BiBERT: Accurate Fully Binarized BERT

ACL 2023

Haotong Qin*[1,4], Yifu Ding*[1,4], Mingyuan Zhang*[2], Qinghua Yan[1], Aishan Liu[1],
Qingqing Dang[3], Ziwei Liu[2], Xianglong Liu[1]

[1]State Key Lab of Software Development Environment, Beihang University
[2]S-Lab, Nanyang Technological University [3]Baidu Inc.
[4]Shen Yuan Honors College, Beihang University

Background

- **Compact deep learning model**
  - Pre-trained language model
    - Great power in various natural language processing (NLP) tasks
    - BERT (Devlin et al., 2018) significantly improves the state-of-the-art performance,
    - Massive parameters hinder their widespread deployment on edge devices in the real world
  - Model compression
    - Alleviate resource constraint issues
    - Quantization, distillation, pruning, parameter sharing, etc
  - Quantization
    - Obtain compact model by compressing parameters to lower bit-width representation
    - Representation limitation and optimization difficulties trigger severe performance drop
  - Knowledge distillation
    - Common remedy in quantization as an auxiliary optimization approach to tackle the performance drop
    - Encourages quantized model to mimic full-precision model to exploit knowledge in teacher's representation (Bai et al., 2020)

Haoli Bai et al. Binarybert: Pushing the limit of BERT quantization. ACL. 2021.

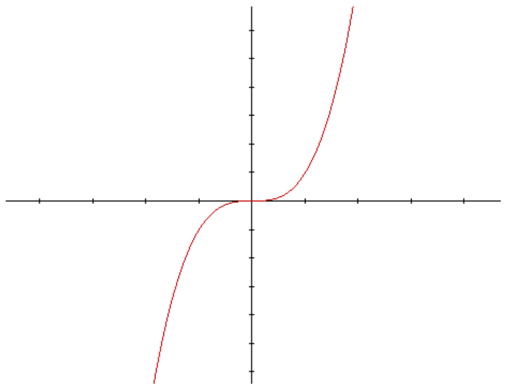
Background

- **Quantization**
  - Convert continuous value to discrete value in certain interval
  - Continuous value
    - IEEE 754 Format
    - FP32

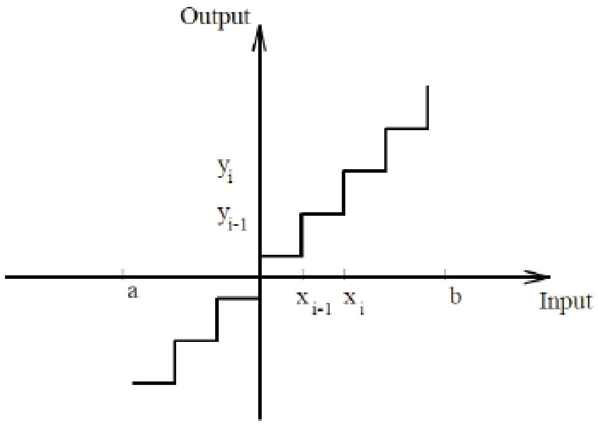
      Standard type for Training, High Resource, Slow Inference
    - FP16

      Standard type for Inference, Low Resource, Fast Inference

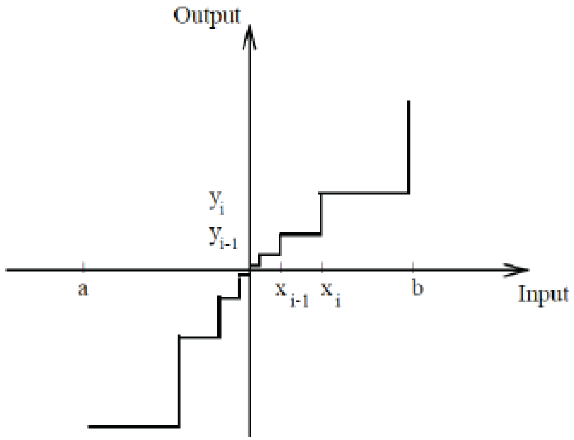
$$F : X \to Y$$
$$\mathbb{R} \to \mathbb{U}$$



Continuous function



Uniform quantization function



Non-uniform quantization function

T. Chen. 'Seismic Data compression: a tutorial.' 1995.
Hao Wu, Patrick Judd, Xiaojie Zhang, Mikhail Isaev, and Paulius Micikevicius. Integer Quantization for Deep Learning Inference: Principles and Empirical Evaluation. arXiv preprint arXiv:2004.09602, 2020.

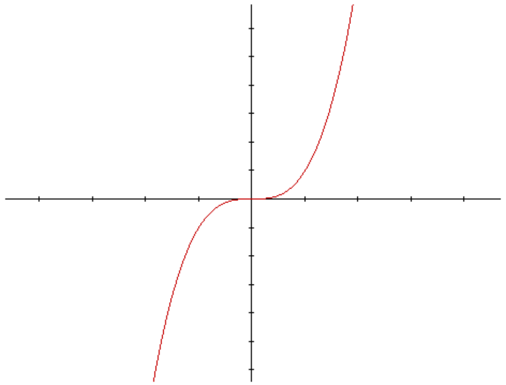
Background

- **Quantization**
  - Convert continuous value to discrete value in certain interval
  - Discrete value
    - Uniform quantization
    - Non-uniform quantization

