BVT-IMA: Binary Vision Transformer with Information-Modified Attention

Zhenyu Wang[1*], **Hao Luo**[4, 5†], **Xuemei Xie**[2, 3†], **Fan Wang**[4], **Guangming Shi**[2]

[1]Hangzhou Institute of Technology, Xidian University, Hangzhou 311200, China
[2]Guangzhou Institute of Technology, Xidian University, Guangzhou 510700, China
[3]Pazhou Lab, Huangpu, 510555, China
[4]DAMO Academy, Alibaba group, 310023, Hangzhou, China
[5]Hupan Lab, 310023, Hangzhou, China
zy_wang1995@outlook.com, {michuan.lh, fan.w}@alibaba-inc.com, xmxie@mail.xidian.edu.cn, gmshi@xidian.edu.cn

Abstract

As a compression method that can significantly reduce the cost of calculations and memories, model binarization has been extensively studied in convolutional neural networks. However, the recently popular vision transformer models pose new challenges to such a technique, in which the binarized models suffer from serious performance drops. In this paper, an attention shifting is observed in the binary multi-head self-attention module, which can influence the information fusion between tokens and thus hurts the model performance. From the perspective of information theory, we find a correlation between attention scores and the information quantity, further indicating that a reason for such a phenomenon may be the loss of the information quantity induced by constant moduli of binarized tokens. Finally, we reveal the information quantity hidden in the attention maps of binary vision transformers and propose a simple approach to modify the attention values with look-up information tables so that improve the model performance. Extensive experiments on CIFAR-100/TinyImageNet/ImageNet-1k demonstrate the effectiveness of the proposed information-modified attention on binary vision transformers.

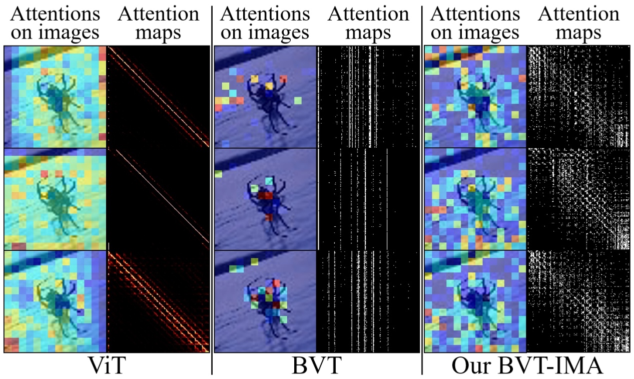


Figure 1: Attention shifting in binary ViT. ‘BVT’ is short for ‘binary ViT’. The attention maps and their corresponding average attention on the input image of different heads in the same block are visualized. All results are sampled from DeiT-Tiny on TinyImageNet. Compared to that in ViT attention maps, the highlight attention of BVT seriously shifts.

Introduction

Benefiting from the powerful long-range modeling capability of multi-head self-attention (MSA) module, vision transformer (ViT) (Dosovitskiy et al. 2020) and its variants (Touvron et al. 2021; Liu et al. 2021a) have achieved promising performance against the convolution neural networks in a variety of computer vision tasks (Amini, Periyasamy, and Behnke 2021; Ding et al. 2022; He et al. 2021). Although ViTs provide an architecture with better feature representation, the high computational cost and massive parameters restrict their application in resource-limited devices (Chen et al. 2021; Chuanyang et al. 2022). Thus, the ViT compression technology attracts wide attention, emerging various compression methods such as structured pruning (Yang et al. 2021; Yu et al. 2021; Yin et al. 2023), token reduction (Ryoo et al. 2021; Liang et al. 2021; Bolya et al. 2022), weights sharing (Lan et al. 2019; Zhang et al. 2022a), and quantization (Liu et al. 2021c; Yuan et al. 2022; Li et al. 2022).

Model binarization is a special compression method that quantizes the activations and weights from 32-bits to 1-bit, resulting in an almost 32× reduction in memory consumption as well as significant speed-up induced by replacing the float-point multiplications with bit-wise operations. Unfortunately, the performance drop caused by its poor representation capability and optimization difficulties restrict the development of such a method. To tackle the bottlenecks of model binarization, several studies have been conducted in convolutional neural networks (CNNs). A variety of optimization schemes are proposed from the perspectives of the representation ability (Liu et al. 2020b,a; Zhang et al. 2022b), the quantization error (Rastegari et al. 2016; Xu et al. 2023), and the gradient approximation (Bai, Wang, and Liberty 2019; Qin et al. 2020), which have significantly narrowed the performance gap between binary CNNs and real-value ones. For transformers, researchers have also made progress on BERT (Kenton and Toutanova 2019) for natural language processing (NLP) tasks by migrating some

[*]Work done during an internship at DAMO Academy.
[†]Corresponding author (Xuemei Xie is the primary one)

CNN binarization approaches (Liu et al. 2022) and correcting the attention value range mismatch (Qin et al. 2022). Although these methods perform well on NLP tasks, recent studies (He et al. 2023a; Gao et al. 2023) find it is still challenging to binarize ViTs due to the more complex features in vision tasks and indicate that the improper binarization in has a large impact on the model performance, which means correcting MSA is the key to improve the accuracy of binary ViTs.

In this paper, we observe a phenomenon named attention shifting that occurs in binarized ViT, which may be a reason for the binary-attention-induced accuracy degradation. Compared to the attention maps in the real-value model, the positions of highlight attention values shift in the binary ViT as shown in Figure 1. Such a deviation changes the regions of interest to ViT and thus prevents the model from extracting discriminative features from key regions in images, making the binary ViTs hard to be optimized. Further analyzing and comparing the calculation processes of self-attentions in two models (real-value ViT and binary ViT), we find the difference between the moduli of their tokens may be a reason for the attention shifting. Given a query Q and a key K, the attention value $A = QK^T = \|Q\|_2 \|K\|_2 \cdot cos\langle Q, K\rangle$ is determined by two factors, the normalized similarity ($cos\langle Q, K\rangle$) between tokens along with the moduli of tokens ($\|Q\|_2 \|K\|_2$) which are positively correlated to the information quantity (detailed analyses in subsequent sections). For the real-value model, two factors dynamically change with the query and key. However, for binary ViTs, the constant moduli of queries and keys caused by binarization can not represent the changes of information quantities in different tokens, leaving only one dynamic factor (the normalized similarity) in the attention value and thus losing much information.

