UCViT: Hardware-Friendly Vision Transformer via Unified Compression

HongRui Song*, Ya Wang*, Meiqi Wang, and Zhongfeng Wang
School of Electronic Science and Engineering, Nanjing University, Nanjing, China
Email: {hrsong, ywang, mqwang}@smail.nju.edu.cn, {zfwang}@nju.edu.cn

Abstract—**Vision Transformer (ViT) has emerged as a powerful model with its extraordinary performance on multiple computer vision applications. However, the huge model size and the enormous energy consumption incurred by the dense matrix multiplications make ViT hard to be implemented on edge devices. To tackle these challenges, we develop a unified compression framework for Vision Transformer (UCViT), whose main focus is on compressing the original ViT model by incorporating the low bit-width quantization and the dense matrix decomposition. To maximally reduce the energy expenditure, we propose a dedicated design by leveraging aggressive quantization, in which the majority of the matrix multiplications are converted to the hardware-friendly shift and addition operations. Besides, we incorporate a small module into the quantized model by harnessing the unique characteristic of multi-head attention during matrix decomposition, which achieves significant accuracy recovery from the deeply compressed model with minimal impact on the energy efficiency. Benefited from the effective fusion of different compression techniques and the hardware-friendly operations, the proposed model can save up to 98% energy consumption in inference compared to the original ViT model. Experiments on CIFAR-10 and CIFAR-100 image classification tasks show that the proposed model obtains a highly compact structure with a competitive compression ratio (up to 6.7×), while causes small loss (less than 1%) on the accuracy.**

Index Terms—**Vision Transformer (ViT), Image Classification, Unified Compression, Neural Network**

I. Introduction

As the Transformer-based models achieve remarkable performance in Natural Language Processing (NLP) tasks, their applications in Computer Vision (CV) tasks also present decent performance among many missions such as image classification [1] [2], object detection [3], and semantic segmentation [4] [5]. However, the tremendous memory and computation consumption makes it challenging for these models to be developed on resource-limited devices like portable phones. Thus, the compressed versions for the aforementioned models are desperately needed for better usage in the industry.

Recently, a considerable literature has grown up around the theme of compression methods for neural networks, among which decomposition and quantization stand out due to their reliable performance. Decomposition has already been introduced to Transformer-based models in [6] in order to decrease computational complexity for its implementation on resource-limited devices. However, the mentioned work suffers from the time-consuming training process, in which an extra decomposition module is introduced and needs to be trained from scratch. As an alternative, quantization approaches can also notably reduce computing cost by converting the floating-point operations to integer operations with the predefined bit-width, generating a great deal of compressed Transformer-based models [7] [8] [9]. Nevertheless, these works suffer from varying degrees of accuracy loss when configured with the corresponding bit-width. Furthermore, the potential of a high-level sparse structure combining the above compression methods is merely explored. Therefore, a unified compression design is urgently desired to distill reliable information from the pre-trained model, fuse the advantages of the above compression methods, and achieve a better trade-off between accuracy performance and energy efficiency.

In this paper, we propose a Unified Compression version of Vision Transformer, termed UCViT. The proposed model is mainly focused on reducing the huge energy consumption, shrinking the overall model size, and achieving reliable accuracy performance. Main contributions of this work can be summarized as follows:

- We propose a unified compression method by effectively fusing decomposition, quantization, and shift, in which the high-cost dense matrix multiplication are converted to the hardware-friendly computation dominated by shifts and additions.
- We incorporate a small module with high-precision into the proposed compact model, which achieves significant accuracy recovery with negligible impact on the training efficiency.
- The unique characteristic of Transformer is thoughtfully considered and leveraged in our design in the process of backward propagation, which further improves the accuracy performance of the proposed model and reduces the overall memory footprint.

