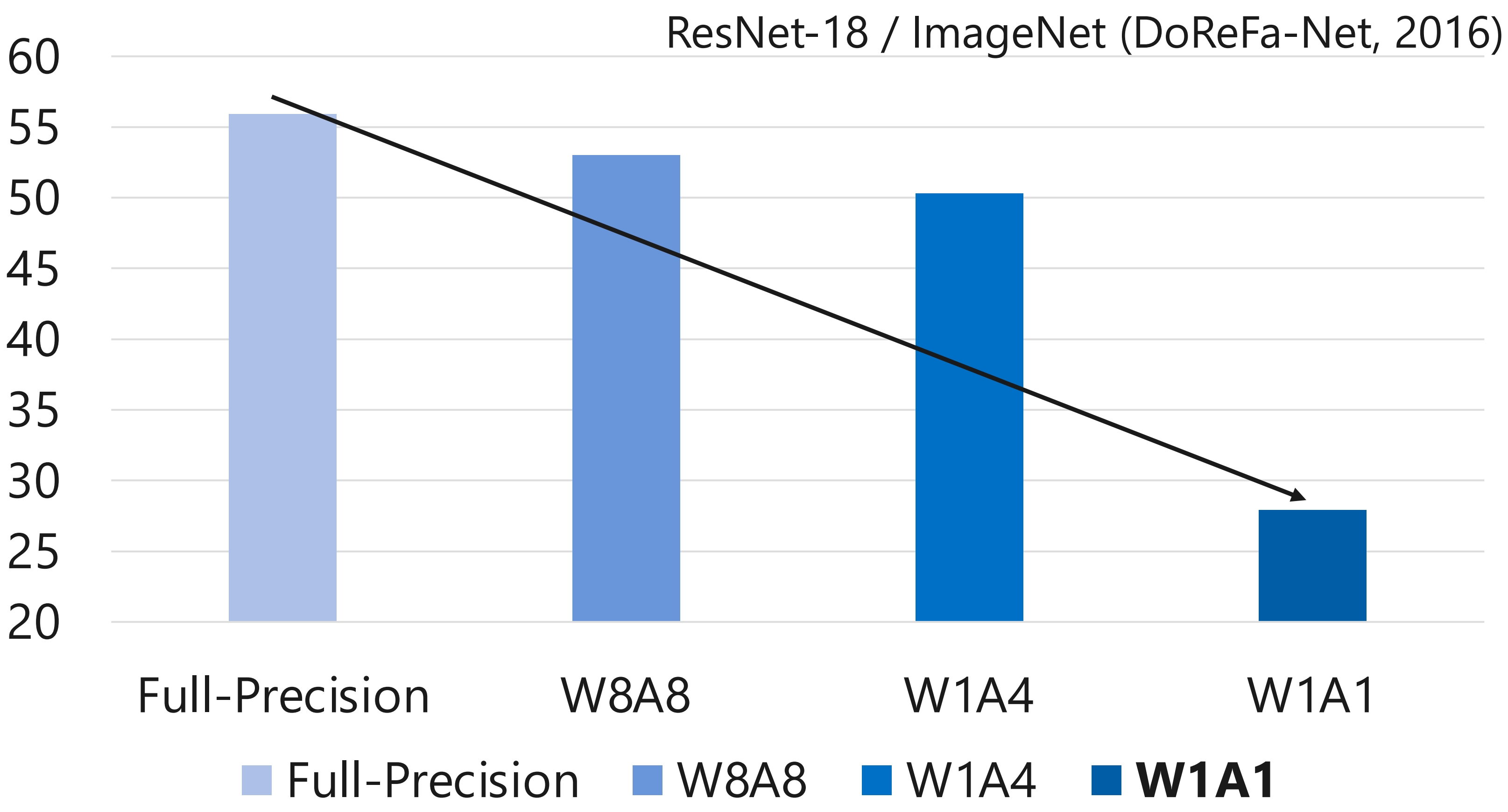
Efficient Instructions

Low Power Consumption

Network Binarization: challenges

Goal: accurate extreme-low bit quantization (binarization)

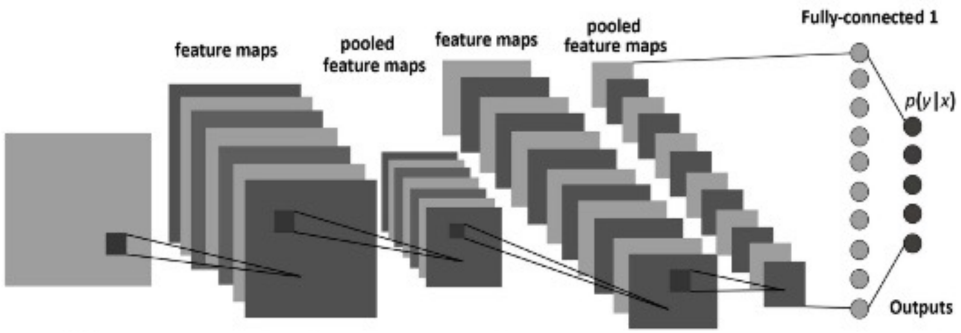
1. the **accuracy** of the binary network has dropped seriously



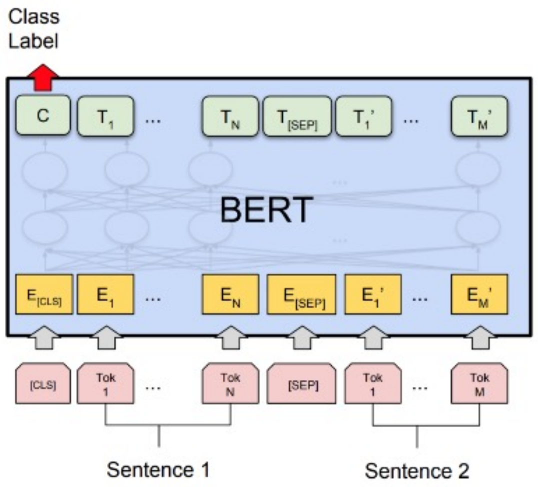
Network Binarization: challenges

Goal: accurate extreme-low bit quantization (binarization)

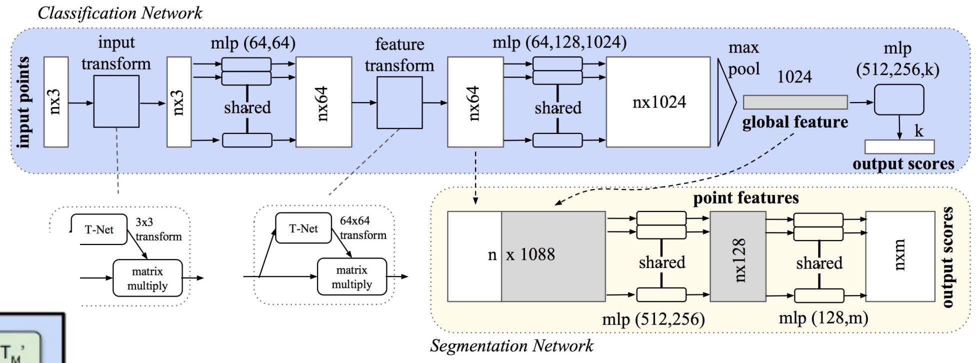
2. binarization methods are not generic across different **architecture**



CNN (ResNet, VGG, ...)



Transformer (BERT, ViT,...)



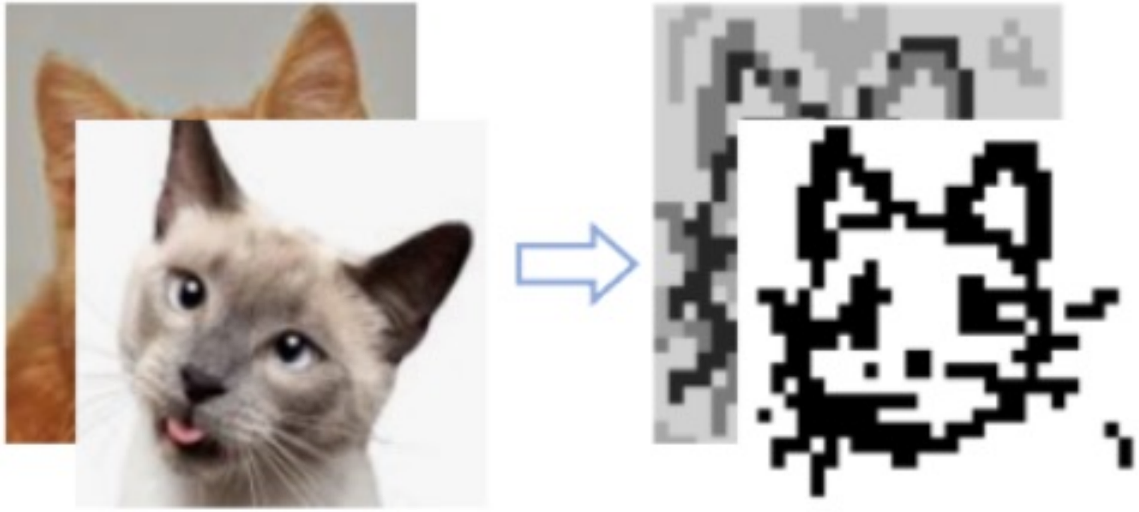
MLP (PointNet, MLP-Mixer, ...)

...

CNN Binarization: information loss and retention

Effects of BNN in the Forward and Backward Propagation

limited representation



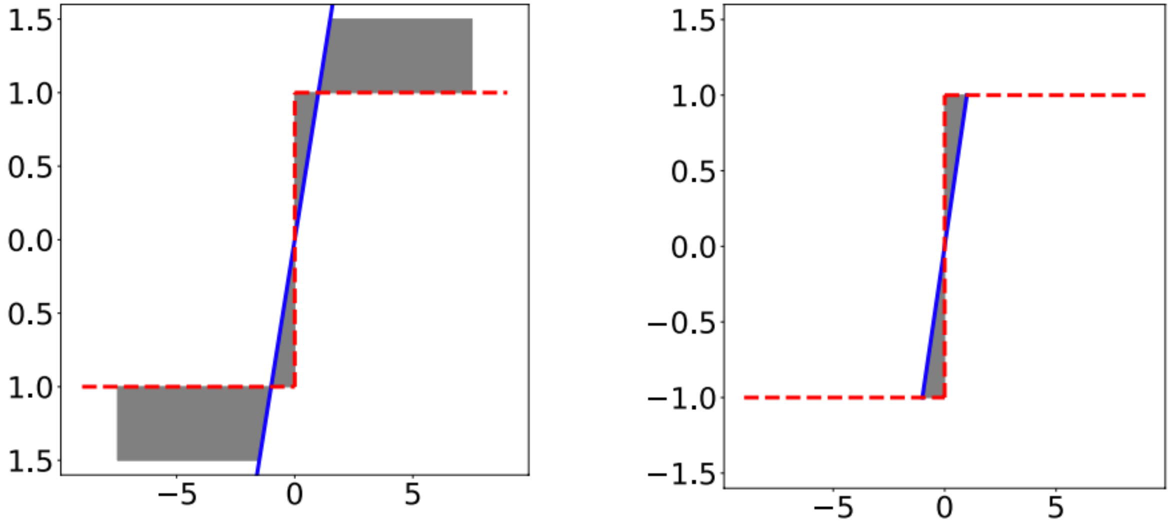
$$\mathbf{B_x} = \mathbf{sign}(\mathbf{x}) = \begin{cases} +1, & if\ \mathbf{x} \geq 0 \\ -1, & otherwise. \end{cases}$$