Based on these analyses, we explore the representation of the information quantity hidden in the attention of binary ViTs. Specifically, the process of computing an attention score in binary models is considered as several Bernoulli trials, for which its probability-dense function (PDF) can be achieved. According to the PDF of attention scores and the limited attention values in binary MSA, the missing information quantity can then be represented by a group of learnable PDF-related modification factors that are optimized together with the model weights. The modification factors of each attention head form an information table, with which the attention maps can be modified by simple looking-up operations. With the help of such an information modification, the attention shifting will be relieved as shown in Figure 1, improving the feature fusion capability of binary ViTs. The code will be uploaded to https://github.com/Daner-Wang/BVT-IMA.git.

Related Work

CNN binarization. Due to the limited representation capacity and the non-differentiable quantizer (*e.g.*, sign function), model binarization will cause serious performance degradation. Several methods have been proposed to increase the accuracy of binary CNNs from perspectives of binarization-friendly architecture (Bulat, Martinez, and Tzimiropoulos 2021; Bethge et al. 2021; Zhang, Zhang, and Lew 2022), optimizing the gradient estimator (Courbariaux and Bengio 2016; Ding, Liu, and Zhou 2022; He et al. 2023b), knowledge distillation (Mishra and Marr 2018; Martinez et al. 2020), etc. XnorNet (Rastegari et al. 2016) proposes to replace the matrix multiplications with bit-wise operations and increase the representation capability by employing float-point scaling factors for binarized weights. ReActNet (Liu et al. 2020b) introduces generalized Sign and PReLU functions for distribution reshaping and shifting. Real-to-binary (Liu et al. 2021b) explores the influence of different training schemes on binary CNNs and reveals that the Adam optimizer can overcome the local optimization induced by zero gradients. IR-Net (Qin et al. 2020) narrows the gradient gap between the quantizer and gradient estimator by smoothing the backward process. ReBNN (Xu et al. 2023) reduces the binarization error through dynamic constraints. Bi-Real Net (Liu et al. 2020a) and PokeBNN (Zhang, Zhang, and Lew 2022) enhance the information capacity of binary models with extra residual architectures. Unfortunately, these approaches have not generalized well to transformer models (Liu et al. 2022), leading to the exploration of binary transformers.

Transformer binarization. In the field of NLP, researchers have proposed some binarized transformers and closed the gap with real-value models to a few percentage points. BinaryBERT (Bai et al. 2021) preliminarily attempts the binarization in words embedding and weights for BERT (Kenton and Toutanova 2019). BiBERT (Qin et al. 2022) firstly proves the practicability of fully binarized transformers and indicates that the attention map should be binarized to $\{0, 1\}$. BiT (Liu et al. 2022) further shows the performance of binarized BERT by the elastic binarization function and multi-distillation. Recently, studies for ViT binarization have also been started, in which the more complex vision features make it a new challenge. BiViT (He et al. 2023a) takes the long-tailed distribution of softmax attention into account and proposes a softmax-aware binarization to reduce the quantization error in attention maps. GSB (Gao et al. 2023) (Group Superposition Binarization) also indicates that the poor feature representation in the binarized self-attention module has a large impact on the performance and introduces a group superposition binarization scheme to increase its feature diversity. Compared to these methods, this paper further analyzes the possible reason for the accuracy decline induced by binary self-attention from the perspective of information quantity and improves the performance with a simple learnable information table, which can keep same float operations during inference.

ViT Binarization

In general, a ViT model is stacked by several transformer blocks including multi-head self-attention (MSA) modules, LayerNorm layers, and the fully connected (FC) layers of Feed-Forward Networks (FFN). Following suggestions of previous binarization studies (Qin et al. 2022; Liu et al. 2022), a binary ViT baseline is built as that shown in Figure 2, in which the FC layers as well as the attention maps in

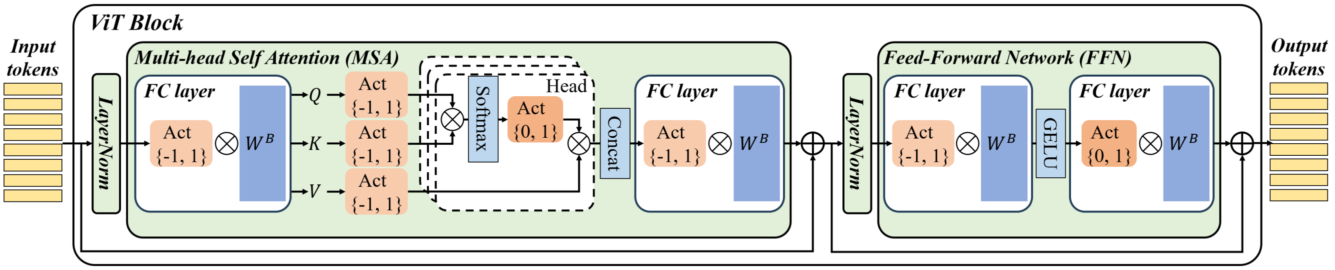


Figure 2: Binarized Vision Transformer block. The weights W^B of fully connected (FC) layers are binarized to $\{-1, 1\}$. The orange circular rectangles denote the binarization for activations. The inputs of MSA/FFN, and the query/key/value in MSA are quantized to $\{-1, 1\}$. The normalized attention maps in MSA and the nonlinear outputs between two FC layers of FFN are binarized to $\{0, 1\}$. The normalization layer with few parameters and the non-parametric residual connections are retained.

all blocks are binarized. The activations other than the attention maps and the outputs of activation functions are quantized to $\{-1, 1\}$ by the sign function, for which the Straight-Through Estimator (STE) (Bengio, Léonard, and Courville 2013) is used to optimize its backward process. The channel-wise learnable bias $\beta^x \in \mathbb{R}^D$ and the layer-wise learnable scale factor $\alpha^x \in \mathbb{R}$ are applied for adjusting the distribution and reducing the quantization error as BiT (Liu et al. 2022) and then the binarization function is formulated as:

$$X^B = \alpha^x \cdot S(X^R - \beta^x), \tag{1}$$

where $S(\cdot)$ means sign-function, $X^R \in \mathbb{R}^{N \times D}$ is the real-value activations, $X^B \in \{-\alpha^x, \alpha^x\}^{N \times D}$ denotes the quantized features, N is the number of tokens, and D is the token dimension. The real-value weights $W^R \in \mathbb{R}^{D \times D'}$ are centralized to zero-mean before binarization for larger information entropy and then quantized as:

$$W^B = \alpha^w \cdot S(W^R - \bar{W}^R), \quad \alpha^w = \frac{\|W^R\|_1}{D \times D'}, \tag{2}$$

where $\bar{W}^R$ is the mean value of weights, D and D' denote the input channels and output channels, respectively. With the quantized weights and inputs, the outputs $X' \in \mathbb{R}^{N \times D'}$ of FC layers can be reformulated as:

$$X' = X^B W^B = \alpha^x \alpha^w \cdot S(X^R - \beta^x) S(W^R - \bar{W}^R), \tag{3}$$

which can be accelerated by the efficient XNOR and Bitcount operators (Rastegari et al. 2016).

