Late Breaking Results: FPGA-Aware Automatic Acceleration Framework for Vision Transformer with Mixed-Scheme Quantization

Mengshu Sun[1]*, Zhengang Li[1]*, Alec Lu[2]*, Haoyu Ma[3], Geng Yuan[1], Yanyue Xie[1], Hao Tang[4], Yanyu Li[1], Miriam Leeser[1], Zhangyang Wang[5], Xue Lin[1], Zhenman Fang[2]
[1]Northeastern University, Boston, MA, United States [2]Simon Fraser University
[3]University of California, Irvine [4]ETH Zurich [5]University of Texas at Austin
[1]{sun.meng, li.zhen, yuan.geng, xie.yany, li.yanyu, xue.lin}@notheastern.edu, mel@coe.neu.edu

ABSTRACT

Vision transformers (ViTs) are emerging with significantly improved accuracy in computer vision tasks. However, their complex architecture and enormous computation/storage demand impose urgent needs for new hardware accelerator design methodology. This work proposes an FPGA-aware automatic ViT acceleration framework based on the proposed mixed-scheme quantization. To the best of our knowledge, this is the first FPGA-based ViT acceleration framework exploring model quantization. Compared with state-of-the-art ViT quantization work (algorithmic approach only without hardware acceleration), our quantization achieves 0.31% to 1.25% higher Top-1 accuracy under the same bit-width. Compared with the 32-bit floating-point baseline FPGA accelerator, our accelerator achieves around 5.6× improvement on the frame rate (i.e., 56.4 FPS vs. 10.0 FPS) with 0.83% accuracy drop for DeiT-base.

ACM Reference Format:

Mengshu Sun, Zhengang Li, Alec Lu, Haoyu Ma, Geng Yuan, Yanyue Xie, Hao Tang, Yanyu Li, Miriam Leeser, Zhangyang Wang, Xue Lin, Zhenman Fang. 2022. Late Breaking Results: FPGA-Aware Automatic Acceleration Framework for Vision Transformer with Mixed-Scheme Quantization. In *Proceedings of the 59th ACM/IEEE Design Automation Conference (DAC) (DAC '22), July 10–14, 2022, San Francisco, CA, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3489517.3530618

1 INTRODUCTION

Transformer, an attention-based encoder-decoder architecture [8], has revolutionized the field of natural language processing (NLP) in recent years. This inspired researchers to adopt transformer-like architecture to computer vision tasks, i.e., vision transformers (ViTs), achieving better performance compared with state-of-the-art convolutional neural networks (CNNs) [3, 7, 9]. However, the complex architecture and enormous computation and storage of ViTs make it challenging for their deployment on resource constrained edge devices. Existing work on ViT accelerators on hardware [4, 6, 12] mainly utilized weight pruning. For quantization, efforts were made on the algorithm level only, and most [1, 10, 11] were on transformers for NLP, while little work [5] has been devoted to ViTs.

* The first three authors contribute equally to this research.

DAC '22, July 10–14, 2022, San Francisco, CA, USA

ACM ISBN 978-1-4503-9142-9/22/07…$15.00
https://doi.org/10.1145/3489517.3530618

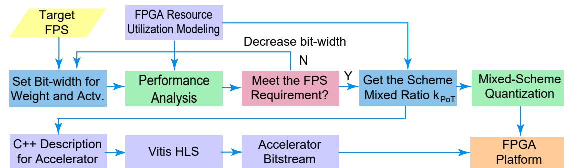

Figure 1: Overview of Auto-ViT-Acc.

This work develops Auto-ViT-Acc, an FPGA-aware automatic acceleration framework for ViTs with mixed-scheme quantization, where fixed-point (Fixed) and power-of-two (PoT) quantization schemes are combined and assigned down to the row (weight filter) level for each layer. The Fixed scheme is used for preserving the accuracy, and multiplications in this scheme can be efficiently implemented with the DSP resources on FPGAs. The PoT scheme is used to further explore the usage of LUT resources on FPGAs, since multiplications in this scheme can be replaced by simple bit shift operations that can be managed by the LUTs. Combining Fixed and PoT enables the potential to improve the FPGA resource utilization efficiency (by utilizing both DSP and LUT resources simultaneously) for inference acceleration while maintaining accuracy. The entire workflow is automated based on a target frame rate (FPS), for a quantized model and an FPGA accelerator. Specifically, an estimation of the FPS of the ViT accelerator is given under a specific bit-width setting, which is adjusted until the target FPS is met. The bit-width and the ratio of Fixed rows over PoT rows can then be optimized to guide the quantization algorithm and the accelerator design that is optimized for ViT multi-head attention.