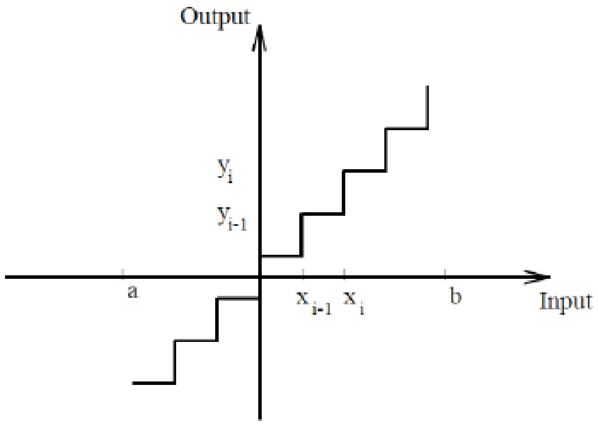
$$F: X \to Y$$
$$\mathbb{R} \to \mathbb{U}$$

$$\{y \mid -128 \leq y \leq 127\}$$

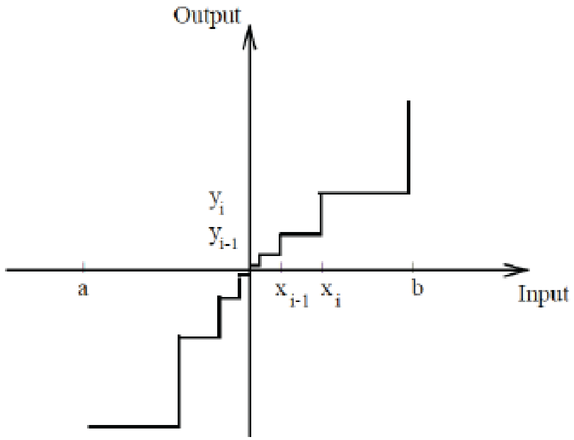
$$f(x) = \begin{cases} 1, & \text{if } 0 \leq x \leq 1 \\ 2, & \text{if } 1 \leq x \leq 10 \\ 3, & \text{if } x \geq 10 \end{cases}$$



Continuous function



Uniform quantization function



Non-uniform quantization function

T. Chen. 'Seismic Data compression: a tutorial.' 1995.
Hao Wu, Patrick Judd, Xiaojie Zhang, Mikhail Isaev, and Paulius Micikevicius. Integer Quantization for Deep Learning Inference: Principles and Empirical Evaluation. arXiv preprint arXiv:2004.09602, 2020.

Background

- **Quantization**
  - Quantization target
    - Weight quantization, Activations quantization, Gradients quantization

Components	Benefits	Challenges
Weight	Smaller model size Faster forward training & inference Less energy	Hard to converge with quantization weights Require Approximate gradients Accuracy Degradation
Activations	Smaller memory foot print during training Allows replacement of dot-product by bitwise operations Less energy	Gradient mismatch problem
Gradients	Communication & memory savings	Convergence requirement

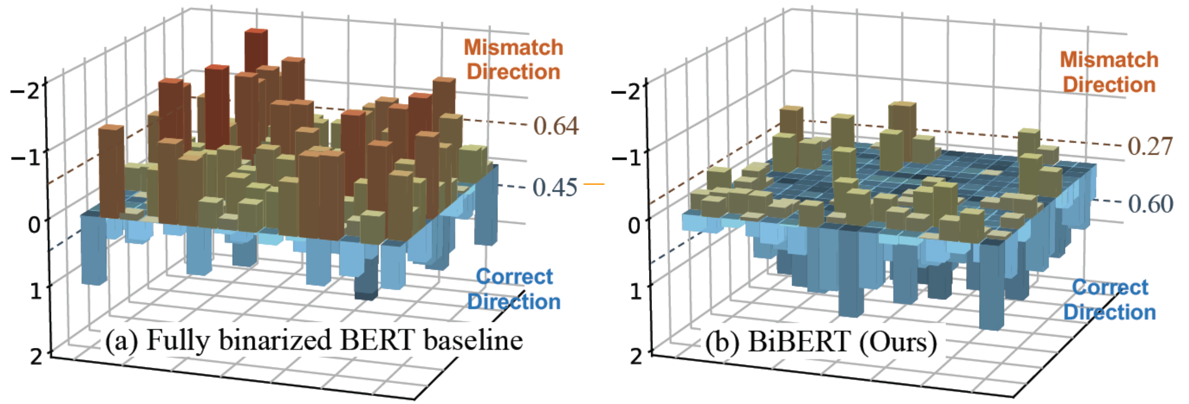
BiBERT

Straight-through estimator (STE, 2013)

Pierre Stock, Armand Joulin, Rémi Gribonval, Benjamin Graham, Hervé Jégou. And the Bit Goes Down: Revisiting the Quantization of Neural Networks. In ICLR 2020. arXiv preprint arXiv:1907.05686, 2020

