Bi-ViT: Pushing the Limit of Vision Transformer Quantization

Yanjing Li[1*], Sheng Xu[1*], Mingbao Lin[2], Xianbin Cao[1†], Chuanjian Liu[3], Xiao Sun[4†], Baochang Zhang[5,6,7]

[1] Beihang University
[2] Tencent
[3] Huawei Noah's Ark Lab
[4] Shanghai Artificial Intelligence Laboratory
[5] Zhongguancun Laboratory
[6] Hangzhou Research Institute, Beihang University
[7] Nanchang Institute of Technology

Abstract

Vision transformers (ViTs) quantization offers a promising prospect to facilitate deploying large pre-trained networks on resource-limited devices. Fully-binarized ViTs (Bi-ViT) that pushes the quantization of ViTs to its limit remain largely unexplored and a very challenging task yet, due to their unacceptable performance. Through extensive empirical analyses, we identify the severe drop in ViT binarization is caused by attention distortion in self-attention, which technically stems from the gradient vanishing and ranking disorder. To address these issues, we first introduce a learnable scaling factor to reactivate the vanished gradients and illustrate its effectiveness through theoretical and experimental analyses. We then propose a ranking-aware distillation method to rectify the disordered ranking in a teacher-student framework. Bi-ViT achieves significant improvements over popular DeiT and Swin backbones in terms of Top-1 accuracy and FLOPs. For example, with DeiT-Tiny and Swin-Tiny, our method significantly outperforms baselines by 22.1% and 21.4% respectively, while 61.5× and 56.1× theoretical acceleration in terms of FLOPs compared with real-valued counterparts on ImageNet. Our codes and models are attached on https://github.com/YanjingLi0202/Bi-ViT/

Introduction

Transformers, which have gained far-flung fame in natural language processing (NLP) area (Devlin et al. 2018; Qin et al. 2022), are also attracting increasing attention in lots of computer vision (CV) tasks, such as object detection (Carion et al. 2020), image classification (Dosovitskiy et al. 2020) and many others (He et al. 2022; Tian et al. 2022), impelling the widespread research on vision transformers (ViTs). There has a natural fit for ViTs to achieve better performance simply by training a larger model on a larger data set. For example, historical records show better performance of a ViT-H model (Dosovitskiy et al. 2020) accompanying with astonishing 632M parameters and 162G FLOPs. Such a high model complexity poses a great challenge to deploy models on platforms with short resource supplies. Therefore, both academia and industry call for an ultimate compression of these large models, and the past years have witnessed some promising techniques such as network pruning (Yang et al. 2021; Chen et al. 2023), low-rank decomposition (Denil et al. 2013), knowledge distillation (Hao et al. 2021; Xu et al. 2022b; Li et al. 2023c), and quantization (Li et al. 2022; Xu et al. 2023a; Li et al. 2023a).

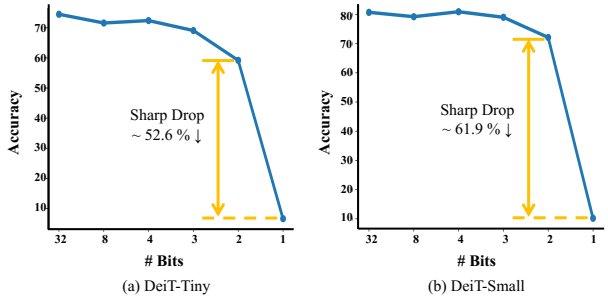


Figure 1: Performance of real-valued and quantized DeiT (Touvron et al. 2021) with varying bit-widths. We report results with (a) DeiT-Tiny and (b) DeiT-Small on ImageNet (Krizhevsky, Sutskever, and Hinton 2012), respectively. Here 8-bit DeiT is quantized with PTQ method (Lin et al. 2022) and 2/3/4 bit DeiT is trained with QAT method (Li et al. 2022). The binarized ViT is conducted with the baseline method Bi-Real Net (Liu et al. 2018).

Network quantization, which represents weights and activations in a low-bit format, has got great earnestness of many researchers for its reduced memory access costs and increased compute efficiency as well as performance benefit. Using the lower-bit quantized data, in particular to the extreme 1-bit case, requires less data movement, both on-chip and off-chip, and therefore reduces memory bandwidth and saves significant energy. Existing documentary records observe 32× less network size and 58× speedups beneficial from xnor and bit-count logics for 1-bit networks (Rastegari et al. 2016). Earlier attempts (Liu et al. 2021b; Lin et al. 2022) apply post-training quantization (PTQ) (Banner,

*These authors contributed equally.
†Corresponding author.

Nahshan, and Soudry 2019; Zhong et al. 2022) directly to ViTs without data-driven fine-tuning, causing sub-optimal performance, in particular to impotent 1-bit ViTs. Therefore, by quantizing while training, quantization-aware training (QAT) methods are more congenial to 1-bit ViTs. Extensive empirical studies (Liu et al. 2020; Xu et al. 2022a; Qin et al. 2022; Xu et al. 2023b) have well demonstrated the efficacy of QAT methods in 1-bit convolutional neural networks (CNNs) or BERTs, however, the application to 1-bit ViTs remains not to be fully explored so far.

In this paper, we first build a fully-binarized ViT baseline, a straightforward solution constructed upon popular binarized QAT method of Bi-Real Net (Liu et al. 2018). Through an empirical study of this baseline, we observe significant performance drops on the ImageNet dataset (Krizhevsky, Sutskever, and Hinton 2012), as shown in Fig. 1. For instance, extending Bi-Real Net to binarize DeiT-Tiny (Touvron et al. 2021) incurs a tremendous performance gap of 52.6% in the Top-1 accuracy compared to the 2-bit quantized counterpart. Similar performance drops occur in DeiT-Small as well. Delving into a deeper analysis, we find that the incompatibility of existing QAT methods mainly stems from the binarized self-attention module in ViTs, where a simple application of existing binarization methods (Liu et al. 2018) leads to severe attention distortion, as plotted in Fig. 2 (a) and Fig. 2 (b), especially in the diagonal scores of the map which are supposed to be the most attentive.