II. Related Work

A. Vision Transformer

Inspired by the outstanding performance that Transformer in NLP tasks has obtained, researchers are thriving to explore the potential of its application in the CV field. The mainstream applications of Transformer in CV tasks fall into the following two categories. The majority of works focus on improving the original ViT structure [2] for better extraction and usage of the local feature. To tackle this challenge, a series of designs for local feature aggregation is proposed [10] [11] [12] to improve the modeling capacity by hierarchical architecture design. Meanwhile, some researchers devote themselves to modifying the computation process of self-attention for better

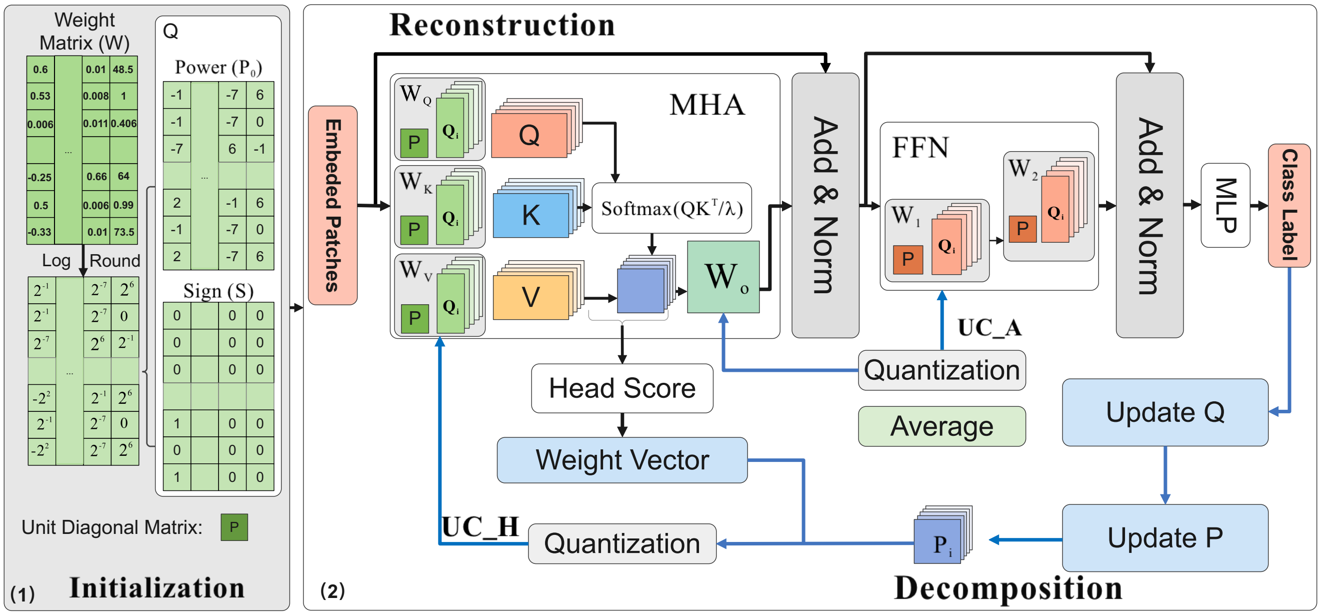

Fig. 1. The overall training process of UCViT. **(1)** During initization, each element in quantized pre-trained weight matrix (W) is rounded to its nearest power-of-two value and stored in matrix Q in the form of 4-bit power and 1-bit sign. **(2)** The black arrow represents the forward propagation including reconstruction process and the blue arrow represents the backward propagation including decomposition process.

global feature extraction ability. DeepViT [13] and XCiT [14] develop cross-head and cross-channel communications, respectively, for self-attention to boost the diversity at each layer. Credit to the reliable performance of Transformer-based models, more and more works are emerged to tackle relevant CV tasks.

