HALOC: Hardware-Aware Automatic Low-Rank Compression for Compact Neural Networks

Jinqi Xiao[1], **Chengming Zhang**[2], **Yu Gong**[1], **Miao Yin**[1],
Yang Sui[1], **Lizhi Xiang**[3], **Dingwen Tao**[2,3], **Bo Yuan**[1]

[1] Rutgers University, [2] Indiana University, [3] Washington State University
jinqi.xiao@rutgers.edu, czh5@iu.edu, yg430@soe.rutgers.edu, miao.yin@rutgers.edu,
yang.sui@rutgers.edu, lizhi.xiang@wsu.edu, ditao@iu.edu, bo.yuan@soe.rutgers.edu

Abstract

Low-rank compression is an important model compression strategy for obtaining compact neural network models. In general, because the rank values directly determine the model complexity and model accuracy, proper selection of layer-wise rank is very critical and desired. To date, though many low-rank compression approaches, either selecting the ranks in a manual or automatic way, have been proposed, they suffer from costly manual trials or unsatisfied compression performance. In addition, all of the existing works are not designed in a hardware-aware way, limiting the practical performance of the compressed models on real-world hardware platforms.

To address these challenges, in this paper we propose HALOC, a hardware-aware automatic low-rank compression framework. By interpreting automatic rank selection from an architecture search perspective, we develop an end-to-end solution to determine the suitable layer-wise ranks in a differentiable and hardware-aware way. We further propose design principles and mitigation strategy to efficiently explore the rank space and reduce the potential interference problem.

Experimental results on different datasets and hardware platforms demonstrate the effectiveness of our proposed approach. On CIFAR-10 dataset, HALOC enables 0.07% and 0.38% accuracy increase over the uncompressed ResNet-20 and VGG-16 models with 72.20% and 86.44% fewer FLOPs, respectively. On ImageNet dataset, HALOC achieves 0.9% higher top-1 accuracy than the original ResNet-18 model with 66.16% fewer FLOPs. HALOC also shows 0.66% higher top-1 accuracy increase than the state-of-the-art automatic low-rank compression solution with fewer computational and memory costs. In addition, HALOC demonstrates the practical speedups on different hardware platforms, verified by the measurement results on desktop GPU, embedded GPU and ASIC accelerator.

Introduction

Model compression is an important deep neural network (DNN) optimization strategy that aims to improve the execution efficiency of deep learning. Motivated by the observation that considerable redundancy exhibits at different levels (e.g., neuron and bit) of DNNs, various compression approaches, such as pruning (Han et al. 2015; Dong et al. 2020; Deng et al. 2021) and quantization (Rastegari et al. 2016; Zhang et al. 2018), have been widely studied and developed to reduce model redundancy with different granularity.

Among the existing compression methods, *low-rank compression* is a unique solution that explores the low-rankness of DNN model at the structure level. By decomposing the original large-size weight matrices or tensors to a series of small cores, low-rank compression can bring significant storage and computational savings. To date, many decomposition-based compression approaches have been proposed. Based on their differences in the factorization schemes, they can be categorized to 2-D *matrix decomposition* based (Klema and Laub 1980) and high-order *tensor decomposition* based (Tucker 1966; Harshman et al. 1970).

Low-Rank Compression in Practice. From the perspective of real-world deployment, practical low-rank DNN compression should satisfy two requirements: ***Automatic Rank Selection*** and ***Hardware-awareness***. First, assume that using Tucker-2 decomposition (Kim et al. 2016) to compress an N-layer DNN with maximum rank value as M per layer, there are total M^{2N} possible rank settings. Consider in practice N is typically tens to hundreds and M can be up to tens, the overall rank selection space is extremely huge. Evidently, automatic rank selection is highly desired to reduce the expensive search efforts and the potential sub-optimality incurred by the heuristic manual setting. Second, because the decomposed DNN models are essentially executed on the practical computing hardware, low-rank compression should be performed towards reducing the hardware-cost metrics, e.g., latency and energy, to obtain the actual benefits in the real-world scenarios. Also, consider the large diversity of current DNN computing hardware (e.g., desktop GPU, embedded GPU, application-specific integrated circuit (ASIC) accelerator, etc.) and different platforms may favor different network structures, the direct feedback from the target devices should be included in the low-rank compression process (Xiang et al. 2022).

Limitations of Existing Works. Measured from the above two practical requirements, the existing low-rank DNN compression approaches have several limitations. First, most of the prior works determine the ranks via rounds of manual trials and adjustments, causing expensive engineering efforts. Second, even for the state-of-the-art solutions with automatic rank selection, due to the insufficient exploration of rank space, their compression performance is still unsatisfied, sometimes even lower than the approaches with heuristic rank settings. Third, all of the existing low-

rank decomposition approaches, no matter determining the ranks manually or automatically, do not consider hardware-awareness in the design process, limiting their performance for deployment on practical hardware platforms.

Technical Contributions and Preview. To address these challenges and promote the widespread adoption of low-rank compression in practice, in this paper we propose HALOC, a hardware-aware automatic low-rank compression framework for compact DNN models. Based on the observation that rank selection process essentially determines the architecture of the low-rank models, we interpret automatic rank selection as the automatic search for the proper low-rank structure, enabling efficient automatic low-rank compression in an differentiable and hardware-aware way. Overall, the contributions of this paper is summarized as follows:

- We identify the hidden connection between setting the layer-wise ranks and searching the network architecture. Based on this interesting discovery, we propose to develop a new low-rank compression strategy with neural architecture search (NAS)-inspired automatic rank selection scheme, which can sufficiently explore the rank space and strong compatibility for hardware-awareness.
- We propose two low-rank-specific design principles to realize efficient exploration in the rank space, leading to significant reduction in the search cost. We also analyze the potential methods for mitigating the interference problem in the search process, and identify the suitable solution to improve compression performance.
- We perform evaluation experiments of compressing various models on different datasets, and also measure the practical speedup across different hardware platforms. On CIFAR-10 dataset, HALOC achieves 0.07% and 0.38% accuracy increase over the original uncompressed ResNet-20 and VGG-16 models with 72.20% and 86.44% FLOPs reduction, respectively. On ImageNet dataset, HALOC brings 0.9% top-1 accuracy increase over the original ResNet-18 model with 66.16% FLOPs reduction. When compressing MoblieNetV2 on ImageNet dataset, HALOC shows 0.66% top-1 accuracy increase over the state-of-the-art low-rank compression method with fewer FLOPs and memory cost. The measurement results on various hardware platforms (desktop GPU, embedded GPU and ASIC accelerator) demonstrate the practical speedups brought by our proposed hardware-aware solution.

Related Work

Low-Rank DNN Compression. Exploring structure-level low-rankness has been well studied for neural network compression. In general, a compact DNN can be obtained via performing low-rank matrix or tensor decomposition on the original large model. For matrix factorization-based methods (Denton et al. 2014; Wen et al. 2017; Deng et al. 2019b; Idelbayev and Carreira-Perpinán 2020; Liebenwein et al. 2021), all the weight tensors, including the 4-D type for the convolutional (CONV) layers, are first flatten to 2-D format and then decomposed to two small matrices; while tensor decomposition-based approaches, including Tucker (Kim et al. 2016; Gusak et al. 2019; Yin et al. 2020, 2021a), CP (Lebedev et al. 2014; Astrid and Lee 2017), tensor train (Novikov et al. 2015; Wang et al. 2019; Deng et al. 2019a; Yin et al. 2022b) and tensor ring (Pan et al. 2019), directly factorize the high-order tensor objectives to a series of tensor and matrix cores. No matter which specific decomposition is adopted, one key challenge is the efficient rank setting. Currently most of the existing works determine the layer-wise ranks via hand-crafted manual trials and attempts, a strategy that requires very expensive engineering efforts.