The contributions of our work are summarized as follows:

- An FPGA-aware mixed-scheme ViT quantization algorithm that can fully leverage heterogeneous FPGA resources while maximally retaining accuracy.
- An automated ViT acceleration framework that optimizes the bit-widths to guide the ViT quantization and automates the workflow from a targeted FPS to an FPGA accelerator implementation.
- The first FPGA-based ViT acceleration framework exploring quantization with significant speedups.

2 PROPOSED AUTO-VIT-ACC FRAMEWORK

Fig. 1 provides the workflow of Auto-ViT-Acc for automatic generations of ViT accelerators. The "FPGA Resource Utilization Modeling" module preforms performance analysis to estimate the FPS of the ViT accelerator given the bit-widths for the Fixed (b) and PoT (b') schemes. The bit-widths are reduced until the target FPS is fulfilled. After deriving the desired ratio k_{PoT} for PoT quantized rows (elaborated later), the mixed-scheme quantization algorithm uses b, b', and k_{PoT} to quantize ViTs, which will be implemented on FPGAs by going through "C++ Description for Accelerator", "Xilinx Vitis High-Level Synthesis (HLS)", and "Accelerator Bitstream".

Table 1: Notations for ViT Accelerator

Notation	Description
T_n	Tiling size for data in input channel dimension in each head
T_m^{Fix} (T_m^{PoT})	Tiling size for Fixed (PoT) data in output channel dimension
N_h	Total number of heads
P_h	Number of heads for computation in parallel
G (G')	Number of data packed as one for activations and Fixed (PoT) weights

Quantization is applied to linear layers of ViT, which involve the most computation-intensive matrix multiplications. For each layer, some of the rows are quantized into the Fixed scheme with b-bit for weights and b-bit for the corresponding activations (Fixed W[b]A[b]), and the rest rows are quantized into the PoT scheme b'-bit for weights and b-bit for corresponding activations (W[b']A[b]). To prevent extra hardware overhead on output shifting among two schemes, we set $2^{(b'-1)} \leq b$, i.e., if b-bit is used for Fixed, then $b' = \lfloor \log_2 b \rfloor + 1$ is used for PoT. The same ratio k_{PoT} is used among different heads of each multi-head self-attention module in ViTs to fully exploit the parallelism of FPGAs. The quantization scheme is assigned down to the row level of a weight matrix based on the weight distribution of a row. If the variance of a row is small, the row is assigned PoT, otherwise Fixed. Detailed quantizer functions can be found in [2].

An FPGA board contains primarily two types of computation resources, namely DSPs and LUTs. Multiplications with fixed-point weights are computed with DSPs, and those with PoT weights can be replaced by shifting operations computed with LUTs. The notations used in ViT accelerators are listed in Table 1. The compute engine can manage $(T_m^{\text{Fix}}+T_m^{\text{PoT}})\cdot P_h \cdot T_n$ multiply-accumulate (MAC) operations in parallel. The parameters to be determined for the accelerator include T_m^{Fix} (T_m^{PoT}), T_n, G (G'), and P_h. On a specific FPGA board, the maximum achievable FPS, denoted by FPS_{max}, can be estimated according to our analysis of FPGA resource utilization and performance. Given the target FPS, denoted by FPS_{tgt}, we first find the precision and scheme combination satisfying $\text{FPS}_{\text{max}} \geq \text{FPS}_{\text{tgt}}$. Under this precision, we fix P_h, T_n, G (G'), and T_m^{Fix}, and adjust T_m^{PoT} to meet the target FPS and obtain the best model accuracy. In detail, P_h is set to a value that can divide N_h exactly for full exploitation of computation resources, i.e., $P_h = 3$ for $N_h = 6$, and $P_h = 4$ for $N_h = 8$ or $N_h = 12$. G is decided based on the FPGA AXI port width and the quantization bit-width of Fixed weights, and is the same for activations in both Fixed and PoT computations as well as weights in Fixed computations. The bit-width of PoT weights is lower, corresponding to G'. T_n is set to the same value as G. The computation parallelism along the output channel dimension is decided by the sum $T_m^{\text{Fix}} + T_m^{\text{PoT}}$, and the model accuracy in quantization is affected by the ratio $k_{\text{PoT}} = \frac{T_m^{\text{PoT}}}{T_m^{\text{PoT}}+T_m^{\text{Fix}}}$, i.e., lower k_{PoT} will result in higher model accuracy. We therefore reduce T_m^{PoT} to make the actual FPS equal to FPS_{tgt} if $\text{FPS}_{\text{max}} > \text{FPS}_{\text{tgt}}$ under this precision, and the actual k_{PoT} ratio will guide the quantization process and the hardware implementations with all these parameters.

