

QUARK: Quantization-Enabled Circuit Sharing for Transformer Acceleration by Exploiting Common Patterns in Nonlinear Operations

Zhixiong Zhao^{3,†}, Haomin Li^{1,†}, Fangxin Liu^{1,2,*}, Yuncheng Lu³, Zongwu Wang^{1,2}, Tao Yang^{4*}
Li Jiang^{1,2}, Haibing Guan¹

1. School of Computer Science, Shanghai Jiao Tong University, 2. Shanghai Qi Zhi Institute
3. Nanyang Technological University, 4. Huawei Technologies Co., Ltd

*Corresponding Author zhixiong003@e.ntu.edu.sg, {liufangxin, ljiang_cs}@sjtu.edu.cn

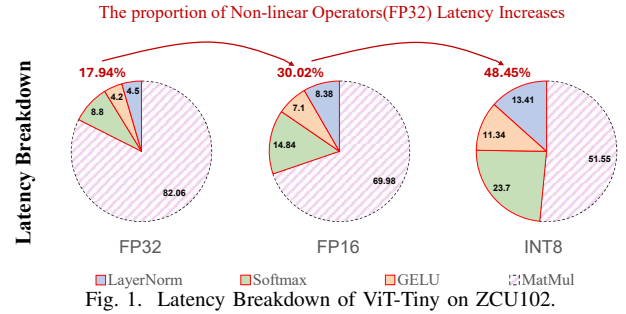
Abstract—Transformer-based models have revolutionized computer vision (CV) and natural language processing (NLP) by achieving state-of-the-art performance across a range of benchmarks. However, nonlinear operations in models significantly contribute to inference latency, presenting unique challenges for efficient hardware acceleration. To this end, we propose QUARK, a quantization-enabled FPGA acceleration framework that leverages common patterns in nonlinear operations to enable efficient circuit sharing, thereby reducing hardware resource requirements. QUARK targets all nonlinear operations within Transformer-based models, achieving high-performance approximation through a novel circuit-sharing design tailored to accelerate these operations. Our evaluation demonstrates that QUARK significantly reduces the computational overhead of nonlinear operators in mainstream Transformer architectures, achieving up to a 1.96× end-to-end speedup over GPU implementations. Moreover, QUARK lowers the hardware overhead of nonlinear modules by more than 50% compared to prior approaches, all while maintaining high model accuracy—and even substantially boosting accuracy under ultra-low-bit quantization.

I. INTRODUCTION

In recent years, Transformer networks, have garnered significant attention in the field of Artificial Intelligence (AI), achieving remarkable results not only in tasks of Natural Language Processing (NLP) but also in Computer Vision (CV) through the Vision Transformer (ViT) [1]. However, the traditional Transformer architecture heavily relies on the attention mechanism, which considerably increases both parameter count and computational complexity compared to earlier neural network models such as Convolutional Neural Networks (CNNs) or Recurrent Neural Networks (RNNs). (e.g., ViT-L contains 307 million parameters and requires 190.7 billion floating-point operations (FLOP) [2]).

Although architectural optimizations for models have been explored [3], most transformer compression efforts focus on

This work was done when Zhixiong Zhao was an intern at Shanghai Jiao Tong University. † These authors contributed equally. This work is supported by the National Key Research and Development Program of China (2024YFE0204300), the National Natural Science Foundation of China (Grant No.62402311), and Natural Science Foundation of Shanghai (Grant No.24ZR1433700). Fangxin Liu and Tao Yang are the corresponding authors.



accelerating linear operators, which are conventionally regarded as the primary computational bottleneck in floating-point models [4], [5]. In contrast, nonlinear operators have received limited attention despite their mathematical complexity and significant challenges in hardware acceleration, particularly as linear operators are quantized from 32-bit floating-point (FP32) to lower-bit integer arithmetic (e.g., 8-bit integer (INT8)). As illustrated in Figure 1, the latency contribution of nonlinear operators increases dramatically. These nonlinear operations are increasingly becoming critical bottlenecks in transformer optimization [6].

Recent studies on nonlinear operators primarily focus on optimizing individual or paired operators, overlooking the compounded computational impact of all nonlinear components (see Figure 1). For instance, approaches like Softmax [7] and NN-LUT [8] target the Softmax operator exclusively, while FQ-ViT [9] and SOLE [10] jointly optimize Softmax and LayerNorm. Although methods such as I-BERT [4] and I-ViT [11] achieve full integer quantization for entire models, and they rely on quantization-aware training (QAT), requiring model retraining or fine-tuning to preserve post-quantization performance. Notably, as model scales increase, the associated retraining costs become prohibitively expensive. Consequently, striking an optimal balance between algorithmic efficiency and hardware constraints remains a critical challenge, particularly given the escalating size of transformer models.

We propose a novel hardware/software co-design framework QUARK to address two critical challenges: 1) Eliminating the reliance on floating-point operations in nonlinear operators

through mathematical reformulation; 2) Reducing quantization errors caused by heterogeneity in activation distributions. Our key innovation lies in a shared sub-operator approximation framework, which enables joint optimization of common non-linear functions such as Softmax, GELU, and LayerNorm. For Softmax, we propose an improved log-sum-exp algorithm [12], transforming computations from base- e to base-2. This enables hardware implementation using shifters and adders, eliminating multiplication, division, and look-up tables (LUTs), and achieving an efficient 8-bit cache design. For GELU, we reformulate its computation as a combination of Softmax and ReLU [13], enabling hardware reuse of the Softmax structure alongside a simple ReLU circuit. For LayerNorm, we simplify its two data accesses to one, approximate the square root using an iterative method [14], and transform the division operation logarithmically. To address the challenges posed by diverse activation distributions, we propose a reordering-based group quantization scheme — a hardware-aware method that clusters channels during offline calibration based on distribution similarity. This approach adapts effectively to the unique characteristics of each layer, such as the power-law distribution in Softmax outputs, asymmetric activation in GELU, and inter-channel variation in LayerNorm. Group quantization scales are derived by shifting a base scale, enabling efficient cross-group alignment via parallel multipliers without requiring any floating-point conversions. In summary, our contributions are as follows:

- **Integer-Only Nonlinear Approximation:** We reformulate exponential, logarithmic, and division operations into low-cost shift-and-add arithmetic, eliminating the need for costly floating-point operations or large LUT-based approximations.
- **Sub-Operator Sharing with Time-Division Multiplexing:** By exploiting common sub-operators (e.g., exponent, log) across Softmax, GELU, and LayerNorm, QUARK unifies them into a single reusable hardware block, significantly reducing resource usage and power consumption.
- **Reorder-Based Group Quantization:** We propose a novel group quantization mechanism that leverages offline reordering and scaled integer alignment to effectively handle diverse output distributions. This approach prevents significant degradation in accuracy under ultra-low-bit settings and simplifies hardware design for multi-group data alignment.