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



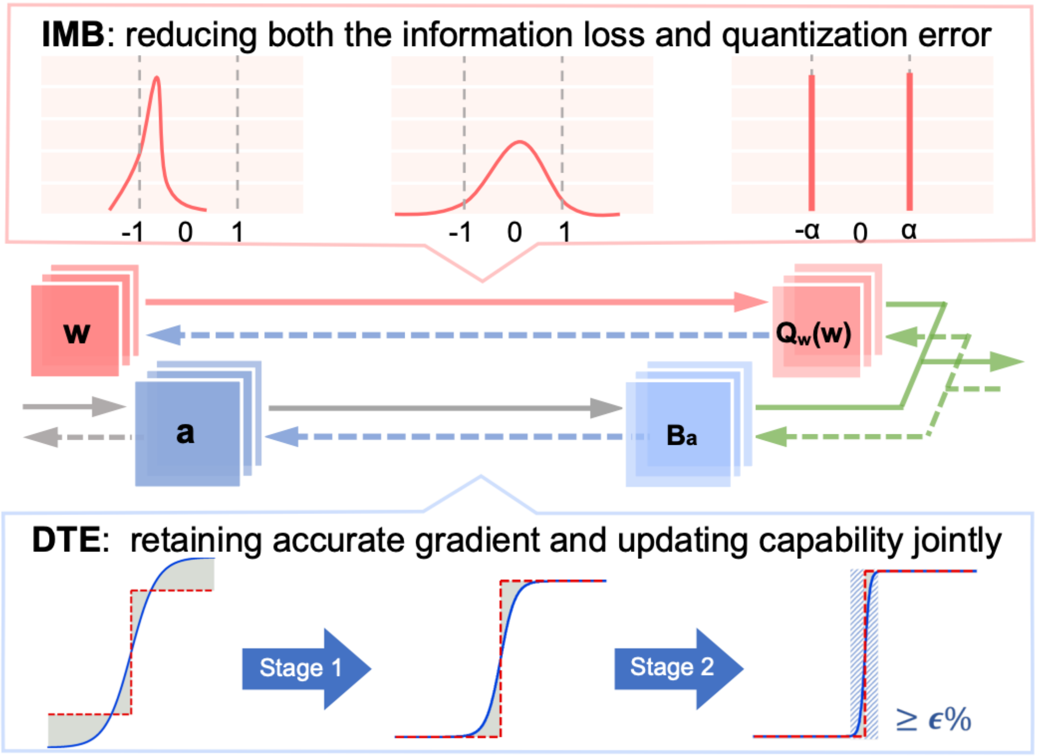
$$\texttt{Identity}: y = x$$

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

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

CNN Binarization: information loss and retention

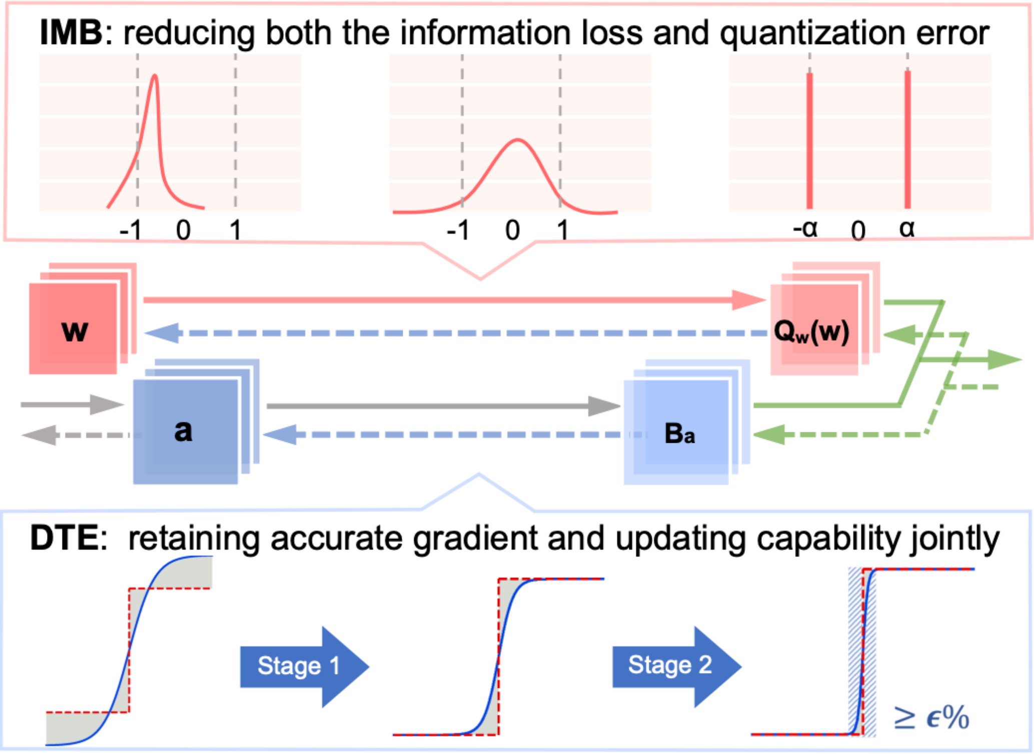
Distribution-sensitive Information Retention (DIR-Net)



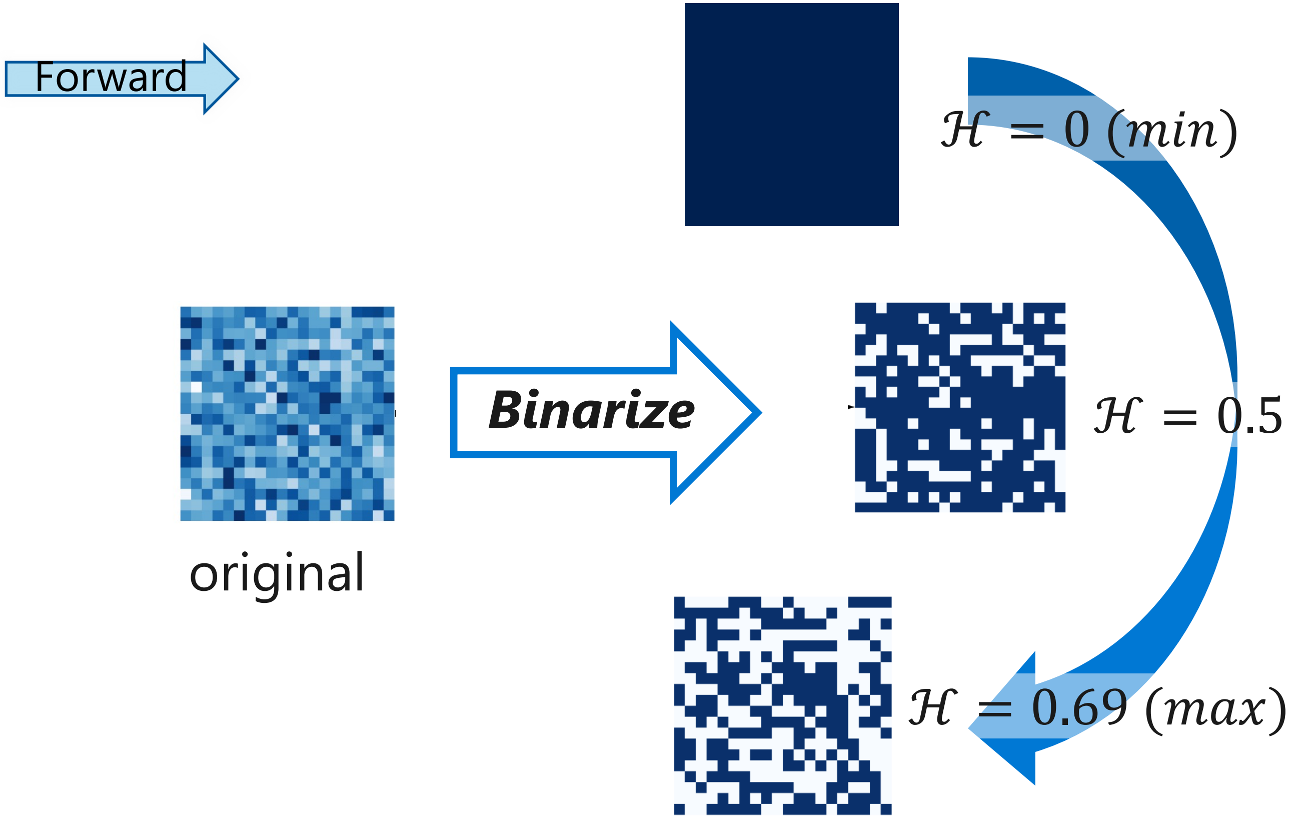
Distribution-sensitive Information Retention for Accurate Binary Neural Network. Haotong Qin, et al. IJCV 2022.

CNN Binarization: information loss and retention

Distribution-sensitive Information Retention (DIR-Net)



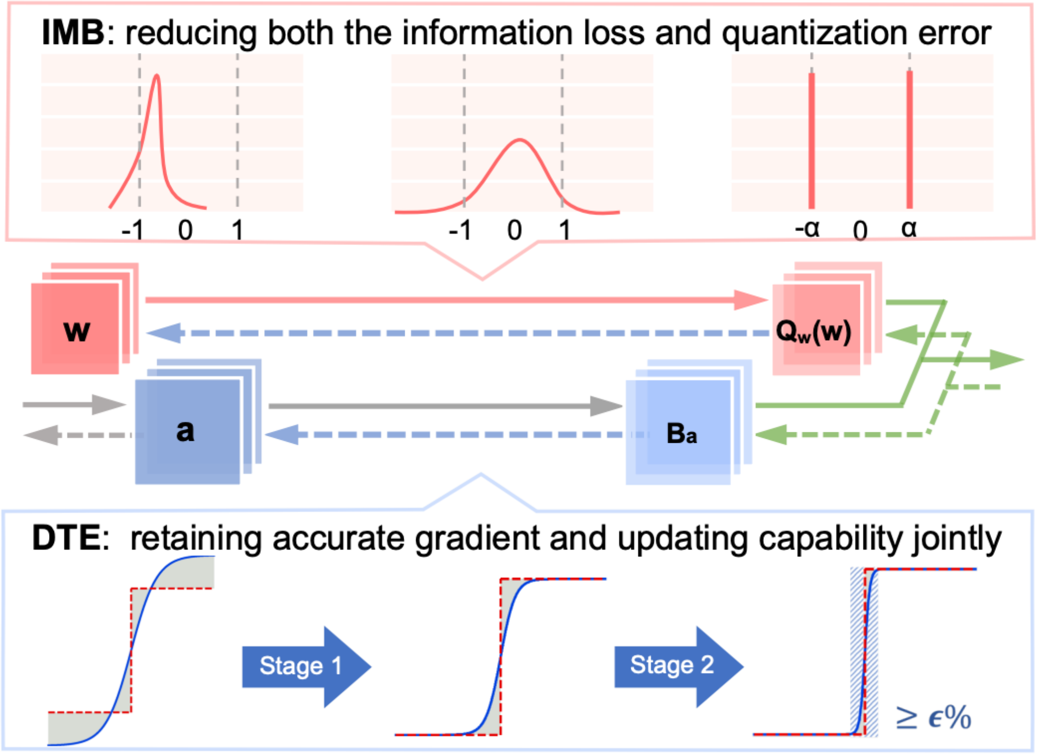
Maximizing the information entropy



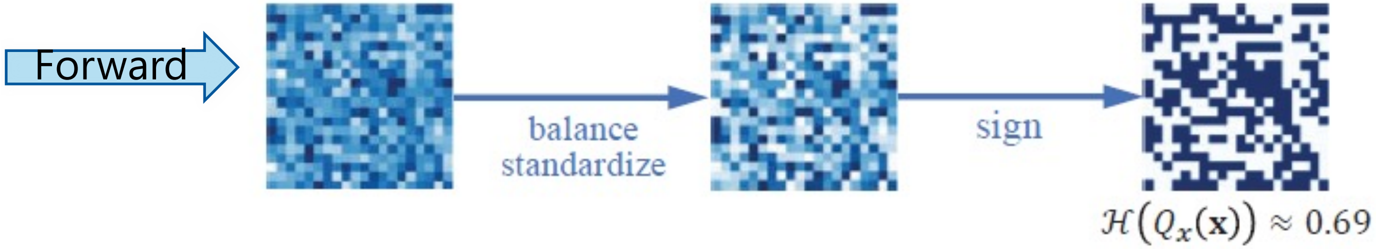
Distribution-sensitive Information Retention for Accurate Binary Neural Network. Haotong Qin, et al. IJCV 2022.