Automatic Rank Selection. Our work is most closely related to (Gusak et al. 2019; Liebenwein et al. 2021; Li et al. 2022; Yin et al. 2022a), the state-of-the-art automatic rank selection solutions. Specifically, (Gusak et al. 2019) proposes to utilize variational Bayesian matrix factorization to determine the ranks of the tensor decomposed DNNs in a multi-stage way. In (Liebenwein et al. 2021), Eckhart-Young theorem (Golub and Van Loan 2013) is used to determine the layer-wise ranks of the factorized weight matrices towards minimizing compression error. Besides, (Li et al. 2022) applies genetic algorithm to the rank search process of tensor ring decomposed DNNs. Compared with the manual trials, these automatic selection approaches indeed facilitate the rank determination procedure. However, their accuracy performance is still limited because of the insufficient exploration of rank space. Also, similar to the existing manual setting-based solutions, these automatic compression approaches are not designed in a hardware-aware way, limiting the performance of the compressed models on the practical hardware platforms.

Preliminaries

Notation. We denote a tensor by using boldface calligraphic script letter, e.g., $\boldsymbol{\mathcal{X}}$. Matrices and vectors are represented in the format of boldface capital letters and lower-case capital letters, e.g. $\boldsymbol{X}$ and $\boldsymbol{x}$, respectively. In addition, non-boldface letters with indices $\mathcal{X}(i_1, \cdots, i_d)$, $X(i, j)$ and $x(i)$ represent the entries for d-dimensional tensor $\boldsymbol{\mathcal{X}}$, matrix $\boldsymbol{X}$ and vector $\boldsymbol{x}$, respectively.

Tucker-2 Decomposed CONV Layer. Without loss of generality, in this paper we study the automatic low-rank compression using Tucker-2 decomposition. In general, for a CONV layer with weight tensor $\mathcal{W} \in \mathbb{R}^{F \times C \times K_1 \times K_2}$, where C is the number of input channels, K_1 and K_2 are the kernel size, and F is the number of output channels, it can be factorized to a core tensor and two matrices along each mode of Tucker-2 decomposition as follows:

$$\mathcal{W}(f, c, i, j) = \sum_{r_1=1}^{r^{(1)}} \sum_{r_2=1}^{r^{(2)}} \mathcal{C}(r_1, r_2, i, j) M_1(r_1, f) M_2(r_2, c), \tag{1}$$

where $\boldsymbol{\mathcal{C}} \in \mathbb{R}^{r^{(1)} \times r^{(2)} \times K_1 \times K_2}$, $\boldsymbol{M}_1 \in \mathbb{R}^{r^{(1)} \times F}$, $\boldsymbol{M}_2 \in \mathbb{R}^{r^{(2)} \times C}$, and $r^{(1)}$ and $r^{(2)}$ denote the tensor rank. Then given an input tensor $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{W \times H \times C}$ and an output tensor $\boldsymbol{\mathcal{Y}} \in \mathbb{R}^{W' \times H' \times F}$, the Tucker-2-format convolution is

performed as follows:

$$\mathcal{K}_1(w,h,r_2) = \sum_{c=1}^{C} M_2(r_2,c)\mathcal{X}(w,h,c), \tag{2}$$

$$\mathcal{K}_2(w',h',r_1) = \sum_{k_1=1}^{K_1}\sum_{k_2=1}^{K_2}\sum_{r_2=1}^{r^{(2)}} \mathcal{C}(r_1,r_2,k_1,k_2)\mathcal{K}_1(w,h,r_2), \tag{3}$$

$$\mathcal{Y}(w',h',f) = \sum_{r_1=1}^{r^{(1)}} M_1(r_1,f)\mathcal{K}_2(w',h',r_1), \tag{4}$$

where $\boldsymbol{\mathcal{K}}_1 \in \mathbb{R}^{W\times H\times r^{(2)}}$ and $\boldsymbol{\mathcal{K}}_2 \in \mathbb{R}^{W'\times H'\times r^{(1)}}$.

Method

Problem Formulation

To realize practical and high-performance low-rank compression, we aim to automatically find the optimal ranks to minimize the accuracy loss and maximize the hardware performance (e.g., inference speed). Mathematically, for an n-layer convolutional neural network, this process can be formulated as a constrained optimization problem as below:

$$\min_{\{\boldsymbol{\mathcal{W}}_i\}_{i=1}^n} \mathcal{L}(\{\boldsymbol{\mathcal{W}}_i\}) \quad \text{s.t.} \quad \sum_{i=1}^{n} \hbar(rank(\boldsymbol{\mathcal{W}}_i)) \le \varepsilon, \tag{5}$$

where $\{\boldsymbol{\mathcal{W}}_i\}_{i=1}^n$ represents the weights of all layers of n-layer CNN model, $rank(\cdot)$ is the function that returns the tensor ranks of the factorized tensor cores, and ε is the specific constraint for the practical hardware performance (e.g., latency). In addition, $\mathcal{L}(\cdot)$ and $\hbar(\cdot)$ denote the loss function and the layer-wise hardware performance, respectively.

Design Challenges

Solving the constrained problem described in Eq. 5 is non-trivial but facing two main challenges. First, because we cannot explicitly construct $rank(\cdot)$, the implicit mechanism that can automatically determine the rank setting is needed. To that end, some methods, such as approximation error-aware minimization and genetic algorithm, have been used to guide the search process. However, a common problem for the existing solutions is the insufficient exploration for the rank space. For instance, because the higher similarity between the original and the decomposed models does not necessarily mean higher accuracy, minimizing the approximation error will severely limit the exploration scope, causing unsatisfied compression performance. Second, the search process of the state-of-the-art rank determination works cannot be extended to consider the hardware performance constraint described in Eq. 5. More specifically, the underlying mechanisms of the existing automatic rank selection methods, by their nature, can only support the differentiable constraint such as compression ratio; while the practical hardware performance, e.g., the measured latency, is non-differentiable, making it challenging to extend the prior solutions to the hardware-aware format.

HALOC: Selecting Ranks as Architecture Search

The above analyzed limitations of the existing rank selection methods call for more efficient solutions. To that end, we propose HALOC, a novel hardware-aware automatic low-rank DNN compression technique. HALOC is built on a key observation – because low-rank compression aims to explore the structure-level redundancy of DNNs, the rank selection process essentially determines the architecture of the compressed models. Based on this perspective, automatic search for the suitable ranks can be interpreted as the automatic search for the proper low-rank structure, opening up the opportunities of designing new rank selection solution guided by the philosophy of neural architecture search (NAS) (Cai, Zhu, and Han 2018; Wu et al. 2019).