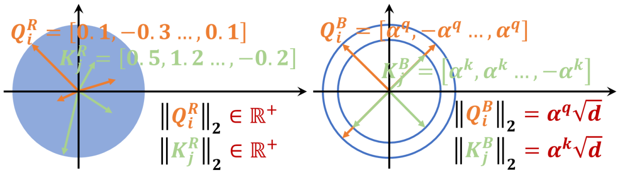
Particularly, because the real-values of the activation function (*e.g.*, GELU) outputs and the attention map after softmax normalization are round zero in the negative semi-axis, these activations are suggested (Liu et al. 2022) to binarize to $\{0, 1\}$:

$$X^B = \alpha^x \cdot \left\lfloor Clip(\frac{X^R - \beta^x}{\alpha^x}, 0, 1) \right\rceil, \tag{4}$$

where $\lfloor \cdot \rceil$ is the round function. The gradient of the scaling factor in such a discontinuous differentiable function is estimated as that in LSQ (Esser et al. 2020).

Information Loss in Binary Attention

As the key module used to fuse information among tokens, the multi-head self-attention has a large impact on ViTs. A special information loss that happens only in the attention map of binary ViTs causes the attention shifting and thus significantly influences the model performance.


$Q_i^R = [0.1, -0.3 \ldots, 0.1]$
$K_j^R = [0.5, 1.2 \ldots, -0.2]$
$\|Q_i^R\|_2 \in \mathbb{R}^+$
$\|K_j^R\|_2 \in \mathbb{R}^+$
$Q_i^B = [\alpha^q, -\alpha^q \ldots, \alpha^q]$
$K_j^B = [\alpha^k, \alpha^k \ldots, -\alpha^k]$
$\|Q_i^B\|_2 = \alpha^q\sqrt{d}$
$\|K_j^B\|_2 = \alpha^k\sqrt{d}$

(a) Real-value query and key (b) Binary query and key

Figure 3: Query and key in different spaces. (a) shows the projections of the real-value query and key in the two-dimensional surface, in which vectors are located on a plane with diverse moduli. (b) demonstrates the projections of the binarized query and key, in which the moduli for queries and keys are constants.

Real-value attention. In real-value ViTs, the attention map of a head can be formulated as:

$$Softmax\left(\frac{A^R}{\sqrt{d}}\right) = Softmax\left(\frac{(Q^R)(K^R)^T}{\sqrt{d}}\right), \tag{5}$$

where Q^R and $K^R \in \mathbb{R}^{N \times d}$ denote the real-value query and key in the head. The attention score $A^R_{i,j}$ in the i-th row, and the j-th column before softmax normalization can be rewritten as:

$$A^R_{i,j} = \|Q^R_i\|_2 \|K^R_j\|_2 \cdot \cos\langle Q^R_i, K^R_j \rangle. \tag{6}$$

This demonstrates the attention value in real-value MSA is determined by the cosine distance as well as the moduli of tokens, both of which change with the query and key.

Binary attention. The binarization will be adopted to the self-attention twice as that shown in Figure 2. Queries and keys are binarized before the matrix multiplication. The achieved attention maps will be binarized after softmax normalization. Because the normalization and the binarization operations will not change the relative magnitudes of attention values, the positions of highlight attentions are determined by the products (attention scores) of binary queries and keys. As demonstrated in Figure 3(b), the elements in queries and keys are binarized to $\left\{-\alpha^q, \alpha^q\right\}$ and $\left\{-\alpha^k, \alpha^k\right\}$ resulting in a fixed modulus for any queries and keys in a block:

$$\left\|Q_i^B\right\|_2 = \alpha^q \sqrt{d}, \left\|K_j^B\right\|_2 = \alpha^k \sqrt{d}, \tag{7}$$

where d is the dimensions of the vector. The attention score before softmax normalization is thus formulated as:

$$\begin{aligned} A_{i,j}^B &= \left\|Q_i^B\right\|_2 \left\|K_j^B\right\|_2 \cdot \cos\left\langle Q_i^B, K_j^B\right\rangle \\ &= \alpha^q \alpha^k d \cdot \cos\left\langle Q_i^B, K_j^B\right\rangle, \end{aligned} \tag{8}$$

where α^q and α^k are constant scaling factors for tokens in the query and key matrices, respectively. Compared to the real-value model with tokens of dynamic moduli in a head (*e.g.*, $\left\|Q_i^R\right\|_2 \in \mathbb{R}^+$), the modulus of any token is a constant in binarized ViTs (*e.g.*, $\left\|Q_i^B\right\|_2 = \alpha^q \sqrt{d}$), leading to the attention scores rely only on the changes of cosine distances between tokens.

Loss of information quantity. The constant moduli in binary MSA are considered to have lost the information quantity hidden in the attention score, leading to the attention shifting to some extent. To explain this, we review the connotation of the moduli in real-value attention from the perspective of information theory. The activations in ViTs with maximum information entropy after full convergence approximately follows zero-mean Gaussian distribution (Qin et al. 2022). Thus, the total information quantity I_i^q of the token Q_i^R can be estimated by:

$$\begin{aligned} I_i^q &= \textstyle\sum_{j=0}^{d-1} - \log \frac{1}{\sqrt{2\pi}\sigma} exp\left(-\frac{\left(Q_{i,j}^R\right)^2}{2\sigma^2}\right) \\ &= \frac{1}{2\sigma^2} \textstyle\sum_{j=0}^{d-1} \left(Q_{i,j}^R\right)^2 + d \log \sqrt{2\pi}\sigma \\ &\Rightarrow \textstyle\sum_{j=0}^{d-1} \left(Q_{i,j}^R\right)^2 \end{aligned} . \tag{9}$$

where $\Rightarrow$ means positive correlation. Meanwhile, The modulus of the token is:

$$\left\|Q_i^R\right\|_2 = \sqrt{\textstyle\sum_{j=0}^{d-1} \left(Q_{i,j}^R\right)^2} \Rightarrow \sum_{j=0}^{d-1} \left(Q_{i,j}^R\right)^2. \tag{10}$$

Obviously, $\left\|Q_i^R\right\|_2 \Rightarrow I_i^q$, which means the moduli of real-value tokens represent the information quantity contained in them. $\left\|Q_i^R\right\|_2 \left\|K_j^R\right\|_2 \Rightarrow I_i^q I_j^k$ can then be regarded as the information quantity of the real-value attention score. Therefore, the attention scores in real-value MSA measure both the correlation between different tokens and the information quantity contained in them. In binary MSA, the attention values with constant modulus lose the attribute of the information quantity. Binary ViTs are thus insensitive to such information and the relative magnitudes among attentions are changed. These biased attention maps influence the information fusion between tokens and finally restrict the performance of binary ViTs.