CNN Binarization: information loss and retention

Distribution-sensitive Information Retention (DIR-Net)



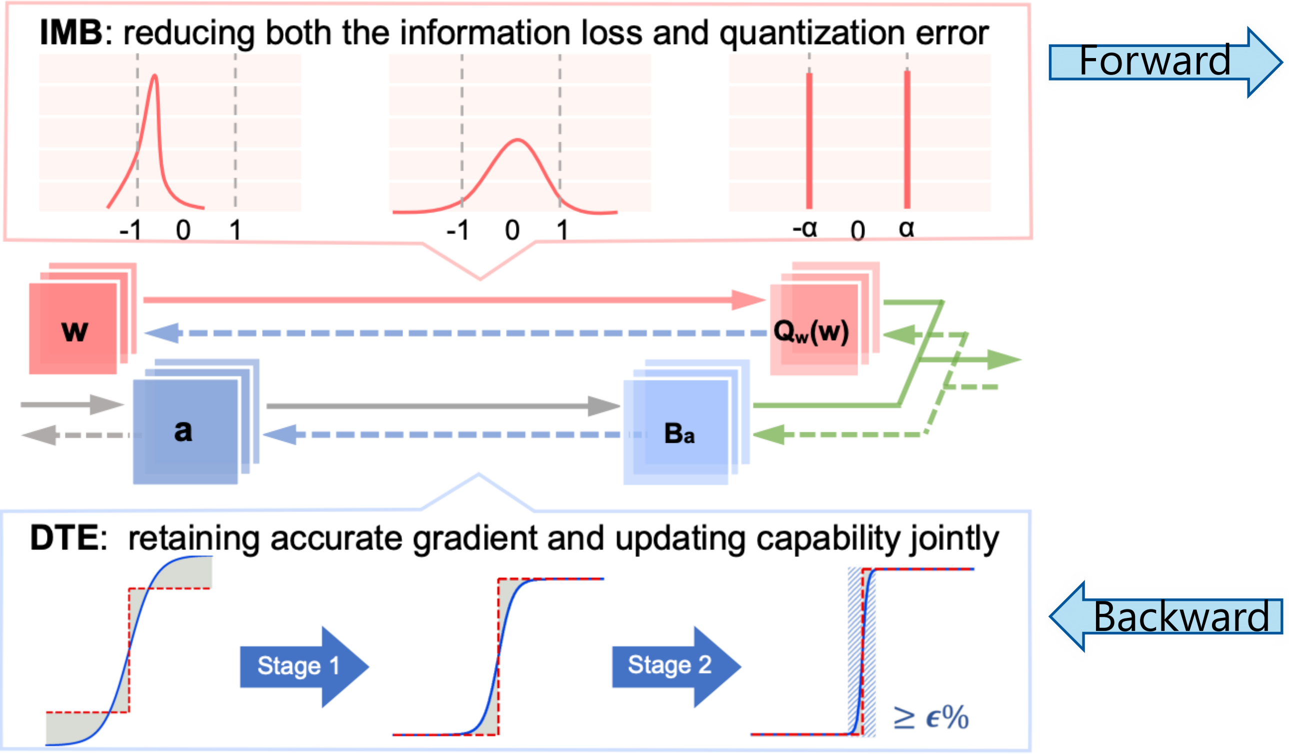
Maximizing the information entropy



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

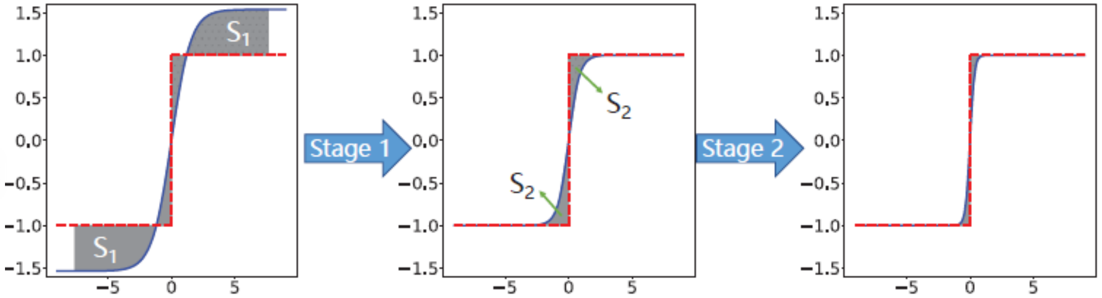
CNN Binarization: information loss and retention

Distribution-sensitive Information Retention (DIR-Net)



Maximizing the information entropy

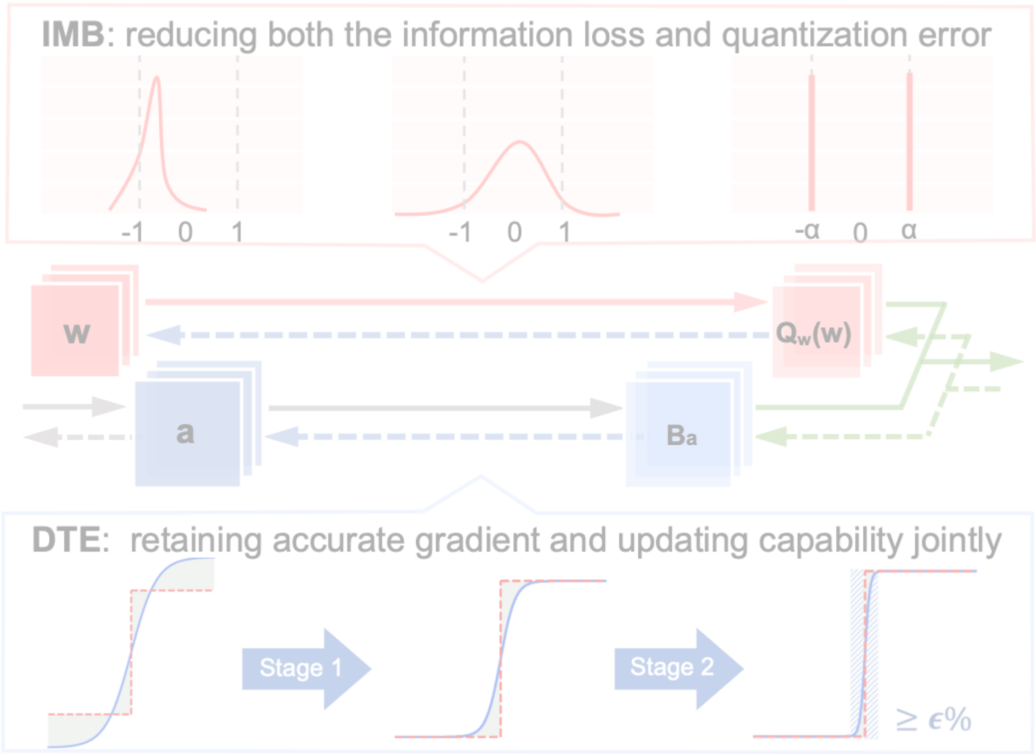
Changing the shape of estimator



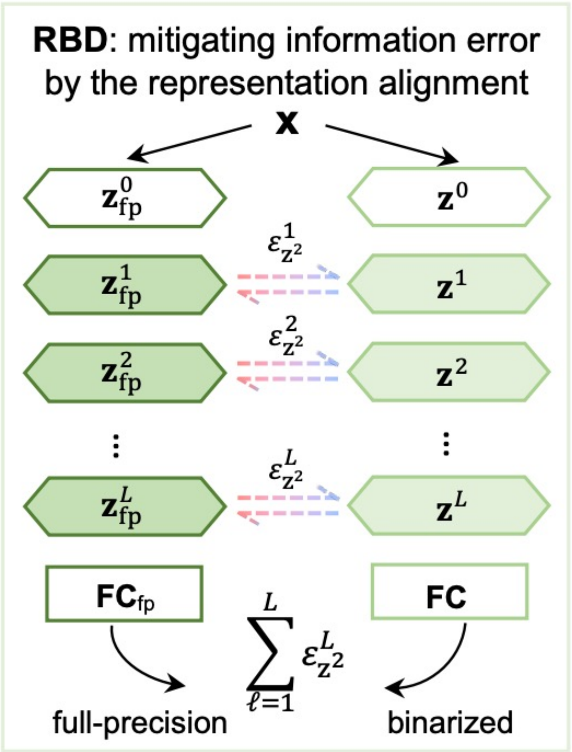
Distribution-sensitive Information Retention for Accurate Binary Neural Network. Haotong Qin, et al. IJCV 2022.

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** ± 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** ± 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** ± 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