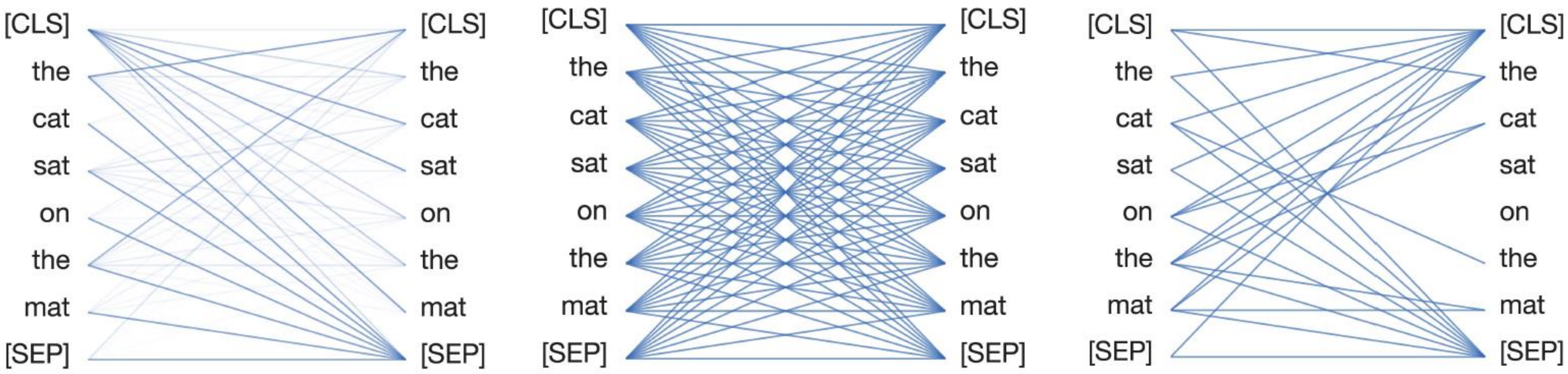
Part 1. # Background

- **Analysis**
  - Direct binarization
    - Degradation of the information of attention weight
    - Invalidation of the selection ability for attention mechanism
    - Severe optimization direction mismatch since the non-neglectable error between the defacto and expected optimization direction



Visualization of direction mismatch

Attention-head view (Vig, 2019)



(a) Full-precision (b) Fully binarized BERT baseline (c) BiBERT (Ours)

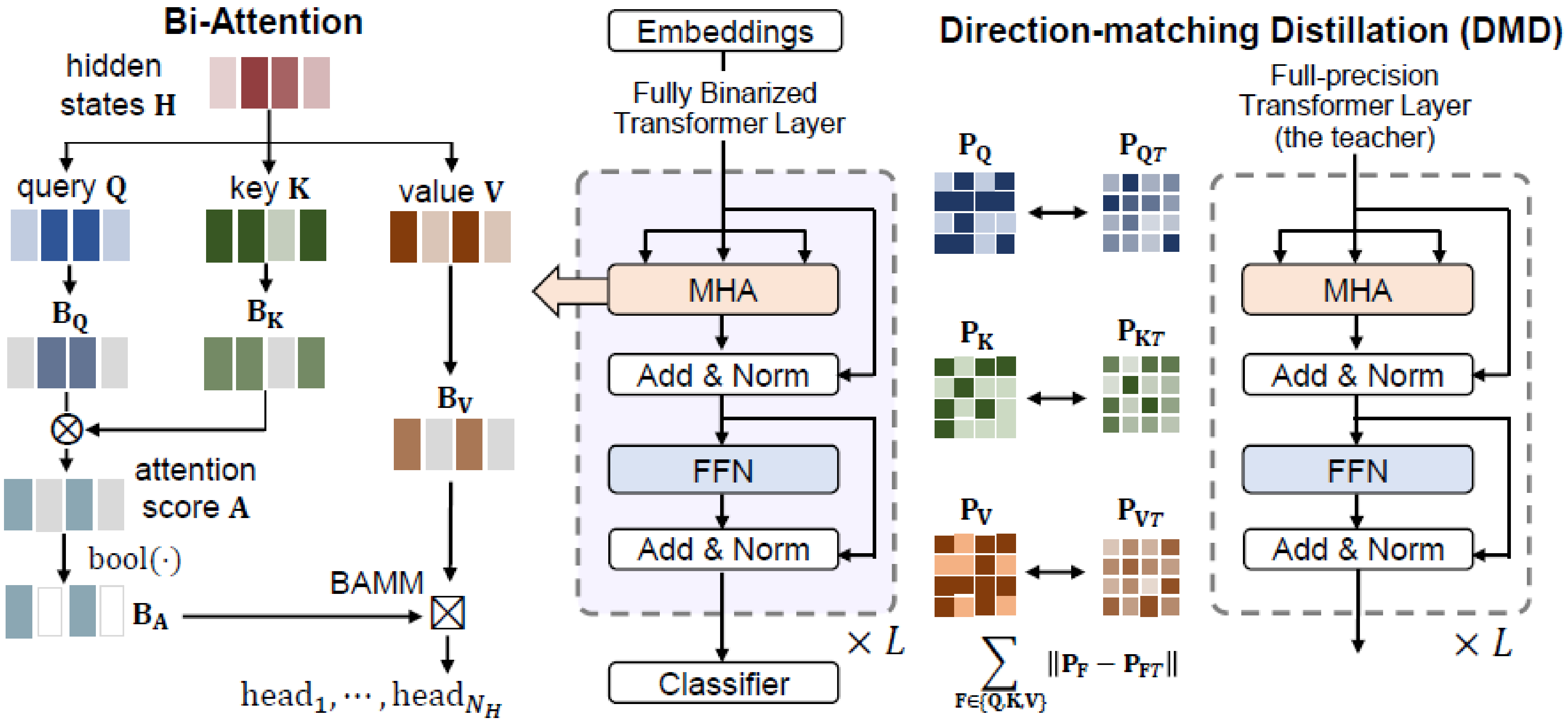
Jesse Vig. A multiscale visualization of attention in the transformer model. ACL. 2019.