Information Modified Binary Attention

To relieve the attention shifting caused by the loss of information quantity in binary MSA, this paper proposes to modify the attention map from the perspective of information theory and extracts the information quantity hidden in binary attention by a simple scheme.

Each element in the attention maps of binary MSA can be calculated by the XNOR operator as follow:

$$\begin{aligned} A_{i,j}^B &= \textstyle\sum_{t=0}^{d-1} Q_{i,t}^B K_{j,t}^B \\ &= \alpha^q \alpha^k d \textstyle\sum_{t=0}^{d-1} S(Q_{i,t}^B) \odot S(K_{j,t}^B), \end{aligned} \tag{11}$$

where $\odot$ denotes the XNOR operator. The XNOR operation between the two elements with only two results ($S(Q_{i,t}^B) \odot S(K_{j,t}^B) \in \{-1, 1\}$) can be considered as the result of one Bernoulli trial (Qin et al. 2022) with the positive probability p, *i.e.*, the probability of $S(Q_{i,t}^B) \odot S(K_{j,t}^B) = 1$. Therefore, the attention score $A_{i,j}^B$ obeys Bernoulli distribution $\mathbb{B}\left(d, p\right)$. Assuming the Hamming distance between the query token and the key token is $n_{i,j}$ which denotes the number of times the result of the XNOR operation is 1 during the calculation of an attention score, the attention score can then be rewritten as $A_{i,j}^B = \alpha^q \alpha^k (2n_{i,j} - d)$. Hence, the probability density function (PDF) of $A_{i,j}^B$ is:

$$f\left(A_{i,j}^B\right) = C_d^{n_{i,j}} (p)^{n_{i,j}} (1-p)^{d-n_{i,j}}. \tag{12}$$

Since the PDF of attention score in binary MSA has been obtained, the corresponding information quantity $I_{i,j}^a$ is formulated as:

$$\begin{aligned} I_{i,j}^a &= -\log f(A_{i,j}^B) = \log \frac{1}{f(A_{i,j}^B)} \\ &\Rightarrow \left(f(A_{i,j}^B)\right)^{-1} \\ &\Rightarrow \left(C_d^{n_{i,j}} (p)^{n_{i,j}} (1-p)^{d-n_{i,j}}\right)^{-1}. \end{aligned} \tag{13}$$

It is obvious that the information quantity of the attention score in binary MSA is related to the Hamming distance $n_{i,j}$ between tokens and the positive probability p. Although the Hamming distance $n_{i,j}$ can be easily achieved from the attention score $A_{i,j}^B$ by $n_{i,j} = \frac{1}{2}(\frac{A_{i,j}^B}{\alpha^q \alpha^k} + d)$, the probability p is different for diverse scores and can not be directly inferred. Fortunately, due to the definite token dimensions and results of XNOR operations, the attention score here involves only $d+1$ values, leading to limited probabilities required to be estimated. Taking advantage of such a property, the information quantity of attention scores in a binary MSA head can be approximated with $d+1$ learnable modification factors $\left\{\gamma_{n_{i,j}} | n_{i,j} = 0, 1, 2, \ldots, d\right\}$:

$$\gamma_{n_{i,j}} = \left(C_d^{n_{i,j}} (p_{n_{i,j}})^{n_{i,j}} (1 - p_{n_{i,j}})^{d-n_{i,j}}\right)^{-m}, \tag{14}$$

where $p_{n_{i,j}}$ is the unknown probabilities, $m \in \mathbb{R}^+$ is a hyper-parameter used during the initialization. Specifically, due to lack of sufficient prior information, the uncertain term $(p_{n_{i,j}})^{n_{i,j}} (1 - p_{n_{i,j}})^{d-n_{i,j}}$ will be set to a constant value 1 during the initialization for simplicity. This will lead to a large range of initial modification factors, which are hard to be optimized. Thus, m is set to $0.5^{\lfloor \log_2(\log_{10}(max(C_d^{n_{i,j}}))) \rfloor}$, scaling the initial modification factors to a range $(0, 1]$ easy to be optimized. As shown in Figure 4, all modification factors form a table used to correct the attention map. For each attention score $A_{i,j}^B$, the corresponding Hamming distance $n_{i,j}$ is also the index used to identify the modification factor $\gamma_{n_{i,j}}$. The attention score is then modified as:

$$\hat{A}_{i,j}^B = A_{i,j}^B \cdot \left|\gamma_{n_{i,j}}\right|. \tag{15}$$

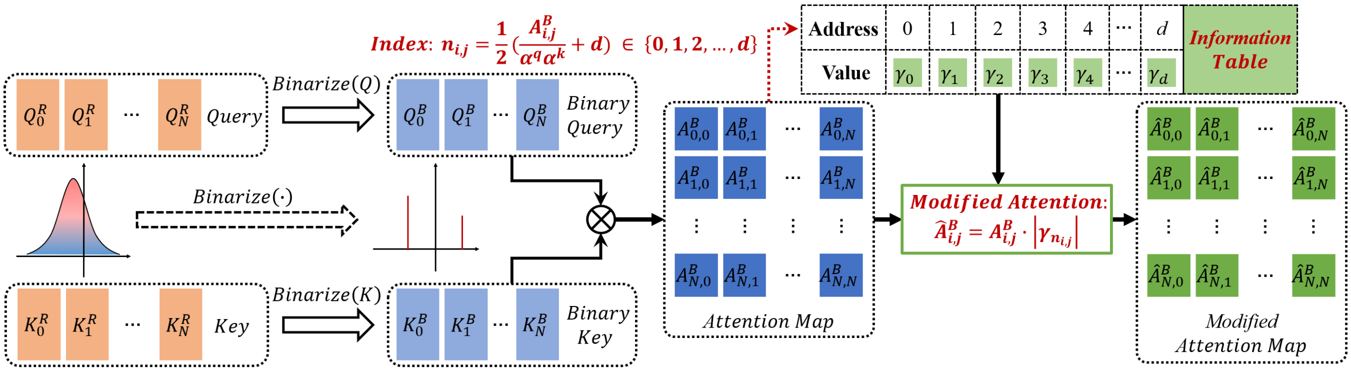

Index: $n_{i,j} = \frac{1}{2}(\frac{A^B_{i,j}}{\alpha^q\alpha^k} + d) \in \{0,1,2,\dots,d\}$
Address
0
1
2
3
4
···
d
Value
γ_0
γ_1
γ_2
γ_3
γ_4
···
γ_d
Information Table
Q^R_0 Q^R_1 ··· Q^R_N Query
Binarize(Q)
Q^B_0 Q^B_1 ··· Q^B_N Binary Query
Binarize(·)
K^R_0 K^R_1 ··· K^R_N Key
Binarize(K)
K^B_0 K^B_1 ··· K^B_N Binary Key
$A^B_{0,0}$ $A^B_{0,1}$ ··· $A^B_{0,N}$
$A^B_{1,0}$ $A^B_{1,1}$ ··· $A^B_{1,N}$
$A^B_{N,0}$ $A^B_{N,1}$ ··· $A^B_{N,N}$
Attention Map
Modified Attention: $\hat{A}^B_{i,j} = A^B_{i,j} \cdot |\gamma_{n_{i,j}}|$
$\hat{A}^B_{0,0}$ $\hat{A}^B_{0,1}$ ··· $\hat{A}^B_{0,N}$
$\hat{A}^B_{1,0}$ $\hat{A}^B_{1,1}$ ··· $\hat{A}^B_{1,N}$
$\hat{A}^B_{N,0}$ $\hat{A}^B_{N,1}$ ··· $\hat{A}^B_{N,N}$
Modified Attention Map

Figure 4: Attention modification for binary MSA. The attention score $A^B_{i,j}$ in a MSA head is revised with $d+1$ learnable modification factors $\{\gamma_0, \gamma_1, \dots, \gamma_d\}$, which denotes the information quantity contained in binary attentions.

Particularly, the revised attention can be further formulated as $\hat{A}^B_{i,j} = \alpha^q\alpha^k(2n_{i,j} - d) \cdot |\gamma_{n_{i,j}}|$, in which $n_{i,j}$ and $\gamma_{n_{i,j}}$ contain $d+1$ one-to-one possible values. Hence, the information table can be updated as:

$$\hat{\gamma}_{n_{i,j}} = |\gamma_{n_{i,j}}| \cdot \alpha^q\alpha^k(2n_{i,j} - d), n_{i,j} = 0, 1, 2, \dots, d. \quad (16)$$

The scaling factors are fused into the table, simplifying the uncorrected attention score $A^B_{i,j}$ and the index $n_{i,j}$ to $A^B_{i,j} = S(Q^R_i - \beta^q)S(K^R_j - \beta^k)^T$ and $n_{i,j} = (A^B_{i,j} + d) >> 1$, respectively. The procedure of the self-attention modification can then be replaced by filling an $N \times N$ map with values $\hat{\gamma}_{n_{i,j}}$ identified by the index $n_{i,j}$.

Experiments

Implementation Details

The proposed method is applied to popular ViT models (DeiT (Touvron et al. 2021), Swin (Liu et al. 2021a), and NesT (Zhang et al. 2022c)) and evaluated on the CIFAR-100 (Krizhevsky and Hinton 2009)/TinyImageNet (Pouransari and Ghili 2014)/ImageNet-1k (Russakovsky et al. 2015) benchmark of 100/200/1000 classes. All experiments are implemented with the Pytorch (Paszke et al. 2019) and TIMM library on NVIDIA-V100 GPUs. The embedding layer and classification head are quantized to 8 bits as GSB (Gao et al. 2023) while the MSA, and FFN modules are binarized. Two-stage training is adopted, in which only weights are binarized in the first stage. Activation binarization and information tables are adopted in the second stage. The Adam optimizer without weight decay is employed. The cosine annealing schedule with 5 epochs of warm-up is applied to adjust the learning rate initialized to $5e-4$. For ImageNet-1k, a two steps binarization scheme [1] is adopted during warm-up to help models to converge on the complex dataset. The knowledge distillation is adopted for each quantized model to learn from its corresponding real-value teacher with the cross entropy loss function and 0.5 distillation factor. Models are trained for 300/150/150 epochs in each stage on CIFAR-100/TinyImageNet/ImageNet-1k. Limited by the GPU memory, the batch size is 128 for DeiT-Tiny in both stages while 128/64 for the other models in the first/second stage. The data augmentation and other hyper-parameters are the same as those in DeiT.

Performance on Full-Attention Transformer

The performances of the proposed method on binary DeiT models are shown in Table 1 and compared with the latest reported ViT binarization studies. We reproduce the architecture of BiT as the binary baseline. On all benchmarks, the proposed BVT-IMA achieves the state-of-the-art (SOTA) performance. Compared to GSB (Gao et al. 2023) that introduces several scaling factors to MSA, BVT-IMA achieves better performance (75.77% vs 71.10%) with less operations (0.15G vs 0.32G). Compared with BiT (Liu et al. 2022), the information modified attention significantly improves the performance with a few more FLOPs. On ImageNet-1k, the proposed method outperforms BiBERT (Qin et al. 2022) and BiViT (He et al. 2023a) in the same quantization setting. The comparison results show that the proposed information modification is a novel and effective way to improve the feature extraction capability of binary ViTs rather than just stacking scaling factors.

Performance on Local-Attention Transformer

In addition to DeiT models, BVT-IMA is also evaluated on popular **l**ocal-**a**ttention **t**ransformers (LATs) (*e.g.* Swin (Liu et al. 2021a), and NesT (Zhang et al. 2022c)) and compared with recently most related studies, of which the results are shown in Table 2, and Table 3. The results of approaches migrated from NLP (*e.g.*, BiBERT, and BiT) are reproduced by BiViT. It can be found that naively migrating the methods that succeed in NLP transformers to the vision field seriously may hurt the model performance (Table 2: only 32.39% $\sim$ 41.89% on TinyImageNet) and even lead to fail convergence (Table 3: BiBERT). Compared to previous methods, our BVT-IMA still obtains the SOTA performances on LATs, which are even better than those of the method (*e.g.*, BiViT) with channel-wise scaling factors.