II. RELATED WORK

a) Optimization of Nonlinear Operators: Prior works primarily target Softmax optimization. Approaches like A3 [15] and NN-LUT [8] leverage LUTs or piecewise linear fitting to approximate exponentials but rely on costly division units. SpAtten [16] and ELSA [17] perform quantized Softmax via floating-point units, which are underutilized and inefficient. CORDIC-based designs [18] improve hardware efficiency but sacrifice accuracy. I-BERT [4] introduces a second-order integer-only GELU approximation, reducing arithmetic complexity compared to floating-point-based error functions and

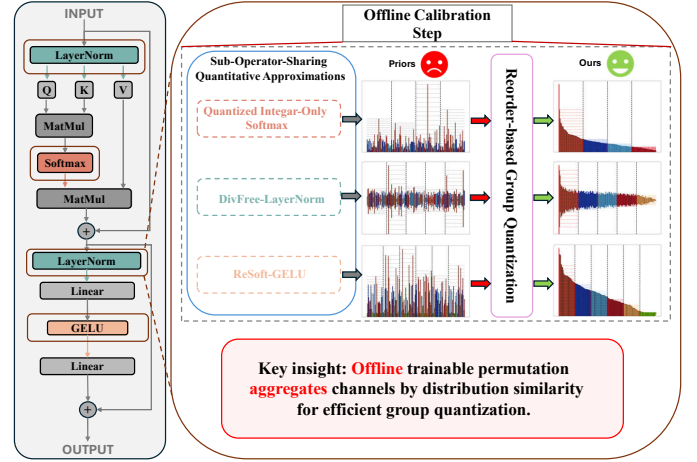


Fig. 2. Overview of QUARK Software.

log-sum-exp methods [19]. However, it still requires 32-bit intermediate caching, incurring high memory overhead. SpAtten also trims input sizes in LayerNorm for performance gains. I-ViT [11] adopts shift-adder-based integer approximations but still depends on expensive divisions and retraining to preserve accuracy.

b) Model Quantization: Quantization has emerged as a crucial technique for reducing model size and computational requirements in LLMs. Current approaches primarily fall into two categories: QAT and PTQ. QAT adapts models to quantization noise through retraining, while PTQ directly converts FP32 models to low-bit formats without requiring training data, offering greater practicality for Transformers. However, ultra-low-bit (≤ 4 -bit) quantization remains challenging due to nonlinear operator heterogeneity. Existing methods like Qbert [20] and VS-Quant [21] apply uniform channel grouping, neglecting inter-channel dynamic range variations and introducing large errors. PEG [22] addresses global statistical divergence but ignores local nonlinear behavior. Quantformer [23] explores search-based grouping in QAT but suffers from a high overhead design, limiting deployment efficiency.

III. QUARK ALGORITHM

In QUARK, we introduce a novel sub-operator-sharing approximation framework specifically designed for three key nonlinear operators in Transformers, enabling efficient deployment of quantized models (Figure 2). First, we detail our sub-operator-sharing approximation methodology. Subsequently, we analyze the resulting activation distributions and derive our reordering-based group quantization approach.

A. Sub-Operator-Sharing Quantitative Approximations

Quantized Integer-Only Softmax. The Softmax operator is widely used as a normalization function in Transformers and can be defined as follows:

$$\text{Softmax}(x_i) = \frac{\exp(x_i)}{\sum_{j=1}^n \exp(x_j)} \quad (1)$$

Due to the nonlinear nature of Softmax, it cannot be directly quantized like linear operators, and integer-only logic typically

does not support exponential arithmetic. Additionally, overflow may occur. To this end, we restrict the range of the exponential operation and eliminate division by using approximate logarithmic and exponential functions, defined as follows:

$$\begin{aligned} \text{Softmax}(x_i) &= \exp\{x_i - \ln[\sum_{j=1}^n \exp(x_j)]\} \\ &= \exp\{(x_i - x_{\max}) - \ln[\sum_{j=1}^n \exp(x_j - x_{\max})]\} \end{aligned} \quad (2)$$

where $x_{\max} = \max\{x_1, x_2, \dots, x_n\}$ and $x_i' = x_i - x_{\max}$. The term x_i' simplifies the expression and is always non-positive. For the exponential operation, we express $\log_2 e$ as a binary approximation, $(1.0111)_b$, to convert from base e to base 2, using only shifters and adders. This reduces hardware complexity. Since the exponential term of 2 may contain a fractional part, we separate it into integer and fractional components:

$$\begin{aligned} \exp(x_i') &= 2^{(x_i' \cdot \log_2 e)} \\ &\approx 2^{[x_i' + (x_i' \gg 1) - (x_i' \gg 4)]} = 2^{q_I \cdot q_F} \end{aligned} \quad (3)$$

where $q_I = [x_i' + (x_i' \gg 1) - (x_i' \gg 4)]$, represents the integer part, which corresponds to the non-positive exponent of 2 rounded upwards, and $q_F = x_i' + (x_i' \gg 1) - (x_i' \gg 4) - q_I \in (-1, 0]$, represent the fractional part. For low-cost computation, we shift the integer part directly, while the fractional part, within the range of $(-1, 0]$, is approximated using a second-order polynomial as follows:

$$2^{q_I \cdot q_F} = 2^{q_F} \ll q_I \quad (4)$$

$$2^{q_F} \approx 0.1713q_F^2 + 0.6674q_F + 0.998 \quad (5)$$

In the logarithm computation, we convert the natural logarithm to a base-2 logarithm for more efficient computation. Since $\ln 2$ is approximated as $(0.1011)_b$, we can use shifts and additions to convert the logarithm. The logarithm of $Sx_i \cdot I'_{x_i}$ is then approximated as:

$$\begin{aligned} \ln(x_i') &= \ln 2 \cdot \log_2(x_i') \\ &\approx \log_2(x_i') - [\log_2(x_i') \gg 2] - [\log_2(x_i') \gg 4] \end{aligned} \quad (6)$$

The term $\log_2(x_i')$ can be reformulated by expressing x_i' as the product of an integer power of 2 and a positive value within the range $[1, 2)$. Using this decomposition and the mathematical properties of the logarithm function, the original logarithmic expression can be separated into the sum of two terms as follows:

$$x_i' = 2^{q_M} \cdot q_N, \text{ where } q_N \in [1, 2) \quad (7)$$

Then, we can rewrite the logarithmic expression as:

$$\log_2(x_i') = \log_2(2^{q_M} \cdot q_N) = q_M + \log_2(q_N) \quad (8)$$

Here, q_M represents the position of the most significant bit (MSB) of x_i' in binary form, and q_N is the normalized factor.

Algorithm 1: Quantized Integer-Only Softmax.

```

1 Input: Integer input  $x_{in}$ 
2 Output: Integer output  $x_{out}$ 
3 Function Appro-Exp( $x_{in}'$ ):  $\triangleright x_{in}' = x_{in} - x_{in_{\max}}$ 
4    $x_{shift} \leftarrow x + (x \gg 1) - (x \gg 4)$   $\triangleright x \cdot \log_2 e$ 
5    $q_I \leftarrow \lceil x_{shift} \rceil$   $\triangleright$  Integer part
6    $q_F \leftarrow x_{shift} - q_I$   $\triangleright$  fractional part
7    $x_a \leftarrow 0.1713q_F^2 + 0.6674q_F + 0.998$   $\triangleright$ Eq. 5
8    $x_{exp} \leftarrow x_a \ll q_I$   $\triangleright$ Eq. 4
9   return  $x_{exp}$   $\triangleright x_{exp} \approx \exp(x_{in}')$ 
10 end
11 Function Appro-Ln( $x_{in}'$ ):  $\triangleright x_{in}' = x_{in} - x_{in_{\max}}$ 
12    $q_M \leftarrow MSB(\log_2(x_{in}'))$ 
13    $q_N \leftarrow x_{in}' \gg q_M$   $\triangleright$  Eq. 7
14    $x_a \leftarrow -0.3369q_N^2 + 1.995q_N - 1.65$   $\triangleright$  Eq. 9
15    $x_{ln} \leftarrow x_a - (x_a \gg 2) - (x_a \gg 4)$   $\triangleright I \cdot \ln 2$ 
16   return  $x_{ln}$   $\triangleright x_{ln} \approx \ln(x_{in}')$ 
17 end
18 Function IntDiv-Free Softmax( $x_{in}$ ):
19    $x_{in}' \leftarrow x_{in} - \max(x_{in})$ 
20    $x_{exp} \leftarrow \text{Appro-Exp}(x_{in}')$ 
21    $x_{ln} \leftarrow \text{Appro-Ln}(\sum x_{in}')$ 
22    $x_0 \leftarrow x_{exp} - x_{ln}$ 
23    $x_{out} \leftarrow \text{Appro-Exp}(x_0)$   $\triangleright$  Eq. 2
24   return  $x_{out}$   $\triangleright x_{out} \approx \text{Softmax}(x_{in})$ 
25 end