In this paper we dig deeper into this attention distortion problem. Through empirical analysis, we find that this phenomenon is mainly caused by gradient vanishing due to the straight-through-estimator (STE) (Bengio, Léonard, and Courville 2013) and non-scaled binarization in self-attention. Meanwhile, a simple distillation utilizing distillation token in DeiT (Touvron et al. 2021) and KL-divergence in ReActNet (Liu et al. 2020) is ineffective in dismissing the ranking disorder, since it neglects the relative order of the attention map between the binarized ViTs and their real-valued counterpart. To address the aforementioned issues, a fully-binarized ViT (Bi-ViT) is developed by reactivating the vanished gradients through a learnable scaling factor in self-attention and a ranking-aware distillation to further effectively rectify the disordered ranking of attention (see the overview in Fig. 3). In addition, we also provide both empirical and theoretical analysis about how our method can rectify the distorted attention and thus promote the optimization of Bi-ViT. The contributions of our work are summarized as:

- We identify the bottleneck of a fully-binarized ViT through empirical analyses and formulate the problem in a theoretical perspective. Based on these, we introduce learnable head-wise scaling factor into binarized self-attention to reactivate the vanished gradients.
- We develop a ranking-aware distillation scheme to eliminate attention distortion. Our distillation method fully utilizes the ranking-aware knowledge from the real-valued teacher to promote the optimization of Bi-ViT.
- Our Bi-ViT is the first promising way to push the limit of ViT quantization to the fully-binarized version. Extensive experiments on the ImageNet benchmark demonstrate that Bi-ViT surpasses both the baseline and prior binarized methods by a significant margin, achieving a remarkable acceleration rate of up to 61.5×.

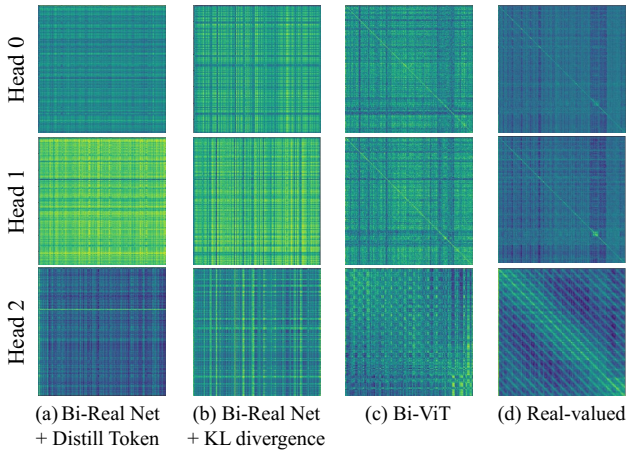


Figure 2: Visualization of the attention map before softmax in the first block of DeiT-Tiny (Touvron et al. 2021) on ImageNet (Krizhevsky, Sutskever, and Hinton 2012). From the left to right, is the baseline method (Liu et al. 2018), previous binarization method (Xu et al. 2022a), our Bi-ViT and real-valued counterpart.

Related Work

Vision Transformer. Unlike traditional CNN-based models, ViTs are capable of capturing long-range visual relationships through the self-attention mechanism, and offer a more generalizable paradigm without inductive bias specific to images. The starting ViT (Dosovitskiy et al. 2020) views an image as a sequence of 16 × 16 patches and uses a unique class token to predict the classification, yielding promising results. Subsequently, many works, such as DeiT (Touvron et al. 2021) and PVT (Wang et al. 2021), have improved upon ViT, making it more efficient and applicable to downstream tasks. However, these high-performing ViTs have also accompanied with a significant number of parameters and high computational overhead, limiting their widespread applications. Thus, designing smaller and faster ViTs has become a new trend. DynamicViT (Rao et al. 2021) proposes a dynamic token sparsification framework to progressively and dynamically prune redundant tokens, achieving a competitive complexity and accuracy trade-off. Evo-ViT (Xu et al. 2022c) proposes a slow-fast updating mechanism that ensures information flow and spatial structure, reducing both the training and inference complexity. While the aforementioned works focus on efficient model design, this paper aims to boost compression and acceleration through binarization.

Network Binarization. Network binarization is a technique originally proposed to train convolutional neural networks (CNNs) with binary weights. BinaryConnect (Courbariaux, Bengio, and David 2015) is the precursor to BinaryNet, where the parameters are binary while the activations re-

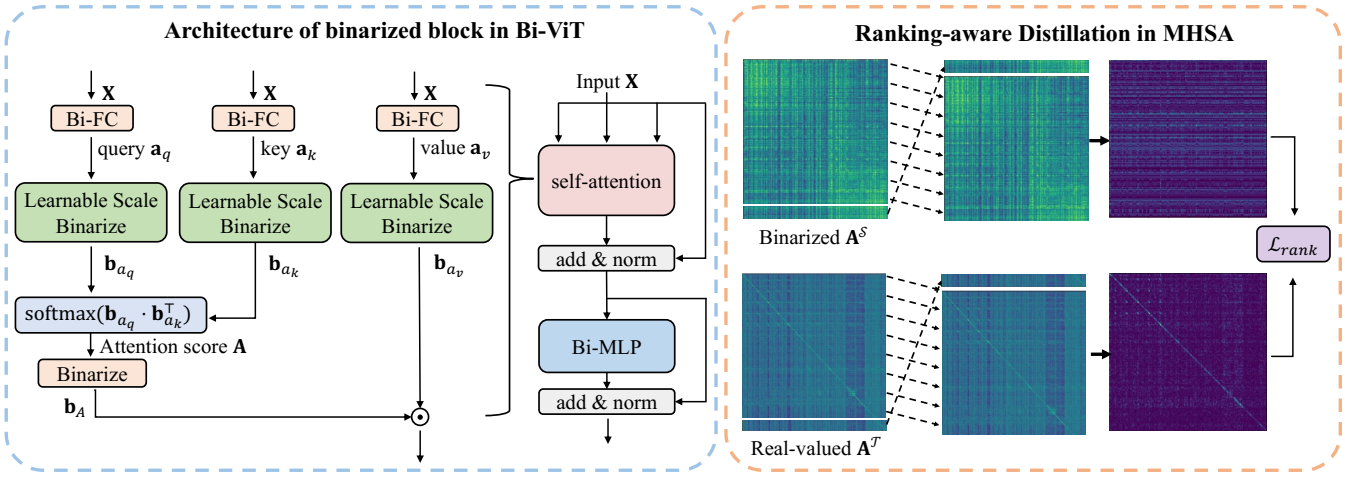


Figure 3: Overview of the proposed Bi-ViT framework. We introduce the learnable scaling factor in an architecture perspective and a ranking-aware distillation scheme incorporated in the optimization process. From left to right, we respectively show the detailed architecture of single block in Bi-ViT and the distillation framework of Bi-ViT.

main in full-precision states. XNOR-Net (Rastegari et al. 2016) was introduced to improve convolution efficiency by binarizing the weights and inputs of convolution kernels. Bi-Real Net (Liu et al. 2018) explores a new variant of residual structure to preserve the information of real activations before the sign function, with a tight approximation to the derivative of the non-differentiable sign function. ReAct-Net (Liu et al. 2020) replaces the conventional PReLU and the sign function of the BNNs with RPReLU and RSign with a learnable threshold, thus improving the performance of BNNs. RBONN (Xu et al. 2022a) introduces a recurrent bilinear optimization to address the asynchronous convergence problem for BNNs, which further improves the performance of BNNs. DCP-NAS (Li et al. 2023b) proposes an architecture with better performance on binarized format than real-valued counterpart. These techniques improve the efficiency and accuracy of binary neural networks (BNNs) and allow them to be applied in practical applications. Majorities of these techniques consider non-scaled binarization in activations, which is beneficial to conventional CNNs while causing gradient mismatch issue for the pecularity of self-attention mechanism in ViTs.