Motivated by this hidden connection between setting the layer-wise ranks and searching the network architecture, we propose to develop efficient low-rank compression with NAS-inspired automatic rank selection. Our key idea is iteratively sampling and evaluating different candidate rank settings to learn the most suitable one in a differentiable way. More specifically, as illustrated in Figure 1, the HALOC framework consists of the following operations.

Step-1. Constructing Low-Rank Search Space. We first build an over-parameterized network $\boldsymbol{\mathcal{N}}$ that consists of multiple candidate rank combinations. Notice that different from the case for NAS methods aiming to select from a group of candidate operators, the n-layer $\boldsymbol{\mathcal{N}}(\mathcal{T}_1, \cdots, \mathcal{T}_n)$ built for HALOC represents the ensemble of the candidate rank settings. More specifically, for the i-th layer of $\boldsymbol{\mathcal{N}}$ as $\mathcal{T}_i$, it consists of $m_1 m_2$ decomposition candidates as $\tau_{i,j}$ with rank setting $(r^{(1)}_{i,j_1}, r^{(2)}_{i,j_2})$, where $j_1 = 1, 2, ..., m_1$, $j_2 = 1, 2, ..., m_2$, and $j = 1, 2, ...m_1 m_2$. At the setup stage of automatic search, each $\tau_{i,j}$ is initialized via performing Tucker-2 decomposition on the i-th layer of the uncompressed model with rank set as $(r^{(1)}_{i,j_1}, r^{(2)}_{i,j_2})$.

Step-2. Updating Probabilities & Weights. Upon the construction and initialization of $\boldsymbol{\mathcal{N}}$, we then alternately update the parameters of $\tau_{i,j}$ and the corresponding selection probability $p_{i,j}$ for the i-th layer. To be specific, because $p_{i,j}$ is calculated as $\boldsymbol{p}_i = Softmax(\boldsymbol{\alpha}_i)$, where $p_{i,j} \in \boldsymbol{p}_i$ and $\boldsymbol{\alpha}_i$ is an learnable vector, the update of all $p_{i,j}$'s can be simultaneously performed via using the backward propagation on the validation dataset. On the other hand, HALOC updates the weights of $\tau_{i,j}$ in a selective way. To reduce the computational cost, guided by the selection probability $p_{i,j}$, each time only one $\tau_{i,j}$ is sampled and updated per layer, and this weight update process is based on the loss function defined on the training dataset. After multiple iterations of alternated update for probabilities and weights, the final selected decomposed candidate for the i-th layer is the $\tau_{i,j}$ with the largest $p_{i,j}$.

Considering Hardware Constraints. As outlined in Eq. 5, the constraints on the hardware performance should be taken into account when designing low-rank compression technique towards practical applications. Unlike the existing automatic low-rank compression works that cannot properly include hardware performance into the design phase, HALOC enjoys an attractive benefit of its inherent com-

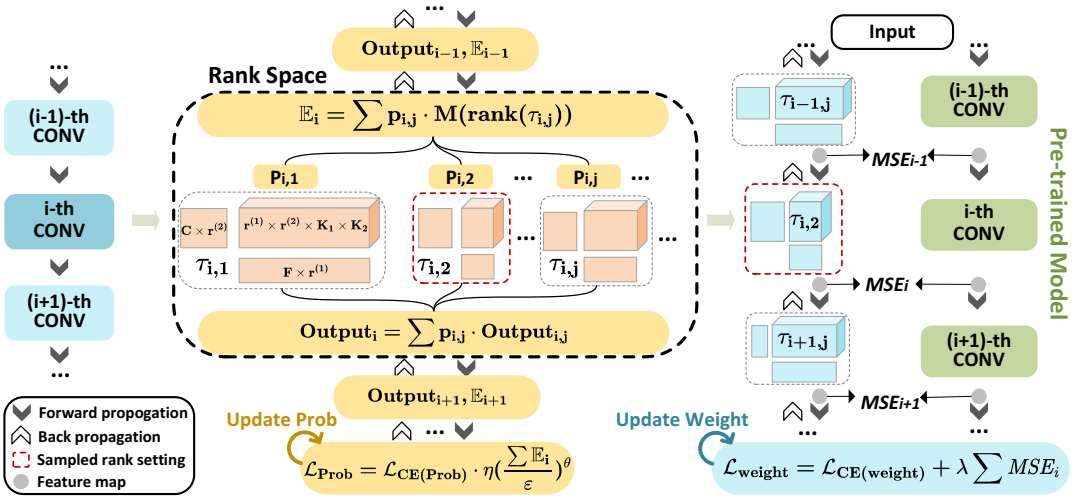


Figure 1: The automatic rank selection process of HALOC via alternately updating rank selection probability and model weight.

patibility for hardware-awareness. More specifically, considering the hardware performance (e.g., latency) is non-differentiable, we use a prediction model to estimate the expected hardware performance of the i-th layer as follows:

$$\mathbb{E}_i = \sum_{j=1}^{m_1 m_2} p_{i,j} \cdot M(rank(\tau_{i,j})), \tag{6}$$

where $M(\cdot)$ denotes the *layer-wise* hardware performance for the decomposition candidate $\tau_{i,j}$ with rank setting $(r_{i,j_1}^{(1)}, r_{i,j_2}^{(2)})$. Here the hardware performance for one candidate can be either pre-measured from the target computing devices or estimated from a regression model. Essentially, the layer-wise prediction model described in Eq. 6 can be viewed as the weighted sum of the performance of the decomposition candidates for the current layer, and it is then integrated to the process for updating the selection probabilities as follows:

$$\mathcal{L}_{Prob} = \mathcal{L}_{CE(Prob)} \cdot \eta(\frac{\sum_{i=1}^{n} \mathbb{E}_i}{\varepsilon})^{\theta}, \tag{7}$$

where $\mathcal{L}_{CE(Prob)}$ is the cross-entropy loss on the validation dataset, and η and θ are the hyperparameters that adjust the impact of hardware constraints on the overall search process (Wu et al. 2019).

Questions to be Answered

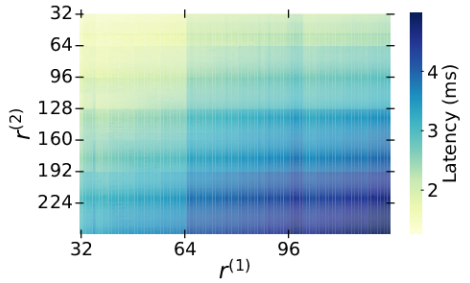
As outlined above, the HALOC framework can be developed from the perspective of architecture search. However, as we will further analyze in this subsection, because HALOC aims to perform automatic low-rank compression, a task that is essentially different from NAS, it is facing several unique design challenges when searching in the rank space. To address these issues and realize automatic rank selection efficiently, next we explore to answer two important questions.

Question 1: *How should we set the proper search scope to realize sufficient exploration in the rank space with affordable search cost?*

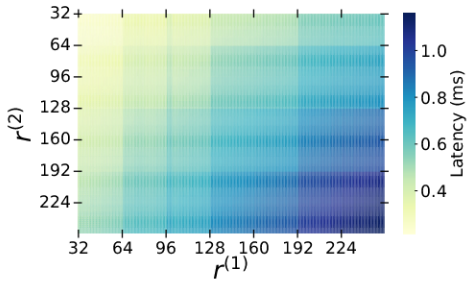
Analysis. A very key challenge for HALOC is the extremely huge search space, which is much larger than the scope explored in the existing NAS works. For instance, there are only 1.5×10^{17} potential network architectures when searching an 18-layer CNN using NAS (Wu et al. 2019); while there exist 2.2×10^{71} candidates of rank combination when performing low-rank compression on a ResNet-18 model. In other words, the rank space for low-rank compression is much larger than the architecture space for NAS, hindering the automatic rank selection in a timely and efficient way.