[1] activations are quantized as: $32bits \rightharpoonup 2bits \rightharpoonup 1bit$

Dataset	Model	Method	W-A	Top1	BOPs (G)	FLOPs (G)	OPs (G)
CIFAR-100	DeiT-Tiny	Real-value	32-32	86.81%	0	1.26	1.26
		BiT (Liu et al. 2022)	1-1	43.59%	1.23	0.036	0.06
		BVT-IMA (Ours)	1-1	**62.46**%	1.23	0.037	0.06
	DeiT-Small	Real-value	32-32	88.80%	0	4.63	4.63
		Q-ViT (Li et al. 2022)	1-1	50.26%	–	–	–
		GSB (Gao et al. 2023)	1-1	71.10%	4.57	0.25	0.32
		BiT (Liu et al. 2022)	1-1	66.41%	4.57	0.072	0.14
		BVT-IMA (Ours)	1-1	**75.77**%	4.57	0.074	0.15
TinyImageNet	DeiT-Tiny	Real-value	32-32	75.11%	0	1.26	1.26
		BiT (Liu et al. 2022)	1-1	24.32%	1.23	0.036	0.06
		BVT-IMA (Ours)	1-1	**39.67**%	1.23	0.037	0.06
	DeiT-Small	Real-value	32-32	78.56%	0	4.63	4.63
		BiT (Liu et al. 2022)	1-1	38.77%	4.57	0.072	0.14
		BVT-IMA (Ours)	1-1	**43.42**%	4.57	0.075	0.15
ImageNet-1k	DeiT-Tiny	Real-value	32-32	74.48%	0	1.26	1.26
		BiT (Liu et al. 2022)	1-1	21.68%	1.23	0.036	0.06
		BVT-IMA (Ours)	1-1	**30.03**%	1.23	0.038	0.06
		BiBERT (Qin et al. 2022)	1-1/32	25.40%	–	–	0.39
		BiViT (He et al. 2023a)	1-1/32	37.90%	–	–	0.39
		BVT-IMA (Ours)	1-1/32	**43.99**%	22.95	0.038	0.40
	DeiT-Small	Real-value	32-32	81.16%	0	4.63	4.63
		BiT (Liu et al. 2022)	1-1	30.73%	4.57	0.072	0.14
		BVT-IMA (Ours)	1-1	**47.98**%	4.57	0.075	0.15
	DeiT-Base	Real-value	32-32	83.38%	0	17.66	17.66
		BiT (Liu et al. 2022)	1-1	38.19%	17.54	0.145	0.42
		BVT-IMA (Ours)	1-1	**62.65**%	17.54	0.150	0.42
		BiBERT (Qin et al. 2022)	1-1/32	67.50%	–	–	5.81
		BiViT (He et al. 2023a)	1-1/32	69.60%	–	–	5.81
		BVT-IMA (Ours)	1-1/32	**74.06**%	365.09	0.150	5.85

Table 1: Performance of binary DeiT on different datasets. 'BOPs' and 'FLOPs' denote the number of bit-wise operations and float-point operations during inference, respectively. 'OPs' is a sum of BOPs and FLOPs, *i.e.*, '$OPs = \frac{BOPs}{64} + FLOPs$' (Liu et al. 2020b). 'W-A' denotes the bit width of weights and activations. '1-1/32' indicates that all weights are binarized while activations in MLP modules are maintained at full precision as BiViT (He et al. 2023a).

Model	Method	W-A	Top1
Swin-Tiny	Real-value	32-32	78.86%
	BiBERT (Qin et al. 2022)	1-1	41.89%
	BiT (Liu et al. 2022)	1-1	40.52%
	BiViT (He et al. 2023a)	1-1	58.66%
	BVT-IMA (Ours)	1-1	**60.85**%
NesT-Tiny	Real-value	32-32	79.94%
	BiBERT (Qin et al. 2022)	1-1	32.39%
	BiT (Liu et al. 2022)	1-1	34.72%
	BiViT (He et al. 2023a)	1-1	52.21%
	BVT-IMA (Ours)	1-1	**64.23**%

Table 2: Performance of binary LATs on TinyImageNet.

Model	Method	W-A	Top1
Swin-Tiny	Real-value	32-32	81.20%
	BiBERT (Qin et al. 2022)	1-1/32	68.30%
	BiViT (He et al. 2023a)	1-1/32	70.80%
	BVT-IMA (Ours)	1-1/32	**72.03**%
NesT-Tiny	Real-value	32-32	81.10%
	BiBERT (Qin et al. 2022)	1-1/32	0.27*%
	BiViT (He et al. 2023a)	1-1/32	68.70%
	BVT-IMA (Ours)	1-1/32	**71.00**%

Table 3: Performance of binary LATs on ImageNet-1k. '*' denotes the results of fail convergence reported by BiViT (He et al. 2023a).

Ablation Study

Influence of different information table settings. The initialization and the size of the information table will affect the attention modification. Thus, we further analyze the influence of different table settings on the model performance. All experiments are conducted on DeiT-Tiny and evaluated on CIFAR-100 and TinyImageNet. Two initialization approaches 'One-init' and 'Bernoulli-init' are evaluated. 'One-

Dataset	w/ Table	Initialization	Top1
CIFAR-100	×	–	43.59%
	✓	One-init	46.04%
	✓	Bernoulli-init	62.46%
TinyImageNet	×	–	24.32%
	✓	One-init	26.20%
	✓	Bernoulli-init	39.67%

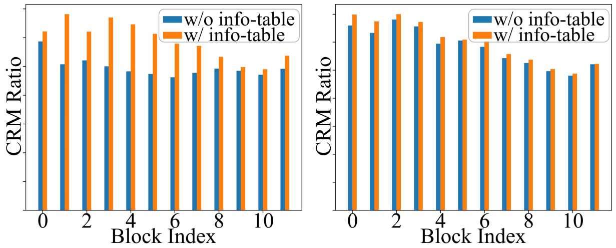
Table 4: Influence of information table initialization on the model performance. The evaluated model is DeiT-Tiny.

Dataset	w/ Table	Table mode	Top1
CIFAR-100	×	–	43.59%
	✓	single-head	62.16%
	✓	multi-head	62.46%
TinyImageNet	×	–	24.32%
	✓	single-head	39.19%
	✓	multi-head	39.67%

Table 5: Influence of multi-head information table on the model performance.

init' denotes initializing all $\gamma_{n_{i,j}}$ to 1 while 'Bernoulli-init' introduces the Bernoulli distribution prior to $\gamma_{n_{i,j}}$ based on Eq. 14. As shown in Table 4, the 'Bernoulli-init' achieves better performance on both benchmarks, which proves the rationality of introducing the Bernoulli prior. Additionally, the impact of whether the information table is shared by all heads in the same MSA module is analyzed as well, *i.e.*, the influence of single/multi-head table. Table 5 demonstrates the results of weather sharing a table for different heads. The diversity induced by the multi-head information table is also slightly beneficial for binary ViT models.