Introduction

- **BiBERT (Fully binarized BERT baseline)**
  - Bi-Attention
    - Efficient structure based on information theory
    - Binarized representations with maximized information entropy, allowing the model to restore the perception of input contents
  - Direction-matching Distillation (DMD)
    - Scheme to eliminate the direction mismatch in distillation
    - Appropriate activation and utilizes knowledge from constructed similarity matrices in distillation to optimize accurately
  - Experiments on GLUE benchmark
    - Average accuracy of BiBERT exceeds 1-1-1 bit-width BinaryBERT, 2-8-8 bit-width Q2BERT
    - Impressive 56.3X and 31.2X saving on FLOPs and model size, respectively

Introduction

- **Architecture Overview**



Method

- **Binarized Architecture**
  - Forward & backward propagation
    - Sign function is applied in the forward propagation
    - Straight-through estimator (STE) (Bengio et al., 2013) is used to obtain the derivative in the backward propagation
  - Weight of binarized linear layers
    - Redistribute the weight to zero-mean for retaining representation information (Rastegari et al., 2016
    - Apply scaling factors to minimize quantization errors (Rastegari et al., 2016)

Binarized linear layers

Scaling factors for weight

$$\text{bi-linear}(\mathbf{X}) = \alpha_{\mathbf{w}}(\text{sign}(\mathbf{X}) \otimes \text{sign}(\mathbf{W} - \mu(\mathbf{W}))), \quad \alpha_{\mathbf{w}} = \frac{1}{n}\|\mathbf{W}\|_{\ell 1}$$

$\otimes$ **Matrix multiplication with bitwise xnor and bitcount**

Yoshua Bengio et al. Estimating or propagating gradients through stochastic neurons for conditional computation. arXiv preprint arXiv:1308.3432. 2013
Mohammad Rastegari et al. Xnor-net: Imagenet classification using binary convolutional neural networks. ECCV. 2016.

Method

- **Binarized Architecture**
  - Multi-Head Attention (MHA) module & Feed-Forward Network (FFN)
    - Sign function is applied in the forward propagation
    - Straight-through estimator (STE) (Bengio et al., 2013) is used to obtain the derivative in the backward propagation

Binarized linear layers

$$\mathbf{Q} = \text{bi-linear}_Q(\mathbf{H}), \quad \mathbf{K} = \text{bi-linear}_K(\mathbf{H}), \quad \mathbf{V} = \text{bi-linear}_V(\mathbf{H})$$

Attention score

$$\mathbf{A} = \frac{1}{\sqrt{D}}\left(\mathbf{B_Q} \otimes \mathbf{B_K}^\top\right), \quad \mathbf{B_Q} = \text{sign}(\mathbf{Q}), \quad \mathbf{B_K} = \text{sign}(\mathbf{K})$$

Binarize the attention weight

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

Yoshua Bengio et al. Estimating or propagating gradients through stochastic neurons for conditional computation. arXiv preprint arXiv:1308.3432. 2013
Mohammad Rastegari et al. Xnor-net: Imagenet classification using binary convolutional neural networks. ECCV. 2016.

Method

- **Bi-Attention For Maximum Information Entropy**
  - The average mutual information Z(X;Y) between input and output

$$I(X;Y) = H(Y) - H(Y \mid X) = H(Y) \qquad p_k = \Pr\{Y_k\} = 1/N.$$

  - Maximum-Output Entropy (MOE) quantizer $D_\theta = \sum_{k=1}^{N} \int_{X_k}^{X_{k+1}} p(x)|x - Y_k|^\theta \, dx,$
  - An approximate relationship for the MAE quantizer $\int_{X_k}^{X_{k+1}} p^{1/(1+\theta)}(x)\, dx \cong \frac{2}{N} \int_0^\infty p^{1/(1+\theta)}(x)\, dx.$
  - If p(x) is uniform or exponential in form (as the Gaussian or Laplacian distributions

$$p^{1/(1+\theta)}(x) = Ap^*(x), \quad \Rightarrow \quad \int_{X_k}^{X_{k+1}} p^*(x)\, dx \cong 1/N,$$

  - MAE quantizer is approximately same as MOE quantizer optimized with respect to p*(x).
  - MOE quantizer is approximately MAE quantizer scaled by the constant multiplicative fac tor C^-1.

$$p^*(x) = C^{-1} p(x/C),$$

D Messerschmitt. Quantizing for maximum output entropy (corresp.). IEEE Transactions on Information Theory, 1971.

