BiBERT: Accurate Fully Binarized BERT

Haotong Qin[1*] Yifu Ding[1*] Mingyuan Zhang[2*] Qinghua Yan[1]
Aishan Liu[1] Qingqing Dang[3] Ziwei Liu[2] Xianglong Liu[1†]

[1]Beihang University [2]Nanyang Technological University [3]Baidu Inc.

Paper: https://openreview.net/forum?id=5xEgrl_5FAJ
Code: https://github.com/htqin/BiBERT
(star is welcome)

1 Introduction: BERT Binarization

- **Large Pre-trained BERT**
  - BERT has achieved remarkable performance on NLP tasks
  - it still suffers expensive FP32 parameters and operations

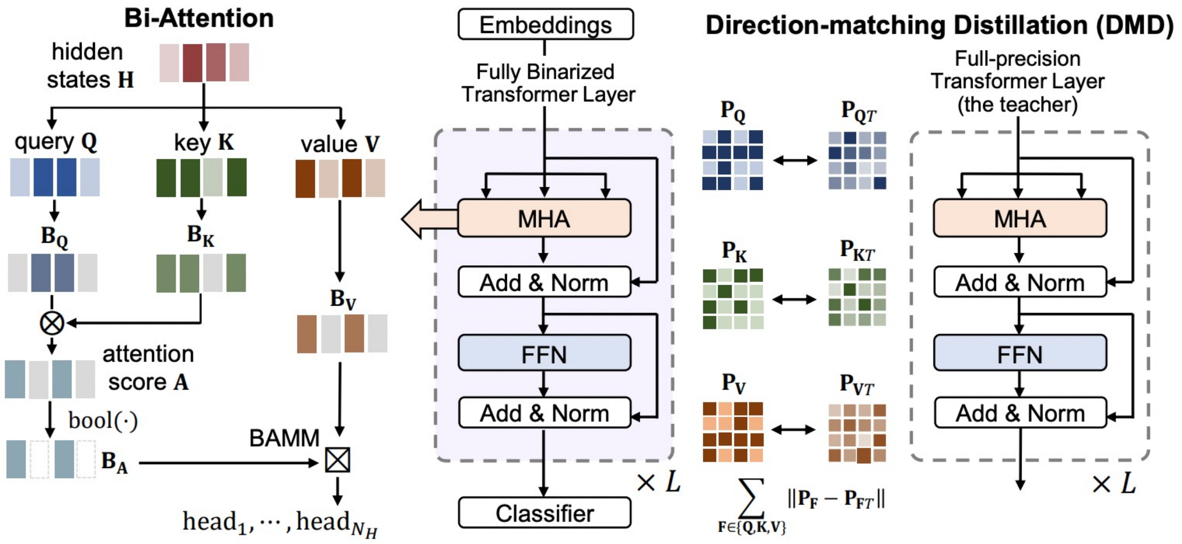
- **Network Binarization**
  - compression by binarizing parameters
  - accelerating by applying bitwise operations

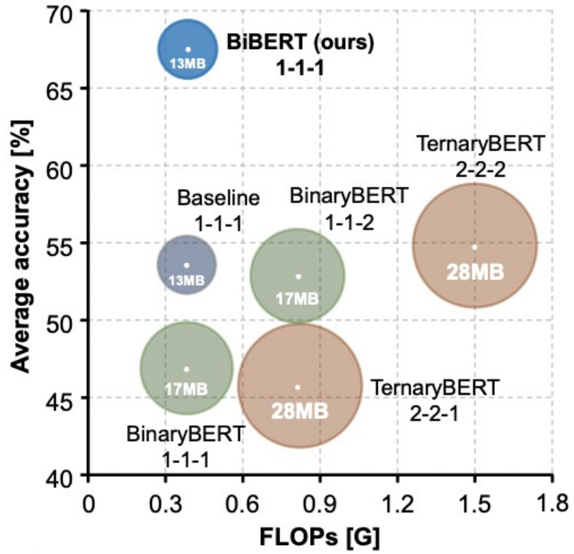
$$Q_{\mathrm{x}}(\mathbf{x}) = \alpha\, \mathbf{B}_{\mathbf{x}}$$

$$\mathbf{B}_{\mathbf{x}} = \operatorname{sign}(\mathbf{x}) = \begin{cases} -1, & \text{if } x \geq 0 \\ 1, & \text{otherwise} \end{cases}$$

$$z = Q_{\mathrm{w}}(\mathbf{w})^{\top} Q_{\mathrm{a}}(\mathbf{a}) = \alpha_{\mathrm{w}}\alpha_{\mathrm{a}}(\mathbf{B}_{\mathbf{w}} \otimes \mathbf{B}_{\mathbf{a}})$$