Background

Multi-Head Self-Attention and Binarization

For a multi-head self-attention (MHSA) module, we denote its query, key, and value set as $\{\mathbf{a}_{\{q,k,v\}} \in \mathbb{R}^{h\times N\times d}\}$, where h denotes head number, N and d represent the patch and channel numbers of each head. Specifically, $N = (W_{in}//W_{in}^P) \times (H_{in}//H_{in}^P)$ where W_{in} and H_{in} are the width and height of the feature, W_{in}^P, H_{in}^P are the width and height of patch maps respectively. Then, the attention score $\mathbf{A}$ and MHSA module output $\mathbf{a}_{out}$ are computed as follows (Vaswani et al. 2017):

$$\begin{aligned} \mathbf{A} &= \mathrm{softmax}[(\mathbf{a}_q \cdot \mathbf{a}_k^\top)/\sqrt{d}], \\ \mathbf{a}_{out} &= \mathbf{A} \cdot \mathbf{a}_v^\top, \end{aligned} \tag{1}$$

where $\mathrm{softmax}(\cdot)$ represents the softmax operation. Intending to represent query, key, value and attention score, *i.e.*, $\mathbf{a}_q$, $\mathbf{a}_k$, $\mathbf{a}_v$ and $\mathbf{A}$, in a 1-bit format, Eq. (1) changes into:

$$\begin{aligned} \mathbf{A} &= \mathrm{softmax}[(\mathbf{b}_{a_q} \cdot \mathbf{b}_{a_k}^\top)/\sqrt{d}], \\ \mathbf{a}_{out} &= \mathbf{b}_A \cdot \mathbf{b}_{a_v}^\top. \end{aligned} \tag{2}$$

We follow the common network binarization methods (Rastegari et al. 2016) that use the sign function $\mathbf{b}_\cdot = \mathrm{sign}(\cdot)$ in the binary forward pass, and STE (Bengio, Léonard, and Courville 2013) $\frac{\partial \mathbf{b}_\cdot}{\partial \cdot} = 1_{|\cdot|\leq 1}$ to compute the gradient for sign function in its backward pass. We omit the non-linear function here for simplicity. For all the projection and linear layers in binarized ViTs, we conduct binarization following (Qin et al. 2022; Liu et al. 2018) as $\mathbf{a}_{out} = \mathbf{b}_{a_{in}} \cdot (\boldsymbol{\alpha}_w \circ \mathbf{b}_w)^\top = \boldsymbol{\alpha}_w \circ (\mathbf{b}_{a_{in}} \cdot \mathbf{b}_w^\top)$ where $\boldsymbol{\alpha}_w = \{\alpha_w^1, \alpha_w^2, ..., \alpha_w^{C_{out}}\} \in \mathbb{R}_+^{C_{out}}$ is known as the channel-wise scaling factor vector (Rastegari et al. 2016) and $\circ$ represents channel-wise multiplication. The matrix multiplication process, *i.e.*, $\mathbf{b}_{a_{in}} \cdot \mathbf{b}_w^\top$, can be executed by the efficient XNOR and Bit-count instructions on edge devices.

Bottleneck of Fully-Binarized ViTs

The high-performing ViTs are built on premise of transformer's supreme ability to model the long-range relationships thanks to the attention mechanism within the MHSA module. Unfortunately, a binarized version of weights and inputs significantly weakens the representation ability. In addition, the sign function and clamp operation also damage the optimization of backpropagation. To be more evident, we perform quantitative ablative experiments where we replace weights or activations in each module of the real-valued DeiT-Tiny (Touvron et al. 2021) with a binarized one and report the resulting Top-1 accuracy drop on the ImageNet dataset (Krizhevsky, Sutskever, and Hinton 2012) after a total of 50 training epochs. Fig. 4 reports the results and we go on a deeper analysis below.

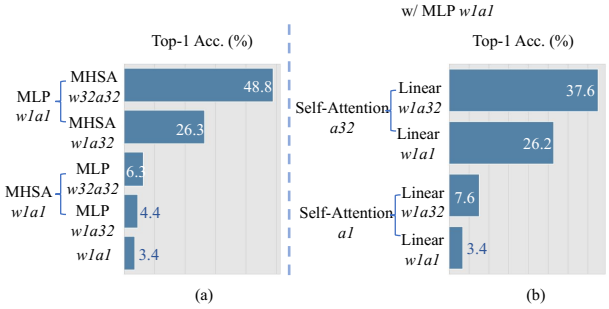
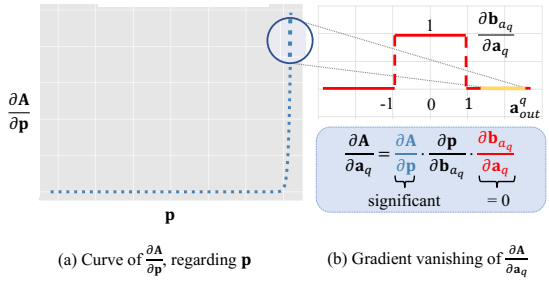


Figure 4: Performance of fully-binarized DeiT-Tiny on ImageNet (Krizhevsky, Sutskever, and Hinton 2012) with different binarized/real-valued settings.



(a) Curve of $\frac{\partial \mathbf{A}}{\partial \mathbf{p}}$, regarding $\mathbf{p}$ (b) Gradient vanishing of $\frac{\partial \mathbf{A}}{\partial \mathbf{a}_q}$

Figure 5: Gradient mismatch between Eq. (5) and Eq. (7).