```

Since q_N lies in the range $[1, 2)$, we approximate $\log_2(q_N)$ using a second-order polynomial as follows:

$$\log_2(q_N) \approx -0.3369q_N^2 + 1.995q_N - 1.65 \quad (9)$$

Algorithm 1 summarizes the integer-only Softmax flow.

Quantized Integer-Only GELU with ReLU and Softmax. GELU [24] is a nonlinear activation function used in Transformer models, defined as:

$$\begin{aligned} \text{GELU}(x_i) &= x_i \cdot \frac{1}{2} \left[1 + \text{erf} \left(\frac{x_i}{\sqrt{2}} \right) \right], \\ \text{where } \text{erf}(x_i) &= \frac{2}{\sqrt{\pi}} \int_0^{x_i} \exp(-t^2) dt \end{aligned} \quad (10)$$

Here, erf is the error function, which is computationally inefficient. Based on the findings in [24], the GELU operator can be approximated using a Sigmoid function as follows:

$$\text{GELU}(x_i) \approx x_i \cdot \sigma(1.702x_i) \quad (11)$$

where $\sigma(\cdot)$ is the Sigmoid function. Since the Sigmoid function requires floating-point arithmetic, it is not suitable for integer-only quantization. However, we observe that the Sigmoid function has a form identical to the first output element of the Softmax function when the input is a binary vector. Thus, we can approximate the Sigmoid using a binary Softmax:

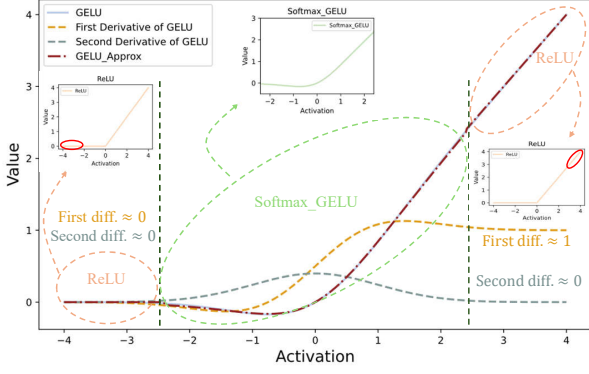


Fig. 3. Characteristics of GELU Operator in Transformers.

$$\begin{aligned} \sigma(1.702x_i) &= \frac{1}{1 + e^{-1.702x_i}} \\ &= \frac{e^0}{e^0 + e^{-1.702x_i}} = \text{Softmax}_2^1([0, -1.702x_i]) \end{aligned} \quad (12)$$

leading to the approximation:

$$\text{GELU}(x_i) \approx 1.702x_i \cdot \text{Softmax}_2^1([0, -1.702x_i]) \quad (13)$$

Here, $(\cdot)_2^1$ represents the first output element with a binary vector as input. However, we observe that this approximation introduces a significant amount of redundant computation when applied to all activation values, as it fails to consider the specific characteristics of the activation values. As shown in figure 3, the second-order derivative of the GELU operator approaches zero for activation values with large absolute magnitudes, making it well-approximated by a linear function. Additionally, the first-order derivative in these regions closely resembles that of the ReLU operator. Based on these observations, we employ the ReLU operator for larger activation values to enhance computational efficiency while preserving accuracy. For the remaining intervals, we retain the approximation algorithm described in Eq. 13. Since this algorithm is applied only to specific intervals, it does not significantly contribute to the overall computational overhead, which ReLU operator as follows:

$$\text{GELU}(x_i) \approx \text{ReLU}(x_i) = \max(0, x_i) \quad (14)$$

For selecting the approximation intervals, the boundary points of the ReLU approximation interval are determined by the activation values where the absolute errors of Eq. 13 and Eq. 14, when compared to Eq. 10 are equal. The ReLU approximation interval is consequently defined as $(-\infty, -2.4] \cup [2.4, \infty)$, ensuring computational efficiency while maintaining accuracy in these regions.

Integer-only LayerNorm without Division LayerNorm is a widely used technique in Transformer models to enhance training stability and model performance [25]. The primary objective of the LayerNorm operator is to normalize the output of each layer to have zero mean and unit variance, thereby mitigating the distributional shifts in the inputs to each layer during training. In the context of Transformers, LayerNorm normalizes the inputs to each layer along the hidden feature

dimension to achieve a unit variance. The LayerNorm operation in Transformer models is formally expressed as follows:

$$\text{LayerNorm}(x_i) = \frac{x_i - \text{Mean}(x_i)}{\sqrt{\text{Var}(x_i)}} \quad (15)$$

where $\text{Mean}(\cdot)$ and $\text{Var}(\cdot)$ represent the mean and variance of the input across the feature dimension.

This process involves several nonlinear operations, including division, squaring, and square roots. Given the rapid rate of change of $\text{Mean}(\cdot)$ and $\text{Var}(\cdot)$ during inference, these values must be computed dynamically at runtime. To minimize data access overhead, we first simplify the variance calculation mathematically, reducing the need for two full data accesses to just one, as follows:

$$\text{Var}(x_i) = \mathbb{E}\{[x_i - \mathbb{E}(x_i)]^2\} = \mathbb{E}(x_i^2) - [\mathbb{E}(x_i)]^2 \quad (16)$$

where $\mathbb{E}(\cdot)$ is the mean value. Previous work [4] optimized the square root operator using an iterative approach; however, it did not address the large number of division operations involved, both in the iterative process and within the LayerNorm operator itself. To overcome this, we optimize the LayerNorm operator by employing shifting and logarithmic division which is the same as Softmax as follows :

$$x_{i+1} = \left(x_i + \left\lfloor \frac{\text{Var}(x)}{x_i} \right\rfloor \right) / 2 = \left(x_i + \left\lfloor \frac{\text{Var}(x)}{x_i} \right\rfloor \right) \gg 1 \quad (17)$$

$$\frac{\text{Var}(x)}{x_i} = \exp(\ln[\frac{\text{Var}(x)}{x_i}]) = \exp(\ln[\text{Var}(x)] - \ln(x_i)) \quad (18)$$

where x_i is the result of the i -th iteration, initialized as $x_0 = 2^{\lfloor \text{bit}(\text{Var}(x))/2 \rfloor}$. This iterative method converges within ten iterations for any INT8 input, avoiding nonlinear operations like division.