Essentially, the ultra-large search space of HALOC results from two sources. First, for each layer, the number of rank candidates is inherently much more than that of operator candidates. For instance, the uncompressed layer (layer3.1.conv1) in ResNet-18 model has full rank size as 256, meaning that there exist at least 256 possible ranks for $r^{(1)}$ or $r^{(2)}$ can be selected for low-rank compression; while the operator for building one layer are only considered and selected from a limited set, e.g., around 10 candidates as indicated in (Liu, Simonyan, and Yang 2018; Cai, Zhu, and Han 2018; Wu et al. 2019). Second, because it is very common that multiple rank modes are needed in low-rank compression, e.g., a Tucker-2 decomposed layer is determined by two ranks ($r^{(1)}$ and $r^{(2)}$ in Eq. 1), the corresponding combinatorial effect further drastically enlarges the overall to-be-explored rank space.

Our Proposal. Based on the above analysis, we propose to perform efficient search in the rank space to achieve good balance between the efficiency and globality of the exploration. First, we reduce the numbers of the rank candidates in a hardware-aware way. More specifically, we analyze the impact of different rank settings of single layer on the measured latency, and we discover that many ranks, though corresponding to different FLOPs, bring very similar latency on the hardware platforms. For instance, as illustrated in Figure 2, when performing Tucker-2 decomposition on a convolutional layer with different rank settings, obvious latency change can only be observed when $r^{(1)}$ and $r^{(2)}$ increase by 32 or 64 – a phenomenon that exists across different types of computing platforms. We hypothesize that the most probable reason causing this phenomenon is the under-utilization



(a) Layer3.0.conv on Nvidia RTX 2080. Batch size is 128.



(b) Layer3.0.conv2 on Nvidia Tesla V100. Batch size is 128.

Figure 2: The heat map of the measured latency on the practical hardware for Tucker-2 format layers of ResNet-18 with different rank settings. Batch processing is used for stable measurement.

of hardware resource for some rank settings. As analyzed in (Qin et al. 2020), when the size of the (decomposed) layer cannot match the size of computing resource (e.g., arrays of processing elements in ASIC accelerator and 32-thread wrap in desktop GPU) very well, many computing units would be idle and under-utilized, causing that the execution of the two low-rank layers with different ranks consumes the same clock cycles. Motivated by this discovery, we propose the following design principle to only perform search on the rank values that correspond to obvious latency changes:

Design Principle-1: *To make good balance between search cost and rank granularity, the rank candidates in HALOC is set as the multiples of a constant (typically 32).*

Notice that when the number of input/output channels of a convolutional layer is small (e.g., 64 or 32), the suggested constant can be set as 16 or 8 to ensure the sufficient coverage for searching in the rank space. More importantly, when the target compressed model size/FLOP is already pre-known, the setting for this constant can be automatically calculated according to the pre-set compression ratio. Please refer to **Appendix** for more details.

Second, in addition to specify the search granularity for $r^{(1)}$ and $r^{(2)}$, we also identify their suitable relationship to further reduce the search space. Based on our in-depth analysis (details described in **Appendix**), we propose the following design principle:

Design Principle-2: *For a Tucker-2-format layer, equal rank setting ($r^{(1)} = r^{(2)}$) can be adopted to simplify the rank search process with good approximation performance.*

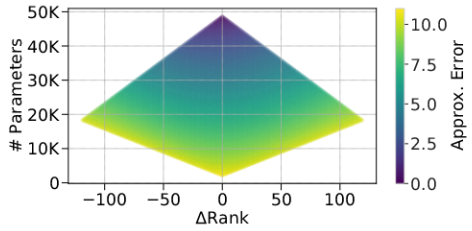


Figure 3: Approximation error of a decomposed convolutional layer (layer2.1.conv2) in ResNet-18 model with different rank discrepancies ($\Delta Rank = r^{(1)} - r^{(2)}$) and target numbers of parameters after decomposition.

Figure 3 empirically illustrates the approximation errors with different rank discrepancy ($\Delta Rank = r^{(1)} - r^{(2)}$) given the same number of weights for a decomposed layer. It is seen that setting $r^{(1)}$ equal to $r^{(2)}$ indeed brings much smaller approximation error than most of other configurations, leading to the good balance between the simplicity and quality of rank search. Overall, guided by these design principles, when using HALOC to compress ResNet-18 model, the number of the possible rank combinations can be reduced from 2.2×10^{71} to 7.8×10^{13}, leading to even smaller search space than that of the existing NAS methods.

Question #2: *What is the proper scheme to mitigate the interference between different selected rank settings?*

Analysis. Another issue that may affect the performance of HALOC is the potential interference between different low-rank structure candidates in the search process. To be specific, because a decomposition candidate $\tau_{i,j}$ in the i-th layer may be connected to different $\tau_{i+1,j}$'s of the $(i+1)$-th layer in different sampling trials, the discrepancy between the gradient directions of τ_{i+1,j_1} and τ_{i+1,j_2} imposed on $\tau_{i,j}$, if large, would bring interference on the weight/probabilities update process, causing insufficient training and/or slow convergence. Notice that similar phenomenon has also been observed in the weight-sharing NAS (Shu, Wang, and Cai 2019; Zhang et al. 2020).

Our Proposal. To address this challenge, we propose to use the prior information of the uncompressed model to mitigate the potential interference. Our key idea lies in the hypothesis that the higher similarity between τ_{i+1,j_1} and τ_{i+1,j_2} renders higher similarity between their gradients directions imposed on $\tau_{i,j}$. Therefore, considering each $\tau_{i+1,j}$ can be roughly interpreted as a type of approximation for the original uncompressed weight tensor $\mathcal{W}_{i+1}$ in the $(i+1)$-th layer, we propose to use the pre-known information of $\mathcal{W}_{i+1}$ to improve the similarity between τ_{i+1,j_1} and τ_{i+1,j_2}. To be specific, we can regularize the search process to force both τ_{i+1,j_1} and τ_{i+1,j_2} to approach $\mathcal{W}_{i+1}$, making them become more similar.