Module Degradation. By gradually replacing the multilayer perceptron (MLP) and MHSA modules with real-valued weights or activations, we have discovered that maintaining the MLP as "*w1a1*" (all weights and activations in the MLP are binarized) still results in satisfactory performance. For instance, keeping MLP as "*w1a1*" while keeping MHSA as "*w1a32*" obtains 26.3% Top-1 accuracy, which might be acceptable comparing to the 55.2% of real-valued DeiT-Tiny when taking into consideration 47.3× acceleration rates. On the contrast, when maintaining MHSA module as "*w1a1*", we observe a significant drop in performance. To be more specific, even when the MLP was maintained as "*w32a32*", we still observe a significant 50.8% decrease in Top-1 accuracy (from 55.2% to 4.4%). This result indicates that using binarized weights and activations in the MHSA module can have a substantial negative impact on the model's performance, even when other parts retain in real-valued states.

Operation Degradation. To better understand the impact of fully-binarized ViT's performance, we conduct further analyses by examining the operations within the MHSA module. Specifically, when we maintain the self-attention activations in Eq. (1) as real-valued ("*a32*"), we observe only a relatively small decrease in performance from 48.8% to 37.6%. However, when the self-attention activations in Eq. (2) are binarized, significant drops in accuracy occur from 48.8% to 7.6%. This finding highlights the importance of the self-attention process within the MHSA module and suggests more efforts to mitigate the negative impact of binarization on the MHSA module.

Our Bi-ViT

In this section, we propose to dismiss the affect of gradient mismatch mentioned in Sec. 4.1 from perspectives of gradient approximation in Sec. 4.2 and intermediate distillation in Sec. 4.3.

Gradient Mismatch in Self-Attention

With conclusion from the experimental results in Sec. 3.2 that self-attention process, *i.e.*, Eq. (2), is the most critical part causing the performance drops. We attempt to analyze the underlying reasons for this phenomenon from an optimization perspective. For simplicity, we derive the gradient mismatch in $\mathbf{a}_q$ as an example, and the analysis can be applicable to explain $\mathbf{a}_k$ as well. We first represent the features before $\mathrm{softmax}(\cdot)$ in Eq. (2) as:

$$\mathbf{p} = (\mathbf{b}_{a_q} \cdot \mathbf{b}_{a_k}^{\top}.)/\sqrt{d}. \tag{3}$$

The gradient of $\mathbf{a}_q^{h_i,n,c}$ *w.r.t.* $\mathbf{A}$ is formulated as:

$$\frac{\partial \mathbf{A}}{\partial \mathbf{a}_q^{h_i,n,c}} = \frac{\partial \mathbf{A}}{\partial \mathbf{p}^{h_i,n,n'}} \cdot \frac{\partial \mathbf{p}^{h_i,n,n'}}{\partial \mathbf{b}_{a_q}^{h_i,n,c}} \cdot \frac{\partial \mathbf{b}_{a_q}^{h_i,n,c}}{\partial \mathbf{a}_q^{h_i,n,c}}, \tag{4}$$

where $h_i \in \mathbb{R}^h$, n & $n' \in \mathbb{R}^N$, $c \in \mathbb{R}^d$ and the gradient of $\mathbf{a}_k$ is likewise. The explicit form of the first item $\frac{\partial \mathbf{A}}{\partial \mathbf{p}^{h_i,n,n'}}$ in Eq. (4) is:

$$\begin{aligned} \frac{\partial \mathbf{A}}{\partial \mathbf{p}_{h_i,n,n'}} &= \frac{\partial\, \mathrm{softmax}(\mathbf{p}_{h_i,n,n'})}{\partial \mathbf{p}_{h_i,n,n'}} \\ &= \mathbf{A}_{h_i,n,n'} \otimes (1 - \mathbf{A}_{h_i,n,n'}), \end{aligned} \tag{5}$$

where $\otimes$ denotes Hadamard product. And the second item is formulated as:

$$\begin{aligned} \frac{\partial \mathbf{p}^{h_i,n,n'}}{\partial \mathbf{b}_{a_q}^{h_i,n,c}} &= \frac{\partial \mathbf{b}_{a_q}^{h_i,n,c} \cdot \mathbf{b}_{a_k}^{\top h_i,c,n'}}{\partial \mathbf{b}_{a_q}^{h_i,n,c}} \\ &= \mathbf{b}_{a_k}^{\top h_i,c,n'}/\sqrt{d}, \end{aligned} \tag{6}$$

result of which is therefore correlated with $\mathbf{b}_{a_k}$. The third item is solved through STE (Bengio, Léonard, and Courville 2013) as:

$$\frac{\partial \mathbf{b}_{a_q}^{h_i,n,c}}{\partial \mathbf{a}_q^{h_i,n,c}} = \mathbf{1}_{|\mathbf{a}_q^{h_i,n,c}| \le 1}. \tag{7}$$

Combing Eq. (5)−Eq. (7), we have the final gradient form in fully-binarized ViTs as:

$$\begin{aligned} \frac{\partial \mathbf{A}}{\partial \mathbf{a}_q^{h_i,n,c}} &= \frac{\partial \mathbf{A}}{\partial \mathbf{p}^{h_i,n,n'}} \cdot \frac{\partial \mathbf{p}^{h_i,n,n'}}{\partial \mathbf{b}_{a_q}^{h_i,n,c}} \cdot \frac{\partial \mathbf{b}_{a_q}^{h_i,n,c}}{\partial \mathbf{a}_q^{h_i,n,c}} \\ &= \mathbf{A}_{h_i,n,n'}(1 - \mathbf{A}_{h_i,n,n'}) \cdot \mathbf{b}_{a_k}^{h_i,c,n'} \cdot \mathbf{1}_{|\mathbf{a}_q^{h_i,n,c}| \le 1}/\sqrt{d}. \end{aligned} \tag{8}$$

Considering $\mathbf{b}_{a_q}^{h_i,n,:} = [1, \cdots, 1]$ and $\cdot \mathbf{b}_{a_k}^{h_i,n',:} = [1, \cdots, 1]$ as the extreme condition, $\mathbf{b}_{a_q}^{h_i,n,:} \cdot \mathbf{b}_{a_k}^{\top h_i,:,n'} = d$. Therefore, a specific element in $\mathbf{b}_{a_q} \cdot \mathbf{b}_{a_k}^{\top}$ is $\in \{-d, \cdots, d\}$.