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**

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**

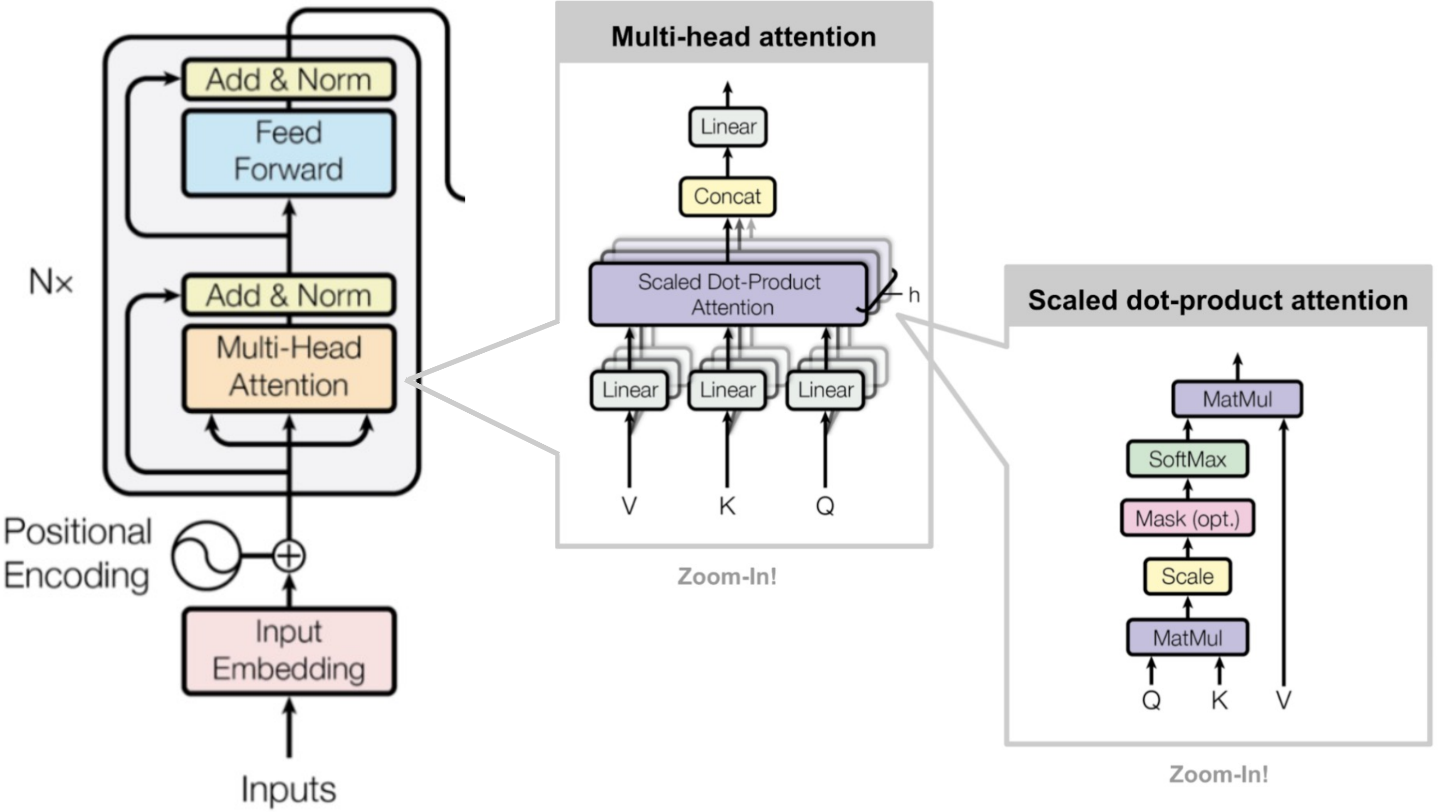
11.1x storage saving

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

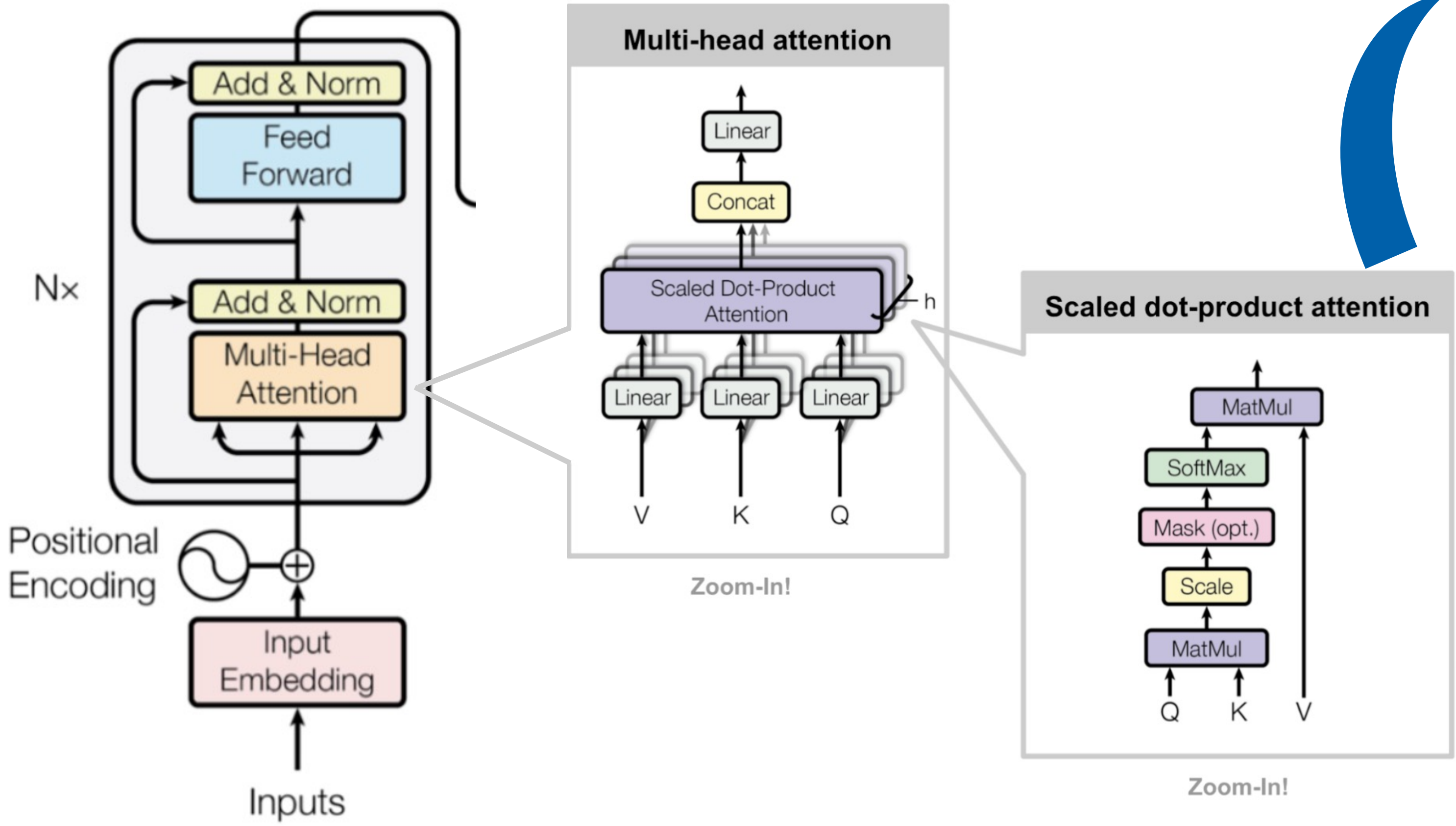


https://deepfrench.gitlab.io/deep-learning-project/

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

Transformer Binarization: attention crash and recovery

Bottlenecks of Fully Binarized BERT Baseline



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

Binarize (directly)

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

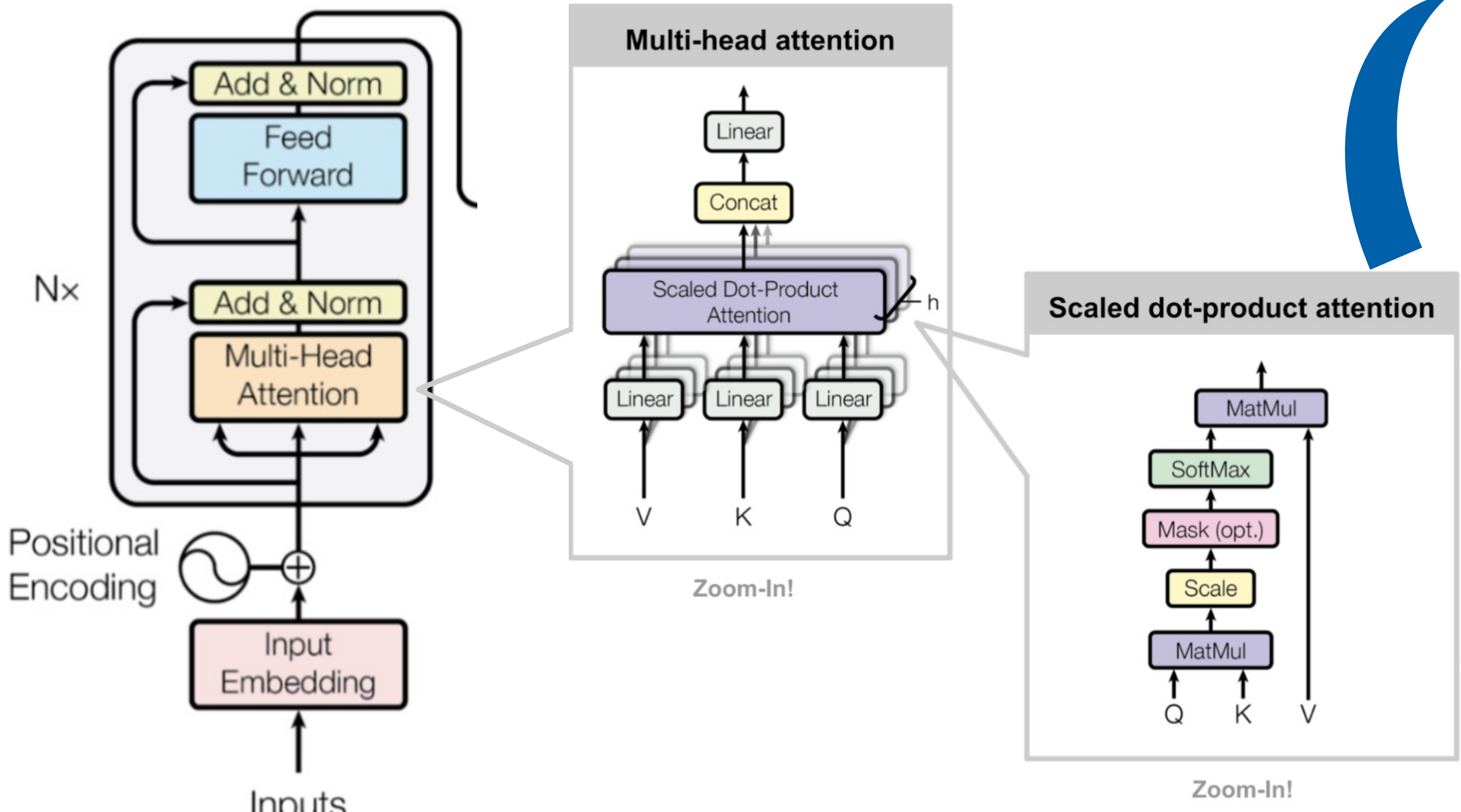
$$\mathbf{A} = \frac{1}{\sqrt{D}}\left(\mathbf{B_Q} \otimes \mathbf{B_K}^\top\right)$$

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

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

Transformer Binarization: attention crash and recovery

Bottlenecks of Fully Binarized BERT Baseline



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

Binarize (directly)

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

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

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

Severely dropped!

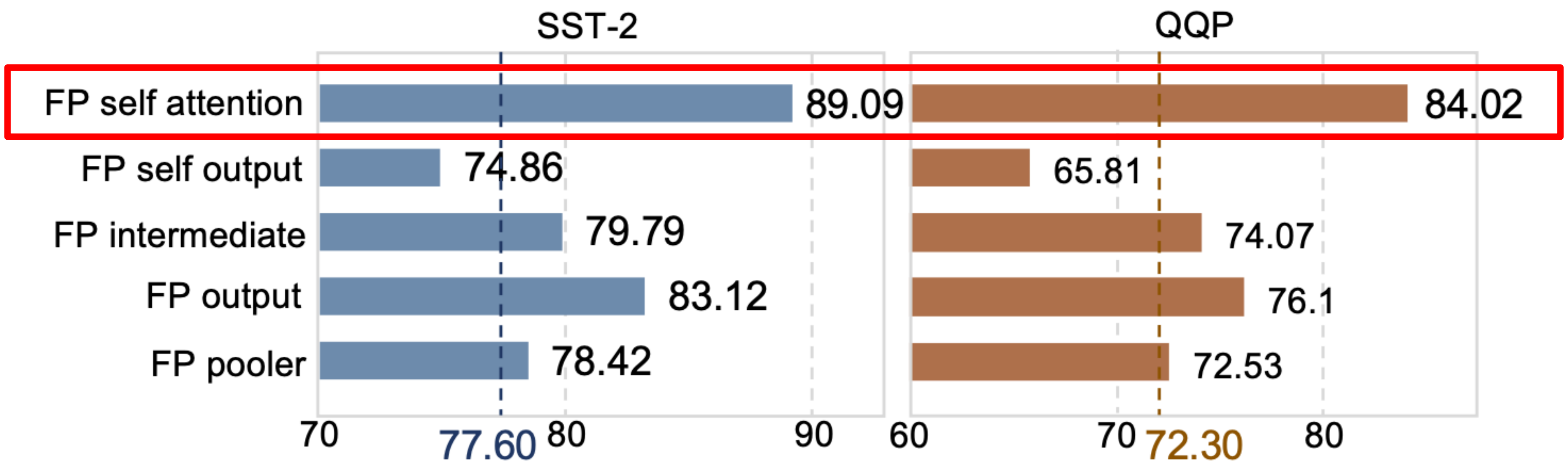
(Avg: 83.9% -> **50.4%**)

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

Transformer Binarization: attention crash and recovery

Bottlenecks of Fully Binarized BERT Baseline

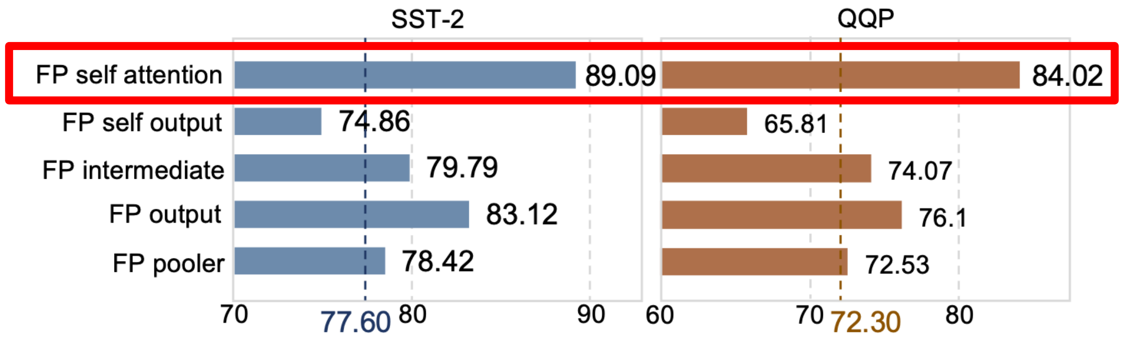
Which part caused the biggest drop?