1 Introduction: Overview

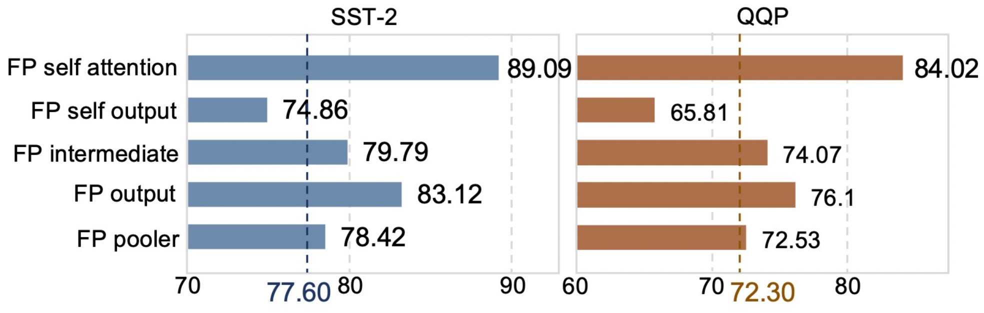




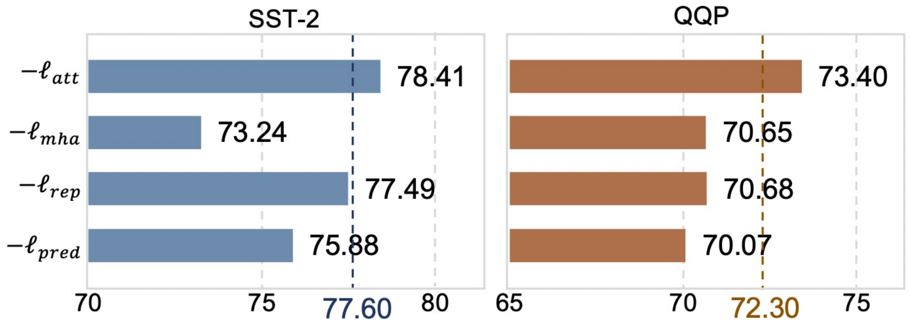
- **Main Contribution**
  - the first full binarization approaches to large pretrained BERTs;
  - identify the challenges that make existing binarization methods difficult to transfer to binarize BERTs, expecially their activation;
  - achieve impressive 56.3× and 31.2× saving on FLOPs and size.

2 The Rise of BiBERT: Bottlenecks of Binarized BERT

- **Binarized BERT Architecture**
  - **Architecture perspective.** Binarizing MHA brings the most significant drop of accuracy among all parts of the BERT. While binarizin FFN and pooler layers brings less harm to tl accuracy.