We plot the curve of a specific element in the first item between $[-64, 64]$ in Fig. 5 (a) as $d = 64$ in DeiT-Tiny (Touvron et al. 2021). We observe $\frac{\partial \mathbf{A}}{\partial \mathbf{p}_{h_i,n,n'}}$ sharply magnified when $\mathbf{p}_{h_i,n,n'}$ increases. As shown in Fig. 5 (b), when $\mathbf{p}_{h_i,n,n'}$ has a large magnitude, $|\mathbf{a}_q| > 1$ and $\frac{\partial \mathbf{b}_{a_q}^{h_I,n,c}}{\partial \mathbf{a}_q^{h_i,n,c}} = 0$. Thus the multiplication of these two items leads to $\frac{\partial \mathbf{A}}{\partial \mathbf{a}_q^{h_i,n,c}} = 0$, likewise for $\mathbf{a}_k$. Therefore we formulate the gradient mismatch phenomenon in the aforementioned theoretical analysis. And such gradient mismatch leads to distorted gradient in the optimization of $\mathbf{a}_q$ & $\mathbf{a}_k$ and therefore degrades performance of fully-binarized ViTs.

Learnable Head-wise Scaling Factor

As one of the solution to the above mentioned problem, we propose a head-wise scaling factor binarization scheme for the self-attention process, where the scaling factors are learned during training to first modify the gradient clip range in Fig. 5(b). Eq. (2) is changed into:

$$\begin{aligned} \tilde{\mathbf{A}} &= \mathrm{softmax}(\tilde{\mathbf{p}}), \\ \tilde{\mathbf{p}} &= (\boldsymbol{\alpha}_q \otimes \boldsymbol{\alpha}_k) \circ (\mathbf{b}_{a_q} \cdot \mathbf{b}_{ak}^{\top})/\sqrt{d} \\ &= \boldsymbol{\alpha}_{q;k} \circ (\mathbf{b}_{a_q} \cdot \mathbf{b}_{a_k}^{\top})/\sqrt{d}, \end{aligned} \tag{9}$$

and

$$\begin{aligned} \tilde{\mathbf{a}}_{out} &= (\boldsymbol{\alpha}_A \circ \mathbf{b}_A) \cdot (\boldsymbol{\alpha}_v \circ \mathbf{b}_{a_v})^{\top} \\ &= (\boldsymbol{\alpha}_A \otimes \boldsymbol{\alpha}_v) \circ (\mathbf{b}_A \cdot \mathbf{b}_{a_v}^{\top}) \\ &= \boldsymbol{\alpha}_{A;v} \circ (\mathbf{b}_A \cdot \mathbf{b}_{a_v}^{\top}), \end{aligned} \tag{10}$$

where $\mathbf{b}_{a_\cdot} = \mathrm{sign}(\frac{\mathbf{a}_\cdot}{\boldsymbol{\alpha}_\cdot})$, $\boldsymbol{\alpha}_q$, $\boldsymbol{\alpha}_k$, $\boldsymbol{\alpha}_v$ and $\boldsymbol{\alpha}_A$ are the head-wise learnable scaling factors in binarized MHSA, where $\boldsymbol{\alpha}_{\{q,k,v,A\}} = \{\alpha^1_{\{q,k,v,A\}}, \alpha^2_{\{q,k,v,A\}}, \cdots, \alpha^h_{\{q,k,v,A\}}\} \in \mathbb{R}^h_+$. The second rows in Eq. (9) & Eq. (10) are established since the scaling factors are aligned with the head dimension, which is independent with the matrix multiplication operation. Thus, $\boldsymbol{\alpha}_{q;k} = \{\alpha^1_{q;k}, \alpha^2_{q;k}, \cdots, \alpha^h_{q;k}\} \in \mathbb{R}^h_+$ and $\boldsymbol{\alpha}_{A;v} = \{\alpha^1_{A;v}, \alpha^2_{A;v}, \cdots, \alpha^h_{A;v}\} \in \mathbb{R}^h_+$.

Consequently, the gradient $\frac{\partial \tilde{\mathbf{A}}}{\partial \mathbf{a}_q^{i,n,c}}$ in Eq. (8) is further formulated in our Bi-ViT as:

$$\frac{\partial \tilde{\mathbf{A}}}{\partial \mathbf{a}_q^{h_i,n,c}} = \underbrace{\tilde{\mathbf{A}}^{h_i,n,n'}(1 - \tilde{\mathbf{A}}^{h_i,n,n'})}_{\frac{\partial \tilde{\mathbf{A}}}{\partial \mathbf{p}^{h_i,n,n'}}} \cdot \underbrace{\boldsymbol{\alpha}_{q;k}^{h_i} \circ \mathbf{b}_{ak}^{h_i,c,n'}}_{\frac{\partial \mathbf{p}^{h_i,n,n'}}{\partial \mathbf{b}_{a_q}^{h_i,n,c}}} \cdot \underbrace{\mathbf{1}_{|\mathbf{a}_q^{h_i,n,c}| \le \boldsymbol{\alpha}_q}}_{\frac{\partial \mathbf{b}_{a_q}^{h_i,n,c}}{\partial \mathbf{a}_q^{h_i,n,c}}}. \tag{11}$$

Since $\mathrm{softmax}(.)$ and $\circ$ are aligned with different dimensions, the value of Eq. (5) remains unchanged ($\mathrm{softmax}(\mathbf{p}) = \mathrm{softmax}(\boldsymbol{\alpha}_{q;k} \circ \mathbf{p})$). As can be seen, the threshold of gradient clip in Eq. (7) changes from 1 into $\boldsymbol{\alpha}_q$, which means that we can surpass the occurance of gradient mismatch by modifying the value of $\boldsymbol{\alpha}_q$. Note that the scaling factor ($\boldsymbol{\alpha}_q$) is to imitate the magnitude of the latent activations. When $\tilde{\mathbf{p}}$ has a large magnitude, *i.e.*, in the circled part of Fig. 5 (a), $\boldsymbol{\alpha}_q$ also tends to be larger and $\mathbf{a}_q^{h_i,n,c}$ locates in the field that $\frac{\partial \mathbf{b}_{a_q}^{h_i,n,c}}{\partial \mathbf{a}_q^{h_i,n,c}} > 0$. Thus the vanishing gradients are reactivated through the introduced learnable scaling factor.