3 EXPERIMENTAL RESULTS

The comparison results of different quantization schemes in terms of accuracy after quantization and performance with resource utilization are listed in Table 2. For DeiT-small, it can be seen that a target FPS of 150 can be met using W4A4+W3A4 quantization precision with PoT ratio $k_{\text{PoT}} = 43\%$ and the Top-1 accuracy reaches 77.78%. For the desired FPS of 100, the implementation using W8A8+W4A8 precision with $k_{\text{PoT}} = 43\%$ can fulfill the requirement with 78.58% accuracy. As for DeiT-base, the accuracy loss incurred by quantization is less than 1%, while 50 FPS with 81.02% accuracy can be achieved using W4A4+W3A4 precision with $k_{\text{PoT}} = 40\%$, and 30 FPS with 81.71% accuracy can be reached using W8A8+W4A8 precision with $k_{\text{PoT}} = 45\%$. Compared with the 32-bit baseline model, our quantized model achieves around 5.6× improvement on frame rate (i.e., 56.4 FPS vs. 10.0 FPS) with only 0.83% accuracy drop.

Table 2: Accuracy and Hardware Results under Different Quantization Schemes for DeiT-small and DeiT-base Models on ImageNet Dataset

Quantization Weight Scheme	Bit-Width (Weight/Actv.)	Accuracy (%) Top-1	Top-5	Resource Util. DSP	kLUT	Power (W)	Thrpt. (FPS)	Energy Effi. (FPS/W)
DeiT-small								
Baseline	W32A32	79.85	94.97	1745	130	8.38	38.9	4.64
PTQ [5] (Fixed)	W8A8	77.47	-	-	-	-	-	-
Fixed	W4A4	78.57	94.41	1933	137	10.44	130.3	12.48
PoT	W3A4	77.24	93.89	13	176	6.55	150.9	23.04
Fixed+PoT	W4A4+W3A4 ($k_{\text{PoT}} = 43\%$)	77.78	94.00	1549	193	10.34	155.8	15.06
Fixed	W8A8	79.69	94.89	1936	122	8.46	78.1	9.23
PoT	W4A8	77.97	94.06	16	175	8.58	91.9	10.71
Fixed+PoT	W8A8+W4A8 ($k_{\text{PoT}} = 43\%$)	78.58	94.43	1552	185	9.63	99.7	10.35
DeiT-base								
Baseline	W32A32	81.85	95.59	1564	120	9.91	10.0	1.01
PTQ [5] (Fixed)	W8A8	80.48	-	-	-	-	-	-
Fixed	W4A4	81.33	95.63	2064	139	11.27	47.5	4.21
PoT	W3A4	80.87	95.57	19	191	8.11	56.8	7.00
Fixed+PoT	W4A4+W3A4 ($k_{\text{PoT}} = 40\%$)	81.02	95.60	1555	179	11.03	56.4	5.11
Fixed	W8A8	81.93	95.90	2066	128	9.40	25.9	2.76
PoT	W4A8	81.51	95.73	20	192	7.24	31.1	4.30
Fixed+PoT	W8A8+W4A8 ($k_{\text{PoT}} = 45\%$)	81.73	95.85	1556	186	9.31	34.0	3.66

4 CONCLUSION

This work proposes an automatic acceleration framework for ViT with mixed-scheme quantization. To the best of our knowledge, this is the first work of quantization-based ViT acceleration on FPGAs.

ACKNOWLEDGMENTS

This work was partly supported by NSF CCF-1901378.

REFERENCES

[1] Haoli Bai et al. 2021. Binarybert: Pushing the limit of bert quantization. In *ACL/IJCNLP (1)*.
[2] Sung-En Chang et al. 2021. Mix and Match: A novel FPGA-centric deep neural network quantization framework. In *HPCA*.
[3] Alexey Dosovitskiy et al. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *ICLR*.
[4] Bingbing Li et al. 2020. Ftrans: energy-efficient acceleration of transformers using fpga. In *ISLPED*.
[5] Zhenhua Liu et al. 2021. Post-Training Quantization for Vision Transformer. In *NeurIPS*. https://openreview.net/forum?id=9TX5OsKJvm
[6] Panjie Qi et al. 2021. Accommodating Transformer onto FPGA: Coupling the Balanced Model Compression and FPGA-Implementation Optimization. In *GLSVLSI*.
[7] Hugo Touvron et al. 2021. Training data-efficient image transformers & distillation through attention. In *ICML*.
[8] Ashish Vaswani et al. 2017. Attention is all you need. In *NeurIPS*.
[9] Li Yuan et al. 2021. Tokens-to-Token ViT: Training Vision Transformers From Scratch on ImageNet. In *ICCV*.
[10] Ofir Zafrir et al. 2019. Q8bert: Quantized 8bit bert. *NeurIPS EMC2 Workshop* (2019).
[11] Wei Zhang et al. 2020. Ternarybert: Distillation-aware ultra-low bit bert. In *EMNLP*.
[12] Xinyi Zhang et al. 2021. Algorithm-hardware Co-design of Attention Mechanism on FPGA Devices. *TECS* (2021).