B. Activation Distribution after Quantized Nonlinear Approximations

In the previous section, we performed quantization approximation on three nonlinear operators in Transformers. However, as shown in Figure 4, the activation distributions of these approximated nonlinear operators exhibit significant non-uniformity. Directly applying identical quantization parameters across entire layers may introduce substantial quantization errors. Through detailed analysis of post-nonlinear activation distributions: 1) **LayerNorm** outputs show severe inter-channel variance - some channels contain prominent outliers where maximum/minimum values exceed those of other channels by orders of magnitude. While per-channel quantization could mitigate this issue, it introduces prohibitive computational overhead. 2) **Softmax** transforms attention scores into (0,1)-bounded probabilities, yet its outputs follow a power-law distribution. In DeiT-S, 96.63% of values reside below 0.1, while the remaining 3.37% encode crucial token dependencies. 3) **Post-GELU** activations demonstrate asymmetric characteristics: negative values cluster near zero forming a sharp peak ($\mathbb{E}[x] \approx -0.1$, $\text{Var}(x) < 0.01$), whereas positive values follow a heavy-tailed distribution extending towards higher magnitudes ($\mathbb{E}[x] \geq 0.8$, $\text{Var}(x) \propto x^2$).

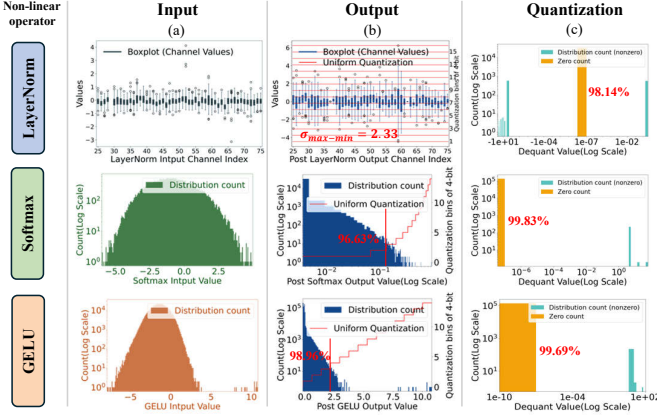


Fig. 4. Visualization of post-nonlinear activations from the 6th layer in DeiT-T, illustrating key Transformer quantization challenges: LayerNorm’s inter-channel variance, Softmax’s zero-collapsing heavy-tailed distribution, and GELU’s asymmetric input range that challenges conventional symmetric quantization schemes.

C. Reorder-based Group Quantization

We propose QUARK, a novel quantization method that employs a reordered grouping quantization scheme specifically designed for post-nonlinear operator activations. This approach adapts to the distinctive characteristics of different nonlinear operators through a unified quantization framework, which performs channel-wise statistical analysis to guide the sorting and grouping of activations. Notably, to eliminate the computational overhead of online reordering, we strategically integrate the reordering process into the model during the offline calibration phase, thereby enabling seamless deployment without runtime penalties.

Group Size Allocation We observe that activations after nonlinear operators in different layers exhibit significant channel-wise distribution divergence, indicating heterogeneous sensitivity to quantization noise across layers. This finding suggests that using a fixed grouping strategy for all layers may lead to suboptimal quantization performance. To address this, we propose optimizing layer-wise group sizes under a bit-wise operations (BOP) constraint to minimize the distribution discrepancy between the quantized and full-precision models. Formally, given a target BOP budget N_{bop} , we formulate the group allocation problem as an integer linear programming (ILP) optimization:

$$\min_{\{g_l\}} \sum_{l=1}^L D_{\text{KL}}(P_{\text{FP}}(a_l) \| P_Q(a_l; g_l)) \quad \text{s.t.} \quad \sum_{l=1}^L B_l(g_l) \leq N_{\text{bop}}, \quad g_l \in \mathbb{Z}^+, l = 1, 2, \dots, L \quad (19)$$

where L denotes the total number of layers in the network, $g_l \in \mathbb{Z}^+$ is the number of groups used for quantizing activations in layer l , $P_{\text{FP}}(a_l)$ and $P_Q(a_l; g_l)$ represent the full-precision and quantized activation distributions of layer l , respectively, and $B_l(g_l) = C_l \cdot b \cdot \log_2 g_l$ models the BOP cost for layer l when using g_l groups, with C_l being the number of channels in layer l and b the quantization bit-width. The

constraint ensures that the total BOP across all layers does not exceed the target budget N_{bop} .

Offline Reordering To enable group quantization by clustering channels with similar statistical characteristics, we first need to reorder activation channels based on their statistical properties. However, performing such reordering during inference introduces additional latency, especially in large-scale models with a large number of channels. Moreover, storing both the original and reordered activation tensors leads to increased memory overhead. To minimize the cost of reordering, we introduce a trainable permutation matrix R , constrained to be orthogonal, and fuse it into adjacent operations (e.g., linear layers or LayerNorm) during the offline calibration stage. To ensure numerical equivalence, we invert the permutation matrix and absorb it into the weight matrix of the current layer. The transformation is expressed as:

$$Y = (X \cdot R) \cdot (R^{-1} \cdot W) = X \cdot W \quad (20)$$

where R is a diagonal matrix and R^{-1} denotes its element-wise inverse along the diagonal. Since the input X typically originates from a preceding linear operation (such as a linear layer or LayerNorm), the permutation matrix R can be fused offline into the parameters of the preceding layer. This design ensures that no significant computational overhead is introduced during inference. As an example, consider reordering the activations after the Softmax operation in the attention mechanism. In this case, the reordering matrix R_{vo} can be fused into the adjacent projection matrix W_V , while its inverse transpose can be absorbed into the subsequent output projection matrix W_O . The output of the attention block becomes:

$$Y_O = [(A_{\text{tnn}} \cdot V) W_V R_{vo} + b_V R_{vo}] R_{vo}^T W_O + b_O \quad (21)$$

Intra-group Quantization and Cross-group Alignment

We divide the activations into groups $G_1, G_2, \dots, G_n$ according to Equation (19). Within each group, normal elements are quantized using a uniform scalar quantizer, while outliers are clamped or frozen to preserve critical information. For a normal element $x \in G_1$, the quantized value is computed as:

$$x_q = \left\lfloor \frac{x}{\Delta_{R_1}} + \frac{1}{2} \right\rfloor \quad (22)$$

For outliers $y \in G_1$ that exceed the threshold, the quantized value is mapped directly to the quantization boundary:

$$y_q = \text{sign}(y) \times Q_{\text{max}} \quad (23)$$

The complete quantization rule can be described as:

$$z_q = \begin{cases} \left\lfloor \frac{z}{\Delta_{R_1}} + \frac{1}{2} \right\rfloor & \text{if } |z| \leq \text{threshold} \\ \text{sign}(z) \times Q_{\text{max}} & \text{if } |z| > \text{threshold} \end{cases} \quad (24)$$

This approach preserves important outlier information while enabling efficient quantization of the remaining values. To support hardware-friendly processing across different groups, we assign a group-wise quantization step:

$$\Delta_{R_i} = 2^{k_i} \cdot \Delta_{R_1}, \quad i = 1, 2, \dots, n \quad (25)$$