Ranking-aware Distillation for Bi-ViT

Fig. 2 illustrates a significant difference in the attention map's relative order between Bi-RealNet (a) and its real-valued counterpart (c). This difference could result in a notable decrease in performance. To address this issue during binarized training, a ranking-aware distillation in a teacher-student framework is introduced:

$$\mathcal{L}_{ranking} = \sum_{l=1}^{L} \|\psi(\mathbf{A}^{\mathcal{T}}) - \psi(\mathbf{A}^{\mathcal{S}})\|_2, \tag{12}$$

where $\mathbf{A}^{\mathcal{T}}$ and $\mathbf{A}^{\mathcal{S}}$ represents the attention scores from the real-valued teacher and binarized student. $\psi(\cdot)$ denotes the function for obtaining the ranking, *i.e.*, relative order of an attention score, which is formulated as:

$$\psi(\mathbf{A}^{:,n,:}) = \begin{cases} \mathbf{A}^{:,n,:} - \mathbf{A}^{:,n-1,:}, & \text{if } \ 0 < n \le N-1 \\ \mathbf{A}^{:,0,:} - \mathbf{A}^{:,N-1,:}, & \text{otherwise}\,. \end{cases} \tag{13}$$

Detailed relative order computation can be seen in the right part of Fig. 3. We implement our Bi-ViT under the teacher-student framework (Touvron et al. 2021), thus the final objective of our method is formulated as:

$$\mathcal{L} = \mathcal{L}_{dist} + \lambda \mathcal{L}_{ranking}, \tag{14}$$

where λ is a hyper-parameter to balance these two loss functions.

Experiments

In this section, we evaluate the performance of the proposed Bi-ViT model for image classification task using popular DeiT (Touvron et al. 2021) & Swin (Liu et al. 2021a) backbones and object detection task using Mask R-CNN (He et al. 2017) & Cascade (Cai and Vasconcelos 2018) Mask R-CNN with Swin-Tiny (Liu et al. 2021a) backbone. To the best of our knowledge, there is no publicly available source codebase on fully-binarized ViTs at this point, so we re-implement the baseline *i.e.*, Bi-Real Net (Liu et al. 2018) methods.

Datasets and Implementation Details

Datasets. The experiments are conducted on the ImageNet ILSVRC12 dataset (Krizhevsky, Sutskever, and Hinton 2012) for image classification task. The ImageNet dataset is challenging due to its large scale and greater diversity. There are 1000 classes and 1.2 million training images, and 50k validation images in it. In our experiments, we use the classic data augmentation method described in (Touvron et al. 2021).

Experimental settings. In our experiments, we initialize the weights of binarized model with the pretrained real-valued model. The binarized model is trained for 300 epochs with batch-size 512 and the base learning rate $5e-4$ without warm-up scheme. For all the experiments, we apply LAMB (You et al. 2020) optimizer with weight decay set as

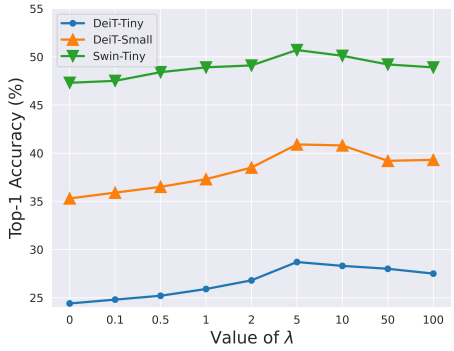


Figure 6: Effect of hyper-parameter λ on ImageNet (Krizhevsky, Sutskever, and Hinton 2012).

0, following DeiT III (Touvron, Cord, and Jégou 2022). Note that we keep the patch embedding (first) layer and the classification (last) layer as real-valued, following (Esser et al. 2019).

Backbone. We evaluate our binarized method on two popular vision transformer networks: DeiT (Touvron et al. 2021) and Swin Transformer (Liu et al. 2021a). The DeiT-Tiny, DeiT-Small, DeiT-Base, Swin-Tiny and Swin-Small are adopted as the backbone models, whose Top-1 accuracy on ImageNet dataset are 72.2%, 79.9%, 81.8%, 81.2%, and 83.2% respectively. For a fair comparison, we utilize the official implementation of DeiT and Swin Transformer.

Ablation Study

Hyper-parameter Selection. We λ of Eq. (14) in this part, with experiments conducted on ImageNet (Krizhevsky, Sutskever, and Hinton 2012) dataset. We show the model performance (Top-1 accuracy) with different setups of hyper-parameter λ in Fig. 6, in which the performances increase first and then decrease with the uplift of λ from left to right. Since λ controls the importance of $\mathcal{L}_{ranking}$, we show that the vanilla baseline ($\lambda = 0$) performs worse than any versions with Ranking-aware Distillation loss ($\lambda > 0$), showing the proposed distillation scheme is necessary. With the varying value of λ, we find $\lambda = 5$ boost the performance of our Bi-ViT, achieving 28.7%, 40.9% and 50.7% Top-1 accuracy on ImageNet (Krizhevsky, Sutskever, and Hinton 2012) with DeiT-Tiny, DeiT-Small and Swin-Tiny backbone, respectively.

Effectiveness of components. We conduct the ablative experiments regarding the proposed components on DeiT-Tiny network. Firstly, we compose the baseline network using the binarization method following Bi-Real Net (Liu et al. 2018). As shown in the third row of Tab. **??**, the baseline networks only obtains 6.6% Top-1 accuracy, which is far from satisfactory. With the introduction of our first novelty, *i.e.*, learnable scaling factor (LSF), the baseline network is boosted by 17.8%, achieving 24.4% Top-1 accuracy. We also observe the other contribution Ranking-aware Disitllation (RD) singly promotes the baseline network by 5.9%, which is also significant on ImageNet dataset. By combining the two main contributions together, we get Bi-ViT, outperforming the vanilla baseline by 22.1%.

Method	#Bits	Top-1$_{(\%)}$
Real-valued	32-32	72.1
Baseline (Bi-Real Net)	1-1	6.6
+ Learnable Scaling Factor (LSF)	1-1	$24.4^{+17.8}$
+ Ranking-aware Distillation (RD)	1-1	$12.5^{+5.9}$
+ LSF + RD (Bi-ViT)	1-1	$\mathbf{28.7}^{+\mathbf{22.1}}$

Table 1: Evaluating the components of Bi-ViT based on DeiT-Tiny (Touvron et al. 2021) backbone. "#Bits" denotes the bit-width of weights and activations.

Results on Image Classification

The experimental results are shown in Tab. **??**. We compare our method with 1-bit methods including BiBERT (Qin et al. 2022), RBONN (Xu et al. 2022a), and Bi-Real Net (Liu et al. 2018) based on the same frameworks for the task of image classification with the ImageNet dataset. We also report the classification performance of the low-bit training-aware quantization method Q-ViT (Li et al. 2022) for further reference. We use model size and OPs following (Liu et al. 2018) in comparison to other bit-width models for further reference. We firstly evaluate the proposed method on DeiT models. For DeiT-Tiny backbone, compared with other binary methods, our Bi-ViT achieves significant performance improvements. For example, our Bi-ViT surpasses the baseline Bi-Real Net (Liu et al. 2018) by 22.1% Top-1 accuracy, which is significant and meaningful for real-world applications. And it is worth noting that the proposed 1-bit model significantly compresses the DeiT-Tiny by 61.5× on OPs. The proposed method also boosts the performance of baseline by 21.7% with the same architecture and bit-width using DeiT-Small bacobone, a significant improvement on the ImageNet dataset. For larger DeiT-B, as shown in Tab. **??**, the performance of the proposed method outperforms the Bi-Real Net by 20.8%, a large margin. Also note that the proposed 1-bit model significantly compresses the DeiT-B by 60.2× on OPs and 28.6× on model size.