B. *Compression of Transformer-Based Models*

As Transformer-based models achieve great success in both NLP and CV fields, a great many researchers are committed to realizing compressed versions of these models to alleviate the energy consumption and the model size. Quantization serves as one of the core compression measures by reducing the bit-width of parameters in the network. In the NLP field, compared to the full-precision baseline, Q-BERT [15] achieves decent performance leveraging a Hessian based mix-precision method, while Q8BERT [8] utilizes symmetric linear quantization to optimize BERT to 8-bit. What's more, a detailed analysis of quantization application on Vision Transformer is proposed by [16], which presents an effective mixed-precision quantization scheme. Decomposition serves as another effective compression measure to boost the training efficiency, which is realized by replacing the original matrices with smaller low-rank sub-matrices. ALBERT [17] leverages decomposition in the word embedding layer obtaining a prominent reduction of parameters. Inspired by these works, we propose a unified compression method fusing the above quantization and decomposition techniques, which can be utilized for the efficient execution of Vision Transformer.

III. Methodology

A. *Overall Framework*

Quantization and decomposition inevitably lead to accuracy loss. In this section, we propose the first trial of a unified compression method, which aims at maximally reducing the energy consumption and the model size with minimal degradation of the accuracy performance. To realize theses goals, we formulate and propose the following algorithm.

As presented in Fig.1, the unified compression training process of UCViT can be divided into three stages: quantization, reconstruction, and decomposition. Specifically, to achieve an ultimate compact data representation, we first implement logarithmic quantization during the initialization. After that, we reconstruct the weight matrix by shift and addition operations in forward propagation. During backward propagation, decomposition is utilized to provide adequate information for update. These stages are implemented iteratively during the corresponding training process.

B. *Ultra Low-Bit Unified-Compressed Training*

Motivated by converting the original high-cost ViT model with huge memory consumption to a hardware-friendly model composed of low-cost computations, we present this unified compression algorithm for the training process whose outline is described in Algorithm 1. The three steps to be iterated are presented below:

Initialization by Quantization Provided with a weight matrix W from the pre-trained model, we first quantize its data representation to low-bit to initialize matrix Q. Additionally, the values of Q after quantization are rounded to logarithm (the powers of 2) to obtain a further compact data representation.

Algorithm 1 Unified Compression Algorithm

```
Initialize $Q \in R^{m \times n}$ by quantizing and rounding pre-trained weight $W \in R^{m \times n}$ to powers of 2 ;
Initialize $P \in R^{r \times r}$ by a unit diagonal matrix $I$;
**During Forward Propagation**:
    Reset W' = 0;
    Tile $Q$ to $Q_i \in R^{m \times r}, i = 1 \sim j$
    **For** $i = 1$; $i < j + 1$; $i + +$ **do**
        $W'_i = P \times Q_i$
        $W' = Concatenate(W', W'_i)$ ;
    **EndFor**
**During Backward Propagation**:
    Update $Q$ by $UC\text{-}Shift$;
    Update $P$ by $UC\text{-}A$ or $UC\text{-}H$;
```

For convenience of update, we separately store the sign bit and the power of each element, forming two sub-matrices S and P_o. To remedy the accuracy degradation brought by aggressive quantization, we introduce an auxiliary accuracy reconstruction matrix P with small size and high precision. The matrix P is initialized by a unit diagonal matrix I.

Reconstruction in Forward Propagation During the forward propagation, we implement the reconstruction of the weight matrix W' tile by tile. Specifically, each tile W'_i is produced by the multiplication of the auxiliary matrix P and the corresponding tile Q_i. It's convenient enough for us to reconstruct W' by concatenating each W'_i tile. Besides, it's worth mentioning that the matrix multiplications in the reconstruction process can be realized by shift and addition operations due to the logarithmic property of matrix Q.