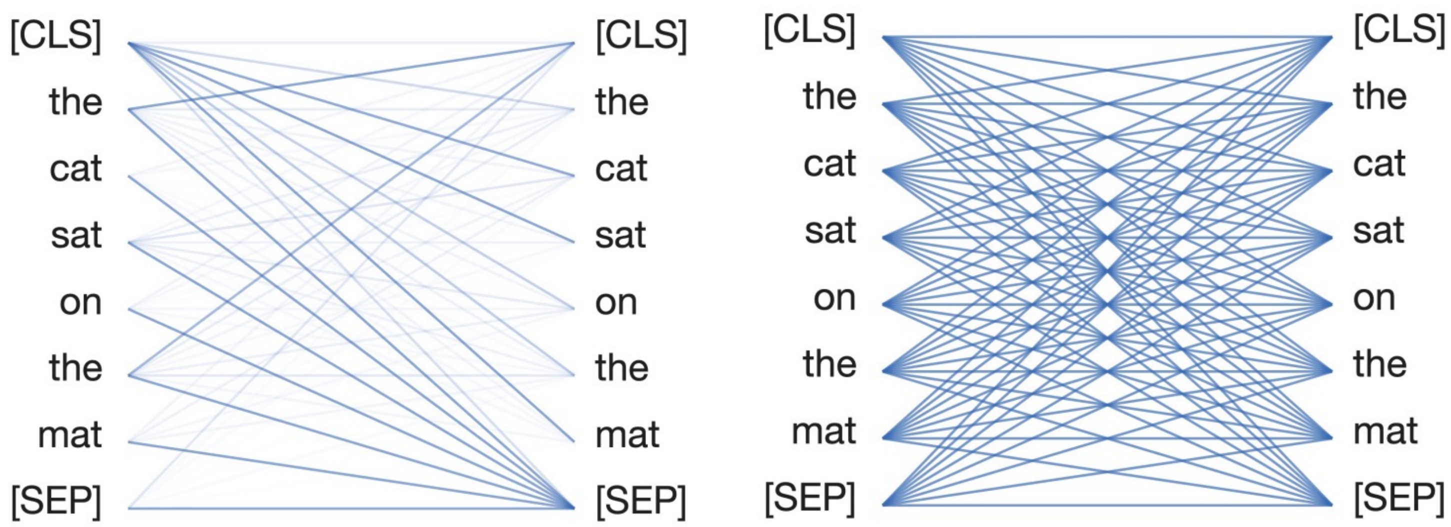


BiBERT: Accurate Fully Binarized BERT. Haotong Qin, et al. ICLR 2022.

Transformer Binarization: attention crash and recovery

Bottlenecks of Fully Binarized BERT Baseline





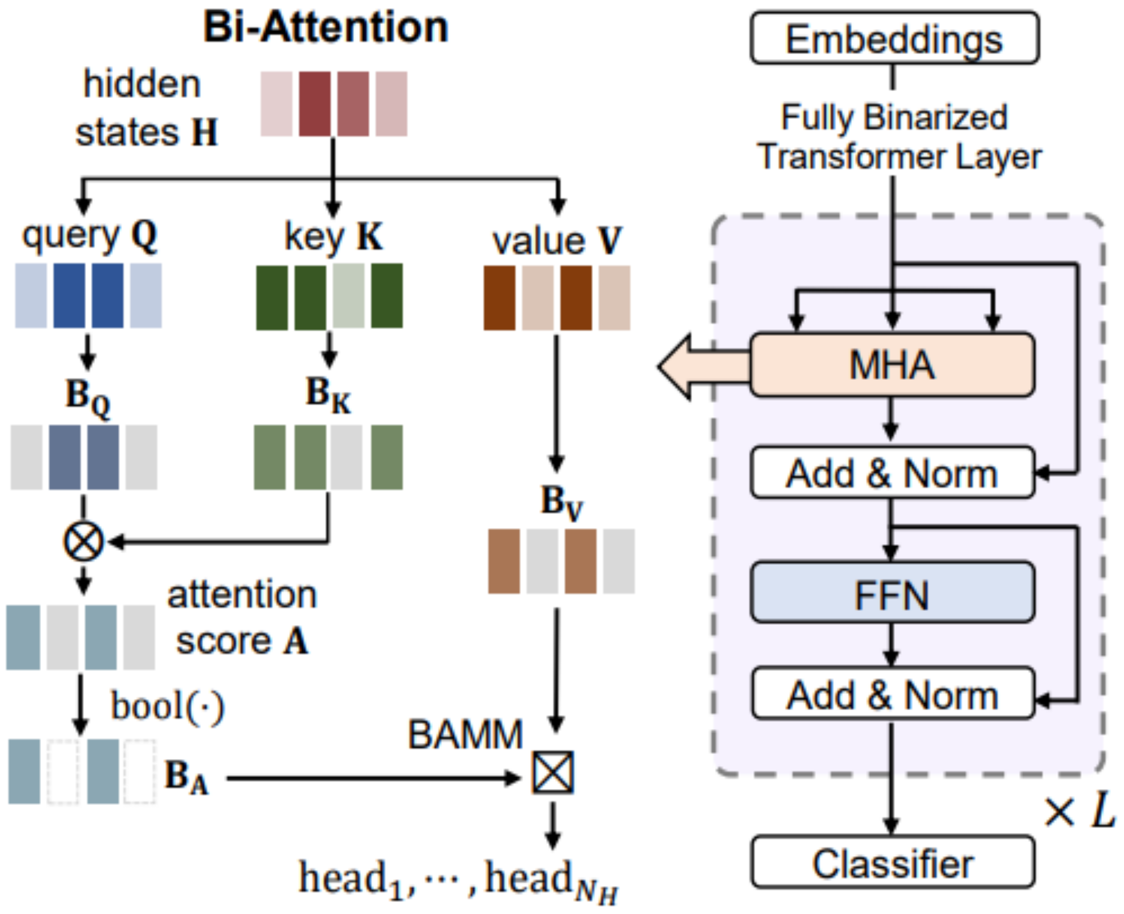
(a) Full-precision (b) Fully binarized BERT baseline

attention mechanism crashed

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

Transformer Binarization: attention crash and recovery

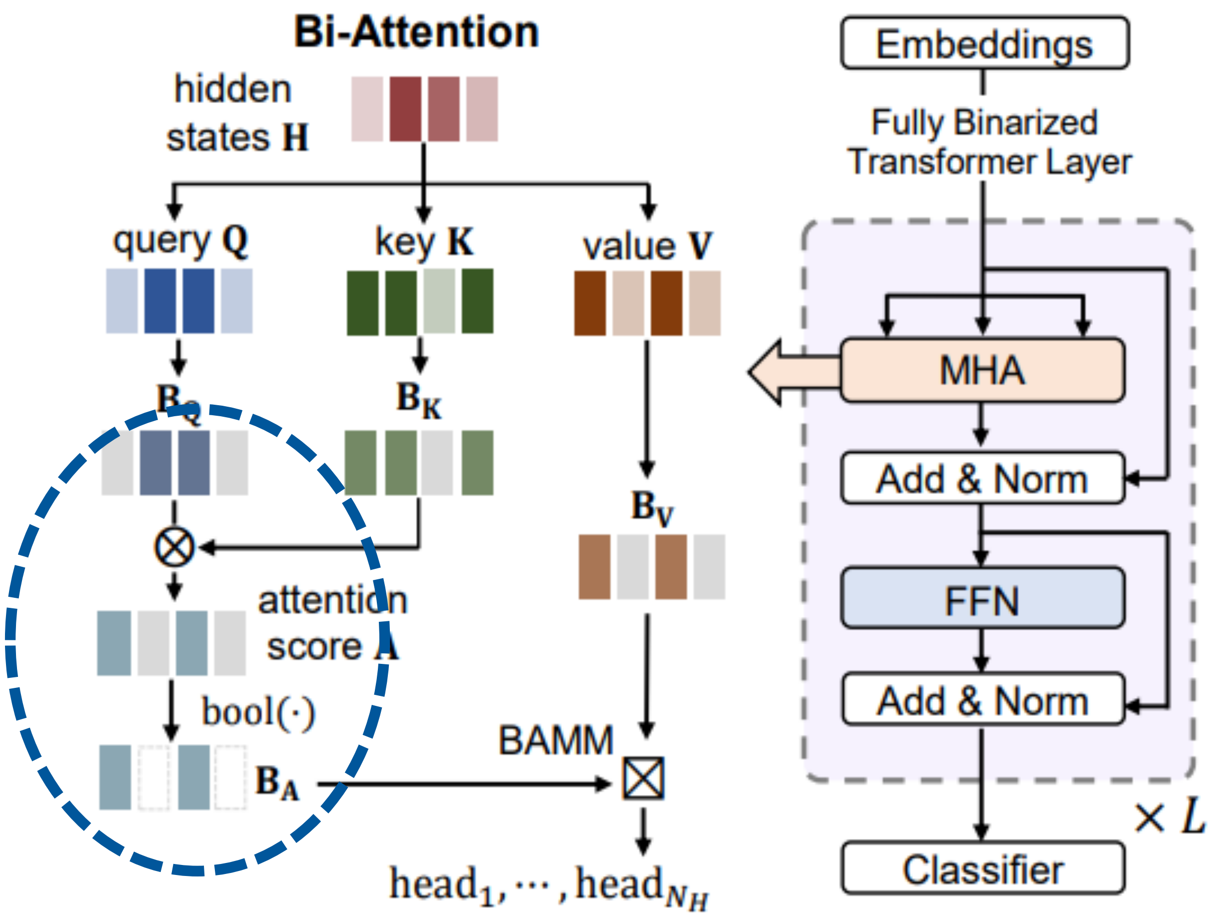
Accurate Fully Binarized BERT (BiBERT)



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

Transformer Binarization: attention crash and recovery

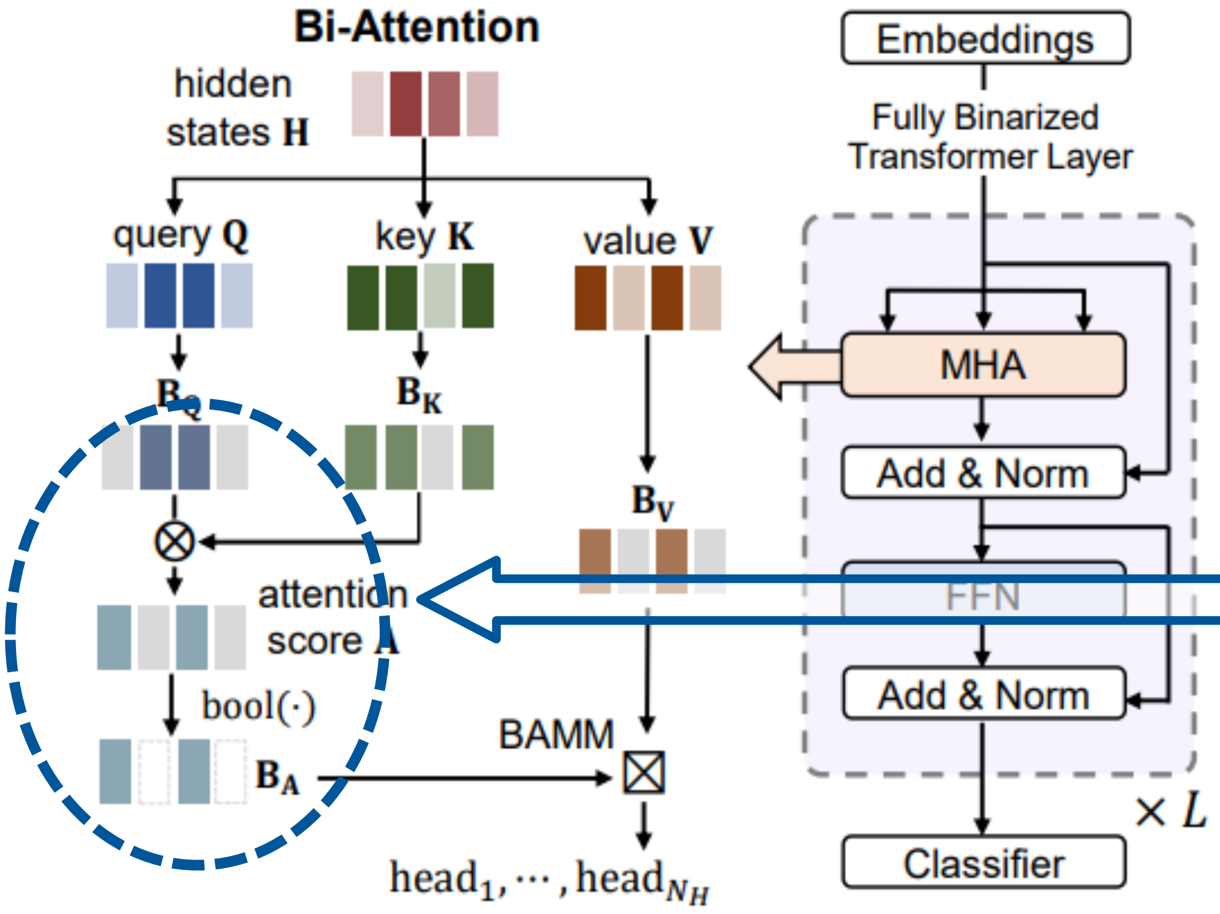
Accurate Fully Binarized BERT (BiBERT)