In general, this uncompressed layer-approaching strategy can be realized via two ways: we can make either the weights or the output feature maps of $\tau_{i+1,j}$ approach those of $\mathcal{W}_{i+1}$. HALOC adopts the feature map-based approaching solution because of two reasons. **1) Low computational costs.** To measure the similarity between the weights of $\tau_{i+1,j}$ and $\mathcal{W}_{i+1}$, we need to reconstruct the approximated

$\mathcal{W}_{i+1}$ from $\tau_{i+1,j}$ via using Eq. 1. Unfortunately, such reconstruction operation is not well supported by the current version of PyTorch/Tensorflow, causing very slow execution speed. On the other hand, calculating the output feature maps of $\mathcal{W}_{i+1}$ is equivalent to performing forward prorogation for the original model, an execution that is much faster and well supported by GPU acceleration. **2) Rich information.** As indicated in (Lin et al. 2020) the feature maps naturally capture and contain the rich and critical information for both model and data. Compared with weight-approaching strategy that only reflects static model characteristics, approaching feature maps can provide additional dynamic data-aware alignment. We observe that such feature map-preferred philosophy is also advocated and adopted in a set of channel pruning works (Hou et al. 2022; Sui et al. 2021; Tang et al. 2020; Lin et al. 2020), demonstrating the importance and usefulness of feature map information. Therefore, we use the following mean square error (MSE)-based loss to measure the feature map similarity:

$$\mathcal{L}_{approach} = \sum_{i=1}^{n} MSE(Fmap_{decomp,i}, Fmap_{org,i}), \quad (8)$$

where $Fmap_{org,i}$ and $Fmap_{decomp,i}$ denote the feature maps of the original weight tensor and the sampled $\tau_{i,j}$ in the i-th layer, respectively. Then the loss function for updating the weights is as follows:

$$\mathcal{L}_{weight} = \mathcal{L}_{CE(weight)} + \lambda \mathcal{L}_{approach}, \quad (9)$$

where $\mathcal{L}_{CE(weight)}$ is the cross-entropy loss defined on the training dataset, and λ is the scaling parameter that controls the impact of feature map-approaching loss.

Experiments

We evaluate the performance of HALOC for compressing different CNN models on CIFAR10 (Krizhevsky, Hinton et al. 2009), and ImageNet (Deng et al. 2009) datasets. We also measure the practical latency of the automatic compressed models on different types of hardware platforms, including desktop GPU, embedded GPU and ASIC accelerator.

Training Details. We use the standard SGD optimizer with Nesterov momentum as 0.9 for model training. The learning rates are initilized as 0.1 and 0.01 for CIFAR-10 and ImageNet, respectively, and they are then scaled down by 0.2 every 55 epochs. In addition, batch size and weight decay are set as 256 and 0.0001, respectively.

CIFAR-10 Results

Table 1 shows the evaluation results for compressing different CNN models on CIFAR-10 dataset. For compressing ResNet-20 model, our method can achieve even 0.07% higher top-1 accuracy than the baseline model with 72.6% FLOPs reduction and 76.1% model size reduction. Compared with ALDS (Liebenwein et al. 2021), the state-of-the-art automatic low-rank compression work, HALOC achieves 0.35% accuracy increase with higher computational and memory cost reductions, demonstrating the outstanding compression performance of our proposed approach.

Method	Comp. Type	Auto. Rank	Top-1 (%)	FLOPs (↓%)	Params. (↓%)
ResNet-20	Baseline	-	91.25	-	-
HALOC	Low-rank	✓	**91.32**	**72.20**	**76.10**
ALDS	Low-rank	✓	90.92	67.86	74.91
LCNN	Low-rank	✓	90.13	66.78	65.38
PSTR-S	Low-rank	✓	90.80	-	60.87
Std. Tucker	Low-rank	✗	87.41	-	61.54
VGG-16	Baseline	-	92.78	-	-
HALOC	Low-rank	✓	**93.16**	**86.44**	**98.56**
ALDS	Low-rank	✓	92.67	86.23	95.77
LCNN	Low-rank	✓	92.72	85.47	91.14
DECORE	Pruning	-	92.44	81.50	96.60
Spike-Thrift	Pruning	-	91.79	-	97.01

Table 1: Comparison with different compression approaches for ResNet-20 and VGG-16 on CIFAR-10. ALDS (Liebenwein et al. 2021), LCNN (Idelbayev and Carreira-Perpinán 2020), PSTR-S (Li et al. 2022), Std. Tucker (Kim et al. 2016), DECORE (Alwani, Wang, and Madhavan 2022), Spike-Thrift (Kundu et al. 2021).

Method	Comp. Type	Auto. Rank	Top-1 (%)	Top-5 (%)	FLOPs (↓%)	Params. (↓%)
ResNet-18	Baseline	-	69.75	89.08	-	-
HALOC	Low-rank	✓	**70.65**	**89.42**	**66.16**	63.64
HALOC	Low-rank	✓	70.14	89.38	63.81	**71.31**
ALDS	Low-rank	✓	69.22	89.03	43.51	66.70
TETD	Low-rank	✗	-	89.08	59.51	-
Stable EPC	Low-rank	✓	-	89.08	59.51	-
MUSCO	Low-rank	✗	69.29	88.78	58.67	-
CHEX	Pruning	-	69.60	-	43.38	-
EE	Pruning	-	68.27	88.44	46.60	-
SCOP	Pruning	-	69.18	88.89	38.80	39.30
MobileNetV2	Baseline	-	71.85	90.33	-	-
HALOC	Low-rank	✓	**70.98**	**89.77**	24.84	40.03
HALOC	Low-rank	✓	66.37	87.02	**45.65**	**62.59**
ALDS	Low-rank	✓	70.32	89.60	11.01	32.97
HOSA	Pruning	-	64.43	-	43.65	27.13
DCP	Pruning	-	64.22	-	44.75	25.93
FT	Pruning	-	70.12	89.48	20.23	21.31

Table 2: Comparison with different compression approaches for ResNet-18 and MobileNetV2 on ImageNet. ALDS (Liebenwein et al. 2021), TETD (Yin et al. 2021b), Stable EPC (Phan et al. 2020), MUSCO (Gusak et al. 2019), CHEX (Hou et al. 2022), EE (Zhang, Gao, and Huang 2021), SCOP (Tang et al. 2020), HOSA (Chatzikonstantinou et al. 2020), DCP (Zhuang et al. 2018), FT (He, Zhang, and Sun 2017).

For compressing VGG-16 model, HALOC also shows high compression performance. It brings 0.38% higher accuracy than the original uncompressed model with 86.44% FLOPs reduction and 98.56% model size reduction. Compared with the state-of-the-art pruning and low-rank compression works, HALOC consistently achieves higher accuracy with more aggressive compression efforts.

Hardware	Method	ResNet-18				MobileNetV2			
		Top-1 (%)	Top-5 (%)	FLOPs (M)	Throughput (images/s)	Top-1 (%)	Top-5 (%)	FLOPs (M)	Throughput (images/s)
NVIDIA Tesla V100	Original	69.75	89.08	1819.07	4362.1	71.85	90.33	314.19	3877.3
	HALOC	69.75	88.93	**553.13**	**6360.5**	70.86	89.77	**245.52**	**3993.6**
NVIDIA Jetson TX2	Original	69.75	89.08	1819.07	86.3	71.85	90.33	314.19	112.1
	HALOC	70.14	89.38	**658.26**	**151.0**	70.80	89.55	**240.99**	**117.0**
ASIC Eyeriss	Original	69.75	89.08	1819.07	121.4	71.85	90.33	314.19	496.3
	HALOC	70.65	89.42	**615.62**	**247.0**	70.83	89.65	**229.13**	**590.2**

Table 3: Measured Speedup for compressed ResNet-18 and MobileNetV2 on different computing platforms. Hardware-aware automatic rank selection is adopted in the low-rank compression process.