Decomposition in Backward Propagation To maintain the compact data representation of matrix Q, we propose the $UC\text{-}Shift$ method to directly update matrix Q in logarithmic representation, whose details are presented below:

- **UC-Shift** As one can see, the logarithmic matrix Q can be represented by two sub-matrices—the sign bit matrix S and the power matrix P_o, where corresponding elements in the three matrices obey $q = s \times 2^{p_o}$. We propose to update them by direct gradient descent. Referring to [18], we set the derivatives of the round function and the sign function to 1. Provided with the result of forward process $Y = W' \times X$, the final gradients from loss C in the backward process can be calculated as

$$\begin{aligned} \frac{\partial C}{\partial P_o} &= \frac{\partial C}{\partial Y} \frac{\partial Y}{\partial P_o} P_o \times ln2, \\ \frac{\partial C}{\partial S} &= \frac{\partial C}{\partial Y} \frac{\partial Y}{\partial S}. \end{aligned} \tag{1}$$

Once obtaining the updated matrix Q and its tiles Q_i, we are adequately prepared for the update of matrix P. To acquire sufficient prior knowledge from the pre-trained model, we construct a series of $P_i \in R^{r \times r}$ by formulating $argmin_{P_i} \|W_i - P_i Q_i\|_2$, where W_i is the tile of pre-trained weight matrix W. Then we can renew matrix P by a weight vector $P = \sum_{i=1}^{j} \alpha_i \times P_i$, where α_i denotes the importance of each P_i. We propose two methods to update the hyper-parameters α named UC-A and UC-H whose details are presented below:

- **UC-A** The UC-A method can be simply realized by applying $\alpha_i = 1/j$, where j denotes the number of P_i, and the result of P stands for the average of all P_i components.
- **UC-H** The UC-H method is inspired by the unique characteristic of self-attention in Transformer. We update the weight tensor α by the normalized sum of absolute values of each head's attention score from the previous self-attention calculation, which can be formulized as $\alpha = Normalize(\|head_i\|_1, i = 1 \sim j)$. Thus, the result of P stands for the weighted summation of P_i based on each head's importance.

It's worth mentioning that the $UC\text{-}A$ method can be widely applied to arbitrary linear layers, while the $UC\text{-}H$ method is more suitable to the calculation of QKV in self-attention with a constrain that r is set to the dimension of each head.

IV. Experimental Results

A. Implementation Details

Datasets The CIFAR-10 and CIFAR-100 datasets for image classification are leveraged in experiments to evaluate the performance of both the proposed methods. The former dataset is composed of 50K training images and 10K validation images with 10 labels, while the latter one consists of 50K training images and 10K validation images with larger 100 labels.

Baseline and experiment settings We implement the unified compression method on the original ViT-B model whose original bit-width is set to be 32. Besides, we utilize the ImageNet1k pre-trained model for post-training. For a fair comparison, the hyper-parameter r in both methods is restricted to the dimension of each head which is equal to 64. We select Radam as the optimizer whose learning rate is 0.001, while the maximum iteration is set to merely 2. The small high-precision matrix P is quantized to 8-bit, and the large aggressively quantized matrix Q is quantized to 5-bit including 1-bit for the sign bit and 4-bit for the data value. Basically, we initialize matrix P to be a unit diagonal matrix and update it 10 or 100 batches at a time for CIFAR-10 and CIFAR-100 tasks, respectively. Additionally, the large matrix Q is initialized by quantizing and rounding the pre-trained weight matrix W and is updated during each batch.

B. Results and Analysis

Model performance The accuracy performance and model size on the development set of CIFAR-10 and CIFAR-100 is demonstrated in Table I. For the CIFAR-100 dataset, the direct implementation of low-bit quantization on the classic ViT model suffers from a great accuracy loss which is 6.7%. When equipped with both methods (UC-A and UC-H), the proposed model can still maintain the small model size, while presents significant accuracy recovery which is 5.3% and 6.0%, respectively. Compared to the original ViT model, we