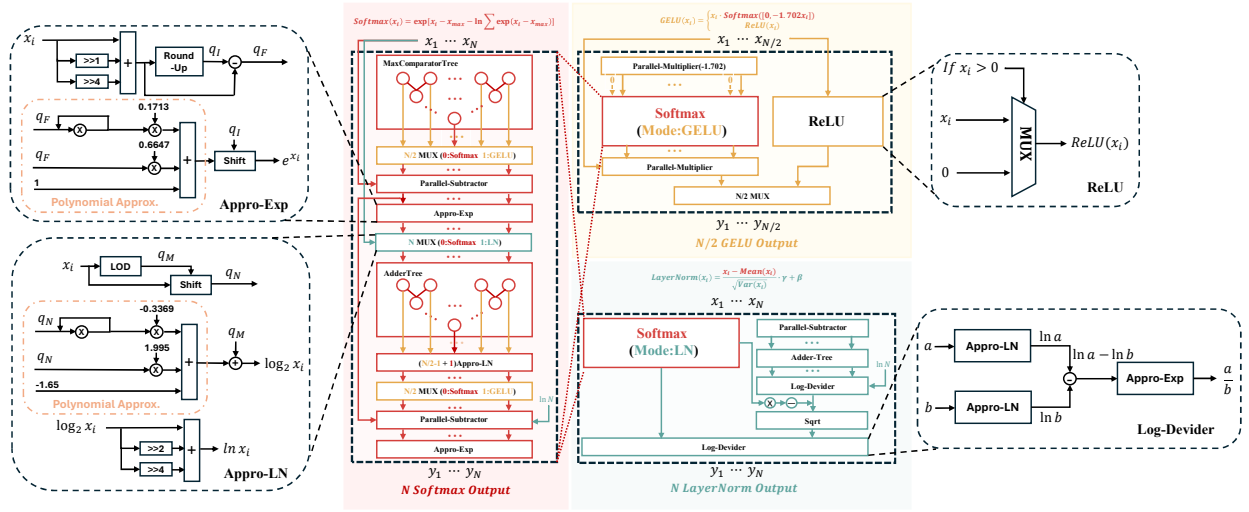


Fig. 5. The shared hardware components and reuse pathways among Softmax, GELU, and LayerNorm.

where k_i is the scaling factor for group G_i , and Δ_{R_1} is the reference step size for group G_1 . This formulation simplifies alignment during matrix multiplications. For example, aligning quantized values $a_q \in G_1$ and $b_q \in G_2$ results in:

$$a_q \cdot \Delta_{R_1} + b_q \cdot \Delta_{R_2} = a_q + b_q \cdot 2^{k_2 - k_1} \cdot \Delta_{R_1} \quad (26)$$

Extending this to n groups, the generalized alignment is:

$$a_q \cdot \Delta_{R_1} + \sum_{i=2}^n b_{q_i} \cdot \Delta_{R_i} = (a_q + \sum_{i=2}^n b_{q_i} \cdot 2^{k_i - k_1}) \cdot \Delta_{R_1} \quad (27)$$

Here, b_{q_i} is a quantized value from group G_i , and k_i is its corresponding scaling factor. This strategy ensures efficient cross-group alignment with minimal computational overhead.

IV. QUARK HARDWARE

QUARK is a pluggable nonlinear unit that connects to the PE array via a shared buffer, making it compatible with existing accelerators and programmable platforms (e.g., FPGAs). This design enables fast Transformer inference while preserving hardware flexibility.

A. Sub-Operator Sharing Unit

QUARK improves efficiency and reduces hardware cost by enabling sub-operator reuse across nonlinear functions such as Softmax, GELU, and LayerNorm. A key innovation is the reformulation of GELU into a binary Softmax-like structure, allowing both to share an exponential computation module, as illustrated in Figure 5. To minimize computational overhead, complex functions (e.g., exp, log, $\div$) are approximated using lightweight adder- and shifter-based logic. This enables a unified arithmetic backend based on adders and shift registers to support multiple nonlinear operators.

Given that these functions are invoked at different stages of the Transformer pipeline—e.g., attention, activation, normalization—QUARK applies time-division multiplexing (TDM). TDM reuses hardware units sequentially across pipeline stages, minimizing area and energy overhead without performance compromise. Figure 5 summarizes the shared arithmetic paths across operators enabled by this design.

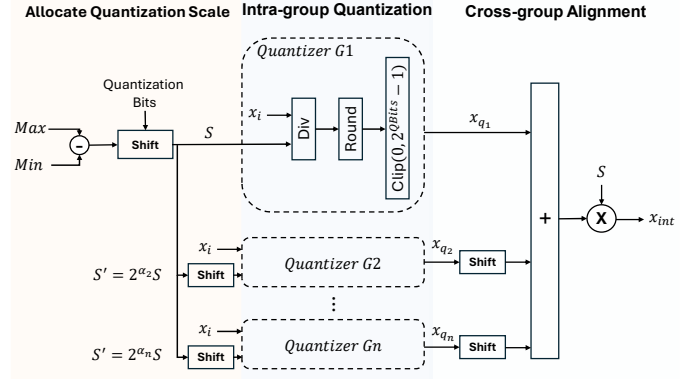


Fig. 6. Architecture of Group Quantization Unit.

B. Group Quantization Unit

The Group Quantization Unit supports efficient processing of post-activation values by implementing a three-stage, group-wise quantization architecture. Based on offline calibration, the unit first selects a base group in the Quantization Scale Allocation stage, deriving its reference scale from statistical features, while estimating scales for other groups via shift-based approximations. In the Intra-group Quantization stage, channels within each group—having similar value distributions—share integer-only quantization parameters. Dynamic bit-width adjustment is applied to balance quantization precision and hardware cost. Finally, in the Cross-group Alignment stage, to support matrix accumulation across groups with different quantization scales, inverse shifting is applied to non-base groups, aligning all values using bit-shifting alone.

As illustrated in Figure 6, this hardware-friendly design avoids floating-point operations, relying instead on simple shift and add primitives. The architecture supports parallel group processing and synergizes with other modules in the QUARK framework to ensure low energy consumption and high throughput without sacrificing quantization accuracy.

V. EXPERIMENTS

A. Experiment Setting

Software Setup To validate our algorithm, we conducted experiments on both CV and NLP tasks. For CV tasks, we

TABLE I

COMPARISON WITH STATE-OF-THE-ART (SOTA) METHODS ON IMAGENET. OUR METHOD ACHIEVES MINIMAL ACCURACY DROP WHILE MEETING BOTH CONSTRAINTS OF NO RETRAINING AND INTEGER-ONLY INFERENCE. NOTABLY, UNDER ULTRA-LOW BIT-WIDTH SETTINGS (4/4), IT SIGNIFICANTLY OUTPERFORMS EXISTING METHODS IN ACCURACY. ("PREC. (W/A)" INDICATES THE BIT-WIDTHS FOR WEIGHTS AND ACTIVATIONS, RESPECTIVELY.)