Influence of the information table on attention maps. We also employ quantitative analysis of the changes in attention maps after modification and propose a criterion to measure the quality of the attention maps, in which the maps of high consistency to those in real-value models are considered to be high-quality maps. As described in previous sections, the attention shifting changes the relative magnitude between attention scores, misleading ViTs to pay less attention to significant tokens. Hence, the quality of attention maps in binary ViTs can be measured by how many relative magnitudes are retained the same as those in real-value models. For example, given an attention score $A^R_{i,j}$ larger than another score $A^R_{i,t}$ in the real-value model. If the corresponding score $A^B_{i,j}$ is still larger than $A^B_{i,t}$ in the binary model, then the relative magnitude between such two attention values is considered to be retained. This case and its opposite case are respectively recorded as 1 and 0 by the relative magnitude consistency $c_{i,j,t}$ [2]. Based on this, we introduce a criterion named consistent relative magnitude (CRM) ratio to evaluate the quality of the attention maps after binarization as $CRM = \frac{1}{N^3}\sum_{i=0}^{N-1}\sum_{j=0}^{N-1}\sum_{t=0}^{N-1} c_{i,j,t}$. The higher the CRM ratio is, the better the quality of attention maps is. As shown in Figure 5, the CRM ratio of models with information tables is higher (orange bars), which proves that the proposed method indeed relieves the attention shifting and improves the attention map quality in binary ViTs.



(a) CRM of DeiT-Tiny (b) CRM of DeiT-Small

Figure 5: Consistent relative magnitude ratio of softmax attentions. 'CRM' denotes the consistent relative magnitude which measures how many attention values in binary models retain their magnitude relationships as those in real-value models. The results are evaluated on CIFAR-100.

More details and results are reported in the supplementary material, including Pytorch implementation of IMA, visualization of attention maps, and so forth.

Conclusion and Discussion

In this paper, we indicate the attention shifting that occurs after ViT binarization, which has a large impact on the feature fusion and the model performance. By comparing and analyzing the differences in self-attention modules between binary and real-value ViTs, we further propose that the deficiency of the information quantity in binary ViTs may be a reason for such a phenomenon. Then, a simple approach is introduced to represent the information quantity hidden in the attention score of a binarized ViT with limited learnable modification factors, which form information tables for different attention heads. With these tables, the missing information can be efficiently achieved by looking up operations. Beneficial from these information modification factors, the attention shifting as well as the performance in binary ViTs are improved. Finally, we also introduce a criterion named consistent relative magnitude (CRM) ratio to measure the quality of attention maps. Experiments on different benchmarks demonstrate that our information-modified attention (IMA) is more suitable to binary ViTs, leading to even more than 20% improvement in accuracy. As a primary attempt to optimize binary ViTs from the perspective of information theory, more future studies can be explored to design more reasonable architectures based on it.

Acknowledgments

This work is supported by the Guangzhou Key Laboratory of Scene Understanding and Intelligent Interaction (No. 202201000001), Pazhou Lab (Huangpu) (2022K0903), and Alibaba Group through Alibaba Research Intern Program.

[2] $c_{i,j,t} = max(S(A^R_{i,j} - A^R_{i,t}) \cdot S(A^B_{i,j} - A^B_{i,t}), 0)$

References

Amini, A.; Periyasamy, A. S.; and Behnke, S. 2021. T6d-direct: Transformers for multi-object 6d pose direct regression. In *DAGM German Conference on Pattern Recognition*, 530–544. Springer.

Bai, H.; Zhang, W.; Hou, L.; Shang, L.; Jin, J.; Jiang, X.; Liu, Q.; Lyu, M.; and King, I. 2021. BinaryBERT: Pushing the Limit of BERT Quantization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 4334–4348.

Bai, Y.; Wang, Y.-X.; and Liberty, E. 2019. ProxQuant: Quantized Neural Networks via Proximal Operators. In *International Conference on Learning Representations*.

Bengio, Y.; Léonard, N.; and Courville, A. C. 2013. Estimating or Propagating Gradients Through Stochastic Neurons for Conditional Computation. *CoRR*, abs/1308.3432.

Bethge, J.; Bartz, C.; Yang, H.; Chen, Y.; and Meinel, C. 2021. MeliusNet: An Improved Network Architecture for Binary Neural Networks. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 1438–1447.

Bolya, D.; Fu, C.-Y.; Dai, X.; Zhang, P.; Feichtenhofer, C.; and Hoffman, J. 2022. Token Merging: Your ViT But Faster. In *The Eleventh International Conference on Learning Representations*.

Bulat, A.; Martinez, B.; and Tzimiropoulos, G. 2021. High-Capacity Expert Binary Networks. In *International Conference on Learning Representations*.

Chen, T.; Cheng, Y.; Gan, Z.; Yuan, L.; Zhang, L.; and Wang, Z. 2021. Chasing sparsity in vision transformers: An end-to-end exploration. *Advances in Neural Information Processing Systems*, 34.

Chuanyang, Z.; Li, Z.; Zhang, K.; Yang, Z.; Tan, W.; Xiao, J.; Ren, Y.; and Pu, S. 2022. SAViT: Structure-Aware Vision Transformer Pruning via Collaborative Optimization. In Oh, A. H.; Agarwal, A.; Belgrave, D.; and Cho, K., eds., *Advances in Neural Information Processing Systems*.

Courbariaux, M.; and Bengio, Y. 2016. BinaryNet: Training Deep Neural Networks with Weights and Activations Constrained to +1 or -1. *CoRR*, abs/1602.02830.

Ding, L.; Lin, D.; Lin, S.; Zhang, J.; Cui, X.; Wang, Y.; Tang, H.; and Bruzzone, L. 2022. Looking outside the window: Wide-context transformer for the semantic segmentation of high-resolution remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*.

Ding, R.; Liu, H.; and Zhou, X. 2022. IE-Net: Information-enhanced binary neural networks for accurate classification. *Electronics*, 11(6): 937.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.

Esser, S. K.; McKinstry, J. L.; Bablani, D.; Appuswamy, R.; and Modha, D. S. 2020. Learned Step Size Quantization. In *International Conference on Learning Representations*.

Gao, T.; Xu, C.; Zhang, L.; and Kong, H. 2023. GSB: Group Superposition Binarization for Vision Transformer with Limited Training Samples. *CoRR*, abs/2305.07931.

He, S.; Luo, H.; Wang, P.; Wang, F.; Li, H.; and Jiang, W. 2021. Transreid: Transformer-based object re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15013–15022.