ImageNet Results

Table 2 shows the performance of different compression methods on ImageNet dataset. For compressing ResNet-18 model, HALOC achieves 0.9% higher top-1 accuracy than the baseline model with 66.16% FLOPs reduction and 63.64% model size reduction. Compared with the state-of-the-art automatic low-rank compression method ALDS, HALOC enjoys 1.4% higher top-1 accuracy with much lower computational costs. Compared with the state-of-the-art pruning work CHEX (Hou et al. 2022), HALOC also achieves 0.54% accuracy increase with much fewer FLOPs.

Besides, our approach also shows impressive performance for compressing MobileNetV2, a task that is conventionally challenging for low-rank compression approach. From Table 2 it is seen that HALOC can achieve higher performance than the existing low-rank compression and pruning works. In particular, compared with ALDS, the method that reports the best performance of compressing MobileNetV2 among all the existing low-rank solutions, HALOC shows 0.66% top-1 accuracy increase with higher FLOPs reduction and model size reduction.

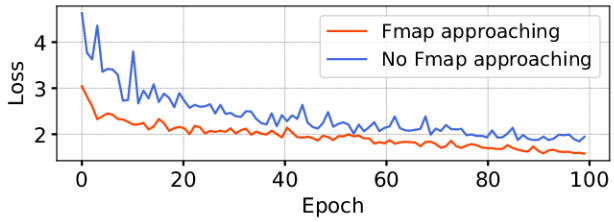
Practical Speedups on Hardware Platforms

We also measure the practical speedups brought by our hardware-aware solution on various computing hardware, including Nvidia Tesla V100, Nvidia Jetson TX2, and ASIC accelerator Eyeriss (Chen, Emer, and Sze 2016). Here the performance of Eyeriss is reported via using Timeloop (Parashar et al. 2019) with 45nm CMOS technology setting. As shown in Table 3, the ResNet-18 and MobileNetV2 models that are compressed by HALOC shows considerable measured speedups across different platforms, demonstrating the practical effectiveness of our proposed hardware-aware low-rank compression solution.

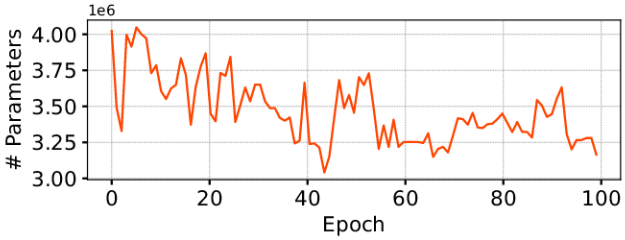
Analysis & Discussion

To obtain in-depth understanding and verify the effectiveness of HALOC, we perform some ablation study and experimental analysis for compressing ResNet-18 model on ImageNet dataset. First, we examine the impact of the proposed feature map-approaching strategy on the search process. As shown in Figure 4 (a), the use of feature map-based regularization brings significant lower training loss, which further translates to higher accuracy for the final compressed model. In addition, Figure 4 (b) shows the change of the total number of parameters during the search process. It is seen

(a) The curve of training loss ($\mathcal{L}_{weight}$)



(b) The change of the model size in the rank search process



(c) The final rank distribution after automatic rank search

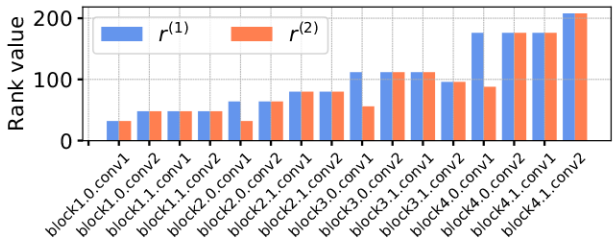


Figure 4: Experimental analysis for compressing ResNet-18 on ImageNet dataset using HALOC.

that HALOC indeed dynamically explores the rank space towards reducing the overall model complexity. The final rank distribution of the automatically compressed ResNet-18 model is visualized in Figure 4 (c).

Conclusion

This paper proposes HALOC, a hardware-aware automatic low-rank compression framework for compact DNN models. By interpreting rank selection as architecture search process, HALOC enables efficient automatic rank determination in a differentiable way, realizing high-performance low-rank compression. Experimental results on both algorithm and hardware aspects demonstrate the effectiveness of our proposed approach.

References

Alwani, M.; Wang, Y.; and Madhavan, V. 2022. DECORE: Deep Compression With Reinforcement Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 12349–12359.

Astrid, M.; and Lee, S.-I. 2017. Cp-decomposition with tensor power method for convolutional neural networks compression. In *2017 IEEE International Conference on Big Data and Smart Computing (BigComp)*, 115–118. IEEE.

Cai, H.; Zhu, L.; and Han, S. 2018. Proxylessnas: Direct neural architecture search on target task and hardware. *arXiv preprint arXiv:1812.00332*.

Chatzikonstantinou, C.; Papadopoulos, G. T.; Dimitropoulos, K.; and Daras, P. 2020. Neural network compression using higher-order statistics and auxiliary reconstruction losses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 716–717.

Chen, Y.-H.; Emer, J.; and Sze, V. 2016. Eyeriss: A spatial architecture for energy-efficient dataflow for convolutional neural networks. *ACM SIGARCH Computer Architecture News*, 44(3): 367–379.

De Lathauwer, L.; De Moor, B.; and Vandewalle, J. 2000. A multilinear singular value decomposition. *SIAM journal on Matrix Analysis and Applications*, 21(4): 1253–1278.

Deng, C.; Sui, Y.; Liao, S.; Qian, X.; and Yuan, B. 2021. GoSPA: An Energy-efficient High-performance Globally Optimized SParse Convolutional Neural Network Accelerator. In *2021 ACM/IEEE 48th Annual International Symposium on Computer Architecture (ISCA)*, 1110–1123.

Deng, C.; Sun, F.; Qian, X.; Lin, J.; Wang, Z.; and Yuan, B. 2019a. TIE: Energy-efficient tensor train-based inference engine for deep neural network. In *Proceedings of the 46th International Symposium on Computer Architecture*, 264–278.

Deng, C.; Yin, M.; Liu, X.-Y.; Wang, X.; and Yuan, B. 2019b. High-performance Hardware Architecture for Tensor Singular Value Decomposition: Invited Paper. In *2019 IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, 1–6.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Denton, E. L.; Zaremba, W.; Bruna, J.; LeCun, Y.; and Fergus, R. 2014. Exploiting linear structure within convolutional networks for efficient evaluation. *Advances in neural information processing systems*, 27.

Dong, P.; Wang, S.; Niu, W.; Zhang, C.; Lin, S.; Li, Z.; Gong, Y.; Ren, B.; Lin, X.; and Tao, D. 2020. Rtmobile: Beyond real-time mobile acceleration of rnns for speech recognition. In *2020 57th ACM/IEEE Design Automation Conference (DAC)*, 1–6. IEEE.

Golub, G. H.; and Van Loan, C. F. 2013. *Matrix computations*. JHU press.

Gusak, J.; Kholiavchenko, M.; Ponomarev, E.; Markeeva, L.; Blagoveschensky, P.; Cichocki, A.; and Oseledets, I. 2019. Automated multi-stage compression of neural networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 0–0.

Han, S.; Pool, J.; Tran, J.; and Dally, W. 2015. Learning both weights and connections for efficient neural network. *Advances in neural information processing systems*, 28.

Harshman, R. A.; et al. 1970. Foundations of the PARAFAC procedure: Models and conditions for an" explanatory" multimodal factor analysis.