Also, our method obtains convincing results on Swin-transformers. As shown in Tab. **??**, the performance of the proposed method with Swin-Tiny outperforms the baseline method by 21.4%, a large margin. For larger Swin-Small, the performance of the proposed method outperforms the 1-bit baseline by 21.5%. Also note that our method theoritically accelerates the network by 58.3×, which demonstrates the effectiveness and efficiency of our Bi-ViT.

Conclusion

In this paper, we present Bi-ViT, an improved version of fully-binarized ViTs that offers a high compression ratio and acceptable performance. Initially, we establish a empirical framework for fully-binarized ViT and analyze the bottlenecks of the baseline. Our empirical analysis shows that attention distortion in MHSA is the primary cause of the significant drop in ViT binarization, which results from gradient vanishing and ranking disorder. To address these is-

Network	Method	#Bits	Size$_{(\mathrm{MB})}$	OPs$_{(10^8)}$	Top-1$_{(\%)}$	Top-5$_{(\%)}$
DeiT-Tiny	Real-valued	32-32	22.8	12.3	72.2	91.1
	Q-ViT (Li et al. 2022)	4-4	3.0	1.6	74.3	91.7
		3-3	2.3	0.8	71.5	91.2
		2-2	1.7	0.4	59.0	81.8
	BiBERT (Qin et al. 2022)	1-1	1.0	0.2	5.9	16.0
	RBONN (Xu et al. 2022a)				6.3	16.9
	Bi-Real Net (Liu et al. 2018)				6.6	17.1
	Bi-ViT				**28.7**$^{+22.1}$	**51.7**$^{+34.6}$
DeiT-Small	Real-valued	32-32	88.2	45.5	79.9	95.0
	Q-ViT (Li et al. 2022)	4-4	11.4	5.8	80.9	94.9
		3-3	8.7	3.0	79.0	94.2
		2-2	6.0	1.5	72.1	90.3
	BiBERT (Qin et al. 2022)	1-1	3.4	0.8	17.4	29.7
	RBONN (Xu et al. 2022a)				18.5	30.0
	Bi-Real Net (Liu et al. 2018)				19.2	30.3
	Bi-ViT				**40.9**$^{+21.7}$	**65.0**$^{+34.7}$
DeiT-Base	Real-valued	32-32	346.2	174.7	81.8	95.6
	Q-ViT (Li et al. 2022)	4-4	44.1	22.0	83.0	96.1
		3-3	33.4	11.1	81.0	95.1
		2-2	22.7	5.7	74.2	92.2
	BiBERT (Qin et al. 2022)	1-1	12.1	2.9	24.5	36.3
	RBONN (Xu et al. 2022a)				26.1	38.6
	Bi-Real Net (Liu et al. 2018)				26.5	38.8
	Bi-ViT				**47.3**$^{+20.8}$	**72.8**$^{+34.0}$
Swin-Tiny	Real-valued	32-32	114.2	44.9	81.2	95.5
	Q-ViT (Li et al. 2022)	4-4	14.6	5.8	82.5	97.3
		3-3	11.2	3.0	80.9	96.1
		2-2	10.0	1.6	74.7	92.5
	BiBERT (Qin et al. 2022)	1-1	4.2	0.8	34.0	46.9
	RBONN (Xu et al. 2022a)				33.8	46.7
	Bi-Real Net (Liu et al. 2018)				34.1	46.9
	Bi-ViT				**55.5**$^{+21.4}$	**79.4**$^{+32.5}$
Swin-Small	Real-valued	32-32	199.8	87.5	83.2	96.2
	Q-ViT (Li et al. 2022)	4-4	25.3	11.1	84.4	98.3
		3-3	19.2	5.6	82.7	97.5
		2-2	13.0	2.9	76.9	94.9
	BiBERT (Qin et al. 2022)	1-1	6.9	1.5	39.4	53.0
	RBONN (Xu et al. 2022a)				39.0	52.7
	Bi-Real Net (Liu et al. 2018)				39.2	52.8
	Bi-ViT				**60.7**$^{+21.5}$	**83.9**$^{+31.1}$

Table 2: Experiments with DeiT (Touvron et al. 2021) and Swin (Liu et al. 2021a) on ImageNet (Krizhevsky, Sutskever, and Hinton 2012). "#Bits" denotes the bit-width of weights and activations. We report the Top-1$_{(\%)}$ and Top-5$_{(\%)}$ accuracy performances. The **bold** denotes the best result with binarized weights and activations.

sues, we introduce a learnable scaling factor that reactivates vanished gradients, which we illustrate through both theoretical and experimental analysis. Additionally, we propose ranking-aware distillation for Bi-ViT, which rectifies disordered ranking in a teacher-student framework. Our work provides a comprehensive analysis and effective solutions for the crucial issues in ViT full binarization, paving the way for the extreme compression of ViT.

Acknowledgements

This research was supported by Zhejiang Provincial Natural Science Foundation of China under Grant No. D24F020011, Beijing Natural Science Foundation L223024, National Natural Science Foundation of China under Grant 62076016 and under Grant 61827901, Foundation of China Energy Project GJNY-19-90. The work was also supported by the National Key Research and Development Program of China (Grant No. 2023YFC3300029) and "One Thousand Plan" projects in Jiangxi Province Jxsg2023102268 and ATR key laboratory grant 220402. 232-CXCY-A01-08-06-01.

References

Banner, R.; Nahshan, Y.; and Soudry, D. 2019. Post training 4-bit quantization of convolutional networks for rapid-deployment. In *Proc. of NeurIPS*, 7950–7958.

Bengio, Y.; Léonard, N.; and Courville, A. 2013. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*.

Cai, Z.; and Vasconcelos, N. 2018. Cascade r-cnn: Delving into high quality object detection. In *Proc. of CVPR*, 6154–6162.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-end object detection with transformers. In *Proc. of ECCV*, 213–229.

Chen, M.; Lin, M.; Li, K.; Shen, Y.; Wu, Y.; Chao, F.; and Ji, R. 2023. Cf-vit: A general coarse-to-fine method for vision transformer. In *Proc. of AAAI*, 1–13.