Method

- **Bi-Attention For Maximum Information Entropy**
  - To Maximize information entropy of binarized representation zero-mean pre-binarized attention weight multiply Value

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

Since the information entropy of $\mathrm{sign}(\mathrm{softmax}_{K+1}(A_1 - \tau_{K+1}))$ is maximized, we have

$$\mathcal{H}(\mathbf{B}) = -\sum_B p(B)\log p(B), \qquad \underbrace{\int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty}}_{K-1} \int_{-\infty}^{\infty} \prod_{i=1}^{K} p_{A_i}(a_i)[\mathrm{softmax}_K(a_1) \le \tau_K]\, da_i = 0.5,$$

  - Binarized attention score approximates the Gaussian distribution

$$A_{ij} = \sum_{l=1}^{D} B_{\mathbf{Q},il} \times B_{\mathbf{K},jl}, \qquad B_{\mathbf{Q},il} \times B_{\mathbf{K},jl} = \begin{cases} 1, & \text{if } B_{\mathbf{Q},il} \veebar B_{\mathbf{K},jl} = 1 \\ -1, & \text{if } B_{\mathbf{Q},il} \veebar B_{\mathbf{K},jl} = -1. \end{cases}$$

$$p_A(2i - D) = C_D^i\, p_e^i (1 - p_e)^{D-i} \simeq \frac{1}{\sqrt{2\pi D p_e (1 - p_e)}} e^{-\frac{(i - D p_e)^2}{2 D p_e (1 - p_e)}},$$

D Messerschmitt. Quantizing for maximum output entropy (corresp.). IEEE Transactions on Information Theory, 1971.

Method

- **Bi-Attention For Maximum Information Entropy**
  - Binarize the attention score A can maximize the information entropy of representation

Theorem 2. *When the binarized query* $\mathbf{B_Q} = \mathrm{sign}(\mathbf{Q}) \in \{-1, 1\}^{N \times D}$ *and key* $\mathbf{B_K} = \mathrm{sign}(\mathbf{K}) \in \{-1, 1\}^{N \times D}$ *are entropy maximized in binarized attention, the probability mass function of each element* $\mathbf{A}_{ij}$, $i, j \in [1, N]$ *sampled from attention score* $\mathbf{A} = \mathbf{B_Q} \otimes \mathbf{B_K}^\top$ *can be represented as* $p_A(2i - D) = 0.5^D C_D^i$, $i \in [0, D]$, *which approximates the Gaussian distribution* $\mathcal{N}(0, D)$.

  - By applying bool function, the elements in attention weight with lower value are binarized to 0, thus trivially get that $\phi(\tau, \mathbf{A}) = 0$

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

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

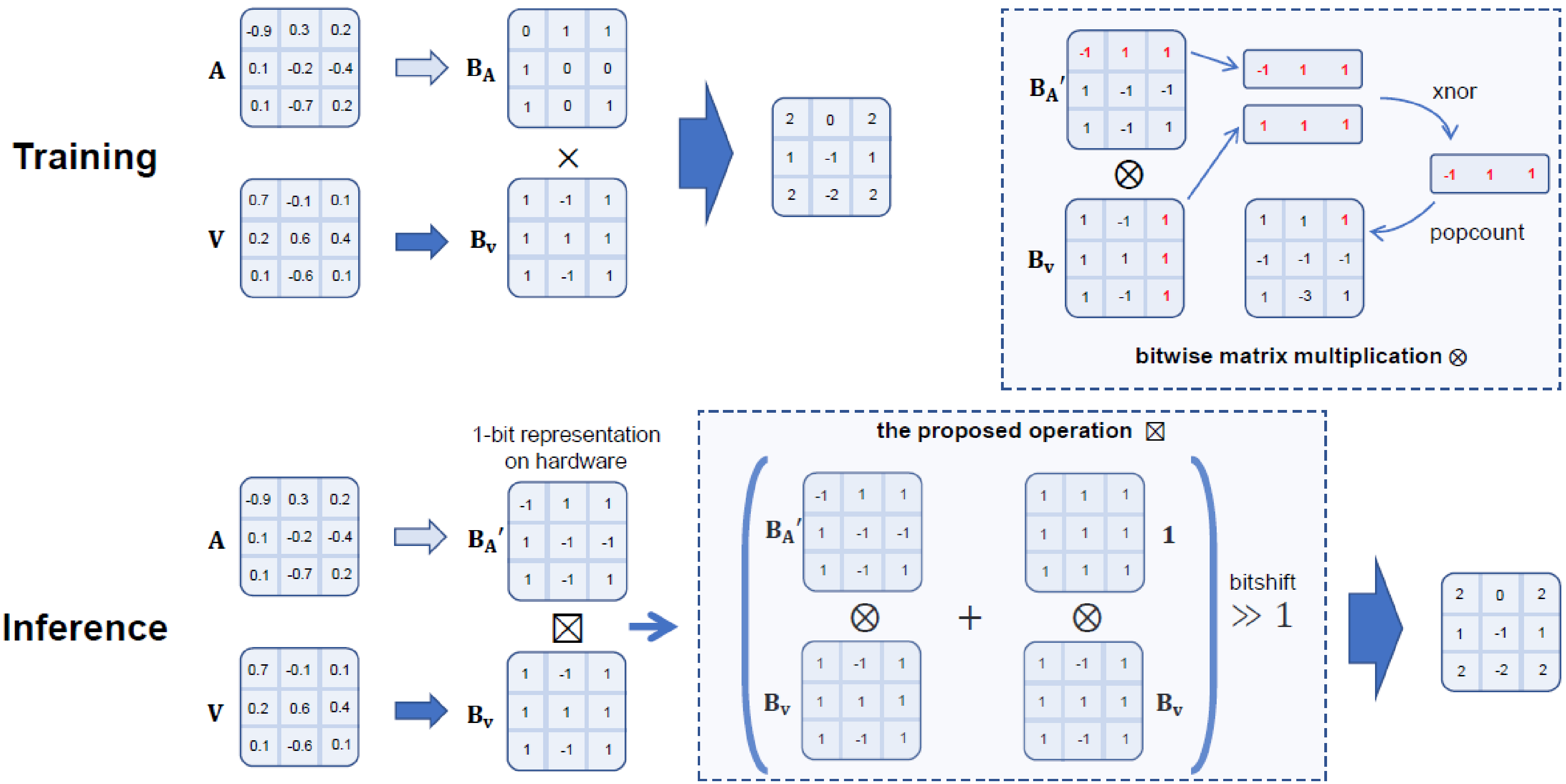
$$\mathbf{B_A} = \mathrm{bool}\,(\mathbf{A}) = \mathrm{bool}\left(\frac{1}{\sqrt{D}}\left(\mathbf{B_Q} \otimes \mathbf{B_K}^\top\right)\right)$$