He, Y.; Lou, Z.; Zhang, L.; Liu, J.; Wu, W.; Zhou, H.; and Zhuang, B. 2023a. BiViT: Extremely Compressed Binary Vision Transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 5651–5663.

He, Y.; Zhang, L.; Wu, W.; and Zhou, H. 2023b. Binarizing by Classification: Is Soft Function Really Necessary? *IEEE Transactions on Circuits and Systems for Video Technology*, 1–1.

Kenton, J. D. M.-W. C.; and Toutanova, L. K. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, 2.

Krizhevsky, A.; and Hinton, G. 2009. Learning multiple layers of features from tiny images. *Handbook of Systemic Autoimmune Diseases*, 1(4).

Lan, Z.; Chen, M.; Goodman, S.; Gimpel, K.; Sharma, P.; and Soricut, R. 2019. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. In *International Conference on Learning Representations*.

Li, Y.; Xu, S.; Zhang, B.; Cao, X.; Gao, P.; and Guo, G. 2022. Q-ViT: Accurate and Fully Quantized Low-bit Vision Transformer. In Oh, A. H.; Agarwal, A.; Belgrave, D.; and Cho, K., eds., *Advances in Neural Information Processing Systems*.

Liang, Y.; Chongjian, G.; Tong, Z.; Song, Y.; Wang, J.; and Xie, P. 2021. EViT: Expediting Vision Transformers via Token Reorganizations. In *International Conference on Learning Representations*.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021a. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10012–10022.

Liu, Z.; Luo, W.; Wu, B.; Yang, X.; Liu, W.; and Cheng, K.-T. 2020a. Bi-real net: Binarizing deep network towards real-network performance. *International Journal of Computer Vision (IJCV)*, 128(1): 202–219.

Liu, Z.; Oguz, B.; Pappu, A.; Xiao, L.; Yih, S.; Li, M.; Krishnamoorthi, R.; and Mehdad, Y. 2022. BiT: Robustly Binarized Multi-distilled Transformer. In Oh, A. H.; Agarwal, A.; Belgrave, D.; and Cho, K., eds., *Advances in Neural Information Processing Systems*.

Liu, Z.; Shen, Z.; Li, S.; Helwegen, K.; Huang, D.; and Cheng, K.-T. 2021b. How do adam and training strategies help bnns optimization? In *International Conference on Machine Learning*. PMLR.

Liu, Z.; Shen, Z.; Savvides, M.; and Cheng, K.-T. 2020b. ReActNet: Towards Precise Binary Neural Network with Generalized Activation Functions. In *European Conference on Computer Vision (ECCV)*.

Liu, Z.; Wang, Y.; Han, K.; Zhang, W.; Ma, S.; and Gao, W. 2021c. Post-Training Quantization for Vision Transformer. In Beygelzimer, A.; Dauphin, Y.; Liang, P.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems*.

Martinez, B.; Yang, J.; Bulat, A.; and Tzimiropoulos, G. 2020. Training binary neural networks with real-to-binary convolutions. In *International Conference on Learning Representations*.

Mishra, A.; and Marr, D. 2018. Apprentice: Using Knowledge Distillation Techniques To Improve Low-Precision Network Accuracy. In *International Conference on Learning Representations*.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Pouransari, H.; and Ghili, S. 2014. Tiny ImageNet Visual Recognition Challenge.

Qin, H.; Ding, Y.; Zhang, M.; Yan, Q.; Liu, A.; Dang, Q.; Liu, Z.; and Liu, X. 2022. BiBERT: Accurate Fully Binarized BERT. In *International Conference on Learning Representations (ICLR)*.

Qin, H.; Gong, R.; Liu, X.; Shen, M.; Wei, Z.; Yu, F.; and Song, J. 2020. Forward and Backward Information Retention for Accurate Binary Neural Networks. In *IEEE CVPR*.

Rastegari, M.; Ordonez, V.; Redmon, J.; and Farhadi, A. 2016. Xnor-net: Imagenet classification using binary convolutional neural networks. In *European conference on computer vision*, 525–542. Springer.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3): 211–252.

Ryoo, M. S.; Piergiovanni, A.; Arnab, A.; Dehghani, M.; and Angelova, A. 2021. TokenLearner: Adaptive Space-Time Tokenization for Videos. In Beygelzimer, A.; Dauphin, Y.; Liang, P.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems*.

Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; and Jégou, H. 2021. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, 10347–10357. PMLR.

Xu, S.; Li, Y.; Ma, T.; Lin, M.; Dong, H.; Zhang, B.; Gao, P.; and Lu, J. 2023. Resilient Binary Neural Network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 10620–10628.

Yang, H.; Yin, H.; Molchanov, P.; Li, H.; and Kautz, J. 2021. NViT: Vision Transformer Compression and Parameter Redistribution. *CoRR*, abs/2110.04869.

Yin, M.; Uzkent, B.; Shen, Y.; Jin, H.; and Yuan, B. 2023. GOHSP: A Unified Framework of Graph and Optimization-Based Heterogeneous Structured Pruning for Vision Transformer. In *AAAI*, 10954–10962. AAAI Press.

Yu, S.; Chen, T.; Shen, J.; Yuan, H.; Tan, J.; Yang, S.; Liu, J.; and Wang, Z. 2021. Unified Visual Transformer Compression. In *International Conference on Learning Representations*.

Yuan, Z.; Xue, C.; Chen, Y.; Wu, Q.; and Sun, G. 2022. PTQ4ViT: Post-training Quantization forVision Transformers withTwin Uniform Quantization. *Springer, Cham*.

Zhang, J.; Peng, H.; Wu, K.; Liu, M.; Xiao, B.; Fu, J.; and Yuan, L. 2022a. MiniViT: Compressing Vision Transformers with Weight Multiplexing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12145–12154.

Zhang, J.; Su, Z.; Feng, Y.; Lu, X.; Pietikäinen, M.; and Liu, L. 2022b. Dynamic Binary Neural Network by Learning Channel-Wise Thresholds. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1885–1889.

Zhang, Y.; Zhang, Z.; and Lew, L. 2022. PokeBNN: A Binary Pursuit of Lightweight Accuracy. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 12465–12475. Los Alamitos, CA, USA: IEEE Computer Society.

Zhang, Z.; Zhang, H.; Zhao, L.; Chen, T.; ; Arık, S. ; and Pfister, T. 2022c. Nested Hierarchical Transformer: Towards Accurate, Data-Efficient and Interpretable Visual Understanding. In *AAAI Conference on Artificial Intelligence (AAAI)*.