TABLE I
COMPARISON OF PROPOSED UCVIT MODELS WITH OTHER COMPRESSION MODELS FOR IMAGE CLASSIFICATION

Model	Dataset	Method	Weight_bit	Activation_bit	Model Size(MB)	Top-1 Accuracy	Accuracy Recovery
ViT-B	CIFAR-10	Original ViT	32	32	344	97.71	/
		Pure Quantization (baseline)	6	6	64.5	93.48	0
		UCViT (UC-A)	5(Q) / 8(P)	16	**51.4**	**96.83**	3.35
		UCViT (UC-H)	5(Q) / 8(P)	16	**51.4**	**97.38**	**3.90**
		MP Quantization	6MP	6MP	64.6	96.83	3.35
		MP Quantization	8MP	8MP	86.0	97.79	3.96
	CIFAR-100	Original ViT	32	32	344	87.30	/
		Pure Quantization (baseline)	6	6	64.5	80.56	0
		UCViT (UC-A)	5(Q) / 8(P)	16	**51.6**	**85.91**	5.35
		UCViT (UC-H)	5(Q) / 8(P)	16	**51.6**	**86.57**	**6.01**
		MP Quantization	6MP	6MP	64.4	83.99	3.43
		MP Quantization	8MP	8MP	86.5	85.76	5.20

obtain a significant compression ratio up to 6.7x, while the final accuracy loss is merely 1.4% and 0.7%. Furthermore, compared to other compression work of ViT [16], our model achieves higher accuracy performance under the similar model size. For the CIFAR-10 dataset, our model also presents decent performance. Compared to the baseline, our methods achieve 3.4% and 3.9% accuracy recovery, which is competitive to the prior work. The above demonstrates UCViT's superiority of effectiveness.

Energy efficiency analysis A key feature relevant to the aggressively quantized matrix Q is that all of its values are presented in the logarithmic form, which enables us to implement the reconstruction of W' in a shift and addition way. Furthermore, after the dedicated design of the calculation order in the inference process, the ratio of hardware-friendly operations can increase three times. Specifically, we design to multiply input or activation with the small matrix P first. Then, we multiply the outcome with the logarithmic matrix Q. The final energy consumption and the count of multiplication and accumulation (MAC) operations in inference are shown in Table II. Our model achieves significant energy efficiency by converting the high-cost multiplication to the low-cost shift, saving up to 98% energy consumption compared to the original ViT model. Thus, due to the ultra-low energy cost in inference and the small model size, the proposed model presents prominent hardware-friendly feature.

TABLE II
COMPARISON OF ENERGY CONSUMPTION IN INFERENCE

Model	FloatMM	FixMM	Add.	Shift	Energy cost
Baseline	17.4G	0	17.4G	0	4002.0J
Quantization	0	17.4G	17.4G	0	200.1J
UCViT	0	2.2G	19.0G	16.8G	**70.5J**

Flexible and better accuracy-speed tradeoffs As shown in Fig.2, we carry out a series of experiments on different update frequencies of matrix P to seek for the sweet point of decent accuracy performance and training time. Our observation is that we can further achieve at least 5.7× and 4.2× training efficiency with negligible accuracy loss for CIFAR-10 and CIFAR-100 tasks. It's also worth mentioning that during the post-training, both of our methods demonstrate excellent performance on convergence speed, since we merely require at most two epochs for the accuracy convergence. What's more, the change of update frequency shows approximately no impact on the convergence speed of our methods.