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

D Messerschmitt. Quantizing for maximum output entropy (corresp.). IEEE Transactions on Information Theory, 1971.

Method

- **Bi-Attention For Maximum Information Entropy**
  - Apply bool function to obtain the binarized attention weights B with values of 0 and 1
  - 1-bit matrix stored in the hardware is unified into the same form (with binary values of 1 and -1) and is supported by most existing hardware



Method

- **Distillation For Binarized BERT**
  - Optimization approach to alleviate the performance drop of quantized BERT
  - Loss function
    - Use mean squared errors (MSE)
    - Measure the difference between student and teacher networks

Attention Loss

$$\ell_{\text{att}} = \sum_{l=1}^{L} \text{MSE}(\mathbf{A}_l, \mathbf{A}_{Tl}),$$

Multi Head Attention Loss

$$\ell_{\text{mha}} = \sum_{l=1}^{L} \text{MSE}(\mathbf{M}_l, \mathbf{M}_{Tl}),$$

Hidden States Loss

$$\ell_{\text{hid}} = \sum_{l=1}^{L} \text{MSE}(\mathbf{H}_l, \mathbf{H}_{Tl})$$

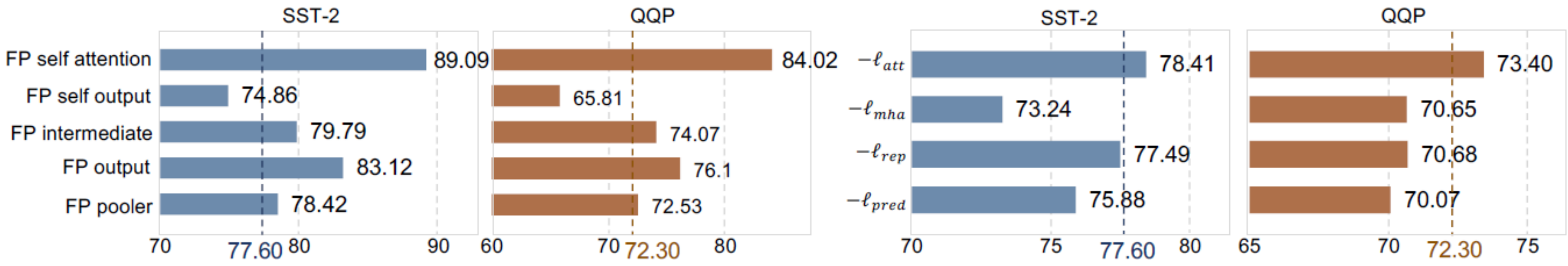
Prediction Loss

$$\ell_{\text{pred}} = \text{SCE}\left(\mathbf{y}, \mathbf{y}_T\right)$$

Loss function

$$\ell_{\text{distill}} = \ell_{\text{att}} + \ell_{\text{mha}} + \ell_{\text{hid}} + \ell_{\text{pred}}$$

Loss function



(a) Architecture perspective (b) Optimization perspective

Method

- **Direction-matching Distillation**
  - Direction mismatch for optimization in the fully binarized BERT baseline

Direction-matching Distillation Loss

$$\mathbf{P}_{\mathbf{Q}} = \frac{\mathbf{Q} \times \mathbf{Q}^\top}{\|\mathbf{Q} \times \mathbf{Q}^\top\|} \quad \mathbf{P}_{\mathbf{K}} = \frac{\mathbf{K} \times \mathbf{K}^\top}{\|\mathbf{K} \times \mathbf{K}^\top\|} \quad \mathbf{P}_{\mathbf{V}} = \frac{\mathbf{V} \times \mathbf{V}^\top}{\|\mathbf{V} \times \mathbf{V}^\top\|} \quad \ell_{\mathrm{DMD}} = \sum_{l \in [1,L]} \sum_{\mathbf{F} \in \mathcal{F}_{\mathrm{DMD}}} \|\mathbf{P}_{\mathbf{F}l} - \mathbf{P}_{\mathbf{F}Tl}\|$$

Loss function

$$\ell_{\mathrm{distill}} = \ell_{\mathrm{att}} + \ell_{\mathrm{mha}} + \ell_{\mathrm{hid}} + \ell_{\mathrm{pred}}$$

$$\ell_{\mathrm{distill}} = \ell_{\mathrm{DMD}} + \ell_{\mathrm{hid}} + \ell_{\mathrm{pred}}$$