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

Transformer Binarization: attention crash and recovery

Accurate Fully Binarized BERT (BiBERT)



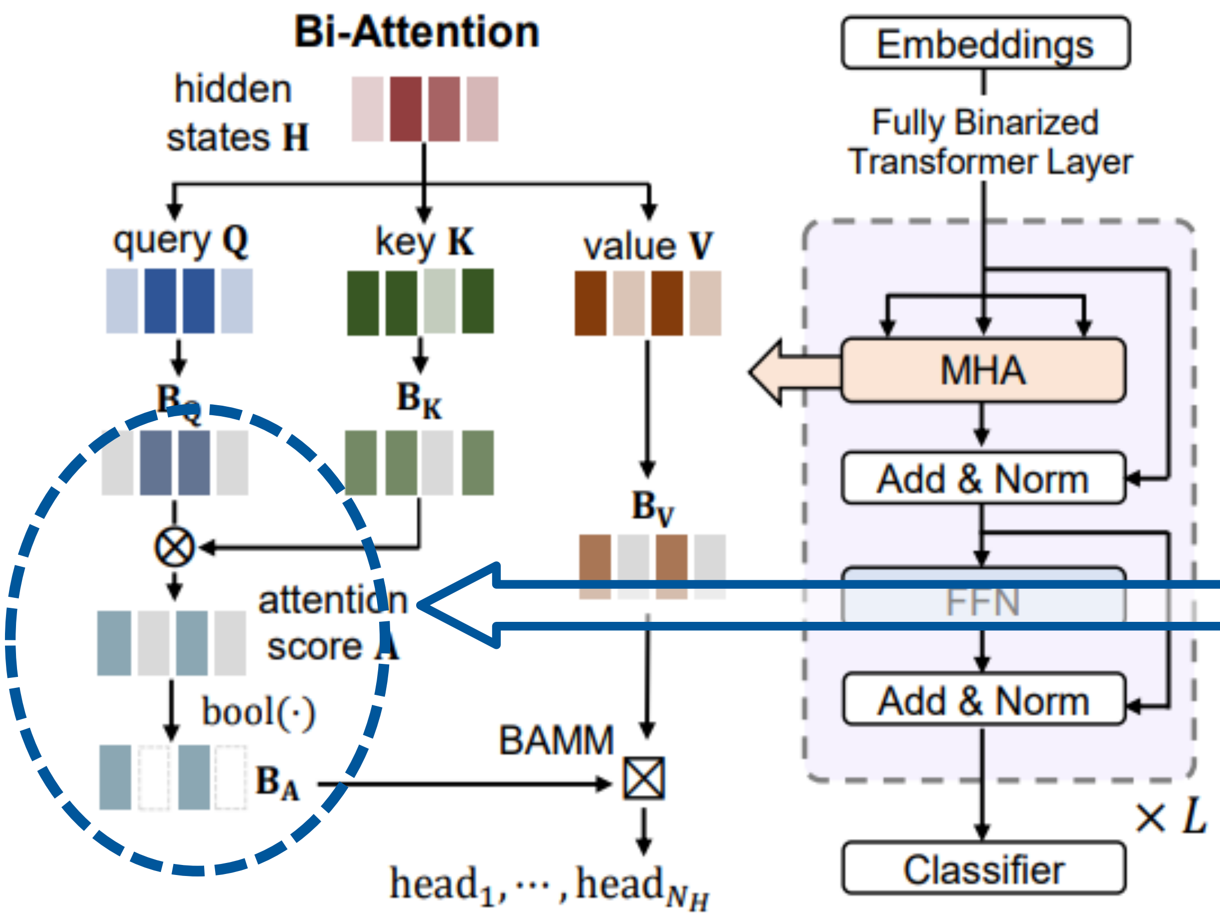
$$\mathrm{bool}(x) = \begin{cases} 1, & \text{if } x \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

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

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

Transformer Binarization: attention crash and recovery

Accurate Fully Binarized BERT (BiBERT)



1. ~~SoftMax~~

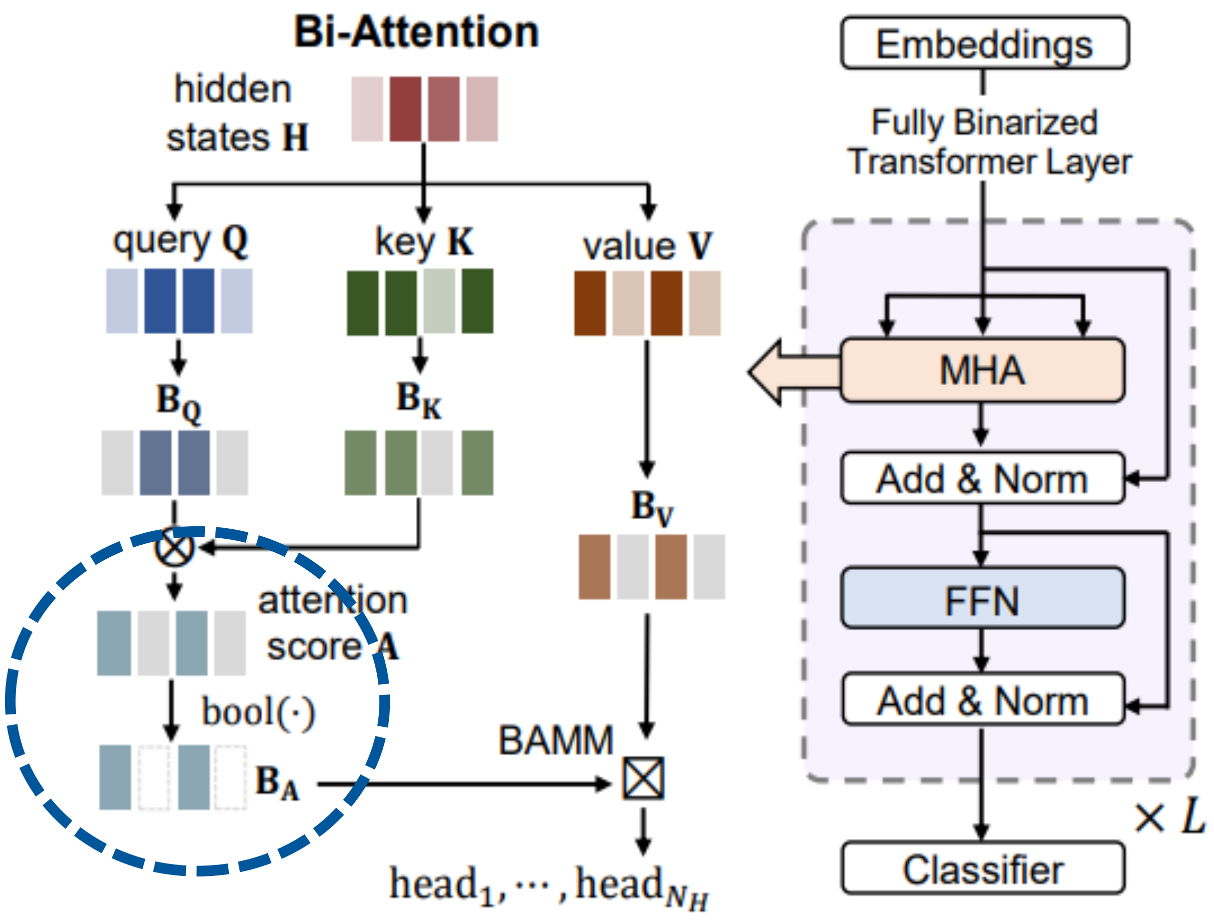
$$\operatorname{bool}(x) = \begin{cases} 1, & \text{if } x \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

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

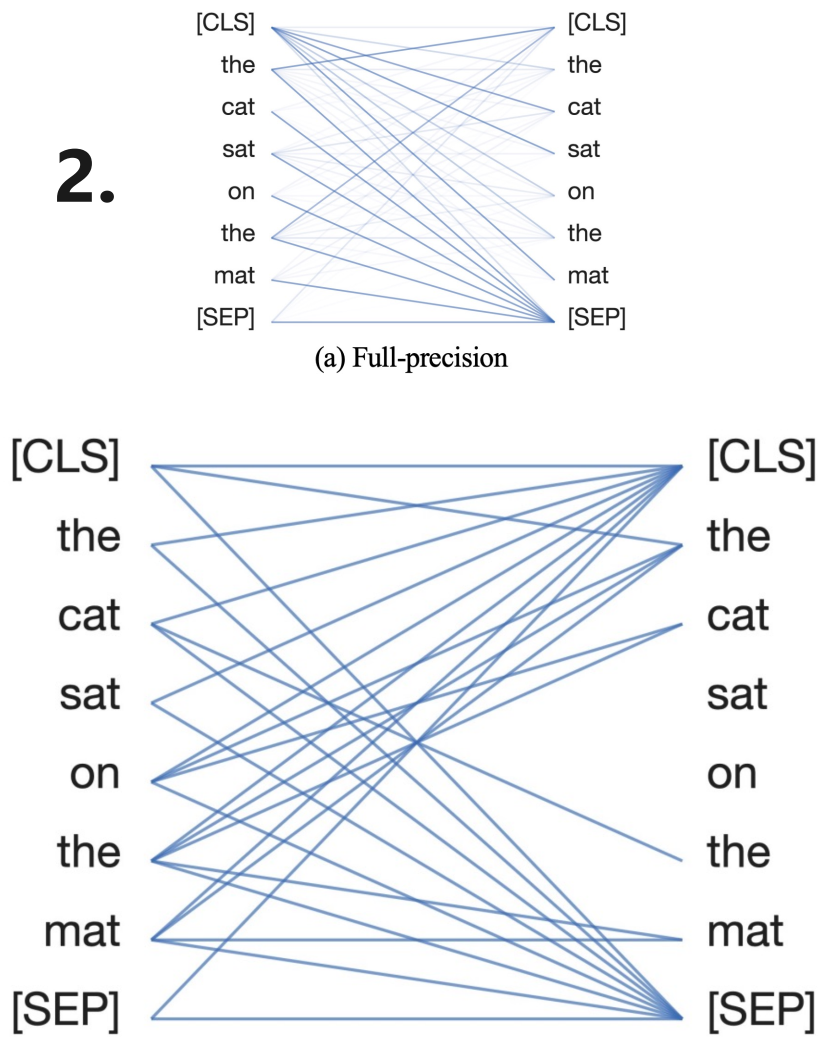
BiBERT: Accurate Fully Binarized BERT. Haotong Qin, et al. ICLR 2022.

Transformer Binarization: attention crash and recovery

Accurate Fully Binarized BERT (BiBERT)



2.