Method	Prec.(W/A)	Training-free	Int.-only	ViT-S	ViT-B	DeiT-T	DeiT-S	DeiT-B	Swin-T	Swin-S	Swin-B
Full-Precision	32/32	-	-	81.39	84.54	72.21	79.85	81.80	81.35	83.23	85.27
I-BERT [4]	8/8	×	✓	80.47	83.70	71.33	79.11	80.79	80.15	81.86	-
I-ViT [11]	8/8	×	✓	81.27	84.76	72.24	80.12	81.74	81.50	83.01	-
FQ-ViT [26]	8/8	✓	×	78.68	82.76	70.92	78.44	81.12	79.97	82.38	82.53
PTQ4ViT [27]	8/8	✓	×	81.00	84.25	71.72	79.47	81.48	81.23	83.10	85.14
RepQ-ViT [28]	8/8	✓	×	81.19	84.39	72.03	79.68	81.77	81.19	83.05	85.12
QUARK(ours)	8/8	✓	✓	80.72	83.78	71.29	79.40	81.52	81.06	82.81	85.03
FQ-ViT [26]	6/6	✓	×	4.26	0.10	58.66	45.51	64.63	48.63	66.50	52.09
PTQ4ViT [27]	6/6	✓	×	78.63	81.65	69.68	76.28	80.25	80.47	82.38	84.01
RepQ-ViT [28]	6/6	✓	×	80.43	83.62	70.76	78.90	81.27	80.89	82.79	84.57
QUARK(ours)	6/6	✓	✓	80.12	83.28	70.29	78.60	81.42	81.08	82.93	84.78
FQ-ViT [26]	4/4	✓	×	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10
PTQ4ViT [27]	4/4	✓	×	42.57	30.69	36.96	34.08	64.39	73.48	76.09	74.02
RepQ-ViT [28]	4/4	✓	×	65.05	68.48	57.43	69.03	75.61	72.39	79.45	78.32
QUARK(ours)	4/4	✓	✓	68.84	74.56	60.68	72.83	78.76	77.85	81.44	83.12

TABLE II
PTQ PERFORMANCE ON GLUE BENCHMARK.

Method	Prec. (W/A)	CoLA (Matt.)	MNLI (acc/mm)	MRPC (f1/acc)	QNLI (acc)	QQP (f1/acc)	RTE (acc)	SST-2 (acc)	STS-B (Pear./Spear.)	Avg.
BERT	32-32	59.60	84.94/84.76	91.35/87.75	91.84	87.82/90.91	72.56	93.35	89.70/89.22	83.83
Percentile [29]	6-6	37.32	72.40/71.69	85.09/79.90	79.37	72.58/80.19	61.73	87.27	86.38/87.29	72.93
EasyQuant [30]	6-6	38.16	75.82/75.66	82.51/77.45	84.94	75.31/81.81	65.34	87.27	85.50/86.33	74.49
Outlier Sup. [31]	6-6	54.40	82.02/81.69	87.45/83.33	89.82	84.69/88.94	70.76	91.86	88.65/88.55	81.19
QUARK(ours)	6-6	61.41	82.78/81.61	88.19/83.56	90.20	86.88/89.83	70.15	91.09	88.38/87.50	82.25
Percentile [29]	4-4	-4.15	32.73/32.95	81.22/68.38	50.36	23.16/60.71	47.29	50.92	-9.18/-8.63	35.64
Outlier Sup. [31]	4-4	-1.63	27.09/26.55	68.00/60.78	59.64	23.34/65.87	52.71	70.41	0.48/-0.29	39.50
QUARK(ours)	4-4	43.76	39.15/40.64	80.43/67.12	67.08	35.99/72.76	51.77	84.39	63.67/64.82	59.91
RoBERTa	32-32	62.50	87.75/87.23	93.10/90.44	92.68	88.78/91.60	80.51	95.18	91.04/90.72	86.40
Percentile [29]	6-6	20.73	72.23/73.68	84.43/78.43	77.16	82.21/87.44	62.82	88.19	79.41/79.64	70.98
EasyQuant [30]	6-6	9.28	74.96/75.87	84.31/76.47	74.04	85.52/89.12	62.45	89.56	80.89/82.38	70.01
Outlier Sup. [31]	6-6	46.35	83.38/83.32	87.50/83.33	86.82	86.86/90.01	67.51	92.20	86.83/86.93	79.62
QUARK(ours)	6-6	49.49	83.33/82.68	87.65/83.42	86.98	86.34/89.03	78.75	92.72	90.02/89.92	81.77
Percentile [29]	4-4	0.01	31.82/31.72	77.22/63.38	50.54	38.15/52.92	47.29	50.92	-1.52/-0.99	36.89
Outlier Sup. [31]	4-4	-3.24	32.75/32.37	76.76/63.84	49.5	15.75/60.36	52.35	51.15	-3.15/-2.52	35.98
QUARK(ours)	4-4	10.49	36.26/37.33	75.44/66.72	52.45	40.75/58.39	73.95	90.28	88.71/87.03	59.02

evaluated image classification on the ImageNet [32] dataset with different model variants, including ViT [1], DeiT [33], and Swin Transformer [34]. For NLP tasks, we performed experiments on the GLUE [35] benchmark using BERT-Base [36] and RoBERTa-Base [37] models. To ensure fair comparisons with prior works, we randomly selected 32 samples from the ImageNet dataset for image classification tasks and 256 samples in NLP tasks to calibrate quantization parameters.

Hardware Setup To evaluate the impact of the proposed method on hardware overhead, we synthesized and implemented the approach on a ZCU102 board using Vivado. At a clock frequency of 300 MHz, the hardware resource utilization was obtained from the place-and-route results. The results were then compared with existing FPGA implementations.

B. Accuracy Results

a) Image Classification on ImageNet: As shown in Table I, our method outperforms I-BERT under W8A8 quantization and nearly matches I-ViT (< 1% accuracy loss), despite I-ViT requiring costly retraining. For W6A6, our approach surpasses FQ-ViT and PTQ4ViT and performs comparably to RepQ-ViT. Unlike PTQ methods, our integer-only framework eliminates

hardware challenges. For W4A4, we set a new state-of-the-art, achieving a 6.08% accuracy gain on ViT-B and > 3% average improvement across various networks.

b) Language Understanding on GLUE: Table II summarizes our results on the GLUE benchmark. At 6-bit quantization, QUARK achieves significant improvements, e.g., 7.01% gain for BERT on CoLA. At 4-bit, QUARK outperforms prior methods, improving BERT’s average GLUE score by 20.41% and RoBERTa’s by 23.04%. These results demonstrate the robustness of QUARK for language tasks.

c) Ablation study: We evaluate the effectiveness of nonlinear operator approximation and reordering-based group quantization across various bit-widths, using linear-layer-only quantization as the baseline. As shown in Figure 7, individual nonlinear operator approximation causes minimal accuracy loss, with an average drop below 0.2% for W8A8 and W6A6 quantization. When applying three approximations simultaneously, the accuracy loss remains at 0.26% for W8A8 and 0.42% for W6A6, and only 0.4% for W4A4 quantization on the DeiT-Small model. Notably, our reordering-based group quantization not only compensates for approximation errors

TABLE III
PERFORMANCE COMPARISON WITH PREVIOUS WORKS.