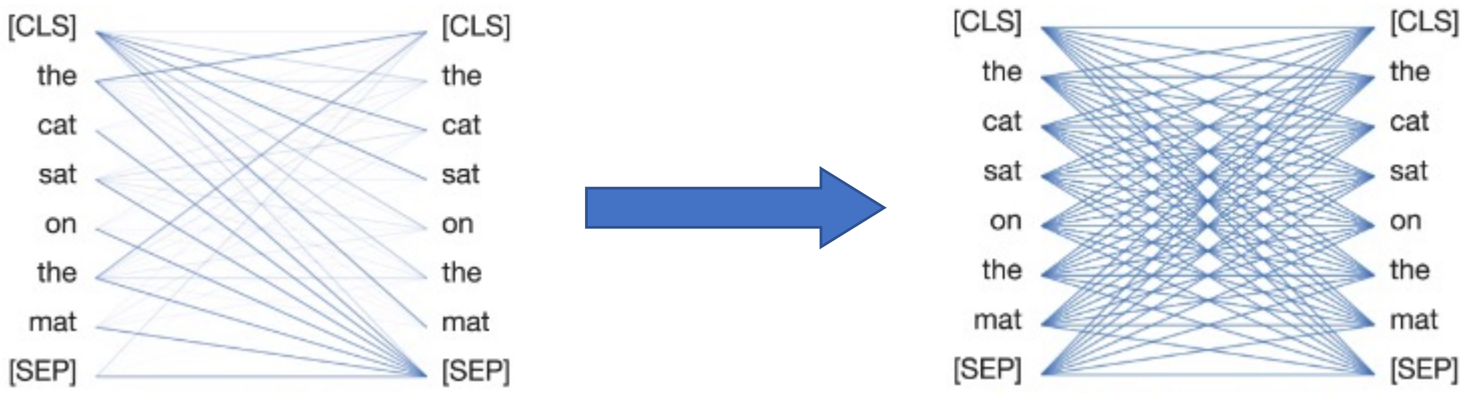


- **Distillation for Binarized BERT**
  - **Optimization perspective.** For most distillation terms, solely removing them in the distillation will harm the performance, however, the performance increases when the distillation loss of attention score is removed.



2 The Rise of BiBERT: Bi-Attention

- **Information Degradation in Attention Structure**
  - **Theorem 1**: Given $\mathbf{A} \in \mathbb{R}^{\boldsymbol{k}}$ with Gaussian distribution and the variable $\widehat{\mathbf{B}}_{\mathbf{A}}^{S}$ generated by $\widehat{\mathbf{B}}_{\mathbf{w}}^{A} = \mathrm{sign}(\mathrm{softmax}(\mathbf{A} - \tau))$, the threshold τ, which maximizes the information entropy $\mathcal{H}(\widehat{\mathbf{B}}_{\mathbf{A}}^{S})$, is negatively correlated to the number of elements k.



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

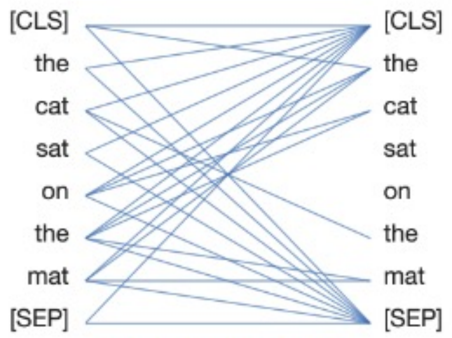
Information Degradation !

2 The Rise of BiBERT: Bi-Attention

- **Bi-Attention for Maximum Information Entropy**
  - The Bi-Attention binarize the attention weight into the Boolean value, while our design is driven by information entropy maximization:

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

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



(c) BiBERT (Ours)

  - where $\boxtimes$ is a Bitwise-Affine Matrix Multiplication (BAMM) operator composed by XNOR-Bitcount and bit-shift to align training and inference representations and perform efficient bitwise calculation.

2 The Rise of BiBERT: Direction-Matching Distillation

- **Direction Mismatch**
  - **Theorem 3**: Given the variables X and X_T follow $\mathcal{N}(0, \sigma_1), \mathcal{N}(0, \sigma_2)$ respectively, the proportion of optimization direction error is defined as $p_{\mathrm{error}Q-\mathrm{bit}} = p(\mathrm{sign}(\mathrm{X} - X_T) \neq \mathrm{sign}(\mathrm{quantize_Q}(X) - X_T))$, where $\mathrm{quantize_Q}$ denotes the $Q - \mathrm{bit}$ symmetric quantization. As Q reduces from 8 to 1, $p_{\mathrm{error}Q-\mathrm{bit}}$ becomes larger.

Bits (Q)	1	2	3	4	5	6	7	8
Proportion (%)	14.36%	6.42%	4.35%	3.30%	2.76%	2.56%	2.51%	2.49%

  - Besides, the activation scales in binarized and FP32 BERTs are significantly different since application the discrete binarization function.