Method

- **Mismatched Direction**

Theorem 4. *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 $p_{\text{error }Q\text{-bit}} = p(\text{sign}(X - X_T) \neq \text{sign}(\text{quantize}_Q(X) - X_T))$, where quantize_Q denotes the Q-bit symmetric quantization. As Q reduces from 8 to 1, $p_{\text{error }Q\text{-bit}}$ becomes larger.*

Proof. Given the random variables $X \sim \mathcal{N}(0,\sigma_1)$ and $X_T \sim \mathcal{N}(0,\sigma_2)$, the Q-bit symmetric quantization function quantize_Q is expressed as (take X as an example)

$$\text{quantize}_Q(X) = \begin{cases} -L, & \text{if } x < -L, \\ \lfloor \frac{(2^Q-1)X}{2L} + 0.5 \rfloor \frac{2L}{2^Q-1}, & \text{if } -L \leq X \leq L, \\ L, & \text{if } x > L, \end{cases} \tag{61}$$

where the $\lfloor \cdot \rfloor$ denotes the round down function, and the range $[-L, L]$ is divided into $2^Q - 1$ inter. The optimization direction error occurs when $\text{sign}(X - \hat{X}) = \text{sign}(\text{quantize}_Q(X) - X_T)$, i.e., $X > X_T$ *and* $\text{quantize}_Q(X) < X_T$ **or** $X < X_T$ *and* $\text{quantize}_Q(X) > X_T$.

Method

- **Mismatched Direction**

(1) When $-L < X < L$, $\lfloor \frac{(2^Q-1)X}{2L} \rfloor \frac{2L}{2^Q-1} <$ quantize$_Q(X) < \lfloor \frac{(2^Q-1)X}{2L} + 1 \rfloor \frac{2L}{2^Q-1}$.

a) If $X_T < \lfloor \frac{(2^Q-1)X}{2L} \rfloor \frac{2L}{2^Q-1}$, since $\lfloor \frac{(2^Q-1)X}{2L} \rfloor \frac{2L}{2^Q-1} < X$, we have $X_T < X$. And since $\lfloor \frac{(2^Q-1)X}{2L} \rfloor \frac{2L}{2^Q-1} <$ quantize$_Q(X)$, $X_T <$ quantize$_Q(X)$. Thus, the optimization direction is always right in this case.

b) If $X_T > \lfloor \frac{(2^Q-1)X}{2L} + 1 \rfloor \frac{2L}{2^Q-1}$, since $\lfloor \frac{(2^Q-1)X}{2L} + 1 \rfloor \frac{2L}{2^Q-1} > X$, we have $X_T > X$. And since $\lfloor \frac{(2^Q-1)X}{2L} \rfloor \frac{2L}{2^Q-1} >$ quantize$_Q(X)$, $X_T >$ quantize$_Q(X)$. Thus, the optimization direction is always right in this case.

c) If $\lfloor \frac{(2^Q-1)X}{2L} \rfloor \frac{2L}{2^Q-1} \leq X_T \leq \lfloor \frac{(2^Q-1)X}{2L} + 1 \rfloor \frac{2L}{2^Q-1}$, first, the probability of $X > X_T$ *and* quantize$_Q(X) < X_T$ can be calculated as:

Method

- **Mismatched Direction**

$$p_{\text{error1}\ Q} = \int_{-L}^{L} \int_{\lfloor \frac{(2^Q-1)X}{2L} + 0.5 \rfloor \frac{2L}{2^Q-1}}^{X} f_{X,X_T}(X, X_T) \tag{62}$$

$$\left[\lfloor \frac{(2^Q-1)X}{2L} + 0.5 \rfloor < X \right] dX_T dX, \tag{63}$$

where $f_{X,X_T}(\cdot,\cdot)$ is the probability density function of the joint probability distribution for $\{X, X_T\}$, and $[\cdot]$ denotes the $Iverson\ bracket$ as defined in Eq. (26).

Then we get the probability of $X < X_T$ *and* $\text{quantize}_Q(X) > X_T$ as

$$p_{\text{error2}\ Q} = \int_{-L}^{L} \int_{X}^{\lfloor \frac{(2^Q-1)X}{2L} + 0.5 \rfloor \frac{2L}{2^Q-1}} f_{X,X_T}(X, X_T) \tag{64}$$

$$\left[\lfloor \frac{(2^Q-1)X}{2L} + 0.5 \rfloor > X \right] dX_T dX. \tag{65}$$

Method

- **Mismatched Direction**

Since $f_{X,X_T}(X, X_T) \geq 0$ is constant established, $p_{\text{error1 } Q}$ and $p_{\text{error2 } Q}$ increases as Q becomes smaller.

(2) When $X > L$, $\text{quantize}_Q(X) = L$. $X_T > X > L = \text{quantize}_Q(X)$ is constant established when $X_T > X$, and when $X_T < X$, the probability of $X_T > \text{quantize}_Q(X) = L$ is also constant based on the given distribution of X_T. Thus, the optimization direction is always right in this case.

(3) When $X < -L$, $\text{quantize}_Q(X) = -L$. $X_T < X < -L = \text{quantize}_Q(X)$ is constant established when $X_T < X$, and when $X_T > X$, the probability of $X_T > \text{quantize}_Q(X) = -L$ is also constant based on the given distribution of X_T. Thus, the optimization direction is always right in this case.