Method	FPGA Platform	Model	Prec. (W/A)	DSP Utilization	LUT Utilization	Freq.	Latency	Throughput
ViA [38]	Alveo U50	Swin-T	F16	2420/5952(40.6%)	258/872(29.6%)	300MHz	-	309.6GOPs
Auto-ViT-Acc [39]	ZCU102	DeiT-S	W8/4 A8	1552/2520(61.6%)	185/274(67.5%)	150MHz	-	907.8GOPs
TCAS-I'23 [40]	ZCU102	ViT-S	W8A8	1268/2520(50.3%)	144/274(52.7%)	300MHz	11.15ms	762.7GOPs
QUARK(ours)	ZCU102	ViT-S	W8A8	1181/2520(46.9%)	130/274(47.4%)	300MHz	10.68ms	787.5GOPs
HotaQ [41]	ZCU102	ViT-S	W4A4	1933/2520(76.7%)	157/274(57.3%)	200MHz	7.59ms	733.7GOPs
QUARK(ours)	ZCU102	ViT-S	W4A4	1056/2520(41.9%)	147/274(53.6%)	300MHz	6.59ms	864.7GOPs

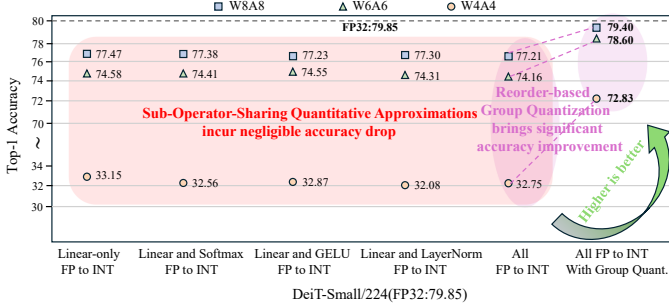


Fig. 7. The impact of individual nonlinear operators (Softmax, GELU, and LayerNorm) and the Reorder-based Group Quantization on model accuracy across different bit configurations (W8A8, W6A6, W4A4).

but also significantly enhances quantization precision. Compared to full-precision models, it achieves only 0.45% accuracy loss under W8A8 quantization and delivers over 40% accuracy improvement in ultra-low-bit quantization (W4A4).

C. Hardware Evaluation

a) Overall Performance: Table III shows significant advancements over SOTA approaches. Our design achieves a peak throughput of 787.5 GOP/s with minimal hardware overhead, outperforming existing solutions across multiple dimensions. Compared to HotaQ [41], which employs hardware-oriented token adaptive quantization for embedded FPGAs, our approach delivers superior network accuracy despite maintaining higher hardware efficiency. While ViA [38] optimizes data locality through ViT structure analysis and partitioning strategies, its FP16-based implementation achieves only 309.6 GOP/s, significantly lower than our design. Auto-ViT-Acc [39] surpasses our throughput through mixed-precision quantization including PoT formats, but this approach compromises accuracy and introduces substantial training overhead.

b) Nonlinear Operators Resource Utilization: As shown in Table IV, our design achieves significant hardware resource savings compared to conventional solutions, reducing LUT utilization to 20%-80% and FF consumption by approximately 70%. Traditional implementations rely on memory-intensive LUT methods or extensive division operations. In contrast, our design leverages the TDM features of nonlinear operators across Vision Transformer stages to enable dynamic resource sharing. Specifically, we introduce a time-multiplexed shared circuit that dynamically reuses computational resources across processing stages, eliminating the hardware redundancy incurred by dedicated circuits for individual nonlinear operators.

c) Latency Comparison: Figure 8 (a) shows QUARK's performance on the ZCU102 FPGA platform compared to

TABLE IV
COMPARISON OF HARDWARE OVERHEAD WITH SOTA FPGA IMPLEMENTATIONS. * INDICATES THE RESOURCE CONSUMPTION OF NON-SHARED COMPONENTS, EXCLUDING SHARED RESOURCES IN NONLINEAR OPERATORS.

Method	Platform	Hardware Resource	Resource Utilization		
			Softmax	GELU	LayerNorm
Swat [42]	Alveo U50	LUT	135,093	11,219	10,731
		FF	-	-	-
		DSP	98	196	0
ViA [38]	Alveo U50	LUT	32,783	3,325	2,817
		FF	32,383	3,161	2,145
		DSP	131	20	7
SOCC'20 [43]	Virtex UltraScale + VXU13P	LUT	21,190	-	10,551
		FF	32,623	-	5,325
		DSP	0	-	129
TCAS-I'23 [40]	ZCU102	LUT	22,865	10,163	10,558
		FF	21,770	5,992	4,038
		DSP	128	32	7
QUARK(ours)	ZCU102	LUT	21,804	2,772*	4,959*
		FF	8,184	1,815*	1,015*
		DSP	144	0*	3*

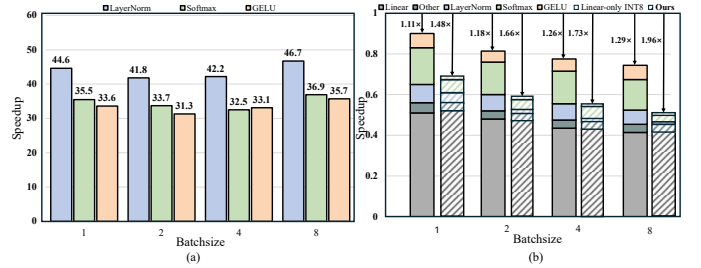


Fig. 8. (a) Speedup over GPU on all nonlinear operators. (b) End-to-end speedup and latency breakdown, the latency is normalized with respect to FP32 implementation.

RTX A5000 GPU, evaluating all nonlinear operators in DeiT-Tiny (batch sizes 1-8, token length 197). QUARK achieves substantial speedups of 41.8x-46.7x for LayerNorm, 32.5x-36.9x for Softmax, and 31.3x-35.7x for GELU, enabled by hardware-friendly approximations and optimized pipelined datapath designs. Figure 8 (b) reveals that while INT8 models with linear layer-only quantization show limited GPU speedups (1.11x-1.29x) due to dominant nonlinear operations, QUARK's nonlinear operator optimization enables remarkable 1.48x-1.96x end-to-end speedup.

VI. CONCLUSION

This work presents QUARK, a software/hardware co-design solution for accelerating quantized Transformer-based models on FPGA platforms. By integrating lightweight integer-only arithmetic modules and a reorder-based group quantization. Evaluations show that QUARK delivers state-of-the-art efficiency gains for vision and language Transformers under various integer-only inference settings.