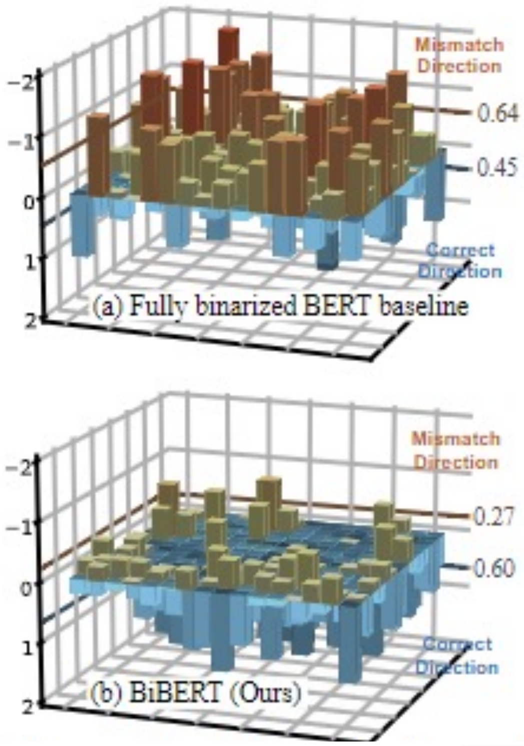
2 The Rise of BiBERT: Direction-Matching Distillation

- **DMD for Accurate Optimization**
  - DMD is designed to solve the optimization direction mismatch in the distillation of the BERT full binarization.
  - We first reselect the distilled activations for DMD, and then =construct similarity pattern matrices for distilling activation, which can be expressed as

$$\mathbf{P_Q} = \frac{\mathbf{Q} \times \mathbf{Q}^\top}{\|\mathbf{Q} \times \mathbf{Q}^\top\|}, \quad \mathbf{P_K} = \frac{\mathbf{K} \times \mathbf{K}^\top}{\|\mathbf{K} \times \mathbf{K}^\top\|}, \quad \mathbf{P_V} = \frac{\mathbf{V} \times \mathbf{V}^\top}{\|\mathbf{V} \times \mathbf{V}^\top\|},$$

$$\ell_{\text{distill}} = \ell_{\text{DMD}} + \ell_{\text{hid}} + \ell_{\text{pred}}, \qquad \ell_{\text{DMD}} = \sum_{l \in [1,L]} \sum_{\mathbf{F} \in \mathcal{F}_{\text{DMD}}} \|\mathbf{P}_{\mathbf{F}l} - \mathbf{P}_{\mathbf{F}Tl}\|,$$

  - By applying the DMD in BiBERT, we mitigate the direction mismatch of output caused by binarization.



Experiments: Precision Performance

Table 2: Comparison of BERT quantization methods without data augmentation.

Quant	#Bits	Size (MB)	FLOPs (G)	MNLI$_{m/mm}$	QQP	QNLI	SST-2	CoLA	STS-B	MRPC	RTE	Avg.
Full Precision	32-32-32	418	22.5	84.9/85.5	91.4	92.1	93.2	59.7	90.1	86.3	72.2	83.9
Q-BERT	2-8-8	43.0	6.5	76.6/77.0	–	–	84.6	–	–	68.3	52.7	–
Q2BERT	2-8-8	43.0	6.5	47.2/47.3	67.0	61.3	80.6	0	4.4	68.4	52.7	47.7
TernaryBERT	2-2-8	28.0	6.4	83.3/83.3	90.1	–	–	50.7	–	87.5	68.2	–
BinaryBERT	1-1-4	16.5	1.5	83.9/84.2	91.2	90.9	92.3	44.4	87.2	83.3	65.3	79.9
TernaryBERT	2-2-2	28.0	1.5	40.3/40.0	63.1	50.0	80.7	0	12.4	68.3	54.5	45.5
BinaryBERT	1-1-2	16.5	0.8	62.7/63.9	79.9	52.6	82.5	14.6	6.5	68.3	52.7	53.7
TernaryBERT	2-2-1	28.0	0.8	32.7/33.0	74.1	59.3	53.1	0	7.1	68.3	53.4	42.3
Baseline	1-1-1	13.4	0.4	45.8/47.0	73.2	66.4	77.6	11.7	7.6	70.2	54.1	50.4
Baseline$_{50\%}$	1-1-1	13.4	0.4	47.7/49.1	74.1	67.9	80.0	14.0	11.5	69.8	54.5	52.1
BinaryBERT	1-1-1	16.5	0.4	35.6/35.3	66.2	51.5	53.2	0	6.1	68.3	52.7	41.0
BinaryBERT$_{50\%}$	1-1-1	13.4	0.4	39.2/40.0	66.7	59.5	54.1	4.3	6.8	68.3	53.4	43.5
BiBERT (ours)	**1-1-1**	**13.4**	**0.4**	**66.1/67.5**	**84.8**	**72.6**	**88.7**	**25.4**	**33.6**	**72.5**	**57.4**	**63.2**
Full Precision $_{6L}$	32-32-32	257	11.3	84.6/83.2	71.6	90.4	93.1	51.1	83.7	87.3	70.0	79.4
BiBERT$_{6L}$ (ours)	**1-1-1**	**6.8**	**0.2**	**63.6/63.7**	**83.3**	**73.6**	**87.9**	**24.8**	**33.7**	**72.2**	**55.9**	**62.1**
Full Precision $_{4L}$	32-32-32	55.6	1.2	82.5/81.8	71.3	87.7	92.6	44.1	80.4	86.4	66.6	77.0
BiBERT$_{4L}$ (ours)	**1-1-1**	**4.4**	**0.03**	**55.3/56.1**	**78.2**	**71.2**	**85.4**	**14.9**	**31.5**	**72.2**	**54.2**	**57.7**