Courbariaux, M.; Bengio, Y.; and David, J.-P. 2015. Binaryconnect: Training deep neural networks with binary weights during propagations. In *Proc. of NeurIPS*, 3123–3131.

Denil, M.; Shakibi, B.; Dinh, L.; Ranzato, M.; and De Freitas, N. 2013. Predicting parameters in deep learning. In *Proc. of NeurIPS*, 2148–2156.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

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 *Proc. of ICLR*, 1–22.

Esser, S. K.; McKinstry, J. L.; Bablani, D.; Appuswamy, R.; and Modha, D. S. 2019. Learned step size quantization. *arXiv preprint arXiv:1902.08153*.

Hao, Z.; Guo, J.; Jia, D.; Han, K.; Tang, Y.; Zhang, C.; Hu, H.; and Wang, Y. 2021. Learning Efficient Vision Transformers via Fine-Grained Manifold Distillation. In *Proc. of NeurIPS*, 1–11.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2022. Masked autoencoders are scalable vision learners. In *Proc. of CVPR*, 16000–16009.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask r-cnn. In *Proc. of ICCV*, 2961–2969.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *Proc. of NeurIPS*, 1097–1105.

Li, Y.; Xu, S.; Cao, X.; Sun, X.; and Zhang, B. 2023a. Q-DM: An Efficient Low-bit Quantized Diffusion Model. In *Proc. of NeurIPS*, 1–12.

Li, Y.; Xu, S.; Cao, X.; Zhuo, L.; Zhang, B.; Wang, T.; and Guo, G. 2023b. DCP–NAS: Discrepant Child–Parent Neural Architecture Search for 1-bit CNNs. *International Journal of Computer Vision*, 131(11): 2793–2815.

Li, Y.; Xu, S.; Lin, M.; Yin, J.; Zhang, B.; and Cao, X. 2023c. Representation Disparity-aware Distillation for 3D Object Detection. In *Proc. of ICCV*, 6715–6724.

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 *Proc. of NeurIPS*, 1–12.

Lin, Y.; Zhang, T.; Sun, P.; Li, Z.; and Zhou, S. 2022. FQ-ViT: Fully Quantized Vision Transformer without Retraining. In *Proc. of IJCAI*, 1173–1179.

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 *Proc. of ICCV*, 10012–10022.

Liu, Z.; Shen, Z.; Savvides, M.; and Cheng, K.-T. 2020. ReActNet: Towards Precise Binary Neural Network with Generalized Activation Functions. In *Proc. of ECCV*, 143–159.

Liu, Z.; Wang, Y.; Han, K.; Zhang, W.; Ma, S.; and Gao, W. 2021b. Post-training quantization for vision transformer. In *Proc. of NeurIPS*.

Liu, Z.; Wu, B.; Luo, W.; Yang, X.; Liu, W.; and Cheng, K.-T. 2018. Bi-real net: Enhancing the performance of 1-bit cnns with improved representational capability and advanced training algorithm. In *Proc. of ECCV*, 722–737.

Qin, H.; Ding, Y.; Zhang, M.; Yan, Q.; Liu, A.; Dang, Q.; Liu, Z.; and Liu, X. 2022. BiBERT: Accurate Fully Binarized BERT. In *Proc. of ICLR*, 1–24.

Rao, Y.; Zhao, W.; Liu, B.; Lu, J.; Zhou, J.; and Hsieh, C.-J. 2021. Dynamicvit: Efficient vision transformers with dynamic token sparsification. *Proc. of NeurIPS*, 1–14.

Rastegari, M.; Ordonez, V.; Redmon, J.; and Farhadi, A. 2016. Xnor-net: Imagenet classification using binary convolutional neural networks. In *Proc. of ECCV*, 525–542.

Tian, Y.; Xie, L.; Wang, Z.; Wei, L.; Zhang, X.; Jiao, J.; Wang, Y.; Tian, Q.; and Ye, Q. 2022. Integrally Pre-Trained Transformer Pyramid Networks. *arXiv preprint arXiv:2211.12735*.

Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; and Jégou, H. 2021. Training data-efficient image transformers & distillation through attention. In *Proc. of ICML*, 10347–10357.

Touvron, H.; Cord, M.; and Jégou, H. 2022. Deit iii: Revenge of the vit. In *Proc. of ECCV*, 516–533.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Proc. of NeurIPS*, 1–11.

Wang, W.; Xie, E.; Li, X.; Fan, D.-P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; and Shao, L. 2021. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proc. of ICCV*, 568–578.

Xu, S.; Li, Y.; Lin, M.; Gao, P.; Guo, G.; Lü, J.; and Zhang, B. 2023a. Q-DETR: An Efficient Low-Bit Quantized Detection Transformer. In *Proc. of CVPR*, 3842–3851.

Xu, S.; Li, Y.; Ma, T.; Lin, M.; Dong, H.; Zhang, B.; Gao, P.; and Lu, J. 2023b. Resilient binary neural network. In *Proc. of AAAI*, 10620–10628.

Xu, S.; Li, Y.; Wang, T.; Ma, T.; Zhang, B.; Gao, P.; Qiao, Y.; Lü, J.; and Guo, G. 2022a. Recurrent bilinear optimization for binary neural networks. In *Proc. of ECCV*, 19–35.

Xu, S.; Li, Y.; Zeng, B.; Ma, T.; Zhang, B.; Cao, X.; Gao, P.; and Lu, J. 2022b. IDa-Det: An Information Discrepancy-aware Distillation for 1-bit Detectors. In *Proc. of ECCV*, 346–361.

Xu, Y.; Zhang, Z.; Zhang, M.; Sheng, K.; Li, K.; Dong, W.; Zhang, L.; Xu, C.; and Sun, X. 2022c. Evo-vit: Slow-fast token evolution for dynamic vision transformer. In *Proc. of AAAI*, 2964–2972.

Yang, H.; Yin, H.; Molchanov, P.; Li, H.; and Kautz, J. 2021. Nvit: Vision transformer compression and parameter redistribution. *arXiv preprint arXiv:2110.04869*.

You, Y.; Li, J.; Reddi, S.; Hseu, J.; Kumar, S.; Bhojanapalli, S.; Song, X.; Demmel, J.; Keutzer, K.; and Hsieh, C.-J. 2020.

Large batch optimization for deep learning: Training bert in 76 minutes. *Proc. of ICLR*, 1–37.

Zhong, Y.; Lin, M.; Chen, M.; Li, K.; Shen, Y.; Chao, F.; Wu, Y.; and Ji, R. 2022. Fine-grained data distribution alignment for post-training quantization. In *Proc. of ECCV*, 70–86.