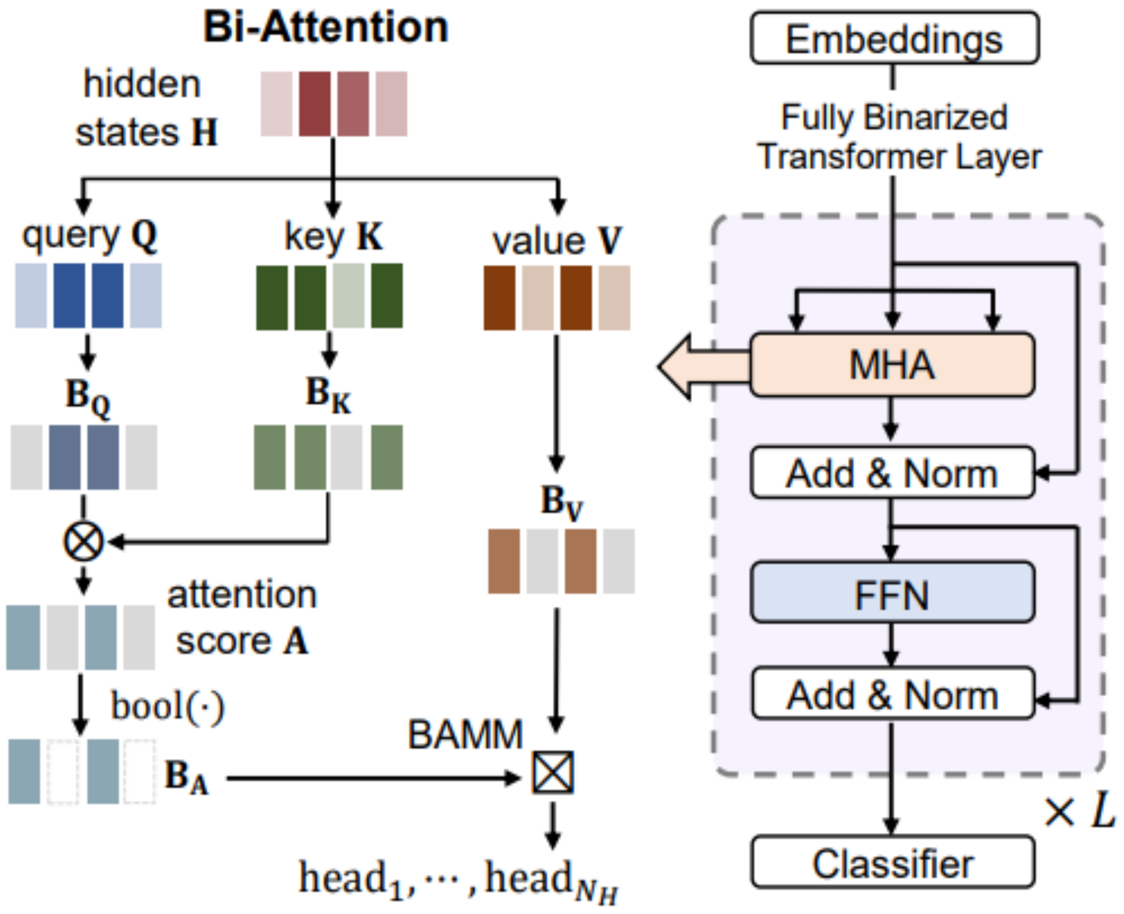
(a) Full-precision

(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)



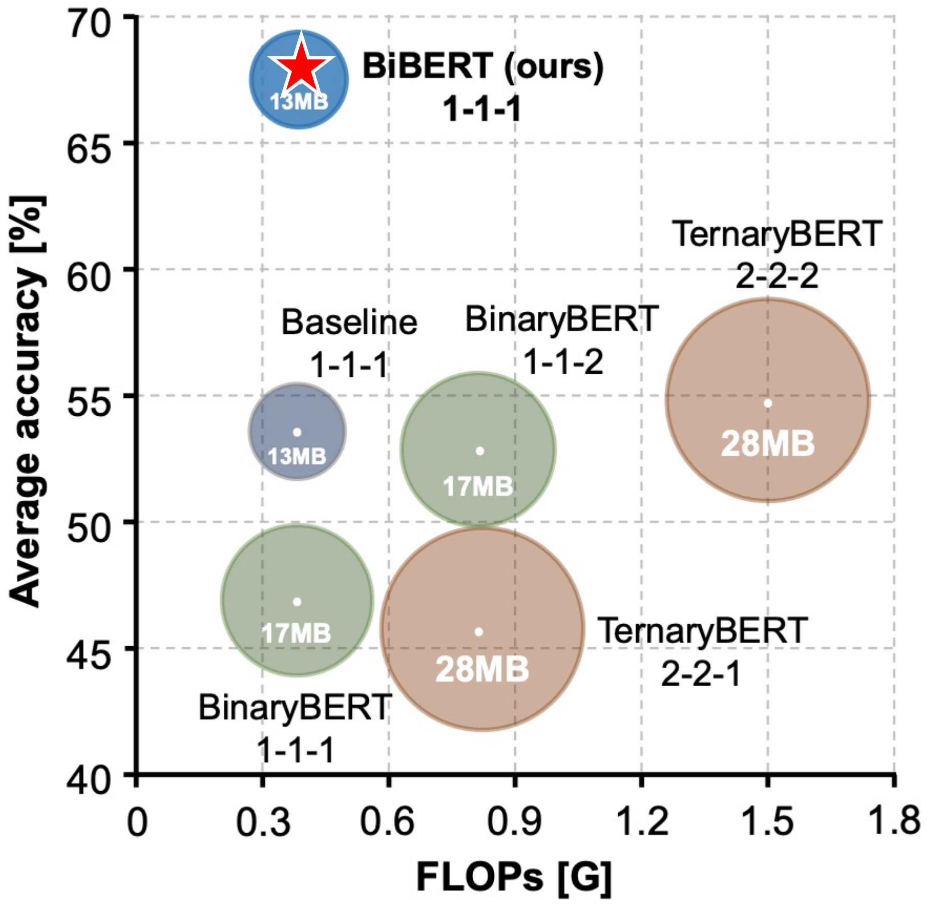
$$\mathbf{B_A} = \text{bool}(\mathbf{A}) = \text{bool}\left(\frac{1}{\sqrt{D}}\left(\mathbf{B_Q} \otimes \mathbf{B_K}^\top\right)\right)$$

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

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

Transformer Binarization: attention crash and recovery

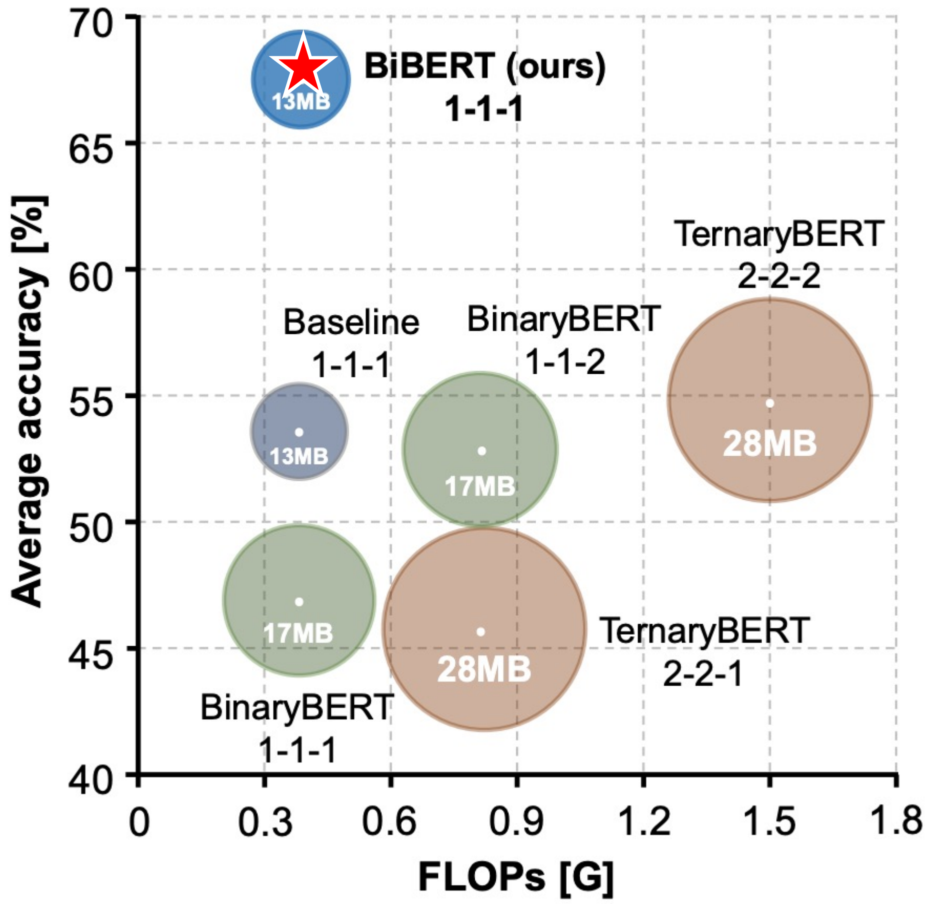
Performance

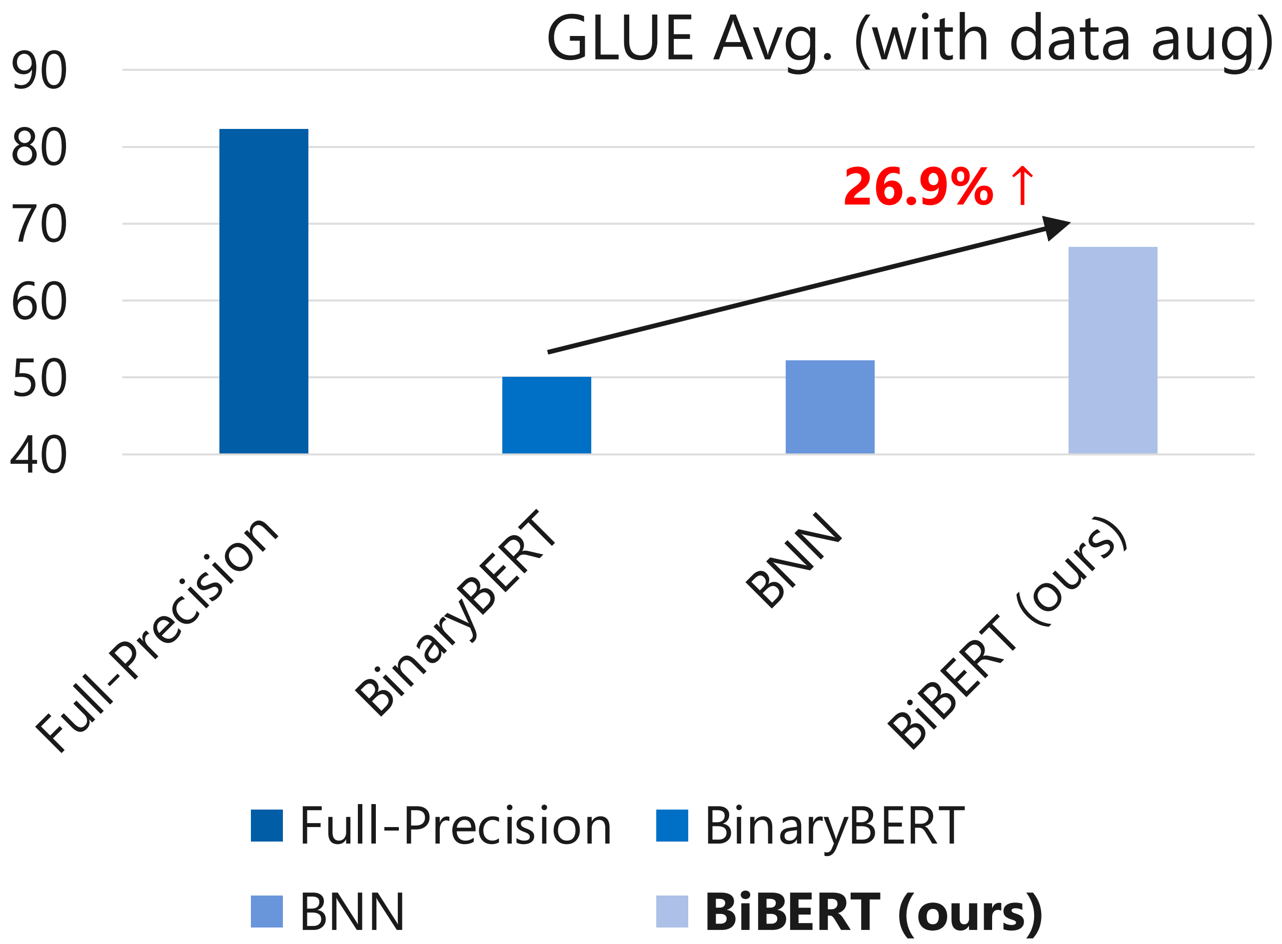


BiBERT: Accurate Fully Binarized BERT. Haotong Qin, et al. ICLR 2022.