REFERENCES

- [1] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," 2021. [Online]. Available: <https://arxiv.org/abs/2010.11929>
- [2] F. Liu, W. Zhao, Z. He, Y. Wang, Z. Wang, C. Dai, X. Liang, and L. Jiang, "Improving neural network efficiency via post-training quantization with adaptive floating-point," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 5261–5270.
- [3] N. P. Jouppi, C. Young, N. Patil, D. Patterson, G. Agrawal, R. Bajwa, S. Bates, S. Bhatia, N. Boden, A. Borchers *et al.*, "In-datacenter performance analysis of a tensor processing unit," in *ISCA*, 2017.
- [4] S. Kim, A. Gholami, Z. Yao, M. W. Mahoney, and K. Keutzer, "I-bert: Integer-only bert quantization," in *International conference on machine learning*. PMLR, 2021, pp. 5506–5518.
- [5] F. Liu, N. Yang, Z. Song, Z. Wang, H. Li, S. Huang, Z. Song, S. Pei, and L. Jiang, "Inspire: Accelerating deep neural networks via hardware-friendly index-pair encoding," in *DAC*, 2024, pp. 1–6.
- [6] F. Liu, N. Yang, H. Li, Z. Wang, Z. Song, S. Pei, and L. Jiang, "Spark: Scalable and precision-aware acceleration of neural networks via efficient encoding," in *2024 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*. IEEE, 2024.
- [7] J. R. Stevens, R. Venkatesan, S. Dai, B. Khailany, and A. Raghunathan, "Softmax: Hardware/software co-design of an efficient softmax for transformers," in *2021 58th ACM/IEEE Design Automation Conference (DAC)*. IEEE, 2021, pp. 469–474.
- [8] J. Yu, J. Park, S. Park, M. Kim, S. Lee, D. H. Lee, and J. Choi, "Nn-lut: neural approximation of non-linear operations for efficient transformer inference," in *Proceedings of the 59th ACM/IEEE Design Automation Conference*, 2022, pp. 577–582.
- [9] Y. Lin, T. Zhang, P. Sun, Z. Li, and S. Zhou, "Fq-vit: Post-training quantization for fully quantized vision transformer," 2023. [Online]. Available: <https://arxiv.org/abs/2111.13824>
- [10] W. Wang, S. Zhou, W. Sun, P. Sun, and Y. Liu, "Sole: Hardware-software co-design of softmax and layernorm for efficient transformer inference," in *2023 IEEE/ACM International Conference on Computer Aided Design (ICCAD)*. IEEE, 2023, pp. 1–9.
- [11] Z. Li and Q. Gu, "I-vit: Integer-only quantization for efficient vision transformer inference," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 17 065–17 075.
- [12] R. Hu, B. Tian, S. Yin, and S. Wei, "Efficient hardware architecture of softmax layer in deep neural network," in *2018 IEEE 23rd International Conference on Digital Signal Processing (DSP)*. IEEE, 2018, pp. 1–5.
- [13] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proceedings of the fourteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 2011, pp. 315–323.
- [14] R. Crandall and C. Pomerance, *Prime numbers: a computational perspective, volume 182*. Springer Science & Business Media, 2006.
- [15] T. J. Ham, S. J. Jung, S. Kim, Y. H. Oh, Y. Park, Y. Song, J.-H. Park, S. Lee, K. Park, J. W. Lee *et al.*, "A³: Accelerating attention mechanisms in neural networks with approximation," in *HPCA*. IEEE, 2020, pp. 328–341.
- [16] H. Wang, Z. Zhang, and S. Han, "Spatten: Efficient sparse attention architecture with cascade token and head pruning," in *HPCA*. IEEE, 2021, pp. 97–110.
- [17] T. J. Ham, Y. Lee, S. H. Seo, S. Kim, H. Choi, S. J. Jung, and J. W. Lee, "Elsa: Hardware-software co-design for efficient, lightweight self-attention mechanism in neural networks," in *ISCA*. IEEE, 2021.
- [18] F. Spagnolo, S. Perri, and P. Corsonello, "Aggressive approximation of the softmax function for power-efficient hardware implementations," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 69, no. 3, pp. 1652–1656, 2021.
- [19] C. Peltekis, K. Alexandridis, and G. Dimitrakopoulos, "Reusing softmax hardware unit for gelu computation in transformers," in *AICAS*, 2024.
- [20] S. Shen, Z. Dong, J. Ye, L. Ma, Z. Yao, A. Gholami, M. W. Mahoney, and K. Keutzer, "Q-bert: Hessian based ultra low precision quantization of bert," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 05, 2020, pp. 8815–8821.
- [21] S. Dai, R. Venkatesan, M. Ren, B. Zimmer, W. Dally, and B. Khailany, "Vs-quant: Per-vector scaled quantization for accurate low-precision neural network inference," *Proceedings of Machine Learning and Systems*, vol. 3, pp. 873–884, 2021.
- [22] Y. Bondarenko, M. Nagel, and T. Blankevoort, "Understanding and overcoming the challenges of efficient transformer quantization," *arXiv preprint arXiv:2109.12948*, 2021.
- [23] Z. Wang, C. Wang, X. Xu, J. Zhou, and J. Lu, "Quantformer: Learning extremely low-precision vision transformers," *TPMAI*, vol. 45, no. 7, pp. 8813–8826, 2022.
- [24] D. Hendrycks and K. Gimpel, "Gaussian error linear units (gelus)," *arXiv preprint arXiv:1606.08415*, 2016.
- [25] J. Lei Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *ArXiv e-prints*, pp. arXiv–1607, 2016.
- [26] Y. Lin, T. Zhang, P. Sun, Z. Li, and S. Zhou, "Fq-vit: Post-training quantization for fully quantized vision transformer," *arXiv preprint arXiv:2111.13824*, 2021.
- [27] Z. Yuan, C. Xue, Y. Chen, Q. Wu, and G. Sun, "Ptq4vit: Post-training quantization for vision transformers with twin uniform quantization," in *European conference on computer vision*. Springer, 2022, pp. 191–207.
- [28] Z. Li, J. Xiao, L. Yang, and Q. Gu, "Repq-vit: Scale reparameterization for post-training quantization of vision transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 17 227–17 236.
- [29] J. L. McKinstry, S. K. Esser, R. Appuswamy, D. Bablani, J. V. Arthur, I. B. Yildiz, and D. S. Modha, "Discovering low-precision networks close to full-precision networks for efficient embedded inference," *arXiv preprint arXiv:1809.04191*, 2018.
- [30] D. Wu, Q. Tang, Y. Zhao, M. Zhang, Y. Fu, and D. Zhang, "Easyquant: Post-training quantization via scale optimization," *arXiv preprint arXiv:2006.16669*, 2020.
- [31] X. Wei, Y. Zhang, X. Zhang, R. Gong, S. Zhang, Q. Zhang, F. Yu, and X. Liu, "Outlier suppression: Pushing the limit of low-bit transformer language models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 17 402–17 414, 2022.
- [32] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.
- [33] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *International conference on machine learning*. PMLR, 2021, pp. 10 347–10 357.
- [34] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.
- [35] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman, "Glue: A multi-task benchmark and analysis platform for natural language understanding," *arXiv preprint arXiv:1804.07461*, 2018.
- [36] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 2019, pp. 4171–4186.
- [37] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.
- [38] T. Wang, L. Gong, C. Wang, Y. Yang, Y. Gao, X. Zhou, and H. Chen, "Via: A novel vision-transformer accelerator based on fpga," *IEEE TCAD*, vol. 41, no. 11, pp. 4088–4099, 2022.
- [39] M. Sun, Z. Li *et al.*, "Fpga-aware automatic acceleration framework for vision transformer with mixed-scheme quantization: Late breaking results," in *DAC*, 2022, pp. 1394–1395.
- [40] M. Huang, J. Luo, C. Ding, Z. Wei, S. Huang, and H. Yu, "An integer-only and group-vector systolic accelerator for efficiently mapping vision transformer on edge," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 70, no. 12, pp. 5289–5301, 2023.
- [41] X. Shen, Z. Han, L. Lu, Z. Kong, P. Dong, Z. Li, Y. Xie, C. Wu, M. Leeser, P. Zhao, X. Lin, and Y. Wang, "Hotaq: Hardware oriented token adaptive quantization for large language models," *IEEE TCAD*, pp. 1–1, 2024.
- [42] Q. Dong, X. Xie, and Z. Wang, "Swat: An efficient swin transformer accelerator based on fpga," in *ASP-DAC*, 2024, pp. 515–520.
- [43] S. Lu, M. Wang, S. Liang, J. Lin, and Z. Wang, "Hardware accelerator for multi-head attention and position-wise feed-forward in the transformer," in *SOCC*, 2020, pp. 84–89.