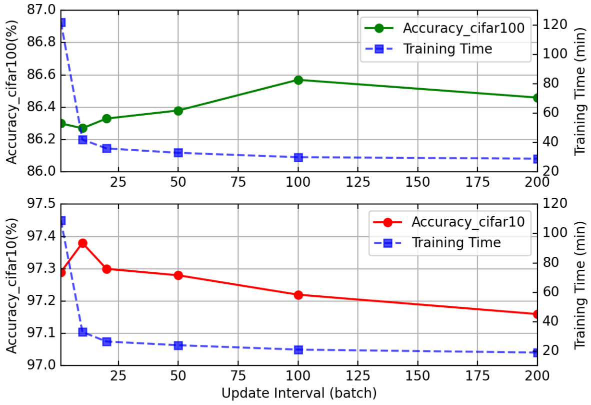

Fig. 2. Accuracy and Inference Efficiency Tradeoff

V. CONCLUSION AND FUTURE WORK

In this paper, we propose UCViT, a unified compressed version of Vision Transformer, which is the first trial to effectively fuse quantization and decomposition during post-training. On CIFAR-10 and CIFAR-100 tasks, the proposed model achieves up to 6.7× compression ratio with 98% energy reduction in inference. Benefited from the dedicated design considering the unique characteristic of multi-head attention, the proposed model achieves significant accuracy recovery over the directly quantized ViT model, and the results significantly outperform prior work with the similar model size on CIFAR-100 task. Our future works will address the hardware design of the proposed hardware-friendly model. Furthermore, we plan to explore the adaptability of UC-A method on the general matrix multiplication of other tasks.

REFERENCES

[1] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herv'e J'egou. Training data-efficient image transformers & distillation through attention. In *ICML*, 2021.

[2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv*, abs/2010.11929, 2021.

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. *ArXiv*, abs/2005.12872, 2020.

[4] Robin A. M. Strudel, Ricardo Garcia Pinel, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. *ArXiv*, abs/2105.05633, 2021.

[5] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip H. S. Torr, and Li Zhang. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *CVPR*, 2021.

[6] Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. *ArXiv*, abs/2106.13008, 2021.

[7] Gabriele Prato, Ella Charlaix, and Mehdi Rezagholizadeh. Fully quantized transformer for machine translation. In *FINDINGS*, 2020.

[8] Ofir Zafrir, Guy Boudoukh, Peter Izsak, and Moshe Wasserblat. Q8bert: Quantized 8bit bert. *2019 Fifth Workshop on Energy Efficient Machine Learning and Cognitive Computing - NeurIPS Edition (EMC2-NIPS)*, pages 36–39, 2019.

[9] Aishwarya Bhandare, Vamsi Sripathi, Deepthi Karkada, Vivek V. Menon, Sun Choi, Kushal Datta, and Vikram A. Saletore. Efficient 8-bit quantization of transformer neural machine language translation model. *ArXiv*, abs/1906.00532, 2019.

[10] Kai Han, An Xiao, Enhua Wu, Jianyuan Guo, Chunjing Xu, and Yunhe Wang. Transformer in transformer. *ArXiv*, abs/2103.00112, 2021.

[11] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Ching-Feng Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *ArXiv*, abs/2103.14030, 2021.

[12] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Francis E. H. Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. *ArXiv*, abs/2101.11986, 2021.

[13] Daquan Zhou, Bingyi Kang, Xiaojie Jin, Linjie Yang, Xiaochen Lian, Qibin Hou, and Jiashi Feng. Deepvit: Towards deeper vision transformer. *ArXiv*, abs/2103.11886, 2021.

[14] Alaaeldin El-Nouby, Hugo Touvron, Mathilde Caron, Piotr Bojanowski, Matthijs Douze, Armand Joulin, Ivan Laptev, Natalia Neverova, Gabriel Synnaeve, Jakob Verbeek, and Hervé Jégou. Xcit: Cross-covariance image transformers. *ArXiv*, abs/2106.09681, 2021.

[15] Sheng Shen, Zhen Dong, Jiayu Ye, Linjian Ma, Zhewei Yao, Amir Gholami, Michael W. Mahoney, and Kurt Keutzer. Q-bert: Hessian based ultra low precision quantization of bert. In *AAAI*, 2020.

[16] Zhenhua Liu, Yunhe Wang, Kai Han, Siwei Ma, and Wen Gao. Post-training quantization for vision transformer. *ArXiv*, abs/2106.14156, 2021.

[17] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. *ArXiv*, abs/1909.11942, 2020.

[18] Mostafa Elhoushi, Farhan Shafiq, Ye Henry Tian, Joey Yiwei Li, and Zihao Chen. Deepshift: Towards multiplication-less neural networks. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2359–2368, 2021.