Method

• Mismatched Direction

Table 4: Simulation of error proportion under the Q-bit

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%

- It is difficult to directly give an analytical representation of the error proportion of Gaussian distribution input under Q-bit quantization

- Monte Carlo algorithm
  - Simulate the probability of directional error caused by Q-bit by the error proportion of the pre-quantized data $\mathbf{X} \in \mathbb{R}^{10000}$
  - Each element in $\mathbf{X}$ is sampled from the standard normal distribution

- Experimental Result
  - Probability of direction mismatch increases rapidly in 1-bit quantization

Experiments

- **Comparison with SOTA Methods**
  - Baseline: BERT_BASE, TinyBERT6L, TinyBERT4L / Dataset: GLUE benchmark
  - 50% : Maximize the information entropy by the 50% quantile thresholdBERT

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**

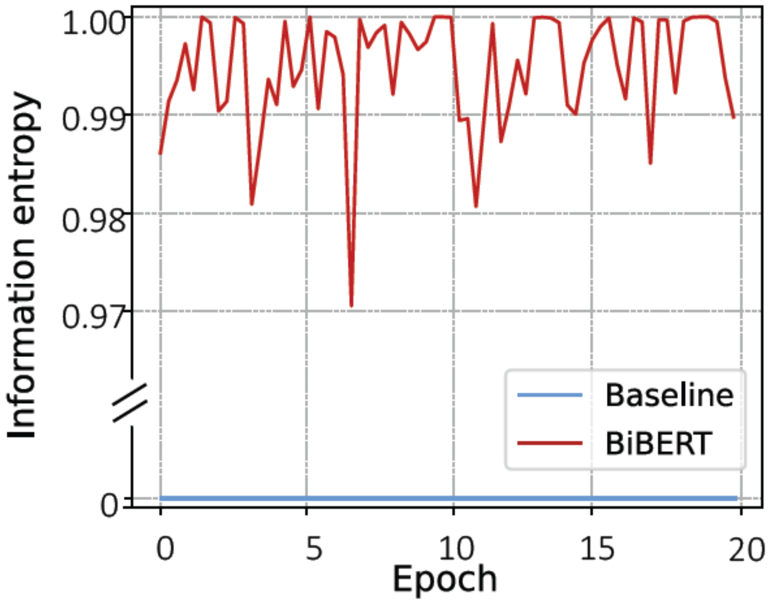
Experiments

- **Ablation Study**
  - Bi-Attention and DMD can improve the performance when used alone
  - Improve BiBERT and close the performance gap between fully binarized BERT and full-precision counterpart

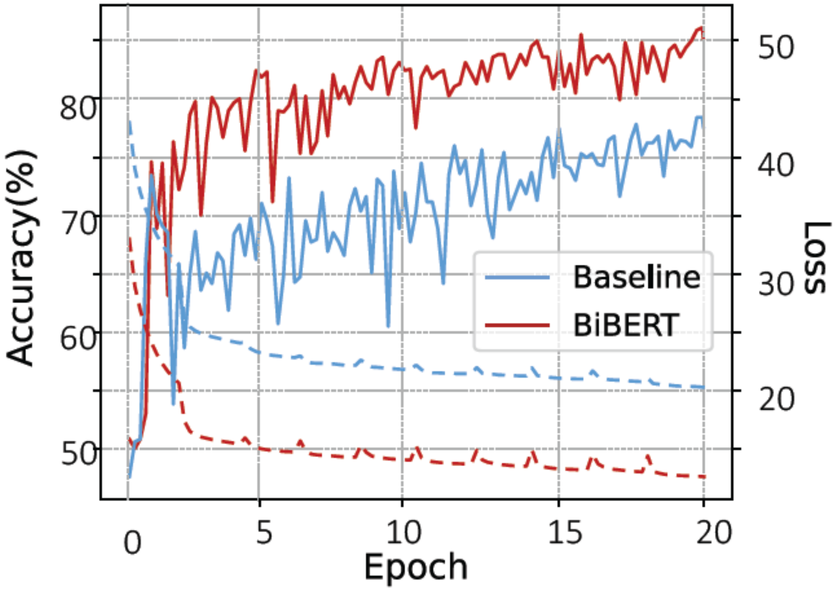
Quant	#Bits	DA	SST-2	MRPC	RTE	QQP
Full Precision	32-32-32	–	93.2	86.3	72.2	91.4
Baseline	1-1-1	✗	77.6	70.2	54.1	73.2
Bi-Attention	1-1-1	✗	82.1	70.5	55.6	74.9
DMD	1-1-1	✗	79.9	70.5	55.2	75.3
BiBERT (ours)	1-1-1	✗	**88.7**	**72.5**	**57.4**	**84.8**
Baseline	1-1-1	✓	84.0	71.4	50.9	-
Bi-Attention	1-1-1	✓	85.6	73.2	53.1	-
DMD	1-1-1	✓	85.3	72.5	56.3	-
BiBERT (ours)	1-1-1	✓	**90.9**	**78.8**	**61.0**	-

Experiments

- **Analysis**
  - Information Performance
    - Take the first heads in layer 0 of each model
    - Information entropy of attention weight in BiBERT fluctuates in a small range and is almost maximized
  - Loss function
    - Achieve faster convergence rate and higher accuracy



(a) Information performance



(b) Training curves

Conclusion

- **BiBERT (Fully binarized BERT baseline)**
  - Propose Bi-Attention and DMD in BiBERT to improve performance
  - Outperforms existing BERT quantization methods, giving an impressive 56.3X FLOPs and 31.2X model size saving