He, Y.; Zhang, X.; and Sun, J. 2017. Channel pruning for accelerating very deep neural networks. In *Proceedings of the IEEE international conference on computer vision*, 1389–1397.

Hitchcock, F. L. 1928. Multiple invariants and generalized rank of a p-way matrix or tensor. *Journal of Mathematics and Physics*, 7(1-4): 39–79.

Hou, Z.; Qin, M.; Sun, F.; Ma, X.; Yuan, K.; Xu, Y.; Chen, Y.-K.; Jin, R.; Xie, Y.; and Kung, S.-Y. 2022. CHEX: CHannel EXploration for CNN Model Compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12287–12298.

Idelbayev, Y.; and Carreira-Perpinán, M. A. 2020. Low-rank compression of neural nets: Learning the rank of each layer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8049–8059.

Kim, Y.-D.; Park, E.; Yoo, S.; Choi, T.; Yang, L.; and Shin, D. 2016. Compression of deep convolutional neural networks for fast and low power mobile applications. *arXiv preprint arXiv:1511.06530v2*.

Klema, V.; and Laub, A. 1980. The singular value decomposition: Its computation and some applications. *IEEE Transactions on automatic control*, 25(2): 164–176.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images.

Kundu, S.; Datta, G.; Pedram, M.; and Beerel, P. A. 2021. Spike-Thrift: Towards Energy-Efficient Deep Spiking Neural Networks by Limiting Spiking Activity via Attention-Guided Compression. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 3953–3962.

Lebedev, V.; Ganin, Y.; Rakhuba, M.; Oseledets, I.; and Lempitsky, V. 2014. Speeding-up convolutional neural networks using fine-tuned cp-decomposition. *arXiv preprint arXiv:1412.6553*.

Li, N.; Pan, Y.; Chen, Y.; Ding, Z.; Zhao, D.; and Xu, Z. 2022. Heuristic rank selection with progressively searching tensor ring network. *Complex & Intelligent Systems*, 8(2): 771–785.

Liebenwein, L.; Maalouf, A.; Feldman, D.; and Rus, D. 2021. Compressing neural networks: Towards determining the optimal layer-wise decomposition. *Advances in Neural Information Processing Systems*, 34: 5328–5344.

Lin, M.; Ji, R.; Wang, Y.; Zhang, Y.; Zhang, B.; Tian, Y.; and Shao, L. 2020. Hrank: Filter pruning using high-rank

feature map. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1529–1538.

Liu, H.; Simonyan, K.; and Yang, Y. 2018. Darts: Differentiable architecture search. *arXiv preprint arXiv:1806.09055*.

Novikov, A.; Podoprikhin, D.; Osokin, A.; and Vetrov, D. P. 2015. Tensorizing neural networks. *Advances in neural information processing systems*, 28.

Pan, Y.; Xu, J.; Wang, M.; Ye, J.; Wang, F.; Bai, K.; and Xu, Z. 2019. Compressing recurrent neural networks with tensor ring for action recognition. 4683–4690.

Parashar, A.; Raina, P.; Shao, Y. S.; Chen, Y.-H.; Ying, V. A.; Mukkara, A.; Venkatesan, R.; Khailany, B.; Keckler, S. W.; and Emer, J. 2019. Timeloop: A systematic approach to dnn accelerator evaluation. In *2019 IEEE international symposium on performance analysis of systems and software (ISPASS)*, 304–315. IEEE.

Phan, A.-H.; Sobolev, K.; Sozykin, K.; Ermilov, D.; Gusak, J.; Tichavskỳ, P.; Glukhov, V.; Oseledets, I.; and Cichocki, A. 2020. Stable low-rank tensor decomposition for compression of convolutional neural network. In *European Conference on Computer Vision*, 522–539. Springer.

Qin, E.; Samajdar, A.; Kwon, H.; Nadella, V.; Srinivasan, S.; Das, D.; Kaul, B.; and Krishna, T. 2020. Sigma: A sparse and irregular gemm accelerator with flexible interconnects for dnn training. In *2020 IEEE International Symposium on High Performance Computer Architecture (HPCA)*, 58–70. IEEE.

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

Shu, Y.; Wang, W.; and Cai, S. 2019. Understanding architectures learnt by cell-based neural architecture search. *arXiv preprint arXiv:1909.09569*.

Sui, Y.; Yin, M.; Xie, Y.; Phan, H.; Aliari Zonouz, S.; and Yuan, B. 2021. CHIP: CHannel Independence-based Pruning for Compact Neural Networks. In Ranzato, M.; Beygelzimer, A.; Dauphin, Y.; Liang, P.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems*, volume 34, 24604–24616. Curran Associates, Inc.

Tang, Y.; Wang, Y.; Xu, Y.; Tao, D.; Xu, C.; Xu, C.; and Xu, C. 2020. Scop: Scientific control for reliable neural network pruning. *Advances in Neural Information Processing Systems*, 33: 10936–10947.

Tucker, L. R. 1966. Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31(3): 279–311.

Wang, D.; Zhao, G.; Li, G.; Deng, L.; and Wu, Y. 2019. Lossless Compression for 3DCNNs Based on Tensor Train Decomposition.

Wen, W.; Xu, C.; Wu, C.; Wang, Y.; Chen, Y.; and Li, H. 2017. Coordinating filters for faster deep neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*, 658–666.

Wu, B.; Dai, X.; Zhang, P.; Wang, Y.; Sun, F.; Wu, Y.; Tian, Y.; Vajda, P.; Jia, Y.; and Keutzer, K. 2019. Fbnet: Hardware-aware efficient convnet design via differentiable neural architecture search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10734–10742.

Xiang, L.; Yin, M.; Zhang, C.; Sukumaran-Rajam, A.; Sadayappan, P.; Yuan, B.; and Tao, D. 2022. TDC: Towards Extremely Efficient CNNs on GPUs via Hardware-Aware Tucker Decomposition. *arXiv preprint arXiv:2211.03715*.

Yin, M.; Liao, S.; Liu, X.-Y.; Wang, X.; and Yuan, B. 2020. Compressing recurrent neural networks using hierarchical tucker tensor decomposition. *arXiv preprint arXiv:2005.04366*.

Yin, M.; Liao, S.; Liu, X.-Y.; Wang, X.; and Yuan, B. 2021a. Towards extremely compact rnns for video recognition with fully decomposed hierarchical tucker structure. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12085–12094.

Yin, M.; Phan, H.; Zang, X.; Liao, S.; and Yuan, B. 2022a. Batude: Budget-aware neural network compression based on tucker decomposition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 1.

Yin, M.; Sui, Y.; Liao, S.; and Yuan, B. 2021b. Towards Efficient Tensor Decomposition-Based DNN Model Compression With Optimization Framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10674–10683.

Yin, M.; Sui, Y.; Yang, W.; Zang, X.; Gong, Y.; and Yuan, B. 2022b. HODEC: Towards Efficient High-Order DEcomposed Convolutional Neural Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 12299–12308.

Zhang, D.; Yang, J.; Ye, D.; and Hua, G. 2018. Lq-nets: Learned quantization for highly accurate and compact deep neural networks. In *Proceedings of the European conference on computer vision (ECCV)*, 365–382.