Table 3: Comparison of BERT quantization methods with data augmentation.

Quant	#Bits	Size (MB)	FLOPs (G)	QNLI	SST-2	CoLA	STS-B	MRPC	RTE	Avg.
Full Precision	32-32-32	418	22.5	92.1	93.2	59.7	90.1	86.3	72.2	82.3
TernaryBERT	2-2-8	28.0	6.4	90.0	92.9	47.8	84.3	82.6	68.4	77.8
BinaryBERT	1-1-4	16.5	1.5	91.4	93.7	53.3	88.6	86.0	71.5	80.8
TernaryBERT	2-2-2	28.0	1.5	50.0	87.5	20.6	72.5	72.0	47.2	58.3
BinaryBERT	1-1-2	16.5	0.8	51.0	89.6	33.0	11.4	71.0	55.9	52.0
TernaryBERT	2-2-1	28.0	0.8	50.9	80.3	6.5	10.3	71.5	53.4	45.5
Baseline	1-1-1	13.4	0.4	69.2	84.0	23.3	14.4	71.4	50.9	52.2
BinaryBERT	1-1-1	16.5	0.4	66.1	78.3	7.3	22.1	69.3	57.7	50.1
BiBERT (ours)	**1-1-1**	**13.4**	**0.4**	**76.0**	**90.9**	**37.8**	**56.7**	**78.8**	**61.0**	**67.0**
Full Precision $_{6L}$	32-32-32	257	11.3	90.4	93.1	51.1	83.7	87.3	70.0	79.2
BiBERT$_{6L}$ (ours)	**1-1-1**	**6.8**	**0.2**	**76.0**	**90.7**	**35.6**	**62.7**	**77.9**	**57.4**	**66.7**
Full Precision $_{4L}$	32-32-32	55.6	1.2	87.7	92.6	44.1	80.4	86.4	66.6	76.2
BiBERT$_{4L}$ (ours)	**1-1-1**	**4.4**	**0.03**	**73.2**	**88.3**	**20.0**	**42.5**	**74.0**	**56.7**	**59.1**

Conclusion

- **Novel Technique:** the first full binarization approaches for large pretrained BERT models.
- **Theoretical Analysis:** present theoretical formulations of the phenomenons (Information Degradation in Attention Structure & Dirsction Mismatch) applying full binarization for BERTs.
- **Good Precision:** show improvements of full BERT binarization than existing methods across several mainstream NLP tasks.
- **High efficiency:** achieves impressive $\mathbf{56.7}\times$ computational FLOPs and $\mathbf{31.2}\times$ storage saving.

Thank you!

Paper: https://openreview.net/forum?id=5xEgrl_5FAJ
Code: https://github.com/htqin/BiBERT
(star is welcome)