Transformer Binarization: attention crash and recovery

Performance



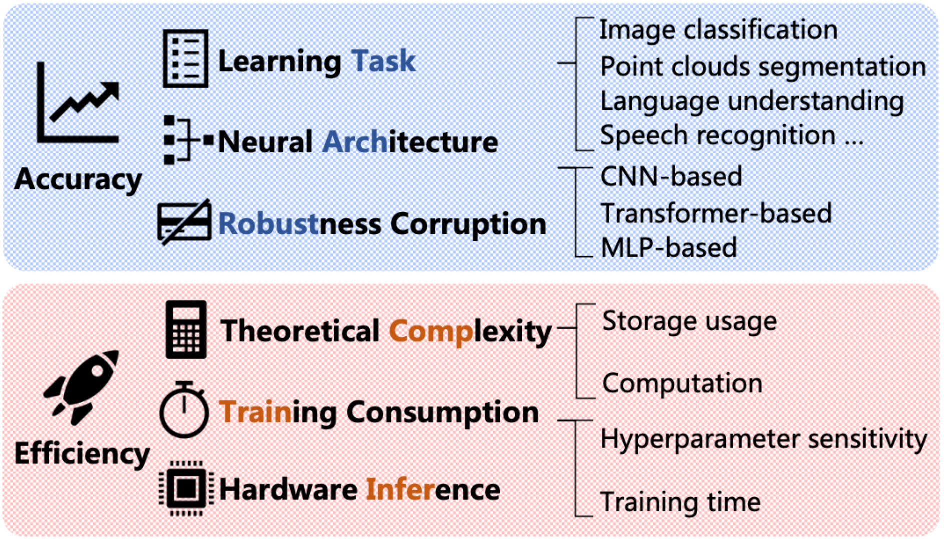


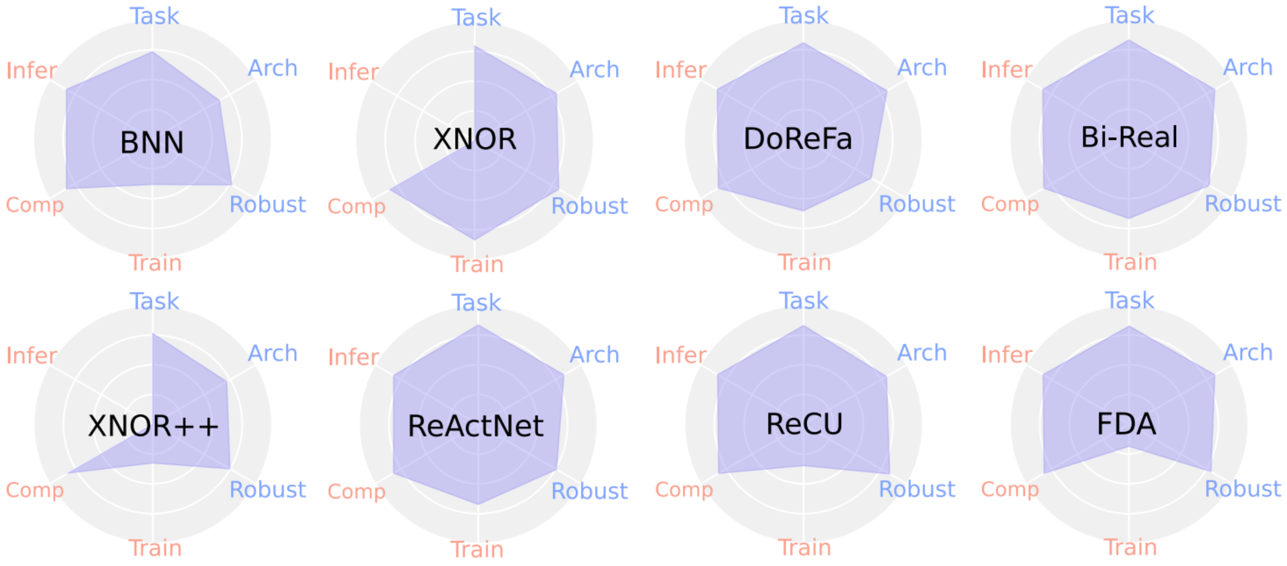
BiBERT: Accurate Fully Binarized BERT. Haotong Qin, et al. ICLR 2022.

Network Binarization: benchmark

BiBench*: Benchmarking and Analyzing Network Binarization*

Evaluation Tracks for 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

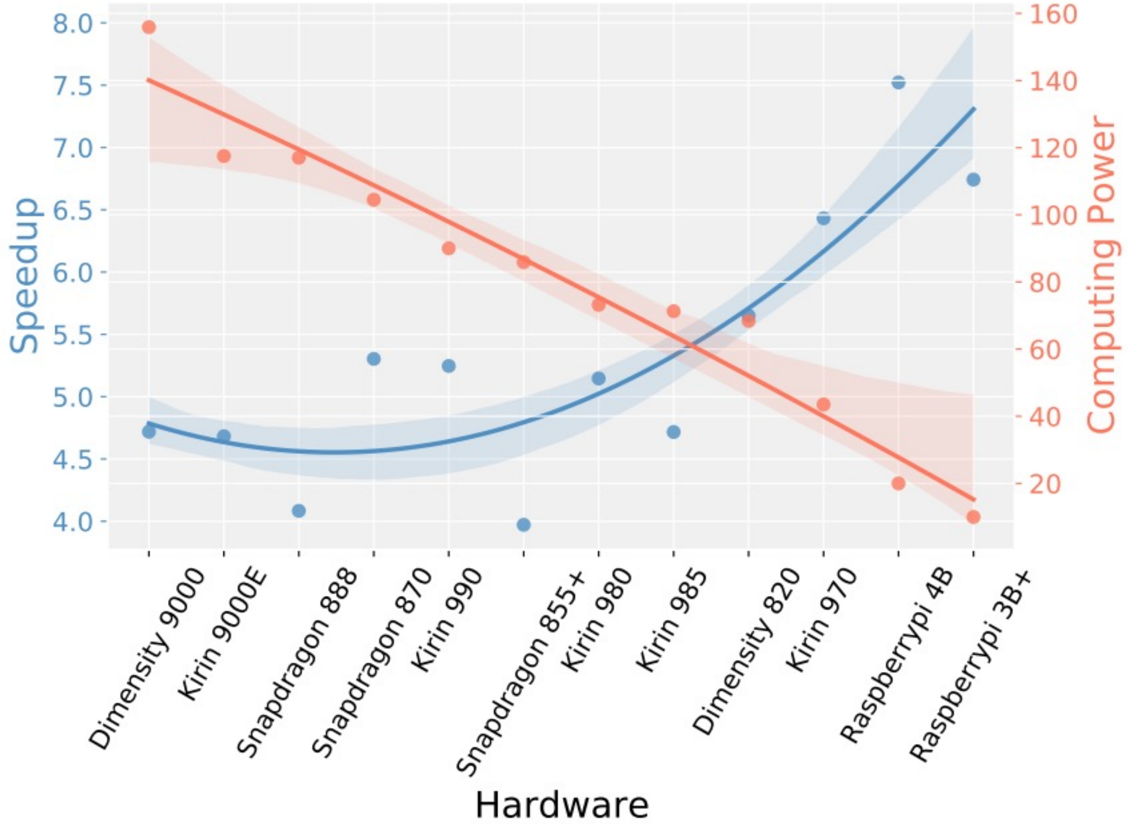
***BiBench**: Benchmarking and Analyzing Network Binarization*

The 3 Most Effective Techniques for Generic Binarization:

(1) Soft gradient approximation

(2) Channel-wise scaling factors

(3) Prebinarization parameter redistributing



Interesting Finding for Binarization:
Born for Edge

Network Binarization: survey

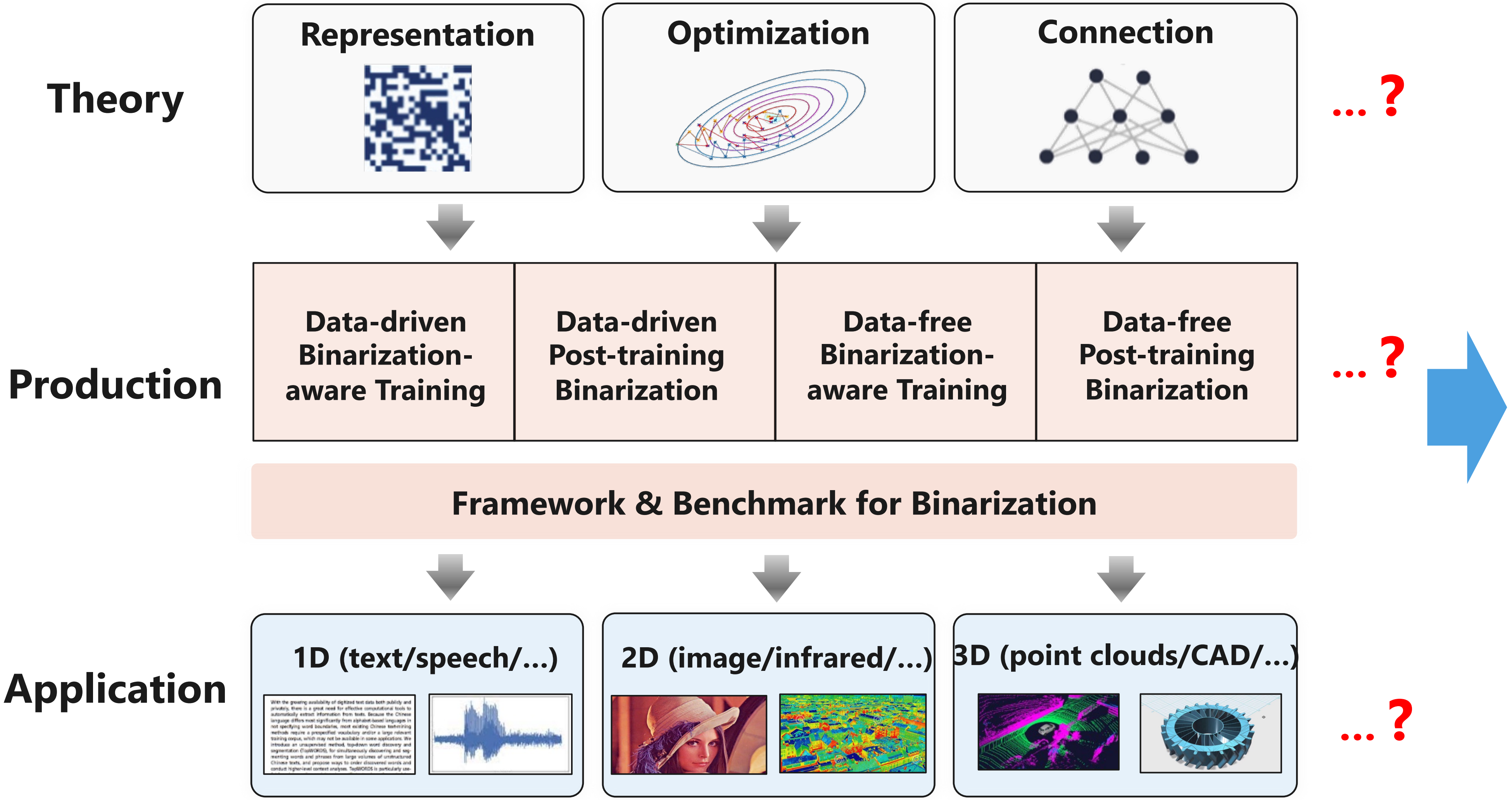
Binary Neural Networks: A Survey

Haotong Qin, et al.
Pattern Recognition
Volume 105, 2020
(132 citations)

Naive BNNs		BinaryConnect
		BNN
		Bitwise Neural Network
Optimization Based BNNs	**Minimize the Quantization Error**	XNOR-Net
		DoReFa-Net
		BWHN
		WRPN
		…
	Improve Network Loss Function	LAB
		RAD
		…
	Reduce the Gradient Error	HWGQ
		Bi-Real Net
		…

https://github.com/htqin/awesome-model-quantization/

Network Binarization: future



Paradigm of binarization

Data-free Training-free

Vision
Language
...

Thank you!

Q&A

Haotong Qin
ETH Zürich CVL & Beihang University