Zhang, Y.; Gao, S.; and Huang, H. 2021. Exploration and Estimation for Model Compression. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 487–496.

Zhang, Y.; Lin, Z.; Jiang, J.; Zhang, Q.; Wang, Y.; Xue, H.; Zhang, C.; and Yang, Y. 2020. Deeper insights into weight sharing in neural architecture search. *arXiv preprint arXiv:2001.01431*.

Zhuang, Z.; Tan, M.; Zhuang, B.; Liu, J.; Guo, Y.; Wu, Q.; Huang, J.; and Zhu, J. 2018. Discrimination-aware Channel Pruning for Deep Neural Networks. In Bengio, S.; Wallach, H.; Larochelle, H.; Grauman, K.; Cesa-Bianchi, N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Appendix

Rank Space with Pre-set Compression Ratio. *Design Principle-1* of the main manuscript discusses the suggested empirical setting of the rank granularity. Here we present a more analytic solution for a special case – when the target model compression ratio is pre-known. Our key idea is to first use the overall compression ratio to obtain a relaxed range for the estimated layer-wise rank value reduction (as compared to full rank), which can be then used to construct the rank space, and then we leverage the rank search process described in the main manuscript to identify the accurate rank setting that satisfies with the compression ratio requirement.

More specifically, suppose the overall target compression ratio is α, then the average compression ratio of each layer is also first roughly set as α. For a CONV layer with weight tensor $\mathcal{W} \in \mathbb{R}^{F\times C\times K_1\times K_2}$, the number of parameters for original CONV and Tucker-2 Decomposed CONV can be calculated as $n_{org} = FCK_1K_2$ and $n_{tucker} = Fr^{(1)} + Cr^{(2)} + r^{(1)}r^{(2)}K_1K_2$, respectively. Then we have

$$\begin{cases} n_{tucker} = \dfrac{C^2}{\alpha_{rank}} + \dfrac{F^2}{\alpha_{rank}} + \dfrac{C^2F^2}{\alpha_{rank}^2} \\ \dfrac{n_{org}}{\alpha} = \dfrac{FCK_1K_2}{\alpha} = n_{tucker} \end{cases}$$
$$\alpha_{rank} = \frac{2FCK_1K_2}{\sqrt{(C^2+F^2)^2 + \frac{4(FCK_1K_2)^2}{\alpha}} - C^2 - F^2}, \tag{10}$$

where α_{rank} is a scaling parameter which represents the reduction of the assigned ranks from the original full rank value, $r^{(1)} = \frac{F}{\alpha_{rank}}$ and $r^{(2)} = \frac{C}{\alpha_{rank}}$. Then we can use Algorithm 1 to determine the search rank space for Tucker-2 decomposition.

Determining Discrepancy between $r^{(1)}$ and $r^{(2)}$. To explore the best suitable $|\Delta Rank = r^{(1)} - r^{(2)}|$, we aim to minimize the approximation error given the same number of post-decomposition parameters. According to the Higher-Order Singular Value Decomposition (HOSVD) (Hitchcock 1928; De Lathauwer, De Moor, and Vandewalle 2000), we use the percentage of the sum of the truncated singular values to evaluate the degree of information preservation after decomposition as follows:

$$\rho = \frac{\sum_{i=1}^{r^{(1)}} s_{1,i}}{\sum_{i=1}^{F} s_{1,i}} \cdot \frac{\sum_{j=1}^{r^{(2)}} s_{2,j}}{\sum_{j=1}^{C} s_{2,j}}, \tag{11}$$

where $s_{1,i}$ and $s_{2,j}$ indicate the singular values, and ρ is the percentage of the sum of the truncated singular values. Without loss of generality, we simplify the calculation by assuming that the $s_{1,i} = s_{2,j}$, then we have

$$\rho = \frac{r^{(1)}s_{1,1}}{Fs_{1,1}} \cdot \frac{r^{(2)}s_{2,1}}{Cs_{2,1}} = \frac{r^{(1)}r^{(2)}}{CF}. \tag{12}$$

Recall that our goal can be described as the following optimization problem:

$$\begin{aligned} &\max \rho(r^{(1)}, r^{(2)}) \\ &\text{s.t. } n_{tucker} = Cr^{(2)} + K_1K_2r^{(1)}r^{(2)} + Fr^{(1)}. \end{aligned} \tag{13}$$

From Eq. 12 and the constraint of the above optimization problem, we can have:

$$\begin{aligned} \rho &= \frac{n_{tucker} - Fr^{(1)} - Cr^{(2)}}{K_1K_2CF} \\ \Rightarrow \frac{\partial\rho}{\partial r^{(1)}} &= \frac{-1}{K_1K_2C}, \ \frac{\partial\rho}{\partial r^{(2)}} = \frac{-1}{K_1K_2F}. \end{aligned} \tag{14}$$

Consider 1) Eq. 14 shows that ρ monotonically decreases when $r^{(1)}$ or $r^{(2)}$ increases; and 2) when $r^{(1)}$ or $r^{(2)}$ increases, the constant n_{tucker} means the corresponding $r^{(2)}$ or $r^{(1)}$ decreases, increasing the value of $|\Delta Rank|$. We can find that smaller $|\Delta Rank|$ brings larger ρ, i.e., more information is preserved after decomposition. increases. Therefore, when constructing the rank search space, we suggest $|\Delta Rank|$ should be small, e.g., $r^{(1)} = r^{(2)}$.

Algorithm 1 Rank Search Space Determination

Inputs: Overall compression ratio α,
List of the number of input channels $\{C_i\}$,
List of the number of out channels $\{F_i\}$,
Output: Low-rank space $\{\{(r_{i,j}^{(1)}, r_{i,j}^{(2)})\}\}$.
# Initialize step size
if $\max(\{C_i\} \cup \{F_i\}) \geq 128$ **then**
  # Corresponding number of channels is
  # {16, 32, 64, 128, 256, 512}
  $\{s_k\} \leftarrow \{4, 8, 16, 16, 32, 32\}$
else
  # For the number of channels less than 128,
  # the setp size is 4
  $\{s_k\} \leftarrow \{4\}$
end if
Calculate α_{rank} by Eq. 10
# Generating rank space for each layer
for each $i \in [0, len(\{C_i\})]$ **do**
  $t \leftarrow min(C_i, F_i),\ k \leftarrow \frac{t}{16} - 1$
  # Relaxing the range of scaling parameter α_{rank}
  Generate $\{r_j\}$ from $[\frac{t}{\alpha_{rank}+2}, \frac{t}{\alpha_{rank}-2}]$ with step size s_k
  # The size relation between $r^{(1)}$ and $r^{(2)}$ is consistent
  # with the original channels, except $r^{(1)} = r^{(2)}$
  if $C_i \geq F_i$ **then**
    $m \leftarrow \frac{C_i}{F_i}$
    $\{(r_{i,j}^{(1)}, r_{i,j}^{(2)})\} \leftarrow \{(r_j, \frac{r_j}{m})\} \cup \{(r_j, r_j)\}$
  else
    $m \leftarrow \frac{F_i}{C_i}$
    $\{(r_{i,j}^{(1)}, r_{i,j}^{(2)})\} \leftarrow \{(\frac{r_j}{m}, r_j)\} \cup \{(r_j, r_j)\}$
  end if
end for